Encyclopedia of
Hydrological Sciences

Encyclopedia of
Hydrological Sciences

Encyclopedia of
Hydrological Sciences

Editor-in-Chief

Malcolm G Anderson
Department of Geography, University of Bristol, Bristol, UK

Senior Advisory Editor

Jeffrey J McDonnell
Department of Forest Engineering, Oregon State University, Corvallis, OR, US

1

WILEY

The Atrium,
Southern Gate,
Chichester,
West Sussex,
PO19 8SQ, England

Telephone: (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street,
Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street,
San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12,
D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street,
Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 5353 Dundas Street West, Suite 400,
Etobicoke, Ontario, Canada M9B 6HB

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of hydrological science/editor-in-chief, Malcolm G. Anderson.
p. cm.
ISBN 0-471-49103-9 (cased) – ISBN 0-470-84894-4 (obook)
1. Hydrology – Encyclopedias. I. Anderson, Malcolm G.
GB655.E527 2005
551.48′03 – dc22

2005022541

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library
ISBN 10: 0-471-49103-9 (HB)
ISBN 13: 978-0-471-49103-9

Typeset in 9½/11½ pt Times Roman by Laserwords Private Ltd., Chennai, India.
Printed and bound in Great Britain by Bath Press, Bath, UK.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Editorial Board

Contents

VOLUME 1

VOLUME 2

VOLUME 3

VOLUME 4

VOLUME 5

Contributors

Michael B Abbott
Knowledge Engineering B.V. Belgium and UNESCO-IHE Delft, Delft, The Netherlands

Richard G Allen
Department of Civil Engineering and Department of Biological and Agricultural Engineering, University of Idaho, Kimberly, ID, US

William M Alley
United States Geological Survey, Office of Ground Water, San Diego, CA, US

Doug Alsdorf
Department of Geological Sciences, Ohio State University, Columbus, OH, US

Mustafa S Altinakar
Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

Jaime M Amezaga
Centre for Land Use and Water Resources Research, Institute for Research on the Environment and Sustainability, University of Newcastle, Newcastle upon Tyne, UK

Ulf Andrae
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

George W Annandale
Engineering and Hydrosystems Inc., Denver, CO, US

Aronne Armanini
Department of Civil and Environmental Engineering, University of Trento, Trento, Italy

Geoffrey L Austin
Department of Physics, The University of Auckland, Auckland, New Zealand

Roni Avissar
Department of Civil and Environmental Engineering, Duke University, Durham, NC, US

Bruce Aylward
Deschutes Water Exchange Program, Deschutes Resources Conservancy, Bend, OR, US

Vladan Babovic
DHI Water & Environment, Agern Alle, Hørsholm, Denmark

Darren L Bade
Center for Limnology, University of Wisconsin, Madison, WI, US

Andy Baker
School of Geography, Earth and Environmental Sciences, The University of Birmingham, Birmingham, UK

Lawrence E Band
University of North Carolina, Chapel Hill, NC, US

András Bárdossy
Universität Stuttgart, Institut für Wasserbau, Stuttgart (Vaihingen), Germany

Luis A Bastidas
Department of Civil and Environmental Engineering/Utah Water Research Laboratory, Utah State University, Logan, UT, US

Paul D Bates
School of Geographical Sciences, University of Bristol, Bristol, UK

Jacob Bear
Department of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa, Israel

Roger Beckie
Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada

Anton Beljaars
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Kenneth E Bencala
United States Geological Survey, Menlo Park, CA, US

Sandra L Berry
Formerly of: Cooperative Research Centre for Greenhouse Accounting, Research School of Biological Sciences, Institute of Advanced Studies, The Australian National University, Canberra, Australia & Currently at: School of Resources, Environment and Society, Faculty of Science, The Australian National University, Canberra, Australia

Richard A Betts
Hadley Centre for Climate Prediction and Research, Exeter, UK

Keith J Beven
Department of Environmental Science, and Lancaster Environment Centre, Lancaster University, Lancaster, UK

Charon Birkett
Earth Science Interdisciplinary Center, University of Maryland, College Park, MD, US

David M Bjerklie
United States Geological Survey, Hartford, CT, US

Günter Blöschl
Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Vienna, Austria

Axel Bronstert
Institute for Geoecology, University of Potsdam, Potsdam, Germany

James D Brown
Universiteit van Amsterdam, Nieuwe Achtergracht, Amsterdam, The Netherlands

Ross D Brown
Meteorological Service of Canada, Dorval, QC, Canada

LA Sampurno Bruijnzeel
Department of Hydrology and Geo-Environmental Sciences, Vrije Universiteit, Amsterdam, The Netherlands

Reto Burkard
University of Bern, Bern, Switzerland

Timothy P Burt
Department of Geography, University of Durham, Durham, UK

Jim M Buttle
Department of Geography, Trent University, Peterborough, ON, Canada

Ian R Calder
Centre for Land Use & Water Resources Research, University of Newcastle, Newcastle upon Tyne, UK

Terri Camesano
Department of Chemical Engineering, Worcester Polytechnic Institute, Worcester, MA, US

Claire L Campbell
Centre for Ecology and Hydrology, Edinburgh, UK

Pamela L Campbell
ETI, St Petersburg, FL, US

Liz Chalk
Environment Agency, North East Region, York, UK

Alfred TC Chang
NASA/Goddard Space Flight Center, Greenbelt, MD, US
†Deceased May 26, 2004

Deborah V Chapman
Environmental Research Institute and Department of Zoology, Ecology and Plant Science, University College Cork, Cork, Ireland

Thomas N Chase
Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, US

Limin Chen
Civil and Environmental Engineering Department, Syracuse University, Syracuse, NY, US

Bernard Chocat
INSA de Lyon, URGC Hydrologie Urbaine, Villeurbanne Cedex, Lyon, France

Robin T Clarke
Instituto de Pesquisas Hidráulicas, Porto Alegre RS, Brazil

T Prabhakar Clement
Department of Civil Engineering, Auburn University, Auburn, AL, US

Cory C Cleveland
Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, US

J Graham Cogley
Department of Geography, Trent University, Peterborough, ON, Canada

Frederick S Colwell
Biological Sciences Division, Idaho National Laboratory, Idaho Falls, ID, US

Josefino C Comiso
NASA/Goddard Space Flight Center, Greenbelt, MD, US

Dennis L Corwin
USDA-ARS, George E. Brown, Jr. Salinity Laboratory, Riverside, CA, US

Justin F Costelloe
CRC for Catchment Hydrology and Department of Civil and Environmental Engineering, The University of Melbourne, Victoria, Australia

Alan P Covich
Institute of Ecology, University of Georgia, Athens, GA, US

Christopher B Craft
School of Public and Environmental Affairs, Indiana University, Bloomington, IN, US

Susanne Crewell
Meteorological Institute, University of Munich, Munich, Germany

Jacob H Dane
Department of Agronomy and Soils, Auburn University, Auburn, AL, US

Lisa S Darby
NOAA Environmental Technology Laboratory, Boulder, CO, US

Jorge Soares David
Instituto Superior de Agronomia, Tapada da Ajuda, Lisboa, Portugal

Bruce Davison
Environment Canada, National Hydrology Research Centre, National Water Research Institute, Saskatoon, SK, Canada

John Dearing
Department of Geography, University of Liverpool, Liverpool, UK

William E Dietrich
Department of Earth and Planetary Science, University of California, Berkeley, CA, US

S Lawrence Dingman
Earth Sciences Department, University of New Hampshire, Durham, NH, US

Kim N Dirks
Department of Physiology, The University of Auckland, Auckland, New Zealand

Stefan H Doerr
Geography Department, University of Wales, Swansea, UK

Shawan Dogramaci
Department of Environment, Salinity and Land Use Impacts Branch, Resource Science Division, East Perth, Australia

Albertus Johannes Dolman
Department of Hydrology and Geo-Environmental Sciences, Faculty of Earth and Life Sciences, Vrije Universiteit, Amsterdam, The Netherlands

Brendan Dooher
Environmental Restoration Division, Lawrence Livermore National Laboratory, Livermore, CA, US

Charles A Doswell III
University of Oklahoma, Norman, OK, US

Charles T Driscoll
Civil and Environmental Engineering Department, Syracuse University, Syracuse, NY, US

Ian G Droppo
National Water Research Institute, Environment Canada, Burlington, ON, Canada

Matthias Drusch
European Centre for Medium-Range Weather Forecasts, Reading, UK

Ralph Dubayah
Department of Geography, University of Maryland, College Park, MD, US

Wolfgang Durner
Institute of Geoecology, Department of Soil Physics, Braunschweig Technical University, Braunschweig, Germany

S Eden
US Climate Change Science Program Office, Washington, DC, US

Anthony C Edwards
Formerly of: Catchment Management Group, Macaulay Institute, Aberdeen, UK & Currently at: Nether Backhill, Aberdeen, UK

William E Eichinger
IIHR Hydroscience and Engineering, University of Iowa, Iowa City, IA, US

Michael B Ek
National Centers for Environmental Prediction/Environmental Modeling Center, Suitland, MD, US

Brenda Ekwurzel
Global Environment Program, Union of Concerned Scientists, Washington DC, US

J Bryan Ellis
Urban Pollution Research Centre, Middlesex University, Enfield, UK

Theodore A Endreny
Program in Hydrological Systems Science & Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY, US

Anthony W England
University of Michigan, Ann Arbor, MI, US

Alan Ervine
Department of Civil Engineering, University of Glasgow, Glasgow, UK

Werner Eugster
Swiss Federal Institute of Technology, Zürich, Switzerland

Souheil M Ezzedine
University of California, Lawrence Livermore National Laboratory, Livermore, CA, US

Roger A Falconer
Cardiff School of Engineering, Cardiff University, Cardiff, UK

Kathy Fallon-Lambert
Ecologic: Analysis and Communications, Quechee, VT, US

Graham D Farquhar
Cooperative Research Centre for Greenhouse Accounting, Research School of Biological Sciences, The Australian National University, Canberra, Australia

Ty PA Ferré
University of Arizona, Tucson, AZ, US

Massimiliano Ferronato
University of Padova, Padova, Italy

Jonathan W Finch
Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

Andrea I Flossmann
Laboratoire de Météorologie Physique/OPGC, Université Blaise Pascal/CNRS, Aubière Cedex, France

Hannes Flühler
Institute of Terrestrial Ecology, Swiss Federal Technical University at Zürich, Schlieren, Switzerland

Tim Forsyth
Development Studies Institute, London School of Economics and Political Science, London, UK

James L Foster
NASA/Goddard Space Flight Center, Greenbelt, MD, US

Luigi Fraccarollo
CUDAM and Department of Civil and Environmental Engineering, University of Trento, Trento, Italy

Janet Franklin
San Diego State University, San Diego, CA, US

James K Fredrickson
Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, US

Sherilyn C Fritz
Department of Geosciences & School of Biological Sciences, University of Nebraska, Lincoln, NE, US

John Gallant
Butler Laboratory, Commonwealth Scientific and Industrial Research Organisation, Land and Water, Canberra, Australia

Giuseppe Gambolati
University of Padova, Padova, Italy

John HC Gash
Centre for Ecology & Hydrology, Wallingford, Oxfordshire, UK

Charles P Gerba
Department of Soil, Water and Environmental Science, University of Arizona, Tucson, AZ, US

Thomas W Giambelluca
Geography Department, University of Hawaii at Manoa, Honolulu, HI, US

Timothy R Ginn
Department of Civil and Environmental Engineering, University of California, Davis, CA, US

Christopher J Gippel
Fluvial Systems Pty Ltd, Stockton, Australia

Scott J Goetz
Woods Hole Research Center, Woods Hole, MA, US

Jaime Gómez-Hernández
Department of Hydraulics and Environmental Engineering, Universidad Politécnica de Valencia, Valencia, Spain

Barry E Goodison
Meteorological Service of Canada, Downsview, Canada

David C Goodrich
Southwest Watershed Research Center, ARS USDA, Tucson, AZ, US

Lars Gottschalk
Department of Geosciences, University of Oslo, Oslo, Norway

Walter H Graf
Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

Rodger B Grayson
CRC for Catchment Hydrology and Department of Civil and Environmental Engineering, The University of Melbourne, Victoria, Australia

Vincent Guinot
Hydrosciences Montpellier, Université Montpellier 2, Montpellier Cedex 5, France

Hoshin V Gupta
Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, US

Dorothy K Hall
NASA/Goddard Space Flight Center, Greenbelt, MD, US

Michael J Hall
Department of Water Engineering, UNESCO-IHE, Delft, The Netherlands

Robin L Hall
Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

John Hallett
Division of Atmospheric Sciences, Desert Research Institute, Reno, NV, US

Matthew C Hansen
Geographic Information Science Center of Excellence, South Dakota State University, Brookings, SD, US

R Michael Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, US

David Harding
Geodynamics Branch, NASA/Goddard Space Flight Center, Greenbelt, MD, US

Holly C Hartmann
Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, US

Bent Hasholt
University of Copenhagen, Copenhagen, Denmark

Richard W Healy
United States Geological Survey, Denver, CO, US

Marius Heinen
Alterra, Wageningen, The Netherlands

Henry Hengeveld
Meteorological Service of Canada, Toronto, ON, Canada

Gerard BM Heuvelink
Wageningen University, Wageningen, The Netherlands

Larry D Hinzman
Water and Environmental Research Center, Institute of Northern Engineering, University of Alaska, Fairbanks, AK, US

Regine Hock
Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden

Forrest Hoffman
Oak Ridge National Laboratory, Oak Ridge, TN, US

Fred Hoffman
Environmental Restoration Division, Lawrence Livermore National Laboratory, Livermore, CA, US

Albert AM Holtslag
Meteorology and Air Quality Wageningen University, Wageningen, The Netherlands

Ekkehard Holzbecher
Humboldt Universität, Inst. of Freshwater Ecology (IGB), Berlin, Germany

Jan W Hopmans
Department of Land, Air and Water Resources, University of California, Davis, CA, US

Bryn Hubbard
Institute of Geography and Earth Sciences, University of Wales, Aberystwyth, UK

Alfredo R Huete
Department of Soil, Water and Environmental Science, University of Arizona, Tucson, AZ, US

George J Huffman
Science Systems and Applications, Inc. and NASA/Goddard Space Flight Center Laboratory for Atmospheres, Greenbelt, MD, US

George A Isaac
Cloud Physics and Severe Weather Research Division, Meteorological Service of Canada, Toronto, ON, Canada

Thomas J Jackson
USDA ARS Hydrology and Remote Sensing Lab, Beltsville, MD, US

Peter Jansson
Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden

Mathieu Javaux
Department of Environmental Sciences and Land Use Planning, Université Catholique de Louvain, Louvain-la-Neuve, Belgium and Agrosphere Institute, (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Graham Jewitt
School of Bioresources Engineering and Environmental Hydrology, University of KwaZulu-Natal, Pietermaritzburg, South Africa

H Gerald Jones
INRS-ETE, Environnement, Université du Quebec, Sainte-Foy, QC, Canada

Julia Jones
Department of Geosciences, Oregon State University, Corvallis, OR, US

Andreja Jonoski
Department of Hydroinformatics and Knowledge Management, UNESCO-IHE Institute for Water Education, Delft, The Netherlands

Per Kallberg
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

George Kallos
Atmospheric Modeling and Weather Forecasting Group, School of Physics, University of Athens, Athens, Greece

Douglas L Kane
Water and Environmental Research Center, Institute of Northern Engineering, University of Alaska, Fairbanks, AK, US

Maarten Keijzer
DHI Water & Environment, Agern Alle, Hørsholm, Denmark

Richard EJ Kelly
University of Maryland, Baltimore, MD, US

Andreas Kemna
Agrosphere Institute (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Soon-Thiam Khu
Centre for Water Systems, School of Engineering, Computer Science and Mathematics, University of Exeter, Exeter, UK

John S Kimball
Department of Ecosystem and Conservation Sciences, NTSG University of Montana, Missoula, MT, US and Flathead Lake Biological Station, Division of Biological Sciences, The University of Montana, Polson, MT, US and Numerical Terradynamic Simulation Group, The University of Montana, Missoula, MT, US

Mike Kirkby
School of Geography, University of Leeds, Leeds, UK

Mary Beth Kirkham
Department of Agronomy, Kansas State University, Manhattan, KS, US

Randal D Koster
NASA/Goddard Space Flight Center, Greenbelt, MD, US

Bommanna G Krishnappan
Aquatic Ecosystem Impacts Research Branch, National Water Research Institute, Environment Canada, Canada Centre for Inland Waters, Burlington, ON, Canada

Charles Kroll
Environmental Resources and Forest Engineering, SUNY College of Environmental Science and Forestry (ESF), Syracuse, NY, US

Jacob Cornelis Jan Kwadijk
WL Delft Hydraulics, Delft, The Netherlands

James W La Baugh
United States Geological Survey, Office of Ground Water, Reston, VA, US

Rob Lamb
JBA Consulting – Engineers & Scientists, Skipton, North Yorkshire, UK

Luca G Lanza
Department of Environmental Engineering, University of Genoa, Genoa, Italy

Michele Larcher
CUDAM and Department of Civil and Environmental Engineering, University of Trento, Trento, Italy

Andrew RG Large
Centre for Land Use and Water Resources Research, University of Newcastle, Newcastle upon Tyne, UK

RG Lawford
International GEWEX Project Office, Silver Spring, MD, US

George Leavesley
United States Geological Survey, Denver, CO, US

Graham Leeks
Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

Michael Lehning
WSL, Swiss Federal Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland

Dennis P Lettenmaier
Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, US

Gro Lilbæk
Centre for Hydrology, Department of Geography, University of Saskatchewan, Saskatoon, SK, Canada

Binliang Lin
Cardiff School of Engineering, Cardiff University, Cardiff, UK

Harry F Lins
United States Geological Survey, Office of Ground Water, Reston, VA, US

Kai Lipsius
Institute of Geoecology, Department of Soil Physics, Braunschweig Technical University, Braunschweig, Germany

Yongqiang Liu
Forestry Sciences Laboratory, USDA Forest Service, Athens, GA, US

Keith Loague
Department of Geological and Environmental Sciences, Stanford University, Stanford, CA, US

David Long
Electrical and Computer Engineering Department, Brigham Young University, Provo, UT, US

Charles H Luce
Rocky Mountain Research Station, Boise, ID, US

Henrik Madsen
River & Flood Management, DHI Water & Environment, Hørsholm, Denmark

David R Maidment
Department of Civil Engineering, University of Texas, Austin, TX, US

Jiri Marsalek
National Water Research Institute, Burlington, ON, Canada

Philip Marsh
NWRI Saskatoon, Saskatchewan, SK, Canada

G Richard Marzolf
United States Geological Survey, Office of Ground Water, Reston, VA, US

Michael D Mastrandrea
Interdisciplinary Program in Environment and Resources, Stanford University, Stanford, CA, US

Joseph R McConnell
Division of Hydrologic Sciences, Desert Research Institute, University and Community College System of Nevada, Reno, NV, US

Steven C McCutcheon
Faculty of Engineering and Warnell School of Forest Resources, University of Georgia on assignment from the US EPA National Exposure Research Laboratory, Athens, GA, US

Kyle C McDonald
Water and Carbon Cycles Group, Science Division, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, US

Jeffrey J McDonnell
Department of Forest Engineering, Oregon State University, Corvallis, OR, US

Brian McGlynn
Department of Land Resources and Environmental Sciences, Watershed Hydrology Laboratory, Montana State University, Bozeman, MT, US

Ian G McKendry
Department of Geography, The University of British Columbia, Vancouver, BC, Canada

James P McKinley
Chemical Sciences Division, Pacific Northwest National Laboratory, Richland, WA, US

Diane M McKnight
Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, US

Walt McNab
Environmental Restoration Division, Lawrence Livermore National Laboratory, Livermore, CA, US

David Meko
Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ, US

Guillermo F Mendoza
New York Department of Environmental Protection, Flushing, NY, US

Michel Meybeck
Sisyphe/CNRS, University of Paris, Paris, France

Hans Middelkoop
Department of Physical Geography, Utrecht University, The Netherlands

Anthony W Minns
Marine & Coastal Management, WL Delft Hydraulics, Delft, The Netherlands

Glenn E Moglen
Department of Civil and Environmental Engineering, University of Maryland, College Park, MD, US

Hans-Martin Münch
Agrosphere Institute (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Arthur E Mynett
WL Delft Hydraulics & UNESCO-IHE and Delft University of Technology, Delft, The Netherlands

Mark Nearing
Southwest Watershed Research Center, Tucson, AZ, US

Kirk E Nelson
Department of Civil and Environmental Engineering, University of California, Davis, CA, US

Malcolm D Newson
School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne, UK

Mary Nichols
Southwest Watershed Research Center, Tucson, AZ, US

Peter Nienow
Institute of Geography, University of Edinburgh, Edinburgh, UK

Bart Nijssen
Departments of Hydrology and Water Resources/Civil Engineering and Engineering Mechanics, University of Arizona, Tucson, AZ, US

John R Nimmo
United States Geological Survey, Menlo Park, CA, US

Robert Oglesby
NASA/Marshall Space Flight Center, Huntsville, AL, US

Atsumu Ohmura
Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

J Philip O'Kane
Environment Institute, National University of Ireland, Cork, Ireland

Taikan Oki
International Center for Urban Safety Engineering, Institute of Industrial Science, University of Tokyo, Tokyo, Japan

Dani Or
Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, US

Thomas C Pagano
Department of Agriculture, Natural Resources Conservation Service, Portland, OR, US

Anthony John Parsons
Department of Geography, University of Leicester, Leicester, UK

Gareth Pender
School of the Built Environment, Heriot–Watt University, Edinburgh, UK

Norman E Peters
United States Geological Survey, Georgia Water Science Center, Atlanta, GA, US

Birgit Peterson
Department of Geography, University of Maryland, College Park, MD, US

Roger A Pielke Sr
Department of Atmospheric Science, Colorado State University, Fort Collins, CO, US

Alain Pietroniro
Environment Canada, National Hydrology Research Centre, National Water Research Institute, Saskatoon, SK, Canada

Rachel T Pinker
Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, US

John W Pomeroy
Centre for Hydrology, Department of Geography, University of Saskatchewan, Saskatoon, SK, Canada

Rajiv Prasad
Environmental Technology Division, Pacific Northwest National Laboratory, Richland, WA, US

Terry D Prowse
Department of Geography, National Water Research Institute of Environment Canada, University of Victoria, Victoria, BC, Canada

Ioannis Pytharoulis
Atmospheric Modeling and Weather Forecasting Group, School of Physics, University of Athens, Athens, Greece

Peter A C Raats
Wageningen University and Research Centre, Wageningen, The Netherlands

Paolo Reggiani
Inland Water Systems, Foundation Delft Hydraulics, Delft, The Netherlands

Thomas E Reilly
United States Geological Survey, Office of Ground Water, Reston, VA, US

Ken Renard
Southwest Watershed Research Center, Tucson, AZ, US

Philippe Renard
Centre for Hydrogeology, University of Neuchâtel, Neuchâtel, Switzerland

Joshua Rhoads
Department of Geography, University of Maryland, College Park, MD, US

Jeffrey E Richey
School of Oceanography, University of Washington, Seattle, WA, US

Jerry C Ritchie
United States Department of Agriculture, Agricultural Research Service Hydrology and Remote Sensing Laboratory, Beltsville, MD, US

Professor Coen J Ritsema
Soil Science Center, Wageningen, The Netherlands

John Roads
Scripps Institution of Oceanography, UCSD La Jolla, CA, US

John Roberts
Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

Dale M Robertson
United States Geological Survey, Water Science Center, Middleton, WI, US

Franklin R Robertson
NASA/Marshall Space Flight Center, Huntsville, AL, US

Mark Robinson
Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

Michael L Roderick
Cooperative Research Centre for Greenhouse Accounting, Research School of Biological Sciences, The Australian National University, Canberra, Australia

Dan Rosbjerg
Environment & Resources DTU, Technical University Denmark, Kongens Lyngby, Denmark

Steven W Running
Department of Ecosystem and Conservation Sciences, NTSG University of Montana, Missoula, MT, US

John Ruprecht
Department of Environment, Salinity and Land Use Impacts Branch, Resource Science Division, East Perth, Australia

Jasmine E Saros
Department of Biology, University of Wisconsin, La Crosse, WI, US

Dragan Savic
Centre for Water Systems, School of Engineering, Computer Science and Mathematics, University of Exeter, Exeter, UK

Ted A Scambos
National Snow and Ice Data Center (NSIDC), University of Colorado, Boulder, CO, US

Marcel G Schaap
Department of Environmental Sciences, University of California, Riverside, CA, US

Timothy D Scheibe
Environmental Technology Directorate, Pacific Northwest National Laboratory, Richland, WA, US

Jaap Schellekens
Inland Water Systems, Foundation Delft Hydraulics, Delft, The Netherlands and Earth and Life Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

C Adam Schlosser
Joint Program on the Science and Policy of Global Change Massachusetts Institute of Technology, Cambridge, MA, US

Thomas J Schmugge
Formerly of: USDA/ARS Hydrology & Remote Sensing Lab, Beltsville, MD, US & Currently at: College of Agriculture, New Mexico State University, Las Cruces, NM, US

Stephen H Schneider
Department of Biological Sciences, Stanford University, Stanford, CA, US

Gerrit Schoups
Department of Land, Air and Water Resources, University of California, Davis, CA, US

Jan Seibert
Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden

Boris Sevruk
Institut für Atmosphäre und Klima (ETH), Zürich, Switzerland

Martin Sharp
Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, AB, Canada

Chris A Shuman
NASA/Goddard Space Flight Center, Greenbelt, MD, US

W James Shuttleworth
Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ, US

Martin J Siegert
Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol, UK

Adrian Simmons
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Jirka Šimůnek
Department of Environmental Sciences, University of California, Riverside, CA, US

Murugesu Sivapalan
Formerly of: Centre for Water Research, Department of Environmental Engineering, University of Western Australia, Crawley, Australia & Currently at: Department of Geography, University of Illinois, Urbana, IL, US

Michael C Slattery
Institute of Environmental Studies, Texas Christian University, Fort Worth, TX, US

David Smiles
CSIRO Land and Water, Canberra, Australia

Roger E Smith
Civil Engineering Department, Colorado State University, Fort Collins, CO, US

Dimitri P Solomatine
UNESCO-IHE Institute for Water Education, Delft, The Netherlands

Shaul Sorek
Ben Gurion University of the Negev, J. Blaustein Institutes for Desert Research, Sede Boker, Israel

Soroosh Sorooshian
Department of Civil and Environment Engineering, University of California, Irvine, CA, US

Manfred Stähli
Swiss Federal Research Institute WSL; Water, Soil, and Rock Movements, Birmensdorf, Switzerland

Fritz Stauffer
Swiss Federal Institute of Technology, Zurich, Switzerland

Matthias Steiner
Princeton University, Princeton, NJ, US

Guus S Stelling
Department of Civil Engineering, Technical University of Delft, Delft, The Netherlands

Grahame Stephens
Atmospheric Sciences, Colorado State University, Fort Collins, CO, US

David A Stonestrom
United States Geological Survey, Menlo Park, CA, US

Hans von Storch
Institute of Coastal Research, GKSS Research Centre, Geesthacht, Germany

Julienne Stroeve
National Snow and Ice Data Center, University of Colorado, Boulder, CO, US

Zhongbo Su
International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands

Alexander Y Sun
CNWRA, Southwest Research Institute, San Antonio, TX, US

Ne-Zheng Sun
Department of Civil and Environmental Engineering, University of California, Los Angeles, CA, US

Peter W Swarzenski
United States Geological Survey, St Petersburg, FL, US

Christina L Tague
San Diego State University, San Diego, CA, US

Pietro Teatini
University of Padova, Padova, Italy

Axel Tillmann
Agrosphere Institute (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Ezio Todini
Department of Earth and Geo-Environmental Sciences, University of Bologna, Bologna, Italy

Sylvia S Tognetti
Consultant, Environmental Science and Policy

G Clarke Topp
Agriculture and Agri-Food Canada, Ottawa, ON, Canada

Alan R Townsend
Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, US

Martyn Tranter
Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol, UK

Ilja Tromp-van Meerveld
Ecole Polytechnique Fédérale de Lausanne, School of Architecture, Civil & Environmental Engineering, Lausanne, Switzerland

Taro Uchida
Research Center for Disaster Risk Management, National Institute for Land & Infrastructure Management, Asahi, Tsukuba, Japan

Michael H Unsworth
Atmospheric Sciences Group, College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, US

Sakari Uppala
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Fernanda Valente
Instituto Superior de Agronomia, Tapada da Ajuda, Lisboa, Portugal

Marnik Vanclooster
Department of Environmental Sciences and Land Use Planning, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

J Vanderborght
Agrosphere Institute, (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Harry Vereecken
Agrosphere Institute (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Anne Verhoef
Department of Soil Science, The University of Reading, Reading, UK

Arre Verweerd
Agrosphere Institute (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Adri Verwey
WL Delft Hydraulics, Delft, The Netherlands

Pedro Viterbo
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Huib J de Vriend
WL Delft Hydraulics, Delft, The Netherlands

Jasper A Vrugt
Formerly of: Institute for Biodiversity and Ecosystem Dynamics – Physical Geography, University of Amsterdam, Amsterdam, The Netherlands & Currently at: Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, US

Thorsten Wagener
Formerly of: Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, US & Currently at: Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, US

Desmond E Walling
Department of Geography, University of Exeter, Exeter, UK

Markus Weiler
Departments of Forest Resources Management and Geography, University of British Columbia, Vancouver, BC, Canada

Micha Guido Franciscus Werner
WL Delft Hydraulics, Delft, The Netherlands

Andrew W Western
CRC for Catchment Hydrology and Department of Civil and Environmental Engineering, The University of Melbourne, Victoria, Australia

Robert G Wetzel
Department of Environmental Sciences and Engineering, The University of North Carolina, Chapel Hill, NC, US
† Deceased April 18, 2005

Sue White
Institute of Water and Environment, Cranfield University, Silsoe, Bedfordshire, UK

Mark Wigmosta
Environmental Technology Division, Pacific Northwest National Laboratory, Richland, WA, US

Peter de Willigen
Alterra, Wageningen, The Netherlands

Ian Willis
Department of Geography, University of Cambridge, Cambridge, UK

Ming-ko Woo

School of Geography and Geology, McMaster University, Hamilton, ON, Canada

Ross Woods

Catchment Processes and Water Resources Group, National Institute of Water and Atmospheric Research, Christchurch, New Zealand

Brian D Wood

Environmental Engineering, Oregon State University, Corvallis, OR, US

Peter C Young

Centre for Research on Environmental Systems and Statistics, Lancaster University, Lancaster, UK and Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia

Paul V Zimba

United States Department of Agriculture, Agricultural Research Service Catfish Genetics Research Unit, Stoneville, MS, US

Preface

The field of hydrological science deals with the occurrence, distribution, movement, and properties of water on the earth. The science of hydrology holds a unique and central place in the field of earth system science, intimately linked with other water-related disciplines such as meteorology, climatology, geomorphology, hydrogeology, and ecology. Beyond basic scientific interest, water quantity and water quality have become two of the most pressing environmental issues of our time. The first comprehensive hydrological studies began in the late 1600s with Pierre Perrault's field studies of the hydrological cycle and Edmund Halley's experiments on evaporation. However, it was not until the mid-1850s that Henry Darcy quantified the hydraulics of groundwater flow and the linear relation between velocity and hydraulic gradient. Since then, and especially through the latter twentieth century, knowledge generation in the hydrological sciences exploded with new discoveries in each of the components of the hydrological cycle. Despite scientific and technological advances, the field of hydrological sciences has been highly fragmented across the engineering-science interface, between basic and applied studies, mathematical and descriptive work, and field and laboratory investigations. Such fragmentation has left many aspects of the hydrological cycle and its underlying mechanics still poorly understood. Such understanding is an essential prerequisite for hydrological prediction and for defining the level of uncertainty around these predictions.

The Encyclopedia of Hydrological Sciences (EHS) is the first definitive research-level multivolume encyclopedia in the hydrological sciences where the full scope of research in the field of hydrological sciences is distilled – from engineering approaches to basic science, from process studies to mathematical modeling, and from field investigations to the laboratory. EHS thus provides an inclusive reference source for the field, defining what we know, what we think we know, and what we need to know about the future. EHS brings together information on hydrological processes from the sub-catchment to the global scale with a focus on research-level analysis.

We have organized the encyclopedia into 17 themed parts that cover the breadth of current major research activity in the field. The 203 articles in EHS have been written by 316 authors from 22 countries, all experts in their fields. Each article has been peer-reviewed and then edited by an Associate Editor, each among the top scientists in their sub-disciplines. We have tried to organize EHS to maximize ease of use for its readers. Each of the five volumes contains a table of contents. The last volume contains a subject index pertaining to the complete work.

The accompanying online version of the Encyclopedia includes hypertext links to appropriate software sites. The online format extends the usefulness of EHS through rapid search capability, and facilitates both updating and subsequent expansion of text material. The hard copy version of EHS is a foundational set of knowledge that will be progressively tuned in the online version, both tracking the core science foci and delivering the latest research findings.

We owe a considerable debt to the author and editor team, colleagues who are among the most active research workers in their respective fields. The Associate Editors of the 17 themed parts have managed both procurement of articles and the substantive review process that such a project demands. Such activity has been supported throughout the project life by staff at John Wiley; in particular, we acknowledge the support given to us by Sally Wilkinson, who has overseen the project from conception, whilst Sue Amesbury has given unfailing support to ourselves, the Associate Editors, and authors. Her contribution has been pivotal in ensuring timely delivery of the project.

We hope that EHS will provide both a sound comprehensive platform of current knowledge and also clear pointers to future research directions in the field.

Malcolm G Anderson
Editor-in-Chief

Jeffrey J McDonnell
Senior Advisory Editor

Abbreviations and Acronyms

1-NAA	1-Naphthylacetic Acid
2-D	Two-Dimensional
3-D	Three-Dimensional
3DVAR	Three-Dimensional Variational Method
3G	Third-Generation
4DDA	4-Dimensional Data Assimilation
AABW	Antarctic Bottom Water
ABA	Abscisic Acid
ABL	Atmospheric Boundary Layer
ABRACOS	Anglo-Brazilian Climate Observation Study
ACLs	Agent Communication Languages
ACPs	Atmospheric Circulation Patterns
ACSYS	Arctic Climate System Study
AD	Automatic Differentiation
ADC	Areal Distribution Curve
ADCP	Acoustic Doppler Current Profiler
ADCs	Analog-to-Digital Converters
ADEOS	Advanced Environmental Observation Satellite
ADV	Acoustic Doppler Velocimeters
ADZ	Aggregate Dead Zone
AEM	Airborne Electromagnetic Systems
AERI	Atmospheric Emitted Radiance Interferometer
AET	Actual Evapotranspiration
AFI	Antecedent Flow Index
AFM	Atomic Force Microscopy
AFORISM	EU Funded Flood Forecasting Research Project
AFOSM	Advanced First-Order Second-Moment
AGBM	Above-Ground Biomass
AGCMs	Atmospheric General Circulation Models
AGNPS	Agricultural Nonpoint Source
AI	Artificial Intelligence
AIC	Akaike's Information Criterion
AIRS	Atmospheric Infrared Sounder
ALMA	Assistance for Land Surface Modelling Activities
ALOS	Advanced Land Observing System
ALW	Earth Life Sciences and Research Council
AMI	Active Microwave Instrument
AMIP	Atmospheric Model Intercomparison Project
AMMA	African Monsoon Multidisciplinary Analysis
AMO	Atlantic Multidecadal Oscillation
AMS	Accelerator Mass Spectrometry
AMSR	Advanced Microwave Scanning Radiometer
AMSR-E	Advanced Microwave Scanning Radiometer – Earth Observing System
AMSU	Advanced Microwave Sounding Unit
AMV	Active Mixing Volume
ANC	Acid-Neutralizing Capacity
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ANPP	Aboveground NPP
ANSWERS	Areal Nonpoint Source Watershed Environmental Response Simulation
AO	Arctic Oscillation
AOSIS	Alliance of Small Island States
APAR	Absorbed PAR
APN	Asia–Pacific Network for Global Change Research
ARF	Area Reduction Factor
ARIMA	Auto Regressive Integrated Moving Average
ARM	Atmospheric Radiation Measurement
ARMA	Auto-Regressive Moving Average
ARMAX	Auto-Regressive Moving Average Exogenous
ARM-SGP	Atmospheric Radiation Measurement Southern Great Plains
ARPS	Advanced Regional Prediction System
ARS	Agricultural Research Service
ARX	Autoregressive Exogenous Variables Models
ASA	Aggregated Simulation Areas
ASAR	Advanced Synthetic Aperture Radar
ASB	Alternatives to Slash and Burn
ASCE	American Society of Civil Engineers
ASDC	Atmospheric Sciences Data Center
ASL	Atmospheric Surface Layer
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
ATOVS	Advanced TIROS Operational Vertical Sounder
ATSR	Along Track Scanning Radiometer
AVHRR	Advanced Very High Resolution Radiometer

AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
BAHC	Biological Aspects of the Hydrologic Cycle
BALTEX	The Baltic Sea Experiment
BARE	Bayesian Recursive Estimation
BAS	Bulk Atmospheric Similarity
BASINS	Better Assessment Science Integrating Point and Nonpoint Sources
BATEA	Bayesian Total Error Analysis
BATS	Biosphere Atmosphere Transfer Scheme
BBM	Building Block Methodology
BEA	Bureau of Economic Analysis
BFI	Baseflow Index
BIC	Bayesian Information Criterion
BLUE	Best Linear Unbiased Estimator
BMPs	Best Management Practices
BMBF	German Federal Ministry of Education and Research
BMWP	Biological Monitoring Working Party
BNF	Biological Nitrogen Fixation
BOA	Bottom of the Atmosphere
BOD	Biochemical/Biological Oxygen Demand
BOR	Bureau of Reclamation
BOREAS	Boreal Ecosystem–Atmosphere Study
BP	Before Present
BRDF	Bidirectional Reflectance Distribution Function
BREB	Bowen Ratio Energy Balance
BRF	Bidirectional Reflectance Factor
BSRN	Baseline Surface Radiation Network
CAAA	Amendments of the Clean Air Act
CALIPSO	Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations
CALM	Circumpolar Active Layer Monitoring
CAM	Crassulacean Acid Metabolism
CAP	Common Agricultural Policy
CARE	Conservation, Amenity, Recreation and Environment
CART	Cloud and Radiation Testbed
CASCC	Caltech Active Strand Cloudwater Collector
CASMM	Catchment Average Soil Moisture Monitoring
CATCH	Coupling of the Tropical Atmosphere and the Hydrological Cycle
CATHALAC	Regional Water Centre for the Humid Tropics of Latin America and the Caribbean
CAVE	CERES Program/Arm Validation Experiment
CBD	Convention on Biological Diversity
CBH	Canopy Base Height
CBIAC	Columbia Basin Interagency Committee
CCM2 or 3	Community Climate Model
CCN	Cloud Condensation Nuclei
CCRS	Canada Centre for Remote Sensing
CDAS	Climate Data Assimilation System
CDC	Center for Disease Control and Prevention
CDE	Convection Dispersion Equation
cdf	Cumulative Distribution Function
CDF	Cumulative Density Function
CDR	Climate-Data-Record
CEH	Centre for Ecology and Hydrology
CEOP	Coordinated Enhanced Observing Period
CEOS	Committee on Earth Observation Satellite
CERCLA	Comprehensive Emergency Response, Compensation, and Liability Act
CERES	Clouds and Earth's Radiant Energy System
CES	Conveyance Estimation System
CF	"Concentration" Factor
CFCs	Chlorofluorocarbons
CFD	Computational Fluid Dynamics
CFT	Colloid Filtration Theory
CGIAR	Consultative Group of International Agricultural Research
CGMS-IPWG	Coordination Group for Meteorological Satellites International Precipitation Working Group
CGSTAB	Conjugate Gradient Stabilized Method
CHAMP	Challenging Minisatellite Payload
CHP	Canopy Height Profile
CHR	Commission for the Hydrology of the River Rhine
CHy	Commission on Hydrology
CI	Capping Inversion
CIP	Classical Inverse Problem
CIPEL	Commission Internationale pour la Protection des Eaux du Léman Contre la Pollution
CISK	Convective Instability of the Second Kind
CIS	Central in Space
CK	Coefficient of Kurtosis
CL	Certification and Labeling
CLASS	Canadian Land Surface Scheme
CLASSIC	Climate and Land Use Scenario Simulation in Catchments
CLAWPACK	Conservation LAWs Package
CliC	Climate and Cryosphere
CLIMAP	Climate: Long-Range Investigation, Mapping, and Prediction

CLIPS	Climate Information and Prediction Services
CLIVAR	Climate Variability and Predictability
CLM	Community Land Model
CLPX	Cold Land Processes Experiment
CLS	Constrained Linear Systems
CLT	Convective Lognormal Transport
CMAP	Merged Analysis of Precipitation
CMC	Canopy Moisture Content
CMDL	Climate Monitoring and Diagnostics Laboratory
CMG	Climate-Modeling Grid
CMIS	Conical Scanning Microwave Imager/ Sounder
CMP	Common Midpoint
CN	Curve Number
CNES	Centre National d'Etudes Spatiales
COC	Colloidal Organic Carbon
COD	Chemical Oxygen Demand
CoD	Coefficient of Determination
COLE	Coefficient of Linear Extensibility
COMSTECH	Standing Committee on Scientific and Technological Cooperation
COPES	Coordinated Observing and Prediction of the Earth System
COUP	Coupled Heat and Mass Transfer
CPC	Climate Prediction Center
CPE	Cytopathogenic Effects
CPOM	Coarse Particulate Organic Matter
CPWC	Cooperative Programme on Water and Climate
CRIM	Complex Refractive Index Method
CRR	Conceptual Rainfall-Runoff
CRYSYS	Cryosphere System
CSA	Chinese Space Agency
CSD	Commission on Sustainable Development
CSEs	Continental Scale Experiments
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CSM	Climate System Model
CSO	Combined Sewer Overflow
CSS	Coarse Suspended Sediments
CS	Collection System
CT	Total Ice Concentration
CV	Coefficient of Variation
CW	Cloud Water
CWA	Clean Water Act
CWI	Catchment Wetness Index
CWSI	Crop Water Stress Index
DAAD	German Academic Exchange Service
DAI	Distributed Artificial Intelligence
DARX	Dynamic Autoregressive Exogenous
DAS	Department of Atmospheric Sciences
DCE	1,1-Dichloroethene
DCI	Deep Convective Index
DCIA	Directly Connected Impervious Area
DDF	Depth–Duration–Frequency
DDM	Data-Driven Modeling
DEM	Digital Elevation Map/Model
DES	De-Randomized Evolutionary Strategy
DFID	Department for International Development
DFIR	Double Fence Intercomparison Reference
DGPS	Differential Global Positioning System
DGVMs	Dynamic Global Vegetation Models
DHSVM	Distributed Hydrology Soil Vegetation Model
DIAL	Differential Absorption Lidar
DIAS	Dynamic Information Architecture System
DIC	Dissolved Inorganic Carbon
DIN	Deutsche Industrie-Normen
DIP	Dissolved Inorganic Phosphorus
DIRB	Dissimilatory Iron-Reducing Bacteria
DISORT	Discrete Ordinates Radiative Transfer
DLVO	Derjaguin Landau Verwey Overbeek
DM	Dry Moderate
DMRT	Dense Media Radiative Transfer
DMSP	Defense Meteorological Satellite Program
DNAPLs	Dense Nonaqueous Phase Liquids
DO	Dissolved Oxygen
DO-BOD	Dissolved Oxygen-Biochemical Oxygen Demand
DOC	Dissolved Organic Carbon
DOE	Department of Energy
DOM	Dissolved Organic Matter
DON	Dissolved Organic Nitrogen
DOP	Dissolved Organic P
DORIS	Doppler Orbitography and Radiopositioning Integrated by Satellite
DOW	Doppler-on-Wheels
DP	Dry Polar
DPHP	Dual-Probe Heat-Pulse
DRIFT	Downstream Response to Imposed Flow Transformation
DSA	Direct Solution Approach
DSDR	Direct Sampling Digital Radiometer
DSD	Drop-Size Distribution
DSI	Drought Severity Index
DT	Dry Tropical

DYNIA	Dynamic Identifiability Analysis
DYRESM	Dynamic Reservoir Simulation Model
EARLINET	European Aerosol Research Lidar Network
EAs	Evolutionary Algorithms
EASM	East Asia Summer Monsoon
EBBR	Energy Balance Bowen Ratio
EC	Eddy Covariance/Electrical Conductivity
ECA	Economic Commission for Africa
ECE	Economic Commission for Europe
ECLAC	The Economic Commission for Latin America
ECMWF	European Center for Medium-Range Weather Forecasts
EDAS	Eta Data Assimilation System
EDB	Ethylene Dibromide
EDC	Earth Resources Observation and Science Data Center
EDG	Earth Observing System Data Gateway
EDNA	Elevation Derivatives for National Application
EDS	Expanded Downscaling
EFFORTS	European Flood Forecasting Operational Real Time System
EFFS	European Flood Forecasting System
EHEC	Enterohemorrhagic *E. coli*
EIA	Environmental Impact Assessment
EIP	Extended Inverse Problem
EKF	Extended Kalman Filter
ELA	Equilibrium Line Altitude
EM	Electromagnetic
EMC	Event Mean Concentration
EM-DAT	Emergency Events Database
EMI	Electromagnetic Induction
EMICs	Earth-System Models of Intermediate Complexity
ENSO	El Niño Southern Oscillation
EOF	Empirical Orthogonal Function
EOS	Earth Observation System
EPA	Environmental Protection Agency
EPM	Equivalent Porous Medium
EPS	Extracellular Polymeric Substances
ERA-40	ECMWF 40-year ReAnalysis
ERBE	Earth Radiation Budget Experiment
ERS	European Remote Sensing
ERT	Electrical Resistivity Tomography
ES	Evolutionary Strategy
ESA	European Space Agency
ESCAP	Economic Commission for Asia and the Pacific
ESCAT	ERS-1/2 Scatterometer
ESCWA	Economic and Social Commission for West Asia
ESE	Earth Science Enterprise
ESMA	Explicit Soil Moisture Accounting
ESMF	Earth System Modeling Framework
ESMR	Electrically Scanning Microwave Radiometer
ESP	Ensemble Streamflow Prediction
ESRI	Environmental Systems Research Institute
ESSA	Environmental Science Services Administration
ESSP	Earth System Science Pathfinder
ESTAR	Electronically Scanned Thinned Array Radiometer
ET	Evapotranspiration
ETM+	Enhanced Thematic Mapper Plus
EU	European Union
EUMETSAT	European Meteorological Satellite Agency
EUROSEM	Kinematic Runoff and Erosion 2 Model
EUROTAS	European River Flood Occurrence and Total Risk Assessments System
EVI	Enhanced Vegetation Index
EVP	Elastic–Viscous–Plastic
EWT	Equivalent Water Thickness
EZ	Entrainment Zone
FAC	Facultative
FACU	Facultative Upland
FACW	Facultative Wet
FAO	Food and Agriculture Organisation
FAU	Formazin Attenuation Unit
FC	Fecal Coliform
FCF	Flood Channel Facility
FD	Flux Data
fdc	Flow Duration Curve
FDM	Finite Difference Method
FE	Finite Elements
FEF	Fernow Experimental Forest
FEH	Flood Estimation Handbook
FEM	Finite Element Method/Model
FF	False Alert
FFT	Fast Fourier Transform
FGV	Fraction of Green Vegetation
FHP	Foliar Height Profile
FIFE	First ISLSCP Field Experiment
FINA	Flow Impeding, Neutral, or Accelerating
FIPA	Foundation for Intelligent Physical Agents
FIR	Finite Impulse Response
FJC	"Freely Jointed Chain"
FM	Fogmonitor

FNMOC	Fleet Numerical Meteorological and Oceanographic Center
FONAFIFO	National Fund for Forest Financing
FONAG	Fondo Del Agua
FORM	First-Order Reliability Method
FP	Forward Problem
FPOM	Fine Particulate Organic Matter
FRAMES	Framework for Risk Analysis in Multimedia Environmental Systems
FRBS	Fuzzy Rule-Based Systems
FRIEND	Flow Regimes from International Experimental and Network Data
FSC	Fractional Snow Cover
FSR	Flood Studies Report
FSS	Fine Suspended Sediments
FSSP	Forward-Scattering Spectrometer Probe
FTCS	Forward-Time Centered Space
FTU	Formazin Turbidity Units
FV	Finite Volumes
FW	Flood Works
GA	Genetic Algorithm
GAC	Global Area Coverage
GAIM	Global Analysis, Integration and Modelling
GAME	GEWEX Asian Monsoon Experiment
GAPP	GEWEX Americas Prediction Project
GARNET	Global Applied Research Network
GARP	Global Atmospheric Research Program
GA–SQP	A Simple GA and Powell's Sequential Quadratic Programming
GATE	Atlantic Tropical Experiment
GCIP	GEWEX Continental-Scale International Project
GCM	General/Global Circulation/Climate Model
GCOS	Global Climate Observing System
GCTE	Global Change Terrestrial Ecosystems
GDRs	Geophysical Data Records
GEBA	Global Energy Balance Archive
GECAFS	Global Environmental Change and Food Systems
GEF	Global Environment Facility
GEMS	Global Environmental Monitoring System
GEO	Geostationary Earth Orbit
GEOS	Goddard Earth Observing System
GEOSS	Global Earth Observing System of Systems
GER	Global Environmental Research
GEWEX	Global Energy and Water Cycle Experiment
GF	Generic Framework
GFDL	Geophysical Fluid Dynamics Laboratory
GFO	Geosat Follow-on
GHCC	Global Hydrology and Climate Center
GHCN	Global Historical Climatology Network
GHG	Greenhouse Gas
GHP	GEWEX Hydrometeorology Panel
GHS	Geochemical Hydrograph Separation
GIP	Generalized Inverse Problem
GIS	Geographic Information System
GISS	Goddard Institute for Space Studies
GIUH	Geomorphologic Instantaneous Unit Hydrograph
GIWA	Global International Waters Assessment
GLACE	Global Land–Atmosphere Coupling Experiment
GLAS	Geoscience Laser Altimeter System
GLASS	Global Land Atmosphere System Study
GLDAS	Global Land Data Assimilation System
GLIMS	Glacier Land Ice Measurements from Space
GLM	Generalized Linear Models
GLOCOPH	Global Continental PalaeoHydrology
GLOFs	Glacier Lake Outburst Floods
GLOWA	Global Change in the Hydrologic Cycle
GLSE	Generalized Least Squares Estimator
GLUE	Generalized Likelihood Uncertainty Estimation
GMS	Geostationary Meteorological Satellite
GNIP	Global Network of Isotopes in Precipitation
GOALS	Global Ocean–Atmosphere–Land System
GODC	Godunov-Centered
GOES	Geostationary Operational Environmental Satellite
GOOS	Global Ocean Observing System
GOS	Global Observing System
GOSIC	Global Observing Systems Information Center
GP	Guelph Permeameter
GPCC	Global Precipitation Climatology Center
GPCP	Global Precipitation Climatology Project
GPD	Generalized Pareto Distribution
GPI	GOES Precipitation Index
GPM	Global Precipitation Measurement
GPP	Gross Primary Production Productivity
GPROF	Goddard Profiling Algorithm
GPR	Ground Penetrating Radar
GPS	Global Positioning System
GR	Gradient Ratios
GRACE	Gravity Recovery and Climate Experiment
GRDC	Global Runoff Data Center
GREEN	Global Rivers Environmental Education Network
GRF	General Radial-Flow

GRID	Global Resource Information Database
GRIP	Greenland Ice Core Program
GRP	Global Radiation Panel
GSFC	Goddard Space Flight Center
GsTL	Geostatistical Template Library
GSWP	Global Soil Wetness Project
GTF	Generalized Transfer Function
GTN-H	Global Terrestrial Network for Hydrology
GTOS	Global Terrestrial Observing System
GTS	Global Telecommunications System
GUH	Geomorphological Unit Hydrograph
GVaP	Global Water Vapor Project
GVMI	Global Vegetation Moisture Index
GWP	Global Water Partnership
GWS	Global Water System
GWSP	Global Water System Project
HA	Hydraulic Analysis
HAV	Hepatitis A Virus
HCDN	Hydro-Climatic Data Network
HCP	Horizontal Coplanar
HDR	Hot Dry Rock
HDS	Heat Dissipation Sensors
HELP	Hydrology for the Environment, Life and Policy
HEPEX	Hydrological Ensemble Prediction Experiment
HFBA	Hierarchical Foreground and Background Analysis
HG	Hydraulic Geometry
HGA	Hybrid Genetic Algorithm
HGM	Hydrogeomorphic
HH	Horizontal Transmit–Horizontal Receive
HIRS	High Resolution Infrared Sounder
HLA	Hydrological Landscape Analysis
HMC	Hybrid Metric-Conceptual
HMLE	Heteroscedastic Maximum Likelihood Estimator
HMW	High Molecular Weight
HMWB	Heavily Modified Water Body
HOF	Horton Overland Flow
HOME	Height of Median Energy
HOPC	Hydrology Observation Panel for Climate
HOST	Hydrology of Soil Types
HPV	Heat-Pulse Velocity
HRU	Hydrologic Response Unit
HSIA	Hydrologically Significant Impermeable Area
HSPF	Hydrologic Simulation Program – Fortran
HSRL	High Spectral Resolution Lidar
HSRS	Hydrosuction Removal System
HTLC	Height to Live Crown
HUS	Hemolytic-Uremic Syndrome
HWRP	Hydrology and Water Resources Programme
HYDROS	Hydrosphere State Mission
IAA	Indole-3-acetic Acid
IAD	Institutional Analysis and Development
IAEA	International Atomic Energy Agency
IAEH	International Association for Environmental Hydrology
IAH	International Association of Hydrogeologists
IAHR	International Association of Hydraulic Engineering and Research
IAHS	International Association of Hydrological Sciences
IAHS/ICSI	The International Association of Hydrological Sciences–International Commission on Snow and Ice
IAI	Inter-American Institute for Global Change Research
IAPs	Invasive Alien Plants
IARF	Infinite Acting Radial Flow
IASI	Infrared Atmospheric Sounding Interferometer
IASWS	International Association for Sediment Water Science
IBI	Index of Biological Integrity
IBL	Instance-Based Learning
ICCE	International Commission on Continental Erosion
ICCLAS	International Commission on the Coupled Land–Atmosphere System
ICCORES	International Coordinating Committee on Reservoir Sedimentation
ICESat	Ice, Cloud and Land Elevation Satellite
ICGW	International Commission on Groundwater
ICID	International Commission on Irrigation and Drainage
ICMS	Interactive Component Modeling System
ICOLD	International Commission on Large Dams
ICRCCM	Intercomparison of Radiation Codes in Climate Models
ICRS	International Commission on Remote Sensing
ICSI	International Commission on Snow and Ice
ICSU	International Council for Science

ICSW	International Commission on Surface Water
ICT	International Commission on Tracers
ICTs	Information and Communication Technologies
ICWE	International Conference on Water and the Environment
ICWQ	International Commission on Water Quality
ICWRS	International Commission on Water Resources Systems
ID	Internal Drainage
IDA	Infiltrated Depth Approximation
IDF	Intensity–Duration–Frequency
IDS	Interdisciplinary Science Projects
IEESAs	Integrated Economic and Environmental Satellite Accounts
IEP	Isoelectric Point
IETC	International Environmental Technology Center
IFIM	In-stream Flow Incremental Methodology
IFS	Integrated Forecasting System
IGBP	International Geosphere–Biosphere Programme
IGOS	Integrated Global Observing Strategy
IGOS-P	Integrated Global Observing Strategy – Partnership
IGRAC	International Groundwater Resource Assessment Center
IGWCO	Integrated Global Water Cycle Observations
IGWMC	International Ground Water Modeling Center
IH	Institute of Hydrology
IHA	Index of Hydrological Alteration
IHD	International Hydrological Decade
IHDM	Institute of Hydrology Distributed Model
IHDP	International Human Dimensions Programme on Global Environmental Change
IHE	Institute for Infrastructural, Hydraulic and Environmental Engineering
IHP	International Hydrological Programme
IHSs	Isotopic Hydrograph Separations
iLEAPS	Integrated Land Ecosystem–Atmosphere Process Study
ILU	Incomplete Lower-Upper
ILWRM	Integrated Land and Water Resources Management
IM	Inverse Modelling
IMS	Ice Mapping System
IN	Ice Nuclei
INBO	International Network of Basin Organizations
INCA	Integrated Catchments
InHM	Integrated Hydrology Model
INQUA	International Union for Quaternary Research
InSAR	Interferometric Synthetic Aperture Radar
INWRDAM	Inter-Islamic Network on Water Resources Development and Management
IO	Input–Output
IOC	Intergovernmental Oceanographic Commission
IOP	Intensive Observation Periods
IP	Induced Electrical Polarization
IPCC	Intergovernmental Panel on Climate Change
IPM	Instrumental Product Matrix
IPO	International Project Office
IPT	Integrated Profiling Technique
IPTA	Interferometric Point Target Analysis
IR	Infrared
IRBM	Integrated River Basin Management
IRD	Ice-Rafted Debris
IRGA	Infrared Gas Analyzer
IRN	International Rivers Network
IRTCES	International Research and Training Center on Erosion and Sedimentation
ISARM	International Shared Aquifer Resources Management
ISAs	Impervious Surface Areas
ISCCP	International Satellite Cloud Climatology Project
ISDR	International Strategy for Disaster Reduction
ISLSCP	International Satellite Land Surface Climatology Project
ISO	International Organization for Standardization
ITASE	International Trans Antarctic Scientific Expedition
ITCZ	Intertropical Convergence Zone
IUCN	International Union for Conservation
IUGG	International Union of Geodesy and Geophysics
IUH	Instantaneous Unit Hydrograph
IUSR	Illinois Urban Storm Runoff
IWA	International Water Association
IWC	Ice Water Content
IWE	Institute for Water Education
IWHA	International Water History Association
IW: LEARN	International Waters Learning Exchange and Resource Network

IWMI	International Water Management Institute
IWRA	International Water Resources Association
IWRM	Integrated Water Resource Management
IWV	Integrated Water Vapor Density
JAXA	Japan Aerospace Exploration Agency
JERS-1	Japanese Earth Resources Satellite
JIIHP	Joint International Isotopes in Hydrology Program
JISAO	Joint Institute for the Study of the Atmosphere and Ocean
JMP	Joint Monitoring Programme for Water Supply and Sanitation
JPL	Jet Propulsion Lab
JSC	Joint Scientific Committee
JWGASF	Joint Working Group on Air/Sea Fluxes
KDD	Knowledge Discovery in Databases
KINEROS2	Water Erosion Prediction Project
KQML	Knowledge Query and Manipulation Language
L/UMCF	Elfin Cloud Forest
LAC	Local Area Coverage
LAI	Leaf Area Index
LAS	Large Aperture Scintillometers
LBA	Large-Scale Biosphere Atmosphere
LDAS	Land Data Assimilation System
LDC	Link Discontinuity Concept
LDM	Lateral Distribution Method
LE	Latent Heat Exchange
LEO	Low Earth Orbit
LES	Large-Eddy Simulation
LEW and REW	Left and Right Edges of Water
LFV	Lower Fraser Valley
LGA	Lattice Gas Automata
LGM	Last Glacial Maximum
LIA	Little Ice Age
LID	Low Impact Development
LIDAR	Light Detection and Ranging
LIRAD	Lidar-Radiometer
LIS	Land Information System
LISEM	Limburg Soil Erosion Model
LLJ	Low-Level Jet
LLNL	Lawrence Livermore National Laboratory
LME	Local Mass Equilibrium
LMW	Low Molecular Weight
LNAPLs	Light Nonaqueous Phase Liquids
LOICZ	Land Ocean Interactions in the Coastal Zone
LPB	La Plata Basin
LPS	Lipopolysaccharides
LRTAP	Long Range Transboundary Air Pollution
LS	Land Surface
LSE	Least Square Estimation
LSM	Land Surface Model
LSP	Land Surface Process
LSS	Land-Surface Schemes
LST	Land Surface Temperature
LSWI	Land Surface Water Index
LTER	Long-Term Ecological Research
LUCC	Land Use and Land Cover Change
LVDT	Linear Variable Differential Transducer
LVIS	Laser Vegetation Imaging Sensor
LWC	Liquid Water Content
LWF	Liquid Water Flux
LWP	Liquid Water Path
LWR	Locally Weighted Regression
MAGIC	Model of Acidification of Groundwater in Catchment
MAGS	McKenzie GEWEX Study
MAIRS	Monsoon Asia Integrated Regional Study
MAP	Mean Annual Precipitation
MAR	Mean Annual Runoff
MASs	Multi Agent Systems
MCA	Medieval Climate Anomaly
MCAT	Monte Carlo Analysis Toolbox
MCC	Mesoscale Convective Complex
MCDEP	Montgomery County Department of Environmental Protection
MCL	Maximum Contaminant Level
MCMC	Markov Chain Monte Carlo
MCS	Monte Carlo Simulation
MCSM	Monte Carlo Set Membership
MCSs	Mesoscale Convective Systems
MDB	Murray Darling Basin
MD-DNR	Maryland Department of Natural Resources
MDGs	Millennium Development Goals
MEMS	Micro-Electro-Mechanical Systems
MERIT	Magneto-Electrical Resistivity Imaging Technique
METEOSAT	Meteorological Satellites
MEWIN	Middle East Water Information Network
MF	Multiple Flow
MFOSM	Mean-Value First-Order Second-Moment
MGPs	Manufactured Gas Plants
MH	Metropolis Hastings
MICCP	Mixed-Integer Chance Constrained Programming
MIM	Mobile–Immobile Transport Model

MIMS	Multimedia Integrated Modeling Systems
MINC	Multiple Interacting Continua
MIP	Major Intrinsic Protein
MIRAS	Microwave Imaging Radiometer by Aperture Synthesis
MISO	Multiple Input Single Output
MISP	Mutually Interactive State and Parameter
MISR	Multiangle Imaging SpectroRadiometer
MJO	Madden–Julian Oscillation
ML	Mixed Layer
MLBMA	Maximum Likelihood Bayesian Model Averaging
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
MLS	Microwave Limb Sounder
MM	Moist Moderate
MMOC	Modified MOC
MMS	Modular Modeling System
MNA	Monitored Natural Attenuation
MNCPPC	Maryland-National Capital Park and Planning Commission
MOA	Mosaic of Antarctica
MOC	Method of Characteristics
MODIS	Moderate Resolution Imaging Spectroradiometer
MOGA	Multiple Objective Genetic Algorithm
MOM	Monin–Obukhov Similarity Method
MOPEX	Model Parameter Estimation Experiment
MORECS	Met Office Rainfall and Evaporation Calculation System
MOS	Monin–Obukhov Similarity
MOSCEM-UA	Multiobjective Shuffled Complex Evolution Metropolis
MOU	Memorandum of Understanding
MP	Moist Polar
MPA	Multisatellite Precipitation Analysis
MPLNET	Micropulse Lidar Network
MPS	Marketable Permit Systems
MR	Multiple Radii
MRM	Moisture Retention Model
MRT	Mean Residence Time
MSA	Metropolitan Statistical Areas
MSC	Meteorological Service of Canada
MSE	Mean Squared Error
MSFC	Marshall Space Flight Center
MSG	Meteosat Second Generation
MSI	Moisture Stress Index
MSL	Meso-Scale Model
MSO	Multistep Outflow
MSS	MultiSpectral Scanner
MSU	Microwave Sounder Unit
MT	Moist Tropical
MTB	Modified Turning Band
MTBE	Methyl-*tert*-butyl-ether
MUSCL	Monotone Upstream-Centered Scheme for Conservation Laws
MUSLE	Modified Universal Soil Loss Equation
mwp-1A	Meltwater Pulse 1A
N_K	Kjeldahl N
NABIR	Natural and Accelerated Bioremediation Research Program
NADW	North Atlantic Deep Water
NAO	North Atlantic Oscillation
NAP	Nonarboreal Pollen
NAPLs	Nonaqueous Phase Liquids
NASA	National Aeronautics and Space Administration
NASDA	National Space Development Agency
NAVSTAR	Navigation Signal Timing and Ranging
NAWQA	National Water Quality Assessment
NCALM	National Center for Airborne Laser Mapping
NCAR	National Center for Atmospheric Research
NCDC	National Climatic Data Center
NCED	National Center for Earth-Surface Dynamics
NCEP	National Centers for Environmental Prediction
NDDSSs	Network Distributed Decision Support Systems
NDII	Normalized Difference Infrared Index
NDOP	National Digital Orthophoto Program
NDSI	Normalized Difference Snow Index
NDVI	Normal Difference Vegetation Index
NED	National Elevation Dataset
NEE	Net Ecosystem Exchange
NEMI	National Environmental Methods Index
NEP	Net Ecosystem Production
NESDIS	National Environmental Satellite, Data, and Information Service
NFFS	National Flood Forecasting System
NGOs	Non-Governmental Organizations
NGRIP	North Greenland Ice Core Project
NH	Northern Hemisphere
NHD	National Hydrography Dataset
NIC	National Ice Center
NIES	National Institute for Environmental Studies
NIH	National Institutes for Health
NIR	Near-Infrared
NIST	National Institute of Standards and Technology
NLCD	National Land Cover Database
NLDAS	North American Land Data Assimilation System

NMR	Nuclear Magnetic Resonance
NOAA	National Oceanic and Atmospheric Administration
NOAA-AVHRR	National Oceanic and Atmospheric Administration–Advanced Very High Resolution Radiometer
NOEL	No Observable Effects Limit
NOHRSC	National Operational Hydrologic Remote Sensing Center
NPDES	National Pollution Discharge Elimination System
NPGA	Niched Pareto GA
NPOESS	National Polar Orbiting Environmental Satellite System
NPP	Net Primary Production
NPV	Nonphotosynthetic Vegetation
NRC	Nuclear Regulatory Commission
NRCS	Natural Resources Conservation Service
NSCAT	NASA Scatterometer
NSF	National Science Foundation
NSGA-II	Nondominated Sorted GA-II
NSIDC	National Snow and Ice Data Center
NSIPP	NASA Seasonal-to-Interannual Prediction Project
NuCM	Nutrient Cycling Model
NWO	Netherlands Organization for Scientific Research
NWP	Numerical Weather Prediction
NWS	National Weather Service
NWSRFS	National Weather Service River Forecast System
NYCT	New York City Tunnels
OAP	Optical Array Probe
OBER	Office of Biological and Environmental Research
ODA	Overseas Development Agency
ODEs	Ordinary Differential Equations
OI	Optimum Interpolation
OIC	Organization of the Islamic Conference
OLR	Outgoing Long-Wave Radiation
OLS	Ordinary Least Squares
OM	Organic Matter
OMS	Object Modeling System
OpenMI	Open Modeling Interface and Environment
OPI	OLR Precipitation Index
OSL-SAR	Optically Stimulated Luminescence and Single Aliquot Regeneration
OSTDS	On-Site Treatment and Disposal Systems
OSU	Oregon State University
OTIS	One-Dimensional Transport with Inflow and Storage
PACS	Pan American Climate Studies
PAGES	Past Global Changes
PAHs	Polycyclic Aromatic Hydrocarbons
PAM	Primary Amoebic Meningoencephalitis
PAR	Photosynthetically Active Radiation
PARCA	Program for Arctic Regional Climate Assessment
PBL	Planetary Boundary Layer
PC	Plant Coefficient
PCA	Principal Components Analysis
PCBs	Polychlorinated Biphenyls
PCG	Preconditioned Conjugate Gradient Method
PCM	Parallel Climate Model
PCR	Polymerase Chain Reaction
PDAs	Personal Digital Assistants
PDE	Partial Differential Equation
pdf	Probability Density Functions
PDISC	Probability-Distributed Interacting Storage Capacity
PDM	Probability Distributed Model
PDO	Pacific Decadal Oscillation
PDSI	Palmer Drought Severity Index
P-E	Precipitation Relative to Evaporation
PE	Potential Evapotranspiration
PEEP	Photoelectronic Erosion Pins
PEM	Prediction Error Minimization
PEP	Phosphoenol Pyruvate Molecule
PERSIANN	Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks
PES	Payment for Environmental Services
PET	Potential Evapotranspiration
PFC	Projected Foliage Cover
PFTs	Plant Functional Types
PHABSIM	Physical Habitat Simulation
PIC	Particulate Inorganic Carbon
PILPS	Project for Intercomparison of Land-Surface Parameterization Schemes
PIP	Precipitation Intercomparison Projects
PIRATA	Pilot Research Moored Array in the Tropical Atlantic
PM	Penman–Monteith
PMIP	The Paleoclimate Modeling Intercomparison Project
PMP	Probable Maximum Precipitation
PMS	Particle Measurement System
PMW	Passive Microwave
PNA	Pacific–North America
POC	Particulate Organic Carbon
PO.DAAC	Physical Oceanography Distributed Active Archive Center
POES	Polar Orbiting Environmental Satellite

POLDER	Polarization and Directionality of the Earth Reflectances
POM	Particulate Organic Matter
PON	Particulate Organic N
POP	Parallel Ocean Program
POPs	Persistent Organic Pollutants
POT	Peaks-Over-Threshold
PP	Particulate P
PPCPs	Pharmaceuticals and Personal Care Products
ppmv	parts per million by volume
PPT	Measured Precipitation
PR	Precipitation Radar
PRARE	Precise Range and Range-Rate Equipment
PRB	Permeable Reactive Barriers
PRF	Pulse Repetition Frequency
PRISM	Parameter-Elevation Regressions on Independent Slopes Model
PRMS	Precipitation-Runoff Modeling System
PRTF	Physically Realizable Transfer Function
PS-SQP	Pattern Search with Sequential Quadratic Programming
PS	Permanent Scatterers
PTFs	Pedotransfer Functions
PUB	Prediction in Ungauged Basins
PW	Precipitable Water
QM	Quantitative Models
QoLC	Quality-of-Life Capital
QPF	Quantitative Precipitation Forecasts
RACMO	Regional Atmospheric Climate Model
Radar	Radiowave Detection and Ranging
RAMP	RADARSAT Antarctic Mapping Project
RAMS	Regional Atmospheric Modeling System
RANS	Reynolds Averaged Navier–Stokes
RAP	Cooperative Research Centre for Catchment Hydrology River Analysis Package
RBF	Radial Basis Function
RBM	Rules Based Models
RCC	River Continuum Concept
RCM	Regional Climate Models
RCRA	Resource Conservation and Recovery Act
RCS	Regional Curve Standardization
RCTWS	Regional Center for Training and Water Studies of Arid and Semi-Arid Zones
RCUWM	Regional Center on Urban Water Management
REA	Representative Elementary Area
REBEX	Radiobrightness Energy Balance Experiments
REV	Representative Elementary Volume
REW	Representative Elementary Watershed
RF	Rainfall-Flow
RFFS	Real Time Flood Forecasting System
RFI	Radio Frequency Interference
RH	Relative Humidity
RHBN	Reference Hydrometric Basin Network
RHESSys	Regional HydroEcological Simulation System
RHS	River Habitat Surveys
RIV	Refined Instrumental Variable
RL	Raman Lidar
RLIWG	River and Lake Ice Working Group
RLSE	Regularized Least Squares Estimator
RMS	Root-Mean-Square
RMSD	Root Mean Squared Difference
RMSE	Root Mean Square Error
RMT	Radio-Magnetotellurics
RMV	Representative Macroscopic Volume
RNN	Recurrent Neural Networks
ROL	Radial Oxygen Loss
ROTO	Routing Outputs to Outlets
RPCA	Rotated Principal Component Analysis
RR	Rainfall-Runoff
RRCs	Regional Research Centers
RRNs	Regional Research Networks
RRSs	Regional Research Sites
RS	Remote Sensing
RSA	Regional Sensitivity Analysis
RSF	Random Space Function
RT	Radiative Transfer
RTD	Residence Time Distribution
RTE	Radiative Transfer Equation
RTT	Radiative Transfer Theory
RUSLE1	Revised Universal Soil Loss Equation
RVA	Range of Variability Approach
RZ	Riparian Zones
SA	Simulated Annealing
SAHRA	Semi-Arid Hydrology and Riparian Areas
SALPEX	Southern Alps Experiment
SAP	Strategic Action Programs
SAR	Synthetic Aperture Radar
SARB	Surface and Atmospheric Radiation Budget
SAT	Soil–Aquifer Treatment
SBL	Stable Boundary Layer
SBUH	Santa Barbara Urban Hydrograph
SCA	Snow-Covered Area
SCAN	Soil Climate Analysis Network
SCE	Snow Cover Extent
SCEM	Shuffled Complex Evolution Metropolis
SCLS	Synthesized Constrained Linear Systems

SCOR	Scientific Committee on Ocean Research
SCOWAR	Scientific Committee on Water Resources
SCWG	Snow and Climate Working Group
SDC	Snowcover Depletion Curves
SDD	Sustainable Development Design
SDFN	Stochastic Discrete Fracture Networks
SDNP	Sustainable Development Networking Programme
SDP	State-Dependent Parameter
SDR	Sediment Delivery Ratio
SDRs	Sensor Data Records
SEB	Surface Energy Balance
SEBAL	Surface Energy Balance Algorithm for Land
SEBS	Surface Energy Balance System
SEDEM	Sediment Delivery Model
SEI	Stockholm Environment Institute
Sf	Stemflow
SFC	"Standard" Fog Collectors
SFF	Spectral-Feature-Fitting
SFGL	Surficial Fine-Grained Laminae
SFLA	Shuffled Frog Leap Algorithm
SG	Satellite-Gauge
SGP	Southern Great Plains
SGSIM	Sequential Gaussian Simulation
SH	Southern Hemisphere
SHAW	Simultaneous Heat and Water
SHB	Stem Heat Balance
SHDOM	Spherical Harmonical Discrete Ordinate Method
SHE	Systéme Hydrologique Européen
SIA	Social Impact Assessment
SiB	Simple Biosphere
SIDA	Swedish International Development Agency
SIDS	Small Island Developing States
SIL	Societas Internationalis Limnologiae
SIM	Split Inversion Method
SIR-C	Shuttle Imaging Radar C-Mission
SISIM	Sequential Indicator Simulation
SIU	Sensor Input Unit
SIWSI	Shortwave Infrared Water Stress Index
SL	Surface Layer
SLA	Shuttle Laser Altimeter
SLEMSA	Soil Loss Estimator for Southern Africa
SLICER	Scanning Lidar Imager of Canopies by Echo Recovery
SLR	Satellite Laser Ranging
SLS	Simple Least Square
SLURP	Semidistributed Land-Use-Based Runoff Processes
SMA	Spectral Mixture Analysis
SMACEX	Soil Moisture–Atmospheric Coupling Experiment
SMC	Soil Moisture Content
SMCF	Subtropical Montane Rain Forest
SMD	Soil Moisture Deficit
SMGA	Structured Messy GA
SMMR	Scanning Multichannel Microwave Radiometer
SMOS	Soil Moisture and Ocean Salinity
SMRF	Lower/Upper Montane Cloud Forest
SMS	Surface-Water Modeling System
SNIA	Sequential Noniterative Approach
SNMR	Surface Nuclear Magnetic Resonance
SNNS	Stuttgart Neural Network Simulator
SNODAS	Snow Data Assimilation System
SNOTEL	SNOpack TELemetry
SNOWMIP	Snow Models Inter-Comparison Project
SO	Southern Oscillation
SOD	Sediment Oxygen Demand
SODA	Simultaneous Optimization and Data Assimilation
SOF	Saturation Overland Flow
SOFM	Self-Organizing Feature Maps
SOI	Southern Oscillation Index
SOM	Soil Organic Matter
SOS	Southern Oxidants Study
SPAW	Soil Plant Atmosphere Water
SPCZ	Southern Pacific Convergence Zone
SPDEs	Stochastic Partial Differential Equations
SPI	Standardized Precipitation Index
SPM	Suspended Particulate Matter
SPOT	Systeme Probatoire pour l'Observation de la Terre
SPR	Standard Percentage Runoff
SRB	Surface Radiation Budget
SRDP	Successive Reaching Dynamic Programming
SRES	Special Report on Emissions Scenarios
SRF	Spatial Random Function
SRIV	Simplified Refined Instrumental Variable
SRM	Snowmelt Runoff Model
SRTM	Shuttle Radar Topography Mission
SS	Suspended Solids
SSARR	Streamflow Simulation and Reservoir Regulation Model
S-SEBI	Simplified Surface Energy Balance Index
SSF	Subsurface Stormflow
SSG	Steady State Gain
SSM/I	Special Sensor Microwave Imager
SSS	Separate Sewer Systems
SST	Sea Surface Temperature
SSTA	Space Science and Technology Alliance
SSU	Stratospheric Sounder Unit

STANMOD	STudio of ANalytical MODels
STAR	Synthetic Thinned Array Radiometry
START	Global Change System for Analysis, Research and Training
STATSGO	State Soil Geographic Data Base
STC	Sediment Trapping Capability
STM	Stream-Tube Models
STRF	Seasonal Tropical Rain Forest
STVF	Surface-Tension Viscous-Flow
STW	Sewage Treatment Works
SUFI	Sequentual Uncertainty Fitting Algorithm
SURFRAD	Surface Radiation Monitoring Network
SVAT	Soil Vegetation Atmosphere Transfer
SVE	Soil Vapor Extraction
SVIWG	Snow–Vegetation Interactions Working Group
SVM	Support Vector Machines
SVS	Soil Vapor Surveys
SW	Split-Window
SW2D	Shallow Water 2D
SWAP	Soil–Water–Atmosphere–Plant
SWAT	Soil and Water Assessment Tool
SWATS	Soil Water and Temperature System
SWB	Soil Water Balance
SWCT	Soil Water Characteristics from Texture
SWE	Snow Water Equivalence
SWH/ATT	Significant Wave Height-Attitude
SWI	Soil Wetness Index
SWIM	Soil and Water Integrated Model
SWMM	Storm Water Management Model
SWOs	Surface Water Outfalls
SWRRB	Simulator for Water Resources in Rural Basins
SZA	Solar Zenith Angle
TATE	Time–Area Topographic Extension
TB	Brightness Temperatures
TC	Total Coliform
TCA	1,1,1-Trichloroethane
TCI	TRMM Combined Instrument
TDA	Transboundary Diagnostic Analysis
TD-EM	Time-Domain Electromagnetic
TDR (FDR)	Time (or Frequency) Domain Reflectometry
TDS	Total Dissolved Solids
TEAPs	Terminal Electron-Accepting Processes
TES	Temperature Emissivity Separation
TF	Transfer Function
TFM	Transfer Function Model
THIR	Temperature Humidity Infrared Radiometer
THORPEX	The Observing System Research and Predictability Experiment
TI	Tension Disc Infiltrometer
TIN	Triangular Irregular Network
TIR	Thermal Infrared
TIROS-1	Television Infrared Operational Satellite-1
TKE	Turbulent Kinetic Energy
TM	Thematic Mapper
TMCF	Tropical Montane Cloud Forests
TMDL	Total Maximum Daily Load
TMI	TRMM Microwave Imager
TN	Total N
TOA	Top of Atmosphere
TOC	Total Organic Carbon
TOGA COARE	Tropical Ocean Global Atmosphere Coupled Ocean Atmosphere Response Experiment
TOMS	Total Ozone Mapping Spectrometer
TOPC	Terrestrial Observation Panel for Climate
TOVS	TIROS Operational Vertical Sounder
TP	Total Phosphorus
TPW	Total Precipitable Water
TRMM	Tropical Rainfall Measuring Mission
TRRL	Transport and Road Research Laboratory Method
TSA	Time Series Analysis
TSI	Total Solar Irradiance
TSM	Transient Storage Model
TSS	Total Suspended Sediment/Solids
TVC	Trail Valley Creek
TVD	Total Variation Diminishing
TVP	Time-Variable Parameter
TWI	Topographic Wetness Index
UCOWR	Universities Council on Water Resources
UCUR	Cincinnati Urban Runoff Model
UGN	Ultra Giant Nuclei
UH	Unit Hydrograph
UIUC	University of Illinois Urbana-Champaign
UKF	Unscented Kalman Filter
UKMO	United Kingdom Meteorological Office
UN	United Nations
UNCCD	United Nations Convention to Combat Desertification
UNCED	United Nations Conference on Environment and Development
UNDP	United Nations Development Program
UN/ECE	United Nations Economic Council for Europe
UNEP	United Nations Environment Program
UNESCO	United Nations Education, Scientific and Cultural Organization
UNFCCC	United Nations Framework Convention on Climate Change

UNICEF	United Nations International Children's Emergency Fund
UNU	United Nations University
UNU-INWEH	United Nations University-International Network on Water, Environment and Health
USDA	US Department of Agriculture
US EPA	United States Environmental Protection Agency
USGS	United States Geological Survey
USLE	Universal Soil Loss Equation
USRA	Universities Space Research Association
UTH	Upper-Tropospheric Humidity
UV	Ultraviolet
UWBPP	The Upper Wharfedale Best Practice Project
VAR	Variable Rainrate Precipitation
VCA	Voluntary Contractual Arrangements
VCP	Vertical Coplanar
VDICS	Vertical Distribution of Intercepted Surfaces
VEGA	Vector-Evaluated GA
VHMW	Very High Molecular Weight
VI	Vegetation Index
VIC	Variable Infiltration Capacity
VIIRS	Visible Infrared Imager/Radiometer Suite
VLBI	Very Long Base Line Interferometry
VLF	Very Low Frequency
VLMW	Very Low Molecular Weight
VMD	Volume-Weighted Mean Droplet Diameter
VNIR	Visible and Near-Infrared Sensors
VNIRA	Visible and Near-Infrared Analysis
VOCs	Volatile Organic Compounds
VPD	Vapor Pressure Deficit
VRTE	Vector Radiative Transfer Equation
VSA	Variable Source Area
VSM	Volumetric Soil Moisture
VSMOW	Vienna Standard Mean Ocean Water
VSP	Vertical Seismic Profiling
VTMX	Vertical Transport and Mixing Experiment
VTPR	Vertical Temperature Profiler Radiometer
VWI	Vegetation Water Indices
WAF	Weighted Average Flux
WARSMP	Watershed and River System Management Program
WATUP	Water Uptake Model
WBCSD	World Business Council for Sustainable Development
WBI	Water Band Index
WCMC	World Conservation Modeling Center
WCP	World Climate Programme
WCRP	World Climate Research Program
WDCM	World Data Center for Meteorology
WDN	Water Distribution Network
WDPT	Water Drop Penetration Time
WDR	Width–Depth Ratio
WEAP	Water Evaluation and Assessment Planning
WEDC	Water, Engineering and Development Centre
WEF	Water Environment Federation
WFD	Water Framework Directive
WGEW	Walnut Gulch Experimental Watershed
WGI	World Glacier Inventory
WGMS	World Glacier Monitoring Service
WHC	Water Holding Capacity
WHO	World Health Organization
WHYCOS	World Hydrological Cycle Observing System
WLC	Wormlike Chain
WLS	Weighted Least Squares
WLSE	Weighted Least Squares Estimator
WMO	World Meteorological Organization
WMS	Watershed Modeling System
WRAP	Water Resources Application Project
WRC	Water Retention Curve
WRCSEAP	Regional Humid Tropics Hydrology and Water Resources Center for South-East Asia and the Pacific
WRI	World Resources Institute
WRR	World Resources Report
WSHP	Water, Sanitation and Health Program
WT	Water Table
WTF	Water-Table Fluctuation
WUA	Weighted Useable Area
WUE	Water Use Efficiency
WWAP	World Water Assessment Project
WWC	World Water Council
WWDR	World Water Development Report
WWF	World Water Forum
WWRP	World Weather Research Programme
WwTWs	Wastewater Treatment Works
WWW	World Weather Watch
XML	eXtended Markup Language
YIC	Young Information Criterion
ZFP	Zero-Flux Plane
ZOH	Zero-order Hold
ZWD	Zenith Wet Delay

PART 1

Theory, Organization and Scale

1: On the Fundamentals of Hydrological Sciences

GÜNTER BLÖSCHL

Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Vienna, Austria

Although there is no universal theory in hydrology that starts from first principles, the various branches of hydrology show numerous common threads. They relate to the nature of the processes including their space–time variability, the general principles of hydrological measurements, and the types of methods for representing hydrological processes in a quantitative way, either statistically or deterministically. The purpose of this article is to provide some common ground for this Encyclopedia to highlight the particularities of the hydrological sciences.

WHAT ARE THE FUNDAMENTALS IN HYDROLOGY?

Hydrology is the science that deals with the waters above and below the land surfaces of the Earth; their occurrence, circulation and distribution, chemical and physical properties, and their interaction with their environment, including their relationship to living beings (NRC, 1991). Owing to its central focus on water, the science of hydrology holds a unique place in the field of earth system science, intimately intertwined with other water-related disciplines such as meteorology, climatology, geomorphology, hydrogeology, and ecology. As an applied science, hydrology is highly relevant to the management of the world's water resources and water quality and for the prediction and mitigation of water-related natural hazards such as floods and droughts. Thus, hydrology is an exciting field of study.

What now are the theoretical underpinnings of hydrology, what are the fundamentals? In many disciplines, a treatise on the fundamentals starts with a universal "big picture" theory on which there is consensus among scientists. From theory one would then move into the specific questions of how to measure, how to conceptualize more specific processes and how to model them. The theory would give guidance on all of these and would be further developed on the basis of feedbacks from them. Hydrology is different in this respect from some other natural sciences. There is no universal theory of hydrology that starts from first principles. There are different concepts for different parts of the hydrologic cycle and different spatial and temporal scales. The various branches of hydrology, however, do show remarkable parallels. The nature of hydrological variability is remarkably similar for different processes and the measurement techniques available to probe them have similar characteristics as well. Both have distinctly shaped the descriptive and predictive methods that have evolved in this discipline over the years and they ultimately control the accuracy of hydrological predictions. The common threads of the various hydrological subdisciplines may hence be a useful starting point for a presentation of fundamentals in the hydrological sciences. These are the subject of **Part 1: Theory, Organization and Scale** of this Encyclopedia. The objective of Part 1 and this chapter in particular, is to provide some common ground for the remainder of the Encyclopedia and to bring out some of the hydrological concepts that are common to them.

HYDROLOGICAL PROCESSES – WATER CYCLES AND WHY ORGANIZATION IS AN ISSUE

The most influential concept in hydrology has undoubtedly been the water cycle that links the movement of water on the Earth's surface with subsurface waters and water in the atmosphere. It not only provides a general layout

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

of the main mechanisms but also allows formulation of how much water there is in the different compartments (in the atmosphere, on the land surface, and in the subsurface) and how fast the exchange takes place (*see* **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1**; **Chapter 25, Global Energy and Water Balances, Volume 1**; **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**; **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; **Chapter 103, Terrestrial Ecosystems, Volume 3**; and **Chapter 173, Global Water Cycle (Fundamental, Theory, Mechanisms), Volume 5**). Obviously, the movement of water is more complex than an exchange of water between different boxes. In fact, one could argue that there are many water cycles, as water moves around at many space and time scales. There is a multitude of different pathways (*see* **Chapter 4, Organization and Process, Volume 1**; **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**; and **Chapter 113, Hyporheic Exchange Flows, Volume 3**). Water may fall as rain in the same regions as it is evaporated, a process termed *local moisture recycling*, and water may remain much longer in the ground in some places than in others, so there are huge differences in the time scales as well. The global water cycle is linked to the global energy cycle through evapotranspiration on the land surface. Understanding the water cycle is also a key element in understanding fluxes of matter (e.g. nutrients, sediments) that are driven by the water fluxes (*see* **Chapter 79, Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport, Volume 2**; **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**; and **Chapter 96, Nutrient Cycling, Volume 3**).

One of the fascinating observations on hydrological processes is their astounding variability at all scales, in both space and time. At the smallest scales of interest in hydrology, water fluxes and composition may vary between individual pores of the soil, and climate and hydrological processes vary over continental scales as well. Infiltration may vary over seconds and groundwater tables may vary over decades and more. Within these limits, variability abounds (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**, **Chapter 7, Methods of Analyzing Variability, Volume 1**). Virtually any quantitative approach to this problem requires the selection of a limited set of spatial and temporal scales. Any particular choice of time and space scales has a major influence on which aspects of this hydrological variability are perceived (*see* **Chapter 8, Fractals and Similarity Approaches in Hydrology, Volume 1**; **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**; and **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3**).

Hydrological variations are driven by variations in physiographic factors such as climate, soils, vegetation, topography, geology, as well as by human activity. These externally driven variations then propagate through hydrological systems (Sivapalan *et al.*, 2001), leading to an extremely rich variety of hydrological patterns apparent at different temporal and spatial scales, in different physical settings. This means that the patterns of variability are linked to their causal processes. Although Schumm (1991) notes that due to nonlinearities multiple processes can lead to the same form, patterns and form should be able to provide an indication of the processes that have led to them. Examining patterns will hence assist in making more representative measurements and more accurate predictions (*see* **Chapter 8, Fractals and Similarity Approaches in Hydrology, Volume 1**, **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**).

The drivers also imply that the variability one encounters in hydrology is usually not fully random but organized in various ways (Gutknecht, 1993). Types of organized variability include continuity (Baird, 1996) in time and space which is often related to storage processes. Another type is the presence of zones with boundaries between them (Woo, 2004). Still another type of organization that seems to exist at all scales is preferential flow – in the voids of the soil, in macropores, and in both porous and hard rock aquifers at a range length-scales (*see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**; **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; and **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**). On the land surface, preferential flow occurs from micro rills to streams in the landscape (Rinaldo *et al.*, 1993). The counterpart to preferential flow in the time domain is episodic behavior, that is, a concentration of activity over short periods or events in a range of processes including runoff, erosion, and sediment transport. Other types of organization include self-similar organization, where small-scale variability looks similar to large-scale variability; the observation that extremes or outliers occur more often than would be expected on the basis of standard statistical distributions (Hurst, 1951; Mandelbrot and Wallis, 1968); and periodic variability at diurnal, annual, and multiannual scales (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**). Clearly, these organized patterns are linked to the processes that drive and modulate them.

The presence of spatial and temporal organization in hydrologic variability has important ramifications for measurements. If it were not for surface runoff concentrated in streams, it would be almost impossible to measure the water flowing from a catchment area. On the other hand, preferential flow in the soils and aquifers tends to

make point samples unrepresentative. Organization also has important ramifications for representing the variability in a quantitative way. The presence of organization or patterns in hydrological systems has been seen as an indication that they are, what Dooge (1986) refers to as *middle number systems* or *systems of intermediate complexity*. In these systems, there are too many components to be dealt with by classical (deterministic) mechanics and just not enough components to be dealt with by statistical methods similar to those of statistical mechanics. Thus in hydrology, both statistical and deterministic methods are appropriate depending on the type of variability one means to capture as well as the questions one asks. This type of system behavior also means that interactions of processes will be important at many scales (*see* **Chapter 4, Organization and Process, Volume 1**) such as interactions between surface water and groundwater (*see* **Chapter 113, Hyporheic Exchange Flows, Volume 3**); between soils, vegetation, and the atmosphere (*see* **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**); between land surface hydrology and terrestrial ecosystems at large (*see* **Chapter 101, Ecosystem Processes, Volume 3**, **Chapter 103, Terrestrial Ecosystems, Volume 3**); between evaporation and flood generation (*see* **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**; Sivapalan *et al.*, 2005); between floods and stream morphology (*see* **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**); between snow processes and boundary-layer atmospheric processes (*see* **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**); between catchment hydrology and soil development (*see* **Chapter 4, Organization and Process, Volume 1**); and between runoff and landscape evolution (*see* **Chapter 4, Organization and Process, Volume 1**). Not all the feedbacks will be apparent to an observer, as often observations are limited to a set of scales that reveal only a few of the many processes that are present in the hydrological environment (*see* **Chapter 6, Principles of Hydrological Measurements, Volume 1**, **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**).

FUNDAMENTAL EQUATIONS – ARE THERE ANY?

Yes, there are fundamental equations in hydrology and the most important one is the mass balance of water over a given volume and time interval. This is termed *the water balance equation*. It is so central to hydrology that some observers have noted that the task of hydrology is to solve the water balance equation. The classic example of its application is the estimation of the average evapotranspiration of a catchment over a long period from rainfall and streamflow measurements, but it is widely applied at a range of space and time scales. The water balance equation is the only equation that can be called a *hydrological equation* in its full right and is applicable to the scales hydrologists are interested in.

Another fundamental equation is energy balance, which is mainly used when interfacing with the atmospheric sciences and plays a key role in hydrology in the context of evaporation and snow processes. The remaining balance equation of classical mechanics, momentum balance, is mainly used in representing open channel flow in a fluid mechanics context (*see* **Chapter 5, Fundamental Hydrologic Equations, Volume 1**, **Chapter 135, Open Channel Flow – Introduction, Volume 4**). These balance equations – fundamental as they are – are not sufficient to fully describe the dynamics of hydrologic systems. Hence additional equations, termed *empirical flux laws*, are needed. Most of them have four characteristics:

1. Many of the flux laws used in hydrology are based on flux–gradient type relationships. Examples are Darcy's law (water flux in aquifers – hydraulic potential gradient), Fick's law (matter flux both in aquifers and in surface waters – concentration gradient); the flux-gradient method (vertical water vapor flux in the atmosphere – vapor pressure gradient); and the Chezy equation (water flux in surface waters – energy gradient).
2. Most of them have some element of empiricism, although derivations from more fundamental laws are possible. For example, Darcy's law can be derived from the Hagen–Poiseulle equation for laminar flow in capillaries. The assumptions in the derivations may imply that their applicability is limited to particular conditions, which may not always be clear. Through simplifications additional empiricism may creep in. In the Darcy example, the geometry of soil pores is far more complex than a bundle of tubes. Thus, an empirical element will usually be involved in the flux laws, for example, through empirical parameters in the flux – gradient relationships.
3. Most of the flux laws have been taken from other disciplines such as fluid mechanics, soil physics, and the atmospheric sciences, and hence
4. Many of them apply to the point scale, that is, a sample size that is small relative to the systems hydrologists are interested in. They have been derived for minute control volumes that are amenable to laboratory experiments and the application of continuum mechanics (Hubbert, 1956), rather than for the objects of interest in hydrology (catchments, aquifers, river reaches, regions, etc.). In principle, equations can be formulated for lumped systems at larger scales and catchment models are a good example (*see* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**). However, repeatable experiments under exactly controlled conditions are not possible at these scales, which challenges the universality of these equations. Another example is the stream order laws of Horton (Horton, 1945) and

other authors that describe the statistical characteristics of the map view patterns of stream networks. While important in fluvial geomorphology they have had limited influence on the hydrological sciences themselves.

Out of the four points listed, the fourth is probably the most important one for hydrology as a science (Blöschl and Sivapalan, 1995). Point scale equations can be straightforwardly extended to catchments, aquifers, reaches, and so on *provided* the boundary conditions are known and the media characteristics are known spatially (e.g. uniform) at the scale of the equations. This means there may be finer scale variability such as grains and voids not resolved by the equations, but at larger scales where the equations apply the media are considered uniform. For example, the mass balance equation for small volumes in aquifers can be combined with Darcy's law, which gives a diffusion type differential equation. It is then possible to use concepts from continuum mechanics to solve for the variable of interest (e.g. hydraulic head), given the initial and boundary conditions (*see* **Chapter 5, Fundamental Hydrologic Equations, Volume 1**).

The challenge in the hydrological sciences is that hydrological systems are never completely uniform in terms of their parameters, fluxes, and states and are often not even approximately uniform. Although there are ways of dealing with their variability – either explicitly through distributed (deterministic) models or implicitly through upscaling methods – it is not a straightforward exercise. Additional assumptions need to be made about the variability, both in space and time and, often most importantly, about the nature and locations of the flow paths, but much of this information may be "unknowable" in practice (Savenije, 2001). One is then far removed from the fundamental equations and on the "thin ice" of models for a particular application. This is also one of the reasons why models generally need to be calibrated to the particular site of interest (Freeze and Harlan, 1969).

There are two classical paradoxes in hydrology – dispersion in the subsurface tends to deviate from Fick's law (Levy and Berkowitz, 2003), and runoff events mainly contain old (pre-event) water (Kirchner, 2003). Both paradoxes are related to small-scale equations not being applicable at the larger hydrological scale because of media heterogeneities (*see* **Chapter 13, Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, Volume 1**, **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**).

The issues of heterogeneity relate to the empirical flux laws but not to the balance equations as the latter are valid at any scale. Issues of heterogeneity also arise in specifying initial and boundary conditions from measurements. Notwithstanding these problems, point scale equations along with continuum mechanics are an essential basis for hydrology (*see* **Chapter 5, Fundamental Hydrologic Equations, Volume 1**) both for understanding system dynamics and for making quantitative predictions.

HYDROLOGICAL MEASUREMENTS – WHY SIZE MATTERS

With the exception of the laboratory case, experiments are not repeatable under exactly the same boundary and initial conditions in hydrology; it is nature that does the experiments (Dunne, 1998; Zehe and Blöschl, 2004). Because of this, observations generally depend on the climatic and hydrological context. From the 1960s, there have been numerous national and international programs, initially on experimental catchments, to examine similarities and differences across different climatic and hydrological conditions (*see* **Chapter 121, Intersite Comparisons of Rainfall-runoff Processes, Volume 3**, **Chapter 203, A Guide to International Hydrologic Science Programs, Volume 5**). These programmes have provided valuable insights but generalizing the findings beyond the areas of interest has always been difficult (*see* **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**). Each aquifer, catchment, and river reach – in fact each episode – seems to have particularities that cannot be specified in full detail. Because of this, in addition to going into process detail for a single site (which has been the traditional approach), contrasting different catchments and different aquifers based on what has been termed *comparative hydrology* has recently been singled out as an important avenue to progress in hydrology (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**, **Chapter 121, Intersite Comparisons of Rainfall-runoff Processes, Volume 3**), with the eventual goal of a common method for assessing and quantifying hydrological similarity.

A key to the progress in the natural sciences is the ability to measure variables to an accuracy that is useful and at the scales one is interested in. This is another challenge in hydrology as, in many instances, the processes of interest are of a scale that is not directly amenable to the measurement techniques available (Klemeš, 1983). Most measurements are collected by point samples, while processes occur over catchments, aquifers, and landscapes. In the time domain, one is often more interested in (temporal) averages (e.g. sediment and nutrient loads) than in the snapshots as can be obtained in dedicated experiments (*see* **Chapter 92, Water Quality Monitoring, Volume 3**). Because of this, much of hydrology is constrained by measurement techniques (*see* **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**). This is particularly the case for spatial distributions which are more difficult to sample than time series (Grayson and Blöschl, 2000), especially for hydrological dynamics that take place beneath the ground surface.

Over the years, hydrologists have developed ways of dealing with the space–time variability and the scale incompatibility of measurements in various ways. The most efficient methods have been ways of aggregating the variability by prudent measurements. The classical examples are measurements of runoff from catchments that aggregate the within-catchment variability and pumping tests of aquifer transmissivities that aggregate the subsurface hydraulic variability within the depression cone (Anderson, 1997). More elaborate measurements require either laboratory analyses with typical sample sizes of 1 dm^3, although tracer experiments and irrigation/flume experiments can deal with somewhat larger scales of tens of meters. Because of this, the sampling design in terms of the space and time scales is critically important for capturing the natural variability in a representative way in addition to ensuring the accuracy of the instruments (*see* **Chapter 6, Principles of Hydrological Measurements, Volume 1**). In long-term monitoring, where networks are operated by national hydrographic services, space and time scales are usually large, while in dedicated field experiments organized by groups of scientists, space and time scales tend to be small, although, recently, a number of large-scale field experiments have been undertaken (*see* **Chapter 203, A Guide to International Hydrologic Science Programs, Volume 5**). In the latter, remote sensing methods play an important role as they are able to sample at finer spatial scales and wider areas than has been traditionally possible in hydrology (*see* **Chapter 47, Sensor Principles and Remote Sensing Techniques, Volume 2**). With recent advances in monitoring techniques from small-scale computer tomography to large-scale remote sensing methods as well as better logistics (*see* **Chapter 6, Principles of Hydrological Measurements, Volume 1**), measurements are increasingly able to capture wider scale ranges, but a scale problem remains for which statistical (nonprocess-based) and deterministic (process-based) upscaling methods have been developed (*see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**, **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**).

THE STATISTICAL APPROACH – STATISTICS, SELF SIMILARITY, AND UP/DOWNSCALING

There are two types of approaches to representing hydrologic systems, statistical, and deterministic. In both of them, data play an important role and both of them have their merits. The statistical approach is warranted if random variability (i.e. variability we are unable to interpret/predict in detail) prevails (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**; **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**; **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**; and **Chapter 125, Rainfall-runoff Modeling for Flood Frequency Estimation, Volume 3**). It represents the bulk information (frequency, distribution, dependence) not the details (spatial and temporal occurrence, dynamics). In the statistical approach it is not usually possible to take causal processes into account, which renders the extrapolation potential more limited than that of the deterministic approach, but extrapolation may not be needed for the application at hand and it is the main method used in many applied hydrologic problems. On the other hand, the statistical approach may be able to deal with systems that are too complex to be dealt with in a deterministic way.

A range of statistical techniques for representing variability are in use in hydrology. Typical steps in a sequential analysis of statistical variability are (i) looking at the data, (ii) analyzing the statistical distribution of the data, (iii) analyzing the first and second order moments (including an analysis of statistical dependence), and (iv) analyzing the data by more elaborate methods such as series expansion (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**). In these steps increasingly more complex descriptions are introduced. The second moments (variance and correlation coefficients) are of particular importance in hydrology as they are a measure of spread and hence variability of a variable. The classical example is the representation of the spread of a plume of concentration by the second moments. In a temporal (and spatial) context, the second moments can be used to represent the continuity of correlated time series (and correlated random fields) through correlation functions or variograms. In the time domain, the correlations can be used for stream flow forecasting (time series analysis), in the space domain for spatial estimation using geostatistical methods (*see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**). Series expansions go a step further by representing the variable of interest by a sum of deterministic functions of random variables. To the latter type of methods belong spectral analysis, wavelet analysis, principal component analysis, and empirical orthogonal functions. Their main value in hydrology lies in the reduction of the dimensionality of the system to assist in identifying the main controls if the patterns are not apparent in large data sets (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**).

The presence of patterns or organized variability is not always considered a favorable property in statistical analyses. They can involve nonstationarity, outliers, non-Gaussian (non-normal) behavior, and thresholds, which are all characteristics commonly encountered in hydrological data but not compatible with the usual statistical methods (*see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**). This is particularly an issue when extreme values (floods, precipitation extremes, low flows, extreme concentrations) are analyzed by statistical methods (*see* **Chapter 37, Rainfall Trend Analysis:**

Return Period, Volume 1; **Chapter 125, Rainfall-runoff Modeling for Flood Frequency Estimation, Volume 3**). Analyses of extremes are of particular relevance in applied hydrology and water resources management. On the other hand, organization tends to produce striking similarities across scales. Little wiggles look like big ones, statistically, and short ones like long ones. The remarkable thing is that whatever hydrologic variable is examined, it more often than not turns out that there exists similarity to the same variable examined at a different scale, at least over a certain range of scales. This is termed *self similar or fractal behavior* (*see* **Chapter 8, Fractals and Similarity Approaches in Hydrology, Volume 1**) and is related to the more general observation, that there is variability at all scales with the strength of variability (e.g. quantified in terms of the second moment) increasing with scale. In the past decades, statistical fractals have been widely used in many branches of hydrology, as they are appealing because of three main reasons. First, they deal with the presence of variability over a wide range of scales, which is consistent with observations. Second, this type of behavior can be related, at least qualitatively, to the dynamic behavior of nonlinear systems, which is an interesting paradigm for hydrological processes. Third, and perhaps most important for practical applications, fractal concepts lead to parsimonious descriptions of rainfall, landscapes, drainage networks, geologic media, and so on. This means that the statistical models only involve a few parameters and these can be estimated more robustly than the more numerous parameters of traditional concepts. Some of the fractal methods are based on the first and second moments (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**) but others involve more complex descriptions.

Statistical methods, including fractal concepts, can be used efficiently to address the scale incompatibility of hydrological processes, measurements, and predictions. They lend themselves to transferring information between various scales, for example, between point scale measurements and catchment scale prediction; or large-scale model output and small-scale predictions. These methods are termed *upscaling and downscaling methods* (*see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**). The first generic task of upscaling/downscaling is to derive the statistics of a variable at one scale from the statistics of the same (or another) variable at another scale. Methods range in complexity from regressions between the variables at different scales to upscaling theory of stochastic hydrogeology (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). The second generic task is to generate spatial patterns (or time series) given the statistical characteristics of the variable one means to represent. This can be either through interpolation between a number of samples or, alternatively, various disaggregation methods where one is interested in obtaining a number of realizations of the variable of interest that all exhibit the same statistics as the data. Some of these methods focus on the second moments by making use of correlation functions or variograms (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**, **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**).

A range of statistical methods are available for upscaling point rainfall to catchments, for disaggregating rainfall in time, for downscaling the output of global circulation models to the scale of catchments, to relate the flood characteristics of catchments of different sizes, to transfer soil moisture across scales both in a catchment and climate modeling context, and for characterizing and generating subsurface media (*see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**). One of the more general observations used in many of the statistical upscaling methods is that aggregation makes processes appear smoother, so variability decreases with aggregation area.

The statistical upscaling/downscaling approach does not attempt to represent the processes in full detail but rather relies on (lumped) summary descriptions of variability. Similar to other statistical methods, this has the benefit of robustness but at the expense of limited extrapolation potential. Alternative upscaling methods exist that involve equations of the underlying process dynamics in various branches of hydrology including dynamic hydrologic models (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**), land–atmosphere interactions (*see* **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**), soil water flow (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**), and stochastic subsurface hydrology (*see* **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**, **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**).

THE DETERMINISTIC APPROACH – MODEL CONCEPTS AND WHY UPSCALING AND DOWNSCALING IS NEEDED

The alternative to the statistical approach is the deterministic approach, which, likewise, has numerous merits. Deterministic relationships can be formulated in a causal way by making use of the fundamental equations and hence can be used for examining "what happens if" questions in a more reliable way than is typically possible with statistical methods. The downside, however, is that the processes may easily become too complex, so care needs to be taken to limit the models to those processes that are tractable and/or for which sufficient data are available. In terms of their

application, there are two main uses of deterministic models – explanatory models for furthering our understanding of a particular system and predictive models for producing estimates of some future or changed state. In both instances, there exists a range of model types, from lumped to spatially distributed, from low dimensional to high dimensional involving many parameters. The simplest ones are based on input–output relationships of the area of interest, the intermediate ones on some degree of understanding (conceptual models), and the most complex ones are based on the fundamental equations discussed earlier (*see* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**; **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**; and **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3**).

Model conceptualization and building usually follows a set number of steps including collecting and examining data and other evidence, assessing which processes may be important for the problem at hand, designing a scheme of the most important process dynamics in the modeler's mind, designing a mathematical model to represent these concepts, calibrating the model by using the data of the region, and testing the model by a separate data set of the same region. If the testing satisfies the modeler's expectations, then the model is ready for use, otherwise one or more of the steps need to be repeated (*see* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**; **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**; and **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**). An important component of the model building process and model application is the assessment of the model and data uncertainty to create confidence in the reliability of the model and model predictions (*see* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**, **Chapter 79, Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport, Volume 2**).

In many subdisciplines of hydrology, spatially distributed deterministic models are currently used in a routine way both for addressing practical water resources issues and for more theoretical analyses. The tremendous computing power that is available today facilitates the application of high-resolution models and there exists sophisticated software, particularly for subsurface hydrology and open channel flow. While the usefulness of these models is undisputed, there does remain significant uncertainty with the predictions for several reasons including data limitations and the model formulation (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**, **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**, Grayson *et al.*, 1992). Specifically, the scale issues discussed earlier in the context of fundamental hydrological equations play an important role here, as the empirical flux laws used in these models are indeed point scale equations (Beven, 1989). Awareness of these issues has triggered research into upscaling methods that are able to deal with unknown (small scale) spatial variability in the context of deterministic models. Much of the recent interest started in the 1970s with the early work of A. Freeze and L. Gelhar on aggregating the groundwater flow equation, based on a stochastic approach (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**), and picked up additional momentum in the 1980s when it was realized that the spatial heterogeneity of the land surface is important for atmospheric models (Gelhar *et al.*, 1977; Eagleson, 1986; Shuttleworth, 1988; **Chapter 32, Models of Global and Regional Climate, Volume 1**, **Chapter 177, The Role of Large-Scale Field Experiments in Water and Energy Balance Studies, Volume 5**). Those branches of hydrology where the basic equations are known with some degree of confidence (e.g. groundwater flow and transport) have had significant progress, but in other areas such as catchment hydrology and hill slope hydrology progress has been slower (Blöschl, 2001). The upscaling methods are either based on volume averaging or ensemble averaging (i.e. averaging all possible realizations on the same location) of the underlying equations. The aggregation methods tend to work very well if (i) the scale of the natural variability to be averaged (such as grains) is small as compared to the scale of the variability to be explicitly represented (such as geologic formations), and (ii) if the small-scale variability is random and does not exhibit organized patterns. Hydrologic variability tends to exhibit organized patterns such as preferential flow and variability tends to occur at all scales, so the upscaling methods have not been used as widely in practice as would be merited by their theoretical underpinnings. The variability within each grid cell of distributed models is hence dealt with in a number of alternative ways including the effective parameter method and statistical schemes for representing this variability (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**). The most sophisticated methods of dealing with these scale issues have been developed in subsurface hydrology (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**; and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**), and to a lesser degree in land–atmosphere interactions (*see* **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**), although promising research is underway in catchment hydrology as well (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**).

LINKING IT ALL TOGETHER – FEEDBACKS AND ELEMENTS OF A THEORY

Over the past decades there has been a trend in the hydrological sciences for a more comprehensive representation of hydrological processes, moving from an isolated description of one particular component of the hydrologic cycle to integrating hydrology with biogeochemical processes (KNAW, 2005). To a large degree, this is reflected in the many articles of this encyclopedia that deal with process links. This trend has been triggered both by the increase in computing power and a realization that feedbacks in the hydrological cycle may be more important than traditionally acknowledged, particularly for climate impacts. Changes in development paradigms in society seem to have played a major role also **Chapter 203, A Guide to International Hydrologic Science Programs, Volume 5**; Falkenmark, 1991). Feedbacks are manifold and occur at many scales, and they involve a range of other disciplines. In the feedback between the water and energy balances at the land surface, soil moisture plays a crucial role. At longer time scales, there exist feedbacks between hydrology and landscape evolution and feedbacks between hydrology and soil formation (*see* **Chapter 4, Organization and Process, Volume 1**). Feedbacks between hydrological water dynamics and biological processes occur at many scales and in many ways, for example, in subsurface flow and transport through microbial activity (*see* **Chapter 105, Microbial Transport in the Subsurface, Volume 3**), in soil formation (Jenny, 1980), in the vegetation dynamics at the land–atmosphere interface (*see* **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**) and in erosion processes (*see* **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**), and through biofilm and macrophyte dynamics in open channel flow and transport (e.g. Battin and Sengschmitt, 1999; Stephan and Gutknecht, 2002). One avenue to address feedbacks has been to link the various processes by coupled models (Bronstert *et al.*, 2005). The strength of this avenue is that the experience with models for each of the processes to be coupled is usually available, but model complexity may limit the practical applicability. The other avenue has been to make the feedbacks themselves a focus of theoretical, quantitative research and this seems to be an emerging area of hydrology, particularly the interactions of vegetation, land surface hydrology and climate (*see* **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**). Through transpiration and photosynthesis, the vegetation links the energy, water, and biogeochemical cycles. A number of strategies have been put forward to explain the functioning of vegetation dynamics such as those based on ecological optimality hypotheses (Eagleson, 1998; **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**).

This line of research focusing on feedbacks may assist in addressing a certain tendency of fragmentation of the subdisciplines in hydrology with papers "digging the same hole deeper" prevailing over comprehensive views as pointed out by some analysts (Burges, 1998). Although diversity of approach has great advantages and has probably been one of the strengths of the hydrological sciences, it is also important to stimulate a process of seeing how one picture fits with another. Hydrologists are now actively thinking about what may be the elements of a theory of hydrological sciences (*see* **Chapter 13, Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, Volume 1**). Most hydrologists would probably agree that theories for some hydrologic processes exist – a linear theory of the rainfall runoff relationship (Dooge, 1973), a theory of infiltration (e.g. Smith *et al.*, 2002), a theory of stochastic hydrogeology (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**), but a comprehensive hydrologic theory in its own right is still lacking. It is important that a theory is different from a model in many respects – a theory would have to apply to a variety of circumstances, for example, a range of climates, geological settings, and a range of scales. Similar to a model, one would expect it to be predictive and it must be falsifiable. From a theory one would expect that it has been so thoroughly tested and developed that we know there is indeed some range of phenomena for which they give correct predictions *every time*. With current hydrologic models this does not seem to be the case. These theories would always remain part of our understanding of hydrology, even when new findings take us beyond them in certain ways. The theory would not be invalidated, but rather extended, by new findings. Clearly, the status of a theory is more than that of a model.

There may be still some way until a formulation of this theory becomes viable, but there is value in speculating about elements that may assist in putting it together (*see* **Chapter 13, Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, Volume 1**). A new theory may involve an increased focus on interactions and feedbacks between different processes such as those involving vegetation, as this would entail a broadening of the scientific perspectives. To address the generalization issue, a theory may use comparative hydrology (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**) to develop a common method for assessing and quantifying hydrological similarity through comparisons between catchments in different hydrologic regimes. The theory would have to be valid for all these regimes. Patterns of hydrological response should perhaps be given particular attention to isolate the processes that have led to them, to reconcile the catchment functioning

with the observations, and for testing hypotheses about process interactions and feedbacks. The level of complexity of a theory will clearly be an important consideration. An elegant, parsimonious theory may be favored over a more complex one provided it captures the essential complexity as suggested by the Occam's razor principle (Sivapalan *et al.*, 2003). In this respect, characteristic scales and scaling concepts (Skøien *et al.*, 2003) that focus on the order of magnitudes (similar to fluid dynamics) may assist. In a broader Earth Science context one may wonder where the place is of hydrology in the realms of physics and biology (Sivapalan, 2003). Harte (2002) noted: "Physicists seek simplicity in universal laws. Ecologists revel in complex interdependencies. A sustainable future for our planet will probably require a look at life from both sides". In a similar vein it is likely that hydrology will have to adopt some of the more complex and less universal concepts ecology is rich in, in addition to the traditional quest for physical concepts. For a hydrological theory to become influential, it will likely have to combine elements from both physics and ecology.

Acknowledgments

I would like to thank the many individuals who have shaped my own evolving ideas on hydrology over the years, in particular, Murugesu Sivapalan, Rodger Grayson, and Dieter Gutknecht. I owe much to all of them. I am indebted to Murugesu Sivapalan who contributed to this article through numerous discussions and to Alfred Paul Blaschke, Uwe Haberlandt, Ralf Merz, Dan Moore, Kurt Roth, Jan Szolgay, David Tarboton, Andrew Western, and Ross Woods for offering advice and many constructive criticisms, which have helped to substantially improve the article. Funding from the Austrian Academy of Sciences Project HÖ 18 is gratefully acknowledged.

REFERENCES

Anderson M.P. (1997) Characterization of geological heterogeneity. In *Subsurface Flow and Transport: A Stochastic Approach*, Dagan G. and Neuman S.P. (Eds.), International hydrology series. University Press: Cambridge, pp. 23–43.

Baird A.J. (1996) Continuity in hydrological systems. In *Contemporary Hydrology*, Wilby R.L. (Ed.), John Wiley and Sons: Chichester, pp. 25–58.

Battin T.J. and Sengschmitt D. (1999) Linking sediment biofilms, hydrodynamics, and river bed clogging: evidence from a large river. *Microbial Ecology*, **37**(3), 185–196.

Beven K.J. (1989) Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology*, **105**, 157–172.

Blöschl G. (2001) Scaling in hydrology. *Hydrological Processes*, **15**, 709–711.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling – a review. *Hydrological Processes*, **9**(3–4), 251–290.

Bronstert A., Carrera J., Kabat P. and Lütkemeier S. (Eds.) (2005) *Coupled Models for the Hydrological Cycle*, Springer: Berlin, Heidelberg, New York, p. 356.

Burges S.J. (1998) Streamflow prediction: capabilities, opportunities, and challenges. In *Hydrologic Sciences – Taking Stock and Looking Ahead*, NRC (Ed.), National Academy Press: Washington, pp. 101–134.

Dooge J.C.I. (1973) *Linear Theory of Hydrologic Systems*, Technical Bulletin No. 1468, US Department of Agriculture, Agricultural Research Service, Washington.

Dooge J.C. (1986) Looking for hydrologic laws. *Water Resources Research*, **22**, 46S–58S.

Dunne T. (1998) Hydrologic science ... in landscapes ... on a planet ... in the future. In *Hydrologic Sciences – Taking Stock and Looking Ahead*, NRC (Ed.), National Academy Press: Washington, pp. 10–43.

Eagleson P.S. (1986) The emergence of global-scale hydrology. *Water Resources Research*, **22**, 6S–14S.

Eagleson P.S. (1998) *Ecohydrology*, Cambridge University Press.

Falkenmark M. (1991) Environmental management – what is the role of hydrologists? *Proceedings of the International Symposium to Commemorate 25 Years of the IHD/IHP*, UNESCO: Paris, pp. 61–80.

Freeze R.A. and Harlan R.L. (1969) Blueprint for a physically-based, digitally simulated hydrologic response model. *Journal of Hydrology*, **9**, 237–258.

Gelhar L.W., Bakr A.A., Gutjahr A.L. and MacMillan J.R. (1977) Comments on 'A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media' by R. Allan Freeze. *Water Resources Research*, **13**, 477–479.

Grayson R.B. and Blöschl G. (Eds.) (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Cambridge University Press: Cambridge, p. 404.

Grayson R.B., Moore I.D. and McMahon T.A. (1992) Physically-based hydrologic modelling, 2. Is the concept realistic? *Water Resources Research*, **28**, 2659–2666.

Gutknecht D. (1993) Grundphänomene hydrologischer Prozesse (Basic characteristics of hydrological processes). In *Current Issues in Hydrology*, Zürcher Geographische Schriften. 53, Geographical Institute ETH: Zürich, pp. 25–38.

Harte J. (2002) Toward a synthesis of the Newtonian and Darwinian world views. *Physics Today*, **55**(10), 29–35.

Horton R.E. (1945) Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology. *Geological Society of America Bulletin*, **56**, 275–370.

Hubbert M.K. (1956) Darcy's law and the field equations of the flow of underground fluids. *Transactions of the American Institute of Mining and Metallurgical Engineering*, **207**, 222–239.

Hurst H.E. (1951) Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 770–808.

Jenny H. (1980) *The Soil Resource*, Springer: New York, p. 377.

Kirchner J.W. (2003) A double paradox in catchment hydrology and geochemistry. *Hydrological Processes*, **17**, 871–874.

Klemeš V. (1983) Conceptualisation and scale in hydrology. *Journal of Hydrology*, **65**, 1–23.

KNAW (2005) *Turning the Water Wheel Inside out, Foresight Study on Hydrological Science in the Netherlands*, Royal Netherlands Academy of Arts and Science: Amsterdam.

Levy M. and Berkowitz B. (2003) Measurement and analysis of non-Fickian dispersion in heterogeneous porous media. *Journal of Contaminant Hydrology*, **64**(2003), 203–226.

Mandelbrot B.B. and Wallis J.R. (1968) Noah, Joseph, and operational hydrology. *Water Resources Research*, **4**, 909–918.

NRC, National Research Council (1991) *Opportunities in the Hydrologic Sciences*, National Academy Press: Washington, p. 348.

Rinaldo A., Rodríguez-Iturbe I., Rigon R. and Bras R.L. (1993) Self-organized fractal river networks. *Physical Review Letters*, **70**, 822–825.

Savenije H.H.G. (2001) Equifinality, a blessing in disguise? *Hydrological Processes*, **15**, 2835–2838.

Schumm S.A. (1991) *To Interpret the Earth: Ten Ways to Be Wrong*, Cambridge University Press: Cambridge, p. 143.

Shuttleworth W.J. (1988) Macrohydrology – the new challenge for process hydrology. *Journal of Hydrology*, **100**, 31–56.

Sivapalan M. (2003) Process complexity at hillslope scale, process simplicity at the watershed scale: is there a connection? *Hydrological Processes*, **17**, 1037–1041.

Sivapalan M., Blöschl G., Merz R. and Gutknecht D. (2005) Linking flood frequency to long-term water balance: incorporating effects of seasonality. *Water Resources Research*, **41**, doi:10.1029/2004WR003439.

Sivapalan M., Blöschl G., Zhang L. and Vertessy R. (2003) Downward approach to hydrological prediction. *Hydrological Processes*, **17**, 2101–2111.

Sivapalan M., Kumar P. and Harris D. (2001) Nonlinear propagation of multiscale dynamics through hydrologic subsystems. *Advances in Water Resources*, **24**(9–10), 935–940.

Skøien J.O., Blöschl G. and Western A.W. (2003) Characteristic space scales and timescales in hydrology. *Water Resources Research*, **39**(10), 1304, 10.1029/2002WR001736.

Smith R.K., Smettem R.J., Broadbridge P. and Woolhiser D.A. (2002) *Infiltration Theory for Hydrologic Applications*, *Water Resources Monograph Series*, American Geophysical Union: Vol. 15: p. 210.

Stephan U. and Gutknecht D. (2002) Hydraulic resistance of submerged flexible vegetation. *Journal of Hydrology*, **269**(1–2), 27–43.

Woo M. (2004) Boundary and border considerations in hydrology. *Hydrological Processes*, **18**, 1185–1194.

Zehe E. and Blöschl G. (2004) Predictability of hydrologic response at the plot and catchment scales: role of initial conditions. *Water Resources Research*, **40**, 21, W10202, doi:10.1029/2003WR002869.

2: The Hydrologic Cycles and Global Circulation

TAIKAN OKI

International Center for Urban Safety Engineering, Institute of Industrial Science, University of Tokyo, Tokyo, Japan

The role of hydrological cycles in earth system, the amount of water on the Earth's surface and its distribution in various reserves are first introduced together with the water cycles on the Earth. The concept of mean residence time, water stored in various parts over the Earth surface as various phases, such as glacier, soil moisture, water vapor, and water flux among these reserves, such as precipitation, evaporation, transpiration, and runoff, are briefly explained with their role in global climate system, and their quantitative estimates are presented. Detailed annual water balances over land estimated by land surface models are introduced, as well. Water balance requirements over land, atmosphere, and their coupled system are explained with some application of the concept, and the role of rivers in the global hydrological cycle is quantitatively shown. Finally, the impact of increasing magnitude of anthropogenic activities on global water system and its relationships with global changes are briefly discussed.

EARTH SYSTEM AND WATER

The Earth system is unique in that water exists in all three phases, that is, water vapor, liquid water, and solid ice, compared to the situations in other planets. The transport of water vapor is regarded as the energy transport because of its large amount of latent heat exchange during phase change to liquid water (approximately $2.5 \times 10^6\,\mathrm{J\,kg^{-1}}$); therefore, water cycle is closely linked to energy cycle. Even though the energy cycle on the Earth is an "Open System" driven by solar radiation, the amount of water on the Earth does not change on shorter than geological timescales (Oki, 1999), and the water cycle itself is a "Closed System."

On global scale, hydrologic cycles are associated with atmospheric circulation, which is driven by the unequal heating of the earth's surface and atmosphere in latitude (Peixóto and Oort, 1992).

Annual mean absorbed solar energy at the top of the atmosphere is maximum near equator with approximately $300\,\mathrm{W\,m^{-2}}$, decreases suddenly at higher latitudes, and is approximately $60\,\mathrm{W\,m^{-2}}$ at Arctic and Antarctic regions. Emitted terrestrial radiative energy from the Earth at the top of the atmosphere is approximately $250\,\mathrm{W\,m^{-2}}$ for ± 20 degree north and south, gradually decreases at higher latitudes, and is approximately $175\,\mathrm{W\,m^{-2}}$ at Arctic region and $150\,\mathrm{W\,m^{-2}}$ at Antarctic region. As a consequence, the net annual energy balance is positive (absorbing) for tropical and subtropical regions in ± 30 degree north and south, and negative in higher latitudes (Dingman, 2002) (*see* **Chapter 39, Surface Radiation Balance, Volume 1**).

If there is no atmospheric and oceanic circulation on the Earth, temperature difference on the Earth should have been more drastic; temperature in equatorial zone should have been high enough that the outgoing terrestrial radiation balances the absorbed solar energy and temperature in the polar regions, which should have been low enough, as well. In reality, there are atmospheric circulations and oceanic circulations that lessen this expected temperature gradation in the absence of circulations.

Both atmosphere and ocean carry energy from the equatorial region toward both the polar regions. In the case of atmosphere, the energy transport consists of sensible heat and latent heat fluxes (Masuda, 1988). The global water circulation is this latent heat transport itself, and water plays active role in the atmospheric circulation; it is not a passive compound of the atmosphere, but it affects atmospheric circulation by both radiative transfer and latent heat release of phase change.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

WATER RESERVES, FLUXES, AND RESIDENCE TIME (*see* Chapter 173, Global Water Cycle (Fundamental, Theory, Mechanisms), Volume 5)

The total volume of water on the Earth is estimated as approximately 1.4×10^{18} m^3, and it corresponds to a mass of 1.4×10^{21} kg. Compared with the total mass of the Earth (5.974×10^{24} kg), the mass of water constitutes only 0.02% of the planet, but it is critical for the survival of life on the Earth and the Earth is called *Blue Planet* and *Living Planet*.

There are various forms of water on the Earth's surface. Approximately 70% of its surface is covered with salty water, the oceans. Some of the remaining areas (continents) are covered by freshwater (lakes and rivers), solid water (ice and snow), and vegetation (which implies the existence of water). Even though the water content of the atmosphere is comparatively small (approximately 0.3% by mass and 0.5% by volume of the atmosphere), approximately 60% of the Earth is always covered by cloud (Rossow *et al.*, 1993). The Earth is the planet whose surface is dominated by the various phases of water.

Water on the Earth is stored in various reserves, and various water flows transport water from one to another. Water flow (mass or volume) per unit time is also called *water flux*.

The mean residence time in each reserve can be simply estimated from total storage volume in the reserve and the mean flux rate to and from the reserve;

$$T_{\text{mean}} = \frac{\text{Total Storage Volume}}{\text{Mean Flux Rate}} \qquad (1)$$

there is even a distribution of flux rate coming in and going out from the storage (Chapman, 1972). The last column of Table 1 presents some values of the global mean residence time of water. Evidently, the water cycle on the Earth is a "Stiff" differential system with variability on many timescales, from a few weeks through thousands of years.

The mean residence time is also important to consider when water quality deterioration and restoration are discussed, since the mean residence time can be an index of how much water is turned over. Apparently, river water or surface water is more vulnerable than groundwater to be polluted; however, any measure to recover better water quality works faster for river water than groundwater. Since major interests of hydrologists have been the assessment of volume, inflow, outflow, and chemical and isotopic composition of water, the estimation of the mean residence time of certain domain has been one of the major targets of hydrology.

EXISTENCE OF WATER ON THE EARTH (*see* Chapter 4, Organization and Process, Volume 1; Chapter 25, Global Energy and Water Balances, Volume 1)

Table 1 (simplified from a table in Korzun 1978) introduces how much water is stored in which reserves on the Earth:

- The proportion in the *ocean* is large (96.5%). Even though classical hydrology has traditionally excluded ocean processes, the global hydrological cycle is never closed without including them. The ocean circulation carries huge amounts of energy and water. The surface ocean currents are driven by surface wind stress, and the

Table 1 World water reserves. Simplified from Table 9 of "World water balance and water resources of the earth" by UNESCO Korzun, 1978. The last column, mean residence time, is from Table 34 of the report

Form of water	Covering Area (km^2)	Total Volume (km^3)	Mean Depth (m)	Share (%)	Mean Residence Time
World ocean	361 300 000	1 338 000 000	3700	96.539	2500 years
Glaciers and permanent snow cover	16 227 500	24 064 100	1463	1.736	1600 years
Ground water	134 800 000	23 400 000	174	1.688	1400 years
Ground ice in zones of permafrost strata	21 000 000	300 000	14	0.0216	10 000 years
Water in lakes	2 058 700	176 400	85.7	0.0127	17 years
Soil moisture	82 000 000	16 500	0.2	0.0012	1 years
Atmospheric water	510 000 000	12 900	0.025	0.0009	8 days
Marsh water	2 682 600	11 470	4.28	0.0008	5 years
Water in rivers	148 800 000	2120	0.014	0.0002	16 days
Biological water	510 000 000	1120	0.002	0.0001	a few hours
Artificial Reservoirs		8000			72 days
Total water reserves	510 000 000	1 385 984 610	2718	100.00	

atmosphere itself is sensitive to the sea surface temperature. Temperature and salinity determine the density of ocean water, and both factors contribute to the overturning and the deep ocean general circulation (*see* **Chapter 174, Global Water Budgets – Fundamental Theory and Mechanisms, Volume 5**).

- Other major reserves are solid water on the continent (glaciers and permanent snow cover) and groundwater. *Glacier* is accumulation of ice of atmospheric origin generally moving slowly on land over a long period. Glacier forms discriminative U-shaped valley over land, and remains moraine when it retreats. If a glacier "flows" into an ocean, the terminated end of the glacier often forms an iceberg. Glaciers react in comparatively longer timescale against climatic change, and they also induce isostatic responses of continental scale upheavals or subsidence in even longer timescale. Even though it is predicted that the thermal expansion of oceanic water dominates the anticipated sea level rise due to the global warming, glaciers over land are also a major concern as the cause of sea level rise associated with global warming (*see* **Chapter 162, Hydrology of Snowcovered Basins, Volume 4**; **Chapter 164, Role of Glaciers and Ice Sheets in Climate and the Global Water Cycle, Volume 4**).
- *Groundwater* is the subsurface water occupying the saturated zone. It contributes to runoff in its low-flow regime, between floods. Deep groundwater may also reflect the long term climatological situation. Groundwater in Table 1 includes both gravitational and capillary water. Gravitational water is water in the unsaturated zone (vadose zone) which moves under the influence of gravity. Capillary water is water found in the soil above the water table by capillary action, a phenomenon associated with the surface tension of water in soils acting as porous media. Groundwater in Antarctica (roughly estimated as $2 \times 10^6\,\mathrm{km}^3$) is excluded from Table 1 (*see* **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4**).
- *Soil moisture* is the water being held above groundwater table. It influences the energy balance at the land surface as a lack of available water suppresses evapotranspiration and as it changes surface albedo. Soil moisture also alters the fraction of precipitation partitioned into direct runoff and percolation. The water accounted for in runoff cannot be evaporated from the same place, but the water infiltrated into soil may be uptaken by suction, and evaporated again (*see* **Chapter 72, Measuring Soil Water Content, Volume 2**).
- The atmosphere carries *water vapor*, which influences the heat budget as latent heat. Condensation of water releases latent heat, heats up the atmosphere, and affects the atmospheric general circulation. *Liquid water in the atmosphere* is another result of condensation. *Clouds* significantly change the radiation in the atmosphere and at the Earth's surface. However, as a volume, liquid (and solid) water contained in the atmosphere is quite little, and most of the water in the atmosphere exists as water vapor. *Precipitable water* is the total water vapor in atmospheric column from land surface to the top of the atmosphere. Water vapor is also the major absorber in the atmosphere of both short-wave and long-wave radiation (*see* **Chapter 196, The Role of Water Vapor and Clouds in the Climate System, Volume 5**).
- *Water in rivers* is very tiny as a stored water at each time, however, the recycling speed, which can be estimated as the inverse of the mean residence time, of river water (river discharge) is relatively high, and it is important because most social applications ultimately depend on water as a renewable and sustainable resource.

The amount of water stored transiently in a soil layer, in the atmosphere, and in river channels is relatively minute, and the time spent through these subsystems is short, but, of course, they play dominant roles in the global hydrological cycle.

WATER CYCLE ON THE EARTH (*see* Chapter 4, Organization and Process, Volume 1)

The water cycle plays many important roles in the climate system, and Figure 1 schematically illustrates the various flow path of water (Oki, 1999). Values are taken from Table 1 and also calculated from the precipitation estimates by Xie and Arkin (Xie and Arkin, 1996). Precipitable water, water vapor transport, and its convergence are estimated using ECMWF objective analyses, obtained as four-year mean from 1989 to 1992. The roles of these water fluxes in the global hydrologic system are now briefly introduced (*see* **Chapter 182, The Hydrological Cycle in Atmospheric Reanalysis, Volume 5**):

- *Precipitation* is water flux from atmosphere to land or ocean surface. It drives the hydrological cycle over land surface and changes surface salinity (and temperature) over the ocean and affects its thermohaline circulation. *Rainfall* refers to the liquid phase of precipitation. Part of it is intercepted by canopy over vegetated area, and remaining part reaches the Earth surface as through-fall. Highly variable, intermittent, and concentrated behavior of precipitation in time and space domain compared to other major hydrological fluxes mentioned below makes the observation of this quantity and the aggregation of the process complex and difficult (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**).
- *Snow* has special characteristics compared with rainfall. Snow may be accumulated, the albedo of snow is quite high (as high as clouds), and the surface temperature

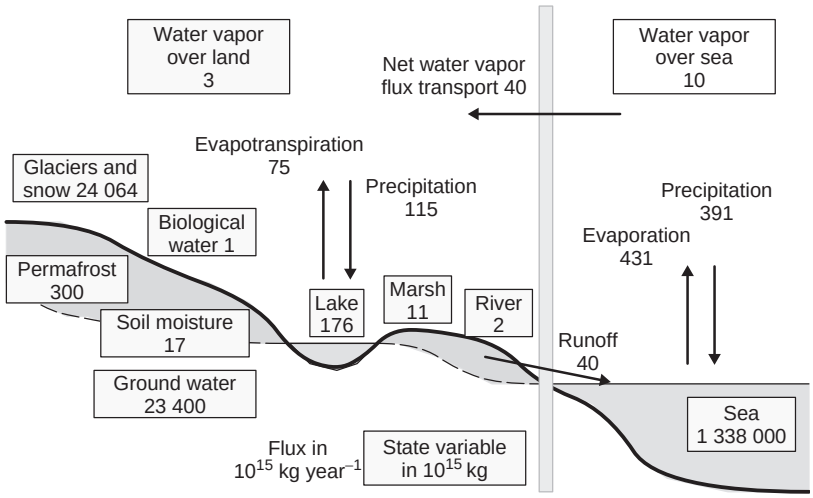


Figure 1 Schematic illustration of the water cycle on the Earth Oki, 1999. Values are taken from Table 1 and calculated from atmospheric water vapor data by European Centre for Medium-Range Weather Forecasts (ECMWF) and precipitation by Xie and Arkin (Xie and Arkin, 1996) for 1989–92

will not rise above 0 °C until the completion of snow melt. Consequently the existence of snow changes the surface energy and water budget enormously. A snow surface typically reduces the aerodynamic roughness, so that it may also have a dynamical effect on the atmospheric circulation and hydrologic cycles (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**; **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**; **Chapter 45, Actual Evaporation, Volume 1**).

- *Evaporation* is the return flow of water from the surface to the atmosphere and gives the latent heat flux from the surface. The amount of evaporation is determined by both atmospheric and hydrological conditions. From the atmospheric point of view, the fraction of incoming solar energy to the surface leading to latent and sensible heat flux is important. Wetness at the surface influences this fraction because the ratio of actual evapotranspiration to the potential evaporation is reduced due to drying stress. The stress is sometimes formulated as a resistance, and such a condition of evaporation is classified as hydrology-driven. If the land surface is wet enough compared to the available energy for evaporation, the condition is classified as atmosphere-driven.
- *Transpiration* is the evaporation of water through stomata of leaves. It has two special characteristics different from evaporation from soil surfaces. One is that the resistance of stomata is related not only to the dryness of soil moisture but also to the physiological conditions of the vegetation through the opening and the closing of stomata. Another is that roots can transfer water from deeper soil than in the case of evaporation from bare soil. Vegetation also modifies the surface energy and water balance by altering the surface albedo and by intercepting precipitation and evaporating this rain water (*see* **Chapter 42, Transpiration, Volume 1**).
- *Runoff* returns water to the ocean which may have been transported in vapor phase by atmospheric advection for inland. The runoff into oceans is also important for the freshwater balance and the salinity of the oceans. Rivers carry not only water mass but also sediment, chemicals, and various nutritional matters from continents to seas. Without rivers, global hydrologic cycles on the earth will never close (*see* **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**; **Chapter 114, Snowmelt Runoff Generation, Volume 3**; **Chapter 111, Rainfall Excess Overland Flow, Volume 3**; **Chapter 112, Subsurface Stormflow, Volume 3**).

  Runoff at hill-slope scale is nonlinear and a complex process. Surface runoff could be generated when rainfall or snow melt intensity exceeds the infiltration rate of the soil or precipitation falls over saturated land surface. Saturation at land surface can be formed mostly by topographic concentration mechanism along hill-slopes. Infiltrated water in the upper part of the hill-slope flows down the slope and discharges at the bottom of the hill-slope. Because of the highly variable heterogeneity of topography, soil properties such as conductivity and porosity, and precipitation, basic equations, such as Richard's Equation, which can express the runoff process fairly well at a point scale or hill-slope scale, cannot be directly applied for macroscale because of its nonlinearity.

The global water cycle unifies these components consisting of the state variables (precipitable water, soil moisture, etc.) and the fluxes (precipitation, evaporation, etc.).

WATER BALANCE REQUIREMENTS (*see* Chapter 5, Fundamental Hydrologic Equations, Volume 1)

The conservation law of water mass in any arbitrary control volume implies a water balance. In this section, the water balance over land, for atmospheric column, and their combination are presented (Oki, 1999). Some applications are introduced, as well.

Water Balance at Land Surface

In the field of hydrology, river basins have commonly been selected for study, and water balance has been estimated using ground observations, such as precipitation, runoff, and storage in lakes and/or groundwater.

The water balance over land is described as,

$$\frac{\partial S}{\partial t} = P - E - R_o - R_u \tag{2}$$

where S represents the water storage within the area, t is time, $(\partial S/\partial t)$ is the change of total water storage with time, P is precipitation, E is evapotranspiration, R_o is surface runoff, and R_u is the groundwater movement. S includes snow accumulation in addition to soil moisture, groundwater, and surface water storage including retention water within the control volume, defined by the area of concern over land with bottom generally at the impermeable bedrock. These terms are shown in Figure 2(a). Equation (1) means water storage over land increases by precipitation, and decreases by evapotranspiration, surface runoff, and groundwater movement.

If the area of water balance is set within an arbitrary boundary, R_o represents the net outflow of water from the region of consideration (i.e. the outflow minus total inflow from surrounding areas). Generally it is not easy to estimate groundwater movement R_u, and the net flux per unit area within a large area is expected to be comparatively small. If all groundwater movement is considered to be that observed at the gauging point of a river ($R_u = 0$), and equation (2) becomes:

$$\frac{\partial S}{\partial t} = P - E - R_o \tag{3}$$

This assumption is generally valid at the outlet of a catchment. In most cases, surface river runoff R_o becomes river discharge through river channel network, and can be observed at a point unlike other fluxes, such as P and E that should be spatially measured.

Water Balance in the Atmosphere

Atmospheric water vapor flux convergence gives water balance information that can complement the traditional hydrological elements such as precipitation, evapotranspiration, and discharge. The basic concept and an application of using atmospheric data to estimate the terrestrial water balance was presented by Starr and Peixóto (1958).

The atmospheric water balance for a column of atmosphere from the bottom at land surface to the top of the atmosphere is described by the equation,

$$\frac{\partial W}{\partial t} = Q + (E - P) \tag{4}$$

where, W represents precipitable water (*i.e.* column storage of water vapor), Q is the convergence of water vapor flux in the atmosphere. Since the water content in the atmosphere in the solid and liquid phases are generally small, only vapor phase of water is considered in equation (4). The balance is schematically illustrated in Figure 2(b), and describes that the water storage in an atmospheric column is increased by the horizontal convergence of water vapor and evapotranspiration from bottom of the column (land surface), and decreases by precipitation which goes out from the bottom of the atmosphere to land.

Combined Atmosphere–River Basin Water Balance

Equations (3) and (4) can be combined into:

$$-\frac{\partial W}{\partial t} + Q = (P - E) = \frac{\partial S}{\partial t} + R_o \tag{5}$$

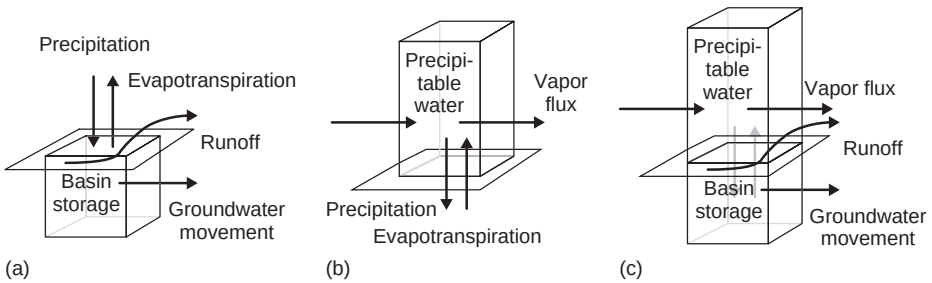


Figure 2 (a) Terrestrial water balance, (b) Atmospheric water balance, and (c) Combined atmosphere-land surface water balance. (a), (b), and (c) correspond to equations (3), (4), and (6), respectively

Figure 2(c) illustrates the balance in this equation, and illustrates that the difference of precipitation and evapotranspiration is equal to the sum of the decrease of atmospheric water vapor storage and horizontal convergence, and also to the sum of the increase of water storage over land and runoff. Theoretically, equation (5) can be applied for any control volume of land area combined with the atmosphere above, even though practical applicability depends on the accuracy and availability of atmospheric and hydrologic information.

The following further assumptions are often employed in annual water balance computations:

- Annual change of atmospheric water vapor storage is negligible ($(\partial W/\partial t) = 0$).
- Annual change of water storage at land is negligible ($(\partial S/\partial t) = 0$).

With these assumptions, equation (5) simplifies to:

$$Q = (P - E) = R_o \tag{6}$$

If a river basin is selected as the water balance region, R_o is simply the discharge from the basin. The simplified equation 6 demands that the water vapor convergence, "precipitation-evaporation", and net runoff should balance over the annual period when the temporal change of storage terms can be neglected.

Estimation of Large-scale Evapotranspiration

The equation

$$E = \frac{\partial W}{\partial t} - Q + P \tag{7}$$

obtained from equation (4) can be applicable over periods shorter than a year, unlike the assumption adopted for equation (6). If atmospheric data with precipitation data are available over short timescales such as months or days, evapotranspiration can be estimated at the corresponding timescales, of course, subject to severe limitations imposed by the accuracy of the data. The region over which the evapotranspiration is estimated is not limited to a river basin but depends only on the scales of the available atmospheric and precipitation data.

Estimation of Total Water Storage in a River Basin

Equations (3) and (4) give

$$\frac{\partial S}{\partial t} = -\frac{\partial W}{\partial t} + Q - R_o \tag{8}$$

which indicates that the change of water storage in the control volume over land can in principle be estimated from atmospheric and runoff data. Although an initial value is required to obtain the absolute value of storage, the atmospheric water balance can be useful in estimating the seasonal change of total water storage in large river basins.

Estimation of Zonally averaged Net Transport of Fresh Water

The meridional (north–south direction) distribution of the zonally averaged annual *energy* transports by the atmosphere and the oceans have been evaluated, even though there are quantitative problems in estimating such values (Trenberth and Solomon, 1994). However, the corresponding distribution of *water* transport has not often been studied although the cycles of energy and water are closely related. Wijffels *et al.*, (1992) used values of Q from Bryan and Oort (1984) and discharge data from Baumgartner and Reichel (1975) to estimate the freshwater transport by oceans and atmosphere, but their results seem to have large uncertainties and they did not present the freshwater transport by rivers.

The annual freshwater transport in the meridional (north–south) direction can be estimated from Q and river discharge with geographical information such as the location of river mouths and basin boundaries (Oki *et al.*, 1995). Results are introduced in the next section.

RIVERS IN GLOBAL HYDROLOGICAL CYCLE

The freshwater supply to the ocean has an important effect on the thermohaline circulation because it changes the salinity and thus the density. The impacts of freshwater supply to ocean are enhanced in the case of large river basins because they concentrate freshwater from large area to their river mouths.

It also controls the formation of sea ice and its temporal and spatial variations. Annual freshwater transport by rivers and the atmosphere to each ocean is summarized in Table 2 based on the atmospheric water balance (Oki, 1999). Some part of the water vapor flux convergence remains in the inland basins. There are a few negative values in Table 2, suggesting that net freshwater transport occurs from the ocean to the continents. This is physically impossible and is caused by errors in the source data. Although a detailed discussion of the values in Table 2 may not be meaningful, it is nevertheless interesting that such an analysis does make at least qualitative sense using the atmospheric water balance method with geographical information on basin boundaries and the location of river mouths. In this analysis, it should be noted that the total amount of freshwater transport into the oceans from the surrounding continents has the same order of magnitude as the freshwater supply that comes directly from the atmosphere, expressed by Q.

The annual freshwater transport in the meridional direction has been also estimated based on atmospheric water

Table 2 Annual freshwater transport from continents to each ocean (10^{15} kg year^{-1}) mean for 1985–88. 'Inner' indicates the runoff to the inner basin within Asia and Africa. $-\nabla_H \cdot \vec{Q}$ indicates the direct freshwater supply from the atmosphere to the ocean. N.P., S.P., N.At., and S.At. represent North Pacific, South Pacific, North Atlantic, and South Atlantic Ocean

		N.P.	S.P.	N.At.	S.At.	Indian	Arctic	Inner	Total
	Asia	4.7	0.4	0.2		3.3	2.7	0.1	11.4
	Europe			1.7		0.0	0.7		2.4
From	Africa			−0.2	0.9	−0.2		−0.4	0.1
rivers	N. America	2.9		4.8			1.1		8.8
	S. America	0.5	0.4	5.7	8.3				14.9
	Australia		0.1			0.1			0.2
	Antarctica		1.0		0.1	0.8			1.9
From	Total	8.1	1.9	12.2	9.3	4.0	4.5	−0.3	39.7
atmosphere	$-\nabla_H \cdot \vec{Q}$	9.9	−11.1	−12.7	−14.0	−14.0	2.2		−39.7
	Grand Total	18.0	−9.2	−0.5	−4.7	−10.0	6.7	−0.3	0.0

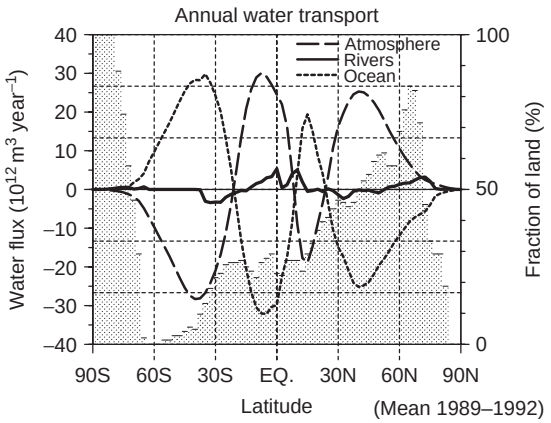


Figure 3 The annual freshwater transport in the meridional (north–south) direction by atmosphere, ocean, and rivers (land) Oki *et al.,* 1995. Water vapor flux transport of 20×10^{12} m^3 year^{-1} corresponds to approximately 1.6×10^{15} W of latent heat transport. Shaded bars behind the lines indicate the fraction of land at each latitudinal belt

balance with results shown in Figure 3. The estimates in Figure 3 are the net transport, that is in the case of oceans, it is the residual of northward and southward freshwater flux by all ocean currents globally, and it cannot be compared directly with individual ocean currents such as the Kuroshio and the Gulf Stream. It should be noted that the directions of river flows are mostly steady unlike ocean or atmospheric circulations, and concentrates the freshwater in one direction through out the year.

Transport by the atmosphere and by the ocean have almost the same absolute values at each latitude but with different signs. The transport by rivers is about 10% of these other fluxes globally (this may be an underestimation because Q tends to be smaller than the river discharge observed at a land surface). The negative (southward) peak by rivers at 30 °S is mainly due to the Parana River in South America, and the peaks at the equator and 10 °N are due to rivers in South America, such as the Magdalena and Orinoco. Large Russian rivers, such as the Ob, Yenisey, and Lena, carry the freshwater towards the north between 50–70 °N.

These results suggest that the hydrological processes over land play nonnegligible roles in the climate system, not only by the exchange of energy and water at the land surface, but also through the transport of freshwater by rivers, which affects the water balance of the oceans and forms a part of the hydrological circulation on the Earth among the atmosphere, continents, and oceans.

GLOBAL WATER BALANCE ESTIMATED BY LAND SURFACE MODELS

The values quoted in Table 1 and Figure 1 are estimated based on various observations with some assumptions in order to obtain global perspectives. These values are sometimes different in other references probably because source of observed data, methodology to estimate, and assumptions are different. In some cases, global water balance are estimated using empirical relationship of evapotranspiration to precipitation in each latitude (Baumgartner and Reichel, 1975).

Recently, under an international research project, land surface models (LSMs) were used to estimate global water and energy balances for 1986 through 1995 in order to obtain global distribution of surface soil moisture, which is not easy to obtain but relevant for understanding the land-atmosphere interactions (IGPO, 2002). The project was called the *Global Soil Wetness Project* (GSWP) and its goal was to produce state-of-the-art global data sets of land surface fluxes, state variables, and related hydrologic quantities (*see* **Chapter 178, Modeling of the Global Water Cycle: Numerical Models (General Circulation Models), Volume 5**; **Chapter 201,**

Land-Atmosphere Models for Water and Energy Cycle Studies, Volume 5).

In the second phase of the project (GSWP-2), Meteorological forcing are hybrid products of National Center for Environmental Prediction (NCEP)/Department of Energy (DOE) reanalysis observational data and satellite data, and provided at 3-hourly time step for a period of thirteen and half years from July 1982 to December 1995. The first three and half years data is used for spin up. The land surface parameters are specified from Earth Resources Observation and Science Data Center (EDC) for land-cover data and International Geosphere–Biosphere Programme Data Information System (IGBP-DIS) for soil data. Both land surface parameters and meteorological forcing are at one degree resolution for all land grids excluding Antarctica.

Figure 4 illustrates the model derived global water balance over global land excluding ice, glacier, and lake. Numerics in the box corresponds to the 10-year mean annual value of 8 LSMs participated to GSWP2 (Oki *et al.*, 2005). The vertical ranges shown above and below the boxes indicate the maximum and minimum values in interannual variation of mean annual value among 8 LSMs. The horizontal ranges shown left and right of the boxes indicate the maximum and minimum values of intermodel variation of 10-year mean value by 8 LSMs. Generally speaking, intermodel variation exceeds the interannual variation, which suggests that the uncertainties associated with the selection of a model or a procedure is larger than the sampling error of estimating global water balance. In the case of rainfall, intermodel variation is small because common precipitation forcing was given to LSMs and the differences among LSM estimates were caused by the different judgment of rain/snow recognition by each modeling group.

The advantage of using models estimating global water balance is the ability to have more detailed insights than observation of only estimates. For example, snow over land excluding ice and glacier area is approximately 10% of total precipitation, and the ratio of surface runoff and subsurface runoff is approximately 2 : 3 in Figure 4. In some LSM, neither surface nor subsurface runoff process is considered and that is the reason why minimum values are zero. Even though now it is impossible to assess the validity of these breakdowns since there is no observational information of either the separated amount of the snow and rain, or the surface and subsurface runoff. However, on the other hand, such estimates will stimulate interest to collect and compile global information on these quantities in the future.

Further, evapotranspiration were estimated separately by bare soil evaporation (Es), evaporation from intercepted water on leaves (Ei), evaporation from open water (Ew), and transpiration from vegetation (Et) as in Figure 5, even though the intermodel variation is quite large partially because some LSM does not consider one or more of these path of evapotranspiration. Even though the values in Figure 5 are not definitive, it is interesting to see that bare soil evaporation and transpiration from vegetation are closely comparative, and interception loss is approximately

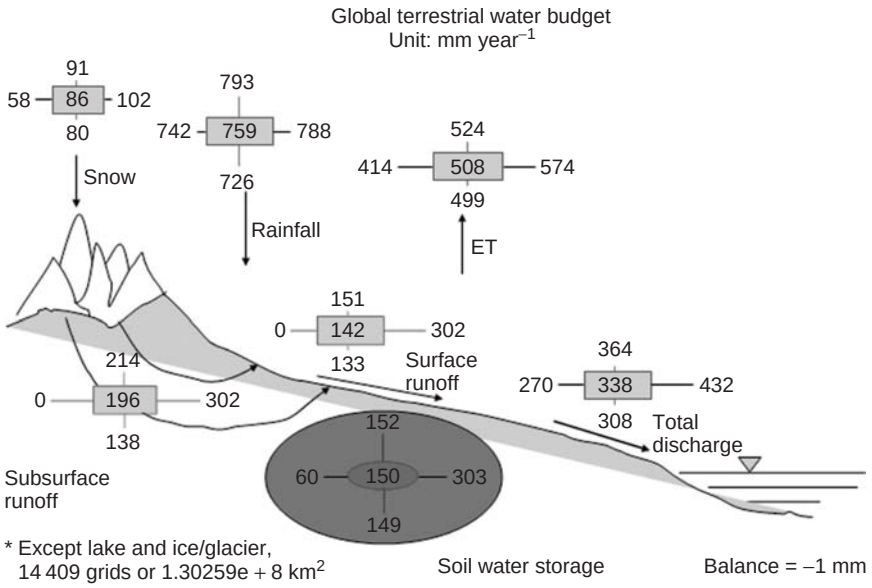


Figure 4 Global terrestrial water balance averaged for 1986–1995 estimated by eight land surface models in boxes. Interannual variation range (vertical) for 1986 through 1995 and intermodel discrepancies (horizontal) among eight models are presented for the annual mean estimates. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

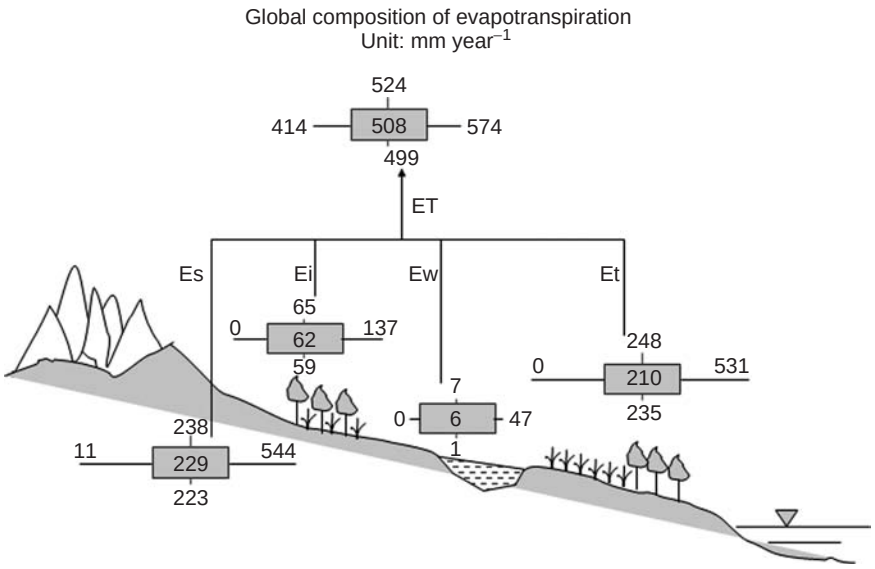


Figure 5 Global composition of evapotranspiration averaged for 1986–1995 estimated by eight land surface models in boxes. Interannual variation range (vertical) for 1986 through 1995 and intermodel descrepancies (horizontal) among eight models are presented for the annual mean estimates. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

10% of the total evapotranspiration. It would be exciting if these estimates are revised and validated by some observational measures, and intermodel discrepancies are reduced.

ANTHROPOGENIC EFFECTS ON THE GLOBAL WATER CYCLE (*see* Chapter 34, Climate Change – Past, Present and Future, Volume 1)

Global water cycles are essential in the Earth System because of flux exchange, mass and heat transport, and control on biogeochemical cycles (*see* **Chapter 189, Land Use and Water Resources Under a Changing Climate, Volume 5**).

According to the paradigm shift of research in natural sciences, after the wide recognition of global environmental problems, it is the era "The Anthropocene" for geosciences to study the real situation of the Earth (Crutzen, 2002) including the various impact of anthropogenic activities. Water cycle is one of the most exposed nature and vulnerable to human impacts. Therefore hydrological science should deal with water cycles on the Earth, its impact on human society, and the anthropogenic impact on water cycles on the Earth (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**; **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**; **Chapter 119, Land Use and Landcover Effects on Runoff Processes: Forest Harvesting and Road Construction, Volume 3**).

Human impacts on hydrologic cycles are various. Land use/land-cover transforms, topographical modification and compression of soil layers, including building cities and cultivation of different species of plants (agricultural activities), have large impacts on water cycles through changing the boundary conditions. Water withdrawals and uptake for irrigation, and municipal and industrial water usages modify water cycle significantly for both quantitative and qualitative aspects (*see* **Chapter 132, Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change, Volume 3**; **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**).

These anthropogenic impacts on surface/subsurface water cycles may have indirect effects on atmospheric circulation and regional climate, for example, deforestation may have caused long-term decrease of precipitation in particular months when large-scale circulation (such as Asian Monsoon) is not dominant for precipitation but the local boundary condition matters (Kanae *et al.*, 2001).

Figure 6 schematically illustrates the impacts of increasing population and economic activities associated with consumptive life style on hydrological cycles, water withdrawals, and resulting change in water stress. Water withdrawals are increased directly by the increase in population and water usage per capita, and indirectly through the increase in food production. Food production also changes land use, and land use is changed by industrialization, as well. Increased industrial activities and land use change are increasing the emission of the Green House Gases (GHGs), and changing climate. Any

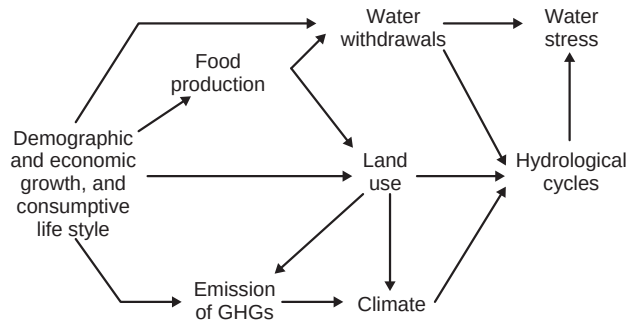


Figure 6 Diagram illustrating major pathways how demographic and economic growth have influence on the changes in hydrological cycles and water withdrawals through changes in land use, water withdrawals, and climate related to food production and the emission of the Green House Gases (GHGs)

change in both supply side (hydrological cycle) and demand side (water withdrawals) will incur adaptation in water resources management, and it will be serious if the climate change will be associated with more intense and intermittent precipitation, which will cause more frequent occurrence of floods (Milly *et al.*, 2002) and droughts (Manabe *et al.*, 2004).

Finally, globalization has increased worldwide trades, and it is associated with true water transport and "virtual water" trade. True water transport is the water transport contained in food, beverage, and other industrial products, and it occurs in local, regional, and international scales. The "virtual water" trade is not the transport of physical water, but it is a concept to consider the external cost of water consumption; namely, the virtual water content of goods is equal to the amount of water required if the transferred goods are produced in the importing/consuming region or country (Allan 1998; Oki *et al.*, 2003). Even though virtual water trade does not correspond to the amount of physical water transport, the concept is useful to assess the real water scarcity in each region, and will be utilized when water resources management issues are concerned. Hydrological sciences are often applied for such issues, and should well cooperate with such socioeconomic concepts of water (Oki *et al.*, 2004).

REFERENCES

Allan J.A. (1998) *'Virtual Water': An Essential Element in Stabilizing the Political Economies of the Middle East*, *No. 103 in Forestry & Environmental Studies Bulletin*, Yale University.

Baumgartner F. and Reichel E. (1975) *The World Water Balance: Mean Annual Global, Continental and Maritime Precipitation, Evaporation and Runoff*, Ordenbourg: München, p. 179.

Bryan F. and Oort A. (1984) Seasonal variation of the global water balance based on aerological data. *Journal of Geophysical Research*, **89**, 11 717–11 730.

Chapman T.G. (1972) *Estimating the Frequency Distribution Of Hydrologic Residence Time*, Vol. 1, IAHS-UNESCO-WMO. pp. 136–152.

Crutzen P.J. (2002) Geology of mankind – the Anthropocene. *Nature*, **415**, 23.

Dingman S.L. (2002) *Physical Hydrology Second Edition*, Prentice-Hall: p. 646.

IGPO (2002) *GSWP-2: The Second Global Soil Wetness Project Science and Implementation*, Technical report, International GEWEX Project Office, Silver Spring.

Kanae S., Oki T. and Musiake K. (2001) Impact of Deforestation on regional precipitation over the Indochina Peninsula. *Journal of Hydrometeorology*, **2**, 51–70.

Korzun V.I. (1978) World water balance and water resources of the earth, *Studies and Reports in Hydrology*, Vol. 25, UNESCO.

Manabe S., Milly P.C.D. and Wetherald R. (2004) Simulated long-term changes in river discharge and soil moisture due to global warming. *Hydrological Sciences Journal*, **49**(4), 625–642.

Masuda K. (1988) Meridional heat transport by the atmosphere and the ocean; analysis of FGGE data. *Tellus*, **40A**, 285–302.

Milly P.C.D., Wetherald R.T., Dunne K.A. and Delworth T.L. (2002) Increasing risk of great floods in a changing climate. *Nature*, **415**(6871), 514–517.

Oki T. (1999) The global water cycle. In *Global Energy and Water Cycles*, Browning K. and Gurney R. (Eds.), Cambridge University Press: pp. 10–27.

Oki T., Entekhabi D. and Harrold T. (2004) The global water cycle. In *State of the Planet: Frontiers and Challenges in Geophysics*, *No. 150 in Geophysical Monograph Series*, Sparks R. and Hawkesworth C. (Eds.), AGU Publication: p. 414.

Oki T., Hanasaki N., Shen Y., Kanae S., Masuda K. and Dirmeyer P.A. (2005) Global water balance estimated by land surface models participated in the GSWP2. Proceedings of the 19th Conference on Hydrology, San Diego, CA, USA, Amer. Met. Soc.

Oki T., Musiake K., Matsuyama H. and Masuda K. (1995) Global atmospheric water balance and runoff from large river basins. *Hydrological Processes*, **9**, 655–678.

Oki T., Sato M., Kawamura A., Miyaka M., Kanae S. and Musiake K. (2003) *Virtual Water Trade to Japan and in the World*, *No. 12 in Value of Water Research Report Series*, IHE: Delft, pp. 221–235.

Peixóto J.P. and Oort A.H. (1992) *Physics of Climate*, American Institute of Physics. p. 520.

Rossow W.B., Walker A.W. and Garder L.C. (1993) Comparison of ISCCP and Other Cloud Amounts. *Journal of Climate*, **6**, 2394–2418.

Starr V.P. and Peixóto J. (1958) On the global balance of water vapor and the hydrology of deserts. *Tellus*, **10**, 189–194.

Trenberth K.E. and Solomon A. (1994) The global heat balance: heat transports in the atmosphere and ocean. *Climate Dynamics*, **10**, 107–134.

Wijffels S.E., Schmitt R.W., Bryden H.L. and Stigebrandt A. (1992) Transport of freshwater by the oceans. *Journal of Physical Oceanography*, **22**, 155–162.

Xie P. and Arkin P.A. (1996) Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *Journal of Climate*, **9**, 840–858.

3: Hydrologic Concepts of Variability and Scale

ROSS WOODS

Catchment Processes and Water Resources Group, National Institute of Water and Atmospheric Research, Christchurch, New Zealand

All hydrological phenomena have significant variations in time and space. Typically, these variations are driven by variations in physiographic factors such as climate, soils, vegetation, topography, geology, as well as by human and animal activities. These externally driven variations then propagate through hydrological systems, leading to an extremely rich variety of hydrological variability apparent at different temporal and spatial scales, in different physical settings. Virtually any quantitative approach to this problem requires the selection of a limited set of spatial and temporal scales within a particular physiographic setting. Any particular choice of time and space scales has a major influence on which aspects of this hydrological variability are perceived. This article surveys hydrological variability in both time and space, across a range of scales.

INTRODUCTION

Variability

Variability is the change in hydrological quantity when comparing one spatial location with another, or one time with another. Variability occurs naturally, and also because of human activity (e.g. land cultivation, urbanization, and forest management). We can see evidence of this variability in measurements of rainfall, air temperature, soil moisture, snow cover, groundwater level, and streamflow, or any other hydrological quantity.

Everyday life already provides us with an intuitive understanding of the many aspects of variability, for example, from the common experience that air temperature is variable with space and time. We know about the time variation of temperature on at least three different scales: first, it is generally cooler at night and warmer during the day; second, daytime temperatures are cooler on cloudy days; and third, it is generally cooler in winter and warmer in summer. These three examples of temporal variability are at different timescales: the first is at the daily timescale, the second has no particular timescale (cloudiness may last for seconds or for days), and the third is at the annual timescale (we will define the concept of scale more carefully in the following text). We also know that air temperature is variable from place to place: it is cooler in the shade of a leafy tree than standing out in the sun, and it is cooler on the mountaintops than in nearby lowlands. We thus have experience of spatial variability for at least two space scales, the plant scale (of the order 1 m) and the landscape scale (perhaps 10 km).

Scale in this article is used to mean a spatial or temporal measure over which a hydrologic variable is being considered. For example, we may think of the amount of water held in the rooting zone of a soil at an instantaneous timescale, or an average value over the timescale of a day or a year or other period. When we choose a scale, this affects how we perceive soil moisture, or whatever other phenomena we care to think of. If we look at the moment-to-moment variability of instantaneously measured soil moisture, we see a particular variation. If we examine daily averaged soil moisture at the same place over the same period, we see something different. This difference is the effect of timescale on variability (e.g. Figure 1). Similarly, we can consider this soil water at the spatial scale of a "point" or a field average, or a catchment average. Again, a change in the scale of observation causes a change in the perceived variability. When one considers measurements or model estimates that are obtained at discrete times and locations, then the concept of scale needs to be expanded: Blöschl and Sivapalan (1995) recognized this need, and defined a "scale triplet" that quantifies the spacing, the extent, and the support of a measurement or model estimate. **Chapter 6, Principles**

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

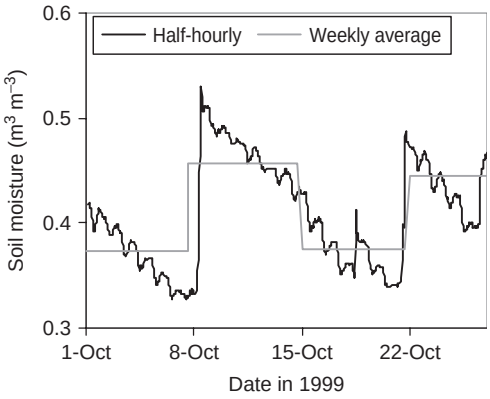


Figure 1 A comparison of half-hourly and weekly average soil moisture in the Mahurangi catchment, northern New Zealand, during October 1999, showing that temporal variability depends on our choice of temporal scale. Storms on the 8th and 21st of October caused soil moisture to rise sharply. Soil moisture also has variability at longer and shorter timescales that are not visible here

of Hydrological Measurements, Volume 1 presents this notion in more detail.

This article discusses concepts that are used to describe hydrological variability, and then examines the forms of variability typically encountered in hydrology, looking at variability in time, in space, and finally simultaneous variability in both time and space. Within each of the sections following, material is generally organized by scale, beginning with phenomena at finer scales. Examples of each type of variability are given, using measurements of hydrological quantities. Since these examples are frequently drawn from New Zealand, they are particular manifestations of a more general concept. It is not practical in this brief article to survey the full gamut of hydrologic variability at every scale across the globe. The purpose of the examples is to simply provide an illustration of each concept. Although this article focuses on the occurrence and movement of water in hydrological systems, many of the same concepts are also used to describe the occurrence and movement of energy and of substances transported by water, such as sediments and nutrients. The focus here is generally qualitative, introducing the nature of hydrological variability and the concepts hydrologists use to deal with variability. **Chapter 7, Methods of Analyzing Variability, Volume 1** presents quantitative methods for analyzing the types of variability introduced in the following text.

Defining the Quantity of Interest

Hydrological variability is much more intricate than our simple air temperature and soil moisture examples would suggest. We have already noted that the variation we observe depends on space scale, time scale, and physiographic setting. In addition, to make meaningful statements about hydrological variability, we need to specify the hydrological quantity we are interested in. For example, we may be interested in a water flux such as precipitation, snowmelt, throughfall, infiltration, evaporation, recharge, or streamflow, or a store of water such as the snowpack, canopy storage, root-zone soil water storage, or water level in an aquifer, river, lake, or wetland. For example, we would expect the temporal variability of groundwater level at a location to be quite different to the temporal variability in rainfall, because of the typically damped response of groundwater to weather. The hydrological cycle for water shown in **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1** explains these fluxes and storages, and their relationship to one another. Since each of them possess a different type of variation, we will need to be specific about the flux or store when using a description of variability.

Hydrological Variability at a Range of Scales

Hydrological variability makes the hydrologist's task both interesting and challenging. The same phenomenon has to be treated differently depending on the space and timescales of the problem being considered. At the simplest level, hydrologists treat variability by measuring or estimating the variations at scales that are relevant to the problem at hand (and can be measured), neglecting other variations. This distinction is also referred to as *resolved* and *unresolved* variability. The selection of appropriate time and space scales at which to resolve variability is often a challenging task. Figure 2 provides some guidance by showing the range of time and space scales that are relevant to several hydrological processes. For a comprehensive review of scale issues, see Blöschl and Sivapalan (1995).

The spatial or temporal scale we use has a great effect on the variability we perceive: the scale can act as a filter, which lets us see some aspects of hydrology and masks out others. We can of course make very detailed observations over long time periods or over large regions, encompassing many sources of variability (e.g. hourly rainfall measurements for many years, or satellite imagery at 5 m resolution over thousands of square kilometers). However, humans do not generally comprehend all these scales at once, and typically take steps to reduce or compress the amount of information, perhaps by reducing resolution (e.g. using time series of monthly rainfall), or by reducing the extent of the data set (e.g. using only a day of 15-s rainfall data).

The Nature of Variability: Random and Deterministic

Descriptions of variability can be divided into two main types: random (happening by chance, unexplained, stochastic, probabilistic) and deterministic (caused by preceding events or natural laws, predictable, cyclic, trend, pattern).

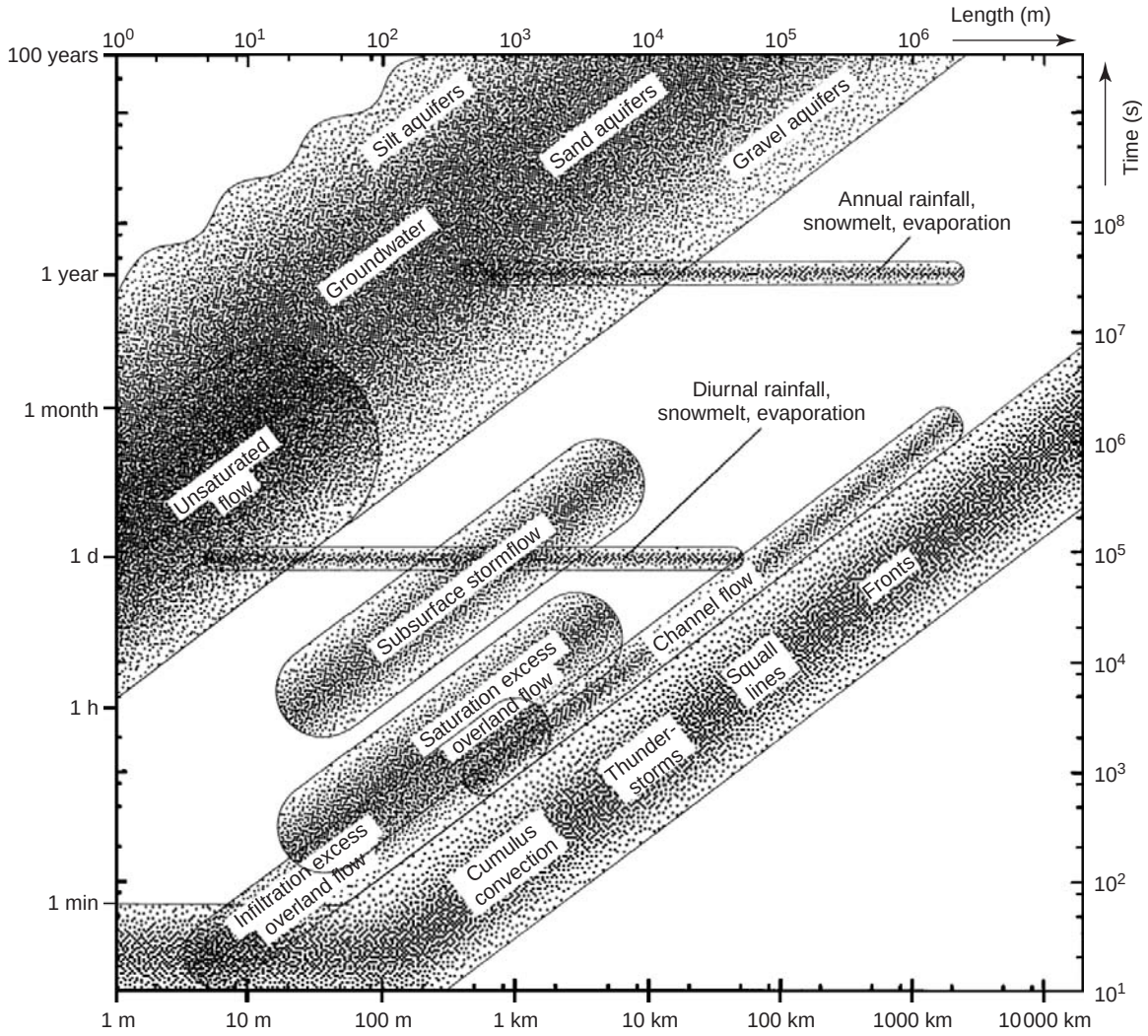


Figure 2 Schematic relationship between spatial and temporal process scales for many hydrological processes (Reproduced from Blöschl and Sivapalan (1995) by permission of John Wiley)

One extreme worldview is that all hydrological variability is deterministic, because every hydrological event has a cause that is knowable at least in principle. However, there are many situations in which the deterministic approach is impractical. There is a long history of treating hydrological quantities as random phenomena, not because they are intrinsically unpredictable, but simply because the random approach is convenient for some tasks. There are also situations in which detailed knowledge of variations and their causes is less helpful than identifying an effective descriptor, such as a statistical parameter or distribution, which captures the essential behavior of the system, without requiring a detailed enumeration of every part of the system. A well-known analogy is the use of thermodynamics to describe the net effect of many interacting molecules, and indeed several attempts have been made to construct hydrological theories using this approach.

The same physical variable can be treated as a random quantity in one context, and as deterministic in another. As an example, consider the rain falling on a small area in a severe storm. For the purpose of understanding a catastrophic flooding event caused by a particular storm, it is appropriate to use detailed measurements of how much rain fell at what time and over which locations to understand the movement of storm runoff and the subsequent flooding. This is a deterministic approach to storm rainfall. However, when designing a structure to withstand severe storms, it is often more useful to consider storm occurrence to be a random phenomena and make a statistical description of storm rainfall, assessing the probability that an event of a certain magnitude might occur. Random and deterministic approaches can also be combined, so that, for example, one might consider the total depth of rain in a storm to be a random variable, but use one or more deterministic patterns

to describe the expected temporal variation of rain intensity during any storm.

Random and Deterministic Temporal Variability

When considering temporal variability, a natural separation between random and deterministic variability is often quite clear. The cyclical movement of the Earth, each day and each year, leads to regular cycles of incoming solar radiation at the top of the Earth's atmosphere. These cycles are conveniently described deterministically, for example, using harmonic functions. These regular cycles of incoming solar radiation are reflected in hydrological variability, to a greater or lesser extent depending on which variables and locations are considered.

The daily cycle is driven by the Earth spinning on its axis, causing the Sun to (apparently) rise and set, so that incoming solar radiation increases and then decreases each day. These regular changes in energy cause daily cycles in evaporation and melting that in turn may cause daily cycles in soil moisture content, snowpack storage, water table position, and streamflow. In some settings, the daily cycle of solar radiation causes thunderstorms or strong winds at particular times of the day, which may then cause other daily cycles in hydrological variables. The two hydrological variables most conveniently described with a deterministic daily cycle are evaporation and snowmelt. At this same timescale, the occurrence and intensity of precipitation is the most obvious example of temporal variability that is conveniently described as random (not withstanding the tendency for afternoon thunderstorms in the tropics).

A similar chain of causes and effects also occurs each year, as the Earth orbits about the Sun. The Earth's axis is tilted at 23.5 degrees to the orbital plane, so in July when the North Pole points towards the Sun, the Northern Hemisphere faces the sun more directly (summer), while the Southern Hemisphere sees the Sun more obliquely (winter). Six months later, when the North Pole points away from the Sun, the seasons are reversed. The consequent seasonal changes of incoming solar radiation cause changes in the hydrological cycle, generally becoming more pronounced as we move away from the equator. There is typically less energy available for evaporation and melting in winter than in summer. During the time of year when evaporation and melt rates are less than precipitation, water may accumulate in storages (e.g. in soil, aquifers, snowpacks, wetlands or lakes) and/or run off as streamflow. As one moves into summer, the energy available for evaporation and melting may increase enough to exceed the precipitation rate so that stored water is evaporated or melted. In very wet or very dry locations where either precipitation or available energy is dominant all year round, this seasonal cycle plays a minor role in the hydrological system.

At some timescales, hydrological variables are understood mainly as random variables. We can illustrate this using both very short and very long timescales, in which our understanding of cause-and-effect is typically the weakest. At the finest space and timescales, highly intermittent turbulent atmospheric processes produce variations in rainfall. As a result, rainfall at those scales appears random.

Figure 3(a) shows 30-s rainfall accumulations that fluctuate strongly in time, with structures at many different timescales. At the other extreme, rainfall at timescales longer than a year typically shows little structure that can be represented deterministically (but see the discussion of the El Nino Southern Oscillation (ENSO) in the following text). In Figure 3(b), the year-to-year variation of annual rainfall at Warkworth in northern New Zealand provides an example of this apparently random phenomena.

Random and Deterministic Spatial Variability

Detailed spatial data for hydrology is much less comprehensive than the corresponding temporal data, mainly because of the difficulties of making spatial measurements covering areas large enough to be hydrologically significant. As a result, hydrologists' current treatment of spatial variability is still maturing. There are very few direct spatial analogues

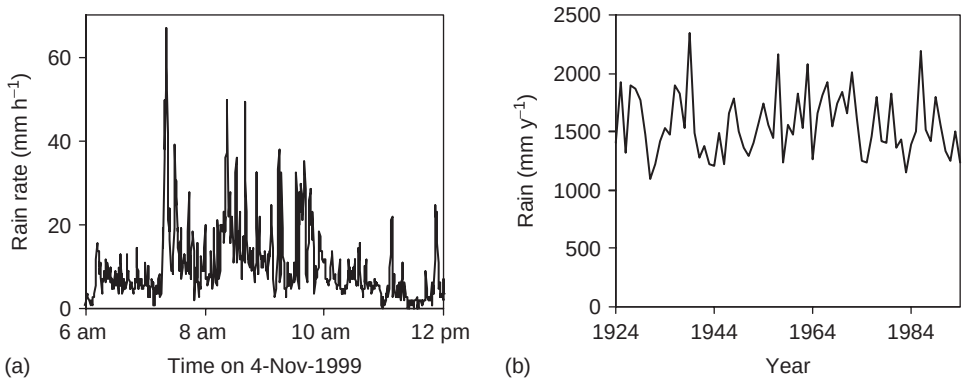


Figure 3 Apparently random variability: (a) Rainfall depths measured every 30 s during a storm over a 6 h duration; (b) annual rainfall depths at one rain gauge over a 70 year duration.

of the regular daily and annual cycles, except perhaps the wave-like landform-vegetation complexes known as *tiger-bush*, where a succession of parallel crests and troughs lead to repeated patterns of moisture content and vegetation growth (e.g. Bromley *et al.*, 1997).

Spatial trends are perhaps more common than spatial cycles – for example, the tendency for precipitation to increase and temperature to decrease, as elevation increases. Figure 4 shows one example of this correspondence in spatial patterns for accumulated rainfall in a small area of northern New Zealand. At the global scale, there is a trend for temperature to decrease as one moves to higher latitudes, as a consequence of the fact that land at high latitudes only faces the Sun obliquely, and thus receives less solar radiation per unit land area.

Much more common than either trends or cycles are irregular formations, sometimes with a repeated spatial structure (e.g. hillslope–floodplain–stream bed) at a variety of spatial scales. An irregular spatial structure is characteristic of almost every hydrological variable. Such systems show a level of organization that is intermediate between deterministic and random – scientifically intriguing, yet elusive. Since the influential work of Mandelbrot (1982), substantial research effort has been directed into recognizing and quantifying patterns in observed hydrology, especially using concepts based on this idea of self-similarity. Although this methodology has been used to describe the scaling of many types of natural variability (e.g. topography, rainfall, river networks), further progress is needed to integrate these patterns into a coherent theory of hydrology. The ideas of self-similarity do allow a concise description of how variability changes from one scale to another, but at present there is little insight into why these relationships exist, or how they might be used to improve our knowledge of the occurrence and movement of water within hydrological systems.

Some spatial patterns in hydrology have a well-understood cause, such as the soil moisture patterns caused by a well-observed storm, or the occurrence of an aquifer that has developed as a result of geological structures that are well mapped. Some current-day patterns are not the result of current conditions at all, but are a legacy of historical climate or tectonic activity. Our knowledge of these causal relationships is frequently valuable in obtaining and interpreting information on spatial variability in complex hydrological settings.

Hydrological Variability Depends on the Physical Setting

It is important to remember when reading this very general introduction, that there are many different manifestations of variability besides those illustrated here. For example, the dominant forms of variability observed in a temperate landscape are very different from those seen in regions of extreme cold or aridity. Global hydrological variability is much more intricate than is commonly presented in textbooks, and although some themes are repeated across the planet, there is a rich diversity between locations. **Chapter 121, Intersite Comparisons of Rainfall-runoff Processes, Volume 3** provides an introduction to intercomparison in the watershed setting.

For further information, the reader is encouraged to consult a reference that takes a comparative approach, such

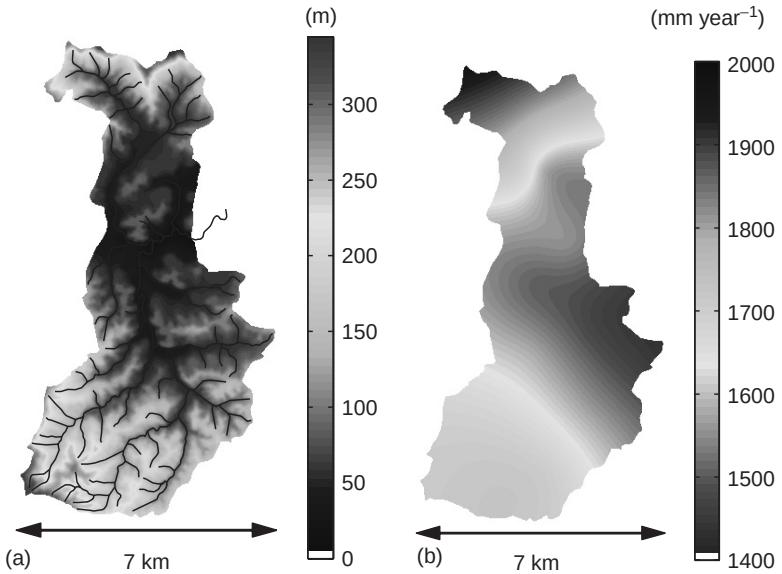


Figure 4 Spatial patterns of elevation (from 20-m elevation contours) and rainfall (interpolated from 19 months of data at 13 rain gauges). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

as Falkenmark and Chapman (1989). There, the authors propose a global hydrological classification, using maps of topography (sloping and flat land), climatic aridity (humid and dry climates), and potential evaporation (cold, temperate, and warm regions). They then provide illustrative examples of the hydrology for each of the classes in the classification system. That approach provides a structured introduction to the astonishing variety of hydrological systems present on our planet, but is beyond the scope of this article.

TEMPORAL VARIABILITY

Weather and climate is the most common direct cause of observed temporal variation in hydrological phenomena such as soil moisture, snow cover, river flow, or groundwater levels. Hydrological systems will modulate the meteorological variations, damping some of them and amplifying others, but the ultimate cause of temporal variability can usually be traced back to meteorological processes that control precipitation and evaporation. The other common source of temporal variability is water or land management by humans. Regardless of the source, we almost always have temporal variability at a wide range of timescales, both the relatively regular diurnal and annual fluctuations, as well as irregular fluctuations.

Diurnal Cycle

Figure 5 shows three examples of diurnal cycles in hydrology. Soil moisture content measured near Warkworth in subtropical northern New Zealand increases during the morning as incoming radiation increases, typically reaching a minimum level shortly after midday each day. The details of the timing at any particular site will depend on the exposure of the site to radiation, and the response characteristics of the soil and vegetation. The groundwater level on the Heretaunga Plains of the temperate eastern North Island of New Zealand reaches its maximum each morning around 7 A.M. local time, and the minimum level occurs about 12 h later. The diurnal fluctuations at this site are typically 0.2 m: the timing and amplitude of the diurnal cycle can vary greatly from one location to another, depending on factors such as how close the groundwater level is to the ground surface and the moisture status of the soil just above the groundwater table. If groundwater is being

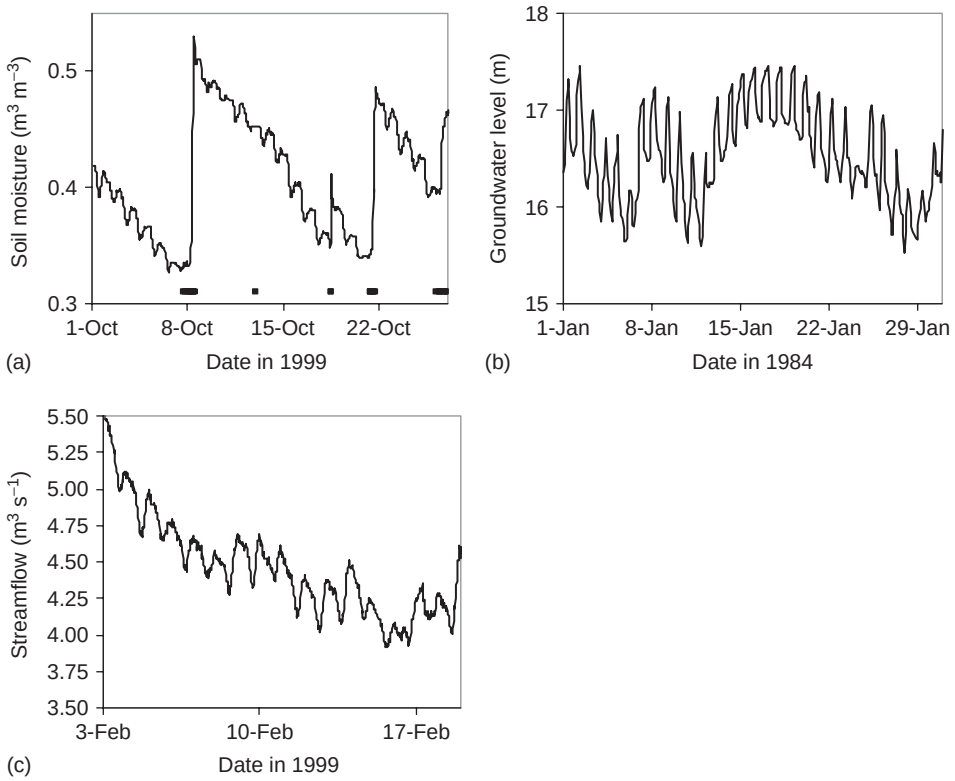


Figure 5 Examples of hydrological systems with strong diurnal cycles: (a) variations in soil moisture near Warkworth, northern New Zealand, caused by diurnal cycle in evaporation (horizontal bars indicate periods of rainfall); (b) variations in shallow groundwater level in Hawke's Bay, eastern North Island of New Zealand, caused by diurnal cycle in evaporation; (c) variations in streamflow on Jollie River at Mt Cook Station, central South Island of New Zealand, caused by diurnal cycle in snowmelt

pumped, then this can change the diurnal cycle significantly. River flow in the Maryburn River in the subalpine central South Island of New Zealand reaches a maximum in mid-afternoon when the peak of the snowmelt from its 52-km^2 catchment reaches the river flow-recording site. The cycle is most pronounced on rainless days in early summer when the seasonal snowpack is melting.

Storm Event

When precipitation falls as a discrete storm, separate from other storms, this is known as an *event*. There are many ways to define storms, so the timescale for events is not precisely defined, but typically ranges from a few minutes to a few days. At even finer timescales of a few seconds, there are significant effects of kinetic energy of raindrops in causing surface erosion. However, although the process might be understood at this timescale, it is treated in practice by estimating a time-averaged input of kinetic energy. The time-varying response of a hydrological system varies tremendously depending on the time pattern of the rainfall, the physical attributes of the land (including its management), the stored water at the beginning of the storm, and the spatial distribution of the rainfall. Figure 6 shows an example from the Mahurangi River at Sheepworld, a 2.6-km^2 catchment near Warkworth, northern New Zealand. For the purposes of this article, the message to be taken from Figure 6 is merely that streamflow responds nonlinearly to rainfall, that is, there is not a direct proportionality between rainfall and streamflow. Thus, temporal variability in rainfall is transformed or filtered by catchments. Understanding the transformation from rainfall to streamflow has occupied thousands of hydrologists for decades, because of its practical engineering significance and its apparent simplicity. Articles **Chapter 111, Rainfall Excess Overland Flow, Volume 3** and **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3** explore this topic in detail.

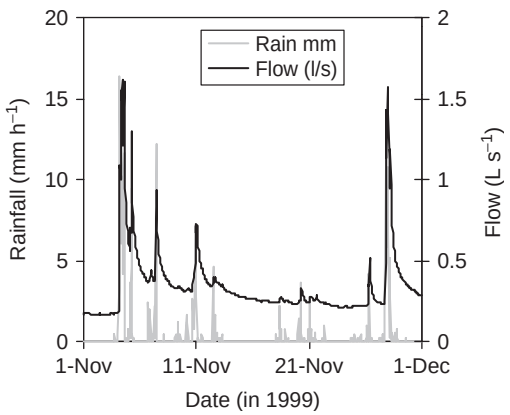


Figure 6 Event-scale variability in hourly rainfall and streamflow from the Mahurangi River at Sheepworld. The response to the small rainfalls over 18–21 November is disproportionately low

Extremes

An important and dramatic practical expression of temporal variability is the occurrence of extreme hydrological events – most obviously floods, severe snowstorms, and very high lake and groundwater levels, and also droughts and very low levels in lakes, rivers, and groundwater. The temporal distribution of floods depends greatly on the physical setting, with some regions having floods that occur at consistent times of the year because of the seasonal weather cycle and its interaction with the landscape and with water resource management, while others can have floods at any time of year, or in fact have years without any significant floods, or even any streamflow at all. The year-to-year variation in flood magnitude is very often treated using statistical techniques to assist in extrapolation from flood records of a few decades' duration up to events that might only be expected once a century, especially as part of the design of engineering structures. **Chapter 125, Rainfall-runoff Modeling for Flood Frequency Estimation, Volume 3** provides a detailed treatment of this subject.

Low river flows typically show less interannual variability than floods, mainly because the distributions of river flows are typically skewed towards lower values, and measured river flows cannot be lower than zero.

In the plots of flow data shown in Figure 7, two rather different examples of interannual variability are shown. The Grey River at Dobson, whose catchment lies in a very high rainfall area, has little year-to-year variation in flood size. In contrast, the much drier Hakataramea River only produces significant floods occasionally. One simple way to quantify interannual variability is to consider the largest or smallest value from every year of record, and then calculate the coefficient of variation (CV, i.e. standard deviation/mean) of this set of observations. Small values of CV (less than one) indicate that the years are similar to one another. The CV values for annual floods are shown in Table 1: for each river, the CV for low flows is less than the CV for floods.

More apparent in Figure 8 is the difference in the range of flows for the two sites. The range of flows in the Grey, from lowest to highest, is less than a factor of 100, whereas the range in the Hakataramea is more than a factor of 1000. Methods for characterizing and comparing the distributions of hydrological variables such as river flows are given in **Chapter 7, Methods of Analyzing Variability, Volume 1**.

Annual Cycle and Seasonality

The seasonal cycle is perhaps the most ubiquitous in hydrology: it is evident in many (but not all) hydrological systems. The causes and effects of seasonal cycles vary considerably

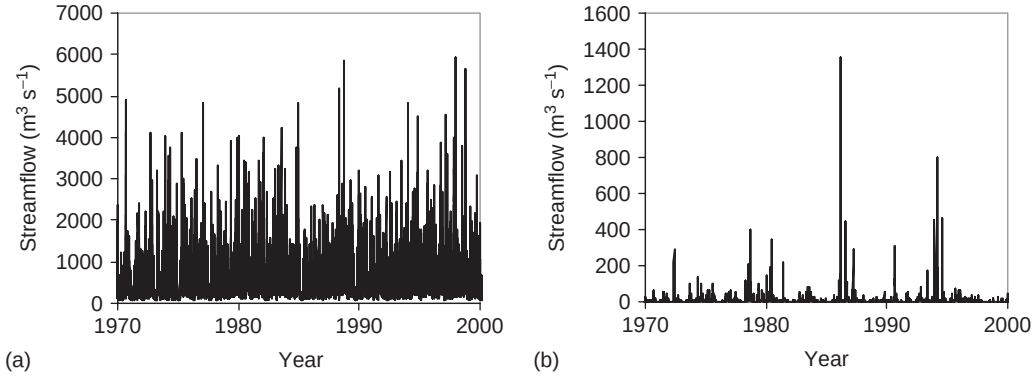


Figure 7 Illustrations of interannual variability in flood peaks: (a) floods vary little from year to year on the Grey River at Dobson (3830 km^2), on the west coast of New Zealand's South Island, where rainfall is frequent and substantial (catchment average annual rainfall exceeds 3500 mm y^{-1}); (b) substantial interannual variability in flood peaks on the Hakataramea River above Main Highway Bridge (899 km^2), on the much drier east coast of New Zealand's South Island (average annual rainfall is ca. 700 mm y^{-1}) (Data from both sites are recorded every 15 min, and the plots show daily maxima)

Table 1 Coefficients of variation for floods and one-day low flows in two contrasting rivers in New Zealand

River measurement site	CV of annual maximum floods	CV of annual minimum one-day flows
Grey River at Dobson	0.24	0.16
Hakataramea River above Main Highway Bridge	1.40	0.48

from region to region, and also across different elements of the hydrological cycle. For example, in regions where temperature is both above and below freezing, seasonal hydrological responses are driven by the accumulation of snow in the cold season and its release as meltwater in the warm season (Figure 9a). In more temperate regions without snow, where rainfall exceeds evaporation in winter, but the reverse is true in summer, water typically accumulates in soils and aquifers over winter (Figure 9b, c), possibly to the extent that water flows into rivers under gravity drainage (Figure 9d). In summer, the water stored in soils and groundwater reduces as plants draw on it to transpire, and gravity drainage gradually exhausts all the pore water it is able to influence. In both these settings, the observed temporal pattern of storage is as much the same (high in winter, low in summer), but the mechanism is different. Of course, there are many regions where neither or both of these descriptions may apply, and it is outside the scope of this article to provide a detailed classification and description of hydrological regimes.

Seasonal hydrological cycles are of tremendous biological, economic, and cultural significance to almost any phenomena that is affected by freshwater, and has response times of months or more. Hydrological seasonality is not only important to hydrologists! To give just a few examples, seasonal changes in soil moisture and rivers affect the growth and viability of almost every kind of terrestrial plant

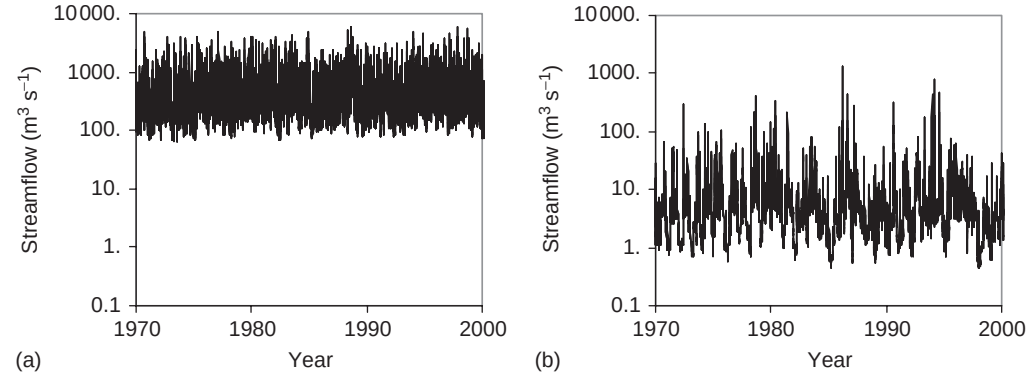


Figure 8 Interannual variability in low flows for the same two rivers shown in Figure 7: these two rivers have rather similar interannual variability in low flows (Data from both sites are recorded every 15 min, and the plots show both daily minima and maxima each day)

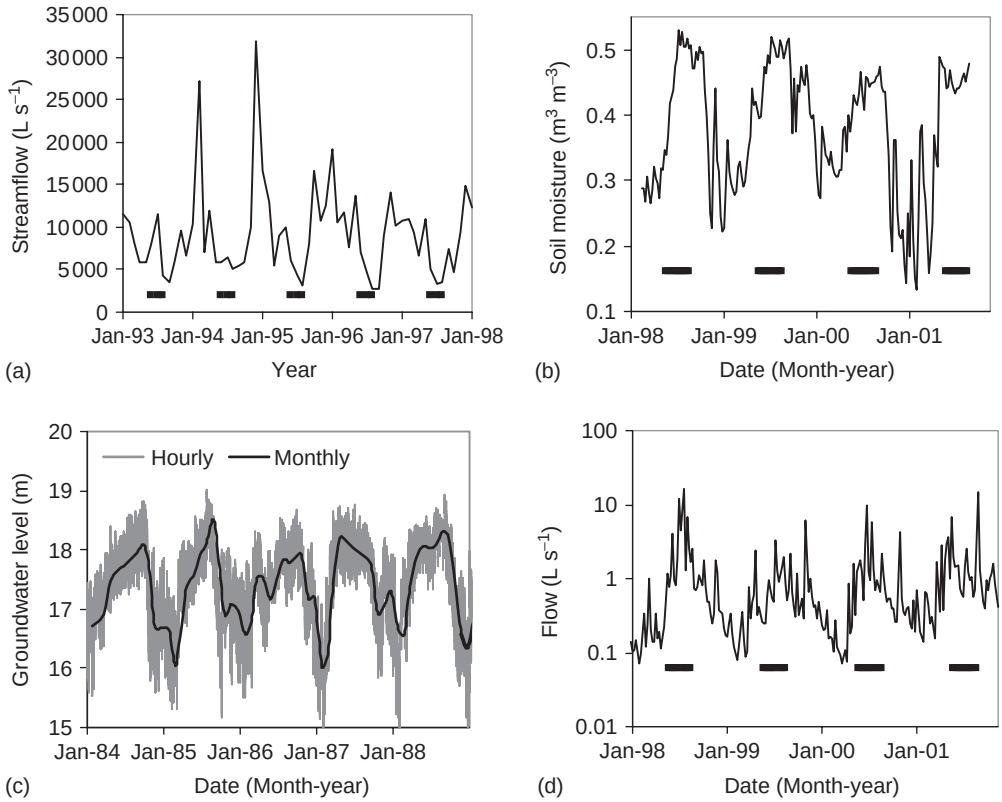


Figure 9 Examples of hydrological systems with annual cycles (horizontal bars indicate winter): (a) streamflow (monthly averages) in the predominantly snow-fed Jollie River, central South Island, New Zealand; (b) soil moisture (weekly averages) near Warkworth; (c) groundwater level (hourly data and monthly averages) on the Heretaunga Plains in Hawke's Bay, eastern North Island, New Zealand; (d) streamflow (weekly averages), for the Mahurangi River at College (46 km^2)

and creature. The availability of water in the warm growing season affects the economic viability of many kinds of terrestrial food production. Finally, many human activities and cultural constructs are centered on the seasons, for example, numerous forms of recreation that use freshwater and snow. It is not hard to see that any changes in the seasonal structure of hydrology would have very far-reaching effects, right across the planet.

Interannual Variability

At timescales longer than a year, detailed understanding of hydrological variability tends to be limited to those systems with long characteristic response times, such as deep groundwater systems and large lakes, which respond relatively gradually to changes in external forcing. In those systems, the impact of a particular event (e.g. a large input of water from a flood) may take years to propagate through the system, and so that system's response may be well understood. In more rapidly responding hydrological systems, our understanding is crucially dependent on understanding the behavior of the external systems such as climate. At timescales longer than a year, atmosphere–ocean phenomena, such as ENSO, PDO (Pacific Decadal Oscillation), and others, are understood to provide significant controls on interannual and interdecadal variations in rainfall, temperature, and solar radiation, all of which can significantly influence hydrology. Statistical analyses have shown that important linkages do exist between these ocean–atmosphere phenomena and local weather (e.g. Salinger and Mullan, 1999).

Finally, at multidecadal and century timescales, climate change can completely alter the hydrological cycle. For example, during previous ice ages, large portions of the earth's surface were permanently covered in ice. Over the last hundred years, a gradual increase in temperatures has occurred, and the most likely scenario is for this temperature increase to continue. This is likely to produce a long-term trend in recorded data for water temperature, snow occurrence, glacier size, and freshwater input to high-latitude oceans. In addition, but with less certainty, forecasts have been made that this global warming will lead to an accelerated hydrological cycle, with more intense rain and more evaporation (Watson *et al.*, 2001).

Hydrological Treatments of Temporal Variability

To summarize, hydrologists have a range of strategies for responding to temporal variability. We frequently resolve the variability down to days, hours, minutes, or even seconds using high-resolution recording instruments or explicit time steps in detailed simulation models. If this approach is impractical or inappropriate, we may select particular time periods of interest, and confine our study to those periods – the study of particular floods or droughts in hydrological systems is an example of this. Finally, where interest is predominantly in providing answers at longer timescales, but short timescale effects are known to be relevant, we may statistically summarize the short timescale variation and use mathematical techniques (e.g. Eagleson, 1978; Milly, 1994) to integrate over this variability.

SPATIAL VARIABILITY

Spatial phenomena in hydrology are mainly driven externally by spatial patterns in climate, soils, vegetation, topography, and geology. However at very long timescales, complex spatial organization develops which is created by the internal dynamics of the hydrological system. In surface hydrology, this self-organization appears to manifest itself in (i) consistent upslope-to-downslope structure of soils and hillslope geomorphology (e.g. Chappell and Ternan, 1992); (ii) regularities in channel network geomorphology (e.g. Rodriguez-Iturbe *et al.*, 1992, but see also Kirchner, 1993); and (iii) braided river networks (Sapozhnikov and Foufoula-Georgiou, 1997). However, our knowledge of how and why such regularities emerge remains very limited, and as a result, they (and other emergent features) have yet to find a place in a coherent theory of hydrology.

Geological variation is a dominant source of spatial variability in groundwater-dominated systems, and several distinctive approaches have been developed to conceptualize and quantify subsurface variability in hydrology (e.g. Anderson, 1997). A distinction between deterministic and random approaches is once again useful. The random (or stochastic) approach has been used to quantify a continuous model of heterogeneity, while more deterministic methods have typically been used when a discrete model of geological facies is employed.

For both surface and subsurface hydrology, the extremely complex nature of the problem and the difficulties in making comprehensive measurements have limited the amount of progress made, although in both cases very sophisticated approaches have been developed. Frequently, the level of sophistication in the concepts has outstripped our ability to collect data that might allow us to choose amongst the competing ideas.

In the light of the more complex, less well-understood situation with spatial variability, here some characteristic spatial patterns in a number of the driving variables known to influence spatial variability in hydrology are outlined. However, since it is not clear how these factors combine, we are unable to provide a clear synthesis at this stage. We conclude this section by summarizing a number of approaches currently taken by hydrologists to address the resulting challenges.

Climate

Climate and weather variability in space occurs in all three spatial dimensions, right down to the scale of a raindrop. Vertical variability is a relatively well-understood matter for meteorologists, but is generally neglected in hydrology, where the emphasis is generally on exchanges of water and energy at the ground surface. At the very fine scale of a raindrop, instantaneous rainfall is patchy and evaporative forcing is turbulent. However, if one considers average fluxes over timescales of minutes or more, then the corresponding spatial scales are correspondingly larger.

Within a rainstorm, there are numerous nested evolving structures, typically within any characteristic scale between the smallest turbulent eddy and the extent of the entire storm. Figure 10 shows several snapshots of a rainstorm at scales from hundreds of kilometers down to a few hundred meters. Structure and complexity is visible at every scale. This intricate system is generally either conceptualized as a continuous hierarchy of scales, linked by some form of self-similarity, or as a discrete hierarchy of macro-, meso-, and microscales.

In locations where the vertical relief provided by hills and mountains is significant, an additional feature is typically found – rainfall tends to be greater at higher elevations. More specifically, rain is typically greater on the upwind side of mountain ranges, where orographic uplift causes rapid cooling of the air. Such cooled air has less capacity to hold moisture, and excess water is precipitated as rain or snow. On the downwind side of the range, the air is able to warm again, and rainfall may be lower, because of the increased moisture capacity of the air. In cases of extremely steep mountains, the downwind rainfall may also be lower if much of the moisture in the air column is rained out on the upwind side.

At the very largest spatial scales, one finds a climate regime – that is, a region with a characteristic set of seasonal weather patterns. Thus, one has, for example, a Mediterranean climate, with cool wet winters and warm dry summers, in contrast to the humid tropical climates found near the equator. The Köppen climate classification (Trewartha and Horn, 1980) provides a formal list of climate types with operational definitions that have been mapped for the globe.

The pragmatic response of hydrology to within-storm spatial variability is to make a very limited attempt to resolve the within-storm spatial patterns. Only recently

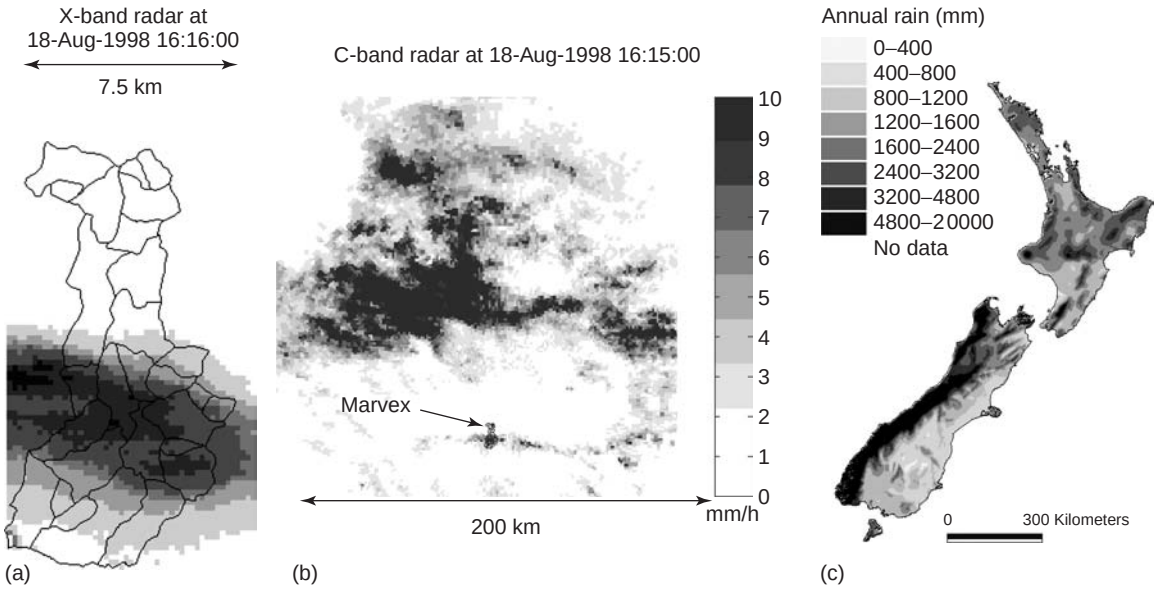


Figure 10 Spatial rainfall patterns for three space-time combinations (a) hourly rain over a 15-km radius near Warkworth, northern New Zealand; (b) hourly rain over a 200-km radius near Auckland, New Zealand; (c) estimated average annual rainfall over New Zealand

have weather radar and numerical modeling technology begun to show promise of reliable quantitative estimates of rainfall. The lack of spatial knowledge has greatly restricted progress in understanding surface hydrology. However, it has had little effect on groundwater hydrology, which is relatively unresponsive at the timescale of a storm. Spatial patterns driven by orographic effects are better understood, but spatial gradients are substantial and often poorly measured. Climate regions are generally well resolved in space and time – over small study areas, the existence of a single region is generally taken for granted.

Soils

Soils develop as a result of interactions between geology, climate, topography, vegetation, and biological processes; this happens over periods that are long in comparison to many hydrological timescales. In geologically young landscapes, the evolution of soils is actively linked to the hydrological cycle. As a result of the numerous factors affecting soil formation and development, soils show complex spatial patterns.

The soil pedon, at a scale from a millimeter up to no bigger than a fist, provides discrete three-dimensional objects around which water may flow preferentially; such flow is a significant cause of complex water and solute movement. The study of macropore flow is an active area of research; suitable research concepts are beginning to emerge, but much remains to be done in order to provide a quantitative assessment of the effect of macropore flow. Water flow within a pedon, known as *matrix flow*, is a well-advanced topic of soil physics, with a relatively mature understanding of the dominant processes. However, the application of this knowledge at larger scales is problematic, because pure matrix flow usually occurs only over short spatial scales.

As one moves downward from the ground surface, soils typically show sharp vertical contrasts in their physical properties, including porosity and permeability. These layers of soil are known as *horizons*, and they play a major role in surface water hydrology, through their ability to limit vertical flow of water and redirect it downslope. Once we descend more than a few meters below the surface, horizons tend to be thicker, but the material is now geological, not soil. We discuss geology in the next section.

In the presence of sloping land, soils typically show a topographically related structure, with shallower, lighter soils on the upper slopes of hills, and deeper, heavier soils at the base of the slope, where soils are generally wetter. This spatial arrangement is known as a *soil catena*. The direct observation of soil properties using core or point-based measurements is rarely practical as a method of obtaining highly resolved spatial soils data. Within limited spatial extents, the above spatial associations between soil and topographic position provide a potential method for mapping spatial patterns of soils. Remote sensing techniques, for example, microwave, ground-penetrating radar, and electromagnetic induction or resistivity, do provide alternative methods for obtaining spatial soils information, though not always in traditional formats.

Except in experimental settings, hydrologists tend to use rather generalized soil maps developed by soil scientists. Soil maps typically define a type of soil within a defined

region. However, since the hydrologist is typically concerned with water (and solute) storage and movement, recourse is needed to a table of physical properties for each type of soil, known as a *pedo-transfer function*. This rather generalized approach has tended to limit progress.

Geology

Geology has all the complexity of soils, but over a much deeper and less accessible region, well below the earth's surface. Anderson (1997) has recently reviewed the geological heterogeneity of sedimentary systems and provides two complementary views of heterogeneity: continuous and discrete. In the continuous view, statistical descriptions are used to describe random fields of hydrogeological properties such as permeability and porosity. The discrete view, on the other hand, emphasizes recognizable structures, known as *facies*, which are effectively homogeneous. In either case, the essential feature is that some regions permit more rapid flow and transport than others, and that the relative locations and extents of these regions are important. *See* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1** for more details on this topic.

Mathematically sophisticated models of flow and transport have been developed on the basis of statistical models of geology. These models enable predictions of heterogeneous flow fields. However, the challenges of obtaining enough geological data to parameterize these models still make the task a challenging one.

Vegetation

Vegetation interacts strongly with the hydrologic cycle, both in influencing water movement and being affected by water availability. In this respect, it is rather like soil. However, the timescales for vegetation are much shorter in some cases, given that the lifetime of a plant may be as short as a few weeks. On the other hand, plant communities can continue at the same location for centuries, if suitable conditions persist.

At the finest scale of an individual plant, the spatial pattern of plant roots (and water uptake) or the presence or absence of a plant's canopy over the ground surface (controlling throughfall) can cause spatial patterns of moisture content at scales of centimeters, below and in between plants. At scales of tens of meters in landscapes with sloping land, one may find the same arrangement as found in a soil catena, with different types and amounts of vegetation at the base of the slope, because of the relative abundance of soil moisture at the bottom of the slope. In addition, differences in the topographic aspect of a hillside may cause differences in solar radiation, and thus in evaporation.

At scales from a few meters to many kilometers, plants may form communities with particular mixtures of species that form a community. The plant canopy will typically show some vertical structure, with one or more upper stories and an understory. At larger scales than this, plants tend to reflect the climate and physiography of the region, and community composition may change in response to long-term changes in climate.

Topography

The rich structure of topographic variability interacts with hydrology at scales ranging from microtopography (1–1000 mm wide depressions and rills) to hillslopes (10–1000 m long) and stream environments (0.1–1000m wide), up to channel networks (1–10 000 km long). Each environment hosts distinctive hydrological processes and spatial complexity.

Microtopographic features are generally sufficiently numerous and variable, in relation to most field studies and practical applications, that they are treated using either a representative uniform value per unit area or by statistical approaches. For example, depressions in the land surface that may provide storage of water can be approximated as an equivalent storage capacity per unit land area, while rills could be assumed to occur at regularly spaced intervals across a hillside.

Hillslopes and streams are large enough in relation to some hydrological studies that their spatial structure may be resolved in two or three dimensions down to the meter or submeter scale (e.g. from topographic survey), in order that the spatial structure of water movement along or within them is explicitly investigated. For studies at larger scales, more generalized approaches are often taken, where, for instance, a hillslope may be represented using a sloping plane or another suitable geometric model. Similarly, a generalized stream geometry may be used to represent the broadscale details, using perhaps a trapezoidal cross section and planar long section.

Similarly, the tremendously rich three-dimensional geometric detail of channel networks extending over thousands of kilometers can be represented either explicitly (albeit only down to the resolution of the available data) or in summary form, or as a statistical summary of the distances (or travel times) which water moving along the channels must cover in order to traverse the network.

Figure 11 shows the predominantly hilly topography of New Zealand at a range of spatial scales, from a few hectares to thousands of square kilometers. Since surface topography is readily accessible to many measurement techniques, enormously rich data is now available in some locations. As a result, topography is a common source of spatially detailed data in hydrological studies, often without equivalent detail in other hydrologically relevant fields, for example, precipitation, whose spatial patterns are often very important but difficult to measure.

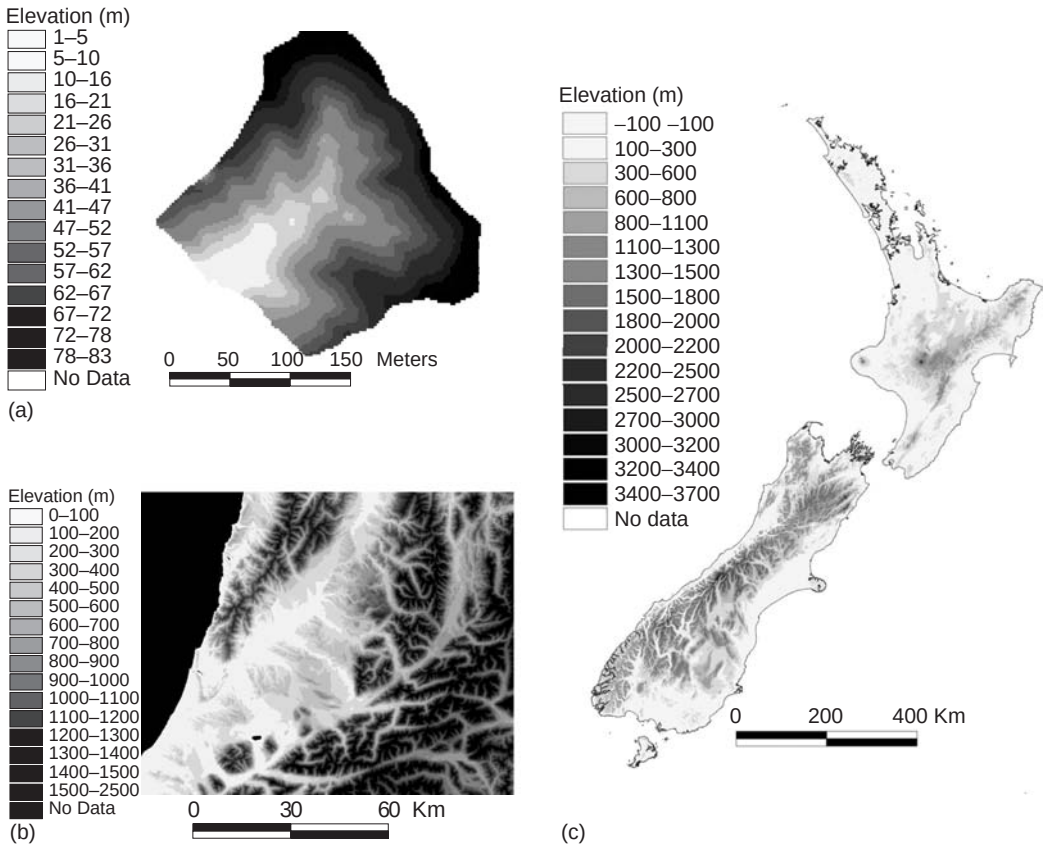


Figure 11 Elevation maps at three spatial scales (a) hillslope topography at Maimai (2 m grid); (b) Grey River catchment, New Zealand (100 m grid); (c) New Zealand (100 m grid). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Catchments and Aquifers

Catchments and aquifers are the dominant human-scale hydrological systems in which all the factors listed above can interact. With so many different sources of spatial variability all influencing water storage and movement, each with multiple spatial scales, it is perhaps no surprise that spatial variability in hydrology is only dimly understood at present.

What do we know about spatial variability in hydrological responses such as soil moisture, evaporation, groundwater, and streamflow? We know that almost all these patterns are difficult to observe in detail, and in the common situation when the observational data is limited, the patterns are difficult to interpret. Putting aside these very serious limitations for one moment, we can say that large differences from place to place in precipitation (e.g. a factor of two or more) generally lead to pronounced differences in hydrology. In fact, such strong spatial gradients of precipitation are generally sufficient to overwhelm almost any other differences, say in soils, vegetation, geology, or topography. However, in places where the spatial precipitation gradients are subtler, all the other factors are potentially of great importance, and it is not yet possible to make general statements about the detailed interactions of all these factors.

Instead, we may make some statements about the effects of patterns in each of these factors, assuming no variability in any of the others. Other things being equal, one may generally find that forested areas have more evaporation, drier soils, and lower streamflows, when compared to areas with short vegetation. Similarly, areas with a high proportion of clay soils can only infiltrate water slowly; if the rain falls mainly as intense, infrequent storms, then soil moisture is only occasionally replenished. The flow of streams in these areas will fall quickly as summer comes. Steeper terrain causes water to flow towards flatter locations, where water tends to accumulate and move more slowly. Along the same lines, geological structures with thicker layers or larger pores can store more water and will respond more slowly to temporal changes in inflows. Most of these generalizations are based in interpretations of rather limited spatial data sets, often in the presence of confounding effects from other variables. For detailed examinations of spatial patterns in hydrology, see the collection of studies by Grayson and Blöschl (2000).

In response to these difficulties, a number of strategies have been evolved. The first step to managing the complexity is usually to limit the spatial extent of the study area as much as possible, and to assume simple boundary conditions so that flow enters and leaves the study area in as few ways as possible. Thus, catchment boundaries and impervious strata are both commonly used no-flow boundaries. The second step is to assume that at most only one or two of the above sources of variability are significant. Thus, one might build a groundwater model with detailed geological heterogeneity, but assuming that rainfall and recharge are effectively spatially uniform. Or, one might model a catchment using a detailed spatial representation of precipitation, but assuming that soil and vegetation characteristics are effectively uniform.

More sophisticated treatments of spatial patterns may seek to identify a characteristic spatial scale over which spatially averaged quantities are calculated. The assumption here is that the details of finer-scale heterogeneity may be neglected without losing important information. The representative elementary volume (REV) idea (Hubbert, 1956; Blöschl *et al.*, 1995) provides a clear example of this approach, with the REV being defined in order to average over the pores and grains in a porous medium, and an equivalent continuum representation for flow being assumed to apply at the scale of the REV. This concept forms the conceptual basis for much of groundwater hydrology. However, it relies on the assumption that variability at scales larger than the REV will be resolved – in practice, this is not always simple because of the difficulty of collecting sufficient data. However, the continuum approach remains a generally accepted paradigm for groundwater.

Inspired by the REV, an entirely analogous concept, the representative elementary area (REA), has been suggested for catchment hydrology (Wood *et al.*, 1988). It was suggested that the REA might provide an averaging scale suitable for a continuum representation of hydrology, and this idea has been further extended by Reggiani *et al.* (1998), to a representative elementary watershed. However, considerable challenges still remain in (i) identifying an area that is truly representative, rather than simply being an arbitrary averaging area and (ii) finding physically meaningful ways to parameterize the effects of subarea heterogeneity: this latter task will remain extremely challenging until characteristic forms are determined for the heterogeneity.

Following several reviews of REA ideas, some attempts to generalize from the REA concept have included the Dominant Length Scales idea of Seyfried and Wilcox (1995), the Dominant Processes Concept of Grayson and Blöschl (2000), and dominant sources of space-time variability by Woods and Sivapalan (1999). However, at this stage, it is fair to say that our concepts of spatial pattern in surface hydrology remain incomplete, being either too qualitative or too limited in applicability to form the basis of a widely adopted approach.

SPACE-TIME VARIABILITY

The ultimate challenge in the context of this article is to combine temporal and spatial variability at a wide range of scales. If a hydrological system is highly damped, then it may be reasonable to view the system as a succession of steady states. However, in most cases, space variability is intimately linked with temporal evolution. At this early stage of understanding, most of our knowledge on space-time variability in hydrology is preliminary based on a small number of hydrological datasets and on simulation models that are not validated in detail. The concepts are perhaps easiest to grasp if one thinks of a time sequence of images, each showing a bird's-eye view of a catchment or aquifer at a moment in time. What is the nature of the movie formed from these images? What do the images look like? How does this change if we look closely at part of the image? What controls the way the images change over time? Do they change smoothly or abruptly in time?

Aquifers are generally dissipative systems, using their storage capacity to filter out high-frequency variability in both space and time. However, heterogeneity, such as localized areas of fracturing or high transmissivity, will cause localized changes in flow, distorting the smoothness that might otherwise predominate.

Similar concepts apply for catchments as well, but across a wider range of timescales. At the very short timescales of storms in river catchments, the complex time-space patterns of weather are concentrated in space by catchments (via downslope flow in hillsides towards valley bottoms and via flow along river networks towards larger streams), and also filtered in time (temporal oscillations are typically damped by the variety of travel times present in a catchment) (Sivapalan *et al.*, 2001). To take a specific example, imagine a catchment with travel times in the river network that range from 3 h for the most remote parts of the catchment, down to zero for places at the catchment outlet. Water that is in the river at the catchment outlet at 3 P.M. is then a mixture of rain that fell at the outlet at 3 P.M., and rain that fell at noon in the upper reaches of the network (and every space-time combination in between). The spatial extent of the catchment, combined with the travel times through it, provides this space-time filtering. Water from different places and times are mixed together by the catchment, damping out both the spatial and temporal variability.

To understand these complex systems, we need to understand both the space-time features of the driving forces and the processes by which they are transformed into hydrological responses. At present, it is common to think of rainfall and radiation as providing the space-time drivers, while soils, vegetation, topography, and geology are viewed

as static, and provide only spatial variability. Of course at longer timescales, even these "static" environmental variables do change significantly over time, for example, through land use changes, soil development, and landscape evolution. The interactions among all these fields lead to space-time fields of hydrological variables such as evaporation, snowpack, soil moisture, streamflow, and groundwater.

Rainfall

As a space-time phenomenon, rainfall is astonishingly variable. As well as the fundamental differences between storm types that cause rain (e.g. thunderstorm vs. stratiform rainfall), rainfall has complex space and time structures right down to the raindrop. The dominant paradigm is that of a complex spatial structure that both moves in space and evolves over time. The spatial element plays the dominant role in the description and time acts as a modifying agent over which spatial patterns change. Somewhat analogous to the treatment of geological heterogeneity as either continuous or discrete, there are two conceptual approaches to the spatial aspect of rainfall models. Fractal and multifractal models (e.g. Seed *et al.*, 1999) use a cascade of nested structures: a simple relationship explains the connections between spatial scales purely as a function of the ratio of those two scales. In this model, the same essential spatial structure is repeated over and over again at a continuum of spatial scales. Discrete conceptual models are still nested spatial models, but these instead use just a few discrete concepts, such as storm, large mesoscale area, small mesoscale area, and storm cell (e.g. Sivapalan and Wood, 1987).

In general, such models typically have parameters that must be estimated from observed rainfall, rather than directly evaluated from knowledge of atmospheric physics. This has limited the transferability of such models to regions without space-time rainfall data, and they remain an active area of research.

Streamflow

Observations of the detailed space-time response of streamflow, or indeed the entire hydrological cycle, are extremely rare, and as a result, our understanding of such fields is limited. Several theoretical constructs have been suggested (e.g. Woods and Sivapalan, 1999), but all such theories await comprehensive data suitable for validation. Most attempts to construct these theories concentrate on a limited range of timescales (e.g. flood only) and are relevant to a limited range of hydrological settings. Much theoretical work remains to be done in this area, but little progress can be expected without more comprehensive data. An attempt to measure multiple sources of hydrological variability at multiple scales is described in Woods *et al.* (2001), but no conclusive results have emerged at this stage. Figure 12 shows the time-evolution of spatial patterns in both rainfall and streamflow, showing some indication of the spatial damping role played by catchments, and also some unexplained variability.

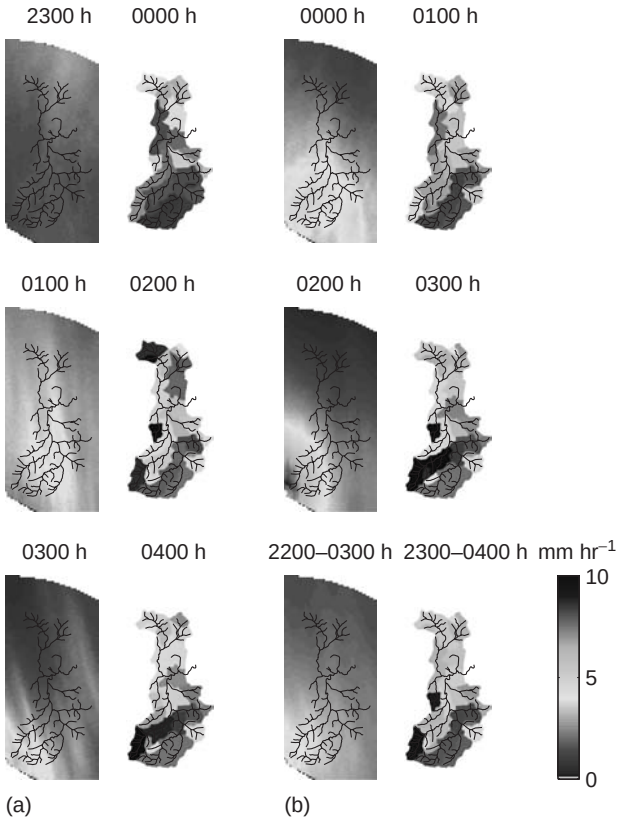


Figure 12 A time sequence of spatial patterns of rainfall and streamflow in the Mahurangi catchment in northern New Zealand. Rainfall (a) and flow (b) are shown in pairs, with rainfall for the previous hour associated with each hour of flow, since the catchments have a response time of the order one hour. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Soil Moisture

Detailed and spatially extensive space-time observations of soil moisture have permitted significant progress in understanding its variability. Grayson and Western (1998) have identified a consistent spatial structure over time in small experimental areas, and proposed the observational concept of catchment average soil moisture monitoring (CASSM). The idea is that some locations in the landscape are consistently good indicator sites to represent spatially averaged soil moisture. Such an approach provides spatial information through the assumption that the site is representative, and provides the temporal information by monitoring continuously at just the CASSM site. Additional observations have suggested that spatial patterns themselves may show a

switching behavior over time in some settings, from a predominantly topographically controlled pattern when soils are relatively wet to a predominantly soil-controlled pattern during drier periods (Grayson *et al.*, 1997).

Recently, a relatively comprehensive theoretical framework has been proposed for space-time variability of soil moisture (Albertson and Montaldo, 2003), and this may go some way to interpreting the time-evolution of statistical spatial variability in soil moisture.

SUMMARY

Variability is fundamental to hydrology. Without it, we would have solved most of our great challenges decades ago. To make useful quantitative statements about variability, we need to quantify and summarize these place-to-place or time-to-time changes. Table 2 provides some examples of the approaches taken to conceptualize and quantify hydrological variability in space and time.

Early efforts to interpret variability have been confounded by the changes in variability that occur as scale changes. The recognition of spatial scale as a controller on the expression of variability has at least allowed us to understand why spatial hydrology seems so complex, but much remains to be done both in terms of measurement and theoretical constructs. There are numerous options for making progress on hydrological variability, depending on the nature of the variability.

At the coarsest level, we lack a fundamental context in which to place our knowledge of variability. For example, suppose you are somehow able to unravel the intricacies of time and space variability of some aquifer or catchment. How do we decide whether your hard-won knowledge can be applied elsewhere? Surely some of your conclusions will carry over to other similar systems elsewhere, but how do we define "similar"? The scientific hydrology community has no globally agreed way of assessing similarity, and this is a considerable impediment to the transfer of information and the assembly of meaningful international data sets for comparative studies (McDonnell and Woods, 2004). Securing agreement on this ought not to be onerous or time-consuming if we set realistic expectations for a relatively general, but widely accepted, system for determining broad similarity. Once the basic elements of such a classification method is that in place, they can be refined to suit local or specialist needs as required.

At the technical level, much is still to be gained from improving our measurement technology, in order to observe variability in hydrology (e.g. evaporation, soil moisture, mountain catchments) and in the key drivers (e.g. precipitation, geological structure). Recent advances in measurement have stimulated new conceptual approaches, and we must remain optimistic that this will remain a fruitful source of inspiration. Most difficult, but potentially rewarding, would be improvements in subsurface measurement techniques, for there we are often working almost "blind", generally with at best a peephole view of the hydrology.

It seems that many of the questions about temporal variability are well understood, at least for timescales from seconds up to years. This is in part due to the relative ease with which we can measure at high temporal resolution for years at a time, and in part due to the small number of dominant sources of temporal variability (mainly climate). At timescales of decades and longer, we are at more of a disadvantage simply because we do not have good quality observational data over sufficient time with which to investigate the questions. This problem seems unlikely to be fully resolved for several decades, unless longer historical records of hydrological response can be reconstructed.

Every step forward in understanding variability at various spatial and temporal scales is crucial to advancing the theoretical underpinnings of hydrology. We strongly suspect that some of our concepts are inappropriate or incomplete, and frequently this is because our treatment of variability is suspect. For example, the effect of macropores on hillslope water flow remains poorly developed, and as a result, our detailed models of water movement do not usually include macropores. It is possible that much of this difficulty occurs because we do not know how to represent the spatial occurrence and connectivity of these pathways. If that problem

Table 2 Examples of deterministic and random views of hydrological variability

	Deterministic	Random/statistical
Temporal	• Time series (for many variables) • Constant value (e.g. temperature of deep groundwater) • Diurnal cycle (e.g. solar radiation, evaporation, snowmelt) • Annual cycle (e.g. soil moisture, snowmelt, streamflow)	• Rain bursts • Occurrence and magnitude of storm events • Occurrence and magnitude of floods and droughts • Interannual variability in climate
Spatial	• Map (where possible, e.g. vegetation type, topographic elevation) • Constant value • Transect along hillslope catena • Transect across mountain range (orographic uplift and rain shadow)	• Climate statistics • Soil • Geology • Microtopography

were solved, it is possible that familiar dynamical equations could then be solved in some as-yet unquantified geometry. Then again, a completely new approach may be needed. In a more general sense, there is still a considerable legacy of thought that the idea of an equivalent homogeneous medium exists, which has essentially the same hydrological characteristics as the spatially heterogeneous medium of the real world. This concept has pervaded both surface and subsurface hydrology for many decades, but now seems certain to be superseded by approaches that acknowledge the crucial role of spatial heterogeneity. There are many hydrological questions, especially those connected with the transport of solutes and particulates, for which a typical medium is not especially helpful. Often, the key attribute of such a transport system is in fact not the typical path but the fastest or most effective transport path.

Given the very limited current observational capabilities for spatial hydrology, it seems almost certain that the next major advances ought to come in this area. Some of this progress seems likely to emerge from the fusion of simulation models of the hydrological cycle with remotely sensed data. Hydrologists will need to continue to actively engage with the remote sensing design process in order to ensure that the data collected not only has desirable space-timescales but is also as sensitive as possible to both water and energy.

As hydrologists gradually solve the problem of characterizing spatial and temporal variability at a wide range of scales, across a variety of hydrological settings, we will be ready to focus our attention on the dynamics. Hydrological dynamics went through a tremendous period of growth and innovation in the 1960s and 1970s, to some extent because temporal variations were sufficiently well observed and conceptualized that the time was ripe to explore the dynamics. Let us hope that we will soon be in the same position with regard to space-time variability!

Acknowledgments

I wish to thank the many organizations that provided data used in this article, including New Zealand's National Institute of Water and Atmospheric Research (NIWA), the University of Auckland, Auckland Regional Council, Hawke's Bay Regional Council, University of Melbourne, and Meridian Energy Limited. I also wish to acknowledge the advice and patience of my editors, Günter Blöschl and Murugesu Sivapalan.

REFERENCES

Albertson J.D. and Montaldo N. (2003) Temporal dynamics of soil moisture variability: 1. Theoretical basis. *Water Resources Research*, **39**, 1274.

Anderson M.P. (1997) Characterization of geological heterogeneity. In *Subsurface Flow and Transport: A Stochastic Approach*, Dagan G. and Neuman S.P. (Eds.), Cambridge University Press: pp. 23–43.

Blöschl G., Grayson R.B. and Sivapalan M. (1995) On the representative elementary area (REA) concept and its utility for distributed rainfall-runoff modelling. *Hydrological Processes*, **9**, 313–330.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling - a review. *Hydrological Processes*, **9**, 251–290.

Bromley J., Brouwer J., Barker A.P., Gaze S.R. and Valentin C. (1997) The role of surface water redistribution in an area of patterned vegetation in a semi-arid environment, south-west Niger. *Journal of Hydrology*, **198**, 1–29.

Chappell N. and Ternan L. (1992) Flow path dimensionality and hydrological modelling. *Hydrological Processes*, **6**, 327–345.

Eagleson P.S. (1978) Climate, soil and vegetation. 1. Introduction to water balance dynamics. *Water Resources Research*, **14**, 705–712.

Falkenmark M. and Chapman T. (1989) *Comparative Hydrology*, UNESCO: Paris, p. 479 Also available on the World Wide Web: `http://www.siwi.org/downloads/downgeneral.html`.

Grayson R.B. and Blöschl G. (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Cambridge University Press: New York.

Grayson R.B. and Western A.W. (1998) Towards areal estimation of soil water content from point measurements: time and space stability of mean response. *Journal of Hydrology*, **207**, 68–82.

Grayson R.B., Western A.W., Chiew F.H.S. and Blöschl G. (1997) Preferred states in spatial soil moisture patterns: local and nonlocal controls. *Water Resources Research*, **33**, 2897–2908.

Hubbert M.K. (1956) Darcy's Law and the field equations of the flow of underground fluids. *Transactions of the American Institute of Mining and Metallurgical Engineers*, **207**, 222–239.

Kirchner J.W. (1993) Statistical inevitability of Horton's laws and the apparent randomness of stream channel networks. *Geology*, **21**, 591–594.

Mandelbrot B.B. (1982) *The Fractal Geometry of Nature*, W. H. Freeman: New York.

McDonnell J.J. and Woods R.A. (2004) On the need for catchment classification. *Journal of Hydrology*, **299**, 2–3.

Milly P.C.D. (1994) Climate, interseasonal storage of soil water, and the annual water balance. *Advances in Water Resources*, **17**, 19–24.

Reggiani P., Sivapalan M. and Hassanizadeh S.M. (1998) A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. *Advances in Water Resources*, **22**, 367–398.

Rodriguez-Iturbe I., Ijjasz-Vasquez E.J., Bras R.L. and Tarboton D.G. (1992) Power law distributions of discharge mass and energy in river basins. *Water Resources Research*, **28**, 1089–1093.

Salinger M.J. and Mullan A.B. (1999) New Zealand climate: temperature and precipitation variations and their links with atmospheric circulation 1930–1994. *International Journal of Climatology*, **19**, 1049–1071.

Sapozhnikov V. and Foufoula-Georgiou E. (1997) Experimental evidence of dynamic scaling and self-organized criticality in braided rivers. *Water Resources Research*, **33**, 1983–1991.

Seed A.W., Srikanthan R. and Menabde M. (1999) A space and time model for design storm rainfall. *Journal of Geophysical Research*, **104**(D24), 31623–31630.

Seyfried M.S. and Wilcox B.P. (1995) Scale and the nature of spatial variability: field examples having implications for spatial variability. *Water Resources Research*, **31**, 173–184.

Sivapalan M., Kumar P. and Harris D. (2001) Nonlinear propagation of multi-scale dynamics through hydrologic subsystems. *Advances in Water Resources*, **24**, 935–940.

Sivapalan M. and Wood E.F. (1987) A multidimensional model of nonstationary space-time rainfall at the catchment scale. *Water Resources Research*, **22**, 1289–1299.

Trewartha G.T. and Horn L.H. (1980) *An Introduction to Climate*, McGraw-Hill: New York.

Watson R.T. and The Core Writing Team (Eds) (2001) *Climate Change 2001: Synthesis Report*, Intergovernmental Panel on Climate Change, Geneva, Switzerland. Also available on the World Wide Web: **`http://www.ipcc.ch/pub/un/syreng/spm.pdf`**.

Woods R.A., Grayson R.B., Western A.W., Duncan M.J., Wilson D.J., Young R.I., Ibbitt R.P., Henderson R.D. and McMahon T.A. (2001) Experimental design and initial results from the Mahurangi River Variability Experiment: MARVEX. In *Observations and Modelling of Land Surface Hydrological Processes*, Lakshmi V., Albertson J.D. and Schaake J. (Eds.), Water Resources Monographs, American Geophysical Union: Washington, pp. 201–213.

Woods R.A. and Sivapalan M. (1999) A synthesis of space-time variability in storm response: rainfall, runoff generation and routing. *Water Resources Research*, **35**, 2469–2485.

Wood E.F., Sivapalan M., Beven K.J. and Band L.E. (1988) Effects of spatial variability and scale with implications to hydrologic modeling. *Journal of Hydrology*, **102**, 29–47.

4: Organization and Process

MIKE KIRKBY

School of Geography, University of Leeds, Leeds, UK

Water, its movement, changes of state, and chemistry, is fundamental to the nature of planet Earth. It is involved in the most important processes near the surface and supports life as we know it. Hydrological sciences potentially address all aspects of water movement and transformation, although with major divisions between hydrology, which is essentially focused on the principle of conservation of mass; hydraulics, based on conservation of momentum (Newton's laws of motion); and meteorology, based on the laws of thermodynamics. This review of processes takes a primarily hydrological viewpoint, concerned with where water is, its pathways, and impacts on other aspects of the environment. A good understanding of these processes is an essential prerequisite to modeling and forecasting the behavior of water, and much of the history of hydrological science reflects changes in our conceptual understanding of how and where water moves. The main theme of this article is the way in which the landscape influences the flow of water, particularly surface hydrology, and how the hydrology in turn influences the form of the landscape.

Introduction – The Hydrological Cycle

Water has a number of properties that give it an overwhelming influence on the environment and life of planet Earth. The equilibrium temperature of the Earth's surface at about 15 °C is a balance between incoming solar radiation and outgoing long-wave radiation at the ambient temperature, with important secondary corrections for atmospheric and albedo effects (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**). Water is able to change in state between solid, liquid, and vapor at a range of temperatures that are close to this equilibrium, and so plays a large part in providing a global thermostat, with substantial areas of ice, open water, and atmospheric vapor that interact to control regional temperatures and energy flows between low and high latitudes.

On geological timescales, water is exchanged with the earth's crust, principally through burial in sediment, subduction along tectonic plate boundaries, and release in volcanic eruptions, leading to the large volumes of water in the oceans today. In the global circulation of water (Figure 1) over shorter time spans, a few large reservoirs dominate the storage of water and, through their slow turnover, act to moderate the circulation processes and the weather (UNESCO, 2000). The smaller stores, and those most intimately associated with surface water hydrology, are the atmosphere, rivers, and soil moisture. With high fluxes and short residence times, these stores are characterized by high variability in time and space, giving surface hydrology its dynamic character and large regional contrasts.

Catchments have residence times that increase with their catchment areas. For small catchments, residence time is limited by the response time for soil moisture that stores and releases precipitation, whereas the response time for larger catchments ($>>100\,\mathrm{km}^2$) is determined more by travel time through the channel network. Because of their longer residence times, larger catchments respond more strongly to storms that both cover greater areas and have longer durations, whereas small catchments typically respond most strongly to localized intense storms.

The atmosphere ultimately draws its water from the greater evaporative potential of the oceans, and can only maintain its moisture supply over the continents by exchanging water and thermal energy with the land and vegetation through rainfall and evapotranspiration. The global distribution of hydrological regimes can be characterized by the interaction of precipitation and evapotranspiration, which is strongly related to temperature. The seasonal patterns of these variables immediately give a strong indication of the overall pattern of hydrological responses (Figure 2).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

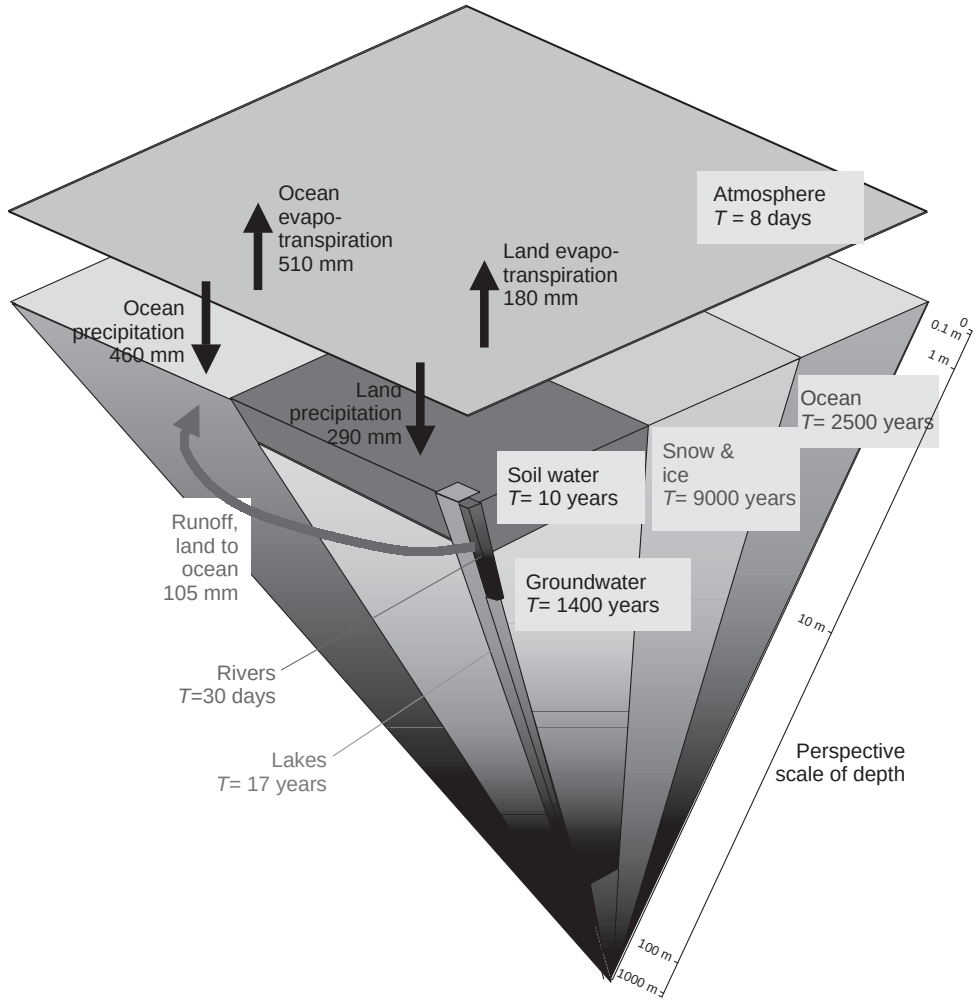


Figure 1 The global hydrological cycle on an annual scale. The area of each block indicates its areal extent, and the depth, on a perspective scale and its mean depth over its area. *T* indicates residence time in each store, and arrows show major transfers between the principal stores. Geological and other long-term exchanges are not included. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Where potential evapotranspiration exceeds precipitation, there is little soil moisture and most runoff is associated with infiltration of excess overland flow. Where potential evapotranspiration is less than precipitation, runoff is associated mainly with saturated conditions. In between these extremes, the role of groundwater is potentially greatest, with significant recharge that can then act as a dominant supplier of runoff where bedrock lithology and structure allow significant transmission of water. Finally, cold conditions lead to precipitation as snow and frozen soil, giving regimes where seasonal melting becomes the dominant hydrological process. Many areas show a significant seasonality so that the dominant processes may change throughout the year, as indicated for some of the example areas sketched in the figure. These issues are pursued in greater detail below, with their implications for and dependencies on the landscape form.

HYDROLOGICAL PATHWAYS

All of the various hydrological processes and pathways discussed here are examined completely elsewhere in the Encyclopedia. The focus here is on the interactions with geomorphological and other processes that help to determine the importance of each pathway. With only few and generally minor exceptions, the surface skin of the landscape acts as a one-way system that transports water, sediment, and solutes downhill and downstream, ultimately towards the sea. Gravity drives this one-way system and water is usually its direct or indirect agent. Its gravitational potential is constantly being renewed by the solar energy that drives evapotranspiration, and about 0.1% of this energy is used to erode the land. Only very small amounts of material are recycled in precipitation, and the continual downward movement of earth materials would

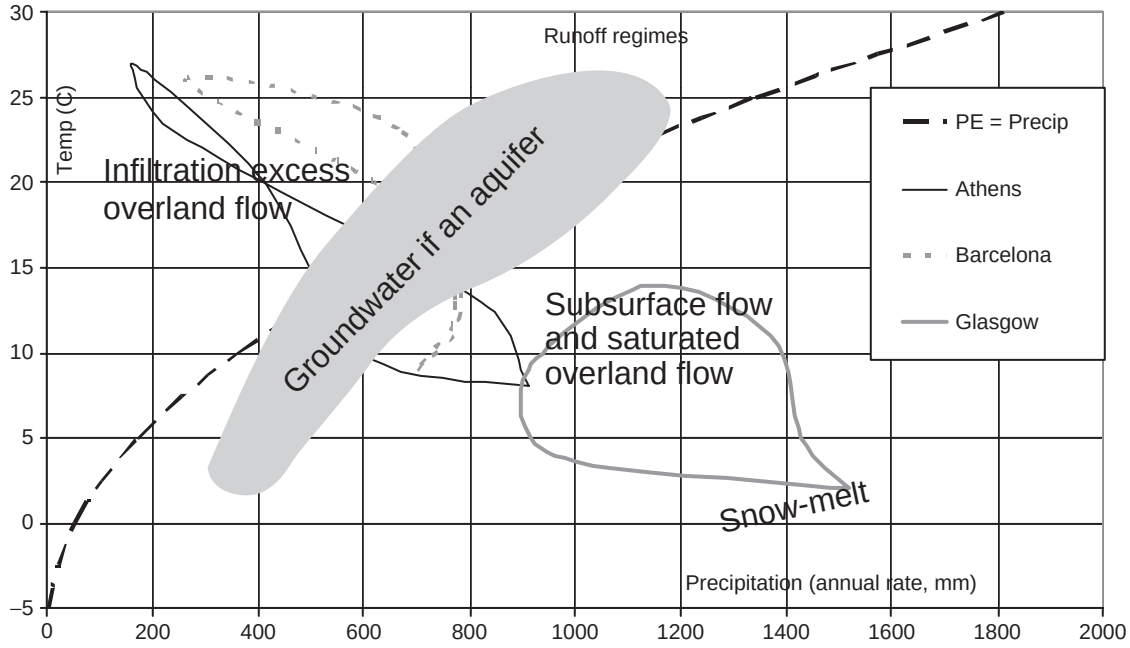


Figure 2 Hydrological regimes in terms of seasonal variations in temperature and precipitation. For the three example areas, monthly values for precipitation (P) and evapotranspiration (E), associated with temperature show the overall hydrological regime over the year. Arid conditions occur where P >> E, humid conditions where P <= E, and cold-dominated regimes where the temperature is below freezing. Groundwater is most important, in the presence of an aquifer, where P and E are similar, because there is substantial infiltration without dominance by shallow subsurface flow

therefore reduce the continents to a low-lying plain over 10–100 million years if they were not periodically uplifted through tectonic activity.

During the denudation of the land, geomorphological processes bring about a structural organization of the land surface that is broadly related to the hydrological regimes shown in Figure 2. The style of the landscape is a response to the mechanisms and potential to export water and earth materials to the oceans. In permanently cold regions, at high latitudes and at high elevations, landscapes are dominated by glaciers and ice caps that convey water as ice, both across broad ice shelves and in relatively narrow ice streams and glaciers focused on topographic depressions that convey most of the ice and the eroded sediment. In humid regions, landscapes are organized into a strong hierarchy of branching river networks, with the 12 largest rivers draining half of the land area. For both ice streams and rivers, the organization into dominant paths indicates a nonlinearity in the driving processes. This nonlinearity is much more marked for sediment transport than for rivers, and the valleys in which they flow are necessarily produced as eroded depressions in the land so that structural organization may be primarily a response to sediment flow rather than hydrology on its own (Kirkby, 1993).

Only the most arid areas show a lack of organization, with internal drainage and sediment accumulation where all precipitation can evaporate without draining to the sea, and wind (that does not respect the gradient of the land) may be the most important agent of sediment transport. Even arid areas, however, eventually develop drainage networks, partly as sediment fills depressions and creates continuously sloping pathways to the sea, and partly as a relic of previous, more humid climates, as global climates have changed and tectonic plates have moved. On the boundary between humid and arid regimes, limestone regions may mimic aridity by allowing water to percolate downwards, so that drainage need not rely on a surface channel network, although many limestone areas have both surface and subsurface drainage routes, reflecting their evolutionary history.

Here we focus on areas that are dominated by river flow, where there are generally connected dendritic channel networks that periodically or continuously carry flows of water, solute, and sediment. Although the same hydrological principles apply everywhere within this system, the balance between processes differs enough to alter the dominant processes and behaviors in different parts of the system and under different climatic regimes. Hillslopes are dominated by processes of runoff generation and the potential for areas to be connected or isolated from drainage channels. Channelways are dominated by the routing of flow, by relationships between channel and floodplain, and by issues of

sediment transport. Groundwater is also dominated by the transmission system and its evolution.

Hillslopes

Hillslopes are distinguished from channelways by the lack of continuous drainage lines, but it will be seen that this separation is somewhat arbitrary, because channel heads change in position over time during storms and over longer periods, and because there may be subsurface drainage lines that have no surface expression. A more functional definition may be in terms of surface morphology, defining hillslopes as the upstream areas without continuous surface channels, and channels by the existence of eroded banks to give a morphological feature. This definition remains difficult to apply in practice, particularly where channels are delimited without field survey, from maps, Digital Elevation Models (DEMs), or remotely sensed images. However, the functional definition can be linked to the stability of the landscape form, in the sense of linearity for sediment transport processes, and will be used here.

Figure 3 shows the main pathways for water on the hillslope (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**). Precipitation falls on the vegetation and the soil surface. Some water is intercepted on foliage and some held in puddles on the surface as depression storage. These stores, together with soil moisture, are available for evaporation, and plants also draw water up from the soil through their roots and transpire it through their leaves, transporting nutrients, supporting photosynthesis, and maintaining the rigidity of soft tissue. Water penetrates the soil surface as infiltration, both across a saturated interface in wet depressions and through the direct entry

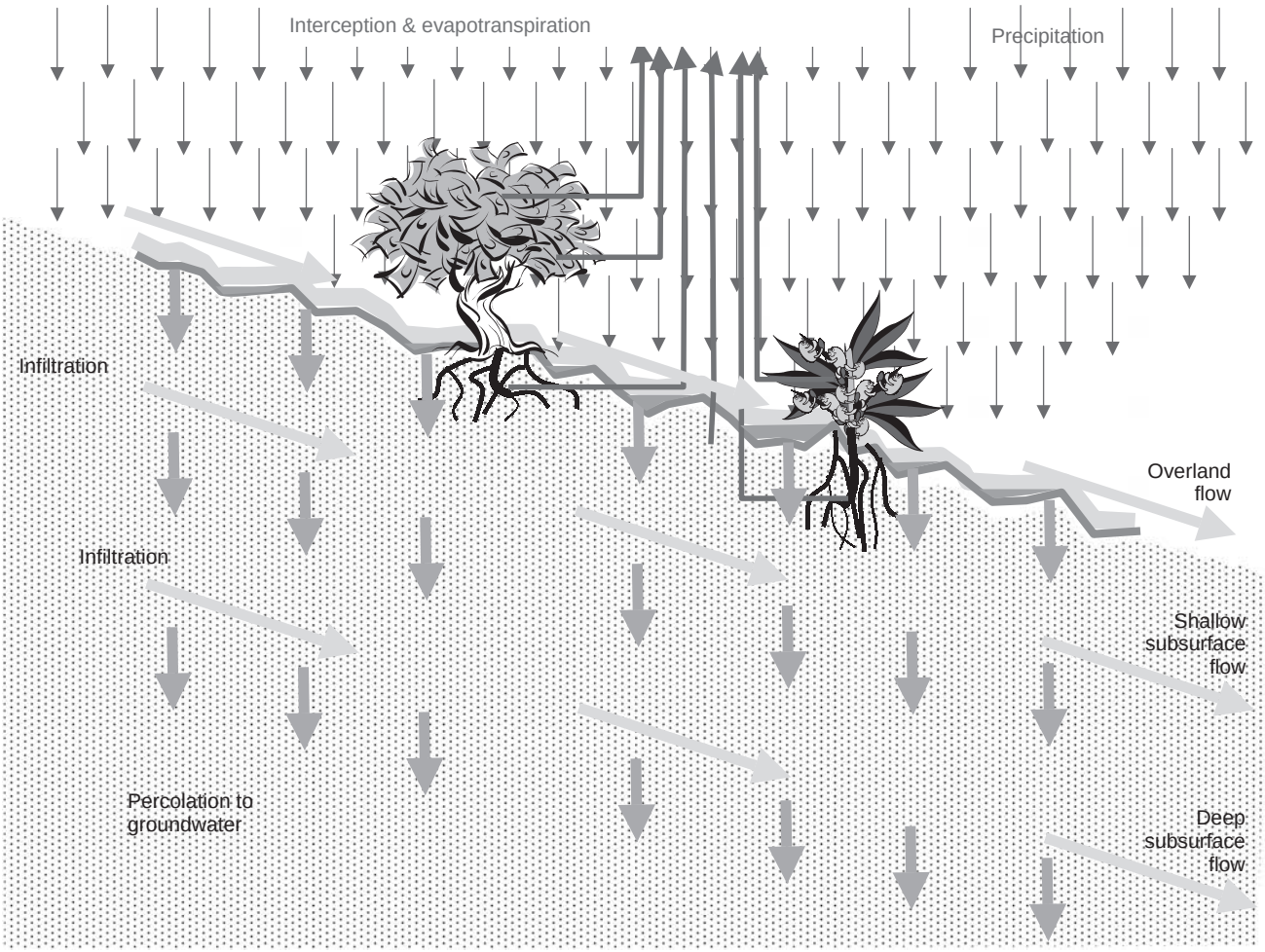


Figure 3 The principal pathways for precipitation on the hillslope. Water is most strongly diverted downslope at levels where the infiltration capacity is decreasing with depth, often associated with soil horizons, and may be extracted by evapotranspiration over the full depth of plant roots. Subsurface flow can take place both through the soil matrix and along discrete macropores. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of raindrops in unsaturated areas. Percolation continues down into the soil, which generally becomes less permeable with depth, although with important exceptions to this rule. Where permeability is maintained into the bedrock, there is scope for recharging groundwater within the rock.

At the soil surface, and progressively through the soil profile during the infiltration process, flow may be diverted laterally to increment downslope discharge and may also be lost to evaporation and root uptake. At the surface, any water diverted becomes overland flow, which has the greatest impact on storm hydrograph response because of its relatively high flow velocity. Overland flow rarely exceeds half of the rainfall and is generally less than 10%. At the surface and at all levels in the soil, water is diverted when the soil beneath becomes saturated, and this may occur either because of decreasing vertical infiltration rates or because of saturation by laterally flowing water from upslope. The balance between these processes varies significantly with soil types, but more strongly with climate, which has a consistent influence on the spatial and temporal patterns of hillslope hydrology, and its interaction with geomorphology.

Humid Areas

Humid areas are those in which precipitation exceeds potential evapotranspiration over the year as a whole so that there is generally a reservoir of soil moisture. If there is a dry season, there may be some alternation between humid and semiarid behavior through the year. The excess of rainfall over evapotranspiration generally leads to substantial infiltration. The infiltrated water may percolate down to groundwater where the parent material acts as a good aquifer, but elsewhere must contribute to lateral subsurface flow at some depth within the soil profile.

Where there is an aquifer, percolating water can flow down through open fissures, as for example, in most karstic limestones, or more slowly through matrix pores as in many sandstones, so that the time delay between rainfall and the corresponding rise in groundwater piezometric levels varies substantially with rock lithology and joint structures, from a few days to many years. In addition, alternating and/or spatially varying lithological sequences, dipping strata, and disconformities between aquifers and impermeable layers can greatly complicate the relationships in any area. Fissures in karstic limestones are enlarged by solution of the limestone along preferential flow lines, initially very slowly and then more rapidly as they grow in size, and may generate flow connections that have little relation to surface topography. Thus, although recharge rates can be estimated, detailed local studies are needed to understand the relationships between groundwater recharge and discharge sites. However, in a uniform aquifer, the uniform recharge is eventually returned to streams under low flow conditions, and the piezometric surface therefore drains towards the channels, with a form that broadly mirrors surface topography, but with much reduced relief.

Where there is no aquifer, then the soil and rock at depth is, in principle, saturated, but transmits a negligible amount of water. Thus, the subsurface moisture profile shows saturation at depth, and, on average a general decrease in the proportion of saturation upward towards the surface. However, there may be consistent differences in this profile, for example, with more frequent saturation above a relatively impermeable iron pan horizon and less frequent saturation below it. There are also differences over time, for example, when intense rainfall wets the surface, driving a wetting front down into unsaturated soil below.

Movement of water in the soil is driven by gravity and by the gradient in hydraulic potential, generally drawn from wetter to drier zones by capillary forces. A simple theory assumes that a soil has a uniform hydraulic conductivity at any moisture content, and that water advances into drier areas along a uniform front. However, soils generally have two or more populations of pores that can be distinguished as matrix and macropores. For a soil consisting of randomly arranged aggregates, the interior of each aggregate acts as a matrix, while the spaces between the aggregates behave as macropores. For a cracking soil, such as heavy clay, the mass of clay may behave as the matrix, while shrinkage cracks between them providing macropores that vary in size with the shrink-swell history. Other macropores can be initiated as animal burrows, surface channels reroofed by bank collapse, or through internal erosion (hydraulic piping) by water, and these processes may reinforce one another.

Flow in the soil matrix approximately follows the simple theory, but flow in macropores bypasses matrix flow so that the overall wetting front below a saturated surface generally consists of a shallow uniform wetting front with spikes of much greater penetration in a thin film of water following the walls of macropores, along which secondary wetting fronts spread into parts of the matrix adjacent to each macropore. In some cases, the overall flux is largely dominated either by matrix infiltration or by macropore flow, but there are many cases where both are important (Beven and Germann, 1982; Brammer and McDonnell, 1996);

Where the soil is unsaturated, the total potential gradient (gravitational plus hydraulic) is much greater in the vertical than lateral directions and flow is predominantly vertical, but where water reaches the saturated level, the potential gradient becomes greatest in a downslope direction and flow becomes predominantly lateral. This most commonly occurs close to the surface, but in some cases may give rise to more than one such perched water table within the regolith.

In a humid area with low permeability bedrock, every point down a slope profile contributes additional water

to the shallow subsurface discharge. For a uniform soil, this subsurface runoff contribution, j, will be more or less constant downslope so that the total subsurface discharge at unit area a from the divide is given by:

$$q = ja \tag{1}$$

where the unit area a is defined as the area drained per unit contour width and is the same as distance from the divide where contours are parallel.

If the saturated subsurface discharge is considered to take place in a layer with fixed hydraulic conductivity K, and the gradient of the flowing layer is s (that will be approximately the same as the surface gradient in many cases), then the depth h of flowing water over the impermeable bedrock is obtained by equating the two expressions for subsurface discharge:

$$q = ja = Khs \tag{2}$$

And the depth of subsurface flow is:

$$h = \frac{j}{K}.\frac{a}{s} \tag{3}$$

It can be seen from this expression that the depth of flow tends to increase with the "topographic index", a/s. If soil depths are more or less uniform downslope, then the degree of saturation is positively linked to the flow depth, h. It follows that the degree of saturation is generally more or less constant on the upper part of the slope profile, which is usually convex, so that s increases with a, and the degree of saturation increases sharply towards the base of the slope, where a tends to increase because of the plan convergence towards stream heads and s tends to decrease because of profile concavity.

Levels of saturation vary over time depending on the sequence of previous rainfalls. Under wet conditions, flow velocities near the surface are high enough to respond during the course of a storm, and under dry conditions, velocities deeper in the soil are slow enough to maintain low flows for several months after the last rainfall. Once the soil becomes saturated by this wedge of subsurface flow, additional rainfall cannot infiltrate however low its intensity, and flow is diverted laterally as "saturation excess" overland flow. Clearly this condition is met most frequently near the base of the slope, leading to time-varying areas of saturation excess overland flow and the concept of the dynamic "saturated contributing area" (Beven and Kirkby, 1979).

Semiarid Areas

In contrast, semiarid areas are characterized by an excess of potential evapotranspiration over precipitation. Thus, much of the rainfall that infiltrates into the soil is used by plants for transpiration, and, even where surface crown cover is sparse, the broader root network is able to utilize soil water efficiently. Under these conditions, the main movement of soil moisture is vertical, with little or no lateral subsurface flow. The conditions for saturation excess overland flow therefore rarely or never occur, and overland flow occurs only when and where rainfall intensity exceeds the infiltration capacity of the soil, which decreases as the wetting front penetrates into the soil. This is called *infiltration excess overland flow* or *Hortonian overland flow* after Horton (1933), who made this concept widely known.

Where the soil surface is directly exposed to raindrop impact, without protection by overhanging vegetation or stones, the surface layer of aggregates can be broken into their constituent particles that are then packed into the surface, creating an almost impermeable crust and sealing macropores. Crusts can also form in depressions where fine particles are washed in and below the surface where the fines are concentrated below a layer of sand or stable aggregates (Valentin and Bresson, 1992). In all cases, they greatly reduce infiltration rates and therefore increase infiltration excess overland flow (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**). Over time, crusts may be broken by, for example, plant shoots, ploughing, and wetting/drying cycles, so that their development and survival are highly dynamic in relation to weather and crop or natural plant growth.

Although semiarid areas are dominated by infiltration excess overland flow and vertical exchanges of soil water, there is a transition between the typical "humid" and "arid" responses. Infrequent storms and high intensities favor this Hortonian behavior, whereas frequent, low intensity storms favor subsurface flow and saturation excess behavior (Figure 4). Some areas show regular seasonal transitions between these modes, and exceptional periods of weather can lead to nontypical behavior for any area (Beven, 2002). Thus, summer thunderstorms can give infiltration excess overland flow in Britain and winter periods with persistent rainfall can generate subsurface flow in southern Spain.

With the dominance of vertical soil moisture exchanges in semiarid areas, the effect of topography is less important than for humid areas, and differences in overland flow response are more focused on areal differences in land use and soil properties.

In storms, most water infiltrates into the soil, even though local runoff coefficients may be high. As storm size is increased, more water infiltrates and penetrates deeper into the soil and some will be out of the reach of plant roots and evapotranspiration. This very small fraction of the rainfall is then available to percolate to depth, eventually recharging groundwater. Hillslope recharge is significant only in areas of well-developed karst, where surface water can reenter the soil through enlarged joints. Thus, in principle there is always a saturated zone at depth, but this interacts minimally with surface hydrology except in karstic areas.

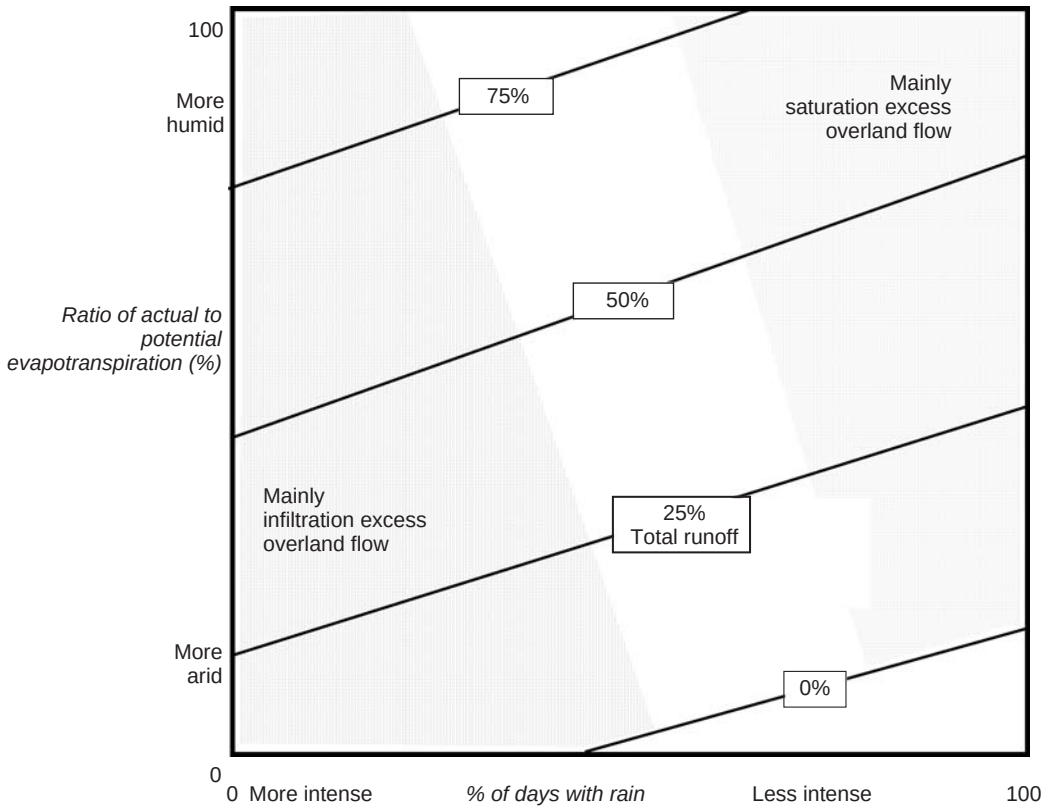


Figure 4 Generalized dependence of runoff coefficient and style of overland flow on degree of aridity and on storm rainfall intensities. Saturation excess overland flow is more common in humid areas, with a high ratio of rainfall to potential evapotranspiration, and under low rainfall intensities, indicated by low values of mean rain per rain day. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Connectivity

Overland flow is generated locally by infiltration and saturation excess mechanisms according to the climatic regime and local surface properties. However, not all of this flow necessarily reaches a stream channel, but may reinfiltrate on the way. This applies to overland flow generated from seminatural areas and uniform cropland, and also from artificial features, particularly roadways and cultivation artifacts such as wheel-tracks and headlands in ploughed fields.

The concept of connectivity considers whether a flux from a point A reaches a down-flow point B. The flux may be of any quantity, but in the present context refers to water flow. At any moment connection is either on or off but, over a period, there is a frequency distribution that varies from point to point, and connection may be expressed in terms of other hydrological state variables, usually with some uncertainty.

In humid conditions, there is generally good subsurface connectivity that determines the dynamics of saturated areas. Normally, the topographic index, a/s (from equation 3) increases steadily downslope along a flow path, but this is not necessarily the case. Where the topographic index shows a continuous increase and soils are more or less uniform, then saturation at a point can only occur when all points down-flow are already saturated, guaranteeing complete connectivity between every point and the catchment outlet. Where, however, differences in soil properties or intervening topography give a reduction in saturation along the flow path, then connectivity of surface flow only occurs if every point along the path is saturated. Figure 5 shows an example of the forecast pattern for an area in the Yorkshire Dales, UK, based on a 2 m DEM, where anomalous down-flow decreases in the topographic index occur at the margins of a summit plateau and between incised side slope gullies.

In general, humid areas generate runoff only in the narrow and variable saturated area close to the streams. In any storm overland flow, discharge is initially zero or very low and rises more or less linearly across the saturated area. Summed across the frequency distribution of storms, discharge can begin to grow far upslope in the most extreme storms, but only closer and closer to the stream in smaller and more frequent storms. The general form of increase

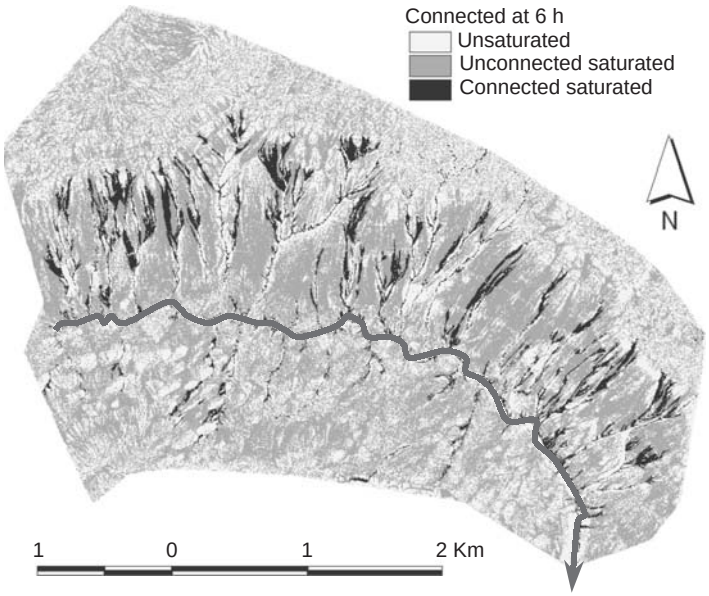


Figure 5 Example of partially connected overland flow, Wharfedale, UK (Reproduced from Lane *et al.*, 2004, by permission of John Wiley & Sons Ltd)). In many areas, saturated areas at any moment may not be connected to the catchment outlet by a continuously saturated pathway. Such disconnected areas are not contributing strongly to peak flows, although flow from them helps to saturate other areas downslope. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

in runoff production downslope may be expressed in a relationship such as

$$j = j_0 \exp\left(\frac{-x_0}{x}\right) \tag{4}$$

where j is the local rate of runoff production at distance x downslope, x_0 is a scale distance that decreases as storm size increases, and the discharge is obtained by summing this expression downslope.

In semiarid conditions, without the unity provided by subsurface flow, connectivity is generally much less complete. The two main reasons are thought to be spatial variability at various scales and the structure of rainstorms. In uncultivated areas, vegetation cover is generally incomplete in semiarid areas with bare patches between perennial plants. Bare patches may be crusted and have less of the organic litter and faunal activity that is concentrated under the vegetation, and therefore have lower infiltration and higher runoff. Bare patches may be joined to form a connected drainage path, but frequently drain into a vegetated area so that their runoff is lost and they do not connect with the streams. At a larger scale, fields may form a tapestry of different runoff generation rates so that high runoff areas may not connect with streams.

Storms in many semiarid areas consist of brief showers of intense rain with longer periods of lower intensity between them. Runoff may be generated widely during a shower and flow can begin to move downslope, but when the intensity drops this flow may reinfiltrate, with connected flow only reaching the stream from a narrow band along its banks. Average overland flow velocities are commonly only $1-2\,\mathrm{cm\,s^{-1}}$ so that flow travels only 18–36 m during a 30 min shower.

Both patchiness (i.e. spatial variability in runoff generation) and the short duration of intense showers limit the connectivity of overland flow, so that discharge increases less than linearly with increasing drainage area downslope. Instead, discharge in a storm initially increases downslope and finally reaches an asymptotic upper limit. The effective connection distance is generally greater in larger storms and discharge increases to a higher asymptotic value. Taking a weighted average across all storms, there is therefore a continuous increase of average discharge downslope, but at an ever decreasing rate, perhaps corresponding to a downslope increase in runoff production j and discharge q, of the form:

$$j = \frac{j_1}{(1 + x/x_1)}$$
$$q = j_1 x_1 \ln\left(1 + \frac{x}{x_1}\right) \tag{5}$$

Thus the downslope buildup of overland flow discharge in humid and arid regions takes the strongly contrasting forms sketched in Figure 6, both giving partial area contribution but generated through different mechanisms. In humid areas, saturation excess mechanisms concentrate runoff generation at the base of the slope. In semiarid areas,

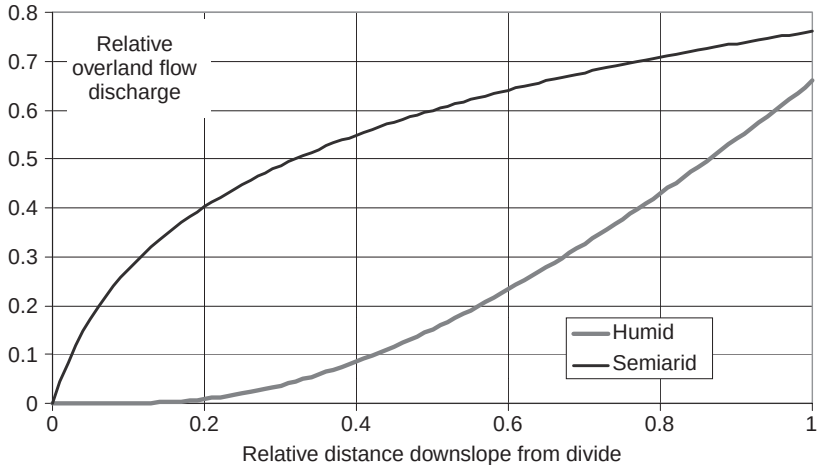


Figure 6 Sketch showing buildup of overland flow downslope in humid and semiarid climates. In humid areas, with saturation excess overland flow, local runoff generation generally increases downslope so that overland flow discharge increases more than linearly with hillside catchment area. In semiarid areas, although overland flow may be generated widely, flow from areas far upslope commonly reinfiltrates before reaching the stream, as storm intensity decreases. Overland flow discharge therefore increases less than linearly with hillside catchment area. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

with infiltration excess overland flow, runoff is generated everywhere but reinfiltrates because of the short connection distances (Kirkby *et al.*, 2002).

Roads and similar linear features tend to generate high volumes of local runoff, with an impermeable surface and local intersection of subsurface flow paths in cuttings. Like natural streams they provide efficient conduits to channel flow that collects in them, and in some areas they can significantly increase the effective drainage density and connectivity, particularly in forest areas (Croke and Mockler, 2001).

The form of the hillslope hydrograph is generally highly nonlinear, and typically shows a lower threshold for storm runoff, as well as increasing fraction of runoff as storm size increases. In humid regions, this is associated with an increase in the saturated area, and in semiarid areas with reductions in infiltration over time and the short duration of intense rainfall pulses, as sketched in Figure 7, although with wide variations according to the intensity structure of the storm. In the most extreme storms, all hillslopes would theoretically give 100% runoff irrespective of soil and climate, but the largest observed storms generally fall far short of this extreme, so that hillslope hydrographs generally show much greater and earlier response to rainfall as storm size increases, in strong contrast to the more linear response of stream channels.

Because of the different flow velocities along different primary pathways, the response time and peak flow generated along different pathways also varies greatly. Figure 8 illustrates the differences in observed peak flows for infiltration and saturation excess overland flow, and for shallow subsurface flow. It can be seen that they differ by more than an order of magnitude. Delays to peak show corresponding strong contrasts. In humid areas, both saturation excess overland flow and subsurface flow contribute to flood hydrographs. In some cases, the two pathways can generate double peaked hydrographs but, more commonly, subsurface flow supports the tail of the overland flow peak, giving a highly skewed hydrograph form, in which shallow subsurface flow is able to maintain low flows for many weeks after rainfall and provides sufficient water in some headwater areas to supply regional stream base flows. For arid areas, there is no subsurface contribution, and runoff commonly reinfiltrates as rainfall intensity decreases so that hydrographs not only show a rapid rise with a burst of intense rainfall, but also an abrupt decline as rainfall ceases. A hydrograph form is therefore more symmetrical for a pulse of rainfall and more generally shows a pattern of rise and fall, which closely mirrors the storm profile. Although there are wide variations, hillslope hydrographs in humid areas commonly peak in about 1 h and fall again over 24 h or more; whereas arid hillslope hydrographs peak within 15 min and fall again within 30 min.

Channelways

In their upper reaches, channels generally provide good transmission of the water flow delivered to them by hillslope hydrological pathways (*see* **Chapter 143, Mountain Streams, Volume 4**). This is because they are eroded forms, with thin or absent regolith cover, so that there is little opportunity for reinfiltration except over aquifers.

Natural and artificial channels typically intersect overland and shallow subsurface flow, and their effect on hydrological response results from the partially opposing effects of

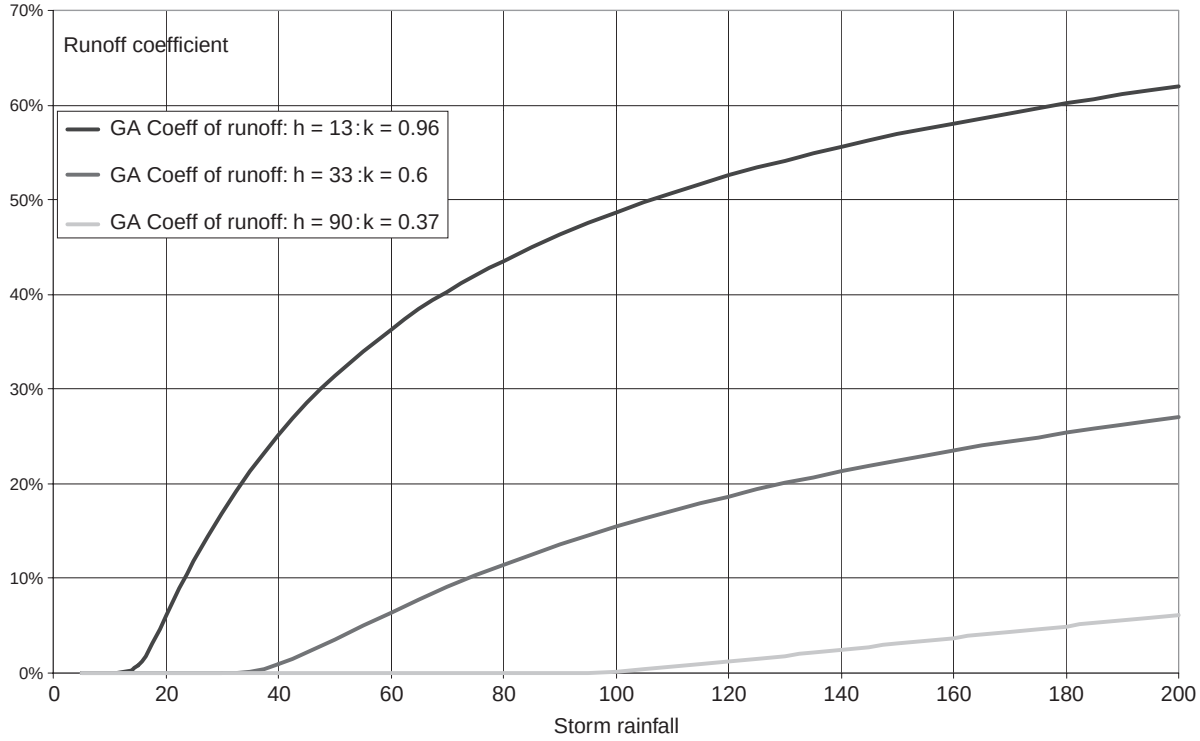


Figure 7 Generalized runoff coefficients for infiltration excess overland flow, for soils with different infiltration characteristics. Comparing sites with different infiltration rates, total storm runoff increases as infiltration rates fall. For comparable conditions, plots of accumulated runoff against accumulated rainfall show a family of nonoverlapping curves as infiltration rates are changed. Such behavior is described empirically by the SCS curve method (xx), as well as by applying infiltration equations (here Green and Ampt, 1911). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the surface and subsurface regimes. Firstly, direct channel precipitation provides an important component of rapid response to storms. Secondly, overland flow will be transmitted with lower reinfiltration losses within a channel. Thirdly, a high water table will be depressed locally by a channel cutting through it so that the initial response to rainfall may be a recharge of the water table, with only a slow release of the added water: this effect is strongest where the water table depression extends across the whole interchannel area (Holden *et al.*, 2004). Thus, an increase in channel density is likely to increase storm runoff in semiarid areas through the first two effects, but can reduce peak flows in humid areas with a high water table through the dominance of the third effect. Nevertheless, in wet climates, shallow water tables in the riparian zone may be very dynamic during events and contribute to subsurface stormflow (Sklash and Farvolden, 1979).

Farther downstream, channelways are increasingly associated with floodplain deposits that may absorb and carry part of the flow below the surface in endorheic flow and that allow high flows to spread over a much broader area. Where the sediment load is high relative to the transporting capacity of the channelway, channels are mobile, with frequent lateral shifting and often with multiple threads of flow in braided floodplains and deltas.

Permanent Flow

Channel flow is an almost linear process and the hydrology of large basins corresponds fairly closely to the assumptions of the Instantaneous Unit Hydrograph (IUH, **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**), or Geomorphological Unit Hydrograph (GUH), in which hillslope input hydrographs are linearly related to downstream output hydrographs, using the network width function (Surkan, 1968; Kirkby, 1976; Valdes *et al.*, 1979) to transform input to output within an area of uniform inputs. The assumption of linearity is physically based on a constant downstream routing velocity that is equivalent to a straight line relationship between discharge Q and cross-sectional area A:

$$Q = c(A - A_0) \tag{6}$$

where c is the routing velocity (Beven, 1979). This relationship empirically compares favorably for many rivers with the power law expression ($Q \propto A^{1.6}$) more commonly used.

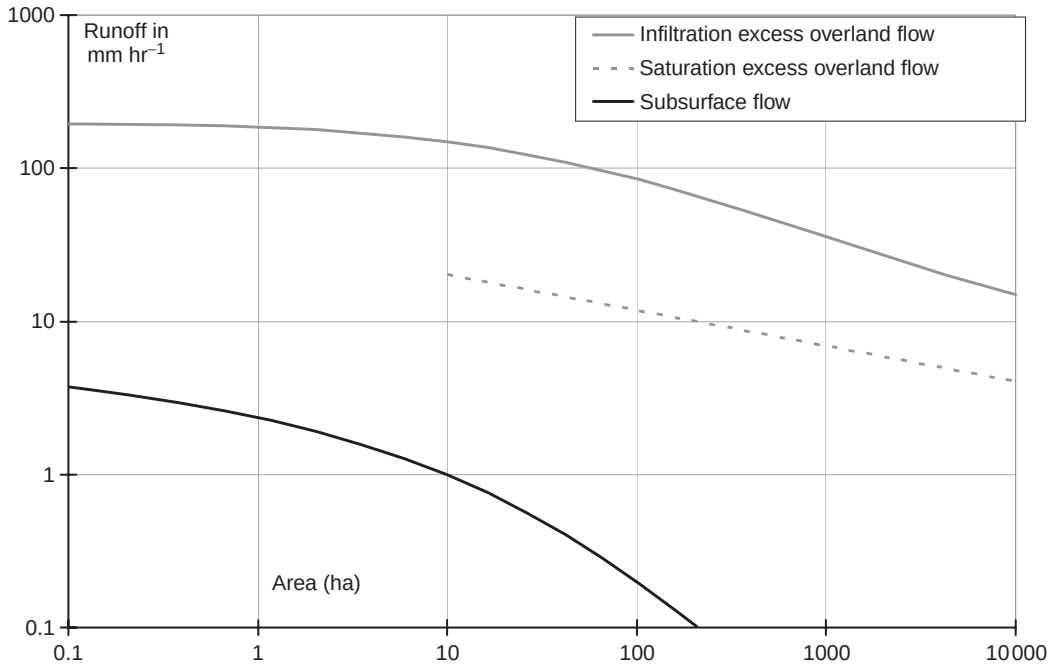


Figure 8 Peak hillslope responses from different runoff generation mechanisms (data from Dunne, 1978). The curves summarize empirical data that shows that the three runoff generation mechanisms, infiltration excess overland flow, saturation excess overland flow, and subsurface flow respectively generate lesser peak flows from a given area. Data also shows that runoff from these mechanisms generates respectively longer delays to peak flow

Clearly, equation (6) does not fully express the physical relationship between channel morphology and discharge, and, to that extent the linear assumption breaks down, and therefore does not provide the best available routing model. However, it does clearly demonstrate and explain the close relationship between channel network morphology and hydrograph response. Provided that attention is paid to the spatial and temporal patterns of rainfall, then catchments draining more than about 100 km^2 show a strong relationship between network morphology and flood hydrograph form, while smaller catchments show less of a link. The critical crossover catchment area varies somewhat with conditions, and is essentially the area above which the response time due to channel flow becomes greater than the response time of the hillslope hydrograph. For areas below this crossover, the nonlinearity of hillslope response then plays an increasing role in determining the form of the outlet hydrograph.

Where storms move across a catchment, then the movement of the storm cell, or the generation of new cells also has a strong impact on flood magnitudes. Where the storm moves downstream at a rate comparable to the celerity of the flood wave, there is very strong reinforcement of the flood peak, since storm rainfall from all parts of the catchment arrives at the outlet more or less simultaneously. If, however, the storm moves upstream, then the flood peak is attenuated and its duration extended.

Ephemeral Flow

In areas where streams are dry for part of the year, because of aridity or in karst areas, flow only occurs above a threshold determined by the subsurface flow capacity, in either deep karstic aquifers or in fluvial sediments within the channelway. Where storms are local, then flow may also be local, building up within the storm area and declining again through bed transmission losses downstream. In these environments, the simple near-linear behavior of humid streams is suppressed.

In extremely arid areas, or along channels with extensive gravels or karst sinkholes, floods may only occur with storms, and many events produce a flow in only part of the catchment. Other areas show a transitional behavior, with connected flow for part of the year and dry beds, sometimes with moisture persisting in bed gravels.

Since bursts of intense rain may produce a flood of only short duration and local extent, the movement of storm cells, and their sequence over time, becomes even more critical for the generation of floods over a catchment of more than a few km^2. Clearly, larger storms have more bursts of intense rain, and each burst may be larger, so that the potential for reinforcement increases with overall storm size, adding to the nonlinearity of arid catchment response.

Although recharge is generally small in semiarid areas, it is greatest along the stream channels because flow persists longest there. In many cases, this recharge supports

moisture within a local aquifer formed by sediment along the channelway. Where there is a uniform regional aquifer, however, recharge occurs primarily along the channels so that the piezometric surface is highest along the stream network, with a general form that is a low-gradient inversion of the surface relief.

IMPACTS ON THE LANDSCAPE

It has been shown above that the hydrological response of the landscape is strongly structured by the climatic regime, as well as by the form of the landscape. Turning to the impacts of the hydrology on the landscape, it will be seen below that the hydrology organizes and structures both the vegetation cover and the landscape morphology, although changes in morphology are only slowly implemented through the action of erosion and sediment transport, generally over periods of tens to hundreds of thousand years. Since climate is changing continuously, and on more rapid timescales, it is geologically normal to see an imbalance between the landscape forms and the contemporary hydrology, with landscapes rarely if ever in phase with climate and hydrology.

The flow of water is, as explained above, strongly influenced by the form of the landscape, expressed in hillslope plan and profile form, and in the density and morphology of the channel network. It is also strongly influenced by the climatic regime, expressed most relevantly by the balance between precipitation and potential evapotranspiration. The climatic link is significantly mediated by the vegetation cover that interacts strongly and dynamically with the hydrology, both influencing the response of the surface to rainfall and growing in response to available water.

Water flow is also, directly or indirectly, the main driver of sediment and solute transport, and the distribution and flow of water is therefore responsible for the erosion of the landscape, creating a second important impact and feedback. However, significant erosion of the landscape generally takes thousands to hundreds of thousand years so that the current landscape is not molded by current hydrology, but has been molded by a set of previous climates that commonly differ appreciably from the present. It is therefore important to consider the response times of different elements of the landscape and the implications of better or worse coupling between present hydrology and the present landscapes.

Vegetation Growth

Plant crown cover protects the soil surface, and its extent determines the proportion of the surface exposed to raindrop impact. Large drops fall at a terminal velocity of up to $10\,m\,s^{-1}$, and their momentum is largely absorbed where they fall on living or dead vegetation, or on surface litter or mulch on the soil surface. Drops that fall from the leaves may approach terminal velocities from high forest canopies ($>10\,m$), but their impact is generally much reduced below multistorey canopies or lower vegetation.

The vegetation, in turn, grows at a rate that is closely related to the transpiration stream, although there are variations according to the type of metabolic pathways (C3 or C4 plants), the level of atmospheric and soil CO_2, and the availability of nutrients, particularly nitrogen and phosphorus. Thus, many of the differences between humid and semiarid hydrological responses are directly related to natural vegetation growth, and artificial management of vegetation cover can create similar contrasts (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**–**Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**). This is most relevant in the context of human land management. Cropland is kept artificially clear of cover through tillage and harvesting for part of the year. Rangeland is grazed, keeping vegetation cover artificially low and is cleared periodically by fire to promote fresh growth in many regions. Widespread farming practices thus commonly allow seasonal crusting of the soil surface and infiltration excess overland flow in areas that would naturally be dominated by saturation excess runoff year round.

In semiarid areas, natural vegetation cover is generally incomplete, except for a carpet of annuals following rain. Perennials tend to be denser in areas of greater moisture, and therefore concentrated particularly along channelways, where flow lasts longest and moisture persists in the bed sediments.

Sediment and Solute Transport

Hillslope hydrology is vital to most sediment transport. Although hillslopes are not generally the most active part of the landscape, they provide almost all of the material that eventually leaves a river catchment through the more active channelways (*see* **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**). The processes by which material is weathered and transported to the streams are therefore vital to an understanding of how the catchment works. The regolith is also the raw material from which soils are developed. Geomorphological processes form an essential part of this crustal recycling that continually renews the surface, and is itself ultimately driven by the hydrological cycle (Holden, 2004; Kirkby *et al.*, 2002).

Water helps to break up the rocks into soil in the processes of weathering and drives the sediment transport processes that carry soil materials down to the ocean, progressively eroding the land. The balance between the rates of the various sediment transport processes has a very strong influence on the form of both the landscape and its soils, and plays a large part in the distinctive appearance of

landscapes in different climatic regions. Hillslope erosion also provides the raw material that rivers transport through their valleys to the oceans. Changes in hillslope erosion rate may not immediately be matched by similar changes in river sediment transport, resulting in either floodplain aggradation and widening, or valley incision.

Material transport processes are of two very broad types, (i) weathering and (ii) transport of the regolith. Within each of these types, there are a number of separate processes that may be classified by their particular mechanisms into three main groups. Most slope processes are greatly assisted by the presence of water that helps chemical reactions, makes masses slide more easily, carries debris as it flows, and supports the growth of plants and animals. For both weathering and transport, the processes can conveniently be distinguished as chemical, physical, and biological.

Weathering Processes

Weathering is the in-place transformation of parent materials into regolith and the further transformation of regolith materials. Chemical transformations change the chemical composition of the minerals in the regolith. The net effect of these changes is to remove the more soluble constituents of the rock minerals and change them into a series of new minerals, which become more and more like clay-forming minerals. Rock composition can be simplified as a combination of bases, silica, and sesquioxides (Carson and Kirkby, 1972). Bases are the most soluble so that weathering removes them first. Silica, although at least 10 times less soluble than bases, is itself at least 10 times more soluble than the sesquioxides. The changes due to chemical weathering can be shown on a triangular diagram (Figure 9),

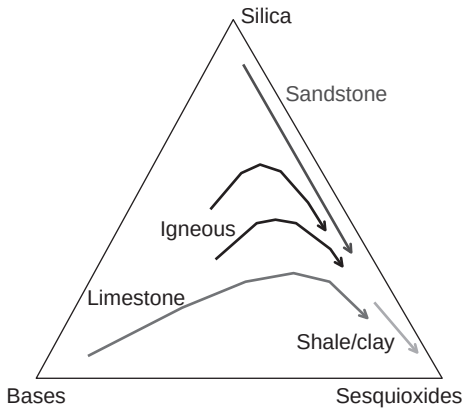


Figure 9 Schematic weathering of rocks, represented as a mixture of bases, silica ($Si0_2$), and sesquioxides (Al_2O_3, Fe_2O_3). The arrows illustrate courses of weathering for typical rock types and their lengths indicate relative rates of change. Thus, in the general course of weathering, bases are lost first, followed by sesquioxides, and long continued weathering leads to clay and eventually lateritic residual soils. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

which shows the proportional composition in terms of these three components. Weathering first reduces the proportion of bases and then the proportion of silica so that all parent materials eventually end up in the sesquioxide corner, essentially as clays, which are the end products in equilibrium with the pressures and temperatures in the regolith. In general, minerals that are formed close to the surface, such as the clays, are less liable to change in this way, and so undergo less chemical weathering than, for example, igneous rock minerals that were formed at high temperatures. Chemical weathering usually produces new minerals that are physically weaker than before, and so makes the material more easily transported by physical processes.

Physical weathering transforms rock and regolith materials by mechanically breaking them into smaller fragments. In some cases, chemical and biological processes help this mechanical breakdown. The most important processes of physical weathering are freeze–thaw and salt weathering. Biological processes are also effective through a combination of biochemical and mechanical methods, mainly through root action and faunal digestion (e.g. worms and termites).

Transport Processes

Where there is a plentiful supply of material, and the process which moves it can only move a limited amount for a short distance, then the rate of transport is limited by the *Transporting Capacity* of the process, which is defined as the maximum amount of material that the process can carry (Kirkby, 1971). Other processes are limited, not by their capacity to transport but by the supply of suitable material to transport, and are described as *Supply Limited*. Although there is not always a completely rigid line between Transport Limited and Supply Limited processes, it is an important distinction, and has a substantial impact on the way in which landscapes evolve. Landscapes that are dominated by transport limited removal are generally covered by a good layer of soil and vegetation, and slope gradients tend to decline through time. Landscapes where removal of material is mainly supply limited, on the other hand, tend to have sparse vegetation, thin soils, and steep slopes which tend to remain steep throughout their evolution. There is a general tendency for humid landscapes to be dominated by transport limited removal and semiarid landscapes by weathering limited removal.

Solute and Nutrient Transport The process of solution is closely linked to chemical weathering. As material is altered in place by weathering, the lost material is removed in solution. Rocks that weather rapidly therefore also lose material in solution rapidly. Rain falls on the soil, where it picks up solutes from the regolith, at a concentration roughly proportional to the amount of each constituent in the regolith, and its solubility. Some water is lost to evapotranspiration, and this carries little or no

solutes so that the remaining overland flow and subsurface flow runoff has an increased concentration of solutes. This distillation effect becomes marked in relatively arid climates, where the evapotranspiration is high. In extreme cases, some of the soluble material reaches its maximum saturated concentration, and any further evapotranspiration leads to redeposition of the dissolved material near the surface, most commonly for calcium, which is often found to form crusts of calcrete near the surface in arid and semiarid areas. The concentration of solutes is therefore generally highest in dry climates, but the total amounts removed in solution are less than for humid areas.

Where a flow of water contains dissolved material, the rate at which the solutes are removed is determined by their concentration in the runoff water. This leaching is generally the most important process in carrying solutes down the slope and into the rivers. Once material is leached out, it generally travels far downstream and its rate is *supply limited.*

When material is physically transported down a hillslope, it may travel as a mass or as independent particles. In a mass movement, a block of rock or soil moves as a single unit, although there may be limited relative movement within the block. The movement of the block is mainly determined by the forces on the block as a whole, and the individual rock or soil fragments within the block are in close contact so that they are moved selectively according to their size, shape, or density. The alternative to a mass movement is a particle movement, in which grains move one, or a few, at a time, and do not significantly interact with one another as they move. For a particle movement, forces act on each particle separately, and they move selectively, mainly depending on their sizes and also on other factors such as shape and density. Both mass movements and particle movements can occur at a range of rates. In general, however, movements that are driven by large flows of water tend to be more rapid than drier movements and tend to be supply limited, whereas drier and slower movements tend to be transport limited.

Mass Movements Mass movement of material is decided by a balance of forces, some of which promote movement and some of which resist movement. For mass movements, the forces act on the block of material that is about to move, and for particle movement on each individual particle. The main forces promoting movement are those of gravity and water detachment. On a slope, there is always a component of the weight of the material that tends to pull it downslope, and this applies equally to particle and mass movements. Friction and cohesion provide resistance to movement.

There are many names for different types of rapid and slow mass movements. In rapid mass movements, the crucial distinction is between *slides*, in which the moving mass essentially moves as a block over a failure surface, and *flows* in which different parts of the mass move over each other with differential movement or *shear*. It is usually found that flows occur in masses with more water mixed into the moving mass, in proportion to the amount of regolith or rock material. In a slide, water is often very important in reducing the frictional resistance, and allowing movement to begin, but there is little water within the moving mass. In a flow, there is usually almost at least as much water as solids, and sometimes many times more. Water and regolith materials can be mixed together in almost any proportions if they are moving fast enough, although coarse materials (sand, gravel, and boulders) can only remain suspended in the mixture in the fastest or most viscous flows.

Initiation of rapid mass movements occurs when gravity forces overcome frictional resistance and cohesion, either in a narrow zone or throughout the depth of the mass. Initial movement dilates the failure zone, separating the moving solids and reducing frictional resistance so that movement initially accelerates. Comparing mass movements with increasing proportions of water in the mass, the intergranular resistance during movement decreases progressively as grain to grain contacts become less frequent. Flow continues until either the driving forces decrease, generally by running out onto a lower gradient, or the frictional forces increase because of drainage of the water.

The essential contrast with slow mass movement (seasonal soil creep) is that it does not involve movement bounded by a discrete slip surface. Failures occur between individual soil aggregates and not over the whole of an area. Movements are usually driven either by "heaves" of expansion or contraction, or by apparently random movements between aggregates. Heaves are usually caused by freezing and thawing of soil water, or by wetting and drying of the soil. Random movements are usually caused by biological activity that mixes the soil in all directions. In all cases, these movements would not cause any net downhill movement on level ground but, on a slope, the steady action of gravity causes more downhill than uphill movement, and there is a gradual downhill transport of regolith material, at a rate which increases with slope gradient. These three main drivers for soil creep, wetting–drying, freeze–thaw, and biological mixing, can all be of similar magnitudes, although one or other dominates in any particular site. Tillage erosion is an additional form of slow mass movement, driven by ploughing and other cultivation that turns and dilates the topsoil at rates 100–1000 times more rapidly than in soil creep, moving material downslope without the direct agency of water.

Particle Movements Where there are cliffs, rockfall from bedrock slopes and its incorporation into talus slopes provides one important category of particle movement. The debris builds up screes of coarse debris that eventually weathers into finer material and may be removed.

Water is directly responsible for the other main process of particle movement, in the group of wash or soil erosion processes. In these processes, material may be detached by the two processes of raindrop impact and flow traction, and transported either by jumping through the air or in a flow of water. Combinations of these detachment and transport processes give rise to the three different processes, rainsplash, rainwash, and rillwash.

Raindrops detach material through the impact of drops on the surface. Drops can be as large as 6 mm in diameter and fall through the air at a terminal velocity that is related to their size. For the largest drops, the terminal velocity is $10\,m\,s^{-1}$, but they only attain this after falling through the air for about 10 m. If their fall is interrupted by hitting the vegetation, drops hit the ground at a much lower speed and have much less effect on impact. As drops hit the surface, their impact creates a shock wave that dislodges grains of soil or small aggregates and projects them into the air in all directions. Grains that are as large or larger than the raindrop which moved them may be detached, so that grains of up to 10 mm can be detached by raindrop impact. The total rate of detachment increases rapidly with the energy or momentum of the raindrops, and therefore with the rainfall intensity. As a working rule, the rate of detachment is roughly proportional to the square of the rainfall intensity. Where the raindrops fall into a layer of surface water that is deeper than the raindrop diameter (6 mm), the impact of the drop on the soil surface is largely lost. Impact through thinner films can still detach aggregates into the water, and other grains jump into flowing water films that can then transport grains which they do not have the power to detach. These forces are also responsible for breaking up the surface aggregates and packing them into the soil surface to form a crust, as described above, which severely limits subsequent infiltration and increases infiltration excess overland flow runoff.

Transportation through the air, in a series of hops, is able to move material both up- and downslope, but there is a very strong downslope bias on slopes of more than about 5°. As a rough guide, the net rate of transportation (downhill minus uphill) increases more or less linearly with slope gradient and inversely with the grain size transported. The gross rates of material transport, for rainsplash, are generally similar to those for soil creep. Rainsplash, however, is strongly particle size selective and operates only on the surface, whereas soil creep operates over a significant depth of soil and carries material together as a coherent mass. Protection from raindrop impact, either by vegetation or by stones, strongly suppresses rainsplash and crusting by reducing the impact velocity of raindrops. Microtopography, including tillage features are gradually smoothed out as rainsplash redistributes material, eroding high points, and filling depressions.

Once there is overland flow over the soil, material can be carried in the flow, and some material can move much farther than through the air in rainsplash. The presence of overland flow provides a thin layer of water on the soil surface, generally distributed rather unevenly following the microtopography that may attenuate the impact of raindrops. In shallow flows, the combination of detachment by raindrop impact and transport by the flowing water is the most effective transport mechanism and is known as *rainflow*. This process is active in inter-rill areas, and provides a significant fraction of the material carried into and along rills and larger channels.

If flow is deeper (i.e. more than 6 mm), raindrop detachment becomes ineffective, and detachment is related to the tractive stress of the flowing water. Sediment is detached when the downslope component of gravity and the fluid entrainment forces overcome frictional and any cohesive resistance in the soil, that is, when the safety factor falls below its critical value. It can be seen that detachment increases with discharge and gradient, and decreases with grain size except where cohesion is significant. Flows that are powerful enough to detach material generally suppress raindrop detachment, and detached material is also carried by the flow. This combination of processes is called *rillwash*, and is responsible for most of the erosion by running water in major storms. Much of the material exported from an eroding field is the direct product of channel enlargement during the storm, and almost all of the material detached by raindrop impact must also eventually leave the area through these channels.

Combining the effects of these three wash processes that are active during storms under a sparse vegetation cover, much of the area is subject only to rainsplash that feeds into areas, some spatially disconnected, with thin films of water where rainflow is dominant. These areas in turn provide sediment to the eroding channels where rillwash is actively detaching material and enlarging the channels. In larger storms, the areas of rillwash and rainflow increase and become better connected to the channels. The runoff generated per unit area and the area contributing runoff to the outlet therefore both increase, giving a more than linear response of runoff to increased storm rainfall. Because sediment transport also increases more than linearly with discharge, the nonlinearity of the sediment load dependence on rainfall is much stronger than for the water flow.

Selective transportation removes fine material from the soil, leaving a lag of coarse material that armors the surface. As the surface is lowered by erosion, the armor layer consists of the coarsest fraction in the layer of soil that has been eroded, and so develops more and more over time. The coarse armor progressively begins to protect the soil by reducing detachment rates, increasing infiltration by preventing soil crusting and providing an increased resistance to flow, and all of these effects reduce the rate

of erosion until some equilibrium is approached. In this equilibrium, local differences in sediment transport rate balance differences in armor grain size. This effect is most commonly seen in a relationship between surface grain size and gradient, with coarser material on steeper slopes.

The effects of selective transportation are only evident where the regolith contains some coarse material. This usually consists of weathered bedrock, but may consist of fragments of calcrete or other indurated soil horizons. Thus, the erosion of deep loess deposits or deeply weathered tropical soils, for example, that contain little or no coarse material, is not affected by the development of armoring, and may continue unchecked to great depths, often allowing the formation of extensive gully systems. On shallow, stony soils weathering from bedrock on the other hand, the effect of armoring is increased because, as the surface erodes, lower layers of the regolith contain less and less fines, and the end point of erosion may be a rocky desert. Some rocks, for example, coarse sandstones and granites, produce a discontinuous distribution of grain sizes in their weathering products, dominated by joint-block boulders of weathered rock and the sand grains that are produced as the boulders breakdown. On these rocks, desert slopes often show a sharp break in slope at the base of steep hillsides, between straight slopes close to the angle of rest and the basal concavity. If grain size is plotted against gradient for these slopes, the sharp break in slope represents missing gradients that correspond to the gap in the grain size distribution.

It would be hard to overemphasize the extreme importance of land cover and agriculture for wash processes. Through the actions of vegetation in preventing crusting and improving soil texture though the addition of organic matter, there is a very strong casual link from vegetation cover to increased infiltration capacity and reduced runoff and erosion. As forests have been cleared for agriculture, erosion rates have greatly increased, removing the most fertile topsoil or stripping mountains to bare rock and causing sedimentation along rivers. In the Middle East, Northern Europe, North America, and Brazil, this process of severe erosion and eventual partial recovery is evident at various stages. Today, land management for food production and amenity remains an important issue, requiring sensitive choices to limit erosion while maintaining productivity. Furthermore, through crusting, armoring, and the promotion of infiltration excess overland flow, wash processes have a powerful impact on hydrological response.

Channel Cutting and Infilling

Many sparsely vegetated areas develop rills and ephemeral gullies that are defined as temporary channels, formed during storms and destroyed by infilling between storms. In cultivated fields, infilling is generally through tillage, sometimes deliberately after each storm and otherwise following the annual cultivation calendar. In uncultivated areas, natural processes of wetting and drying, or freezing and thawing, create a loose surface layer that accumulates downslope along the depressed rill lines and gradually obliterates them. Rills are small channels, generally 5–10 cm deep that are formed on a smooth hillside and are not associated with a depression. Over a series of storms, the rills reform in different locations and gradually lower the whole hillside more or less evenly. Ephemeral gullies form along shallow depressions, and therefore reform along the same line in each storm, gradually enlarging and deepening the depression, while the infilling processes bring material from the sides and widen the depression.

In a particularly large storm, channels may form, which are too large to be refilled before the next event. These channels then collect runoff in subsequent events, leading to further enlargement, and may become permanent additions to the channel network. Where the soil is stronger close to the surface than in the horizons beneath, either due to the presence of a tough vegetation root-mat or due to the exposure of an indurated soil layer, then gullies that breach the surface layer may incise rapidly into the weaker horizons beneath. As material is exported, undercutting of the surface layer can lead to further rapid growth of a linear or branching gully system that disrupts agriculture and roads and may be very difficult to heal. These features can also intercept groundwater and may enlarge through subsurface piping, particularly in sodic soils.

Hillside erosion, either by wash erosion or mass movements, may deliver more sediment to the main channels than they are able to carry. In the short term, high erosion rates initiated by reductions in vegetation cover generally increase downstream transporting capacity less than the increase in load, so that channelways aggrade, giving broad valley floors, often with braided channels. Contrariwise, reductions in hillslope erosion, for example, by converting agricultural land back to forest, may cause downstream incision. In the long-term, channel systems adapt to carry changed sediment loads, but historical changes have shown that this adjustment may take hundreds of years (*see* **Chapter 84, Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands, Volume 2**).

Sediment transport in rivers is highly episodic, and sediment concentrations increase more than linearly with discharge so that sediment flux is strongly concentrated on the largest flood events, and more than half of the annual sediment load is carried in 5% of the time. In a flood event, layers of sediment are entrained roughly in the proportions in which they lie on the stream bed (equal mobility), where fine material is released when the flow entrains the coarser material that shelters it. However, material is deposited selectively according to grain size. The coarsest material typically travels short distances, as bed load, and its movement is usually transport limited.

Finer material is carried progressively farther in each flood event, and for the finest material is effectively supply limited, with transport limited by the lack of available material. Thus, silt and clay wash load and solute transport are largely controlled by the hillslope source areas for these materials. The progressive reduction in stream travel distance with increasing grain size is responsible for the much coarser grain size distributions found on stream beds than in hillslope soils, as similar volumes of material pass through every part of the system, from weathered bedrock to soils to channel deposits and down to the sea.

FEEDBACKS

The landscape form shapes the hydrological response to precipitation, both on the hillslopes and through the channel network, particularly in flood events. The form of flood hydrographs has a strong impact on the sediment transport throughout the catchment and, in the long-term, erodes and shapes the catchment. This nonlinear feedback loop reinforces some components of the hydrological response and some landscape features to give the familiar characteristics of fluvial landscapes, and the differences between landscapes that are often characteristic of particular lithologies, climates, and tectonic settings.

Landscapes evolve over time in response to the internal redistribution of sediment, usually with some net removal of material to rivers or the ocean. The way in which channel networks and hillslopes evolve depends on their initial form, the slope processes operating, and the boundary conditions that determine where and how much sediment is removed.

The assemblages of sediment and solute processes tend to fall into two groups (Figure 10), corresponding to the dominance of infiltration excess overland flow or subsurface flow.

In humid areas, where subsurface flow is dominant, wash processes are less important than large and small mass movements for transporting sediment. Subsurface flow carries dissolved material, often removing more material in solution than is lost as particulate sediment. This leads to good soil development, dominance of transport limited removal, and a strong hydrological reinforcement of hollows (plan concavities). Hollows preferentially generate subsurface flow so that solution is greatest there, leading to enlargement of the hollows, and this is particularly striking in areas with soluble rocks. These hollows then generate saturation excess overland flow, enlarging them by some surface erosion. Mass movement processes give rise to convex (creep) or rectilinear (landslides) profiles so that the characteristic slope profile consists of a broad convexity and a narrow concavity that is often also concave in plan, forming a hollow around stream heads.

In contrast, semiarid areas are dominated by infiltration excess overland flow; surface wash processes are the most

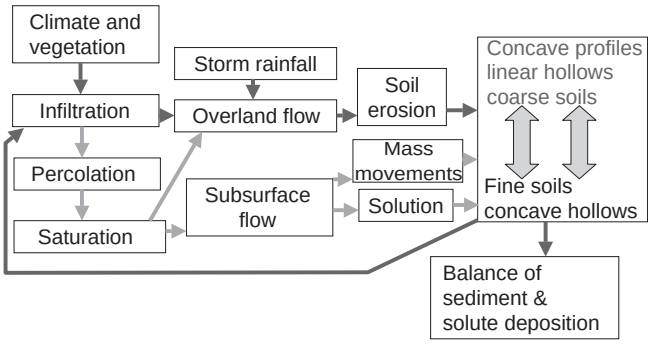


Figure 10 Overland and subsurface flow pathways and their geomorphic consequences. In semiarid conditions, hillslope development is dominated by overland flow and soil erosion, with relatively low chemical weather, and slopes are typically concave in profile, with coarse-textured soils. In humid areas, subsurface flow dominates, promoting weathering and mass movements, and tending to produce fine grained soils on mainly convex slope profiles, often with strong plan convergence around stream heads. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

important and solution less rapid. Soils are less weathered, thinner, and with lower clay content, and removal is commonly supply limited. The dominance of wash processes gives rise to concave slope profiles with only a narrow convexity around divides and, with none of the reinforcement processes to emphasize hollow development, landscapes tend to be more rectilinear in plan form.

Climate and hydrology also influence the density of stream channels within a landscape. Channels may be considered as a balance between hillslope processes that tend to fill them with sediment and channel processes that tend to clear them out. Both sets of processes are episodic so that stream heads may be filled with mass movement deposits for thousands of years before they are evacuated in a major flood. Thus, there are elements of chance in the position of observed stream heads due to the incidence of prior major flood events. The observed position may therefore fluctuate over time, but the persistence of a major valley represents some kind of time average of the balance between processes.

Theoretically (Smith and Bretherton, 1972), this balance is related to the balance between processes that increase less than linearly with drainage area (creep and rainsplash), for a given gradient, and those that increase more than linearly (certainly rillwash and perhaps landslides), that is, between processes that are not driven directly by water flow and those that are. For an unchanneled hillslope, it can be shown that there is a strong correspondence between the development of profile concavity and the dominance of water-driven processes.

As with hillslope forms, it is the balance between processes that determines the observed drainage density.

Two generalizations can be made, although the determining factors are still not fully understood. The first observation is that drainage density is generally higher in semiarid than in humid areas (e.g. Melton, 1957), and this seems to be a natural corollary of the narrower convexities in semiarid areas. The second observation is that drainage density is higher where gradient around the stream head is steeper. This result seems to indicate that the rate of water-driven processes increases more rapidly with gradient than for creep and rainsplash. The first of these effects is the stronger and creates regional contrasts, resulting in drainage densities of $>100\,\mathrm{km\,km^{-2}}$ in some semiarid areas, and down to $1-5\,\mathrm{km\,km^{-2}}$ in many humid areas. The gradient effect is generally weaker with about a 10 times range in the area needed to support a stream head, and is observed within as well as between areas following local topography.

CONCLUSION

In following the influences on catchment hydrology, it is clear that the form of the landscape is vitally important in affecting the response of the landscape to a storm event. The form of the landscape, however, is already a product of the hydrology over a long period, determining the soil, vegetation, hillslope forms, and channel network morphology. The climate has never remained constant, and is now changing, perhaps more rapidly than ever before. In understanding the structure of hydrological processes and their dynamic organization with respect to vegetation and landscape morphology, it is necessary to understand which parts of the physical system are truly constant, and which are dynamically changing in response to differences in both land use and climate.

FURTHER READING

Bull L.J. and Kirkby M.J. (Eds.) (2002) *Dryland Rivers*, John Wiley: Chichester, p. 388.

REFERENCES

Beven K.J. (1979) On the generalised kinematic routing method. *Water Resources Research*, **15**, 123.

Beven K.J. (2002) Runoff generation in semi-arid areas. In *Dryland Rivers*, Bull L.J. and Kirkby M.J. (Eds.), John Wiley: Chichester, pp. 57–105.

Beven K.J. and Germann P. (1982) Macropores and water flow in soils. *Water Resources Research*, **18**(5), 1311–1325.

Beven K.J. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Brammer D.D. and McDonnell J.J. (1996) An evolving perceptual model of hillslope flow at the Maimai catchment. In *Advances in Hillslope Processes*, Anderson M.G. and Brooks S.M. (Eds.), John Wiley: Chichester, pp. 35–60.

Carson M.A. and Kirkby M.J. (1972) *Hillslope Form and Process*, Cambridge University Press: p. 475.

Croke J. and Mockler S. (2001) Gully initiation and road-to-stream linkage in a forested catchment, southeastern Australia. *Earth Surface Processes and Landforms*, **26**(2), 205–217.

Dunne T. (1978) *Field studies of hillslope flow processes, in Hillslope Hydrology*, Kirkby M.J. (Ed.), John Wiley, Chichester, pp. 227–293.

Green W.H. and Ampt G. (1911) Studies of soil physics, part I – the flow of air and water through soils. *J. Ag. Sci.* **4**, 1–24.

Holden J. (Ed.) (2004) *An Introduction to Physical Geography and the Environment*, Pearson: Europe, p. 696.

Holden J., Chapman P.J. and Labadz J.C. (2004) Artificial drainage of peatlands: hydrological and hydrochemical process and wetland restoration. *Progress in Physical Geography*, **28**(1), 95–123.

Horton R.E. (1933) The role of infiltration in the hydrological cycle. *Transactions, American Geophysical Union*, **14**, 446–460.

Kirkby M.J. (1971) Hillslope process-response models based on the continuity equation. In *Slopes Form and Process*, Brunsden D. (Ed.), Institute of British Geographers: Special Publication No3, pp. 15–30.

Kirkby M.J. (1976) Tests of random network model, and its application to basin hydrology. *Earth Surface Processes and Landforms*, **1**(3), 197–212.

Kirkby M.J. (1993) Long term interactions between networks and hillslopes. In *Channel Network Hydrology*, *Ninth Edition*, Beven K.J. and Kirkby M.J. (Eds.), John Wiley: Chichester, pp. 255–293.

Kirkby M., Bracken L.J. and Reaney S. (2002) The influence of land use, soils and topography on the delivery of hillslope runoff to channels in SE Spain. *Earth Surface Processes and Landforms*, **27**(13), 1459–1473.

Lane S.N., Brookes C.J., Kirkby M.J. and Holden J. (2004) A network-indexbased version of TOPMODEL for use with high-resolution digital topographic data, *Hydrological Processes*, **18**(1), 191–201.

Melton M.A. (1957) *An Analysis of the Relations Among Elements of Climate, Surface Properties and Geomorphology*, Technical Report 11, Project NR 389–042, Office of Naval research, Columbia University.

Sklash M.G. and Farvolden R.N. (1979) Role of groundwater in storm runoff. *Journal of Hydrology*, **43**(1–4), 45–65.

Smith T.R. and Bretherton F.P. (1972) Stability and the conservation of mass in drainage basin evolution. *Water Resources Research*, **8**(6), 1506–1529.

Surkan A.J. (1968) Synthetic hydrographs – effects of variations in network geometry. *Transactions, American Geophysical Union*, **49**(1), 164.

UNESCO (2000) **http://www.unesco.org/science/waterday2000/Cycle.htm**

Valdes J., Fiallo B.Y. and Rodrigues-Iturbe I. (1979) A rainfall-runoff analysis of the geomorphological IUH. *Water Resources Research*, **15**(6), 1421–1434.

Valentin C. and Bresson L.M. (1992) Morphology, genesis and classification of surface crusts in loamy and sandy soils, *GEODERMA*, **55**(3–4), 225–245.

5: Fundamental Hydrologic Equations

ROGER BECKIE

Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada

In this article our goal is to present an overview of the fundamental principles that are the basis of most models used in hydrology. We develop the fundamental principles of mass, momentum, and energy conservation and express them in mathematical form. We first outline the general approach that can be used to develop a mathematical statement of a conservation law, using a so-called Eulerian framework, where we consider volumes fixed in time and space through which material may flow. We then derive the general conservation equations for mass, momentum, and energy for the case of flowing fluids. We next provide examples from hydrology that illustrate the application of the general conservation principles. We begin with relatively straightforward applications of the conservation equations and progress to more complex and less direct applications. Our first and simplest example is the advection–dispersion equation, which is a relatively transparent application of the conservation of mass principle, augmented with a so-called gradient-flux model, Fick's law, which describes the dispersion and diffusion of solute mass within the bulk flowing fluid. Next we present the Navier–Stokes equations, which are the conservation of momentum equations for a Newtonian fluid. The next suite of examples involves flow in porous media, which is described by more than one conservation principle applied simultaneously. Our last example is from engineering hydraulics, the Saint Venant equations, which are gross but practical simplifications of the general conservation statements.

INTRODUCTION

The principles of mass, momentum, and energy conservation are fundamental to quantitative hydrology. Most physically based models in hydrology respect these principles, and those models that violate them are viewed with great suspicion. These fundamental conservation principles are applied in the form of mathematical expressions so that they can be communicated precisely and succinctly and used for quantitative analysis. One may view mathematics as a "grammar" and "language" to communicate the "literature" of physical principles. In this article, we shall therefore focus mostly upon the mathematical development of the conservation principles and their manifestation in hydrology. In this way, we hope that our development will give physical meaning to the mathematical expressions. We will assume that the reader has an elementary understanding of multivariate calculus and differential equations.

We first state the fundamental principles and then outline the general ways in which mathematical expressions for these principles are formulated. We then develop general mathematical equations for each principle for the specific case of a flowing fluid, which in hydrology is usually water. These general equations are then typically used as the basis for expressions used in hydrology. We therefore conclude the article with a sequence of examples from hydrology to show how the principles are manifested in expressions that describe hydrologic processes.

FUNDAMENTAL CONSERVATION PRINCIPLES

The three fundamental principles are:

1. Conservation of mass: Mass is neither created nor destroyed.
2. Conservation of momentum: The momentum of a body is conserved unless it is acted upon by a force, in which case the rate of change of momentum equals the net force on the body.
3. Conservation of energy: Energy is neither created nor destroyed.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

There are two general ways in which these fundamental principles are applied in hydrologic analysis. In one way, the principles are used to establish an accounting or bookkeeping scheme to keep track of the mass, momentum, or energy moving into and out of a finite-sized closed region called a *control volume*. When applied to control volumes, the resulting conservation laws typically lead to algebraic expressions that quantify the gross behavior of the entire control volume. In the second general usage, conservation statements are written that apply at every mathematical point in a continuous medium. The conservation statements are then written in the form of differential equations, which when solved, yield detailed, point-wise information about the system. These continuous conservation statements can be understood and developed as limiting cases of finite-sized control volume statements. Generally, the control-volume approach is simpler to apply, but leads to coarser, bulk results. The second approach, using differential equations, has the potential to provide much more detailed information at the cost of a more difficult computation. Indeed, for many problems, analytical solutions do not exist for the differential equations and analytical simplifications or numerical approximations are required.

All the three conservation principles follow the same essential theme. When applied to a control volume, the conservation principles imply that *the amount of the conserved quantity (mass, momentum, or energy) entering a control volume during a specified time period, less the amount leaving during that time period, must equal the net change in conserved quantity stored in the control volume during that time period.*

For example, consider the control volume shown in Figure 1, which contains a catchment. The dotted line represents the surface of the control volume. If we let P be the precipitation rate, and ET be the rate at which water evaporates or transpires from plants out of the catchment, and G_{in} and G_{out} be the rates at which groundwater enters and leaves the catchment, and Q_{in} and Q_{out} be the rates at which surface water enters and leaves the catchment, then we can write the following statement for the conservation of mass (water) for an arbitrary time period Δt:

$$(P - ET + G_{\text{in}} - G_{\text{out}} + Q_{\text{in}} - Q_{\text{out}})\Delta t = \Delta S \qquad (1)$$

where ΔS is the change in mass stored over the time period Δt and where each term in the brackets has dimension of mass per unit time $[M/T]$. This algebraic expression of mass conservation is called a *water balance*. The terms in the water balance may depend upon other variables, leading to a more complex expression.

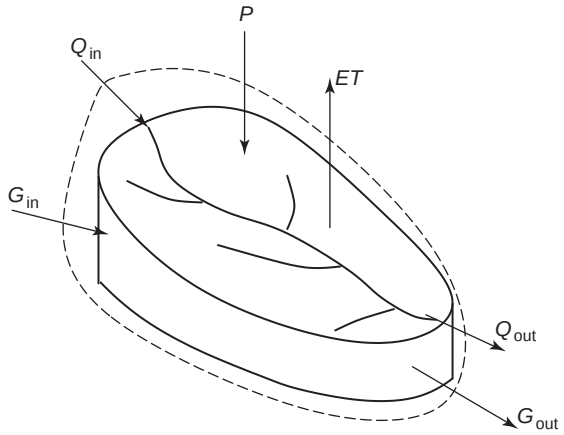


Figure 1 A control volume is any closed region where fluxes across the volume boundaries and changes in internal storage of mass, momentum, or energy are accounted for over specified time intervals. The catchment above can be considered a finite-sized control volume. P is precipitation flux, ET is the evapotranspiration flux, Q_{in} and Q_{out} are surface water flow into and out of the catchment, and G_{in} and G_{out} are subsurface fluxes into and out of the catchment

DIFFERENTIAL FORM OF CONSERVATION EXPRESSIONS

Differential representations of the fundamental conservation principles for flowing fluids are typically developed in one of two ways. In the first way, one refers to a control volume and then considers how a physical quantity, such as the mass contained in the control volume, changes with time. In this case, the control volume is usually fixed in space and time while fluid moves through it. This type of description is called an *Eulerian description*. In the second approach, a system under examination is considered to be composed of so-called material particles, which are small elements of constant mass. One then describes how properties such as the energy content, momentum, and geometry of each particle change in time. This type of description is known as a *material* or *Lagrangian description* of the system. In a sense, one tags along with an element of fluid mass and describes how it moves and its geometry, momentum, and energy change. The two different descriptions lead to different mathematical representations for the same conservation concepts. Convenience dictates the preference, although the Eulerian approach appears to be most common in hydraulics and hydrology. We shall develop Eulerian expressions here.

The derivations for the differential form of the mass, momentum, and energy conservation expressions all follow the same basic pattern. We consider a finite-sized control volume, fixed in space with fluid moving through it, and a time interval from t to $t + \Delta t$. We then write expressions

for: (i) $\{Stored\}_t$ and $\{Stored\}_{t+\Delta t}$, the amounts of the conserved quantity stored in the control volume at times t and $t + \Delta t$; (ii) $\{Net\ flux\}_{t \to t+\Delta t}$, the net amount of the conserved quantity carried into or out of the control volume by the flowing fluid during the time interval t to $t + \Delta t$; and (iii) $\{Sources\}_{t \to t+\Delta t}$, the net amount of the conserved quantity added to or removed from the control volume by other processes during the time from t to $t + \Delta t$. Processes that can add or remove a conserved quantity to or from the control volume are typically specific to a conserved quantity. For example, the process of heat conduction can transport energy into and out of the control volume. With the expressions described above, the conservation principle for the control volume is written:

$$\{Net\ flux\}_{t \to t+\Delta t} + \{Sources\}_{t \to t+\Delta t} = \{Stored\}_{t+\Delta t} - \{Stored\}_t \quad (2)$$

If we then divide this expression by the volume of the control volume and the time interval Δt, and shrink the volume and time interval to a point in space and time, then a differential conservation expression results.

The differential conservation equations are idealizations that apply at the mathematical point and instant in timescale, whereas the finite-volume conservation equations, such as the catchment-scale water balance (equation 1), apply to finite volumes for discrete intervals of time. Accordingly, if the differential conservation equation holds at each point in a region of space over a continuous time interval, then a finite-volume conservation expression can also be developed for any closed finite subvolume and time interval in that region.

The general conservation equations that we develop are valid for any continuous media such as solids, gases, and liquids. In the case of hydrology, the most common medium will be water. The only critical assumption is that the media be continuous, meaning that its physical properties change smoothly in time and space, such that derivatives of these properties exist at all points. We shall see later in the examples where we develop equations for flow in porous media that we have to first define a continuum to apply these principles. We next derive the differential mass, momentum, and energy conservation equations.

Conservation Of Mass

The mass conservation principle states that mass is neither created nor destroyed. Einstein showed that this is not exactly true, since mass and energy are equivalent. However, for almost all processes of interest to hydrology, energy levels are not high enough for a measurable exchange between mass and energy. Consequently, mass is essentially conserved in hydrologic systems.

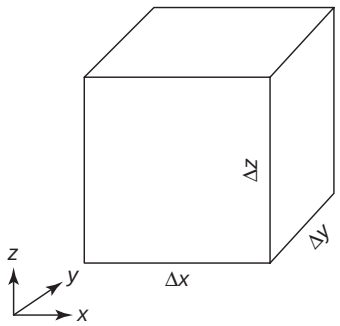


Figure 2 A control volume that is used to develop differential forms of the conservation principles. We conceptually shrink this finite-sized control volume to a mathematical point to develop differential forms of the conservation principles

To develop a differential expression, we examine mass conservation over an interval of time from t to $t + \Delta t$ for a rectangular control volume of size $\Delta x \Delta y \Delta z$ fixed in space at a point (x, y, z) (Figure 2). To avoid confusing subtleties, we first assume that the control volume is always fully saturated with water. Later we will consider more than one fluid phase when we develop expressions describing the unsaturated zone in soils. We also assume that we can define mathematically continuous parameters such as density ρ and dependent variables such as velocity with components (v_x, v_y, v_z). To derive the mass conservation expression, we must compute the net mass flux into the control volume and equate it with the change in mass storage over the time interval Δt.

Consider the two faces of the control volume that are perpendicular to the x coordinate direction, one located at x and one at $x + \Delta x$ (Figure 3). Over the time interval, Δt the mass carried into the control volume by the flowing fluid on the x face is $\rho v_x|_x \Delta y \Delta z \Delta t$, and the mass carried out of the $x + \Delta x$ face is $\rho v_x|_{x+\Delta x} \Delta y \Delta z \Delta y$ (the notation $\rho v_x|_{x+\Delta x}$ is to be read "ρv_x evaluated at the

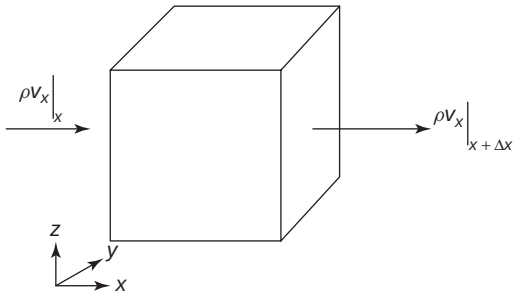


Figure 3 The control volume used to develop the conservation of mass equation. Here, ρ is the fluid density and v_x is the x-component of the fluid flow velocity. Accordingly, ρv_x is the x-component of the mass flux vector

Table 1 Conservation of mass flux terms

Face	Inlet mass flux	Outlet mass flux
x	$\rho v_x\|_x \Delta y \Delta z \Delta t$	$\rho v_x\|_{x+\Delta x} \Delta y \Delta z \Delta t$
Y	$\rho v_y\|_y \Delta x \Delta z \Delta t$	$\rho v_y\|_{y+\Delta y} \Delta x \Delta z \Delta t$
z	$\rho v_z\|_z \Delta x \Delta y \Delta t$	$\rho v_z\|_{z+\Delta z} \Delta x \Delta y \Delta t$

point located at $x + \Delta x$"). In the same fashion, we can express the mass flux across the faces in the y- and z-direction. The fluxes across each face are summarized in Table 1.

The mass contained in the volume at time t and time $t + \Delta t$ is simply the density times the volume at those times, or $\rho \Delta x \Delta y \Delta z|_t$ and $\rho \Delta x \Delta y \Delta z|_{t+\Delta t}$. Referring to the general conservation principle in equation (2), where we assume that there are no sources of mass inside the volume, we can write the finite-volume mass conservation expression

$$\begin{aligned}(\rho v_x|_x - \rho v_x|_{x+\Delta x})\Delta y \Delta z \Delta t \\ + (\rho v_y|_y - \rho v_y|_{y+\Delta y})\Delta x \Delta z \Delta t \\ + (\rho v_x|_z - \rho v_z|_{z+\Delta z})\Delta x \Delta y \Delta t \\ = (\rho|_{t+\Delta t} - \rho|_t)\Delta x \Delta y \Delta z \end{aligned} \quad (3)$$

Dividing by the volume $\Delta x \Delta y \Delta z$ and the time interval Δt we have

$$\begin{aligned}\frac{(\rho v_x|_x - \rho v_x|_{x+\Delta x})}{\Delta x} + \frac{(\rho v_y|_y - \rho v_y|_{y+\Delta y})}{\Delta y} \\ + \frac{(\rho v_z|_z - \rho v_z|_{z+\Delta z})}{\Delta z} = \frac{(\rho|_{t+\Delta t} - \rho|_t)}{\Delta t} \end{aligned} \quad (4)$$

Last, recognize that as the dimensions of the volume and the length of the time interval shrink, we can interpret terms as derivatives $(\rho v_x|_x - \rho v_x|_{x+\Delta x})/\Delta x \rightarrow -\partial(\rho v_x)/\partial x$, where $\partial(\rho v_x)/\partial x$ is the partial derivative of ρv_x with respect to x, and can be thought of as the rate at which the value of ρv_x changes in the x-direction. The result is the mass conservation equation for a flowing fluid,

$$-\frac{\partial(\rho v_x)}{\partial x} - \frac{\partial(\rho v_y)}{\partial y} - \frac{\partial(\rho v_z)}{\partial z} = \frac{\partial \rho}{\partial t} \quad (5)$$

This equation is also known as the *continuity equation*. The continuity equation is written using the so-called vector notation as

$$-\nabla \cdot (\rho \overline{v}) = \frac{\partial \rho}{\partial t} \quad (6)$$

where $\overline{v}$ is the notation for the water velocity vector, and the divergence operator $\nabla.()$ is defined for Cartesian coordinates as $\nabla \cdot (\rho \overline{v}) = \partial(\rho v_x)/\partial x + \partial(\rho v y)/\partial y + \partial(\rho v_z)/\partial z$.

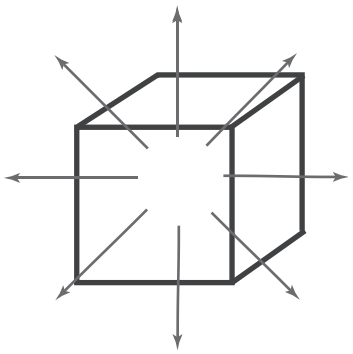


Figure 4 Divergence of a quantity is positive at a point when the flux of the quantity is diverging from a point. In the absence of sources of the quantity at that point, the amount of the quantity at that point must be decreasing

A conservation expression is often betrayed by the divergence operator, in vector form $\nabla \cdot (\overline{f})$, where $\overline{f}$ is the flux rate of a conserved quantity into a point by some transport process (e.g. in the conservation of mass, equations (5) or (6), the mass flux rate is $\rho \overline{v}$ and the transport process is the flow of water). As shown in Figure 4, a positive divergence at a point indicates that the flux of the conserved quantity is *diverging* or flowing away from the point. Accordingly, the amount of the conserved quantity at that point must be decreasing with time. Mathematically, if for example, the divergence of the mass flux is positive at a point, $\nabla \cdot (\rho \overline{v}) > 0$, then by equation (6), $\partial \rho / \partial t < 0$ and the mass at that point must be decreasing. The divergence is easy to recognize when written in vector form such as in equation (6).

Essentially, the differential continuity equation is a balance between the net mass flux rate into a mathematical point (the three terms on the left-hand side of equation 5) and the rate at which the mass stored in the point changes in time (the term on the right-hand side). However, because the equation is derived from a finite-sized control volume, it is sometimes easier to interpret the differential conservation equations with a small control volume in mind. We will show in examples to follow that a conservation principle is concealed in many of the common expressions used in hydrology.

An important, special case of the continuity equation is that for an *incompressible fluid*. If the fluid is incompressible, $\partial \rho / \partial t = 0$ and so $\nabla \cdot (\rho \overline{v}) = 0$, and since $\nabla \rho = 0$ also holds for an incompressible fluid, the continuity equation (6) becomes

$$\nabla \cdot \overline{v} = 0 \quad (7)$$

Conversely, at any moment when the divergence of the fluid velocity is not zero, $\nabla \cdot \overline{v} \neq 0$, then the fluid must be compressing or expanding. Indeed, we state without proof

that $\nabla \cdot \overline{v}$ equals the volumetric rate of deformation of a fluid element (Panton, 1984).

Conservation Of Momentum

The relationship between momentum and force is given by Newton's second law, which states that a force $\overline{F}$ (a vector) on an isolated body equals the mass of the body times the acceleration. Acceleration is defined as rate of change of the velocity vector $\mathrm{d}\overline{v}/\mathrm{d}t$ so

$$\overline{F} = m\frac{\mathrm{d}\overline{v}}{\mathrm{d}t} = \frac{\mathrm{d}(m\overline{v})}{\mathrm{d}t} \tag{8}$$

where the momentum is $m\overline{v}$. The conservation of momentum principle states that *the momentum of a body is constant unless it is acted upon by a force, in which case the rate of change of momentum equals the net force* (equation 8). The conservation of momentum equations can be developed for a flowing fluid in much the same fashion as the continuity equation. There are actually three equations, since momentum is a vector quantity, unlike mass, which is scalar.

To develop the momentum balance, we must consider all forces acting on the fluid, since forces are a source of momentum. It is sometimes difficult to determine what forces are relevant to a problem, especially when the reference frame (coordinate system) that is used for the problem is accelerating. Such a coordinate system is called a *non-inertial reference frame*. For large-scale problems in geophysical fluid dynamics, in the order of tens of kilometers and more, the acceleration of the earth's surface, relative to the fixed stars, means that a coordinate system fixed to the earth will be non-inertial. For smaller-scale problems, the effect of the earth's rotation is negligible and can be ignored.

We derive the differential form of the conservation of momentum for a flowing fluid in an inertial (nonaccelerating) reference frame. For simplicity, we will only present the development for the x-component of momentum in detail. We follow the same procedure that we used to develop the mass conservation equation: we consider a small control volume of size $\Delta x \Delta y \Delta z$ fixed in space at a point (x, y, z) and compute the momentum flux by fluid flow into the volume over a time period from t to $t + \Delta t$, plus any momentum-generating forces acting on the fluid, and equate these with the change in momentum stored in the volume.

The x-momentum contained in the volume at time t and time $t + \Delta t$ is the fluid density times the velocity in the x-direction (i.e. the x-component of momentum) times the volume, or $\rho v_x \Delta x \Delta y \Delta z|_t$ and $\rho v_x \Delta x \Delta y \Delta z|_{t+\Delta t}$. Following the argument for the conservation of mass, the $x-$component momentum fluxes carried into and out of the volume by the flow are summarized in Table 2.

Table 2 Conservation of momentum flux terms

Face	Inlet x-momentum flux	Outlet x-momentum flux
x	$\rho v_x v_x\|_x \Delta y \Delta z \Delta t$	$\rho v_x v_x\|_{x+\Delta x} \Delta y \Delta z \Delta t$
y	$\rho v_x v_y\|_y \Delta x \Delta z \Delta t$	$\rho v_x v_y\|_{y+\Delta y} \Delta x \Delta z \Delta t$
z	$\rho v_x v_z\|_z \Delta x \Delta y \Delta t$	$\rho v_x v_z\|_{z+\Delta z} \Delta x \Delta y \Delta t$

Two other similar tables could be listed for the y- and z-component fluxes. Each term in this table is actually an impulse, that is, a force through time, which according to equation (8), is an increment of momentum, that is, $\overline{F} \times \mathrm{d}t = \mathrm{d}(m\overline{v})$.

The terms in Table 2 quantify the flux of momentum into the control volume by the flow of the fluid. To complete conservation law, we must account for remaining sources of momentum, namely, the forces acting on the fluid.

The forces acting on the fluid include gravity, pressure gradients, and friction (viscous stresses). Here it is best to conceptualize the conservation equations with a control volume in mind (a point-scale volume for the differential equations). In that context, forces acting on a fluid are often conceptualized as body forces, acting on the bulk material or as surface forces, typically pressure and viscous drag effects acting on the boundaries of the control volume. We next compute the impulses contributed by these forces on the fluids.

The x-component of impulse contributed to the fluid in the control volume by gravity is simply the mass of fluid times gravity (the gravity force) times the time interval, $\rho g_x \Delta x \Delta y \Delta z \Delta t$ where g_x is the x-component of the acceleration of gravity.

To show how pressure leads to a net force, we must consider the small control volume again (Figure 5). Pressure over an area is a force, so there is an inward directed force on each face of the control volume. If the pressure

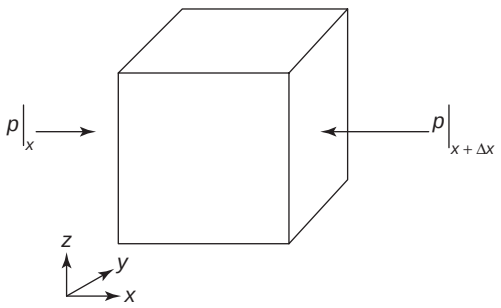


Figure 5 Pressure on a face creates a force that is directed toward the center of the volume. If pressures on opposite sides of the volume are the same, then each generates the same force and the net force is zero. Pressure forces therefore result from pressure differences, which are quantified by the pressure gradient

is the same everywhere, then there will be no net force on the volume. A net force results when there is a difference in pressure force in a given direction. Accordingly, the net x-component impulse contributed by fluid pressure is $(p|_x - p|_{x+\Delta x})\Delta y \Delta z \Delta t$. As we show next, a stress imbalance across the control volume also leads to a net force.

Internal shearing and viscosity lead to stresses that are felt on each face of the control volume. Consider the control volume in Figure 6. On the x face at $x + \Delta x$ is shown three components of the viscous stress τ. For example, on the face perpendicular to the x-direction, there is a normal viscous stress τ_{xx} and two shear components, one pointing in the z-direction, τ_{zx} and one in the y-direction, τ_{yx}. The viscous stress state in the fluid can be described by a 3×3 *viscous stress tensor* that is given in Cartesian components as

$$\overline{\overline{\tau}} = \begin{bmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{bmatrix}$$

A tensor is a generalization of the concept of a vector, and has magnitude and directional properties (Bird *et al.*, 1960). For our purposes, each component of the stress tensor can be thought of as a force per unit area such that a net force on a surface from a stress component is the stress component times the area over which it acts. Following the example for pressure from above, (see Figure 5) the net impulse from the viscous stress force in the x-direction results from the differences in normal stress components at x and $x + \Delta x$, $(-\tau_{xx}|_x + \tau_{xx}|_{x+\Delta x})\Delta y \Delta z \Delta t$, plus the x-component of the net shear stress forces contributed by differences in shear forces at the faces perpendicular to y-direction, $(-\tau_{yx}|_y + \tau_{yx}|_{y+\Delta y})\Delta x \Delta z \Delta t$ and the z-direction, $(-\tau_{zx}|_z + \tau_{zx}|_{z+\Delta z})\Delta x \Delta y \Delta t$. Note that stress here is positive for tension, in contrast to pressure, which is positive for compression; in some developments stress is positive for compression however.

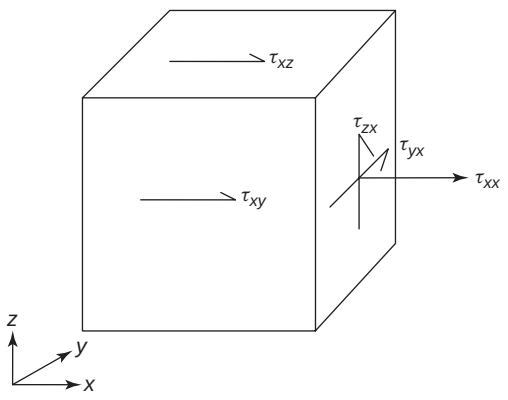


Figure 6 Stress over an area is a force. The net x-component force is a sum of the force caused by the normal stress on the face normal to the x-direction, τ_{xx} plus the shear stresses on the other faces, such as τ_{xy} and τ_{xz}

We now collect the expressions for the flux of momentum through the boundary of the control volume, the sources of momentum from surface and body forces, and equate the sum to the change in the momentum stored in the control volume at times t and Δt to give

$$\begin{aligned}
&(\rho v_x v_x|_x - \rho v_x v_x|_{x+\Delta x})\Delta y \Delta z \Delta t \\
&\quad + (\rho v_x v_y|_y - \rho v_x v_y|_{y+\Delta y})\Delta x \Delta z \Delta t \\
&\quad + (\rho v_x v_z|_z - \rho v_x v_z|_{z+\Delta z})\Delta x \Delta y \Delta t \\
&\quad + \rho g_x \Delta x \Delta y \Delta z \Delta t + (p|_x - p|_{x+\Delta x})\Delta y \Delta z \Delta t \\
&\quad + (-\tau_{xx}|_x + \tau_{xx}|_{x+\Delta x})\Delta y \Delta z \Delta t \\
&\quad + (-\tau_{yx}|_y + \tau_{yx}|_{y+\Delta y})\Delta x \Delta z \Delta t \\
&\quad + (-\tau_{zx}|_z + \tau_{zx}|_{z+\Delta z})\Delta x \Delta y \Delta t \\
&\quad = (\rho v_x|_{t+\Delta t} - \rho v_x|_t)\Delta z \Delta y \Delta z
\end{aligned} \tag{9}$$

Dividing by the volume and time interval, and shrinking the control volume and time interval as we did in equations (3) to (5), we arrive at the following conservation of the x-component of momentum equation

$$-\frac{\partial(\rho v_x v_x)}{\partial x} - \frac{\partial(\rho v_y v_x)}{\partial y} - \frac{\partial(\rho v_z v_x)}{\partial z} - \frac{\partial p}{\partial x} + \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z}\right) + \rho g_x = \frac{\partial(\rho v_x)}{\partial t} \tag{10}$$

Similarly, there are two more equations for the y- and z-components of the momentum vector

$$-\frac{\partial(\rho v_x v_y)}{\partial x} - \frac{\partial(\rho v_y v_y)}{\partial y} - \frac{\partial(\rho v_z v_y)}{\partial z} - \frac{\partial p}{\partial y} + \left(\frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z}\right) + \rho g_y = \frac{\partial(\rho v_y)}{\partial t} \tag{11}$$

$$-\frac{\partial(\rho v_x v_z)}{\partial x} - \frac{\partial(\rho v_y v_z)}{\partial y} - \frac{\partial(\rho v_z v_z)}{\partial z} - \frac{\partial p}{\partial z} + \left(\frac{\partial \tau_{zx}}{\partial x} + \frac{\partial \tau_{zy}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z}\right) + \rho g_z = \frac{\partial(\rho v_z)}{\partial t} \tag{12}$$

All three equations can be written in a compact vector form as

$$-\nabla\cdot(\rho \overline{v}\overline{v}) - \nabla p + \nabla\cdot\overline{\overline{\tau}} + \rho \overline{g} = \frac{\partial(\rho \overline{v})}{\partial t} \tag{13}$$

The terms in equation (13) can be read from left to right as: (i) the net flux of momentum into the point by the flow of the fluid (note the divergence operator here); (ii) the source or sink of momentum per unit volume by the pressure gradient (a net force caused by a force imbalance);

(iii) the source or sink of momentum per unit volume due to viscous stress gradients (again a force imbalance); (iv) the source of momentum due to the gravitational force, all summed and balanced by; (v) the time rate of change of momentum at the point.

Equation (13) is sometimes more specifically called the *conservation of linear momentum* to distinguish it from the conservation of angular momentum equation. We will not discuss angular momentum here. The three equations contained in (13), one for each coordinate direction, are often the *equations of motion* and can be thought of as statements of Newton's second law.

Conservation of Energy

The conservation of energy principle states that energy is neither created nor destroyed. For a control volume, this implies that *the net rate at which energy is transferred into or out of the control volume by heat or work must balance the net rate of change of energy stored in the control volume*. Energy can be stored in three forms: (i) internal energy u, which is the energy stored by molecular motions and vibrations of the material in the control volume; (ii) kinetic energy, given by $mv^2/2$ for a body and by $v^2/2$ on a per unit mass basis, where v^2 is the magnitude of the velocity vector of the bulk fluid squared; and (iii) potential energy. Gravity can be treated in terms of potential energy or as work against a force. Potential energy is "stored" by work against a potential field ϕ, where the potential field is related to the acceleration of gravity by $\nabla\phi = \overline{g}$. Therefore, for the situation where z is pointed up and the acceleration of gravity is uniform and directed down, the gravitational potential is $\phi = gz$, where g is the magnitude of the gravity vector and z is measured from a specified plane in a direction opposite to the gravitational field. Instead of following the potential approach, we will develop the fundamental conservation of energy equation where we treat gravity in terms of work.

We again consider a fixed control volume $\Delta x\Delta y\Delta z$. The internal and kinetic energy inside the control volume at time t and $t+\Delta t$ is $\rho\left(u+1/2v^2\right)\big|_t \Delta x\Delta y\Delta z$ and $\rho\left(u+1/2v^2\right)\big|_{t+\Delta t} \Delta x\Delta y\Delta z$. Energy is transported into the control volume by the bulk fluid flow. Over the time interval Δt, the energy transport by the bulk fluid is given by the terms in the Table 3.

Next we determine the energy transfers by heat and work. Here we let the vector $\overline{q}_h$ be the heat-flux vector with units of energy per unit area per unit time. The net heat conduction into the control volume over the time interval Δt is then given by

$$(q_{hx}|_x - q_{hx}|_{x+\Delta x})\Delta y\Delta z\Delta t + (q_{hy}|_y - q_{hy}|_{y+\Delta y})\Delta x\Delta z\Delta t + (q_{hz}|_z - q_{hz}|_{z+\Delta z})\Delta x\Delta y\Delta t \quad (14)$$

We consider work by the bulk fluid by body forces such as gravity, and work done by the fluid at the surface of the control volume by pressure and viscous stress. Work by definition is a force through a distance in the direction of the force. Work *by* the fluid on the surroundings is negative (energy loss) and work *on* the fluid is positive (energy gain). The work by the x-component of gravity is the gravity force (which is the acceleration of gravity in the x-direction, g_x, times the mass of fluid $\rho\Delta x\Delta y\Delta z$) multiplied by the distance moved in the direction of the force ($v_x\Delta t$). Summing the work by all components gives:

$$\rho(v_x g_x + v_y g_y + v_z g_z)\Delta x\Delta y\Delta z\Delta t \quad (15)$$

The fluid gains energy (work on the fluid) when the velocity vector and gravity vectors are oriented in the same direction.

The energy added by pressure forces can be described in simple terms as a force multiplied by the velocity at the surface of the control volume. The force is the pressure times the area. Summing from all sides, the energy entering over the time interval Δt from pressure work is

$$\{(pv_x)|_{x+\Delta x} - (pv_x)|_x\}\Delta y\Delta z\Delta t + \{(pv_y)|_{y+\Delta y} - (pv_y)|_y\}\Delta x\Delta z\Delta t + \{(pv_z)|_{z+\Delta z} - (pv_z)|_z\}\Delta x\Delta y\Delta t \quad (16)$$

We will examine the meaning of the pressure work term in more detail later.

Viscous-stress work takes the form of a force (difference in stress × area) times distance in the force direction (velocity × time interval). For example, the net work from normal stresses on the x faces is ($[\tau_{xx}|_{x+\Delta x} -$

Table 3 Conservation of energy flux terms

Face	Inlet energy flux	Outlet energy flux
x	$\rho v_x\left(u+\frac{1}{2}v^2\right)\Big\|_x \Delta y\Delta z\Delta t$	$\rho v_x\left(u+\frac{1}{2}v^2\right)\Big\|_{x+\Delta x} \Delta y\Delta z\Delta t$
y	$\rho v_y\left(u+\frac{1}{2}v^2\right)\Big\|_y \Delta x\Delta z\Delta t$	$\rho v_y\left(u+\frac{1}{2}v^2\right)\Big\|_{y+\Delta y} \Delta x\Delta z\Delta t$
z	$\rho v_z\left(u+\frac{1}{2}v^2\right)\Big\|_z \Delta x\Delta y\Delta t$	$\rho v_z\left(u+\frac{1}{2}v^2\right)\Big\|_{z+\Delta z} \Delta x\Delta y\Delta t$

$\tau_{xx}|_x]\Delta y\Delta z)(v_x\Delta t)$. We must account for both normal and shear stress components. Summing the stress-related work on all faces, and rearranging, we have

$$\begin{aligned}
&\{(\tau_{xx}v_x + \tau_{xy}v_y + \tau_{xz}v_z)|_{x+\Delta x} \\
&\quad - (\tau_{xx}v_x + \tau_{xy}v_y + \tau_{xz}v_z)|_x\}\Delta y\Delta z\Delta t \\
&\quad + \{(\tau_{yx}v_x + \tau_{yy}v_y + \tau_{yz}v_z)|_{y+\Delta y} \\
&\quad - (\tau_{yx}v_x + \tau_{yy}v_y + \tau_{yz}v_z)|_y\}\Delta x\Delta z\Delta t \\
&\quad + \{(\tau_{zx}v_x + \tau_{zy}v_y + \tau_{zz}v_z)|_{z+\Delta z} \\
&\quad - (\tau_{zx}v_x + \tau_{zy}v_y + \tau_{zz}v_z)|_z\}\Delta x\Delta y\Delta t \qquad (17)
\end{aligned}$$

The physics of energy transfer by radiation is complex. We represent it by R, the radiation flux density per unit volume. The net energy lost by radiation over the time interval Δt is then

$$R\Delta x\Delta y\Delta z\Delta t \qquad (18)$$

The conservation of the energy equation is now obtained by summing the terms for the flux of internal energy across the boundaries of the control volume, the net transfer of energy by conduction, pressure and viscous work, and radiation and equating that with the difference in energy stored at times t and $t + \Delta t$. Following the familiar procedure, we divide by the volume and time interval to arrive at the conservation of energy equation with $e_t = u + 1/2v^2$,

$$\begin{aligned}
&-\frac{\partial(\rho v_x e_t)}{\partial x} - \frac{\partial(\rho v_y e_t)}{\partial y} - \frac{\partial(\rho v_z e_t)}{\partial z} \\
&+ \rho(v_x g_x + v_y g_y + v_z g_z) - \frac{\partial q_{hx}}{\partial x} - \frac{\partial q_{hy}}{\partial y} - \frac{\partial q_{hz}}{\partial z} \\
&- \frac{\partial(pv_x)}{\partial x} - \frac{\partial(pv_y)}{\partial y} - \frac{\partial(pv_z)}{\partial z} \\
&+ \frac{\partial}{\partial x}(\tau_{xx}v_x + \tau_{xy}v_y + \tau_{xz}v_z) \\
&+ \frac{\partial}{\partial y}(\tau_{yx}v_x + \tau_{yy}v_y + \tau_{yz}v_z) \\
&+ \frac{\partial}{\partial z}(\tau_{zx}v_x + \tau_{zy}v_y + \tau_{zz}v_z) - R = \frac{\partial(\rho e_t)}{\partial t} \qquad (19)
\end{aligned}$$

In vector notation,

$$\begin{aligned}
&-\nabla\cdot(\rho\overline{v}e_t) + \rho(\overline{v}\cdot\overline{g}) - \nabla\cdot\overline{q}_h - \nabla\cdot(p\overline{v}) + \nabla\cdot(\overline{\overline{\tau}}\cdot\overline{v}) - R \\
&\qquad = \frac{\partial(\rho e_t)}{\partial t} \qquad (20)
\end{aligned}$$

Although the equation appears impenetrable, it has a simple interpretation. Reading each term in the vector-notation equation (20) starting at the left: (i) the net rate of energy input by transport with flowing fluid; (ii) the rate of work done on the fluid per unit volume by gravitational forces; (iii) the rate of energy input per unit volume by heat conduction; (iv) the rate of work done on the fluid per unit volume by pressure forces; (v) the rate of work done on the fluid by viscous forces; (vi) the rate of energy received per unit volume by radiation; and (vii) the right-hand side which equals the net rate of energy gain per unit volume.

To better appreciate the pressure work term $\nabla\cdot(p\overline{v})$, we chain out the derivative to give

$$p(\nabla\cdot\overline{v}) + \overline{v}\cdot\nabla p \qquad (21)$$

The first term can be interpreted as a contribution to internal or thermal energy. Because $(\nabla\cdot\overline{v})$ is equal to the volumetric rate of deformation of a fluid, $p(\nabla\cdot\overline{v})$ can be interpreted as a force (pressure) times a deformation. The term thus represents the heating or cooling of the fluid because of compression or expansion. In the second term $\overline{v}\cdot\nabla p$, the pressure gradient can be interpreted as a force imbalance across a fluid element that is multiplied component-wise with the velocity, and thus is a contribution to the kinetic energy.

The conservation of energy is often further divided into a conservation of a mechanical energy equation and a conservation of a thermal energy equation. The mechanical energy equation is given by

$$-\nabla\cdot\left(\rho\overline{v}\frac{v^2}{2}\right) + \rho\overline{v}\cdot\overline{g} - \overline{v}\cdot\nabla p + \overline{v}\cdot(\nabla\cdot\overline{\overline{\tau}}) = \frac{\partial}{\partial t}\left(\rho\frac{v^2}{2}\right) \qquad (22)$$

The terms of the mechanical energy equation (22) represent: (i) the transport of kinetic energy into the volume by the bulk fluid flow; (ii) the work on the fluid by the gravitational force; (iii) a velocity times a force gradient, or pressure work; (iv) a velocity time a stress (force) gradient; equals (v) the rate of change of kinetic energy at the point.

The conservation of thermal energy is then given by subtracting the mechanical energy equation from the total energy equation to give

$$-\nabla\cdot(\rho\overline{v}u) - \nabla\cdot\overline{q}_h - p\nabla\cdot\overline{v} + \overline{\overline{\tau}} : \nabla\overline{v} - R = \frac{\partial(\rho u)}{\partial t} \qquad (23)$$

where $\overline{\overline{\tau}} : \nabla\overline{v}$ is notation for the complex expression

$$\begin{aligned}
\overline{\overline{\tau}} : \nabla\overline{v} = {} & \tau_{xx}\frac{\partial v_x}{\partial x} + \tau_{yx}\frac{\partial v_y}{\partial x} + \tau_{zx}\frac{\partial v_z}{\partial x} \\
& + \tau_{xy}\frac{\partial v_x}{\partial y} + \tau_{yy}\frac{\partial v_y}{\partial y} + \tau_{zy}\frac{\partial v_z}{\partial y} \\
& + \tau_{xz}\frac{\partial v_x}{\partial z} + \tau_{yz}\frac{\partial v_y}{\partial z} + \tau_{zz}\frac{\partial v_z}{\partial z}
\end{aligned}$$

The terms in the thermal energy equation (23) represent: (i) the transport of thermal energy into the control volume by the bulk fluid flow; (ii) the flux of heat into the volume by conduction; (iii) the heating or cooling of the fluid by volume expansion; (iv) the heating (always) of the fluid by viscous dissipation because of fluid–fluid friction; (v) the net transport of energy into the volume by radiation equals (vi) the rate of change of internal energy within the volume.

Summary of Conservation Equations

1. Conservation of Mass:

$$-\nabla\cdot(\rho\overline{v}) = \frac{\partial\rho}{\partial t} \tag{24}$$

2. Conservation of Linear Momentum (three equations):

$$-\nabla\cdot(\rho\overline{v}\overline{v}) - \nabla p + \nabla\cdot\overline{\overline{\tau}} + \rho\overline{g} = \frac{\partial(\rho\overline{v})}{\partial t} \tag{25}$$

3. Conservation of Total Energy:

$$-\nabla\cdot(\rho\overline{v}e_t) + \rho(\overline{v}\cdot\overline{g}) - \nabla\cdot\overline{q}_h - \nabla\cdot(p\overline{v}) + \nabla\cdot(\overline{\overline{\tau}}\cdot\overline{v}) - R = \frac{\partial(\rho e_t)}{\partial t} \tag{26}$$

   (a) Conservation of Mechanical Energy:

$$-\nabla\cdot\left(\rho\overline{v}\frac{v^2}{2}\right) + \overline{v}\cdot(\nabla\cdot\overline{\overline{\tau}}) - \overline{v}\cdot\nabla p + \rho\overline{v}\cdot\overline{g} = \frac{\partial}{\partial t}\left(\rho\frac{v^2}{2}\right) \tag{27}$$

   (b) Conservation of Thermal Energy:

$$-\nabla\cdot(\rho\overline{v}u) - p\nabla\cdot\overline{v} + \overline{\overline{\tau}} : \nabla\overline{v} - \nabla\cdot\overline{q}_h - (\varepsilon - \alpha) = \frac{\partial(\rho u)}{\partial t} \tag{28}$$

HYDROLOGY EXAMPLES

We conclude this article with several examples of these general conservation principles from hydrology. Most of the models used in hydrology are based upon one or more of the fundamental equations above, typically augmented by semiempirical expressions that describe specific processes. Here we focus upon the equations and their relationship to the fundamental conservation statements. We refer the reader to the other articles in this volume for a comprehensive development of the physical principles behind these hydrology examples. In the examples that follow, we begin with relatively straightforward applications of the conservations and progress to more complex and less direct applications.

Our first and simplest example is the advection–dispersion equation, which is a relatively transparent application of the conservation of mass principle, augmented with a so-called gradient-flux model, Fick's law, that describes the dispersion and diffusion of solute mass within the bulk flowing fluid. Our second example is the Navier–Stokes equations, which are mathematically complex and challenging, but are essentially just the conservation of momentum equations augmented with a model that describes the dissipation of momentum to heat caused by internal friction (viscosity) within the fluid.

The next suite of examples involves flow in porous media, which is described by more than one conservation principle applied simultaneously. First we develop the groundwater flow equation, which results when the conservation of momentum equations are substituted into the mass conservation equation. The conservation of momentum equations appears in the form of Darcy's law, which was only recognized as a conservation of momentum equation after its original formulation. We then consider the simultaneous flow of more than one fluid in porous media, which can be described with a coupled system of equations, sometimes called *the multiphase flow equations*. Often, in the air-water systems found in unsaturated porous media (e.g. soils), the multiphase flow equations can be simplified, flow of air ignored, and a single equation called *Richard's equation* can be used to describe the flow of water.

Our last example is from engineering hydraulics, the Saint Venant equations which are perhaps the most difficult to directly connect to the fundamental equations as developed above. The Saint Venant equations are gross simplifications of the general conservation statements, but prove to be quite practical and are widely used in application to flood routing and open-channel flow.

Advection–Dispersion Equation

An equation that arises frequently in hydrology is the advection–dispersion equation, which describes the transport of solutes such as a salt or dissolved species in a flowing fluid. It is applied in open fluids, such as streams, lakes, and rivers, as well as in porous media (*see* **Chapter 69, Solute Transport in Soil at the Core and Field Scale, Volume 2** and **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**).

To develop the advection–dispersion equation, we first define the solute concentration c as the mass of the solute per volume of solution. The mass of solute in a control volume (assuming that the solution fills the entire control volume) is then the concentration multiplied by the volume of the control volume, $c\Delta x\Delta y\Delta z$. The corresponding

expression from the earlier conservation of fluid mass development is $\rho\Delta x\Delta y\Delta z$ (cf. equation 3).

Solute is transported by the flow of the bulk fluid and by a combination of diffusion and dispersion relative to the bulk flow. The mass of solute transported by the bulk flow of water that moves at mean velocity v_x across the face perpendicular to the x-direction in time period Δt is $v_x c\Delta y\Delta z\Delta t$. Recall that the corresponding expression from the conservation of fluid mass development is $v_x\rho\Delta y\Delta z\Delta t$ (Table 1).

Mass transport relative to the bulk flow by diffusion and dispersion is often modeled using Fick's law, which for one-dimensional transport in the x-direction is given by $-D\partial c/\partial x$, with dimensions of $[M/L^2T]$ or mass crossing unit area in unit time, where D is the diffusion/dispersion coefficient with dimensions of $[L^2/T]$. The negative sign indicates that solute mass moves in the direction from high concentrations to low concentrations. In three dimensions, Fick's law is written using the diffusion/dispersion coefficient tensor $\overline{\overline{D}}$ and the solute concentration gradient vector ∇c as

$$-\overline{\overline{D}}\cdot\nabla c \tag{29}$$

The mass flux across the control volume face normal to the x-direction over time period Δt by diffusion and dispersion is then $-\left(D_{xx}\partial c/\partial x + D_{xy}\partial c/\partial y + D_{xz}\partial c/\partial z\right)\Delta y\Delta z\Delta t$, which simplifies for the case of a diagonal diffusion/dispersion coefficient tensor to $-D_{xx}\partial c/\partial x\Delta y\Delta z\Delta t$.

The net flux of mass across the control volume face perpendicular to the x-direction is then $\left(v_x c - D_{xx}\partial c/\partial x - D_{xy}\partial c/\partial y - D_{xz}\partial c/\partial z\right)\Delta y\Delta z\Delta t$. The conservation of solute mass is therefore written (cf. equation 24)

$$\nabla\cdot(\overline{v}c - \overline{\overline{D}}\cdot\nabla c) = \frac{\partial c}{\partial t} \tag{30}$$

This is called the *advection–dispersion equation*. This equation is often simplified for the case of incompressible fluid, where $\nabla\cdot\overline{v} = 0$. Using this fact, after we apply the chain rule to the first term in equation (30), we get $\nabla\cdot(\overline{v}c) = (\nabla\cdot\overline{v})c + \overline{v}\cdot\nabla c = \overline{v}\cdot\nabla c$ and the advection–dispersion equation becomes

$$\overline{v}\nabla c - \nabla\cdot(\overline{\overline{D}}\cdot\nabla c) = \frac{\partial c}{\partial t} \tag{31}$$

We can also modify this equation to take into account other sources of solute. For example, if a solute is dissolving from a solid mineral into the solution at a rate W, which has dimensions of mass of solute entering the solution, per volume of solution per unit time $[M/L^3T]$, then the advection–dispersion equation becomes

$$\overline{v}\nabla c - \nabla\cdot(\overline{\overline{D}}\cdot\nabla c) + W = \frac{\partial c}{\partial t} \tag{32}$$

If W is positive, then solute is dissolving into solution.

Navier–Stokes Equations

The Navier–Stokes equations describe the motions of a class of fluids that have the same mechanical behavior (response to stress). The Navier–Stokes equations are fundamentally the conservation of momentum equations (equations 25). To describe the motions of a fluid with the conservation of momentum equations (25), we must first provide a theory for the viscous behavior of the fluid. That is, we need to provide a model for $(\nabla\cdot\overline{\overline{\tau}})$ in terms of primary dependent variables and perhaps additional parameters. Such a theory is called a *constitutive theory* or a *closure theory* and the model for these terms is called a *constitutive model* or *constitutive equation*. In the case of fluids, one of the most common models for viscous behavior is that of a *Newtonian fluid*. When the Newtonian model for viscous stress is used in the conservation of momentum equations, the resulting equations are called *the Navier–Stokes equations*.

According to the Newtonian model of viscosity, the viscous stress is linearly proportional to the strain rate. Specifically, in Cartesian coordinates,

$$\tau_{xx} = 2\mu\frac{\partial v_x}{\partial x} - \frac{2}{3}\mu(\nabla\cdot\overline{v}) \tag{33}$$

$$\tau_{yy} = 2\mu\frac{\partial v_y}{\partial y} - \frac{2}{3}\mu(\nabla\cdot\overline{v}) \tag{34}$$

$$\tau_{zz} = 2\mu\frac{\partial v_z}{\partial z} - \frac{2}{3}\mu(\nabla\cdot\overline{v}) \tag{35}$$

$$\tau_{xy} = \tau_{yx} = \mu\left(\frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x}\right) \tag{36}$$

$$\tau_{xz} = \tau_{zx} = \mu\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x}\right) \tag{37}$$

$$\tau_{zy} = \tau_{yz} = \mu\left(\frac{\partial v_z}{\partial y} + \frac{\partial v_y}{\partial z}\right) \tag{38}$$

where μ is the *viscosity* of the fluid. The terms for the strain rate can be understood intuitively. For example, if $\partial v_x/\partial x$ is not zero, then the fluid is stretching or compressing in the x-direction. Terms such as $\partial v_x/\partial y$ describe shearing motion. For example, if $\partial v_x/\partial y$ is greater than zero, then the x-component of velocity increases as the y coordinate value increases. The difference in x velocity of fluid at two points separated from each other in the y-direction leads to shearing, and this shearing motion is the principal source of internal friction in fluids.

To derive the Navier–Stokes equation in their traditional form, the conservation of momentum equations are first rearranged. Consider the x-component of equation (10). Using the continuity equation (5), the x-component of the

conservation of momentum equation may be written

$$-\rho\left(v_x\frac{\partial v_x}{\partial x}+v_y\frac{\partial v_x}{\partial y}+v_z\frac{\partial v_x}{\partial z}\right)-\frac{\partial p}{\partial x}+\left(\frac{\partial \tau_{xx}}{\partial x}+\frac{\partial \tau_{xy}}{\partial y}+\frac{\partial \tau_{xz}}{\partial z}\right)+\rho g_x=\rho\frac{\partial v_x}{\partial t} \quad (39)$$

Now, we substitute in the Newtonian model, equations (33)–(38) for the viscous stress tensor $\overline{\overline{\tau}}$ to get, (for the x-component)

$$-\rho\left(v_x\frac{\partial v_x}{\partial x}+v_y\frac{\partial v_x}{\partial y}+v_z\frac{\partial v_x}{\partial z}\right)-\frac{\partial p}{\partial x}+\mu\left(\frac{\partial}{\partial x}\left(2\frac{\partial v_x}{\partial x}-\frac{2}{3}(\nabla\cdot\overline{v})\right)+\frac{\partial}{\partial y}\left(\frac{\partial v_x}{\partial y}+\frac{\partial v_y}{\partial x}\right)+\frac{\partial}{\partial z}\left(\frac{\partial v_x}{\partial z}+\frac{\partial v_z}{\partial x}\right)\right)+\rho g_x=\rho\frac{\partial v_x}{\partial t} \quad (40)$$

Similarly, for the y and z coordinate directions,

$$-\rho\left(v_x\frac{\partial v_y}{\partial x}+v_y\frac{\partial v_y}{\partial y}+v_z\frac{\partial v_y}{\partial z}\right)-\frac{\partial p}{\partial y}+\mu\left(\frac{\partial}{\partial x}\left(\frac{\partial v_x}{\partial y}+\frac{\partial v_y}{\partial x}\right)+\frac{\partial}{\partial y}\left(2\frac{\partial v_y}{\partial y}-\frac{2}{3}(\nabla\cdot\overline{v})\right)+\frac{\partial}{\partial z}\left(\frac{\partial v_y}{\partial z}+\frac{\partial v_z}{\partial y}\right)\right)+\rho g_y=\rho\frac{\partial v_y}{\partial t} \quad (41)$$

$$-\rho\left(v_x\frac{\partial v_z}{\partial x}+v_y\frac{\partial v_z}{\partial y}+v_z\frac{\partial v_z}{\partial z}\right)-\frac{\partial p}{\partial z}+\mu\left(\frac{\partial}{\partial x}\left(\frac{\partial v_x}{\partial z}+\frac{\partial v_z}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{\partial v_z}{\partial y}+\frac{\partial v_y}{\partial z}\right)+\frac{\partial}{\partial z}\left(2\frac{\partial v_z}{\partial z}-\frac{2}{3}(\nabla\cdot\overline{v})\right)\right)+\rho g_z=\rho\frac{\partial v_z}{\partial t} \quad (42)$$

There are four primary unknowns in these equations: the three velocity components and the fluid pressure. A closed system of equations is formed with the three Navier–Stokes equations (40), (41), and (42), the continuity equation (6), and the relationships between pressure and density and viscosity and density. These equations are extremely challenging to solve, even numerically, and are rarely tackled in the form given above. Typically, a simplified set of equations is used.

If the fluid can be assumed to be incompressible, and viscosity does not change, then the continuity equation (7) applies and the more common form of the Navier–Stokes equations results, here in vector form,

$$-\rho\overline{v}\cdot\nabla\overline{v}-\nabla p+\mu\nabla^2\overline{v}+\rho\overline{g}=\rho\frac{\partial\overline{v}}{\partial t} \quad (43)$$

The Navier–Stokes equations are thus the conservation of momentum equations for the case of a Newtonian fluid. The Navier–Stokes equations make a surreptitious appearance in our next suite of examples, which involve flow in porous media.

Flow In Porous Media

Flow in porous media that is fully saturated with water is described by the groundwater flow equation. In this section, we derive the groundwater flow equation by combining the conservation of mass equation with Darcy's law. Darcy's law is an empirical law that relates the flux of water to the forces that drive water. We then develop the equations that describe the flow of both air and water in porous media (the multiphase flow equations), and finally simplify these equations to arrive at what is called *Richards equation*, a model of the flow of water in partially saturated porous media (see Part 6: Soils and Part 13: Groundwater).

Groundwater Flow Equation

To write equations describing flow in porous media in differential form requires that we first define continuous porous media properties, such as porosity, that are valid at a mathematical point and vary smoothly in space. Because pores and solid grains in aquifers can be rather large (compared to molecules in a fluid), it is not clear how to define the porosity or other porous media properties at a mathematical point. This issue has received much attention in the literature (see Baveye and Sposito, 1984) so we will only outline the basic steps involved.

Continuous porous media properties are defined in terms of volume averages centered at a specific point. See Figure 7 for a cartoon of the idea. Consider V, a finite-sized spherical averaging volume that contains porous media and is centered at point (x, y, z). The volume contains solids of volume V_s and voids of volume V_v, where $V = V_s + V_v$. The porosity n at the point (x, y, z) is defined as the ratio of the volume of voids in the volume, to the total volume,

$$n(x, y, z) = \frac{V_v}{V} \quad (44)$$

If the center of the averaging volume is moved infinitesimally to a nearby point $(x + \Delta x, y, z)$, then we can define the porosity at the new point in the same way. Moving the averaging volume throughout the domain allows us to define the porosity at all points in the area under investigation. Because it is defined using a finite-sized averaging volume, porosity is a scale-dependent quantity, which means that the numerical value of the porosity in most natural materials depends upon the averaging volume size. Note that we never shrink the averaging volume to a point – it must be finite in size to produce continuous properties.

To develop the groundwater flow equation, we first write the conservation of mass equation for porous media.

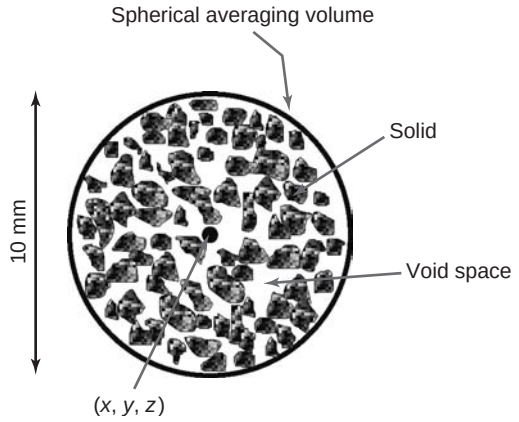


Figure 7 Most variables that describe porous media properties are defined with respect to a finite-sized averaging volume. For example, the value of the porosity at a point (x, y, z) is defined as the ratio of the volume of voids in the averaging volume to the total volume of the averaging volume. When properties are defined in this way, they are continuous at the mathematical point scale, such that derivatives of them are well defined. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Consider the same control volume that we used to develop the conservation of mass equation (Figure 2). In our earlier development, the volume was filled with fluid only; now it is filled with both solids and fluid. Here we are concerned with the conservation of fluid mass in the volume. If the porous media is fully saturated with water, then the volume of water in the control volume is equal to the volume of voids, $V_w = V_v = n\Delta x\Delta y\Delta z$ and the mass of water in the sample is $\rho n\Delta x\Delta y\Delta z$.

In porous media, water flow is quantified by the specific discharge vector $\overline{q}$. Consider an averaging volume in Figure 8, and the cross-sectional area A_x perpendicular to the x-direction. If the volume of water that crosses the area A_x in a time interval Δt is Q_x, then we define the x-component of the specific discharge vector, q_x as

$$q_x = \frac{Q_x}{A_x \Delta t} \tag{45}$$

that has dimensions of $[L/T]$. The other components of the specific discharge vector are defined in a similar manner. Although the specific discharge has the same dimensions as flow velocity, it is a volumetric flux. The mass flux into a control volume across the face perpendicular to the x-direction is then $\rho q_x \Delta y \Delta z$. Accordingly, the conservation of mass equation in porous media becomes (cf. equation 5)

$$-\frac{\partial(\rho q_x)}{\partial x} - \frac{\partial(\rho q_y)}{\partial y} - \frac{\partial(\rho q_z)}{\partial z} = \frac{\partial(\rho n)}{\partial t} \tag{46}$$

The specific discharge is modeled in porous media with Darcy's law. Darcy proposed his famous law in 1856 on the basis of an experiment consisting of a sand packed column with two water reservoirs on either end maintained at a constant elevation, shown schematically in Figure 9. Darcy found that the specific discharge through the column depended inversely upon the length of the column and linearly upon the difference in vertical elevation between the water levels in the two reservoirs, with water flowing from the reservoir with higher water elevation to that with lower, or $q_x \propto -\mathrm{d}h/\mathrm{d}x$ for x aligned with the axis of the column, where the minus sign indicates that the

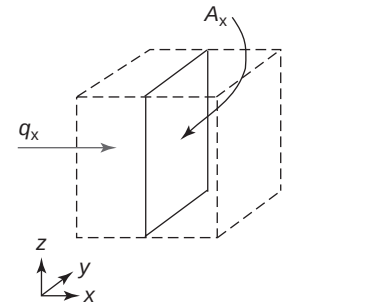


Figure 8 Each component of the specific discharge vector $\overline{q}$ is defined using a finite-sized averaging area oriented perpendicular to the component to be defined as the ratio of the volume of water crossing the area in a time interval divided by the size of the averaging area and the length of the time interval. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

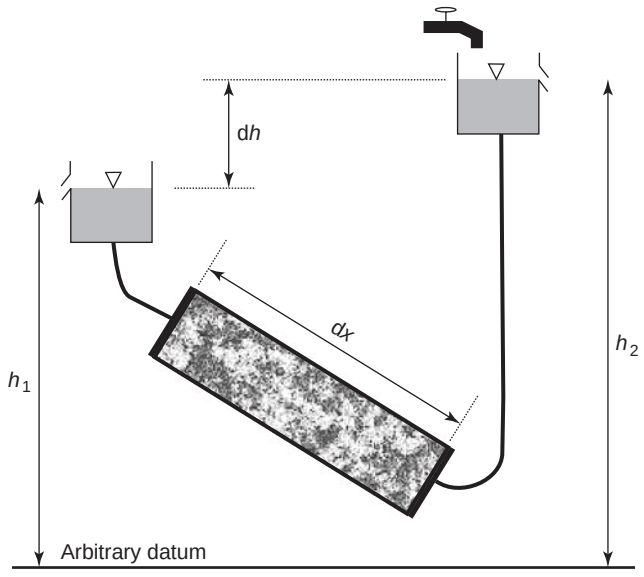


Figure 9 A cartoon of Darcy's column experiment. The column is filled with porous media and connected to two reservoirs in which the water levels are maintained at a constant elevation. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

direction of flow is opposite to the direction in which h is increasing. The quantity h is called *the hydraulic head*, and $\mathrm{d}h/\mathrm{d}x$ is the x-component of the hydraulic head gradient. Hydraulic head is a measure of the mechanical energy of the fluid and is only properly defined for irrotational flows, which in practice means, for fluids with constant density or with density that is a function of fluid pressure alone. The value of the constant of proportionality in the relation between flux and the head gradient depends upon the porous media (high for sand, low for silt) and the fluid (low for a relatively viscous fluid, high for a relatively inviscid fluid). In one dimension, for the flow of water, $q_x = -K\,\mathrm{d}h/\mathrm{d}x$ where K is the hydraulic conductivity. In natural porous media, stratification of sediments can result in a directionally dependent hydraulic conductivity. Darcy's law is then written in terms of the hydraulic conductivity tensor as

$$\begin{Bmatrix} q_x \\ q_y \\ q_z \end{Bmatrix} = -\begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \begin{Bmatrix} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial y} \\ \frac{\partial h}{\partial z} \end{Bmatrix} \tag{47}$$

or in vector notation, $\overline{q} = -\overline{\overline{K}}\cdot\nabla h$, where

$$\overline{\overline{K}} = \begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix}$$

is the hydraulic conductivity tensor. Darcy's law works well for most situations encountered in the field, although it loses accuracy whenever inertial forces become important, for example, when head gradients or pore sizes are very large (Bear, 1972). A more general Darcy's law expression applies to fluids other than water and to situations where hydraulic head is not properly defined,

$$\overline{q} = -\frac{\overline{\overline{k}}}{\mu}\cdot(\nabla p - \rho\overline{g}) \tag{48}$$

where $\overline{\overline{k}}$ is the permeability tensor with dimensions $[L^2]$, and μ is the dynamic fluid viscosity, with dimensions $[M/LT]$. The permeability is related to the hydraulic conductivity by $K = k\rho g/\mu$.

Darcy's law can be conceived of as a form of the conservation of momentum equations. Consider equation (48), rewritten as follows

$$\overline{\overline{k}}^{-1}\,\mu\overline{q} = -(\nabla p - \rho\overline{g}) \tag{49}$$

where $\overline{\overline{k}}^{-1}$ is the inverse of the permeability tensor. The term on the left side of equation (49) quantifies the loss of momentum through friction with the pore walls, whereas the terms on the right are forces that drive the fluid. Although it was originally proposed on the basis of empirical observation, Darcy's law has subsequently been justified from the Navier–Stokes equations by considering slow flow (called *creeping flow*) in pores (Bear, 1972).

To develop the equation describing the flow of water in saturated porous media, we begin with the conservation of mass equation (equation 46) and substitute for the specific discharge with Darcy's law (equation 47),

$$\nabla\cdot(\rho\overline{\overline{K}}\cdot\nabla h) = \frac{\partial(\rho n)}{\partial t} \tag{50}$$

We state without derivation that when the porous media is fully saturated with water, and a hydraulic head formulation is valid, the storage term $\partial(\rho n)/\partial t$ can be written in terms of hydraulic head as $\partial(\rho n)/\partial t = \rho S_s \partial h/\partial t$ (Freeze and Cherry, 1979), where S_s is the specific storage coefficient, which characterizes how the aquifer and pore fluids compress in response to changes in stress. If spatial gradients in density are assumed sufficiently small (a good assumption in fresh waters), then $\nabla\cdot(\rho\overline{\overline{K}}\cdot\nabla h) = \rho\nabla\cdot(\overline{\overline{K}}\cdot\nabla h)$, and the groundwater flow equation results,

$$\nabla\cdot(\overline{\overline{K}}\cdot\nabla h) = S_s\frac{\partial h}{\partial t} \tag{51}$$

Because we have factored out density, this equation can be viewed as a conservation of volume equation, combined with the conservation of momentum equations in the form of Darcy's law. The term on the left is the volume of water entering or leaving a point, per unit volume of porous media per unit time (flux divergence), and the term on the right side is the change in the volume of water at a point per unit volume of porous media per unit time. With the appropriate boundary conditions, this equation can be solved in a problem domain to determine the hydraulic head at every point.

Multiphase Flow And Richards Equation

If more than one fluid is present in the porous media, we can generalize the single flow equation to the multiphase flow equations. The approach is conceptually straight forward: we write distinct mass conservation and Darcy's law statements for each fluid, as well as a statement that says the total fluid volume cannot exceed the porosity. The end result is a coupled system of equations that describe the flow of each fluid. We develop the equations for a two-phase system consisting of air and water. We show how the coupled system of equations from the multiphase formulation can be simplified into Richards equation if we assume that the air phase is infinitely mobile.

We note at this point that there are many possible ways to formulate the multiphase flow equations and Richards equation. Our presentation here should not be considered a comprehensive development of the theory, but as an example of the application of the fundamental equations.

In multiphase flow, each individual fluid is a distinct phase, meaning that it has unique physical properties and is separated from other phases by an interface. A pressure difference caused by interfacial tension exists across the interface between each pair of fluids (e.g. air and water) and is called a *capillary pressure* p_c,

$$p_c = p_a - p_w \tag{52}$$

where p_a, p_w are the air and water fluid pressures. For a single interface as in an individual pore, the capillary pressure is given by $p_c = 2\sigma_{aw}/r$ where σ_{aw} is the interfacial tension between these fluids and r is the radius of curvature of the interface between the fluids (Bear, 1972).

Following equation (46), the conservation of mass equations for air and water are

$$-\frac{\partial}{\partial x}(\rho_a q_{ax}) - \frac{\partial}{\partial y}(\rho_a q_{ay}) - \frac{\partial}{\partial z}(\rho_a q_{az}) = \frac{\partial(\rho_a \theta_a)}{\partial t} \tag{53}$$

$$-\frac{\partial}{\partial x}(\rho_w q_{wx}) - \frac{\partial}{\partial y}(\rho_w q_{wy}) - \frac{\partial}{\partial z}(\rho_w q_{wz}) = \frac{\partial(\rho_w \theta_w)}{\partial t} \tag{54}$$

where the subscripts a, w refer to the air and water phases, and θ_a and θ_w are fluid contents, analogous to porosity and defined for an averaging volume V as (e.g. for the air phase),

$$\theta_a = \frac{V_a}{V} \tag{55}$$

where V_a is the volume of air in the averaging volume. Each fluid phase thus requires an independent conservation of mass equation. The sum of all the fluid contents must sum to the porosity,

$$n = \theta_a + \theta_w \tag{56}$$

The fluid content is related to the capillary pressure by a nonunique and hysteretic (history-dependent) pressure-saturation characteristic relationship (Bear, 1972),

$$\theta_\alpha = \theta_\alpha(p_c),\ \alpha = a, w \tag{57}$$

This relationship also implies that the capillary pressure is a function of the fluid contents.

Each fluid phase is described by a distinct Darcy's law with its own permeability that depends upon fluid content $\overline{\overline{k}}(\theta_\alpha)$, $\alpha = a, w$ (Bear, 1972). Typically, the permeability felt by a fluid phase drops to essentially zero when the phase no longer fills a connected pathway through the pores. For multiphase flow, Darcy's law is written for each phase as,

$$\overline{q}_\alpha = -\frac{\overline{\overline{k}}(\theta_\alpha)}{\mu_\alpha}\cdot(\nabla p_\alpha - \rho_\alpha \overline{g}),\ \alpha = a, w \tag{58}$$

Substituting Darcy's law into the conservation of mass equations gives the following two equations for air and water (for the case where the off diagonal components of the permeability tensor are zero)

$$\begin{aligned}&\frac{\partial}{\partial x}\left[\rho_\alpha \frac{k_{xx}(\theta_\alpha)}{\mu_\alpha}\left(\frac{\partial p_\alpha}{\partial x} - \rho_\alpha g_x\right)\right] \\ &\quad + \frac{\partial}{\partial y}\left[\rho_\alpha \frac{k_{yy}(\theta_\alpha)}{\mu_\alpha}\left(\frac{\partial p_\alpha}{\partial y} - \rho_\alpha g_y\right)\right] \\ &\quad + \frac{\partial}{\partial z}\left[\rho_\alpha \frac{k_{zz}(\theta_\alpha)}{\mu_\alpha}\left(\frac{\partial p_\alpha}{\partial z} - \rho_\alpha g_z\right)\right] \\ &\quad = \frac{\partial(\rho_\alpha \theta_\alpha)}{\partial t},\ \alpha = a, w\end{aligned} \tag{59}$$

or, in vector notation,

$$\nabla\cdot\left(\rho_\alpha \frac{\overline{\overline{k}}(\theta_\alpha)}{\mu_\alpha}\cdot(\nabla p_\alpha - \rho_\alpha \overline{g})\right) = \frac{\partial(\rho_\alpha \theta_\alpha)}{\partial t},\ \alpha = a, w \tag{60}$$

For the case of two fluids, there are four unknowns: p_a, p_w, θ_a, θ_w. These can be determined by solving the four independent equations: (i) the two fluid flow equations (equations 60); (ii) the capillary pressure relationship (equation 52); and (iii) the constraint on the pore volume (equation 56). In addition, the permeability-fluid content relationships $\overline{\overline{k}}(\theta_\alpha)$ must be known *a priori*. These coupled equations are very challenging to solve because they are strongly nonlinear.

Richards equation results when we simplify the air–water multiphase flow equations. Richards equation is typically used to describe water flow in the unsaturated zone, the zone where water and air are present in the subsurface. First, we assume that air is infinitely mobile. Consequently, no pressure gradients develop in the air phase and the pressure in the air phase is a constant everywhere p_a. Under this assumption, the capillary pressure is essentially equivalent to the water pressure: $-p_w = p_c - p_a$ and the capillary pressure–fluid-content relationship (equation 57) simplifies to a single relationship that is a function of water pressure alone: $\theta_w = \theta_w(p_w)$.

Richards equation is conventionally written in terms of hydraulic head, which is related to the water pressure through

$$h = z + \psi \tag{61}$$

where z is the elevation of the point of measurement and $\psi = p_w/\rho_w g$ is the pressure head. Darcy's law for water in the unsaturated zone can be written using hydraulic head as

$$\overline{q}_w = -K(\theta_w)\cdot\nabla h \tag{62}$$

where $K(\theta_w)$ is the water-content dependent hydraulic conductivity. Substituting this into the conservation of mass equation for water, and assuming that the fluid density is constant, yields the so-called mixed form of Richards equation,

$$\nabla\cdot(\overline{\overline{K}}(\theta_w)\cdot\nabla h) = \frac{\partial\theta_w}{\partial t} \tag{63}$$

There are two unknowns in this equation: the hydraulic head h and the water content θ_w. These unknowns can be determined by solving Richards equation with a specified water-content pressure relationship $\theta_w = \theta_w(p_w)$ or, as a function of pressure head, $\theta_w = \theta_w(\psi)$.

Richards equation can also be written in terms of hydraulic head. To do this, we eliminate the water content θ_w from the mixed form, equation (63), by substituting the water-content pressure relationship $\theta_w(\psi)$ and then express water pressure in terms of pressure head to give

$$\nabla\cdot(\overline{\overline{K}}(\psi)\cdot\nabla h) = C(\psi)\frac{\partial\psi}{\partial t} \tag{64}$$

where $\partial\theta_w/\partial t = \mathrm{d}\theta_w/\,\mathrm{d}\psi\,\partial\psi/\partial t = C(\psi)\partial\psi/\partial t$ and $C(\psi) = \mathrm{d}\theta_w\,\mathrm{d}\psi$ is the specific moisture capacity. The specific moisture capacity plays the same role in the unsaturated zone as the specific storage coefficient does in the saturated zone and characterizes the incremental change in water content with an incremental change in pressure head. Richards equation is also nonlinear: to solve the equation one must know the value of the hydraulic conductivity and specific moisture capacity at each point, but these cannot be specified without knowing the pressure head.

Saint Venant Equations For Open-channel Flow

Open-channel flow is often modeled using a one-dimensional form of the conservation of mass and conservation of momentum (*see also* **Chapter 135, Open Channel Flow – Introduction, Volume 4**). For this application, the equations are called the *Saint Venant equations*. The Saint Venant equations are simplified versions of the conservation of mass and momentum, where the following assumptions are made (Chow *et al.*, 1988):

1. Flow is one-dimensional, depth and velocity vary only in the longitudinal direction of the channel.
2. Flow is gradually varying along the channel so that hydrostatic conditions exist and vertical flow is negligible.
3. The longitudinal axis of the channel is approximated by a straight line.
4. The bottom slope of the channel is small. The channel bed is fixed.
5. Resistance coefficients for steady, uniform turbulent flow are applicable.
6. The fluid is incompressible and density is constant.

The Saint Venant equations are developed for a control volume positioned across the channel, which has cross-sectional area A and width along the channel $\mathrm{d}x$ (Figure 10). The volume of the control is then $A\,\mathrm{d}x$. For such a control volume, the conservation of mass is written for the time increment from t to $t + \Delta t$:

$$\rho(Q|_x - Q|_{x+\Delta x})\Delta t + \rho q\ \ \mathrm{d}x\Delta t = \rho A\ \ \mathrm{d}x|_{t+\Delta t} - \rho A\,\mathrm{d}x|_t \tag{65}$$

where $Q[L^3/T]$ is the flow in the channel such that $V = QA$ and $q[L^3/LT]$ is the lateral inflow with units of flow per unit length of the channel. Dividing by the length of the control volume ($\mathrm{d}x$) and time increment Δt and density ρ, we arrive at a one-dimensional conservation of mass equation:

$$-\frac{\partial Q}{\partial x} + q = \frac{\partial A}{\partial t} \tag{66}$$

To describe the flow in the channel, we must also provide a conservation of momentum equation. To develop this equation, we must sum the forces on the control volume, account for the flux of momentum into the control volume by the flow of the fluid, and then equate these with the change in momentum in the control volume. The forces considered in the Saint Venant equations are: (i) gravity; (ii) friction along the bottom and sides of the channel; (iii) expansion forces produced by changes in the channel cross section; (iv) wind shear force; and (v) pressure forces. We present the mathematical expressions for these forces without derivation (see Chow *et al.*, 1988). The gravity force on the control volume is given by $\rho g A S_0 \mathrm{d}x$, where S_0 is the slope of the channel bottom. The friction force

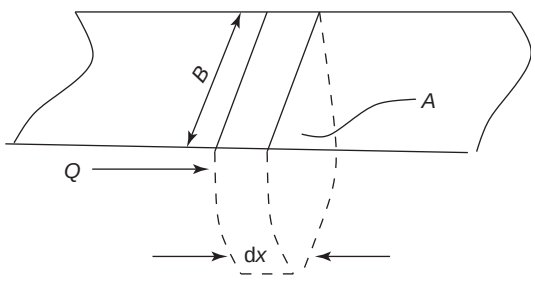


Figure 10 A schematic of the channel used to develop the Saint Venant equations

on the control volume is $-\rho g A S_f \mathrm{d}x$, where S_f is the so-called friction slope. The expansion/contraction forces are incurred when the channel contracts or expands suddenly. It is given by $-\rho g A S_e \mathrm{d}x$, where S_e is called *the eddy loss slope*. Wind shear on the free surface is given by $-\rho W_f B \mathrm{d}x$, where W_f is the wind shear factor and B is the width water surface perpendicular to the flow direction. The pressure force is given by $-\rho g A \partial y/\partial x \, \mathrm{d}x$, where y is the depth of water in the channel.

An impulse of momentum enters the control volume from upstream as $\rho\beta V Q \mathrm{d}t|_x$, where $\beta = 1/V^2 A \int\int v^2 \, \mathrm{d}A$ is the Boussinesq coefficient that accounts for nonuniform velocity at a cross section and v is the velocity through a small element $\mathrm{d}A$ of the cross section. Momentum also enters from lateral inflow into the channel with velocity v_x as $+\rho\beta v_x q \mathrm{d}x \mathrm{d}t$. The momentum outflow at the downstream boundary is $\rho\beta V Q \mathrm{d}t|_{x+\Delta x}$. The conservation of momentum for the control volume can thus be written as

$$\begin{aligned} &\rho[\beta V Q|_x - \beta V Q|_{x+\Delta x}] \, \mathrm{d}t + \rho\beta v_x q \, \mathrm{d}x \, \mathrm{d}t \\ &\quad + \rho g A S_0 \, \mathrm{d}x \, \mathrm{d}t - \rho g A S_f \, \mathrm{d}x \, \mathrm{d}t - \rho g A S_e \, \mathrm{d}x \, \mathrm{d}t \\ &\quad - \rho W_f B \, \mathrm{d}x \, \mathrm{d}t - \rho g A \frac{\partial y}{\partial x} \mathrm{d}x \, \mathrm{d}t \\ &= \rho V A \, \mathrm{d}x|_{t+\Delta t} - \rho V A \, \mathrm{d}x|_t \end{aligned} \tag{67}$$

Dividing by the length of the control volume $\mathrm{d}x$ and the time interval gives the Saint Venant conservation of momentum equation,

$$\begin{aligned} &-\frac{\partial(\rho\beta V Q)}{\partial x} + \rho\beta v_x q + \rho g A S_0 - \rho g A S_f \\ &- \rho g A S_e - \rho W_f B - \rho g A \frac{\partial y}{\partial x} = \frac{\partial(\rho Q)}{\partial t} \end{aligned} \tag{68}$$

FURTHER READING

Batchelor G.K. (1967) *Introduction to Fluid Dynamics*, Cambridge University Press: Cambridge, p. 615.

REFERENCES

Baveye P. and Sposito G. (1984) The operational significance of the continuum hypothesis in the theory of water movement through soils and aquifers. *Water Resources Research*, **20**, 521–530.

Bear J. (1972) *Dynamics of Fluids in Porous Media*, Originally published in 1972 by Elsevier: New York, republished by Dover: 1988.

Bird R.B., Stewart W.E. and Lightfoot E.N. (1960) *Transport Phenomena*, Wiley: New York, p. 780.

Chow V.T., Maidment D.R. and Mays L.W. (1988) *Applied Hydrology*, McGraw Hill: New York, p. 572.

Freeze R.A. and Cherry J.A. (1979) *Groundwater*, Prentice Hall, Engelwood Cliffs, NJ, 604 p.

Panton R.L. (1984) *Incompressible flow*, Wiley: New York, pp. 780.

6: Principles of Hydrological Measurements

ANDREW W WESTERN, RODGER B GRAYSON AND JUSTIN F COSTELLOE

CRC for Catchment Hydrology and Department of Civil and Environmental Engineering, The University of Melbourne, Victoria, Australia

This article highlights several key considerations in a successful measurement programme. These are:

- *The need to have a clear set of objectives or hypotheses to be tested*
- *An understanding of the temporal and spatial variability of the phenomena of interest.*
- *An understanding of the characteristics of the instruments available for measurement, including errors.*
- *Matching of the temporal and spatial characteristics of the measurement with those of the processes of interest.*
- *Considering the trade-offs inherent in sampling including the number of measurements and accuracy.*
- *Considering a range of practical issues related to setup in the field and the need to check and manage the data as it is collected.*

Successful monitoring programs are well designed, well resourced, and analysis will occur during the monitoring program. Timely analysis allows an ongoing review of performance and appropriate modification, which is critical to success.

New and emerging sensor technologies, as well as the ever-increasing availability of remote sensing information, bode well for the future of hydrological measurement. The synthesis of "smart measurements" with theoretical developments and modeling will yield the greatest advances in the coming years.

INTRODUCTION

There are at least three main reasons why we might want to undertake measurements of hydrological phenomena:

- improving the understanding of processes (e.g. a plot or small catchment study) involving hypothesis testing
- quantifying a resource (e.g. stream gauging)
- compliance (e.g. regular water quality measurement).

In all cases we are seeking data to provide some insight into the "truth" about a hydrological phenomenon, but the detail of information we wish to obtain from the data is quite different. In Case 1, we might want to characterize a particular process, determine the pathways of water to a stream, or define the heterogeneity of an aquifer. This information may be used to develop predictive models, test theories, or even develop new theories. Case 1 generally places the most rigor on the associated sampling design since it is likely that a number of measurements of different fluxes or states will be involved, each with different spatial and temporal characteristics. Similar considerations may be involved in Case 2, albeit simpler. Case 3 is often a prescriptive type of measurement where the basic measurement, frequency of sampling, and possibly the sample size are specified as part of the compliance requirement. Nevertheless, an understanding of the principles of measurement, sampling, and statistical analysis is required for proper design of sampling strategies for compliance monitoring. In this article, we focus on fundamental principles underlying measurements, with examples most relevant to Cases 1 and 2.

The most important question to answer before any measurements program can be designed is "What are the objectives of this data collection – that is what are the hypotheses that are being tested"? This sounds like a simple question but it is all too often ignored, or answered in only a cursory manner. A detailed answer will provide the basics

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

needed to design a sampling strategy including:

- phenomena to be measured
- key spatial and temporal scales of interest
- required accuracy and
- available resources.

This article is based on the assumption that a clear answer to the question has been provided, but such "*trite assumptions*" should *never* be made in a real sample design! Indeed, the objectives should be returned to many times, since, as discussed below, sampling design is an iterative process. Hydrological measurement is also generally an expensive activity involving a significant capital cost of equipment, personnel for design, construction and maintenance, and travel and infrastructure costs. Hence, there are strong incentives for the design of measurement systems to be well targeted to the questions at hand.

The article is presented in four main sections. In the first section, we summarize some common hydrological instruments and logging equipment and discuss recent advances. We then present some fundamental issues related to scale, followed by a discussion of considerations in planning a measurement system. Finally, we address key practical matters to be considered in the measurement of hydrological variables.

This article focuses on fundamental principles, providing examples to assist the explanation; but it is not a comprehensive review of measurement instruments or data logging equipment. There are a number of reference texts that provide excellent discussions of hydrological measurements and details on particular measurement methods and sampling protocols. A selection of these texts is included in the "References" section. In addition, the Internet provides access to information from the major manufacturers of instrumentation, most of whom can provide technical assistance and guidance for using their equipment.

INSTRUMENTS AND DATA LOGGING SYSTEMS

Instruments

A compilation of the most common hydrological measurement techniques is provided in Table 1. This list is by no means exhaustive but covers the techniques in general used for measurement of the most commonly required data related to the wide range of activities that are part of hydrological monitoring. The measurements have been divided into four broad categories; meteorological, surface water, porous media, and physico-chemical techniques.

The Table is organized so that the phenomenon being measured is identified in the first column, with the actual parameter that represents the phenomenon in the second column. The technique(s) used for each parameter are then identified in the third column and the typical time-step and resolution of the technique/instrument are noted, as well as its typical error margin. These error margins are generally manufacturer specified and typically represent performance under ideal conditions. Actual performance in the field often depends on installation, calibration, site characteristics, and maintenance. Finally, a brief comment is made about each technique. Where possible, a suitable reference that provides greater detail about the technique is noted.

The techniques range from time-honored methods that have been in use for decades to those that take advantage of the latest scientific advances in monitoring and recording of data. All are point measurements collected *in situ* unless otherwise stated. Techniques with extensive calibration and increased complexity are also identified.

Many of the techniques and sensors listed in Table 1 can be interfaced to a data logging system. Data loggers vary in the number of channels available, the input signal formats that can be measured, measurement precision and accuracy, measurement rates, storage (memory) capacities, memory volatility, robustness, programming language, and power requirements. The choice of a data logger is dictated in part by the number, variety, and characteristics of the sensor outputs to be used and by the conditions in which the logger is to be deployed. Some sensors have requirements for high sensitivity, such as net radiometers, which typically produce voltage output in the μV range and thus require analogue input channels with sufficient sensitivity and accuracy in this range. These are typically not available on the cheaper loggers. Counter channels on data loggers typically have lower and upper limits on input frequencies. Systems such as eddy correlation systems also typically require higher end loggers with relatively high measurement rates and processing speeds. Many data loggers can be interfaced with modems and telemetry systems. There is also an emerging suite of microsensors and processors with inbuilt low-power radio systems that can develop self-configured networks (sensor webs). Applications of such systems to hydrological monitoring are likely to develop over the next decade.

The measurement techniques described in Table 1 are mostly ground-based, physical measures. The burgeoning use of satellite-based remote sensing and chemical (especially isotopic) measurement techniques form a relatively new and exciting frontier for measuring hydrological processes (see Section "Advances in techniques"). For the sake of brevity, they have been omitted from Table 1. Remote sensing measurement techniques are more discussed in detail in Part 5 (**Chapter 46, Principles of Radiative Transfer, Volume 2** to **Chapter 61, Estimation of River Discharge, Volume 2**) and chemical/isotopic techniques

Table 1 Common hydrological measurement techniques and their characteristics

Measurement	Parameter	Instrument	Time-step	Typical resolution	Instrument accuracy	Comment
Meteorological						
Precipitation	Rainfall depth (mm) and intensity ($mm\,h^{-1}$)	Rain gauge	Minutes to daily	Sub mm to mm	5–15% error	Range from manually read daily totals, tipping-bucket mechanisms that measure intensity and universal gauges that measure both weight and outflow of precipitation (including frozen precipitation) from a pan. Errors occur due to wind field, siting characteristics, precipitation type, temperature. Heated tipping-bucket gauges can be used for measuring frozen precipitation but have problems with high-power requirements and evaporation and deflection of precipitation. Dingman (2002), Maidment (1992).
	Rain drop size, total precip. depth and intensity	Disdrometer, optical, and Doppler radar precipitation gauges	Continuous to minutes	Sub mm	1–5%	Disdrometers convert mechanical momentum of raindrop to an electrical impulse with amplitude proportional to rain drop diameter. Optical and Doppler radar gauges measure variations in scintillations of infrared beam or Doppler frequency shifts caused by precipitation and this is converted to rainfall intensity. Can differentiate between rain, snow, and mixed precipitation. Also can provide raindrop-size analysis (with 5% accuracy of diameter) and requires a drop size >0.5 mm for accurate measurement. Dingman (2002).
	Reflectivity	Radar	Sub min to min	Sub mm	2–3 $mm\,hr^{-1}$ or a factor of 2	Requires extensive calibration using a raingauge network to convert reflectivity to precipitation intensity. Error figures here are for hourly rainfall of moderate intensity with storm-by-storm calibration against a good raingauge network. Accuracy depends on rainfall type and intensity. A remote technique measuring a field and the only technique to provide two-dimensional coverage of rainfall events. Meischner (2004).
Frozen precipitation on ground	Pressure, depth, and water equivalent	Pressure and acoustic sensors, snow tube	Single to minutes	mm	4–10% or 2.5 mm for pressure sensors	The water equivalent pressure of snow (or other frozen precipitation) is measured using snow pillows and other pressure sensors. Acoustic, optical, and radioactive sensors can be used to measure depth of snow to ground surface or as a flux rate of snowfall (Lehning *et al.*, 2002). Manual method determines depth and weight of snow using aluminium tubes of known volume. Measured in sections as snow courses and requires estimation of snow density to convert to water equivalent depth. Dingman (2002).

(continued overleaf)

Table 1 *(continued)*

Measurement	Parameter	Instrument	Time-step	Typical resolution	Instrument accuracy	Comment
Mean wind sensors	Wind velocity and direction	Anemometers and Doppler sonars and lasers	Continuous	0.1–0.5 m s^{-1} over range of 0.3–75 m s^{-1} 1° direction	Vel <3% Dir: 1°	Anemometers can use rotating cup or propellers, differential pressure or thermoelectric (hot wire) and acoustic/sonic devices to measure wind velocity and direction. Anemometers with rotating cups/propellers or using differential pressure need to be in plane of wind for most accurate measurement and error increases in wind gusts or low velocity winds. Acoustic/sonic/Doppler sonar anemometers can measure three-dimensional wind movement. Kaimal & Finnigan (1994).
Mean air temperature	Degrees Celsius	Thermometer (electrical)	Continuous	<0.1 °C	0.05–0.5 °C	Platinum resistance, thermocouple, thermistor, and quartz thermometers use variations in resistance, voltage, or frequency with temperature. Quartz thermometers provide the highest accuracy but are mostly used in laboratory applications. Kaimal & Finnigan (1994).
		Infrared thermometer	Continuous	<1 °C	<0.1 °C	Remotely senses temperature using infrared emissivity.
Humidity	Absolute and relative humidity	Hygrometer, psychrometer, capacitance, radiation absorption	Continuous	<0.1 °C	<0.5 °C (hygrometer) 0.5–1.0 °C (others)	Absolute humidity is measured by dew-point hydrometers while relative humidity is measured by wet and dry bulb (psychrometric) temperature sensors or capacitance sensors. Humidity fluctuations can also be measured using ultraviolet or infrared radiation absorption sensors. Most methods require periodic cleaning and checks on calibration. Kaimal & Finnigan (1994).
Air pressure	Atmospheric air pressure (hPa)	Barometer	Continuous	0.1–1 hPa	0.3–1 hPa	Aneroid, mercury, and piezoresistive sensor barometers measure changes in air pressure. The piezoresistive barometer records and logs data digitally and is small, mobile, and relatively rugged relative to traditional and accurate aneroid and mercury barometers. The aneroid barometer can be attached to analogue recorders for logging.
Heat flux	Solar radiation and net radiation flux density	Net radiometer, Pyranometers, Pyrgeometer	>30 s Daily	10–35 W m^{-2}	10% daily total	Thermopile or solar cells measure incoming shortwave or long wave (infrared) radiation (solar radiation) or incoming and outgoing radiation (net radiation flux density). Measured radiation decreased by wind. Sensors can measure direct (tracking movement of sun), diffuse or total radiation. Kaimal & Finnigan (1994).

	Sensible heat flux	Soil heat flux plates	>4 min	50 μV W^{-1} m^{-2}		Thermopile measures temperature difference between two metal plates separated by material with known conductivity. Large errors (up to 50%) can occur in drying soils where soil evaporation is occurring beneath the soil surface. Kaimal & Finnigan (1994).
	Latent heat flux	Lysimeter	Continuous	1–100 g	10–100 g	Measures changes in weight of a block of soil and collects runoff and percolation. Can range in size from 1 m^2 to >10 m^2.
Cloudiness and sunshine hours	Amount of cloud	Visual or spectral radiometers, Campbell-Stokes recorder	Single or Continuous	10%	>10%	Cloudiness is commonly visually estimated but can be quantitatively estimated using radiometers (comparing difference between clear sky and cloudy skies) or analyzing digital photographs. Sunshine hours can be measured using a Campbell-Stokes glass dome that burns a trace on a card.
Actual evapo-transpiration.	Bowen ratio	Wet and dry bulb or dew-point sensors at two heights	>30 min averages		<30 Wm^{-2} mean daily latent heat flux	Temperature and vapor pressure are measured at two heights to by wet and dry bulb (psychrometric) or by temperature and dew-point (hygrometric) sensors to determine the vertical energy transport defined by the ratio of sensible heat flux over the latent heat flux (the ''Bowen ratio''). Requires relatively large fetch of surface being measured. Drexler *et al.* (2004), Kaimal & Finnigan (1994).
	Eddy covariance	Triaxial sonic anemometer, humidity-temperature sensors	10–30 min periods		Generally within 10% of Bowen ratio method	Sonic anemometers measure turbulent eddy flux. The covariance between the vertical wind velocity component and the measured temperature and humidity is used to estimate the sensible and latent heat flux. Drexler *et al.* (2004), Kaimal & Finnigan (1994).
	Surface renewal	High-frequency temperature sensors, sonic anemometer (for calibration)	Continuous		30–50 Wm^{-2}	High-frequency temperature sensors measure the duration and amplitude of temperature ramplike variations. This is used to estimate the sensible heat flux and the latent heat flux is estimated as the residual of the heat balance equation. This method needs to be calibrated against a sonic anemometer to allow for unequal heating of air below the sensor height. Drexler *et al.* (2004).
	Water vapor mixing ratio	Raman LIDAR	Continuous	1.5 m spatial resolution	4%	A LIDAR system measures the backscattered radiation from a laser beam. The wavelength of the radiation and the amount of backscattering at that wavelength provides estimates of temperature, concentrations of atmospheric constituents (e.g. water vapor) and wind. The water vapor mixing profiles are used to estimate heat fluxes. Drexler *et al.* (2004).

(continued overleaf)

Table 1 *(continued)*

Measurement	Parameter	Instrument	Time-step	Typical resolution	Instrument accuracy	Comment
Potential evapo-transpiration.		Evaporation pans. Manual or water level logger	Daily	<1 mm	1 mm	Traditional method of measuring evaporation loss from large open pans. A "pan factor" needs to be applied to the measurements to convert to a potential evapotranspiration applicable to larger areas (including larger water bodies). Water loss can be measured manually or using a water level logger. Various models or formulae (e.g. Penman, Penman Monteith, Priestly-Taylor) can be used to estimate potential evapotranspiration from meteorological data.
Surface water						
Water level	Stage (m)	Mechanical, electronic, acoustic, pressure gauges	Subhourly	<1 mm	0.2–3.0 mm	Mechanical devices use a float in a stilling well that is attached by cables to a rotational encoder that records changes in water level. Commonly used in permanent gauging stations. Water level variations result in changes in capacitance along a sealed cable for electronic gauges or pressure variations in a pressure transducer, bubble gauges use changes of pressure induced in a mercury reservoir and acoustic/laser sounders measure changes in stage above a stable section and pressure. Herschy (1995), Boiten (2000).
Channel shape	Two-dimensional position (m)	Dumpy level or total station	Single	mm	mm	Dumpy level used in conjunction with tapes or chains to measure positions relative to a datum (horizontal) level. Total station is able to rotate theodolite in vertical plane when measuring positions relative to station location.
Discharge	Average velocity (m/s)	Current meter	Single	Depth >0.15 m Velocity >0.1 m s^{-1}	5–10% (25% in turbulent streams)	Velocities measured as a series of profiles in a cross section across channel. Gordon *et al.* (1992), Herschy (1995).
		Acoustic Doppler velocimeter	Single or Continuous	0.01–2.5 m s^{-1}	1% of measured range	Measures velocity and direction from Doppler shift of acoustic signals backscattered from particles in flow. Measured as a cross-section across channel or as a 3-D field around objects when measuring turbulent flow. Herschy (1995), McLelland & Nicholas (2000).
	Tracers	Dilution gauging	Single			A tracer (chemical, radioactive, fluorescent) is added at an upstream point and its diluted concentration is measured at a downstream point assuming full mixing. Used in turbulent mountain streams. Tracer can be added as a slug or at a constant rate. Gordon *et al.* (1992), Herschy (1995).

Porous media						
Bulk density and total porosity	Dry mass/volume (Mg/m^3)	Oven drying samples	Single	<mg	mg	A ''wet'' weight of the soil sample (core or solid clumps) is taken and then the sample is oven dried and a dry weight taken. The volume of the sample is taken by measurement (core) or by a water displacement method. Porosity is measured by dividing the bulk density by the particle density, the latter usually taken from the literature. McKenzie *et al.* (2002).
Soil water characteristic		Suction plates	Single			Relationship between volumetric water content (%) and matric potential (-m) of sample. Mostly measured in the laboratory using suction plates to determine points on curve for given matric potentials. McKenzie *et al.* (2002).
Matric potential		Heat dissipation or electrical resistance sensors	Single			Heat dissipation or electrical resistance of a porous matrix in equilibrium with the soil is measured and related to matric potential by previous calibrations. Allows matric potential to be measured in the field. Scanlon *et al.* (2002).
Saturated and unsaturated hydraulic conductivity (K_s)	K (mm/hr)	Twin ring and tension infiltrometer, well permeameter	Single	$<1\,mm\,hr^{-1}$	Coefficient of variation <0.5 with 20 readings	The steady-state rate of infiltration from water filled rings or discs (twin ring and tension infiltrometers) are used to measure soil hydraulic conductivity while a well permeameter uses maintains a constant head of water in an auger hole and the rate of outflow is measured from the steady-state inflow from the well permeameter and used to calculate K_s. Useful for measuring K_s at depths from 0.15–2.0 m below ground surface. Infiltrometers can also be used to measure the hydraulic conductivity of soil cores in the laboratory. McKenzie *et al.* (2002).
Hydraulic conductivity	K (mm/hr)	Rainfall simulator	Single	$<1\,mm\,hr^{-1}$		Rainfall simulator applies ''rain'' at a constant rate to a sealed, *in situ* plot. Runoff is measured by a calibrated container or tipping bucket and these measurements allow the calculation of hydraulic conductivity. Advantages of being applied to larger areas than other techniques and integrates soil heterogeneities.
	K (mm/hr)	Rising and falling head slug tests	Single	$mm\,hr^{-1}$		A slug test removes (rising head) or adds (falling head) a volume of water from the piezometer and monitors the time taken for the water level to recover to its predisturbance level to determine saturated hydraulic conductivity.

(continued overleaf)

Table 1 *(continued)*

Measurement	Parameter	Instrument	Time-step	Typical resolution	Instrument accuracy	Comment
	K (mm/hr)	Cone permeameters	Single	$mm\,hr^{-1}$		Cone permeameters inject water radially into the borehole at a constant rate and the pressure gradient away from the injection site is measured and used to estimate the permeability and hydraulic conductivity at that point. Allows measurements at a number of points within a borehole.
Piezometer	Hydrostatic groundwater level (m)	Manual or logger – see surface water	Single or sub hourly	<1 mm −10 mm	<1–10 mm	Depth to groundwater from a datum level is manually measured using tape or water level loggers (see surface water section) can be installed.
Soil moisture	% volumetric water content	Time domain (TDR) and Frequency-domain (FDR) reflectometers	Continuous	1%	3%	TDR measures the dielectric permittivity of soil by measuring the time it takes for an electromagnetic wave to propagate through the soil, while FDR does this by determining the resonant frequency with greatest amplitude. The volumetric water content is determined through calibration with the dielectric permittivity.
	% volumetric water content	Neutron probes	Single and logged			A decaying source emits fast neutrons into the soil. The neutrons are backscattered and slowed by the presence of water. A detector in the probe measures the amount of slowed neutrons and via calibration this provides a measure of the soil water content.
physico-chemical						
Conductivity	Siemens/cm	Resistance and induction conductivity probes	Continuous	0.1 μS	0.5%	Resistance between two electrodes or induced current and resulting magnetic field measured by a coil is used as a measure of water conductivity. Temperature usually measured in conjunction with conductivity. Induction probes have the advantage of not having exposed electrodes.
pH	pH scale	Potentiometric (electrode) pH meter	Continuous	0.01 unit	0.02–0.1 unit	Potentiometric electrodes monitor changes in voltage caused by changes in the activity of hydrogen ions in a solution to determine the pH of the solution.

Dissolved oxygen (DO)	Mg/L or % saturation	Polarographic DO meter	Continuous	0.01 mg L^{-1}	0.3 mg L^{-1} or 2%	Dissolved oxygen diffuses through a membrane and is then measured by a sensor cell. Membrane requires regular maintenance.
Turbidity	nephelometric turbidity units (NTU)	Photoelectric turbidimeters	Continuous	0.01–0.2 NTU		Light is emitted from the sensor and the amount of backscattered radiation is converted to a measure of turbidity in nephelometric turbidity units.
Water Clarity	Depth (cm)	Secchi disc	Single	0.5 cm	0.5 cm	A disc is lowered into the water and the depth that it disappears from sight
Water samplers	Suspended sediments, water quality	Sediment sampler – depth or point integrated, rising stage sampler	Single			Samples collected from a point in the stream section or lowered through the water column to collect a velocity-weighted sample through the section. Sample bottles can also be used with siphons or one-way valves to collect samples during the rising stage. Gordon *et al.* (1992).
Bed load	Bed load sediments	Pits and baskets, pressure-difference sampler	Single			Various pits and baskets installed into the stream bed are used to collect bed load samples. Pressure-difference samplers use a pressure drop at the exit ensuring that the entrance velocity is the same as the stream and that separation of bed load material occurs at the exit and is collected in a mesh bag.
Erosion and deposition		Pins and chains	Single			Pins and chains can be inserted into the bank and channel bed and rates of scour and fill can be quantitatively monitored by surveying changes in bed-bank position relative to the pins or chains.
		Pressure meters	Continuous – daily	Mg	mg	Pressure plates placed onto the stream bed or embedded into the sediments can measures changes in overlying pressure to estimate changes in sediment coverage.

Note: DO, dissolved oxygen; TDR, time domain reflectometers; FDR, Frequency domain reflectometers; NTU, nephelometric turbidity units.

are discussed in Part 10 (**Chapter 116, Isotope Hydrograph Separation of Runoff Sources, Volume 3**).

Advances in Techniques

In recent times, there have been some major advances in measurement techniques, and this continues to be an area of active development. The number of satellite-based remote sensing platforms has increased dramatically and they are more focused on environmental applications (*see* **Chapter 47, Sensor Principles and Remote Sensing Techniques, Volume 2** and Part 5 generally). Important components of the water balance such as soil moisture (e.g. via Time Domain Reflectometry-based devices) and evapotranspiration (e.g. via eddy covariance equipment) are now much cheaper and simpler to measure. Logging equipment is getting cheaper and smaller, and the cost of telemetry has dramatically reduced, particularly where cellular phone coverage is available or distances permit low-power radio-based communication networks. This has led to the development of "sensor webs" where large numbers of compact sensor/logger/telemetry units can be deployed with an automatic feed of information to Internet-connected computers, making real-time data available at relatively low cost. In some cases, sensors are considered "disposable" such as those dropped ahead of fire fronts to provide real-time information to fire fighters. This sort of technology is rapidly developing and is certain to find hydrological applications, where the limitation of a relatively small number of point measurements has constrained our ability to observe spatial patterns. Similarly, the ground truthing of satellite observations will be enhanced by more spatially dense sensor networks.

Streamflow measurement remained almost unchanged for 100 years with discharge measured using a propeller meter at points across a cross section. But this has been revolutionized with the advent of Acoustic Doppler Current meters and profilers, providing fast, accurate information on velocity fields over a wide range of depths. Sensors measuring important water quality parameters *in situ* are also developing to include nutrients and other measures of ecological significance, although the costs of these devices are still high.

Techniques involving tracers to assist in defining pathways and residence times of water, sediment, and other constituents in the environment have been available for many years but are gaining more widespread use in hydrology (e.g. Kendall and McDonnell, 1998). These methods are particularly useful because they provide *complementary* information to that obtained from more traditional methods. For example, a study into hill slope processes of water movement would commonly include wells and soil moisture equipment to measure time series of responses in the saturated and unsaturated zone, but also knowing how long water had been in the soil, or perhaps what pathway it had taken is likely to add a great deal to understanding of the hill slope behavior. In addition, the combined use of time series of discharge data and conservative tracers, such as chloride, can substantially decrease parameter value uncertainty during the calibration of catchment models (Kuczera and Mroczkowski, 1998). There are other examples of complementary measurements such as remote sensing and *in situ* sensors collecting time series. The former provides detailed spatial information but only a snapshot in time, whereas the latter is "dense" in time but only at a point. Together they can help build some detail in both time and space.

Many of the developments in measurement are aimed at providing more data at lower cost and this is generally achieved by making a "surrogate" measure that is related to the phenomena of interest and is quick to make. Calibration between the surrogate and the actual phenomena is therefore of critical importance with many of these new sensor technologies. Our experience is that "factory calibrations" should always be checked under the conditions in which the device is to be used, and rechecked at regular intervals to correct for any drift (see also later discussion).

We look forward to more developments in measurement technology. Hydrology would greatly benefit from nonintrusive measurements of shallow subsurface flow, preferred pathways, and hydraulic properties of soils that can be applied over large areas. Other methods to indicate the pathways of water such as isotopic and other tracers will increase in use as they provide new information of particular importance to environmental issues. Remote sensing will continue to improve, and the synthesis of different data sources into integrated products will produce information in which we will have much improved confidence. The development of new instrumentation enabling higher density of measurements in space and time or the ability to cover large areas will also continue. But to exploit all this new capability fully, and avoid getting "lost in a sea of data" we will still require a careful design of measurement programs based on some fundamental concepts.

FUNDAMENTAL ISSUES OF SCALE

Designing an efficient program for measurement of hydrological variables involves a conundrum. We make measurements to inform us about a process or resource, but to make those measurements efficiently, we need to know about that process or resource. Designing measurement programs is therefore an iterative activity and one that often draws on past experience from other places. We use measurements to tell us about the phenomena and as we learn more, we can use this knowledge to design a better sampling program. Similarly, we can set up a hypothesis, test it with some data, and then refine the hypothesis and measurement

program. We can often add modeling to this iterative loop to further enhance the information that we glean from the data collected. For example, if we know that a particular water quality parameter varies rapidly during a runoff event but only slowly during baseflow conditions, we can use this knowledge (a "model" if you like) to design a structured sampling approach that provides more information than the same number of samples obtained with a naive random or uniform sample spacing in time. Such a design can be statistically sound, thereby maximizing the value of the data. A variety of sophisticated random approaches exist that can incorporate existing knowledge as well as information on sampling costs to maximize the efficiency of a sampling effort.

Sampling theory is a well-developed field in itself and texts such as Thompson (2002) provide a range of useful sampling schemes. The following is based on Thompson (2002). They key advantages of a statistically sound sampling scheme are that:

- it allows statistical inference of the population characteristics with known confidence;
- it minimizes the likelihood of the inadvertent biasing of measurements associated with subjectively choosing "representative" sites;
- it provides a representative data set derived from sites that are objectively chosen; and
- well-designed sampling schemes can incorporate existing knowledge to maximize the efficiency (i.e. minimize uncertainty), minimize costs, and to be robust to errors in the existing knowledge.

Fundamentally, a sampling design consists of assigning probabilities of selection into the sample for all members of a population, then randomly selecting from the population. Simple random sampling assigns an equal probability of selection to every member of a population (say every point in a spatial field). More sophisticated approaches assign unequal probabilities based on existing information about a phenomena and/or about the cost of sampling in such a way as to maximize the value of the information gained from the measurements. In this case some points are more likely than others to be selected, for example, it might be worth allocating a higher probability of selection to high flows when sampling water quality for load estimation or using randomly selected transects for spatial sampling to minimize costs of moving between measurement points. The interpretation of the measurements then takes into account the probability of selection when inferences about the population characteristics (say the mean) are made.

Some useful sampling designs include stratified random sampling, cluster sampling, a variety of model-based sampling schemes including designs that utilize existing or more densely sampled auxiliary data (say topographic information in a regression), sampling in the presence of spatial (or temporal) correlation, and adaptive sampling, among others. Stratified random sampling allows a population to be stratified, say on the basis of geology, soil type, or elevation, and then samples are taken randomly within each stratum (group). Adaptive sampling uses information already collected during the measurement process to aid in selecting the next points in the sample. An example might be a sampling design that samples more intensively around a point when specific (usually rare) conditions are encountered, say a preferential recharge area of coarse sediment on a floodplain otherwise dominated by fine sediments. It is useful where clusters of rare events occur. This allows the sampling process to focus on a particular interest while still obtaining useful information about the population.

It is important to note that structured sampling approaches are always based on some form of model or understanding of a variable's behavior, and that if this turns out to be wrong then poor results can be obtained if such approaches are applied naively; however, there are designs that are robust to model errors. Of course, this is clearly also the case with choosing "representative" sites using some subjective approach. Another issue relating to sample selection that can arise is related to systematic sampling in the presence of a periodic process. If the sampling interval closely matches the period (or an integer multiple thereof), then biases in estimates of both mean properties and variability are likely to result. This is an issue with sun-synchronous remote sensing platforms, as they always observe at the same point in the diurnal cycle.

A prerequisite for useful hydrological measurements is that the temporal and spatial scales of the measurement appropriately match those of the phenomena of interest. For example, if we are interested in the total annual runoff from a small arid-zone ephemeral stream, a sensor that takes an instantaneous water level every week will not provide useful information. The sampling frequency (a week) is too large compared to the timescale of the process (probably minutes to hours in this example), whereas a week may be quite sufficient for measuring a slow-changing groundwater level. There is some basic theory that provides a framework for these scale considerations. The spatial (and temporal) dimensions of measurements can be characterized by three scales as depicted in Figure 1. These scales are the spacing, the extent, and the support, and have been termed the *scale triplet* by Blöschl and Sivapalan (1995).

The *spacing* refers to the distance (or time) between samples, the *extent* refers to the overall coverage of the data (in time or space), and the *support* refers to the averaging volume or area (or time) of the samples. All three components of the scale triplet are needed to uniquely specify the space and the time dimensions of measurements. For example, for TDR soil moisture samples in a research catchment, the scale triplet in space may have typical values of, say, 10 m spacing (between the samples), 200 m extent

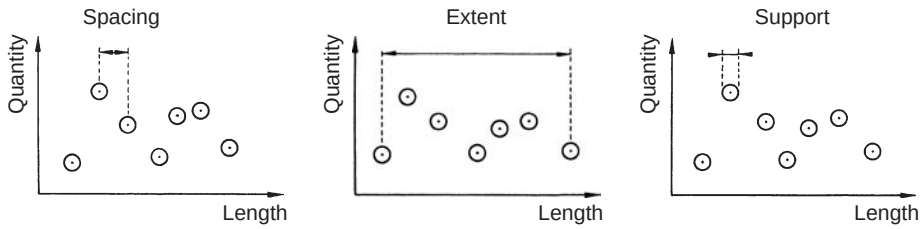


Figure 1 Definition of the scale triplet (spacing, extent, and support). After Blöschl and Sivapalan (1995)

(i.e. the length of the plot sampled), and 10 cm support (the diameter of the region of influence of a single TDR measurement). Similarly, for a remotely sensed image, the scale triplet in space may have typical values of, say, 30 m spacing (i.e. the pixel size), 10 km extent (i.e. the overall size of the image), and 40 m support (i.e. the "footprint" of the sensor). The footprint of the sensor is the area over which it integrates the information to record one pixel value. It is usually on the order of the pixel size but not identical to it. There are more complex cases such as measurements of evapotranspiration using eddy covariance equipment where the support is difficult to define and may vary in time due, for example, to different wind conditions. Similarly, the support for a groundwater measurement well may not be clearly defined. While the terms spacing, extent, and support are commonly used in spatial analysis, the analogous terms in time series analyses are sampling interval, length of record, and smoothing or averaging interval (e.g. Blackman and Tukey, 1958).

Ideally, measurements should be taken at a scale that is able to resolve all the variability (in both space and time) that influences the hydrological features in which we are interested. In general, because of logistical constraints, this will not be the case and so the measurements will not reflect the full natural variability. For example, if the *spacing* of the data is too large, the small-scale variability will not be captured and the measurements will appear "noisy" or discontinuous. If the *extent* of the data is too small, the large-scale variability will not be captured and will translate into a trend in the data. If the *support* is too large, most of the variability will be smoothed out. These examples are depicted schematically in Figures 2 and 3. Figure 2 is for the temporal domain, where the sine wave relates to the natural variability of some hydrological variable and the wavelength is related to the scale of the true hydrological features. The points in Figure 2(a) relate to the scale triplet of the measurements. Similar concepts apply to the spatial domain (Figure 3). We are often particularly limited with capturing spatial variability since many of our techniques are essentially point measurements. Measurements of soil hydraulic conductivity are good examples where the measurement support is limited – just a few cm^3 for laboratory measurements of soil cores to perhaps a few m^3 for small-scale permeameter tests to perhaps 10 000 m^3 or more for large-scale well pumping tests in sandy aquifers. A large number of tests would be required to fully characterize a particular location. Often in such instances, a statistical approach is taken where sampling is designed to define the statistical properties of the phenomena and the data are used in a stochastic rather than a deterministic manner.

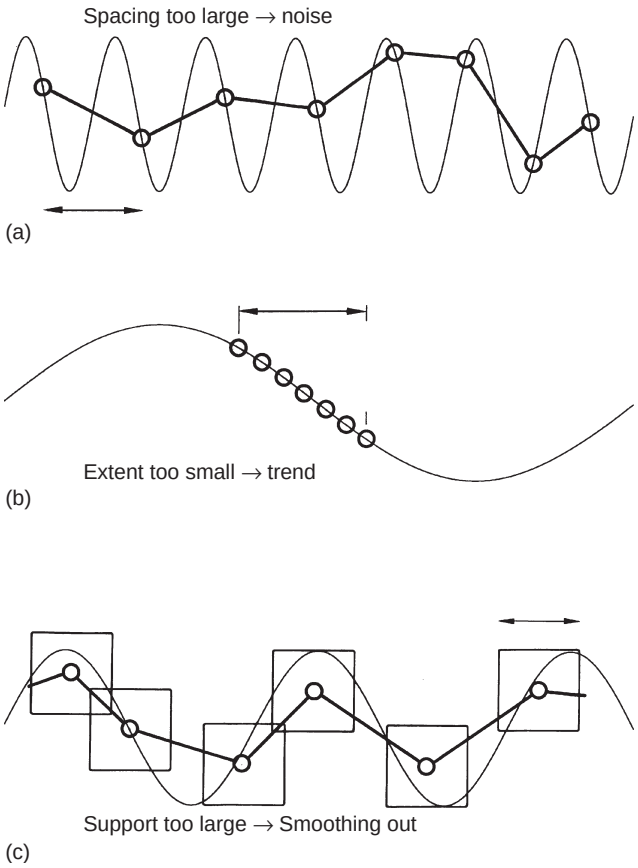


Figure 2 Pattern of a generic temporal phenomena showing the effects of too large a spacing (a); too small an extent (b) and too large a support (c). After Blöschl and Sivapalan (1995)

It is clear that measurement is a sort of filtering, that is, the true spatial and temporal patterns are filtered by the properties of the measurement, which are then reflected in

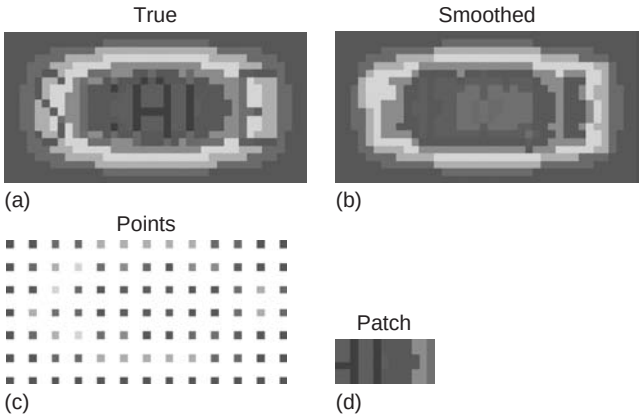


Figure 3 (a) shows the pattern of generic spatial phenomena; (b) shows sampling over the whole extent but with a large *support* thereby smoothing the small-scale variability; (c) shows sampling over the whole extent with a small *support* and medium *spacing* that captures the basics of the pattern but not the detail; and (d) shows sampling with too small an *extent* where we see a trend and some detail of a small section of the pattern. Redrawn from Western *et al.* (2002). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the data. The effect of this filtering is to introduce errors into the observed variability if the scale of the measurement does not match the scale of the process. There is a substantial body of literature that deals with methods for defining and predicting the way in which variability is captured (or not captured) by the measurement characteristics (e.g. Wiener, 1949; Krige, 1951; Matheron, 1965, 1973; Blackman and Tukey, 1958; Federico and Neuman, 1997). An important practical outcome of that work is the development of methods to (i) assess how many measurements are needed to capture (to a certain accuracy and under particular assumptions) a natural pattern and (ii) to quantify the bias in variability introduced by filtering (e.g. Journel and Huijbregts, 1978; Vanmarcke, 1983; *see* **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**).

PLANNING A MEASUREMENT SYSTEM

The hypotheses being tested, or objectives of the sampling exercise will define the processes (or states and fluxes) that need to be measured. But do the instruments available have the characteristics to measure what you really need and how do you match the type of measurement to the process of interest?

Matching of Measurement to Process

While often a quantitative treatment will not be needed, a qualitative consideration of the scale of the natural variability and that of the measurements is important to assess at least the magnitude of information on variability not resolved by the sampling. In Figure 4, the spatial spacing and extent of several different types of measurement devices are plotted versus their typical temporal spacing and extent. The shaded area refers to the domain between spacing and extent of the measurements. For daily rain gauges, the domain covers ranges, in time, from 1 day to, say, 100 years, and in space, from 10 km (average spacing of the gauges) to 2000 km (size of the region). Figure 4 also shows the typical scales of TDR measurements of soil moisture in research catchments as well as a number of space-borne sensors relevant to hydrology.

Figure 5 is a similar plot of spatial and temporal scales related to key hydrological processes (after Blöschl and Sivapalan, 1995). Figure 4 can be compared with Figure 5 and the parts of the space-time domain that overlap are those where we have measurement techniques that are appropriate for describing the process of interest, whereas areas that do not overlap are not described well. In other words, from a particular measurement one can only "see" processes within a limited window (determined by the scale triplet), and processes at larger and smaller scales will not be reflected in the data. For example, daily rain gauges cannot capture atmospheric dynamics at the 10-km scale as the temporal spacing is too large, but on the other hand the Meteosat satellite sensor is commensurate with atmospheric processes from thunderstorms to fronts and one would expect it to capture these processes with little bias as a result of scale incompatibility. The comparison also indicates that TDR measurements can potentially capture runoff generation processes in a small research catchment setting. Figures 4 and 5 are used here just as examples, but the concept of matching measurements with processes in space and time is critical to efficient measurement design.

Accuracy and Error

The preceding discussion highlighted that poor sampling design can introduce considerable error or uncertainty but that these can be minimized by considering the spatial and temporal characteristics of the phenomena of interest to ensure that measurements are representative of the variability. Once the preferred temporal and spatial characteristics of a measurement are defined, consideration must be given to the accuracy of the measurement itself. Measurements can be a direct measure of a hydrological variable (such as the stage of a stream, rainfall depth, or snow water equivalent measured by weighing a snow core), or they can be indirect measures where some feature that is closely related to the variable of interest is recorded. In either case, the measurement accuracy can be defined via knowledge of the sensor response and quality of calibration.

If the measurement error is large relative to the overall magnitude of the signal, the spatial or temporal patterns apparent in the data will be a poor representation of the

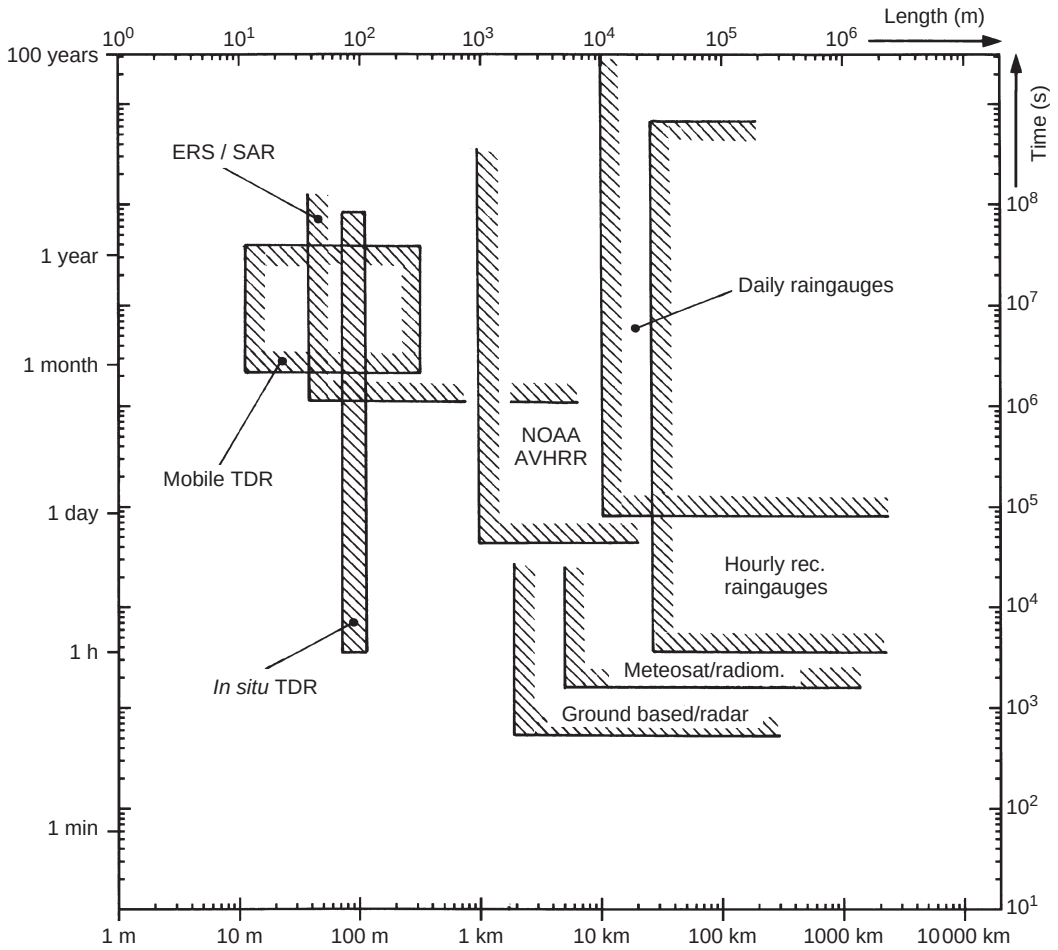


Figure 4 Space and timescales of rainfall and soil moisture variability that can be captured by different instruments represented as the domain between spacing and extent of the measurements. After Blöschl and Grayson (2000)

true underlying pattern. For example, random errors in TDR measurement of soil moisture are typically 2% v/v for field applications. In a temperate environment, soil moisture may vary by over 30% v/v so the relative error is low, but in dry times when soil moisture is close to wilting point, the range may be only 5% v/v and the measurement error is significant. Measurement of water level in a stream can be quite accurate (within a few mm), but there may be considerable error in the rating curve that converts depth to discharge because of, for example, extrapolation of the curve above the highest gauged water level.

There are two types of measurement errors, systematic and random. A systematic measurement error may be introduced either by an improper measurement setup (such as the catch deficit of rain gauges caused by wind exposure) or by improper rating functions (e.g. incorrect TDR calibration curves). Systematic errors can also be introduced by systematic sampling. For example, sun-synchronous satellite orbits lead to sampling at a consistent time during the diurnal cycle, thereby potentially introducing a bias in the observed variables. In many cases, it will be possible to correct for such systematic errors, provided additional (more accurate) data are available for comparison. A random measurement error is inherent in the stochastic processes of nuclear decay and neutron scattering, which underlie the operation of neutron soil moisture meters or it may be introduced by inaccurate readings of an observer who reads off the stage of a stream gauge. While for random errors it is not possible to apply a correction, taking multiple measurements of the same variable significantly reduces random errors and allows the statistical uncertainty to be assessed. For example, if there is a measurement error variance of 3 $(\%V/V)^2$ attached to a single TDR measurement, 10 such measurements at the same location and time pooled together only have a measurement error of 0.3 $(\%V/V)^2$, provided the errors of these 10 measurements are statistically independent. More generally speaking, the measurement error variance decreases with the inverse of the number of samples that are aggregated (see any basic statistics text e.g. Kottegoda and Rosso, 1996).

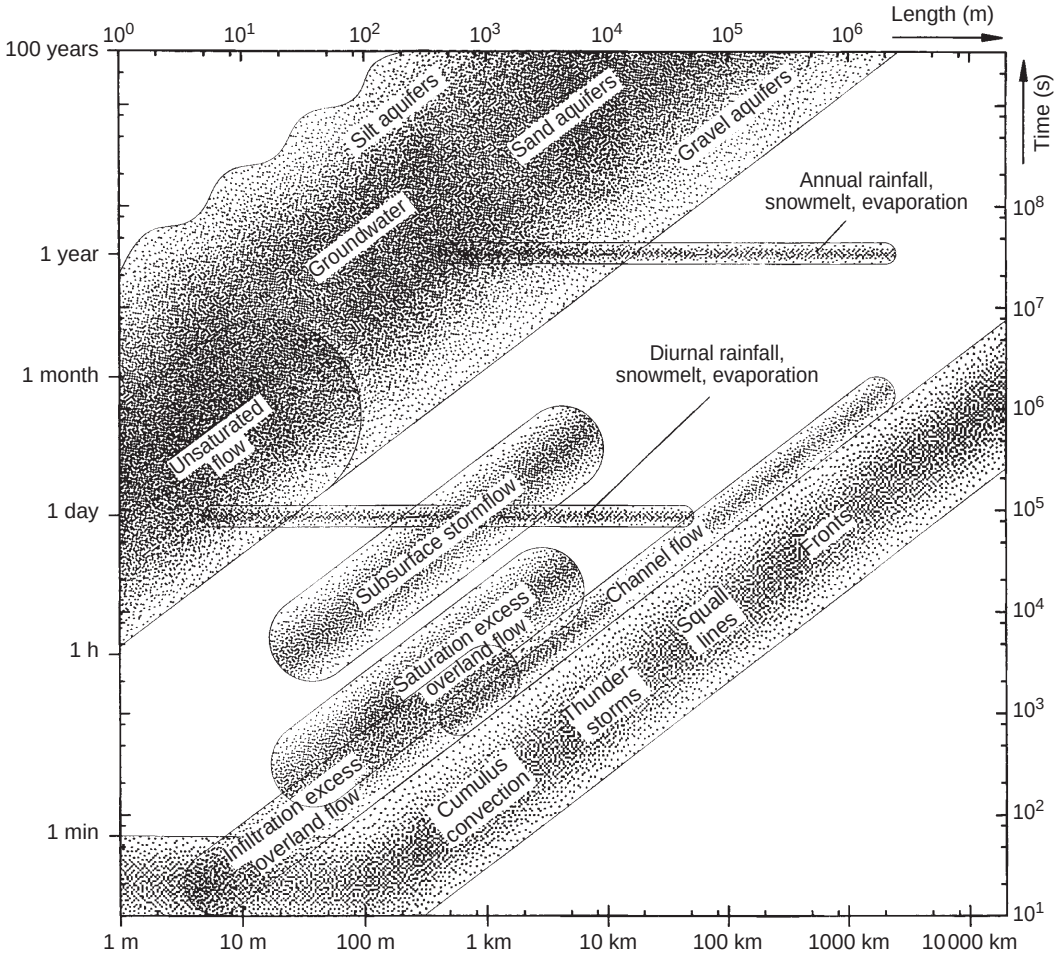


Figure 5 Schematic relationship between spatial and temporal process scales for a number of hydrological processes (after Blöschl and Sivapalan, 1995)

Trade-offs in Measurement

There is often a trade-off between great accuracy and few points (and hence a poor resolution/coverage), and poorer accuracy and lots of points. An important example in hydrology is the use of remote sensing data. For example, weather radar does not measure rainfall, but radar reflectivity, which is correlated with rainfall intensity, also depends on other factors (such as drop-size distribution and the presence of snow), only some of which are known. As a consequence, there is often a substantial error introduced when converting reflectivity to rainfall and good measurement systems integrate rain gauge and radar data. Other examples in remote sensing include soil moisture as estimated from SAR sensors where a huge number of points (pixels) in space are available but correlations between the SAR backscatter and soil moisture tend to be poor. The same is true with some ground data. For example, in an Alpine environment, it typically takes of the order of 3 min to measure snow depth but it may take 30 min or more to collect a sample of snow water equivalent. Similarly, total suspended solids (TSS) in a stream is best measured by collecting a sample, filtering a known volume in the laboratory, and weighing the dried solids, but stream turbidity, which can be measured continuously using attenuation of light, is often related to TSS concentration, albeit with some error. The optimum sampling strategy will often therefore involve some combination of a few measurements of high accuracy (to test surrogate relationships and calibrations) and many measurements of lower accuracy (to provide the spatial and temporal spacing required to capture the processes of interest).

Trialling and Pilot Studies

The preceding sections provide some basic concepts to assist in designing a measurement system, but there is nothing like a trial to see how those ideas translate to reality. There are two basic forms of pilot studies, small-scale field applications and synthetic studies.

Field trials are designed to test the real-world performance of sensors or monitoring networks to see if they do indeed perform as expected. Is the information sufficient to meet the objectives or test the hypotheses? Are the errors greater than was thought? Is the real-world variability greater or smaller than expected? Essentially, this is part of the iterative process referred to earlier where the field-pilot study is the opportunity to learn more about the phenomena of interest and refine the sampling before fully committing to an expensive program.

Synthetic studies can also be useful, particularly in the early planning phases of a measurement program. The idea is to use either similar data from elsewhere or simulated data with the characteristics of data expected from the instruments in the field to test how well the objectives of a sampling program are likely to be met. For example, if we are interested in estimating the total annual load of suspended solids from a river, we might use data from a similar river with high-resolution flow and TSS measurement to trial a range of different sampling strategies. This is done by "subsampling" from the TSS and flow data to see how the annual load estimates vary depending on how many samples are taken at what time. Is regular sampling the best or is it better to separate base flow from high flow? What is the minimum number of samples needed during high flow to estimate storm loads within the accuracy that is needed? Synthetic studies can also be used to assess the impact of error on the usefulness of the data. This is common in remote sensing applications where a "perfect pattern" is generated and then noise is added to see whether the resulting data are still useful for the intended applications.

Network Design and Site Selection

Network design and site selection are critical to quality monitoring outcomes. Network design covers aspects of site density and distribution and has been discussed earlier from a theoretical perspective, to maximize the likelihood of the data meeting the measurement objectives. In practice, accessibility, availability of power, coverage by cellular phone for telemetry, and agreement by landholders should also be considered.

Site selection is associated with the specific characteristics a site should have, which vary between measurement types, and must be determined from an understanding of the characteristics of the measurement systems. Examples of key requirements include fetch requirements for eddy correlation measurements and the desirability of locating a stream gauge at a hydraulic control. Often a site is chosen to be "representative" of a broader area. A highly accurate measurement at a point can introduce considerable error to a survey if it is extrapolated to a larger area of which that point is not representative. For example, piezometers sited along fault structures are likely to be unrepresentative of groundwater conditions in the same geological unit away from fault structures. It may be necessary to undertake a broader survey or use other data to quantify this "representativeness". Soils, geology, vegetation, and topographic mapping can be used to define the distributions of these characteristics in a study catchment compared to a broader area. In addition, remotely sensed data (both satellite and airborne geophysical data) may be required to define representative sites, particularly for subsurface studies. Similarly, the shape of a flow-duration curve indicates important features of streamflow response that can be used to compare catchments.

PRACTICAL ISSUES

To achieve trouble-free operation and high-quality data, a number of important practical issues must be addressed. These include:

- power
- time and date
- telemetry
- quality assurance
- calibration
- data archiving and documentation
- security and protection
- occupational health, safety, and site access
- capital/operation cost trade-offs.

It is difficult to give prescriptive answers on how these issues should be addressed, partly because technologies are changing rapidly and partly because they are often very situation-dependent. Thus, the following discussion aims to identify common issues rather than providing detailed solutions.

Power requirements of modern electronics are usually low and it is possible to run monitoring systems off batteries that can be periodically replaced. However, for extended operation, where telemetry is required, or for some higher power consumption sensors, such a solution is not sufficient. In these cases, solar panels offer a reasonably cheap and reliable power source for many environments. Solar panel charging circuits should include appropriate voltage regulation. It is valuable to monitor battery voltages to detect problems early, especially in telemetered systems, and to use loggers with nonvolatile data storage so that loss of data due to power problems is limited. It is still the case that a large proportion of reliability problems with instrumentation are related to power supply reliability, so attention to detail is important here as in other aspects.

Time and date: the golden rules are to make sure that they are set properly, that different data loggers (and computers used to access and reset data loggers) are synchronized, and that "summer" time issues are dealt with. Our preference is to always use the local standard time or UTZ. It is useful if

the clock in a data logger continues to run during a power outage because solar charging systems would allow loggers to restart measurements. This requires a backup power source (usually an internal logger battery) to maintain the core logger operation.

A variety of *telemetry* options are available and this is an area of rapid change with developments such as novel uses of modern cellular phone systems including text messaging, and innovations like sensor webs becoming available. The options at present include landlines, cellular phones, various satellite uplink options, and radio networks. Each of these options has advantages and disadvantages in terms of data rates, capital and operating costs, coverage (an issue in remote areas), and reliability. The key general advantages of telemetry are near real-time data access and archiving, early identification of instrument problems, and a reduced cost of station maintenance due to reduced site visit requirements.

Quality assurance is a critical part of any measurement campaign and it has pre-, co-, and post-measurement aspects. Many factors go to ensuring high-quality data. Elements that need to be considered here are sensor quality, sensor and logger calibration, sensor installation (poor installation = poor measurements), sensor and site maintenance, data checking, data documentation, and data archiving.

Calibration involves both ensuring that the data logging system measures and records the sensor response correctly, as well as relating the sensor response to the variable being measured. The golden rule is "do not rely on factory calibrations supplied with an instrument", especially where the environmental characteristics affect the sensor response (e.g. soil type effects on soil moisture sensor response). Calibration is an ongoing process because for most sensors, relationships drift over time. The frequency of recalibration required for different types of sensors varies greatly as a result of different instrument stabilities. Where multiple measurements of the same variable are being made, instrument cross-calibration and intercomparison are important for maximizing our ability to compare measurements. In some cases, there can be a dependence of the sensor response on the installation, as is often the case with soil moisture sensors. In such cases, field calibration is preferred. It is also important to ensure appropriate site maintenance occurs, particularly where some site exclusion causes vegetation conditions to be markedly different to maintenance of the surrounding area. Appropriate maintenance for sensors (e.g. cleaning) and loggers (e.g. battery changes) also needs to be considered.

Data checking and archiving are also important components of quality assurance, but is something that is often not done well. Gross data checking can be undertaken automatically by comparing measurements to expected ranges, by checking rates of change over time, and by comparisons to appropriate climatologies. This should be done as soon as possible after data collection and ideally in real time. It is no use realizing a year later that there was an instrument problem that could have been fixed immediately. Other important aspects of data checking include plotting data, comparing data to models, and intercomparisons between sites and sensors. Once data are checked they need to be stored (usually in digital format) in well-documented formats, with meta data about the measurement and quality assurance procedures and site characteristics, including geographic coordinates (with information stating the map coordinate system used). This is time consuming but critical to the long-term utility of your data. To ensure long-term database integrity, appropriate backup, archiving to stable media, and planned media migration are necessary. Again this is an area that is often not done well, particularly in the research environment. It is salutary to consider how much data has been generated during PhD and other research studies around the world, but is for all intents and purposes lost. The cost of digital storage is constantly reducing and the Internet provides an ideal avenue for making data widely available. Hopefully, this will encourage better archiving of data over time.

Physical security of instrumentation from interference by people, animals, and the physical environment is important. Prevention of interference by people and animals is dependent on using adequate enclosures, housings, and cable protection systems (e.g. cable conduits). Animals can be curious and can cause damage to equipment, especially cables, or disturb the site, which can be problematic if it influences the variable being measured (Figure 6). An example is destruction of vegetation cover under a net radiometer. Insects such as ants can cause problems by colonizing instrument housings. Issues to consider related to the physical environment include flooding during extreme events, adequate water proofing, anchoring of sensors (and cables) in streams, wind loads on masts and tripods, lightening protection, and control of operating temperature for some instruments.

A variety of *occupational health and safety* issues exist with any field activity. These range from issues such as traffic accident risks, being stuck in remote areas due to vehicle breakdown, exposure to the environment (heat and cold), and exposure to risks associated with local fauna and flora (e.g. snake bites) to issues specifically associated with setting up and maintaining measurement stations. This latter group can include elevated working platforms, excavations, machinery, and plant related risks (e.g. drilling rigs), and risks associated with stream gaugings by wading, from boats and from aerial cableways, among others. Modern occupational health and safety procedures generally require risk assessment of work activities, appropriate risk mitigation approaches, and appropriate training of personnel. It is always useful, while in the warm security of your office, to think through all of the possible problems

Figure 6 Before and after – curious animals around a reflector for aircraft overpass of a microwave sensor. Photos Rodger Young. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

you may encounter in the field and plan a solution to each, being sure you pack the necessary materials or equipment to carry out your plan.

Given many of the practical issues discussed above, it is clear that there can often be significant trade-offs between capital and operating costs. For example, telemetry systems generally reduce time requirements and travel cost for infield logger maintenance. Good quality installations with robust enclosures and housings also generally require a lesser basic maintenance of equipment. Therefore when planning a measurement system, this trade-off needs to be considered early in the process as it affects site design and equipment purchasing.

Back in the Office

A critical element to successful measurement programs is checking and using the data as soon as possible after it is collected. Part of the planning for a measurement program is designing what will be done with the data to actually test the hypothesis. Analysis methods may have even been tested as part of synthetic or pilot field studies. Early use of the "real" data will quickly expose instrument errors or flaws in the measurement program that may prevent your objectives being met. It is difficult to overstate the importance of this step. When in the middle of major field programs, it is very easy to be overwhelmed by the immediate practical matters of the measurement network and put off initial analysis of the data. High-quality measurement programs are regularly reviewed by reconsidering the objectives, analyzing whether the data are best meeting those objectives, ensuring that the data quality and archiving processes are sound, and so on.

SUMMARY

This article highlights several key considerations in a successful measurement programme. These are:

- The need to have a clear set of objectives or hypotheses to be tested.
- An understanding of the temporal and spatial variability of the phenomena of interest.
- An understanding of the characteristics of the instruments available for measurement, including errors.
- Matching of the temporal and spatial characteristics of the measurement with those of the processes of interest.
- Considering the trade-offs inherent in sampling, including the number of measurements and accuracy.
- Considering a range of practical issues related to setup in the field and the need to check and manage the data as it is collected.

These steps are summarized in the boxed section below.

Successful monitoring programs have a number of common features that contribute to the success. They are well designed and well resourced financially and in terms of expertise both in the initial stages and for the full measurement period. Good monitoring programs tend to be comprehensive in terms of measuring a number of complementary responses. Analysis will occur during the monitoring program and will look at the interrelationships between measurements to take advantage of complementarity and to test against the theoretical framework underlying the experimental design. However, it is important that the analysis is not blinkered by the assumed theoretical framework as this may disguise important results. For instance, the analysis of chemical (conservative ion and isotopic) techniques in conjunction traditional streamflow data is challenging previous conceptions of rainfall-runoff processes in a number of humid catchments (see discussion by McDonnell, 2003). Timely analysis allows an ongoing review of performance and appropriate modification, which is critical to success.

There are a number of new and emerging sensor technologies, as well as the ever-increasing availability of remote sensing information that bodes well for the future of hydrological measurement. But as has always been the case,

Box Basic questions for the design of a measurement program

A) What are the processes we are trying to capture (or hypotheses we are testing) with the measurement programme?

- What is the variability in time and space of the feature of the process that we will be measuring?
- Which variables should we measure and how representative are they of the process?
- What is the typical length scale of the feature of interest?
- How quickly does the feature change and are there particularly important timescales (e.g. diurnal, seasonal etc.)?
- What are the minimum and maximum values that are expected to be measured?
- Do we have *predictive methods* (e.g. relationships to auxiliary data) for defining the variation of the feature and how accurate are these?

B) How will the actual measurement be made?

- What measurement device (or devices) could be used?
- What is the accuracy of the chosen devices?
- What is the sampling support (time and space) of the device?
- Over what extent (time and space) do we need to make measurements?
- What level of calibration is needed for the device (initial and ongoing)?
- What are the practical constraints related to time, cost, and the logistics of the measurements to be made? Will telemetry be used? (Answers will indicate the possible number of samples or sites and enable the network to be planned.)
- Are there alternative variables to be measured that perhaps are less representative of the process or less accurate, but can be more easily collected?

C) We then need to try and match the needs of the measurement exercise with the variability in the feature being measured and the characteristics of the measurement device. In this step we need to recall that:

- if the spacing is too big compared to the feature of interest, we will not characterize small-scale variability (it will become ''noise'');
- if the extent is too small, we will miss out on the big scale pattern and instead measure a trend;
- if the support is too large, small-scale variability is smoothed out;
- if the sampling error is large compared to the variance of the feature, we will not detect the pattern.

Compromises will always be needed and, because of lack of knowledge, there will be some guesswork. Pilot trials in the field or using synthetic data can assist in this step. There will ultimately be constraints imposed by equipment, finances, and so on, and hence the final measurement system will not be ''ideal.'' A realistic assessment is needed of whether the original objectives will still be met given these pragmatic limitations, or whether the objectives themselves need to be revised.

D) Implementation of the measurement network

- how will the power be supplied?
- what amount of data storage or telemetry speed is needed?
- what physical protection is required?
- what are the capital and operating costs of various options?
- what procedures are in place for data checking and archiving?

The next step then is to look at the data set as it becomes available and ask how well did the measurement programme meet the objectives or enable the hypotheses to be tested? Are revisions to the program required? And so the process begins again as we understand more about both the underlying processes and the practical performance of our design.

it is the synthesis of "smart measurements" with theoretical developments and modeling that will yield the greatest advances in the coming years.

REFERENCES

Blackman R.B. and Tukey J.W. (1958) *The Measurement of Power Spectra*, Dover Publications: New York.

Blöschl G. and Grayson R. (2000) Spatial observations and interpolation. In *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Chap. 2, Grayson R. and Blöschl G. (Eds.), Cambridge University Press: Cambridge, pp. 17–50.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling – a review. *Hydrological Processes*, **9**, 251–290.

Boiten W. (2000) *Hydrometry*, AA Balkema: Rotterdam.

Dingman S.L. (2002) *Physical Hydrology*, *Second Edition*, Prentice-Hall.

Drexler J.Z., Snyder R.L., Spano D. and Paw U.K.T. (2004) A review of models and micrometeorological methods used to estimate wetland evapotranspiration. *Hydrological Processes*, **18**(11), 2071–2101.

Federico V.D. and Neuman S.P. (1997) Scaling of random fields by means of truncated power variograms and associated spectra. *Water Resources Research*, **33**(5), 1075–1085.

Gordon N.D., McMahon T.A. and Finalyson B.L. (1992) *Stream Hydrology: An introduction for ecologists*, John Wiley & Sons: Wichester.

Herschy R.W. (1995) *Streamflow Measurement*, *Second Edition*, E & FN Spon: London.

Journel A.G. and Huijbregts C.J. (1978) *Mining Geostatistics*, Academic Press: London.

Kaimal J.C. and Finnigan J. (1994) *Atmospheric Boundary Layer Flows: Their Structure and Measurement*, Oxford University Press: New York.

Kendall C. and McDonnell J. (1998) *Isotope Tracers in Catchment Hydrology*, Elsevier: Amsterdam.

Kottegoda N.T. and Rosso R. (1996) *Introductory Statistical, Probability and Reliability Methods for Civil and Environmental Engineers*. McGraw-Hill, NewYork.

Krige D.G. (1951) A statistical approach to some basic mine evaluation problems on the Witwatersrand. *Journal of Chemical and Metallurgical Society of South Africa*, **52**, 119–139.

Kuczera G. and Mroczkowski M. (1998) Assessment of hydrologic uncertainty and the worth of multiresponse data. *Water Resources Research*, **34**(6), 1481–1489.

Lehning M., Naaim F., Naaim M., Brabec B., Doorschot J., Durand Y., Guyomarc'h G., Michaux J.L. and Zimmerli M. (2002) Snow drift: acoustic sensors for avalanche warning and research. *Natural Hazards and Earth System Sciences*, **2**, 121–128.

Maidment D.R. (1992) *Handbook of Hydrology*, McGraw-Hill: New York.

Matheron G. (1965) *Les variables Régionalisées et leur Estimation*, Masson: Paris.

Matheron G. (1973) The intrinsic random functions and their applications. *Advances in Applied Problems*, **5**, 438–468.

McDonnell J.J. (2003) Where does water go when it rains? Moving beyond the variable source area concept n rainfall-runoff response. *Hydrological Processes*, **17**, 1869–1875.

McKenzie N., Coughlan K. and Cresswell H. (2002) *Soil Physical Measurement and Interpretation for Land Evaluation*, CSIRO Publishing: Melbourne.

McLelland S.J. and Nicholas A.P. (2000) A new method for evaluating errors in high-frequency ADV measurements. *Hydrological Processes*, **14**, 351–366.

Meischner P. (2004) Weather Radar: principles and advanced applications. *Physics of Earth and Space Environments*, Springer.

Scanlon B.R., Andraski B.J. and Bilskie J. (2002) Miscellaneous methods for measuring matric or water potential. In *Methods of Soil Analysis, Part 4, Physical Methods*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: No. 5, pp. 643–670.

Thompson S.K. (2002) *Sampling*. John Wiley & Sons, New York, p. 367.

Vanmarcke E. (1983) *Random Fields: Analysis and Synthesis*, The MIT Press: Cambridge.

Western A.W., Grayson R.B. and Blöschl G. (2002) Scaling of soil moisture: a hydrologic perspective. *Annual Review of Earth and Planetary Sciences*, **205**, 20–37.

Wiener N. (1949) *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, The MIT Press: Cambridge.

7: Methods of Analyzing Variability

LARS GOTTSCHALK

Department of Geosciences, University of Oslo, Oslo, Norway

In an introductory part basic concepts from probability theory, and specifically from the theory of random processes, are introduced as a basis for the characterization of variability of hydrological time series, space processes and time-space processes. A partial characterization of the random process under study is adopted in accordance with three different schemes:

1. *Characterization by distribution function (one dimensional),*
2. *Second moment characterization, and*
3. *Karhunen–Loève expansion, that is, a series representation in terms of random variables and deterministic functions of a random process.*

The article follows the same division into three major sections. In the first one, distribution functions of frequent use in hydrology are shortly described as well as the flow duration curve. The treatment of second-order moments includes covariance/correlation functions, spectral functions and semivariograms. They allow establishing the structure of the data in space and time and its scale of variability. They also give the possibility of testing basic hypothesis of homogeneity and stationarity. By means of normalization and standardization, data can be transformed into new data sets owing these properties.

The section on Karhunen–Loève expansion includes harmonic analysis, analysis by wavelets, principal component analysis, and empirical orthogonal functions. The characterization by series representation in its turn assumes homogeneity with respect to the variance–covariance function. It is as such a tool for analyzing spatial-temporal variability relative to the first- and second-order moments in terms of new sets of common orthogonal random functions.

INTRODUCTION

Observations from studies of hydrological systems at an appropriate scale are characterized by complex variation patterns in time and space and reflect regularity in a statistical sense. It is commonly reasonable to apply concepts from statistics and probability theory to be able to properly describe these observations and model the system. The theory of random processes is of particular interest.

The basic ideas of probability theory and random processes are well known. Experiments are basic elements of probability theory and statistics and are defined as actions aimed at investigating some unknown phenomenon or effect. The result is, as a rule, a set of values in a region in space and/or an interval in time. An experiment in a laboratory can be repeated and different realizations can be obtained under the same conditions (*experimental data*). In hydrology, it is the nature that performs experiments and therefore it is not possible to control the conditions (*historical data*). The historical data at hand are considered as *samples* (or *realizations*) from some very large or even infinite *parent population*. In the following small letters, say x, will denote the sample while capital letters, X, will denote the corresponding theoretical population. The structure of the available data guides the method to be used for analyzing variability. It is possible to distinguish between three different situations:

1. An observation point in space is fixed and only the development in time is observed at this point, as

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

illustrated in Figure 1, where the annual streamflow for the Göta River for the period 1807–1938 is shown. This is referred to as a *time series* $x_k = x(t_k),\ k = 1, \ldots, N$, where t_k denotes the (regular) observation points in time (here years) and N is the number of observations (in the example, $N = 131$). The process is characterized by an active and inherent dynamic uncertainty, the properties in different points change with time in a random manner. In the general case, the order in which time series is sampled is of outmost importance as the temporal fluctuations show *persistence*, that is, adjacent observations show dependence. The opposite situation when data are independent and can be reshuffled without loss of information is an important special case.

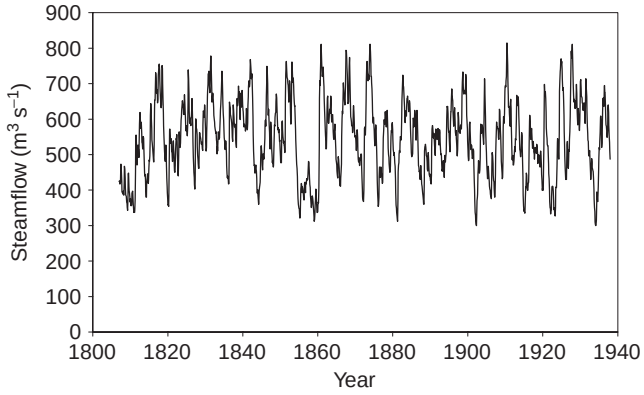


Figure 1 Time series: Streamflow of the Göta River at Sjötorp, Sweden 1807–1938

2. The second situation – a *space process* $x_i = x(\mathbf{u}_i),\ i = 1, \ldots, M$- is illustrated by georadar measurements reflecting the geological structure of the top layer of the Gardermoen delta deposit at Moreppen (Norway) along transects of some 10th of meters. $\mathbf{u}_i$ denotes the position in space of the i-th measurement, of totally M (Figure 2). In the example, observations are made at a regular grid in space. More common is the case of spatial measurements from an irregular observation network. It may be assumed in this case that the changes over time of the system are small (at a human timescale). The uncertainty in the description

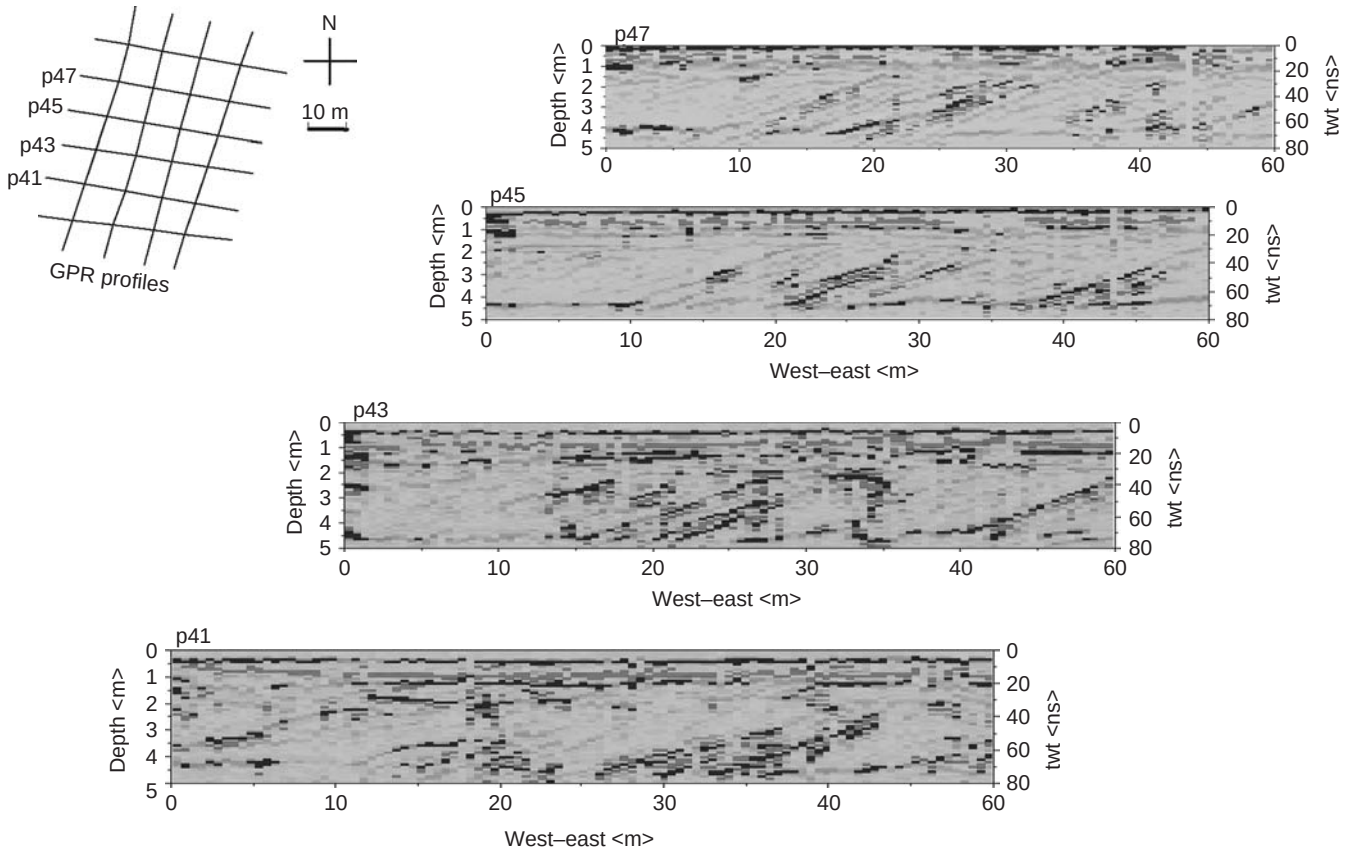


Figure 2 Space process: The geological structure of the top layer of the Gardermoen delta deposit at Moreppen (Norway), as reflected by georadar measurements (ground penetrating radar (GPR) signals for four profiles. The strong reflectors in the dipping forest unit are from silty layers with high soil moisture content. Yellow and green colors reflect drier sand. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of the properties of a disordered system, for which the development in time does not matter, is of passive nature. Though the changes in time of a characteristic at a given position might be negligible, its value is unknown until it is measured. Measurements in all possible points are, as a rule, neither feasible nor economic, and the information value is lowered by the measurement errors as well. Persistence in data is also of relevance for spatial data, that is, the order in which they are sampled in the two-dimensional space is of vital importance.

3. Figure 3 offers an attempt to illustrate of the most general case showing the space-time development of streamflow (monthly values) along the different branches of the Rhône River (French part). In reality, observations of such a *time-space process* $x_{ik} = x(\mathbf{u}_i, t_k),\ i = 1, \ldots, M; k = 1, \ldots, N$ represent measurements in discrete irregular points along the river network at discrete regular times (and not as the fully reconstructed space-time development as in Figure 3), that is, a vector of data where columns represent different points in space and rows time. These observations are used to get an idea about the pattern of variation of the whole system by means of reconstructing the past development in time and space and/or forecasting the future.

The scale problem is fundamental in all description and modeling of time-space processes. A phenomenon that seems to contain mainly deterministic elements in a micro scale might at a larger scale demonstrate characteristics that vary much and demand a probabilistic approach for their description. At a still larger scale (macro scale), the same structure can appear to be a part of an object that can be described by its mean value or by classes. Variation of precipitation intensity at annual and daily timescales seems to behave totally at random, while, at a finer timescale, (minutes) this variation assumes a dynamically varying pattern (Figure 4). At the monthly scale, a seasonal variation might be present. In other words, there is, as a rule, the lower and the upper boundary for the variation range (distance or time) within which a model for characterization of the patterns of variations has a practical value. This is the point of departure for the application of classical theory of random processes in hydrology, which has been mainly applied to study stationary (time independent) random processes like annual or submonthly quantities (upper three graphs, and lower graph in Figure 4). Possible large-scale elements like "trends" and "periods" were looked upon as "deterministic", and, as such, identified and subtracted from the original data (e.g. Hansen, 1971; Yevjevich, 1972). This perspective contrasts with the current view accepting "... irregularly changes, for unknown reasons on all timescales" (National Research Council, 1991). Random process models need to be changed accordingly. The scale problem is of course not only limited to processes in time. Referring to the geological structure in Figure 2, the patterns of variability and its character will change drastically both when going down in scale as when going up. The topic of scale is further developed in **Chapter 6, Principles of Hydrological Measurements, Volume 1** and **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**).

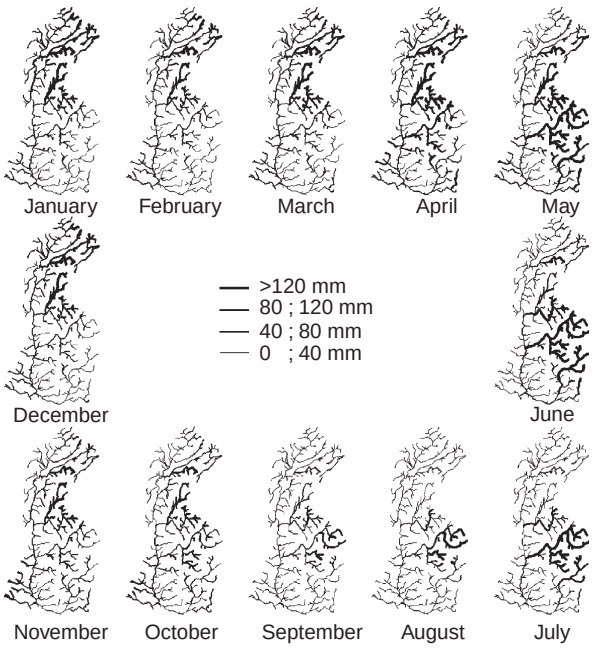


Figure 3 Time-space process: Estimated streamflow of the Rhône River for 12 months of the year (Reproduced from Sauquet and Leblois, 2001, by permission of Societe Hydrotechnique de France)

Let us turn back to probability theory terminology and continue formalizing the description of the outcome of an experiment. In the elementary case, the population is described in terms of a random variable and, in more complex situations, as a random process (random field). Figure 3 may be used to illustrate our basic model, where each point in the sample space $\mathbf{u} \in \Omega$ (points along rivers) maps into a time function $X(\mathbf{u},t)$. A point $\mathbf{u}_k$ in space can be specified, which results in a random process in time, a time series, $X(t) = X_k(t)$, like the one shown in Figure 1. If the data of the series fulfill the condition of having an independent identical distribution (i.i.d.), they can be treated as a sample of a random variable X. Only in this latter case, the one-dimensional probability distribution $F_X(x)$ will give a complete characterization of X. The time $t = t_i$ can be frozen, which leads to a random process in space only $X(\mathbf{u}) = X_i(\mathbf{u})$, illustrated in Figure 2. Also, in this case, a one-dimensional distribution describes variations across space. The i.i.d. condition should, of course, be fulfilled to give a complete description.

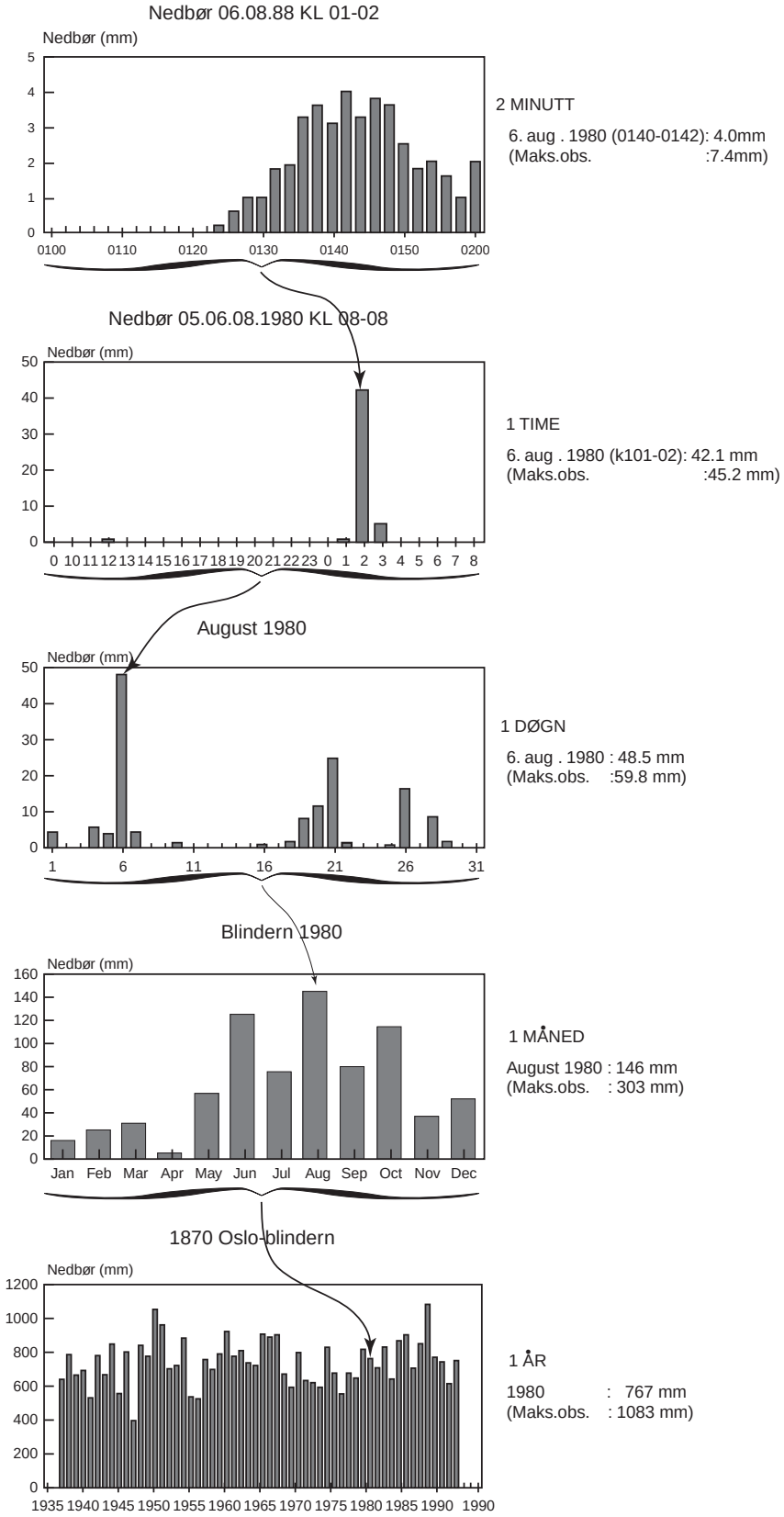


Figure 4 Observed precipitation at Blindern (Oslo) with 5 different time resolutions 2 min, 1 h, 1 day, 1 month, and 1 year

Many important characteristics of random processes viz. *homogeneity* (*stationarity*), *isotropy*, and *ergodicity*, permit a more effective use of the limited data amount available for estimation of important properties of the process. The strict definitions of these characteristics can be formulated with the help of the multivariate (M-dimensional) distribution function $F_M(\mathbf{x})$ (abbreviated df). A random process is called *homogeneous* if all multivariate distributions do not change with the movement in the parameter space (translation, not rotation). This implies that all probabilities depend on relative and not absolute positions of points in the parameter space. The term "*stationary*" instead of homogeneous is usually used for one-dimensional random processes (time series), that is, the df does not change with time. A process is called *isotropic* if the multivariate distribution function remains the same even when the constellation of points is rotated in the parameter space. A random process is *ergodic* if all information about this multivariate distribution (and its parameters) is contained in a single realization of the random field. It is important to note that this property is also related to the characteristic scale of variability of the process. If the process is observed over a time interval (or region in space), which in its extension is of the same order of magnitude as the characteristic scale (or smaller), the estimate of the variability of the process will by necessity be negatively biased. The process will not be able to show its whole range of patterns of variability. A rule of thumb has been to say that a process needs to be observed for a period of time that is at least 10 times the characteristic scale of the process, in order to eliminate the negative bias in the variance. In times when environmental and climate change are in focus and accepting that the process shows variability on a range of scales, the dilemma related to the ergodicity problem is obvious. Do the observed data reveal the real variability of the natural processes under study? In **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**, this topic is brought further and the process scale related to the natural variability is confronted with the measurement scale, defined in terms of *extent* (coverage), *spacing* (resolution), and *support* (integration volume (time) of a data set).

The parameter space of a random process $X(\mathbf{u},t)$ in the general case includes an unlimited and infinite number of points. Characterization by means of distributions functions is therefore only of a theoretical value. When complex variation patterns are concerned, a possibility of a direct estimation of the underlying multivariate distribution function is not tractable. The conventional way of handling this difficulty is to accept a *partial characterization*. The two most widely used are as follows:

1. Characterization by distribution function (one dimensional).
2. Second-moment characterization.

In a characterization by the distribution function, only the first-order probability density is specified. In a characterization by distribution function in the general case, a multivariate distribution would be needed for a complete characterization. The one-dimensional distribution constitutes in this case the *marginal distribution* of the data. The *flow duration curve* (fdc) widely used in hydrology is a good example. In a second-moment characterization, only the first and second moments of the process are specified, that is, mean values, variances, and covariances. Random processes, which are postulated to be homogeneous (stationary), in practice satisfy this condition only in a weak sense and not strongly, which means that they possess this property only with respect the to the first- and second-order moments (*weak homogeneity, weak stationarity*). A further possibility is to apply

3. Karhunen–Loève expansion, that is, a series representation in terms of random variables and deterministic functions of a random process.

The deterministic functions can either be postulated as for harmonic analysis and analysis by wavelets or they can be determined from the data themselves by analysis in terms of empirical orthogonal functions (eof) or principal components (pca).

In this article, these three ways for representation of a random process will be followed, thus defining three methods for describing variability of hydrologic variables. The development of relations between variability and scale is treated in **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**, although some aspects of the problem are touched upon here as well. Before going into a detailed statistical analysis the importance of "looking" at data should be stressed. A visual inspection of graphical plots of the observed data like those shown in Figures 1, 2, and 4 is a natural point of departure when analyzing variability. In our age of nearly unlimited computing power, this visual graphical data exploration is becoming increasingly important. A further step is an exploratory data analysis, where different hypotheses concerning the structure of the data are tested (Tukey, 1977).

CHARACTERIZATION BY DISTRIBUTION FUNCTION

Restricting ourselves to the one-dimensional case, the basic problem is the following: find a distribution function (df) $F_X(x)$ (probability density function $f_X(x)$, pdf), which is a good model for the parent data $x_1, x_2, \ldots, x_N$. From probability theory, it is well known that this distribution only gives a full description of phenomena in case data can fulfill the condition of being independent identically distributed (i.i.d.). In many applications in hydrology, the

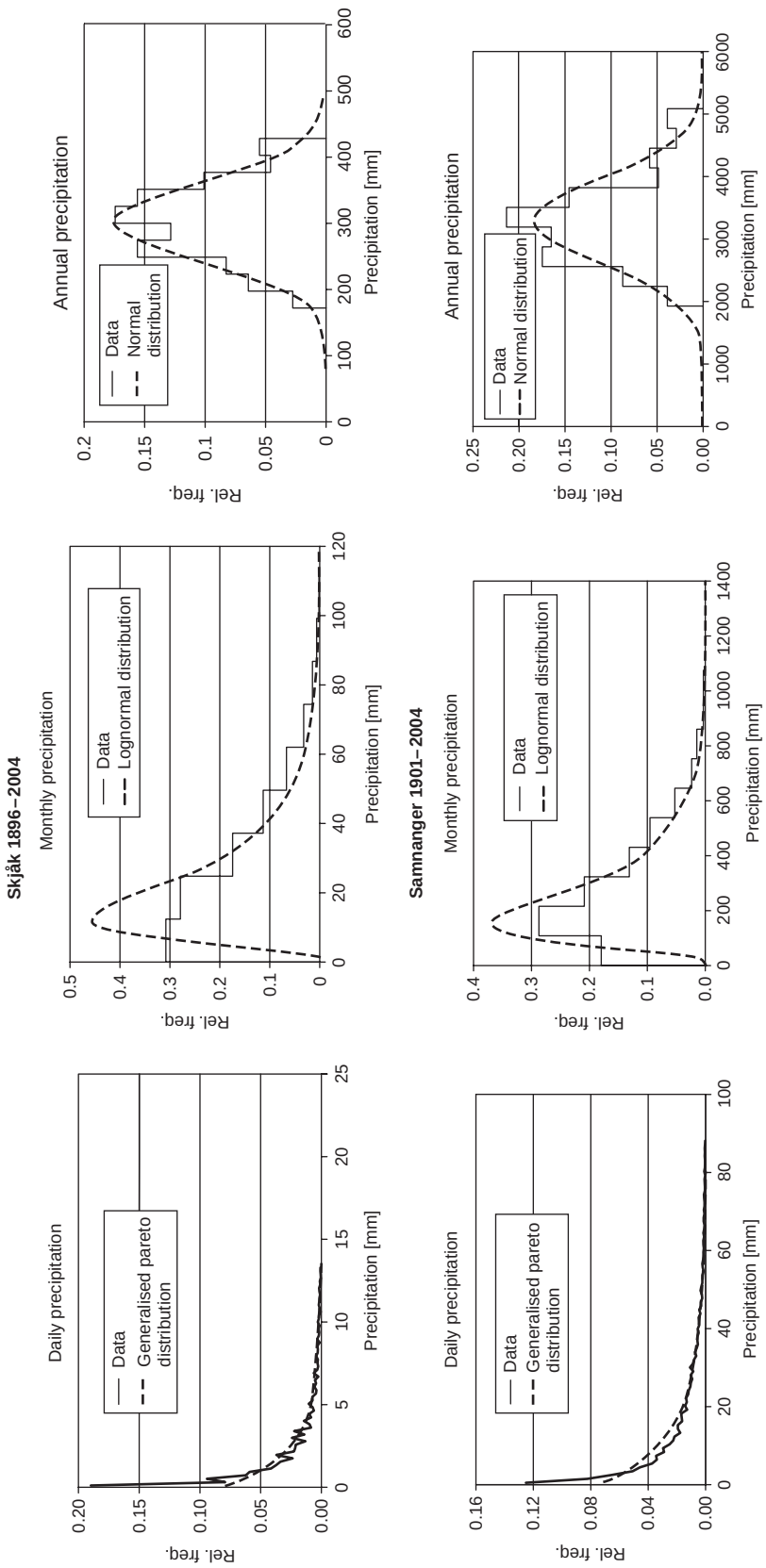


Figure 5 The distribution (pdf) of rainfall at two Norwegian rainfall stations (Skjåk and Samnanger) for three different durations 1 day, 1 month, and 1 year

i.i.d. assumption is rather postulated than really tested and the one-dimensional distribution is to be interpreted as a partial characterization (the marginal distribution function of a multivariate one). Anyhow, this marginal distribution might be a proper tool to study the data. The application of the normal distribution for frequency analysis of runoff data by Hazen (1914) symbolizes the start of the fitting a theoretical distribution to observed data in hydrology. Somewhat later, it became obvious that the river runoff distribution is not symmetrical and also the gamma distribution was introduced in hydrological analysis (Foster, 1924; Sokolovskij, 1930). Important benchmarks in the utilization of probability theory and statistical methods in hydrology were the developments by Kritskij and Menkel (1946), who suggested a transformation of the gamma distribution and Chow (1954), who introduced the lognormal distribution. Figure 5 illustrates the change in the distribution (pdf) of precipitation with changing time step for data from to two stations in Norway, starting from a highly skewed distribution for daily data, to a lognormal shape for monthly data and ending with a symmetric normal distribution for annual data. This example provides an illustration of the central limit theorem in statistics that states that the distribution of a sum of random variables converges to normal distribution as the number of elements in the sum approaches infinity. How quickly the sum converges (and also if it converges) depends on how well certain assumptions are fulfilled. Still, it is important to note that data follow statistical laws and that knowledge of these laws helps when analyzing and interpreting results as well as in choosing an appropriate model. The list of theoretical distributions applied to hydrological data since Hazen can be made very long. A remark might be that the advantage of using a more complex distribution with many parameters instead of the classical ones (the normal, the gamma, the lognormal) is usually minor in relation to the small data samples commonly available and thereby related uncertainty.

The flow duration curve (fdc) represents the relationship between the magnitude and frequency of daily, weekly, monthly (or some other time interval) of streamflow for a particular river basin, providing an estimate of the percentage of time a given streamflow was equaled or exceeded over a historical period (Vogel and Fennessey, 1994). The fdc has a long tradition of application for applied problems in hydrology. The first paper on this topic, in accordance with Foster (1933), is the one published by Herschel in 1878. The interpretation of fdc by Foster is "(Frequency and) duration curves may be considered as forms of probability curves, showing the probability of occurrence of items of any given magnitude of the data". In this respect, an fdc is a plot of the empirical quantile function X_p, that is, the p-th quantile or percentile of streamflow for a certain duration versus exceedance probability p, where p is defined by

$$p = 1 - P\{X \leq x\} = 1 - F_X(x) \tag{1}$$

Foster sees two distinct uses of the fdc: (i) if treated as a probability curve, it may be used to determine the probability of occurrence of future events; and (ii) it can be used merely as a conventional tool for studying of the data. Mosley and McKerchar (1992) look at the problem from a different point of view: "It (a flow duration curve) is not a probability curve, because discharge is correlated between successive time intervals, and discharge characteristics are dependent on season of the year. Hence the probability that discharge on a particular day exceeds a specific value depends on the discharge on proceeding days and on the time of the year". Indeed, the fdc gives a static and incomplete description of a dynamic phenomenon in terms of cumulative frequency of discharge. To have a complete description, it is necessary to turn over to a multivariate distribution, which defines the parent distribution of the data. Anyhow, the marginal distribution of this parent distribution is the fdc. It is a natural point of departure when analyzing streamflow data, which is evident from its wide practical application (Foster, 1933; Vogel and Fennessey, 1995; Holmes *et al.*, 2002).

Foster, in his original paper, compared daily, monthly, and annual fdcs and recognized the fact that the differences between the curves for different durations (timescale) changed with the type of river basins. Searcy (1959) performed a similar comparison. With present computer technology, it is usually taken for granted that the fdc is founded on daily data records (Fennessey and Vogel, 1990; Yu and Yang, 1996; Holmes *et al.*, 2002), although exceptions exist where monthly and 10-day period data are applied to determine the fdc (Mimikou and Kaemaki, 1985; Singh *et al.*, 2001).

SECOND-MOMENT CHARACTERIZATION

The characterization of a random process by means of moments is an alternative to the characterization by means of distribution function. When analyzing hydrologic data as realizations of random processes, it is as a rule neither tractable nor of interest to formulate models in terms of distribution functions. The most common situation is the one when only one realization of the random process is at hand. In order to be able to solve problems of identification, interpolation, and extrapolation, it is usually assumed that the following conditions are satisfied, namely that the random process studied is ergodic, homogeneous, and isotropic. Usually, the second-order homogeneity is a sufficient condition, which is also called weak homogeneity (weak stationarity), as explained earlier. The classical methods for the second-order analyses of stationary stochastic

processes are based on the works by Wiener (1930, 1949) and Khinchin (1934, 1949), where similarity and relationship between autocorrelation function (acf) and spectral function (sf) have been explained. Correlation and persistence (memory, inertia) described by means of acf and sf, which are statistical moments of the second order, belong to the most important characteristics of random processes. If a random process has a normal distribution, then the information of the first (mean value) and second order (acf, sf) are sufficient for an application to multidimensional problems, that is, in this case, weak stationarity implies strict stationarity.

As stated earlier, the point of departure for our study is a space-time random process. Depending on the way data are sampled from this general process, the observations may be looked upon as a random variable, a time series, a random vector, and a dynamically coupled time-space process, respectively.

Random Variable

It is a common situation in hydrology that observed data are treated as a sample from a random variable, as already mentioned in the previous section. The typical situation is data sampled over time with regular intervals at a fixed site in space – $x_1, x_2, \ldots, x_N$. The situation of data sampled on a regular or irregular network in space at a fixed time is also of interest, for example, snow or soil moisture surveys.

If X is a random variable with cumulative distribution function $F_X(x)$, the first moment is the *mean value* or expected value of X:

$$m = m_X = E[X] \tag{2}$$

The second-moment $E[X^2]$ is the mean square of X. Central moments are obtained as the expected values of the function $g(X) = (X - m)^n$. The first central moment is zero. The second central moment is by definition the *variance* of X:

$$\sigma^2 = \sigma_X^2 = \mathrm{Var}[X] = E[(X - m)^2] = E[X^2] - m^2 \tag{3}$$

The square root of the variance σ_X^2 is the *standard deviation* σ_X of X. If $m = 0$, the standard deviation is equal to the root of the mean square. When $m \neq 0$, the variation of X is usually described by means of the *coefficient of variation*:

$$V = V_X = \frac{\sigma_X}{m_X} \tag{4}$$

The *skewness coefficient* γ_1 is defined from the third-order central moment

$$\gamma_1 = \frac{E[(X - m_X)^3]}{\sigma_X^3} = \frac{E[X^3] - 3m_X E[X^2] + 2m_X^3}{\sigma_X^3} \tag{5}$$

Moments are used to describe the random variable and its distribution. The mean value is a measure of central tendency, that is, it shows around which value the distribution is concentrated. Other alternative measures are (i) the *median, Me*, the value of which for X corresponds to $F(x) = 0.5$ (i.e. the middle value in the distribution) and (ii) the *mode*, M, which corresponds to the value of x when the pdf is at maximum (i.e. the most frequent value). The variance, alternatively the standard deviation, describe how concentrated is the distribution around its center of gravity, the mean. The skewness describes how symmetrical the distribution is. If $\gamma_1 = 0$ the distribution is totally symmetrical, while if $\gamma_1 > 0$, it has a "tail" to the right (towards large x values) and if $\gamma_1 < 0$, it has a "tail" to the left (towards small values of x). The parameters m_X, σ_X, and γ_1 offer an acceptable approximation of the (marginal) distribution function $F_X(x)$ of the variable X for most applications in hydrology. In the applied case, m_X, σ_X, and γ_1 are substituted by the corresponding *sample moments* $\overline{x}$, s, g_1, respectively:

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x(t_k) \tag{6}$$

$$s_X{}^2 = \frac{1}{N}\sum_{k=1}^{N} x(t_k)^2 - \overline{x}^2 \tag{7}$$

$$g_1 = \frac{\left(\frac{1}{N}\sum_{k=1}^{N} x(t_k)^3 - 3\overline{x}\frac{1}{N}\sum_{k=1}^{N} x(t_k)^2 + 2\overline{x}\right)}{s^3} \tag{8}$$

The moments of the sample are accompanied with *sampling errors* (*standard errors*) and biases. The well-known formula for the standard error of the mean is

$$\sigma_{\overline{x}} = \frac{\sigma_X}{\sqrt{N}} \tag{9}$$

or as estimated from the sample

$$s_{\overline{x}} = \frac{s_X}{\sqrt{N}} \tag{10}$$

The standard error of the mean is not dependent on the distribution of the population. Standard errors of higher-order moments are related to the theoretical distribution (Kendall *et al.*, 1987). The moments developed here are obviously of the same importance whether or not the necessary i.i.d. conditions are satisfied. On the other hand, it is important to clearly state whether the moments of the one-dimensional distribution of a random variable are considered or the moments of the marginal distribution of a random process. In the latter case, the standard errors and the bias of the moments are related to the

structure (covariance function) of the data (*see* "Time Series" below). The classical methods for statistical tests for random variables are thus not directly applicable in this latter case.

Time Series

A time series is a sequence of data that are sampled over time with regular intervals at a fixed site in space or data that are sampled along a line in space at regular intervals at a fixed time – $x(t_1), x(t_2), \ldots, x(t_N)$ like in Figure 1. The difference, compared to the previous section, is that the order in which data are sampled in the parameter space $t_1, t_2, \ldots, t_N$ is of vital importance here. The two first-order moments m_X and σ_X determine the marginal distribution function $F_X(x)$ of the random process $X(t)$, like in the case of a random variable. However, for a random process, a descriptor of the random structure of this process needs to be added and the autocovariance function (acf) (the second-order mixed moment) determines this structure as an acceptable approximation. The covariance $B(t, t')$ of the state of a random process between two different points in time $X(t)$ and $X(t')$ defines this *autocovariance* function (of t and t'):

$$\begin{aligned} B(t, t') &= B_X(t, t') = \mathrm{Cov}[X(t), X(t')] \\ &= E[X(t), X(t')] - m(t)m(t') \end{aligned} \tag{11}$$

Similarly, the *autocorrelation* function is defined by

$$\rho(t, t') = \rho_X(t, t') = \frac{B(t, t')}{\sigma(t)\sigma(t')} \tag{12}$$

which is the correlation coefficient between $X(t)$ and $X(t')$. A weakly stationary random process has the following properties:

$$\begin{aligned} &E[X(t)] = m(t) = m, \\ &\mathrm{Var}[X(t)] \text{ exists for all } t, (\mathrm{Var}[X(t)] < \infty), \\ \text{and} \quad &B(t, t') = B(|t - t'|) = B(\tau) \end{aligned} \tag{13}$$

$\tau = |t - t'|$ describes the relative distance between the two points in time. The autocovariance function, as well as the autocorrelation function, depend therefore only on relative positions in time, and not absolute ones. A simple measure of the scale of variability can be defined as the integral under the autocorrelation for $\tau > 0$, the integral scale θ_X of a stationary process, that is,

$$\theta_X = \int_0^\infty \rho(\tau)\, \mathrm{d}\tau \tag{14}$$

For stationary processes, a characterization using the *spectral function* (sf) $S_X(f)$, where f is the frequency, is equivalent to the covariance function characterization:

$$S_X(f) = \int_{-\infty}^{\infty} B_X(\tau)\, \mathrm{e}^{i2\pi\ f\tau}\, \mathrm{d}\tau \tag{15}$$

and

$$B_X(f) = \int_{-\infty}^{\infty} S_X(f)\, \mathrm{e}^{-i2\pi\ f\tau}\, \mathrm{d}f \tag{16}$$

These two equations are usually referred to as the Wiener–Khinchin relations.

The stationarity assumption equation (13), which also is the theoretical background for the derivation of the Wiener–Khinchin equations, demands finite variance. The existence of a process with variation at many scales might violate this assumption theoretically. The demand of a weak stationarity is then too strong. An alternative is then to put the demand of stationarity on the differences $[X(t_i) - X(t_j)]$, that is, that its mean value is constant and variance is finite and independent on absolute position. This means mathematically:

$$E[X(t_i) - X(t_j)] = 0, \tag{17}$$

$$\begin{aligned} \text{and} \quad \mathrm{Var}[X(t_i) - X(t_j)] &= E[(X(t_i) - X(t_j))^2] \\ &= 2\gamma(|t_1 - t_2|) = 2\gamma(\tau) \end{aligned} \tag{18}$$

The conditions above are called the *intrinsic hypothesis*. In classical works on turbulence and also in meteorology, the function equation (18) is named *structure function* (Kolmogorov, 1941). It is also called *variogram* (Matheron, 1965) and $\gamma(\tau)$ is accordingly called *semivariogram* (sv), that is,

$$\gamma(\tau) = \tfrac{1}{2} E[(X(t_1) - X(t_2))^2] \tag{19}$$

The intrinsic assumption is more general than the demand for second-order stationarity (weak stationarity). If the condition for weak stationarity is satisfied, that is, the variance exists and equals $B(0)$, the following relation between the semivariogram and covariance functions can be established:

$$\gamma(\tau) = B(0) - B(\tau) \tag{20}$$

The semivariogram is not commonly used for analyzing temporal data in hydrology. It is treated more in detail in the section for the analysis of spatial data below.

The covariance $\hat{B}(k)$ and correlation functions $r(k)$ of the sample are estimated by (stationarity is assumed):

$$\hat{B}(k) = \hat{B}(k\ \Delta t) = \frac{1}{N-k} \sum_{k=1}^{N-k} x(t_j)x(t_{j+k}) - \overline{x}^2; \quad k = 0, \ldots, K \tag{21}$$

$$r(k) = \frac{r(k\ \Delta t) = \hat{B}(k)}{s^2};\ k = 0, \ldots, K \tag{22}$$

where k is the time lag in terms of the interval Δt between observations in time and K is the maximum lag. The standard errors grow with increasing lag as fewer and fewer observations are available for estimation. A rule of thumb is to set $K = 0.1{*}N$ as only a few foremost terms can be known with some acceptable confidence. An estimate of the sample semivariogram $\hat{\gamma}(k)$ when data are available at regular time intervals is

$$\hat{\gamma}(k) = \hat{\gamma}(k\ \Delta t) = \frac{1}{2(N-k)} \sum_{k=1}^{N-k} (x(t_j) - x(t_{j+k}))^2; \quad k = 0, \ldots, K. \tag{23}$$

The estimation of the sample spectral function is more complicated. There exist two principal approaches. One relies on the Wiener–Khinchin equation (15) and the spectral function is calculated from the sample autocorrelation function by numerical integration of this equation. The other is founded on an expansion of the observed data in terms of Fourier series, resulting in a so-called *periodogram*. Very fast algorithms are available, that is, FFT (fast Fourier transform) for this purpose. The highest frequency that can be represented by the sample spectral function is the so-called *Nyquist frequency* $f_N = 1/(2\Delta t)$. Low-frequency components with wavelengths in the order of magnitude of the total period of observations ($T = N{*}\Delta t$) or larger are estimated with poor accuracy and are numerically filtered out. The sample spectrum thus cover frequencies for the range $\frac{K}{N} 1/T < f < 1/(2\Delta t)$, where K as before is the truncation level for the time lag. Figure 6 illustrates the estimated acf, sv, and sf of the discharge data from the Göta River, Sweden (Figure 1). This river drains a large lake Vänern, which explains the very high autocorrelation in these series both for the annual and the monthly time step. A seasonal variation pattern as well as a strong between year variability is noted. Note that the Nyquist frequency of 0.5 in the lower left diagram for annual data relates to a period of 2 years, while the corresponding lower right diagram for monthly data relates to a period of 2 months.

How far can the information about the variability of the studied process contained in the acf, sv, and sf be interpreted, that is, which model can be applied? This question is linked to the discussion in the introductory section about the existence of variability in data across a range of scales. A traditional approach, as already commented on, has been to divide the variability in time series into two parts, namely, one related to deterministic variations and the remaining one to random fluctuation. The goal in this approach has been to be able to describe the random part by a stationary random process. The deterministic part in its turn may be described by long-term trends as well as purely periodic fluctuations. A common model for this case would look as follows:

$$X(t) = D_t + P_t + S_t \varepsilon(t) \tag{24}$$

where D_t denotes the trend, P_t and S_t represent the periodic elements in the mean value and standard deviation respectively, and $\varepsilon(t)$ is the random fluctuations. The random component is found by means of rearrangement:

$$\varepsilon(t) = \frac{X(t) - D_t - P_t}{S_t} \tag{25}$$

which then is assumed to be a stationary process. It can be described by a simple Markov process like an autoregressive model of order one, AR(1), or two, AR(2). The correlation function for these latter two models has the form:

$$\rho(\tau) = \rho(1)^\tau \tag{26}$$

$$\rho(\tau) = \frac{(1-\rho(1)^2)\rho(\tau-2) - (\rho(1) - \rho(1)\rho(2))\rho(\tau-1)}{\rho(2) - \rho(1)^2} \tag{27}$$

The first one exponentially decays towards zero. For the second one, it is more difficult to deduct its performance. It can be shown that it also decays exponentially for large lags towards zero from the positive side if both the lag one and lag two correlations are positive or as a periodically damped oscillation around zero when the two correlations have different signs. The corresponding spectra have the form:

$$S_X(f) = \sigma_X^2 \frac{1-\rho(1)^2}{1+\rho(1)^2 - 2\rho(1)\cos(2\pi f)} \tag{28}$$

$$S_X(f) = \sigma_X^2 \frac{(1-b_1^2)(1+b_2)/(1-b_2) - b_2^2}{1+b_1^2+b_2^2 - 2b_1(1+b_2)\cos(2\pi f) - 2b_2\cos(4\pi f)} \tag{29}$$

where b_1 and b_2, are the parameters of the AR(2) scheme $x_t = b_1 x_{t-1} + b_2 x_{t-2} + \varepsilon_t$. This is simplified to $x_t = b_1 x_{t-1} + \varepsilon_t$ for the AR(1) case for which $b_1 = \rho(1)$.

The AR(1) and AR(2) models are both examples of Markov models or "short memory processes" and as such they are members of the larger family of ARIMA (auto regressive integrated moving average) models (Box and Jenkins, 1970), which have been widely in use in hydrology.

Another development in stochastic hydrology has been initiated by the findings of Hurst (1951) when analyzing the long record of water level information from the Nile River and also other long-term geophysical observation series. Hurst studied asymptotic behavior of the range of cumulative departures from the mean for a given sequence

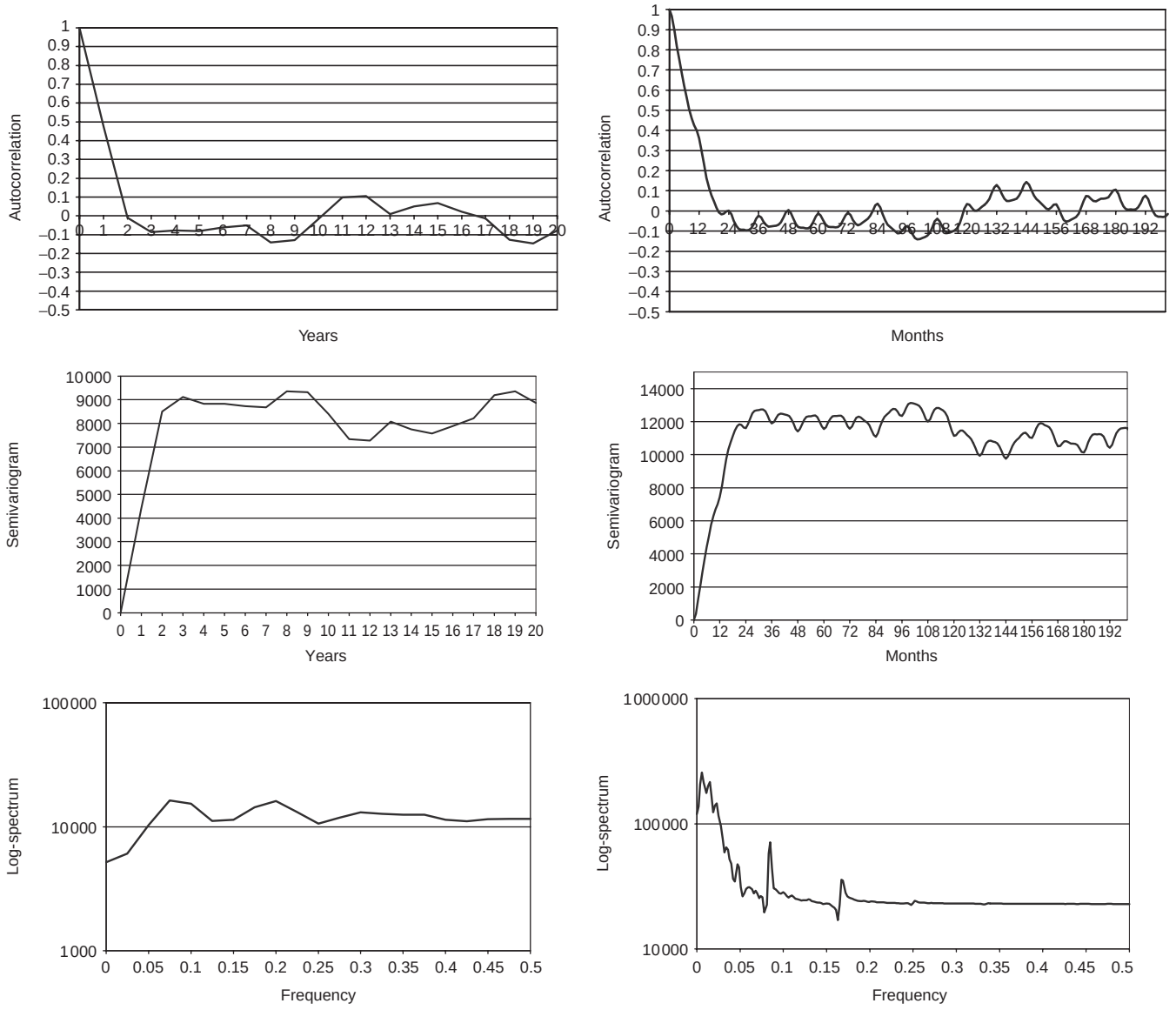


Figure 6 Autocorrelation, semivariograms, and spectral functions for annual and monthly data from the Göta River, Sweden. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of runoff for N years. In case of a Markov model, this statistics will grow as $N^{0.5}$, where N is the number of years of observations. Hurst found that the growth of the rescaled range with the number of years rather followed a relation N^H, with $H > 0.5$, where H is named the *Hurst coefficient*. A scientific discussion followed with a focus on the ability of random models to reproduce the so-called Noah and Joseph effects of natural series, that is, the ability to reproduce extreme extremes and the tendency of long spells of dry and wet years. A natural mechanism inducing a long memory to the system was proposed as a possible explanation to this behavior and a fractional Gaussian noise (fGn) model containing such a "long memory" component was developed by Mandelbrot and Wallis (1968, 1969a,b). fGn is able to generate synthetic data with H different from 0.5. The correlation function and the spectral function respectively have the expressions:

$$\rho(\tau) = \tfrac{1}{2}[(\tau - 1)^{2H} - (\tau + 1)^{2H}] - \tau^{2H} \qquad (30)$$

$$S(f) = cf^{1-2H} \qquad (31)$$

A recent contribution to this discussion is provided by Koutsoyiannis (2002, 2003), who especially criticizes the notion of "deterministic trends" in the model equation (18) and also gives an alternative interpretation of the Hurst phenomenon: "It relies on an 'absence of memory' concept

rather than a 'long memory' concept. The hypotheses proposed is that not only does the system disremember what the value of the process was 100 years ago, but it further forgets what the process mean value at that time was". The idea is thus a *composite random process* with variations at several timescales (Vanmarcke, 1988). Koutsoyiannis (2003) shows that a Markovian underlying process with random fluctuations of the process mean at different scales yields a process very similar to fGn, the composite process being stationary. In the same sense, it can be argued that trends cannot be deterministic, but rather reflect this variability at a range of scales. Looking back at the times series in Figure 1, a trend can for sure be identified in the data for a period of time of, say, 30 years. For the next 30-year period, the trend has changed and continues to change for subsequent 30-year periods.

Are then the periodic components in hydrologic time series to be considered as deterministic ones? Indeed, the astronomic periodic fluctuations originating from the Sun and the Moon ranging from half a day to 10th of thousands of years with the annual cycle being the most important are of a deterministic character and might contribute to the variability of hydrologic records. The strength of these signals to the outer atmosphere is filtered through a complex chain of processes, resulting in an often weak oscillation entering the surface hydrologic system with an intensity that may changeover time. The annual cycle, the seasonal variation, is anyhow strong for most climates. River flow regimes, and related seasonal patterns in precipitation, show large variations around the globe and maybe useful for classification of hydrological regional features. The regime is usually defined as the average seasonal pattern over many years of observations. The assumption is that the hydrologic regime is stable and shows the same average pattern from year to year. There indeed exist flow regimes for which this is true, for example, snowmelt fed regimes in cold climates. However, many flow regimes show instability (Krasovskaia and Gottschalk, 1992; Krasovskaia, 1996, Krasovskaia *et al.*, 1999), that is, the seasonal patterns alternate among several flow regime patterns during individual years in a chaotic way. Climate change accentuates this instability and gives rise to a change over time in the frequency of different seasonal patterns observed at a site.

It is important to underline the difference in working with the statistics of a random variable and that of a random process (here time series). The autocovariance in data, independent of whether it reflects a short or a long memory process, introduces a loss in the precision in parameter estimation and also in biases. Accepting that our process is described by an AR(1) model, the standard error of the mean value is (Hansen, 1971):

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{N}} \left[1 + \frac{2\rho(1)}{N} \frac{N(1-\rho(1)) - (1-\rho(1))^N}{(1-\rho(1))^2} \right]^{1/2} \tag{32}$$

This standard error can be compared with that for the mean of a random variable equation (11). This comparison allows defining the "equivalent number of independent observations", N_e, that is, the number of independent observations that have the same precision in the estimate of the mean:

$$N_e = N \left[1 + \frac{2\rho(1)}{N} \frac{N(1-\rho(1)) - (1-\rho(1))^N}{(1-\rho(1))^2} \right]^{-1} \tag{33}$$

For an AR(1) process, the variance estimate is

$$\tilde{\sigma}_X^2 = \sigma_X^2 \left[1 - \frac{2\rho(1)}{N(N-1)} \frac{N(1-\rho(1)) - (1-\rho(1)^N)}{(1-\rho(1))^2} \right] \tag{34}$$

which is a negatively biased estimate. The fact that data are correlated does thus hamper the process from showing its full range of variability during a short observation period. For an AR(1) process, the characteristic integral scale is $\theta_X = -\Delta t/\ln[\rho(1)] \cong 0.5\Delta t(1+\rho(1))/(1-\rho(1))$, where Δt is the time step. It was noted when discussing the ergodicity concept that the process needs to be observed during an interval that is at least 10 times this scale to secure that the true variability of the process is not underestimated.

Turning back to the Göta River data with a lag one correlation coefficient of 0.476, this means that the 131 years of observations is replaced by 47 years of equivalent independent data. The estimated timescale is $\theta_X = 1.35$ years and the variance is underestimated by about one per cent.

Koutsoyiannis (2003) gives the corresponding standard error estimate of the mean for a "long memory" process:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{N^{1-H}} \tag{35}$$

The corresponding "equivalent number of independent years" would then be

$$N_e = N^{2-2H} \tag{36}$$

A corresponding expression for a "long memory" process for an unbiased estimate of the variance is (Beran, 1994):

$$\tilde{\tilde{\sigma}}_X^2 = \frac{N-1}{N - N^{2H-1}} \sigma_X^2 \tag{37}$$

that is, an underestimation of the real variability of the process.

Random Vector

Data sampled at several fixed sites in space over time with regular intervals: $x(\mathbf{u}_i, t_k)$ at M stations at points

$\mathbf{u}_i$, $i = 1, \ldots, M$ at N points of time, t_k, $k = 1, \ldots, N$ might be considered as a realization of a random vector. This is a common situation in meteorology and hydrology. It is possible to determine the first- and second-order moments at each site and between sites from this data set. A special case is when only one observation at each site is available, common in hydrogeology. It is then not possible to estimate individual moments for a single site. The characterization of variability between sites is done in terms of a semivariogram. The general situation is treated first.

Let the M random variables $X_1, X_2, \ldots, X_M$ denote the elements of the random vector **X.** The mixed second order moment – the covariance B_{ij} –between two elements of X_i and X_j is defined as the expected value of the product of the deviations from respective mean values:

$$\begin{aligned} \mathrm{Cov}[X_i, X_j] = B_{ij} &= E[(X_i - m_i)(X_j - m_j)] \\ &= E[X_i X_j] - m_i m_j \end{aligned} \tag{38}$$

If we divide B_{ij} by $\sigma_i \sigma_j$, the dimensionless correlation coefficient between X_i and X_j is obtained as follows:

$$\rho_{ij} = \rho_{X_i X_j} = \frac{\mathrm{Cov}[X_i, X_j]}{\sigma_i \sigma_j} = \frac{B_{ij}}{\sigma_i \sigma_j} \tag{39}$$

It follows from the definition that $B_{ij} = B_{ji}$, $\rho_{ij} = \rho_{ji}$ and that $|\rho_{ij}| \leq 1$ (and consequently that $B_{ij} \leq \sigma_i \sigma_j$). Independence of two random variables means that there is no correlation, while the opposite is not valid. The correlation coefficient is a measure of the degree of linear dependence. The covariance of M random variables $X_1, X_2, \ldots, X_M$ (which are elements of the vector **X**) can be arranged in a symmetrical M by M covariance matrix $\mathbf{B} = \mathbf{B}_X$:

$$\mathbf{B} = \begin{bmatrix} B_{11} & B_{12} & \ldots & B_{1M} \\ B_{21} & B_{22} & \ldots & B_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ B_{M1} & B_{M2} & \ldots & B_{MM} \end{bmatrix} \tag{40}$$

The Spatial Correlation Function

In the applied situation, first- and second-order sample moments are determined from the observations $x(\mathbf{u}_i, t_k)$ in M stations at points $\mathbf{u}_i$, $i = 1, \ldots, M$ at k points of time, t_k, $k = 1, \ldots, N$. As a first step, the time means can be calculated for each station as

$$\overline{x_i} = \frac{1}{N} \sum_{k=1}^{N} x(u_i, t_k), \quad i = 1, \ldots, M \tag{41}$$

The variance can be obtained as

$$\hat{B}_{ii} = s_i^2 = \frac{1}{N} \sum_{k=1}^{N} x(u_i, t_k)^2 - \overline{x}_i^2, \qquad i = 1, \ldots, M \tag{42}$$

pairwise covariances as

$$\hat{B}_{ij} = \frac{1}{N} \sum_{k=1}^{N} x(u_i, t_k) x(u_j, t_k) - \overline{x}_i \overline{x}_j; \qquad i, j = 1, \ldots, M \tag{43}$$

and pairwise correlation coefficients as

$$r_{ij} = \frac{\hat{B}_{ij}}{(s_i s_j)}; \qquad i, j = 1, \ldots, M \tag{44}$$

The only condition for these calculations is that observations are stationary in time.

A second step would be to investigate the relationship between moments for two points in dependence on the distance between them, direction, and also special physiographic characteristics at these points. Figure 7 shows diagrams illustrating pairwise correlation coefficients r_{ij} in dependence on the distance h_{ij} between the observation points for precipitation events in the Oslo region. The data sample has been divided in dependence on the precipitation type into two parts – frontal and convective, respectively.

A third step in the analysis is a check of the validity of the assumptions. Performing such a check, it is important to bear in mind that moments estimated empirically might have statistical errors. A large scatter in the correlation coefficients' values can be noted in Figure 7, but, in general, these values lie within the boundaries of a confidence interval for a theoretical correlation function. The premises of homogeneity and isotropy are hardly always satisfied. It is common that the correlation structure demonstrates homogeneity, while the covariances do not. An appropriate model in this case will be

$$B_{ij} = \sigma_i \sigma_j \rho(h_{ij}) \tag{45}$$

We can cope with the condition of anisotropy by a simple linear transformation of the coordinate scale, assuming an elliptical form of the direction dependence. The problem of nonhomogeneity in the mean and variance can be handled by means of a respective normalization and standardization of the initial data. This is actually not a complete solution as it is necessary to find an approach for interpolation of the mean and, alternately, mean and variance. These statistical parameters, however, can be expected to have a more even and uniform spatial distribution than the initial observations and their map representation does not absolutely require application of stochastic interpolation methods.

The model equation (45) assumes that it is possible to find an analytical expression $\rho(h_{ij})$ that can be fitted to the ensemble of points in the diagrams in Figure 7. A choice of this correlation function is not totally free, however. The following conditions must be satisfied (Christakos, 1984):

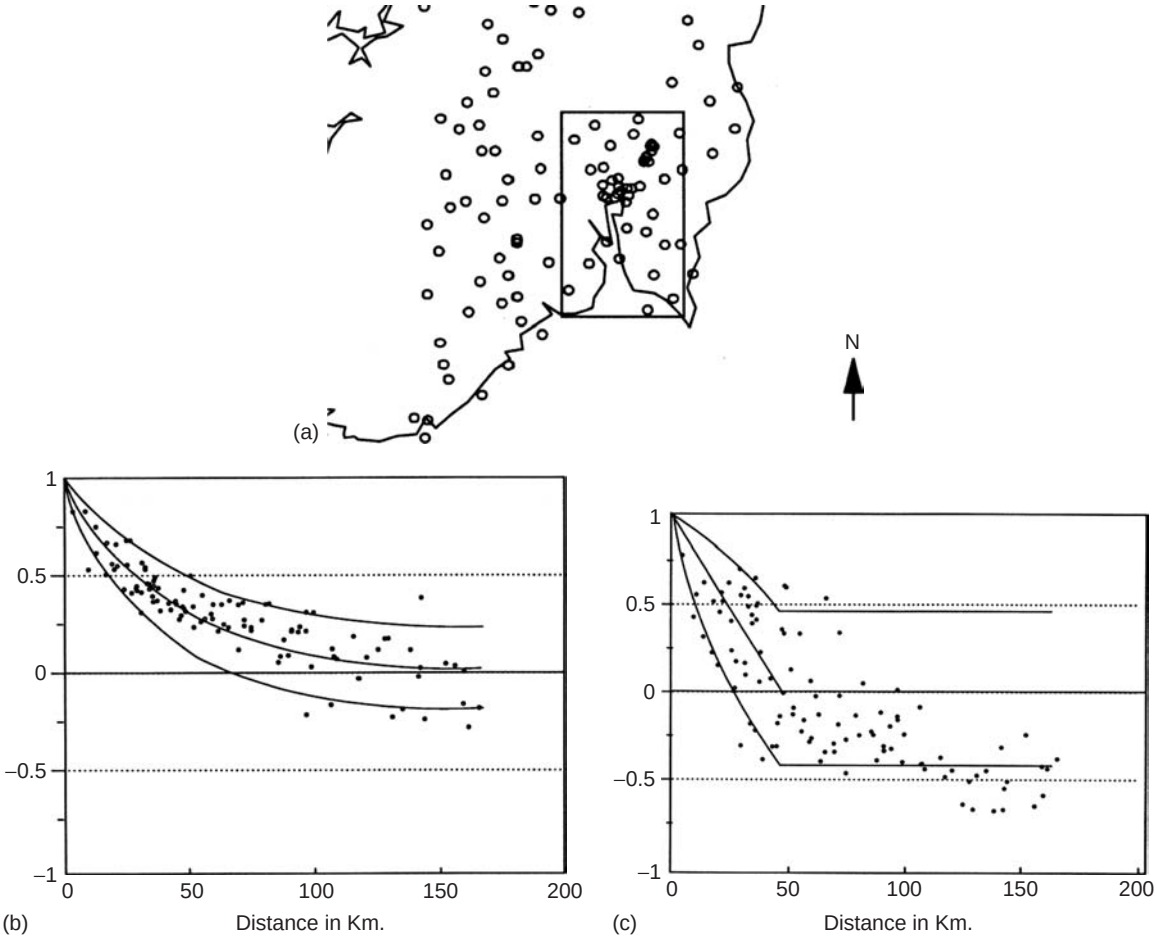


Figure 7 Dependence of the values of pairwise correlation coefficients on the distance between observation points for precipitation events of frontal precipitation (a) and convective precipitation (b). Data from the Oslo region, marked by the rectangle on the map, have been used. The circles on the map show the location of the observation points. (from Skaugen T, Personal communication 1993)

1. The standardized covariance $\rho(h)$ must be a real, even, and continuous function (possibly, except for $h = 0$) for which it is valid for each h that

$$\rho(-h) = \rho(h) \tag{46}$$

   Thus, only functions of positive h can be considered.
2. The standardized covariance $\rho(h)$ always has an upper boundary

$$|\rho(h)| \leq \rho(0) = 1 \tag{47}$$

3. The decay for $h \to \infty$ is determined by the following expression:

$$\lim_{h\to\infty} \frac{\rho(h)}{|h|^{(1-D/2)}} = 0 \tag{48}$$

   where D is the dimension of the vector $\mathbf{u}$ (thus, here $D = 2$).
4. Variances of linear combinations of variables $X(\mathbf{u}_i)$, $i = 1, \ldots, M$, should be positive, which is guaranteed if the standardized covariance function is positively definite, that is,

$$\sum_{j=1}^{M}\sum_{k=1}^{M} \lambda_j \lambda_k \rho(h_{jk}) > 0 \tag{49}$$

   for all M and real coefficients $\lambda_1, \ldots, \lambda_M$ (different from zero).

Matérn (1960) presents different methods to derive "appropriate" isotropic correlation functions. Below follow some frequently used expressions for correlation functions:

$$\rho(h) = \exp(-\alpha^2 h^2) \qquad \text{"Gaussian"} \tag{50}$$

$$\rho(h) = \left(\frac{1+h^2}{\alpha^2}\right)^{-n}, \quad n > 0 \tag{51}$$

$$\rho(h) = \frac{1}{2^{n-1}\Gamma(n)}(\alpha h)^n K_n(\alpha h), \alpha > 0, n > 0 \quad (52)$$

where α and n are parameters and K_n is the modified Bessel function of the second type. The latter expression describes the so-called Matérn class of correlation functions for an isotropic random process (Handcock and Stein, 1993). $\alpha > 0$ is a scale controlling the range of correlation. The smoothness parameter $n > 0$ (which, for the general case, is a real number) directly controls the smoothness of the random field. The following cases are of a special interest:

$$n = \frac{1}{2};\ \rho(h) = \exp(-\alpha h) \quad (53)$$

that is, the exponential function which in one dimension represents a fist-order autoregressive process (Figure 8a);

$$n = 1;\ \rho(h) = \alpha\ hK_1(\alpha\ h) \quad (54)$$

which corresponds to equation (53) in two dimensions (Figure 8b);

$$n = 1\frac{1}{2};\ \rho(h) = (1 + \beta h)\exp(-\alpha h) \quad (55)$$

the linear exponential function which corresponds to a second-order autoregressive process in one dimension (Figure 8c).

As $n \to \infty$, the expression equation (52) approaches the Gaussian function equation (50). This model forms the upper limit of smoothness in the class and will rarely represent natural phenomena because realizations from it are infinitely differentiable. Parameters α and n, in principle, can be determined by means of least square methods for an ensemble of pairwise correlations. In practice, a manual "try and error" fitting is applied, relying totally on visual criteria. Data are usually too sparse to allow identifying the true structure of the studied spatial process. The importance of the choice of theoretical covariance model is revealed only when using the identified covariance structure for simulation of spatial fields. Figure 9 shows the results of simulations with equal spatial integral scales of the model, but with the two extremes of smoothness, exponential, and Gaussian, respectively. The difference is striking and underlines the importance of the choice of theoretical correlation to be used for simulations.

The Semivariogram

In the situation when only one observation $x(\mathbf{u}_i)$ is at hand at each site in space $\mathbf{u}_i$; $i = 1, \ldots, M$, site-specific mean values and standard deviations remain unknown. This is a frequent case in (hydro)geology. The semivariogram

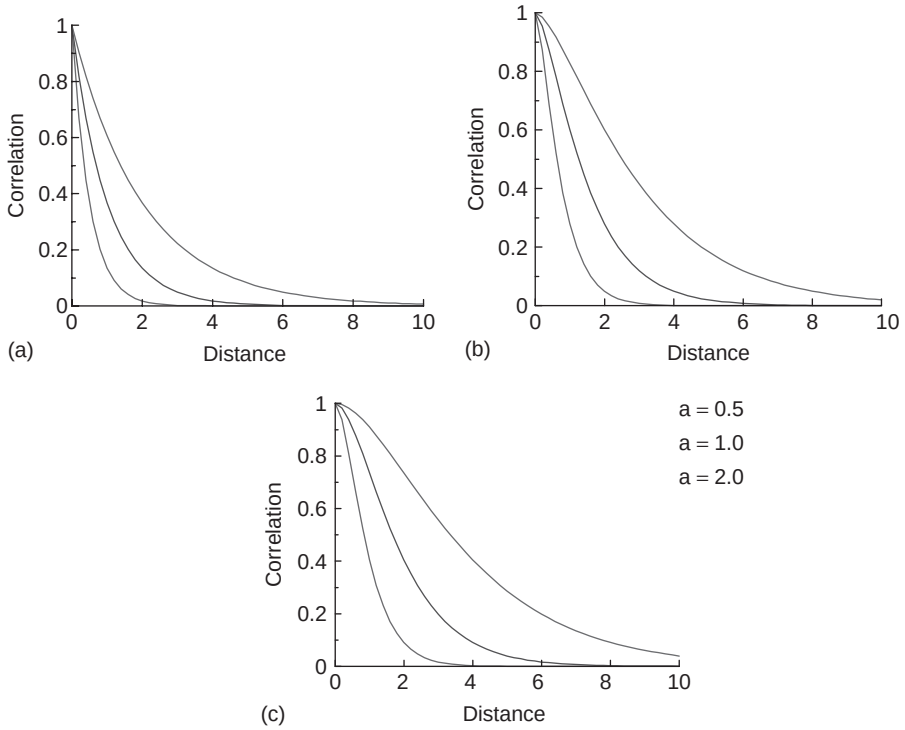


Figure 8 Examples of theoretical correlation functions: exponential function (equation 53) (a); modified Bessel function (equation 54) (b); linear exponential function (equation 55) (c). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

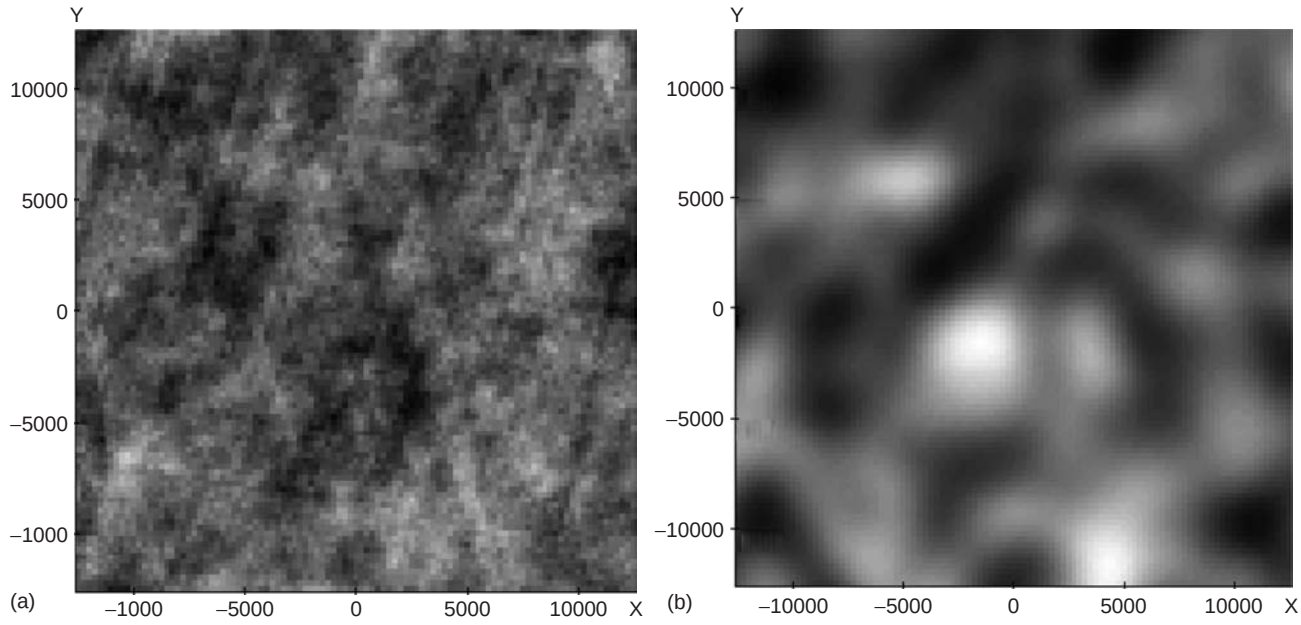


Figure 9 Simulated data in space with the turning band method (a) with an exponential correlogram model and (b) with a Gaussian correlogram model

is the appropriate alternative to the covariance (which assumes these moments to be known) to describe spatial dependency, that is,

$$\gamma(\mathbf{h}) = \frac{1}{2}E[(X(\mathbf{u}_i) - X(\mathbf{u}_i + \mathbf{h}))^2] \tag{56}$$

The sample semivariogram is estimated form empirical data as

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)\in R(\mathbf{h})}^{N} [x(\mathbf{u}_i) - x(\mathbf{u}_j)]^2 \tag{57}$$

where $R(\mathbf{h}) = \{(i, j) : |\mathbf{h}\text{-}\varepsilon| \leq |\mathbf{u}_i\text{-}\mathbf{u}_j| \leq |\mathbf{h} + \varepsilon|\}$ and $N(\mathbf{h})$ is the number of elements in distance class $R(\mathbf{h})$. A diagram that shows $\hat{\gamma}(\mathbf{h})$ and a corresponding value of $\mathbf{h}$ is in a general case a function of vector $\mathbf{h}$, which can depend on both magnitude and direction of $\mathbf{h}$. In the latter case, when semivariograms are different for different directions, the phenomenon studied demonstrates anisotropy. An anisotropic semivariogram has to be transformed into an isotropic one in order to be used (Journel and Huijbregts, 1978). In the following, isotropy will be assumed and, thus, distance can be handled as a scalar h.

It can be expected that the difference $[X(i) - X(i + h)]$ increases with the distance between observations, h. Observations, situated in the vicinity of each other, can be expected to be more alike than those far away. However, in practice, $\gamma(h)$ often approaches some positive value C_0, called the *nugget-effect*, when h approaches zero. This value reveals a discontinuity in the semivariogram in the vicinity of the origin of coordinates at the distance that is shorter than the shortest distance between observation points. This discontinuity is caused by variability in a scale smaller than the shortest distance and also by observation errors.

The procedure of estimation of a theoretical model to an experimental variogram depends on discrete empirical values at specified distances. It can involve a certain degree of subjectivity. Procedures for automatic estimation are least square estimation (LSE), maximum likelihood (ML), and Bayesian estimators (Smith, 2001). Such automatic methods are becoming more frequent in hydrology but, still, subjective trial-and-error fitting dominates. An experimental variogram is very sensitive to errors and uncertainties in data (Gottschalk *et al.*, 1995).

Similar to the premises that have been formulated for a theoretical correlation function, the following conditions should be satisfied for a theoretical semivariogram (Christakos, 1984):

1. A semivariogram $\gamma(h)$ should be a real, even, and continuous function (possibly with the exception of $h = 0$) for which the following condition is valid for any h:
$$\gamma(-h) = \gamma(h) \tag{58}$$
2. The value of this function should follow the following dependency when h approaches infinity:
$$\lim \frac{\gamma(h)}{|h|^2} = 0|h| \longrightarrow \infty \tag{59}$$
3. Variance of linear combinations of variables $X(\mathbf{u}_i)$, $i = 1, \ldots, M$, must be positive, which is satisfied when

– $\gamma(h)$ is conditionally positively definite, that is,

$$\sum_{j=1}^{M}\sum_{k=1}^{M}\lambda_j\lambda_k\gamma(h) > 0 \text{ when } \sum_{j=1}^{M}\lambda_i = 0 \text{ for all } M \tag{60}$$

In this case, it can be noted that there is no demand that $\gamma(h)$ should have some upper boundary as in the case of the correlation (covariance) function, which might be seen as an advantage of using the semivariogram.

Depending on the behavior of the semivariogram for large values of h, theoretical models can be subdivided into two categories: models with a so-called "sill" and those without it. Presence of a sill means that h increases from zero to a specified value, sill, and is constant thereafter (Figure 10). The value of $\gamma(h)$ at this distance h, say, a, is approximately equal to the observation's variance. a is called *range*. The range a is an important characteristic as this distance indicates that observations are correlated within it, while they are independent at larger distances. If the range a is smaller than the shortest distance between observations, a pure nugget-effect is observed, that is, data are independent. The phenomenon studied demonstrates in this case a completely random pattern with respect to the distances between observation points available.

Below, five often used theoretical semivariograms are presented: (i) linear; (ii) spherical; (iii) Gaussian; (iv) exponential; and (v) fractal. The nugget-effect is denoted by C_0, sill is denoted by $C_0 + C_1$ and range by a:

1. Linear (Figure 11a)

$$\gamma(h) = C_0 + B * h0 \le h < a$$
$$\gamma(h) = C_0 + C_1 \; h \ge a \tag{61}$$

where B is the slope for $0 \le h \le a$.

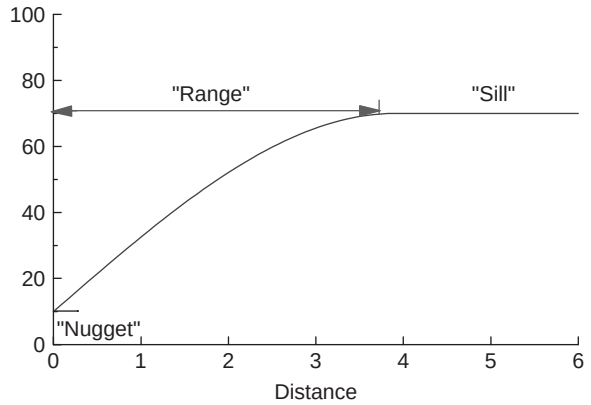


Figure 10 A semivariogram and its parameters. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2. Spherical (Figure 11b)

$$\gamma(h) = C_0 + C_1\left(\frac{3}{2}\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3 \; 0 \le h \le a$$
$$\gamma(h) = C_0 + C_1 h \ge a \tag{62}$$

3. Gaussian (Figure 11c)

$$\gamma(h) = C_0 + C_1\left[1 - \exp\left(\frac{-h^2}{a_0^2}\right)\right] \tag{63}$$

The parameter a_0 is related to the range as $a_0 = 1/(\sqrt{3}a)$, where a is estimated visually as a distance at which the experimental semivariogram becomes stable.

4. Exponential (Figure 11d):

$$\gamma(h) = C_0 + C_1\left[1 - \exp\left(\frac{-h}{a_0}\right)\right] \tag{64}$$

Exponential semivariogram converges asymptotically towards a sill.

5. Fractal (Figure 11e):
Finally, there are models without a sill that describe the phenomenon with, in principle, infinite variance and therefore these models can be associated with phenomena variations at all scales, a fractal process. A theoretical semivariogram model has the following form:

$$\gamma(h) = C_0 + A * h^b; \quad 0 < b < 2 \tag{65}$$

The parameter b must be strictly greater than zero and strictly smaller than 2. It can be shown that it is related to Hurst H, which has been used before for characterization of fractal phenomena, as $b = 2H$.

The existence of an upper boundary, sill, means that the variance is limited. Therefore, the relationship $\gamma(h) = B(0) - B(h)$, where $B(0)$ corresponds to the sill, C_0, is also valid. Furthermore, if the covariance is standardized as $\rho(h) = B(h)/B(0) = 1 - \gamma(h)/B(0)$, the theoretical expressions for spatial correlation function $\rho(h)$ (equations 50–55) can be directly related to those for $\gamma(h)$ above. It can be noted that the Gaussian model equation (63) corresponds to equation (50) and the exponential model (equation 64) to equation (53). No discontinuity for $h = 0$ (nugget) has been included in the correlation functions, however. The choice of models is, thus, somewhat different in meteorology/hydrology and hydrogeology. It is, of course, possible to choose any of the semivariogram models referred to earlier and/or correlation functions with or without a nugget. The only exception is the fractal model equation (65), which does not have a bounded variance.

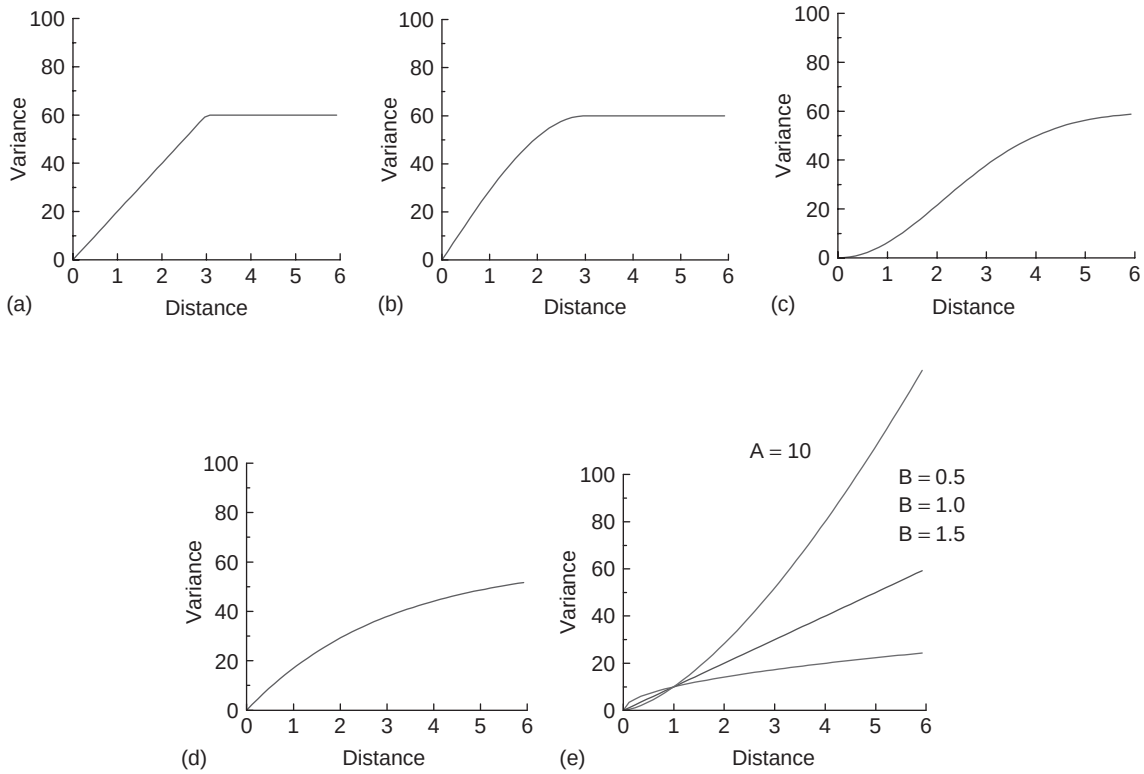


Figure 11 Examples of theoretical semivariograms: (a) linear, (b) spherical, (c) Gaussian, (d) exponential, (e) fractal. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

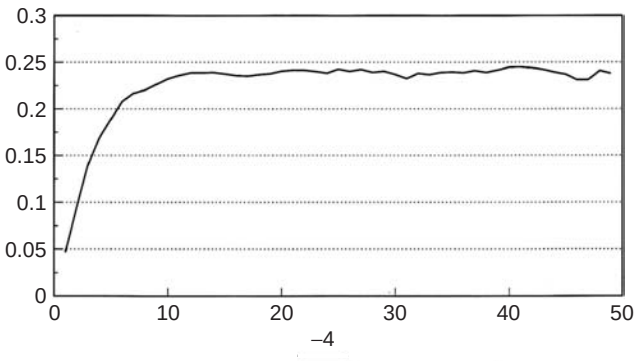


Figure 12 Estimated semivariogram of the ''sand class'' of GPR data (Figure 2) for at tilt of 4 degrees of the geological layers

Figure 12 shows the estimated empirical semivariogram of the "sand class" of GPR data illustrated in Figure 2. The amount of data is for this case extensive (1 00 000 data points) and they are observed in a regular network. That is why the resulting empirical semivariogram is so regular. We note that in case of a geological structure soil classes are described by set functions (present–not present) and need to be studied separately (e.g. Langsholt *et al.*, 1998).

A more common situation is illustrated in Figure 13 with groundwater levels in Gardermoen aquifer northeast of Oslo, Norway. Groundwater levels were recorded in 185 observation points in May 1993. The observations are relatively few, irregular, and clustered, and this has a significant influence on the resulting estimated empirical semivariograms (Figure 14).

The estimated global semivariogram from these data is presented in Figure 14a. Figure 14b shows the four directional semivariograms (note the difference in scale for the variogram between the two figures). Irregular and limited amount of data results in irregular empirical semivariograms. It is an evidence of the uncertainty in specifying spatial dependence in the same manner as the scatter of estimates of spatial correlation in Figure 7.

Spatial-temporal Random Processes

The most general case is a random process in time as well as in space. The observations are thus not realizations of a random vector as in the previous case, but rather a set of time series for each site of observation. In the general case, the joint time-space dynamics of the process under study needs to be considered in terms of two-dimensional time-space covariance functions. Common examples are the development of precipitation over an area at a timescale

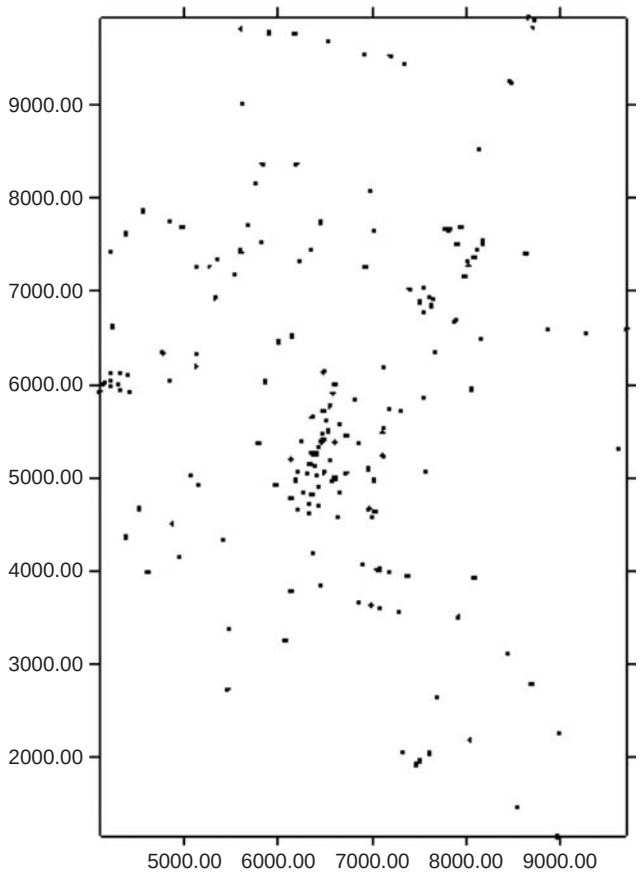


Figure 13 Location of 185 observation points of groundwater levels in the Gardermoen area in May 1993 (Reproduced from Engen, 1995, by permission of University of Oslo)

of minutes or hours and runoff at an hourly timescale for a small catchment. At larger timescales say a week, a month, or a year in most hydrologic situations there exists no dynamic link between the processes in space and time and they can be treated separately by a covariance structure in space and another in time. The methods that are developed below in Section "Karuhnen–Loève expansion" of this article are well suited to handle this latter situation. For the general case, a dynamic spatial-temporal model needs to be formulated. Models for precipitation may serve as examples (e.g. Northrop *et al.*, 1999). For simple linear hydrologic systems expressed in terms of an ordinary or partial differential equation, it is possible to directly derive a theoretical covariance function from these equations (Gottschalk, 1977, 1978). Here, only some simple examples of time-space correlation functions are treated.

For the general formulation of a time-space process, Vanmarcke (1988) distinguishes between three important special types of two-dimensional covariance functions, namely,

- the covariance structure is separable, that is, $\sigma^2\rho(h,\tau) = \sigma^2\rho(h)\rho(\tau)$ and an example is $\sigma^2\rho(h,\tau) = \sigma^2\exp\{-(h/k_1)^2 - (\tau/k_2)^2\}$
- the correlation structure is isotropic, that is, the covariance structure can be expressed in terms of the "radial" covariance function where $r = \sqrt{h^2+\tau^2} : \sigma^2\rho^R(r) = \sigma^2\rho(r,0) = \sigma^2\rho(0,r) = \sigma^2\rho(h,\tau)$
- the covariance structure is ellipsoidal, that is, by appropriate scaling and rotation of the coordinate axes random fields with ellipsoidal covariance structure can be reduced to isotropic random fields.

Gandin and Kagan (1976) suggest a covariance model similar to the second type for use in meteorology and climatology:

$$\mathrm{Cov}[h,\tau] = \sigma^2\rho\left(\left|\left(\frac{h}{v}\right)+\tau\right|\right) \tag{66}$$

where v is a velocity and h/v can be interpreted as a time of travel. Bass (1954) refers to a similar expression for application in turbulence theory.

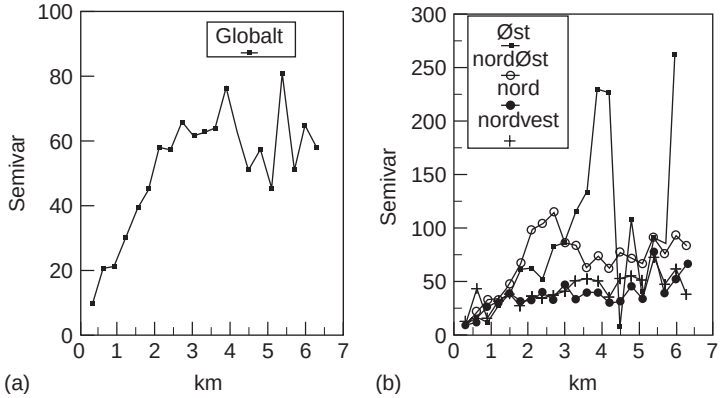


Figure 14 Estimated global semivariogram (a) and four direction variograms (b) (Reproduced from Engen, 1995 by permission of University of Oslo)

KARUHNEN–LOÈVE EXPANSION

Introduction

A technique of expanding a deterministic function $f(t)$, defined on a finite interval $[a, b]$, into a series, based on some special deterministic orthogonal function (e.g. a Fourier series), is well known from mathematics. The problem can be generalized also to two dimensions, where function $f(\mathbf{u})$ is defined for an area Ω (e.g. two-dimensional Fourier series). In a similar manner, a random function can be expanded in terms of random variables and deterministic orthogonal functions. The "proper orthogonal decomposition theorem" (Loève, 1945) states the following: "A random function $X(t)$ continuous in quadratic mean on a closed interval has on this interval an orthogonal decomposition:

$$X(t) = \sum \lambda_n \ \xi_n \ \psi_n \ (t) \tag{67}$$

with

$$E[\xi_m \xi_n] = \delta_{mn}, \tag{68}$$

$$\int \psi_m(t) \ \psi_n(t) \, \mathrm{d}t = \delta_{mn} \tag{69}$$

if, and only if, the $|\lambda_n|^2$ are the proper values (eigenvalues) and the $\psi_n(t)$ are the orthonormalized proper functions (eigenfunctions) of its covariance. Then the series converges in quadratic mean uniformly on I".

The eigenfunctions, which, as noted from the above theorem, are the eigenfunctions of the covariance, are thus determined from the following equation (a Fredholm's integral equation of the first type):

$$\int B(t, t') \ \psi_n(t') \, \mathrm{d}t' = |\lambda_n|^2 \ \psi_n(t) \tag{70}$$

Furthermore, the covariance can be written as a series expansion

$$\mathrm{Cov}[X(t) \ X(t')] = B(t, t') = \sum |\lambda_n|^2 \ \psi_n(t) \ \psi_n(t') \tag{71}$$

which for the variance reduces to the simple expression

$$\mathrm{Var}[X(t)] = B(t, t) = \sum |\lambda_n|^2 \tag{72}$$

The ξ_n s are random variables and are derived from the relation

$$\lambda_n \xi_n = \int X(t) \ \psi_n(t) \, \mathrm{d}t \tag{73}$$

The covariance between the random function $X(t)$ and ξ_n is

$$\mathrm{Cov}[X(t) \ \xi_n] = \lambda_n \ \psi_n(t) \tag{74}$$

The representation equation (67) of a random function is widely used under the name "Karhunen–Loève expansion". It appears to have been introduced independently by a number of scientists (see Lumley, 1970): Kosambi (1943), Loève (1945), Karhunen (1946), Pougachev (1953) and Obukhov (1954). In case of normally distributed data, the statistical orthogonality equation (68) is equivalent to independence and the projection equation (73) is equivalent to conditioning. Karhunen–Loève expansion is used for analyzing data in terms of a "spectral" representation for reconstruction and simulation of data. It might also be an effective tool for dimensionality reduction of large data sets to eliminate redundant information.

Harmonic (Spectral) Analysis

In physics, the most important orthogonal decomposition is the harmonic one, for, loosely speaking, it yields "amplitudes", and hence "energies", corresponding to various parts of the "spectrum" of the random function. Following Loève's strict formulation, the random function has an imaginary as well as a real part. A more physical approach avoids complex number algebra (Vanmarcke, 1988). In this latter case, the stationary random function $X(t)$ is expressed as a sum of its mean $m = c_0/2$ and $2K$ sinusoids with positive and negative frequencies $f_k = \pm k f_1$, random amplitudes c_k and phase angles θ_k, $k = 1, \ldots, K$.

$$X(t) = \frac{1}{2} c_0 + \sum_{k=-K}^{K} c_k \cos(2\pi f_k t + \theta_k) \tag{75}$$

All random amplitudes and phase angles are mutually independent random variables and the phase angles are uniformly distributed over $[0,2\pi]$. Every single term in the sum has a mean zero and the variance $\sigma_k^2 = (1/2)E[c_k^2]$., and the total variance

$$\mathrm{Var}[X(t)] = \sum \frac{1}{2} E[c_k^2] \tag{76}$$

Equations (75) and (76) have there direct parallels to equations (67) and (72) above. Generalizing to a continuous spectrum for a homogeneous process, that is, with $B(t, t') = B(|t - t'|)$, we derive the two-sided spectral function equation (16). It can be rewritten with the help of known trigonometric relationships and taking into consideration that $S(f)$ is an even function, as

$$S(f) \, \mathrm{e}^{i2\pi f t} = \int_{-\infty}^{\infty} B(|t' - t|) e^{i2\pi f t'} \, \mathrm{d}t' \tag{77}$$

This expression can be compared with equation (70). Similar results can be obtained if $[a, b]$ is finite, as long as the spectrum $S(f)$ is a rational function. Analytical

expressions for this case can be found in Davenport and Root (1958) and Fortus (1973).

Wavelet Analysis

In the Fourier series representation, the orthonormal base function $\psi_n(t)$ is generated by dilation of a single function $\psi(t) = e^{it}$, that is, $\psi_n(t) = \psi(nt)$. For any integer n with large absolute value, the wave has high frequency, and for n with small absolute value, the wave has low frequency. So every function is composed of waves with different frequencies. The sinusoidal function is defined on one period of 2π and the condition for the existence of a series expansion is that the random function is absolute integrable over this period.

$$\int_0^{2\pi} |X(t)|^2 \, \mathrm{d}t < \infty \tag{78}$$

In case of wavelets, the point of departure is also an orthogonal decomposition in accordance with equation (67) (Chui, 1992). The difference first of all lies in the fact that the random function is defined on the real line and that the random function thus satisfies the condition

$$\int_{-\infty}^{\infty} |X(t)|^2 \, \mathrm{d}t < \infty \tag{79}$$

The two function spaces are quite different since, in particular, the local average of every function must decay to zero at $\pm\infty$ and the sinusoidal (wave) functions do not satisfy this condition. Waves that can satisfy this condition should decay to zero at $\pm\infty$, and, for all practical purposes, the decay should be fast. Small waves or "wavelets" are thus appropriate. It is preferred to have one single generating function, like in Fourier series (mother wavelet or analyzing wavelet $\psi(t)$). But, if the wavelet has a very fast decay, how can it cover the whole real line? The answer is to allow shifts along the real line. The power of two is used for frequency partitioning

$$\psi(2^j t - k) \tag{80}$$

It is obtained from a single wavelet function by binary dilation (i.e. dilation by 2^j) and dyadic translation (of $k/2^j$). Normalization results in

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \tag{81}$$

The normalizing equation corresponding to equation (69) has the form

$$\int_{-\infty}^{\infty} \psi_{j,k}(t)\psi_{l,m}(t) \, \mathrm{d}t = \delta_{jl} \, \delta_{km} \tag{82}$$

The series expansion is written as

$$X(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \beta_{j,k}\psi_{j,k}(t) \tag{83}$$

The coefficients in the series expansion are determined from

$$\beta_{j,k} = \int X(t) \; \psi_{j,k}(t) \, \mathrm{d}t \tag{84}$$

A simple example is the Haar function

$$\psi_H(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2} \\ -1 & \frac{1}{2} \le t < 1 \\ 0 & \text{else} \end{cases} \tag{85}$$

Similar to the harmonic analysis, we can turn over to a continuous representation and define a wavelet transform as (Chui, 1992):

$$W(\tau, s) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{|s|}} \psi * \left(\frac{t - \tau}{s} \right) \mathrm{d}t \tag{86}$$

where $\psi^*(t)$ is the complex conjugate of ψ (t). The inverse transform of equation (86) for reconstruction of $x(t)$ is written down as

$$X(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(\tau, s) \frac{1}{s^2} \psi \left(\frac{t - \tau}{s} \right) \mathrm{d}\tau \, \mathrm{d}t \tag{87}$$

where C_ψ is a constant of admissibility, which depends on the wavelet used and needs to satisfy the condition:

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} \mathrm{d}\omega < \infty \tag{88}$$

($\hat{\psi}(\omega)$ is the Fourier transform of ψ (t)).

Basic works introducing wavelets are those by Grossman and Morlet (1984) and Meyer (1988). Wavelet decomposition has found many applications in, for instance, image processing and fluid dynamics and turbulence. Wavelets also provide a convenient tool for studying scaling characteristics of a random process (Mallat, 1989; Wornell, 1990; Kumar and Foufoula-Georgiou, 1993). A simple example from Feng (2002) illustrates the application to hydrologic data (Figure 15). In this case, a quadratic spline function is used as mother wavelet to be able to reconstruct and simulate observed periodic hydrological time series.

Principal Component Analysis (pca)

The basic matrix equation for the principal component analysis of a random function is expressed as

$$\mathbf{B}_X \Psi = \Psi \Lambda \tag{89}$$

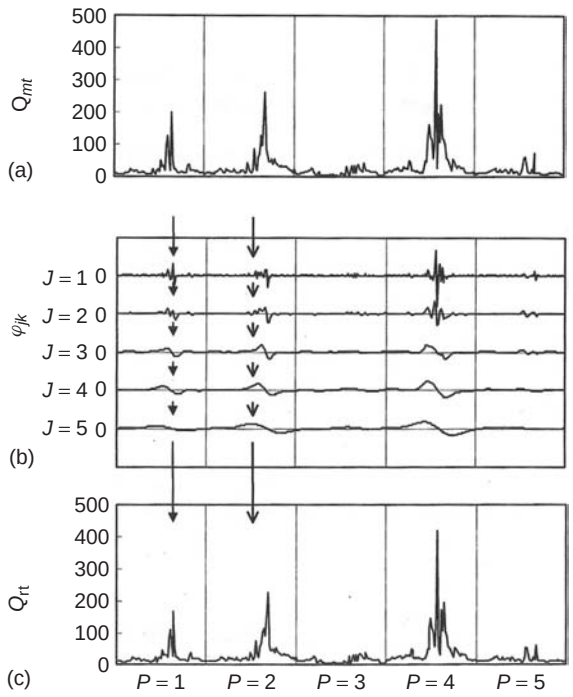


Figure 15 Illustration of traditional wavelet decomposition and reconstruction: (a) shows measured time series with five main periods; (b) decomposed wavelet series $\psi_n(t)$; and (c) reconstructed series (from Feng, 2002)

where in the general case $\mathbf{B}_X$ is a covariance matrix, Ψ a coefficient matrix of eigenvectors, and Λ a diagonal matrix of eigenvalues. Each M by M symmetrical positively definite covariance matrix $\mathbf{B}_X$ has a set of M positive eigenvalues. Furthermore, there exists a linear transformation $\mathbf{Z} = \Psi^{\mathbf{T}}\mathbf{X}$ of the original observation matrix $\mathbf{X}$, which has a diagonal covariance matrix $\mathbf{B}_Z$. $\Psi^{\mathbf{T}}$ is an M by M coefficient matrix representing the eigenvectors of $\mathbf{B}_X$. The variables $\mathbf{Z}$ are named *Principal Components*. The observation matrix $\mathbf{X}$ can now be expressed as a linear combination of the principal components:

$$\mathbf{X} = \Psi\ \mathbf{Z} \tag{90}$$

The coefficient matrix Ψ is orthogonal, that is, $\Psi\Psi^{\mathrm{T}} = \mathbf{I}$. The principal components $\mathbf{Z}$ have the following covariance matrix:

$$\mathbf{B}_Z = \frac{1}{n}\mathbf{Z}\ \mathbf{Z}^{\mathrm{T}} = \Psi^{T}\ \mathbf{B}_X\Psi = \Lambda \tag{91}$$

where $\mathbf{B}_X$ is, as before, the covariance matrix of $\mathbf{X}, \mathbf{B}_X = (1/n)\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and Λ is the diagonal matrix of eigenvalues $\lambda^2{}_j$, $j = 1, \ldots, M$. If equation (90) is written as a sum:

$$x_{jt} = \sum_{k=1}^{M} \psi_{jk} z_{kt} = \sum_{k=1}^{M} \psi_{jk}\lambda_k z'_{kt};$$
$$j = 1, \ldots, M; t = 1, \ldots, n \tag{92}$$

where z'_k is the normalized values of z_k with respect to its variance λ_k. Parallels to equation (67) can be seen clearly by a simple exchange of symbols. From equation (91), we have

$$\frac{1}{n}\sum_{t=1}^{n} z_{jt}z_{kt} = \lambda_k^2\delta_{jk};\ \frac{1}{n}\sum_{t=1}^{n} z'_{jt}z'_{kt} = \delta_{jk}j, k = 1, \ldots, M \tag{93}$$

and using the condition of orthogonality of the coefficient matrix Ψ we get

$$\frac{1}{M}\sum_{l=1}^{M} \psi_{jl}\psi_{lk} = \delta_{jk}; \quad j, k = 1, \ldots, M \tag{94}$$

Finally, multiplying equation (91) by Ψ from the left yields

$$\sum_{l=1}^{M} B_{jk}\psi_{lk} = \lambda_j\psi_{jk},\ j = 1, \ldots, M \tag{95}$$

Also, here there are parallels to equations (68), (69), and (70), respectively, if a symbol change is done.

The method of pca representation is usually carried out in terms of the solution of the matrix equation (89) of very general applicability. Principal component analysis (factor analysis) has its root in psychometrics. The classical work is the one by Hotelling (1933). The generality of the method might be a strength for many applications, but with caution. Hydrological time series data are as a rule collected at regular time intervals. For this case, we can apply the matrix equation as a simple approximation of the more general equation (70). In the case of application to spatial data in meteorology, as pointed out by Buell (1971), there are very strong geometrical elements (in the general case, the covariance matrix represents covariances between irregularly spaced observation points) that can be advantageous, but which are missing in the matrix formulation. Hydrological applications have, as a rule, the same strong geometrical elements and $\mathbf{B}_X$ is usually the covariance matrix (equation 40) with elements $B_{ij} = \mathrm{E}[(x_i - m_i)(x_j - m_j)]$, $i, j = 1, \ldots, M$, covariances between values x_i measured at point $\mathbf{u}_i$ and x_j measured at point $\mathbf{u}_j$.

It is found appropriate to differ between situations when geometrical aspects of the problem are strong and when they are not. In the latter case, the pca matrix formulation equation (89) is appropriate. In the other case, the point of departure is equation (70) and, for discrete data, a numerical solution of this equation should be developed. The name for this situation that will be used here is empirical orthogonal functions (eof).

Empirical Orthogonal Functions (eof)

The notion of *empirical orthogonal functions* was first introduced in the classical work by Lorenz (1956). Other

earlier applications of this technique in meteorology are those by Obukhov (1960) and Holmström (1963) (without using the name "eof"). Already Lorenz mentions the parallel in the problem with that of factor analysis (principal component analysis) used by psychologists quoting the classical work by Hotelling (1933). The point of departure for the development of this technique for Lorenz was dimensionality reduction, that is, getting rid of the large amount of redundant information contained in meteorological data. In psychology, on the other hand, the central point was to interpret psychological tests into observed behavior of their patients. Today both approaches are used in meteorology and climatology as well as in hydrology and a difference can be traced in the interpretation of the empirical functions – are those only mathematical constructions or can they be interpreted in some process oriented way. In the latter case, this has led forward to the use of rotations of the principal components yielding possibilities for a better interpretation of results (*see* Richman, 1986 and Jolliffe, 1990, 1993 for an overview). Long unfruitful discussions about the possible distinction between pca and eof can be found in the climatological literature. One such distinction should be related to how the normalization is performed.

The point of departure in the works by Lorenz and Holmström is a random process $X(\mathbf{u},t)$ that develops in space $\mathbf{u} = (u_1, u_2)$ over a certain domain in space Ω and time t. Equation (70) is for this case generalized to a process in space:

$$\int_{\Omega} B(\mathbf{u}, \mathbf{u}')\ \psi_n(\mathbf{u}')\, \mathrm{d}\mathbf{u}' = |\lambda_n|^2\ \psi_n(\mathbf{u}) \tag{96}$$

It is important to emphasize that this integral equation formulation is the appropriate one for problems like in meteorology and hydrology dealing with a random function X on a continuum in space. The geometrical relations involving the domain of integration and the relations between the points $\mathbf{u}_i$, $i = 1, \ldots, M$ are completely ignored in the matrix formulation. The fact that function values are obtained from measurements at discrete points (perhaps sparsely located) is a practical limitation to the numerical solution of the problem (e.g. Obled and Creutin, 1986).

$X(\mathbf{u},t)$ is expanded into double orthogonal series of the form

$$X(\mathbf{u}, t) = \sum_n z_n(t)\ \psi_n(\mathbf{u}) \tag{97}$$

where eigenfunctions $\psi_j(\mathbf{u})$ and $\psi_k(\mathbf{u})$ as before are analytically orthogonal:

$$\int_{\Omega} \psi_k(\mathbf{u}) \psi_j(\mathbf{u})\, \mathrm{d}\mathbf{u} = \delta_{kj} \tag{98}$$

that is, $\psi_k(\mathbf{u})$ can be regarded as a deterministic function. It is determined by numerical solution of equation (96). The functions $\psi_k(\mathbf{u})$ depend both on the covariance function and the area Ω.

$z_j(t)$ and $z_k(t)$, on the other hand, are statistically orthogonal or uncorrelated:

$$E[z_k(t) z_j(t)] = \delta_{kj} \lambda_k^2 \tag{99}$$

where δ_{kj} is Kronecker's delta and λ_k^2 is the eigenvalue as before.

The function $z_k(t)$ is obtained as

$$z_k(t) = \int_{\Omega} X(\mathbf{u}, t) \psi_k(\mathbf{u})\, \mathrm{d}\mathbf{u} \tag{100}$$

that is, by means of projection of the realization at time t on the k-th eigenfunction.

For a given analytical expression for the autocorrelation function, eigenfunctions corresponding to it can be found. Fortus (1975) and Braud (1990) show an analytical solution for the case when Ω is a circle. $B(\mathbf{u},\mathbf{v})$ can be written as a series expansion:

$$B(\mathbf{u}, \mathbf{v}) = \sum_n |\lambda_n|^2 \psi_n(\mathbf{u}) \psi_n(\mathbf{v}) \tag{101}$$

in correspondence with equation (71). The expansion equation (97) can be truncated till, say, N terms as

$$\hat{X}_N(\mathbf{u}, t) = \sum_{n=1}^{N} z_n(t) \psi_n(\mathbf{u}) \tag{102}$$

which minimizes the variance in the estimation error of X by $\hat{X}_N$:

$$E\left[\int_{\Omega} \left\{X(\mathbf{u}, t) - \hat{X}_N(\mathbf{u}, t)\right\}^2\right] \mathrm{d}\mathbf{u} \tag{103}$$

and which obtains the value $\sum_{n=N+1}^{\infty} \lambda_k^2$. In case of high redundancy in the data the expansion equation (102) converges rapidly, which indicates a possibility of truncating the series expansion after rather few terms. This is the idea behind the use of eof for dimensionality reduction. The functions $z_k(t)$, often called *amplitude functions*, represent time series that are not linked to any specific points of the domain Ω. On the other hand, they can be used to construct $X(\mathbf{u}_k, t)$ for a point $\mathbf{u}_k$ if the functions $\Psi_n(\mathbf{u}_k)$, $n = 1, \ldots, N$ are known at this point. Figure 16 illustrates the principle where the eof method is applied to monthly runoff data from the Rhône basin in France. Amplitude functions are shown as well the results of prediction of the runoff pattern for independent stations to the right (Sauquet *et al.*, 2000). Figure 3 in the introduction of this article shows a prediction of the monthly flow patterns for the whole river system.

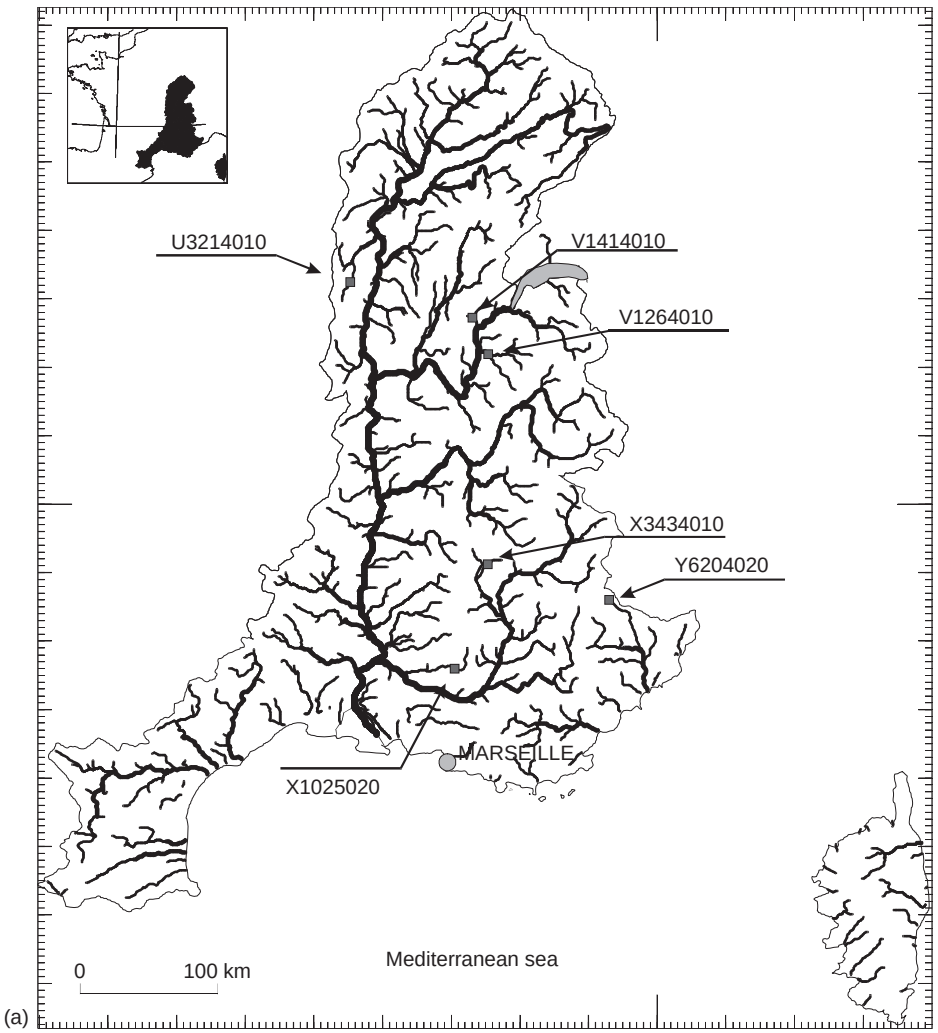


Figure 16 Example of spatial interpolation of monthly runoff patterns for the Rhone basin in France (a). The first six amplitude functions are shown in (b) and the result of prediction of the runoff pattern for independent stations in (c) (Reproduced from Sauquet *et al.,* 2000 by permission of the European Geosciences Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CONCLUDING REMARKS

In the introductory part of this article, it was adopted to use a *partial characterization* of the random process under study in accordance with three different schemes:

1. Characterization by distribution function (one dimensional).
2. Second-moment characterization.
3. Karhunen–Loève expansion, that is, a series representation in terms of random variables and deterministic functions of a random process.

This should not be understood so that one replaces the other. On the contrary, these three schemes for partial characterization complement each other. All methods of analyzing variability developed in this article are applied in practice without consideration of the parent distribution of the data. On the other hand, all these methods have a strong theoretical base if normality can be assumed. For instance, as already noted, normally distributed data weak homogeneity equals strict homogeneity. For normally distributed data, statistical orthogonality is equivalent to independence and a projection onto a system of orthogonal axes is equivalent to conditioning. Furthermore, the assumption of normality opens up for related statistical tests. Analyzing the distribution function of the data is therefore a logical first step (after "looking at data"). If the data are far from normally distributed, it might be worthwhile to utilize a transformation to normal. The following transformations to normality

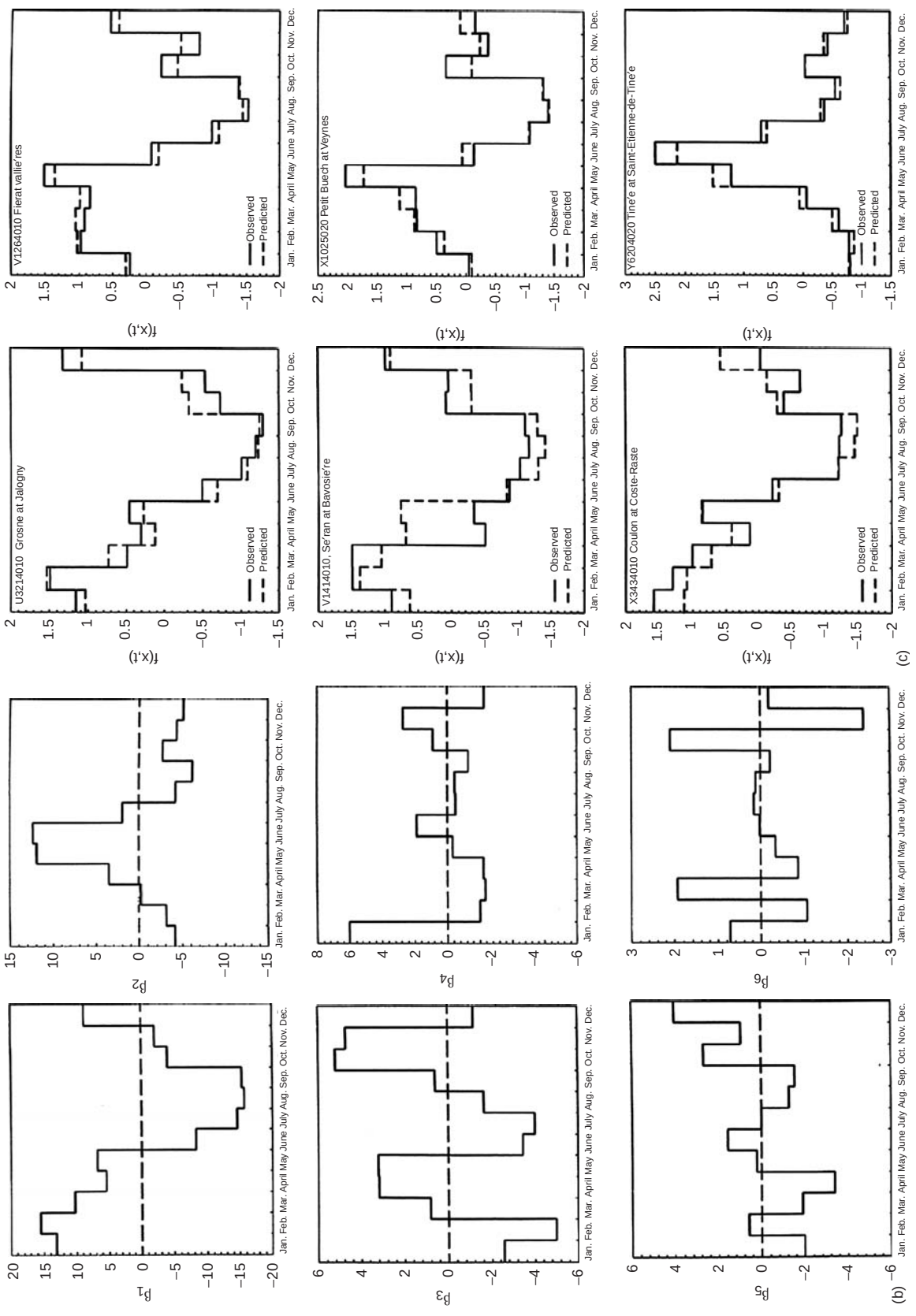


Figure 16 (*Continued*)

are often used in hydrology: (i) ln(x) in case of lognormally distributed data; (ii) The cube root $(x/\overline{x})^{1/3}$ (Wilson and Hilferty, 1931) transformation in case of gamma distributed data; and (iii) the more general power transformation $[(x+c)^h - 1]/h$, where h and c are parameters (Box and Cox, 1964).

The characterization by second moments allows establishing the structure of the data in space and time and its scale of variability. It also gives the possibility of testing basic hypothesis of homogeneity and stationarity. By means of normalization and standardization, data can be transformed into new data sets owing these properties. The characterization by series representation in its turn assumes homogeneity with respect to the variance–covariance function. It is as such a tool for analyzing spatial-temporal variability relative to the first- and second-order moments in terms of new sets of common orthogonal random functions.

The conclusions is thus that the approaches developed here form logical steps in a sequential analysis of variability: (i) looking at data, exploring data; (ii) analyzing the distribution of data; (iii) analyzing first- and second-order moments; (iv) analyzing data by means of series expansion.

REFERENCES

Bass J. (1954) *Space and Time Correlations in a Turbulent Fluid, Part I*, University of California Press: Berkely and Los Angeles, p. 83.

Beran J. (1994) *Statistics for Long Memory Processes*, *Vol. 61 of Monographs on Statistical and Applied Probability*, Chapman and Hall: New York.

Box G.E.P. and Cox D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, **26**, 211.

Box G.E.P. and Jenkins G.M. (1970) *Time Series Analysis, Forecasting and Control*, Holden Day: San Fransisco.

Braud I. (1990) *Etude Méthodologique de l'Analyse en Composantes Principales de Processes Bidiminsionelles*, Doctorat INPG: Grenoble.

Buell E.C. (1971) Integral equation representation for factor analysis. *Journal of the Atmospheric Sciences*, **28**, 1502–1505.

Chow V.T. (1954) The log-probability law and its engineering application. *Proceedings of ASCE*, Vol. 80, separate.

Christakos G. (1984) On the problem of permissible covariance and variogram models. *Water Resources Research*, **20**(2), 251–265.

Chui C.K. (1992) *An Introduction to Wavelet*, Academic Press: Boston.

Davenport W.B. and Root W.L. (1958) *An Introduction to the Theory of Random Signal and Noise*, McGraw-Hill.

Engen T. (1995) Stokastisk interpolasjon av grunnvannspeilet i israndsdeltaet på Gardermoen. (in Norwegian). *Hovedfagsoppgave i Hydrologi*, Institutt for Geofysikk, Universitet i Oslo.

Feng G. (2002) A method for simulation of periodic hydrological time series using wavelet transform. In *Hydrological Models for Environmental Management*, *NATO Science Series, 2. Environmental Security – Vol. 79*, Bolgov V., Gottschalk L., Krasovskaia I. and Moore R.J. (Eds.), Kluwer Academic Publishers: Dordrecht.

Fennessey N.M. and Vogel R.M. (1990) Regional flow duration curves for ungauged sites in Massachusetts. *Journal of Water Resources Planning and Management*, **116**(4), 530–549.

Fortus M.I. (1973) Statistically orthogonal functions of a finite interval of a stochastic process (in Russian). *Fizika Atmosfery I Okeana*, **9**(1), 34–46.

Fortus M.I. (1975) Statistically orthogonal functions of stochastic fields defined for a finate area (in Russian). *Fizika Atmosfery I Okeana*, **11**(11), 1107–1112.

Foster A. (1924) Theoretical frequency curves and their application to engineering problems. *Transactions of the American Society of Civil Engineers*, **87**, 142–173.

Foster A. (1933) Duration curves. *Transactions of the American Society of Civil Engineers*, **99**, 1213–1267.

Gandin L.S. and Kagan P.L. (1976) *Statistical Methods for Interpretation of Meteorological Observations (in Russian)*, Gidrometeoizdat: Leningrad.

Gottschalk L. (1977) Correlation structure and time scale of simple hydrological systems. *Nordic Hydrology*, **8**, 129–140.

Gottschalk L. (1978) Spatial correlation of hydrologic and physiographic elements. *Nordic Hydrology*, **9**, 267–276.

Gottschalk L., Krasovskaia I. and Kundzewicz Z.W. (1995) Detecting outliers in flood data with geostatistical methods. In *New Uncertainty Concepts in Hydrology and Water resources*, Z.W. Kundzewicz (Ed.), Cambridge University Press: pp. 206–214.

Grossman A. and Morlet J. (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, **15**(4), 723–736.

Handcock M.S. and Stein M.L.S.tein (1993) A Bayesian analysis of kriging. *Technometrics*, **35**(4), 403–410.

Hansen E. (1971) Analyse af hydrologiske tidsserier. (in Danish) Polyteknisk forlag, København.

Hazen A. (1914) The storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, **77**, 1542–1669.

Holmes M.G.R., Young A.R., Gustard A. and Grew R. (2002) A region influence approach to predicting flow duration curves within ungauged basins. *Hydrology and Earth System Sciences*, **6**(4), 721–731.

Holmström I. (1963) On a method for parametric representation of the state of the atmosphere. *Tellus*, **15**(2), 127–149.

Hotelling H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417–441 498–520.

Hurst H.E. (1951) Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 770–799.

Jolliffe I.T. (1990) Principal component analysis: a beginner's guide – I. Introduction and application. *Weather*, **45**, 375–382.

Jolliffe I.T. (1993) Principal component analysis: a beginner's guide – II. Pitfalls, myths and extensions. *Weather*, **48**, 246–252.

Journel A. and Huijbregts C.hJ. (1978) *Mining Geostatistics*, Academic Press: New York.

Karhunen K. (1946) Zur spektraltheorie stochastic prozesse. Annales Academiae Scientarium Fennicae Series A I. *Mathematica-Physica*, **34**, 1–7.

Kendall M., Stuart A. and Ord J.K. (1987) *Kendall's Advanced Theory of Statistics (Ch. 10 Standard Errors)*, Charles Griffin: London.

Khinchin A.I. (1949) *Mathematical Foundations of Statistical Mechanics*, Dover: New York.

Khinchin A.Ya. (1934) Korrelationstheorie der stationären stochastichen Prozesse, *Math. Ann.* **109**(4), 604–615.

Kolmogorov A.N. (1941) The local turbulence structure of an incompressible viscous liquids for very big Reynolds numbers (in Russian). *Doklady Academii Nauk SSSR*, **30**(4), 299–303.

Kosambi D.D. (1943) Statistics in function space. *Journal of Indian Mathematical Society*, **7**, 76–88.

Koutsoyiannis D. (2002) The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal*, **47**(4), 573–595.

Koutsoyiannis D. (2003) Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences Journal*, **47**(4), 573–595.

Krasovskaia I. (1996) Sensitivity of the stability of river flow regimes to small fluctuations in temperature. *Hydrological Sciences Journal*, **41**(2), 251–264.

Krasovskaia I. and Gottschalk L. (1992) Stability of river flow regimes. *Nordic Hydrology*, **23**, 137–154.

Krasovskaia I., Gottschalk L. and Kundzewicz Z.W. (1999) Dimensionality of Scandinavian river flow regimes. *Hydrological Sciences Journal*, **45**(5), 705–723.

Kritskij S.N. and Menkel M.F. (1946) On models for studies of random variations in river runoff (in Russian). *Runoff and Hydrological Calculations*, Ser. IV, Vyp. 29, Gidrometeoizdat: Leningrad.

Kumar P. and Foufoula-Georgiou E. (1993) A new look at rainfall fluctuations and scaling properties of spatial rainfall using orthogonal wavelets. *Journal off Allied Meteorology*, **32**, 209–222.

Langsholt E., Kitteröd N.-O. and Gottschalk L. (1998) Development of 3-dimensional hydrostratigraphical models based on soft and hard data. *Ground Water*, **36**(1), 104–111.

Loève M. (1945) Fonctions aleatoire de second ordre. *Comptes Rendus de l'Academie des Sciences, Paris*, **1**, 220.

Lorenz E.N. (1956) *Empirical Orthogonal Functions and Statistical Weather Prediction*, Statistical Forecasting Project, Scientific Report No. 1, MIT Department of Meteorology: Cambridge.

Lumley J.L. (1970) *Stochastic Tools in Turbulence*, Academic Press: New York and London.

Mallat S.G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.

Mandelbrot B.B. and Wallis J.R. (1968) Noah, Joseph and operational hydrology. *Water Resources Research*, **4**(5), 909–918.

Mandelbrot B.B. and Wallis J.R. (1969a) Computer experiments with fractional Gaussian noises. Part 1: averages and variances. Part 2: Rescaled ranges and spectra. Part 3: mathematical appendix. *Water Resources Research*, **5**(1), 228–241.

Mandelbrot B.B. and Wallis J.R. (1969b) Some long run properties of geophysical records. *Water Resources Research*, **5**(1).

Matérn B. (1960) Spatial variation. *Meddelande från Statens Skogsforskningsinstitut, band*, **49**(5).

Matheron G. (1965) *Les Variables Régionalisées et Leur Estimation*, Masson: Paris.

Meyer Y. (1988) Ondelettes et Operateurs. Hermann.

Mimikou M. and Kaemaki S. (1985) Regionalization of flow duration characteristics. *Journal of Hydrology*, **82**, 77–91.

Mosley M.P. and McKerchar A.I. (1992) Chapter 8 Streamflow. In *Handbook of Hydrology*, Maidment D.R. (Ed.), MCGraw-Hill: NewYork, p. 8.27.

National Research Council (1991) *Opportunities in the Hydrologic Sciences*, National Academy Press: Washington DC.

Northrop P.J., Chandler R.E., Isham V.S. Onof C. and Wheater H.S. (1999) *Spatial-Temporal Stochastic Rainfall Modelling for hydrological design*, IAHS Publication No. 255, IAHS: pp. 225–235.

Obled C. and Creutin J.D. (1986) Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *Journal of Applied Meteorology*, **25**(9), 1189–1204.

Obukhov A.M. (1954) Statisticheskoe opisanie nepereryvnikh polej (Statistical description of continuous fields). *Akademii Nauk SSSR, Trudy Geofizicheskogo Instituta*, **24**(151), 3–42.

Obukhov A.M. (1960) O statisticheski ortogonalnykh razlo zheniyakh empiricheskikh funktsii (On statistical orthogonal expansions of empirical functions). *Izvestija Akademii Nauk SSSR, Ser. Geofizicheskaja*, **3**, 432–439.

Pougachev V.S. (1953) Obschaya teoriya korrelatsii sluchainykh funktsii (A general theory of correlation of random functions). *Izvestija Akademii Nauk SSSR, Serija Matematicheskaja*, **17**, 401–420.

Richman M.R. (1986) Rotation of principal components. *Journal of Climatology*, **6**, 293–335.

Sauquet E., Krasovskaia I. and Leblois E. (2000) Mapping mean monthly runoff patterns using EOF analysis. *Hydrology and Earth System Science*, **4**(1), 79–73.

Sauquet E. and Leblois E. (2001) Mapping runoff within the GEWEX-Rhone project. *La Houille Blanche*, **2001**(6–7), 120–129.

Searcy J.K. (1959) *Flow Duration Curves*, U.S. Geological Survey Water Supply Paper 1542-A, p. 33.

Singh R.D., Mishra S.K. and Chowdhary H. (2001) Regional flow-duration models for large number of ungauged Himalayan catchment for planning microhydro projects. *Journal of Hydrologic Engineering*, **6**(4), 310–316.

Smith R.L. (2001) *Environmental Statistics*, Department of Statistics, University of North Carolina: Web Reference: **`/www.stat.unc.edu/postscript/rs/envnotes.ps`**, July 2001.

Sokolovskij D.L. (1930) *Application of Distribution Curves to Determine Probability of Fluctuations in Annual Runoff for*

Rivers in the European Part of USSR (in Russian), Gostechizdat: Leningrad.

Tukey J.W. (1977) *Exploratory Data Analysis*, Addison Wesley: Reading.

Vanmarcke E. (1988) *Random Fields: Analysis and Synthesis*, MIT Press: Cambridge Mass, Third Printing.

Vogel R.M. and Fennessey N.M. (1994) Flow duration curves I: new interpretation and confidence intervals. *Journal of Water Resources Planning and Management*, **120**(4), 485–504.

Vogel R.M. and Fennessey N.M. (1995) Flow duration curves II: application in water resources planning. *Water Resources Bulletin*, **31**(6), 1029–1039.

Wiener H. (1930) Generalized harmonic analysis. *Acta Mathematics*, **55**, 117–258.

Wiener H. (1949) *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, MIT Press: Cambridge.

Wilson E.B. and Hilferty M.M. (1931) The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, **17**, 684.

Wornell G.W. (1990) A Karhunen-Loève-like expansion for 1/f processes via wavelets. *IEEE Transactions on Information Theory*, **36**(4), 859–861.

Yevjevich V. (1972) *Stochastic Processes in Hydrology*, Water Resources Publications: Forth Collins.

Yu P.-S. and Yang T.-C. (1996) Synthetic regional flow duration curve for southern Taiwan. *Hydrological Processes* **10**, 373–391.

8: Fractals and Similarity Approaches in Hydrology

LUCA G LANZA[1] AND JOHN GALLANT[2]

[1]Department of Environmental Engineering, University of Genoa, Genoa, Italy
[2]Butler Laboratory, Commonwealth Scientific and Industrial Research Organisation, Land and Water, Canberra, Australia

The application of fractals and similarity concepts in hydrology has given rise to a better understanding of the space-time organization of forms and processes that are relevant to the hydrologic cycle. Indeed, a variety of literature results support the conjecture that scaling holds for most hydrological variables in time and space.

This chapter concentrates on fields where fractals and similarity approaches have proved helpful in fostering advances in specific hydrological studies. In particular, precipitation and drainage network morphology are addressed. A few specific applications to the study of natural forms and patterns relevant to the hydrological sciences are presented first, while the application of scaling concepts to the study of processes themselves is later discussed.

Because of its highly irregular behavior, the rainfall process is one ideal candidate to be approached by means of self-similarity and/or (multi)fractal description models. Such models involve increasing complexity and computational burden as soon as the interest moves from the one-dimensional time series to the three-dimensional case, where the full space-time pattern of rainfall is considered. Examples of recent interesting results are presented with reference to the one-, two-, and three-dimensional approaches to rainfall modeling based on similarity concepts.

INTRODUCTION

As a geophysical science, hydrology investigates both natural forms and processes. Forms are the underlying physical matrix where hydrological processes actually take place. Forms may be shaped by the host processes themselves, for example, according to some energy expenditure criterion. Therefore, both forms and processes evolve in time and space, although the (space and/or time) scales of significant variations of forms are usually orders of magnitude larger than those of the dynamics of most hydrological processes (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**).

The wide range of self-similarity approaches, presently exploited in hydrology, found on the pioneering intuition by Mandelbrot (1983) that the description of natural forms and processes can hardly be bounded within the limits of the regular Euclidean geometry, since, for example, "clouds are not spheres…", with analogue statements applying to most of the objectives of scientific investigation in the geophysical sciences. He developed the idea of fractals, or fractal sets, in order to achieve a more suitable mathematical description – and, therefore, a deeper interpretation and understanding – of the observed forms and processes in nature.

Similarity approaches in general allow the transfer of information between scales (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**) or between places, and are frequently based on similarity indices – ratios or dimensionless numbers – that characterize some aspect of the hydrological processes. Two catchments with the same set of similarity indices should have a similar hydrological response. Fractals describe the situation where the similarity indices change predictably with scale.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Indeed, the application of fractals and similarity concepts in hydrology has given rise to a better understanding of the space-time organization of many processes within the hydrologic cycle, including precipitation (*see* **Chapter 28, Clouds and Precipitation, Volume 1**), evaporation (*see* **Chapter 45, Actual Evaporation, Volume 1**), soil moisture (*see* **Chapter 72, Measuring Soil Water Content, Volume 2**), runoff production (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**), and groundwater flow (*see* **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4**). This chapter concentrates on a few examples of cases where fractals and similarity approaches have proved helpful to foster advances in specific hydrological studies, although it is evident that analogous "scaling" approaches actually span over the whole panorama of hydrological investigations with promising results and, sometimes, very exciting perspectives. Such fields of investigation also include, for example, stream chemistry, where chloride concentration in stream flows (*see* **Chapter 91, Water Quality, Volume 3**) is shown to exhibit fractal scaling (Kirchner *et al.*, 2000), and snow hydrology (*see* **Chapter 159, Snow Cover, Volume 4**), the related scaling issues being discussed by Blöschl (1999).

The basic theoretical concepts of fractals and self-similarity are initially recalled in the following, and additional resources for deeper insights into such fundamentals are provided. Specific applications to the study of some natural forms and patterns relevant to the hydrological sciences are then presented, focusing mainly on the supporting physical matrix of most hydrological processes, that is, the drainage network in case of surface processes. The application of scaling concepts to the study of processes themselves is later discussed with reference to the rainfall process in both time and space (*see* **Chapter 28, Clouds and Precipitation, Volume 1**), this being the driving motor for the land phase of the hydrological cycle (*see* **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1**). Finally, a few comments are provided in the conclusions about the role of fractals and similarity approaches based on the results obtained in the various fields of investigation and their effective contribution in advancing our knowledge in the hydrological sciences.

FRACTALS AND SELF-SIMILARITY CONCEPTS

The concept of similarity is used to describe any object (and its generating process) that is invariant under ordinary geometric similarity (change of scale), and the object is therefore termed *self-similar*. This implies that the whole object can be split into N parts, obtained from it by a similarity of ratio r (after some geometric transformation) so that the latter is expressed as:

$$r(N) = \frac{1}{N^{1/D}} \tag{1}$$

and the similarity dimension is:

$$D = -\frac{\log N}{\log r(N)} \tag{2}$$

This definition also applies for common objects such as a line, a square, or a cube. Self-similar objects are also called *scaling objects* in the literature.

In nature, objects are seldom self-similar in the strict sense; rather, they may show statistical self-similarity, in the sense that random variations may affect the similarity of the smaller parts with respect to the original shape. Also, these smaller parts may be skewed by uneven reduction of scales in the different directions, and are called *self-affine objects* in this case.

A stochastic process P is said to be self-similar according to the above definitions when, once averaged at two different scales, it displays the following property:

$$\{P_{\lambda T,\lambda X}(t,x)\} \overset{d}{=} \lambda^H \{P_{T,X}(t,x)\} \tag{3}$$

where the symbol $\overset{d}{=}$ indicates equality in the probability distribution, λ is a scale factor, and H is the scaling exponent, which under suitable conditions can be related to D with simple relationships (Feder, 1988). Equation (3) indicates that the probability distribution of the averaged process is independent of the scale used for integration, and, therefore, the averaged quantities $P_{T,X}(t,x)$ and $P_{\lambda T,\lambda X}(t,x)$ display the same distribution when rescaled by a factor of λ^H. As a consequence of this property, we can also state that the raw moments of any order are scale invariant, so that $E[P^q_{\lambda T,\lambda X}] = \lambda^{qH} E[P^q_{T,X}]$, with q denoting the moment order. These properties are also referred to as simple scaling in the literature.

It is important to note that scaling properties usually apply in nature over a certain range of scales commonly bounded by a lower limit (the inner scale) and an upper limit (the outer scale).

The scaling exponent H is also termed the *Hurst exponent*, since the fractal behavior is an indication of long-term persistence or large spatial correlations, which is indicated as the Hurst phenomenon in the literature (see Hurst, 1951; Mesa and Poveda, 1993 for a discussion on the estimation of the Hurst exponent).

Hurst observed that the annual flows of rivers in successive years were not statistically independent but exhibited long-term dependence, with years of high and low flow tending to cluster together rather than be randomly

interspersed. This has important implications for reservoir design, and the cause of the phenomenon continues to be the subject of much debate – is it due to long-term persistence or does it result from an overlaying of multiple sources of variation having different timescales?

Similarity approaches were initially developed as empirical relationships based on observations of, for example, drainage network properties (Horton, 1945) and annual flow in rivers (Hurst, 1951). These relationships were given a mathematical foundation by Mandelbrot (1967, 1983) who developed the idea of fractals – processes or objects with fractional dimension whose distinguishing characteristics are variability and structure at all scales and the absence of a characteristic scale.

The similarity dimension equals the fractal or Hausdorff dimension (see Mandelbrot, 1983 for definitions) only for strictly self-similar fractals. However, fractals needs not be self-similar in general, and the Hausdorff dimension is a more general definition of the fractal dimension.

In a Euclidean framework, the topological dimension D_T – in simplified terms the "intuitive" dimension that represents the number of independent coordinates required in order to identify a generic location on the object in hand – is always an integer, while the so-called "critical" Hausdorff dimension D need not be an integer. They just satisfy the inequality $D \geq D_T$.

Following Mandelbrot (1983), a fractal is, by definition, a set for which the Hausdorff dimension D strictly exceeds the topological dimension D_T, so that in such cases, $D > D_T$. Every set with a noninteger D is a fractal (e.g. the Cantor set), although fractals may have an integer D (e.g. the Brownian motion), still not greater than the Euclidean dimension but strictly greater than D_T.

The most common methods to determine the Hausdorff dimension of a physical object are the Richardson (1961) and the box counting methods. The first one is also known as the *compass method*, as it simply requires to represent in a log–log plot, for several different compass sizes, the number of compass lengths needed to cover the object in hand. In case of scale-invariant behavior, this exercise will result in a straight line with negative slope that can be represented as a power law, and the slope itself is the sought fractal dimension. This method was first applied by Richardson to handle the popular issue of measuring the length of a coastline.

In the box counting method, a grid is placed over the investigated object and the number of boxes covering at least part of the fractal object are counted. The procedure is repeated after varying the dimension of the grid, and a log–log plot of the box counts over the grid size is used to derive a power law as in the previous method. Again, the slope of the derived straight line is assumed to be the fractal dimension of the object investigated.

Only the power law probability distribution is scale invariant and is, therefore, occasionally termed the *fractal distribution*. A characteristic scale in terms of space and/or time (see e.g. Skøien *et al.*, 2003 for a discussion based on the analysis of variograms) is, on the contrary, involved in other commonly used distributions such as the Gaussian or the Exponential. Fractal or self-similar processes are sometimes described by non-Gaussian distributions, and the observed power law behavior of a given variable is often the signature of some underlying scaling characteristic of the associated physical process. In general, however, a fractal has a power law probability distribution in the limit, that is, this behavior is observed as an asymptotic limit for the tail of the distribution (see e.g. Harris *et al.*, 1996).

In the particular case when the spatial scales of length, L_i and L_j, and the temporal scales, T_i and T_j, of a physical system are governed by a power law in the form

$$\left(\frac{T_i}{T_j}\right) = \left(\frac{L_i}{L_j}\right)^z \tag{4}$$

then the system is characterized by scale invariance in the dynamic sense or dynamic scaling, and z is the dynamic scaling exponent.

Another basic aspect of interest for hydrological applications is the concept of multifractality, or multiscaling behavior. The multifractal theory (see e.g. Feder, 1988; Schertzer *et al.*, 2002) was initially developed for describing the fluctuations of the velocity field in turbulent flows and later extended to geophysical processes such as rainfall and river networks.

Multiscaling can be seen as a departure from simple scaling, for example, in case the relation between raw moments at different scales is given as $E[P^q_{\lambda T,\lambda X}] = \lambda^{\varphi(q)} E[P^q_{T,X}]$, with the scaling exponent $\varphi(q) = qH \cdot \alpha(q)$ now being a function of the moment order q. Simple scaling is here the special case when $\alpha(q) = 1$, while for dissipative systems, $\alpha(q)$ is a convex function of q.

In order to further argument on this concept, let us introduce the so-called structure function of order q, that for a given set of observed data y_i of size N is defined as:

$$S_q(\tau) = \langle |y_{i+\tau} - y_i|^q \rangle, \qquad q > 0 \tag{5}$$

with $\langle . \rangle$ denoting the ensemble average. As it is evident from the case when $q = 2$, the structure function is a generalized correlation function, although q in the general definition can be any real (positive or negative) number.

In case the structure function should obey a scaling law, that is, the process is scale invariant, it would be written in the general form:

$$S_q(\tau) \propto \tau^{\zeta(q)} \tag{6}$$

where the exponent $\zeta(q)$ indicates a multifractal behavior in case it is a nonlinear function of the moment order q. This

behavior is termed *anomalous scaling* or *multifractality*. In short, it can be said that multifractality is evidenced by a curve when plotting $\zeta(q)$ over q, while a straight line in the same graph would indicate a mono-fractal behavior. Note that the Hurst exponent is here a function of q, namely $\varphi(q) = \zeta(q)/q$, and calculation of $\varphi(q)$ allows the identification of persistence as well as the mono-fractal or multifractal nature of the process.

Among the many other methods available to investigate the scaling properties of a physical process, the power spectrum is worth mentioning here due to its wide use, for example, in the study of rainfall time series. Again, if the frequency spectrum $E(f)$ of a given signal can be expressed by a power law in the form:

$$E(f) \propto f^{-\beta} \tag{7}$$

with f the frequency and β a suitable exponent, this indicates the absence of a characteristic scale in the range where the power law holds, and is, therefore, an indication of a possible scale-invariant behavior.

NATURAL FORMS AND PATTERNS

Mandelbrot's inspiration for the development of fractal geometry was the complexity of natural forms that are poorly described by the building blocks of Euclidean geometry: straight lines, smooth curves, spheres, and so on. Mandelbrot's demonstration that complex natural forms and patterns could be described and reproduced using simple mathematical techniques inspired a wide range of applications of fractal methods in the natural sciences, including hydrology and geomorphology. Two of the inspirations for Mandelbrot's theories were from these fields: Hurst's description of long-term persistence of river-flow rates over multiannual time periods (Hurst, 1951), and Richardson's reporting of the variation of the apparent length of coastlines when measured at different scales (Richardson, 1961).

The fractal model can be applied to natural terrain in two different ways: the fractal characteristics of the terrain surface itself, and the fractal characteristics of natural drainage networks.

Fractal Terrain

Elevation appears to obey power law scaling across a broad range of scales: the variance of elevation increases with increasing distance, and the relationship obeys a power law. Mandelbrot and others demonstrated that this relationship was consistent with a fractal model and much use was made of the fractal model as an explanation of the scaling properties of topography.

The scaling exponent, and hence the fractal dimension of terrain surfaces, is usually measured using spectral analysis

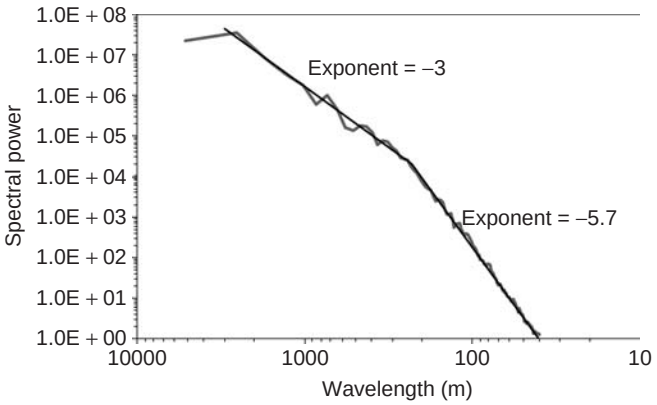


Figure 1 Power spectrum of surface topography based on 20 m resolution data showing the different scaling behaviour at coarse and fine scales (Livingstone Creek area near Wagga Wagga, NSW)

(Turcotte, 1989) or semivariogram (Mark and Aronson, 1984) techniques. One advantage of the spectral technique is its ability to identify surfaces smoother than the fractal model permits, indicated by a spectral exponent less than -3 (Gallant *et al.*, 1994). Many studies have found that the scaling exponent of the power law varies with scale, with a more rapid change of variance with distance found at distances less than about 200 m (Mark and Aronson, 1984; Gilbert, 1989; Tate, 1998). Over distances larger than about 200 m, the power law scaling conforms to a fractal model, but at the finer scales, the surface is smoother than the fractal model allows, as shown in Figure 1.

The length scale of about 200 m appears to correspond to the length of hillslopes in the landscape. The relative smoothness at scales finer than the hillslope length is probably due to diffusive processes that tend to obliterate small variations (Culling and Datko, 1987). At the broader scale where the drainage network dominates the form of the landscape, fluvial processes tend to exaggerate and perpetuate small variations.

It is worth noting that Mandelbrot's analysis of coastlines used length scales of 1 km and longer, so the relative smoothness of the landscape at finer scales was not noticed in his development of fractal models for topography.

Although land surfaces do not exactly fit the fractal model, the fractal dimension does appear to be a useful geomorphic parameter for distinguishing different landscape types, especially if used in conjunction with a measure of relief at a reference scale (Klinkenberg, 1992; Outcalt *et al.*, 1994).

The implications of a fractal model (at least over certain scales) for topography are many. The most important implication is that any finite resolution representation of the land surface fails to resolve the full variability of the real surface. This has the further implication that any estimate involving the derivatives of the surface, such as

slope or curvature, will vary with the resolution of the data. Measured slope, for example, has been demonstrated to consistently decrease with coarser resolution data (Moore *et al.*, 1993).

Drainage Networks

The self-similar properties of drainage networks were recognized from the 1940s (Horton, 1945; Strahler, 1952; Shreve, 1967) with the observation that the dendritic drainage network of a small catchment is similar to that of the larger catchment containing it. Horton established that the similarity across scales goes beyond appearances: the number of stream segments, their length, and the area drained all depend on stream order in a systematic way.

The ratios of these quantities between successive stream orders tend to remain constant over several levels of stream order and have been used to characterize the geomorphology of drainage networks. This self-similarity of drainage networks is a clear example of a natural fractal, and the fractal dimension of the drainage network can be readily derived from the Horton ratios.

The fractal dimension d of individual streams (a measure of the tendency towards meandering of a water course) and the fractal dimension D of the stream network (a measure of the tendency of river networks to fill the space when depicted as plan views) are usually of interest in hydrology, since they "can be used to investigate the scaling properties of the attributes and parameters describing drainage basin form and process" (La Barbera and Rosso, 1989). The following expressions were derived by La Barbera and Rosso (1989) and Rosso *et al.* (1991) on the basis of a quantitative analysis of river networks by means of the Horton's laws of network composition:

$$d = \max\left(1, 2\frac{\log R_L}{\log R_a}\right) \tag{8}$$

$$D = \min\left(2, \max\left(1, \frac{\log R_b}{\log R_L}\right)\right) \tag{9}$$

with R_a, R_b, R_L being respectively the Horton's area, bifurcation, and length ratios. A comprehensive although synthetic review of these and further developments in the fractal theory of stream networks can be found, for example, in the work by Schuller *et al.* (2001).

Self-similarity has also been found in regional drainage systems containing multiple independent but conterminous drainage basins along a coastline in both theoretical studies based on network generation models (Sun *et al.*, 1995) and measurements of natural drainage basins (La Barbera and Lanza, 2001).

In the latter work, the cumulative probability distribution of basin areas along a coastline, that is usually expressed in the form of a power law (see e.g. Perera and Willgoose (1998), La Barbera and Lanza (2000)) as:

$$P[A > a] \propto a^{-\gamma} \tag{10}$$

is related to the regional area ratio R_A and the multiplicative factor R_C defined as the regional analogues of the Horton's area and bifurcation ratios, in the form:

$$\gamma = \frac{\log R_C}{\log R_A} \tag{11}$$

Figure 2(a), 2(b) shows the observed variability across scales of the R_A and R_C ratios for independent drainage basins in the Liguria region of Italy. Figure 3 illustrates the cumulative distribution function of catchment areas for the same geographic region.

Variations in the Horton ratios from one catchment to another should affect the hydrological response, and this link was formalized by Rodriguez-Iturbe and Valdes (1979) then extended by several authors, culminating in the major work of Rodriguez-Iturbe and Rinaldo (1997).

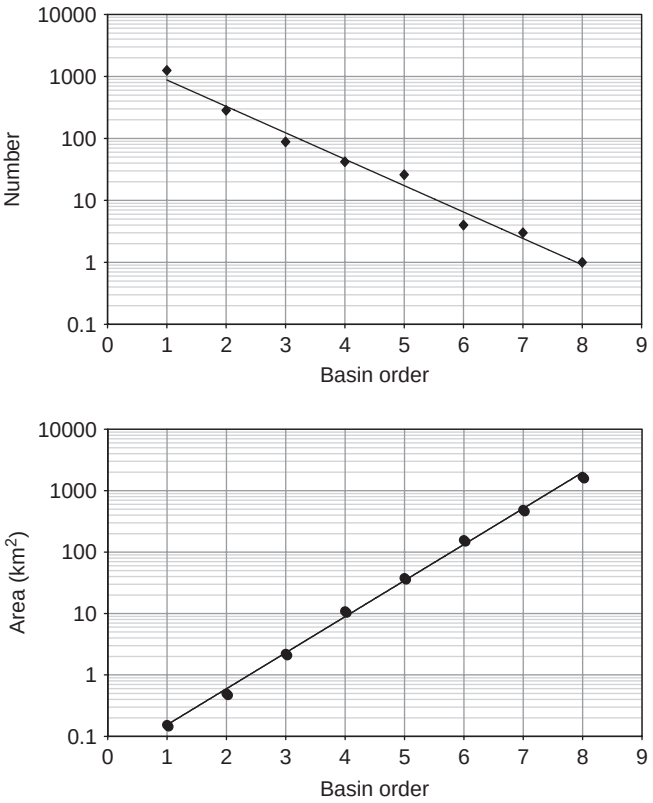


Figure 2 Scaling properties of a regional drainage network in terms of (a) the multiplicative factor R_C and (b) the area ratio R_A, defined in analogy with the Horton's bifurcation and area ratios, for independent drainage basins along the coastline of the Liguria region of Italy

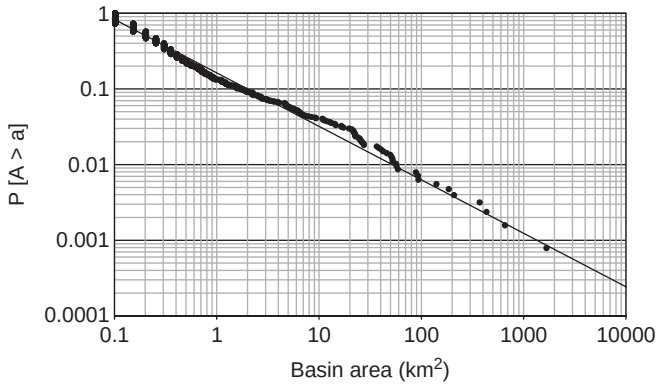


Figure 3 Power law fitting of the cumulative distribution function of catchment areas for independent drainage basins along the coastline of the Liguria region of Italy

The hillslope length or area of first-order catchments acts as the fine-scale limit of fractal scaling for both topographic surfaces and the drainage network. However, no other connection has been established between the fractal dimensions of surface topography and the drainage network except for theoretical models of drainage basin evolution (e.g. Sun *et al.*, 1994).

Hydraulic Geometry

Hydraulic geometry describes the relationships between channel form and flow regimes such as mean annual discharge or mean annual flood. The underlying principle of hydraulic geometry is that channels adjust to the flow within them reflecting the dynamic equilibrium between erosional power (*see* **Chapter 82, Erosion Prediction and Modeling, Volume 2**) and resistance, with bank-full discharge acting as the channel-forming flow. This condition is generally true only for alluvial channels that are in equilibrium with hydrologic conditions (Mosley and McKerchar, 1992).

Channels tend to become wider and deeper and have lower slope as flow increases further down the catchment, and the variations of width, depth, and slope with flow (or catchment area) tend to follow a consistent pattern (Leopold *et al.*, 1964, chapter 7; Sweet and Geratz, 2003). The relationships between channel width, depth, slope, and catchment area provide the basis for extrapolating measured streamflow at one location to different parts of a catchment, for inferring streamflow in a nearby catchment with similar rainfall, geology, and geomorphology or for estimating channel form from terrain analysis (typically catchment area and slope) for the purpose of modeling hydrologic response. In many situations, the imprecision of these relationships prevents accurate estimation of unknown quantities.

Catchment Similarity Indices

The use of similarity is not always connected with similarity across scales. Many different ratios and indices can be calculated for a catchment that can be used to assess similarity in hydrologic response, ranging from simple topographic ratios such as circularity and slope to complex hydroclimatic indices. Catchments with similar index values are expected to have a similar hydrologic response, although the degree to which this works in practice depends on the degree to which the catchment matches the assumptions of the index. Hydrologic similarity indices can also be computed at different locations within catchments to determine hydrologic similarity of particular sites, such as the $\ln(a)/\tan(\beta)$ index (topographic wetness index) originated by Beven and Kirkby (1979). This index is the basis of the TOPMODEL approach to hydrological modeling (*see* **Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**) (Quinn *et al.*, 1995).

Milly (1993, 1994) and Woods (2003) developed several catchment similarity indices relating directly to hydrologic response. Woods' starting point is the climate dryness index R defined as the ratio of mean evaporation rate to mean rainfall rate. Seasonality of the balance between rainfall and evaporation is measured by the index S, with small values indicating little seasonality. A uniformly wet climate is characterized by $R < 1$ and $S < 1 - R$ while a seasonally wet climate has $S > \max(1 - R, R - 1)$. Other indices represent the effects of canopy interception, root zone storage, subsurface flow, and surface saturation, and the analysis leads to a combined storage-seasonality index S_r^* that is claimed to provide a concise definition of similarity among catchments.

These indices are yet to be widely adopted but have the potential to provide a common set of measures used to describe catchments and a means for identifying the dominant sources of variations in hydrologic processes between catchments.

Porous Media

The physical properties of porous media controlling the subsurface flow of fluids (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**) and the transport of contaminants (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**) are seldom characterized by smoothly varying functions of position, and they rather exhibit spatial fluctuations over a wide range of scales (Hewett and Behrens, 1989) (*see* **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**). The spatial variability of the hydraulic conductivity K has been widely investigated in order to understand flow and transport processes in heterogeneous systems.

In this context, fractal models are generally characterized by variograms or correlation measures of the relevant

variables that increase as a power law with the separation distance. A unifying framework for scaling models of such heterogeneities in porous media has been recently proposed by Painter (2001) and suitable references can be found therein.

The development of scaling models largely derives from the empirical evidence provided by the analysis of measured hydraulic conductivity fields. However, Liu and Molz (1997) observed that a mono-fractal model for the log-conductivity $logK$ cannot represent all types of heterogeneities and suggested to treat $logK$ as a multifractal field. Boufadel *et al.* (2000) argued that K itself should be treated as multifractal.

Tennekoon *et al.* (2003) analyzed the scaling properties of the hydraulic conductivity of three sites in North America and, by investigating the structure function of the data while varying the moment order q, found that K exhibits multifractality in both the vertical and the horizontal directions.

Mukhopadhyay and Cushman (1998) studied the effect of heterogeneities on the spreading of pollutants from a nonaqueous phase liquid trapped in a soil by modeling heterogeneities as a self-similar fractal process. Zhu and Sykes (2000) developed expressions for flow variances and macrodispersivities in a fractal semiconfined aquifer in terms of the leakage factor. A fractal model for total solute transport, assuming power law retention times, is proposed by Schumer *et al.* (2003) who provide a wide set of references for further reading on this subject.

HYDROLOGICAL PROCESSES

Rainfall

The rainfall process, as many other geophysical processes deriving from the superposition and interaction of different scales of motion, presents high space-time variability (*see* **Chapter 28, Clouds and Precipitation, Volume 1**) and non-Gaussian probability distributions for most of its characteristic variables (rain intensity, event duration, interarrival periods, etc.). The actual variability and complexity of the phenomenon are, however, often masked by the strong limitations and sampling characteristics of the observation methods (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**, **Chapter 63, Estimation of Precipitation Using Ground-based, Active Microwave Sensors, Volume 2** and **Chapter 64, Satellite-based Estimation of Precipitation Using Microwave Sensors, Volume 2**). Additionally, the shortage of data strongly impinges upon the conclusions derived. The uncertainties associated with our many requirements in terms of simulation and forecasting are usually dealt with by modeling rainfall as a random process (*see* **Chapter 31, Models of Clouds, Precipitation and Storms, Volume 1**).

Due to the above-mentioned characteristics, the rainfall process is one ideal candidate to be approached by means of self-similarity and/or fractal description models. Such models involve increasing complexity and computational burden as soon as the interest moves from the one-dimensional case, that is, the analysis of rainfall time series at a single point in space, to the two-dimensional case, that is, the analysis of the rain field in space at a frozen instant in time, and, finally, to the three-dimensional case where the full space-time pattern of rainfall is considered. Examples are given below, together with a short mention of an interesting application of the scaling approach to the description of rainfall extremes in terms of their intensity, duration, and frequency characteristics for application in the hydrological practice.

Time Series

In this section, we concentrate on the statistical behavior of rainfall in time without explicitly considering the spatial distribution and organization of the precipitation field.

The investigation of the highly irregular behavior of the local rainfall process led many authors to obtain some positive evidence of the fractal nature of rainfall in time – see, for example, Lovejoy and Mandelbrot (1985), Hubert *et al.* (1993), Olsson (1995), Svensson *et al.* (1996), and Tessier *et al.* (1996) – and provided a widely accepted indication that a multifractal approach seems to be appropriate for the description of rainfall time series. Evidence of this is reported in Figure 4 for a 10-min time series recorded in Genova (Italy) in terms of the normalized scaling exponent of the structure function $\zeta(q)/\zeta(2)$ as a function of the moment order q, while in Figure 5, the normalized spectral density function for the

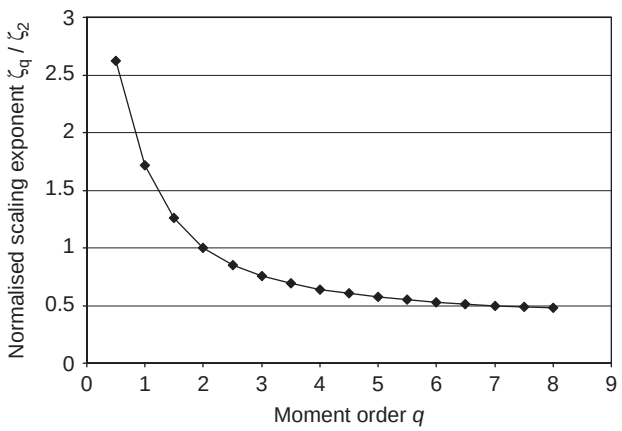


Figure 4 Example of the multiscaling behavior of 10-min rainfall data over a 10-year period at the rain gauge station of Genova (after Molini, 2001)

same data set is presented (see e.g. Molini, 2001 and Molini *et al.*, 2005a).

The conceptual relevance of such indication is the possibility of evaluating useful properties or synthetic parameters over a wide range of aggregation scales, in this way answering the basic requirement of many hydrological applications where the rainfall input is known (measured) at a different scale than those of interest for the problem in hand – see, for example, the downscaling (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**) operational problem addressed in Molini *et al.* (2005b). On the modeling side, the observed multiscaling behavior yields the possibility of generating suitable local rainfall scenarios by employing random cascade algorithms, based on the analogy with the energy cascade that is used to describe the dynamic behavior of turbulence (see e.g. Menabde and Sivapalan, 2000; Olsson, 1995, 1998; Svensson *et al.*, 1996; Güntner *et al.*, 2001).

An interesting approach to rainfall modeling in time has been recently proposed by Veneziano and Iacobellis (2002), who presented a pulse-based representation of temporal rainfall with multifractal properties at small scales and deviations from scaling in the transient regime from small to large scales. The model is able to reproduce the observed on/off properties as well as the behavior of rain-intensity fluctuations with a parsimonious parameterization (i.e. employing few parameters that can be calibrated more easily and make the model more robust than non-fractal models) in the range from 20 min to a few days.

Design Rainfall

The similarity approach has direct practical implications in the statistical analysis of extreme rainfall events, aimed at the assessment of design rainfall as a function of duration, area, and the return period (*see* **Chapter 37, Rainfall Trend Analysis: Return Period, Volume 1**) – namely, the Depth-Duration-Frequency (DDF) curves. This also allows the derivation of a theoretical formulation of the ARF (Area Reduction Factor), a widely used parameter in hydrology and water resources-related disciplines where area-average rainfall statistics are required for application and design purposes.

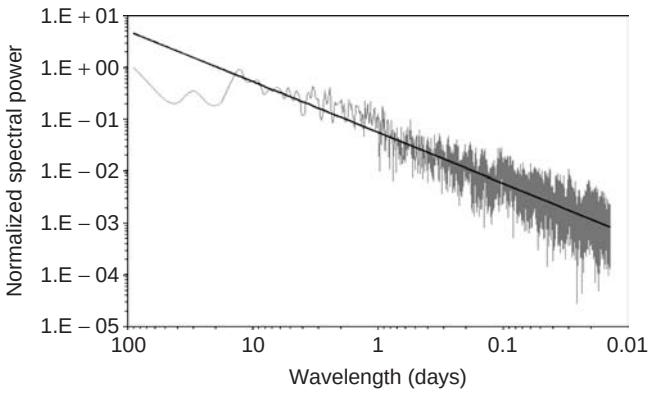


Figure 5 Normalized spectral density function for 10-min rainfall data over a 10-year period at the rain gauge station of Genova (after Molini, 2001)

A comprehensive approach to the derivation of a theoretical formulation for the DDF curves based on a scaling framework has been provided by Burlando and Rosso (1996) and Menabde *et al.* (1999). In the first work, the scale invariance properties of extreme storm probabilities are shown to underlie the derivation of a distribution-free formulation for the DDF curves based on the simple scaling and multiscaling properties of temporal rainfall, for example, in the first case in the form:

$$h(d, T) = a(1 + cv \cdot K_T) \cdot d^n \qquad (12)$$

where h is rainfall depth, d is the duration, T the return period, cv the coefficient of variation, K_T the frequency factor (Chow, 1951), and a, n two suitable parameters.

Veneziano and Furcolo (2002) recently presented an in-depth discussion of the above mentioned simple scaling behavior of the DDF curves and evidences of multifractality in rain time series, and provided a link between the scaling exponents of the curves and the moment scaling function of the rainfall time series.

The advantage of the scaling approach over the traditional one lies in the possibility of estimating the DDF parameters from relatively short time series of coarse rainfall data and to be able to use them for prediction of extreme storm characteristics over a range of technical interest that usually spans from 1 to 24 h in duration.

Space-Time Rainfall

The modeling of rain fields is traditionally approached by investigating the spatial statistical properties of rainfall accumulation on a fixed time window, without explicitly taking into account the evolution in time of the observed spatial patterns.

The conjecture that the rain field is scale invariant over a broad range of spatial and temporal scales is inspired by the statistical theory of turbulence, with reference to the atmospheric processes that originate rainfall, and supported by increasing empirical evidence.

Time-frozen or space-frozen rain fields can be suitably interpreted within a scaling framework, and a few examples are described, for example, by Lovejoy and Mandelbrot (1985) and Lovejoy and Schertzer (1985). However, it is evident that such a frozen field interpretation is not able to fully characterize the variety of observed rainfall structures (Deidda *et al.*, 2004) and suitable models are required to reproduce statistical properties jointly observed at different space and time scales.

An extensive analysis of space-time scaling properties of the rain field was recently performed by Deidda *et al.*

(2004) based on data from the well-known TOGA-COARE experiment, with results showing evidence of multifractality under self-similar transformations in space and time.

Multifractal models are available in the literature in order to provide a tool for parsimonious generation of realistic space-time rainfall scenarios (see e.g. Gupta and Waymire, 1990; Marsan *et al.*, 1996; Over and Gupta, 1996; Venugopal *et al.*, 1999a,b).

The space-time pattern of rainfall is indeed relevant to hydrological applications such as flood forecasting (*see* **Chapter 123, Rainfall-runoff Models for Real-time Forecasting, Volume 3**), especially when the meteorological predictions provided by physically based prognostic models of the atmosphere are obtained at much coarser scales than those required by catchment-scale hydrologic rainfall-runoff models used for flood warning purposes, so that suitable downscaling must be performed. The knowledge of scale-invariant properties in high-intensity storms is suitable to accomplish the need of rainfall downscaling methods in order to disaggregate large-scale rainfall forecasts to the smaller response scales of hydrological catchment models (see e.g. Deidda, 2000; Venugopal *et al.*, 1999a; Turner *et al.*, 2004).

CONCLUSIONS

Hydrological processes operate over a wide range of scales in time and space. Describing these processes and how they vary across scales has been a significant challenge in hydrology. Similarity approaches look for robust relationships between scales and hydrological parameters to allow the estimation of properties at a range of scales based on observations or measurements at one or few scales, and a variety of literature results support the conjecture that scaling holds for most hydrological variables in time and space.

The wide range of scales present in hydrology and other natural systems can be thought about in two distinct ways: as a series of distinct hierarchical levels or as a continuum of scales.

The hierarchy approach conceives of the range of scales as nested levels and focuses on the separation of different scales of variation and different processes at different scales. Such an approach commonly lists a number of hierarchical levels, each with its distinct characteristics. In hydrology, these might range from the scale of the experimental column through plots, hillslopes, small catchments, large catchments to the global hydrological cycle.

The continuum view conceives of the range of scales as a continuum with a focus on a single characterization across all (or a wide range of) scales. The fractal model is one such model that seeks to explain (or at least describe) the range of variations at different scales using a single model with one or two parameters. The continuum view tends to assume that the characterization is valid across all scales; this assumption has the corollary that there is no characteristic scale.

The weakness of the hierarchical view is that real systems may not exhibit a consistent hierarchy (some levels may be missing, or duplicated) and the division between levels of the hierarchy may be difficult to discern. The corresponding weakness of the continuum scaling view is that, in practice, scaling laws only operate over restricted ranges of scales. These two views can be reconciled by considering levels of the hierarchy as scale regimes within which scaling laws apply (Lewis, 1995). This composite view usually leads to a hierarchy with fewer levels, and to scaling laws with well-defined limits of applicability.

REFERENCES

Beven K.J. and Kirkby M.J. (1979) A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Blöschl G. (1999) Scaling issues in snow hydrology. *Hydrological Processes*, **13**, 2149–2175.

Boufadel M.C., Lu S., Molz F.J. and Lavallee D. (2000) Multifractal scaling of the intrinsic permeability. *Water Resources Research*, **36**(11), 3211–3222.

Burlando P. and Rosso R. (1996) Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. *Journal of Hydrology*, **187**, 45–64.

Chow V.T. (1951) A general formula for hydrologic frequency analysis. *Transactions American Geophysical Union*, **32**, 231–237.

Culling W.E.H. and Datko M. (1987) The fractal geometry of the soil-covered landscape. *Earth Surface Processes and Landforms*, **12**, 369–385.

Deidda R. (2000) Rainfall downscaling in a space-time multifractal framework. *Water Resources Research*, **36**(7), 1779–1794.

Deidda R., Badas M.G. and Piga E. (2004) Space-time scaling in high-intensity Tropical Ocean Global Atmosphere Coupled Ocean-Atmosphere Response Experiment (TOGA-COARE) storms. *Water Resources Research*, **40**, W02506.

Feder J. (1988) *Fractals*, Plenum Press: New York.

Gallant J.C., Moore I.D., Hutchinson M.F. and Gessler P.E. (1994) Estimating the fractal dimension of profiles: a comparison of methods. *Mathematical Geology*, **26**(4), 455–481.

Gilbert L.E. (1989) Are topographic data sets fractal? *Pure and Applied Geophysics*, **131**(1/2), 241–254.

Güntner A., Olsson J., Calver A. and Gannon B. (2001) Cascade-based disaggregation of continuous rainfall time series: the influence of climate. *Hydrology and Earth System Sciences*, **5**(2), 145–164.

Gupta V.K. and Waymire E. (1990) Multiscaling properties of spatial rainfall and river flow distributions. *Journal of Geophysical Research*, **95**, 1999–2009.

Harris D., Menabde M., Seed A.W. and Austin G.L. (1996) Multifractal characterization of rain fields with a strong orographic influence. *Journal of Geophysical Research*, **101**(D21), 26405–26414.

Hewett T.A. and Behrens R.A. (1989) Scaling laws in reservoir simulation and their use in a hybrid finite difference/streamtube approach to simulating the effects of permeability heterogeneity. *Reservoir Characterization*, Academic Press: San Diego, pp. 402–441.

Horton R.E. (1945) Erosional development of streams and their drainage basins: hydrological approach to quantitative geomorphology. *Geological Society of America Bulletin*, **56**, 275–370.

Hubert P., Tessier Y., Lovejoy S., Schertzer D., Schmitt F., Ladoy P., Carbonnel J.P., Violette S. and Desurosne I. (1993) Multifractals and extreme rainfall events. *Geophysical Research Letters*, **20**, 931–934.

Hurst H.E. (1951) Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineering.*, **116**, 770–808.

Kirchner J.W., Feng X. and Neal C. (2000) Fractal stream chemistry and its implications for contaminant transport in catchments. *Nature*, **403**, 524–527.

Klinkenberg B. (1992) Fractals and morphometric measures: is there a relationship? *Geomorphology*, **5**, 5–20.

La Barbera P. and Lanza L.G. (2000) Comment on: "A physical explanation of the cumulative area distribution curve" by H. Perera and G. Willgoose. *Water Resources Research*, **36**(3), 815–817.

La Barbera P. and Lanza L.G. (2001) On the cumulative area distribution of natural drainage basins along a coastal boundary. *Water Resources Research*, **37**(5), 1503–1509.

La Barbera P. and Rosso R. (1989) On the fractal dimension of stream networks. *Water Resources Research*, **25**, 735–741.

Leopold L.B., Wolman M.G. and Miller J.P. (1964) *Fluvial Processes in Geomorphology*, Freeman: San Francisco.

Lewis, A. (1995) *Scale in Spatial Environmental Databases*, Ph.D. thesis, Centre for Resource and Environmental Studies, Australian National University.

Liu H.H. and Molz F.J. (1997) Multifractal analyses of hydraulic conductivity distributions. *Water Resources Research*, **33**, 2483–2488.

Lovejoy S. and Mandelbrot B.B. (1985) Fractal properties of rain and a fractal model. *Tellus*, **37A**, 209–232.

Lovejoy S. and Schertzer D. (1985) Generalized scale invariance and fractal models of rain. *Water Resources Research*, **21**, 1233–1250.

Mandelbrot B.B. (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, **155**, 636–638.

Mandelbrot B.B. (1983) *The Fractal Geometry of Nature*, Freeman: San Francisco.

Mark D.M. and Aronson P.B. (1984) Scale-dependent fractal dimensions of topographic surfaces: an empirical investigation, with applications in geomorphology and computer mapping. *Mathematical Geology*, **16**(7), 671–683.

Marsan D., Schertzer D. and Lovejoy S. (1996) Causal space-time multifractal processes: predictability and forecasting of rain fields. *Journal of Geophysical Research*, **101**, 26333–26346.

Menabde M., Seed A. and Pegram G. (1999) A simple scaling model for extreme rainfall. *Water Resources Research*, **35**(1), 335–339.

Menabde M. and Sivapalan M. (2000) Modelling of rainfall time series and extremes using bounded random cascades and levy-stable distributions. *Water Resources Research*, **36**, 3293–3301.

Mesa O.J. and Poveda G. (1993) The Hurst effect: the scale of fluctuation approach. *Water Resources Research*, **29**(12), 3995–4002.

Milly P.C.D. (1993) An analytic solution of the stochastic storage problem applicable to soil-water. *Water Resources Research*, **29**(11), 3755–3758.

Milly P.C.D. (1994) Climate, soil water storage and the average annual water balance. *Water Resources Research*, **30**, 2143–2156.

Molini A. (2001) *Analysis of NonLinear Features of the Rain Process with Emphasis on the Intermittency Structure (In Italian)*, PhD Dissertation, University of Genova.

Molini A., La Barbera P. and Lanza L.G. (2005a) Correlation patterns and information flows in rainfall fields. *Journal of Hydrology*, (in press).

Molini A., Lanza L.G. and La Barbera P. (2005b) The impact of tipping-bucket raingauge measurement errors on design rainfall for urban-scale applications. *Hydrological Processes*, **19**, 1073–1088.

Moore I.D., Lewis A. and Gallant J.C. (1993) Terrain attributes: estimation methods and scale effects. In *Modelling Change in Environmental Systems*, Chap. 8, Jakeman A.J., Beck B. and McAleer M. (Eds.), John Wiley & Sons: London, pp. 189–214.

Mosley M.P. and McKerchar A.I. (1992) Streamflow. In *Handbook of Hydrology*, Chap. 8, Maidment D.R. (Ed.) McGraw-Hill: New York.

Mukhopadhyay S. and Cushman J.H. (1998) Diffusive transport of volatile pollutants in nonaqueous phase liquid contaminated soil: a fractal model. *Transport in Porous Media*, **30**(2), 125–154.

Olsson J. (1995) Limits and characteristics of the multifractal behaviour of a high-resolution time series. *Nonlinear Processes in Geophysics*, **2**, 23–29.

Olsson J. (1998) Evaluation of a cascade model for temporal rainfall disaggregation. *Hydrology and Earth System Science*, **2**, 19–30.

Outcalt S.I., Hinkel K.M. and Nelson F.E. (1994) Fractal Physiography? *Geomorphology*, **11**, 91–106.

Over T.M. and Gupta V.K. (1996) A space-time theory of mesoscale rainfall using random cascades. *Journal of Geophysical Research*, **101**, 26319–26331.

Painter S. (2001) Flexible scaling model for use in random field simulation of hydraulic conductivity. *Water Resources Research*, **37**(5), 1155–1163.

Perera H. and Willgoose G. (1998) A physical explanation of the cumulative area distribution curve. *Water Resources Research*, **34**, 1335–1343.

Quinn P., Beven K. and Lamb R. (1995) The $\ln(a)/\tan(\beta)$ index: How to calculate it and how to use it within the TOPMODEL framework. *Hydrological Processes*, **9**, 161–182.

Richardson L.F. (1961) The problem of contiguity: an appendix of statistics of deadly quarrels. *General Systems Yearbook*, **6**, 139–187.

Rodriguez-Iturbe I. and Rinaldo A. (1997) *Fractal River Basins: Chance and Self-Organisation*, Cambridge University Press: New York.

Rodriguez-Iturbe I. and Valdes J.B. (1979) The geomorphologic structure of hydrologic response. *Water Resources Research*, **15**, 1409–1420.

Rosso R., Bacchi B. and La Barbera P. (1991) Fractal relation of main-stream length to catchment area in river networks. *Water Resources Research*, **27**, 381–387.

Schertzer D., Lovejoy S. and Hubert P. (2002) An introduction to stochastic multifractal fields. Mathematical problems in environmental science and engineering. In *Series in Contemporary Applied Mathematics*, Ern A. and Weiping L. (Eds.), Higher Education Press: Beijing, pp. 106–179.

Schuller D.J., Rao A.R. and Jeong G.D. (2001) Fractal characteristics of dense stream networks. *Journal of Hydrology*, **243**, 1–16.

Schumer R., Benson D.A., Meerschaert M.M. and Baeumer B. (2003) Fractal mobile/immobile solute transport. *Water Resources Research*, **39**(10), 1296–1308.

Shreve R.L. (1967) Infinite topologically random channel networks. *Journal of Geology*, **75**, 178–186.

Skøien J.O., Blöschl G. and Western A.W. (2003) Characteristics space scales and timescales in hydrology. *Water Resources Research*, **39**(10), 1304–1323.

Strahler A.N. (1952) Hypsometric (area-altitude) analysis of erosional topography. *Geological Society of America Bulletin*, **63**, 1117–1142.

Sun T., Meakin P. and Jossang T. (1994) The topography of optimal drainage basins. *Water Resources Research*, **30**(9), 2599–2610.

Sun T., Meakin P. and Jossang T. (1995) Minimum energy dissipation river networks with fractal boundaries. *Physical Review E*, **51**(6), 5353–5359.

Svensson C., Olsson J. and Berndtsson R. (1996) Multifractal properties of daily rainfall in two different climates. *Water Resources Research*, **32**, 2463–2472.

Sweet W.S. and Geratz J.W. (2003) Bankfull hydraulic geometry relationships and recurrence intervals for North Carolina's coastal plain. *Journal of the American Water Resources Association*, **39**(4), 861–871.

Tate N.J. (1998) Maximum entropy spectral analysis for the estimation of fractals in topography. *Earth Surface Processes and Landforms*, **23**, 1197–1217.

Tennekoon L., Boufadel M. C., Lavallee D. and Weaver J. (2003) Multifractal anisotropic scaling of the hydraulic conductivity. *Water Resources Research*, **39**(7), 1193–1205.

Tessier Y., Lovejoy S., Hubert P., Schertzer D. and Pecknold S. (1996) Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. *Journal of Geophysical Research*, **101**, 26427–26440.

Turcotte D.L. (1989) Fractals in geology and geophysics. *Pure and Applied Geophysics*, **131**, 171–196.

Turner B.J., Zawadzki I. and Germann U. (2004) Predictability of Precipitation from Continental Radar Images. Part III: Operational Nowcasting Implementation (MAPLE). *Journal of Applied Meteorology*, 231–247.

Veneziano D. and Furcolo P. (2002) Multifractality of rainfall and scaling of intensity-duration-frequency curves. *Water Resources Research*, **38**(12), 1306.

Veneziano D. and Iacobellis V. (2002) Multiscaling pulse representation of temporal rainfall. *Water Resources Research*, **38**(8), 1138.

Venugopal V., Foufoula-Georgiou E. and Sapozhnikov V. (1999a) A space-time downscaling model for rainfall. *Journal of Geophysics Research*, **104**, 19,705–19,721.

Venugopal V., Foufoula-Georgiou E. and Sapozhnikov V. (1999b) Evidence of dynamic scaling in space-time rainfall. *Journal of Geophysics Research*, **104**, 31,599–31,610.

Woods R. (2003) The relative roles of climate, soil, vegetation and topography in determining seasonal and long-term catchment dynamics. *Advances in Water Resources*, **26**(3), 295–309.

Zhu J. and Sykes J.F. (2000) Head variance and macrodispersivity tensor in a semiconfined fractal porous medium. *Water Resources Research*, **36**(1), 203–212.

9: Statistical Upscaling and Downscaling in Hydrology

GÜNTER BLÖSCHL

Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Vienna, Austria

Upscaling and downscaling methods are needed to transfer information from small-scale data to large-scale predictions and vice versa. There are two types of methods; methods focusing on how the equations and parameters of dynamic models change with scale, which are treated in other articles of this Encyclopedia, and methods focusing on how to best represent variability statistically, which is the subject of this article. This article gives a brief overview of scale concepts and illustrates first-order effects of upscaling and downscaling. The most important statistical upscaling/downscaling methods are then reviewed for various hydrological processes – rainfall, floods, soil moisture, and subsurface flow and transport.

INTRODUCTION

Hydrological analyses and predictions build on two strands of information; prior knowledge, which is usually embodied in some kind of model, and local data. The local data collected in the area of interest and for the time period of interest are used to specify the exact form of the model outcome through use of the inputs and choice of the model parameters (by model calibration), and sometimes through the choice of the model structure as well (**Chapter 10, Concepts of Hydrologic Modeling, Volume 1**). The spatial and temporal scales at which the predictions are needed may be different from those of the data, and in some instances model output from one scale needs to be combined with models at a different scale. In both cases, some scale adjustment or scale transfer is needed, either upwards in scale (termed *upscaling*) or downwards in scale (termed *downscaling*).

There exists an immense body of literature on upscaling and downscaling methods in the various subdisciplines of hydrology. Summary reviews include Blöschl and Sivapalan (1995) and Bierkens *et al.* (2000). More specific reviews of land surface parameters and porous media are provided in Shuttleworth *et al.* (1997) and Farmer (2002), respectively, and conference proceedings or collections of papers include Rodríguez-Iturbe and Gupta (1983), Gupta *et al.* (1986), Kalma and Sivapalan (1995), Stewart *et al.* (1996), Blöschl *et al.* (1997), Sposito (1998), and Pachepsky *et al.* (2003).

There are two generic types of upscaling and downscaling methods. The first type of method involves dynamic models of parts of the hydrologic cycle where the upscaling and downscaling issue is how the model equations and model parameters change with scale. This is beyond the scope of this article and is dealt with in articles **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**, **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**, **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**, **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**, and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**.

The second type of method consists of statistical descriptions where the focus is on how to best represent random variability in both space and time at various scales. This is the focus of this article. The section "Scale, upscaling, and downscaling" of the present article gives a brief overview of scale concepts and introduces the simplest case of statistical upscaling/downscaling methods (first-order effects). The section "Types of statistical upscaling/downscaling methods" summarizes the most important types of methods. The

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

section "Statistical upscaling and downscaling methods" reviews statistical upscaling/downscaling methods that are currently of most interest in different subareas of the hydrological sciences, that is, rainfall (Sections "Upscaling point rainfall to catchments, temporal disaggregation of rainfall, statistical downscaling of the output of global circulation models"), floods (Section "Flood frequency as a function of catchment scale"), soil moisture (Section "Upscaling and downscaling soil moisture"), and subsurface hydrology (Section "Subsurface media characterization and generation"). Section "Concluding remarks" summarizes the main aspects of the methods.

SCALE, UPSCALING, AND DOWNSCALING

Processes, Sampling, and Modeling

Natural variability can be characterized statistically by spatial or temporal probabilities or their moments (such as the mean and the variance). In the simplest case of focusing on the second moments (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**), the scale of the underlying variability can be represented by the integral scale which is the average distance over which a variable is correlated (Skøien *et al.*, 2003). A sampling exercise will rarely reveal the underlying natural variability in full detail because of instrument error and because the spatial and temporal dimensions of the instruments or measurement setup will always be finite (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1** and **Chapter 6, Principles of Hydrological Measurements, Volume 1**). Blöschl and Sivapalan (1995) suggested to term the dimensions of the measurements *the sampling scale triplet* consisting of the spacing, extent, and support of the data. In dedicated studies, soil hydraulic conductivity measurements, for example, may have spacings of decimeters, while rain gauges in a region are typically spaced at tens of kilometers. The extent, which is the overall size of the domain sampled, may again range from meters to hundreds of kilometers in hydrological applications. The support is the integration volume or area of the samples ranging from, say, $1\,\mathrm{dm}^2$ in the case of time domain reflectometry *in situ* probes (*see* **Chapter 72, Measuring Soil Water Content, Volume 2**), to hectares in the case of groundwater pumping tests (Anderson, 1997) or micrometeorological studies of the atmosphere (Schmid, 2002), and square kilometers in the case of remotely sensed data (Western *et al.*, 2002). In hydrology, the catchment area can also be thought of as a support scale (Bierkens *et al.*, 2000).

In case of a model, the notion of a scale triplet is similar. For example, for a spatially distributed hydrologic model the scale triplet may have typical values of, say, 25 m spacing (i.e. the grid spacing), 1 km extent (i.e. the size of the catchment or aquifer to be modeled), and 25 m support (the cell size). Analogue scales apply to the temporal domain.

The important point in the context of upscaling and downscaling is that the sampling scale triplet will have some bearing on the data and the modeling scale triplet will have some bearing on the predictions. Generally, if the spacing of the data is large, the small-scale components of the natural variability will not be captured by the measurements (*see* **Chapter 6, Principles of Hydrological Measurements, Volume 1**). If the extent of the data is small, the large-scale variability will not be captured and will translate into a trend in the data. If the support is large, most of the variability will be smoothed out and the data will appear very smooth. These sampling scale effects can be thought of as some sort of filtering in that the true patterns are filtered by the properties of the measurements (Cushman, 1984, 1987; Beckie, 1996; Di Federico and Neuman, 1997; Blöschl, 1999). In case of the modeling, the scale effects can be conceptualized in a similar way.

The different types of upscaling and downscaling (depending on what component of the scale triplet is changed) are illustrated in Figure 1. Downscaling in terms of spacing (i.e. decreasing the spacing) is usually referred to as *interpolation*, with the opposite being *singling out*. Upscaling in terms of extent (i.e. increasing the extent) is usually referred to as *extrapolation*; the opposite is, again, *singling out*. Upscaling and downscaling in terms of the support are referred to as *aggregation* and *disaggregation* respectively, particularly if the spacing is changing at the same time as the support. The scheme in Figure 1 can relate to both the sampling step (upscaling/downscaling from the underlying distribution to the data) and to the modeling step (upscaling/downscaling from the data to the model predictions). These two steps are conceptually similar.

First-order Scale Effects

To illustrate the effects of the scale triplet, both in sampling and modeling, the simplest case of linear upscaling and downscaling of a two-dimensional stationary random field such as that in Figure 1(a) will be examined first. It is assumed that the field can be fully characterized by a variogram of the form:

$$\gamma(h) = \sigma^2 \left(1 - \exp\left(\frac{-h}{\lambda}\right)\right) \tag{1}$$

where λ is the correlation length, σ^2 is the variance and h is the distance between two points in the random field (*see* equation 15 in **Chapter 7, Methods of Analyzing Variability, Volume 1**). For clarity, the random field is assumed to have zero mean and unit variance $\sigma^2 = 1$. An important assumption for first-order scale effects is that the variable of interest aggregates linearly or in other words, simple arithmetic averaging applies. In hydrology there are

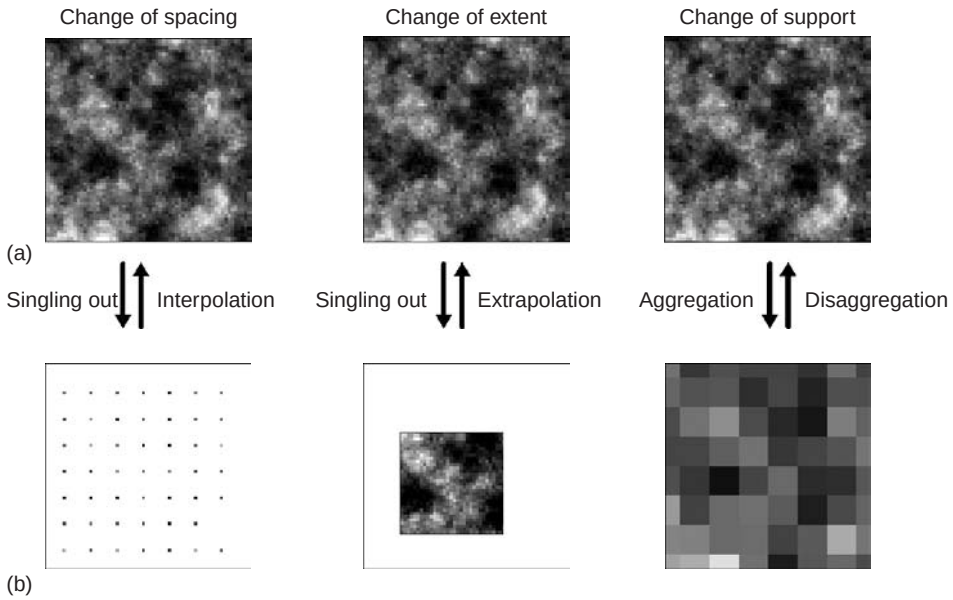


Figure 1 Schematic of upscaling and downscaling by changing the scale triplet. (a) represents the underlying natural variability with the sizes of the patches of the patterns being related to the integral scale. (b) shows the actual information reflected in the samples (or a model). Modified from Bierkens *et al.* (2000)

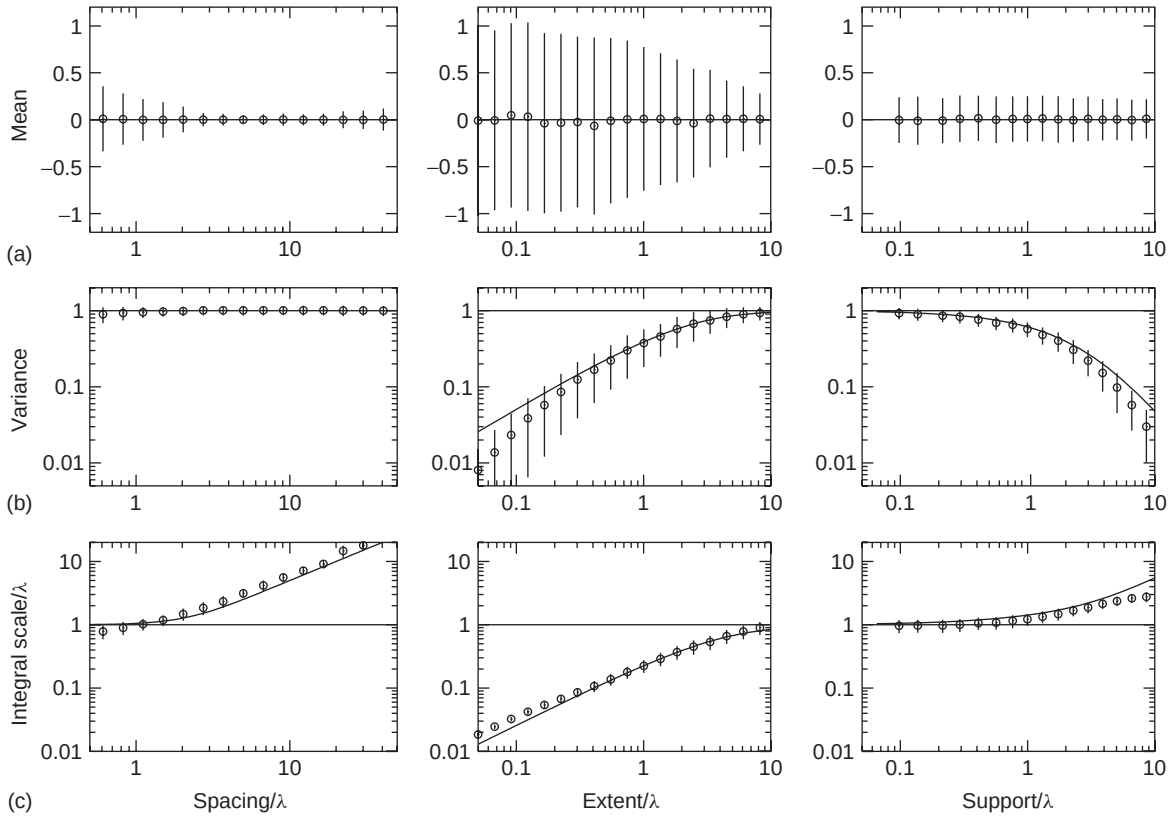


Figure 2 Effect of the sampling scale triplet on the sample mean, variance, and integral scale for the case of a two-dimensional stationary random field. λ is the correlation length of the random field. The circles show the ensemble mean and the error bars represent the standard deviation around the ensemble mean for 100 gridded samples from a Monte Carlo study. The solid lines show the predictions for the ensemble mean. From Skøien and Blöschl (2005)

variables where linear averaging is meaningful, such as precipitation, because a conservation law (e.g. of mass) holds. Other variables such as hydraulic conductivity do not average linearly. In Darcy's law, for example, the average hydraulic conductivity over an area does not give the average flux over the same area, so alternative, nonlinear methods are needed (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1** and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**).

To examine the effects of the scale triplet, Skøien and Blöschl (2005) performed a Monte Carlo analysis and generated a large number of realizations of two-dimensional random fields from which hypothetical samples at fixed locations conforming to a certain sampling scale triplet were drawn. These were used to estimate the sample mean, spatial variance, and integral scale for each realization. For each of these three characteristics, the mean over all realizations (the ensemble mean) and the standard deviation around the ensemble mean were calculated. Figure 2(a) shows the results for the sample mean. If the ensemble mean of the sample mean is zero, the estimates of the mean are unbiased. If the standard deviations (shown as error bars) are small, the uncertainty of the mean is small. The results suggest that the mean of the samples will be unbiased irrespective of the sampling scale triplet, but it will be highly uncertain if the extent of the domain is small relative to the correlation length of the underlying variability. This is because all samples are heavily correlated and very little is learned about the variability of the population. Figure 2(b) shows the results for the sample variance. If the ensemble mean of the sample variance is unity, the estimates of the variance are unbiased. Figure 2(b) indicates that large spacings (relative to the correlation length of the underlying variability) do not bias the variance, but small extents do and will lead to an underestimation of the variance. Large supports will reduce the variance and this is related to the smoothing effect of the support mentioned earlier.

Figure 2(c) shows the results for the sample integral scale. The integral scale is the average distance over which a variable is correlated and in the case of a variogram of the form of equation 1, the integral scale of the random field is equal to the correlation length λ (*see* equation 10 in **Chapter 7, Methods of Analyzing Variability, Volume 1**). In Figure 2, the sample integral scale has been scaled by λ, so if the ensemble mean of the scaled sample integral scale is unity, the estimates of the integral scale are unbiased. The results suggest that the integral scale will always be biased. It will be overestimated in the case of large spacings and large supports and underestimated in the case of small extents relative to the correlation length of the underlying variability. This is because in the case of large spacings and supports, the small-scale variability is not sampled, while in the case of small extents, the large-scale variability is not sampled. Figure 2 is for the case of gridded sampling. There are numerous other sampling schemes (Thompson, 2002) with slightly different biases and uncertainties for the case of large spacings. Random sampling, for example, will reduce the biases of the integral scale but will increase its uncertainty. The analysis in Figure 2 examines how well the characteristics of the underlying random field (i.e. the population) can be estimated from a limited number of samples. Skøien and Blöschl (2005) also examined how well the characteristics of a single (unknown) realization can be estimated and indicated that some of the biases and random errors will be different from the case shown here.

It is clear that spacing, extent, and support are all scales, but they have a different role in upscaling and downscaling methods. For example, the variance of, say, precipitation tends to increase with scale if scale is defined as extent, but decreases with scale if scale is defined as support because of the filtering involved.

To a first approximation, the effect of the filter will be closely related to the ratio of the sampling scale and the characteristic scale of the underlying process (the correlation length), which is consistent with dimensional reasoning. Geostatistical methods (Journel and Huijbregts, 1978) allow the estimation of the sampling biases in a consistent manner (see lines in Figure 2, taken from Western and Blöschl, 1999) and are the basis for some of the statistical upscaling/downscaling methods. The reduction in variance as a result of increasing support is widely used for linear aggregation methods and may also give some guidance for nonlinear cases such as extreme value analyses of rainfall and model parameter aggregation in subsurface hydrology.

TYPES OF STATISTICAL UPSCALING/ DOWNSCALING METHODS

Any upscaling or downscaling exercise involves the following steps: (i) analyzing the local data and scrutinizing the literature to decide on the model type; (ii) estimating the parameters from the data; (iii) verifying the upscaling/downscaling model against an independent data set; and (iv) performing the actual downscaling and upscaling step. Methods of upscaling and downscaling differ in terms of how they represent hydrologic variability.

Representing Variability

The simplest option is to represent the variability as a random variable that is fully characterized by its covariance (or equivalently its variogram, as the variogram can be expressed by the covariance and *vice versa*). Again, the simplest assumption in this case is that the random variable is multi-Gaussian and stationary which means that all linear

combinations of the variable sampled at any set of locations are normally distributed and there is no spatial/temporal trend. In zero dimensions (no spatial/temporal correlations), this type of variability is used in distribution models of soil moisture downscaling; in one dimension, for disaggregating annual rainfall into monthly values; in two dimensions (Figure 1a), for estimating catchment rainfall; and in three dimensions, for representing subsurface media characteristics. Representations of multi-Gaussian variability used in hydrology usually exhibit one single scale of variability (e.g. λ, if the variogram is of the form of equation 1). An alternative is to recognize that variability may occur at many scales. The variability may still be fully characterized by a variogram, but this time it is of the form:

$$\gamma(h) = a \cdot h^b \tag{2}$$

where h is the spatial (or temporal lag) and a and b are constants. This is termed *fractal variability* and is nonstationary, that is, the variance increases gradually with the extent of the field (*see* **Chapter 8, Fractals and Similarity Approaches in Hydrology, Volume 1**). This representation of variability can be used in a similar way as the multi-Gaussian concept discussed earlier, although with additional mathematical complexity, and has been widely used in hydrology, in particular, for rainfall where the presence of variability at many scales is most obvious.

There are more complex types of variability than those that can be exhaustively represented by a variogram (*see* **Chapter 1, On the Fundamentals of Hydrological Sciences, Volume 1** and **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**). The most important case is discontinuous (or intermittent) variability. Rainfall is discontinuous in both time and space (wet and dry), and subsurface media are discontinuous when there is a sharp transition of media properties between different facies or different soil types. This type of variability can be represented by Boolean methods where objects (such as storms, sand lenses, fractures) are placed in space (and in time in the case of storms) according to specified statistical distributions. Alternatively, discontinuous variability can be represented by transition probabilities (i.e. Markov chains) where a value is generated as a function of its neighboring value, both in time (in rainfall) and in space (e.g. subsurface media).

More complex forms of organization include the presence of connectivity or a tree structure, which can be represented by percolation characteristics or, in the case of soil moisture, by terrain indices. Another type of variability, gleaned from turbulence, is multifractal behavior where most of the variability is concentrated in singularities (Sreenivasan, 1991). This means, it is the peaks that control the upscaling/downscaling behavior which has been found to represent a wide range of hydrological characteristics very well, including rainfall, streamflow, and subsurface characteristics (Lovejoy and Schertzer, 1991; Gupta *et al.*, 1994; Menabde *et al.*, 1997; Veneziano and Essiam, 2003). It has been suggested that there may exist a causal relationship with turbulent processes in all of these cases. Multifractal model types include multiplicative cascades with the additional advantage of a small number of parameters, which facilitates parameter estimation.

Whatever the type of variability that is present, the important thing is that the methods of upscaling and downscaling represent these variabilities well. In other words, selection of one or the other method should be guided by an analysis of the underlying variability of interest in a particular case.

Modes of Application

Methods of upscaling and downscaling also differ in the way they are applied in the upscaling/downscaling step depending on what information – statistical characteristics or space-time patterns – is required.

The first generic task is to derive the statistics of a variable at one scale from the statistics of the same (or another) variable at another scale. Here, scale is usually the support scale in both instances. Examples include the downscaling of the output from global circulation models (GCMs), the areal reduction from point to catchment rainfall, and the effective parameter problem in subsurface and catchment hydrology (*see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**; **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**, and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). The first task can be upscaling or downscaling depending on which way one proceeds. Methods range in complexity from regressions between the variables at different scales to upscaling theory of stochastic hydrogeology (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**).

The second generic task is to generate spatial patterns (or time series) given the statistical characteristics of the variable one means to represent. This can be either through interpolation between a number of samples (using methods such as kriging with or without auxiliary data) or disaggregation. In the disaggregation case there are three variants (Deutsch and Journel, 1992; Bierkens *et al.*, 2000):

1. *Unconditional simulations (no data points given)*: Here the interest is in obtaining a number of realizations of the variable of interest that all exhibit the same statistics (e.g. the variogram) as inferred from the data. The patterns are not conditioned by the data points which means that the values of the simulated patterns at the locations of the data may be different from the data. Methods include the turning bands method for generating random fields and stochastic rainfall models for generating sequences of rainstorms.

2. *Conditional simulations (data points given)*: These are similar to unconditional simulations but they fit exactly to the individual data points. There is also some similarity with interpolation, but with the important difference that in interpolation one is interested in the most likely pattern while in conditional simulations one is interested in obtaining many realizations with the most realistic statistical characteristics.

3. *Conditional simulations (averages given)*: These are similar to the above but now the patterns are conditioned on the averages over support areas or time periods. Examples include the generation of subsurface media patterns by disaggregating pumping test data and the generation of monthly rainfall time series by disaggregating annual rainfall.

The opposite scale transfer of obtaining averages from patterns is nontrivial only in the case of processes that average nonlinearly as most model parameters do. The two main types of averaging methods are volume averaging (where a model parameter and/or an equation is averaged over a support volume) and ensemble averaging (where the model parameters and/or equations are averaged over all possible realizations of an ensemble at a single point). This, however, is beyond the scope of this article and the reader is referred to **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1** and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**.

STATISTICAL UPSCALING AND DOWNSCALING METHODS

This section reviews statistical upscaling and downscaling methods for a number of hydrological application areas. Research into upscaling and downscaling rainfall, historically, has followed three routes. The first has been driven by the need for estimating catchment averages of extreme rainfall from point data for hydrologic design, the second from the need of estimating, say, daily rainfall from monthly data both in applied and theoretical contexts, and the third has been triggered by the interest of deriving local rainfall from estimates of average rainfall over tens of thousands of square kilometers provided by GCMs in the context of climate impact assessments (Sections "Upscaling point rainfall to catchments", "Temporal disaggregation of rainfall", and "Statistical downscaling of the output of global circulation models"). In flood frequency analysis, the challenge has been, and still is, to transfer flood characteristics between catchments of different scales and here the main focus of the more involved methods has been on a better understanding of regional flood generation processes (Section "Flood frequency as a function of catchment scale"). An important issue in catchment hydrology and, increasingly, in climate modeling is how to upscale and downscale measurements and simulation results of near-surface soil moisture (Section "Upscaling and downscaling soil moisture"). In subsurface hydrology, one of the main issues is how to represent the media characteristics be it in soils, porous aquifers, or in fractured rocks from data at different scales as the heterogeneities control the dynamics of flow and transport (Section "Subsurface media characterization and generation").

Upscaling Point Rainfall to Catchments

One of the classical upscaling problems in catchment hydrology is how to obtain catchment rainfall from observed point rainfall data to drive runoff models. Because of the averaging effects, extremes of catchment rainfall tend to be smaller than those of point rainfall. To account for this effect, areal reduction factors (ARFs) have been introduced which are defined as the ratio of catchment rainfall and point rainfall. Two kinds of ARFs are presently in use (Srikanthan, 1995).

1. Fixed-area (also known as *geographically fixed*) ARFs relate rainfall at any arbitrary point, that is, a point rainfall estimate to an average over a catchment, which is fixed in space. They are estimated by constructing from all available station rainfall data, the time series of catchment average rainfall (e.g. using the Thiessen polygon method), performing an extreme value analysis on them as well as on the point data, and finally relating the catchment rainfall intensities to the point values, for the same return period and duration.
2. Storm-centered ARFs refer to a given storm. They represent the ratio of average storm depth over an area (defined by the rainfall isohyets) and the maximum rainfall depth for the storm (at the storm center). Storm-centered ARFs are usually somewhat smaller than fixed-area ARFs. Storm-centered ARFs are used more commonly in the estimation of probable maximum floods, while the fixed-area ARFs are used for designing hydraulic structures for flood control, for example, bridges and culverts.

The main controls on reduction factors are the storm type with small-scale convective storms producing a more pronounced decrease of the ARF with area than large-scale storms, and duration with shorter durations being associated with a stronger decrease of the ARF with area (Srikanthan, 1995). The original concept of the ARFs has been a purely empirical one but a number of recent contributions have attempted to provide a sounder theoretical basis for them. Bacchi and Ranzi (1996) derived ARFs on the basis of the crossing properties of the rainfall process aggregated in space and time assuming that the number of crossings of high rainfall intensity levels is Poisson-distributed. The

ARFs so obtained showed a power-law decay with respect to area and duration of the storm, and they showed a slight decrease with respect to the return period. In a somewhat related analysis, Skaugen *et al.* (1996) analyzed the fraction of a catchment with rainfall depths greater than a threshold for different thresholds, which allowed them to derive catchment average rainfall. De Michele *et al.* (2001) derived ARFs from power law or fractal characteristics of the underlying process in space and time. In an alternative approach, Sivapalan and Blöschl (1998) derived ARFs from the spatial correlation structure of rainfall. They averaged the parent distribution of point rainfall over a catchment area making use of the variance reduction with support (Figure 2) and then transformed it to the corresponding extreme value distribution, using the asymptotic extreme value theory of Gumbel (1958). Figure 3 shows the ARFs so obtained for a duration of 24 h. It is interesting that, similar to Bacchi and Ranzi (1996), there is a tendency for the ARFs to decrease with return period, which is related to decreasing correlation lengths of rainfall as the storms increase in magnitude.

There have been a number of attempts at linking the areal reduction factor concept to other representations of areal precipitation. Booij (2002), for example, linked the Sivapalan and Blöschl (1998) ARFs to statistical downscaling methods (see Section "Statistical downscaling of the output of global circulation models"), and Venugopal *et al.* (1999) proposed a downscaling scheme that preserves both the spatial and temporal correlation of rainfall as grid size changes, which is closely related to deriving ARFs from the correlation structure of rainfall. In principle, ARFs can be derived from multisite stochastic rainfall models (see Sections "Temporal disaggregation of rainfall", "Statistical downscaling of the output of global circulation models"). The latter may be physically more realistic but involve significantly more parameters than empirically derived ARFs. A large number of parameters may be difficult to estimate from limited data.

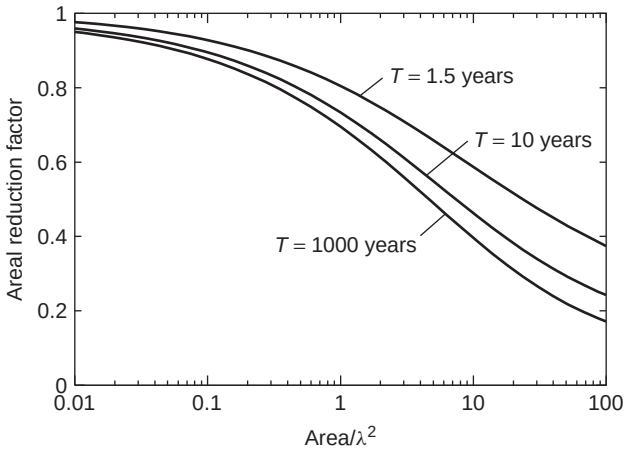


Figure 3 Areal reduction factors derived from the rainfall spatial correlation structure (Sivapalan and Blöschl, 1998). The time interval is 24 h, A/λ^2 is the catchment area scaled by the square of the rainfall correlation length and T is the return period

Temporal Disaggregation of Rainfall

Rainfall data with low temporal resolution (large temporal supports) often are available at more stations and over a longer time period than high-resolution records. Disaggregating them in time by stochastic rainfall models is therefore attractive for a range of practical purposes including the simulation of reservoir operation, urban drainage, and design, but is also interesting from a theoretical perspective for understanding rainfall processes. As with most other downscaling methods, there are two basic cases – disaggregating observed time series under mass constraints for each observed time step, that is, conditional simulations, and generating time series with small temporal supports with given statistical characteristics, that is, unconditional simulations. A range of model types have been developed (Foufoula-Georgiou and Georgakakos, 1991; Foufoula-Georgiou and Krajewski, 1995; Srikanthan and McMahon, 2001):

1. *Linear disaggregation models*: The classic model of disaggregating annual rainfall into monthly rainfall has been proposed by Valencia and Schaake (1973). It is a linear disaggregation model which assumes that the monthly values can be estimated by a linear combination of the annual values plus an additive error term. These are conditional simulations. The model has been extended in a number of ways (see Salas, 1993, pp. 19–32) and has probably been the most widely used scheme for stochastic disaggregation problems in hydrological applications. However, schemes of this kind are not suitable for the disaggregation of rainfall for timescales finer than monthly due to the skewed distributions and the intermittent nature of the rainfall process at fine timescales (*see* **Chapter 7, Methods of Analyzing Variability, Volume 1**).

2. *Two-part models*: Two-part models consist of a model for the occurrence of dry and wet days and a model for the generation of rainfall amount on wet days. Models of rainfall occurrence are of two main types, those based on Markov chains and those based on alternating renewal processes. Markov chains specify the state of each day as "wet" or "dry" and develop a relation between the state of the current day and the states of the preceding days. Long-term persistence can be represented by Hidden State Markov models (Thyer and Kuczera, 2003). In alternating renewal models (Grace and Eagleson, 1966, Sivapalan *et al.*, 2005), sequences of event and between-event periods are simulated. The wet and dry spells are usually assumed to be independent and may conform to different distributions

that may also vary with the season. The rainfall amount models use various distribution types such as the Gamma distribution or the Levy-stable distribution (Menabde and Sivapalan, 2000). Total event rainfall amount can then be disaggregated to, say, hourly time steps by traditional mass curves. In mass curves, the traces of cumulative storm depths normalized by total storm depth versus cumulative time since the beginning of a storm normalized by the storm duration are assumed to follow a set pattern (Huff, 1967). Within-event rainfall patterns can also be obtained by random disaggregation (Woolhiser and Osborn, 1985).

3. *Point process models*: These are models where the stochastic process is completely characterized by the position of its events (Eagleson, 1978; Waymire and Gupta, 1981; Smith and Karr, 1985). Among the point process models, the most widespread methods have been the cluster-based models such as the Neyman-Scott rectangular pulse model (Favre *et al.*, 2004) and the Bartlett-Lewis rectangular pulse model (Onof and Wheater, 1993). In these models, storm arrival follows a Poisson process and each storm gives rise to a cluster of rain cells with each cell having a random time location, duration, and intensity. These models have been extended to multiple sites (e.g. Fowler *et al.*, 2005) but this is at the cost of increased mathematical complexity and potential problems with parameter estimation. The seasonal variation in rainfall is an important factor. Some of the models deal with seasonality by assuming that the parameters vary seasonally, in other models the parameters have been linked to weather types, and there are also combinations of these two options (see Section "Statistical downscaling of the output of global circulation models"). **Chapter 125, Rainfall-runoff Modeling for Flood Frequency Estimation, Volume 3** discusses point process models in the context of flood frequency analysis. Point process models are usually applied as unconditional simulations.

4. *Multiplicative cascades*: The idea in multiplicative cascades is that an initial rainfall depth over a time period is split into two subperiods and these subperiods are split again and so forth. In an analogous way, rainfall over an area is split into subsequently smaller subareas. Multiplicative cascades are based on the notion that the underlying variability is multifractal. The main strength of multiplicative cascade models is that they have relatively few parameters and hence these can be estimated more robustly than is possible with point process models. The other advantage of multiplicative cascade models is that they are more easily able to reproduce observed rainfall characteristics over a wide scale range than other models (Foufoula-Georgiou and Krajewski, 1995). Most of the recent research has hence focused on models where at least one variability aspect is represented by multiplicative cascades. Examples include Olsson and Berndtsson (1998), Güntner *et al.* (2001), and Seed *et al.* (1999) who proposed a space and time model for design storm rainfall. Multiplicative cascade models lend themselves naturally to disaggregating storm rainfall (i.e. conditional simulations) and have been linked to the mass curves concept by Koutsoyiannis and Foufoula-Georgiou (1993) and Koutsoyiannis and Mamassis (2001).

5. *Combined models*: The above classification of models has been based on their statistical concepts, and there a number of models that combine some of these concepts. Jothityangkoon *et al.* (2000), for example, use a Markov chain in the time domain and multiplicative cascades in the space domain. Another example has been furnished by Koutsoyiannis *et al.* (2003) who proposed a multisite rainfall generation framework based on a combination of methods. There exists a range of rather simple and practical disaggregation methods that combine still other concepts. These methods are used in a number of countries at a national level and include regression methods (Canterford *et al.*, 1987) and the quadrant method (Grebner, 1995) where within-day temporal patterns of rainfall are transposed from the neighboring recording raingauge stations on the basis of a similarity measure. The strength of these simple methods are their robustness, although their accuracy relative to other methods has not yet been fully evaluated.

Statistical Downscaling of the Output of Global Circulation Models

Outputs of climate simulations from global circulations models (GCMs) cannot be directly used for hydrological impact studies of climate change because of a scale mismatch. The spatial grid resolution (i.e. support) of GCMs in use today is on the order of tens of thousands of square kilometers. The useful grid box size is even larger as GCMs are inaccurate at the scale of a single grid box. In contrast, the spatial scale at which inputs to hydrologic impact models are needed is on the order of tens or hundreds of square kilometers. Because of the scale mismatch, the statistical characteristics of the GCM output may be vastly different from those of the local (surface) variable that shares the same name. For example, the maximum rainfall intensities simulated by GCMs tend to be much smaller than those at the point scale. There will also be local effects induced by topography, land cover, and so on, which are not captured in the GCM. Downscaling methods can be used to transfer the large-scale GCM output to small-scale variables and to account for local effects.

There are two approaches to downscaling GCM output (IPCC, 2001; Yarnal *et al.*, 2001). The first is dynamic downscaling where deterministic regional climate models are nested into GCMs. This means the initial conditions

and boundary conditions to drive the regional climate model are taken from the GCMs. Dynamic downscaling is discussed in **Chapter 32, Models of Global and Regional Climate, Volume 1** and is not dealt with in this article.

The second approach is empirical or statistical downscaling, which will be briefly reviewed here in a hydrological context. In empirical or statistical downscaling, explicit relationships between the large-scale GCM output and the observed small-scale or local station data such as precipitation are used. Unlike dynamic downscaling, statistical downscaling methods are computationally inexpensive. They can thus be used to generate a large number of realizations to assess the uncertainty of predictions. They can also use climate data from individual stations directly, so local information can be accounted for in an efficient way. The method, however, hinges on the assumption that the statistical relationships developed for present day climate also hold under the different forcing conditions of possible future climates, and this assumption is essentially unverifiable. The relationships can indeed be unstable for many reasons as short-term relationships can be conditional on longer-term variations in the climate system (Charles *et al.*, 2004). Application of the method consists of four steps as follows (Yarnal *et al.*, 2001):

1. *Selection of the local variable*: The most common variables to be estimated by downscaling procedures are precipitation and air temperature, either at daily or monthly timescales. The corresponding data are usually collected at individual climate stations in the area.

2. *Selection of the large-scale GCM-derived variable or variables (termed "predictors")*: This selection step is less obvious and depends on the downscaling relationship used. Ideal predictors exhibit high predictive power, are accurately simulated by the GCM, and are associated with relationships that are time stable. The most widely used predictor is sea level pressure because of its long record, which aids model development. Geopotential heights are also frequently used. Most common is 500 hPa, a level representing midtropospheric circulation and storms tracks. When empirical downscaling of precipitation is the goal, it is also useful to include a predictor of atmospheric moisture because changes in the hydrologic cycle are likely to be the underlying cause of future changes in precipitation. It is important to note, however, that GCMs tend to provide less accurate simulations of atmospheric moisture than of surface pressure and geopotential heights (Yarnal *et al.*, 2001). Wilby and Wigley (2000) and Wilby *et al.* (2002) discuss the relative merits of predictors currently in use.

3. *Deriving relationships between the observed small-scale or local station data and the large-scale GCM-derived variable*: The relationships (i.e. the downscaling models), generally, belong to one out of three types of methods (IPCC, 2001).

(a) Regression techniques are the most widely used methods. The simplest is linear regression using grid-cell values from the GCMs as predictors (e.g. Wilby *et al.*, 1998) or principal components of the predictor fields (Hewitson and Crane, 1996). Canonical correlation analysis and singular-value decomposition are other methods that condense the spatial patterns into a few values amenable to the regressions (von Storch and Zwiers, 1999). In addition to linear regression, nonlinear models such as nonlinear interpolation (Brandsma and Buishand, 1997) and artificial neural networks (Hewitson and Crane, 1996) have been developed. In the regression models, for each time step the "best" value of the local variable is generally estimated from the large-scale variables at the same time step.

(b) Stochastic models and weather typing: Stochastic models have been discussed in Section "Temporal disaggregation of rainfall" of this article. They do not estimate a best value for each time step but generate possible realizations of time traces of precipitation that are consistent with the statistics of the observations, given the probabilities derived from the GCM output. In the context of GCM downscaling, most models of this type condition the probabilities of the local variable of interest on weather types (e.g. Hay *et al.*, 1991). A number of weather classification schemes are in use (*see* **Chapter 26, Weather Patterns and Weather Types, Volume 1**, also see Tveito and Ustrnul, 2003), such as the Lamb weather types for the United Kingdom (Conway and Jones, 1998) and the Central European Großwetterlagen (Stehlik and Bardossy, 2002). The classification of each day can be either manual or automatic, involving various statistical methods based on the patterns of geopotential pressure heights and other synoptic variables (Yarnal *et al.*, 2001). The time step of the stochastic models is usually a day. Wilks and Wilby (1999) and Srikanthan and McMahon (2001) review stochastic climate models with an emphasis on precipitation.

(c) The analog method is, again, based on a classification procedure, but no stochastic model is involved. The method consists of identifying, from a pool of historical circulation patterns, the one that is most similar to the circulation pattern on the day of interest (Lorenz, 1969). The local variables observed on the most similar day within the pool are then assigned to the day of interest. The similarity of the large-scale circulation patterns can be defined in various ways. Zorita and von Storch (1999), for example, proposed Empirical Orthogonal Functions to reduce the

degrees of freedom of the large-scale atmospheric circulation. While there may rarely exist perfect analogs from the past, the method does not require any transformation of the local variable, which is a particular advantage for the case of precipitation. This method generates physically feasible spatial patterns of the surface variables but the predicted pattern is restricted to the observed values.

All of the methods need to be tested whether the relationships remain stable over time, which is usually done by split-sample testing involving nonoverlapping periods for parameter estimation and verification.

4. As a final step in the downscaling procedure, the relationships derived are applied to the GCM output for changed climate scenarios to estimate the local variables for a changed climate. These local variables (mainly precipitation and air temperature) can then be used to drive hydrological models in an impact assessment.

Flood Frequency as a Function of Catchment Scale

Regional analyses of flood frequency are used in practical contexts for estimating floods at ungauged sites and improving flood estimates at gauged sites (Cunnane, 1988), and in more theoretical contexts for advancing the understanding of the spatial variability of hydrologic fluxes (Pilgrim, 1983; Blöschl and Sivapalan, 1995; Gupta, 2004). To sharpen the focus on the pure scale effect, one usually assumes that the spatial trends of the flood characteristic are either small, or average out when a sufficiently large ensemble of catchments is considered. Spatial trends are dealt with separately in regional flood frequency analysis (e.g. Institute of Hydrology(IH), 1999; Merz and Blöschl, 2005).

The flood frequency curve is likely to change with catchment size because of a number of factors that are related to rainfall (e.g. rainstorms of different sizes may become important) and catchment processes (e.g. different runoff processes may become important). The main effect of catchment size is that the specific mean annual flood (i.e. the mean of the maximum annual flood peaks divided by catchment area) tends to decrease with catchment scale (Eaton *et al.*, 2002) and this is related to the smoothing effects of support scale as illustrated in Figure 2. The steepness of the flood frequency curve (which is closely related to the coefficient of variation, CV) changes with catchment size in a less predictable way and hence has recently attracted considerable attention in the literature.

Some practical methods, such as the index flood method (Dalrymple, 1960), imply that the CV in any homogeneous region does not change with catchment area (Gupta *et al.*, 1994), but there is evidence to the contrary from two avenues of enquiry.

1. The first are quantile-based analyses where floods of a given return period (i.e. flood quantiles) and catchment area are often formulated as a power law:

$$Q_T = c(T) \cdot A^{\theta(T)} \tag{3}$$

where Q_T are the flood quantiles, c is a factor, T is the return period, A is catchment area, and θ is an exponent. Data analyses from a range of climates suggest that there is a tendency for θ to decrease with return period (e.g. Gupta and Waymire, 1998). Typically, θ varies between 0.6 and 0.9 for $T =$ 2 years and between 0.4 and 0.6 for $T >$ 100 years. This dependence on the return period implies a decrease of CV with catchment area.

2. The second type of analysis examines the flood moments. Flood data in the central Appalachians indicated that the CVs tend to increase with catchment scale from 1 to 100 km^2 and decrease between 100 and 25 000 km^2 (Smith, 1992). Smith (1992) proposed two interpretations. First, the data on the small basins may be unrepresentative due to possible systematic errors in stream gauging or sampling artifacts, which would suggest that CV tends to always decrease with catchment scale. Second, the peak in CV at a catchment size around 100 km^2 may be real and is related to the organization of extreme storm rainfall and the downstream development of the channel/floodplain system. Gupta and Dawdy (1995) provided an additional interpretation. On the basis of rainfall runoff simulations they suggested CV at small catchment scales to be controlled by basin response, and based on a comparison of snow melt and rainfall dominated catchments they suggested CV at large scales to be controlled by rainfall variability. A similar interpretation has been furnished by Robinson and Sivapalan (1997) based on a derived flood frequency model. They suggested that the increase in CV in small catchments is due to the interaction between timescales of storm duration and catchment response while the decrease in CV in large catchments is due to the spatial rainfall characteristics. Blöschl and Sivapalan (1997) used a similar model and noted that the interactions between timescales of storm duration and catchment response may be hidden by other processes such as nonlinear runoff generation.

While these findings were obtained by an upward approach (Sivapalan *et al.*, 2003) of deriving the flood frequency statistics from rainfall statistics, an alternative, downward route for explaining the change of the flood frequency curve with catchment scale has been taken by Merz and Blöschl (2003). They classified 12 000 flood peaks in Austria into long-rain floods, short-rain floods, flash-floods, rain-on-snow floods, and snow-melt floods and then examined the flood statistics separately for each of the groups (Figure 4). They found that the CV of the snow-melt flood type exhibited the flattest decrease with

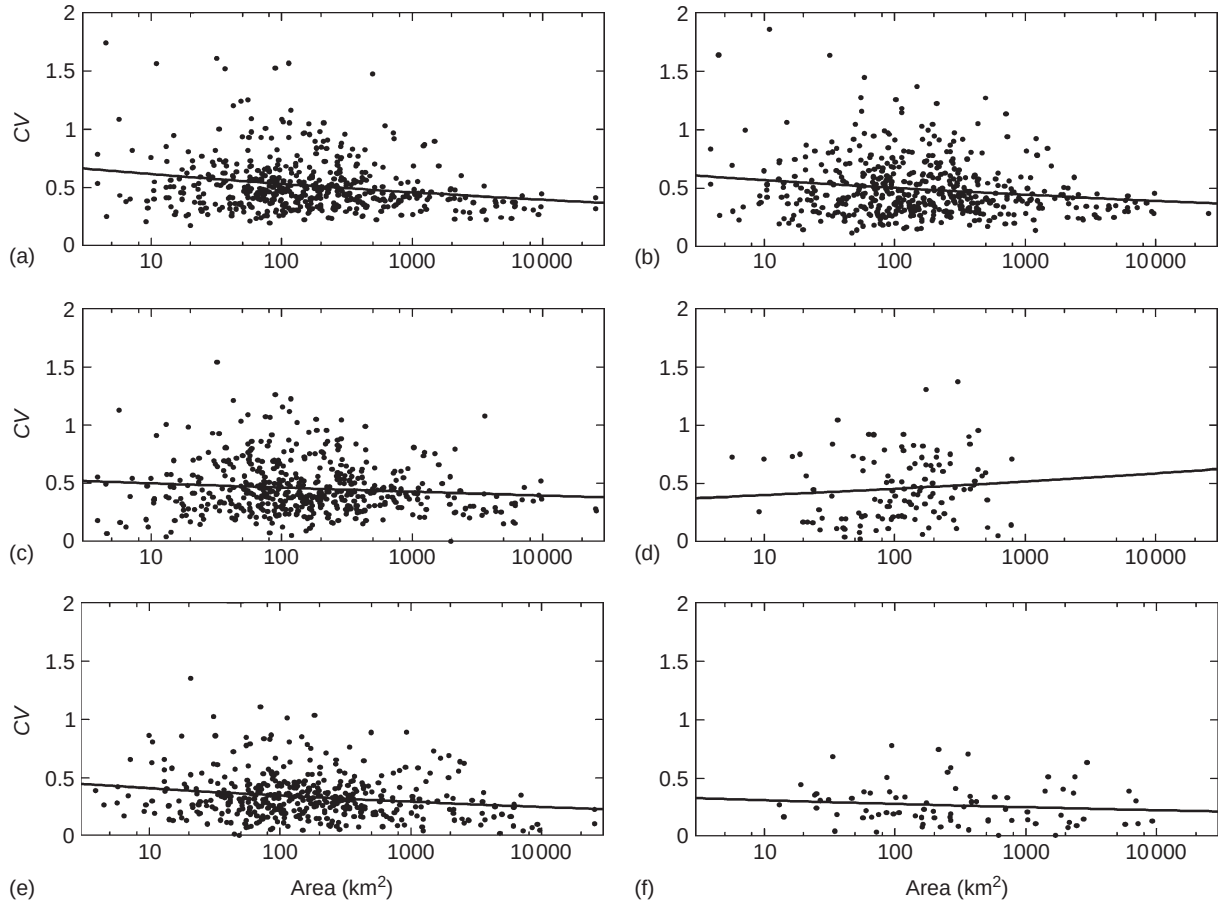


Figure 4 Coefficient of variation of annual floods in Austria stratified by process type and plotted versus catchment area. (a) all flood types; (b) long-rain floods; (c) short-rain floods; (d) flash-floods; (e) rain-on-snow floods; and (f) snow-melt floods. Regression lines are shown. From Merz and Blöschl (2003)

catchment area (Figure 4f) which is consistent with the usually large extent of snow melt. The *CV* of the flash-flood process type, however, tended to increase with catchment area (Figure 4d), which was interpreted as been related to the nonlinearity of runoff generation associated with fast hillslope response.

An extension of the moment-based analyses to higher order moments has been performed by a number of authors (e.g. De Michele and Rosso, 2002). These analyses generally suggest that the flood statistics exhibit fractal or multifractal characteristics depending on how the exponents change with the order of the moment (*see* **Chapter 8, Fractals and Similarity Approaches in Hydrology, Volume 1**). These methods of analysis are similar to those used in rainfall disaggregation based on multiplicative cascades (Section "Temporal disaggregation of rainfall").

Upscaling and Downscaling Soil Moisture

An accurate representation of the spatial variability of near-surface soil moisture is critically important both for representing hydrological fluxes in the subsurface at various scales (Zehe and Blöschl, 2004) and for linking hydrological processes with atmospheric processes (*see* **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**; also see Ronda *et al.*, 2002; Montaldo and Albertson, 2003). There are two basic issues of soil moisture upscaling/downscaling in the hydrological sciences. The first is how to best estimate average catchment soil moisture (or spatial distributions) from point soil moisture measured in the field (*see* **Chapter 6, Principles of Hydrological Measurements, Volume 1**), the second is how to best estimate patterns of soil moisture from catchment average soil moisture as simulated by a hydrologic or atmospheric model. The first issue is an upscaling task while the second one is downscaling. In both instances one can use deterministic process-based models which is dealt with in **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**; **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**,

and in the literature (e.g. Blöschl and Grayson, 2000; Pitman, 2003). Alternatively, one can use simplified statistical descriptions that aim at representing the most important controls and these will be briefly reviewed here. These methods can either exploit the spatial statistics of soil moisture or make use of auxiliary information in terms of a moisture index (Western *et al.*, 2002, 2003).

1. *Methods based on the spatial statistics*: A number of authors have suggested that the spatial distribution function of soil moisture can be approximated by a normal distribution although the shape of the distribution does change with climate (e.g. Mohanty *et al.*, 2000; Nyberg, 1996). The variance of the spatial distribution of soil moisture, when taking the numerous studies around the world together (Western *et al.*, 2003), tends to depend on mean catchment moisture, indicating a pattern of variance that increases from near zero at wilting point to a peak at moderate moisture levels and then decreases to near zero as the mean moisture approaches saturation. Understanding of the spatial distribution of soil moisture has been used in distribution models to estimate runoff generation and evaporation (Beven, 1995; Wood *et al.*, 1992; Zhao, 1992). It is interesting that the shape of the distribution functions varies widely between models, but the models are similarly successful in predicting catchment runoff.

Studies on the spatial correlation of soil moisture have been summarized by Western *et al.* (2004). Typical correlation lengths vary between 1 m and 600 m and there is a tendency for the correlation lengths to increase with extent and spacing of the data as would be expected given the sampling scale effects illustrated in Figure 2. While some of the small-scale catchment studies suggest that the spatial soil moisture variability is stationary (similar to equation 1), analyses of remotely sensed soil moisture have found a fractal behavior (e.g. Hu *et al.*, 1997). Ground-based point data collected over large areas in the former Soviet Union, Mongolia, China, and the United States suggest that soil moisture variation could be represented as a stationary field with a correlation length of about 400 to 800 km (Entin *et al.*, 2000). Part of the differences in correlation lengths in small-scale and large-scale studies may again be related to sampling effects, although there also appear to exist important changes in the process controls with scale causing such differences. Over short scales, climate may be relatively uniform and the variation may be mainly related to differences in soils and vegetation (Seyfried, 1998), while at larger scales climate may be a dominant source of soil moisture variability. Methods of upscaling and downscaling that are based on the spatial statistics involve a wide spectrum of geostatistical methods to obtain spatial patterns or averages from point data or to obtain spatial patterns from (simulated) catchment average soil moisture (e.g. Deutsch and Journel, 1992). These methods include conditional simulation methods based on the assumption that soil moisture is a Gaussian random field. Geostatistical methods can also be used to derive analytical estimates of how, say, the runoff contributing area in a distributed model will change with grid size (Western and Blöschl, 1999).

2. *Index approach*: In the index approach, spatial organization can be imposed on the soil moisture field that goes beyond the Gaussian random field of the previous method by using landscape characteristics. These characteristics are usually condensed into an index for numerical efficiency guided by the understanding one has about the movement of water in the landscape (Moore *et al.*, 1991). In humid climates, lateral redistribution of moisture by shallow subsurface flow can be an important process and in this case, indices reflecting upslope area, slope, or convergence should be related to the soil moisture. The most commonly used index is the topographic wetness index of Beven and Kirkby (1979) (see also O'Loughlin, 1986) which is defined as

$$w = \ln\left(\frac{a}{\tan(b)}\right) \tag{4}$$

where a is the specific contributing area and b is the surface slope. Terrain data are widely available and there exists sophisticated terrain analysis software (*see* **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**; Wilson and Gallant, 2000). Because of this, equation 4 is widely used for upscaling and downscaling soil moisture. The index involves a number of assumptions, some of which have been relaxed recently. Barling *et al.* (1994), for example, relaxed the steady-state assumption and Woods *et al.* (1997) relaxed the assumption of uniform recharge. Western *et al.* (1999) examined the predictive ability of various terrain indices against soil moisture data collected in a small humid catchment in southeastern Australia. The wetness index (equation 4) typically explained 50% of the spatial soil moisture variance during the wet season (Figure 5a) and there were other indices that showed a similar performance such as the tangent curvature of the terrain (Figure 5b). The largest soil moisture values were collected in the gullies that exhibit large specific contributing areas and strongly negative tangent curvature. However, as the catchment dried out, the explanatory power of the indices dropped off rapidly. Western *et al.* (1999) also summarized tests of terrain indices in various climates and noted that their predictive ability varies substantially, depending on whether their main assumptions are satisfied.

All of these indices can be used to estimate a spatial pattern (or a spatial distribution) from average catchment soil moisture (i.e. downscaling) and to estimate a spatial

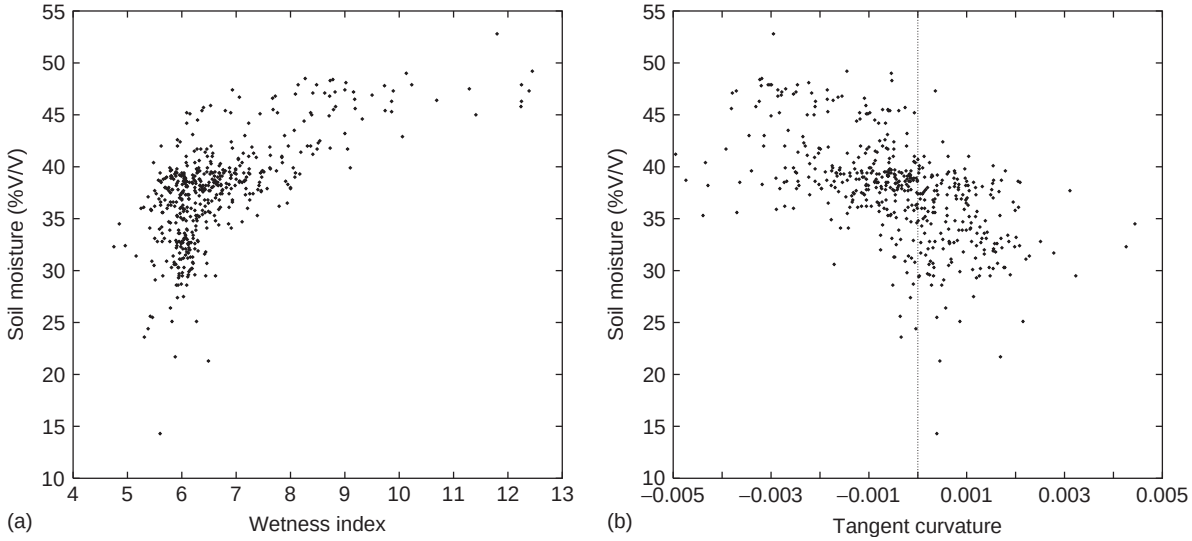


Figure 5 Relationship between volumetric soil moisture in the top 30 cm of the soil profile and wetness index, and tangent curvature for September 27, 1995 in the Tarrawarra catchment, Australia. From Western *et al.* (1999)

pattern from point measurements. In the latter case, similar geostatistical methods can be used as discussed above, but they are extended to accommodate the index as an auxiliary variable. Methods include external drift kriging, cokriging, and georegression (e.g. Grayson and Blöschl, 2000). Examples of applications include those by Viney and Sivapalan (2004) who disaggregated areal average soil moisture into spatial patterns and by Green and Erskine (2004) who compared a geostatistical analysis with linear georegression using terrain indices. In spite of the considerable progress that has been made in the past decades on soil moisture upscaling and downscaling, there is still significant uncertainty involved because of the large natural variability in soil moisture and its controls. If only a few point soil moisture measurements are available in a catchment, the errors associated with upscaling them to catchment averages can be enormous (Grayson *et al.*, 2002). An interesting extension of the index approach has therefore been proposed by Grayson and Western (1998). They suggested that concepts of time stability, applied to catchments with significant relief, could be used to identify certain parts of the landscape that consistently exhibit mean behavior irrespective of the overall wetness. They denoted these areas as *catchment average soil moisture monitoring* (CASMM) sites. This approach promises to assist in the upscaling issue if point measurements of soil moisture can be located in these areas.

Subsurface Media Characterization and Generation

One of the important tasks in subsurface hydrology is to reconstruct the spatial distribution of subsurface characteristics from limited data from boreholes, outcrops, water tables, and perhaps tracer tests. Knowledge on the subsurface characteristics is critical for understanding the movement of contaminants in the subsurface be it through conceptual analyses or through groundwater flow and transport models. Once the type of variability is known, one can derive the aggregated characteristics of the flow system either by analytical methods of stochastic hydrogeology or by performing flow and transport Monte Carlo simulations based on the generated media (*see* **Chapter 147, Characterization of Porous and Fractured Media, Volume 4** and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). Methods of characterization and generating subsurface media have been summarized by Koltermann and Gorelick (1996) and Anderson (1997). Findikakis (2002) and Zhang (2002) provide a collection of recent contributions, and Wang (1991) summarizes methods for the case of fractured rocks. Nonfractured sedimentary deposits can be represented by either continuous or discrete methods:

1. *Continuous heterogeneity*
   (a) Gaussian models for representing the spatial variability have been the starting point in stochastic theories of estimating hydraulic conductivity and dispersion coefficients for various supports (Gelhar and Axnes, 1983; Dagan, 1989; **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). A range of methods are in use to generate media including conditional and unconditional stochastic simulations (Deutsch and Journel, 1992).
   (b) Fractal models: Subsurface heterogeneity may occur at a range of scales for which a fractal representation

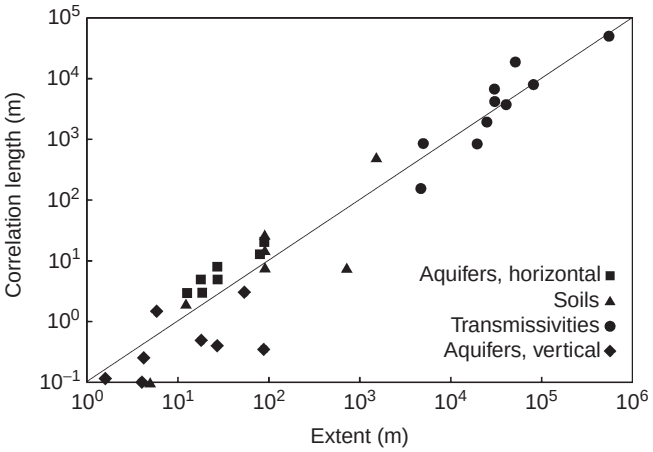


Figure 6 Correlation lengths of natural log hydraulic conductivities and transmissivities at various sites versus extents of the fields. Data from Gelhar (1993, Table 1). Line shows correlation length equal to 10% of the extent

may be more appropriate. Correlation lengths of hydraulic conductivity of aquifers (Figure 6) tend to increase with the extent of the study areas, which has been interpreted in two ways. The first interpretation is the existence of a fractal variogram of the form of equation 2. This type of variability has been used widely for media characterization and generation (Hewett, 1986; Hewett and Deutsch, 1996; Rubin and Bellin, 1998) and for stochastic theories of upscaling (e.g. Neuman, 1990, 1993). However, the conceptual difficulty with this is that geologic features observable in the field do indicate the existence of distinct scales in cores, fluvial deposits, alluvial basins, and in interbasin aquifers (Anderson, 1991). As an alternative interpretation, Gelhar (1993, p. 295) suggested that the variogram may be nested, consisting of a linear combination of variograms of the form of equation 1 with different λ_i which in the limit for many scales may resemble equation 2. It should also be noted that there may exist sampling biases and inference issues (Gallant *et al.*, 1994; Gray, *et al.*, 1993). Indeed, Figure 2 indicates that for a limited number of samples, the correlation length will always be on the order of 10% of the extent of the domain irrespective of the underlying correlation structure.

2. *Discrete heterogeneity in nonfractured sedimentary deposits*

(a) At some scale, geological media are not continuous but exhibit facies transitions and/or geological structures that affect the hydraulic conductivity field. Facies are units of similar characteristics. Geological structures include large-scale features such as faults and finer scale features such as bedding planes (Bierkens and Weerts, 1994). Conceptual models of the shapes and relative relations of the facies in a particular geological setting can be used for both upscaling and downscaling by assigning media characteristics to each of the facies (Anderson, 1997).

(b) Indicator geostatistics have been used to exploit these facies models for generating media characteristics in a consistent way. Indicator variograms are variograms of binary values (1 or 0 depending on whether the variable is above or below a threshold) and so may exhibit different shapes for different magnitudes of the variable of interest. These variograms can then be used for unconditional and conditional simulations as well as for interpolation between borehole data. Desbarats (1987), for example, generated a dual conductivity field of sandstone with small-scale shale features based on the indicator approach. Ritzi *et al.* (2000) compared the indicator approach with other methods of representing facies in a buried-valley aquifer setting.

(c) The spatial connectivity of high conductivity regions, both in soils and in aquifers, can be critically important to flow and solute movement at various scales. While it has been suggested that the indicator approach can also capture connectivity (e.g. Journel and Alabert, 1989; Gómez-Hernández and Wen, 1997), Western *et al.* (1998, 2001) noted that its ability is rather limited and it may be more efficient to use methods that directly capture connectivity such as connectivity functions based on percolation characteristics. Such a connectivity function represents the probability that two points are connected as a function of distance and hence is a reflection of preferential flow paths. Media patterns can then be generated by optimization methods such as simulated annealing where an image is generated iteratively by swapping pixels in a way that a given connectivity function is reproduced (Koltermann and Gorelick, 1996). An alternative to it is the use of Markov chains where transition probabilities between facies sequences are used to generate three-dimensional media with realistic internal architecture (Weissmann *et al.*, 1999). These transition probabilities can be estimated from boreholes and outcrops. Other methods are Boolean methods where objects (such as sand lenses) are randomly placed onto the domain to generate patterns with a given connectivity (Scheibe and Freyberg, 1995; Jussel *et al.*, 1994). The probability distribution of the spacings and sizes of the objects can again be estimated from outcrops. Subsurface media have also been characterized by the multifractal approach (Veneziano and Essiam, 2003), which is able to represent high conductivity features.

(d) Depositional simulation: While the above methods do not explicitly represent the physics that has lead to the subsurface deposits, there are also methods that attempt to mimic the depositional processes (Koltermann and Gorelick, 1996). The model of Tetzlaff and Harbaugh (1989), for example, simulates alluvial fans and deltas using the physical equations of water and sediment movement. This type of model is driven by paleoclimate which, however, may be very difficult to specify in a realistic way. Anderson (1997) also noted that there are difficulties with calibrating this type of model.

3. *Fractured rocks*

(a) Discrete-fracture approach: The most obvious method of generating fractured rock media is to explicitly consider flow through discrete fractures. Fractures provide conduits for the movement of groundwater and contaminants through an otherwise relatively impermeable rock mass. Each fracture in the network is specified by its location, shape, orientation, and hydraulic characteristics. The fractures are often represented by disks (Long *et al.*, 1985). To reduce computation time in solving the groundwater flow and transport equations, the three-dimensional network of disks has been replaced by a network of pipes (Cacas *et al.*, 1990). Of particular interest is the distribution of the connectivity of the fractures for which fractal concepts have been applied (Acuna and Yortsos, 1995; Renshaw, 1999). The media can be used directly to drive simulations models.

(b) Continuum approach: An alternative is the continuum approach where the fracture network is represented as if it were a granular porous medium. Often, multiple interacting continua are used, each of them representing fractures of a certain size class and one of them representing the solid rock mass (Bai *et al.*, 1993). The effective (upscaled) properties can be derived from the characteristics of the fractures (e.g. Bogdanov *et al.*, 2003) or can be found by calibrating a groundwater model.

CONCLUDING REMARKS

Even though the physical mechanisms of the various processes discussed here are clearly different, there exists a remarkable similarity in the statistical upscaling and downscaling methods. Most methods have traditionally started from assuming Gaussian random variability because of mathematical tractability. Here the most important scale effect is the reduction of variance with increasing support scale. Recognizing that hydrological variability can be more complex, the methods have subsequently evolved into more sophisticated ways of representing variability. Fractal representations are appealing because of their parsimony and because of the obvious presence of natural variability at all scales, but there is some controversy about the physical processes causing such type of variability. Multifractals are very attractive and appear to still merit further development in the different branches of hydrology because of their ability to represent singularities – spikes controlling the upscaling and downscaling characteristics – and again the parsimony and hence robustness of parameter estimation. There is also growing awareness that different parts of the subsurface (facies) and different situations (weather types, flood types) operate differently and there are merits of dealing with them separately.

There also seems to exist some convergence of methods in different application areas. In rainfall upscaling and downscaling, for example, the three historic strands – catchment rainfall estimation, temporal disaggregation, and climate model downscaling – seem to converge through the use of stochastic space-time rainfall models, although some of them are still at an early research stage. Similarly, geostatistical models that allow physically realistic structure to be represented – be it in soils or aquifers – seem to converge in the various application areas. Focusing on upscaling and downscaling methods may help cross-fertilize ideas within the hydrological subdisciplines (Blöschl, 2001).

Finally, as noted in the introduction, the focus of this article has been on how statistical upscaling and downscaling methods represent random variability in space and time at various scales. An alternative view is the dynamic upscaling and downscaling approach where the interest resides in how the model equations and model parameters change with scale, which is dealt with in **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**; **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**, and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**.

Acknowledgments

I would like to thank the numerous individuals who have contributed to the organization of the ideas in this article through discussions over the years, including Murugesu Sivapalan, Rodger Grayson, Andrew Western, Ralf Merz, Jon Skøien, Dieter Gutknecht, Erwin Zehe, and Lars Gottschalk. I am also grateful to Marc Bierkens for providing very helpful comments on the manuscript. Funding from the Austrian Academy of Sciences Project HÖ 18 is gratefully acknowledged.

REFERENCES

Acuna J.A. and Yortsos Y.C. (1995) Application of fractal geometry to the study of networks of fractures and their pressure transient. *Water Resources Research*, **31**(3), 527–540.

Anderson M.P. (1991) Comment on "Universal Scaling of Hydraulic conductivities and dispersivities in geologic media" by S.P. Neuman. *Water Resources Research*, **27**, 1381–1382.

Anderson M.P. (1997) Characterization of geological heterogeneity. In *Subsurface Flow and Transport: A Stochastic Approach*, *International Hydrology Series*, Dagan G. and Neuman S.P. (Eds.), University Press: Cambridge, pp. 23–43.

Bacchi B. and Ranzi R. (1996) On the derivation of the areal reduction factor of storms. *Atmospheric Research*, **42**(1–4), 123–135.

Bai M., Elsworth D. and Roegiers J.-C. (1993) Multiporosity/multipermeability approach to the simulation of naturally fractured reservoirs. *Water Resources Research*, **29**(6), 1621–1634, 10.1029/92WR02746.

Barling R.D., Moore I.D. and Grayson R.B. (1994) A quasi-dynamic wetness index for characterizing the spatial distribution of zones of surface saturation and soil water content. *Water Resources Research*, **30**(4), 1029–1044.

Beckie R. (1996) Sampling scale, network sampling scale, and groundwater model parameters. *Water Resources Research*, **32**(1), 65–76.

Beven K. (1995) Linking parameters across scales: subgrid parameterizations and scale dependent hydrological models. *Hydrological Processes*, **9**, 507–525.

Beven K.J. and Kirkby N.J. (1979) A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Bierkens M.F.P., Finke P.A. and de Willigen P. (2000) *Upscaling and Downscaling Methods for Environmental Research*, Kluwer Academic Press: Dordrecht, p. 190.

Bierkens M.F.P. and Weerts H.J.T. (1994) Block hydraulic conductivity of cross-bedded fluvial sediments. *Water Resources Research*, **30**, 2665–2678.

Blöschl G. (1999) Scaling issues in snow hydrology. *Hydrological Processes*, **13**, 2149–2175.

Blöschl G. (2001) Scaling in hydrology. *Hydrological Processes*, **15**, 709–711.

Blöschl G. and Grayson R. (2000) Spatial observations and interpolation. In *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson R. and Blöschl G. (Eds.), Cambridge University Press: Cambridge, Chap. 2, pp. 17–50.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling - a review. *Hydrological Processes*, **9**, 251–290.

Blöschl G. and Sivapalan M. (1997) Process controls on regional flood frequency: coefficient of variation and basin scale. *Water Resources Research*, **33**(12), 2967–2980.

Blöschl G., Sivapalan M., Gupta V.K. and Beven K. (Guest Editors) (1997) Scale problems in hydrology. *Water Resources Research*, **33**(12), (special issue) 2881–2999.

Bogdanov I.I., Mourzenko V.V., Thovert J.-F. and Adler P.M. (2003) Effective permeability of fractured porous media in steady state flow. *Water Resources Research*, **39**(1), 1023, doi:10.1029/2001WR000756.

Booij M.J. (2002) Extreme daily precipitation in Western Europe with climate change at appropriate spatial scales. *International Journal of Climatology*, **22**, 69–85, DOI: 10.1002/joc.715.

Brandsma T. and Buishand T.A. (1997) Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature. *Journal of Hydrology*, **198**, 98–123.

Cacas M.C., Ledoux E., de Marsily G. and Tillie B. (1990) Modeling fracture flow with a stochastic discrete fracture network: calibration and validation. 1: the flow model. *Water Resources Research*, **26**, 479–489.

Canterford R.P., Pescod N.R., Pearce H.J. and Turner L.H. (1987) Design intensity-frequency-duration rainfall. In *Australian Rainfall and Runoff*, Pilgrim D.H. (Ed.), The Institution of Engineers, Barton, ACT: Australia, pp. 15–40.

Charles S.P., Bates B.C., Smith I.N. and Hughes J.P. (2004) Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, **18**(8), 1373–1394.

Conway D. and Jones P.D. (1998) The use of weather types and air flow indices for GCM downscaling. *Journal of Hydrology*, **212/213**, 348–361.

Cunnane C. (1988) Methods and merits of regional flood frequency analysis. *Journal of Hydrology*, **100**, 269–290.

Cushman J.H. (1984) On unifying the concepts of scale, instrumentation, and stochastics in the development of multiphase transport theory. *Water Resources Research*, **20**(11), 1668–1676.

Cushman J.H. (1987) More on stochastic models. *Water Resources Research*, **23**, 750–752.

Dagan G. (1989) *Flow and Transport in Porous Formations*, Springer-Verlag: Heidelberg Berlin New York, p. 465.

Dalrymple T. (1960) *Flood Frequency Analysis*, Water Supply Paper 1543_a, US Geol. Survey: Reston.

De Michele C., Kottegoda N.T. and Rosso R. (2001) The derivation of areal reduction factor of storm rainfall from its scaling properties. *Water Resources Research*, **37**(12), 3247–3252.

De Michele C. and Rosso R. (2002) A multi-level approach to flood frequency regionalisation. *Hydrology and Earth System Sciences*, **6**, 185–194.

Desbarats A.J. (1987) Numerical estimation of effective permeability in sand-shale formations. *Water Resources Research*, **23**, 273–286.

Deutsch C.V. and Journel A.G. (1992) *GSLIB Geostatistical Software Library and User's Guide*, Oxford University Press: New York, p. 340.

Di Federico V. and Neuman S.P. (1997) Scaling of random fields by means of truncated power variograms and associated spectra. *Water Resources Research*, **33**(5), 1075–1085.

Eagleson P.S. (1978) Climate, soil, and vegetation: 2. The distribution of annual precipitation derived from observed storm sequences. *Water Resources Research*, **14**, 713–721.

Eaton B.C., Church M. and Ham D. (2002) Scaling and regionalization of flood flows in British Columbia, Canada. *Hydrological Processes*, **16**(16), 3245–3263.

Entin J.K., Robock A., Vinnikov K.Y., Hollinger S.E., Liu S. and Namkhai A. (2000) Temporal and spatial scales of observed soil moisture variations in the extratropics. *Journal of Geophysical Research*, **105**(D9), 11865–11877.

Farmer C.L. (2002) Upscaling: a review. *International Journal for Numerical Methods in Fluids*, **40**, 63–78.

Favre A.-C., Musy A.- and Morgenthaler S. (2004) Unbiased parameter estimation of the Neyman-Scott model for rainfall simulation with related confidence interval. *Journal of Hydrology*, **286**(1–4), 168–178.

Findikakis A.N. (Ed.) (2002) *Proceedings of the Groundwater Symposium 2002*. International Association of Hydraulic Research, Madrid 623 pp.

Foufoula-Georgiou E. and Georgakakos K.P. (1991) Recent advances in space-time precipitation modeling and forecasting. In *Recent Advances in the Modeling of Hydrologic Systems*, Bowles D.S. and O'Connell P.E. (Eds.), Kluwer: Dordrecht, pp. 47–65.

Foufoula-Georgiou E. and Krajewski W.F. (1995) *Recent Advances in Rainfall Modeling, Estimation, and Forecasting, Reviews of Geophysics*, U.S. National Report to International Union of Geodesy and Geophysics: pp. 1991–1994, pp. 1125–1137.

Fowler H.J., Kilsby C.G., O'Connell P.E. and Burton A. (2005) A weather-type conditioned multi-site stochastic rainfall model for the generation of scenarios of climatic variability and change. *Journal of Hydrology*. in press.

Gallant J.C., Moore I.D., Hutchinson M.F. and Gessler P.E. (1994) Estimating fractal dimension of profiles: a comparison of methods. *Mathematical Geology*, **26**, 455–481.

Gelhar L.W. (1993) *Stochastic Subsurface Hydrology*, Prentice-Hall: Englewood Cliffs, p. 390.

Gelhar L.W. and Axnes C.L. (1983) Three-dimensional stochastic analysis of macrodispersion in aquifers. *Water Resources Research*, **19**, 161–180.

Gómez-Hernández J.J. and Wen X.-H. (1997) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. *Advances in Water Resources*, **21**, 47–61.

Grace R.A. and Eagleson P.S. (1966) *The Synthesis of Short-time-increment Rainfall Sequences*, Report No. 91, Hydrodyn. Lab., Mass. Inst. of Technol., Cambridge.

Gray T.A., Chork C.Y. and Taggart I.J. (1993) *Pitfalls in the Fractal Analysis of Reservoir Property Data*, Soc. Pet. Eng.: SPE paper 26421, pp. 45–60.

Grayson, R.B. and Blöschl G. (Eds) (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*. Cambridge University Press: Cambridge, p. 404.

Grayson R., Blöschl G., Western A. and McMahon T. (2002) Advances in the use of observed spatial patterns of catchment hydrological response. *Advances in Water Resources*, **25**, 1313–1334.

Grayson R.B. and Western A.W. (1998) Towards areal estimation of soil water content from point measurements: time and space stability of mean response. *Journal of Hydrology*, **207**, 68–82.

Grebner D. (1995) *Klimatologie und Regionalisierung Starker Gebietsniederschläge in der nordalpinen Schweiz*, Zürcher Geographische Schriften 59, ETH Zürich.

Green T.R. and Erskine R.H. (2004) Measurement, scaling, and topographic analyses of spatial crop yield and soil water content. *Hydrological Processes*, **18**(8), 1447–1465.

Gumbel E.J. (1958) *Statistics of Extremes*, Columbia Univ. Press: New York.

Güntner A., Olsson J., Calver A. and Gannon B. (2001) Cascade-based disaggregation of continuous rainfall time series: the influence of climate. *Hydrology and Earth System Sciences*, **5**, 145–164.

Gupta V.K. (2004) Emergence of statistical scaling in floods on channel networks from complex runoff dynamics. *Chaos, Solitons and Fractals*, **19**, 357–365.

Gupta V.K. and Dawdy D.R. (1995) Physical interpretations of regional variations in the scaling exponents of flood quantiles. *Hydrological Processes*, **9**, 347–361.

Gupta V.K., Mesa O.J. and Dawdy D.R. (1994) Multiscaling theory of flood peaks: regional quantile analysis. *Water Resources Research*, **30**, 3405–3421.

Gupta V.K., Rodríguez-Iturbe I. and Wood E.F. (Eds.) (1986) *Scale Problems in Hydrology*. D. Reidel Publ., Dordrecht, p. 246.

Gupta V.K. and Waymire E. (1998) Spatial variability and scale invariance in hydrologic regionalization. In *Scale Dependence and Scale Invariance in Hydrology*, Edited by Sposito G. (Ed.), Cambridge University Press, London, pp. 88–135.

Hay L.E., McCabe G.J., Wolock D.M. and Ayers M.A. (1991) Simulation of precipitation by weather type analysis. *Water Resources Research*, **27**, 493–501.

Hewett T.A. (1986) *Fractal Distributions of Reservoir Heterogeneity and their Influence on Fluid Transport*, SPE paper 15386, Soc. Pet. Eng..

Hewett T.A. and Deutsch C.V. (1996) Challenges in Reservoir Forecasting. *Mathematical Geology*, **28**(7), 829–842.

Hewitson B.C. and Crane R.G. (1996) Climate downscaling: Techniques and application. *Climate Research*, **7**, 85–95.

Hu Z., Islam S. and Cheng Y. (1997) Statistical characterisation of remotely sensed soil moisture images. *Remote Sensing of the Environment*, **61**, 310–318.

Huff F.A. (1967) Time distribution of rainfall in heavy storms. *Water Resources Research*, **3**(4), 1007–1019.

Institute of Hydrology (IH) (1999) *Flood Estimation Handbook*. Institute of Hydrology, Wallingford.

IPCC (2001) *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton, J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge and New York, p. 881.

Jothityangkoon C., Sivapalan M. and Viney N.R. (2000) Tests of a space-time model of daily rainfall in southwestern Australia based on nonhomogeneous random cascades. *Water Resources Research*, **36**(1), 267–284.

Journel A.G. and Alabert F.G. (1989) Non-gaussian data expansion in the earth sciences. *Terra Nova*, **1**, 123–134.

Journel A.G. and Huijbregts C.J. (1978) *Mining Geostatistics*, Academic Press: London, p. 600.

Jussel P., Stauffer F. and Dracos T. (1994) Transport modeling in heterogeneous aquifers: 2. Three-dimensional transport model and stochastic numerical tracer experiments. *Water Resources Research*, **30**(6), 1819–1832, 10.1029/94WR00163.

Kalma J.D. and Sivapalan M. (Eds.) (1995) *Scale Issues in Hydrological Modelling. Advances in Hydrological Processes*, John Wiley: Chichester, p. 489.

Koltermann C.E. and Gorelick S.M. (1996) Heterogeneity in sedimentary deposits: a review of structure-imitating, process-imitating and descriptive approaches. *Water Resources Research*, **32**, 2617–2658.

Koutsoyiannis D. and Foufoula-Georgiou E. (1993) A scaling model of storm hyetograph. *Water Resources Research*, **29**(7), 2345–2361.

Koutsoyiannis D. and Mamassis N. (2001) On the representation of hyetograph characteristics by stochastic rainfall models. *Journal of Hydrology*, **251**(1–2), 65–87.

Koutsoyiannis D., Onof C. and Wheater H.S. (2003) Multivariate rainfall disaggregation at a fine timescale. *Water Resources Research*, **39**(7), 1173, doi:10.1029/2002WR001600.

Long J.C.S., Gilmour P. and Witherspoon P.A. (1985) A model for steady state flow in random, three- dimensional networks of disk-shaped fractures. *Water Resources Research*, **21**(8), 1150–1115.

Lorenz E.N. (1969) Atmospheric predictability as revealed by naturally occurring analogs. *Journal of the Atmospheric Sciences*, **26**, 639–646.

Lovejoy S. and Schertzer D. (1991) Multifractal analysis techniques and rain and cloud fields from 10^{-3} to 10^{6} m. In *Scaling, Fractals and Non-linear Variability in Geophysics*, Schertzer D. and Lovejoy S. (Eds.), Kluwer: Dordrecht, pp. 111–144.

Menabde M. and Sivapalan M. (2000) Modeling of rainfall time series and extremes, using bounded random cascades and Levy-stable distributions. *Water Resources Research*, **36**(11), 3293–3300.

Menabde M., Harris D., Seed A., Austin G. and Stow D. (1997) Multiscaling properties of rainfall and bounded random cascades. *Water Resources Research*, **33**(12), 2823–2830.

Merz R. and Blöschl G. (2003) A process typology of regional floods. *Water Resources Research*, **39**(12), 1340, doi:10.1029/2002WR001952.

Merz R. and Blöschl G. (2005) Flood frequency regionalisation – spatial proximity vs. catchment attributes. *Journal of Hydrology*, **302**(1–4), 283–306.

Mohanty B.P., Skaggs T.H. and Famiglietti J.S. (2000) Analysis and mapping of field-scale soil moisture variability using high-resolution, ground-based data during the Southern Great Plains 1997 (SGP97) Hydrology Experiment. *Water Resources Research*, **36**(4), 1023–1031.

Montaldo N. and Albertson J.D. (2003) Temporal dynamics of soil moisture variability: 2. Implication for land surface models. *Water Resources Research*, **39**(10), Art. No. 1275, doi:10.1029/2002WR001618, 2003.

Moore I.D., Grayson R.B. and Ladson A.R. (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5**, 3–30.

Neuman S.P. (1990) Universal Scaling of Hydraulic conductivities and dispersivities in geologic media. *Water Resources Research*, **26**, 1749–1758.

Neuman S.P. (1993) Comment on "A critical review of data on field-scale dispersion in aquifers" by L. W. Gelhar, C. Welty, and K. R. Rehfeldt. *Water Resources Research*, **29**, 1863–1865.

Nyberg L. (1996) Spatial variability of soil water content in the covered catchment at Gårdsjön, Sweden. *Hydrological Processes*, **10**, 89–103.

O'Loughlin E.M. (1986) Prediction of surface saturation zones in natural catchments by topographic analysis. *Water Resources Research*, **22**(5), 794–804.

Olsson J. and Berndtsson R. (1998) Temporal rainfall disaggregation based on scaling properties. *Water Science and Technology*, **37**(11), 73–79.

Onof C. and Wheater H.S. (1993) Modelling of British rainfall using a random parameter Bartlett-Lewis rectangular pulse model. *Journal of Hydrology*, **149**, 67–95.

Pachepsky Y.A. Radcliffe D.E. and Selim H.M. (Eds.) (2003) *Scaling Methods in Soil Physics*. CRC Press, Boca Raton, p. 434.

Pilgrim D.H. (1983) Some problems in transferring hydrological relationships between small and large drainage basins and between regions. *Journal of Hydrology*, **65**, 49–72.

Pitman A.J. (2003) The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology*, **23**, 479–510, doi: 10.1002/joc.893.

Renshaw C.E. (1999) Connectivity of joint networks with power law length distributions. *Water Resources Research*, **35**(9), 2661–2670.

Ritzi R.W. Jr, Dominic D.F., Slesers A.J., Greer C.B., Reboulet E.C., Telford J.A., Masters R.W., Klohe C.A., Bogle J.L. and Means B.P. (2000) Comparing statistical models of physical heterogeneity in buried-valley aquifers. *Water Resources Research*, **36**(11), 3179–3192.

Robinson J.S. and Sivapalan M. (1997) An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research*, **33**(5), 1045–1060.

Rodríguez-Iturbe I. and Gupta V.K. (1983) Scale problems in hydrology. *Journal of Hydrology*, **65**,(special issue) 159–174.

Ronda R.J., van den Hurk B.J.J.M. and Holtslag A.A.M. (2002) Spatial heterogeneity of the soil moisture content and its impact on surface flux densities and near-surface meteorology. *Journal of Hydrometeorology*, **3**(5), 556–570.

Robinson J.S. and Sivapalan M. (1997) An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research*, **33**(5), 1045–1060.

Rubin Y. and Bellin A. (1998) Conditional simulation of geologic media with evolving scales of heterogeneity. In: *Scale Dependence and Scale Invariance in Hydrology*, Sposito G. (Ed.), Cambridge University Press, London pp. 398–420.

Salas J.D. (1993) Analysis and modeling of hydrologic time series. In *Handbook of Hydrology*, Maidment D. (Ed.), McGraw-Hill: New York, chap. 19.

Scheibe T.D. and Freyberg D.L. (1995) Use of sedimentological information for geometric simulation of natural porous media structure. *Water Resources Research*, **31**(12), 3259–3270.

Schmid H.P. (2002) Footprint modeling for vegetation atmosphere exchange studies: a review and perspective. *Agricultural and Forest Meteorology*, **113**, 159–183.

Seed A.W., Srikanthan R. and Menabde M. (1999) A space and time model for design storm rainfall. *Journal of Geophysical Research*, **104**(D24), 31623–31630.

Seyfried M. (1998) Spatial variability constraints to modeling soil water at different scales. *Geoderma*, **85**(2–3), 231–254.

Shuttleworth W.J., Yang Z.-L. and Arain M.A. (1997) Aggregation rules for surface parameters in global models. *Hydrology and Earth System Sciences*, **1**, 217–226.

Sivapalan M. and Blöschl G. (1998) Transformation of point rainfall to areal rainfall: intensity-duration-frequency curves. *Journal of Hydrology*, **204**, 150–167.

Sivapalan M., Blöschl G., Merz R. and Gutknecht D. (2005) Linking flood frequency to long-term water balance: incorporating effects of seasonality, *Water Resources Research*, **41**, in press.

Sivapalan M., Blöschl G., Zhang L. and Vertessy R. (2003) Downward approach to hydrological prediction. *Hydrological Processes*, **17**, 2101–2111.

Skaugen T., Creutin J.D. and Gottschalk L. (1996) Reconstruction and frequency estimates of extreme daily areal precipitation. *Journal of Geophysical Research*, **101**(D21), 26287–26295.

Skøien J.O. and Blöschl G. (2005) Sampling scale effects in random fields and implications for environmental monitoring. *Environmental Monitoring and Assessment*, **105**, in press.

Skøien J.O., Blöschl G. and Western A.W. (2003) Characteristic space scales and timescales in hydrology. *Water Resources Research*, **39**(10), 1304, 10.1029/2002WR001736.

Smith J.A. (1992) Representation of basin scale in flood peak distributions. *Water Resources Research*, **28**, 2993–2999.

Smith J.A. and Karr A.F. (1985) Statistical inference for point process models of rainfall. *Water Resources Research*, **21**(1), 73–79.

Sposito G. (1998) *Scale Dependence and Scale Invariance in Hydrology*, Cambridge University Press: p. 423.

Sreenivasan K.R. (1991) Fractals and multifractals in fluid turbulence. *Annual Reviews Fluid Mechanics*, **23**, 539–600.

Srikanthan R. (1995) *A Review of the Methods for Estimating Areal Reduction Factors for Design Rainfalls*, Report 95/3, Cooperative Research Centre for Catchment Hydrology, Melbourne, p. 36.

Srikanthan R. and McMahon T.A. (2001) Stochastic generation of annual, monthly and daily climate data: a review. *Hydrology and Earth System Sciences*, **5**, 653–670.

Stehlik J. and Bardossy A. (2002) Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation. *Journal of Hydrology*, **256**(1–2), 120–141.

Stewart J.B., Engman E.T., Feddes R.A. and Kerr Y. (1996) *Scaling up in Hydrology using Remote Sensing*, John Wiley: Chichester, p. 255.

Tetzlaff D.M. and Harbaugh J.W. (1989) *Simulating Clastic Sedimentation*, Van Nostrand Reinhold: New York, p. 202.

Thompson S.K. (2002) *Sampling. Wiley Series in Probability and Statistics*, Wiley: New York, p. 367.

Thyer M. and Kuczera G. (2003) A hidden Markov model for modelling long-term persistence in multi-site rainfall time series. 2. Real data analysis. *Journal of Hydrology*, **275**(1–2), 27–48.

Tveito O.E. and Ustrnul Z. (2003) *A Review of the Use of Large-scale Atmospheric Circulation Classification in Spatial Climatology*, Norwegian Meteorological Institute Report 10/03, Norwegian Meteorological Institute, Oslo.

Valencia D. and Schaake J.C. (1973) Disaggregation processes in stochastic hydrology. *Water Resources Research*, **9**(3), 211–219.

Veneziano D. and Essiam A.K. (2003) Flow through porous media with multifractal hydraulic conductivity. *Water Resources Research*, **39**(6), 1166, doi:10.1029/2001WR001018.

Venugopal V., Foufoula-Georgiou E. and Sapozhnikov V. (1999) A space-time downscaling model for rainfall. *Journal of Geophysical Research*, **104**(D16), 19705–19721.

Viney N.R. and Sivapalan M. (2004) A framework for scaling of hydrologic conceptualizations based on a disaggregation-aggregation approach. *Hydrological Processes*, **18**(8), 1395–1408.

von Storch H. and Zwiers F.W. (1999) *Statistical Analysis in Climate Research*, Cambridge University Press. p. 494.

Wang J.S.Y. (1991) Flow and transport in fractured rocks. *Reviews of Geophysics*, **29**(S), 254–262.

Waymire E. and Gupta V.K. (1981) The mathematical structure of rainfall representations, 3, Some applications of the point process theory to rainfall processes. *Water Resources Research*, **17**, 1287–1294.

Weissmann G.S., Carle S.F. and Fogg G.E. (1999) Three-dimensional hydrofacies modeling based on soil surveys and transition probability geostatistics. *Water Resources Research*, **35**(6), 1761–1770.

Western A.W. and Blöschl G. (1999) On the spatial scaling of soil moisture. *Journal of Hydrology*, **217**, 203–224.

Western A.W., Blöschl G. and Grayson R.B. (1998) How well do indicator variograms capture the spatial connectivity of soil moisture? *Hydrological Processes*, **12**, 1851–1868.

Western A.W., Blöschl G. and Grayson R.B. (2001) Towards capturing hydrologically significant connectivity in spatial patterns. *Water Resources Research*, **37**(1), 83–97.

Western A., Grayson R. and Blöschl G. (2002) Scaling of soil moisture: a hydrologic perspective. *Annual Review of Earth and Planetary Sciences*, **30**, 149–180.

Western A.W., Grayson R.B., Blöschl G., Willgoose G.R. and McMahon T.A. (1999) Observed spatial organisation of soil moisture and its relation to terrain indices. *Water Resources Research*, **35**(3), 797–810.

Western A.W., Grayson R.B., Blöschl G. and Wilson D.J. (2003) Spatial variability of soil moisture and its implications for scaling. In *Scaling Methods in Soil Physics*, Pachepsky Y., Radcliffe D.E. and Selim H.M. (Eds.), CRC Press: Boca Raton, pp. 119–142.

Western A.W., Zhou S.-L., Grayson R.B., McMahon T.A., Blöschl G. and Wilson D.J. (2004) Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes. *Journal of Hydrology*, **286**(1–4), 113–134.

Wilby R.L., Conway D. and Jones P.D. (2002) Prospects for downscaling seasonal precipitation variability using conditioned weather generator parameters. *Hydrological Processes*, **16**, 1215–1234.

Wilby R.L. and Wigley T.M.L. (2000) Precipitation predictors for downscaling: observed and general circulation model relationships. *International Journal of Climatology*, **20**, 641–661.

Wilby R.L., Wigley T.M.L., Conway D., Jones P.D., Hewitson B.C., Main J. and Wilks D.S. (1998) Statistical downscaling of general circulation model output: a comparison of methods. *Water Resources Research*, **34**, 2995–3008.

Wilks D.S. and Wilby R.L. (1999) The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, **23**, 329–357.

Wilson J.P. and Gallant J.C. (Eds.) (2000) *Terrain Analysis: Principles and Applications*, John Wiley: New York.

Wood E.F., Lettenmaier D.P. and Zartarian V.G. (1992) A land-surface hydrology parameterization with subgrid variability for general circulation models. *Journal of Geophysical Research*, **97**(D3), 2717–2728.

Woods R.A., Sivapalan M. and Robinson J.S. (1997) Modeling the spatial variability of subsurface runoff using a topographic index. *Water Resources Research*, **33**, 1061–1073.

Woolhiser D.A. and Osborn H.B. (1985) A stochastic model of dimensionless thunderstorm rainfall. *Water Resources Research*, **21**(4), 511–522.

Yarnal B.A., Comrie C., Frakes B. and Brown D.P. (2001) Developments and prospects in synoptic climatology. *International Journal of Climatology*, **21**, 1923–1950.

Zehe E. and Blöschl G. (2004) Predictability of hydrologic response at the plot and catchment scales: role of initial conditions. *Water Resources Research*, **40**, W10202, doi:10.1029/2003WR002869.

Zhang D. (2002) *Stochastic Methods for Flow in Porous Media*, Academic Press: San Diego, p. 368.

Zhao R.-J. (1992) The Xinanjiang model applied in China. *Journal of Hydrology*, **135**, 371–381.

Zorita E. and von Storch H. (1999) The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of Climate*, **12**, 2474–2489.

10: Concepts of Hydrologic Modeling

DAN ROSBJERG[1] AND HENRIK MADSEN[2]

[1]*Environment & Resources DTU, Technical University Denmark, Kongens Lyngby, Denmark*
[2]*River & Flood Management, DHI Water & Environment, Hørsholm, Denmark*

Hydrological modeling is introduced as an indispensable tool for testing new hypotheses and obtaining a better understanding of hydrological processes and their interaction. The difficulties in developing generic model tools because of natural heterogeneity and multiscaling are emphasized, and the concept of appropriate modeling is introduced for the application of a parameter parsimonious model that ensures a realistic simulation or prediction including assessment of the uncertainty. In model development, the basic conceptualization of a model is governing the crucial choices of scale, dimension, discretization, and process delimitation. Hydrological models are classified according to their primary application area. In model calibration and validation, the use of split sampling techniques and automatic calibration procedures is underlined. The sources of model uncertainty are presented and different assessment techniques including first-order analysis, Monte Carlo-based methods, the GLUE framework, and Bayesian procedures are briefly introduced along with data assimilation techniques to enhance the predictive capabilities.

INTRODUCTION

Hydrological models for descriptive and predictive purposes have been developed over more than 100 years. Early examples are the rational method for peak flow prediction (Mulvaney, 1850), the diagram for determination of storage requirements (Rippl, 1883) and the unit hydrograph method for transformation of rainfall into runoff (Sherman, 1932). Analytical and simple numerical models were gradually developed, but the emergence of electronic computers changed the situation completely. The development and application of models are now moving ahead at a speed that was unthinkable 50 years ago.

Today, the progress in hydrological sciences is closely connected to modeling. Although experimental hydrology is extremely important, it is in combination with modeling that the real new insight is achieved. Modeling is a framework for testing new theories and hypotheses in order to improve our understanding of the hydrological processes and how the different processes interact.

A model is a simplified description of nature developed or adjusted for a specific goal. Models can be more or less general, but so far no model can be assumed universal. There are many reasons for that. So many processes are involved in the cycling of the water that only the most important ones can be taken into account in the modeling efforts. In a given application, the most important ones will depend on the purpose of the modeling. Here, we shall only emphasize a few basic facts that make hydrological modeling a never-ending story. The hydrological cycle covers an immense range of spatial and temporal scales that make a single model approach impossible. Moreover, at any scale there is a subset of smaller scales such that the heterogeneity of the land surface and the subsurface can never be completely described. Finally, it should be emphasized that many of the physical processes governing the hydrological cycle are far from being completely known.

Thus, we may consider hydrological modeling as an art, in which the artist has to decide on very crucial matters regarding scale, heterogeneity, temporal, and spatial discretization and so on. In some cases, a continuum approach may suffice, while in other situations a very detailed description of the heterogeneity may be needed. Again the purpose of the modeling indirectly gives the answer.

For process understanding, the descriptive abilities of hydrological models are very important. A good fit of a model to measured data, however, may be obtained

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

through an over-parameterization of the different processes involved. Thus, the concept of parameter parsimony in process description is paramount. This is maybe even more important when applying models for predictive purposes. Reliable predictions can only be achieved with models using a set of well-identified parameters that reflect the fundamental governing mechanisms.

We shall end these introductory remarks by introducing the concept of *appropriate modeling*. This means the development or selection of a model with a degree of sophistication that reflects the actual needs for modeling results. It means a parameter parsimonious model with a discretization in space and time that ensures a realistic simulation or prediction of the requested variables. Although distributed physically based models have great advantages, a highly distributed model does not necessarily always ensure the most reliable answer because of the many parameters involved. In some cases, a simple lumped description may do a better job. Conversely, if a detailed picture of state variables in space or time is requested, the distributed models will have their force. For one particular site the simulation may not be very precise but, despite that, the general variation in space and time may appear satisfactory.

MODEL DEVELOPMENT

Basically, any kind of modeling can be looked upon as an input to a system that transforms the input into an output. To exemplify, the well-known unit hydrograph model is a linear transformation of the effective rainfall into runoff.

The system is identical to the mathematical/numerical formulation of the hydrological processes included in the modeling efforts. The modeling system can have a firm physical basis built on theories and physical laws of the various hydrological processes, for example, the Boussinesq equation for groundwater flow or the St. Venant equations for channel flow. Alternatively, one can analyze the input and output data and build empirical models that describe the observed relations using, for example, deterministic artificial neural networks (ANN) or autoregressive-moving average (ARMA) linear stochastic models. Between these two opposite approaches, a broad spectrum of modeling systems can be formulated using different process conceptualizations that have a certain degree of physical content, but need observation data to tune or calibrate the model parameters. Examples of different modeling systems are shown in Figure 1.

Depending on the degree of physical considerations involved in the process description, it has been common to divide models into black box (empirical), grey box (conceptual), and white box (physically based) models. This, however, may not suffice today. A grey box model at one particular scale may tend to become a white box model at another scale and vice versa. A model must be appropriate for its particular use and not chosen because of a black, grey, or white label. Moreover, white box models are in reality an illusion. A purely physically based model does not exist. Simplifications of the physics and lack of complete knowledge about the processes are problems we always have to face.

In the development of a hydrological model, a number of crucial questions have to be addressed. As already emphasized above, the answers to these questions should be governed by the contemplated use of the model. First, which part of the hydrological cycle is incorporated in the model? Second, in what detail should the involved hydrological processes be described? And third, what will be the input or, in other words, the initial and boundary conditions forcing the model?

Hydrology is strongly related to a number of other geoscience disciplines (geology, meteorology, etc.), and there is a multitude of internal relations between the state variables of hydrological processes that can be either included or disregarded by the modeler. We shall mention just a few of the most important aspects to consider in the choice of process description, or, in other words, in the parameterization or conceptualization of the model.

There are fundamental decisions to be taken regarding the temporal as well as the spatial scale in a model. In the temporal regime, the model can be event based (as the unit hydrograph model) or it can be developed for continuous time application. Should the model describe a steady-state situation, or will the variability in time be important? If the time-step chosen is less than a year, the annual cycle must be addressed. The answers to these questions will point to very different conceptualizations.

In the spatial domain, the options are several. The most simple is the lumped description with only one unit to consider. This can be extended gradually from combinations of lumped sub-models to models with 1D, 2D, or 3D grid cells of any size. Some components may have a 1D spacing, while other elements may have 2D or 3D grids. The multitude of choices is obvious already at this stage, as any of the temporal solutions can be combined with any spatial distribution.

Further, complexity will be added depending on which processes are taken into account in the modeling. Should only surface or subsurface processes be included, or should one consider an integrated approach where the different processes are dynamically linked? Is it only the liquid water that is our concern, or are the other phases in the form of snow and ice and water vapor also important to describe? Moreover, do we need to consider the constituents transported by the water, or even to expand basic hydrology into, for example, ecohydrology by including modeling of habitats and so on? For each process component that is included, different model descriptions can be chosen, ranging from complex descriptions that solve the governing

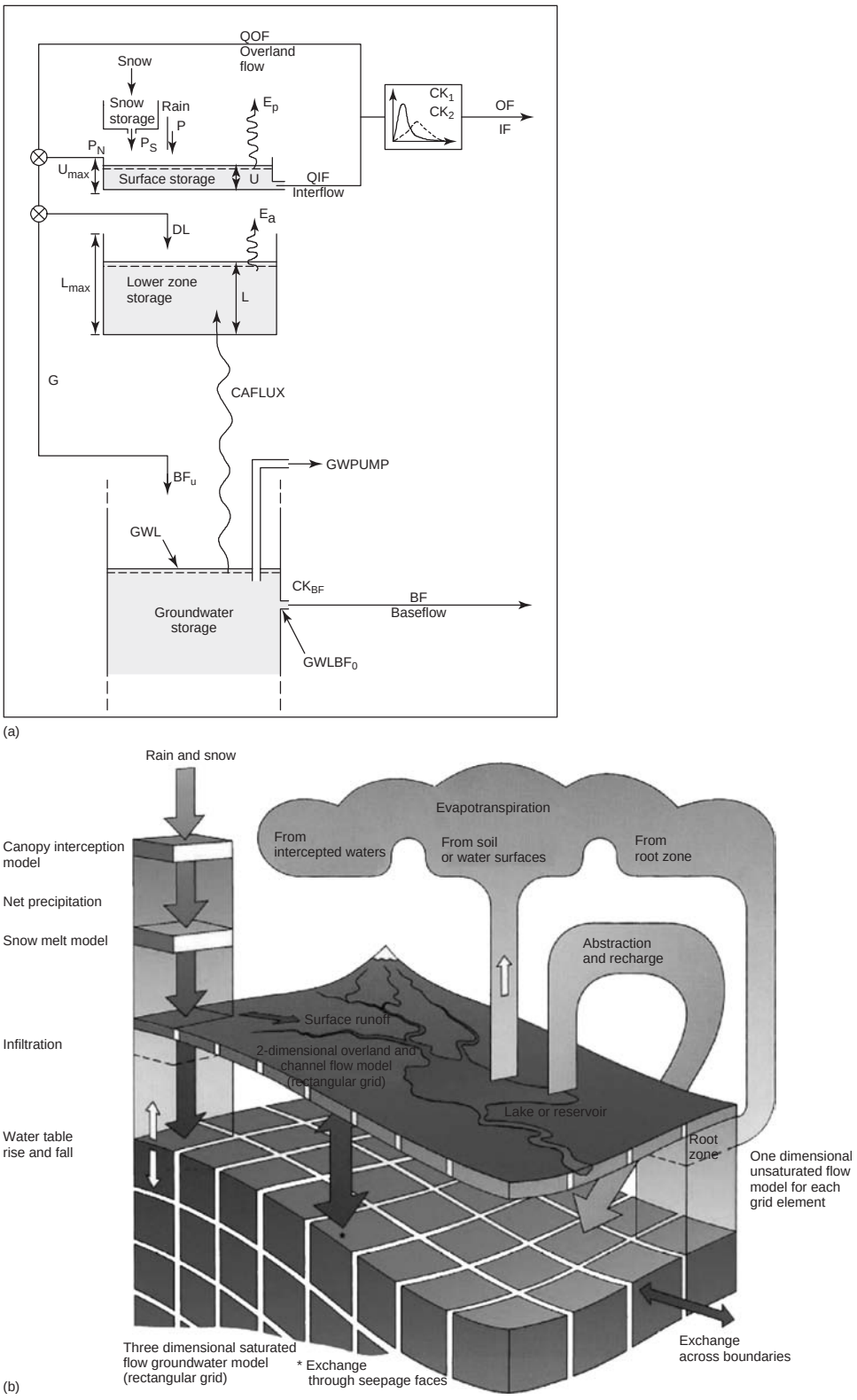


Figure 1 Examples of different hydrological modeling systems. (a) the MIKE 11/NAM lumped, conceptual rainfall-runoff model (Havnø *et al.*, 1995). (b) the MIKE SHE distributed and integrated modeling system (Refsgaard and Storm, 1995). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

partial differential equations to more simple, conceptual box-model descriptions.

In many cases, the forcing of a hydrological model consists of meteorological time series. In river or reservoir studies, it may suffice to use runoff time series as input, and in groundwater modeling the forcing is often dominated by constant head or flux boundaries. The initial values of the state variables are crucial in some model applications, but in other cases the first part of the simulation should be considered a warming-up period ensuring that the state variables gradually attain realistic values.

In the development of a hydrological model, it is very important to be aware of its intended future application. It is of crucial importance to know whether the model will be run with forcing similar to the data used in the development phase, or the model will be forced with input outside this range. In studies of climate change, impact scenarios are developed that deviate substantially from the present-day situation. This put much stronger requirements to the model conceptualization and parameterization. The validation procedure will be much more cumbersome as well, as illustrated in a later section.

MODEL CLASSIFICATION

Many different schemes for classification of hydrological models have been introduced during time. Often models have been classified according to contrasting properties like deterministic versus stochastic, lumped versus distributed, steady state versus time variable and so on; see, for example, the model monographs by Singh (1995) and Abbott and Refsgaard (1996). As shown above, these properties are extremely important to consider in the model development phase, but we prefer to classify the final products, the models, according to their primary application area. In the following, a set of model classes will be presented along with a brief description of the main features of each class.

River and Reservoir Models River models can be simple routing models like the Muskingum model, used in the HEC-1 Flood Hydrograph Package (Feldman, 1995), but also advanced hydrodynamic models that solve the full St. Venant equations, for example, the MIKE 11 model (Havnø *et al.*, 1995). They are well suited for calculation of flood wave propagation and the impact of structures in rivers. They can be expanded to interact with the flood plain and amended with modules for sediment transport and river morphology as well as advection–dispersion processes and water quality indicators. In recent years, wetland and habitat models are being developed, reflecting the growing interest in combining hydrological and ecological expertise. Reservoir models for design are simple continuity models, which, however, make use of advanced methods for generation of synthetic inflow series. River models in forecast mode are used on-line for managing flood waves and as a basis for issuing emergency warnings.

Rainfall-runoff Models, or in a wider sense *catchment* or *watershed models*, can be unit hydrograph models or simple lumped conceptual models like the Sacramento model (Burnash *et al.*, 1973; Burnash, 1995), the HBV model (Bergström and Forsman, 1973; Bergström, 1995) and the NAM model (Nielsen and Hansen, 1973; Havnø *et al.*, 1995). On the basis of meteorological data, the temporal variation of runoff at the outlet of a catchment can be calculated with a quite good precision, given a well-calibrated model. With a much higher degree of spatial resolution, the so-called *distributed models* have the advantage of being able to take into account the variation of several state variables at a number of locations distributed in the catchment. The data requirements, however, grow rapidly with increasing spatial resolution. In order to manage the data efficiently, the use of GIS-based methods is now becoming standard. Well-known examples of distributed catchment models are the TOPMODEL (Beven *et al.*, 1995; Beven, 1997), the MIKE SHE model (Refsgaard and Storm, 1995) and the SWAT model (Arnold *et al.*, 1998). These models take into account the complete land phase of the hydrological cycle, but use quite different approaches. While the TOPMODEL puts emphasis on a detailed topographical description, the MIKE SHE uses an integrated approach where emphasis can be put on different processes, including for instance a full 3D groundwater component, a 2D diffusive wave approximation for the overland flow, and a 1D full dynamic component for the flow in the river system. Depending on the actual use of the various models, these differences can become more or less pronounced. Catchment models can be combined with river models and expanded to include a number of transported constituents. Soil erosion processes can be included as well. As an example, SHETRAN (Bathurst, 2002) has been used for sediment transport and sediment yield modeling. Recently, some catchment models are being further developed for coupling with atmospheric models. This requires a joint description of both water and energy fluxes as done in the so-called *land-surface schemes*, which for long time has been used in general and regional circulation models. The increasing resolution of the atmospheric models now makes it possible to use an advanced hydrological model in the coupling.

Groundwater models are confined to the subsurface part of the hydrological cycle. There are several model systems available, and among them many are built on the well-known MODFLOW concept (e.g. Anderson and Woessner, 1992). When abstraction of groundwater is of primary interest, a steady-state groundwater model is usually applied. In the analysis of pollution plume movement in combination with different degrees of remediation action, a steady-state

groundwater model is usually supplemented with a transient advection–dispersion module, and in many cases also with modules for geochemistry and biochemical degradation. Groundwater models that take into account reactive transport are now becoming standard tools (Kinzelbach and Schäfer, 1994). Also, models with transient flow descriptions are now more and more frequently used, often in a combined modeling of the unsaturated and the saturated zone. In groundwater modeling, the main problem is the description of the heterogeneity in the soil and the aquifers. A complete knowledge can never be obtained, and the optimal use of the available pedological and geological information is therefore paramount. To this effect, geostatistically based methods are often included in the models (e.g. de Marsily, 1986).

Design and Management Models In the assessment of hydrological design values, frequency models have traditionally been used for return periods up to 500–1000 years. With the development of advanced statistical tools it has now become standard not just to calculate the T-year event estimate, but also to assess the uncertainty of the estimate (e.g. Cunnane, 1987; Stedinger *et al.*, 1993; Rosbjerg and Madsen, 1998). By including regional information, design at ungauged sites can be accomplished, and the precision can be improved at sites with short data records (e.g. Hosking and Wallis, 1997; Madsen and Rosbjerg, 1997a,b). Design for very rare events requires comprehensive modeling tools that combine meteorological and hydrological assessments. How to do this, and how to avoid both under- and over-design, are still unsolved questions in hydrology. In the management of the water resources, dedicated water allocation models in a GIS framework have shown their strength, especially when supplemented with water quality modules. The use of systems analysis has a long tradition in hydrology. The new generation of powerful computers has ensured a much wider application of these tools. Decision support and expert systems, in many cases combined with optimization routines, which a few years ago was impossible to apply within a reasonable time frame, are now coming into practice. Sustainable water resources management is a primary objective in policy development and in the implementation of water action plans (e.g. Loucks, 2001). This has brought rural and urban hydrology closer together and has led to hydrological management models that link up to, or even integrate socioeconomic models.

Emerging Models Although still valid, the above classification of hydrological models may become less obvious in the future. A new generation of hydrological models are being developed, where a specific model is being created for each specific purpose, that is, flexible and tailor-made models can be assembled without the huge efforts necessary so far. For the different hydrological processes to be included in the analysis, a number of different model options will be available in order to ensure an optimal combination given the actual application. Eventually, this may help in achieving the ultimate goal of appropriate modeling.

MODEL CALIBRATION AND VALIDATION

In the section on model development, the first steps leading to model application were delineated. First, the model frame must be conceptualized and parameterized, and second the involved hydrological processes must be given an appropriate mathematical/numerical formulation corresponding to the chosen temporal and spatial discretization. The outcome of these efforts is a model code that should be verified to the extent possible. In cases where analytical solutions exist, the complete code, or subsets of it, should be able to reproduce the analytical results with a reasonable accuracy. This code verification should not be mixed up with the model validation described below.

In a given model application, the actual value of the model parameters are generally unknown, or only known to be within a certain range of realistic values. Thus, there is a need for calibration of the model, that is, a procedure leading to adoption of an acceptable set of model parameters. Usually the available data are split into two subsets, such that one of the subsets can be used for calibration and the other for a subsequent validation. This highly recommendable split-sample technique should always be used, if the amount of data allows it. Following the split-sample validation, a final calibration using all available data should be accomplished.

In the calibration process, the parameters are varied such that certain chosen norms for correspondence between the simulated and measured outcome are fulfilled to the extent possible. This can be a simple visual judgment of the goodness of fit, but often some mathematical measures are used, for example, the correlation coefficient, the mean square error, or the Nash–Sutcliffe model efficiency (Nash and Sutcliffe, 1970). By changing the parameters iteratively, the calibration result is approached.

Until recently, the dominating calibration procedure was "trial and error". After each model run, the modeler changes the parameter values in order to get a better fit. For an experienced modeler, it is usually possible to obtain a good calibration result without too much effort, but for less experienced hydrologists it can be a very time-consuming method. Now automated procedures for determination of optimized parameter values are getting strong attention. Here the parameter set, or a subset of parameters that minimizes one or more mathematically formulated objective functions is selected as the calibration result.

While gradient search and other local search techniques often work quite well in groundwater modeling (e.g. Carrera, 1988), they usually fail when applied to catchment models. However, new global search techniques based on population-evolution algorithms have been developed and

shown to have a good performance in cases where the local search breaks down (e.g. Duan *et al.*, 1992; Vrugt *et al.*, 2003). Automated calibration has a number of advantages. First of all, automatic calibration allows a fast and objective calibration. In addition, it clearly shows if there is a problem with the identifiability of a model parameter, and it gives an assessment of the accuracy by which a parameter is estimated. The main disadvantage is related to the definition of an appropriate objective function that takes into account all the characteristics of the modeled system that are used by the hydrologist for evaluating the goodness of fit. The recent advancement of multiobjective calibration, where several objective functions can be optimized simultaneously (e.g. Gupta *et al.*, 1998; Madsen, 2000), allows the hydrologist to formulate specific objectives to be optimized depending on the model application being considered. State-of-the-art automatic calibration procedures perform favorably compared to expert manual calibrations (Gupta *et al.*, 1998; Madsen *et al.*, 2002; Madsen, 2003).

In order to test the generality of the hydrological model, and to assess if the estimated parameter values in fact are generic values, the calibrated model should be run with an independent set of data. If the correspondence between the simulated and measured outcome is acceptable in comparison to the result obtained during the calibration period, the model is said to be validated. In the strict sense, it can be questioned whether a model can be validated at all. What is achieved is only that based on the given level of information on which the model performs satisfactorily, and for that reason it should not be rejected. Thus, if more information becomes available, it may then lead to rejection of the model, or, more likely, to restriction of the range within which the model should be applied.

Depending on the anticipated use of a model, the validation criteria should be more or less demanding. Obviously, if a model is calibrated for humid conditions and is to be applied for arid conditions, it must be validated on such a data set. A comprehensive set of validation schemes has been developed by Klemes (1986) and applied in a systematic way by Refsgaard and Knudsen (1996).

MODEL UNCERTAINTIES

Uncertainty is an unavoidable and inherent element in hydrological modeling. The main uncertainty sources are related to (i) errors in the model forcing, (ii) use of an incomplete model structure, (iii) use of nonoptimal model parameters, and (iv) errors in the measurements used for the model calibration. It is now generally accepted that these uncertainties and their effect on the model predictions should be quantified as part of a hydrological modeling exercise. A reliable estimate of the prediction uncertainty is crucial for making efficient decisions on the basis of model simulations.

How to include and combine appropriately all the different error sources in hydrological modeling is a very complex and yet unsolved problem. Ideally, the joint probability distribution of all the important sources should be quantified and used as input to the model to derive probability distributions of the model predictions. However, quantification of the probability distributions for the different error sources is a difficult task, and hence simplifications are needed.

The classical statistical approach for evaluating model prediction uncertainty is a first-order analysis where the covariance of the model input (forcing and/or model parameters) are propagated through the model using a first-order Taylor series approximation of the model operator. For assessing parametric uncertainty, the covariance is usually estimated using a multinormal approximation of the probability density function of the model around the estimated optimum based on a gradient-based search.

While the first-order approach is an efficient solution to quasi-linear models, the method fails when applied to highly nonlinear models. In such cases, Monte Carlo-based procedures have been put forward. In these methods, the statistical properties of the model input are represented by an ensemble estimate, and this ensemble is then propagated through the model to produce an ensemble of model outputs from which confidence limits on the model predictions can be derived. The main shortcoming of the Monte Carlo approach is the slow convergence of the ensemble to the true probability distribution (proportional to the inverse of the ensemble size), and hence the method has huge computer processing (CPU) requirements. Alternative sampling strategies have been proposed such as Latin Hypercube Sampling, which allows a more efficient sampling in the model input space.

Another shortcoming of the Monte Carlo-based procedure is the explicit use of a joint probability distribution for the model input. Other sampling procedures have been proposed that are conditioned on the observations of the modeled system for evaluating the model prediction uncertainties. These methods include importance sampling such as the generalized likelihood uncertainty estimation (GLUE) procedure (Beven and Binley, 1992) and Markov Chain Monte Carlo sampling (e.g. Kuczera and Parent, 1998; Vrugt *et al.*, 2003).

The GLUE framework addresses the problem of nonuniqueness in model calibration. That is, many different parameter combinations are equally acceptable in reproducing the observed system behavior, and these parameter sets may often come from very different regions in the parameter space. Beven and Binley (1992) refer to this situation as model equifinality. In such cases, the multinormal approximation of the model parameters around an estimated optimum is inadequate to represent the parameter uncertainty.

Most of the model calibration and uncertainty propagation procedures implicitly assume a correct model structure, and only parameters within that structure are allowed to vary. When significant model structural errors are present, such procedures will provide biased model parameter estimates and unreliable uncertainty predictions. Model structural errors may be qualitatively addressed within a multiobjective calibration framework. The tradeoff between different objective functions of the Pareto optimal solutions will provide an indication of possible model errors (Gupta *et al.*, 1998). Quantification of these errors is much more complicated, since model structural errors are difficult to isolate from the errors originating from parameter uncertainty. Bayesian statistical inference methods have been proposed to consider jointly the different error sources in a simplified functional form (e.g. Kennedy and O'Hagan, 2001). For groundwater flow and transport modeling, Neuman (2003) proposed a Bayesian model averaging procedure to combine the predictions of several alternative conceptual hydrogeological models and to assess their joint probability. The GLUE method and Markov Chain Monte Carlo sampling offer another framework to address jointly different error sources, but so far have considered mainly parameter uncertainty. Butts *et al.* (2004) proposed a general modeling framework for considering different model structures to address the effect on model predictions and compared with the prediction uncertainties from other error sources. They showed that model structure uncertainty is at least as large as uncertainties caused by errors in model parameters, model forcing, and observations.

As an alternative to error propagation using the deterministic model equations, the governing differential equations can be treated as stochastic differential equations and error propagation can be analyzed by means of perturbation theory. This approach has been particularly addressed in groundwater flow and transport modeling. Under certain simplifying assumptions, analytical solutions for the prediction uncertainties can be developed (Dagan, 1989; Gelhar, 1993). For transient phenomena, Dagan (1989) has calculated approximate moments in the time domain, while (Gelhar, 1993) has focused on the asymptotic behavior using spectral theory.

Although discrepancies between model results and measured values may be minimized in the calibration process, the predictive capabilities of the hydrological model may be significantly improved by adaptively updating the model when new data become available. This is particularly important in real-time applications, where measurements are used for updating the model prior to the time of forecast and for correcting model predictions in the forecast period. Model updating or, in more general terms, data assimilation can be classified according to the variables that are modified in the data assimilation feedback process, that is, input variables (model forcing), model states, model parameters, and output variables (WMO, 1992). Updating of output variables, also known as *error correction*, is the most widely used procedure in operational hydrological forecasting (WMO, 1992; Refsgaard, 1997; Toth *et al.*, 1999). Recent methodological advances of data assimilation in meteorology and oceanography, and the advent of new data sources from remote sensing instruments has put more focus on the use of data assimilation in hydrological modeling (McLaughlin, 2002).

CONCLUDING REMARKS

Hydrological models are being developed for application at widely different temporal and spatial scales. A basic request is that the models efficiently address the dominant hydrological processes at the chosen scales such that sufficient descriptive and predictive abilities are preserved. This requires parameter parsimonious models and appropriate procedures for model calibration. Moreover the model must pass a validation test, if the results should be considered reliable. Finally, an up-to-date modeling effort necessitates an assessment of the uncertainty range for the output variables.

This can be summarized in the request for *appropriate modeling*, where the actual purpose of the modeling is governing the choice of scales, the sophistication level in process description and parameterization, the calibration and validation procedures, and the uncertainty assessment.

Model development will continue, and still more comprehensive models will emerge taking into account, for example, multiscale heterogeneities including methods for assessing the effects of sub-grid heterogeneity and water quality including chemical and biological processes. The predictive abilities of the models will become a keypoint with strong requests for reliable predictions even in a changed climate. This puts high demands on hydrological modelers and sets the scene for a challenging future development of hydrological models.

REFERENCES

Abbott M.B. and Refsgaard J.C. (Eds.) (1996) *Distributed Hydrological Modelling*, Kluwer Academic Publishers.

Anderson M.P. and Woessner W.W. (1992) *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*, Academic Press.

Arnold J.G., Srinivasan R., Muttiah R.S. and Williams J.R. (1998) Large area hydrologic modeling and assessment – Part 1: model development. *Journal of the American Water Resources Association*, **34**(1), 73–89.

Bathurst J.C. (2002) Physically-based erosion and sediment yield modelling: the SHETRAN concept. In *Modelling*

Erosion, Sediment Transport and Sediment Yield, Summer W. and Walling D.E. (Eds.), UNESCO-IHP Technical Documents in Hydrology No. 60, UNESCO-IHP, pp. 47–68.

Bergström S. (1995) The HBV model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, pp. 443–476.

Bergström S. and Forsman A. (1973) Development of a conceptual deterministic rainfall-runoff model. *Nordic Hydrology*, **4**(3), 147–170.

Beven K.J. (Ed.) (1997) *Distributed Hydrological Modelling: Applications of the TOPMODEL Concept*, John Wiley & Sons.

Beven K.J. and Binley A.M. (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**(3), 279–298.

Beven K.J., Lamb R., Quinn P., Romanowicz R. and Freer J. (1995) TOPMODEL. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, pp. 627–668.

Burnash R.J.C. (1995) The NWS river forecast system – catchment modelling. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, pp. 311–366.

Burnash R.J.C., Ferral R.L. and McGuire R.A. (1973) *A Generalized Streamflow Simulation System – Conceptual Modelling for Digital Computers*, US Department of Commerce, National Weather Service and State of California, Department of Water Resources.

Butts M.B., Payne J.T., Kristensen M. and Madsen H. (2004) An evaluation of the impact of model structure and complexity on hydrological modelling uncertainty for streamflow prediction. *Journal of Hydrology*, **298**, 242–266.

Carrera J. (1988) State of the art of the inverse problem applied to the flow and solute transport equations. *NATO ASI Series C*, Vol. 224, Reidel Publishing, pp. 549–583.

Cunnane C. (1987) Review of statistical models for flood frequency estimation. In *Hydrologic Frequency Analysis*, Singh V.P. (Ed.), Reidel, pp. 49–95.

Dagan G. (1989) *Flow and Transport in Porous Formations*, Springer: New York.

de Marsily G. (1986) *Quantitative Hydrogeology: Groundwater Hydrology for Engineers*, Academic Press.

Duan Q., Sorooshian S. and Gupta V. (1992) Effective and efficient global optimisation for conceptual rainfall-runoff models. *Water Resources Research*, **28**(4), 1015–1031.

Feldman A.D. (1995) HEC-1 flood hydrograph package. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, pp. 119–150.

Gelhar L.W. (1993) *Stochastic Subsurface Hydrology*, Prentice Hall.

Gupta H.V., Sorooshian S. and Yapo P.O. (1998) Toward improved calibration of hydrological models: multiple and noncommensurable measures of information. *Water Resources Research*, **34**(4), 751–763.

Havnø K., Madsen M.N. and Dørge J. (1995) MIKE 11 – a generalized river modelling package. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, pp. 733–782.

Hosking J.R.M. and Wallis J.R. (1997) *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge University Press.

Kennedy M.C. and O'Hagan A. (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series A*, **63**, 425–464.

Kinzelbach W. and Schäfer W. (1994) Reactive transport in heterogeneous media. In *Computational Methods in Water Resources X*, Vol. 1, Peters A., Wittum G., Herrling B., Meissner U., Brebbia C.B., Gray W.G. and Pinder G.F. (Eds.), Kluwer Academic Publishers, pp. 637–650.

Klemes V. (1986) Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, **31**(1), 13–24.

Kuczera G. and Parent E. (1998) Monte Carlo assessment and parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology*, **211**, 69–85.

Loucks D.P. (2001) Water resource systems modelling: its role in planning. *Encyclopedia of Life Support Systems*, Vol. II, pp. 1349–1360.

Madsen H. (2000) Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of Hydrology*, **235**, 276–288.

Madsen H. (2003) Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources*, **26**(2), 205–216.

Madsen H. and Rosbjerg D. (1997a) The partial duration series method in regional index-flood modeling. *Water Resources Research*, **33**(4), 737–746.

Madsen H. and Rosbjerg D. (1997b) Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. *Water Resources Research*, **33**(4), 771–781.

Madsen H., Wilson G. and Ammentorp H.C. (2002) Comparison of different automated strategies for calibration of rainfall-runoff models. *Journal of Hydrology*, **261**, 48–59.

McLaughlin D. (2002) An integrated approach to hydrologic data assimilation: interpolation, smoothing and filtering. *Advances in Water Resources*, **25**, 1275–1286.

Mulvaney T.J. (1850) On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and of flood discharges in a give catchment. *Proceedings of the Institution of Civil Engineers Ireland*, **4**, 18–31.

Nash J.E. and Sutcliffe J.V. (1970) Riverflow forecasting through conceptual models. Part 1: a discussion of principles. *Journal of Hydrology*, **10**, 282–290.

Neuman S. (2003) Accounting for conceptual model uncertainty via maximum likelihood averaging. In *Calibration and Reliability in Groundwater Modelling: A Few Steps Closer to Reality*, Kovar K. and Hrkal Z. (Eds.), IAHS Publication 277, IAHS, pp. 303–313.

Nielsen S.A. and Hansen E. (1973) Numerical simulation of the rainfall-runoff process on a daily basis. *Nordic Hydrology*, **4**(3), 171–190.

Refsgaard J.C. (1997) Validation and intercomparison of different updating procedures for real-time forecasting. *Nordic Hydrology*, **28**(2), 65–84.

Refsgaard J.C. and Knudsen J. (1996) Operational validation and intercomparison of different hydrological models. *Water Resources Research*, **32**(7), 2189–2202.

Refsgaard J.C. and Storm B. (1995) MIKE SHE. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, pp. 809–846.

Rippl W. (1883) The capacity of storage reservoirs for water supply. *Proceedings of the Institution of Civil Engineers*, **71**, 270–278.

Rosbjerg D. and Madsen H. (1998) Design with uncertain design values. In *Hydrology in a Changing Environment*, Wheater H. and Kirby C. (Eds.), John Wiley & Sons, pp. 155–163.

Sherman L.K. (1932) Streamflow from rainfall by the unit-graph method. *Engineering News-Record*, **108**, 501–505.

Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Water Resources Publications.

Stedinger J.R., Vogel R.M. and Foufoula-Georgiou E. (1993) Frequency analysis of extreme events. In *Handbook of Hydrology*, Chap. 18, Maidment D. (Ed.), McGraw-Hill: New York.

Toth E., Brath A. and Montanari A. (1999) Real-time flood forecasting via combined use of conceptual and stochastic models. *Physics and Chemistry of the Earth, Part B,***24**(7), 793–798.

Vrugt J.A., Gupta H.V., Bouten W. and Sorooshian S. (2003) A shuffled complex evolution metropolis algorithm for optimisation and uncertainty assessment of hydrological model parameters. *Water Resources Research*, **39**(8), 1201, doi:10.1029/2002WR001642.

WMO (1992) *Simulated Real-Time Intercomparison of Hydrological Models*, Operational Hydrology Report No. 38, World Meteorological Organisation, Geneva.

11: Upscaling and Downscaling – Dynamic Models

MARK WIGMOSTA AND RAJIV PRASAD

Environmental Technology Division, Pacific Northwest National Laboratory, Richland, WA, US

Transferring information from one spatial or temporal scale to another (scaling) presents one of the most daunting scientific challenges in hydrology. This is particularly true in hydrological modeling where disparities in scale exist in nearly every phase of model development and application. This article discusses scales, hydrological processes, heterogeneity, upscaling, and downscaling as applied to dynamic hydrological modeling. It is not meant to be an exhaustive discussion covering the entire range of scale issues in hydrology, or of hydrological modeling, rather it attempts to summarize the relevant issues and illustrate some approaches using specific examples to address upscaling and downscaling in hydrological model development and application.

INTRODUCTION

Transferring information from one spatial or temporal scale to another (scaling) presents one of the most daunting scientific challenges in hydrology. This is particularly true in hydrological modeling where disparities in scale exist in nearly every phase of model development and application. Models are typically developed for a specific temporal (subdaily, daily, storm event, monthly, etc.) and spatial (point, laboratory, hillslope, basin, etc.) scale, while they are often calibrated and applied at larger or smaller scales. Observations used to drive a model, such as precipitation, may be collected at a small scale, and then distributed at a larger scale for model input. A watershed model may be run at a smaller scale (e.g. DEM grid or hillslope) and validated against discharge at the basin scale. Conversely, a groundwater model may be run at a larger scale and validated against small-scale borehole measurements.

The dominant hydrological processes may also vary with scale. At the laboratory scale, the flow through pores of the soil matrix may be the dominant mode of water transport. At the hillslope scale, macropores associated with cracks or root holes may transport the bulk of the flow. However, soil moisture and precipitation intensity must be great enough to allow the macropore network to become connected. Under these circumstances, upscaling a model formulated at the laboratory scale to the hillslope scale would require a different mathematical representation of subsurface flow. Such a formulation should also account for the influence of soil moisture and precipitation intensity on the effectiveness of the macropore network. Under the above circumstances, the original model, however, cannot actually be scaled up or simply extrapolated, but must be reformulated to include previously neglected processes.

Hydrological processes are normally represented in a model using mathematical equations composed of state variables (e.g. soil moisture) and parameters such as the soil-saturated hydraulic conductivity. The model is driven using observed or estimated input variables, localized (made appropriately suitable for application at a specific site) using estimated or calibrated parameters, and produces outputs that may be validated with observations. The large natural heterogeneity in catchment features and spatial/temporal variability in hydrological processes may require changes in model parameters, state variables, and even the fundamental equations when a model developed for one scale is upscaled to a larger scale or downscaled to a smaller scale.

Scaling is defined as a procedure that transfers information and implementations from one scale to another. Upscaling is defined as the procedure that transfers information and implementations from a small scale (higher resolution) to a larger scale (lower resolution). Downscaling is

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

the opposite, and transfers information and implementations from a large scale to a smaller scale. This definition of scaling is broader than that usually applied to the aggregation and disaggregation of model outputs. As defined here, scaling also has implications for model formulations, since data as well as process descriptions undergo a scale change simultaneously.

Scale can refer to the *extent* of the modeling domain: for example, with respect to space, a field plot, a small zero-order subbasin, a large river basin, and so on, and with respect to time, the total period of concern, seconds, days, years, or decades. Scale can also refer to *resolution* (e.g. a 30-m digital elevation model, a hydrography coverage mapped at 1 : 24 000 scale, etc.). A finite element of space and a finite duration of time are always implicitly involved in descriptions of all observations, hydrological processes, and model implementations, whether acknowledged explicitly or not. All hydrological observations require a control volume (for variables that refer to a three-dimensional piece of space, for example, soil moisture content), a control area (for variables that refer to a two-dimensional piece of area, e.g. leaf area), or a control length (for variables that refer to a one-dimensional curvilinear piece, e.g. fetch length) and are made over a sampling time period (e.g. hourly precipitation, daily pan evaporation, monthly discharge, annual change in reservoir storage, etc.). In this context, scale can also refer to *support*. It is the largest volume, area, length, or time interval for which the property is considered homogeneous within the interval. Thus, when applied to any aspect of hydrological modeling, scale can refer to extent, resolution, or support of observations, of processes, or of model implementations (Blöschl, 1996).

This article discusses upscaling and downscaling as it applies to physically based hydrological modeling. It is not meant to be an exhaustive discussion covering the entire range of scale issues in hydrology (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1** for more detail), or hydrological modeling (*see* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**), rather it attempts to summarize the relevant issues and illustrate some approaches used to address upscaling and downscaling in model development and application. This discussion draws heavily from the work of Bierkens *et al.* (2000), Blöschl (1996), Blöschl and Sivapalan (1995), and Grayson and Blöschl (2000). The first section discusses scales, hydrological processes, and heterogeneity in the context of model development and application. The second gives an overview of hydrological models and defines some basic terminology required for following discussions on upscaling and downscaling. Upscaling and downscaling are discussed next. The article concludes with a discussion of scaling issues and possible future directions.

SCALES AND PROCESS REPRESENTATION

Natural catchments typically exhibit a great degree of heterogeneity in physical features such as topography, soil characteristics, geology, and vegetation. Likewise, precipitation may also exhibit great spatial and temporal variability. Often, relatively small-scale heterogeneities are responsible for the most dramatic hydrological response. For example, in many locations the average soil permeability exceeds the average storm rainfall intensity. It is those locations with lower than average infiltration capacities that generate overland flow during peak rainfall. At the same time, the bulk of the soil erosion associated with this runoff may be transported in rills that occupy a small fraction of the surface area. Thus, the averaging associated with moving the model to a larger spatial/temporal scale may cause the model to ignore this runoff mechanism unless model input (spatial/temporal distribution of precipitation), soil parameters, and state variables (e.g. soil moisture) are developed to "capture" the influence of the (now) smaller-scale variability.

Hydrological processes occur over a wide range of spatial and temporal scales from the relatively small-scale process of infiltration to large-scale groundwater movement. Runoff generated from infiltration excess operates on short spatial (point) and temporal scales, beginning in portions of the catchment where the rainfall intensity exceeds the local infiltration capacity. Saturation excess runoff and return flow operate at the hillslope scale and longer temporal scales associated with the time required for a perched water table to develop and intersect the ground surface (Blöschl and Sivapalan, 1995). Regional groundwater flow has large spatial scales and temporal scales from a few months to hundreds of years. It has been noted that small temporal scales are typically associated with small spatial scales and vice versa (e.g. Blöschl and Sivapalan, 1995).

At the small to medium size catchment scale (tens to hundreds of km^2), hillslope processes dominate the timing and magnitude of the runoff hydrograph in most watersheds. At this spatial scale, the residence time for water in the channel system is small compared to the surrounding hillslopes and, in most cases, the channel network plays a minor role in the hydrological response of the watershed. At larger spatial scales, the residence time in the channel system is greater and there is a stronger influence of the channel networks on the timing and shape of the outflow hydrograph. Although hillslope processes still determine the volume of water input to the channel system, the spatial distribution of hillslope runoff production becomes less important. Therefore, at the larger scale, the importance of hydrological processes begins to shift from the hillslope to the channel network.

HYDROLOGICAL MODELS

Hydrological models are always implemented at specific spatial and temporal scales, whether explicitly stated as so or not (Beven and Binley, 1992). A hydrological model is driven using observed (sometimes also estimated or synthesized) data (the input variables), it keeps track of state variables, is localized (made appropriately suitable for application at a specific site) using a set of estimated or calibrated parameters, and produces outputs that are validated using observations of one or more state variables. The spatial and temporal scales of all inputs, state variables, and parameters must match. For example, an hourly time step, regular grid-based hydrological model must be inputted using hourly observations at the grid scale. The model implementation usually ensures that state variables are tracked and output is generated at the modeling scale. The process descriptions must also be developed (and their corresponding parameters estimated) at the modeling scale. Spatial and temporal modeling scales are readily obvious in model implementations, usually called *spatial and temporal resolutions of the model* (modeling element size used to discretize the modeling domain and the time step used for solution of model equations).

Input variables and output variables can change in space and in time. The state variables are related to the input and output variables through the model parameters. Model parameters (e.g. saturated hydraulic conductivity) represent the intrinsic properties of the model. They may vary in space and most typically are constant in time, although some models may use parameters that also vary with time, such as temporal changes in Leaf Area Index to represent seasonal changes in deciduous canopy cover. The values of parameters can also change with scale as they connect variables that may change in space and in time (Bierkens *et al.*, 2000). Constants (e.g. acceleration due to gravity) do not change in time or in space. The form of the model, expressed in terms of its process representation (i.e. descriptive equations) may need to be changed with scale. For example, a Darcy's Law–based approximation of soil matrix flow may be adequate to represent saturated flow at the laboratory scale, but may be insufficient to account for preferential flow paths at the hillslope scale.

Models may also be classified by how they represent spatial heterogeneities in model input variables, parameters, and state variables. Lumped models typically treat the entire system (e.g. a hillslope) as a single unit, either effectively ignoring spatial variability or tying to incorporate the effect of spatial variability into lumped parameters, while distributed models attempt to explicitly account for spatial variability at a given scale by subdividing the landscape into a number of model elements at that scale. Physically based models employing distributed parameters for describing system variability are the most challenging to scale.

Empirical models, such as the Rational Method simply define an input–output relationship with limited or no attempt to describe the physical behavior of the underlying hydrological processes. Conceptual models (e.g. Hydrologic Simulation Program Fortran, USEPA, 1984) represent (or conceptualize) basic hydrological processes using simplified algorithms with parameters and state variables that may not have a physical basis. Physically based models (e.g. SHE, Abbott *et al.*, 1986a,b) attempt to represent the basic biophysical processes of the modeled system using parameters with physical analogs that can be measured.

Each type of model introduces a somewhat distinct problem with respect to scaling. Empirical models should not be scaled beyond the original range of development, conceptual models may introduce more or fewer phenomena as scale is changed, and physical models have to confront the problem of variability as the resolution changes. It is generally easier, in the third type, to go upward by averaging or aggregating outputs obtained at finer resolution, but very difficult – perhaps impossible – to project down from a coarse rendition (e.g. estimation of local or watershed-scale precipitation from global precipitation predictions obtained by atmospheric models). The problem is one of recovering details that were not present at the original scale. Physical models are usually formulated with the presumption that all relevant processes are included, or at least the neglected processes are appreciated.

Physically based modeling involves the consideration of definite mass and flux balance equations relating specific physical processes believed to describe the system of interest at a particular scale. Mathematical scaling is required to extrapolate hydrological balance equations. Note that statistical scaling, however, involves the search for probabilistic invariance of a system's description and is applicable to a description in terms of statistics of random components comprising the representation. This subject is discussed in **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**, and is not addressed here, except to recognize that the variable quantities defining some needed model parameters may naturally or expediently be described as random (in space or in time).

Physically based models imply a certain deterministic interrelationship of component physical processes; however, their variables or parameters may often be conveniently viewed as randomly distributed or variable in a certain sense. Thus, in performing scaling it is helpful to employ mathematical tools that can be of utility in quantifying the variability of parameters. A reader will find that certain mathematical theories or methods have come to be associated with the scaling problem: probability theory and certain basic geostatistical methods, fractals, fuzzy sets, genetic algorithms, chaos theory, and others. Moreover, sometimes, analog models of mathematical structure can be substituted for the actual systems because they

have well-characterized mathematical attributes that can be studied instead of working directly with an incompletely known actual hydrological system. For instance, tree models of stream–river networks connected with their drainage basins are studied mathematically to understand the origins of similarity and scaling properties, or the invariances that are discovered in the measurements of the system response to inputs.

When process equations of the model remain nearly invariant under downscaling or upscaling of its variables, the model can be said to be scalable. Even if the form of model equations remains the same under scaling, the parameters that describe the basic process interactions may, and likely will, require reevaluation to encompass the influences of spatial or temporal variability encountered at the extrapolated scale. Scaling techniques must be used to properly account for the variability that was not originally incorporated in the model's conceptualization. This article discusses some of those mathematical techniques.

UPSCALING

Depending on the application and the model involved, there are four general methods that are typically employed for upscaling:

- *Averaging of input or output* data if the model is linear, or if it is viewed as amenable to a direct averaging procedure applied over many locations or time steps.
- *Effective parameters* if the model has the same form when extrapolated to another scale, using either deterministic or stochastic parameter values.
- *Average model equations* if the larger-scale model can be analytically derived from the smaller-scale model.
- *Fully Distributed Numerical Modeling* if no other appropriate upscaling techniques yield sufficiently resolved predictions or if detailed spatial information (input and parameters) is required to achieve the modeling objectives.

These approaches are discussed in more detail below.

Averaging Input or Output

Simple averaging of input or output provides a viable technique for upscaling under a limited set of conditions. It may be used to upscale linear models such as the Rational Method to estimate steady-state peak discharge from a homogeneous surface under constant rainfall and infiltration:

$$Q = CAP \tag{1}$$

where Q is the peak discharge ($m^3\ s^{-1}$), C is a runoff coefficient, A is the area (m^2), and P is the constant rate of precipitation ($m\,s^{-1}$). For example, consider discharge Q_2 from support A_2 composed of n smaller support areas (tributaries), A_i. The discharge can be obtained either by summing the output of n applications of equation (1):

$$Q_2 = \sum_{i=1}^{n} C_i A_i P_i, \tag{2}$$

or through the use of an average runoff coefficient

$$C_2 = \sum_{i=1}^{n} C_i Ar_i Pr_i \tag{3}$$

in equation (1), where Ar_i is the ratio of subarea i to the total area and Pr_i is the ratio of precipitation on subarea i to precipitation on A_2 (equals 1.0 for uniform precipitation over A_2). If the model is nonlinear, but can be applied independently at many locations and/or time steps, the model is run at the smaller scale and the output at the larger scale is obtained by averaging (or summing) the outputs from the smaller scale.

Effective Parameters

Most hydrological models, in particular those that are physically based, are too complex and nonlinear to allow simple averaging of the parameters as a means for upscaling. However, if the equations used to describe processes at the smaller scale remain valid at the larger scale, it may be possible to upscale the model using effective parameters. An effective parameter is used to capture the influence of smaller-scale heterogeneity at the larger scale. They are single values that reproduce the bulk behavior of a finite area or volume (Grayson and Blöschl, 2000). If the small-scale variation is known completely, deterministic methods may yield effective values. Typically, information about the parameter is only known at a limited number of locations (or times) and some function must be assumed that describes the spatial or temporal variation of the property at the smaller scale within the larger scale. This function can be either deterministic or stochastic.

Bierkens *et al.* (2000) illustrate the general approach and limitations associated with estimating effective parameters, in this case finding the representative conductivity for blocks in a numerical model of groundwater flow. The intent is to upscale fine resolution (small support) conductivity to the larger model blocks to decrease run time. Darcy's law is used to describe groundwater flow as:

$$q(x) = -k(x)\nabla h(x) \tag{4}$$

where x is the spatial coordinate, q is the groundwater flux vector, h is the hydraulic head, ∇h is the hydraulic

head gradient vector, and k is a tensor containing the fine scale hydraulic conductivities (in the different directions). Let $\overline{k(x)}$, $\overline{\nabla h(x)}$, and $\overline{q(x)}$ denote the average conductivity, average hydraulic head gradient vector, and average groundwater flux vector, within a larger-scale block respectively. The values of each of these variables within the block can be written as an average plus a deviation from the average: $k(x) = \overline{k(x)} + k'(x)$, $\nabla h(x) = \overline{\nabla h(x)} + \nabla h'(x)$, and $q(x) = \overline{q(x)} + q'(x)$, where the average of each deviation term is zero (e.g. $\overline{k'(x)} = 0$). Substituting these representations into Darcy's Law yields:

$$\overline{q(x)} + q'(x) = -(\overline{k(x)} + k'(x))(\overline{\nabla h(x)} + \nabla h'(x)) \quad (5)$$

The block average flux is obtained from equation (5) by taking the average of both sides (and noting the average of the deviation terms is zero):

$$\overline{q(x)} = -(\overline{k(x)}\ \overline{\nabla h(x)}) - \overline{k'(x)\nabla h'(x)} \quad (6)$$

Looking for a representative parameter with $\overline{q(x)} = K_e\overline{\nabla h(x)}$ implies from equation (6) that this effective conductivity must obey the relation:

$$K_e\overline{\nabla h(x)} = (\overline{k(x)}\ \overline{\nabla h(x)}) + \overline{k'(x)\nabla h'(x)} \quad (7)$$

Thus, the expression for representative conductivity consists of a term containing the block averages of the conductivity and head gradient, and another term involving the averaged covariation of the (small scale) deviations of the conductivity and head gradient. The solution depends on both properties of the porous medium and on the boundary conditions. As noted by Bierkens *et al.* (2000), specification of boundary conditions in the third term of equation (7) requires knowledge of flow at the small scale (i.e. for $\nabla h'(x)$), the very problem we were trying to avoid. In addition, these boundary conditions depend on both conditions within the block and on conditions outside the block, making the solution "nonlocal". To overcome this problem, a particular set of boundary conditions is assumed (e.g. uniform flow), allowing a local solution. Analytical solutions require further simplified representations of the porous medium. As a result, most solutions are both nonunique and approximative.

Deterministic Methods

If information about the small-scale heterogeneity is sufficient, it may be possible to derive effective parameters under limited conditions through theoretical considerations or from experience. For example, it is known from various applications (and from theory) that the effective saturated hydraulic conductivity for uniform, two-dimensional flow (K_e) is well represented by the geometric average of the smaller-scale conductivities, K_i (Bierkens *et al.*, 2000; Desbarats, 1991):

$$K_e = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln K_i\right] \quad (8)$$

When information about the parameter, however, is only known at a limited number of locations (or times), some function must be assumed that describes the spatial or temporal variation of the property at the smaller scale within the larger scale. Then a deterministic function (e.g. inverse distance weighting, spline functions, or some other such interpolation) yields one upscaled value to represent the (unknown) small-scale variation associated with an area or volume in the larger domain. Such deterministic effective parameters are nonunique, because they depend on the particular interpolation used. Equation (8) is an example of deterministic upscaling, because it yields a single set of parameters. This upscaling is based on a supposition that data is available to directly compute a set of effective parameters, which are assumed to be deterministically established, but are not necessarily uniquely known. Therefore, the term deterministic applies only to the scaling rule that yields a single upscaled value given a specific set of parameter values at the smaller scale. Also, parameters obtained by inverse methods of matching a model's predictions with measurements or observations of its outputs for given inputs can be included as being of a deterministic effective type. Here the model itself provides the averaging metric. Moreover, the effective parameters obtained by inverse fitting are dependent of the hydrological model's conceptualization, or, are specific to the model.

Stochastic Methods

Stochastic methods acknowledge the nonuniqueness associated with interpolation and attempts to account for this uncertainty by fitting a family of functions through the observations. Each function has similar statistical properties that are estimated from the observations (i.e. mean, variance, etc.), and is assumed to have the same probability of representing the true small-scale variation. The family of equally probable functions is called the *stochastic function*, and each particular function within the family is called a *realization* (Bierkens *et al.*, 2000). Instead of a single parameter value for a given portion of the larger-scale domain, the stochastic approach yields estimates of the small-scale values corresponding to each realization of the stochastic function, resulting in a probability distribution expressing the uncertainty about the effective conductivity due to the imperfectly known small-scale variation.

Upscaling rules for stochastic effective parameters are applied to each realization of the small-scale values to produce a realization of the upscaled value for the larger scale. This approach contrasts with applying equation (8)

as a deterministic upscaling procedure, because only a single effective parameter set was obtained. The probability distribution of the upscaled values can be estimated from these realizations, reflecting the uncertainty in the upscaled values due to the imperfectly known small-scale variation. Each realization of the upscaled value can then be used with the model in a Monte Carlo simulation to generate a suite of model outcomes. These alternative outcomes can then be used to estimate the probability distribution expressing the uncertainty in model output or state variables associated with the upscaled effective parameter. The probability distribution estimated for the upscaled value as well as that for the model output or state variables is just an estimate of the true probability distribution of these quantities. It is, in general, impossible to know the true probability distribution of these quantities. Nevertheless, some indication of the accuracy of model predictions can be established by comparing the distribution of predictions to the corresponding observed hydrological response.

Averaging Model Equations

In some cases the model itself can be upscaled, instead of finding effective parameters or averaging output. The upscaling is achieved through spatial or temporal averaging of the basic model equations at a smaller scale. Spatial averaging of infiltration has been the focus of considerable research (Maller and Sharma, 1981; Dagan and Bresler, 1983; Montoglou and Gelhar, 1987; Chen *et al.*, 1994, Nachabe *et al.*, 1995). The general approach is to use a local (point) infiltration equation and specify the spatial frequency distribution of its parameters. This approach typically involves defining the areal soil hydraulic conductivity frequency, $f(K_s)$. The lower tail of $f(K_s)$ represents soils with low conductivity, whereas the upper tail may represent a fraction of the surface with high conductivity. Nachabe *et al.* (1995) assume $f(K_s)$ follows a lognormal distribution (Nielsen *et al.*, 1973) and increases with elevation. They use Philip's infiltration equation:

$$I_p(t) = \frac{1}{2}\left(\frac{H\Delta\theta K_s}{b}\right)^{1/2}[t - (t_p - t_e)]^{-1/2} + K_s \qquad (9)$$

where $I_p(t)$ is the ponded infiltration rate at time t, H is the effective capillary drive, $\Delta\theta$ is the difference between initial and saturated soil moisture contents, $b \cong 0.55$ for most soils, t_p is the time to ponding, and t_e is an equivalent ponding time. At time t, the ponded fraction of the surface ($F_p(t)$) is:

$$F_p(t) = \int_0^{K_s'(t)} f(K_s)\,dK_s \qquad (10)$$

where $K_s'(t) = r[1 - (1 - t/(t + a))^{1/2}]$, the soil hydraulic conductivity at and below which ponding occurs, and $a = H\Delta\theta/4br$. The mean infiltration rate over the ponded area is:

$$\overline{I_p(t)} = \frac{\int_0^{K_s'(t)} I(K_s, t) f(K_s)\,dK_s}{F_p(t)} \qquad (11)$$

where r is the rainfall rate and $I(K_s, t)$ is given through equation (9). The area-averaged infiltration rate ($I_a(t)$) at t is:

$$I_a(t) = r[1 - F_p(t)] + \overline{I_p(t)}F_p(t) \qquad (12)$$

The first and second terms on the right-hand side of equation (12) represent the contributions of nonponded and ponded infiltration, respectively.

Nachabe *et al.* (1995) also consider cases where K_s decreased with elevation, and where K_s was assumed to vary randomly. In both cases, a spillover of excess rain from ponded areas to nonponded areas occurs. The fraction of ponded surface increased the fastest when K_s decreased with elevation, followed by the case of random K_s. They also used an effective K_s equal to the arithmetic and geometric means of the spatial conductivity field. The arithmetic mean was found to overestimate the infiltration rate, while the geometric mean underestimated the infiltration rate. It should be noted that this solution, like most upscaled equations, is nonunique and is an approximation to the true infiltration rate.

Fully Distributed Numerical Modeling

If the biophysical system to be modeled is one for which no appropriate upscaling techniques yield sufficiently resolved predictions or detailed spatial information is needed (which may be lost in an upscaling method), fully distributed modeling is required. Distributed parameter models attempt to represent hydrological variability by subdividing the model domain into a number of subareas or model elements. Input and model parameter values are required for each element. The response for individual elements is coupled by routing equations (e.g. Darcy's Law for saturated subsurface flow).

Hydrological processes with length scales smaller than the element size are represented implicitly, while processes with length scales larger than the element are represented explicitly by element-to-element variations (Blöschl and Sivapalan, 1995). The simplest (and perhaps most common) approach to expressing subelement variability of model parameters is to assume that a single value is valid for the whole element. The initial value is typically taken from limited measurements, perhaps at a different scale than the model element, with the final value obtained through calibration. With this approach, the parameter no longer represents a point value, rather, an effective parameter that attempts to reproduce the bulk behavior of the element (Grayson and Blöschl, 2000).

Subelement variability can also be expressed using distribution functions, where a distribution of values is used in the model equations, rather than a single value. Normally, this approach is applied to a limited number of input variables or parameters that control the dominant processes that operate at a subelement scale. Moore (1985) used a probability distribution of storage elements to model runoff production. Entekhabi and Eagleson (1989) used distribution functions of rainfall intensity and infiltration rate to calculate infiltration excess runoff. In many cases, the distribution is represented with a number of discrete classes. For example, elevation classes are used to express important subelement differences in air temperature and precipitation rates that govern snow accumulation and melt in mountainous regions.

The subelement variability may also be parameterized directly. Moore and Burch (1986) represented the effect of rills in a model element by the lumped equation:

$$R = F\ A^m \tag{13}$$

where R is the hydraulic radius (m), A is the cross-sectional area of flow (m^2), and F and m are parameters. The parameter m can be derived directly for simple cross-sectional geometries, or both F and m can be calculated from a graph of R versus A obtained from a microtopographic survey (Grayson and Blöschl, 2000).

DOWNSCALING

Downscaling is often required in hydrology to transfer large-scale observations or model predictions to a smaller scale. Examples include (i) estimation of local soil moisture from the catchment average (typically this example involves downscaling in space at a given instant in time), (ii) estimation of rainfall intensity pattern from bulk storm properties (typically this example involves downscaling in time at a specific location in space), and (iii) estimation of monthly streamflow at several locations within a catchment when only annual or seasonal catchment aggregate outflow may be available (this example involves downscaling both in space and in time).

Usually, downscaling involves disaggregation. Note, however, that some underlying properties of the system or laws governing the physical processes must be known to disaggregate the aggregate value at the larger scale into constituents at smaller scale. When the underlying system properties or governing laws can be expressed deterministically using mathematical relationships, these relationships can be used to disaggregate the larger-scale value into a unique pattern at the smaller scale. This procedure is classified as *deterministic downscaling*. However, when the underlying system properties can only be expressed probabilistically, there may be infinitely many disaggregated patterns that all obey the probabilistic system laws and aggregate up to the same larger-scale value. This latter procedure is called *stochastic downscaling* (the original definition of the term stochastic refers to random processes that change in time, but here we use the term more broadly that it may also include time independence).

Deterministic

The TOPMODEL Example

TOPMODEL was developed as a catchment-scale hydrological model (Beven and Kirby, 1979) that also contains a parameterization for local soil moisture deficit as a function of the catchment average. There are certain simplifying assumptions made (see Barling *et al.*, 1994) that allow this elegant analytical formulation. This formulation can be used to estimate the local soil moisture deficit at any modeling time step from the catchment average soil moisture deficit. It also allows prediction of areas of surface saturation within a catchment.

The final form of the soil moisture deficit relationship can be expressed as

$$d_i = d_{\text{avg}} + \frac{1}{f}\left[\frac{1}{A}\int_0^A \ln\left(\frac{a_i}{\tan\ \beta_i}\right)\,\mathrm{d}A - \ln\left(\frac{a_i \cdot T}{T_i \cdot \tan\ \beta_i}\right)\right] \tag{14}$$

where d_i is the local depth to the perched water table at location i (m), d_{avg} is the catchment average depth to the perched water table (m), f is a catchment average parameter that describes the exponential rate of reduction in soil transmissivity with depth, A is the catchment area (m^2), a_i is specific contribution area at location i (m), β_i is the angle of topographic slope at location i, T_i is soil transmissivity at location i ($m^2\ s^{-1}$), and T is the catchment average soil transmissivity ($m^2\ s^{-1}$). The distribution of the topographic index ln(a/tan β) has since been used as a surrogate for the spatial pattern of soil moisture within catchments.

As pointed out by Barling *et al.* (1994), the simplifying assumptions used during the development of equation (14) are not always appropriate. The most limiting of these, the assumption of steady-state subsurface flow, can be relaxed as shown by Barling *et al.* (1994), which allows for dynamic redistribution of soil moisture under variable rainfall rates while retaining most of the simplicity and elegance of the formulation. Beven and Freer (2001) developed a modified version of TOPMODEL that includes a dynamic specific contribution area in its governing equations. In the original TOPMODEL, there would always be some downslope flow, regardless of the soil moisture deficit. In the modified version, an additional parameter was used to indicate a limiting soil moisture deficit for downslope flow to start. The modifications also include downslope moisture redistribution between groups of landscape elements that

are hydrologically distinct. These landscape elements must be obtained using some form of classification based on distributions of landscape, soil, and vegetation information. Connectivity among these elements was determined from digital terrain analysis. The model retained the ability to estimate local soil moisture deficit.

Storm Precipitation Pattern

The use of design storms is a common occurrence in surface water hydrology. Foster and Lane (1987) suggest that the maximum information required to represent a design storm consists of the total storm amount and duration, average intensity, ratio of peak intensity to average intensity, and time to peak intensity. A double exponential function is fitted to the normalized intensity pattern of each storm, $i(t)$,

$$i(t) = \begin{cases} i_p \exp(b(t - t_p)) & 0 < t \le t_p \\ i_p \exp(-d(t - t_p)) & t_p < t \le 1 \end{cases} \quad (15)$$

where i_p is the ratio of peak intensity to the storm average intensity, t_p is the ratio of time to peak intensity to the storm duration, and parameters b and d are to be determined. Intensity-duration-frequency data developed from detailed precipitation data are typically used to prescribe i_p on the basis of storm duration and frequency. Integration of equation (15) over the appropriate limits results in two equations

$$1 - \exp(bt_p) = \frac{bt_p}{i_p} \qquad \text{at } t = t_p \quad (16)$$

$$1 - \exp(d(1 - t_p)) = \frac{d(1 - t_p)}{i_p} \qquad \text{at } t = 1 \quad (17)$$

that must be solved for b and d. From equation (15), $i(0)$ is equal to $i(1)$ so that $d = bt_p/(1 - t_p)$, and we only need to solve equation (16) for b to have the entire solution. Newton's method may be used to solve for b, given i_p and t_p.

The integral $I(t)$ of equation (15) can be written as

$$I(t) = \begin{cases} \frac{a}{b}[\exp(bt) - 1] & 0 < t \le t_p \\ \frac{c}{d}[\exp(-d(t_p - t)) - 1] & t_p < t \le 1 \end{cases} \quad (18)$$

where from above $a = i_p \exp(-bt_p)$ and $c = i_p \exp(dt_p)$. Notice that it must be that $0.0 < I(t) < 1.0$. If we subdivide the $I(t)$ interval [0,1] into n equal subintervals and call the right endpoint of these subintervals $F_1, F_2, \ldots, F_n$, then we can define specific time values as $T_1, T_2, \ldots, T_{n+1}$. These values of $T_1, T_2, \ldots$ can be defined by inverting the $I(t)$ function, such that $T_i = \text{inverse} I(F_i)$, for i from 1 to n.

The average normalized intensity over the interval $[T_i, T_{i+1}]$ is then calculated as $I_i = (F_{i+1} - F_i)/(T_{i+1} - T_i)$. The result of these calculations is an array of ordered pairs $[T_i, I_i]$ that are normalized time-intensity values. However, because the values of F_i are on a regular subinterval, the time intervals $(T_{i+1} - T_i)$ vary inversely with $I(t)$. That is, when $I(t)$ is high, then $(T_{i+1} - T_i)$ is small and when $I(t)$ is low, $(T_{i+1} - T_i)$ is large. This means that the method of disaggregation described here is unique in that values of F_i are all equal but the values of $\Delta t = (T_{i+1} - T_i)$ vary with rainfall intensity so that the highest intensity portion of the storm is always described by small Δt's. Moreover, a small number of intervals (n) can be used to describe storms or various durations. If we use n subintervals, then the dimensionless disaggregated time-intensity data are: $T_1, I_1, \ldots, T_{n+1}, I_{n+1}$. To restore the original dimensions, one must multiply each T_i by the storm duration and each I_i by the ratio of total storm precipitation and the storm duration. Calculations for a 6-h, 30.2 mm design storm with $t_p = 0.5$ and $i_p = 18$ are shown in Table 1.

Downscaling of RCM Precipitation

Hydrological modeling is often required to estimate the impact of climate change for water resources management and planning. This impact assessment requires us to drive

Table 1 Disaggregated 6 h, 30.2 mm design storm with $t_p = 0.5$ and $i_p = 18$

Dimensionless time Ti (1)	Dimensionless average intensity Ii (2)	Dimensioned time (min) (3)	Dimensioned intensity (mm hr^{-1}) (4)	Amount of rainfall during interval (mm) (5)
0.0	0.22	0.0	1.11	3.02
0.455	5.19	163.9	26.1	3.02
0.475	8.88	170.8	44.7	3.02
0.486	12.5	174.9	63.0	3.02
0.494	16.1	177.8	81.2	3.02
0.500	16.1	180.0	81.2	3.02
0.506	12.5	182.2	63.0	3.02
0.514	8.88	185.1	44.7	3.02
0.525	5.19	189.2	26.1	3.02
0.545	0.22	196.1	1.11	3.02
1.0	0.0	360.0	0.0	0.0

hydrological models using meteorological predictions from regional climate models (RCM). Typically, RCM grid cells are large compared to the extent of regions where water resources impacts need to be evaluated. While using spatially distributed hydrological models, a basis for downscaling the RCM predictions to the hydrological model grid scale is needed. Here we present an example that shows how precipitation predictions at the RCM grid scale may be downscaled to a hydrological model grid scale.

Typically, RCMs can only use a simplified description of land surface topography. Precipitation is highly correlated with topographic features. Land surface topography can be highly heterogeneous within the extent of an RCM grid cell. Owing to the simplified representation of topography, the RCM is only able to account for large-scale topographic features. A monthly precipitation patterns dataset, called Parameter-elevation Regressions on Independent Slopes Model (PRISM), for the continental United States has been developed using multivariate regression analysis of long-term precipitation observations and other topographic properties of the region (Daly *et al.*, 1994), and are available from Oregon State University.

Since the RCM grid cell is large compared to the hydrological model grid cell, a probability distribution of elevation as well as precipitation exists within the RCM grid cell. We make a simplifying assumption that the RCM-predicted precipitation can be assumed to correspond to the median of the distribution of precipitation within the RCM grid cell. This assumption gives us a basis to downscale the RCM-predicted precipitation from the RCM grid scale to the hydrological model grid scale using the probability distribution of the PRISM precipitation. Figure 1 shows the schematic representation of the downscaling approach.

Consider that for any time step P_{RCM} is the precipitation predicted by the RCM. We find the median PRISM precipitation, P_{PRISM}, from the distribution of PRISM precipitation for the same month to which the time step belongs, within the RCM grid cell. The downscaled precipitation at any hydrological model grid cell P_g is then expressed as

$$P_g = \frac{P_{\mathrm{PRISM},g}}{P_{\mathrm{PRISM}}} \times P_{\mathrm{RCM}} \tag{19}$$

where $P_{\mathrm{PRISM},g}$ is the PRISM monthly precipitation at the hydrological model grid cell g. Note that in this downscaling approach, the spatial pattern within the RCM grid cell as defined by the PRISM dataset is preserved during downscaling. Also, as the RCM predictions of precipitation change with time, the downscaled precipitation field also changes, although the spatial pattern remains the same for time steps that belong to the same month. The hydrological model is inputted with this downscaled precipitation field.

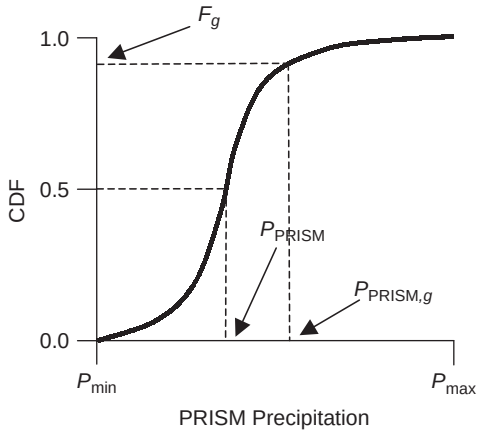


Figure 1 Schematic representation of an RCM precipitation downscaling approach. The cumulative density function (CDF) represents the probability distribution of precipitation within an RCM grid cell. P_{min} and P_{max} define the range of precipitation within the RCM grid cell. P_{PRISM} is the median PRISM precipitation within the RCM grid cell and $P_{\mathrm{PRISM},g}$ is the PRISM precipitation (corresponding to a cumulative density of F_g) at a hydrological model grid cell *g*, located within the RCM grid cell

Stochastic

Downscaling Hydraulic Conductivity in Space

One of the most frequent problems in spatially distributed hydrological modeling is the estimation of hydraulic conductivity of soil in space. Usually, only a few point measurements are available, while the model needs specification of hydraulic conductivity at all grid cells. Since measurement of hydraulic conductivity at the grid scale is infeasible, the point measurements must be downscaled to produce hydraulic conductivity at the grid scale.

The spatial correlation structure of hydraulic conductivity is used as the underlying law governing its distribution. Under isotropic conditions, a single, direction-independent semivariogram may be developed from measurements (e.g. Clark, 1979). A semivariogram quantifies the expected difference between values of hydraulic conductivity at two points in space. An interpolation scheme, such as kriging (see Journel and Huijbregts, 1978; Delhomme, 1978) can then be used to estimate the hydraulic conductivity values over the modeling grid.

An alternative to estimating the hydraulic conductivity field or map at the grid scale using kriging is to simulate these fields using statistical properties of the semivariogram with the Monte-Carlo procedure. These simulations produce equally probable fields of hydraulic conductivity by using the probability distribution of hydraulic conductivity while preserving measured values (Desbarats, 1996). These simulated fields can be used to quantify the uncertainty in the behavior of the hydrological system due to our imperfect knowledge of subsurface properties.

Another approach toward determination of hydraulic conductivity fields is to condition these fields on auxiliary measurements of state variables like the pressure head, using a mathematical subsurface flow model. These methods are also referred to as solutions to the inverse problem (estimation of subsurface properties given measurements of state variables); a large body of literature exists on this topic and more details can be found in **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**.

One of the most critical problems related to estimation or simulation of spatial distribution of hydraulic conductivity is the occurrence of contiguous regions of high or low conductivity (e.g. macropores, fractures, and sand or clay lenses). These "anomalies" result in preferential flow pathways within the subsurface. Statistical methods that rely on one-point (e.g. the probability density function) and two-point statistics (e.g. correlation, semivariogram, anisotropy) described above do not address these structures or "organization" adequately (Blöschl, 1996).

FUTURE DIRECTIONS FOR SCALING

There is no substitute for having sufficient data to perform hydrological modeling. Ideally, hydrological modeling requires measurements rich in both spatial as well as temporal coverage. However, many hydrological datasets that are rich in spatial coverage have poor temporal coverage and the reverse is also true. This limitation in hydrological datasets stems from several reasons: (i) there is a lack of measurement techniques that can provide both spatially as well as temporally rich datasets, (ii) there are very few accurate inferential methods that can be applied to infer variables of interest from observations of surrogates, and (iii) maintaining a hydrological measurement network is expensive. Until there are significant advances in hydrological measurement programs that address these issues, hydrologists will have to deal with extrapolation in space and in time.

One consequence of extrapolation (both while upscaling and downscaling) is that observations at the extrapolated scale may not be readily available to validate the predictions directly. It may be possible to obtain measurements at intermediate scales and validate the predictions at these intermediate scales to provide some indication whether the extrapolation produces reasonable agreements with observations across scales. In the absence of rigorously derived mathematical scaling laws, we must acknowledge that uncertainty in modeling predictions may be introduced as a result of this extrapolation. A detailed uncertainty analysis can then be performed to provide a measure of the uncertainty introduced into the predictions.

Since the goal of model scaling is an extrapolation of predictions, certain modern mathematical paradigms or constructs may find utility when applied to this problem. It seems unlikely that any scaling procedure can introduce accurate fundamental information that was not already present in an adequate measurement of input variables or estimation of the parameters describing the variability present in a hydrological system. However, there may be some mathematical methods that can at least facilitate the scaling extrapolation. Some mathematical constructs might even help a further understanding of the implications of scaling and some specific mathematical tools can be pointed out as potential contributors to treating the scaling problem. In addition to those mentioned below, certainly there are many other mathematical tools that are useful in hydrological model building and evaluating scaling properties, but we limit the following discussion to a few modern paradigms that show some special insights or utilities for the future of the scaling problem.

The contribution of the fractal theory in describing invariances under the self-similarity of a system's geometry may be useful to scaling. A fractal perspective has contributed to understanding rainfall sequences over space and time (Deidda, 2000), for example. Fractal structure can also be used to describe river networks (Rodriguez-Iturbe and Rinaldo, 1997; Tarboton *et al.*, 1988, 1992).

Fuzzy set theory may allow the variability of parameters and variables to be built into a hydrological model directly, without the need to consider the probability distribution of outputs associated with random inputs. It might provide a formulation to help accomplish scaling while accounting for variability. Bardossy (1996) has discussed the possible utility of fuzzy mathematics applied to hydrology. As another example, fuzzy set theory has found considerable use with the climate downscaling problem. Fuzzy conceptual rainfall-runoff models were described by Ozelkan and Duckstein (2001).

Neural networks have the potential for simplified and more efficient calculation of the nonlinear interactions between subunits of a system. The scaling of subunits into to a composite model unit might be made more computationally manageable. Whitley and Hromadka (1999) give an example of the use of neural networks for the study of flood prediction.

Percolation theory has found use in the study of the scale-dependent behavior of hydraulic conductivity (Hunt, 2003). It might find application in a further understanding of transport of water-borne constituents at larger scales.

An understanding of chaos theory has the possibility of replacing a complicated probability distributed description of hydraulic parameters by a more efficient deterministic generation. Chaos theory is finding use in hydrology, especially with evaluating a rainfall time series (Sivakumar, 2000). It may aid in understanding the scaling of rainfall over several timescales.

Kalman filtering may help with forecasting and improving the predictions of hydrological models. This may

help in testing model scaling and correcting predictions. An example of a Kalman filter application in evaluating soil moisture distribution is discussed by Walker *et al.* (2002).

Genetic algorithms may enhance the calculation of effective parameters in hydrological models to fit more extensive data records. Wang (1991) describes the use of a genetic algorithm to calibrate a rainfall-runoff model. McKinney and Lin (1994) discuss the use of a genetic algorithm to treat groundwater management and parameter selection.

At this time, we foresee scaling as the most challenging problem in hydrological modeling. We hope that this article provides a review of current practices in hydrology as they apply to scaling, fosters greater thinking on part of the reader, and generates spirited discussions within the hydrological community.

Acknowledgment

We appreciate the careful review and improvements to the manuscript suggested by C.S. Simmons.

FURTHER READING

Moore R.J. and Clarke R.T. (1985) A distribution function approach to rainfall-runoff modelling. *Water Resources Research*, **17**, 1367–1382.

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986a) An introduction to the European Hydrological System – Syst'eme Hydrologique Europ'een, SHE. 1. History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 45–59.

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986b) An introduction to the European Hydrological System – Syst'eme Hydrologique Europ'een, SHE. 2. Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

Bardossy A. (1996) The use of fuzzy rules for the description of elements of the hydrologic cycle. *Ecological Modelling*, **85**, 59–65.

Barling R.D., Moore I.D. and Grayson R.B. (1994) A quasi-dynamic wetness index for characterizing the spatial distribution of zones of surface saturation and soil water content. *Water Resources Research*, **30**, 1029–1044.

Beven K.J. and Binley A.M. (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K. and Freer J. (2001) A dynamic TOPMODEL. *Hydrological Processes*, **15**, 1993–2011.

Beven K.J. and Kirby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**(1), 43–69.

Bierkens M.F.P., Finke P.A. and Willigen P.D. (2000) *Upscaling and Downscaling Methods for Environmental Research*, *Developments in Plant and Soil Science, 88*, Kluwer Academic: Dordrecht, Boston, London.

Blöschl, G. (1996) *Scale and Scaling in Hydrology (Habilitationsschrift)*, Vol. 132, Institut fur Hydraulik, Gewasserkunde und Wasserwirtschaft, Wiener Mitteilungen, Wasser-Abwasser-Gewasser, Technical University of Vienna: p. 346.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrologic modeling – a review. *Hydrological Processes*, **9**, 99–109.

Clark I. (1979) *Practical Geostatistics*, Applied Science Publishers: London, pp. 129.

Chen Z.Q., Govindaraju R.S. and Kavvas M.L. (1994) Spatial averaging of unsaturated flow conditions over areally heterogeneous fields, 1 – development of models. *Water Resources Research*, **30**(2), 523–533.

Dagan G. and Bresler E. (1983) Unsaturated flow in spatially variable fields, 1, derivation of models of infiltration and redistribution. *Water Resources Research*, **19**, 413–420.

Daly C., Neilson R.P. and Phillips D.L. (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, **33**, 140–158.

Deidda R. (2000) Rainfall downscaling in a space-time multifractal framework. *Water Resources Research*, **36**(7), 1779–1794.

Delhomme J.P. (1978) Kriging in the hydrosciences. *Advances in Water Resources*, **1**(5), 251–266.

Desbarats A.J. (1991) Spatial averaging of transmissivity. *Proceedings of the 5th Canadian/American Conference on Hydrogeology: Parameter Identification and Estimation for Aquifer and Reservoir Characterization*, National Water Well Association: Dublin, pp. 139–154.

Desbarats A.J. (1996) Modeling spatial variability using geostatistical simulation. In *Geostatistics for Environmental and Geotechnical Applications, ASTM STP 1283*, Srivastava R.M., Rouhani S., Cromer M.V., Johnson A.I. and Desbarats A.J. (Eds.), American Society for Testing and Materials.

Entekhabi D. and Eagleson P.S. (1989) Land surface hydrology parameterization for atmospheric general circulation models including subgrid scale spatial variability. *Journal of Climate*, **2**, 816–831.

Foster G.R. and Lane L.J. (1987) *User Requirements: USDA-Water Erosion Prediction Project (WEPP)*, NSERL Report No. 1, USDA-ARS Misc. Pub. No. 1275, USDA-ARS Misc. Publications, pp. 147–153.

Grayson R. and Blöschl G. (2000) *Spatial Patterns in Catchment Hydrology*, Cambridge University Press: Cambridge.

Hunt A.G. (2003) Some comments on the scale dependence of the hydraulic conductivity in the presence of nested heterogeneity. *Advances in Water Resources*, **26**, 71–77.

Journel A.G. and Huijbregts C. (1978) *Mining Geostatistics*, Academic Press: London, p. 600.

Maller R.A. and Sharma M.L. (1981) An analysis of areal infiltration considering spatial variability. *Journal of Hydrology*, **52**, 25–37.

McKinney D.C. and Lin M.-D. (1994) Genetic algorithm solution of groundwater management models. *Water Resources Research*, **30**(6), 1897–1906.

Montoglou A. and Gelhar L.W. (1987) Stochastic modeling of large-scale transient unsaturated flow systems. *Water Resources Research*, **23**(1), 37–46.

Moore R.J. (1985) The probability distributed principle and runoff at the point and basin scales. *Hydrological Sciences Journal*, **30**, 263–297.

Moore I.D. and Burch G.J. (1986) Sediment transport capacity of sheet and rill flow: application of unit stream power theory. *Water Resources Research*, **22**, 1350–1360.

Nachabe M.H., Illangasekare T.H., Morel-Seytoux H.J., Ruan H. and Mapa R.B. (1995) Infiltration over a heterogeneous surface. In *Proceedings of the 15th Annual American Geophysical Union Hydrology Days*, Morel-Seytoux H.J. (Ed.) Hydrology Days Publications: Atherton, pp. 195–207.

Nielsen D.R., Biggar J.W. and Erh K.T. (1973) Spatial variability of field measured soil-water properties. *Hilgardia*, **42**, 215–260.

Ozelkan E.C. and Duckstein L. (2001) Fuzzy conceptual rainfall-runoff models. *Journal of Hydrology*, **253**, 41–68.

Rodriguez-Iturbe I. and Rinaldo A. (1997) *Fractal River Basins: Chance and Self-Organization*, Cambridge University Press: New York.

Sivakumar B. (2000) Chaos theory in hydrology: important issues and interpretations. *Journal of Hydrology*, **227**, 1–20.

Tarboton D.G., Bras R.L. and Rodriguez-Iturbe I. (1988) The fractal nature of river networks. *Water Resources Research*, **24**(8), 1317–1322.

Tarboton D.G., Bras R.L. and Rodriguez-Iturbe I. (1992) A physical basis for drainage density. *Geomorphology*, **5**(1/2), 59–76.

U.S. Environmental Protection Agency (USEPA) (1984) *Hydrologic Simulation Program – Fortran, HSPF, User's Manual for Release 8.0*, EPA-600/3-84-006, Environment Research Laboratory: Athens.

Walker J.P., Willgoose G.R. and Kalma J.D. (2002) Three-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: simplified Kalman filter covariance forecasting and field application'. *Water Resources Research*, **38**(12), 1301–1313.

Wang Q.J. (1991) The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resources Research*, **27**(9), 2467–2471.

Whitley R. and Hromadka T.V. (1999) Approximate confidence intervals for design floods for a single site using a neural network. *Water Resources Research*, **35**(1), 203–209.

12: Co-evolution of Climate, Soil and Vegetation

SANDRA L BERRY[1], GRAHAM D FARQUHAR[2] AND MICHAEL L RODERICK[3]

[1]Formerly of: Cooperative Research Centre for Greenhouse Accounting, Research School of Biological Sciences, Institute of Advanced Studies, The Australian National University, Canberra, Australia & Currently at: School of Resources, Environment and Society, Faculty of Science, The Australian National University, Canberra, Australia
[2]Cooperative Research Centre for Greenhouse Accounting, Research School of Biological Sciences, The Australian National University, Canberra, Australia
[3]Cooperative Research Centre for Greenhouse Accounting, Research School of Biological Sciences, The Australian National University, Canberra, Australia

The myriad of interactions between the vegetation, the soil, the climate, and the air involve transfers of energy and matter between the atmosphere and the lithosphere. In this article, these interactions are investigated mainly from the perspective of the vegetation. Most of the water taken up from the soil by plants is evaporated from the leaves via stomata during the process commonly referred to as transpiration. Transpiration of water is necessary for the uptake of carbon dioxide from the atmosphere for photosynthesis. Through transpiration and photosynthesis, the vegetation links the energy, water, and carbon and other biogeochemical cycles. This linkage provides the backbone to the article. Sections deal with: energy and water limitations to transpiration and plant growth at the catchment scale; the relationship between plant water use and photosynthesis; ecological strategies and temporal variability of water use and volume of soil involved in exchange with the atmosphere; soil water holding capacity and mineral nutrient availability; and catchment scale relationships between vegetation, soil, and climate. In the final sections, the concept of optimality between climate, vegetation, soil, and air is introduced, and the potential application of the optimality approach to hydrology is reviewed.

INTRODUCTION

The vegetation, the soil, the air, and the climate evolve together through a series of interactions involving transfers of energy and matter (Figure 1). The hydrological cycle is intimately involved in all of these interactions. Although each component in Figure 1 affects, and is affected by, all of the others, studies of the interactions have traditionally focused on particular interactions, for example, vegetation–climate (Budyko, 1974; Holdridge, 1967; Walter, 1979), plant–soil (Eyre, 1968; Black, 1968; Russell, 1973), soil–climate (Jenny, 1941; Brady and Weil, 2002) and air–climate interactions (see any climatology text). The interactions between the air and the vegetation and soil, and the feedbacks to the climate, have only recently become a research priority. This has been driven primarily by interest in the effects of the increasing concentration of carbon dioxide in the atmosphere resulting from anthropogenic combustion of fossil fuels.

As with any "new" field, the participants are on a steep learning curve. The prime difficulty is that one has to come to grips with several disciplines, for example, hydrology, ecology, climatology, atmospheric physics, biogeochemistry, and so on. This has prompted a need for interdisciplinary studies (e.g. Eagleson, 1986). Accordingly, the aim of this article is to present a brief overview of the principal interactions between the various components in Figure 1. Our overall approach is mainly from the point of view of the vegetation. Our hope is that by following this approach we will present the biological component of

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

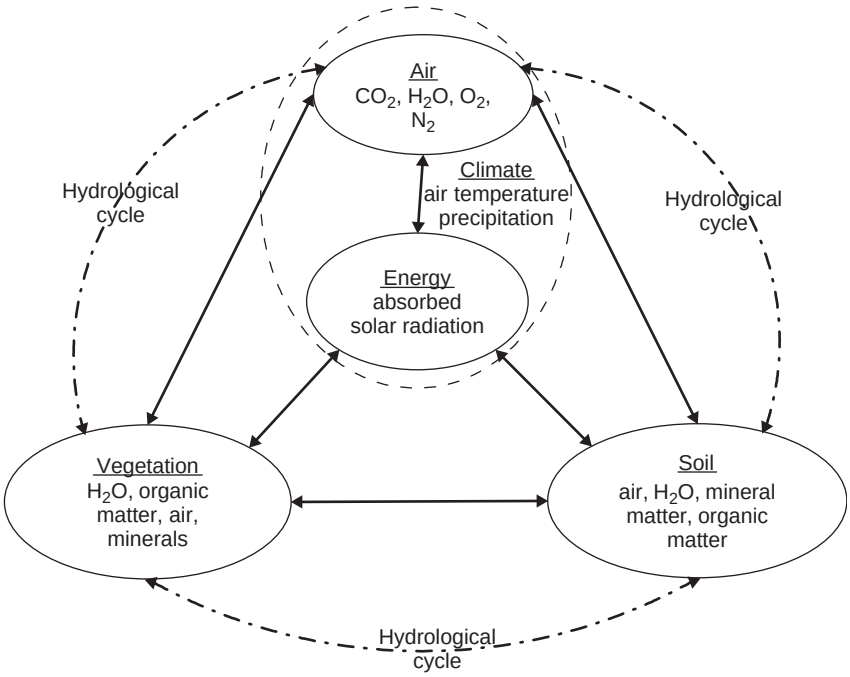


Figure 1 Relationship between the vegetation, the soil, the air, and climate. The climate is generally described in terms of the air temperature and precipitation, factors that depend on the energy and moisture content of the air. The most important components of the air, soil, and vegetation are italicized

Figure 1 in a way that is sympathetic to the needs of both practicing and research hydrologists. For example, we only briefly consider those topics that are traditionally close to hydrology (e.g. surface energy balance), but conversely, we spend more time on, for example, the biochemical basis for the observed differences in water-use efficiency between different vegetation types.

MASS AND ENERGY BALANCE

The processes in Figure 1 can be described in terms of transfers and stores of energy and mass (e.g. hydrological cycle, carbon cycle, nitrogen cycle). As mass and energy are conserved, the change in the store of mass and/or energy over a period of time within the air, the vegetation, or the soil must be equal to the difference between the input and the output. For a more detailed description of global energy and water balances, (*see* **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1**, **Chapter 25, Global Energy and Water Balances, Volume 1**).

Surface Energy Balance

The absorbed solar radiation is the ultimate source of energy for the terrestrial processes in Figure 1. The amount of solar radiation absorbed at the surface of the Earth is dependent on the quantity of solar radiation received at the top of the atmosphere, and consequently on the solar output and the geometry of the orbit of the Earth around the Sun, as well as on the Earth's atmospheric and surface properties. Only a portion of the solar irradiance arriving at the top of the atmosphere, Q_a, actually reaches the surface. The shortwave radiation flux that arrives at the surface is referred to as the global irradiance, Q_s. The remainder is absorbed or backscattered by molecules, aerosols, or clouds within the atmosphere. Data presented by Roderick (1999) indicate that on cloudless days in places having no atmospheric pollution, Q_s/Q_a may reach $\sim$0.80, but with increasing cloud the ratio may fall to $\sim$0.10. A proportion of Q_s is reflected back to space, and the fraction reflected back is referred to as the surface albedo, α. In addition to Q_s, the surface receives a flux of long-wave radiation $R_l \downarrow$ emitted by the molecules, clouds, and aerosols in the atmosphere. The surface also emits an upward flux of long-wave radiation, $R_l \uparrow$.

The overall radiation balance at the surface determines the net radiation, R_n, defined as the difference between the absorbed and emitted fluxes:

$$R_n = Q_s - \alpha Q_s + R_l \downarrow - R_l \uparrow \qquad (1)$$

The outgoing fluxes in equation (1) (αQ_s and $R_l \uparrow$) are largely controlled by surface properties. Values of α for terrestrial surfaces (Sturman and Tapper, 1996) vary from a maximum of $\sim$0.95 for clean, fresh snow to $\sim$0.05

for dark and wet soil and dense evergreen forest. As the vegetation canopy cover (or surface soil wetness) decreases, α increases from 0.05 towards the dry soil or desert value which can range from 0.20 to 0.45. Although the canopy cover of forests and cereal crops may be similar, forests often have a lower albedo than cereal crops, possibly because radiation reflected by leaves within forest canopies is absorbed by other leaves (Budyko, 1974).

The surface temperature largely determines $R_l \uparrow$ as given by the Stefan–Boltzmann law:

$$R_l \uparrow = \varepsilon \sigma T_s^4 \tag{2}$$

where ε is the emissivity of the surface relative to that of a full radiator or black body (for which ε is unity), σ ($5.67 \times 10^{-8}\,\mathrm{Wm^{-2}K^{-4}}$) is the Stefan–Boltzmann constant and T_s is the absolute temperature of the surface. There is a tendency for ε to decrease with increasing α (see data presented in Oke, 1987). T_s is dependent on the partitioning of net radiation through the heat balance and on the thermal properties of the soil. A detailed treatment of this subject is provided by Oke (1987).

The net flux of radiation to the surface (R_n) provides the energy to evaporate water (potential energy), to convectively heat the overlying air (kinetic energy), to heat the soil near the lithosphere/atmosphere interface (thermal energy), and to drive photosynthesis. In terms of these processes, the energy balance at the surface over a given time interval is:

$$R_n \cong \lambda E + H + G + A \tag{3}$$

where R_n is the net radiation ($\mathrm{J\,m^{-2}\,s^{-1}}$), λ is the latent heat of vaporization ($\mathrm{J\,kg^{-1}}$ H_2O), E is the evaporation rate ($\mathrm{kg^{-1}}$ $H_2O\,\mathrm{m^{-2}\,s^{-1}}$), H is the flux of energy convected away from the surface as sensible heat, G is the heat flux into the soil and A is the flux of energy transformed into energy of chemical bonds during photosynthesis ($\mathrm{J\,m^{-2}\,s^{-1}}$). As A is very much smaller than R_n, it is often ignored. However, A is coupled to the transpiration component of λE (see Section "Catchment mass balance"). Over the annual period:

$$G \cong 0 \tag{4}$$

and equation (3) can be simplified to:

$$R_n \cong \lambda E + H. \tag{5}$$

Catchment Mass Balance

The water balance of the active layer of the soil where water exchange generally occurs can be written as:

$$\frac{dS}{dt} = P - E - r \tag{6}$$

where P is the precipitation, E is the evaporation, r is the net (surface and ground) runoff and S is the storage of water in the active layer.

The evaporation (E) can be partitioned into three components: evaporation of intercepted water (E_I), evaporation from the soil (E_S), and evaporation via the stomata of plants, commonly referred to as transpiration (E_T).

$$E = E_I + E_S + E_T \tag{7}$$

The measurement and calculation of water balance parameters including these evaporation fluxes are described in **Chapter 35, Rainfall Measurement: Gauges, Volume 1**, **Chapter 40, Evaporation Measurement, Volume 1**, **Chapter 42, Transpiration, Volume 1**, **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**.

The vegetation can potentially modify all parameters in equations (6) and (7) above. The roots of plants commonly extend the thickness of the active layer of the soil involved in water exchange. Water is extracted by the roots from this extended active layer to furnish the transpiration requirements (E_T) of the vegetation. The transpired vapor may contribute to subsequent precipitation. For example, it has been estimated that about 48% of the annual transpiration in the Amazon Basin is recycled as rain (Salati and Vose, 1984). The canopy serves to increase the surface area for precipitation interception and consequently increases E_I. By reducing the exposure of the soil surface, the canopy usually reduces E_S. For example, for a rainforest in the Amazon Basin, E_T, E_I and E_S respectively comprised 75%, 19%, and 6% of the annual mean E of $1526\,\mathrm{mm\,yr^{-1}}$ (Costa and Foley, 1997). The values for grassland in the same region ($E = 1292\,\mathrm{mm\,yr^{-1}}$) were E_T61%, E_I20%, and E_S19%.

The catchment mass balance is a very useful tool in hydrology. For example, Costa and Foley (1997) observed a 12% reduction in annual E following the replacement of woody vegetation with grassland in the Amazon basin. In terms of the mass balance (equation 6), the reduction in E must be balanced by an increase in runoff and/or soil moisture storage in the active zone, or a decrease in P. These feedbacks are discussed in more detail in the Section "Individual plants and ecological strategy".

Budyko's Approach to Catchments

By assuming that over annual and longer periods the change in water storage (dS/dt) is zero, equation (6) can be simplified to:

$$P = E + r \tag{8}$$

In order to explain the geographic zonality of soils and vegetation, Budyko (1974) combined equations (5) and (8).

Using that approach:

$$E \cong \frac{R_n - H}{\lambda} \cong P - r \tag{9}$$

In equation (9), evaporation over the annual period is constrained by the availability of energy or water. The energy constraint limits the maximum possible evaporation to the value of R_n/λ which, according to Budyko (1974, Figure 99), is approximately equivalent to the annual potential evaporation, E_p, calculated from the atmospheric humidity deficit and temperature differences between the active surface and the air. As data for catchment r and E were not commonly available, Budyko derived a semiempirical equation (based on the available catchment measurements) to estimate E from P and R_n over the Earth's surface at the catchment scale from annual data:

$$E = \left[\frac{R_n P}{\lambda} \tanh\frac{\lambda P}{R_n}\left(1 - \cosh\frac{R_n}{\lambda P} + \sinh\frac{R_n}{\lambda P}\right)\right]^{1/2} \tag{10}$$

Very similar numerical values of E are obtained using the computationally simpler formulation, (Choudhury, 1999):

$$E = \frac{P}{\left[1 + \left(\frac{P}{R_n/\lambda}\right)^{2.6}\right]^{1/2.6}} \tag{11}$$

There is close agreement between E calculated with equations (10) and (11), and E determined by field observations for nine regions representing a wide range of global environments in the data set collated by Choudhury (1999, Figure 2). The relationship between calculated E (equation 11) and the water and energy budget over the annual period is summarized in Figure 2.

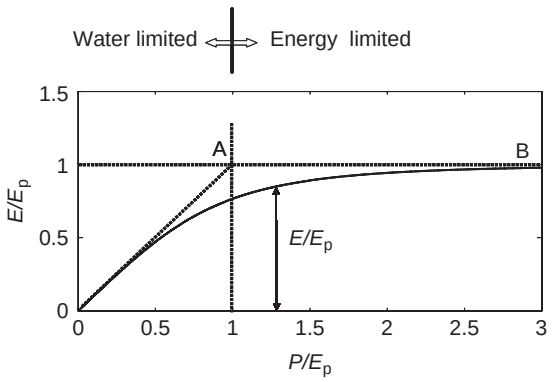


Figure 2 Dependence of the ratio of mean annual evaporation (E) on precipitation (P) and the radiation balance (E_p, where $E_p = R_n/\lambda$) following the approach of Budyko (1974). The line AB shows the energy limitation to evaporation. The solid line is drawn from equation (11)

Budyko (1974) named the ratio $R_n/\lambda P$. the *radiative index of dryness* and subsequently constructed a map of this index for the entire global land surface. By comparing this map with the available vegetation and soil maps he demonstrated that the availability of water and energy greatly influenced the geographic distribution of vegetation and soil types. The relationship between the radiative index of dryness and the geobotanic zonality of the Australian vegetation is shown in Figure 3. The relationships in Figure 3 between the vegetation structure and energy/water show the same pattern as those described by Budyko (1974).

VEGETATION

Most of the water that is removed from the soil by plants is ultimately transpired. As noted in Section "Surface energy

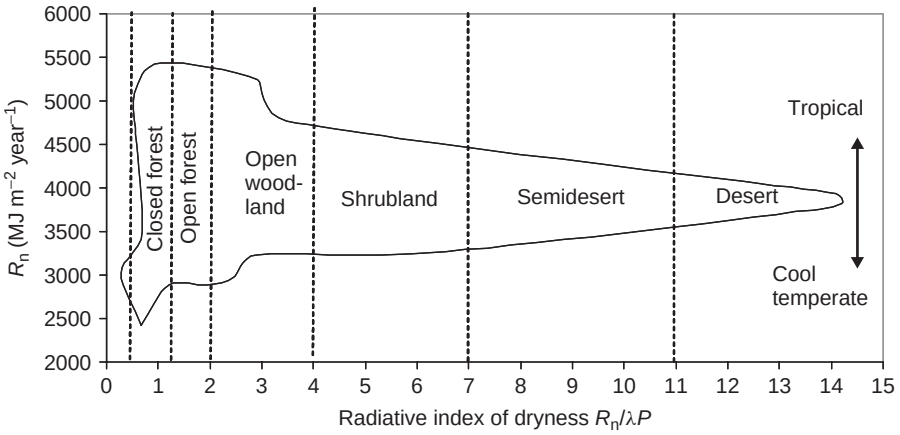


Figure 3 Relationship between vegetation structure, Budyko's radiative index of dryness and annual sum of net radiation for Australia. The continuous line describes the boundaries of values for the land surface. The net radiation and radiative index of dryness values were determined from gridded (0.05° × 0.05° latitude/longitude) data over the continent. Vegetation cover data are from the Natural Vegetation Map (AUSLIG, 1990). Data are from Berry and Roderick (2004)

balance" a very much greater amount of energy is involved in transpiration compared to that which is transformed to energy of covalent bonds by photosynthesis. However, the two processes are tightly coupled. Hence, in order to understand transpiration it is necessary to understand photosynthesis. For terrestrial plants, transpiration is a necessary part of photosynthesis (Section "Biological overview"), the source of energy and organic carbon molecules that sustains all life forms. The biochemistry of photosynthesis is similar in all plant species. However, modifications to the process of CO_2 uptake (Section "Leaf biochemistry") allow some plants to minimize transpiration per unit CO_2 uptake in water-limited environments, or to more efficiently scrub the air for CO_2 when this is in short supply. The amount of photosynthesis that can be done depends on the availability of the resources required for the construction and operation of the photosynthetic tissues. A range of ecological strategies have evolved (Section "Individual plants & ecological strategy") to allow plants to efficiently utilize resources. Resource availability is affected by soil properties (Section "Soil: water and nutrient cycles") as well as climate dynamics. Variations in photosynthesis and ecological strategies, and soil properties all ultimately affect the dynamics of the hydrological cycle over intra-annual and annual periods (Section "Catchment").

Biological Overview

The terrestrial plant consists of three basic parts: a canopy of leaves that act to capture light and carbon dioxide from the atmosphere; a network of roots which serve to anchor the plant into the soil, and to obtain water that is required by the leaves and mineral nutrients required for biochemical processes; and a stem which serves to elevate the leaves into the light, and is a conduit for the passage of water and nutrients from the soil to the leaves (through xylem conducting tissues), and a sugar solution from the leaves to the roots (via the phloem conducting network).

Through the process of photosynthesis, the canopy converts a small amount of the solar radiation that reaches the Earth's surface into the chemical energy of high energy C–C and C–H bonds. The source of the carbon is atmospheric CO_2 taken up by the leaves. The source of hydrogen is H_2O and this is taken up mostly by the roots (a small amount may be absorbed by the leaves). The numerous reactions of photosynthesis can be summarized:

$$n\mathrm{CO_2} + n\mathrm{H_2O} \xrightarrow{\text{light}} (\mathrm{CH_2O})_n + n\mathrm{O_2} \qquad (12)$$

Although the reactions of photosynthesis require water either directly (equation 12), or indirectly as the solvent, approximately 90% (Raven *et al.*, 1971) of the water absorbed by the roots is transpired. Water that is absorbed by the plant roots passes into the vascular tissues where it is transported by an elaborate hydraulic network (xylem vessels and/or tracheids) to the leaves. At the termini of the vascular bundles (i.e. the ends of the xylem conduits in the leaves), the water moves (by capillary flow) into the cell walls of the cells that surround the intercellular air spaces of the stomatal cavity (Crafts *et al.*, 1949) as shown in Figure 4. At the same time, CO_2 from the air within the stomatal cavity dissolves into the water that wets these cell walls.

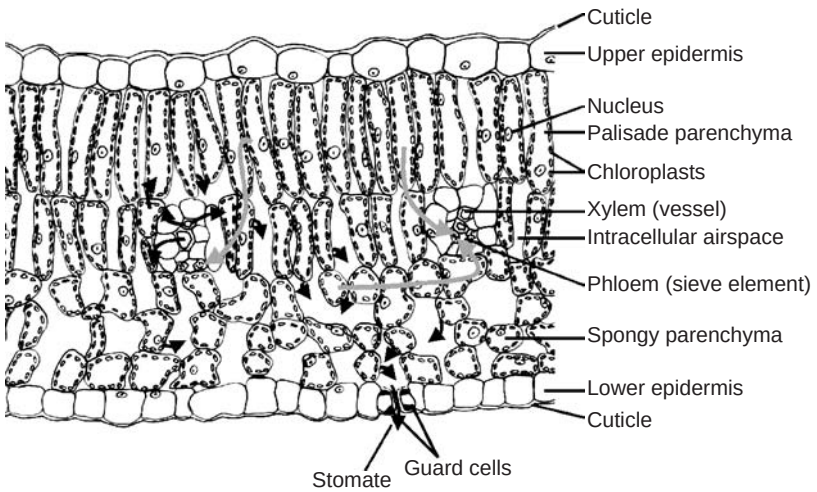


Figure 4 Diagram showing the arrangement of cells within a leaf and the pathways followed by water molecules (black arrows) and sugar molecules manufactured during photosynthesis (grey arrows). Photosynthesis takes place in the chloroplasts. The pathway of CO_2 molecules through the intracellular airspace is in the opposite direction to the water molecules

The dissolved CO_2 passes through the interstices of the cell wall, across the plasma membrane and into the chloroplasts where photosynthesis occurs. Whilst the stomata are open to allow CO_2 to diffuse into the intercellular airspace from the air outside of the leaf, the air within the intercellular airspace is almost saturated with water vapor, and there is a net evaporative flux of water molecules from the wet cell surfaces to the overlying air, and a net flux of water vapor to the exterior of the leaf. This flux of water vapor from the plant tissues to the air constitutes the transpiration. Transpiration is controlled by the guard cells at the entrance to the stomatal cavity. By closing the stomata, the guard cells block the flow of CO_2 into, and H_2O out of the intercellular airspace, and thus allow the plant to make conservative use of the available water. The ratio of net CO_2 uptake to net H_2O loss is referred to as the water-use efficiency of photosynthesis (WUE_{ph}), a parameter which links the carbon and water cycles. Sugar molecules manufactured in the chloroplasts, that are superfluous to the needs of the local tissues, are transported through the interconnected network of living parenchyma cells to the phloem tissues of the vascular system. From there the sugar molecules may be transported to developing stem and leaf tissues, or to the roots.

The energy obtained by the plant through photosynthesis is subsequently metabolized to build new molecules and to provide energy to maintain cellular processes. Metabolic processes that involve the oxidative breakdown of C–C and C–H bonds to release energy (and CO_2) are also referred to as respiration (Raven *et al.*, 1971). The overall reaction for the respiration of glucose to CO_2 and H_2O is the reverse of photosynthesis:

$$(CH_2O)_n + nO_2 \longrightarrow nCO_2 + nH_2O + \text{energy} \qquad (13)$$

The chemistry of photosynthesis and metabolism involves a wide array of biomolecules that contain a range of chemical elements in addition to carbon, hydrogen, and oxygen, as integral structural and/or functional components. These elements are generally classified as macronutrients (nitrogen, potassium, calcium, phosphorus, magnesium, and sulfur) required in relatively large quantities, and micronutrients (e.g. iron, chlorine, copper, manganese, zinc, molybdenum, boron), also essential for plant function but required in lesser amounts. Some plants also require cobalt and/or sodium (Raven *et al.*, 1971). For a comprehensive review of plant mineral nutrition, see Chapin (1980). The primary source of these nutrients is the soil, and their availability for uptake by plants is dependent on factors including soil parent material, physical and chemical properties, and the presence of soil microbes. The availability of nutrients may be modified by the anthropogenic addition of fertilizers or substances to amend excess soil acidity or alkalinity. Soil moisture is of paramount importance to nutrient availability as it is required to sustain hydration of soil organisms as well as being the solvent for soil chemical processes.

When all resources required for plant growth are well supplied the vegetation can form a dense canopy that intercepts most of the photosynthetically active radiation incident upon it. If resources are generally well supplied, except for an essential nutrient such as nitrogen or phosphorus, a less luxuriant canopy having lower photosynthetic capacity will result. Whilst limitations to photosynthesis and growth arising from insufficient availability of mineral nutrients and light are important, at the catchment scale the availability of water usually has the greatest impact. It is for this reason that Budyko's index of radiative dryness is correlated with vegetation cover (Figure 3).

Leaf Biochemistry

In water-limited environments the rate of transpiration relative to the rate of carbon assimilation in leaf tissues is affected by the biochemical pathways initially utilized by the leaf tissues prior to the bulk of the photosynthesis reactions that are invariant between plant species. The various initial pathways, known as C_3, C_4, and (crassulacean acid metabolism) CAM, reflect evolutionary processes that confer enhanced probability of survival and growth on plants under a range of water and atmospheric CO_2 concentration scenarios. The C_3 pathway is discussed first, as this is utilized by most plant species.

The equation for photosynthesis (equation 12) summarizes a complex series of reactions that occur within the chloroplasts, organelles of specialized cells (photosynthetic parenchyma) within plant leaves. The reactions take place in two stages, a light dependent and a light-independent stage. Energy generated by the light reactions is used to reduce carbon dioxide molecules to sucrose during the dark (light independent) reactions. During the initial stages of the dark reactions, a molecule of CO_2 is bound onto the enzyme (ribulose 1,5-biphosphate carboxylase/oxygenase) commonly referred to as Rubisco. Rubisco catalyzes the reaction of the molecule of CO_2 with the 5 carbon ribulose 1,5-biphosphate (RuBP) molecule to produce 2 molecules of the 3-carbon compound, 3-phosphoglycerate. When most plants are exposed to $^{14}CO_2$ in the light, the first detectable radioactive compound produced is 3-phosphoglycerate (PGA). These plants are referred to as C_3 plants and they comprise about 93% of plant species (Nobel, 1994).

In C_3 plants, not all of the carbon that is initially taken up by the Rubisco enzyme is incorporated into glucose molecules. Some is reoxidized to CO_2 by a process known as *photorespiration* and this reduces the WUE_{ph}. Photorespiration occurs because the Rubisco enzyme can also take up O_2 molecules. When an O_2 molecule reacts with the RuBP molecule, one molecule of the 3-carbon

PGA and one molecule of the 2-carbon phosphoglycolate are formed. Only the PGA molecule enters into the dark reactions of photosynthesis. Through a series of reactions, some of the carbon in the phosphoglycolate molecule is saved and the rest is ultimately oxidized. The overall stoichiometry is that 2 oxidations causes the release of one molecule of CO_2 (Farquhar *et al.*, 1980a).

The amount of O_2 that is taken up by Rubisco, and consequently the amount of CO_2 released by photorespiration, depends on the relative concentrations of O_2 and CO_2. As O_2 is approximately 600 times more abundant in the atmosphere than CO_2, the ratio of O_2:CO_2, is much more sensitive to a small change in the concentration of CO_2 than O_2. As the concentration of CO_2 in the atmosphere has increased from $\sim$280 μmol mol^{-1} to $\sim$375 μmol mol^{-1} over the past two centuries, there should have been a concomitant increase in the WUE_{ph} of the vegetation over this period. Where evaporation is limited by water supply, this increase in WUE_{ph} would allow the available water to support more vegetation and more photosynthesis per unit area. Where evaporation is limited by energy supply and further increases in canopy cover are suppressed by light limitation, an increase in photosynthesis would still be expected. However, this may not translate into an increase in WUE_{ph} as transpiration may also serve to dry out the soil (and hence increase aeration) in the root zone in wet environments. This would be expected in environments in which evaporation is limited by energy supply, but photosynthesis is limited by the availability of nutrients and/or root respiration is limited by the availability of oxygen.

Losses to photorespiration are avoided or minimized in the 1% of species that utilize the C_4, and the 6% that utilize the CAM photosynthetic pathway (Nobel, 1994). Although few species utilize the C_4 pathway they are responsible for approximately 25% of the primary productivity of the global vegetation (Sage, 2004). A large component of the primary productivity of C_4 plants (e.g. maize, millet, sugarcane, and sorghum) is appropriated by humans for food for themselves or their livestock.

C_4 photosynthesis is a series of biochemical processes that act to concentrate the absorbed CO_2 around the Rubisco enzyme (Hatch *et al.*, 1971; recently reviewed by Sage, 2004). In C_4 plants, as the CO_2 passes into the cytoplasm of the mesophyll cells, it is converted to HCO_3^- by the enzyme, carbo-anhydrase, and this is subsequently bound to a (3-carbon) phosphoenol pyruvate molecule (PEP) in a reaction catalyzed by the PEP carboxylase enzyme (Larcher, 1995). The resultant 4-carbon compound is transferred to chlorenchyma cells (away from direct contact with the air, and O_2) and subsequently decarboxylated to form pyruvate and CO_2. The CO_2 is then taken up by Rubisco and bound to a RuBP molecule. The reactions of photosynthesis then proceed as for C_3 plants.

The C_4 pathway may be advantageous in environments having a low concentration of CO_2 in the atmosphere because the mesophyll cells of C_4 plants are able to extract almost all of the CO_2 from the intercellular air (Ehleringer *et al.*, 1991). The pathway may also be advantageous in arid environments, as the combination of effective scrubbing of CO_2 from intercellular air, and the absence of photorespiration, result in greater water-use efficiency of C_4 plants (Ehleringer *et al.*, 1991). This reduced requirement for water could also confer advantages in saline or toxic environments. As less Rubisco is needed for C_4 photosynthesis, the nutrient requirements of C_4 plants are potentially reduced (Larcher, 1995). Additionally, the C_4 pathway appears to be energetically more efficient than the C_3 pathway in environments where the concentration of CO_2 in the atmosphere is similar to or less than the present ambient concentration (Nobel, 1994).

The C_4 pathway is utilized by plants that grow in a range of environmental conditions. It is used by annual herbaceous plants, especially those that grow in tropical regions, and summer annuals further from the equator. It also occurs in some aquatic plants and some shrubs (Larcher, 1995), in addition to the perennial grasses that inhabit arid tropical and subtropical environments. C_4 photosynthesis makes it possible for perennial grasses in semiarid and arid regions to take up CO_2 when there is very little water available for transpiration. C_3 shrubs and trees survive in these regions by having very deep root systems that enable them to access water stored several meters below the surface. By contrast, the C_4 grasses can survive in dry conditions by using less water.

C_4 plants are also commonly found in tropical and temperate environments where there is abundant but seasonal precipitation (Ehleringer *et al.*, 1991). This suggests that the C_4 pathway may provide plants with other benefits in addition to increased WUE_{ph}. The increased efficiency of water and nutrient use appears to be of little benefit to C_4 plants such as maize and sugarcane. These crop plants are profligate consumers of both water and nutrients (Masefield *et al.*, 1969). Under optimal conditions and current atmospheric CO_2 concentration ($\sim$370 ppm), the maximal rate of net CO_2 uptake for fast-growing, highly productive, cultivated C_4 crops is about 1.25 times the rate for C_3 crops grown under comparable conditions (Nobel, 1994). The C_4 pathway may benefit these species that grow in dense stands by allowing them to assimilate more CO_2 (due to more efficient scrubbing of CO_2 from the atmosphere within the stand) during the day.

While C_4 plants separate CO_2 uptake from the photosynthetic carbon reduction cycle (i.e. the dark reactions of photosynthesis) in space, CAM (crassulacean acid metabolism) plants separate it in time (Larcher, 1995). The $\sim$6% of plant species that utilize the CAM pathway include most cacti and succulent leaved plants (Nobel, 1994). The stomata of CAM

plants open for CO_2 uptake at night. The initial process of CO_2 fixation to produce 4-carbon acids is the same as in C_4 plants. However, whilst in C_4 plants, the 4-carbon acids are exported to other cells, in CAM plants the acids are stored overnight in the vacuole of the same cell. The following day (while the stomata are closed), the acid is decarboxylated to provide CO_2 to the chloroplasts of the same cells. The CO_2 is then taken up by Rubisco and the reactions of photosynthesis proceed as for C_3 plants.

By taking up CO_2 at night, plants that utilize the CAM pathway are able to minimize water loss. According to Nobel (1994), the water-use efficiency of CAM plants is about six times that of typical C_3 plants. During periods of severe drought CAM plants are able to keep their stomata closed day and night, and capture and recycle CO_2 released by respiration (Larcher, 1995). Plants that use the CAM pathway can also take up CO_2 during the day using the C_3 pathway (Nobel, 1994). The CAM pathway provides an alternative mechanism by which plants can minimize water and nutrient use (because of the reduced Rubisco requirement) and thus survive in extreme (arid or saline) environments.

Individual Plants and Ecological Strategy

With the exception of the C_4 and CAM modifications to CO_2 uptake, the basic process of photosynthesis is reasonably conserved across the plant kingdom. Over a unit area, the availability of resources (carbon dioxide, water, light, mineral nutrients) at a given time to the plant tissues that require them limits the potential amount of photosynthesis and transpiration that can occur and this limits the extent of the vegetation canopy.

There are many possible ways in which the photosynthetic tissues forming the canopy can be arranged in space and time; these different arrangements can be conveniently called *ecological strategies*. Successful ecological strategies allow individuals, or species, to persist at a site over years to decades to centuries. The main strategies are summarized in Table 1. One or more of these strategies may be utilized by the vegetation at a site or within a catchment. The Budyko framework (Section "Budyko's approach to catchments") focuses on energy and water and does not apparently need to consider the ecological strategy of the vegetation in a catchment. This seems to be a useful approximation at the annual scale. However, hydrologists are also interested in the time course of events, for example, the magnitude and timing of peak and base flows, and these can be altered by the ecological strategy of the vegetation in a catchment (Figure 5). That is the subject of this section.

Dynamics of the catchment water flows are affected by the ecological strategies in two basic ways: (i) through separation in time of the evaporation flux from the precipitation flux; and (ii) through changes in the volume of the active soil layer that is involved in water exchange. Plants that utilize the annual or ephemeral strategy produce leaves and carry out their photosynthesis and transpiration when there is sufficient moisture and energy content in the upper soil profile. When annual or ephemeral plants can no longer access moisture because the soil is drying, or because the increased viscosity of cold water (Roderick and Berry, 2001) slows the flow of water sufficiently to drought the leaf tissues (Daubenmire, 1974), the stems and leaves die. The leaves of annuals and ephemerals are capable of very high rates of photosynthesis but are short-lived (weeks to <6 months). The deciduous strategy is similar to the annual strategy, except that the stems persist from year to year, and the woody root system allows deciduous plants to obtain moisture from a larger volume. The evergreen woody, perennial, and cacti/succulent strategies allow for plant survival by using the available water at a slower rate over a longer period (e.g. Hollinger, 1992). By retaining a canopy, these plants can exclude some light from deciduous and annual plants. The longer-lived leaves typical of evergreen plants have less tissue involved in photosynthesis and more structural tissue. Depending on the proportions of these two basic tissue types, the evergreen vegetation may be mesic (relatively low proportion of structural tissue) or sclerophyllous (relatively high proportion of structural tissue) (Berry and Roderick, 2002b). As perennial herbaceous plants obtain their moisture requirements from upper soil layers they must survive dry periods by conservative use of water. Some woody evergreens may also utilize this strategy. However, many woody evergreen plants growing in water-limited environments have extensive root systems to access water several meters below the soil surface. Water may also be stored in plant tissues for later use by the plant. For example, Waring *et al.* (1979) estimated the total water storage capacity of plant tissues in four stands of Scots pine forest in northeastern Scotland to be 124, 147, 153, and 212 $m^3\,ha^{-1}$. They found that the transpiration flux from these forests during the summer rarely exceeded 3 $mm\,day^{-1}$, and the stored water contributed 30–50% of

Table 1 Ecological strategies utilized by plants

Leaf longevity	Plant form and ecological strategy
Usually <6 months	Herbaceous (grasses and forbs) **Annuals and ephemerals** Woody (trees, shrubs, and vines) **Deciduous**
1 year or longer	Herbaceous (grasses and forbs) **Perennial herbs** Woody (trees, shrubs, and vines) **Evergreens**
No leaves, use stems for photosynthesis	Store water in fleshy stems **Cacti and succulents** Leafless woody (trees, shrubs, and vines) **Evergreens**

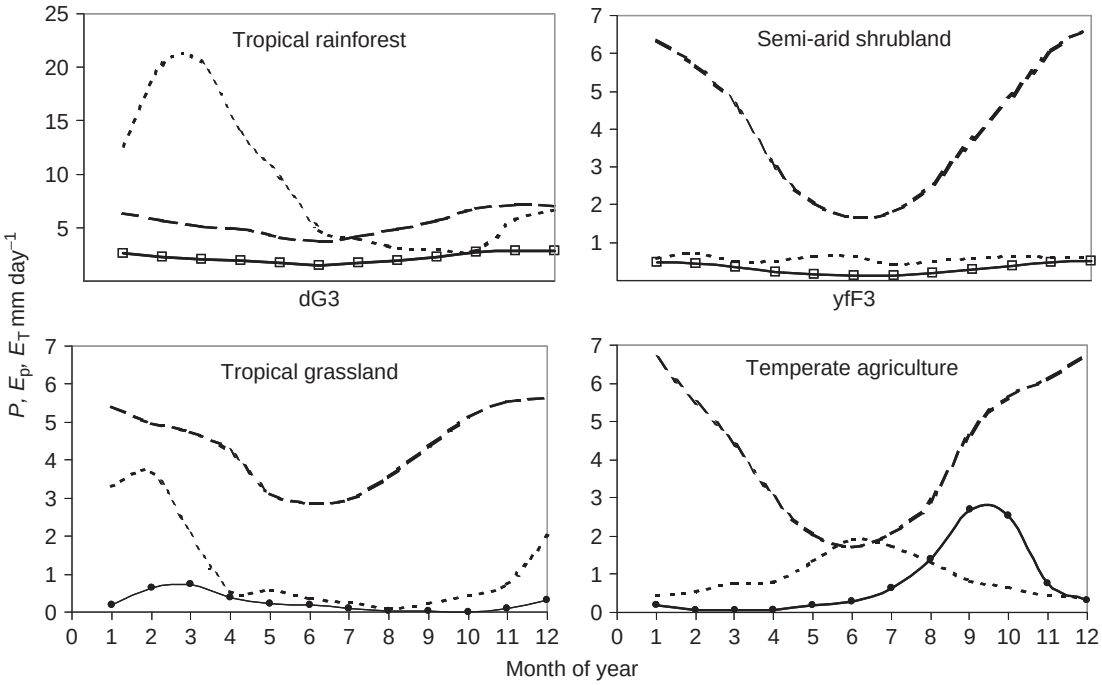


Figure 5 Estimated annual course of transpiration (E_T), precipitation (P) and potential evaporation (E_p) for four vegetation types of contrasting ecological strategy. Solid line, transpiration; open squares, evergreen woody; filled circles, annual herbaceous; Short dashes, precipitation; Long dashes, Potential evaporation. Data from Berry and Roderick (2004). Site locations are tropical rainforest 145.9 °E, 17.9 °S; tropical grassland 142.2 °E, 21.1 °S; semiarid shrubland 140.1 °E, 32.9 °S; temperate agriculture (winter wheat) 118.6 °E, 32.1 °S

the transpired water. In these forests, assuming a 30–50% contribution of stored water to a 3-mm day^{-1} transpiration flux, 150 m^3 ha^{-1} of stored water in plant tissues (equivalent to a 15-mm depth of water over the surface) would be depleted in 10–15 days. In water-limited environments, the utilization of water stored within the plant tissues or well below the soil surface allows the evaporation of water during periods of little or no precipitation (Fritts, 1976; O'Grady *et al.*, 1999).

Whenever evaporation is limited by water availability (see Figure 2), changes to the dominant ecological strategy (Table 1) associated with land use change may alter the hydrological balance. It has been commonly observed that the stream flow increases with reductions in forest canopy cover associated with fire or timber harvesting (Bosch and Hewlett, 1982). This has been observed in environments where evaporation is limited by water availability or energy availability (Malmer, 1992). Subsequent forest regeneration results in a decrease in stream flow to preclearance amounts (Fahey and Jackson, 1997; Langford, 1976).

In terms of the catchment mass balance, the observed changes in stream flow (runoff, r) could result from changes in $\mathrm{d}S/\mathrm{d}t$, P, or E (see equation 6). For practical hydrological purposes, the store of water in plant tissues may be considered part of $\mathrm{d}S/\mathrm{d}t$. As changes in r have been observed in paired catchment experiments (i.e. having a control and a treatment site; e.g. Malmer, 1992; Fahey and Jackson, 1997; Cornish and Vertessy, 2001) that allow P to be normalized, the residual (and large) changes in r must arise from changes in $\mathrm{d}S/\mathrm{d}t$ and E.

A decrease in the amount of water stored in the active layer where exchange occurs, $\mathrm{d}S/\mathrm{d}t$, would be expected when deep rooted vegetation is replaced by shallow rooted vegetation. This decrease would result from a reduction in the thickness of the layer of soil involved in active exchange and a reduction in the storage capacity of plant tissues. This is commonly observed when woody vegetation is cleared for pasture and agriculture, and replaced with annual herbaceous vegetation. When that happens, the dynamics of water flow are affected by both the decrease in the volume of soil involved in exchanges with the atmosphere, and the increase in the temporal variance of the transpiration flux (and a decrease in E) over the annual period (see the temperate agricultural example in Figure 5). This often results in a rise in the height of the water table (the top of the zone of water saturation) and the attendant problems of wet season waterlogging in low-lying parts of the landscape (Brady and Weil, 2002).

In environments where evaporation is limited by the availability of energy the soil profile typically remains wet throughout the year. Consequently, $\mathrm{d}S/\mathrm{d}t$ remains near zero and any change in r must result from a change in E.

In that case, a change in E may result from a change in net radiation (R_n, equation 5), arising from a change in albedo (α, equation 1) associated with the land cover change. This effect will be small when Q_s is small, for example, at high latitudes where the top of atmosphere solar irradiance is small, and/or very cloudy or polluted locations where the atmospheric transmittance is low.

Soil: Water and Nutrient Cycles

The formation of soil results from the interaction of the climate and the biota (vegetation and organisms) with the parent rock material. The soil comprises material in the solid (mineral and organic particles), liquid (water, minerals, and organic matter in solution), and gaseous (soil air) phases. The volume fractions of these three phases changes over time through the accumulation and leaching of solid particles, and with wetting and drying cycles.

Water in soils may be (i) bound to the soil solids as water of hydration (bound water), (ii) hydrogen bonded to other water molecules in capillary spaces (capillary water), or (iii) associated only with other water molecules or solutes (i.e. salts in solution), and free to flow through the soil pores (free water). The free water is only present when the soil is saturated, thus all of the pores are filled with water. When the soil is at field capacity, the macropores are filled with air and the water in the soil is held against gravity as bound and/or capillary water. At field capacity, only the capillary water is available to plant roots (Brady and Weil, 2002).

The water holding capacity (WHC) of soils is greatly affected by soil properties including the soil porosity, and the soil texture. The porosity is the volume fraction of the soil that is not occupied by solids. It can be estimated from measurement of the bulk density, which is the mass of dry soil per unit of bulk volume (soil plus pore space). The soil texture is a measure of the size distribution of soil particles.

There is a consistent trend of many soil properties to vary with particle size. Some of these are shown in Figure 6. Sand particles pass through a 2-mm sieve but are retained by a 0.05-mm sieve (Brady and Weil, 2002). They are usually comprised of quartz (SiO_2), so provide few mineral nutrients for plants. Being relatively large and globular in shape, they have a low ratio of surface area: volume (SA:V). Sandy soils lacking in organic matter have a low WHC. The large pores between the sand particles allow water to flow through rapidly, so sandy soils tend to drain freely following saturation as there are few capillary spaces.

Silt particles pass through a 0.05-mm sieve but are retained by a 0.002-mm sieve. These particles are basically microsand particles, dominated by quartz minerals (Brady and Weil, 2002). Although the particles are similar in shape to sand, they have a much larger SA:V ratio. The presence of capillary-sized spaces between the particles gives silt a much greater WHC than sand.

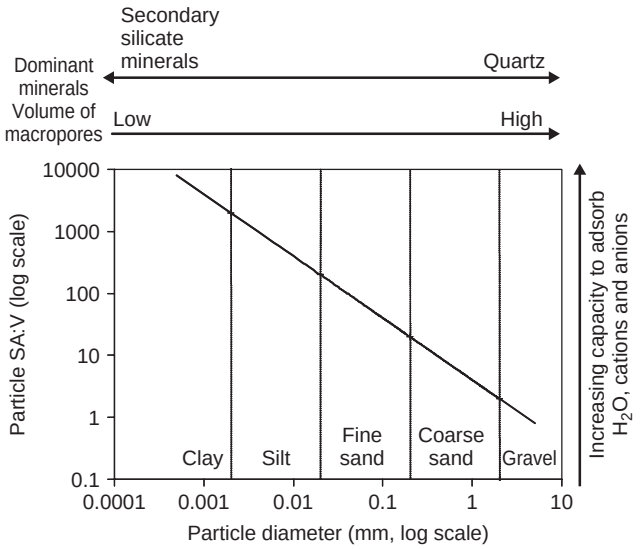


Figure 6 Relationship between soil particle size and general trends in soil properties

Clay particles are smaller than 0.002 mm and have a very large ratio of SA:V (Figure 6). Clay particles have internal as well as external surfaces, as they are structured as a stack of adhering layers (like a stack of plates). Consequently, there is an extremely large surface area onto which water molecules, anions, and cations can be adsorbed. The pores in clay soils are minute, and some kinds of clay swell on wetting. These properties give clays a very large WHC. Some of the capillary water is not available to plants in fine-textured clay soils as some of the capillary spaces are too small to allow root access. Consequently, the wilting point (i.e. the moisture content of the soil below which water is no longer accessible to plant roots) in clay soils occurs at higher soil moisture content than in the coarser soil types.

Although in the forgoing discussion the properties of individual soil particle size fractions were considered, soils may comprise one or a mixture of these particle size classes and thus display properties intermediate to the extremes. The mass fractions (dry mass basis) of the three components can be used to describe textural classes. The textural classes are generally illustrated in soil textbooks (e.g. Brady and Weil, 2002) using a ternary diagram, also known as the *texture triangle*. Whilst the soil texture is broadly related to soil WHC, soil aeration, and the potential availability of mineral nutrients required by plants, many other factors including climate, land use, and management, the chemistry of the parent material, and the vegetation also influence these soil properties. A more detailed discussion of these processes is provided by Brady and Weil (2002) and Schlesinger (1997).

The vegetation modifies soil properties by: (i) creating macropores through, for example, old root channels;

(ii) providing food for the soil biota; and (iii) producing organic compounds that resist decay and form the soil humus fraction. The formation of macropores by root penetration assists the percolation of water to deeper soil layers, reducing surface runoff. Soil biota serve many functions including the decomposition of organic compounds to release mineral nutrients, and improved soil aeration (e.g. through worm tunnels). Particles of humus have a similar size and ratio of SA:V to fine clay particles. The presence of humus greatly increases the WHC and nutrient holding capacity of soils. In low nutrient, excessively wet (anaerobic) environments the organic matter accumulates to form the upper soil horizons.

Catchment

At the catchment scale, the relationship between the vegetation, soil, and climate is sufficiently consistent to allow coarse resolution soil maps to be constructed from vegetation distributions. (For a detailed discussion of the relationships between vegetation formations and soil types see Eyre, 1968). In energy-limited environments (Figure 2), the vegetation has access to a plentiful supply of soil moisture throughout the year. In these environments, provided that the soil is not waterlogged and that the water is not frozen (e.g. as permafrost or snow), the vegetation forms a closed canopy (closed forest, Figure 3). The leaf properties (e.g. composition, morphology, capacity for photosynthesis and longevity) of the canopy reflect the availability of carbon (atmospheric CO_2 and light) relative to mineral nutrients (soil factors). As the concentration of CO_2 does not vary spatially at the catchment scale, the availability of mineral nutrients largely determines the type of vegetation forming the closed forest. The availability of nutrients is affected by factors such as the decomposition rate of organic material, as well as properties related to soil texture.

In water-limited environments (Figure 2), the canopy becomes increasingly more open as the water deficit increases (e.g. $R_n/\lambda P > 1$, Figure 3). In these environments, there is seasonal variability in the availability of water resulting from seasonality of precipitation and/or run-on and/or seasonal variation in solar radiation. As the environment becomes increasingly arid, the ecological strategy of the vegetation is dictated by the availability of soil moisture to plants (Stephenson, 1990). The plant available soil moisture (PAM, the difference between the moisture content at field capacity and at the wilting point) is affected by soil properties (see Section "Soil: water and nutrient cycles") and the volume of soil involved in water exchange with the atmosphere, as well as by the climate. The ecological strategies (Table 1) of the natural vegetation reflect the PAM and the inter- and intra-annual variability of PAM. This variability can be expressed as the coefficient of variation in PAM, or V_{PAM}, which is the standard deviation of PAM expressed as a percentage of the mean. Over the annual period, PAM is highly correlated with P/E_p (Walker and Langridge, 1997). (According to Budyko, 1974, the potential evaporation, E_p, can be equated to R_n/λ. Thus P/E_p is the reciprocal of Budyko's radiative index of dryness, $R_n/\lambda P$ in Figure 3.) Woody evergreen trees occur when P/E_p is moderate to low and plants can access moisture (perhaps at variable depths within the soil profile) throughout the year. With increasing aridity evergreen trees are replaced by evergreen shrubs. Deciduous trees and shrubs dominate at low to moderate P/E_p at higher V_{PAM}. This variability in PAM may be due to seasonal water deficit within the rooting zone of the plants associated with a soil moisture deficit (e.g. in monsoonal wet/dry tropics), or a seasonal energy deficit (and perhaps freezing of the soil) at high latitudes. As water cools, its viscosity increases and the water molecules become more strongly bound to the soil particles and the adjoining hydration layers. Thus, although water is present in the soil it may not be available to the plant at the rate required to meet the transpiration demand (Daubenmire, 1974). The herbaceous strategies dominate where woody plants cannot survive. C_3 perennial herbaceous plants dominate at moderate to high P/E_p where the soil is seasonally saturated in winter and PAM is low in summer due to a soil moisture deficit exacerbated by clay soils. C_4 perennial herbaceous plants dominate in environments that are too arid for woody vegetation but where PAM is sufficiently available (Section "Leaf biochemistry"). When there is high within year V_{PAM} resulting from the interaction of seasonal precipitation and clay soils, and P/E_p is moderate to low the annual herbaceous strategy is dominant. In low P/E_p environments having high within year V_{PAM} but reliable (but low) seasonal annual rainfall, cacti and succulents survive by storing water in their stems. In low P/E_p environments having high between-year V_{PAM} only, the ephemeral herbaceous strategy is viable.

TOWARDS OPTIMALITY

From the observation that there are repeatable patterns between the natural vegetation cover, climate, and soil properties (Sections "Budyko's approach to catchments" and "Catchment"), and between canopy cover and soil moisture availability (Stephenson, 1990; Specht, 1972; Woodward, 1987), it can be inferred that natural vegetation tends to some type of "optimality". Of course, that should not be a surprise because the Darwin–Wallace concept of natural selection ("survival of the fittest") includes the idea of optimality. However, the Darwin–Wallace theory is still largely qualitative. Hence, it is not surprising that many different optimality schemes have been used to try and predict the vegetation in terms of climate, along with the properties of the soil and the air. In some respects, these approaches are pragmatic in that they use the concept of optimality to simplify the problem at hand. For example,

one widely used idea is that the "natural" vegetation is arranged in space/time so that the photosynthetic uptake is optimized (e.g. Woodward, 1987). This, or similar approaches, are fundamental to many vegetation models (Schimel *et al.*, 1996; Schimel *et al.*, 1997).

In this section, we briefly review how the idea of optimality can be used in hydrology. Following that, we present some of the previous work on vegetation optimality from hydrology (Eagleson, 1982, 2002; Eagleson and Tellers, 1982), plant physiology and biophysics (Cowan and Farquhar, 1977), and ecology (Berry and Roderick, 2002a). After that, we briefly review a new and somewhat different approach that is based on a combination of fluid mechanics, thermodynamics, and an engineering philosophy of purposeful design (Bejan, 2000). This latter approach has not to our knowledge been applied in classical hydrology (or ecology), but we believe that it has the potential to be applicable throughout the hydrological sciences.

Some Simple Optimality Concepts

Perhaps the simplest way to introduce the idea of "optimality" is to return to the mass and energy balances of the Section "Mass and energy balance". For example, when E is limited by the availability of water, the photosynthetic uptake is limited by the capacity of the vegetation to access water for transpiration (E_T). In this case, an optimal vegetation cover (and mixture of ecological strategies) will act to maximize the appropriation of precipitation to the transpiration flux.

Alternatively, when E is limited by the availability of energy, the photosynthetic uptake is limited by the capacity of the canopy to access light (photosynthetically active radiation). In this case, an optimal vegetation cover (and mixture of ecological strategies) would maximize energy capture. One way of doing this is to increase the total single-sided leaf area per unit of land surface (known as the *leaf area index*). In light-limited environments a multilayered canopy is common.

Optimal Canopy Cover Based on Energy and Water Limitation to Plant Growth

Eagleson (1982) proposed theory and formulated hypotheses to predict the optimal projected foliage cover of the vegetation (PFC) from average annual water balance parameters and some basic soil hydraulic properties. In that and subsequent formulations it was assumed that there is always a nonlimiting supply of CO_2 and mineral nutrients. Also, the PFC is considered to be a surrogate for productivity which is in turn proportional to E_T. It was hypothesized (Eagleson, 1982) that when water limited, the optimal vegetation cover has a PFC that minimizes the "water demand stress" (i.e drought) under the local climate conditions. This requires that the optimal canopy uses the available water conservatively so that soil moisture is maximized whilst PFC and productivity are also maximized. The formulation was tested using catchment-scale estimates of the water balance along with canopy cover. The results were encouraging (Eagleson and Tellers, 1982). This is the second paper from the series and this has resulted in more research along this line (Eagleson, 2002).

One of the main aims for the original work by Eagleson was to avoid the need to collect vast amounts of detailed soil and vegetation information which is either impractical or impossible to achieve. Hence, the work must be viewed in that context. Further, the pioneering nature of the work meant that Eagleson made many simplifying assumptions. For example, that nutrients and CO_2 are nonlimiting. Hence, one could not use the Eagleson approach to assess the impact of increasing atmospheric CO_2 on the transpiration from catchments. Despite any shortcomings (e.g. Kerkhoff *et al.*, 2004), the Eagleson approach does demonstrate that the idea of optimality is potentially useful in hydrology and ecology.

Optimality of Canopy Water-use Efficiency of Photosynthesis

Cowan and Farquhar (1977) and Cowan (1982) proposed that there has been an evolutionary tendency for optimization of CO_2 uptake relative to water loss by plants that exist in water-limited environments. This is achieved by varying the transpiration rate throughout the day by movements of the stomata so that a given CO_2 uptake (and carbohydrate production) is achieved with the lowest possible loss of water. Equivalently, for a given water loss, the maximum amount of CO_2 is taken up. In their formulation, the optimum occurs when the partial derivative of transpiration (E_T) with respect to assimilation (A) is equal to a constant, τ;

$$\frac{\partial E_T}{\partial A} = \tau \tag{14}$$

If real leaves (and canopies) followed equation (14), then the assimilation would be a maximum for a given amount of transpiration, despite the within-day variations that may occur in the concentration of CO_2 and the vapor pressure of H_2O in the air surrounding the leaf. There is substantial experimental evidence (e.g. Farquhar *et al.*, 1980b; Williams, 1983; Hari *et al.*, 1999) that the behavior of stomata over an entire day is consistent with the Cowan–Farquhar optimality model. The outcome from this optimization is more rapid growth without a reduction in the probability of survival.

Unlike the Eagleson approach, nutrients and CO_2 are both explicitly included in the Cowan–Farquhar scheme via the involvement of nutrients and CO_2 in photosynthesis. The Cowan–Farquhar scheme has recently been further developed to incorporate optimization of nitrogen allocation

within the canopy to maximize the carbon gain from photosynthesis (Farquhar *et al.*, 2002; Buckley *et al.*, 2002). This allows for the prediction of changes in canopy leaf area in response to changes in the availability of water.

A Qualitative Approach Based on Plant Community Dynamics

One of the distinguishing features of natural vegetation is that monocultures are rare. Instead, ecological communities are usually composed of a mixture of species having different ecological strategies (Section "Individual plants and ecological strategy"). This can lead to difficulties for the Eagleson-type approach. For example, the productivity of a shrubland growing on a nutrient impoverished soil will be much less than that of a forest having the same PFC but growing on relatively fertile soil. The differences between the shrubland and forest noted above are largely described by differences in the longevity of the leaves (Roderick *et al.*, 2000) and hence the ecological strategy. Describing these differences is the main thrust of community ecology.

One recent empirical approach that explicitly recognizes the importance of different ecological strategies has used satellite observations to separate canopies into three classes of leaves according to leaf longevity (Berry and Roderick, 2002a,b). The basic tenet of this approach is that many of the physiological attributes of leaves are known to be consistently correlated with leaf longevity (Roderick *et al.*, 2000), and as noted above, are thereby also correlated with the ecological strategy. By comparing their results with existing vegetation maps, Berry and Roderick (2002a) found that the satellite-derived estimates of leaf longevity were realistic. The implication is that the properties of soil need not be known as they are already contained in the observations of the canopy. (This is consistent with one of the basic tenets of the Eagleson approach.) Berry and Roderick also showed how the framework could be used to make predictions, for example, what would be the impact of changes in disturbance or atmospheric CO_2 on the vegetation assemblage. The utility of their approach is that the leaf longevity classes provide a direct link back to the underlying physiology and the ecological strategy. Thus, this largely empirical approach could be used as the basis for understanding the relationships between ecological strategy and the dynamics of the water balance at landscape (and larger) scales.

One criticism of the Berry–Roderick approach is that optimality is implicitly assumed but is not made explicit because there is as yet no formal mathematical framework like the Eagleson and Cowan–Farquhar approaches that can be applied to a community of plants. This remains a major challenge to ecologists (and hydrologists) and one possible approach may be through the application of the "constructal" theory of Bejan which we discuss in the following section.

Potential Hydrological and Ecological Applications of Bejan's Theory

"Constructal" theory is the name given by A. Bejan to an approach that seeks to rationalize both engineered and natural flow systems (Bejan, 2000). Bejan has proposed that both engineered and natural flow systems "evolve" towards optimal states and has derived optimality criteria to describe those states. Note that the engineered systems "evolve" because the engineers designing the systems are always seeking to improve the system performance. The theory was originally conceived and developed in traditional engineering applications, and uses methods familiar to engineers (e.g. fluid mechanics, thermodynamics) but has not to our knowledge yet been applied in traditional hydrological applications (e.g. water balance, runoff estimation). Nevertheless, the theory (Bejan, 2000) provides a number of examples of flow systems directly relevant to hydrology. Some examples include the geometric configuration of river flow, flow through (engineered and biological) trees and networks of channels, along with derivations of the scaling laws that govern those flows (see Bejan (2000) for further examples). Because the Bejan theory addresses flow systems, and the optimal states for those flow systems, we believe that it has many potential applications in hydrology (and ecology). However, we only give a very brief introduction to this new work here. Interested readers should consult Bejan (2000) for more information.

The basic principal of constructal theory is that engineered (e.g. heat exchangers) and natural flow (e.g. botanical trees) systems contain regions of low and high flow resistances. For example, a typical heat exchanger has a solid casing (region of high resistance) and fluid flowing through channels (regions of low resistance). Constructal theory considers both the flow of matter and energy and proposes that the high and low resistance regions are geometrically distributed in an optimal way. The term "constructal" is derived from this idea, that is, real, finite-sized systems are "constructed" from smaller components.

The approach of Bejan (2000) has potential application to the formulation of many of the difficult practical problems in catchment-scale hydrology. For example, a long-standing problem in hydrology is how to characterize the permeability (in the Darcy sense) of a catchment. As is well known, the difficulty is that many catchments are composed of regions of relatively impermeable material but have smaller regions of preferential flow, for example, flow of water through root channels in the soil. Consequently, it is very difficult (if not impossible) to estimate the catchment permeability using point scale measurements because the regions of preferential flow cannot be adequately sampled. Prof. Jeff McDonnell (personal communication to MLR, October, 2003) has proposed that we might be able to use Bejan's approach to characterize the permeability of a catchment by assuming that a natural catchment would

"evolve" towards an optimal state. The optimal state would have a balance between the storage of water (the high resistance flow path) and drainage of water from the catchment (the low resistance flow path). Hence, catchments in very wet places would "evolve" to shed the water by the low resistance flow path, and so on. While speculative, this idea seems worth pursuing.

As stated above, we believe that Bejan's ideas may have widespread applicability in hydrology. The reason is that there is already some evidence for this from ecology – specifically – the flow of water through tree stems. For example, wood scientists and foresters have long known that the density of the solid material that makes up wood is near constant at about $1.5\,g\,cm^{-3}$. (This is more or less the density of cellulose that is listed in any materials handbook.) However, wood scientists and foresters have also long known that the density of dry wood (dry mass per unit total volume) varies from about 0.1 up to $1.1\,g\,cm^{-3}$. It follows that these differences must arise because of differences in the volume fraction of the solid material. For example, a "light" wood like Balsa has a very low volume fraction of solid material while a heavy wood like a tropical hardwood has a very high volume fraction of solid material. The significance is that if the tropical hardwood has a high volume fraction of solid, then it must have a low volume fraction of liquid and/or gaseous phases, and these phases are where most of the flow (and biochemistry) occurs. (The reverse case applies to Balsa.) Quite independently from Bejan's theory, Roderick and Berry (2001) used the above facts to derive physical expressions to describe the flow rate of water through tree stems assuming laminar (Poiseuille) flow. The resulting theory predicted amongst other things that (i) for a given pressure gradient the flow of water through wood would increase as the wood density declined and (ii) if the flowing water were cold, and therefore had a higher viscosity, that the wood density should decline. The common occurrence of "softwoods" (i.e. low density wood) in cold climates is consistent with the above prediction. Subsequent work on this theory in the field (Barbour and Whitehead, 2003; Atwell *et al.*, 2003) and in laboratory experiments (Thomas *et al.*, 2004) has found the theory to be in accord with observations. At the time that the above-noted theory was derived, the authors were not aware of Bejan's constructal theory. Nevertheless, in retrospect, the Roderick and Berry (2001) theory is in general accord with the ideas underlying Bejan's constructal theory.

CONCLUSION

Primarily as a consequence of interest in the effects of the increasing concentration of carbon dioxide in the atmosphere, an understanding of interactions between the vegetation, soil, climate, and air has now become a research priority. These interactions are complex and in this article we have sought to describe them through transfers of energy and matter across the atmosphere-lithosphere boundary, mainly from the perspective of the vegetation. Through this approach, the role of the terrestrial vegetation as an active (rather than a passive) link between the atmosphere, lithosphere, and hydrosphere becomes evident. By altering the surface properties, the vegetation can modify the outgoing fluxes of solar and terrestrial radiation (Section "Surface energy balance") and the partitioning of net radiation into latent and sensible heat (Section "Surface energy balance" and "Individual plants & ecological strategy"). The vegetation can also potentially modify all components of the hydrological balance (Section "Catchment mass balance"). Rainfall that may otherwise accumulate in the soil store or contribute to runoff if vegetation was absent is returned to the atmosphere by the transpiration flux from the leaves. The ecological strategy of the vegetation affects the dynamics of catchment water flows through temporal separation of the evaporative (transpiration) flux from the precipitation flux, and through changes in the volume of the active layer of soil involved in exchange of water with the atmosphere (Section "Individual plants & ecological strategy"). Organic matter has a high water holding capacity and its incorporation into the soil thus alters the soil hydrological properties (Section Soil: water and nutrient cycles). In the first three sections of this article, we discussed the vegetation in a general way. However, in the eco-hydrological context, vegetation implies a collection of individual plants that grow to form repeatable structural patterns (representing one or more ecological strategy) that relate to the availability of water and energy and soil properties. These repeatable patterns have long been recognized by natural scientists. For example, most vegetation mapping is based on the idea of repeatable (structural) assemblages. We discussed how these repeatable patterns had been described in terms of "optimality" (Section "Towards optimality"). One widely used approach to "optimality" is to assume an extremum principle based on, for example, maximization of the conversion of energy from solar radiation into chemical energy of covalent bonds (i.e. gross primary productivity, GPP). Other optimality approaches might also prove useful and we expect this to be a key research thrust of both hydrologists and ecologists over the coming decades.

FURTHER READING

Roderick M.L., Berry S.L. and Noble I.R. (1999a) The relationship between leaf composition and morphology at elevated CO_2. *New Phytologist*, **143**, 63–72.

Roderick M.L., Berry S.L., Saunders A.R. and Noble I.R. (1999b) On the relationship between the composition, morphology, and function of leaves. *Functional Ecology*, **13**, 696–710.

Rosenzweig M.L. (1968) Net primary productivity of terrestrial communities: prediction from climatological data. *The American Naturalist*, **102**, 67–74.

REFERENCES

Atwell B.J., Henery M.L. and Whitehead D. (2003) Sapwood development in *Pinus radiata* trees grown for three years at ambient and elevated carbon dioxide partial pressures. *Tree Physiology*, **23**, 13–21.

AUSLIG (1990) *Vegetation*, Australian Government Publishing Service: Canberra.

Barbour M.M. and Whitehead D. (2003) A demonstration of the theoretical prediction that sap velocity is related to wood density in the conifer *Dacrydium cupressinum* (rimu). *New Phytologist*, **158**, 477–488.

Bejan A. (2000) *Shape and Structure, from Engineering to Nature*. Cambridge University Press, Cambridge.

Berry S.L. and Roderick M.L. (2002a) CO_2 and land use effects on Australian vegetation over the last two centuries. *Australian Journal of Botany*, **50**, 511–531.

Berry S.L. and Roderick M.L. (2002b) Estimating mixtures of leaf functional types using continental-scale satellite and climatic data. *Global Ecology and Biogeography*, **11**, 23–40.

Berry S.L. and Roderick M.L. (2004) Gross primary productivity and transpiration flux of the Australian vegetation from 1788 to 1988 AD: effects of CO_2 and land use change. *Global Change Biology*, **10**, 1884–1898.

Black C.A. (1968) *Soil-plant relationships*, *Second Edition*, John Wiley & Sons: New York.

Bosch J.M. and Hewlett J.D. (1982) A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. *Journal of Hydrology*, **55**, 3–23.

Brady N.C. and Weil R.R. (2002) *The Nature and Properties of Soils*, *Thirteenth Edition*, Prentice-Hall: Upper Saddle River.

Buckley T.N., Miller J.M. and Farquhar G.D. (2002) The mathematics of linked optimisation for water and nitrogen use in a canopy. *Silva Fennica*, **36**, 639–669.

Budyko M.I. (1974) *Climate and Life*, Academic Press: New York.

Chapin F.S. III (1980) The mineral nutrition of wild plants. *Annual Review of Ecology and Systematics*, **11**, 233–260.

Choudhury B.J. (1999) Evaluation of an empirical equation for annual evaporation using field observations and results from a biophysical model. *Journal of Hydrology*, **216**, 99–110.

Cornish P.M. and Vertessy R.A. (2001) Forest age-induced changes in evapotranspiration and water yield in a eucalypt forest. *Journal of Hydrology*, **242**, 43–63.

Costa M.H. and Foley J.A. (1997) Water balance of the Amazon Basin: dependence on vegetation cover and canopy conductance. *Journal of Geophysical Research*, **102**, 23,973–23,989.

Cowan I.R. (1982) Regulation of water use in relation to carbon gain in higher plants. In *Physiological Plant Ecology II: Water Relations and Carbon Assimilation*, Lange O.L., Nobel P.S., Osmond C.B. and Zeigler H. (Eds.), Springer-Verlag: Berlin, pp. 589–613.

Cowan I.R. and Farquhar G.D. (1977) Stomatal function in relation to leaf metabolism and environment. In *Integration of Activity in the Higher Plant*, Jennings D.H. (Ed.) Society for Experimental Biology, Cambridge University Press: Cambridge, pp. 471–505.

Crafts A.S., Currier H.B. and Stocking C.R. (1949) *Water in the Physiology of Plants*, Chronica Botanica Company: Waltham.

Daubenmire R.F. (1974) *Plants and Environment; a Textbook of Plant Autoecology*, *Third Edition*, Wiley: New York.

Eagleson P.S. (1982) Ecological optimality in water-limited natural soil-vegetation systems 1. Theory and hypothesis. *Water Resources Research*, **18**, 325–340.

Eagleson P.S. (1986) The emergence of global-scale hydrology. *Water Resources Research*, **22**, 6S–14S.

Eagleson P.S. (2002) *Ecohydrology : Darwinian Expression of Vegetation Form and Function*, Cambridge University Press: Cambridge.

Eagleson P.S. and Tellers T.E. (1982) Ecological optimality in water-limited natural soil-vegetation systems 2. Tests and applications. *Water Resources Research*, **18**, 341–354.

Ehleringer J.R., Sage R.F., Flanagan L.B. and Pearcy R.W. (1991) Climate change and the evolution of C_4 photosynthesis. *Trends in Ecology & Evolution*, **6**, 95–99.

Eyre, S.R. (1968) *Vegetation and Soils: A World Picture*, Edward Arnold, London.

Fahey B. and Jackson R. (1997) Hydrological impacts of converting native forests and grasslands to pine plantations, South Island, New Zealand. *Agricultural and Forest Meteorology*, **84**, 69–82.

Farquhar G.D., Buckley T.N. and Miller J.M. (2002) Optimal stomatal control in relation to leaf area and nitrogen content. *Silva Fennica*, **36**, 625–637.

Farquhar G.D., von Caemmerer S. and Berry J.A. (1980a) A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species. *Planta*, **149**, 78–90.

Farquhar G.D., Schulze E.-D. and Kuppers M. (1980b) Responses to humidity by stomata of *Nicotiana glauca* L. and *Corylus avellana* L. are consistent with the optimisation of carbon dioxide uptake with respect to water loss. *Australian Journal of Plant Physiology*, **7**, 315–327.

Fritts H.C. (1976) *Tree Rings and Climate*, Academic Press: London.

Hari P., Mäkelä A., Berninger F. and Pohja T. (1999) Field evidence for the optimality hypothesis of gas exchange in plants. *Australian Journal of Plant Physiology*, **26**, 239–244.

Hatch M.D., Osmond C.B. and Slatyer R.O. (Eds.) (1971) *Photosynthesis and Photorespiration*, John Wiley & Sons, New York.

Holdridge L.R. (1967) *Life Zone Ecology*, Tropical Science Center: San Jose.

Hollinger D.Y. (1992) Leaf and simulated whole-canopy photosynthesis in two co-occurring tree species. *Ecology*, **73**, 1–14.

Jenny H. (1941) *Factors of Soil Formation*, *First Edition*, McGraw-Hill: New York.

Kerkhoff A.J., Martens S.N. and Milne B.T. (2004) An ecological evaluation of Eagleson's optimality hypothesis. *Functional Ecology*, **18**, 404–413.

Langford K.J. (1976) Change in yield of water following a bushfire in a forest of *Eucalyptus regnans*. *Journal of Hydrology*, **29**, 87–114.

Larcher W. (1995) *Physiological Plant Ecology*, *Third Edition*, Springer: Berlin.

Malmer A. (1992) Water-yield changes after clear-felling tropical rainforest and establishment of forest plantation in Sabah, Malaysia. *Journal of Hydrology*, **134**, 77–94.

Masefield G.B., Wallis M., Harrison S.G. and Nicholson B.E. (1969) *The Oxford book of food plants*, Oxford University Press: Oxford.

Nobel P.S. (1994) *Remarkable agaves and cacti*, Oxford University Press: New York.

O'Grady A.P., Eamus D. and Hutley L.B. (1999) Transpiration increases during the dry season: patterns of tree water use in eucalypt open-forests of northern Australia. *Tree Physiology*, **19**, 591–597.

Oke T.R. (1987) *Boundary Layer Climates*, *Second Edition*, Methuen: London.

Raven P., Evert R.F. and Curtis H. (1971) *Biology of Plants*, Worth: New York.

Roderick M.L. (1999) Estimating the diffuse component from daily and monthly measurements of global radiation. *Agricultural and Forest Meteorology*, **95**, 169–185.

Roderick M.L. and Berry S.L. (2001) Linking wood density, tree growth and environment: a theoretical analysis based on the motion of water. *New Phytologist*, **149**, 473–485.

Roderick M.L., Berry S.L. and Noble I.R. (2000) A framework for understanding the linkage between environment and vegetation based on the surface area to volume ratio of leaves. *Functional Ecology*, **14**, 423–437.

Russell E.W. (1973) *Soil Conditions and Plant Growth*, *Tenth Edition*, Longman: London.

Sage R.F. (2004) The evolution of C_4 photosynthesis. *New Phytologist*, **161**, 341–370.

Salati E. and Vose P.B. (1984) Amazon Basin: a system in equilibrium. *Science*, **225**, 129–138.

Schimel D.S., Braswell B.H. and McKeown R. (1996) Climate and nitrogen controls on the geography and timescales of terrestrial biogeochemical cycling. *Global Biogeochemical Cycles*, **10**, 677–692.

Schimel D.S., Braswell B.H. and Parton W.J. (1997) Equilibration of the terrestrial water, nitrogen, and carbon cycles. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 8280–8283.

Schlesinger W.H. (1997) *Biogeochemistry: an analysis of global change*, *Second Edition*, Academic Press: San Diego.

Specht R.L. (1972) Water use by perennial evergreen plant communities in Australia and Papua New Guinea. *Australian Journal of Botany*, **20**, 273–299.

Stephenson N.L. (1990) Climatic control of vegetation distribution: the role of water balance. *The American Naturalist*, **135**, 649–670.

Sturman, A.P. and Tapper, N.J. (1996) *The Weather and Climate of Australia and New Zealand*, Oxford University Press, Oxford.

Thomas D.S., Montagu K.D. and Conroy J.P. (2004) Changes in wood density of *Eucalyptus camaldulensis* due to temperature – the physiological link between water viscosity and wood anatomy. *Forest Ecology and Management*, **193**, 157–165.

Walker B.H. and Langridge J.L. (1997) Predicting savanna vegetation structure on the basis of plant available moisture (PAM) and plant available nutrients (PAN): a case study from Australia. *Journal of Biogeography*, **24**, 813–825.

Walter H. (1979) *Vegetation of the Earth and Ecological Systems of the Geo-Biosphere*, Springer-Verlag: Berlin.

Waring R.H., Whitehead D. and Jarvis P.G. (1979) The contribution of stored water to transpiration in Scots pine. *Plant Cell and Environment*, **2**, 309–317.

Williams W.E. (1983) Optimal water-use efficiency in a California shrub. *Plant Cell and Environment*, **6**, 145–151.

Woodward F.I. (1987) *Climate and Plant Distribution*, Cambridge University Press: London.

13: Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale

MURUGESU SIVAPALAN

Centre for Water Research, The University of Western Australia, Crawley, Australia

Catchment hydrology is presently operating under an essentially reductionist paradigm, dominated by small-scale process theories. Yet, hydrology is full of examples of highly complex behavior, including strong nonlinearities and thresholds, and paradoxes that defy causal explanation through these small-scale process theories. There are strong interactions and feedbacks between processes, leading to apparent simplicities in the overall catchment response, yet the laws governing these feedbacks are not well understood. Routine measurements and specialized field experiments have been valuable for observing catchment responses and understanding the underlying process controls, but there has been little progress in extrapolating the local knowledge and understanding gained from these well studied (or gauged) catchments to ungauged catchments. Efforts at generalization are hampered by the lack of an appropriate quantitative framework, for example, a classification system, to help identify interesting and useful patterns in the observations. There are many theories governing different elements of catchment hydrology, but not a unified theory that connects these seemingly disparate elements. This article presents the broad outlines of an emerging new, unified theory of hydrology at the catchment scale, and the approaches being used to develop it. The new theory embraces multiscale heterogeneities as a natural and intrinsic part of catchment hydrology. Instead of relying solely on current process theories, it seeks to discover new catchment-scale process theories that embed within them the effects of natural heterogeneities. Instead of attempting to prescribe in detail the actual patterns of heterogeneity in every catchment, it will seek to incorporate the geomorphic or landforming processes that may have generated them in the first place, and their ecological, pedological, and geomorphological functions. Instead of using our rather meagre observations to calibrate complex models that are based on small-scale theories, the theory will emphasize the use of patterns in the observations to formulate and test alternative hypotheses about the underlying process controls. Instead of using field measurements to learn more and more about individual catchments, it will seek to find connections between observations in different catchments, to identify broad-scale or general patterns. The defining feature of the new theory of catchment hydrology will be a sharp focus on the interconnection and feedbacks between pattern and process, over a range of scales, and their interpretation in terms of their "function", that is, the reason that these connections arise. The renewed focus on pattern, process, and function will revolutionize hydrology, elevate its place within the earth system sciences, and strengthen the scientific foundations of its practice.

INTRODUCTION

Owing to its focus on water, the science of hydrology holds a unique and central place in the field of earth system science, intimately intertwined with other water-related disciplines such as meteorology, climatology, geomorphology, hydrogeology, and ecology. As an applied science, hydrology is highly relevant to the management of the world's water resources and water quality, and for the prediction, prevention, and amelioration of water-related

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

natural hazards, such as floods and droughts. Thus, hydrology should be an exciting field of study, yet it appears to be fragmented, deeply rooted in empiricism, and struggling to realize its full potential.

Hydrology boasts of many theories, for each of its many constituent processes (e.g. infiltration, evaporation, overland flow, groundwater flow etc.), but there is an almost complete lack of an holistic theory unique to hydrology itself, unifying these many varied theories. In spite of the sophistication of the individual process theories, there has been little progress toward understanding the laws governing the interactions and feedbacks between these processes, so much so that models based on current process theories often cannot explain or reproduce key patterns of observed hydrological behavior (Tromp-van Meerveld and McDonnell, 2005). The reliance on individual process theories, and the lack of a unifying theory governing process interactions and feedbacks, have led to a proliferation of complex hydrological models, which suffer from overparameterization and high predictive uncertainty. As an applied science, hydrology derives many of its methods of analysis and predictive tools from the experience gained through specializations such as engineering hydrology, agricultural hydrology, urban hydrology, and so on. However, there is little, if any, common ground amongst these different perspectives, nor between these perspectives and the advances made in fundamental process understanding. Scientific progress and advances in hydrological practice have both been hampered due to the lack of a unified theory of hydrology at the catchment scale.

There have been frequent calls for a new unified theory of hydrology, as it would considerably improve our understanding of hydrological phenomena, including a more holistic understanding of their function within the entire earth system, and improve the scientific management of water resources, water quality, and water-related natural hazards (Dooge, 1986; Dunne, 1998; Sivapalan, 2003a). There has been an increasing recognition of the presence of natural, multiscale heterogeneities in hydrology, and of the need for an holistic, rather than fragmented description of these heterogeneities in hydrological theory and practice (Gupta, 2000). It has been suggested that with the advent of a new unified theory, we would not need to appeal to different, and often contradictory, conceptual models to explain (physical and chemical) phenomena that coexist in the same catchment (Kirchner, 2003).

This article presents a broad review of the current theoretical foundations of hydrology, the possible approaches to developing a more coherent and unified theory, and a brief survey of the progress that has already been made. On the basis of this review and a survey of current global trends in the scientific arena, the article identifies new opportunities and challenges that may be poised to accelerate the development of a new unified hydrological theory.

Subject Matter of Catchment Hydrology

In this article, we limit ourselves to the theory of hydrology pertaining to catchments, which are widely recognized as being the most fundamental landscape unit for the cycling of water, sediments, and dissolved geochemical and biogeochemical constituents. Catchments integrate all aspects of the hydrological cycle within a clearly defined area in a way that can be studied, quantified, and acted upon (Wagener *et al.*, 2004). It is for this reason that we choose catchments as the building block for the development of a new hydrological theory.

While practitioners of catchment hydrology approach the field from many different perspectives, all of them still have as their basis, the need to understand, manage and/or deal with space-time variability of catchment responses to climatic inputs (water and energy) at the land surface. Understanding of the spatial and temporal variability of hydrological processes aggregated to the catchment scale, their extremes, and their scaling behavior both in time and space, is important for a number of applications: for example, flood estimation, drought mitigation, water resources systems analysis. The pathways that water takes in its passage through the catchment, their spatial and temporal variabilities, and the associated residence times, are important for water quality predictions and for managing the health of aquatic ecosystems.

Hydrologists are also concerned with the need to understand and predict alterations to these hydrological responses due to changes at the earth's land surface and to the earth system as a whole, due to human impacts and any global change. Predicting the effects of human impacts, such as urbanization and deforestation, is important from the perspectives of water resources assessment, mitigation of natural hazards, and water quality management. Exchanges of water and energy between the land surface and the atmosphere, their sensitivity to long-term climate changes, and their impact on global water and energy circulations and teleconnections, are important for the study of global hydrology and of the global climate system. Therefore, improvements to the theory of catchment hydrology will have positive ramifications beyond hydrology, contributing to the sustainable management of land and water resources and aquatic ecosystems, and to managing global change.

Catchments as "Complex Systems with Some Degree of Organization"

Hydrological processes arise as a result of interactions between climate inputs and landscape characteristics that occur over a wide range of space and timescales (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**). In the time domain, these may range from a few seconds needed to capture turbulent exchanges of mass, energy, and momentum between the land surface

and the atmosphere, to intermediate timescales governing runoff generation processes during storm events, for example, overland flow and subsurface stormflow, and long timescales governing deep groundwater flow, seasonal variations of climate and annual water balances, and interannual and interdecadal variabilities. In the space domain, the length scales may range from an individual soil pore, leaf blade or surface gully, to small hillslopes, to river basins as large as the Mississippi, to whole climatic or geographic regions, all the way to the entire globe.

Due to the tremendous heterogeneities in landscape properties and climatic inputs, the resulting hydrological processes are highly variable and complex at all scales. It is not practical, or even feasible, to routinely observe hydrological processes at the scale of a soil pore or a surface gully, or at the scale of a hillslope, in all catchments. For both scientific and practical reasons, routine observations of hydrological processes are made only at the catchment scale, leading to a gap between the scales at which processes actually occur, and the scale at which routine observations are made and predictions are required. Catchments thus qualify as *complex* or *poorly defined* systems. This means that while process understanding at all scales, especially at scales smaller than catchment scale, is very valuable in guiding or underpinning predictions of catchment responses, actual predictions must still be based and/or conditioned on observations at the catchment scale.

On the other hand, the catchment is a self-organizing system, whose form, drainage network, ground and channel slopes, channel hydraulic geometries, soils, and vegetation, are all a result of adaptive, ecological, geomorphic or landforming processes. As a result, they lend themselves to regular geometric patterns, which, if understood and embraced, may actually lead to a simplification of catchment descriptions to be used in analysis and predictions. Dooge (1986) categorizes catchments as "complex systems with some degree of organization" (Figure 1).

Hence a holistic theory of hydrology at the catchment scale must be founded on a synthesis of process understanding and process theories *at all scales*, with empirical theories derived from the analysis of observations at the catchment scale, mediated by theories governing the natural organization and self-similarity underlying the spatial heterogeneities in landscape properties. The next three subsections will give a brief overview of the current status of process theories, empirical theories, and theories relating to natural organization of landscape properties.

Current Status of Process Theories

In view of the central role of both water movement and storage in the catchment, which take place on the land surface including in river channels, in various parts of the soil, as well as in vegetation, catchment hydrology currently derives many of its laws from sister disciplines such as open

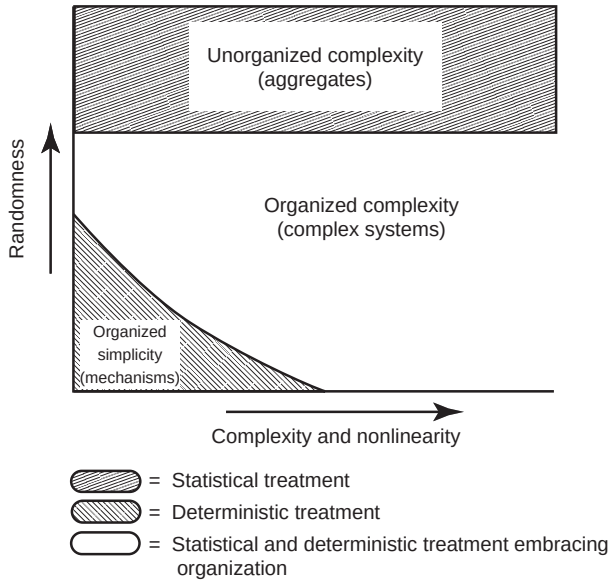


Figure 1 Catchments as complex environmental systems, that is, complex systems with some degree of organization. Adapted from Weinberg (1975) and Dooge (1986)

channel hydraulics, soil physics and chemistry, groundwater flow, crop micrometeorology, plant physiology, boundary layer meteorology, and so on. These laws and associated governing equations are used to quantitatively describe hydrological processes such as overland flow, snowmelt, channel flow, infiltration, recharge to the water table and capillary rise, evaporation, root water uptake and transpiration, and contaminant movement. Some examples include Darcy's law, Fick's law of diffusion, Manning and Chezy equations, the Saint Venant equations governing surface flows, Richards equation governing subsurface water movement, and the Penman and Penman–Monteith equations governing evaporation and transpiration. Detailed descriptions of these governing equations and their derivations have been presented elsewhere in this encyclopedia (*see* **Chapter 5, Fundamental Hydrologic Equations, Volume 1**), and in many standard textbooks, and will not be repeated here.

Considerable research has been carried out in the last 50 years toward the development and application of the governing equations, leading to significant advances in the understanding and description of many individual hydrological processes. Sophisticated numerical models have been developed on the basis of these governing equations, and the resulting numerical models have gained the status of physically-based models. However, it is often overlooked that the constituent process theories are essentially derived at the laboratory or other small scales. They are underpinned by assumptions of homogeneity, and uniformity, and time invariance of various flow paths, over the land surface and in channels, and through soils and vegetation (e.g. roots, stems, and leaves). In reality, catchments are

highly heterogeneous, dynamic and evolving entities (with respect to vegetation, soil structure, and morphology), responding dynamically to climatic inputs, which also exhibit tremendous variability in both space and time. One way that the catchment response can be modeled is by splitting the catchment into elements that are small and homogeneous enough so that the process theories can still be deemed to be applicable: this is the current paradigm (e.g. Abbott *et al.*, 1986a,b; Wigmosta *et al.*, 1994 and *see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**). Another way is to develop the balance equations for mass, momentum, and energy directly at the catchment scale. Some progress has been made in this direction (Reggiani *et al.*, 1998; Reggiani *et al.*, 1999); however, the needed closure relations at the catchment scale, to replace the current ones based on Darcy's law, Fick's law, and so on still remain to be developed to complete the specification of the governing equations.

The difficulty, or even fallacy, of the reductionist paradigm that has dominated hydrological science for the past 25 years has been discussed and debated at length in recent times (Beven, 1989a, 1993, 2000a,b, 2001, 2002). For example, it is highly impractical, using current and even future technologies, to describe in full the natural heterogeneity exhibited in catchments; in the unlikely case in which we can, the resulting models will be overly complex, and pose a huge computational burden. In the more likely situation in which the model parameters cannot be estimated *a priori* from observable landscape properties, the resulting models will pose a huge parameter estimation problem (Beven, 1989a).

An even greater difficulty arises from the fact that, on their own, traditional process theories cannot account for processes and process interactions that may occur at the catchment scale in the presence of natural heterogeneities and the natural self-organization underlying these. For example, while our best models are predicated on porous media flow theory based on Darcy's law and Richards equation, we observe non-Darcian flow in the field (Tromp-van Meerveld and McDonnell, 2005) due to the presence of preferred pathways such as macropores. The use of passive tracers, for example, stable isotopes, chloride, and so on has demonstrated in many catchments that while streamflow responds promptly to rainfall inputs, fluctuations in the passive tracers are strongly damped, indicating the stormflow is mostly "old" water (Sklash, 1990; Turner and Macpherson, 1990; Buttle, 1994). The old water paradox – the fact that catchments store water for considerable periods of time and then release it promptly during storm events – cannot be explained by these small-scale process theories (Kirchner, 2003). A variety of concepts have been invoked in attempts to explain this phenomenon – piston flow, kinematic waves, transmissivity feedback, exchange between matrix and macropores (Beven, 1989b; McDonnell, 1990; Bishop, 1991; Kendall *et al.*, 1999), but with limited success. Clearly, mechanisms other than contained in our current process theories, must be at work here.

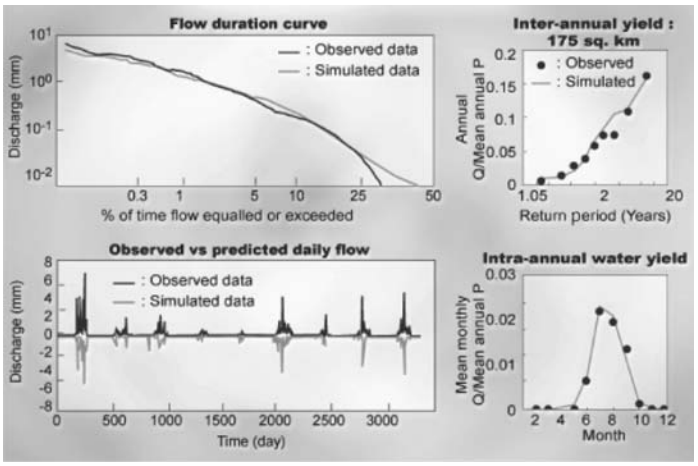
Current Status of Empirical Theories

Given that hydrological processes vary over a broad range of space and timescales, the business of hydrology is to understand, explain, and characterize hydrological variability, in space and time, including how this variability changes with time or space scale. Usually, this variability is characterized as a space-time field of the quantity of interest, be it streamflow, soil moisture, groundwater table depth, or rates of evaporation. Increasingly, we are also interested in the pathways that the water takes to arrive at the catchment outlet, the distribution of travel times, and the age of the water that exits the catchment as these are indicators of the underlying space-time variability of hydrological processes, and have implications for water quality predictions (Vache and McDonnell, 2005).

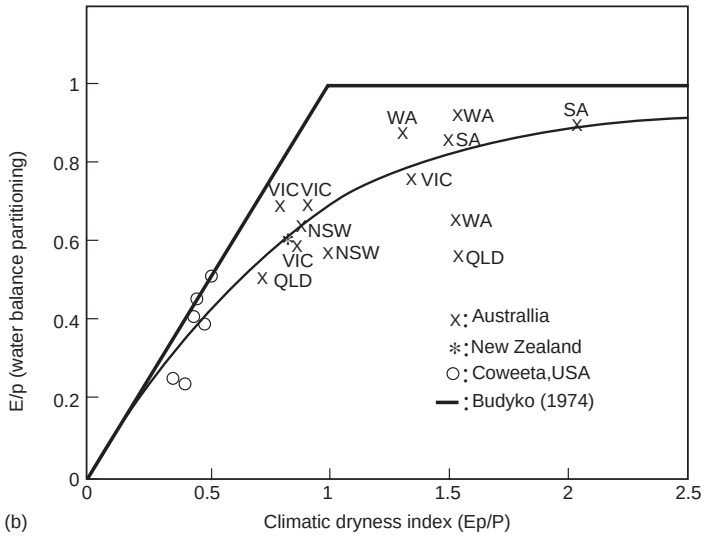
In the context of empirical data analysis, the role of theory is to provide a robust, quantitative, and reproducible framework to relate descriptors or *signatures* of hydrological variability to properties of the catchments and climatic inputs, which we might call *predictor variables*. When such robust relationships are established, we can then hope to predict the responses of catchments with confidence, given only the relevant climatic parameters and catchment properties. Therefore, the value of these signatures is not so much what they tell us about individual catchments, which is still considerable, but what they can tell us about differences between catchments. A robust and reproducible theory will evolve only when we broaden the search from ever more detailed explorations of processes *within* individual catchments toward quantitative and causal explanations of the differences *between* catchments.

Descriptors (Signatures) of Hydrological Variability

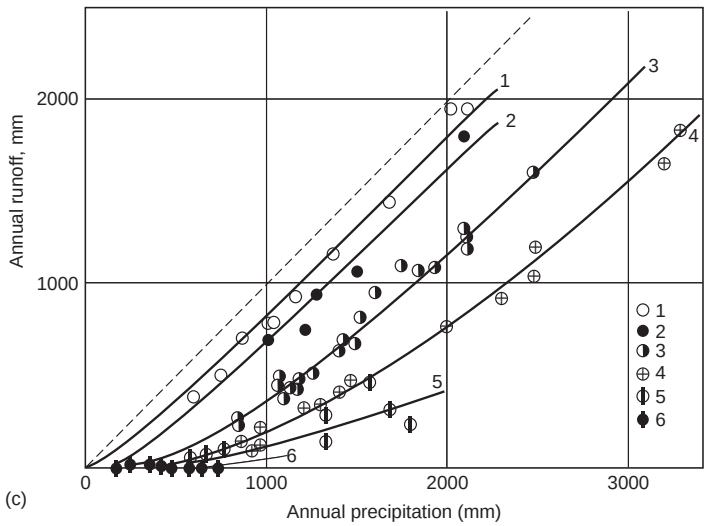
In order to develop coherent theories to underpin empirical data analysis, we need *signatures* of variability that are physically meaningful, and also useful in a practical context (*see*, **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**). Focusing on streamflow response, some of the commonly used measures of hydrological variability include interannual (between-year) variability of annual streamflow, and intraannual (within-year) variabilities, such as mean monthly variation (i.e. regime curve), the flow duration curve and annual rainfall-runoff relationships. These describe the character of temporal streamflow variability in one catchment, examples of which are presented in Figure 2(a). Other measures of temporal variabilities include the flood frequency curve, and the low flow (drought) distribution. In hydrology we are also interested in between-catchment variabilities of the measures listed earlier in the text, either within the same region, or between different hydroclimatic regions. Figure 2(b) presents the



(a)



(b)



(c)

Figure 2 Signatures of hydrological variability: (a) interannual, intraannual (mean monthly variation and flow duration curve); (b) mean annual water balance as a function of the climatic dryness index, E_p/P: crosses refer to locations in Australia and New Zealand; and (c) geographical variation of annual runoff versus annual precipitation for South and South-East Asia: the numbers refer to different vegetation types (Reproduced from L'vovich, 1979 by permission of American Geophysical Union)

geographical variation of E/P, the ratio of mean annual actual evaporation to mean annual precipitation, as a function of the ratio of the mean annual potential evaporation (surrogate for net radiant energy) and annual precipitation, E_p/P. This is known as *the Budyko curve* (Budyko, 1974), and shows that climate, as exemplified by E_p/P, is a good first-order predictor of annual water balance. Geographical variations of the relationship between annual runoff and annual precipitation are presented in Figure 2(c), taken from L'vovich (1979), with the differences between different regions attributed to differences in climate (seasonality, storminess etc.), soils and vegetation, including the way that native vegetation adapts to water stress, such as leaf shedding and deep rooting. Other measures of spatial variability include scaling behavior of flood frequency curves, with respect to the size of catchments in the same region. Analysis of data from around the world has thrown up interesting patterns in many of these signatures. For example, L'vovich (1979) and McMahon *et al.* (1992) have presented a compendium of interhemispherical and interregional comparisons of interannual variability of annual runoff volumes, intraannual variations of streamflow (the regime and flow duration curves), and annual maximum flood peaks. There is tremendous value in exploring their underlying process controls, which will assist in developing a coherent new theory of hydrology.

Other signatures, besides measures of streamflow variability, include distributions of residence time or water age, temperature, isotopic composition, concentrations of tracer chemicals such as chloride or nitrate, although measurements of these are not as widespread as streamflow. Some of these chemical signatures may be strongly distinctive, and diagnostic of important differences between catchments and need to be predicted correctly. Patterns of vegetation cover, in both space and time, can also be excellent indicators of hydrological variability since they provide a window into the underlying water balance (Boer, 1999; Boer and Puigdefábregas, 2005), although, in the past, they have only been utilized as prescribed inputs to hydrological models. In a similar vein, patterns of other hydrological response data such as soil moisture and snow cover have been shown to provide powerful indicators to the understanding of catchment behavior (Grayson and Blöschl, 2000).

Predictors of Runoff Variability

Examples of the *predictors* of catchment responses include, but are not limited to:

- Climate: aridity/humidity, seasonality, especially, the relative seasonality of precipitation and potential evaporation, measures of storminess, ratio of interstorm period to storm duration, nature of within-storm variability of rainfall intensity,and so on;
- Catchment area and shape, drainage density;
- River network: length and shape of channel network;
- Soil properties: soil depth, soil texture (well drained or poorly drained soils, saturated hydraulic conductivity etc.), and their spatial distributions;
- Geology: fractured or monolithic rock, its influence on the subsurface hydrogeology, layering, relationship to topography, and so on;
- Topography: steepness (surface and stream slopes), mean elevation, curvature;
- Vegetation: type and density, spatial patterns, and temporal variability.

At the present time, our ability to infer or learn from observations through systematic data analysis is not well advanced. Progress in developing robust, quantitative relationships between the signatures of hydrological variability described above and the predictor variables (various climate parameters and landscape properties), toward the development of a general and reproducible theory at the catchment scale, has been hampered for a number of reasons. Firstly, the signatures of variability presented above can be thought of as reflecting processes occurring at or below the catchment scale, and providing a window into the interactions and feedbacks between different processes and between the various constituent elements of the catchment. To date, most of these signatures are not yet understood, and have not been explored in terms of the underlying processes. One consequence of this is that, as yet, we are not even able to choose predictor variables with clear, causal connections to the signatures.

Secondly, a prerequisite for making inferences from observations is the availability of a physically meaningful classification system, which can be used to guide empirical data analysis, to organize the data in such a way as to elicit interesting and useful patterns. Such a classification system, and a theory of inference based on the analysis of patterns in the observed data, is almost nonexistent in catchment hydrology at the present time (McDonnell and Woods, 2004; Woods, 2002). Dooge (1986) has suggested that hydrology is in the same position of confusion as the field of hydraulics had been before the Reynolds and Froude numbers were proposed; almost two decades on, there has been no real advance in this direction, notwithstanding the work of Milly (1994).

Much of the data analysis that is presently carried out is model focused, for example, during the calibration of models based on small-scale process theories, constrained by assumptions about processes upon which the models are based. This may explain why data analyses in the past have not been very revealing, and possibly even why there are too many models. Indeed, it can be said that the increasing sophistication of models based on small-scale process-based theories and the increasing power of computers may even have contributed to a neglect of systematic and thoughtful data analyses, and the role of data has been relegated to its use in model calibration only.

Current State of Theories Regarding Organization

The role of organization in catchment response has been widely acknowledged over the years (e.g. Blöschl *et al.*, 1993; Blöschl and Sivapalan, 1995). At the most fundamental level, climate acts as the unifying global force in the coevolution of landscapes and vegetation. An illustration of this, is the fact that the world's broad vegetation classes can be predicted by the combination of just two climatic variables: temperature and rainfall (Figure 3a, Shuttleworth, 1983; Woodward, 1987). Observed precipitation patterns demonstrate space-time variability over a wide range of scales, including but not limited to such definable units as cells, small mesoscale areas, large mesoscale areas,

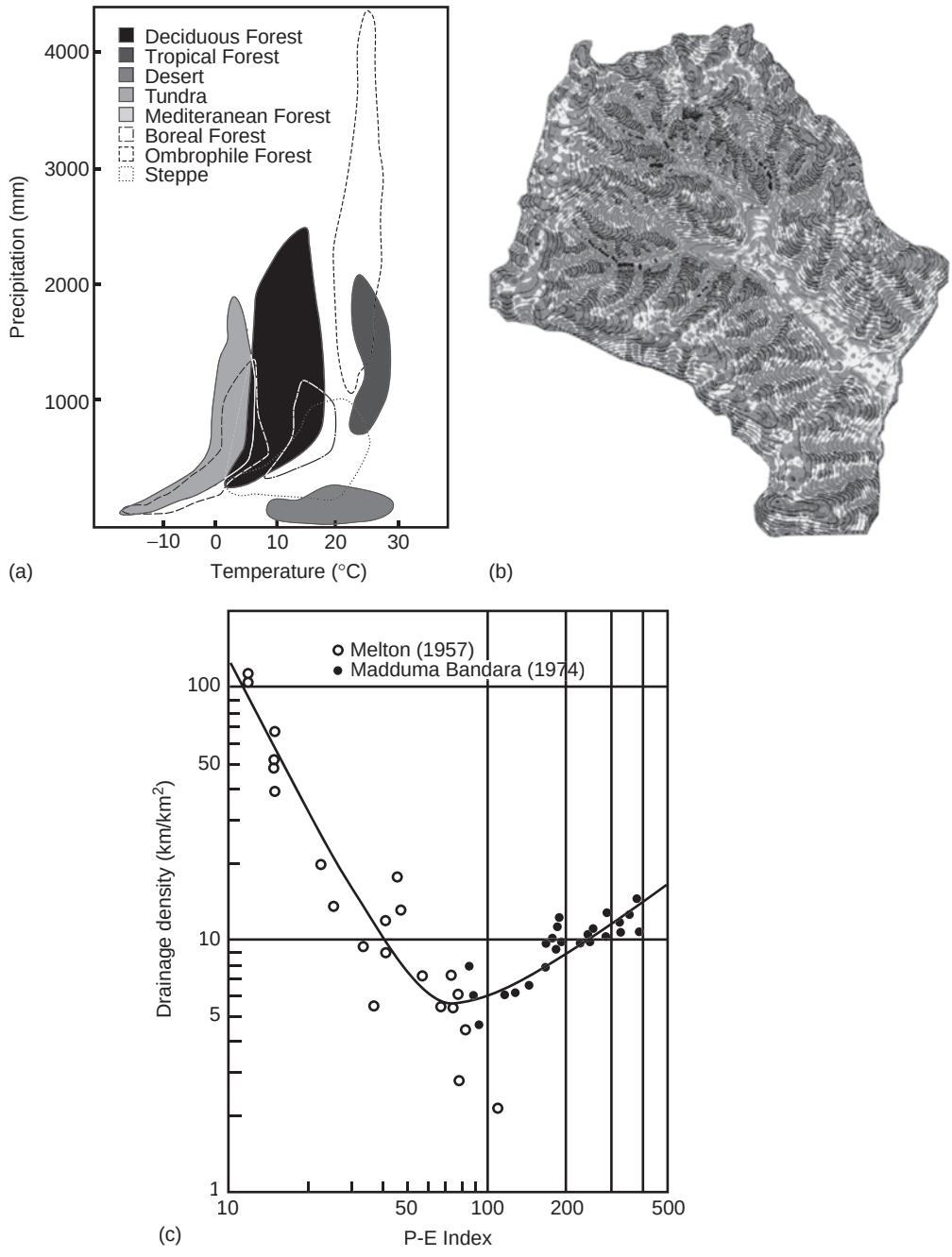


Figure 3 Patterns of landscape properties due to self-organization: (a) climatic influence on major vegetation zones of the world (from Shuttleworth, 1983); (b) soil catena – simulated patterns of soil depth within a catchment (from Dietrich *et al.,* 1995); (c) global relationship between drainage density of landforms against the precipitation–evaporation (P-E) index (Reproduced from Abrahams, 1984 by permission of American Geophysical Union)

synoptic areas, and so on as they develop, mature, move, and dissipate. Rivers carve the landscape into intricate shapes called *river networks*, which appear to embody a deep sense of symmetry (Rodriguez-Iturbe and Rinaldo, 2000). Soils tend to develop in response to state controls such as topography, with different parts of a basin (nose, slope, hollow) being formed by different processes and having different functions (e.g. concentration, storage, and evapotranspiration of water). The resulting soil patterns (soil catena) exhibit a common form of organization and symmetry, with water movement clearly being an active agent in their formation (Figure 3b, Dietrich *et al.*, 1995). Drainage densities observed around the world in different hydroclimatic regions demonstrate a robust relationship with the so-called precipitation–evaporation (P–E) index, the difference between annual precipitation and evaporation (Abrahams, 1984), the particular U-shape of this relation, and its minima (Figure 3c) being caused by the armoring imparted to the soil by the presence of vegetation roots.

The interactions between climate, soils, vegetation, and topography thus contribute to the generation of the interesting patterns that we see in natural catchments, which must contain valuable information about the way they function. Therefore, a fundamental aim of theoretical hydrology must be to recognize these patterns, decipher the underlying order or symmetry occurring over a wide range of scales, and explore the mechanisms that may have generated them. The field of hydrogeomorphology is attempting to discover the mechanisms underlying the order or symmetry in terms of quantitative measures of drainage network composition (Rodriguez-Iturbe and Rinaldo, 2000). In a similar vein, theories and associated models of soil formation, shallow landsliding, and erosion have been pursued in order to explain the observed patterns in soil properties (Jenny, 1941; Willgoose *et al.*, 1991; Dietrich *et al.*, 1995). On the other hand, the field of ecohydrology is attempting to discover rules or organizing principles governing spatial patterns of vegetation density and type, linking these to underlying water and energy balances, consistent with a Darwinian natural selection process that is optimal for growth and reproduction within the prevailing climate and geology (Rodriguez-Iturbe *et al.*, 1999; Eagleson, 2002). However, to complete the development of a theory of catchment hydrology, current understanding of self-organized patterns in landscape properties must be extended to produce insights into the interactions and feedbacks between hydrological processes at the catchment scale. Progress would be made, for example, when such natural self-organization in landscape properties could be confirmed to be the cause of simple process descriptions or closure relations extracted from observations at the catchment scale (Savenije, 2001).

The most that has been done to accommodate the natural self-organization has been with respect to the choice of model structure, that is, the way the various components of the landscape are organized and interconnected, that underlies many current hydrological models. For example, a typical model structure may consider a catchment to consist of a population of hillslopes, of different sizes, shapes and steepnesses, wrapped around the stream network, which is its most distinctive element (Troch *et al.*, 2003). The channel hydraulic properties, collectively known as *hydraulic geometry* (*HG*), may be allowed to vary systematically with flow at a single site, as well as in the downstream direction. In the vertical direction, the catchment (and the associated hillslopes) is assumed to consist of the land surface, and the soil and bedrock beneath it, including any vegetation that is contained within it. The subsurface is further characterized by an unsaturated zone, underlain by one or more saturated zones, underlain by bedrock. In larger catchments, to account for large-scale spatial variations of climatic and landscape properties, the catchment may be divided into a number of subcatchments which are hierarchically organized around the stream network, before being further divided into hillslopes (Gupta and Waymire, 1998; Reggiani *et al.*, 1998). Recent work has suggested that the treatment of hillslopes as monolithic entities suppresses important functional variations that may occur within them, and has advocated their partitioning into upland and riparian zones (McGlynn and McDonnell, 2003). Other studies have advocated the introduction of a hyporheic zone between the riparian and the stream zones.

However, this kind of representation is essentially a static stratification of the landscape, and does not necessarily recognize or incorporate the dynamic mechanisms that sustain those subsystems. In particular, the processes that occur within these subsystems continue to be described in terms of small-scale process theories. The effects of the natural self-organization of soil properties, vegetation, and topography have not been embedded in the process conceptualizations that appear in most current hydrological models. Notable exceptions to this are the body of work that led to the development of TOPMODEL (Beven and Kirkby, 1979; Sivapalan *et al.*, 1987) and the related TOPOG model (O'Loughlin, 1986). In this case, the two models captured the effects of topographic convergence through the use of a topographic wetness index, along with assumptions made to characterize in a simple way the interactions between upslope and downslope regions of hillslopes, including the accumulation of water near the stream zone and the generation of dynamic saturation areas generating saturation excess runoff. The kind of spatial organization of soil moisture assumed in TOPMODEL has since been supported, to some extent, through a number of field studies at the small catchment scale (e.g. Western *et al.*, 1999), although spatial patterns have also been observed which are not consistent with TOPMODEL predictions. Another example of the use of organization in developing appropriate process

conceptualization is the now popular Xinanjiang or variable infiltration capacity model (Zhao *et al.*, 1980; Wood *et al.*, 1992; Liang *et al.*, 1994; Sivapalan *et al.*, 1997). In this case, the self-organization present within the catchment is expressed in terms of a statistical distribution of soil depths or infiltration capacity, which implicitly also accounts for the position on the hillslope. A third and final example is the development of the geomorphological instantaneous unit hydrograph (GIUH) of catchments on the basis of the organization present within the stream network in the form of, for example, Horton's order ratios (Rodriguez-Iturbe and Valdes, 1979; Rinaldo *et al.*, 1991; Snell and Sivapalan, 1994; Saco and Kumar, 2002). In this case, the dispersion imparted to the incoming rainfall by the stream network is quantified in terms of measures of the drainage network structure.

Status of Theories of Catchment Hydrology: Impasse!

The paradigm underpinning current hydrological theories at the catchment scale is shown in Figure 4, and is essentially *reductionist*. In this framework, the natural organization present within catchments is partly embraced through model structures that may reflect the presence of a stream network, a set of hillslopes or subcatchments organized around this network, distinct saturated and unsaturated zones, and a further possible partitioning of the hillslopes into upland, riparian, and hyporheic zones. However, the conceptualizations of hydrological processes are still dominated by small-scale process theories, unrelated and unconnected to the nature of self-organization present within the catchment. A set of balance equations at the scale of a representative catchment, and respecting the natural organization that is present within catchments has been presented recently (Reggiani *et al.*, 1998, 1999); the lack of appropriate catchment-scale closure relations to close this set of equations is hampering efforts to turn these into a new blueprint for distributed modeling at the catchment scale (Lee *et al.*, 2005; Zehe *et al.*, 2005). Even the best models based on current process theories are found to be inadequate to predict catchment responses since they demand complete knowledge of climatic inputs and landscape (soils and vegetation) characteristics, which is not routinely available. In fact, due to the strong heterogeneities of climatic landscape and climate properties, hydrology is replete with examples of highly complex behavior, including strong nonlinearities and threshold behavior, and paradoxes that defy causal explanation by models based on small-scale theories.

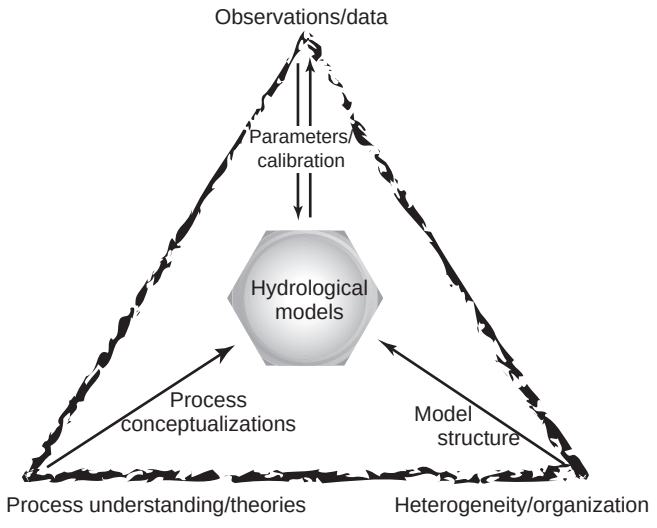


Figure 4 Current state of theory in catchment hydrology – reductionist with a reliance on calibration

Catchments being "poorly defined systems with some degree of organization", predictions of catchment responses must be conditioned or founded on empirical observations. Yet, in the current framework, the main role of observations and data appears to be, with a few exceptions, to assist in the calibration of models that are based on small-scale process theories, and *a priori* model structures, that is, "grist to the calibration mill" (Sivapalan, 1997). Indeed, hydrologists have not demonstrated the collective will, skill in experimental design, and the required clarity in posing scientific hypotheses and testing them on comprehensive datasets. The lack of holistic process theories and the inadequacies of current small-scale process theories, have meant that data collection to formulate alternative hypotheses has been limited and the analyses of existing datasets have not been so revealing. We do not yet have a sound quantitative framework, for example, a classification system, based on a set of predictor variables and governed by the understanding of underlying process controls, that can help us recognize interesting patterns in the data.

Because of the overparameterization in relation to the meagre datasets against which they are calibrated, current models suffer from the problem of equifinality (Beven, 1989a, 2001, 2002; Savenije, 2001), which expresses the fact that infinite combinations of parameters can give rise to model predictions that provide a good match to the observations. The lack of a holistic theory at the catchment scale has led to a plethora of alternative models that are overly complex, overparameterized, and uncertain. The end result is confusion not clarity, and stagnation and not real progress, in spite of the explosion of new knowledge related to various individual hydrological processes.

It is clear that catchment hydrology is trapped in a dead-end track, a theoretical impasse! We urgently require a new holistic and unified theory that overcomes the limitations of the current theories in dealing with processes, organization, and data analysis.

TOWARD A NEW UNIFIED HYDROLOGICAL THEORY AT THE CATCHMENT SCALE

Scope of a Unified Theory of Catchment Hydrology

In the discussions that follow, we will adopt a simple, but overarching definition of *theory* as the set of ideas or concepts that is best able to describe or explain the system of interest, the catchment, its presence in the landscape, its behavior, and its function in relation to other systems. In this context, theory is seen as more than the sum total of all knowledge, but as *distilled* knowledge, and, as knowledge that is causally interlinked, that is, every piece of knowledge must make sense with regard to all other pieces. Theory helps to connect the specific to the general, the local to the global, and the past to the future. Theory provides a framework to assess what we know and what we do not know. Theory provides the avenues to seek the knowledge that we do not possess.

Given the ubiquitous nature of hydrological variabilities at multiple space – time scales, the main role or purpose of a coherent hydrological theory is then to:

- help explain observed patterns of hydrological behavior over multiple space-timescales in terms of the underlying climate, soil, vegetation, and topography interactions, in this way providing a robust framework for a dialogue with nature;
- guide us to make appropriate measurements to further improve our understanding and our ability to generalize or extrapolate in space and time, and assist with the design of observational networks and/or focused field experiments; and
- guide us to make better predictions, into the future or to other points in space, that are based on *a priori* understanding and not just calibration, and in this way help establish the practice of hydrology on firm scientific foundations.

Given the nature and basis of hydrological science and its place at the center of a number of earth science disciplines, the desired hydrological theory at the catchment scale will combine ideas and concepts from:

- natural sciences such as physics, chemistry, and biology: examples include Newton's laws of motion, the 2nd law of thermodynamics, biological laws including theories of evolution, chemical laws of reaction and transformation;
- earth sciences, because of the overlap and interactions with other branches of earth system science: examples include theories of soil physics, micrometeorology, open channel hydraulics, geomorphology, ecology;
- empirical science: hydrology is fundamentally an empirical science, and depends crucially on inferences made from observations at all scales, all the way from laboratory to global;
- applied science: hydrology is also an applied science and elements of hydrological theory may derive from the sharing of experiences from its applications, for example, in agricultural hydrology, engineering hydrology, forest hydrology, and so on.

Hydrological theory will thus derive from the actual practice of the science, as natural science, empirical science, earth system science, and as applied science. Hydrological theory will not arise through mere speculation of the human mind, or, as Klemeŝ (1986) put it, "... the logic of hydrological processes cannot be deduced from algebra". By the same token, we cannot postpone the practice of the science of hydrology until a theory is ready.

Approaches to a New Unified Theory of Hydrology at the Catchment Scale

The observed patterns of variability of hydrological behavior at the catchment scale arise out of interactions between space-time variability of climatic inputs, for example, precipitation, solar radiation, atmospheric humidity, wind, and so on, the natural multiscale heterogeneity of landscape properties, such as the soils, vegetation, topography, and so on, and any alterations to these due to human impacts. The natural heterogeneities of the landscape properties, in turn, themselves arise through geomorphic (i.e. landforming), and ecological processes that occur over a much longer period of time, compared to the typical timescales of hydrological processes.

The new unified hydrological theory must ultimately consist of a set of organizing principles or natural laws governing:

- the ways that catchments are organized in space and time, in terms of their constituent landscape elements, including the geomorphic and ecological processes that may have led to them;
- the ways that catchments respond to climatic inputs and the nature of the interactions between the heterogeneities in the climatic inputs and the landscape properties;
- the resulting fundamental hydrological processes, their space-time variabilities, including the pathways, fluxes and stores of water, energy and other constituents, and the interactions between them;
- the way that the different constituent parts of the catchments, and the catchments as a whole, function, interact with, and feedback on each other; and
- the way that catchments respond to human-induced changes in the climate inputs and the landscape properties, in terms of both their form and function (e.g. storage of water, primary production etc.), in the short-term and in the long-term.

Considering the self-organized aspect of catchments, their constituent landscape elements, the processes that generated them in the first place, the way they interact and feedback on each other, and the resulting impact on hydrological processes at all scales, it is clear that feedbacks between pattern and process, must be the defining feature of any new, unified theory of catchment hydrology. It is also clear that the feedbacks between pattern and process are highly relevant in different contexts also, due to the critical role played by water in climatological, ecological, geomorphological, and pedological processes. Therefore, the ecological, geomorphological, and other "function" of the feedbacks between hydrological patterns and processes is essential for a deeper understanding of hydrological variability. For this reason, it is argued that pattern, process, and function must be the key elements of a new theory of hydrology at the catchment scale.

Pattern: instead of using observations and data for calibration of *a priori* constructed models, seek and identify patterns (both within-catchment and between-catchment) in the data or observations, to formulate and test hypotheses about processes, process interactions and feedbacks, including the mechanisms that contribute to natural self-organization;

Process: discover or explore new processes, process interactions, and feedbacks at all scales, and descriptions that embrace or embed within them explicitly or implicitly the effects of landscape and climatic heterogeneities, including any simplification that comes about due to the feedbacks and self-organization underlying these heterogeneities;

Function: investigate the processes that lead to the heterogeneities and self-organization exhibited by landscape properties, and explore the laws or organizing principles governing their ecological, geomorphological, or pedological "function", with the idea that these laws could act as constraints to both the process descriptions (within-catchment), and broad-scale patterns of behavior (between-catchment, regional etc.).

An example of an ecological function is the provision of physical habitat, or of food supply. The geomorphological function of watercourses may be the efficient movement of water and sediment. This function entails both conveyance and storage, which are critical to the healthy functioning of the stream. A stream may attain a shape, form, or pattern that permits the necessary movement of water and sediment with the energy available (i.e. slope). The landscape functioning may be related to maintaining the stability of the landscape, it may develop pipe flow as a mechanism for fast release of water to prevent too frequent landslides. It may develop a vegetation-soil association to keep erosion to a minimum.

By combining pattern, process, and function, the new theory will lead to process descriptions that respect patterns of observed behavior, it will lead to more parsimonious models with much-reduced parameterizations, it will encourage a scientific culture of learning from observations, instead of using them for calibration, and it will encourage the formulation of rigorous hypotheses to underpin future experimental campaigns and data collection exercises. By branching out to embrace organizing principles or natural laws from neighboring disciplines such as geomorphology, pedology, and ecology, it will also broaden and enrich the hydrological perspective.

Downward and Upward Approaches to Theory Development

Klemeš (1983) proposed two alternative approaches for pursuing the organizing principles or laws that might constitute the theory of hydrology at the catchment scale: the "upward or bottom-up approach" and the "downward or top-down approach", and the eventual reconciliation of the outcomes of these two approaches. Dooge (1986), in a similar vein, proposed parameterization of microscale effects (upward), and the search for general laws at the macroscale (downward) as alternative approaches to the discovery of hydrological laws at the catchment scale. Acknowledging the need for a reconciliation of these two approaches and considering the importance of scale and the adaptive or self-organized character of catchments, Dooge argued for the discovery and exploration of scaling laws in hydrological behavior as a third, alternative method that helps to find links across catchments of different sizes.

These ideas also resonate with recent developments in other related fields, as exemplified by the theoretical vision for earth system science proposed by Harte (2002). Harte considered systems having characteristics that apply equally well to catchments: poorly defined, unique, and continuously evolving; self-organized, characterized by strong feedbacks and interdependencies; requirement not just to characterize but also to generalize and extrapolate, so that the behavior in response to climate changes and/or land use changes can be predicted. Similar to Dooge, Harte proposed a theoretical framework that involves a combination of (i) simple falsifiable models, (ii) a search for patterns and laws, and (iii) the science of the place. Harte (2002) characterizes this theoretical framework as a synthesis of the Newtonian and Darwinian worldviews, combining "particularity and contingency, which characterize the ecological sciences, and generality and simplicity, which characterize the physical sciences".

Figure 5 illustrates the application of the downward and upward approaches, seeking connections between patterns and processes. Within-catchment investigations deal with specific catchments, and attempt to explain the temporal patterns of variability in terms of the underlying process

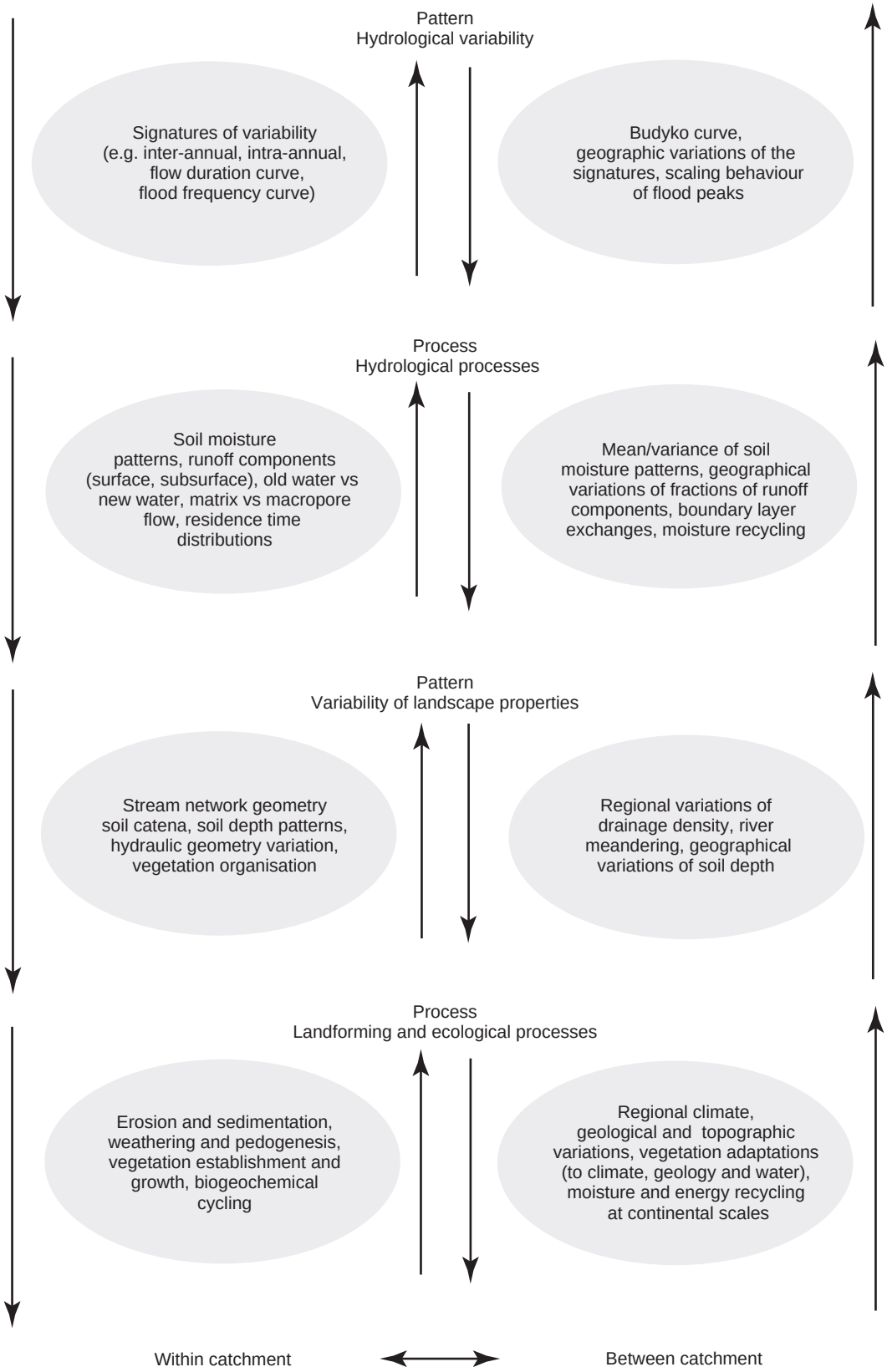


Figure 5 Downward and upward approaches to theory development in catchment hydrology – exchanges of knowledge and understanding at multiple levels

controls, and *vice versa*. In an analogous manner, between-catchment studies deal with spatial patterns, that is, differences between catchments located in the same region, or in different regions. Particular attention must be paid to scaling behavior exhibited by hydrological variables, both in the space and time domains.

Klemeš (1983) defines the downward approach as the "route that starts with trying to find a distinct conceptual node directly at the level of interest (or higher) and then looks for the steps that could have led to it from a lower level" (*see* also Sivapalan *et al.*, 2003a; and *see* **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3**). Along this downward path, we start at the patterns of observed hydrological variability at the catchment scale, and characterize these using appropriate measures of variability, as outlined before. The role of hydrological theory then is to explore the underlying process controls behind these patterns. At the next level, the approach will first involve observing hydrological processes at multiple space scales, not just the catchment scale, but both smaller and larger scales, and characterizing their variability through appropriate measures. The next step is to interpret the observed process variabilities in terms of the underlying heterogeneities in landscape properties, for example, soils, vegetation, topography, network geomorphology, HG and so on, and climate inputs. At the final, deeper level, the approach is to first characterize the natural spatial heterogeneities in the landscape properties, and quantify the heterogeneities through appropriate quantitative measures. The role of hydrological theory then is to explore the underlying, that is, hydrological, landforming, and ecological, process controls.

On the other hand, Klemeš (1983) defines the upward approach as "the route that attempts to combine, by mathematical synthesis, the empirical facts and theoretical knowledge available at a lower level of scale, into theories capable of predicting events to be expected at a higher, in our case hydrological, level". Along the upward path, we start at the deeper level, utilize existing knowledge about hydrological, landforming, and ecological processes toward development of process-based models capable of generating realistic patterns of landscape heterogeneities and validate these against observed patterns. At the next higher level, the approach will characterize the spatial heterogeneities in the landscape and climate properties, for example, soils, vegetation, topography, HG, network geomorphology and rainfall, and combine these with available small-scale process theories. Theory development will proceed by testing the predictions of the models against any observations of hydrological processes at the catchment scale. At the final level, the approach will observe and characterize hydrological processes at a wide range of scales, and aggregate these to generate patterns of hydrological variability at the catchment scale. The hydrological theory will evolve through testing or matching these predictions against patterns of behavior observed in the field.

Reconciliation Between the Upward and Downward Approaches

It is clear that the development of the catchment-scale theory will involve, at each level, almost symmetrical exchanges, upward in scale as well as downward, of both knowledge and understanding between observed patterns and the underlying process controls on the one hand, and between observed processes and the underlying patterns, on the other. As indicated previously in Figure 5, in the downward direction, theory development involves:

1. knowing the observed patterns, testing hypotheses about alternative processes that may have led to them, and
2. knowing the observed processes, testing hypotheses about alternative patterns that may have contributed to them.

In the upward direction, theory development involves:

1. knowing the processes, learning through constraining the patterns that they produce to match observed patterns, and
2. knowing the patterns, constraining the processes that they generate to match those that are observed.

The methodologies associated with the application of the downward and upward approaches, and the role of data, are presented in Figure 6. The upward approach starts with complex process descriptions and patterns, and whittles away the complexity and/or heterogeneity by constraining the model predictions using observed patterns and processes (Dooge, 1986; Sivapalan, 2003b). When the upward approach is repeated in different catchments, the descriptions will therefore evolve from specific to general catchment behavior. The downward approach involves identifying patterns of behavior or global relationships at the larger scale, for example, catchment scale, and looking for the processes that may have produced them, trying to connect the identified patterns or global relationships ultimately to such factors as soils, vegetation, drainage networks, and rainfall patterns (Dooge, 1986). Typically, the downward approach will start with simple process descriptions or patterns, and gradually adds complexities, through learning from observed patterns and/or process complexities. As we add more details in this way, the resulting models and process descriptions will evolve from the general or universal behavior toward behavior of specific catchments.

These objectives of the upward and downward approaches to the development of a new theory of catchment hydrology bear remarkable resemblance to equivalent themes reflected in the fields of ecology. In a recent review, Levin (1992), an ecologist, suggested, *inter alia*, that:

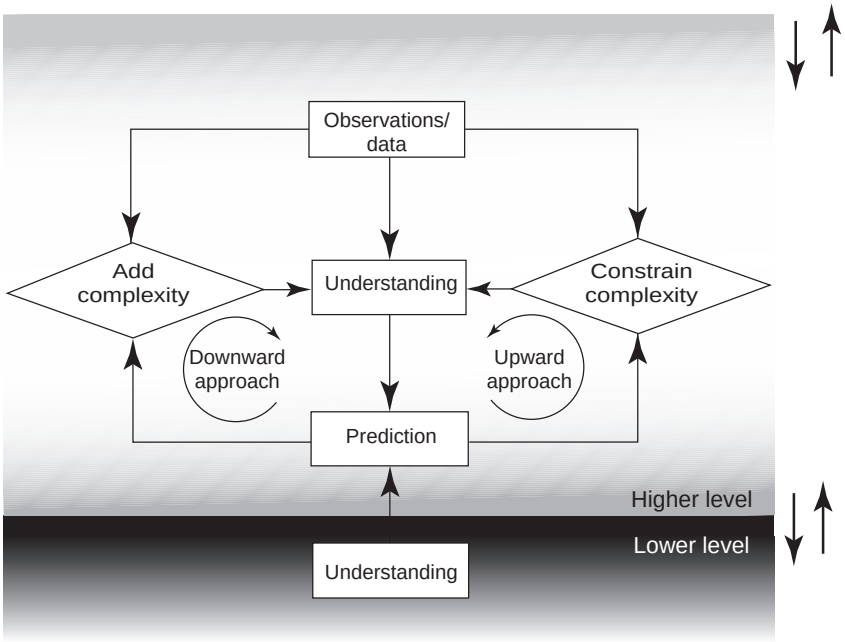


Figure 6 Methodologies of the downward and upward approaches to theory development. Adapted from Dooge (1986)

> "To scale from the leaf to the ecosystem to the landscape and beyond ... we must understand how information is transferred from fine scales, and vice versa. We must learn how to aggregate and simplify, retaining essential information without getting bogged down in unnecessary detail. The essence of modeling is, in fact, to facilitate the acquisition of this understanding, by abstracting and incorporating just enough detail to produce observed patterns... the objective of a model should be to ask how much detail can be ignored without producing results that contradict specific sets of observations, on particular scales of interest".

Clearly, there is much to learn from neighboring disciplines such as ecology, geomorphology, and pedology; the issues and the challenges are remarkably similar across the disciplines.

In fact, there is no reason for us not to use both the upward and downward approaches to generate and test *analogous* hypotheses regarding hydrological behavior at catchment scale. Independent application of the upward and downward approaches may, however, throw up conflicting outcomes, as illustrated schematically in Figure 7, which might still require reconciliation. One possibility is to altogether abandon any pretence to smaller scale process theories, and look for laws that may be sufficient to explain processes and/or patterns at the larger scale (Hatton *et al.*, 1997). Another possibility is that such conflicts or paradoxes might trigger further investigations leading to discoveries of new concepts or laws underpinning hydrological mechanisms that transcend multiple space-timescales, such as laws governing the ecological or geomorphological or pedological *function* of the catchment or parts of it, and in this way bring about the needed

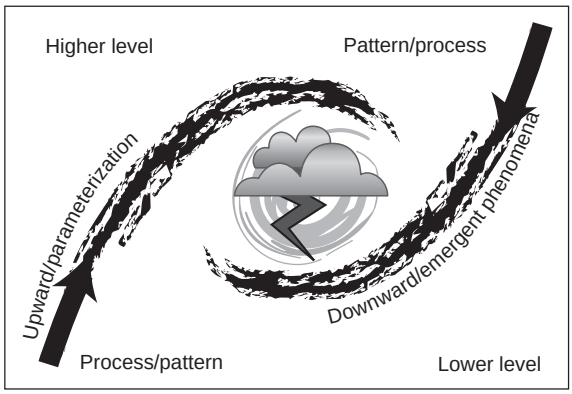


Figure 7 Reconciliation of downward and upward approaches – break or transition from a reliance on averaging and parameterization of lower level features to a culture of discovery and explanation of emergent phenomena at the higher level. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

reconciliation. Examples of organizing principles or natural laws that are currently debated in earth system science include, principle of minimum energy expenditure in geomorphological systems (Rodriguez-Iturbe and Rinaldo, 2000), ecological optimality in vegetation systems (Eagleson, 1978a,c,g, 1982, 2002; Cowan and Farquhar, 1977; Cowan, 1982; *see* **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**), maximum entropy production in climate change and climate-vegetation feedbacks (Ozawa *et al.*, 2003; Kleidon, 2004; Bejan, 2000), and

self-organized criticality in general complex systems (Bak, 1996; Rodriguez-Iturbe and Rinaldo, 2000; Hallet, 1990; De Boer, 2001). Any discoveries of this type can only come about through a broadening of the hydrological perspective to include multidisciplinary perspectives. These will also require or could be triggered through radically new observations that throw light on hitherto unknown mechanisms. Clearly, there is a lot of room for innovation and creativity; while the necessary conditions for breakthrough can be thus prescribed, success itself is *a priori* not predictable or guaranteed.

LEARNING FROM PATTERNS – METHODOLOGY OF THE DOWNWARD APPROACH, AND EXAMPLES

The methodology of the downward approach (Figure 6) may include the following steps: (i) identify an interesting pattern of behavior or an aspect of process observable at the catchment scale; (ii) devise alternative hypotheses, which may be potential explanations, or organizing principle, for the observed pattern or process characteristic; (iii) build the simplest possible predictive model based on this explanation or organizing principle; (iv) devise a numerical experiment (or several of them) with the simple numerical model, make a prediction of the pattern or process, and compare the predictions against data available at the higher scale; (v) on the basis of this test, either confirm or falsify the organizing principle or explanation as being correct or not; (vi) recycle the above steps, depending on the outcomes, making alternative hypotheses, or making further subhypotheses or sequential hypotheses to help resolve the possibilities that remain, developing a new numerical model and experiment to reflect the alternative hypothesis, or add more complexity to the previous model to reflect the subhypothesis or sequential hypothesis, and testing their predictions against the same data or additional data at the higher scale. Two examples are presented below as illustrations of the downward approach.

Pattern to Process: Spatial Scaling Behavior of Flood Frequency

The quantity of interest here is the annual maximum flood peak, and its dependence on catchment area. Empirical studies around the world have revealed that the annual flood peaks of a given return period T, Q_T, scale with catchment area, A, in terms of a relationship of the type $Q_T = cA^{\theta}$, with an exponent θ in the range between -0.10 and -0.40 (Jothityangkoon and Sivapalan, 2001). The power function relationship, and the exponent θ, can be seen as emergent behavior, that integrates information about complex rainfall–runoff–flood processes operating within a given region, how these change with increasing catchment size (Gupta and Dawdy, 1995; Robinson and Sivapalan, 1997a; Jothityangkoon and Sivapalan, 2001; Gupta, 2004), and about how the underlying process controls differ between different regions.

A number of studies have attempted to explore the physical basis of this scaling relationship, through the use of simple models that nevertheless captured the dominant process controls. Robinson and Sivapalan (1997a) approached this problem with the use of a simple rainfall-runoff model based on the unit hydrograph concept, with rainfall inputs that were scale-dependent, a mean catchment residence time also dependent on catchment area. With the use of this model, they showed that the interactions between two timescales, namely, rainfall duration and catchment response time, lay at the heart of the observed scaling relationship. Subsequent work by Robinson and Sivapalan (1997b) found that the observed apparent log–log linearity of the relationship between annual maximum flood peaks and catchment area is in fact caused by more complex interactions. At small catchment scales, within-storm patterns of rainfall variability interacted with the associated small mean resident times to increase the flow peaks. At large catchment scales, within-storm patterns were not important; instead, longer timescales in the rainfall field such as seasonality and the carry-over of storage between storms interacted with the longer residence times and again increased the magnitude of the flood peaks. Thus, the observed log–log linearity of the $E[Q_p]$ versus A relationship is a result of a "resonance" between the increasing catchment response time and the changing timescales associated with rainfall variability. Figure 8, adapted from Robinson and Sivapalan (1997b) illustrates this phenomenon. On the other hand, the spatial scaling of the rainfall intensities had a small but nevertheless significant contribution at all scales.

This then gives rise to a phenomenon which can be described as representing a *space to time connection* – an apparently simple spatial scaling behavior, an emergent property, being generated by complex interactions and feedbacks in the time domain between rainfall and runoff processes (Jothityangkoon and Sivapalan, 2001). The situation is further complicated when the relationship of the catchment response time to A depends on the relative dominance of hillslope and channel network in controlling the response time (Robinson *et al.*, 1995; Jothityangkoon and Sivapalan, 2001). Where the hillslope response time is dominant, θ approached zero, whereas in catchments where hillslope residence time is small, θ approached -0.40, an exponent in the relationship between channel length and A. Jothityangkoon and Sivapalan (2001) also demonstrated that the space-time connection and θ were also affected by the underlying long-term water balance regime, through its control of the antecedent soil moisture. Blöschl and Sivapalan (1997) found that the effects of resonance and the space-time connection, while important in a given hydrological setting, tend to be swamped by the other factors in larger, nonhomogeneous regions. They classified the

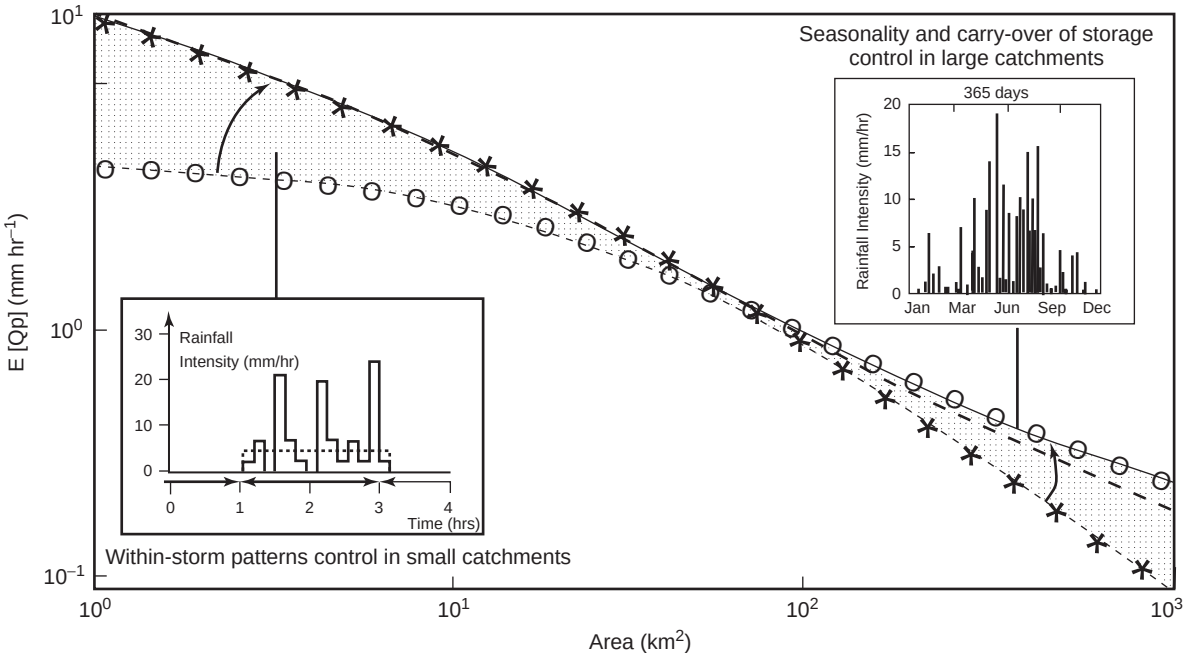


Figure 8 Process controls on scaling behavior of flood frequency. Power law relationship of mean annual flood with catchment area as an emergent property – complex interactions in the time domain leading to apparently simple pattern of behavior at the catchment scale

scaling behavior in Austria in terms of the differences in the underlying hydrological regime, which encapsulates the long-term water balance in a region.

A downward route for explaining the change of the flood frequency curve with catchment scale has also been taken by Merz and Blöschl (2003). They classified 12 000 flood peaks in Austria into long-rain floods, short-rain floods, flash-floods, rain-on-snow floods and snowmelt floods and then examined the flood statistics separately for each of the groups. They found that the coefficient of variation (CV) of the snowmelt flood type exhibited the flattest decrease with catchment area, which is consistent with the usually large extent of snowmelt. The CV of the flash-flood process type, however, tended to increase with catchment area, which was interpreted as being related to the nonlinearity of runoff generation associated with fast hillslope response.

These studies have demonstrated: (i) that apparently simple behavior may come about due to complex process interactions, and therefore the critical importance of these interactions; (ii) the power of simple models to elucidate the underlying process controls; and (iii) the use of these simple models to decipher broad-scale patterns and explain them in term of the underlying process controls.

Pattern to Process: Scaling of Hydraulic Geometry and Links to River Meandering

The dependence of channel hydraulic properties on streamflows have been known for a long time, and empirically described by the notion of HG introduced by Leopold and Maddock (1953). HG refers to the power law relationships relating the channel width, mean flow depth, and mean velocity to streamflow (discharge). These power law relationships have been observed to hold either for different discharges at a single cross section (called *at-a-station HG*), or for different downstream locations related through characteristic discharges having a constant frequency of occurrence (denoted as downstream *HG*). In their original work, Leopold and Maddock (1953) looked at HG in an *average* sense, ignoring the scattering around the proposed power laws.

Recently, Dodov and Foufoula-Georgiou (2004a) carried out extensive work on the *HG* relationships, paying particular attention to the scatter in the observed power laws. Through careful analysis, they showed that the exponents of the at-a-station *HG* systematically depended on catchment area, and that the exponents of downstream *HG* depended on the frequency. To quantify this empirical finding, they presented a lognormal multiscaling model, which was used to derive revised at-station HG whose coefficients are now explicit functions of catchment area. This generalized HG model was fitted to 85 gaging stations in Oklahoma and Kansas, and shown to reproduce the empirical trends extremely well. These revised HG relationships, being caused by streamflows in a self-organizing manner, therefore, represent an example of an emergent property, a phenomenon that only emerges at the catchment

scale, possibly due to complex process interactions and feedbacks.

Subsequent work by Dodov and Foufoula-Georgiou (2004b) set out to explore the underlying process controls of the empirically determined and statistically described scale-dependence of at-site *HG*. They presented an analysis of fluvial instability (Parker, 1976) as a function of catchment area, and showed that channel planform geometry (e.g. sinuosity, curvature, and wavelength) and, particularly, the transition between straight and meandering channels, are scale-dependent. To relate channel planform geometry and channel shape, they used the numerical river model of Johannesson and Parker (1989) to calculate the bed topography of representative meander bends of a given Strahler order, and subsequently the *HG* of these bends. In this way, Dodov and Foufoula-Georgiou (2004b) showed that the at-site *HG* that emerges from this physical model is scale-dependent, and agrees with the empirical trends and the proposed multiscaling statistical model. On the basis of these findings, they concluded that the scale-dependent *HG* is caused by the systematic increase of channel asymmetry downstream, induced by scale-dependent fluvial instability; an example of an apparently simple and useful relationship being brought out by complex process interactions at smaller scales.

LEARNING FROM MODELS – METHODOLOGY OF THE UPWARD APPROACH, AND AN EXAMPLE

The upward approach starts with the most complex model based on the most current or appropriate process descriptions. The objective then is to discover natural rules or organizing principles that may act to constrain the combinations of parameters or process interactions, into permissible ranges that are consistent with observed patterns in real catchments. The methodology of the upward approach, as presented in Figure 6, may include the following steps: (i) choose or identify the most detailed or appropriate process model for the problem and catchment of interest; (ii) devise alternative hypotheses regarding parameter combinations or process interactions; (iii) devise a numerical experiment (or several of them) with the chosen model, constrain the combination of model parameters or process interactions; (iv) compare the resulting model predictions against observations available at the higher scale; (v) on the basis of the above, confirm, or falsify the organizing principle as being correct or not; (vi) recycle the above steps, depending on the outcomes, making alternative hypotheses, or making further subhypotheses or sequential hypotheses to help define the possibilities that remain.

The use of models in this manner renders them "virtual reality models" (Weiler and McDonnell, 2004; Wood *et al.*, 2005), used to gain insights and generate hypotheses that can be tested with observations, and thus lead to gains in understanding. This is quite different from traditional calibration exercises aimed at choosing the parameter combination that produces the best match to observations, which are not meant to generate understanding. One example involving the upward approach is presented in the next section as an illustration of the method.

Process to Patterns: Climate-soil-vegetation Interactions and Ecological Optimality

The interactions between climate, soils, and vegetation in controlling a catchment's water balance and subsequently the drainage characteristics have been highlighted earlier through the Budyko (1974) curve (*see* Figure 2b), which suggests that in spite of differences in geology, soils, and vegetation, the annual water balance is governed, to a large extent, by climate, indicating that the soils and the vegetation that develop in a catchment are already adapted to the climate. Figure 3(c) from Abrahams (1984) was equally suggestive of the role of vegetation in controlling drainage density, through the role of soil armoring through vegetation roots. In other words, there are strong feedbacks between climate, soil and vegetation, through both the water balance and erosional stability.

Eagleson (1978a,b,c,d,e,f,g) carried out a pioneering study of climate-soil-vegetation controls on annual water balance, and feedbacks that develop between climate, soil and vegetation; his work represents the best example of the upward approach. For this, Eagleson utilized a comprehensive hydrological (water balance) model consisting of physically-based conservation equations governing each of the constituent processes: infiltration, exfiltration, transpiration, percolation to groundwater, and capillary rise from the water table. The climate inputs were intermittent rainfall events, separated by interstorm periods. Because all of the surface and subsurface fluxes depend on soil moisture content, which is time variable in response to the climatic inputs, an equilibrium soil moisture concentration s_0, which represents the spatially and temporally averaged state of the soil, was chosen as the state variable, and all fluxes were estimated in terms of the assumed equilibrium moisture content. By solving the resulting annual water balance equation, the unknown equilibrium moisture content s_0 was estimated as a function of the climatic variables and the soil and vegetation characteristics. The problem was cast within a statistical-dynamic framework by introducing the probability density functions of the climatic inputs, and Eagleson derived probability distribution functions of the annual water balance components: surface runoff, evaporation, and groundwater runoff.

These distribution functions expressed the mean water balance partitioning in terms of the independent climatic variables (rainfall intensity, duration, and interstorm period), and the soil hydraulic properties (mainly

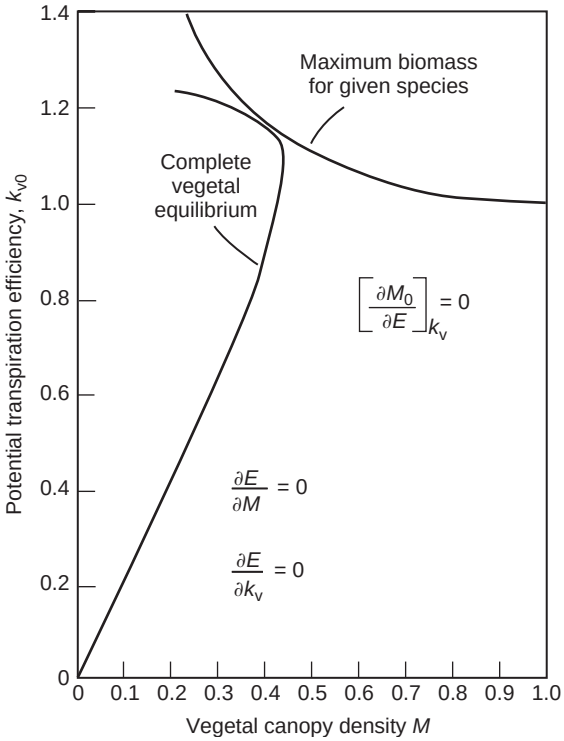


Figure 9 Ecological optimality constraints on climate-soil-vegetation interactions and the resulting water balance, and thus narrow down the space within which vegetation and soil parameters of a complex process-based model can vary (adapted from Eagleson, 1982)

the saturated hydraulic conductivity) and vegetation characteristics (density M, and species type k_v). In principle, according to the model, any combination of climate, soils, and vegetation is possible; apart from the climate inputs and process constraints, no constraints apply as yet to the soil-vegetation combinations that can be included in the model. In line with the methodology of the upward approach, Eagleson (1978d,e,f) invoked three constraints regarding the expected state of vegetation in a natural undisturbed ecosystem in an equilibrium state, which he called *the ecological optimality hypotheses* (*see* Figure 9).

Hypothesis 1: Over short timescales the vegetation canopy density, M, will equilibrate with the climate and soil parameters to minimize the water stress of the component plants, which is equivalent to a maximization of the equilibrium soil moisture, s_0.

Hypothesis 2: Over long timescales, species will be selected whose transpiration efficiency, k_v, maximizes the equilibrium soil moisture, s_0, which is equivalent to minimizing the total evapotranspiration, E.

Hypothesis 3: Over much longer timescales, vegetation will alter soil properties (saturated hydraulic conductivity, K_s, and the pore disconnectedness index, c, to maximize the optimal canopy density, M_0, derived from Hypothesis 1.

The invoking of these ecological constraints effectively limits the set of climate-soil-vegetation combinations to a small subset of the entire set of climate-soil-vegetation combinations that are possible in principle. Furthermore, through the invoking of the ecological optimality constraints, the theory can now be inverted; given the observed water balance in a given climate, the properties of the soil-vegetation system can be derived through an inversion procedure. Eagleson's ecological optimality hypotheses thus represent a predictive and testable theory (Hatton *et al.*, 1997; Kerkhoff *et al.*, 2004). Eagleson successfully tested his theory using observed data from catchments in different climates. In subsequent work, Eagleson (1982) argued that, with time, vegetation modifies (moderates) the hydraulic characteristics of parent material toward values which maximize vegetation production, that is, sandy soils become richer and more water-retentive; clayey soils become more porous and conductive. His ecological optimality hypotheses are capable of reproducing these observed features.

When confirmed on a wide variety of catchments, the ecological optimality hypotheses could constitute the elements of the new theory of hydrology. While recent analysis of the theory have cast doubts on the three

ecological optimality hypotheses (Kerkhoff *et al.*, 2004) based on ecological considerations, they do not invalidate the approach adopted.

SUMMARY AND FUTURE PERSPECTIVES

The foregoing review has highlighted the difficulties with the reductionist paradigm that underpins the current theory of hydrology at the catchment scale. Current process theories using the derivatives of mass, momentum and energy balances, along with empirically derived closure relations, such as Darcy's law and Manning's equation, are sufficient to capture only a small fraction of all the hydrological variability that occurs within catchments. Due to the uniformity assumption underlying the small-scale process theories, the resulting distributed models are highly complex, data hungry, and overparameterized compared to the meagre observations on which they can be calibrated. This reductionist framework and the focus on calibration leads hydrologists to study specific catchments in ever more detail, instead of exploring, and learning from, differences between catchments. In spite of the high level of model complexity and the inputs of enormous amounts of data on climatic variables and landscape properties, the models based on current theories still cannot reproduce some key or defining aspects of observed behavior; the old water paradox is just one example. Paradoxes like this abound in hydrology, and in the absence of a new theory of hydrology that can unify the different perspectives, different experiences, and observations made in different contexts, models are being made more and more complex to accommodate these differences. Current textbooks on hydrology propagate the same fragmented vision of hydrology, organized by process, and written in the form of recipes, for example, 10 different formulas for estimating infiltration, potential evaporation, and so on. The situation is literally analogous to "a cacophony of noises ... not a harmonious melody" (Sivapalan, 1997; Sivapalan *et al.*, 2003b). There is a clear and urgent need to develop a new, unified, and holistic theory of hydrology at the catchment scale that overcomes these limitations.

Elements of the New Unified Theory of Hydrology

On the basis of this review, it is clear that feedbacks will play a central, defining role in the new theory of hydrology at the catchment scale. These include feedbacks between different processes, between patterns and processes, between different parts of catchments, and feedbacks in time (through memory effects). With that focus on feedbacks, the new theory will include the following basic elements:

Pattern: The new theory will seek and identify patterns (both within-catchment and between-catchment) in the data or observations as part of the learning process, to formulate and test hypotheses about underlying processes, process interactions and feedbacks, including the mechanisms that contribute to natural self-organization. Increased attention will be given to structured learning from observations and data.

Process: The new theory will seek to discover or explore new processes, process interactions, or mechanisms at all scales that embed within them either explicitly or implicitly the effects of landscape and climatic heterogeneities, including any simplification that might come about due to natural self-organization that may underlie these heterogeneities.

Function: The new theory will investigate the processes that lead to heterogeneities and the natural self-organization exhibited by landscape properties, and explore the laws or organizing principles governing their ecological function, with the idea that these laws will act as constraints on both the process descriptions (within-catchment), and broad-scale patterns of behavior (between-catchment, regional etc.).

Holistic: The new theory will be holistic, treating pattern, process, and function as parts of a whole continuum – processes lead to patterns, which in turn lead to other processes, with the interactions and feedbacks between pattern and process being mediated by "function", a seemless transition between these three elements, as illustrated schematically in Figure 10.

Multiscale: The new theory will accommodate heterogeneities and variabilities of the catchment system and its responses over multiple space and timescales. It is not limited to making connections between just two scales, a small scale (microscale) and a larger scale (macroscale), but to

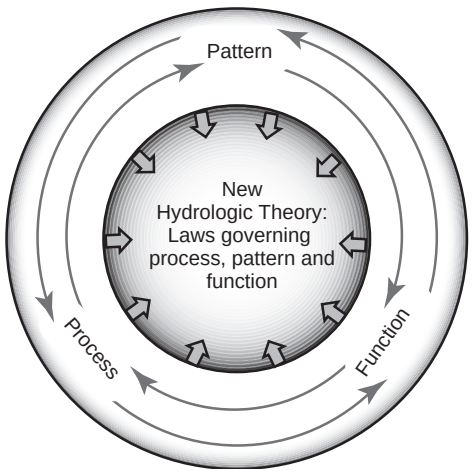


Figure 10 Pattern, process and function, feeding back on each other – elements of a new holistic theory of hydrology at the catchment scale

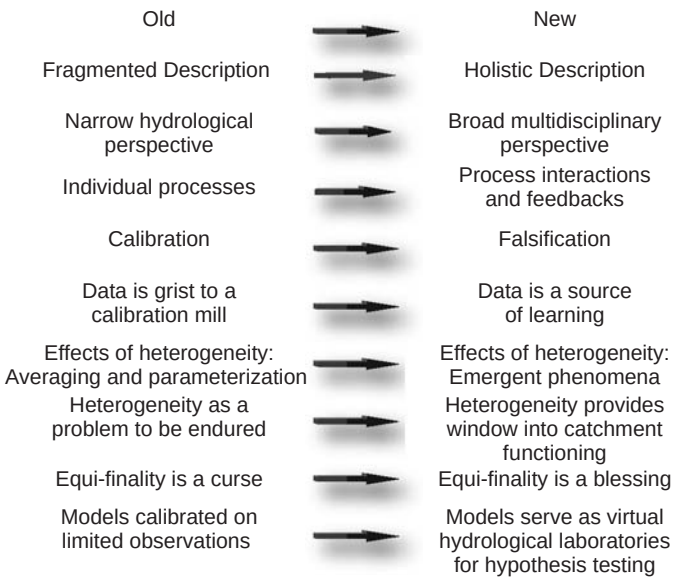


Figure 11 Paradigm shifts accompanying new theory of hydrology at the catchment scale

identifying and quantifying patterns that span a wide range of scales, and to exploring mechanistic explanations for and the common concepts behind these patterns.

Multidisciplinary: By broadening the nature of scientific inquiry to deal with "function", the new theory will necessarily deal with issues and questions that are at the interface between hydrology and neighboring disciplines such as geomorphology, ecology, pedology, and climatology. In this way, it will necessarily involve a broadening of the hydrologic perspective to include multidisciplinary perspectives.

Combination of "place-based" and comparative studies: Most studies exploring processes must be conducted in actual catchments, in real places. Studies of pattern require observations in many catchments, in the same region, and/or in different regions. Exploration of function underlying pattern and process will therefore require a combination of both place-based and comparative studies.

The development of a new theory of hydrology will therefore require nothing short of fundamental or paradigmatic shifts in both research and practice, as illustrated in Figure 11.

Process of Theory Development and Needed Infrastructure

Exploration of Puzzles and Paradoxes

A new theory will come about through answering specific questions regarding catchment behavior in real places, and solving what appear to be puzzles or paradoxes, that is, through a "dialogue with nature". Hydrologists have become very proficient at solving the "what" and "how" questions, but a new theory requires the answering of broader "why" type of questions related to "function". Examples of some unsolved puzzles or paradoxes include, but are by no means limited to the following questions.

- The old water paradox (Kirchner, 2003): How do catchments store "old" water for long periods, but then release it rapidly during storm events, and vary its chemistry according to the flow regime?
- Nature is replete with preferential flow at many scales, ranging from fingering, macropores, fractures, rills and gullies, all the way to the river network. What is the ecological or landscape function of these preferred pathways? Can we predict their occurrence and their spatial densities in terms of the underlying climate and geology alone?
- Why do landscapes evolve under the movement of water into intricate shapes, for example, river networks, soil catena, a signature of which is also present in vegetation and soil moisture patterns? What are the rules underlying their natural symmetry? What is their function?
- How does natural vegetation evolve and adapt itself to limitations of water, energy, and nutrients? What are the underlying organizing principles? Can the natural vegetation pattern and its functioning be predicted on the basis of climate, soils, and the water balance? Can the observed vegetation pattern in space and time give us clues to the underlying water balance?
- How does vegetation and other biotic elements generate and modify the soils to maximize their own ecological functioning? What are the underlying organizing principles?

Hydrological Infrastructure

Explorations of these puzzles and paradoxes will require sustained, painstaking work by individual hydrologists, perhaps working in small groups (but not in committees!). Nevertheless, this kind work will benefit from the presence of a supportive infrastructure, which helps to multiply and link the work of individual hydrologists. Infrastructure and organization focused on observations, new measurement technologies, and advances in modeling capability, all aimed at predictions in ungauged catchments worldwide will help advance fundamental theory development, as illustrated schematically in Figure 12.

Hydrological observatories: A number of highly focused, and detailed field experiments must be carefully designed and carried out in different regions of the world, in *nested* fashion, in order to observe the multiscale hydrological processes at the plot, hillslope and basin scales, assemble the necessary internal and surrogate data needed to make inferences about the underlying mechanisms over a wide range of scales, and their connections to landscape and climatic heterogeneities. By definition, these have to be established, and maintained over of long periods of time; this can only be done at the regional, national, or international level.

New measurement technologies: One of the difficulties that has hampered the development of a unified theory of hydrology has been the inability to observe processes over a wide range of scales, and at the same time, monitor internal variables such as soil moisture storage, groundwater levels, saturation areas, and so on, so that catchment water balance can be closed with confidence. Inadequate data resolution leads to incomplete or even wrong theories, and partly explains the stagnation in the development of a unified theory. There is a crying need for breakthroughs in measurement technology, on the ground as well as through remote sensing (e.g. radar, satellite, geophysical instruments, tracers). Along with new sources and types of data, we also need a revolution in techniques for data processing, and for identifying patterns in the data, extending beyond the present standard methods, such as cluster or principal component analysis, fractal analysis, artificial neural networks, and so on.

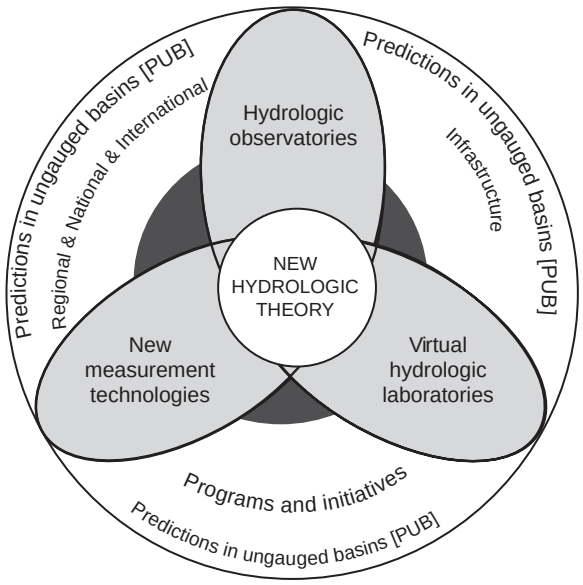


Figure 12 National and international initiatives provide the motivation and urgency toward theoretical advances in catchment hydrology

Virtual hydrological laboratories: In hydrology, we already have access to some of the most detailed, distributed, physically-based hydrological, hydrometeorological, eco-hydrological and landform evolution models, containing within them the most up-to-date process descriptions currently available. There is much to gain from their use as "virtual reality models", by implementing them to test alternative hypotheses about the constraints that need to be imposed on the underlying heterogeneities of landscape properties to match observed patterns of behavior. This kind of intellectual activity can play a critical role in the development of a new theory of hydrology, and along the way it will lead to improvements of the models for predictive purposes because of the possibility that such constraints will lead to a simplification of the models.

Methodological Framework

Two approaches, upward and downward, and their reconciliation, were outlined for the development of a new theory. Along the downward route, we begin with observed patterns at the catchment scale and explore the underlying process controls. Along the upward route, we combine existing knowledge about hydrological, landforming, and ecological processes in complex hydrological models, and explore their ability to reproduce observed patterns. Both methods, systematic data analysis and modeling investigations, require organization and a commonality of purpose. In this respect, we include three community level activities that are crucial to the success of the theory development effort.

Catchment classification: The downward approach involves, essentially, making inferences from patterns in the observations. To make progress, in order to decipher meaningful patterns in the observations, hydrology requires a globally agreed classification system capable of predicting, to first order, the dominant controls on water fluxes and pathways from amongst the entire range of mechanisms that are possible. The classification system, a prerequisite for any attempt at theory development, must be established urgently as part of a broad community effort, even though it will evolve with time as the field matures (Dooge, 1986).

Such a classification system would itself provide an important organizing principle, complementing the concept of the hydrological cycle and the principle of mass conservation (McDonnell and Woods, 2004; Woods, 2002). To provide meaningful distinctions between catchments, the classification scheme should be based on characteristic measures of fluxes, storages, and response times: a measure of climate dryness such as the ratio of mean annual potential evaporation to rainfall, E_p/P; the average volumes of water stored in different compartments, that is, snow and glaciers, pore water (in soils, and rocks), and in open water (lakes, wetlands, river channels); and characteristic response times of these catchment stores (Skøien *et al.*, 2003).

New balance equations for nested river basins: To make progress along the upward route to conceptualization, the new theory of hydrology must embrace a quantitative framework that is able to connect components of the hydrological cycle and the catchment system across multiple space-timescales. A primary requirement of this framework is that it be distributed, explicitly respecting the self-organized river network structure that connects different parts of the catchment, including the organized and random heterogeneities that develop within it (e.g. topography, soil catena, vegetation). This quantitative framework should be in the form of a set of coupled balance equations for mass (water), momentum and energy, that is consistent with the organized heterogeneities that are present within the catchment. The most obvious natural building blocks for the derivation of the balance laws are a nested set of subcatchments associated with individual stream links, although other building blocks may also be considered (Troch *et al.*, 2003). Gupta and Waymire (1998) presented a coupled set of mass balance equations for flow within a stream network, and for runoff processes within associated subcatchments, all expressed in the form of a set of hierarchical (nested) difference equations. Reggiani *et al.* (1998, 1999) derived a more complete set of coupled mass, momentum, and energy balance equations for the stream network and associated subcatchments, which they called *the representative elementary watersheds* (*REWs*). These are indeed significant advances toward a consistent quantitative framework that is required for catchment hydrology. Nevertheless, the balance equations derived in these studies must still be supplemented with numerous closure relations for various exchange fluxes (Lee *et al.*, 2005). The determination of such closure relations, reflecting the natural self-organization that is present within catchments, is a significant and as yet incomplete element of the new theory of hydrology (Zehe *et al.*, 2005).

Scaling laws in hydrological behavior: Observed patterns of heterogeneity of climatic inputs landscape elements and hydrological behavior are based on information from disparate sources: experimental plots, field surveys, weather radar reflectivities, and satellite data. Increasingly, much information relevant to hydrology is able to be extracted from satellite remote sensing, including aspects such as soil moisture and vegetation patterns. Coarser, larger-scale field surveys may reflect processes occurring over longer timescales. For example, tree rings and paleobiological and paleohydrological data may unearth information about past changes over very long timescales (Harte, 2002). If simple, robust scaling patterns exist among these variabilities, they will let us connect and extend insights between different scales. Fractals, multifractals, and random cascades are modern stochastic techniques, which exploit concepts such as geometric and statistical self-similarity to quantify the relationship between the variabilities present at different scales, and have the power to describe apparently complex forms of heterogeneity occurring over a wide range of scales with a small number of parameters (Lovejoy and Schertzer, 1995). Multifractal concepts have already been used successfully to characterize space-time variability of rainfall fields and the structure of stream networks, and are already beginning to be used to describe spatial patterns of soils and vegetation, soil moisture, and a number of other geophysical and biophysical phenomena. The search for scaling laws is a key component of the development of a new theory of hydrology (Dooge, 1986; Levin, 1992; Harte, 2002; West and Brown, 2004), and must happen independently of the other two approaches.

Predictions in Ungauged Basins (PUB)

The development of a new theory of hydrology receives additional impetus due to the urgent need to predict catchment behavior, for the sustainable management of water resources and water quality, and the prevention and amelioration of water-related natural hazards such as floods and droughts. The International Association of Hydrological Sciences (IAHS) has launched the IAHS Decade on Predictions in Ungauged Basins (2003–2012), or prediction of ungauged basins (PUB), a new global initiative, aimed at formulating and implementing appropriate science programs to engage and energize the scientific community toward achieving major advances in the capacity to make predictions in ungauged basins (Sivapalan *et al.*, 2003b). PUB emphasizes (i) improved understanding of multiscale variabilities of hydrological behavior at catchment scale, (ii) increased use of advanced technologies, and (iii) development and application of sophisticated numerical models that depend less on calibration and more on understanding. The urgency engendered by PUB forces us to challenge and critically evaluate existing approaches to making hydrological predictions. In addition, a number of other ongoing national and international initiatives act as strong catalysts toward triggering significant theoretical breakthroughs (*see* **Chapter 203, A Guide to International Hydrologic Science Programs, Volume 5**). Parallel national programs

are needed in many individual countries and in regions to provide the necessary infrastructure and funding for the associated research activities.

Current Intellectual Environment Surrounding Hydrology

Three contemporary scientific movements are providing a supportive environment for the theoretical advances in catchment hydrology, providing coherence and respectability to the efforts in this direction, as depicted schematically in Figure 13: (i) global change science, which provides support toward comparative hydrology on a global scale, helping to monitor large-scale patterns and understand the causes of these patterns; (ii) ecohydrology and earth system science, which help broaden the nature of hydrological inquiry, through exploration of process interactions and feedbacks, including interactions between hydrological, ecological, geomorphological, pedological, and climatological processes; and, (iii) complex systems science, an emerging interdisciplinary field spanning many fields, including mathematics, statistics, physics, ecology, and earth system science, which helps to generate new mathematical or analytical tools to deal with pattern dynamics and emergent phenomena that arise from nonlinear interactions and feedbacks in complex systems.

Apart from these scientific trends, the thrust toward a new theory of hydrology is also considerably aided by advances in technology: (i) vast improvements in our ability to measure and monitor hydrological parameters and state variables (and even fluxes through indirect methods) at large scales over the whole earth: these include satellite remote sensing, geophysical, and electronic measurements of increasing capability to plumb the subsurface hydrological condition, increasing sophistication of environmental isotope and chemical tracers to measure and monitor fluxes and state variables at the catchment scale; and (ii) increasing speed and storage capacity of digital computers that will enable hydrologists to run massively complex numerical models, in virtual reality mode, and improved methods of communication and sharing of data and results through the Internet which allows scientists to more easily cooperate and interact amongst large groups.

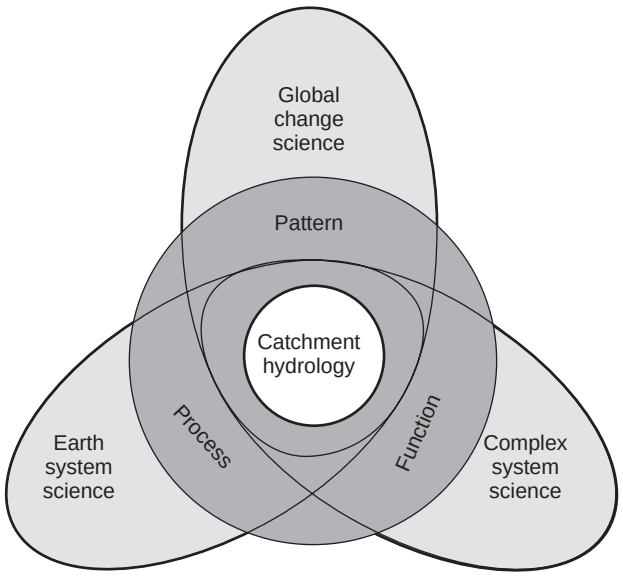


Figure 13 Current global scientific and intellectual environment providing a strong catalyst toward development of a new and holistic theory of hydrology at the catchment scale

Concluding Remarks

The urgency for a new, unified theory of catchment hydrology arises partly from the increasing frustration with the difficulties faced with existing hydrological models based on current theories, including the inability to satisfactorily resolve observed paradoxes. The urgent need to make satisfactory predictions of water quantity and water quality in ungauged basins, including the effects of climatic changes and human impacts requires a sound theory of hydrology that is solidly based on understanding and not calibration (Sivapalan *et al.*, 2003b). Therefore, the time is ripe for fundamental research that will set the stage for major advances in our predictive capabilities. Many national programs, and international initiatives such as PUB, are beginning to provide the necessary infrastructure, leadership, and coordination for groups of scientists to work together to address some of the most fundamental questions related to the development of a new theory. The increasing focus on earth system science, global change science, and complex system science is providing the necessary intellectual framework for a cross-fertilization of ideas across diverse disciplines, which can only benefit hydrology. For these reasons, there is real hope and excitement that catchment hydrology will leave behind the empiricism and fragmentation that has bedeviled it for so long, move forward toward a more unified and holistic theory that is fully accommodative of broad multidisciplinary perspectives, and will evolve from the "cacophony of noises to a harmonious melody" as anticipated in Figure 14. Hydrologists should rise up to these challenges and make use of the exciting opportunities that the pursuit of a new unified theory will generate.

Acknowledgments

Many of the ideas presented in this article are the result of discussions I have had with numerous colleagues and students over the past few years. As such, they are not just personal views and, to a large extent, they reflect the emerging views of many in the hydrology community. In particular, I am indebted to Günter Blöschl, Christoph Hinz,

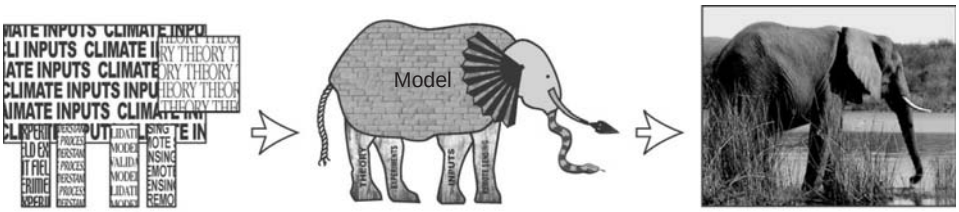


Figure 14 From a fragmented to a holistic view of catchment hydrology, as an integral part of earth system science (Reproduced from Sivapalan *et al.*, 2003b by permission of IAHS. Illustration of elephant reproduced with permission from Jason Hunt © 1999)

Gavan McGrath, Jeff McDonnell, Hubert Savenije, Stan Schymanski, Ross Woods, Matthias Boer, Greg Hancock, and Patricia Saco, for freely sharing their ideas with me, and for offering advice and many constructive criticisms, which have helped to substantially improve the article.

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986a) An introduction to the European Hydrological System–Systeme Hydrologique Europeen, "SHE", 1: history and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 45–59.

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986b) An introduction to the European Hydrological System–Systeme Hydrologique Europeen, "SHE", 2: structure of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

Abrahams A.D. (1984) Channel networks: a geomorphological perspective. *Water Resources Research*, **20**(2), 161–188.

Bak P. (1996) *How Nature Works, the Science of Self-Organised Criticality*, Springer-Verlag: New York, p. 212.

Bejan A. (2000) *Shape and Structure, From Engineering to Nature*, Cambridge University Press: Cambridge, p. 324.

Beven K.J. (1989a) Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (1989b) Interflow. In *Unsaturated Flow in Hydrologic Modeling*, Morel-Seytoux H.J. (Ed.), Kluwer: Amsterdam, pp. 191–219.

Beven K.J. (1993) Prophecy, reality and uncertainty in distributed hydrological modeling. *Advances in Water Resources*, **16**, 41–51.

Beven K.J. (2000a) On the future of distributed modeling in hydrology. *Hydrological Processes*, **14**, 3183–3184.

Beven K.J. (2000b) Uniqueness of place and the representation of hydrological processes. *Hydrology and Earth System Sciences*, **4**, 203–213.

Beven K.J. (2001) On landscape space to model space mapping. *Hydrological Processes*, **15**, 323–324.

Beven K.J. (2002) Towards a coherent philosophy for modelling the environment. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **458**, 1–20.

Beven K.J. and Kirkby M.J. (1979) A physically-based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Bishop K.H. (1991) *Episodic Increases in Stream Acidity, Catchment Flow Pathways and Hydrograph Separation*, Ph.D. dissertation, University of Cambridge, Cambridge.

Boer M.M. (1999) *Assessment of Dryland Degradation: Linking Theory and Practice Through Site Water Balance Modelling*, Netherlands Geographical Studies: Utrecht, p. 294.

Boer M.M. and Puigdefábregas J. (2005) Assessment of dryland condition using remotely sensed anomalies of vegetation index values. *International Journal of Remote Sensing*, in press.

Budyko M.I. (1974) *Climate and Life*, Academic Press: New York.

Buttle J.M. (1994) Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins. *Progress in Physical Geography*, **18**, 16–41.

Blöschl G., Gutknecht D., Grayson R.B., Sivapalan M. and Moore I.D. (1993) Organisation and randomness in catchments and the verification of hydrologic models. *EOS, Transactions of the American Geophysical Union*, **74**, 317.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling – a review. *Hydrological Processes*, **9**, 251–290.

Blöschl G. and Sivapalan M. (1997) Process controls on regional flood frequency. Coefficient of variation and basin scale. *Water Resources Research*, **33**(12), 2967–2980.

Cowan I.R. (1982) Regulation of water use in relation to carbon gain in higher plants. In *Physiological Plant Ecology II: Water Relations and Carbon Assimilation*, Lange O.L., Nobel P.S., Osmond C.B. and Zeigler H. (Eds.), Springer-Verlag: Berlin, pp. 589–613.

Cowan I.R. and Farquhar G.D. (1977) Stomatal function in relation to leaf metabolism and environment. In *Integration of Activity in the Higher Plant*, Jennings D.H. (Ed.), Society for Experimental Biology, Cambridge University Press: Cambridge, pp. 471–505.

De Boer D.H. (2001) Self-organization in fluvial landscapes: sediment dynamics as an emergent property. *Computational Geosciences*, **27**, 995–1003.

Dietrich W.E., Reiss R., Hsu M.-L. and Montgomery D.R. (1995) A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. *Hydrological Processes*, **9**, 383–400.

Dodov B. and Foufoula-Georgiou E. (2004a) Generalized hydraulic geometry: derivation based on a multiscaling formalism. *Water Resources Research*, **40**, W06302, doi:10.1029/2003WR002082.

Dodov B. and Foufoula-Georgiou E. (2004b) Generalized hydraulic geometry: insights based on fluvial instability analysis and a physical model. *Water Resources Research*, **40**, W12201, doi:10.1029/2004WR003196.

Dooge J.C.I. (1986) Looking for hydrologic laws. *Water Resources Research*, **22**(9), 46S–58S.

Dunne T. (1998) Wolman lecture: hydrologic science .... in landscapes .... on a planet ... in the future. *Hydrologic Sciences: Taking Stock and Looking Ahead*, National Academy Press: Washington, p. 138.

Eagleson P.S. (1978a) Climate, soil, and vegetation, 1. Introduction to water balance dynamics. *Water Resources Research*, **14**, 705–712.

Eagleson P.S. (1978b) Climate, soil, and vegetation, 2. The distribution of annual precipitation derived from observed storm sequences. *Water Resources Research*, **14**, 713–721.

Eagleson P.S. (1978c) Climate, soil, and vegetation, 3. A simplified model of soil moisture movement in the liquid phase. *Water Resources Research*, **14**, 722–730.

Eagleson P.S. (1978d) Climate, soil, and vegetation, 4. The expected value of annual evapotranspiration. *Water Resources Research*, **14**, 731–740.

Eagleson P.S. (1978e) Climate, soil, and vegetation, 5. A derived distribution of storm surface runoff. *Water Resources Research*, **14**, 741–748.

Eagleson P.S. (1978f) Climate, soil, and vegetation, 6. Dynamics of the annual water balance. *Water Resources Research*, **14**, 749–764.

Eagleson P.S. (1978g) Climate, soil, and vegetation, 7. A derived distribution of annual water yield. *Water Resources Research*, **14**, 765–776.

Eagleson P.S. (1982) Ecological optimality in water-limited natural soil-vegetation systems 1. Theory and hypothesis. *Water Resources Research*, **18**, 325–340.

Eagleson P.S. (2002) *Ecohydrology, A Darwinian Expression of Vegetation Form and Function*, Cambridge University Press: Cambridge, p. 443.

Grayson R.B. and Blöschl G. (Eds.) (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Cambridge University Press: Cambridge, p. 404.

Gupta V.K. (2000) *A Framework for Reassessment of Basic Research and Educational Priorities in Hydrologic Sciences*, A report to the U. S. National Science Foundation. `http://cires.colorado.edu/hydrology`

Gupta V.K. (2004) Emergence of statistical scaling in floods on channel networks from complex runoff dynamics. *Chaos, Solitons, and Fractals*, **19**(2), 357–365.

Gupta V.K. and Dawdy D.R. (1995) Physical interpretations of regional variations in the scaling exponents of flood quantiles. *Hydrological Processes*, **9**, 347–361.

Gupta V.K. and Waymire E. (1998) Spatial variability and scale invariance in hydrologic regionalization. In *Scale Dependence and Scale Invariance in Hydrology*, Sposito G. (Ed.), Cambridge University Press: Cambridge, pp. 88–135.

Hallet B. (1990) Spatial self-organization in geomorphology – from periodic bedforms and patterned-ground to scale-invariant topography. *Earth Science Reviews*, **29**, 57–75.

Harte J. (2002) Toward a synthesis of the Newtonian and Darwinian worldviews. *Physics Today*, **55**(10), 29–34.

Hatton T.J., Salvucci G.D. and Wu H.I. (1997) Eagleson's optimality theory of an ecohydrological equilibrium, quo vadis? *Functional Ecology*, **11**, 665–674.

Hunt J. (1999) `http://www.naturalchild.com/jason/blind_men_elephant.html`.

Jenny H. (1941) *Factors of Soil Formation*, *First Edition*, McGraw-Hill: New York.

Johannesson H. and Parker G. (1989) Linear theory of river meandering. In *River Meandering*, *AGU Water Resources Monograph 12*, Ikeda S. and Parker G. (Eds.), AGU: pp. 181–214.

Jothityangkoon C. and Sivapalan M. (2001) Temporal scales of rainfall-runoff processes and spatial scaling of flood peaks: space-time connection through catchment water balance. *Advances in Water Resources*, **24**(9–10), 1015–1036.

Kendall K.A., Shanley J.B. and McDonnell J.J. (1999) A hydrometric and geochemical approach to test the transmissivity feedback hypothesis during snowmelt. *Journal of Hydrology*, **219**, 188–205.

Kerkhoff A.J., Martens S.N. and Milne B.T. (2004) An ecological evaluation of Eagleson's optimality hypotheses. *Functional Ecology*, **18**, 404–413.

Kirchner J.W. (2003) A double paradox in catchment hydrology and geochemistry. *Hydrological Processes*, **17**, 871–874.

Kleidon A. (2004) Beyond Gaia: thermodynamics of life and earth system functioning. *Climatic Change*, **66**, 271–319.

Klemeŝ V. (1983) Conceptualization and scale in hydrology. *Journal of Hydrology*, **65**, 1–23.

Klemeŝ V. (1986) Dilettantism in hydrology: transition or destiny? *Water Resources Research*, **22**, 177S–188S.

Lee H., Sivapalan M. and Zehe E. (2005) Representative Elementary Watershed (REW) approach, a new blueprint for distributed hydrologic modelling at the catchment scale: the development of closure relations. In *Predicting Ungauged Streamflow in the Mackenzie River Basin: Today's Techniques & Tomorrow's Solutions*, Spence C., Pomeroy J. and Pietroniro A. (Eds.), Canadian Water Resources Association (CWRA): Ottawa.

Leopold L.B. and Maddock T. (1953) The hydraulic geometry of stream channels and some physiographic implications. *US Geological Survey Proffesional Paper*, **252**, 9–16.

Levin S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**(6), 1943–1967.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, **99**(D7), 14 415–14 428.

Lovejoy S. and Schertzer D. (1995) Multifractals and rain. In *New Uncertainty Concepts in Hydrology and Water Resources*, Kundzewicz Z.W. (Ed.), Cambridge University Press: Cambridge, pp. 62–103.

L'vovich M.I. (1979) *World Water Resources and their Future*, English Translation, American Geophysical Union: Washington, p. 415.

McDonnell J.J. (1990) A rationale for old water discharge through macropores in a steep, humid catchment. *Water Resources Research*, **26**, 2821–2832.

McDonnell J.J. and Woods R.A. (2004) On the need for catchment classification. *Journal of Hydrology*, **299**, 2–3.

McGlynn B.L. and McDonnell J.J. (2003) Quantifying the relative contributions of riparian and hillslope zones to catchment runoff. *Water Resources Research*, **39**(11), 1310, doi: 10.1029/2003WR002091.

McMahon T.A., Finlayson B.L., Haines A.T. and Srikanthan R. (1992) *Global Runoff – Continental Comparisons of Annual Flows and Peak Discharges*, Catena-Verlag: Cremlingen-Destedt, p. 166.

Merz R. and Blöschl G. (2003) A process typology of regional floods. *Water Resources Research*, **39**(12), 1340, doi:10.1029/2002WR001952.

Milly P.C.D. (1994) Climate, interseasonal storage of soil water, and the annual water balance. *Advances in Water Resources*, **17**, 19–24.

O'Loughlin E.M. (1986) Prediction of surface saturation zones in natural catchments by topographic analysis. *Water Resources Research*, **22**, 794–804.

Ozawa H., Ohmura A., Lorenz R.D. and Pujo T. (2003) The second law of thermodynamics and the global climate system: a review of the maximum entropy production principle. *Reviews of Geophysics*, **41**(4),, 1018, doi:10.1029/2002RG000113.

Parker G. (1976) On the cause and characteristic scales of meandering and braiding in rivers. *Journal of Fluid Mechanics*, **76**(3), 457–480.

Reggiani P., Hassanizadeh S.M., Sivapalan M. and Gray W.G. (1999) A unifying framework for catchment thermodynamics. Constitutive relationships. *Advances in Water Resources*, **23**(1), 15–39.

Reggiani P., Sivapalan M. and Hassanizadeh S.M. (1998) A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy, entropy and the 2nd law of thermodynamics. *Advances in Water Resources*, **22**(4), 367–398.

Rinaldo A., Marani A. and Rigon R. (1991) Geomorphological dispersion. *Water Resources Research*, **27**(4), 513–525.

Robinson J.S. and Sivapalan M. (1997a) An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research*, **33**(5), 1045–1059.

Robinson J.S. and Sivapalan M. (1997b) Temporal scales and hydrological regimes: implications for flood frequency scaling. *Water Resources Research*, **33**(12), 2981–2999.

Robinson J.S., Sivapalan M. and Snell J.D. (1995) On the relative roles of hillslope processes, channel routing and network geomorphology in the hydrological response of natural catchments. *Water Resources Research*, **31**(12), 3089–3101.

Rodriguez-Iturbe I., D'Odorico P., Porporato A. and Ridolfi L. (1999) Tree–grass coexistence in savannas: the role of spatial dynamics and climate fluctuations. *Geophysical Research Letters*, **26**, 247–250.

Rodriguez-Iturbe I. and Rinaldo A. (2000) *Fractal River Basins*, Cambridge University Press: Cambridge.

Rodriguez-Iturbe I. and Valdes J.B. (1979) The geomorphologic structure of hydrologic response. *Water Resources Research*, **15**, 1409–1420.

Saco P.M. and Kumar P. (2002) Kinematic dispersion in stream networks. 1. coupling hydraulic and network geometry. *Water Resources Research*, **38**(11), 1244, doi: 10.1029/2001WR000694.

Savenije H.H.G. (2001) Equifinality, a blessing in disguise? *Hydrological Processes*, **15**, 2835–2838, doi: 10.1002/hyp.494.

Shuttleworth W.J. (1983) Evaporation models in the global water budget. In *Variations in the Global Water Budget*, Street-Perrott A. and Beran M. (Eds.), D. Reidel: Hingham, pp. 147–171.

Sivapalan M. (1997) Computer models of watershed hydrology. Book Review. *Catena – Journal of the International Society of Soil Science*, **29**(1), 88–90.

Sivapalan M. (2003a) Prediction of ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes*, **17**(15), 3163–3170.

Sivapalan M. (2003b) Process complexity at hillslope scale, process simplicity at the catchment scale: is there a connection? *Hydrological Processes*, **17**, 1037–1041, doi: 10.1002/hyp.5109.

Sivapalan M., Beven K. and Wood E.F. (1987) On hydrologic similarity. 2. A scaled model of storm runoff production. *Water Resources Research*, **23**(12), 2266–2278.

Sivapalan M., Blöschl G., Zhang L. and Vertessy R. (2003a) Downward approach to hydrological prediction. *Hydrological Processes*, **17**, 2101–2111, doi: 10.1002/hyp.1425.

Sivapalan M., Takeuchi K., Franks S.W., Gupta V.K., Karambiri H., Lakshmi V., Liang X., McDonnell J.J., Mendiondo E.M., O'Connell P.E., *et al.* (2003b) IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. *Hydrological Sciences Bulletin*, **48**(6), 857–880.

Sivapalan M., Woods R.A. and Kalma J.D. (1997) Variable bucket representation of TOPMODEL and investigation of the effects of rainfall heterogeneity. *Hydrological Processes*, **11**(9), 1307–1330.

Sklash M.G. (1990) Environmental isotope studies of storm and snowmelt runoff generation. In *Process Studies in Hillslope Hydrology*, Anderson M.G. and Burt T.P. (Eds.), John Wiley & Sons: Chichester, pp. 401–435.

Skøien J.O., Blöschl G. and Western A.W. (2003) Characteristic space scales and timescales in hydrology. *Water Resources Research*, **39**(10), 1304, 10.1029/2002WR001736.

Snell M. and Sivapalan M. (1994) On geomorphological dispersion in natural catchments and the geomorphological unit hydrograph. *Water Resources Research*, **30**(7), 2311–2323.

Troch P.A., Paniconi C. and van Loon E. (2003) Hillslope-storage Boussinesq model for subsurface flow and variable source areas along complex hillslopes: 1. Formulation and characteristic response. *Water Resources Research*, **39**(11), 1316, doi: 10.1029/2002WR001728.

Tromp-van Meerveld H.I. and McDonnell J.J. (2005) Threshold relations in subsurface storm flow 1. A 147 storm analysis of the Panola hillslope trench. *Water Resources Research*, in review.

Turner J.V. and Macpherson D. (1990) Mechanisms affecting streamflow and stream water quality: an approach via stable isotope, hydrogeochemical, and time-series analysis. *Water Resources Research*, **26**(12), 3005–3019.

Vache K.B. and McDonnell J.J. (2005) Discharge, streamwater residence time and distributed soil water residence time as evaluative criteria for runoff modeling. *Water Resources Research*, in review.

Wagener T., Sivapalan M., McDonnell J.J., Hooper R., Lakshmi V., Liang X. and Kumar P. (2004) Predictions in Ungauged Basins (PUB) – A catalyst for multi-disciplinary hydrology. *EOS, Newsletter of American Geophysical Union*, **85**(44), 451–457.

Weiler M. and McDonnell J.J. (2004) Virtual experiments: a new approach for improving process conceptualization in hillslope hydrology. *Journal of Hydrology*, **285**, 3–18.

Weinberg G.M. (1975) *An Introduction to General Systems Thinking*, Wiley-Interscience: New York.

West G.B. and Brown J.H. (2004) Life's universal scaling laws. *Physics Today*, **57**(9), 36–42.

Western A.W., Grayson R.B., Blöschl G., Willgoose G.R. and McMahon T.A. (1999) Observed spatial organisation of soil moisture and its relation to terrain indices. *Water Resources Research*, **35**(3), 797–810.

Wigmosta M.S., Vail L. and Lettenmaier D.P. (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, **30**, 1665–1679.

Willgoose G.R., Bras R.L. and Rodriguez-Iturbe I. (1991) A physically based coupled network growth and hillslope evolution model. 1. theory. *Water Resources Research*, **27**, 1671–1684.

Wood E.F., Boll J., Bogaart P. and Troch P.A. (2005) The need for a virtual hydrologic laboratory for PUB. In *Predictions in Ungauged Basins: International Perspectives on State-of-the-Art and Pathways Forward, Proceedings of the Australia-Japan Workshop on Pub Working Groups*, Franks S.W., Sivapalan M., Takeuchi K. and Tachikawa Y. (Eds.), IAHS Press: Wallingford.

Wood E.F., Lettenmaier D.P. and Zartarian V.G. (1992) A land-surface hydrology parameterization with subgrid variability for general circulation models. *Journal of Geophysical Research*, **97**, 2717–2728.

Woods R.A. (2002) Seeing catchments with new eyes. *Hydrological Processes*, **16**, 1111–1113.

Woodward F.I. (1987) *Climate and Plant Distribution*, Cambridge University Press: Cambridge.

Zehe E., Lee H. and Sivapalan M. (2005) Derivation of closure relations and commensurate state variables for mesoscale models using the REW approach. In *Predictions in Ungauged Basins: International Perspectives on State-of-the-Art And Pathways Forward*. In *Proceedings of the Australia-Japan Workshop on PUB Working Groups*, Franks S.W., Sivapalan M., Takeuchi K. and Tachikawa Y. (Eds.), IAHS Press: Wallingford.

Zhao R.J., Zhang Y.L., Fang L.R., Liu X.R. and Zhang Q.S. (1980) The Xinanjiang model. In *Hydrological Forecasting, Proceedings of the Oxford Symposium*, IAHS Publication No. 129, IAHS: pp. 351–356.

PART 2

Hydroinformatics

14: Hydroinformatics and its Contributions to Hydrology: From Computation to Communication

MICHAEL B ABBOTT[1], ARTHUR E MYNETT[2] AND J PHILIP O'KANE[3]

[1] *Knowledge Engineering B.V. Belgium and UNESCO-IHE Delft, Delft, The Netherlands*
[2] *WL Delft Hydraulics & UNESCO-IHE and Delft University of Technology, Delft, The Netherlands*
[3] *Environment Institute, National University of Ireland, Cork, Ireland*

Hydroinformatics is concerned with the application of information and communication technologies for the planning, management, and conservation of the aquatic environment. With the rapidly increasing capabilities of computer-based systems, considerable advances were made during the past decade or more. Some historic developments are outlined in this article and the development of hydroinformatics as a sociotechnology is explained. The transition from computation to communication is illustrated by examples from present day flood simulation and flood early warning systems; emerging applications of data-mining and evolutionary computing are demonstrated for the fields of ecohydraulics and ecohydrology. The impact of hydroinformatics on the future role of hydrology in a multiuser environment is discussed. The contours of a next (fifth) generation software systems making use of agent-based communication over a network of computers and developers are outlined. The scope and content of all subsequent articles within this chapter are presented; these can be broadly grouped into areas of (i) physics-based numerical modeling; (ii) data-driven modeling and evolutionary computing; and (iii) design and decision support systems.

INTRODUCTION

Hydroinformatics is quite often identified as "the application of information, computation, and communication technologies in the various fields of hydroscience and engineering". With the advent of computer-based modeling and stimulating the development of the SHE system (as will be elaborated later), hydroinformatics has been involved with hydrology already for quite some time. However, according to Abbott (1992):

> *"No theory of science – let alone a theory of sciences – can itself be constituted as a scientific theory. Similarly, that which serves a technology need not necessarily be a technology – and indeed it is most commonly not a technology. Hydrology, quite consentaneously, is itself not a science, and neither is it itself a technology, even though it deals in the productions of many sciences and serves many technologies. Accordingly, when viewed scientifically or technologically, it appears to be bereft of a body of theory and so lacking all intentionality. It appears to be without an own form and without an own direction. But if it is not a science and it is not a technology, what is the status of hydrology? It is now the time and the place to state, provocatively but with serious intent, that hydrology is a rhetoric waiting for a grammar."* (Abbott, 1992).

So much has happened in computer-based modeling in the years since this was written, and yet so little seems to have happened in the field of hydrology as a whole. Since that time, *hydroinformatics* has taken wing to become an indispensable adjunct and support to most major construction works in the aquatic environment, to have its own major biannual conferences, its own Journal of Hydroinformatics affiliated with all three international water associations, and to have been accepted generally as a major discipline in its own right, to be taught in many universities

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

around the world. Hydroinformatics started off as a technology by expanding the field of computational hydraulics in river, urban, and coastal engineering. However, hydroinformatics was also concerned with the field of computational hydrology through its involvement in the development of distributed physically based model codes, as represented by the European Hydrologic System/Système Européen Hydrologique (SHE – see Figure 1), which constituted a "quantum jump" in complexity as compared with any other code so far known in hydrology, as discussed by Abbott and Refsgaard (1996).

The ethos of the physics-based hydrological modeling systems that follow this tradition can be described as *essentialist*, while the regressions to devices that disregard many, if not all, of the productions of the physical, including biological, sciences, such as are still very actively propagated even today, can continue to be described as *intrumentalist*, in that they are only concerned to reproduce some observed results and even then only for particular combinations of processes occurring in specific places at specific times, and often further only under specific anterior conditions. As regressions to an earlier era, preceding that of modern science, they have the same concerns as had the ancient Greek astronomers, of saving the appearances ($\sigma\acute{\omega}\zeta\varepsilon\iota\nu\ \tau\acute{\alpha}\ \phi\alpha\iota\nu\acute{o}\mu\varepsilon\nu\alpha$) of certain observed phenomena, while disregarding all, or almost all, concern with understanding. It has been one of the disappointments of hydroinformatics that much excellent work that has been done, for example, in the area of data mining technologies, has been misused in this way and not least in the area of hydrology. Indeed, it is still apposite to repeat the observations of Klemes (1986; see also Abbott, 1992) on this unfortunate habit of misappropriating techniques borrowed from other disciplines:

> *"Hydrology, having no solid foundation of its own and moving clumsily along on an assortment of crutches borrowed from different disciplines, has always been an easy victim of this practice. Every mathematical tool has left behind a legacy of misconceptions invariably heralded as scientific breakthroughs. The Fourier analysis, as was pointed out by Yevjevich (in 1968), had seduced the older generation of hydrologists into decomposing hydrologic records into innumerable harmonics in the vain hope that their reconstruction would facilitate prediction of future hydrologic fluctuations (fortunately few computers were available at the time, so that the Fourier fever did not become an epidemic); various statistical methods developed for evaluation of differences in repeatable experiments have been misused to create a scientific analysis of unrepeatable hydrologic events; linear algebra has been used to transform the idea of a unit hydrograph from a crude but useful approximation of a soundly based concept into a pretentious masquerade of spurious rigour now exercised in the modelling of flood events;*

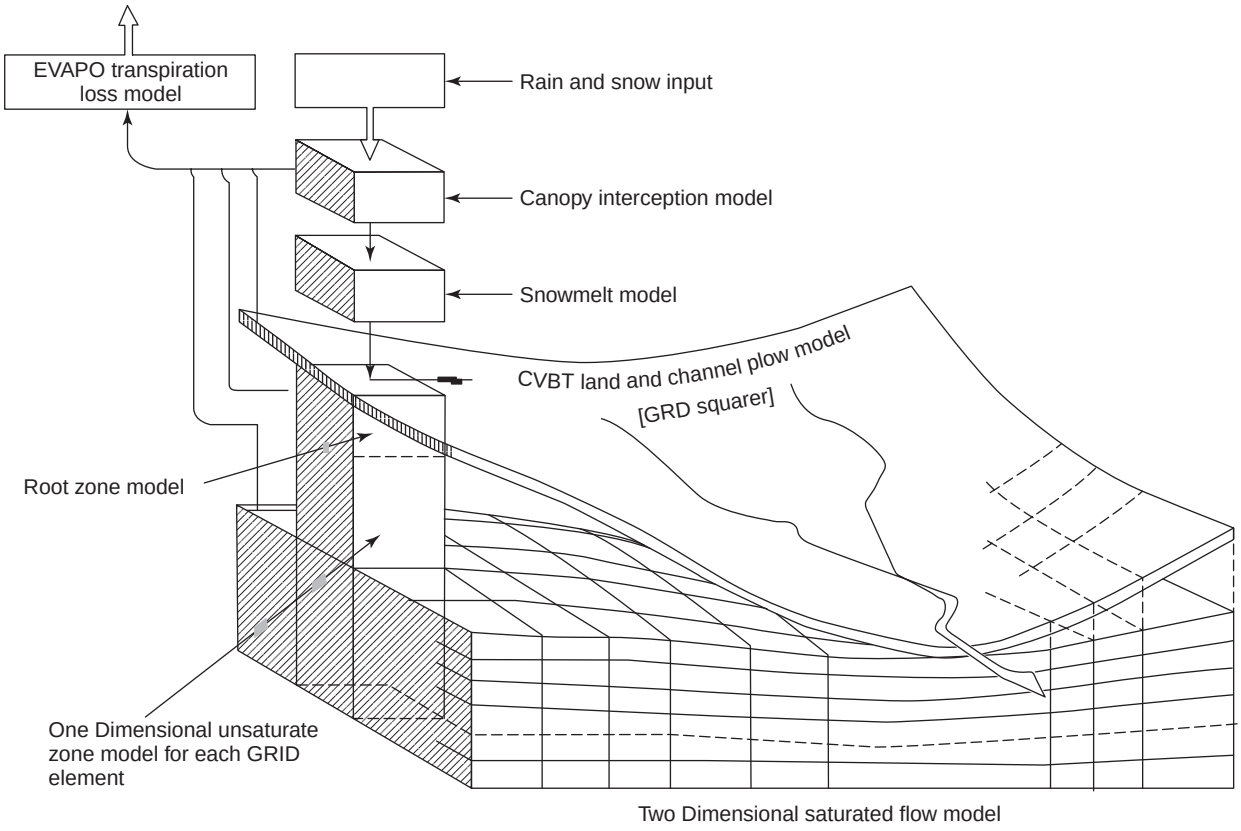


Figure 1 Schematic presentation of MIKE SHE (Courtesy DHI Water & Environment)

time series analysis has been used to remake inadequate 20-year stream flow records into 'adequate' 1,000-year records, or even more 'adequate' 10 000-year records; and the illusion of pattern recognition is now being courted in the vain hope that it will lend legitimacy to the unscientific concept of mindless fitting that dominates contemporary hydrologic modelling. In all these cases, mathematics has been used to redefine a hydrologic problem rather than to solve it". [Klemes, 1986]

Clearly, those who are concerned with modeling for engineering practice do not want to waste too much of their time on refuting these, as they see it, regressions to primitive forms of prescientific instrumentalism: they much prefer to persevere further along the path of modern science as this now comes to presence in an era of such greatly extended possibilities for communication. And, indeed, as will shortly be explained, it is in just this era that a modern-scientific hydrology can attain to its full potential.

HYDROINFORMATICS DEVELOPING INTO A SOCIOTECHNOLOGY

Hydroinformatics, being in essence "market-driven", so far has had relatively few dealings with the individual sciences that are linked together through hydrology and that constitute its scientific-hydrological foundation. Abbott and Refsgaard (1996) considered the marked discrepancies between developments in hydrology that could, in principle, be catalyzed by developments in hydroinformatics and developments in fields like river and coastal engineering, river basin, and coastal zone management, where already at that time hydroinformatics had triggered major advances.

It had by then become clear that only distributed physically based models like the SHE systems could make use of the rapid advances in instrumentation, supervisory control and data acquisition (SCADA) systems with real-time data transmission, data assimilation techniques, remotely sensed data and its interpretation systems, seamless geographical information systems (GIS) interfacing, advances in geodetic surveying incorporating GPS-based position-fixing equipment, cartographic transformation packages, intranetted and extranetted communication systems, and the many other developments that hydroinformatics had already woven together to obtain important synergies in areas like river basin and coastal zone management. By way of explanation of this situation, Abbott and Refsgaard (1996) identified the following four areas of difficulty in the application of such physics-based systems in hydrology:

"1. Data Availability

A prerequisite for making full use of the distributed physically-based models is the existence and easy accessibility of large amounts of data, including detailed spatial information on natural parameters such as geology, soil and vegetation and man-made impacts such as water abstractions, agricultural practices and discharges of pollutants. In most cases, all such relevant data does not exist and even the existing data is most often not easily accessible due to lack of suitably computerised data bases.

A further complication in this regard is the administrative problem created by the fact that these models, in addition to the traditional hydrometeorological data, require and can make use of many other data sources, such as those arising from agricultural, soil science, and geological investigations. Another important development gradually improving the availability of data is the application of GIS technology, which is particularly suitable for couplings with distributed hydrological models.

2. Lack of Scientific-Hydrological Understanding

With the introduction of a new modelling paradigm and concurrent research in process descriptions, new shortcomings in the scientific-hydrological understanding have emerged, especially with regard to flow, transport, and water quality processes at small scales and their up-scaling to describe larger areas. [Some of the key scientific problems were highlighted in several chapters of Abbott and Refsgaard (1996)]. These scientific shortcomings have, on the one hand, constrained the practical applications of distributed hydrological models and, on the other hand, the existence and application of such models have put a new focus on some of these problems, thus contributing to advances in the scientific-hydrological understanding.

3. Traditions in Hydrology and Water Resources Understanding

The distributed physically-based model codes, as represented by the European Hydrologic System/Système Européen Hydrologique (SHE), constituted a 'quantum jump' in complexity as compared with any other code so far known in hydrology. Moreover it used technologies, such as had been developed in computational hydraulics, with which few hydrologists were familiar. Although the numerical algorithmic problems could be largely overcome through the development of the codes into user-friendly fourth generation modelling systems with well proven algorithms, such as the MIKE SHE, the problem was then only shifted back to one of comprehending the fully integrated complexity of the physical system that was being modelled together with the constraints that were inherent in the modelling procedures.

Very few professional engineers and managers were, and still are, educated and trained with the necessary integrated view of hydrological processes in anything like their real-world physical complexity. This difficulty is exacerbated by the very nature of hydrology itself, whereby most professionals possess only a limited view of the physical processes involved. Soil physicists, plant physiologists, hydrogeologists and others usually have only a very partial view on the whole system, while there are few organisations that have available both the full range of such specialists and the more broader-ranging professionals that are needed in many situations to exploit the potential of distributed physically-based codes to such a degree that this exploitation is economically justified.

4. Technological Constraints

In order to achieve a wide dissemination of a new modelling technology to a considerable part of the professional community (and not only to experienced hydrological modellers) experience from hydraulic engineering shows that fourth generation (i.e. user-friendly software products) are required. Furthermore, it is believed that fifth generation systems are required to realise their full potential in terms of practical applications. The fifth generation systems [will be] hydroinformatics based." [Abbott and Refsgaard, 1996].

Thus, although already one decade ago the further development of physically realistic computer-based hydrological

modeling systems was seen to evolve from developments in hydroinformatics, the formidable institutional problems that could be blocking these developments in many areas were also foreseen already at the time. These institutional problems have become increasingly serious since then and have now, in their turn, led to serious problems within the hydrologic community itself. These problems appear as a divergence within the community concerning the very nature of hydrologic modeling and practice, and with this the role that hydrology has to play, or indeed can play, in society as a whole. Such divisions strike at the heart of the scientific foundations of hydrology, raising questions concerning its very nature as a discipline based upon the results of modern science.

In the years that have elapsed since the above evaluation was written, hydroinformatics has passed through two major transformations that have in turn been quite closely related. The first of these was the realization that hydroinformatics was something more again than a technology connecting together other technologies and sciences to provide applications. As the environmental movement grew ever more vociferous and exercised ever more power at the level of government, hydroinformatics was drawn increasingly into the public debates of the matters that it treated. It had then necessarily to develop new systems, both to provide means for governments to satisfy the engaged populace and to represent the interests of this populace at the level of government.

Hydroinformatics became increasingly involved in drafting legislation that would empower the concerned populace as genuine stakeholders in water resources, and this necessitated the development of information and knowledge systems for transforming this legislated empowerment into real-time, on-line operational realities. Hydroinformatics not only explored ongoing technical developments, such as those associated with rapidly expanding technologies like the Internet, mobile telephony, applications of Java, and other communication-enhancing technologies, but it also became inseparably involved in the so-called "technologies of persuasion" and other, previously separated, social aspects. It entered into a field where social and technical aspects were so closely interwoven that no development within the one could proceed at all without a corresponding development in the other – thus hydroinformatics increasingly became a *sociotechnology*.

FROM COMPUTATION TO COMMUNICATION

This chain of development was, of course, in all cases inseparable from the second major development: a change in emphasis in hydroinformatics *from computation to communication*. The computational development continued, of course, but its products were communicated in new forms directed to satisfy the needs for instantaneous information and supporting knowledge to a totally new kind of nonprofessional, but highly engaged, public, commonly represented by influential nongovernment organizations (NGOs). This public, whether as individuals or NGOs, needed to know the precise movements of water, sediments, and other materials not only at the very moment that these were occurring, but also as these would occur in the

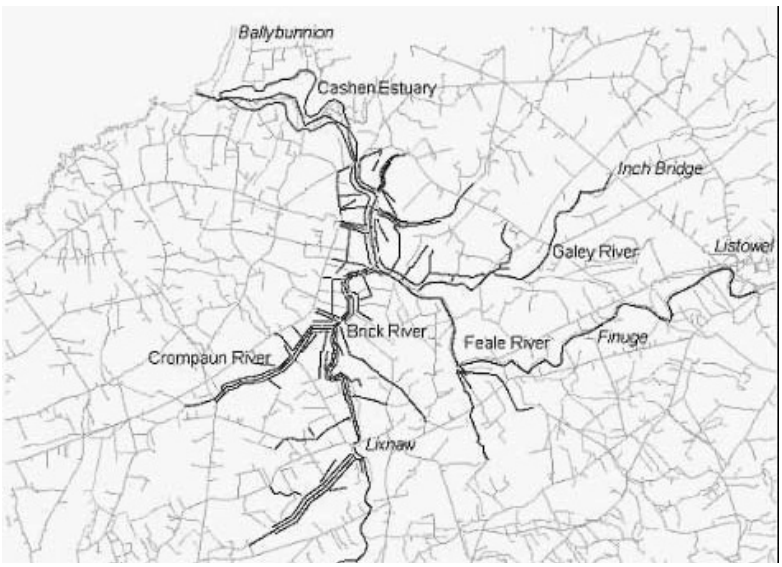


Figure 2 Lower Feale river network, Ireland. The Feale debouches through the short Cashen estuary into the outer Shannon at Ballybunnion. The Upper Feale enters the 250 km^2 Feale polders (15) at Listowel where it forms a torrent at high flows. Source: Martin (2002) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

foreseeable future under existing plans. Only under these circumstances could this public, through its NGOs, be responsibly empowered by legislation to modify, or even stop, the ongoing works if these posed threats to its perceived interests. Hydroinformatics thus entered into an ongoing process of ever increasing interaction between long-established representative democratic forms of government and certain new forms of participative democracy, as catalyzed by rapidly ongoing developments in communication technologies. It is only recently that the consequences of this development for hydrology have become clear.

If we define a science conventionally as a"totality established through an interconnection of true propositions", then each science can be regarded as the body of knowledge that is expressed by, or encapsulated in, this totality. When we speak about the physically and biologically based sciences upon which hydrology draws for its sustenance, we speak in the spirit of modern science. The development of modern science from Buridan's recognition of the law of conservation of momentum – of impetus – onwards, has been recorded so extensively that no further commentary will be given here. Similarly, the introduction of empirical, but carefully measured, coefficients into these universal laws in such a way that these coefficients also apply to all processes of a specific kind in all places and at all times, has been greatly advanced over the centuries: see Cunge (2003) as an example from hydraulics with an obvious relevance to hydrology.

O'Kane *et al.* (2004) followed this approach to explore the hydrology and hydraulics of a pumped polder in North Kerry, Ireland, by setting up a distributed physics-based modeling system for the lower Feale catchment in southwest Ireland – see Figure 2 – that is subject to annual flooding. Two of the 15 hydraulically independent polders making up the catchment were instrumented as part of a pump experiment. Several state-of-the-art, high frequency instruments monitor the pump experiment and provide input data for a very detailed hydrologic-hydraulic model. An electrical resistivity investigation and a ground-penetrating radar survey were carried out to collect subterranean data and a number of deep boreholes were drilled to calibrate these two surveys. A dynamic geographical database, built in ArcGIS (ESRI), was set up to store the wide range of data collected and the results of the different surveys carried out (see Figure 3).

An integrated hydrological-hydraulic model of the sub-catchment that includes the two polders was developed to

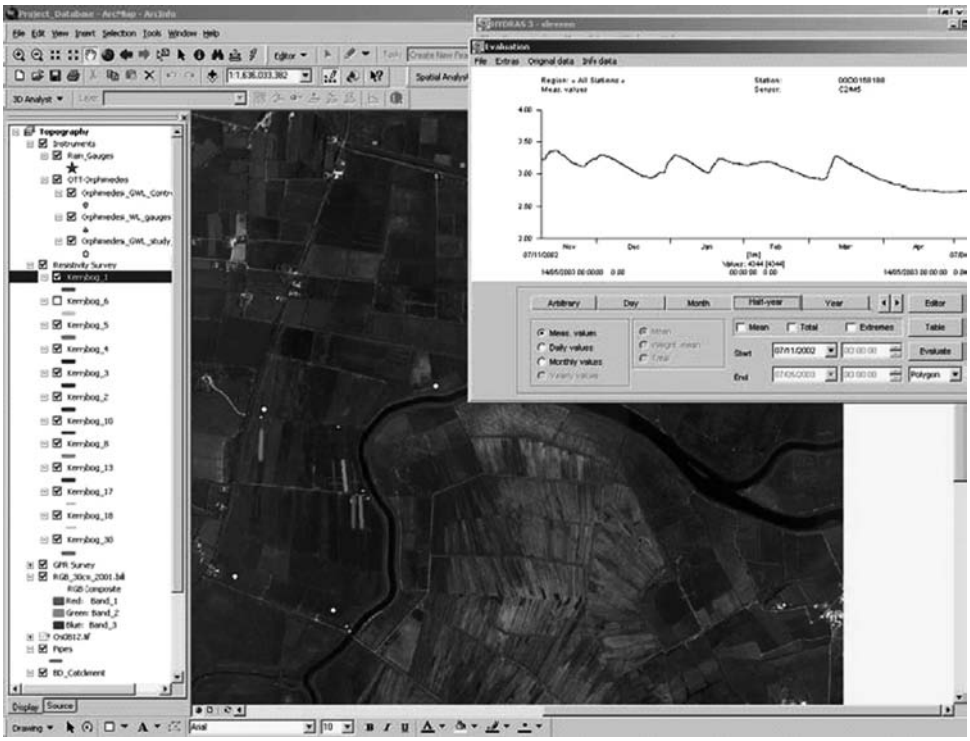


Figure 3 ArcGIS database and ArcMap layout of the centre of the Lower Feale – see figure 2. The window on the right shows the measurements of groundwater level during winter at a gauge in the pumped polder to the west of the Brick tributary. Source: Migliori (2004), O'Kane and Migliori (2004). A color version of this image is available at http://www.mrw.interscience. wiley.com/ehs

evaluate the effectiveness of pumping, for eventual implementation in a flood mitigation strategy in the remaining polders. The model of the study site was implemented using the combined Mike SHE/Mike11 modeling system of DHI Water and Environment. It simulates the surface drainage network, saturated and unsaturated groundwater zones, all hydraulic structures, and incorporates a high-resolution digital elevation model.

The main objective of the study was to properly understand the arterial drainage infrastructure, and, in particular, the performance of the sluiced culverts system and its relationship with flooding. A detailed evaluation of various degrees of engineering interventions for the alleviation of flooding (dredging, pumping, etc.) was carried out with the computer simulation model. A network of pumps was found to be the most effective solution from both engineering and economical point of view. On the basis of the graphical display capabilities of the (sociotechnical) hydroinformatics system, the results could easily be communicated with the public and provided useful support to the decision makers involved.

Yet another example of the combined use of high-resolution digital elevation (DEM) data, advanced numerical simulation systems, and the powerful visualization capabilities of GIS is demonstrated by McGrath *et al.* (2003), exploring the implications of sea level rise on the flood risk assessment for the city center of Cork, Ireland. With access to high-resolution DEM data produced by the German Aerospace Agency (DLR), the visualization of likely natural flooding behavior becomes much more achievable. Using the High Resolution Stereo Camera Airborne (HRSC-A) equipment originally intended for use in the Mission to Mars space program, the National University of Ireland, Cork, applied it to remote sensing problems on the surface of the Earth. In May 2001, an aerial survey using this camera was conducted over Cork City and the Lower River Lee. The resultant elevation datasets cover an area of 325 km^2, coregistered and geo-referenced to the Irish National Grid, having a −1 m spatial resolution and a −10 cm vertical resolution.

At present, the River Lee regularly reaches the level of the city center quays and occasionally overwhelms its banks and floods the central island and adjacent areas including the Central Business District. The usual scenario leading to such events is the joint occurrence of a spring tide with a low pressure weather system to the southwest of Ireland, which usually results in precipitation and Southerly winds driving the water up from the harbor. Such effects are presently mitigated by the controlled release of water into the river system by the Electricity Supply Board at the Inishcarra and Carrigadrohid dams to the west. It should be noted, however, that the elevation of the docklands regions is even lower than that of the city center. These docklands are in fact built in an urban polder and risk being completely inundated in the event of a general sea level rise due to global warming. This will be particularly worrying as The Cork Docklands Development Strategy foresees an investment of €38 000 0000 on infrastructure to facilitate high-density housing as well as improved transport facilities.

Because of this, detailed investigations into the hydrological and hydraulic implications were carried out using DEM and GIS in combination with advanced numerical simulation tools, presenting the results by easy-to-grasp visualizations of the likely flooding effects due to sea level rise, as presented in Figures 4a and 4b. These "hydroinformatics systems" offer powerful instruments for decision support on the development of any flood defence scheme and on the efficacy of potential mitigation measures. More details on these various aspects are provided in subsequent articles by

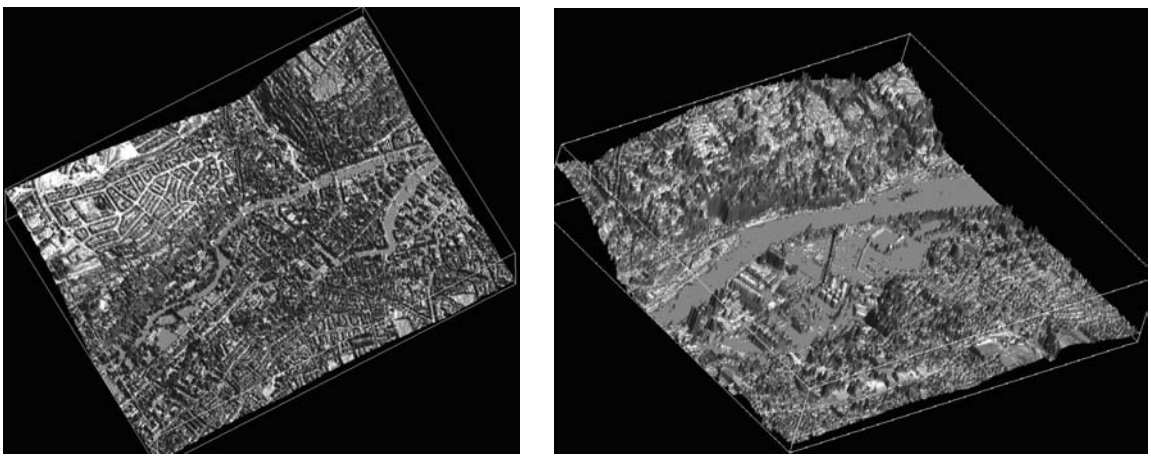


Figure 4 Perspective view of flooding extents in Cork city center for a 1-m high tide level rise Source: McGrath *et al.* (2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Moglen and Maidment (*see* **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**), Stelling and Verwey (*see* **Chapter 16, Numerical Flood Simulation, Volume 1**), Lin and Falconer (*see* **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**), as well as Guinot (*see* **Chapter 18, Shallow Water Models with Porosity for Urban Flood Modeling, Volume 1**). A recent and ongoing application of flood early warning system (FEWS) development in the United Kingdom is elaborated in the article by Werner *et al.* (*see* **Chapter 23, Flood Early Warning Systems for Hydrological (sub) Catchments, Volume 1**).

FLOW RESISTANCE DUE TO VEGETATION OBTAINED FROM GENETIC PROGRAMMING

Since hydroinformatics is much concerned with planning, management, and conservation of the aquatic environment, applications in the area of ecohydraulics and ecohydrology arise quite naturally. The scope and techniques of environmental hydroinformatics were outlined by Mynett (1999, 2002) and some practical applications of hydroinformatics tools in eco-compatible adaptive water resources management were indeed a topic of discussion during the Virtual Water Forum at the WWF3 in Japan (Mynett, 2003a) as well as during the latest Hydroinformatics Conference in Singapore (Mynett, 2004a). Also, within the ecohydraulics community, it is recognized that eco-compatible adaptive water management strategies can only be developed and implemented by making full use of advances in hydroinformatics system development as demonstrated by Jorde (2004) and Hardy *et al.* (2004). For the design of river restoration measures and adaptive river basin management procedures, evolutionary algorithms and fuzzy rule-based systems prove extremely valuable (Chen, 2004a).

Since considerable emphasis is being put on flood simulation, proper modeling of overland flows in wetlands and vegetated floodplains is of great practical importance both in flood early warning and in river restoration. Many research initiatives have been undertaken in order to improve on the description of the relationship between flow resistance and the temporal and spatial distribution of vegetation. Both analytical and experimental studies of vegetation-related resistance to flow and the equivalent resistance coefficients have shown that the resistance coefficients are very much waterdepth- (and hence solution) dependent. Consequently, the traditional approach of using a single resistance coefficient fails to describe correctly the (nonlinear) physical behavior of this phenomenon. One way of improving the description is updating the equivalent resistance coefficient on the basis of the computed water depth. However, in order to do so, a relation between the vegetation characteristics, bed resistance, water depth, and equivalent resistance coefficient is needed. Genetic programming (GP) was used by Rodriguez (2004) following up from Babovic and Keijzer (1999, 2000) to derive such relation.

When refining a model of a physical process, a scientist focuses on the agreement of theoretically predicted and experimentally observed behavior. If these agree in some accepted sense, then the model is considered "correct" within that context. Here, the inverse problem to verification of theoretical models is considered: (how) can we obtain the governing equations directly from measurements? To do this, Babovic and Keijzer (2000) extended the notion of qualitative information contained in a sequence of observations to consider directly the underlying mechanisms. They showed that, using this information, one can deduce the effective governing equations directly from the data. In this way, the deterministic component of the observed behavior can be obtained to an *a priori* specified level of correctness or accuracy.

One way of obtaining a detailed account of resistance description of flow through and above vegetation is to perform detailed numerical simulations on the basis of a one-dimensional turbulence model for the vertical (1DV) direction (Uittenbogaard, 2003). The 1DV model assumes that the flow is locally uniform in the horizontal directions and calculates the orthogonal horizontal velocities $u(z)$ and $v(z)$ as a function of the vertical coordinate z. The 1DV model is a simplification of the full 3D Navier–Stokes equations by decoupling the vertical from the horizontal flow conditions. By including the effects of plants into the commonly used k-ε model for turbulence closure, the following equation of motion arises:

$$\rho_0 \frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} = \frac{\rho_0}{1 - A_\mathrm{p}} \frac{\partial}{\partial z} \left((\nu + \nu_\mathrm{T}) \frac{\partial u}{\partial z} \right) - F \qquad (1)$$

where $(1 - A_\mathrm{p})$ denotes the specific area occupied by the fluid and F is the drag force exerted by the plants. The k-equation in the k-ε model can be formulated to account for the effect of plants, as described in detail by Uittenbogaard (2003). The results of the 1DV k-ε model can be used to obtain the detailed roughness description of resistance to the flow caused by vegetation and to obtain the commonly used roughness values like the Manning (n), Chézy (C), or Nikuradse (k_s) coefficients. Water-depth-dependent roughness relationship as well as the water-level slope are plotted in Figure 5 against the ratio of plant height (k) and water depth (h). Two different conditions can be identified: (i) unsubmerged vegetation, when the plants height exceeds the water depth and (ii) submerged vegetation, when the water depth exceeds the plants height.

Clearly, for unsubmerged flow conditions, the water-level slope is much higher than for submerged conditions implying that the resistance of the vegetation is higher.

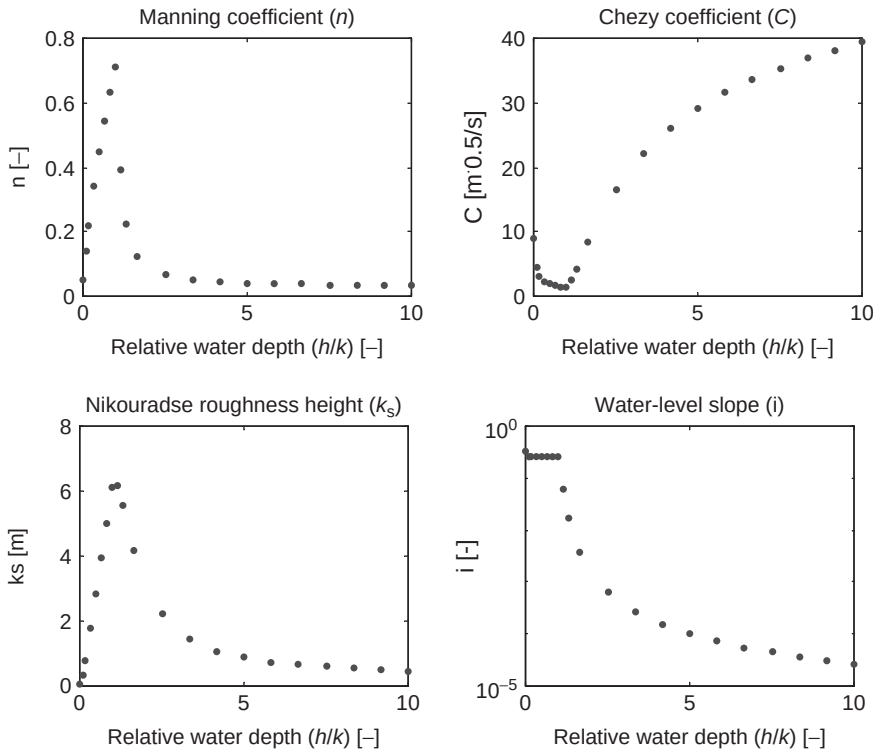


Figure 5 Vegetation-related resistance coefficients and water-level slope versus water depth. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Unsubmerged flow conditions can be successfully treated analytically (Rodriguez, 2004). If the water level is high enough, flow through the vegetation is negligible compared to the flow above, as can be seen from Figure 5; in the transition zone, both the flow through the vegetation as well as above it are both relevant and, consequently, all resistance coefficients are depth-dependent.

Genetic algorithms, evolution strategies, evolutionary programming as well as differential evolution represent just few examples of evolutionary algorithms that have been explored by a number of scientists during the past decades. However, it was rather recently that Koza (1992) of Stanford University proposed a special kind of evolutionary algorithm: GP, which involves symbolic expressions. Applications of GP in hydrosciences were introduced by (Babovic and Abbott, 1997) and (Babovic and Keijzer, 2000). Also, (Keijzer and Babovic, 1999) developed the *dimensionally aware GP* algorithm, using the observations together with their units of measurement. By requiring dimensional correctness as constraints – in what can be called a *weakly typed* or *implicit casting* approach – dimensional correctness is promoted rather than enforced.

The role of the (expert) user is then to choose his most suitable formulation to further analyze the proposed relationships. The user can exploit background knowledge or implement some belief about the problem domain. The final step lies in examining the selected equation(s) in order to interpret them. When a reasonable explanation for the apparent goodness-of-fit of such an equation is produced, the user's belief in the correctness of the equation is enhanced. The equation then no longer functions as a black box for making accurate predictions but as a genuine empirical equation that can be used with more confidence than mere statistical accuracy: the equation and its interpretation are amenable for review by experts and peers.

The dimensionally aware GP approach was applied to a set of 990 calculations carried out with the 1DV numerical model for submerged vegetation, using the input variables as presented in the table below (Rodriguez, 2004). For dimensional consistency, a slightly adapted Chézy's coefficient was used by virtue of which time-related units of measurements can be avoided and the resistance coefficient becomes solely a function of the geometry of the system. GP was then employed in a multiobjective sense, simultaneously optimizing the following three objectives:

1. root mean square error (RMSE) (RMSE – measure of the overall accuracy of the formula)
2. coefficient of determination (CoD) – measure of the goodness of shape of the formula and
3. dimensional error – measure of the dimensional consistency of the formulae.

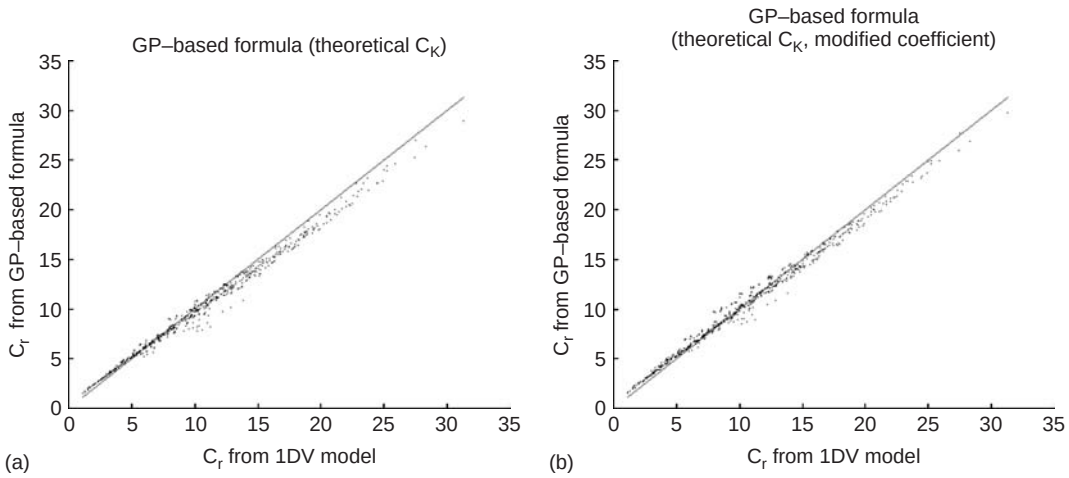


Figure 6 Scatter plots for C_r (GP-based formula with theoretical C_k for different coefficients). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The results presented in Figure 6 indicate that quite a good agreement was achieved.

HARMFUL ALGAL BLOOM PREDICTION USING FUZZY LOGIC AND CELLULAR AUTOMATA

In case of flooding over large areas and/or prolonged periods of time, pollution is likely to become an issue. Algal blooms may develop that can significantly affect water quality and cause ecosystem deterioration or even threaten human health. Predicting algal blooms is an ambitious and difficult topic because of the complexity of the aquatic ecosystem behavior, the insufficient knowledge available on the detailed processes and mechanisms involved, and the shortage of high quality data. However, artificial intelligence and machine learning techniques like fuzzy logic (FL) have the ability to deal with imprecise, uncertain or ambiguous data or relationships among data, and, hence, have the potential to be a useful and practical method in algal bloom modeling.

In order to explore this, a robust fuzzy logic approach was developed by Chen and Mynett (2004b) deriving (inter) relationships directly from measurements, using expert knowledge as a reference. This allows the combination of partial knowledge on processes with partially available data from observations. In collaboration with the European Commission project Harmful Algal Bloom Expert System (EU-HABES), this approach was applied to the North Sea for modeling Chlorophyll-a (Chl-a) concentrations, which provide a good indication of the likely expectance of algal bloom events.

In Dutch coastal waters, spring algal blooms, in particular, *Phaeocystis globosa* occur almost every year. The blooms cause great adverse impacts on shellfish farming and recreation. In order to predict the possible occurrence of spring algal blooms, Chen and Mynett (2004b) developed an integrated numerical and fuzzy cellular automata model combining information on irradiance, nutrients availability, and neighborhood conditions. The model was developed to predict phytoplankton biomass – a very good indicator for the potential occurrence of algal blooms in the Dutch coastal waters. The numerical Delft3D-WAQ (water quality) model was used to simulate the flow conditions, water column irradiance, nitrogen and phosphorus concentrations discharged from the River Rhine.

Membership functions are constructed by: (i) collecting available heuristic knowledge and setting up a reference rule base; (ii) clustering input and output data into characteristic clusters or subsets using, for example, self-organizing feature map (SOFM) techniques; (iii) defining membership functions for each variable by selecting proper function types and mean values for each cluster; (iv) inducing fuzzy rules by either *feature reasoning* or *case-based reasoning* (Chen and Mynett, 2004b). Sample membership functions for total inorganic nitrogen (TIN) and Chlorophyll-a for conditions along the Dutch North Sea coast are presented below in Figure 8.

The fuzzy logic module developed for the Dutch coast was used to predict algal biomass using the computed abiotic factors from the numerical Delft3D model. In order to take into account the spatial heterogeneity and local behavior, and to capture patchiness dynamics, a cellular automata paradigm was implemented using the available model grid configuration. The simulation results for the year 1995 as presented in Figure 7 were found to exhibit similar features as obtained from satellite images (Mynett, 2003b). As to the temporal evolution of bloom events, the modeled and observed increase in Chl-a concentration are seen to correspond quite well (Figure 9), given the complexity of the processes involved.

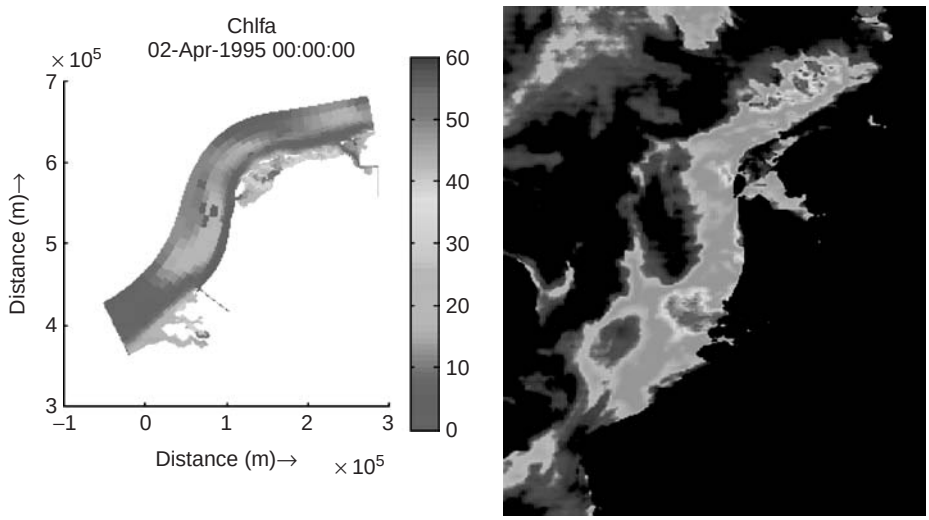


Figure 7 Spatial distribution of bloom prediction and satellite observations along the Dutch coast. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

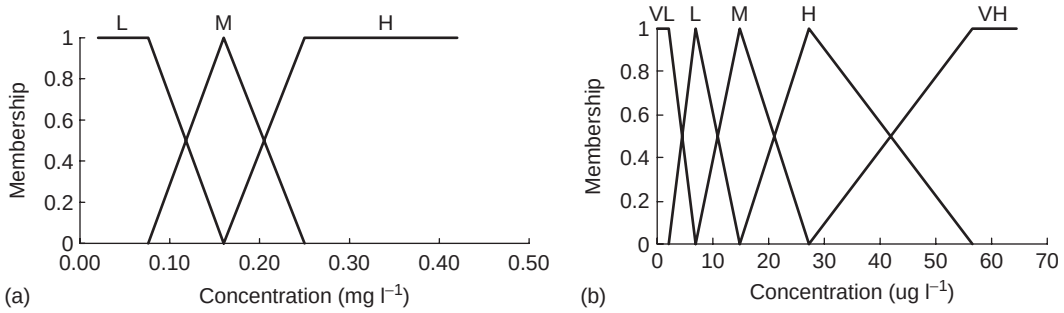


Figure 8 Membership functions for TIN (a) and Chl-a (b)

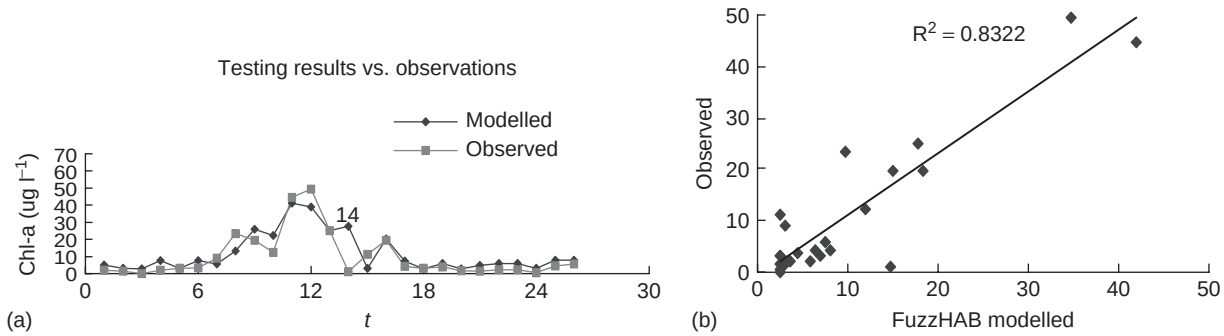


Figure 9 Time evolution of modeled and observed Chl-a concentration (a) and scatter plot (b). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The overall conclusion is that hydroinformatics is seen to play an increasingly important role in water quality and ecosystem modeling by providing tools and technologies (often in terms of software systems) that can integrate expert knowledge with measurement data and numerical simulation packages, as elaborated by Mynett *et al.* (2004c).

THE "HOLY GRAIL" OF THE FIFTH GENERATION SOFTWARE SYSTEM

It may be recalled that the first generation software was confined to the programming of methods originally intended for human computation, so that these methods were "people friendly" and often not at all suited to solutions using digital

machines. The second generation, from 1955 onwards, was consequently directed to more machine-friendly methods that were, correspondingly, not particularly "people friendly". These methods then necessitated the construction of a theoretical apparatus to enable the programmer to translate from people-friendly methods (e.g. partial differential equations) to computer-friendly methods (e.g. finite-difference, finite-element, finite-volume methods etc.). The second generation restricted itself almost exclusively to physics-based models that were distributed in space and time but each of which was specialized to a particular, domain-specific, physical object (a particular river, a particular coast, particular settling tank, etc.). The third generation, from 1970 onwards, took the methods developed for the second generation but encapsulated them into devices that could themselves construct any domain-specific model from a prescribed formatted description of the particular domain of interest.

The first three generations of software in this area were all restricted to models of movements of water and all that this moving water transported. The business area that was built up around this technology was correspondingly a business that was restricted to supplying the *results* obtained from running these models and was not a business in supplying the models themselves. It was only with the introduction of fourth-generation modeling systems, from 1985 onwards, that a part of the business passed over to selling models to end-users, so that at least a part of the business organizations involved passed from being "performers of projects" to being "providers of products". The project departments themselves grew as the products became more widely diffused and as the possibilities as well as the difficulties of modeling became more clearly delineated: "the projects followed the products" as sources of revenue to the tool-making/tool-using organizations. Also, the development of fourth-generation modeling introduced a new business model in this field: products had to be marketed and sold, training had to be provided, and service had to be available locally, in the local language. This in turn led to agent-network arrangements of considerable proportions, so that even at these early stages, activities in this area took on a *socio*technical complexion.

Almost simultaneously, with the development of fourth-generation modeling was the rise of hydroinformatics: the much more general application of information, and subsequently also communication, technologies to water problems. An increasing variety of data acquisition, data transmission, data processing, and data storage systems were elaborated; geographical information system (GIS) were ever more closely integrated with numerical simulation algorithms through "seamless interfaces"; these integrations greatly facilitated the incorporation of remotely sensed data into the domain description; global positioning systems (GPS) were integrated into hydraulic-geographic-photogrammetric systems; software for transformations between different cartographic coordinate systems and rapid advances in geodetic surveying were further incorporated; complex information management systems were installed to make the resulting information applicable to a wide range of end-users; elementary knowledge management systems appeared as means to accommodate the stream, growing to a flood, of information; and so on.

This complex of developments was the birthplace and the nursery of hydroinformatics. It quickly became clear (typically from 1995 onwards), however, that few, and ever fewer, of the problems with which hydroinformatics was confronted would be usefully treated as purely technical problems, but the greater number, and an ever greater number, were essentially *socio*technical. Whether this was in the area of self-management systems that allowed two or more contractors to work simultaneously in the same space under design-and-build contractual arrangements, or for enabling work schedules to be synchronized accurately with determinant hydraulic and environmental events, or for such operations as bathing, fishing, and mussel farming to be sustained in the vicinity of disturbances due to large constructions, which could extend over 100 km or more, or – most important of all – for gaining and retaining public tolerance, acceptance, and even support for interventions in nature that affected the interests of this public, the sociotechnical came ever more to the fore.

The introduction of *communication* (and primarily Internet-enabled) systems that made it possible for the general public to participate actively in all the activities that concerned this public became the *conditio sine qua non* for the practice of such activities. The duties and rights to exercise this participation were then anchored in the enabling legislation itself. It thereby became increasingly apparent that the communication processes were becoming ever more prominent as compared with the purely computational processes. Correspondingly, ever more persons were employed on enabling the communicative processes than were employed on the computational processes.

AGENT-BASED COMMUNICATION

These developments have led some hydroinformaticians to turn their attention increasingly towards communication processes, all of which are in their essence sociotechnical processes. They have done this from the two sides, both from the social side and from the technical side, even while understanding how the processes studied on these sides conjoin, merge or are "woven together" in the center of the sociotechnical complex. One part of the resulting study has been directed to societies that are themselves composed of those artifacts, commonly described as "agents", that

have been introduced above. These agents interact so as to form societies rather like "human agents" do, and these interactions are inevitably realized through processes of communication. It was long ago recognized that as models, for example, become taxonomically more extended, the communication aspects increasingly take precedence over the computational processes (Abbott, 1992, 1993). These studies have now resulted in the identification of a new approach to software applications in the water sector, and one that is in no way restricted to stand-alone numerical modeling but is much broader in its scope.

The fifth generation paradigm starts out from the observation that the communication aspects of software systems have now become the predominant ones in future developments. This is not to say, of course, that the computational aspects can now be ignored – that is far from being the case – but that the strategic developments must now proceed primarily through communication paradigms. Thus, to take the most familiar example of the numerical model, the main thrust of developments resides in allowing the various computational elements in this model to communicate with one another and with the world outside the model in much more efficient ways. For the sake of simplicity, it is possible to think of the fifth generation paradigm as being one in which software systems generally, with models as just one example, are composed of numerical components, regarded as "agents". These communicate with one another through what can be regarded as a kind of intranet and which, correspondingly, communicate with other, and normally very different, "agents" in their outer world through what can be regarded as a kind of extranet. Since all the "agents" involved then communicate and thereby cooperate by passing messages between them, the resulting architecture is commonly described as a "multiagent, message-passing" architecture.

This communication architecture can also be applied to the elaboration of new hydrologic modeling tools and new hydrologic modeling practices where this approach to modeling is most obviously beneficial. However, this approach may be hampered by the restrictive communication architectures of existing distributed physics-based hydrological systems. The reason for this is simple: once having accepted that the communications within every model generated by the new system should use the same technology as the communications between this model and its environment, the use of a common communication technology is almost mandatory. However, the communication processes occurring in, for example, MIKE SHE are, so to say, "hard wired", being largely fixed for all models, with only limited facilities for customization to specific modeled conditions. Although advanced for its time (1977–1978), it achieved its reliability at a cost in inflexibility that is not warranted by recent developments in communication technologies.

Despite its antiquity, the integration provided by the MIKE SHE modeling system is still judged higher than in any other hydrological modeling system currently available (e.g. Kaiser-Hill, 2001). With the various add-ons now available, it remains a powerful tool for modeling many water and environmental problems. Moreover, developments are already under way to extend the integrated modeling capacity even more, by interfacing with other models that were previously designed for modeling entirely separated water-related problems. For example, work continues on interfacing MIKE SHE with the MIKE MOUSE modeling system, used for urban drainage and sewer systems (where leakages into ground water commonly occur). This demonstrates the increasing demand for integrated modeling.

However, by the same token, since this system can currently be operated only through its graphical user interface, having no facilities to communicate with its outside world in any other way, it has no capacity to communicate over the Internet: in a word, it is not web-enabled. Thus, although the MIKE SHE remains a formidable instrument, it belongs to a paradigm that is rapidly becoming too restricted to remain competitive with future generation communication-based systems.

HYDROLOGY AS A CONJUNCTIVE KNOWLEDGE IN A MULTIKNOWLEDGE ENVIRONMENT

Multiknowledge environments are those in which two or more different knowledges have to function simultaneously and interactively in supporting one and the same activity. For this purpose, the participating knowledges alone cannot suffice, but a quite other knowledge is required to connect and coordinate their interactions. Such a knowledge is called *a conjunctive knowledge* (Abbott, 1993). Thus, for example, *sociotechnology* is a conjunctive knowledge that connects and coordinates (or *conjoins*) interacting social and technical processes, weaving these together to provide knowledge products that are not accessible from the social and technical domains independently. In such cases, the one knowledge is said to be *implected in*, or *implexively contained within*, the other through the intercession of the conjunctive knowledge. Clearly, hydrology is also a conjunctive knowledge, weaving together the strands of meteorology, plant physiology, soil physics, and other such sciences within its multiknowledge environment with its intention always directed towards the world of practice, of technology. The structure within a professional space implies the relevance of a specific architecture, which is called most generally an *agent-orientated architecture*. Then, in the words of Jonoski (2002, p203):

> *"What then is the essence of the new approach to conceptualization that is based on agent-orientation? As a first step, the new*

approach 'expands' the traditional conceptualization of knowledge domains beyond their descriptions in terms of different objects, with their properties and relations. In fact it 'replaces' the basic notion of 'knowledge domain' with a much broader notion of an 'environment', where the basic conceptualization units are not objects, but new kinds of entities called agents. The fundamental property of the agents is that they are embedded in their environment and are able to perceive and affect the environment. In addition to that, in their interaction with the environment they express goal-like behaviour. This is in fact very close to a general definition of an agent provided by Russel and Norvig (1995) who define an agent as any entity which operates in some environment (physical, virtual, cyberspace), which possess sensors to perceive this environment, effectors to act on this environment, and goals of its own which may or may not be explicitly represented in the agent itself. In general terms, an agent can be envisaged as a "creator of objects" within a fluctuating content."

"One of the radical changes that this shift in the mode of conceptualisation brings about is that it allows the agents to have their own representation of knowledge domains, which may or may not be part of their own environment. An agent's environment is therefore not to be confused with a knowledge domain. In some sense, the exclusive role of the (human) observer in the traditional, purely rationalistic, approach, of conceptualising and representing domains in terms of objects, has now been attributed to these new kinds of entities called agents. This is certainly one of the most fundamental changes brought about with the new approach, primarily because it allows for different perspectives (or points of view) between the human observer and the agent, and/or between the different agents themselves."

Thus, instead of the crude *positivism* associated with an object orientation, one has a much more flexible and subtle *phenomenology* associated with an agent orientation. In this approach, an agent can be attributed to, for example, the unsaturated zone in the soil, and this will then be interrogating its environment of precipitation, infiltration, evaporation, transpiration, and root water uptake as defined by its position in the professional space in order to decide upon *its own most appropriate physical and biological behavior*. Other domain agents will have other "concerns" again, and each agent will be monitoring *its own environment* within the professional space from *its own point of view*, corresponding to *its own perceptions*, as colored by *its own intentions*. It is the interaction between these *agents with different perceptions corresponding to different intentions* that generates an *active cooperation between the agents* in the representation of the physical system, as each of these does, so to say, "what it most wants to do" within the constraints of its physical and biological environment. Thus, whereas an object-oriented approach leads to totally managed, fixed hierarchical structures, *which correspond to exostructure*s, an agent-oriented approach has the potential to provide emergent, self-managing structures *that constitute endostructures*.

The future (fifth generation) software systems can be characterized as "network centered" as opposed to the previous "desktop-centered" or even earlier "host-centered" systems. This means that the platform for its development and deployment will not be a single machine, but a *network of computers* and developers. Initially, the primary focus will be the utilization of the internal network of computers that functions as the central, or kernel, organization's intranet. In due time, but most likely sooner than most are expecting – and hydrologists better take active part in these developments or they will find themselves overtaken by developments in the combined areas of meteorology and hydraulics – these new approaches will lead to an entirely new generation of computer-based hydrological modeling in practice.

OUTLINE AND SCOPE OF SUBSEQUENT ARTICLES CONTAINED IN THE REMAINDER OF THIS CHAPTER

Maidment and Moglen (*see* **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**) elaborate on the powerful capabilities of the use of digital elevation data (DEM) and geographic information systems (GIS), already introduced above for some practical applications. The basic properties of all DEMs are introduced as well as the basic concepts of GIS technology. The use of GIS and DEM in a hydrologic context is demonstrated; the principles of flow direction, flow accumulation, watershed delineation, flow lengths, slopes, and travel times as derived from DEM data within a GIS are presented.

Stelling and Verwey (*see* **Chapter 16, Numerical Flood Simulation, Volume 1**) describe the state-of-the-art of numerical flood simulation, in particular, of hydraulic models based upon the laws of physics. The concepts of kinematic and diffusive wave approximations are discussed in order to provide better ways of understanding the physical behavior of flood wave propagation and to demonstrate the link between hydraulics and hydrology. This contribution briefly introduces the impact of new data collection techniques on numerical flood model schematization.

Lin and Falconer (*see* **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**) focus on the hydrological and environmental modeling of transport processes in rivers and estuaries. Numerical solutions of the governing hydrodynamic, solute, and sediment transport equations form the basis of a number of software tools. Details are given of three-, two-, and one-dimensional flow-field formulations and the transport of water quality indicators as well as sediment transport processes within the water column. The implications are discussed for hydrological models used to specify the pollutant discharge loads from land use models.

Guinot (*see* **Chapter 18, Shallow Water Models with Porosity for Urban Flood Modeling, Volume 1**) introduces the concept of porosity in combination with hydrodynamic flow modeling in order to model urban flooding where one has to deal with highly variable geometries and flow parameters. Since the application of distributed, physically based models to refined urban flood modeling is not yet feasible, an alternative approach could be to use two-dimensional macroscopic models on the basis of a modified version of the classical shallow water equations, where the fraction of the land surface occupied by buildings is accounted for via porosity. Such models provide a practical alternative to classical semidistributed models, as illustrated by a computational example on a small-sized urban area.

Solomatine (*see* **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**) elaborates on the use of data-driven modeling and computational intelligence methods in hydrology. With the advent of rapidly increasing data volumes, there is a strong emphasis in the field of Hydroinformatics already for a decade or more, on exploring so-called *data-driven modeling paradigms*. These are based on methods from artificial intelligence, computational intelligence, and machine learning that are becoming more and more feasible for practical application when increasing amounts of data are available that describe the phenomenon of interest. Various aspects of data-driven modeling are discussed, including data preparation and a brief overview of popular techniques – neural networks, regression and model trees, instance-based learning, nonlinear dynamics.

Minns and Hall (*see* **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**) elaborate in more detail the use of Artificial Neural Network (ANN) concepts in hydrology. The solution of many applied hydrological problems, such as the forecasting of floods, has for several decades been based upon the concepts of linear systems analysis. However, the introduction of informatics tools, such as ANNs, with their origins in cognitive sciences and pattern recognition, has made available new lines of investigation. Provided proper attention is paid by the expert modeler, ANNs invariably seem to provide a model whose goodness-of-fit to an independent testing data set is superior to that of parameter-based hydrological modeling systems.

Babovic and Keijzer (*see* **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1**) demonstrate the use of GP, yet another technique from computational intelligence, for modeling rainfall-runoff processes that are often highly nonlinear, time-varying, spatially distributed, and hence not easily described by simple models. By using data-driven models such as GP, one can attempt to model runoff on the basis of available hydrometeorological data. The capabilities of GP are discussed for creating rainfall-runoff models on the basis of data alone, as well as in combination with conceptual models (i.e. taking advantage of knowledge about the problem domain).

Savic and Khu (*see* **Chapter 22, Evolutionary Computing in Hydrological Sciences, Volume 1**) provide an extensive overview of applications of Evolutionary Computing (EC) in the Hydrological Sciences. These applications range from the use of EC to assist in the understanding of hydrological processes, to improve the performance of simulation models and, most recently, to quantify risk and uncertainty with the aim of providing decision support. Examples are drawn from hydrologic processes research, rainfall-runoff modeling, reservoir operations and control, groundwater systems, urban water systems, and water quality modeling.

Werner *et al.* (*see* **Chapter 23, Flood Early Warning Systems for Hydrological (sub) Catchments, Volume 1**) demonstrate the impact of hydroinformatics on the increasingly relevant practical (sociotechnical) issue of providing early warning as an effective strategy in reducing flood damage and loss of life. By combining computer-based simulation techniques, computer-based data processing capabilities, and computer-based communication facilities, they show the present state-of-the-art capabilities of setting up FEWS for hydrological (sub) catchments.

Jonoski (*see* **Chapter 24, Network Distributed Decision Support Systems and the Role of Hydrological Knowledge, Volume 1**) outlines a next generation Network Distributed Decision Support environment for decision-making processes in the fields of water and environment. These (NDDSS) systems are designed to enable active participation of large numbers of interested stakeholders, including the general public, using electronic networks such as the Internet and wireless mobile telephony. Core components of such systems are instantiated hydrological models, open for wider integration with, for example, ecological or socioeconomic models.

SOFTWARE LINKS

ARCGIS (ESRI)

DEM data (check Moglen-Maidment)

DHI website software systems
WL website software systems
HRW website software systems
HEC website software systems
USGS ???

FURTHER READING

Abbott M.B. (1979) *Computational Hydraulics*, Pitman Publishing: London.

Chen Q., Mynett A.E. and Teng L. (2004c) Integrated numerical and decision tree modelling of Phaeocystis globosa bloom along the Dutch coast. In *Proceedings of the 6th Hydroinformatics Conference*, Liong S.-Y., Phoon K.-K., Babovic V. (Eds.), World Scientific Publishing: Singapore, Vol. 1, pp. 315–323.

Martin J. and O'Kane J.P. (2000) Development of a high resolution hydrodynamic flood mapping model using one-dimensional hydraulic models integrated with high resolution digital elevation models, GIS-1. *Hydroinformatics – 4th International Conference, IAHR ∼ AIRH*, IAHR: Iowa, p. 163.

Mynett A.E. (2004b) Living with floods in the Mekong Basin: on international cooperation and hydroinformatics technologies. *Proceedings of the International Conference on "Advances in Integrated Mekong River Management"*, Vientiane, 25–27 October, 2004, the Lao PDR, RR 2002-(6) Research Group.

Zagonjolli M., Mynett A.E. and Verwey A. (2005) Dam break modelling of Bovilla dam, near Tirana, Albania, Submitted to *Third International Symposium on Flood Defence*, Nijmegen, 25–27 May.

REFERENCES

Abbott M.B. (1992) The theory of the hydrologic model, or: the struggle for the soul of hydrology. In *Advances in Theoretical Hydrology*, O'Kane J.P., (Eds.), Elsevier, Amsterdam.

Abbott M.B. (1993) The electronic encapsulation of knowledge in hydraulics, hydrology and water resources, *Adv. Water Resources*, **16**, pp. 3–14.

Abbott M.B. and Refsgaard J.R. (1996) *Distributed Hydrological Modelling*, Kluwer, Dordrecht.

Babovic V. and Abbott M.B. (1997) The evolution of equations from hydraulic data, Part I: Theory, and Part II: Application. *Journal of Hydraulic Research*, **35**(3), 397–430.

Babovic V. and Keijzer M. (1999) Computer supported knowledge discovery – a case study in flow resistance induced by vegetation, *Proceedings of the XXVIII Congress of International Association for Hydraulic Research*, Graz.

Babovic V. and Keijzer M. (2000) Genetic programming as a model induction engine. *Journal of Hydroinformatics*, **2**(1), 35–60.

Chen Q. (2004a) *Cellular Automata and Artificial Intelligence in Ecohydraulics Modelling*, PhD thesis, UNESCO-IHE, Taylor & Francis Group plc, London, ISBN: 90 5809 696 3.

Chen Q. and Mynett A.E. (2004b) Predicting algal blooms along the Dutch coast by integrated numerical and fuzzy logic approaches. In *Proceedings of the 6th Hydroinformatics Conference*, Liong S.-Y, Phoon K.-K. and Babovic V. *et al.* (Eds.), Singapore.

Cunge J.A. (2003) Of data and models, *J. Hydroinformatics*, **5**(2), pp. 75–98.

Hardy T., Combs M. and Gowing I. (2004) Comparative Evaluation of Rapid Assessment Methodology for Instream Flow Assessments with Intensive Physical Habitat Simulation (PHABSIM) Approaches. *Proceedings of the 5th International Symposium on Ecohydraulics*, 12-17 September 2004, IAHR Publishing: Madrid.

Jonoski A. (2002) Hydroinformatics as Sociotechnology: Promoting Individual Stakeholder Participation by using Network Distributed Decision Support Systems, UNECO-IHE, Delft and Swets and Zeitlinger, Lisse.

Jorde K. (2004) Conceptual Framework for Assessment of Ecosystem Losses due to Reservoir Operations. In *Proceedings of the 5th International Symposium on Ecohydraulics*, 12-17 September 2004, IAHR Publishing: Madrid.

Kaiser-Hill, (2001) Model Code and Scenario Selection Report: Site-Wide Water Balance, Rocky Flats Environmental Technology Site, Report No. 01-RF- 0037, **http://www.dhisoftware.com/mikeshe/Download/RFETS_2-20-01.pdf**.

Keijzer M. and Babovic V. (1999) Declarative and preferential bias in GP-based scientific discovery. *Genetic Programming and Evolvable Machines*, **3**, 41–79.

KlemesV. (1986) Dilettantism in hydrology: tradition or destiny. *Water Resources Research* **22**(9) pp. 177–188.

Koza J.R. (1992) *Genetic Programming: on the Programming of the Computes by Means of Natural Selection*, MIT Press.

Martin J. (2002) *De-watering the Lower Feale – "A Virtual Water World"*, Ph.D. Thesis, National University of Ireland, Cork (University College Cork).

McGrath J., Barry K., O'Kane J.P. and Kavanagh R.C. (2003) High Resolution DEM and Sea Level Rise in the Centre of Cork City – Blue City Project. *Proceedings of the National Hydrology Seminar 2003 "Urban Hydrology: Stormwater Management"*. Office of Public Works, Dublin, **http://www.opw.ie/hydrology/index.asp**.

Migliori L. (2004) The hydrology and hydraulics of a pumped polder in North Kerry – a case study in hydroinformatics. Ph.D. Thesis. National University of Ireland, Cork (University College Cork).

Mynett A.E. (1999) *Art of Modelling – water systems in their natural environment*; inaugural address on Environmental Hydroinformatics, The International Institute for Infrastructural, Hydraulic and Environmental Engineering (IHE): Delft, The Netherlands.

Mynett A.E. (2002) Environmental hydroinformatics: the way ahead. In *Proceedings of the 5th Hydroinformatics Conference*, Falconer R.A. *et al.* (Eds.), IWA Publishing: London, Cardiff, Vol.1, pp. 31–36.

Mynett A.E. (2003a) Hydroinformatics in ecosystem restauration and management, *Final Report, Science, Technology & Management Panel, 3rd World Water Forum*, Kyoto-Osaka-Shiga.

Mynett A.E. (2003b) Feature extraction from remote sensing images for water resources management, *Proceedings of the International Conference on GIS and Remote Sensing in Hydrology, Water Resources and Environment*, Yichang.

Mynett A.E. (2004a) Environmental hydroinformatics tools for ecohydraulics modeling. In *Proceedings of the 6th Hydroinformatics Conference*, Liong S.-Y, Phoon K.-K. and Babovic V. (Eds.), Singapore, Vol. 1, pp. 13–23.

Mynett A.E., Chen Q. and Babovic V.M. (2004c) Artificial intelligence techniques in environmental hydrodynamics: the

role of expert knowledge, Keynote address, *Proceedings of IAHR Asian Pacific Division Environmental Hydrodynamics Conference*, Hong Kong, 15–18 December.

O'Kane J.P. and Migliori L. (2004) "The hydrology and hydraulics of a pumped polder in North Kerry – a case study in hydroinformatics". *Proceedings National Hydrology Seminar on "The Water Framework Directive – Monitoring & Modelling Issues for River Basin Management"*. Office of Public Works: Dublin, **http://www.opw.ie/hydrology/**.

Russel S.G. and Norvig P. (1995) Artificial Intelligence: A Modern Approach, Prentice Hill, New Jersey.

Rodriguez J.R. (2004) *Evaluation of Physically Based and Evolutionary Data Mining Approaches for Modelling Resistance due to Vegetation in SOBEK1D-2D*, M.Sc thesis HH485, UNESCO-IHE, Delft, The Netherlands.

Uittenbogaard R. (2003) Modelling turbulence in vegetated aquatic flows, *Riparian Forest Vegetated Channels Workshop*, Trento.

15: Digital Elevation Model Analysis and Geographic Information Systems

GLENN E MOGLEN[1] AND DAVID R MAIDMENT[2]

[1]Department of Civil and Environmental Engineering, University of Maryland, College Park, MD, US
[2]Department of Civil Engineering, University of Texas, Austin, TX, US

This article presents the technology of Geographic Information Systems (GIS) applied to a particular form of representation of topography, the digital elevation model (DEM). It first defines what a DEM is and introduces the basic properties of all DEMs, as well as several common sources of DEM data. Then the basic concepts of GIS technology are presented, defining what a GIS is, what fundamental data types it can store, and what kind of manipulations it can perform. Some of the most common GIS software packages are discussed. The remainder of the article focuses on the use of GIS technology for the interpretation of DEM in a hydrologic context. The principles of flow direction, flow accumulation, watershed delineation, flow lengths, slopes, and travel times as derived from DEM data within a GIS are presented. A brief case study is presented illustrating a number of the concepts presented in this article.

INTRODUCTION

The hydrologic sciences depend on quantifying the characteristics of watersheds as derived from topographic maps. Such maps reveal both the horizontal extent of the watersheds as well as their relief and drainage structure. This article focuses on two relatively recent technological innovations that allow the interpretation of topography to be performed quickly, systematically, and at a very detailed level. In a sense, this article does not contribute any new ideas to the hydrological sciences. Rather, the focus here is on cutting edge software and data products that work together to provide the hydrologist the information necessary to apply fundamental hydrologic concepts at a high spatial resolution and over a large spatial extent. Present-day technology allows these concepts to be applied quickly and uniformly so that the hydrologist can focus on the physical system, while the drudgery of the mechanics of the analysis is taken care of by the computer.

The first innovation in technologies that has emerged over the recent decade or so, the development of digital elevation models or DEMs, is in essence nothing more than a digital representation of a topographic map. Scientists and engineers have worked with paper forms of topographic maps for a long time. Digital representations of topography are even more powerful because algorithms allow for the automated interpretation of the topography, rendering results reproducible between different individuals.

The second innovation is the technological advance brought on by the advent of the geographic information system or GIS. This software technology provides a common structure within which spatial data can be stored, displayed, and analyzed. It is not necessary to use a GIS to handle DEM data, but most GIS packages provide standard algorithms for the hydrologic interpretation of DEMs. These algorithms automate the determination of flow directions, drainage areas, watershed delineation, channel slopes, and travel distances and time. The GIS also includes visualization tools allowing the hydrologist to see the spatial distribution and organization of all of the above quantities. This visual component aids in the development of an intuitive understanding of the systems being studied, and in the presentation of information to others.

This article gives a broad overview of the basics of DEMs. The focus is on the standard characteristics of all DEMs and on sources for obtaining this type of data. A

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

very natural way to interact with DEM data is through the use of GIS technology; so this article will also provide a brief introduction to this technology focusing on the central concepts common to all geographic information systems. The main body of this article will provide a detailed description of how DEM data is interpreted in a hydrologic context within the GIS. A case study illustrating a packaged system that integrated DEM data within a GIS environment is demonstrated in the final section of this article, briefly illustrating the kind of the DEM-based analysis that can be automated within a GIS. Relevant software links are also provided in the article.

DIGITAL ELEVATION MODELS

A digital elevation model (DEM) is a raster-based presentation of topography. Each cell of the DEM contains a representative value of the elevation related to the areal extent of the cell.

DEM Data Sources and Products

In the United States, the standard digital elevation model is the National Elevation Dataset (NED) (USGS, 2003a). This is a seamless 1-arc second (30 m) grid whose elevation values are expressed in floating point meter units. It was compiled by converting to a grid the topographic contours from standard 1 : 24 000 maps sheets covering the nation, and merging the resulting grids and filling in small gaps in coverage between the map sheets. NED data can be obtained for any user-selected area from a seamless data server (**`http://seamless.usgs.gov/`**).

A series of hydrologic data products called the *Elevation Derivatives for National Application* (EDNA) have been derived from the National Elevation Dataset (**`http://edna.usgs.gov/`**) (USGS, 2003b). These include the flow direction and flow accumulation grids, a stream network defined for cells with more than 5000 cells upstream, and a catchment set with one catchment for each stream reach. The average size of these catchments is approximately 8 km^2, and they form with the stream network a dendritic representation of the drainage network of the nation.

In a similar manner to the National Elevation Dataset, Figure 1 shows the United States Geological Survey's (USGS) EROS Data Center has created a 30 arc-second (1 km) digital elevation model of the earth called *GTOPO30* (**`http://edcdaac.usgs.gov/gtopo30/gtopo30.html`**) (USGS, 2003c). This dataset was produced by converting the topographic information from the 1 : 1 000 000 scale Digital Chart of the World, and supplemented with other data sources.

The EROS Data Center has compiled a series of hydrologic derivative products called *HYDRO1K* for GTOPO30, using the same process as for the Elevation Derivatives

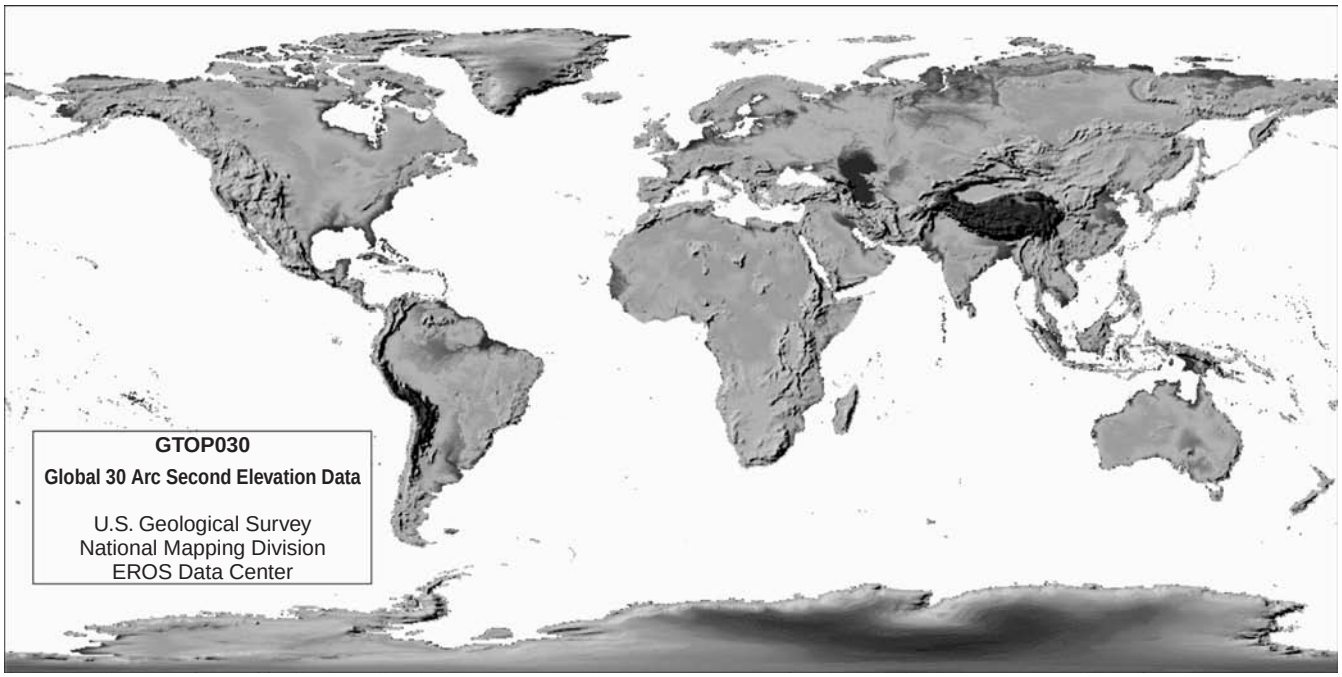


Figure 1 The GTOPO30 elevation dataset produced globally by the United States Geological Survey (Data available from United States Geological Survey, EROS Data Center, Sioux Falls, SD; source: http://edcdaac.usgs.gov/gtopo30/dem_img.html). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

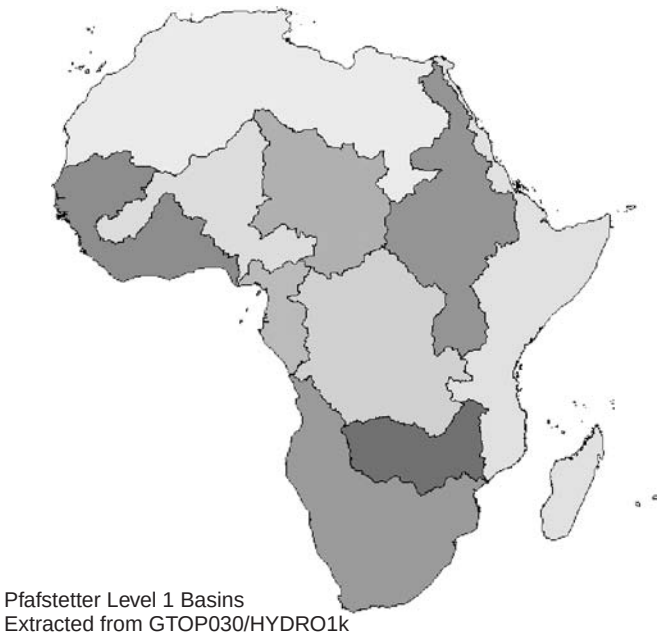


Figure 2 Level 1 drainage basins in Africa determined from GTOPO30 elevation dataset (Data available from United States Geological Survey, EROS Data Center, Sioux Falls, SD; source: http://edcdaac.usgs.gov/gtopo30/hydro/af_basins.html). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

for National Application (see Figure 2). Thus, DEM-based hydrology can be applied at a coarse resolution anywhere in the world.

The level of accuracy of global terrain information was dramatically increased by the Shuttle Radar Topography Mission (SRTM) (NASA, 2003), which took place from February 11–22 in the year 2000 (**http://www.jpl.nasa.gov/srtm/**). Figure 3 shows an image from this SRTM dataset.

Earth surface terrain was sensed by using two radar emitters mounted at a distance from one another on the space shuttle, where each sent out radar waves and received reflected waves from the land surface. By using radar interferometry, a principle similar to the human eyes being able to determine depth by coordinating two parallel images, the radar images are converted into terrain grids, at 30 m resolution for the United States and 90 m for the rest of the earth, although for scientific purposes 30 m data can be obtained for regions outside the United States.

DEMs are useful for hydrologic analysis through the many automated procedures commonly incorporated in GIS software. These procedures allow the hydrologist to, among other tasks, interpret flow paths, delineate watersheds, derive channel networks, and measure slopes.

Figure 3 SRTM image of the Bosphorus Strait, Turkey (Courtesy NASA/JPL-Caltech; Source: http://photojournal.jpl.nasa.gov/catalog/PIA03349). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

GEOGRAPHIC INFORMATION SYSTEMS

A geographic information system (GIS) is an assembly of geospatial data and functions by operating on those data whose purpose is to describe the landscape and facilitate its analysis. Geospatial data describe the earth using a geographically referenced set of spatial coordinates. When the term "hydrologic data" is used, hydrologists have long understood this term to mean data on hydrologic observations, such as rainfall or streamflow. As the use of GIS in hydrology has matured, the term hydrologic data has broadened to include geospatial data describing the hydrologic environment, including land surface terrain, soils, land use, river networks, water body hydrography, and the like. Of these geospatial data, terrain information in the form of digital elevation models has had the greatest impact because it enabled for the first time an automated delineation of an interrelated set of watersheds and stream networks, and thus created the basis for a greater degree of detail and spatial scope of hydrologic modeling than had been possible previously.

Geospatial data can describe the earth in either continuous or discrete space. Digital elevation models are an example of a continuous spatial representation, in which the spatial domain is covered by a regular grid or raster of square cells, in each cell of which is stored a value, in this case the land surface elevation. This representation is also appropriate for other continuous spatial variables

in hydrology, such as rainfall or evaporation. A discrete space or vector representation is used where the landscape is subdivided into distinct spatial units, such as watersheds, where each unit has a boundary that divides it from its neighbors. The boundary is represented by a connected set of lines that close to form an area or polygon feature. Similarly, stream networks are defined by connected sets of line features, where a line is formed as a sequence of points. Gauging stations are represented as point features defined by a single pair of (x, y) coordinates. The term vector data arises from the fact that a vector can be drawn from the origin (0,0) to any (x, y) point, and thus from one point to the next in a line, and so on. A comprehensive hydrologic database of a region contains both raster and vector data. Conversions between these two different data formats can be made (see Figure 4), such that a single raster cell corresponds to a point, and a set or zone of raster cells corresponds to a line or an area. Thus, watersheds and stream networks can be defined by raster DEM analysis and then converted to vector format for further analysis.

GIS databases are built using layers of data representing different types of information (see Figure 5). This concept is derived from the traditional process of map printing in which the "blue lines" represent hydrography, the "brown lines" represent topography, and so on. Each of these layers of information is prepared on a separate sheet and maps are printed by repeatedly running the map sheet through a press and applying the different colors in succession. By extension, each data "theme" in a GIS database is represented as a separate layer in which all the geographic features in the landscape of that type are gathered.

However, GIS is more than just a digital map – the features also have data or "attributes" associated with them. Thus, the name of the watershed, its drainage area, average slope, and rainfall can be attached as attributes of a watershed data layer (see Figure 6). Attributes are collected in rectangular tables, in which each row represents one spatial feature, and each column or field represents an attribute. These tables can in turn be related to other tables containing more details where necessary. Thus, a GIS integrates both geospatial and tabular information about hydrologic features of the landscape.

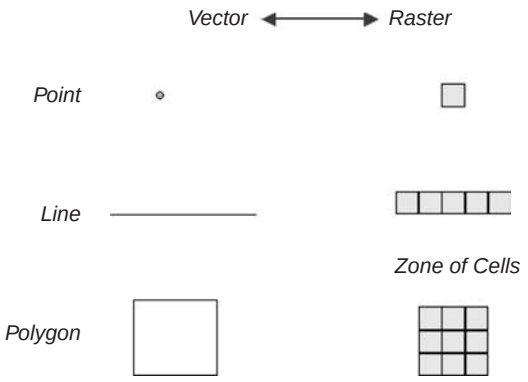


Figure 4 Comparison of vector and raster representations of point, line, and polygon geographic features. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

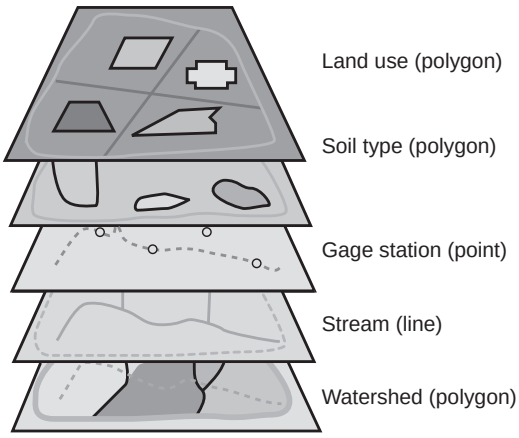


Figure 5 Illustration of multiple GIS layers at a given location depicting various forms of hydrologically relevant data. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The most widely used GIS products are the ArcView and ArcInfo systems produced by the Environmental Systems Research Institute (ESRI) (**http://www.esri.com**) of Redlands, California (ESRI, 2003). These systems have recently been merged into a new form called *ArcGIS*, and a customization of this system for water resources has been developed called *Arc Hydro* (Maidment, 2002) (**http://gis.esri.com/esripress/display/index.cfm?fuseaction=display&websiteID=56&moduleID=0**). Arc Hydro integrates the various data layers together using database relationships, and also integrates geospatial and temporal water resources information into a single database for the first time in a GIS setting. This lays the foundation for the development of hydrologic information systems, which synthesize water resources data and hydrologic models to describe the functioning of hydrologic processes.

Map Projections

The earth is spherical in shape but maps are flat, so to relate locations on the earth in latitude and longitude coordinates to locations on a map in (x, y) coordinates, a coordinate transformation is needed. This involves first a reduction in scale from the earth to a globe (a map scale of 1 : 100 000 means that 1 mm on the map corresponds to 100 000 mm or

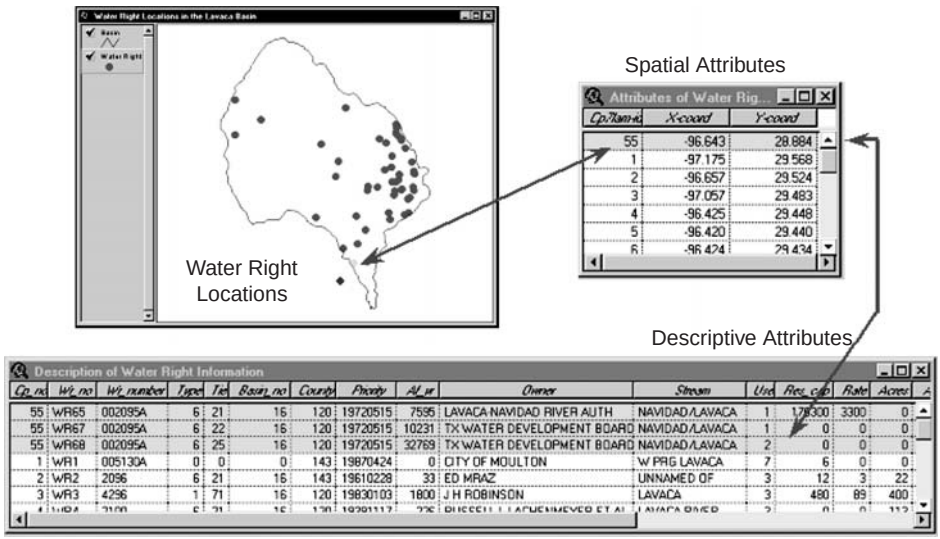


Figure 6 Illustration of the relationship between geospatial and tabular data as managed within a GIS environment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

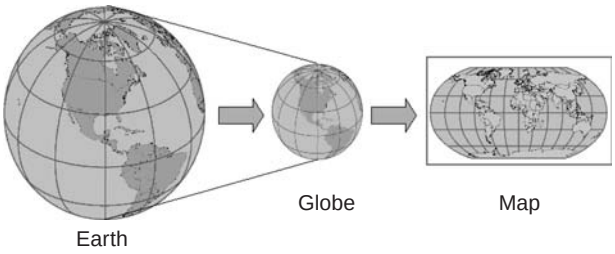


Figure 7 Illustration of concept of map projections. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

100 m on the earth's surface), and then a projection from a curved globe to a flat map (see Figure 7). Hydrologic analysis of digital elevation models requires map projection because its computations depend on the slope between cells, and thus the (x, y, z) coordinates must all be expressed in length units. For hydrologic analysis of digital elevation models, the most common map projection is the Albers Equal Area projection, which preserves true earth surface area on the projected map.

A good freeware program that illustrates the concepts of map projections and coordinate systems is produced by the United States Army Corps of Engineers – Topographic Engineering Center. The program, called *CORPSCON*, may be obtained at `http://crunch.tec.army.mil/software/corpscon/corpscon.html` (USACE, 2003a). Although this is not, strictly speaking, GIS software, this program is useful in gaining an appreciation for the concepts of map projections and may be used to project a point or a series of points from one projection/coordinate system to another.

HYDROLOGIC INTERPRETATION OF A DEM

DEMs may be employed in many different contexts depending on the user's intended application. Interpretation of the DEM varies naturally along with this application. For instance, a road designer or land developer would tend to interpret the elevations reported by the DEM at face value, and compare these predevelopment DEM elevations against the desired postdevelopment elevations to determine "cut and fill" requirements. In contrast, the hydrologist views the DEM as a surface that naturally sheds water and directs its flow to channels located at relatively low points in the topography.

Flow Direction

There are numerous algorithms employed to interpret DEMs. Probably, the most common algorithm is the "D8" algorithm, where the flow direction for every cell within the watershed is determined by considering the surrounding eight neighboring cells. The local slope in each of the eight directions of these neighboring cells is calculated by taking the difference in elevation indicated by the DEM value at each of these eight neighboring locations and the value at the cell being examined. This difference in elevation is then divided by the center-to-center distance between these cells (this distance will be the cell size in the cardinal directions and the cell size $\cdot\sqrt{2}$ in each of the diagonal directions). The direction that yields the steepest downhill slope is the inferred direction of water flow. The concept of flow direction is very important because it allows for the inference of drainage areas, flow lengths, and the automated delineation of watersheds as will be illustrated later in this article.

13	11	12
9	10	11
6	6	8

Figure 8 Small sample 3 × 3 DEM to illustrate calculation of flow direction

Consider the 3 × 3 DEM in Figure 8 with a horizontal resolution of 30 m and reported in vertical units of meters.

Notice that in general, the elevations decrease from the top row to the bottom row so the general sense of flow that we anticipate is from the top to the bottom of Figure 8. Let us calculate the slope in each of the eight directions from the central cell:

Due East:

$$\text{Slope} = \frac{(10-11)}{30} = -0.0333 \tag{1}$$

Northeast:

$$\text{Slope} = \frac{(10-12)}{30}\sqrt{2} = -0.0471 \tag{2}$$

Due North:

$$\text{Slope} = \frac{(10-11)}{30} = -0.0333 \tag{3}$$

Northwest:

$$\text{Slope} = \frac{(10-13)}{30}\sqrt{2} = -0.0707 \tag{4}$$

Due West:

$$\text{Slope} = \frac{(10-9)}{30} = 0.0333 \tag{5}$$

Southwest

$$\text{Slope} = \frac{(10-6)}{30}\sqrt{2} = 0.0943 \tag{6}$$

Due South:

$$\text{Slope} = \frac{(10-6)}{30} = 0.1333 \tag{7}$$

Southeast:

$$\text{Slope} = \frac{(10-8)}{30}\sqrt{2} = 0.0471 \tag{8}$$

13	14	14	13	12
12	13	15	14	11
12	11	13	10	10
10	10	12	11	9
8	9	11	8	10

Figure 9 Sample 5 × 5 DEM used to illustrate hydrologic interpretation methods throughout article

Notice that the slopes in the east, northeast, north, and northwest directions are negative (indicating that the flow directions are uphill) from the center cell. These flow directions can be immediately discarded as possible directions for this particular cell. The largest positive slope is in the due south direction with a value of 0.1333. This is the inferred flow direction for the center cell. The reader will notice that the elevation in the southwest direction is the same as in the due south direction, however, because of the greater center-to-center distance, the slope in southwest direction is slightly less steep. Owing to the $\sqrt{2}$ factor, it is even possible in some cases for the elevation in one of the cardinal directions to be slightly higher than in a diagonal direction and yet, the steepest slope may still correspond to the cardinal direction. Let us examine a slightly larger example shown in Figure 9.

Consider the 5 × 5 DEM in Figure 9 where we add the additional stipulation that the flow cannot leave the boundaries of the 5 × 5 grid (i.e. it is as if there are vertical walls bounding the overall grid). Before solving for the flow directions, the following observations can be made concerning this grid:

- As in the 3 × 3 grid, there is a general trend of decreasing elevations as we scan from the top rows to the bottom rows (therefore we would expect that most of the drainage directions will have a southerly component: south, southeast, or southwest).
- The middle column has relatively higher elevations than those in adjacent columns (therefore we would expect that this column will act as a divide between the right and left sides of the grid).
- There is a tie for the lowest elevations in the grid with a value of "8". We will expect these two locations to act as "sinks" for this system.

The reader should be able to verify that the flow directions are as those shown in Figure 10.

Notice that each cell has only one flow direction. However, some cells, for example, cell $C4$, receive flow from cells $B3$, $B4$, and $C3$. Cells $E1$ and $E4$ were the locations of a tie for the lowest elevation. The flow directions here

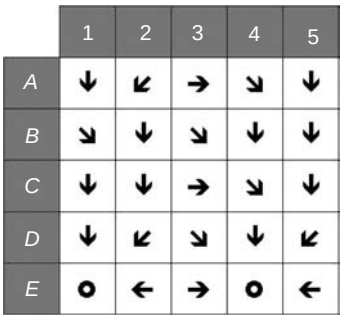


Figure 10 Arrows indicate visually the flow directions throughout the 5 × 5 sample DEM

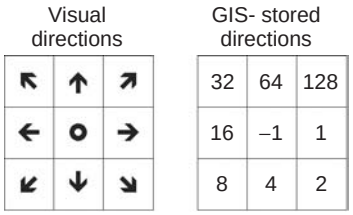


Figure 11 An illustration of the transformation of visually meaningful flow direction to GIS-stored information

are undefined (or these cells can be considered to be sinks). Here, we indicate their flow direction (or lack of one) by the "○" symbol. The reader should notice that the other observations made before determining flow directions have been borne out. There is a general tendency for southerly flow and there are no arrows (flow directions) that cross through the plane separating columns 2 and 3, confirming the existence of a drainage divide between these columns.

The arrows indicating flow directions are useful for the human eye to interpret, but unfortunately, the GIS needs some other method to "understand" which way the flow is going at a given cell. The method used to denote these flow directions is arbitrary. The ArcView GIS uses powers of 2 increasing clockwise from due east: 1, 2, 4, 8, 16, 32, 64, and 128. We add the numerical value, −1, to indicate the existence of a sink or undefined flow direction as shown in Figure 11.

Applying this GIS's nomenclature for flow directions, we obtain the flow direction matrix for the 5 × 5 DEM example shown in Figure 12.

It is worthwhile to briefly comment on some practical difficulties that must often be overcome in determining flow directions from a DEM. The D8 algorithm presented here assumes that flow directions always exist. In areas where the topography is flat or if the vertical resolution of the DEM is poor, it is not unusual to have many adjacent cells all with the same elevation. These are referred to as "flats" and secondary rules, beyond those illustrated above, must be incorporated in order to determine flow directions in

	1	2	3	4	5
A	4	8	1	2	4
B	2	4	2	4	4
C	4	4	1	2	4
D	4	8	2	4	8
E	–1	16	1	–1	16

Figure 12 GIS-stored flow directions for 5 × 5 sample DEM

these areas. These rules are somewhat arbitrary, but they generally work by taking a more broad-ranging view of the DEM and assign flow directions in flat areas consistent with these larger scale tendencies in the topography. A more severe problem in hydrologically interpreting DEMs is the presence of "pits" or local low points in the topography. In the case of pits, the assumption of the flow direction algorithm is that all flow drains off the edge of the DEM defined extent of the topography. Thus, any internal pits must somehow be removed. The typical algorithm is to fill these pits until the elevation within the pit is the same as that of the lowest neighboring cell that drains elsewhere. These pits then become flats that are resolved as described above. A by-product of the flow direction algorithm, not shown here, is usually an updated DEM in which the pits have been removed by this filling process.

Flow Accumulation

Once flow directions are determined, this information is used to determine the flow accumulation. Adjusted to proper units, the flow accumulation is synonymous with drainage area. We begin by defining a two-rule algorithm for determining flow accumulation at any cell:

1. If the cell has no neighboring cells draining to it, a value of "1" is assigned.
2. If the cell receives drainage from any of the eight immediate neighboring cells, it is assigned the value of "1" plus the sum of the flow accumulation draining from each of these neighbors.

Notice that Rule 2 amounts to a recursive definition. In practice, one must start determining flow accumulation at the upstream end of all flow paths and work downstream, otherwise the drainage area of all neighboring cells to the cell in question may not be known. Rules 1 and 2 are repeated across the entire DEM. Flow accumulation is a powerful GIS capability because calculating it as a spatially distributed quantity allows us to determine drainage area not at just one point, but at *any* point within the domain of the original DEM field.

	1	2	3	4	5
A	1	1	1	2	1
B	3	1	1	1	4
C	1	5	1	4	5
D	2	6	1	1	10
E	10	1	1	15	1

Figure 13 Flow accumulation for 5 × 5 sample DEM. Cells that have a flow accumulation of 5 cells or greater are shaded here to illustrate the hypothetical beginnings of the channel network. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Having determined the flow directions, we are now ready to determine the "flow accumulation" for the 5 × 5 sample DEM illustrated in the previous section. Applying Rules 1 and 2 to the 5 × 5 DEM results in the solution shown in Figure 13.

Let us examine several cells in the above example to see how they illustrate the flow accumulation algorithm rules. Note that cell $A1$ has no arrow (flow) going into it – therefore Rule 1 applies and this cell receives a value of "1". In contrast, cell $D5$ receives flow from cells $C4$ and $C5$, which carry flow accumulations of 4 and 5 cells respectively. Therefore, the flow accumulation for cell $D5$ is $4 + 5 + 1 = 10$. Notice that the values in our two sink cells, cells $E1$ and $E4$, are 10 and 15 respectively. The sum of these two values is 25, which is equal to the area (in cells) of our 5 × 5 DEM grid. Further, notice that the value at $E1$ (10) is the area of this 2 column by 5 row "watershed" (recall the divide between columns 2 and 3 noted earlier) and the value at $E4$ (15) is the area of this 3 column by 5 row "watershed". It is always the case that the flow accumulation at the outlet cell of a watershed is equal to the drainage area (in cells) of the watershed it is draining. If we know the resolution of the DEM being examined (e.g. 30 m cells), then the drainage area of the rightmost or eastern watershed in this 5 × 5 DEM is:

$$\mathrm{DA} = 15 \text{ cells} \times \frac{(30\mathrm{m})^2}{\text{cell}} = 13\,500\,\mathrm{m}^2 \qquad (9)$$

A simple way to infer the location of streams or rivers is to have the GIS indicate all cells with a flow accumulation in excess of a fixed value. This fixed value physically corresponds to a source area and is interpreted as the minimum source area needed to support the existence of a stream. This approach is illustrated for the 5 × 5 sample DEM. A hypothetical minimum source area of five cells was applied to this DEM, and those cells draining this area or greater are shown shaded in Figure 13. At this scale it is difficult to envision, but these two shaded lengths can be physically interpreted as the very beginnings of the channel network. In the examples and case study to follow where a drainage network is shown, this is the approach used to infer the location of the drainage network.

We should note that the above flow accumulation algorithm is premised on the D8 model of flow directions. If a more complicated flow direction algorithm is used such as those developed by Costa-Cabral and Burges (1994) or Tarboton (1997), then determining flow accumulation is also somewhat more complicated. The main difference in the outcomes from these other algorithms is that the resulting flow accumulation will no longer work out to be in integer counts of cells, but instead flow accumulation will generally be a real quantity. In any case, the physical interpretation remains the same that flow accumulation corresponds to drainage area.

Watershed Delineation

Flow accumulation and watershed delineation are closely related concepts. This idea was alluded too briefly in the 5 × 5 sample DEM illustration, but merits further discussion here. In the above section, a claim was made that the flow accumulation at the outlet cell of a watershed is equal to the drainage area (in cells) of the watershed it is draining. In other words, if we can identify all the cells that are draining to a common location or outlet, this set of cells represents the watershed associated with that outlet. In the days before GIS and the availability of DEM data, watershed delineation often required painstaking and time-consuming manual effort on the part of the hydrologist to determine watershed boundaries, hand-drawing curves emanating from the watershed outlet, and always traveling perpendicular to contour lines of elevation. Today's commonly available GIS-based tools have reduced the efforts of watershed delineation to a simple "mouse-click" at the location of the desired watershed outlet, as will be illustrated in the case study section to follow.

Flow Length

The flow length is the distance from any point in the watershed to the watershed outlet. This distance is measured along the flow paths determined from the topography not "as the crow flies". In GIS, the flow length of an arbitrary cell is determined by summing the incremental distances from center-to-center of each cell along the flow path from the selected cell to the outlet cell. The flow length assigned to the outlet cell is zero.

The concept of flow lengths is important to hydrologists. When it rains, a drop of water landing somewhere in the basin must first travel some distance before reaching the outlet. Assuming constant flow velocities (an assumption we will relax later), the cell with the greatest flow length to the outlet represents the hydrologically most remote cell.

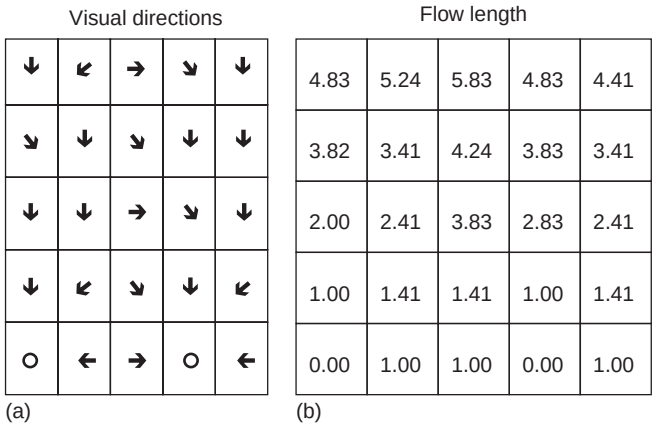


Figure 14 Visual flow directions and flow lengths for the 5 × 5 sample DEM

Its flow length divided by the flow velocity represents a representative lag time for the basin. The lag time quantifies how long before the entire basin is contributing to surface flow at the outlet, and is a representative timescale for the basin. Returning to the 5 × 5 DEM sample used earlier, let us examine this sample in terms of flow lengths.

The left grid in Figure 14 shows schematically the flow directions that were determined from the original DEM. If the resolution of the original DEM is 1 unit, then the right grid shows the downstream flow lengths from each cell to the grid outlets. (Recall that we have two outlets as shown by grid cells shown with a ○ symbol.) The flow lengths at the outlet cells themselves have been set to zero.

There are three ways that we might choose to look at travel times within the GIS environment:

1. As a mapped iso-distance ("contour" lines of travel length).
2. As a frequency distribution.
3. As the longest flow path within the watershed.

Considering the first perspective enumerated above, a map of shaded iso-distance lines is shown in Figure 15 for a small watershed near Washington DC, measured at 30 m resolution.

As would be expected, the smallest flow lengths occur in the proximity of the watershed outlet and generally increase away from the outlet. However, notice that the shaded areas do not radiate out in concentric circles from the watershed outlet, but instead have jagged edges and are organized along the drainage paths of main channels within the watershed as indicated by the dark lines within the figure. This perspective is useful for gaining insight into the overall spatial distribution of flow lengths within the watershed and how these flow lengths are organized along the principal channels within the watershed.

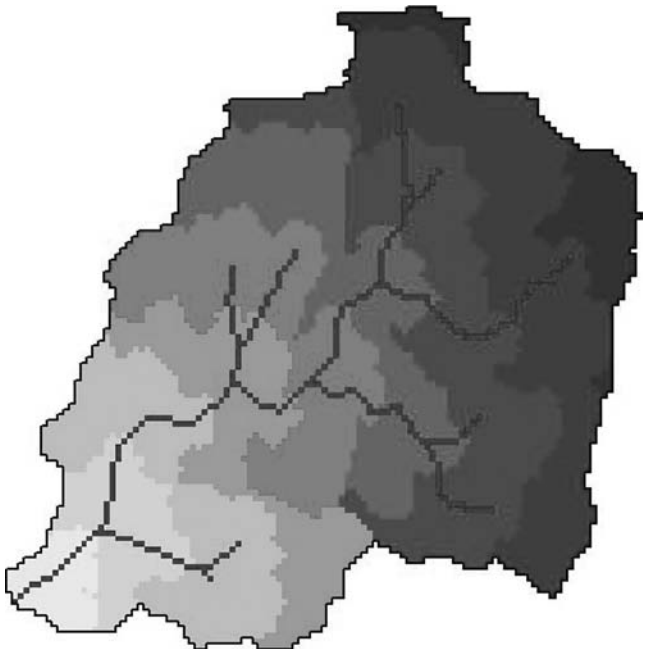


Figure 15 Shaded map of flow lengths in a small watershed. Darker shades correspond to longer flow lengths. The inferred drainage network is shown to help in visualizing the overall organization of flow paths. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The second perspective to analyze flow lengths involves creating a frequency distribution of travel lengths with each cell within the watershed contributing a single travel length "observation". This distribution, first discussed by Surkan (1968) is referred to in the geomorphology literature as the width function. A coarse frequency plot corresponding to the flow lengths shown in Figure 15 is provided in Figure 16.

The overall shape and time base of this distribution is often strongly correlated with the unit hydrograph for the same watershed. We will return to this idea later as we examine travel times in contrast to the flow lengths we are considering here.

The third and final perspective we will consider is the location of the longest flow path within the watershed. This is the path that the flow from the cell that is located the greatest distance from the watershed outlet will take as it drains to this outlet. The length of this longest flow path is indicated by the upper-bound of the last bin of the frequency distribution and its location is hinted at by the contour lines of iso-distance. The actual path is determined readily using GIS techniques and can be likened to placing a drop of water on this most remote cell and then tracing its path as it flows towards the watershed outlet. Figure 17 shows this longest flow path for the watershed examined for the previous two perspectives.

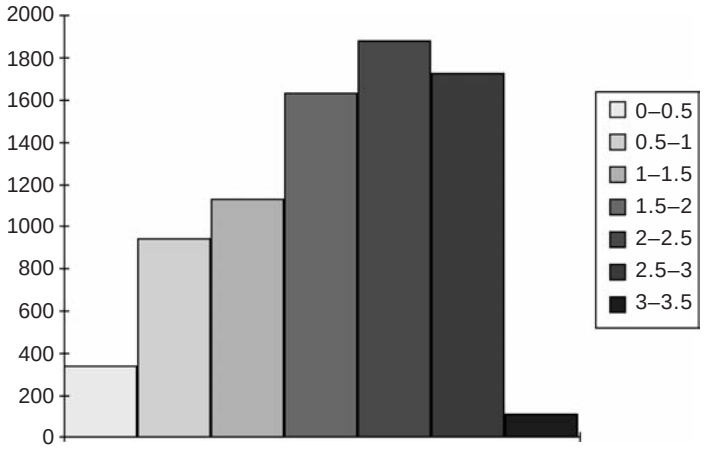


Figure 16 Frequency distribution of flow lengths for watershed pictured in Figure 15. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 17 The longest flow path within the watershed pictured in Figure 15. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Velocity and Travel Times

In the previous section, we considered the frequency distribution of flow lengths within the watershed and noted that this is essentially the width function defined by Surkan (1968). The shape of the frequency distribution shown in Figure 16 bears some similarity to a hydrograph, although its negative skew is counter to the shape typically encountered by hydrologists.

Let us assume a uniform velocity throughout the watershed; another interpretation of this width function is that it is synonymous with the "Instantaneous Unit Hydrograph" or IUH. The IUH is defined as the hydrograph that would be observed at the watershed outlet if a unit pulse of water were instantaneously placed uniformly over the entire watershed at a given instant. With the travel lengths known and a single uniform velocity of flow observed throughout the watershed, the travel time ($t_{t,i}$) to the outlet for any randomly chosen cell, i, would be given by:

$$t_{t,i} = \frac{d_i}{v} \tag{10}$$

where d_i is the distance from the ith cell to the watershed outlet and v is the uniform watershed flow velocity. This is a powerful concept, because we have now used the DEM to discern the flow organization of the watershed and its unique hydrologic signal, the unit hydrograph, which is dependent on the watershed size, shape, and connectivity. In this way, we can determine a unique unit hydrograph for any watershed, being no longer bound by the dimensionless hydrographs offered by the Natural Resources Conservation Service (NRCS) (SCS, 1985), Clark (1943), or the Snyder (1938) model, to name a few.

We note that this uniform velocity assumption simply amounts to a scaling factor on the width function. We are apparently no closer to a typical hydrograph shape than we were before. What have we neglected? Immediately, the reader may recognize that we have assumed a uniform velocity everywhere throughout the watershed. In reality, we would expect much greater velocities in the channels than on the hillslopes. This velocity difference can either be conceptualized as a difference in the Manning's n roughness of the flow surface encountered on hillslopes versus channels, or as the difference between surface runoff in the channels and subsurface flow on the hillslopes. In any event, we expect the difference in velocities could easily be on the order of 10 to 100 times greater in the channels than on the hillslopes. Let us rewrite equation (10) now with this

Flow accumulation		
1	2	1
1	1	4
1	4	5
1	1	10
1	15	1

Travel time		
3.66	2.66	2.24
2.97	2.56	1.24
2.56	1.56	0.24
1.41	1.00	0.14
1.00	0.00	1.00

Figure 18 Flow accumulation and travel times for rightmost three columns of sample 5×5 DEM. Shaded cells are interpreted here to correspond to the very head of a channel. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

two-velocity model:

$$t_{t,i} = \frac{d_{\mathrm{H},i}}{v_{\mathrm{H}}} + \frac{d_{\mathrm{C},i}}{v_{\mathrm{C}}} \tag{11}$$

where $d_{\mathrm{H},i}$ and $d_{\mathrm{C},i}$ are the flow distances for cell i from along the hillslope and channel portions of the overall flow path, respectively. Let us return to the rightmost three columns of the 5×5 DEM analyzed earlier. We will define a channel as any cell with a cumulative area of five cells or greater. For simplicity, we will take each cell to have a side length of 1 unit and use a channel velocity (v_{C}) of 10 units-t^{-1}, and a hillslope velocity (v_{H}) of 1 unit-t^{-1}. The result is shown in Figure 18. The left side of Figure 18 repeats the flow accumulation grid for these columns and the right side of Figure 18 shows the distribution of travel times. Notice the very small travel times associated with the cells draining 5 and 10 cells towards the bottom right of the grid. These two cells, by virtue of equaling or exceeding the channel threshold of 5 cells, develop velocities that are 10 times greater than any other cells in the grid. The result is that the travel times to the outlet associated with these two cells are greatly diminished. It is instructive to compare these travel times to the flow lengths (which can be equated to travel times for a uniform velocity of 1 unit-t^{-1} throughout) shown previously in Figure 14. Notice the much smaller values in Figure 18 for the rightmost three columns compared to those in Figure 14. If we think of the watershed as a system whose objective is to drain water as quickly as possible, then the presence of channel cells with high travel velocities represents an efficiency within the system as indicated by the small travel times associated with these cells.

Figure 19 illustrates the distribution of travel times resulting from the application of equations (10) and (11) to a small watershed draining to Rock Creek immediately north of Washington DC.

The left image corresponds to either the single velocity model of equation (10) applied to the watershed, or to equation (11) with $v_{\mathrm{C}} = v_{\mathrm{H}}$. The right image corresponds to the case where the channel velocity, v_{C}, is 20 times greater than the hillslope velocity and the channels have been defined according to a minimum source area concept, requiring $250 - 30$ m cells to form a channel. The channel network is indicated by the dark cells in both images. The reader should note the strong tendency in the right image for small travel times to cluster around the drainage network. This tendency is absent in the left image where, because of the single velocity applied throughout, the channel network does not play an accentuated role in the spatial distribution of travel times. Figure 20 provides the frequency distributions of travel times for the two cases shown in Figure 19.

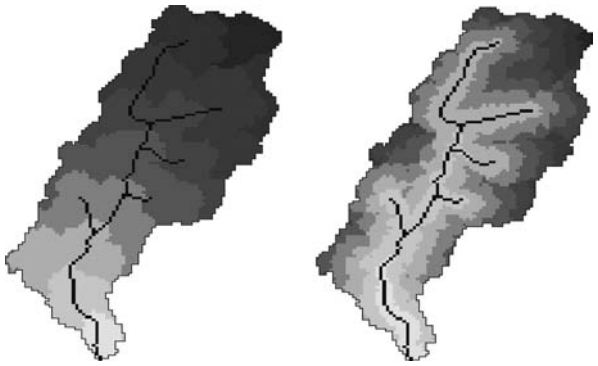


Figure 19 Isochrones of travel time for a small watershed draining to Rock Creek immediately north of Washington, DC. Image (a) corresponds to a single velocity applied uniformly across the entire watershed. Image (b) corresponds to the two-velocity model where channels flow faster than hillslopes. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The use of the two-velocity model has had the effect of making the skew of the travel time distribution more positive. Notice also that since the single velocity model quantified in Figure 19(a) employed a large velocity throughout the watershed, the overall time base of this distribution is considerably shorter than in Figure 19(b) where the hillslope velocity was 20 times smaller than the channel velocity.

Watershed and Channel Slopes

The interpretation of the DEM to this point has largely focused on horizontal information contained in the DEM. Vertical information was used to determine flow directions, which have served to determine all subsequent information such as watershed boundaries, drainage areas, and flow paths. We return now to examining the vertical component of the watershed as described by the DEMs.

Hydrologists have historically quantified the vertical characteristics of the watershed in a number of ways. In the context of a GIS analysis, we will examine several such characterizations here: channel slope, watershed slope,

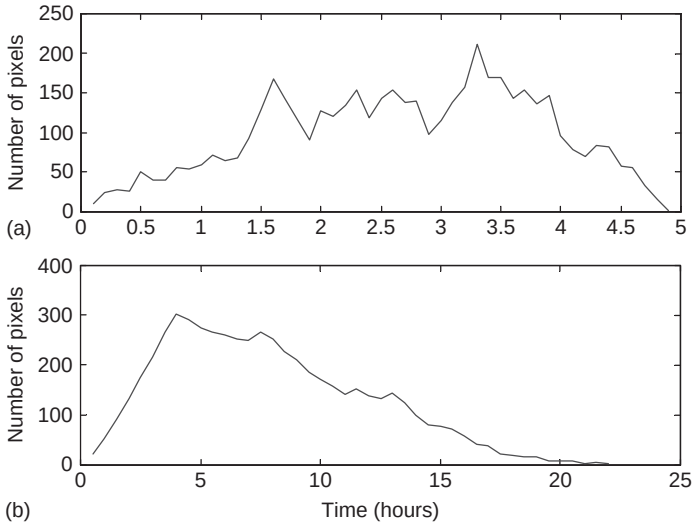


Figure 20 Frequency distribution of travel times for the two cases shown in Figure 19. Images (a) and (b) of Figure 19 result in the frequency distributions shown in (a) and (b), respectively. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

land slope, and basin relief. We start with the channel slope, which is defined as the difference in elevation from the upstream to the downstream end of the main channel divided by the length of the main channel:

$$S_c = \frac{\Delta E}{L_c} \tag{12}$$

where ΔE is the change in elevation from the upstream to downstream end of the main channel and L_c is the length of the main channel. From a hydrologic standpoint, the calculation of the channel slope as defined by equation (12) is straightforward. However, from a GIS perspective there is an ambiguity to consider. One may readily assume the outlet of the watershed to indicate the downstream extent of the channel, but where does the channel begin? This can be a matter of interpretation depending on the scale of the map or aerial photograph being used. Such information might easily differ with field observations of the same channel. The "blue lines", such as those defined by the USGS in the National Hydrography Dataset (USGS, 2003d), are often taken to indicate the upstream extent of a stream, however, these lines vary with map scale and are subject to the personal aesthetics of those digitizing the streams (Leopold, 1994). In the end, some objective, reproducible approach must be employed by the GIS user to define the upstream extent of a stream. For instance, the minimum source area described earlier and illustrated in Figure 13 for a value of 5 cells might be one way to define the upstream extent of a stream. (Please note that the 5 cell threshold used in the 5×5 DEM example is purely for illustrative purposes and is not intended to imply an appropriate value for 30 m DEMs or any other resolution.) Coupling whatever rule for defining the upstream end of the channel with the watershed's outlet location, the GIS can then readily determine the channel slope as given in equation (12).

The watershed slope is calculated in much the same way as the channel slope except that the upstream end of the "channel" is chosen to begin at the watershed divide. The flow path (and thus the flow length) is determined exactly as shown in Figure 17 describing the calculation of the "longest flow path". Such a definition removes the ambiguity of defining an upstream end of the "channel" and makes the determination of this slope more straightforward than the calculation of the channel slope.

The land slope is defined by the NRCS (SCS, 1985) as the average of all the individual slopes determined over the grid system placed over the watershed. This definition precedes the advent of GIS technology, but the grid is clearly analogous to a raster characterization of the watershed. The average slope is thus determined by using the GIS to calculate the local cell-to-cell slopes, as was illustrated earlier in the flow direction section. These local slopes are determined throughout the watershed and then averaged to derive the land slope.

Watershed relief is another measure of the vertical character of the watershed. It is defined as the average of all elevations within the watershed minus the elevation at the watershed outlet. Because this quantity is averaged over the watershed, it is a better indicator of the potential energy available to runoff than is the maximum basin relief that only considers the most extreme elevation within a watershed. The basin relief is readily calculated within the GIS using only the DEM and knowledge of the watershed boundaries.

Clearly DEM and GIS are extremely helpful tools to provide an indication, or a first approximation, of the real hydrological and hydraulic behavior of a watershed or river basin. Such approximation can prove quite adequate (and perhaps even suffice) if the interest is in the overall basin characteristics. However, if detailed simulations are required of dynamic processes (like in case of flood routing or flood early warning) then often a more physically based description is to be used, deriving flows from physical conservation principles of mass, momentum, and energy, as described in the articles by Verwey and Stelling (2005), Lin and Falconer (2005), Werner *et al.* (2005) and elaborated in more detail in **Chapter 13, Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, Volume 1** of this Encyclopedia.

CASE STUDY – MARYLAND'S GISHYDRO2000

The users of GIS technology and DEM data for hydrologic analysis and modeling has steadily grown outward from academia into a community of practitioners that use these tools and data on a daily basis. Along these lines, specialized GIS-based environments have evolved that help streamline the types of hydrologic analyses that practitioners are performing. Such tools include HEC-GeoRAS (USACE, 2003b) and the Watershed Modeling System (WMS) (USACE, 2003c). These are GIS-based tools that serve as interactive "front-ends" to hydrologic or hydraulic models that depend heavily on topographic and other geographic data. These tools take advantage of GIS technology to aid the user in developing input data for computer models that might otherwise require time-consuming or repetitive manipulations before an analysis can be performed.

We focus here on a specialized environment, GISHydro2000, developed at the University of Maryland (Ragan, 1991; Moglen and Casey, 1998; Moglen and Kosicki, 2000; Moglen, 2003) for use in conducting hydrologic analyses anywhere in the State of Maryland. This tool runs within the ArcView GIS environment and depends critically on DEM data in order to function. GISHydro2000 is unique in that it contains both the data and software tools necessary to perform a hydrologic analysis all within a single interface. A brief sample analysis will be demonstrated here to illustrate the merger of a GIS-based tool with DEM data.

The first step of the analysis process is the selection of data to sufficiently cover the watershed being studied. In GISHydro2000, this step is performed either graphically or by textually selecting the areal extent of the analysis. This is shown in Figure 21 where the user specifies the analysis extent by selecting one or more USGS 7.5 minute

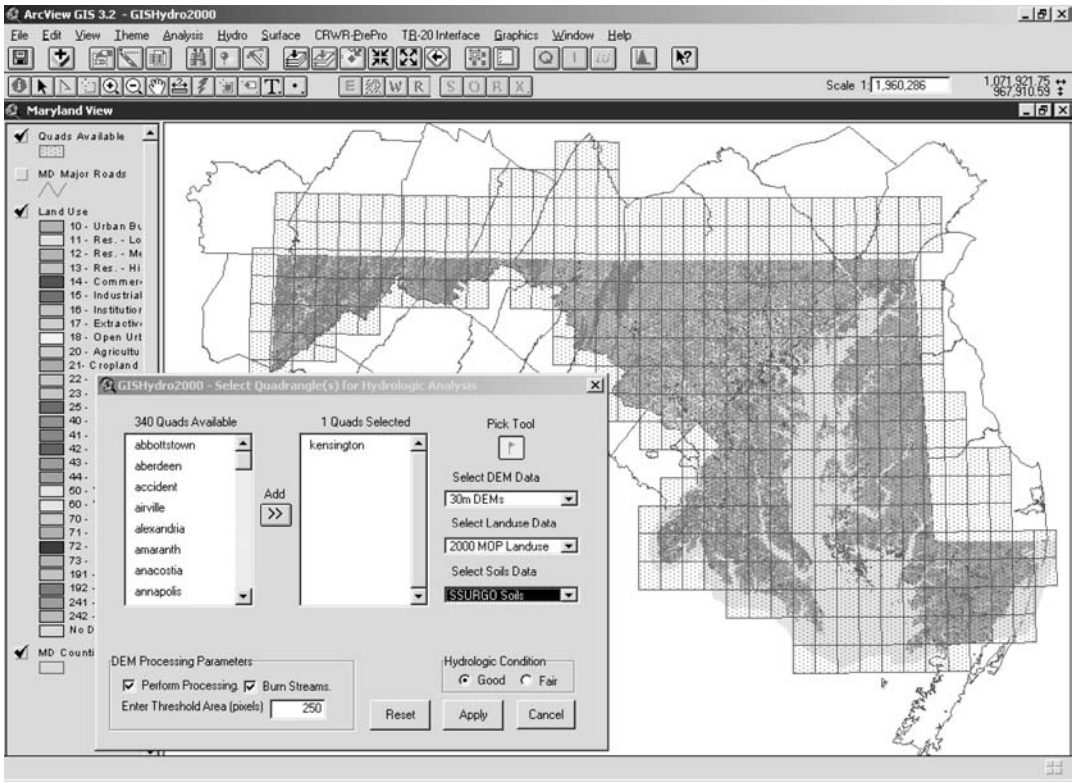


Figure 21 The data selection step in GISHydro2000. The dialogue box at lower right allows for graphical or textual selection of quadrangles. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

quadrangles (the Kensington quadrangle is shown selected in the dialogue box).

In addition to specifying the analysis extent, the user is able to specify different resolutions of DEMs (30 m, 90 m, or a mixture), different dates or sources of land use information, and different sources of soils information. Finally, the user indicates whether only these raw data are desired or if the DEM is to be interpreted hydrologically for flow directions and flow accumulation as described earlier in this article. When all the above information is specified, the user presses the "Apply" button which then extracts the requested data and performs the hydrologic interpretation of the DEM if indicated. For a single 7.5-min quadrangle, the entire process of determining flow directions and flow accumulation may take on the order of 30 – 60 s depending on the hardware that is being used.

The next step, delineation of the watershed, requires the user to indicate the watershed outlet. The GIS then delineates the watershed draining to that outlet with a typical result shown in Figure 22.

Notice in this figure the concept of overlapping information described earlier. The shaded background surrounding the watershed is the DEM with darker shades corresponding to higher elevations. The solid region superimposed on this layer is the watershed itself. Within the watershed, the inferred stream network is indicated. This network was inferred by having the GIS indicate all cells that have a flow accumulation in excess of a fixed amount that is typical of the source area needed to form a stream in this region. (Please note that these cells are not the same information as the "blue lines" that are part of the National Hydrography Dataset (NHD) USGS, 2003d). Blue lines are a vector quantity that are manually digitized by cartographers. These inferred channel cells are a raster quantity that is automatically derived from the DEM as described earlier in this article.) Finally, the major road network is shown as a set of black vector lines spanning the entire area of interest.

The outcome of the selection of a menu choice is the next illustrated step that produces the dialogue box shown in Figure 23.

The upper portion of this dialogue reflects back to the user the data and parameters that were used for the analysis. (For instance, the reader here can confirm that the data are for a watershed in the Kensington, MD quadrangle with an outlet at the indicated location in the Maryland state plane coordinate system.) Under "Findings", a suite of GIS-based calculations have been performed, many of which have been described earlier in this article. The drainage area has been determined (2.8 $mile^2$, 7.3 km^2), the channel slope (62.8 ft $mile^{-1}$, 11.9 m km^{-1}) and land

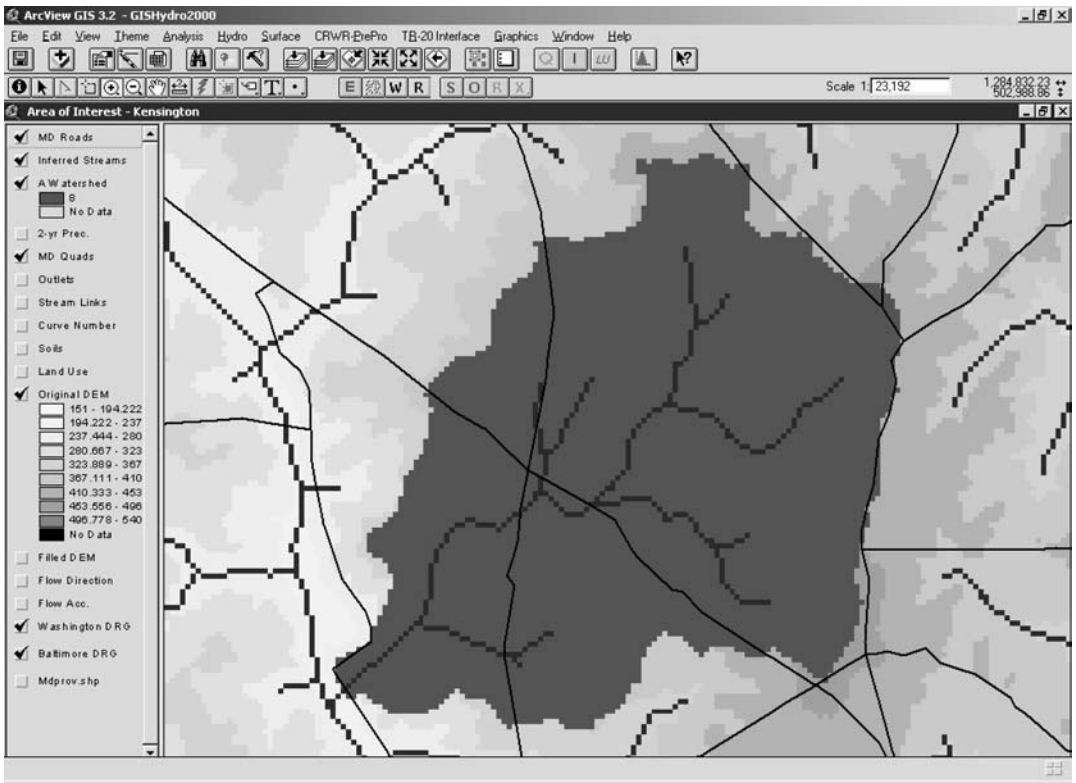


Figure 22 The delineation of a watershed within GISHydro2000. Visible are the surrounding shaded DEM data, the watershed boundary, the inferred drainage network, and the major road network. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

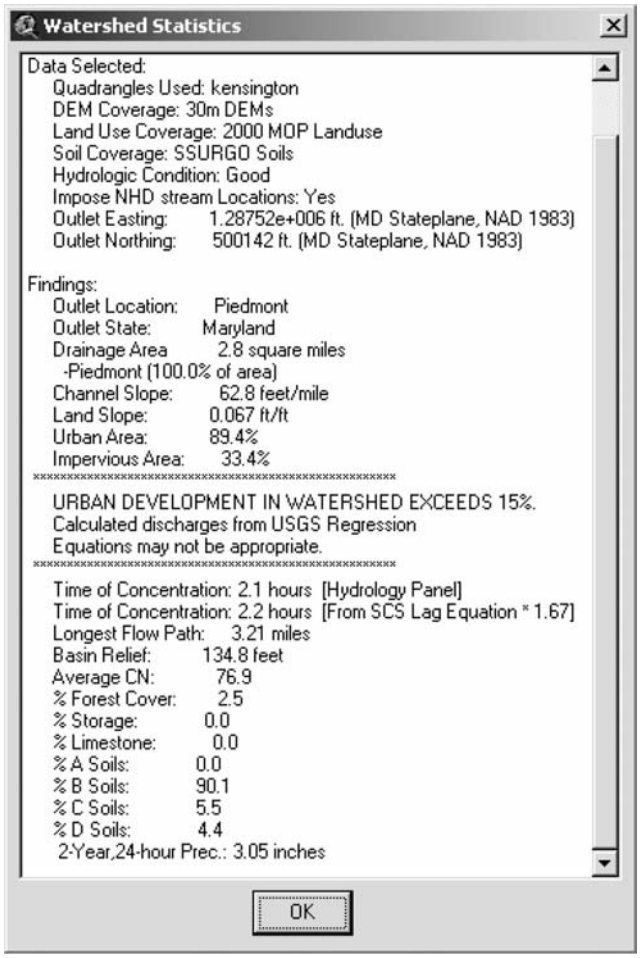


Figure 23 The watershed statistics dialogue in GISHydro2000 corresponding to the watershed delineated in Figure 22. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

slopes have been calculated (0.067 ft ft^{-1}, 0.067 m m^{-1}), the longest flow path is 3.21 miles (5.77 km), and the basin relief is 134.8 ft (41.1 m). By virtue of the land use and soils data also used in the analysis, the reader should note that the statistics include information about the land use and soils distributions in the watershed. The NRCS (SCS, 1985) curve number can also be determined readily from land use and soils information with the average value for the watershed shown here (76.9). This information is all quickly determined within the GIS by using the watershed boundary as a "cookie-cutter" or mask to include only those data that are located within the watershed in any reported calculation. Note that the NRCS lag-based time of concentration is easily calculated. This value is calculated defined as (SCS, 1973)

$$t_c = \frac{L^{0.8}[(1000/CN) - 9]^{0.7}}{441 \cdot S^{0.5}} \tag{13}$$

where t_c is the time of concentration in hours, L is the length of the longest flow path (in meters), CN is the basin-averaged curve number, and S is the land slope in percent. The calculation of several of these quantities was described earlier in this article. The reader should verify that the resulting time of concentration, t_c, is approximately 2.2 h for the data reported in the dialogue box shown in Figure 23. The automation of exactly these types of calculations is easily performed within the GIS.

Although GISHydro2000 also serves as a front-end to the NRCS TR-20 (SCS, 1984) rainfall-runoff model, we will end our case study illustration here with a second dialogue box, shown in Figure 24, that shows the peak flow estimates based on the USGS regression equations applied to this case study watershed.

The current USGS peak flow regression equations (Dillow, 1996) estimate the 2- through 500-year peak flows at ungauged rural sites on the basis of the watershed characteristics shown in Figure 23. For instance, the 2-year peak discharge in the Maryland Piedmont province is calculated as:

$$Q_2 = 6.979A^{0.635}(F + 10)^{-0.266} \tag{14}$$

where Q_2 is the 2-year peak discharge in m^3 s^{-1}, A is the drainage area in km^2, and F is the percent forest cover. In this case, the drainage area is 7.25 km^2 (2.8 mile^2) and the forest cover is 2.5%. The reader can confirm that the discharge is essentially 12.5 m^3 s^{-1} (442 ft^3 s^{-1}) (to within truncation error in the reporting of both inputs to equation (14) and the resulting discharge estimate output). In Figure 24, the GIS has simply been used as a tool to quickly apply these regression equations to the case study watershed. To learn more about this particular analysis and

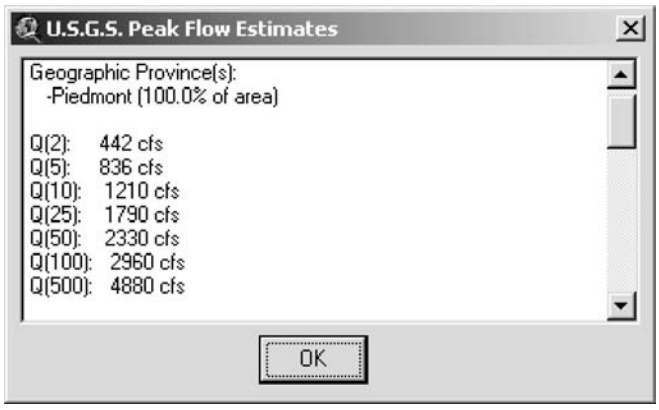


Figure 24 The USGS rural regression equations peak discharge dialogue in GISHydro2000 corresponding to the watershed delineated in Figure 22. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

modeling tool and to download a copy of this program, please visit **http://www.gishydro.umd.edu**.

We should emphasize that none of the results presented here are profound or novel. Rather, the novelty lies in the means used to arrive at these results. By virtue of GIS technology and being driven heavily by DEM data, the illustrative case study could be performed very rapidly (less than 2 min) with the user only needing to specify an analysis extent, the location of the watershed outlet, and then selecting two menu choices to request the watershed characteristics and peak flow estimates. A similar analysis done entirely by hand using the older technologies of a planimeter or a light table and graph paper could easily take an hour or longer just for this small watershed. If the watershed were more substantial in area, the time necessary to do a similar analysis within the GIS would still remain negligible while the old technology analysis time might stretch to weeks or even months. Furthermore, the results from the GIS are objective and reproducible where hand analyses are probably not.

Hence, DEM and GIS prove to be powerful tools when it comes to a rapid assessment of overall watershed/river basin characteristics. This may often already be quite adequate for hydrological analysis purposes. However, if more detailed processes are to be derived from sufficiently available hydrological data (e.g. rainfall-runoff relations), then other technologies are available based on, for example, data-driven modeling, as described by Solomatine (2005) and Artificial Neural Network concepts in hydrology, as elaborated by Minns and Hall (2005).

SUMMARY AND CONCLUSIONS

In this article, the DEM data and GIS technologies have been presented with a specific emphasis on their relevance to hydrologic analysis. DEMs were defined and several sources for these data were provided. GIS technology was introduced with a basic description of both raster and vector data models. The database-map presentation linkage of GIS was illustrated and a brief discussion of map projections was provided. The majority of this article focused on the hydrologic interpretation of DEMs. Flow directions, flow accumulation, watershed delineation, flow lengths, travel times, and measures of slopes were discussed and illustrated through several GIS-based examples. This article concluded with a brief illustration of a GIS-based computer environment developed for the State of Maryland that uses DEM and other spatial data to automate hydrologic analysis. A sample watershed was delineated, its main physical characteristics were quantified, and its peak flow behavior was estimated using rural peak discharges provided by the USGS. The purpose of this illustration was to show how DEM data and the GIS environment can be integrated to streamline hydrologic analyses and rapidly provide both visual and textual information about any watershed the hydrologist wishes to study. Moreover, these computer-based methodologies allow analyses to be performed in a fraction of the time spent by traditional hand methods, with greater consistency and reproducibility. DEM and GIS are typical examples of the importance and capability of hydroinformatics tools and technologies in the field of hydrology.

REFERENCES

Clark C.O. (1943) Storage and the unit hydrograph. *Proceedings of the American Society of Civil Engineers*, **9**, 1333–1360.

Costa-Cabral M.C. and Burges S.J. (1994) Digital Elevation Model Networks (DEMON) – A model of flow over hillslopes for computation of contributing and dispersal areas. *Water Resources Research*, **30**(6), 1681–1692.

Dillow J.J.A. (1996) *Technique for Estimating Magnitude and Frequency of Peak Flows in Maryland*, Water Resources Investigations Report No. 95-4154,.U.S. Geological Survey.

Environmental Systems Research Institute (2003) *The GIS Software Leader* Redlands, **http://www.esri.com**

Leopold L.B. (1994) *A View of the River*, Harvard University Press: Cambridge.

Lin B. and Falconer R.A. (2005) Hydrological and environmental modelling in rivers and estuaries. *Encyclopedia of Hydrological Sciences*, Wiley Publishing.

Maidment D.R. (2002) *Arc Hydro: GIS for Water Resources*, ESRI Press: Redlands.

Minns A.W. and Hall M.J. (2005) Artificial neural networks concepts in hydrology. *Encyclopedia of Hydrological Sciences*, Wiley Publishing.

Moglen G.E. (2003) *GISHydro: A GIS-Based Hydrologic Modeling Tool*, College Park, **http://www.gishydro.umd.edu**

Moglen G.E. and Casey M.J. (1998) A perspective on the use of GIS in hydrologic and environmental analysis in Maryland. *Infrastructure*, **3**(4), 15–25

Moglen G.E. and Kosicki A. (2000) GISHydro2000: performing automated hydrologic analyses in Maryland. *TR News*, **210**, 18–19. Transportation Research Board, National Academy of Sciences. Washington. Also available on the World Wide Web: **http://gulliver.trb.org/publications/trnews/rpo/rpo.trn210.pdf**

National Aeronautics and Space Administration (2003) *Shuttle Radar Topography Mission*, Jet Propulsion Laboratory, Pasadena, **http://www.jpl.nasa.gov/srtm/**

Ragan R.M. (1991) *A Geographic Information System to Support State-wide Hydrologic and Nonpoint Pollution Modeling*, Technical Report No. FHWA/MD-91/02, Department of Civil Engineering, University of Maryland, College Park, MD.

Snyder F.F. (1938) Synthetic unit-graphs. *Transactions-American Geophysical Union*, **19**, 447–454.

Soil Conservation Service (1973) *A Method for Estimating Volume and Rate of Runoff in Small Watersheds*, TP-149. U.S. Department of Agriculture: Washington.

Soil Conservation Service (1984) *Computer Program for Project Formulation*, Technical Release 20, Washington.

Soil Conservation Service (1985) *National Engineering Handbook*, Supplement A, Section 4, Hydrology. U.S. Department of Agriculture: Washington.

Solomatine D.P. (2005) Data-driven modelling and computational intelligence methods in hydrology. *Encyclopedia of Hydrological Sciences*, Wiley Publishing.

Surkan A.J. (1968) Synthetic hydrographs: effects on network geometry. *Water Resources Research*, **5**(1), 112–128.

Tarboton D.G. (1997) A new method for the determination of flow directions and upslope areas in gird digital elevation models. *Water Resources Research*, **33**(2), 309–319.

United States Army Corps of Engineers (2003a) *Corpscon 5.x for Windows*, U.S. Army Topographic Engineering Center: Alexandria, **http://crunch.tec.army.mil/software/corpscon/corpscon.html**.

United States Army Corps of Engineers (2003b) *HEC-GeoRAS*, Hydrologic Engineering Center: Davis, **http://www.hec.usace.army.mil/software/hec-ras/hecras-hec_georas.html**

United States Army Corps of Engineers (2003c) *Watershed Modeling System*, Coastal and Hydraulics Laboratory: Vicksburg, **http://chl.wes.army.mil/software/wms/**

United States Geological Survey (2003a) *USGS EDC: National Elevation Dataset Home Page*, EROS Data Center: Sioux Falls, **http://gisdata.usgs.net/ned/default.asp**

United States Geological Survey (2003b) *Elevation Derivatives for National Applications (EDNA)*, EROS Data Center: Sioux Falls, **http://edna.usgs.gov**

United States Geological Survey (2003c) *GTOPO30 – Global Topographic Data*, EROS Data Center: Sioux Falls, **http://edcdaac.usgs.gov/gtopo30/gtopo30.html**

United States Geological Survey (2003d) *National Hydrography Dataset Home Page*, Reston, **http://nhd.usgs.gov/**

Verwey A. and Stelling G.S. (2005) Title of forthcoming contribution on numerical flood simulation. *Encyclopedia of Hydrological Sciences*, Wiley Publishing.

Werner M., Kwadijk J. and Schellekens J. (2005) Title of forthcoming contribution on flood early warning systems. *Encyclopedia of Hydrological Sciences*, Wiley Publishing.

16: Numerical Flood Simulation

GUUS S STELLING[1] AND ADRI VERWEY[2]

[1]Department of Civil Engineering, Technical University of Delft, Delft, The Netherlands
[2]WL Delft Hydraulics, Delft, The Netherlands

This contribution describes the state of the art of numerical flood simulation, in particular, the models based upon hydraulic laws. Both theory and practical aspects are discussed. In the introduction, various types of applications of these models are listed with their requirements in terms of robustness, accuracy, speed of operation, and engineering staff time. After introducing the 1D Saint Venant unsteady flow equations and describing the meaning of their terms, the concepts of kinematic and diffusive wave approximations are discussed. These provide better ways of understanding the physical behavior of flood waves and demonstrate the link between hydraulic and hydrologic flood propagation models. Numerical discretization of the unsteady flow equations is introduced with special attention given to the correct and robust modeling of rapidly varying flow, as occurring around hydraulic structures, discontinuities in channel beds, and flood-generating breaches of river embankments and reservoir dams. Proceeding with the 2D flow description and its numerical discretization principles, the emergence of hybrid 1D/2D models is discussed with their different options of linking model components of flood-prone areas. The quality of models also relies very much on the availability of detailed and correct data. This contribution briefly introduces the impact of new data collection techniques on numerical flood model schematization.

INTRODUCTION

The risk and impact of floods have always challenged scientists and engineers to master the knowledge on this phenomenon. Over the past decades, for a number of reasons, considerable energy has been spent towards achieving this end. In the first place, there was the challenge of protecting a rapidly increasing world population with a strong drive to settle in flood-prone coastal zones and river flood plains. In addition, there was the increasing awareness about climate changes, which appear to bring more and more precipitation to river catchments. There also are numerous intellectual challenges, as scientific and technological developments in a wide range of fields open up many new ways of supporting flood related studies.

In this contribution, we are limiting ourselves to physically based hydrodynamic or hydraulic models. The underlying principles are conservation of volume, momentum, and energy. The equations were already developed some centuries ago by Newton (1687), who introduced clear mathematical formulations based on physical conservation principles, and De Saint Venant (1871), who formulated the mathematical equations for modern day river flow simulations. However, it required the development of powerful computer equipment to obtain suitable numerical techniques to solve these equations for practical applications. This has led to a large number of modeling systems for the simulation of unsteady channel flow. In the area of numerical flood simulation, the frequently used tools currently available in the market are: the integrated 1D/2D modeling system SOBEK of Delft Hydraulics (**www.wldelft.nl** and **www.sobek.nl**), the Mike11 (1D) and Mike21 (2D) modeling systems of the Danish Hydraulics Institute (**www.dhi.dk**), ISIS of Wallingford Software (**www.wallingfordsoftware.com**), and HEC-RAS of the US Army Corps of Engineers (**www.hec.usace.army.mil/software/hec-ras**).

Although numerical methods for solving unsteady flow equations were developed already half a century ago, for example, by Arakawa (1966), it was only recently that numerical techniques behind flood simulation models reached an acceptable level of perfection. Hydrologists are now able to model flow in channels and over flood plains, irrespective of their bathymetric or topographic

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

complexity, the number and location of embankments for flood protection, roads and railways, and the number and complexity of hydraulic structures and the way we control these.

Depending on their objective, flood simulation models may differ in their requirements. Criteria for the selection of the appropriate tool are often based on the required engineering staff time for model development, overall consultancy time for product delivery, speed of computation, completion time for a simulation, accuracy level of results, data requirements, numerical robustness, user-friendliness of the software, and possibly others, depending on the objective of the model. These objectives may be related to flood risk analysis, flood forecasting, and flood control and may be based upon a variety of causes, such as storms, dam or dike breaks, hurricanes, typhoons, or similar low atmospheric pressure phenomena. Recently (2004) attention has also been drawn once again to the devastating effects of tsunamis. All these application areas of numerical models have their own requirements, as will be discussed briefly in the sequel.

In many countries, insurance companies are using flood risk maps, sometimes based upon relatively simple and quick estimates obtained via simple rules in GIS. Generally, this does not justify the more complex laws defining the flow of water. For this reason, a first improvement is found by applying 1D (one-dimensional) steady flow models with GIS postprocessing to develop topography based flood frequency contour lines (e.g. FEMA procedures: `www.floodmaps.fema.gov/fhm/`). However, there is a tendency now to base flood risk analysis on more detailed 1D and 2D unsteady flow models for flood-prone areas with valuable assets and a complex infrastructure. Federal and local governments have also become more aware of the potential of using such models for evacuation planning. In The Netherlands, for example, more than 60% of the country is subject to flood risk and for most of these areas, combined 1D and 2D hydrodynamic models have been developed to study the effects of potential dike breaks to provide guidelines to authorities in setting up evacuation plans. Besides generating flood depths, these models have to be capable of providing accurate estimates of flood wave propagation celerities over dry beds.

Flood forecasting sets quite different requirements. The speed of producing a forecast is one of the most important criteria, especially in areas where flash floods occur. For this reason, numerical models behind a river–catchment flood forecasting system are usually 1D hydrodynamic models, gradually replacing the simpler hydrological routing techniques. There is a tendency to include partly 2D hydrodynamic models, which is already common practice in flood forecasting systems for coastal areas and seas. Numerical models for flood forecasting are usually embedded in a flood forecasting platform, such as the Delft FEWS system (Werner *et al.*, 2004), which has recently been installed in the UK to provide flood forecasts for nearly all river basins in the country.

Important criteria for numerical models supporting flood control are accuracy, flexible schematization options, numerical robustness, and consultancy time for model development and use. Currently, state of the art for flood control is the use of combined 1D and 2D models (e.g. Hesselink *et al.*, 2003). The former use of flood cells has been replaced by complete 2D flow descriptions, whereas subgrid channel flow is still better described in 1D. Flood control models should be based upon reliable physical descriptions and schematizations, as part of their use is in extrapolation of calibrated models to extreme situations that have never occurred. One of the reasons to build models for flood control is the study of downstream impacts, especially cross-border effects. Downstream impacts of flood control are changed flood wave celerity and changed flood peak attenuation. Higher flood wave celerities result from deepening of the river and the construction of embankments. This, in turn, leads to increased peak floods downstream. The construction of flood retention areas has opposite impacts and may be used to compensate the negative impacts. Model selection criteria then follow from the detail in which potential economic, environmental, and social impacts have to be studied.

The analysis of floods caused by dam- and dike breaks requires extremely robust numerical methods, especially for the description of flooding of dry areas and the correct propagation of the wave front. Moreover, model accuracy, partly based upon the ability to describe the full hydrodynamic equations, is important, as will be discussed in the section on software and model validation. As dam- and dike break simulations are nearly always made for the prediction of their potential effects, data for model calibration is rarely available. The quality of the model fully depends on its descriptive capabilities of the physical system in terms of topographic and roughness data, the representativeness of the equations, and the numerical methods applied. However, it has to be kept in mind that the overall model accuracy also follows from the quality of the description of the dam failure mechanism and the assumptions made here.

Floods generated by the passage of low atmospheric pressure zones such as hurricanes, typhoons, and the geologically induced tsunamis require the modeling of 2D flow in coastal zones, seas, and oceans and may set requirements such as the description of Coriolis forces, the use of spherical coordinates and curvilinear grids, the specification of moving atmospheric pressure fields, special ways of handling initial data, and so on. A possible integrated use of 1D, 2D and 3D models may provide advantages here.

In addition to numerical method developments, new technologies for the collection of data have recently led to complete changes in the selection of the type of numerical models. In particular, the development of global positioning system (GPS) and differential GPS (DGPS) technology has led to far cheaper methods of collecting bathymetric and topographic data (Moglen and Maidment, **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**). In turn, this has led to the gradual replacement of the hydrograph-based hydrologic models by the bathymetric- and topographic-based hydrodynamic or hydraulic models. Similarly, the collection of detailed digital terrain data in river and coastal flood plains has led to the replacement of 1D models by 2D models.

Let us first consider the impact of LIDAR (Light Detection And Ranging). The use of this laser technology, scanning the earth's surface with laser beams from airplanes or helicopters, has provided the means to generate highly accurate digital elevation models at a relatively low cost. As an example, the whole area of The Netherlands has been remapped over the past years, with an accuracy of approximately 10 cm in the vertical at a density of 1 point per 16 m^2. Total cost of this project was approximately 10 million euros, or approximately 250 euros per km^2 (Verwey, 2001). This new topographic information has been essential for the flood risk analyses and the evacuation planning studies mentioned earlier.

Also river bathymetries are obtained at a relatively low cost now by boats equipped with multibeam echo sounders. Here also, the position of the boat is recorded via DGPS, while the spatial sound signals record bottom depths relative to the boat at an accuracy of approximately 10 cm in the vertical. At a typical boat speed of 4 m s^{-1} and a swat width of the order of magnitude of the river depth, the bathymetry of quite large riverbeds can be obtained in a few days. Experiments are also being made with laser beams in the green range, as these are able to pass through clean water and enable the application of LIDAR technology for river bathymetries also.

In similar ways, other technologies are advancing rapidly, enabling large amounts of data to be collected at relatively low cost. Worth mentioning are the Acoustic Doppler Current Profiler (ADCP) technology for tidal discharge measurements and the increased precision of spatially distributed precipitation measurements with radar.

ONE-DIMENSIONAL FLOW

By using one-dimensional models (1D) in flood simulation, it is assumed that the phenomenon can be described satisfactorily as unsteady flow in one spatial dimension. In this approach, the flood is assumed to be correctly defined by the state variables or dependent variables, discharge Q and water level h, as a function of the independent variables t for time and x for space.

Basic assumptions and/or limiting conditions are:

- the discharge is sufficiently well defined as the integral of the velocities through a cross section, perpendicular to the x-axis and perpendicular to the flow velocity vectors in the flood plain;
- the water level is constant along the cross section. This implies that at any time the water level at all points along a given cross section should rise or fall at the same rate. This assumption is generally justified when the widths of river and flood plain are of the same scale and free of obstacles such as natural levees or embankments;
- the water level slope or gradient in the x-direction is constant along the cross section. However, it should be noted that in case this condition is not satisfied, correction factors may be applied to reduce significantly errors in the model parameters, as discussed in this contribution.

Quantitative analysis of the two state variables requires two independent equations. Usually, the following equations are used:

- the continuity equation, based upon volume conservation in a control volume defined between two successive points along the channel axis;
- the momentum equation, based upon the conservation of momentum, including the effect of impulses generated by forces acting upon the water contained in the control volume.

Making, furthermore, the following assumptions that

- the pressure distribution in the vertical is hydrostatic;
- the resistance relationship for steady flow is also applicable for unsteady flow; and
- the bed slope is moderately steep so that the cosine of the slope can be replaced by unity,

De Saint Venant (1871) derived the following equations (presented in a slightly adapted form here):

$$\frac{\partial A_t}{\partial t} + \frac{\partial Q}{\partial x} = q_{\text{lat}} \tag{1}$$

$$\frac{1}{gA}\left\{\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{Q^2}{A}\right)\right\} + \frac{\partial \zeta}{\partial x} + \frac{Q|Q|}{K^2} = 0 \tag{2}$$

where A_t is the cross-sectional area representative of storage over a control volume (m^2), t is the time (s), Q is the discharge (m^3 s^{-1}), x is the position along the channel axis (m), q_{lat} is the lateral discharge per unit length of channel (m^2 s^{-1}), A the flow-conveying cross-sectional area

(m^2), ζ the water level above a selected horizontal reference plane (m) and K the channel conveyance ($m^3\ s^{-1}$).

The first term of equation (1) accounts for the rate of change in volume stored over a unit length of channel. The second term represents the rate at which the discharge changes along the channel per unit time. The term at the right-hand side represents the effect of lateral flow per unit time and length into the channel.

Equation 2 is the momentum equation which has been simplified by dividing all terms by a density, assumed to be constant. In its original form, the first term represents the change of momentum in a control volume of unit length of channel and reflects the inertia of the water mass present in that control volume. The second term is called the *convective momentum term* and reflects the balance of momentum flowing through the control volumes' upstream and downstream cross section. The third term combines the effect of the impulses generated by differences in upstream and downstream hydrostatic forces (*hydrostatic pressure term*) and the gravity acting on the mass in the control volume (*gravity term*). Finally, the last so-called *friction term* represents the effect of channel wall friction.

It is readily observed that all terms in the momentum equation are written in dimensionless form and represent a gradient of a certain quantity along the channel axis. In its simplest form, equation (2) may be reduced to the familiar steady flow conveyance relationship

$$Q = K\sqrt{I} \tag{3}$$

with the conveyance K expressed, as

$$K = CA\sqrt{R} \quad \text{or} \quad K = \frac{1}{n}AR^{2/3} \tag{4}$$

by equivalence with the Chezy equation and with the Manning equation, respectively, where I is the channel bed slope (−), C the Chezy resistance coefficient ($m^{1/2}\ s^{-1}$), R the hydraulic radius (m); and n the Manning friction coefficient ($s\,m^{-1/3}$).

The set of equations (1) and (3) forms the so-called *kinematic wave approximation* for flood propagation, which, after substitution of (3) into (1) and neglecting the lateral flow term, can be further simplified to the form

$$\frac{\partial Q}{\partial t} + c\frac{\partial Q}{\partial x} = 0 \tag{5}$$

with a flood wave celerity c ($m\,s^{-1}$) expressed as

$$c = \frac{1}{b_s}\frac{dQ}{dh} \tag{6}$$

It is important to realize that, as an essential assumption in the derivation of this kinematic waveform, the *channel bed slope* has been used in the conveyance relationship. Equation (5) has the form of an advection equation, which expresses that flood waves propagate with a celerity c which is inversely proportional to the available channel storage width b_s (m) and a linear function of the derivative of the local flow *rating curve*. The characteristic celerity of this kinematic wave is lower than that of the dynamic wave characteristic in the same direction. As discussed by Abbott (1979) it is this mechanism that leads to roll waves at flood wave fronts, limiting their propagation speed (*see also* Stoker, 1957).

The first-order partial differential equation (5) also expresses that along its characteristic celerity c the discharge remains constant, and so does the peak of the flood wave. In other words, there is no dampening effect of the flood peak. Though this is approximately true for rivers with steep slopes, the stretches with milder slopes require a lesser simplification of the Saint Venant equations. Following, for example, Chaudhry (1993) and defining I of equation (3) as the *water level slope* $\partial\zeta/\partial z$, substitution of the full last two terms of equation (2) into equation (1) gives the so-called *diffusive wave approximation*

$$\frac{\partial Q}{\partial t} + c\frac{\partial Q}{\partial x} = D\frac{\partial^2 Q}{\partial x^2} \tag{7}$$

with a flood wave diffusion coefficient D ($m^2\ s^{-1}$) derived as

$$D = \frac{K}{2b_s\sqrt{I}} \tag{8}$$

Including the lateral flow term, the diffusive wave approximation reads

$$\frac{\partial Q}{\partial t} + c\frac{\partial Q}{\partial x} = D\frac{\partial^2 Q}{\partial x^2} + cq \tag{9}$$

Returning to equation (3) again, the variable I represents the bed slope in the case of the kinematic wave approximation and the water level slope in the case of the diffusive wave approximation. The description based upon the full set of equations (1) and (2) is defined as the *full dynamic wave description*.

Equation (9) shows that along the characteristic line $c = dx/dt$ the flood peaks are dampened by the integral of the diffusion term, whereas they are increased by the integral of the lateral flow term. Although equations (5) and (9) are rarely used in discretized form anymore, they are useful for the flood modeler to provide insight into the physical nature of flood wave propagation. Moreover, it links hydrologic and hydraulic flood routing techniques. By applying a Taylor's series expansion to the Muskingum equation, Cunge (1969) has demonstrated that the well-known hydrological Muskingum flood routing technique emulates the solution

of the advection–diffusion equation (7). This insight provides guidelines for a suitable choice of the Muskingum parameters on the basis of the expressions for c and D (equations (6) and (8) respectively).

The numerical Muskingum method belongs to the class of hydrological routing techniques. These methods are based upon the notion that flood wave propagation characteristics can be derived from measured flood hydrographs along the river, rather than from detailed topographic information, as discussed in the World Meteorological Organization (WMO) report on flood forecasting models by Serban *et al.* (2005). The basis of these methods is that flood wave propagation and diffusion behavior can be represented by a limited number of parameters that can be calibrated from observed hydrographs. Usually there are only two parameters, as in the case of the Muskingum routing method. The limitation of this assumption is clearly shown by analyzing the stage-dependent expressions for celerity and diffusion given by equations (6) and (8). Although the current methods for flood routing are far more precise, there are still many situations where the use of hydrological flood routing techniques is still justified. In many river catchments, and especially in tributaries, the collection of detailed information on river bathymetry and floodplain topography is economically not always justified, despite the emergence of relatively cheap, new technologies.

An alternative to hydrological forecasting methods is provided by newly developed artificial neural network (ANN) concepts (Minns & Hall, **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**). This technology also relies on measured hydrographs to derive relationships between inflowing and out flowing hydrographs, though in practice ANNs are mostly applied to the development of relationships between rainfall and river catchment runoff.

Both hydrological routing and ANN techniques provide limited reliability for the range of events outside those used for calibration. However, this is exactly the range one is interested in when dealing with extreme flood events, which rarely occur and for which measurements are even more rarely available. So, even though ANNs and other data-driven modeling techniques may prove quite valuable in, for example, determining rainfall-runoff relations (Solomatine, **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**), physically based descriptions, such as those provided by hydraulic routing techniques and based upon the full use of equations (1) and (2), offer better extrapolation possibilities than hydrological routing methods or ANN-based models. For this reason, we will focus on hydraulic modeling techniques in the sequel. However, it should be realized that simplified equations, such as equation (7), remain useful, as these offer us a good insight into the physical nature of flood propagation, in particular, the concepts of flood peak arrival time and the attenuation of peak discharges.

NUMERICAL SOLUTIONS

Hydraulic modeling techniques for 1D are currently based upon the numerical solution of the Saint Venant equations, including the full convective momentum term. These numerical solutions are nearly exclusively based upon implicit finite difference methods. These offer the advantage of unconditional numerical stability, while the various robustness problems of the past, relating to nonlinear effects and flooding and drying of channels and flood plains, have been solved satisfactorily. The application of finite elements and finite volume techniques does not provide specific advantages as, in 1D modeling, most of these techniques lead to equivalent forms derived through finite difference formulations.

Numerical methods for the Saint Venant equations may be based upon so-called *staggered* and *nonstaggered schemes*. The first category represents formulations where the dependent variables Q and ζ are defined alternately at successive grid points along the x-axis. For nonstaggered schemes, however, the variables Q and ζ are defined at the same grid points. At first sight, this last definition offers advantages through the availability of the state variables discharge and water level at the same points along the channel axis. It has been shown, however, that the staggered grid approach offers distinct advantages over nonstaggered grids by guaranteeing the convergence of numerical solutions and the better ability to handle flooding and drying of grid sections, as shown by Stelling *et al.* (1998).

For the numerical solution of the Saint Venant equations (1) and (2), we will consider their Eulerian form per unit width of channel by first neglecting the lateral flow and further simplifying the equations to

$$\frac{\partial \zeta}{\partial t} + \frac{\partial (uh)}{\partial x} = 0 \tag{10}$$

and

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + g\frac{\partial \zeta}{\partial x} + c_f \frac{u|u|}{h} = 0 \tag{11}$$

where ζ is the water level defined as $\zeta = h + z_b$ with h defined as the local water depth (m) and z_b as the local bottom level (m), u the flow velocity ($\mathrm{m\,s^{-1}}$), and c_f the dimensionless bottom friction coefficient.

Referring to Figure 1 and to Stelling and Duinmeijer (2003) for further details, the staggered grid approach requires that alternately at ζ- and u-points, equations (10) and (11) are transformed into finite difference form (*see also* Abbott, 1979; Cunge *et al.*, 1986; Hirsch, 1990; Toro, 1999). Taking a ζ-point as the only feasible choice for

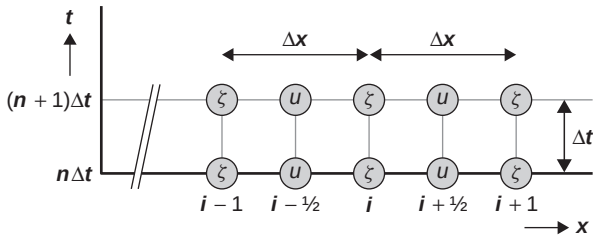


Figure 1 Staggered grid for unsteady channel flow

the transformation of the continuity equation, the finite difference form relates three successive unknowns $u_{i-1/2}^{n+1}$, ς_i^{n+1} and $u_{i+1/2}^{n+1}$ defined at the time level $(n+1)\ \Delta t$ to known values at time level $n\Delta t$ as

$$\frac{\varsigma_i^{n+1} - \varsigma_i^n}{\Delta t} + \frac{{}^*h_{i+1/2}^n\, u_{i+1/2}^{n+\theta} - {}^*h_{i-1/2}^n\, u_{i-1/2}^{n+\theta}}{\Delta x} = 0 \quad (12)$$

where

$$u^{n+\theta} = (1-\theta)u^n + \theta u^{n+1} \quad (13)$$

and

Δt = the time step along the t-axis (s)
Δx = the space step along the x-axis (m)
n = a superscript denoting the time-step number along t
i = a subscript denoting the space step number along x
θ = the time-step weighting coefficient.

The symbol * in equation (12) indicates that the value of h at this grid point has to be approximated by its nearest available upstream value along the grid. For a positive flow direction, for example, this means that $h_{i+1/2}$ is approximated by the value of h_i.

Equation (12) can be reformulated as

$$\alpha 1_i u_{i-1/2}^{n+1} + \beta 1_i \varsigma_i^{n+1} + \gamma 1_i u_{i+1/2}^{n+1} = \delta 1_i \quad (14)$$

where $\alpha 1_i$, $\beta 1_i$, $\gamma 1_i$, and $\delta 1_i$ are the coefficients of the linearized implicit finite difference scheme.

Equation (12) can also be rewritten to provide a choice of time step that guarantees the computation of positive water depths, as:

$$h_i^{n+1} = \left(1 - u_{i+1/2}^{n+\theta}\frac{\Delta t}{\Delta x}\right) h_i^n + u_{i-1/2}^{n+\theta}\frac{\Delta t}{\Delta x} h_{i-1}^n \quad (15)$$

for

$$u_{i-1/2}^{n+\theta} \geq 0 \quad \text{and} \quad u_{i+1/2}^{n+\theta} \geq 0 \quad (16)$$

Equation (15) shows that for positive water velocities and for positive water depths, the *velocity Courant number* condition

$$u_{i+1/2}^{n+\theta}\frac{\Delta t}{\Delta x} \leq 1; \text{ or } \Delta t \leq \frac{\Delta x}{u_{i+1/2}^{n+\theta}} \quad (17)$$

can be applied to provide a sufficient condition for obtaining positive water depths at the new time level. This implies that for the time-step limitation of equation (17), newly computed water levels can never fall below the bottom of the channel. As a consequence, no artificial bottom slots or other arrangements are required to avoid numerical robustness problems for small water depths. A similar condition can be derived for negative flow directions.

Following the same procedure as applied to the continuity equation, the momentum equation (11) can now be defined at a velocity point, by the finite difference form

$$\frac{u_{i+1/2}^{n+1} - u_{i+1/2}^n}{\Delta t} + a(u^n, u^n) + g\frac{\varsigma_{i+1}^{n+\theta} - \varsigma_i^{n+\theta}}{\Delta x} + c_f \frac{\left|u_{i+1/2}^n\right| u_{i+1/2}^{n+1}}{{}^*h_{i+1/2}^n} = 0 \quad (18)$$

where the symbol * has the same meaning as in equation (12) and $a(u^n, u^n)$ is a generalization of the discretization of the convective momentum term.

Appropriate formulations for $a(u^n, u^n)$ follow from the physical conditions, as detailed by Stelling and Duinmeijer (2003). In that paper, it is shown that the correct formulation of the convective momentum term depends on the way in which the convective speed of momentum is interpolated on the grid. As a rule, the discretization is based upon a transformation of the convective momentum term to

$$u\frac{\partial u}{\partial x} = \frac{1}{h}\left\{\frac{\partial (uq)}{\partial x} - u\frac{\partial q}{\partial x}\right\} \quad (19)$$

and its discretization

$$u\frac{\partial u}{\partial x} \simeq \frac{1}{\overline{h}_{i+1/2}}\left(\frac{{}^*u_{i+1}\overline{q}_{i+1} - {}^*u_i\overline{q}_i}{\Delta x} - u_{i+1/2}\frac{\overline{q}_{i+1} - \overline{q}_i}{\Delta x}\right) \quad (20)$$

where

$$\overline{h}_{i+1/2} = \frac{h_i + h_{i+1}}{2}; \quad \overline{q}_i = \frac{q_{i-1/2} + q_{i+1/2}}{2}$$

and $\quad q_{i+1/2} = {}^*h_{i+1/2}u_{i+1/2}$

Also here the value of $^*h$ has the same meaning as in equation (12). The values of $^*u_i$ are missing at ζ-points on the grid and are approximated by first-order unwinding as

$$^*u_i = u_{i-1/2}, \text{ if } \frac{q_{i-1/2} + q_{i+1/2}}{2} \geq 0$$

$$\text{and} \quad ^*u_i = u_{i+1/2}, \text{ if } \frac{q_{i-1/2} + q_{i+1/2}}{2} < 0 \quad (21)$$

For positive flow direction, this yields the simple expression

$$u\frac{\partial u}{\partial x} \simeq \frac{\bar{q}_i}{\bar{h}_{i+1/2}}\left(\frac{u_{i+1/2} - u_{i-1/2}}{\Delta x}\right) \quad (22)$$

This momentum conservative approximation is applied in cases of gradually varying flow or in flow expansions. Referring to Stelling and Duinmeijer (2003) again, this formulation should not be used in strong flow contractions, as this will breed energy. Instead, the energy head conservation formulation should be used, given by

$$u\frac{\partial u}{\partial x} + g\frac{\partial \varsigma}{\partial x} = g\frac{\partial}{\partial x}\left(\frac{u^2}{2g} + \varsigma\right) \tag{23}$$

and, for positive values of u, the energy head conservative upwind discretization

$$u\frac{\partial u}{\partial x} \simeq \frac{u_{i+1/2}^2 - u_{i-1/2}^2}{2\Delta x} = \frac{1}{2}(u_{i-1/2} + u_{i+1/2})\left(\frac{u_{i+1/2} - u_{i-1/2}}{\Delta x}\right) \tag{24}$$

For negative flow velocity, this complete expression is shifted one grid point in the positive x-direction. Comparison of equations (22) and (24) leads to the conclusion that the difference in obtaining momentum or energy conservation lies in the way in which the convective velocity of momentum is interpolated from the flow field.

Terms in equation (18), including (20) or (24) can be collected to give the generalized relation

$$\alpha 2_i \varsigma_i^{n+1} + \beta 2_i u_{i+1/2}^{n+1} + \gamma 2_i \ \varsigma_{i+1}^{n+1} = \delta 2_i \tag{25}$$

By successive elimination of all equations at u-points, the remaining set is given by

$$\alpha_i \varsigma_{i-1}^{n+1} + \beta_i \varsigma_i^{n+1} + \gamma_i \ \varsigma_{i+1}^{n+1} = \delta_i \tag{26}$$

and is solved by applying elimination or conjugate gradient techniques, as discussed further down in this contribution.

The numerical discretization is second-order accurate for all terms of the Saint Venant equations, except for the convective momentum term, which is first-order accurate. For many practical applications, the convective momentum term is of lesser importance and this lower discretization accuracy is quite acceptable then. However, near second-order accuracy can be achieved by up-winded second-order extrapolation of u-values, combined with slope limiters, as described by Stelling and Duinmeijer (2003), again.

In rapidly varying flow, the convective momentum term becomes locally dominant. Modeling these types of flow requires an appropriate implementation of the convective momentum term. By doing so, Delft Hydraulics' SOBEK package has enabled the highly accurate and robust modeling of phenomena such as supercritical flow in steep channels and moving hydraulic jumps.

The numerical scheme given by equations (14) and (25) only provides linear equations at internal grid points. At channel boundaries, additional equations are required. As the Saint Venant equations are of second-order hyperbolic type, one boundary condition is required for each of its characteristic lines entering the computational domain. Boundary condition requirements for 1D river models were discussed extensively by Cunge *et al.* (1986). Here we will limit ourselves by stating that in most practical applications, inflowing discharges are specified at the upstream ends of channels entering the flood model domain and water levels or rating curves at channels leaving the model domain. At internal boundaries, such as channel junctions, usually a modified continuity equation is applied, jointly with water level compatibility at all channel boundaries at that junction.

HYDRAULIC STRUCTURES

Nearly all flood simulation models deal with hydraulic structures, such as dams, weirs, bridges, and so on. The length of the hydraulic structure along the x-axis is usually supposed to be negligible on the scale of the river or channel, and therefore the structure can be seen as one single point along x, where both water- and energy levels are discontinuous and the Saint Venant equations do not apply. At this point a relation is established between the upstream water level, the discharge through the structure, and the downstream water level. In this relation, it is assumed that the upstream water level is taken at the nearest point just upstream of the structure, where vertical accelerations can still be neglected. Similarly, the downstream water level is taken at the nearest location just downstream of the structure, where the flow can be seen as nearly horizontal and the structure flow no longer contributes to the energy dissipation. Usually, it is also assumed that storage of water around the structure is negligible, so the relationship is applicable at any point in time. On this basis, there is a variety of ways in which the flow through hydraulic structures can be formulated:

- in most cases, hydraulic structure descriptions are based upon empirical laws relating the discharge through the structure to the upstream and downstream water level. The relation has the general form,

$$Q_{\text{crest}} = Q(\varsigma_{\text{up}}, \varsigma_{\text{down}}) \tag{27}$$

  where ζ_{up} is the water level upstream of the structure (m), Q_{crest} the discharge through the structure ($\text{m}^3\ \text{s}^{-1}$), and ζ_{down} the water level downstream of the structure (m). This formulation includes the possibility of using energy levels instead of water levels. Equation (27) may be linearized to

$$\alpha \varsigma_{\text{up}} + \beta Q_{\text{crest}} + \gamma \varsigma_{\text{down}} = \delta \tag{28}$$

where α, β, γ, and δ are the local values of nonlinear coefficients at a given state of the flow. A wide variety of structure descriptions is available from literature, for example, the classic books of Chow (1959), Henderson (1966) and, more recently, of Chanson (1999);

- as an alternative, the state of the flow in the section just upstream of the structure, where the flow is contracted and vertical accelerations occur, can be described by an energy conservation principle. Similarly, the state of the flow just downstream of the structure, where the flow expands with energy losses associated to it, can be described by a momentum and impulse balance relationship. By internal elimination of unknowns in these relationships, the water level at the structure crest or at its most contracted section can be eliminated, and a relation similar to equation (28) is obtained;
- in cases where it is difficult to define the state of the flow in terms of equations, laboratory experiments may be set up to define a matrix relating upstream and downstream water levels to the structure discharge. As an alternative, the structure may be modeled in detail by a 3D numerical code, leading to a similar set of matrix coefficients. Conditions are that a fine grid is used, and that the code is based upon an appropriate numerical description of convective momentum terms and the effect of turbulence.

In all cases discussed previously, the structure equations could be based upon energy levels, instead of water levels. Through suitable transformations, these relations can be reworked to the form of equation (28). Similar descriptions can be made for local energy loss descriptions along a channel.

Appropriate linearization of equation (27), or of the other structure flow descriptions, leads to a numerical scheme of the form of equation (25). In the total set of equations for channel flow, this structure relationship replaces the momentum equation that is generally applied between successive ζ-points.

In the case of a closed structure or a discharge specified by a pump, the coefficients α and γ of equation (25) are both zero, and the structure presents itself as an internal boundary condition with a discharge given. In the case of free flow in the positive x-direction, only the coefficient γ equals zero, showing also in the numerical relationship that the structure discharge is only dependent on the upstream water level.

In the literature, an abundant number of structure equations can be found (Chow, 1959; Henderson, 1966; and more recently, Chanson, 1999). Implementing these in a numerical code requires attention for discharge compatibility among the various flow states, especially at the transition between free and submerged flow. Incompatibility in the definition of the discharge may lead to oscillations in the computed flow, in particular, at the moment of flow reversal. Problems may be suppressed by defining some inertia to the structure flow. This is implemented by adding a small term to the coefficients β and δ. The numerical structure behavior is generally better when applying an energy relationship in the accelerating flow section upstream of the structure combined with a momentum–impulse relationship for the section downstream of the structure. Where appropriate, energy losses may be added to the description of flow in the contracted section. Internal elimination of local variables leads, again, to a relationship as given by equation (25). In all cases, the different flow states only affect the computation of the coefficients α, β, γ, and δ and have no bearing on the solution algorithm of the resulting system of equations.

TWO-DIMENSIONAL MODELING

Following the same principles as those leading to the 1D flow equations, the two-dimensional (2D), shallow-water equations read

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} + \frac{\partial (vh)}{\partial y} = 0 \tag{29}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + g\frac{\partial (h + z_b)}{\partial x} + c_f \frac{u\sqrt{u^2 + v^2}}{h} = 0 \tag{30}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + g\frac{\partial (h + z_b)}{\partial y} + c_f \frac{v\sqrt{u^2 + v^2}}{h} = 0 \tag{31}$$

where we now also introduce the y-axis, orthogonal to the x-axis, with its flow velocity v ($\mathrm{m\,s^{-1}}$) associated to it. Basic assumptions are similar to those given for the 1D equations, as far as applicable in this form of schematization.

Referring to the 2D grid shown in Figure 2, a volume conservative finite difference form of the continuity equation is given by:

$$\frac{\zeta_{i,j}^{n+1} - \zeta_{i,j}^{n}}{\Delta t} + \frac{{}^{*}h_{i+1/2,j}^{n} u_{i+1/2,j}^{n+\theta} - {}^{*}h_{i-1/2,j}^{n} u_{i-1/2,j}^{n+\theta}}{\Delta x} + \frac{{}^{*}h_{i,j+1/2}^{n} v_{i,j+1/2}^{n+\theta} - {}^{*}h_{i,j-1/2}^{n} v_{i,j-1/2}^{n+\theta}}{\Delta y} = 0 \tag{32}$$

$$\frac{u_{i+1/2,j}^{n+1} - u_{i+1/2,j}^{n}}{\Delta t} + a_{11}(u^n, u^n) + a_{12}(v^n, u^n) + g\frac{\zeta_{i+1,j}^{n+\theta} - \zeta_{i,j}^{n+\theta}}{\Delta x} + c_f \frac{u_{i+1/2,j}^{n+1}\sqrt{\left(u_{i+1/2,j}^{n}\right)^2 + \left(\overline{\overline{v_{i+1/2,j}^{n}}}\right)^2}}{{}^{*}h_{i+1/2,j}^{n}} = 0 \tag{33}$$

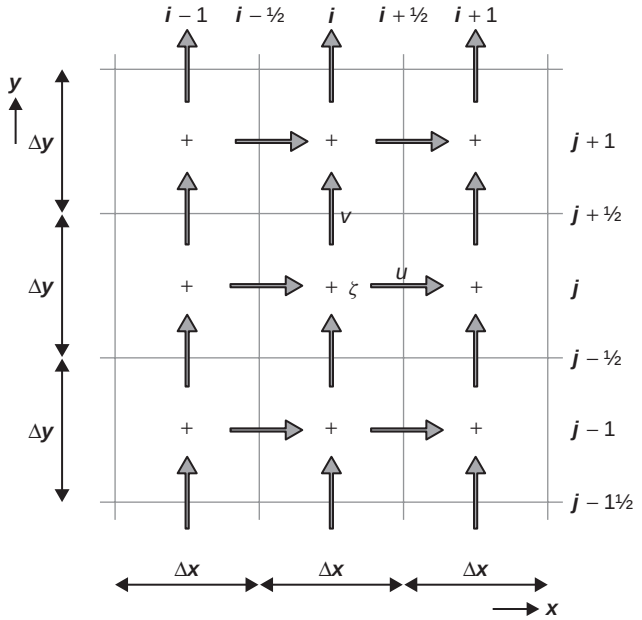


Figure 2 Staggered grid for 2D flow simulations

$$\frac{v_{i,j+1/2}^{n+1} - v_{i,j+1/2}^{n}}{\Delta t} + a_{21}(u^n, v^n) + a_{22}(v^n, v^n) + g\frac{\varsigma_{i,j+1}^{n+\theta} - \varsigma_{i,j}^{n+\theta}}{\Delta y} + c_f \frac{v_{i,j+1/2}^{n+1}\sqrt{\left(\overline{\overline{u_{i,j+1/2}^{n}}}\right)^2 + \left(v_{i,j+1/2}^{n}\right)^2}}{{}^{*}h_{i,j+1/2}^{n}} = 0 \quad (34)$$

where the symbol * has the same meaning as in equation (12) again and $a_{11}(u^n, u^n)$, $a_{12}(v^n, u^n)$, $a_{21}(u^n, v^n)$, and $a_{22}(v^n, v^n)$ are generalizations of the discretization of the convective momentum term. The long double bar over the velocity in the friction term means that this velocity is obtained by averaging over values at four surrounding grid points. The friction term requires special treatment in case of flooding of dry terrain. At the wave front the water velocity rapidly accelerates from zero. Overshoot of velocities can be prevented by a predictor–corrector approach.

The convective momentum terms are subject to the same principles as discussed for the 1D approximations. For example, for positive flow velocities the momentum conservative discretization of the term $a_{12}(u^n, v^n)$ is given by

$$a_{12}(v^n, u^n) \simeq \frac{\overline{{}_{v}q}^{x}_{i+1/2,j-1/2}}{\overline{h}^{x}_{i+1/2,j}} \left(\frac{u_{i+1/2,j} - u_{i+1/2,j-1}}{\Delta y}\right) \quad (35)$$

whereas, it is given by

$$a_{12}(v^n, u^n) \simeq \overline{v}^{x}_{i+1/2,j-1/2} \left(\frac{u_{i+1/2,j} - u_{i+1/2,j-1}}{\Delta y}\right) \quad (36)$$

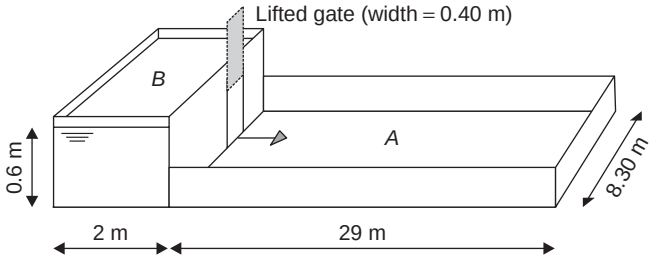


Figure 3 Layout of the instantaneous dam break experiment

for the energy conservative discretization. In the first expression $\overline{{}_{v}q}^{x}$ means the specific discharge in y-direction averaged over two surrounding points along the local x-axis. In the last expression $\overline{v}^{x}$ has the same meaning in relation to the velocity v.

The treatment of the convective momentum terms shown in the preceding text is numerically very robust and allows for the correct description of the effects of sudden expansions and contractions and similar changes in the topography, such as steps in the bed level. Moreover, it allows for the 2D simulation of supercritical flows and the propagation of hydraulic jumps.

As described by Stelling and Duinmeijer (2003), the correct modeling of these phenomena has been demonstrated in a software validation study, where results of the numerical scheme of equations (32), (33) and (34) were compared with video monitored measurements in a physical model (Figure 3). The setup consists of two reservoirs with different water levels, separated by a wall. The wall contains a gate which can be lifted. The width of both reservoirs is 8.30 m, the length of the upper reservoir 2 m, and the length of the lower reservoir 29 m. The gate has a width of 0.4 m and has been placed at the middle of the wall.

The numerical experiment was made on a grid of 0.1 × 0.1 m, giving a total of approximately 25 000 grid cells. The time step was set to 0.005 s. The gate was lifted at a speed of 0.16 m s^{-1}, to produce a flow spreading out into a two-dimensional plain. Initial data were set at a depth of 0.60 m for the upstream reservoir and a depth of 0.05 m for the downstream reservoir.

Figure 4 shows the results of this simulation. Figure 4(a) presents a video recorded view from above. Figure 4(b) presents the computed results. It is clearly seen that the front propagation, the propagation of the hydraulic jump, and the side-spreading of the wave are represented reasonably well. Figure 5 shows a comparison of the measured and computed position of the wave front at various times.

Simulations were made with various Manning roughness coefficients and both for an initially wet and a dry downstream reservoir. For the propagation of the flood on the dry bed, the Manning roughness turned out to be a sensitive parameter. If, for dam break models, a reliable

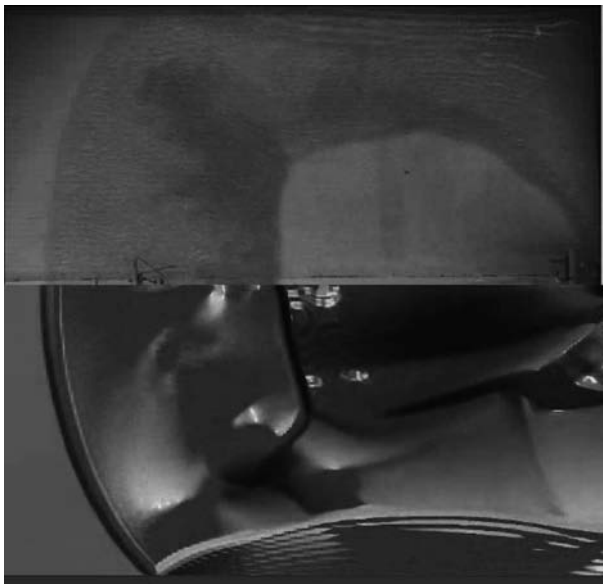


Figure 4 Top view of results of the numerical scheme of equations (32), (33), and (34) (b) compared with video monitored measurements in a physical model (a). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

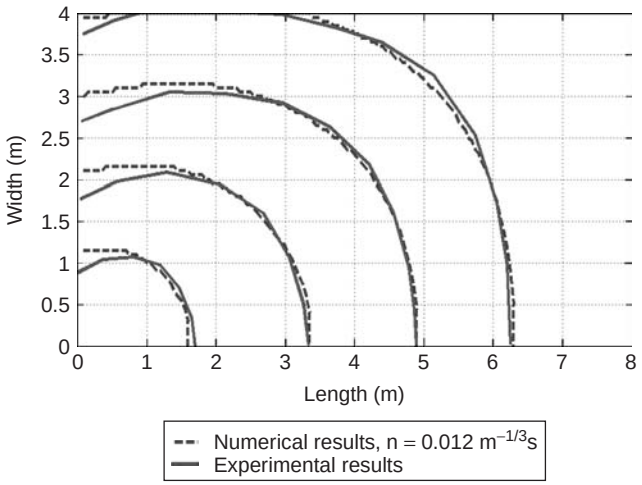


Figure 5 Comparison of measured and computed wave front position at various times (see case Figure 4). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

topography is available, the roughness parameter remains the only parameter to be estimated. Currently, researchers are focusing on better descriptions of roughness parameters by deriving depth-dependent relationships on the basis of vegetation characteristics; see for example, Uittenbogaard (2003), Rodriguez Uthurburu (2004) and Baptist (2005).

INTEGRATED 1D/2D MODELING

In flood modeling, there are numerous practical examples where flows are best described by combinations of 1D and 2D schematizations. An obvious example is the flooding of delta areas, often characterized by a flat topography with complex networks of natural levees, polder dikes, drainage channels, elevated roads and railways, and a large variety of hydraulic structures. Flow over the terrain is best described by the 2D equations, whereas channel flow and the role of hydraulic structures are satisfactorily described in 1D. Flow over higher elevated line elements, such as roads and embankments, can be reasonably modeled in 2D by raising the bottom of computational cells to embankment level. However, for a higher accuracy of the numerical description, adapted formulations have to be applied, such as energy conservation upstream of overtopped embankments.

Another example is the flood propagation in a meandering river, with shortcuts via the flood plain when over-bank flow occurs. In large-scale models, the flow between the riverbanks is satisfactorily described by the Saint Venant equations, solved with 1D grid steps several times the width of the channel. An equivalent accuracy of description in 2D would require a large number of grid cells with step sizes of a fraction of the channel width. However, flow in the flood plain may be better described in 2D and may allow for 2D grid steps often exceeding the width of the river, as made plausible for the flood plain shown in Figure 6.

For this reason, hybrid 1D and 2D schematizations are often used. Basically there are two approaches: one with

Figure 6 Small river with a large flat flood plain in the south of The Netherlands. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

interfaces defined between 1D and 2D along vertical planes and the other with schematization interfaces in almost horizontal planes.

Coupling along vertical planes gives a full separation in the horizontal space of the 1D and 2D modeled domains. In the 1D domain, the flow is modeled with the Saint Venant equations applied over the full water depth. The direction of flow in the 1D domain is assumed to follow the channel x-axis and in the model, it carries its momentum in this direction, also above bank level. Without special provisions, there is no momentum transfer accounting applied between the 1D and 2D domains. Momentum and volume entering or leaving the 2D domain at these interfaces are generated by the compatibility condition applied. As a result, the coupling cannot be expected to be momentum-conservative. Depending on the numerical solution applied, the linkage may either be on water level or on discharge compatibility. Particular care has to be taken in applying this form of schematization if water quality processes are to be included in the model.

In a model coupled along an almost horizontal plane, 2D grid cells are placed above the 1D domain, as shown in Figure 7. In this schematization, the Saint Venant equations are applied only up to bank level. Above this level, the flow description in the 2D cell takes over. For relatively small channel widths compared to the 2D cell size, errors in neglecting the effect of momentum transfer at the interface are minor. This would be the case for a 1D/2D model built for the situation shown in Figure 6. For wider channels, it is recommended to modify each 2D cell depth used in the momentum equation by adding a layer defined by the local hydraulic radius for that part of the 1D cross section which underlies a 2D cell. Further refinements are possible, including terms describing the momentum transfer between the 1D and 2D domains.

Numerical solutions are obtained by discretizing separately the 1D and 2D domains. Assuming that for both

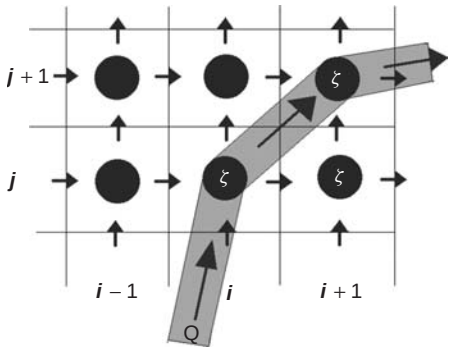


Figure 7 Coupling of 1D and 2D domains in SOBEK. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

domains implicit numerical schemes are applied, the interface compatibility conditions can be modeled either as an explicit or an implicit link. Applying explicit links, first the solutions for the 1D and 2D domains are generated sequentially. Subsequently, exchange flows are computed and added as lateral flows at the next time step. Implicit links are based upon water level compatibility. These equations are then added to the complete sets of equations generated separately for the 1D and 2D domains. There are many approaches to solving the complete set of equations. With the current state of the art, it is no longer necessary to apply for the 1D domain different solvers for so-called *simply-* or *multiply connected channel networks*. Similarly, in 2D there is no real need anymore for alternating direction algorithms, as the efficiency of the conjugate gradient solvers has increased significantly over the past years.

As an example, Delft Hydraulics has developed its combined 1D/2D package SOBEK for the modeling of integrated freshwater systems. The 1D part of hybrid models is based upon the numerical scheme of equations (12) and (18). The 2D part is described by the single step 2D scheme given by equations (32), (33), and (34). For efficiency reasons, the continuity equations for the 1D and 2D domains are combined into one single equation at points where 1D grid sections underlie a 2D cell. As a first step in reducing the total number of equations, SOBEK eliminates all equations at velocity grid points. The second step in the solution algorithm is the elimination of a large number of unknowns by applying a minimum connection search between unknown water levels. As a rule, this leads to an efficient elimination of nearly all unknowns of the 1D domain and a substantial number of unknowns in the 2D domain. This direct solver carries its elimination on, until nearly every second equation in the 2D domain has been eliminated. Beyond this point, it is more economical to apply the conjugate gradient solver to solve the remaining set of equations.

Apart from its efficiency, an additional advantage of eliminating nearly every second 2D equation is the improved conditioning of the resulting matrix. This follows from the fact that elimination of an unknown water level at a 2D grid point has the effect of increasing the spatial distance between the remaining adjacent points, where water levels are still unknown. This, in turn, reduces Courant numbers and as a consequence leads to changed coefficients at the main diagonal of the matrix, which is now more dominant in relation to the other diagonals; see for example, Verwey (1994).

An example of a combined 1D/2D model is shown in Figure 8. It represents the schematization of a model of the Eem Valley area in The Netherlands applied in a study of the potential effects of a River Rhine dike breach. This model has been used to provide information on warning lead times and flood depths for evacuation planning. The

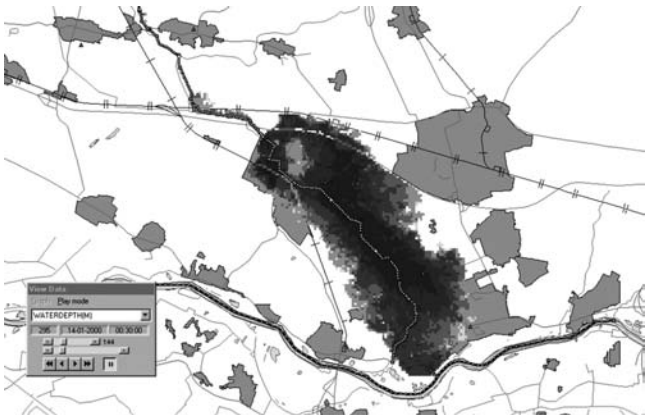


Figure 8 Flood modeling of the Vallei and Eem area, The Netherlands. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Rhine branch upstream of the breach has been modeled in 1D. At its upstream end a design hydrograph was specified, whereas the downstream boundary condition of this relatively short branch is given by a rating curve. Such a short downstream reach is permissible as the rating curve automatically corrects for most of the effects of flow deviated through the breach at this boundary. The breach itself has been described in 1D as a structure with a velocity dependent breach growth. North of the dike, the 1D link discharges into the 2D domain, given by a 100×100 m grid with bottom levels derived from a digital elevation model and resistance coefficients derived from land use maps. Elevated roads and railways are presented as flow barriers by raising the underlying cell bottom levels up to the levels of these embankments. The resulting flood depths presented in Figure 8 clearly show the effect of the 1D channel in the schematization. Because of their greater depth, flood waves propagate faster in these channels than over land. Further downstream, this leads to first signs of the progressing flood wave already one or two days before the main flood arrives.

CONCLUDING REMARKS

The state of the art of numerical flood simulation has progressed significantly over the last decades of the twentieth century and has come to the point where a quite realistic picture of potential flood threats can be produced at very reasonable costs. On the one hand, new data collection techniques have emerged which alleviate the traditional problem of lack of data for model construction. On the other, numerical techniques have matured, providing robustness and efficiency in model simulation.

Such simulation models have proven to be of great value for flood forecasting, flood control, dam break analysis, and flood damage assessment. In flood forecasting, the numerical flood simulation models, or hydraulic models, gradually replace the traditional hydrologic modeling methods such as Muskingum. In flood control, dam break analysis, and flood damage assessment, there is a much longer tradition of using hydraulic models. In these areas we see a gradual replacement of 1D models by integrated 1D/2D models. Typical applications are design of flood protection works, both in rural and urban areas, design and operation of flood retention basins, studies of the effect of climate change, evacuation planning, and GIS-based postprocessing of model results for damage assessment and flood insurance support.

Over the past decades significant progress has been made in the numerical solution of the unsteady flow equations, both in 1D and in 2D. On the basis of bathymetric and topographic information, robust and efficient simulations of flood wave propagation can be produced. Numerical models are now able to model without problems, flooding and drying of floodplains and all sorts of sudden variations in river or channel bathymetry or floodplain topography. Also, there is no longer a problem in modeling sudden transitions between sub- and supercritical flow states, such as frequently encountered in dam break analyses.

Robustness of numerical flood simulation models is not only based upon the use of unconditionally stable numerical schemes. Because of the highly nonlinear nature of the unsteady flow equations, additional conditions have to be imposed, such as the velocity Courant number condition, to avoid negative water depths. Sudden variations in channel cross-sections are modeled better now by applying energy conservation principles in strong flow contractions and correct momentum principles in sudden flow expansions. In addition, other time-step limiters or artificial inertia may be necessary to avoid problems, such as nonconvergence of numerical solutions, mass balance problems, oscillating hydraulic structure flow, and so on. Currently, the nature of most of these problems is well understood.

The construction of numerical flood models has been facilitated much by the development of new data collection equipment, such as LIDAR, ADCPs, multibeam echo sounders, and advanced radar technology. The use of these technologies, jointly with the increased speed of computers, has led to the gradual replacement of 1D models by 2D models for floodplain, wetland, and shallow lagoon modeling. As some water bodies, such as rivers and canals are, for large-scale flood routing problems, still modeled better in 1D, the use of hybrid 1D/2D models has emerged. Coupling of these models may be along vertical planes or along a nearly horizontal plane. This last option provides the most realistic representation of physical processes.

The speed of computers is still increasing continuously. Currently we are able to run a model in one single minute, where this took one hour one decade ago. This has a

strong impact on the practical applications of models. In the coming years, much of this additional speed will be mobilized for simulations on finer grids, to make better use of the details provided by digital elevation models and the use of more refined roughness descriptors, such as those based upon vegetation characteristics. This area still leaves much space for further research. Much research also is needed on improved flood generating breach development descriptions, based upon the physical description of soil–water interactions.

Over the past decade the emerging field of hydroinformatics has provided many new technologies supporting data acquisition, processing, and mining leading to the development of new knowledge, such as more correct empirical relationships, improved optimization methods, and new simulation techniques (*see* **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1**; **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**; **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**; **Chapter 23, Flood Early Warning Systems for Hydrological (sub) Catchments, Volume 1**; **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**; **Chapter 60, Estimation of River and Water-Body Stage, Width and Gradients Using Radar Altimetry, Interferometric SAR and Laser Altimetry, Volume 2**; **Chapter 61, Estimation of River Discharge, Volume 2**; **Chapter 138, Unsteady Flow, Volume 4**; **Chapter 139, Numerical Modeling of Unsteady Flows in Rivers, Volume 4**; **Chapter 141, Computer Modeling of Overbank Flows, Volume 4** and in particular, abbott and mynett, **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**). These developments are likely to have further great impact on the way hydrologists are composing their tool boxes in the years to come.

REFERENCES

Abbott M.B. (1979) *Computational Hydraulics–Elements of the Theory of Free Surface Flows*, Pitman: London/San Francisco/Melbourne.

Arakawa A. (1966) Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow, Part I. *Journal of Computational Physics*, **1**(1), 119–143.

Baptist M.J. (2005) *Modelling Floodplain Biogeomorphology*, Ph.D. thesis, ISBN 90-407-2582-9, Delft University of Technology, Faculty of Civil Engineering and Geosciences, Section Hydraulic Engineering, p. 193.

Chanson H. (1999) *The Hydraulics of Open Channel Flow*, Arnold Publishers/Wiley: Paris/New York.

Chaudhry M.H. (1993) *Open-Channel Flow*, Prentice Hall: Anglewood Cliffs.

Chow V.T. (1959) *Open Channel Hydraulics*, McGraw-Hill: New York.

Cunge J.A. (1969) On the subject of a flood propagation computation method (Muskingum method). *Journal Hydrology Research*, **7**, 2.

Cunge J.A., Holly F.M. and Verwey A. (1986) *Practical Aspects of Computational River Hydraulics*, Pitman Publishing, London, Great Britain, 1980. Reprinted at Iowa Institute of Hydraulic Research, p. 420. (Also translated into Russian).

De Saint Venant B. (1871) Théorie du mouvement non-permanent des eaux avec application aux crues des rivieres et a l'introduction des marees dans leur lit, *Comptes Rendus de l'Académie des Sciences*, **73**, 148–154, 237–240.

Henderson F.M. (1966) *Open Channel Flow*, McMillan Company: New York.

Hesselink A.W., Stelling G.S., Kwadijk J.C.J. and Middelkoop H. (2003) Inundation of a Dutch river polder, sensitivity analysis of a physically based inundation model using historic data. *Water Resources Research*, **39**(9), 1234.

Hirsch C. (1990) *Numerical Computation of Internal and External Flows*, Wiley: New York.

Newton I. (1687) *Mathematical Principles of Natural Philosophy*, translated by Motte A., Cajuri F. (Ed.), Berkeley University Press: California (1947).

Rodriguez Uthurburu R. (2004) *Evaluation of Physically Based and Evolutionary Data Mining Approaches for Modelling Resistance Due to Vegetation in SOBEK 1D-2D*, M.Sc. thesis HH 485, UNESCO-IHE, Delft.

Serban P., Crookshank N.L. and Willis D.H. (2005) *Intercomparison of Forecast Models for Streamflow Routing in Large Rivers*, WMO: Geneva.

Stelling G.S. and Duinmeijer S.P.A. (2003) A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *International Journal for Numerical Methods in Fluids*, **43**, 1329–1354.

Stelling G.S., Kernkamp H.W.J. and Laguzzi M.M. (1998) Delft flooding system: a powerful tool for inundation assessment based upon a positive flow simulation. In *Hydroinformatics '98*, Babović V. and Larsen L.C. (Eds.), Balkema: Rotterdam, pp. 449–456.

Stoker J.J. (1957) *Water waves. Pure and Applied Mathematics*, Interscience Publishers: New York, Vol. IV.

Toro E.F. (1999) *Riemann Solvers and Numerical methods for Fluid Dynamics*, Springer: Berlin.

Uittenbogaard R. (2003) Modelling turbulence in vegetated aquatic flows, *International Workshop on RIParian FORest Vegetated Channels: Hydraulic, Morphological and Ecological Aspects*, Trento, 20-22 February 2003.

Verwey A. (1994) Linkage of physical and numerical aspects of models applied in environmental studies, keynote lecture in: *Proceedings of the Conference on Hydraulics in Civil Engineering*, Brisbane.

Verwey A. (2001) Latest developments in floodplain modelling – 1d/2d integration, keynote lecture in: Proceedings of the 6th Conference on Hydraulics in Civil Engineering, Hobart.

Werner M.G.F., van Dijk M. and Schellekens J. (2004) DELFT-FEWS: an open shell flood forecasting system. In *Proceedings of the 6th International Conference on Hydroinformatics*, Liong S.Y., Phoon K.K. and Babovič V. (Eds.), World Scientific Publishing Company: Singapore, pp. 1205–1212.

17: Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries

BINLIANG LIN AND ROGER A FALCONER

Cardiff School of Engineering, Cardiff University, Cardiff, UK

Hydroinformatics tools are increasingly being used for predicting hydrodynamic, hydrological, and water quality processes in river and estuarine waters. The numerical solutions of the governing hydrodynamic and solute and sediment transport equations form the basis of these software tools. Details are given herein of the governing equations for three-, two- and one-dimensional flow-field predictions and the transport of water quality indicators and sediment transport processes within the water column. Two case studies are then discussed. In the first of these studies linked 1-D and 2-D models are used to predict the velocity and fecal coliform distributions for a relatively large river and estuarine system, including the Ribble Estuary, its tributaries, and the Fylde coast. In the second of these studies, the velocity, suspended sediment, and heavy metal concentration distributions are predicted along the Humber Estuary using 2-D and 3-D models, with dynamic partitioning coefficients being used to link the metal concentrations and sediments. Details are also given of the hydrological models used to specify the pollutant discharge loads from land used models.

INTRODUCTION

In recent years there has been growing public concern about the quality of water within many river and estuarine systems, particularly in those parts of the world where rivers and estuaries have become increasingly used as receiving water bodies for the discharges of domestic effluents, industrial by-products, agricultural waste, and urban drainage. Clearly this has been an important driver for computer-based modeling paradigms in hydrology, as described by Abbott *et al.* (2005; *see* **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**). Human and aquatic life is often threatened by the transport of pollutants through riverine systems to coastal waters, and it is therefore not surprising to find that from a water quality point of view, rivers have been studied more extensively and for longer periods of time than any other bodies of water (Thomann and Mueller, 1987). This is partly due to the fact that many people live close to, or interact with, rivers and streams.

An estuary is a semienclosed coastal body of water that has a free connection with the open sea and within which seawater is measurably diluted with freshwater derived from land drainage and so on (Dyer, 1997). It includes a mixing zone that moves in position according to factors such as tide, wind, and freshwater flow. Consequently, water quality in estuaries usually exhibits marked spatial and temporal variations and involves a range of abiotic as well as biotic processes, as outlined by Mynett (2002). Salinity levels increase from zero at the upstream tidal limit to approximately the same value as that of the open sea. For a well-mixed estuary, these variations produce a water quality gradient along the estuary (Wallingford, 1996). The high, suspended sediment load in most rivers and estuaries provides a large surface area for the attachment of chemical and biological species, including heavy metals arising from industrial waste discharges and bacterial microorganisms arising from urban domestic wastewater discharges.

With increasing awareness of all aspects of hydroinformatics and hydroecological pollution, there has been a marked increase in recent years in the development and application of numerical models to predict river and estuarine water quality characteristics. One of the main reasons for this increase is that numerical models offer powerful tools for investigating the complex interactions between velocities, sediment fluxes, and water quality indicators

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

(Falconer *et al.*, 2001). Water quality modeling originally focused on river and stream pollution, with one-dimensional numerical models often being used for modeling river flow and transport processes (Stelling and Verwey, 2005; *see* **Chapter 16, Numerical Flood Simulation, Volume 1**) as well as water quality assessment. Because of the spatial variation in the distributions of flow velocities and solute concentration distributions, two- and three-dimensional (2-D and 3-D) models are often used for estuarine problems.

The primary processes occurring in rivers and estuaries are usually classified separately as hydrodynamic and transport processes, thus these two processes are also often numerically treated separately. The main objectives of this chapter are to present the governing differential equations and parameters describing the hydrodynamic, water quality, and sediment transport processes occurring in rivers and estuaries and to review recent applications of hydroinformatics tools for modeling water quality indicators. Details are given of 1-D, 2-D and 3-D modeling approaches, with examples being given on the capabilities of such tools.

HYDRODYNAMIC MODELING

Three-dimensional Flows

The numerical models used by water and environmental engineers and managers to predict the flow, water quality, and sediment and contaminant transport processes in rivers and estuaries are based on first solving the governing hydrodynamic equations. For a Cartesian coordinate system, with the main body of the flow in the x-direction, the corresponding 3-D Reynolds equations for mass and momentum conservation can be written in a general conservative form as (Falconer, 1993; Falconer *et al.*, 2001):

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \tag{1}$$

$$\underbrace{\frac{\partial u}{\partial t}}_{1} + \underbrace{\frac{\partial u^2}{\partial x} + \frac{\partial uv}{\partial y} + \frac{\partial uw}{\partial z}}_{2} = \underbrace{X}_{3} - \underbrace{\frac{1}{\rho}\frac{\partial P}{\partial x}}_{4} - \underbrace{\frac{\partial \overline{u'u'}}{\partial x} + \frac{\partial \overline{u'v'}}{\partial y} + \frac{\partial \overline{u'w'}}{\partial z}}_{5} \tag{2}$$

where u, v, w, are the velocity components in x, y, z coordinate directions respectively, t = time, X = body force in x-direction, ρ = fluid density, P = fluid pressure, and $\overline{u'u'}, \overline{u'v'}, \overline{u'w'}$ = Reynolds stresses in x-direction on x, y, z planes respectively.

Equations similar to (2) can be written to evaluate the velocity components v and w in the y and z directions respectively. The numbered terms in equation (2) refer to: local acceleration (term 1), advective acceleration (2), body force (3), pressure gradient (4) and turbulent shear stresses (5).

In modeling estuarine flows in two and three dimensions, the effects of the earth's rotation need to be included, giving for the body force components

$$\left.\begin{aligned} X &= 2v\varpi \sin\phi \\ Y &= -2u\varpi \sin\phi \\ Z &= -g \end{aligned}\right\} \tag{3}$$

where ϖ = speed of earth's rotation, ϕ = earth's latitude and g = gravitational acceleration. The main effects of the earth's rotation, give rise to the Coriolis acceleration, which can set up transverse water surface slopes across an estuary and enhance the effects of secondary currents and meandering.

For three-dimensional flow predictions, either the full three-dimensional governing equations are solved, which leads to a complex numerical formulation to evaluate the pressure P, or, more usually, a hydrostatic pressure distribution is assumed to occur in the vertical (z) direction, which leads to an expression for P of the following form:

$$P(z) = \rho g(\zeta - z) + P_a \tag{4}$$

where ζ = water surface elevation above (positive) datum and P_a = atmospheric pressure. The corresponding derivative of equation (4), for inclusion in equation (2), gives:

$$\frac{\partial P}{\partial x} = \rho g \frac{\partial \zeta}{\partial x} + \frac{\partial P_a}{\partial x} \tag{5}$$

A similar representation can be written for the pressure gradient in the y-direction. The effects of the atmospheric pressure gradient are generally small in riverine and estuarine flows and are neglected.

As a consequence, the only unknown terms remaining in equation (2) are the Reynolds stresses, which need to be related to the 3-D velocity field before solving for the water levels and the three-dimensional velocity components. In solving for these stresses, Boussinesq (Goldstein, 1938) proposed that they could be represented in a diffusive manner, giving:

$$\left.\begin{aligned} -\overline{u'u'} &= \nu_t \left[\frac{\partial u}{\partial x} + \frac{\partial u}{\partial x}\right] \\ -\overline{u'v'} &= \nu_t \left[\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right] \\ -\overline{u'w'} &= \nu_t \left[\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right] \end{aligned}\right\} \tag{6}$$

where ν_t = kinematic eddy viscosity.

The eddy viscosity coefficient can be obtained in several ways (Falconer and Chen, 1996). The simplest approach is to assume a constant value based on field data. However,

whilst this approach may be adequate for predicting velocity distributions in large water bodies, such as coastal basins, lakes, or reservoirs, it is not particularly accurate for 3-D model simulations in rivers and estuaries, where such models are generally only used for predicting complex velocity field distributions in the vicinity of structures (such as bridge piers) and short river reaches. Another approach is to apply a zero-equation turbulence model, similar to that prescribed by Prandtl's mixing length hypothesis (Goldstein, 1938), wherein:

$$\nu_t = \ell^2 J \tag{7}$$

where $\ell =$ a characteristic mixing length and $J =$ magnitude of local velocity gradients in x, y, z directions. Using this approach, the mixing length can readily be determined for a typical logarithmic type velocity profile, but for the more complex flow fields where 3-D models are appropriate, the velocity distribution is unlikely to be primarily logarithmic in form, and strong secondary currents often lead to a more complex and less well-defined mixing length.

Hence, for most practical problems, where 3-D models are appropriate for river and estuarine simulations, the turbulent stresses given in equation (2) need to be solved using either a two-equation turbulence model of the k-ε type, or an algebraic stress type model wherein the Reynolds stress terms are solved directly (Lin and Shiono, 1995). For the more usual approach, using either the linear or nonlinear k-ε model, the eddy viscosity is defined as:

$$\nu_t = \frac{C_\mu k^2}{\varepsilon} \tag{8}$$

where $C_\mu =$ turbulent model coefficient, $k =$ turbulent kinetic energy and $\varepsilon =$ dissipation rate of turbulent kinetic energy. Transport equations are derived for k and ε (Rodi, 1984), which, in general, include: transport by advection, production, and dissipation.

Two-dimensional Flows

For many practical problems where there are significant variations across the streamwise flow direction, it is commonplace for the velocity field to be determined using a 2-D depth-integrated numerical model. To determine the hydrodynamic velocity field using a 2-D model, the governing three-dimensional equations are integrated over the depth, giving for equations (1) and (2) respectively:

$$\frac{\partial \zeta}{\partial t} + \frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} = 0 \tag{9}$$

$$\frac{\partial q_x}{\partial t} + \beta\left[\frac{\partial U q_x}{\partial x} + \frac{\partial V q_x}{\partial y}\right] = f q_y - gH\frac{\partial \zeta}{\partial x} + \frac{\tau_{xw}}{\rho} - \frac{\tau_{xb}}{\rho} + 2\frac{\partial}{\partial x}\left[\bar{\nu}_t \frac{\partial q_x}{\partial x}\right] + \frac{\partial}{\partial y}\left[\bar{\nu}_t\left(\frac{\partial q_x}{\partial y} + \frac{\partial q_y}{\partial x}\right)\right] \tag{10}$$

where q_x, $q_y =$ discharges per unit width in x, y directions, U, V are the depth-averaged velocities in x, y directions; $\beta =$ momentum correction factor for nonuniform vertical velocity profile; τ_{xw}, τ_{xb} are surface and bed shear stress components respectively in x-direction; and $\bar{\nu}_t =$ depth-averaged eddy viscosity.

The momentum correction factor can be estimated either from field data, which is preferable, or alternatively by assuming a velocity profile, such as a logarithmic distribution, to give:

$$\beta = \left[1 + \frac{g}{C^2\kappa^2}\right] \tag{11}$$

where $C =$ Chezy bed roughness coefficient and $\kappa =$ von Karman's constant ($= 0.41$).

For the surface wind stress a quadratic friction law is generally assumed (Wu, 1969), giving:

$$\tau_{xw} = C_s \rho_a W_x W_s \tag{12}$$

where $C_s =$ air-water resistance coefficient, $\rho_a =$ air density, $W_x =$ wind velocity component in x-direction and $W_s =$ wind speed $(= \sqrt{W_x{}^2 + W_y{}^2})$, where $W_y =$ wind velocity component in y-direction.

Bed friction is also generally represented in the form of a quadratic friction law, as given by:

$$\tau_{xb} = \rho g q_x \frac{V_s}{C^2 H} \tag{13}$$

where $V_s =$ depth-average fluid speed $(= \sqrt{U^2 + V^2})$, and $H =$ total depth of water.

To determine the Chezy value for the bed roughness, most of the widely used models tend to use the Manning formula, which expresses C in terms of the local depth as follows:

$$C = \frac{H^{1/6}}{n} \tag{14}$$

where $n =$ Manning roughness coefficient with typical values of n being in the range from 0.012 for smooth lined channelized rivers to 0.04 or more for meandering rivers with vegetation and so on. Although the Manning coefficient is primarily measured in practice for 1-D river reaches, this parameter has been widely used in 2-D flow fields with high levels of accuracy often being obtained for relatively complex flow fields.

Although this approach is appropriate for most rivers and estuaries, the Manning representation assumes that the

flow is rough turbulent flow and that the local head-loss is dependent only on the size and characteristics of the bed roughness, that is, form drag dominates. However, for low velocity flows on shallow floodplains and wetlands, Reynolds number effects may be significant, reflecting the increased influence of skin friction. This complex hydrodynamic phenomenon can be represented using the more comprehensive friction formulation given by the Colebrook–White equation (Henderson, 1966), given as:

$$C = -17.715 \log_{10}\left[\frac{k_s}{12H} + \frac{0.282C}{R_e}\right] \tag{15}$$

where k_s = Nikuradse equivalent sand grain roughness and R_e = Reynolds number for estuaries $= 4V_s/\nu$, where ν = kinematic laminar viscosity. The other advantage in using the Colebrook–White formulation to represent the bed roughness, rather than the Manning formulation, is that the physical roughness parameter k_s can be directly related to the height of bed features, such as ripples or dunes, rather than being based on a descriptive representation of the bed characteristics as for the Manning formulation.

Finally, for the depth-averaged eddy viscosity $\bar{\nu}_t$, this parameter can preferably be estimated from field data of the vertical velocity profile or assuming bed-generated turbulence dominates over free shear layer turbulence, then a logarithmic velocity profile can be assumed giving (Elder, 1959):

$$\bar{\nu}_t = 0.167\kappa U_* H \tag{16}$$

where U_* = shear velocity ($\sqrt{g}V_s/C$). However, field data by Fischer (1973) showed that the turbulent diffusion coefficient in straight, fairly uniform rivers is generally much higher and is more accurately represented by:

$$\bar{\nu}_t = 0.15U_* H \tag{17}$$

For most practical estuaries, even this value can be low compared with the measured data recorded in rivers, with values for $\bar{\nu}_t/U_*H$ typically ranging from 0.42 to 1.61 (Fischer *et al.*, 1979). Recent results of using data-mining techniques for deriving formulations for variable (viz. water depth dependent) roughness coefficients are described by Babovic (2005) (*see* **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1**) who used Genetic Programming techniques to obtain a formulation of the variable roughness that can be elegantly incorporated in the numerical formulation, as demonstrated by Stelling and Verwey (2005) (*see* **Chapter 16, Numerical Flood Simulation, Volume 1**).

One-dimensional Flows

In simulating flows within rivers and narrow estuaries, one-dimensional models may be used where the longitudinal flow features dominate the system. The governing St. Venant equations of motion are obtained by integrating the 3-D equations of motion (i.e. equations (1) and (2)) over the area of flow to give:

$$T\frac{\partial \zeta}{\partial t} + \frac{\partial Q}{\partial x} = q \tag{18}$$

$$\frac{\partial Q}{\partial t} + \beta\frac{\partial (Q^2/A)}{\partial x} = -gA\frac{\partial \zeta}{\partial x} - gAS_f \tag{19}$$

where T = surface width of flow, Q = discharge, q = lateral inflow or outflow per unit length of river, A = area of flow, and S_f = friction slope, or slope of total energy line, which is given by the following representation:

$$S_f = \frac{Q|Q|}{C^2A^2R} \tag{20}$$

where R = hydraulic radius for conveyance segment and C (i.e. the Chezy coefficient) may be estimated from either equations (14) or (15). More details can be found in the accompanying article by Stelling and Verwey (2005) (*see* **Chapter 16, Numerical Flood Simulation, Volume 1**).

WATER QUALITY MODELING

Three-dimensional Flow Fields

In modeling numerically the flux of water quality indicators, sediments, or contaminants within a river or estuary, the conservation of mass equation can first be written in general terms for a three-dimensional flow field as given by (Harleman, 1966):

$$\underbrace{\frac{\partial \varphi}{\partial t}}_{1} + \underbrace{\frac{\partial \varphi u}{\partial x} + \frac{\partial \varphi v}{\partial y} + \frac{\partial \varphi w}{\partial z}}_{2} + \underbrace{\frac{\partial}{\partial x}\overline{u'\varphi'} + \frac{\partial}{\partial y}\overline{v'\varphi'} + \frac{\partial}{\partial z}\overline{w'\varphi'}}_{3} = \underbrace{\varphi_S + \varphi_d + \varphi_k}_{4} \tag{21}$$

where φ = solute concentration, φ_S = source or sink solute input (e.g. an outfall), φ_d = solute decay or growth term, and φ_k = kinetic transformation rate for solute. The individual terms in equation (21), generally referred to as *the advection-diffusion equation*, refer to: local effects (term 1), transport by advection (2), turbulence effects (3), and source (or sink), decay (or growth), and kinetic transformation effects (4).

The cross-produced terms $\overline{u'\varphi'}$ and so on represent the mass flux of solute due to the turbulent fluctuations and, by analogy with Fick's law of diffusion, it can be assumed that

this flux is proportional to the mean concentration gradient and is in the direction of decreasing concentration. Hence, the third terms can be written as:

$$\left.\begin{aligned}\overline{u'\varphi'} &= -D_{tx}\frac{\partial\varphi}{\partial x}\\ \overline{v'\varphi'} &= -D_{ty}\frac{\partial\varphi}{\partial y}\\ \overline{w'\varphi'} &= -D_{tz}\frac{\partial\varphi}{\partial z}\end{aligned}\right\} \tag{22}$$

where D_{tx}, D_{ty}, D_{tz} = turbulent diffusion coefficients in x, y, z directions. These coefficients are often associated with the eddy viscosity ν_t and a Schmidt number, with its value found to vary between 0.5 and 1.0 (Lin and Shiono, 1995). For estuarine flows it is common to assume isotropic turbulence and to approximate the horizontal diffusion terms to the depth mean coefficients as given by Fischer (1973), whereby, in the absence of field data, these terms are often equated to:

$$D_{tx} = D_{ty} = 0.15U_*H \tag{23}$$

Likewise, for the vertical diffusion coefficient, in the absence of stratification and field data, it is common to assume a linear shear stress distribution and a logarithmic velocity profile giving (Vieira, 1993):

$$D_{tz} = U_*\kappa z\left(1 - \frac{z}{H}\right) \tag{24}$$

However, as indicated for hydrodynamic modeling, 3-D models only tend to be used for simulating water quality processes in which the vertical variations of a water quality indicator, such as density, are significant.

Two-dimensional Flow Fields

For practical problems where there are significant variations in the velocity field across the river or estuary, such as over floodplains or mangrove forests, a 2-D numerical model solution is more commonly used, together with the depth-integrated advective-diffusion equation (Falconer, 1991) given as:

$$\begin{aligned}\frac{\partial\phi H}{\partial t} + \frac{\partial\phi q_x}{\partial x} + \frac{\partial\phi q_y}{\partial y} - \frac{\partial}{\partial x}\left[HD_{xx}\frac{\partial\phi}{\partial x} + HD_{xy}\frac{\partial\phi}{\partial y}\right]\\ - \frac{\partial}{\partial y}\left[HD_{yx}\frac{\partial\phi}{\partial x} + HD_{yy}\frac{\partial\phi}{\partial y}\right]\\ = H[\phi_s + \phi_d + \phi_k]\end{aligned} \tag{25}$$

where ϕ = depth-average solute concentration, k = decay rate constant and so on, ϕ_s, ϕ_d, ϕ_k are the depth-average concentrates corresponding to φ_s, φ_d, φ_k, in equation (21) and D_{xx}, D_{xy}, D_{yx}, D_{yy} are the depth-average longitudinal dispersion and turbulent diffusion coefficients in x, y directions.

For the dispersion-diffusion terms, these coefficients can be shown to be of the following form (Preston, 1985):

$$\left.\begin{aligned}D_{xx} &= \frac{(D_\ell U^2 + D_t V^2)H\sqrt{g}}{V_s C} + D_w\\ D_{yy} &= \frac{(D_\ell V^2 + D_t U^2)H\sqrt{g}}{V_s C} + D_w\\ D_{xy} = D_{yx} &= \frac{(D_\ell - D_t)UVH\sqrt{g}}{V_s C} + D_w\end{aligned}\right\} \tag{26}$$

where D_ℓ = depth-average longitudinal dispersion constant, D_t = depth-average turbulent diffusion constant, and D_w = wind-induced dispersion coefficient. For values of D_ℓ and D_t, these dimensionless constants can be preferably obtained from field data, or alternatively minimum values can be obtained by assuming a logarithmic velocity profile, wherein $D_\ell = 5.93$ (Elder, 1959) and $D_t = 0.15$ (Fischer, 1973). However, in practical studies these values tend to be rather low (Fischer *et al.*, 1979), with measured values for D_ℓ and D_t ranging from 8.6 to 7500 and 0.42 to 1.61 respectively. For present-day capabilities of data-driven modeling and data-mining techniques, reference is made to, for example, Solomatine (2005) (*see* **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**), Babovic (2005) (*see* **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1**), and Savic and Khu (2005) (**Chapter 22, Evolutionary Computing in Hydrological Sciences, Volume 1**).

One-dimensional Flow Fields

As for hydrodynamic studies, water quality studies of rivers and narrow well-mixed estuaries are often based on a 1-D numerical model. For this purpose the 3-D advective-diffusion equation (21) is integrated over an arbitrary cross-sectional area A to give:

$$\frac{\partial\phi A}{\partial t} + \frac{\partial\phi Q}{\partial x} - \frac{\partial}{\partial x}\left[AD_l\frac{\partial\phi}{\partial x}\right] = S_s + S_d + qS_L \tag{27}$$

where ϕ = area-average solute concentration, S_s = area-average source term, S_d = area-average decay term and S_L = solute concentration of lateral input (or output) (see also equation (18)).

WATER QUALITY PROCESSES

In modeling water quality processes in rivers and estuaries, a range of water quality parameters are often modeled, including physical, chemical, and biological indicators

(Falconer *et al.*, 2001). General discussions on water quality processes can be found in Part 8 of this Encyclopedia: Water Quality & Biogeochemistry (McCutcheon, 2005) (*see* **Chapter 100, Water Quality Modeling, Volume 3**). In this article, details are given of the procedures used for modeling fecal coliform (FC) and heavy metals.

Fecal Coliform Modeling

Total coliform (TC) has been used for many years as the main indicator in evaluating bathing-water quality with respect to domestic waste. However, because of the difficulties associated with the occurrence of non-fecal bacteria in tests, the use of the TC test is being gradually replaced by FC and fecal streptococci as the main bacteriological indicators (Thomann and Mueller, 1987). The FC bacteria group are indicative of organisms from the intestinal tract of humans and other animals. In recent years, FC has been used by several researchers to assess the quality of bathing water and urban streams for both outfall and/or nonoutfall sources (Wyer *et al.*, 1997; Thackston and Murr, 1999; Young and Thackston, 1999).

In modeling FC, the decay terms in equations (25) and (27) are generally expressed as a first-order decay function and are typically included in 1-D and 2-D models respectively via the following equations:

$$S_d = -k_B \phi A \quad \text{and} \quad \phi_d = -k_B \phi \tag{28}$$

where k_B = coliform decay rate (day^{-1}). The response time, that is, the time required for the water body (i.e. river or estuary etc.) to complete a fixed percentage of its recovery, is also commonly used to represent the growth/decay of bacteria. This parameter can be easily derived from a first-order decay formulation of the form $\phi = \phi_0 e^{-kt}$, where ϕ_0 = initial concentration at $t = 0$. In practice, T_{90} is a commonly used parameter for this purpose and is defined as the time during which the original organism population would reduce by 90%. It can readily be shown that T_{90} can be calculated from (Chapra, 1997):

$$T_{90} = \frac{2.303}{k_B} \tag{29}$$

Several factors may influence the population of the organisms in a water body, and thus, in reporting the decay rate, sampling conditions are usually specified. These factors are mainly sunlight intensity and duration, temperature and salinity levels, suspended particulate matter, and concentrations of toxic substances. A wide range of decay rates for TCs and FCs and fecal streptococci have been reported in the literature. For example, TC and FC decay rate values have been reported from 0 to 2.4 day^{-1} for 2–18% salinity and dark conditions and 2.5 to 6.1 day^{-1} for 15% salinity and sunlight conditions. Likewise, a range of 37 to 110 day^{-1} has also been reported for FC decay rates in seawater under good sunlight conditions (see Thomann and Mueller, 1987, Table 9).

Sediment Transport Modeling

On the basis of the general advective-diffusion equation (21), the governing equation for suspended sediment transport processes is generally written as:

$$\frac{\partial S}{\partial t} + \frac{\partial}{\partial x}(uS) + \frac{\partial}{\partial y}(vS) + \frac{\partial}{\partial z}[(w - w_s)S] - \frac{\partial}{\partial x}\left(D_{tx}\frac{\partial S}{\partial x}\right) - \frac{\partial}{\partial y}\left(D_{ty}\frac{\partial S}{\partial y}\right) - \frac{\partial}{\partial z}\left(D_{tz}\frac{\partial S}{\partial z}\right) = S_T \tag{30}$$

where S = suspended sediment concentration and S_T = source or sink term. It should be noted that the source or sink term is introduced through the bed boundary conditions. For cohesive sediments, the following bed conditions have been used (Wu *et al.*, 1999):

$$-w_s S - D_{tz}\frac{\partial S}{\partial z} = q_{dep} \quad \text{when } \tau_b \le \tau_d \quad \text{(deposition)} \tag{31a}$$

$$-w_s S - D_{tz}\frac{\partial S}{\partial z} = q_{ero} \quad \text{when } \tau_b \ge \tau_e \quad \text{(erosion)} \tag{31b}$$

$$-w_s S - D_{tz}\frac{\partial S}{\partial z} = 0 \quad \text{when } \tau_d < \tau_b < \tau_e \quad \text{(equilibrium)} \tag{31c}$$

where τ_b = bed shear stress, τ_d = critical shear stress beyond which no further deposition occurs, τ_e = critical shear stress for erosion, and q_{dep}, q_{ero} are the deposition and erosion rates respectively at the bed.

Heavy Metal Modeling

As heavy metals can exist in both dissolved and adsorbed particulate phases in rivers and estuaries, the distribution between these two phases is usually described by a partitioning coefficient K_D. For transport of metals in the dissolved phase, the three-dimensional advective-diffusion equation is given as:

$$\frac{\partial C}{\partial t} + \frac{\partial}{\partial x}(uC) + \frac{\partial}{\partial y}(vC) + \frac{\partial}{\partial z}(wC) - \frac{\partial}{\partial x}\left[D_{tx}\frac{\partial C}{\partial x}\right] - \frac{\partial}{\partial y}\left[D_{ty}\frac{\partial C}{\partial y}\right] - \frac{\partial}{\partial z}\left[D_{tz}\frac{\partial C}{\partial z}\right] = (C_d + C_t) \tag{32}$$

where C = concentration of heavy metals dissolved in water column, C_d = source or sink of dissolved heavy metal, and C_t = transformation flux from, or to, adsorbed particulate phase onto the sediments. The adsorbed

particulate phase is transported with the sediments and this process may be described by the following equation:

$$\frac{\partial SP}{\partial t} + \frac{\partial}{\partial x}(uSP) + \frac{\partial}{\partial y}(vSP) + \frac{\partial}{\partial z}[(w - w_s)SP] - \frac{\partial}{\partial x}\left[D_{tx}\frac{\partial SP}{\partial x}\right] - \frac{\partial}{\partial y}\left[D_{ty}\frac{\partial SP}{\partial y}\right] - \frac{\partial}{\partial z}\left[D_{tz}\frac{\partial SP}{\partial z}\right] = (SP_d + SP_t) \quad (33)$$

where SP = concentration of heavy metal adsorbed on the suspended sediments, w_s = apparent sediment settling velocity, SP_d = source or sink of adsorbed particulate heavy metal, SP_t = transformation flux from, or to, dissolved phase in water column. The transformation processes between the dissolved and adsorbed particulate phases are very complex.

Millward and Turner (1995) have established that the partitioning of heavy metals between the particulate and dissolved phase is dependent upon salinity. They established from field data an empirical relationship linking the partitioning coefficient with salinity of the following form:

$$\log_e(K_D) = b\log_e(S + 1) + \log_e(K_D^0) \quad (34)$$

where K_D = partitioning coefficient, S = salinity and K_D^0 = partition coefficient in freshwater.

MODELING APPLICATIONS

In this section, details are given of two modeling studies undertaken by the authors in applying the above governing equations for predicting the hydrodynamic, water quality, and sediment transport processes in river and estuarine waters in two UK basins. Some other recently published studies that are related to the modeling of riverine and estuarine processes are also highlighted.

Modeling Bathing Water Quality in the Ribble Estuary

A numerical model study was undertaken to establish the water quality of the EU-designated bathing waters located near the mouth of the Ribble Estuary, located along the northwest coast of England (Kashefipour *et al.*, 2002). The numerical model domain included three tidal rivers, namely, the Ribble, Darwen and Douglas rivers, and the Ribble Estuary, with the upstream boundaries being at the tidal limits of these three rivers and with the downstream boundary being located around the 25-m depth contour in the Irish Sea (see Figure 1). The length of the seaward boundary was 41.2 km, with the width of the river boundaries being generally less than 10 m. Thus, in this study a 1-D model was dynamically linked to a 2-D model to create a single model in which the computations for the hydrodynamic and water quality variables were undertaken simultaneously across the entire model domain.

Six sets of hydrodynamic and water quality data were collected during the winter of 1998 and summer of 1999, thereby including different weather and tide conditions.

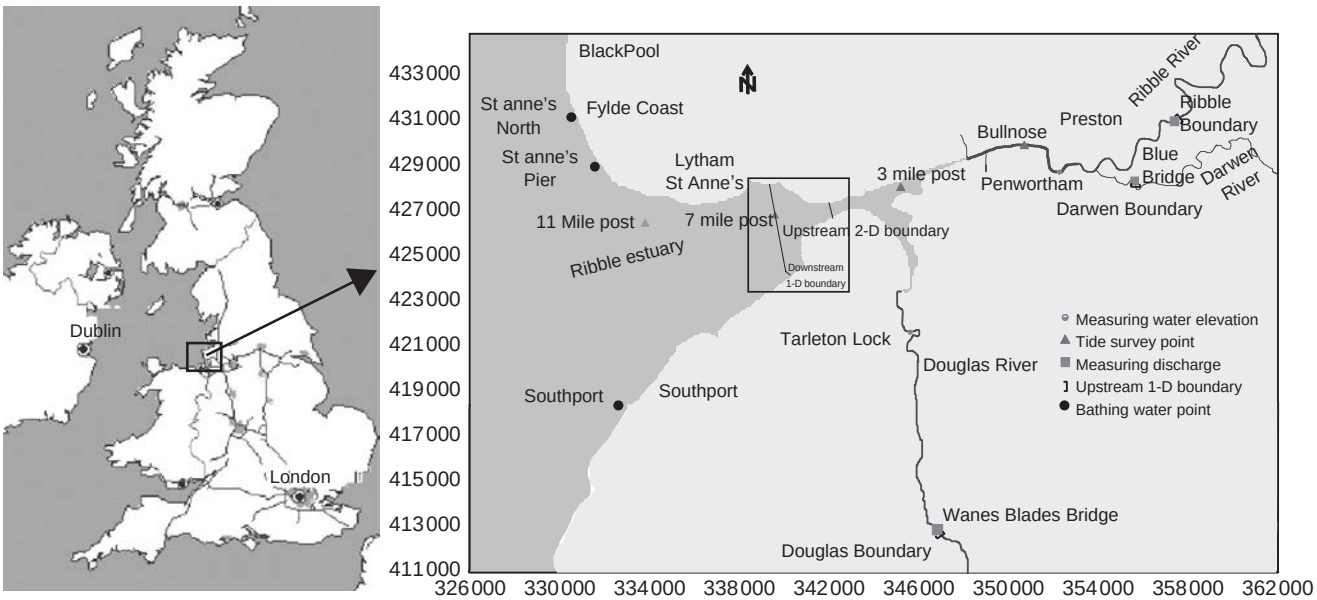


Figure 1 Fylde Coast, Ribble Estuary and its tributaries. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Comprehensive data were collected, including water depths, current speeds and directions, salinity levels, and concentrations of suspended solids, fecal and total coliforms, and fecal streptococci.

Fecal indicator organisms in river and estuarine waters could derive from human and/or animal sources. The former could be effluents discharged to rivers and estuaries and spills from combined sewer overflows (CSOs). The latter could be diffuse sources in riverine catchments related to agriculture activities (Loague, 2005) (*see* **Chapter 94, Point and NonPoint Source Pollution, Volume 3**). Field mapping of land use was undertaken over the catchments using a classification scheme (Fewtrell *et al.*, 1998) that divided the land use into 25 classes, including improved pasture, rough grazing, arable, woodland and built-up land and so on. A stepwise multiple regression analysis was used to model relationships between geometric mean indicator concentrations and land use types for base and high flows.

As an example, the results obtained for the survey on 3 June 1999 are discussed herein. This survey was carried out during a wet period for a mean tidal range. Figures 2 (a–c) show comparisons at a calibration point, that is, 7 milepost (see Figure 1), between predicted water elevations and current speeds and directions respectively. Figure 2 (a) shows that good agreement was achieved between the predicted and measured water levels at this site, with an average error of 0.11 m when compared with the measured tidal range. Figures 2 (b, c) illustrate relatively good agreement between the predicted current speeds and directions with the corresponding measured data, with an average error of 0.13 m s^{-1} and 8.1° respectively.

In this study, FC was used as the main water quality indicator organism. To reflect all of the environmental conditions affecting the fate of microorganisms, different decay rates were used for (i) day and night times (ii) dry and wet weather conditions, and (iii) sea and river waters. FC inputs during the 3 June 1999 survey along the rivers and estuary were categorized into four groups, including open boundaries, wastewater treatment works (WwTWs), CSOs, and the other inputs. As can be seen from Figure 3 the most significant FC load came from the river inputs. Statistical analysis showed similar results for the other surveys, both for wet and dry events.

Predicted and measured FC concentrations at 7 milepost for the survey on 3 June 1999 are compared in Figure 4. Relatively good agreement between both sets of data can be seen from this figure, with an average error of 30.5%. Figure 5 shows another example for the linked model application to the Ribble Estuary and compares the predicted and measured FC concentrations at 3 milepost for the 19 May 1999 survey, which was carried out for a dry event and a spring tidal range. As can be seen from this figure, both sets of data agreed well, with an average error of 27.6%.

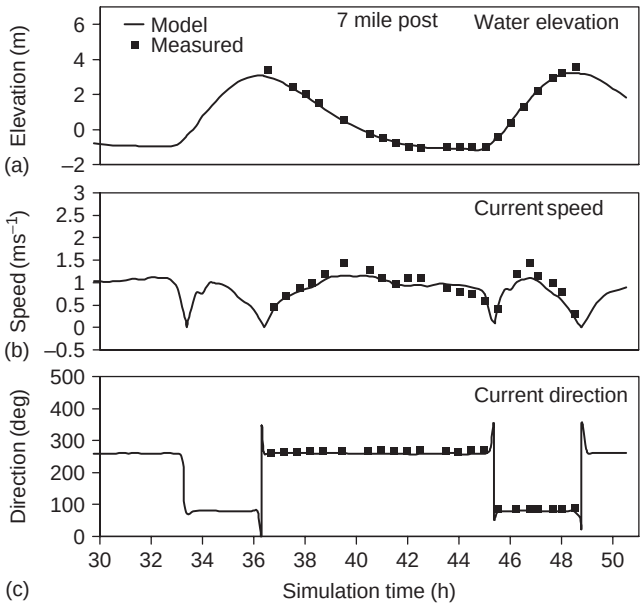


Figure 2 Comparison of predicted and measured: (a) water levels, (b) current speeds, and (c) current directions at 7 milepost, on 3 June, 1999

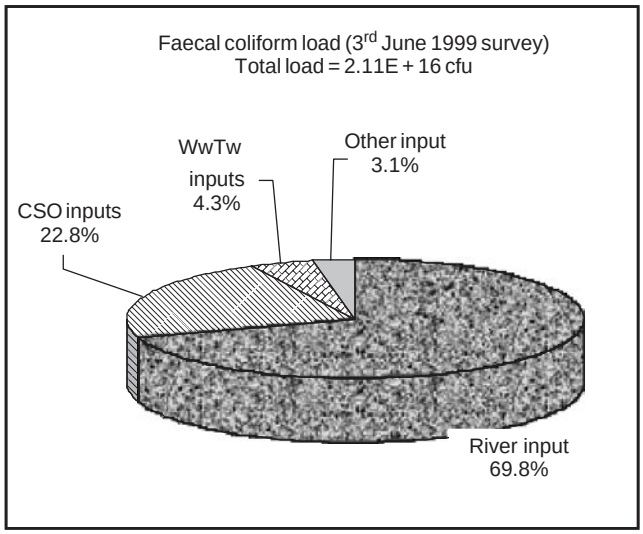


Figure 3 Fecal coliform inputs to Ribble Estuary on 3 June, 1999

Model Application to Humber Estuary

The Humber Basin is a well-mixed estuarine system located along the east coast of northern England and providing an outlet to the North Sea for the rivers Trent and Ouse and shipping access to a number of ports, including Hull, Immingham, and Grimsby (see Figure 6). The management and monitoring of water quality in the Humber Basin is coordinated by the UK Environment Agency (Edwards *et al.*, 1987). Field measurements of water elevations,

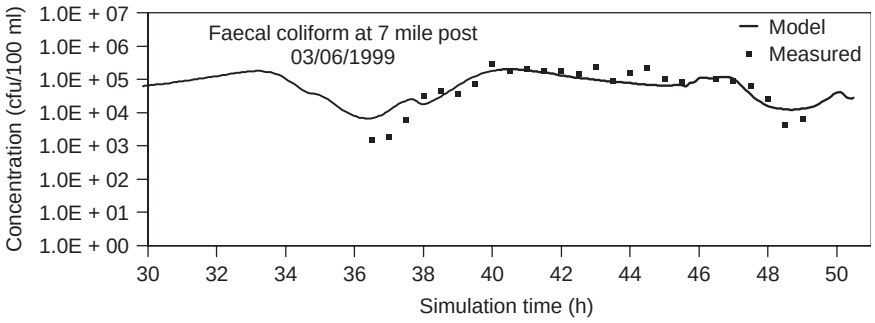


Figure 4 Comparison of predicted and measured fecal coliform concentrations at 7 milepost, on 3 June, 1999

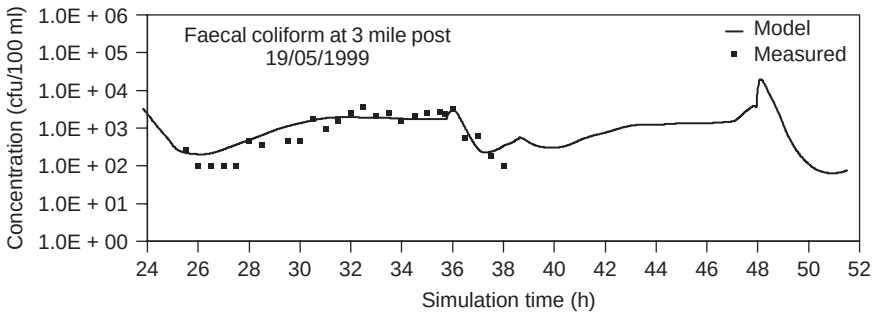


Figure 5 Comparison of predicted and measured fecal coliform concentrations at 3 milepost, on 19 May 1999

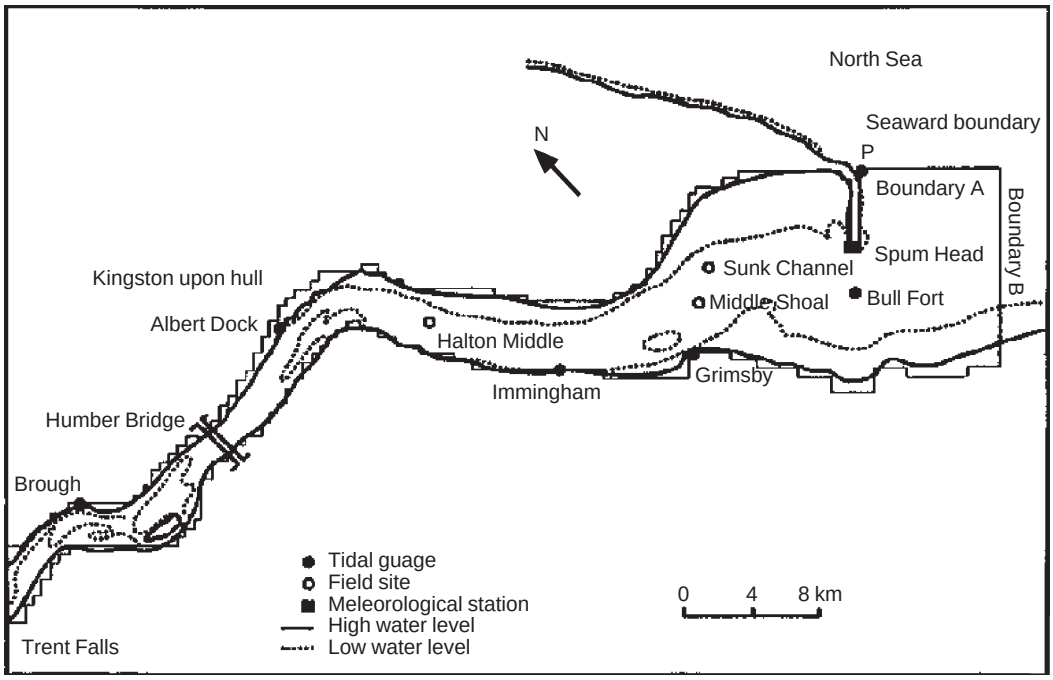


Figure 6 Plan of Humber Estuary

velocities, and various water quality constituents were also available at several sites in between the two open boundaries. Data were available for different tidal ranges at each site, with a complete tidal cycle being monitored at quarter hour intervals. For the water quality constituent inputs from the rivers Wharfe, Aire, Don, Trent, and Ouse, the boundary values used were obtained by summing up the individual components. For the loadings from the various

sewage treatment works at Hull and others, corresponding values were obtained from Yorkshire Water plc.

In general, it is always necessary to be able to check the water levels and velocity field predictions obtained from a numerical model with field data before proceeding with any analysis of the predictive results obtained, with the same being true for a laboratory model investigation. For the model studies undertaken for this basin, comparisons were first made using 3-D and 2-D models of the water level, and velocity predictions and measurements were taken at several sites along the estuary. For the 3-D model predictions, a typical set of results are shown in Figure 7, for comparison, at Halton Middle. As can be seen, the corresponding typical comparison shows good agreement between both sets of results. Predictions were made using the water quality formulations cited previously, and comparisons between measured and predicted results along the estuary are shown for salinity and dissolved oxygen in Figures 8 and 9 respectively. As can be seen from these results, the agreement between the measured and predicted concentration levels was encouraging for all of the individual constituents considered.

Finally, the sediment transport and water quality models were extended to predict trace metal contaminant levels, including the sorption and desorption between dissolved and particulate forms in the water column and the sediments. The 2-D model was applied to simulate the distribution, behavior, and characteristics of the dissolved and particulate metal phases in the Humber for different partition coefficients. These scenarios included (i) no partitioning of contaminants between the dissolved and adsorbed phases, (ii) partitioning of contaminants with no salinity dependence, and (iii) partitioning with salinity dependence.

A series of numerical experiments were undertaken to study the impact of sediment transport on the fate of heavy metals. Figure 10 shows a typical dissolved metal concentration distribution predicted by the numerical model, 496 h after an outfall spill (Gunapala, 2002). The period of spill was assumed to be 12.4 h, with the discharge, and metal concentrations being $2.75\,m^3\ s^{-1}$ and $1.32\,\mu g\,l^{-1}$,

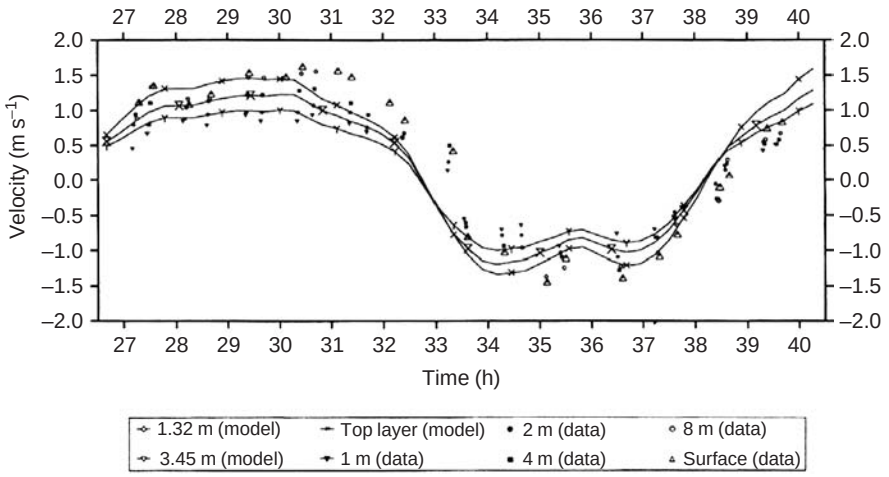


Figure 7 Comparison of predicted and measured velocities for a spring tide at Halton Middle

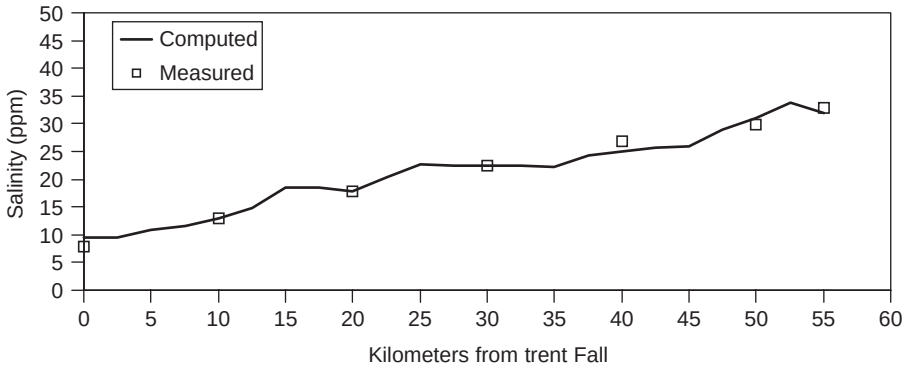


Figure 8 Comparison of predicted and measured salinity levels along the Basin

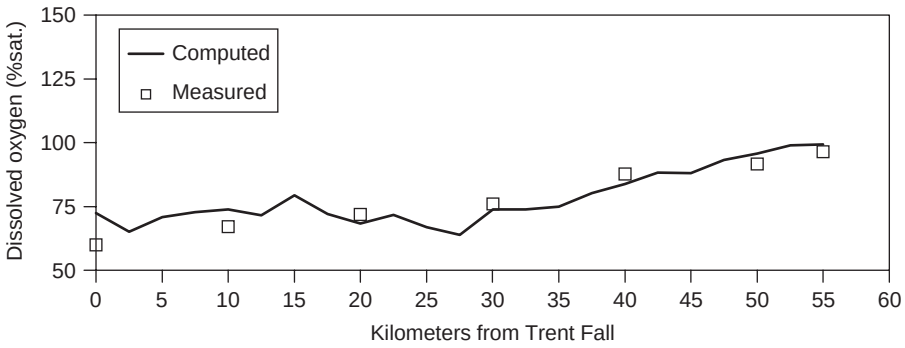


Figure 9 Comparison of predicted and measured dissolved oxygen levels along the Basin

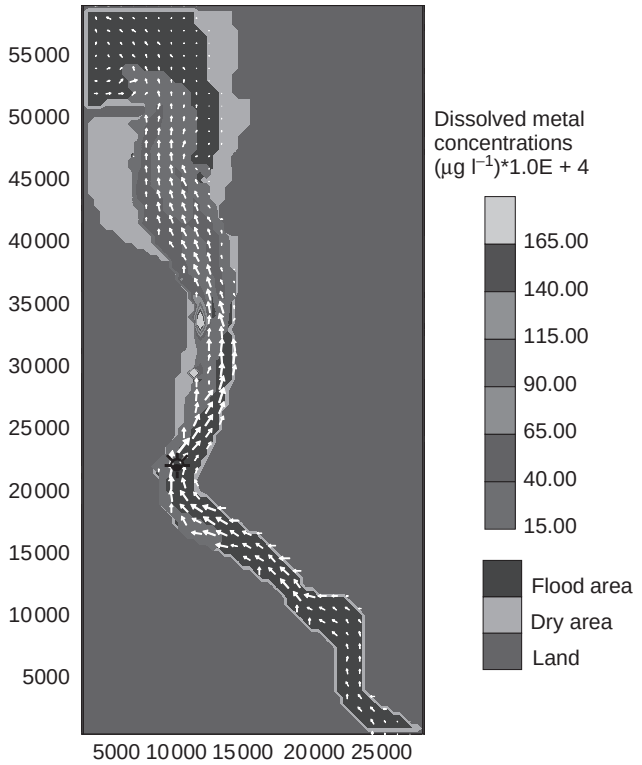


Figure 10 Dissolved metal concentrations along the Humber estuary at a $K_D = 6600$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

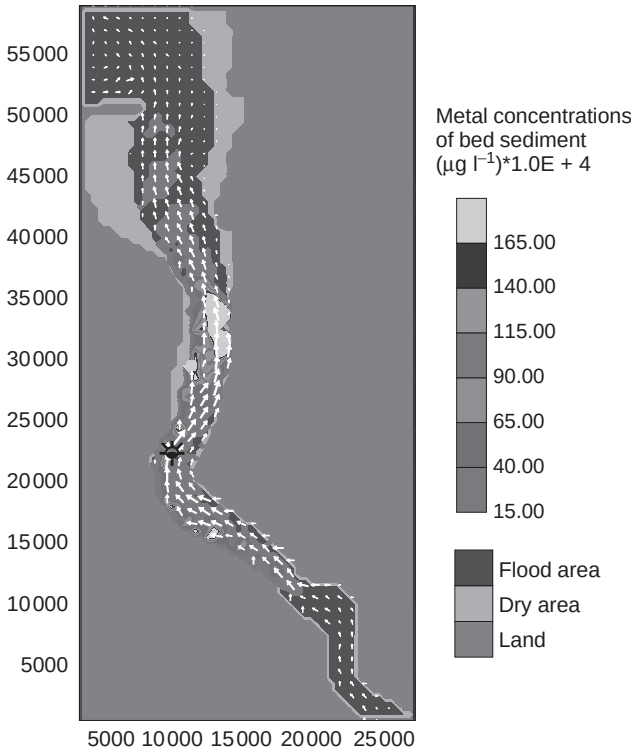


Figure 11 Metal concentrations of bed sediment along the Humber estuary at a $K_D = 6600$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

respectively. As the sediment suspends and deposits from the bed in a cyclical manner with tidal movement, suspension and deposition occurs at high and low velocities respectively. However, because the residual velocity is towards the lower boundary, this suspension, deposition, and resuspension of the sediments results in slowly transporting the metals seawards. From Figure 11 it can be seen that heavy metals adsorbed onto the bed sediments have moved seawards, but at a much slower pace when compared with the dissolved metals.

Discussion

Details are given of two modeling studies to illustrate how integrated hydroinformatics tools can be used in the assessment of river and estuarine water quality. However, it is worthwhile noting that water quality models are now widely used in river and estuarine environmental impact assessment studies, and, in particular, three-dimensional numerical models are increasingly being refined for this purpose.

In a separate, but related, modeling study of the Humber Estuary, Naden *et al.* (2002) used three types of models to

estimate these input fluxes, including (i) a diffuse source model, (ii) a point source model, and (iii) an in-stream model. A catchment delivery model was used to predict the diffuse input of water, sediment, and selected nutrients and other contaminants. The delivery model was spatially distributed to represent the regional variation in contaminant loss. The simulated fluxes from diffuse sources were delivered into a one-dimensional river model simulating transfer and transformation processes in the river network.

Huang and Spaulding (1995) developed a three-dimensional model to predict tidal circulation and mixing processes induced by surface discharges in estuarine and coastal waters. The model consists of a hydrodynamic, pollutant water quality and turbulence modules. Model simulations were compared with analytical solutions and laboratory experiments, and good comparisons were obtained for tidal, wind, and density forcing, as well as salinity intrusion.

In recent years, more effort has also been focused on developing full 3-D hydrodynamic models that do not use the assumption of a hydrostatic pressure distribution in the vertical direction (Casulli and Stelling, 1996). These types of models have been shown to be able to simulate flows with large vertical accelerations and are particularly useful for density stratified estuarine flows.

A 3-D numerical model developed by Chau and Jiang (2002) was used to predict the hydrodynamic and pollutant transport processes in the Pearl River Estuary, China. The model is based on an orthogonal coordinate system in the horizontal direction and a sigma coordinate system in the vertical direction. The model was used to predict salinity and Chemical Oxygen Demand (COD) concentrations along the estuary.

For predicting heavy metal transport processes in riverine systems, Ji *et al.* (2002) applied coupled hydrodynamic, sediment transport, and contaminant transport models to the Blackstone River in the United States. The model was used to predict five types of metals during storm events to clarify the impact of contaminant sources and sediment resuspension processes on the river water quality.

CONCLUSIONS

As a result of the rapid advances in both computer power and numerical solution techniques, numerical models are increasingly used to predict hydrodynamic, hydrological, and water quality processes in river and estuarine waters. Details are given in this article of the governing hydrodynamic equations for flow-field predictions and the corresponding advective-diffusion equation for the transport of water quality indicators, suspend sediments, and contaminant solutes. Emphasis has been focused on the equations that are widely used for unsteady river and estuarine flow conditions. In applying the advective-diffusion equation to solute transport processes, details are given of the kinetic reactions and decay rates for FC. For simulating heavy metal fluxes, details are given of the procedures for modeling both sediment erosion and deposition, and metal adsorption and desorption processes.

In detailing the modeling of FC concentrations for the Ribble Estuary, UK, it has been shown that dynamically linked 1-D and 2-D models can provide accurate predictions of the transport processes of FCs along complex river basins. This conclusion is consistent with that obtained by Stelling and Verwey (2005) (*see* **Chapter 16, Numerical Flood Simulation, Volume 1**) on numerical flood modeling. It has also been shown that field surveys and hydrological models are both important for accurately specifying bacterial loads from boundaries, outfalls, and diffuse sources associated with the manner in which the catchment land is being used. In modeling the concentration distributions of water quality indicators and sediments in the Humber Estuary, UK, it has been shown that the model predictions generally agree well with field-measured data. In modeling the fate of heavy metals, it has been shown that the tidal environment, because of the adsorption and desorption of metals onto the sediments, and the erosion and deposition processes play a critical role in the retention time of contaminants within the estuary. Finally, some related 3-D modeling studies were briefly described to highlight that 3-D models are increasingly applied to predict river and estuarine water quality characteristics.

Acknowledgments

The research studies reported herein were mainly funded by the Natural Environment Research Council and the Engineering and Physical Sciences Research Council. The authors are grateful to North West Water Ltd (NWW) and the Environmental Agency for funding one of the projects and for the provision of data for the Ribble basin. They are also grateful to Dr Seyed Kashefipour and Ms Emma Harris and Professor David Kay and Dr Carl Stapleton (University of Wales, Aberystwyth) for their contribution to this study.

FURTHER READING

Guinot V. (2005) Advances in numerical modelling capabilities in hydrology. *Encyclopaedia of Hydrological Sciences*, Part 2, Hydroinformatics, John Wiley & Sons.

Krone R. B. (1962) *Flume Studies of the Transport of Sediment in Estuarial Processes*, Final Report, Hydraulic Engineering Laboratory and Sanitary Engineering Research Laboratory, University of California: Berkeley.

Turner A., Millward G. E., Bale A. J. and Morris A. W. (1993) Application of the K_D concept to the study of trace metal removal and desorption during estuarine mixing. *Estuarine, Coastal and Shelf Science*, **36**, 1–13.

van Rijn L. C. (1984a) Sediment transport part 1: bed load transport. *Journal of Hydraulic Engineering ASCE*, **10**, 1431–1456.

van Rijn L. C. (1984b) Sediment transport part 2: suspended load transport. *Journal of Hydraulic Engineering*, **10**, 1613–1641.

REFERENCES

Abbott M.B., Mynett A.E. and O'Kane P. (2005) Computer-based modelling paradigms in hydrology. *Encyclopaedia of Hydrological Sciences*, Part 2, Hydroinformatics, John Wiley & Sons.

Babovic V. (2005) Genetic programming of hydrological processes. *Encyclopaedia of Hydrological Sciences*, Part 2, Hydroinformatics, John Wiley & Sons.

Casulli V. and Stelling G. S. (1996) Simulation of three-dimensional, non-hydrostatic free-surface flows for estuaries and coastal waters. In *Proceedings of the 4th International Conference on Estuarine and coastal Modelling*, Spaulding M. L. and Cheng R. T. (Eds.), ASCE: pp. 1–12.

Chapra S.C. (1997) *Surface Water Quality Modelling*, McGraw-Hill Companies, p. 844.

Chau K.W. and Jiang Y.W. (2002) Three-dimensional pollutant transport model for the Pearl river Estuary. *Water Research*, Vol. 26, IWA, pp. 2029–2039.

Dyer K.R. (1997) *Estuaries: A Physical Introduction*, *Second Edition*, John Wiley & Sons: Chichester.

Edwards A., Freestone R. and Urquhart C. (1987) *The Water Quality of the Humber Estuary*, Report of the Humber Estuary Committee, Yorkshire Water Authority, p. 96.

Elder J.W. (1959) The dispersion of marked fluid in a turbulent shear flow. *Journal of Fluid Mechanics*, **5**(4), 544–560.

Falconer R.A. (1991) Review of modelling flow and pollutant transport processes in hydraulic basins. *Proceedings of First International Conference on Water Pollution: Modelling, Measuring and Prediction*, Computational Mechanics Publications: Southampton, pp. 3–23.

Falconer R.A. (1993) An introduction to nearly horizontal flows. In *Coastal, Estuarial and Harbour Engineers' Reference Book*, Chap. 2, Abbott M.B. and Price W.A. (Eds.), E.&F.N. Spon: London, pp. 27–36.

Falconer R.A. and Chen Y. (1996) Modelling sediment transport and water quality processes on tidal floodplains. In *Floodplain Processes*, Chap. 11, Anderson M.G., Walling D.E. and Bates P.D. (Eds.), John Wiley and Sons: Chichester, pp. 361–398.

Falconer R.A., Lin B. and Kashefipour S.M. (2001) Modelling water quality processes in riverine systems. In *Validation of Water Quality Model*, Chap. 14, Anderson M.G. and Bates P.D. (Eds.), John Wiley & Sons: pp. 358–387.

Fewtrell L., Kay D., Wyer M., Crowther J., Carbo P. and Mitchell G. (1998) *Faecal Indicator Budgets Discharging to the Ribble Estuary*, Report to the Environment Agency, North West Region, p. 52.

Fischer H.B. (1973) Longitudinal dispersion and turbulent mixing in open channel flow. *Annual Review of Fluid Mechanics*, **5**, 59–78.

Fischer H.B., List E.J., Koh R.C.J., Imberger J. and Brooks N.H. (1979) *Mixing in Inland and Coastal Waters*, Academic Press: San Diego, p. 483.

Goldstein S. (1938) *Modern Development in Fluid Dynamics*, Vol. 1, Oxford University Press: Oxford.

Gunapala, S.R. (2002) *Fate and Behaviour of Heavy Metals in the Humber Estuary*, PhD Thesis, Cardiff University, p. 243.

Harleman D.R.F. (1966) Diffusion processes in stratified flow. In *Estuary and Coastline Hydrodynamics*, Chap. 12, Ippen A.T. (Ed.), McGraw-Hill Book: New York, pp. 575–597.

Henderson F.M. (1966) *Open Channel Flow*, Collier-Macmillan Publishers: London, p. 522.

Huang W. and Spaulding M. (1995) 3D Model of estuarine circulation and water quality induced by surface discharges. *Journal of Hydraulic Engineering ASCE*, **121**, 300–311.

Ji Z., Hamrick J.H. and Pagenkopf J. (2002) Sediment and metals modeling in shallow river. *Journal of Environmental Engineering ASCE*, **128**, 105–119.

Kashefipour S.M., Lin B., Harris E.L. and Falconer R.A. (2002) Hydro-environmental modelling for bathing water compliance of an estuarine basin. *Water Research*, Vol. 36, IWA, pp. 1854–1868.

Lin B. and Shiono K. (1995) Numerical modelling of solute transport in compound channel flows. *Journal of Hydraulic Research IAHR*, **33**, 773–788.

Loague K. (2005) Point and non-point source pollution. *Encyclopaedia of Hydrological Sciences*, Part 8, Water Quality and Biogeochemistry, John Wiley & Sons.

McCutcheon S. (2005) Water quality modelling. *Encyclopaedia of Hydrological Sciences*, Part 8, Water Quality and Biogeochemistry, John Wiley & Sons.

Millward G.E. and Turner A. (1995) Trace metals in estuaries. In *Trace Metals in Natural Waters*, Salbu B. and Steinnes E. (Eds.), CRC Press: Baton Rouge, pp. 223–245.

Mynett A.E. (2002) Environmental hydroinformatics: the way ahead. *Proceedings of the Fifth International Conference on Hydroinformatics*, Cardiff, Vol. 1, IWA: London, pp. 31–36.

Naden P.S., Cooper D.M. and Boorman D.B. (2002) Modelling large-scale river basins. In *Land-Ocean Interaction*, Huntly D.A., Leeks G.J.L. and Walling D.E. (Eds.), IWA Publishing: pp. 105–142.

Preston R.W. (1985) *The Representation of Dispersion in Two-Dimensional Shallow Water Flow*. Central Electricity Research Laboratories, US, Report No. TPRD/U278333/N84, p. 13.

Rodi W. (1984) *Turbulence Models and their Application in Hydraulics*, *Second Edition*, International Association for Hydraulics Research: Delft, p. 104.

Savic D. and Khu S.T. (2005) Evolutionary computing in hydrological sciences. *Encyclopaedia of Hydrological Sciences*, Part 2, Hydroinformatics, John Wiley & Sons.

Solomatine D. (2005) Genetic algorithms and multi-criteria optimisation techniques. *Encyclopaedia of Hydrological Sciences*, Part 2, Hydroinformatics, John Wiley & Sons.

Stelling G. and Verwey A. (2005) Numerical flood simulation. *Encyclopaedia of Hydrological Sciences*, Part 2, Hydroinformatics, John Wiley & Sons.

Thackston E.L. and Murr A. (1999) CSO control model modifications based on water quality studies. *Journal of Environmental Engineering ASCE*, **125**, 979–987.

Thomann R.V. and Mueller J.A. (1987) *Principles of Surface Water Quality Modelling Control*, Harper Collins Publisher: New York, p. 644.

Vieira J.K. (1993) Dispersive processes in two-dimensional models. In *Coastal, Estuarial and Harbour Engineers' Reference Book*, Chap. 14, Abbott M.B. and Price W.A. (Eds.), E.&F.N. Spon: London, pp. 179–190.

Wallingford H.R. (1996) *Estuary Processes and Morphology Scoping Study*, Report SR 446, HR Wallingford Ltd., UK, p. 104.

Wu J. (1969) Wind-stress and surface roughness at air-sea interface. *Journal of Geophysical Research*, **74**, 444–455.

Wu Y., Falconer R.A. and Uncles R.J. (1999) Modelling of water flows and cohesive sediment fluxes in the Humber Estuary, UK. *Marine Pollution Bulletin*, **37**, 182–189.

Wyer M.D., Oneill G., Kay D., Crowther J., Jackson G. and Fewtrell L. (1997) Non-outfall sources of faecal indicator organisms affecting the compliance of coastal waters with directive 76/160/EEC. *Water Science and Technology*, Vol. 35, IWA, pp. 151–156.

Young K.D. and Thackston E.L. (1999) Housing density and bacterial loading in urban streams. *Journal of Environmental Engineering ASCE*, **125**, 1177–1180.

18: Shallow Water Models with Porosity for Urban Flood Modeling

VINCENT GUINOT

Hydrosciences Montpellier, Université Montpellier 2, Montpellier Cedex 5, France

The modeling of urban floods has to cope with highly variable geometries and flow parameters. Among the various types of models available, distributed, physically based models offer the best insight into the processes. However, the application of such models to refined urban flood modeling is not foreseeable in the near future. An alternative approach could be to use two-dimensional macroscopic models based on a modified version of the classical shallow–water equations, in which the fraction of the land surface occupied by the buildings is accounted for via a porosity. In such models, a cell covers part or totality of a house block. Such models provide a viable alternative to classical semidistributed models, as illustrated by a computational example on a small-sized urban area.

INTRODUCTION

With about 60% of the world total population expected to live in urban areas by the year 2030 (United Nations, 2002), urban floods may be seen as a very good illustration of the gradual shift "from nature-dominated to human-dominated" (Messerli *et al.*, 2000) environmental hazards. Hydraulic/hydrological modeling has proved its usefulness in the assessment of urban flood-related risk. However, it is hampered with a number of difficulties.

1. The meteorological forcing and its impact on urban runoff production (hence on the flood) has been observed to be highly variable in time and space (Ragab *et al.*, 2003). The typical temporal and spatial resolutions needed for efficient monitoring of urban meteorological forcing were identified (Schilling, 1991; Berne *et al.*, 2004) to be 1 to 5 min and 1 km^2 respectively. Radar-based rainfall measurement and its coupling with distributed hydrological models is currently being seen as a promising research path (Einfalt *et al.*, 2004).
2. Owing to the artificial nature of the urban geometry, the flow is highly variable in space and time and the role of inertial effects is nonnegligible, especially in the neighborhood of junctions (crossroads, roundabouts, etc.). The topographical features are identified as a major source of uncertainty in the modeling output of urban floods using distributed hydraulic models (Paquier and Mignot, 2003).
3. The variability of the flow is often related to variations in the hydraulic regime (such as transcritical flow or hydraulic jumps), which are not handled equally well by all existing modeling software packages. Besides the possible limited validity of the assumptions and equations attached to the various types of models, significant differences may be found between the outputs of models solving the same set of equations when different solution techniques are used, as shown by the results of the IMPACT benchmarking tests (see e.g. IMPACT, 2001).

Three main types of urban flood modeling approaches can be identified.

The first type of approach aims to determine a reduced set of specific flow variables (such as a peak discharge or a total flood hydrograph for a specifically delimited urban zone). The simulation results may then be interpreted in the light of the local topographical features in order to produce risk maps. Empirical models such as the unit hydrograph method (Nash, 1957) are often used because of their simplicity and rapidity. Gremillion *et al.*

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

(2000) investigated runoff production over an urbanizing catchment in Florida using a hydrograph separation method. Calabro (2001) applied a unit hydrograph method to the prediction of urban runoff over catchments in Italy and Yugoslavia. The use of reservoir models is also reported in the literature (e.g. Kang *et al.*, 1998; Cheng and Wang, 2002).

In a second approach, a conceptual distinction is made between runoff production and flood routing. In semidistributed, conceptual models the urban area is decomposed into subcatchments that contribute individually to runoff production. The runoff at the outlet of the various subcatchments is routed to the catchment outlet by means of a reservoir-based method (Boyd *et al.*, 1996; Hsieh and Wang, 1999) or a kinematic wave-based approach (Aronica and Cannarozzo, 2000). This approach has the advantage over the former in that the flow distribution (and consequently the distribution of the estimated risk) can be spatialized in greater detail. However the parameters of the routing methods used in such models do not often bear a physical meaning (Hsieh and Wang, 1999), which makes the analysis of urban development scenarios rather difficult.

The third category is that of physically based, distributed modeling. Physically based models have the advantage that the embedded flow parameters have a direct physical meaning and that scenario-based analyses thus become possible. Such models may be one-dimensional (Mark *et al.*, 2004), two-dimensional (Haider *et al.*, 2003; Turner-Gillespie *et al.*, 2003), or a combination of both (Hsu *et al.*, 2000). The use of three-dimensional Computational Fluid Dynamics (CFD) models in urban areas is also reported for the detailed investigation of the local features of the flow (Ma *et al.*, 2002). More integrated attempts focus on the coupling between the dynamics of surface water and the other compartments of the hydrological cycle (Valeo and Moin, 2000a,b; Jia *et al.*, 2001).

The strong variability of the urban geometry has a direct consequence on the variability of the flow patterns. In most cases, the flow characteristics cannot be considered to be uniform and the classical one-dimensional approach is not valid any more and the usual assumptions of equality of levels or heads at junctions (Cunge *et al.*, 1980) does not hold. Conversely, the refined simulation of the flow patterns in urban areas (e.g. using two-dimensional models) require an amount of data and a computational power that is hardly compatible with the constraints associated with classical engineering studies. This is illustrated by an example of a refined, two-dimensional urban flood simulation over a 290 m × 280 m area in the center of Nice, France (Gourbesville *et al.*, 2004). The cell size of the Digital Elevation Model (DEM) is 1 m × 1 m (Figure 1). The effect of the injection of a hypothetical hydrograph with a peak value 100 m^3 s^{-1} along a main street is simulated using the Shallow Water 2D (SW2D) computational code (Guinot, 2004) that solves the two-dimensional shallow-water equations using Godunov-type algorithms.

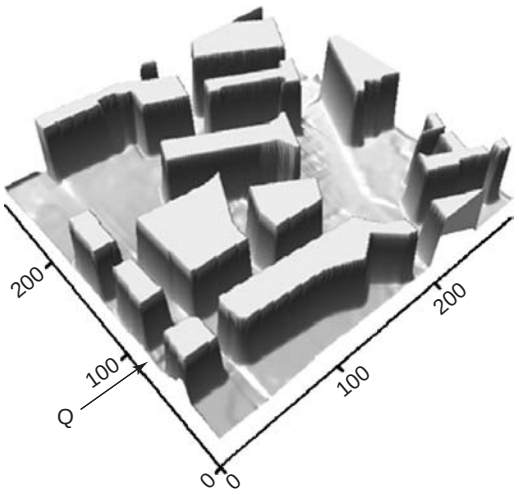


Figure 1 Perspective view of the DEM used for the refined simulation; distances in meters (By courtesy of Philippe Gourbesville, University of Nice, France)

Figure 2 shows the water depths computed at intervals of 60 s. The resolution of the DEM and the refinement of the computational grid allow the flow patterns to be investigated in detail (note the multiple front reflection across the street in the N-E part of the model at $t =$ 2 min). However the typical ratio of CPU to simulated time for the 80 000 cell model is about 6 on a 3.2 GHz Pentium 4 processor with 1 GB RAM. Clearly, the refined simulation of flow transients over entire urban areas cannot be foreseen to become a daily engineering practice in the near future.

Shallow Water Models with Porosity

A reasonable alternative to refined two-dimensional modeling consists in using large-scale, or so-called macroscopic flow models, in which a computational cell covers part or the totality of a house block. Such models may be used to represent the broad characteristics of the flow and its distribution between the various districts of a given urban area. The computational results can be used as boundary conditions for smaller-scale, nested models with a more refined grid, in which all the details of the urban geometry are represented. The formulation of large-scale urban hydrodynamics differs from the classical free-surface flow formulation in that part of the land surface is occupied by buildings or structures and is therefore not available to the flow. This is accounted for by introducing a so-called porosity. The shallow-water equations with porosity

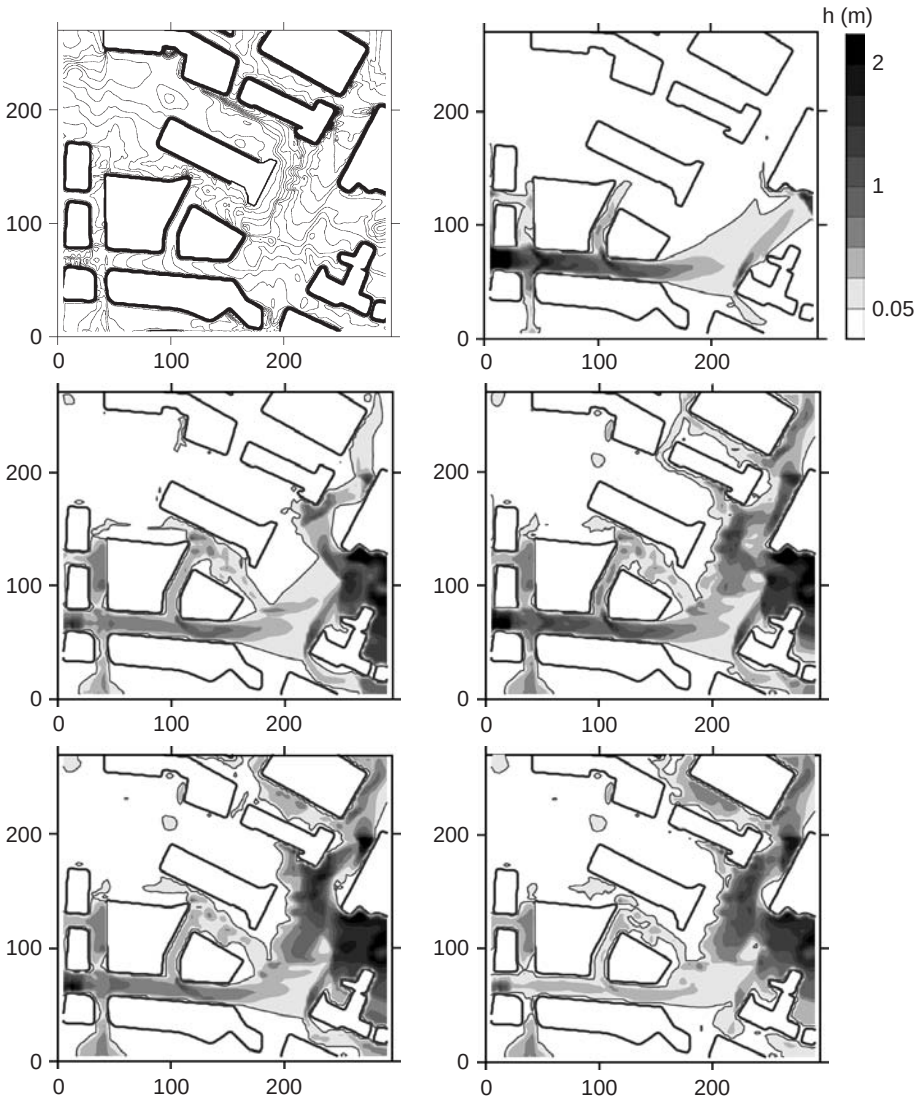


Figure 2 Model topography and water depths simulated using the SW2D computational code. Time interval between the plots: 60 s. Topographical contour line spacing: 0.25 m

were introduced in the mid-1990s (Defina *et al.*, 1994; Hervouet *et al.*, 2000). They can be written in conservation form as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = \mathbf{S} \tag{1}$$

where the conserved vector variable **U**, the vector fluxes **F** and **G** and the source term **S** are defined as

$$\mathbf{U} = \begin{bmatrix} \phi h \\ \phi uh \\ \phi vh \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} \phi uh \\ (u^2 h + gh^2/2)\phi \\ \phi uvh \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} \phi vh \\ \phi uvh \\ (v^2 h + gh^2/2)\phi \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 0 \\ (S_{0,x} - S_{f,x})\phi gh \\ (S_{0,y} - S_{f,y})\phi gh \end{bmatrix} \tag{2}$$

where g is the gravitational acceleration, h is the water depth, u and v are the x- and y-velocities respectively, $S_{0,x}$ and $S_{0,y}$ are the source terms arising from the bottom slope and the porosity gradient in the x- and y-directions respectively, $S_{f,x}$ and $S_{f,y}$ are the friction slopes in the x- and y-directions respectively, and ϕ is the porosity. The porosity is defined as

$$\phi = \frac{V}{Ah} \tag{3}$$

where A is the total plan view area and V is the volume actually occupied by the water. Note that Ah is the total volume of water in the absence of structures or objects on the ground surface. The source terms $S_{0,x}$ and $S_{0,y}$ are

defined as

$$\left.\begin{aligned} S_{0,x} &= -\phi g h \frac{\partial z_b}{\partial x} + g\frac{h^2}{2}\frac{\partial \phi}{\partial x} \\ S_{0,y} &= -\phi g h \frac{\partial z_b}{\partial y} + g\frac{h^2}{2}\frac{\partial \phi}{\partial y} \end{aligned}\right\} \quad (4)$$

Equation (1) can also be written in characteristic form as

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} + \mathbf{B}\frac{\partial \mathbf{U}}{\partial y} = \mathbf{S} \quad (5)$$

where $\mathbf{A}$ and $\mathbf{B}$ are the Jacobian matrices of $\mathbf{F}$ and $\mathbf{G}$ respectively with respect to $\mathbf{U}$. Equation (2) leads to the following expressions for $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & 0 \\ -uv & v & u \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 1 & 0 \\ -uv & v & u \\ c^2 - v^2 & 0 & 2v \end{bmatrix} \quad (6)$$

These are the characteristic matrices for the classical, two-dimensional shallow-water equations (see e.g. Daubert and Graffe, 1967; Katopodes and Strelkoff, 1979; Gerritsen, 1982; Guinot, 2003a,b for the derivation of the equations and analysis of the solution properties). The wave celerities and the stability properties of explicit discretizations of the conservation form (1) are identical to those of the classical shallow-water equations. The problems induced by the discretization of (1) mainly arise from the discretization of the fluxes and source terms in the presence of strong variations in the bottom level and the porosity. These parameters can be expected to vary strongly in real-world applications. Inadequate discretization of the source terms may lead to the instability of the numerical solution. The most common available options for the discretization of source terms arising from bottom slope are the so-called well-balanced approach and associated techniques (Greenberg and Le Roux, 1997; Alcrudo and Benkhaldoun, 2001), the quasi-steady wave propagation algorithm (LeVeque, 1998) and source term upwinding (Bermudez and Vasquez-Cendon, 1994).

In the proposed approach, equation (1) is discretized using the finite-volume formalism (Figure 3). The computational domain is discretized into cells, over which the average value of the conserved variable $\mathbf{U}$ is known. Integrating equation (1) over a cell between two time levels n and $n+1$ yields the following formula

$$\left.\begin{aligned} \mathbf{U}_i^{n+1,h} &= \mathbf{U}_i^n - \frac{\Delta t}{A_i}\sum_{j\in \mathrm{N}(i)} (\mathbf{F}_{i,j}^{n+\frac{1}{2}} n_{i,j}^{(x)} + \mathbf{G}_{i,j}^{n+\frac{1}{2}} n_{i,j}^{(y)}) \\ &\qquad w_{i,j} + {\mathbf{S}_{0_i}}^{n+\frac{1}{2}} \Delta t \\ \mathbf{U}_i^{n+1} &= \mathbf{S}_f(\mathbf{U}_i^{n+1,h}) \end{aligned}\right\} \quad (7)$$

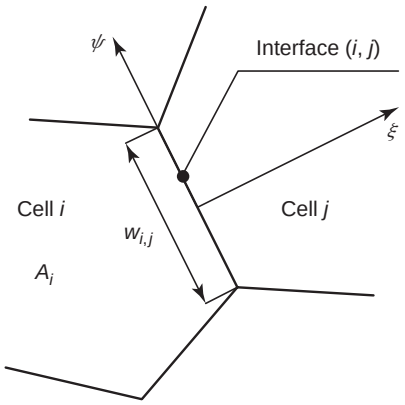


Figure 3 Definition sketch of the geometry

where A_i is the area of the cell i, $\mathrm{N}(i)$ is the set of neighbor cells of the cell i, $n_{i,j}^{(x)}$ and $n_{i,j}^{(y)}$ are the x- and y-components of the normal unit vector between the cells i and j (oriented from i to j), $w_{i,j}$ is the width of the interface (i, j) between the cells i and j, $\mathbf{U}_i^n$ is the average value of $\mathbf{U}$ over the cell i at the time level n, $\mathbf{F}_{i,j}^{n+1/2}$ and $\mathbf{G}_{i,j}^{n+1/2}$ are the average values of $\mathbf{F}$ and $\mathbf{G}$ respectively between the time levels n and $n+1$ over the interface (i, j), ${\mathbf{S}_{0_i}}^{n+1/2}$ and ${\mathbf{S}_{f_i}}^{n+1/2}$ are the average values of the source terms over the cell i between the time-steps n and $n+1$, and Δt is the computational time-step. The superscript h denotes the solution obtained at the time level $n+1$ after the solution of the hyperbolic terms only. Computing the values of $\mathbf{F}_{i,j}^{n+1/2}$ and $\mathbf{G}_{i,j}^{n+1/2}$ at all the cell interfaces allows $\mathbf{U}_i^{n+1,h}$ to be computed for all the cells. In the present approach, the fluxes and source terms are computed explicitly from the average values $\mathbf{U}_i^n$ at the time level n. Equation (6) is rewritten in the local coordinate system as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial \xi} = \mathbf{S}_0 \quad (8)$$

where ξ denotes the coordinate in the normal direction to the interface. The flux $\mathbf{F}$ is computed using a one-dimensional Riemann solver, namely, a procedure that solves the following initial-value problem, also called *Riemann problem*

$$\mathbf{U}(\eta, t=0) = \begin{cases} \mathbf{U}_\mathrm{L} & \text{for } \xi < 0 \\ \mathbf{U}_\mathrm{R} & \text{for } \xi > 0 \end{cases} \quad (9)$$

where $\mathbf{U}_\mathrm{L}$ and $\mathbf{U}_\mathrm{R}$ are the (constant, non necessarily equal) left and right states of the Riemann problem. The solution algorithm consists of the following steps:

1. For each interface (i, j), convert the variable $\mathbf{U}$ to the local coordinate system (ξ, ψ) formed by the normal and tangent unit vectors to the interface. Define

the Riemann problem in the direction normal to the interface (i, j).
2. Solve the Riemann problem using an exact or approximate Riemann solver such as Roe's solver (Roe, 1981), the HLLC solver (Toro *et al.*, 1994) or any approximate state solver (Glaister, 1988; Guinot, 2003a). This yields the constant value $\mathbf{F}_{i,j}^{n+1/2}$ at the location $\xi = 0$.
3. Carry out the mass balance on the cell i according to equation (6).
4. Add the effect of the source term in the ξ-direction using the source term upwinding approach (Bermudez and Vasquez-Cendon, 1994).
5. Use the solution obtained from the previous four steps as a starting point to incorporate the effects of friction.

The robustness of the porosity approach is illustrated by simulating the effects of a flash flood wave propagating into an initially dry plain, in the middle of which the extension of an existing urban area is planned. The site chosen for this application is the urban area of Saint Mathieu (South of France). The area under study is rectangular, with dimensions 2700 m × 2200 m (Figure 4). The ground levels range from 88 m to 120 m. The city is currently located on the right bank of the valley. The shallow-water model with porosity is used to simulate the propagation of a flash flood wave into the valley. Two simulations are presented hereafter. In the first simulation the valley is free from any buildings. In the second simulation the urban zone is extended into the valley, as indicated by the thick dashed line in Figure 5. The urban area is characterized by a porosity of 25%. The same input hydrograph is used in both simulations. The peak discharge is equal to $100\,m^3\ s^{-1}$. The DEM has a grid size of 25 m × 25 m. The effects of the urbanized area on the floodplain dynamics are clearly visible in Figure 5. The reduction in the cross-sectional area and in storage capacity results in a head loss that triggers a rise in the water levels upstream of the urban area and contributed to the retardation of the flood (see the maps at $t = 600$ s and $t = 1200$ s). As a result, the stream is partly diverted to the west and flows across a saddle to fill the western part of the area.

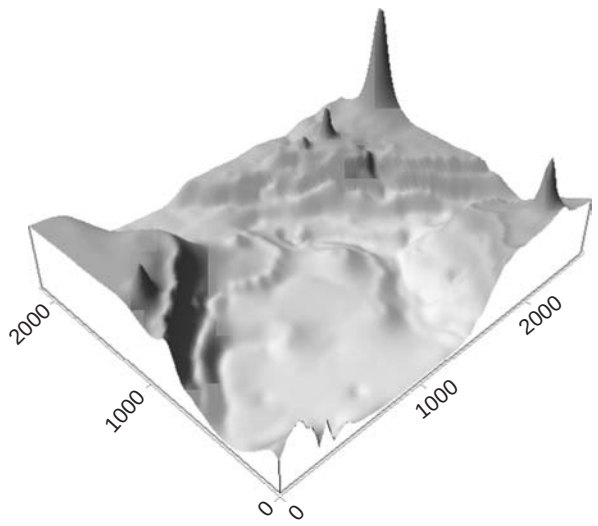


Figure 4 Perspective view of the DEM for the Saint Mathieu application. Ratio of vertical to horizontal scale 1 : 10

Table 1 Characteristics of the refined and large-scale simulations

	Refined model (Nice)	Large-scale model (Saint Mathieu)
Modeled area	7.6 ha	59 000 ha
Number of cells	75 600	9504
DEM grid size	1 m	25 m
Ratio CPU/real-time per unit area	8.5×10^{-1} ha^{-1}	1.9×10^{-7} ha^{-1}

The CPU time needed by the SW2D computational code to simulate the propagation of the 1 h hydrograph is approximately 40 s for both configurations. Note that the 3.2 GHz processor with 1.0 GB RAM used for these two simulations is the same as for the refined simulation in the center of Nice. Table 1 summarizes the characteristics of both models. Obviously, the geometries of the two test cases are not comparable. However, the differences between the CPU times show that the porosity approach is a viable alternative for the investigation of the flow patterns over urbanized areas.

CONCLUSIONS

Conceptual and semidistributed conceptual models probably represent the most widely used approaches in urban watershed modeling. Such models have the advantage that they do not require as much data as do distributed, physically based models. They are also simpler to operate and easier to tune. However, the parameters of such models do not bear the physical meaning that is attached to the more refined, and more computationally demanding, physically based approach. Therefore, conceptual and semidistributed models can hardly be operated in view of scenario-based analyses. Shallow water models with porosity appear as a viable compromise between refined CFD approaches and conceptual models for the investigation of surface water transients in urban areas. The outputs from such models may be used as inputs for local, refined physically based models in specific areas.

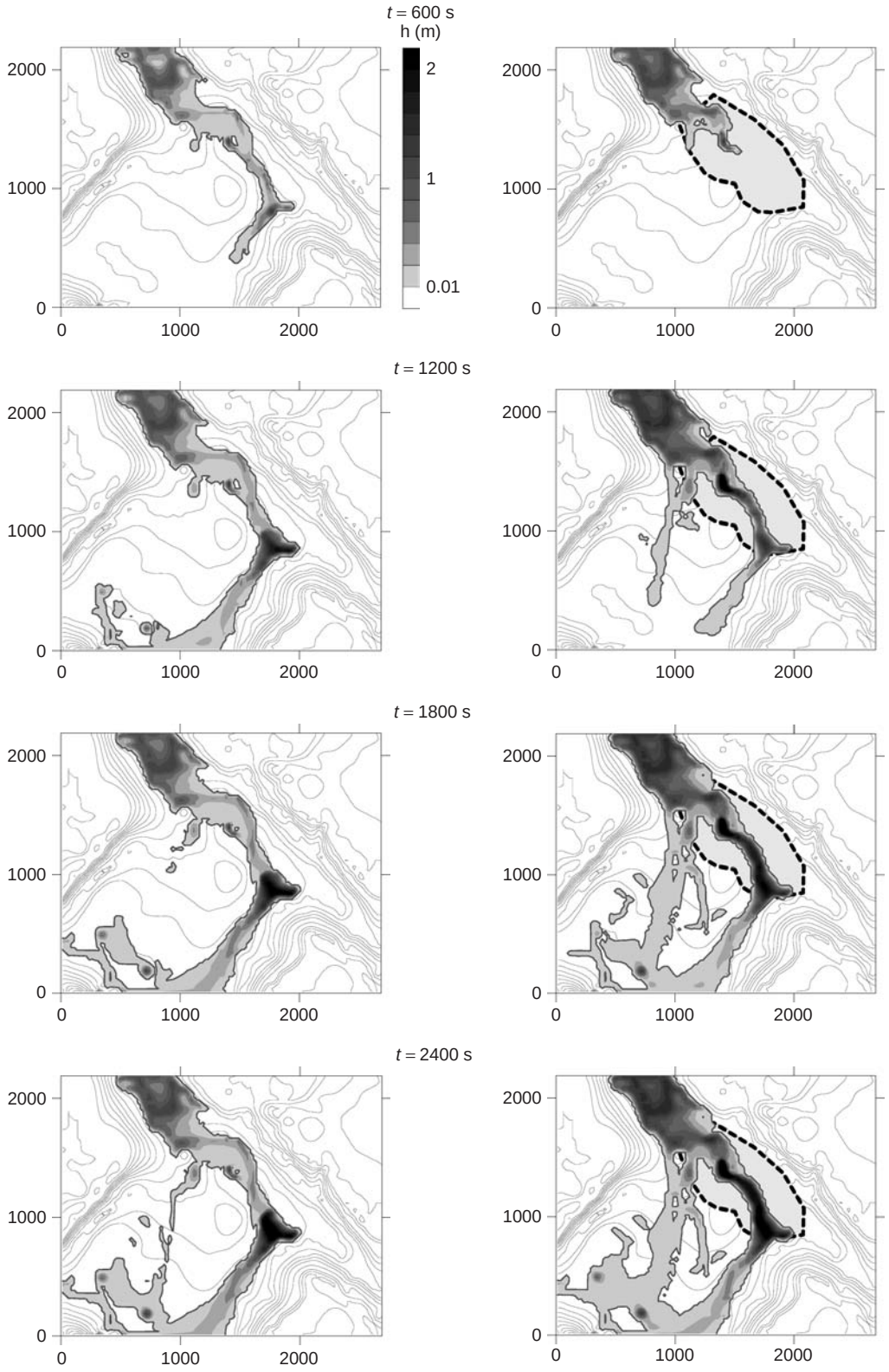


Figure 5 Influence of the urbanization of the plain of Saint Mathieu (south of France). Water depths simulated at various times using the shallow water with porosity without (a) and with (b) the hypothetical urbanized area. The planned urban extension is indicated by the thick dashed line in the maps on the right-hand side. Topographical contour line spacing 2.5 m; distances in meters

Acknowledgments

The DEM used for the refined flow simulation in Nice was kindly provided by Dr Philippe Gourbesville from the University of Nice-Sophia Antipolis (France).

REFERENCES

Alcrudo F. and Benkhaldoun F. (2001) Exact solutions to the Riemann problem of the shallow water equations with a bottom step. *Computers and Fluids*, **30**, 643–671.

Aronica G. and Cannarozzo M. (2000) Studying the hydrological response of urban catchments using a semi-distributed linear non-linear model. *Journal of Hydrology*, **238**, 35–43.

Bermudez A. and Vasquez-Cendon E. (1994) Upwind methods for hyperbolic conservation laws with source terms. *Computers and Fluids*, **23**, 1049–1071.

Berne A., Delrieu G., Creutin J.D. and Obled C. (2004) Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology*, **299**, 166–179.

Boyd M.J., Rigby E.H. and VanDrie R. (1996) WBNM – a computer software package for flood hydrograph studies. *Environmental Software*, **11**, 167–172.

Calabro P.S. (2001) Cosmoss: conceptual simplified model for sewer system simulation. A new model for urban runoff quality. *Urban Water*, **3**, 33–42.

Cheng S.-J. and Wang R.-Y. (2002) An approach for evaluating the hydrological effects of urbanization and its application. *Hydrological Processes*, **16**, 1403–1418.

Cunge J.A., Holly F.M. Jr and Verwey A. (1980) *Practical Aspects of Computational River Hydraulics*, Pitman Publishing.

Daubert A. and Graffe O. (1967) Quelques aspects des écoulements presque horizontaux à deux dimensions en plan et non permanents. Application aux estuaires. *La Houille Blanche*, **22**, 847–860 (in French).

Defina A., D'Alpaos L. and Matticchio B. (1994) A new set of equations for very shallow water and partially dry areas suitable to 2D numerical domains, *Proceedings of the Conference on 'Modelling of Flood Propagation Over Initially Dry Areas'*, Milan, 29 June–1 July.

Einfalt T., Arnbjerg-Nielsen K., Golz C., Jensen N.-E., Quirmbach M., Vaes G. and Vieux B. (2004) Towards a roadmap for use of radar rainfall data in urban drainage. *Journal of Hydrology*, **299**, 186–202.

Gerritsen H. (1982) *Accurate Boundary Treatment in Shallow Water Flow Computations*, Ph.D. Thesis, University of Twente, Netherlands.

Glaister P. (1988) Approximate Riemann solutions of the shallow water equations. *Journal of Hydraulic Research*, **26**, 293–306.

Gourbesville P., Pisot N., Le Fur H. and Lindberg S. (2004) High resolution digital elevation models: a major interest for urban flooding management, *Proceedings of the 6th Hydroinformatics Conference*, Singapore, Vol. 1, pp. 621–628, 21–24 June.

Greenberg J.M. and Le Roux A.Y. (1997) A well balanced scheme for the numerical processing of source term in hyperbolic equations. *SIAM Journal of Numerical Analysis*, **43**, 1980–2007.

Gremillion P., Gonyeau A. and Wanielista M. (2000) Application of alternative hydrograph separation models to detect changes in flow paths in a watershed undergoing urban development. *Hydrological Processes*, **14**, 1485–1501.

Guinot V. (2003a) *Godunov-type Schemes: an Introduction for Engineers*, Elsevier: p. 508.

Guinot V. (2003b) Riemann solvers and boundary conditions for two-dimensional shallow–water simulations. *International Journal for Numerical Methods in Fluids*, **41**, 1191–1219.

Guinot V. (2004) *Shallow Water 2D (SW2D). User manual, module and data structure, formulation and algorithms, verification manual*, Research report, Hydrosciences Montpellier, December 2004.

Haider S., Paquier A., Morel R. and Champagne J.-Y. (2003) Urban flood modelling using computational fluid dynamics. *Water and Maritime Engineering*, **156**, 1–8.

Hervouet J.M., Samie R., and Moreau B. (2000) Modelling dambreak flood waves in dambreak flood wave numerical simulation, *Proceedings of the International Seminar and Workshop on Rescue Action Based on Dam-Break Flood Analysis*, Seinäjoki, 1–6 October.

Hsieh L.S. and Wang R.Y. (1999) A semi-distributed parallel-type linear reservoir rainfall – runoff model and its application in Taiwan. *Hydrological Processes*, **13**, 1247–1268.

Hsu M.H., Chen S.H. and Chang T.J. (2000) Inundation simulation for urban drainage basin with storm sewer system. *Journal of Hydrology*, **234**, 21–37.

IMPACT (2001) *Investigation of Extreme Flood Processes and Uncertainty*, EC Contract EVG1-CT-2001-00037 IMPACT available at `http://www.impact-project.net`.

Jia Y., Ni G., Kawahara Y. and Suetsugi T. (2001) Development of WEP model and its application to an urban watershed. *Hydrological Processes*, **15**, 2175–2194.

Kang I.S., Park J.I. and Singh V.P. (1998) Effect of urbanization on runoff characteristics of the On-Cheon Stream watershed in Pusan, Korea. *Hydrological Processes*, **12**, 351–363.

Katopodes N. and Strelkoff T. (1979) Two-dimensional shallow water-wave models. *Journal of Engineering Mechanics Division (ASCE)*, **105**, 317–344.

LeVeque R.J. (1998) Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *Journal of Computational Physics*, **146**, 346.

Ma L., Ashworth P.J., Best J.L., Elliott L., Ingham D.B. and Whitcombe L.J. (2002) Computational fluid dynamics and the physical modelling of an upland urban river. *Geomorphology*, **44**, 375–391.

Mark O., Weesaku S., Apirumanekul C., Aroonnet S.B. and Djordjevic S. (2004) Potential and limitations of 1D modelling of urban flooding. *Journal of Hydrology*, **299**, 284–299.

Messerli B., Grosjean M., Hofer T., Núñez L. and Pfister C. (2000) From nature-dominated to human-dominated environmental changes. *Quaternary Science Reviews*, **19**, 459–479.

Nash J.E. (1957) *The Form of Instantaneous Unit Hydrograph*, *IAHS Publication* 51, International Association of Scientific Hydrologists: Gentbrugge, pp. 546–557.

Paquier A. and Mignot E. (2003) Use of 2D models to calculate flood water levels: calibration and sensitivity analysis, *Proceedings of 30th IAHR Congress*, Thessaloniki, Vol. C2, pp. 95–102, August.

Ragab R., Bromley J., Rosier P., Cooper J.D. and Gash J.H.C. (2003) Experimental study of water fluxes in a residential area: 1. Rainfall, roof runoff and evaporation: the effect of slope and aspect. *Hydrological Processes*, **17**, 2409–2422.

Roe P.L. (1981) Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, **43**, 357–372.

Schilling W. (1991) Rainfall data for urban hydrology: what do we need? *Atmospheric Research*, **27**, 5–22.

Toro E.F., Spruce M. and Speares W. (1994) Restoration of the contact surface in the HLL-Riemann solver. *Shock Waves*, **4**, 25–34.

Turner-Gillespie D.F., Smith J.A. and Bates P.D. (2003) Attenuating reaches and the regional flood response of an urbanizing drainage basin. *Advances in Water Resources*, **26**, 673–684.

United Nations (2002) *World Urbanization Prospects: the 2001 Revision*, United nations: New York.

Valeo C. and Moin S.M.A. (2000a) Variable source area modelling in urbanizing watersheds. *Journal of Hydrology*, **228**, 68–81.

Valeo C. and Moin S.M.A. (2000b) Grid-resolution effects on a model for integrating urban and rural areas. *Hydrological Processes*, **14**, 2505–2525.

19: Data-driven Modeling and Computational Intelligence Methods in Hydrology

DIMITRI P SOLOMATINE

UNESCO-IHE Institute for Water Education, Delft, The Netherlands

Along with the physically based (process) models based on mathematical descriptions of hydrological processes, the so-called data-driven models (DDM) are becoming popular. They are based on the use of methods of computational intelligence and machine learning and assume the presence of considerable amount of data describing the modeled phenomenon. The article covers various aspects of DDM, including the data preparation and a brief overview of the used techniques – neural networks, regression and model trees, instance-based learning, and nonlinear dynamics. A number of references to the successful application of data-driven methods are provided along with the links to the relevant web sites and software packages. An example where several data-driven methods were used in a hydrologic forecasting problem is provided.

INTRODUCTION

In the field of hydrology a range of modeling techniques can be applied. A *model* can be defined as a simplified representation of reality with the objective of explanation or prediction. Modeling includes studying the system, formulating its behavior, collecting and preparing data, building the model, testing it, using it, interpreting the results, and, possibly, iterating the entire procedure. In hydrology and hydraulics, the following types of models are commonly distinguished:

1. a *physical* or *scale* model (since it is usually smaller than the real system) built from material components or objects; in particular, in hydraulics these types of models have been well established;
2. a *mathematical* model based on the description of its behavior (often based on first-order principles from physics) of a phenomenon or system, referred to later as *process model* (also called *knowledge-driven, simulation, behavioral* or *physically based* model). In hydrology, the following simulation models are typically distinguished:
   (a) lumped conceptual models that operate with different but mutually interrelated storages representing physical elements in a catchment;
   (b) distributed physically based models that describe the natural system using the basic mathematical representations of the flows of mass, momentum, and various forms of energy;
3. an *empirical* model that involves mathematical equations that have been assessed not from the physical process in the catchment but from analysis of concurrent input and output time series. Typical examples here are the unit hydrograph method and various statistical models – for example, linear regression, multilinear, ARIMA, and so on. During the last decade, the area of empirical modeling received an important boost due to developments in the area of computational intelligence (CI), particularly, machine learning (ML). It can be said that it has now entered a new phase and deserves a special name – *data-driven modeling* (DDM).

DDM is based on the analysis of all the data characterizing the system under study. A model can then be defined on the basis of the connections between the system state variables (input, internal, and output variables) with only a limited number of assumptions about the "physical" behavior of the system. The methods used nowadays can go much further than the ones used in conventional empirical modeling: they allow for solving numerical prediction problems, reconstructing highly nonlinear functions, performing classification, grouping data, and building rule-based systems.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

CONTRIBUTORS TO DATA-DRIVEN MODELING

The following areas contributing to DDM can be mentioned: artificial intelligence (AI), data mining (DM), knowledge discovery in databases (KDD), CI, ML, intelligent data analysis (IDA), soft computing (SC), and pattern recognition – all these areas overlap, with often similar focus and application areas. Terminologies and research areas sometimes compete to name the same interdisciplinary area. It is difficult, if not impossible, to provide a formal definition of disparate areas with their own established individualities such as fuzzy logic, neural networks, pattern recognition, evolutionary computation, ML, Bayesian reasoning, and so on. It is really difficult to find clear-cut differences between them. Still certain definitions can be formulated:

- CI incorporates three large areas: neural networks, fuzzy systems, and evolutionary computing. CI tends to incorporate more and more areas that are also considered in AI (except perhaps symbolic methods) and ML.
- SC is an area that is very close to CI, but with a special focus on fuzzy rule-based systems (FRBS) induced from data.
- ML is an area of computer science, which was for a long time considered a subarea of AI and concentrates on the theoretical foundations used by CI and SC. Classification (pattern recognition) problems are addressed by ML more often than regression (numerical prediction) problems.
- DM and KDD are focused often at very large databases and are associated with applications in banking, financial services, and customer resources management. DM is seen as a part of a wider KDD. Methods used are mainly from statistics and ML.
- IDA is a relatively new term and seems to denote studies concentrating more on the data analysis in medicine and research. Methods used are also from statistics and ML.

In this connection we see DDM as an approach to modeling that focuses on using the CI (particularly ML) methods in building models (often of natural systems) that would complement or replace the "knowledge-driven" models describing behavior of physical systems. DDM uses methods developed in the fields mentioned earlier and tunes them to particular application areas. Examples of the most popular methods used in DDM of hydrological systems are: statistical methods, artificial neural networks (ANN) and FRBS. Among popular CI methods are also genetic algorithms (GA), evolutionary algorithms, and other global optimization algorithms; they are not, however, modeling paradigms or function approximation methods but constitute an approach to optimization that can be used in model calibration or model structure optimization. For more on evolutionary methods, *see* **Chapter 22, Evolutionary Computing in Hydrological Sciences, Volume 1**.

MACHINE LEARNING AS THE BASIS FOR DDM

One could see ML as the main source of method for DDM. An ML method (Figure 1) is an algorithm that estimates so-far-unknown mappings (or dependencies) between a system's inputs and its outputs from the available data (Mitchell, 1997). By data we understand known samples that are combinations of inputs and corresponding outputs. As such a dependency (*viz.* mapping, or "model") is discovered (induced), which can be used to predict (or effectively deduce) the future system's outputs from the known input values.

By data we usually understand a set K of examples (or instances) represented by tuples $\langle \mathbf{x}_k, \mathbf{y}_k \rangle$, where $k = 1, \ldots, K$, vector $\mathbf{x}_k = \{x_1, \ldots, x_n\}_k$, vector $\mathbf{y}_k = \{y_1, \ldots, y_m\}_k$, $n =$ number of inputs, $m =$ number of outputs. The process of building a function (or "mapping", or "model") $y = f(\mathbf{x})$ is called *training*. Very often only one output is considered, so $m = 1$ and the model to build takes the following form: $\mathbf{y} = f(\mathbf{x})$.

In statistics the following four types of data are considered: nominal (e.g. color, symbolic labels); ordinal (e.g. flood severity expressed as low, medium, or high); interval (e.g. temperature), and ratio (real-valued numeric, e.g. water level or flow). In ML there is a tendency to speak only of two data types: nominal (that includes also ordinal and sometimes integer numeric), and real-valued numeric.

There are four main styles of learning considered:

- classification – on the basis of classified examples, a way of assigning a class label to the unseen examples is to be found. In this case the output variable y takes nominal values;
- association – association between all variables characterizing the system is to be identified (typically based on finding combinations of values that are most frequent);

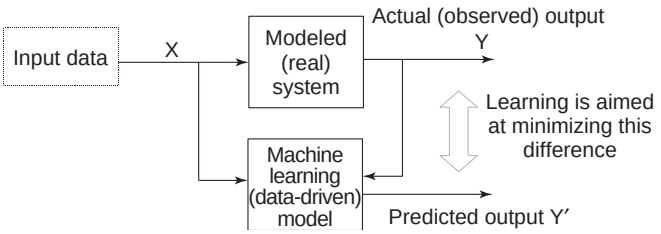


Figure 1 Learning in data-driven modeling. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

- numeric prediction (regression) – outcome is not a class, but a numeric real value or a vector ($\mathbf{y}_k \in \Re^m$). In engineering applications, inputs are often numeric as well, so $\mathbf{x}_k \in \Re^n$;
- clustering – groups of objects (examples) that are "close" in space of input variables are to be identified. In this case there is no explicit output variable.

Many numerical prediction methods can also be used for classification but the data should be rearranged. The number of outputs is made equal to the number of possible class labels, the nominal values of the output (class labels) in the training data should be replaced by a vector of real values with zeros for all outputs except the one that has the number corresponding to the original class label. When the model is trained, then it can predict the class label of a new example: the number of the output with the maximum value is taken as the class label number.

THE PROCESS OF DATA-DRIVEN MODELING

The following list of steps in model building is often distinguished (e.g. Pyle, 1999):

1. Select a clearly defined problem that the model will help to resolve.
2. Specify the type of solution to the problem.
3. Define how the solution delivered is going to be used in practice.
4. Learn the problem, collect the domain knowledge, and understand it. Clearly define assumptions, discuss them with the domain knowledge experts.
5. Let the problem drive the selection of modeling techniques.
6. Make the model as simple as possible, but no simpler. This rule is formulated sometimes in different ways – like KISS, for example ("Keep It Sufficiently Simple", or "Keep It Simple, Stupid!"). More generally, this idea is widely known as the *Occam's Razor principle* – formulated by William of Occam in 1320 in the following form: "shave all unneeded philosophy off the explanation".
7. Build (train) the model.
8. Refine the model iteratively (try different options until the model seems as good as it is going to get).
9. Test the model and evaluate the results.
10. Explore instabilities in the model (critical areas where small changes in inputs lead to large changes in output).
11. Define uncertainties in the model (critical areas and ranges in the data set where the model produces low confidence predictions).
12. Put the model into operation. Review it, if necessary, when experience is obtained.

TRAINING, CROSS-VALIDATION, VERIFICATION (TESTING)

In reality, the process of modeling is not linear but continuous with feedback loops. For example, the lack of particular data may lead to a change in the modeling method selected. Minimizing the difference between observed and predicted outputs actually has a double meaning. One should distinguish:

- minimizing the error during model training (calibration); and
- minimizing the error during model operation.

A model that has been trained may perform well (with low error) on the training data set, but when new instances, previously unseen by the model, are fed into the model, the error may be high. This means that it is necessary to have a separate data set (called *cross-validation* set), which does not contain instances from the training set and is used to judge the model error. A training set is sometimes called *in-sample data set.* Data that do not contain instances from the training set is often called *out-of-sample data set* (Figure 2).

As the model gradually improves as a result of the training process, the model error on the training (in-sample) set continuously decreases. However, the error on the out-of-sample set first decreases, but then starts to increase. This effect of "being trained too long" results in *overfitting* – the model tries to follow all data points (possibly noisy) too closely "forgetting" actually about the underlying trend in the data. It is said that the error on the out-of-sample set presents the *generalization accuracy* of a model – the accuracy with which it fits examples beyond the training data.

The presented example prompts for the necessity of using a cross-validation data set along with the training set. Training actually should stop when the error on cross-validation data set starts to increase – this is an indicator of the start of overfitting.

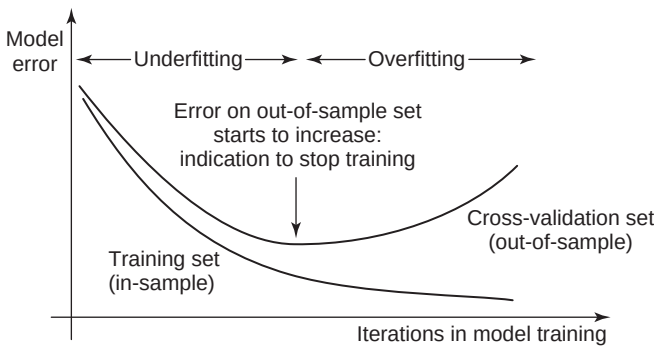


Figure 2 Use of cross-validation (out-of-sample testing) during training

When a model is put into operation, it will confront input instances that it never encountered before. In order to test its performance before the model is put into operation, it has to be verified (by some form of error measurement). For this purpose, yet another data set is used, called *verification (test) data set*. It allows to see how the model would perform if new data is fed into it.

Construction of these three data sets on the basis of available data, should follow the following principle: the three sets should be statistically similar, that is, they should have at least similar ranges and variability. This can be achieved by the careful selection of examples for each data set to ensure such statistical similarity, or for example, by random sampling of data from the whole data set.

The following summarizes the use of these three data sets: the training set is used to train the model; the cross-validation set is used to determine the moment of stopping the training process and possibly additional tuning of model parameters; the verification (test) set is used to measure the error of the trained model before the model is put into operation.

MODELS DRIVEN BY NOMINAL DATA: CLUSTERING AND CLASSIFICATION

Classification is often treated as attributing class labels to data points $\{x_k\} \in R^n$. Classes must be such that points in a class are close to each other in some sense, and classes are far from each other. *Clustering* is finding groups (subsets) of data without assigning them to particular classes. These types of models are in essence a mapping from the space of input data to classes or groups.

Popular methods for clustering are: partition-based clustering based on Euclidean distance, for example, k-means; fuzzy c-means; self-organizing feature maps (SOFM) also called *Kohonen networks* (Duda *et al.*, 2001). For classification, the following methods can be mentioned: k-nearest neighbors method, Bayesian classification, Decision trees (DTs) classification (Quinlan, 1992; Witten and Frank, 2000), support vector machines (SVM) classification (Vapnik, 1998), and ANN, (*see also* **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**). These last two methods can be used both for classification and for prediction.

Examples of Applications

In hydrology, clustering and classification are used much less frequently than prediction algorithms. However, a number of successful examples of their use were reported:

- Frapporti *et al.* (1993) used the method of fuzzy c-means clustering in the problem of classifying shallow Dutch groundwater sites into homogeneous groups.
- Hall and Minns (1999) used SOFM for classifying catchments into groups based on their characteristics, and then applying ANN to model the regional flood frequency.
- Hannah *et al.* (2000) used clustering for finding groups of hydrographs on the basis of their shape and magnitude; clusters are then used for classification by experts.
- Harris *et al.* (2000) applied clustering to identify the classes of river regimes.
- Velickov *et al.* (2000) used SOFM (Kohonen networks) as clustering methods, and SVM as classification method in aerial photograph interpretation with the purpose of subsequent construction of flood severity maps.
- Solomatine and Maskey (2005) used DTs and k-NN in the classification of river flow levels according to their severity in a flood forecasting problem in Nepal.

In this last problem a medium-sized foothill-fed river in the Bagmati basin was considered, having an area of about $3700\,\mathrm{km}^2$. Time series data of rainfall at three stations within the basin with daily sampling over eight years (1988 to 1995) were collected. Daily flows were recorded at one station, so this precluded modeling the routing. Weight factors were calculated using the Thiessen polygon. The daily evapotranspiration was computed using the modified Penman method recommended by FAO. A dependency analysis of input and output variables was done by visual inspection and the interdependencies between variables and the lags τ were established using correlation and average mutual information analyses. The classification model predicting the flow class one day ahead on the basis of five variables was set to be of the form

$$Q_{t+1} = f(RE_{t-2}, RE_{t-1}, RE_t, Q_{t-1}, Q_t)$$

where RE_{t-k} and Q_{t-k}, are the lagged effective rainfall and runoff respectively.

Eight years of data sets (2919 records) were split as follows: the first 919 points (3 January 1988 to 7 July 1990) were used as testing data set, and the remaining points (8 July 1990 to 30 December 1995) as training data set. Each instance was represented by a vector in a five-dimensional space (since there are five inputs) accompanied by the associated value of its output variable.

The forecasting problem was formulated in two ways: as a classification problem and as a prediction problem. For classification, the runoff data was discretized on the basis of the statistical properties and the following five classes were distinguished: "very low" (flow $<50\,\mathrm{m}^3\,\mathrm{s}^{-1}$), "low" (flow $50-300\,\mathrm{m}^3\,\mathrm{s}^{-1}$), "medium" (flow $300-750\,\mathrm{m}^3\,\mathrm{s}^{-1}$), "medium high" (flow $750-1350\,\mathrm{m}^3\,\mathrm{s}^{-1}$), and "high" (flow $>1350\,\mathrm{m}^3\,\mathrm{s}^{-1}$). Decision tree (DT) and k-nearest neighbor classifiers were trained – their errors in testing error were found to be 89% and 96% respectively.

MODELS DRIVEN BY REAL-VALUED DATA: SIMPLE FUNCTIONS COMBINED

Most hydrologic modeling problems are formulated with the view of real-valued data. The problem of prediction of real-valued variables is also called a *regression problem*. Since ML aims at finding a function that would best approximate some given function, it can also be seen as a problem of function-fitting, and this prompts for the use of the corresponding methods already available like linear regression, polynomial functions like splines or orthogonal polynomial functions, for example, Chebyshev polynomials. Current trend in DDM, however, is in combining many simple functions. It has been mathematically proven that adding up simple functions allows for a universal approximation of more complex functions (Kolmogorov, 1957).

Radial basis functions (**RBF**) could be seen as a sensible alternative to the use of complex polynomials. The idea is to approximate some function $y = f(x)$ by a superposition of J functions $F(x, \sigma)$, where σ is a parameter characterizing the span, or "width" of the function in the input space (Figure 3).

Functions F are typically "bell-shaped" (e.g. a Gaussian function is often used) so that they are defined in the proximity to some "representative" locations (centers) w_j in n-dimensional input space, $j = 1, \ldots, J$ and their values are close to zero far from these centers. Then the approximation becomes

$$f(x) = \sum_{j=1}^{J} b_j \exp(-\delta_j^2/\sigma_j^2)$$

where δ is the distance from point x to center w_j understood in Euclidean sense:

$$\delta_j = \sqrt{\sum_{i=1}^{n}(x_i - w_{ij})^2}$$

Finding the positions of centers w_j and the parameter σ of the functions $f(x, \sigma)$ is the aim of learning and can be done by building a *RBF neural network*; its training allows to identify these unknown parameters. Ideologically, RBF networks are close to the modular models considered in the following text.

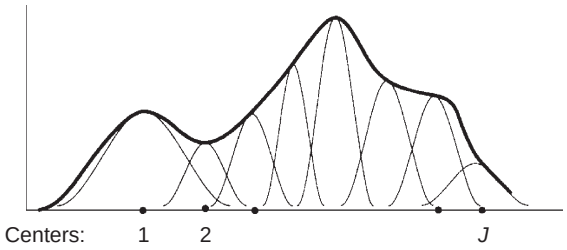


Figure 3 Using a number of radial basis functions (RBFs) as an approximation (an example for one input variable)

Multilayer perceptron (**MLP**) is a typical example of an ANN. It consists of several layers of connected nodes; each node calculates the weighted sum of its inputs and subjects the result to a nonlinear transformation (sigmoid of hyperbolic tangent function) which then serves as the input for all nodes in the next layer. Typically there are two layers. The weights are found by training, that is, by solving the problem of minimizing the model error in the space of weights (Haykin, 1999).

In the 1970–1980s the so-called backpropagation method for efficient training of MLPs was found and perfected, and this made this type of ANN the most popular ML tool of today. Various types of ANNs are widely used for prediction and classification.

A typical MLP network is presented in Figure 4.

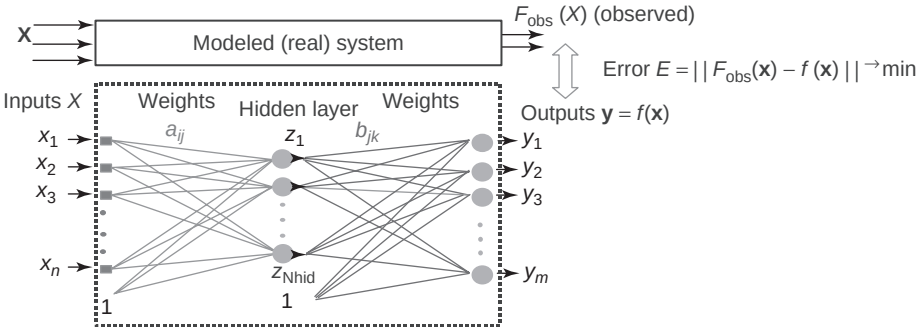


Figure 4 Structure of a multilayer perceptron (MLP) network. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Intermediate ("hidden") nodes calculate the weighted sum of the inputs and then perform a nonlinear transformation F of the form:

$$z_j = F\left(a_{oj} + \sum_{i=1}^{N_{inp}} a_{ij} x_i\right)$$
$$j = 1, \ldots, N_{hid}$$

The output nodes perform a similar transformation but use different weights:

$$y_k = F\left(b_{ok} + \sum_{i=1}^{N_{hid}} b_{jk} y_j\right)$$
$$k = 1, \ldots, m$$

The function F enables the presence of nonlinearity in the model and "squashes" the output signal. Often the so-called logistic function (Figure 5) is used:

$$F(u) = \frac{1}{1 + e^{-u}}$$

Training of MLP is in fact an optimization procedure aimed at finding such weight vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ that minimize error E.

Examples of Applications

The use of ANNs is known to have several dozens of successful applications in hydrology and related problems:

- modeling rainfall-runoff processes: Hsu *et al.* (1995), Minns and Hall (1996), Dawson and Wilby (1998), Dibike *et al.* (1999), Abrahart and See (2000), Govindaraju and Ramachandra Rao (2001);
- replicating the behavior of hydrodynamic/hydrological models of a river basin where ANNs are used in model-based optimal control of a reservoir (Solomatine and Torres, 1996);
- building ANN-based intelligent controller for real-time control of water levels in a polder (Lobbrecht and Solomatine, 1999); and
- modeling stage-discharge relationships (Sudheer and Jain, 2003; Bhattacharya and Solomatine, 2005).

For more on the use of ANNs in hydrology, *see also* **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**.

COMBINING MODELS: LOCAL MODELS, BOOSTING, COMMITTEE MACHINES, AND TREES

Hydrologic phenomena are multistationary and are composed of a number of processes, and their accurate modeling sometimes is not possible by building of one single ("global") model. In this case, the training set can be split into a number of subsets (possibly, statistically sampled) and separate models are trained on these subsets. It can also be said that the input space can be divided into a number of subspaces or regions for each of which a separate specialized model is built. These models are called *local*, or *expert models*, or *experts* (not to confuse with the human domain experts). Such modular model is often called a *committee machine* (CM) (Haykin, 1999).

When building, training and using a CM, two decisions have to be made: (A) which module should receive which training pattern (splitting problem); and (B) how the outputs of the modules should be combined to form the output of the final output of the system (combining problem). Accordingly, two decision units have to be built, or one unit performing both functions. Such a unit is called an *integrating unit*, or a *gating network* (a reference to a neural network often used for this purpose). It should be delivered to the user of the final model, along with the trained modules. Note that the functioning of units A and B could be different during training and operation. Figure 6 illustrates the principle.

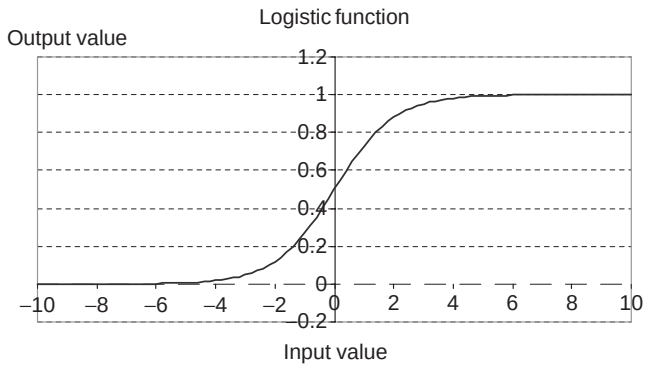


Figure 5 Logistic function A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

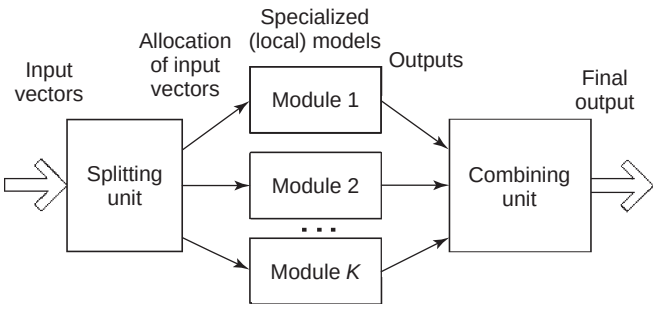


Figure 6 Splitting inputs, local models and combining outputs

Soft Splitting of the Training Set (Input Space)

The group of the statistically driven approaches with "soft" splits of input space are represented by *mixtures of experts* (Jordan and Jacobs, 1995), *bagging* (Breiman, 1996), and *boosting* (Freund and Schapire, 1997). Here we will briefly introduce boosting only.

Boosting (its advanced version, *AdaBoost*, is described by Freund and Schapire, 1997) can be seen as a method of building a series of modular models using soft splits. It was designed to improve the performance of the so-called *weak classifiers*, that is, the ones that are just marginally better than guessing; however, it has a more general applicability. In the first iteration, the basis model is trained (this will be the first module) on the whole data set and prediction error is calculated for each data point. The probabilities for each data vectors to be selected for the next iteration is adjusted: it is increased if prediction for this data vector was poor. Using this distribution, the new data set of the same size is sampled from the original data set. This process is repeated n times thus resulting in n modules, each trained on different (intersecting) subsets. Combining unit B uses the weighted sum of the modules, weight being dependent on the accuracy of the module. During training, splitting unit A arranges recalculation of the distribution and proper resampling, and during operation simply distributes each new input vector to all modules. Boosting was originally developed for binary classification problems and was later extended to solve multiclass classification problems (*AdaBoost.M2*) and regression problems (*AdaBoost.R*). Yet another version of boosting for regression is *AdaBoost.RT* by Solomatine and Shrestha (2004). They have demonstrated on two problems of rainfall-runoff modeling that this algorithm is more accurate than other data-driven methods.

Hard Splitting of the Training Set (Input Space)

A number of methods do not combine the outputs of different models but explicitly use only one of them, the most appropriate one (a particular case when the weights of other expert models are zero). Such methods use "hard" splits of input space into regions. Each individual local model is trained individually on subsets of instances contained in these regions, and finally the output of only one specialized expert is taken into consideration. This can be done manually by experts on the basis of domain knowledge. Another way is to use information theory to perform such splits and to perform splitting progressively; examples are: DTs, regression trees, MARS (Breiman *et al.*, 1984), and M5 model trees (MTs) (Quinlan, 1992).

Regression Trees and M5 Model Trees

These ML techniques use the following idea: split the parameter space into areas (subspaces) and build in each of them a separate regression model of zero or first order (Figure 7). If models in the leaves are of zero order (numeric constants) then this model is called a *regression tree* (Breiman *et al.*, 1984), if the models are of first order (linear regression models) then the model is referred to as a M5 MT (Quinlan, 1992). "M5" stands for "Model trees, version 5". Tree-based models are constructed by a divide-and-conquer method. The set T is either associated with a leaf, or some test is chosen that splits T into subsets corresponding to the test outcomes and the same process is applied recursively to the subsets. The splitting criterion for the M5 MT algorithm is based on treating the standard deviation of the output values that reach a node as a measure of the error at that node, and calculating the expected reduction in this error as a result of testing each attribute at that node.

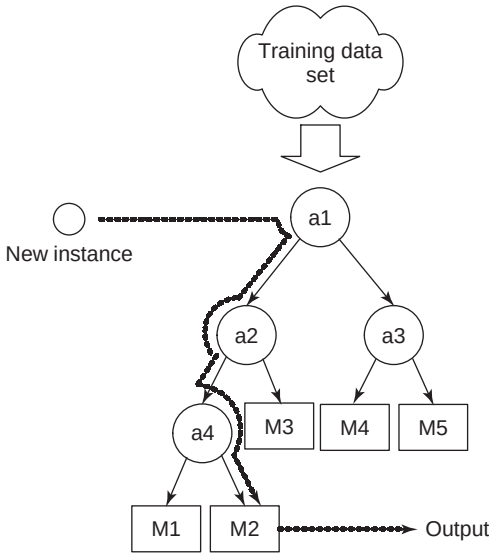


Figure 7 Building a tree-like modular model

In case of numeric inputs, the Boolean tests a_i used to split the data set have the form "$x_i < C$" where i and C are chosen to minimize the standard deviation in the subsets resulting from the split. *Mn* are local specialized models built for subsets filtered down to a given tree leaf. In fact, the resulting model can be seen as a committee of linear models being specialized on certain subsets of the training set belonging to particular regions of the input space.

This idea is not new; a combination of specialized models ("local" models) is used in modeling quite often. One can find a clear analogy between MTs and a combination of linear models used in dynamic hydrology already in the 1970s – a notable paper on multilinear models is by Becker and Kundzewicz (1987). However, the M5 MT approach – based on the principle of information theory – makes it possible to split the multidimensional parameter space and generate the models automatically according to the overall

quality criterion; it also allows for varying the number of models.

Each leaf in an MT represents a local model and, in principle, is (locally) more accurate than a global model (even a nonlinear one, e.g. a neural network) trained on the whole data set. The linear regression method is based on an assumption of linear dependencies between input and output. In an M5 MT a step toward nonlinearity is made by building a model that is locally linear, but overall is nonlinear. MTs may serve as an alternative to nonlinear models like ANNs (which are *global* models) and are often almost as accurate as ANNs but have some important advantages:

- training of MTs is much faster than ANNs, and it always converges;
- the results can be easily understood by decision makers;
- by applying *pruning* (that is making trees smaller by combining subtrees in one node) it is possible to generate a range of MTs – from an inaccurate but simple linear regression (one leaf only) to a much more accurate but complex combination of local models (many branches and leaves).

Probably the first application of M5 MTs in hydrologic forecasting was reported by Kompare *et al.* (1997). Solomatine and Dulal (2003) used M5 MT in rainfall-runoff modeling of a catchment in Italy (*see* the example at the end of this article).

Hybrid Models

Note that the models (modules) in Figure 6 may not be necessarily data-driven ones but rather have various nature, and may include expert judgments. If an overall model uses various types of models, it can be called a *hybrid model*. One of the important problems is incorporation of domain knowledge into the modeling process. A typical ML algorithm minimizes the training (cross-validation) error, seeing it as the ultimate indicator of the algorithm's performance, and so is purely data driven. Domain experts, however, may have other considerations in judging the quality of the model, and want to have certain control over the decisions (A) and (B) and over the choice of models used in each unit. These models could be not only data-driven ones based on ML but also physically based models based on the description of the underlying physical processes.

Xiong *et al.* (2001) combined several physically based forecasting models with the help of a fuzzy system. Solomatine and Maskey (2005) built a committee on the basis of several types of data-driven models including instance-based models, M5 MTs and neural networks.

In a study where flow predictions in the Huai river basin (China) were made on the basis of previous flows and precipitation, Solomatine and Xue (2004) built a committee hybrid model. The problem was to predict Q_{t+1} flow one day ahead. The following notations are used: flows on the previous and current day as Q_{t-1} and Q_t, respectively; precipitation on the previous day as P_{t-1}; moving average (two days) of the precipitation two days before as $P\text{mov}2_{t-2}$; moving average (three days) precipitation four days before as $P\text{mov}3_{t-4}$.

As a first step, domain experts were asked to identify several hydrological conditions (rules), used to split the input space into three regions:

1. $Q_{t-1} \geq 1000\,\text{m}^3\,\text{s}^{-1}$ (high flows)
2. $Q_{t-1} < 1000\,\text{m}^3\,\text{s}^{-1}$ and $Q_t \geq 200\,\text{m}^3\,\text{s}^{-1}$ (medium flows)
3. $P_{t-1} > 50$ and $P\text{mov}2_{t-2} < 5$ and $P\text{mov}3_{t-4} < 5$ (flood condition due to the short but intensive rainfall after a period of dry weather).

For each of these regions separate local models were built (M5 MTs and ANNs).

Addressing the problem of incorporating domain knowledge in DDM, Solomatine and Siek (2004) presented *M5flex* algorithm, an approach allowing for a more active role of an expert in building M5 trees, and demonstrated its accuracy in hydrologic modeling.

Complementary Models

Models can be combined not only to model the same process but also to complement each other. In such combined model a data-driven model can be used to correct errors of the primary model (this would be typically a physically based model, but can be a data-driven model as well), *see* Figure 8. Such an approach was employed in a number of hydrologic studies. Shamseldin and O'Connor (2001) used ANNs to update runoff forecasts; the simulated flows from a model and the current and previously observed flows were used as input, and the corresponding observed flow as the target output. Updates of daily flow forecasts for a lead-time of up to four days were made. It was reported that ANN models gave more accurate improvements than autoregressive models. Lekkas *et al.* (2001) showed that error forecasting provides improved real-time flow forecasting, especially when the forecasting model is poor. Abebe and Price (2004) used this approach to correct the errors of a routing model of the River Wye in UK by an ANN. Solomatine and Maskey (2005) built an ANN-based rainfall-runoff model where its outputs were corrected by an instance-based model.

NO EXPLICIT MODELS: INSTANCE-BASED LEARNING

In instance-based learning (IBL) (Mitchell, 1997) no model is built: classification or prediction is made directly by

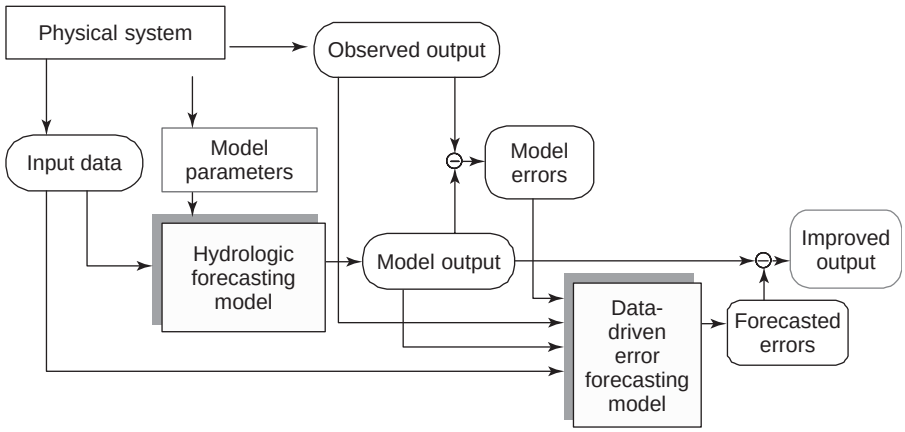


Figure 8 Complementary modeling. The output of the hydrologic forecasting model (physically based or data driven) is corrected by the data-driven model. (Adapted from Abebe and Price (2004).) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

combining instances from the training data set that are close to the new vector $\mathbf{x}_q$ of inputs. In fact, IBL methods construct a local approximation to the modeled function that applies well in the immediate neighborhood of the new query instance encountered; IBL methods are never intended to construct an approximation designed to perform well over the entire instance space. Thus it describes a very complex target function as a collection of less complex local approximations. IBL often demonstrates competitive performance when compared with, for example, ANNs.

The nearest neighbor classifier approach is one of the simplest and oldest methods of classification. It classifies an unknown pattern $\mathbf{x}_q$ by choosing the class of the nearest example x in the training set as measured by some distance metric, typically Euclidean.

Generalization of this method is the k-nearest neighbor (k-NN) method. For a discrete valued target function, the estimate will just be the most common value among k training examples nearest to $\mathbf{x}_q$. For real-valued target functions, the estimate is the mean value of the k-nearest neighboring examples. The k-NN algorithm can be improved by weighting each of the k neighbors according to their distance to the query point $\mathbf{x}_q$.

Further extensions are known as *locally weighted regression* (LWR) when the regression model is built on k-nearest instances. The training instances are assigned weights according to their distance to the query instance and the regression equations are generated on the weighted data.

Applications of IBL in hydrology mainly refer to the simplest method, *viz.* k-NN. Karlsson and Yakowitz (1987) were probably the first to use this method in hydrology, focusing, however, only on (single-variate) time series forecasts. Galeati (1990) demonstrated the applicability of the k-NN method (with the vectors composed of the lagged rainfall and flow values) for daily discharge forecasting and favorably compared it to the statistical ARX model. Shamseldin and O'Connor (1996) used the k-NN method for adjusting the parameters of the linear perturbation model for river flow forecasting. Toth *et al.* (2000) compared the k-NN approach with other time series prediction methods in a problem of short-term rainfall forecasting. Solomatine and Maskey (2005) considered IBL on a wider context of ML and explored several methods, tested their applicability in short-term hydrologic forecasting and compared their performance to other methods on a number of case studies and benchmark data sets; the LWR with Gaussian kernels appeared to be the most accurate IBL method.

OTHER METHODS

Chaos theory and nonlinear dynamics appear to be excellent predictive tools for time series that are of sufficient length. A single-variate model uses only the time series itself, without any use of other related variables, so it is applicable when the time series carries enough information about the behavior of the system (Lorenz 1963; Abarbanel 1996). Let a time series $\{x_1, x_2, \ldots, x_t, \ldots, x_n\}$ be given (e.g. a sequence of river water levels). The main idea is to represent the state of the system at time t by a vector in m-dimensional state space:

$$\mathbf{y_t} = \{x_t, x_{t-\tau}, x_{t-2\tau}, \ldots, x_{t-(m-1)\tau}\} \quad (1)$$

where τ is the delay time. The whole time series can then be represented by a sequence of such vectors $\{\mathbf{y}_t\}$:

$$\{\boldsymbol{y}_m, \boldsymbol{y}_{\mathrm{m}+1}, \ldots, \boldsymbol{y}_n\}.$$

If the original time series exhibits so-called chaotic properties, then its equivalent trajectory in the phase space has properties allowing for accurate predictions of the future

values of $\boldsymbol{y}$, and hence x. In practical applications, the delay time τ and the dimension m need to be appropriately chosen in order to fully capture the dynamic structure of the time series. Multivariate models embody time series representing several variables.

The predictive capacity of chaos theory, based on an idea that the system behaves in the future in a similar manner as in the past, supersedes the class of linear models like ARIMA. Solomatine *et al.* (2000) and Velickov *et al.* (2003) used chaos theory to predict the surge water level in the North Sea close to Hook of Holland; the data set included measurements of surge for 5 years with a 10-min interval. For a two-hourly prediction, the error was as low as 10 cm and was at least on par with the accuracy of hydrodynamic models. Babovic *et al.* (2000) used a chaos theory-based approach for predicting water levels at the Venice lagoon, and Phoon *et al.* (2002) used it for forecasting hydrologic time series. These methods do not have universal applicability: they can be successfully applied only when time series (or their combination) exhibit properties of chaotic behavior and when time series are of adequate (considerable) length.

Fuzzy Rule-based Systems (FRBS)

After fuzzy logic (as opposed to the classical "crisp" logic) was introduced by Lotfi Zadeh(Zadeh 1965), it found multiple successful applications, mainly in control theory (*see* Kosko, 1997). FRBS can be built by interviewing human experts, or by processing historical data and thus forming a data-driven model. The basics of the latter approach and its use in a number of water-related applications can be found in Bárdossy and Duckstein (1995). FRBS were effectively used for drought assessment (Pesti *et al.*, 1996), prediction of precipitation events (Abebe *et al.*, 2000b), analysis of groundwater model uncertainty (Abebe *et al.*, 2000a), control of water levels in polder areas (Lobbrecht and Solomatine, 1999), and modeling rainfall-discharge dynamics (Vernieuwe *et al.*, 2005). One of the limitations of FRBS is that the demand for computer memory grows exponentially with an increasing number of input variables.

Genetic Programming (GP) and Evolutionary Regression

GP is a regression method where various elementary mathematical functions, constants, and arithmetic operations are combined in one function. Such combination is sought to form a tree structure that is optimized by a multiextremum nonderivative based optimization method – usually a genetic (evolutionary) algorithm. More on GP and its applications can be found in **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1**.

Evolutionary regression is a method similar to GP, but in which the elementary functions are chosen from a limited set, and the structure of the overall function is fixed. Typically, polynomial regression equation is used, and the coefficients are found by genetic (evolutionary) algorithm. Giustolisi and Savic (2005) used this method for modeling groundwater level and river temperature.

Support Vector Machines (SVM)

SVM is a relatively new important method based on the extension of the idea of identifying a line (or a plane or some surface) that separates two classes in classification. It is based on statistical learning theory initiated by V. Vapnik in the 1970s (Vapnik, 1998, Kecman, 2001). This classification method was also extended to solving prediction problems, and in this capacity was used in hydrology-related tasks. Dibike *et al.* (2001) and Liong and Sivapragasam (2002) reported using SVMs for flood management and in prediction of water flows and stages.

AN EXAMPLE OF APPLYING DATA-DRIVEN MODELING TO HYDROLOGIC FORECASTING

A problem of short-term flow forecasting was posed for the Sieve River – a tributary of Arno River located in the Central Italian Apennines, Italy. The basin covers mostly hills, forests and mountainous areas except in the valley, with an average elevation of 470 m above sea level. A total of 2160 records of hourly discharge downstream, precipitation, and potential evapotranspiration data were available (December 1969; January and February 1970), which represent various types of hydrological conditions. The discharge data were available only for one location. The Arno basin that includes the Sieve catchment was extensively studied and was used as a case study for various physically based hydrologic modeling exercises (Todini 1996; Marsigli *et al.*, 2002). Data-driven methods were used by Solomatine and Dulal (2003). The objective is to forecast the future flow 1 h and 3 h ahead (Q_{t+1} and Q_{t+3} respectively) on the basis of previous values of effective rainfalls $RE_{t-\tau}$ and flows Q_{t-k} (where τ and k are time lags).

One of the important steps of DDM in hydrologic forecasting is data preparation and selection of the relevant input variables and time lags τ and k. Visual inspection of a number of rainfall events makes it possible to identify the time lags between peak rainfall and runoff which is around 5 to 7 h for this particular catchment. The highest correlation between lagged rainfall and discharge is at a lag of 6 h. The autocorrelation value for the discharge is 0.989 for a lag of 1 h (and drops as the time lag increases). Additionally, average mutual information can be used for the same purpose. Such analysis allows for identification of the proper lags τ and k (Figure 9).

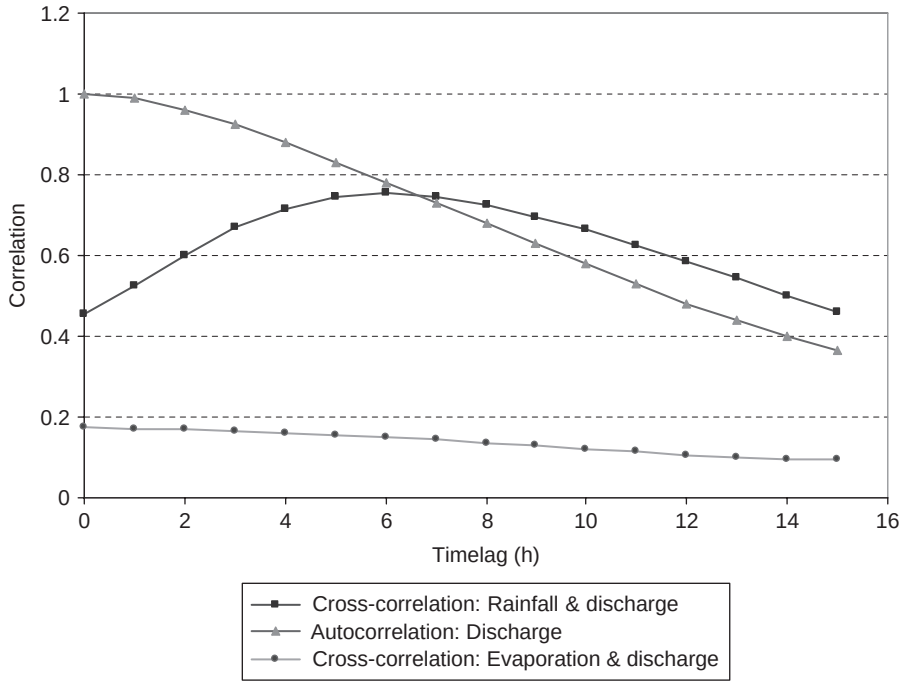


Figure 9 Correlation analysis used to assess the proper lags

The following models with eight and six input variables were built:

$$Q_{t+1} = f(RE_t, RE_{t-1}, RE_{t-2}, RE_{t-3}, RE_{t-4}, RE_{t-5}, Q_{t-1}, Q_t)$$

$$Q_{t+3} = f(RE_t, RE_{t-1}, RE_{t-2}, RE_{t-3}, Q_{t-1}, Q_t)$$

The first 300 examples were used as test data and the rest of data as training data set. The split was made in such a way that both sets have data of similar range and variability. The problem of classification, namely, predicting classes of flow rather than their numerical values was posed as well. The output was discretized into three levels: "low" (flow $<50\,m^3\,s^{-1}$), "medium" (flow $50-300\,m^3\,s^{-1}$), and "high" (flow $>300\,m^3\,s^{-1}$).

A number of DDM methods were applied: MLP ANN, RBF ANN, M5 MTs(Solomatine and Dulal, 2003; Solomatine and Siek 2004), IBL, LWR (Solomatine and Maskey, 2005). Software packages *NeuralMachine* [**http://www.data-machine.com**] and *NeuroSolutions* [**http://www.nd.com**] were used for the ANN modeling; MLP architecture was used. A nonlinear activation function, the hyperbolic tangent function (bounded between -1 and $+1$) was used in the hidden layer. A linear activation function was used in the output layer because it is unbounded and is able, to a certain extent, to extrapolate beyond the range of the training data. The number of hidden nodes was determined by trial-and-error optimization (performed automatically by the software). For training, the back propagation algorithm was used with momentum rule and stopping after 5000 epochs or when the mean squared error (MSE) reached 0.0001.

For M5 MT construction, *Weka* software was used (Witten and Frank, 2000). An example of a pruned (reduced) MT (to 3 rules) from the total of 16 rules is shown in the following:

```
if Q(t) <= 59.4 then
   if Q(t) <= 32.5 then LM1 (1011 examples)
                   else LM2 ( 396 examples)
if Q(t) > 59.4     then LM3 ( 447 examples)

LM1: Q(t+1) = 0.0388 + 0.0108RE(t)
              + 0.0535RE(t-1) + 0.0173RE(t-2)
              + 0.0346RE(t-3) + 1.01Q(t)
              - 0.0127Q(t-1) + 0.00311Q(t-2)
LM2: Q(t+1) = -0.221 + 0.0108RE(t)
              + 1.68RE(t-1) + 0.0626RE(t-2)
              + 7.3RE(t-3) + 1Q(t)
              - 0.0127Q(t-1) + 0.00311Q(t-2)
LM3: Q(t+1) = 3.04 + 2.46RE(t)
              + 4.97RE(t-1) - 0.04RE(t-2)
              + 1.75Q(t) - 1.08Q(t-1)
              + 0.265Q(t-2)
```

It can be seen that actually a mixture of three linear models was built; each of the models is trained on nonintersecting data sets corresponding to low- (below $32.5\,m^3\,s^{-1}$), medium- (between 32.5 and $59.4\,m^3\,s^{-1}$), and high (above $59.4\,m^3\,s^{-1}$) flows. The simple structure of the

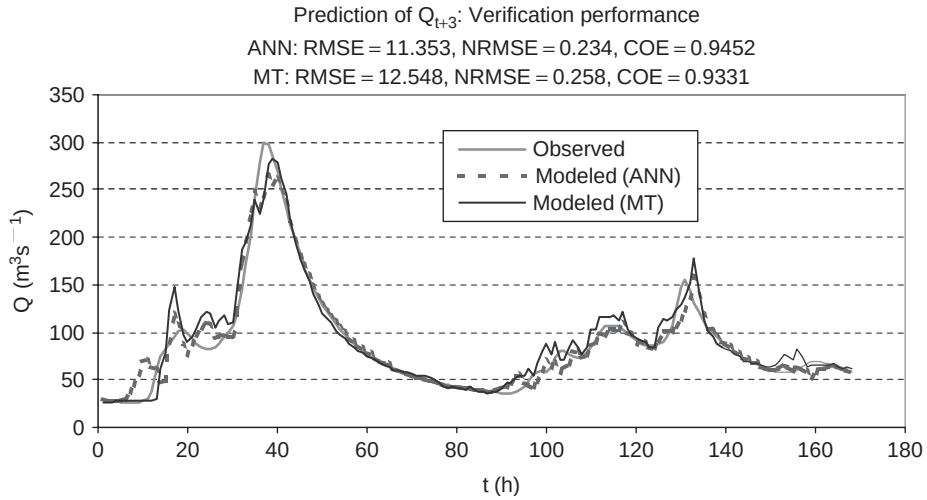


Figure 10 Flow forecasting using ANN and M5 MT

resulting individual models makes it possible to easily see the relative influence of all input variables and perform sensitivity analyses. The equations can be simplified even more by removing the terms with small coefficients. A similar model was generated for predicting Q_{t+3}.

One of the resulting plots comparing the performance of ANN and M5 MT in forecasting Q_{t+3} is shown in Figure 10. In solving the problem of classifying the conditions leading to a particular flow level (low, medium, and high), DTs and SVMs were used. The experiments show the applicability and high accuracy of DDM in short-term hydrologic forecasting.

CONCLUSIONS

DDM and CI methods have proven their applicability to various hydrologic problems: modeling, short-term forecasting, classification of hydrology-related data, building flood severity maps based on aerial or satellite photos, and so on. Normally, any particular (scientific) domain area will benefit from DDM if: (i) there is a considerable amount of data available; (ii) there are no considerable changes to the system during the period covered by the model; (iii) it is difficult to build knowledge-driven simulation models (e.g. due to lack of understanding of the underlying processes); or (iv) in particular cases, when available models are not adequate enough; and (v) there is a necessity to validate the simulation results of physically based models with other types of models.

It is always advisable to apply various types of DDMs and to compare and/or combine the results. For example, M5 MTs, combining local and global properties, could very well complement ANNs, and be more easily accepted by decision makers because of their reliance on simple linear models.

The future is seen in using the hybrid models-combining models of different types and following different modeling paradigms, including the combination with physically based models. It can be foreseen that computational intelligence (machine learning) will be used not only for building data-driven models, but also for building optimal and adaptive *model structures* of such hybrid models.

LINKS

www.data-machine.com (see software description)
www.datamining.ihe.nl
www.kdnuggets.com

REFERENCES

Abarbanel H.D.I. (1996) *Analysis of Observed Chaotic Data*, Springer-Verlag: New York.

Abebe A.J., Guinot V. and Solomatine D.P. (2000a) Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. *Proceedings of the 4th International Conference on Hydroinformatics*, Cedar Rapids.

Abebe A.J., Solomatine D.P. and Venneker R. (2000b) Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. *Hydrological Sciences Journal*, **45**(3), 425–436.

Abebe A.J. and Price R.K. (2004) Information theory and neural networks for managing uncertainty in flood routing. *ASCE Journal of Computing in Civil Engineering*, **18**(4), 373–380.

Abrahart R.J. and See L. (2000) Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecast in two contrasting catchments. *Hydrological Processes*, **14**, 2157–2172.

Babovic V., Keijzer M. and Stefansson M. (2000) Optimal embedding using evolutionary algorithms. *Proceedings of*

the 4th International Conference on Hydroinformatics, Cedar Rapids.

Bárdossy A. and Duckstein L. (1995) *Fuzzy Rule-Based Modeling with Applications to Geophysical, Biological and Engineering Systems*, CRC Press: Boca Raton.

Becker A. and Kundzewicz Z.W. (1987) Nonlinear flood routing with multilinear models. *Water Resources Research*, **23**, 1043–1048.

Bhattacharya B. and Solomatine D.P. (2005) Neural networks and M5 model trees in modelling water level–discharge relationship. *Neurocomputing*, **63**, 381–396.

Breiman L. (1996) Stacked regressor. *Machine Learning*, **24**(1), 49–64.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth International: Belmont.

Dawson C.W. and Wilby R. (1998) An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, **43**(1), 47–66.

Dibike Y., Solomatine D.P. and Abbott M.B. (1999) On the encapsulation of numerical-hydraulic models in artificial neural network. *Journal of Hydraulic Research*, **37**(2), 147–161.

Dibike Y.B., Velickov S., Solomatine D.P. and Abbott M.B. (2001) Model induction with support vector machines: introduction and applications. *ASCE Journal of Computing in Civil Engineering*, **15**(3), 208–216.

Duda R.O., Hart P.E. and Stork D.G. (2001) *Pattern Classification*, John Wiley & Sons: New York.

Frapporti G., Vriend P., Van Gaans P.F.M. (1993) Hydrogeochemistry of the shallow Dutch groundwater: Interpretation of the national groundwater quality monitoring network. *Water Resources Research*, **29**(9), 2993–3004.

Freund Y. and Schapire R. (1997) A decision-theoretic generalisation of on-line learning and an application of boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139.

Galeati G. (1990) A comparison of parametric and nonparametric methods for runoff forecasting. *Hydrological Sciences Journal*, **35**(1), 79–94.

Giustolisi O. and Savic D.A. (2005) A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics*, **7**, (in press).

Govindaraju R.S. and Ramachandra Rao A. (Eds.) (2001) *Artificial Neural Networks in Hydrology*, Kluwer: Dordrecht.

Hall M.J. and Minns A.W. (1999) The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal*, **44**, 693–704.

Hannah D.M., Smith B.P.G., Gurnell A.M. and McGregor G.R. (2000) An approach to hydrograph classification. *Hydrological Processes*, **14**, 317–338.

Harris N.M., Gurnell A.M., Hannah D.M. and Petts G.E. (2000) Classification of river regimes: a context for hydrogeology. *Hydrological Processes*, **14**, 2831–2848.

Haykin S. (1999) *Neural Networks: A Comprehensive Foundation*, McMillan: New York.

Hsu K.L., Gupta H.V. and Sorooshian S. (1995) Artificial neural network modelling of the rainfall-runoff process. *Water Resources Research*, **31**(10), 2517–2530.

Jordan M.I. and Jacobs R.A. (1995) Modular and hierarchical learning systems. In *The Handbook of Brain Theory and Neural Networks*, Arbib M. (Ed.), MIT Press: Cambridge.

Karlsson M. and Yakowitz S. (1987) Nearest neighbour methods for nonparametric rainfall runoff forecasting. *Water Resources Research*, **23**(7), 1300–1308.

Kecman V. (2001) *Learning and Soft Computing*, MIT Press: Cambridge.

Kolmogorov A.N. (1957) On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, **114**, 953–956.

Kompare B., Steinman F., Cerar U. and Dzeroski S. (1997) Prediction of rainfall runoff from catchment by intelligent data analysis with machine learning tools within the artificial intelligence tools. *Acta Hydrotechnica*, **16/17**, 79–94, (in Slovene).

Kosko B. (1997) *Fuzzy Engineering*, Prentice-Hall: Upper Saddle River.

Lekkas D.F., Imrie C.E. and Lees M.J. (2001) Improved nonlinear transfer function and neural network methods of flow routing for real-time forecasting. *Journal of Hydroinformatics*, **3**(3), 153–164.

Liong S.Y. and Sivapragasam C. (2002) Flood stage forecasting with SVM. *Journal of American Water Resources Association*, **38**(1), 173–186.

Lobbrecht A.H. and Solomatine D.P. (1999) Control of water levels in polder areas using neural networks and fuzzy adaptive systems. In *Water Industry Systems: Modelling and Optimization Applications*, Savic D. and Walters G. (Eds.), Research Studies Press: Baldock, pp. 509–518.

Lorenz E.N. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Science*, **20**, 130–141.

Marsigli M., Todini F., Diomede T., Liu Z. and Vignoli R. (2002) Calibration of rainfall-runoff models. Deliverable 9.1 of the Project report *MUSIC–Multiple-Sensor Precipitation Measurements, Integration, Calibration and Flood Forecasting*, EU Contract No. EVK1-CT-2000-00058.

Minns A.W. and Hall M.J. (1996) Artificial neural network as rainfall-runoff model. *Hydrological Sciences Journal*, **41**(3), 399–417.

Mitchell T.M. (1997) *Machine Learning*, McGraw-Hill: New York.

Pesti G., Shrestha B.P., Duckstein L. and Bogárdi I. (1996) A fuzzy rule-based approach to drought assessment. *Water Resources Research*, **32**(6), 1741–1747.

Phoon K.K., Islam M.N., Liaw C.Y. and Liong S.Y. (2002) A practical inverse approach for forecasting nonlinear hydrological time series. *ASCE Journal of Hydrologic Engineering*, **7**(2), 116–128.

Pyle D. (1999) *Data Preparation for Data Mining*, Morgan Kaufmann: San Francisco.

Quinlan J.R. (1992) Learning with continuous classes. *Proceedings of AI'92, 5th Australian Joint Conference on Artificial Intelligence*, Adams A. and Sterling L. (Eds.), World Scientific: Singapore, pp. 343–348.

Shamseldin A.Y. and O'Connor K.M. (1996) A nearest neighbour linear perturbation model for river flow forecasting. *Journal of Hydrology*, **179**, 353–375.

Shamseldin A.Y. and O'Connor K.M. (2001) A nonlinear neural network technique for updating of river flow forecasts. *Hydrology and Earth System Science*, **5**(4), 557–597.

Solomatine D.P. and Dulal K.N. (2003) Model tree as an alternative to neural network in rainfall-runoff modelling. *Hydrological Sciences Journal*, **48**(3), 399–411.

Solomatine D.P. and Maskey M. (2005) Instance-based learning compared to other data-driven methods in hydrologic forecasting. *Journal of Hydrology*, **319**, (submitted).

Solomatine D.P., Rojas C., Velickov S. and Wust H. (2000) Chaos theory in predicting surge water levels in the North Sea. *Proceedings of the 4th International Conference on Hydroinformatics*, Cedar-Rapids.

Solomatine D.P. and Shrestha D.L. (2004) AdaBoost.RT: a boosting algorithm for regression problems. *International Joint Conference on Neural Networks*, Budapest.

Solomatine D.P. and Siek M.B. (2004) Flexible and optimal M5 model trees with applications to flow predictions. *Proceedings of the 6th International Conference on Hydroinformatics*, World Scientific: Singapore.

Solomatine D.P. and Torres L.A. (1996) Neural network approximation of a hydrodynamic model in optimizing reservoir operation. *Proceedings of the 2nd International Conference on Hydroinformatics*, Balkema: Rotterdam, pp. 201–206.

Solomatine D.P. and Xue Y. (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *ASCE Journal Hydrologic Engineering*, **9**(6), 491–501.

Sudheer K.P. and Jain S.K. (2003) Radial basis function neural network for modeling rating curves. *ASCE Journal of Hydrologic Engineering*, **8**(3), 161–164.

Todini E. (1996) The ARNO rainfall–runoff model. *Journal of Hydrology*, **175**, 339–382.

Toth E., Brath A. and Montanari A. (2000) Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, **239**, 132–147.

Vapnik V.N. (1998) *Statistical Learning Theory*, Wiley & Sons: New York.

Velickov S., Solomatine D. and Price R.K. (2003) Prediction of nonlinear dynamical systems based on time series analysis: issues of entropy, complexity and predictability. *Proceedings of the XXX IAHR Congress*, Thessaloniki.

Velickov S., Solomatine D.P., Yu X. and Price R.K. (2000) Application of data mining techniques for remote sensing image analysis. *Proceedings of the 4th International Conference on Hydroinformatics*.

Vernieuwe H., Georgieva O., De Baets B., Pauwels V.R.N., Verhoest N.E.C. and De Troch F.P. (2005) Comparison of data-driven Takagi–Sugeno models of rainfall–discharge dynamics. *Journal of Hydrology*, **302**(1–4), 173–186.

Witten I.H. and Frank E. (2000) *Data Mining*, Morgan Kaufmann: San Francisco.

Xiong L.H., Shamseldin A.Y. and O'Connor K.M. (2001) A nonlinear combination of the forecasts of rainfall–runoff models by the first-order Takagi-Sugeno fuzzy system. *Journal of Hydrology*, **245**(1–4), 196–217.

Zadeh L.A. (1965) Fuzzy sets. *Information and Control*, **8**, 338–353.

20: Artificial Neural Network Concepts in Hydrology

ANTHONY W MINNS[1] AND MICHAEL J HALL[2]

[1]*Marine & Coastal Management, WL Delft Hydraulics, Delft, The Netherlands*
[2]*Department of Water Engineering, UNESCO-IHE, Delft, The Netherlands*

The solution of many applied hydrological problems, such as the forecasting of floods, has for several decades been based upon the concepts of linear systems analysis. However, the introduction of informatics tools, such as Artificial Neural Networks (ANNs), with their origins in cognitive sciences and pattern recognition, has made available new lines of investigation. Nevertheless, despite their apparent structural simplicity, the use of ANNs to encapsulate the transformation of rainfall over a catchment into streamflow at its outlet requires at least as much hydrological insight as a conventional physical/conceptual hydrological model. Of particular importance are: the choice of the input data streams; and the maintenance of the ability of the ANN to generalize and extrapolate beyond its training data set. Attention to these aspects invariably provides a model whose goodness-of-fit to an independent testing data set is superior to that of parameter-based hydrological modeling systems.

INTRODUCTION

Hydrology has always suffered from the dichotomy of being a geophysical science on the one hand, and an applied science for the solution of engineering problems on the other. In its latter guise, hydrology is essentially data dependent, with the optimal use being made of whatever information is available at the time of the investigation. The application of linear systems theory (e.g. Dooge, 1973; Domenico, 1972), developed during the 1950s and 1960s, provides a ready example of an approach that is still widely applied. However, the last decade has seen the gradual introduction of informatics tools, such as artificial neural networks (ANNs), into hydrology, hydrogeology, and water resources planning and management. An ANN is also often referred to as a *data-driven method* (*see* **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**), and, taken at face value, should be similar in utility to linear systems analysis, but with less onerous constraints in application.

Unfortunately, despite the many and various applications of data-driven methods that have appeared over the last decade, a broad appreciation of their potential has been slow to develop. This apparent reluctance to consider such approaches could be partly ascribed to unfamiliarity with the concept of devices such as ANNs, a circumstance that has not been assisted by the somewhat esoteric nomenclature associated with neural networks, stemming from their origins as computational devices for pattern recognition and classification in artificial intelligence and cognitive sciences. However, there is indeed a growing awareness of these techniques as most clearly demonstrated by the very existence of the ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (ASCE-TCANNH, 2000a,b) and the recent publication of several textbooks dedicated solely to this subject (e.g. Abrahart *et al.*, 2004; Govindaraju and Rao, 2000).

The parallel processing architecture and multiple interconnections of processing elements that make up the structure of an ANN have an obvious analogy with biological systems, as explained in any introductory textbook on the subject (e.g. Aleksander and Morton, 1990; Beale and Jackson, 1990; Hertz *et al.*, 1991; Haykin, 1999). The discomfort of the potential user tends to be increased by the many different types of neural networks that are available. However, the emphasis of this review is on *applications*, although inevitably some reference has to be made to the available architectures and calibration methods. In this

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

context, a cursory review of the literature is sufficient to reveal the wide variety of hydrological problems to which ANNs have been applied successfully. The following listing, expanded from Minns and Hall (2004), is intended to be illustrative rather than exhaustive:

- as submodels of complex processes within a larger physically based framework; the estimation of daily solar radiation from daily maximum and minimum air temperatures and precipitation (Elizondo *et al.*, 1994), and the modeling of drying water retention curves for sandy soils by Schaap and Bouten (1996) provide ready examples of this approach;
- as a replacement for, or for modeling the results obtained from, more complex, physically based computer models that impose heavy demands on computing resources (often referred to as *emulating* complex systems); examples include the river salinity forecasting model of Maier and Dandy (1996), and the determination of optimum pumping scenarios for groundwater remediation schemes by Rogers and Dowla (1994), Rogers *et al.* (1995), and Aly and Peralta (1999);
- as models of analytically intractable relationships, such as the approximation for the confidence limits to the quantiles from a flood frequency distribution derived by Whitley and Hromadka (1999);
- as a method of avoiding the constraints associated with standard techniques, such as multiple linear regression analysis, in deriving relationships between the parameters of a regional flood-frequency distribution and catchment characteristics (Muttiah *et al.*, 1997; Hall and Minns, 1998; Hall *et al.*, 2002);
- as a screening model, for example, for the identification of critical realizations of log conductivity fields for a single realization groundwater remediation management model (Ranjithan *et al.*, 1993);
- for the modeling of water table fluctuations (Coulibaly *et al.*, 2001) and the migration of pollutant plumes in groundwater (Hassan and Khaled, 2001);
- as a method of pattern completion, for example, the neural kriging method of Rizzo and Dougherty (1994) and spatial (Lin and Chen, 2004a) and temporal infilling (Khalil *et al.*, 2001);
- as a classification tool for spatial variation and severity of droughts (Shin and Salas, 2000);
- as models of various aspects of sediment transport (Abrahart and White, 2001; Tayfur, 2002);
- as a model of the preferences of decision-makers in multiobjective optimization (see Wen and Lee, 1998, for an example);
- as a forecasting device, such as that for predicting rainfall fields proposed by French *et al.* (1992) and Luk *et al.* (1998, 2000), and for integrating radar images (Hessami *et al.*, 2003) or radiosonde and satellite information (Kim and Barros, 2001) with rainfall gauge measurements;
- for the classification and interpretation of remote sensing data (Islam and Kothari, 2000; Matheussen and Thorolfsson, 2003);
- as an alternative to parameter-intensive physical/conceptual models in applications that do not require a detailed understanding of the system dynamics.

The last category is typified by the problem of modeling the relationship between rainfall and runoff (*see* **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**). However, as the following review is intended to demonstrate, the application of ANNs to the development of a rainfall-runoff model demands as much, if not more, hydrological insight than the calibration of a standard physical/conceptual model (see also Maier and Dandy, 1999).

The details of 18 studies in which ANNs have been employed to develop rainfall-runoff models are presented in Table 1, adapted from Minns and Hall (2004). A similar tabulation has also been provided by Dawson and Wilby (2001). The approach that has been overwhelmingly favored has been the multilayer perceptron (MLP) network with the error back-propagation learning algorithm. Several authors reported favorable results using radial basis function (RBF) networks, largely because the training time of RBF networks was usually significantly less than that for equivalent MLP networks. Furthermore, the RBF networks appear to provide a superior performance over MLPs when dealing with only small numbers of input data sets. However, Dibike *et al.* (1999b) reported that the generalization properties of the RBF networks deteriorated as the number of input data sets was increased and the RBF networks were subsequently out-performed by the MLP networks in this case.

ARTIFICIAL NEURAL NETWORKS

In brief, an ANN consists of layers of processing units (representing biological neurons – see Hopfield, 1994) where each processing unit in each layer is connected to all processing units in the adjacent layers (representing biological synapses and dendrites). The architecture of the various types of ANNs have been well described elsewhere (for example, Beale and Jackson, 1990; Aleksander and Morton, 1990; Hertz *et al.*, 1991; Haykin, 1999). The selection of an appropriate architecture for an ANN depends upon both the problem to be solved and the type of learning algorithm to be applied. For example, the use of Kohonen networks (Kohonen, 1997) for unsupervised classification of patterns, and the use of Hopfield networks for recalling previously learned patterns are two approaches commonly used in pattern recognition.

Table 1 Summary of previous studies in which ANNs have been employed to model rainfall-runoff relationships

References	Time unit	Catchment	Area, km^2	Technique
Abrahart and Kneale (1997)	hour	Wye, UK	11	MLP/BP
Campolo *et al.* (1999)	hour	Tagliamento, Italy	2480	MLP/BP
Carriere *et al.* (1996)	30 s	Laboratory catchment	2.1 m^2	MLP/BP
Dawson and Wilby (1998)	15 min	Amber, UK	139	MLP/BP
	15 min	Mole, UK	142	MLP/BP
Fernando and Jayawardena (1998)	10 min	Kamihonsha, Japan	3.12	RBF/OLS; MLP/BP
Hall and Minns (1993)	5 s	Laboratory catchment	26.8 m^2	MLP/BP
	one min	Doncaster, UK	5.14 ha	MLP/BP
Hsu *et al.* (1995)	day	Leaf River, USA	1949	MLP/LLSSIM
Jayawardena and Fernando (1998)	h	Kamihonsha, Japan	3.12	RBF/OLS; MLP/BP
Lange (1998)	h	Zeller Bach, Germany	20	MLP/BP
	h	Windach, Germany	344.7	MLP/BP
Lorrai and Sechi (1995)	month	Araxisi, Italy	121	MLP/BP
Minns and Hall (1997)	30 min	Dollis Brook, UK	24	MLP/BP
	30 min	Silk Stream, UK	31.25	MLP/BP
Poff *et al.* (1996)	day	Independence, USA	230	MLP/BP
	day	Little Patuxent, USA	97	MLP/BP
See and Openshaw (1998)	h	Ouse, UK	3286	MLP/BP with KN
Shamseldin (1997)	day	Sunkosi, Nepal	18 000	MLP/CG
	day	Shiquan, China	3092	MLP/CG
	day	Yanbain, China	2350	MLP/CG
	day	Bird Creek, USA	2344	MLP/CG
	day	Wolombi Creek, Australia	1580	MLP/CG
	day	Brosna, Ireland	1207	MLP/CG
Teegavarapu (1998)	10 day	Malaprabha, India	?	MLP/BP; RBF/OLS
Thirumalaiah and Deo (2000)	H	Bhatsa Dam, India	391	MLP/BP/CG/CC
	Day	Godavari River, India	313 000	
Tokar and Markus (2000)	Month	Fraser River, USA	458	MLP/BP
	Day	Raccoon Creek, USA	960	
	Day	Little Patuxent, USA	98	
Zealand *et al.* (1999)	quarter-month	Namakan	19 270	MLP/BP

Note: MLP – multilayer perceptron; BP – error back-propagation; RBF – radial basis function; OLS – ordinary least squares; LLSSIM – linear least squares with multistart simplex operation; CG – conjugate gradient; KN – Kohonen network; CC – cascade correlation.

However, when applied more generally to systems identification, the purpose is to train an ANN to provide a correct output response to a given input stimulus. In particular, for rainfall-runoff modeling, the input stimulus corresponds to the measured rainfall and the output response to the measured runoff from a catchment. A multilayer, feed-forward, perceptron-type ANN is one of the most suitable types of ANN for learning the stimulus-response relationship for a given set of measured data. Figure 1 shows a general schematization of a three-layer, feed-forward ANN.

The working of an ANN can best be described with the aid of Figure 1 by following the operations involved during training and computation. An input signal, consisting of an array of numbers x_i is introduced to the input layer of processing units or nodes.

The signals are carried along connections to each of the nodes in the adjacent layer, and can be either amplified or inhibited through weights, w_i, associated with each connection. The nodes in the adjacent layer act as summation devices for the incoming (weighted) signals (see Figure 2).

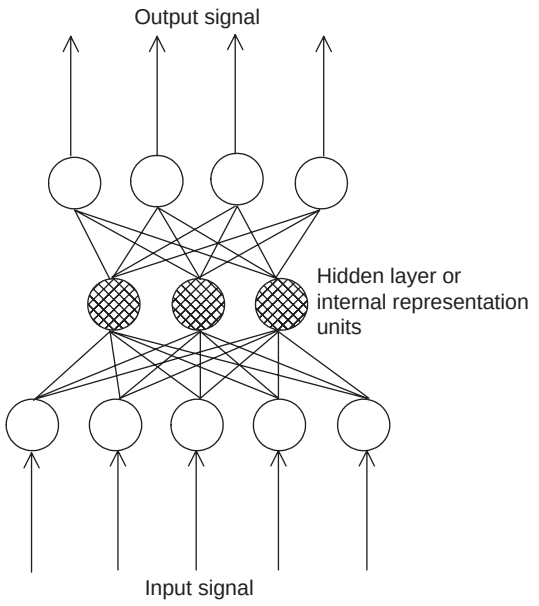


Figure 1 Representation of a multilayer, feed-forward artificial neural network (ANN)

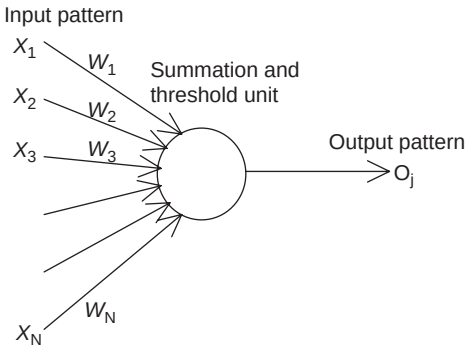


Figure 2 A typical ANN node

The incoming signal is transformed into an output signal, O_j, within the processing units by passing it through a threshold function. A common threshold function for the ANN depicted in Figure 1 is the sigmoid function defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

which provides an output in the range $0 < f(x) < 1$. In most routines, the threshold function usually takes the form of a single-valued, hard-delimiter. The sigmoidal threshold function is chosen for mathematical convenience because it resembles a hard-limiting step-function for extremely large positive and negative values of the incoming signal and also gives useful information about the response of the processing unit to inputs that are close to the threshold value. Furthermore, the sigmoid function has a very simple derivative that makes the subsequent implementation of the learning algorithm much easier.

The output from the processing unit is then:

$$O_j = \frac{1}{1 + e^{-\sum x_i w_i}} \tag{2}$$

This output signal is subsequently carried along the weighted connections to the following layer of nodes and the process is repeated until the signal reaches the output layer. The one or more layers of processing units located between the input and output layers have no direct connection to the outside world and are referred to as *hidden layers*. The output signal can then be interpreted as the response of the ANN to the given input stimuli.

The ANN can be *trained* (i.e. calibrated) to produce known or desired output responses for given input stimuli. The ANN is first initialized by assigning random numbers to the interconnection weights. An input signal is then introduced to the input layer and the resulting output signal is compared to the desired output signal. The interconnection weights are then adjusted to minimize the error between the ANN output and the desired output. This process is repeated many times with many different input/output tuples until a sufficient accuracy for all data sets has been obtained. The adjustment of the interconnection weights during training employs a method known as *error back-propagation* in which the weight associated with each connection is adjusted by an amount proportional to the strength of the signal in the connection and the total measure of the error (see Rumelhart and McClelland, 1986). The total error at the output layer is then reduced by redistributing this error value backwards through the hidden layers until the input layer is reached. The next input/output tuple is then applied and the connection weights readjusted to minimize this new error. In this way, the back-propagation algorithm can be seen to be a form of gradient descent for finding the minimum value of the multidimensional error function. This procedure is repeated until all training data sets have been applied. The whole process is then repeated starting from the first data set once more and continued until the total error for all data sets is sufficiently small and subsequent adjustments to the weights are inconsequential. In many situations, it is recommended to utilize a *cross-validation* data set in addition to the training data set. The error between the network output and the cross-validation data is continuously monitored during the training process. Although the error on the training data may continue to decrease, the error on the cross-validation data may begin to increase. By stopping the training procedure at this time, the problem of *overfitting* can be avoided (*see* **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**).

The ANN is now said to have *learned* a relationship between the input and output training data sets. The exact form of this relationship cannot be extracted from the ANN but, rather, is encapsulated in the stored series of weights and connections between nodes. The absolute values of the individual weights cannot usually be interpreted to have any deeper physical meaning (Minns, 1995), although some success has recently been claimed in studying the pattern of weights (Dibike *et al.*, 1999a; Minns, 2000; Wilby *et al.*, 2003).

Although the error back-propagation method does not guarantee convergence to an optimal solution since local minima may exist, it appears in practice that the back-propagation method leads to solutions in almost every case (Rumelhart *et al.*, 1994). In fact, Hornik *et al.* (1989) concluded that standard multilayer, feed-forward networks are capable of approximating any measurable function to any desired degree of accuracy. They further state that errors in representation appear to arise only from having insufficient hidden units or the relationships themselves being insufficiently deterministic.

For details of other ANN architectures, which include recurrent neural networks (RNN), self-organizing feature maps (SOFM), Hopfield networks, radial basis function

(RBF) networks, support vector machines (SVM), and so on, the reader is referred to Hertz *et al.* (1991), Haykin (1999), and Kecman (2001). Several commercially available software packages provide a wide variety of network architectures and learning algorithms. For Windows-based PCs, the NeuroSolutions package is probably the most advanced and flexible package available. **http://www.neurosolutions.com/products/ns/index.html**

For users familiar with the Matlab working environment, the Matlab Neural Network Toolbox provides an excellent alternative. **http://www.mathworks.com/products/neuralnet/index.html**

The Stuttgart Neural Network Simulator SNNS may be downloaded free-of-charge, but this package is only available for Unix/Linux workstations. **http://www-ra.informatik.uni-tuebingen.de/SNNS/**

RAINFALL-RUNOFF RELATIONSHIPS

The majority of hydrological textbooks (e.g. Bras, 1990; Hornberger *et al.*, 1998) provide succinct descriptions of the physics of hydrograph generation. The forcing function to the catchment system is obviously precipitation in general, and the variations of rainfall intensity over time in particular. However, the relationship between the rainfall intensity and the response of the catchment in terms of changes in discharge at the outlet is primarily dependent on the action of the intervening processes within the hydrological cycle. Overland flow supply and subsurface flow are essentially functions of the soils and vegetative cover, and are, therefore, dependent on the state of wetness of the catchment prior to the rainfall. The contribution of groundwater is a function of geology and the height of the phreatic surface in relation to the channel system (*see* **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4**). The slope, roughness, and geometry of the latter then shape the formation of the outlet hydrograph.

Such influences are well appreciated in a qualitative sense, and models of individual processes, such as interception on a vegetal canopy or soil moisture depletion, are readily available. Their integration into spatially distributed, physically based models of the land phase of the hydrological cycle has been pursued vigorously for almost two decades (*see* **Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**). Perhaps the most widely known of the modern generation of physically based, distributed catchment modeling systems is the Système Hydrologique Européen (SHE), the original structure of which was described by Abbott *et al.* (1986). The structure of a typical SHE model may be visualized as a three-dimensional grid, with the vertical layers running from the deep groundwater through the surface layers and the vegetal cover to the overlying atmospheric boundary layer. As a general guideline, the pixels forming the horizontal grid should be about one per cent of the total catchment area (see Bathurst, 1986). However, in their account of an application of SHE to a river basin of some 4955 square km in India, Refsgaard *et al.* (1992) have acknowledged that their use of 2 km by 2 km grid squares still did not provide a fully physically based and fully distributed description of the basin, even though it was entirely sufficient for the practical application in question. There remained a certain degree of empiricism in the representation of particular hydrological processes, even in these systems, so that process identification and the associated determination of parameter values by direct measurement continues to necessitate the use of extensive calibration procedures.

These and similar reported experiences lead to the conclusion that, for many problems of rainfall-runoff modeling involving, for example, record extension or forecasting, without any significant changes in land use or other such factors and over a certain range and distribution of antecedent soil conditions, simpler models would in most situations be equally accurate and much cheaper to apply. However, in fairness to the distributed, physically based models, their use is primarily directed to the representation of processes other than simple rainfall-runoff. Indeed, from the point of view of practical applications, problems of waste disposal, erosion, changes in vegetation, and so on, have wider societal implications than rainfall-runoff alone.

In contrast, hydrology in general and rainfall-runoff modeling in particular, provides ample opportunities to take advantage of data-driven techniques, such as artificial neural networks. The principal advantage of ANNs is that, even if the precise relationship between input and output data streams is unknown but is acknowledged to exist, the network can be trained to learn that relationship. The use of the data as recorded, that is, the total rainfall volumes instead of the rainfall excess volumes and the recorded discharges instead of the direct runoff rates, is an added incentive to avoid unnecessary empiricism. However, the user must be assured at the outset that the relevant input and output data sets have been selected in the first place.

To date, ANNs have been applied to model the rainfall-runoff relationships of anything from laboratory catchments (Hall and Minns, 1993; Carriere *et al.*, 1996) to drainage areas in excess of 19 000 square km (Zealand *et al.*, 1999) – see Table 1. For the larger sizes of catchment, the use of stage or flow records for sites upstream of the outlet may be possible, so that the ANN is implicitly routing hydrographs as part of the learning process (e.g. See and Openshaw, 1998). Indeed, ANNs have been applied directly for the routing of flood and stage hydrographs by Zhu and Fujita (1994), Raman and Sunilkumar (1995), Thirumalaiah and Deo (1998), and Teegavarapu (1998). Minns (1998, 2000) showed that for these types of simple

advection and dispersion processes, an ANN is capable of encapsulating the same knowledge that is contained in the governing partial differential equations. In fact, the governing continuum equations could actually be restored for the particular choice of ANN configuration by analyzing the weights of the ANNs that had been trained with measured data.

The following discussion, however, is more specifically concerned with the art of rainfall-runoff modeling, which, despite being a fertile area for exploration, offers several less-than-obvious traps for the unwary. The first choice to be made by the ANN modeler is the mode of presentation of the data to the network, that is, how are the input and output patterns to be defined? One possibility is to take the n successive ordinates of the rainfall hyetograph and feed these into the n input nodes of a network whose m output nodes carry the m successive ordinates of the flow hydrograph. This was the approach followed by both Smith and Eli (1995) and Lange (1998), but in the former case, the outputs were the coefficients from a truncated harmonic series representation of the hydrograph, which had the added advantage of already being standardized within the interval ± 1.

An alternative method of defining patterns is the so-called dynamic approach in which the input is a set of concurrent ordinates of (say) the rainfall totals from p rain gauges within the catchment and the output is the concurrent rate of outflow. In this mode, each time step defines a pattern, and, therefore, a single storm event provides as many exemplars as there are runoff ordinates, rather than only a single input-output pairing. This is the approach that has been adopted by the majority of writers on rainfall-runoff modeling using ANNs, but requires a much higher level of hydrological insight into the working of the catchment system.

Within the dynamic approach, Minns and Hall (2004) identified three types of model:

- naïve dynamic model

$$Q(t) = f(r(t), r(t-1), \ldots) \tag{3}$$

- rainfall-runoff simulation model

$$Q(t) = f(r(t), r(t-1), \ldots, Q(t-1), Q(t-2), \ldots) \tag{4}$$

- "auto-regressive" model

$$Q(t) = f(Q(t-1), Q(t-2), \ldots) \tag{5}$$

where $Q(t)$ is the outflow at time level t and $r(t)$ is the rainfall ordinate at time level t.

The simplest, naïve dynamic rainfall-runoff model (equation 3) would consist of an ANN with inputs from one or more rain gauges at time t, and an output of concurrent flow. A simple scatter plot of input(s) against output is sufficient in this case to indicate that the description of the input pattern is inadequate. Some improvement is obtainable by allowing for the time lag between the occurrence of the flow and the incidence of the causative rainfall. Since the flow at any instant is effectively composed of contributions from different subareas whose time of travel to the outlet covers a range of values, both the concurrent and antecedent rainfalls can be considered to be contributing to the outflow. Use of a moving window of rainfall at time t and the k previous intervals provides some improvement (e.g. Hall and Minns, 1993; Karunanithi *et al.*, 1994).

The choice of length for the moving window of rainfall can significantly affect the accuracy of the resulting ANN model. If the window is too short, the input data does not contain enough information about the entire rain event that is contributing to the concurrent outflow. Output hydrographs tend to represent only the shape of the rainfall bursts, and there is a very poor representation of the rising limbs and the recession limbs. Conversely, if the window is too long, the input contains too much information, which may even include historical rainfall events whose effects have long since passed out of the catchment, and so are no longer contributing to the concurrent flow. In this case, an ANN can no longer generalize the relationship between rainfall and runoff. There is simply too much data being presented at the input layer and – to maintain the biological analogy – the ANN becomes "confused". Hall and Minns (1993) showed that the most accurate results are obtained using a moving window length that broadly encompasses the range of centroid-to-centroid lag times of the training data.

A further problem with the naïve model is that the simplistic input patterns may easily result in ambiguous results. More specifically, intervals with zero rainfall inputs are encountered in two contrasting situations: immediately prior to the beginning of the storm at the start of the rising limb; and some time after the end of the storm event when flows are moderately high and in recession. The ANN has no information to discriminate between these two "no-rainfall" conditions and once more becomes "confused". These conditions have to be differentiated by the addition of another input if the ANN is to achieve the correct mapping. The most obvious candidate is a flow ordinate, which is most easily provided by the output at time t to become an input at time $t+1$ (Hall and Minns, 1993; Minns and Hall, 1996; and similar use of stage outputs by Campolo *et al.*, 1999). This approach is described above as model (equation 4). In effect, the flow (or stage) ordinate is employed as a crude measure of catchment wetness.

Model (equation 5) is then the logical extension of models (equation 3) and (equation 4) into purely "auto-regressive" time series prediction. This model does not make use of any rainfall data at all but uses only antecedent

outflow values as input to the ANN to predict the concurrent outflow. Minns (1998) demonstrated that model (equation 5) suffers from a persistent phase error. In this case, the ANN has no information available that tells it at which level the rising limb should stop until the actual measurements indicate that this is so. That is, at the top of the rising limb, the output from the ANN wishes to continue rising in magnitude based only upon the pattern of the preceding flows. It is not until several time steps have passed for which the measured values are all constant that the ANN "recognizes" that the equilibrium level has been reached. Similarly, the phase error that occurs at the beginning of the recession limb is caused by the ANN having no knowledge about the cessation of the rainfall until one or two time steps after the actual measured values have started to decrease.

In general, the plethora of literature involving the application of ANNs to rainfall-runoff modeling confirms the exceptional accuracy of ANN models for short forecasting intervals. Longer forecasting intervals may be obtained by utilizing $Q(t+1), Q(t+2), \ldots$ and so on, as outputs during the training of the ANN. Unfortunately, the performance of the ANN decreases quite rapidly with an increasing prediction time horizon (Campolo *et al.*, 1999; Zealand *et al.*, 1999). Another approach is to use a trained ANN with a "feedback" loop in which the predicted output is used directly as input data for the subsequent time step. Unfortunately, the error accumulation associated with this approach also means that the performance of the ANN deteriorates quite rapidly after only one or two iterations (van den Boogaard *et al.*, 1998). Although Abrahart (1998) describes a method to deal with the accumulated error, the most promising approach would appear to involve the use of (partial) recurrent neural networks, which contain feedback loops in both the training and recall modes of the network. Hertz *et al.* (1991) describe the architecture of these so-called Jordan/Elman Recurrent Networks. Proaño *et al.* (1998), van den Boogaard *et al.* (1998), and Chiang *et al.* (2004) show significant improvements in long-term predictions using recurrent neural networks. In addition, Varoonchotikul *et al.* (2002b) and Varoonchotikul (2003) have demonstrated that a Jordan Recurrent Network can be trained more quickly and can provide an improvement in performance over a standard MLP-type ANN. van den Boogaard (2004) describes the theoretical background of partial recurrent neural networks and provides two applications that demonstrate the practical relevance of these tools for modeling dynamic hydrological systems.

In terms of the number of patterns that can be extracted from a given data set, the above-mentioned dynamic models are superior to those based simply on the use of the whole hyetograph as the input and the complete hydrograph as the output. However, the problem then arises as to the set of time points for which those patterns are extracted. Here, a clear perspective is required as to the purpose of the modeling, since with any series of discharges, the (positive) skewness of its marginal distribution tends to increase as the time interval at which the data are recorded reduces. This effect is manifested in the appearance of sustained recessions to the hydrographs as the time interval becomes shorter, such that with (say) daily data, they become the dominant features of the time series. In these circumstances, the rising limbs and peaks of the storm hydrographs form only a small portion of the total number of patterns in the time series, and the mapping of inputs to outputs is biased in favor of the recession behavior. If, therefore, the purpose of the modeling is to capture the essence of the flood regime of the catchment, then the inputs should be restricted to the major storm events. For convenience, these inputs and outputs may be arranged in the form of an artificial time series in which the flow transitions between successive events are smoothed to provide continuity. In the absence of seasonal influences, this approach has been found to work satisfactorily (Hall and Minns, 1993; Minns and Hall, 1996, 1997; Minns, 1996; Campolo *et al.*, 1999). If the full range of flow behavior is of interest, then an alternative approach might be to carry out *a priori* classification of event types or hydrograph features – perhaps employing a Kohonen network – and to develop a separate ANN rainfall-runoff model for each class (e.g. See and Openshaw, 1998). Hsu *et al.* (2002) have also proposed a novel structure of ANN incorporating an SOFM to classify the input vectors.

Dawson and Wilby (1998) concluded that ANNs for long flow series at short time intervals should ideally be calibrated and validated on data for a common period of the year. Seasonal influences can, of course, be incorporated by extending the list of input variables. For example, Abrahart and Kneale (1997) and Abrahart (1998) employed an annual hour count converted into its sine and cosine equivalents to denote "time of year". Alternatively, Zealand *et al.* (1999) added a "period of the year" (in effect, a week number) and cumulative precipitation since the previous 1st November to the current time period, up to 1st April, to their inputs, the latter being intended as a measure of winter snowpack accumulation. Yet another approach to incorporating a seasonal variable is to use temperature data as an additional input (Lorrai and Sechi, 1995; Poff *et al.*, 1996; Zealand *et al.*, 1999).

Despite the potential significance of seasonal influences, the majority of ANN rainfall-runoff models have tended to rely on combinations of current and antecedent rainfall totals and antecedent flow (or stage) ordinates as inputs (see, e.g. Hsu *et al.*, 1995; Minns and Hall, 1996, 1997; Jayawardena and Fernando, 1998; Fernando and Jayawardena, 1998; Campolo *et al.*, 1999). Gautam *et al.* (2000) included soil moisture data as inputs and Bin Zhang and Govindaraju (2003) incorporated knowledge of catchment geomorphology in the construction of their ANNs. More

elaborate inputs derived from the basic records have been introduced in some studies, such as the derivative of the rainfall intensity and the integral of the rainfall intensity over the previous five time steps employed by Mason *et al.* (1996). Shamseldin (1997) defined a series of rainfall indices consisting of weighted sums of previous rainfall ordinates, the weights being derived from the ordinates of a gamma distribution. Burian and Durrans (2002) have also applied ANNs to disaggregate hourly rainfall records into subhourly increments in order to improve runoff hydrograph predictions.

The predominant objective of the rainfall-runoff models that have been developed using some form of ANN has been that of forecasting future flows given the knowledge of past flows, rainfalls, and other relevant variables. Interest in this topic has continued unabated (e.g. Hu *et al.*, 2001; CiGizoGlu, 2003; Chibanga *et al.*, 2003; Campolo *et al.*, 2003; Rajurkar *et al.*, 2002, 2004; Lin and Chen, 2004b). Comparisons between ANNs and other approaches are not always completely favorable to the former (e.g. Sivakumar *et al.*, 2002; Xiong and O'Connor, 2002; Laio *et al.*, 2003) largely because the basic MLP-type of ANN is not ideally suited to this application. The sigmoidal activation function adopted by many authors imposes a scaling on the network output such that the network is incapable of predicting a flow larger in magnitude than the one contained within the training data set. This effect is amply demonstrated by Minns and Hall (1996) based upon trials with synthetic data. Consequently, if an ANN were to be applied to a real catchment, even if the training data included all the available measurements, there remains always a small but nonnegligible probability that an extreme event beyond the range of recorded experience might occur in future.

An alternative approach suggested by Minns (1996) and Minns and Hall (1997) is to use the *change* in flow as the output rather than the absolute magnitude of the discharge. This variable was used independently by Zhu and Fujita (1994) for forecasting purposes, but without explanation. Change in stage was adopted as the forecast variable by See and Openshaw (1998), presumably because of the scaling problem outlined above. However, Karunanithi *et al.* (1994) claim that a (clipped) linear activation function allows extrapolation. This property arises due to the unbounded nature of the linear activation output. However, if this type of ANN is applied with little or no hydrological insight, this apparent luxury of "unlimited" extrapolation may lead to quite unacceptable linear extrapolations of some very nonlinear hydrological processes. The results may then be not only meaningless, but also quite dangerous to apply.

A further alternative suggested by Varoonchotikul *et al.* (2002a) and Varoonchotikul (2003) involves a careful selection of the standardization range of the activation function based upon the maximum and minimum contained in the raw data set. In order to provide "room" for extrapolation, an arbitrary factor is applied to the maximum. A value of two was found to provide a reasonably satisfactory compromise, allowing extrapolation up to peak discharges with return periods of between 33 and 100 years, depending upon the rapidity of catchment response. In contrast, Imrie *et al.* (2000) proposed the addition of a *guidance system* to the output layer of the cascade correlation architecture of ANN to overcome the extrapolation problem. Unfortunately, this approach depends upon the use of information from the testing data set, thereby disrupting the training and verification cycle.

Another possibility, proposed by Hettiarachchi *et al.* (2005), is the incorporation of independent domain knowledge into the modeling process. Such knowledge can be developed from the standardized procedures for estimating design flood hydrographs that are available in many countries. Since such procedures are invariably based on regionalization exercises on data from many catchment areas, they incorporate information beyond that of a single river basin. Such procedures can be employed to develop (say) the 100-year design flood hydrograph and its associated rainfall hyetograph, the ordinates of which can then be added to the training data set. Hettiarachchi *et al.* (2005) demonstrate that an Estimated Maximum Flood, which provides an approximation to the physical upper limit of runoff of which the catchment is capable, can be included in training without affecting the ability of the network to generalize. The authors also paid attention to the method of standardization of the input and output data, including the use of logarithmic transformations. Bowden *et al.* (2003) have also examined the use of different standardization procedures for the input and output data, but found little improvement on the standard linear transformation approach.

CONCLUDING REMARKS

The results of all of the numerical experiments reported to date indicate that suitably configured ANNs are capable of identifying usable relationships between runoff discharges and antecedent rainfall depths to an exceptional degree of accuracy. The relationships are obtained using only the raw, measured data and do not require the use of any derived or artificial calibration parameters. In particular, the ANN model provides these exceptional results unhindered by constraints of volume continuity in the input and output data and, in fact, the units of the data are chosen simply for convenience of measurement and representation (e.g. rainfall depths in mm, discharges in m^3/s). Furthermore, simple, nonhydrological parameters such as the percentage of impervious area may be easily incorporated into the model at the discretion of the modeler. These types of parameters may be derived from simple measurements or may even be highly intuitive, and are likewise unrestricted in

terms of conditions of dimension or hydrological-physical consistency. Clearly, the understanding of the underlying physical processes is enhanced if dimensionally consistent parameters are used (*see* **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1**). In general, applications of evolutionary computing in hydrological sciences benefit greatly if available knowledge and understanding of the underlying physical processes is incorporated before introducing *artificial* intelligence methods (*see* **Chapter 22, Evolutionary Computing in Hydrological Sciences, Volume 1**).

Despite the expanding literature on neural network modeling, certain errors of procedure remain all too common. Once a suitable combination of input and output parameters has been chosen – in itself not necessarily a simple task, since this already requires some insight into the mutual interactions of not yet fully understood processes – attention needs to be turned to the choice of the network architecture. Assuming that an MLP-type of ANN is being applied, the key features are the number of hidden layers and the number of processing elements in each. Previous experience has indicated that little advantage accrues from having more than one hidden layer in hydrological applications (e.g. Minns and Hall, 1996). The number of nodes in the hidden layer may then be determined by trial and error, but care is necessary to ensure that the number of training exemplars is well in excess of the number of processing elements, perhaps by a factor of 4 or 5 (see, e.g. Walczak and Cerpa, 1999). With large numbers of processing elements, the danger exists of each becoming associated with a subset of input tuples. In these circumstances, the ability of the trained ANN to generalize is compromised, and its performance on an independent verification data set falls well below that of the training data set.

A similar problem can arise if the training is continued for too long. The network then "learns" both signal and noise, and the generalization properties of the network are again impaired. However, this effect may be controlled by choosing a third set of data for *cross-validation*. Training is stopped at arbitrary intervals, and the fitting error of the cross-validation set is examined. If the error in both the training and the cross-validation data sets decreases, then the training is continued. However, once the error in the cross-validation data set increases, training is stopped. Coulibaly *et al.* (2000) provides a detailed description of this problem and Smith (1993) suggests procedures for overcoming it (*see* **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**). The optimal division of the available records into the three data subsets has been investigated by Bowden *et al.* (2002) using both SOFMs and Genetic Algorithms. An alternative approach involving bootstrapping of the data was proposed by Abrahart (2003).

The discussion above has demonstrated that, although an ANN may be regarded as some form of "black-box" model (Minns and Hall, 1996), the potential user is not absolved from devoting some thought to the mode of presentation of data to the network. The principal question to be posed is: what exactly constitutes the pattern of inputs that produces the pattern of outputs? Moreover, do the selected input and output patterns contain additional information that is not strictly relevant to the purpose of the exercise? These questions are inevitably problem-dependent, but in all cases, the selection of inputs, whether data as recorded or variables derived from operations on recorded data, requires the application of hydrological insight as much as any conventional physical/conceptual rainfall-runoff model. Provided that such insight is applied, the performance of ANN models on independent testing data sets can be undoubtedly superior to conventional hydrological models in situations that do not require more detailed knowledge of the hydrological system.

FURTHER READING

There are many excellent web-based resources that may provide additional information and software for neural network practitioners. A good starting point is the neural network FAQ located at:

ftp://ftp.sas.com/pub/neural/FAQ.html

An international overview of research groups involved in neural networks and associated sciences (including links to relevant homepages) can be found at:

http://www.eleceng.adelaide.edu.au/Groups/PCON/ngroups.html

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986) An introduction to the European hydrological system – système hydrologique Européen, "SHE", 2: structure of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

Abrahart R.J. (1998) Neural networks and the problem of accumulated error: an embedded solution that offers new opportunities for modelling and testing. *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotterdam, pp. 725–731.

Abrahart R.J. (2003) Neural network rainfall-runoff forecasting based on continuous resampling. *Journal of Hydroinformatics*, **5**(1), 51–61.

Abrahart R.J. and Kneale P.E. (1997) Exploring neural network rainfall-runoff modelling. In: *Proceedings of the 6th National Hydrological Symposium (Salford, UK)*, British Hydrological Society: London, pp. 9–35–9–44.

Abrahart R.J., Kneale P.E. and See L.M. (Eds.) (2004) *Neural Networks for Hydrological Modelling*, Balkema: Leiden, p. 306.

Abrahart R.J. and White S. (2001) Modelling sediment transfer in Malawi: comparing back-propagation neural network solutions against a multiple linear regression Benchmark using small data sets. *Physics and Chemistry of the Earth (B)*, **26**, 19–24.

Aleksander I. and Morton H. (1990) *An Introduction to Neural Computing*, Chapman & Hall: London.

Aly A.H. and Peralta R.C. (1999) Optimal design of aquifer cleanup systems under uncertainty using a neural network and a genetic algorithm. *Water Resources Research*, **35**(8), 2523–2532.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000a) Artificial neural networks in hydrology. I: preliminary concepts, *Journal of Hydrologic Engineering*, **5**(2), 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000b) Artificial neural networks in hydrology. II: hydrologic applications, *Journal of Hydrologic Engineering*, **5**(2), 124–137.

Bathurst J.C. (1986) Physically-based distributed modelling of an upland catchment using the Système Hydrologique Européen. *Journal of Hydrology*, **87**, 79–102.

Beale R. and Jackson T. (1990) *Neural Computing: An Introduction*, Institute of Physics: Bristol.

Zhang B. and Govindaraju R.S. (2003) Geomorphology-based artificial neural networks (GANNs) for estimation of direct runoff over watersheds. *Journal of Hydrology*, **273**, 18–34.

van den Boogaard H.F.P. (2004) The use of partial recurrent neural networks for auto-regressive modelling of dynamic hydrological systems. In *Neural Networks for Hydrological Modelling*, Chap. 7, Abrahart R.J., See L. and Kneale P.E. (Eds.), Balkema: Leiden, pp. 115–138.

van den Boogaard H.F.P., Gautam D.K. and Mynett A.E. (1998) Auto-regressive neural networks for the model-ling of time series. In *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotterdam, pp. 741–748.

Bowden G.J., Dandy G.C. and Maier H.R. (2003) Data transformation for neural network models in water resources applications. *Journal of Hydroinformatics*, **5**(4), pp. 245–258.

Bowden G.J., Maier H.R. and Dandy G.C. (2002) Optimal division of data for neural network models in water resources applications. *Water Resources Research*, **38**(2), 2.1–2.11, doi: 10.1029/2001WR000266.

Bras R.L. (1990) *Hydrology. An Introduction to Hydrologic Science*, Addison-Wesley: Reading.

Burian S.J. and Durrans S.R. (2002) Evaluation of an artificial neural network rainfall disaggregation model. *Water Science and Technology*, **45**(2), 99–104.

Campolo M., Andreussi P. and Soldati A. (1999) River flood forecasting with a neural network. *Water Resources Research*, **35**(4), 1191–1197.

Campolo M., Soldati A. and Andreussi P. (2003) Artificial neural network approach to flood forecasting in the River Arno. *Hydrological Sciences Journal*, **48**(3), 381–398.

Carriere P., Mohaghegh S. and Gaskari R. (1996) Performance of a virtual runoff hydrograph system. *Journal of Water Resources Planning and Management*, **122**(6), 421–427.

Chiang Y.-M., Chang L.-C. and Chang F.-J. (2004) Comparison of static-feedforward and dynamic-feedback neural networks for rainfall-runoff modelling. *Journal of Hydrology*, **290**, 297–311.

Chibanga R., Berlamont J. and Vandewalle J. (2003) Modelling and forecasting of hydrological variables using artificial neural networks:the Kafue River sub-basin. *Hydrological Sciences Journal*, **48**(3), 363–369.

CiGizoGlu H.K. (2003) Estimation, forecasting and extrapolation of river flows by artificial neural networks. *Hydrological Sciences Journal*, **48**(3), 349–361.

Coulibaly P., Anctil F., Aravena R. and Bobée B. (2001) Artificial neural network modelling of water table depth fluctuations. *Water Resources Research*, **37**(4), 885–896.

Coulibaly P., Anctil F. and Bobée B. (2000) Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology*, **230**, 244–257.

Dawson C.W. and Wilby R. (1998) An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, **43**, 47–66.

Dawson C.W. and Wilby R. (2001) Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, **25**, 80–108.

Dibike Y., Minns A.W. and Abbott M.B. (1999a) Applications of artificial neural networks to the generation of wave equations from hydraulic data. *Journal of Hydraulic Research, IAHR*, **37**(1), 81–97.

Dibike Y.B., Solomatine D. and Abbott M.B. (1999b) On the encapsulation of numerical-hydraulic models in artificial neural networks. *Journal of Hydraulic Research*, **37**(2), 147–161.

Domenico P.A. (1972) *Concepts and models in groundwater hydrology*, McGraw-Hill Book: New York, pp. 405. pp.

Dooge J.C.I. (1973) *Linear Theory of Hydrologic Systems*, US Department Agricultural Research Service Technical Bulletin 1468, 327.

Elizondo D., Hoogenboom G. and McClendon R.W. (1994) Development of a neural network model to predict daily solar radiation. *Agricultural and Forest Meteorology*, **71**, 115–132.

Fernando D.A.K. and Jayawardena A.W. (1998) Runoff forecasting using RBF networks with OLS algorithm. *Journal of Hydrologic Engineering*, **3**(3), 203–209.

French M.N., Krajewski W.F. and Cuykendall R.R. (1992) Rainfall forecasting in space and time using a neu-ral network. *Journal of Hydrology*, **137**, 1–31.

Gautam M.R., Watanabe K. and Saegusa H. (2000) Runoff analysis in humid forest catchment with artificial neural network. *Journal of Hydrology*, **235**, 117–136.

Govindaraju R.S. and Rao A.R. (Eds.) (2000) *Artificial Neural Networks in Hydrology*, Kluwer Academic: Dordrecht, p. 329.

Hall M.J. and Minns A.W. (1993) Rainfall-runoff modelling as a problem in artificial intelligence: experience with a neural network. *Proceedings of the 4th National Hydrological Symposium (Cardiff, UK)*, British Hydrological Society: London, pp. 5.51–5.57.

Hall M.J. and Minns A.W. (1998) Regional flood frequency analysis using artificial neural networks. *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotterdam, pp. 759–763.

Hall M.J., Minns A.W. and Ashrafuzzaman A.K.M. (2002) The application of data mining techniques for the regionalization of hydrological variables. *Hydrology and Earth System Sciences*, **6**(4), 685–694.

Hassan A.E. and Khaled H.H. (2001) Prediction of plume migration in heterogeneous media using artificial neural networks. *Water Resources Research*, **37**(3), 605–623.

Haykin S. (1999) *Neural Networks. A comprehensive foundation*, *Second Edition*, Prentice-Hall International: Upper Saddle River, pp. 842.

Hertz J., Krogh A. and Palmer R.G. (1991) *Introduction to the Theory of Neural Computation*, Addison-Wesley: Redwood City.

Hessami M., Anctil F. and Viau A.A. (2003) An adaptive neuro-fuzzy inference system for the post-calibration of weather radar rainfall estimation. *J. Hydroinformatics*, **5**(1), 63–70.

Hettiarachchi P., Hall M.J. and Minns A.W. (2005) The extrapolation of artificial neural networks for the modelling of rainfall-runoff relationships. *Journal of Hydroinformatics*, **7**.

Hopfield J.J. (1994) Neurons, dynamics and computation. *Physics Today*, **47**(2), 40–46.

Hornberger G., Raffensperger J., Wiberg P. and Eshelman K. (1998) *Elements of physical hydrology*, Johns Hopkins Univ. Press: Baltimore, pp. 314.

Hornik K., Stinchcombe M. and White H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Hsu K., Gupta H.V., Gao X., Sorooshian S. and Imam B. (2002) Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modelling and analysis. *Water Resources Research*, **38**(12), 1302, doi: 10.1029/2001WR00795.

Hsu K., Gupta H.V. and Sorooshian S. (1995) Artificial neural network modelling of the rainfall-runoff process. *Water Resources Research*, **31**(10), 2517–2530.

Hu T.S., Lam K.C. and Ng S.T. (2001) River flow time series prediction with a range- dependent neural network. *Hydrological Sciences Journal*, **46**(3), 729–745.

Imrie C.E., Durucan S. and Korre A. (2000) River flow prediction using artificial neural networks: generalization beyond the calibration range. *Journal of Hydrology*, **233**, 138–153.

Islam S. and Kothari R. (2000) Artificial neural networks in remote sensing of hydrologic processes. *Journal of Hydrologic Engineering*, **5**(2), 138–144.

Jayawardena A.W. and Fernando D.A.K. (1998) Use of radial basis function type artificial neural networks for runoff simulation. *Computer-aided Civil and Infrastructure Engineering*, **13**, 91–99.

Karunanithi N., Grenney W.J., Whitley D. and Bovee K. (1994) Neural networks for river flow prediction. *Journal of Computing in Civil Engineering*, **8**(2), 201–220.

Kecman V. (2001) *Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models*, MIT Press: Cambridge, p. 541.

Khalil M., Panu U.S. and Lennox W.C. (2001) Groups and neural networks based streamflow data infilling procedures. *Journal of Hydrology*, **241**, 153–176.

Kim G. and Barros A.P. (2001) Quantitative flood forecasting using multisensor data and neural networks. *Journal of Hydrology*, **246**, 45–62.

Kohonen T. (1997) *Self-Organizing Maps*, *Second Edition*, Springer-Verlag, Berlin, p. 362.

Lange N.T.G. (1998) Advantages of unit hydrograph derivation by neural networks. *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotterdam, pp. 783–789.

Laio F., Porporato A., Revelli R. and Ridolfi L. (2003) A comparison of non-linear flood forecasting methods. *Water Resources Research*, **39**(5), 1129, doi: 10.1029/2002WR001551.

Xiong L. and O'Connor K.M. (2002) Comparison of four updating models for real-time river flow forecasting. *Hydrological Sciences Journal*, **47**(4), 621–639.

Lin G.-F. and Chen L.-H. (2004a) A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. *Journal of Hydrology*, **288**, 288–298.

Lin G.-F. and Chen L.-H. (2004b) A nonlinear rainfall-runoff model using radial basis function network. *Journal of Hydrology*, **289**, 1–8.

Lorrai M. and Sechi G.M. (1995) Neural nets for modelling rainfall-runoff transformations. *Water Resources Management*, **9**, 299–313.

Luk K.C., Ball J.E. and Sharma A. (1998) Rainfall forecasting through artificial neural networks. *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotterdam, pp. 797–804.

Luk K.C., Ball J.E. and Sharma A. (2000) A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *Journal of Hydrology*, **227**, 56–65.

Maier H.R. and Dandy G.C. (1996) The use of artificial networks for the prediction of water quality parameters. *Water Resources Research*, **32**(4), 1013–1022.

Maier H.R. and Dandy G.C. (1999) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Journal Environmental Modelling and Software*, **15**(1), 101–123.

Mason J.C., Price R.K. and Tem'me A. (1996) A neural network model of rainfall-runoff using the radial basis function. *Journal of Hydraulic Research*, **34**(4), 537–548.

Matheussen B.V. and Thorolfsson S.T. (2003) Estimation of snow covered area for an urban catchment using image processing and neural networks. *Water Science and Technology*, **48**(9), 155–164.

Minns A.W. (1995) Analysis of experimental data using artificial neural networks. *Hydra 2000, Proceedings of XXVI Congress IAHR*, Vol 1, Thomas Telford Services: London, pp. 218–223.

Minns A.W. (1996) Extended rainfall-runoff modelling using artificial neural networks. *Proceedings of the Hydroinformatics '96, 2nd International Conference On Hydroinformatics (Zurich, Switzerland)*, Balkema: Rotterdam, pp. 207–213.

Minns A.W. (1998) *Artificial Neural Networks as Subsymbolic Process Descriptors*, Balkema: Rotterdam.

Minns A.W. (2000) Subsymbolic methods for data mining in hydraulic engineering. *Journal of Hydroinformatics*, **2**(1), 3–13.

Minns A.W. and Hall M.J. (1996) Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal*, **41**, 399–417.

Minns A.W. and Hall M.J. (1997) Living with the ultimate black box: more on artificial neural networks. *Proceedings of the 6th National Hydrological Symposium (Salford, UK)*, British Hydrological Society: London, pp. 9.45–9.49.

Minns A.W. and Hall M.J. (2004) Rainfall-runoff modelling. In *Neural Networks for Hydrological Modelling*, Chap. 9, Abrahart R.J., See L. and Kneale P.E. (Eds.), Balkema: Leiden, pp. 157–175.

Muttiah R.S., Srinivasan R. and Allen P.M. (1997) Prediction of two-year peak stream discharges using neural networks. *Journal of the American Water Resources Association*, **33**, 625–630.

Poff N.L., Tokar S. and Johnson P. (1996) Stream hydrological and ecological responses to climate change assessed with an artificial neural network. *Limnology and Oceanography*, **41**(5), 857–863.

Proaño C.O., Minns A.W., Verwey A. and van den Boogaard H.F.P. (1998) Emulation of a sewerage system computational model for the statistical processing of large numbers of simulations. In *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotter-dam, pp. 1145–1152.

Rajurkar M.P., Kothyari U.C. and Chaube U.C. (2002) Artificial neural networks for daily rainfall-runoff modelling. *Hydrological Sciences Journal*, **47**(6), 865–877.

Rajurkar M.P., Kothyari U.C. and Chaube U.C. (2004) Modelling of the daily rainfall-runoff relationship with artificial neural network. *Journal of Hydrology*, **285**, 96–113.

Raman H. and Sunilkumar N. (1995) Multivariate modelling of water resources time series using artificial neural networks. *Hydrological Sciences Journal*, **40**, 145–163.

Ranjithan S., Eheart J.W. and Garrett J.H. (1993) Neural network-based screening for groundwater reclamation under uncertainty. *Water Resources Research*, **29**(3), 563–574.

Refsgaard J.C., Seth S.M., Bathurst J.C., Erlich M., Storm B., Jørgensen G.H. and Chandra S. (1992) Application of SHE to catchments in India, Part 1. General results. *Journal of Hydrology*, **140**, 1–23.

Rizzo D.M. and Dougherty D.E. (1994) Characterization of aquifer properties using artificial neural networks: neural kriging. *Water Resources Research*, **30**(2), 483–497.

Rogers L.L. and Dowla F.U. (1994) Optimization of groundwater remediation using artificial neural networks with parallel solute transport modelling. *Water Resources Research*, **30**(2), 457–481.

Rogers L.L., Dowla F.U. and Johnson V.M. (1995) Optimal field-scale groundwater remediation using artificial neural networks and the genetic algorithm. *Environmental Science and Technology*, **29**, 1145–1155.

Rumelhart D.E., McClelland J.L. and the PDP Research Group (1986) *Parallel Distributed Processing – Explorations in the Microstructure of Cognition*, Vol. 2, MIT Press: Cambridge, p. 547, (Vol. 1) p. 611, (Vol. 2).

Rumelhart D.E., Widrow B. and Lehr M.A. (1994) The basic ideas in neural networks. *Communications of the Acm*, **37**(3), 87–92.

Schaap M.G. and Bouten W. (1996) Modelling water retention curves of sandy soils using neural networks. *Water Resources Research*, **32**(10), 3033–3040.

See L. and Openshaw S. (1998) Using soft computing techniques to enhance flood forecasting on the River Ouse. *Proceedings of the Hydroinformatics '98, 3rd International Conference On Hydroinformatics (Copenhagen, Denmark)*, Vol. 2, Balkema: Rotterdam, pp. 725–731.

Shamseldin A.Y. (1997) Application of a neural network technique to rainfall-runoff modelling. *Journal of Hydrology*, **199**, 272–294.

Shin H.-S. and Salas J.D. (2000) Regional drought analysis based on neural networks. *J. Hydrologic Engineering*, **5**(2), 145–155.

Sivakumar B., Jayawardena A.W. and Fernando T.M.K.G. (2002) River flow forecasting: use of phase space reconstruction and artificial neural networks approaches. *Journal of Hydrology*, **265**, 225–245.

Smith M. (1993) *Neural Networks for Statistical Modelling*, Van Nostrand Reinhold, New York.

Smith J. and Eli R.N. (1995) Neural-network models of rainfall-runoff process. *Journal of Water Resources Planning and Management*, **121**(6), 499–508.

Tayfur G. (2002) Artificial neural networks for sheet sediment transport. *Hydrological Sciences Journal*, **47**(6), 879–892.

Teegavarapu R. (1998) Input structures for a neural network model used for streamflow forecasting. *Hydrology in a Changing Environment*, Vol. III, Proceedings of the British Hydrological Society International Conference (Exeter, UK), Wiley: Chichester, pp. 105–114.

Thirumalaiah K. and Deo M.C. (1998) River stage forecasting using artificial neural networks. *Journal of Hydrologic Engineering*, **3**(1), 26–32.

Thirumalaiah K. and Deo M.C. (2000) Hydrological forecasting using neural networks. *Journal of Hydrologic Engineering*, **5**(2), 180–189.

Tokar A.S. and Markus M. (2000) Precipitation-runoff modelling using artificial neural networks and conceptual models. *Journal of Hydrologic Engineering*, **5**(2), 156–161.

Varoonchotikul P. (2003) *Flood forecasting using artificial neural networks*, Swets and Zietlinger: Lisse, p. 102.

Varoonchotikul P., Hall M.J. and Minns A.W. (2002a) Extrapolation management for artificial neural network models of rainfall-runoff relationships. In *Hydroinformatics 2002, Vol. 1: Model Development and Data Management*, Falconer R.A., Lin B., Harris E.L. and Wilson C.A.M.E. (Eds.), IWA Publishing: London, pp. 673–678.

Varoonchotikul P., Hall M.J. and Minns A.W. (2002b) Flood forecasting using Jordan recurrent artificial neural networks. *Flood Defence 2002, Proceedings of the. 2nd International Symposium on Flood Defence, Beijing*, Vol. II, Science Press: Beijing, pp. 908–914.

Walczak S. and Cerpa N. (1999) Heuristic principles for the design of artificial neural networks. *Information and Software Technology*, **41**, 109–119.

Wen C.-G. and Lee C.-S. (1998) A neural network approach to multiobjective optimization for water quality management in a river basin. *Water Resources Research*, **34**(3), 427–435.

Whitley R. and Hromadka T.V. (1999) Approximate confidence intervals for design floods for a single site using a neural network. *Water Resources Research*, **35**(1), 203–209.

Wilby R.L., Abrahart R.J. and Dawson C.W. (2003) Detection of conceptual model rainfall-runoff processes inside an artificial neural network. *Hydrological Sciences Journal*, **48**(2), 163–181.

Zealand C.M., Burn D.H. and Simonovic S.P. (1999) Short-term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, **214**, 32–48.

Zhu M.L. and Fujita M. (1994) Comparisons between fuzzy reasoning and neural network methods to forecast runoff discharge. *Journal of Hydroscience and Hydraulic Engineering*, **12**(2), 131–141.

21: Rainfall-runoff Modeling Based on Genetic Programming

VLADAN BABOVIC AND MAARTEN KEIJZER

DHI Water & Environment, Agern Alle, Hørsholm, Denmark

The runoff formation process is believed to be highly nonlinear, time varying, spatially distributed, and not easily described by simple models. Considerable time and effort has been directed to model this process, and many hydrologic models have been built specifically for this purpose. All of them, however, require significant amounts of data for their respective calibration and validation. Using physical models raises issues of collecting the appropriate data with sufficient accuracy. In most cases, it is difficult to collect all the data necessary for such a model.

By using data-driven models such as genetic programming (GP), one can attempt to model runoff on the basis of available hydrometeorological data. This work addresses the use of GP for creating rainfall-runoff (R-R) models both on the basis of data alone, as well as in combination with conceptual models (i.e taking advantage of knowledge about the problem domain).

INTRODUCTION

The runoff formation process is believed to be highly nonlinear, time varying, spatially distributed, and not easily described by simple models. Considerable time and effort has been devoted to model these processes, and many hydrologic models have been built specifically for this purpose. These models are generally referred to as rainfall-runoff (R-R) models. The R-R model is a hydrologic model, which basically determines the runoff signal that leaves the watershed basin from the rainfall signal received by this basin. According to the traditional hydrologic classifications R-R models are grouped into three categories, namely: empirical black-box models, lumped conceptual models, and distributed physically based modeling systems. The great majority of the R-R modeling systems used in practice are from the first two categories.

Empirical black-box models are entirely lacking in an explicitly well-defined representation of the physical processes involved in the transformation of rainfall into runoff. A large number of black-box models have their origin in the unit hydrograph theory of (Sherman, 1932) and are considered to be at the lower end of the scale in terms of inclusion of physical laws into the model structure. These models depend on rainfall and discharge observations for the estimation of their parameters and for further refinement of their structure. It is believed that the black-box models do not work very well outside the conditions used for their development and calibration. However, experience over decades has shown that these models are useful operational tools, and indeed, they are the only option in cases where there are no other meteorological data available except rainfall, or these data are of poor quality.

Conceptual R-R models are designed to approximate (in some physically realistic manner) the general internal subprocesses and physical mechanisms that govern the hydrologic cycle. Conceptual models are usually based on simplified forms of the physical laws and are generally nonlinear, time-invariant, and deterministic with parameters that are representative of the watershed characteristics. Such models ignore the spatially distributed, time-varying, and stochastic nature of the rainfall-runoff process and attempt to incorporate realistic representations of the major nonlinearities inherent in the R-R relationships. Again, despite their simplicity, many such models have proven quite successful in representing an already measured hydrograph. However, the implementation and calibration of such a model typically presents various difficulties, requiring sophisticated

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

mathematical tools, significant amounts of calibration data, and some degree of expertise and experience with the model.

While there are a large number of existing black-box and conceptual models, there are only a few distributed physically based hydrologic modeling systems suitable for research purposes and for real world projects. Deterministic models are explicitly based on our current understanding of the physics of the constituent hydrological processes. Perhaps the most widely known such system is the Systeme Hydrologique Europeen (SHE) (Abbott *et al.*, 1987) created jointly by the Institute of Hydrology, the Danish Hydraulic Institute, and SOGREAH. SHE is a general, physically based, distributed modeling system for constructing and running models of all or any part of the land phase of the hydrological cycle for any geographical area. These types of modeling systems have extensive data demands. They utilize quite a large number of parameters in their operation, which have a direct relation to physical catchment characteristics (topography, soil, vegetation, and geology) and operate within a distributed framework to account for the spatial variability of both physical characteristics and meteorological conditions. Even so, deterministic models also need calibration mainly because the parameters they require could not or are not directly measured everywhere in the modeled basin. The physically based distributed models do not have the applicability shortcomings of the models from the first two groups. In general, they are not directed only towards studying the R-R processes but also some other processes like erosion, conjunctive use of ground water and surface water, and environmental impacts of land use changes related to the agricultural and forestry practices, which are much more important than R-R alone. To model the runoff of a certain river basin using physical models raises issues of collecting the appropriate data with sufficient accuracy. In most cases, it is difficult to collect all the data necessary for such a model. Furthermore, this kind of model requires significant amounts of data for their calibration and validation.

An alternative to the outlined approaches may be to use new data-driven black- or grey-box type techniques that can model the process using only basic hydrometeorological data. Artificial neural networks (ANNs) have already gained much popularity in hydrologic circles (Minns and Hall, 1996). Another such technique is genetic programming (GP) (Koza, 1992). GP is a relatively new domain-independent method for evolving computer programs for solving or approximately solving problems. GP's learning algorithm is inspired by the theory of natural evolution and by our current understanding of biology and natural evolution.

The road map for the rest of the paper is as follows. First, evolutionary algorithms (EAs) as a method for constructing equations on the basis of data are described. Then, a case study, Orgeval catchment, is described in greater detail, and finally the rainfall-runoff process in the catchment is modeled using GP. Several approaches are presented and discussed in the concluding sections.

EQUATION BUILDING

When refining a model of a physical process, a scientist focuses on the agreement of theoretically predicted and experimentally observed behaviour. If these agree in some accepted sense, then the model is "correct" within that context. Here, we consider the problem inverse to verification of theoretical models: how can we obtain the governing equations directly from measurements? To do this, we will extend the notion of qualitative information contained in a sequence of observations to consider directly the underlying dynamics. We will show that, using this information, one can deduce the effective governing equations. The latter summarize up to an *a priori* specified level of correctness or accuracy, the deterministic portion of the observed behaviour. The observed behaviour on short timescales unaccounted for by the reconstructed equations will be considered as extrinsic noise.

Evolutionary Computation

According to the Darwinian theory of evolution, all animals and plants inhabiting our planet are actually descendants of a few primitive progenitors. Darwin, in the illustrious work *The Origin of Species by Means of Natural Selection* (Darwin, 1859), claims that all complex and intricate life forms that surround us are actually direct offsprings of these original prototypes. However, the offsprings differ from the original ancestors. They are not exact copies of their ancestors, but rather variations that possibly provide competitive advantages over other, similar, specimens in the same environment. And so, claims Darwin, through the process of copying (reproduction) with variations (mutation) and competition for resources, the organisms that evolve possess capabilities that are best adapted to the environment they are situated in. Survival of the fittest thus results in a situation in which a given environment is populated with the best adapted (most fit) organisms.

Evolutionary algorithms (EAs) are processes that are closely inspired by the Darwinian theory of evolution and have one principal objective: to evolve solutions to the problems, rather than to solve problems directly. The fundamental idea is no more original than plagiarism of natural processes, which corresponds to providing "algorithmic organisms" with hereditary capabilities, allowing them to reproduce and let them, through competition for resources, evolve those traits that maximize their benefits in a given environment. The environment to which entities adapt in the EA context is actually formed by a problem

domain for which solutions are being evolved. Thus, EAs attempt to mirror evolutionary processes from nature that allow for adaptation of evolving entities to the problem domain, which in turn emerges as a solution to a problem in question.

Here, we first outline properties of natural evolution, and then attempt to mirror those in an artificial media, as exemplified through EAs.

Properties of Natural Evolution

Natural evolution has been extremely successful in creating many "useful" things. Technology can be nothing but jealous about the successes of natural evolution. The success of adaptation achieved by living organisms to their environment can hardly be matched by human creations. For example, the rate of energy consumption for a given speed of any modern submarine, let alone surface vessel, exceeds that of a fish swimming in the water by several orders of magnitude. What are the processes that enabled natural evolution to construct such effective creations? According to the prevalent views, there are three main criteria for an evolutionary process to occur (Maynard-Smith, 1975):

Criterion of heredity	Offsprings are similar to their parents: the genotype copying process maintains a high fidelity;
Criterion of variability	Offsprings are not exactly the same as their parents: the genotype copying process is not perfect;
Criterion of fecundity	Variants leave a different number of offspring; specific variations have an effect on behavior and behavior has an effect on reproductive success.

The three requirements above are the necessary and sufficient conditions for an evolutionary process to occur. The criterion of heredity assures that offspring inherits information from parents, assuring their similarity. Variability is ensured through mutations, whereas the criterion of fecundity provides, on an average, more fit individuals with possibilities to reproduce more often, thus generating more and better-surviving offspring.

Evolutionary Algorithms

EAs are engines simulating grossly simplified processes occurring in nature and implemented in artificial media – such as a computer. The family of EAs today is divided into four main streams: evolution strategies (Schwefel, 1981), evolutionary programming (Fogel *et al.*, 1966), genetic algorithms (GAs) (Holland, 1975), and genetic programming (Koza, 1992). Although different and intended for different purposes, all EAs share a common conceptual base (schematized in Figure 1). In principle, an initial population of individuals is created in a computer and allowed to evolve using the principles of inheritance (so that offspring resemble parents), variability (the process of offspring creation is not perfect – some mutations occur), and selection (more fit individuals are allowed to reproduce more often whereas less fit, less often so that their "genealogical" trees disappear in time). One of the main advantages of EAs is their domain independence. EAs can evolve almost anything, given an appropriate representation of evolving structures. Similarly, in processes observed in nature, one should distinguish between an evolving entity's genotype and its phenotype. The genotype is basically a code to be executed (such as a code in a DNA strand), whereas the phenotype represents a result of the execution of this code (such as any living being). Although the information exchange between evolving entities (parents) occurs at the level of genotypes, it is the phenotypes in which one is really interested.

The phenotype is actually an interpretation of a genotype in a problem domain. This interpretation can take the form of any feasible mapping. For example, for optimization and constraint satisfaction purposes, genotypes are typically interpreted as independent variables of a function to be optimized. Along these lines, one can employ

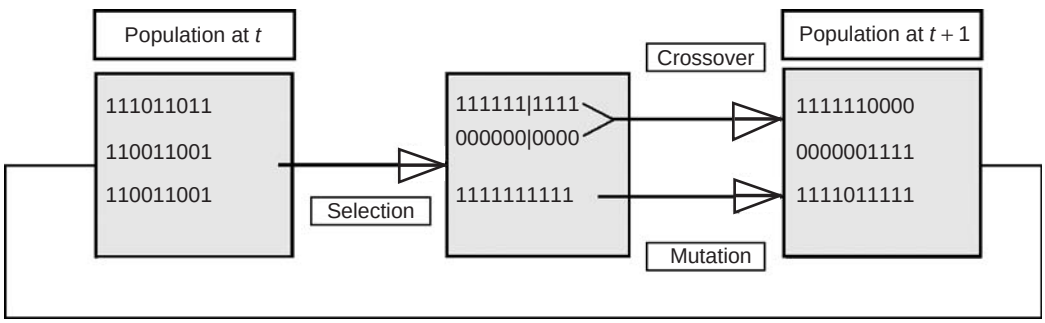


Figure 1 Schematic illustration of an evolutionary algorithm

mappings in which genotypes are interpreted as roughness coefficients in a free surface pipe flow model with the GAs directed toward the minimization of the discrepancies between model output and measured water level and discharge values. The resulting GA represents an automatic calibration model of hydrodynamic systems (Babovic *et al.*, 1994) (Madsen, 2000). Several other applications of GAs, which make use of various kinds of genotype-phenotype mappings and with a specific emphasis on water resources, are described, for example, in (Babovic, 1996).

Genetic Programming

Genetic Programming is one instance of the EA family. In GP, the evolutionary force is directed towards the creation of models that take a symbolic form. In this evolutionary paradigm, evolving entities are presented with a collection of data and the evolutionary process is directed towards the creation of a closed-form symbolic expression describing the data. In its primitive form, GP lends itself quite naturally to the process of induction of mathematical models based on observations: GP is an efficient search algorithm that need not assume the functional form of the underlying relationship. Given an appropriate set of basic functions, GP discovers a (sometimes very surprising) mathematical model that approximates the data well.

Individual solutions in GP are computer programs represented as parse trees (Figure 2). The population of the very first generation is usually generated through a random process. However, subsequent generations are evolved through genetic operators of selection, reproduction, crossover, and mutation. GP thus iteratively applies variation and selection on a population of evolving parse trees representing symbolic expressions. Standard variation operators in GP are subtree mutation (replace a randomly chosen subtree with a randomly generated subtree) and subtree crossover (replaces a randomly chosen subtree from a formula with a randomly chosen subtree from another formula – Figure 3). For a detailed description, see, for example, (Babovic and Keijzer, 2000).

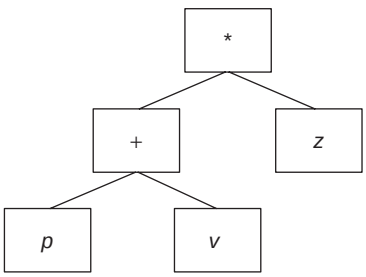


Figure 2 An equation $(p + v) \cdot z$ represented as a parse tree

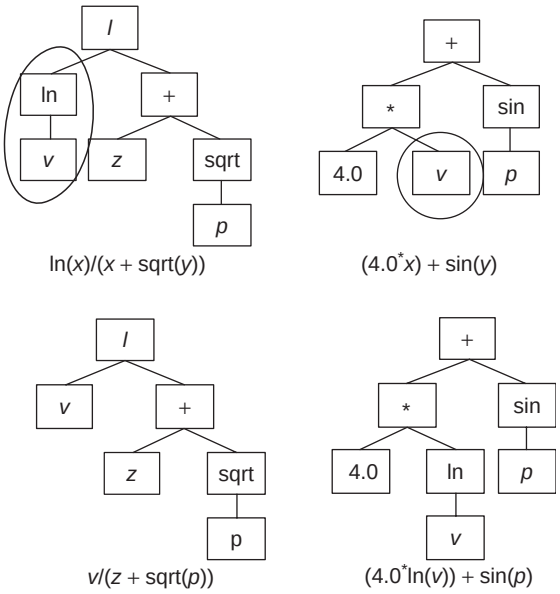


Figure 3 The action of the crossover operator: subtrees of selected parents (above) are swapped in crossover to generate the offspring (below)

The types of functions used in this tree structure are user-defined. This means that they can be algebraic operators, such as sin, log, +, −, and so on, but they can also take the form of if-then-else rules, making use of logical operators such as OR, AND, and so on.

The search process in GP is guided by fitness (i.e., a measure of accuracy). Determination of the fitness function to be adopted is an important aspect in GP since its performance largely depends upon how well this fitness function represents the objective or goal of the problem at hand. In the present work, we adopt a multiobjective approach in which both root mean squared error (RMS) and Coefficient of Determination (CoD) are used as fitness functions. The evolutionary process is then guided towards simultaneously minimizing RMS and maximizing CoD towards the value of unity. It has been shown empirically (Babovic and Keijzer, 2000) that this combination of objective functions implicitly promotes parsimony and results in simpler expressions.

A number of applications of GP has been reported, such as studies in which salt intrusion data were analyzed (Babovic and Minns, 1994), experimental data for bed concentration of suspended sediment (Babovic and Keijzer, 1999), analysis of roughness forces induced by vegetation (Babovic and Keijzer, 2000), as well as rainfall-runoff modeling (Babovic and Abbott, 1997a) (Khu *et al.*, 2001). In all of the above-mentioned studies, GP-induced relationships provided more accurate descriptions of data than those obtained using more conventional methodologies. An extensive survey of the applications of GP in water

resources is provided in (Babovic and Abbott, 1997b) and (Babovic, 1996).

Symbolic Regression

Regression – linear or nonlinear – plays a central role in the process of finding empirical equations. In its most general form, regression techniques proceed by selecting a particular model structure and then estimating the accompanying coefficients based on the available data. The model structure can be linear, polynomial, hyperbolic, logarithmic and so on. The only requirement in such an approach is that the coefficients in the model can be estimated using an optimization technique. In generalized linear regression for instance, the only requirement is that the model is linear in the coefficients. The model itself can consist of any functional form. Another technique may be a nonlinear regression where the only requirement is that the model is differentiable both in the inputs and in the coefficients. Supervised ANNs belong to this class of regression techniques.

Genetic programming can also be understood as a regression technique, a so-called *symbolic regression*. The specific model structure is not chosen in advance, but is part of the search process. In this algorithm, both model structure and coefficients are searched for simultaneously. The user has to define some basic building blocks (function and variables to be used); the algorithm tries to build a model using only those specified blocks. As a space of model structures is, in general, not smooth, not differentiable, nor linear in any useful sense (it is in fact highly discontinuous), standard optimization techniques fail when trying to find both the model structure and the coefficients.

CASE STUDY

The catchment under consideration is the Orgeval catchment, in France (Figure 4), which has been studied extensively in the World Meteorological Organization's intercomparison project (WMO, 1992). The catchment is located about 80 km east of Paris and the main river that drains the catchment runoff is the Orgeval. The catchment has an area of about 104 km^2. The catchment comprises mainly rural area, with only 1% of the total being urban areas or roads, and 18% of the total being covered by forest.

In this study, a total of 10 storm events – from 1972 to 1974 – hourly flow record are selected for training the GP while a total of six storm events (denoted as Storms 1–6) between 1979–1980 are selected and used for the verification of the updating procedure.

Forecast based on Conceptual Model – NAM

In order to establish grounds for intercomparison, the widely used R-R simulation model NAM (Nielsen and Hansen, 1973) is used to simulate the runoff for the entire period of interest. Since the main interest is the investigation of the skill related to the modeling of runoff processes (i.e. runoff as a response to forcing by rain), in all cases a so-called *ideal rainfall forecast* (measured rainfall was used in place of forecasted values) is assumed.

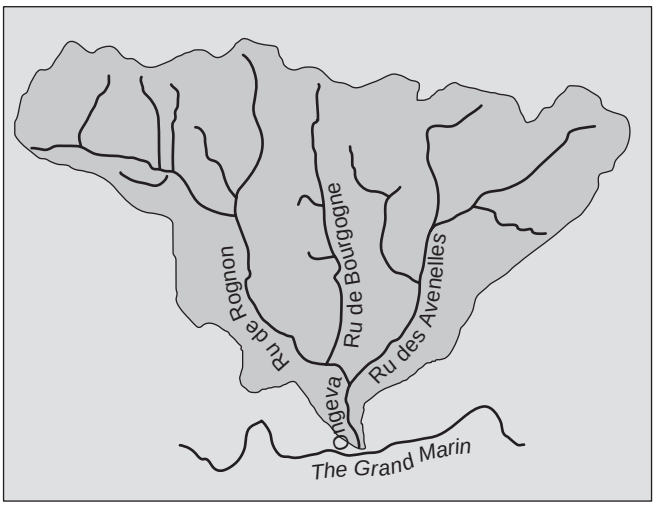


Figure 4 The Orgeval catchment

NAM represents a model of an R-R process. Given the ideal rainfall forecast, the quality of runoff forecast will not deteriorate the forecast horizon. For the present case, the forecast skill is summarized in Table 1.

Naïve Forecast

Another, almost trivial, possibility is to use a so-called *naïve forecast*: one simply assumes a forecast value which is exactly the same as the presently observed discharge. Owing to the strong autocorrelation, the forecast skill is expected to be good for very short lead times, but also to quickly deteriorate with forecast horizons. The results for the naive forecast are summarized in Table 1.

Forecast Based on Genetic Programming

A forecasting system is based on information of the past and current states of hydrometeorological and catchment conditions as inputs, as well as forecasted values of forcing term (rainfall R in this case) in order to forecast the catchment's response (runoff $\widehat{Q}$) in the future. Mathematically, this relationship can be expressed as:

$$\widehat{Q}(t+1) = F(Q_{\text{obs}}(t), Q_{\text{obs}}(t-1), \ldots, Q_{\text{obs}}(t-5), R(t+1), R(t), \ldots, R(t-5)) \quad (1)$$

In the present case, the choice of orders for $Q_{\text{obs}}(t)$ and $R(t)$ of the immediate past 5 time-steps were based on

Table 1 Statistical measures of accuracy (mean absolute error – *MAE*, Correlation coefficient – r, and Pearson's R^2) for the GP forecast as well as for the naïve forecast 1–12 hours

Forecast Horizon	GP Forecast			Naïve forecast		
	MAE	r	R^2	*MAE*	r	R^2
1 h	0.0161	0.9995	0.9991	0.0483	0.9973	0.9946
2 h	0.0245	0.9990	0.9980	0.0954	0.9899	0.9798
3 h	0.0322	0.9984	0.9969	0.1412	0.9787	0.9578
4 h	0.0390	0.9979	0.9957	0.1857	0.9646	0.9305
5 h	0.0445	0.9973	0.9947	0.2285	0.9484	0.8995
6 h	0.0495	0.9969	0.9938	0.2698	0.9307	0.8662
7 h	0.0537	0.9966	0.9932	0.3099	0.9118	0.8315
8 h	0.0571	0.9964	0.9927	0.3482	0.8921	0.7958
9 h	0.0597	0.9962	0.9924	0.3851	0.8715	0.7594
10 h	0.0623	0.9960	0.9920	0.4202	0.8500	0.7224
11 h	0.0647	0.9958	0.9916	0.4533	0.8276	0.6850
12 h	0.0682	0.9955	0.9911	0.4846	0.8044	0.6471

the catchment's concentration time, which varies up to a maximum of 5 h, that is, 5 time-steps (WMO, 1992).

For forecasts that extend longer into the future (α time-steps into the future), a slightly different, so-called *iterative approach* was utilized:

$$\widehat{Q}(t+\alpha) = F(\widehat{Q}(t+\alpha-1), \widehat{Q}(t+\alpha-2), \ldots, \widehat{Q}(t+\alpha-5), R(t+\alpha+1), R(t+\alpha), \ldots, R(t+\alpha-5)) \quad (2)$$

To be precise, in the present case, GP was utilized to forecast the temporal difference between the current and the future discharge $dQ(t+1)$, rather than the absolute value of the discharge $Q(t+1)$. There are two strong reasons for adopting such a setup. Firstly, due to a very strong autocorrelation of the discharges, there is a pronounced local optimum for forecasting discharges of the form $Q(t+1) = \beta Q(t)$, β being a constant, typically smaller than one. Such a local optimum may be statistically very accurate, but the associated phase error discredits its use as a forecasting tool. Once the temporal differences are introduced, the strong autocorrelation is removed, and the GP is forced to approximate change in response $dQ(t+1)$ as a function of forcing terms (rainfall R) and past discharges $Q(t)$. Secondly, temporal differencing of time series of discharges $Q(t)$ removes any of the possible trends that may exist in raw data and, consequently, yields a less biased forecast.

$$dQ(t+1) = (Q(t) - Q(t-1)) * \sqrt{\frac{R(t) + \sqrt{Q(t-1) + Q(t-2)^2}}{\frac{Q(t)}{Q(t)+Q(t-1)+Q(t)}}} + Q(t) \quad (3)$$

Equation (3) fundamentally models $dQ(t+1)$ by multiplying $dQ(t) = Q(t) - Q(t-1)$ with a nonlinear, time-varying correction factor. The correction factor is based on past discharges $Q(t)$, $Q(t-1)$ and $Q(t-2)$ as well as forecast rainfall intensity $R(t)$, and this explains the absence of phase error (see Figure 5). For longer lead times, the quality of iterative forecast deteriorates only due to errors introduced through calculated discharge. The ultimate result is that an approach based on GP outperforms NAM even for lead times of 12 h (see Figure 6).

Updating

The previous two sections demonstrated that forecast based on data deteriorates as a function of forecast lead time. At the same time, the forecast based on NAM was not as accurate; however, the quality of the forecast did not deteriorate with the forecasting horizon. A logical idea is to combine the two and provide a hybrid in which the best of the two approaches is combined, yielding a highly accurate forecast that does not deteriorate with forecasting lead times. This corresponds to a form of data assimilation, namely, the one of error-correction (for more details see (Refsgaard, 1997)).

This method is particularly interesting in real-time forecasting, where the originally forecasted values may be updated or modified as measured data become available and, thus, prediction errors can be determined and used to improve forecast skill. In real-time runoff forecasting with rainfall-runoff simulation models, rainfall time series up to the desired runoff forecast horizon must be available. A similar idea has been utilized before for hydrological problems (see e.g. Khu *et al.*, 2001; Madsen *et al.*, 2000) as well as in marine problems, albeit using neural networks, and for the forecast of current speed in Danish coastal waters (Øresund) (Babovic *et al.*, 2001).

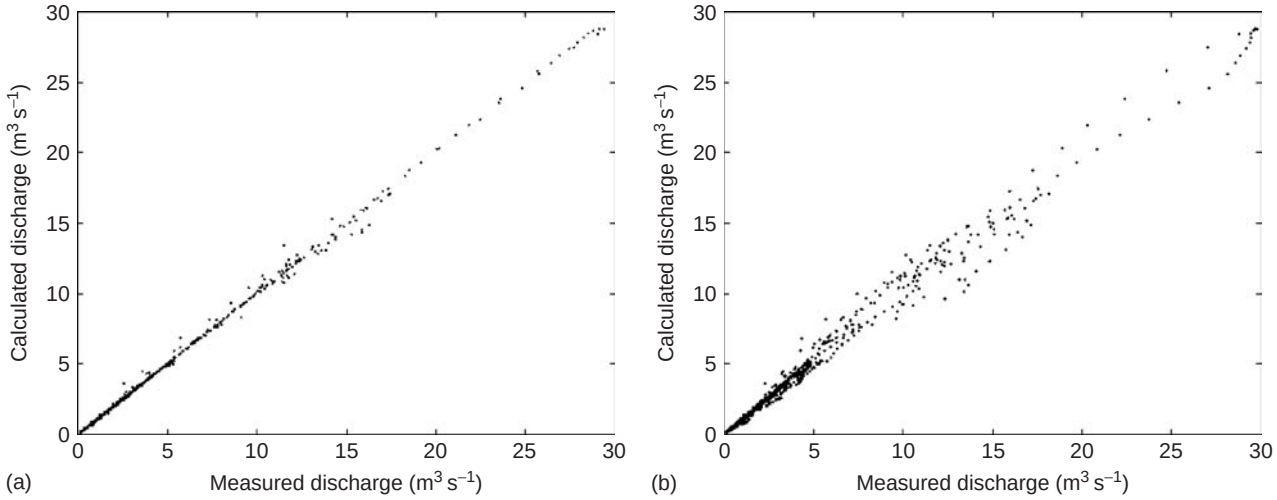


Figure 5 Scatter plots for GP based forecast utilizing equation (3) for lead time of (a) 1 h and (b) 12 h

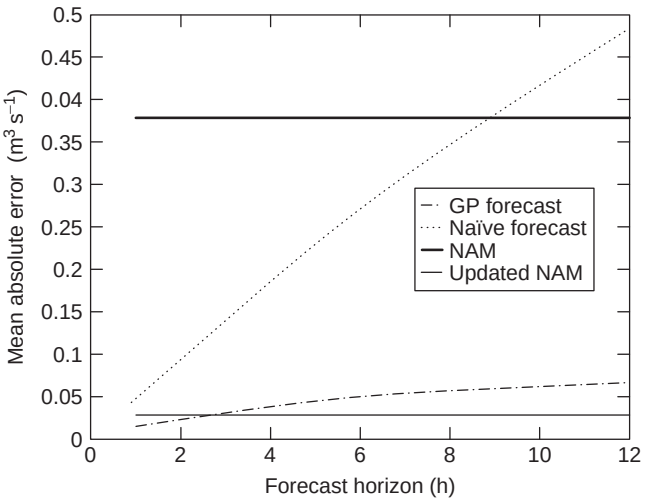


Figure 6 Evolution of mean absolute error as a function of forecasting lead time. The performances is calculated for 6 verification storm events. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Here, NAM is first used to simulate the discharge, Q_{sim}, for the entire period of interest based on the rainfall data R. Then the prediction error ϵ obtained by comparing the simulated discharge Q_{sim} with the observed discharge, Q_{obs} is computed. The improved discharge $\widehat{Q}$ is computed by adjusting Q_{sim} for each forecast lead time within forecast horizon. Mathematically, the measured discharge $Q_{\text{sim}}(t)$ can be expressed as

$$Q_{\text{obs}}(t) = Q_{\text{sim}}(t) + \epsilon(t) \tag{4}$$

Obviously,

$$\epsilon(t) = Q_{\text{obs}}(t) - Q_{\text{sim}}(t) \tag{5}$$

Genetic programing can then be used to approximate the functional relationship between the prediction error and the simulated discharges, the past simulation errors up to the current time as well as rainfall intensity up to forecast horizon. For lead time of 1 h, the functional relationship for the prediction error $\widehat{\epsilon}$ may be expressed as follows:

$$\begin{aligned}\widehat{\epsilon}(t+1) = F(&Q_{\text{sim}}(t+1), Q_{\text{sim}}(t), \ldots, Q_{\text{sim}}(t-5),\\ &\epsilon(t), \epsilon(t-1), \ldots, \epsilon(t-5),\\ &R(t+1), R(t), \ldots, R(t-5))\end{aligned} \tag{6}$$

while the forecast for improved discharge $\widehat{Q}(t+1)$ can be calculated as:

$$\widehat{Q}(t+1) = Q_{\text{sim}}(t+1) + \widehat{\epsilon}(t+1) \tag{7}$$

For longer lead times of $2, 3, \ldots, \alpha$ hours, the recursive form of equation (7) can be written as

$$\begin{aligned}&\widehat{\epsilon}(t+\alpha)\\ &= F(Q_{\text{sim}}(t+\alpha), Q_{\text{sim}}(t+\alpha-1), \ldots, Q_{\text{sim}}(t+\alpha-5),\\ &\widehat{\epsilon}(t+\alpha-1), \widehat{\epsilon}(t+\alpha-2), \ldots, \widehat{\epsilon}(t+\alpha-5),\\ &R(t+\alpha), R(t+\alpha-1), \ldots, R(t+\alpha-5))\end{aligned} \tag{8}$$

Note the use of error estimates $\widehat{\epsilon}$ instead of true error ϵ. The forecast for improved discharge $\widehat{Q}(t+\alpha)$ can now be calculated as

$$\widehat{Q}(t+\alpha) = Q_{\text{sim}}(t+\alpha) + \widehat{\epsilon}(t+\alpha) \tag{9}$$

The real-time flood forecast updating procedure for 1-h lead time could be summarized as follows:

Table 2 Statistical measures of accuracy (Mean absolute error – *MAE*, Correlation coefficient – *r*, and Pearson's R^2) for 'raw' NAM values as well as for updated model. Lead time 1 h

Statistic	NAM	NAM after update
MAE	0.3784	0.0279
r	0.9028	0.9612
R^2	0.8150	0.9240

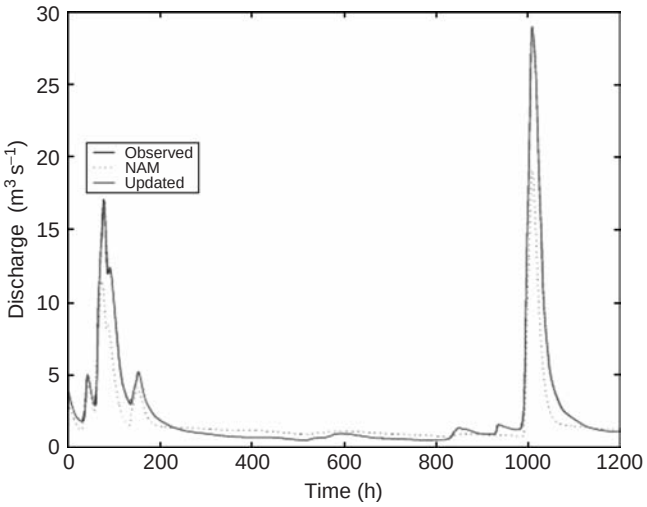


Figure 7 Time series of observed discharge, the one calculated using NAM and the one obtained through updating for two validation events. Lead time is 1 h and the difference between observed and the updated cases is so small that it cannot be optically distinguished

1. Surface runoff has been simulated with parameter values of the NAM model calibrated on 1972–1974 period for the validation period of 1979–1980.
2. The prediction errors, ϵ between the NAM simulated and observed runoff for each time interval are computed.
3. GP is then used to derive the functional relationship between the present prediction error $\widehat{\epsilon}$, the NAM simulated discharge Q_{sim}, the past prediction errors $\widehat{\epsilon}$, and rainfall intensities $R(t)$ as given in equation (6).
4. The improved simulated discharge, $\widehat{Q}$, is finally calculated, using equation (7).
5. For $t > 1$, the above procedure is repeated following equations (8) and (9).

As before and for the same reasons, GP was actually used to approximate a temporal difference in error evolution $\mathrm{d}E(t+1)$ rather than $E(t+1)$.

$$\begin{aligned}&\mathrm{d}\epsilon(t+1)\\&= \epsilon(t) - \epsilon(t-1) + ((\epsilon(t-3) - \epsilon(t-1))\\&\times (0.054 - \epsilon(t-1)) - 0.054(0.064(\epsilon(t) - 6.536)\\&+ ((2\epsilon(t) - 0.0644)R(t-3)Q_{\text{sim}}(t-3))) \qquad (10)\end{aligned}$$

The first two terms in equation (10) are the error at the t^{th} time-step $\epsilon(t) - \epsilon(t-1)$. This error is then corrected by introducing a high-order correction term which utilizes past rainfall $R(t-3)$ as well as output of conceptual model $Q_{\text{sim}}(t-3)$. Once this equation is used to calculate the error, and this error is in turn appended to NAM output Q_{sim}, the updated results provide a manyfold improvement over raw model outputs (see Table 2 and Figure 7).

CONCLUSIONS AND DISCUSSION

Several issues have emerged in the preceding sections.

1. Forecasting on the basis of data is possible, and in some cases can do considerably better in short term than forecasting and modeling on the basis of (conceptual) models.
2. The quality of forecasts created on the basis of data alone deteriorates with forecast horizon. This is perfectly reasonable since the initial conditions (in this case, observed discharges) are "washed out" and replaced by calculated discharges. Through the iterative process, inaccuracies are introduced, which amplify with the forecast horizon.
3. Modeling on the basis of our (albeit, conceptualized) insights about the physical processes cannot match short-term forecast skills created on the basis of data alone. However, the quality of such forecasts does not deteriorate with time.
4. It appears that the best approach is to combine the best of the two worlds: Use the data to improve the short-term forecast and use knowledge (in the form of a conceptual model) to help in extending the forecasting horizon without deterioration of the forecast skill. It is rather interesting to observe that the updated model is more accurate than the NAM model alone for lead times well beyond the catchment's concentration time (in this case, around five hours). This is due to the fact that GP forecasts errors created by NAM, and in principle "explains" phenomena not resolved by a conceptual model.
5. Genetic programming proves to be a powerful tool in the context of the rainfall forecast. The convenience of a single and simple equation, yet of extreme accuracy defends its use as an approach to short-term forecast.
6. Finally, it is very important to emphasize that it is the updating approach that provides the most accurate results. This clearly demonstrates that an amalgamation of knowledge (in the form of a conceptual R-R model) with a data-driven approach (in the present case, in

the form of GP) provides the best forecast skill. The authors strongly believe that it is a combination of the two approaches that will enable us to gain new insights, which may ultimately lead to better and more accurate rainfall-runoff models.

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J., O'Connel P.E. and Rasmunsen J. (1987) An introduction to the european hydrological system – systeme hydrologique europeen (she) 1: History and philosophy of physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 45–59.

Babovic V. (1996) *Emergence, Evolution, Intelligence: Hydroinformatics*, Balkema: Rotterdam.

Babovic V. and Abbott M.B. (1997a) Evolution of equation from hydraulic data: Part ii – applications. *Journal of Hydraulic Research*, **35**, 15–34.

Babovic V. and Abbott M.B. (1997b) Evolution of equation from hydraulic data: Part I – theory. *Journal of Hydraulic Research*, **35**, 1–14.

Babovic V., Cañizares R., Jensen H.R. and Klinting A. (2001) Artificial neural networks as a routine for updating of numerical models. *ASCE Journal of Hydraulic Engineering*, **127**, 181–193.

Babovic V. and Keijzer M. (1999) Data to knowledge – the new scientific paradigm. In *Water Industry Systems: Modelling and Optimisation Applications*, Savic D. and Walters G. (Eds.), Research Studies Press: Exeter, pp. 3–14.

Babovic V. and Keijzer M. (2000) Genetic programming as a model induction engine. *Journal of Hydroinformatics*, **2**, 35–60.

Babovic V., Larsen L.C. and Wu Z. (1994) Calibrating hydrodynamic models by means of simulated evolution. In *Proceedings of the First International Conference on Hydroinformatics*, Verwey A., Minns A.W., Babovic V. and Maksimovic C. (Eds.), Balkema: Rotterdam, pp. 193–200.

Babovic V. and Minns A.W. (1994) Use of computational adaptive methodologies in hydroinformatics. In *Proceedings of the First International Conference on Hydroinformatics*, Verwey A., Minns A.W., Babovic V. and Maksimovic C. (Eds.), Balkema: Rotterdam, pp. 201–210.

Darwin C. (1859) *The Origin of Species by Means of Natural Selection*, *Sixth Edition*, John Murray: London.

Fogel L., Owens A. and Walsh M. (1966) *Artificial Intelligence Through Simulated Evolution*, Ginn: Needham Height.

Holland J. (1975) *Adaptation in Natural and Artificial Systems*, University of Michingan: Ann Arbor.

Khu S.T., Liong S.-Y., Babovic V., Madsen H. and Nuttil N. (2001) Genetic programming and its application in real-time runoff forecasting. *Journal of American Water Resources Association*, **37**(2), 439–451.

Koza J.R. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press: Cambridge.

Madsen H. (2000) Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of Hydrology*, **235**, 276–288.

Madsen H., Butts M., Khu S.T. and Liong S.Y. (2000) Data assimilation in rainfall runoff forecasting. *Proceedings of 4th International Conference on Hydroinformatics*, Cedar Rapids: Iowa.

Maynard-Smith J. (1975) *The Theory of Evolution*, *Third Edition*, Penguin books: Harmondsworth.

Minns A.W. and Hall M.J. (1996) Artificial neural networks as rainfall-runoff models. *Journal of Hydrological Sciences*, **41**, 399–417.

Nielsen S. and Hansen E. (1973) Numerical simulation of rainfall runoff process on a daily basis. *Nordic Hydrology*, **4**, 171–190.

Refsgaard J.C. (1997) Validation and intercomparison of different updating procedures for real-time forecasting. *Nordic Hydrology*, **28**, 65–84.

Schwefel H.-P. (1981) *Numerical Optimization of Computer Models*, Wiley: Chichester.

Sherman L.K. (1932) Streamflow from rainfall by the unit-graph method. *Engineering News Record*, **108**, 501–505.

WMO (1992) *Simulated Real-Time Intercomparison of Hydrological Models*, WMO Operational Hydrology Report 38 – WMO No. 779, World Meteorological Organisation, Geneva.

22: Evolutionary Computing in Hydrological Sciences

DRAGAN SAVIC AND SOON-THIAM KHU

Centre for Water Systems, School of Engineering, Computer Science and Mathematics, University of Exeter, Exeter, UK

With the advent of data acquisition techniques and data processing capabilities (computers + algorithms), evolutionary computing (EC) has found its way into a wide range of applications in hydrological science as demonstrated in this article. These applications range from the use of EC to assist in the understanding of hydrological processes, to improve the performance of simulation models for almost all aspects of hydrologic applications, and most recently, to quantify risk and uncertainty with the aim of providing decision support.

In this article, an attempt is made to provide a detailed review of the applications of EC in different hydrological science topics. References are divided into six main categories: hydrologic processes research, rainfall-runoff modeling, reservoir operations and control, groundwater systems, urban water systems, and water quality studies. These are broad categorizations, and many applications of EC span across multiple groupings.

Evolutionary Computing is a continuously and rapidly growing area of academic research for computer science, mathematics, economics and management, engineering, physical science, and many others. As such, it is futile to try to provide a comprehensive overview of the currently available EC tools and techniques and therefore, no such attempt is made in this article.

INTRODUCTION

With the advent of data acquisition techniques and data processing capabilities (computers + algorithms), evolutionary computing (EC) has found its way into a wide range of applications in hydrological science and indeed, forms one of the main thrust of a new water discipline known as *Hydroinformatics*.

EC refers to a broad range of computing techniques that uses a common algorithmic framework consisting of the following salient features:

- utilizing a set of solutions instead of a single solution;
- generating new solutions from previous solutions through heuristics; and
- modifying previous solutions through a combination of subtle and sudden changes;

In the process of generating new solutions and modifying old ones, EC imitates the mechanisms of heredity, variation, and selection at an abstract level. As such, EC has the ability to tolerate deteriorations of the quality of previous solutions during the search process and overcome local suboptima in order to find global or near global solutions.

Very generally, EC mimics the processes of biological evolution with its ideas of natural selection and survival of the fittest to provide effective solutions for optimization problems. EC uses algorithms based on natural evolution to solve a wide range of problems, which may not be solvable by standard and/or conventional optimization techniques. Examples of EC are genetic algorithms (Holland, 1975; Goldberg, 1989), genetic programming (Koza, 1992), evolutionary strategies (Schwefel, 1981), evolutionary programming (Fogel *et al.*, 1966), shuffled complex evolution (Duan *et al.*, 1992), and many other hybrid methods. Readers are encouraged to refer to the Handbook of Evolutionary Computation (Baeck *et al.*, 1997) for a general understanding of EC.

The applications of EC in hydrological science range from the use of EC to assist in the understanding of hydrological processes, to improve the performance of simulation

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

models for almost all aspects of hydrologic applications, and most recently, to quantify risk and uncertainty with the aim of providing decision support. For the purpose of clarity, the research in hydrological sciences may be grouped in the following sections:

- hydrological processes including evaporation, transpiration, rainfall, fundamental understanding of hydraulics and fluid flows, sediment transport, and so on.
- rainfall-runoff simulation and modeling including catchment model calibration, hybrid modeling with artificial neural network (ANN) and data-driven methods, process modeling, and so on;
- reservoir control and operations including reservoir planning and operations, irrigation, flood and drought control, and so on;
- groundwater (hydrogeological) system including groundwater remediation, monitoring and sampling network design, inverse problems or parameter identification/estimation, and so on;
- urban water systems including water distribution systems, urban drainage systems, and wastewater systems; and
- water quality issues including water pollution control, waste load allocation, health risks, and so on.

The above grouping is purely for the convenience of the discussion of application of evolutionary computation in hydrological science and is not an attempt to draw clear distinction between different branches of HS. The reader should understand that there are considerable overlaps between different groupings, or research may span across multiple groupings.

EVOLUTIONARY COMPUTING IN HYDROLOGICAL PROCESSES RESEARCH

One of the most widely researched topics in hydrological process is infiltration and indeed, the application of EC also found its way to optimize, calibrate, and model infiltration. Zeigler *et al.* (1996) proposed a fuzzy system designed by a genetic algorithm (GA) to solve the Green–Ampt equation without using the iteration method. They found that the GA fuzzy system was very efficient compared to traditional methods such as the Newton–Raphson iterative search. Vrugt *et al.* (2001) used a hybrid GA to calibrate the root water uptake model, HYDRUS, for 1D, 2D, and 3D cases. A simple GA was used to estimate the approximate global model parameter and fine-tuned using a simplex algorithm. In Ines and Droogers (2002), GA was used to estimate the parameters for Richards' equation to model soil water transport in the saturated zone where most plant activities are concentrated. A modified μGA (Krishnakumar, 1989) was developed incorporating creep mutation.

Other applications of EC in hydrological processes research include the use of GP to induce simulation models for flow through idealized vegetation (Harris *et al.*, 2003). Formulae or expressions were derived automatically on the basis of laboratory data collected in a flume with steady flow over a deep channel with relatively shallow vegetated floodplains. The resultant expression indicated that parameters describing the shading due to vegetation dominated the hydrodynamic behavior of the compound channel. This was in close agreement with other studies and support the belief that GP is a feasible tool for knowledge induction;

Nicklow *et al.* (2003) used the modified GA coupled with HEC-6 sediment transport simulation model to minimize sediment degradation and deposition in order to control the change in river channel bed morphology. The modification to GA was in the form of a blending crossover operator (Haupt and Haupt, 1998). Kisi (2004) used a differential evolution (Storn and Price, 1997) coupled with fuzzy logic to model suspended sediment transport behavior.

In the area of rainfall modeling, Yoo *et al.* (2003) used a simple GA to estimate six (out of nine) parameter values of the Waymire-Gupta-Rodriguez-Iturbe (WGR) multidimensional rainfall model. The results were compared with those using other rainfall models as well as the WGR model parameters estimated using the Davidson–Fletcher–Powell algorithm. The proposed method was applied to characterize the monthly and regional variation of rainfall fields of the Han River basin in Korea. They found that the parameter values estimated by GA were more consistent and better.

Evolutionary Computing in Rainfall-runoff Simulation and Modeling

There have been a large number of research works applying EC to calibrate runoff models, and one of the earliest applications of EC in hydrological science is the seminal work by Wang (1991) who calibrated the Xinanjiang conceptual rainfall-runoff (CRR) model (essentially a soil moisture accounting model) using a simple binary GA. In Wang (1991), GA was used to estimate the values of seven parameters (five water balance parameters and two linear transformation parameters) when the Xinanjiang model was applied to the Bird Creek catchment in the United States. The objective function used was a modified residual variance formulation taking the initial variance into account. The results obtained from GA were further tuned using the sequential simplex method. Wang (1991) had indicated that the simple binary GA had performed adequately in terms of locating good solutions and fine-tuning of GA may be useful if the GA is trapped in some local optima. He also concluded that GA could also be used to calibrate other hydrological models.

Many studies involving simple GA and runoff simulation have been performed and some of the works published in peer-reviewed journals are summarized here.

Liong *et al.* (1995) used a single-objective GA to calibrate the SWMM urban runoff model and applied it to a small urbanized catchment in Singapore. The GA used was the simple binary GA, known as *GENESIS*. Sixteen storm events of durations ranging from 5 h to 30 h were used for calibration and validation. They found that good results could be obtained for both calibration and validation storm events. Balascio *et al.* (1998) calibrated the SWMM model using a GA with manual fine-tuning option. They found that better results were obtained using manual fine-tuning after a limited number of GA searches. Cluckie *et al.* (1999) and Yuan *et al.* (1999) used a mutation-lead GA to calibrate the RHINO urban drainage model and applied to the Bolton Town Centre in the northwest of England. The mutation operator in their GA is directed (biparous) consisting of "positive" (forward) and "negative" (backward) steps, taken with reference to a random single gene position at any population. This scheme is similar to EC in some way. However, the results from their studies were not discussed and therefore difficult to assess the performance of their calibration algorithm.

Franchini (1996) also suggested a two-step procedure, GA–SQP (a simple GA and Powell's sequential quadratic programming local search) in which the GA was used to locate globally good solutions and the local search was used for fine-tuning the GA search result. They calibrated the 11-parameter CRR model, known as *A Distributed Model* (ADM), which was analyzed for two cases: error free data and actual data. It was shown that GA–SQP was able to achieve 100% success rate on the error free data and only slightly poorer on the actual data. Further research was performed to compare 11 GA schemes with various selection, reproduction, and mutation operators in order to analyze the effect of different GA structure on its capacity to find regions of optimum solutions. They found that GA seemed to be robust where the results obtained from different schemes did not differ from each other significantly. However, the authors also noted, "every objective function has its own intrinsic characteristics and therefore, the most appropriate optimisation algorithm may vary from case to case". With that, a further study was carried out to compare GA with two commonly used calibration techniques: pattern search with sequential quadratic programming (PS-SQP) (Henrickson *et al.*, 1988) and shuffled complex evolution (SCE-UA) (Duan *et al.*, 1992). They found that SCE-UA was able to provide slightly more consistent estimates of parameter values even though none of the algorithms were able to locate the optimal solution for a complex catchment with known parameter values.

Kuzcera (1997) compared different probabilistic optimization algorithms including GA and found that SCE-UA to be superior to the GA. However, since the simple GA was used, Kuzcera (1997) pointed out that "improved genetic algorithm performance is therefore possible" without indicating the method of improving GA.

Mohan (1997) used a simple binary GA to estimate the parameter values of the three-parameter nonlinear Muskingum equation and compared the results with those of traditional gradient-based optimization techniques. They found that GA was able to estimate the model parameters with reasonable accuracy.

Wang (1997) investigated the usefulness and robustness of GA when used to calibrate the expanded nine-parameter Xinanjiang model. The technique was applied to three medium-sized catchments: Bird Creek (United States); Wollombi Brook (Australia); Kizu (Japan); and a small catchment: Halton (Australia). In order to evaluate the robustness of GA, the runoffs were synthetically generated using known parameter values. The models were then calibrated using GA, assuming that the parameter values were unknown, in order to determine whether GA was able to locate the exact parameter values. The results showed that GA was, in general, capable of locating the regions where the exact parameter values lay. However, the author highlighted that, at times, GA was trapped in local optima closed to the known global optima. This is perhaps an indication that either the Xinanjiang model is over-parameterized (resulting in numerous parameter values with similar objective function values), or the objective function used in the study was not able to constrain the search (hence requiring more objective functions).

Cooper *et al.* (1997) calibrated the TANK CRR model using three algorithms: GA, SCE-UA, and simulated annealing (SA). They were concerned with the efficiency of the algorithms, as measured by the number of model simulations required, and accuracy of the parameter values. Four objective functions were considered as performance measures and they were: Nash coefficient, root-mean-square-error of the peak discharges, root-mean-square-error of the average flow, and ordinary least squares. It was found that SCE-UA performed better than GA, which in turn is better than SA. However, for the FLW criteria, the performance of GA was less variable compared to SCE-UA. This study showed that the choice of performance criteria would affect the calibration outcome.

Ndiritu and Daniell (1999) used a hybrid elitist GA to calibrate the 10-parameter MODHYDROLOG CRR model and applied it to a south Australian catchment. The hybrid GA consisted of a combination of fine-tuning (reducing and refocusing of search space), periodic hill-climbing, and subpopulation searches with shuffling. The strategy of forming subpopulation and shuffling was similar to that used in SCE-UA (Duan *et al.*, 1992). Ten separate

simulation runs were performed and the results indicated that the hybrid GA was able to locate the "optimum" parameter values that were very similar to each other for each run. The same algorithm was later applied to calibrate the SIXPAR CRR model with artificially generated runoff time series (Ndiritu and Daniell, 2001). Results showed that the hybrid GA was effective in locating the optimum parameter value, but was less efficient compared to SCE-UA.

Senbeta *et al.* (1999) also endorsed the need for local search to be coupled with GA. The simple binary GA was used, followed by the Rosenbrook method (Rosenbrock, 1960) and the Simplex method (Nelder and Mead, 1965). This sequential search technique was used to calibrate the probability-distributed interacting storage capacity (PDISC) model and applied to six catchments with daily data including three large catchments in Nepal and China and three midsize catchments in Bangladesh, Nepal, and Australia. The PDISC model consists of two submodules: a six-parameter water balance module and a three-parameter routing module. They found that the sequential calibration technique was able to provide satisfactory parameter estimates.

In most recent years, how to improve the efficiency and effectiveness of GA as a calibration technique has been the main focus of the following research.

Seibert (2000) used GA for multicriteria calibration of the HBV model. The method was applied to a synthetic runoff series generated by the HBV model and later to two catchments with different geological characteristics. Seibert (2000) took advantage of the population-based characteristics of GA and formed three subpopulations of parameter sets. Each of the subpopulation was used to optimize a different objective function. After a certain number of iterations, part of each subpopulation was exchanged between the subpopulations. In this way, characteristics of good parameter sets were transferred, and information regarding different objective functions was passed to different subpopulation. Furthermore, the performance criteria (one for runoff and the other for groundwater level) were fuzzified and combined into a single measure taken as the geometric mean of the two fuzzy measures. Seibert (2000) found that the use of multicriteria and additional data helped to constrain the ranges of the parameter values and provided a more realistic representation of the actual process. The same GA was also applied to calibrate three catchments of different sizes, each nested in the other (Seibert *et al.*, 2000). The purpose of this study was to investigate the issues of scale dependency of the similar catchments and simultaneous calibration of runoff series from different sub-catchments. They found that all three sets of optimal parameter were similar in values, indicating that these parameters accounted for the similarity in catchment characteristics but not size/scale. Variations existed in two parameters but they were found to be insensitive within the optimal range. One of the parameters also differs for all three catchments and it could be accounted for by the difference in the slope of the three catchments. It was found that the multicriteria approach had provided good estimates for the catchment parameter values.

Liong *et al.* (2001) proposed a hybrid multicriteria method to calibrate HydroWorks, a urban runoff/drainage model. A modified GA (known as *ACGA*) was used to generate the Pareto front (or trade-off surface), and an artificial neural network was used to provide an inverse mapping of the front. The purpose of the hybrid approach was to generate additional points on the Pareto front for the purpose of decision making. The approach was applied to calibrate the eight-parameter HydroWorks for the Upper Bukit Timah catchment in Singapore. Comparisons were made with a number of multiobjective genetic algorithms, and ACGA was found to be more efficient in their study.

Cheng *et al.* (2002) used a simple GA with fuzzy objectives to calibrate the 17-parameter Xinanjiang model applied to Shuangpai catchment in south China. Thirty four flood events were used for calibration and 11 flood events for validation. Three objectives were considered: peak discharge, time to peak, and total runoff volume. All these objectives are fuzzified and grouped into a matrix and a value of membership degree was assigned. Selection and reproduction were based on a weighted distance that took into account the membership degree matrix. The authors reported that fuzzification of the objectives had assisted the GA to evaluate and simplify the otherwise difficult multiobjective calibration problem.

EC has also been applied in the area of modeling rainfall-runoff process. Babovic and Abbott (1997) used GP to simulate the rainfall-runoff transformation generated from RORB CRR model. They compared GP with ANN and found that both techniques gave equally good coefficient of determinations. However, GP was preferred because more physical insight may be obtained compared with ANN. Savic *et al.* (1999) applied GP to a real case study in the United Kingdom, and compared their results with those obtained from a nine-parameter HYRROM CRR model, a 35-parameter land-use model, and ANN. Again, GP was preferred because of its ability to provide more insight into the form of the rainfall-runoff relationship. Whigham and Crapper (2001) applied GP on two catchments with very different runoff characteristics: (i) Glan Teifi catchment in Wales, a catchment with quick response, but also with large subsurface contribution (ii) Namoi river catchment in Australia, where the runoff ratio was less than 10% and was highly variable depending on seasons. GP results were compared with the IHACRES model results and it was found that GP was able to simulate runoff satisfactorily for both catchments, but IHACRES had problem simulating low flows in the Namoi river catchment. Liong *et al.* (2002)

and Davidson *et al.* (2003) modified standard GP in order to simulate the rainfall-runoff process better. Dorado *et al.* (2003) proposed a coupled ANN–GP approach to simulate rainfall-runoff in an urban catchment.

The use of ANN as a rainfall-runoff simulator is gaining popularity, but researchers often faced the problem of deciding the ANN network architecture to use. Abrahart *et al.* (1999) addressed this problem by using GA to optimize the network configuration. They compared GA with a network pruning technique and found that GA offered distinct advantage in terms of flexibility in network design. See and Abrahart (2001) used GA to optimize a hybrid fuzzy neural network for the purpose of continuous river level forecasting using both rainfall and upstream river level as inputs. Khu *et al.* (2001) used GP as a real-time forecasting technique to update the discharge obtained from MIKE11/NAM model. The technique was applied to a small catchment in France and it was found that GP outperformed all standard updating techniques including ARMA and Kalman filter. Good results were obtained for real-time forecasting horizon of up to 6 h. The use of ANN in hydrology is discussed in **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**.

Evolutionary Computing In Reservoir Control and Operation

In the area of water supply management, most applications of EC are for reservoir operations and control with a small number of applications in irrigation management. One of the earliest applications of EC in reservoir control was by Esat and Hall (1994). They applied a GA to a four-reservoir problem in order to maximize the benefits from power generation and irrigation water supply subjected to constraints on reservoir storages and releases. Esat and Hall (1994) demonstrated advantages of using GA over standard dynamic programming techniques in terms of computational requirements. Subsequently, there were a number of works in or related to this area.

Oliveira and Loucks (1997) developed a GA-based approach to search for effective operating policies for dynamic models of multipurpose multireservoir systems. They found that their approach provided decision makers with significant flexibility and benefits in defining operating rules and policies.

Wardlaw and Sharif (1999) extended the work of Esat and Hall (1994) by examining different GA schemes to solve the four-reservoir problem. The different GA schemes considered different arrangements in terms of representation, selection, reproduction and mutation methods, and the use of elitism. They found that a real-value representation incorporating tournament selection, elitism, uniform crossover, and modified uniform mutation produced the best results. They also applied this GA scheme to a 10-reservoir problem. Sharif and Wardlaw (2000) applied GA to optimize a real multireservoir problem – Brantas Basin in Indonesia. Four case studies were considered: (i) maximizing hydropower returns; (ii) maximizing hydropower and irrigation returns; (iii) same as (ii) but including a future water resources development scenario; and (iv) same as (iii) but including more reservoirs in the system.

Cai *et al.* (2001) used GA combined with LP (linear programming) to improve the efficiency and effectiveness of solving highly nonlinear systems. They applied the GA and LP approach to: (i) a five-reservoir operation problem for hydropower generation, water supply, flood control, and flow augmentation; and (ii) a long-term dynamic, multi-period river basin planning model for irrigation planning and water allocation. The authors found that the multi-period planning problem was too complex even for GA and LP to solve with limited resources and time and the solutions obtained from GA and LP can be improved using gradient-based nonlinear programming.

Merabtene *et al.* (2002) used GA to derive reservoir operation rules (release) for drought management. The approach was applied to the Fukuoka city water supply system, where drought is a major concern due to water sharing conflicts among users – pumping, streamflow depletion for irrigation, water quality deterioration, and slow recovery rate. A decision–support system was developed and GA was used to generate risk-based scenarios. It was found that the optimal solution was biased towards certain scenarios and weight allocations. The authors proposed that future work should include the possibility of fuzzy inferencing to take into account the experience of engineers.

There are a number of research applications on using fuzzy logic to form operation rules and then optimized using GA. Chang and Chen (1998) used a real-code GA to locate good fuzzy control rules for operation of a simplified flood control reservoir. Chang and Chang (2001) used GA to locate the optimal reservoir operating rules based on a given inflow series, which was then used to train the adaptive network-based fuzzy inference system for water release operations. In Miles *et al.* (2002), GA was used to locate the optimal rule set for training a classifier system that was used for deriving control strategies of reservoirs. In Ponnambalam *et al.* (2002, 2003), GA was used to optimize the parameters in the neural fuzzy inference system for a single reservoir operation under stochastic inflow condition.

Chaves *et al.* (2003) proposed a GA coupled with DP approach to optimize reservoir storage taking water quantity and quality into account. Firstly, the reservoir water quality model, PAMOLORE, was calibrated using GA. Stochastic dynamic programming and fuzzy inference system was used to optimize the water quantity model. Then the water quality and quantity models were linked and optimized using deterministic DP and GA.

Chang *et al.* (2003) applied a simple GA to optimize and determine the rules used for flushing a reservoir to avoid excessive sedimentation. A flush model was developed and linked to a reservoir operations model. An optimal level of flushing was desired in order to flush out the sediments, but a high concentration of sediments may create a large impact on the ecological system and water uptake downstream.

Cui and Kuczera (2003) compared the use of GA and SCE (Duan *et al.*, 1992) to optimize the management of a urban water supply headworks system. The headworks system consisted of one reservoir and had three decision variables to be optimized. They found that SCE was more efficient compared to GA, but because of its implicit parallelism, GA was preferred.

In summary, the benefits of using GA for reservoir control and operations can be stated (Sharif and Wardlaw, 2000)as follows:

- Discretization of state and decision space is not required;
- Initial trial state trajectories are not required;
- Complex reservoir systems need not be decomposed as in successive approximation-based approaches;
- It has the ability to work with highly complex, even discontinuous objective functions;
- Noninvertible systems can be easily handled without introducing dummy state variables as in DP-based approaches.

In the area of irrigation management, Chen (1997) applied a simple GA to maximize the economic benefit that could be gained by distributing water resources to each district. In Kuo *et al.* (2000), GA was used to optimize the on-farm irrigation demand on the basis of a climate-soil-plant model. The approach was applied to an actual irrigation scheduling project in Delta, Utah, United States, by optimizing the crop area water allocation thereby maximizing crop production benefits. In a similar application, Nixon *et al.* (2001) applied a GA for optimizing off-farm irrigation scheduling in order to find the optimum delivery schedules. In Montesinos *et al.* (2001), GA was used to locate the optimal schedule for furrow irrigation in a maize field located in Cordoba, Spain. Since the problem was heavily constrained, the authors found that the level of penalty imposed was crucial to the success of GA. Nevertheless, GA was able to find good solutions compared to standard scheduling (heuristic approach).

Recently, Vink and Schot (2002) presented a case study looking at generating drinking water production strategies that allows minimization of production cost and environmental impact. The case study was performed on a catchment scale taking into account variable cost such as drawdown options, damage to wetland vegetations, impact on agriculture sector, reduction of river baseflow, soil subsidence, and so on. A multiobjective GA was used to optimize the model and applied to the water supply system in south of the Netherlands.

EVOLUTIONARY COMPUTING IN GROUNDWATER SYSTEMS

Applications of Evolutionary Computing in groundwater and hydrogeological systems have seen a steady growth since the early 1990s and the hydrogeological community has introduced various new developments in EC such as new fine-tuning techniques and the use of heuristics and multiobjective optimization to the hydrological science community. Three subgroupings are considered hereafter.

Evolutionary Computing in Groundwater Remediation

It is extremely costly to implement a groundwater remediation scheme and thus, every effort is made to identify the most cost-effective way of implementation (Cunha, 2002). In this aspect, EC is ideally placed to offer an effective means to optimize the cost of groundwater remediation. This is certainly true judging from the large number of research publications in this area. Leitao (1998) offers a detailed discussion on the various measures currently available for groundwater remediation.

McKinney and Lin (1994) and Ritzel and Eheart (1994) are generally regarded as pioneers in applying GA to groundwater remediation problems. In McKinney and Lin (1994), GA was used to solve three problems: maximize the pumping rate from an aquifer; minimize the water supply development cost; and minimize the total aquifer remediation cost. In the aquifer remediation problem, the objective was to obtain a steady-state policy/schedule for a pump-and-treat design problem, taking capital and operating cost of treatment, extraction, and injection systems into account. They found that GA was able to locate a better solution than the nonlinear programming approach.

In Ritzel and Eheart (1994), the concept of multiobjective optimization using GA was introduced to solve the groundwater pollution containment problem (using the pump-and-treat method). The authors compared several GA and multiobjective GA schemes (including vector-evaluated GA (VEGA) and a Pareto rank-based GA) with those from mixed-integer chance constrained programming (MICCP). Even though the performance of MOGA(multiple objective genetic algorithm) closely matched that from MICCP, the MOGA formulation had the advantage of easily incorporating different fixed costs that maybe associated with the pump-and-treat system.

There are a considerable number of applications of EC to groundwater remediation (mainly pump-and-treat system) and only some of them are summarized here.

El Harrouni *et al.* (1996) used a binary GA to optimize the pumping rates in a homogeneous aquifer simulated using a dual reciprocity boundary element method. The results were compared with an analytical and a FEM(finite element model) (optimized by LP) solution, and were found to be very close to the analytical solution and indeed better than the FEM/LP solution.

Huang and Mayer (1997) and Wang and Zheng (1997, 1998) used simple GA to determine optimal dynamic policies in which the policies change with each management period. Their objectives were to minimize the remediation cost of contaminated plumes with specific well configurations while minimizing cross contamination. Hsiao and Chang (2002) used a simple GA coupled with constrained differential dynamic programming (Murray and Yakowitz, 1979) to optimize the pumping wells setup, while considering both fixed well-installation costs and operating cost of time-varying pumping simultaneously.

In Aly and Peralta (1999a, b), GA was used to solve a nonlinear mixed-integer problem aimed at minimizing the capital and operating cost of pump-and-treat system. Single and multiple planning periods were considered as different scenarios for optimization. They compared GA with a formal mixed-integer nonlinear programming approach and found that GA was clearly better. Gumrah *et al.* (2000) solved a similar problem, but using a groundwater transport simulator based on Method of Characteristics.

Yoon and Shoemaker (1999) dealt with optimizing *in situ* groundwater bioremediation instead of the pump-and-treat method. They compared a number of EA schemes such as real-code or binary GA, and de-randomized evolutionary strategy (DES), and also direct search methods. They found that DES performed well both in terms of speed and accuracy, but more importantly, no one algorithm was consistently accurate for all three test problems. In a later study (Yoon and Shoemaker, 2001), they found that with suitable modification (such as directed reproduction and screened replacement), the real-code GA performed significantly better than binary GA.

Recognizing the drawback of simple GA in optimizing pump-and-treat system, a number of improvements were made to GA, mainly to increase the computational efficiency of GA. Guan and Aral (1999a, 1999b) coupled their GA with a groundwater simulator (known as *progressive GA*), and this coupled model was then used to optimize the full numerical model. The progressive GA seemed to work well and was able to reduce the computation time significantly. Chan Hilton and Culver (2000, 2001) developed new constraint handling techniques and coupled them with the simple GA. They found that the multiplicative penalty method of constraint handling worked well for both cases of pump-and-treat problem. Morshed and Kaluarachchi (2000) coupled a simple GA with three different types of enhancements on the basis of heuristics developed from better understanding of the performance of simple GA on groundwater problems. The results showed that all three enhancements were able to improve the performance of the simple GA considerably.

In Smalley *et al.* (2000), a noisy GA was used to determine cost-effective options for reducing risk to acceptable levels in an uncertainty context. The risk-based corrective action approach was optimized, and they found that noisy GA was capable of identifying highly reliable designs from only a small number of samples, providing a significant advantage for computationally intensive groundwater management models.

Maskey *et al.* (2002) presented the use of GA to determine the optimal combination of pumping rates and well locations for the removal of a contaminant plume using particle tracking. They used MODFLOW and MODPATH for groundwater flow simulation and particle tracking. They compared GA with (i) controlled random search, (ii) multistart algorithm, and (iii) adaptive cluster covering techniques and found that GA seemed to be the most accurate technique despite requiring more computational effort than other techniques.

Erickson *et al.* (2002) used a multiobjective GA (known as *niched Pareto GA* (NPGA)) to the pump-and-treat groundwater remediation problem. NPGA was applied to simultaneously minimize two objectives: (i) remedial design cost and (ii) contaminant mass remaining at the end of the remediation horizon.

In order to reduce the simulation time required to solve the groundwater remediation problem, ANN was used as a fast surrogate model. Several researchers have used GA coupled with the ANN to find optimal remediation strategies (Rogers and Dowla, 1994; Rogers *et al.*, 1995; Fedra and Jamieson, 1996; Morshed and Kaluarachchi, 1998a, b). Rogers and Dowla (1994) and Rogers *et al.* (1995) used a GA to select pumping wells by maximizing the fitness defined by the combination of regulatory constraints on water quality, amount of extracted pollutant, and normalized cost. They used a trained artificial neural network instead of the direct use of simulation model to reduce the computational burden.

Evolutionary Computing in Groundwater Sampling Design and Monitoring

The applications in this section includes design and planning of groundwater recharge systems, design of monitoring points for contaminated land, and so on. In Cieniawski *et al.* (1995), MOGA was used to determine the optimal locations of monitoring wells by simultaneously minimizing the contaminated area and maximizing reliability. Four different MOGA formulation were tested and they were: (i) weighted aggregate objective function; (ii) vector-evaluated GA; (iii) Pareto rank-based MOGA; and (iv) a combination of VEGA and rank-based formulation. The GA approach

was also compared with the simulated annealing method. The authors found that the combined formulation (iv) outperformed the other formulations in terms of the ability to generate the entire Pareto front. In addition, the parameters of the combined formulations allowed decision makers to focus on the particular portion of the trade-off curve they are most interested in.

Wagner (1995) investigated the groundwater monitoring problem subjected to uncertainty. They minimized the uncertainties of model prediction with a constraint of fixed budget for design in order to identify the optimal sampling locations in a groundwater aquifer, which improves the reliability of groundwater and the contaminant transport model. The GA method was compared to the branch-and-bound algorithm. GA was also used by Katsifarakis *et al.* (1999) to minimize the pumping cost for any number of wells under various constraints. Reed *et al.* (2000, 2001) studied the long-term groundwater monitoring problem as a multiobjective problem. The goal of their application was to minimize long-term monitoring costs while still providing accurate information on the mass of contaminants present. They used the nondominated sorted GA-II (NSGA-II) (Deb, 2001) for the optimization and found that this algorithm was much more efficient than the previously known MOGA (Reed *et al.*, 2003).

Tang and Mays (1998) used a simple GA to optimize the operations of the soil aquifer treatment system, which is a commonly used method for combined groundwater recharge and municipal wastewater removal. They found that the advantage of GA for such a problem was that the upper bound cycle time for the reactor (a constraint) does not need to be prespecified, thereby providing flexibility in the solutions. With this in mind, the authors felt that even though the runtime for GA was more compared to traditional successive linear quadratic algorithm, possible enhancements to GA (such as progressive optimal algorithm (Howson and Sancho, 1975)) may warrant further research.

Evolutionary Computing in Inverse Modeling/Parameter Estimation

Inverse modeling is a commonly adopted technique for the determination of aquifer(s) parameters in order to characterize the aquifer(s) for modeling groundwater flow and contaminant/solute transport and for groundwater/water resource management. Despite issues such as range of applicability, transferability and nonuniqueness of parameter values, reliability and robustness of results, and so on, which cast doubts about the methodology of inverse modeling, it is still a technology that is widely acknowledged as practical and useful, especially when parameters are not directly measurable.

Simple, single-objective GA has been applied to estimate parameter values for a wide range of groundwater models, such as, single aquifer parameters (Lingireddy, 1998), transmissivities in zoned aquifers (Katsifarakis *et al.*, 1999; Prasad and Rastogi, 2001), groundwater model calibration (Solomatine *et al.*, 1999), determination of locations of accretion (Gentry *et al.*, 2001), velocity and dispersivity parameters (Giacobbo *et al.*, 2002), transmissivity, storativity, conductivity, and aquifer thickness (Lee *et al.*, 2002), and locations of small-scale aquitard leakage, (Gentry *et al.*, 2003) amongst others.

In Heidari and Ranjithan (1998), GA was combined with the truncated-Newton search technique to estimate groundwater parameters for a confined steady-state groundwater model. Although the combined technique was successful in estimating the parameter values, it was found that the use of prior information about the parameters was important in obtaining the global optimum. Acknowledging the complex nature of the objective function in groundwater parameter estimation, Karpouzos *et al.* (2001) proposed a multi-population-based GA approach to estimate the transmissivity values. The authors highlighted that for multi-population-based approach, an important parameter for GA is the emigration rate. If the emigration rate is too high, the diversity in different subpopulations could not be maintained; if the migration rate is too low, the convergence rate of GA will decrease and the solutions may be far from optimal.

In Aral *et al.* (2001), the progressive GA (Guan and Aral, 1999b) was used to solve the contaminant source identification problem by minimizing the difference between the observed and simulated contaminant concentration at different locations. The source location coordinates and the source release histories were regarded as unknowns and had to be determined through the optimization process. The authors found that progressive GA was a robust approach that could handle errors in the observations, and errors obtained from the calibrated model stayed within the known bounds of the measurement errors. Tsai *et al.* (2003) used a hybrid GA to determine and locate the number and locations as well as the values of the basis points associated with the Voronoi tessellation. The hybrid GA consisted of a simple GA fine-tuned by grid search and quasi-Newton's algorithm.

In the area of model induction, Hong and Rosen (2002) applied GP to model the dynamic behavior of groundwater fluctuation in a fracture rock aquifer. The results showed that the GP-induced model was capable of predicting groundwater fluctuation due to storm water infiltration and at the same time, provided insight into the dynamic behavior of the modeled system. Because the resultant model was highly efficient in terms of computational runtime, numerous simulations can be made to allow hydrogeologists to extract information for the studied area.

EVOLUTIONARY COMPUTING IN URBAN WATER SYSTEMS

Water Supply/Distribution Systems

There has been a long tradition of applying GA to optimize, calibrate, design, and improve the performance of water distribution network or system. One of the first civil engineering applications of GA is the optimization of a pipeline (Goldberg and Kuo, 1987). However, the first widely referred work in the area of water distribution network (WDN) is that by Simpson *et al.* (1994). Since then, there have been considerable efforts in applying EC to WDN, examples ranging from new applications and case studies, to improvement to GA and development of new algorithms. We will attempt to review works in refereed journal publications according to the different purpose of the applications.

In the area of using GA to design new WDN, Savic and Walters (1997) applied a simple GA to optimize the cost of designing the two-loop network and the Hanoi network. The two-loop network consists of six nodes and eight pipes and full enumeration of all possible designs requires 1.48×10^9 evaluations. Better solutions were found using only 0.017% of the full enumeration. The Hanoi network consists of 32 nodes and 34 pipes organized in three loops. A full enumeration requires 2.87×10^{26} model evaluations, which is not possible even with the present-day computing power. A better solution was found using a maximum of 1×10^6 model evaluations. Gupta *et al.* (1999) also worked on the two-looped network and used a modified heuristic-based GA instead. The heuristics used are: (i) division of WDN into diameter zones according to the judgment of the designer, (ii) equivalent pipe diameters, and (iii) local adjustment to pipe diameters based on head-loss per unit length and pipe flow velocity. They also applied the modified GA to a 23-nodes and 38-pipes WDN.

A number of researchers recognized that the optimal design of WDN is often a complex multiple-objective process involving trade-offs between the total cost and its reliability, functionality, water quality, level of service and other factors (Rowell and Barnes, 1982; Goulter and Bouchart, 1990; Gupta and Bhave, 1994; Todini, 2000). Savic (2002) showed some shortcomings of the use of single-objective optimization for water distribution system design and introduced a GA multiobjective model that promised to ease the difficulties in applying optimization and providing decision support for that important problem. The optimization model used simple and intuitive objectives and constraints that were not difficult to formulate in mathematical terms, for example, cost, number of nodes with pressure deficit, and the extent of pressure deficit. These objectives allow a decision-maker to visualize the trade-offs between different benefits and costs, and more importantly to consider uncertainty in future demands and performance levels. This type of optimization could also take into account the fact that the system needs to be implemented in stages.

In the area of network rehabilitation, Simpson *et al.* (1994) used a simple GA to optimize the rehabilitation cost of a two-supply-source gravity fed network (Gessler, 1985) and found that GA was able to provide near-optimal solutions at a fraction of the full enumeration cost.

Dandy *et al.* (1996) applied an improved GA for the New York City tunnels (NYCT) network. Variable power scaling of the fitness function, Gray coding, and creep mutation operators were used to improve the GA. Results indicated that the improved GA was significantly better than simple GA. Lippai *et al.* (1999) noted that most research work in optimization of WDN used in-house software and proposed a method to link EPANET with several commercial solvers. Their applications included the NYCT and a 15-loop Almos network. A number of researchers also applied different GAs to the NYCT network:

- Montesinos *et al.* (1999) proposed an index-based deterministic selection operator;
- Vairavamoorthy and Ali (2000) proposed to use real-coded GA with a variable penalty for constraint violation. They also applied their method to the Hanoi network;
- Wu and Simpson (2002) proposed using flexible penalty factor for constrained optimization and developed heuristic rules for adjusting the lower and upper bounds of the penalty factor. This method, known as *self-adaptive penalty factor*, was coupled with the fast messy GA;
- Eusuff and Lansey (2003) proposed a hybrid EC, known as *Shuffled Frog Leap Algorithm* (SFLA), which is a cooperative search method inspired by natural memetics. The SFLA draws ideas from different meta-heuristic methods such as, parallel searching – from GA; multiple population – from island model; information exchange – from Shuffled Complex Evolution (SCE) (Duan *et al.*, 1992) and local search – from particle swarm optimization (Eberhart and Kennedy, 1995). They also applied SFLA to the Hanoi network.

Simple GAs usually represent solutions as fixed length strings. Goldberg *et al.* (1989) proposed a variable length representation for GA, known as *messy GA*, which allows the string length to vary within population. Goldberg *et al.* (1989) demonstrated that the messy GA was able to identify good clusters of string patterns leading to good solutions. Halhal *et al.* (1997) applied the structured messy GA (SMGA) to the rehabilitation of a hypothetical network and Moroccan network, a real-world large WDN. This was a restricted form of messy GA in that the string lengths across populations were allowed to change but not within population. The SMGA starts with a complete

enumeration of all single elements to form the initial population. The best elements were retained to form the essential building block for subsequent populations with longer strings. The SMGA was able to take advantage of the problem structure and reduce the search space significantly. Moreover, they had included Pareto ranking as the basis of formulating objective functions, therefore, SMGA can be considered as a multiobjective GA. Walters *et al.* (1999) later applied SMGA to the hypothetical "Anytown" network with good results. The "Anytown" network is one of the more challenging networks, which simulates a full WDN expansion programme with duplications of pipes, addition of new pipes, addition and/or extension of pumping stations, and provision of new reservoir storage at any location. Wu and Simpson (2001) applied a fast messy GA to three case studies: the Gessler network, the NYCT network, and the Moroccan network, and found that messy GA gave consistent results in all the networks.

To make the implementation of GA more accessible and closer to engineering practice, Morley *et al.* (2001) demonstrated the integrated usage of GA optimization, GIS, and a hydraulic network solver, StruMap, on a real-world, large network-expansion and rehabilitation. Despite the complexity of the problem, the proposed method was able to find solutions that provided up to 35% cost savings.

Noting the inherent parallelism of GA and the advancement in network technology, Balla and Lingireddy (2000) implemented an optimization model based on distributed GA on a peer-to-peer network of PCs linked using Ethernet cards running on MS Windows 95/NT operating system. The network was tested to calibrate three WDN (system A, B, and C) with different complexity, that is, 52 pipes, 435 pipes, and 850 pipes respectively. They found that the computational advantage of the PC network roughly corresponded to the number of PCs within the network.

Water distribution systems, like other infrastructures, will age and are subject to deterioration over time, and hence require a long-term perspective in terms of design, maintenance, and rehabilitation scheduling. Dandy and Engelhart (2001) demonstrated the use of GA to find a near-optimal schedule for the replacement of pipes in order to minimize the present value of capital, repair, and damage cost. They applied their method to a case study in Adelaide, Australia.

In the area of network layout design, Walters and Lohbeck (1993) used a GA to evolve tree-like structures to identify good designs for a directed WDN. Smith and Walters (2000) extended the algorithm to undirected networks. Davidson (1999) applied a modified GA to optimize the layout of networks constrained to a fully looped geometry. Fully looped geometry is one in which node isolation of basic vertices will not occur when any single link in the network fails. The modified GA had customized crossover and mutation operators to ensure that newly generated solutions are contiguous and feasible.

The implications of uncertainty in input data and model parameters are not always understood or considered when designing water distribution systems. Therefore, there is a need to develop design methods that can model such uncertainties and produce "robust" designs. Here, robustness of a system is defined as its ability to provide adequate service to customers despite fluctuations in some or all of the design parameters. Kapelan *et al.* (2003a) and Babayan *et al.* (2003) formulated and solved the problem of stochastic (i.e. robust) water distribution system design under condition of uncertainty. The two methods differ in how the uncertainty is quantified within the EC algorithm; the former used reduced Monte Carlo sampling, while the latter used a numerical integration scheme to convert the stochastic problem into its deterministic equivalent. The objective in both cases was to minimize total design costs subject to a target level of system robustness. System robustness is defined as the probability of simultaneously satisfying minimum pressure head constraints at all nodes in the network. The source of uncertainty analyzed in their work is future water consumption. Later on, the same authors formulated and solved the optimal design problem by using a multiobjective GA with the second objective being the level of robustness achieved (Babayan *et al.*, 2004; Kapelan *et al.*, 2004).

In the area of network sampling design, Meier and Barkdoll (2000) addressed the optimal sampling locations for locating flow monitors within the WDN. A simple GA was applied to a network model for a small town I Ohio, US, to optimize the sampling design. de Schaetzen *et al.* (2000) proposed the use of GA with Shannon entropy function as a measure of optimality for the selection of pressure monitoring point locations in WDN. They compared their methods with the shortest path algorithm and expert design and found that there were potential cost savings based on more efficient equipment deployment as suggested by GA. Kapelan *et al.* (2003b) formulated the sampling design problem as a multiobjective optimization problem with relevant constraints. Two sampling design models were developed to solve this problem. They both used GA as a search engine. The first model is based on the single-objective approach in which two objectives (cost and calibration model accuracy) are combined into a single one using appropriate weights. The second model uses the multiobjective GA approach and is based on Pareto ranking. Both the sampling design models were applied to two case studies (literature and real-life problem). Results from the case studies showed the advantages of the multiobjective GA sampling design model when compared with the single-objective and several previously published sampling design approaches.

Water quality monitoring and control in urban water supply systems can also be formulated as an optimization problem, where the aim is to estimate the source dosages, locations of disinfectant boosters, and/or scheduling of boosters. Munavalli and Mohan Kumar (2003) presented a thorough study on the use of GA to estimate the optimal dosage for multiple locations in three actual WDNs. Different chlorine dosage model were studied and the three problems varied in network complexity. GA was found to be well suited for optimal scheduling of multiple chlorine sources.

Leakage detection and calibration of hydraulic models are important issues for the management of water and other distribution networks. Savic and Walters (1997) introduced GAs in the area of water distribution network calibration. The inverse transient problem (i.e. calibration of transient WDS models) was first solved with the use of GAs by Vitkovsky and Simpson (1997). They solved the inverse transient problem by linking GAs to a forward transient solver. Kapelan *et al.* (2003c) developed an inverse transient model based on a hybrid search technique. The inverse transient procedure is formulated as a constrained optimization problem of the weighted least-squares type. Initially, two optimization techniques were tested: the GA and the Levenberg–Marquardt method. After examining their performance, a new hybrid genetic algorithm (HGA) was implemented to exploit the advantages of combining the two methods. The resulting HGA-based inverse transient model is compared with the GA and Levenberg-Marquardt-based inverse transient models using two case studies. The HGA-based inverse transient model proved to be more stable than the Levenberg–Marquardt-based model and more accurate and much faster than the GA-based inverse transient model.

Evolutionary Computing in Urban Drainage/Wastewater Systems

Urban flood control often involves the design of channels, detention, and storage facilities and so on, for the purpose of peak flow attenuation. Flood detention facilities are the most common engineering approach to control the impact of storm water runoff (both quantity and quality), and the design concepts for a single storage component have been well established. However, any single method of flood control is often insufficient, and local approaches may aggravate responses in other parts of the drainage system. Hence, an integrated approach with multiple structures is often required. As a result, the problem of urban flood control becomes highly complex where full enumeration for optimal solutions may not be feasible, and multiple objectives must be considered.

Many urban flood control problems can be formulated as nonlinear mixed-integer problems and hence can be readily tackled through the use of EC. Rauch and Harremoes (1999a) discussed the potential of genetic algorithms in urban drainage modeling and highlighted model calibration and model predictive control as the main applications. Two simple examples, one of each case, were provided. They suggested that the application of GA in multicriteria decision analysis is a potential area in urban drainage modeling, which had been overlooked, and provided hypothetical benchmark examples to illustrate their point.

Yeh and Labadie (1997) applied multiobjective genetic algorithm to generate nondominated solutions for system cost and detention effect for the Pazam catchment in Taiwan. Three multiobjective formulations were considered: VEGA, Pareto ranked GA (Goldberg, 1989), and MOGA (Fonseca and Fleming, 1993). The results were compared with those of a heuristic-based dynamic programming approach known as *successive reaching dynamic programming* (SRDP). They found that although SRDP was computationally more efficient, multiobjective GA had the ability of accommodating more than two objectives easily and hence was preferable.

On the other hand, the use of single-objective GA in urban drainage and wastewater system design is more common. Diogo *et al.* (2000) applied GA to optimize the design of a 3D urban drainage. Crossover and mutation operators were modified to produce dendritic or tree-like network layout. In Harrell and Ranjithan (2003), a simple GA was used for the design and siting of detention ponds for a water supply catchment in North Carolina, US. Since the problem was subjected to pollution loading constraints, the multiplicative penalty function was implemented in the GA.

In the area of wastewater system, Rauch and Harremoes (1999b) used a simple GA to optimize and control the performance of an urban wastewater system in order to minimize pollution. The authors were able to achieve good results with hypothetical problem sets and indicated that GA showed great potential in nonlinear model predictive controls for wastewater systems. Tait *et al.* (2003) discussed the presence of high level of suspended solids in sewer especially during alternating wet and dry flow period for a combined sewer system. They proposed a new process model for in-sewer deposit erosion in order to predict sewer flow quality and used a simple GA to calibrate the proposed model. A commercial software "Evolver" was chosen for the optimization, but a large number of evaluations were required in order to achieve good calibration results.

In a separate predictive modeling application, GP was used to model the dynamic performance of a wastewater treatment plant (Hong and Bhamidimarri, 2003). The predictive accuracy of GP was compared with ANN and a well-known IAWQ ASM2 model. The authors highlighted that the strength of the GP approach are: (i) the induced model structure is interpretable and (ii) the transparency of

the models allow further development/modification of the models for scenario generation and strategy development.

In order to reduce the computational cost of running the simulator for optimization, ANN is commonly used (as can be seen in previous sections). In the area of urban water systems, Sanchez *et al.* (1998) used an ANN as a surrogate model for the nonfulfillment time (exceedance time of fecal coliforms concentration limits) in a sewer system. The ANN was applied to emulate the sewer system for the metropolitan area of Gijon in Spain. The singular value decomposition method was used to optimize the ANN weights, and a simple GA was used to optimize the ANN structure/architecture.

EVOLUTIONARY COMPUTING IN WATER QUALITY RESEARCH

In the area of water quality research, GAs have been applied to optimize waste load allocation and ANN surrogate models, calibrate quality models, generate quality models, and model induction.

Waste load allocation refers to the process of determining the required pollutant removal levels at a number of either point or nonpoint sources to attain a satisfactory water quality response in a receiving water body. Burn and Yulianti (2001) applied a Pareto rank-based multiobjective GA to a waste load allocation problem in order to evaluate the trade-off between total treatment cost and to maintain various steady-state water quality conditions. Vasques *et al.* (2000) noted that Burn and Yulianti (2001) did not account for natural variability and parameter uncertainty and proposed to evaluate the model reliability instead. They used the first-order reliability method (FORM) to estimate the model reliability and a GA to generate the trade-off between cost and reliability estimate. Both studies were applied to a case study for the Willamette River basin in Oregon, US.

In the area of water quality model calibration, uncertainty of estimates was also taken into account in some research works. Chen and Chang (1998) used a simple GA coupled with a fuzzy objective function to calibrate a water quality model taking uncertainty into account. Their approach was applied to a case study of water quality management and control problem for the Tseng–Wen catchment in Taiwan. In Mulligan and Brown (1998), the potential use of GA for calibrating the steady-state Streeter–Phelps dissolved oxygen model for predicting DO concentration in streams was investigated. Synthetic data with and without error were used, and GA's performance was verified with a case study using the QUAL2E model applied to a river stretch in New England, US. The use of multiresponse as objective functions was also investigated and found to be superior to single response data. More recently, Ng and Perera (2003) used a simple GA to calibrate both QUAL2E model and the Australian Yarra River Water Quality model. They investigated the effect of different GA operator settings for calibration as well as for sensitivity analysis of the model parameters.

There are a number of applications of GA to optimize ANN surrogate water quality models. Neelakantan *et al.* (2002) employed ANN to predict the risk of protozoa (Cryptosporidium and Giardia) by relating risk to other biological, chemical, and physical parameters in surface waters. Different ANN training algorithms such as steepest descent, conjugate gradient, quadratic programming, and GA were compared. GA was found to be the more robust compared with the other training algorithms. In Wang and Jamieson (2002), ANN was employed to replicate a process-based water quality simulation model, TOMCAT (Bowden and Brown, 1983) and trained using a GA. Recently, Bowden *et al.* (2002) proposed a systematic approach for the optimal division of data for ANN models. They compared three different approaches of data division: arbitrary division, using GA, and using self-organizing feature maps (SOFM) and found that SOFMs were significantly better when these three methods were applied to an ANN which was used to forecast salinity in the River Murray, South Australia.

Bobbin and Recknagel (2001) used evolutionary strategy (ES) combined with rule-set generator to predict algal blooms in a Japanese lake. ES was used to optimize the parameters in the rule-set generator and to construct the rule-based model. The application of process-based model improvement and model induction was examined in Whigham and Rechnagel (2001). A GA was firstly used to calibrate the difference equation within the lake ecosystem model, SALMO, which predicts algal growth. GP was then applied to the model to evolve new expression for the different terms (such as photosynthesis, respiration, or grazing) in the difference equation. Initial findings were very encouraging and the authors were going to extend their work to process understanding of the evolved terms.

FURTHER READING

Chen L. (2003) Real coded genetic algorithm optimisation of long term reservoir operation. *Journal of American Water Resources Association*, **39**(5), 1157–1165.

Damas M., Salmeron M., Ortega J., Olivares G. and Pomares H. (2001) Parallel dynamic water supply scheduling in a cluster of computers. *Concurrency and Computation: Practice and Experience*, **13**(15), 1281–1302.

Dorn J.L. and Ranjithan S. (2003) Evolutionary multiobjective optimization in watershed water quality management. In *Evolutionary Multi-Criteria Optimization, Lecture Notes in Computer Science (LNCS)*, Fonseca C.M., Fleming P.J., Zitzler E., Deb K., Thiele L, *et al.* (Eds.), Springer-Verlag: Berlin, Vol. 2632, pp. 692–706.

Fonlupt C. (2001) Solving the ocean colour problem using a genetic programming approach. *Applied Soft Computing*, **1**(1), 63–72.

Franchini M. and Galeati G. (1997) Comparing several genetic algorithm schemes for the calibration of conceptual rainfall-runoff models. *Journal of Hydrological Sciences*, **42**(3), 357–379.

Franchini M., Galeati G. and Berra S. (1998) Global optimisation techniques for the calibration of conceptual rainfall-runoff models. *Journal of Hydrological Sciences*, **43**(3), 443–458.

Fujiwara O. and Khang D.B. (1990) A two-phase decomposition method for optimal design of looped water distribution networks. *Water Resources Research*, **26**(4), 539–549.

Huang W.-C., Yuan L.-C. and Lee C.-M. (2002) Linking genetic algorithm with stochastic dynamic programming to the long-term operation of a multi-reservoir system. *Water Resources Research*, **38**(12), 401–409.

Knaapen M.A.F. and Hulscher S.J.M.H. (2002) Regeneration of sand waves after dredging. *Coastal Engineering*, **46**(4), 277–289.

Knaapen M.A.F. and Hulscher S.J.M.H. (2003) Use of a genetic algorithm to improve predictions of alternative bar dynamics. *Water Resources Research*, **39**(3), 1231, doi:10.1029/2002WR001793.

Kuo S.-F., Liu C.-W. and Chen S.-K. (2003) Comparative study of optimisation techniques for irrigation project planning. *Journal of American Water Resources Association*, **39**(1), 59–73.

Lingireddy S. and Ormsbee L.E. (2002) Hydraulic network calibration using genetic optimization. *Civil Engineering and Environmental Systems*, **19**(1), 13–39.

Mayer A.S., Kelley C.T. and Miller C.T. (2002) Optimal design for problems involving flow and transport phenomena in saturated subsurface systems. *Advances in Water Resources*, **25**(8–12), 1233–1256.

Morelissen R., Hulscher S.J.M.H., Knaapen M.A.F., Nemeth A.A. and Bijker R. (2003) Mathematical modelling of sand wave migration and the interaction with pipelines. *Coastal Engineering*, **48**(3), 197–209.

Obregon N., Sivakumar B. and Puente C.E. (2002) A deterministic geometric representation of temporal rainfall: sensitivity analysis of a storm in Boston. *Journal of Hydrology*, **269**, 224–235.

Reis L.F.R., Porto R.M. and Chaudhry F.H. (1997) Optimal location of control valves in pipe networks by genetic algorithm. *Journal of Water Resources Planning and Management, ASCE*, **123**(6), 317–326.

Savic D. A. and Walters G. A. (1995a) Integration of a model for hydraulic analysis of water distribution networks with an evolution program for pressure regulation. *Microcomputers in Civil Engineering*, **10**(3), 219–229.

Savic D.A. and Walters G.A. (1995b) An evolution program for optimal pressure regulation in water distribution networks. *Engineering Optimization*, **24**(3), 197–219.

Schielen R.M.J., Doelman A. and de Swart H.E. (1993) On the nonlinear dynamics of free bars in straight channels. *Journal of Fluid Mechanics*, **252**, 325–356.

Sen Z. and Oztopal A. (2001) Genetic algorithms for the classification and prediction of precipitation occurrence. *Journal of Hydrological Sciences*, **46**(2), 255–267.

Sztobryn M. (2003) Forecast of storm surge by means of artificial neural network. *Journal of Sea Research*, **49**, 317–322.

Tung C.-P., Hsu S.-Y., Liu C.-M. and Li J.-S. (2003) Application of the genetic algorithm for optimising operation rules of the LiYuTan Reservoir in Taiwan. *Journal of American Water Resources Association*, **39**(3), 649–657.

Wang Q.J., Chiew F.H.S. and McMahon T.A. (1995) Calibration of environmental models by genetic algorithms, *Proceedings of the MODSIM 95, International Congress on Modelling and Simulation*, Newcastle, NSW, Australia, Vol. 3, pp. 185–190.

Yu P.-S. and Yang T.-C. (2000) Fuzzy multi-objective function for rainfall – runoff model calibration. *Journal of Hydrology*, **238**, 1–14.

REFERENCES

Abrahart R.J., See L. and Kneale P.E. (1999) Using pruning algorithms and genetic algorithms to optimise network architectures and forecasting inputs in a neural network rainfall-runoff model. *Journal of Hydroinformatics*, **1**(2), 103–114.

Aly A.H. and Peralta R.C. (1999a) Comparison of a genetic algorithm and mathematical programming to the design of groundwater cleanup systems. *Water Resources Research*, **35**(8), 2415–2425.

Aly A.H. and Peralta R.C. (1999b) Optimal design of aquifer cleanup systems under uncertainty using a neural network and a genetic algorithm. *Water Resources Research*, **35**(8), 2523–2532.

Aral M.M., Guan J. and Maslia M.L. (2001) Identification of contaminant source location and release history in aquifers. *Journal of Hydraulic Engineering, ASCE*, **6**(3), 225–234.

Babayan A., Savic D.A. and Walters G.A. (2004) Multiobjective Optimization of Water Distribution Systems Design under Uncertain Demands. In *The 6th International Conference on Hydroinformatics, Singapore*, Liong S.-Y., Phoon K.-K. and Babovic V. (Eds.), World Scientific Publishing: New Jersey, Vol. 1, pp. 906–913, 21–24 June.

Babayan A.V., Savic D.A. and Walters G.A. (2003) *Least-cost Design of Water Distribution Networks Under Uncertain Demand, Advances in Water Supply Management*, Maksimovic C., Butler D. and Memon F. (Eds.), A.A. Balkema Publishers: pp. 139–146.

Babovic V. and Abbott M.B. (1997) The evolution of equations from hydraulic data, Part II: Applications. *Journal of Hydraulic Research*, **35**, 411–430.

Baeck T., Fogel D.B. and Michalewicz Z. (1997) *Handbook of Evolutionary Computation*, IOP Publ. Co. & Oxford University Press.

Balascio C.C., Palmeri D.J. and Gao H. (1998) Use of a genetic algorithm and multi-objective programming for calibration of a hydrologic model. *Transactions of the ASAE*, **41**(3), 615–619.

Balla M.C. and Lingireddy S. (2000) Distributed genetic algorithm model on network of personal computers. *Journal of Computing in Civil Engineering, ASCE*, **14**(3), 199–205.

Bobbin J. and Recknagel F. (2001) Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms. *Ecological Modelling*, **146**, 253–262.

Bowden G.J., Maier H.R. and Dandy G.C. (2002) Optimal division of data for neural network models in water resources applications. *Water Resources Research*, **38**(2), 201–211.

Bowden K. and Brown S.R. (1983) Relating effluent control parameters to river quality objectives using a generalised catchment simulation model. *Water Science and Technology*, **16**, 22–31.

Burn D.H. and Yulianti J.S. (2001) Waste-load allocation using genetic algorithms. *Journal of Water Resources Planning and Management, ASCE*, **127**(2), 121–129.

Cai X., McKinney D.C. and Lasdon L.S. (2001) Solving nonlinear water management models using a combined genetic algorithm and linear programming approach. *Advances in Water Resources*, **24**, 667–676.

Chan Hilton A.B. and Culver T.B. (2000) Constraint handling for genetic algorithms in optimal remediation design. *Journal of Water Resources Planning and Management, ASCE*, **126**(3), 128–137.

Chan Hilton A.B. and Culver T.B. (2001) Sensitivity of optimal groundwater remediation designs to residual water quality violations. *Journal of Water Resources Planning and Management, ASCE*, **127**(5), 316–323.

Chang F.-J. and Chen L. (1998) Real-coded genetic algorithm for rule-based flood control reservoir management. *Water Resources Management*, **12**, 185–198.

Chang F.-J., Lai J.-S. and Kao L.-S. (2003) Optimisation of operation rule curves and flushing schedule in a reservoir. *Hydrological Processes*, **17**, 1623–1640.

Chang L.-C. and Chang F.-J. (2001) Intelligent control for modelling of real-time reservoir operation. *Hydrological Processes*, **15**, 1621–1634.

Chaves P., Kojiri T. and Yamashiki Y. (2003) Optimisation of storage reservoir considering water quantity and quality. *Hydrological Processes*, **17**, 2769–2793.

Chen H.S. and Chang N.-B. (1998) Water pollution control in the river basin by fuzzy genetic algorithm-based multiobjective programming modelling. *Water Science and Technology*, **37**(8), 55–63.

Chen Y.-M. (1997) Management of water resources using improved genetic algorithms. *Computers and Electronics in Agriculture*, **18**, 117–127.

Cheng C.T., Ou C.P. and Chau K.W. (2002) Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall-runoff model calibration. *Journal of Hydrology*, **268**, 72–86.

Cieniawski S.E., Eheart J.W. and Ranjithan S. (1995) Using genetic algorithms to solve a multiobjective groundwater monitoring problem. *Water Resources Research*, **31**(2), 399–409.

Cluckie I.D., Lane A. and Yuan J. (1999) Modelling large urban drainage systems in real-time. *Water Science and Technology*, **39**(4), 21–28.

Cooper V.A., Nguyen V.T.V. and Nicell J.A. (1997) Evaluation of global optimization methods for conceptual rainfall-runoff model calibration. *Water Science and Technology*, **36**(5), 53–60.

Cui L.-J. and Kuczera G. (2003) Optimising urban water supply headworks using probabilistic search methods. *Journal of Water Resources Planning and Management, ASCE*, **129**(5), 380–387.

Cunha M.C. (2002) Groundwater cleanup: the optimisation perspective (a literature review). *Engineering Optimization*, **34**(6), 689–702.

Dandy G.C. and Engelhart M. (2001) Optimal scheduling of water pipe replacement using genetic algorithms. *Journal of Water Resources Planning and Management, ASCE*, **127**(4), 214–223.

Dandy G.C., Simpson A.R. and Murphy L.J. (1996) An improved genetic algorithm for pipe network optimisation. *Water Resources Research*, **32**(2), 449–458.

Davidson J.W. (1999) Evolution program for layout geometry of rectilinear looped networks. *Journal of Computing in Civil Engineering, ASCE*, **13**(4), 246–253.

Davidson J.W., Savic D.A. and Walters G.A. (2003) Symbolic and numerical regression: experiments and applications. *Information Sciences*, **150**, 95–117.

de Schaetzen W.B.F., Walters G.A. and Savic D.A. (2000) Optimal sampling design for model calibration using shortest path, genetic and entropy algorithms. *Urban Water*, **2**, 141–152.

Deb K. (2001) *Multi-objective Optimisation Using Evolutionary Algorithms*, John Wiley: Hoboken.

Diogo A.F., Walters G.A., de Sousa E.R. and Graveto V.M. (2000) Three-dimensional optimisation of urban drainage systems. *Computer Aided Civil and Infrastructure Engineering*, **15**, 409–426.

Dorado J., Rabunal J.R., Pazos A., Rivero D., Santos A. and Puertas J. (2003) Prediction and modelling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. *Applied Artificial Intelligence*, **17**(4), 329–343.

Duan Q., Sorooshian S. and Gupta V.K. (1992) Effective and efficient global optimisation for conceptual rainfall-runoff models. *Water Resources Research*, **24**(7), 1163–1173.

Eberhart R.C. and Kennedy J. (1995) A new optimiser using particles swarm theory. *Proceedings of the 6th International Symposium On Micro Machine and Human Science*, IEEE Service Centre: Piscataway, pp. 39–43.

El Harrouni K., Ouazar D., Walters G.A. and Cheng A.H.-D. (1996) Groundwater optimisation and parameter estimation by genetic algorithm and dual reciprocity boundary element method. *Engineering Analysis with Boundary Elements*, **19**, 287–296.

Erickson M., Mayer A. and Horn J. (2002) Multi-objective optimal design of groundwater remediation systems: application of the niched Pareto genetic algorithm (NPGA). *Advances in Water Resources*, **25**(1), 51–65.

Esat V. and Hall M.J. (1994) Water resources system optimisation using genetic algorithms. *Proceedings of the 1st International Conference on Hydroinformatics*, Balkema: Rotterdam, pp. 225–231.

Eusuff M.M. and Lansey K.E. (2003) Optimisation of water distribution network design using the shuffled frong leaping algorithm. *Journal of Water Resources Planning and Management, ASCE*, **129**(3), 210–225.

Fedra K. and Jamieson D.G. (1996) The 'WaterWare' decision-support system for river basin planning 2: planning capability. *Journal of Hydrology*, **177**, 177–198.

Fogel L.J., Owens A.J. and Walsh M.J. (1966) *Artificial Intelligence through Simulated Evolution*, John Wiley: New York.

Fonseca C.M. and Fleming P.J. (1993) Genetic algorithms for multi-objective optimisation; optimisation, formulation discussion and generalization, *Proceedings of the 5th International Conference on Genetic Algorithms and Their Application*, Illinois at Urbana-Champaign, pp. 416–423.

Franchini M. (1996) Use of a genetic algorithm combined with a local search method for the automatic calibration of conceptual rainfall-runoff models. *Journal of Hydrological Sciences*, **41**(1), 21–39.

Gentry R.W., Camp C.V. and Anderson J.L. (2001) Use of GA to determine areas of accretion to semiconfined aquifer. *Journal of Hydraulic Engineering, ASCE*, **127**(9), 738–746.

Gentry R.W., Larsen D. and Ivey S. (2003) Efficacy of genetic algorithm to investigate small scale aquitard leakage. *Journal of Hydraulic Engineering, ASCE*, **129**(7), 527–535.

Gessler J. (1985) Pipe network optimization by enumeration. *Proceedings of the Computer Applications for Water Resources*, ASCE: New York, pp. 572–581.

Giacobbo F., Marseguerra M. and Zio E. (2002) Solving the inverse problem of parameter estimation by genetic algorithms: the case of a groundwater contaminant transport model. *Annals of Nuclear Energy*, **29**, 967–981.

Goldberg D.E. (1989) *Genetic Algorithms in Search, Optimisation and Machine Learning*, Addison-Wesley: Reading.

Goldberg D.E., Korb B. and Deb K. (1989) Messy genetic algorithms: motivation, analysis and first results. *Complex System*, **3**, 493–530.

Goldberg D.E. and Kuo C.H. (1987) Genetic algorithms in pipeline optimization. *Journal of Computing in Civil Engineering, ASCE*, **1**(2), 128–141.

Goulter I. and Bouchart F. (1990) Reliability-constrained pipe network model. *Journal of Hydraulic Engineering, ASCE*, **116**(2), 211–229.

Guan J. and Aral M.M. (1999a) Optimal remediation with well locations and pumping rates selected as continuous decision variables. *Journal of Hydrology*, **221**, 20–42.

Guan J. and Aral M.M. (1999b) Progressive genetic algorithm for solution of optimisation problems with nonlinear equality and inequality constraints. *Applied Mathematical Modeling*, **23**, 329–343.

Gumrah F., Erbas D., Oz B. and Altintas S. (2000) Genetic algorithms for optimising the remediation of contaminated aquifer. *Transport in Porous Media*, **41**, 149–171.

Gupta I., Gupta A. and Khanna P. (1999) Genetic algorithm for optimisation of water distribution systems. *Environmental Modelling and Software*, **14**, 437–446.

Gupta R. and Bhave R. (1994) Reliability analysis of water distribution systems. *Journal of Environmental Engineering, ASCE*, **120**(2), 447–460.

Halhal D., Walters G.A., Ouazar D. and Savic D.A. (1997) Water network rehabilitation with structured messy genetic algorithm. *Journal of Water Resources Planning and Management, ASCE*, **123**(3), 137–146.

Harrell L.J. and Ranjithan S.R. (2003) Detention pond design and land use planning for watershed management. *Journal of Water Resources Planning and Management, ASCE*, **129**(2), 98–106.

Harris E.L., Babovic V. and Falconer R.A. (2003) Velocity predictions in compound channels with vegetated floodplains using genetic programming. *International Journal of River Basin Management*, **1**(2), 117–123.

Haupt R.L. and Haupt S.E. (1998) *Practical Genetic Algorithms*, Wiley & Sons: New York.

Heidari M. and Ranjithan S.R. (1998) Hybrid optimisation approach to the estimation of distributed parameters in two-dimensional confined aquifers. *Journal of American Water Resources Association*, **34**(4), 909–920.

Henrickson J.D., Sorooshian S. and Brazil L. (1988) Comparison of Newton type and direct search algorithm for calibration of conceptual rainfall-runoff models. *Water Resources Research*, **24**(5), 691–700.

Holland J.H. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press: Ann Arbor, p. 1975.

Hong Y.-S. and Bhamidimarri R. (2003) Evolutionary self-organising modelling of a municipal wastewater treatment plant. *Water Research*, **37**, 1199–1212.

Hong Y.-S. and Rosen M.R. (2002) Identification of an urban fractured-rock aquifer dynamics using an evolutionary self-organising modelling. *Journal of Hydrology*, **259**(1–4), 89–104.

Howson H.R. and Sancho N.G.F. (1975) A new algorithm for the solution of multi-state dynamic programming problems. *Mathematical Programming*, **8**, 104–116.

Hsiao C.-T. and Chang L.-C. (2002) Dynamic optimal groundwater management with inclusion of fixed costs. *Journal of Water Resources Planning and Management, ASCE*, **128**(1), 57–65.

Huang C. and Mayer A.S. (1997) Pump and treat optimisation using well locations and pumping rates as decision variables. *Water Resources Research*, **33**(5), 1001–1012.

Ines A.V.M. and Droogers P. (2002) Inverse modelling in estimating soil hydraulic functions: a genetic algorithm approach. *Hydrology and Earth System Sciences*, **6**(1), 49–65.

Kapelan Z., Savic D.A. and Walters G.A. (2004) A multiobjective approach to rehabilitation of water distribution networks under uncertainty. *Proceedings of the 6th International Symposium on Systems Analysis and Integration Assessment*, WATERMATEX, IWA: Beijing, Nov. 3–5, 2004.

Kapelan Z.S., Savic D.A. and Walters G.A. (2003a) A hybrid inverse transient model for leakage detection and roughness calibration in pipe networks. *Journal of Hydraulic Research*, **41**(5), 481–492.

Kapelan Z.S., Savic D.A. and Walters G.A. (2003b) Multiobjective sampling design for water distribution model calibration. *Journal of Water Resources Planning and Management, ASCE*, **129**(6), 466–479.

Kapelan Z.S., Savic D.A. and Walters G.A. (2003c) *Robust Least Cost Design of Water Distribution Systems Using GAs, Advances in Water Supply Management*, Maksimovic C., Butler D. and Memon F. (Eds.), A.A. Balkema Publishers: pp. 147–155.

Karpouzos D.K., Delay F., Katsifarakis K.L. and de Marsily G. (2001) A multipopulation genetic algorithm to solve the inverse problem in hydrogeology. *Water Resources Research*, **37**(9), 2291–2302.

Katsifarakis K.L., Karpouzos D.K. and Theodossiou N. (1999) Combined use of BEM and genetic algorithm in groundwater

flow and mass transport problems. *Engineering Analysis with Boundary Elements*, **23**, 555–565.

Khu S.T., Liong S.Y., Babovic V., Madsen H. and Muttil N. (2001) Genetic programming and its application in real-time flood forecasting. *J. American Water Resources Assoc.*, **36**(2), 439–452.

Kisi O. (2004) Daily real-time suspended sediment modelling using hybrid fuzzy differential evolution approach. *Journal of Hydrological Sciences*, **49**(1), 183–197.

Koza J.R. (1992) *Genetic Programming: On The Programming Of Computers By Means Of Natural Selection*, MIT Press.

Krishnakumar K. (1989) *Micro-genetic Algorithms for Stationary and Non-stationary Function Optimization*. SPIE: Intelligent Control and Adaptive Systems, Vol. 1196, Philadelphia.

Kuo S.-F., Merkley G.P. and Liu C.-W. (2000) Decision support for irrigation project planning using a genetic algorithm. *Agricultural Water Management*, **45**, 243–266.

Kuzcera G. (1997) Efficient subspace probabilistic parameter optimisation for catchment models. *Water Resources Research*, **33**(1), 177–185.

Lee T.-C., Perina T. and Lee C.-Y. (2002) Validation of aquifer parameter determination by extrapolation fitting and treating thickness as an unknown. *Journal of Hydrology*, **265**(1), 15–33.

Leitao T.E. (1998) A proposal for a decision flow chart for the selection of technologies for rehabilitation of polluted aquifers. In *Environmental Contamination and Remediation Practices at Former and Present Military Bases*, Fonnum F., *et al.* (Eds.), Kluwer Academic Publishers.

Lingireddy S. (1998) Aquifer parameter estimation using genetic algorithms and neural networks. *Civil Engineering and Environmental Systems*, **15**, 125–144.

Liong S.Y., Chan W.T. and ShreeRam J. (1995) Peak-flow forecasting with genetic algorithm and SWMM. *Journal of Hydraulic Engineering, ASCE*, **121**(8), 613–617.

Liong S.Y., Gautum T.R., Khu S.T., Babovic V., Keijzer M. and Muttil N. (2002) Genetic programming: a new paradigm in rainfall runoff modelling. *Journal of American Water Resources Association*, **38**(3), 705–718.

Liong S.Y., Khu S.T. and Chan W.T. (2001) Derivation of Pareto front with genetic algorithm and neural network. *Journal of Hydraulic Engineering, ASCE*, **6**(1), 52–61.

Lippai I., Heaney J.P. and Laguna M. (1999) Robust water system design with commercial intelligent search optimizers. *Journal of Computing in Civil Engineering, ASCE*, **13**(3), 135–143.

Maskey S., Jonoski A. and Solomatine D.P. (2002) Groundwater remediation strategy using global optimisation algorithms. *Journal of Water Resources Planning and Management, ASCE*, **128**(6), 431–440.

McKinney D.C. and Lin M.D. (1994) Genetic algorithm solution of groundwater management models. *Water Resources Research*, **30**(6), 1897–1906.

Meier R.W. and Barkdoll B.D. (2000) Sampling design for network model calibration using genetic algorithms. *Journal of Water Resources Planning and Management, ASCE*, **126**(4), 245–250.

Merabtene T., Kawamura A., Jinno K. and Olsson J. (2002) Risk assessment for optimal drought management of an integrated water resources system using a genetic algorithm. *Hydrological Processes*, **16**, 2189–2208.

Miles J.C., Peggs T. and Moore C.J. (2002) Deriving reservoir operating policies using classifier systems. *Civil Engineering and Environmental Systems*, **19**(4), 285–310.

Mohan S. (1997) Parameter estimation of nonlinear Muskingum models using genetic algorithm. *Journal of Hydraulic Engineering, ASCE*, **123**(2), 137–142.

Montesinos P., Camacho E. and Alvarez S. (2001) Seasonal furrow irrigation model with genetic algorithms (OPTIMEC). *Agricultural Water Management*, **52**, 1–16.

Montesinos P., Garcia-Guzman A. and Ayuso J.L. (1999) Water distribution network optimisation using a modified genetic algorithm. *Water Resources Research*, **35**(11), 3467–3473.

Morley M.S., Atkinson R.M., Savic D.A. and Walters G.A. (2001) GA *net*: genetic algorithm platform for pipe network optimisation. *Advances in Engineering Software*, **32**, 467–475.

Morshed J. and Kaluarachchi J.J. (1998a) Application of artificial neural network and genetic algorithm in flow and transport simulations. *Advances in Water Resources*, **22**(2), 145–158.

Morshed J. and Kaluarachchi J.J. (1998b) Parameter estimation using artificial neural network and genetic algorithm for free product migration. *Water Resources Research*, **34**(5), 1101–1113.

Morshed J. and Kaluarachchi J.J. (2000) Enhancements to genetic algorithm for optimal groundwater management. *Journal of Hydraulic Engineering, ASCE*, **5**(1), 67–73.

Mulligan A.E. and Brown L.C. (1998) Genetic algorithms for calibrating water quality models. *Journal of Environmental Engineering, ASCE*, **124**(3), 202–211.

Munavalli G.R. and Mohan Kumar M.S. (2003) Optimal scheduling of multiple chlorine sources in water distribution systems. *Journal of Water Resources Planning and Management, ASCE*, **129**(6), 493–504.

Murray D.M. and Yakowitz S.J. (1979) Constrained differential dynamic programming and its application to multireservoir control. *Water Resources Research*, **15**(5), 1017–1027.

Ndiritu J.G. and Daniell T.M. (1999) Assessing model calibration adequacy via global optimisation. *Water SA*, **25**(2), 317–326.

Ndiritu J.G. and Daniell T.M. (2001) An improved genetic algorithm for rainfall-runoff model calibration and function optimisation. *Mathematical and Computer Modelling*, **33**, 695–705.

Neelakantan T.R., Lingireddy S. and Brion G.M. (2002) Effectiveness of different artificial neural network training algorithms in predicting protozoa risks in surface waters. *Journal of Environmental Engineering, ASCE*, **128**(6), 533–542.

Nelder J.A. and Mead R. (1965) A simplex method for function optimization. *Computer Journal*, **7**, 308–313.

Ng A.W.M. and Perera B.J.C. (2003) Selection of genetic algorithm operators for river water quality model calibration. *Engineering Applications of Artificial Intelligence*, **16**, 529–541.

Nicklow J.W., Ozkurt O. and Bringer J.A. Jr. (2003) Control of channel bed morphology in large-scale river network suing a genetic algorithm. *Water Resources Management*, **17**, 113–132.

Nixon J.B., Dandy G.C. and Simpson A.R. (2001) A genetic algorithm for optimizing off-farm irrigation scheduling. *Journal of Hydroinformatics*, **3**(1), 11–22.

Oliveira R. and Loucks D.P. (1997) Operating rules for multi-reservoir systems. *Water Resources Research*, **33**(4), 839–852.

Ponnambalam K., Karray F. and Mousavi S. (2002) Optimisation approaches for reservoir systems operation using computational intelligence tools. *Systems Analysis Modelling Simulation*, **42**(9), 1347–1360.

Ponnambalam K., Karray F. and Mousavi S. (2003) Minimizing variance of reservoir systems operations benefits using soft computing tools. *Fuzzy Sets and Systems*, **139**, 451–461.

Prasad K.L. and Rastogi A.K. (2001) Estimating net aquifer recharge and zonal hydraulic conductivity values for Mahi Right Bank canal project area, India by genetic algorithm. *Journal of Hydrology*, **243**(3–4), 149–161.

Rauch W. and Harremoes P. (1999a) On the potential of genetic algorithms in urban drainage modelling. *Urban Water*, **1**, 79–89.

Rauch W. and Harremoes P. (1999b) Genetic algorithms in real time control applied to minimise transient pollution from urban wastewater systems. *Water Research*, **33**(5), 1265–1277.

Reed P., Minsker B. and Goldberg D.E. (2001) A multi-objective approach to cost effective long-term groundwater monitoring using an elitist non-dominated sorted genetic algorithm with historical data. *Journal of Hydroinformatics*, **3**(2), 71–89.

Reed P., Minsker B. and Goldberg D.E. (2003) Simplifying multi-objective optimisation: an automated design methodology for the nondominated sorted genetic algorithm-II. *Water Resources Research*, **39**(7), Technical Note, TNN2-1-5.

Reed P., Minsker B. and Volocchi A.J. (2000) Cost – effective long-term groundwater monitoring design using a genetic algorithm and global mass interpolation. *Water Resources Research*, **36**(12), 3731–3741.

Ritzel B.J. and Eheart J.W. (1994) Using genetic algorithms to solve a multiple objective groundwater pollution containment problem. *Water Resources Research*, **30**(5), 1589–1603.

Rogers L.L. and Dowla F.U. (1994) Optimization of groundwater remediation using artificial neural networks with parallel solute transport modelling. *Water Resources Research*, **30**(2), 457–481.

Rogers L.L., Dowla F.U. and Johnson V.M. (1995) Optimal field-scale groundwater remediation using neural networks and the genetic algorithm. *Environmental Science and Technology*, **29**(5), 1145–1156.

Rosenbrock H.H. (1960) An automatic method for finding the greatest or least value of a function. *Computer Journal*, **3**, 175–184.

Rowell W.F. and Barnes J.W. (1982) Obtaining the layout of water distribution systems. *Journal of Hydraulics Division, ASCE*, **108**(1), 137–148.

Sanchez L., Arroyo V., Garcia J., Koev K. and Revilla J. (1998) Use of neural networks in design of coastal sewage systems. *Journal of Hydraulic Engineering, ASCE*, **124**(5), 457–464.

Savic D.A. (2002) Single-objective vs. multiobjective optimisation for integrated decision support. In *Integrated Assessment and Decision Support, Proceedings of the First Biennial Meeting of the International Environmental Modelling and Software Society*, Rizzoli A.E. and Jakeman A.J. (Eds.), iEMSs: Manno, Switzerland, ISBN 88-900787-0-7, pp. 7–12.

Savic D.A. and Walters G.A. (1997) Genetic algorithms for least-cost design of water distribution networks. *Journal of Water Resources Planning and Management*, **123**(2), 67–77.

Savic D.A., Walters G.A. and Davidson J.W. (1999) A genetic programming approach to rainfall-runoff modelling. *Water Resources Management*, **13**, 219–231.

Schwefel H.P. (1981) *Numerical Optimisation of Computer Models*, John Wiley: Chichester.

See L. and Abrahart R.J. (2001) Multi-model data fusion for hydrological forecasting. *Computers and Geosciences*, **27**, 987–994.

Seibert J. (2000) Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, **4**(2), 215–224.

Seibert J., Uhlenbrook S., Leibundgut C. and Halldin S. (2000) Multiscale calibration and validation of a conceptual rainfall-runoff model. *Physics and Chemistry of the Earth (B)*, **25**(1), 59–64.

Senbeta D.A., Shamseldin A.Y. and O'Connor K.M. (1999) Modification of the probability distributed interacting storage capacity model. *Journal of Hydrology*, **224**, 149–168.

Sharif M. and Wardlaw R. (2000) Multireservoir systems optimisation using genetic algorithms: case study. *Journal of Computing in Civil Engineering, ASCE*, **14**(4), 255–263.

Simpson A.R., Dandy G.C. and Murphy L.J. (1994) Genetic algorithms compared to other techniques for pipe optimisation. *Journal of Water Resources Planning and Management, ASCE*, **120**(4), 423–443.

Smalley J.B., Minsker B.S. and Goldberg D.E. (2000) Risk-based in-situ bioremediation design using a noisy genetic algorithm. *Water Resources Research*, **36**(10), 3043–3052.

Smith D.K. and Walters G.A. (2000) An evolutionary approach for finding optimal trees in undirected networks. *European Journal of Operational Research*, **120**, 593–602.

Solomatine D.P., Dibike Y.B. and Kukuric N. (1999) Automatic calibration of groundwater models using global optimisation techniques. *Journal of Hydrological Sciences*, **44**(6), 879–894.

Storn R. and Price K. (1997) Differential evolution - a simple and efficient heuristic for global optimisation over continuous spaces. *Journal of Global Optimization*, **11**, 341–359.

Tait S.J., Chebbo G., Skipworth P.J., Ahyerre M. and Saul A.J. (2003) Modelling in-sewer deposit erosion to predict sewer flow quality. *Journal of Hydraulic Engineering, ASCE*, **129**(4), 316–324.

Tang A. and Mays L.W. (1998) Genetic algorithms for optimal operation of soil aquifer treatment systems. *Water Resources Management*, **12**, 375–396.

Todini E. (2000) Looped water distribution networks design using a resilience index based heuristic approach. *Urban Water*, **2**, 115–122.

Tsai F.T.C., Sun N.-Z. and Yeh W.W.G. (2003) A combinatorial optimisation scheme for parameter structure identification in groundwater modelling. *Ground Water*, **41**(2), 156–169.

Vairavamoorthy K. and Ali M. (2000) Optimal design of water distribution systems using genetic algorithms. *Computer Aided Civil and Infrastructure Engineering*, **15**, 374–382.

Vasques J.A., Mailer H.R., Lence B.J., Tolson B.A. and Foschi R.O. (2000) Achieving water quality system reliability

using genetic algorithms. *Journal of Environmental Engineering, ASCE*, **126**(10), 954–962.

Vink K. and Schot P. (2002) Multiple-objective optimisation of drinking water production strategies using a genetic algorithm. *Water Resources Research*, **38**(9), 1181, doi:10.1029/2000WR000034.

Vitkovsky J.P. and Simpson A.R. (1997) *Calibration and Leak Detection in Pipe Networks Using Inverse Transient Analysis and Genetic Algorithms*, Department of Civil and Environmental Engineering, University of Adelaide: Adelaide, p. 97.

Vrugt J.A., van Wijk M.T., Hopmans J.W. and Simunek J. (2001) One-, two, and three-dimensional root water uptake function for transient modelling. *Water Resources Research*, **37**(10), 2457–2470.

Wagner B.J. (1995) Sampling design methods for groundwater modelling under uncertainty. *Water Resources Research*, **31**(10), 2581–2591.

Walters G.A., Halhal D., Savic D.A. and Ouazar D. (1999) Improved design of "Anytown" distribution network using structured messy genetic algorithms. *Urban Water*, **1**(1), 23–38.

Walters G.A. and Lohbeck T. (1993) Optimal layout of tree networks using genetic algorithms. *Engineering Optimization*, **22**(1), 27–48.

Wang C.G. and Jamieson D.G. (2002) An objective approach to regional wastewater treatment planning. *Water Resources Research*, **38**(3), 1022, 10.1029/2000WR000062.

Wang M. and Zheng C. (1997) Optimal remediation policy selection under general conditions. *Ground Water*, **35**(5), 757–764.

Wang M. and Zheng C. (1998) Groundwater management optimisation using genetic algorithms and simulated annealing: formulation and comparison. *Journal of American Water Resources Association*, **34**(3), 519–530.

Wang Q.J. (1991) The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resources Research*, **27**(9), 2467–2471.

Wang Q.J. (1997) Using genetic algorithms to optimise model parameters. *Environmental Modelling and Software*, **12**(1), 27–34.

Wardlaw R. and Sharif M. (1999) Evaluation of genetic algorithms for optimal reservoir system operation. *Journal of Water Resources Planning and Management, ASCE*, **125**(1), 25–33.

Whigham P.A. and Crapper P.F. (2001) Modelling rainfall-runoff using genetic programming. *Mathematical and Computer Modelling*, **33**, 707–721.

Whigham P.A. and Rechnagel F. (2001) Predicting chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling*, **146**, 243–251.

Wu Z.Y. and Simpson A.R. (2001) Competent genetic-evolutionary optimisation of water distribution systems. *Journal of Computing in Civil Engineering, ASCE*, **15**(2), 89–101.

Wu Z.Y. and Simpson A.R. (2002) A self-adaptive boundary search genetic algorithm and its application to water distribution systems. *Journal of Hydraulic Research*, **40**(2), 191–203.

Yeh C.-H. and Labadie J.W. (1997) Multiobjective watershed-level planning of storm water detention systems. *Journal of Water Resources Planning and Management, ASCE*, **123**(6), 336–343.

Yoo C., Jung K.-S. and Ahn J.H. (2003) Estimating characteristics of rainfall and their effects on sampling schemes: case study for Han River Basin, Korea. *Journal of Hydraulic Engineering, ASCE*, **8**(3), 145–157.

Yoon J.-H. and Shoemaker C.A. (1999) Comparison of optimization methods for groundwater bioremediation. *Journal of Water Resources Planning and Management, ASCE*, **125**(1), 54–63.

Yoon J.-H. and Shoemaker C.A. (2001) Improved real-coded GA for groundwater remediation. *Journal of Computing in Civil Engineering, ASCE*, **15**(3), 224–231.

Yuan J.M., Tilford K.A., Jiang H.Y. and Cluckie I.D. (1999) Real-time urban drainage system modelling using weather radar rainfall data. *Physics and Chemistry of the Earth Part B*, **24**(8), 915–919.

Zeigler B.P., Moon Y., Lopes V.L. and Kim J. (1996) DEVS approximation of infiltration using genetic algorithm optimisation of a fuzzy system. *Mathematical and Computer Modelling*, **23**(11–12), 215–228.

23: Flood Early Warning Systems for Hydrological (sub) Catchments

MICHA GUIDO FRANCISCUS WERNER, JAAP SCHELLEKENS AND JACOB CORNELIS JAN KWADIJK

WL Delft Hydraulics, Delft, The Netherlands

Provision of early warning is an effective strategy in reducing flood damage and loss of life due to flooding. Flood forecasting and warning systems are important instruments in supporting relevant authorities issuing appropriate warnings. With the advent of computer-based simulation techniques, computer-based data processing capabilities, and computer-based communication facilities, the impact of hydroinformatics on flood forecasting in support of flood warning is manifest. In this article, an overview is given of flood forecasting systems for providing warnings in hydrological catchments. This overview is given from the perspective of the position of flood forecasting within the process of detection, forecasting, warning, and response. A categorization of flood forecasting systems is introduced, and the elements of a flood forecasting system are discussed.

INTRODUCTION

Floods pose a major natural hazard, and extensive losses from major and minor flood events are unfortunately commonplace. Recent flood events such as in the Rhine basin in 1995, the Mississippi in 1993, the Elbe basin in 2002, and the Yangtze in 1998 have firmly established the realization that natural hazards such as those posed by flooding cannot be brought under control simply through engineering measures. While the hazard cannot be avoided, mitigation of the consequences can help significantly reduce the risk posed. A thorough understanding of how flood events develop is fundamental to help mitigate these, and the development of this understanding is an important topic of much hydrological research. In the operational context, however, effective mitigation of the consequences of flooding can be provided through the provision of effective flood warning (Krzysztofowicz *et al.*, 1992; Parker and Fordham, 1996; Haggett, 1998; Penning-Rowsell *et al.*, 2000; De Roo *et al.* 2003).

The basic objective of the development of operational flood warning systems is to provide timely warning such that loss of property and life can be reduced, if not avoided completely, through implementation of simple, yet effective measures such as evacuation, temporary relocation, or implementation of a flood control strategy (Carpenter *et al.*, 1999). An operational flood warning system combined with these simple measures provide a much more cost-effective way of reducing flood risk and may help avoid large investments in traditional engineering flood control measures such as the raising of dykes or the building of flood control dams (Krzysztofowicz *et al.*, 1992). To be effective, the ability to provide timely warning must be complemented by awareness and preparedness by those at risk. A good example of how losses can be reduced was shown in the Meuse basin. Financial losses due to flooding in the Netherlands during the event of December 1993 greatly exceeded those of the event of January 1995, despite the two events being very comparable in magnitude. The reduction was partly attributed to the provision of early warning and the subsequent response (Wind *et al.*, 1999).

Operational flood warning systems have been developed or are under development in several river basins both in Europe (e.g. Parker and Fordham, 1996; Sprokkereef, 2001; Bürgi, 2002) and overseas (e.g. Grijsen *et al.*, 1992; Burnash, 1995; Jørgensen and Høst-Madsen, 1997; Du Plessis, 2002). All of these rely on the detection of floods through hydrometeorological observation networks, and the use of observation data is a primary element of a flood warning system. To increase the potential utility of the flood

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

warning service through extension of the lead time with which a flood event can be predicted, the state-of-the-art systems also incorporate some form of (model based) flood forecasting (Parker and Fordham, 1996). A wide range of techniques may be employed to provide flood forecasting capabilities, ranging from data-driven modeling techniques (e.g. Zealand *et al.*, 1999; Laio *et al.*, 2003; Young, 2003), through conceptual modeling approaches (e.g. Bürgi, 2002), to more complex networks of conceptual and physical models (e.g. Moore and Jones, 1998; Grijsen *et al.*, 1992, Van Kalken *et al.*, 2004).

How appropriate different types of techniques are is very much determined by the given requirement of delivering a forecast for a forecasting point at a minimum lead time. Lead time in the context used here defines how much warning can reliably be given of imminent flooding. The lead time at which effective warning can be delivered is clearly dependent on the lag times between precipitation falling and the flood peak reaching the point of interest. Warning lead times, therefore, vary greatly even in a single basin; for example, in the Rhine catchment warning lead times in the upper basin may only be in the order of 24 h (Bürgi, 2002), while in the lower basin, forecast lead times of up to four days can be provided using only hydrometeorological observations and a rainfall-runoff and routing model (Sprokkereef, 2001). The requirement to provide effective warning, together with the objective of increasing lead time and accuracy, poses new challenges in the implementation of flood warning systems.

This article explores flood forecasting systems from the perspective of its position within the flood warning process. A method for classifying the different approaches taken in flood forecasting is introduced before the elements of a present-day flood forecasting system are discussed in detail. Finally, the state of the art in developing flood forecasting systems is addressed including how to deal with specific challenges posed.

FLOOD FORECASTING AS A PART OF THE FLOOD WARNING PROCESS

The diversity of flood forecasting systems in operation or proposed in the literature is great. Flood warning systems are typically tailor-made to suit the specific requirements for the location(s) for which the warnings are to be provided, ranging from fast-responding local warning systems in the headwaters of a river (e.g. Krzysztofowicz *et al.*, 1992) or urban areas (e.g. Koussis *et al.*, 2003), to flood warning systems for lower reaches of large river basins (e.g. Sprokkereef, 2001). Often, flood warning systems are developed to cover all rivers within an administrative boundary, depending on the responsibilities of the authorities whose obligation it is to deliver the forecast (e.g. Moore *et al.*, 1990). The structure of the system in adjacent regions or countries may be very different (compare for example Moore *et al.* (1990) to Dobson and Davies (1990), who describe forecasting systems used for two adjacent regions in the United Kingdom). In some cases, differences may even occur within the same river system, such as in the Rhine basin. No less than 25 flood forecasting and/or flood warning centers are involved from the source to mouth, each with a different approach to flood forecasting (Steinbach and Wilke, 2000).

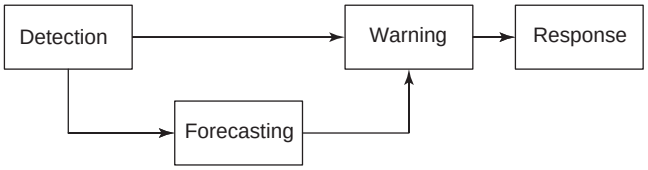


Figure 1 Stages in the delivery of effective flood warning

Despite the apparent variety in approaches, the elements on which the effective flood warning depends is an optimal combination of activities as depicted in Figure 1 (Parker and Fordham, 1996; Haggett, 1998);

1. *Detection*. In the detection stage, real-time data on processes that could generate a flood event are monitored. This includes primarily, monitoring of hydrological and meteorological conditions in the catchment through online information gathered through telemetry systems, climate stations, weather radar, and so on. The detection stage may also include techniques such as storm tracking and nowcasting (e.g. weather radar propagation), though this could also be thought to be a part of the forecasting stage.
2. *Forecasting*. In the forecasting stage, predictions are made of levels and flows, as well as the time of occurrence of possible forthcoming flood events. Typically, this involves the use of hydrological models, driven by both the real-time data gathered in the detection phase and forecasts of meteorological conditions such as rainfall and temperature. These are often obtained through external meteorological forecasts.
3. *Warning and dissemination*. The warning stage is a key factor in the success of operational flood warning. Using information derived from the detection and forecasting stages, the decision to warn appropriate authorities and/or properties at risk must be taken. The warning must be such that it gives an unambiguous message on the imminent flood potential.
4. *Response*. Response to flood warnings issued is vital for achieving the aims of operational flood warning. If the objective is to reduce damage through flood preparedness, an appropriate response by relevant authorities and affected persons must be taken following a warning, if it is to be realized.

Although the flood forecasting stage is an important part of the flood warning process, the most basic warning

systems do not include an explicit forecasting step and issue flood warnings on the basis of observations such as gauged rainfall and flows, combined with the judgment and experience of the forecasters (Cluckie, 2000). The detection step, together with the warning and response steps are the most important for a flood warning system to attain its objectives. This is reflected in the criteria used by Parker and Fordham (1996) in assessing and comparing the maturity of various operational flood warning systems in operation across Europe. Most of these criteria address the dissemination of the flood warning and the organizational embedding of the system. Indeed criteria were introduced to compare the importance of the flood warning element to the flood forecasting element. Systems where flood forecasting dominated and less importance was given to flood warning were deemed to be in an early stage of development. Failure of flood warnings to reach the public (Penning-Rowsell and Tunstall, 1999), or a large number of warnings issued to the public that prove to be false due to low accuracy of forecasts (Krzysztofowicz *et al.*, 1992), will lead to performance deterioration of the flood warning service as a whole.

Despite the need to view flood forecasting from within the perspective of the flood warning system as a whole, the flood forecasting step is one where hydroinformatics and an understanding of the hydrological processes involved can contribute to the improvement of a flood warning system. This is particularly relevant in view of the requirement to deliver a warning with a given lead time, with the required lead time achieved using forecasting techniques that may range from expert (hydrological) interpretation of the monitored data to more formal hydrological modeling setups. In view of this requirement, a categorization of flood forecasting methods can be developed on the basis of the hydrological and meteorological processes involved, which must be explicitly considered in the forecasting step to achieve the desired lead time.

Classification of Flood Forecasting Systems

A simple classification of flood forecasting systems can be found through comparing the desired lead time T_d to the hydrological response time T_p at the location for which the forecast is to be provided. The desired lead time is the time that is needed to effectively undertake the two final steps of *warning* and *response*. From the point of view of different lead times for forecasts with different objectives, we separate between nowcasting (a few hours), short-range forecasting (hours to two days), medium-range forecasting (2–10 days), and long-range and seasonal forecasting (between 10 days and a year). The hydrological response time T_p is similar to the more general hydrological concept of the *"time to peak"*. In the context of initiating a flood warning, it is more appropriate to consider the time between the rainfall event occurring and the crossing of a flood warning threshold. On crossing of such a threshold, a warning may need to be issued and appropriate response initiated. The hydrological response time can be further subdivided into the time that water needs to flow through the river channel (T_c) and the time the water needs to flow from the land phase into the river (T_s). The division between the land phase and the river channel is perhaps somewhat arbitrary, but generally, the river channel is considered to be the main river (system), while the response of the land phase is the response of (sub)catchments before the water flows into the main river system (see Figure 2).

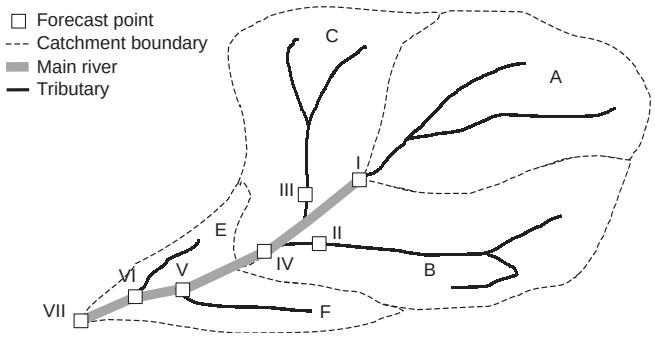


Figure 2 Schematic layout of a catchment, including the main river, tributaries, and catchments

On the basis of these characteristic times, four situations can be recognized (adapted from Lettenmaier and Wood, 1993);

1. $T_d < T_c$ or $T_s << T_c$. The desired lead time is such that the warning will be issued on the basis of water that is already in the main river channel; or, the time the water needs to flow from the land phase into the river is insignificant compared to the time the water needs to flow through the main river. This may be the case for forecast point VII in Figure 2, assuming that catchments E and F have only a minor contribution in flood genesis.
2. $T_d < T_p$ and $T_c \cong T_s$. The desired lead time is such that the warning will be issued on the basis of water that is still on the land phase, and the response time is determined both by the time this water needs to flow from the land phase into the river channel and by the time the water will need to flow through the main river. This may be the case for forecast point IV in Figure 2.
3. $T_d < T_p$ and $T_s >> T_c$. The desired lead time is such that the warning will be issued on the basis of water that is still on the land phase and the response time is mainly determined by the time this water needs to flow from the land phase into the river channel. This may be the case for forecast point I in Figure 2.
4. $T_d > T_p$. The desired lead time is such that warning may be issued on the basis of water that has not yet

fallen as rain. In this case also, the weather forecast is needed for a timely forecast.

Cases 1–3 are typically applied for short-range forecasting in medium and larger basins. Case 4 is typically applied in either medium to long-range forecasting in larger river basins or for forecasting in small (flashy) river basins.

Elements of a Flood Forecasting System

The choice of an appropriate forecasting method incorporated in the flood forecasting system used in the forecast stage will depend very much on how the desired lead time and the hydrological response time compare for the location at which a flood warning service is to be provided. Typically, the forecasting system will employ one or more models to simulate discharges and stages in the river system. These are simulated as a function of observed meteorological conditions and possibly short- or medium-term forecasts of these, depending on how the desired lead time compares with the characteristic hydrological response times. The models are obviously employed in a real-time environment, and this will pose specific requirements on the model in terms of robustness and efficiency. More important though is that the model is seen as a part of the flood forecasting system. Madsen *et al.* (2000) list key elements of a forecasting system operating in a real-time environment:

1. Real-time data acquisition for observed meteorological and hydrological conditions. This element is provided through the detection stage, and it is clear that the data must be made available. For use within the forecasting system, however, the data must be subjected to stringent data validation procedures to ensure integrity of the data used in model simulations on which warnings may be issued.
2. Hydrologic and hydraulic models for simulation. A wide range of modeling approaches may be applied where models take information on the current and past states of the system, and forecasts are made for the desired lead time as a function of boundary inputs on the system.
3. Forecast of meteorological conditions. Where necessary, forecasts of meteorological conditions are required to allow issue of warnings at sufficient lead time. Generally, these forecasts are sourced from meteorological services, and must be integrated within the flood forecasting system.
4. Updating and data assimilation. Through the process of data assimilation and updating, simulated data are combined with real-time data to provide a more accurate forecast. The use of an updating or data assimilation technique is an important aspect of applying models in real time.

REAL-TIME DATA ACQUISITION

Figure 1 shows the important position of real-time data acquisition within the flood warning process. Monitoring of hydrometeorological conditions in the catchment is paramount in the running of an operational flood warning and forecasting system. The availability of real-time observed data is also important to the flood forecasting stage, as these provide the boundary conditions to the models used. The observed data is used not only to bring the models up-to-date, but also for keeping the models in track with reality through data assimilation. Particularly, for forecasting systems of the first and second category described above, where lag times are such that desired lead times can be achieved through calculating runoff and routing of observed rainfall and flow, the availability of real-time data is essential.

Most monitoring data is typically provided from telemetry sites, including river gauges and rain gauges. These sites are typically used for a more general purpose than flood warning alone, providing data also to navigation authorities, water resources planners, irrigation agencies, and others. For flood warning, the timely availability of the data is obviously an important factor in determining the utility of the data, and also reliability of data transfer is an important aspect. Communication may simply be through telephone (e.g. Dobson and Davies, 1990), or where reliability of telephone networks is insufficient, through dedicated VHS or UHF radio links (e.g. Grijsen *et al.*, 1992). The advances in communication technology have in recent years significantly increased the reliability of data transfer, with off-the-shelf technology such as GSM networks and Internet providing reliable transfer mechanisms.

Although the problems of data transmission are being resolved through advances in communication technology, the reliability of the data itself is an issue that warrants close attention. Validation and checking of all real-time data prior to their use in a forecasting model is important to avoid warnings being issued on the basis of doubtful data. The data used is mainly that of rainfall accumulations from rain gauges and levels from river gauging stations. For upstream boundaries, water levels are typically transformed into discharge through stage-discharge relationships. Particularly in severe floods, these stage-discharge relationships may contain a significant degree of error, and the reliability of these relationships at high flows should be carefully assessed. For catchment rainfall estimates derived from rain gauges, the problem of scale arises, as the point measurements must be up-scaled to catchment rainfall estimates. Various interpolation techniques are available, but errors are introduced because of the correlation structure of rainfall and the low density of real-time rain gauges. Rain-gauge locations may also dominate in, for example, valleys in mountainous regions rather than in more inhospitable high

mountain areas (Bürgi, 2002), while the higher mountain regions will generally receive most rainfall.

The use of weather radar partially solves the problem of spatial coverage of rainfall estimates, and has been growing in popularity in recent decades (Todini, 2001). Weather radar provides a spatial estimate of precipitation as inferred through the radar reflectivity of the rainfall, and a reliable relationship between reflectivity and rainfall rate must be applied. Unfortunately, no unique relationship is available (Sharif *et al.*, 2002). In addition, there are a number of physical factors influencing uncertainties in radar derived rainfall estimates. These are not only due to atmospheric distortions such as the melting layer (Sharif *et al.*, 2002), but also due to, for example, radar beam shielding in mountainous areas (Borga *et al.*, 2000). An overview of potential sources of error can be found in Smith and Krajewski, (1991). In practice, integration of rain-gauge data and radar rainfall estimates can be combined through, for example, block Kriging to provide more reliable estimates of distribution and volume of observed rainfall (Todini, 2001).

An alternative method for estimating rainfall quantities is based on the analysis of clouds and cold-cloud duration, as shown by images from a geostationary satellite. This may provide some estimate of rainfall in areas otherwise inaccessible, as demonstrated successfully in the upper Nile basin by Grijsen *et al.*, (1992). However, this technique has not yet proved sufficiently reliable for application in small- or medium-sized catchments outside of tropical areas (Todini, 2001).

HYDROLOGIC AND HYDRAULIC MODELS FOR SIMULATION

Of the large number of hydrologic and hydraulic models that exist, a significant number is used in forecasting systems. However, the requirements are such that for a certain class of forecasting systems the choice of model concepts is usually limited. For an in-depth discussion on the different kinds of (hydrological) models that may be used, the reader is referred to **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**. The following model types used in forecasting can be distinguished:

- Correlation-based models use a linear or nonlinear relation between two or more locations to get to a forecast level or discharge. Water levels or discharge for one location (the forecasting site) can be derived from the level or discharge at one or more support sites (see example in Figure 3). A support site can be an upstream gauge, a gauge in a neighboring catchment, or a combination thereof. For peak flow forecasting, these systems are usually based on a database of events for which the peak flow and date-time information are stored in a database. When new events are recorded, they can be added to the database to improve the results. Depending on the river characteristics and the amount of historical events that are available, a correlation system can provide accurate forecasts. However, the lead time is limited to the travel time between the forecast location and the support location. Within the classification given above, the correlation model is therefore suitable only in the first class of forecasting systems, where the desired lead time for the downstream site is such that the event has been recorded at the upstream site.
- Transfer function models have been used widely in flood forecasting system. There are many different types of transfer functions, the most commonly used types being the Multiple Input Single Output (MISO), dynamic autoregressive exogenous (DARX) models (Lees, 2000a). Typically, transfer functions will be established to define a relationship between rainfall and runoff, or between flow at a downstream site and one or more upstream sites (comparable to the correlation models described earlier). The main attraction of transfer functions is that the model structure is identified using available data only, thus leading to efficiently parameterized model structures (Young, 2003). Transfer function models are also relatively simple to implement and computationally very efficient. To solve some of the difficulties found in representing the nonlinear rainfall-runoff response using linear transfer functions, a nonlinear rainfall loss term can be introduced prior to the linear transfer function to represent antecedent catchment conditions (Lees, 2000a). Young (2003) describes the Data-Based Mechanistic model, which extends the transfer function concept but allows the input parameters to be adaptive as a function of changes in flow, with flow effectively acting as a surrogate measure for catchment storage (see Young, 2003; Young and Tomlin, 2000). Transfer functions have been applied in several operational flood forecasting systems (e.g. Cluckie, 2000). Their reliance on availability of observed data does, however, mean that the effective lead times that can be achieved using transfer function is in the order of the response time of the catchment.
- Nonlinear forecasting methods such as nonlinear prediction (e.g. Laio *et al.*, 2003), Artificial Neural Networks (Laio *et al.*, 2003), and Genetic Programming (Babovic, 1997) extend the principles of transfer function modeling in identifying model structure and parameters through mining the available data. Despite good performance demonstrated in academic research, they are yet to be applied in true operational systems. More details can be found in **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1** and **Chapter 20, Artificial Neural Network Concepts in Hydrology, Volume 1**.

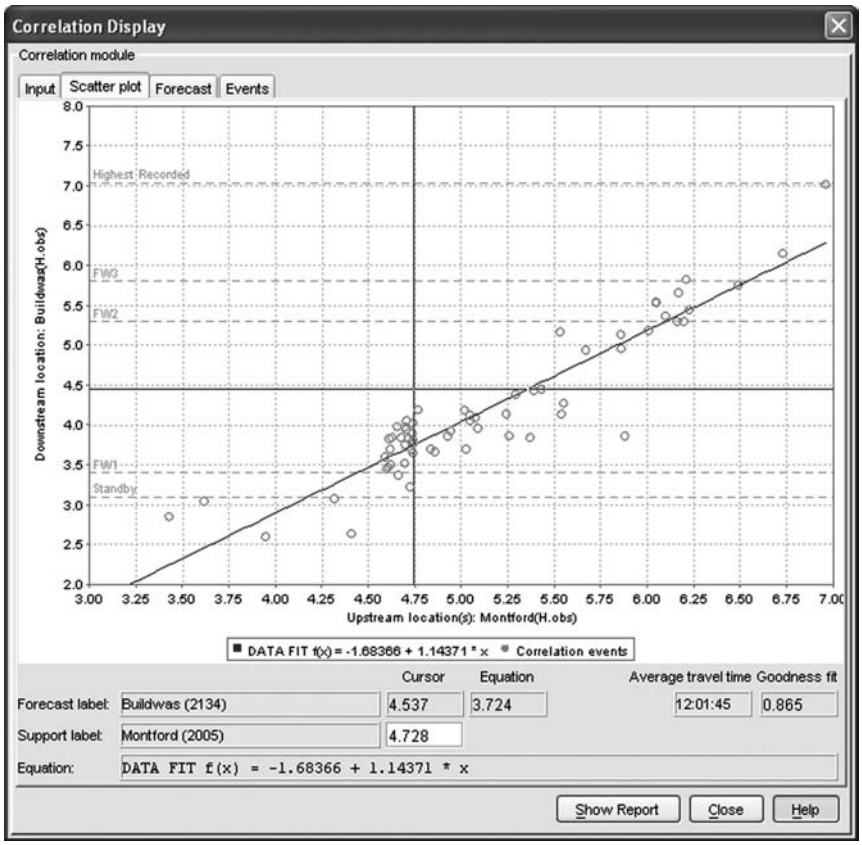


Figure 3 An example of a correlation established between an upstream and a downstream site. The recorded level at the upstream site can be used to derive a forecast for the downstream site through the relationship established. This example is taken from the UK Environment Agency Midlands Emergency Level Forecasting System, where it is used both as a backup and as a sanity check to the suite of hydrologic runoff and routing models normally used. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

- Conceptual hydrological models have also found wide application within flood forecasting systems. These models simplify the hydrological system in a number of concepts, each of which represents a specific part of the hydrological system. In many cases, the different types of dynamic response of the catchment are represented as linear stores. There are a great number of conceptual models in use in flood forecasting systems, including models such as NAM (Van Kalken *et al.*, 2004), Sacramento (Grijsen *et al.*, 1992), HBV (Bürgi, 2002), and PDM (Moore and Jones, 1998). An example of a conceptual model is the HBV model (Bergström, 1995), the structure of which is shown in Figure 4. These models can be applied from small to very large catchments, with larger catchments often being subdivided into several small catchments such that each has different model parameters. In this case, the models may be linked by an external routing model or include a simple hydrological routing procedure in the model itself. Recently, physically based hydrological models have also found application in flood forecasting systems, for example, the LISFLOOD model applied as a distributed model for a pan-European flood forecasting system (De Roo *et al.*, 2003). A comprehensive discussion on conceptual and physically based hydrological models is given in **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**. As both conceptual and physically based models claim to model the behavior of the catchment in response to rainfall (and snowmelt), these models can be applied both in forecasting systems with short desired lead times and those with long lead times, where the response of the catchment to medium to long-range precipitation forecasts is required.
- Once the flow has left the land phase and entered the river (as arbitrary as such a crisp division may be), flow routing is applied to propagate the flood wave from the catchment outlet to the forecasting location, for which a flood warning is to be established. A comprehensive overview of flood routing methods is given in Chow *et al.*, (1988), broadly dividing concepts into lumped routing concepts and distributed routing concepts. The lumped concepts include methods such

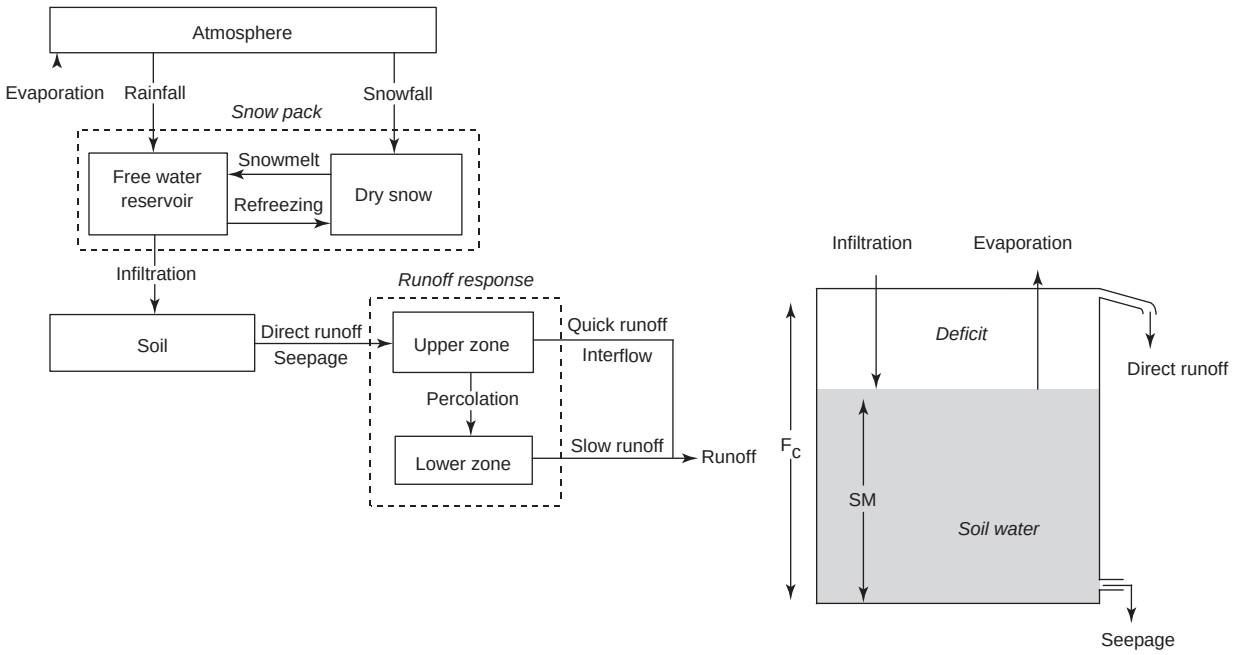


Figure 4 Schematic representation of the main components of the HBV model

as Muskingum routing, while distributed routing models include kinematic wave routing, diffusion wave routing, and full hydrodynamic routing (see **Chapter 16, Numerical Flood Simulation, Volume 1** for further details). The suitability of each of these concepts is very much dependant on the hydraulic conditions of the reach to be modeled. For steeper reaches, the simpler kinematic wave or Muskingum routing methods provide sufficient accuracy, and owing to their simplicity and computational efficiency, they have clear advantages over more complex full hydrodynamic modeling. These have been successfully applied in operational flood forecasting systems (e.g. Dobson and Davies, 1990; Moore *et al.*, 1990). For flatter reaches where backwater effects due to structures, confluences, or tidal influences affect the propagation of the flood wave, full hydrodynamic modeling is more appropriate. With the advance of computer processing power, these models can also be easily incorporated into flood forecasting systems, and have been employed in several operational system (e.g. Grijsen *et al.*, 1992; Jørgensen and Høst-Madsen, 1997). The use of routing models is important in forecasting systems of all lead times, where flood wave propagation through the river system is observed. The relative importance of these models in determining the overall accuracy of the forecasting will, however, decrease as the desired lead times increases beyond the response time of the catchment.

- Floodplain inundation models, with the potential of providing a real-time prediction of spatial flood extent have, with the increase of computing power, also been considered for application in operational flood forecasting systems. Jones and Fulford (2002) describe the use of a hydrodynamic code running in "near-real-time", where inundation maps are created as a parallel process to the operational forecasting system, delivering results a few hours later. De Roo *et al.* (2003) and Beven *et al.* (2000) demonstrate a simple inundation model for use in flood forecasting. The latter explicitly include model uncertainties in the predicted flood extent. This is an important aspect of issuing warnings on the basis of modeled flood extent, as although technically feasible, the use of inundation maps in operational warning must be considered carefully. The potential for misinterpretation of the visually powerful information in a flood map is not to be underestimated (Clark, 1998).

FORECAST OF METEOROLOGICAL CONDITIONS

As the desired lead times for flood warning increase (see the classification of flood forecasting systems in previous articles), the requirement to use quantitative precipitation forecasts (QPF) to provide boundary conditions to the hydrological runoff models increases. Obtaining a QPF at appropriate spatial and temporal scales is not an easy task as rainfall is one of the most difficult elements of the hydrological cycle to forecast (French *et al.*, 1992; Collier and Krzysztofowicz, 2000). For short lead

times, observation-intensive approaches such as nowcasting through, for example, propagation of observed radar rainfall (Golding, 2000) can be applied. Time series analysis techniques using stochastic models or artificial neural networks (Toth *et al.*, 2000; French *et al.*, 1992) can also provide short-term predictions.

For longer lead times, improvements of skill levels in numerical weather prediction models (NWP) have led to these being integrated with flood forecasting systems. Numerical weather prediction models are available either as global models or for a particular section of the atmosphere. The latter are typically local area models with a higher resolution and are nested in the global models. Examples of global models are the European Centre for Medium-Range Weather Forecasting (ECMWF) deterministic and ensemble prediction systems (Buizza *et al.*, 1999) and the German Weather Service global model (Majewski *et al.*, 2002). Many national meteorological agencies operate local area models, for example, the ARPS University of Oklahoma, (Xue *et al.*, 2000), DWD-LM (German Weather Service, Damrath *et al.*, 2000), and HIRLAM (Danish Meteorological Institute, amongst others, Sass *et al.*, 2000). Golding (2000) describes a range of products operated by the UK Met. Office, spanning short, medium, and long-range NWP. NWP products are easily linked to hydrological models, as demonstrated for several of these by De Roo *et al.*, (2003). Although useful in flood forecasting, systematic errors in NWP models, including underestimation, overestimation, and location and scale errors, will have significant impact on the accuracy of derived flood forecasts (Ebert and McBride, 2000).

Longer term flow predictions have also been attempted, though these predictions are primarily concerned with water management issues rather than flood forecasting. These predictions are often based on statistical time series analysis and may use indices such as El Niño/Southern Oscillation(ENSO) as indicators for predicted rainfall (Sharma, 2000; Anderson *et al.*, 2001; Chiew *et al.*, 2003)

UPDATING AND DATA ASSIMILATION

It is important to realize that a fixed model, calibrated to historical rainfall and flow data, cannot be expected to perform equally well when set in a real-time forecasting environment (Young, 2003). Both the model and the available real-time data must be seen as sources of information on the behavior of the catchment, and both will contain some degree of uncertainty, resulting in differences between the observed data and the simulated data. In short-term forecasting, these differences must be taken explicitly into account through some sort of feedback mechanism (Refsgaard, 1997), with the objective of reducing uncertainty in the forecast. Commonly referred to as *data assimilation or updating*, this feedback mechanism combines model results and observations and is a fundamental element of a forecasting system (Grijsen *et al.*, 1992; Kachroo, 1992; Madsen *et al.*, 2000). Four main approaches to data assimilation can be identified (WMO, 1992; Refsgaard, 1997). Figure 5 shows where these different approaches interact with the model. In all cases, the updating procedure is applied as a consequence of the comparison of model outputs and observed values. Model outputs are a function of the input variables, the model states, and the model parameters.

1. Updating of input variables (approach A in Figure 5). Particularly for hydrological rainfall-runoff models used in flood forecasting, the input variables are seen as the dominant source of error. These input variables, such as precipitation and temperature, or inflow discharge, are adjusted to minimize the differences between model output and observed variables. An advantage of input updating is that through adjustment of the input variables the model states are also updated. The two main drawbacks of the methods are that the model itself is in the optimization loop, and that the number of degrees of freedom in adjusting the input parameters results in a badly posed optimization problem. As model inputs are adjusted to minimize errors in model output, the model is considered as perfect. This is generally not the case, but the model should represent catchment behavior relatively accurately to avoid corrections to inputs to be unrealistic. The method of input updating is most applicable to the first three categories of forecasting system. Despite this, there are few practical examples of input updating applied in operational forecasting systems.
2. Updating of state variables (approach B in Figure 5). On the basis of the observed residuals, the state variables of the model are adjusted. These could be for example, soil moisture deficit (SMD) values in a hydrological model, or water levels and discharges at the computational nodes of a hydrodynamic model. A number of approaches can be followed, ranging from simple to complex statistical filters. The simplest method is direct insertion, where a state variable is substituted by an observed variable. This is, however, not always conceptually correct because of the differences in interpretation of what state variables and observed values represent, as well as issues of scale. More advanced methods include Kalman filtering approaches where state variables are adjusted in a physically consistent way using statistical assumptions of the spatial and temporal correlation of model errors (Brummelhuis, 1996). State updating has been applied widely in (operational) flood forecasting systems, varying from simple methods (Moore *et al.*, 1990; Van Kalken *et al.*, 2004) to more advanced Kalman filter approaches in precipitation estimation (Todini, 2001), hydrological models (Georgakakos, *et al.*, 1988; Grijsen *et al.*, 1992), and

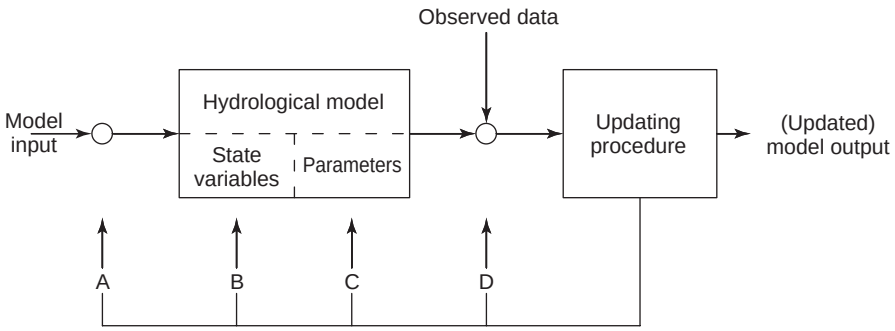


Figure 5 Schematic representation of four approaches in data assimilation (after Refsgaard, 1997)

hydraulic models (Grijsen *et al.*, 1992; Van Kalken *et al.*, 2004; El-Serafy and Mynett, 2004).

3. Updating of model parameters (approach C in Figure 5). Model error is minimized through adjusting model parameters as a function of the errors found in model outputs. The model is thus allowed to adjust to changes in response to small changes in catchment behavior not accurately identified when the model was initially set up (Young, 2003). This approach is particularly applicable to data-driven modeling concepts, and has been applied to transfer functions and data-based Mechanistic Modeling (Young, 2003). Application to more physically based models is less apparent. Here model parameters such as roughness coefficients and runoff coefficients can be considered for adjustment. The procedure could be seen as constant recalibration of the model. If, however, some physical meaning is to be attributed to these parameters, it is unlikely that these will vary with the same dynamics as model errors. Kachroo (1992) notes "it is intrinsically difficult to accept the operation of any hydrological system can change significantly over such a short interval as the observation time".
4. Updating of output variables or error prediction (approach D in Figure 5). Rather than correct the model or its inputs, in this approach, the statistical structure of the model error is considered, and based on the structure found a forecast of the model error is made. The forecast variable obtained from the model is then adjusted with the forecast of the error to obtain an updated forecast (see Figure 6). Commonly, statistical models such as an auto-regressive moving average (ARMA) model is used. The method is conceptually simple, and has the advantage that it is computationally efficient as no additional model evaluations are required. For models used in continuous forecasting, care must be taken that the model gives a reasonable representation of catchment behavior. Contrary to the previous methods, the model state is not updated, and if the model behaves poorly, simulated and updated results will diverge. Although the forecast may be reasonable, this will then rely primarily on the statistical updating method. Error correction is particularly applicable in a chain of models such as in the second and third categories of flood forecasting system. As the statistical correlation of model errors is often not very persistent, the method is less applicable where the desired lead time is relatively long as in the fourth category of forecasting system. Examples of application can be found in Moore *et al.*, 1990, Dobson and Davies, 1990, Refsgaard, 1997, and Madsen *et al.*, 2000.

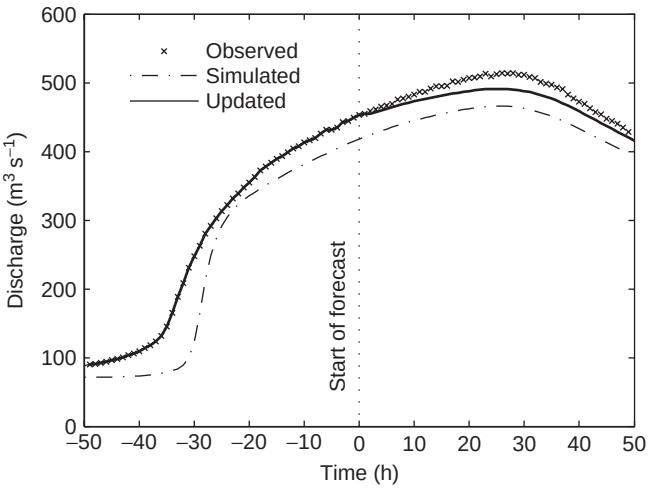


Figure 6 Example of updating simulated flows using error correction. Observed values up to the start of forecast (time zero) were used in identifying the error forecast model. This was then applied in updating the simulated model outputs after the start of forecast

EVALUATING FORECAST PERFORMANCE AND DEALING WITH UNCERTAINTIES

The level of confidence that can be placed in a model-derived forecast is an important factor, if a rational decision is to be made by the flood forecaster on whether or not

to issue a warning on the basis of that forecast. The uncertainty in the forecast may arise as a result of data errors, model structure, uncertainty in model parameters, and most importantly, uncertainties in the forecast model boundary conditions (Lees, 2000a; Beven *et al.*, 2000). Where meteorological forecasts are used to extend forecast lead time, particularly in the last category of desired lead time described in previous sections, the uncertainty of forecast precipitation will dominate resulting uncertainties.

To help make uncertainties explicit, numerous techniques have been applied, although most of these have yet to prove their value in operational flood forecasting, and remain largely academic exercises. A simple method to deal with uncertainties in input data used operationally has been the definition of so-called what-if scenarios. In these, the confidence level given by the meteorological forecaster to the precipitation forecast is implicitly translated into running a number of flow forecast scenarios with amended forecast precipitation (e.g. 10% additional or 10% less precipitation). Uncertainty in the observed precipitation, depending on the balance of desired lead time against hydrological run time, will also influence uncertainties in the forecast, and can be treated in the same fashion. Krzysztofowicz and Herr (2001) introduce a more formal way of translating a probabilistic precipitation forecast into a probabilistic river stage forecast. They use a Bayesian formulation to quantify hydrological uncertainty given a perfect precipitation forecast. This is then combined with the probability of precipitation to derive the probabilistic river stage forecast.

Integration of flood forecasting systems with weather radar nowcasting and numerical weather prediction models has led to the use of uncertainty of inputs in precipitation and temperature being derived from the use of

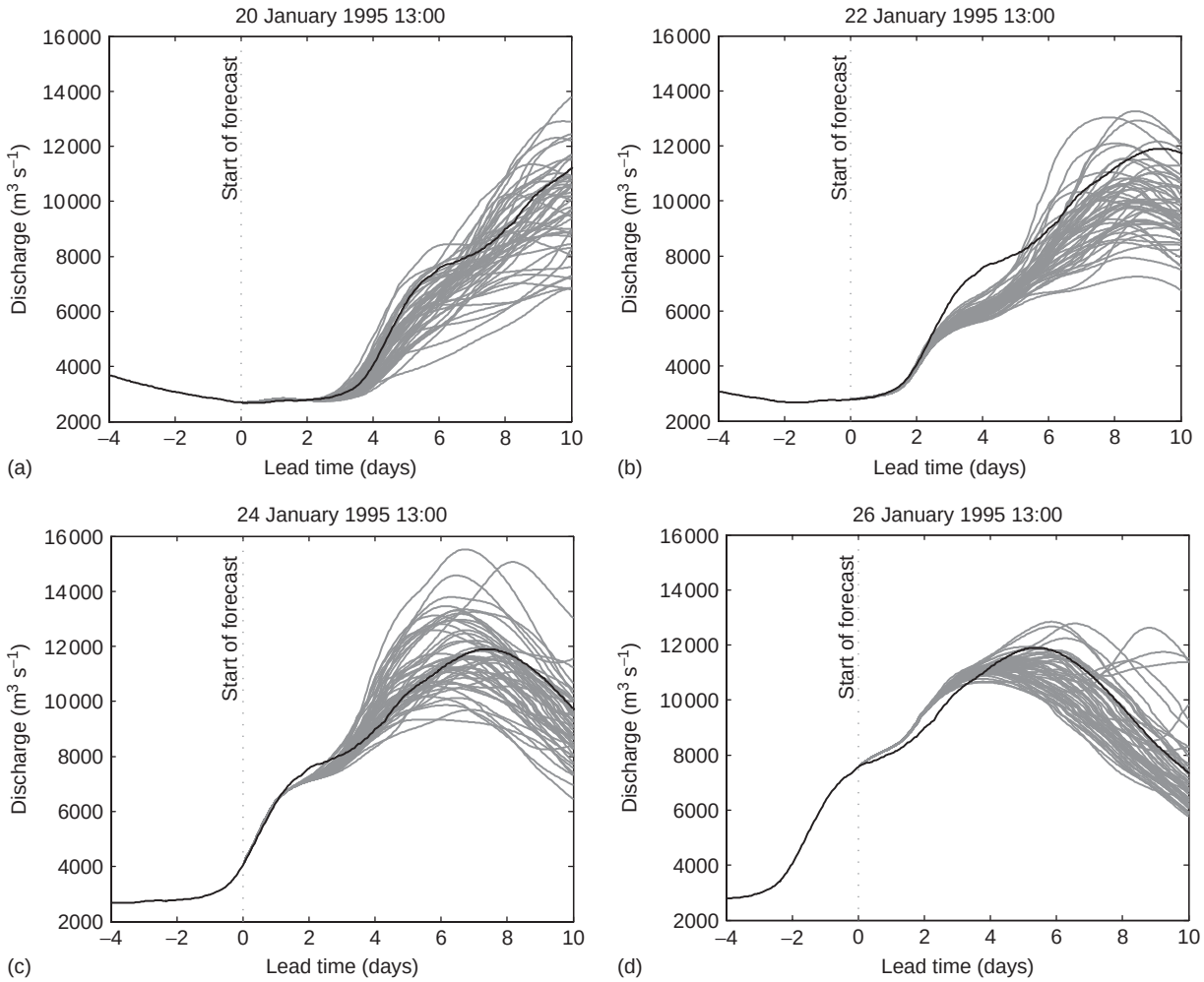


Figure 7 Ensemble forecasts for the January 1995 event on the River Rhine at Lobith. Forecasts shown are issued at two day intervals starting 20 January 1995 at 13:00. Grey lines show the results of ensemble runs, showing updated (simulated) discharge before the start of forecast and ensemble forecast results after the start of forecast. The black line is the observed discharge (after Werner *et al.*, 2004b)

different local or global weather models (De Roo *et al.*, 2003). Ensemble predictions such as those provided by the (ECMWF) Ensemble Prediction System (Buizza *et al.*, 1999), or the American National Weather Service (NWS) Ensemble Streamflow Prediction system (Mullusky *et al.*, 2003) have made the uncertainty in precipitation forecasting even more explicit. The EPS creates 50 ensemble members for a 10-day lead time weather forecast through perturbation of the initial conditions for each (see De Roo *et al.*, 2003 for details). Each ensemble member can then be used to derive precipitation and temperature boundary conditions for hydrological models of a catchment, as demonstrated for the Meuse catchment in De Roo *et al.*, (2003). Figure 7 shows the results of a series of ensemble forecasts for the forecasting location at Lobith on the River Rhine leading up to the January 1995 event. This location is at the lower end of the Rhine basin on the Dutch/German border. This explains why the ensemble weather prediction has almost no influence on the first two days of the forecast, as at these short lead times flows at Lobith are dominated by water already in the main river. At increased lead time, the influence of the ensemble spread starts to dominate the discharge prediction.

The influence of model uncertainties on the total uncertainty in the forecast may also be significant. As in the use of multiple meteorological models, multiple hydrological models may be applied, with each model giving different forecasts. More importantly, the effect of uncertainty in model parameters will influence model-derived predictions. A number of techniques are available for estimating this uncertainty. Lees (2000b) uses a Kalman Filter approach to produce confidence intervals for models derived using the data-based mechanistic approach. More common are Bayesian techniques for estimating posterior parameter distributions from parameter prior distributions, as conditioned on the available calibration data. Examples are the generalized likelihood uncertainty estimation (GLUE) algorithm described by Beven and Binley, (1992) and the shuffled complex evolution Metropolis (SCEM-UA) algorithm described by Vrugt *et al.*, (2003). The posterior distributions established in calibration can then be applied in the forecasting phase, thus creating an ensemble of model runs, with each run using a different parameter set sampled from the posterior parameter distributions.

Irrespective of the method used to quantify uncertainty, the confidence in the forecast must be made explicit by applying a suitable performance measure. Most of the performance measures, or objective functions, used in calibrating hydrological models such as described by Green and Stephenson (1986) may be less suitable for evaluating the performance of model-based forecasts. Although these can be used to give a general indication of model performance, statistics such as skill scores, correctly/incorrectly predicting thresholds, and the timings of threshold crossings may be more relevant in evaluating the performance of the forecasting system (see e.g. of performance measures (Moore *et al.*, 1990) and Krzysztofowicz *et al.*, 1992).

ADVANCES IN FLOOD FORECASTING SYSTEM DEVELOPMENT

The foundations of a scientific approach to flood warning systems had been laid down by numerous researchers during the 1970s and early 1980s (Krzysztofowicz, 1995). During those years, hydrological and hydraulic modeling made some impressive steps forward, partly driven by the rapid increase in computing power as a result of the developments in microchip technology. These models form the basis of many forecasting systems.

As a result, many forecasting systems have evolved around a specific (set of) models. For performance reasons, this made good sense; the forecasting systems is a calibrated model enhanced with the minimum requirements to transform it into an operational forecasting system. However, these systems are not very flexible. Adding new locations may be possible relatively easily, but adding new types of on-line data (that may increase the forecasting performance) is usually very difficult. Most important, however, is that new models cannot be included as the model itself lies in the heart of the system. This is not only an issue when new, potentially better models are considered, but also in the many cases where agencies strive to harmonize forecasting systems across different regions.

Besides these developments in modeling, the availability of data for use in flood forecasting systems has increased dramatically in recent decades. Particularly, the increasing operational availability of local and medium-range meteorological forecasting (De Roo *et al.*, 2003), with the potential of increasing flood forecast lead time, has shifted the challenge in developing flood forecasting systems not only run to the hydrological and hydraulic models as required, but to the handling, quality checking, and processing of large data sets, and finally integrating these with the models. To meet these challenges, the focus in developing flood forecasting systems is moving away from the tailor-made model-centered approach to the generic open-systems approach. These open systems must allow open integration of modeling techniques. More important is that these systems cater for the specific needs of handling data within the context of flood forecasting, including a broad range of data handling techniques such as validation, interpolation (spatial and gap filling), transformation, and algorithms to allow integration of models and data across spatial and temporal scales. Once established, different types of modeling approach can be easily integrated in the flood forecasting process. The choice of the most appropriate modeling strategy can then be established on the basis of the efficiency with which the

forecasting requirement in terms of lead time and accuracy is met, rather than through the simple argument that the flood forecasting system can run with only one particular model. While there are some early example of flood forecasting systems following this "open" approach (e.g. Moore *et al.*, 1990), these are still very much proprietary software packages. Advances in standard mechanisms for data exchange and communication such as the eXtended Markup Language (XML), and Internet communication protocols have more recently helped these systems to comply with international standards in systems development.

A relatively new system that sets the use of hydrometric, meteorological, and forecast data central, while providing both generic data handling utilities and an open interface to models, is DELFT-FEWS as described by Werner *et al.* (2004a). This system is applied in several operational flood forecasting schemes (e.g. Bürgi, 2002; Sprokkereef, 2001), and forms the basis of the National Flood Forecasting System (NFFS) under development for the UK Environment Agency. To illustrate some of the issues addressed through migrating from a model-centered approach to the data-centered approach, the remainder of this section describes the differences between the new "open" NFFS system (using DELFT-FEWS) for the UK Midlands regions and the existing FFS2 system currently in use for that region.

The FFS2 system, a first version of which is described in detail in Dobson and Davies (1990), was regarded as highly advanced at the time it was installed. It is constructed with a hydrological runoff and a flow routing model at the heart of the system. The hydrological (MCRM) model is a bespoke conceptual rainfall-runoff model that includes interception, evapotranspiration, a snow pack routine, and simple flow routing, as well as the ability to consider reservoirs in catchments. The flow routing model (DODO) is a two-layer Muskingum model. Both the hydrological and the routing models are equipped with an error model that updates the simulated discharge using observed discharge at each river gauging station, and the updated trace is subsequently routed into the downstream section. The models have been extensively calibrated over the years and perform well in most cases. The system also polls the telemetry system directly before each forecast, so that the most up-to-date data is available prior to running a forecast. The system holds 147 hydrological gauges, 124 meteorological gauges, and 272 forecasting points that may or may not coincide with a gauge location. Forecasts are made every 12 h but more frequently if alarm conditions occur. Forecasts (up to 50 h lead time) are made on the basis of a 6-h lead time precipitation forecast derived from radar image propagation (NIMROD, see Golding, 2000), extrapolated temperature, and a standard profile for evapotranspiration. SMD values in the hydrological models are updated manually (through direct insertion) using weekly SMD values provided by the UK Met. Office. The entire system currently runs on a DEC-VAX and was written in VAX-FORTRAN.

Given the fact that the current system generally works satisfactorily, would there be any reason to migrate to a new system apart from the fact that the number of people with VAX knowledge is declining rapidly? If migrating to a new system would mean replacing the current models by new ones for the entire region, (111 catchments models and 95 reach model) an enormous effort would be required, without necessarily improving the quality of forecasts. Moreover, the step of going to a new system *and* new models may prevent operational forecasters doing their work properly for a significant length of time. Therefore, a new system should be able to use the current models. In addition, most desires for improvements to the systems are related to newly available data and the desire to include hydrodynamic modeling in the tidal reaches where the flow routing procedure is not suitable. For example, the current numerical weather predictions made by the UK Met Office can provide improved lead times (e.g. 36 h) compared to the radar forecasts and also provide forecasts of temperature and evapotranspiration. This alone can improve the forecasts for longer lead times significantly.

Within the new system, all the functionality of the current system has been replicated, and the original model (MCRM and DODO) are now integrated through an XML interface under control of the DELFT-FEWS system. This XML interface is identical to all models, meaning that where a hydrodynamic model proves more reliable, the flow routing model can simply be replaced for that reach by a hydrodynamic module, without implicating the structure of the forecasting systems, nor the skills or procedures required in operational use. The relative ease with which this is accomplished has already been demonstrated with the limited effort required to include a hydrodynamic ISIS model for the tidal reaches of the Severn River. The system allows also for the hydrodynamic module to run alongside the current flow routing predictions so that the relative improvements by including the new model can be explicitly evaluated. The next steps in gradually improving the system while guaranteeing operational continuity include integration of more hydrodynamic model stretches and the use of NWP forecasts.

CONCLUDING REMARKS

This article focused on a comprehensive overview of flood forecasting systems from the perspective of these systems as a part of the flood warning process. It is clear from the large number of techniques given as examples that there exists a great variation in such systems, often tailor-made to cater for a specific forecasting requirement. Although a wide range of techniques is presently available, and much research has been directed at advancing flood forecasting

techniques, most operational flood forecasting schemes used today are rudimentary.

An important reason for this is that in contrast with the development of models used in design work, making flood forecasting systems operational is a much more complex task. Because of its position as an integrated step in the flood warning process, the difficulty in developing flood forecasting capabilities are more often institutional than technical. The flood forecasting system used by a forecasting authority will have profound influence on the operational procedures employed by that authority in delivering the forecasting and warning requirement. As a consequence, institutional development is as an important aspect to consider when a flood forecasting system is established as a step in the flood warning process. For an operational organization, the impact of a change in the models used in flood forecasting will be much less than the impact of adapting to changes in procedures following the introduction of a new or alternative flood forecasting and warning system. The challenge to hydroinformatics is therefore not only to make the latest techniques in, for example, quantitative precipitation forecasting and hydrologic or hydraulic modeling available for use in operational flood forecasting, but also to do so in such a way that fits into the whole process of flood warning. Where flood warning is already pursued operationally, the continuity of the operational warning is of great importance when introducing advances. Developing open systems that allow advances in sciences to be introduced smoothly into operational procedures is a first step in the right direction, and also closes many of the gaps between academic research on flood forecasting and warning and the procedures actually used operationally.

Acknowledgment

Erik Sprokkereef of the Institute of Inland Water Management, the Netherlands, is thanked for providing the data on which the ensemble forecasts in Figure 7 are based. Tim Harrison of the UK Environment Agency, Midlands Region, is thanked for his comments on the discussion of the migration of the Midland Region flood forecasting system.

REFERENCES

Anderson M.L., Kavvas M.L. and Mierzwa M.D. (2001) Probabilistic/ensemble forecasting: a case study using hydrologic response distributions associated with El Niño/Southern Oscillation (ENSO). *Journal of Hydrology*, **249**, 134–147.

Babovic V. (1997) On the modeling and forecasting of nonlinear time series. In *Operational Water Management*, Refsgaard J.C. and Karalis E.A. (Eds.), Balkema: Rotterdam 195–202.

Bergström S. (1995) The HBV model2. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: pp. 443–476.

Beven K.J. and Binley A. (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K.J., Romanowicz R. and Hankin B. (2000) Mapping the probability of flood inundation (even in real time). In *Flood Forecasting: What does current research offer the practitioner?*, Lees M. and Walsh P. (Eds.), BHS: Occasional Paper No. 12, pp. 56–63.

Borga M., Anagnostou E.N. and Frank E. (2000) On the use of real-time radar rainfall estimates for flood prediction in mountainous basins. *Journal of Geophysical Research*, **105**, 2269–2280.

Buizza R., Hollingsworth A., Lalaurette E. and Ghelli A. (1999) Probabilistic Predictions of Precipitation: Using the ECMWF Ensemble Prediction System. *Weather and Forecasting*, **14**, 168–189.

Brummelhuis P.G.J.ten (1996) Kalman filtering in real-time control systems. In *Proceedings of the 2nd Hydroinformatics conference*, Müller A. (Ed.), Balkema: Rotterdam.

Bürgi T. (2002) Operational flood forecasting in mountainous areas - an interdisciplinary challenge. In *International Conference in Flood Estimation*. Spreafico M. and Weingartner R. (Eds.), CHR Report II-17. CHR, Bern, pp. 397–406.

Burnash R.J.C. (1995) The NWS river forecasting system – catchment modelling. In *Computer Models of Watershed Hydrology*. Singh V.P. (Ed.), Water Resources Publications, New York. pp. 311–366.

Carpenter T., Sperfslage J., Georgakakos K., Sweeney T. and Fread D. (1999) National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *Journal of Hydrology*, **224**, 21–44.

Chiew F.H.S., Zhou S.L. and McMahon T.A. (2003) Use of seasonal streamflow forecasts in water resources managements. *Journal of Hydrology*, **270**, 135–144.

Chow V.T., Maidment D.R. and Mays L.W. (1988) *Applied Hydrology*, McGraw-Hill Book Company: New York.

Clark M.J. (1998) Putting water in its place: a perspective on GIS in hydrology and water management. *Hydrological Processes*, **12**, 823–834.

Cluckie I. (2000) Fluvial flood forecasting. *Journal of the Chartered Institution of Water and Environmental Management*, **14**, 270–276.

Collier C. and Krzysztofowicz R. (2000) Quantitative precipitation forecasting. *Journal of Hydrology*, **239**, 1–2.

Damrath U., Doms G., Früwald D., Heise E., Richter B. and Steppeler J. (2000) Operational quantitative precipitation forecasts at the German weather service. *Journal of Hydrology*, **239**, 260–285.

De Roo A., Gouweleeuw B., Thielen J., Bartholmes J., Bongioannini-Cerlini P., Todini E., Bates P., Horritt M., Hunter N., Beven K., Pappenberger F., Heise E., Rivin G., Hills M., Hollingsworth A., Holst B., Kwadijk J., Reggiani P., van Dijk M., Sattler K. and Sprokkereef E. (2003) Development of a European flood forecasting system. *International Journal of River Basin Management*, **1**, 49–59.

Dobson C. and Davies G.P. (1990) Integrated real time data retrieval and flood forecasting using conceptual models. In *International Conference on River Flood Hydraulics*, White W.R. (Ed.), John Wiley & Sons., pp. 21–30.

Du Plessis L. (2002) A review of effective flood forecasting, warning and response systems for application in South Africa. *Water SA*, **28**, 129–137.

Ebert E.E. and McBride J.L. (2000) Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, **239**, 179–202.

El-Serafy G.Y. and Mynett A.E. (2004) Comparison of EKF and EnKF in SOBEK-River: Case study Maxau-IJssel. In *Proceedings of the 6th international conference on hydroinformatics*, Liong S.-Y., Phoon K.K. and Babovic V. (Eds.), World Scientific Publishing Company: Singapore 513–520.

French M.N., Krajewski W.F. and Cuykendall R.R. (1992) Rainfall forecasting in space and time using a neural network. *Journal of Hydrology*, **137**, 1–31.

Georgakakos K.P., Rajaram H. and Li S.G. (1988). On Improved Operational Hydrologic Forecasting of Streamflows, IIHR Report No. 325, Iowa Institute of Hydraulic Research, The University of Iowa, Iowa City, Iowa, pp. 162.

Golding B.W. (2000) Quantitative precipitation forecasting in the UK. *Journal of Hydrology*, **239**, 286–305.

Green I. and Stephenson D. (1986) Criteria for comparison of single event models. *Hydrological Sciences Journal*, **31**, 395–411.

Grijsen J., Snoeker X., Vermeulen C., El Amin Moh. Nur M. and Mohamed Y. (1992) An information system for flood early warning. In *Floods and Flood Management*, Saul A. (Ed.), Kluwer Academic Publishing, pp. 263–289.

Haggett C. (1998) An integrated approach to flood forecasting and warning in England and Wales. *Journal of the Chartered Institution of Water and Environmental Management*, **12**, 425–432.

Jones J.L. and Fulford, J.M. (2002) Near-real time flood modeling and mapping for the internet, In *Proceedings of the Second Interagency Hydrologic Modeling Conference*, *Las Vegas*.

Jørgensen G.H. and Høst-Madsen J. (1997) Development of a flood forecasting system in Bangladesh, In *Operational Water Management*, Refsgaard J.C. and Karalis E.A. (Eds.), Balkema, Rotterdam 137–145.

Kachroo R. (1992) River flow forecasting. Part 1: a discussion of principles. *Journal of Hydrology*, **133**, 1–15.

Koussis A., Lagouvardos K., Mazi K., Kotroni V., Sitzmannm D., Lang J., Zaiss H., Buzzi A. and Malguzzi P. (2003) Flood forecast for urban basin with integrated hydro-meteorological model. *Journal of Hydrological Engineering*, **8**, 1–11.

Krzysztofowicz, R. (1995) Recent advances associated with flood forecast and warning systems *Reviews of Geophysics* **33**, Suppl, American Geophysical Union, **http://www.agu.org/revgeophys/krzysz00/krzysz00.html**.

Krzysztofowicz R. and Herr H.D. (2001) Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation dependent model. *Journal of Hydrology*, **249**, 46–68.

Krzysztofowicz R., Kelly K.S. and Long D. (1992) Reliability of flood warning systems. *Journal of Water Resources Planning and Management*, **120**, 906–926.

Laio F., Porporato A., Revelli R. and Ridolfi L. (2003) A comparison of nonlinear flood forecasting methods. *Water Resources Research*, **39**(5), 1129, doi: 10.1029/2002WR001551.

Lees M.J. (2000a) Advances in transfer function based flood forecasting. In *Flood Forecasting: What does current research offer the practitioner?* Lees M. and Walsh P. (Eds.), BHS: Occasional Paper No. 12, pp. 41–55.

Lees M.J. (2000b) Data-based mechanistic modeling and forecasting of hydrological systems. *Journal of Hydroinformatics*, **2**, 15–34.

Lettenmaier D.P. and Wood E.F. (1993) Hydrologic Forecasting. In *Handbook of Hydrology*, Maidment R.D. (Ed.), McGraw-Hill: pp. 26.1–26.30.

Madsen H., Butts M., Khu S. and Liong S. (2000) Data Assimilation in rainfall runoff forecasting, *Proceedings of the 4th Hydroinformatics Conference*, IAHR, Iowa.

Majewski D., Liermann D., Prohl P., Ritter B., Buchhold M., Hanisch T., Paul G., Wergen W. and Baumgardner J. (2002) The operational global Icosahedral-hexagonal gridpoint model GME: Description and high-resolution tests. *Monthly Weather Review*, **130**, 319–338.

Mullusky, M., Demargne, J., Welles, E., Wu, L., and Schaake, J. (2003) *Hydrologic Applications of Short and Medium Range Ensemble Forecasts in the NWS Advanced Hydrologic Prediction Services (AHPS)*, **http://ams.confex.com/ams/pdfpapers/72137.pdf**.

Moore R. and Jones D. (1998) Linking hydrological and hydrodynamic forecast models and their data. In *RIBAMOD River Basin Modeling, Management and Flood Mitigation: Proceedings of the First Workshop*, Casale R., Havnø K. and Samuels P. (Eds.), European Community: pp. 37–56, EUR17456EN.

Moore R.J., Jones D.A., Bird P.B. and Cottingham M.C. (1990) A basin-wide flow forecasting system for real time flood warning, river control and water management. In *International Conference on River Flood Hydraulics*, White W.R. (Ed.), John Wiley & Sons.: UK, pp. 21–30.

Parker D. and Fordham M. (1996) Evaluation of flood forecasting, warning and response systems in the European Union. *Water Resources Management*, **10**, 279–302.

Penning-Rowsell E.C. and Tunstall S. M. (1999) The weak link in the chain: Flood warning dissemination. In *RIBAMOD River Basin Modeling, management and flood mitigation: Proceedings of the workshop/expert meeting*, Casale R., Borga M., Baltas E. and Samuels P. (Eds.), European Community: pp. 129–146, EUR18853EN.

Penning-Rowsell E., Tunstall S., Tapsell S. and Parker D. (2000) The benefits of flood warnings: real but elusive, and politically significant. *Journal of the Chartered Institution of Water and Environmental Management*, **14**, 7–14.

Refsgaard J. (1997) Validation and intercomparison of different updating procedures for real-time forecasting. *Nordic Hydrology*, **28**, 65–84.

Sass B.H., Nielsen N.W., Jörgensen J.U., Amstrup B. and Kmit, M. (2000) *The Operational DMI-HIRLAM System*. Danish Meteorological Institute Technical Report 00–26, Danish Meteorological Institute, Available from **www.dmi.dk**.

Sharif H.O., Ogden F.L., Krajewski W.F. and Xue M. (2002) Numerical simulations of radar rainfall error propagation. *Water Resources Research*, **38**(8), 15-1–15-14 doi:10.1029/2001WR00525.

Sharma A. (2000) Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 – a nonparametric probabilistic forecast model. *Journal of Hydrology*, **239**, 249–258.

Smith J.A., Krajewski W.F. (1991) Estimation of the mean field bias of radar rainfall estimates. *Journal of Applied Meteorology*, **30**, 397–412.

Sprokkereef E. (2001) *Extension of the Flood Forecasting Model FloRIJN. NCR Publication*,12–2001, ISSN no. 1568234X.

Steinbach G. and Wilke K. (2000) Flood forecasting and warning on the River Rhine. *Journal of the Chartered Institution of Water and Environmental Management*, **14**, 39–44.

Todini E. (2001) A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hydrology and Earth System Sciences*, **5**(2), 187–199.

Toth E., Brath A. and Montanari A. (2000) Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, **239**, 132–147.

Van Kalken T., Skotner C. and Madsen H. (2004) A new generation, GIS based, open flood forecasting system. *Proceedings of the 8th National conference on hydraulics in Water Engineering*, The institute of Engineers: Australia.

Vrugt J., Gupta H., Bouten W. and Sorooshian S. (2003) A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, **39**, 1201, doi:10.1029/2002WR001642.

Werner M.G.F., Reggiani P., De Roo A., Bates P.B. and Sprokkereef E. (2004b) Flood forecasting and warning at the river basin and at the European Scale. *Natural Hazards*, in press.

Werner M.G.F., van Dijk M. and Schellekens J. (2004a) DELFT-FEWS: An open shell flood forecasting system, In *Proceedings of the 6th international conference on hydroinformatics*, Liong S.-Y., Phoon K.K. and Babovic V. (Eds.), World Scientific Publishing Company, Singapore 1205–1212.

Wind H., Nierop T., De Blois C. and De Kok J. (1999) Analysis of flood damages from the 1993 and 1995 Meuse floods. *Water Resources Research*, **35**, 3459–3465.

WMO (1992) *Simulated Real-time Intercomparison of Hydrological Models*, Technical Report 38, World Meteorological Organisation, Geneva.

Xue M., Droegemeier K.K. and Wong V. (2000) The Advanced Regional Prediction System (ARPS) – A multi-scale non hydrostatic atmospheric simulation and prediction model. Part I: Model dynamics and verification. *Meteorology Atmosphere Physics*, **75**, 161–193.

Young P.C. (2003) Advances in real-time flood forecasting. *Transactions of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences*, **360**, 1433–1450.

Young P.C. and Tomlin C.M. (2000) Data-based mechanistic modeling and adaptive flow forecasting. In *Flood Forecasting: What does current research offer the practitioner?* Lees M. and Walsh P. (Eds.), BHS: Occasional Paper No. 12, pp. 26–40.

Zealand C.M., Burn D.H. and Simonovic S.P. (1999) Short-term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, **214**, 32–48.

24: Network Distributed Decision Support Systems and the Role of Hydrological Knowledge

ANDREJA JONOSKI

Department of Hydroinformatics and Knowledge Management, UNESCO-IHE Institute for Water Education, Delft, The Netherlands

This article introduces the hydroinformatics concept of new kinds of environments for supporting decision-making processes in the fields of water and the environment, under the name of Network Distributed Decision Support Systems (NDDSSs). These environments are designed to enable active participation of large numbers of interested stakeholders including the general public. The platforms for their development and deployment are the electronic networks such as the Internet and the wireless mobile telephony networks. Core components of such systems are instantiated hydrological models, acting as main fact providers for the participants' judgments and evaluations of the proposed courses of actions. For achieving the required performance within NDDSSs, these models need to be distributed (viz. capable of representing spatial variability as encountered in real situations) and capable of simulating integrated physical processes in the hydrological system, and open for wider integration with other kinds of models such as ecological or socioeconomic models. A brief overview of the software agent technology is presented, with its potential for the design and development of these new decision support environments. The manners in which full-fledged NDDSSs can be implemented into our present societies are also presented.

INTRODUCTION: THE NEW CONTEXTS FOR HYDROLOGY

During the last couple of decades, the developments in hydroinformatics have revolutionized the ways in which we deal with our contemporary water- and environment-related problems. This has been achieved by exploiting the explosive advances in the Information and Communication Technologies (ICTs) for maximizing the benefits of generation, communication, and application of water-related knowledge. The fundamental process through which hydroinformatics has been utilizing the advances in ICT is the process of knowledge *encapsulation*. Primary examples of knowledge encapsulators are the hydroinformatics modeling systems which encapsulate *generic* knowledge in a given discipline or integrated set of disciplines (e.g. hydrology, hydraulics, water quality, and so on; *see also* **Chapter 16, Numerical Flood Simulation, Volume 1**; **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**; **Chapter 18, Shallow Water Models with Porosity for Urban Flood Modeling, Volume 1**; *and also* **Chapter 23, Flood Early Warning Systems for Hydrological (sub) Catchments, Volume 1**). Through implementation of advanced user interfaces and a variety of data integration components, the application of these modeling systems for specific cases (building instantiated models) has been made both efficient and widely available. Nowadays, full hydroinformatics systems integrate observation and data gathering components, data storage facilities, integrated modeling systems, and their embedding into wider decision support tools and environments.

If we make an attempt to generalize the kind of transformation that these developments have brought about to the water-related knowledge, we can notice that what hydroinformatics did in this area is actually a part of a much wider

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

transformation of knowledge that we are witnessing in our contemporary societies. As the ICT developments enabled the processes of knowledge encapsulation, transmission, and application generally, the very nature of knowledge has been transformed. The term knowledge *production,* which is in frequent use at our present times, already suggests the kind of transformation that we are dealing with. Following Lyotard (1979, 1984), we will introduce the term "*knowledge commodification*" for describing this process. This term suggests that in our postmodern societies, knowledge is increasingly considered a commodity, and the direct intention of knowledge *trading* is becoming dominant legitimation for the knowledge "production" processes. The very notion of "knowledge products" and the possibility for their buying and selling is inextricable from the ICTs. The pervasion of these technologies to an extent that they would transform our traditional and established understanding of knowledge was clearly foreseen by Lyotard (1979, 1984) almost three decades ago:

> "The nature of knowledge cannot survive unchanged within this context of general transformation. It can fit into the new channels, and become operational, only if learning is translated into quantities of "information." We can predict that anything in the constituted body of knowledge that is not translatable in this way will be abandoned and that the direction of new research will be dictated by the possibility of its eventual results being translatable into computer language. The "producers" and users of knowledge must now, and will have to, possess the means of translating into these languages whatever they want to invent or learn. . . Along with the hegemony of computers comes a certain logic, and therefore a certain set of prescriptions determining which statements are accepted as "knowledge" statements".

Hydrological knowledge is certainly not excluded from these fundamental transformations, and is in fact, quite central in the above-described hydroinformatics developments. The processes of externalization and the consequent transmission and application of hydrological knowledge, which used to be limited to texts, images, and mathematical descriptions, are increasingly carried out by means of hydrological models and modeling systems. Much of the present research in hydrology (as many articles in this encyclopedia exemplify), and quite a lot more in hydroinformatics (as demonstrated in this article) is oriented towards further sophistication of these modeling systems in terms of their capabilities for encapsulating the hydrological processes in question, but also towards diversification of the modes of their application and making them easily accessible through improved user interfaces and data integration. The potential of the new communication media for these purposes, such as the Internet and the mobile telephony, is just beginning to be exploited. With the underlying intention of continuously increasing the number of users and broadening the user bases for these modeling systems, all such developments form a clear picture of hydrological knowledge commodification.

Analyzing only the side of knowledge production, however, will give an incomplete picture of the transformations within which knowledge commodification emerges. In fact, it is only made possible if corresponding transformations also appear on the side of knowledge application, or more broadly, the knowledge *demand*. If we again start with hydrological knowledge, which is our immediate concern here, we notice that recent decades have brought exponential growth in the number of problems where hydrological knowledge became indispensable. There is a reasserted relevance of hydrology for understanding global climatic or environmental phenomena and the effects of human interventions in the natural system. Even more significantly, there is now a clear awareness that any human intervention at regional and local level into the water-based natural "economy" has far-reaching consequences in many different, and rather diverse sectors of our societies. Consequently, increased numbers of social groups, institutions, as well as individuals or groups of individuals show direct interest in water and increasingly claim a stake in water-related affairs. This situation necessitates provision of efficient access to hydrological knowledge and capabilities for its application in very diverse application contexts.

At the same time, this rediscovered nature of our contemporary water-related problems puts new demands for combining and integrating hydrological knowledge with all kinds of other types of knowledge, such as, ecological, economic, social, political, and legal. Very often the framework of integration needs to include a kind of social knowledge that is variously described as "traditional", "custom", or "narrative", but has in fact, a unifying characteristic of not being fully susceptible to rational analysis in a modern scientific sense. These new requirements for integrating hydrological knowledge with such diverse types of knowledges (Used deliberately in plural following many writers on issues related to postmodern epistemology; see, e.g. Gergen and Thatchenkery (1997), and also in more familiar context, Abbott *et al.* in **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**) bring formidable difficulties regarding the design of conceptual integrating frameworks or the development of environments, where the required knowledge exchange and integration might be enabled.

This relatively new situation on the application side inevitably changes and determines the nature of hydrological knowledge. Since the integrated approach described here is conceivable only by employing the power of enabling ICTs, the questions such as hydrological knowledge encapsulation, interfacing with other knowledges, and communication to a variety of users through appropriate translation are increasingly penetrating the hydrological agenda. In addition to the advances on the scientific fronts of hydrology (including those dealing with monitoring and observation of hydrological systems), which are also

critically dependent on ICTs, the above mentioned questions are becoming very important for the role and social relevance of hydrological knowledge.

Recognition of the *diversity of knowledge and knowledge needs*, together with the realization of the necessary mobilization of different knowledges when dealing with our contemporary problems is a second general characteristic of the postmodern knowledge transformation, in addition to the knowledge commodification. Again, quite independently from developments in hydrology and hydroinformatics, these transformations were steered by a mixture of social, technical, and historical developments during the last century. The catastrophes of the two world wars, the subsequent cold war, and the powerful awareness of the scales of environmental destruction brought about by modern development had strong influence on shaping this transformation. Explosive development of the mass media contributed to the widespread awareness of our existence in a "global village," which is at the same time sustained by cultural and social diversity, in a manner that is not so different from the sustaining role of biodiversity for natural ecosystems. *Diversity* has entered the scene in virtually all domains of human thought and action, and knowledge diversity appears as an inevitable consequence.

Recognizing that there are diverse knowledges as well as diverse interests and stakes in water affairs may have been the easier task, when compared to the need for integration of this diversity for purposes of improved decision making. As mentioned earlier, in the fields of water and the environment we are faced with immense integration-related difficulties at both conceptual and at implementation level. One attempt to capture and deal with the complexity of these integration problems has been through the introduction of the paradigms of Integrated Water Resources Management (IWRM) and Integrated River Basin Management (IRBM). These paradigms have been quite successful in providing the basic descriptive framework for analyses of water-related problems and as fundaments for shaping national, regional, and international policies on water to an extent that in some cases they are incorporated in legislation (e.g. the European Water Framework Directive, of the European Union). These frameworks have also been quite beneficial for understanding the new contexts in which hydrological knowledge has to be generated and applied (Environmental Impact Assessment – EIA and Social Impact Assessment – SIA frameworks, when dealing with water-related projects, bring similar broadening of the contexts for hydrological knowledge).

When it comes to implementation actions, however, these frameworks are simply not enough for achieving the desired results. Firstly, a lot of concepts are simply not possible to be properly framed in such paradigms. One good example is the proper institutional arrangement, which remains an outstanding problem even in most developed societies and countries (see Loucks, 2003). Secondly, some of the concepts in these paradigms are contradictory among themselves (integration/optimization, which presupposes cooperation and valuation consensus, versus stakeholder, and public participation, which inevitably brings value disagreements and even conflicts which require negotiations). Or, some concepts are rather detached from actual realities with no directions as to how they can be realized in practice. A clear example from IRBM is the frequently forgotten fact that the very concept of a river basin as a unit of analysis is still alien to many communities and, consequently, to many already established institutions. Rivers were (and still are), more often than not, natural boundaries and borders between different communities, regions, and countries, rather than geographical features of integration. River basins are rarely intimated to individuals and communities as something toward which they develop sense of belonging, and in view of their other social allegiances they may very often turn out to be less important, if not irrelevant.

In view of this situation, rather than imposing theoretical frameworks, the approach advocated by hydroinformatics is to proceed with the implementation of the required integrative approaches through design, development, and deployment of new kinds of ICT-based systems and environments. These systems would be specifically intended to enable direct participation of all stakeholders (including the "general public", which consists of a large number of *individuals*) in the decision-making and assessment processes. The envisaged deployment of these systems or environments is, in concrete projects, for enabling the participants to analyze the consequences of the proposed courses of action regarding issues of their *immediate* concern. These products must go beyond the mere exhortations concerning the need for environmental preservation, sustainable development, or proper integrated management of their resources. As Woods (1993) explained this within the context of his analysis of ICT contribution to development:

> Educating people *about* natural preservation is not enough. The education and training to achieve that preservation is needed. Interactive technology has immense potential as a tool for assisting communities to enter their own data on their own resources and to process it and weigh up alternative solutions. It carries the learning process into practical application in specific situations and can then provide technical instructional materials to suit the conditions of these situations. This capacity to help individuals, groups and communities to manage their own resource is being shown to have a remarkable 'mobilizing' effect.

At present, the obvious platforms for development and deployment of such tools are the electronic and wireless networks, such as the Internet and mobile telephony. The working title that has been given to such systems is Network Distributed Decision Support Systems (NDDSSs). Although the word "network" refers here primarily to the

fact that these systems will use the electronic and wireless network as their platforms, recent hydroinformatics studies emphasized the fact that the design and development of such systems cannot be considered as a purely technical task. Thus, the electronic networks are only technologies that enable the realization of these essentially sociotechnical systems. The systems themselves enable the creation of networks of people and institutions engaged in the processes of knowledge exchange for the purposes of improved assessment and decision making.

In what follows in this article we will try to present the basic concepts of NDDSSs, the role of hydrological knowledge, particularly the role of hydrological models within these systems, and the potential of a specific software design and development technology, known as *software agent technology*, for the realization of these systems. At the end we will identify the potential, social support for their future development. The theoretical basis, and the detailed description of NDDSSs have been introduced in Jonoski (2002).

The studies on NDDSSs are not yet in a mature stage of development, lacking full-fledged practical implementations. At this juncture, we need to stress the relevance of these studies in redefining hydroinformatics as a sociotechnical discipline and in highlighting the developments in hydrology from such sociotechnical perspective. As shown in most recent writings about future hydroinformatics directions, such analyses and redefinition are centered on the questions of knowledge, knowledge encapsulation and communication, and knowledge legitimation in general (*see* **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**). By taking this meta-perspective on the fields of hydroinformatics and hydrology, NDDSSs studies can be considered as postmodern, theorizing attempts to "tell it as it might become" rather then "telling it like it is".

NETWORK DISTRIBUTED DECISION SUPPORT SYSTEMS (NDDSSs)

The basic idea of NDDSSs is to develop Internet-based environments as virtual platforms that will enable transmission of knowledge for three distinct purposes(Strictly speaking, "environment" is a more appropriate word for describing NDDSSs compared to the word "system", since there is no clear delineation of what constitutes the "system" in this case, or what its inputs and outputs are. These two words will be used interchangeably here, depending on the intended meaning in the local context). Firstly, they should enable all interested parties in the decision-making process to gather relevant knowledge and data about a given water-related system and particularly about consequences of any proposed, or projected interventions in the same system. Secondly, by combining this factual knowledge with their own beliefs, attitudes, and interests, these participants should be able to formulate judgments about the proposed interventions and even suggest new courses of action. Thirdly, after aggregating the judgments and positions of all interested parties, the system needs to act as a platform for negotiation and collaboration activities among all participating parties through which the decision-making process would move towards commonly complying actions. Although this is a rather general description, it connects the requirements of NDDSSs to fields such as environmental impact assessment and social impact assessment, risk assessment, and participatory policy analysis. The design of these systems will therefore need to rely on many concepts already developed in these diverse fields. It should also become clear that these systems have much more to offer in the areas of planning, design, and strategic development, while being less relevant when it comes to operational management of water-based systems.

Since we are concerned here with water-related problems, an NDDSS is best conceived as dealing with decision problems related to a particular spatially bounded object such as a river basin, a coastal region, or a groundwater system of aquifers, or even some or all of such features combined. At the simplest and most general level, an NDDSS can be seen as consisting of a "knowledge center", and a large number of user nodes distributed along a "user periphery" (Figure 1).

The knowledge center is the primary repository of all what we consider as scientific, or, more broadly, factual knowledge. This is primarily, electronically encapsulated knowledge, and first and foremost this means various models (hydraulic, hydrological, ecological etc.). The

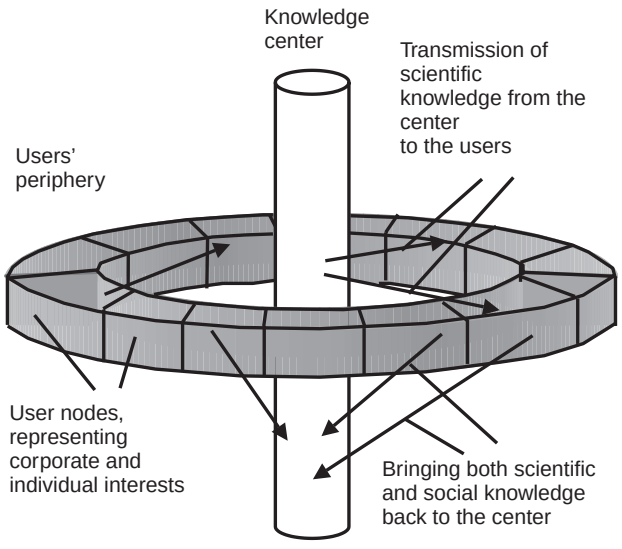


Figure 1 The general concept of NDDSS (adapted from Jonoski and Abbott, 1998)

knowledge center is also the repository of all relevant data that are needed for generating or supporting these models. In many cases, we may conceive that the knowledge center will have to maintain continuous links with all kinds of measurement facilities (both from data gathering equipment on the ground and from remote sensing data such as satellite imagery, radar data for rainfall prediction etc.). Although the gathering and distribution of this kind of knowledge is the primary role of the knowledge center, other kinds of knowledge that are not directly related to water sciences should also be located here. For example, any complete center will eventually have to encapsulate legal knowledge bearing upon the problems in question.

The users of the system may now be visualized as being distributed over the user nodes and connected to the knowledge center. These users are both corporate bodies with their own perceived interests, as well as individuals or groups of individuals grouped around some common concern.

It is very important to stress that the knowledge center should not be seen as a physical center of the system, or, even worse, to assume that this is a center only in technical terms and that it can be reduced to one, or even several server computers. The word "center" is used here only conceptually, and the models and data provided by this center might well reside at different physical locations, and, quite commonly, in different organizations.

Within this basic concept of the NDDSS environment, we can introduce its three functional components that will enable the three principal processes of knowledge transmission from the beginning of this section. These three functional components are correspondingly: *the fact engine, the judgment engine*, and a *collaboration and negotiation platform*. The principal processes of knowledge transmission in turn determine the main characteristics and tasks for these three functional components. Later in this section, we will introduce these features, together with some suggestions as to how these three components can be realized. In the following section we will focus mainly on the fact engine component and the important role of model-encapsulated hydrological knowledge.

Fact Engines

The fact engine component of the NDDSS is located in the knowledge center and is responsible for the production of the (scientific) facts, primarily by providing storage and access to a collection of models and databases related to various aspects of the physical system in question. Most of these models will initially be traditional numerical models, such as various hydraulic, hydrological, ecological, or water quality models. Since the main role of these models is to provide the facts about the physical system in question, we classify them under the category of fact engines. The access to these models is now necessarily going to be through the Internet, or some similar kind of electronic, or wireless network. Therefore the traditional user interfaces designed for using models as stand-alone applications will have to be decoupled from their computational engines and embedded in other media, such as, for example, a Web-browser.

However, the envisaged broad user base of NDDSSs indicates that these models may need to be stripped from their traditional user interfaces and replaced by new ones, which will be more adaptable to the diversity of the users. In fact, alternative kinds of user interfaces that will be able to represent the same facts in different forms must be provided. Because this aspect is actually closely related to the judgment-forming process that occurs at the user nodes and is to be handled by the judgment engines, the user interfaces within an NDDSS are not primary considerations in the design of the fact engine component. The existing user interfaces will certainly be indispensable for setting up or "instantiating" the models as fact engines – a task to be carried out by the modelers who are working at the knowledge center. However, the interfaces to these instantiated models that will be available at the user nodes will need to be reconstructed in such a way that their primary goal is to enable the judgment-forming process, and not the modeling process *per se*. This implies the representation of the modeling results in different forms that will support the diverse backgrounds and knowledge of the users. It also places limitations on the number of model inputs (or parameters) that can be manipulated. In fact, these manipulations must be limited to those inputs that are of direct concern for the actual user/participant in the process.

The processes of fact generation are still rather diverse, and apart from running models and preparing appropriate material for transmission, they will involve tasks such as interrogating measurement stations, requiring and interpreting satellite images, or querying databases and knowledge bases for other non "water-related" facts. It is clear that by far the greater part of the scientific and other factual material is available and can be supplied by existing organizations which will become in their turn "knowledge providers" to the center. Efficient interoperation of such diverse fact engines, located at different knowledge providers will necessarily require new kinds of institutional arrangements.

Judgment Engines

The judgment engine component will usually and primarily be distributed over the users' periphery, where we may envisage that all the users who will participate in the impact assessment process and the decision-making process will be situated. The judgment engines have the task of assisting the users in making their own judgments about a particular intervention in the system on the basis of the interaction

between their beliefs, intentions, and interests, and the received facts from the knowledge center.

Because of the diversity of the users/participants involved in the process, the first prerequisites in the formation of judgments are related to the ways in which the facts coming from the knowledge center are delivered. As we have already mentioned, the user interfaces of the fact engines may have to be significantly adapted for this purpose. The main orientation in the new design must be towards *customization* and *personalization* of the content delivered from the fact engines. Various techniques for provision of customized content, albeit in different areas and for different purposes, are already used on the World Wide Web by many content providers. In this case, the customization is related both to the *kind* of facts that are relevant to a particular user for making his or her judgment and to the *form* of delivery of these facts. The design of these interfaces should especially be oriented towards assisting users that are not directly familiar with the water- and environment-related scientific knowledge in their judgment-forming process.

The delivery of the relevant facts is, however, only one step in the judgment-forming process. The main role of the judgment engines is in enabling the *combination* of the facts with the users' beliefs and interests for forming judgments. It should be realized that the judgment engines of an NDDSS are envisaged as primarily supporting and facilitating the formation of judgments of *value*. The diversity in values and in value gradients of the users/participants requires that the judgment-forming process becomes understood within a framework where this diversity is respected, and the tendencies towards restriction of certain values, or towards reduction of all values to one (usually monetary) value, are discouraged.

In any NDDSS, there will be issues for which the valuation is relatively straightforward, and even uniform, across all the participants in the process. This is especially the case for those issues that are easily represented in the form of quantitative information. However, the problems arise when there are large number of issues involved, and especially when many of them cannot be represented as quantitative information. If we look at the approach taken in various environmental impact assessment procedures and tools, the value judgments for such issues are usually encapsulated in the so-called "*value-functions*", which are relationships between certain facts and user-perceived "environmental qualities" or "values". Each "environmental quality" or "value" is usually expressed on a scale between 0 (very bad) and 1 (very good). These relationships are usually preprepared by experts and the user only implements them in the assessment procedure.

This kind of expert knowledge should be provided in the judgment engine components of NDDSSs as well, but in a more appropriate form and mainly as information needed to support the judgment-forming process. The temptation to "lock" the judgments in the value-functions should however be avoided. Many participants will have specific knowledge about their local conditions, or have a different reasoning altogether as compared to the one underlying this expert advice, and this may lead to different patterns of valuation. In fact, the judgment engines of NDDSSs will have to allow for judgments which are in a sense "looser" as compared with the structured ones provided by traditional tools and systems for impact assessment, which are based on formalized procedures. Frequently, the only meaningful judgment provided by a given user about a certain state of affairs may be in simple linguistic terms, such as "good" or "bad". This indicates the potential for employing fuzzy sets for conversion of such linguistic terms into fuzzy numbers that can be used for further processing of the provided judgments. Some background on using fuzzy concepts in modelling are provided by Solomatine in **Chapter 19, Data-driven Modeling and Computational Intelligence Methods in Hydrology, Volume 1**.

When it comes to the fundamental theories in support of the actual design of the judgment engine, the basic and starting consideration is that the judgment-forming process is equally dependent on the judged objects and the mental model of the judging agent. It is a view that is sometimes named *interactionist* (see Klinkenberg, 1996, p. 100), and may be contrasted with the traditional dichotomy of the subjective/objective view. The answer to the questions of *how* this interaction actually occurs must be sought in the study of processes of *signification and communication*, as the embodiments and transmissions of all experience. General semiotics is the field that studies these processes at a level that can provide many guidelines for the design of judgment engines. The appropriate use of *signs and sign vehicles* will become a central question for good design of judgment engines, which enables the participants to make *autonomous* judgments.

This field is relatively new to hydroinformatics, and has not been studied sufficiently. In the design of modeling systems orientated predominantly towards specialists in water and environment, the need for a systematic study of the design of user interfaces has not presented such a crucial requirement. The discourse of this user base has been relatively limited and there have not been serious misunderstandings in this process (although we are capable of still distinguishing "good" from "bad" user interfaces). The broad discourse of an NDDSS requires a much more systematic approach to the designing of judgment engines, particularly when the provided "content" needs to be customized and personalized for different users.

A reminder that we are dealing for the most part with judgments of value may highlight how sensitive the proper choice of semiotic devices is. While the designers of the system will inevitably present a structure of meanings to

all users by virtue of the chosen sign vehicles, certain signification structure may steer the valuation process in some predetermined direction (the one of the most powerful interest involved in the process, or the one with the strongest valuation consensus). This situation may act as sophisticated obstruction to autonomous judgment forming, and discredit the whole concept of NDDSS. The study of such employment of "technologies of persuasion" by contemporary media, and particularly by the advertising industry, may be indispensable for the proper design of judgment engines in order to prevent such kinds of NDDSS abuse.

An integral part of the judgment engine is the further processing of judgments after their formation by the users/participants. Two main tasks require such further processing: Firstly, it may be expected that an individual user/participant of the system will provide judgments about several issues of his or her concern. This means that these judgments will have to be aggregated in such a way that an overall position of that user towards the proposed courses of action will be presented both to that user and to all other interested users. Secondly, in order to facilitate the processes of negotiation and collaboration that should follow the judgment-forming processes, the judgments and evaluations of all users need to be aggregated. This aggregation should somehow provide the overall response of the whole community involved in the decision-making process (the so-called *social landscape*). In this way, any individual participant will be able to analyze his or her position against the "background" of all the others participating in the process. These tasks can be supported by the well-developed methods of multicriteria analysis, with the important difference in employment of their results as starting points for negotiation and collaboration, instead of "ultimate solutions".

Collaboration and Negotiation Platforms

The collaboration and negotiation platform is required to support human interaction. The NDDSS concept supports an interactive, emergent approach to decision making, where interactions between participants may lead to options – and therefore decisions – which could not have been foreseen. The collaboration and negotiation platform must be designed in such a way that it will support open and transparent approaches to negotiation and limit as much as possible the opportunities for abusing the process.

As individual participants make judgments on possible courses of action and with the judgment engine assisting by aggregating those judgments into a social landscape, alliances may form on the basis of shared interests and values. This process can be facilitated by providing measures of various kinds of *closeness* between a given participant and other groups or individual participants within the social landscape of judgments made by all participants mentioned earlier. For this purpose, various techniques may be used from the field of data-driven approaches for knowledge discovery, and then particularly those dealing with classification problems (*see e.g.* **Chapter 21, Rainfall-runoff Modeling Based on Genetic Programming, Volume 1** and **Chapter 22, Evolutionary Computing in Hydrological Sciences, Volume 1**). Many recent hydroinformatics studies have explored such techniques for more standard problems of modeling various hydraulic, hydrological, and ecological systems. This valuable experience can be used in the future for designing important support components of the negotiation and collaboration platform.

Still, these processes of collaboration, cooperation, and negotiation take the form of a highly nonlinear discourse and are not at all amenable to formalization. This component of an NDDSS cannot be designed as an encapsulation of a given set of procedures to be followed. Individual-to-individual communication will entail some synthesis of email, instant messaging, web logs, and web fora, enriched with support for the management of content, including scenarios, tools for visualizing positions in a social landscape, and other similar devices.

HYDROLOGICAL MODELS AS FACT ENGINES

When dealing with NDDSSs specifically designed for water- and environment-related problems, hydrological knowledge becomes crucial for providing facts about the physical systems in question. Quite obviously, the main encapsulators of hydrological knowledge that will be incorporated into NDDSSs will be the hydrological computer models. It is not straightforward, however, to identify any particular model or even class of models that will immediately satisfy any decision-making context for which a given NDDSS is developed. There are, on the other hand, several key properties that hydrological models need to possess in order to be treated as candidates for becoming fact engines within any NDDSS.

The first required property is that these models need to be *distributed*. It seems rather obvious that this property is essential for the given purpose. All hydrological processes are spatially distributed and the requirement for distributed modeling is considered necessary solely for more accurate process representation and better prediction of hydrological responses. However, there is an additional reason that requires distributed models within the context of NDDSSs. It is the spatial distribution of the users of the system with their spatially distributed, and very often quite localized, interests. This is an essential characteristic of NDDSSs, which practically excludes all nondistributed hydrological models as candidates for becoming core fact engine components. Nondistributed hydrological modeling components

may find their role only in some cases, for minor, supporting tasks to the main distributed hydrological models.

Dealing with distributed hydrological modeling is nowadays almost exclusively done with the support of Geographical Information Systems (GIS). Some hydrological modeling tasks can even be carried out with a sophisticated GIS system alone (*see* **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**), without employing separate hydrological models. Other hydrological models depend on interfacing with GIS data for their construction and execution like in flood modeling (*see* **Chapter 16, Numerical Flood Simulation, Volume 1**). Within the NDDSSs, the power of GIS will also need to be harnessed for designing the judgment engine and the platform for negotiation and collaboration. The term "social landscape" introduced in the previous section already suggests that the presentation and analysis of the user's own judgments and positions versus those aggregated for the whole community can be best achieved through GIS-based dynamic maps (For developments in this direction, see Jankowski, 2000).

In relation to the distributed property of the employed hydrological models and the GIS support, it is important to stress that within an NDDSS the knowledge about the effects of proposed courses of action needs to be provided at a suitable, sufficiently high resolution. It can be expected that the real value of any given NDDSS to an individual user/participant will depend upon the possibility thereby provided to follow effects on certain features that correspond to his or her individual interests, such as the flooding of an individual house or street, or the lowering of the groundwater table at an individual field. While for many other corporate participants this level of resolution may not be that relevant, bringing NDDSS to the level of individual participants will be critically dependent on it (For an example related to flooding problems, see Martin and O'Kane, 2000).

The second property of hydrological models, which is essential for fact generation within an NDDSS, is that they are *physically based.* Although different terms have sometimes been used in literature, such as process-based, or theory-based models, the term chosen here, suits best for describing the needed property. Physically based hydrological models are based on mathematical representations of physical processes that occur in the hydrological system. Sometimes, the argument is put forward that most hydrological models have this property, because they are all based on satisfying at least *some* physical laws. It should become obvious that when the term physically based is used in the context of NDDSSs, we are considering models that include *as many relevant physical processes as possible*, and that they include them in an integrated manner. At present, the most advanced physically based models in this sense are those based on the concept of SHE (Systéme Hydrologique Européen), (Abbott *et al.*, 1986a,b), which allows for simulation of the whole land phase of the hydrological cycle. Examples of modeling systems of this kind are the MIKE SHE commercial package of the Danish Hydraulic Institute (`http://www.dhisoftware.com/mikeshe/`) and SHETRAN, a modeling package developed by the Water Resources Systems Research Laboratory, at the University of Newcastle, UK (`http://www.ncl.ac.uk/wrgi/wrsrl/models.html`).

If we question the reason why this property is so critical for a hydrological model to be a fact engine in an NDDSS, the answer is quite straightforward. Considering that NDDSSs are primarily going to be developed for decision-making contexts that involve planning and design, it is essential that the employed models are capable of simulating *changes* in the physical system in question. Only distributed, physically based models allow such simulations. If the processes are represented as components in an integrated manner, changes in one process component (for example, the groundwater system) which cause effects in other processes (such as the surface water component or the unsaturated zone) can be simulated and analyzed. Larger number of included processes bring more possibilities for simulating *diverse changes* in the hydrological system. The obvious importance of this property for the purposes of NDDSSs is related to the possibility of meeting diverse knowledge needs of the potential users of the system. While for many traditional hydrological problems there are situations in which a number of different models can be applicable, the new context of NDDSSs requires use of specific, distributed, and physically based modeling systems.

At the same time, however, these types of models are most difficult to be built, and most complex to understand. They are still associated with a number of problems (overparameterization being clearly one of these problems), and they need to be continuously developed. Particularly, in future they need to take the best advantage from the possibilities offered by new technologies for data collection (remote sensing, both satellite and airborne, but also the various experiences accumulated in fields such as oil exploration and meteorology). For the purposes of NDDSSs, these developments are necessary because there are no other hydrological modeling alternatives that can deliver the required performance.

This leads us to the third key property of hydrological models as fact engines, which is *integrated modeling*. Traditionally, and in full accordance with the established modern scientific approach, separate models have been developed for separate kinds of water-related problems. Hydraulic, hydrologic, water distribution, groundwater, or water quality models – all deal with separate kinds of problems. Within each class of these models there are numerous variations in terms of spatial and temporal

scales, as well as underlying assumptions, and indeed there are classes of models that are particularly suitable for particular classes of problems. Only more recently has much effort been put into providing integrated modeling systems that try to combine as many models as possible into one conceptual and computational framework. Beyond this again, seamless interfaces between modeling systems and GISs have already been extensively developed as means of integrating water-related results to corresponding geographical circumstances.

Within the framework of an NDDSS, integrated modeling gains an importance that proceeds beyond the need for providing more relevant, reliable, and accurate results. It also goes beyond the integration of hydrological processes only. The large number – and the diversity – of the needs of the users of the system imply integration of different domains as well, such as those of hydrology, ecology, economics, and even models of social behavior. An intervention in one part of the physical system in question can have such far-reaching consequences to other parts, and many modelers are already aware of this aspect. But with NDDSSs a need arises to simulate how such consequences spread across different domains.

There are many developments towards integrated modeling, and some existing modeling systems are already employing quite integrative approaches, but the above mentioned requirements will require a level of integration that is on a completely different level again. Water movement in all different phases of the hydrological cycle will increasingly have to be combined with water quality models, models of urban networks for drinking and wastewater, and models of water and wastewater treatment processes. All these water-related models will then have to be combined with ecological or other socioeconomic models. Even if a general framework of this integration is inconceivable outside of a given decision-making context, the development of the fact engine components of NDDSSs will be greatly facilitated if such integrated modeling is achieved.

The realization that hydrological models, and even more generally water- and environment-related models, will not be the only fact engines in the fact engine component, and yet they will have to be somehow integrated with other kinds of fact engines, leads us to the key problem of the fact engine design: achieving *software interoperability*, which will be sufficiently general for a given decision-making context. This problem already arises at the level of integrated modeling, of course, but in the light of the existence of other fact-providing software the extent of this problem becomes even clearer. A fact engine component of a particular system may contain dozens or even hundreds of different individual fact engines that would have to work with large dependencies on each other. Achieving their interoperability is then a formidably difficult task. General recipes for achieving software interoperability do not exist and various technologies may be employed for different tasks and domains. In the following section we will describe one recently proposed framework based on a so-called *agent orientation*. As we shall see, the framework is in fact much broader because, in addition to providing support for the key problem of the fact engine component, it can also support the design of the other two NDDSS components – the judgment engine and the platform for collaboration and negotiation.

SOFTWARE AGENTS FOR NDDSSs

Although the study of agents and agency has a long history, the emergence of agent orientation as a new conceptualization paradigm can be attributed to several recent and concurrent developments in different scientific and technological disciplines. Advances in such diverse fields as Distributed Artificial Intelligence (DAI), studies of complex adaptive systems with their applications in different areas such as biology, sociology, or economics, network distributed software development, and the explosive growth of the Internet, have all contributed to the development of this new paradigm. The notion of an agent and the understanding of the meaning, the rationale, and the justification for adopting the agent-oriented paradigm therefore vary very much, depending on the respective field of application. The common underlying ground for all these different disciplines and application fields is sought in the new kind of abstract conceptualization offered by the new paradigm of agent orientation.

As a first step, the new approach "expands" the traditional conceptualization of knowledge domains beyond their description solely in terms of different objects, with their properties and relations. In fact, it "replaces" the basic notion of "knowledge domain" with a much broader notion of an "environment", where the basic conceptualization units are not objects, but new kinds of entities called *agents*. The fundamental property of the agents is that they are embedded in their environment, and they are able to perceive and affect the environment. In addition to that, in their interaction with the environment they express goal-oriented behavior. This is in fact very close to a general definition of an agent provided by Russell and Norvig (1995, p. 31) who define an agent as any entity which operates in some environment (physical, virtual, cyberspace) and which possesses sensors to perceive this environment, effectors to act in this environment, and goals of its own which may or may not be explicitly represented in the agent itself.

One of the radical changes that this shift in the mode of conceptualization brings about is that it allows the agents to have their own representation of knowledge domains, which *may or may not be part of their own environment*. An agent's environment is therefore not to be confused with a knowledge domain. In some sense, the exclusive role of

the (human) observer in the traditional, purely rationalistic, approach of conceptualizing and representing domains in terms of objects has been now attributed to these new kinds of entities called agents. This is certainly one of the most fundamental changes brought about with the new approach, primarily because it allows for different perspectives (or points of view) between the human observer and the agent, and/or between the different agents themselves.

As we have mentioned, this same paradigm is being developed and applied within several very different disciplines. For example, the very broad area of computer-simulation-using agents has already been applied in many different fields, such as biology, ecology, sociology, economics, business or industrial process management, transport, and others. In all these fields, the agent-based simulation is used to observe some emergent phenomena which result from the individual actions of large number of individual agents. Of course, in different domain areas, the agents represent different entities (individual organisms in ecological simulations, human agents in social simulations, customers, and producers in economic simulations, etc.). For examples of agent-based simulations for water resources management combined with socioeconomic aspects, see for example, the European research project FIRMA (`http://firma.cfpm.org/`).

However, the advantages of agent orientation for the design and development of NDDSSs are in another wide area, namely, that of *software agents*. In the area of software engineering, the shift towards agent orientation means primarily the use of software agents as more useful, higher level mental abstractions, which can be beneficially used for software development. There are two general streams of software agent development. The first one is concerned with the most important characteristic of complex distributed software systems, which is the one of *interaction*. This stream deals with *collections* of agents, which are most commonly described as Multi Agent Systems (MASs). The second stream of software agent development is concerned with construction of individual agents, where the stress is much more on developing *intelligent* software component capable of learning and adaptation. A very important application area for these agents is the construction of intelligent user interfaces for human interactions with computer systems. Of course, both classes can benefit from each other's properties (MASs involving learning and adaptive agents, and individual agents using interaction with other agents for improved performance), and the developments frequently overlap.

Software agents are best defined as "computer systems *situated* in some *environment*, and capable of *autonomous* action in this environment in order to meet their design objectives" (Wooldridge and Jennings, 1995). The environments of software agents are determined by the nature and tasks in an actual application, while they are generally described with various properties and the degree to which these properties are satisfied (accessibility, determinism, static versus dynamic, discrete versus continuous, etc; Russell and Norvig, 1995, p. 46).

The important term in the above definition of a software agent is "autonomous action". This is one basic property of software agents that distinguishes them from the more familiar software *objects*. As summarized by Wooldridge (2002, p 21):

- Agents embody a stronger notion of autonomy than objects, and, in particular, they decide by themselves whether or not to perform an action on request from another agent;
- Agents are capable of flexible (reactive, pro-active, social) behaviour, and the standard object model has nothing to say about such behaviour
- A multiagent system is inherently multithreaded, in that each agent is assumed to have at least one thread of control

It should be noted that the autonomous action also covers the area of interagent communication, collaboration, competition, and other types of social behavior. Such behaviors in agent orientation are considered possible only via autonomous action of the involved agents. The overall objectives of the designers of the systems are achieved as emerging from interactions of the individual agents. For designing complex software systems, in which it is rather difficult to specify the operation of the system as a whole, the agent orientation approach is a viable alternative.

Without going into the details of the various agent architectures, we will now focus on the properties of MASs and of individual, and particularly *interface* agents, which are relevant for design and development of NDDSSs (For an overview of different agent architectures, and their different applications see Russell and Norvig (1995); Wooldridge and Jennings (1995); Nwana (1996); and Wooldridge (2002).)

The social ability of agents in MASs is dependent on the ability of the agents to communicate with each other. This is usually achieved by making use of specially designed, highly expressive Agent Communication Languages (ACLs). Agent communication is of such importance in MASs that some researchers propose that the ability to communicate through a high-level communication language is the first criterion for agenthood. The most widely used ACL is the so-called *KQML* (Knowledge Query and Manipulation Language), developed by Stanford University. It is based on speech act theory, where, like most of the approaches used in developing ACLs, the message exchanged between the agents is divided into two parts. The first part conveys the *intent* of the sender to the receiver (whether it is a request, question, response etc.) for which KQLM uses the so-called *reserved performatives*. They determine the kinds of interactions that are possible with a KQLM-speaking agent. The second part of the message is the message itself, usually called the *content* of the message

that is attached to a performative. Another similar ACL language is the so-called *FIPA-ACL*, developed by an Agents Standards body called the *Foundation for Intelligent Physical Agents* (FIPA).

Second key components for social interactions within MASs are the so-called *ontologies*. Ontologies are the formal specifications of conceptualization of any particular domain about which the agents communicate. They are the "vocabularies" that are used to build the content of the messages in a consistent way. They are in fact very important for interagent communication, regardless of the ACL in question, because they contain the descriptions of the domains about which the agents exchange messages. At the same time, however, the "ontology problem" is probably one of the most significant in the whole field of agent-oriented software engineering (see Nwana and Ndumu, 1999). It relates to the fact that the creation of an ontology for a given domain, which can then be used by various agents, still requires an *explicit definition of every concept that needs to be represented*. Moreover, same domains can have different ontological descriptions for different *tasks*. General-purpose ontologies are therefore still practically impossible, and this turns the developers towards building domain-specific ontologies.

We now turn to the primary problem of the fact engine component of a given NDDSS – the problem of software interoperability – and see how this can be approached from the MAS perspective. The heterogeneous software components introduced earlier, which constitute the fact engine component such as the various models, databases, documents, or expert systems, can become equipped with communication capabilities using ACL. Existing software applications may actually be converted into agents, by "wrapping" them with such a communication layer, while new applications can possibly be equipped with communication capabilities of this kind from the beginning.

Such a design approach would then mean that each agent from the agent component would basically become capable of two kinds of actions, which we can call *private* and *communicative*. The private actions are those related to the internal tasks for which a given software component is built (execution of a model run, extraction of a query result from a database or a knowledge base etc.). For its private actions each agent operates within its own, internal, and rather restricted domain, and the way in which such agents are implemented is independent of the processes in the agent communication layer.

In order to be able to exchange meaningful messages, the agents will need to use a common ontology (or several common ontologies). In a concrete decision-making context, the fact engine component will still operate in a relatively controlled and bounded environment, where both the domain and the tasks can be well defined, and the use of common ontologies will be feasible.

In analyzing the second class of individual agents, we will focus on the so-called *interface agents*. These are not predominantly developed for operation in a society with other agents but as assistants to human users when using a particular application. These agents change the way in which the creation of user interfaces for software applications is understood. Most existing user interfaces are built for tasks that involve closed, static, relatively small, and well-structured application domains. Furthermore, they rely completely on direct manipulation by their users. Interface agents (sometimes called user assistants, or personal assistants, or expert assistants) assume that the applications operate in an open, dynamic, and sometimes unstructured and unpredictable environment (e.g. the Internet). Interface agents are active without their users' direct manipulation and they try to match the user's interests, habits, and preferences with the changing conditions that they experience in their dynamic environment. Crucial properties of interface agents are their ability to learn from the actions of their human users and adapt to changes in the environment or to changes in user preferences or behavior.

These properties of interface agents are actually quite adequate for the requirements of the judgment engine component of an NDDSS. Firstly, in order to achieve its main function, the judgment engine component has to provide customized and even personalized interfaces for users who can have very diverse beliefs, intentions, and interests. At the same time, the whole "background" against which the users are forming their judgment is dynamic and may change over time (users with new kinds of interests join in, new facts are being provided, the social landscape changes as a result of negotiations, etc.). This means that these interface agents will have to deal both with the changing attitudes of their users and with such changing situations in this "background environment". These aspects make the judgment component of an NDDSS an "open" application domain, open also to its human agents.

The most recommendable approach would certainly be if these agents could be built with the abilities to learn and adapt to continuous changes in both the user attitudes and in the "background environment". The advantages from having learning and adaptive interface agents will be twofold. Firstly, such interface agents will be able to act as true "personal assistants", as they will be able to suggest beneficial actions to the user in a changing dynamic environment. Secondly, such agents would in principle be capable of "replacing" the human users and continuing to act autonomously on their behalf when they are absent. The achievement of this second advantage involves the difficult issue of trust that the human users can put on such autonomous applications.

It has to be realized, however, that building interface agents with such sophisticated learning and adaptation capabilities is extremely difficult. As Nwana and Ndumu (1999)

point out, the design of such agents has to meet the so-called *intelligent tutoring and mentoring problem*, for which the agent is required to have four key knowledge modules: domain knowledge on the application, a model of its user, strategies for assistance, and some sort of a catalogue of the typical problems (situations) that arise within the particular environment. These requirements suggest that such agents need necessarily to use deliberative architectures, with explicit representation of the required knowledge in these four modules. The alternative approach of Maes (1994) of using more reactive architectures built out of competing competence modules that produce emergent learning and adaptive behavior, has been successful in several applications, but in domains that are still rather simple. This area is still under active research, but it seems that some kind of hybrid architecture will be needed in order to meet the requirements of complex application domains like the judgment engine component of an NDDSS.

Because such applications are still an intention rather than a reality, the judgment engine interface agents can, as in many other cases besides, take a more simplified approach. Instead of continuously learning every "move" of its user, the agent can use the so-called *user profiles* created from the user input, and then try to "match" situations occurring in the "background environment" with such representations of the user. This approach is similar to applications for providing customized content, as exemplified by many existing Internet applications.

An agent-based NDDSS design will also need to address the problem of connecting the interface agents to the heterogeneous agents that constitute the fact engine component. With the envisaged large number of interface agents that will serve all the users/participants and the diversity of the facts that they will require for supporting the judgment-forming process, it may become very difficult to maintain direct connections between the interface agents and the fact engine agents. Such direct connections may be very expensive in terms of computational resources and prohibitively increase the complexity at the implementation level.

In order to overcome problems of this kind, one common solution is that of employing another kind of agents, the so-called *facilitators*. Instead of using direct communications, the interface agents communicate through these agents, which thus act as their intermediaries. The facilitators, envisaged to be located in the knowledge center, establish and maintain connections across all interface agents and ensure a coordinated exchange of messages between them and the heterogeneous agents acting as fact engines. In communicating with the fact engines, the facilitators then need to be capable of using the same ACL as all other heterogeneous agents. All the agents involved can then be organized in a so-called *federated system*, where the communication overhead is significantly reduced.

The combination of fact engines interoperating through an ACL, and interface agents communicating with them via facilitator agents, is, in fact, also suitable for the requirements of the platform for collaboration and negotiation. The collaboration and negotiation processes involving "many parties and many issues", are impossible to be automated and will essentially be carried out by the human participants. Nevertheless, the system still needs to provide support for these processes and then both at the users' nodes and in the knowledge center. At the "users" nodes, the interface agents can have their role extended so as to provide both the social landscape emerging from the interactions of all participants and the embedded position of that particular user in such a landscape. For this purpose, the interface agents need to act as a kind of "medium" through which the "moves" of all the participants are made transparent. This transparency should enable the users to discern the actions by which they can orientate their attitudes relative to one another, and therefore enhance the collaboration process. This implies that the interface that is actually presented to each user will depend on the actions of all the other users/participants, which means that the interface agents need to interact *among themselves.* With the large number of interface agents, the so-called *broadcast message-passing* approach seems again infeasible for this purpose, and the solution needs to be sought again in the federated system, making use of the facilitator agents that act as intermediaries between the interface agents.

Apart from this basic technical task of coordinating the communication exchange between the interface agents, the facilitator agents will need to encapsulate all the necessary knowledge for supporting the collaboration and negotiation processes. These agents therefore need to take upon themselves the necessary tasks of aggregation of the individual judgments and positions, as well as the creation and updating of the social landscape presented to the users through the interface agents. The support in terms of calculating similarities of users' positions that can be used for coalition and alliance building, or the determination of alternatives with mutual gains, are also tasks to be handled by the facilitator agents. The facilitator agents may integrate these additional tasks within their own agent architecture, or they can maintain separate connections with other "back-end" agents specially built for this purpose.

This general overview of the software agents' potential for designing NDDSSs does not touch any of the numerous difficult implementation issues. Any concrete NDDSS design and development problem will certainly open up many such implementation issues, and possibly even raise the question of whether the agent-oriented approach is suitable at all at the implementation level. Nevertheless, there seems to be a lot of value in adopting the agent-oriented approach at the conceptual level. The essential problem in

the design of NDDSSs is the integration of the three functional components into one design framework. The tasks that these three components need to perform are very different, and the use of more traditional design methodologies may lead to design specifications that are too complex. This in turn will make the implementation even more difficult. The adoption of an agent-based approach offers the possibility to provide an abstract, high-level conceptual design, which is comparatively less complex, and which emphasizes the key integration issues. Once the entire system is decomposed and specified in terms of agents, many implementation issues can be approached separately for each agent.

In conclusion, it is obvious that fully developed NDDSSs, as they are proposed here, are still some way removed in the future. Their design and development still requires a lot of research, testing, and prototyping with different agent architectures for different decision-making problems. It is entirely possible that in some cases the agent-oriented approach may not be required at all. As a general approach to NDDSSs design and development, however, the software agent technology clearly has a lot of potential, and its advances elsewhere need to be followed closely and constantly related to the present applications.

THE FUTURE FOR NDDSSs

While the concept of NDDSSs, including the technical possibilities for their development, is quite developed, the question of their introduction in practice still remains open. The need for systems like NDDSSs becomes more and more apparent as we are facing increased complexity in the water-related decision-making problems. On the other hand, there seems to be enough human and technical potential for developing full-fledged real-world NDDSS applications. As with any new development of this kind, the actual realization may critically depend on the existence of certain supportive social ethos. We can postulate that the locus of possible social support for such systems with sufficient inclination towards their real-world application may emerge in three different ways:

1. Introduction of Internet-distributed NDDSS-like systems in the form of multiplayer role-play games. This approach is also known as the *edutainment* (education + entertainment) approach. In order to be successful, this approach requires strategic alliance between the developers of more traditional hydroinformatics tools and systems and professional game developers. It is envisaged that both groups may benefit from such a development, even though difficulties may arise as to what the right balance between "real-world" and "entertainment" content is. If the balance is not appropriate, such games can teach unintended lessons, and those lessons may be counterproductive. On the other hand, successful developments of such scenarios would mean that through the penetration of such games into the culture of our societies, the realization of the potential benefits of the NDDSS approach might be significantly increased.
2. The second locus of support for development and introduction of such systems may come from governmental organizations and various Non-governmental Organizations (NGOs) involved in water- and environment-related issues. In most countries, water- and environment-related problems fall mainly under governmental jurisdiction, even if they involve many other, and different, interests. This makes the governmental organizations an obvious place from where such systems can be disseminated. Particular NGOs, on the other hand, may become natural supporters of such systems, since they promote most of the things for which these organizations stand for. One could even envisage partnership between governmental organizations and NGOs on development of such systems.
3. The third approach for developing of such systems may be directly through the market. This is a more likely scenario for developed countries and societies, where concerned citizens and institutions can afford to pay for services offering relevant water-related knowledge and information. It is an approach that looks directly at the opportunities offered by the processes of knowledge commodification. It is very often stressed how these processes are driven by technology, such as continuously improving Internet solutions, or the introduction of third-generation (3G) telephone networks. At the same time, this process of constant introduction of new, more efficient, technologies is countered by the lack of appropriate and useful "content" that is to be delivered through them. Water- and environment-related problems are certainly one area where fast and always-available personalized content delivery is needed. The initiatives for such businesses need to come again primarily from organizations involved in the development of hydroinformatics systems, but now the strategic alliances need to be sought with market-oriented companies that are promoting new technologies and are searching for appropriate and useful content. The involvement of government in water-related issues indicates that such attempts may need to be developed through strategic public-private partnerships.

The three proposed approaches only point towards the potential areas from where the introduction of NDDSSs into our societies may be taken up. Clearly, for different decision-making contexts, and especially for different societies and cultures, different approaches or combinations of approaches will be suitable.

As a general conclusion for future developments, it can be envisaged that forthcoming developments in the field of information and communication technologies (as understood in the term *informatics*) are likely to continue to change all fields of hydrological sciences – which in essence, is the contribution from the field of hydroinformatics.

Acknowledgment

The author is grateful to Prof. Michael B. Abbott for his inspiring ideas, which enabled the research of Network Distributed Decision Support Systems, and to UNESCO-IHE Institute for Water Education for supporting this rather unconventional research topic within the framework of a PhD study.

SOFTWARE LINKS

http://www.dhisoftware.com/mikeshe/
Integrated hydrological modeling software package MIKE SHE, DHI, Water, and Environment
http://www.ncl.ac.uk/wrgi/wrsrl/models.html
Integrated hydrological modeling software package SHET RAN, Water Resources Systems Research Laboratory, University of Newcastle, UK.

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986a) An introduction to the European hydrological system – système hydrologique Européen 'SHE', 1: history and philosophy of a physically based distributed modelling system. *Journal of Hydrology*, **87**, 45–59.

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986b) An Introduction to the European hydrological system – système hydrologique Européen 'SHE', 2: structure of a physically based distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

FIRMA – 2003 Freshwater Integrated Resources Management with Agents (European research project) **http://firma.cfpm.org/**.

Gergen K.J. and Thatchenkery T. (1997) Organizational science in a postmodern context. *Journal of Applied Behavioral Science*, **32**, 356–377.

Jankowski P. (2000) Collaborative spatial decision making in environmental restoration management: an experimental approach. *Journal of Hydroinformatics*, **2**(3), 197–206.

Jonoski A. (2002) *Hydroinformatics as Sociotechnology: Promoting Individual Stakeholder participation by Using Network Distributed Decision Support Systems*, Sweets & Zeitlinger B.V.: Lisse.

Jonoski A. and Abbott M.B. (1998) Network distributed decision support systems as multi-agent constructs. In *Hydroinformatics '98*, Babovic V.M. and Larsen L.-c (Eds.), Balkema: Rotterdam.

Klinkenberg J.-M. (1996) *Précis de Sémiotique Générale*, De Boeck & Larcier S.A..

Loucks D.P. (2003) Managing America's rivers: who's doing it? *International Journal of River Basin Management*, **1**, 21–31.

Lyotard J.-F. (1979) *La Condition Postmoderne: Raport Sur le Savoir*, Minuit: Paris; (1984) *The Postmodern Condition: A Report on Knowledge*, Manchester University Press.

Maes P. (1994) Modeling adaptive autonomous agents. *Artificial Life Journal*, Langton C. (Ed.), MIT Press: Vol. 1, No. 1 & 2, pp. 135–162.

Martin J. and O'Kane J.P. (2000) Development of a high resolution hydrodynamic flood mapping model using one-dimensional hydraulic models integrated with high resolution digital elevation models, *4th International Conference Hydroinformatics 2000*, Cedar Rapids.

Nwana H.S. (1996) Software agents: an overview. *Knowledge Engineering Review*, **11**(3), 205–244.

Nwana H.S. and Ndumu D.T. (1999) A perspective on software agents research. *Knowledge Engineering Review*, **14**(2), 1–18.

Russell S.J. and Norvig P. (1995) *Artificial Intelligence: A Modern Approach*, Prentice Hall: New Jersey.

Woods B. (1993) *Communication, Technology, and the Development of People*, Routledge: London.

Wooldridge M. (2002) *An Introduction to Multiagent Systems*, John Wiley & Sons: Chichester.

Wooldridge M. and Jennings N.R. (1995) Intelligent agents: theory and practice. *Knowledge Engineering Review*, **10**(2), 115–152.

PART 3

Meteorology and Climatology

25: Global Energy and Water Balances

ATSUMU OHMURA

Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

The earth's energy and water balances are summarised with the most recent observational and computational results. A special reference is given to the earth's surface. After a brief discussion of the limits of the observational accuracy, the author explains the mean state of the energy balance starting with the sun, through the atmosphere and to the earth's surface. The main contents are the Total Solar Irradiance (earlier called solar constant) and its variability, the climatic effect of the earth's orbital change, extinction processes of solar radiation in the atmosphere, the distributions of solar radiation on the earth's surface, reflection and albedo, terrestrial radiation from the atmosphere and the earth's surface, net radiation, sensible and latent heat fluxes, subsurface heat flux, and latent heat of melt. The distributions of the individual components are presented in maps. The important points of geographical distributions and energy fluxes are explained in the text. The energy balance climatology in the present article differs from earlier works mainly on three topics: 1) larger atmospheric absorption of solar radiation; 2) smaller solar radiation on the earth's surface; and 3) larger terrestrial counter radiation. These new situations are the results of the recent advances in spectrometry, observational technology both on the earth's surface and in the space, and also in computational skill. Therefore, the historical development of the understanding of the earth's energy balance is added. The article ends with a brief description of the mean hydrological cycle, allowing unsteady conditions due to the melt of glaciers during the twentieth century.

INTRODUCTION

Solar radiation is practically the only energy source for the earth's climate system. The global climate is to a great extent determined by the equilibrium maintained between the solar radiation absorbed by the atmosphere and earth's surface system (planetary system) on one hand and the emission of terrestrial radiation to space on the other. A destabilization of the equilibrium leads to climate changes. The entire cycle of energy flow in the climate system begins with the absorption of solar radiation by the atmosphere and the earth's surface, whereby the latter plays a dominant role. The major transformation of the solar radiation into enthalpy and latent heat happens at the earth's surface. Of these two components, the latent heat of evaporation is by far more important. Further, the water vapor contributes upon condensation for heating the middle and upper troposphere. This is also the process through which the energy and hydrological balances are closely connected. Globally averaged, the earth's surface gains more energy than it loses as a result of radiation exchanges. In the meridional direction, the radiation balance of the planetary system creates a net energy gain in lower latitudes and a loss in higher latitudes. This is how the net flow of enthalpy and water vapor takes place from the earth's surface to the atmosphere and from the lower to higher latitudes. The major climatic and hydrological phenomena are closely associated with this equalizing process of uneven energy balance. These processes will be quantitatively presented. In considering energy fluxes, accuracy is of prime importance. The absolute limit in accuracy for radiation measurements is the accuracy limit of the World Radiometric Reference (WRR). The current limit in the WRR is estimated at $\pm 0.3‰$ or approximately $\pm 0.3\,\mathrm{W\,m^{-2}}$. The highest accuracy in the operational measurements of all energy fluxes is attributed to direct solar radiation. The smallest unsystematic error of direct solar radiation is about $1\,\mathrm{W\,m^{-2}}$. To achieve this accuracy, a pyrheliometer of the absolute cavity type must be used without a filter. If a commercially available pyrheliometer with a thermopile-type sensor and a quartz filter are used, errors larger than $\pm 2\,\mathrm{W\,m^{-2}}$ must be expected. Pyranometers for 2π-irradiance measurements

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

have an unsystematic error larger than ±2.5% or typically ±5 W m^{-2}. The same error for a pyrgeometer for longwave measurement is estimated at ±8 W m^{-2}. The heat flux of conduction into and out of the ground can be determined with a typical error of ±3 W m^{-2}, mainly because of the nature of the small absolute value of this flux. The errors of turbulent heat fluxes are usually much larger and more difficult to estimate than for the fluxes of radiation and conduction. It is not uncommon to experience differences of up to 30% between two instruments located side by side. These errors are estimated with the data obtained under favorable conditions. In reality, the influence of adverse weather conditions such as rain, snow, and frost on instruments reduces the above accuracy. The most important factor for the accuracy, however, is the experience, skill, and dedication of the observers. Energy fluxes are difficult to measure without personal interest and an alert mind. For a general discussion of the observational errors for irradiances, see Ohmura *et al.* (1998).

THE SUN

The intensity of the solar radiation received without atmosphere at the mean sun/earth distance is defined as solar constant. The recent direct observation of the solar constant from space provided evidence that the solar constant is in fact variable. As a result, we apply today to this concept a new term, Total Solar Irradiance (TSI). The TSI has been monitored from space since 1978 with a comparable accuracy and scale to the present level. Since then a quarter century has passed. The time series in Figure 1 which is due to Fröhlich (2004) and a recent supplement by C. Fröhlich (personal communication) shows four important facts: (i) TSI changes with 11 year sunspot cycle; (ii) the amplitude of the 11 year cycle is the order of 0.6 W m^{-2}; (iii) beyond the 11 year periodicity, there is no statistically significant trend; and (iv) its mean value is 1366.0 W m^{-2}.

There are some publications that indicate an increasing trend in TSI, offering an alternative explanation for the present warming (Willson, 1997). These works were subsequently found to have ignored certain key corrections of the instruments. This does not mean that TSI does not change over a longer period. It was simply not detected during the last 25 years of observation. There have been a number of attempts to relate the long-term variability of TSI to some observable phenomena such as sunspot number, length of the sunspot cycle, or ^{10}Be and ^{14}C concentrations in the atmosphere, but all these hypotheses are highly speculative. Because of the spherical shape of the earth, one-quarter of TSI, 324 W m^{-2} can be regarded as the mean solar irradiance at the top of the atmosphere (TOA), and hence the primary energy source for the climate system.

EARTH'S ORBITAL EFFECT

The solar irradiance at TOA is completely describable with three orbital parameters, eccentricity, longitude of perihelion (seasonal variation of the distance to the sun), and the tilt of ecliptic. The present eccentricity of the earth's orbit is 0.0167 (Berger, 1978), a rather small value within the

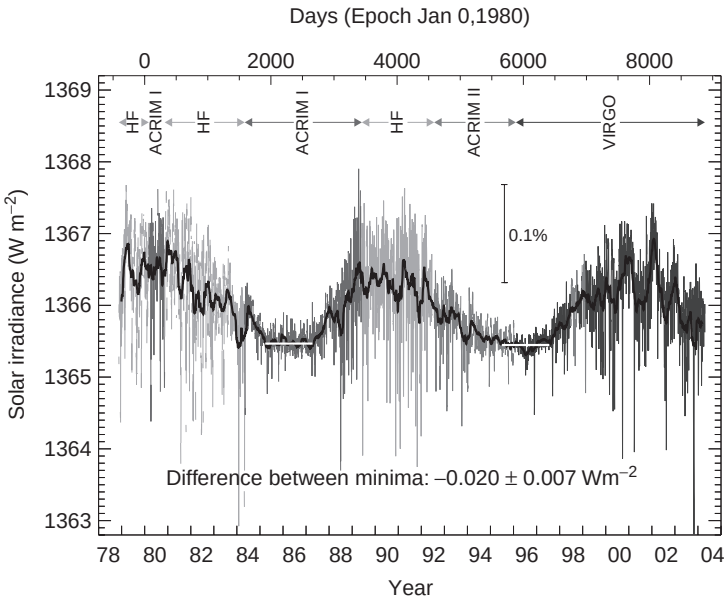


Figure 1 Observed total solar irradiance (solar constant) from space.Fröhlich (2004) supplemented by C. Fröhlich (personal communication). The long-term mean TSI is 1366 W m^{-2}. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

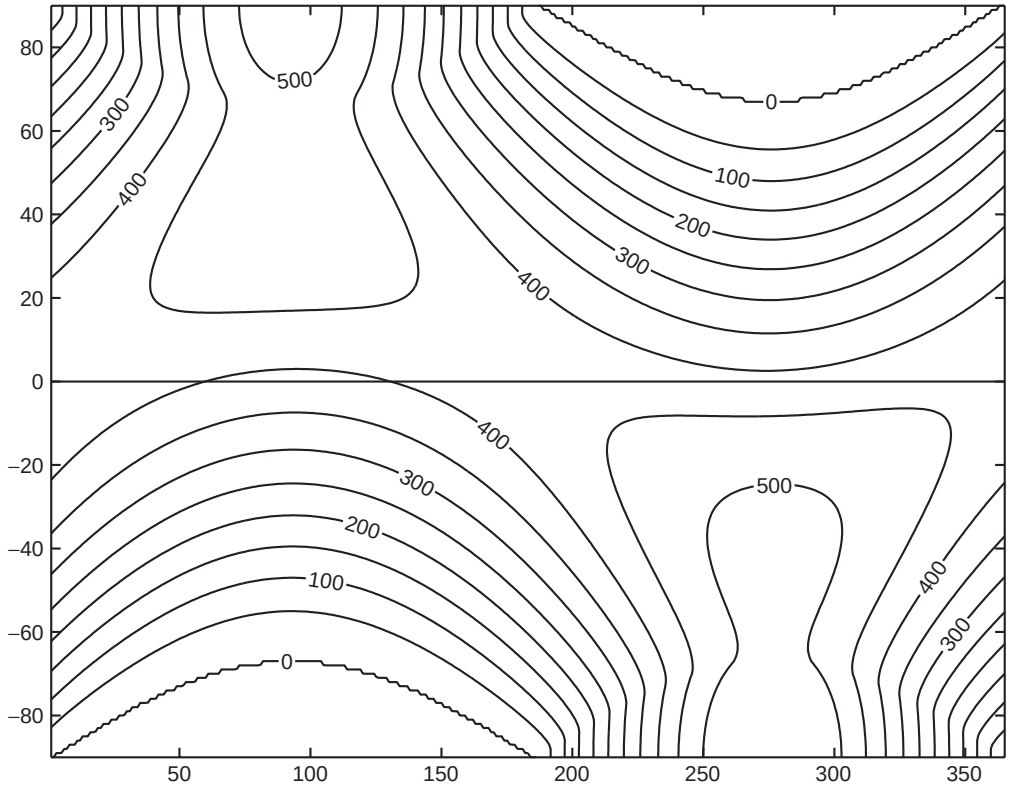


Figure 2 Irradiance at the top of the atmosphere under the present orbital parameters. The asymmetry between both hemispheres is due to the present eccentricity of 0.0167

entire Quaternary. This small eccentricity, however, yields a significant variation of the solar irradiance of ±3.5% between the perihelion and aphelion at TOA. The seasonal course of the TOA solar irradiance is presented in Figure 2. The earth is presently at the perihelion around January 4 and at the aphelion around July 4. Consequently, the Southern Hemisphere receives 7% more TOA solar radiation than the Northern Hemisphere, in their respective summers. Whetherald and Manabe (1975) show that 1% variation of the TOA solar irradiance causes approximately 1 K temperature change at the earth's surface or at the bottom of the atmosphere (BOA). The orbital effect produces a rather significant contrast in climate between both the Hemispheres. For example, the serious situation of the Ultraviolet radiation in summer in the Southern Hemisphere is caused mainly by this astronomical condition, rather than its proximity to the ozone hole over the Antarctic.

EFFECT OF THE ATMOSPHERE

Solar radiation undergoes a substantial transformation in the atmosphere before reaching the earth's surface. The important processes in the atmosphere are Rayleigh scattering due to molecules, Mie scattering by aerosol, reflection by clouds, and absorption by all these components. The climate system loses 30% (Barkstrom *et al.*, 1990) of the primary solar radiation back to space by reflection. The reflective loss is 3% due to Rayleigh scattering, 1% due to Mie scattering, and 22% due to reflection by clouds, the remaining 5% being the reflection from the earth's surface that will be discussed later in detail. This total reflectance of the earth is called a *planetary albedo*.

The atmospheric absorption of solar radiation is of medium strength. This is a process presently far better understood than it was 20 years ago. Earlier the earth's atmosphere was considered very transparent for solar radiation. An absorption rate of 17 to 20% of the primary solar radiation was most often suggested. Recent investigations based on field experiments and numerical computations indicate that the atmosphere absorbs more than 25% of the primary solar radiation (Budyko *et al.*, 1988; Ohmura and Gilgen, 1993). The most likely atmospheric absorption is $28 \pm 2\%$. The gross underestimation in the past is due to an improper treatment of the near-infrared (IR) absorption by water vapor and the neglect of a number of weak absorption bands in trace gases. Although they exert only weak absorptions individually, the effect of their sum can be significant. It has also become known recently that the effect of clouds on absorption of solar radiation was often overestimated in the past (Ohmura and Raschke, 2005). The strength of the

Table 1 Atmospheric absorption with MODTRAN

$[H_20]25$ mm we	70.1 W m^{-2}
$[O_3]300$ DU	13.1
$[CO_2]370$ ppm	3.8
$[CH_4]1.7$ ppm	1.1
Aerosol $\tau = 0.1$, $\omega = 0.95$	5.1
Cloud (ISCCP)	2.8
Total absorption	96.0

Source: TOA solar irradiance (342 W m^{-2}).

absorption by individual components of the atmosphere is presented in Table 1. Of all these components, water vapor and clouds deserve further discussion. Water vapor, which occupies only 0.24% of the atmosphere, plays the most important role in the radiation budget of the atmosphere. It absorbs 70 W m^{-2} solar radiation. Previous underestimation of this effect was often due to a heavy reliance on Fowle's equation (Fowle, 1915). Many works also neglected the absorption in the wavelength range beyond 2.8 μm, where more than 30 W m^{-2} of TSI is contained and water vapor absorption is strong. The effect of clouds on atmospheric absorption is rather modest. Clouds have two counteracting effects. Firstly, they enhance the absorption of solar radiation through the elongated optical paths within the cloud layer as a result of the multiple scattering by cloud droplets. The increased optical paths make weak and intermediately strong absorptions, strong enough to be considered. A minor contribution is also made through the absorption by cloud droplets. Secondly, strong and diffuse reflection at the upper layer of the clouds promotes the absorption by the atmosphere above the cloud layer. Finally, the existence of the cloud layer deprives the atmosphere below of potential absorption because so much solar radiation is already lost before reaching this level. The recent database from the International Satellite Cloud Climatology Project (ISCCP), Level D2 (personal communication by W.C. Rossow) indicates that the global mean cloud effect on the solar radiation absorption in the atmosphere is 3 W m^{-2}, that is, just about 1% of the primary solar irradiance. The other substances, such as ozone, carbon dioxide, methane, and aerosol, absorb 23 W m^{-2}, making the total absorption of the solar irradiance 96 W m^{-2}, corresponding to 28% of the primary radiation from the sun.

SOLAR RADIATION AT THE EARTH'S SURFACE

Solar radiation at a point on the earth's surface is determined by TSI, the earth–sun distance, the solar zenith angle, and the extinction effect of the atmosphere. Solar radiation arrives at the earth's surface in two components, direct solar and diffuse sky radiation. The sum of the two is referred to as *shortwave incoming radiation* or, in brief, *global radiation*. On a cloudless day in a clear atmosphere, up to 90% of the global radiation is made of direct radiation, while on an overcast day 100% of the global radiation can be due to diffuse sky radiation. Globally averaged, direct and diffuse radiation divide global radiation at about 50 to 50%. Global radiation is presently the best-known component among all fluxes at the earth's surface. The wide range of the proposed values of global radiation in earlier literature can be narrowed down with presently available observational data of high quality. Satellite investigations have fixed the long-term mean value of the planetary albedo at 30%. The analysis based on Table 1 shows that the atmosphere absorbed as much as 28% of TOA solar irradiance. These values state that the absorption of global radiation by the earth's surface must be around 42% of the TSI, that is, between 142 and 147 W m^{-2}. This consideration eliminates all works other than Budyko (1963), Budyko *et al.* (1988), Ohmura and Gilgen (1993), Ohmura and Raschke (2004), and the numerical simulation by ECHAM4 (Wild *et al.*, 1998), and indicates that globally averaged shortwave incoming radiation at the surface must be in the range of 170 W m^{-2}, which is supported by the recent reevaluation of global radiation (Ohmura and Raschke, 2004). Further, global radiation has large spatial and seasonal variations. The following discussions on geographical distribution of the fluxes will be made on the basis of ECHAM4 simulations. This simulation is considered to represent most accurately the seasonal distribution of energy fluxes in the global scale.

ECHAM4 is a general circulation model (GCM) developed at the European Centre for Medium-range Weather Forecasts (ECMWF), adapted for climate simulations at the Max-Planck-Institute for Meteorology at Hamburg. This model is equipped with one of the most advanced radiation codes. The model was used to simulate the present climate in a high-resolution grid-equivalent of 120 km by using the climatological sea-surface temperature during the period 1979–1988. In many regards, the fluxes computed in these simulations can be considered to offer one of the most realistic distributions of surface fluxes.

The annual mean shortwave incoming radiation (global radiation) at the earth's surface in Figure 3 ranges from the smallest value of 70 W m^{-2} over the Barents Sea in the North Atlantic to the largest value of nearly 300 W m^{-2} in the subtropical East Pacific. Generally, the largest global radiation is not found under the equator, despite its largest TOA irradiance in the world, because of the cloud development associated with the Intertropical Convergence Zone (ITC). The largest global radiation appears in the subtropical horse latitudes where potential solar radiation at TOA is still large and the atmosphere is relatively dry and cloud-free. On land this is the major desert regions of the world, such as, the Sahara, Arabian, Kalahari, Australian, Mojave, and Atacama. In these regions, the annual global

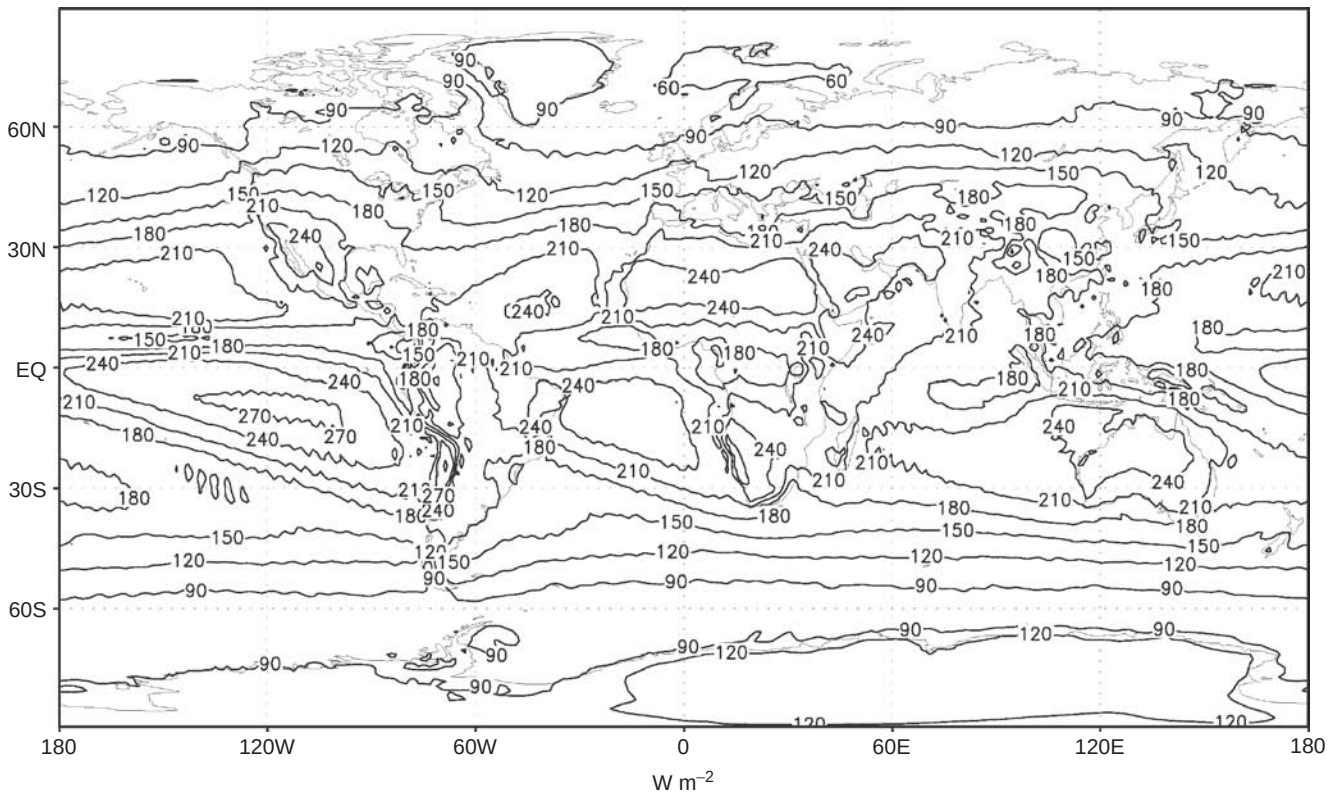


Figure 3 Annual global radiation computed by ECHAM4T106. Among all general circulation models (GCM), ECHAM4 developed at the European centre for medium-range weather forecasts (ECMWF), adapted for climate simulation at max-planck-institute (MPI) for meteorology, Hamburg most closely simulates the present distribution of shortwave incoming radiation (global radiation). The global annual mean on this figure is 170 W m^{-2}

radiation exceeds 250 W m^{-2}. The Gobi and Takla Makan are excluded from this category because of their locations at higher latitudes. The largest instrumentally measured global radiation on land was observed at Riyadh, on the Arabian Peninsula and at Tammanrasset in the Sahara, where global radiation reaches 260 W m^{-2}. Yet, the world largest global radiation is probably not found on land, but over an ocean. The area of the tropical East Pacific between the ITC and the Southern Pacific Convergence Zone (SPCZ) is considered to be the sunniest spot in the world. In this area, the annual global radiation exceeds 270 W m^{-2}. The largest annual global radiation of 298 W m^{-2} is reported at Canton, on Canton Island in Kiribati. This is basically a westward extension of the desert climate to the ocean, but the atmosphere contains much less aerosol in comparison with the sites in the continental desert. The middle latitudes are the regions of a large gradient of global radiation with a rate of -3 W m^{-2} per one degree of latitude. The smallest global radiation does not appear right at the pole but in subpolar regions from latitudes 60° to 70° in both Hemispheres. These are regions of larger cloud amount owing to the subpolar cyclones. From here toward the poles, global radiation increases. The higher surface albedo in the Arctic Ocean and the Antarctic continent also enhances global radiation through multiple reflection between the surface and the atmosphere. At the North Pole, annual global radiation is estimated at about 90 W m^{-2}, while at the South Pole the long-term mean of the observed annual global radiation is found to be 130 W m^{-2}. The difference in annual global radiation between North and South Poles is caused by the earth's orbital effect the difference in altitudes, cloud conditions, and the surface albedo.

Globally averaged, the earth's surface receives 7% more global radiation in January than in July because of the occurrence of the perihelion in January and the present eccentricity of the earth's orbit around the sun amounting to 0.0167.

The seasonal fluctuation of global radiation is very large except for the equatorial zone as is witnessed in the example of Singapore shown in Figure 4. In regions poleward of 10° N and S, the global radiation experiences considerably large seasonal fluctuations. The distribution of global radiation in winter is strongly affected by the solar elevation, hence latitudes, while in summer it is greatly influenced by the distribution of clouds. Thus, in

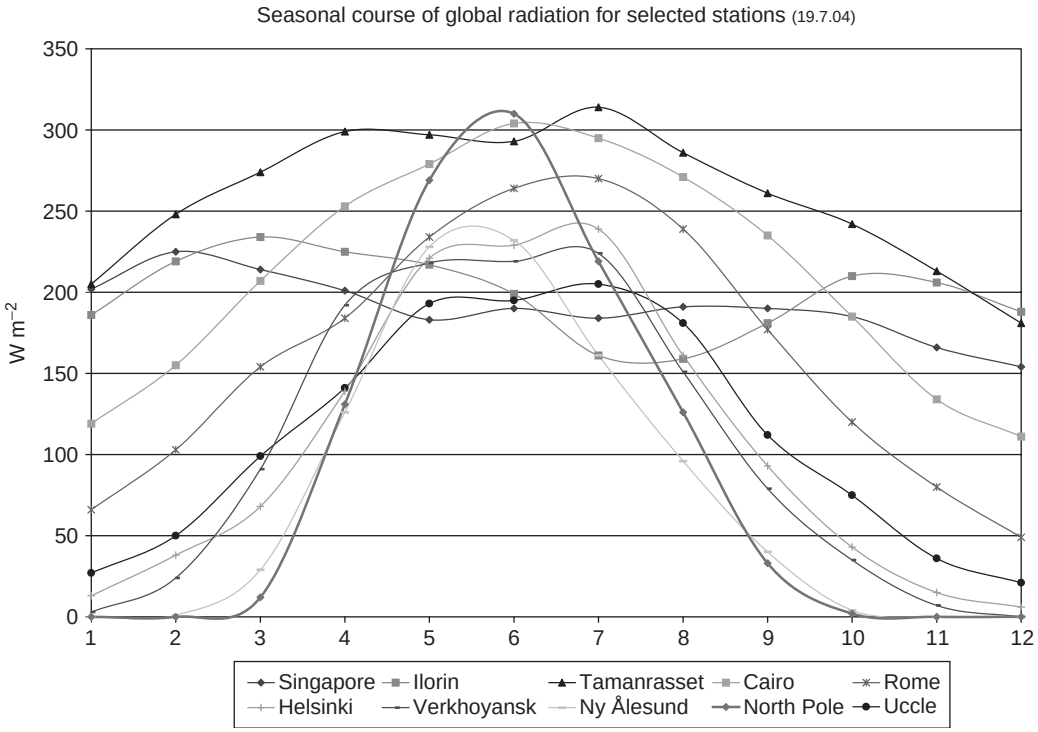


Figure 4 Monthly mean global radiation for selected stations. The figure represents the meridional variation of the seasonal global radiation based on the long-term measurements at 10 stations. These stations are spaced approximately by 10° in latitude, starting at the equator for Singapore and ending at north pole. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the months of extreme solar declinations of June (Northern Hemisphere) and December (Southern Hemisphere) the latitudinal distribution of global radiation forms a maximum near the pole and a secondary maximum at around 30° of latitude in the summer hemisphere. During these months the summer hemisphere receives 70% of the total global solar radiation. In December, the world's largest monthly global radiation of $450\,W\,m^{-2}$ is observed on the plateau of East Antarctica and the lowest in the regions pole-ward of 70° N, where the monthly global radiation is zero due to the polar night. In June, the largest monthly global radiation is found in the central region of Greenland ($380\,W\,m^{-2}$), while in the region of the North Pole it remains smaller at around $300\,W\,m^{-2}$ because of larger optical air mass and cloud amount, and lower albedo of the melting sea ice. Across the equator in June, the global radiation declines into the Southern Hemisphere at a rate of $3\,W\,m^{-2}$/degree of latitude, reaching zero at 70° S, which closely touches the Antarctic coast of the Eastern Hemisphere.

REFLECTION OF SOLAR RADIATION AT THE EARTH'S SURFACE

Reflection is influenced to a great extent by albedo, a ratio of reflected to global radiation. The earth's surface albedo is of fundamental importance for energy balance, as this quantity greatly influences net radiation. The albedo values depend on the material of the surface and its condition on one hand, and on the incident direction and the spectral composition of the incoming radiation on the other. Under natural conditions, the albedo of the earth's surface is strongly influenced by vegetation, wetness, solar elevation, and the amount of clouds. Depending on the nature of the surface, the existence of the snow cover has a variable effect on the albedo. In Table 2, representative albedo values for the entire solar spectral range are presented for various surfaces.

The global distribution of albedo is strongly influenced by the location of continents and oceans. The lowest albedo is seen over water bodies such as oceans and lakes. The albedo of the ocean increases with solar zenith angle. While the minimum monthly albedo of 0.06 is expected over the ocean at the equator in months with equinoxes, the ocean albedo increases beyond 0.1 between 47° and 50° of latitude in the same months. The albedo of the water surface just off the sea–ice margin at the beginning of the melt season is also about 0.1. The monthly mean albedo of the ocean water surface is summarized in Table 3. The albedo in high latitudes is further enhanced by sea ice. The seasonal variation of the sea ice albedo is summarized in Figure 5. On

Table 2 Albedo for global radiation (total albedo) (monthly values, not instantaneous values; albedo-ranges given in brackets in the second lines are those with a snow cover)

Surfaces	Albedo values	
	Observed range	Most frequent value
Glacier, accumulation area	0.57–0.85	0.75
Glacier, equilibrium line area (during melt)	0.37–0.77	0.65
Glacier, ablation area (after snow melt)	0.11–0.70	0.35
Seasonal snow cover, clean (before melt)	0.65–0.85	0.7
Seasonal snow cover, clean (during melt)	0.3–0.6	0.4
Dirty snow cover	0.2–0.4	0.3
Tundra (in snow-free period)	0.07–0.22	0.15
Natural boreal forest, close crown (Canada)	0.04–0.17	0.1
Natural boreal forest, open woodland (Canada)	0.07–0.19	0.12
Natural boreal forest, close crown (Siberia)	0.12–0.21	0.16
Burned boreal forest, close crown (Siberia)	0.10–0.18	0.14
Managed forest (Pinus) (Germany)	0.09–0.12	0.1
	(0.15–0.25)	
Managed forest (Abies) (Switzerland)	0.05–0.06	0.06
	(0.21–0.30)	
Bare rock	0.09–0.28	0.21
Bog and Muskeg	0.05–0.17	0.11
Deciduous forest, before fall	0.10–0.20	0.15
Deciduous forest, after fall	0.08–0.15	0.11
Step	0.09–0.25	0.17
Dry step	0.22–0.26	0.25
Desert	0.24–0.28	0.25
Tall grass	0.22–0.33	0.27
Tropical forest	0.11–0.15	0.12
Agricultural field, bare soil	0.08–0.14	0.1
Agricultural field, with plants	0.14–0.27	0.17
	(0.24–0.73)	
Short grass (meteorol. stations)	0.21–0.25	0.22
Cities	0.11–0.18	0.15
	(0.18–0.37)	
Lakes	0.02–0.16	0.08
Ocean (open water)	0.03–0.45	0.08
(Albedo of water depends heavily on the incidence angle of radiation.)		
Sea ice before melt (excl. leads)	0.75–0.88	0.83
Sea ice during melt (excl. leads and ponds)	0.25–0.82	0.50–0.65
Melting white ice	0.55–0.72	0.65
Melting blue ice	0.25–0.32	0.28
Ponds on sea ice	0.10–0.30	0.2

Sources: BSRN, Cutler and Munro (1996), Davies (1965), GEBA, Giambelluca *et al.* (1999), Grenfell *et al.* (1981, 1994), A. Kessler and L. Jaeger (personal communication), Kobayashi *et al.* (2001), Kondrat'ev (1973), Kondratyev (1969), Lee (1978), Liljequist (1956), Oguntoyinbo (1970), Ohmura (1982, 2001), Oke (1978), Payne (1972), Perovich *et al.* (2002), Pinker (1982), Ross *et al.* (1998), Roth (1985), Sellers (1965), Shuttleworth (1989), Shuttleworth *et al.* (1984), Sie (1994), Takeuchi (2002), Takeuchi *et al.* (2001), Tsay *et al.* (1998), Z'graggen (2001).

the continents, the albedo also increases toward the poles because of the snow cover and glaciers. The annual mean albedo of the earth's surface ranges from the lowest value of 0.065 at the equator over the ocean to the highest value of 0.82 to 0.84 in the central plateaus of Greenland and Antarctica. Consequently, the meridional distribution of the surface albedo increases very steeply in higher latitudes and forms an important basis for the formation of the present climate. The annual mean albedo for the global oceans including the sea ice is estimated at 0.082. The mean albedo of the land surfaces including seasonal snow covers is estimated at 0.31 and is considerably higher than that of the oceans. The global mean earth's surface albedo is most likely to be 0.15.

The distribution of the reflected solar radiation is the product of the global radiation and the albedo, which was already discussed in detail. In terms of annual means, the regions with more than $30\,W\,m^{-2}$ of reflected solar radiation are witnessed only in the polar and arid regions. In the former, the coverage by snow and ice and the lower solar elevation angle are responsible for larger reflected

Table 3 Albedo of water for clear sky by Grishenko (Cogley, 1979), Payne (1972) and Ter-Markaryantz (Minorova, 1973) (The albedo values are entered for each month in the order of the authors. Empty places indicate no information)

Latitude	January	February	March	April	May	June	July	August	September	October	November	December
90°	–	–	30.1	29.3	17.1	14.8	16.0	24.6	34.2	–	–	–
80°	–	30.1	31.9, 33	22.5, 14	16.0, 10	13.1, 9	14.5, 8	20.6, 8	29.4, 12	30.5	–	–
70°	30.1	33.8, 41, 27.1	22.9, 15, 20.3	14.8, 10, 12.5	11.6, 8, 10.0	11.2, 7, 8.9	11.4, 7, 9.2	13.4, 9, 10.5	20.2, 11, 15.8	31.3, 25, 25,6	30.1, 24.3	–
60°	33.9, 28, 27.0	24.0, 12, 20.3	15.5, 9, 13.0	10.5, 7, 8.9	8.8, 7, 7.4	8.4, 7, 7.4	8.6, 6, 7.7	9.8, 7, 8.0	13.6, 7, 10.2	21.6, 10, 16.8	32.1, 16, 25.6	35.5, 44, 27.4
50°	22.0, 11, 18.0	16.1, 10, 14.2	10.8, 8, 9.6	8.4, 7, 7.7	7.5, 6, 7.2	7.3, 6, 6.7	7.4, 6, 6.6	8.0, 7, 6.2	9.9, 7, 8.1	14.4, 8, 11.0	21.0, 11, 16.0	24.1, 12, 22.1
40°	14.5, 10, 12.6	11.1, 9, 9.1	8.5, 7, 7.5	7.3, 7, 6.3	6.8, 6, 6.1	6.7, 6, 5.7	6.8, 6, 5.7	7.1, 6, 6.4	8.0, 7, 8.1	10.3, 8, 10.7	13.8, 10, 11.4	16.1, 11, 13.3
30°	10.3, 9, 9.3	8.6, 7, 7.7	7.3, 6, 6.7	6.7, 6, 5.7	6.5, 6, 5.8	6.4, 6, 5.6	6.4, 6, 5.6	6.6, 6, 5.6	7.1, 6, 6.3	8.2, 7, 5.8	10.0, 8, 8.2	11.1, 9, 9.7
20°	8.3, 7	7.4, 6	6.7, 6	6.4, 6	6.3, 6	6.3, 6	6.3, 6	6.4, 6	6.6, 6	7.2, 6	8.1, 7	8.7, 7
10°	7.2, 7	6.7, 6	6.4, 6	6.3, 6	6.4, 6	6.4, 6	6.4, 6	6.3, 6	6.3, 6	6.6, 6	7.1, 6	7.4, 7
0°	6.6, 6	6.4, 6	6.3, 6	6.4, 6	6.6, 6	6.8, 6	6.7, 6	6.4, 6	6.3, 6	6.4, 6	6.6, 6	6.8, 6

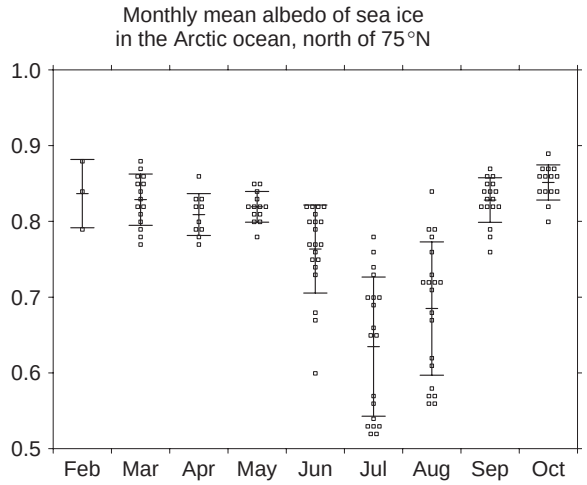


Figure 5 Monthly mean albedo of sea ice in the Arctic ocean, observed at north pole stations

radiation. In the arid regions that have a generally high albedo of the dry surfaces, the larger global radiation due to the lack of clouds and the location in lower latitudes are responsible for larger reflection. In this area, the greater portion of the reflected radiation is lost directly into space, causing the negative net radiation of the atmosphere/earth's surface in the lower latitudes. Globally averaged, the annual mean reflected radiation falls between 24 $W m^{-2}$ (ISCCP D2) and 30 $W m^{-2}$ (Kiehl and Trenberth, 1997). The most likely estimate is 25 $W m^{-2}$ (revision after Ohmura and Gilgen, 1993). The uncertainty is due to the different estimates of the albedo for water, which covers 67% of the earth's surface.

LONGWAVE ATMOSPHERIC RADIATION

Minor atmospheric components, such as water vapor, carbon dioxide, methane, dinitrogen oxide (nitrous oxide), ozone, and a number of chlorofluorocarbons and their substitutes have strong absorption bands in the wavelength range between 4 and 50 μm where terrestrial emission takes place. Because the Kirchhoff law applies to the troposphere and stratosphere, these molecules emit in the same wavelengths. Of these constituents, water vapor plays the most important role, followed by carbon dioxide. The total emission by these gases and clouds, which we call longwave incoming radiation, is as large as TOA solar radiation. The longwave incoming radiation is the source of the atmospheric greenhouse effect. The longwave radiation received at the earth's surface is the result of the emission and absorption by the entire atmosphere and clouds. Presently, there is a set of commercially produced pyrgeometers that can provide the accuracy of $\pm 3\,W m^{-2}$ (Philipona and Ohmura, 2001). The computation of the longwave incoming radiation involves essentially the numerical integration of the Schwarzschild equation and the atmospheric profile from the stratopause to the earth's surface. In practice, however, in the absence of clouds it is the bottom 1 km of the atmosphere that contributes more than 90% of the longwave incoming radiation at the surface.

The longwave incoming radiation is numerically the second largest component in the earth's surface energy balance, next to the longwave outgoing radiation. The absolute magnitude of this flux was until recently grossly underestimated. The underestimation was caused by an insufficient knowledge of the water vapor continuum in the computation. The underestimation was also induced by the convective heat loss in pyrradiometers that were often used before pyrgeometers were developed. The progress in infrared spectrometry, numerical computations, and the development of the instruments during the last 20 years contributed to achieve more realistic values. With an annual global mean of 345 $W m^{-2}$, it is considerably larger than the absorbed solar radiation of 144 $W m^{-2}$. The regional difference of this flux is also much larger than that of the global radiation as is presented in Figure 6. In the Northern Hemisphere at sea level, it ranges from 225 $W m^{-2}$ at the North Pole to 420 $W m^{-2}$ in cloudy regions under the equator. The Hemisphere's observed minimum is recorded in the central region of the Greenland ice sheet (Summit: 167 $W m^{-2}$). The distribution of the longwave incoming radiation is also strongly influenced by the distribution of the continents and oceans. While drier continents tend to shift isopleths southwards, relatively large fluxes are seen over the eastern halves of the oceans. In the Southern Hemisphere, the distribution is less affected by the distribution of continents and oceans. The isopleths run more or less parallel to latitudes. The longwave incoming radiation decreases from about 420 $W m^{-2}$ at 10 °S to 230 to 250 $W m^{-2}$ in the coastal region of Antarctica at a rate of 3 $W m^{-2}$/one degree in latitude. It decreases further in the interior of Antarctica. The central plateau region of Antarctica receives the smallest flux. At about 100 $W m^{-2}$, there is no other region on the earth where the longwave incoming radiation underscores this value.

The longwave incoming radiation decreases in general with altitude, as a result of the reduction of air temperature, water vapor, and other greenhouse gas concentrations. The vertical gradient for the midlatitude mountainous regions is $-3\,W m^{-2}$/100 m (Marty *et al.*, 2002).

In terms of monthly means, the maximum measured longwave incoming radiation is 420 $W m^{-2}$ over equatorial oceans. The minimums are located in the interiors of Greenland and Antarctica where they reach 130 $W m^{-2}$ and 75 $W m^{-2}$ respectively in midwinter (Kuhn *et al.*, 1977; T. Yamanouchi, personal communication). It is indeed in these high-latitude regions where a large annual amplitude of the longwave incoming radiation is also registered. The July

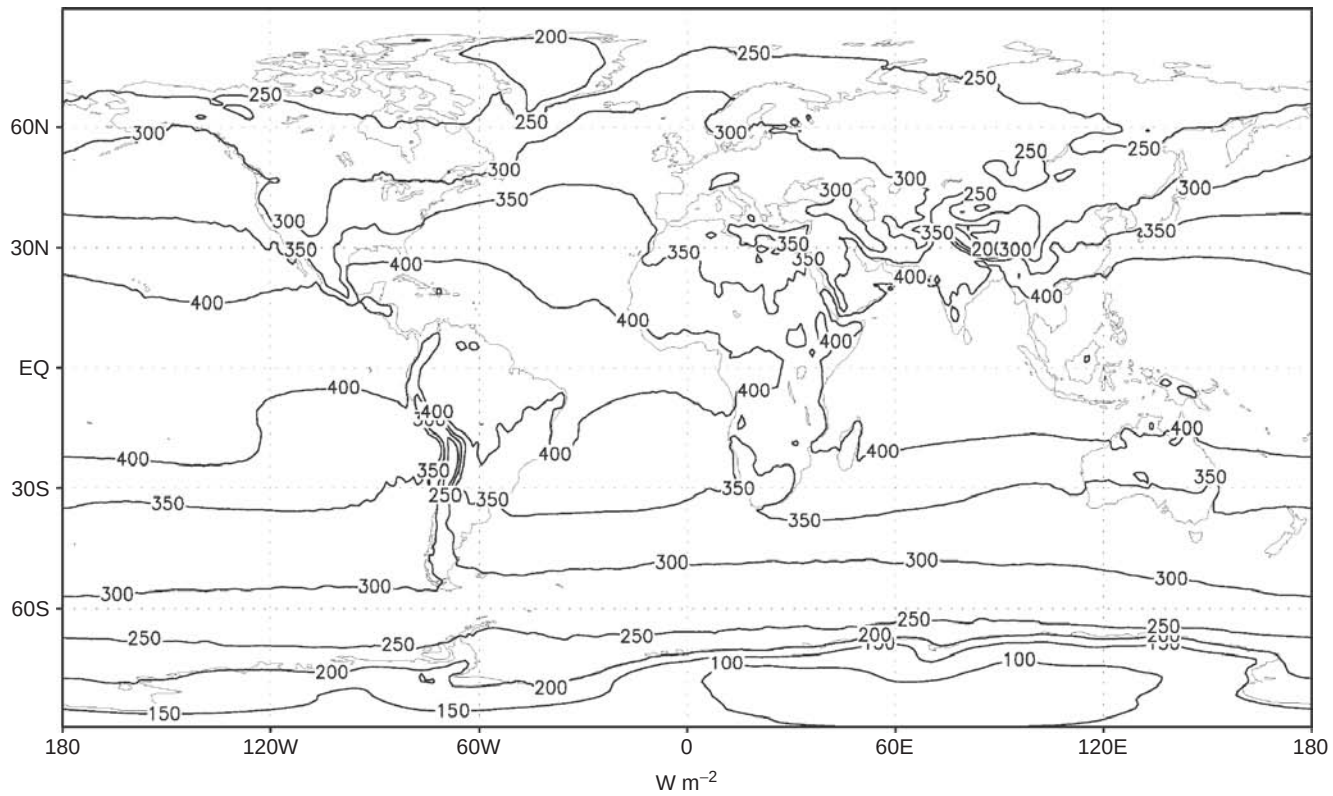


Figure 6 Annual longwave incoming radiation computed by ECHAM4T106

and January mean fluxes reach 220 W m^{-2} and 190 W m^{-2} respectively (Hoch *et al.*, 2004; Putnins, 1970; Rusin, 1964) in the interior of Greenland and on the Antarctic plateau in summer. Although located in higher latitudes than Greenland, the Arctic Ocean receives a considerably larger longwave flux owing to the higher air temperature and absolute humidity (145 W m^{-2} in January and 300 W m^{-2} in July). The unpublished data in the present section are from Global Energy Balance Archive (GEBA) and Baseline Surface Radiation Network (BSRN).

LONGWAVE OUTGOING RADIATION

The longwave outgoing radiation is the ultimate reaction of the earth's surface under a given energy balance condition. This term is passive and determined by all other fluxes. The best instrument to measure this component is longwave radiometers specified as pyrgeometers directed to the surface. The measurement of the brightness temperature by a radiation thermometer together with the Planck equation with the Kirchhoff law also gives a good estimate of the longwave outgoing radiation. The uncertainty of the emissivity does not play a significant role in this matter, as the longwave outgoing flux contains the reflection of the longwave incoming radiation with the reflectivity which is complementary to surface emissivity.

Since measurements of the surface temperature by radiation thermometers are not widely available, a possible use of a standard meteorological element, such as air temperature should be investigated. The effective temperature of the surface as mentioned above can differ a great deal from the standard meteorological air temperature observed at the standard screen-level. If short-term temperatures such as one minute means are compared, a discrepancy of up to 10 °C can occur for radiatively strongly cooled or heated dry surfaces. Differences of up to 15 °C are measured under strongly positive or negative net radiation and calm conditions. Theoretically, one can expect a difference up to 20 °C (Ohmura, 1981). If sufficiently long-term observations are compared, such as monthly means, the difference between the surface effective temperature and the standard meteorological air temperature decreases to within 1 °C (Z'graggen, 2001). This is a useful feature in calculating the longwave outgoing radiation for climatological purposes, as the surface effective temperature is usually not measured. This feature also helps analyze the geographical distribution of the longwave outgoing radiation in terms of climatological air temperature, which is widely available. The best estimate of the longwave outgoing radiation is 385 W m^{-2} which corresponds to a mean surface temperature of 14 °C (Jones *et al.*, 1999).

NET RADIATION

The global distribution of the annual mean net radiation is presented in Figure 7. Net radiation at the earth's surface is generally larger over the ocean due to the lower albedo. The maximum annual net radiation of 180 W m^{-2} is seen at the equator over the ocean. The annual net radiation on oceans is on an average 30 to 50 W m^{-2} larger than on the continents of the same latitude. The annual net radiation declines from the equator to the pole at an average rate of 2 W m^{-2}/degree in latitude, rarely falling into negative values on the earth's surface. The earth's surface is basically an energy source. There are three regions of limited surface areas where negative net radiation is observed on an annual basis. They are the central region of the Greenland ice sheet above 2500 m a.s.l., the relatively cloud-free regions of the Antarctic ice sheet, and the upper accumulation area of mountain glaciers.

Various independent sources indicate that the global mean annual net radiation at the earth's surface lies between 100 and 105 W m^{-2} (Budyko *et al.*, 1988; Ohmura and Gilgen, 1993; Kiehl and Trenberth, 1997).

Seasonal fluctuations in net radiation are large. The summer (JJA or DJF) hemisphere is characterized entirely by positive net radiation including the high-altitude zones of the major ice sheets and the accumulation areas of mountain glaciers. There is a weak equator-to-pole gradient of 1.5 W m^{-2}/degree in latitude. The largest net radiation is seen over the subtropical oceans. In this region, the mean net radiation during the three summer months reaches 220 W m^{-2}. The energy exchange with the atmosphere is, however, relatively inactive, as almost half of the net radiation is kept in the ocean for the release during winter, not only in the same area but also in the higher latitudes. During the winter (DJF or JJA), 25% of the hemispheric surface, mostly the regions pole-ward of 50° N, S lie under negative net radiation, with a mean equator to pole gradient of 4 W m^{-2}/ degree in latitude. In most regions of the negative net radiation, the magnitude remains rather small, except for the following regions. The ocean areas in high latitudes tend to be strongly negative (-30 W m^{-2} or less) owing to the heat release by the ocean. In some areas on the land, net radiation falls into the strongly negative range. They are the high-latitude/high-altitude regions with a small cloud amount, such as the Brooks Range in Alaska, the McKenzie Mountains, the northern Rocky Mountains and the parts of the Canadian Arctic Islands with less cloud amounts, Greenland, Scandinavia, and a part of East Antarctica. In the lower altitudes, the regions such as the basins of the Lena and Yenisey with relatively small cloud amounts in winter show strongly negative net radiation. In these areas,

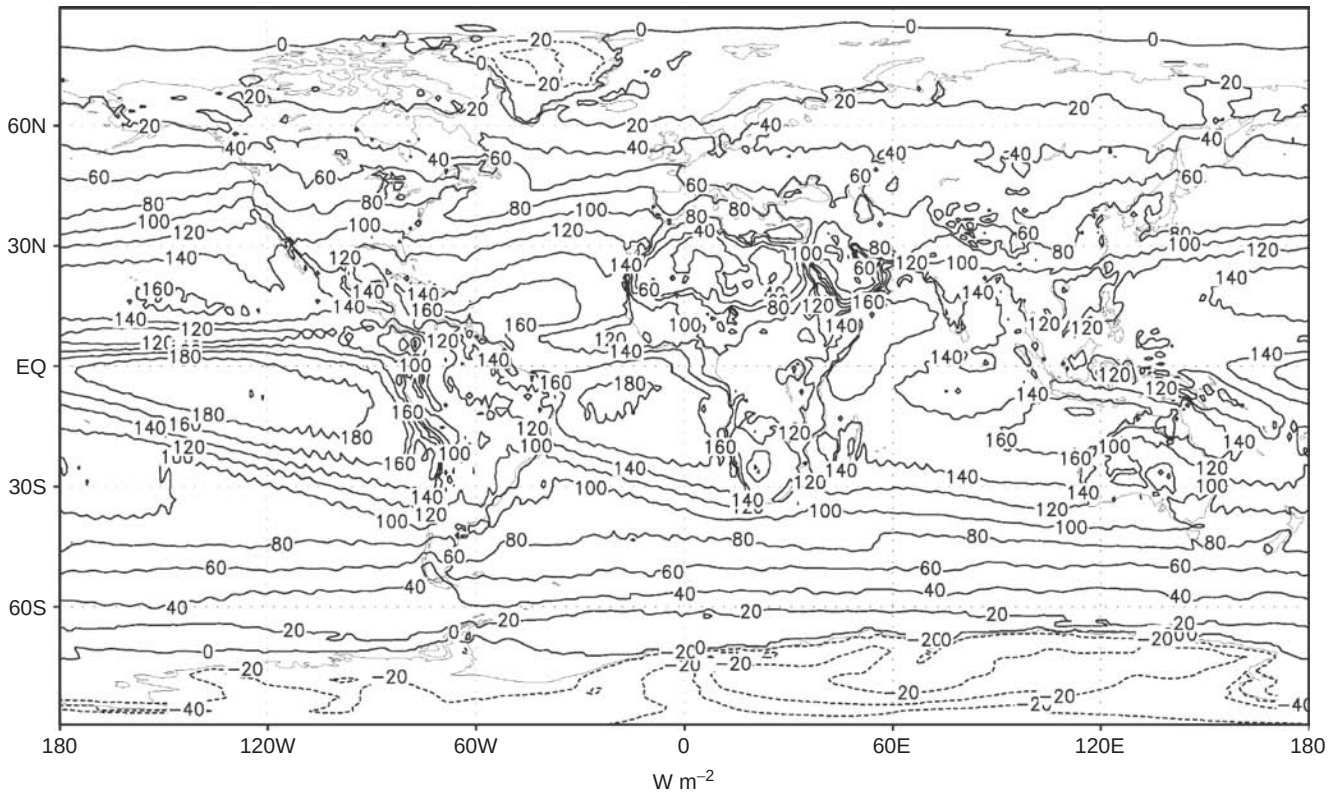


Figure 7 Annual net radiation computed by ECHAM4T106

sensible heat from the atmosphere is the main heat source during winter.

Net radiation strongly influences evapotranspiration. This feature is used to determine evapotranspiration with rather limited instrumentation. Well-known examples are the Bowen ratio/energy balance method (Bowen, 1926) and the Penman equation (Penman, 1948). Naturally, both methods are subsequently refined to improve the accuracy (Doorenbos and Pruitt, 1977; Ohmura, 1982). The strength of these methods is the fact that they do not rely heavily on aerodynamic methods, since the major error for turbulent heat fluxes is induced by wind measurements. The Bowen ratio/energy balance and the Penman methods rely on net radiation, which is accurately measurable with an inexpensive instrument such as a net radiometer. These methods are especially suited for long-term observations with minimum maintenance requirements. The nature of these fluxes will be detailed in the next section.

TURBULENT HEAT FLUXES

Figure 8 shows the annual mean sensible (a) and latent (b) heat fluxes. For the global mean, latent heat flux of vaporization with $85\,W\,m^{-2}$ accounts for majority of net radiation at the surface. Sensible heat flux accounts for ($19\,W\,m^{-2}$) only 18% of surface net radiation. These fluxes correspond to 25% and 6% of the primary solar radiation at the TOA, respectively. It is noteworthy that these two turbulent heat fluxes that are carried by similar eddies and propelled by the same energy source, namely the surface net radiation, differ by as much as factor 4. The main reason is found not on the surface but in the middle troposphere where water vapor condenses to form clouds, generating heat. This wet convection is responsible for creating two fundamentally different vertical gradients in the lower troposphere, the sharp decrease in water vapor concentration, and the increase in sensible heat with altitude. This situation results in different geographical distributions of latent and sensible heat fluxes. While latent heat flux is directed upward almost everywhere on the earth's surface, sensible heat flux is often directed downward. Henceforth, in a large area in high latitudes, pole-ward of 70° N and S sensible heat flux is directed to the surface. Sensible heat flux can easily flip downward even in an area with positive net radiation. In lower latitudes when an arid continent is located windward, sensible heat flux turns downward as is witnessed over the Arabian Sea. The lower SST on the equator suffices to reduce the magnitude of evaporation but causes in many regions a downward flow of sensible heat.

The ocean evaporates 88% of global atmospheric water. Further, the surface of the ocean within 25° N and S of the equator accounts for 50% of global evaporation. The land surface contributes only 12%. Per unit area the latent heat flux on the ocean ($110\,W\,m^{-2}$) is far larger than over the land ($36\,W\,m^{-2}$). This difference is due to the limit in surface water availability on the land. Hence the land surface generates sensible heat more easily than on the ocean: $26\,W\,m^{-2}$ on land versus $15\,W\,m^{-2}$ on ocean. Taking the respective surface areas into account, 40% of the global sensible heat is transferred into the atmosphere from the land, and the remaining 60% from the ocean.

Seasonally analyzed, however, the largest sensible heat flux is seen on the ocean in winter and on the desert in summer. In winter, in the monsoon regions where strong cold air flows from the continents on to the surface of the warm western boundary currents, some of the largest sensible heat fluxes can occur. In the regions off New England and Japan, sensible heat fluxes in January can reach more than $100\,W\,m^{-2}$. The outbursts of the winter monsoon are also dry and the same regions also see the largest latent heat flux of evaporation in the world ($250\,W\,m^{-2}$). Both these large fluxes are supported by a strong convection in the ocean mixing layer that delivers in places as much as $400\,W\,m^{-2}$-ocean surface heat flux. Oceans are in general very active evaporating regions in winter. Although often overlooked, both hemispheres evaporate substantially more during winter than in summer (Ohmura and Wild, 2002). This situation shows that temperature alone does not determine large-scale evaporation. The strength of the atmospheric circulation is equally influential in determining evaporation.

As winter is a season with active evaporation on oceans, summer is the season for actively generating sensible heat on land. The year's peak of sensible heat flux on land is usually observed in early summer (May and June in NH, October and November in SH). During these months, the sensible heat flux in arid regions of both hemispheres reaches 50 to $100\,W\,m^{-2}$, and becomes a heat sink that is similar but more powerful than longwave net radiation.

The computed turbulent fluxes are very difficult to validate with observations, as there are not many high-quality flux observations carried out long enough to be climatological data. It is possible, however, to check the quality of the global estimate of latent heat or evaporation. Various numerical experiments showed a close relationship between the annual global latent heat flux and the annual global net radiation as is presented in Figure 9. Since the surface net radiation is much more accurately known than the latent heat flux, the figure allows a first-hand estimation of the global mean latent heat flux. The latent heat flux corresponding to the most likely value of the surface net radiation (100 to $105\,W\,m^{-2}$) is $85\,W\,m^{-2}$. There is a second quality check. Since annual mean global evaporation must be very close to the annual mean global precipitation, these two components can be compared. The water equivalent of $85\,W\,m^{-2}$ is 1100 mm, which fits between the estimations of annual

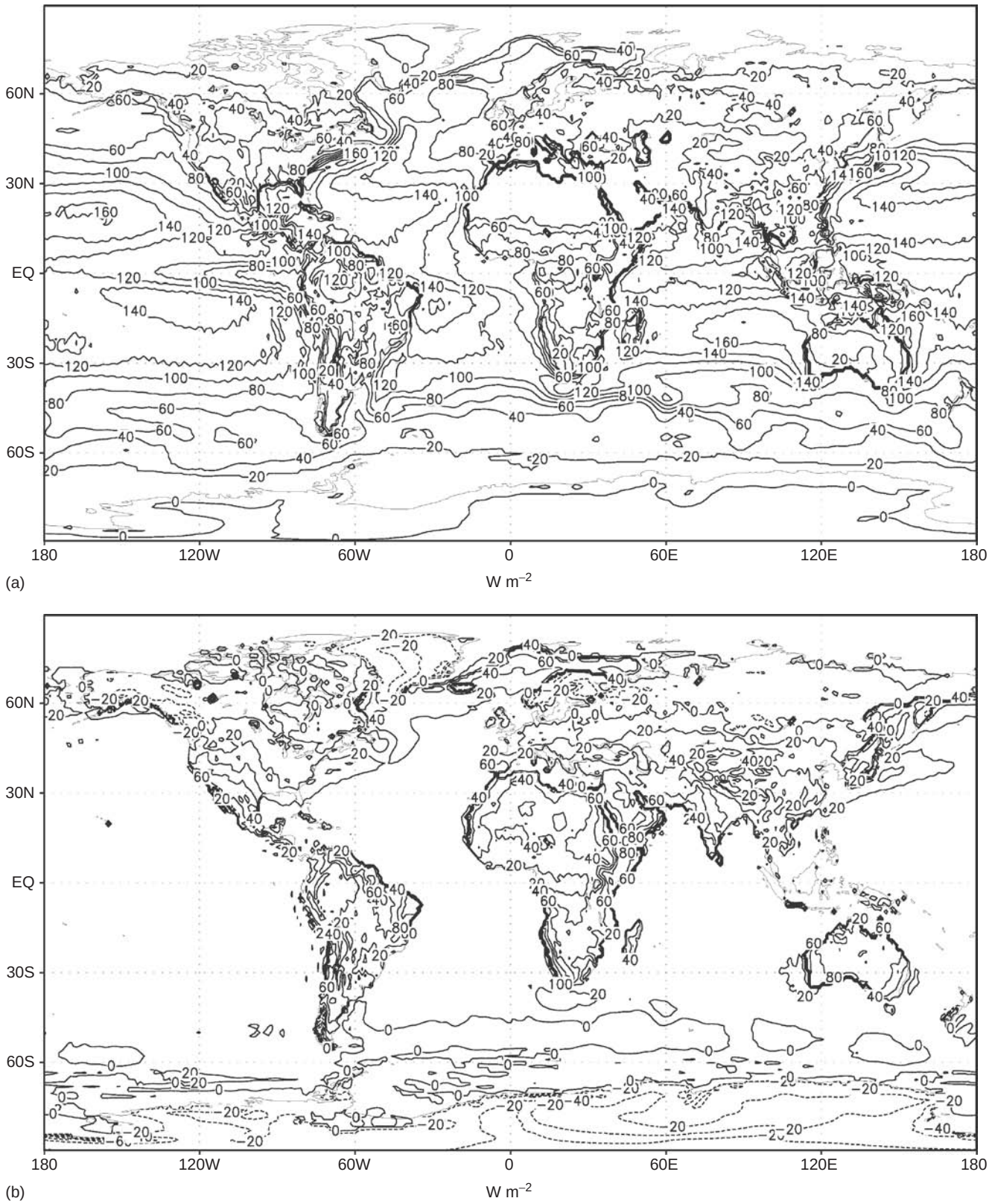


Figure 8 Annual latent heat flux (a) and sensible heat (enthalpy) flux (b)

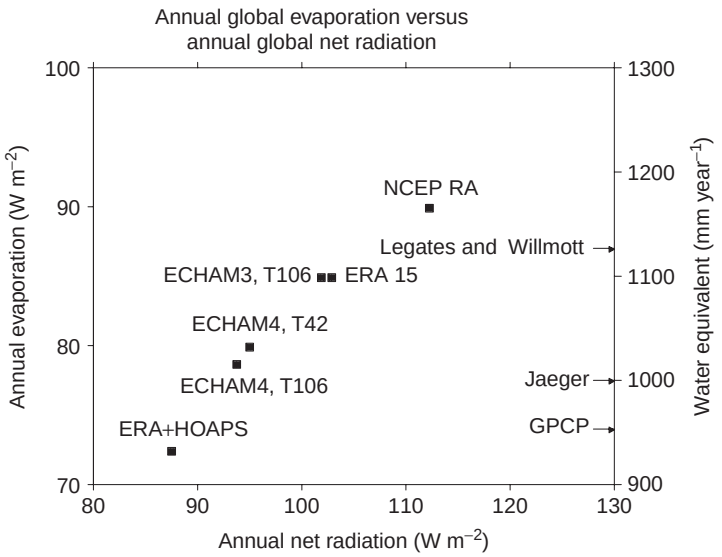


Figure 9 Relationship between the annual global mean latent heat flux and surface net radiation in various models

global precipitation by Jaeger (1976) and Legates and Wilmott (1990). This agreement indicates that the present estimation is realistic.

SUBSURFACE HEAT FLUX

Since the subsurface heat flux has fundamentally different mechanisms on land and on a water body, it will be discussed separately for these surfaces. Ground heat flux is normally one order of magnitude smaller than other fluxes and it plays a limited role in surface energy balance. The main reason for this is the relatively small thermal diffusivity of the material composing the ground in comparison with the eddy diffusivity of the near-surface atmosphere. The thermal diffusivity of common soils and rocks is 10^{-4} to 10^{-6}, when usually observed values for the eddy diffusivity in the first 10 m of the atmosphere are taken as unity. Nevertheless, this heat flux determines the thermal conditions of the near-surface layers of the pedo- and lithosphere. Ground heat flux turns downward in spring and reaches the year's maximum heat flow in early summer, usually one month ahead of the summer solstice. The flux declines throughout the summer and turns upward in autumn until the following spring. At the year's maximum, it attains about 10% of the net radiation. The phase and the magnitude of the ground heat flux can be seriously affected by the snow cover, the thickness, the timing of its onset, and the melt. The annual mean ground heat flux observed for more than several years is very close to zero or becomes the same as the geothermal heat flux.

The subsurface heat on the ocean surface is the ocean surface heat flux or net downward heat flux in oceanography (Figure 10). This component is substantial in diurnal and annual surface energy balance as well as in climate changes. The importance of the ocean in the climate system is due to its large heat storing function. This function is not only due to the large heat capacity of water, but to a great extent to the transparency of ocean water for solar radiation, and to the oceanic turbulence and convection which promote diffusion. The annual amplitude of the ocean surface heat flux reaches up to $200\,\mathrm{W\,m^{-2}}$ on the midlatitude ocean. In the course of a year, the ocean flips between a heat source and a sink on the hemispheric scale. Globally observed are three belts of heat sinks for the surface energy balance (sources for the ocean), one on the equator and the two others in the higher midlatitudes around 40° to 50° N and S where a substantial amount of heat flows downward. Of these three, the region under the equator is a permanent energy sink. In the equatorial ocean, large shortwave incoming radiation, supported by relatively weak longwave net radiation (due to cloud amount) and smaller turbulent heat losses (due to higher humidity and lower SST), form the basis of the permanent energy sink for the surface (source for the ocean). The other two belts in the higher midlatitudes are sinks on an annual basis, but are in a delicate balance between the sink in summer and the source in winter. Unlike ground heat flux, the annual mean ocean surface heat flux assumes large values. The eastern half of the higher midlatitude ocean surfaces where the temperature is lower is especially important as a sink region. Further in higher latitudes, in the Arctic Ocean (Central Polar Ocean including marginal oceans, such as Norwegian, Barents, Kara, Leptev, East Siberian, and Beaufort Sea) and the Antarctic Seas near the continent, ocean surface flux is an important heat source keeping the polar and subpolar regions relatively mild.

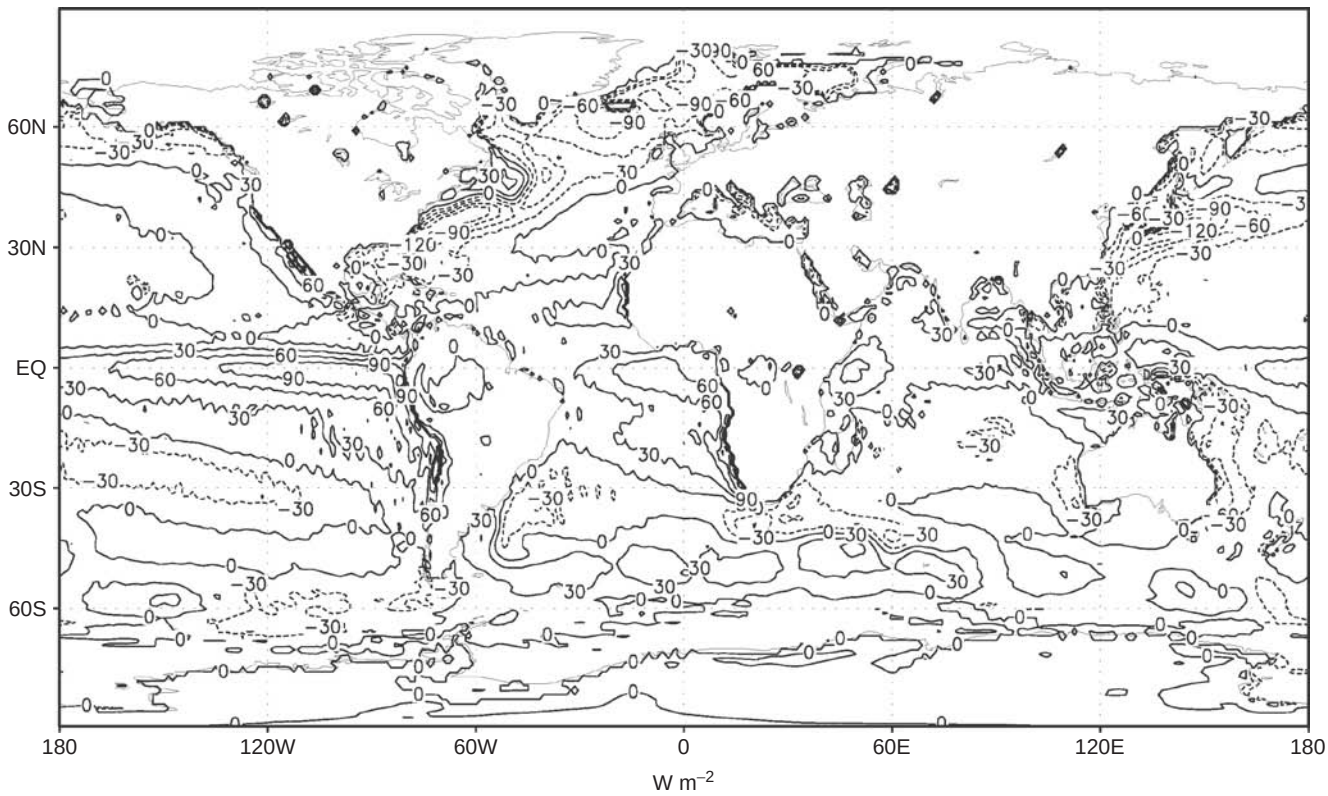


Figure 10 Annual subsurface heat flux computed by ECHAM4T106. Signs are taken positive when directed downward

LATENT HEAT OF MELT

Latent heat of melt (fusion) globally averaged is the order of 1% of the extraterrestrial solar radiation. Because this value gives an impression of numerical insignificance, the latent heat of fusion is often not considered in energy balance climatology (Budyko, 1956). Regionally viewed, however, the latent heat of melt plays an extremely important role. This component is the main heat sink and averages at around 100 W m^{-2} during the melt season on glaciers, sea ice, and snow cover. The reason for the low air temperature of the Arctic in summer, despite one of the largest global radiation of the hemisphere, is due not only to the high albedo but also due to the heat consumption by the latent heat of melt. Latent heat of melt is determined in the field either by measuring the lowering of the surface or the discharge of the melt water. In theory, one can explain the amount of the melt by evaluating all terms of the equation in the equation of energy balance. In practice, the melt is so closely related to the air temperature that the air temperature alone can become a powerful index for estimating the rate of the melt. The relative importance of the energy source terms for the melt varies considerably depending on the site characteristics, such as altitude, albedo, air temperature, and humidity. One common feature of all energy balance measurements made over the melting snow and ice is the fact that the longwave incoming radiation is by far the most important energy source for the melt. It was pointed out earlier that the majority of the longwave radiation originates from the atmospheric layers very close to the surface. This is the basic reason why air temperature alone is such a good indicator of the melt rate.

MEAN STATE OF THE GLOBAL ENERGY BALANCE

The above discussion is summarized in Figure 11 to represent the present mean state of the energy exchange in the climate system. The global radiation at the surface, 169 W m^{-2}, accounts for 49% of the primary energy source from the sun. This value is smaller than previously used, 186 W m^{-2} by Sellers (1965), 185 W m^{-2} by Salby (1996), and 198 W m^{-2} by Kiehl and Trenberth (1997). The discrepancy of 15 to 30 W m^{-2} for global radiation is one of the most important findings of recent years. The present value can be compared with 168 W m^{-2} proposed earlier by Budyko *et al.* (1988). The new value presented here is the result of the improved network of direct observations of global radiation at the earth's surface. The earth's surface albedo was also newly evaluated and found to converge at around 15%. This value is somewhat larger than 14% by Budyko (1956), 11% by Sellers (1965), and 9% by Salby

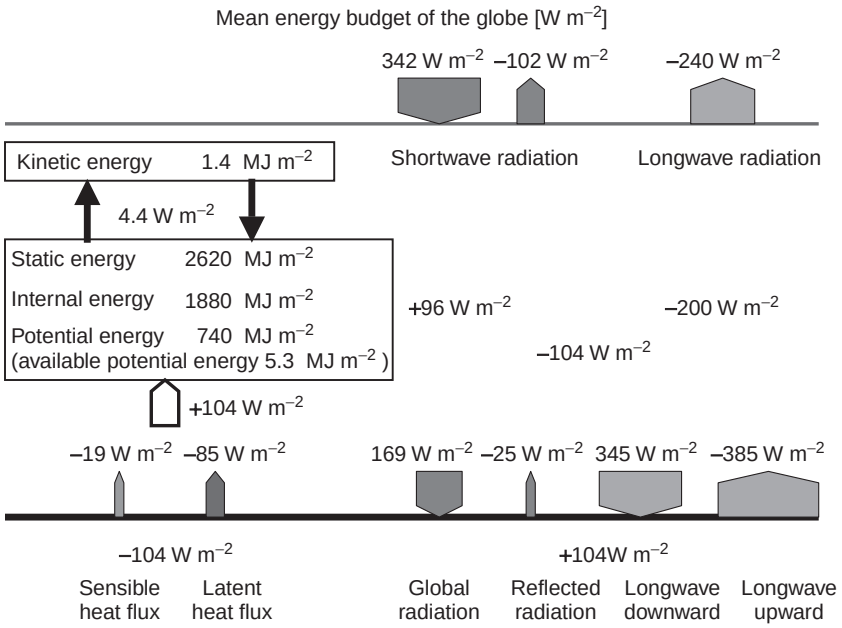


Figure 11 Mean state of the global energy balance. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

(1996). The present value of 15% coincides with those proposed by Budyko *et al.* (1988) and Kiehl and Trenberth (1997). The present value of the earth's surface albedo is mainly from the accumulated data on albedo measurements over a variety of surfaces by terrestrial and airborne instruments, the long-term mapping of the distributions of sea ice and snow cover obtained by satellites, and the detailed investigations of the albedo of water under variable solar elevation angle and cloud conditions. Keen readers probably have noticed that the terrestrial incoming radiation from the atmosphere in Figure 8 is considerably larger than in previously published sources. The recent advances in infrared radiometry have corrected previous underestimations. A proper treatment of the water vapor continuum and the effect of clouds are also responsible for the present improvement. The longwave outgoing radiation is calculated on the basis of the recently obtained earth's surface temperatures by Jones *et al.* (1999).

The difference in the net radiation at BOA and TOA is separately evaluated for solar and terrestrial radiation. The divergence of solar radiation is negative due to atmospheric absorption; it amounts to $-96\,\mathrm{W\,m^{-2}}$. This component was previously grossly underestimated. The most frequently quoted value for the atmospheric absorption of solar radiation is about $60\,\mathrm{W\,m^{-2}}$ (or 17% of the TOA solar irradiance) by Sellers (1965). The absorption due to water vapor beyond $2.8\,\mu\mathrm{m}$, the collective effect of a number of minor absorption bands, aerosol, and clouds supports the present result of larger absorption. The divergence of longwave radiation is positive and is equivalent to radiative cooling of $200\,\mathrm{W\,m^{-2}}$. The net divergence of atmospheric radiation is positive; it amounts to $104\,\mathrm{W\,m^{-2}}$. It is balanced by the convergence of vertical convective fluxes (sensible and latent).

The present state of the observation network does not allow the evaluation of the global mean turbulent heat fluxes based solely on observations. The evaluation of turbulent heat fluxes is a case for which model-based computations play an important role. The present evaluation is mainly based on the computations by ECHAM4 and ERA15. Latent heat flux is, however, better understood than the sensible heat flux, as the evaporation is measured at more sites than the sensible heat flux. The global mean evaporation can also be estimated from the global precipitation. Summing up the above discussions, the annual mean global latent heat of evaporation is very likely to be $85\,\mathrm{W\,m^{-2}}$. The sensible heat flux of $19\,\mathrm{W\,m^{-2}}$ is estimated as the difference between the net radiation and latent heat flux. The kinetic energy, and the conversion rate of the available potential energy to kinetic energy are due to Kung (1988).

The mean energy flow through the climate system can be interpreted in the following manner: The earth receives a quarter of TSI, $342\,\mathrm{W\,m^{-2}}$, of which $102\,\mathrm{W\,m^{-2}}$ leaves the planet by reflection. The remaining $240\,\mathrm{W\,m^{-2}}$ is absorbed by the atmosphere and the earth's surface. The atmosphere absorbs solar radiation of $96\,\mathrm{W\,m^{-2}}$, leaving the remaining $144\,\mathrm{W\,m^{-2}}$ to be absorbed by the surface. The terrestrial net radiation at the surface is $-40\,\mathrm{W\,m^{-2}}$. The net radiation at the earth's surface ($104\,\mathrm{W\,m^{-2}}$) is transferred to the atmosphere primarily as latent heat of vaporization ($85\,\mathrm{W\,m^{-2}}$)

and somewhat less as enthalpy flow (19 W m^{-2}). The sum of these two terms (104 W m^{-2}), augmented by the atmospheric absorption of solar radiation (96 W m^{-2}) and net longwave radiation from the surface (40 W m^{-2}), is emitted by the Earth into space at the rate of 240 W m^{-2}, exactly the same amount as solar radiation originally absorbed by the planetary system. The errors of the radiative fluxes of Figure 11 differ for the individual components. Ohmura and Raschke (2005) estimate $\pm 5-7$ W m^{-2} as error for the TOA net radiation.Ohmura *et al.* (1998) estimate the error at $\pm 5-10$ W m^{-2} for the BOA net radiation, and ± 12 W m^{-2} for the total atmospheric radiation flux divergence.

MEAN STATE OF THE GLOBAL HYDROLOGICAL CYCLE

The global mean hydrological cycle is summarized on the basis of the preceding and following discussions. The global annual precipitation is estimated at 1030 ± 90 mm year^{-1} by combining the estimations by Jaeger (1976), Legates and Wilmott (1990), Roads (2002), and model computations mentioned earlier. On an average, the precipitation over an ocean (1150 mm year^{-1}) is factor 1.5 larger than on continents (740 mm year^{-1}). Because 71% of the earth's surface is the ocean, the total precipitation on oceans and the continents are estimated at 415 and 110 km^3 year^{-1}, respectively. The average flux density of evaporation is almost exactly three times stronger over oceans (1285 mm year^{-1}) than continents (415 mm year^{-1}). Total evaporation over the ocean is estimated at 463 km^3 year^{-1}, while that from the land surface is 62 km^3 year^{-1}. This situation causes a large imbalance in the hydrological budget between the ocean surface and the land. The ocean surface generates an evaporation surplus of 48 km^3 year^{-1}, while on land surfaces there is exactly the same magnitude of the precipitation surplus that results in a discharge from the continents into oceans.

The hydrological balance of the globe is, however, presently not in equilibrium. One of the most important events in the recent hydrological cycle is the systematic loss of ice from glaciers. Presently, 15.9×10^6 km^2 or 11% of the land surface is covered by glaciers. The total mass of water held in glaciers is 30.1×10^6 km^3 water equivalent (w.e.) which corresponds to about 80% of the fresh water on the earth. With respect to ice mass, 92% is in the Antarctic, 8% in Greenland, and the remaining 0.15% is distributed among all mountain glaciers and small ice caps of the world. The geographic statistics are given in Table 4 (Ohmura, 2004). During the second half of the twentieth century, glacier mass balance was observed on more than 100 glaciers including short-term measurements. There are only about 40 glaciers with more than 30 years of mass balance observations. The fact that the mass budget of the Greenland is negative has just become known in recent years. The sign of the mass budget of the Antarctic has not yet been established. A crude estimation suggests a small negative budget, however. The estimations of mass balance for three categories of glaciers, Antarctic, Greenland, and mountain glaciers and small ice caps are presented in Table 4 for the second half of the twentieth century. According to this estimate, the glacier mass balance contributed to the sea-level rise at a rate of 0.7 mm year^{-1}. Among these three glacier types, the contribution by the mountain glaciers and small ice caps was the most important.

The global water balance discussed above is graphically summarized in Figure 12. The figure represents the hydrological balance of the second half of the twentieth century, as this is the only period with sufficient data to discuss this subject. About 10% of the ocean evaporation is contributed to the land area where there is a systematic imbalance. The hydrological imbalance on the land surface is caused by the consumption of the storage in glacier ice at a rate of 300×10^3 km^3 or 18 mm year^{-1} in favor of the sea-level rise at a rate of 0.7 mm year^{-1}. The negative mass balance on glaciers, however, was not constant. As the analyses of the seasonal mass balances on mountain glaciers and small ice caps indicate so far, the mass loss was accelerated at a rate of 12 mm year^{-2}, which was caused by

Table 4 Main features and mass balance of glaciers during the second half of the twentieth century Flux unit is in km^3 year^{-1} w.e.; values in () are in mm year^{-1} w.e

	Antarctic	Greenland	Mountain glaciers & small ice caps	Total (mean)
Surface area [10^6 km^2]	13.59	1.75	0.51	15.58
Ice volume [10^6 km^3 w.e.]	27.61[1)]	2.43[2)]	0.05	30.01
Accumulation	2500 (184[3)])	519 (297)	469 (919)	3488 (220)
Ablation, melt	0 (0)	295 (169)	611 (1198)	906 (57)
Ablation, calving	2550 (188)	315[4)] (180)	0 (0)	2865 (181)
Change in storage	−50[5)] (−4)	−91 (−52)	−142 (−279)	−283 (−18)
Sea-level contribution [mm year^{-1}]	0.06*	0.25	0.39	0.7*

Values marked with an asterisk ''*'' took into account the effect of the floating ice.
Sources: Drewry (1983), Weidick (1995), Giovinetto *et al.* (1992), Weidick *et al.* (1992) and Reeh (1994), Rignot and Thomas (2002), All others are due to Ohmura (2004).

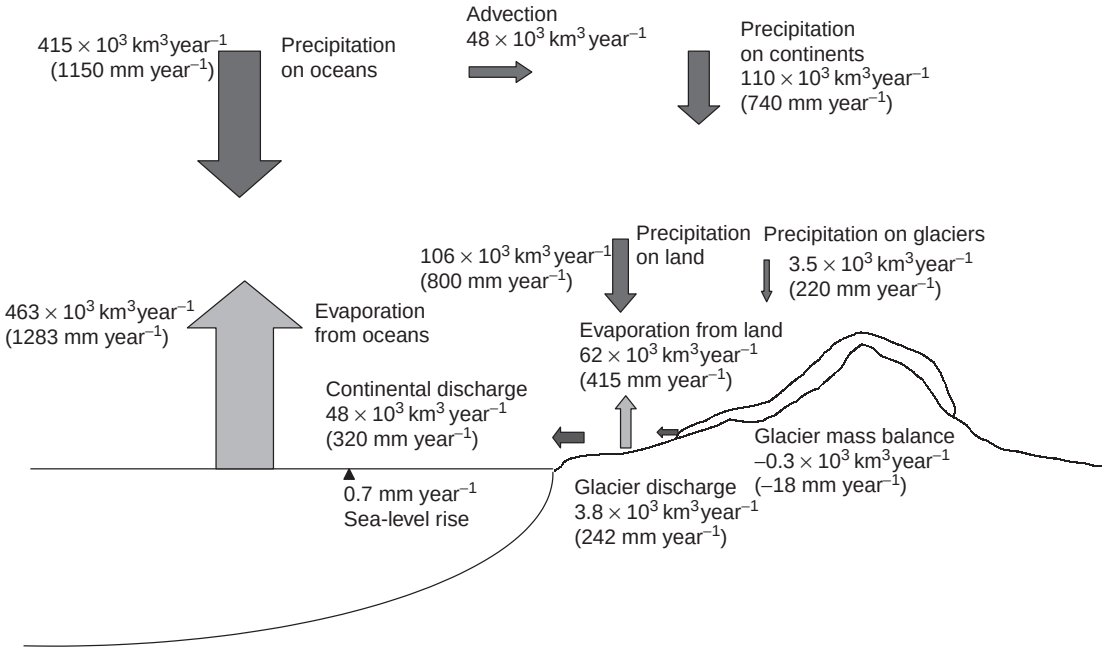


Figure 12 Mean state of the global hydrological balance. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the acceleration of the melt at a rate of 13 mm year^{-2}. The change in the accumulation during the same period was statistically insignificant and the accumulation on the glacierized regions of the world can be considered as constant (Ohmura, 2004).

REFERENCES

Barkstrom B.R., Harrison E.F. and Lee R.B. III (1990) Earth radiation budget experiment, preliminary seasonal results. *EOS*, **71**, 304–305.

Berger A. (1978) Long-term variation of daily insolation and quaternary climate changes. *Journal of the Atmospheric Sciences*, **35**, 2362–2367.

Bowen I.S. (1926) The ratio of heat losses by conduction and by evaporation from any water surface. *The Physical Review*, **27**, 779–787.

Budyko M.I. (1956) Das Wärmehaushalt der Erdoberfläche (Translated from Russian by E. Pelzl), Fachliche Mitteilungen, Nr. 100, Luftwaffenamt, Porz-Wahn, p. 282.

Budyko M.I. (Ed.) (1963) *Atlas Teplovogo Balansa Zemnogo Shara*, (Atlas of the Heat Balance of the Earth), Akademiya Nauk SSSR, Presidium Mezhvedomstvennyi Geofizichesyi Komitet: Moscow, p. 69.

Budyko M.I., Golitsyn G.S. and Izrael Y.A. (1988) *Global Climatic Catastrophes*, (English translation by V. G. Yanuta), Springer: Berlin, p. 99.

Cogley J.G. (1979) The albedo of water as a function of latitude. *Monthly Weather Review*, **107**, 775–781.

Cutler P.M. and Munro S. (1996) Visible and near-infrared reflectivity during the ablation period on Peyto Glacier, Alberta, Canada. *Journal of Glaciology*, **22**, 333–340.

Davies J.A. (1965) Albedo investigations in Labrador-Ungava. *Archives for Meteorology Geophysics and Bioclimatology Series B*, **13**, 137–151.

Doorenbos J. and Pruitt W.O. (1977) *Guidelines for Predicting Crop Water Requirements*, FAO Irrigation and Drainage Papers, No. 24.

Drewry D. (Ed.) (1983) *Antarctica: Glaciological and Geophysical Folio*, Cambridge University Press.

Fowle F.E. (1915) The transparency of aqueous vapor. *Astrophysical Journal*, **42**, 394–411.

Fröhlich C. (2004) Solar irradiance variability. *Geophysical Monograph*, **141**, 97–110.

Giambelluca T.W., Fox S., Yanasarn S., Onibutr P. and Nullet M.A. (1999) Dry-season radiation balance of land covers replacing forest in northern Thailand. *Agricultural and Forest Meteorology*, **95**, 53–65.

Giovinetto M., Bromwich D.H. and Weller G. (1992) Atmospheric net transport of water vapor and latent heat across 70 °S. *Journal of Geophysical Research*, **97D**, 917–930.

Grenfell T.C., Perovich D.K. and Ogren J.A. (1981) Spectral albedo of an alpine snow pack. *Cold Regions Science and Technology*, **4**, 121–127.

Grenfell, T.C., Warren, S.G., and Mullen, P.C., (1994) Reflection of solar radiation by the antarctic snow surface at ultraviolet, visible and near-infrared wavelengths. *Journal of Geophysical Research*, **99**(D9), 18669–18684.

Hoch S.W., Schelander P., Bourgeois C.S., Ohmura A. and Calanca P. (2004) Energy budget at Summit, Greenland.

Geophysical Research, Abstracts, Abstract No. EGU04-A-04254.

Jaeger L. (1976) *Monatskarten Des Niederschlags Für Die Ganze Erde*, Berichte des Deutschen Wetterdienstes: Nr. 139, p. 38.

Jones P.d., New M., Parker D.E., Martin S. and Rigor I.G. (1999) Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics*, **37**, 173–199.

Kiehl J.T. and Trenberth K.E. (1997) Earth's annual global mean energy budget. *Bulletin of the American Meteorological Society*, **78**, 197–208.

Kobayashi Y., Machimura T., Iwahana G., Fukuda M. and Fedorov A.N. (2001) Fire effect on flux and active layer dynamics in Taiga forest over East Siberia permafrost region. In *Research Report of Permafrost Disturbance and Induced Emission of Greenhouse Gases in 2000*, Fukuda M. and Kobayashi Y. (Eds.), pp. 156–161.

Kondtatyev K.Y.a (1969) *Radiation in the Atmosphere*, Academic Press: New York, p. 912.

Kondrat'ev K.Y.a (1973) *Radiation Characteristics of the Atmosphere and the Earth's Surface*, Amerind Publication: New Delhi, p. 580.

Kung E.C. (1988) Spectral energetics of the general circulation and time spectra of transient waves during the FGGE year. *Journal of Climate*, **1**, 5–19.

Kuhn M., Kundla L.S. and Stroschein L.A. (1977) The radiation budget at plateau station, Antarctica, 1966–1967. In Businger J.A. (Ed.) *Meteorological Studies at Plateau Station*, Antarctic Research Series, 25: Antarctica, pp. 41–73.

Lee R. (1978) *Forest Microclimatology*, Columbia University Press: New York, p. 276.

Legates D.R. and Willmott C.J. (1990) Mean seasonal and spatial variability in gauge-corrected, global precipitation. *International Journal of Climatology*, **10**, 111–127.

Marty C., Philipona R., Fröhlich C. and Ohmura A. (2002) Altitude dependence of surface radiation fluxes and cloud forcing in the Alps: results from Alpine surface radiation budget network. *Theoretical and Applied Climatology*, **72**, 137–155.

Mironova Z.F. (1973) Albedo of the earth's surface and clouds. In *Radiation Characteristics of the Atmosphere and the Earth's Surface*, Chap. 4, Kondrat'ev K.Y.a (Ed.), Amerind Publication: New Delhi, pp. 192–247.

Oguntoyinbo J.S. (1970) Reflection coefficients of natural vegetation, crops and urban surfaces in Nigeria. *Quarterly Journal of Royal Meteorological Society*, **96**, 430–441.

Ohmura A. (1981) *Climate and Energy Balance of Arctic Tundra*, Geographical Institute ETH: Zurich, p. 448.

Ohmura A. (1982) Climate and energy balance on the arctic tundra. *Journal of Climatology*, **2**, 65–84.

Ohmura A. (2001) Physical basis for the temperature-based melt-index method. *Journal of Applied Meteorology*, **40**, 753–761.

Ohmura A. (2004) *Cryosphere During the 20th Century*, Geographical Monograph, American Geophysical Union: Washington.

Ohmura A. and Gilgen H. (1993) *Re-Evaluation of the Global Energy Balance*, *Geophysical Monograph, 75*, Vol. 15, IUGG: pp. 93–110.

Ohmura A. and Raschke E. (2004) Energy budget at the earth's surface. In *Observed Global Climate*, Chap. 10, Hantel M. (Ed.), Springer-Verlag: Berlin.

Ohmura A. and Raschke E. (2005) Energy budget at the earth's surface. In *Observed Global Climate*, Chap. 10, Series Landolt-Börnstein, Hantel M. (Ed.), Springer-Verlag: Berlin.

Ohmura A. and Wild M. (2002) Is the hydrological cycle accelerating? *Science*, **298**, 1345–1346.

Ohmura A., Dutton E., Forgan B., Fröhlich C., Gilgen H., Hegner H., Heimo A., König-Langlo G., McArthur B., Müller G., Philipona R., Pinker R., Whitlock C.H., Dehne K. and Wild M. (1998) Baseline surface radiation network (BSRN/WCRP): new precision radiometry for climate research. *Bulletin of the American Meteorological Society*, **79**, 2115–2136.

Oke T.R. (1978) *Boundary Layer Climates*, Methuen: London and New York, p. 372.

Payne R.E. (1972) Albedo of the sea surface. *Journal of the Atmospheric Sciences*, **29**, 959–970.

Penman H.L. (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Meteorological Society*, **193**, 120–145.

Perovich D.K., Grenfell T.C., Light B. and Hobbs P.V. (2002) Seasonal evolution of the albedo of multiyear arctic sea ice. *Journal of Geophysical Research*, **107C**, 8044, doi:10.1029/2000JC000438.

Philipona R. and Ohmura A. (2001) Pyrgeometer absolute calibration and the quest for a world radiometric reference for longwave irradiance measurements. In *Current Problems in Atmospheric Radiation*, Smith W.L. and Timofeyev Y.uM. (Eds.), A. Deepak Publishing: Hampton, pp. 461–464.

Pinker R.T. (1982) The diurnal asymmetry in the albedo of tropical forest vegetation. *Forest Science*, **28**, 297–304.

Putnins P. (1970) The climate of Greenland. In *Climates of the Polar Regions*, Vol. 14, Chap. 2, Orvig S. and Landsberg H. (Eds.), World Survey of Climatology, Elsevier: Amsterdam, pp. 3–128.

Reeh N. (1994) Calving from Greenland glaciers: observations, balance estimates of calving law. In *Workshop on the Calving Rate of West Greenland Glaciers in Response to Climatic Change*, Danish Polar Centre: Copenhagen, pp. 85–102.

Rignot E. and Thomas R. (2002) Mass balance of polar ice sheets. *Science*, **297**, 1502–1506.

Roads J. (2002) Closing the water budget. *GEWEX News*, **12**, 1–8.

Ross J.L., Hobbs P.V. and Holben B. (1998) Radiative characteristics of the regional hazes dominated by smoke from biomass burning in Brazil: closure tests and direct radiative forcing. *Journal of Geophysical Research*, **103**(D24), 31925–31941.

Rusin N.P. (1964) *Meteorological and Radiation Regime of Antarctica* (translated from Russian). Israel Program for Scientific Translations: Jerusalem, p. 355.

Salby M.L. (1996) *Fundamentals of Atmospheric Physics*, *International Geophysics Series, 61*, Academic Press: New York, p. 627.

Sellers W.D. (1965) *Physical Climatology*, University of Chicago Press: Chicago, p. 272.

Shuttleworth W.J. (1989) Micrometeorology of temperate and tropical forest. *Philosophical Transactions of the Royal Society of London. Series B*, **324**, 299–334.

Shuttleworth W.J., Gash J.H.C., Lloyd C.R., More C.J., Roberts J., Marques Filho A.O., Fisch G., Silva Filho V.P., Ribeiro M.N.G., Molion L.C.B., *et al.* (1984) Eddy correlation measurements of energy partition for Amazonian forest. *Quarterly Journal of Royal Meteorological Society*, **110**, 1143–1162.

Sie R. (1994) *Die Albedo des Mittellandes*, ETH Zurich: Diplomarbeit, p. 135.

Takeuchi N. (2002) Surface albedo and characteristics of cryoconite on an Alaska glacier, Gulkana Glacier in the Alaskan Range. *Bulletin of Glaciological Research*, **19**, 63–70.

Takeuchi N., Kohshima S., Shiraiwa T. and Kubota K. (2001) Characteristics of cryoconite and surface albedo of a Patagonian glacier, Tyndall glacier, Southern Patagonian Icefield. *Bulletin of Glaciological Research*, **18**, 65–69.

Tsay S., King M.D., Arnold G.T. and Li J.Y. (1998) Airborne spectral measurements of surface anisotropy during SCAR-B. *Journal of Geophysical Research*, **D24**, 31943–31953.

Weidick A. (1995) Greenland. In *Satellite Image Atlas of Glaciers of the World*, Williams R.S. and Ferrigno J.G. (Eds.), US Geological Survey Professional Paper, 1386-C, US Government Ptinting Office: Washington, D.C.

Weidick A., Egede-Bøggild C. and Knudsen N.T. (1992) *Glacier Inventory Atlas of West Greenland*, Report 158, Geological Survey Greenland: Copenhagen.

Whetherald R.T. and Manabe S. (1975) The effect of changing the solar constant on the climate of a general circulation model. *Journal of the Atmospheric Sciences*, **32**, 2044–2059.

Wild M., Ohmura A., Gilgen H., Roeckner E., Giorgetta M. and Morcrette J.J. (1998) The disposition of radiative energy in the global climate system: GCM versus observational estimates. *Climate Dynamics*, **14**, 853–869.

Willson R. (1997) Total solar irradiance trend during solar cycles 21 and 22. *Science*, **277**, 1963–1965.

Z'graggen L. (2001) *Strahlungsbilanz Der Schweiz*, Doctoral Dissertation, No. 14158, Swiss Federal Institute of Technology (ETH), Zurich, p. 196.

26: Weather Patterns and Weather Types

IAN G MCKENDRY

Department of Geography, The University of British Columbia, Vancouver, BC, Canada

In popular usage, the terms "weather pattern" and "weather type" are used variously and often imprecisely to describe states of the atmosphere. An understanding of weather processes and patterns has important applications in such diverse areas as air-quality management, hydrology, water management, human health, forestry, agriculture, energy demand, transportation safety, the insurance industry, economics, and tourism. These terms are formally defined and the various classification techniques that form the basis of synoptic climatology are described. In summarizing the various methods used to identify weather patterns and types, it is abundantly clear that underlying such approaches are significant assumptions about the state of the atmosphere together with varying degrees of subjectivity inherent in the techniques themselves. It is important that these constraints are adequately acknowledged in all applications of the techniques outlined. Finally, the impacts of computing advances, new techniques, and enhanced datasets on synoptic climatology are described.

INTRODUCTION

"Weather" is defined as the instantaneous state of the atmosphere at a particular location and is typically characterized by a range of meteorological variables (or weather elements) such as pressure, temperature, humidity, wind, cloudiness, and precipitation (Barry and Chorley, 1995). Weather at any particular location is a consequence of the movement and evolution of "weather systems". These may be considered to operate at several time and space scales, ranging from the microscale (e.g. small-scale eddies recognizable as gustiness in winds) to the mesoscale (e.g. sea breezes), through the synoptic scale (e.g. midlatitude cyclones and tropical storms), and finally to the planetary scale (e.g. jet streams and Rossby waves). Variations in weather occur fundamentally because the atmosphere is a highly dynamic fluid that moves both horizontally and vertically in response to the differential heating of the earth by solar radiation. Such variability has significant impacts on human activities. Consequently, an understanding of weather processes and patterns has important applications in such diverse areas as air-quality management, hydrology, water management, human health, forestry, agriculture, energy demand, transportation safety, the insurance industry, economics, and tourism. Given such important applications, it is not surprising that considerable scientific effort has been invested in identifying the linkages between particular states of the atmosphere (as manifested by weather patterns and types) and the environment. This field of investigation, characterized by a variety of classification techniques, forms a significant part of the subdiscipline of *synoptic climatology* (Yarnal, 1993) and will be the focus of this entry.

In popular usage, the terms "weather pattern" and "weather type" are used variously and often imprecisely to describe states of the atmosphere. For example, in the popular media a tornado might be described as an "unusual weather pattern" that is associated with a particular type of weather (severe). Alternatively, an El Nino event in the central Pacific may be referred to as part of a global "weather pattern" that results in particular weather "types" (e.g. wet winters in California or drought in Australia). Although such usage does implicitly recognize the inherent scales, regularity, and interconnectedness of weather processes and systems, use of these terms in the atmospheric and hydrological sciences tends to be somewhat more specific, most commonly referring to the association of a characteristic combination of weather elements (e.g. marked by specific states of the atmosphere with respect to temperature, pressure, cloudiness, wind, and temperature) with particular weather systems. Furthermore, "weather patterns" and "weather types" tend to be most commonly linked to the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

synoptic scale of atmospheric phenomena. At this scale, the day-to-day movement and evolution of weather systems modulates and influences a wide range of human activities to a considerable degree. Implicit in both terms are notions of classification. In this context, the terms "pattern" and "type" are often used interchangeably and with sometimes subtle nuances in meaning. For example, "pattern" can be used in both its spatial and temporal senses. A repeatable sequence of weather may be regarded as a weather pattern, or more commonly, "pattern" refers to a particular arrangement of isobars on a weather map (sometimes referred to as a map pattern, synoptic type, or circulation type).

THE ORIGINS OF PATTERN

Throughout human history there has been recognition of pattern within atmospheric behavior. At the simplest level, this can be seen in the regularity of the seasons or in the association of particular wind directions with temperature (e.g. cool winds from the north in the Northern Hemisphere). At larger scales, sailors have noted and exploited the remarkable persistence of the easterly "trade" winds in the tropics, in marked contrast to the constant and somewhat chaotic succession of cyclonic systems forming and dissipating as they move eastward through the midlatitudes. In the twentieth century, development of a conceptual model of the midlatitude cyclone formalized patterns in weather elements associated with the passage of such low-pressure systems (or depressions). For example, passage of a cold front is typically associated with a particular sequence of cloud types, winds, pressure changes, precipitation, and temperature changes. Typically, in the midlatitudes, cold front passage is marked by the arrival from the west of a band of cloud (often convective in form) accompanied by rainfall and sharply decreasing temperatures, a change in wind direction (often from the south-west to the north-west in the Northern Hemisphere). Behind such a front there is often rapid clearing, rising pressures, and occasional showers. Although cold fronts exhibit considerable variability, the weather pattern described above is recognizable to most inhabitants of the midlatitudes whether in the Northern or Southern Hemispheres, and forms the basis of simple predictions that can be made on the basis of cloud and/or pressure observations.

Given the constancy of Earth's orbital characteristics, solar output, boundary conditions (earths distribution of land and water and atmospheric constituents), and axial tilt, and the universal application of forces governing atmospheric motion (i.e. linked to gradients in atmospheric pressure, friction, and the earth's rotation), it is not at all surprising that patterns in atmospheric behavior emerge across a wide spectrum of spatial and temporal scales. For example, toward the poles, the seasonal cycle increases in amplitude and significance because of the tilt of the earth on its axis. Furthermore, at the global scale, broad latitudinal bands can be identified in which particular types of weather predominate and which arise as a result of the differential heating of the planet and the application of the particular forces influencing atmospheric motion. This average condition is known as the *general circulation* of the atmosphere and in essence represents a "gigantic heat engine" (Barry and Chorley, 1995, P97). Although interrupted and modified by the global distribution of land and water, the dominant features are persistent winds from an easterly quarter in the tropics (the trade winds) and the broad zone of midlatitude westerly winds in which midlatitude depressions are constantly forming and dissipating. These features are represented for the idealized case of the Northern Hemisphere in Figure 1. Essentially the same patterns are evident in the Southern Hemisphere (i.e. south-easterly trade winds and midlatitude westerlies). However, significant differences between the two hemispheres with respect to the distribution of land (continents) and water results in significant differences in the nature and variability of the major circulation features. For example, the absence of significant landmasses in the Southern Hemisphere mid- and high latitudes, results in much stronger and persistent midlatitude westerlies than observed in the Northern Hemisphere (hence, the southern latitudes renown for their strong westerly winds: the "roaring forties", "furious fifties", and so on).

The particular scales of motion (both time and space) mean that the atmosphere assumes patterns that are geographically constrained and quasi-repetitive. Superimposed on these broad geographical patterns is variability in a range of scales. For example, west coast midlatitude continental locales (e.g. Washington State and British Columbia, southern Chile, The British Isles, and Norway) are exposed to a succession of eastward moving cyclonic systems during winter. Then, during summer, the subtropical high-pressure systems expand poleward providing some relief from the characteristic cool, moist, and highly variable weather patterns that tend to dominate in these regions due to their continental location and latitudinal position. In other areas (e.g. southeast Asia and Africa), monsoon patterns may dominate the annual weather cycle (such patterns are associated with a continental-scale seasonal wind reversal accompanied by distinct changes in precipitation). Furthermore, at large scales, natural modes of oscillation in the atmosphere/ocean circulation, as manifested in the El-Nino-Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), the Pacific Decadal Oscillation (PDO), the Arctic Oscillation (AO) and the Madden–Julian Oscillation, play an important role in modulating weather patterns at interannual and decadal timescales.

Perhaps the best known are the oscillatory patterns associated with ocean–atmosphere interactions in the Pacific

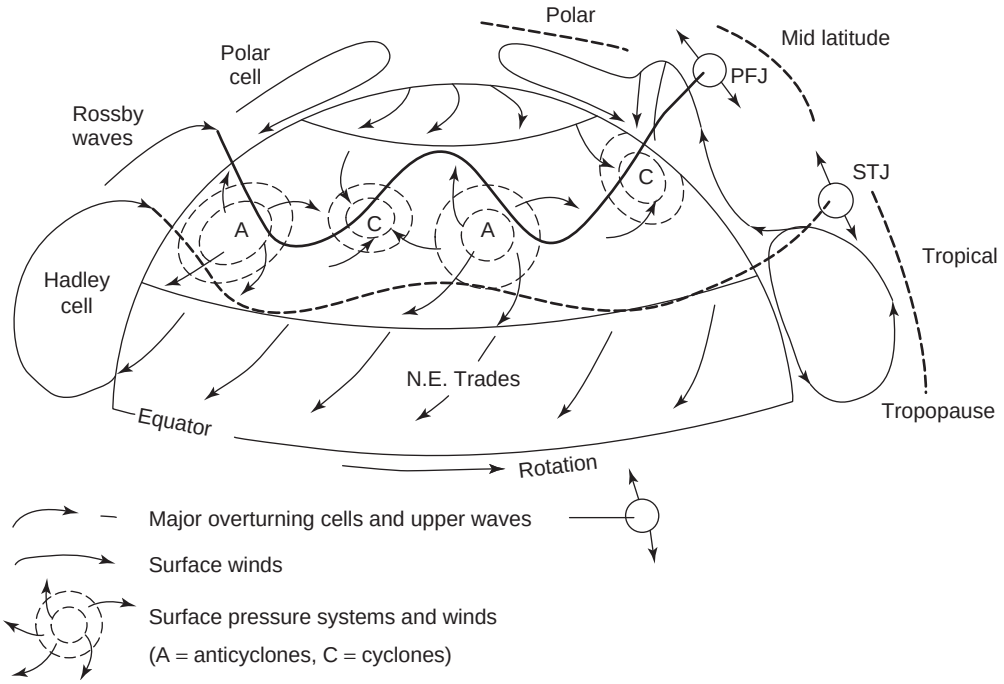


Figure 1 A model of the general circulation of the atmosphere for the northern hemisphere. Note that this is essentially a tropospheric view of the atmosphere with the vertical extent of the circulation constrained to the bottom ~20 km of the atmosphere. **STJ** and **PFJ** refer to the subtropical and polar front jets respectively (Reproduced from Oliver and Hildore, 2002, © Pearson Education)

Basin. Clear evidence exists that ocean–atmosphere interactions in the entire Pacific Basin induce significant climatic variability at a range of timescales with an influence that extends well beyond the Basin itself. Because of the sheer size of the Pacific Ocean, these interactions represent the clearest signal of such atmosphere–ocean interactions globally. These interactions are by nature cyclic and are manifested in changing spatial patterns of sea surface temperature (SST) that in turn drive changes in tropical convection and precipitation that ultimately influence the position and intensity of the jet stream in the midlatitudes (an example of a *teleconnection*). Consequently, midlatitude storm tracks, storm frequency, and intensity are affected (Cayan *et al.*, 1998, 1999). At timescales of years to decades, two principal modes of variability have been identified in the Pacific Basin:

El Niño-Southern Oscillation (ENSO): a well-known phenomenon characterized by an east–west "seesaw" pattern in tropical SSTs that operates on timescales of months to years (the last major El Niño events occurred in 1983, 1992–1993, and 1997–1998). The atmospheric mode of the phenomenon is known as the *Southern Oscillation*, the intensity and phase of which is measured by the southern oscillation index (SOI – the surface pressure difference across the tropical pacific) (between Tahiti and Darwin). During *El Niño* events (negative SOI), SSTs in the tropical eastern Pacific (off Peru) are higher than "normal" and produce excess convection and precipitation in that region. *La Nina* events (positive SOI) are characterized by unusually warm SSTs in the western Pacific and produce strong convection and precipitation over those regions.

Pacific Decadal Oscillation (PDO): a recent discovery, the PDO is manifested by El Niño–like changes (so-called *regime shifts*) in the SST distribution over the tropical and northern Pacific evident at decadal timescales (Mantua *et al.*, 1997). The warm (positive) phase of the PDO is characterized by below-normal SSTs in the central and western north Pacific and unusually warm SSTs along the west coast of North America. The cold phase (negative) produces the reverse distribution.

Individual phases of this oscillation typically last for 23–35 years resulting in a 50–75 year cycle (Minobe, 1997). This century, "cool" PDO regimes prevailed from 1890 to 1924 and again from 1947 to 1976, while "warm" PDO regimes dominated from 1925 to 1946 and from 1977 through (at least) the mid-1990s (Figure 2). Studies utilizing tree-ring analysis suggest that the PDO has played an important role in Pacific climate since at least the 1700s (D'Arrigo *et al.*, 2001; Biondi *et al.*, 2001).

This pattern is strongly linked to atmospheric circulation over North America and the North Pacific as commonly expressed by the Pacific North American (PNA) atmospheric pressure index (PNA). Low values of the PNA index are associated with a weak Aleutian low pressure,

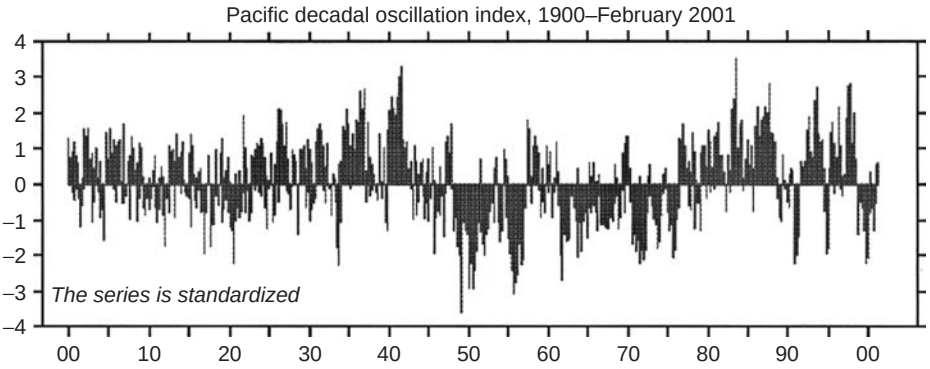


Figure 2 The Pacific Decadal Oscillation Index 1900–2001 (Reproduced from **http://www.jisao.washington.edu/data_sets/pdo/**)

while high values are associated with a strong Aleutian low. High values of the PNA tend to be associated with the warm phase of PDO (e.g. 1977 to about 1998) and El Nino events (Hsieh and Tang, 2001).

North Atlantic Oscillation (NAO)

Many of the elements of atmosphere–ocean variability found in the very large Pacific Basin are also evident in the Atlantic Ocean (Hurrell and van Loon, 1997). The NAO is the dominant mode of winter climate variability in the North Atlantic region encompassing central North America, Europe, and much of Northern Asia. It represents a large-scale seesaw in atmospheric mass (pressure) between the subtropical high and the polar low. The NAO index exhibits considerable year-to-year variability, but also shows a tendency to remain in one phase for several years. The positive NAO index phase has a stronger-than-usual subtropical high-pressure center and a deeper-than-normal Icelandic low. This leads to a higher frequency of, and more intense, winter storms crossing the Atlantic Ocean. As a consequence, winters in Europe tend to be warm and wet, while Canada and Greenland experience dry, cold conditions. In contrast, the negative NAO index phase is characterized by a weak subtropical high and a weak Icelandic low. This condition typically brings moist air into the Mediterranean, cold air to northern Europe, and a higher frequency of cold air outbreaks to the eastern United States of America.

TECHNIQUES TO IDENTIFY AND CLASSIFY WEATHER PATTERNS AND WEATHER TYPES

Given the obvious tendency of the atmosphere to exhibit patterned and repetitive (if not necessarily predictable) behavior at a range of spatial and temporal scales, it is hardly surprising that climatologists have attempted to identify and classify such pattern in order to provide a linkage between myriad human activities and/or the surface environment and atmospheric circulation. This has been an ongoing and developing research focus in the atmospheric and hydrological sciences over the twentieth century and provides the basis for the scientific meaning of the terms "weather pattern" and "weather type". Many techniques have been developed in order to identify and classify weather patterns and conditions. These are summarized in an excellent primer on the topic by Yarnal (1993) and updated in Yarnal *et al.* (2001). Yarnal's terminology and taxonomy of techniques is discussed in the following sections.

Manual Classification

In this approach, "weather patterns" and weather types" are established on the basis of a subjective classification of synoptic data (i.e. weather observations of various types or weather maps) by the researcher. Because of the current predominance of computer-assisted approaches, this technique has lost favor somewhat. However, subjective classification of weather/circulation patterns or types from visual analysis of synoptic weather maps remains the foundation of three well-known classification schemes that have demonstrated remarkable longevity: (i) the Muller classification of synoptic types over the United States (Muller, 1977) (ii) the Lamb Catalog for the British Isles (Lamb, 1972), and (iii) the Grosswetterlagen (Hess and Brezowsky, 1977) for central Europe.

The Lamb catalog neatly illustrates the manual approach. Using daily weather maps, Lamb (1972) recognized seven basic weather types based on the curvature and orientation of surface isobars:

A. Anticyclonic
C. Cyclonic
W. Westerly
NW. Northwesterly
N. Northerly

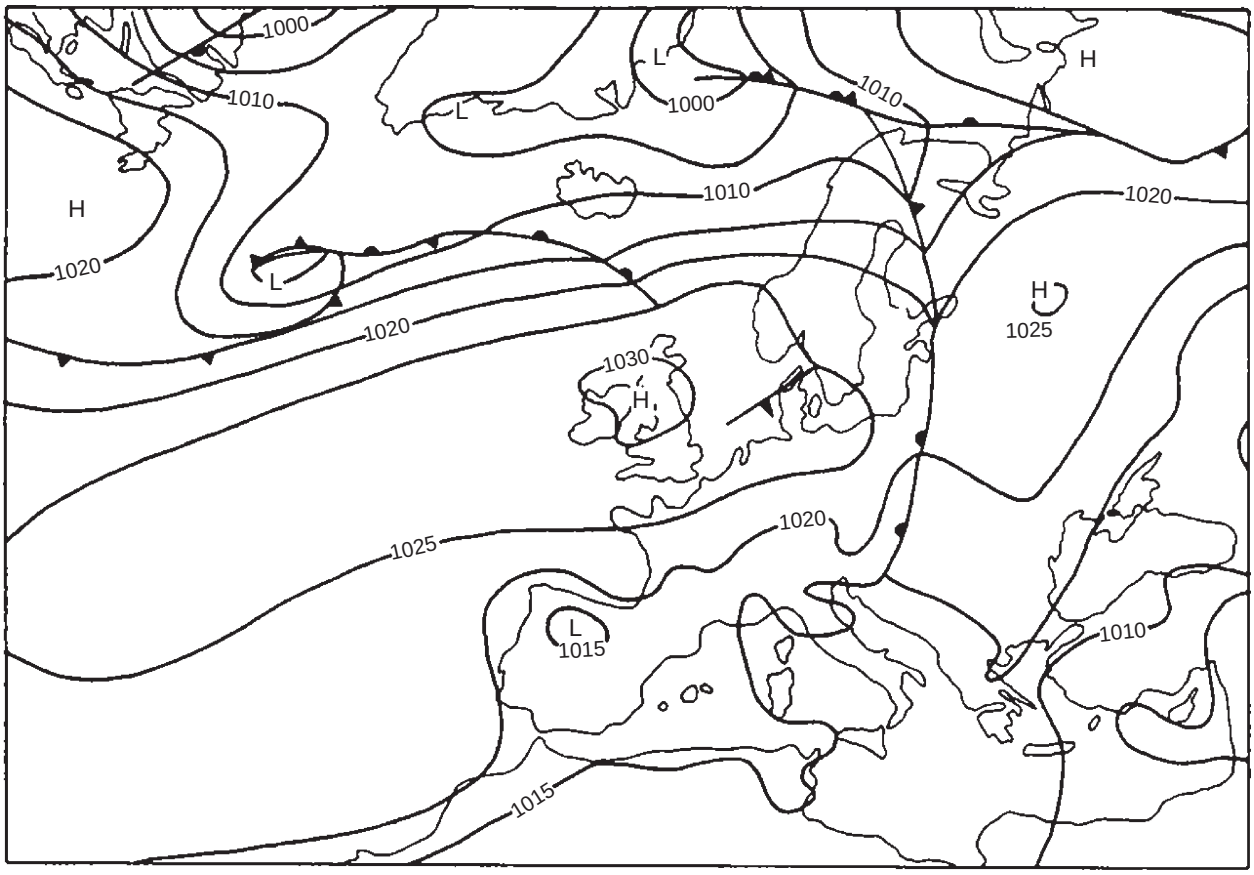


Figure 3 An example of Lamb's weather type A (Anticyclonic) (Reproduced from Henderson-Sellers and Robinson, 1986)

D. Easterly
S. Southerly

For example, Type A was associated with high-pressure areas dominating over the British Isles (Figure 3). Typical weather conditions include light winds, warm weather in summer, and cold conditions in winter. The catalog was further extended on the basis of unclassifiable and hybrid days. Consequently, anticyclonic northwesterly isobaric curvature over the region would give rise to ANW types and could be distinguished from CNW types. On this basis, a total of 27 types could be identified. Lamb subjectively classified each day's weather over the British Isles from 1861 to February 1997. All the years' data have been published, except for 1996–1997, in Lamb (1972) and Hulme and Barrow (1997). Since Lamb's death, an objective version of the scheme has been developed by Jenkinson and Collison (1977) and continues to be updated.

Although labor intensive, subjective, and not easily reproduced, manual approaches to identifying weather patterns and types are still utilized (see for example, Buishand and Brandsma, 1997), primarily as a result of the degree of control and specificity they offer the researcher.

Correlation-based Approaches

In many respects, correlation-based approaches simply represent an automated version of the manual approaches described above. In its simplest sense, the approach involves the correlation of synoptic weather maps (pressure fields) with one another in order to form weather types. Correlation threshold values are used to determine the membership of discrete types or categories. This approach was first introduced by Lund (1963) using Pearson product moment correlations and later by Kirchhofer (1973) using a sum-of-squares algorithm that has been shown to be essentially identical to Lund's technique.

In this procedure, each grid of pressure values (a weather map) is first normalized to remove the seasonal cycle:

$$Z_i = \frac{x_i - \overline{x_i}}{s} \tag{1}$$

Where Z_i is the normalized value of the grid point i, x the mean of the N-point grid, and s the standard deviation of the grid. Each normalized grid (map) is then compared

with all other grids by the sum-of-squares equation:

$$S = \sum_{i=1}^{N} (Z_{ai} - Z_{bi})^2 \qquad (2)$$

Where S is the Kirchhofer score, Z_{ai} the normalized grid value of point i on day a, Z_{bi} the grid value of point i on day b, and N the number of data points. Days a and b represent any pair of maps from which the Kirchhofer score S is calculated.

Subscore values for each row and column of the matrices are also calculated using equation (2) in order to ensure pattern similarities in all areas of the grid. Maps are considered to be similar, based on predetermined thresholds. Within the period of maps being considered (usually several years), the day that has most S values meeting the threshold criteria associated with it is designated *keyday1*. This key day is then removed from the analysis as are all days that met thresholds of correlation with that day, and all days associated with those days. This analysis is then repeated with all remaining days to find *keyday 2*, and so on, until all days are classified into m groups of five days or more. The remainder are deemed "unclassified".

Once the *keydays* are established, equation (2) is then used to compare each of the *keydays* with all other grids (maps). Each day is then assigned to the class associated with the *keyday* for which it produces the lowest S value. This represents the final step in establishing classes.

Although still involving subjective decisions, correlation-based approaches are intuitive, simple to implement and reproducible. Consequently, they have been used extensively in a wide range of applications. For example, McKendry (1994) uses a combination of correlation-based approaches and compositing (see the following text) to elucidate meteorological impacts on air quality in the Lower Fraser Valley (LFV) of British Columbia. Twelve of the 17 types identified for mean sea level maps by the Kirchhofer approach are shown in Figure 4. The relation between weather types and ozone concentrations is shown in Figure 5 and demonstrates the association of elevated concentrations of ozone with anticyclonic conditions and an appropriately positioned ridge of high-pressure at 500 hPa.

The critical importance of coupling between the surface and upper-level flow patterns to O_3 concentrations in the LFV (McKendry, 1994) suggest that traditional synoptic climatological approaches that emphasize classification of a single atmospheric level (usually the surface) within a static framework may have limited applications. This LFV ozone analysis highlights the extent to which particular surface circulation types may be associated with a range of 500 hPa types. Consequently, only by explicit consideration of the three-dimensional structure of the atmosphere can the meteorological controls on O_3 concentrations be satisfactorily explained. Furthermore, the LFV example confirms that meteorological controls on air quality can only be examined within a dynamic framework. Elevated O_3 concentrations appear to be strongly controlled by antecedent synoptic conditions and the degree of persistence in the evolving synoptic state. Within this context, the sequencing and compositing techniques developed by Comrie (1992) are particularly useful.

Eigenvector-based Approaches

With the growth and accessibility of computing power, more sophisticated, elegant, parsimonious, and statistically robust techniques have emerged for the classification of weather types and patterns. These approaches exploit a variety of eigenvector-based techniques that permit the isolation of clusters in large and diverse data sets and include principal components analysis (PCA), empirical orthogonal functions (EOF), discriminant analysis and cluster analysis. Application of such techniques requires a high degree of statistical competency, but provides significant reward with respect to statistical rigor and interpretation. Recent examples of the application of such multivariate techniques include Kidson's (2000) investigation of weather regimes affecting New Zealand and the study of Romero *et al.* (1999a, b) on precipitation patterns over the Spanish Mediterranean.

Eigenvector-based techniques have found greatest application in synoptic type classifications (where a range of weather variables such as temperature, cloud cover, wind speed, and direction are classified) rather than map-pattern classification. This may be attributed to the fact that maps of component loadings are less easily interpreted than the pressure fields that result from other classification techniques (Yarnal, 1993).

Kalkstein and Corrigan (1986) nicely illustrate the approach as developed primarily by Kalkstein and his collaborators at the University of Delaware. With a view to examining the meteorological controls on sulfur dioxide concentrations, they took seven weather variables (recorded 4 times per day) over five winters (451 days) to create a 28 by 451 P-mode data matrix. This was then converted to a correlation matrix and an unrotated PCA applied. Five principal components were retained (explaining 78% of the variance). A clustering technique was then used to reduce the component scores matrix to 10 synoptic types. Means of the weather variables were then calculated for each of the types. Representative weather maps were also examined for each type as an aid to understanding and interpretation.

Despite the complexity and rigor of the eigenvector-based approaches, the researcher exploiting this approach is faced with myriad decisions regarding mode of decomposition, types of rotation, and type of dispersion matrix. In essence, each application is, therefore, unique.

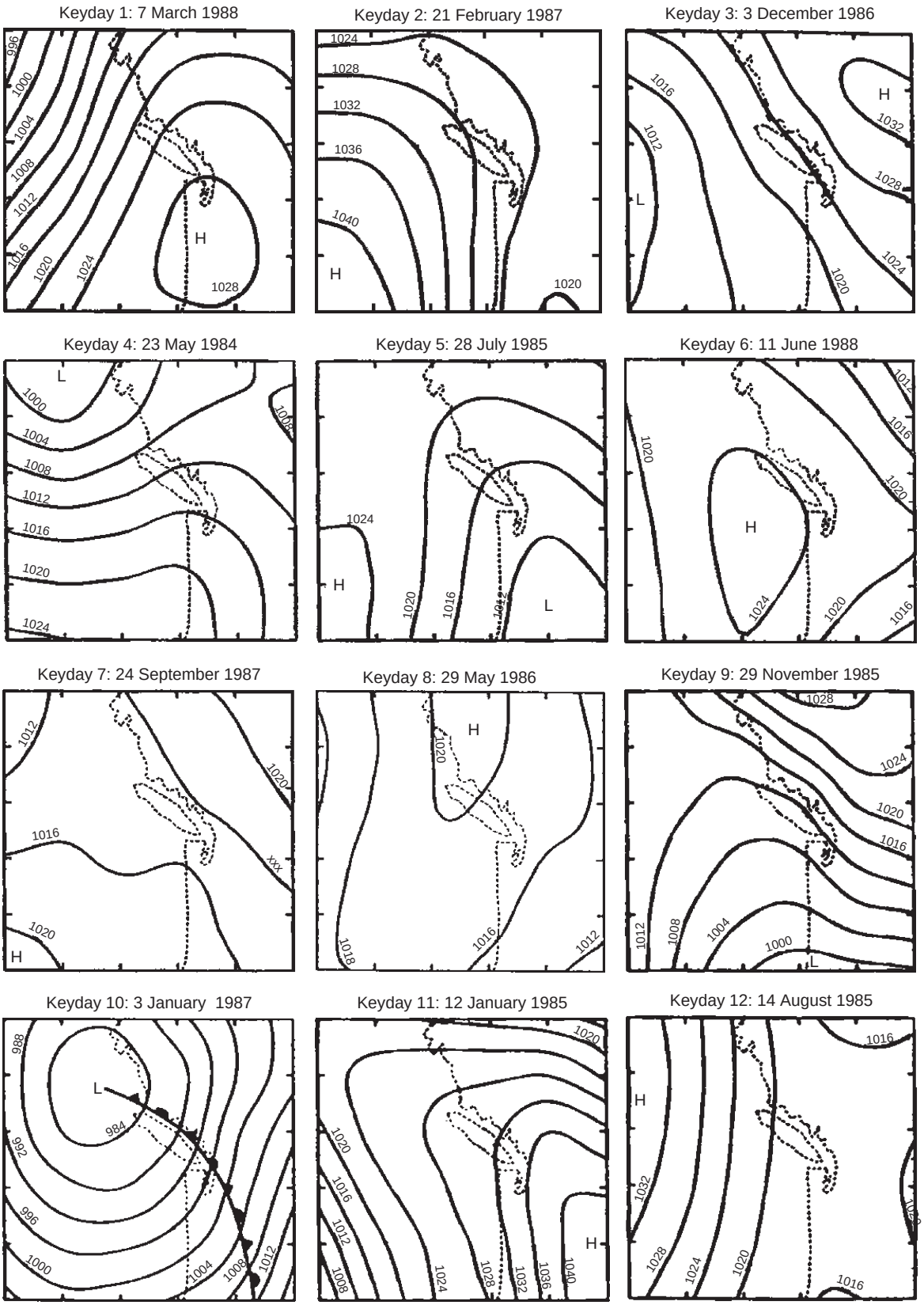


Figure 4 Keydays representing 12 of 17 Mean Sea Level synoptic types identified using the Kirchhofer Scheme for British Columbia (Reproduced from McKendry, 1994, © American Meteorological Society)

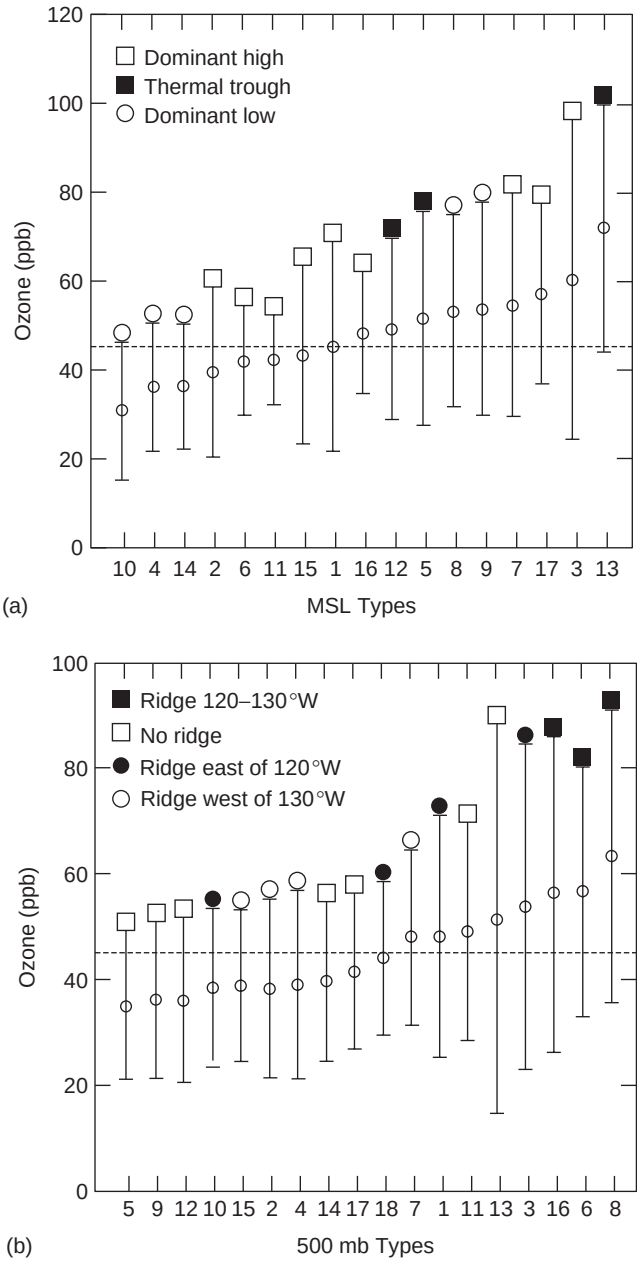


Figure 5 Mean (with standard deviation bars) daily maximum ozone concentrations for a Vancouver, B.C., station 1978–93 (May–September) by (a) MSL and (b) 500 hPa synoptic types (Reproduced from McKendry, 1994)

Compositing and Indexing

Composites represent "average" weather maps or weather patterns and are easily developed on the basis of a particular criterion. For example, gridded pressure fields (synoptic weather maps) for days on which ozone concentrations or precipitation totals reach a particular threshold over a region, may be averaged to produce a composite map. For example, McKendry (1994) produced composite map and weather condition sequences associated with exceedances of the 82 ppb Canadian ozone "objective". One such example is shown in Figure 6.

Indexing, on the other hand, involves the assignment of a single number or "index" to capture the essence of a particular weather pattern or type (Yarnal, 1993). Using this approach, one can develop a time-series that can be compared/correlated with other indexes such as those used with large-scale climatic fluctuations like ENSO, PDO, or AO. Recent examples of this approach include studies by Greene *et al.* (1999), who constructed a synoptic climatology of air pollution in the United States by using a synoptic index, and Speer and Leslie (1997), who matched frequencies of Australian ridging events with ENSO variations.

New Developments

To date, approaches to the specification of weather patterns and types have spanned a continuum, from simple manual classification techniques to highly sophisticated computer-assisted statistical approaches. Both extremes have their advantages. Recently, new approaches have been developed that exploit the simplicity of the manual approach while maintaining the rigor of the automated statistical methodologies. An example of such a hybrid scheme is the Spatial Synoptic Classification (SSC2) as described by Sheridan (2002) and produced for the coterminous USA. This approach exploits six subjectively defined weather types (similar to the air mass types defined by Bergeron (1930) as developed by Kalkstein *et al.* (1996). These are dry polar (DP), dry moderate (DM), dry tropical (DT), moist polar (MP), moist moderate (MM), and moist tropical (MT). Each weather type is defined for all locations across the continent using statistical analysis of meteorological data for each station (temperature, dew point wind, cloud cover, and sea level pressure). Output of year-round weather types for each day is achieved by an equally weighted-sum of squared z-scores and results in a daily calendar of weather types for each station. These can be found at `http://sheridan.geog.kent.edu/sscnow.html` for stations across Canada and the United States.

Modifications to the traditional classification approaches described above continue to be developed, but they are largely variations on a theme. Perhaps the most promising is the exploitation of Geographical Information Systems (GIS) and artificial intelligence approaches (e.g. artificial neural networks) in the identification of weather patterns and types. Yarnal *et al.* (2001) note, that as yet, GIS has made few implicit inroads into synoptic climatology. In the areas of data visualization and data management, GIS has contributed to the evolution of the discipline, but has made little progress analytically.

The free availability of large high quality data sets (such as the National Centers for Environmental Protection

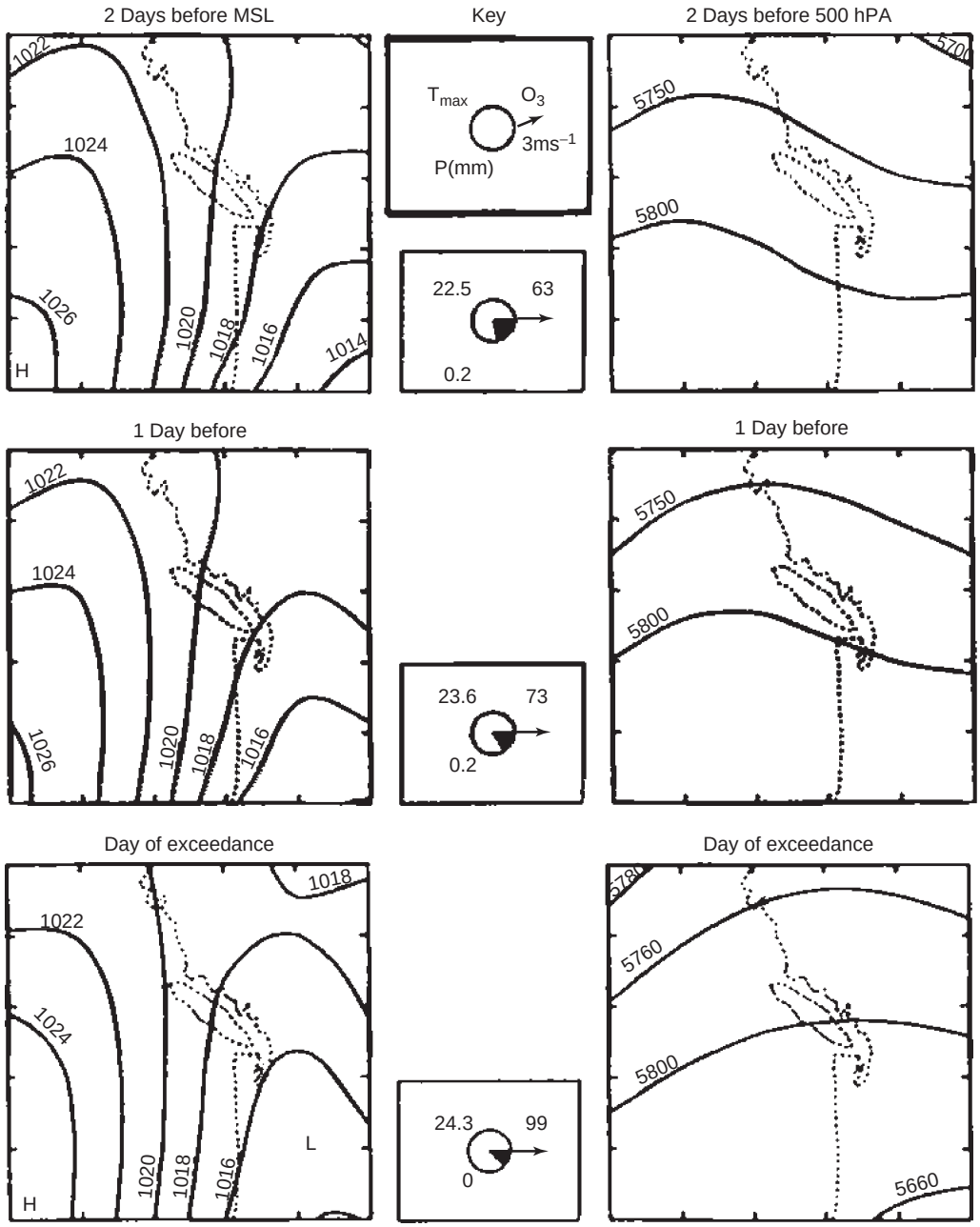


Figure 6 A three-day composite synoptic sequence showing MSL and 500 hPa maps associated with severe ozone episodes (''exceedances'') in south-western British Columbia. Also shown are station circles representing the average weather (cloud cover, wind, maximum temperature (°C), precipitation and ozone concentration (ppb)) associated with the composites (Reproduced from McKendry, 1994, © American Meteorological Society)

(NCEP) reanalysis) together with improved computational power and user-friendly computer interfaces has made synoptic typing a much more accessible activity. An excellent example is the "Synoptic Typer", an application designed in Australia to perform automated objective synoptic typing using a pattern recognition scheme. A history of grid point analyses is used to develop the synoptic types using both principal component and cluster analysis techniques. Real-time NWP model output is then classified to produce synoptic type forecast guidance (**`http://www.bom.gov.au/inside/cosb/mss/projects/synoptictyper/`**).

Assumptions and Limitations

In the field of synoptic climatology, considerable energy is being spent in the identification of weather types and

weather patterns. In doing so, much has been achieved with respect to improving understanding of the means by which weather modulates human activities, together with a broad range of environmental variables. Furthermore, this enterprise has provided a firm basis for prediction in such areas as air quality, water management, tourism, and transportation. However, Yarnal (1993) draws attention to important assumptions that should be borne in mind, when considering weather patterns and types as derived by the various techniques above. Generally, it is assumed that

1. *atmospheric circulation is a critical determinant of the surface environment* – this is clearly the case with respect to air quality, but may not be the case for some environmental processes (e.g. tectonic activity);
2. *current conceptual models of the midlatitude cyclone are correct* – there are still uncertainties with respect to the structure and behavior of midlatitude systems (e.g. the behavior and structure of rainbands and fronts);
3. *the atmosphere can be partitioned into discrete, nonoverlapping intervals (i.e. weather types)* – Yarnal (1993) suggests that the atmosphere is a multidimensional continuum and therefore, patterns and types really represent important clusters in multidimensional space;
4. *classification identifies all important map patterns or synoptic types* – establishment of weather patterns and types necessarily involves subjective decisions about the extent of within-class variability that is acceptable;
5. *classification techniques achieve what the investigator really expects* – classification with complex statistical packages often is akin to a "black-box" approach. The onus is on the climatologist to make informed decisions and interpretations;
6. *the temporal and spatial scales of observations and atmospheric circulation processes match* – for example, local air quality is strongly influenced by local winds such as sea breezes and slope winds. Such flows are not necessarily reflected in synoptic scale maps. Consequently, a perfect match between synoptic scale processes and pollutant concentrations should not be expected.
7. *variability within classes of weather patterns or types is not problematic* – in reality such variability is both problematic and inevitable.

For each assumption, Yarnal (1993) identifies significant concerns, and illustrates that in practical synoptic climatologies in which weather patterns and types are identified, each of these assumptions is violated to some degree. In essence, weather patterns and types can only be regarded as approximations, and they neglect the fact that the atmosphere really represents a continuum in which weather is infinitely varying.

CONCLUSIONS

Despite the marked variability and seemingly chaotic behavior of the atmosphere, it is, in fact, characterized by quite a remarkable order and regularity at a range of scales. This permits the identification of recognizable and repeatable weather patterns and types at particular localities. These form the basis of climate and are best expressed at the synoptic scale, the scale of weather variability seen over regions on timescales of hours to days. Such variability, expressed for example, in the passage of a midlatitude cyclone, a tropical cyclone, or the arrival of monsoon rains, has important implications for myriad human activities and environmental variables. Synoptic climatologists have adopted a variety of techniques of varying degrees of sophistication in order to identify and classify such synoptic weather patterns and types. These approaches range from the simple, subjective classification of weather maps into classes, to complex statistical analyses of multiple meteorological variables. Whatever the approach, important linkages between the atmospheric circulation and the environment have been identified. The range of applications is enormous. Such approaches have been successfully exploited to explore connections between weather patterns and human health, air pollution, precipitation, agricultural production, fluvial hydrology, severe weather events, to name just a few.

However, in summarizing the various methods used to identify weather patterns and types, it is abundantly clear that underlying such approaches are significant assumptions about the state of the atmosphere (Yarnal, 1993; Yarnal *et al.*, 2001) together with varying degrees of subjectivity inherent in the techniques themselves. It is important that these constraints are adequately acknowledged in all applications of the techniques outlined herein.

The science of identifying weather patterns and types is not static. The field continues to grow and change. Combinations of traditional techniques are being used to considerable advantage (e.g. hybrid classifications) while new approaches are emerging. As computing power continues to increase, together with application of new techniques and enhanced datasets, it is likely that synoptic climatology will grow in terms of its potential applications and sophistication. The possibility of classifying a fully three-dimensional atmosphere has become a realistic goal, and the availability of online synoptic climatologies and typing software holds the promise of much wider exploitation of the approach.

FURTHER READING

Jones P.D., Hulme M. and Briffa K.R. (1993) A comparison of Lamb circulation types with an objective classification scheme. *International Journal of Climatology*, **13**, 655–663.

Lamb H.H. (1991) British Isles daily wind and weather patterns 1588, 1781–86, 1972–91 and shorter early sequences (in 1532, 1570 and other years, notably 1688, 1689, 1694, 1697, 1703, 1717, 1783–4, 1791, 1792, 1795, 1822, 1825, 1829, 1845, 1846, 1849, 1850, 1854–5. *Climate Monitor*, **20**, 47–70.

REFERENCES

Barry R.G. and Chorley R.J. (1995) *Atmosphere, Weather and Climate*, *Sixth Edition*, Routledge: p. 392.

Bergeron T. (1930) Richtlinien einer dynamischen klimatologie. *Meteorologische Zeitung*, **47**, 246–262.

Biondi F., Gershunov A. and Cayan D.R. (2001) North Pacific decadal variability since 1661. *Journal of Climate*, **14**, 5–10.

Buishand T.A. and Brandsma T. (1997) Comparison of circulation classification schemes for predicting temperature and precipitation in the Netherlands. *International Journal of Climatology*, **17**, 875–889.

Cayan D.R., Dettinger M.D., Diaz H.F. and Graham N.E. (1998) Decadal Variability of precipitation over western North America. *Journal of Climate*, **11**, 3148–3165.

Cayan D.R., Redmond K.T. and Riddle L.G. (1999) ENSO and Hydrologic extremes in the western United States. *Journal of Climate*, **12**, 2881–2893.

Comrie A.C. (1992) An enhanced synoptic climatology of ozone using a sequencing technique. *Physical Geography*, **13**, 53–65.

D'Arrigo R., Villalba R. and Wiles G. (2001) Tree-ring estimates of Pacific decadal variability. *Climate Dynamics*, **18**, 219–224.

Greene J.S., Kalkstein L.S., Ye H. and Smoyer K. (1999) Relationships between synoptic climatology and atmospheric pollution at four US cities. *Theoretical and Applied Climatology*, **62**, 163–174.

Henderson-Sellers A. and Robinson P.J. (1986) *Contemporary Climatology*, Longman Scientific and Technical: p. 439.

Hess P. and Brezowsky H. (1977) Katalog der Grosswetterlagen Europas (1881-1976). *Berichte des Deutschen Wetterdienstes*, **15**(113), (Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main, Germany).

Hsieh W.W. and Tang B. (2001) Interannual variability of accumulated snow in the Columbia basin, British Columbia. *Water Resources Research*, **37**, 1753–1759.

Hulme M. and Barrow E. (Eds.) (1997) *Climate of the British Isles: Present, Past and Future*, Routledge: London, p. 454.

Hurrell J.W. and van Loon H. (1997) Decadal variations in climate associated with the North Atlantic Oscillation. *Climatic Change*, **36**, 301–326.

Jenkinson A.F. and Collison F.P. (1977) An initial climatology of gales over the North Sea. *Synoptic Climatology Branch Memorandum No. 62*, Meteorological Office: Bracknell.

Kalkstein L.S. and Corrigan P. (1986) A synoptic climatological approach for environmental analysis: Assessment of sulphur dioxide concentrations. *Annals of the Association of American Geographers*, **13**, 68–75.

Kalkstein L.S., Nichols M.C., Barthel C.D., Greene J.S. (1996) A new spatial synoptic classification: application to air mass analysis. *International Journal of Climatology*. **16**, 983–1004.

Kidson J.W. (2000) An analysis of New Zealand synoptic types and their use in defining weather regimes. *International Journal of Climatology*, **20**(3), 299–316.

Kirchhofer W. (1973) Classification of European 500mb patterns, *Arbeitsbericht der Schweizerischen Meteorologischen Zentralanstalt* Nr. 43, Geneva.

Lamb H.H. (1972) British Isles Weather types and a register of daily sequence of circulation patterns, 1861–1971. *Geophysical Memoir 116*, HMSO: London, p. 85.

Lund I.A. (1963) Map-pattern classification by statistical methods. *Journal of Applied Meteorology*, **2**, 56–65.

Mantua N.J., Hare S.R., Zhang Y., Wallace J.M. and Francis R.C. (1997) A Pacific Inter-decadal Climate Oscillation with Impacts on Salmon Production. *BAMS*, **78**, 1079.

Minobe, S. 1997 A 50–70 year climatic oscillation over the North Pacific and North America. *Geophysical Research Letters* **24**(6), 683–686.

McKendry I.G. (1994) Synoptic circulation and summertime and ground-level ozone concentrations at Vancouver, British Columbia. *Journal of Applied Meteorology*, **33**(5), 627–641.

Muller R.A. (1977) A synoptic climatology for environmental baseline analysis: New Orleans. *Journal of Applied Meteorology*, **2**, 56–65.

Oliver J.E. and Hildore J.J. (2002) *Climatology: An Atmospheric Science*, *Second Edition*, Prentice Hall: p. 410.

Romero R., Ramos C. and Guijarro J.A. (1999a) Daily rainfall patterns in the Spanish Mediterranean area: an objective classification. *International Journal of Climatology*, **19**(1), 95–112.

Romero R., Sumner G., Ramis C. and Genovés A. (1999b) A classification of the atmospheric circulation patterns producing significant daily rainfall in the Spanish Mediterranean area. *International Journal of Climatology*, **19**(7), 765–785.

Sheridan S.C. (2002) The redevelopment of a weather-type classification scheme for North America. *International Journal of Climatology* **22**, 51–68.

Speer M.S. and Leslie L.M. (1997) A climatology of coastal ridging over south-eastern Australia. *International Journal of Climatology*, **17**, 831–845.

Yarnal B. (1993) *Synoptic Climatology in Environmental Analysis: A Primer*, Belhaven Press: London, p. 195.

Yarnal B., Comrie A.C., Frakes B. and Brown D.P. (2001) Developments and prospects in synoptic climatology. *International Journal of Climatology*, **21**, 1923–1950.

27: Storm Systems

CHARLES A DOSWELL III

University of Oklahoma, Norman, OK, US

The concept of a storm is defined as a weather disturbance that, in the context of hydrological science, produces precipitation or affects the formation and distribution of precipitation. Storm systems are reviewed in terms of their spatial and temporal scale, with the dominant storm system on large scales in midlatitudes being the so-called extratropical cyclones. Not only do such storms produce considerable weather in their own right, but they provide a setting in which the so-called mesoscale storm systems can develop. Mesoscale weather is not dominated by any particular form of disturbance, but can take many different forms. The intensity of the weather, including precipitation rates, in mesoscale systems can be much greater than in extratropical cyclones although the duration of such weather is typically less than in large-scale systems. Small-scale storms are dominated by convective storms, which can produce the most violent weather of all, including large hail, strong winds, tornadoes, and torrential rainfalls.

INTRODUCTION

The word "storm" is generally defined as a disturbance in the weather. In the context of hydrological science, the most obviously relevant type of weather disturbance is one that produces precipitation. The deposition of precipitation is the result of a number of processes, but the most basic requirement for precipitation is the ascent of air that contains water vapor. Ascending air cools by adiabatic expansion, and condensation begins when this cooling results in relative humidities at or near 100% (*see* **Chapter 28, Clouds and Precipitation, Volume 1**). This condensation produces clouds that can develop precipitation following continued ascent of the air (see Lamb (2001) for a more complete description of how precipitation is formed). Once it begins, the instantaneous rate of precipitation, R, is roughly proportional to the product Ewq, where E is the efficiency at which water vapor is converted to precipitation that reaches the surface, w is the ascent rate, and q is the water vapor mixing ratio (the mass of water vapor per unit mass of air) in the ascending air (Doswell *et al.*, 1996). When the average precipitation rate, $\overline{R}$, is multiplied by the duration of the precipitation, D, the result is the total accumulated precipitation during the storm. Typically, the efficiency of precipitation production by storms is not particularly high, perhaps being on the order of 50%, occasionally much more, and often much less, and can vary from one storm to the next, as well as during the life cycle of a single storm. Precipitation efficiency depends strongly on the environment in which a storm occurs – dry environments increase the likelihood that precipitation will evaporate before it reaches the surface. Strong variation of the horizontal wind with height, called *vertical wind shear*, is also thought to reduce precipitation efficiency.

The water vapor needed for precipitation is present in the air mostly as a result of the process of *evapotranspiration* – the combined effects of evaporation (*see* **Chapter 45, Actual Evaporation, Volume 1**) from open liquid water (frozen forms of water do not release much water vapor) and transpiration of water vapor by vegetation (*see* **Chapter 42, Transpiration, Volume 1**). However, precipitation does not always fall at or near the place where water vapor is first introduced into the air. Rather, the moving air, or *wind*, often transports that added water vapor away from its source before it ultimately falls out somewhere else as precipitation. Therefore, even nonprecipitating weather systems are hydrologically pertinent, because they are associated with the atmospheric water vapor *transport* component of the hydrologic cycle. As a result of this transport, precipitation can fall in geographic locations that have little or no local evapotranspiration (e.g. the polar regions, or deserts), for lack of vegetation and/or open

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

liquid water. The movement of air also alters the transpiration *rate*, so any discussion of hydrologically important storm systems must also consider those processes that control the wind.

Storm systems operate on a variety of spatial and temporal scales, and that is the basis for an orderly consideration of them herein. In order to understand storm systems, it is essential to know the basic physics that govern atmospheric motion. Air within the lower atmosphere is a mixture of many gases, but is mostly nitrogen (about 78%), oxygen (about 20%), argon (about 1%), carbon dioxide (less than 1%), and water vapor (variable, 1% or less). Of these gases, only the water vapor content varies much within the lower atmosphere. The fact that water vapor is a constituent gas that, unlike the others, can change phase to liquid or solid forms within the range of temperatures and pressures found within the atmosphere is a critical factor in the evolution of weather systems. The magnitude of this contribution to the total energy of storm systems by water substance is due to the relatively high *latent heat* of water. When condensation of water vapor occurs, latent heat is released, often enhancing the processes leading to upward air motions that initiated condensation in the first place. Storms systems on the Earth are stronger than they would be if the planet was dry and without significant latent heat release from condensation (like the atmosphere of Mars). Clouds produced by condensation also influence strongly the local radiation balance and so alter the spatial temperature distribution. Incoming solar radiation is the ultimate energy source for all weather, and clouds are important factors in the local energy budget.

Atmospheric processes are governed by a complex set of equations that are the mathematical expression of physical conservation laws: the conservation of momentum (Newton's laws of motion), the conservation of energy, and the conservation of mass, including the total mass of water in any form. Also included is a thermodynamic equation of state – for most practical purposes, the atmosphere is well-approximated as an ideal gas. The complete set of equations (see e.g. Holton, 1992) describing atmospheric motions can be simplified in different ways, depending on the temporal and spatial scale of atmospheric processes under consideration. By reducing the complexity of the governing equations, it is possible to gain a qualitative understanding of those storm systems associated with a particular scale of motion. Although the real atmosphere makes no such simplifications, the dominance of certain processes at specific scales is simply a result of the integrated dynamics. The fact that the dynamical system describing the atmospheric is substantially nonlinear means that a quantitative treatment of the atmosphere is only possible via computer simulations that are necessarily only approximations to the mathematical equations. These mathematical equations are, in turn, only approximations of the real atmosphere. Nonlinear dynamics is why weather forecasting is so widely recognized as challenging and subject to considerable uncertainty (Lorenz, 1993). Since our understanding of the atmosphere is incomplete and our measurement of atmospheric variables is neither perfectly accurate nor at infinite spatial and temporal resolution, forecasting storm systems is never going to be perfect. This uncertainty is tied to the nonlinearity of the dynamical system and so cannot be circumvented at any time in the foreseeable future.

The vertical structure of the atmosphere is broadly described in terms of layers. From the surface to a height of about 10 km, temperature decreases with height at a rate of roughly $6\,\mathrm{K\,km^{-1}}$. At a height of 10 km, the pressure has fallen to about 30% of the surface value. This layer from the surface to roughly 10 km is called the *troposphere*, the actual height of which varies, being generally lowest in the polar regions and deepest near the Equator. Above the troposphere is the *stratosphere*, within which the temperature generally increases with height up to around 40–50 km, where the pressure has fallen to less than one percent of the surface value. The boundary between the troposphere and the stratosphere is called the *tropopause*. Generally, most of what we call "weather" is confined within the troposphere and lower portions of the stratosphere. The depth of the troposphere compared to size of the Earth is comparable to the thickness of the skin on an apple. Thus, the complexities of storm systems are mostly confined to a very thin layer, and the fact that the atmosphere is so thin relative to the planetary scale is an important factor in the dynamics of large-scale storms.

LARGE-SCALE STORM SYSTEMS

Seen from space, the cloud patterns (Figure 1) show that cloudy regions in middle latitudes are broadly associated with relatively large disturbances that rotate cyclonically – that is, counterclockwise in the northern hemisphere and clockwise in the southern hemisphere. The main large-scale weather systems of middle latitudes are called *extratropical cyclones*, to distinguish them from tropical cyclones, which have very different dynamics. Outside of the tropics, the dominant contributions to the dynamics of the atmosphere are: gravity, planetary rotation (manifested by the so-called *Coriolis Force* that varies from a maximum at the poles to zero at the Equator), the sphericity of the Earth, the character of the topography (notably, the mountains and oceans), and the unequal distribution of temperature resulting from solar heating. Because storm systems of large scale within the troposphere are very flat, the airflow in such systems, therefore, is predominantly horizontal and the vertical motions are so weak (on the order of a few $\mathrm{cm\,s^{-1}}$) as to defy accurate routine measurement.

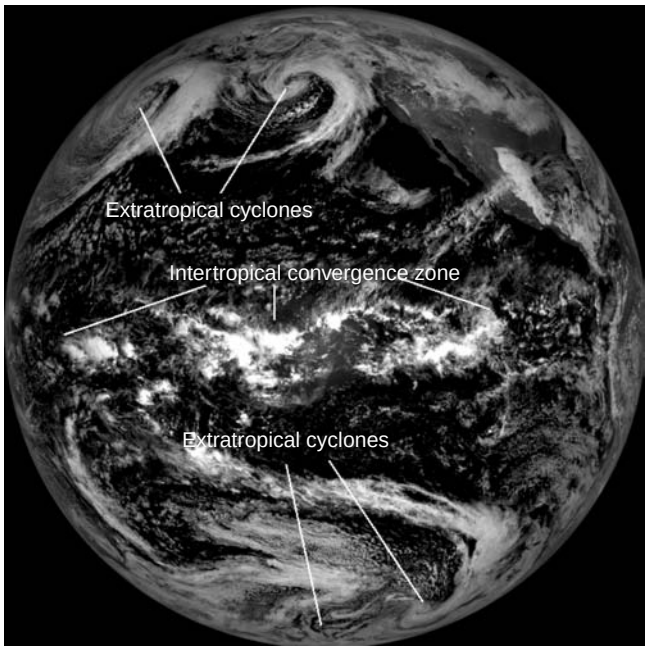


Figure 1 Full-disk image of the earth from the geostationary satellite GOES-12 on 11 March, 1997, showing two major extratropical cyclones in the northern hemisphere's Pacific Ocean: one approaching the west coast of North America, and another immediately behind it, in the central Pacific. Frontal cloud bands spiral inward toward their centers in a counterclockwise rotation. Others can also be seen rotating the opposite way in the southern hemisphere. A bright band of thunderstorm clusters marks the intertropical convergence zone between the northern and southern hemispheric circulations (NOAA image)

Broadly speaking, the fact that the Earth is a sphere means that incoming solar radiation per unit area is always largest in the tropics, decreasing as one moves poleward. This radiational imbalance is the key factor in large-scale weather of midlatitudes (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**). The tilt of the Earth's axis, combined with the rotation around the Sun, produces the seasonal changes poleward from the tropics. Within the tropics, the effect of seasonality is much reduced compared to middle and polar latitudes. Hence, although the Earth is always warmer in the tropics than near the poles, the temperature difference between the poles and the tropics changes with the season, being at its maximum in late winter and its minimum in late summer. Cyclonic storms operate on the energy available as a result of the temperature contrast between the poles and the tropics. The extratropical cyclones are, therefore, most intense during the cool season and at their weakest in the summer. A cyclonic disturbance transports cold air from the polar regions toward the tropics, and warm air from the tropics toward the poles, thereby acting to reduce the horizontal temperature contrast.

Another important factor in the temperature distribution is the difference between oceans and land surfaces. Water has a much higher specific heat than that of land, so a given amount of incoming solar radiation changes the temperature of water much less than that of a comparable land surface. This damps the magnitude of the diurnal and seasonal temperature cycles over the oceans compared to that over land, and creates *land/sea breezes* on a daily timescale and *monsoons* on a seasonal timescale. Moreover, the ocean basins include large-scale currents (such as the Gulf Stream) of their own that modulate the structure of the overlying air and, in turn, are influenced by the airflow over the oceans. The oceans and the atmosphere together are a coupled dynamical system.

The existence of a north-south (or *meridional*) temperature gradient results in, among other things, a band of strong westerly winds at a height of about 10 km (near the tropopause) in midlatitudes called the *jet stream*. The meridional temperature contrast between the Equator and the poles is not evenly distributed but tends to be concentrated in a relatively narrow zone called the *Polar Front*. As a result of all the complicating factors already mentioned, the Polar Front itself varies in location and intensity from day to day and generally migrates poleward in summer and equatorward in winter. The overlying jet stream is in fact tied dynamically to the strength and location of the Polar Front.

On large scales, it turns out that extratropical cyclones are the size they are because disturbances of that size are most efficient at transporting heat from the equatorial regions poleward (and, equivalently, transporting cold air from the polar regions equatorward). The leading edges of cold air masses traveling equatorward are *cold fronts*, whereas the leading edges of warm air masses traveling poleward are *warm fronts*. Extratropical cyclones also provide some *vertical* transport of heat, as warm air involved in the storm typically ascends, whereas the cold air descends. The effect of the vertical heat transport is to carry the excess heat from solar radiation at the surface upward. The weak vertical motions on this scale are efficient at vertical heat transport only because of the large size of the air masses involved. On the average, it only takes about 3 to 6 extratropical cyclones per hemisphere to maintain the observed mean hemispheric thermal structure in the face of the continuing unequal solar heating (Palmén and Newton, 1969). Without extratropical cyclones, the tropics would become much warmer than they now are, while the polar regions would cool still more, perhaps to the point where the habitable portion of the Earth would be confined to a narrow strip in middle latitudes.

Generally speaking, extratropical cyclones have a life cycle (Figure 2) that includes a time of development, during which the increasing kinetic energy associated the winds of the storm is drawn from the potential energy available from

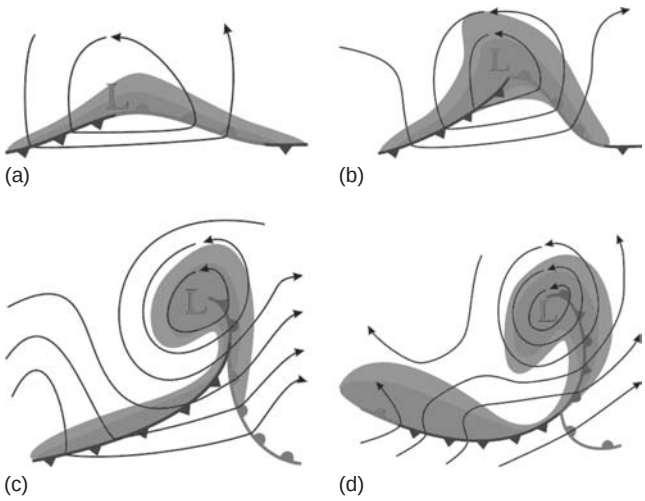


Figure 2 Schematic evolution of an extratropical cyclone (the center of low pressure is indicated by the letter ''L'') in the northern hemisphere, showing the development of a wave on the polar front (a), the amplification of that wave (b), the mature phase of the cyclone, which is beginning to ''occlude'' (c), and the beginning of the dissipation of the cyclone (d), as the occlusion process proceeds. Cold fronts are heavy lines with triangles, warm fronts are heavy lines with half circles, and occluded fronts are heavy lines with alternating triangles and half circles; darker shading indicates regions of precipitation and lighter shading indicates clouds; the black lines with arrows indicate the surface pressure with winds roughly parallel to the pressure contours (Schematic drawings provided courtesy of NOAA). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the unequal distribution of temperature. The life cycle of an extratropical cyclone unfolds over several days. Given that these storms generally move from west to east in middle latitudes, owing to the generally westerly flow in which they are embedded, a new extratropical cyclone passes through a given region on the order of every few days. This results in a cycle whereby locations along the track of such storms experience warming as the extratropical cyclone approaches and cooling as the storm passes. The extratropical storm track is generally along the average location of the Polar Front, which can sometimes persist in the same area for many weeks, or it can shift as the overall airflow pattern changes.

When an extratropical cyclone develops, its horizontal and vertical wind speeds increase – the kinetic energy of the cyclone increases as it draws down the available potential energy. The airflow tends to organize itself in "conveyor belts" (see Browning, 1985) where the ascending currents become cloudy and the descending currents tend to be cloud-free, resulting in the characteristic cloud patterns associated with extratropical cyclones (Figure 1). The development of the storm also increases its overall rotation rate, drawing the clouds into spiral patterns.

Generally, the air to the east of an extratropical storm is rising, as well as moving poleward. If this poleward-moving warm air is also moist at low levels because at some point in the past it has passed over a moisture source (e.g. the warm waters of the Gulf of Mexico), then the likelihood of precipitation is relatively large. On the other hand, if the source of the warm air is from a dry region (like the Sahara Desert), then the likelihood of precipitation is relatively small. Conversely, the equatorward-moving cold air to the west of an extratropical cyclone is usually sinking, so the weather tends to be relatively fair. Depending on the local topography, various exceptions to these tendencies can be encountered.

It is noteworthy that the northern hemisphere has several large mountain ranges and other regions of relatively high terrain, and also has most of the land mass on the planet. On the other hand, the southern hemisphere has a much higher percentage of its area covered by oceans, with only one major mountain range, the Andes Mountains of South America. This asymmetry is also reflected in the complexity of the pattern of cyclones in their respective hemispheres: northern hemisphere weather patterns are far more complex, on average, than those in the southern hemisphere, where westerly flow (also called *zonal* flow) predominates in upper levels. The presence of complex topography also influences the distribution of weather within passing extratropical cyclones. Where airflows toward rising terrain, it is forced upward, which can cause precipitation if that ascending air is also moist (*see* **Chapter 30, Topographic Effects on Precipitation, Volume 1**). Thus, the western slopes of the Americas, with their coastal mountain ranges facing predominantly westerly winds from the Pacific Ocean, usually have abundant rainfall. When airflows downslope, it tends to dry out – deserts or semiarid regions are generally found to the east of the North American mountain ranges, for instance.

MESOSCALE STORM SYSTEMS

The distribution of precipitation within an extratropical cyclone (as illustrated schematically in Figure 2) can vary substantially from one situation to the next, depending on the availability of water vapor in the air currents, and from time to time within the life cycle of the storm system. Embedded within an extratropical cyclone are smaller storm systems that occur on an intermediate scale, often referred to as *mesoscale* storm systems. The horizontal scale of such storm systems ranges from about 100 km to about 1000 km and the timescale ranges from a few hours to a few days. The specific conditions that favor the development of mesoscale storms within an extratropical cyclone are modulated by the details of structure and evolution for that particular large-scale system. Whereas large-scale meteorology is dominated by extratropical cyclones, mesoscale storm

systems include a larger variety of processes than observed at large scales in midlatitudes. Hence, it is difficult to make broad generalizations about mesoscale storm systems. Although there are tendencies for certain weather patterns associated with extratropical cyclones (Figure 2), exceptions to those tendencies are common, because the distribution of weather within an extratropical cyclone depends on many factors (*see* **Chapter 26, Weather Patterns and Weather Types, Volume 1**). Extratropical cyclones are always present in midlatitudes, because there is always the temperature difference between the poles and the equator to drive them, but the smaller scale storm systems created within them are only present intermittently. Mesoscale storms themselves generally act to remove whatever large-scale conditions caused them in the first place. These mesoscale storm systems are complex in their variety, since on this scale, virtually no dynamical factor is always negligible. Mesoscale weather is therefore the most complicated and, therefore, the least understood – see Houze (1993) for a discussion of clouds and precipitation systems across a range of scales. Another factor complicating the understanding of mesoscale processes is the lack of quantitative observations of the needed resolution. Mesoscale storm systems can be broken down in two broad classes: (i) those tied to some topographic feature such as large lakes, coastal areas, orographic features, and so on and (ii) those resulting from inherently mesoscale internal atmospheric mechanisms (Emanuel, 1986).

1. *Topographically driven mesoscale storm systems*: Some examples of mesoscale storm systems associated with topography are lake-effect snowstorms, upslope precipitation, mountain precipitation systems, sea/land breeze systems, and so on. A critical issue in such storms is their intimate connection to processes on larger scales. For example, if we consider lake-effect snow storms (Figure 3), the occurrence and location of the snow depends very much on the wind direction of the large-scale flow relative to the lake in question. These mesoscale snow events arise as cold airflows across warm water, becoming moist and unstable at low levels as a result of sensible and latent heat flux from the water. As a large-scale system moves by, the direction of the prevailing low-level winds changes. Change that large-scale flow and the lake-effect snow will cease in one place but may commence in another. Hence, during the passage of an extratropical cyclone, the distribution of lake-effect precipitation will evolve and different areas could receive heavy snowfall on different days. Since each extratropical cyclone is different, the mesoscale details will vary, but large lakes are fixed topographical features, so there are *preferred* areas for lake-effect snow. Exactly which areas will be affected and at what time depends on the detailed structure of the particular extratropical cyclone.

Upslope rain events are another example of a topographically driven mesoscale storm. As moist air is forced upslope, it condenses and forms first clouds and then rain. As with lake-effect snows, the direction of the flow at large scales interacts with the topography to produce important weather. Upslope rain involves thunderstorms when the air flowing upslope is moist and unstable. Since the situations in which upslope thunderstorms form can persist for many hours, the result can be prodigious rainfall rates, up to $200\,\mathrm{mm\,h^{-1}}$, for extended periods

The fact that topographically driven mesoscale storm systems result from an interaction between topographic features and large-scale storms systems make them somewhat easier to predict. The greatest accuracy in weather forecasting is generally associated with the largest scale systems, so that when armed with a detailed knowledge of topography, it can be fairly straightforward to anticipate the general character of the topographically forced mesoscale features

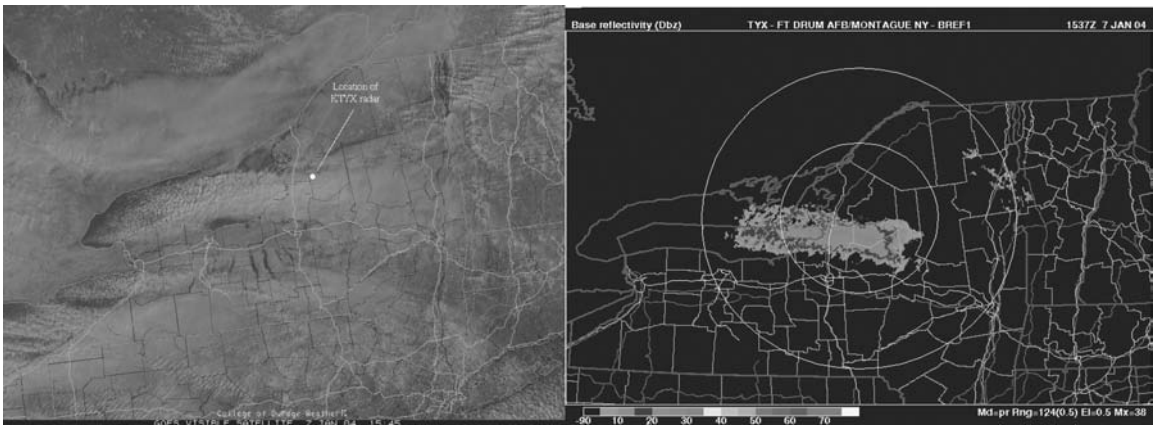


Figure 3 Visible satellite image (a) and radar-observed precipitation (b) associated with a lake-effect snow storm, showing flow from the west-northwest across lake Ontario, with multiple cloud bands producing snow over the eastern part of the lake and on into the state of New York (NOAA images provided by College of Dupage). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

that will occur during the passage of large-scale weather systems. Forecasting such systems is still neither easy nor perfectly accurate, because in forecasting, the small details are inevitably very important. The fact that topographic features are fixed or change only slowly means that the mesoscale storm systems tied to them can be somewhat more predictable than the other category of mesoscale storm systems – those that are driven by instabilities associated with internal atmospheric processes.

2. *Free mesoscale storm systems*: Although extratropical cyclones, the dominant large-scale storm systems, are the size they are as a result of a known scale selection mechanism that maximizes their efficiency at transporting heat meridionally, no such dominant scale selection dynamic process is known for mesoscale storm systems in general. Curiously, the so-called *fronts* that are characteristic of extratropical cyclones have a somewhat ambiguous scale: *along* such a front, they are clearly large-scale processes, extending for 1000 km or so, but perpendicular to a front, the front itself is mesoscale, with characteristic widths for the frontal zone being 10–100 km. The dynamics of fronts are reasonably well known, and frontal zones can include many complicated mesoscale structures that are important for modulating the weather but are only poorly understood. Fronts are an example of a mesoscale process driven by the dynamics of large-scale weather systems.

Mesoscale cyclonic storms called *frontal waves* are often found in association with fronts. These might at times represent the early stages of a developing extratropical cyclone or they might remain within the mesoscale size range, being relatively transient features that, nevertheless, can influence the weather during their mesoscale life cycle. Theory says that such mesoscale perturbations within an extratropical cyclone tend to be confined to near the surface, are shallow, and have shorter life cycles than the extratropical cyclones in which they occur (Gall, 1976). Note that fronts themselves can be important in development of mesoscale regions of ascending air, leading to precipitation. Mesoscale storm systems have vertical motions that can be 10–100 times as strong as those associated with extratropical cyclones – those vertical motions, therefore, are a few tens of cm s^{-1} to perhaps 1 m s^{-1}. Such relatively strong ascent is concentrated in mesoscale regions and contributes to much higher precipitation rates than would be found on the scale of the typical extratropical cyclone. In some regions, at certain times of the year, these mesoscale storm systems produce heavy rain and snow falls, such as within the East China Sea region in the wintertime, during the so-called *Baiu Front* season (Ninomiya *et al.*, 1988).

Although individual thunderstorms are small enough to be considered "small-scale" weather systems and, therefore, are considered in the next section, under certain circumstances (described in Maddox, 1983), many individual thunderstorms become organized into what are called *mesoscale convective systems* (MCSs) (see Fritsch and Forbes (2001) for more details). In many cases, the thunderstorms are linked together into lines of individual storms that interact strongly with each other to produce a mesoscale system, examples of which is shown in Figure 4. Such systems can persist for many hours, and sometimes even for days. They can be associated with very heavy precipitation as well as severe convective weather (discussed in the next section). MCSs arise in a variety of ways, but some of the largest and most persistent examples, given the special name of *mesoscale convective complex* (MCC), appear to have a

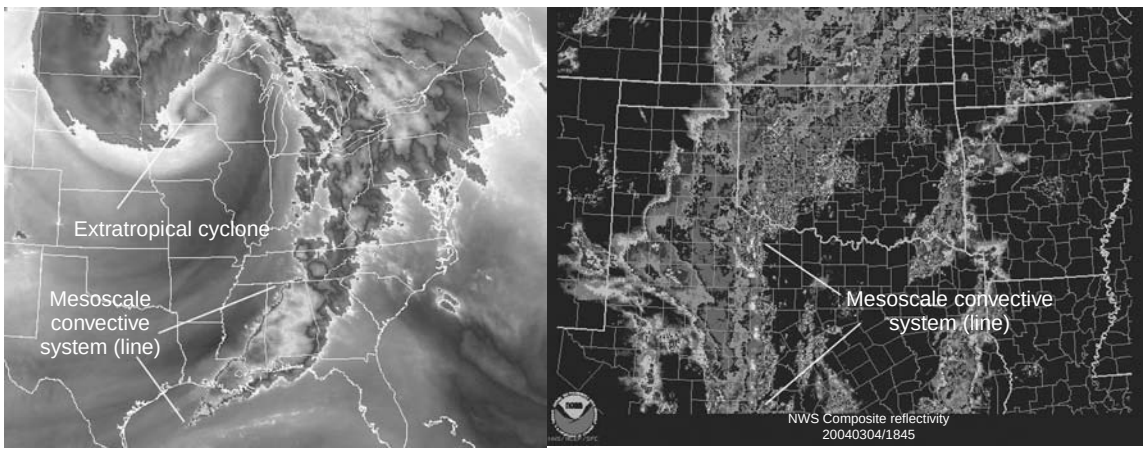


Figure 4 A mesoscale convective system (MCS) along the southern coast of the United States as seen in an enhanced image from a geostationary satellite (a – NCAR image, used by permission) and a radar image of a similar MCS-associated line of thunderstorms (b – NOAA image). Note the size of the MCS affecting the states along the southern coast, compared to the extratropical cyclone system that fills the image, centered west of the Great Lakes. Radar reveals the locations of the strong convective cells (in white) that are powering the MCS, whereas satellite images show the high, cold cloud tops near the tropopause that cover the whole convective system

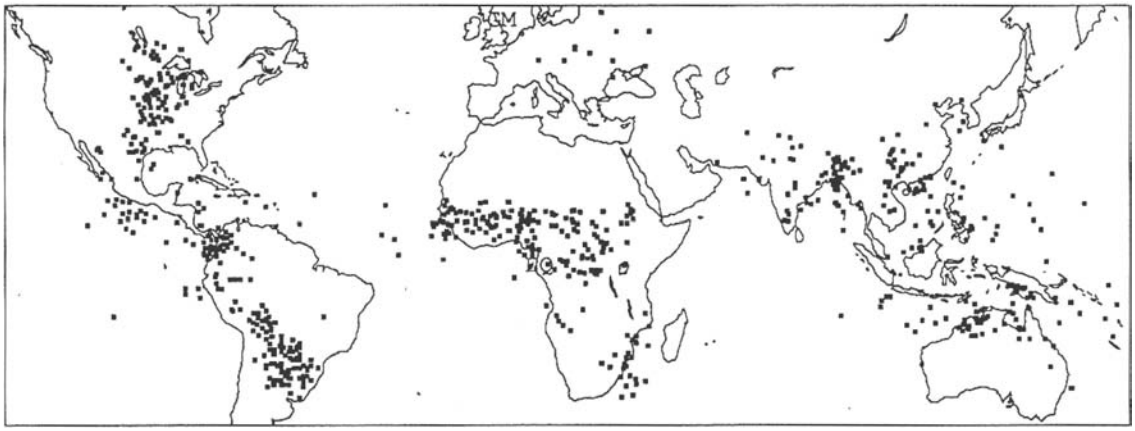


Figure 5 The global distribution of mesoscale convective complexes (MCCs), based on satellite imagery. Small squares indicate location of MCC at time of maximum extent; adapted from Laing and Fritsch (1997)

preference for certain regions of the world (Figure 5), which suggests that their occurrence might be linked to topographic effects. Note that many MCCs are observed in the tropics, as well as in middle latitudes.

Mesoscale features are often embedded within extratropical cyclones at any time of the year, including the winter. In addition to lake-effect snows, there can be mesoscale regions of intense snowfall that travel along with the extratropical cyclone, producing swaths of heavy snow (and/or freezing precipitation) along their tracks. Such bands of winter precipitation are typically on the order of a few hundred km in width (Figure 6), and so represent the track of traveling mesoscale winter precipitation maxima that are associated with processes within a much larger extratropical cyclone.

Within the wintertime polar airstreams often can be found the so-called *polar lows* (Rasmussen and Turner, 2003). These are mesoscale storm systems that take on two rather different forms: when occurring over the ice-free ocean waters, some have been found to have many characteristics in common with tropical cyclones, including relatively weak horizontal temperature contrasts, warm cores, and occasionally even cloud-free "eyes", despite their occurrence at high latitudes. Other types of polar lows are clearly small cousins to the extratropical cyclone, forming in association with strong horizontal temperature contrasts, sometimes associated with topographic features. Both can produce intense snowfalls in association with localized high winds.

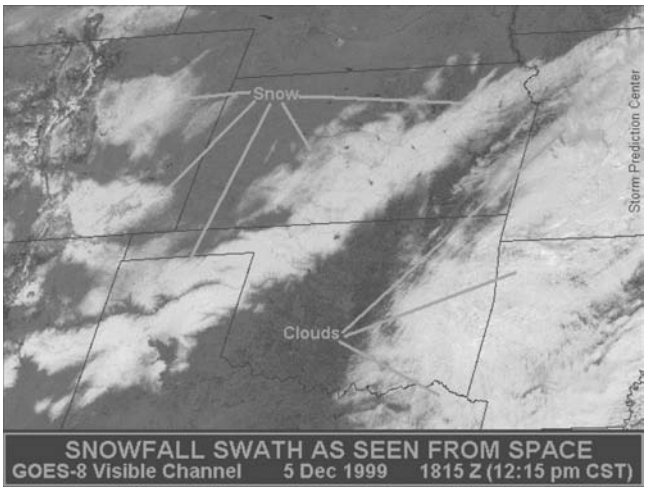


Figure 6 Satellite image from 05 December, 1999, 1815 UTC, showing a snow band of mesoscale width (roughly 100–150 km), from southeast New Mexico, across the Texas Panhandle, Oklahoma, and Kansas. The Canadian River valley can be seen as the dark line within the snow band in the Texas Panhandle, as can several large man-made lakes (the dark spots) within the snow band in Oklahoma and Kansas (NOAA Image provided by the Storm Prediction Center). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

SMALL-SCALE STORM SYSTEMS

As the scale of atmospheric phenomena decreases below that of the mesoscale, it again becomes possible to make simplifying approximations. For example, the curvature and rotation of the Earth can be neglected for many small-scale phenomena. Events that might plausibly be called storms on this scale are primarily *thunderstorms*. Thunderstorms arise when the input of latent and sensible heat at low levels cannot be moderated fast enough by processes on larger scales (Doswell, 2001). Thus, the occurrence of thunderstorms and the weather they produce (strong winds, hail, tornadoes, extreme rainfall rates, and lightning strikes) is associated mostly with land surfaces (Figure 7), as illustrated by the global distribution of lightning. Thunderstorms typically develop during the daytime within the warm season.

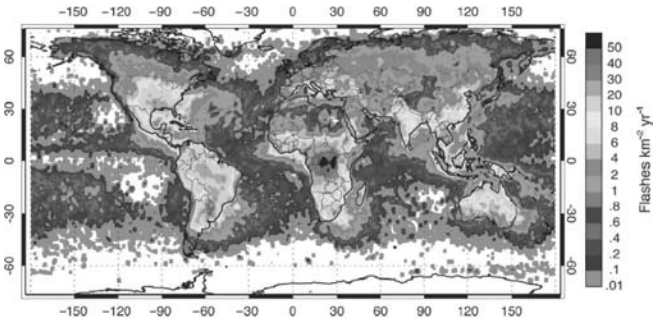


Figure 7 Global lightning flash distribution (from Christian, H.J., R.J. Blakeslee, D.J. Boccippio, W.L. Boeck, D.E. Buechler, K.T. Driscoll, S.J. Goodman, J.M. Hall, W.J. Koshak, D.M. Mach, and M. F. Stewart, 2003: Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. *J. Geophys. Res.*, 108 (D1), 4005–4019. Reproduced by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Exceptions to this can be found, owing to topographic details, as in the relatively high thunderstorm frequencies over the warm waters of the Gulf Stream, east of the United States. When the low levels are warmed and the air contains sufficient moisture, the atmosphere is said to become gravitationally unstable – under such circumstances, plumes of heated air rising from near the surface reach condensation, becoming *towering cumulus* clouds (Figure 8a). These evolve rapidly into mature thunderstorm clouds, called *cumulonimbus* (Figure 8b). After maturity, ordinary thunderstorms dissipate (Figure 8c). The entire life cycle of such a prototypical *thunderstorm cell* is roughly 20–30 min (Byers and Braham, 1949), and most thunderstorms include a sequence of such clouds forming, moving through their life cycles, and dissipating. Thus, they are termed *multicell* thunderstorms and they can be rather ordinary, or they can become severe thunderstorms. The criteria for storms being called *severe* are generally arbitrary; in the United States, if a thunderstorm produces winds of 50 knots ($25\,m\,s^{-1}$) or stronger, hailstones of diameter 3/4 inch (2 cm) or larger, or a tornado, it is deemed a severe thunderstorm (Galway, 1989). Heavy rainfall is not considered officially severe in the United States.

The physical process driving thunderstorms is buoyancy and that buoyancy is strongly dependent on latent heat release. Thus, for thunderstorms to occur, the heat content at low levels must be high relative to that in middle and upper levels (instability), some process to cause air from low levels to rise to a height where it becomes buoyant is needed (lift), and there must be sufficient moisture in the ascending air to maintain the buoyancy. Without any one of these three ingredients (moisture, instability, and lift), thunderstorms cannot form. Thunderstorms act to reduce the instability by redistributing the heat (sensible and latent) from low

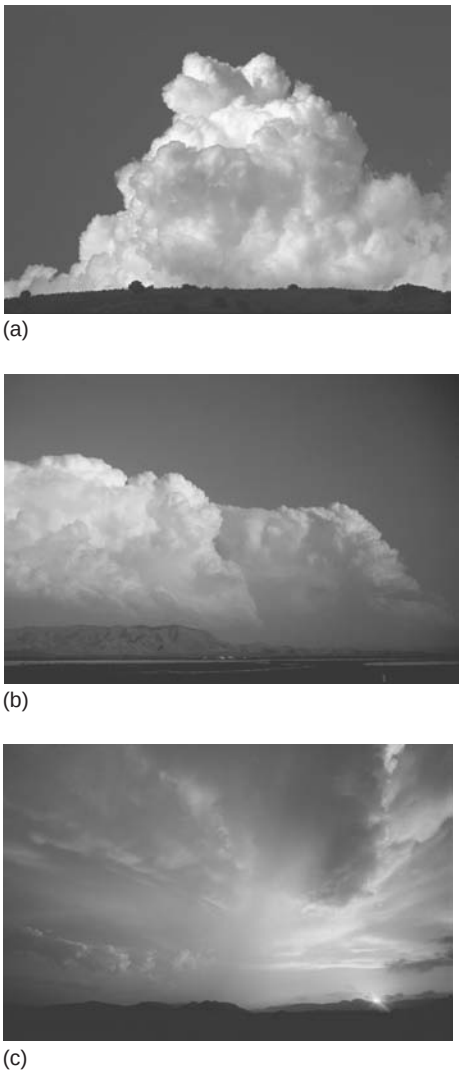


Figure 8 Life cycle stages of ordinary thunderstorms: (a) towering cumulus stage, (b) mature cumulonimbus stage, and (c) dissipating stage (Photographs © C. Doswell, used by permission). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

levels upward. Once they have accomplished the needed heat redistribution, then thunderstorm activity ceases.

Most thunderstorms are ordinary, in the sense that they do not produce phenomena that meet the criteria to be called *severe thunderstorms*. Perhaps a few percent of all thunderstorms become severe, with most of them only barely meeting the criteria. Out of all the severe thunderstorms, a small fraction (on the order of a few percent) of them are *supercell* thunderstorms. Almost all supercells (around 95%) produce one form or another of severe weather. A supercell develops under conditions that favor thunderstorms in the presence of strong vertical wind shear. Thunderstorms that develop in strong vertical wind shear transform that vertical wind shear into rotation about a vertical axis, at times visually evident

Figure 9 Tornadic supercell thunderstorm on 3 June, 1999, showing structures characteristic of a rotating storm, in which air in the storm is moving left to right in the foreground, into the photograph on the right-hand edge of the clouds, and right to left in the background, spiraling inward to the tornado, which is partially obscured by precipitation (Photograph © C. Doswell, used by permission). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

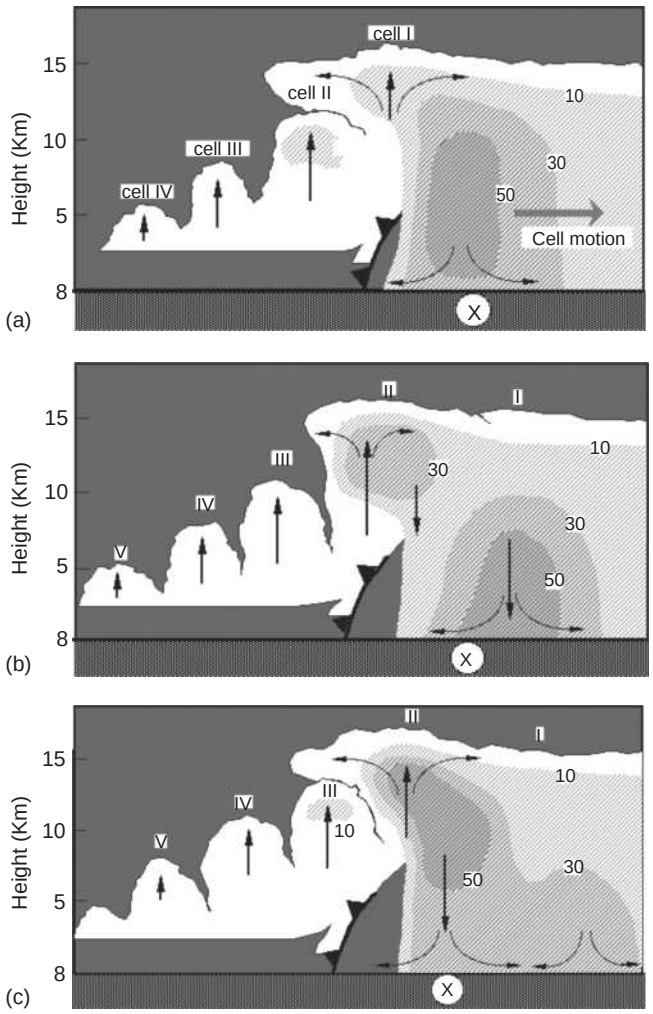


Figure 10 Schematic of the ''training'' effect. (a) At this time, there are four numbered thunderstorm cells in various stages of development. Cell I is mature, with both updrafts and downdrafts, and heavy rain is about to commence at point ''X''. Cells II, III, and IV are still developing, and have only updrafts. Cell II has precipitation forming aloft. The hatched contours are radar reflectivity, in units of dBz, which is related to the rainfall rate. (b) About 15 min later, Cell I's updraft is dissipated, and it is now dominated by downdraft. Heavy rain continues at ''X'' while Cell II is maturing and developing a downdraft. Cells III, IV, and now V are still immature. (c) About 15 more minutes have elapsed. Cell I's rainfall is continuing but it is now nearly dissipated, while Cell II is entering late maturity. It is still raining at ''X'' but now the rainfall is from Cell II, and heavy rain from Cell II is descending from aloft. Now Cell III is developing its first precipitation aloft. Cells IV and V are still immature (Doswell *et al.* (1996). © 1996 American Meteorological Society). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

in a spiral structure to the storm (Figure 9). In the northern hemisphere, this interaction results in counterclockwise rotation concentrated in the so-called *mesocyclone*. (Note that this term is something of a misnomer, since it seems to suggest a mesoscale process, but the horizontal scale of a mesocyclone is on the order of a few kilometers, too small to be properly considered mesoscale.) It is the presence of a deep, persistent mesocyclone that distinguishes supercells from ordinary thunderstorms. Owing to the organizing dynamics of mesocyclones, supercells can become quasi-steady and persist for several hours, much longer than the typical ordinary thunderstorm cell. They also can be responsible for the most violent severe thunderstorm phenomena: families of strong-to-violent tornadoes, hailstones with diameters exceeding 2 inches (5 cm), and wind gusts exceeding 65 knots ($32\,\mathrm{m\,s^{-1}}$). This is the result of dynamical processes that produce the supercell and enhance the vertical motions in such storms beyond that expected from buoyancy effects alone.

Thunderstorms also can become organized into lines of interacting thunderstorm cells, sometimes called *squall lines*. These often are large enough to be considered mesoscale in extent and so are properly termed *MCSs*. However, this tendency for linear organization of storms persists into scales arguably near or below any particular arbitrary threshold separating "mesoscale" from "small-scale". Any time that thunderstorms can become organized, the potential for severe weather increases. Although many heavy rain-producing thunderstorms are not severe by official criteria used in the United States (see above), they can be organized to produce dangerous flash floods under the right

hydrological circumstances. Generally, thunderstorms that produce heavy rainfall are associated with the repeated passage of mature thunderstorm cells over the same general region; this is the so-called *training effect*, illustrated schematically in Figure 10. Training of thunderstorm cells results in localized heavy rainfall amounts and is by far the most common evolution associated with flash-flood producing storms. Supercells often produce large instantaneous rainfall rates owing to their intense updrafts, but they typically do not remain in one place long enough to create high rainfall totals. Nevertheless, when the rainfall rate reaches $200\,\mathrm{mm\,h^{-1}}$, which some supercells have attained, they can produce dangerous urban flooding in as little as 15 min (Smith *et al.*, 2001).

Nearly stationary rainstorms can also develop when airflows upslope, as noted above. Sometimes, such rainstorms occur with little or no lightning, while at other times, upslope rainstorms are associated with considerable lightning and thunder. Whether they are thunderstorms or not, they can produce small-scale regions of very heavy rainfall when the upslope slow persists for many hours.

On rare occasions, some snowstorms are accompanied by lightning and thunder. In such events, the instantaneous snowfall rates can be quite high (instantaneous rates might be as high as $25\,\mathrm{cm\,h^{-1}}$, but such rates are not typically sustained for long). Although intense winter storms such as lake-effect snowstorms and upslope snowstorms are likely to be more properly considered mesoscale events, the peak values associated with such storms might be less than 100 km in spatial scale and so represent small-scale events embedded within a mesoscale storm system.

REFERENCES

Browning K.A. (1985) Conceptual models of precipitation systems. *Meteorological Magazine*, **114**, 293–319.

Byers H.R. and Braham R.R. Jr (1949) *The Thunderstorm*, U.S. Government Printing Office: Washington, 287 pp.

Christian H.J., Blakeslee R.J., Boccippio D.J., Boeck W.L., Buechler D.E., Driscoll K.T., Goodman S.J., Hall J.M., Koshak W.J., Mach D.M. *et al.* (2003) Global frequency and distribution of lightning as observed from space by the optical transient detector. *Journal of Geophysical Research*, **108**(D1), 4005–4019.

Doswell C.A. III (2001) Severe convective storms – an overview. *Severe Convective Storms. Meteorological Monographs*, Vol. 28, No. 50, American Meteorology Society: pp. 126.

Doswell C.A. III, Brooks H.E. and Maddox R.A. (1996) Flash flood forecasting: an ingredients-based methodology. *Weather and Forecasting*, **11**, 560–581.

Emanuel K.A. (1986) Overview and definition of mesoscale meteorology. In *Mesoscale Meteorology and Forecasting*, Ray P.S. (Ed.), American Meteorological Society: pp. 117.

Fritsch J.M. and Forbes G.S. (2001) Mesoscale convective systems. *Severe Convective Storms. Meteorological Monographs*, Vol. 28, No. 50, American Meteorology Society: pp. 323357.

Gall R. (1976) Structural changes of growing baroclinic waves. *Journal of the Atmospheric Sciences*, **33**, 374–390.

Galway J.G. (1989) The evolution of severe thunderstorm criteria with the weather service. *Weather and Forecasting*, **4**, 585–592.

Holton J.R. (1992) *An Introduction to Dynamic Meteorology, Third Edition*, Academic Press: 511 pp.

Houze R.A. (1993) *Cloud Dynamics*, Academic Press: 573 pp.

Laing A.G. and Fritsch J.M. (1997) The global population of mesoscale convective complexes. *Quarterly Journal of the Royal Meteorological Society*, **123**, 389–405.

Lamb D. (2001) Rain production in convective storms. *Severe Convective Storms. Meteorological Monographs*, Vol. 28, No. 50, American Meteorology Society: pp. 299321.

Lorenz E.N. (1993) *The Essence of Chaos*, University of Washington Press: 227 pp.

Maddox R.A. (1983) Large-scale meteorological conditions associated with midlatitude, mesoscale convective complexes. *Monthly Weather Review*, **111**, 1475–1493.

Ninomiya K., Akiyama T. and Ikawa M. (1988) Evolution and fine structure of a long-lived meso-a-scale convective system in a Baiu front zone, Part I: evolution and meso-a-scale characteristics. *Journal of the Meteorological Society of Japan*, **66**, 331–350.

Palmén E. and Newton C.W. (1969) *Atmospheric Circulation Systems*, Academic Press: 603 pp.

Rasmussen E.A. and Turner J. (2003) *Polar Lows: Mesoscale Weather Systems in the Polar Regions*, Cambridge University: 624 pp.

Smith J.A., Baeck M.L., Zhang Y. and Doswell C.A. III (2001) Extreme rainfall and flooding from supercell thunderstorms. *Journal of Hydrometeorology*, **2**, 469–489.

28: Clouds and Precipitation

GEORGE A ISAAC[1] AND JOHN HALLETT[2]

[1]Cloud Physics and Severe Weather Research Division, Meteorological Service of Canada, Toronto, ON, Canada
[2]Division of Atmospheric Sciences, Desert Research Institute, Reno, NV, US

Clouds and the precipitation that comes from them are important elements of the hydrological cycle. Clouds provide a blanket for our Earth, both shielding it from radiation from the Sun and trapping heat escaping from the surface. They also generate precipitation through the condensation-coalescence mechanism, which involves only liquid cloud drops, or through ice initiation leading to large ice particles and eventually snow or rain. Mean annual precipitation amounts reach a maximum near the equator, near 8 mm day^{-1}, and decrease poleward to about 1 mm day^{-1}. A good understanding of both cloud and precipitation processes is very important for climate and weather predictions. This paper outlines some of the most important processes and provides reference material where more detailed information can be obtained.

INTRODUCTION

Earth is covered with clouds as any satellite photo of our planet will show. Figure 1 shows a cloud climatology from the GEWEX Surface Radiation Budget data set (Whitlock *et al.*, 1995). Almost everywhere, the annual mean cloud cover is greater than 50% and in some large areas it is greater than 80%. These clouds literally provide a blanket for our Earth, both shielding it from radiation from the Sun and trapping heat escaping from the surface. Understanding clouds is extremely important for both climate and weather predictions because their presence or absence can strongly affect surface temperatures. Figure 2 shows that for a station in northern Canada, the presence of cloud in the wintertime makes the surface warmer, while it provides a cooling effect in the summertime.

Life cannot exist without water and in most cases we get our water from precipitation. Figure 3 shows the annual mean precipitation rate as compiled by Xie and Arkin (1996, 1997) using gauge observations, satellite estimates and numerical model outputs. Precipitation amounts reach a maximum near the equator, near 8 mm day^{-1}, and decrease poleward to about 1 mm day^{-1}. The higher temperatures in the tropical regions produce strong convection and more precipitation, with cloud base temperatures being greater than 20 °C. It should also be mentioned that the area of the earth in the tropical regions is much greater than that at the poles, which also accounts for a greater amount of precipitation in that region. At the poles, the temperatures are low and the stratiform clouds that exist barely produce precipitation, less than 1 mm per day.

There are many uncertainties in our knowledge about clouds and precipitation. However, much progress has been made and this article will briefly summarize our current knowledge, and point to more comprehensive articles where additional information can be found. Earlier textbooks on cloud physics which provide useful information include those written by Fletcher (1962) and Mason (1971). More recently, a general textbook on cloud physics has been written by Rogers and Yau (1989) and a more detailed book on the microphysics of clouds and precipitation was published by Pruppacher and Klett (1997).

CLOUD FORMATION AND TYPES

Clouds can exist in many forms in the atmosphere. See the World Meteorological Organization International Cloud Atlas (WMO, 1975, 1987), the AMS Glossary (Glickman, 2000) and Scorer (1972) for a full description of cloud types. Cirrus clouds form at temperatures below −40 °C, generally occur above 5 km, and they cover wide areas in a

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

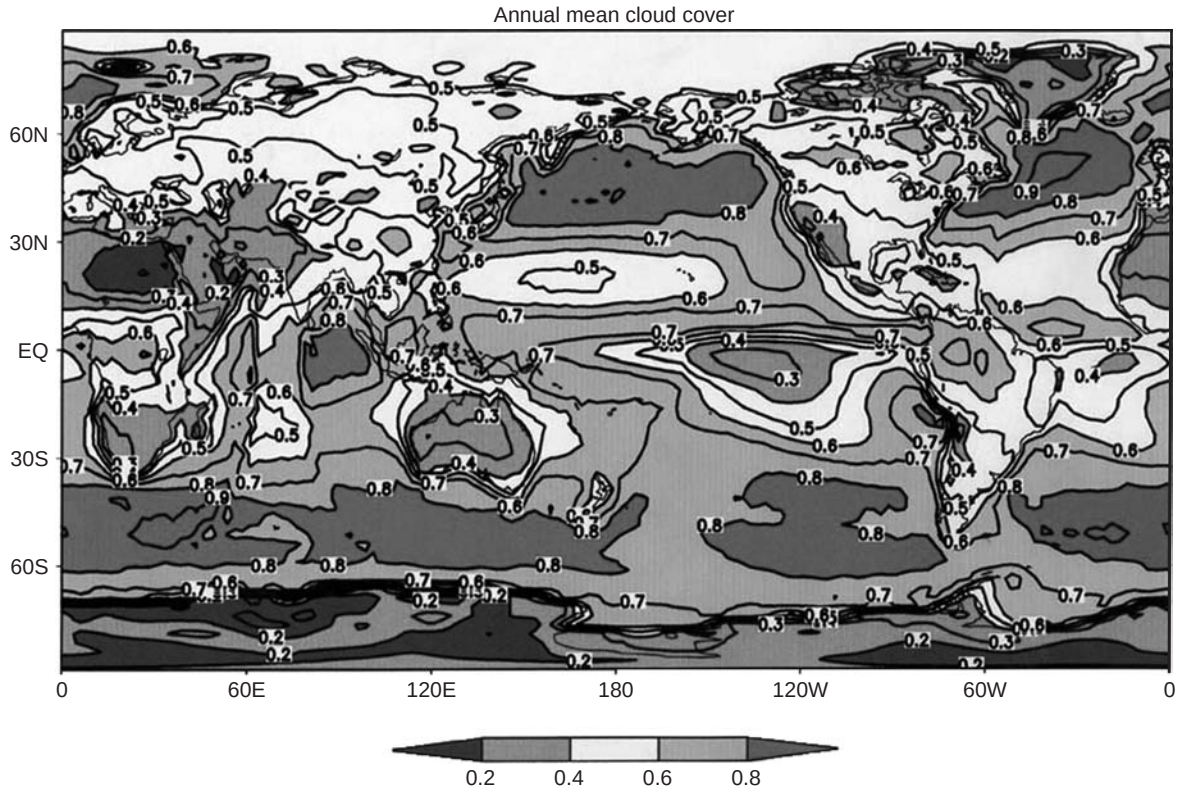


Figure 1 Mean annual cloud cover as seen from satellite as a function of latitude and longitude. From the GEWEX surface radiation budget data set (Whitlock *et al.*, 1995. © 1995 American Meteorological Society). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

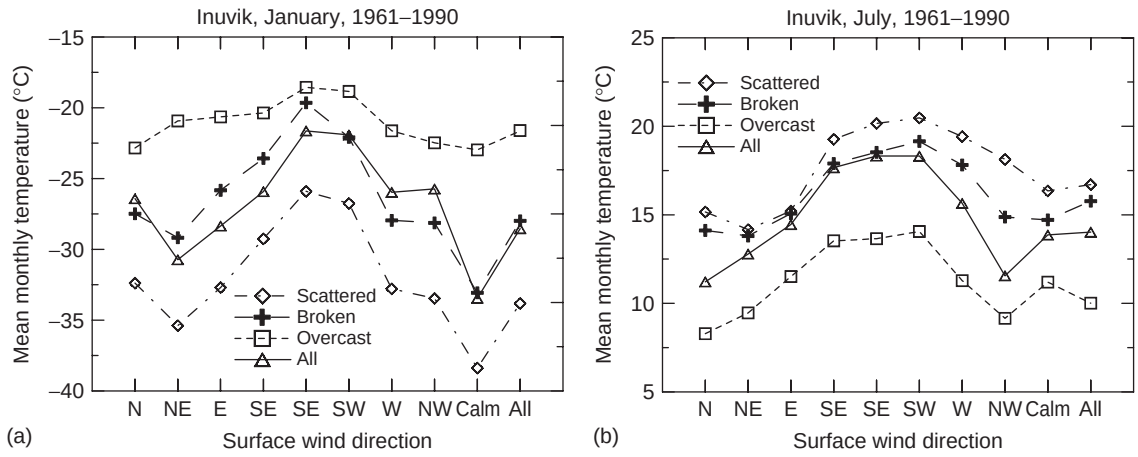


Figure 2 January and July mean monthly temperature at Inuvik, Northwest Territories, Canada, as a function of cloud cover and surface wind direction. ''Scattered'' indicates 0–1 tenth sky, ''broken'' 2–8 tenths, and ''overcast'' 9–10 tenths sky coverage (Isaac and Stuart, 1996. © 1996 American Meteorological Society)

sheet form, often with a fibrous aspect (Figure 4a). Below this temperature, ice crystals form even in the absence of insoluble nuclei in cloud drops and a few degrees lower in diluted haze droplets. Cirrus can form from broadscale uplift, in the outflow of thunderstorms, or even from the vapor trails caused by high-flying jet aircraft (Figure 4b). These clouds generally do not create precipitation that reaches the ground and they are primarily formed of ice crystals. However, they can start to precipitate (Figure 4c) and the falling ice crystals can "seed" lower layers of

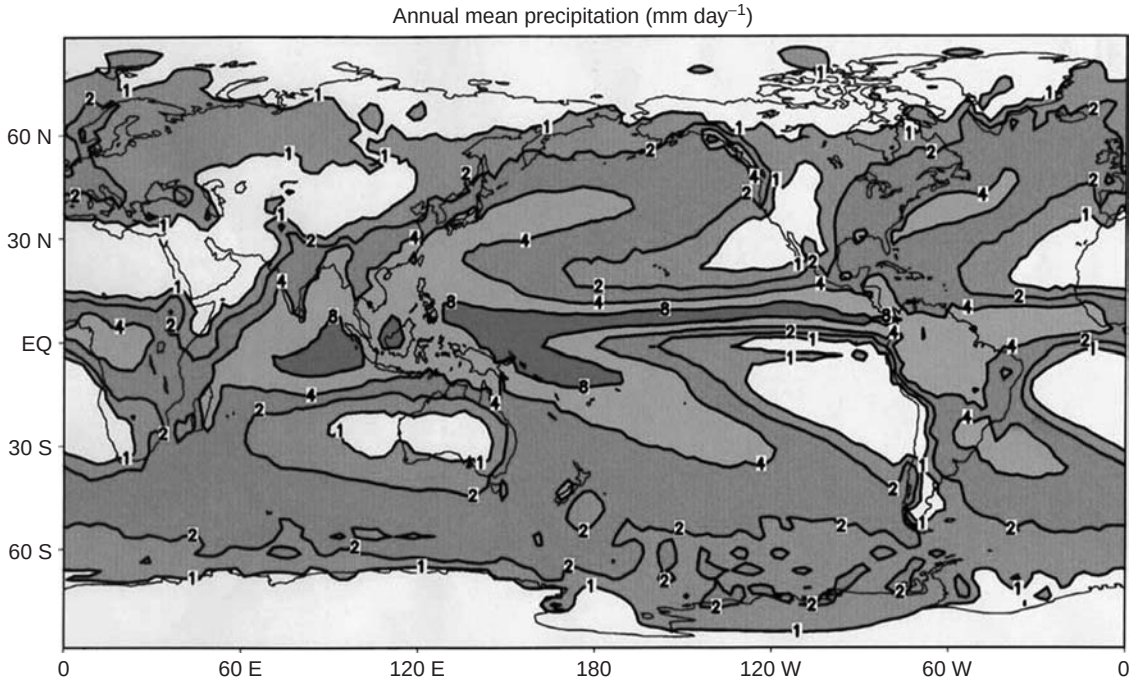


Figure 3 Annual mean precipitation (mm day^{-1}) as compiled by Xie and Arkin (1996, 1997) using gauge observations, satellite estimates and numerical model outputs (Xie and Arkin, 1996. © 1996 American Meteorological Society). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

cloud and initiate precipitation. Mid-level clouds such as altostratus or altocumulus are also formed by broadscale lifting, often produced by frontal systems. They can be composed of either ice crystals or liquid water, but they do not account for much of our precipitation. Lower-level cloud types such as stratus, and stratocumulus (Figure 4d, e) are associated with precipitation, either snow or rain. In midlatitudes, especially in winter, these cloud types account for most of our precipitation. Cumulus clouds (Figure 4f) and especially thunderstorms (Figure 4e) create most of our precipitation in the summer at mid- and lower-latitudes.

Clouds are generally formed in circulation around low pressure cyclonic storm systems, with precipitation taking place at fronts where higher temperature, moist air if lifted above cooler air. Mesoscale complexes and hurricanes form towards the tropics and easterly waves form convergence zones for precipitation in low-latitude tropical regions.

WATER PHASES, LATENT HEAT

Water in the atmosphere exists in three phases: vapor, solid, and liquid. Cloud and precipitation formation involve transforming vapor into the other two phases. A vapor to liquid phase change occurs through condensation and the reverse process involves evaporation. A vapor to solid phase change is called *deposition*, while the reverse is called *sublimation*. The liquid to solid phase change involves the freezing process, while melting occurs when solid water changes to liquid. Phase changes occur with the corresponding release or uptake of latent heat. For example, in order for water to change from the liquid to vapor state, latent heat is required to break the hydrogen bonds between water molecules in the liquid. This is called *the latent heat of vaporization*. Similarly, *the latent heat of fusion* is required to change from the solid to liquid state. The *latent heat of sublimation* is released when vapor changes directly into ice. Normally, phase changes from liquid or solid to vapor occur when the air is subsaturated with respect to liquid water or ice, and the reverse happens when the air is supersaturated.

The Clausius–Clapeyron equation, one of the most important in cloud physics, relates the saturation vapor pressure with respect to water (e_s) or ice (e_i) to the latent heat of either vaporization (L_v) and or sublimation (L_s) to temperature (T). For the saturation vapor pressure over water, the equation may be written:

$$\frac{de_s}{dT} = \frac{L_v e_s}{R_v T}$$

where R_v is the gas constant for water vapor (461.5 J kg^{-1} K^{-1}).

Table 1 shows the saturation vapor pressure over water and ice, and the latent heats of vaporization (condensation)

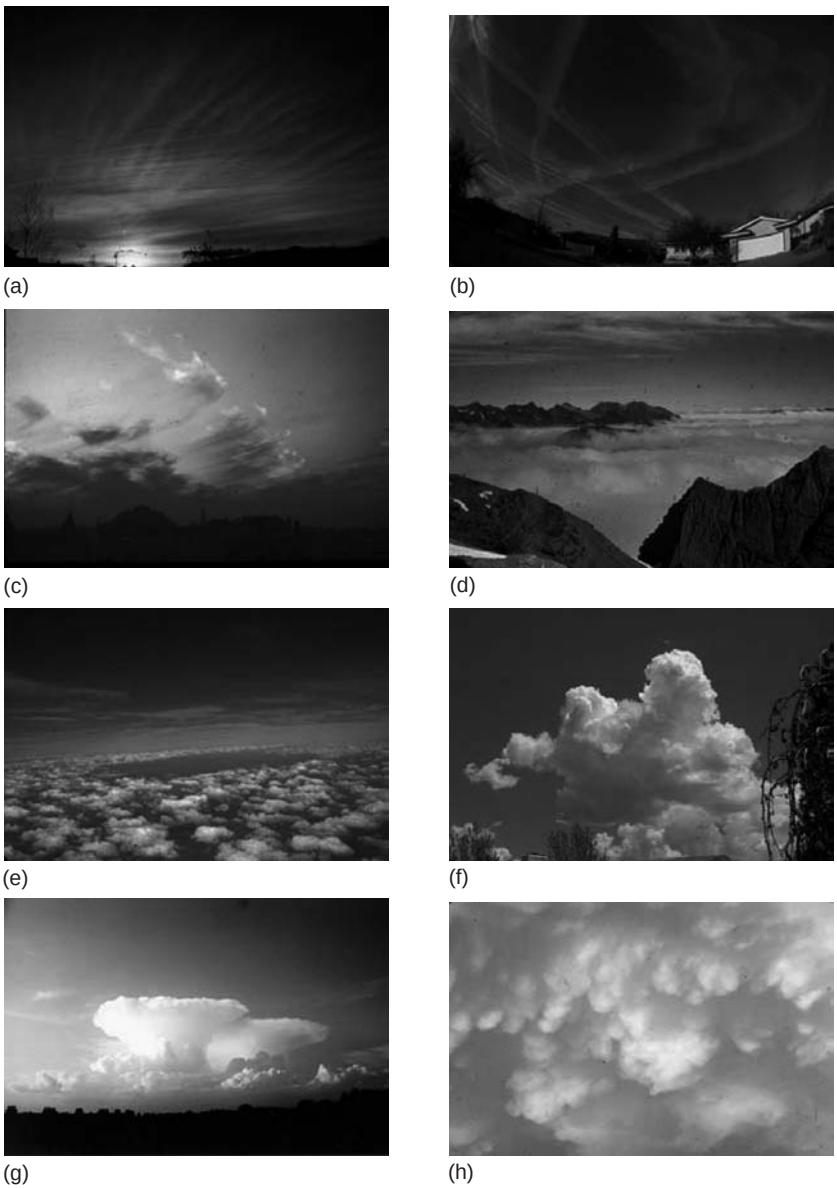


Figure 4 Some examples of cloud types such as: (a) cirrus cloud, (b) cirrus clouds produced by high-flying aircraft (contrails), (c) cirrus clouds with falling ice crystal streaks, (d) stratus clouds below an inversion, (e) stratocumulus deck, (f) cumulus cloud, (g) thunderstorm, and (h) mammatus. For full definitions see American Meteorological Society (AMS) Glossary (Glickman, 2000), World Meteorological Organization Cloud Atlas (WMO, 1975, 1987), and Scorer (1972). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs (© John Hallett)

and sublimation as a function of temperature. The latent heat of fusion, which is released when ice changes to liquid, is the difference between L_v and L_s. The high values of the latent heat of vaporization of ice and water, compared with the latent heat of fusion (melting/freezing) show that water molecules are mostly bonded in both liquid and solid, and the bonding only changes by about 12% on melting. It is clear from Table 1 that warm air can hold much more water than cold air. For example, at $-10\,^{\circ}\mathrm{C}$, $0\,^{\circ}\mathrm{C}$, $+10\,^{\circ}\mathrm{C}$ and $+20\,^{\circ}\mathrm{C}$ and 1000 mb, the saturation mixing ratio over water is approximately 1.8, 3.8, 7.8 and 15.0 g of vapor per kg of air, respectively (List, 1968). This helps explain why there is more precipitation in the Tropics than in the Arctic.

ATMOSPHERIC STABILITY AND CLOUD FORMATION

In order to consider how clouds form in the atmosphere, it is necessary to understand atmospheric stability and the

Table 1 The saturation vapor pressure over water and ice, and the latent heats of vaporization (condensation) and sublimation as a function of temperature (from Rogers and Yau, 1989). Papers by Fukuta and Gramada (2003) and Marti and Mauersberger (1993) are examples of recent works that attempt to measure the saturation vapor pressure over water and ice more accurately

T (°C)	e_s (Pa)	e_i (Pa)	L_v (J g^{-1})	L_s (J g^{-1})
−40	19.05	12.85	2603	2839
−30	51.06	38.02	2575	2839
−20	125.63	103.28	2549	2838
−10	286.57	259.92	2525	2837
0	611.21	611.15	2501	2834
10	1227.94	–	2477	–
20	2338.54	–	2453	–
30	4245.20	–	2430	–
40	7381.27	–	2406	–

normal temperature profile of the atmosphere. Typically, as you go further up in height, the pressure falls and the temperature decreases. For a parcel moving upward, without condensation occurring, its temperature cools down following a dry adiabatic lapse rate of approximately 0.98 °C per 100 m. For a parcel where condensation has commenced, the release of latent heat slows the cooling of the parcel as it ascends and the parcel follows a wet adiabat (0.65 °C per 100 m at 0 °C and 1000 mb) approaching the dry adiabatic lapse rate at temperatures below −40 °C where the amount of vapor is minimal.

Figure 5 shows atmospheric soundings for two severe weather events. On 3 May 1999, intense tornadoes were generated by severe thunderstorms and killed 36 people near Norman Oklahoma (Thompson and Edwards, 2000; Brooks and Doswell, 2002). The sounding is unstable above about 1.5 km (a saturated parcel would keep rising through buoyancy), showing the potential for convection. The sounding on 9 January 1998 for Gray, Maine, was during an intense and very prolonged freezing rainstorm (DeGaetano, 2000). The atmosphere is stable in this sounding but shows a deep saturated layer with a warm section reaching about +12 °C near 1.5 km with a below freezing surface layer.

WARM RAIN PROCESS

Water generally condenses in the atmosphere around hygroscopic particles known as *cloud condensation nuclei* (CCN) Table 2 shows the relative sizes and fall velocity dependence on diameter of cloud nuclei (CN) which are only active at very high supersaturations, CCN around which most cloud droplets form, and Ultra Giant Nuclei, (UGN) which can form the nuclei of large cloud drops. Effective CCN are sodium chloride containing particles formed in sea spray or ammonium sulfate particles from anthropogenic sources, possibly also contaminated further by equally deliquescent organic materials. These CCN deliquesce at conditions less than about 80% relative humidity, or when the air is subsaturated with respect to water, resulting in haze. More typically in clouds, at supersaturations of less than 1%, small droplets are quickly formed around these CCN reaching sizes of 5 to 15 μm. The actual number activated depends on the chemical composition of the CCN particles. More efficient CCN, as are typically found in maritime environments, activate at lower supersaturations. Larger updrafts tend to create higher supersaturations, with more CCN activated, and thus higher droplet number concentrations. Typically, maritime clouds have lower droplet

Table 2 Function relationships for terminal velocity for different particle types. As a first approximation, spheres in the range from 1 to about 100 μm follow Stokes law, with velocity proportional to diameter squared. Smaller and large particles follow a linear relation; particles larger than 0.5 cm follow a square root relationship (assuming a constant drag coefficient) up to 10 cm hail. Snow flakes and larger ice crystals follow Stokes law below about 100 μm but have constant fall velocity with size beyond a few mm. Brownian motion gives an inverse relation for molecular cluster size pollution particles, and inverse square root relation for particles greater than 0.02 μm. The fall velocity and Brownian displacement random direction velocity are comparable for particles of diameter near a fraction of a micron, depending on particle density.

Diameter (d)		Terminal velocity or displacement dependence	Regime
cm	Rain, hail	$\sqrt{d}$	Constant drag coefficient
500 μm	Small rain, drizzle	d	
10 μm	Cloud droplets	d^2	Stokes
10 μm – 1 μm	UGN	d^2	Stokes–Cunningham
1 μm – 0.1 μm	CCN		
10^{-2} μm	CN	$\frac{1}{\sqrt{d}}$	Brownian motion
10^{-3} μm	Molecular cluster	$\frac{1}{d}$	Stokes/Cunningham/self diffusion

Note: CN – cloud nuclei, CCN – cloud condensation nuclei, UGN – Ultra Giant Nuclei

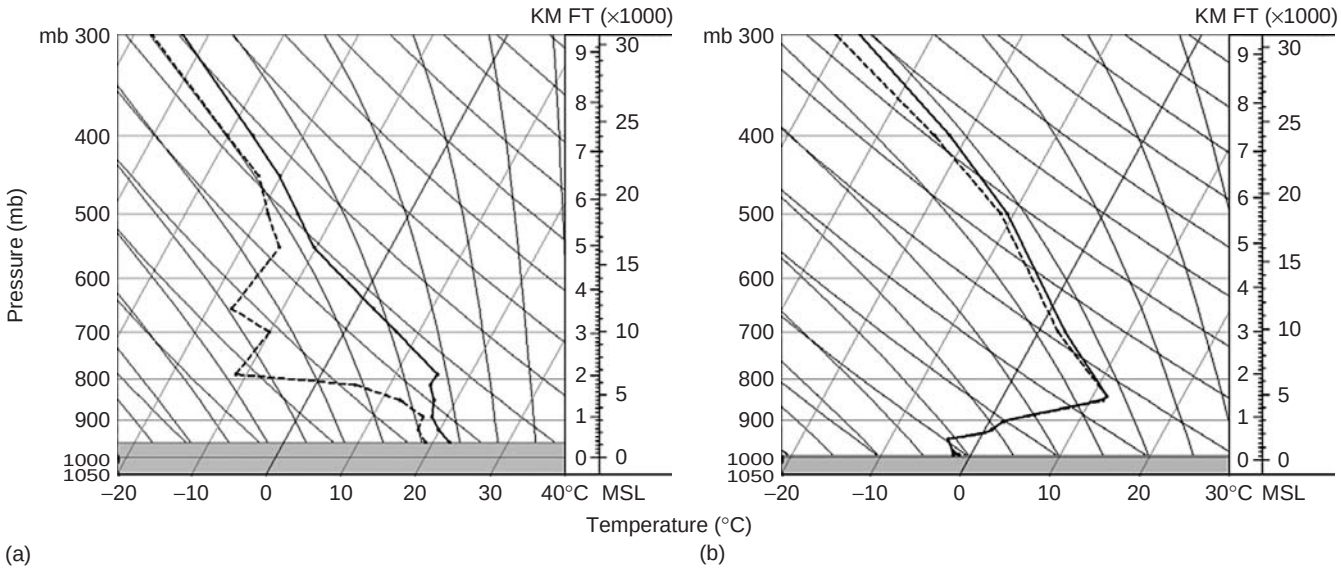


Figure 5 Pressure versus temperature atmospheric soundings for 4 May 1999 at 00 UTC for Norman, Oklahoma (a) and 9 January 1998 at 00 UTC for Gray, Maine (b). The horizontal lines represent pressure, the lines running from the bottom up to the right represent constant temperature, the curved lines running upwards towards the left represent wet adiabat lines, while the approximately straight lines running up and to the left represent dry adiabats. The dark line on the right represents the temperature profile in the atmosphere while the dashed line on the left represents the dewpoint. The altitudes below the station elevation are shaded. As an example, the 9 January sounding has a surface temperature near −2 °C, and it reaches a maximum temperature at +12 °C near 850 mb (1.5 km), and then gets colder than 0 °C at about 750 mb (3.5 km). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

number concentrations and larger droplet sizes than continental clouds (see Table 3).

The cloud droplets that form in a cloud grow by condensation in the presence of a persistent updraft that generates a small supersaturation. As they grow, the droplets also begin to collide with one another and grow by the collision-coalescence mechanism. This mechanism is most efficient when some of the droplets are larger than others (20 to 40 μm). It is more efficient in maritime clouds where for the same liquid-water content, because of the reduced droplet number concentration, the drops tend to be larger (see Table 3). As Rogers and Yau (1989) indicate "the central task of precipitation physics is to explain how raindrops can be created by condensation and coalescence in times as short as 20 min." Rain appears to form in such a time interval in nature. There are considerable uncertainties involved in this "warm rain" process but it is critical that the drop-size distribution widens, so that it permits the coalescence process to work effectively. It will take about 10^5 small cloud droplets to form a raindrop 1 mm in diameter. Drops have to be larger than 500 microns in order to be considered raindrops. Drops between 200 and 500 μm are considered as drizzle (AMS Glossary–Glickman, 2000). Figure 6 shows the relative sizes of cloud aerosols, cloud droplets and raindrops that are involved in the warm rain process.

Although the warm rain process is commonly used to describe the process of condensation followed by coalescence in clouds >0 °C, it also works in supercooled clouds down to at least −20 °C and probably below (e.g. Rasmussen *et al.*, 1995; Cober *et al.*, 1996; Kajikawa *et al.*, 2000; Lawson *et al.*, 2001). Such large supercooled drops forming at cold temperatures are important for aircraft icing processes and precipitation formation in tropical convection.

BERGERON–FINDEISEN OR ICE PRECIPITATION FORMATION PROCESS

Clouds exist in the atmosphere at temperatures colder than 0 °C. Many of these clouds contain liquid water in the supercooled state. When ice is also present, a precipitation formation process can commence whereby ice crystals grow by aggregation of ice particles or riming of cloud droplets. The resulting precipitation can fall to the ground as either snow or rain, depending upon the freezing level. Mixed-phase clouds which contain both liquid and ice occur quite frequently in the atmosphere, as the summary by Korolev *et al.* (2003) has shown. Liquid water typically will not freeze until about −40 °C unless it has some impurities in it known as *ice nuclei* (IN), such as clay mineral particles (Roberts and Hallett, 1968). Although ice melts at 0 °C, it does not form in the

Table 3 Cloud microphysical summaries, using 30 s or 3 km pathlength averages, for maritime and continental cases in terms of static temperature (Ta), droplet number concentration (Nd), total water content (TWC) and median volume diameter (MedVD). The ice crystal concentration is represented as I. As an example, 25% of the maritime droplet concentrations (Nd) were less than 16 cm^{-3}. These data were collected as part of the Canadian Freezing Drizzle Experiment (I and III) and the Alliance Icing Research Study I (from Isaac *et al.*, 2001) and they represent stratiform mid latitude winter clouds. Reproduced from Isaac *et al.*, 2001, © Canadian Aeronautics & Space Institute

	1%	25%	50%	75%	99%
Maritime		Points = 1154[a]	Ta $\leq 0\,^{\circ}C$ I $\leq 1 L^{-1}$	TWC $\geq 0.005\,g\,m^{-3}$	
Ta (°C)	−20.6	−5.8	−4.1	−2.0	0.0
Nd (cm^{-3})	1	16	52	108	406
TWC ($g\,m^{-3}$)	0.01	0.07	0.13	0.20	0.47
MedVD (μm)	10	18	24	34	527
	1%	25%	50%	75%	99%
Continental		Points = 4759[a]	Ta $\leq 0\,^{\circ}C$ I $\leq IL^{-1}$	TWC $\geq 0.005\,g\,m^{-3}$	
Ta (°C)	−24.7	−9.1	−6.2	−3.2	−0.2
Nd (cm^{-3})	2	55	121	233	643
TWC ($g\,m^{-3}$)	0.01	0.05	0.11	0.21	0.49
MedVD (μm)	10	13	17	22	643

[a]Liquid and mixed-phase, in-icing conditions.

atmosphere at any well-defined temperature. As described in cloud physics texts, generally insoluble IN can activate the process by allowing vapor to go directly into the ice phase (deposition nuclei), by condensation onto the IN followed by freezing (condensation freezing nuclei), by IN contact with a supercooled droplet (contact nuclei), or by IN immersion into a droplet followed by freezing (immersion nuclei). A typical concentration of IN might be 1 per liter near −20 °C, with a variability of X10 in concentration not being uncommon. Once ice is initiated, the concentration can increase via several ice multiplication mechanisms. For example, evaporation and melting of ice particles creates locally thinner regions leading to breakup, or during the riming process in the presence of graupel and large cloud drops, the freezing and shattering of favorably accreted drops can occur (Hallett and Mossop, 1974). Many have tried to show that ice nuclei concentrations are dependent on temperature (Fletcher, 1962), or both temperature and supersaturation (Meyers *et al.*, 1992). These relationships are often used in numerical models. However, observations often show (e.g. Gultepe *et al.*, 2001) that ice particle concentrations in the atmosphere do not depend on temperature, and appear to be the same in different geographic regions. Ice nuclei and ice particle concentrations found in clouds are not easily related to each other because of possible ice multiplication mechanisms and the redistribution of ice particles within clouds. The measurement uncertainties for both IN and small ice crystals are also substantial. Ice initiation in the atmosphere is a complicated process that is not well understood. It should be noted that the existence of any extensive region of supercooled drops depends on an absence of ice particles and nuclei, while the universal presence of ice suggests a remarkably efficient redistribution system, as in a mature hurricane or an extensive occluded cold, deep, midlatitude low pressure system.

Small ice crystals have many shapes generally based on hexagonal or sixfold symmetry. These can include columns, needles, plates, dendrites, and so on (Figure 6). Magono and Lee (1966) discuss ice particle shapes as a function of temperature (0 to −40 °C) and supersaturation. Bailey and Hallett (2004) have recently updated the work at temperatures colder than −20 °C and extended the temperature range to −70 °C. Much of this work on ice particle shape was done in the laboratory under controlled conditions. Measurements in natural clouds, especially lower-level layered cloud, show that ice particles tend to have quite irregular shapes, either consisting of faceted polycrystalline particles or sublimating (solid to vapor) ice particles with smooth curving sides and edges (Korolev *et al.*, 1999, 2000; Figure 7). More recent work on quantifying shapes has found that ice particles tend to be "rounder" when they are smaller, except for a small fraction which grow as pristine single crystals. As the remainder grow larger, having many crystal orientations, they tend to get more shaped (Korolev and Isaac, 2003).

The saturation vapor pressure over ice is less than the saturation vapor pressure over water for all temperatures (Table 1), resulting from the greater bonding of water molecules in the ice lattice as compared with supercooled liquid water. Consequently, in a supercooled cloud, once ice crystals begin to form, they grow rapidly by deposition while the surrounding droplets evaporate. If the cloud is not growing with a substantial updraft, the ice crystals will dominate and can quickly convert all the liquid to ice in a process called *glaciation*. Once the ice crystals reach a certain size they begin to fall and accrete water droplets through a process known as *riming*. They can

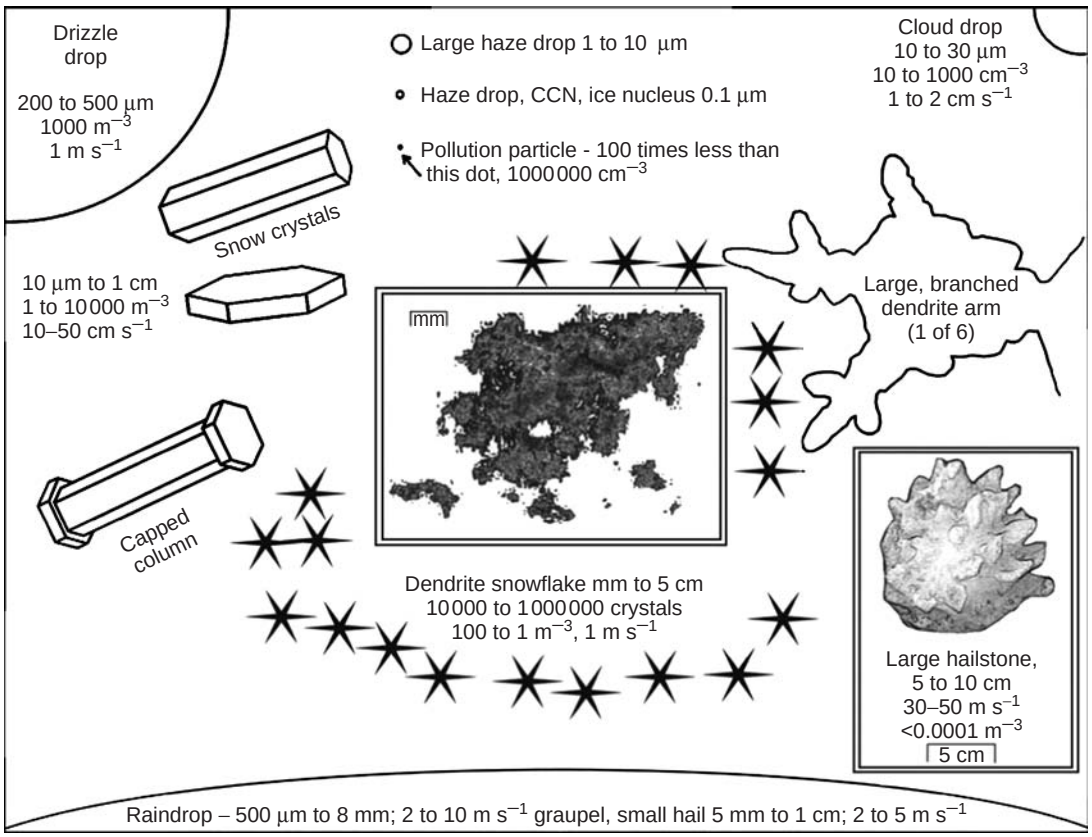


Figure 6 Relative sizes and shapes of cloud aerosols (haze particles, ice nuclei), cloud droplets, raindrops, ice crystals, snowflakes, graupel and hail (Adapted from Wallace and Hobbs, 1977; McDonald, 1958). Typical diameters, concentration, and fall velocities are given. For an additional sizing reference, a human hair is about 100 μm in diameter. The ice crystals shown are the classical hexagonal shapes, depending on the growth conditions. Columns form near −4 °C and plates at higher and lower temperatures, which sometimes lead to a capped column. Dendrite arm side branches may be symmetric or otherwise depending on the ambient turbulence (Hallett and Knight, 1994). Most ice particles have irregular shapes (e.g. Korolev *et al.*, 1999, 2000) resulting from different nucleation and changing growth conditions

also aggregate with other ice crystals to form snowflakes. Formation of precipitation via this mechanism accounts for most of the precipitation that forms in mid- and high latitudes. This theory was proposed by Bergeron (1935) and further refined by Findeisen (1938). Because it builds on the early work of Wegener (1911), it is referred to as the Bergeron–Findeisen–Wegener theory, or it is often shortened to the Bergeron–Findeisen process. As an illustration of this process, cirrus clouds often produce large ice crystals that begin to fall. If these ice crystals fall into lower-level supercooled liquid clouds, they can initiate precipitation. Much of the physical basis for cloud seeding or weather modification is to create the initial ice crystals by providing artificial ice nuclei, or through the use of a coolant such as evaporating dry ice which can produce temperatures well below −40 °C, near −80 °C.

This process produces rain because the resulting ice crystals typically melt when they fall into warmer air near the surface, with melting usually being complete for temperatures above +4 or +5 °C. However, most of our snow, when such a warm layer is not present, is also formed by the Bergeron–Findeisen process.

PRECIPITATION TYPES, DEW AND FROST

Precipitation as Rain

Precipitation may be defined as those particles falling from clouds in the atmosphere which reach the ground and remain for sufficient time to leave an observable residue of water. Table 2 shows how the fall velocities of different precipitation particles depend on their size. Precipitation it may be in the form of liquid-rain or drizzle as defined in Sections "Warm rain process" and "Bergeron–Findeisen or ice precipitation process" , either above 0 °C or supercooled below 0 °C. Drizzle and rain drops are spherical for sizes <0.8 mm. Larger raindrops deform during fall with the lower part flattening by the airflow (Figure 8). In contrast, cloud drops fall at a velocity of less than 100 cm s^{-1}

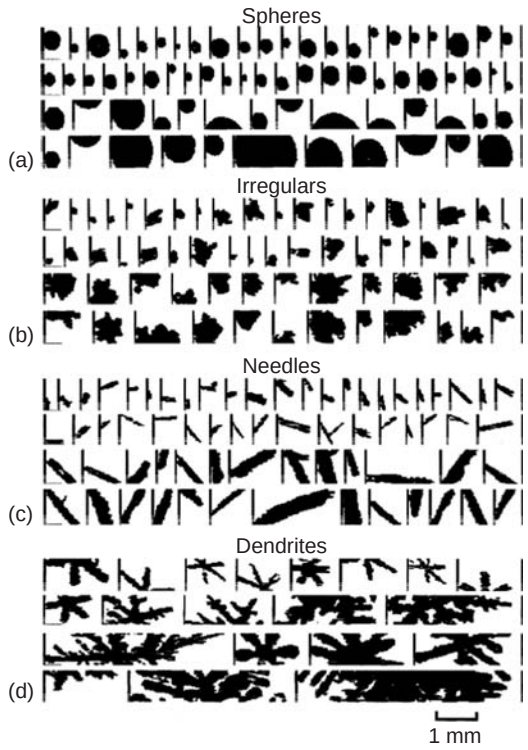


Figure 7 An example of different cloud particles as seen by a Particle Measuring System 2D imaging probe (Knollenberg, 1981) while flying through different clouds. (From Korolev *et al.*, 2000 by permission of The Royal Meteorological Society).

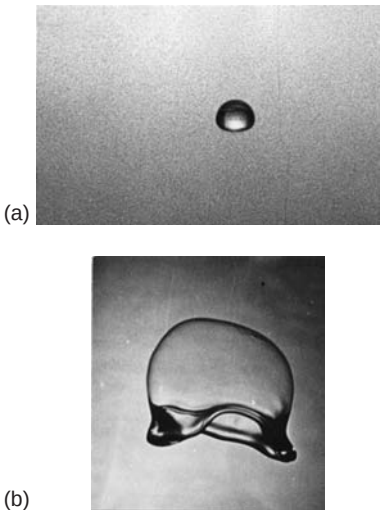


Figure 8 Shapes of drops. Drops in free fall in the atmosphere are in balance between their weight and drag forces resulting from both viscosity of the air and a hydrodynamic effect. With increasing size above about 20 μm, flow around the drop becomes increasingly asymmetrical, with a standing eddy wake and eventually shedding vortices, peaking by resonance with the drop surface tension for a size near 1 mm. Much below this size drops are spherical (Figure 9). Near this size resonance occurs and the drops move sideways during fall. At larger sizes the drop base becomes increasingly flatter (Figure 8a, 3 mm horizontal diameter), and eventually turn inside out and a large bubble forms, becoming greater than a centimeter in diameter (Figure 8b). As it blows up further, the film breaks and forms many small drops with intermediate size drops being formed from the lower ring remnant (© John Hallett)

dependent on the square of the size. Beyond about 6 mm, depending on ambient turbulence, the flat bottom may deform by the airflow, turning the drop inside out as a thin water film bag some 1–2 cm dimension with a ring at its base, which quickly breaks up into much smaller drops (Figure 8). This process limits raindrop size to 6 mm in more turbulent air, and up to >9 mm under quiescent conditions. The fall velocity of such particles increases to a maximum between 9 and 10 m s^{-1}. Figure 9 shows how the fall velocity of liquid particles increases from <1 cm s^{-1} for 10 μm diameter, approximately 100 cm s^{-1} for drizzle drops, >4 m s^{-1} for >1 mm raindrops. For a size a little less than 1 mm, the drop resonates with eddies shed from its rear and falls at angles as much as 40 degrees to the vertical. Larger and smaller drops fall straight down in quiescent air.

Drops splash and break on impact with the ground or vegetation, the splash depending on whether the ground is dry or wet and the depth of any liquid film. This spreads spores from fungi to considerable distances and aids their dispersal (Levin and Hobbs, 1971). Splash is different in fresh or ocean water and is highly complex (Hallett and Christensen, 1984). The drop impact yields first a crater, which subsequently retracts to an upward vertical jet which breaks. Small drops and drops from the breaking jet fall back into the water and penetrate to as much as 10 cm below the surface – but less in ocean water because freshwater has buoyancy (Figure 10).

Optical phenomena, such as rainbows, corona, and glory yield information on the nature of liquid cloud and precipitation particles from simple observation. Figure 11(a) shows the transition of snow scatter of white light to a well-defined rainbow as melting proceeds. The rainbow formed from large raindrops is a refraction process, but for smaller drops and cloud drops diffraction effects are increasingly important. Table 4 summarizes observations of common optical phenomena leading to inference of the form (drop size, crystal, shape, fall orientation) of the cloud or precipitation particles responsible. Distinction is made between phenomena viewed toward the sun, as ice halo and cloud corona and to phenomena opposite the sun as the rainbow or glory. Distinction is further made between phenomena having constant angular diameter as the primary and secondary rainbow (42, 50 degrees), rainbow and ice crystal halo (23 and 46 degrees) and variable diameter

Table 4 Inferences from common optical phenomena into the properties of atmospheric particulates. Distinction is made between phenomena viewed towards the sun direction (right column, ice halo, corona), opposite the sun direction (left column, rainbow, glory), and at 90 degrees, middle. Distinction is also made between phenomena depending on particle shape having near constant angle to the sun or antisun viewing direction (rainbow, halo) and phenomena having changing angles and colors depending on particle size (corona, glory). See Minnaert (1954), Greenler (1989), Tape (1994), Lynch and Livingston (2001)

Viewing direction		
Away from the sun direction (look at shadow of head)	At 90 degrees to sun direction	Toward sun direction
Primary rainbow: 42 degrees from sun; blue toward sun; red away from sun	Polarized light all directions; Haidinger's Brush Molecular scattering	22 degrees solar halo; red toward sun; blue away from sun
Two internal reflections inside near spherical raindrop $>100\,\mu$m diameter		Refraction by 120 degrees in hexagonal, ice crystals $>20\,\mu$m thick
Separation by Alexander's dark band		Dark outside of halo
Secondary rainbow 50 degrees away from antisun; inside red; outside blue		46 degrees halo; red toward sun; blue away from sun
Three internal reflections and refraction inside near spherical raindrop $>100\,\mu$m		Refraction by 90 degrees in hexagonal prisms $>20\,\mu$m in size
White rainbow; drop diameters 50–100 μm		Sun dog; enhanced colors at low angles in haloes
		Falling hexagonal crystals oriented horizontal by attached vortices
Diffraction effects and spread of drop-size wash out colors		
Supernumary rainbows, inside primary (brighter) and outside secondary bow, resulting from diffraction of different size drops leading to colors and different bow diameters. Intense with narrow drop-size spread >100um, < mm diameter		
Drops may oscillate with changing electric fields of a lightning discharge causing the rainbow to shake		
Away from sun/moon Glory (as around a shadow of head or aircraft)		Toward sun/moon Corona (as around sun or moon)
Colored rings around sun or moon resulting from diffraction effects from cloud droplets (occasionally ice spheres) A diffraction effect giving a wide range of ring diameters and colors		
Colors closer to primary colors (two adjacent primary colors missing)		Colors closer to complementary, white less primary interference color
Optical interference for $n\lambda/2 = d\sin\theta$ (λ = wavelength, d = droplet diameter, θ = half angle sun to ring, n = order of ring; sometimes $>n = 5$ visible). Smaller drops give larger diameter ring; sun angular diameter = 1/2 degree, rings visible to >10 degrees. Uniform droplet size gives intense colors, a spread gives washed out colors. Droplets size a few um to $>20\,\mu$m. Colors not uniquely related to drop size for values near few μm.		
For spatially uniform droplets, colors are in a well-defined ring. For clouds with drop size changing with viewing angle, as in clouds growing over a mountain, both corona and glory may change angle from place to place around the viewing direction. As a special case, for lenticular clouds up to >15 degrees from the sun, with limited dimension (5 degrees), the sun angle change is small compared with the drops in the cloud, which grow upstream and have a minimum size at the cloud leading and trailing edge and a maximum size in the middle. This leads to uniform colored regions tracing the cloud edge with constant d sin θ and changing orders of color. Turbulent cloud may also be colored (mother of pearl clouds) but with changing patchiness.		

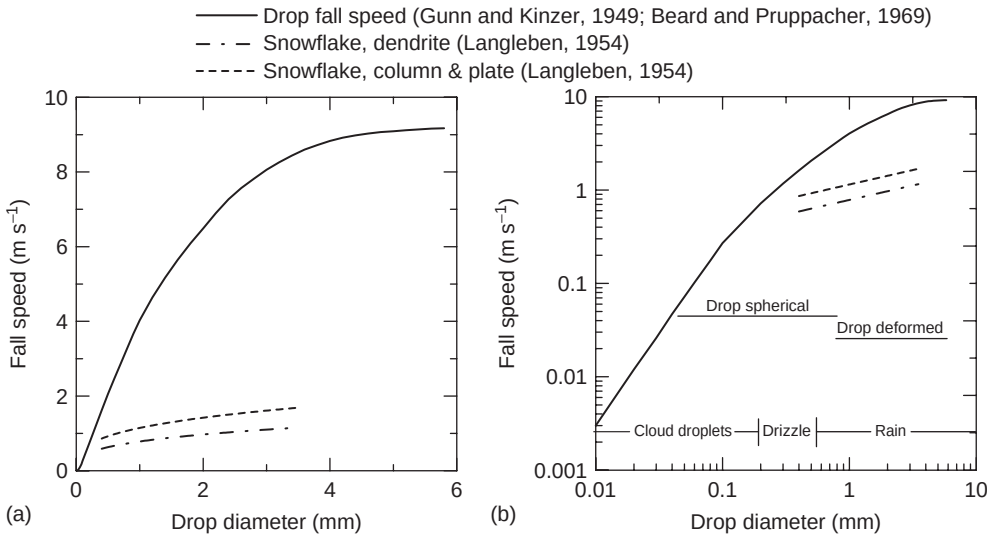


Figure 9 Fall velocity of drops, ice crystals and snow aggregates. Part (a) shows the fall velocity as a function of diameter on a linear scale, while part (b) shows it on a log scale in order to capture the full range of drops existing in the atmosphere. For the snowflakes, the melted equivalent diameter is used. Drops are spherical up to about 1 mm, increasingly deform by aerodynamic forces up to about 6–8 mm (depending on turbulence) and break up for larger size. Drops shed eddies and have a resonance with surface tension deformation with a resonance frequency of 320 hz for a diameter near 985 μm. They side slip during fall. Fall velocities decrease with deformation as compared with a frozen sphere

as the corona and glory. The former rely on the spherical shape of smaller raindrops and fixed angle of an hexagonal ice crystal, whereas the latter rely on different diameters of cloud drops in the viewing direction. A corona is observed looking toward the sun, and a glory is observed away from the sun with a distinction in the perceived color in the form of a ring for a uniform droplet size (Figure 11b and c). A smaller diameter implies larger cloud drops. Should cloud drop size be related to location (as in a lenticular wave cloud) well away from the sun, the contour of the cloud identifies the colors (Figure 11d and e).

Precipitation as Snow

Precipitation as ice occurs in many forms. Ice forms individual crystals as hexagonal columns or dendrites. These are ideal, pristine forms. It also occurs as rimed snow, graupel, or hail, as frozen drops, or as aggregates of other snow particles. Snow pellets and snow grains are variously defined (AMS Glossary – Glickman, 2000) and are more complex particles of ice precipitation often formed from multiple crystals grown from a frozen drop as an initial nucleus. The vapor growth habit of individual and also the component crystals of polycrystals (defined as the relative lengths of hexagonal direction compared with that at right angles) is related to growth temperature and growth supersaturation (relative humidity with respect to ice $>100\%$, up to about 140% for growth in cold supercooled clouds at temperatures approaching $-40\,^{\circ}$C). The local supersaturation also increases with fall velocity, which is dependent on the particle size and shape through the drag coefficient and the particle mass through its distribution of density. The density of snowflakes range from <0.1 to $0.4\,\mathrm{g\,ml^{-1}}$.

Figure 12(a–c) show individual crystals. Figure 12(d, e) show composite crystals resulting from growth under changing temperature. Snow flakes may be composed of pristine crystals or rimed crystals or mixtures of both. The presence of pristine crystals is readily inferred from observation by the presence of specific optical effects, Crystals in cirrus produce haloes of well-defined angles around sun or moon (Table 4, Figure 11f, g) and also sun dogs (parhelia) should the crystals become oriented during fall by attached wake eddies (Figure 11h). Thin section analysis in polarized light reveals a complex growth regime from the distribution of crystals and bubbles in small ice particles (Figure 12f, 12g). Fall velocities of ice particles are also shown in Figure 9. Ice particles and snowflakes fall much slower than liquid drops of the same size.

As snow falls to the surface, lower layers tend to compact because of the weight of the snow above. With temperature from some -15 to $-20\,^{\circ}$C and below, crystals retain their original habit for days or longer. Metamorphosis occurs by vapor evaporation from edges and redeposition at contact points. These processes eventually lead to bubbly ice, precursor to glacier ice and its ability to flow under the overlying stress. Bubbles shrink as the pressure increases and eventually disappear as the bubble air transforms to solid hydrate at hundreds of meters depth under pressures of many atmospheres. At higher temperatures, this change

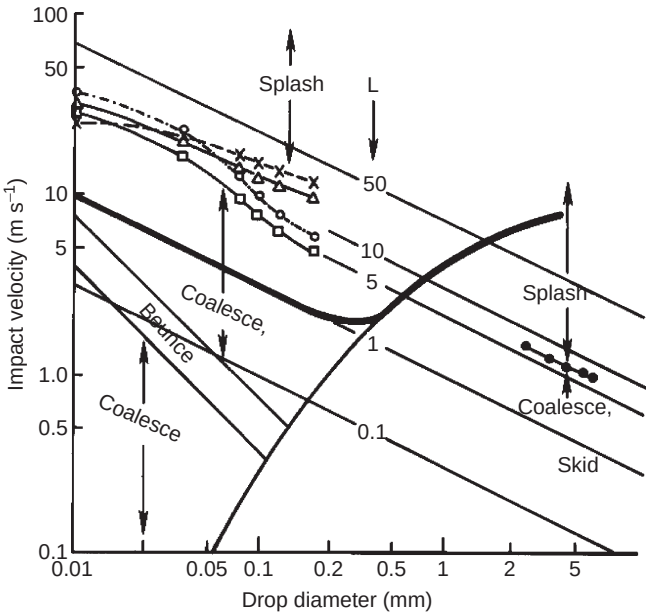


Figure 10 Drop Breakup and Splash Characteristics. Fog, drizzle or rain drops falling into a flat water surface, as a lake or ocean, behave in different ways depending on size and impact velocity. Assuming a terminal impact velocity (Figure 9), a drop, diameter <0.6 mm, forms a vortex ring which propagates into the liquid. The energy for propagation is the impact kinetic energy and the drop surface energy; the former dominates for large drops, the latter for small drops. L = ratio of kinetic to surface energy. Larger drops (diameter between 0.6 and 1.1 mm) leave a crater which collapses and projects a jet upwards which in turn breaks into smaller drops to fall back and reenter at a velocity well below terminal. Drops larger than 1.1 mm form a turbulent and wider jet which rises a shorter distance and also breaks before falling back. The jets are composed of liquid of the original drop and should sea water drops be lofted and evaporate under suitable wind conditions, provide large and ultragiant nuclei for cloud and drizzle drop formation. Drops diameter greater than 2 mm produce a rising thin film crown outward from their entry periphery, which may break up and produce a multitude of smaller drops, which provide smaller salt nuclei for smaller cloud drop formation. Some larger drops (>3 mm) entering the surface, without any local perturbation, produce a crown which closes at the top resulting in an air pressure reduction from the growth of the crater below and ultimately leave air bubbles, some 2 cm diameter with a flat base. These are readily seen in puddles and larger bodies of water under heavy rain conditions and persist a few seconds before draining and breaking. The splash is modified for a water layer of thickness comparable with the drop size (no crater or jet); freshwater entering sea water produces buoyant vortex rings which return toward the surface. Heavy rain on the ocean has a calming effect by providing a light water layer over a dense water layer and also transporting momentum from the top layer downwards. Local winds influence the splash pattern in even more complicated ways as does drop impact on wet sloping land, which influences the erosion of underlying soil

is much more rapid, leading to the granular (fraction of mm) nature of recently fallen snow near 0 °C. Vertical temperature gradients within the snow pack may lead to layers of vapor grown crystals close by layers of low-density snow, particularly near -4 and -15 °C. Such events are preludes to avalanches along these layers of weakness in mountainous terrain.

Dew and Frost Formation

Dew forms at the surface as drops up to 1–2 mm diameter, and may or may not wet the surface. They differ from guttation drops, typically formed on grass tips, formed by exuding water from the plant itself. In either case, a flux of vapor from moist air above or from damp soil vegetation below may be the source of dew or inhibit the evaporation of guttation drops. Dew formation follows radiative cooling of the surfaces when exposed to a clear sky or as moist air is advected over a cooled surface. With temperature below 0 °C, drops may supercool and nucleate over a temperature of several degrees below 0 °C. Most drops are frozen by -5 °C. With the initiation of ice, crystals grow from the vapor, in many ways having the same variability as snow crystals, although in the case of frost the ventilation varies from zero to ambient wind speed. Higher wind speed tends to mix out the cold surface layer producing the supersaturation. Most spectacular frost occurs with radiative surface cooling below 0 °C with moist air above, cooling and supersaturation spreading upward from the surface to give crystals at varying levels on surrounding vegetation (Figure 13a–c). Figure 13(d) shows ice boules that formed over a stream with the air temperature below 0 °C but the water temperature a few degrees above freezing.

Freezing Precipitation

Freezing drizzle and rain is a significant weather hazard involving rain falling into a layer colder than 0 °C at the surface, and then freezing upon contact. It can form either through the warm rain process, entirely at cold temperatures (Huffman and Norman, 1988), or by snow falling into a warm layer, melting, and then the rain supercooling in the cold boundary-layer air. Climatologies of freezing rain have been compiled for the US and Canada (Stuart and Isaac, 1999; Cortinas *et al.*, 2004). Stuart and Isaac (1999) showed a maximum in freezing precipitation occurrence at St. John's, Newfoundland, averaging about 150 h year^{-1}. The 5–9 January 1998 ice storm, which covered portions of northeastern US and southeastern Canada, yielding 90 mm of rain across a wide area, caused $4.4 billion in damages (see Gyakum and Roebber, 2001). An atmospheric profile during this storm is shown in Figure 5.

Graupel and Hail

Graupel and hail microphysical formation mechanisms have been summarized in Pruppacher and Klett (1997). Graupel

Figure 11 Optical phenomena that can yield information on particle type, size, shape and orientation (© John Hallett). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

is simply a heavily rimed snow particle that can either be lumpy or conical in shape, and is often considered as small hail. It can form in weakly convective clouds. Hailstones are greater than 5 mm in diameter up to the size of "grapefruit", and they form in large thunderstorms and convective complexes (Figure 14). Graupel particles have densities near 0.1 to 0.8 $g\,ml^{-1}$, while hailstones have density approaching pure ice (0.92 $g\,ml^{-1}$). Marwitz (1972) recognized the different types of storms and attempted to categorize them as supercell (one large updraft), multicell, squall line, and so on. It is clear that the formation mechanism of hail requires strong updrafts to keep them aloft. The largest hailstones can have terminal velocities of approximately 50 $m\,s^{-1}$ which implies updrafts of similar magnitude (Pruppacher and Klett, 1997). The interactions between the dynamics and microphysics is complex and not completely understood. However, hailstorms can cause large amounts of property and crop damage, mostly resulting from extensive regions of modest 0.5 to 3 cm diameter hail. Changnon and Burroughs (2001) compared the US annual average of $445 million in damages the $1.9 billion caused by one storm (2001 dollar values).

Thin sections of hailstones reveal the presence of air bubbles of different size and concentration resulting from rejection of dissolved air as ice crystals grow from accreted supercooled drops. Drop size and concentration changes from place to place in the updraft environment; the bubble characteristics reflect these changes giving alternation regions of clear and bubbly ice. Further information on the growth process can be obtained by use of polarized

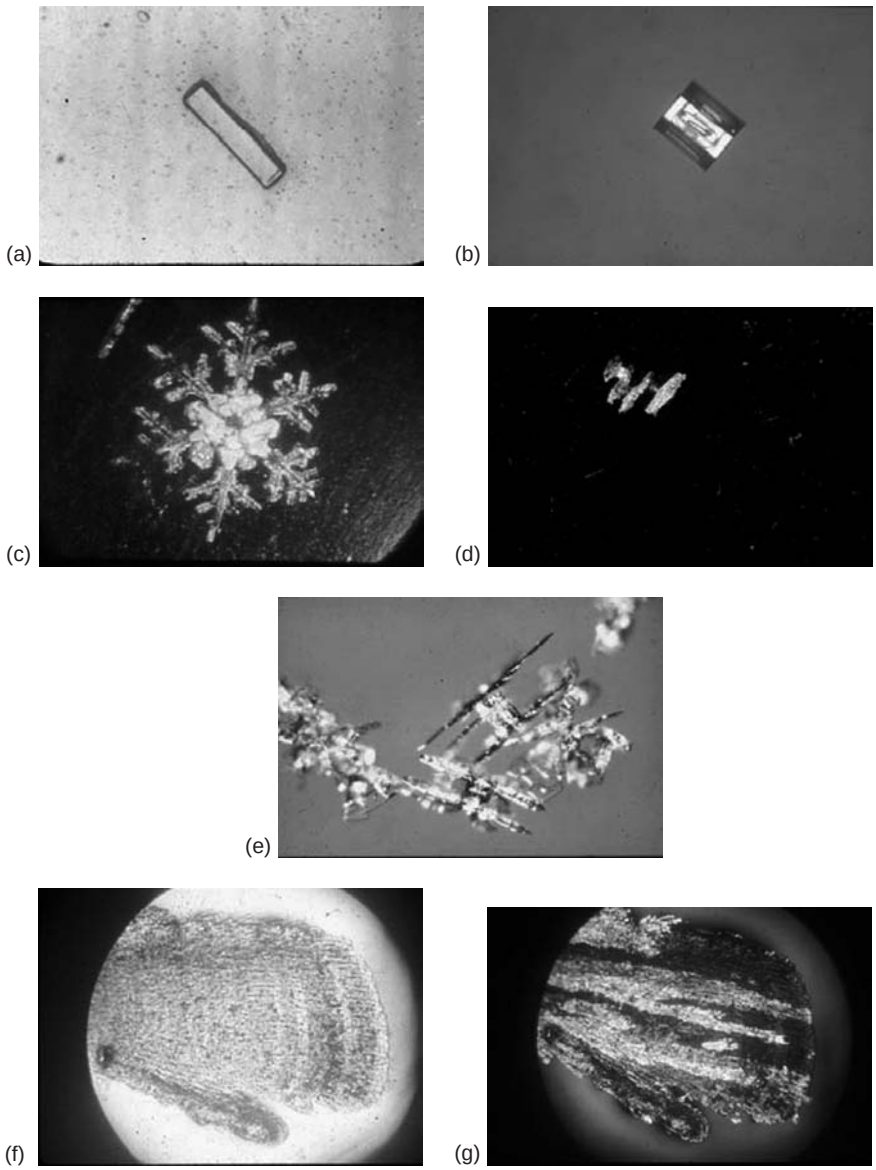


Figure 12 (a) Ice particles which formed in the atmosphere. (b) Growth is influenced by temperature, supersaturation and the accretion of cloud drops. (c) A dendrite crystal photographed by using the crystal as a mirror showing internal plates. (d) and (e) Composite crystals having combined plates and columns as conditions changed; (f) and (g) Part of a rimed crystal, in ordinary and polarized light. (b) Courtesy Yoshi Furukawa, Sapporo; (e) Courtesy Tom Henderson, Fresno (© John Hallett). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

light, distinguishing between wet growth when latent heat evolution is sufficiently rapid to prevent complete freezing and ice grows in to a liquid surface layer, and dry growth when supercooled drops are captured and freeze as individuals. Under the former conditions, large crystals are formed as shown by uniform color of regions of the thin section; in the latter case bubbles and small crystals may be formed. This may also occur under simultaneous accretion of mixed-phase ice and supercooled water drops. The correlation of crystals is complicated as it depends on several things, temperature, water and ice content, particle size distribution and fall velocity of the hailstone (Brownscombe and Hallett, 1967). Figure 14 shows thin sections of two stones. The first (Figure 14a, b) formed originally on a conical graupel particle. The second (Figure 14c, d) formed on a frozen raindrop, which grew first as graupel and then as a hailstone. The circular gaps around the core of the hailstone (Figure 14c) were regions grown as spongy ice, only partly frozen, during growth under high liquid-water content. Water was lost prior to collection and storage at low temperature. The contours of bubbles and changing ice size result from changing cloud conditions for growth. The

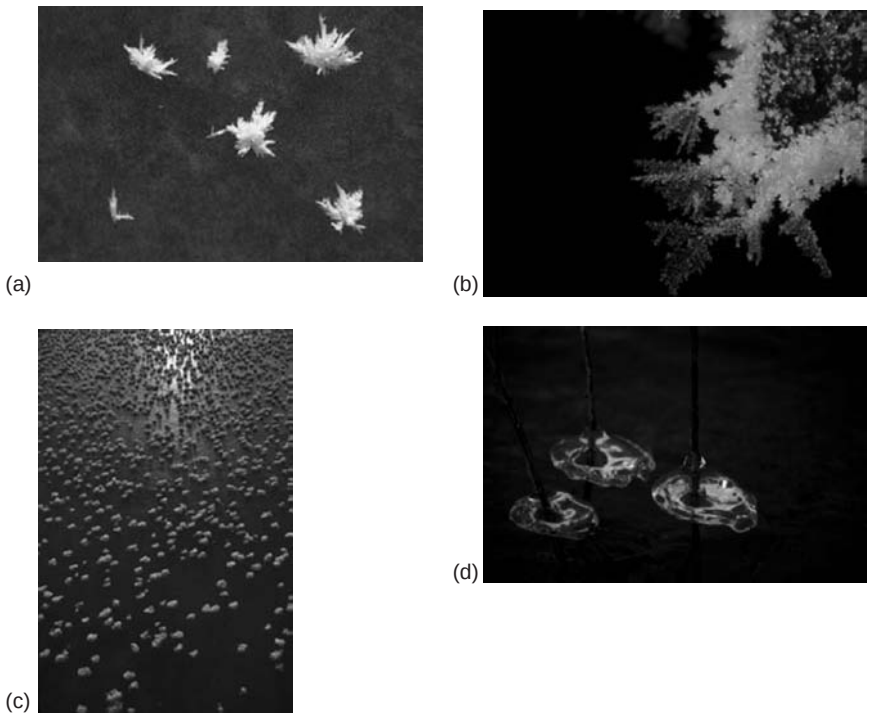


Figure 13 Examples of frost forming by radiative cooling with habit dependent on growth temperatures (a, b, and c) and ice boules formed over a stream by splashing of warmer water into cold air above (d) (© John Hallett). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

larger the hailstone, the larger the fall velocity, and the greater likelihood of its wet growth in high water content clouds.

CLOUD LIQUID-WATER CONTENT AND PRECIPITATION AS A FUNCTION OF TEMPERATURE

There have been several summaries of large data sets showing cloud properties as a function of temperature. Figure 15 shows how the cloud liquid-water content changes as a function of temperature (Gultepe and Isaac, 1997) for mid latitude stratiform clouds. Similar results were obtained by Feigelson (1978), Mazin (1995) and later again by Gultepe *et al.* (2002).

Similarly, precipitation amounts tend to be larger when the temperatures get warmer as shown in Figure 16 for some stations in the Mackenzie river valley (Isaac and Stuart, 1996). However, these relationships only hold for simple stratiform clouds. For precipitation from convective clouds, which often occur during the summertime in the same area, the amounts get smaller when the temperatures are cooler. Evidence is found for this in Figure 16 when the temperature gets warmer than $+10\,^{\circ}\mathrm{C}$. Isaac and Stuart (1992) came to the same conclusion in a study of Canadian precipitation-temperature relationships for different seasons.

Table 1 provides justification for why cloud water content and precipitation amount should depend on temperature. Warmer air has the capacity to hold more moisture and the potential for greater precipitation is thus enhanced. However, radiative influences may also be important in the precipitation-temperature relationship. For example, when convection occurs in the afternoon during the summertime, temperatures can drop because the incoming solar radiation is blocked.

CLOUD AND PRECIPITATION PARTICLE SIZE DISTRIBUTIONS

There have been many studies of the particle size distributions in clouds and in precipitation. Figure 17 shows results of averaging 2037 spectra averaged over 3 km in all liquid stratiform clouds found at temperatures between 0 and $-34\,^{\circ}\mathrm{C}$. The measurements were made using five different Particle Measurement System (PMS) probes (Knollenberg, 1981) and span a diameter range of a few microns to several millimeters. Of most relevance to hydrology is the particle size distribution of rain and snow at the surface. The early work of Marshall and Palmer (1948) and Gunn and Marshall (1958) assumed an exponential distribution of the

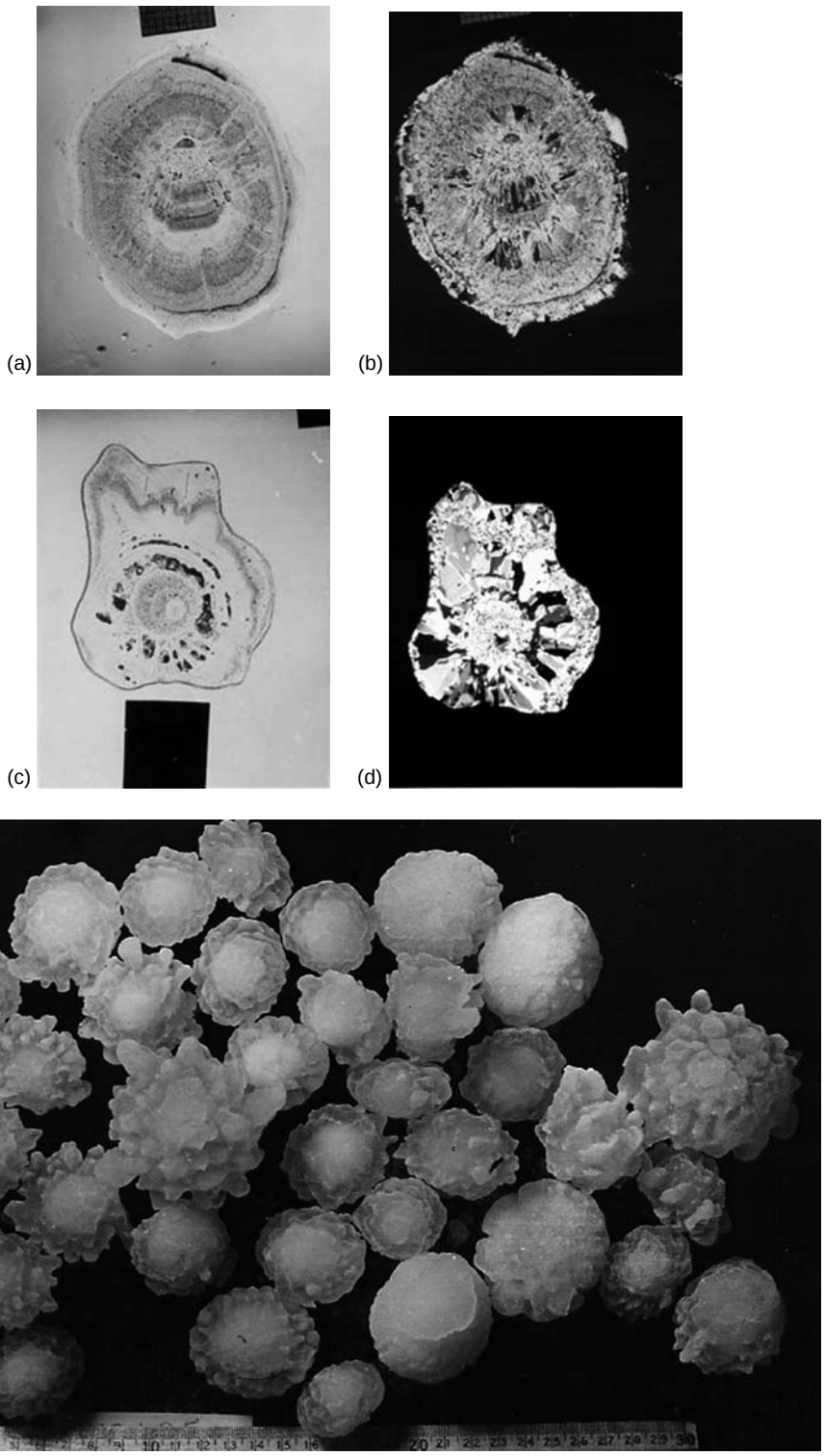


Figure 14 Hailstones. (a, b) thin section of hailstone showing conical center as observed with normal light and under polarized light (© John Hallett); (c,d) thin section of hailstone with a drop center as observed with normal light and polarized light (© John Hallett); (e) an example of hailstones with rough surfaces (© George Isaac; Barge and Isaac, 1973). The scale on (a) and (c) is 1 cm wide. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

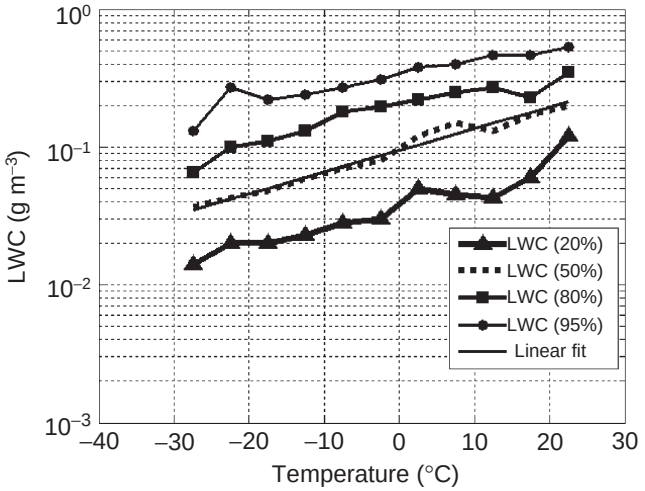


Figure 15 Cloud liquid-water content as a function of temperature for given percentile values. The solid line is the best fit to the median values. (Gultepe and Isaac, 1997. © 1997 American Meteorological Society)

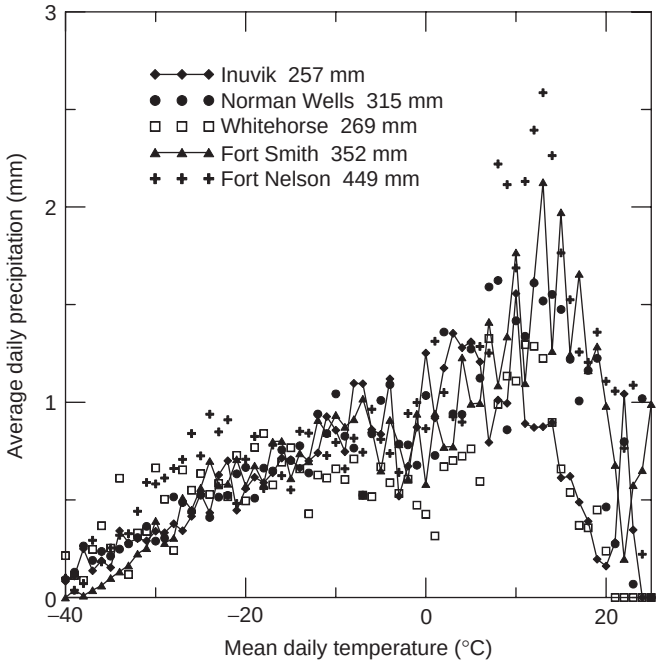


Figure 16 Relationship between mean daily surface temperature and average daily precipitation amounts for several stations in the Mackenzie river valley (Isaac and Stuart, 1996. © 1996 American Meteorological Society)

form:

$$N(D) = N_o \exp(-\Lambda D)$$

with $N_o = 8 \times 10^3\,\mathrm{m^{-3}\,mm^{-1}}$ ($\Lambda = 41R^{-0.21}$) for rain and $N_o = 3.8 \times 10^3 \quad R^{-0.87}\,\mathrm{m^{-3}\,mm^{-1}}$ ($\Lambda = 25.5R^{-0.48}$) for melted diameters of snow where R is in $\mathrm{mm\,h^{-1}}$.

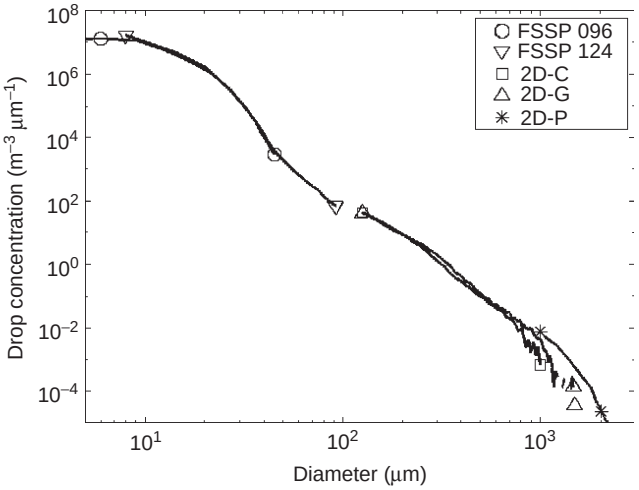


Figure 17 Results from averaging 2037 spectra made in all liquid Canadian winter clouds using an instrumented aircraft (Cober *et al.*, 2003) © 2003 by Environment Canada. Published by the American Institute of Aeronautics and Astronautics, Inc.

More recently, at least for raindrop-size distributions, both the gamma (Ulbrich, 1983) and lognormal distributions (Feingold and Levin, 1986) have been used. The gamma distribution is given as:

$$N(D) = N_o D^{\mu} \exp(-\Lambda D)$$

and it defaults to the exponential distribution when the curvature parameter, μ, goes to zero. The gamma distribution is a useful way of describing the raindrop size–distribution (e.g. Bringi *et al.*, 2003), but the coefficients depend on climate regime and type of rainfall (e.g. convective or stratiform). However, the older Marshall–Palmer parameterization is still being used and can give a reasonable first approximation (e.g. Sheppard and Joe, 1994).

Such distributions may only effectively describe extended data sets and thus are useful in a climatological sense. For specific short periods or small data sets, their use must be treated with caution because local turbulent regions may lead to particle sorting and physical processes of a specific scale (as the width of a cirrus trail) may lead to significant local deviations.

CLOUD SYSTEMS

The main purpose of this article is to describe cloud formation and the microphysical processes leading to precipitation. Other sections will handle storms and storm systems. However, Figure 18 shows the types of clouds and cloud systems that can generate precipitation around the world. In polar regions, as Figure 3 shows, little precipitation actually falls, and, what does comes from stratiform clouds and

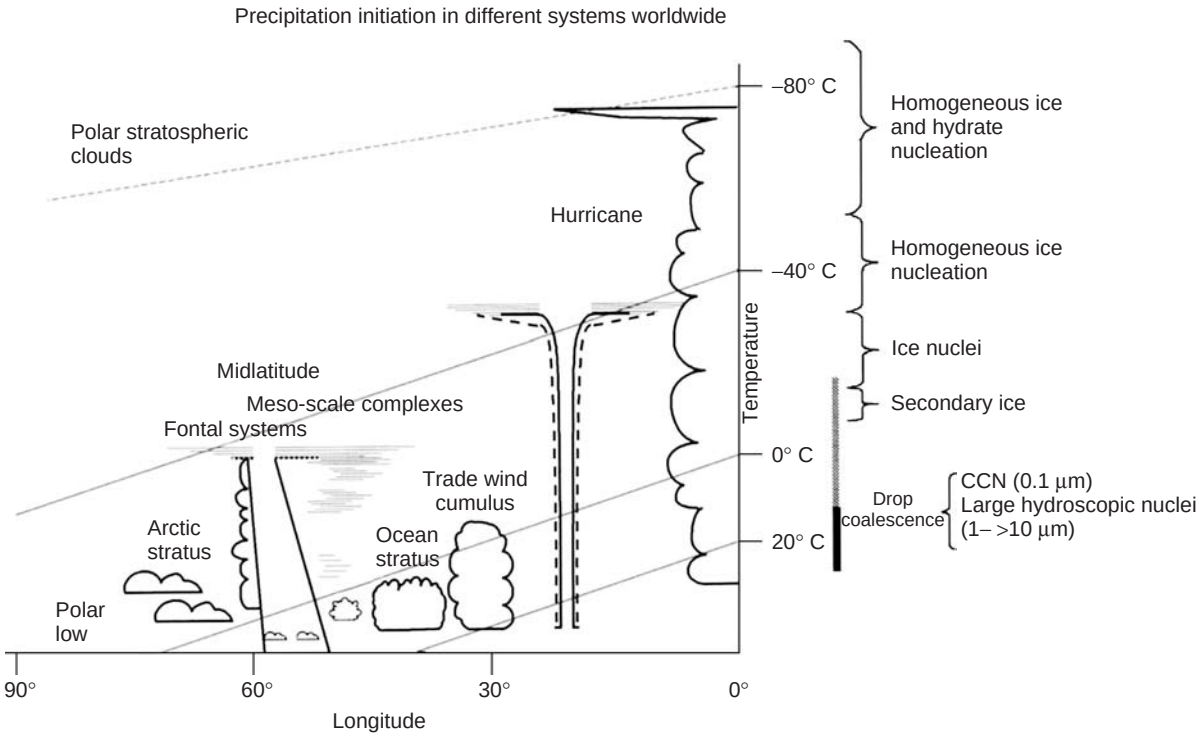


Figure 18 Examples of precipitation mechanisms, as a function of latitude and altitude, on a worldwide basis (Hallett and Isaac, 2001. © 2001 American Meteorological Society)

associated systems such as polar lows and frontal systems. Frontal systems and convective storms dominate the mid latitude belt, with the occasional hurricane. In the tropics, deep convective clouds and their associated systems account for the large amount of precipitation that falls within this latitude zone (see Figure 3). The dynamics of such systems is very complicated but obviously very important for cloud development and precipitation formation.

CONCLUDING REMARKS

The driving engine for most clouds systems is the latent heat release that occurs through phase changes. For example, for each mm of precipitation, $25\,\mathrm{W\,m^{-2}}$ of latent heat is released into the atmosphere. It is essential to know how such precipitation is formed from clouds in order to completely understand the energy cycle of the atmosphere. Our climate is particularly sensitive to cloud properties. For example, Slingo (1990) stated that the radiative forcing at the top of the atmosphere due to a doubling of carbon dioxide concentrations could be balanced by relatively modest increases of 15–20% in the amount of low clouds and 20–35% in liquid-water path, and by decreases of $\sim$ 15–20% in mean drop radius. Rotstayn (1999) suggested that a 1% increase in cloudiness, a 6% increase in liquid-water path, and a 7% decrease in droplet effective radius may result in a radiative forcing of about $-2.1\,\mathrm{W\,m^{-2}}$ in the heat budget of the atmosphere. In comparison, the Intergovernmental Panel on Climate Change (2001) estimated that the radiative forcing due to increases in greenhouse gases from preindustrial times to 1998 was $2.4\,\mathrm{W\,m^{-2}}$.

It should be emphasized that the greatest driving force for the atmospheric engine is generated in the tropics, since this is the region where cloud bases have the highest temperature and water contents aloft, and therefore most of the atmospheric latent heat is released in the convection occurring in this area. The transport of this energy away from the tropics drives the atmospheric system.

However, there are many uncertainties in how clouds and precipitation form. Hallett and Isaac (2001) summarized a Panel discussion at the 13th International Conference on Clouds and Precipitation (Reno, 2001), which considered this topic. We need to extend our knowledge of CCN for larger size particles, and determine the geographic variability of such particles. Our current knowledge of how ice forms in the atmosphere from ice nuclei to ice multiplication needs further study. The development of precipitation through the warm rain process also has gaps in our understanding. A better insight into the physical (particle growth) and dynamical (air motions) processes is required. We also need to gain more insight through studies

that compare model simulations to observations. Finally, we need a multidisciplinary approach to these questions because they require a balance between, theory, modeling, laboratory experimentation, instrument design and field observational work.

REFERENCES

Bailey M. and Hallett J. (2004) Growth rates and habits of ice crystals between −20 and −70°C. *Journal of the Atmospheric Sciences*, **61**, 514–544.

Barge B.L. and Isaac G.A. (1973) The shape of Alberta hailstones. *Journal de Recherches Atmospheriques*, **7**, 11–20.

Beard K.V. and Pruppacher H.R. (1969) A determination of the terminal velocity and drag of small water drops by means of a wind tunnel. *Journal of the Atmospheric Sciences*, **26**, 1066–1072.

Bergeron T. (1935) On the physics of clouds and precipitation. *Proces Verbaux de l'Association de Météorologie*, International Union of Geodesy and Geophysics, Paris, pp. 156–178.

Bringi V.N., Chandrasekar V., Hubbert J., Gorgucci E., Randeu W.L. and Schoenhuber M. (2003) Raindrop size distribution in different climatic regimes from disdrometer and dual-polarized radar analysis. *Journal of the Atmospheric Sciences*, **60**, 354–365.

Brooks H.E. and Doswell C.A. (2002) Deaths in the 3 May 1999 Oklahoma City tornado from a historical perspective. *Weather and Forecasting*, **17**(3), 354–361.

Brownscombe J.L. and Hallett J. (1967) Experimental and field studies of precipitation particles formed by the freezing of supercooled water. *Quarterly Journal of Royal Meteorological Society*, **93**, 455–472.

Changnon S.A. and Burroughs J. (2001) The tristate hailstorm: the most costly on record. *Monthly Weather Review*, **131**(8), 1734–1739.

Cober S.G., Isaac G.A., Shah A.D. and Jeck R. (2003) Defining characteristic cloud drop spectra from in-situ measurements. *AIAA 41st Aerospace Science Meeting and Exhibit*, AIAA: Reno Nevada, 6–10 January 2003, AIAA 2003–0561.

Cober S.G., Strapp J.W. and Isaac G.A. (1996) An example of supercooled drizzle droplets formed through a collision coalescence process. *Journal of Applied Meteorology*, **35**, 2250–2260.

Cortinas J.V., Bernstein B.C., Robbins C.C. and Strapp J.W. (2004) An analysis of freezing rain, freezing drizzle and ice pellets across the United States and Canada: 1976–90. *Weather and Forecasting*, **19**, 377–390.4.

DeGaetano A.T. (2000) Climatic perspective and impacts of the 1998 northern New York and New England ice storm. *Bulletin of the American Meteorological Society*, **81**(2), 237–254.

Feigelson E.M. (1978) Preliminary radiation model of a cloudy atmosphere. Part I: structure of clouds and solar radiation. *Beitrage zur Physik der Atmosphäre*, **51**, 203–229.

Feingold G. and Levin Z. (1986) The lognormal fit to raindrop spectra from frontal convective clouds in Israel. *Journal of Climate and Applied Meteorology*, **25**, 1346–1363.

Findeisen W. (1938) Kolloid-meteorologische Vorgänge bei Neiderschlags-bildung. *Meteorologiche Zeitschrift*, **55**, 121–133.

Fletcher N.H. (1962) *Physics of Rain Clouds*, Cambridge University Press. p. 386.

Fukuta N. and Gramada C.M. (2003) Vapor pressure measurement of supercooled water. *Journal of the Atmospheric Sciences*, **60**, 1871–1875.

Glickman T.S. (Ed.) (2000) *Glossary of Meteorology*, *Second Edition*, American Meteorological Society, p. 745.

Greenler R. (1989) *Rainbows, Halos, and Glories*, New York Cambridge University Press: Cambridge, p. 195.

Gultepe I. and Isaac G.A. (1997) Relationship between liquid water content and temperature based on aircraft observations and its applicability to GCMs. *Journal of Climate*, **10**, 446–452.

Gultepe I., Isaac G.A. and Cober S.G. (2001) Ice crystal number concentration versus temperature. *International Journal of Climatology*, **21**, 1281–1302.

Gultepe I., Isaac G.A. and Cober S.G. (2002) Cloud microphysical characteristics versus temperature for three Canadian field projects. *Annales Geophysicae*, **20**, 1891–1898.

Gunn R. and Kinzer G.D. (1949) The terminal velocity of fall for water drops in stagnant air. *Journal of Meteorology*, **6**, 243–248.

Gunn K.L.S. and Marshall J.S. (1958) The distribution with size of aggregate snowflakes. *Journal of Meteorology*, **15**, 452–461.

Gyakum J.R. and Roebber P.J. (2001) The 1998 ice storm – analysis of a planetary-scale event. *Monthly Weather Review*, **129**(12), 2983–2997.

Hallett J. and Christensen L. (1984) Splash and penetration of drops in water. *Journal de Recherches Atmospheriques*, **18**, 225–242.

Hallett J. and Isaac G.A. (2001) Perspectives in cloud physics. *Bulletin of American Meteorological Society*, **82**, 2259–2263.

Hallett J. and Knight C. (1994) On the symmetry of snow dendrites. *Atmospheric Research*, **32**, 1–11.

Hallett J. and Mossop S.C. (1974) Production of secondary ice particle during the riming process. *Nature*, **249**, 26–28.

Huffman G.J. and Norman G.A. (1988) The supercooled warm rain process and the specification of freezing precipitation. *Monthly Weather Review*, **116**(11), 2172–2182.

Isaac G.A., Cober S.G., Strapp J.W., Korolev A.V., Tremblay A. and Marcotte D.L. (2001) Recent Canadian research on aircraft in-flight icing. *Canadian Aeronautics and Space Journal*, **47–3**, 213–221.

Isaac G.A. and Stuart R.A. (1992) Temperature-precipitation relationships for Canadian stations. *Journal of Climate*, **5**, 822–830.

Isaac G.A. and Stuart R.A. (1996) Relationships between cloud type and amount, precipitation and surface temperature in the Mackenzie River valley – Beaufort Sea area. *Journal of Climate*, **9**, 1921–1941.

Kajikawa M., Kikuchi K., Asuma Y., Inoue Y. and Sato N. (2000) Supercooled drizzle formed by condensation-coalescence in the mid-winter season of the Canadian Arctic. *Atmospheric Research*, **52**, 293–301.

Knollenberg R.G. (1981) Techniques for probing cloud microstructure. In *Clouds: their Formation, Optical Properties and Effects*, Hobbs P.V. and Deepak A. (Eds.), Academic: San Diego, pp. 15–92.

Korolev A.V., Isaac G.A. and Hallett J. (1999) Ice particle habits in Arctic clouds. *Geophysical Research Letters*, **26**, 1299–1302.

Korolev A., Isaac G.A. and Hallett J. (2000) Ice particle habits in stratiform clouds. *Quarterly Journal of the Royal Meteorological Society*, **126**, 2873–2902.

Korolev A.V. and Isaac G.A. (2003) Roundness and aspect ratio of particles in ice clouds. *Journal of Atmospheric Sciences*, **60**, 1795–1808.

Korolev A.V., Isaac G.A., Cober S.G., Strapp J.W. and Hallett J. (2003) Observations of the microstructure of mixed phase clouds. *Quarterly Journal of Royal Meteorological Society*, **129**, 39–65.

Langleben M.P. (1954) The terminal velocity of snowflakes. *Quarterly Journal of Royal Meteorological Society*, **80**, 174–181.

Lawson P.R., Baker B., Schmitt C.G. and Jensen T. (2001) An overview of microphysical properties of Arctic clouds observed in May and June 1998 during FIRE ACE. *Journal of Geophysical Research*, **106**(D14), 14989–15014.

Levin Z. and Hobbs P.V. (1971) Splashing of water drops on solid and wetted surfaces; hydrodynamics and charge separation. *Philosophical Transactions of the Royal Society of London Series A*, **269**, 555–590.

List R.J. (1968) *Smithsonian Meteorological Tables*, *Sixth Edition*, Smithsonian Institution.

Lynch D.K. and Livingston W. (2001) *Color and Light in Nature*, *Second Edition*, Cambridge University Press: Cambridge, New York, p. 277.

Magono C. and Lee C. (1966) Meteorological classification of natural snow crystals. *Journal of the Faculty of Science*, Hokkaido University Series VII, **2**, 321–335.

Marshall J.S. and Palmer W.M.cK. (1948) The distribution of raindrops with size. *Journal of the Atmospheric Sciences*, **5**, 165–166.

Marti J. and Mauersberger K. (1993) A survey and new measurements of ice vapor pressure at temperatures between 170 and 250 K. *Geophysical Research Letters*, **20**, 363–366.

Marwitz J.D. (1972) The structure and motion of severe hailstorms. Part III: severely sheared storms. *Journal of Applied Meteorology*, **11**(1), 189–201.

Mason B.J. (1971) *The Physics of Clouds*, *Second Edition* Clarendon Press, p. 671.

Mazin I.P. (1995) Cloud water content in continental clouds of middle latitudes. *Journal of Atmospheric Research*, **35**, 283–297.

McDonald J.E. (1958) Physics of cloud modification. *Advances in Geophysics*, **5**, 244.

Meyers M.P., DeMott P.J. and Cotton W.R. (1992) New primary ice-nucleation parameterizations in an explicit cloud model. *Journal of Applied Meteorology*, **31**, 708–721.

Minnaert M. (1954) *The Nature of Light and Colour in the Open Air*, Dover: New York, Reprint. p. 362.

Pruppacher H.R. and Klett J.D. (1997) *Microphysics of Clouds and Precipitation*, Kluwer Academic Publishers, ISBN 0-7923-4211-9.

Rasmussen R.M., Bernstein B.C., Murakami M., Stossmeister G., Reisner J. and Stankov B. (1995) The 1990 Valentine's Day Arctic outbreak. Part I: Mesoscale and microscale structure and evolution of a Colorado front range shallow upslope cloud. *Journal of Applied Meteorology*, **34**, 1481–1511.

Roberts P. and Hallett J. (1968) A laboratory study of the ice nucleating properties of some mineral particulates. *Quarterly Journal of the Royal Meteorological Society*, **94**, 25–34.

Rogers R.R. and Yau M.K. (1989) *A Short Course in Cloud Physics*, *Third Edition*, Buterworth-Heinemann Publications, ISBN 0-7506-3215-1.

Rotstayn L.D. (1999) Climate sensitivity of the CSIRO GCM: effect of cloud modelling assumptions. *Journal of Climate*, **12**, 334–356.

Scorer R.S. (1972) *Clouds of the World: A Complete Color Encyclopedia*, Stackpole Books: Harrisburg.

Sheppard B.E. and Joe P.I. (1994) Comparison of raindrop size distribution measurements by a Joss-Waldvogel distrometer, a PMS 2DG spectrometer, and a POSS Doppler radar. *Journal of Atmospheric and Oceanic Technology*, **11**, 874–887.

Slingo A. (1990) Sensitivity of the earth's radiation budget to changes in low clouds. *Nature*, **343**, 49–51.

Stuart R.A. and Isaac G.A. (1999) Freezing precipitation in Canada. *Atmosphere Ocean*, **37–1**, 87–102.

Tape W. (1994) *Atmospheric Halos*, *Antarctic Research Series. Vol. 64*, American Geophysical Union: Washington, p. 143.

Thompson R.L. and Edwards R. (2000) An overview of environmental conditions and forecast implications of the 3 May 1999 tornado outbreak. *Weather and Forecasting*, **15**(6), 682–699.

Ulbrich C.W. (1983) Natural variations in the analytical form of the raindrop size distribution. *Journal of Climate and Applied Meteorology*, **22**, 1764–1775.

Wallace J.M. and Hobbs P.V. (1977) *Atmospheric Science, an Introductory Survey*, Academic Press: New York, p. 467.

Wegener A. (1911) *Thermodynamik der Atmosphäre*, Leipzig.

Whitlock C.H., Charlock T.P., Staylor W.F., Pinker R.T., Laszlo I., Ohmura A., Gilgen H., Konzelman T., DiPasquale P.C., Moats C.D., *et al.* (1995) First global WCRP shortwave radiation budget data set. *Bulletin of the American Meteorological Society*, **76**, 905–922.

WMO (1975) *International Cloud Atlas*, Vol. 1, World Meteorological Organization, p. 155.

WMO (1987) *International Cloud Atlas*, Vol. 2, World Meteorological Organization, p. 212.

Xie P. and Arkin A. (1996) Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *Journal of Climate*, **9**, 840–858.

Xie, P. and Arkin A. (1997) Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society*, **78**, 2539–2558.

29: Atmospheric Boundary-Layer Climates and Interactions with the Land Surface

ALBERT AM HOLTSLAG[1] AND MICHAEL B EK[2]

[1] *Meteorology and Air Quality Wageningen University, Wageningen, The Netherlands*
[2] *National Centers for Environmental Prediction/Environmental Modeling Center, Suitland, MD, US*

The purpose of this article is to provide an overview of the relevant processes and interactions in the Atmospheric Boundary Layer (ABL) over land. Modeling principles for the ABL are summarized and discussed, including the formulation and role of the surface fluxes. It is shown that both the ABL and the coupled soil-vegetation-atmosphere system can be well represented in the well-defined cases examined. The interactions between the ABL and the land surface are illustrated with the role of soil moisture on boundary-layer cloud initiation. As such, a new analytical development is discussed and illustrated for the tendency of relative humidity at boundary-layer top. Typically, increased soil moisture leads to a higher potential for boundary-layer clouds, but in some cases, the largest potential for boundary-layer clouds is predicted over dry soils. For readers not familiar with atmospheric turbulence and meteorological definitions, some background is given as well.

INTRODUCTION

The characteristics of the Earth's surface strongly influence the motions and processes in the lower part of the atmosphere, known as the *Atmospheric Boundary Layer* (ABL). Usually, friction of the wind causes turbulence in the ABL, and this increases with the roughness of the surface. Over relatively warm surfaces, the ABL is heated from below, and convection may strongly amplify the environmental turbulence. On the other hand, over relatively cold surfaces, the presence of turbulence may vanish completely owing to the cooling of the ABL. Thus, the climate in the atmospheric boundary layer strongly depends on the presence of turbulence in response to the surface characteristics and the atmospheric flow above the ABL (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**).

Turbulence directly impacts the transfer between the surface and the atmosphere of momentum, sensible heat, water vapor, ozone, and methane, among many other quantities. Turbulence also defines the mixing of properties inside the atmospheric boundary layer, the transfer of quantities between the boundary layer and the clear or cloudy atmosphere aloft, and the mixing inside clouds. It is characteristic that the surface temperature over sea is rather constant on the timescale of a day, while over land, the surface temperature may vary considerably owing to solar downward radiation and long-wave cooling of the surface. This has important implications for the turbulence and processes in the ABL (e.g. Stull, 1988).

In this article, we focus on the processes and the resulting climate of the ABL over land. Figure 1 (after Ek and Holtslag, 2004) gives a schematic illustration of the relevant processes and interactions in an ABL over land during daytime conditions. In such conditions, the sensible heat arising from the surface supports a growing boundary layer with a rising temperature. This may be amplified by the mixing of warmer air from above the boundary layer, known as "entrainment". Given a certain value of specific humidity in the ABL, a rising temperature in the ABL implicates a lower value for the relative humidity (*RH*). This normally implies a higher evaporation or transpiration (latent heat flux) from the surface given that sufficient soil moisture is available. However, the energy for sensible and latent heat at the surface is limited by the available net radiation and soil heat flux. These variables depend in turn on the characteristics of the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

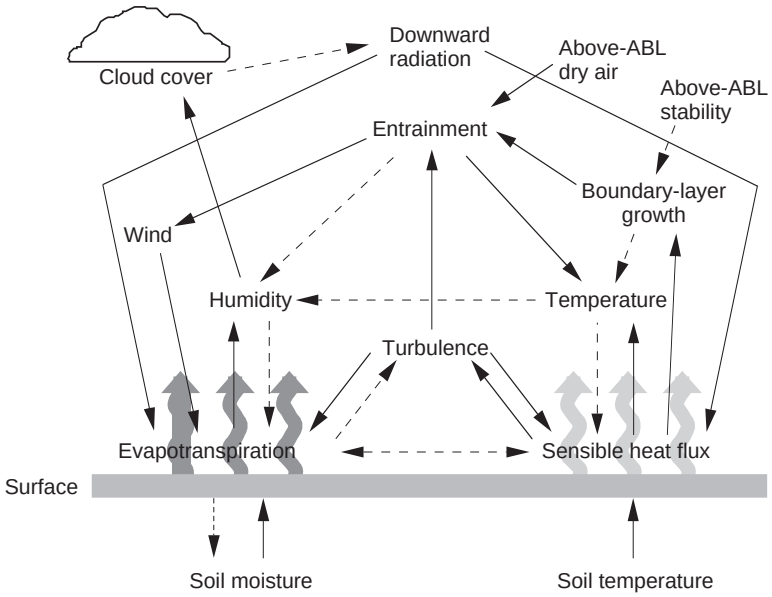


Figure 1 Schematic showing important interactions between the land surface and atmospheric boundary layer for conditions of daytime surface heating (Ek and Holtslag, (2004). © 2004 American Meteorological Society). Full lines indicate the direction of feedbacks, which are normally positive (leading to an increase of the recipient variable), while dashed lines indicate negative feedbacks. Two consecutive negative feedbacks make a positive one. Note the many positive and negative feedback loops, which may lead to increased or decreased humidity and cloud cover

surface and the temperature and moisture conditions in the soil.

The preceding text illustrates that the coupling of the ABL to the land surface is rather complex and depends on many variables. More background on this is given by Betts (2000), Margulis and Entekhabi (2001), Bonan (2002), Findell and Eltahir (2003), and Kabat *et al.* (2004), among many others studies. Ek and Holtslag (2004) quantified the different factors influencing cloud initiation in the ABL, with an analytical development. The findings of the latter study are summarized below (Section 6).

The purpose of this article is to provide an overview of the relevant processes and interactions in the ABL over land as well as the modeling of these. For readers not familiar with atmospheric turbulence and meteorological definitions, some background is given in Section 2. Subsequently, Section 3 provides an overview of the boundary-layer characteristics over land, and Section 4 summarizes the modeling principles for the ABL. Then, Section 5 deals with the surface fluxes and Section 6 with the interactions between the ABL and the land surface, in particular, the role of soil moisture on boundary-layer cloud initiation.

BACKGROUND

There is an enormous range of scales in atmospheric motion, with length scales from millimeters to the atmospheric motion on global scales (e.g. Stull, 1988; Garratt, 1992; Holtslag, 2002). As such, there is a need to separate the relatively small scales of atmospheric turbulence (length scales up to orders of kilometers) from the larger-scale motions. Let C denote an atmospheric variable, such as specific humidity. Then, $\overline{C}$ represents a mean or "smoothed" value of C, typically taken on a horizontal scale of order 10 (or more) kilometers and a corresponding timescale on the order of 10 min to 1 h. A local or instantaneous value of C would differ from $\overline{C}$. Thus, we have

$$C = \overline{C} + c \tag{1}$$

Here c represents the smaller-scale fluctuations (note that we use lower case for the latter). In principle, the fluctuations around the mean motion also reflect gravity waves and other smaller-scale motions, in addition to turbulence. Gravity waves often coexist with turbulence or are generated by turbulence. If the wind at the same time is weak, there may be no turbulence at all. Anyhow, if turbulence exists, it is usually more important for most atmospheric applications, because it mixes more efficiently than the other small-scale motions.

To make the mathematical handling of c tractable, it must satisfy the so-called Reynolds postulates. These require, for example, that $\overline{c} = 0$ and that small- and larger-scale values must not be correlated. After a quantity has been averaged to create a larger-scale quantity, further averaging should produce no further changes, in order for this postulate to apply. The mean of the summation of two variables A

and C will produce $\overline{A \pm C} = \overline{A} \pm \overline{C}$. A further condition is that a mean variable $\overline{C}$ must be differentiable, since differentials show up in the atmospheric equations (see below). In practice, not all these conditions are rigorously satisfied. If the Reynolds postulates are fulfilled, then the averaging for the product of two variables provides

$$\overline{AC} = \overline{A}\,\overline{C} + \overline{ac} \tag{2}$$

The second term at the right-hand side of equation (2) is known as the *turbulent covariance*. Similarly, the turbulent variance of a quantity is given by $\overline{C^2} - (\overline{C})^2$ (which is the square of the standard deviation).

If, in equation (2), the variable A represents one of the velocity components (U,V,W in the x,y,z direction, respectively), then $\overline{AC}$ is the total flux of C and the second term at the right-hand side of equation (2) represents a turbulent flux of C. For instance, $\overline{uc}$ and $\overline{wc}$ are the horizontal and vertical turbulent fluxes of the variable C, respectively. Here u and w are the turbulent fluctuations of the horizontal and vertical velocities. Near the surface, the mean vertical wind $\overline{W}$ is usually small, and thus the total vertical fluxes are normally dominated by the turbulent contributions. However, important exceptions occur over heterogeneous and sloping terrain.

Turbulent fluxes and related variables are directly influenced by the vertical variation of temperature, humidity, and wind (see below). Here the variation of temperature in the atmospheric boundary layer plays an important role in particular. Since pressure decreases with altitude, air parcels, which are forced to rise (sink), do expand (compress). According to the first law of thermodynamics, a rising (sinking) parcel will cool (warm) if there is no additional energy source such as condensation of water vapor. Then, this is called a *dry adiabatic process*.

It can be shown that in the atmospheric boundary layer, the temperature (T) variation with height for a dry adiabatic process is $\mathrm{d}T/\mathrm{d}z = -g/C_\mathrm{p}$ (where g is the gravity constant and C_p is the specific heat for air at constant pressure). The value for g/C_p is approximately 1 K per 100 m. An atmospheric layer that has such a temperature variation with height is called *neutral for dry air* (at least when there is no convection arising from other levels). In that case, $\Theta = T + (g/C_\mathrm{p})z$ is constant, where Θ is called the *potential temperature* (note that the previous definition for potential temperature is not accurate above the boundary layer). Since air normally contains water vapor and because moist air is lighter than dry air, we have to correct for the influence of this on vertical motions. Consequently, a virtual potential temperature is defined as $\Theta_v = \Theta\,(1 + 0.61q)$, where q is the specific humidity (defined as the mass of water vapor per unit mass of moist air).

In a neutral layer with constant Θ_v, vertical motions of moist (not saturated) air can maintain themselves. If the virtual potential temperature of the atmospheric layer increases with height, vertical displacements are suppressed. This is called a *stable condition* (or "inversion"). On the other hand, when the virtual potential temperature decreases with height, vertical fluctuations may be accelerated. Consequently, this is called an *unstable condition*. Thus, in considerations with turbulent fluctuations and atmospheric stability, we have to deal with the (virtual) potential temperature and not with the actual temperature.

THE ATMOSPHERIC BOUNDARY LAYER OVER LAND

The mean structure of the ABL over land depends strongly on the surface fluxes on scales of 5 km or larger. Figure 2 (after Stull, 1988), provides the typical, idealized, mean vertical profiles for temperature T, potential temperature Θ, specific humidity q, in addition to the horizontal wind M (defined by $M^2 = U^2 + V^2$). These profiles apply for an atmospheric boundary layer over land in clear-sky conditions in the afternoon and around midnight, respectively.

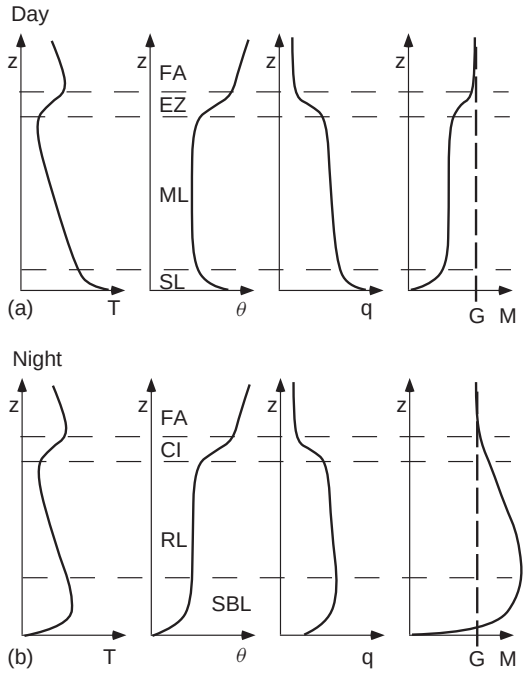


Figure 2 Idealized vertical profiles of mean variables in the atmospheric boundary layer (ABL) over land in fair weather (Stull, 1988, with kind permission of Springer Science & Business Media). In (a), SL refers to the surface layer (lowest 10% of the ABL), ML is the mixed layer, EZ is the entrainment zone, FA is the free atmosphere. In (b), SBL is stable boundary layer, RL is residual layer, CI is the capping inversion. For further explanation, see text and Figure 3

Note that in the free atmosphere the horizontal wind is mostly a result of the acting of the larger-scale pressure differences and the Coriolis force due to the rotation of the earth (but other effects may play a role as well). The resulting wind is known as the *Geostrophic* wind, and is indicated with G in Figure 2 (dashed line). In the daytime boundary layer, the actual wind is smaller due to surface friction, while at clear nights the actual wind away from the surface may be substantially stronger than G due to inertial effects (resulting in the so-called low level jet).

An idealized "clear-sky" picture for the temporal variation of the boundary layer over land is given in Figure 3 (after Stull, 1988). Here the arrows with local time indications refer to the day and nighttime cases of Figure 2. The depth of the dry ABL can vary over land between tens of meters during night up to kilometers during daytime in clear-sky conditions. Over sea, the depth is often only a few hundred meters and rather constant on the timescale of a day. In overcast situations, and in cases with relatively high wind speeds (say over $6\,m\,s^{-1}$ at the 10-m level), the diurnal cycle in the ABL over land is relatively small (as over sea). So, realize that Figures 2 and 3 are not generally valid and that reality is more complex and less ideal!

In any case, the temporal variation of the boundary layer and its characteristics over land is, in particular, substantial in cases with strong diurnal variation of solar incoming radiation (as in spring and summer time) and in cases with nighttime cooling at the land surface (when clear skies and low winds prevail). During daytime, the turbulent boundary layer grows into the nonturbulent "free atmosphere" (indicted as FA in Figure 1). At night, the turbulent part of the stable boundary layer (SBL) may only extend up to a few hundred meters or less (the lowest dashed line in the lower figure).

Figure 2 indicates that the ideal boundary layer during daytime shows a three-layer structure: an unstable "surface layer (SL)", a "well-mixed layer (ML)", and the "entrainment zone (EZ)". The surface layer refers to the lowest 10% of the ABL, in which the turbulent fluxes are rather constant with height but in which the mean profiles normally show the strongest variation with height due to the presence of the surface. In contrast, in the "well-mixed layer (ML)" the (virtual) potential temperature and wind profiles are rather uniform with height due to the strong turbulent mixing, although other variables may still show significant structure. The latter is influenced by the exchange processes in the entrainment zone (EZ), in which turbulence acts to exchange heat, momentum, water vapor, and trace gasses between the boundary layer and the free atmosphere in dependence on the mean values inside and above the boundary layer. The mean values above the boundary layer may vary strongly in time and space depending on the actual larger-scale weather patterns.

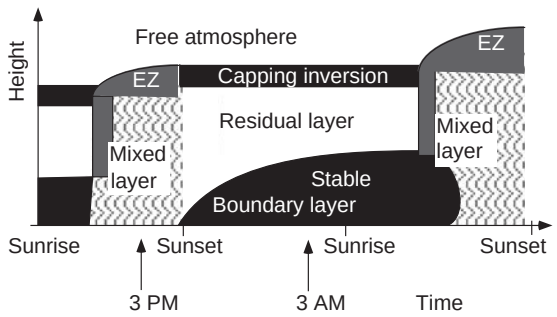


Figure 3 Idealized diurnal evolution of the atmospheric boundary layer over land in fair weather (Stull, 1988, with kind permission of Springer Science & Business Media). The different layers are also indicated in Figure 2

During nighttime, often, the vertical structure of the previous day persists above the SBL. As such, a "residual layer (RL)" with sporadic turbulence (remaining from the previous day) can be identified as well as a "capping inversion (CI)". Also, a surface layer can still be identified as the lowest 10% of the SBL, but then the surface layer is rather shallow in terms of actual height for the ideal case (therefore not shown in the Figures 2 and 3).

Overall turbulence in the ABL is mainly due to the mechanical turbulence by vertical wind shear and convection. Most of the atmosphere above the ABL is not turbulent, although turbulence can occur throughout the whole atmosphere. For instance, cumulus-type clouds, which may grow into thunderstorms, are always turbulent through convection produced by the heat released due to the condensation of water vapor. Turbulence can also occur in clear air above the ABL; most of this is produced in layers of strong vertical wind shear at the boundary between air masses (so-called Clear-Air Turbulence).

Because of the mixing capacity of turbulence, modeling the ABL is also relevant for many practical applications. For instance, chimney plumes are diluted and spread over larger volumes than they would be without turbulence. Turbulent fluctuations in the horizontal motions during severe storms can be fatal to tall buildings or bridges, particularly if resonance (e.g. forcing of a system at its natural frequency) occurs.

ATMOSPHERIC BOUNDARY-LAYER MODELING

The challenge of modeling the atmospheric boundary layer is the prediction of the temporal variation of the vertical and horizontal structures in response to the influence of the major processes acting in the atmosphere and at the earth's surface. As such, the governing equations have to be integrated. In practice, the variables are split into "mean" larger-scale motions and smaller-scale fluctuations as in

equation (1). Inserting this into the basic equations and after averaging this provides a set of equations for the behavior of the larger-scale (mean) variables. The larger-scale variables are then used explicitly in atmospheric models. This can be demonstrated as given below (after the treatments in text books as Stull, 1988; Garratt, 1992).

The general character of any of the budget equations dealing with atmospheric motions is

$$\frac{DC}{Dt} = S_i \tag{3a}$$

Here S_i represents the subsequent sources and sinks for the variable C (such as radiation or chemistry effects). The notation DC/Dt represents the total rate of change for the variable C by local changes ($\partial/\partial t$), and changes transported with the fluid motion in the three directions. As such, we have

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} + V\frac{\partial C}{\partial y} + W\frac{\partial C}{\partial z} = S_i \tag{3b}$$

(Recall that U,V,W are the wind speed components in the three directions x,y,z respectively).

If in the atmospheric motion, each variable is split into a mean component and a fluctuation, then equation (3b) provides after Reynolds-averaging, some algebraic manipulations and simplifying assumptions, a budget equation for the mean variable $\overline{C}$. This reads as

$$\begin{aligned}\frac{D\overline{C}}{Dt} &= \frac{\partial \overline{C}}{\partial t} + \overline{U}\frac{\partial \overline{C}}{\partial x} + \overline{V}\frac{\partial \overline{C}}{\partial y} + \overline{W}\frac{\partial \overline{C}}{\partial z} \\ &= \overline{S_i} - \frac{\partial \overline{uc}}{\partial x} - \frac{\partial \overline{vc}}{\partial y} - \frac{\partial \overline{wc}}{\partial z}\end{aligned} \tag{4}$$

We may note that in the derivation of equation (4), single terms representing fluctuations have disappeared (as above in equation 2). However, terms involving the product of two fluctuations remain.

Thus, because the basic equations are nonlinear, the budget equations for the mean variables contain terms involving smaller-scale motions. The latter terms are of the form of a divergence of fluxes produced by such motions in the three directions and appear as the last three terms in equation (4). These motions are said to be subgrid and, consequently, closure formulations or parameterizations are needed to introduce mixing by the smaller-scale (subgrid) motions into the equations for the larger-scale motions (as resolved by the model). Note that additional terms may also appear in equation (4) when the source or sink term S_i incorporates nonlinear effects (such as in the case of clouds and chemistry). These additional factors are not considered here for simplicity.

A special and simple form of equation (4) arises for horizontally homogeneous conditions. In such cases, the terms including horizontal derivatives are negligible. If in addition the mean vertical wind is small and if there are no other sources and sinks, then equation (4) provides

$$\frac{\partial \overline{C}}{\partial t} = -\frac{\partial \overline{wc}}{\partial z} \tag{5}$$

This equation is known as the one-dimensional, vertical "diffusion" equation. It shows that the local time rate of change for the mean of a variable (such as temperature or wind) at a certain height is only given by the divergence of the turbulent (e.g. corresponding heat or momentum) flux in the vertical direction at that height. As such, information on the turbulent flux may produce a local forecast of the variation of a mean variable (but only under the simplifications mentioned). Often, this equation is used for illustrative purposes in comparison with observations on so-called Golden days. However, normally, the other terms in equation (4) are also relevant, in particular, the terms with mean wind speed (the so-called advection terms). This means that in general the budget equations for momentum, heat, and the various scalars are closely coupled in any atmospheric model.

To solve the budget equation (4) for all the mean atmospheric variables involved, the terms involving turbulent fluxes need to be provided or "parameterized". As mentioned before, this means that the fluxes need to be expressed in terms of available mean model quantities, both in the atmosphere and at the surface. Note that the surface fluxes enter as boundary conditions when solving the budget equations for all the relevant mean variables. The atmospheric model equations are integrated starting with proper initial values (taken from observations). New values are then calculated for the following time step and so on.

The most frequently used parameterization for environmental and atmospheric models is known as first-order closure, often also called *K-theory*. In this theory, it is assumed that the flux $\overline{wc}$ of a variable C in the vertical direction z is down the vertical gradient of the mean concentration of C per unit mass. Thus,

$$\overline{wc} = -K_c\frac{\partial \overline{C}}{\partial z} \tag{6}$$

Here, K_c is known as the *eddy-diffusivity* or mixing coefficient for the variable C. Similarly, the horizontal fluxes can be represented in terms of horizontal gradients. Note that the corresponding eddy-diffusivities typically are not constant, but that they generally depend on properties of the flow and the variable of interest. This also means that normally no analytic solutions are possible and numerical methods must be used.

In atmospheric boundary layers with strong atmospheric convection, the turbulent flux of a conserved quantity is

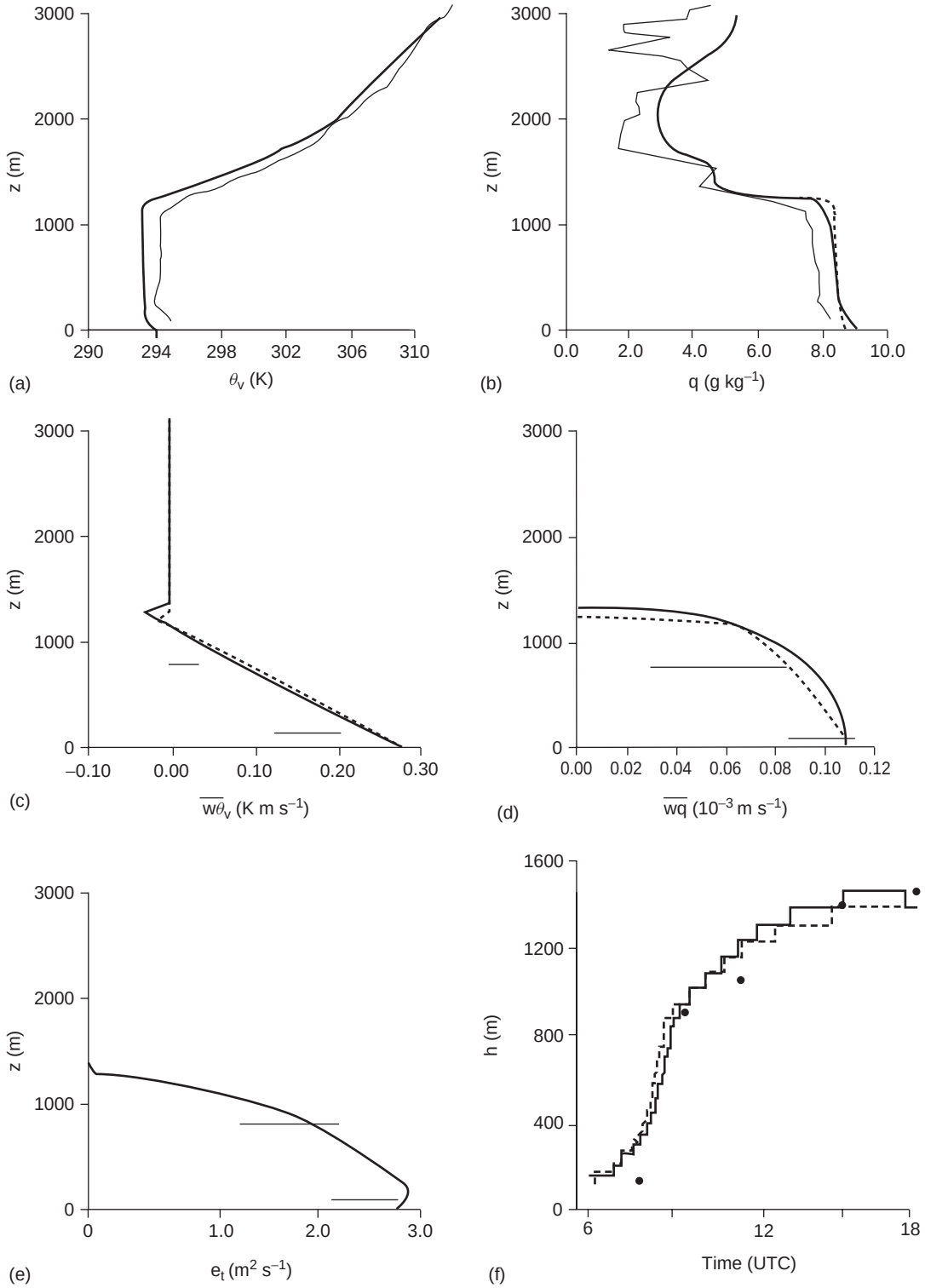


Figure 4 A comparison of observations made on 8 July 1986, in the ''HAPEX-MOBILHY'' field experiment (thin lines) with model outputs for the atmospheric boundary layer in fair weather using two different turbulence parameterizations (Cuxart *et al.*, 1994, with kind permission of Springer Science & Business Media). The models are based on a ''1.5 order closure'' (thick lines) and a ''transilient turbulence closure'' (dashed lines). The results are for mean virtual potential temperature (a), specific humidity (b), the vertical virtual potential heat flux (c), the vertical specific humidity flux (d), the turbulent kinetic energy (e), and the thickness of the turbulent boundary layer (f), The horizontal thin lines in (c), (d), and (e) indicate the range of the corresponding observations, and the dots in (f) refer to observations for the boundary-layer depth

typically not proportional to the local gradient alone as predicted by equation (6). In fact, in a large part of the ABL, the mean gradients are small in conditions with dry convection, in particular, for potential temperature (see Figure 2). Then, the fluxes depend mostly on the mixing characteristics of the larger turbulent motions ("eddies") across the depth of the ABL. Theories are available (see e.g. Stull, 1988; Garratt, 1992, or Holtslag, 2002 for overviews), which have modified K-theory to allow for the influence of convection. To reproduce larger eddy mixing, additional terms on the right-hand side of equation (6) can be incorporated. This has important impacts on reproducing the global land surface climate (e.g. Holtslag and Boville, 1993).

Figure 4 gives an example of a short-range forecast produced by an atmospheric boundary-layer model in comparison with observations (after Cuxart *et al.*, 1994). Here the observations are taken from the "HAPEX-MOBILHY" experiment for July 8, 1986, in France. In this case, a composite of observations for the surface fluxes was used to act as surface boundary conditions for the solving of the budget equations (4) for temperature and humidity, and the initial profiles and advection terms were taken from a meso-scale analysis (see Cuxart *et al.*, 1994). The model outputs for the atmospheric boundary layer apply for fair weather using two different turbulence parameterizations. The models are based on two alternative parameterizations, namely, a so-called 1.5 order closure (e.g. Garratt, 1992) (thick lines) and a so-called transilient turbulence closure (e.g. Stull, 1988) (dashed lines). The results are for the mean variables, the vertical fluxes, the turbulent kinetic energy (defined by $e_t = (\overline{u^2} + \overline{v^2} + \overline{w^2})/2$), and the depth of the turbulent boundary layer. It is seen that overall the agreement with observations is rather good in this case, in particular, because of the details known in the atmospheric forcing conditions and because the surface fluxes were prescribed. However, in general, the agreement may be much less due to lack of proper representation of atmospheric and land surface factors and for long integration times (see below).

The atmospheric model equations can also be applied on much smaller spatial and temporal scales than discussed here, for instance, by using vertical and horizontal grid elements of 10 to 100 m, and time steps of seconds only. It is important to realize that in such cases a significant part of the turbulent fluctuations are resolved by the model equations. This type of modeling is nowadays known as "Large-Eddy Simulation (LES)". This has become a powerful and popular tool in the last decade to study turbulence in clear and cloudy boundary layers under well-defined conditions (e.g. Wyngaard, 1997). It is important to realize that in the case of LES the simplifying assumptions leading to equation (2) are normally not valid.

SURFACE FLUXES OVER LAND

Near the surface, the turbulent fluxes for sensible and latent heat are given by $H = \rho C_p \overline{w\theta_0}$ and $LE = \rho L \overline{wq}_0$. These fluxes (or more specifically flux densities) represent the energy per time and unit area of the involved variables at the surface. Here ρ is the density of the air (in $\mathrm{kg\,m^{-3}}$), C_p is the specific heat at constant pressure (in $\mathrm{J\,K^{-1}}$ kg), L is latent heat of vaporization (in $\mathrm{J\,kg^{-1}}$), and the subscript (o) indicates a surface value (note that here specific humidity q should be given in $\mathrm{kg\,kg^{-1}}$) (*see* **Chapter 201, Land-Atmosphere Models for Water and Energy Cycle Studies, Volume 5**).

Over land, the surface heat fluxes are related through the available energy at the surface

$$H + LE = Q_* - G \tag{7}$$

where Q_* is the surface net radiation and G is the soil heat flux (all in $\mathrm{W\,m^{-2}}$). Here the net radiation is given by the sum of downward ($K^{\downarrow}$) and reflected ($K^{\uparrow}$) short wave, and downward ($L^{\downarrow}$) and emitted ($L^{\uparrow}$) long-wave radiation components at the surface (*see* **Chapter 39, Surface Radiation Balance, Volume 1**). This reads as

$$Q_* = K^{\downarrow} - K^{\uparrow} + L^{\downarrow} - L^{\uparrow} \tag{8}$$

As an example, Figures 5 and 6 give observations and model results of the diurnal variation for the surface fluxes and the related radiation and atmospheric variables over land. The findings apply for the "Golden day" of May 31, 1978, during daytime over a grass-covered surface at Cabauw, The Netherlands (after Ek and Holtslag, 2004). In fact, the observations in Figure 5 are fairly typical for a grass-covered surface in mid-latitude summertime conditions with sufficient soil moisture. In such cases, most of the available net radiation is distributed to the latent heat flux. Consequently, the Bowen ratio *H/LE* is about 35% and the evaporative fraction $LE/(H + LE)$ is about 70%. The soil heat flux is typically small for this type of vegetated surfaces during daytime (order 10% of net radiation Q_*). However, these ratios may be significantly different for other values of soil moisture and for other surface conditions such as in the case of tall plants, forests, and urban landscapes (e.g. Bonan, 2002; Kabat *et al.*, 2004). For instance, in a desert, the latent heat flux (or evaporative fraction) is typically small and the soil heat flux is relatively large (e.g. Heusinkveld *et al.*, 2004).

In Figure 5, also several model results are indicated of a land surface (LS) scheme. The simplest case is given by a LS scheme in stand-alone mode. In such a case, the observations of atmospheric temperature, specific humidity, wind speed, the (incoming and reflected) solar radiation as well as the downward long-wave radiation of Figure 6 were

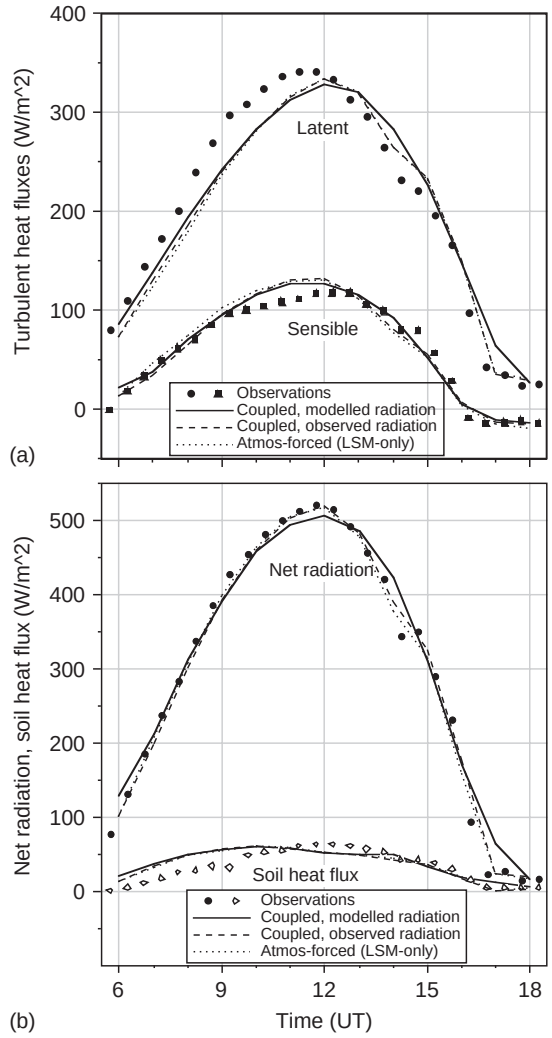


Figure 5 The diurnal variation of the energy balance components for May 31, 1978, at Cabauw, NL (Ek and Holtslag (2004). © 2004 American Meteorological Society)

used as input to calculate the surface fluxes. Consequently, the case is referred as the "atmospheric forcing" case. To calculate the surface fluxes, equations (7) and (8) are utilized in combination with the usual soil diffusion equation, and a "transfer equation" for the surface fluxes.

The transfer equation reads similar as equation (6), but rather than the atmospheric gradient the difference of atmospheric and surface properties is used for the quantity of interest. For heat this reads as (e.g. Beljaars and Holtslag, 1991)

$$\overline{w\theta_0} = \beta_t u_{*0}(\Theta_0 - \Theta_a) \tag{9}$$

Here Θ_0 and Θ_a are the values of the (potential) temperature at the surface and in the air, respectively; β_t is a transfer parameter and u_{*0} is the well-known surface friction velocity. The latter is directly related to the surface momentum fluxes (or stresses), as defined by

$$u_{*0}^4 = ((-\overline{uw_0})^2 + (-\overline{vw_0})^2) \tag{10}$$

It is important to realize that near the surface the average wind must vanish because the mean wind is zero at the earth's surface. On the other hand, we know from observations that the fluxes of heat, momentum, and trace gasses are nonzero. Consequently, equation (9) represents the "effective" surface flux of sensible heat $\overline{w\theta_0}$ due to the combined effect of molecular diffusion and turbulence at the surface. As such, β_t depends on surface characteristics and the atmospheric state near the surface. Similar relations apply for the latent heat flux and other surface fluxes. In any case, the surface value for the involved quantity needs special consideration as well as the formulation of the transfer parameter β_t. The latter is normally part of the land surface scheme (e.g. Beljaars and Holtslag, 1991; Holtslag and Ek, 1996; Bonan, 2002).

Figure 5 shows that the surface fluxes can be realistically modeled in the "off-line" case. The next step is to couple the land surface scheme to the ABL model. Here the ABL model of Troen and Mahrt (1986) and Holtslag and Boville (1993) is used (see Ek and Holtslag, 2004 for a list of recent updates). It is seen that the coupled model results are very similar to the "off-line" results for this case, indicating realistic behavior of both the land surface and boundary-layer scheme for the present "ideal" case Figure 6. Using a sophisticated land surface scheme allows the land–atmosphere system the freedom to respond interactively with the ABL where many processes and important feedback mechanisms are represented (Figure 1). However, in general, the correct modeling is not at all trivial, certainly not for the more complex situations that occur in reality (see Bonan, 2002 for more information on this issue).

In any case, to understand and to correctly model boundary-layer climates as well as the feedbacks in the coupled land–atmospheric system, a realistic model of both the atmospheric boundary-layer (ABL) and the land surface (LS) is needed.

LAND SURFACE AND BOUNDARY-LAYER INTERACTIONS

On the basis of the coupled land surface – atmospheric boundary-layer model set up as discussed above – we can explore the interaction of the land surface with the ABL and the effect on boundary-layer cloud development. We focus on the role of soil moisture and, consequently, we make a series of model runs (a "reference" set) where we change the soil moisture from dry to wet conditions with initial conditions and forcing the same as in the coupled model

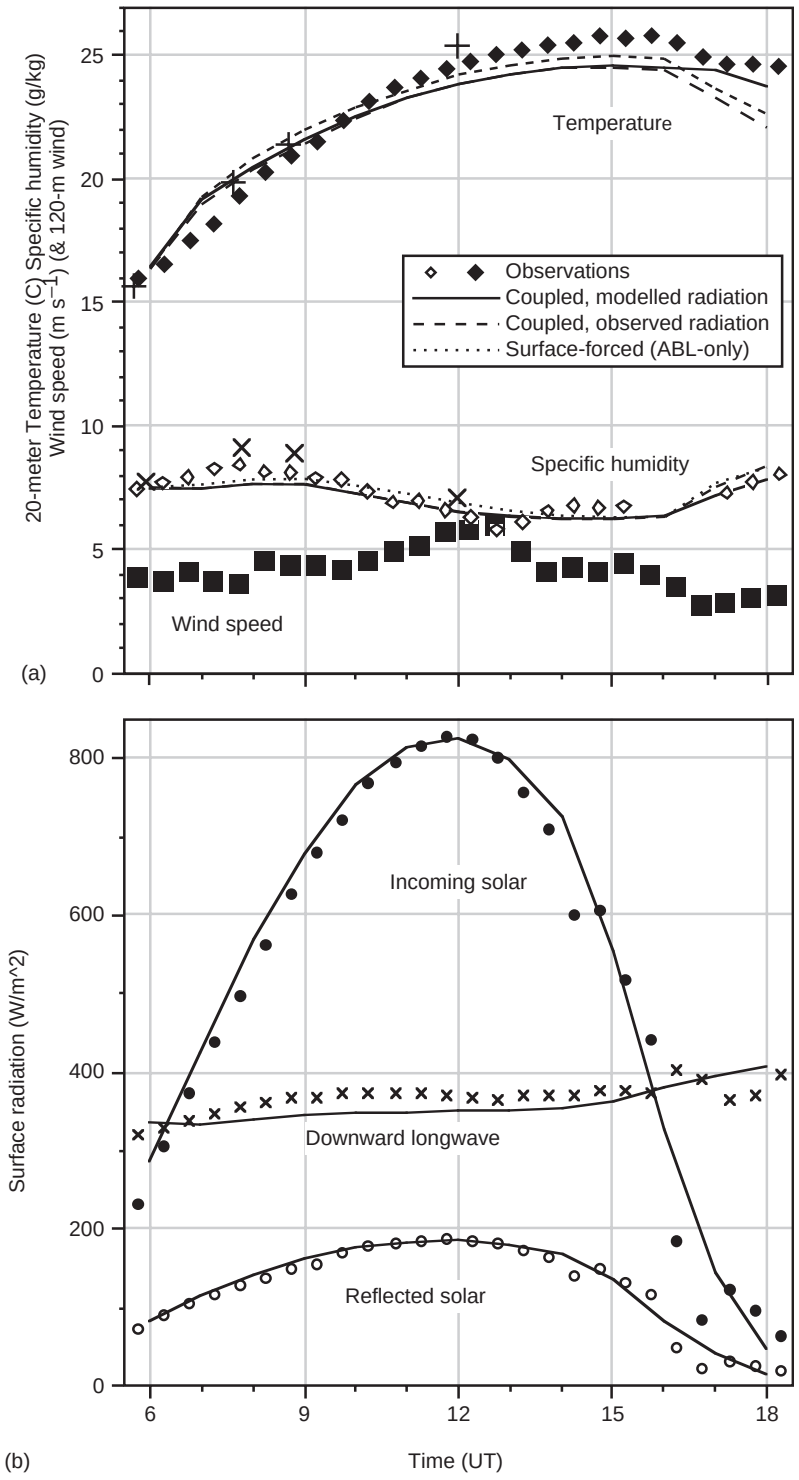


Figure 6 The diurnal variation of the radiation components and atmospheric variables (T, q, U) for May 31, 1978, at Cabauw, NL (Ek and Holtslag (2004). © 2004 American Meteorological Society)

runs (as above). As such, we vary soil moisture from below the wilting point (quite dry) to near saturation (quite wet).

When the initial soil moisture is decreased from intermediate soil moisture values (close to the Cabauw observations with volumetric soil moisture ≈0.43) to below the wilting point, the ABL cloud cover decreases to zero (Figure 7). This confirms intuition. However, as we increase the initial soil moisture from intermediate soil moisture values to near

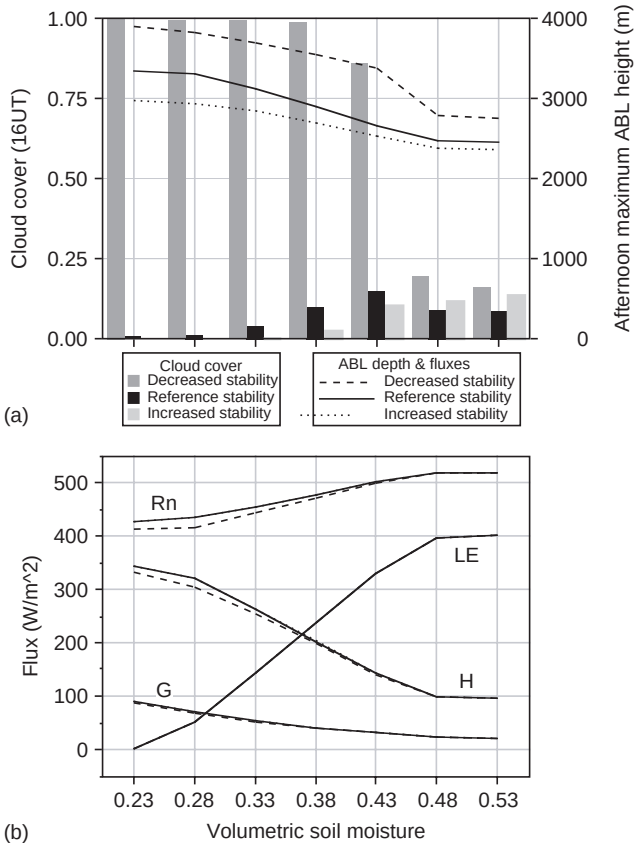


Figure 7 The impact of variation in volumetric soil moisture on the ABL depth and cloud cover (a), and components of the surface energy budget (b) for May 31, 1978, at Cabauw, NL (Ek and Holtslag (2004). © 2004 American Meteorological Society). The different sets of model runs refer to a reference case, and cases with increased and decreased stability above the ABL (see text for full explanation)

saturation, ABL cloud cover decreases slightly, a somewhat counterintuitive result. Certainly, there are a number of processes that account for this behavior, that is, interactions between the land surface, atmospheric boundary layer (including ABL clouds), free atmosphere, and initial ABL conditions (Figure 1).

Before attempting an explanation of this response, we also examine the role of atmospheric stability (that is the strength of the capping inversion) above the ABL. Note that the atmospheric stability above the ABL, in general, has a strong influence on the boundary-layer depth and on its temporal variation (see Garratt, 1992; Stull, 1988 for more background on this). To illustrate this effect, we make two additional sets of model runs as above, namely, a run in which we prescribe an increased atmospheric stability above the afternoon boundary-layer top and another run with decreased atmospheric stability. We then examine the resulting afternoon ABL depth and fractional cloud cover and the mid-day surface energy budget as it changes with changing prescribed initial soil moisture (see Figure 7).

The set of model runs with stronger atmospheric stability have a shallower ABL depth than the reference set and less cloud cover for drier soils, with increasing cloud cover for model runs with increased soil moisture (Figure 7). However, in great contrast, the set of model runs with weaker atmospheric stability above the ABL have a deeper ABL depth (as one would expect) and yet a much greater cloud cover for drier soils, with decreasing cloud cover for increasing soil moisture. These effects are not easily expected beforehand.

To further understand the role of soil moisture and other factors on ABL cloud development, Ek and Holtslag (2004) examine a useful new equation for relative humidity (*RH*) tendency at the ABL top:

$$\frac{\partial RH}{\partial t} = \frac{(Q_* - G)}{\rho L h q_s}[ef + ne(1 - ef)] \qquad (11)$$

Here $Q_* - G$ is the available energy at the surface, h is ABL depth, and q_s is saturation-specific humidity just below the ABL top (again ρ is air density and L is latent heat of vaporization). The surface evaporative fraction (the surface energy available for evaporation), *ef*, is already defined above. Furthermore, *ne* reflects the direct effects of nonevaporative processes on relative humidity tendency (see Ek and Holtslag, 2004). Basically, *ne* consists of three factors: ABL-top dry-air entrainment (a negative contribution to the tendency of relative humidity at the ABL top), the boundary-layer growth (a positive contribution), and boundary-layer heating through surface warming and ABL-top warm-air entrainment (a negative contribution).

From equation (11), we see that the relative humidity tendency is proportional to available energy and inversely proportional to ABL depth and temperature (via saturation-specific humidity), while the sign of the relative humidity tendency is determined by the sign of $ef + ne\ (1 - ef)$. On examining equation 11, it is apparent that the direct role of *ef* is to increase the ABL-top relative humidity, while the indirect role of surface evaporation (via reduced surface heating, and diminished ABL growth and entrainment) is found in the expression $ne\ (1 - ef)$. Figure 8 shows how $ef + ne\ (1 - ef)$ depends on *ef* versus *ne*, where $ef + ne\ (1 - ef)$ is simply the relative humidity tendency, $\partial RH/\partial t$, normalized by the available energy term $(Q_* - G)/\rho L h q_s$. In Figure 7 also, some data points are given for the Cabauw case (labeled with their time of occurrence).

For the case where $ne < 1$, $\partial RH/\partial t$ increases as the evaporative fraction (ef) increases, confirming intuition. (For the range $0 < ne < 1$, $\partial RH/\partial t > 0$ and increases with increasing *ef*, while for $ne < 0$, $\partial RH/\partial t > 0$ only when *ef* exceeds some threshold value that increases for increasingly negative values of *ne*). Here soil moisture acts to increase ABL-top relative humidity tendency and thus increases the

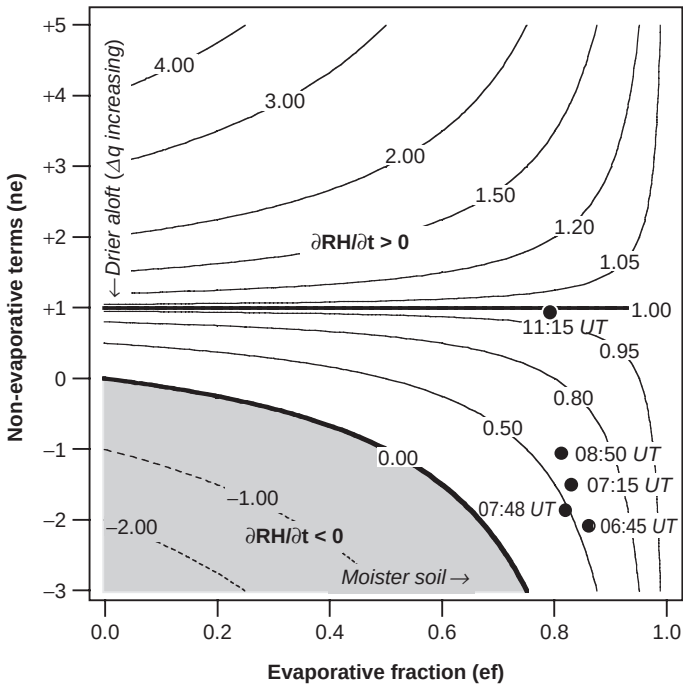


Figure 8 Relative humidity tendency (normalized by the available energy term), with dependence on evaporative fraction versus nonevaporative terms (Ek and Holtslag (2004). © 2004 American Meteorological Society). See text for full description

probability of ABL cloud initiation, given a sufficient initial ABL relative humidity. This is the case when the soil is sufficiently moist for unrestricted surface evaporation, and the environment above the ABL is not too dry and atmospheric stability is not too weak.

For the case where $ne > 1$, $\partial RH/\partial t$ increases as ef decreases (which is somewhat counterintuitive) so that here soil moisture acts to limit the increase of ABL-top relative humidity and thus decreases the probability of ABL cloud initiation. This is the case when the environment above the ABL is again not too dry but atmospheric stability is rather weak, so for drier soils, surface evaporation is lower with boundary-layer growth less restricted than with moister soils. Note that the largest values of $\partial RH/\partial t$ are achieved for $ne > 1$, suggesting that the greatest potential for ABL cloud initiation is not over moist soils, but rather over dry soils with weak stability (and air not too dry above the ABL).

We may note that the outcome of equation 11 (as presented in Figure 8) agrees well with the output of the coupled model (confirmed by more than a thousand model runs), as long as the ABL is sufficiently well mixed (as reflected by rather uniform potential temperature and large heat fluxes). For the set of model runs with increased (stronger) atmospheric stability, ABL depth is shallower (as one would expect), and since $ne < 1$, there is a decrease in relative humidity tendency at the ABL-top. This implies less cloud cover for drier soils (Figure 8), with increasing cloud cover for increasing soil moisture ($ne \approx 0$). In contrast, for the set of model runs with decreased (weaker) atmospheric stability, ABL depth is deeper (as one would expect), and yet since $ne > 1$ there is an increase in ABL-top relative humidity tendency. This implies more cloud cover for drier soils, with decreasing cloud cover for increasing soil moisture ($ne \gg 1$), as was suggested in the relative humidity tendency development described above.

The indicated findings are qualitatively consistent with Ek and Mahrt (1994) for HAPEX-MOBILHY data (summer 1986, Southwest France), which found that a day with strong atmospheric stability above the ABL and a large observed evaporative fraction (via higher soil moisture) gave a similar mid-day relative humidity at the ABL-top as a case nine days later, with weaker atmospheric stability and soil moisture that had decreased by 20%.

SUMMARY

In this article, we have focused on atmospheric boundary-layer climates over land. Special emphasis was given to the basic characteristics and approaches in use for modeling and parameterization of the ABL in relation with the land surface. Examples are given for "Golden days" in the "HAPEX-MOBILHY" experiment in France (after Cuxart *et al.*, 1994), and at Cabauw, The Netherlands (after Ek and Holtslag, 2004). It is shown that both the ABL and

the coupled soil-vegetation-atmosphere system can be well represented in the well-defined cases examined. At the same time, it is realized that reality is often much more complex (e.g. Bonan, 2002; Kabat *et al.*, 2004).

The role of soil moisture in the development of ABL clouds is explored via model runs and an analytical development. This is done in terms of tendency equation of relative humidity (*RH*) at the ABL-top, which involves a number of land-atmosphere interactions (Ek and Holtslag, 2004). It is shown that the effect of soil moisture is to increase ABL-top *RH* tendency and thus potential for ABL cloud formation (confirming intuition), but only if the stability above the ABL is not too weak (and given sufficient initial *RH* in the ABL and air above the ABL that is not too dry). Alternately, for weak stability above the ABL, drier soils yield a greater ABL-top *RH* tendency and thus potential for ABL cloud formation (somewhat counter-intuitive), where in this case soil moisture acts to limit the increase of relative humidity at ABL-top. Then, the largest values of ABL-top *RH* tendency are predicted over dry soils. The new relative humidity tendency equation presented (after Ek and Holtslag, 2004) may provide a useful quantitative framework for hydrometeorological studies involving land surface interaction with the atmospheric boundary layer.

Acknowledgments

Part of the material in this article is based on earlier works by the authors, notably Holtslag (2002) and Ek and Holtslag (2004).

REFERENCES

Beljaars A.C.M. and Holtslag A.A.M. (1991) Flux parameterization over land surfaces for atmospheric models. *Journal of Applied Meteorology*, **30**, 327–341.

Betts A.K. (2000) Idealized model for equilibrium boundary layer over land. *Journal of Hydrometeorology*, **1**, 507–523.

Bonan G. (2002) *Ecological Climatology; Concepts and Applications*, Cambridge University Press: p. 678.

Cuxart J., Bougeault P., Lacarre P., Noilhan J. and Soler M.R. (1994) A comparison between transilient turbulence theory and the exchange coefficient model approaches. *Boundary-Layer Meteorology*, **67**, 251–267.

Garratt J. (1992) *The Atmospheric Boundary Layer*, Cambridge University Press: p. 316.

Ek M. and Mahrt L. (1994) Daytime evolution of relative humidity at the boundary-layer top. *Monthly Weather Review*, **122**, 2709–2721.

Ek M.B. and Holtslag A.A.M. (2004) Influence of soil moisture on boundary cloud development. *Journal of Hydrometeorology*, **5**, 86–99.

Findell K.L. and Eltahir E.A.B. (2003) Atmospheric controls on soil moisture-boundary layer interactions. Part I: framework development. *Journal of Hydrometeorology*, **4**, 552–569.

Holtslag A.A.M. and Boville B.A. (1993) Local versus nonlocal boundary-layer diffusion in a global climate model. *Journal of Climate*, **6**, 1825–1842.

Holtslag A.A.M. and Ek M. (1996) Simulation of surface fluxes and boundary layer development over the pine forest in HAPEX-MOBILHY. *Journal of Applied Meteorology*, **35**, 202–213.

Holtslag A.A.M. (2002) *Atmospheric Boundary Layers: Modelling and Parameterization*, *Encyclopedia of Atmospheric Sciences*, Vol. 1, Holton J.R., Pyle J. and Curry J.A. (Eds.), Academic Press: pp. 253–261.

Heusinkveld B.G., Jacobs A.F.G., Holtslag A.A.M. and Berkowicz S.M. (2004) Surface energy balance closure in an arid region: role of soil heat flux. *Agricultural and Forest Meteorology*, **122**, 21–37.

Kabat P., Claussen M., Diemeyer P.A., Gash J.H., Bravo de Guenni L., Meybeck M., Pielke R.A., Vorosmarty C.J., Hutjes R.W.A. and Lutkemeier S. (Eds.) (2004) *Vegetation, Water, Humans and the Climate: A New Perspective on An Interactive System*, Springer: p. 600.

Margulis S.A. and Entekhabi D. (2001) Feedback between the land surface energy balance and atmospheric boundary layer diagnosed through a model and its adjoint. *Journal of Hydrometeorology*, **2**, 599–620.

Stull R.B. (1988) *An Introduction to Boundary-Layer Meteorology*, Kluwer: Dordrecht, p. 666, (reprinted 1999).

Troen I. and Mahrt L. (1986) A simple model of the atmospheric boundary layer: sensitivity to surface evaporation. *Boundary-Layer Meteorology*, **37**, 129–148.

Wyngaard J.C. (1997) Review of simulation and modeling of turbulent flows. *Bulletin of the American Meteorological Society*, **78**, 1785–1787.

30: Topographic Effects on Precipitation

GEOFFREY L AUSTIN[1] AND KIM N DIRKS[2]

[1] *Department of Physics, The University of Auckland, Auckland, New Zealand*
[2] *Department of Physiology, The University of Auckland, Auckland, New Zealand*

Topography and orography can drastically alter the wind patterns in the atmosphere that in turn can have a major effect on the geographical distribution of precipitation. The processes involved include the upslope enhancement of existing rainfall or the triggering of new rainfall in mountainous regions, and the effect of temperature or humidity contrasts between boundary layer air masses over land and oceanic areas. These processes can result in rainfall patterns ranging from a single triggered thunderstorm to widespread torrential rainfall, as occur when tropical cyclone air masses are pushed up steep coastal mountain ranges. Since the magnitude and distribution of the resulting precipitation pattern can depend critically on the wind speed and direction, as well as the terrain itself, the topographical precipitation processes often determine not only the climatic rainfall patterns of a region but also the interannual variability of this distribution.

INTRODUCTION

Topography leads to a diverse range of modifications to the wind patterns of the Earth's atmosphere. These modifications, in the form of topographically induced circulations, can modify cloud and precipitation-producing processes. The net effect of the topography and orography can be to redistribute and modify the amount of precipitation that would have occurred in the absence of complex terrain and, in some cases, to produce precipitation that would otherwise not have occurred at all.

Processes that modify the airflow can be due to the movement of moist air over or around hills or mountain ranges. In such situations, a variety of circulations and precipitation patterns are possible depending on the stability and humidity of the air and the height and horizontal extent of the topographical barrier. Clearly, if the atmosphere is unstable, the topography may induce convective overturning, whereas if the air mass is stable, then the flow may exhibit lee waves.

The modification of airflows can also be caused by thermal gradients due to changes in the nature of the Earth's surface. Sea-breeze circulations and monsoons are examples of such phenomena. Both of these can also result in triggered convection (convection resulting from potentially unstable air being lifted as a result of topography). A particularly well-studied example of smaller-scale sea-breeze circulations is the development of thunderstorms over Florida (Pielke, 1974; Nicholls *et al.*, 1991), where the sea breeze may provide enough lifting for thunderstorms to develop. An example of a much larger system is the Asian monsoon where large contrasts in temperature develop between the Indian Ocean and the continent, leading to substantial amounts of rainfall (Huang-Hsiung and Liu, 2003).

Probably the most useful accounts of mesoscale circulations, both from observational and theoretical points of view, are contained in Atkinson (1989) and Ray (1986) to which the reader is referred. With the exception of monsoon circulations, all of the thermally and orographically induced circulations described below fit into the category of mesoscale circulations. The development of radar and satellite remote-sensing systems have allowed better depictions of these circulations by viewing cloud and precipitation patterns that were too small for global networks to reveal. These technologies have been supplemented in projects around the world with high spatial-resolution networks of rain gauges and meteorological towers. In situations of complex terrain, such networks often provide the most accurate measures of the spatial distribution of rainfall.

Over the years, a number of experimental programmes have been undertaken in various parts of the world to

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

investigate the effects of mountainous topography on precipitation patterns, following the pioneering work in the United Kingdom South of Wales Experiment (Browning, 1980). Of particular note are the recent GAME project of the Asian Monsoon (**http://www.hyarc.nagoya-u.ac.jp/game/**), investigations into the effect of the Rocky Mountains on rainfall in the Northwestern United States (Klimowski *et al.*, 1998), the effect of the European Alps (Frei and Schär, 1998), and rainfall patterns resulting from the Southern Alps in the South Island of New Zealand (Wratt *et al.*, 1996). It is apparent from these studies that where the scale of the mountains or the stability or wind speed differ, the precipitation-forming mechanisms can be quite different from one another. For example, for South Wales with relatively modest mountains and high humidity, the presence of snow aloft frequently induces large rainfalls by the seeder-feeder process (see later), whereas for other locations this process may be unimportant.

In this chapter, we attempt to describe some of the basic aspects of topographically and orographically induced flows and their impact on precipitation patterns. Some of the experimental programmes carried out to investigate these mechanisms are also described.

TOPOGRAPHICALLY INDUCED FLOW

Land–Sea Breezes

One of the most studied mesoscale circulations is the land–sea breeze system. Land and sea breezes result from thermal discontinuities at the interface between the sea and land and dominate the flow during calm, clear conditions in coastal environments. In the morning, solar radiation heats the land more quickly than the sea, causing the air above it to expand. This results in a slope in the isobars from land to sea, leading to a light flow of air towards the land at the ground surface and seaward flow aloft, along with rising air above the land with subsiding air above the sea, as shown in Figure 1. At night, the reverse flow occurs, though it is generally weaker than during the day, and the vertical extent of the circulation is generally reduced. Sea breezes are strongest under calm wind conditions, clear skies and when the difference between the temperature of the sea surface and the land is at its maximum. For a narrow isthmus of land in the ocean, a convergence zone may develop at the boundary between two opposing sea breezes, resulting in the potential for triggered convection along the convergence zone.

A particularly striking example of such a sea-breeze circulation is that which frequently develops over Florida during the wet summer season. The convergence of moist winds from both the Atlantic Ocean and the Gulf of Mexico results in convection and clouds, often with cloud-free air over the water. If the atmosphere is conditionally

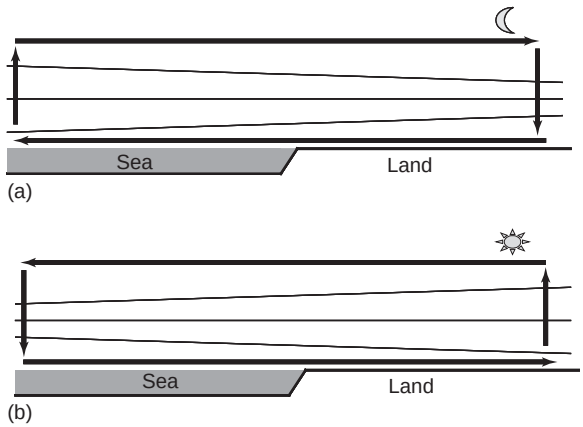


Figure 1 The land-sea breeze circulation. The differential heating of the land and sea results in offshore winds near the surface at night (a) and onshore winds near the surface during the day (b). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

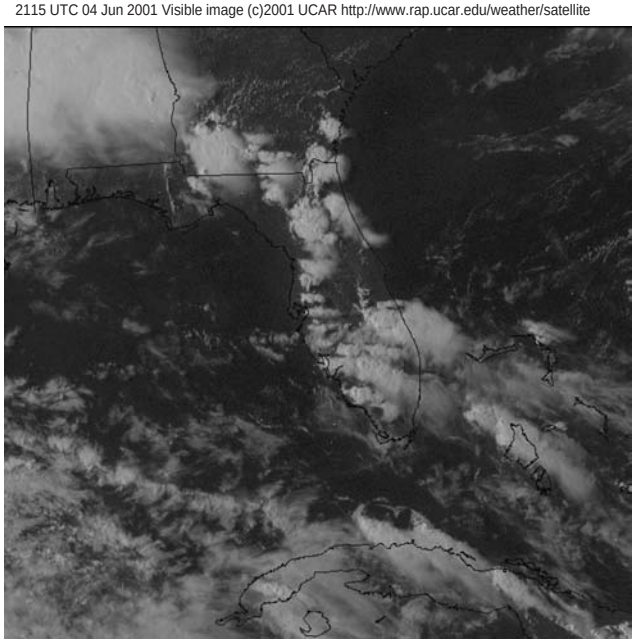


Figure 2 Satellite imagery of thunderstorm development over Florida (The source of this material is the University Corporation for Atmospheric Research (UCAR). © 2002 All Rights Reserved)

unstable, then thunderstorms may be formed by the sea breeze convergence, as originally suggested by Byers and Rodebush (1948). More clear evidence of the development of thunderstorms as a result of the sea breezes over Florida can be seen from satellite imagery, as shown in Figure 2.

Monsoon Circulations

Monsoon circulations are large-scale shifts in seasonal wind patterns caused by differential heating between continents

and the oceans which reverse between summer and winter. In some ways, monsoon circulations are similar to large-scale sea-land breezes that occur on a yearly cycle rather than a daily cycle. The Asian monsoon, the Singapore monsoon, the West African monsoon, and the Arizona monsoon are examples of monsoon systems found around the world. In the case of the Asian monsoon, during the winter, the air mass over South Asia becomes much cooler than the air mass over the Indian Ocean. This results in the development of a large high-pressure system over land, and a clockwise circulation of air that flows out over the Indian Ocean, as shown in Figure 3. The northeasterly winds over southern and eastern Asia generally bring dry air and clear skies. In contrast, during the summer, the air mass over South Asia becomes much warmer than the air mass over the Indian Ocean so a low-pressure system tends to develop over land, leading to moist southwesterly winds from the ocean. The lifting of these air masses as a result of the topography leads to heavy rain and thunderstorms at this time of the year.

The Asian monsoon is the most intense and the most complicated of the monsoons because of the existence of the Himalaya–Tibetan plateau (Chou, 2003). Many studies carried out over the past 30 years have shown the importance of the Tibetan plateau as an elevated heat source (Hahn and Manabe, 1975; He *et al.*, 1987) and the impact this has on the precipitation patterns resulting from thermal triggering. The livelihood of nearly a billion people is dependent on the arrival of monsoon rains, while at the same time the monsoon can result in devastating floods.

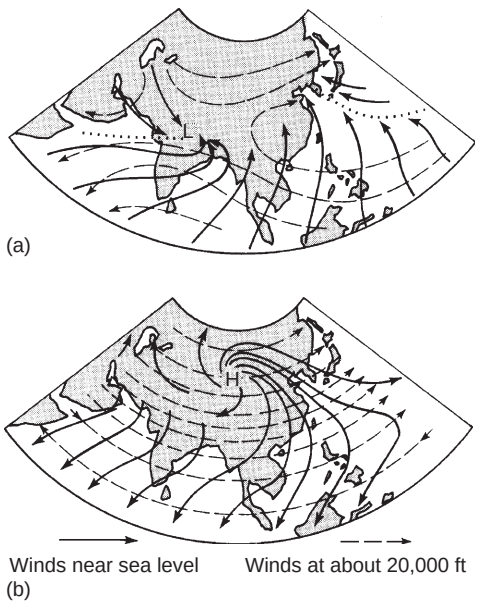


Figure 3 The Asian monsoon circulation (a) during the summer and (b) during the winter (Reproduced from Petterssen, 1969 with premission of the McGraw Hill Companies)

OROGRAPHIC AIRFLOWS

Updrafts can be significantly increased when wind flows are lifted over mountain barriers. In the case of preexisting rainfall, this can result in increased rainfall and the possibility of convective rainfall. The effects of this rainfall-generating process over mountain barriers can be readily recognized in the geographical distribution of rainfall in the form of enhancements on the upwind side of ranges facing moist prevailing winds. An example of this is the enhanced rainfall on the windward side of the Southern Alps of New Zealand often observed under moist northwesterly wind conditions.

Whether or not an air mass is raining before it arrives at a mountain, it will cool as a result of the uplift, and, on reaching saturation, produce an orographic cloud and/or become unstable and trigger convection. In the case of a stable airmass, the orographic cloud will eventually produce precipitation as the cloud droplets combine to give precipitation-sized particles. Early observations and theoretical calculations of this process suggest that it may take as long as 20 min to produce rain in this way. Findeisen (1939) carried out some of the first calculations of the growth of cloud droplets by coalescence while Langmuir (1948) refined the calculations and discussed the role of drop breakup in generating fragments suitable for further growth. Langmuir's calculations were later confirmed by Bowen (1950) using radar observations. If the rain barrier is of modest horizontal extent (less than about 50 km or so), this 20 min or so required for rain-sized drops to be produced from an orographic cloud may be insufficient to generate rain at the surface of the orographic barrier. Microphysical enhancement processes which allow these clouds to rain are discussed below in the section on cloud microphysical processes.

For hills or mountains of modest horizontal extent, several different mechanisms may lead to a modest enhancement or redistribution of preexisting precipitation or the development of precipitation from moist air. However, for the case where a strongly developing tropical cyclone is incident on a similar orographic barrier, the large volumes of warm saturated air that are lifted over the barrier can lead to torrential rainfall.

Houze (1993) describes seven basic types of orographic mechanisms depicted in Figure 4. They are: "(a) The seeder-feeder mechanism of Bergeron (1950, 1968) in which low orographic clouds provide liquid water 'food' to enhance the growth of precipitation particles falling from mid-level clouds; (b) condensation by forced upslope flow; (c) upslope triggering of convection; (d) upstream cloud formation resulting from orographic flow-blocking or vertically propagating gravity waves; (e) thermal triggering by an elevated heat source; (f) lee-side convergence owing to flow around a 3D obstacle; and (g) lee-side enhancement

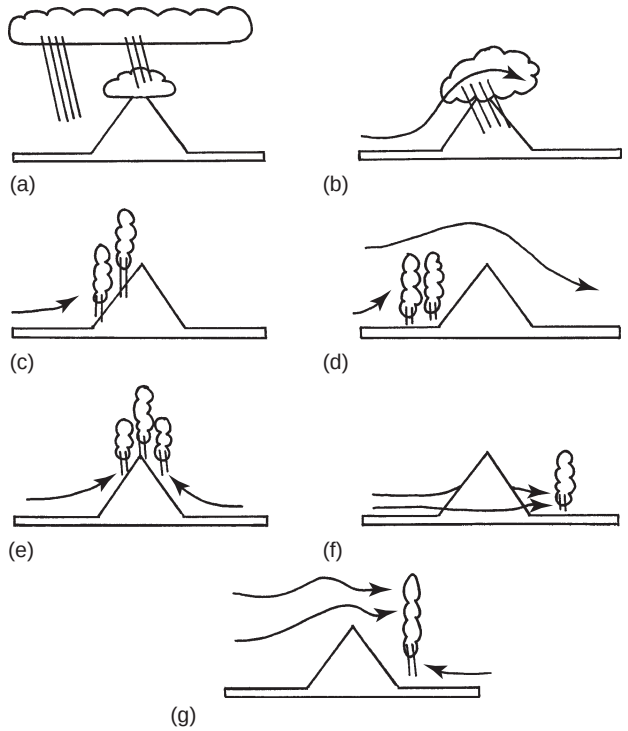


Figure 4 Mountain airflows leading to precipitation and/or precipitation enhancement by (a) the seeder-feeder mechanism, (b) condensation by forced upslope flow, (c) upslope triggering of convection, (d) upstream cloud formation resulting from orographic flow-blocking or vertically propagating gravity waves, (e) thermal triggering by an elevated heat source, (f) lee-side convergence owing to flow around a 3D obstacle, and (g) lee-side enhancement of precipitation by mountain-induced gravity waves upslope convection (Reprint from Houze, 1993 with permission from Elsevier)

of precipitation by mountain-induced gravity waves. In addition to these mechanisms, there is the important effect of concentration of the windward flow into valleys. These convergent flows produce extreme climatological maxima, ... and concentrate the precipitation and runoff to produce extreme flooding events".

In some situations of low hills of 300 m or so, it has been observed that there is a deficit in precipitation upwind of the peak and a relative excess downwind, inconsistent with the predictions of the above-mentioned mechanisms. In this case, it may be that there has not been any enhancement in precipitation as such, and that the observed spatial variability is simply a redistribution of precipitation as a result of a perturbation in the trajectories of raindrops due to a distortion in wind flow. This rainfall redistribution mechanism is discussed in more detail below.

Of particular interest in the area of predictive meteorology is the role of orographic barriers in the formation of cyclones on their lee side and the subsequent precipitation that may result. Petterssen (1956) and Pierrehumbert (1986) have shown this to occur in the lee of the Rocky Mountains and the European Alps, respectively. There appears to be no completely satisfactory theory on the mechanism of lee cyclogenesis, although it clearly involves the modification of a barotropic instability by the flow over the mountain. The reader is referred to Pierrehumbert (1986) for further discussion on this topic.

CLOUD MICROPHYSICAL PROCESSES

As suggested by Houze (1993) "A common feature of these orographic precipitation mechanisms is the sensitive interdependence of orographically-induced flow dynamics (channeling, blocking and gravity-wave responses) and cloud microphysics. The microphysical effects are profound in regions of complex terrain. Heavy orographic precipitation often involves the growth and fallout of ice particles, which exhibit a wide range of fall velocities. Vapor-grown crystals and aggregates of ice crystals fall at speeds of 0.3–1.5 m/s. The more the particles grow by riming of supercooled water, the more dense they are and the faster they fall. Heavily rimed particles (graupel) fall at speeds of 1–3 m/s, while hail falls at 1–50 m/s. These fall speeds become crucial to determining precipitation patterns in complex terrain. Since cloud ice and snow particles are so readily carried by the horizontal wind, slight differences in the fallspeeds of ice particles determine which side of the mountain or in which valley (i.e. hydrological basin) precipitation ultimately reaches the ground."

Seeder-feeder Mechanism

The seeder-feeder mechanism was first proposed by Bergeron (1950, 1965) to explain orographic enhancements of up to a factor of 0.5 over small hills. Conventional microphysical analysis would suggest that the air traversing small terrain features of a few hundreds of meters or so is much too quick for the formation of raindrops from clouds produced by the hills themselves. The mechanism has two basic requirements: the presence of an upper-level seeder cloud and the existence of a lower-level feeder cloud. The feeder cloud is produced as a result of a low-level near-saturated air mass being forced to rise as it meets an orographic barrier. The water vapor in the air condenses, forming a cloud on the windward side of the hill. A stratiform seeder cloud provides precipitation particles to seed the lower-level feeder cloud and results in the removal of many cloud droplets of the feeder cloud that would otherwise be carried over the mountain. There must be a high liquid water content in the feeder cloud for the seeder-feeder mechanism to be effective, and there must also be a strong low-level flow in order to replenish the moisture lost due to the wash-out of the feeder cloud. This effect is found to be more significant when the "seeder" cloud is in the form

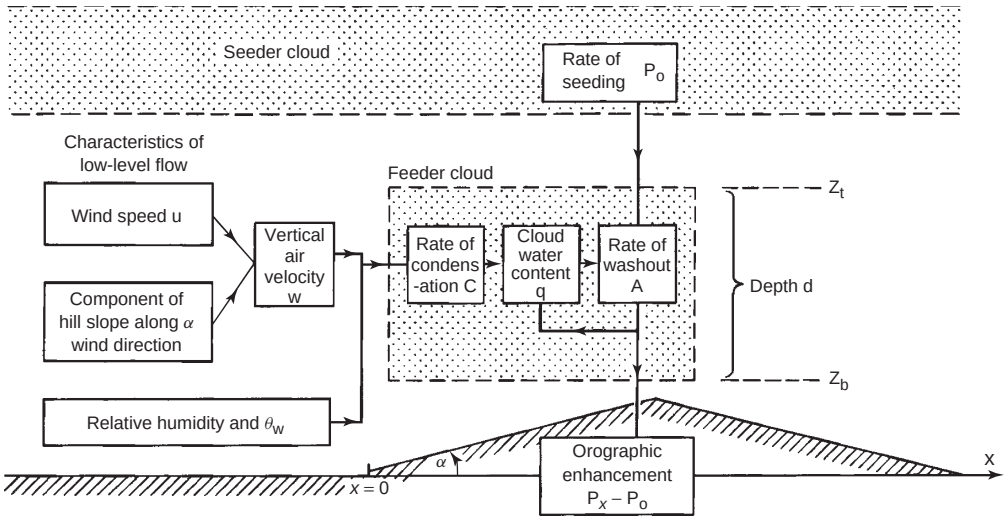


Figure 5 Schematic diagram illustrating the seeder-feeder mechanism (From Browning, 1980)

of snow (Choularton and Perry, 1986). Figure 5 illustrates the principle of the seeder-feeder mechanism as depicted by Browning (1980).

In situations of small hills, such as in South Wales, precipitation enhancements tend to coincide with areas where there is preexisting rain. The seeder-feeder effect plays a much lesser role in the scale of terrain features found on the west coast of the United States. In this situation, the majority of the orographic rainfall is the result of convection embedded in rainbands where a significant amount of orographic precipitation can fall without the presence of any preexisting rainfall.

During 1994–1996, observational field campaigns were undertaken over the Southern Alps of New Zealand as part of the Southern Alps Experiment (SALPEX) programme. The main aim of the programme was to develop a better understanding of the processes leading to enhancements of up to a factor of four observed on the windward side of the mountain range (Henderson, 1993), and the factors influencing the amount of spillover to the lee side of the mountains. (The mountain range regularly exceeds 2 km in altitude and is approximately 800 km in length.) Annual accumulations of rainfall on the windward western slopes often exceed 10 m, with rainfall accumulations on the lee side averaging at about 2 m.

Purdy *et al.*, (in press) present the results of a study showing radar evidence of orographic enhancement due to the seeder-feeder mechanism. SALPEX provided data from a vertically pointing radar (shown in Figure 6) showing clear evidence of the seeder-feeder mechanism at work. In this case, shallow rainfall at a height of approximately 2500 m has been enhanced by snowfall from above. However, it was found that the majority of the precipitation was due to the existence of drizzle drops of about 40 μm in diameter in the saturated airmass which approached the mountains. Thus the rain could form very quickly because the initial part of the 20 min buildup time had already occurred.

Rainfall Redistribution

An mentioned previously, in some regions with low hills it is observed that there is a deficit in rainfall upwind of the peak and a relative excess downwind, inconsistent with the predictions of most common rainfall production processes. In these cases, it may be that there has not been any enhancement in rainfall as such, and that the observed spatial variability is simply a redistribution of rainfall from a perturbation in the raindrop trajectories, due to the distortion of wind flow, as they pass over the hill. Such rainfall patterns were observed in a study on Norfolk Island in the South Pacific where a high-density rain-gauge network was installed in order to investigate high-resolution rainfall patterns (Stow and Dirks, 1998). The network consisted of 13 rain gauges distributed over the 8 km by 5 km island dominated by a 300 m high hill. In this study, Bradley *et al.*, (1998) suggested that the rainfall patterns were simply a result of a redistribution of raindrops due to perturbations of wind flows over the hill, as shown in Figure 7. In this figure, the bold line represents a transect of the topography of the island, while the other lines running left to right represent the wind streamlines. The deflection of the wind flow results in a deflection of raindrop trajectories (illustrated by the lines running from the top left to the bottom right of the diagram). Areas where there is a large spacing between the trajectories as the drops hit the ground represent deficits in rainfall, whereas small spacings between trajectories represent increases. These rainfall patterns were modeled using the analytic solutions for potential flow over a cylinder to find streamlines similar to the topographical transects through the hill and aligned with the wind direction. The results indicate that rainfall

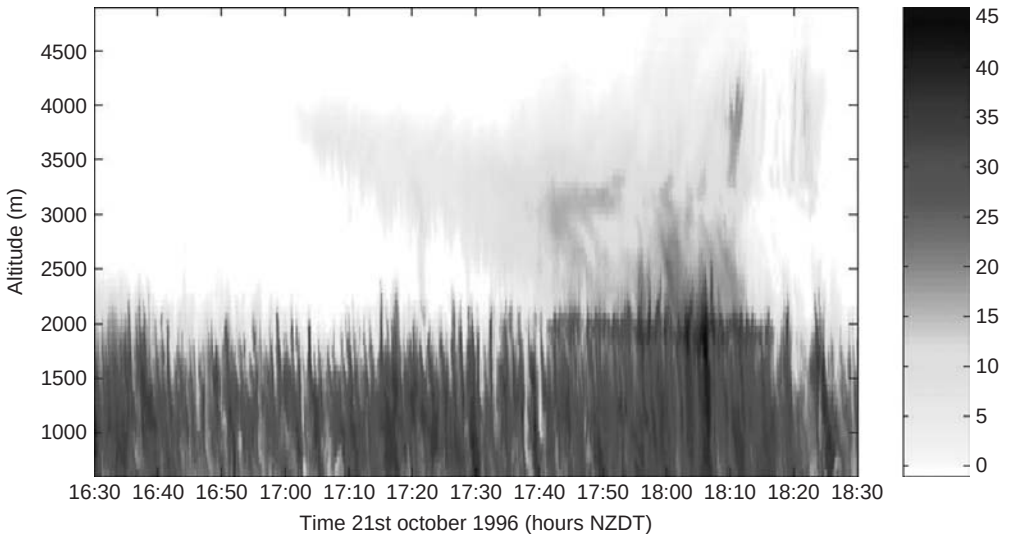


Figure 6 Evidence of the seeder-feeder mechanism at work from radar observations (with permission of Joanne Purdy). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

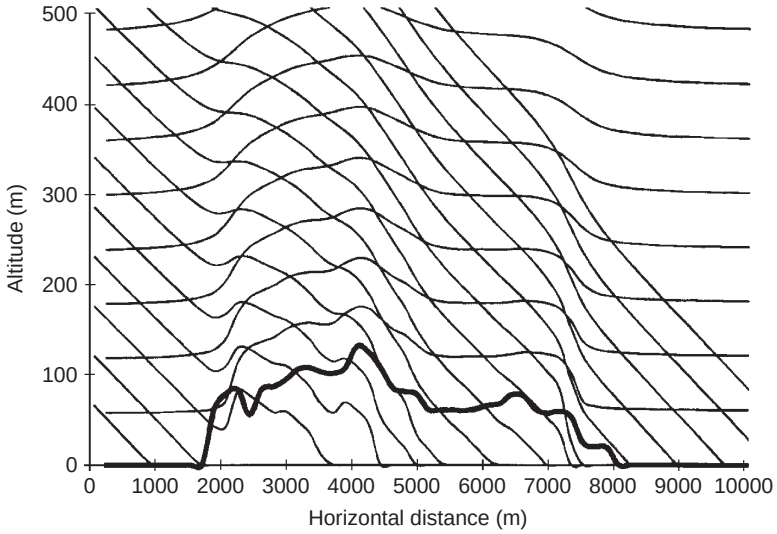


Figure 7 Predictions from the general three-dimensional potential flow model for a ratio of drop fall speed to wind speed equal to 0.075. The topography is shown as an overlay together with streamlines (running from left to right) and drop trajectories (running from upper left to lower right) (From Bradley *et al.,* 1998)

redistribution is a plausible mechanism for the observed rainfall patterns for small hills in the absence of low-level moisture.

CONCLUSIONS AND FURTHER DIRECTIONS

The effects of topography and orography on precipitation are largely dictated by the effects of the terrain on wind flows. In the simplest case, the modified low-level flow causes the precipitation to be redistributed. In more fundamental cases, the topography may result in mesoscale circulations due to the modified thermal properties or the slope of the boundary layer. These mesoscale air circulations can result in triggered convection, upslope rainfall, or completely new mesoscale systems, such as lee depressions. In the more stratiform cases, the long times required for precipitation development from the newly formed clouds may not always minimize the effects of topography. In South Wales, upper level snowfall often scavenges rain from the lower orographically induced cloud, and in the New Zealand Southern Alps, often the inflowing air mass has

large drizzle drops already formed before orographic lifting occurs.

The complex and important effects of topography on precipitation will continue to be studied because they are a dominant process in determining the geographical distribution of rainfall worldwide. Moreover, they are the origin of several hazardous conditions, including thunderstorms running up valleys or canyons in mountainous regions, and the landfall of both tropical and subtropical cyclones over coastal mountain ranges which can lift very moist precipitating air, leading to a potential for disastrous flooding.

REFERENCES

Atkinson B.W. (1989) *Meso-Scale Atmospheric Circulations*, Academic Press: London.

Bergeron T. (1950) Uber der mechanismus der ausgeibigan niederschlage. *Berichte Des Deutschen Wetterdienstes*, **12**, 225–232.

Bergeron T. (1965) On the low level redistribution of atmospheric water caused by orography. *Supp. Proc. Int. Conf. Cloud Physics*, IAMAP/WMO: Tokyo, pp. 96–100, May 1965.

Bergeron T. (1968) On the low-level distribution of atmospheric water caused by orography. *Presented at International Cloud Physics Conference*, Toronto.

Bowen E.G. (1950) The formation of rain by coalescence. *Australian Journal of Scientific Research Series A-Physical Sciences*, **3**, 193–196.

Bradley S.G., Dirks K.N. and Stow C.D. (1998) High-resolution studies of rainfall on Norfolk Island, Part III: A model for rainfall redistribution. *Journal of Hydrology*, **208**, 194–203.

Browning K.A. (1980) Structure, mechanism and prediction of orographically enhanced rain in Britain. In *Orographic effects in Planetary Flows*, Hide R. and West P.W. (Eds.), GARP Publication Series No. 23, WMO: pp. 85–114.

Byers H.R. and Rodebush H.R. (1948) Causes of thunderstorms of the Florida peninsula. *Journal of Meteorology*, **5**, 275–280.

Chou C. (2003) Land-sea heating contrast in an idealized Asian summer monsoon. *Climate Dynamics*, **21**, 11–25.

Choularton T.W. and Perry S.J. (1986) A model for orographic enhancement of snowfall by the seeder-feeder mechanism. *Quarterly Journal of the Royal Meteorological Society* **112**, 335–345.

Findeisen W. (1939) Zur Frage der Regentropfenbildung in reinen Wasserwolken. *Meteorologische Zeitschrift*, **56**, 365–370.

Frei C. and Schär C. (1998) A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology*, **18**(8), 873–900.

Hahn D.G. and Manabe S. (1975) The role of mountains in the south Asian monsoon circulation. *Journal of the Atmospheric Sciences*, **32**, 1515–1541.

He H., McGinnis J.W., Song Z. and Yanai M. (1987) Onset of the Asian monsoon in 1979 and the effect of the Tibetan Plateau. *Monthly Weather Review*, **115**, 1966–1995.

Henderson R.D. (1993) Extreme storm rainfalls in the Southern Alps, New Zealand, extreme hydrological events: precipitation, floods and droughts. *Proceedings of the Yokohama Symposium*, JAHS Publication No. 213, JAHS, July 1993.

Houze R.A. Jr (1993) Orographic Clouds. *Cloud Dynamics*, Academic Press: pp. 502–538.

Huang-Hsiung H. and Liu X. (2003) Relationship between the Tibetan Plateau heating and East Asian summer monsoon rainfall. *Geophysical Research Letters*, **30**(20), 2066–2075.

Klimowski B.A., Becker R., Betterton E.A., Bruintjes R., Clark T.L., Hall W.D., Orr B.W., Kropfli R.A., Piironen P., Reinking R., *et al.* (1998) The 1995 Arizona Program: toward a better understanding of winter storm precipitation development in mountainous terrain. *Bulletin of the American Meteorological Society*, **79**, 799–813.

Langmuir I. (1948) The production of rain by a chain reaction in cumulus clouds at temperatures above freezing. *Journal of Meteorology*, **5**, 175–183.

Nicholls M.E., Pielke R.A. and Cotton W.R. (1991) A two-dimensional numerical investigation of the interaction between sea-breezes and deep convection over the Florida peninsula. *Monthly Weather Review*, **119**, 298–323.

Petterssen S. (1956) *Weather Analysis and Forecasting*, McGraw-Hill: New York.

Petterssen S. (1969) *Introduction to Meteorology*, *Third Edition*, McGraw-Hill: New York.

Pielke R.A. (1974) A three-dimensional numerical model of the sea breezes over south Florida. *Monthly Weather Review*, **102**, 115–139.

Pierrehumbert R.T. (1986) Lee Cyclogenesis. In *Mesoscale Meteorology and Forecasting*, Ray P.S. (Ed.) Publication of the American Meteorological Society: Boston, pp. 493–511.

Purdy J., Austin G.L., Seed A.W. and Cluckie I.D. (in press) Radar evidence of orographic enhancement due to the seeder feeder mechanism, Dominant processes in New Zealand's Southern Alps, Physis Department University of Auckland, pp. 145.

Ray P.S. (1986) *Mesoscale Meteorology and Forecasting*, Ray P.S. (Ed.) Publication of the American Meteorological Society: Boston.

Stow C.D. and Dirks K.N. (1998) High-resolution studies of rain fall of Norfolk Island, Part I: the spatial variability of rainfall. *Journal of Hydrology*, **208**, 163–186.

University Corporation for Atmospheric Research (2002).

Wratt D.S., Ridley R.N., Sinclair M.R., Larsen H.R., Thompson S.M., Henderson R., Austin G.L., Bradley S.G., Auer A., Sturman A.P., *et al.* (1996) The New Zealand Southern Alps experiment. *Bulletin of the American Meteorological Society*, **77**, 683–692.

31: Models of Clouds, Precipitation and Storms

ANDREA I FLOSSMANN

Laboratoire de Météorologie Physique/OPGC, Université Blaise Pascal/CNRS, Aubière Cedex, France

Clouds play an important role for life on earth. Apart from influencing, for example, the radiative balance of the atmosphere and the lifetime of atmospheric trace constituents, they are the essential element in the hydrological cycle. Clouds transport the evaporated water of the oceans to the continents where the precipitation releases the water load. This release of water can be more or less vigorous depending on the energy stored in the cloud in the form of condensed hydrometeors (liquid or solid). This energy depends on the amount of available moisture and the way the vertical lifting necessary for cloud formation proceeds. We distinguish here mainly two different forms with varying extensions on the horizontal scale: the gentle uplift associated to the large-scale lifting, for example at a frontal zone, and the vigorous small-scale ascent associated with convection, knowing that also mixed forms of the lifting occur.

This article provides an introduction to the complex subject of modeling clouds, the production of precipitation, and the development of cloud and storm systems. The elements intervening in cloud modeling are exposed, starting from a description of the physical phenomena. On the basis of the occurring scale problem, a number of approaches for simplification are presented. These simplifications concern the dynamics as well as the microphysics. Bulk and bin modeling approaches are explained, as well as cumulus parameterizations. Some numerical problems are discussed. This approach gives an insight into current state-of-the-art cloud modeling and the necessary balance between the degree of parameterization, the number of physical and chemical processes relevant to a particular problem, and the available computing resources.

INTRODUCTION

Clouds play an important role for life on earth. Apart from influencing, for example, the radiative balance of the atmosphere and the lifetime of atmospheric trace constituents, they are *the* essential element in the hydrological cycle (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**). Clouds transport the evaporated water of the oceans to the continents where the precipitation releases the water load. This release of water can be more or less vigorous depending on the energy stored in the cloud in the form of condensed hydrometeors (liquid or solid). This energy depends on the amount of available moisture and the way the vertical lifting necessary for cloud formation proceeds. We distinguish here mainly two different forms with varying extensions on the horizontal scale: the gentle uplift associated to the large-scale lifting, for example, at a frontal zone, and the vigorous small-scale ascent associated with convection, knowing that also mixed forms of the lifting occur.

Consequently, when deciding to model clouds, precipitation, and storms, we immediately encounter a problem of scale: in order to model correctly the large-scale dynamics, we need to consider a horizontal domain of several thousand kilometers. And in order to consider correctly the form and the size of the condensed hydrometeors, we need also to take into account processes that take place on the micrometer scale.

Below, we quickly review the type and scale of microphysical processes that need to be considered (Flossmann and Laj, 1998) before presenting the different types of approaches used in cloud modeling in order to account for the scale problem.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

MICROPHYSICAL PROCESSES

For a complete description see, for example, Pruppacher and Klett (1997), Rogers and Yau (1989), Cotton and Anthes (1989), and Houze (1993) (*see* **Chapter 28, Clouds and Precipitation, Volume 1**).

Nucleation of Drops

Our atmosphere does not allow the formation of drops by agglomerating water vapor molecules (homogeneous nucleation) alone. In order to form a tiny drop of $2 \times 10^{-3}\,\mu$m radius, assembled of 800 molecules, it would require a relative humidity of 200%. Consequently, in the atmosphere droplets form on already existing nuclei (heterogeneous nucleation). A subset of the aerosol particles present in every air mass provides these necessary nuclei.

Aerosol particles typically encountered in the atmosphere have a size range between 10^{-3} and 10 μm. Their total number concentration varies between 100 and 1 00 000 cm^{-3} and their mass can reach up to several hundreds of $\mu\mathrm{g\,m}^{-3}$. It has been proposed to distinguish three different size ranges for the particles corresponding to different formation and destruction mechanisms:

Table 1 points out that Aitken and large aerosol particles are formed by condensation processes from the gas phase and aggregation. The gases can be of natural or anthropogenic origin. However, giant aerosol particles have a mechanical origin, which is mostly natural.

Depending on their size, chemical composition and the ambient relative humidity aerosol particles take up a certain amount of water (Köhler curve, see Pruppacher and Klett, 1997). So even in the absence of a cloud, the atmosphere is full of tiny droplets representing swollen aerosol particles, which can become visible, for example, in an atmospheric haze.

At a given size, particles carrying more hygroscopic material can take up more water before achieving a critical size at a critical relative humidity (supersaturation below 5%). When a given aerosol particle has passed its critical size, we call it activated. It serves as a cloud condensation nucleus (CCN). Then, it does not require a further increase in relative humidity to make it grow and it is considered a cloud droplet as long as supersaturation prevails. At a given ambient supersaturation, all aerosol particles whose critical supersaturation is below the ambient one will be activated and transformed into drops.

Condensation and Evaporation

Once moist aerosol particles have been activated to form drops, they grow according to the droplet growth equation

$$\frac{\mathrm{d}m_{\mathrm{con/eva}}}{\mathrm{d}t} = \frac{\left(\mathrm{s_{vw}} - \dfrac{2M_{\mathrm{w}}\sigma_{\mathrm{s,a}}}{RT\rho_{\mathrm{w}}r} - \dfrac{\upsilon\phi_{\mathrm{s}}\varepsilon_{\mathrm{v}}M_{\mathrm{w}}\rho_{\mathrm{s}}r_{\mathrm{N}}^3}{M_{\mathrm{s}}\rho_{\mathrm{w}}(r^3 - r_{\mathrm{N}}^3)}\right) f_{\mathrm{v}}}{\dfrac{\rho_{\mathrm{w}}RT}{e_{\mathrm{sat,w}}D_{\mathrm{v}}'M_{\mathrm{w}}} + \dfrac{l_{\mathrm{vw}}\rho_{\mathrm{w}}}{K'T}\left(\dfrac{l_{\mathrm{vw}}M_{\mathrm{w}}}{RT} - 1\right)} \tag{1}$$

Equation (1) describes the change of the drop mass m (radius r) as a function of the supersaturation $s_{\mathrm{vw}} = e_{\mathrm{a}}/e_{\mathrm{sat,w}}$ (ratio of the vapor pressure and the saturation vapor pressure). Herein, T: temperature, l_{vw}: latent heat of evaporation, D_{v}': modified diffusion coefficient, K': modified thermal conductivity, $\sigma_{\mathrm{s,a}}$: surface tension of the solution droplet, ρ_{w}, ρ_{s}: density of water and salt respectively, M_{w}, M_{s} molecular weight of water and salt, υ: number of ions, ϕ_{s}: osmotic coefficient, R: universal gas constant, ε_{v}: mass fraction of soluble material, r_{N}: radius of the insoluble aerosol nucleus, f_{v}: is the ventilation coefficient which is due to the falling motion of drops in the atmosphere (for more information see Pruppacher and Klett, 1997).

The growth rates given by this formula are rapid for small droplets. However, they decrease with increasing radius and become quite small for drops larger than 10 μm radius (see Pruppacher and Klett, 1997). Consequently, the larger the drop, the longer it will take for it to grow further, leading to a halt in growth by condensation at around 30 μm in radius. Consequently, in order to form precipitation-sized drops other growth mechanisms interfere as detailed below.

Fall Speed of Drops

The terminal velocity of a drop is determined by a balance between buoyancy-corrected gravitation and the drag force

Table 1 Classification of aerosols after Junge and Whitby and the different formation and removal processes most important for the various size classes of aerosol particles

Size classes (μ m)	Junge (1963)	Whitby (1978)	Formation	Removal
$r < 0.1$	Aitken particles	Nuclei or transient mode	Condensation from the gas phase	Aggregation
$0.1 < r < 1$	Large particles	Accumulation mode	Condensation, aggregation	Washout and rainout
$r > 1$	Giant particles	Coarse mode	Sea spray, wind- blown dust, volcanoes, plant debris, diesel engines	Sedimentation

Table 2 Approximate fall velocities of water drops for the conditions at the surface (1013 hPa, 20 °C)

Radius (r) in μm	50	100	200	500	1000	2000
Terminal velocity (V_∞) in ms^{-1}	~0.25	~0.8	~1.5	~4	6.5	8.5

acting on the drop. In Table 2, some values for the terminal velocity as a function of drop radius are given.

These values pertain to normal surface conditions. At higher altitudes the fall speed increases as a result of a decrease in air density. These fall speeds are determined in wind tunnel measurements because an analytical calculation using the balance between the acting forces is not possible. This is due to the fact that large drops are no longer spherical. They get deformed by the airstream and develop an indentation on the upwind side. Furthermore, they start to oscillate because of internal and external vortices that develop and because of electrical charges in the atmosphere and the drop. These features do not allow a correct analytical formulation of the drag.

Collision and Coalescence of Drops

Because of the difference in their size, drops have different terminal velocities. Thus, they can overtake each other while falling and collide to form a larger drop with the sum of the original masses. The probability for a small drop r_2 that is located inside the geometrical sweep-out volume of the larger drop r_1 to collide with this drop is given by the collision efficiency E_{coll} (Figure 1). This collision efficiency is strongly dependent on the size of the two drops. If the large drop is smaller than 20 μm, a collision is essentially impossible, as the terminal velocity is too small and the small drop is just carried around the large drop by the airstream. In a wide-size range, the collision efficiency takes values of unity. When the two drops are roughly of equal size, E_{coll} can even exceed unity because of the wake capture of the trailing drop (see Pruppacher and Klett, 1997, for details).

$$E_{\text{coll}} = \frac{\pi y_c^2}{\pi (r_1 + r_2)^2} \tag{2}$$

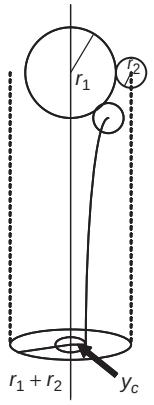


Figure 1 Schematical display of the collision efficiency

The process of collision is not always accompanied by coalescence, a fact that is described by the coalescence efficiency. Colliding drops may bounce apart, coalesce, coalesce temporarily, and then separate again later, or coalesce temporarily and then shatter.

The collection efficiency of the process of drop–drop collision can be written as the product of the collision and the coalescence efficiency:

$$E(m_1, m_2) = E_{\text{coll}} E_{\text{coal}} \tag{3}$$

and this efficiency will determine the rate of precipitation formation in clouds that do not develop an ice phase.

It is important to note that collection is enhanced by a broad droplet size spectrum that depends essentially on the dynamics of the clouds but also on the nature and sizes of the aerosol particles, as well as turbulence in clouds, and even radiative cooling effects. The broadness of the distribution is influenced not only by the number of CCN but also by the presence of giant or ultragiant particles, which can serve as coalescence embryos.

Breakup of Drops

Drops in the atmosphere do not grow indefinitely. The upper size limit is around 6 mm in diameter. Drops arriving at this size range develop a tendency to break up. We can distinguish two main mechanisms:

- *Spontaneous breakup:* the deformation of large drops and the induced oscillation becomes so important that the drop becomes hydrodynamically unstable and breaks up
- *Collisional break up:* during the process of collision/coalescence, the newly formed drop is too unstable and breaks up forming two large drops, slightly smaller than the original drops, and a number of small satellite droplets (see discussion on coalescence above)

Nucleation of Ice Particles

In the atmosphere, liquid drops can exist even at temperatures well below 0 °C, that is, in a supercooled state. In fact, significant numbers of ice particles start to form only below −5 °C coexisting mostly still with liquid drops. Homogeneous freezing of liquid droplets depends on the size; large droplets can freeze homogeneously at temperatures of around −33 °C, whereas by −40 °C even the smallest

droplets freeze homogeneously. Thus, at $-40\,^{\circ}C$ the last liquid has disappeared. In the temperature range between -5 and $-40\,^{\circ}C$, the presence of ice-forming nuclei is necessary to initiate the formation of an ice crystal. These ice nuclei (IN) are aerosol particles that can act in four main ways:

- *Deposition mode:* water is adsorbed directly from the vapor phase onto the surface of an IN where it is transformed into ice
- *Condensation–freezing mode:* this is a hybrid process that requires supersaturation with respect to water. Here, the CCN that has formed the drop acts now as an IN. This process seems far more effective than the deposition mode.
- *Freezing mode:* the IN, scavenged by the drop, initiates the ice phase from within a supercooled water droplet
- *Contact mode:* the IN initiates the ice phase at the moment of contact with the supercooled drop

The number of IN depends on the chemical properties of the aerosol particles. It has been found that there exists a dependency on supersaturation (Meyers *et al.*, 1992) and also on temperature (Fletcher, 1962). In contrast to CCN, a good IN should be insoluble and have already a crystalline-type structure to facilitate the formation of the ice lattice (e.g. silicate). Some bacteria have also been identified as excellent IN. For a comprehensive review of the biogenic versus anthropogenic sources of IN, see Szyrmer and Zawadzki (1997).

Deposition and Sublimation

Water droplets in the atmosphere have a spherical appearance facilitating their analytical treatment. Ice crystals present more problems since they appear in a number of different shapes depending on temperature and the grade of water supply. Here we find densely packed structures like needles, columns, and plates, as well as light dendritic structures (for details see e.g. Pruppacher and Klett, 1997). The actual growth rate of these crystals by water vapor deposition depends on the crystal form as well as temperature and humidity.

This deposition growth can take place at the same time as the condensation growth of droplets, if the ambient relative humidity is high enough to maintain supersaturation over liquid-water and over ice. However, the equilibrium curves of H_2O also allow occurrences of air supersaturated with respect to ice and subsaturated with respect to liquid water. Then, the drops evaporate and the vapor condenses onto the ice crystals (Bergeron–Findeisen effect; Pruppacher and Klett, 1997).

Fall Speed of Ice Particles

Ice particles like droplets have a terminal velocity that depends heavily on the size of the particle. Additionally, they depend on the shape of the particle and its density, which can take values between $0.98\,g\,cm^{-3}$ and $0.2\,g\,cm^{-3}$. A light dendritic structured ice particle will, thus, have a lower terminal velocity than a dense plate-like crystal even if both have the same diameter. For aggregated and rimed structures, the same applies.

Aggregation and Riming

For droplets we noted that condensation alone cannot develop precipitation, and we need the process of collision/coalescence. The same applies for ice crystals. Their growth by deposition of water vapor alone is also limited. Thus, a process of collision is necessary to produce larger aggregates. Two different mechanisms are possible:

- *The collision of ice crystals among themselves:* this process is called *aggregation*. It is responsible for the formation of snow crystals. This process is efficient in two different temperature regimes, around -10 to $-15\,^{\circ}C$ and around $0\,^{\circ}C$. Around -10 to $-15\,^{\circ}C$, the crystals develop a dendritic structure and, thus, upon the collision of two crystals they easily succeed in interlinking with their branches. Around $0\,^{\circ}C$, the crystals develop a "pseudo liquid layer" (a micro layer of melting ice) at their surface which enables them to create a link with the other crystal via freezing. Outside these specific temperature regions, the collision of two ice crystals is rarely successful, thus, not resulting in a snow flake.
- *The collision of ice crystals with liquid drops:* this process is called *riming*. Here, the drop is frozen upon collision with the crystal and several of these collisions transform the ice crystal into a graupel and then into a hail particle. This is a very efficient way of precipitation production. Sometimes the latent heat involved in this riming process is so important that not all the total captured liquid can be frozen. Normally, the remaining liquid will be incorporated into the solid structure forming a "spongy" ice. In a current mesoscale model(MSL), this is not treated. Thus, the excess liquid is shed and reattributed to the liquid water, a process called *wet growth*.

Secondary Ice Particle Formation

During the above-discussed processes of aggregation and riming, an ice multiplication mechanism can occur. If the crystals involved in the collision have a dendritic structure, parts can easily break off and form new ice crystals. Hallett and Mossop (1974) also proposed a process leading to

secondary ice formation during the process of riming. If the drop involved freezes from the outside-in, then first a solid shell will form with a liquid core. When the liquid core starts to freeze, the volume inside the ice shell is too small to receive the forming ice. This fact will explode the shell and eject liquid material in the air, which will immediately freeze and form new ice crystals. This process is called *splintering*.

Melting

Even though a precipitating particle might have formed through riming of ice crystals, it can arrive at the ground as a liquid raindrop. This depends on the altitude of the 0 °C level in the atmosphere. Below this level, ice aggregates will start to melt and, depending on the fall distance, will arrive at the ground as partly or completely melted hydrometeors. A 3-mm ice particle can fall at a distance between 1 and 3 km before complete melting, depending on its initial density and its warm environment (see Pruppacher and Klett, 1997, for details).

THE SCALE PROBLEM

Models are an assembly of equations that describe the phenomenon to be studied. For a cloud model, this includes equations describing the change in time of temperature, humidity, pressure, density, and three-dimensional (3-D) wind field, and some variables that represent the cloud, such as drop or crystal numbers or liquid water content. These equations emerge from the concepts of conservation of mass and energy, Newton's law, and the perfect gas law and can be written in a general form:

$$
\begin{aligned}
&\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho \vec{v}) = 0 \\
&\frac{\partial \rho T}{\partial t} + \mathrm{div}(\rho \vec{v} T) = -\frac{1}{c_\mathrm{p}} \mathrm{div} \vec{F}_T + \frac{1}{c_\mathrm{p}} \frac{\mathrm{d}p}{\mathrm{d}t} + \frac{L}{c_\mathrm{p}} C_\mathrm{con} \\
&\qquad + \frac{L_\mathrm{sub}}{c_\mathrm{p}} C_\mathrm{sub} + \frac{L_\mathrm{melt}}{c_\mathrm{p}} C_\mathrm{melt} \\
&\frac{\partial}{\partial t} \rho \vec{v} + \mathrm{div}(\rho \vec{v} \vec{v}) = \mathrm{grad}\ p - \rho \vec{g} + \mathrm{div}\tau \\
&\frac{\partial}{\partial t} \rho q_\mathrm{v} + \mathrm{div}(\rho \vec{v} q_\mathrm{v}) = -\mathrm{div} \vec{F}_\mathrm{v} - C_\mathrm{con} - C_\mathrm{sub} - C_\mathrm{melt} \\
&\frac{\partial N_\mathrm{drop}}{\partial t} + \mathrm{div}(\vec{v} N_\mathrm{drop}) + \frac{\partial}{\partial z}(V_{\infty,\mathrm{drop}} N_\mathrm{drop}) \\
&\qquad = -\mathrm{div} \vec{F}_\mathrm{drop} + S_\mathrm{drop} \\
&\frac{\partial N_\mathrm{crystal}}{\partial t} + \mathrm{div}(\vec{v} N_\mathrm{crystal}) + \frac{\partial}{\partial z}(V_{\infty,\mathrm{crystal}} N_\mathrm{crystal}) \\
&\qquad = -\mathrm{div} \vec{F}_\mathrm{crystal} + S_\mathrm{crystal} \\
&\qquad p/\rho = nRT
\end{aligned}
\tag{4}
$$

with ρ: density of air, $\vec{v}$: 3-D wind field, T: temperature, q_v: water vapor mixing ratio, c_p: specific heat at constant pressure, $\vec{F}_T$: sensible heat-flux, p: pressure, L, L_sub, L_melt: latent heat of condensation, sublimation, and melting, $C_\mathrm{con}, C_\mathrm{sub}, C_\mathrm{melt}$: rate of phase change of condensation, sublimation, and melting, $\vec{g}$: acceleration of gravity, τ: friction tensor, N_drop, N_crystal: number of drops and crystals of a given size per unit volume, $\vec{F}_\mathrm{drop}$, $\vec{F}_\mathrm{crystal}$: diffusion fluxes of drops and crystals, S_drop, S_crystal: source, sinks, and transfer terms of drops and crystals, n: number of moles, R: universal gas constant. (Instead of temperature, models often calculate potential temperature $\theta = T(p_0/p)^{(R/cp)}$.)

These equations are, then, solved numerically at a number of points in space and time, while the points represent a certain interval in time or space. As the equations result from continuity principles, these intervals should be infinitely small. In practice, the grid boxes have a finite size. Thus, to be correct, we need to average the equations over finite increments in space and time:

$$
\begin{aligned}
&\overline{\psi} = \frac{1}{\Delta t \Delta x \Delta y \Delta z} \int_t^{t+\Delta t} \int_x^{x+\Delta x} \int_y^{y+\Delta y} \int_z^{z+\Delta z} \psi \, \mathrm{d}z \, \mathrm{d}y \, \mathrm{d}x \, \mathrm{d}t \\
&\psi = \overline{\psi} + \psi'
\end{aligned}
\tag{5}
$$

This modifies the equations in such a way that we now calculate average values over a grid box that modifies the equations so that a new term appears in every prognostic equation taking into account the subgrid changes of space and time inside the box:

$$
\begin{aligned}
&\frac{\partial \overline{\rho}}{\partial t} + \mathrm{div}(\overline{\rho} \overline{\vec{v}}) = 0 \\
&\frac{\partial \overline{\rho} \overline{T}}{\partial t} + \mathrm{div}(\overline{\rho} \overline{\vec{v}} \overline{T}) - \frac{1}{c_\mathrm{p}} \mathrm{div} \vec{F}_T^\mathrm{turb} = -\frac{1}{c_\mathrm{p}} \mathrm{div} \overline{\vec{F}}_T + \frac{1}{c_\mathrm{p}} \frac{\mathrm{d}\overline{p}}{\mathrm{d}t} \\
&\qquad + \frac{L}{c_\mathrm{p}} \overline{C}_\mathrm{con} + \frac{L_\mathrm{sub}}{c_\mathrm{p}} \overline{C}_\mathrm{sub} + \frac{L_\mathrm{melt}}{c_\mathrm{p}} \overline{C}_\mathrm{melt} \\
&\frac{\partial}{\partial t} \overline{\rho} \overline{\vec{v}} + \mathrm{div}(\overline{\rho} \overline{\vec{v}} \overline{\vec{v}}) - \mathrm{div} \tau^\mathrm{turb} = \mathrm{grad} \overline{p} - \overline{\rho} \vec{g} + \mathrm{div} \overline{\tau} \\
&\frac{\partial}{\partial t} \overline{\rho} \overline{q}_\mathrm{v} + \mathrm{div}(\overline{\rho} \overline{\vec{v}} \overline{q}_\mathrm{v}) - \mathrm{div} \vec{F}_\mathrm{v}^\mathrm{turb} = -\mathrm{div} \overline{\vec{F}}_\mathrm{v} \\
&\qquad - \overline{C}_\mathrm{con} - \overline{C}_\mathrm{sub} - \overline{C}_\mathrm{melt} \\
&\frac{\partial \overline{N}_\mathrm{drop}}{\partial t} + \mathrm{div}(\overline{\vec{v}}\, \overline{N}_\mathrm{drop}) + \frac{\partial}{\partial z}(\overline{V}_{\infty,\mathrm{drop}} \overline{N}_\mathrm{drop}) \\
&\qquad - \mathrm{div} \vec{F}_{Nd}^\mathrm{turb} = -\mathrm{div} \overline{\vec{F}}_\mathrm{drop} + \overline{S}_\mathrm{drop} \\
&\frac{\partial \overline{N}_\mathrm{crystal}}{\partial t} + \mathrm{div}(\overline{\vec{v}}\, \overline{N}_\mathrm{crystal}) + \frac{\partial}{\partial z}(\overline{V}_{\infty,\mathrm{crystal}} \overline{N}_\mathrm{crystal}) \\
&\qquad - \mathrm{div} \vec{F}_{Nc}^\mathrm{turb} = -\mathrm{div} \overline{\vec{F}}_\mathrm{crystal} + \overline{S}_\mathrm{crystal} \\
&\qquad \overline{p}/\overline{\rho} = nR\overline{T}
\end{aligned}
\tag{6}
$$

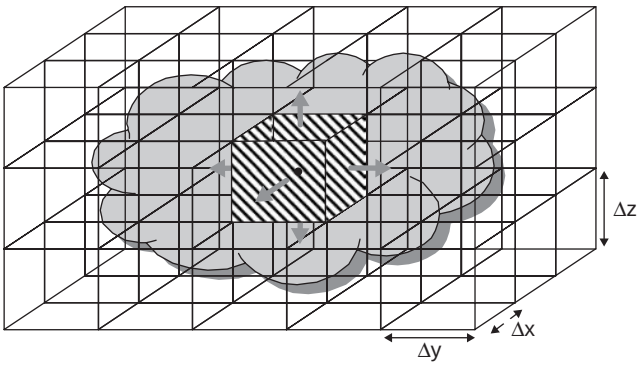


Figure 2 Scheme of grid boxes that resolve a cloud

The new terms have the form:

$$\vec{F}_{\psi}^{\,\mathrm{turb}} = -\overline{\rho\psi'\vec{v}'}, \quad \tau^{\mathrm{turb}} = -\overline{\rho\vec{v}'\vec{v}'} \tag{7}$$

and depend on the subgrid fluctuations ψ' in space and time of the considered variables. These terms need to be parameterized as a function of the variables $\overline{\psi}$ of the resolved scale. This approach is known under the term *closure*, whereby different theories of closure (first, second order, etc.) can be found in the literature Stull (1991). In the following, for simplicity of writing, the overbar will be dropped; however, all equations are averaged equations.

Considering, that an appropriate closure approach has been selected, the foregoing equations can be solved numerically in a type of grid such as that displayed in Figure 2, covering the entire cloud area and the affected environment. The averaging done in equations (6) and (7) now allows finite grid increments. However, the microphysical processes discussed above make it evident that essential processes in clouds operate at the scale of micrometers, requiring very small intervals in space. But, on the other hand, clouds sometimes form systems of several thousand kilometers, requiring a large spatial domain to be covered. One is, thus, confronted with the problem that actual computer power does not allow us to cover domains of $1000\,\mathrm{km}^2$ with grid increments of $1\,\mathrm{m}^2$ (this would require 10^{12} grid boxes).

A similar argumentation applies to the time increment. Time and space increments are linked in an explicit numerical treatment of the system of equations by the Courant–Friedrich–-Levy (CFL) criterium (Jacobson, 1999):

$$\left(\frac{\Delta t}{\Delta x}\right) U \leq 1 \tag{8}$$

with U being the maximum transport velocity considered. This criterium insures that the fasted signal does not jump a grid box in one time step. The fasted transport velocities in the atmosphere are those of the sound. They are not physically relevant in cloud models and are, thus, filtered by suppressing local density fluctuations in the equations (an elastic or soundproof assumption) (Ogura and Phillips, 1962).

However, the numerical effort for a high resolution of the processes stays considerable. Consequently, strategies to reduce the simulation time need to be developed. These consist in compromises concerning the dynamics, the microphysics, and the resolution.

SIMPLIFICATION IN THE DYNAMICS

3-D Models

The appropriate way to simulate the dynamics of the atmosphere is a 3-D representation of space. Here, a classical coordinate system would be a Cartesian one with x, y, and z in the directions of space (*see* Figure 2). In order to take into account the topography, often the vertical coordinate z is replaced by one that follows the terrain. Phillips (1957) designed such a terrain-following coordinate system for use in hydrostatic, numerical prediction models by defining:

$$\sigma = \frac{p}{p_{\mathrm{s}}} \tag{9}$$

where p_s represents surface pressure and σ varies from a value $\sigma = 1$ at the ground to $\sigma = 0$ at the top of the atmosphere. Nowadays, however, most of the 3-D models (Clark, 1977; Dudhia, 1993; Pielke *et al.*, 1992) devoted to cloud simulations use the vertical coordinate:

$$z^* = H\frac{z - z_{\mathrm{s}}}{H - z_{\mathrm{s}}} \tag{10}$$

The surface height above some reference level is given by z_s (e.g. sea level) and the height of the model top is given by H.

Three-dimensional models that cover an entire continent equally need to take into account in the horizontal grid the deformation of the coordinate system by the curvature of Earth and the large-scale Coriolis force. These models use stereographic coordinate systems. Some of the most widely used 3-D models are currently RAMS (Pielke *et al.*, 1992; Cotton *et al.*, 2003), MM5 (Dudhia, 1993), and Clark (1977), among others (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**).

2-D Models

For special case studies concerning phenomena that are mainly two-dimensional (2-D; flow over a mountain range, e.g.) or in order to reduce computer calculation times, 3-D dynamics can be reduced to two dimensions. This eliminates one coordinate (e.g. y in a Cartesian coordinate system) (e.g. Orville and Kopp, 1977; Soong and Ogura, 1980)

Another configuration, especially adapted in simulating isolated cumulus clouds consists of a cylindrical model with a symmetry with respect to the angular coordinate. Such a model configuration has been used, for example in the models of Shiino (1983), Murray and Koenig (1975), and Reisin *et al.* (1996).

1-D and 0-D Models

The cylindrical model philosophy has also been applied in the 1.5-D model of Asai and Kasahara (1967) where in fact two one-dimensional models are connected to represent the updraft and the downdraft region of a convective cloud.

Nowadays, one-dimensional models are mainly used to test microphysical schemes. In this case, there exists just a vertical axis z, subdivided into several layers, representing the updraft region of the cloud.

Finally, the least sophisticated dynamics is that of an air parcel. In this concept, we consider a volume of air separated from the environment by an immaterial surface. For the entraining air parcels, they exchange mass and heat with the environment; otherwise, they are considered as adiabatic. Air parcels can represent an interesting concept if they are driven by a 3-D flow field, because they permit highly resolved microphysics to be followed in a larger-scale dynamics that does not consider microphysics (Feingold *et al.*, 1998a). Here, care needs to be taken to ensure that the parcel bulk properties are adequately constrained to those from the driving 3-D model.

MODELS OF CLOUD MICROPHYSICS

Drops and ice crystals exist in a cloud in varying number concentrations and with different sizes. We distinguish cloud drops of radii between 1 μm and 30 μm, drizzle drops of up to 200 μm, and rain drops up to 6 mm. Ice crystals have diameters of up to 100 μm, snow flakes, graupel, and hail particles have no clear upper size but can become larger than raindrops. The number of hydrometeors per unit volume of a given size is fundamental for the evolution of the cloud and, thus, should be considered in a cloud model.

Explicit or Bin-resolving Cloud Models

The most logical approach is to introduce an explicit number density distribution function $f(m)\ \mathrm{d}m$, which follows the number per unit volume of a given type of hydrometeors in a given mass class between m and $m + \mathrm{d}m$ (see Figure 3). Inside a bin class most models assume a uniform value. There exist, however, also models that assume inside a drop bin class a distribution of aerosol particles or chemical compounds (e.g. Wobrock *et al.*, 2002).

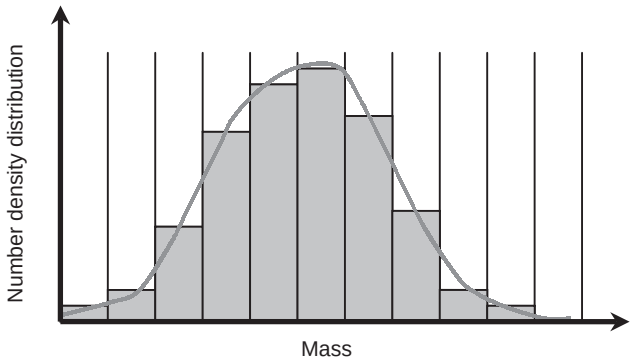


Figure 3 Number density distribution function $f(m)\,\mathrm{d}m$ which follows the number per unit volume of air of a given type of hydrometeors in a given mass class between m and $m + \mathrm{d}m$

Warm Clouds

Most ground fogs or low-level clouds contain only liquid drops, and no ice crystals. If they develop precipitation, this will take place by the collision and coalescence of liquid drops. These clouds are called *warm* clouds. In order to model the size distribution of droplets in warm clouds, a drop number density distribution function $f_{\mathrm{d}}(m)$ is defined. It allows the continuous distribution of drops over the entire spectrum to be captured in a discrete function.

Since the liquid hydrometeors need to cover a range from 1 μm drops up to 6 mm raindrops, which covers at least three orders of magnitude, an equidistant grid in size or mass is not appropriate. Consequently, Berry and Reinhardt (1974a–c) have introduced a logarithmic size grid which allows a doubling of mass every *JRS* category:

$$r(j) = r(1)2^{(J-1)/3JRS} \tag{11}$$

$r(1)$ gives the minimum radius (e.g. 1 μm) and JRS can be 1, 2, 3, or more. In most models $JRS = 2$ is chosen, as a compromise between the necessary resolution of drop spectrum (Silverman and Glass, 1973) and the computer times of the simulation. For each of the resulting drop-size classes (on the order to 70), the following prognostic equation has to be solved:

$$\begin{aligned}\frac{\partial f_{\mathrm{d}}(m,\vec{r},t)}{\partial t} &= -\mathrm{div}(\vec{v}\, f_{\mathrm{d}}(m,\vec{r},t)) \\ &+ \frac{\partial}{\partial z}(V_\infty f_{\mathrm{d}}(m,\vec{r},t)) + \mathrm{div}(\vec{F}_{fd}^{\,\mathrm{turb}}) \\ &+ \left[\frac{\partial}{\partial t} f_{\mathrm{d}}(m,\vec{r},t)\right]_{\mathrm{nuc}} + \left[\frac{\partial}{\partial t} f_{\mathrm{d}}(m,\vec{r},t)\right]_{\mathrm{con/eva}} \\ &+ \left[\frac{\partial}{\partial t} f_{\mathrm{d}}(m,\vec{r},t)\right]_{\mathrm{coll}} + \left[\frac{\partial}{\partial t} f_{\mathrm{d}}(m,\vec{r},t)\right]_{\mathrm{breakup}}\end{aligned} \tag{12}$$

The term on the left-hand side calculates the change of the number of drops at a certain location $\vec{r}$ per unit volume between a mass m and $m + \mathrm{d}m$ per time interval $\mathrm{d}t$ because of the physical effects represented by the terms on the right-hand side which are: transport with the scale velocity $\vec{v}$, the sedimentation with the terminal velocity V_∞, transport through turbulence, the change in number concentration due to nucleation of drops, due to condensation or evaporation, due to collision and coalescence among the drops, and due to breakup of drops. The mathematical expressions for these terms follow the physics detailed above.

A simple way to calculate the nucleation of drops is to prescribe the number of cloud condensation nuclei as a function of supersaturation:

$$N_{\mathrm{CCN}} = c s_{\mathrm{vw}}^{k} \tag{13}$$

with c, k constants which are adapted to the air mass considered, as, for example, in Twomey (1959). In addition, an assumption is necessary on the size of the freshly nucleated drops. More sophisticated models (Flossmann *et al.*, 1985) follow the number of aerosol particles and their chemical composition in order to calculate as a function of supersaturation (Köhler equation, Pruppacher and Klett, 1997) the number and the size of activated aerosol particles.

Once formed, the drops grow or shrink by condensation or evaporation respectively:

$$\left[\frac{\partial}{\partial t} f_{\mathrm{d}}(m)\right]_{\mathrm{con/eva}} = -\frac{\partial}{\partial m}\left(\frac{\mathrm{d}m_{\mathrm{con/eva}}}{\mathrm{d}t} f_{\mathrm{d}}(m)\right) \tag{14}$$

where the individual droplet growth rate is given from equation (1). The process of collision and coalescence of drops is calculated following Berry and Reinhardt (1974a–c):

$$\left[\frac{\partial}{\partial t} f_{\mathrm{d}}(m)\right]_{\mathrm{coll}} = \int_0^{m/2} f_{\mathrm{d}}(m - m')\, f_{\mathrm{d}}(m')\, K(m - m', m')\, \mathrm{d}m' - \int_0^{\infty} f_{\mathrm{d}}(m)\, f_{\mathrm{d}}(m')\, K(m, m')\, \mathrm{d}m' \tag{15}$$

Here, the first term on the right-hand side describes the gain of m-sized drops due to collision between drops of sizes $(m - m')$ and m'. The second term describes the loss of m-sized drops due to all types of collision. K represents the coalescence kernels that give the probability of such a collision to happen. E is the collection efficiency.

$$K(m, m') = \pi (r + r')\, |V_{\infty,\mathrm{drop}}(m) - V_{\infty,\mathrm{drop}}(m')|\, E(m, m') \tag{16}$$

The breakup term can be parameterized either as a spontaneous breakup (see e.g. Hall, 1980; Srivastava, 1971) or as a collisional breakup (Low and List, 1982a,b).

In solving these equations coupled with the 1, 2, or 3–D dynamics as detailed above, the calculation of the evolution of the drop-size distribution as a function of time and space inside and below the cloud, as well as on the ground is allowed.

However, because of the large number of variables to be updated at every time step, only a few models available can calculate explicit cloud microphysics in a 3-D framework (e.g. Feingold *et al.*, 1994 Kogan, 1991 Khvorostyanov, 1995). Most models are restricted to two dimensions (Hall, 1980) or one dimension (Ogura and Takahashi, 1973). The parcel model discussed above is an exception, since it does not allow precipitation to be calculated. It can, thus, only be used to simulate the evolution of a nonprecipitating cloud (Feingold *et al.*, 1998a), or at the most estimate the amount of precipitable water.

Cold Clouds

Most clouds in midlatitudes form ice particles and the formation of precipitation involves the ice phase. The ice phase can also be modeled by a bin approach if a certain number of additional size distributions are considered – one distribution function to follow the pristine (non or little rimed) crystals: f_{i}, one distribution function for the heavily rimed particles like graupel or hail f_{h}, and one distribution function for snow flakes, that is, aggregates of crystals f_{s}. As ice particles generally can become larger than liquid drops, more categories need to be considered. Also, because of the different possible forms of ice crystals (columns, needles, plates, etc.) additional distribution functions can be added. However, most models follow just one species of crystal shape (plate-like; e.g. Alheit *et al.*, 1990). Only a few models calculate the explicit cold microphysics in three-dimensions (Ovtchinnikov and Kogan, 2000). Most models use reduced dynamics, for example, 2-D (Respondek *et al.*, 1995), 1 – D, or parcel models (Alheit *et al.*, 1990). Takahashi had the first paper on bin-resolved microphysics of mixed-phase clouds in the 1970s. Reisin *et al.* (1996) is another review on mixed-phase bin models of considerable sophistication. It has been used in 2-D models of cumuli and also Arctic stratus clouds (Harrington *et al.*, 2000).

Bulk Parameterizations

Between 1965 and 1975, when cloud modeling started, due to the limitations of early computers, bin modeling in a reasonable dynamic framework was completely out of reach.

Warm Clouds

Consequently, Kessler (1969) proposed a simple parameterization of the microphysics of a warm cloud (Figure 4).

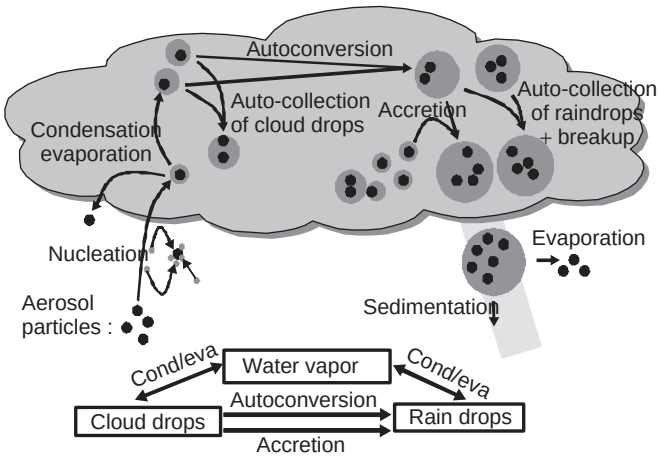


Figure 4 Schematic display of warm cloud processes and the simplification of the Kessler parameterization

It aimed to simulate a convective cloud and just calculated the amount of liquid water attached to the small, nonprecipitating drops (r smaller than 50 μm) and the amount of liquid water attached to precipitation-sized drops (r larger than 50 μm). However, due to its simplicity this scheme was quickly adapted to all types of clouds, and still has its place in many weather prediction models.

The basic equations of this model allow calculation of the evolution with time of the water vapor mixing ratio q_v, the cloud water mixing ratio q_c, and the rain water mixing ratio q_R.

$$\begin{aligned}
\frac{\partial \rho q_v}{\partial t} &= -\mathrm{div}(\vec{v}\rho q_v) + \mathrm{div}(\vec{F}_v^{\mathrm{turb}}) \\
&\quad - C_{\mathrm{con},c} + C_{\mathrm{eva,R}} \\
\frac{\partial \rho q_c}{\partial t} &= -\mathrm{div}(\vec{v}\rho q_c) + \mathrm{div}(\vec{F}_c^{\mathrm{turb}}) \\
&\quad - P_{\mathrm{auto}} - P_{\mathrm{acc}} + C_{\mathrm{con},c} \\
\frac{\partial \rho q_R}{\partial t} &= -\mathrm{div}(\vec{v}\rho q_R) + \frac{\partial}{\partial z} V_R \rho q_R \\
&\quad + \mathrm{div}(\vec{F}_R^{\mathrm{turb}}) + P_{\mathrm{auto}} + P_{\mathrm{acc}} - C_{\mathrm{eva,R}} \qquad (17)
\end{aligned}$$

Here, Kessler (1969) proposes for the condensation rate of cloud drops $C_{\mathrm{con},c}$ a saturation adjustment. This means that all dynamically generated supersaturation is immediately converted to cloud water. And in cases of subsaturation, an amount of cloud water is evaporated that restores saturation, if possible. Obviously, this is a rather crude treatment that does not consider the role of the *CCN* population and the supersaturation.

P_{auto} is the autoconversion rate, that is, the collision of cloud water that forms rain water. It is assumed to be a linear function of the cloud water content, only influenced by threshold values q_{crit} determining the onset of the autoconversion process. In the literature, numerous different values for q_{crit} have been proposed.

$$P_{\mathrm{auto}} = k(\rho q_{\mathrm{crit}} - \rho q_c) \quad k = 10^{-3} s^{-1} \text{ if } \rho q_c > \rho q_{\mathrm{crit}},$$
$$\text{otherwise } k = 0 \qquad (18)$$

The accretion rate P_{acc} details the amount of rainwater created by further collision with cloud water. Kessler (1969) calculates the accretion rate by:

$$P_{\mathrm{acc}} = \int_0^\infty \frac{\mathrm{d}m_{\mathrm{acc}}}{\mathrm{d}t} N_R(r)\,\mathrm{d}r \qquad (19)$$

by calculating the increase of mass of one raindrop m_{acc} of radius r due to the collision with a cloud water population presumed monodisperse represented by q_c with a collision efficiency of E.

$$\frac{\mathrm{d}m_{\mathrm{acc}}}{\mathrm{d}t} = \pi r^2 E V_R(r) \rho q_c \qquad (20)$$

In assuming that the raindrops obey a Marshall–Palmer distribution, (Marshall and Palmer, 1948):

$$N_R(r) = N_{R,0} e^{-2\lambda r} \qquad (21)$$

where λ: shape parameter, and assuming a simple analytical function for the drop terminal velocity:

$$V_R(r) = 130\sqrt{2r} \qquad (22)$$

yielding

$$V_R = \frac{\int_0^\infty V_R(r) N_R(r)\,\mathrm{d}r}{\int_0^\infty N_R(r)\,\mathrm{d}r} \qquad (23)$$

in $\mathrm{m\,s^{-1}}$.

Autocollection of cloud or raindrops are not considered in this approach. The evaporation rate of raindrops below cloud base can be calculated by

$$C_{\mathrm{eva,R}} = -\int \frac{\mathrm{d}m_{\mathrm{con/eva}}}{\mathrm{d}t} N_R(r)\,\mathrm{d}r \qquad (24)$$

Numerous authors have since modified the terms for the parameterizations of the different processes, keeping, however, the basic concept (Berry,1965; Orville and Kopp, 1977, among others).

Cold Clouds

The basic concept of the warm cloud bulk parameterization was soon extended to cold clouds, distinguishing two or three categories of ice. One, corresponding to nonprecipitating cloud drops is called *cloud ice*. For the precipitating

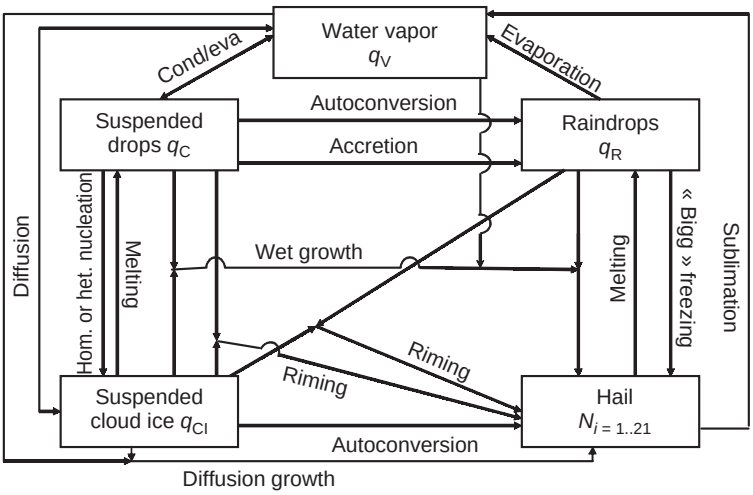


Figure 5 Scheme of the interaction taken into account in a hybrid scheme combining a bulk approach for cloud water, cloud ice, and rain drops and an explicit approach for hail (Adapted from Wobrock *et al.*, 2003)

cloud ice, one can, for example, distinguish *graupel* from *snow*.

An example for cloud ice and graupel can be found in Orville and Kopp (1977), an example for cloud ice, graupel, and snow can be found in Lin *et al.* (1983), and an example using pristine, such as non- or little-rimed crystals (ice A) and rimed crystals (ice B) can be found in Koenig and Murray (1976). A new approach to couple bulk and explicit microphysics can be found in Farley and Orville (1986), which was recently applied to a 3-D model (Figure 5, Wobrock *et al.*, 2003). It should be mentioned that the approach of Farley and Orville is a hybrid approach in which hail is simulated in bins, although continuous accretion approximations are used.

Examples for pure bulk ice physics in a 3-D model can be found in Reisner *et al.* (1998) and for a 2-D model in Grabowski *et al.* (1996).

The Semispectral Microphysics Parameterizations

As a compromise between the bin schemes that follow the spectra of hydrometeors in discrete size classes or bins and the bulk parameterizations that follow just the water mass associated to 5 or 6 classes of hydrometeors, recently, semispectral parameterizations have been developed.

These parameterizations, similar to the bulk ones, follow a limited number of categories. Within one category, however, they assume a size distribution of a log-normal or gamma type, and then calculate the time evolution of the first (total number) and third (mass) moment of these distribution functions. Example of these approaches can be found in Ferrier (1994), Meyers *et al.* (1997), Cohard and Pinty (2000), and Caro *et al.* (2003). Actually, Meyers *et al.* was already extended to the ice phase. Moreover, their approach used approximate solutions or look-up table solutions to the stochastic collection equations as opposed to the continuous accretion approximations to collection. Subsequently, Feingold *et al.* (1998b) extended the emulation of a bin model in a bulk scheme to autoconversion and sedimentation.

More recently Saleeby and Cotton (2004) added a second mode to the cloud droplet spectrum which permitted a more accurate representation of collection (autoconversion) and sedimentation, and permitted the explicit activation of CCN (including a concept for giant CCN) and IN. Others such as extended this approach to a three-moment scheme in which all degrees of freedom in the specified basis functions are predicted. This approach offers an interesting compromise between the detailed bin approach and bulk schemes. The reduction in computational effort will probably mean that even operational forecast models will move towards these schemes in the future.

CUMULUS PARAMETERIZATIONS

If the size of the grid increments becomes larger than 10 km, it is no longer possible to resolve the convective clouds explicitly (stratiform clouds with large horizontal extension can still be resolved). But since the energetic processes associated with convective clouds are important, their net impact on the scale values needs to be parameterized, in addition to the turbulence parameterization already discussed. This gives rise to cumulus parameterizations which can be traced back to the development of numerical prediction models. Smagorinsky (1956) was the first to introduce a cumulus parameterization when he adjusted the vertical derivative of an "effective static stability" which included the heat released during condensation. This gave rise to the first approaches of cumulus parameterizations called *convective adjustment* (e.g. Krishnamurti and Moxim, 1971;

Manabe *et al.*, 1965; Kurihara, 1973). As these approaches turned out too crude, a second generation of schemes attempted to link the convection to the large-scale mass, moisture, or energy convergence, while using some form of "cloud model" (mostly rising plumes and compensating downdrafts) to vertically distribute mass, moisture, and energy by a flux scheme (e.g. Kuo, 1965, 1974; Anthes, 1977, Fritsch and Chappel, 1980; Tiedke, 1989). Another large-scale control was introduced by Arakawa and Schubert (1974). They assumed that the rate of stabilization by an ensemble of cumulus clouds balances the rate at which the large scale makes buoyant energy available for convection. For a complete review of the approaches, see Emanuel and Raymond (1993).

All cumulus parameterization schemes depend severely on the size of the grid box, which they need to stabilize completely or only partially. Consequently, they will only apply to the case for which they were developed and cannot easily be generalized for other grid sizes.

A recent approach, for example, Grabowski *et al.* (2001) and Khairoutdinov and Randall (2001) concerns the development of so-called "super-parameterizations" that have now been implemented in several global circulation models. This approach consists basically in running a 2-D cloud-resolving model at each grid point and average the cloud-resolving model data to determine heating, moistening, and precipitation rates.

NUMERICS

The forgoing physical treatments of the different processes involved in clouds enter the balance equations. These equations, then, need to be solved numerically on a computer. However, before coding, the differential expressions in the terms need to be discretized in order to be linked to the scale variables. The real art of cloud models these days lies here. Numerous studies have been performed on the best numerical scheme to solve the transport equation (positive-definite schemes; e.g. Smolarkiewicz, 1984; Bott, 1989) and the integrals in the drop–drop collision/coalescence process (e.g. Berry and Reinhardt, 1974a-c; Bott, 2000), and it is the best way to distribute the calculated values in a grid, among others. Another serious numerical problem lies in the calculation of the supersaturation due to dynamical and microphysical tendencies. These tendencies should be solved at the same time and not one after the other as in the models. As these tendencies are very sensible, they require extremely small time steps, which in connection with the timescales of the other processes introduce a numerical stiffness problem. Numerous approaches concerning this problem have been proposed in the literature. Next to the explicit method, implicit methods were also developed. Equally, the state-of-the-art cloud models now dispose of an interactive two-way grid nesting. This allows zooming with finer resolution into a domain of special interest and transferring this information to the larger grid outside afterwards. These techniques create a new way to address the scale problem, in following the large-scale dynamics in a coarse grid and then, to zoom with a finer grid into the cloud region. Another technique that advances cloud modeling is the assimilation of remote sensing data (e.g. from satellite) into a model in order to improve model initialization. Furthermore, the calculated evolution can be constrained by observation or outputs of other models. For a summary of current modeling techniques, see Jacobson (1999).

CONCLUSION

This article attempts to provide an introduction to the complex subject of modeling clouds, the production of precipitation, and the development of cloud and storm systems. The elements intervening in cloud modeling have been exposed, starting from a description of the physical phenomena. On the basis of the occurring scale problem, a number of approaches for simplification were presented. These parameterizations are generic and can be combined as a function of the addressed problem. In addition, other modules that have not been discussed in the article need to be added, such as a radiation model, a surface model, a chemistry model, and so on. The main problem is achieving a balance between the degree of parameterization, the number of physical and chemical processes relevant to a particular problem, and the available computing resources. Owing to the increasing capacity of computers, however, less and less parameterizations are necessary and the resulting models are more and more complete. Consequently, the current state-of-the-art MSMs (e.g. RAMS, MM5, Clark, etc.) have large capacities. Essentially the same models can be used to study problems as diverse as rainfall over large areas and over many hours, resulting from frontal cloud systems, the development of local storms with damaging hail and flooding, transport of pollutants by small convective clouds, and photochemical reactions in fog. Next to these models a variety of different models exist which are more or less restricted to a specific question.

REFERENCES

Alheit R.R., Flossmann A.I. and Pruppacher H.R. (1990) A theoretical study of the wet removal of atmospheric pollutants. Part IV: The uptake and redistribution of aerosol particles through nucleation and impaction scavenging by growing cloud drops and ice particles. *Journal of the Atmospheric Sciences*, **47**, 870–887.

Anthes R.A. (1977) A cumulus parameterization scheme utilizing a one-dimensional cloud model. *Monthly Weather Review*, **105**, 270–286.

Arakawa A. and Schubert W.H. (1974) Interaction of a cumulus cloud ensemble with the large-scale environment. Part I. *Journal of the Atmospheric Sciences*, **31**, 674–701.

Asai T. and Kasahara A. (1967) A theoretical study of the compensating downward motions associated with cumulus clouds. *Journal of the Atmospheric Sciences*, **24**, 487–496.

Berry E.X. (1965) Cloud droplet growth by collection. *Journal of the Atmospheric Sciences*, **24**, 688–701.

Berry E.X. and Reinhardt R.L. (1974a) An analysis of cloud drop growth by collection. Part I: Double distributions. *Journal of the Atmospheric Sciences*, **31**, 1814–1824.

Berry E.X. and Reinhardt R.L. (1974b) An analysis of cloud drop growth by collection. Part II: Single initial distributions. *Journal of the Atmospheric Sciences*, **31**, 1825–1831.

Berry E.X. and Reinhardt R.L. (1974c) An analysis of cloud drop growth by collection: Part III: Accretion and self-collection. *Journal of the Atmospheric Sciences*, **31**, 2118–2126.

Bott A. (1989) A positive definite advection scheme obtained by nonlinear renormalisation of the advective fluxes. *Monthly Weather Review*, **117**, 1006–1015.

Bott A. (2000) A flux method for the numerical solution of the stochastic collection equation: extension to two dimensional particle distributions. *Journal of the Atmospheric Sciences*, **57**, 284–294.

Caro D., Wobrock W., Flossmann A.I. and Chaumerliac N. (2003) A two-moment parameterization of aerosol nucleation and impaction scavenging for warm cloud microphysics: description and two-dimensional results. *Atmospheric Research*, **70**(3–4), 171–208, DOI 10.1016/j.atmosres.2004. 01.002.

Clark T.L. (1977) A small-scale dynamic model using a terrain-following coordinate transformation. *Journal of Computational Physics*, **24**, 186–215.

Cohard J.-M. and Pinty J.-P. (2000) A comprehensive two-moment warm microphysical bulk scheme. Part I: description and selective tests. *Quarterly Journal of the Royal Meteorological Society*, **126**, 1815–1842.

Cotton W.R. and Anthes R.A. (1989) *Storm and Cloud Dynamics*, Academic Press: San Diego, p. 880.

Cotton W.R., Pielke R.A. Sr, Walko R.L., Liston G.E., Tremback C.J., Jiang H., McAnelly R.L., Harrington J.Y., Nicholls M.E., Carrio G.G., McFadden J.P. (2003) RAMS 2001: current status and future directions. *Meteorology and Atmospheric Physics*, **82**, 5–29.

Dudhia J. (1993) A non-hydrostatic version of the PENN State-NCAR meso-scale model: validation tests and simulation of an Atlantic cyclone and cold fronts. *Monthly Weather Review*, **121**, 1493–1413.

Emanuel K.A. and Raymond D.J. (1993) The representation of cumulus convection in numerical models. *Meteorological Monographs*, Vol. 24, American Meteorological Society.

Farley R.D. and Orville H.D. (1986) Numerical modelling of hailstorms and hailstone growth. Part I: Preliminary model verification and sensitivity test. *Journal of Climate and Applied Meteorology*, **25**, 2014–2035.

Feingold G., Kreidenweis S.M. and Zhang Y. (1998a) Stratocumulus processing of gases and cloud condensation nuclei 1. Trajectory ensemble model. *Journal of Geophysical Research*, **103**, 19527–19542.

Feingold G., Stevens B., Cotton W.R. and Walko R.L. (1994) An explicit cloud microphysics/LES model designed to simulate the Twomey effect. *Atmospheric Research*, **33**, 207–234.

Feingold G., Walko R.L., Stevens B. and Cotton W.R. (1998b) Simulations of marine stratocumulus using a new microphysical parameterization scheme. *Atmospheric Research*, **47–48**, 505–528.

Ferrier B. (1994) A double moment multiphase four-class bulk ice scheme. Part I: description. *Journal of the Atmospheric Sciences*, **51**, 249–280.

Fletcher N.H. (1962) *The physics of rainclouds*, Cambridge University Press: London.

Flossmann A.I., Hall W.D. and Pruppacher H.R. (1985) A theoretical study of the wet removal of atmospheric pollutants. Part I: the redistribution of aerosol particles captured through nucleation and impaction scavenging by growing cloud drops. *Journal of the Atmospheric Sciences*, **44**, 2912–2923.

Flossmann A.I. and Laj P. (1998) *Aerosols, gases and microphysics of clouds*, ERCA, Vol. 3, Boutron C.F. (Ed.) EDP Sciences: pp. 90–119.

Fritsch J.M. and Chappel C.F. (1980) Numerical Prediction of Convectively driven mesoscale pressure systems. Part I: convective parameterizations. *Journal of the Atmospheric Sciences*, **37**, 1722–1762.

Grabowski W.W. (2001) Coupling cloud processes with the large-scale dynamics using the cloud-resolving convection parameterization (CRCP). *Journal of the Atmospheric Sciences*, **58**, 978–997.

Grabowski W., Wu X. and Moncrieff M.W. (1996) Cloud resolving modelling of tropical cloud systems during Phase III of GATE. Part I: two-dimensional experiments. *Journal of the Atmospheric Sciences*, **53**, 36843709.

Hall W.D. (1980) A detailed microphysical model within a two-dimensional dynamic framework: model description and preliminary results. *Journal of the Atmospheric Sciences*, **37**, 2486–2507.

Hallett J. and Mossop S.C. (1974) Production of secondary ice particles during the riming process. *Nature*, **249**, 26–28.

Harrington J.Y., Reisin T., Cotton W.R. and Kreidenweiss S.M. (2000) Cloud resolving simulations of arctic stratus Part II: transition-season clouds. *Atmospheric Research*, **55**, 45–75.

Houze R.A. (1993) *Cloud dynamics*, Academic Press: San Diego.

Jacobson M.Z. (1999) *Fundamentals of Atmospheric Modelling*, Cambridge University Press: p. 656.

Junge C.E. (1963) *Air Chemistry and Radioactivity*, Academic Press.

Kessler E. (1969) On the distribution and continuity of water substance in atmospheric circulations. *Meteorological Monographs*, Vol. 32, American Meteorological Society: Boston.

Khairoutdinov M.F. and Randall D.A. (2001) A cloud resolving model as a cloud parameterization in the NCAR community climate system model: preliminary results. *Geophysical Research Letters*, **28**, 3617–3620.

Khvorostyanov V.I. (1995) Meso scale processes of cloud formation, cloud-radiation interaction and their modelling

with explicit cloud microphysics. *Atmospheric Research*, **48**, 1–67.

Koenig L.R. and Murray F.W. (1976) Ice-bearing cumulus cloud evolution. Numerical simulation and general comparison against observation. *Journal of Applied Meteorology*, **15**, 747–762.

Kogan Y.L. (1991) The simulation of a convective cloud in a 3-D model with explicit microphysics. Part I: model description and sensitivity experiments. *Journal of the Atmospheric Sciences*, **48**, 1160–1189.

Krishnamurti T.N. and Moxim W.J. (1971) On parameterization of convective and non convective latent heat release. *Journal of Applied Meteorology*, **10**, 3–13.

Kuo H.L. (1965) On the formation and intensification of tropical cyclones through latent heat release by cumulus convection. *Journal of the Atmospheric Sciences*, **22**, 40–63.

Kuo H.L. (1974) Further studies of the parameterization of the influence of cumulus convection on large-scale flow. *Journal of the Atmospheric Sciences*, **31**, 1232–1240.

Kurihara Y. (1973) A scheme of moist convective adjustment. *Monthly Weather Review*, **101**, 547–553.

Lin Y.L., Farley R. and Orville H. (1983) Bulk parameterization of the snow filed in a cloud model. *Journal of Climate and Applied Meteorology*, **22**, 1065–1092.

Low T.B. and List R. (1982a) Collision, coalescence and breakup of raindrops: Part I: experimentally established coalescence efficiencies and fragment size distribution in break up. *Journal of the Atmospheric Sciences*, **39**, 1591–1606.

Low T.B. and List R. (1982b) Collision, coalescence and break up of raindrops: Part II: parameterizations of fragment size distributions. *Journal of the Atmospheric Sciences*, **39**, 1607–1618.

Manabe S., Smagorinsky J. and Strickler R.R. (1965) Simulated climatology of a general circulation model with a hydrological cycle. *Monthly Weather Review*, **93**, 769–798.

Marshall J.S. and Palmer W.M. (1948) The distribution of raindrops with size. *Journal of Meteorology*, **5**, 165–166.

Meyers M., DeMott P.J. and Cotton W.D. (1992) New primary ice-nucleation parameterization in an explicit cloud model. *Journal of Applied Meteorology*, **31**, 708–721.

Meyers M., Walko R., Harrington J. and Cotton W. (1997) New RAMS cloud microphysics parameterization. Part II: the two-moment scheme. *Atmospheric Research*, **45**, 3–39.

Murray F.W. and Koenig L.R. (1975) Cumulus cloud energetics as revealed in a numerical model of cloud dynamics: Part I. Theoretical development. *Pure and Applied Geophysics*, **113**, 909–923.

Ogura Y. and Phillips N.A. (1962) Scale analysis of deep and shallow convection in the atmosphere. *Journal of the Atmospheric Sciences*, **19**, 173–179.

Ogura Y. and Takahashi T. (1973) The development of warm rain in a cumulus model. *Journal of the Atmospheric Sciences*, **30**, 262–277.

Orville H.D. and Kopp F.J. (1977) Numerical simulation of the life history of a hailstorm. *Journal of the Atmospheric Sciences*, **34**, 1596–1618.

Ovtchinnikov M. and Kogan Y.L. (2000) An investigation of ice production mechanisms in small cumuliform clouds using a 3-D model with explicit microphysics. Part I: model description. *Journal of the Atmospheric Sciences*, **57**, 2989–3003.

Pielke R.A., Cotton W.R., Walko R.L., Tremback C.J., Lyons W.A., Grasso L.D., Nicholls M.E., Moran M.D., Wesley D.A., Lee T.J., *et al.* (1992) A comprehensive meteorological modeling system – RAMS. *Meteorology and Atmospheric Physics*, **49**, 69–91.

Phillips N.A. (1957) A coordinate system having some special advantages for numerical forecasting. *Journal of Meteorology*, **14**, 184–185.

Pruppacher H.R. and Klett J.D. (1997) *Microphysics of Clouds and Precipitation; Second Revised and Enlarged Edition*, Kluwer Academic: p. 953.

Reisin T., Levin Z. and Tzivion S. (1996) Rain production in convective clouds as simulated in an axisymmetric model with detailed microphysics. Part I: description of model. *Journal of the Atmospheric Sciences*, **53**, 497–519.

Reisner J., Rasmussen R. and Bruintjes R. (1998) Explicit forecasting of supercooled liquid water in winter storms using MM5 mesoscale model. *Quarterly Journal of the Royal Meteorological Society*, **124**, 1071–1107.

Respondek P.S., Flossmann A.I., Alheit R.R. and Pruppacher H.R. (1995) A theoretical study of the wet removal of atmospheric pollutants. Part V: the uptake, redistribution and deposition of $(NH_4)_2SO_4$ by a convective cloud containing ice using a two-dimensional cloud dynamics model with detailed microphysics. *Journal of the Atmospheric Sciences*, **52**, 2121–2132.

Rogers R.R. and Yau M.K. (1989) *A Short Course in Cloud Physics*, Pergamon: p. 293.

Saleeby S.M. and Cotton W.R. (2004) A large-droplet mode and prognostic number concentration of cloud droplets in the Colorado State University Regional Atmospheric Modeling System(RAMS). Part I: module descriptions and supercell simulations. *Journal of Applied Meteorology*, **43**, 182–195.

Shiino J. (1983) Evolution of raindrops in an axisymmetric cumulus model. Part I. Comparison of the parameterized with non-parameterized microphysics. *Journal of the Meteorological Society of Japan*, **61**, 629–655.

Silverman B.A. and Glass M. (1973) A numerical simulation of warm cumulus clouds. Part I: parameterized vs non-parameterized microphysics. *Journal of the Atmospheric Sciences*, **30**, 1620–1637.

Smagorinsky J. (1956) On the inclusion of moist adiabatic processes in numerical prediction models. *Berichte des Deutschen Wetterdienstes*, **5**, 82–90.

Smolarkiewicz P.K. (1984) A fully multidimensional positive definite advection transport algorithm with small implicit diffusion. *Journal of Computational Physics*, **54**, 325–362.

Soong S.T. and Ogura Y. (1980) Response of tradewind cumuli to large-scale processes. *Journal of the Atmospheric Sciences*, **37**, 2035–2050.

Srivastava R.C. (1971) Size distribution of raindrops generated by their breakup and coalescence. *Journal of the Atmospheric Sciences*, **28**, 410–414.

Stull R.B. (1991) *An Introduction to Boundary Layer Meteorology*, Kluwer Academic Publishers: p. 666.

Szyrmer W. and Zawadzki I. (1997) Biogenic and anthropogenic sources of ice-forming nuclei: a review. *Bulletin of the American Meteorological Society*, **78**(2), 209–227.

Tiedke M. (1989) A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, **117**, 1779–1800.

Twomey S. (1959) The nuclei of natural cloud formation II. The supersaturation in natural clouds and the variation of cloud droplet concentration. *Geofisica Pura et Applicata*, **43**, 243–249.

Whitby K.T. (1978) The physical characteristics of sulphur aerosols. *Atmospheric Environment*, **12**, 135–159.

Wobrock W., Flossmann A.I. and Farley R.D. (2003) Comparison of observed and modelled hailstone spectra during a severe storm over the Northern Pyrenean foothills. *Atmospheric Research*, **67–68**(1–4), 685–703, DOI 10.1016/S0169-8095(03)00081-4, (Elsevier).

Wobrock W., Flossmann A.I., Monier M., Pichon J.M., Cortez L., Fournol J.F., Schwarzenböck A., Heintzenberg S.J., Laj P., Orsi G., *et al.* (2002) The cloud ice mountain experiment CIME 1998: experiment overview and modelling of the microphysical processes during the seeding by isentropic gas expansion. *Atmospheric Research*, **58**, 231–266.

32: Models of Global and Regional Climate

HANS VON STORCH

Institute of Coastal Research, GKSS Research Centre, Geesthacht, Germany

The concept of climate simulations with quasi-realistic climate models is discussed and illustrated with examples. The relevant problem of deriving regional and local specifications is considered as well.

When we speak about "climate", we refer to the statistics of weather. The statistics of weather can be described by first and second moments, that is, by time means and time variability on different timescales and its spectrum, by covariability between different variables, characteristic patterns, and the like. The climate is thought to be conditioned by external forcing, such as the presence of greenhouse gases in the atmosphere, changing solar output, and other factors. Thus, external forcings cause changes in the statistics of weather, but weather itself varies independently of the presence of changing external factors.

Key methods to unravel the dynamics of the climate system are the analysis of observed data and experimentation with climate models of varying complexity.

"Observed" data cover a wide range of data sets. Examples are *in situ* readings of precipitation, wind-speed and other variables, stream flow in rivers, conventional oceanographic and atmospheric vertical soundings, and also pixel data derived from satellite retrievals, and sophisticated "analyses". The latter are subjective or empirical (kriging-based) spatial interpolations of point observations, or model simulations into which the observed data have been assimilated using the concept of *state-space* modeling. All weather maps are such analyses. Important data sets of 6-hourly weather maps since 1948 or 1960 have been prepared by National Centers for Environmental Prediction (NCEP) and by European Center for Medium Range Forecast (ECMWF) (e.g. Kalnay *et al.*, 1996; Gibson *et al.*, 1997).

Climate models are process-based dynamical models (for further reading, refer to Müller and von Storch, 2004), which operate on the entire globe or in limited regions of the world. Climate models describe several compartments of the climate system, as for instance, the atmosphere, the oceans, the cryosphere, the surface hydrology, the vegetation, or cycles of matter. Thus, climate models may have different qualities of "complexity" – they may describe fewer components, but describe these components in greater detail. GCM-based models (GCM stands for "General Circulation Model"; such models operate with the "Primitive Equations", which describe the relevant atmospheric dynamics in detail – for further information refer to the references given), which are named quasi-realistic in the following, are of that sort (e.g. Washington, 1999). Another modeling strategy is to consider more components but in less detail – an example is the CLIMBER model (e.g. Ganopolski *et al.*, 1997). The former are often called *complex* and the latter *medium complexity* – these terms are in use but are not really precise semantics in describing the differences between the two classes of models.

In the following, we discuss the utility of "quasi-realistic" models. There are many books and articles on this subject. The books by Washington and Parkinson (1986), McGuffie and Henderson-Sellers (1997) and von Storch *et al.* (1999) describe the challenges of numerical modeling on a technical level, while the monograph by Müller and von Storch (2004) deals more with the philosophical problems related to the usage of such models. Also the collection of papers offered by Trenberth (1993) or von Storch and Flöser (2001), the articles by Bengtsson (1997) and Manabe (1997), and the description of the state-of-the-art in the Intergovernmental Panel on Climate Change (IPCC) reports (Houghton *et al.*, 1996, 2001) may be helpful for the interested reader.

Section "Quasi-realistic climate models (surrogate reality)" discusses the construction and validation of quasi-realistic global models, Section "Free simulations and forced simulations for reconstruction of historical climate"

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the performance of such models in reconstructing historical climate, and Section "Climate change simulations" climate change simulations. The problem of how to infer a description of the impact-relevant regional and local climate is dealt with in Section "Downscaling"; the major downscaling tools are regional models; the construction of such models is considered in Section "Regional climate modeling". The success in reconstructing the climate of the past decades of years is demonstrated in Section "Reconstructions", and scenarios of plausible future climate change are discussed in Section "Regional scenarios". The article concludes with the Section "Conclusions".

The examples used throughout the text are chosen subjectively – and with a bias towards work done in the mostly European academic milieu of the author. It would have been equally possible to write this article with a very different set of examples, without compromising the representativity and usefulness of this article.

QUASI-REALISTIC CLIMATE MODELS (SURROGATE REALITY)

Models that can realistically simulate the sequence of weather events are called *quasi-realistic* climate models. They comprise circulation models of the atmosphere and the ocean and other components such as the land surface and sea ice. The components of such a model are sketched in Figure 1 (Hasselmann, 1990).

Such models are complex models – their degree of complexity is a compromise of computation possibilities and the required length of the integration. If the model is supposed to be integrated for 1000 years, then a coarser spatial resolution is chosen and some processes are described in a less detailed manner. For such an integration, a spatial grid size of about 300 km is often used. In order to achieve a higher spatial resolution, so-called downscaling methods have to be applied (see below; von Storch, 1999).

In the climate system, processes are operating at all timescales. On the other hand, the numerical formulation of the dynamical equations requires a cut-off at a certain scale. Figure 2 sketches the situation for atmospheric dynamics – with faster processes on smaller scales, and slower processes on larger scales. The space/time truncation, sketched in Figure 2 by hatching, leads to the disregard of many processes such as cumulus convection. These processes are, however, essential for the formation of the general circulation of the atmosphere – therefore they are included into the numerical equations as "parameterizations". That is, the expected effect of such processes on the resolved processes conditional upon the resolved state is specified. All models, atmospheric and oceanic, global and regional, contain many of these parameterizations, and they are a major cause for the different performance of dynamical models.

The skill of models in describing the real world depends on the spatial scale. Phenomena on larger scales are better described than smaller scales. Grid point values are

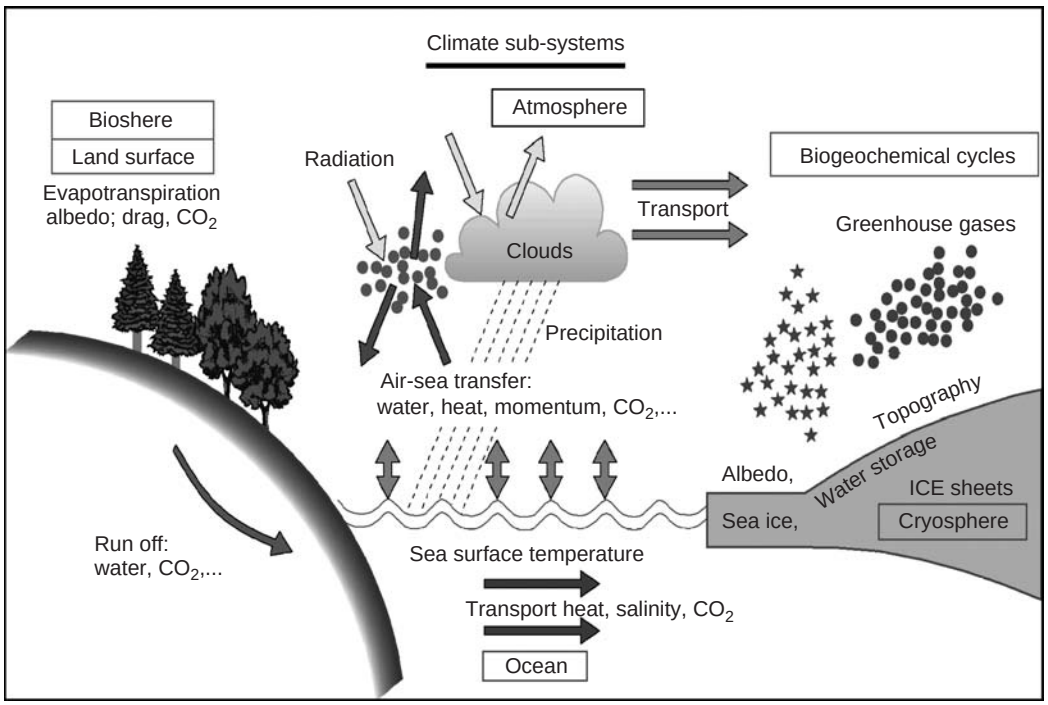


Figure 1 Components of a dynamical climate model (Hasselmann, 1990, © JCB Mohr, Tübingen). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

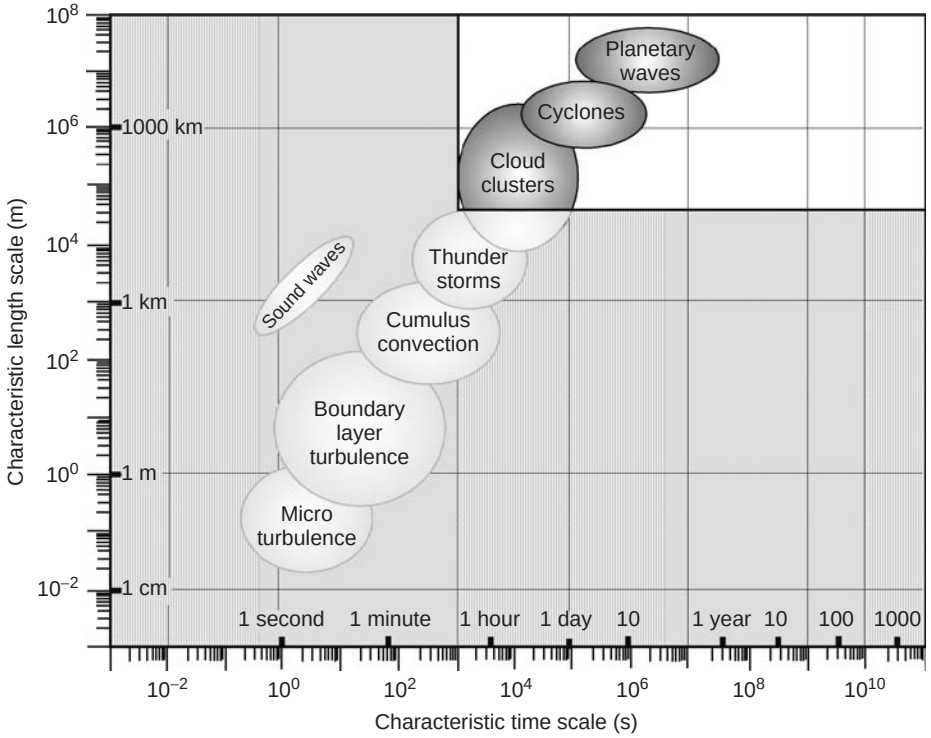


Figure 2 Resolved and unresolved processes and scales in a contemporary atmospheric model (Müller and von Storch, 2004, © Springer Verlag)

usually meaningful only if the variables are smooth so that the grid point value is representing a larger area. However, when the considered variables vary strongly from grid point to grid point, such as rainfall, the sequence of events at a grid point and at the geographic location, formally corresponding to the grid point, likely will not compare well. Grid point values do not represent local values when there is a great amount of spatial variability of scales of the grid size and smaller. For larger areas, represented by many grid boxes, this is no longer a problem.

Such models have been shown to have considerable skill in reproducing many aspects of contemporary climate, such as the annual cycle, the level of stochastic variability, and the formation of extratropical storms. These models strive to be as realistic as possible. Since these models are, nevertheless, significant simplification of the complex real system, we (Müller and von Storch, 2004) use the term *quasi-realistic* for such models.

Figure 3 provides an example of a sophisticated characteristic of atmospheric dynamics, namely, the "storm track" in the North Atlantic. The storm track is conveniently defined by the intensity of the band-pass filtered variance of 500 hPa geopotential height. The variations on timescales between 2.5 and 6 days are shown in the diagrams – variations on these timescales are related to the formation and migration of baroclinic storms. The model generates a pattern and intensity of the storm track (Figure 3, bottom) which is very similar to pattern and intensity derived from ECMWF analyses (Figure 3, top). The intensity in the model output is smaller than in the analyses – but the difference is usually considered acceptable within the range of uncertainties.

Figure 4 provides another example of a validation of GCMs. It compares the performance the analyses of rainfall determined in the ERA-40 data set (prepared by the European Center for Medium Range forecast; ECMWF) with the precipitation simulated in many GCMs (Kharin *et al.*, 2004). Specifically, the spatial distributions of the time-mean precipitation and 20-year return values are studied and compared with their counterparts derived from ERA-40.

The diagram is not easily understood, but it provides a compact description of the skill of a set of models. In the first step, the spatial average of the spatial distributions is subtracted – so that "anomaly fields" are obtained. From these anomaly fields, three characteristic numbers are calculated and displayed in Figure 4 by one symbol for each model. The trick is that in this diagram three characteristic numbers are displayed by one symbol in a two-dimensional diagram.

- The *mean-squared difference* between the anomaly field of the considered model and the ERA-40 reference

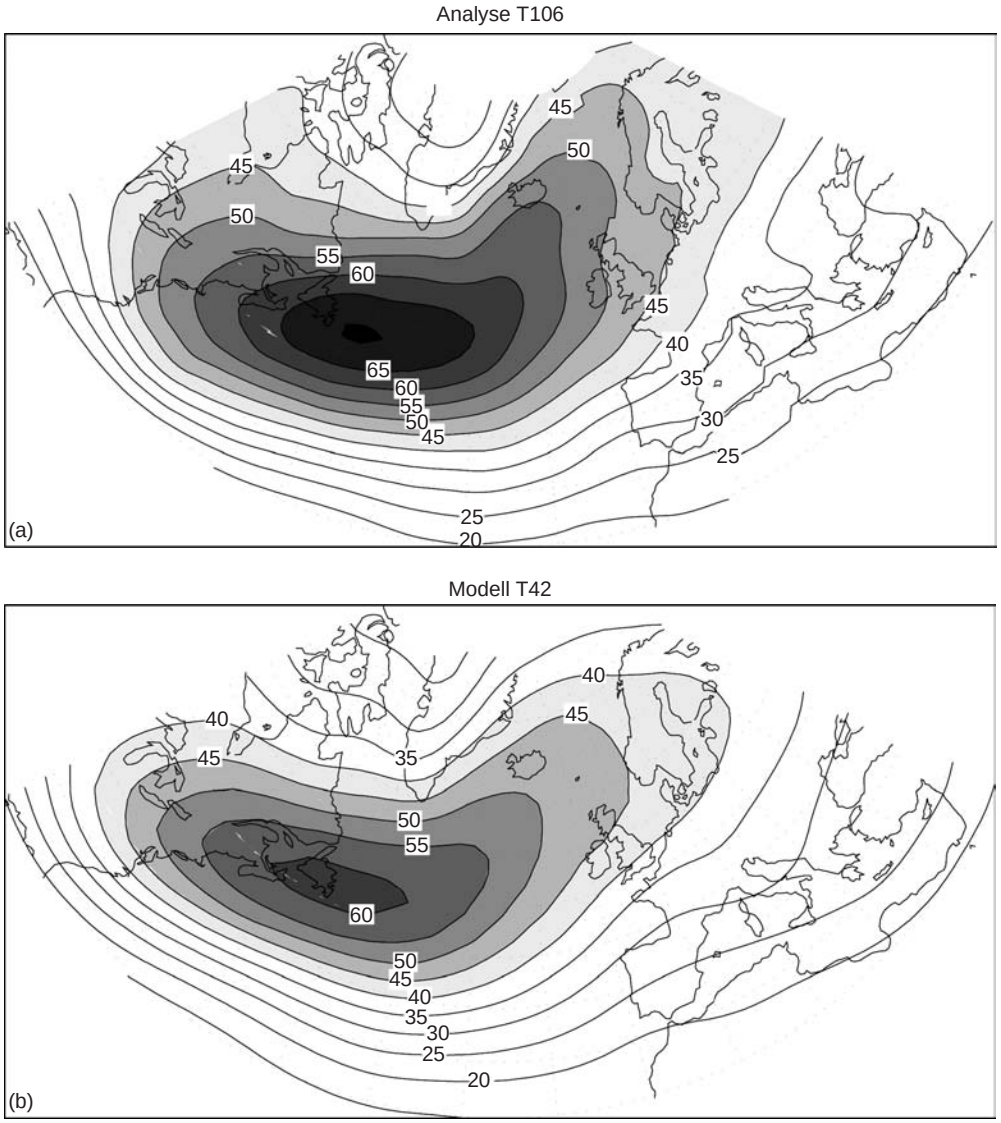


Figure 3 North Atlantic storm track as given by the band – passed filtered variance of 500 hPa geopotential height (band – pass: 2–5 to 6 day variability is retained) in ECMWF analyses (a) and in an extended simulation with the T42 ECHAM3 model (b) (von Storch *et al.*, 1999, © Springer Verlag)

anomaly. To facilitate easier comparison, the mean-squared difference is normalized by the variance of the reference anomaly. This characteristic number is given by the light blue circles emanating from the horizontal axis, where the units are given.

- The *ratio of variances* of model anomaly and of the reference anomaly (given by green dashed circles emanating from right vertical axis).
- The *correlation* indicated of the model and reference anomalies. This is given by the by pink straight lines emanating from left vertical axis.

The ERA-40-reference itself has a ratio of variances of one, a mean-squared difference of zero and a pattern correlation of one – its dot is placed on the lower margin of the diagram.

Usually the spatial variance of the simulated *time means* is underestimated (dashed circles, 50–80%), while the normalized mean-squared difference is moderate (light blue circles, 20–40%). The pattern correlation is high (typically 80%). The spread for the *20-year return values* is much larger (circles in Figure 4). Some of the models are doing similarly well as the mean precipitation, while other models produce a much too small variability (less than 20%) but a large high mean-square difference (100%). The pattern correlation is less than 80%, in some cases, as little as 50 and less percent. Interestingly, the comparison of the ERA-40 analysis with other analyses by NCEP

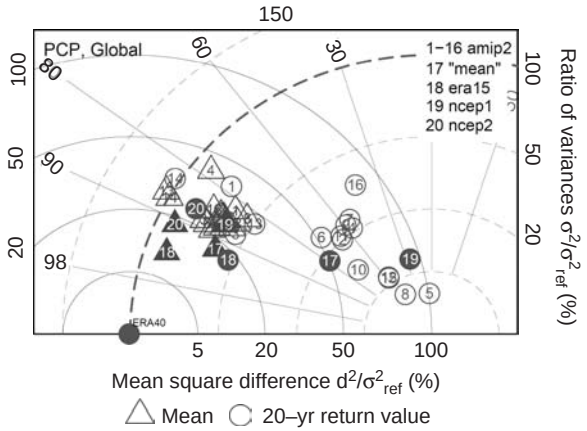


Figure 4 Comparison of the simulation of the time mean (triangles) and 20 year return values (circles) of precipitation in a series of 16 GCMs with the ERA-40 analysis (Kharin *et al.*, 2004). Also shown are the comparison with two NCEP-reanalyses, an earlier, shorter ERA reanalysis and the average of all 16 considered GCMs. Three characteristic measures are shown – the mean-squared difference, the ratio of spatial variances and the anomaly correlation. The mean-squared difference of the ERA-40 anomaly field and the model anomaly field is indicated by the light blue circles emanating from the units given on the horizontal axis. This parameter is normalized by the spatial variance of the reference anomaly field. The ratio of spatial variances of the reference and model anomaly fields is given by the green dashed circles emanating from the vertical axis. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

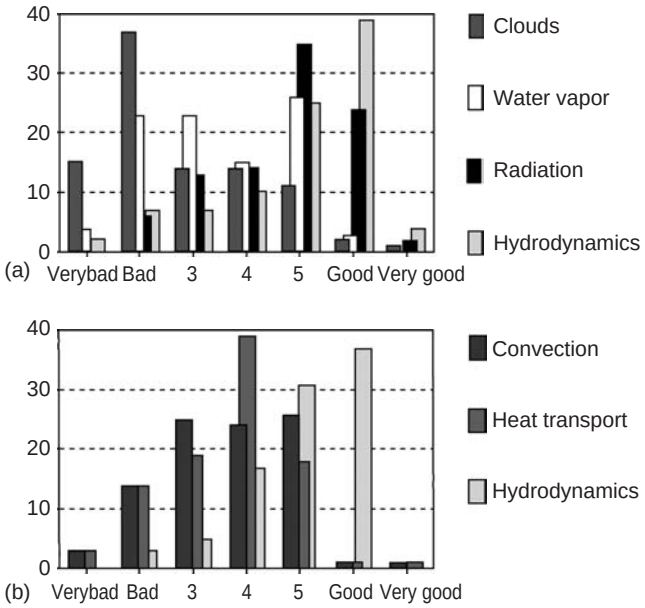


Figure 5 Result of a survey among climate modelers on the confidence into the description of processes in atmospheric (a) and oceanic (b) models. Answers were requested on a scale varying between ''very bad'' and ''very good''. The units on the vertical axis are in percent. (Bray and von Storch, 1999, © 1999 AMS)

(numbers 18 and 19) indicates substantial difference among reanalyzes. Obviously, the estimation of precipitation in reanalyzes provides further improvements, if space/time details are needed.

Figure 5 displays the outcome of a survey among 104 climate modelers, who have been asked to subjectively assess the skill of contemporary climate models in the end of the 1990s in describing a number of processes (Bray and von Storch, 1999). They were requested to respond to a seven-graded scale, varying between "very good" and "very bad". For obvious reasons, the response "very good" is almost never heard. Hydrodynamics, that is, the implementation of the laws of conservation of mass and momentum, is considered to be well reproduced. However, thermodynamic processes, related to convection or clouds, are assessed by many experts as being insufficiently represented. Of course, this assessment is partly reflecting the wish of modelers to continue their work in improving their models, but the outcome of the survey is also strong evidence that models really need to be improved.

Quasi-realistic models are considerably less complex than reality, but nevertheless, *very* complex. They can react in ways that cannot be foreseen by simple conceptual models. This is a virtue of such models, as they make them to laboratories to test hypothesis with – they constitute a *virtual* or *substitute reality* (Müller and von Storch, 2004).

FREE SIMULATIONS AND FORCED SIMULATIONS FOR RECONSTRUCTION OF HISTORICAL CLIMATE

Climate models are run in different modes. There are "free" simulations and "forced" simulations.(The wording "free" and "forced" is somewhat misleading. All climate simulations are, in a sense, forced, as they are exposed to a series of prescribed factors external to the model. In case of "free" simulations, these external factors do not vary, except for a fixed annual cycle. Variations in "free" simulations are therefore entirely due to the internal dynamics of the model and cannot relate to specific external factors. In contrast, "forced" simulations respond to a forcing, which varies irregularly. Thus, such model simulations exhibit a mix of externally induced variability and internally generated variability.) The former are useful to generate purely internal variability, whereas the latter allow the analysis of the effect of external factors. Figure 6 shows an example of a free simulation. The climate model ECHO-G was integrated over 1000 years – with continuously repeated annual cycles of solar insolation and no other external factor

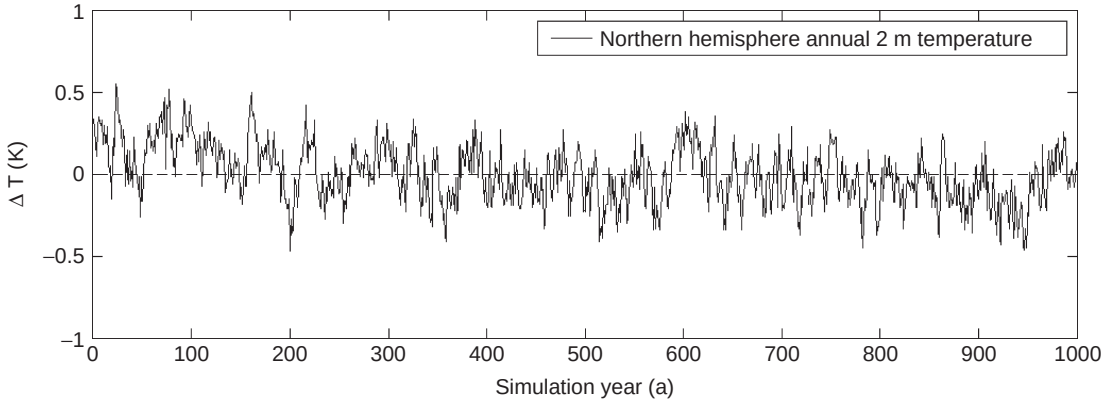


Figure 6 Air temperature anomalies (deviation form the long-term mean) simulated in a 1000-year ''free'' simulation (Wagner *et al.*, 2005, © Springer Verlag)

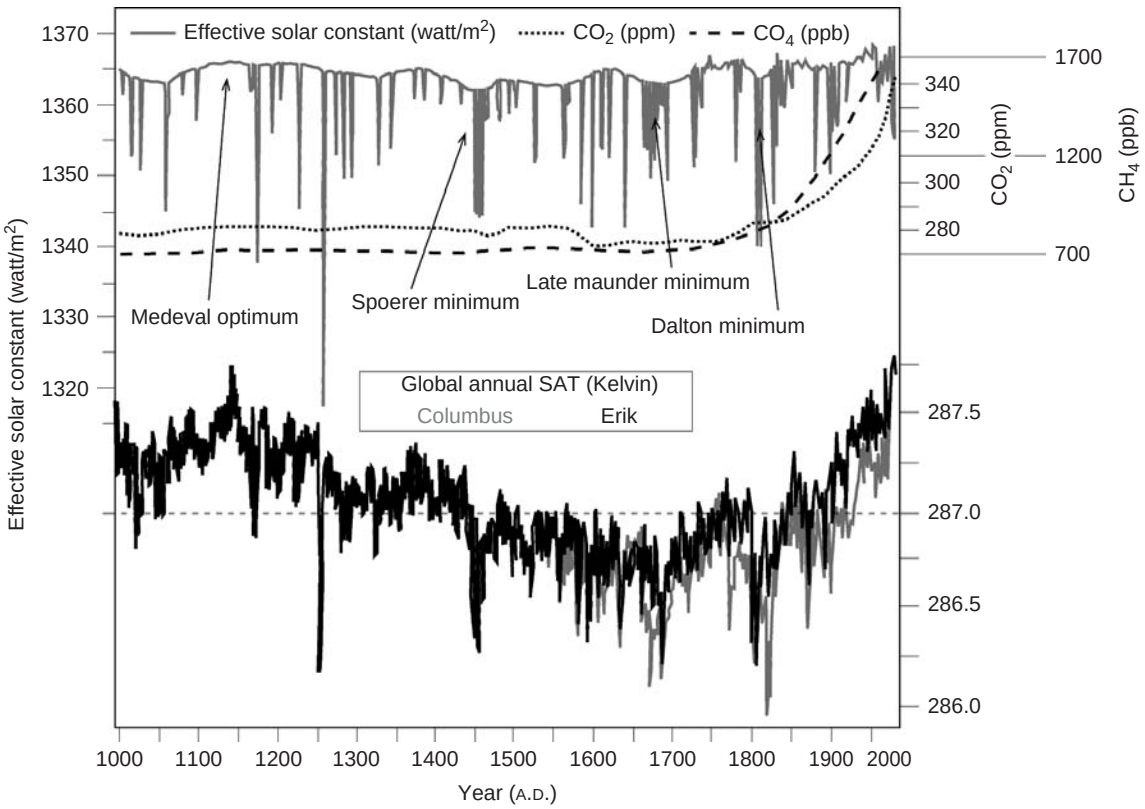


Figure 7 Time variable forcing (a) and temperature (b) in a 1000-year simulation and in a 500-year ''forced'' simulation

(Wagner *et al.*, 2005). The variable shown is air temperature averaged across the Northern Hemisphere. Obviously, the temperature undergoes significant variations, which cannot be traced back to "causes". The reason for this "smoke without fire" effect is the presence of myriads of nonlinear chaotic processes. The sum of all these chaotic processes may be conceptualized by the mathematical concept of stochastic noise (e.g. von Storch *et al.*, 2001). This noise is integrated by the slow components in the climate system, so that variations on all timescales appear with a first-order approximation red spectrum (Hasselmann, 1976).

On the other hand, characteristic cause-and-effect features emerge when external time-variable factors are added (González-Rouco *et al.*, 2003; Lionello *et al.*, 2004). Figure 7 shows time series of atmospheric forcing by time-variable solar output and the effect of stratospheric volcanic

aerosols, and by variable atmospheric concentration of the two greenhouse gases carbon dioxide and methane. Here, the radiative effect of the volcanic aerosols is accounted for by reducing the solar insolation for a short time. The time series of global mean air temperature is also displayed. This series is composed of variations unrelated to the forcing, like in the free simulation, and to variations excited by the forcing. A close inspection reveals that the variable output of the sun (including the volcanic effect) is the dominant factor until the middle of the nineteenth century. Since then, the effect of the ever-increasing greenhouse gas concentration is becoming dominant.

The overall development of the simulated temperature during the last millennium is consistent with the historical account, but the range of the variations is larger than what has been reconstructed from proxy-data, like tree rings.

The emergence of variability unrelated to external forcing factors makes also a forced simulation to a random experiment – the resulting weather stream is not determined by the forcing, but *conditioned*. For instance, the details of cyclones and anticyclones will vary from one simulation to the next, but the statistics of the formation of cyclones and anticyclones will be similar in any two realizations. In order to get a robust statistic, several simulations with identical forcing are preferable (*ensemble* simulations). To make them different realizations, several measures are possible; a popular method is to use a slightly different initial state.

CLIMATE CHANGE SIMULATIONS

In climate change simulations, assumed changes of the forcing are administered to the model. These changes are "scenarios" of possible and plausible changes. In most cases they refer to the emission of greenhouse gases, sometimes to the emission of anthropogenic aerosols. These emissions themselves are based on scenarios of economic and social development. The output of the climate models is then named a "climate change scenario" of a possible and plausible future climate.

The scenarios are not predictions; they do not describe the most probable development; instead usually several different scenarios are presented, which differ significantly from each other. Scenarios are plausible and consistent images of a possible future; have an impact on the future itself. (The movie "The day after tomorrow" provides a story of future climate change; it is, however, not the scenario as it describes a climate which is impossible to emerge as it is not consistent with the physical laws of climate.) Thus, scenarios are not only depictions of possible futures, but also active agents forming the future. (In the context of global warming, various scenarios of possible future development are prepared to emphasize the severity of the threat of global warming. As such, they help the formation of a climate policy mitigating the envisaged anthropogenic climate change.)

The consensus of the models given a specific emission scenario on larger scales is illustrated in Figure 8. A total

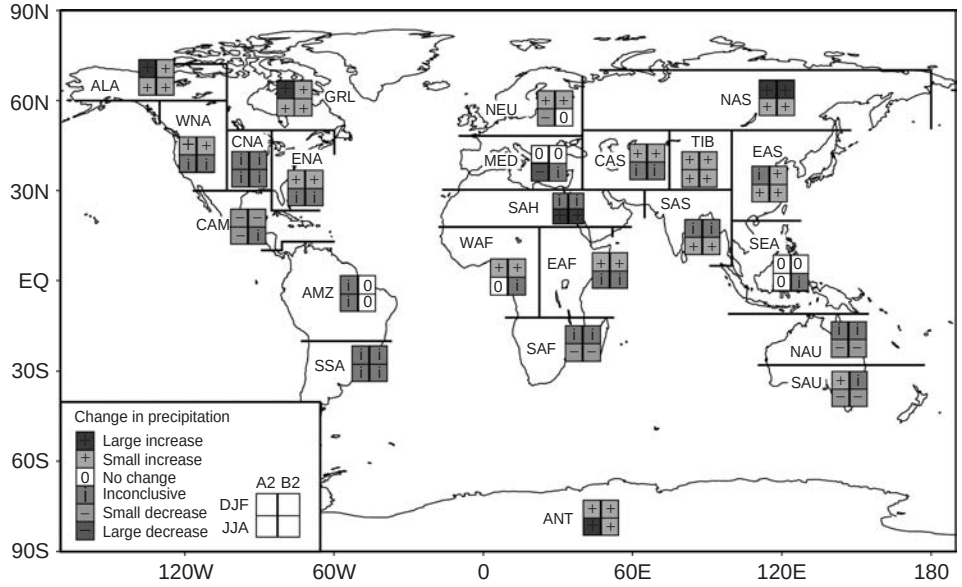


Figure 8 Convergence of climate models in simulating the same regional change in precipitation according to the greenhouse gas emission emissions scenarios A2 (left two boxes) and B2 (right two boxes) at the end of the twenty-first century. The top two boxes refer to northern winter (DJF), the bottom to the northern summer (JJA). For further details, refer to the text (Reproduced from Giorgi *et al.,* 2001a by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of 9 models was analyzed with respect to their similarity in the change of precipitation averaged over subcontinental areas (Giorgi *et al.*, 2001a). All nine models have been forced with the same Special Report on Emissions Scenarios (SRES) scenarios provided by the IPCC for its Third Assessment Report. Whenever 7 out of the 9 models produced similar responses, the models were considered agreeing in envisaging a small or large increase or small or large decrease of precipitation, or no change in precipitation. The symbol "i" was introduced in Figure 8 if the models were found to generate conflicting assessments. This exercise was done for a series of subcontinental areas, for both scenarios A2 and B2 and for the two seasons December-January-February (DJF) and June-July-August (JJA) (A2 is a scenario, which describes a rather steep increase in the usage of fossil fuels and emissions into the atmosphere. B2, on the other hand, assumes more efficient measures to curb emissions.). For each of the regions, a square provides the assessment for the two scenarios and two seasons. Obviously, the models agree in most cases – and indeed the pattern of responses is the same, also if a number of earlier climate change experiments exploiting somewhat different scenarios are checked (not shown; Giorgi *et al.*, 2001a).

The argument seems to indicate that the similarity among models would be a proof for the reality of the response. This is certainly not so; the arguments certainly demonstrate the stability of the response across models – but since the models are not developed independently of each other, they may all suffer from the same limitations.

The climate change simulations provide useful information on large scales. Here "large" means global, continental, and subcontinental scales. A rough rule is that contemporary models are skillful on scales of 10^7 km^2. On smaller scales, the model output will often depend on the specifics of the considered model.

This is insofar a severe limitation as the effect of changing climate is felt on a regional scale; assessing the impact of climate change requires scenarios on the regional and even local scale. Thus, extra efforts are required to derive the required impact-relevant regional scenarios. Tools for that purpose are discussed next.

DOWNSCALING

The idea of downscaling is that the smaller scale climate may be understood as the outcome of an interaction of larger-scale dynamics and smaller-scale physiographic details (e.g. von Storch, 1999). The concept is based on the observation that the global scale circulation is already formed on an aqua planet without any physiographic features (a planet entirely covered by the ocean, without any land and topography); the formation of stationary planetary features needs the presence of the gross land–sea contrast and the largest mountain ranges.

There are several downscaling methods in use (Giorgi *et al.*, 2001b).

One main group is utilizing empirically determined transfer functions, which relate variables of regional or local interest to well-simulated large-scale variables (for an overview, refer to Giorgi *et al.*, 2001). Such transfer functions are often regression equations, but also nonlinear techniques like neural nets are in use. Sometimes the transfer functions relate statistical parameters to each other, such as intramonthly percentiles of an impact variable and monthly mean air pressure fields. Another approach is to directly relate meteorological state variables like free tropospheric temperature and humidity to relevant surface variables (Wilby and Wigley, 1997). Weather generators are also used for downscaling, with their parameters conditioned on the large-scale state (e.g. Busuioc and von Storch, 2003).

The other group of methods is based on the use of regional climate models (Giorgi and Mearns, 1991). In most cases, the 6-hourly large-scale weather stream generated by a global climate model is enforced on the regional domain; the dynamical model is constructing a regional-scale weather stream, which is consistent with both the global weather stream and the physiographic details of the considered region. In the following, we will deal with this approach in more detail.

One has to keep in mind that downscaling operates with the assumption that the large scales are properly represented by the global simulation or the global analysis. This is usually not a problem in case of analyses, but for (free or forced) global simulations this is a nontrivial assumption. For instance, the formation of blocking situations, which may be considered large scale in certain downscaling applications, is not sufficiently simulated if the global model has too low a resolution.

REGIONAL CLIMATE MODELING

Regional climate modeling (for a recent overview, refer to Wang *et al.*, 2004) is in most cases just regional atmospheric modeling with some basic parameterization of the thermodynamics of the upper soil layer. The other climatically relevant state at the surface of the earth, in particular, the sea surface temperature, sea ice and lake ice conditions, the state of the vegetation – are in these cases prescribed. Since a few years, significant efforts are made to construct coupled regional models, which feature regional oceans and lakes, run-off or vegetation explicitly together with the regional atmosphere. For instance, the model system BALTIMOS (Jacob; personal communication), designed for Baltic Sea catchments studies, is made up of the Baltic Sea ocean model, a hydrological model and the regional atmospheric model REMO (Jacob *et al.*, 1995). The Swedish Rossby Center is working with

a system featuring a Baltic Sea model, regional hydrology, and a regional atmosphere (Räisänen *et al.*, 2004)

Commonly, the regional models are forced by boundary conditions along the lateral boundaries and, as discussed above, at the surface of water bodies. The lateral conditions are enforced with the help of a "sponge" zone (Davies, 1976) of a few grid points. In the sponge zone, the simulated state is nudged to the externally given state, with stronger nudging coefficients near the model's margin, and weaker ones in the interior. That this concept is practically working fine has been demonstrated convincingly by the "Big Brother Experiment" by Denis *et al.* (2002). In that experiment, a 50-km grid, regional model covering a large area ("big brother") was run over an extended time; then a smaller domain within the larger domain was chosen. The meteorological variables simulated in the large domain along the margin of the small area were selected. The same regional model, with a 50-km grid, was then run on the smaller domain ("little brother"), forced with the boundary values provided by the large-area model after "coarsening", that is, the data was not given every 50 km along the margin, but every 100, 200 or even 500 km. The research question was whether the fine scale features simulated in the smaller domain in the big brother setup would be recovered by the little brother setup. The answer was positive; after a few days, differences between the large-area simulation and the small-area simulation were small. The area considered was the well-flushed Eastern North America and Western North Atlantic. (A region is "well flushed" if the information is quickly advected from the boundaries into and through the interior.)

Mathematically, the problem of inferring the dynamical state of a fluid by providing lateral boundary conditions is not a well-posed problem. The lateral boundaries do not determine a unique "solution" in the interior; instead, several different states in the interior are consistent with a given set of lateral boundary conditions. The tendency to form very different solutions in the interior as a response to the same boundary conditions depends on how well the region is flushed, that is, how efficiently a boundary steering is established. In midlatitudes, such as Europe or in Denis *et al.*, (2002) case, the regions are mostly well flushed; in areas with little "through-flow", like the Arctic, this is not so. Thus, any two extended simulations that are run with the same boundary values but slightly different initial states (which may simply be two observed states 12 hours apart) will generate more or less frequently very different behavior. For a region like Europe, such a "divergence" is rare (Weisse *et al.*, 2000), but Caya and Biner (2004) report a dramatic case in eastern North America. For the Artic, such divergence is more frequent (Rinke and Dethloff, 2000). This phenomenon of intermittent divergence is reflecting the conflicting influences of control by inflow boundary conditions and of regional chaotic dynamics.

Figure 9 shows an example of this intermittent divergence. The observed zonal wind at a location in the German Bight is shown together with grid-box-simulated zonal winds. (There is a problem of comparing local wind affected by local particularities with grid box averages. Deviations between simulated numbers and observed numbers may be due to local effects not described by the model's resolution.) In this case, six simulated time series are shown. They are generated by the same regional model, forced with the same lateral boundary conditions but with slightly different initial values. The reason for these different developments is not that the initial conditions would be very different, leading to different forecasts; instead miniscule differences in the initial conditions excite the chaotic divergence of the dynamical system "regional atmosphere". After a few days, after January 8, the divergence has ceased and the development is the same in all six simulations. This convergence following an episode of divergence reflects the fact that the system moves into a configuration with a more westerly weather regime, so that the information provided with lateral boundary conditions is efficiently "flushed". After several months, a similar divergence episode emerges (not shown).

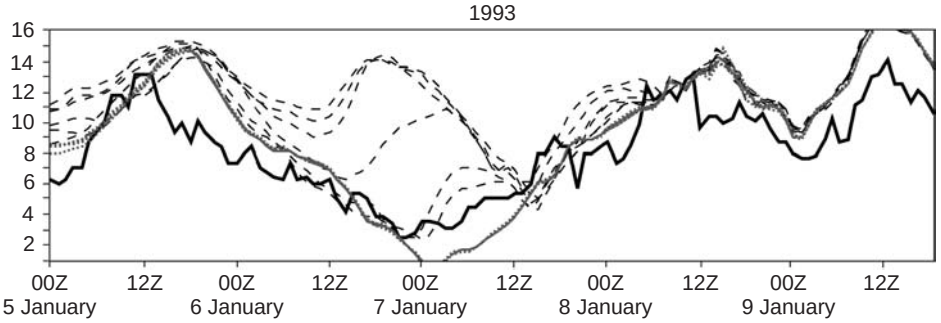


Figure 9 Intermittent divergence in a regional atmospheric model. Shown is the zonal wind at a location in the German Bight, as observed (solid) and as simulated in six simulations with a regional model with conventional lateral boundary forcing (dashed) and with spectral nudging (grey) (Reproduced from Weisse and Feser, 2003, by permission of Elsevier)

A method to overcome this intermittently emerging divergence is to cast the whole regional modeling problem not as a boundary problem but as a state space problem (e.g. Müller and von Storch, 2004), in which the dynamical model is used to augment existing knowledge about the regional state of the atmosphere. The latter is knowledge about the large-scale state of the atmosphere above a certain vertical level, where the influence of the regional physiographic details is small. This concept leads to *spectral nudging* (von Storch *et al.*, 2000; Miguez-Macho *et al.*, 2004), which consists of the addition of penalty terms in the equations of motion. These terms are getting large if the simulated large-scale state deviates from the prescribed large-scale state, but vanish if the model remains close to the prescribed large-scale state. The method has been tested, and it is found that this approach is better in capturing regional details than when the model is forced only with lateral conditions (a case demonstrating this claim is provided by von Storch *et al.*, 2000). In addition, the emergence of intermittent divergence is suppressed (Weisse and Feser, 2003). This is demonstrated by the other set of curves in Figure 9, displaying the development in the six simulations, starting with the same set of initial conditions as in the conventional lateral boundary forcing cases – the curves differ so little that they appear as a somewhat broader grey line.

RECONSTRUCTIONS

One important application of regional models is the high-resolution reconstruction of the weather stream of the past 40 or 50 years (Feser *et al.*, 2001 – "Feser reconstruction"). Using the spectral nudging technique, using the European weather stream on spatial scales of 1500 km and more, reliably analyzed by NCEP since 1948 (Kalnay *et al.*, 1996) as constraint, regional details of the atmospheric state were reconstructed continuously with the regional atmospheric model REMO for 40 and more years on a 50-km grid. The data was stored once an hour.

The added value of this exercise is an increased resolution in space and time; thus it is expected that the tails of the distributions (i.e. of climate) are better described. Figure 10 demonstrates that this improvement has indeed been achieved, at least for wind over the sea (Sotillo, 2003). Quantiles are derived for wind-speed-time series recorded at two buoys. They are compared with quantiles derived from the NCEP reanalysis and from the Feser-reconstruction. In one case, both model quantiles are very similar to the observed quantiles; in the other, only the Feser-reconstruction exhibits the right level of strong windiness. In the former case, the buoy data have entered the NCEP reanalysis, but in the latter the buoy data have not. Thus, it may be concluded that the reconstruction using a regional atmospheric reconstruction together with a spectral nudging approach, is recovering relevant detail to regional climate statistics. However, further analysis of the added value, in particular, in terms of precipitation and wind over land, needs to be done.

This added value is used in assessment studies, for instance, about ocean wave conditions (EU project HIPOCAS; Soares *et al.*, 2002). An example of successfully reproducing local wave conditions at an island in the North Sea is shown in Figure 11. The high wave results, obtained as response to the Feser-winds, are in very good agreement with the local observations, recorded either by a local buoy or by a local radar system. In fact, the wind data set is being increasingly used by regional decision makers.

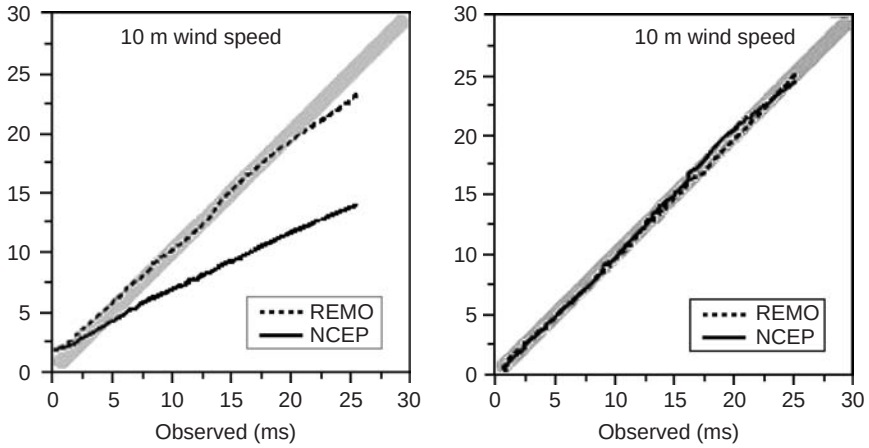


Figure 10 Quantile–quantile plot of 10 m-wind speed at two buoy locations in the east Atlantic (a) and in the Ionian Sea (b). The vertical axis represents the quantiles from the buoy data, the vertical the quantiles derived either from NCEP reanalyses (solid) or from the regional model reconstruction (dashed). Note that the data from the Atlantic buoy has been assimilated into the NCEP reanalysis, while the data from the Ionian Sea buoy are independent (Sotillo, 2003)

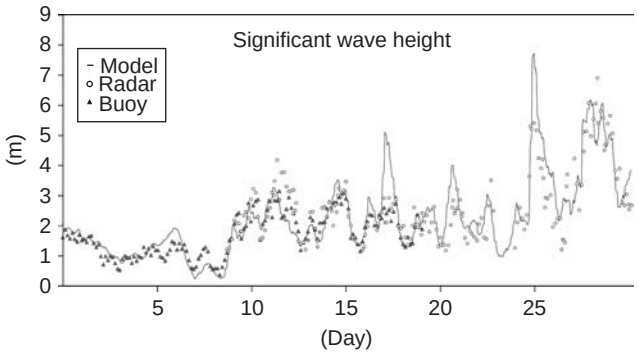


Figure 11 Simulated significant wave height off the island of Heligoland in the German Bight during one month. The line is the significant wave height obtained by running a wave model with the Feser-reconstructed winds in the North Sea domain; the triangles are wave height recorded by a local buoy and the open dots estimates derived from a local radar system (By courtesy of Gaslikova)

The data set has so far not been systematically studied with respect to the reconstruction of precipitation. However, it is already clear that a one-to-one association of grid box precipitation in the model to 50 km by 50 km real averages derived from observations is not possible. Only averages over several grid boxes are meaningful. Figure 12 shows an example for the catchment of the river Odra (Messal, personal communication). The similarity is not perfect but encouraging, considering the uncertainty in the "observed" rainfall. Other aspects which have been examined are related to cloud cover and cloud amounts (Meinke *et al.*, 2004)

REGIONAL SCENARIOS

In the European project PRUDENCE (Christensen *et al.*, 2002), the same set of global climate change scenarios are processed with a large number of regional climate models. Most of the models are purely atmospheric models, but some have added oceanic and hydrological components.

The global scenarios were prepared by the model HadAM3 of the Hadley Center, using A2 and B2 emission scenarios. The boundary values as well as the sea surface temperature and the sea ice conditions from the global run during a 1961–1990 control and during the interval 2071–2100 were used to force the regional models, which were integrated over 30 years. Additionally, in the 2071–2100 runs, the radiative conditions in the regional models were changed according to the emission scenario.

So far, the process of comparing the responses of the various models is not yet completed. First results indicate that during winter the regional models deviate little from each other. The simulated expected changes due to global warming coincide across most models in terms of strong winterly windiness and heavy summer rainfall events (Beniston *et al.*, 2005).

The added value produced by the regional models is expected to consist in a better simulation of the spatially and temporarily smaller scales. In fact, a better description of

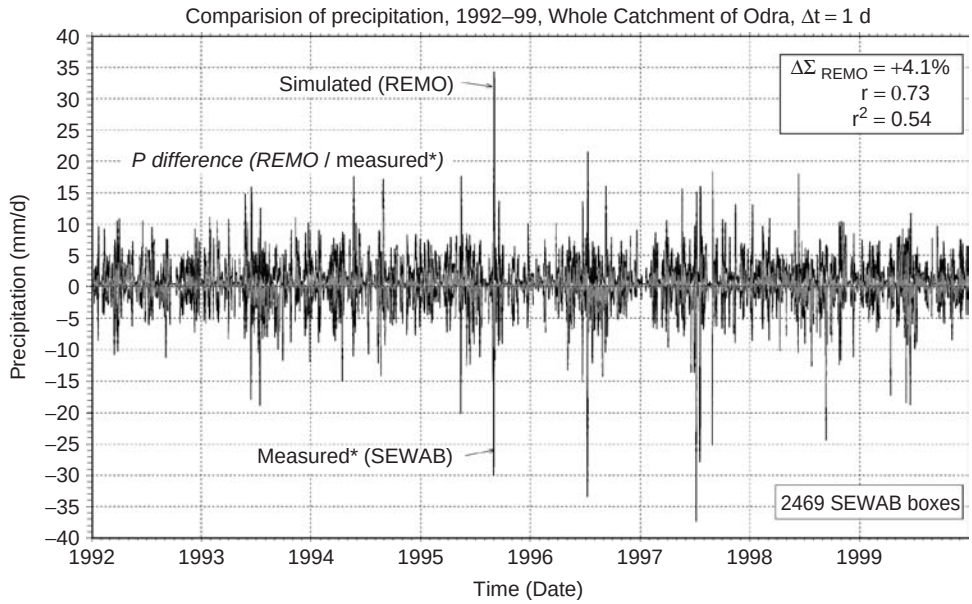


Figure 12 Simulated (upward; black) and analyzed (downward; black) daily precipitation in the catchment of the Odra River in 1997. The correlation is 0.73, the amount of variance described 54%. The grey curve shows the difference. The bias is 4%, which is within the limit of uncertainty (By courtesy of Hillmar Messal)

the details of frequency distributions is obtained (Beniston *et al.*, 2005).

As an example, the precipitation during summer time has been examined, (Christensen and Christensen, 2003) as well as wind conditions over the North Sea (Woth, personal communication). In both cases, a similar result is obtained, namely, the mean conditions are weakened – that is, the total amount of precipitation was found to be decreased, but the intensity of rare events was found to be increased by up to 40% (Figure 13). Similarly, the mean wind speed over the North Sea is envisaged to become slightly weaker on average, while strong westerly winds may increase by a few percent (not shown).

The PRUDENCE experience seems to indicate that the differences resulting from the use of the same global climate change scenario but different regional models are moderate. On the other hand, experiences at the Rossby–Center (e.g. Bergström *et al.*, 2001) indicate that the use of different global climate change scenarios and the same regional model induces much larger uncertainties. Figure 14 shows simulated regional annual precipitation changes in Europe derived by postprocessing (downscaling) two global climate change scenarios (prepared by the Hadley Center and the Max–Planck–Institute of Meteorology) with the Rossby–Center regional model. While broad features are similar, like more rainfall in the northern part and less in the southern, regional details contradict each other, for example, for the terrain of Poland.

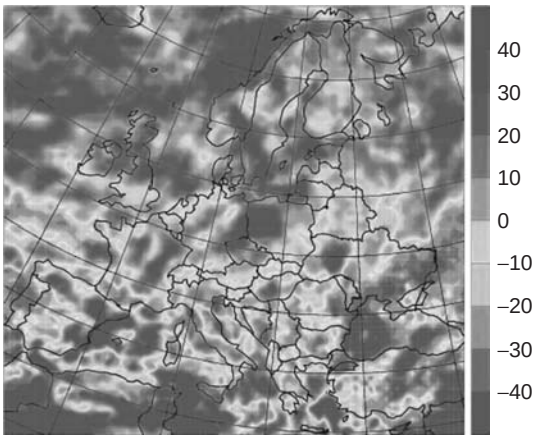


Figure 13 Change in precipitation intensity or the rare summer events as envisaged by a regional climate model for the end of the twenty-first century. The quantity shown is the change in five-day mean exceeding the 99th percentile (Christensen and Christensen, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CONCLUSIONS

Climate modeling is a standard exercise, which has matured in the past years after its introduction into the 1960s by pioneers like Manabe and Bryan (1969). Climate modeling is commonly understood as the space-time detailed modeling of at least the atmosphere, the ocean, and the sea ice. In such "quasi-realistic" models, the considered components are described in as much detail as is consistent with the anticipated application of the model (in particular the length of the integration time), and is feasible, given the computational platform. The atmospheric components describes baroclinic instability and the associated macroscale vortices (extratropical storms), while the dynamically relevant role of eddies in the ocean is parameterized by a certain type of diffusion (one could say, a climate model's ocean is not filled with water but with mustard).

Nowadays, such models are extended to contain more components of the earth system, in particular, surface hydrology, pathways and cycles of matter, vegetation and ice shelves and sheets.

Global models are meant to simulate phenomena of several grid length sizes; phenomena on scales of a few

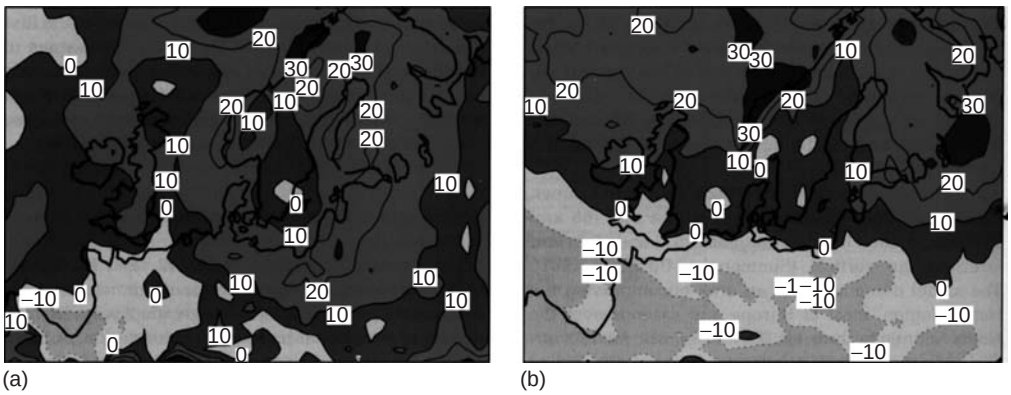


Figure 14 Simulated changes (%) in annual precipitation in Europe in two similar greenhouse gas emission scenario runs with different global climate models (Reproduced from Bergström *et al.*, 2001 by permission of Inter-Research). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

100 km or less are usually not simulated reliably. Thus, global model output is of limited utility for areas such as the North Sea, Colorado, or Taiwan. A thumb of rule gives 10^7 km^2 as threshold for skillful presentations for global models at this time. (This length scale is certainly a moving target. With increasing computer power, global models will be run with higher spatial resolution, and the skillful spatial scales will further decrease.) If smaller scaled descriptions and scenarios are needed, then one has to resort to a downscaling method, in particular, to the regional climate modeling. Such regional models are also readily available nowadays, and are presently extended – as their global siblings – to take into account more and more other components of the earth system.

For scientists not working with such climate models, the following items may be useful to remember:

- Global climate modeling allows the representation of global, continental, and subcontinental scales. Global models are not designed for, and thus not well suited for the regional and local scale.
- Global climate is varying because of both internal dynamics as well as external forcing.
- Scenarios of future climate change hinge on the validity of economic scenarios.
- Simulation of regional climate is a downscaling problem and not a boundary value problem.

Acknowledgments

I am thankful to Nikolaus Groll, Brigit Hüncke, Viacheslav Kharin, Hillmar Messal, Sebastian Wagner, and Katja Woth for their helpful comments. Many thanks to Beate Gardeike, who adapted many diagrams for this article. Two reviewers provided useful advice.

FURTHER READING

von Storch H. (2001) Models. In *Models in Environmental Research*, von Storch H. and Flöser G. (Eds.), Springer Verlag: pp. 17–33.

REFERENCES

Bengtsson L. (1997) A numerical simulation of anthropogenic climate change. *Ambio*, **26**(1), 58–65.

Beniston M., Stephenson D.B., Christensen O.B., Ferro C.A.T., Frei C., Goyette S., Halsnaes K., Holt T., Jylhä K., Koffi B., *et al.* (2005) Future extreme events in European climate: an exploration of regional climate model projections. *Climatic Change*, in review.

Bergström S., Carlsson B., Gardelin M., Lindström G., Petterson A. and Rummukainen M. (2001) Climate change impacts on runoff in Sweden – assessment by global climate models, dynamical downscaling and hydrological modeling. *Climate Research*, **16**, 101–112.

Bray D. and von Storch H. (1999) Climate Science. An empirical example of postnormal science. *Bulletin of the American Meteorological Society*, **80**, 439–456.

Busuioc A. and von Storch H. (2003) Conditional stochastic model for generating daily precipitation time series. *Climate Research*, **24**, 181–195.

Caya D. and Biner S. (2004) Internal variability of RCM simulations over an annual cycle. *Climate Dynamics*, **22**, 33–46.

Christensen J.H., Carter T. and Giorgi F. (2002) PRUDENCE employs new methods to assess European climate change. *EOS*, **83**, 147.

Christensen J.H. and Christensen O.B. (2003) Severe summertime flooding in Europe. *Nature*, **421**, 805–806.

Davies H.C. (1976) A lateral boundary formulation for multi-level prediction models. *Quarterly Journal of the Royal Meteorological Society*, **102**, 405–418.

Denis B., Laprise R., Caya D. and Cote J. (2002) Downscaling ability of one-way nested regional climate models: the big brother experiment. *Climate Dynamics*, **18**, 627–646.

Feser F., Weisse R. and von Storch H. (2001) Multidecadal atmospheric modeling for Europe yields multi-purpose data. *EOS*, **82**, 305–310.

Ganopolski A., Rahmstorf S., Petoukhov V. and Claussen M. (1997) Simulation of modern and glacial climates with a coupled global climate model. *Nature*, **391**, 351–356.

Gibson J.K., Kallberg P., Uppala S., Nomura A., Serrano E. and Hernandez A. (1997) *ERA Description. ECMWF Reanalysis Project Report 1: Project Organisation*, Technical Report, European Centre of Medium Range Weather Forecast, Reading.

Giorgi F., Hewitson B., Christensen J., Hulme M., von Storch H., Whetton P., Jones R., Mearns L. and Fu C. (2001b) Regional climate information – evaluation and projections. In *Climate Change 2001. The Scientific Basis*, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press, pp. 583–638.

Giorgi F. and Mearns L.O. (1991) Approaches to the simulation of regional climate change: a review. *Reviews of Geophysics*, **29**, 191–216.

Giorgi F., Whetton P.H., Jones R.G., Christensen J.H., Mearns L.O., Hewitson B., von Storch H., Fransico R. and Jack C. (2001a) Emerging patterns of simulated regional climatic changes for the 21st century due to anthropogenic forcings. *Geophysical Research Letters*, **28**(17), 3317–3320.

González-Rouco J.F., Zorita E., Cubasch U., von Storch H., Fischer-Bruns I., Valero F., Montavez J.P., Schlese U. and Legutke S. (2003) Simulating the climate since 1000 A. D. with the AOGCM ECHO-G. *Proceedings of the ISCS 2003 Symposium, 'Solar Variability as an Input to the Earth's Environment'*, Tatranská Lomnica, ESA SP-535, p. 329–338, 23–28 June 2003.

Hasselmann K. (1976) Stochastic climate models. Part I. theory. *Tellus*, **28**, 473–485.

Hasselmann K. (1990) How well can we predict the climate crisis? In *Environmental Scarcity – the International Dimension*, Siebert H. (Ed.), JCB Mohr: Tübingen, pp. 165–183.

Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.) (2001) *Climate Change 2001. The Scientific Basis*, Cambridge University Press.

Houghton J.T., Meira Filho L.G., Callander B.A., Harris N., Kattenberg A. and Maskell K. (Eds.) (1996) *Climate Change 1995. The Science of Climate Change*, Cambridge University Press: ISBN 0 521 56436-0, p. 572.

Jacob D., Podzun R. and Claussen M. (1995) REMO – A Model for Climate Research and Weather Prediction. *International Workshop on Limited-Area and Variable Resolution Models*, Beijing, China, October 23–27, 1995, 273–278.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G., Woollen J. *et al.* (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77**(3), 437–471.

Kharin V.V., Zwiers F.W. and Zhang X. (2004) Intercomparison of near surface temperature and precipitation extremes in AMIP-2 simulations. *Journal of Climate*, (in review).

Lionello P., de Zolt S., Luterbacher J. and Zorita E. (2004) Is the winter European climate of the last 500 years conditioned by the variability of the solar radiation and volcanism? *Climate Dynamics*, in review.

Manabe S. (1997) Early development in the study of greenhouse warming: the emergence of climate models. *Ambio*, **26**(1), 47–51.

Manabe S. and Bryan K. (1969) Climate calculations with a combined ocean-atmosphere model. *Journal of the Atmospheric Sciences*, **26**, 786–789.

McGuffie K. and Henderson-Sellers A. (1997) *A Climate Modeling Primer*, *Second Edition*, John Wiley & Sons: Chichester, p. 253, ISBN 0–471-95558-2.

Meinke I., von Storch H. and Feser F. (2004) A validation of the cloud parameterization in the regional model SN-REMO. *Journal of Geophysical Research*, **109**(D13205), 11, doi:10.1029/2004JD004520.

Miguez-Macho G., Stenchikov G.L. and Robock A. (2004) Spectral nudging to eliminate the effects of domain position and geometry in regional climate model simulations. *Journal of Geophysical Research*, **109**(D13), D13104, 10.1029/2003JD004495.

Müller P. and von Storch H. (2004) *Computer Modeling in Atmospheric and Oceanic Sciences – on the Building of Knowledge*, Springer Verlag: Heidelberg, p. 304, ISN 1437-028X.

Räisänen J., Hansson U., Ullerstig A., Döscher R., Graham L.P., Jones C., Meier H.E.M., Samuelsson P. and Willén U. (2004) European climate in the late 21st century: regional simulations with two driving global models and two forcing scenarios. *Climate Dynamics*, **22**, 13–31.

Rinke A. and Dethloff K. (2000) On the sensitivity of a regional Arctic climate model to initial and boundary conditions. *Climate Research*, **14**, 101–113.

Soares C.G., Weisse R., Carretero J.C. and Alvarez E. (2002) A 40 years hindcast of wind, sea level and waves in European Waters. *Proceedings of OMAE 2002: 21st International Conference on Offshore Mechanics and Arctic Engineering 23–28 June 2002 in Oslo*, Norway, OMAE2002-28604.

Sotillo M.G. (2003) *High Resolution Multi-Decadal Atmospheric Reanalysis in the Mediterranean Basin*, PhD dissertation, UCM, Madrid, p. 216.

Trenberth K. (Ed.) (1993) *Climate System Modeling*, Cambridge University Press, p. 788.

von Storch H. (1999) The global and regional climate system. In *Anthropogenic Climate Change*, von Storch H. and Flöser G. (Eds.), Springer Verlag: ISBN 3-540-65033-4, pp. 3–36.

von Storch H. and Flöser G. (Eds.) (2001) Models in environmental research. *Proceedings of the Second GKSS School on Environmental Research*, Springer Verlag: ISBN 3-540-67862, p. 254.

von Storch H., Güss und S. and Heimann M. (1999) *Das Klimasystem und seine Modellierung. Eine Einführung*, Springer Verlag: ISBN 3-540-65830-0, p. 255.

von Storch H., Langenberg H. and Feser F. (2000) A spectral nudging technique for dynamical downscaling purposes. *Monthly Weather Review*, **128**, 3664–3673.

von Storch H., von Storch J.-S. and Müller P. (2001) Noise in the climate system – ubiquitous, constitutive and concealing. In *Mathematics Unlimited – 2001 and Beyond. Part II*, Engquist B. and Schmid W. (Eds.), Springer Verlag, pp. 1179–1194.

Wagner S., Legutke S. and Zorita E. (2005) European winter temperature variability in a long coupled model simulation: the contribution of ocean dynamics. *Climate Dynamics*, (in press).

Wang Y., Leung L.R., McGregor J.L., Lee D.-L., Wang W.-C., Ding Y. and Kimura F. (2004) Regional climate modeling: progress, challenges and prospects. *Journal of the Meteorological Society of Japan*, **82**, 1599–1628.

Washington W. (1999) Three dimensional numerical simulation of climate: The fundamentals. In *Anthropogenic Climate Change*, von Storch H. and Flöser G. (Eds.), Springer Verlag: ISBN 3-540-65033-4, pp. 37–60.

Washington W.M. and Parkinson C.L. (1986) *An Introduction to Three-Dimensional Climate Modeling*, University Science Books, p. 422.

Weisse R. and Feser F. (2003) Evaluation of a method to reduce uncertainty in wind hindcasts performed with regional atmosphere models. *Coastal Engineering*, **48**, 211–255.

Weisse R., Heyen H. and von Storch H. (2000) Sensitivity of a regional atmospheric model to a sea state dependent roughness and the need of ensemble calculations. *Monthly Weather Review*, **128**, 3631–3642.

Wilby R.L. and Wigley T.M.L. (1997) Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, **21**, 530–548.

33: Human Impacts on Weather and Climate

STEPHEN H SCHNEIDER[1] AND MICHAEL D MASTRANDREA[2]

[1]Department of Biological Sciences, Stanford University, Stanford, CA, US
[2]Interdisciplinary Program in Environment and Resources, Stanford University, Stanford, CA, US

Climate change is a worldwide environmental, social, and economic challenge. It includes issues of air pollution, land use, toxic waste, transportation, industry, energy, government policies, development strategies, and individual freedoms and responsibilities. Human use of the atmosphere as an unpriced dumping space has led to the buildup of gases and particles such as carbon dioxide and methane that can alter the energy exchange between the Earth's surface and space. There is a great deal of confusion among policy makers and the public about just how much is known about our climate, and what makes it change. However, although many aspects of the science of climate change are laced with major uncertainties, many others are actually scientifically well established. Human modification of the climate system has far-reaching implications for human welfare and the health of our planet. Are our actions causing the climate to change in ways or at rates that will threaten natural systems or make human adaptation difficult? This chapter outlines the current state of scientific knowledge regarding human impacts on the climate system.

INTRODUCTION

Throughout human history, climate has both promoted and constrained human activity. In fact, humans only very recently have been able to substantially reduce their dependence on climate variability through advances in technology and organization, such as high yield agriculture, and food distribution and storage systems which have virtually eliminated famine in most countries with developed or transitioning economies. On the other hand, human action can also affect climate. Human use of the atmosphere as an unpriced dumping space has led to the buildup of gases and particles that can alter the energy exchange between the Earth's surface and space. Carbon dioxide (CO_2), methane (CH_4), and water vapor are the principal heat-trapping greenhouse gases. Carbon, the main element in the top two human-enhanced greenhouse gases (CO_2 and CH_4), is the underpinning of most fuels used in transportation and energy production. Carbon also makes up about half of the dry weight of most vegetation. Thus, the carbon that cycles through species, air, water, and soils is both an essential nutrient and a potential problem. Human modification of the carbon cycle has far-reaching implications for human welfare and the health of the biosphere. Are our actions causing the climate to change in ways or at rates that will threaten natural systems or make human adaptation difficult?

Climate change is a worldwide environmental, social, and economic challenge. It includes issues of air pollution, land use, toxic waste, transportation, industry, energy, government policies, development strategies, and individual freedoms and responsibilities. There is a great deal of confusion among policy makers and the public about just how much is known about our climate, and what makes it change. However, although many aspects of the science of climate change are laced with major uncertainties, many others are actually scientifically well established. This chapter outlines the current state of scientific knowledge regarding human impacts on the climate system.

IS THE CLIMATE CHANGING?

The Global Temperature Record

Modern temperature records, derived from thermometers sufficiently accurate and geographically dispersed to permit computation of a global average temperature, date back to the mid-nineteenth century. Extracting a global average from the data is complicated by many factors ranging

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

from the growth of cities, with their "heat island" warming of formerly rural temperature measuring stations, to such mundane effects as changes in the types of buckets used, to sample seawater temperature from ships (Harvey, 2000, Chapter 5). Early data suffer from a dearth of measurements and a bias toward the more developed regions of the planet. But climatologists understand how to account for these complications, and essentially all agree that Earth's average temperature increased by approximately 0.6 °C since the mid-nineteenth century (IPCC, 2001a). Throughout this chapter, many research findings we refer to are taken from Intergovernmental Panel on Climate Change (IPCC) Assessment Reports, which present the worldwide consensus on climate change science every five years. Figure 1 shows the global temperature record as a plot of the yearly deviations from the 1961–1990 average temperature.

A glance at Figure 1 shows that Earth's temperature is highly variable, with year-to-year changes often masking the overall rise of approximately 0.6 °C. Nevertheless, the long-term upward trend is obvious. Especially noticeable is the rapid rise at the end of the twentieth century. Indeed, all but three of the ten warmest years on record occurred in the 1990s, with 1998 marking the all-time record high through 2000. There is good reason to believe that the 1990s would have been even hotter had the eruption of Mt. Pinatubo in the Philippines not put enough dust into the atmosphere to cause global cooling of a few tenths of a degree for several years. Looking beyond the top ten years, Figure 1 shows that the 20 warmest years include the entire decade of the 1990s and all but three years from the 1980s as well. Clearly the recent past has seen substantial surface warming.

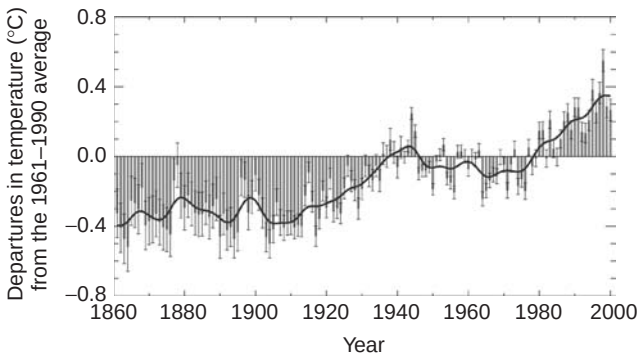


Figure 1 Variation in Earth's average global temperature from 1860 to 1999. Data are taken from global networks of thermometers, corrected for a variety of effects, and combined to produce a global average for each year. Wider, solid bars represent temperature deviations for each year, relative to the 1961–1990 average temperature, and narrow gray bars show uncertainties in the yearly temperatures. Black curve is a best fit to the data (Adapted from Wolfson and Schneider, 2002)

A Natural Climate Variation?

Could the warming shown in Figure 1, especially of the past few decades, be a natural occurrence? Might Earth's climate undergo natural fluctuations that could result in the temperature record of Figure 1? Increasingly, we are finding that the answer to that question is "no". We would be in a better position to determine whether the temperature rise of the past century is natural if we could extend the record further back in time. Unfortunately, direct temperature measurements of sufficient accuracy or geographic coverage simply do not exist before the mid-1800s. But by carefully considering other quantities that do depend on temperature, climatologists can reconstruct approximate temperature records that stretch back hundreds, thousands, and even millions of years.

Figure 2 shows the results of a remarkable study that attempts to push the Northern Hemisphere temperature record back a full thousand years (Mann *et al.*, 1999). In this work, climatologist Michael Mann and colleagues performed a complex statistical analysis involving 112 separate indicators related to temperature. These included such diverse factors as tree rings, the extent of mountain glaciers, changes in coral reefs, and many others. The resulting temperature record of Figure 2 is a "reconstruction" of what one might expect had thermometer-based measurements been available. Although there is considerable uncertainty in the millennial temperature reconstruction, as shown by the error band in Figure 2, the overall trend is most consistent with a gradual temperature decrease over the first 900 years, followed by a sharp upturn in the twentieth century. That upturn is a compressed representation of the

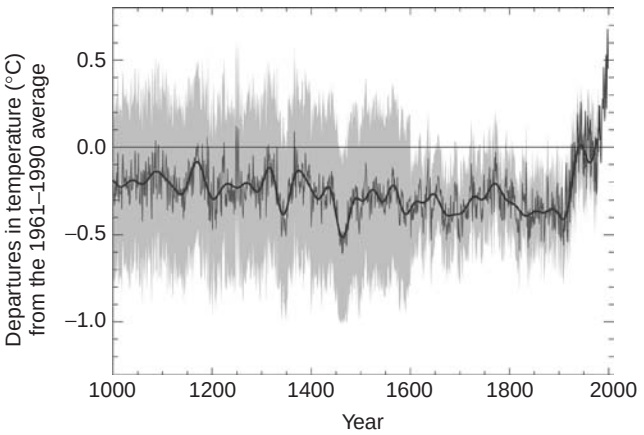


Figure 2 Reconstruction of the 1000-year temperature record for the Northern Hemisphere. Black curve is a best fit to the millennial temperature record; gray is the 95% confidence interval, meaning that there is a 95% chance that the actual temperature falls within this band. Date from the mid-nineteenth century on are from the thermometer-based temperature record of Figure 1 (Adapted from Wolfson and Schneider, 2002)

thermometer-based temperature record shown in Figure 1. Among other things, Figure 2 suggests that the 1990s was the warmest decade not only of the twentieth century but also of the entire millennium. Taken in the context of Figure 2, the temperature rise of last century clearly is an unusual occurrence.

ARE WE CHANGING THE CLIMATE?

Is the change documented in Figure 2 caused by human activities or natural variability? Or is it a combination of both? Mann *et al.* approached that question by correlating their temperature reconstruction with several factors known to influence climate, including solar activity, volcanism, and humankind's release of heat-trapping gases (greenhouse gases). They found that solar variability and volcanism (natural influences) were the dominant influences in the first 900 years of the millennium, but that much of the twentieth century variation could be attributed to human activity. Given the indirect, statistical nature of the study, this result can hardly be taken as conclusive evidence that humans are to blame for twentieth century global warming. But the Mann *et al.* result does provide independent corroboration of computer climate models that also indicate a human influence on climate.

''Fingerprint'' Analysis

Methods used to attribute observed warming trends to human-induced climate change are often called "fingerprint analysis". Although each fingerprint may be a circumstantial piece of evidence for human-induced climate changes taken together, several such lines of evidence have led the IPCC to conclude that there is a discernible human fingerprint on climate change (IPCC, 2001a). These different types of fingerprints include the fact that the Earth's stratosphere cooled while the surface warmed – a fingerprint of changes due to increased atmospheric CO_2 rather than, for example, a fingerprint of an increase in the heat output of the sun which should have warmed all altitudes in the atmosphere (Schneider, 2003).

Figure 3 shows three different attempts, using the same basic climate model (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**), to reproduce the historical temperature record of Figure 1. In the model runs of Figure 3(a), only estimates of solar variability and volcanic activity – purely natural forcings, or climatic influences – were included in the model. The projected temperature variation, represented by a thick band indicating the degree of uncertainty in the model calculations, does not show an overall warming trend and clearly is a poor fit to the actual surface temperature record. The runs of Figure 3(b) include only anthropogenic, or human-induced,

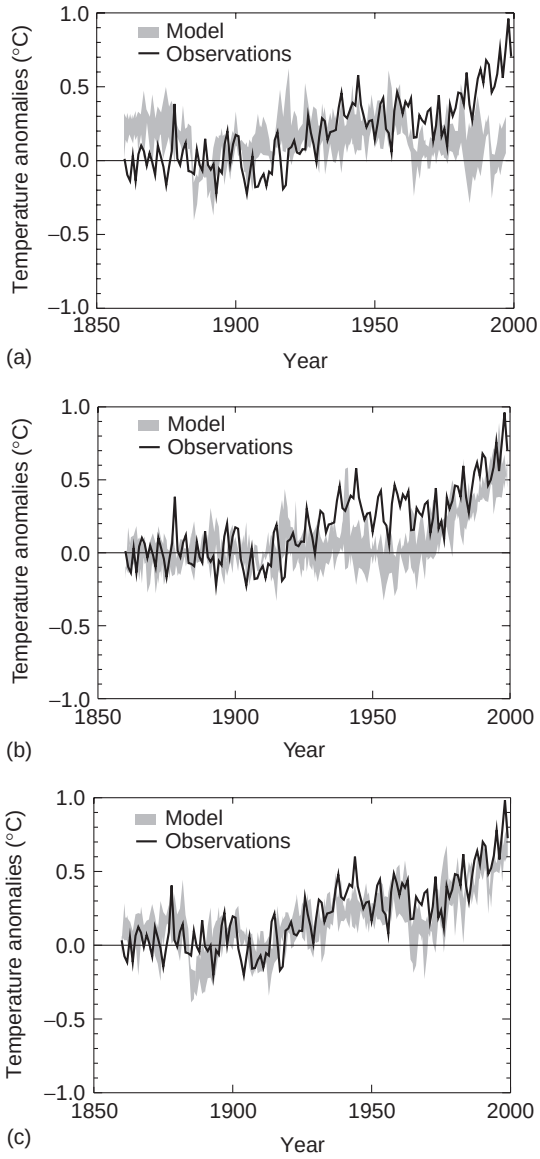


Figure 3 Attempts to model Earth's temperature from the 1860s using different model assumptions. In all three graphs, the solid curve is the observed surface temperature record of Figure 1. Gray bands represent model projections. In each graph, the bands encompass the results of four separate model runs. In (a), only natural forcings – volcanic activity and solar variability – are included. Clearly this simulation lacks the upward trend in the observed temperature record, suggesting that the temperature rise of the last century and a half is unlikely to have a purely natural explanation. Simulation (b), including only anthropogenic forcings, does much better, especially with the rapid temperature increase of the late twentieth century. Simulation (c), combining both natural and anthropogenic forcings, shows the best agreement with observations (Adapted from Wolfson and Schneider, 2002. Data from IPCC, 2001a)

forcings such as emissions of greenhouse gases and particulates. This clearly does a much better job, especially in the late twentieth century, but deviates significantly from the historical record around mid-century. Finally, Figure 3(c) shows the results from runs that include both natural and anthropogenic forcings. The fit is excellent, and it suggests that we can increase our confidence in this model's projections of future climate. Furthermore, the model runs of Figure 3 taken together strongly suggest that the temperature rise of the past few decades is unlikely to be explained without invoking anthropogenic greenhouse gases as a significant causal factor. Thus, the "experiments" of Figure 3 illustrate one way of attempting to pry an anthropogenic climate signal from the natural climatic noise. In other words, Figure 3 provides substantial circumstantial evidence of a discernible human influence on climate, and, with the other evidence mentioned, supports the IPCC conclusion that "most of the warming observed over the last 50 years is attributable to human activities" (IPCC, 2001a).

Keeping a Planet Warm

How can human activities affect the Earth's climate? What ultimately determines climate, and specifically, the Earth's temperature? That question is at the heart of climate science and of the issues surrounding human-induced climate change.

The gases that comprise the Earth's atmosphere are largely transparent to visible light. Therefore, much of the incident sunlight penetrates the atmosphere to reach the surface where it is absorbed, warming the surface, which then reemits the energy as infrared radiation. But the atmosphere is not so transparent to infrared radiation. Certain naturally occurring gases and particles – particularly clouds (*see* **Chapter 196, The Role of Water Vapor and Clouds in the Climate System, Volume 5**) – absorb infrared radiation and hinder its ability to escape from the atmosphere. That infrared energy that is trapped in the atmosphere is reemitted, both up to space and back down towards the surface – the latter primarily adding heat to the lower layers of the atmosphere. As a result, the Earth's surface warms further, emitting infrared radiation at a still greater rate, until the emitted radiation is in balance with the incident sunlight and the other forms of energy coming and going from the surface (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**). The difference between the upward and downward energy flows, in the steady state, is just the right amount to maintain energy balance between absorbed solar radiation, evaporation, thermal energy lost via rising plumes of heated air, and the net infrared radiation balance. So, the Earth is in nearly perfect energy balance but with a surface temperature significantly higher than it would be in the absence of greenhouse gases. This is what accounts for the 33 °C difference between the Earth's actual surface air temperature and that which it would be were there no such gases in the atmosphere.

Because the atmosphere functions, in a crude sense, like the heat-trapping glass of a greenhouse, this heating has earned the nickname greenhouse effect, and the gases responsible are greenhouse gases. (The greenhouse analogy is not such a good one; a greenhouse traps heat primarily by preventing the wholesale escape of heated air, with the blockage of infrared playing only a minor role). The 33 °C warming due to natural greenhouse gases is the natural greenhouse effect, and is a good thing because it makes our planet much more habitably warm than it would be otherwise. However, human activities have been and are continuing to enhance the greenhouse effect by adding additional greenhouse gases into the atmosphere. Such emissions add to the "blanket" of heat-trapping gases, further increasing the Earth's temperature. It is important to understand that the basic greenhouse phenomenon is well understood and solidly grounded in basic science. More controversial is the extent to which we have already caused climate change and by how much we will enhance future climate disturbance.

CO_2 and Past Climate

Human activities add to the atmospheric concentrations of a number of naturally-occurring greenhouse gases, and introduce other potent greenhouse gases which are not naturally occurring. The greenhouse gas that has been most affected by anthropogenic emissions is CO_2. In the last 140 years, atmospheric CO_2 concentrations (see Figure 4) have increased by 30% from 280 to 370 parts per million (ppm) (Neftel *et al.*, 1994; Keeling and Whorf, 2000). The reality of this CO_2 increase is unquestioned, and virtually all climatologists agree that the cause is human activity, predominantly the burning of fossil fuels, and to a lesser extent deforestation and other land use changes, along with industrial activities such as cement production.

Figures 1 and 4 taken together show concurrent increases in global temperature and CO_2 concentration, both occurring during an era of rapid industrialization. So, are anthropogenic CO_2 emissions a direct cause of recent warming? As the study summarized in Figure 2 suggested, it looks increasingly like the answer is "yes". But the connection between the past 140 years' warming and the coincident rise in CO_2 is not conclusive. For example, global temperature actually declined in the period after World War II, a time of rapid industrialization when CO_2 concentrations began an especially rapid increase. On the other hand, CO_2 increases do not induce immediate changes in temperature, so we should not expect to find that recent temperature and CO_2 are instantaneously correlated. Moreover, there are other factors that can influence climate fluctuations or trends, and all of these are confounded in the data shown in Figures 1 and 4. Separating the anthropogenic "signal"

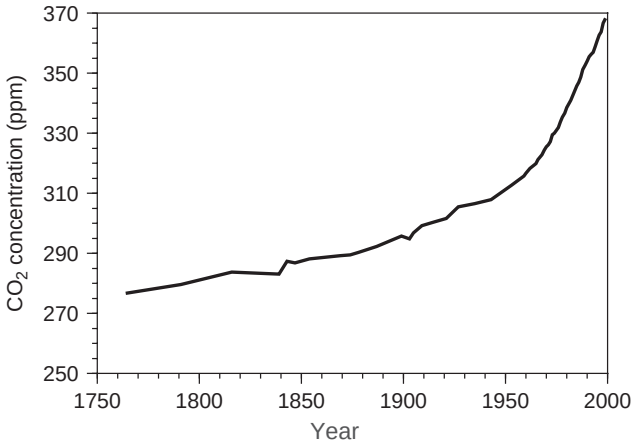


Figure 4 Atmospheric carbon dioxide has increased by more than 30% since preindustrial times (Adapted from Wolfson and Schneider, 2002. Data are from Neftel *et al.*, 1994; and Keeling and Whorf, 2000)

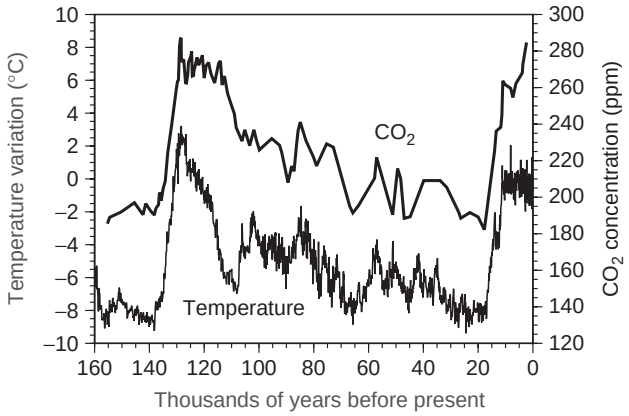


Figure 5 Atmospheric carbon dioxide (upper curve) and temperature variation (lower curve) over the past 160 000 years, from ice cores taken at Vostok, Antarctica. The record shows long stretches of low temperature (ice ages) separated by brief, warm interglacial periods. The correlation between CO_2 and temperature is quite obvious. Note also the small change, averaging perhaps 6 °C, between the present warm climate and the recent ice age. Data do not extend to the present, but stop well before the industrial era (Adapted from Wolfson and Schneider, 2002. CO_2 data are from Petit *et al.*, 2000; temperature data from Jouzel *et al.*, 1987, as reproduced in the Carbon Dioxide Information Analysis Center)

of climate change from the "noise" of natural fluctuations is a complex process.

A clearer relationship emerges when past climate is examined (*see* **Chapter 34, Climate Change – Past, Present and Future, Volume 1**). Ice cores bored from the Greenland and Antarctic ice sheets provide estimates of both temperature and atmospheric CO_2 going back hundreds of thousands of years (Petit *et al.*, 2000; Jouzel *et al.*, 1987). Variations in ice density associated with seasonal snowfall patterns provide a year-to-year calibration of the age of given points in the ice core, and analysis of air bubbles trapped in ancient ice gives an indication of CO_2 concentration. Temperature inference is accomplished by comparing a ratio of oxygen isotopes that is sensitive to temperature. Temperature differentially affects the concentration of the isotopes in evaporating water, and therefore their concentration in precipitation and in the ice itself. The result of such an ice core analysis, shown in Figure 5, gives dramatic evidence that temperature and carbon dioxide concentration are correlated over the long term.

It is not clear from the graph alone that the CO_2 variations in Figure 5 are the cause of the temperature changes. Sometimes a CO_2 increase precedes a warming, but sometimes not. In fact, climatologists suspect a feedback process whereby a slight increase in temperature, probably caused by subtle changes in the Earth's orbit, results in an increase in atmospheric CO_2 through a variety of mechanisms such as the release of CO_2 dissolved in the oceans (IPCC, 2001a; *see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**). The increased atmospheric CO_2, in turn, leads to greenhouse warming, amplifying the initial temperature increase. The result is a nearly simultaneous and substantial increase in both CO_2 and temperature. Eventually, orbital changes trigger a modest temperature decrease, and again feedback mechanisms amplify the decrease, driving down both CO_2 and temperature. Some paleoclimatologists also posit biotic feedback mechanisms, such as an initial cooling, which causes a drying of the continents, producing more windblown dust (IPCC, 2001a). This dust contains minerals needed by phytoplankton in the oceans. As dust settles on the ocean surface, it fertilizes these tiny oceanic organisms. The phytoplankton, in turn, increase their productivity by drawing down atmospheric CO_2, thus making the move toward an ice age even more rapid and deep. But despite such complexities, there is still much regularity in the record. The pattern of varying temperature and carbon dioxide concentration shown in Figure 4 is believed to repeat on a timescale of roughly every 100 000 years over most of the past million years, at least in part as a result of periodic changes in the Earth's orbit and the inclination of its polar axis.

Note that Figure 5 shows brief periods of warmth punctuated by much longer, cooler ice ages, which are characterized by dramatically different climatic conditions, with ice sheets 2 km thick covering what is now Canada, the northeastern United States, and northwestern Europe, and engulfing high mountain plateaus all around the world. Today we enjoy the warmth of an interglacial period, but not long ago, geologically speaking, conditions were very different. Figure 5 shows that the difference in temperature between an interglacial and an ice age is on the order

of 6 °C. Climate models driven by standard assumptions about population, land use, and energy consumption project a warming over the next century of 1.5 °C to 6 °C. The difference between the higher and lower ends of this range has substantial implications for sea level rise, extreme weather, redistribution of species ranges, and other impacts. Policymakers and the general public often ask how a few degrees can matter all that much. Figure 5 provides one striking answer: Downward changes on the same order as the largest projected warming are enough to make the difference between our current climate and an ice age. A few degrees, sustained in time and taken over the entire globe, can make a significant difference – especially when the change occurs many times faster than the comparable changes of the last ice age.

A second important point follows from comparing Figures 4 and 5. Note in Figure 5 that the maximum CO_2 concentration in the ice core record of the past 160 000 years is less than 300 ppm. This does not include the very recent past, but only the preindustrial period. The present-day concentration of 370 ppm in Figure 4 is far above anything the Earth has seen, probably for millions of years. Figure 6 shows the effect of adding the recent rise in CO_2 to the ice core data. Clearly, the anthropogenic increase in CO_2 concentration is unprecedented in both its size and its rapidity. We have made truly dramatic changes in the Earth's atmosphere over the past century or so, and we can almost certainly expect significant climate change to result. To begin to predict the extent of this change, we must examine all of the changes human activities are making to the climate system.

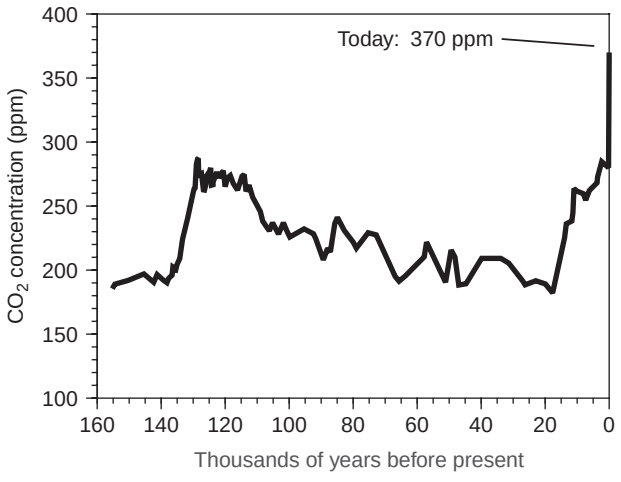


Figure 6 The CO_2 record of Figure 5, with data up to 1999 included. The CO_2 rise of Figure 3 is shown here as a dramatic jump to levels not seen on Earth for hundreds of thousands (and probably millions) of years (Adapted from Wolfson and Schneider, 2002)

HOW ARE WE CHANGING THE CLIMATE?

Although CO_2 is the most significant anthropogenic influence on the climate system, accounting for some 60% of the enhanced infrared blockage, a host of other influences also result from human activities: emissions of other greenhouse gases; feedback effects, whereby human-induced greenhouse warming may influence other processes that either exacerbate or dampen the warming; and emissions of particulate pollution from cars and fossil-fueled power plants, that can result in regional cooling that may mask or reduce the effects of greenhouse warming. We quantify these influences below.

Greenhouse Gases and Radiative Forcing

Carbon Dioxide

CO_2 is the most important of the anthropogenic greenhouse gases in terms of its direct effect on climate, but other gases play a significant role too. On a molecule-to-molecule basis, most other greenhouse gases (except water vapor) are far more potent absorbers of infrared radiation than is CO_2, but they are released in much lesser quantities, so their overall effect on climate is smaller. Climatologists characterize the effect of a given atmospheric constituent by its radiative forcing, the rate at which it alters absorbed solar or outgoing infrared energy. Currently, anthropogenic CO_2 produces a radiative forcing estimated at about 1.5 watts for every square meter of Earth's surface (all forcings cited in this section are from IPCC, 2001a). Relative to the 235 W m^{-2} of solar energy that is absorbed by the Earth and its atmosphere, the CO_2 forcing is a modest perturbation of the overall energy balance; but regardless of its size, any forcing affecting the energy balance will shift climate over time.

Methane

The second most prevalent anthropogenic greenhouse gas is methane (CH_4), produced naturally and anthropogenically, when organic matter decays anaerobically. Such anaerobic decay occurs in swamps, landfills, rice paddies, land submerged by hydroelectric dams, the guts of termites, and the stomachs of ruminants such as cattle. Methane is also released by oil and gas drilling, coal mining, volcanic eruptions, and the warming of methane-containing compounds on the ocean floor. One methane molecule is roughly 30 times more effective at blocking infrared than is one CO_2 molecule, although this comparison varies with the timescale involved, and the presence of other pollutants. Whereas CO_2 concentration increases tend to persist in the atmosphere for centuries or longer, the more chemically active methane typically disappears in decades, making its warming potential relative to that of CO_2 lower on longer timescales. Currently, methane accounts for about

$0.5\,W\,m^{-2}$ of anthropogenic radiative forcing, about one-third that of CO_2.

Nitrous Oxide and Halocarbons

Other anthropogenic greenhouse gases include nitrous oxide, produced from agricultural fertilizer and industrial processes, and the halocarbons used in refrigeration. (A particular class of halocarbons – the chlorofluorocarbons (CFCs) – is also the leading cause of stratospheric ozone depletion. Newer halocarbons do not cause severe ozone depletion but are still potent greenhouse gases.) Together, nitrous oxide and halocarbons account for roughly another $0.5\,W\,m^{-2}$ of radiative forcing. A number of other trace gases contribute roughly $0.05\,W\,m^{-2}$ of additional forcing. All the gases mentioned so far are well mixed, meaning that they last long enough to be distributed in roughly uniform concentrations throughout the lowest 10 km of the atmosphere.

Ozone

Another greenhouse gas is ozone (O_3), familiar because of its depletion by anthropogenic CFCs. Ozone occurring naturally in the stratosphere (some 10–50 km above the surface) absorbs incoming ultraviolet radiation and protects life from UV-induced cancer and genetic mutations, hence the concern about ozone depletion and in particular the polar "ozone holes". Unfortunately, ozone depletion and global warming have become confused in the public mind, even among political leaders and some environmental policymakers. But the two are very distinct problems. The ozone depletion problem is not the same as the global warming problem. Ozone depletion eventually will come under control because of the 1987 Montreal Protocol and its subsequent strengthening accords, international agreements that ban the production of the chlorinated fluorocarbons that destroy stratospheric ozone.

Because ozone is a greenhouse gas, there are some direct links between greenhouse warming and anthropogenic changes in atmospheric ozone. Ozone in the lower atmosphere – the troposphere – is a potent component of photochemical smog, resulting largely from motor vehicle emissions. Tropospheric ozone contributes roughly another $0.35\,W\,m^{-2}$ of radiative forcing, although, unlike the well-mixed gases, tropospheric ozone tends to be localized where industrialized society is concentrated. In the stratosphere, the situation is reversed. Here the anthropogenic effect has been ozone depletion, resulting in a negative forcing of approximately $-0.15\,W\,m^{-2}$. Thus stratospheric ozone depletion, on its own, would cause a slight global cooling. Taken in the context of the more substantial positive forcings of other gases, though, the effect of stratospheric ozone depletion is a slight reduction of the potential for global warming, an effect that will diminish as the ozone layer gradually recovers under the Montreal Protocol's ban on chlorofluorocarbons. The net effect of all anthropogenic ozone (both tropospheric and stratospheric) probably amounts to a slight positive forcing. The net forcing to date from all anthropogenic gases probably is about $3\,W\,m^{-2}$, and is expected to become much larger if business-as-usual development scenarios are followed in the twenty-first century.

Aerosols

Fuel combustion, and to a lesser extent agricultural and industrial processes, produce not only gases but also particulate matter. Coal-fired power plants burning high-sulfur coal, in particular, emit gases that become sulfate aerosols that reflect incoming solar radiation and thus result in a cooling trend. Natural aerosols from volcanic eruptions and the evaporation of seawater also produce a cooling effect. However, diesel engines and some biomass burning produce black aerosols such as soot, which can warm the climate. Recent controversial estimates suggest that these could offset much of the cooling from sulfate aerosols, especially in polluted parts of the subtropics (Jacobson, 2001). The IPCC estimates the total radiative forcing resulting directly from all anthropogenic aerosols very roughly at about $-1\,W\,m^{-2}$. However, this figure is much less certain than the radiative forcings caused by the greenhouse gases. Furthermore, aerosol particles also exert an indirect effect, in that they act as "seeds" for the condensation of water droplets to form clouds (*see* **Chapter 196, The Role of Water Vapor and Clouds in the Climate System, Volume 5**). Thus, the presence of aerosols affects the size and number of cloud droplets. An increase in sunlight reflected by these aerosol-altered clouds may result in $-2\,W\,m^{-2}$ of potential radiative forcing. Similarly, soot particles mixed into clouds can make the droplets absorb more sunlight, producing some warming. Taken together, aerosols add an element of uncertainty into anthropogenic radiative forcing of about $1\,W\,m^{-2}$ and complicate attempts to discern an anthropogenic signal of climatic change from the noise of natural climatic fluctuations.

Solar Variability

Another important influence on the climate system not affected by human activities should be mentioned here, variation in the Sun's energy output. Variations caused by the 22-year solar activity cycle amount to only about 0.1% and are too small and occur too rapidly to have a significant climatic effect. However, long-term solar variations, either from variability at the Sun itself, or from changes in the Earth's orbit and inclination, have substantially affected Earth's climate over geologic time. Although accurate, satellite-based measurements of solar output are available for only a few decades, indirect evidence of solar activity allows us to estimate past

variations in solar energy output (Hoyt and Schatten, 1997). Such evidence suggests that solar forcing since preindustrial times amounts to about $0.3\,W\,m^{-2}$, which is enough to contribute somewhat to observed global climate change, but far below what is needed to account for the warming of recent decades. However, there is some speculation that magnetic disturbances from the Sun can influence the flux of energetic particles impinging on the Earth's atmosphere, which in turn affect stratospheric chemical processes and might thus indirectly alter the global energy balance. These speculations have led some to declare the warming of the past century to be wholly natural, but this notion is discounted by nearly all climatologists for two reasons: first, there is no demonstrated way in which solar energetic particles can have a large enough effect to account for the recent warming and, second, because it is unlikely that such solar magnetic events happened only in the past few decades and not any other time over the past 1000 years. Assessment groups such as the IPCC are convened to sort out such claims and to weigh their relative probabilities. That is why we report primarily the IPCC assessments rather than the claims of a few individual scientists.

Radiative Forcing: The Overall Effect

Figure 7 summarizes our current understanding of radiative forcings caused by greenhouse gases, aerosols, land use changes, solar variability, and other effects since the start of the industrial era. The negative forcings from some of these anthropogenic changes might appear sufficient to offset the warming caused by anthropogenic greenhouse gases. This implication is misleading, however, because the effects of aerosols are short-lived and geographically localized compared with the long-term, global effects of the well-mixed greenhouse gases. The most advanced climate models (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**, **Chapter 201, Land-Atmosphere Models for Water and Energy Cycle Studies, Volume 5**) are driven by a range of plausible assumptions for future emissions of all types and make it clear that the overall effect of human activity is almost certainly a net positive forcing.

Feedback Effects

Knowing the radiative forcing caused by changes in atmospheric constituents would be sufficient to project future climate if there were no additional climatic effects beyond the direct change in energy balance. But a change in climate caused by simple forcing can have significant effects on atmospheric, geological, oceanographic, biological, chemical, and even social processes. These effects, in turn, can further alter the climate. If that further alteration is in the same direction as its initial cause, then the effect is called a *positive feedback*. If the further alteration tends to counter the initial change, then it is a negative feedback. In reality, numerous feedback effects greatly complicate the full description of climate change. Here we list just a few to give a sense of their variety and complexity.

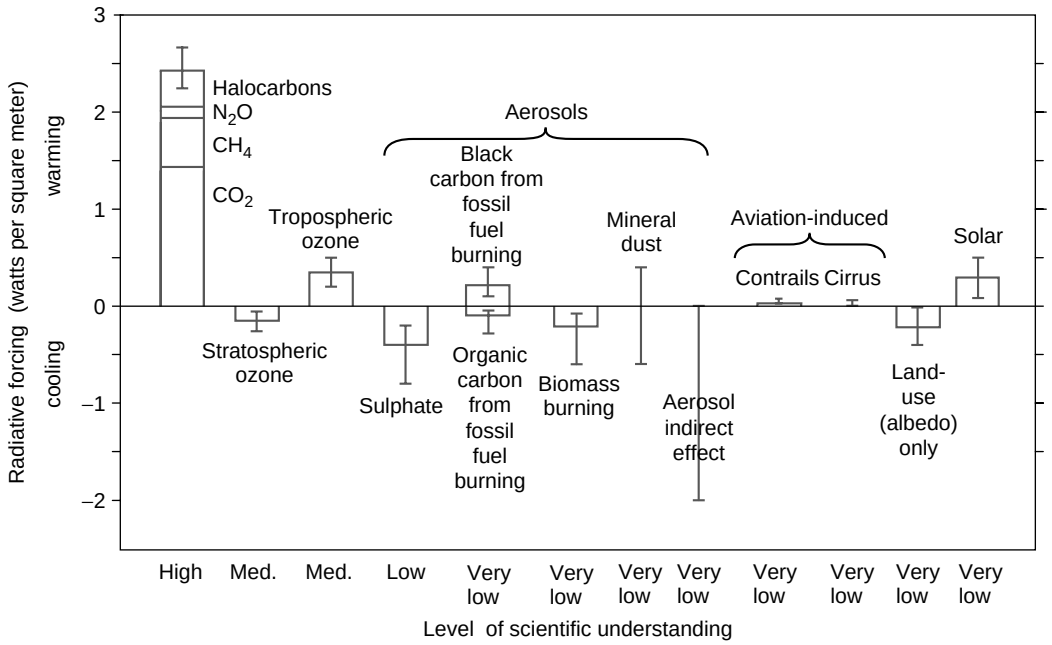


Figure 7 Radiative forcings caused by anthropogenic greenhouse gases, particulate emissions (aerosols), and other processes. Vertical bars indicate relative uncertainties, and the overall level of scientific understanding of and confidence in these processes is listed below the graph. (Adapted from Wolfson and Schneider, 2002. Data from IPCC, 2001a)

Ice-albedo Feedback

Albedo is a planet's reflectance of incident sunlight. The Earth's albedo is about 0.31, meaning that 31% of incident sunlight is reflected back to space. A decrease in that number would mean more sunlight absorbed, which would increase global temperature. One likely consequence of rising temperature is the melting of some ice and snow, that would eliminate a highly reflective surface and expose the darker land or water beneath the ice. The result is a decreased albedo, increased energy absorption, and additional heating. This is a positive feedback.

Hydrologic Feedbacks

Rising temperature also results in increased evaporation of water from the oceans (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**). That means more water vapor in the atmosphere. Because water vapor is itself a greenhouse gas, this effect results in still more warming and is thus a positive feedback. But increased water vapor in the atmosphere might mean more widespread cloudiness, which reflects sunlight and thus raises the albedo, resulting in less energy absorbed by the Earth-atmosphere system. The result is a negative feedback, tending to counter the initial warming. On the other hand, clouds also absorb outgoing infrared, resulting in a warming – a positive feedback. There are actually a number of processes associated with clouds, some of which produce warming, and some, cooling (*see* **Chapter 196, The Role of Water Vapor and Clouds in the Climate System, Volume 5**). These effects vary with the type of cloud, the location, and the season. Our limited understanding of cloud effects is one of the greatest sources of uncertainty in global climate sensitivity and thus in climate projections. However, the best estimates suggest that the overall effect of increased water vapor is a positive feedback that causes a temperature increase at least 50% higher than would occur in the absence of this feedback mechanism (Harvey, 2000, Chapter 9).

Biological Feedbacks

Some feedbacks are biological. For example, increased atmospheric CO_2 stimulates plant growth, and plants in turn remove CO_2 from the atmosphere. This is a negative feedback. On the other hand, warmer soil temperatures stimulate microbial action that releases CO_2 – a positive feedback effect. Drought and desertification resulting from climate change can alter the albedo of the land by replacing dark plant growth with lighter soil and sand. Greater reflection of sunlight results in cooling, so this is a negative feedback. But here, as so often with the climate system, the situation is even more complex. If sand is wet, as on a beach, then it is darker, and therefore absorbs more sunlight than dry sand. Yet dry sand is hotter. The resolution of this conundrum is that the wet sand is cooler because of the cooling effects of evaporation, but the Earth is warmed by the wet sand because the evaporated water condenses in clouds elsewhere and puts the heat back into the overall system. Thus, cooling or warming of the Earth-atmosphere system does not always imply cooling or warming of the Earth's surface at that location. There are even social feedbacks. For example, rising temperature causes more people to install air conditioners. The resulting increase in electrical consumption means more fossil fuel-generated atmospheric CO_2 – again giving a positive feedback.

Accounting for all significant feedback effects entails not only identifying important feedback mechanisms but also developing a quantitative understanding of how those mechanisms work. That understanding often includes research at the boundaries of disciplines such as atmospheric chemistry and oceanography, biology and geology, even economics and sociology.

With positive feedback, there is a danger of runaway warming, whereby a modest initial warming triggers a positive feedback that results in additional warming. This feedback could lead to extreme climate change. That is what has happened on Venus, where the thick, CO_2-rich atmosphere produced a runaway greenhouse effect that gives Venus its abnormally high surface temperature. Fortunately, we believe that the conceivable terrestrial feedbacks, at least under the Earth's current conditions, are incapable of such dramatic effects. But while that means we are not going to boil the oceans away, it does not preclude potentially disruptive climatic change.

HOW MUCH WILL THE CLIMATE CHANGE?

Climate Model Projections

Uncertainty in greenhouse gas emissions, the magnitude of their influence on the changing climate, and the host of other processes affected by increases in radiative forcing from greenhouse gas emissions demonstrate that projected future climate change is a complex task. The most sophisticated tools we have are global models of the climate system. Today's climate models (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**) provide geographic resolution down to the scale of a small country. Not only can they reproduce global temperature records, as shown in Figure 3, but the best model results reproduce, although not completely, the detailed geographic patterns of temperature, precipitation, and other climatic variables seen on a regional scale. These pattern-based comparisons of models and reality provide further confirmation of the models' essential validity.

No one model validation experiment alone is enough to give us high confidence in future climate projections. But considered together, results from the wide

range of published experiments probing the validity of climate models give considerable confidence that these models are treating the essential climate-determining processes with reasonable accuracy. Therefore, we can expect from them moderately realistic projections of future climate, given credible emission scenarios. However, we still expect variations in the projections of different models. And because future greenhouse gas emissions depend on human behavior, future projections will differ depending on what assumptions modelers make about the human response to global climate change. The uncertainties in projections of human behavior cause about as much spread in estimates of future warming as do uncertainties about the sensitivity of the climate system to radiative forcings.

Consequences from Global Climate Change

Taking all of these issues into account, the current consensus modeling estimate, as mentioned previously, is a global average temperature increase over the current century of anywhere between 1.5 and 6 °C (IPCC, 2001a). This increase may seem modest, but as we noted, it could imply quite serious impacts. What might be the consequences? The most sophisticated climate models speak to a wide variety of possible impacts from global climate change. Recall that a 6 °C temperature drop means the difference between the Earth's present climate and an ice age. Fortunately, it does not appear that a comparable rise will have consequences as devastating as 2-km-thick ice sheets over populated areas of the Northern Hemisphere. But that does not mean the consequences of a few degrees' global warming will not be substantial and disruptive. How will the temperature rise be distributed in time and in space? In fact, global climate change will vary substantially from one geographical region to another, and it will have different effects during night and day, winter and summer, and on land and sea.

Climate models provide rough consensus on many temperature-related projections. In general, projected temperature rises are greatest in the polar regions, and they affect the polar winter more dramatically than the summer. Similarly, nighttime temperatures are projected to rise more than daytime temperatures. Land temperatures are projected to rise more than oceans for the most part, influencing the patterns of monsoons and life-giving rains (and deadly floods) that they engender (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**). Other temperature-related consequences include increases in maximum-observed temperatures and more hot days, increases in minimum temperatures and fewer cold days, and longer growing seasons owing to earlier last frosts and later first frosts. All these trends have already been seen in the climate change of the past few decades, and all are projected to continue through the present century. Climatologists' assessed confidence in these projections ranges from "likely" (two-thirds to 90% probability) to "very likely" (90 to 99% probability). Table 1 summarizes these and other effects of global climate change, and gives the IPCC's quantitative estimates of the probability of each effect (IPCC, 2001b).

The broadest impacts of direct temperature effects on human society are likely to be in agriculture and water supplies. However, health effects, including the spread of lowland tropical diseases, vertically upward to plateaus and mountains and horizontally into temperate regions, may also be significant depending on the effectiveness of adaptive measures to reduce the threat. Natural ecosystems may also respond adversely to global climate change. With temperatures changing much more rapidly than in most natural sustained climatic shifts, temperature-sensitive plant species may find themselves unable to migrate fast enough to keep up with the changing climate. Even though their suitable habitat may shift only a few hundred kilometers, if plant species cannot reestablish themselves fast enough then they – and many animal species that depend on them – will go at least locally extinct. This is not just theory. Recent analyses of over 1000 published studies have shown that, among other impacts, birds are laying eggs a few weeks earlier, butterflies are moving up mountains, and trees are blooming earlier in the spring and dropping their leaves later in the fall. In her capacity as a lead author for IPCC Working Group II, Terry Root led a group that combed recent literature to conclude that the most consistent explanation for these observed changes in environmental systems over the past few decades is warming on a global scale, and it appears that there is a discernible impact of regional climate change on wildlife and other environmental systems (Root and Schneider, 2002). The responses observed are, in about 80% of the cases, in the direction expected with warming.

Rising temperature also means rising sea level. A popular misconception holds that this is because of melting arctic ice. Actually, ice now floating on the oceans has no direct effect on sea level if it melts. Glaciers and the large ice sheets covering Greenland and Antarctica are a different story, as meltwater from these sources does increase sea level. But the bulk of sea level rise observed to date or expected in the next century comes from the simple thermal expansion of seawater – the same process that drives up the liquid in a mercury thermometer. Determining a global average level for the world ocean is difficult, but measurements suggest that sea level rose some 10–20 cm during the twentieth century. Climate models suggest that the rate of rise should increase as much as fourfold through the current century, resulting in a rise most likely near half a meter. This may not seem like much, but it adds to the highest tides and to

Table 1 Projected effects of global climate change in the twenty-first century (Adapted from Wolfson and Schneider, 2002)

Projected effect	Probability estimate	Examples of projected impacts with high confidence of occurrence (67–95% probability) in at least some areas
Higher maximum temperatures, more hot days and heat waves over nearly all land areas	Very likely (90–99%)	• Increased deaths and serious illness in older age groups and urban poor • Increased heat stress in livestock and wildlife • Shift in tourist destinations • Increased risk of damage to a number of crops • Increased electric cooling demand and reduced energy supply reliability
Higher minimum temperatures, fewer cold days, frost days and cold waves over nearly all land areas	Very likely (90–99%)	• Decreased cold-related human morbidity and mortality • Decreased risk of damage to a number of crops, and increased risk to others • Extended range and activity of some pest and disease vectors • Reduced heating energy demand
More intense precipitation events	Very likely (90–99%) over many areas	• Increased flood, landslide, avalanche, and mudslide damage • Increased soil erosion • Increased flood runoff increasing recharge of some floodplain aquifers • Increased pressure on government and private flood insurance systems and disaster relief
Increased summer drying over most mid-latitude continental interiors and associated risk of drought	Likely (67–90%)	• Decreased crop yields • Increased damage to building foundations caused by ground shrinkage • Decreased water resource quantity and quality • Increased risk of forest fire
Increase in tropical cyclone peak wind intensities, mean and peak precipitation intensities	Likely (67–90%) over some areas	• Increased risks to human life, risk of infectious disease epidemics and many other risks • Increased coastal erosion and damage to coastal buildings and infrastructure • Increased damage to coastal ecosystems such as coral reefs and mangroves
Intensified droughts and floods associated with El Niño events in many different regions	Likely (67–90%)	• Decreased agricultural and rangeland productivity in drought- and flood-prone regions • Decreased hydropower potential in drought-prone regions
Increased Asian summer monsoon precipitation variability	Likely (67–90%)	• Increase in flood and drought magnitude and damages in temperate and tropical Asia
Increased intensity of mid-latitude storms	Uncertain (current models disagree)	• Increased risks to human life and health • Increased property and infrastructure losses • Increased damage to coastal ecosystems

Source: IPCC, 2001b.

the surges associated with major storms (the intensity of which is also expected to increase – see Table 1). Given that much of the world's population lives close to sea level, even a half-meter rise could have serious consequences in some regions, particularly those such as Bangladesh, which possess minimal resources and infrastructure to adapt to rising seas and higher storm surges. However, slow processes such as glacial melting would go on for many centuries, even after greenhouse gas emissions had long been replaced with nonemitting alternative energy systems. Thus, if humans use a substantial fraction of remaining fossil fuels and dump the greenhouse gases produced from their combustion into the atmosphere, then sea level is expected to go on rising, perhaps by several meters or more, over the thousand years that would follow the end of the fossil fuel era (IPCC, 2001c).

Other weather-related projections include increased frequency of intense precipitation events, more heat waves in which the temperature remains at high levels for an extended time, fewer cold waves, more summer droughts,

Table 2 Regional Adaptive capacity, vulnerability, and key concerns[a,b] (relevant sections of the technical summary of IPCC, 2001b for each example are given in square brackets)

Region	
Africa	• Adaptive capacity of human systems in Africa is low due to lack of economic resources and technology, and vulnerability high as a result of heavy reliance on rain-fed agriculture, frequent droughts and floods, and poverty. [5.1.7] • Grain yields are projected to decrease for many scenarios, diminishing food security, particularly in small food-importing countries *(medium to high confidence*[6]). [5.1.2] • Major rivers of Africa are highly sensitive to climate variation; average runoff and water availability would decrease in Mediterranean and southern countries of Africa *(medium confidence*[6]). [5.1.1] • Extension of ranges of infectious disease vectors would adversely affect human health in Africa *(medium confidence*[6]). [5.1.4] • Desertification would be exacerbated by reductions in average annual rainfall, runoff, and soil moisture, especially in southern, North, and West Africa *(medium confidence*[6]). [5.1.6] • Increases in droughts, floods, and other extreme events would add to stresses on water resources, food security, human health, and infrastructures, and would constrain development in Africa *(high confidence*[6]). [5.1] • Significant extinctions of plant and animal species are projected and would impact rural livelihoods, tourism, and genetic resources *(medium confidence*[6]). [5.1.3] • Coastal settlements in, for example, the Gulf of Guinea, Senegal, Gambia, Egypt, and along the East–Southern African coast would be adversely impacted by sea level rise through inundation and coastal erosion *(high confidence*[6]). [5.1.5]
Asia	• Adaptive capacity of human systems is low and vulnerability is high in the developing countries of Asia; the developed countries of Asia are more able to adapt and less vulnerable. [5.2.7] • Extreme events have increased in temperate and tropical Asia, including floods, droughts, forest fires, and tropical cyclones *(high confidence*[6]). [5.2.4] • Decreases in agricultural productivity and aquaculture due to thermal and water stress, sea level rise, floods and droughts, and tropical cyclones would diminish food security in many countries of arid, tropical, and temperate Asia; agriculture would expand and increase in productivity in northern areas *(medium confidence*[6]). [5.2.1] • Runoff and water availability may decrease in arid and semiarid Asia but increase in northern Asia *(medium confidence*[6]). [5.2.3] • Human health would be threatened by possible increased exposure to vector-borne infectious diseases and heat stress in parts of Asia (*medium confidence*[6]). [5.2.6] • Sea level rise and an increase in the intensity of tropical cyclones would displace tens of millions of people in low-lying coastal areas of temperate and tropical Asia; increased intensity of rainfall would increase flood risks in temperate and tropical Asia *(high confidence*[6]). [5.2.5 and Table TS-8] • Climate change would increase energy demand, decrease tourism attraction, and influence transportation in some regions of Asia *(medium confidence*[6]). [5.2.4 and 5.2.7] • Climate change would exacerbate threats to biodiversity due to land use and land-cover change and population pressure in Asia *(medium confidence*[6]). Sea level rise would put ecological security at risk, including mangroves and coral reefs *(high confidence*[6]). [5.2.2] • Poleward movement of the southern boundary of the permafrost zones of Asia would result in a change of thermokarst and thermal erosion with negative impacts on social infrastructure and industries *(medium confidence*[6]). [5.2.2]
Australia and New Zealand	• Adaptive capacity of human systems is generally high, but there are groups in Australia and New Zealand, such as indigenous peoples in some regions, with low capacity to adapt and consequently high vulnerability. [5.3 and 5.3.5] • The net impact on some temperate crops of climate and CO_2 changes may initially be beneficial, but this balance is expected to become negative for some areas and crops with further climate change *(medium confidence*[6]). [5.3.3] • Water is likely to be a key issue *(high confidence*[6]) due to projected drying trends over much of the region and change to a more El Niño-like average state. [5.3 and 5.3.1] • Increases in the intensity of heavy rains and tropical cyclones *(medium confidence*[6]), and region-specific changes in the frequency of tropical cyclones, would alter the risks to life, property, and ecosystems from flooding, storm surges, and wind damage. [5.3.4] • Some species with restricted climatic niches and which are unable to migrate because of fragmentation of the landscape, soil differences, or topography could become endangered or extinct *(high confidence*[6]). Australian ecosystems that are particularly vulnerable to climate change include coral reefs, arid and semiarid habitats in southwest and inland Australia, and Australian alpine systems. Freshwater wetlands in coastal zones in both Australia and New Zealand are vulnerable, and some New Zealand ecosystems are vulnerable to accelerated invasion by weeds. [5.3.2]

Table 2 (*continued*)

Region	
Europe	• Adaptive capacity is generally high in Europe for human systems; southern Europe and the European Arctic are more vulnerable than other parts of Europe. [5.4 and 5.4.6] • Summer runoff, water availability, and soil moisture are likely to decrease in southern Europe, and would widen the difference between the north and drought-prone south; increases are likely in winter in the north and south *(high confidence*[6]). [5.4.1] • Half of alpine glaciers and large permafrost areas could disappear by end of the twenty-first century *(medium confidence*[6]). [5.4.1] • River flood hazard will increase across much of Europe *(medium to high confidence*[6]); in coastal areas, the risk of flooding, erosion, and wetland loss will increase substantially with implications for human settlement, industry, tourism, agriculture, and coastal natural habitats. [5.4.1 and 5.4.4] • There will be some broadly positive effects on agriculture in northern Europe *(medium confidence* 6); productivity will decrease in southern and eastern Europe *(medium confidence*[6]). [5.4.3] • Upward and northward shift of biotic zones will take place. Loss of important habitats (wetlands, tundra, isolated habitats) would threaten some species *(high confidence*[6]). [5.4.2] • Higher temperatures and heat waves may change traditional summer tourist destinations, and less reliable snow conditions may impact adversely on winter tourism *(medium confidence*[6]) . [5.4.4]
Latin America	• Adaptive capacity of human systems in Latin America is low, particularly with respect to extreme climate events, and vulnerability is high. [5.5] • Loss and retreat of glaciers would adversely impact runoff and water supply in areas where glacier melt is an important water source (high confidence[6]). [5.5.1] • Floods and droughts would become more frequent with floods increasing sediment loads and degrade water quality in some areas (high confidence[6]). [5.5] • Increases in intensity of tropical cyclones would alter the risks to life, property, and ecosystems from heavy rain, flooding, storm surges, and wind damages (high confidence[6]). [5.5] • Yields of important crops are projected to decrease in many locations in Latin America, even when the effects of CO_2 are taken into account; subsistence farming in some regions of Latin America could be threatened (high confidence[6]). [5.5.4] • The geographical distribution of vector-borne infectious diseases would expand poleward and to higher elevations, and exposures to diseases such as malaria, dengue fever, and cholera will increase (medium confidence[6]). [5.5.5] • Coastal human settlements, productive activities, infrastructure, and mangrove ecosystems would be negatively affected by sea level rise (medium confidence[6]). [5.5.3] • The rate of biodiversity loss would increase (high confidence[6]). [5.5.2]
North America	• Adaptive capacity of human systems is generally high and vulnerability low in North America, but some communities (e.g., indigenous peoples and those dependent on climate-sensitive resources) are more vulnerable; social, economic, and demographic trends are changing vulnerabilities in subregions. [5.6 and 5.6.1] • Some crops would benefit from modest warming accompanied by increasing CO_2, but effects would vary among crops and regions *(high confidence*[6]), including declines caused by drought in some areas of Canada's Prairies and the U.S. Great Plains, potential increased food production in areas of Canada north of current production areas, and increased warm-temperate mixed forest production *(medium confidence* 6). However, benefits for crops would decline at an increasing rate and possibly become a net loss with further warming *(medium confidence*[6]). [5.6.4] • Snowmelt-dominated watersheds in western North America will experience earlier spring peak flows *(high confidence*[6]), reductions in summer flows *(medium confidence*[6]), and reduced lake levels and outflows for the Great Lakes-St. Lawrence under most scenarios *(medium confidence*[6]); adaptive responses would offset some, but not all, of the impacts on water users and on aquatic ecosystems *(medium confidence*[6]). [5.6.2] • Unique natural ecosystems such as prairie wetlands, alpine tundra, and cold-water ecosystems will be at risk and effective adaptation is unlikely *(medium confidence*[6]). [5.6.5] • Sea level rise would result in enhanced coastal erosion, coastal flooding, loss of coastal wetlands, and increased risk from storm surges, particularly in Florida and much of the U.S. Atlantic coast *(high confidence*[6]). [5.6.1] • Weather-related insured losses and public sector disaster relief payments in North America have been increasing; insurance sector planning has not yet systematically included climate change information, so there is potential for surprise *(high confidence*[6]). [5.6.1] • Vector-borne diseases – including malaria, dengue fever, and Lyme disease – may expand their ranges in North America; exacerbated air quality and heat stress morbidity and mortality would occur *(medium confidence*[6]); socioeconomic factors and public health measures would play a large role in determining the incidence and extent of health effects. [5.6.6]

(*continued overleaf*)

Table 2 (*continued*)

Region	
Polar	• Natural systems in polar regions are highly vulnerable to climate change and current ecosystems have low adaptive capacity; technologically developed communities are likely to adapt readily to climate change, but some indigenous communities, where traditional lifestyles are followed, have little capacity and few options for adaptation. [5.7] • Climate change in polar regions is expected to be among the largest and most rapid of any region on the Earth, and will cause major physical, ecological, sociological, and economic impacts, especially in the Arctic, Antarctic Peninsula, and Southern Ocean (*high confidence*[6]). [5.7] • Changes in climate that have already taken place are manifested in the decrease in extent and thickness of Arctic sea ice, permafrost thawing, coastal erosion, changes in ice sheets and ice shelves, and altered distribution and abundance of species in polar regions (*high confidence*[6]). [5.7] • Some polar ecosystems may adapt through eventual replacement by migration of species and changing species composition, and possibly by eventual increases in overall productivity; ice edge systems that provide habitat for some species would be threatened (*medium confidence*[6]). [5.7] • Polar regions contain important drivers of climate change. Once triggered, they may continue for centuries, long after greenhouse gas concentrations are stabilized, and cause irreversible impacts on ice sheets, global ocean circulation, and sea level rise (*medium confidence*[6]). [5.7]
Small Island States	• Adaptive capacity of human systems is generally low in small island states, and vulnerability high; small island states are likely to be among the countries most seriously impacted by climate change. [5.8] • The projected sea level rise of 5 mm per year for the next 100 years would cause enhanced coastal erosion, loss of land and property, dislocation of people, increased risk from storm surges, reduced resilience of coastal ecosystems, saltwater intrusion into freshwater resources, and high resource costs to respond to and adapt to these changes (*high confidence*[6]). [5.8.2 and 5.8.5] • Islands with very limited water supplies are highly vulnerable to the impacts of climate change on the water balance (*high confidence*[6]). [5.8.4] • Coral reefs would be negatively affected by bleaching and by reduced calcification rates caused by higher CO_2 levels (*medium confidence*[6]); mangrove, sea grass beds, and other coastal ecosystems and the associated biodiversity would be adversely affected by rising temperatures and accelerated sea level rise (*medium confidence*[6]). [4.4 and 5.8.3] • Declines in coastal ecosystems would negatively impact reef fish and threaten reef fisheries, those who earn their livelihoods from reef fisheries, and those who rely on the fisheries as a significant food source (*medium confidence*[6]). [4.4 and 5.8.4] • Limited arable land and soil salinization makes agriculture of small island states, both for domestic food production and cash crop exports, highly vulnerable to climate change (*high confidence*[6]). [5.8.4] • Tourism, an important source of income and foreign exchange for many islands, would face severe disruption from climate change and sea level rise (*high confidence*[6]) . [5.8.5]

[a]Because the available studies have not employed a common set of climate scenarios and methods, and because of uncertainties regarding the sensitivities and adaptability of natural and social systems, the assessment of regional vulnerabilities is necessarily qualitative.
[b]The regions listed in Table 2 are graphically depicted in Figure TS-2 of the Technical Summary of IPCC, 2001b.
Note: Adapted from Wolfson and Schneider, 2002.
Source: IPCC, 2001b.

and more wet spells in winter. The intensity of tropical cyclones (hurricanes and typhoons) is likely to increase, although it is less clear whether the frequency or average locations of these storms will change. Hail and lightning are also likely to become more frequent. The large-scale Pacific Ocean fluctuation known as the *El Niño/Southern Oscillation* could become more persistent, which would have a substantial climatic impact on the Americas and Asia. All these projected changes will impact agriculture and may increase flooding and erosion, with concomitant effects on health and on the insurance industry. As shown in Table 1, confidence in this group of consequences ranges from medium (likelihood between one-third and two-thirds) to high (greater than two out of three chances). Keep in mind, however, that the probabilities given in Table 1 are not based on conventional statistical analysis because they refer to future events that do not follow past patterns – and obviously, the future has not occurred yet. Rather, these are subjective odds based on scientific judgment as sound as current understanding permits.

Finally, there is the remote possibility of dramatic changes such as alterations in large-scale patterns of ocean

circulation or the disintegration of the West Antarctic Ice Sheet. These could occur because the climate system is inherently nonlinear, meaning that a small change in some conditions can produce a disproportionately large change in others. Changes in the Gulf Stream – part of the so-called ocean thermohaline circulation – caused by greenhouse gas emissions, could eventually leave northwest Europe with a much colder climate. Climate models predict with high confidence that the thermohaline circulation will weaken over the present century. But they also suggest, fortunately, that wholesale disruption is very unlikely at least before the year 2100. However, the models also warn that what humans do in the twenty-first century can precondition what the ocean currents will do in the twenty-second century and beyond. Potentially irreversible events could be built into the long-term planetary future even if those of us living in the twenty-first century are spared the experience of those effects (Rahmstorf, 1999; Schneider and Thompson, 2000). Similarly, recent studies suggest that the West Antarctic Ice Sheet is likely to remain stable for the foreseeable future, which is a very good thing because its breakup would result in a rise in sea level by some six meters. But that "unlikely possibility" is not ruled out and looms as a potential threat that we need to check for periodically as we advance our understanding of the climate system and its potential for surprises.

There is one final note on the issue of climatic impacts. In the above example of Bangladesh suffering from sea level rises or more intense storms, we mentioned that adaptation would be difficult. This is much less the case for a richer, more technologically advanced country such as the Netherlands. In fact, as is illustrated in Table 2 (in which IPCC, 2001b authors summarize a comprehensive list of potential climate-change impacts for most regions of the world and economic sectors), a consensus is building in the scientific community that the damages that climatic changes might inflict on societies will depend in part on the adaptive capacities of those future societies, which in turn depend on their resource bases and technological and infrastructure capabilities (IPCC, 2001b). This suggests, as Table 2 notes, that damages may be asymmetrically felt across the developed/developing country divide. The scenario where the northern rich countries get longer growing seasons, and the poor tropical nations get more intense droughts and floods is clearly a situation ripe for increasing tensions in the world of the twenty-first century.

We have given a brief description of the causes and anticipated consequences of global climate change. Even if we humans get our greenhouse gas emissions under control – not a likely occurrence in the near future – global temperature will continue to rise toward a new equilibrium value that will take at least many decades – more likely centuries – to become established. The effects of global climate change, in particular sea level rise, will almost certainly continue to increase beyond the end of the twenty-first century, and they may well become far more dramatic over the following centuries.

REFERENCES

Harvey L.D.D. (2000) *Global Warming: The Hard Science*, Prentice Hall: Englewood Cliffs, p. 336.

Hoyt D.V. and Schatten K.H. (1997) *The Role of the Sun in Climate Change*, Oxford University Press: New York.

Intergovernmental Panel on Climate Change (IPCC) (2001a) *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the IPCC*, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge.

Intergovernmental Panel on Climate Change (IPCC) (2001b) *Climate Change 2001: Impacts, Adaptation, and Vulnerability, Contribution of Working Group II to the Third Assessment Report of the IPCC*, McCarthy J.J., *et al.* (Eds.), Cambridge University Press: Cambridge.

Intergovernmental Panel on Climate Change (IPCC) (2001c) *Climate Change 2001: Synthesis Report, Contribution of Working Groups I, II, and III to the Third Assessment Report of the IPCC*, Watson R.T., *et al.* (Eds.), Cambridge University Press: Cambridge.

Jacobson M.Z. (2001) A physically based treatment of elemental carbon: implications for global direct forcing of aerosols. *Geophysical Research Letters*, **27**, 217–220.

Jouzel J, Lorius C., Petit J.R., Genthon C., Barkov N.I., Kotlyakov V.M. and Petrov V.M. (1987) Vostok ice core: a continuous isotope temperature record over the last climatic cycle (160,000 years). *Nature*, **329**, 403.

Keeling C.D. and Whorf T.P. (2000) Atmospheric CO_2 records from sites in the SIO air sampling network. *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory: Oak Ridge.

Mann M.E., *et al.* (1999) Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. *Geophysical Research Letters*, **26**, 759.

Neftel A, *et al.* (1994) Historical CO_2 record from the siple station ice core. *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory: Oak Ridge.

Petit J.R., *et al.* (2000) Historical isotopic temperature record from the Vostok ice core. *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory: Oak Ridge.

Rahmstorf S (1999) Shifting seas in the greenhouse. *Nature*, **399**, 523–524.

Root T.L. and Schneider S.H. (2002) Climate change: overview and implications for wildlife. *Wildlife Responses to Climate Change: North American Case Studies*, Schneider S.H. and Root T.L. (Eds.), National Wildlife Federation, Island Press: Washington.

Schneider S.H. and Thompson S.L. (2000) A simple climate model used in economic studies of global change. *New Directions in the Economics and Integrated Assessment of Global Change*, DeCanio S.J., *et al.* Pew Center on Global Climate Change: Washington.

Schneider S.H. (2003) Imaginable surprise. *Handbook of Weather, Climate, and Water*, John Wiley & Sons: Chichester.

Wolfson R and Schneider S.H. (2002) Understanding Climate Science. In *Climate Change Policy*, Schneider S.H., *et al.* (Eds.), Island Press: Washington.

34: Climate Change – Past, Present and Future

HENRY HENGEVELD

Meteorological Service of Canada, Toronto, ON, Canada

Local climate is a major determinant of both the composition and behavior of a region's ecosystems and of the infrastructure and culture of human society residing within the region. Hence, climatic statistics are an important factor in ecosystem management and planning for socioeconomic development. However, although decision makers often treat these statistics as a constant, there is clear evidence that climate has changed, is changing, and will change. Large natural changes have occurred in the geologic past. However, these have been relatively modest during the past few millennia. On the other hand, human interference with the climate system, primarily through land use change and changes in atmospheric composition, is now adding an unprecedented and increasingly dominant force for change. There is convincing evidence that this interference has already caused a substantial increase in global mean temperatures over the past 50 years. Projected changes for the next century will likely exceed anything yet experienced in human history and could rival the magnitude of very large changes during the past million years, but at a much more rapid rate. Such rapid change will have dramatic implications for, inter alia, the global hydrological cycle and the frequency and severity of extreme weather events.

INTRODUCTION

Climate is commonly defined as average weather. That is, the climate of a particular locale or region is the average of the day-to-day variations in temperature, precipitation, cloud cover, wind, and other atmospheric conditions that normally occur there over an extended period of time (usually three decades or more). But climate is more than just the sum of these average values. It is also defined by the variability of individual climate elements, such as temperature or precipitation, and by the frequency with which various kinds of weather conditions occur. Indeed, any factor that is characteristic of a particular location's weather pattern is part of its climate.

Although the very notion of climate as described above assumes a long-term consistency and stability in these patterns, climate is, nevertheless, a changeable phenomenon. That is because the Earth's climate system is a dynamic system that continuously responds to forces, both internal and external, that alter the delicate balances that exist within and between each of its components. Often, these changes are relatively small in magnitude and short in duration – like a period of cool climate conditions following a large volcanic eruption. However, evidence from the Earth's soils, its ocean, and lake bottom sediments, its coral reefs, its ice caps, and even its vegetation collectively indicate that such forces can also cause major, long-term shifts in climate. For example, over long timescales of hundreds of thousands of years or more, the Earth's climate has undergone very large shifts from glacial to interglacial conditions and back again – changes that caused global redistributions of flora and fauna. However, during the past 10 000 years of the current interglacial, changes in climate have been of relatively small magnitude and, while from time to time disruptive on a regional scale, have allowed global vegetation to flourish over most land masses.

The focus of this article will be on changes that have occurred over the past century of the current interglacial and that can be expected to occur over the next century. It is within the relatively short time span of the past century that the first real evidence of human interference with the natural processes governing the climate system has become apparent. It is within the next century that such interference is expected to pose the real risk of danger to the well being of global ecosystems and society. However, to understand how these human influences affect climate, it is necessary to

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

provide context by first describing how the climate system works and to discussing in greater detail how and why the Earth's climate of the more distant past has varied and changed.

EARTH'S CLIMATE SYSTEM

Global Energy Balance

In many respects, the Earth's climate system can be thought of as a giant heat engine fueled by incoming energy from the sun. As the solar energy passes through the engine, it warms the Earth and surrounding air, setting the atmospheric winds and the ocean currents into motion and driving the evaporation-precipitation processes of the water cycle. The result of these motions and processes is weather and, hence, climate.

To avoid overheating the climate system, the energy entering the climate system must eventually be released again back to space. As long as the same amount of energy leaves the system as enters it, our atmospheric heat engine will be in balance and the Earth's average temperature will remain relatively constant. However, if the amount of energy entering or leaving the climate system changes, the balance will be upset and global temperatures will change until the system adjusts itself and reaches a new equilibrium. This flow of energy through the climate system is largely regulated by the Earth's atmosphere, although the radiative properties of the Earth's surface are also important factors. The fundamental processes involved are described in the paragraphs below (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**).

Incoming Solar Radiation

Averaged around the world, the amount of sunlight entering the atmosphere is about 342 W m^{-2}. However, as illustrated in Figure 1 (Baede *et al.*, 2001), about 31% of this incoming short wave energy is reflected back to space by the atmosphere and the Earth's surface, unused by the climate system. The remaining 69% (about 235 W m^{-2}) is absorbed within the atmosphere (~20%) and by the Earth's surface (~49%) as the fuel that drives the global climate system. The left side of Figure 1 shows how this incoming energy is reflected and absorbed by the atmosphere and surface. The processes are as follows:

Reflection by the atmosphere and Earth's surface – Clouds and aerosols (fine solid particles and liquid droplets) within the atmosphere reflect and scatter a significant amount of incoming solar radiation back to space. Highly reflective aerosols include tiny droplets of sulfuric acid from volcanic eruptions, sulfates from vegetation fires and industrial processes, salt from sea spray and dust. The amount of shortwave radiation returned to space by clouds and aerosols varies considerably with time and space. For example, major volcanic eruptions can abruptly produce large amounts of highly reflecting sulfate aerosols in the stratosphere that can remain there for several years before they settle out due to the forces of gravity. Alternatively, human emissions of sulfate aerosols into the lower atmosphere can significantly increase the reflection of incoming sunshine in industrialized regions compared to less polluted areas

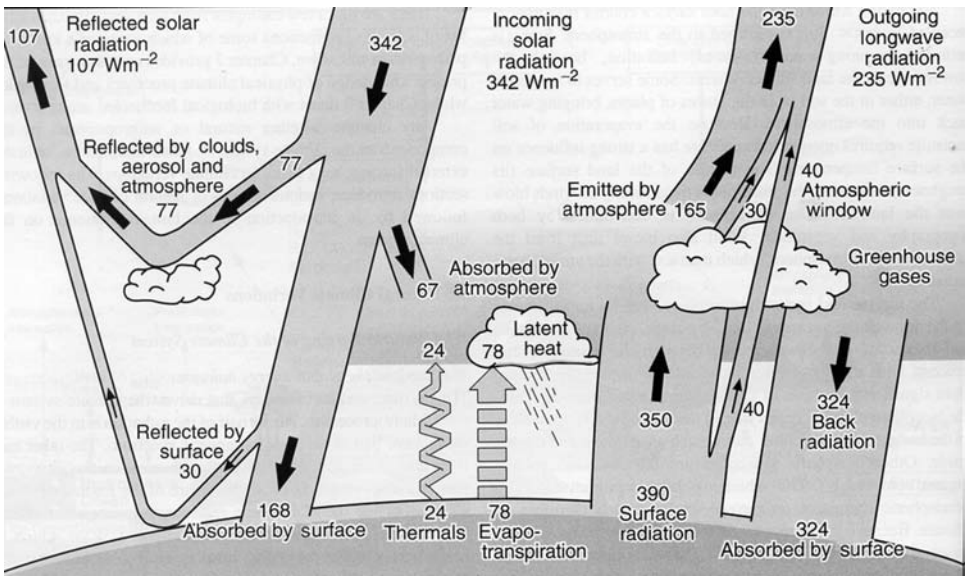


Figure 1 Illustration showing how the atmosphere and earth's surface affects the climate system's incoming and outgoing energy (IPCC Third Assessment Report 2001. Reproduced by permission of Intergovernmental Panel on Climate Change)

of the world. Observational data indicate that, on average, clouds and aerosols currently reflect about 22.5% of incoming radiation back to space. In addition, almost 9% of the solar energy is reflected back to space by the Earth's surface. Like the atmosphere, the albedo (or reflectivity) of the Earth's surface depends on the time of year and the location. That is because snow and ice, which cover much of the Earth's surface in mid to high latitudes in winter, are highly reflective, while ice-free oceans surfaces and bare soils are low reflectors.

Absorption in the atmosphere – In addition to reflecting and scattering solar radiation, the atmosphere also absorbs almost 20% of this energy. About two-thirds of this absorption is caused by water vapor. A second significant absorber is the ozone layer in the stratosphere, which absorbs much of the ultraviolet part of incoming solar energy. Thus, this layer not only protects the Earth's ecosystem from the harmful effects of this radiation but also retains a portion of the sun's energy in the upper atmosphere. Another one-tenth of the absorption can be attributed to clouds. Finally, a small fraction of the absorption can be attributed to other greenhouse gases and to aerosols (particularly dark aerosols such as soot). However, most of the gases of the atmosphere are relatively transparent to incoming sunlight.

Outgoing Heat Radiation

The Earth's atmosphere and surface, heated by the unreflected portion of the sun's rays, eventually release this energy back to space again by giving off long-wave infrared radiation. When the climate system is in equilibrium, the total amount of energy released back to space by the climate system must, on average, be the same as that which it absorbs from the incoming radiation – that is, $235\,W\,m^{-2}$. However, as the infrared radiation tries to escape to space, it encounters several major obstacles. These can be described as follows:

Clouds. Besides reflecting incoming solar radiation, clouds also absorb large quantities of outgoing heat radiation. The energy absorbed by clouds is reradiated, much of it back to the surface. That is why air near the Earth's surface is usually much warmer on a cloudy night than on a clear one. The amount of radiation absorbed and returned depends on the amount, thickness, and type of cloud involved.

Absorbing gases – A number of naturally occurring minor gases within the atmosphere, most of which are relatively transparent to incoming sunlight, absorb most of the infrared heat energy being transmitted by the Earth toward space. This absorbed energy is then reradiated in all directions, some back to the surface and some upward where other absorbing molecules at higher levels in the atmosphere are ready to absorb the energy again. Eventually, the absorbing molecules in the upper part of the atmosphere emit the energy directly to space. Hence, these gases make the atmosphere opaque to outgoing heat radiation, much as opaque glass will affect the transmission of visible light. Together with clouds, they provide an insulating blanket around the Earth, keeping it warm. Because they retain heat in somewhat the same way that glass does in a greenhouse, this phenomenon has been called the *greenhouse effect*, and the absorbing gases that cause it, *greenhouse gases*. Important, naturally occurring greenhouse gases include water vapor, carbon dioxide, methane, ozone, and nitrous oxide.

The magnitude of the thermal-insulating effect caused by greenhouse gases and clouds can be estimated fairly easily. Theoretically, the average radiating temperature required to release $235\,W\,m^{-2}$ to space is $-19\,^{\circ}C$. Yet, we know from actual measurements that the Earth's average surface temperature is more like $+14\,^{\circ}C$, some $33\,^{\circ}C$ higher. This additional warming is a result of the greenhouse effect. It is enough to make the difference between a planet that is warm enough to support life and one like the moon that is not.

Dynamics of the Climate System

Climate and climate variability at the Earth's surface are ultimately consequences of the way the atmosphere and the oceans redistribute and release heat energy that the Earth has absorbed from the sun. Because the intensity of the solar radiation changes with latitude, time of day and seasons, all parts of the planet are not heated equally. The heating effect is greatest in the tropics, where more energy is received from the sun then is reradiated back to space. Temperatures here are subsequently much warmer than the global average, remaining consistently within a few degrees of $30\,^{\circ}C$ during all seasons of the year. At the opposite extreme, the Earth's polar regions experience a net loss of energy to space. The magnitude of this loss varies significantly with time, being largest in winter seasons and weakest in summer. Hence, temperatures in polar regions can vary from highs of nearly $20\,^{\circ}C$ in northern polar summers to a low of $-60\,^{\circ}C$ in southern polar winters.

These large temperature differences between the tropics and the poles are the primary driving force for Earth's atmospheric winds and ocean currents. Essentially, these carry warm air and water from the equator to the poles, while cold air and water move in the opposite direction. This flow is modified, however, by the Earth's spin about its axis and the effects of landmasses and topography to produce a complex pattern of vertical and horizontal circulation of air masses and ocean waters.

Much of the sun's energy absorbed at the Earth's surface is used to evaporate water from ocean and land surfaces and ecosystems. The more heat at the surface and the warmer the air temperature, the greater the amount of water vapor that can be evaporated from the surface and retained within the atmosphere. Once air becomes

saturated with water vapor, it condenses again into tiny water droplets or ice crystals that form clouds. When the conditions are right, these droplets or crystals fall to the ground as precipitation. Where, when, how much, and what type of precipitation falls will depend on the characteristics of a range of local atmospheric and surface factors. Furthermore, since atmospheric moisture is also transported horizontally by air currents, the precipitation patterns that emerge around the Earth are also influenced by the global atmospheric circulation patterns. As a result, the distribution of precipitation around the globe presents an even more complex pattern than that for atmospheric circulation. Some areas receive large surpluses of rainfall, which support very lush, rich ecosystems, while others do not receive enough to nourish vegetation and so become deserts (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**).

Numerous other factors also affect the Earth's climate. In addition to the circulation of the air, the currents of the ocean and the surface properties that influence evaporation processes, we must also consider the effects of clouds and of snow and ice, the influence of topography and the impact of processes and activities within the biosphere. To these we must add the variations in solar heating with time, not only between seasons but also during each day.

All of these elements are interconnected, interacting parts of the climate system. If a change in one of these parts upsets the balance of that system, it is likely to initiate complex reactions in some or all of the other parts as the system adjusts to establish a new equilibrium. Some reactions occur very rapidly, while others occur very, very slowly. Furthermore, some may increase the initial change (a process known as *positive feedback*), while others may oppose and partially offset it (a *negative feedback*). For example, anything that changes the amount of solar energy entering the atmosphere or the amount of this energy absorbed by the atmosphere will alter the amount of input energy that drives the climate system. Likewise, a change in the net amount of energy released by the climate system back to space will cause a change in the Earth's cooling mechanism. Such initial changes will cause reactions and feedbacks in the rest of the system and accompanying changes in surface climates until the system adjusts to a new balance at the top of the atmosphere between the incoming and outgoing energy.

CLIMATES OF THE PAST

Reconstructing Past Climates

Within the rich diversity of living species around the world, there are those that thrive in hot climates and others that prefer cooler and even cold climates. Some like it wet, and some like it dry. The result is a broad range in the composition of regional ecosystems, with the characteristics of each largely determined by its prevailing climate. Therefore, if the climate of a particularly region changes over time, so will its ecological composition. As these species procreate and eventually die, they also leave vestiges of their presence in the surrounding ice, soils, rocks, corals, and/or lake and ocean sediments – letting the paleoclimatologists who later analyze these repositories know that they have been there, and informing these researchers about the environment within which they existed. These indicators of past ecological composition provide an abundance of evidence as to how climates have varied and changed over timescales that vary from hundreds to many millions of years. There are also other nonbiotic proxies for past climate, such as the vertical heat profile in the Earth's crust or the isotopic composition of ice buried in polar or alpine ice sheets. For more recent times, there are the added proxy data available from human anecdotal records – like information on dates of harvest, the types of crops grown, major weather catastrophes – that can help the climate detective reconstruct climate patterns of the past (Beltrami, 2002; Mann, 2002; Shrag and Linsley, 2002).

The Past 1.5 Million Years

During the last 1.5 million years (the Quaternary), Earth's climate has undergone large swings between lengthy periods of extreme minimum temperatures accompanied by extensive glaciation and shorter periods of warm interglacial conditions. Ice core data extracted from the ice sheets of Antarctica provide valuable information about the last four of these large cycles. As illustrated in Figure 2 (Folland *et al.*, 2001a), these data indicate that the glacial-interglacial cycles occurred about once every 1 00 000 years. During each cycle, there was a slow transition from interglacial to glacial conditions, but a rapid deglaciation at the end of each cold period. Over Antarctica, for example, local temperatures warmed by a dramatic 8 to 10 °C over an interval of 5000 years or so during each glacial-interglacial transition. Within this 1 00 000-year cycle, smaller anomalies have occurred with regularity. Similar patterns are found in data extracted from Greenland ice cores and from ocean sediments, although the latter suggest a more modest change in temperatures for tropical regions. Model studies, in fact, suggest that the global average change in temperature during a glacial-interglacial cycle may be a more moderate 4 to 6 °C, producing an average rate of warming during deglaciation of about 0.1 °C century^{-1} (Weaver *et al.*, 1998; Bush and Philander, 1999).

Although many theories have been advanced to explain these large temperature variations, the most widely accepted hypothesis involves changes in the precession, obliquity, and eccentricity of the Earth's orbit around the sun, occurring with frequencies of 22 000, 41 000, and 1 00 000 years

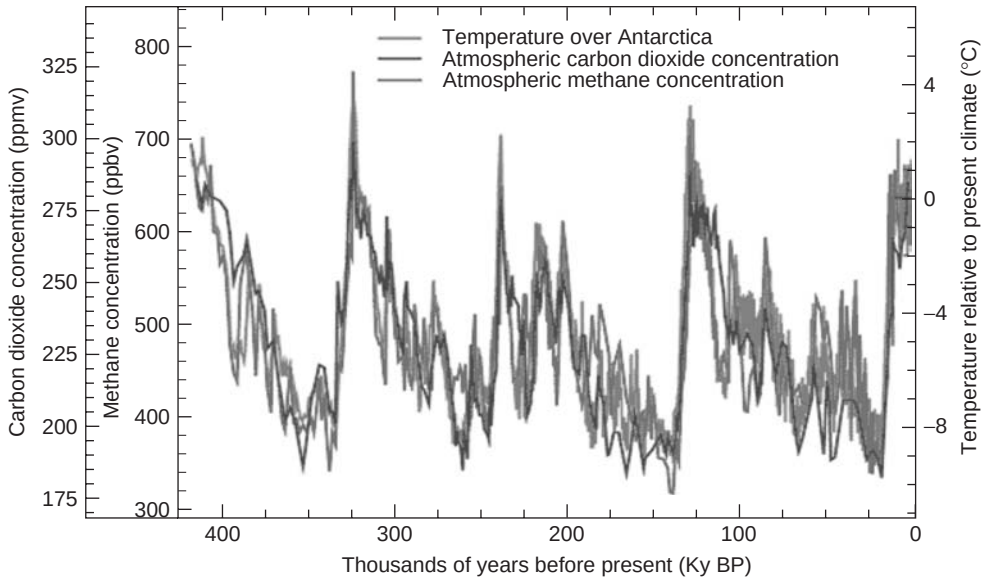


Figure 2 Long-term variations in Antarctic temperatures, carbon dioxide concentrations and methane concentrations over the past 4 20 000 years, as reconstructed from ice core data (IPCC Third Assessment Report 2001. Reproduced by permission of Intergovernmental Panel on Climate Change). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

respectively. These changes affect the seasonal distribution of incoming sunlight across the Earth's surface. However, while the large glacial-interglacial cycles that occur at the 1 00 000-year timescale correlate well with changes in orbit eccentricity, the net annual forcing (i.e. imbalance between incoming and outgoing radiative energy flux at the top of the atmosphere that causes the climate system to adjust until a new balance is achieved) caused by those changes is far too weak to fully explain the amplitude of the glacial-interglacial cycles. Hence, various feedback processes appear to significantly amplify this forcing. Paleo studies indicate that responsive changes in atmospheric greenhouse gas concentrations and altered surface albedos are two such important positive feedback mechanisms. For example, recent analyses of Antarctic and Greenland ice cores indicate a strong correlation between past long-term changes in climate and the natural atmospheric concentrations of carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O), all important greenhouse gases. As Figure 2 indicates, the correspondence between atmospheric carbon dioxide, methane concentrations, and local Antarctic temperatures during the past 4 20 000 years has been remarkable. However, the various processes involved in such millennial scale changes in climate are very complex, and can differ between hemispheres. Furthermore, they may also differ from one cycle to the next, suggesting that past events may not be good analogs for the current interglacial. Some researchers argue that the current interglacial could, in fact, last another 50 000 years. Human factors may further dampen or even prevent the next reglaciation, thus ushering

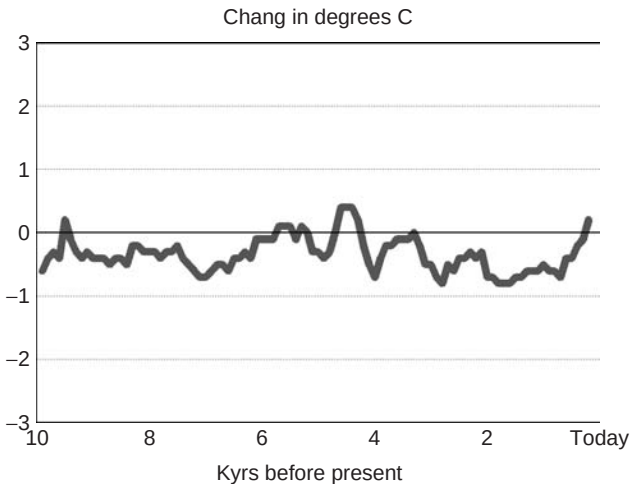


Figure 3 Variations in Antarctic temperatures over the past 10 000 years, as reconstructed from ice core data

in the "Anthropocene" stage of Earth's climate (Berger and Loutre, 2002; Crowley, 2002; Pepin *et al.*, 2001).

Figure 3 (Petit *et al.*, 2000) provides a more detailed depiction of Antarctic temperature variations for the past 10 000 years. During this interglacial period (known as the *Holocene*), average mid- to high-latitude temperatures peaked slightly about 5000 to 6000 years before present, and have gradually cooled since then. This warm peak of the interglacial is commonly referred to as the *Holocene Maximum*. During this high-latitude Holocene Maximum, central North American climates were generally warmer,

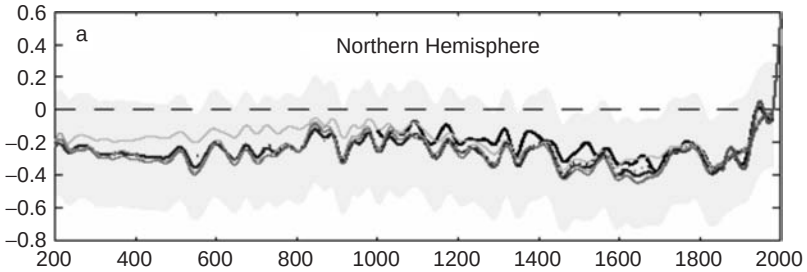


Figure 4 Northern hemispheric climate trends of the last 2000 years, as reconstructed from multiple proxy data sources. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

drier, and windier than that of today. In contrast, European climates during that period were initially warmer and wetter, then they also became drier. Climates in arid regions of Africa and Asia were also significantly wetter than today. However, both paleo data and model studies suggest that mid-Holocene temperatures may have been slightly cooler than today in low-latitude regions. Hence, average global temperatures appear to have been remarkably stable during the entire Holocene. Several "little ice ages", or short periods of cooling, appear superimposed upon the Holocene record at approximately 2500-year intervals, the latest having occurred between about 1400 A.D. and 1900 A.D. (Adams, 1997; Gajewski *et al.*, 2000).

Figure 4 (Mann and Jones, 2003) illustrates how average global temperatures have varied during the past 2000 years. On the basis of multiple proxy data sources at many different locations around the world (although predominantly from the Northern Hemisphere), it suggests that average temperatures today are warmer than at any previous time during the past two millennia. However, it appears likely that current temperatures are still slightly cooler than those of the Holocene maximum, and perhaps 1 to 2 °C cooler than the peak of the last interglacial of 1 35 000 YBP.

For fluctuations during the past several millennia, there have been encouraging results from efforts at correlating changes in temperature with volcanic forcing and solar irradiance cycles. Such comparisons, for example, indicate that volcanic forcing can be important in inducing long-term cooling on a global scale, but less so at the regional scale. In contrast, solar forcing can create significant long-term regional climate changes as well as global scale anomalies similar to the Little Ice Age events noted above. In fact, much of the higher frequency variations in climate over the past 300 years appear to be closely linked to changes in sunspot cycle behavior. However, the mechanisms by which relatively small change in solar irradiance can significantly affect climate are as yet not well understood (Shindell *et al.*, 2003; Rind, 2002).

While Figure 4 indicates that average Northern Hemispheric temperatures about 1000 years ago were still cooler than today, some regions may have been somewhat warmer at that time. These regions include western Europe, Greenland, and eastern Canada. This regional warm period, often referred to as the *Medieval Warm Period*, lasted several hundred years. Hence, during this period, local tree lines and other natural vegetation boundaries gradually moved northwards. Milder Arctic climates brought substantial decreases in sea ice cover. These conditions may have encouraged the migration of Inuit within the Arctic and have been a determining factor in the settlement and survival of European Vikings in Iceland and Greenland. These Vikings appear to have navigated freely throughout much of the Canadian archipelago, and, in Greenland, they were able to carry on a viable agriculture. Ironically, European attempts from the seventeenth to the nineteenth centuries to find a northwest passage to India failed. This was primarily because they began after the Medieval Warm Period gave way to the Little Ice Age, which lasted from 1400 to about 1850 A.D. Since temperatures in some of these regions have now returned to something similar to the Medieval Warm Period, it is likely that the local vegetation and ice regimes that prevailed years ago will return, even if climates do not become any warmer than they are now (Briffa and Osborn, 2002).

Paleo reconstructions can also provide important information on changes in circulation patterns and precipitation regimes. They indicate, for example, that the wind strength of the southwestern Asian monsoon has increased over the past four centuries. While it remains uncertain whether this trend was linked to a warming of global climates, it is consistent with projections for increased intensity of the southwest monsoon under future warming. Changes in monsoonal behavior can have large impacts on regional societies, causing severe droughts when the intensity of the monsoons decrease and severe flooding when their intensity increases (Anderson *et al.*, 2002).

CLIMATES OF THE TWENTIETH CENTURY

Monitoring Recent Climate Behavior

While the Earth's natural environment and human anecdotal information recorded over time have been valuable

sources of proxy data to help provide approximations of past climate behavior, they have major limitations in terms of spatial and temporal details and provide little information on aspects of climate other than temperature and precipitation.

With the advent of instrumental climate record keeping in Europe several centuries ago, systematic observations of temperature, precipitation, and many other climate variables began to remove some of these limitations. Initially the spatial coverage of climate monitoring systems was sparse, particularly in polar regions and parts of North America, Africa, China, and Russia. However, global coverage was much improved by mid-twentieth century. The advent of satellite observing systems some 25 years ago has further added to this coverage.

In addition to the possible bias introduced by changes in observing coverage and density over time, there are other significant challenges in analyzing these data records for trends and variations in regional and global climate conditions. For example, observing methods have changed over time. Furthermore, land use change such as deforestation or increased urbanization has caused significant biases in many local temperature records. Various research groups have worked meticulously to identify possible systematic biases in these records and to adjust them accordingly. Although there continue to be uncertainties in results, the high level of consistency between the various independent analyses undertaken to date and between corrected sea and land data where they abut along coastlines lends considerable confidence in the significance of the trends observed, particularly at the global scale (Folland *et al.*, 2001b).

Analyzing global trends in precipitation and other hydrometeorological variables (including cloud) is even more problematic, since hydrological variables are more significantly influenced by local influences that increase their inhomogeneity over space and time. Furthermore, there is relatively little information for monitoring trends in precipitation over oceans. Hence, while good estimates for precipitation trends are available for some land regions with long records and reasonably dense network of monitoring stations, there are no reliable estimates of global trends (Folland *et al.*, 2001a).

Over the past 50 years, there has also been an increasing array of complementary measurements of meteorological conditions within the atmosphere provided by balloon-borne instrument packages. More recently, satellite systems have added to this database. These data have helped understand global trends in atmospheric conditions better, including cloud cover, humidity, and atmospheric temperatures. Finally, there are many indirect indicators of recent and current trends in climate provided by monitoring of the global cryosphere (snow cover, sea ice, and glaciers) and of behavior of flora and fauna (Folland *et al.*, 2001a).

Temperature Trends

Globally, average surface temperatures have increased by about 0.7 °C (±0.2 °C) over the past century. However, as shown in Figure 5 (National Climatic Data Center, 2004), the observed global trends in temperature have not been uniform in time. While average temperatures changed very little between 1860 and 1920, they increased relatively rapidly over the next two decades. The climate cooled moderately from midcentury until the early 1970s, then warmed rapidly at about 0.15 °C decade^{-1} during the past 30 years. During the more recent warming period, nighttime minimum temperatures have been increasing at about twice the rate of daytime maximum temperatures, thus decreasing the diurnal temperature range. Land surface temperatures have also been rising at about twice the rate of sea surface temperatures. Together, these factors have contributed to a lengthening of the frost-free period over lands in mid to high latitudes (Jones and Moberg, 2003; Stone and Weaver, 2002).

When compared with the proxy data for climate variations of the past two millennia (Figure 4), it seems likely that the twentieth century is now the warmest over that time period, and that the 1990 s was the warmest decade. Furthermore, the *rate* of warming in recent decades appears to be unprecedented over at least that time period. This evidence has led experts to conclude that recent climate behavior has been very unusual and thus increasingly difficult to explain on the basis of natural variability and/or natural forcing factors such as changing intensity of solar irradiance or atmospheric dust loading from volcanic eruptions. In contrast, there is increasing evidence that recent temperature behavior is consistent with how the climate should

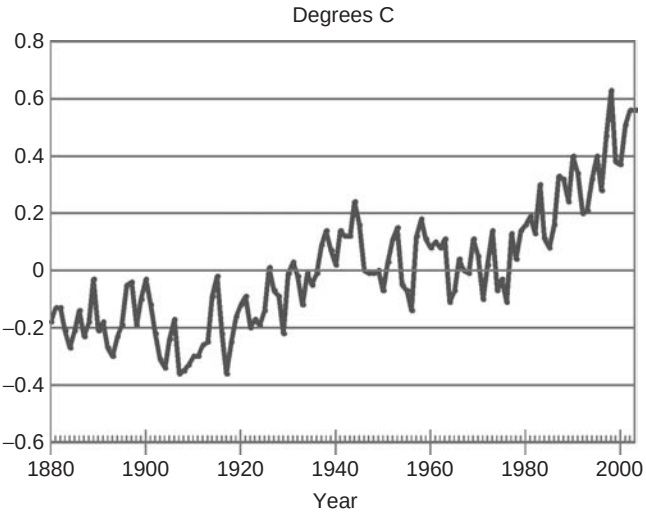


Figure 5 Trends in global average surface temperatures over the past 140 years, shown as departures from the long-term mean. Based on US National Climatic Data Center data files

have responded to increasing human interferences with the climate system, and therefore likely due to anthropogenic causes (Folland *et al.*, 2001b; Mann *et al.*, 2003).

Although the monitoring of temperatures within the Earth's atmosphere has a much shorter history than that for surface temperatures, climatologists now have some 45 years of data directly recorded by radiosondes borne aloft by balloons and almost 25 years of information obtained indirectly by instruments onboard satellites, particularly the Microwave Sounding Unit (MSU). Comparison of the longer radiosonde records with surface observations show that the long-term trend of globally averaged temperatures in the lower atmosphere since 1957 is very similar to that at the surface. However, there are significant and varying differences in trends on shorter timescales. For example, the lower atmosphere warmed more rapidly than the surface between 1957 and 1975, but warmed at a slower rate since that time. The satellite MSU data shows a similar slower atmospheric warming relative to that at the surface since 1979. On the other hand, satellite system data for skin temperatures right at the Earth's surface show trends since 1982 that are very similar to those for the near-surface data collected at the climate stations. Hence, while measurements and analysis errors may be a factor, experts suggest that much of the difference between surface and lower atmospheric trends is real. There are indications that they may be caused by changing atmospheric lapse rates with time, perhaps because of factors such as El Nino Southern Oscillation (ENSO) events, volcanic eruptions, and global warming. Over longer timescales, these differences are expected to average out (Gaffen *et al.*, 2000; Christy *et al.*, 2001; Hegerl and Wallace, 2002; Jin and Dickinson, 2002; Jin and Treadon, 2003; Vinnikov and Grody, 2003).

The radiosonde and satellite data also indicate that the lower stratosphere has been cooling in recent decades. This appears to be related to the combined effects of stratospheric ozone depletion and enhanced greenhouse gas concentrations. The former results in less *in situ* absorption of incoming solar UV radiation, while the latter causes a greater insulation against heat loss from the surface and lower atmosphere to the stratosphere above (Christy *et al.*, 2001).

There is also increasing evidence that all other parts of the global climate system are slowly warming and thus storing up more heat. Snow melt, for example, has been occurring earlier across most of the Northern Hemisphere. Most glaciers and ice sheets in polar and alpine regions have been shrinking, particularly in Alaska and Europe, with many of the small glaciers expected to completely disappear within decades. There is also evidence that at least some of the large ice shelves in Antarctica have been thinning (Arendt *et al.*, 2002; Reichert *et al.*, 2002; Rignot and Thomas, 2002; Thompson *et al.*, 2002; Meier *et al.*, 2003; Shepherd *et al.*, 2003; Stone *et al.*, 2002). Meanwhile, sea ice cover has been retreating dramatically across the Arctic (Comiso, 2002). The rate of heat uptake through these cryospheric melting processes is estimated to be similar to that occurring within the atmosphere. Likewise, borehole temperature measurements of the Earth's lithosphere indicate that this component of the climate system is storing additional heat at similar rates (Beltrami *et al.*, 2002). More dramatically, waters within the upper 3 km of the world's oceans have increased their heat content at rates some 10 times greater than this (Levitus *et al.*, 2001).

Although the above evidence indicates that the world is becoming warmer, the spatial and temporal patterns of these trends are varied and complex. Some regions have warmed much more rapidly than the global average and others much less so, or even cooled. For example, the Antarctic Peninsula has warmed rapidly in recent decades, while other parts of Antarctica have cooled (Turner *et al.*, 2002; Thompson and Solomon, 2002; Doran *et al.*, 2002). Likewise, parts of the Arctic in the northwestern region of the Northern Hemisphere and in Siberia have warmed as much as 3 °C over the past 50 years, while other regions like that for the North Atlantic and North Pacific and the northeastern United States have actually cooled slightly (Przyblak, 2002; Robinson *et al.*, 2002). In general, winter and spring seasons have also warmed more than summer and fall seasons. These complex patterns reflect shifts in global atmospheric circulation patterns that are occurring concurrently with the gradual rise in average temperatures.

While such circulation changes have always occurred as normal climate variability, there are indications that the recent behavior could also be influenced by the general warming of the global climate system (Mortiz *et al.*, 2002; Visbeck *et al.*, 2001).

Precipitation Trends

Available precipitation data records are much less representative of global trends than are those for temperature, since precipitation by its very nature is far more inhomogeneous than temperature. Furthermore, there is scant precipitation data for the world's ocean areas, which represent 70% of its surface. As illustrated in Figure 6 (Folland *et al.*, 2001a), available records indicate a recent increase in annual average precipitation of about 0.5 to 1% per decade over most land areas in the mid- to high-latitudes of the Northern Hemisphere and a somewhat more modest increase over the tropics. There also appears to be a corresponding upward trend in both cloud cover and tropospheric water vapor content over much of the Northern Hemisphere. Relative humidity in the lower troposphere, for example, has increased by about 10% since 1973 over a number of regions, including the United States, Alaska, the Caribbean,

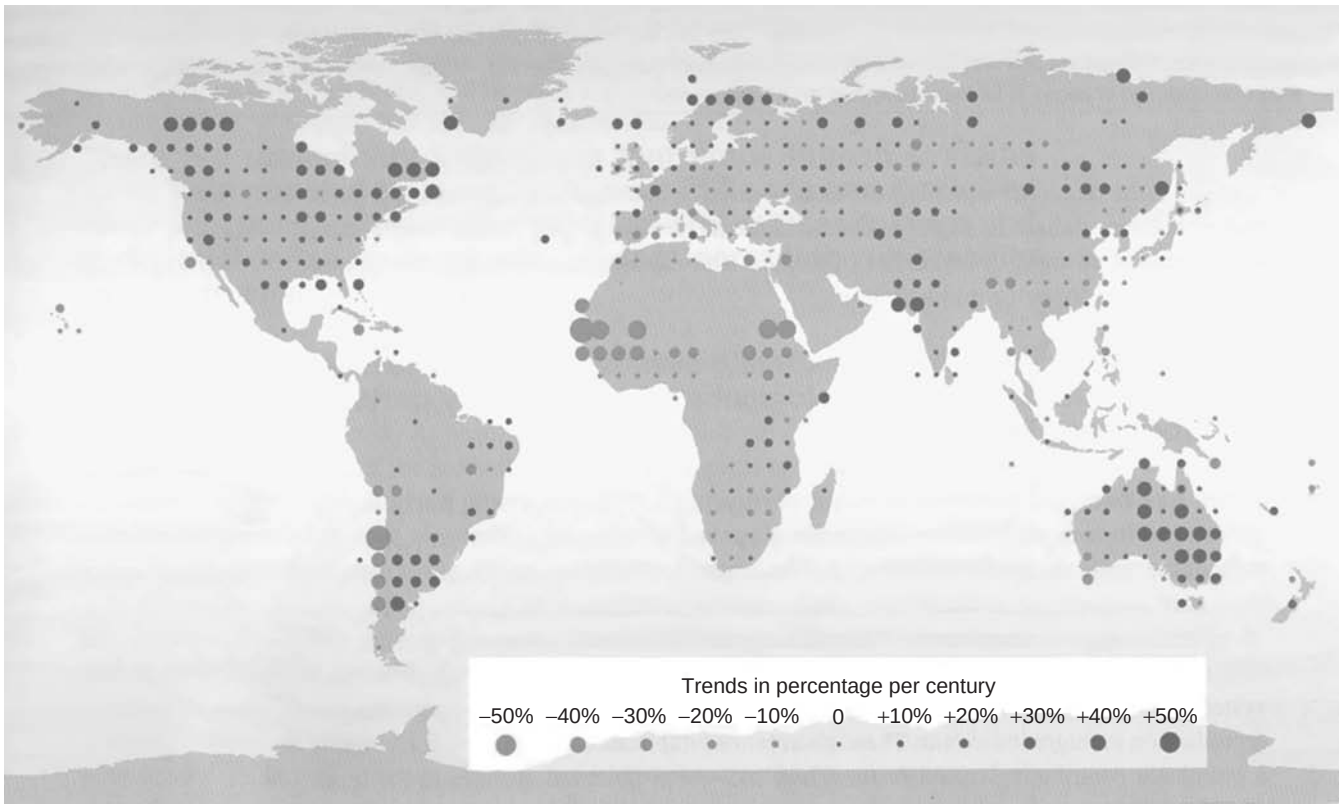


Figure 6 Net change over the twentieth century in annual precipitation amounts in percent estimated for various land regions of the world (IPCC Third Assessment Report 2001. Reproduced by permission of Intergovernmental Panel on Climate Change). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and China. Water content in the comparatively dry stratosphere has also been increasing by about 1% per year over the past four decades. In contrast, there has been a modest decline (about 0.3%/decade) in precipitation over the Northern Hemisphere subtropics. There are no clear indications of precipitation trends in the Southern Hemisphere, although some regions within South America and Africa show decreases. A number of countries have experienced increased number of wet days, and an increased proportion of total precipitation as heavy rain. Most land areas also have increased persistence of wet spells (New *et al.*, 2001; Rosenlof, *et al.*, 2001; Ross and Elliott, 2001; Frich *et al.*, 2002; Milly *et al.*, 2002; Trenberth *et al.*, 2003).

Other Climatological Trends

The melting of global glaciers and ice sheets and the thermal expansion of ocean waters as they warm are both contributing to a rise in average global sea levels. Evidence suggests that sea levels have risen on average by some 1 to 2 cm per decade over the past century. Recent satellite data suggest current rates of increase may have accelerated to about twice that amount (Church *et al.*, 2001; Cabanes *et al.*, 2001). Meanwhile, atmospheric pressures at sea level over the past half century show decreases over the Arctic, Antarctica, and the North Pacific, and increases over the subtropical Atlantic. These observed changes are believed to have caused a northward displacement of dominant storm tracks, thus increasing the frequency and intensity of unusual weather events in Europe in recent decades (Gillett *et al.*, 2003). They are also the likely cause for altered wave action in the North Atlantic, where extreme wave heights have increased in the northeast sector in winter. Significant increases in wave height have also occurred off the Canadian coasts in summer and fall, but other areas and seasons have experienced decreases (Lozano and Swail, 2002; Wang and Swail, 2002).

There is a broad range of indicators to suggest that global ecosystems are already responding to the recent changes in climate discussed above. Recent surveys of related studies indicate that about 80% of changes in behavior of some 1500 to 1700 biological species studied were consistent with that expected due to regional changes in climate. On average, species have shifted their distributions poleward by some 6 km per decade and advanced the onset of spring activities by 2 to 5 days per decade. Tropical ocean corals have also undergone massive bleaching in recent years. If such ecological responses to changes in

climate differ significantly amongst species, this could seriously disrupt their interdependence within ecosystems and effectively tear such communities apart (Parmesan and Yohe, 2003; Root *et al.*, 2003; Walther *et al.*, 2002; Penuelas and Filella, 2001).

Finally, a variety of indicators show significant trends in climate extremes. For example, warm summer nights have become more frequent over the past few decades, particularly in midlatitude and subtropic regions. This has contributed to a reduction in the annual number of frost days and in the intra-annual extreme temperature range. There has also been an increase in some regions in the extreme amount of precipitation derived from wet spells, in the number of heavy rainfall events, and/or in the frequency of drought. Most of the major river basins of the world have experienced a significant shift towards a higher frequency of extreme floods, with three-quarters of extreme events recorded during the past century or so occurring since 1953. Such trends appear to be very unusual, with an estimated 1.3% probability of being entirely due to natural variability. However, they are consistent with expected responses to warmer climates. These results, however, may not apply to more modest flood events and cannot be extrapolated to smaller river basins, where most floods occur (Frich *et al.*, 2002; Milly *et al.*, 2002; Vinnikov *et al.*, 1999). These changes in extreme weather behavior have also caused a global rise in related economic losses. In 2002, for example, record setting floods in Europe and other disasters around the world resulted in economic losses in excess of US$55 billion (Schnur, 2002; Schiermeier, 2003).

FUTURE CLIMATES

Climate Models – the Primary Tool for Projecting Future Climates

As noted in the preceding sections, proxy data and climate observations provide invaluable information on *what* has happened to the Earth's climate in the past and *how* it is behaving today. However, because the climate system is complex and nonlinear in its behavior, these data by themselves are limited in their usefulness to explain *why* the climate system changes with time and in space.

To develop such predictive capability, the complex feedbacks that inextricably couple together the various components of the climate system must be carefully studied with the help of differing types of climate models, ranging from very simple energy budget models to multidimensional physical–chemical models (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**; **Chapter 201, Land-Atmosphere Models for Water and Energy Cycle Studies, Volume 5**). Results of such studies must then be used to approximate these feedbacks as complex mathematical equations that capture the fundamental physical, chemical, and biological processes involved. These can then be integrated within a complex model of the entire dynamic global system. Such climate system models have been under development for some four decades, and are now sufficiently advanced to include a dynamic circulating atmosphere coupled to a circulating global ocean system with a responsive ice cover. Some have now also coupled the dynamics of global ecosystems within the system, thus allowing them to replicate the biogeochemical feedbacks between the climate and biosphere systems. These models, generally referred to as *coupled climate models*, provide intricate approximations of how the Earth's biogeochemical system functions and require the largest and most advanced supercomputers available to run. However, they continue to be limited by computing capacity, by inadequate observational data for describing, calibrating, and verifying the physical and chemical processes included in the models, and by gaps in scientific knowledge of feedbacks within the climate system. Despite these limitations, they have become valuable tools for adding to the scientific understanding of why climate changes, and how it may change in the future.

There is now a large suite of coupled models available internationally to undertake climate change experiments. One of the key performance indicators of these models is model *climate sensitivity* – that is, the change in global temperature simulated by a model in response to a specified imbalance in net radiative flux at the top of the atmosphere. This is usually described as the amount of warming predicted by the model as a result of an equilibrium response to a doubling of carbon dioxide concentrations. Since the models available today vary considerably in how they deal with various climate processes and feedbacks, they also vary considerably in their climate sensitivity. However, in general, all advanced models have a climate sensitivity somewhere between a 1.5 and 4.5 °C equilibrium response to a doubling of atmospheric CO_2 concentrations (Cubasch *et al.*, 2001).

Human Interference with the Climate System

The preceding sections have described how natural variations internal to the climate system as well as external natural factors (primarily solar and volcanic) have affected past climates. These factors will continue to modify future global climates. However, the range of such variations and naturally induced changes during the next century are unlikely to exceed that observed over the past few millennia.

There is, however, another major factor at work on the climate system that is rapidly increasing in magnitude and is expected to dominate over all other natural forces and variations likely to occur over the next century and beyond – that is, human activities. The following paragraphs describe two prominent aspects of the climate system that humans have altered in the past, and how these are likely to change in

the future (*see* **Chapter 33, Human Impacts on Weather and Climate, Volume 1**).

Human Influences on Regional Surface Albedo and Hydrology

Humans have been significantly transforming the Earth's regional landscape ever since the onset of civilization in Asia and Africa some 8000 years ago (Ruddiman, 2003). Over the millennia, they have changed vast areas of forested lands into agricultural fields (and, in some places, back again), dry lands into wetlands or wetlands into dry, and rural landscapes into city environments. Such changes in land use have altered the local albedo of the Earth's surface and hence have influenced how much sunlight is reflected back to space. In some cases, these changes have reduced surface reflection of incoming sunlight and caused a warming influence. In others, particularly in regions covered with winter snow, they have increased surface reflection and thus caused a local cooling. For example, the albedo effects of change from forest to crops landscapes may have caused a net global cooling effect of between 0.09 and 0.22 °C during the past three centuries (Matthews *et al.*, 2003).

In addition to changing albedos, land use change also affects regional evaporation, evapotranspiration, rainfall, atmospheric circulation, and cloud cover. Some experts argue that the net accumulated effect of land use change over the millennia may have been a significant factor in changing climate on a global scale as well. However, the regional and seasonal complexities of this forcing factor are poorly understood and hence its importance in helping to explain past changes in climate is still difficult to quantify (Chase *et al.*, 2001; Zhao and Pitman, 2002; Heck *et al.*, 2001; Matthews *et al.*, 2003).

Future land cover change and hence its effect on surface albedo and hydrology will largely be determined by a range of demographic and climate factors. Although the demographic factors involved are difficult to predict, experts suggest that changes in land use over the next century have finite limits that range between a worst-case scenario of a continued deforestation until the world's virgin forests have been depleted and a best-case scenario of gradually increasing global forest cover from both a decline in deforestation activities and increased reforestation efforts (Watson *et al.*, 2000; Nakicenovic and Swart, 2000).

The effect of future climate change on land cover can vary from enhanced forest cover in polar regions as treelines respond to warmer climates to increased desertification or conversion of forests to grasslands through natural successional processes in others. Advanced climate models are now beginning to include such vegetation feedbacks within their simulations, and hence related albedo and hydrological feedbacks become an inherent part of these simulations (Bergengren *et al.*, 2001; Betts, 2000; Cox *et al.*, 2000; Douville *et al.*, 2000, Foley *et al.*, 2000; Harding *et al.*, 2002).

Changing Atmospheric Composition

The other major way in which humans are interfering with the climate system is through emissions into the atmosphere of greenhouse gases and aerosols that gradually change its composition, and hence its role in regulating the flow of energy into and out of the Earth's climate system. For example:

- Over the past 150 years, humans have cumulatively emitted approximately 1500 billion tons of carbon dioxide into the atmosphere. About two-thirds of these emissions were caused by the combustion of fossil fuels for energy and the remainder by deforestation. Emissions from the latter have been relatively stable in recent decades (at about 6 billion tons of CO_2 per year). However, those from fossil fuel use continue to rise quite rapidly, increasing from an average of about 11 billion tons per year in the 1960s to about 23 billion tons per year during the 1990s. Fortunately, natural processes are removing a significant fraction of these human emissions from the atmosphere through enhanced absorption in surface oceans and increased biological growth on land. However, the amount of CO_2 in the atmosphere still increased at the rate of some 12 billion tons per year during the 1990s. Atmosphere concentrations, which were at a remarkably stable level of about 260 to 280 parts per million by volume (ppmv) throughout the Holocene, have increased to levels of about 374 ppmv by 2002. This is an increase over preindustrial levels of about 33%. There are indications, in fact, that current levels may be unprecedented in the past 20 million years (Carbon Dioxide Information and Analysis Center, 2003; Climate Monitoring and Diagnostic Laboratory, 2002; Prentice *et al.*, 2001).
- Human activities have contributed to dramatic increases in other greenhouse gases as well. Atmospheric methane concentrations have more than doubled over the past century, while those for nitrous oxide have increased by about 15%. Tropospheric ozone has also increased substantially over much of the industrialized world, and entirely new and powerful greenhouse gases such as halocarbons and sulphur hexafluoride are now being added in significant amounts. In contrast, halocarbons have contributed to a decrease in ozone within the stratosphere.
- Finally, there has also been a progressive increase in the anthropogenic emissions of aerosols and their precursors into the atmosphere. While most of these aerosols have relatively short lifetimes within the atmosphere of days to weeks, their continuous production in industrialized regions of the world has resulted in a large and sustained increase in their concentrations over and

down-wind of these regions. From a climate perspective, these aerosols have several important roles. First, they directly affect the amount of incoming sunlight that is reflected back to space or absorbed within the atmosphere. Secondly, fine aerosols also function as condensation nuclei and hence alter the amount and properties of cloud. They thus indirectly affect the absorption and reflection of incoming radiation through the role of these clouds. Finally, as these aerosols settle out of the atmosphere onto the Earth's surface, they can affect the surface albedo. This is particularly true for soot on snow or ice. While there are large uncertainties associated with these effects, experts suggest that the net direct effect of increased concentrations of all aerosols over the past century is likely negative (on the order of $-0.5\,W\,m^{-2}$), hence offsetting some of the warming effects of rising greenhouse gas concentrations. Their indirect effects through altered cloud conditions may be larger, but are even more uncertain.

Radiative studies indicate that the observed increase in greenhouse gas concentrations during the past century have enhanced net average surface energy flux by about $2.5\,W\,m^{-2}$. This has been partially offset by a net cooling by increased aerosol concentrations (direct and indirect), although the magnitude of this offset is poorly understood. Hence, the net radiative forcing from all human causes may be somewhere in the range of 1 to $2\,W\,m^{-2}$. Simulations of past climate response to both these human forcings and those estimated for natural factors (see above) have quite successfully simulated the observed climates of the past 100 years. Furthermore, these model results, together with the paleo evidence for the range of natural climate variations over the past 2000 years, have led researchers to conclude that "most of the observed warming over the past 50 years is attributable to human activities" (Ramaswamy *et al.*, 2001; Mitchell *et al.*, 2001).

As with surface albedo and hydrological processes, projections of future emissions of greenhouse gases and aerosols are highly sensitive to both demographic behavior and climate feedbacks. After carefully considering the range of plausible futures for the various socioeconomic, technological, and biological factors involved, experts have suggested that, unless deliberate measures are taken to curtail future emissions, another 3600 to 8000 billion tons of carbon dioxide will be emitted into the atmosphere over the next century as a result of fossil fuel combustion and land use change. Allowing for terrestrial and ocean removal processes, they project that this will almost certainly result in a doubling of preindustrial atmospheric CO_2 concentrations by 2100, and a tripling could be possible. Over the same time period, projected concentrations of other greenhouse gases also range from modest to significant increases. In contrast, aerosol emissions are projected to eventually decline relative to current levels because of expected measures to curtail their emissions for local environmental reasons. The net climatic effect of these projected changes in both greenhouse gases and aerosols is an increase in net global radiative forcing by 2100 relative to preindustrial levels of between about 4 and $9\,W\,m^{-2}$ (Nakicenovic and Swart, 2000; Cubasch *et al.*, 2001).

Projected Climate Change for the Next Century

Temperature

Climate models are the primary tools used to project how future changes in radiative forcings might affect the climates of the next century. In its Third Assessment Report, the IPCC uses these models to project that average surface temperatures will warm by some 1 to 2.5 °C by 2050, increasing to between 1.4 and 5.8 °C by 2100, relative to current climates (Houghton *et al.*, 2001). Delayed effects from changes in radiative forcings to date already commit the world to 0.5 °C of that warming, even if all emissions of greenhouse gases were to stop immediately (Hansen *et al.*, 2002). Some researchers have attempted to put some probabilities on these projections. In general, they suggest only a 5% probability that global climates will warm by less than 1.7 °C by 2100, or by more than 4.9 °C (Reilly *et al.*, 2001; Webster *et al.*, 2003). As illustrated in Figure 7 (Cubasch *et al.*, 2001), at the lower end of this range, changes will be unprecedented in the history of human civilization, while at the upper end they are comparable to the magnitude of change during the last deglaciation, but at about 50 times the rate.

While the above results indicate that there is considerable agreement between models on the significance of global scale changes in temperature, there is much less agreement with respect to regional changes. However, despite these differences, there are a number of common features amongst all of the models (Cubasch *et al.*, 2001). For example:

- Land areas warm more than ocean surfaces. This is simply a consequence of the thermal inertia of oceans. This ocean inertia is also the primary reason for a significant delay of decades in achieving a new equilibrium climate in response to an enhanced forcing regime.
- The Arctic polar region warms more than the tropics. The primary reason for this polar amplification is the positive feedback provided by snow and ice. Although such amplification is also expected to eventually occur in the Antarctic region, models predict a delayed response of that region's cryosphere.
- Nighttime temperatures will, on average, warm more than daytime temperatures, thus reducing the daily temperature range.

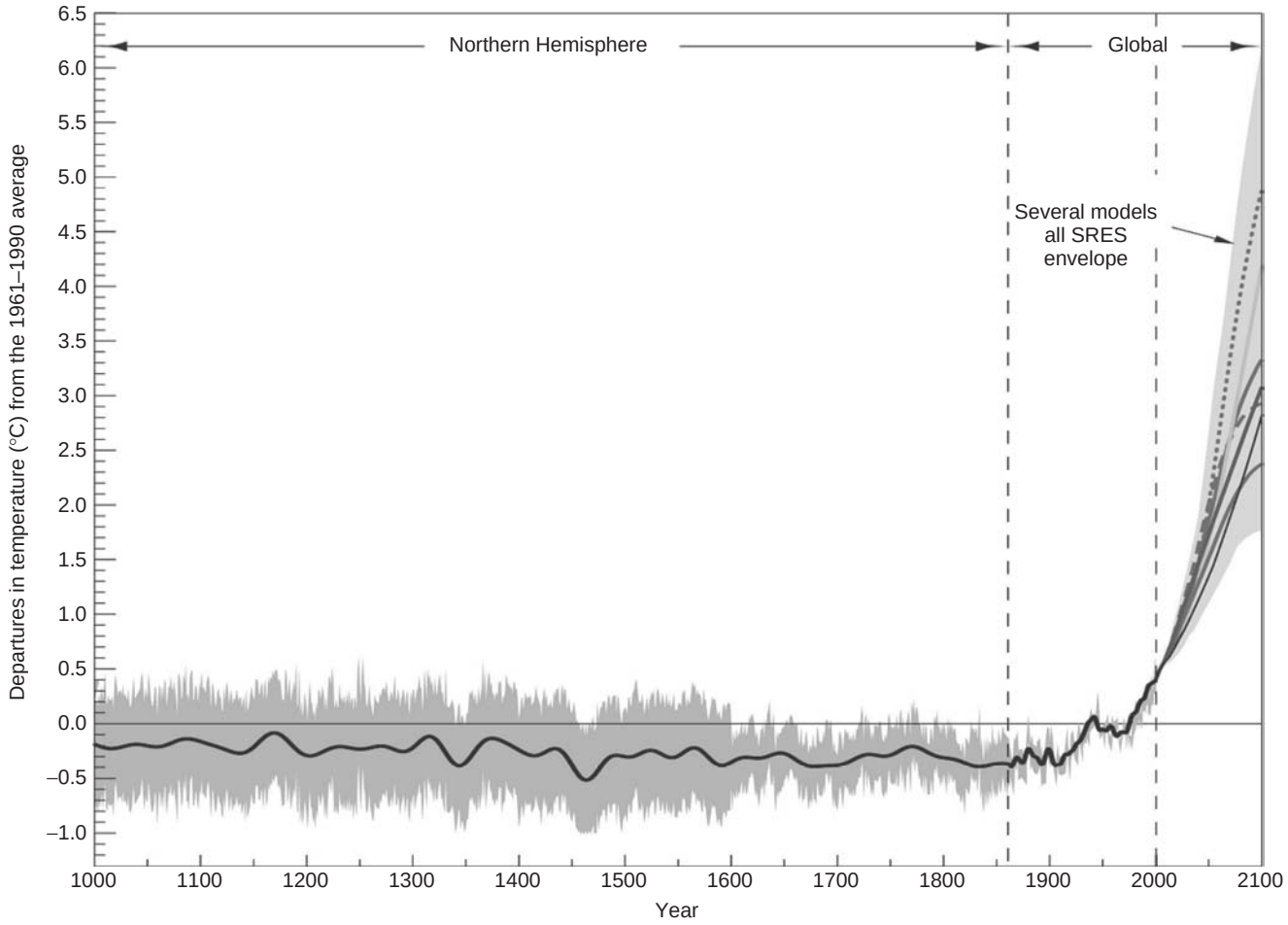


Figure 7 Projected changes in average global temperatures to 2100, in comparison with trends of the past millennium. Individual colored lines represent projections for six representative emission scenarios using a model of intermediate climate sensitivity, while the broader gray shading between 2000 and 2100 is indicative added uncertainty due to varying model sensitivities. Gray shading in the proxy data indicates estimates uncertainty in the data (IPCC Third Assessment Report 2001. Reproduced by permission of Intergovernmental Panel on Climate Change). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

- Ocean circulation is expected to slow down. The turnover of the global oceans is largely determined by thermohaline (i.e., temperature and salinity) processes that affect surface water densities. All models agree that enhanced precipitation in the high latitudes of the Northern Hemisphere is likely to decrease the rate of deep water formation in the North Atlantic and hence weaken the thermohaline circulation system. This may cause some surface ocean regions, like areas of the North Atlantic, to actually cool while the rest of the world warms.
- Natural oscillations in the climate system will be superimposed upon the projected upward trends in temperatures, and hence will modulate both the temporal and spatial response of climates to enhanced radiative forcing. This adds significantly to the uncertainty of the model projections, particularly at the regional scale.
- There is also increasing evidence that the pattern of future warming will increasingly be like that of current El Nino years, with enhanced warming in the central and eastern tropical Pacific relative to the western Pacific. This, in turn, causes global atmospheric circulation patterns to change (Boer *et al.*, 2004).

Precipitation

Warmer surface temperatures will increase surface evaporation and hence enhance the global hydrological cycle (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**). Various model projections suggest global precipitation response to projected warming by mid-century could be anywhere between a small decrease of −0.2% to an increase of 5.6% (Yang *et al.*, 2003). However, in general, the global hydrological cycle (and hence precipitation) is likely to be enhanced by about 1 to 2%

per degree Celsius of warming. The atmospheric holding capacity for water vapor should also increase by about 7% per degree Celsius of warming. This means increased surface water loss through evaporation processes will likely exceed increases in precipitation. It also suggests a greater average increase in atmospheric moisture content per degree of warming for low latitudes than for higher latitudes. Since the additional release of latent heat when water vapor condenses within the atmosphere will further invigorate storms, there will be a greater increase in rate of heavy precipitation than for average precipitation. That is, precipitation events (including snow storms) are likely to become fewer, but more intense. The fraction of precipitation falling as snow will decrease, as will the duration of the snow season. However, concurrent shifts in atmospheric circulation and hence precipitation patterns suggest that these changes will vary significantly from region to region and season to season. In general, the El Nino–like pattern indicated in many of the models suggests an eastward shift of precipitation in the tropical Pacific. Various models also suggest an enhanced positive Arctic Oscillation under warmer climates. Most high-latitude areas are expected to experience significant increases in precipitation, while much of the subtropical regions are likely to experience decreases in mean rainfall. In the tropics, projected changes are smaller and less consistent between models. On average, the intensity of rainfall increases. Most models project an increase in precipitation variability where mean values increase, and an increase in droughtiness of mid-continental regions in summer. Several simulations indicate that, by the time atmospheric CO_2 concentrations double relative to preindustrial levels, the Saharan desert will have shifted northwards by about 0.55° of latitude and will thus have become significantly hotter and drier (Räisänen, 2002; Douville *et al.*, 2000; Gillett *et al.*, 2002; Liu *et al.*, 2002; Wilby and Wigley, 2002; Boer *et al.*, 2004; Trenberth *et al.*, 2003).

Sea, River, and Lake Ice

In the Arctic, sea ice is projected to decrease in extent and thickness, particularly in summer. By 2100, most of the Arctic will likely be entirely free of ice in late summer. While winter ice cover will continue to be extensive, it will consist of thinner and more brittle first-year ice, form much later in autumn, and disappear earlier in the spring. Reduced ice cover will, in general, make Arctic shipping much easier, but could be a major threat for some of the important Arctic species that rely on ice for their habitat. It will also present challenges for the people in the region who rely on these species. Furthermore, reduced sea ice in the margins of the Arctic Ocean will expose coastlines to accelerated erosion caused by an intensification of the ocean wave regimes. The low-lying coastlines of the delta regions of the great Arctic rivers are particularly vulnerable (Kerr, 2002; Armstrong and Brodzik, 2001; Gregory *et al.*, 2002; Falkingham *et al.*, 2003). In Antarctic, where sea ice already disappears in summer today, the changes in ice cover during the next century are expected to be less dramatic. This is primarily because of projected changes in Southern Ocean circulation patterns that are expected to delay ocean surface warming relative to oceans surfaces elsewhere. (Anisimov *et al.*, 2001; Cubasch *et al.*, 2001).

The duration of ice cover and ice thickness on inland lakes and rivers are also expected to decrease substantially. Experts suggest, for example, that ice growth seasons on lakes in the northern United States, could be reduced by up to 100 days, and mean ice thickness by as much as 40 cm (Fang and Stefan, 1998).

Land Ice and Sea Levels

Most temperate glaciers are expected to decline significantly, with many disappearing entirely within the next century. The Greenland ice sheet is also expected to decline slowly. However, in Antarctica, a more maritime type climate may sufficiently enhance snowfall onto the interior of the ice sheets to more than offset enhanced melting at its margins, thus causing a slow increase in net ice volume. Collectively, these changes in land ice volume are expected to add to ocean volume and cause sea levels to rise. Thermal expansion of ocean waters as they warm will add to this rise, resulting in a net sea level rise of somewhere between 9 cm and 88 cm by 2100 and continuing for centuries thereafter (Van Lipzig *et al.*, 2002; Vaughan and Spouge, 2002; Church *et al.*, 2001).

Meanwhile, the permafrost that underlies some two-thirds of the Arctic land mass is expected to undergo extensive thawing, increasing the thickness of the active layer overtop continuous permafrost zones by 30 to 40% in most regions by 2100. Since snow acts as an insulator that slows down the penetration of winter cold temperatures into the ground, any decrease in snow cover under warmer climates could help offset some of permafrost loss. Much of the permafrost would completely disappear in discontinuous zones. These changes would cause significant ground settlement in most affected regions, and catastrophic slope movement in some locations. While much of the melting ground ice may evaporate, it will also have pronounced effects on the permeability of soils, the melting of surface ice and the seasonality of stream flow. These changes, in turn, will affect ecosystems. There is also a risk that permafrost degradation could cause the release of large amounts of methane gas from natural gas hydrates, perhaps explosively (Stieglitz *et al.*, 2003; Nelson *et al.*, 2002; Stendel and Christensen, 2002; Hecht, 2002).

Severe Weather

Much of the potential risks of danger associated with warmer climates relate to the frequency of severe weather events that exceed the tolerance levels of ecosystems and/or socioeconomic systems. These events are often regional and local in scale and by definition occur infrequently. Hence, they are much more difficult to simulate than larger-scale climate phenomena. However, various types of climate studies using model outputs have helped provide some useful clues as to how the behavior of these events might change under warmer climates (Timmermann, 2001; Herbert and Dixon, 2003; Palmer and Raisanen, 2002; Milly *et al.*, 2002; Huntingford *et al.*, 2003). For example:

- Temperature extremes. An increase in average global temperatures will very likely cause a disproportionately larger increase in the frequency and intensity of extreme hot days and a decrease in the probability of extreme cold days. Because of the added discomfort of higher humidity, periods of intense heat stress become even more frequent.
- Precipitation extremes increase more than their means, and intense rainfall events become almost twice as frequent in many regions. Length of wet and dry spells both increase. Extreme rainfall events over the United Kingdom that now only occur, on average, about once every 20 years could occur once every 3 to 5 years within the next century. Over central and northern Europe, winter precipitation extremes that now occur about once in 40 years may increase to once in 8 years by the time of a atmospheric CO_2 concentrations double preindustrial levels. A similar increase occurs for summer monsoonal rainfall extremes over much of southern Asia. However, such events are projected to become less frequent over the Mediterranean and northern Africa.
- While there is no general agreement amongst models, some studies suggest that the total number of storms in midlatitudes is likely to decrease, but that the number of intense storms will rise. Dominant storm tracks may also be displaced poleward.
- Intense El Nino and La Nina events, which can be very disruptive to normal weather patterns, may become more frequent in the future.
- While it is uncertain whether the frequency of these storms will rise, warmer ocean surfaces imply that the potential upper threshold for storm intensity will increase.
- Most large river basins will experience a decrease in average flows, but an increase in extreme flows and related flood risks. These changes in precipitation extremes, together with changes in timing of snowmelt, will have significant impacts on river flows and flood risks. In one study, 6 of 14 large river basins assessed are projected to experience a higher probability of floods, while two show a decrease. In this region, poor farmers and consumers are expected to bear most of the related costs. Figure 8 (Milly *et al.*, 2002) shows results from another study that projects,

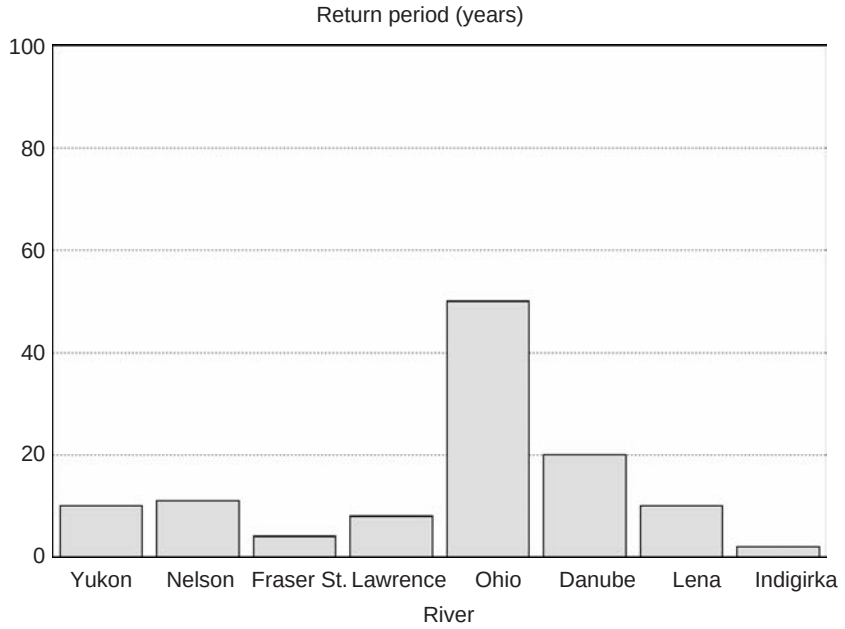


Figure 8 Return periods for the current 100-year flood for selected North American and Russian River basins under a $4 \times CO_2$ scenario. The control bar is the current 100-year flood event

for a quadrupled CO_2 climate, that some extratropical river basins, particularly in Russia, could see current one in 100-year flood events increase in frequency to once every 2 to 3 years. However, most high-latitude rivers will have lower severe spring flood risks because of less snowmelt.

Risks of Large-scale Abrupt Changes in Climate

In general, climate models suggest that global climates will evolve in a continuous transient response to enhanced radiative forcings. However, there is considerable evidence from paleoclimatological data that the Earth's climate can change abruptly from one relatively stable mode to another. Such abrupt changes would be far more catastrophic in nature, since the changes would be large and very rapid, and hence the potential for progressive adaptation of ecosystems and socioeconomic infrastructures would be minimal. Experts have identified three possible processes by which enhanced radiative forcing could cause such abrupt discontinuities in climates over the next decade (Alley *et al.*, 2003; Vaughan and Spouge, 2002; Smith *et al.*, 2001).

The first is the risk of a complete shutdown of the global ocean thermohaline circulation. Since this circulation system is the driver for the Gulf Stream that keeps western Europe some 10 °C warmer than it would be in its absence, its cessation could cause a calamitous cooling in that region even as the rest of the world warms. It would also dramatically alter the global atmospheric circulation patterns and hence the distribution of rainfall.

The second is the possible collapse of the West Antarctic Ice Sheet. This ice sheet is considered to be unstable and its collapse through glacial surging (rather than *in situ* melting) could be triggered within the next century. That ice sheet stores enough water to raise global sea levels by 5 to 6 m. A panel of experts has noted that there is a 5% risk of a collapse of the West Antarctic Ice Sheet within 200 years (which would cause sea levels to rise by 1 m/century) and a 30% chance that rapid decay of the ice sheet would add 20 cm per decade to sea levels.

The third risk relates to the potential release of large amounts of methane contained within the solid hydrates found in the Arctic permafrost zone. Paleo studies indicate that such large releases from hydrates below the ocean bottom likely occurred during an abrupt climate warming some 55 million years ago, and could happen again. Since methane is a potent greenhouse gas, such releases could abruptly launch the Earth into a super-greenhouse effect era.

While the impacts of each of these abrupt discontinuities would be globally catastrophic, experts considered the risks of such events within the next century to be quite small.

Concluding Remarks

While there continue to be significant uncertainties with respect to attributing past climate change to specific causes and in projecting the rate and characteristics of future climate change, there are a number of fundamental conclusions about climate change that now have both wide acceptance within the science research community and significant implications for global ecosystems and society. These include the following:

- Global climate is constantly changing in response to a range of external and internal forces acting upon the climate system.
- The magnitude and rate of average global surface warming during the last 50 years appear to be unprecedented in at least the past two millennia.
- Other aspects of the global climate system have changed significantly in recent decades.
- The balance of evidence indicates that these changes are likely due to human interference with the climate system.
- Projections for future warming suggest that changes in global temperatures by 2100 will be unprecedented in human history.
- Warmer climates will alter many aspects of local and regional weather and thus have major impacts on ecosystems, hydrology, and human infrastructures.

Acknowledgments

In preparing this overview of climate and climate change science, the author has relied heavily on the assessments provided by the Intergovernmental Panel on Climate Change, particularly the Third Assessment completed in 2001. That assessment can be accessed on line at **www.ipcc.ch.** The author also wishes to acknowledge the valuable input and review comments provided by Elizabeth Bush, Patti Edwards, Dr Venkata Neralla, and Bob Whitewood.

REFERENCES

Adams J.M. (1997) *Global Land Environments Since the Last Interglacial*, Oak Ridge National Laboratory: **http://www.esd.ornl.gov/ern/qen/nerc.html**.

Alley R.B., Marotzke J., Nordhaus W.D., Overpeck T.J., Peteet D.M., Pielke R.A., Pierrehumbert R.T., Rhines P.B., Stocker T.F., Talley L.D. Jr, *et al.* (2003) Abrupt Climate Change. *Science*, **299**, 2005–2010.

Anderson D.M., Overpeck J.T. and Gupta A.K. (2002) Increase in the Asian Southwest Monsoon during the past four centuries. *Science*, **297**, 596–599.

Anisimov O., Fitzharris B., Hagen J.O., Jeffries R., Marchant H., Nelson F., Prowse T. and Vaughan D.G. (2001) Polar Regions

(Arctic and Antarctic). In *Intergovernmental Panel on Climate Change 2001 Third Assessment Report of the intergovernmental Panel on Climate Change, WG II. Climate Change 2001: Impacts, Adaptation and Vulnerability*, Chap. 16, McCarthy J.J., Canziani O.F., Leary N.A., Dokken D.J. and White K.S. (Eds.), Cambridge University Press: Cambridge, pp. 801–841.

Arendt A.A., Echelmeyer K.A., Harrison W.D., Lingle C.S. and Valentine V.B. (2002) Rapid Wastage of Alaska Glaciers and Their Contribution to Rising Sea Level. *Science*, **297**, 382–386.

Armstrong R.L. and Brodzik M.J. (2001) Recent Northern Hemisphere snow extent: a comparison of data derived from visible and Microwave Satellite Sensors. *Geophysical Research Letters*, **28**, 3673–3676.

Baede A.P.D., Ahlonsou E., Ding Y. and Schimel D. (2001) The climate system: an overview. In *Climate Change 2001: The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Chap. 1, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskall K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge, pp. 85–98.

Beltrami H. (2002) Earth's long-term memory. *Science*, **297**, 206–207.

Beltrami H., Smerdon J.E., Pollack H.N. and Huang S. (2002) Continental heat gain in the global climate system. *Geophysical Research Letters*, **29**, doi:10.1029/2001GL014310.

Bergengren J.C., Thompson S.L., Pollard D. and Deconto R.M. (2001) Modeling global climate-vegetation interactions in a doubled CO_2 world. *Climatic Change*, **50**(1–2), 31–75.

Berger A. and Loutre M.F. (2002) An Exceptionally Long Interglacial Ahead? *Science*, **297**, 1287–1288.

Betts R.A. (2000) Offset of the potential carbon sink from boreal forestation by decreases in surface Albedo. *Nature*, **408**, 187–190.

Boer G.J., Yu B., Kim S.-J. and Flato G.M. (2004) Is There Observational Support for an El Nino-Like Pattern of Future Global Warming? *Geophysical Research Letters*, **31**, L06201, doi:10.1029/2003GL018722.

Briffa K.R. and Osborn T.J. (2002) Blowing hot and cold. *Science*, **295**, 2227–2228.

Bush A.B. and Philander S.G.H. (1999) The climate of the last glacial maximum: results from a coupled atmosphere-ocean general circulation model. *Journal of Geophysical Research*, **104**, 24509–24525.

Cabanes C., Cazenave A. and Le Provost C. (2001) Sea level rise during the last 40 years determined from satellite and *in situ* observations. *Science*, **294**, 840–842.

Carbon Dioxide Information and Analysis Center, Oak Ridge National Laboratory (2003) Oakland: **http://cdiac.esd.ornl.gov/trends/emis/em_cont.htm**.

Chase T.N., Pielke R.A. Sr, Kittel T.G.F., Zhao M., Pitman A.J., Running S.W. and Nemani R.R. (2001) The relative climatic effects of landcover change and elevated carbon dioxide combined with aerosols: a comparison of model results and observations. *Journal of Geophysical Research*, **106**, 31685–31691.

Christy J.R., Parker D.E., Brown S.J., Macadam I., Stendel M. and Norris W.B. (2001) Differential trends in tropical sea surface and atmospheric temperatures since 1979. *Geophysical Research Letters*, **28**, 183–186.

Church J.A., Gregory J.M., Huybrechts P., Kuhn M., Lambeck K., Nhuan M.T., Qin D., Woodworth P.L., Anisimov O.A., Bryan F.O., *et al.* (2001) Changes in sea level, in *Climate Change 2001: The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Ding Y., Griggs D.J., Noguer N., van der Linden P.J., Dai X., Maskall K. and Johnson C.A. (Eds.), Cambridge University Press, Cambridge, pp. 639–693.

Climate Monitoring and Diagnostic Laboratory (2002) *Carbon Cycle*, *CMDL Summary Report #26*, NOAA, Boulder: **http://www.cmdl.noaa.gov/publications/annrpt26/index.html**.

Comiso J.C. (2002) A rapidly declining perennial sea ice cover in the Arctic. *Geophysical Research Letters*, **29**, 1956, doi:10.1029/2002GL015650.

Cox P.M., Betts R.A., Jones C.D., Spall S.A. and Totterdell I.J. (2000) Acceleration of global warming due to carbon cycle feedbacks in a coupled climate model. *Nature*, **408**, 184–187.

Crowley T.J. (2002) Cycles, cycles everywhere. *Science*, **295**, 1473–1474.

Cubasch U., Meehl G.A., Boer G.J., Stouffer R.J., Dix M., Noda A., Senior C.A., Raper S., Yap K.S., Ouchi A.A., *et al.* (2001) Projections of future climate change. In *Climate Change 2001: The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Ding Y., Griggs D.J., Noguer N., van der Linden P.J., Dai X., Maskall K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge, pp. 525–582.

Doran P.T., Priscu J.C. and Lyons W.B. (2002) Antarctic climate cooling and terrestrial ecosystem response. *Nature*, **415**, 517–520.

Douville H., Planton S., Royer J.F., Stephenson D.B., Tyteca S., Kergoat L., Lafont S. and Betts R.A. (2000) Importance of vegetation feedbacks in doubled-CO_2 climate experiments. *Journal of Geophysical Research*, **105**, 14841–14861.

Falkingham J., Melling H. and Wilson K. (2003) Shipping in the Canadian arctic: possible climate change scenarios. *CMOS Bulletin*, **31**, 68–69.

Fang X. and Stefan H.G. (1998) Potential climate warming effects on ice covers of small lakes in the contiguous U.S. *Cold Region Science and Technology*, **27**, 119–140.

Foley J.A., Levis S., Costa M.H., Cramer W. and Pollard D. (2000) Incorporating dynamic vegetation cover within global climate models. *Ecological Applications*, **10**, 1620–1630.

Folland C.K., Karl T.R., Christy J.R., Clarke R.A., Gruza G.V., Jouzel J., Mann M.E., Oerlemans J., Salinger M.J., Wang S.-W., *et al.* (2001a) Observed climate variability and change. In *Climate Change 2001: The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Cubasch U., Meehl G.A., Boer G.J., Stouffer R.J., Dix M., Noda A., Senior C.A., Raper S., Yap K.S., Ouchi A.A., *et al.* (Eds.), Cambridge University Press: Cambridge, pp. 99–181.

Folland C.K., Rayner N.A., Brown S.J., Smith T.M., Shen S.S.P., Parker D.E., Macadam I., Jones P.D., Jones R.N., Nicholls N.,

et al. (2001b) Global temperature change and its uncertainties since 1861. *Geophysical Research Letters*, **28**, 2621–2624.

Frich P., Alexander L.V., Della-Marta P., Gleason B., Haylock M., Klein Tank A.M.G. and Peterson T. (2002) Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, **19**, 193–212.

Gaffen D.J., Sargent M.A., Habermann R.E. and Lanzante J.R. (2000) Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality. *Journal of Climate*, **13**, 1776–1795.

Gajewski K., Vance R., Sawada M., Fung I., Gignac L.D., Halsey L., John J., Maisongrande P., Mandell P., Mudie P., *et al.* (2000) The climate of North America and adjacent ocean waters *ca* 6 ka. *Canadian Journal of Earth Sciences*, **37**, 661–681.

Gillett N.P., Zwiers F.W., Weaver A.J., Hegerl G.C., Allen M.R. and Stott P.J. (2002) Detecting anthropogenic influence with a multi-model ensemble. *GRL*, **29**(20), 1970, doi:10.1029/2002GL015836.

Gillett N.P., Zwiers F.W., Weaver A.J. and Stott P.A. (2003) Detection of human influence on sea-level pressure. *Nature*, **422**, 292–294.

Gregory J.M., Stott P.A., Cresswell D.J., Rayner N.A., Gordon C. and Sexton D.M.H. (2002) Recent and future changes in arctic sea ice simulated by the HadCM3 AOGCM. *Geophysical Research Letters*, **29**, 2175, doi:10.1029/2001GL014575.

Hansen J., Sato M., Nazarenko L., Ruedy R., Lacis A.A., Koch D., Tegen I., Hall T., Shindell D., Santer B., *et al.* (2002) Climate forcings in Goddard Institute for space studies SI2000 simulations. *Journal of Geophysical Research*, **107**, 4347–4384.

Harding R., Kuhry P., Christensen T.R., Sykes M.T., Dankers R. and van der Linden S. (2002) Climate Feedbacks at the Tundra-Taiga Interface. *Ambio Special Report*, **12**, 47–55.

Hecht J. (2002) Earth's ancient heat wave gives a taste of things to come. *New Scientist*, **173**(2372), 21.

Heck P., Luthi D., Wernli H. and Schar C. (2001) Climate impacts of European-scale anthropogenic vegetation changes: a sensitivity study using a regional climate model. *Journal of Geophysical Research*, **106**, 7817–7835.

Hegerl G.C. and Wallace J.M. (2002) Influence of patterns of climate variability on the difference between satellite and surface temperature trends. *Journal of Climate*, **15**, 2412–2428.

Herbert J.M. and Dixon R.W. (2003) Is the ENSO phenomenon changing as a result of global warming? *Physical Geography*, **23**, 196–211.

Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.) (2001) *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, p. 881.

Huntingford C., Jones R.G., Prudhomme C., Lamb R., Gash J.H.C. and Jones D.A. (2003) Regional climate-model predictions of extreme rainfall for a changing climate. *Quarterly Journal of the Royal Meteorological Society*, **129**, 1607–1621.

Jin M. and Dickinson R.E. (2002) New observational evidence for global warming from satellite. *Geophysical Research Letters*, **29**, doi:10.1029/2001GL013833.

Jin M. and Treadon R.E. (2003) Correcting the Orbit Drift on AVHRR land surface skin temperature measurements. *International Journal of Remote Sensing*, **24**, 4543–4558.

Jones P.D. and Moberg A. (2003) Hemispheric and large-scale surface air temperature variations: an extensive revision and an update to 2001. *Journal of Climate*, **16**, 206–223.

Kerr R.A. (2002) A warmer arctic means change for all. *Science*, **297**, 1490–1492.

Levitus S., Antonov J.I., Wang J., Delworth T.L., Dixon K.W. and Broccoli A.J. (2001) Anthropogenic warming of earth's climate system. *Science*, **292**, 267–270.

Liu P., Meehl G.A. and Wu G. (2002) Multi-model trends in the Sahara induced by increasing CO_2. *Geophysical Research Letters*, **29**(18), 1881, doi:10.1029/2002GL015923.

Lozano I. and Swail V. (2002) The link between wave height variability in the North Atlantic and the storm track activity in the last four decades. *Atmosphere-Ocean*, **40**, 377–388.

Mann M.E. (2002) The value of multiple proxies. *Science*, **297**, 1481–1482.

Mann M.E. and Jones P.D. (2003) Global surface temperatures over the past two millennia. *Geophysical Research Letters*, **30**(15), 1820, doi:10.1029/2003GL017814.

Mann M.E., Rutherford S., Bradley R.S., Hughes M.K. and Keimig F.T. (2003) Optimal surface temperature reconstruction using terrestrial borehole data. *Journal of Geophysical Research*, **108**(D7), 4203, doi: 10.1029/2002JD002532.

Matthews H.D., Weaver A.J., Eby M. and Meissner K.J. (2003) Radiative forcing of climate by historical land cover change. *Geophysical Research Letters*, **30**(2), 1055, DOI:10.1029/2002GL016098.

Meier M.F., Dyurgerov M.B. and McCabe G.J. (2003) The health of glaciers: recent changes in glacier regime. *Climatic Change*, **59**, 123–135.

Milly P.C.D., Wetherald R.T., Dunne K.A. and Delworth T.L. (2002) Increasing risk of great floods in a changing climate. *Nature*, **415**, 514–517.

Mitchell J.F.B., Karoly D.J., Hegerl G.C., Zwiers F.W., Allen M.R., Marengo J., Barros V., Berliner M., Boer G., Crowley T., *et al.* (2001) Detection of climate change and attribution of causes. In *Climate Change 2001: The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Cubasch U., Meehl G.A., Boer G.J., Stouffer R.J., Dix M., Noda A., Senior C.A., Raper S., Yap K.S., *et al.* (Eds.), Cambridge University Press: Cambridge, pp. 697–738.

Mortiz R.E., Bitz C.M. and Steig E.J. (2002) Dynamics of recent climate change in the arctic. *Science*, **297**, 1497–1502.

Nakicenovic N. and Swart R. (Eds.) (2000) *Emission Scenarios 2000*, Special Report of the Intergovernmental Panel on Climate Change, Cambridge University Press: p. 599.

National Climatic Data Center (2004) U.S. National Oceanic and Atmospheric Administration, National Climatic Data Center: Ashville. **ftp://ftp.ncdc.noaa.gov/pub/data/anomalies/annual_land.and.ocean.ts**.

Nelson F.E., Anisimov O.A. and Shiklomanov N.I. (2002) Climate Change and Hazard Zonation in the Circum-Arctic Permafrost Regions. *Natural Hazards*, **26**, 203–225.

New M., Todd M., Hulme M. and Jones P. (2001) Precipitation measurements and trends in the twentieth century. *International Journal of Climatology*, **21**, 1899–1922.

Palmer T.N. and Raisanen J. (2002) Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature*, **415**, 512–514.

Parmesan C. and Yohe G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Penuelas J. and Filella I. (2001) Responses to a warming world. *Science*, **294**, 793–795.

Pepin L., Raynaud D., Barnola J.-M. and Loutre M.F. (2001) Hemispheric roles of climate forcings during glacial-interglacial transitions as deduced from the Vostok record and LLN-2D model experiments. *Journal of Geophysical Research*, **106**, 31885–3191892.

Petit J.R., Raynaud D., Lorius C., Jouzel J., Delaygue G., Barkov N.I. and Kotlyakov V.M. (2000) *Historical Isotopic Temperature Record from the Vostok Ice Core*. **http://cdiac.esd.ornl.gov/trends/temp/vostok/jouz_tem.htm**.

Prentice I.C., Farquhar G.D., Fasham M.J.R., Goulden M.L., Heimann M., Jaramillo V.J., Kheshgi H.S., Le Q., Scholes R.J., Wallace D.W.R., *et al.* (2001) The carbon cycle and atmospheric carbon dioxide. in *Climate Change 2001: The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Cubasch U., Meehl G.A., Boer G.J., Stouffer R.J., Dix M., Noda A., Senior C.A., Raper S., Yap K.S., *et al.* (Eds.) Cambridge University Press: Cambridge, pp. 183–237.

Przyblak R. (2002) Changes in seasonal and Annual High-frequency Air Temperature Variability in the Arctic from 1951–1990. *International Journal of Climatology*, **22**, 1017–1032.

Räisänen J. (2002) CO2-induced changes in interannual temperature and precipitation variability in 19 CMIP2 experiments. *Journal of Climate*, **15**, 2395–2411.

Ramaswamy V., Boucher O., Haigh J., Hauglustaine D., Haywood J., Myhre G., Nakajima T., Shi G.Y., Solomon S., Betts R., *et al.* (2001) Radiative forcing of climate change. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Cubasch U., Meehl G.A., Boer G.J., Stouffer R.J., Dix M., Noda A., Senior C.A., Raper S., Yap K.S., *et al.* (Eds.), Cambridge University Press: Cambridge, pp. 349–416.

Reichert B.K., Bengtsson L. and Oerlemans J. (2002) Recent glacier retreat exceeds internal variability. *Journal of Climate*, **15**, 3069–3081.

Reilly J., Stone P.H., Forest C.E., Webster M.D., Jacoby H.D. and Prinn R.G. (2001) Uncertainty and climate change assessments. *Science*, **293**, 430.

Rignot E. and Thomas R.H. (2002) Mass balance of polar ice sheets. *Science*, **297**, 1502–1506.

Rind D. (2002) The Sun's role in climate variations. *Science*, **296**, 673–678.

Robinson W.A., Reudy R. and Hansen J.E. (2002) General circulation model simulations of recent cooling in the east-central United States. *Journal of Geophysical Research*, **107**, 4748, doi: 10.1029/2001JD001577.

Root T.L., Price J.T., Hall K.R., Schneider S.H., Rosenzweig C. and Pounds J.A. (2003) Fingerprints of global warming on wild animals and plants. *Nature*, **421**, 57–60.

Rosenlof K.H., Oltmans S.J., Kley D., Russell J.M. III, Chiou E.-W., Chu W.P., Johnson D.G., Kelly K.K., Michelsen H.A., Nedoluha G.E., *et al.* (2001) Stratospheric water vapor increases over the past half-century. *Geophysical Research Letters*, **28**, 1195–1198.

Ross R.J. and Elliott W.P. (2001) Radiosonde-based northern hemisphere tropospheric water vapor trends. *Journal of Climate*, **14**, 1602–1612.

Ruddiman W.F. (2003) The anthropogenic greenhouse era began thousands of years ago. *Climatic Change*, **61**, 261–293.

Schiermeier Q. (2003) Insurers left reeling by disaster year. *Nature*, **421**, 6619–6699.

Schnur R. (2002) Climate science – the investment forecast. *Nature*, **415**, 483–484.

Shepherd A., Wingham D., Payne T. and Skvarca P. (2003) Larsen ice shelf has progressively thinned. *Science*, **302**(5646), 856–859.

Shindell D.T., Schmidt G.A., Miller R.L. and Mann M.E. (2003) Volcanic and solar forcing of climate change during the Pre-industrial Era. *Journal of Climate*, **16**, 4094–4107.

Shrag D.P. and Linsley B.K. (2002) Corals, chemistry and climate. *Science*, **296**, 277–278.

Smith J.B., Schellnhuber H.-J., Mirza M.M.Q., Fankhauser S., Leemans R., Erda L., Ogallo L., Pittock B., Richels R., Rosenzweig C., *et al.* (2001) Vulnerability to climate change and reasons for concern: a synthesis. In *Climate Change 2001. Third Assessment Report. WG II. Climate Change 2001: Impacts, Adaptation and Vulnerability*, McCarthy J.J., Canziani O.F., Leary N.A., Dokken D.J. and White K.S. (Eds.) Cambridge University Press, Cambridge, 911–967.

Stendel M. and Christensen J.H. (2002) Impact of global warming on permafrost conditions in a coupled GCM. *Geophysical Research Letters*, **29**(13), 1632, doi:10.1029/2001GL014345.

Stieglitz M., Dery S.J., Romanovsky V.E. and Osterkamp T.E. (2003) The role of snow cover in the warming of arctic permafrost. *Geophysical Research Letters*, **30**(13), 1721, 10.1029/2003GL017337.

Stone R.S., Dutton E.G., Harris M. and Longenecker D. (2002) Earlier spring snowmelt in Northern Alaska as an indicator of climate change. *Journal of Geophysical Research*, **107**(D10), 4089, 10.1029/2000JD000286.

Stone D.A. and Weaver A.J. (2002) Daily maximum and minimum temperature trends in a climate model. *Geophysical Research Letters*, **29**(9), 1356, 10.1029/2001GL014556.

Thompson L.G., Mosley-Thompson E., Davis M.E., Henderson K.A., Brecher H.H., Zagorodnov V.S., Mashiotta T.A., Lin P.-N., Mikhalenko V.N., Hardy D.R., *et al.* (2002) Kilimanjaro ice core records: evidence of Holocene climate change in tropical Africa. *Science*, **298**, 589–593.

Thompson D.W.J. and Solomon S. (2002) Interpretation of recent southern hemisphere climate change. *Science*, **296**, 895–899.

Timmermann A. (2001) Changes of ENSO stability due to greenhouse warming. *Geophysical Research Letters*, **28**, 2061–2064.

Trenberth K.E., Dai A., Rasmussen R.M. and Parsons D.B. (2003) The changing character of precipitation. *Bulletin of the American Meteorological Society*, **84**, 1205–1215.

Turner J., King J.C., Lachlan-Cope T.A. and Jones P.D. (2002) Recent temperature trends in Antarctica. *Nature*, **418**, 291–292.

Van Lipzig N.P.M., Van Meijgaard E. and Oerlemans J. (2002) Temperature sensitivity of the antarctic surface mass balance in a regional atmospheric climate model. *Journal of Climate*, **15**, 2758–2774.

Vaughan D.G. and Spouge J.R. (2002) Risk estimation of collapse of the West Antarctic ice sheet. *Climatic Change*, **52**, 65–91.

Vinnikov K.Y. and Grody N. (2003) Global warming trend of mean tropospheric temperature observed by satellites. *Science*, **302**, 269–272.

Vinnikov K.Y., Robock A., Stouffer R.J., Walsh J.E., Parkinson C.L., Cavalieri D.J., Mitchell J.F.B., Garrett D. and Zakharov V.F. (1999) Global warming and northern hemisphere sea ice extent. *Science*, **286**, 1934–1937.

Visbeck M.H., Hurrell J.W., Polvani L. and Cullen H.M. (2001) The North Atlantic oscillation: past, present and future. *Proceedings of the Academy of Natural Sciences*, **98**, 12876–12877.

Walther G.-R., Post E., Convey P., Menzel A., Parmesan C., Beebee T.J.C., Fromentin J.M., Hoegh-Guldberg O. and Bairlein F. (2002) Ecological responses to recent climate change. *Nature*, **416**, 389–395.

Wang X.L. and Swail V.R. (2002) Trends of Atlantic Wave extremes as simulated in a 40-yr wave hindcast using kinematically reanalyzed wind fields. *Journal of Climate*, **15**, 1020–1035.

Watson R.T., Noble I.R., Bolin B., Ravindranath N.H., Verardo D.J. and Dokken D.J. (Eds.) (2000) *Land Use, Land-Use Change and Forestry*, Intergovernmental Panel on Climate Change, Cambridge University Press.

Weaver A.J., Eby M., Fanning A.F. and Wiebe E.C. (1998) Simulated influence of carbon dioxide, orbital forcing and ice sheets on the climate of the last glacial maximum. *Nature*, **394**, 847–853.

Webster M.D., Forest C.E., Reilly J.M., Babiker M.H., Kicklighter D., Mayer M., Prinn R., Sarofim M.C., Sokolov A., Stone P., *et al.* (2003) Uncertainty analysis of climate change and policy response. *Climatic Change*, **61**, 295–320.

Wilby R.L. and Wigley T.M.L. (2002) Future changes in the distribution of daily precipitation totals across North America. *Geophysical Research Letters*, **29**(7), doi:10.1029/2001GL013048.

Yang F., Kumar A., Schlesinger M.E. and Wang W. (2003) Intensity of hydrological cycles in warmer climates. *Journal of Climate*, **16**, 2419–2423.

Zhao M. and Pitman A.J. (2002) The impact of land cover change and increasing carbon dioxide on the extreme and frequency of maximum temperature and convective precipitation. *Geophysical Research Letters*, **28**, 10.1029/2001GL013476.

PART 4

Hydrometeorology

35: Rainfall Measurement: Gauges

BORIS SEVRUK

Institut für Atmosphäre und Klima (ETH), Zürich, Switzerland

It is difficult to measure precipitation without introducing systematic errors or biases. The quality of any measurement will depend on the type of gauge used and the details of the installation, as well as the characteristics of the gauge site including exposure and the prevailing weather. All precipitation measurements using can-type gauges are subject to systematic errors, mostly due to deformation of the wind field above the precipitation gauge orifice, wetting and evaporation losses, splash-out and splash-in, blowing snow, and so on. Because these measurement errors are systematic, precipitation gauges frequently show different precipitation figures. This is the case when the gauges are of different construction, or the same construction, but from different manufacturers, or are installed at different heights at the same site near each other, or under different gauge site exposure. With the exception of measurements of blowing and drifting snow, they show less precipitation than the "true" precipitation falling on the ground. Correction procedures are different for rain and snow. They were developed for different types of precipitation gauges and various time intervals according to the availability of necessary input data. Corrections are based either on empirical field methods using the World Meteorological Organization reference standard, laboratory experiments, or numerical simulation.

INTRODUCTION

Precipitation measurements are important for many applications in meteorology, hydrology, agriculture, and climate research. In hydrology, they are used to assess the areal precipitation, which is a basic component of water balance and precipitation-runoff models. To achieve these aims, elevated can-type precipitation gauges are operated in networks consisting of several gauge sites selected according to the main characteristics of the area and precipitation fields of interest (*see* **Chapter 36, Precipitation Measurement: Gauge Deployment, Volume 1**). Such simple point measurements are converted to areal precipitation using various interpolation techniques, including empirical and geostatistical (kriging) methods. They are also used to check the results of more complex methods to assess areal precipitation such as radar (*see* **Chapter 63, Estimation of Precipitation Using Ground-based, Active Microwave Sensors, Volume 2**), satellites (*see* **Chapter 64, Satellite-based Estimation of Precipitation Using Microwave Sensors, Volume 2**), and mesoscale forecast model (**Chapter 31, Models of Clouds, Precipitation and Storms, Volume 1**). Methods of precipitation measurements and definitions in English, French, Russian, Chinese, and Spanish are described briefly in WMO (1994), and with more details being given in German by Sevruk (2004). For further discussion and details of the most significant and recent papers, the reader is referred to the review by Sevruk (2004).

TYPES OF PRECIPITATION GAUGES

According to the definition, precipitation gauges are universal instruments to catch all forms of liquid and solid precipitation, including fog, dew, rime, and so on. They consist of a cylindrical collector to catch the precipitation, and a container to accumulate it during the recording time interval. The collector and the container are connected through a funnel with an outflow. Sometimes minor changes are made to measure snowfall in places where snow makes up a large portion of the total precipitation. (The funnel is removed to increase the depth of the collector, or a snow cross is inserted into the collector to reduce the blowing-out of snow.) To ensure the orifice area remains at the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

standard size and to protect it against deformation, the orifice is reinforced by a hard metal rim of a given diameter with sharpened upper edge. The orifice level has to be kept horizontal. To protect it against splash-in, drifting, and blowing snow, it is fixed on a pole at a standard height above the ground, or placed on the ground and surrounded by antisplash material. The precipitation is measured as a volume and converted into depth units, that is, millimetres (or inches) to the accuracy of 0.1 mm (or 0.1 in). There are three basic types of precipitation gauges. They differ according to the recording time interval as follows: recording gauges have a resolution of one minute, standard (daily) gauges of a few hours up to one day, and storage gauges (totalizer) of months up to one year. The latter have a big container to accumulate annual precipitation, and are installed mostly in mountains at the heights well above the depth of snow cover, of at least 3 m above the ground, to be protected against blowing and drifting snow. The standard gauges and storage gauges are measured manually using calibrated cylinders or sticks graduated in 0.1 mm (or 0.1 in). Precipitation depths smaller than 0.05 mm are recorded as a trace. Recording gauges have different systems of measurements. The demand for precipitation data of very short time intervals to record high rainfall intensities (up to 5 mm min^{-1}), which are important variables for many engineering applications, has resulted in various instrument designs and operational data processing and archiving procedures. These designs may use a float to record the water level in the container, or mechanical and electronic weighing systems weigh continuously the water in the container or falling from the collector (tipping-bucket). Drop counters and optical systems are also used. Distrometers are used to measure the drop-size distribution. To melt the snow accumulated in a precipitation gauge, standard daily gauges are brought to a warm room, heating devices are used by recording gauges, and antifreeze liquid is used by storage gauges. The accumulated precipitation in storage gauges is protected against evaporation by a thin oil layer.

Precipitation gauges vary from country to country. Sevruk and Klemm (1989) showed that there were more than 50 types of manual, national standard precipitation gauges in use at that time around the world. Gauges differ considerably in design, shape, size, and material. The orifice area varies from 7 to 1000 cm^2, most gauges having an area of 100–200 cm^2. Materials used are primarily galvanized iron and copper, but plastic is popular as well. The installation height varies between countries from 0.2 to 2.0 m. Up to 90 countries use installation heights of 1 m and less, 2 m was used only in the countries of the former Soviet Union but 0.3 m in the former British Empire. To reduce adverse wind effects, seven types of gauges are permanently fitted with different designs of windshield. In each country, one specific type of gauge installed at the same standard height, is used as the standard gauge. Even so, in regions where snow is common, the installation height is usually bigger. The most widely used standard manual gauges appear to be the German Hellmann gauge, followed by the Chinese type and the English (Snowdon) Mk2 gauge. These types account for one-half of all precipitation gauges (approximately 200 000 worldwide) but they are concentrated in an area of only 31×10^6 km^2. Of the recording precipitation gauges, float systems are the most common, followed by tipping-bucket gauges. Weighing systems are used far less, and drop counter and optical systems are rarely used (Sevruk, 2002). The oldest method of precipitation recording, a pen and time chart, is still used in most countries, followed by modern data logging and other intermediate methods. Punched paper tape is used only in a few countries. The most used method of data transmission is by conventional post, followed by telephone, radio, and satellite. But data loggers, telex, e-mail, telegraph, and PC networks are all also used. In contrast to the standard manual gauges, a particular country may use different recording gauges at different installation heights.

ERRORS OF PRECIPITATION MEASUREMENTS

It is difficult to measure precipitation without introducing systematic errors or biases (Sevruk, 1982, 1986a; WMO, 1994). Some observers have a tendency to roundup smaller fractions of precipitation than 0.1 mm, others round it down. Moreover, even the method of employing the measuring glass or stick during readings can change between different observers. Small deviation in the gauge orifice from the horizontal level can also result in errors (Sevruk, 1984). Both error sources cause random errors. Other errors arise from the fact that the can-type gauges with a horizontal orifice are not able to measure properly all forms of precipitation such as fog, dew, and so on, as well as the oblique precipitation falling on exposed slopes. This causes deficits in certain regions (e.g. Andes, tropical forests, mountains, slopes etc.), and in these cases special instruments should be used to measure precipitation more accurately. On exposed windward slopes, precipitation is frequently falling because of strong winds, perpendicular to the slope or drifting along the slope, and the horizontal orifice does not catch the precipitation properly. The windward side of the gauge collector presents an obstacle to the airflow, and the precipitation particles are carried out over the horizontal orifice. In addition, the horizontal projection of a slope results in a smaller area than the true area of the slope. Consequently, the volumes of precipitation and runoff are in fact greater than assessed from the measurements. Therefore, orifices parallel with the slope are preferably on exposed slopes in hydrological and agrometeorological studies (e.g. erosion, land slides) (Sevruk, 1972; Sharon, 1980). The quality of measurements also depends on the type of gauges used and

how they are installed, as well as the characteristics of the gauge site and the prevailing weather. Generally, recording gauges seem to be less accurate than manual gauges (Sevruk, 1996), in that they tend to measure less precipitation. Particularly, the tipping-bucket gauge is not very accurate (Sevruk, 1996; Upton and Rahimi, 2004). These gauges should be calibrated at least once a year. Losses due to heating in the winter season are large (Zweifel and Sevruk, 2003). Frequent failures such as blocking of the tipping-bucket, clogging of the outflow by insects, leaves, bird droppings, and so on, increase the maintenance costs. Electronic weighing systems seem to be better, but they show also shortcomings such as temperature effects, and software is needed to filter the effects of wind shocks and sudden changes of weight (Sevruk, 2004). However, all precipitation measurements using can-type gauges are subject to systematic errors mostly due to wind field deformation above the precipitation gauge orifice, wetting and evaporation losses, splash-out and splash-in, blowing snow, and so on (Sevruk, 1982, 1993). With the exception of snow blowing and drifting, they show less precipitation than the "true" precipitation falling on the ground. An elevated precipitation gauge systematically distorts the wind field above the gauge orifice and forces the wind speed to increase over the gauge orifice (blocking effect). This phenomenon has been demonstrated many times by experiments in wind tunnels and by numerical simulation (Sevruk *et al.*, 1989; Nespor and Sevruk, 1999). The adverse effect of the wind is that some of the lighter precipitation particles are borne away before reaching the gauge and are lost from the measurement. The wind-induced loss depends on windspeed, the weight of precipitation particles, and the gauge construction (e.g. the shape of the gauge and orifice rim, and the use of windshields). Windshields reduce the height/diameter ratio of the gauge, which improves aerodynamic properties, shift the gauge leeward to the zone of lower windspeed, and displace the airflow downward. In this way, they reduce the wind-induced losses. Consequently, the wind-induced loss is smaller for large intensities (i.e. heavy raindrops), small installation heights (smaller windspeeds), and gauges with windshield and those that are placed at protected gauge sites. (The effect of windshields is generally small for rains, as shown by Duchon and Essenberg, 2001.) In contrast, it is large for small intensities, large installation heights, unshielded gauges, and at exposed gauge sites. The loss amounts on average to 2–10% of measured values of rain and up to 60% of snow for unshielded gauges and windspeeds greater than $4\,\mathrm{m\,s^{-1}}$ (Goodison *et al.*, 1998). Since the wind-induced losses during snowfall can be ten times larger than for rain, the fraction of snow in the total precipitation is a very important variable to consider when assessing the losses. The error resulting from the blowing of snow into the gauge should be considered during snowstorms with windspeed larger than $5\,\mathrm{m\,s^{-1}}$ (Bogdanova *et al.*, 2003). In dry cold regions, the overcatch due to blowing snow can more than compensate for the wind-induced losses and the trace precipitation (Sugiura *et al.*, 2003). Some types of windshields also cause overcatch of snow during small windspeeds. (The zone of concentrated precipitation is shifted over the gauge.) Wetting losses depend on age, material, and the depth of the collector relative to its diameter, as well as on the frequency of events and types of precipitation. Snowfalls cause smaller wetting losses than raindrops, which stick on the inner walls of the collector and container and evaporate. Generally, wetting losses amount to 0.2 and 0.1 mm per rain and snow event, respectively. The evaporation loss from the container depends on the shape of the gauge and weather characteristics (e.g. saturation deficit of the air, windspeed, and duration of evaporation, see Sevruk and Klemm, 1989; WMO, 1994).

CONSEQUENCES OF ERRORS OF PRECIPITATION MEASUREMENT

The systematic measurement errors in precipitation gauges frequently result in different precipitation figures. This is the case when the gauges are of different construction, or the same construction, but from different manufacturers, or when gauges are installed near to each other but at different heights at the same site, or at different gauge sites. (Windspeed increases with the height above the ground.) Because different types of precipitation gauges as used in the same or different countries, the global and local precipitation data sets are often not compatible (Sevruk, 1994). Comparing precipitation figures and intensities among countries, or within a single country, often shows systematic differences. To eliminate these spatial inhomogeneities of precipitation time series, the performance of precipitation gauges has to be checked and the precipitation measurements corrected. Similarly, changing to a different type of gauge at a site, or moving the gauge to a new site with different exposure, can also cause inhomogeneities (Peterson *et al.*, 1998). Sevruk and Klemm (1989) showed that in 50 countries, changes in national standards such as the replacement of a long-used type, or lowering or increasing the installation height, have taken place in the last 50 years. The most decisive change was related to the replacement of a certain number of standard manual precipitation gauges with recording precipitation gauges (Sevruk, 1996). The effect of gauge site exposure on the quality of precipitation measurements is most readily assessed using four classes of exposure (exposed, mostly exposed, mostly protected, and protected). They are based on analysis of metadata stored in archives of meteorological services (De Smedt *et al.*, 2003). Exposed sites are characterized by only small obstacles such as bushes, group of trees, or a small house. Mostly, exposed sites are surrounded by small groups of trees or bushes or one or two houses. They frequently occur on islands, lake

shores, in high mountains, prairies, and so on. Mostly, protected sites are found in parks, forest edges, village centers, farms, groups of houses, or yards. Protected sites are typically young forests, small forest clearings, parks with big trees, city centers, closed deep valleys, strongly rugged terrain, or leeward of big hills and buildings, and so on. The classes can be based on direct measurements of the vertical angle of obstacles. In some cases, fish-eye pictures of gauge sites are also available. Because of considerable wind-induced losses, the first two classes should not be used for hydrological studies, unless precipitation measurements are corrected.

CORRECTIONS OF PRECIPITATION MEASUREMENT ERRORS

Correction procedures are different for rain and snow, and different corrections have been developed for different types of precipitation gauge and various time intervals according to the availability of necessary input data as shown by Sevruk (1982, 1986a). They are applied in many parts of the world (e.g. Scandinavia, Germany, Switzerland, Baltic countries, former USSR, Greenland, Tibet, Slovakia etc.). Input data include windspeed, the precipitation intensity and weather situation (different drop-size distributions exist for different types of rain with the same intensity), temperature, rain/snow amounts, frequency of events, and so on. Corrections are based either on empirical field methods, using the World Meteorological Organization (WMO) reference standard, or on numerical simulation as described by Nespor and Sevruk (1999). In addition, the water equivalent of fresh snow and snow cover and the snow depth have also been used to assess the "true" snowfall and to derive correction procedures (Sevruk, 1983, 1986b; Sevruk *et al.*, 1998). The WMO reference standard consists of a pit gauge with antisplash grid for rain measurements, and the double-fence for snow measurements (Figure 1). The pit gauge is sunk into a pit to make the gauge orifice level with the ground. A metallic or plastic grid protects the gauge against splash-in. Because the windspeed at ground level is reduced almost to zero, the wind-induced error of pit gauges is small. However, such a gauge cannot be used for snow measurements as drifting snow would fill it and falsify the measurements. The double-fence was developed to overcome this problem. It consists of two circular lath-fences of different diameters: 12 m for the outer one and 4 m for the inner one. The respective heights are 3.5 m and 3 m. A precipitation gauge with a windshield is situated in the center and elevated 3 m above the ground. The WMO organized two international intercomparisons using these references with the aim of developing correction procedures for the wind-induced losses, as reported

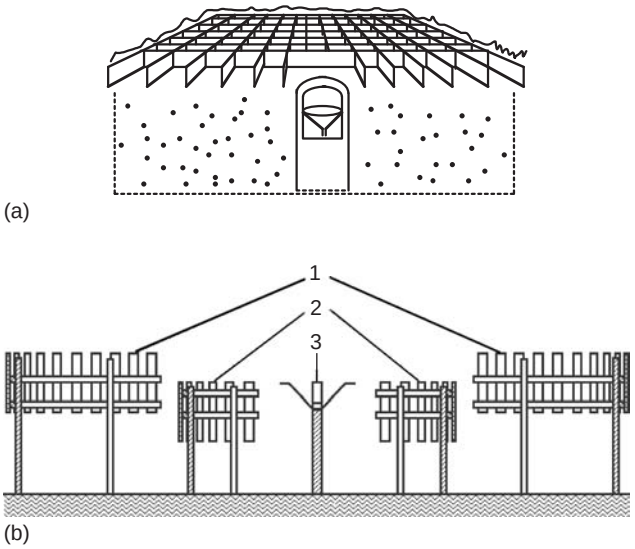


Figure 1 The WMO reference standard precipitation gauges. (a) the pit gauge with antisplash grid for rain measurements and (b) the double-fence for snow measurements. The latter consists of two fences of different diameters of 12 m (the outer one: 1) and 4 m (the inner one: 2). The respective heights are 3.5 m and 3.0 m. 3: indicates the Russian Tretjakov gauge with windshield

by Sevruk and Hamon (1984) and Goodison *et al.* (1998). The conversion factor k, defined as the ratio of precipitation values of the reference, P_k, and the elevated gauge, P_m, is used to correct the wind-induced losses ($P_k = k P_m$,). k is related to the windspeed, and the parameters of precipitation structure such as the mean intensity of rain, temperature, and the fraction Q of solid precipitation in total precipitation. Such a dependency for monthly means is shown in Figure 2 and for daily values in Figure 3. Because the structure of snow depends on temperature intervals, the diagram in Figure 2(b) is valid for snowfalls in the range of temperature from -1 to $-9\,^{\circ}\mathrm{C}$ (Sevruk, 1982). It is important to note that there is a certain threshold value of precipitation intensity for a specific windspeed interval, because below a certain threshold value, despite the same windspeed, the wind-induced error increases rapidly (Sevruk, 1989a, Figure 3). The smaller the actual intensity, the greater the fraction of small precipitation particles, which are blown away even by very small windspeeds. This process is nonlinear as shown in Figure 3. Generally, the intensity of rain i of $0.03\,\mathrm{mm\,min^{-1}}$ is considered as the threshold value for monthly precipitation and all windspeeds (Sevruk, 1982). The fraction of rain in the total monthly precipitation, being below such an intensity is termed the parameter of rain structure N'. Because precipitation intensity measurements are missing at many gauge sites, the parameter N' was derived from the mean monthly temperature, T, the measured precipitation depth, N_g, and number q of days with precipitation as shown in

the insert in the left corner in Figure 2(a) (Sevruk, 1989b). Considering the nature of the wind-induced losses from the point of view of fluid dynamics, it is clear that the results of field intercomparison measurements are valid for a particular precipitation gauge type, defined by relevant parameters. Such international intercomparisons provide important information on the accuracy, reliability, and corrections of the measurements, and good guidance for the selection of the most reliable type of precipitation gauge.

In the case of simulations, the wind-induced error is estimated by computing the wind field around the gauge, using computational fluid dynamics methods and by numerical simulation of precipitation particle trajectories in the computed flow field. Nespor and Sevruk (1999) and Sevruk *et al.* (2000) show the results of simulations for two unshielded precipitation gauges used in Switzerland: the common daily gauge by Hellmann and the automatic station gauge ASTA (tipping-bucket). Simulation computations have many advantages when compared with the rather costly empirical methods. Once the methodology is developed, correction procedures for any type of precipitation gauge and any range of variables and very small time intervals can be made quickly on a computer, whereas, for field tests, years of data are necessary and the range of variables is limited. Usually, the most interesting events with high windspeeds and small intensities of precipitation, showing the greatest wind-induced losses are missing. Moreover, during field tests with heated gauges, the losses of precipitation catch caused by heating should be accounted separately for. This presents a rather difficult problem. The agreement between the results of empirical and theoretical procedures

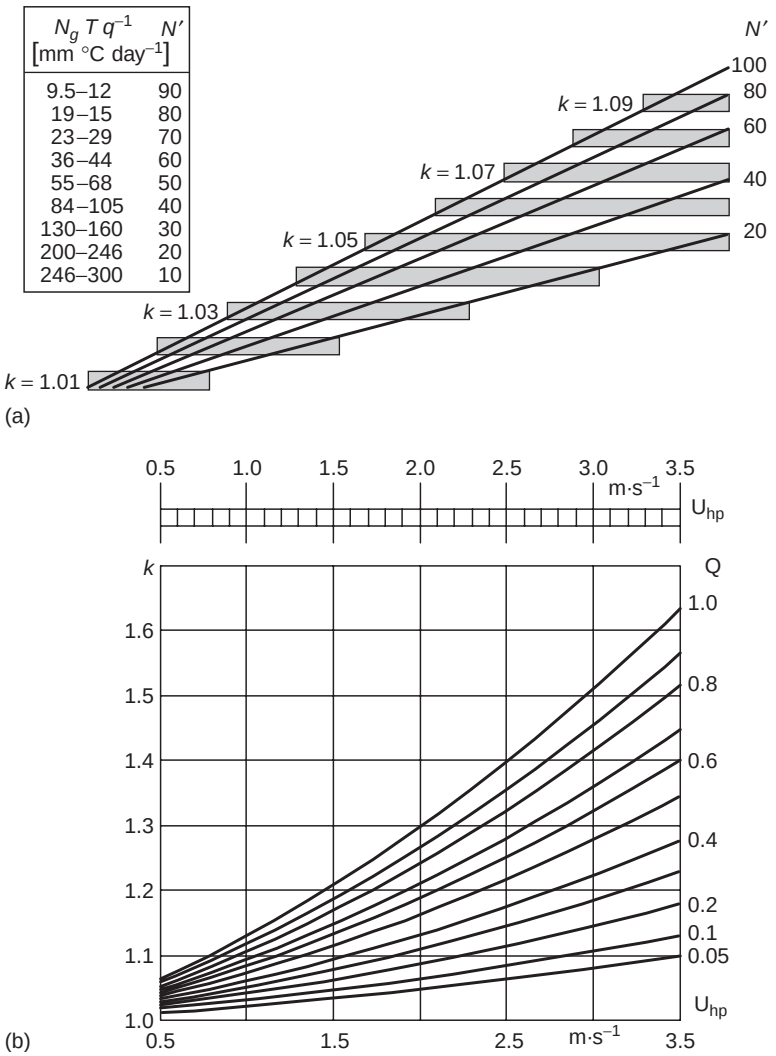


Figure 2 Mean monthly values of conversion factor k of precipitation measurements using the Hellmann gauge to reference measurements as a function of wind speed u_{np} and the structure parameter N' for rain (a) and fraction Q of snow in total precipitation (b). The latter is valid for the monthly temperature range from −1 to −9 °C

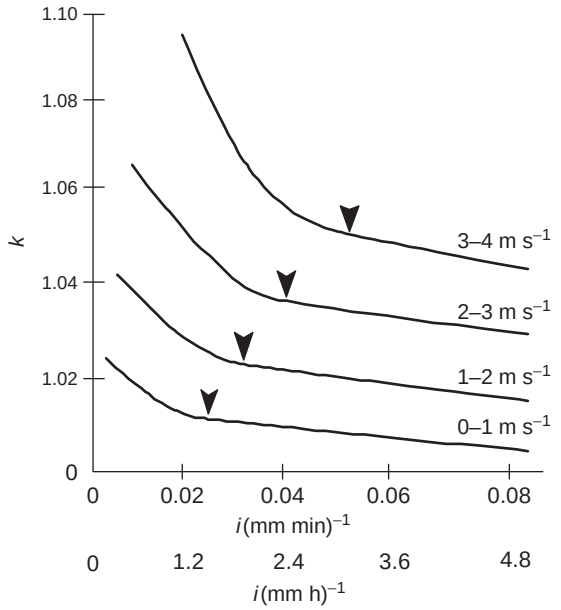


Figure 3 Wind-induced daily conversion factor k as related to the mean daily intensity of rain i and windspeed. The arrows indicate the threshold values of rain intensity. Below the threshold values the conversion factors k increase rapidly. According to Sevruk (1989b)

is fairly good, as shown by Nespor and Sevruk (1999), and Chvíla *et al.*, (2002). Having a developed correction procedure for a given type of gauge, the problem consists primarily in the resolution and availability of input data. This poses no problems for the automatic meteorological stations where hourly values of many variables are available, but it is a serious disadvantage for common gauge sites with only one precipitation observation per day. Here, interpolations are needed of practically all variables (Sevruk, 1986c). Wind measuring instruments are usually installed 10–20 m about the ground, which is up to ten times higher than the precipitation gauge orifice level (1.0–2.5 m). This discrepancy can be accounted for using the reduction procedure on the basis of the logarithmic vertical wind profile and degree of gauge site protection (De Smedt *et al.*, 2003). The latter can be directly measured at a site using a theodolite or assessed from the metadata as discussed above.

It is difficult to measure precipitation without introducing systematic errors or biases. All precipitation measurements using can-type gauges are subject to systematic errors, mostly due to wind field deformation above the precipitation gauge orifice, wetting and evaporation losses, splash-out and splash-in, blowing snow, and so on. The errors depend on the type of gauges used and their installation specifics, as well as the characteristics of the gauge site including exposure and weather situations. At present, precipitation measurements are not corrected regularly for systematic errors by national meteorological and hydrological services using standard procedures – as other meteorological variables, for example, temperature, humidity, and so on, are. Correction procedures are based on field and laboratory tests, or numerical simulations. Corrections have been developed for different types of precipitation gauges and various time intervals, according to the availability of the necessary input data, and are therefore gauge site dependent. Because these procedures are not yet standardized and will develop further in the future, it is important to preserve both the corrected and the original measured precipitation values.

REFERENCES

Bogdanova E.G., Ilyin B.I. and Dragomilova I.V. (2003) Applying a comprehensive model for bias correction of measured precipitation in different climatic conditions. *Proceedings of WCRP Workshop on Determination of Solid Precipitation in Cold Climate Regions*, Fairbanks, June 2003.

Chvíla B., Ondras M. and Sevruk B. (2002) *The Wind-Induced Loss of Precipitation Measurement of Small Time Intervals as Recorded in the Field*, Instruments and Observing Methods Report, WMO/TD-No. 1123 (CD), World Meteorological Organization: Geneva.

De Smedt B., Mohymont B. and Demarée G.R. (2003) Grubs revisited: a statistical technique to differentiate measurement methods of the degree of exposure of rain gauges in a network. *Journal of Hydrology Sciences*, December **48**(6), 871–897.

Duchon C.E. and Essenberg G.R. (2001) Comparative rainfall observations from pit and aboveground rain gauges with and without wind shields. *Water Resources Research*, **37**(12), 3253–3263.

Goodison B.E., Louie P.YT and Yang D. (1998) *WMO Solid Precipitation Measurement Intercomparison*, WMO/TD-No. 872, World Meteorological Organization: Geneva, p. 212.

Nespor V. and Sevruk B. (1999) Estimation of wind-induced error of rainfall gauge measurements using a numerical simulation. *Journal of Atmospheric and Oceanic Technology*, **16**(4), 450–464.

Peterson ThC, Easterling D.R., Karl ThR, Groisman P., Nicolls N., Plummer N., Torok S., Auer I., Boehm R., Gillett D., *et al.* (1998) Homogenity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology*, **18**(13), 1493–1517.

Sevruk B. (1972) Precipitation measurements by means of storage gauges with stereo and horizontal orifices in the Baye de Montreux watershed. *Distribution of Precipitation in Mountainous Areas*, WMO-Publication No. 326(2), World Meteorological Organization: Geneva, pp. 86–95.

Sevruk B. (1982) *Methods of Correction for Systematic Error in Point Precipitation Measurement for Operational Use*, Operational Hydrology Report, 21, WMO-No. 589, World Meteorological Organization: Geneva, p. 91.

Sevruk B. (1983) Correction of measured precipitation in the Alps using the water equivalent of new snow. *Nordic Hydrology*, **14**(2), 49–58.

Sevruk B. (1984) Comments on "Out-of-level" instruments: error in hydrometeor spectra and precipitation measurement. *Journal of Climatology and Applied Meteorology*, **23**(6), 988–989.

Sevruk B. (Ed.) (1986a) *Proceedings of Workshop on the Correction of Precipitation Measurements*, *Zürcher Geographische Schriften*, ETH Zürich, No. 23. (See also Instruments and Observing Methods Report., No. 24, WMO/TD-No. 104, World Meteorological Organization: Geneva, 1985), p. 288.

Sevruk B. (1986b) Conversion of snowfall depths to water equivalents in the Swiss Alps. In *Proceedings of Workshop on the Correction of Precipitation Measurements*, Sevruk B. (Ed.), *Zürcher Geographische Schriften*, Eidgenossische Technische Hochschule ETH: ETH Zürich, No. 23, pp. 81–88.

Sevruk B. (1986c) Correction of precipitation measurements: Swiss experience. In *Proceedings of Workshop on the Correction of Precipitation Measurements*, Sevruk B. (Ed.), *Zürcher Geographische Schriften*, Zürich, Geograhpisches Institut: ETH Zürich, No. 23, pp. 187–196.

Sevruk B. Hertig J.-A. Spiess R. (1989) Wind field deformation above precipitation gauge orifices. In *Atmospheric deposition*, Delleur J.W. (Ed.), International Association of Hydrological Sciences Publication, No. 179, 65–70.

Sevruk B. (1989a) *Wind-Induced Measurement Error for High-Intensity Rains*, WMO Instruments and Observing Methods Report, No. 48, WMO/TD- No. 328, World Meteorological Organization: Geneva, pp. 199–204.

Sevruk B. (1989b) *Precipitation Correction: Parameter Estimation of Precipitation Structure*, WMO Instruments and Observing Methods Report, No. 35, World Meteorological Organization: Geneva, pp. 317–322.

Sevruk B. (1993) Checking precipitation gauge performance. In *Measurement of Airborne Pollutants*, Couling B. (Ed.), Butterworth Heinemann: Oxford, pp. 89–107.

Sevruk B. (1994) Spatial and temporal inhomogeneity of global precipitation data. In *Global Precipitation and Climate Change*, NATO ASI Series, Vol. I 26, Desbois M. and Desalmand F. (Eds.), Springer Verlag: Berlin, pp. 219–230.

Sevruk B. (1996) Adjustment of tipping-bucket precipitation gauge measurements. *Atmospheric Research*, **42**(1–4), 237–246.

Sevruk B. (2002) WMO questionnaire on recording precipitation gauges: state-of-the-art. *Water Science and Technology*, **45**(2), 139–145.

Sevruk B. (2004) Precipitation as the water cycle element. *Theory and Practice of Precipitation Measurement*, (in German), Manuscript to be published.

Sevruk B. and Hamon W.R. (1984) *International Comparison of National Precipitation Gauges with a Reference Pit Gauge*, Instruments and Observing Methods Report, No. 17, World Meteorological Organization: Geneva, p. 135.

Sevruk B. and Klemm S. (1989) *Catalogue of National Standard Precipitation Gauges*, Instruments and Observing Methods Report, WMO/TD-No. 313, World Meteorological Organization: Geneva, p. 50.

Sevruk B., Paulais M. and Roulet Y.-A. (1998) Correction of precipitation measurement using fresh snow as reference. *Proceedings of WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation TECO-98*, WMO/TD-No. 877, World Meteorological Organization, Geneva, pp. 349–352.

Sevruk, B. Roulet Y.-A. and Nespor V. (2000) *Corrections of the Wind Induced Error of Tipping-Bucket Precipitation Gauges in Switzerland Using Numerical Simulation*, Instruments and Observing Methods Report, No. 74 (TECO-2000 Beijing, China, WMO/TD No. 1028, World Meteorological Organization, Geneva, pp. 144–147.

Sharon D. (1980) The distribution of hydrologically effective rainfall incident on sloping ground. *Journal of Hydrology*, **46**(1–2), 165–188.

Sugiura K., Yang D. and Ohata T. (2003) Systematic error aspects of gauge-measured solid precipitation in the Arctic, Barrow, Alaska. *Geophysical Research Letters*, **30**(4), 41–44.

Upton G.J.G. and Rahimi A.R. (2003) On-line detection of errors in tipping-bucket raingauges. *Journal of Hydrology*, **278**(1–4), 197–212.

WMO (1994) *Guide to Hydrological Practices. Data Aquisition and Processing, Analysis, Forecasting and Other Applications*, *Fifth Edition*, WMO-No. 168. World Meteorological Organization, Geneva.

Zweifel A. and Sevruk B. (2003) Comparative accuracy of solid precipitation measurement using heated recording gauges. *Proceedings of WCRP Workshop on Determination of Solid Precipitation in Cold Climate Regions*, Fairbanks, June 2003 (CD Rom).

36: Precipitation Measurement: Gauge Deployment

MARK ROBINSON

Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

The deployment of a network of rain gauges involves two stages: first, the broad distribution of gauges across a region or catchment and, second, the specific location of each gauge. The spatial distribution of gauges across a region will be determined by a combination of: (a) physical characteristics of the region, including the topography and the natural variability of precipitation, and (b) institutional factors, such as the purpose of the network (setting the accuracy and level of detail of data required as well as the resources available). The specific location of gauges will be determined by purely local factors, which would affect the accuracy of the measurements. These are primarily concerned with site exposure and wind effects that could lead to errors in catch and make the data unrepresentative of a larger area.

INTRODUCTION

Precipitation is a major factor controlling the hydrology of a region and is of fundamental importance to many human activities. The number and distribution of rain gauges needed for a particular area will depend upon the natural variability of precipitation and upon the purpose for which the data are collected, since this will determine the detail and accuracy of measurements required. Much more detailed information will, for example, be required for urban storm drainage design and for research purposes than for general water resource and water supply projects. In addition, the form of the precipitation may also influence the network of gauges: snow presents special problems (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**) and may need to be tackled in different ways, including the use of remote sensing and snow lines. In some coastal or mountain areas, large quantities of fine water droplets in low cloud or mist may be deposited directly onto vegetation and other surfaces and special measurements may be needed (*see* **Chapter 38, Fog as a Hydrologic Input, Volume 1**).

The primary measure of precipitation is the catch in a rain gauge. The amount of water collected may be measured by manually emptying a storage rain gauge and noting the amount of accumulated water. This is done at regular intervals, most commonly each day at a set time. Alternatively a recording rain gauge may be used, which automatically registers the rate of accumulation of rainfall. The main type of recording rain gauges uses a tipping bucket mechanism to record increments of rain, typically at a resolution of 0.1 to 0.5 mm, over hourly intervals or less. Guidelines on procedures for collecting and processing rain gauge data are provided by Meteorological Office (1982, 2001) and WMO (1994). Gunston (1998) provides advice particularly for developing countries.

Despite the history of rainfall measurement dating back over 2000 years (Biswas, 1970; NIH, 1990) many problems remain in the collection and accuracy of precipitation data. The major source of error is due to wind turbulence generated around the gauge, which usually results in underestimates (Sevruk, 1982). This may be affected by the type of rain gauge (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**) and, also very importantly, by the local exposure of the site.

The deployment of rain gauges within an area needs to consider two aspects:

- The design of the rain gauge network, defining its purpose, the anticipated variability of precipitation, and the resources available to operate the network.
- The selection of suitable gauge sites within this framework which will provide catches that are accurate and representative

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Deployment of a Network of Gauges

A network of rain gauges represents a finite number of point samples of the two-dimensional pattern of precipitation depths. The accuracy of areal estimation depends on both the total number of gauges and their spatial distribution.

Although there are numerous technical "Guidelines" to the density of gauges needed to define areal rainfall, the actual number of gauges likely to be deployed will depend very much upon two groups of factors:

1. The natural variability of the precipitation distribution across an area. This will be determined by many physical factors, of which the most important include the topography and climate characteristics. Precipitation is more likely to be variable in areas of hilly or mountainous topography and where rainfall results from localized convectional storms rather than widespread frontal systems over flat terrain. In estimating the areal pattern of precipitation from a given gauge network, errors will occur because of the random nature of storms and their paths relative to the gauges. The accuracy of the network will depend on the spatial variability of precipitation; thus more gauges would be required in areas of steeply sloping terrain or in areas prone to random localized thunderstorms rather than to widespread frontal systems.
2. The purpose of the network (including the accuracy required) and funding available for the installation, operation, and maintenance of the gauges will have an important influence on the network of gauges deployed. There are very important nonhydrological constraints and considerations (social, economic) in determining network density. A rain gauge or any other network exists to serve certain objectives in developing and managing water resources, and must do this "on a scale commensurate with the overall level of economic development and environmental needs of the country" (WMO, 1994). The density of gauges required will, for example, depend upon the timescale of interest. For water resource purposes in large basins, monthly or seasonal totals may be sufficient whilst in some studies, such as urban runoff design, rainfall depths over only a few minutes' duration may be required. In short-duration local storms, the spatial distribution of precipitation is very heterogeneous and precipitation gradients may be quite steep. As larger areas and longer time periods are considered, the distribution of precipitation becomes more homogeneous and gradients are less steep. Denser networks of rain gauges are needed for smaller space and time intervals studies. This is why the purpose of a precipitation gauge network design is one of the limiting factors of the accuracy of areal precipitation assessment.

In situations where only manually read gauges are in use, there is an obvious need for these gauges to be at sites near to centers of population and road access. Historically, this has resulted in gauge distributions with a concentration in river valleys and near to towns and villages, and with relatively fewer in adjacent hill areas, despite the often greater precipitation depths. The development of automatic recording gauges has not only provided much increased information on the short-term rates of rainfall, but, importantly, it also means that gauges can be distributed more easily into unpopulated areas, although snow measurement is still an unresolved problem.

In practice, rain gauge distributions generally reflect a mixture of both the physical and social or institutional factors described above and the resulting "network" is simply "an aggregation of gauges or stations that have no coherence in their objectives" (WMO, 1994). Rain gauge network design is usually concerned with modifying an existing network, either to expand it to meet new requirements or to reduce it to lower costs. Only very rarely does it involve creating an entirely new network from scratch. If a new network is envisaged, then there are several different approaches. If an area has no existing gauges to indicate the spatial distribution of precipitation, it is necessary to transpose information on its variations in time and space from a similar area, in order to design a preliminary network. This is more often the case for developing countries. In developed countries, in contrast, some form of network is usually already in place but probably evolved in an arbitrary manner. One approach to network design is to define *a priori* those factors that are likely to control precipitation patterns – such as altitude, distance to the sea, ground slope and aspect – and to divide the area into domains representing classes with different ranges of these geographical and topographic characteristics.

An alternative approach, in an area without rain gauges, would be to initially install a large number of gauges in order to identify the predominant areal pattern of precipitation. Subsequently, "redundant" gauges that are not needed to define the average or the spatial patterns to the desired level of accuracy would be removed (ASCE, 1996). This method is likely to be expensive and time-consuming since the full network of temporary gauges may need to be operated for a considerable time in order to encompass a full range of precipitation conditions.

The areal distribution of gauges in an area should reflect the intended use for the network; thus, if the main purpose of precipitation measurement is for runoff studies, then more gauges should be deployed in those areas that contribute most to runoff. Similarly, if the gauges are to be used to calibrate a weather radar, then their distribution would be influenced by their position (quadrant and distance) relative to the radar.

Deriving Areal Average Values

A basic requirement in many hydrological applications, including water balance and rainfall–runoff studies, is to estimate the mean precipitation depth over a given area in a specified time period. This is often referred to as "areal rainfall". There are a large number of techniques that can be used to calculate areal rainfall from a network of often comparatively widely separated point measurements, and to refine the location and number of gauges required. Singh (1989) provides a detailed discussion of 15 different methods. These include polygonal weighting, inverse distance weighting, isohyetal, trend surface analysis, analysis of variance and kriging.

Whilst the accuracy of areal precipitation estimates will increase with the number of gauges in the network, a dense network is difficult and expensive to maintain. A number of general guidelines for gauge density have been produced. The World Meteorological Organization (Perks *et al.*, 1996) evaluated the adequacy of hydrological networks on a global basis for the *Basic Hydrological Network Assessment Project* and gave the following broad guidelines for the minimum gauge density of precipitation networks in various geographical regions: one rain gauge per 25 km^2 for small mountainous islands with irregular precipitation; 250 km^2 per gauge for mountainous areas; 575 km^2 elsewhere in temperate, Mediterranean, and tropical climates, and 10 000 km^2 for arid and polar climates.

Comparable actual values of the average area (km^2) per storage rain gauge for a range of countries include France (110), Netherlands (130), China (470), India (790), Australia (1010), United States (1040), and Mongolia (47 420). It is, however, salutary to note that even for the United Kingdom, which has one of the highest densities of storage rain gauges in the world – an average of one gauge per 60 km^2 (WMO, 1995) – the total collecting area of all its rain gauges combined is less than one standard football pitch!

1. Arithmetic mean. The simplest technique to calculate areal precipitation is to calculate the arithmetic mean of all the rain gauge totals within the area of interest. This assumes that each gauge is representative of an equal portion of the total area. This may be satisfactory in regions of flat topography with little systematic variation in precipitation and a uniform distribution of rain gauges. Such conditions are not generally found in practice, and there is often a tendency for the distribution of gauges to be in clusters that mirror human populations, which results in the potential for biased estimates (Figure 1). Thus, gauges are often most frequent in river valleys and most widely spaced in mountainous areas where precipitation depths are greatest, and the spatial variability is typically largest due to the influence of altitude and slope orientation.

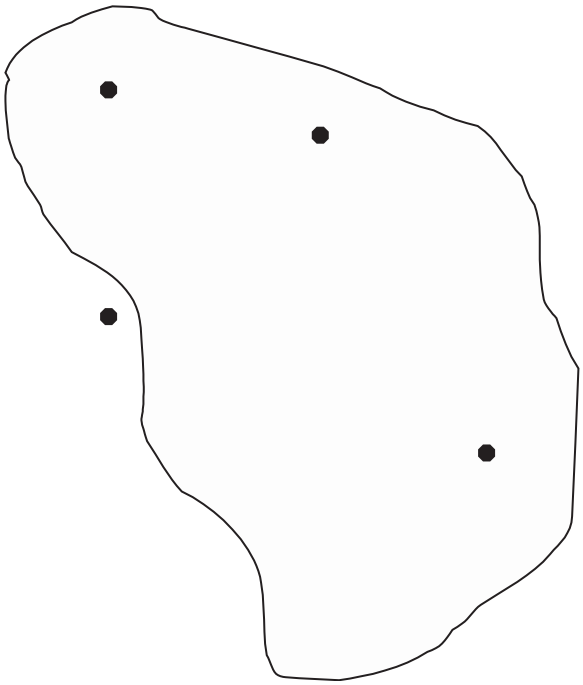


Figure 1 The areal precipitation calculated from arithmetic mean of this uneven spatial distribution of gauges could be seriously biased if there was a strong rainfall gradient across the area. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2. Thiessen. This method is suited to flat areas where each gauge is assumed to be the best measure of the fall on all of the land that is closer to it than to any other gauge. The catch at each gauge is weighted by the proportion of the total basin that it represents to calculate the basin mean rainfall. The Thiessen polygon method (Thiessen, 1911) has been widely adopted as a better method for calculating areal depths than the arithmetic mean. It allows for a nonuniform distribution of gauges by assigning "weights" to the measured depths at each gauge according to the proportion of the catchment area that is nearest to that gauge (Figure 2). The individual weights are multiplied by the gauge reading and the values are summed to obtain the areal mean precipitation. The method may be carried out graphically, or can be programmed for computer application by superimposing a regular grid over the area and allocating each grid point value to the nearest gauge. This assumes that each gauge is representative of its portion of the area, and the resulting precipitation surface is a series of polygonal plateaus with sharp steps between them. A modification is to allocate a region or "domain" to each gauge on the basis of physical factors such as local meteorological conditions and topography, thought likely to influence precipitation. Voronoi interpolation (Gold, 1989) is a development of the Thiessen method and has the advantage for hydrological applications that it

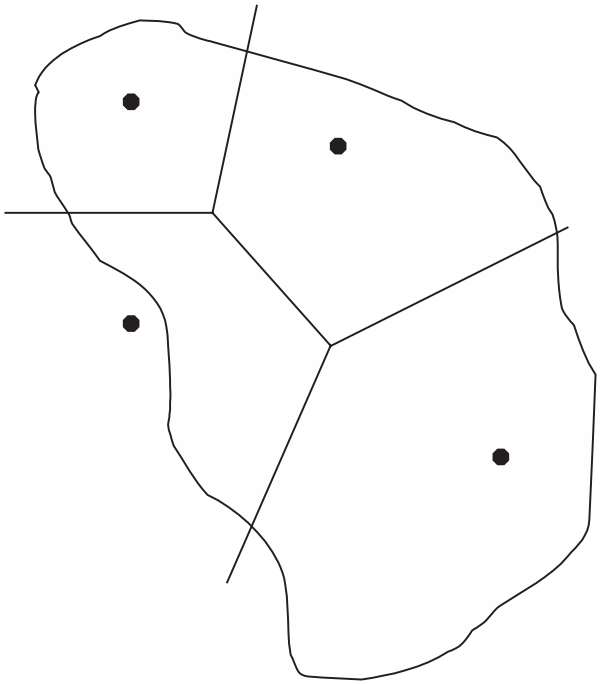


Figure 2 Thiessen polygons allow rain gauge catches to be weighted by the size of area they are nearest to. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

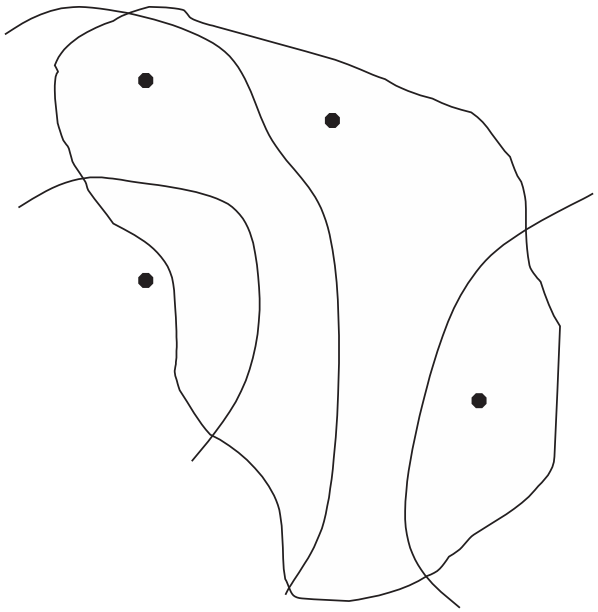


Figure 3 Isohyets between point gauge values allow other qualitative factors to be considered. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

produces a much more realistic surface (BSI, 1996). Neither method allows for precipitation gradients with altitude, slope orientation, and aspect. Their results can be strongly influenced by the positions of the gauges and where they are very unevenly distributed, quite small changes in gauge location may result in large changes in the weighting factors.

3. Isohyetal. This is a graphical technique, which enables topographic effects and any other subjective or "local" knowledge about meteorological patterns to be incorporated. It is probably the most reliable, but subjective, of the standard methods of areal precipitation calculation. It involves drawing estimated lines of equal precipitation (isohyets) on a map between gauges, making allowance for factors such as topography and distance from the sea. It can incorporate local knowledge and experience of factors such as windward and leeward effects related to the prevailing wind direction (e.g. Chater and Sturman, 1998). Areal precipitation is then computed by calculating the areas between the isohyets (Figure 3). This method thus uses all the data and knowledge about precipitation patterns in an area, but the results are subjective and can involve a considerable amount of time to construct the maps.

4. Geostatistical methods. The accuracy of the estimation at each point or grid will be a function of the distance from the nearest gauges. A large number of studies have used techniques including multiple regression and kriging to quantify the statistical structure of precipitation patterns, and to identify a more effective network design (e.g. Periago *et al.*, 1998). The simplest technique is to use an inverse distance weighting of gauge values to each point. In this method, weights are calculated depending on the distances between the location where an estimate is required and the location where the rainfall are measured (Figure 4). For each grid point, weights are assigned to each gauge value based on the distance from the grid point. In using any method based on the correlation between gauges, it is of paramount importance to ensure the homogeneity of record of each gauge. Unrecorded changes in siting or exposure may weaken correlations between gauges, resulting in networks that are denser than necessary. Kriging is a statistical method that uses the variogram of the precipitation field (i.e. the variance between pairs of points at different distances apart) to estimate interpolated values. The resulting precipitation field is optimal in the sense of identifying gauge weightings to minimize the estimation error. It has the advantage that it can be used to generate a map of the standard error of the estimates that indicates where additional gauges would be of most benefit. Bastin *et al.* (1984) use the technique to estimate areal rainfall, to indicate the degree of redundancy in a gauge network and to identify locations where the deployment of additional gauges might be most useful.

Different methods for estimating mean areal rainfall have been used in different studies. The selection of

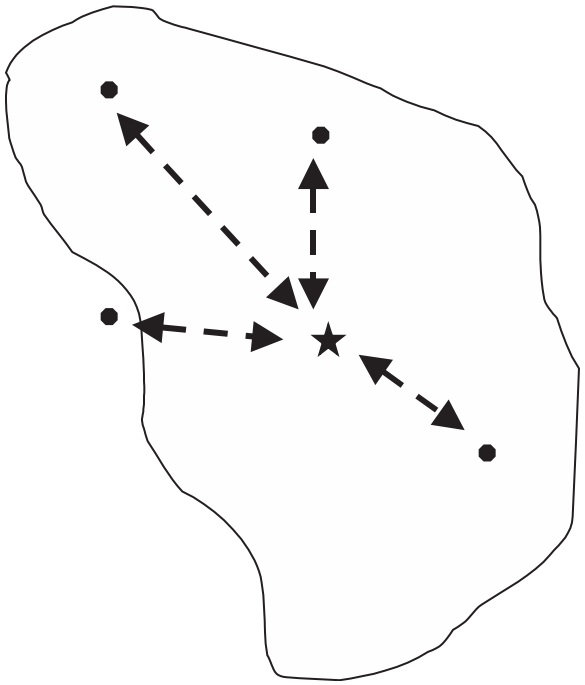


Figure 4 Estimating precipitation at a point by weighting gauge catches by their distance from that point. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the most appropriate one for a particular problem will depend upon a number of factors, including: the time available and the expertise of the hydrologist, the density of the gauge network, and the spatial variability of the precipitation. In general, the accuracy of all the methods for estimating areal rainfall will increase with: (a) the density of gauges, (b) the length of period considered, and (c) the size of area. Comparisons of the methods have perhaps not surprisingly shown smaller differences where time interval selected is long and the topography of the area is subdued (Singh, 1989).

Rain gauge Deployment – Local Site Factors

In addition to considerations of the broad regional pattern of gauge locations, it is necessary to ensure that each point gauge site should give an unbiased catch that avoids purely local (small scale) effects and so is representative of a larger area. The detailed site conditions will be paramount in determining their actual locations. There are a number of sources of error in catch, but by the far the most important is usually wind turbulence across the gauge resulting in underestimation of the true value. Measures that may be taken to reduce wind effects include the design of rain gauge used and the choice of location. Aspects of rain gauge design that may affect its susceptibility to catch errors are discussed separately (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**).

The siting and exposure of a rain gauge is very important for obtaining accurate measurements. Sites should be chosen to minimize the effects of wind on catch. Gauges should not be located close to objects and obstructions that would distort local wind flow patterns; nor should they be so exposed that the gauge catch may be seriously reduced when the wind is strong. The best location is where the gauge is uniformly protected in all directions, by objects that are at a distance of at least twice their height above the gauge (Meteorological Office, 1982). In some open situations exposed to strong winds, it may be necessary to construct a low turf wall around the gauge with a sloping outer edge and a flat crest on the same plane as the gauge rim (see Figure 5).

Rain gauge measurements may be crucially sensitive to the immediate environment surrounding the gauge. As a general rule, the windier the gauge location is, the greater the precipitation error will be. Rodda and Smith (1986) examined the records for 43 sites across the United Kingdom where a standard gauge (rim height 30 cm) operated alongside a ground-level gauge. Record lengths varied between sites but were generally of at least 10 years duration. The long-term percentage error of the standard gauges (assuming the ground-level gauge to represent the "true" rainfall) was generally an average of about -3 to -6%, but at some exposed upland sites this figure rose to -15%, with considerably higher errors over shorter periods and in individual storms. There was a strong correlation between average wind speeds and gauge error at the different sites. Price (1999) found that comparable gauges in upland Scotland had a systematic undercatch of about 10%.

Figure 5 The construction of a low turf wall around the gauge with a crest on the same plane as the gauge rim can help reduce undercatches. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

General advice on rain gauge siting dates back over a century (Symons, 1864; Mills, 1901) and remains broadly the same today in texts such as Meteorological Office (1982); NWS (1989); WMO (1994). A site should not be overexposed to strong winds, nor should it be unduly sheltered by nearby obstacles. Gauges should be positioned in a flat area, away from any obstructions that might cause air turbulence and consequent nonuniform deposition of rain droplets. As a general rule, the gauge should be at a distance of at least twice (and preferably four times) the height of any obstacle. But these guides also recognize that some degree of shelter is needed, since a very open site with no shelter would be too exposed and subject to excessive wind effects. Where this is difficult to achieve, providing good exposure to the most common directions of wind should be the priority.

In practice, there may be considerable constraints placed on those locations where rain gauges can be deployed. Some landowners may refuse to allow gauges on their land, whilst others may place severe restrictions – for example, in farmed areas that gauges are put at the edge of fields rather than in the center. Pressure for gauges to be sited close to the office of the organization responsible for the readings may result in poor exposure (see Figure 6a, b).

Consideration of wind effects on gauge catch is of particular importance when interpreting precipitation records from less than ideal sites. In the eighteenth and nineteenth centuries, rain gauges were often placed on a roof or a high wall to be safe from human or animal interference, thus inadvertently increasing the potential undercatch. This practice is still common today in many developing countries (Farquharson, 2003), and particular care must be made when interpreting such records. For many urban and suburban areas, the only sites available for rain gauges are on rooftops (see Figure 7).

These considerations demonstrate the crucial importance of site visits and metadata – records of site changes, which include instrumentation, and site exposure. These data may be crucial to be able to separate true climate trends from local site effects. Good exposures are not always permanent. Manmade alterations to the area and the growth of vegetation may change an excellent exposure to an unsatisfactory one in a very short time, necessitating the moving of precipitation gauges to sites having better exposures. In such a situation, it is not desirable to move a badly placed gauge immediately, as that would break the long-term homogeneity of the record. The better plan is to place the new gauge nearby in a more suitable location and to observe both gauges for a minimum of a year so that a correlation can be obtained and a correction may be applied to the old observations.

Despite the availability of international guidelines for rain gauge exposure, local and national standards vary widely. In an attempt to address the crucial problem of site exposure, Sevruk and Zahlavova (1994) attempted to devise an objective and quantitative measure relating to the angle above the horizontal from the gauge rim to the surrounding obstacles in eight compass directions. This may be calculated using a fish-eye lens and a digitizer, or alternatively from site photographs, sketches and station reports. De Smedt *et al.* (2003) found this approach was superior to several simpler alternative methods. Unfortunately, the potential use of this approach to correct rain gauge errors is limited as the necessary measurements are available at only a few sites (Ungersböck *et al.*, 2000).

(a)

(b)

Figure 6 In some areas gauges may be located close to the home or office of the person responsible for the readings, at sites with less than ideal exposure. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 7 In urban and suburban areas, the only available sites for rain gauges may be on rooftops as shown in the picture. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Special Site Considerations

There are particular circumstances where the measurement of precipitation poses special problems: these are where the topography is very steep and uneven, where the ground is covered by a continuous expanse of forest, where the precipitation falls as a fine mist or as snow.

In very steep terrain, not only is rainfall spatially very variable but a standard gauge with horizontal rim will tend to undercatch precipitation in winds blowing upslope, and overcatch in winds blowing downslope (Hamilton, 1954). The best solution may be a ground-level gauge with its rim inclined parallel with the ground slope. The effective catch area must then be converted to a horizontal plan area by dividing by the cosine of the slope angle.

About one-third of the Earth's land surface is covered by forests. It is not always possible to measure precipitation in clearings and it may be necessary to install a rain gauge on a tower at tree top (canopy) level. It is possible for catches very similar to ground-level gauge values to be achieved in this way because the airflow disturbance of the gauge is similar to the roughness of the forest canopy, but this is very dependent upon the height of the gauge relative to the canopy. If the gauge is too high it will experience wind-induced undercatch, whilst if it is too low it may suffer from overcatch as a result of drip from adjacent branches or undercatch because of sheltering. It is difficult to define a suitable height for the gauge rim where the trees are of irregular height or the topography is uneven. Furthermore, the level of the gauge must also be regularly raised in-line with growth of the trees (Robinson *et al.*, 2004).

In some parts of the world, appreciable quantities of water may be deposited by fine water droplets carried in the wind impacting on vegetation surfaces (*see* **Chapter 38, Fog as a Hydrologic Input, Volume 1**). This will not be recorded in conventional rain gauges, and so it may be necessary to deploy special cloud water collectors in order to determine the water content of these clouds and then to theoretically calculate their likely rate of deposition onto vegetation surfaces according to their respective aerodynamic roughnesses.

Snow

Information on the spatial distribution of snow was traditionally based on reports from observers at meteorological stations. The difficulties in measuring the amount of snowfall in gauges are even greater than those of rainfall. Snowflakes are even more prone than raindrops to turbulence around gauges, resulting in severe undercatches. Although wind effects can be greatly reduced by using windshields around the gauges, the errors due to undercatches are often still too great to be acceptable (Sevruk, 1982). Windshields may be used to minimize the loss of precipitation. This loss is much greater during snowfall than rainfall, so shields are seldom installed at gauges unless at least 20% of the annual precipitation falls as snow. In areas where heavy snowfall occurs (e.g. mountainous areas in the western USA), gauges may be mounted on towers at a height considerably above the maximum level to which snow accumulates (see Figure 8).

The WMO initiated a comparison of the catches of some of the most widely used precipitation gauges with those of a *Double Fence Intercomparison Reference gauge* (DFIR) comprising a *Tretyakov* gauge within two concentric fence shields (Goodison *et al.*, 1989). Undercatches by the standard gauges relative to the DFIR increased from only a few percent for rain in light winds, up to 50% or more for snowfall in strong winds. Eventually it may be possible to derive correction procedures to reduce the catch errors of standard gauges, but this may require detailed information including wind speed and the discrimination between "solid" and "liquid" precipitation.

Figure 8 In areas of heavy snowfall, gauges may have to be mounted on towers above. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

In many cases, it may be better to measure the depth of snow lying at particular locations. This may be converted to the snow water equivalent using the density of the snow. The density of freshly fallen snow varies from 50 to 200 kg m^{-3}, depending on the temperature during the storm. This increases over time due to settling and compaction as well as to any partial melting and refreezing of the snow pack, and may reach values of up to 600 kg m^{-3}.

The depth of snow may vary greatly with topography and with drifting, resulting in spatial sampling problems that are even more severe than those already discussed for rainfall. For this reason, depth and density measurements may be made along predetermined snow courses, selected to be representative of conditions over a wide area.

Acknowledgments

Many thanks to Dr Howard R. Oliver for information about historical studies of rainfall measurement.

REFERENCES

ASCE (1996) *Hydrology Handbook*, *Second Edition*, American Society of Civil Engineers: New York.

Bastin G., Lovert B., Duque C. and Gevers M. (1984) Optimum estimation of the average areal rainfall and optimal selection of rain gauge locations. *Water Resources Research*, **20**, 463–470.

Biswas A.K. (1970) *History of Hydrology*, North–Holland Publishing Company: Amsterdam, p. 336.

BSI (1996) *Guide to the Acquisition and Management of Meteorological Precipitation Data: Areal Rainfall*, British Standard 7843: Section 2.4, British Standards Institution: London.

Chater A.M. and Sturman A.P. (1998) Atmospheric conditions influencing the spillover of rainfall to lee of the Southern Alps, New Zealand. *International Journal of Climatology*, **18**, 77–92.

De Smedt B., Mohymont B. and Demarée G.R. (2003) Grubbs revisited: A statistical technique to differentiate measurement methods of the degree of exposure of rain gauges in a network *Hydrological Sciences Journal*, **48**, 1–7.

Farquharson F.A.K. (2003) *Personal Communication*, Head of Water Resources, Centre for Ecology and Hydrology: Wallingford.

Gold C.G. (1989) Surface interpolation, spatial adjacency and GIS. In *Three Dimensional Applications in Geographical Information Systems*, Raper J. (Ed.), Taylor and Francis, pp. 21–35.

Goodison B.E., Sevruk B. and Klemm S. (1989) WMO solid precipitation measurement intercomparison: objectives, methodology, analysis. *International Association of Hydrological Sciences Publication*, **179**, 57–64.

Gunston H.M. (1998) *Field Hydrology in Tropical Countries – A Practical Introduction*, Intermediate Technology Publications: London, p. 110.

Hamilton E.L. (1954) *Rainfall Sampling on Rugged Terrain*,*Technical Bulletin No 1096*, US Department of Agriculture: Washington, p. 41.

Meteorological Office (1982) *Observer's Handbook*, *Fifth Edition*, HMSO: London, p. 220.

Meteorological Office (2001) *Rules for Rainfall Observers*, Meteorological Office, Bracknell, p. 12.

Mills H.R. (1901) The development of rainfall measurement in the last 40 years. *British Rainfall 1900*, Edward Stanford: London, pp. 23–39.

NIH (1990) *Hydrology in Ancient India*, National Institute of Hydrology: Roorkee, p. 103.

NWS (1989) *Observing Handbook* Number 2, National Weather Service: Silver Spring.

Periago M.C., Lana X., Fernandez Mills G. and Serra C. (1998) Optimisation of the pluviometric network of Catalonia (North East Spain) for climatological studies. *International Journal of Climatology*, **18**, 183–198.

Perks A., Winkler T. and Stewart B. (1996) *The Adequacy of Hydrological Networks: a Global Assessment, Technical reports in Hydrology and Water Resources, 52*, World Meteorological Organisation: Geneva, p. 56.

Price D. (1999) Systematic error of standard UK rain gauges in the central Scottish highlands. *Weather*, **54**, 334–341.

Robinson M., Grant S. and Hudson J. (2004) Measuring rainfall to a forest canopy: an assessment of the performance of canopy level raingauges. *Hydrology and Earth System Sciences*, **8**, 327–333.

Rodda J.C. and Smith W. (1986) The significance of the systematic error in rainfall measurement for assessing wet deposition. *Atmospheric Environment*, **20**, 1059–1064.

Sevruk B. (1982) *Methods of Correction for Systematic Error in Point Precipitation Measurement for Operational Use*, Operational Hydrology Report No 21, World Meteorological Organisation: Geneva, p. 91.

Sevruk B. and Zahlavova L. (1994) Classification system of precipitation gauge site exposure: evaluation and application. *International Journal of Climatology*, **14**, 681–689.

Singh, V.P. (1989) *Hydrologic Systems II: Watershed Modelling*, Prentice Hall: New Jersey, p. 320.

Symons G.J. (1864) Raingauges and hints on observing them. *British Rainfall, 1863*, Edward Stanford: London, pp. 8–12.

Thiessen A.H. (1911) Precipitation averages for large areas. *Monthly Weather Review*, **39**, 1082–1084.

Ungersböck M., Rubel F., Fuchs T. and Rudolf B. (2000) Bias correction of global daily raingauge measurements. *Physics and Chemistry of the Earth*, **26**, 411–414.

WMO (1994) *Guide to Hydrological Practices*, Report No 168, *Fifth Edition*, World Meteorological Organization: Geneva, p. 735.

WMO (1995) *INFOHYDRO Manual*, Operational Hydrology Report No 28, World Meteorological Organization: Geneva.

37: Rainfall Trend Analysis: Return Period

ROBIN T CLARKE

Instituto de Pesquisas Hidráulicas, Porto Alegre RS, Brazil

This article describes the concept of return period, with particular reference to the analysis of rainfall records. Methods are described for calculating return levels, corresponding to different return probabilities, when rainfall is accumulated over different time intervals. In the case of intense, short-duration rainfalls, threshold models are illustrated. The article points out that the concept of return period is most easily interpreted where stationarity in the rainfall regime can be assumed, so that it is important to establish whether rainfall records show evidence of trends over time. Procedures are discussed for testing whether time trends exist in annual and monthly rainfall, in the occurrence and magnitude of daily rainfalls, and in annual maximum rainfall intensities over different durations. Where a trend in rainfall regime is apparent, an alternative to the concept of return period is suggested, on the basis of the probability of occurrence of an extreme event of specified magnitude, during a period extending into the future during which the observed trend can be assumed to continue.

INTRODUCTION: DEFINITION OF RETURN PERIOD

In the context of this article, the term *return period* refers to the frequency with which a given characteristic of rainfall will occur, over a very long period extending into the indefinite future. The return period is commonly measured in years. In the absence of changing hydrological regime, an event with return period T years is also the event that may occur in any one particular year with probability $1/T$.

To develop the idea of return period, consider the annual rainfall X at a particular measurement site. This varies unpredictably from one year to the next, and observations of the random variable X over a sequence of N years will give recorded values $x_1, x_2, x_3, \ldots x_N$, or $\{x_i\}$, $i = 1, 2, \ldots N$. In many cases, it is reasonable to assume that while the values $\{x_i\}$ vary throughout the period in an unpredictable way, the statistical characteristics of this sequence (such as the mean value μ about which they vary) remain unchanged over time: that is, the observed values are a *realization* of a *stationary process.* In the case of some rainfall characteristics, such as annual total rainfall, physical considerations may suggest that the rain falling in year t is statistically independent of the rain that fell in the preceding year $t-1$, since atmospheric conditions may change rapidly over periods as short as even a few days. By contrast, in the case of rain falling on successive days, meteorological conditions may persist from one day to the next, so that the rainfalls measured on days $t-1$ and t cannot be regarded as statistically independent. This article begins by concentrating on the simple case, like that of annual total rainfall ("annual rainfall", for the sake of brevity), in which the underlying statistical process is (i) stationary, with (ii) observations statistically independent.

By assuming that the mean μ and other parameters defining the statistical characteristics of rainfall are constant, the variability amongst the observed annual rainfalls $\{x_i\}$ may be expressed in terms of a probability density function $f_X(x)$, the form of which is sometimes suggested by theoretical considerations, but more often is taken as an empirical function that satisfactorily describes the characteristics of the observed sequence $\{x_i\}$. Then, in a very long period of rainfall measurement, the proportion of years with annual rainfall lying within the short interval $[x, x + \Delta x]$ is, with very small error, $f_X(x).\Delta x$. The same idea may be expressed in a slightly different way. Since $f_X(x)$ is a probability density function (pdf), the probability that the total rainfall X in any year is equal to or greater than some given value x_0 is

$$P[X \geq x_0] = \int_{x_0}^{\infty} f_X(u)\,\mathrm{d}u \qquad (1)$$

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

where u is any dummy variable, and $P[.]$ denotes the probability of the event in square brackets. In the left-hand side (LHS) of equation (1), x_0 is given, and the quantity on the right-hand side (RHS) is calculated from it using the known properties of $f_X(x)$. But it is equally possible to define the magnitude of the probability on the RHS, say $1/T$, where T is measured in years, and then use an inverse calculation to estimate the rainfall x_0. Thus, the equation to be solved for x_0 (commonly referred to as the *return level*) is

$$\int_{x_0}^{\infty} f_X(u)\,\mathrm{d}u = \frac{1}{T} \tag{2}$$

If, in equation (2), the value of T is set at 100, then $1/T = 0.01$ and the value of x_0 is the annual rainfall with 100-year return period: that is, the rainfall that in any randomly selected year would be exceeded with probability $1/100 = 0.01$. Stated in another way, in a very long period of rainfall record, the annual rainfall would equal or exceed x_0 in a proportion 1/100 years. Stated in yet another way, the annual rainfall will be greater than x_0 once in 100 years, in the long run. Obviously, T may be set to any value, but in practice extreme events (that is events with long return period, i.e., large T) are of more interest. Trivially, if $T = 1$, equation (2) immediately gives $x_0 = 0$, showing that the annual rainfall with 1-year return period is 0 mm. Annual rainfall, by definition, must be nonnegative.

An important phrase in the preceding paragraph is "in the long run". Because annual rainfall varies randomly, it does not follow that in any single consecutive 100-year sequence, the annual rainfall with 100-year return period will occur exactly once. Indeed, it is easy to show that in a period of 100 years, the probability that the 100-year rainfall will occur once *or more* is high, 0.634 or 63%, even though the expected number of occurrences of the T-year rainfall during this period is 1.

Solution of equation (2) gives the value of annual rainfall that is exceeded once in T years in the long run. It is equally possible to work with the lower "tail" of the pdf $f_X(x)$, and in an arid region this is likely to be of greater interest. Thus, once in T years in the long run, annual rainfall will be less than x_0, where x_0 is now given by

$$\int_{u=0}^{x_0} f_X(u)\,\mathrm{d}u = \frac{1}{T} \tag{3}$$

Furthermore, although we have taken the annual rainfall X as an example, the same procedure can be followed for many other rainfall characteristics, of which the following are just a few examples: annual maximum rainfall during a 24-h period; depth of rain falling in the month prior to a crop-planting date; annual maximum rainfall intensity of 15-min duration; length of period of consecutive days without rain in any year. For each example, the form of $f_X(x)$ will be different, but the procedure for calculating the event with T-year return period is basically as set out in equations (2) and (3) above. In each case, too, the units of the calculated x_0 will be those of the data sequence: millimeters in the case of annual rainfall, days in the case of duration of longest dry period, millimeters per hour ($\mathrm{mm\,h^{-1}}$) in the case of rainfall intensities over different durations.

THE PROBABILITY DENSITY FUNCTION $f_X(x)$

It will be clear that the numerical value obtained for the rainfall event x_0 with T-year return period will depend on the function $f_X(x)$ used in equations (2) and (3), and the question arises: How is $f_X(x)$ to be obtained? Three points are relevant to this question.

The (rare) Case where x_0 can be Calculated without Explicit Knowledge of $f_X(x)$

Suppose that the recorded rainfall sequence from N years of record is $\{x_i\}$, $i = 1, 2 \ldots N$, so that $x_1, x_2, \ldots x_N$ are the annual rainfalls (or maximum intensities) in years $1, 2, \ldots N$ of the sequence. Then, provided that (i) $T < N + 1$ and (ii) the values $x_1, x_2, \ldots$ are statistically independent, it may not be necessary to define the function $f_X(x)$, because the following argument can be used. Put the values $\{x_i\}$ in increasing order of magnitude, to give the sequence $x_{(1)}, x_{(2)} \ldots x_{(N)}$, where $x_{(1)}, x_{(N)}$ are the smallest and largest values respectively in the sequence. Then, the probability that (say) annual rainfall in any randomly selected year is greater than $x_{(N)}$ is $1/(N + 1)$, and the probability that it is less than $x_{(1)}$ is also $1/(N + 1)$. Equally, the probability that annual rainfall is greater (less) than $x_{(N-1)}$ ($x_{(2)}$) is $2/(N + 1)$, and so on. By interpolation, it is therefore possible to estimate the annual rainfall x_0 with T-year return period, without specifying $f_X(x)$. However, in many cases, T will be larger than $N + 1$, and the method fails. It then becomes necessary to specify a mathematical form for the distribution $f_X(x)$. This may be indicated by theoretical considerations or, more commonly, by empirical knowledge of what has proved appropriate in the past. In either case, $f_X(x)$ will almost certainly contain *parameters* that must be estimated from whatever rainfall record is available: hence, it is appropriate to write $f_X(x)$ in a more general form $f_X(x, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots \theta_m]$ is the set of m parameters defining the statistical characteristics (e.g., mean value, dispersion about the mean value, skewness, etc.) of the pdf. In many cases, distributions with $m = 2$ or 3 are adequate.

Estimation of $f_X(x)$ by Means of a Kernel Function

It is possible to calculate a form of distribution representing the data using a kernel estimate. Kernel density estimation

is a useful tool for exploring the unknown underlying distribution of a sample; Silverman (1986) gives a general introduction to density estimation by kernel methods, which can be regarded as a formal development of the simple histogram. The kernel method constructs an estimate $f_X(x)$ of the true density function by placing a kernel function $K(x; x_i, h)$ over each observation x_i in the sample $\{x_i\}$, $i = 1, 2, \ldots N$. The kernel function $K(x; x_i, h)$ is a density function with location parameter x_i and scale parameter h, also called *bandwidth* in this context. The density estimate is then given by

$$f_X(x) = \left(\frac{1}{Nh}\right) \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

where N is the sample size. It turns out that the choice of kernel function K is not very critical for the resulting estimate $f_X(x)$ (see Silverman (1986)). The Gaussian kernel is commonly used and is therefore adopted here as kernel function, that is,

$$K(u) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-u^2}{2}\right) \tag{5}$$

For example, Table 1 shows the annual rainfall for the years 1976–2002 at Fortaleza in the Brazilian Northeast. Over the 27-year period, the annual rainfall X varied from 978 mm in 1990 to 2836 mm in 1985. Figure 1 shows the distribution $f_X(x)$ fitted using equation (4); using $T =$ 100 years in equation (2) gives the annual rainfall x_0 with 100-year return period as 3032 mm. Using equation (3), the annual rainfall exceeded in 99 years out of 100, in the long run, is 517 mm. Both the values lie outside the range of observed rainfalls shown in Table 1. The two values are the 1% and 99% quantiles of the pdf. The bandwidth h used in this example was proportional to the standard deviation of the data; the choice of bandwidth requires care, since if h is small, the fitted distribution will be multimodal, while if it is too large, the fitted distribution is oversmoothed.

Table 1 Annual rainfall (mm) at Fortaleza, Northeast Brazil, 1976–2002

Year	1976	1977	1978	1979	1980
P	1490	2020	1557	1191	1216
Year	1981	1982	1983	1984	1985
P	1086	1051	955	2029	2836
Year	1986	1987	1988	1989	1990
P	2457	1260	1862	1862	978
Year	1991	1992	1993	1994	1995
P	1549	1089	1043	2380	2144
Year	1996	1997	1998	1999	2000
P	1708	1143	1012	1347	1673
Year	2001	2002			
P	1554	1742			

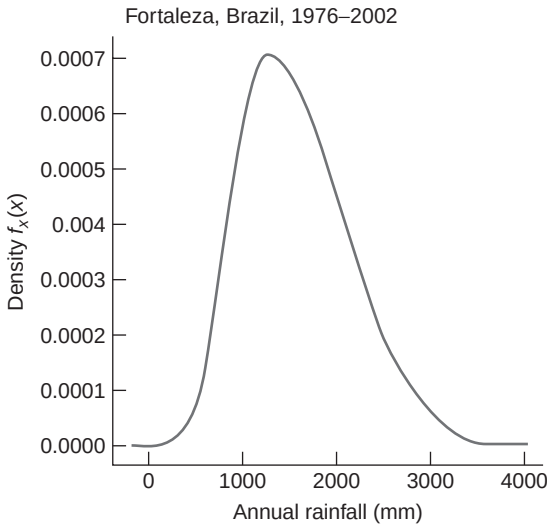


Figure 1 Probability distribution fitted to Fortaleza annual rainfall for period 1976–2002. Distribution was fitted by the kernel method, using a Normal distribution as kernel and bandwidth proportional to standard deviation. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Use of an Explicit Form for $f_X(x)$

Continuing the example using annual rainfall data, a much more common alternative to estimating the pdf by the kernel method is to select a pdf of some known form, with characteristics determined by parameters that are estimated using relevant data in the form of rainfall records. A histogram (not shown) of the Fortaleza data in Table 1 shows that the distribution of annual rainfall has positive skew, suggesting that a Normal (Gaussian) distribution may provide a good fit if the data are first transformed by taking natural logarithms. This is confirmed by Figure 2, which shows a plot of the log of the ordered sample $x_{(1)}, x_{(2)} \ldots x_{(N)}$ (vertical axis) against the quantiles of the standard Normal distribution N(0,1). All 27 points lie within the confidence region shown in the figure, so that $\ln(x_i)$ can be represented by a Normal distribution, whose estimated mean and standard deviation are $\hat{\mu} = 7.3073 \pm 0.0590$ and $\hat{\sigma} = 0.3065 \pm 0.0417$ (the "hats" indicating estimates, obtained from the data, of the relevant parameters μ and σ). Equation (2) then becomes

$$\frac{1}{\{\hat{\sigma}\sqrt{(2\pi)}\}} \int_{x_0}^{\infty} \exp\left[-\frac{(\ln(x) - \hat{\mu})^2}{2\hat{\sigma}^2}\right] \frac{\mathrm{d}x}{x} = \frac{1}{T} \tag{6a}$$

or

$$\Phi\left(\frac{\ln(x_0) - \hat{\mu}}{\hat{\sigma}}\right) = 1 - \frac{1}{T} \tag{6b}$$

where $\Phi(z)$ denotes the cumulative probability $\mathrm{P}[Z < z]$ for the standard N(0,1) distribution. For the annual rainfall

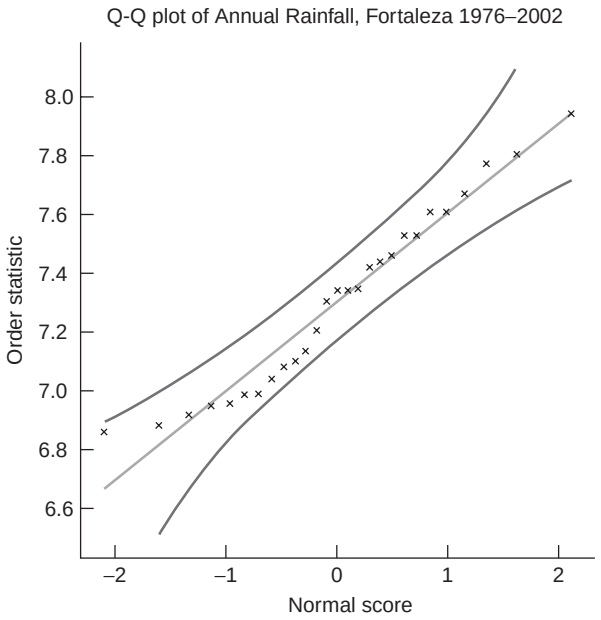


Figure 2 Quantile–Quantile plot, showing goodness of fit between ln(x) and a Normal distribution, where x is annual rainfall at Fortaleza, Northeast Brazil. ''Normal score'' shows the expected values of order statistics from the standardized N(0,1) distribution; ''Order statistic'' shows the 27 values of ln(x) in ascending order. The overall probability of the plotted data lying completely within the confidence bands is approximately 95% under the null hypothesis that ln(x) is a random sample from the normal distribution. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

with $T = 100$-year return period, it is found from a table of Normal deviates that $[\ln(x_0) - \hat{\mu}]/\hat{\sigma} = 2.326$, so that x_0 is estimated as $\exp[\hat{\mu} + 2.326\hat{\sigma}] = \exp(7.3073 + 2.326 \times 0.3065) = 3042$ mm. This compares with the estimate $x_0 = 3032$ mm found by the kernel method described above.

Thus, we have considered three ways of estimating the value a random variable that corresponds to a return period of T years. In the approach described above, the pdf was found to be of simple form after transformation of the data, but in very many cases, distributions other than the Normal will be more appropriate. If, for example, the data $x_1, x_2, x_3, \ldots x_N$ are the maximum daily rainfalls in each of the N years, then it is likely that the Gumbel distribution will be more appropriate; maintaining the notation $f_X(x)$ for the pdf of the random variable X that is under statistical analysis, the Gumbel distribution is

$$f_X(x, \boldsymbol{\theta}) = \sigma^{-1} \exp\left[-\frac{x-\mu}{\sigma} - \exp\left\{-\frac{x-\mu}{\sigma}\right\}\right] \quad -\infty < x < \infty \tag{7}$$

where in this case the two-element vector $\boldsymbol{\theta}$ of parameters is $\boldsymbol{\theta} = [\mu, \sigma]^T$. Putting $y = (x - \mu)/\sigma$, the cumulative distribution function (cdf) has the simple form

$$F_1(y) = \exp(-\exp[-y]) \qquad -\infty < y < \infty \tag{8}$$

which is the first of three max-stable (Cox *et al.*, 2002; Coles, 2001) limiting forms of cdf for maxima of samples of values drawn independently from a probability distribution, as the sample size increases indefinitely. The other two limiting forms

$$\begin{aligned} F_2(y) &= \exp(-y^{-\alpha}) \qquad y \geq 0 \\ &= 0 \qquad\qquad\quad y < 0 \end{aligned} \tag{9}$$

and

$$\begin{aligned} F_3(y) &= \exp(-(-y)^{\alpha}) \qquad y \leq 0 \\ &= 1 \qquad\qquad\quad y > 0 \end{aligned} \tag{10}$$

are the Fréchet and Weibull distributions. All three forms, known respectively as EVI, EVII, and EVIII, are particular cases of the generalized extreme-value (GEV) distribution

$$f_X(x) = \left(\frac{1}{\sigma}\right)\left[\frac{1 + \xi(x-\mu)}{\sigma}\right]^{-(1/\xi)-1} \times \exp\left\{-\left[\frac{1 + \xi(x-\mu)}{\sigma}\right]^{-1/\xi}\right\} \tag{11}$$

defined for x such that $1 + \xi(x - \mu)/\sigma > 0$, with parameters μ, σ, and ξ satisfying $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \xi < \infty$. The three special cases EVI (Gumbel), EVII (Fréchet), and EVIII (Weibull) correspond to the three cases $\xi = 0$, $\xi > 0$, and $\xi < 0$. Where the data consist of annual maximum rainfalls accumulated over time periods of a day or less, distributions such as equation (7) or equation (11) are appropriate for use in equation (2) for determining the value of x_0 with T-year return period.

There is a very large literature concerning procedures for estimating the parameters of distributions (such as equation 7, equation 11, and the log-Normal distribution used in equation 6a). One method with strong theoretical appeal is the method of maximum likelihood (Cox and Hinkley, 1974), which estimates the parameters $\boldsymbol{\theta}$ by maximizing the probability of the observed dataset $\{x_i\}$, $i = 1, 2 \ldots N$. However, a method that has come into widespread use in recent years (see, for example, Institute of Hydrology, 1999) is the method of L-moments, described by Hosking and Wallis (1997). In this method, linear combinations of the order statistics of the data – that is, linear combinations of the values $x_{(1)}, x_{(2)} \ldots x_{(N)}$ – are fitted to their equivalents calculated from the pdf, which are functions of the distribution parameters $\boldsymbol{\theta}$. Hosking *et al.*(1985) argue that in the relatively small data samples

provided by hydrological and meteorological records, the L-moment approach outperforms maximum likelihood in the estimation of GEV parameters (see equation 11) as judged by mean square errors. Cox *et al.*(2002) note that such comparisons depend on the particular choice of parameters, while Coles and Dixon (1999) contend that this apparent superiority is a result of a constraint placed upon the shape parameter by the L-moment approach. Cox *et al.*(2002) also note that "an attractive feature of likelihood-based methods is the flexibility afforded by the possibility of incorporating the effects of temporal and/or spatial non-stationarity and covariates." It is this flexibility that is exploited in later sections for the analysis of trends.

STATISTICAL METHODS FOR DETECTING TREND IN RAINFALL RECORDS

Much of the remainder of this article addresses the problem of testing whether a rainfall record shows evidence of time trend. Spatial trends are not considered, since there exists a very large literature on statistical procedures appropriate for the analysis of spatial trends (Cressie, 1993; Isaaks and Srivastava, 1989; Ripley, 1991). A useful reference on methods for detecting trends in hydrological data in general is Kundzewicz and Robson (2000).

Where it is required to test whether a time trend exists in a rainfall record, and to estimate its magnitude, the methods available depend on the time interval over which rainfall is accumulated. Four cases are considered in the text below: (i) trends in annual total rainfall; (ii) trends in monthly rainfall; (iii) trends in occurrence and/or depth of daily rainfall; and (iv) trends in annual maximum rainfall intensity of different durations, possibly 15 min, 1 h, 3 h, and so on. We note that where a time trend is shown to exist, the assumption that the random variable X varies about a constant mean μ becomes invalid, and the concept of return period becomes of questionable validity. If, for whatever reason, rainfall regime is in process of change, the parameters $\boldsymbol{\theta}$ of the pdf $f_X(x, \boldsymbol{\theta})$, and possibly even its mathematical form, are also changing. The phrase "in the long run" – necessary for defining a return period – then has little meaning. We return to this difficulty later.

Trends in Annual Rainfall Totals

Since annual rainfall is the sum of the daily rainfalls occurring through the year, a reasonable working hypothesis is that the random variable "annual rainfall" is approximately Normally distributed, by virtue of the central limit theorem. In most cases, it will also be reasonable to assume that annual rainfalls are serially independent: this means that the pdf of the random variable "annual rainfall in year t", denoted by X_t, satisfies $f_X(x_t|x_{t-1}) = f_X(x_t)$, showing that knowledge of x_{t-1}, the rain that fell in the preceding year, conveys no information about X_t. Given serial independence, standard linear regression methods can then be used to test for the existence of linear (quadratic, etc.) trend in a record. Where annual totals are not Normally distributed, and cannot be transformed to Normality by transformation (such as the log-transformation used above), nonparametric tests for trend are appropriate.

Where it is required to test for time trends in records from several sites within a region, it will be necessary to allow for the probable existence of spatial correlation between the annual total rainfall registered at neighboring sites. Multivariate regression can be used to test whether a homogeneous time trend exists over the region as a whole, or whether the time trend varies spatially. Thus, a multivariate model for k sites within a region could be proposed, of the form

$$\mathbf{X} = \mathbf{T}\boldsymbol{\beta} + \mathbf{E} \tag{12}$$

where $\mathbf{X}$ is a matrix of dimension $N \times k$ in which the tth row is the annual rainfall in the tth year of the N years of record at each of the k sites; $\mathbf{T}$ is a matrix of dimension $N \times (q+1)$ in which the tth row is $1, t, t^2, \ldots t^q$; $\boldsymbol{\beta}$ is a matrix of dimension $(q+1) \times k$, whose mth column is the vector of regression coefficients for the mth site, namely, $[\beta_0^{(m)}\ \beta_1^{(m)} \ldots \beta_{q+1}^{(m)}]^T$; and $\mathbf{E}$ is a $N \times k$ matrix of residuals, with rows that are independent and multivariate Normal $N_k(\mathbf{0}, \boldsymbol{\Sigma})$, possibly after transformation of the k rainfall records. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be estimated by maximum likelihood, and a likelihood-ratio test can be used to test whether the trend is homogeneous over all k sites. In this test, the null hypothesis is that the m columns of the matrix $\boldsymbol{\beta}$ are equal for all m. The model equation (12) fits polynomial functions of order q in time t at each site, and is best restricted to small values of q (≤ 2) if instabilities are to be avoided. Even if statistically significant polynomial trends are identified in records at the k sites, it would be extremely inadvisable to extrapolate such empirical relations far into the future: it is only possible to extrapolate with any degree of confidence when the physical processes giving rise to the trend are understood and have been expressed in mathematical terms.

Trends in Monthly Rainfall Totals

Various statistical tests for trend are possible, depending on the characteristics of the monthly rainfall record. Linear regression or nonparametric methods can be applied separately to rainfall in each month of the year, giving 12 separate significance tests; however, even if the 12 tests can be considered statistically independent, the probability of at least one apparently significant trend being detected at the 5% level of significance, even where no trend exists in reality, is high, nearly 46%. Erroneous conclusions can

be avoided by using the Bonferroni inequality (e.g., Johnson and Wichern, 1992), which reduces the Type I error in each month (i.e., the probability of concluding that significant trend exists in the month when it does not) from 5% to 5%/12 = 0.42%, the overall Type I error for all 12 tests remaining at 5%. The picture is further complicated if rainfalls in successive months are serially correlated (where, say, a sequence of months all record rainfall that is greater then the corresponding monthly averages) since the 12 tests are then not statistically independent. These complications apply whether the tests used are parametric tests (based on regression theory) or nonparametric (Mann–Kendall).

An alternative approach is to test for trend (in addition to seasonality) in the single sequence of $12N$ monthly rainfalls derived from the N years of record. This approach is useful when, as may often happen, records in some months are missing. Multiple regression models of the form

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_{k+1}\cos\left(\frac{2\pi t}{12}\right) + \beta_{k+2}\sin\left(\frac{2\pi t}{12}\right) + \cdots + \varepsilon_t \tag{13}$$

or

$$x_t = \boldsymbol{\beta}_{\text{tr}}^T \boldsymbol{f}(t) + \boldsymbol{\beta}_{\text{ha}}^T \boldsymbol{h}(t) + \varepsilon_t \tag{14}$$

with Normally and independently distributed residuals $\varepsilon_t \sim N(0, \sigma^2)$ are a point of departure; the vector $\boldsymbol{\beta}_{\text{tr}}^T$ is a vector of trend parameters multiplying a vector $\boldsymbol{f}(t)$ of known trend functions (such as polynomials in t) and $\boldsymbol{\beta}_{\text{ha}}^T$ is a vector of parameters multiplying a vector of known harmonic functions $\boldsymbol{h}(t)$ describing seasonality. A test of whether there is a trend superimposed on seasonality is given by a test of the hypothesis H_0: $\boldsymbol{\beta}_{\text{tr}} = 0$ against the alternative H_1: $\boldsymbol{\beta}_{\text{tr}}$ not all zero. However, where there is marked seasonality in rainfall distribution, it is unlikely that the residual variance σ^2 will be constant; this must be checked by calculating the variances, month by month, of the calculated residuals $\hat{\varepsilon}_t$, obtained after fitting the regression model equation (14). Checks will also be needed to explore whether the calculated residuals are serially correlated, which may occur if months of higher-than-usual rainfall tend to be followed by months in which rainfall is also higher than usual, and conversely. Where this occurs, the regression in equation (14) may need to be modified to allow for the serially correlated residuals. One such model might be:

$$x_t = \boldsymbol{\beta}_{\text{tr}}^T \boldsymbol{f}(t) + \boldsymbol{\beta}_{\text{ha}}^T \boldsymbol{h}(t) + \varepsilon_t \tag{15a}$$

$$\varepsilon_t - \phi_1 \varepsilon_{t-1} - \phi_2 \varepsilon_{t-2} - \phi_3 \varepsilon_{t-3} - \ldots = a_t \tag{15b}$$

where now the a_t are independently distributed N(0,σ_a^2). Extensions in which the residuals ε_t are of type ARMA (p, q) (see, e.g., Box and Jenkins, 1970) are possible, and can be easily fitted using available statistical software (S-plus; GenStat, SAS, and others).

Trends in Occurrence and Depth of Daily Rainfall

The characteristics of daily rainfall are (i) alternating sequences of dry days and wet days, the sequences being of variable length; (ii) the durations of the alternating periods of wet and dry days are determined by a random process, having a structure that must be identified from the daily rainfall record; (iii) the quantities of rain falling on wet days are determined by a random process, which must also be identified from data; and (iv) the random processes that govern rainfall occurrence and rainfall amount vary throughout the year. The hydrological literature has references to many models of daily rainfall; a typical model structure, adopted as a basis for discussion in the present paper, uses a two-state Markov model to describe rainfall occurrence (the two states 0 and 1 representing dry and wet days respectively, so that any record of daily can be represented by a sequence of 0's and 1's), and a skewed distribution, such as exponential or gamma, to describe depth of rain occurring on days when rain falls. As noted above, the 0's and 1's will rarely be independent because meteorological conditions may persist from one day to the next. A convenient way of representing this dependence is by means of a Markov chain, in which probabilities are defined as transitions from the states 0 to 0 (dry day followed by dry day), 0 to 1 (dry day followed by wet day), 1 to 0 (wet followed by dry), and 1 to 1 (wet followed by wet). The transition probabilities of the Markov process for rainfall occurrence, and the parameters of the probability distribution of rainfall depths, are time-variant, typically being represented in terms of harmonic functions $\cos(2\pi kt/365)$, $\sin(2\pi kt/365)$, $k = 1, 2, \ldots$ (Stern and Coe, 1984) to incorporate seasonality.

To fix ideas, suppose that the Markov model of rainfall occurrence is of first order with the four transition probabilities $P_{ij}(t), i, j = 0, 1$, so that $P_{01}(t)$, for example, is the probability that a dry day (0) on day $t-1$ is followed by a wet day (1) on day t of the year, for $t = 1, 2, \ldots 365$. Then since, for each day t, $P_{00}(t) + P_{01}(t) = 1$, $P_{10}(t) + P_{11}(t) = 1$, it is only necessary to model two of the four transition probabilities, say $P_{00}(t)$ and $P_{10}(t)$, typically using

$$\ln\left(\frac{P_{00}(t)}{1-P_{00}(t)}\right) = \boldsymbol{\beta}_1^T \boldsymbol{f} \tag{16a}$$

$$\ln\left(\frac{P_{10}(t)}{1-P_{10}(t)}\right) = \boldsymbol{\beta}_2^T \boldsymbol{f} \tag{16b}$$

where $\boldsymbol{\beta}_i^T = [\beta_{i0}\, \beta_{i1}\, \beta_{i2}\, \beta_{i3} \ldots]^T$, $\boldsymbol{f} = [1\, \cos(2\pi t/365),\, \sin(2\pi t/365),\, \cos(4\pi t/365),\, \sin(4\pi t/365) \ldots]^T$. This is the approach used by Coe and Stern (1982) and Stern and Coe

(1984). These authors modeled rainfall amounts $x(t)$ on day t by means of a gamma distribution with time-varying mean

$$f(x) = \left(\frac{\kappa}{\mu(t)}\right)^{\kappa} x^{\kappa-1} \exp\frac{[-\kappa x/\mu(t)]}{\Gamma(\kappa)} \qquad (17)$$

and the time dependence in the mean $\mu(t)$ was taken to be of the form $\ln(\mu(t)) = \boldsymbol{\beta}_{\mathrm{ha}}^{T}\ \boldsymbol{h}(t)$. Fitting models of the type (16a,b) for the occurrence of rainfall and of the type (17) for the depth of rainfall is straightforward using statistical software for generalized linear models (GLM; McCullagh and Nelder, 1989) and there are now many software packages (GLIM, GenStat, SAS, amongst others) that fit GLMs and provide the necessary diagnostics for assessing goodness of model fit.

Up to this point, the models in equations (16) and (17) model seasonality, but not trend, and the fitting procedures described by Stern and Coe (1984) treat years as replicates (that is, with N complete years of record, the probability $P_{10}(t)$, for example, is estimated by the proportion, out of the N "trials", in which rain on day $t-1$ of the years was followed by no rain on day t). McCullagh and Nelder (1989) point out that if time trends over the years are also present, it is important not to regard the years as replicates; instead, the N years of daily rainfall record are regarded as a single series of $365 \times N$ Bernoulli observations (with due allowance for leap years; see Stern and Coe, 1984) with values 0 or 1 for the absence or presence of rain, indexed by day, year, and previous day, either wet or dry. This increases the computational burden, but the procedure in which the models are fitted as GLMs remains essentially the same and models can still be comfortably fitted using modern PCs. Time trends in the quantity of rain falling can also be modeled by a similar extension to equation (17), with $\ln \mu(t)) = \boldsymbol{\beta}_{\mathrm{tr}}^{T}\boldsymbol{f}(t)$, the vector $\boldsymbol{f}(t)$ now being a vector of known functions describing trend.

As an example of the procedure suggested by McCullagh and Nelder (1989), we take a 42-year record of daily rainfall from the State of Ceará in the Brazilian Northeast. The record has 15 330 days, and a parallel record is constructed in which, for day t, 1 is entered if day $t-1$ was dry and day t was wet; otherwise, a 0 is entered; thus, the example considers trend in the transition probability P_{01}. (A similar calculation, not given here, could be made for the transition probability P_{11}.) To simplify matters, each year is taken to consist of 365 days, observations on 29 February in leap years being excluded. A parameter d_i $(i = 1 \ldots 365)$ is fitted for each day of the year, and a parameter y_j $(j = 1 \ldots 42)$ is fitted for each year. With $P_{ij}(t)$ representing the probability of a transition from states 0 to 1 between days $t-1$ and t, the model taken is

$$P_{ij}(t) = \frac{1}{1 + \exp[-(\mu + d_i + y_j)]} \qquad (18)$$

or

$$\ln\left[\frac{P_{ij}(t)}{1 - P_{ij}(t)}\right] = \mu + d_i + y_j, \quad i = 1 \ldots 365, \quad j = 1 \ldots 42 \qquad (19)$$

In this model (in which the expression on the LHS of equation 19 is termed a *logit*), the parameters $(\mu,\ d_i,\ y_j)$ were estimated with the results shown in Figure 3. This figure shows the changes in the estimates of the year effects y_j, relative to y_1, the effect in the first year of record; the upper and lower lines connected by asterisks show approximate 95% confidence limits for the changes. The upper and lower limits bracket zero for all years, indicating no evidence that the frequencies of occurrence of transitions between the states 0 and 1 varied over the course of the 42-year record.

An alternative approach, which tests specifically for a linear trend, would be to replace the model in equation (19) by

$$\ln\left[\frac{P_{ij}(t)}{1 - P_{ij}(t)}\right] = \beta_0 + \beta_1 t + \beta_2 \cos\left(\frac{2\pi t}{365}\right) + \beta_3 \sin\left(\frac{2\pi t}{365}\right) \qquad (20)$$

and a test of the hypothesis H_0: $\beta_1 = 0$, against the alternative H_1: $\beta_1 \neq 0$ is possible using likelihood-ratio procedures as described by McCullagh and Nelder (1989). The conclusion from the test is that the daily rainfall record at Ceará shows no evidence that transitions in which a dry day (0) is followed by a wet day (1) show no significant evidence of linear increase or decrease, over the 42-year record.

The discussion of this section has addressed the problem of trend detection and estimation in a daily rainfall sequence at a single site. In practice, it will often be necessary to explore whether a common trend in daily rainfall is found at a number of rain-gauge sites within a region, and the analysis becomes appreciably more difficult because of the need to allow for spatial correlation in rainfall occurrence and amount, in addition to the temporal correlations existing at each site separately.

TRENDS IN ANNUAL MAXIMUM RAINFALL INTENSITY OF DIFFERENT DURATIONS

The pdf of annual maximum rainfall intensity for a given duration D hours is commonly taken as an extreme-value distribution, the Gumbel shown in equation (7) above being widely appropriate (e.g., Buishand, 1993) although it is to be noted that the annual maximum intensities for different durations $D_1, D_2, \ldots D_K$ are correlated. The Gumbel distribution in equation (7) has the location and scale parameters

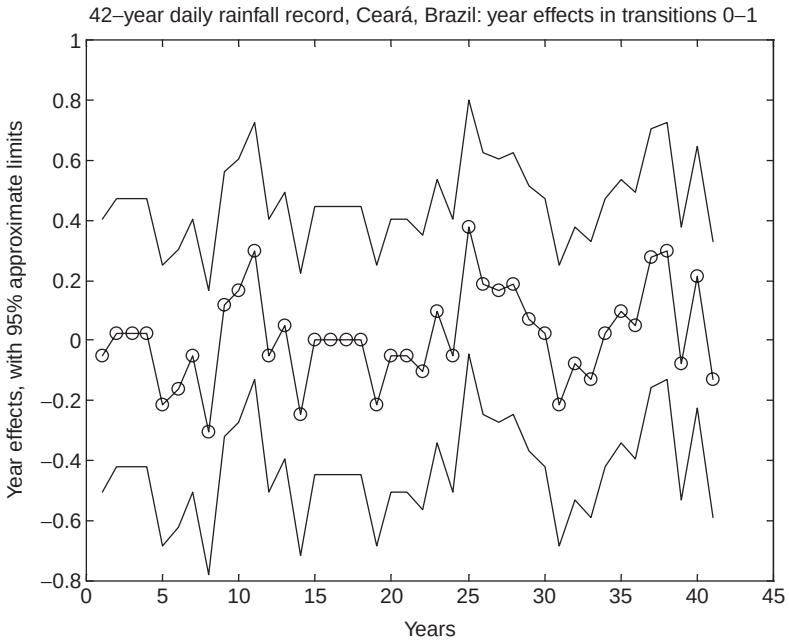


Figure 3 Year effects y_j in the logit model shown in equation (19), for the transition probability $P_{01}(t)$ for a dry day (state 0) followed by a wet day (state 1) in the 42-year record of daily rainfall at Ceará, Northeast Brazil. The 95% confidence limits include zero, indicating absence of trend in $P_{01}(t)$ over the period

μ and σ, while the mean, standard deviation, and skewness are $u + \gamma\sigma, \pi\sigma/\sqrt{(6)}$, and 1.1396 respectively, where γ is Euler's constant, $\gamma = 0.577\,215\ldots$ and $\pi = 3.141\,59\ldots$ One way of allowing for the existence of a time trend in the Gumbel mean is to modify equation (7) to be of the form

$$f(x;\mu,\sigma,\beta) = \left(\frac{1}{\sigma}\right) \exp\left[-\frac{x-\mu-\beta t}{\sigma} - \exp\left\{-\frac{x-\mu-\beta t}{\sigma}\right\}\right] \qquad (21)$$

where t is a time variable, so that equation (19) reduces to equation (7) when $\beta = 0$. The mean of the distribution in equation (21), for a given time t, is $(u + \gamma\sigma) + \beta t$, which as t varies is a line with slope β and intercept $(u + \gamma\sigma)$. More elaborate time trends can obviously be explored by replacing βt in equation (21) by $\boldsymbol{\beta}_{\mathrm{tr}}^T\ \boldsymbol{f}(t)$, with $\boldsymbol{\beta}_{\mathrm{tr}}^T$ a vector of trend parameters. More generally, trends in both position and scale parameters of the Gumbel distribution could be considered, while Coles (2001) provides S-Plus software (available on `http://www.maths.bris.ac.uk/~masgc/ismev/summary.html`) with which trends in all three parameters μ, σ, ξ of the GEV distribution can be explored. There is now a wide choice of software for fitting models for extreme-value data, and drawing inferences from them; GenStat (version 7), for example, has a menu structure, called from a toolbar, with many options for fitting EV distributions. A very useful website leading to other software (evd in R; EVIS in S; Xtreme; and others) can be found on `http://www.maths.lancs.ac.uk/~stephena/software.html`.

PROBLEMS WITH THE CONCEPT OF RETURN PERIOD IN THE PRESENCE OF TREND

The reasoning given above as the basis of return period rested critically on the concept of stationarity: the concept that the record of measurements of rainfall observed in the past provides information about the structure of the random process that will continue unmodified into the indefinite future. Under the assumption of stationarity, it is entirely valid to calculate a value for the rainfall characteristic – whether it be annual total, annual maximum intensity, longest run of consecutive days without rain, or whatever – that will occur in the future, in the long run, with a frequency of once in 10 years, once in 100 years, ...once in T years. Where, however, analysis of a rainfall record shows that trend exists, the assumption of stationarity is inappropriate. It is always possible that a trend detected in a relatively short period of record will prove, in the longer term, to be part of a longer-term fluctuation, so that what appears to be a trend in a 50-year record would, if observation were allowed to continue for say 500 years, appear as part of a longer-term pattern resulting perhaps

from slowly varying climate fluctuations. This is of little consolation, however, when decisions are required and must be based on the limited, apparently nonstationary data currently available. Until the future courses of atmospheric and oceanic processes that give rise to changes of climate can be predicted into the future, estimating how often extreme events will occur in the future will remain a very difficult problem.

Where hydrologic regimes are changing, a different approach to quantifying the probability of occurrence of extreme events is required, which avoids reference to the long-term frequency of occurrence. Recall that, in the first paragraph of this entry, an event with return period T years was defined as the event that may occur in any one year with probability $1/T$; under changing hydrological regime, this probability is no longer constant, and to generalize the concept of return period it is necessary to describe how the probability is changing. Clarke (2003) has suggested the following series of steps by which statements about the future frequency of occurrence of hydrological events become possible, where time trends are found to exist in records. Step 1: identify the extreme event that, if it occurred, would influence the choice of decision. This might be, for example, a particularly intense rainfall over a duration that would lead to severe flooding. Call the magnitude of this event x_{crit}; several values of x_{crit} may be explored. Step 2: Assuming that the trend in regime exhibited in the available hydrologic records continues into the future at the rate hitherto observed, determine the probability distribution of the time to first occurrence of x_{crit}, the event selected at Step 1. Step 3: Determine the probability of the extreme event x_{crit} occurring in the next t years, where t extends up to an acceptable (but limited: see discussion below) planning horizon. Clarke (2003) gave expressions for two cases: first, where trends have been detected in annual maximum rainfall intensity represented by a Gumbel distribution or, more generally, by a GEV distribution, with time-variant means; and second, where trends have been detected in rainfall records consisting of pairs of values $(t,\ x_t)$, where x_t is the magnitude of rainfall intensity exceeding some "threshold" value x_{thresh}, and t is the time at which $x_t > x_{\text{crit}} > x_{\text{thresh}}$ occurs. Thus, Clarke's suggestion, in the presence of trend, is to replace the concept of "event with return period T years", by the concept "the probability that a critical event, suitably defined, will occur at least once during the forthcoming limited period of S years, assuming that the observed trend in the record continues over this limited period at the same rate as that recently observed".

This second case is related, in the rainfall-intensity context, to the "peaks-over-threshold" (POT) approach to modeling the occurrence of floods in rivers; in the case in which x_t denotes a flood discharge, the sequence of pairs $(t,\ x_t)$ is sometimes termed the *partial duration series*. Fitting statistical models to POT-type data, and drawing inferences for both stationary and nonstationary processes, has been the subject of much theoretical analysis (see Smith, 1989; Davison and Smith, 1990). The limiting distribution of the random variable $Y = X - x_{\text{thresh}}$, conditional on $X > x_{\text{thresh}}$, is the generalized Pareto distribution (GPD), which has cdf

$$H(y) = 1 - \left(\frac{1 + \xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{22}$$

defined for $y > 0$, $1 + \xi y/\tilde{\sigma} > 0$. The GPD plays an analogous role, in the analysis of partial duration series, to that of the GEV distribution in the analysis of annual maxima. A very clear account of threshold models is given by Coles (2001, Chapter 4); Chapter 6 of the same work discusses trends in nonstationary sequences, and gives references to work in related fields, such as corrosion technology (Laycock *et al.*, 1990). Methods of nonparametric exploratory analysis of time-varying parameters and return levels for EV distributions have also been presented by Ramesh and Davison (2002) and Lanzante (1996).

In conclusion, changes in hydrological regime – whether as a consequence of climate change or of change in land use – require the concept of return period to be redefined. If further evidence in support of climate change accumulates, this will have important consequences for the many kinds of civil engineering project that have long been designed according to principles based on the return level of events with T-year return period, estimated from data sequences that are realizations of stationary processes.

ANALYSIS OF RAINFALL EXTREMES: AN EXAMPLE

Table 2 shows the maximum one-hour rainfall in each of the years 1970–1986 at Eskdalemuir, in the Southern Uplands of Scotland. Since the data consist of annual maxima, it is appropriate to consider using a GEV distribution; when this is fitted by maximum likelihood, the estimates of the three parameters μ, σ, ξ together with their standard errors, are

$$\hat{\mu} = 9.935 \pm 0.7421; \quad \hat{\sigma} = 1.879 \pm 0.5457;$$
$$\hat{\xi} = -0.03115 \pm 0.2978 \tag{23}$$

Having fitted the GEV distribution, use of equation (3) shows that annual one-hour rainfalls (i.e., return levels) for 5-, 10- and 20-year return periods are 12.69, 14.02, and 15.27 mm respectively. However, the estimate of ξ is almost an order of magnitude less than its standard error, so that this parameter, which describes the shape of the GEV, can safely be set to zero and an EV1 (Gumbel) distribution used. The estimates of μ and σ

Table 2 Annual maximum one-hour rainfalls (mm) at Eskdalemuir, Scotland, 1970–1986

Year	1970	1971	1972	1973	1974
AnnMax_01_hr	11.9	9.8	8.4	7.4	13.4
t	1975	1976	1977	1978	1979
AnnMax_01_hr	10.1	9.3	16.6	10.2	8.4
t	1980	1981	1982	1983	1984
AnnMax_01_hr	12.6	12.6	12.3	14.0	10.0
t	1985	1986			
AnnMax_01_hr	9.5	10.1			

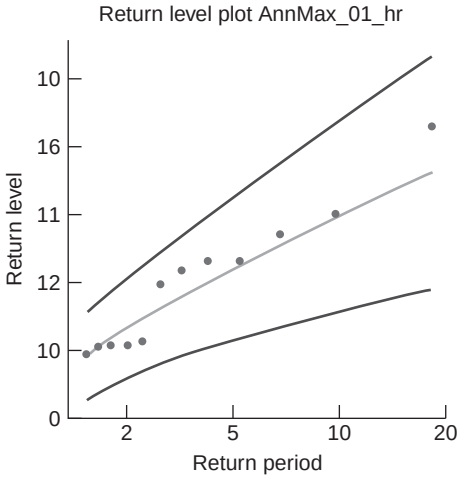


Figure 4 Plot of return level (mm) against return period (years), when an EV1 (Gumbel) distribution is fitted to the 17 annual maximum one-hour rainfalls at Eskdalemuir, 1970–1986 (see Table 2). Some points overlap. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

then become 9.904 ± 0.6720 and 1.860 ± 0.5023 respectively. The return levels for 5-, 10- and 20-year return periods are hardly changed: 12.69, 14.09, and 15.27 mm. A further calculation (see Coles, 2001) gives 95% confidence limits for these estimates as {10.42, 14.96}, {11.18, 17.00}, and {11.86, 18.99}, in units of millimeters. Figure 4 shows the relation between return level and return period, together with the 95% confidence band for predicted return levels.

If the Gumbel distribution is modified as in equation (21) to include a time trend, the trend parameter β is estimated as 0.0979 ± 0.1170 (units: millimeters per year), showing no evidence of trend in annual maximum one-hour rainfalls over the period 1970–1986.

A disadvantage in using annual maxima in an analysis of hourly rainfall extremes is that each year provides only a single observation. A POT analysis, on the other hand, may provide several observations per year that exceed the selected threshold (or, indeed, none, if the threshold is set too high). For the Eskdalemuir record of hourly rainfall 1970–1986, a threshold x_{thresh} was selected as 7.7 mm,

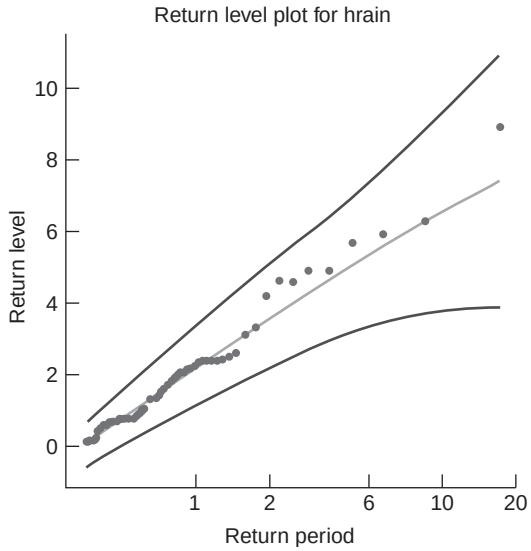


Figure 5 Plot of return level (mm) against return period (years), when a Generalized Pareto distribution is fitted to 49 most intense one-hour rainfalls (''hrain'') exceeding a threshold of 7.7 mm at Eskdalemuir, 1970–1986. Minimum interval between intense rainfalls is taken as 24 h. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

which gave an average of about three events per year. Since the rainfalls exceeding x_{thresh} may occur in clusters, it was also necessary to specify a minimum time interval between clusters of hourly rainfalls to ensure that clusters were independent, and knowledge of weather patterns at Eskdalemuir suggested that this time interval should be taken as 24 h. The variable analyzed was then the peak rainfall in each cluster, and there were 49 clusters in the 17-year period. The parameters σ and ξ of the pdf in equation (22) were estimated as

$$\hat{\sigma} = 2.266 \pm 0.6609; \quad \hat{\xi} = -0.09432 \pm 0.2113 \qquad (24)$$

from which the one-hour rainfalls with 5-, 10- and 20-year return periods were calculated as 13.04, 14.23, and 15.33 mm, slightly greater than where the GEV and EV1 distributions were fitted. Figure 5 shows the relation between return level and return period, together with the 95% confidence band for the predicted return levels; the 49 plotted points correspond to the 49 clusters determined by the selected combination of threshold and time interval between clusters.

Acknowledgments

The Fortaleza data in Table 1 were kindly provided by Mr. Marcio Nobrega; the Eskdalemuir data in Table 2, and the hourly data used in an example, were supplied by the

British Atmospheric Data Centre. The detailed and very constructive comments of two reviewers are acknowledged with gratitude.

REFERENCES

Box G.E.P. and Jenkins G.M. (1970) *Time Series Analysis: Forecasting and Control*, Holden-Day: San Francisco.

Buishand T.A. (1993) Rainfall depth-duration–frequency curves: A problem of dependent extremes. In *Statistics for the Environment*, Barnett Vic and Feridan Turkman K. (Eds.), John Wiley & Sons Ltd: Sussex.

Clarke R.T. (2003) Frequencies of future extreme events under conditions of changing hydrologic regime. *Geophysical Research Letters*, **30**, 3, 10.1029/2002GL016214.

Coles S. (2001) *An Introduction to Statistical Modeling of Extreme Values*, Springer.

Coe R. and Stern R.D. (1982) Fitting models to daily rainfall data. *Journal of Applied Meteorology*, **21**, 1024–1031.

Cox D.R., Isham V.S. and Northrop P.J. (2002) Floods: Some probabilistic and statistical approaches. In *Flood Risk in a Changing Climate: Philosophical Transactions of the Royal Society of London. Series A* 1389–1408.

Cox D.R. and Hinkley D.V. (1974) *Theoretical Statistics*, Chapman & Hall: London.

Coles S.G. and Dixon M.J. (1999) Likelihood-based inference for extreme value models. *Extremes*, **2**, 5–23.

Cressie N.AC (1993) *Statistics for Spatial Data*. John Wiley & Sons Inc: New York; *Flood Estimation Handbook*, (1999) NERC: Swindon.

Davison A.C. and Smith R.L. (1990) Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society*, **B62**, 191–208.

Hosking J.RM and Wallis J.R. (1997) *Regional Frequency Analysis: An Approach Based on L-moments*, Cambridge University Press.

Hosking J.RM, Wallis J.R. and Wood E.F. (1985) Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics*, **27**, 251–275.

Institute of Hydrology. (1999) *Flood Estimation Handbook*, Natural Environment Research Council – NERC.

Isaaks E.H. and Srivastava R.M. (1989) *An Introduction to Applied Geostatistics*, Oxford University Press.

Johnson R.A. and Wichern D.W. (1992) *Applied Multivariate Statistical Analysis*, (*Third Edition*) Prentice–Hall International Inc.

Kundzewicz Z.W., Robson A. (2000) *Detecting Trend and other Changes in Hydrological Data*, WMO World Climate Programme: Data and monitoring. WCDMP-45, WMO/TD-No.1013.

Lanzante J.R. (1996) Resistant, robust and non-parametric techniques for the analysis of climate data. *International Journal of Climatology*, **16**(11), 1197–1226.

Laycock P.J., Cottis R.A. and Scarf P.A. (1990) Extrapolation of extreme pit depths in space and time. *Journal of the Electrochemical Society*, **137**, 64–99.

McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models*, *Second Edition*, Chapman & Hall: London.

Ramesh N.I. and Davison A.C. (2002) Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology*, **256**, 106–119.

Ripley B.D. (1991) *Statistical Inference for Spatial Processes*, Cambridge University Press.

Silverman B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman & Hall: London.

Smith R.L. (1989) Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science*, **4**(4), 367–393.

Stephenson A. `http://www.maths.lancs.ac.uk/~stephena/software.html`

Stern R.D. and Coe R. (1984) A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A*, **147**, 1–34.

38: Fog as a Hydrologic Input

LA SAMPURNO BRUIJNZEEL[1], WERNER EUGSTER[2] AND RETO BURKARD[3]

[1] *Vrije Universiteit, Amsterdam, The Netherlands*
[2] *Swiss Federal Institute of Technology, Zürich, Switzerland*
[3] *University of Bern, Bern, Switzerland*

This article reviews the hydrologic importance of fog in its various forms. Meteorologically, fog is defined as a ground-touching cloud with a visibility in the horizontal of less than 1000 m. The most widely occurring fog types include radiation fog, sea fog and steam fog, and advection fog, but often fog is also referred to by its location of occurrence (coastal, valley, or mountain fog). The physical processes underlying the various types of fog are described briefly. The use, advantages, and limitations of the most common types of fog collectors, fog detectors, and fog droplet spectrometers are dealt with before embarking on a discussion of techniques for the measurement and modeling of fog deposition on forest vegetation. Results of fog deposition measurements made at selected locations (representing coastal vs. inland, lowland vs. montane, and temperate vs. tropical conditions) are compared. The data confirm the importance of fog interception at many coastal and montane sites across the tropical to mediterranean and warm-temperate climatic spectrum. Under more continental conditions, contributions by fog are usually more modest except on windy mountain ridges and summits. Finally, the paper indicates the chief gaps in knowledge regarding the measurement and importance of fog as a hydrologic input.

INTRODUCTION

One of the first scientific publications on fog by Bell (1885) dealt with a practical invention (the fog muskette) to prevent ship collisions with icebergs during foggy conditions. Another example of early scientific interest in fog was documented by Blake (1871), who was interested mostly in the spatial extent of dense fog in the city of London. Even nowadays fog is often associated with negative influences on human activity, such as piracy against ships or aircraft (Ediang, 2001), increased road traffic hazard (Panzram, 1975; Fedorova *et al.*, 2001), or the severe limitations imposed by fog on air traffic (Weinstein, 1974; Mitchell and Suckling, 1987; Robinson, 1989). In addition, persistent fog has been reported to negatively affect the mood and health of patients. Although this may not be solely due to fog, there is a direct relationship between sunlight (or the lack of it) and the length of hospitalization for depression (Benedetti *et al.*, 2001). Fog is also known to negatively influence respiratory conditions of asthmatic children (Kashiwabara *et al.*, 2002).

In hydrology, on the other hand, fog is usually considered a positive entity, and may add much needed moisture to otherwise arid ecosystems, and the animals and people depending on them. An extreme example comes from northern Chile, where frequent occurrence of advective sea fog has given rise to a particular type of forest-like vegetation ("*loma*") that, in the near-complete absence of rainfall, thrives almost exclusively on fog (Aravena *et al.*, 1989; Pinto *et al.*, 2001). It is also believed that there were prehistoric settlements in this area that were sustained by drinking water collected from favorably exposed rock formations ("fog oases"; Larrain *et al.*, 2001). Elsewhere along the American Pacific Coast, moisture additions by fog are sufficiently large to allow the development of tall evergreen forest (including the famous redwoods) in areas where amounts of ordinary rainfall are such that only some sort of Mediterranean scrub vegetation would be expected (Dawson, 1998). Similar examples come from, *inter alia*, East Africa (Hursch and Pereira, 1953), the south Caribbean dry zone (Sugden, 1982; Cavelier and Mejia, 1990), eastern Mexico (Vogelmann, 1973), and Hawaii (Juvik and Nullet, 1995a).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Classification of fog densities in relation to visibility (modified from Wanner, 1979). Selected characteristics of advected mountain fog at Pico del Este, Puerto Rico, in summer 2002 added for comparison: *LWC* = liquid water content, *VMD* = volume-weighted mean droplet diameter, *LWF* = liquid water flux (fog deposition) as measured with the eddy covariance method (R. Burkard and W. Eugster, unpublished data). Negative values denote a net deposition of fog droplets

Visibility (m)	Fog density class	LWC (μm)	VMD	LWF (% of time)	Duration
	(mg m^{-3})		(mg m^{-2} s^{-1})		
<100	Very dense fog	81.9	13.5	−8.90	17.4
100–250	Dense fog	85.4	14.4	−11.58	59.5
251–500	Medium fog	22.6	9.8	−3.08	4.5
501–1000	Light fog	12.6	8.6	−1.45	4.0
1001–2000	Haze	4.1	7.5	−0.57	4.3
>2000[a]	No fog	10.3			

[a]Arbitrary value imposed by the instrument used (Vaisala Present Weather Detector PWD11).

This article reviews the literature on fog as a hydrologic input. Starting off with a working definition of fog and its various forms, the physics and climatology of fog are discussed briefly before presenting a range of techniques to measure fog and quantify fog deposition on vegetation, including a review of current modeling approaches. Key results obtained in different parts of the world are compared and the chief gaps in knowledge identified. Finally, in view of the paucity of reviews of the hydrologic role of fog since Kerfoot (1968) and Golding (1970), the paper concludes with an extensive reference list.

DEFINITION, PHYSICS, AND TYPES OF FOG

According to its meteorological definition, fog is a ground-touching cloud where the horizontal visibility is less than 1000 m (Glickman, 2000). However, this widely adopted definition does not express any quality of the fog other than visibility, such as the spatial extent of the fog in the horizontal and vertical; its liquid water content or droplet size spectrum; whether the droplets are liquid or frozen; or the (minimum) duration of the fog event. This definition has other limitations. Depending on the observer's position, a fog may be reported as a cloud by someone standing at a distance and observing a visibility above 1000 m. Moreover, in public usage, fog and mist have a meaning that clearly extends beyond the rather arbitrary threshold of 1000 m visibility used in the meteorological definition of fog. In the literary sense, fog is always associated with unclear vision, something that obscures an object that would be clearly seen in the absence of fog. Indeed, an electronic search of the scientific literature on "fog" reveals numerous references to "fog of war", that is, the obscuring effect of gunpowder haze. Clearly, this is not a fog in meteorological terms because it consists of smoke and dry dust rather than liquid or frozen particles suspended in air.

Fog as referred to in this article relates to any type of ground-touching cloud having a density that may be expressed (albeit somewhat arbitrarily) in terms of visibility as specified in Table 1 (cf. Wanner, 1979). Several basic fog characteristics associated with each visibility class as observed at a montane site in Puerto Rico (R. Burkard and W. Eugster, unpublished data) have been added to illustrate the changes in fog density and water content associated with different visibility classes.

Fog Physics

Apart from the special case of ice fog, which falls outside the scope of the present article, fog consists of liquid water droplets that are either a condensate of water vapor, or the remnants of larger sea spray droplets or evaporating raindrops. The tiny water droplets blur the vision much more than does gaseous water (atmospheric moisture), although most of the water present in fog is in the gaseous phase rather than as liquid droplets. For example, air at 5 °C can hold as much as 6.8 g of water per m^3, whereas a typical fog, depending on its density, only contains 0.2–0.5 g of liquid water per m^3. The Clausius-Clapeyron relationship between air temperature and the amount of gaseous water that the air can contain is normally used to determine when fog occurs (i.e. when the relative humidity reaches 100%), and when it dissolves again. In the absence of condensation nuclei, relative humidity has to exceed 100% before spontaneous condensation can set in (Curry and Webster, 1999; Figure 1). However, such supersaturation is very rare in practice since even in areas with a minimum of pollution there are usually sufficient condensation nuclei present such that fog droplets start to form as soon as the temperature drops below the local dew point.

The saturation vapor pressure over the curved surface of the droplets decreases as the curvature increases, compared with the saturation vapor pressure over a flat and level water surface (Curry and Webster, 1999). Because the atmosphere tends to be more in equilibrium with the saturation vapor pressure observed over a flat water surface, the droplets will tend to evaporate and the fog will clear. Similarly, an increase in temperature during the day increases the

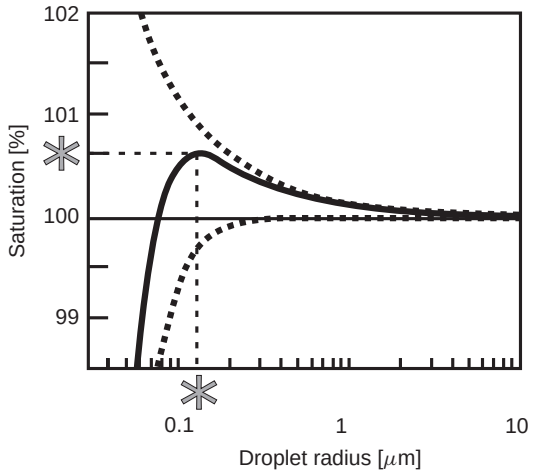


Figure 1 Supersaturation as a function of droplet radius and size of condensation nuclei. Top curve shows the equilibrium saturation ratio in the absence of condensation nuclei. The bottom curve shows the effect of the presence of condensation nuclei, and the solid line in the center shows the resulting 'Köhler' curve, indicating that in the presence of condensation nuclei the droplet radius has to exceed a threshold value r^* before the droplets start growing. Smaller droplets will tend to evaporate even when the atmosphere is saturated with water vapor

potential of the atmosphere to hold water in gaseous form, thereby also promoting the disappearance of the fog. Fog droplets grow primarily owing to condensation – coalescence does not play a significant role unless the nominal diameter of the droplets exceeds a value of ca. 36 μm (Eagleson, 2003). Once coalescence sets in, the droplets can grow much more quickly than by condensation alone, thereby rapidly exceeding the drop sizes that are typically associated with fog (<100 μm diameter). This explains: (i) why fog droplet size spectra tend to show a sharp decline with increasing drop size (Figure 2); and (ii) why fog does not automatically develop into a rain cloud within the typical life cycle of fog (cf. Pruppacher and Klett, 1998). Although fog droplet sizes cover a broad range and can be highly variable in time and space (DeFelice, 2002), the peak of the distribution (or peaks if the distribution is bimodal) usually occurs at a value below 25 μm (Figure 2), reflecting the counteracting processes of reevaporation and condensation of water vapor.

Types of Fog

There is no generally used nomenclature to distinguish between the various types of fog. In the literature, the term "stratus" is often used, regardless of whether the stratiform cloud has a visibility below 1000 m or not. As an illustration, only about one-quarter of scientific articles on the topic of "stratus" actually relate to the various types of fog described below. The nomenclature is particularly variable

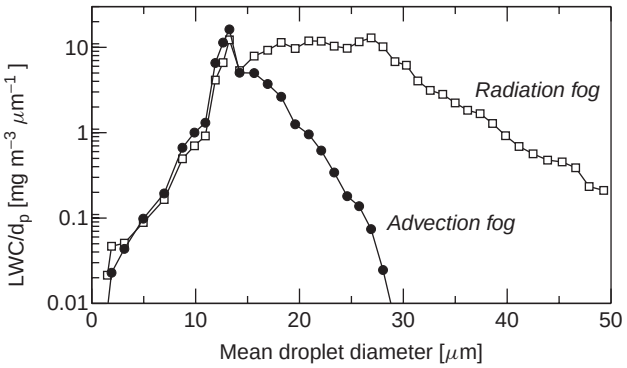


Figure 2 Examples of droplet size distributions in radiation fog (solid circles; fog event of 15 March, 2002) and advection fog (open squares; fog event of 20 February, 2002) at the Lägeren mixed forest site in Switzerland (based on original data from Burkard *et al.*, 2003)

in non-English language publications where authors may refer to a specific type of fog in their own language, which may then become translated into less common English terms (cf. Stadtmüller, 1987). Apart from ice fog and urban fog, seven relatively distinct terms (Figure 3) that are frequently used to describe different types of fog were distilled from the meteorological literature. The first four types of fog are named according to their formation process, the remaining three according to their geography of occurrence. The various types are described briefly below, starting with the most frequently used terms.

Radiation Fog

Radiation fog is the most typical inland fog type. For this kind of fog to develop, the atmosphere must be stable, with a temperature inversion layer almost adjacent to the ground surface (Nakanishi, 2000). The dominant process leading to radiation fog is long-wave radiative cooling of the atmospheric surface layer at night, which reduces the temperature below dew point. The excess water vapor starts to condense on aerosol particles present, thereby forming tiny water droplets with a size distribution typically ranging between 1 and 50 μm (Figure 2). Radiation fog in humid temperate climates occurs most frequently during the cold season, with a reported maximum in fall (Meyer and Lala, 1990; Wanner, 1979; Bendix, 2002). This reflects a combination of sufficient nocturnal cooling and an adequate supply of atmospheric moisture during this time of year. Similarly, in warmer climates radiation fog tends to be most pronounced during the (cool) dry season when cloudiness at night is at a minimum (Liu *et al.*, in press). Because radiation fog can be very shallow and local, it is notorious for causing traffic problems on roads (Panzram, 1975; Fedorova *et al.*, 2001) and airports (Weinstein, 1974; Mitchell and Suckling, 1987; Robinson, 1989). Radiation fog life cycles are generally confined to a relatively short time window, however, being centered mostly

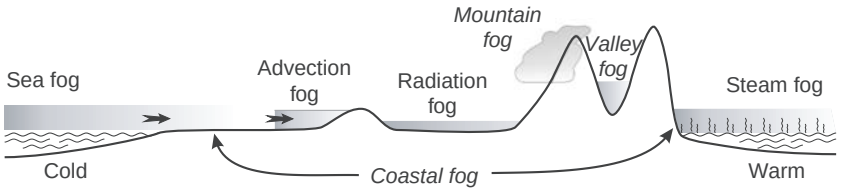


Figure 3 Types of fog. Names in upright font are process-based types, those in oblique font are geographic types

around sunrise when temperatures tend to be at a diurnal minimum (Meyer and Lala, 1990). Initial relative humidity and nocturnal cooling rate are key parameters for forecasting the actual onset of fog formation. Since radiation fog is most likely to occur during calm and cloudless nights, the associated conditions favor the development of a droplet size spectrum with a peak diameter of about 20–25 μm (Figure 2) and relatively high liquid water content (*LWC*) values (up to 500 mg m^{-3}). For example, radiation fog events during the winter of 2001/02 at the Lägeren research site in Switzerland (Burkard *et al.*, 2003) had a significantly higher median *LWC* value (143 mg m^{-3}) than "advection" fog events (see below and Figure 3) owing to the broader droplet size spectrum associated with radiation fog. Radiation fog events had a significant contribution of droplets larger than 20 μm, which were almost absent in advection fog (Figure 2).

Sea Fog

Sea fog occurs mainly in spring and early summer, but occasionally also during summertime. The chief processes leading to the formation of sea fog are cooling and condensation in relation to changes in relative humidity over open sea. Different processes of cooling and condensation create different types of sea fog but, typically, warm, humid air being cooled below its dew point as it moves over a cold water surface may lead to the formation of sea fog (Zhou, 1987; Figure 3). Sea fog constitutes a most hazardous meteorological phenomenon for ships (also because of the enhanced risk of piracy in some parts of the world; cf. Ediang, 2001). Because its occurrence is not restricted to light wind conditions, sea fog can be much more persistent than radiation fog. Also, in contrast to radiation fog over land, sea fog often forms far from the area it is affecting and may thus also be called *advection fog*, depending on the conditions (Roach, 1995; cf. Figure 3). Typical examples include sea fogs that are blown onshore as a dense but shallow layer from which tall buildings may emerge, and where the distance from the coast line may determine the presence or absence of the fog further inland. San Francisco and surroundings probably constitute the most renowned example of a sea-fog affected area (Kerfoot, 1968; Dawson, 1998), but sea fogs have also been reported to limit rice production along the Pacific coast of Japan (Inoue *et al.*, 1997), as well as to affect coastal areas in the UK (Roach, 1995), arctic Alaska (Prudhoe Bay), northern Chile (Cereceda and Schemenauer, 1991), and many other locations around the world.

Steam Fog

Steam fog differs from sea fog in that it typically forms when cold air flows over warmer water, such that vapor evaporating from the sea condenses immediately (Figure 3). In addition, it is caused by the mixing of turbulent eddies with different temperature and humidity (Øfkland and Gotaas, 1995). Visually, the process appears as smoke or steam, hence the occasionally used synonym "sea smoke" (or "*Seerauch*" in German; Tiesel and Foken, 1987). The warm water surface evaporates at much higher rates than the cold air above can take up moisture and the excess vapor condenses to liquid water droplets. A related developing mechanism is thermal convection (Tiesel and Foken, 1987). Nakata (1982) reported that for steam fog to occur in southern Japan (Hiroshima Bay), the minimum temperature difference between the water and the air had to be 9 °C, a prerequisite that was met mostly from October until December. Fenland steam fog is a comparable type of fog that typically forms when a squall line associated with a cold front moves over a relatively warm wetland (such as bogs and fens) (Oliver, 1980).

Advection Fog

Advection fog differs from radiation fog by its typical occurrence in combination with steady winds leading to the horizontal translation ("advection") of the fog layer (Figure 3). While the advective component may occasionally refer to sloping terrain (e.g. Nadezhina and Shklyarevich, 1994), the term is used most frequently in relation to fog occurrence at coastal locations (Goodman, 1977; Cereceda and Schemenauer, 1991), as well as in the context of pollutant transport in industrialized inland areas (Vong *et al.*, 1991). In coastal areas, the strong contrast in surface heating between the land and the ocean during the daytime generates a sea breeze that pushes any fog formed offshore toward the land where it may dissipate again at some distance from the shoreline as a result of the higher temperatures prevailing over land surfaces. Therefore, advection fog in coastal areas is often also referred to as "coastal fog" (see below and Figure 3). Because it is wind-driven, advection fog is usually associated with substantially greater hydrologic (and chemical) fluxes than radiation fogs (Vong *et al.*,

1991), even though advection fog has a much narrower size range (Figure 2). Advection fog is a "warm fog" with liquid droplets that can produce serious surface glazing when occurring over frozen ground (Nadezhina and Shklyarevich, 1994).

Ice Fog

In contrast to the previous fog types (which consist of liquid droplets that may persist as supercooled water even at temperatures well below freezing point; Sakurai and Ohtake, 1979), ice fog is composed of frozen water crystals. Ice fog is basically restricted to very cold arctic and antarctic regions with temperatures below – 25 °C and will not be discussed here.

Coastal Fog

Although fogs are most frequently named according to their formation process or physical characteristics (Figure 3), fog types are often also referred to according to their geography of occurrence (such as coast, valley, mountain, etc.). As indicated earlier, this makes a unique classification of fog types impossible. The four fog types discussed below are the most widely used geographic terms to denote different kinds of fog.

Both advected sea fog coming in with the sea breeze, as well as the reverse (i.e. radiation fog developed over land and subsequently advected to the ocean with the land breeze at night), and even steam fog have been referred to as "coastal fog" in various Californian studies (e.g. Pilié, 1979; Noonkester, 1979; Goodman, 1977). Cereceda and Schemenauer (1991) described a situation with "coastal fog" in northern Chile where a combination of onshore movement of a marine stratocumulus deck and orographic effects resulted in the relatively consistent occurrence of a fog belt between ca. 500 and 1000 m altitude from late afternoon onwards until midmorning. The mean droplet size diameter of this advected sea fog ranged from 10.8–15.3 μm (Schemenauer and Joe, 1989), that is, similar to that observed in advection fog under continental conditions in Switzerland (Figure 2).

Valley Fog

As the name implies, valley fog is radiation fog forming in valleys (Figure 3; Pilié, 1975). Valley fog may occur regardless of the elevation of the valley or the general climate (Fitzjarrald and Lala, 1989; Pristov and Trontelj, 1979; Liu *et al.*, in press), its typical feature being that the life cycle of the fog is strongly governed by the diurnal mountain-valley circulation system. Within a few hours after sunrise, winds moving up the valley tend to erode the core of cold air built up in the valley bottom during the previous night, thereby dissolving the fog layer. The fog may return again after cold air starts to drain from higher ground and the radiative surface energy budget of the valley becomes negative, usually toward sunset (Whiteman and McKee, 1982). The top of the valley bottom radiation fog layer is normally determined by a local thermal inversion that separates the static warmer air above from the cold air in the valley. In climatologic terms, this leads to the occurrence of a warm thermal belt at some elevation along the slope. At higher elevations, fog is frequently observed above the typical lifting condensation level of convective clouds. In Switzerland, for example, valley fog occurrence above the cloud base is much more frequent, especially in summer, than radiation fog forming in valleys lower down. The exact climatology of fog occurrence may differ between areas, depending on local topography and atmospheric humidity regime. A valley with a large hinterland generating cold air draining into the main valley at night tends to experience more frequent radiation fog if the outlet for the air is narrow (e.g. a gorge). Similarly, drainage of cold air will be much more rapid – and valley fog less frequent – in the case of a gradually sloping valley that widens steadily in the downslope direction.

Urban Fog

The occurrence of fog in and around large cities has received particular attention to derive the necessary scientific understanding for fog forecasting, air pollution control (smog!), air traffic planning, and road condition assessment (e.g. London: Blake, 1871; Munich: Sachweh and Koepke, 1997; Shanghai: Zhou, 1991; Zhou and Wang, 1991; Bao *et al.*, 1995; various industrial cities in Central India: Patel *et al.*, 1998; and the surroundings of Mexico City and Puebla in Mexico: Padilla-Gordón, 1998). The various types of urban fog distinguished in the literature can be considered to represent a mixture of the process-based fog types distinguished in Figure 3, with the actual type depending on specific climatologic setting.

Mountain Fog

Mountain fog is an orographically controlled type of fog (Figure 3) that has been shown to strongly affect the hydrologic, nutrient, and pollution budgets of montane forests, both in the temperate zone (e.g. Sigmon *et al.*, 1989; Vong *et al.*, 1991; Collett *et al.*, 2002) and in the so-called tropical montane cloud forest belt (e.g. Gordon *et al.*, 1994a, b; Bruijnzeel and Proctor, 1995; Clark *et al.*, 1998; Hafkenscheid, 2000; Bruijnzeel, 2001). Mountain fogs generally consist of clouds that extend upward along the slope from their lowermost condensation level ("cloud base"). Depending on the height of the cloud base, the vertical extension of the cloud belt, and the height of the mountain, considerable portions of the landscape may thus become more or less regularly immersed in fog. A well-known example of this is the "sea of clouds" that envelops north-facing slopes on the Canary Islands for about one-third of the time during most of the year. The upper limit of the fog belt is controlled by the height of a temperature

inversion layer that generally occurs around 2000 m but is lowered to about 1000 m in summer. Similarly, the cloud base varies seasonally between ca. 700 m in summer and ca. 1100 m in fall. In addition to the strong reduction of the vertical extension of the fog belt in summer, fog duration is also much reduced at that time of year (to ca. 13% of the time; Marzol *et al.*, 1996). The extra moisture provided by the clouds supports tall evergreen forest types within the cloud belt, in strong contrast to the arid vegetation found both above and below the cloud-affected zone (Jiménez *et al.*, 2000).

Because the more or less frequent presence of fog affects both the microclimate and overall water budget of montane vegetation, the altitudinal zonation of montane forest types in general, and of bryophyte cover of trees (mosses) in particular, has been related to the intensity and persistence of fog occurrence on wet tropical mountains (Grubb, 1977; Frahm and Gradstein, 1991; cf. Bruijnzeel, 2001). While tall "Lower Montane Rain Forest" with ca. 10% moss cover is typically associated with largely fog-free conditions, "Lower Montane Cloud Forest" may have up to 25–50 per cent moss cover and experience "frequent" fog incidence. Above the latter so-called "Upper Montane Cloud Forest" is found, having up to 80% bryophyte cover and being subjected to "long and persistent" low cloud. Although quantitative criteria of what constitutes "frequent" or "long persistent" fog are lacking from these definitions (Grubb, 1977), similar zonations are found all over the tropics. In addition, the term "cloud forest" at least reflects the general recognition that fog and low cloud exert a dominant influence on a range of forest hydrologic and ecological processes (Cavelier, 1990; Hutley *et al.*, 1997; Hafkenscheid, 2000).

It is pertinent to note that fog-affected forests in the (sub-)tropics cover a wide altitudinal range in response to changes in prevailing temperatures and atmospheric humidity (rather than rainfall). Away from the equator, the average elevation of the cloud belt tends to decrease with increasing latitude, reflecting the associated reduction in average temperatures. On small oceanic mountains, the fog belt is typically found at much lower elevations than on similar-sized or larger mountains further inland. The resulting compression of vegetation zones on small mountains compared to that on larger mountains is called the "*mass elevation effect*". This somewhat peculiar term relates to the idea that the vertical displacement of the earth's surface in large mountain massifs is capable of raising ambient air temperatures – and therefore cloud condensation levels – above values observed in the free atmosphere at the same elevation, thereby extending the zonation of vegetation in an upward direction (Richards, 1996). However, the fact that the effect is rarely (if ever) observed in the dry tropics and becomes less pronounced on mountains away from the ocean, suggests a major atmospheric humidity effect on cloud condensation and vegetation zonation as well (Stadtmüller, 1987; Bruijnzeel *et al.*, 1993).

Synonyms and Other Types of Fog

Apart from the four process-based (Figure 3) and the four geographic fog types (nos. 5–8) distinguished above, there are numerous other terms in the literature on fog. Fog-related terms as used in the German and French literature are discussed in Stadtmüller (1987) and Dufour (1978), respectively. The following terms may be considered to be synonymous to the respective fog types distinguished in this article, although the consultation of individual scientific papers may reveal minor differences: *marine fog* → sea fog or steam fog; *frontal fog* → advection fog; *ground fog* → radiation fog; *raised fog* → stratus; *offshore fog* → sea fog.

In many places, fog water tends to be enriched with the heavier stable isotopes ^{2}H and ^{18}O compared to rainfall, reflecting differences in the origin and life history of fog and rain (Ingraham, 1998; Scholl *et al.*, 2002; cf. **Chapter 116, Isotope Hydrograph Separation of Runoff Sources, Volume 3**). Recent investigations in the maritime tropics (where mountain fog often occurs together with rain), however, have shown stable isotope contents in fog water to be similar to, or to be depleted even relative to those in rain water (Te Linde *et al.*, 2001; Eugster *et al.*, 2002), indicating that the processes leading to fog formation may differ considerably between sites. Nevertheless, it should be possible to define distinct isotopic signatures for the process-based fog types distinguished in Figure 3, such that the currently not very consistent nomenclature and use of synonyms in the literature may be based on more objective criteria in the future (Ingraham, 2001).

FOG CLIMATOLOGY

Detailed fog statistics that include information on the timing and duration of fog are mostly available for airports but rarely for routine weather stations. Typically, only the number of days with fog is given (Figure 4). However, since fog is not a continuous phenomenon and shows high spatial variability, such counts usually reveal high year-to-year variation (cf. Wanner, 1979). Also, the standard WMO procedure used by most weather services should be kept in mind when interpreting statistical fog data. A modern development is to use satellite imagery (e.g. NOAA, AVHRR) to map the spatial extent of radiation fog layers (e.g. Bendix, 2002) but this approach too has its limitations. For example, it cannot be determined unambiguously whether a cloud layer that looks rather uniform in a satellite image is actually touching the local surface or not. Thus, both ground fog and raised fog are lumped together, whereas in addition only fog fields of a certain minimum spatial extent can be detected in this

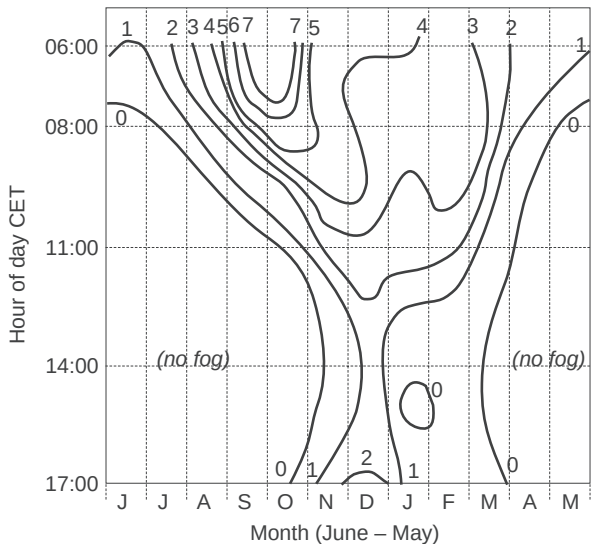


Figure 4 Number of days with fog at the Dübendorf airport, Switzerland, during the period 1938–1944 as a function of time of day and month of year. Redrawn after Zingg (1944)

way. Furthermore, because radiation fog often occurs early in the morning, this may not always match the timing of a satellite's flight over pass. Finally, illumination of the fog in the early morning may be problematic for automatic determination of fog from satellite images. Nevertheless, satellite imagery provides the best spatial overview of radiation fog occurrence, although at the present stage of development the technique completely ignores mountain fog and other fog types, which, unlike radiation fog, do not give a distinct evenly illuminated pattern (Bendix, 2002).

FOG MEASUREMENT TECHNIQUES

There is no single instrument that performs all the possible measurements required or desired in modern fog research. In the following, the main instrument types as well as their chief advantages and disadvantages are discussed briefly.

Fog Collectors

At the lower end of the sophistication range, passive fog gauges are often used to collect fog water, be it for the determination of fog water volume or for subsequent chemical analysis. Many different types of fog gauges are available, ranging from a wire-mesh cylinder placed on top of a rain gauge ("Grunow type"; Russell, 1984), to wire "harps" made of teflon strings within a square frame (Goodman, 1985) or a cylinder (Falconer and Falconer, 1980), cylindrical aluminum louvered screens (Juvik and Nullet, 1995b), and "standard" fog collectors (SFC) consisting of a 1 m^2 polypropylene net (Schemenauer and Cereceda, 1994). Usually, the extra catch recorded by a fog collector over that of an adjacent rain gauge is attributed (and equated) to "fog". However, passive fog collectors suffer various limitations: their efficiency is dependent on prevailing wind speeds; they are not specific to fog in situations where fog and (wind-driven) rain occur simultaneously; and, by nature, they are unable to adequately mimic the complexities of (tall) live vegetation in all but the simplest cases (Joslin *et al.*, 1990; Bruijnzeel, 2001). Two-dimensional screens or harps have the added disadvantage of not presenting the same silhouette and configuration in different wind directions. Adding a protective cover may eliminate some but not all wind-driven rain in particularly wind-exposed situations (Juvik and Nullet, 1995b). For particularly windy situations, Daube *et al.* (1987) proposed a wire harp collector placed within a rain-proof box in which the air flow is restricted by two baffles. The front baffle causes the passing air to accelerate and project heavy rain drops against the rear baffle where they are drained away. The lighter fog droplets are blown onward to impact against the collecting harp (Hutley *et al.*, 1997). Cylindrical gauges, on the other hand, generally represent a much smaller surface area than the "standard" fog collecting screen and the latter may thus generate measurable volumes of fog water where fog liquid water content is low or winds are light (Schemenauer and Cereceda, 1995).

Passive fog collectors are therefore used best as comparative instruments for site characterization, and should ideally be protected against direct rainfall and equipped with a recording device to evaluate the timing and frequency of fog. Nevertheless, caution is needed when interpreting data obtained in this way. For example, using Grunow-type fog gauges to characterize (potential) fog interception as part of a transect study in Panama (Figure 5), Cavelier

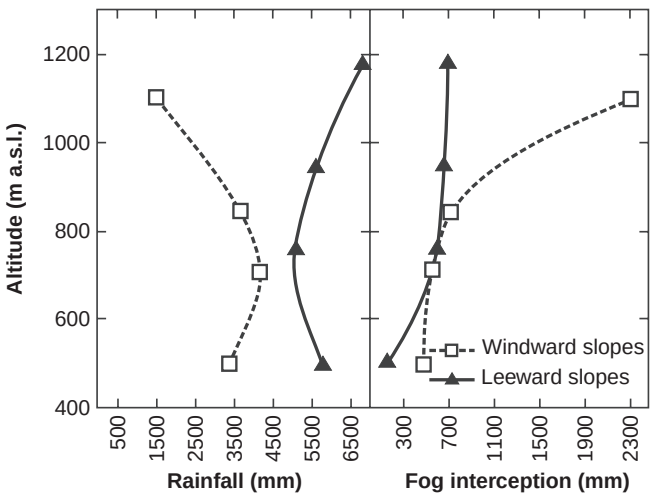


Figure 5 Variation in rainfall and 'fog interception' with altitude in the Central Cordillera of Panamá (modified from Cavelier *et al.*, 1996)

et al. (1996) obtained unrealistically high values (up to ca. 2300 mm year^{-1}) for an exposed site at 1100 m asl near the Pacific-Caribbean water divide. The corresponding rainfall at this very windy site was determined at less than 1500 mm year^{-1} even though rainfall totals at comparable elevations on either side of the divide were above 3500 mm (Figure 5). Such findings suggest severe underestimation of measured rainfall and overestimation of fog interception owing to wind effects (see Bruijnzeel (2001) for a fuller discussion). Apparent values of *LWC* may be derived from information on water volume and wind speed per sampling period if the trapping efficiency of the collector is known, but the fact that collector efficiencies tend to be dependent on wind speed renders the exercise less than straightforward (Mueller and Imhoff, 1989). Collector efficiencies vary widely between types, for example, 0.82–0.87 for the cylindrical Teflon wire harp (for wind speeds between 5.6 and 10.3 m s^{-1}; Lin and Saxena, 1991) versus 0.50–0.66 for the SFC (Schemenauer and Joe, 1989). Dawson (1998) drew attention to the fact that wire harps placed inside a forest collected significantly higher amounts of fog water than similar devices placed in the open. He attributed the difference to higher evaporative losses from the gauges at more exposed locations.

Arguably the best sampling concept is represented by the Caltech collector (Figure 6; Daube *et al.*, 1987). This active sampler is highly efficient and capable of collecting sufficient volumes of fog water even for complex chemical analyses. The instrument may also be used for the derivation of fog *LWC* (see Demoz *et al.* (1996 for details). A further advantage is that rain drops are effectively separated from the sampled fog water (Figure 6). Limitations include the high power consumption of the fan, and the uni-directional intake, which requires measures to ensure the intake is always facing the prevailing wind direction.

Detailed literature is available comparing the performance of the different types of samplers (Collett *et al.*, 1990; Schell *et al.*, 1992). To save power in field applications, fog detectors have been used to decide when to switch on the fog collector (e.g. when visibility drops below 500 m or *LWC* values fall below 20 mg m^{-3} and the efficiency of the Caltech collector becomes greatly reduced; Thalmann *et al.*, 2002).

Fog Detectors

Fog detectors (Krovetz, 1988; Mallant, 1988; Valente, 1989) are widely used in meteorological networks for the monitoring of traffic hazard conditions (fog, icing, heavy precipitation). The detection of fog or cloud is achieved by measuring the degree of optical back- or forward-scattering of emitted light in the near-infrared range. Under clear conditions, there is no such diffuse scattering. Some sensors are more sensitive than others if rain and fog occur at the same time and a careful assessment of the measurement technique is required to meet the requirements of a specific investigation.

There is an inverse (hyperbolic) relationship between visibility and fog liquid water content (Kunkel, 1984; cf. Table 1):

$$LWC = \frac{a}{(b \cdot V - c)^d} \tag{1}$$

where V is the visibility (m), *LWC* liquid water content (mg m^{-3}), and a, b, c, and d are empirical coefficients. The form of equation (1) describes why visibility can drop dramatically in dense fog even if variations in liquid water content are only small. This is illustrated further by the data presented in Table 1. The reason for the breakdown of the apparent relationship between *LWC* and visibility relates to the fact that visibility is mainly a function of droplet size distribution and numbers rather than of *LWC* itself. At low visibilities, there are numerous very small droplets (<2μm) that strongly reduce the visibility but have a limited effect on fog *LWC* (Seinfeld and Pandis, 1998). Visibility and *LWC* data collected at Lägeren, Switzerland suggested average values of $a = 6800$, $b = 1$, $c = 21.5$ and $d = 0.92$. Since the scatter in individual measurements was considerable, a simpler equation with $a = 1\,08\,340$, $b = 1$, $c = 0$, and $d = 1.136$ was equally effective (Burkard *et al.*,

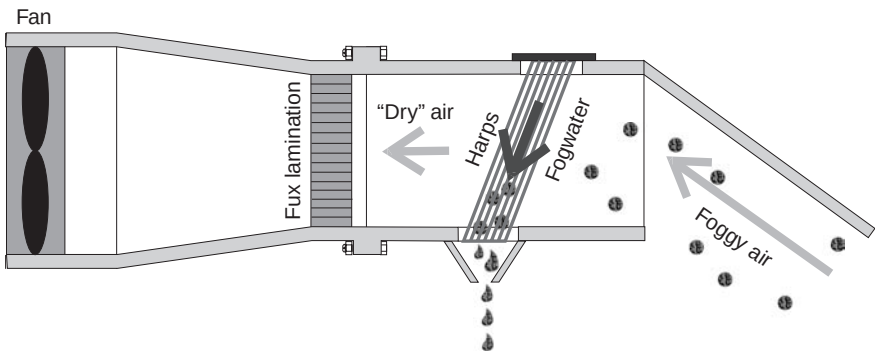


Figure 6 The Caltech Active Strand Cloudwater Collector (CASCC) (redrawn after Daube *et al.*, 1987)

2003). Equation (1) has been used inversely to predict visibility from *LWC* estimates obtained with numerical models (Smirnova *et al.*, 2000).

Fog- and Cloud-Droplet Spectrometers

Droplet spectrometers were initially used on aircraft flying across clouds and their ground-based application in fog research is comparatively recent (Beswick *et al.*, 1991; Kovalski *et al.*, 1997; Burkard *et al.*, 2002). These instruments are relatively expensive (starting at US $ 35 000) and not always designed for continuous long-term usage. The most widely used spectrometers include the Particle Volume Monitor (PVM-100), the Forward-Scattering Spectrometer Probe (FSSP) and its two-dimensional variant the Optical Array Probe (OAP), and the FogMonitor (FM-100). The PVM-100 uses a laser beam whose light is scattered forward at small angles by liquid water droplets present in the probed volume of air. There is a linear relationship between the degree of scattering and fog *LWC* (Arends *et al.*, 1992). The PVM-100 is particularly sensitive to fog droplets in the size range 3–50 μm but not to larger droplets. It is generally considered the reference instrument for fog and cloud physical research (Valente, 1989; Borrmann *et al.*, 1994). The FSSP and OAP also measure drop size distribution photo-electronically. Instrumental designs are available for the probing of precipitation (larger drops) or cloud and fog (smaller drops). However, derived values of *LWC* are quite sensitive to errors related to probe orientation, uncounted particles while the device is processing data, wind gusts, and the rather low air flow rate employed by the FSSP (see Baumgardner, 1983; DeFelice, 1998 for details). The FM-100 is an improved OAP (or FSSP) in which a pump is used to produce an air flow of nominally $13\,\mathrm{m\,s^{-1}}$ across the optical path of the forward-scattering laser (Burkard *et al.*, 2001; Eugster *et al.*, 2001). The FM-100 measures the size of each single drop passing the optical path and counts their number in each of 40 size classes between 1 and 50 μm aerodynamic diameter. The accuracy at small drop sizes is such that the useful range is 2–50 μm. However, because fog droplet spectra tend to tail off at both ends (cf. Figure 2) the poor accuracy for low diameters is not a problem in most applications. The advantage of the FM-100 over earlier instruments is that it uses a clearly measurable flow rate and aperture size of the laser beam, allowing better specification of the sampled air volume. The liquid water content of the fog is computed from the information on droplet sizes and counts, assuming the droplets are spherical. The FM-100 has been applied successfully in eddy covariance fog water flux studies in montane forests in the temperate zone (Eugster *et al.*, 2001; Burkard *et al.*, 2002) and, more recently, in the humid tropics (Holwerda *et al.*, 2005).

Other sensors measure fog *LWC* only (i.e. not its drop size spectrum). The principle of operation is that liquid water collected on a surface is evaporated by heating, and the heating rate required to evaporate the water is used to compute the absolute amount of liquid water. This type of sensor is very accurate because no assumptions on size distribution and shape of the droplets need to be made. However, for studies of fog deposition, knowledge of droplet size spectrum is usually required (see section on modeling below), such that readings of *LWC* with a liquid water sensor are made mostly as a cross-check of *LWC* values derived from droplet size spectral information. The King hot-wire probe and the Rosemount icing probe are examples of liquid water sensors (see DeFelice (1998) for details).

QUANTIFYING FOG INTERCEPTION BY VEGETATION

The fog intercepting capacity of (tall) vegetation has been recognized for centuries. Arguably the most famous example concerns a large laurel-like tree on the island of El Hierro, Canary Islands, which allegedly trapped sufficient water from passing cloud for the entire population (and their cattle) on this otherwise parched island. Such was the importance of this "fountain" tree that not only was it guarded continuously but it also appears on the island's municipal coat of arms. The services of the tree were enjoyed for centuries until it was uprooted in a violent storm in 1610 (Gioda *et al.*, 1995; see also Marzol (in press) for various historic accounts). Other well-known examples include the coastal redwood forests of California and the *fynbos* of South Africa's Western Cape Province, both of which were the subject of fog-related research as early as the first few decades of the twentieth century (see Kerfoot (1968) for a review of early literature), and the Canary pine, tree heath, and *fayal* forests on Tenerife, Canary Islands, which have been investigated increasingly since the 1950s (reviewed by Marzol, in press; Jiménez *et al.*, 2003). Finally, the very high amounts of streamflow emanating from tropical headwater catchments with montane cloud forest are often thought to reflect substantial inputs by fog interception (Zadroga, 1981; for other examples see Bruijnzeel, 2001).

While the hydrologic importance of the extra inputs provided by fog water is thus widely acknowledged, its actual quantification remains notoriously difficult (Kerfoot, 1968; Bruijnzeel, 2001; Holwerda *et al.*, 2005). Key factors affecting rate and amount of fog deposition include wind speed, fog *LWC*, surface area, and geometry (including height) of the vegetation, as well as fog duration (Schemenauer, 1986). Until recently, the two most frequently used approaches have been: (i) the use of passive fog gauges (Cavelier *et al.*, 1996; Hutley *et al.*, 1997; Ataroff, 1998); and (ii) the comparison of canopy drip or net precipitation (P_n, i.e. throughfall T_f + stemflowS_f) as measured

inside the vegetation, with amounts of gross rainfall (P_g) for periods with and without fog (Harr, 1982; Sigmon *et al.*, 1989; Hafkenscheid *et al.*, 2002). The deposition of fog onto vegetation has also been measured directly with the so-called eddy covariance approach (Beswick *et al.*, 1991; Vong and Kovalski, 1995; Kovalski *et al.*, 1997; Holwerda *et al.*, 2005). The two primary transport processes governing the deposition of fog droplets are turbulent diffusion of smaller droplets ($<10\,\mu$m) suspended in the air, and gravitational settling of the larger droplets. The turbulent part of the fog water flux can be measured directly by determining the covariance of vertical wind speed and fog liquid water content. The gravitational part is calculated from Stokes's settling velocities for specific drop diameters and information on fog droplet size distribution (Beswick *et al.*, 1991; Burkard *et al.*, 2002). Before presenting quantitative information on "typical" amounts of fog interception for different forest types and climatic settings, the limitations of the three methods are discussed below. Estimates of fog deposition on to (mostly coniferous) forest canopies using various model approaches are dealt with separately in the next section.

Evaluating Fog Deposition from Measurements of Net Precipitation

The impossibility of equating the (extra) catch of a simple fog gauge with that of a complex forest canopy in most cases (cf. Joslin *et al.*, 1990) has been commented upon already. Because each forest represents a more or less unique situation that defies standardization, the classic approach of comparing gross and net precipitation for events or periods with and without fog at least has the merit of incorporating the influence of the structural characteristics of the forest under consideration. Subtracting amounts of throughfall (T_f) plus stemflow (S_f) from gross rainfall (P_g) gives the amount of precipitation intercepted by the canopy and evaporated back to the atmosphere during and shortly after the event. This process is usually referred to as rainfall interception loss (I) as it represents a *net loss* of water to the forest (cf. **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**):

$$I = P_g - (T_f + S_f) \tag{2}$$

where the terms are as defined above and expressed in mm of water per time period. Where fog occurs in the absence of rainfall, a similar process of cloud water or fog interception (CW) may be defined as:

$$CW = E_{cw} + T_f + S_f \tag{3}$$

Because neither the actual amount of fog interception (CW, often also termed *fog deposition*, D_f) nor that evaporated again from the wetted vegetation (E_{cw}) are easily quantifiable in a direct manner, a more practical approach is to equate net precipitation (P_{net}) during fog-only events to *net* fog interception CW_{net}:

$$CW_{\text{net}} = T_f + S_f = P_{\text{net}} \tag{4}$$

where the term fog interception now implies a *net gain* of water to the ecosystem. In the more complex case of rainfall plus fog, separate knowledge of evaporation from the wetted vegetation ($\sum E_I$) is required to solve the wet-canopy water budget equation for CW:

$$P_g + CW = \sum E_I + T_f + S_f \tag{5}$$

Often, however, CW (strictly speaking, CW_{net}) is simply equated to the increase in the ratio P_{net}/P_g (or T_f/P_g) when going from events or periods without fog to events/periods with fog. In doing so, not only the stemflow component is neglected but also any differences in evaporation rates from a canopy wetted by rainfall or fog are ignored (Harr, 1982; Sigmon *et al.*, 1989; cf. **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). Solving equation (5) for CW under the wet and windy conditions prevailing at many fog-affected forest sites is not easy and the associated error bands of the estimate tend to be wide because of the accumulation of errors in the individual wet-canopy water balance terms (Holwerda *et al.*, 2005). Depending on wind speeds and rainfall intensities, P_g can be severely underestimated because of unaccounted wind-driven rain missed by a standard rain gauge (cf. Figure 5). Although various corrections have been proposed for this (cf. **Chapter 35, Rainfall Measurement: Gauges, Volume 1**), there is the added complication of trees sticking out of the main canopy and catching inclined rainfall (and/or CW) more efficiently than their more sheltered neighbors, thereby increasing the overall catch to a level that exceeds amounts of conventionally measured rainfall in the open (Herwitz and Slye, 1992; Figure 7), particularly in steeply sloping terrain (Sharon, 1980; Holwerda *et al.*, 2005). Furthermore, because net precipitation under such conditions often exceeds P_g it is not possible to estimate $\sum E_I$ in a manner analogous to equation (2). Some investigators have therefore used the wet-canopy variant of the Penman-Monteith equation to approximate $\sum E_I$ although this may result in underestimation owing to problems with advected energy (cf. **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). More importantly, the typically high spatial variability of throughfall in natural forests (particularly tropical ones) requires the use of a large number of gauges (usually >20–30). A "roving" gauge arrangement in which the gauges are relocated regularly is believed to sample "drip" points (where rain or fog drip become concentrated because of peculiarities in the configuration of the trees) in a more representative manner than a fixed gauge arrangement (Lloyd and Marques,

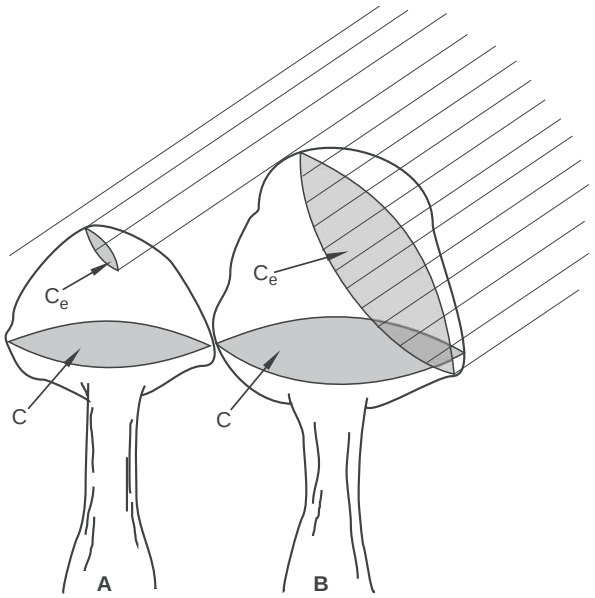


Figure 7 Illustration of the differential interception of inclined rainfall by two neighboring trees (A and B) showing how their effective intercepting crown areas (C_e) differ from their vertically projected crown areas (redrawn after Herwitz and Slye, 1992)

1988). Although the difference in mean T_f catch between the two gauge arrangements (higher T_f for roving gauges) tends to disappear above 25–30 gauges (cf. Czarnowski and Olszewski, 1970), the standard error of the mean is usually much reduced when using roving gauges (Lloyd and Marques, 1988). Because only few throughfall studies have employed a roving gauge technique, errors in published estimates of CW and CW_{net} based on conventionally measured net precipitation are likely to be substantial (see discussions in Hafkenscheid *et al.*, 2002; Holwerda *et al.*, 2005).

It follows that the comparison of P_{net} and P_g for events/periods with and without fog to evaluate CW or CW_{net} only works well when fog contributions are substantial and temporally well-defined; the error bands around the mean T_f estimate sufficiently narrow; and S_f either negligible or quantified separately. However, regression equations linking P_{net} to P_g for conditions with and without fog are often not significantly different at probabilities <0.10 (Sigmon *et al.*, 1989; Hafkenscheid *et al.*, 2002; cf. Holwerda *et al.*, 2005).

Mass-Balance Techniques

Knowing that concentrations of sodium and chloride in fog water are generally much higher than in rainfall but leached in negligible quantities only from within the leaves (Asbury *et al.*, 1994; Sigmon *et al.*, 1989), Hafkenscheid *et al.* (1998) attempted to evaluate the contribution of fog to two nearly adjacent montane cloud forests of different exposure using the sodium mass balance:

$$(P_g \times C_{Pg}) + (CW \times C_{cw}) = (T_f \times C_{Tf}) + (S_f \times C_{sf}) \quad (6)$$

in which C denotes the concentration of sodium (or any other suitable constituent) in the respective components. While a reasonable estimate was obtained for the more exposed forest, an unexpectedly high fog water input was derived for the less exposed forest. Similarly, unrealistically high values for CW were obtained in this way for an exposed ridge-top cloud forest in Puerto Rico (Te Linde *et al.*, 2001; cf. Asbury *et al.*, 1994), suggesting that the approach is less than straightforward in complex mountainous terrain, possibly owing to spatial and temporal variations in dry deposition of sea salts. A similar mass balance approach uses the difference in isotopic composition of rain and fog water (fog often being enriched in the heavier isotopes ^{2}H and ^{18}O relative to rainfall in the same region; Ingraham, 1998; Scholl *et al.*, 2002). Dawson (1998) quantified the importance of CW in this way to the water use of a redwood forest in California. The method gives good results as long as contrasts in the isotope composition of fog and rain are sufficiently large (Dawson, 1998; Te Linde *et al.*, 2001; Eugster *et al.*, 2002).

Attempts have also been made to measure fog deposition using lysimeters (Trautner *et al.*, 1990; Fowler *et al.*, 1990; Cameron *et al.*, 1997). Although such a set-up approximates natural conditions as closely as possible, it is best suited for short-statured vegetation, such as, grasses and crops rather than tall, deep-rooting trees. When lysimeter-based results were compared with estimates derived with micrometeorological techniques, considerable discrepancies were sometimes obtained (Fowler *et al.*, 1990; Cameron *et al.*, 1997). Finally, Chang *et al.* (2002) estimated fog interception by a mossy montane coniferous forest in Taiwan by multiplying the average rate of fog absorption by individual moss samples times their estimated biomass at the stand scale. Although the fog "stripping" capability of individual conifer leaves was about half that of the mosses (0.30 vs. 0.63 g H_2O g^{-1} dry weight h^{-1}), the corresponding leaf biomass was about 20 times larger, giving an inferred fog interception capacity for the canopy as a whole of ca. 2 mm h^{-1} (a very high value) versus ca. 0.17 mm h^{-1} for mosses only (Chang *et al.*, 2002).

Eddy Covariance Method

Although the direct measurement of net fog deposition (i.e. the downward flux to the canopy minus the flux in the opposite direction owing to concurrent fog formation) allowed by the eddy covariance method is promising, the technique is not without problems and cannot be applied in the steep and complex topography in which many fog-affected vegetation types occur. The few studies that have

compared eddy covariance-based estimates of *CW* with traditionally derived values (using equations (4) or (5); Vermeulen *et al.*, 1997; Holwerda *et al.*, 2005) suggested much smaller (3–6 times) values for directly measured net deposition. The reasons for such discrepancies are likely to vary between sites. On the one hand, these include the sometimes large uncertainties in the estimation of P_g and P_{net} (especially at wet and windy locations; cf. Figures 5, 7, and 8), on the other hand there is the problem of vertical flux divergence, that is, the flux as measured several meters above a forest canopy is not equal to that occurring at the canopy level (Kovalski and Vong, 1999; Burkard *et al.*, 2002; Holwerda *et al.*, 2005). The current understanding is that fog formation (through condensation of ascending air) and fog interception by the vegetation are concurring and large components at sloping windy sites. This makes it difficult to directly determine the fluxes at the plant leaf reference level using eddy covariance measurements performed above the canopy, since the technique only measures the net difference between the two components (Kovalski and Vong, 1999; Holwerda *et al.*, 2005). Current fog deposition research is beginning to focus on conditions prevailing within canopies (DeFelice, 2002). So far, only "surface-normal" fluxes have been investigated, and it has been implicitly assumed that horizontal advective impacts during strong winds (Walmsley *et al.*, 1996) lead to a vertical flux that is included in the eddy covariance measurement. This assumption may need to be revised in the future.

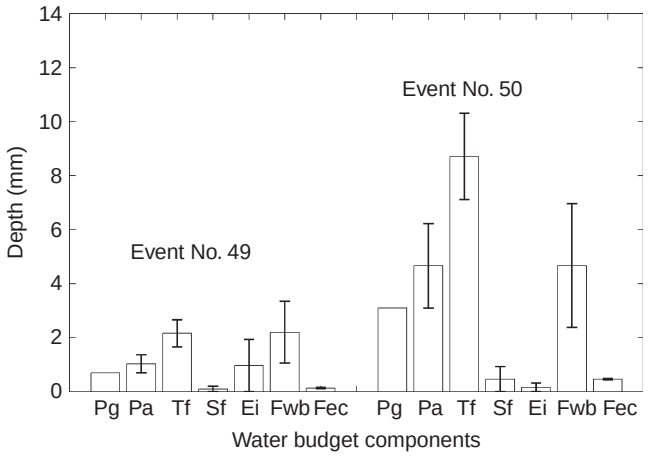


Figure 8 Wet-canopy water budget components and their (estimated) standard errors for two consecutive sampling periods (Nos. 49 and 50) in elfin cloud forest, Pico del Este, Puerto Rico (after Holwerda *et al.*, 2005). Explanation of symbols: *Pg* = gross precipitation, *Pa* = precipitation adjusted for wind- and slope effects, *Tf* = throughfall, *Sf* = stemflow, *Ei* = wet-canopy evaporation, *Fwb* = fog interception according to water budget method, *Fec* = fog deposition according to eddy covariance method

Fog as a Hydrologic Input: Selected Examples

Fog inputs are highly variable in time (Yin and Arp, 1994; Dawson, 1998; Holder, 2003; Liu *et al.*, in press) and in space, particularly in mountainous terrain (Weaver, 1972; Cavelier *et al.*, 1996; Walmsley *et al.*, 1996; cf. Figures 5 and 10). Reported figures are often difficult to compare between locations, however, because of differences in methodology (fog gauges vs. net precipitation or direct eddy covariance measurements) and duration of the measurements (short-term vs. seasonal or annual). Generally, fog deposition (interception) in an area increases with elevation within the cloud belt, mostly as a result of the concurrent

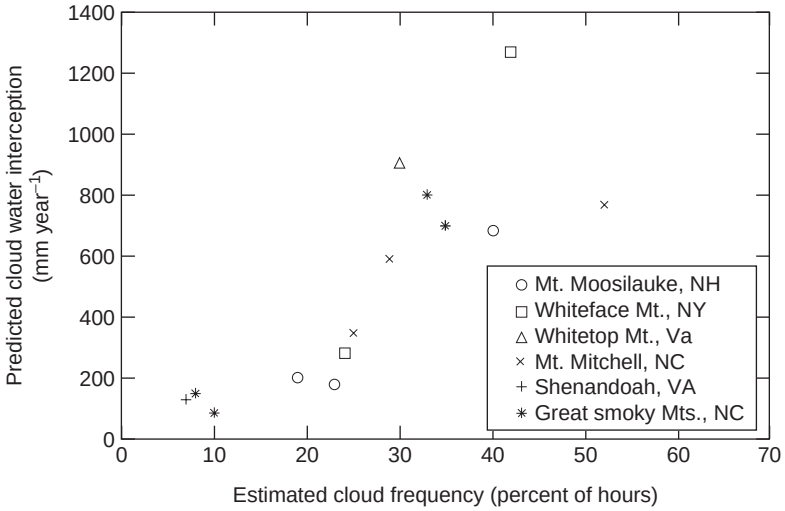


Figure 9 Comparison of fog interception and cloud frequency for several mountains in the eastern United States (drawn from tabled data in Vong *et al.*, 1991)

increases in fog frequency (immersion time) and wind speed (Figures 5, 9, and 10). The reasonably consistent relationship between fog frequency and fog deposition per region (e.g. Figure 9) indicates that the former may be used as an initial screening factor to identify areas that are likely to receive significant hydrologic (and associated chemical) inputs from fog (Vong *et al.*, 1991; Yin and Arp, 1994). Similarly, interception of advection fog in coastal areas tends to decrease with distance from the coast (Cereceda *et al.*, 2001) and Yin and Arp (1994) were able to closely predict the frequency of fog over large areas in eastern Canada as a function of distance from the coast, time of

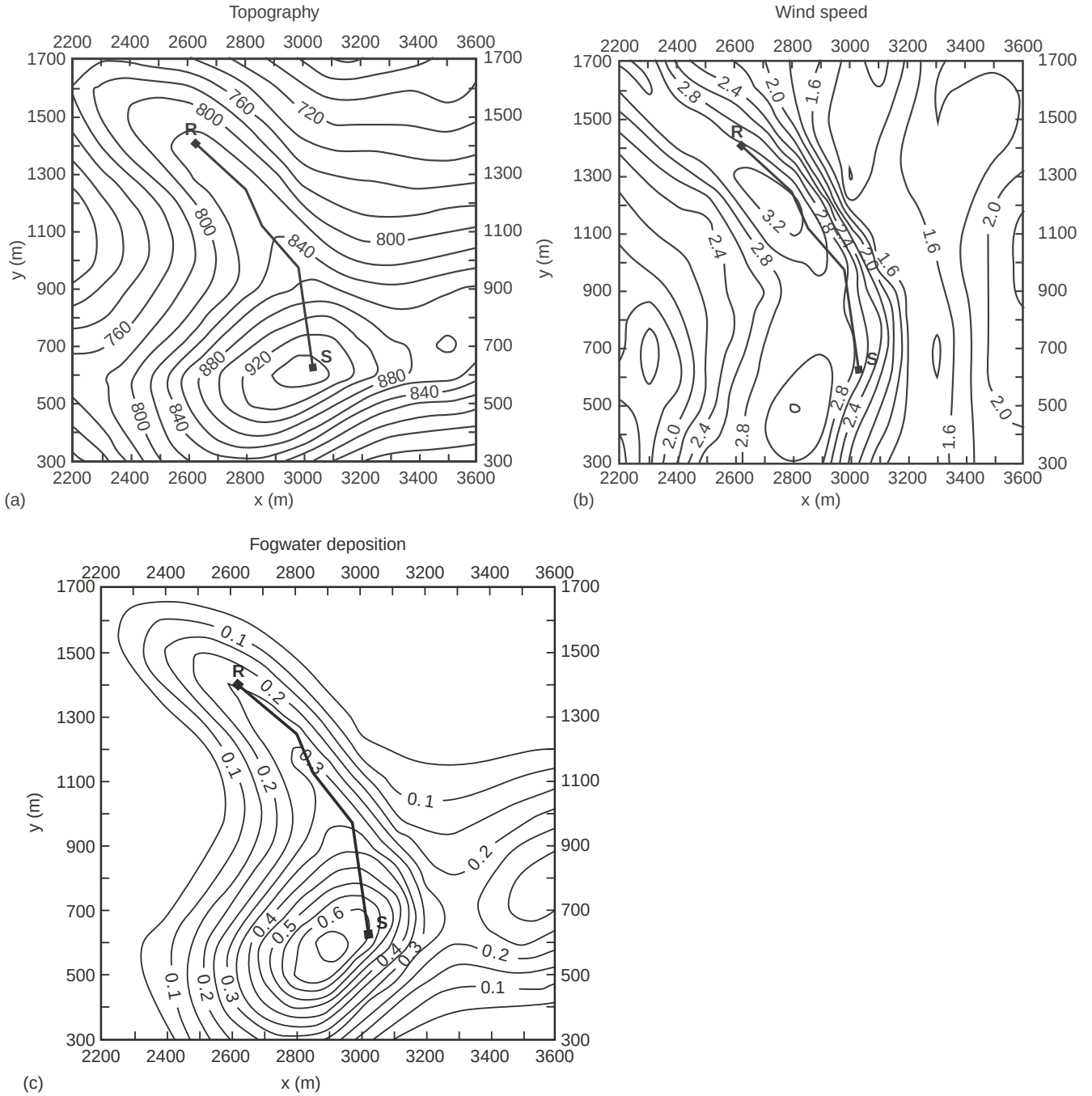


Figure 10 Spatial patterns of (a) topography (contour interval 20 m), (b) modeled wind speeds at 1.5 m above the canopy (isotach interval 0.2 m s^{-1}), and (c) modeled fog water deposition rate (isopleth interval 0.05 L m^{-2} h^{-1}) for the summit (S) and ridge (R) area on Roundtop Mountain, Quebec, Canada. Heavy line indicates ridge crest (redrawn after Walmsley *et al.*, 1996)

Table 2 Amounts of fog deposition/interception at selected sites and under contrasting climatic and topographic conditions. *MAP* = mean annual precipitation; vegetation types: ECF = elfin cloud forest; L/UMCF = lower/upper montane cloud forest; SMRF = subtropical montane rain forest; SMCF = subtropical montane cloud forest; STRF = seasonal tropical rain forest

Location	Latitude	Elevation (m asl)	MAP (mm)	Vegetation type	Duration of experiment	Fog deposition (mm d^{-1})	(% of P)	Methodological comments
Coastal zone and maritime climates								
(Sub-)tropics								
N. Chile[a]	20° 49′S	850	<70	Bare	41 months	8.5	≫100	SFC near coast
	20°12′S	1050	<70	Bare	41 months	1.1	≫100	SFC 12 km inland
Puerto Rico[b]	18° 19′N	1010	4435	ECF	44 days	2.1	17	Wet-canopy water budget; 20 roving collectors
						0.8[r]	6	ECV, no correction
						2.15[r]	18	ECV, corrected for flux divergence
Puerto Rico[c]	18° 19′N		4435	ECF	8 months			$(T_F + S_F)/P_g$; 60 collectors
Windward slope		1000				1.2/1.4	20/24	CW_{net} / CW using 4% E_I[u]
Ridge top		1015				1.8/2.0	35/39	Idem
Leeward slope		930				?/0.2	3	CW using 4% E_I[u]
Costa Rica[d]	10° 18′N	1500	2520	LMCF	12 months	2.45	28	Cylindrical wire harp
Guatemala[e]	15° 05′N	2550	2500	UMCF	12 months	0.65/1.3d	8/53d	CW_{net}; 58 fixed gauges
Hawaii[f]	20° 10′N	1170	4400	LMCF	12 months	2.15	19	CW_{net}; roving troughs
Queensland[g]	28° 13′N	1000	1350	SMRF	12 months	0.95	35	Wet-canopy water budget; 75 fixed T_F troughs
Mediterranean to temperate zone								
Canary Islands[h]	28° 07′N	1300	750	SMCF	12 months	1.8	83	SFC
						0.2	8	P_{net} during fog; 2 collectors
N California[i]	41° 33′N	120	1300	Redwood (35–38 m)	36 months	0.6	17	Cylindrical harp in open
						0.85	23	Idem inside forest
						1.2	34	Very large collectors
Oregon[j]	45° 25′N	955	2000	Conifers (45–55 m)	40 weeks	1.35	25	CW_{net}; 4 large gutters
						2.4	44	CW, E_I = 20%
					63 summer days	ca. 1.65	ca. 80	CW, E_I = 18%

Washington State[k]	48° 18′N	460	2000	Conifers	19 summer days	1.6[r]	–	ECV corrected for flux divergence
Eastern Canada[a]	44–46° N	<150	1375	Mixed	12 months	0.2–0.3	5–8	Simulated; basic climatic data and equation (7)
South Island, N.Z.[m]	45° 50′S	740	1500	Tussock grass	21 days	0.16/0.7[r]	4	ADM; 1000 h of fog year^{-1}
Cumbria, UK[n]	54° 30′N	620	2000?	Moorland	few days	0.32/0.45[r]	–	ADM; 2000 h of fog year^{-1}
Inland zone/continental climates								
South China[o]	21° 55′N	750	1485	STRF	48 months	0.4	5	12 fixed T_F collectors;
Eastern US[p]	35–44° N	900–2000	<2000	Conifer/Deciduous	Various	0.35–3.5	–	Various (model, T_F)
Eastern Germany[q]	50° 44′N	875	980	Spruce	24 months	0.29	11	ECV and water budget
						0.37	14	Lovett (1984) model
SE Germany[s]	50° 08′N	786	1185	Spruce	5.5 months	0.10/0.52[r]	4	ECV
Switzerland[t]	47° 29′N	690		Coniferous	6.5 months	0.05	3.4	ECV

[a]Cereceda *et al.* (2001).
[b]Holwerda *et al.* (2005).
[c]Weaver (1972).
[d]Clark *et al.* (1998).
[e]Holder (2003).
[f]Juvik *et al.* (2002).
[g]Hutley *et al.* (1997).
[h]García-Santos *et al.* (in press).
[i]Dawson (1998).
[j]Harr (1982).
[k]Kovalski and Vong (1999).
[l]Yin and Arp (1994).
[m]Cameron *et al.* (1997).
[n]Fowler *et al.* (1990).
[o]Liu *et al.* (in press).
[p]Vong *et al.* (1991).
[q]Zimmermann and Zimmermann (2002).
[r]per fog day.
[s]Thalmann *et al.* (2002).
[t]Burkard *et al.* (2003).
[u]as measured by Te Linde *et al.* (2001); d = dry season; SFC: standard fog collector; ECV: eddy covariance method; *CW* = fog deposition (cf. equation 3); ADM = aerodynamic method.

year (month) and monthly temperature. Finally, owing to their close association with wind, advection fogs (both in coastal and mountain areas) are typically associated with larger depositions than radiation fogs (see below).

Table 2 lists gross and net fog interception data (i.e. CW and CW_{net} as used in equations 2–5) from selected locations around the world, representing a range of conditions in terms of climatic zone, elevation, proximity to the coast, and vegetation type. Despite the methodological limitations referred to earlier, and the caution that thus needs to be applied in their interpretation, the data in Table 2 confirm that fog is a significant hydrologic input in many coastal and montane situations, regardless of climatic zone. Leaving the very high potential deposition recorded for otherwise barren coastal northern Chile (site 1 in Table 2) aside for the moment, average CW values derived for a range of evergreen forest types across the humid tropical to subtropical and warm-temperate "maritime" climatic spectrum exhibit a fairly narrow range (0.95–2.45 mm d^{-1}), with the highest values being associated with the wettest and windiest locations (Puerto Rico, Costa Rica). Naturally, the (mostly corresponding) values of CW_{net} at these and comparable sites are somewhat lower owing to the effect of wet-canopy evaporation, but they are still considerable (range: 0.65–2.15 mm d^{-1}). Moreover, higher values than these annual means are sometimes observed during the dry season, both for mountain fog (e.g. Guatemala: 1.3 mm d^{-1} vs. 0.64 mm d^{-1} on average; Holder, 2003; cf. García-Santos *et al.*, in press) and valley (radiation) fog (e.g. South China: 0.22 vs. 0.52 mm d^{-1} during the wet and dry season, respectively; Liu *et al.*, in press). Needless to say, such increases render the extra inputs by fog even more important, both from the hydrological and the ecological perspective. Returning to the most extreme example (northern Chile), there are no measurements of fog drip underneath natural vegetation in the area but throughfall measurements made beneath ten *Caesalpina* trees in the subdesertic coastal area of southern Peru suggested an average fog drip of nearly 1 mm d^{-1} (Calamini *et al.*, 1998). In northern Chile, the fog is exploited increasingly for domestic purposes using large screen collectors (cf. Schemenauer and Joe, 1989; Cereceda *et al.*, 2001).

Except at the windiest and coldest mountain sites (cf. Figure 9), fog deposition rates under cool temperate conditions tend to be much more modest (typically $\ll$1 mm d^{-1}), particularly when tall vegetation gives way to grass or moorland (Table 2). Fowler *et al.* (1990) suggested that replacing moorland vegetation in the northern UK (site n in Table 2) by ca. 20-m tall spruce forest might increase fog deposition by a factor of 3.7. In addition, because solute concentrations in cloud water are generally much higher than in rainfall (e.g. Vong *et al.*, 1991), the associated deposition of major ions might even be increased by 50% (Fowler *et al.*, 1990).

MODELING FOG DEPOSITION

Although the measurement of fog deposition on tall vegetation is fraught with difficulty and uncertainty, even for individual sites (cf. Figures 5 and 8), there is a need for sound estimates to be made over long time periods and in regions of complex terrain, both in the context of understanding forest decline in areas with high pollution loading via cloud water deposition (e.g. Vong *et al.*, 1991) and in areas where fog constitutes an important addition to the overall water budget (Vogelmann, 1973; Dawson, 1998). Such difficulties have prompted the development and testing of predictive models of fog deposition of varying complexity. Some of the more widely used models are discussed briefly below.

The general model developed by Shuttleworth (1977) (based in turn on the pioneering work of Hori (1953) and Merriam (1973)) describes the fundamental processes of cloud water deposition to, and evaporation from, a uniform vegetation canopy. In this steady-state, single-layer model, the turbulent transfer of fog droplets to the canopy from the surrounding air is modeled by direct analogy to Ohm's law for electrical circuits, in which a current representing the flux is calculated as the ratio of a potential gradient and a resistance. In this case, the flux is that of fog water to the canopy surface, the potential gradient is the difference in fog droplet concentration between the air and the receiving surface, and the resistance that to droplet deposition (Shuttleworth, 1977). Lovett (1984) adapted the Shuttleworth model by incorporating the vertical variation in an otherwise homogenous canopy structure. A series of layers was introduced in which the vertical turbulent transport of fog droplets is controlled by the aerodynamic resistances between layers, and that between the top layer and the air above the canopy. In more detail, Lovett's model consists of two submodels, the first simulating the turbulent diffusion of fog droplets into the forest and their deposition on to foliar and branch surfaces, and the second simulating evaporation and condensation during conditions of cloud immersion within the forest. The description of the turbulent transport component is based on the similarity hypothesis, which states that momentum and droplets are exchanged between the within-canopy and above-canopy air spaces with the same efficiency, that is, the droplets are assumed to follow the turbulent airstream into the canopy. Although the interior air of large clouds is generally supersaturated, thin clouds and fogs can be slightly undersaturated. Also, net radiation, which is the chief variable controlling evaporative potential, ranges from slightly positive to slightly negative depending on the time of day and depth of the cloud (Pruppacher and Klett, 1998). Therefore, Lovett (1984) simulated evaporation and condensation using a submodel that is very similar to the cloud deposition submodel, except that it accounts for

the partition of available energy into sensible and latent heat (cf. **Chapter 45, Actual Evaporation, Volume 1**). To enable comparisons to be made between model predictions of "gross" fog water interception (deposition), and "net" fog interception derived from measurements of throughfall (cf. equations (2) and (3)), the corresponding amounts of wet-canopy evaporation need to be subtracted from predicted deposition totals (Lovett, 1984; cf. **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**).

Lovett's model was developed specifically to estimate fog deposition in subalpine balsam fir forest in the northeastern US, for which highly detailed forest structural information was available. The model was subsequently generalized for use in spruce forests (Mueller, 1991) and deciduous forests (cf. Vong *et al.*, 1991), also in the eastern US. Modifications included improvements in leaf area and droplet collection efficiency parameterization schemes, diffusivity profiles above and within the canopy, and the relationship between fog droplet size distribution and liquid water content (Vong *et al.*, 1991; Mueller, 1991). Comparative experiments have revealed that the Lovett deposition model is quite sensitive to the specification of LWC and drop size distribution. Indeed, Lovett (1984) himself concluded that uncertainty in the input data, in addition to general temporal and spatial variability aspects, made his model impractical for estimations at the landscape or seasonal scale (see also below).

Another widely used model (Beswick *et al.*, 1991; Gallagher *et al.*, 1992) is the analytical model developed by Slinn (1982). Approximated analytical equations are used to model the dry deposition of droplets to vegetation canopies. The deposition velocity of a droplet is described by the cumulative effects of its sedimentation velocity, the overall drag coefficient of the canopy, wind speeds at reference height and at the top of the canopy, a universal droplet collection efficiency, and a parameter characterizing the wind profile within the canopy. The chief drawback of this model is its dependence on a large number of parameters. Moreover, Kovalski and Vong (1999) examined the validity of the assumed analogy between momentum and fog droplet fluxes and concluded that simpler models based on droplet impaction efficiencies would be physically more robust, although requiring accurate information on within-canopy structure, wind speeds, and size-dependent droplet number concentrations. Such information is rarely available at high spatial resolutions and thus the modeling of fog water deposition in this way will remain a major challenge for some time.

For most hydrologic applications, it is not essential to know the vertical distribution of fog water flux within a canopy. Thus, Unsworth and Wilshaw (1989) proposed a simpler, aerodynamic approach in combination with droplet settling velocities to derive the rate of fog water deposition D_f (expressed in $\mathrm{g\,m^{-2}\,s^{-1}}$):

$$D_f = LWC \left\{ k^2 u(z) / \left[\ln\left(\frac{z-d}{z_0} \right) \right] + v_s \right\} \qquad (7)$$

where k = von Kármán's constant (0.4), $u(z)$ = wind speed ($\mathrm{m\,s^{-1}}$) as measured at reference height z (m), d = zero plane displacement (m), z_0 = roughness length (m), v_s = deposition velocity by sedimentation ($\mathrm{m\,s^{-1}}$), and *LWC* ($\mathrm{mg\,m^{-3}}$) as defined previously (*see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1** for definitions of d and z_0). The Unsworth and Wilshaw model assumes that fog droplets are deposited by way of turbulent diffusion and sedimentation, and that turbulent diffusion proceeds at the maximum rate allowed by momentum transport. This alternative approach achieves a greater degree of spatial integration and its predictions match eddy-covariance flux measurements of fog deposition much better than the more mechanistic approaches discussed above. For example, incorporation of a fog deposition component as derived with equation (7) into the water budgets of six forested watersheds in the Maritime Provinces of eastern Canada effectively resolved earlier discrepancies between simulated and measured streamflow volumes (Yin and Arp, 1994). In areas with a noncontinuous forest cover, there is the added complication that fog interception is much higher at forest edges than toward the interior (Weathers *et al.*, 1995; Yin and Arp, 1994).

In recent years, increased effort has been directed toward assessing the spatial distribution of fog water deposition on forest in complex mountainous terrain. Walmsley *et al.* (1996) developed and applied a model for the estimation of spatial variations in fog water volume "stripped" by a coniferous forest canopy over areas of $2-164\,\mathrm{km^2}$ surrounding Roundtop Mountain (970 m) in Quebec, Canada. Their model acknowledges that amounts of fog water deposition are highly site dependent and largely governed by the following five factors: canopy structure, horizontal wind speed, collection efficiency of the tree crowns, fog liquid water content, and variation of fog frequency with altitude. Forest canopy structure was described as the summation of individual trees, where each tree was assumed to be a vertical cylinder topped by a cone. Spatial variations in wind speed and direction were calculated using the wind flow model of Bridgman *et al.* (1994). The fog collection efficiency of the treetops (i.e. the upper "cones") was assumed to be represented by that of an AES/ASRC passive fog collector, which has been shown to closely simulate fog water collection by similar types of coniferous trees (Joslin *et al.*, 1990). However, fog *LWC* values could not be derived directly from collector sample amounts because of the variation in collector efficiency with wind speed and droplet size referred to earlier (Mueller and Imhoff,

1989). Hence, *LWC* at different elevations was calculated from cloud-base height observations and the assumption that *LWC* above the cloud base was 38% of the value calculated for adiabatic ascent. This simple approach proved to be in "reasonable agreement" with *LWC* values measured earlier in the region (Walmsley *et al.*, 1996). Finally, the variation of fog frequency with altitude was assumed to be linearly related to height above sea level as demonstrated earlier for the area by Schemenauer (1986). Any effects of fog droplet size on deposition velocity (notably for moderate wind speeds; Gallagher *et al.*, 1988) were not considered because calculations had to be made over several tens of km^2 and detailed dropsize distribution data are not available at this scale (Walmsley *et al.*, 1996).

In the application, spatial variations in the wind velocity field just above the canopy were found to be closely related to the main terrain features (valleys, ridges, summits; Figure 10). Fog water deposition rate at the canopy top was specified as a function of topographic height above the cloud base (read: *LWC*) and wind speed. As a result, spatial patterns of fog water deposition strongly reflected the pattern of topographic contours (Figure 10). Although predicted deposition rates were not validated in the field, they were found to be typical of measured values reported in the literature (Walmsley *et al.*, 1996).

The models discussed so far were developed primarily to demonstrate the importance of fog water deposition to high elevation ecosystems. Compared to mountain and advective fog types, comparatively little effort has been put into the modeling of radiation fog, despite the fact that water and ion fluxes associated with such fog may be significant in particular cases (e.g. Liu *et al.*, in press). von Glasow and Bott (1999) presented a sophisticated one-dimensional model describing the (thermo)dynamics and microphysical structure of radiation fog, as well as the associated atmospheric radiative transfers in some detail (cf. Bott *et al.*, 1990).

This short overview has demonstrated, if anything, that prediction of both one-dimensional and spatial patterns of fog water deposition remains a major challenge. Moreover, although some deposition models (notably that of Lovett, 1984) have been applied to a variety of vegetation types, simulated deposition totals have only rarely been evaluated against credible measurements (cf. Vong *et al.*, 1991; Kovalski and Vong, 1999) whereas model applications in tropical montane cloud forests seem to be lacking entirely (Bruijnzeel, 2004; Bruijnzeel *et al.*, in press).

CONCLUDING REMARKS

Although the hydrologic importance of fog to many coastal and montane ecosystems (particularly under subhumid and arid conditions) is well-established, reliable quantitative information on fog deposition is relatively scarce (cf. Table 2). This reflects the intrinsic complexities of the processes involved as well as the large temporal and spatial variability of fog occurrence and deposition rates (Kovalski and Vong, 1999; Lovett, 1984). With the arrival of eddy-covariance equipment, direct measurements of net fog deposition (i.e. the flux to the canopy minus the flux in the opposite direction owing to concurrent fog formation) have become possible, but in the few examples in which the results have been compared with independent measurements (e.g. crown drip), eddy covariance based estimates were invariably much lower. Further comparative work at carefully selected locations is needed.

Similarly, a number of fog deposition models of varying complexity is available but the predicted rates of fog interception have rarely been validated (Vong *et al.*, 1991). In addition, the modeling of spatial patterns of fog deposition over larger areas (tens to hundreds of km^2) is still in its infancy although plausible estimates have been obtained using a topography-based approach (Walmsley *et al.*, 1996). To improve the degree of realism in predictions of spatial fog deposition, state-of-the-art knowledge of cloud physics will need to be combined with (micro)meteorological and hydrologic measurement of cloud-droplet spectra and fog water deposition. In particular, the approach of Walmsley and coworkers needs to be linked to actually measured *LWC* values and drop size spectra (for the computation of the gravitational component of deposition) at different elevations and for contrasting degrees of exposure.

Our understanding of the ecological role of fog interception and absorption by tall vegetation has increased in recent years for some (subhumid) ecosystems, including the *lomas* of northern Chile and Peru (Pinto *et al.*, 2001; Aravena *et al.*, 1989) and the redwoods of California (Dawson, 1998; Burgess *et al.*, 2001), partly through the application of stable isotopes. Much less is known about the influence of fog on the physiologic functioning of tropical montane cloud forests, be it absorption of fog by foliage (Yates and Hutley, 1995) or epiphytes (Hölscher *et al.*, 2004; Zotz and Hietz, 2001), the suppression or even reversal of tree water uptake (Cavelier, 1990; Burgess *et al.*, 2001), and photosynthesis (Jiménez *et al.*, 2003). Such information is especially relevant at "marginal" cloud forest sites where rainfall alone is not sufficient to sustain evergreen rain forest (Vogelmann, 1973; Hutley *et al.*, 1997). Similarly, fears have been expressed that the loss of the cloud-capturing capacity of cloud forests upon conversion to pasture or temperate vegetable cropping will lead to diminished streamflow, particularly during the dry season (Zadroga, 1981), although experimental evidence for this is limited (cf. Ingwersen, 1985; Bruijnzeel, 2001).

Ridge-top cloud forests are potentially under threat by lifting cloud base levels owing to global warming (Still *et al.*, 1999) or, more speculatively, lowland forest conversion to pasture (Lawton *et al.*, 2001). The plants and

animals living in cloud forests are finely adapted to the prevailing (and rather extreme) climatic conditions and a significant decline in the frequency or density of fog may have disastrous consequences for the vitality (or indeed survival) of populations of frogs, toads, and anoline lizards (cf. Pounds *et al.*, 1999) as well as epiphyte and bryophyte communities (cf. Zotz and Hietz, 2001). Under rather more marginal conditions in East Africa, cloud forests are increasingly subject to fire and Hemp (in press) has drawn attention to the potentially large volumes of water that may be missed once the fog intercepting capacity of tall vegetation is lost after fire.

Summarizing, although much is still to be learned (both qualitatively and quantitatively) about the hydrologic and ecological roles of fog in a variety of ecosystems, considerable progress has been made in recent years with the measurement and modeling of fog deposition. In addition, a considerable body of work is currently underway (cf. Schemenauer and Puxbaum, 2001; Bruijnzeel, 2004; Bruijnzeel *et al.*, in press) and the results of these new initiatives are awaited with interest.

REFERENCES

Aravena R., Suzuki O. and Pollastri A. (1989) Coastal fog and its relation to groundwater in the IV region of northern Chile. *Chemical Geology*, **79**, 83–91.

Arends B.G., Kos GPA, Wobrock W., Schell D., Noone K.J., Fuzzi S. and Pahl S. (1992) Comparison of techniques for measurements of fog liquid water content. *Tellus*, **44B**, 604–611.

Asbury C.E., McDowell W.H., Trinidad-Pizarro R. and Berrios R. (1994) Solute deposition from cloud water to the canopy of a Puerto Rican montane forest. *Atmospheric Environment*, **28**, 1773–1780.

Ataroff M. (1998) Importance of cloud water in Venezuelan Andean cloud forest water dynamics. In *Proceedings of the First International Conference on Fog and Fog Collection*, Schemenauer R.S. and Bridgman H.A. (Eds.), International Development Research Centre: Ottawa, pp. 25–28.

Bao B., Shu J. and Zhu B. (1995) Study of physicochemical properties of urban fog in Shanghai. *Journal of Nanjing Institute of Meteorology*, **18**, 114–118.

Baumgardner D. (1983) An analysis and comparison of five water droplet measuring instruments. *Journal of Climate and Applied Meteorology*, **22**, 891–910.

Bell A.G. (1885) Preventing collisions with icebergs in a fog. *Science*, **5**,(122), 460–461.

Bendix J. (2002) A satellite-based climatology of fog and low-level stratus in Germany and adjacent areas. *Atmospheric Research*, **64**, 3–18.

Benedetti F., Colombo C., Barbini B., Campori E. and Smeraldi E. (2001) Morning sunlight reduces length of hospitalization in bipolar depression. *Journal of Affective Disorders*, **62**, 221–223 doi:10.1016/S0165-0327(00)00149-X.

Beswick K.M., Hargreaves K.J., Gallagher M.W., Choularton T. and Fowler D. (1991) Size-resolved measurements of cloud droplet deposition velocity to a forest canopy using an eddy correlation technique. *Quarterly Journal of the Royal Meteorological Society*, **117**, 623–645.

Blake J.V. (1871) The London fog. *American Naturalist*, **5**, 76–79.

Borrmann S., Jaenicke R., Maser R. and Arends B. (1994) Instrument intercomparison study on cloud droplet size distribution measurements: Holography vs. laser optical particle counter. *Journal of Atmospheric Chemistry*, **19**, 253–258.

Bott A., Sievers U. and Zdunkowski W. (1990) A radiation fog model with a detailed treatment of the interaction between radiative transfer and fog microphysics. *Journal of Atmospheric Science*, **47**, 2153–2166.

Bridgman H.A., Walmsley J.L. and Schemenauer R.S. (1994) Modelling the spatial variations of wind speed and direction on Roundtop Mountain, Quebec. *Atmosphere-Ocean*, **32**, 605–619.

Bruijnzeel L.A. (2001) Hydrology of tropical montane cloud forests: A reassessment. *Land Use and Water Resources Research*, **1**, 1.1–1.18 `http://www.luwrr.com`.

Bruijnzeel L.A. (2004) Tropical montane cloud forests: A unique hydrological system. In *Forests, Water and People in the Humid Tropics*, Bonell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge, pp. 462–483.

Bruijnzeel L.A., Juvik J.O., Scatena F.N., Hamilton L.S. and Bubb P. (Eds.) (2006) *Forests in the Mist: Science for Conservation and Management of Tropical Montane Cloud Forests*, Hawaii University Press: Honolulu.

Bruijnzeel L.A. and Proctor J. (1995) Hydrology and biogeochemistry of tropical montane cloud forests: What do we really know? In *Tropical Montane Cloud Forests*, *Ecological Studies*, Vol. 110, Hamilton L.S., Juvik J.O. and Scatena F.N. (Eds.), Springer Verlag: New York, pp. 38–78.

Bruijnzeel L.A., Waterloo M.J., Proctor J., Kuiters A.T. and Kotterink B. (1993) Hydrological observations in montane rain forests on Gunung Silam, Sabah, Malaysia, with special reference to the 'Massenerhebung' effect. *Journal of Ecology*, **81**, 145–167.

Burgess S.S.O., Dubinsky E. and Dawson T.E. (2001) The role of fog and water relations of coast redwood Sequoia sempervirens. In *Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 121–124.

Burkard R., Bützberger P. and Eugster W. (2003) Vertical fogwater flux measurements above an elevated forest canopy at the Lägeren research site, Switzerland. *Atmospheric Environment*, **37**, 2979–2990 doi:10.1016/S1352-2310(03)00254-1.

Burkard R., Eugster W., Wrezinsky T. and Klemm O. (2002) Vertical divergence of fogwater fluxes above a spruce forest. *Atmospheric Research*, **64**, 133–145.

Burkard R., Wrzesinsky T., Eugster W. and Klemm O. (2001) Quantification of fog deposition with two similar set-ups. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 185–188.

Calamini G., Giacomin A., Falciai M., Salbitano F. and Villasante F. (1998) Fog interception and water budget of Caesalpina spinosa trees in the lomas ecosystems of Mejia

(Arequipa, Peru). In *Proceedings of the First International Conference on Fog and Fog Collection*, Schemenauer R.S. and Bridgman H. (Eds.), International Development Research Centre: Ottawa, pp. 473–476.

Cameron C.S., Murray D.L., Fahey B.D., Jackson R.M., Kelliher F.M. and Fisher G.W. (1997) Fog deposition in tall tussock grassland, South Island, New Zealand. *Journal of Hydrology*, **193**, 363–376.

Cavelier J. (1990) Tissue water relations in elfin cloud forest tree species of Serrania de Macuira, Guajira, Colombia. *Trees*, **4**, 155–163.

Cavelier J. and Mejia C. (1990) Climatic factors and tree stature in the elfin cloud forest of Serrania de Macuira, Colombia. *Agricultural and Forest Meteorology*, **53**, 105–123.

Cavelier J., Solis D. and Jaramillo M.A. (1996) Fog interception in montane forests across the Central Cordillera of Panama. *Journal of Tropical Ecology*, **12**, 357–369.

Cereceda P. and Schemenauer R.S. (1991) The occurrence of fog in Chile. *Journal of Applied Meteorology*, **30**, 1097–1105.

Cereceda P., Osses P., Larrain H., Lazaro P., Pinto R. and Schemenauer R.S. (2001) Radiation, advective and orographic fog in the Tarapaca Region, Chile. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 457–459.

Chang S.C., Lai I.L. and Wu J.T. (2002) Estimation of fog deposition on epiphytic bryophytes in a subtropical montane forest ecosystem in northeastern Taiwan. *Atmospheric Research*, **64**, 159–167.

Clark K.L., Nadkarni N.M., Schaeffer D. and Gholz H.L. (1998) Atmospheric deposition and net retention of ions by the canopy in a tropical montane forest, Monteverde, Costa Rica. *Journal of Tropical Ecology*, **14**, 27–45.

Collett J.L., Bator A., Sherman D.E., Moore K.F., Hoag K.J., Demoz B.B., Rao X. and Reilly J.E. (2002) The chemical composition of fogs and intercepted clouds in the United States. *Atmospheric Research*, **64**, 29–40.

Collett J.L., Daube B.C., Munger J.W. and Hoffmann M.R. (1990) A comparison of two cloudwater/fogwater collectors: The rotating arm collector and the caltech active strand cloudwater collector. *Atmospheric Environment*, **24A**, 1685–1692.

Curry J.A. and Webster P.J. (1999) *Thermodynamics of Atmospheres and Oceans*, *International Geophysics Series*, Vol. 65, Academic Press: San Diego.

Czarnowski M.S. and Olszewski J.L. (1970) Number and spacing of rainfall gauges in a deciduous forest stand. *Oikos*, **21**, 48–51.

Daube B., Kimball K.D., Lamar P.A. and Weathers K.C. (1987) Two new ground-level cloud water sampler designs which reduce rain contamination. *Atmospheric Environment*, **21**, 893–900.

Demoz B.B., Collett J.L. and Daube B.C. (1996) On the Caltech active strand cloudwater collectors. *Atmospheric Research*, **41**, 47–62.

Dawson T.E. (1998) Fog in the California redwood forest: Ecosystem inputs and use by plants. *Oecologia*, **117**, 476–485.

DeFelice T.P. (1998) *An Introduction to Meteorological Instrumentation and Measurement*, Prentice Hall: Upper Saddle River NJ.

DeFelice T.P. (2002) Physical attributes of some clouds amid a forest ecosystem's trees. *Atmospheric Research*, **65**, 17–34.

Dufour L. (1978) Le brouillard dans la littérature Francaise. *Publications de l'Institut Royal Météorologique de Belgique, Série B*, **94**.

Eagleson P.S. (2003) *Dynamic Hydrology*, *EGU Reprint Series*, Vol. 2, European Geophysical Union: Katlenburg-Lindau.

Ediang O.A. (2001) Piracy and armed robbery against ships and aircraft in the Niger Delta: Negative impact of fog in Nigeria. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 421–424.

Eugster W., Burkard R., Holwerda F., Bruijnzeel S., Scatena F.N. and Siegwolf R. (2002) Fogwater inputs to a cloud forest in Puerto Rico. *Eos Transactions of the American Geophysical Union*, **83**,(47), Fall Meeting Supplement, Abstract H52A-0828.

Eugster W., Burkard R., Klemm O. and Wrzesinsky T. (2001) Fog deposition measurements with the eddy covariance method. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 193–196.

Falconer R.E. and Falconer P.D. (1980) Determination of cloud water acidity at a mountain top observatory in the Adirondack Mountains of New York State. *Journal of Geophysical Research*, **85**, 7465–7470.

Fedorova N., Dal Piva E. and de Carvalho M.H. (2001) Investigation of radiation fog formation on the south coast of Brazil. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 395–398.

Fitzjarrald D.R. and Lala G.G. (1989) Hudson valley fog environments. *Journal of Applied Meteorology*, **28**, 1303–1328.

Fowler D., Morse A.P., Gallagher M.W. and Choularton T.W. (1990) Measurements of cloud water deposition on vegetation using a lysimeter and a flux gradient technique. *Tellus*, **42B**, 285–293.

Frahm J.P. and Gradstein S.R. (1991) An altitudinal zonation of tropical rain forests using bryophytes. *Journal of Biogeography*, **18**, 669–676.

Gallagher M.W., Beswick K.M. and Choularton T.W. (1992) Measurement and modelling of cloudwater deposition to a snow-covered forest canopy. *Atmospheric Environment*, **26A**, 2893–2903.

Gallagher M.W., Choularton T.W., Morse A.P. and Fowler D. (1988) Measurements of the size dependence of cloud droplet deposition at a hill site. *Quarterly Journal of the Royal Meteorological Society*, **114**, 1291–1303.

García-Santos G., Regalado C.M., Ritter A. and Marzol M.V. (2006) Water balance and fog features in a Laurisilva subtropical montane cloud forest, Garajonay National Park, La Gomera, Canary Islands. In *Forests in the Mist: Science for Conservation and Management of Tropical Montane Cloud Forests*, Bruijnzeel L.A., Juvik J.O., Scatena F.N., Hamilton L.S. and Bubb P. (Eds.), Hawaii University Press: Honolulu.

Gioda A., Maley J., Espejo Guasp R. and Baladón A.A. (1995) Some low elevation fog forests of dry environments:

Applications to African paleoenvironments. In *Tropical Montane Cloud Forests*, *Ecological Studies*, Vol. 110, Hamilton L.S., Juvik J.O. and Scatena F.N. (Eds.), Springer Verlag: New York, pp. 156–164.

von Glasow R. and Bott A. (1999) Interaction of radiation fog with tall vegetation. *Atmospheric Environment*, **33**, 1333–1346.

Glickman T.S. (Ed.), (2000) *Glossary of Meteorology*, *Second Edition*, American Meteorological Society: Boston.

Golding D.L. (1970) The effects of forest on precipitaiton. *The Forestry Chronicle*, **October 1970**, 397–402.

Goodman J. (1977) Microstructure of California coastal fog and stratus. *Journal of Applied Meteorology*, **16**, 1056–1067.

Goodman J. (1985) The collection of fog drip. *Water Resources Research*, **21**, 392–394.

Gordon C.A., Herrera R. and Hutchinson T.C. (1994a) Studies of fog events at two cloud forests near Caracas, Venezuela – I. Frequency and duration of fog. *Atmospheric Environment*, **28**, 323–337.

Gordon C.A., Herrera R. and Hutchinson T.C. (1994b) Studies of fog events at two cloud forests near Caracas, Venezuela – II. Chemistry of fog. *Atmospheric Environment*, **28**, 323–337.

Grubb P.J. (1977) Control of forest growth and distribution on wet tropical mountains: With special reference to mineral nutrition. *Annual Review of Ecology and Systematics*, **8**, 83–107.

Hafkenscheid RLLJ (2000) *Hydrology and Biogeochemistry of Montane Rain Forests of Contrasting Stature in the Blue Mountains of Jamaica*, Ph.D. thesis, Vrije Universiteit, Amsterdam.

Hafkenscheid RLLJ, Bruijnzeel L.A. and De Jeu RAM (1998) Estimates of fog interception by montane rain forest in the Blue mountains of Jamaica. In *Proceedings of the First International Conference on Fog and Fog Collection*, Schemenauer R.S. and Bridgman H.A. (Eds.), International Development research Centre: Ottawa, pp. 33–36.

Hafkenscheid RLLJ, Bruijnzeel L.A., De Jeu RAM and Bink N.J. (2002) Water budgets of two upper montane rain forests of contrasting stature in the Blue mountains, Jamaica. In *Proceedings of the Second International Colloquium on Hydrology and Water Management*, Gladwell J.S. (Ed.), CATHALAC: Panamá City, pp. 399–424.

Harr R.D. (1982) Fog drip in the Bull Run municipal watershed, Oregon. *Water Resources Bulletin*, **18**, 785–789.

Hemp A. (2006) Climatic change and its impact on montane cloud forests: Fire as a determinant in the subalpine zone of Mt. Kilimanjaro, Tanzania. In *Forests in the Mist: Science for Conservation and Management of Tropical Montane Cloud Forests*, Bruijnzeel L.A., Juvik J.O., Scatena F.N., Hamilton L.S. and Bubb P. (Eds.), Hawaii University Press: Honolulu.

Herwitz S.R. and Slye R.E. (1992) Spatial variability in the interception of inclined rainfall by a tropical rainforest canopy. *Selbyana*, **13**, 62–71.

Hölscher D., Köhler L., van Dijk AIJM and Bruijnzeel L.A. (2004) The importance of epiphytes to total rainfall interception by a tropical montane rain forest in Costa Rica. *Journal of Hydrology*, **292**, 308–322.

Holder C.D. (2003) Fog precipitation in the Sierra de las Minas Biosphere Reserve, Guatemala. *Hydrological Processes*, **17**, 2001–2010.

Holwerda F., Burkard R., Eugster W., Scatena F.N., Meesters AGCA and Bruijnzeel L.A. (2005) Estimating fog deposition at a Puerto Rican elfin cloud forest site: Comparison of the water budget and eddy covariance methods. *Hydrological Processes*.

Hori T. (1953) *Studies in Fogs (In Relation to Fog-Preventing Forests)*, Tanne Trading Co.: Sapporo.

Hursch C.R. and Pereira H.C. (1953) Field moisture balance in the Shimba Hills. *East African Agricultural Journal*, **18**, 139–148.

Hutley L.B., Doley D., Yates D.J. and Boonsaner A. (1997) Water balance of an Australian subtropical rainforest at altitude: The ecological and physiological significance of intercepted cloud and fog. *Australian Journal of Botany*, **45**, 311–329.

Ingraham N.L. (1998) Isotopic variations in precipitation. In *Isotope Tracers in Catchment Hydrology*, Kendall C. and McDonnell J.J. (Eds.), Elsevier: Amsterdam, pp. 87–118.

Ingraham N. (2001) The hydrologic origin of fog water: A stable isotopic analysis. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 369–372.

Ingwersen J.B. (1985) Fog drip, water yield, and timber harvesting in the Bull Run municipal watershed, Oregon. *Water Resources Bulletin*, **21**, 469–473.

Inoue K., Yoshimoto M. and Abe H. (1997) Cloud physical property of sea fog induced by Yamase in the Sanriku seashore. *Journal of Agricultural Meteorology*, **53**, 21–28.

Jiménez M.S., González-Rodriguez A.M., Peters J. and Morales D. (2003) Tenerife, the paradise for plant environmental physiology (from the desert to humid habitats), Keynote paper to the 5th *International workshop on Field Techniques for Environmental Physiology*, Tenerife, pp. 1–10, 16–22 March 2003.

Joslin J.D., Mueller S.F. and Wolfe M.H. (1990) Test of models of cloudwater deposition to forest canopies using artificial and living collectors. *Atmospheric Environment*, **24A**, 2893–2903.

Juvik J.O. and Nullet D. (1995a) Relationships between rainfall, cloud-water interception and canopy throughfall in a Hawaiian montane forest. In *Tropical Montane Cloud Forests*, *Ecological Studies*, Vol. 110, Hamilton L.S., Juvik J.O. and Scatena F.N. (Eds.), Springer Verlag: New York, pp. 165–182.

Juvik J.O. and Nullet D. (1995b) Comments on 'A proposed standard fog collector for use in high elevation regions'. *Journal of Applied Meteorology*, **34**, 2108–2110.

Juvik J., Liwai J. and Delay J. (2002) The contribution of direct cloud-water interception to canopy throughfall in a Hawaiian tropical montane cloud forest, Paper presented at the *Association of Tropical Biology Symposium on Tropical forests: Past, Present and Future*, Panamá City, 29 July – 2 August 2002.

Kashiwabara K., Kohrogi H., Ota K. and Moroi T. (2002) High frequency of emergency room visits of asthmatic children on misty or foggy nights. *Journal of Asthma*, **39**, 711–717.

Kerfoot O. (1968) Mist precipitation on vegetation. *Forestry Abstracts*, **29**, 8–20.

Kovalski A.S., Anthoni P.M., Vong R.J., Delany A.C. and MacLean G.D. (1997) Deployment and evaluation of a system for ground-based measurement of cloud liquid water turbulent fluxes. *Journal of Atmospheric and Oceanic Technology*, **14**, 468–479.

Kovalski A.S. and Vong R.J. (1999) Near-surface fluxes of cloud water evolve vertically. *Quarterly Journal of the Royal Meteorological Society*, **125**, 2663–2684.

Krovetz D.O. (1988) Assembly and field testing of a ground-based presence-of-cloud detector. *Journal of Atmospheric and Oceanic Technology*, **5**, 579–581.

Kunkel B.A. (1984) Parameterization of droplet terminal velocity and extinction coefficient in fog models. *Journal of Applied Meteorology*, **23**, 34–41.

Larrain H., Cereceda P., Pinto R., Lazaro P., Osses P. and Schemenauer R.S. (2001) Archeological observations at a coastal fog-site in Alto Patache, South of Iquique, Northern Chile. In *Proceedings of the Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 289–292.

Lawton R.O., Nair U.S., Pielke R.A. Sr and Welch R.M. (2001) Climatic impact of tropical lowland deforestation on nearby montane cloud forests. *Science*, **294**, 584–587.

Lin N.H. and Saxena V.K. (1991) In-cloud scavenging and deposition of sulfates and nitrates: Case studies and parametrization. *Atmospheric Environment*, **24A**, 2302–2320.

Liu W.L., Meng F.R., Zhang Y.P., Liu Y.H. and Li H.M. (2006) Fog drip and fog chemistry in the tropical seasonal rain forest of Xishuangbanna, South-west China. In *Forests in the Mist: Science for Conservation and Management of Tropical Montane Cloud Forests*, Bruijnzeel L.A., Juvik J.O., Scatena F.N., Hamilton L.S. and Bubb P. (Eds.), Hawaii University Press: Honolulu.

Lloyd C.R. and Marques Filho A.de O. (1988) Spatial variability of throughfall and stemflow measurements in Amazonian rain forest. *Forest and Agricultural Meteorology*, **42**, 63–73.

Lovett G.M. (1984) Rates and mechanisms of cloud water deposition to a subalpine balsam fir forest. *Atmospheric Environment*, **18**, 361–371.

Mallant R.K.A.M. (1988) Poor man's optical fog detector. *Annalen der Meteorologie*, **25**, 333–334.

Marzol M.V., Sanchez-Mgia J.L., Valladares P., Perez-Gonzalez P. and Dorta P. (1996) La captacion del agua del mar de nubes en Tenerife. Metodo e instrumental. In *Clima y Agua: La Gestion de un Recurso Climatico*, Marzol M.V., Dorta P. and Valladares P. (Eds.), Tabapress: Madrid, pp. 333–350.

Marzol M.V. (2006) Historical background of fog water collection studies in the Canary Islands. In *Forests in the Mist: Science for Conservation and Management of Tropical Montane Cloud Forests*, Bruijnzeel L.A., Juvik J.O., Scatena F.N., Hamilton L.S. and Bubb P. (Eds.), Hawaii University Press: Honolulu.

Merriam R.A. (1973) Fog drip from artificial leaves in a fog wind tunnel. *Water Resources Research*, **9**, 1591–1598.

Meyer M.B. and Lala G.G. (1990) Climatological aspects of radiation fog occurrence at Albany, New York. *Journal of Climate*, **3**, 577–586.

Mitchell M.D. and Suckling P.W. (1987) Winter fog and air transportation in Sacramento, California. *Climatological Bulletin*, **21**, 16–22.

Mueller S.F. (1991) Estimating cloud-water deposition to subalpine spruce-fir forests – I. Modifications to an existing model. *Atmospheric Environment*, **25A**, 1093–1104.

Mueller S.F. and Imhoff R.I. (1989) Inferring cloud deposition to a forest canopy using a passive cloudwater collector. *Geophysical Research Letters*, **16**, 683–686.

Nadezhina E.D. and Shklyarevich O.B. (1994) Advection fogs and glaze over slope in coastal areas. *Russian Meteorology and Hydrology*, **9**, 13–20.

Nakanishi M. (2000) Large-eddy simulation of radiation fog. *Boundary-Layer Meteorology*, **94**, 461–493.

Nakata T. (1982) Steam fog in the neighbourhood of Hiroshima Bay. *Umi to Sora (Sea and Sky)*, **57**, 187–191.

Noonkester V.R. (1979) Coastal marine fog in Southern California. *Monthly Weather Review*, **107**, 830–851.

Øfkland H. and Gotaas Y. (1995) Modelling and prediction of steam fog. *Beiträge zur Physik der Atmosphäre*, **68**, 121–131.

Oliver H.R. (1980) Fenland steam fog. *Weather*, **35**, 118–120.

Padilla H.P. (1998) *Comparacion de la Composicion Quimica de las Nubes y de la Precipitation Pluvial entre diferentes zonas Montañosas de Mexico*. PhD Thesis, Universidad Nacional Autonoma de Mexico, Mexico city.

Panzram H. (1975) Strassenglätte, Nebel und Kraftverkehr. *Naturwissenschaftliche Rundschau*, **28**, 437–438.

Patel K.S., Tripathi A.N., Chandrawanshi C.K., Aggarwal S.G., Patel R.M., Deb M.K., Agnihotri P.K. and Patel V.K. (1998) In *Proceedings of the First International Conference on Fog and Fog Collection*, Schemenauer R.S. and Bridgman H.A. (Eds.), International Development research Centre: Ottawa, pp. 309–312.

Pilié R.J. (1975) Life cycle of valley fog, Part 1. Micrometeorological characteristics. *Journal of Applied Meteorology*, **14**, 347–363.

Pilié R.J. (1979) Formation of marine fog and the development of fog-stratus systems along the California coast. *Journal of Applied Meteorology*, **18**, 1275–1286.

Pinto R., Larrain H., Cereceda P., Lazaro P., Osses P. and Schemenauer R.S. (2001) Monitoring fog-vegetation communities at a fog site in Alto Patache, South of Iquique, Northern Chile, during 'El NiZo' and 'La NiZa' events (1997–2000). In *Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 293–296.

Pounds J.A., Fogden M.P.A. and Campbell J.H. (1999) Biological response to climate change on a tropical mountain. *Nature*, **398**, 611–615.

Pristov J. and Trontelj M. (1979) Nebelerscheinungen in den Gebirgstälern. *Proceedings of the 7th Mezhdunarodnaya Konferentsiya po Meteorologii Karpat*, Slovensky Akademija: Vied Tatranska Lomnitsa, pp. 387–397, 21–25 Sept., 1975.

Pruppacher H.R. and Klett J.D. (1998) *Microphysics of Clouds and Precipitation*, Kluwer Academic Publishers: Dordrecht.

Richards P.W. (1996) *A Tropical Rain Forest*, *Second Edition*, Cambridge University Press: Cambridge.

Roach W.T. (1995) Back to basics: Fog. Part 3: The formation and dissipation of sea fog. *Weather*, **50**, 80–84.

Robinson P.J. (1989) The influence of weather on flight operations at the Atlanta Hartsfield International Airport. *Weather and Forecasting*, **4**, 461–468.

Russell S. (1984) Measurement of mist precipitation. Techniques Notebook. *The Bryological Times*, **25**, 4.

Sachweh M. and Koepke P. (1997) Fog dynamics in an urbanized area. *Theoretical and Applied Climatology*, **58**, 87–93.

Sakurai K. and Ohtake T. (1979) On the condensation and ice nuclei contained in supercooled droplet and ice fog particles. *Journal des Recherches Atmosphériques*, **13**, 291.

Schell D., Georgii H.-W., Maser R., Jaeschke W., Arends B.G., Kos G.P.A., Winkler P., Schneider T., Berner A. and Kruisz C. (1992) Intercomparison of fog water samplers. *Tellus*, **44B**, 612–631.

Schemenauer R.S. (1986) Acidic deposition to forests: The 1985 Chemistry of High Elevation Fog (CHEF) Project. *Atmosphere-Ocean*, **24**, 303–328.

Schemenauer R.S. and Cereceda P. (1994) A proposed standard fog collector for use in high-elevation regions. *Journal of Applied Meteorology*, **33**, 1313–1322.

Schemenauer R.S. and Cereceda P. (1995) Reply to comments by Juvik and Nullet (1995). *Journal of Applied Meteorology*, **34**, 2111–2112.

Schemenauer R.S. and Joe P.I. (1989) The collection efficiency of a large fog collector. *Atmospheric Research*, **24**, 53–69.

Schemenauer R.S. and Puxbaum H. (Eds.) (2001) *Proceedings of the Second International Conference on Fog and Fog Collection*, International Development Research Centre: Ottawa, St. John's, Canada, 15–20 July 2001.

Scholl M.A., Gingerich S.A. and Tribble G.W. (2002) The influence of microclimates and fog on interpretation of regional hydrology using $*^{18}$O and *D: East Maui, Hawaii. *Journal of Hydrology*, **264**, 170–184.

Seinfeld J.H. and Pandis S.N. (1998) *Atmospheric Chemistry and Physics – From air Pollution to Climate Change*, John Wiley and Sons: Chichester.

Sharon D. (1980) The distribution of hydrologically effective rainfall incident on sloping ground. *Journal of Hydrology*, **46**, 165–188.

Shuttleworth W.J. (1977) The exchange of wind-driven fog and mist between vegetation and the atmosphere. *Boundary-Layer Meteorology*, **12**, 463–489.

Sigmon J.T., Gilliam F.S. and Partin M.E. (1989) Precipitation and throughfall chemistry for a montane hardwood forest ecosystem: Potential contributions from cloud water. *Canadian Journal of Forest Research*, **19**, 1240–1247.

Slinn W.G.N. (1982) Predictions for particle deposition to vegetative canopies. *Atmospheric Environment*, **16**, 1785–1794.

Smirnova W.G., Benjamin S.G., and Brown J.M. (2000) Case study: Verification of RUC/MAPS of fog and visibility forecasts. Preprints, *Ninth Conference on Aviation, Range, and Aerospace Meteorology*, American Meteorological Society: Orlando, pp. 31–36.

Stadtmüller T. (1987) *Cloud Forests in the Humid Tropics. A Bibliographic Review*. The United Nations University: Tokyo and Centro Agronomico Tropical de Investigacion y Ensenanza: Turrialba.

Still C.J., Foster P.N. and Schneider S.H. (1999) Simulating the effects of climate change on tropical montane cloud forests. *Nature*, **398**, 608–610.

Sugden A.M. (1982) The vegetation of the Serrania de Macuira, Guajira, Colombia: A contrast of arid lowlands and an isolated cloud forest. *Journal of the Arnold Arboretum*, **63**, 1–30.

Te Linde A.H., Bruijnzeel L.A., Groen J., Scatena F.N. and Meijer H.A.J. (2001) Stable isotopes in rainfall and fog in the Luquillo Mountains, eastern Puerto Rico: A preliminary study. In *Second International Conference on Fog and Fog Collection*, Schemenauer R.S. and Puxbaum H. (Eds.), International Development Research Centre: Ottawa, pp. 181–184.

Thalmann E., Burkard R., Wrzesinsky T., Eugster W. and Klemm O. (2002) Ion fluxes from fog and rain to an agricultural and a forest ecosystem in Europe. *Atmospheric Research*, **64**, 147–158.

Tiesel R. and Foken T. (1987) Zur Entstehung des Seerauchs an der Ostseeküste vor Warnemünde. *Zeitschrift für Meteorologie*, **37**, 173–176.

Trautner F., Frank G., Tschiersch J. and Voigt G. (1990) A method to quantify wet and occult deposition of aerosols to wheat plants. *Journal of Aerosol Science*, **21**, 295–298.

Unsworth M.H. and Wilshaw J.C. (1989) Wet, occult and dry deposition of pollutants on forests. *Agricultural and Forest Meteorology*, **47**, 221–238.

Valente R.J. (1989) Field intercomparison of ground-based cloud physics instruments at Whitetop Mountain, Virginia. *Journal of Atmospheric and Oceanic Technology*, **6**, 396–406.

Vermeulen A.T., Wyers G.P., Römer F.G., Leeuwen N.F.M.V., Draaijers G.P. and Erisman J.W. (1997) Fog deposition on a coniferous forest in the Netherlands. *Atmospheric Environment*, **31**, 375–386.

Vogelmann H.W. (1973) Fog precipitation in the cloud forests of Eastern Mexico. *BioScience*, **23**, 96–100.

Vong R.J., Sigmon J.T. and Mueller S.F. (1991) Cloud water deposition to Appalachian forests. *Environmental Science and Technology*, **25**, 1014–1021.

Vong R.J. and Kovalski A.S. (1995) Eddy correlation measurements of size dependent cloud droplet turbulent fluxes to complex terrain. *Tellus*, **47B**, 331–352.

Walmsley J.L., Schemenauer R.S. and Bridgman H.A. (1996) A method for estimating the hydrologic input from fog in mountainous terrain. *Journal of Applied Meteorology*, **35**, 2237–2249.

Wanner H. (1979) Zur Bildung, Verteilung und Vorhersage winterlicher Nebel im Querschnitt Jura-Alpen. *Geographica Bernensia*, **G7**, 240. pp.

Weathers K.C., Lovett G.M. and Likens G.E. (1995) Cloud deposition to a spruce forest edge. *Atmospheric Environment*, **29**, 665–672.

Weaver P.L. (1972) Cloud moisture interception in the Luquillo Mountains of Puerto Rico. *Caribbean Journal of Sciences*, **12**, 129–144.

Weinstein A.I. (1974) Projected utilization of warm fog dispersal systems at several major airports. *Journal of Applied Meteorology*, **13**, 788–795.

Whiteman C.D. and McKee T.B. (1982) Breakup of temperature inversions in deep mountain valleys: Part II: Thermodynamic model. *Journal of Applied Meteorology*, **21**, 290–302.

Yates D.J. and Hutley L.B. (1995) Foliar uptake of water by wet leaves of *Sloanea woolsii*, an Australian subtropical rainforest tree. *Australian Journal of Botany*, **43**, 157–167.

Yin X.W. and Arp P.A. (1994) Fog contributions to the water budget of forested watersheds in the Canadian Maritime Provinces: A generalized algorithm for low elevations. *Atmosphere-Ocean*, **32**, 553–566.

Zadroga F. (1981) The hydrological importance of a montane cloud forest area of Costa Rica. In *Tropical Agricultural Hydrology*, Lal R. and Russell E.W. (Eds.), John Wiley & Sons: Chichester, pp. 59–73.

Zhou B. (1987) Numerical modeling of radiation fog. *Acta Meteorologica Sinica*, **45**, 21–29.

Zhou S. (1991) The formation and features of fog in Shanghai urban area. *Quarterly Journal of Applied Meteorology*, **2**, 140–146.

Zhou S. and Wang X. (1991) The urban moisture island and urban fog of Shanghai. *Quarterly Journal of Applied Meteorology*, **2**, 256–263.

Zimmermann L. and Zimmermann F. (2002) Fog deposition to Norway Spruce stands at high-elevation sites in the Eastern Erzgebirge (Germany). *Journal of Hydrology*, **256**, 166–175.

Zingg T. (1944) Die Nebel- und Hochnebelhäufigkeiten in Dübendorf 1938/1944. *Annalen der Meteorologischen Zentralanstalt Zürich*, **1944**, 4. pp.

Zotz G. and Hietz P. (2001) The physiological ecology of vascular epiphytes: Current knowledge, open questions. *Journal of Experimental Botany*, **52**, 2067–2078.

39: Surface Radiation Balance

MICHAEL H UNSWORTH

Atmospheric Sciences Group, College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, US

The surface radiation balance is the algebraic sum of the net solar radiation and the net long-wave radiation received by a surface. It may be measured directly with a net radiometer, deduced from measurements of its upwelling and downwelling components, or modeled from climatological, geographical, and atmospheric data. Methods of modeling and measuring each of the components are discussed. Downwelling solar radiation from cloudless skies may be modeled from knowledge of atmospheric properties, but effects of clouds require more empirical approaches most suitable for monthly means. Downwelling long-wave radiation below cloudless skies is well related to temperature and humidity near the ground. An empirical method to account for clouds is given. Reflection coefficients and emissivities of natural surfaces are tabulated for use in calculating upwelling short and long-wave radiation. Net radiometer designs are reviewed, and examples are given of the surface radiation balance and its components over forest and grass. Issues of spatial variability of the surface radiation balance are discussed.

INTRODUCTION

Radiation from the sun is the initial source of the energy that drives the hydrologic cycle. But transformations of the solar radiation reaching the earth in the atmosphere and at the earth's surface result in there being several streams of upwelling and downwelling radiant energy in various wavebands at the earth's surface, and it is the net balance of these streams (just like the balance in a bank account) that provides the main resource that is partitioned at the surface into latent and sensible heating. So the surface radiation balance Q^* is a fundamental component of the surface-atmosphere energy exchange. In spite of its importance, Q^* is seldom measured as a climatological variable; more commonly it is computed from measured or modeled components. Modeling approaches may include numerically complex schemes such as those that are included in large-scale atmospheric models (e.g. (Niemela *et al.*, 2001a,b), which require vertical profiles of atmospheric properties, or they may rely on more empirical approaches on the basis of observed climatological variables at the surface. This review will therefore include discussion of the measurement and modeling of components of Q^* before considering methods and examples of direct measurements of the surface radiation balance.

DEFINITION

The surface radiation balance Q^* for a horizontal surface is

$$Q^* = K^* + L^* = K_\downarrow - K_\uparrow + L_\downarrow - L_\uparrow \qquad (1)$$

where K and L are the short- and longwave components; arrows represent the direction of the flux, and * signifies a net flux.

COMPONENTS OF THE SURFACE RADIATION BALANCE

Downwelling Shortwave Radiation

The shortwave radiation falling on a horizontal surface is the remainder after processes of scattering and absorption in the atmosphere have had their effect. At the mean distance of the earth from the sun, which is 1.50×10^8 km, the incident radiant flux density (termed the *irradiance*) on a surface held perpendicular to the solar beam is

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

known as the *Solar Constant*. The name is misleading because this quantity, with a mean value of $1366\,W\,m^{-2}$ (Crommelynck *et al.*, 1995), varies by about $\pm 1\,W\,m^{-2}$ in phase with the sunspot number (Frohlich and Lean, 1998; Willson and Mordvinov, 2003). The path length of the solar beam through the atmosphere varies with latitude, season, and time of day in ways that can be predicted from astronomical formulae (Monteith and Unsworth, 1990). As the beam passes through the atmosphere it is modified in quantity, quality, and direction by the processes of scattering and absorption.

Scattering has two main forms. Gas molecules in the atmosphere scatter light almost uniformly in all directions (Rayleigh scattering). The effectiveness is proportional to the inverse fourth power of the wavelength so that blue light (400 nm) is scattered about nine times more effectively than red light. Thus in a cloudless sky some of the direct beam radiation becomes diffuse radiation and reaches the ground from the (blue) hemisphere. Dust in the atmosphere and thin clouds scatter more dominantly forward (Mie scattering), but multiple scattering below thick clouds is more spatially uniform. The efficiency of scattering tends to become independent of wavelength as particle or droplet size increase. Consequently, overcast or very dusty (turbid) skies appear white.

Absorption of solar radiation in the cloudless atmosphere is mostly from ozone, water vapor, carbon dioxide, and oxygen. Clouds may absorb up to 20% of solar radiation.

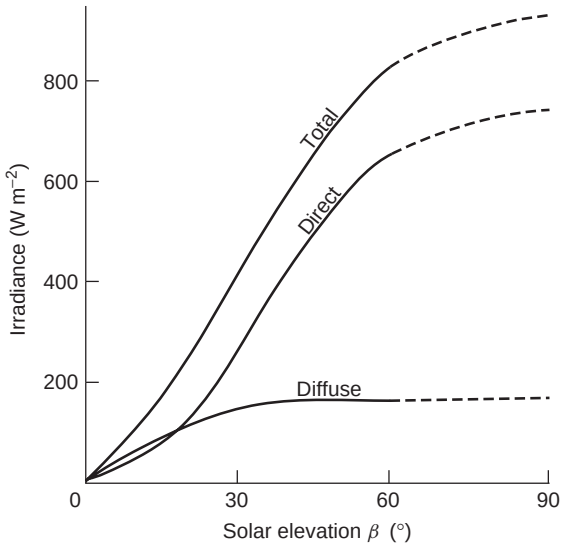


Figure 1 Variation of solar radiation components on a horizontal surface on a cloudless day at Sutton Bonington (53° N, 1° W), showing diffuse and direct irradiance and their sum, the total (global) solar irradiance. The full lines are from measurements and the dashed lines are approximate extrapolations to illustrate how the terms would vary at larger solar elevations (Reprinted from Principles of Environmental Physics, (1990) with permission from Elsevier)

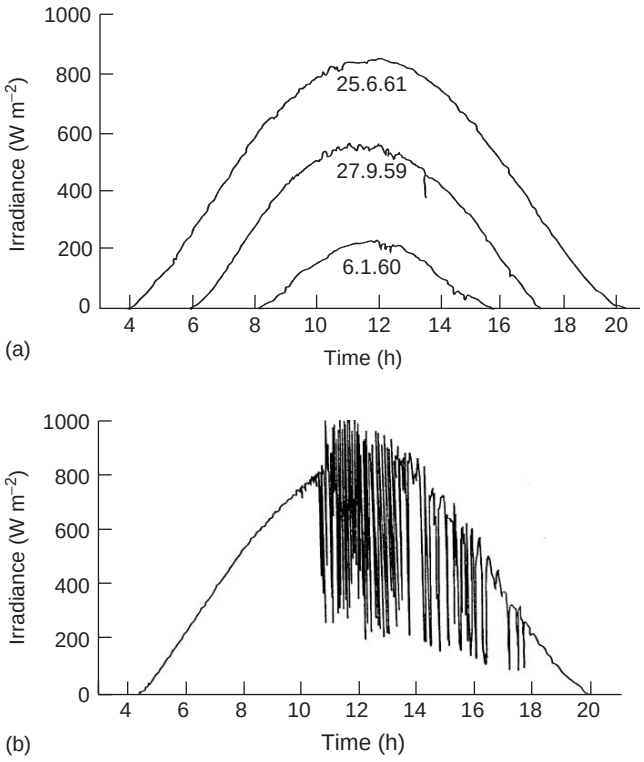


Figure 2 (a) Global solar irradiance on a horizontal surface on three cloudless days at Rothamsted, UK (52° N, 0° W). During the middle of the days, the records fluctuate more than in the morning and evenings, probably because of diurnal changes in dust loading in the atmosphere (Reprinted from Principles of Environmental Physics, (1990) with permission from Elsevier) (b) Global solar irradiance on a horizontal surface on a day of broken cloud at Rothamsted, UK (52° N, 0° W). Very high values of irradiance occurred immediately before and after the sun was occluded by clouds because radiation is scattered downwards from the cloud margins. The minimum values are the diffuse irradiance when the sun was completely obscured (Reprinted from Principles of Environmental Physics, (1990) with permission from Elsevier)

Solar Radiation at the Ground

As a consequence of scattering and absorption, solar radiation has two distinct directional properties when it reaches the ground. Direct radiation arrives from the direction of the sun; diffuse radiation includes all the other solar radiation received from the blue sky and from clouds either by scattering or transmission. The sum of the flux densities for direct and diffuse radiation on a horizontal surface is known as *global or total solar radiation* $K\downarrow$ ($W\,m^{-2}$). Figure 1 shows a typical example of the variation of each of these components with solar elevation on a cloudless day. Figure 2(a) shows the daily variation of $K\downarrow$ on three cloudless days and Figure 2(b) shows the daily variation on a day when broken clouds developed. When there is a broken cloud, the irradiance is enhanced when clouds are close to the sun and scatter radiation toward

the ground, but when clouds obscure the sun the radiation becomes entirely diffuse.

Measuring Solar Radiation

Solar radiation at the ground is in the waveband 0.3–3.0 μm, and the standard instruments for measuring it (pyranometers) should have a uniform response over this waveband so that they are equally effective for global or diffuse (blue sky) radiation. The best-known instruments are made by Eppley (**http://www.eppleylab.com/**) and Kipp and Zonen (**http://www.kippzonen.com/**) and use thermopile sensors coated with a black pigment. An alternative method for measuring global solar radiation is to use a pyranometer with a silicon photovoltaic sensor (**http://env.licor.com/Products/Sensors/rad.htm**), but these sensors detect only a small fraction of the spectral range and so should be used only in unobstructed global radiation (i.e. not for diffuse or reflected radiation unless they are specifically recalibrated for such measurements against a thermopile pyranometer).

Modeling Solar Radiation

Apart from astronomical and geographical factors (latitude, elevation), $K\downarrow$ is influenced by cloud, turbidity, absorption, and scattering. Empirical models based on these factors or other climatological variables have been developed by several investigators including Kasten (1983), Thornton and Running (1999), and Meza and Varas (2000). Methods for calculating instantaneous values of $K\downarrow$ are given by Campbell and Norman (1998). Methods using observations of sunshine hours or cloud cover can be very effective for estimating $K\downarrow$ for monthly mean or other longer period calculations. Iziomon and Mayer (2001) concluded that the Angstrom–Prescott formula agreed with observations within 2.5% for a lowland site and 3.4% for a mountain site when its empirical coefficients were calibrated for the sites. The formula is:

$$\frac{\Sigma K\downarrow}{\Sigma E_0} = A + B\left(\frac{S}{S_0}\right) \qquad (2)$$

where $\Sigma K\downarrow$ and ΣE_0 are the daily totals of global solar radiation and extraterrestrial radiation (MJ m^{-2}) respectively, and S and S_0 denote the corresponding sunshine duration and day length (hrs). Values of E_0 and S_0 can be found in List (1966). The disadvantage of equation (2) is that the coefficients A and B are somewhat site dependent. Iziomon and Mayer (2001) found mean values $A = 0.21 \pm 0.02$ and $B = 0.58 \pm 0.02$ for 12 sites in Germany.

Upwelling Shortwave Radiation $K\uparrow$

Upwelling short-wave radiation $K\uparrow$ is the product of $K\downarrow$ and the shortwave reflection coefficient of the surface ρ, also termed the *albedo*. Typical albedos for several surface types are given in Table 1, where data are taken from Campbell and Norman (1998), Monteith and Unsworth (1990), Offerle *et al.* (2003) and Gash and Nobre (1997). Values of ρ are influenced by the amount of vegetation cover, soil type, soil wetness, and solar elevation; the values in the table should be regarded as approximate for high solar elevations. In particular, tall canopies and smooth water surfaces have albedos that increase at low solar elevation. The albedo of clear, still water is almost constant at about 0.05 when the solar elevation exceeds 45°, but increases rapidly with decreasing elevation, approaching 1.0 at grazing incidence (Monteith and Unsworth, 1990) (*see* **Chapter 44, Evaporation from Lakes, Volume 1**).

Upwelling shortwave radiation can be measured with an inverted thermopile pyranometer. Special care is necessary at low sun angles to avoid measuring radiation that has been reflected and refracted in the instrument domes.

Downwelling Long-wave Radiation $L\downarrow$

Downwelling long-wave radiation $L\downarrow$ at the surface originates from radiatively active gases in the atmosphere (principally water vapor and carbon dioxide) and from clouds when present. Gaseous emission is present in specific wavebands in the range 0.3–30 μm, whereas

Table 1 Reflection coefficients (albedos) for solar radiation

Surface	Reflection coefficient	Surface	Reflection coefficient
Grass	0.24–0.26	Coniferous forest	0.05–0.15
Wheat	0.16–0.26	Tropical rainforest	0.12–0.14
Barley	0.23	Snow, fresh	0.75–0.95
Maize	0.18–0.22	Snow, old	0.40–0.70
Potato	0.19	Soil, wet dark	0.08
Sugar cane	0.15	Soil, dry dark	0.13
Cotton	0.21	Soil, wet light	0.10
Tundra	0.15–0.20	Soil, dry light	0.18
Tropical pasture	0.17–0.19	Sand, dry	0.35
Deciduous forest	0.10–0.20	Urban areas	0.10–0.27
		Water	0.05 (solar elevation >45°)

emission from clouds corresponds to black body radiation at the temperature of cloud base. From cloudless skies, more than half the long-wave flux received at the ground comes from gases in the lowest 100 m, and roughly 90% from the lowest kilometer. Consequently, $L\downarrow$ is often estimated as

$$L\downarrow = \varepsilon_a \sigma T_a^4 \tag{3}$$

where ε_a is an estimated broadband atmospheric emissivity, σ is the Stefan–Boltzmann constant ($5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$), and T_a is a bulk atmospheric temperature (in Kelvin) usually approximated by air temperature near the ground. Formulae for estimating ε_a for cloudless skies have been reviewed by Prata (1996) and Niemela *et al.* (2001a). The Prata equation (Prata, 1996) was most successful

$$\varepsilon_a = 1 - (1 + aw)e^{-(b+cw)^{0.5}} \tag{4}$$

where w is the precipitable water content of the atmosphere, given approximately by

$$w = \frac{4.65 e_a}{T_a} \tag{5}$$

where e_a is the vapor pressure near the ground (the units of w are $\mathrm{kg\,m^{-2}}$ when e_a is in Pa and T_a is in K; the empirical constants are $a = 0.10\,\mathrm{kg^{-1}\,m^2}$, $b = 1.2$, $c = 0.30\,\mathrm{kg^{-1}\,m^2}$). By using equations (4) and (5) with equation (3), values of $L\downarrow$ from cloudless skies can be calculated. A simpler expression for $L\downarrow$ from cloudless skies was given by Monteith and Unsworth (1990)

$$L\downarrow = a + b\sigma T_a^4 \tag{6}$$

where the empirical coefficients determined in central England were $a = -119 \pm 16\,\mathrm{W\,m^{-2}}$ and $b = 1.06 \pm 0.04$. A similar expression was derived by Swinbank (1963) from measurements in Australia.

Effects of clouds on $L\downarrow$ are harder to model because cloud fraction, height, and type have radiative impacts. For climatological averages, Monteith and Unsworth (1990) suggested that the emissivity $\varepsilon_a(c)$ of sky covered with a fraction c of cloud was given by

$$\varepsilon_a(c) = (1 - 0.84c)\varepsilon_a + 0.84c \tag{7}$$

Offerle *et al.* (2003) and Sugita and Brutsaert (1993) developed more complex expressions to account for clouds.

Measuring Downwelling Long-wave Radiation

Downwelling long-wave radiation may be measured with a pyrgeometer: a thermopile instrument that uses a silicon dome with a vacuum-deposited interference filter that allows transmission of long-wave radiation while excluding shortwave radiation. Pyrgeometers (precision infrared radiometers) are made by Eppley (**http://www.eppleylab.com**) and by Kipp and Zonen (**http://www.kippzonen.com/product/cg4.pdf**). There has been concern that pyrgeometers register incorrectly large values of $L\downarrow$ when the domes are exposed to direct solar radiation, which causes heating. Correction procedures have been described (Albrecht and Cox, 1977; Weiss, 1981), and some users prefer to employ a shading disk that avoids sun striking the dome.

Table 2 Long-wave emissivities of natural and artificial surfaces

Surface	Emissivity
Vegetation leaf	0.94–0.99
Soil	0.93–0.96
Water	0.96
Concrete	0.88–0.93

Upwelling Long-wave Radiation $L\uparrow$

It is often assumed that the long-wave emissivity ε_s of most natural surfaces is unity, that is, they are perfect emitters for long-wave radiation, so that $L\uparrow$ can be estimated from knowledge of the surface temperature T_s (in Kelvin) as $L\uparrow = \varepsilon_s \sigma T_s^4$ with $\varepsilon_s = 1$. Although this is often a reasonable assumption for radiation balance estimates, Table 2 shows that the emissivities of natural surfaces are generally a few percent less than unity, and for more precise calculations this should be taken into account as follows. Since long-wave absorptivity is equal to long-wave emissivity, a small fraction of $L\downarrow$ will be reflected from the surface, augmenting the emitted flux. Thus, $L\uparrow$ is given by

$$L\uparrow = \varepsilon_s \sigma T_s^4 + (1 - \varepsilon_s)L\downarrow \tag{8}$$

Calculations using equation (8) with typical values from equations (3–7) can be used to show that the error in assuming an emissivity of unity for the surface is usually very small, particularly under cloudy skies.

Upwelling long-wave radiation may be measured with an inverted pyrgeometer, or it may be estimated using a radiation thermometer assuming that emissivity is equal to 1.0.

SURFACE RADIATION BALANCE Q^*

The surface radiation balance Q^* can be deduced from measurements or models of the separate components described above, or it may be measured directly with a net radiometer. A net radiometer is a thermopile instrument capable of detecting downwelling and upwelling solar and long-wave radiation and performing the algebraic subtraction indicated

in equation (1). The basic design of most net radiometers has changed little since the reviews by Fritschen and Gay (1979) and Woodward and Sheehy (1983). Single-structure instruments have upward and downward facing thermopiles covered with domes of thin polyethylene that transmits short and long-wave radiation. The thermopiles are connected in opposition so that they generate a signal proportional to the difference between the downward and upward radiant fluxes. The sign convention is that fluxes toward the surface are treated as positive. Halldin and Lindroth (1992) compared and evaluated six net radiometer designs, and Hodges and Smith (1997) described the intercalibration of net radiometers at 21 sites. Two commonly used commercial instruments are the REBS Q7-1 (manufactured by Radiation and Energy Systems, Seattle, Washington, and marketed by Campbell Scientific (**http://www.campbellsci.com/solarrad.html**)), and the NR-LITE net radiometer (manufactured by Kipp and Zonen (**http://www.kippzonen.com**)). Recently, an instrument consisting of four independent sensors for the components of Q^* has become available (i.e. with upward and downward facing pyranometers and pyrgeometers configured so that the individual signals and the net radiation are outputs) from Kipp and Zonen (model CNR1). Examples of measurements with this system are shown in Figures 3 and 4.

The Figures show the surface radiation balance over short grass and over an old growth Douglas fir/Western Hemlock forest in the Pacific Northwest of the United States. By day, solar radiation was the dominant influence on the radiation balance. Reflected solar radiation was much less from the forest than the grass. The upwelling long-wave radiation varied less over the forest because the tree canopy was

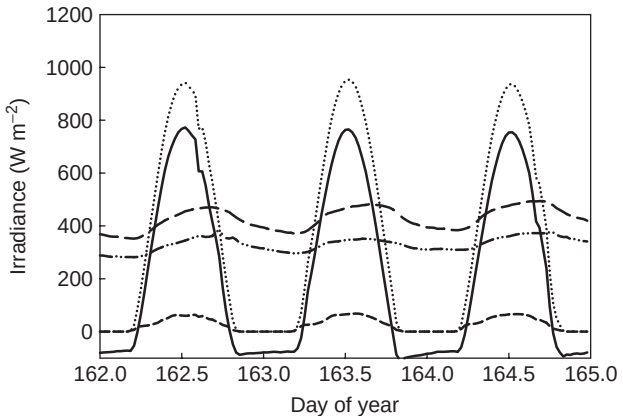


Figure 3 Net radiation and its components measured above an old growth Douglas fir/Western hemlock forest on three almost cloudless days (June 11–13, 2002) at Wind River, Washington, USA (Data courtesy of Ken Bible, University of Washington) Symbols: Q^* ——; K↓ · · · · ; K↑- - - - ; L↓—·—·—·—; L↑———

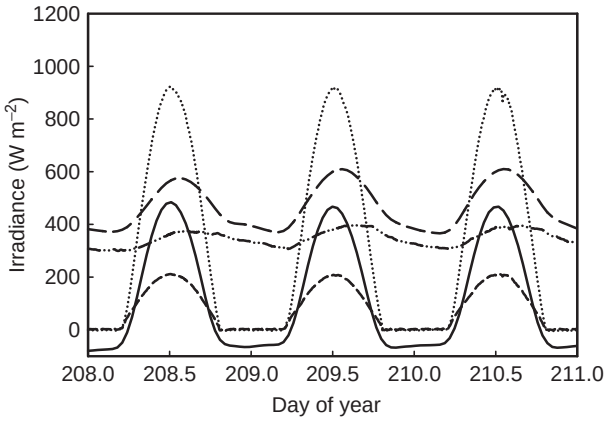


Figure 4 Net radiation and its components measured above an unirrigated field of short grass on three cloudless days (July 27–29, 2003) at Hyslop Farm, Corvallis, Oregon, USA (Data courtesy of Reina Nakamura, Oregon State University) Symbols: Q^* ——; K↓ · · · · ; K↑- - - - ; L↓—·—·—·—; L↑———

well coupled to the atmosphere in the day so that it is not heated as much by the sun as the short grass. At night, the surface radiation balances over both surfaces are similar in magnitude, suggesting that heat flow from the soil was not effective in keeping the grass canopy warmer than the more isolated forest canopy.

Figures 3 and 4 demonstrate that there is a good correlation between Q^* and $K\downarrow$ which can be used to estimate Q^* from global solar radiation records (Kaminsky and Dubayah, 1997). But the principles summarized in this article show that such an approach is not only site dependent, but also surface dependent, so that methods for estimating Q^* from its components are preferable (Offerle *et al.*, 2003).

Influences of Surface Heterogeneity and Large-Scale Averaging

Net radiation measurements are usually made with a single instrument deployed over the surface. For uniform surfaces such as agricultural crops, this single measurement may be a good representation of the spatial average surface radiation budget, suitable for use in energy balance studies. But for spatially inhomogeneous surfaces such as natural ecosystems or urban areas, a point measurement of Q^* may not be appropriate. Anthoni *et al.* (2000) used measurements and a simulation model to estimate the spatial variability of upwelling short- and long-wave radiation in a spatially heterogeneous juniper sagebrush ecosystem, and Offerle *et al.* (2003) compared three methods of estimating Q^* over urban areas. Hodges and Smith (1997) used point measurements of Q^* together with satellite measurements of K^* to produce a map of net radiation over a large part of the region where the BOREAS experiment (*see* **Chapter 177,**

The Role of Large-Scale Field Experiments in Water and Energy Balance Studies, Volume 5) was conducted (50° N–60° N, 110° W–95° W). They found that clouds had a greater influence on the surface radiation budget than ecosystem variability at this scale.

Acknowledgments

I am grateful to Ken Bible, University of Washington, and Reina Nakamura, Oregon State University, for providing the data for Figures 3 and 4 respectively.

REFERENCES

Albrecht B. and Cox S.K. (1977) Procedures for improving pyrgeometer performance. *Journal of Applied Meteorology*, **16**, 188–197.

Anthoni P.M., Law B.E., Unsworth M.H. and Vong R.J. (2000) Variation of net radiation over heterogeneous surfaces: measurements and simulation in a juniper-sagebrush ecosystem. *Agricultural and Forest Meteorology*, **102**, 275–286.

Campbell G.S. and Norman J.M. (1998) *Environmental Biophysics*, *Second Edition*, Springer-Verlag: New York, pp. 286.

Crommelynck D., Fichot A., Lee R.B. and Romero J. (1995) First realisation of the space absolute radiometric reference (SARR) during the Atlas 2 flight period. *Advances in Space Research*, **16**, 17–23.

Fritschen L.J. and Gay L.W. (1979) *Environmental Instrumentation.*, Springer Verlag: New York, pp. 216.

Frohlich C. and Lean J. (1998) The sun's total irradiance: cycles, trends and related climate change uncertainties since 1976. *Geophysical Research Letters*, **25**, 4377–4380.

Gash J.H.C. and Nobre C.A. (1997) Climatic effects of Amazonian deforestation: some results from ABRACOS. *Bulletin of the American Meteorological Society*, **78**, 823–830.

Halldin S. and Lindroth A. (1992) Errors in net radiometry: comparison and evaluation of six radiometer designs. *Journal of Atmospheric and Oceanic Technology*, **9**, 762–783.

Hodges G.B. and Smith E.A. (1997) Intercalibration,objective analysis, intercomparison and synthesis of BOREAS surface net radiation measurements. *Journal of Geophysical Research*, **102**, 28885–28900.

Iziomon M.G. and Mayer H. (2001) Performance of solar radiation models–a case study. *Agricultural and Forest Meteorology*, **110**, 1–11.

Kaminsky K.Z. and Dubayah R. (1997) Estimation of surface net radiation in the boreal forest and northern prairiefrom shortwave flux measurements. *Journal of Geophysical Research*, **102**, 29707–29716.

Kasten F. (1983) Parametriserung der Globalstrahlung durch Bedeckungsgrad und Trubungsfactor. *Annals fur Meteorologie*, **20**, 49–50.

List R.J. (Ed.) (1966) *Smithsonian Meteorological Tables*, *Sixth Edition*, Smithsonian Institution: Washington.

Meza F. and Varas E. (2000) Estimation of mean monthly solar global radiation as a function of temperature. *Agricultural and Forest Meteorology*, **100**, 231–241.

Monteith J.L. and Unsworth M.H. (1990) *Principles of Environmental Physics*, *Second Edition*, Edward Arnold: London, p. 291.

Niemela S.P., Raisanen P. and Savijarvi H. (2001a) Comparison of surface radiative flux parameterizations. Part 1: Longwave radiation. *Atmospheric Research*, **58**, 1–18.

Niemela S.P., Raisanen P. and Savijarvi H. (2001b) Comparison of surface radiative flux parameterizations. Part II: Shortwave radiation. *Atmospheric Research*, **58**, 141–154.

Offerle B., Grimmond C.S.B. and Oke T.R. (2003) Parameterization of net all-wave radiation for urban areas. *Journal of Applied Meteorology*, **42**, 1157–1173.

Prata A.J. (1996) A new longwave formula for estimating downward clear-sky radiation at the surface. *Quarterly Journal of the Royal Meteorological Society*, **122**, 1127–1151.

Sugita M. and Brutsaert W. (1993) Cloud effect in the estimation of instantaneous downward longwave radiation. *Water Resources Research*, **29**, 599–605.

Swinbank W.C. (1963) Longwave radiation from clear skies. *Quarterly Journal of the Royal Meteorological Society*, **89**, 339–348.

Thornton P.E. and Running S.W. (1999) An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. *Agricultural and Forest Meteorology*, **93**.

Weiss A. (1981) On the performance of pyrgeometers with silicon domes. *Journal of Applied Meteorology*, **20**, 962–965.

Willson R.C. and Mordvinov A.V. (2003) Secular total solar irradiance trend during solar cycles 21–23. *Geophysical Research Letters*, **30**, 1199–1202.

Woodward F.I. and Sheehy J.E. (1983) *Principles and Measurements in Environmental Biology*, Butterworths: London.

40: Evaporation Measurement

ANNE VERHOEF[1] AND CLAIRE L CAMPBELL[2]

[1]*Department of Soil Science, The University of Reading, Reading, UK*
[2]*Centre for Ecology and Hydrology, Edinburgh, UK*

This chapter summarises the practical methods available to estimate evaporation, E. It describes the background of methods based on soil physical, micrometeorological and plant physiological concepts. The most appropriate method will mainly depend on required spatial and temporal scales of E, accuracy needed, and the available resources (financial and manpower). These issues are taken into account during the discussion.

INTRODUCTION

The aim of this chapter is to provide an overview of the techniques available for measurement of land surface evaporation. This review will focus on obtaining high-frequency evaporation estimates based on scientific methods and relatively sophisticated instrumentation. The methods used to measure evaporation are based on the concepts of micrometeorology, plant physiology, or the soil water balance (SWB, see equation 1). The most appropriate method(s) will be determined by the scientific research question or the practical purpose for which the evaporation data are required.

Scientific research will generally need values of evaporation at small time steps (e.g. half-hours), and at a variety of spatial scales (leaf, plant, plot, and field). Often, knowledge of the separate components of evaporation (soil and plant) may be required to provide a full understanding of the atmosphere-vegetation-soil interactions related to the evaporation process. Micrometeorological methods are generally the most expensive, but they have the advantage of providing field-scale area-average estimates of evaporation at a fine time resolution. The SWB-based and plant physiological methods will supply small-scale spatial estimates at a range of timescales (half-hourly to weekly). These can be scaled up to provide field-scale evaporation, using estimates of leaf area index (LAI), and soil/vegetation coverage fractions.

We will discuss the methods most commonly used to measure evaporation. For each method, the underlying principles will be described, and the method will be evaluated against relevant experimental design criteria. These criteria include the spatial scale, temporal frequency and accuracy of the data, labor intensity, sources of error, and the relative cost.

The alternative method of estimating evaporation is by modeling. Models range from those calculating actual evaporation by multiplication of the potential evaporation (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**) by a factor depending on crop type, development stage, and soil water status (see Allen *et al.*, 1998), to detailed mechanistic Soil Vegetation Atmosphere Transfer models (*see* **Chapter 45, Actual Evaporation, Volume 1**).

We will begin by describing the two key equations used in measurement of evaporation. First, the water balance for a soil profile, which is given by (see Hillel, 1998):

$$\Delta S_M = P + I_R + C_R - E - D - R \qquad (1)$$

where ΔS_M is the change in soil moisture content during a given period, P is the precipitation, I_R the amount of irrigation, C_R the capillary rise, E the amount of water evaporated, D the deep drainage, and R the run off or run on. P, I_R, and C_R are gain terms, E and D are loss terms, whereas R can be either a gain or a loss. In equation 1, the terms have units of depth of water per unit of time, typically mm H_2O day^{-1}.

Secondly, the surface energy balance (SEB), which is given by:

$$Q^* - G = H + \lambda E \qquad (2)$$

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

where λ is the latent heat of vaporization ($\sim$2.45 MJ kg^{-1}). Equation 2 is an important boundary condition for energy exchange processes at the land surface, determining how available energy (net radiation flux density, Q^*, minus soil heat-flux density, G) is partitioned between sensible heat-flux density, H and latent heat-flux density, λE. The SEB is directly related to the SWB through the evaporation term, λE (see Brutsaert, 1982; Hillel, 1998).

In the SEB, we are dealing with energy fluxes in units of W m^{-2}. E in mm day^{-1} can be converted to λE in W m^{-2} given that 1 kg m^{-2} H_2O = 1 mm H_2O.

METHODS BASED ON THE SOIL WATER BALANCE

Weighing lysimetry and soil physical methods use the SWB to determine evaporation. These methods depend on being able to either ignore (e.g. drainage, D), successfully measure, or estimate the various gain and loss terms in equation 1, so that we can estimate E.

Weighing Lysimetry

The first and most direct method is weighing lysimetry, which involves monitoring weight changes of an extracted, undisturbed soil column. Evaporation is calculated from differences between regularly measured masses corrected for precipitation, irrigation, and often drainage. The top of the lysimeter is mounted flush with the soil surface, and it is assumed that the water loss from the lysimeter is representative of that of the surrounding area.

Lysimeters vary in size and depth, mainly depending on the type of evaporation of interest (total, plant, or soil) and the type of vegetation. Large (surface area of approximately 2–10 m^2 and depths ranging between 2 to 4 m) to medium-sized (usually $<$1 m^2 and maximum depths of 1.5 m) lysimeters are mainly used to measure total evaporation, that is, vegetation plus soil (e.g. Dugas and Bland, 1989; Allen and Fisher, 1990). Additionally, they can measure transpiration if the bare soil under the crop is covered to prevent soil evaporation (see Klocke *et al.*, 1985), or provide estimates of soil evaporation if the spacing between the canopy elements (e.g. trees or crop rows) is large enough. The smallest lysimeters, the microlysimeters (see Boast and Robertson, 1982), are used for measurements of soil evaporation below canopies. These consist of cores, which weigh typically 1–3 kg, with diameters ranging between about 5 and 20 cm, and depths of between 5 and 30 cm.

Lysimeters vary in their weighing method. Large lysimeters will use either balance beam and counterweight mechanisms, or the hydraulic principle (see Allen and Fisher, 1990). Generally, the medium types are completely supported by load-cells. Microlysimeters are usually weighed manually (see Allen, 1990; Daamen *et al.*, 1993).

The spatial scale of the measurement is equal to the lysimeter surface area. Concerning the measurement frequency, most lysimeters will allow continuous weighing, yielding half-hourly estimates of E, whereas microlysimeters are often weighed only once a day, although increased scale accuracy allows a higher frequency, if manpower is available. Most lysimeters are intended for long-term use, for example, to determine interannual variation of seasonal E.

Crucial practical details to consider with lysimetry are the depth of the lysimeter, minimization of soil and plant disturbance during extraction, the boundary to water flow often imposed at the base of the lysimeter (e.g. D and C_R are zero for closed-bottom microlysimeters), minimization of edge effects, and choice of lysimeter wall material. The lysimeter material, particularly if made of metal, can have a significant effect on the soil heat-flux and evaporation regime. If used below a vegetated surface, the depth of the lysimeter should be below the expected maximum root depth. The LAI of the plants in the lysimeter should be representative of the study site (see Dugas and Bland, 1989).

The resolution of lysimeters depends on the weighing equipment and on the resolution of the dataloggers. Lysimeters of the balance beam and hydraulic type have a typical resolution of 0.02 mm. The smaller, load-cell based lysimeters usually have a reduced resolution of about 0.05 mm (Allen and Fisher, 1990). Wind, temperature, and hysteresis effects can influence the accuracy of the method (Allen and Fisher, 1990). For microlysimeters, the magnitude of the errors depends on the width and depth of the microlysimeter and on the time since the core was extracted from the soil profile. Daamen *et al.* (1993) proposed a protocol for the use of microlysimeters in water balance studies.

Finally, the cost and manpower requirements of the facility should be considered. A large-sized lysimeter set up, including datalogging equipment, has a similar cost to the mid-to-high range micrometeorological methods. Manpower requirements are relatively low after installation; the installation of the larger type lysimeters is not straightforward (see Howell *et al.*, 1985). Microlysimeters, often made in-house, are extremely cheap and use standard laboratory scales. However, they put a considerable demand on manpower because of their requirement for manual weighing.

Soil Physical Methods

Another way to determine evaporation on the basis of the SWB is by estimating its other terms separately for an *in situ* soil profile, and inverting the equation to obtain E. This method can be used to measure total evaporation or soil evaporation where the spacing is such that the soil moisture regime between canopy elements is unaffected by root water uptake.

It is relatively easy to quantify P, by employing raingauges. Also, if the surface is relatively flat, R can be ignored. Very often, capillary rise, C_R, is assumed to be negligible, which is a reasonable assumption if the soil is coarse-textured and groundwater levels are deep. Detailed measurements of soil moisture content, θ, over time, are required to derive ΔS_M, the change in profile soil water content during a given period. There are various ways to measure θ, as will be discussed below. However, to get reliable estimates of E, it is crucial to know the partitioning between E and D and there are three methods to achieve this.

1. The simplest method is based on the assumption that the soil moisture content is equivalent to or less than field capacity (*see* **Chapter 72, Measuring Soil Water Content, Volume 2**). Thus, drainage below the root zone is assumed to be negligible, and E can be calculated directly from ΔS_M. This is a reasonable assumption in some cases (e.g. a deep, sandy soil), but generally (slow) drainage will continue for weeks after a considerable rainfall or irrigation event at levels comparable to the water loss by evaporation. Using this method can lead to errors in the estimation of E of up to 30% (Van Bavel *et al.*, 1968). In some cases D is approximated as a certain percentage of rainfall or as a function of soil moisture content (Klay and Vachaud, 1992).
2. In this method, D is found by employing a time-integrated form of Darcy's law:

$$D = \int F \, \mathrm{d}t = \int -k(\theta)\frac{\Delta \Psi_h}{\Delta z} \, \mathrm{d}t \qquad (3)$$

   F is the downward flux of soil water through the bottom of the soil profile, $k(\theta)$ is the hydraulic conductivity as a function of θ, and $\Delta \Psi_h / \Delta z$ is the change of hydraulic potential with depth. This technique requires additional vertical measurements of the matric potential, Ψ_m, using tensiometers, from which Ψ_h can be easily obtained (see Figure 1), as well as estimates of the $k(\theta)$ function. For details of methods for determination of k and Ψ_h, see Rowell (1994). Klay and Vachaud (1992) propose a simplified method, based on equation 3, for use with sandy soils in semiarid conditions.
3. This method is based on the determination of the zero-flux plane, ZFP (see McGowan and Williams, 1980). The zero-flux plane (ZFP) is the level above which upward water flux, that is, evaporation takes place, and below which downward flow due to drainage takes place. It can be located as the depth at which the hydraulic potential gradient, $\Delta \Psi_h / \Delta z$, equals zero (see Figure 1). Once the position of the ZFP has been found, the changes in soil water stored above the ZFP can be attributed to evaporation, and those below the ZFP attributed to drainage.

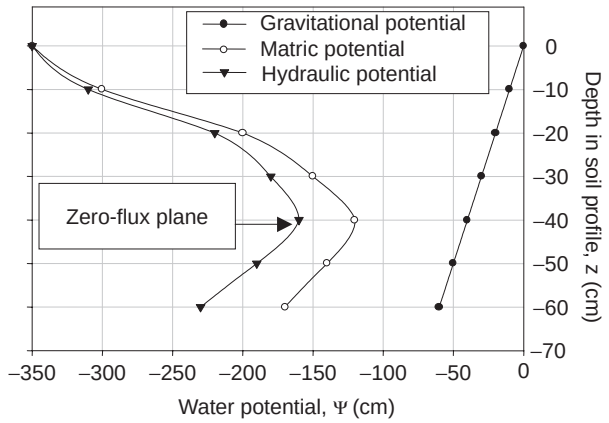


Figure 1 This diagram shows values of the soil hydraulic potential, Ψ_h, against depth in the soil profile. Ψ_h is the sum of Ψ_g (the gravitational potential) and Ψ_m (the matric potential). Ψ_m can be measured using tensiometers or derived from the soil water retentivity curve if values of soil moisture content are known. Ψ_g equals the depth in the profile, if, as in this case, the potential are expressed in length units. This diagram illustrates the concept of the ZFP, that is, the point where the gradient of hydraulic potential with depth, $\Delta \Psi_h / \Delta z$, is zero. Soil water movement above this plane will be upwards, below it the water will flow downwards

Using the ZFP concept to divide between E and D assumes that plant roots are not extracting water from below the ZFP, which is not always realistic. Also, it is very hard to accurately determine the ZFP as root water uptake may be too small to perturb the hydraulic gradient enough to enable unequivocal detection of the ZFP.

A time series of the θ profile (for derivation of ΔS) may be achieved using neutron probes, γ-ray scanners (Hillel, 1998), Time Domain Reflectometry (Topp *et al.*, 1980; Roth *et al.*, 1990), Profile probes or a profile of single Thetaprobes (Delta-T Devices, 1999; 2001, for example). Excepting the Thetaprobe and Time Domain Reflectometry sensors, these instruments require the installation of access tubes. In the first two techniques, the probe is lowered manually and operated at various depths, typically every 7–10 days. The Profile probe estimates θ from the dielectric constant of soil, and is essentially a further development of the widely used Thetaprobes. These may be connected to a data logger to provide continuous records of θ. The 1-m long Profile probe can be installed permanently to provide continuous concurrent measurements at six depths. Each of the soil moisture instruments described has a small sphere of influence (thereby limiting representativity), ranging from about two centimeters maximum diameter for the Profile probes to about 30 cm for the neutron probe or γ-ray scanners. The spatial scale of the evaporation

estimate is given by the diameter of the sphere of influence of the soil moisture measurement. To obtain a representative plot or field average ΔS_M, several access tubes should be distributed over the area of interest.

Temporal frequency of E estimates depends on the frequency of θ measurements, or may be determined by their reliability at short time steps. We recommend that Profile and Thetaprobe soil moisture measurements should not be used for time steps of less than 24 hours, since they can be sensitive to daily temperature variation. Soil type and soil moisture content appear to determine the magnitude of the temperature effect.

Unless the soil is very gravelly or has a high clay content, access tubes and tensiometers are relatively easy to install. Use of the neutron probe or γ-ray scanners is labor intensive, and requires supplementary gravimetric measurements, as these instruments are unreliable within 30 cm of the soil surface. Furthermore, these two instruments can pose a health hazard if handled incorrectly. All methods mentioned above are not valid during times when rainfall or irrigation occurs.

The accuracy of the method depends largely on the errors related to the installation and calibration of the soil moisture equipment and the tensiometers (*see* **Chapter 73, Soil Water Potential Measurement, Volume 2**). Locating the ZFP can be a significant problem. However, measurement of the $k(\theta)$ curve also poses challenges.

MICROMETEOROLOGICAL METHODS

Two types of micrometeorological methods can be defined: direct and indirect methods. The direct method, the Eddy Covariance (EC) method, is based on turbulence theory. The indirect methods include the Flux-Profile method and Bowen ratio energy balance (BREB) method, both of which infer the fluxes from the relationship between gradients of scalar concentration and the vertical scalar flux. The Variance method and Scintillometry use experimental techniques, as explained below, to estimate sensible heat-flux, H. To obtain latent heat-flux (i.e. evaporation), λE, over the scintillometer or Variance method footprint, requires back-calculation from equation 2 using additional estimates of equivalent area-average Q^* and G.

First, a brief introduction on a few important micrometeorological concepts is given. For further reading, see Brutsaert (1982), Stull (1988), Monteith and Unsworth (1990), and Garratt (1992). As the procedures and equations related to determination of H and momentum flux, τ, are very much intertwined with those related to λE, these fluxes will be addressed as well, as this will facilitate the explanation.

The vertical exchange of atmospheric fluxes at the Earth's surface is a result of turbulence, because of forced and free convection. Forced convection is a consequence of airflow experiencing friction with the Earth's surface; its strength is given by a velocity scale called the *friction velocity*, u_* ($\mathrm{m\,s^{-1}}$), which is related to τ by: $u_*^2 = |\tau|/\rho_a$, where ρ_a is the density of air ($\mathrm{kg\,m^{-3}}$). Free convection gives rise to vertical motions in the atmosphere, as a result of differences in air density caused by solar heating of the ground. The relative influence of these two processes will determine the stability of the atmosphere (neutral, stable, or unstable), that is, its propensity to mix atmospheric constituents. This may be expressed by a stability parameter, for example, the ratio of the measurement height, $z(m)$ to the Obukhov length, $L(m)$, see equation 7.

All micrometeorological techniques require extensive homogeneous upwind areas and steady state conditions. The upwind distance from the mast to the nearest change in underlying surface type or roughness is referred to as the available fetch (Gash, 1986). The flux recorded by micrometeorological methods is an integration of sources and sinks of a scalar, for example, water vapor, within a shifting, dynamic source area, upwind of the sensor. The flux source area varies in magnitude with surface roughness, atmospheric stability, and sensor height, and in location with wind direction. Models are available to estimate the upwind fetch and crosswind dimensions of the flux source area (see Schmid, 2002, for a review). Typically, fetches measure a few hundred to a few thousand meters upwind, and a few tens of meters crosswind of a measuring point.

It is important to choose the sensor height such that the flux source area remains within the available fetch, but to also ensure that the sensors are within the inertial sub layer, typically taken to begin at twice the canopy height (see Brutsaert, 1982). The exception is the BREB method which may be applied below the inertial sub layer, as it is independent of turbulence (Moncrieff *et al.*, 1997b).

Eddy Covariance (EC) Method

The random nature of turbulence has resulted in meteorologists using statistical, rather than deterministic, approaches. By application of Reynolds averaging an atmospheric variable A may be separated into its turbulent part, A', and its nonturbulent, average part, $\overline{A}$, ($A = \overline{A} + A'$) (see Stull, 1988, for details of Reynolds averaging rules). Furthermore, the covariance between two variables is defined as $(1/N)\sum_{i=0}^{N-1}(A_i - \overline{A})(B_i - \overline{B})$ where A_i and B_i are instantaneous values of A and B, and N is the number of observations within the period of interest. Using Reynolds averaging, this can also be written as $(1/N)\sum_{i=0}^{N-1} A_i' B_i'$ and hence as $\overline{A'B'}$. For the surface fluxes λE and H, we are interested in the covariance between the vertical wind speed, $w(\mathrm{m\,s^{-1}})$ and the specific humidity, $q(\mathrm{kg_{water}\,kg_{air}^{-1}})$, on the one hand, and w and air temperature, T on the other.

These are also called the *eddy fluxes* and with appropriate multiplication factors ($\rho_a\lambda$ and $\rho_a c_p$) we obtain:

$$\lambda E = \rho_a \lambda \overline{w'q'} + C \quad H = \rho_a c_p \overline{w'T'} + C \qquad (4)$$

Here, c_p is the specific heat capacity of air ($\mathrm{J\,kg^{-1}\,K^{-1}}$). C is a correction term as discussed later. The above explains why this method is called the Eddy Covariance method. It is also often referred to as the Eddy Correlation method, as the covariance normalised by the product of the standard deviations of the two variables (e.g. σ_w and σ_T) equals a correlation coefficient. Generally, the term Eddy Covariance is now preferred over Eddy Correlation, as the first more accurately captures the essence of the method.

By obtaining fast-response measurements of state variables such as w and q or T, we can generate time series for which we can find the perturbation values of the data points (i.e. the fluctuations around the mean, as indicated by the primes). For example, if we subtract the mean air temperature, as averaged over, for example, 30 minutes, from each data point (within that half-hour period) to yield a time series of T' and similarly of w', we can find a time series of $w'T'$ by multiplying the two. The average of this series, $\overline{w'T'}$, is the vertical turbulent (eddy) heat flux (see Stull, 1988).

The EC method therefore requires high-frequency measurements of the three components of wind speed (u, v, and w) and scalars relevant to the fluxes of interest (e.g. temperature, water vapor or carbon dioxide concentration). Fast-response sensors, capable of measuring at 10–20 Hz, are essential to detect the flux carried by high frequency eddies. Typically, a three-axis sonic anemometer and an adjacent Infra Red Gas Analyzer (IRGA) are used to provide the required data for H, λE and carbon dioxide flux estimates (Moncrieff *et al.*, 1997a) (see Figure 2).

Three-axis sonic anemometers measure u, v, and w, and the speed of sound. The "sonic temperature" is derived from variations in the speed of sound, and has been found to be similar to the virtual temperature (Kaimal and Gaynor, 1991). Typically, the sonic temperature (with some corrections) is used to calculate the sensible heat-flux.

IRGAs calculate the concentration of H_2O and CO_2 in the optical path from the difference due to absorption of infrared (*IR*) radiation between the source and the detector (alternatively, fast-response humidity meters, Lyman-α or Krypton hygrometers, can be used to measure the H_2O concentration). IRGAs are available in either a closed path (e.g. Licor 6262) or an open path (e.g. Licor 7500) set up. Open path IRGAs have fewer associated errors, although comparisons have shown that the results can be similar after appropriate corrections are applied (Leuning and Judd, 1996). They do not work in rain, however, and the aerodynamic effects of their presence close to the sensor volume of the sonic anemometer needs to be carefully scrutinized.

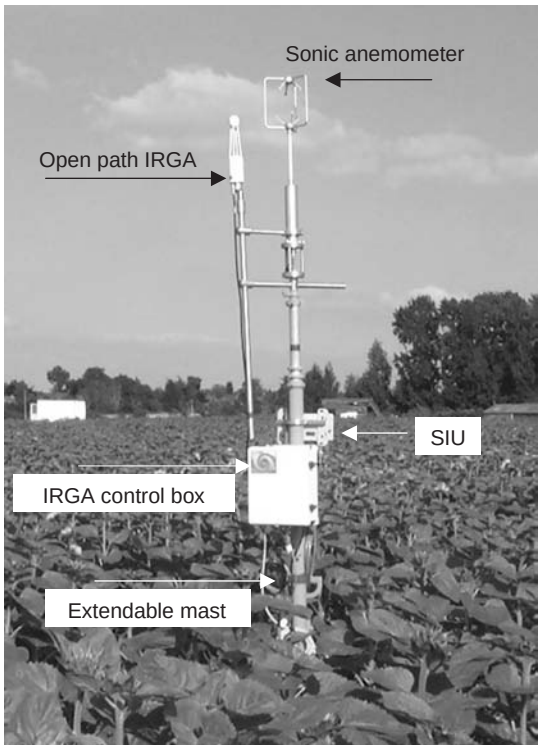


Figure 2 This photograph shows an example of an Eddy Covariance system in use in the field. In this example, the open path Licor 7500 Infra Red Gas Analyzer, and the Gill Solent R3 sonic anemometer were used. With this equipment, the sonic anemometer and IRGA sensor cables are connected to the Gill Sensor Input Unit (SIU) to transmit data to the logging PC. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Acquisition and postprocessing software are required to obtain and record the raw data (about 20 MB a day), to generate the covariances and to apply the corrections. Two Windows programs, *Edisol*, acquisition software, and *Edire*, for postprocessing, were developed at the University of Edinburgh, and may be downloaded from the Web at **`http://www.geos.ed.ac.uk/abs/research/micromet/EdiRe`**.

Eddy Covariance Corrections

The correction terms in equation 4 are due to both physical effects and system artifacts. The physical corrections account for the use of sonic virtual temperature instead of air temperature in calculation of H (Kaimal and Gaynor, 1991; Liu, 2001), and the fact that $\overline{w\rho_a} \neq 0$ due to atmospheric buoyancy (Webb *et al.*, 1980; Paw *et al.*, 2000). The system corrections remove any time lag between instruments, apply a coordinate rotation of the wind velocities (McMillen, 1988; Wilczak *et al.*, 2001; Finnigan *et al.*, 2003; Paw *et al.*, 2000; Finnigan *et al.*, 2003), and correct for frequency loss due to instrumentation effects (Moore, 1986; Laubach and McNaughton, 1998; Massman, 2000).

Standard correction procedures can be found in Moncrieff *et al.* (1997a) and Aubinet *et al.* (2000), and within the recent review by Massman and Lee (2002).

Finally, cosine errors in the sonic anemometer due to flow sheltering by the instrument frame may lead to significant underestimation of fluxes (Gash and Dolman, 2003). Van der Molen *et al.* (2004) have proposed a polynomial correction for Solent sonic anemometers.

Flux-Profile Method

The Flux-Profile method is based on K-theory (see Stull, 1988), which assumes that vertical transport of quantities (such as water vapor) is proportional to the concentration gradient of the quantity. The constant of proportionality is known as the *eddy diffusivity* (much larger than the molecular diffusivity as found in, for example, a laminar sublayer), and given the symbol K.

$$\lambda E = -\lambda \rho_a K_e \frac{\partial \overline{q}}{\partial z}; \quad H = -\rho_a c_p K_h \frac{\partial \overline{T}}{\partial z}; \quad \tau = \rho_a K_m \frac{\partial \overline{u}}{\partial z} \tag{5}$$

where K_x (with $x = e, h$, and m for λE, H, and momentum, respectively) is the turbulent exchange coefficient or eddy diffusivity (with units of $m^2\,s^{-1}$). It is assumed that the turbulent Ks are related to wind shear (through friction velocity, u_*) and buoyancy, via a stability function ϕ. Using dimensional analysis and similarity theory (see Stull, 1988), we find for K:

$$K_x = \frac{\kappa u_* (z - d)}{\phi_x} \tag{6}$$

where κ is the von Kármán constant ($= 0.4$) and $d(m)$ the zero-plane displacement height, a parameter that is introduced to correct the shape of the logarithmic wind profile close to the surface (Tennekes, 1973). ϕ_x is a function of the stability term, the Obukhov length, L:

$$L = \frac{-\overline{T}}{g\kappa} \frac{u_*^3}{\overline{w'T'}(1 + 0.61\overline{w'q'})} \tag{7}$$

with g ($m\,s^{-2}$) being the acceleration due to gravity. Several functions for ϕ_x are in use; however, the most commonly used are those of Dyer (1974) and Paulson (1970).

In practice, the derivatives in equation 5 are difficult to measure. Therefore, discrete differences are used, for example, Δq, ΔT, and Δu. One possibility is to employ a profile above the crop, of at least three levels of colocated, continuous T, q, and u measurements, and to calculate $u_*(= \kappa z \phi_m \Delta\overline{u}/\Delta z)$, H and λE simultaneously. In this case, the set of equations given above are applied in an iterative procedure, as L contains the unknown quantities u_*, $\overline{w'T'}(= H/\rho_a c_p)$ and $\overline{w'q'}(= \lambda E/\rho_a \lambda)$.

Alternatively, the Flux-Profile equations can be written as (see Thom, 1975), in analogous form to Ohm's law:

$$\lambda E = -\rho_a \lambda \frac{\overline{q}(z_2) - \overline{q}(z_1)}{r_e}, \quad H = -\rho_a c_p \frac{\overline{T}(z_2) - \overline{T}(z_1)}{r_h} \tag{8}$$

that is, the flux is proportional to the driving gradient and inversely proportional to the resistance to transport in the atmosphere ($r_x = \int_{z_1}^{z_2} (\,dz/K_x)$, an aerodynamic resistance). Equation 8 is the resistance analog of equation 5 and it indicates that K_x, and thus ϕ_x, has to be integrated from z_1 to z_2. The integrated stability functions, usually denoted with Φ_x, are given by Paulson (1970).

In practice, z_2 is taken as the reference level (often taken as at least two times the canopy height), whereas z_1 refers to the surface (i.e. $z = 0$). This leads to the bulk transfer equations for latent heat and sensible heat transfer, often used for remote sensing applications as it is relatively easy to determine T_s remotely (see Kustas *et al.*, 1989):

$$\lambda E = \rho_a \lambda \frac{q_s - q_r}{r_e}, \quad H = \rho_a c_p \frac{T_s - T_r}{r_h} \tag{9}$$

with the subscripts s and r referring to the value of the atmospheric variable at surface and reference level, z_r, respectively. For λE, a problem is the determination of q_s, the surface humidity, that is, the air's specific humidity at the humidity roughness height, $z_{0,q}$. For unsaturated or drying soil surfaces, and for the canopy surface in particular, q_s is not readily determined (see Garratt, 1992). To remedy this problem, the Penman–Monteith equation has been developed (Monteith, 1965); evaporation is calculated using single-level environmental variables (radiation, vapor-pressure deficit, temperature, and wind speed), and estimates of aerodynamic and surface resistances. For a detailed discussion of this method, (*see* **Chapter 45, Actual Evaporation, Volume 1**).

To describe the aerodynamic resistances r_e and r_h required in equation 9, we use the theoretical vertical profiles of u, T, and q in the atmospheric surface layer. We denote the heights, above d, where the downward extrapolated profiles yield the surface values by $z_{0,m}$, $z_{0,T}$ and $z_{0,q}$ (and usually assume $z_{0,T} = z_{0,q}$). These are known as the surface roughness lengths for momentum, heat and evaporation, respectively. The logarithmic wind profile (Tennekes, 1973) is given by

$$u_* = \kappa u \left[\ln\left(\frac{z_r - d}{z_{0,m}}\right) - \Phi_m\left(\frac{z_r - d}{L}\right) \right]^{-1} \tag{10}$$

The aerodynamic resistance between $(z_{0,m} + d)$ and z_r is then calculated as

$$r_{ax} = \frac{1}{\kappa u_*} \left[\ln\left(\frac{z_r - d}{z_{0,m}}\right) - \Phi_x\left(\frac{z_r - d}{L}\right) \right] \tag{11}$$

However, the total scalar resistance, r_x (with $x = h$ or e) is composed of $r_x = r_{ax} + r_r$, where r_r is often called the *excess resistance*, that is, the resistance for an air layer between $(z_{0,T/q} + d)$ and $(z_{0,m} + d)$. In practice (for remote sensing purposes, for example), the value of r_r is not known and instead relatively simple formulae are used that employ readily available parameters to derive r_r. To achieve this, r_r is related to the dimensionless quantity $\kappa B^{-1} : \kappa B^{-1} = \ln(z_{0,m}/z_{0,T}) = \kappa r_r u_*$ and Brutsaert (1982) and, Verhoef *et al.* (1997a) for example, describe several formulae for the estimation of κB^{-1}. Often, for simplicity, the following very general rules are used: $\kappa B^{-1} = 2.0$ for closed crops, 0 for bare soils and κB^{-1} has values between about 5 and 15 for sparse canopies. Momentum roughness parameters $z_{0,m}$ and d can be found from profile measurements of u (Robinson, 1962) or from empirical relations related to canopy geometry (e.g. Raupach, 1992, Verhoef *et al.*, 1997b).

Bowen Ratio Energy Balance Method (BREB)

The BREB method assumes that $\Phi_h = \Phi_e$ and therefore $r_h = r_e$, eliminating the effect of atmospheric stability and turbulence. This method combines measurements of air temperature and humidity profiles with the SEB. It employs the Bowen Ratio, β, which is defined by $H/\lambda E$. β is derived from air temperature and vapor pressure (e in mbar) differences between the measurement levels:

$$\beta = \gamma \frac{\Delta \overline{T}}{\Delta \overline{e}} \tag{12}$$

where γ is the psychrometric constant in mbar K^{-1} (see Perez *et al.*, 1999). Combining the SEB and β gives

$$\lambda E = \frac{Q^* - G}{1 + \beta} \quad H = \beta \frac{Q^* - G}{1 + \beta} \tag{13}$$

To use the BREB technique, two atmospheric humidity and temperature sensors, a net radiometer and an estimate of soil heat-flux are required.

The accuracy of the BREB method depends on the ability of the instrumentation to measure small gradients of temperature and humidity (and this can be very difficult over rough surfaces such as forests), plus good measurements of Q^* and G. The soil heat-flux can be estimated with a variety of methods, involving soil heat-flux plates or a profile of soil temperature and θ (see Smith and Mullins, 1991). Even more important are reliable estimates of Q^*; relatively large differences can exist between simultaneous recordings with different sensors as a result of poor levelling, lack of recalibration, wind and water drop effects (see Field *et al.*, 1992). At night $(Q^* - G)$ is small, and β can be difficult to determine. The BREB method has mainly been used to measure total λE, although Wallace and Holwill (1997) used it to measure soil λE between strips of sparse vegetation.

Variance Method

Both the Variance method and the scintillometry method are based on Monin–Obukhov similarity theory. Similarity theory is a technique often used by meteorologists when the knowledge of the governing physics is insufficient to derive laws based on first principles. It is based on the organization of meteorological variables into dimensionless groups. A proper choice of groups will allow empirical relationships between these groups that are universally applicable. Monin–Obukhov similarity theory is usually applied to the surface layer and uses variables or groups such as the friction velocity, u_*, and the Obukhov length, L (see Stull, 1988).

Monin–Obukhov similarity theory predicts a universal relationship between the variance of temperature, humidity, and wind speed (σ_T, σ_q, and σ_u), and the dimensionless stability parameter $(z - d)/L$ (Panofsky and Dutton, 1984). Relationships describing sensible and latent heat-flux transfer can be derived using this theory, such that (see de Bruin *et al.*, 1993):

$$H = \rho_a c_p \left[\left(\frac{\sigma_T}{c_{T1}} \right)^3 \left(\frac{\kappa g (z-d)}{T} \right) \left(\frac{1 - c_{T2}(z-d)/L}{-(z-d)/L} \right) \right]^{1/2} \tag{14}$$

$$\lambda E = \lambda \rho_a \left(\frac{\sigma_q}{c_{T1}^{3/2}} \right) \left[\sigma_T \left(\frac{\kappa g (z-d)}{T} \right) \left(\frac{1 - c_{T2}(z-d)/L}{-(z-d)/L} \right) \right]^{1/2} \tag{15}$$

where c_{T1} and c_{T2} are constants. In order to estimate L, an additional profile of anemometers is required to calculate u_* (see Monteith and Unsworth, 1990). Measurement of σ_T can be made with fast-response thermocouples, whereas σ_q can be obtained from an IRGA or a Lyman-α hygrometer, raising the cost significantly. de Bruin *et al.* (1993) showed that estimates of λE using σ_q did not correspond well with EC data, whereas estimates of H using the Variance method did compare well with the EC method (see also Lloyd *et al.*, 1991). Therefore, most researchers would use σ_T to find H and then calculate λE using equation 2 with area-average estimates of Q^* and G.

Scintillometry Method

Monin–Obukhov similarity theory also predicts that the temperature structure parameter C_T^2, which scales directly with H, can be defined as a unique function of $(z - d)/L$:

$$\frac{C_T^2 (z-d)^{2/3}}{(\overline{w'T'}/u_*)^2} = C_{TT1} \left(1 - C_{TT2} \left(\frac{z-d}{L} \right) \right)^{-2/3} \tag{16}$$

with $c_{TT1} = 4.9$ and $c_{TT2} = 9$ (de Bruin *et al.*, 1993).

Once H has been determined, as with the Variance method, λE can be found from $Q^* - H - G$.

C_T^2 can be measured using scintillometry. Temperature fluctuations cause fluctuations of the refractive index of air. By measurement of the fluctuation of the light intensity of a beam transmitted over a horizontal path with known length, this refractive index can be determined. In general, both temperature and humidity fluctuations will cause fluctuations of the refractive index. For operations in the visible or near-infrared range, and at large Bowen ratios, this humidity prediction can be neglected. The light intensity fluctuations are then directly proportional to C_T^2 (Kohsiek, 1982; Green *et al.*, 1994; de bruin *et al.*, 1995).

PLANT-PHYSIOLOGICAL METHODS: MEASUREMENT OF TRANSPIRATION

The main factors determining transpiration are described in **Chapter 42, Transpiration, Volume 1**. Plant physiological methods used for measurement of transpiration comprise sap flow measurements (Allen and Grime, 1995; Smith and Allen, 1996), plant chambers (Goulden and Field, 1994), deuterium tracing (Calder *et al.*, 1992), or a combination of porometry (see Jones, 1992) and calculations (Roberts *et al.*, 1990; Dugas *et al.*, 1993). In some cases, destructive methods have been used to directly obtain transpiration (Calder *et al.*, 1992). The sap flow method has become the standard in estimating whole plant transpiration. This technique is usually automated, so continuous records of plant water use with high time resolution can be obtained. In contrast, the other methods are labor intensive, provide a considerably poorer time resolution, influence the microclimate of the plant or leaf, or are often difficult to interpret (see Smith and Allen, 1996). Therefore, only the sap flow method will be described in detail below.

Sap Flow

The sap flow method measures the rate at which sap ascends stems, by using heat as a tracer for sap movement. These methods can use the heat balance principle (stem heat balance and trunk sector heat balance method), which involves applying continuous heating, or the heat-pulse technique where short pulses of heat are applied. Also, rates of sap flow can be determined empirically, using the thermal dissipation technique, from the temperature of sapwood near a continuously-powered heater implanted in the stem (Granier, 1985).

The stem heat balance method can be used to measure sap flow in both woody (Steinberg *et al.*, 1989) and herbaceous stems, with stem diameters ranging between 2 and 125 mm. The other methods can only be used on woody trunks with diameters larger than 120 mm (trunk sector heat balance), >30 mm (heat-pulse method), or >40 mm (thermal dissipation technique). Such gauges are commercially available (see, Smith and Allen, 1996 for more details) and can be connected to dedicated loggers.

In the stem heat balance method, the gauge comprises a flexible heater, typically a few centimeters in width, which is wrapped around the stem, insulated and weather-shielded. Thermopiles are used to measure the radial temperature gradient away from the heater (ΔT_r), and a set of thermocouples is used to measure temperature gradients ΔT_a and ΔT_b (see Figure 3). These temperature gradients are required to calculate components in the heat balance of the stem (see equation 17). Heat is applied to the entire circumference of the stem encircled by the heater, and the mass flow of sap is obtained from the balance of the fluxes of heat into and out of the stem (see also Figure 3, reprinted from Smith and Allen, 1996), as summarized by:

$$P = q_v + q_r + q_f \tag{17}$$

where P is the electrical power supplied to the heater, q_v is the rate of vertical heat loss by conduction, q_r is the radial heat loss by conduction and q_f is the heat uptake by the moving sap stream. If P, q_v, and q_r are known, q_f can be determined by difference and the mass flow rate of sap (F_m) calculated using

$$F_m = \frac{2q_f}{c_s(\Delta T_a + \Delta T_b)} \tag{18}$$

where c_s is the specific heat capacity of sap, whereas q_v is found from

$$q_v = A_{\mathrm{st}} k_{\mathrm{st}} \left(\frac{\Delta T_b - \Delta T_a}{x} \right) \tag{19}$$

where A_{st} is the cross-sectional area of the heated stem section, k_{st} is the thermal conductivity of the stem and x is the distance between the two thermocouple junctions on each side of the heater. Standard values of k_{st} are $0.42\,\mathrm{Wm^{-1}\,K^{-1}}$ for woody stems and 0.54 for herbaceous stems (Smith and Allen, 1996).

The radial component of the stem heat balance, q_r, is determined from ΔT_r using

$$q_r = K_{\mathrm{sh}} \Delta T_r \tag{20}$$

K_{sh} is the effective thermal conductance of the sheath of materials surrounding the heater.

Gauges should be installed on straight sections of stem without swellings or lumps since they could cause poor contact between the stem surface and the heater or thermocouples. Also, it is important that water ingress between the probe and the stem is prevented. Accurate determination of sap flow rate using this technique depends critically on the correct value of K_{sh}, which can be obtained when sap

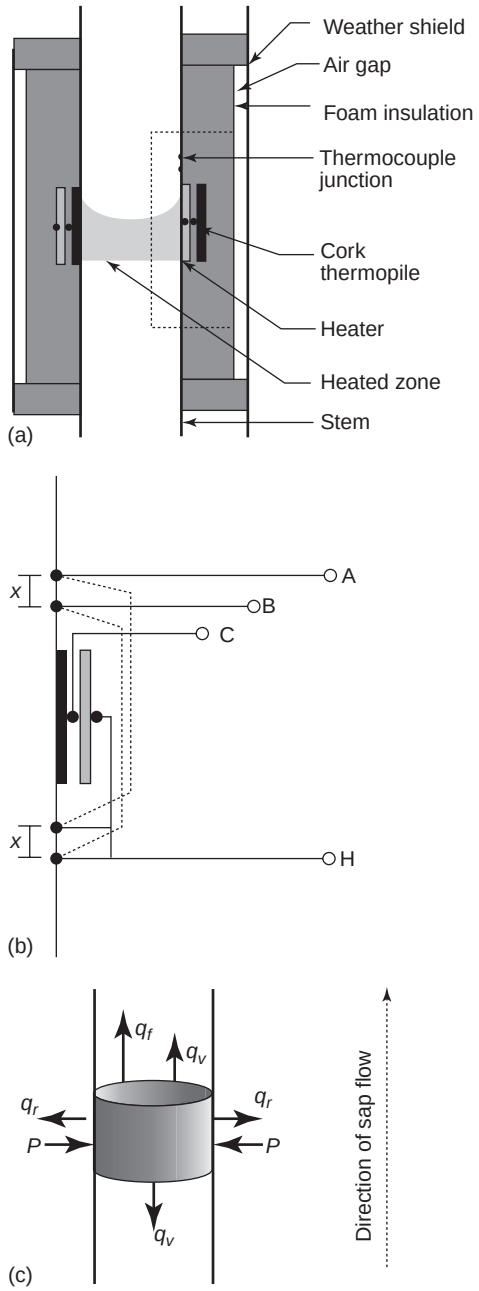


Figure 3 (a) Vertical cross-section through a stem heat balance gauge. The dashed box in (a) is expanded in (b) to show wiring details of thermocouples in the gauge; copper wires are shown as solid lines and constantan wires as dotted lines. For the determination of sap flow, the temperature differentials ΔT_a, ΔT_b, and ΔT_r are obtained from measurements of the voltages across AH, BH, and CH, respectively. (c) The heat balance of the length of stem heated by the gauge, where P is heat-applied to the stem, q_v is the rate of vertical heat loss by conduction, q_r is radial heat loss by conduction and q_f is heat uptake by the sap stream. (Smith DM and Allen SJ, Measurement of sap flow in plant stems. *Journal of Experimental Botany* (1996), **47**(305): 1833–1844, by permission of Oxford University Press)

flow is zero, for example, using data from the hours before dawn (Steinberg *et al.*, 1989), after heavy rainfall (Allen and Grime, 1995), or by wrapping all foliage in plastic to prevent transpiration (Steinberg *et al.*, 1989).

In the heat-pulse method, the mass flow of sap is determined from the velocity of the short heat pulses, v_h, moving along the stem. Each set of heat-pulse probes consists of one heater probe and two sensor probes containing thermistors. Usually, four sets of heat-pulse probes are used, one in each quadrant of the stem. Heater and sensor probes (about 2 mm in diameter) must be installed in parallel holes drilled radially into the sapwood, with one sensor probe placed at a distance of x_d (~5 mm) below the heater and the other probe placed at x_u (~10 mm) above the heater. Short (1–2 s) pulses of heat are periodically released from the heater probe and the sensor probes are monitored continuously to measure the velocity of each pulse as it moves with the sap stream. The volumetric sap flux density per unit cross-sectional area of sapwood, u_v, is given by:

$$u_v = \frac{\rho_{\mathrm{sm}} c_{\mathrm{sm}}}{\rho_s c_s} v_h \quad \text{with} \quad v_h = \frac{(x_d - x_u)}{2t_{\mathrm{e}}} \tag{21}$$

where ρ and c are density and specific heat capacity, with the subscripts s and sm referring to sap and sap plus woody matrix, respectively. Furthermore, t_e is the time at which the temperatures of the upstream and downstream sensor probes are equal after the release of a heat pulse. Mass flow rates of sap through the stem (F_m) are then calculated from the integral of the sap flux-profile over the cross-sectional area of the sapwood. For more details on these methods, see Smith and Allen (1996).

Scaling up from Plant to Stand

Each gauge provides a measure of the transpiration of an individual plant or tree, but to get area-average estimates of transpiration it is necessary to extrapolate to the field scale. Various methods have been used (see Smith and Allen, 1996). Where members of a stand are of similar size, stand transpiration can be calculated from plant density; where size is variable within a stand, individual plants can be scaled up by determining relationships between sap flow rates and stem diameter, stem basal area, sapwood area, or leaf area, and then extrapolating transpiration rates to unit area of land on the basis of surveys of tree size or measurements of LAI.

LAI can be measured destructively, by removing the plant leaves and using either a leaf area meter, a planimeter, or a scanner to measure the area of each leaf. Alternatively, the gap fraction analysis method (see Welles and Cohen, 1996, for details of the method and comparison of available sensors), or digital photography (White *et al.*, 2000) can be used. Finally, it is possible to derive LAI from remote sensing, by use of the Lidar (Laser

altimetry ranging) technique (White *et al.*, 2000). Refer to section **Chapter 51, Spatially Resolved Measurements of Evapotranspiration by Lidar, Volume 2** for details of the use of Lidar to estimate vegetation structure. Further discussion of the statistical and modeling approaches to upscaling in general can be found in sections **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1** and **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1** respectively.

SUMMARY

We have described the principles of each of the commonly used evaporation measurement methods. For any given experiment, the most appropriate method(s) will be determined by the spatial and temporal scales of interest, accuracy, restrictions on applicability, labor intensity, equipment required, and the cost. The methods are summarized with respect to these criteria, below.

The Eddy Covariance, Variance, and Flux-Profile methods have similar footprint dimensions (Horst, 1999; Lloyd *et al.*, 1991), such that the effective fetch is typically of the order of 10's to 100's of meters upwind. Scintillometry may be used to supply H measurements at the regional scale, using path lengths of several kilometers (see de Bruin *et al.*, 1996). Lysimetry may provide λE averaged over an area from $0.002\,m^2$ to $10\,m^2$. The methods based on the water balance can also supply long-term, continuous estimates of evaporation (generally, total or soil evaporation) but they have problems with representativity as they supply small-scale estimates; maximum averaging areas of about $0.3\,m^2$. Sap flow gauges provide transpiration for an individual branch, tiller, or plant.

Half-hourly averaging periods are generally used for the described micrometeorological techniques, plus lysimetry, and the use of sap flow gauges. The BREB method may be applied with shorter averaging periods (Steduto and Hsiao, 1998). Microlysimetry often provides daily averages, whereas soil physical methods provide daily or weekly averages.

EC estimates of H and λE are usually validated by examining the closure of the SEB (equation 2). Typically, the sum of H and λE estimated by eddy covariance underestimates $Q^* - G$ (see Wilson *et al.*, 2002, for a review). Fluxes measured by the EC method are considered to be accurate to 5–20% (Goulden *et al.*, 1996). Micrometeorological methods are typically compared with EC data as proof of accuracy: Scintillometry (Beyrich *et al.*, 2002) and the variance method (Lloyd *et al.*, 1991) all compared reasonably well with the EC method. Transpiration obtained using the sap flow technique appears very accurate and it can be scaled up to provide field-scale estimates of λE, which has been shown to compare well with EC estimates (see Allen and Grime, 1995).

Eddy covariance measurements of some trace gases can become unreliable, particularly at night when wind speeds are low and katabatic drainage is suspected. The Flux-Profile method collapses when gradients become small, for example, close to rough canopies. Furthermore, errors of up to 30% may occur, due to uncertainties in the stability corrections and in the equivalence of the exchange coefficients (Steduto and Hsiao, 1998). The BREB method fails when sensible and latent heat fluxes are equal and opposite (e.g. at sunset and sunrise), during irrigation or precipitation and when the available energy is low (see Perez *et al.*, (1999), for BREB quality control criteria and model). The variance method breaks down during stable conditions when H becomes negative. Sources of error in scintillometry include obstacles within the optical path, and misalignment of the transmitter and receiver. Soil physical methods fail during periods of rainfall or irrigation.

All micrometeorological methods require a medium amount of labor and user intervention. When using the EC method, regular field calibration of the IRGA is recommended. The corrections applied to the covariances must be understood, and applied correctly. Lysimetry requires little manual intervention after installation, except microlysimeters, which require manual weighing. The soil physical methods require little labor, with the exception of the use of the neutron probe or γ-ray scanner for soil moisture measurements. Sap flow gauges may be fully automated, and require relatively little maintenance.

For Eddy Covariance, the total equipment cost is quite substantially larger than that of any other method described. The Flux-Profile method equipment cost is that of the three sets of temperature, wind speed, and humidity instruments, plus data logging. For the Variance method, the equipment required is one fast-response thermocouple and a profile of anemometers, and as such, is inexpensive. The BREB method requires a minimum of two psychrometers, two anemometers, a net radiometer and soil heat-flux plates. A medium-to-large lysimeter is quite costly, but smaller varieties are not. The soil physical methods are also economical.

Acknowledgments

We thank Dr. Jesús Fernández Gálvez for his assistance with Section 2.2 and Dr. Antonio Díaz Espejo for his literature search on lysimetry.

REFERENCES

Allen S.J. (1990) Measurement and estimation of evaporation from soil under sparse barley crops in northern Syria. *Agricultural and Forest Meteorology*, **49**, 291–309.

Allen R.G. and Fisher D.K. (1990) Low-cost electronic weighing lysimeters. *American Society of Agricultural Engineers*, **33**(6), 1823–1833.

Allen S.J. and Grime V.L. (1995) Measurements of transpiration from savannah shrubs using sapflow gauges. *Agricultural and Forest Meteorology*, **75**, 23–41.

Allen R.G., Pereira L.S., Raes D. and Smith M. (1998) *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements – FAO Irrigation and Drainage Paper 56*, Food and Agriculture Organization: Rome.

Aubinet M., Grelle A., Ibrom A., Rannik U., Moncrieff J., Foken T., Kowalski A.S., Martin P.H., Berbigier P., Bernhofer Ch., *et al.* (2000) Estimates of annual net carbon and water exchange of forests: the EUROFLUX methodology. *Advances in Ecological Research*, **30**, 13–175.

Beyrich F., Richter S.H., Weisensee U., Kohsiek W., Lohse H., de Bruin H.A.R., Foken Th., Gockede M., Berger F., Vogt R., *et al.* (2002) Experimental determination of turbulent fluxes over the heterogeneous LITFASS area: selected results from the LITFASS-98 experiment. *Theoretical and Applied Climatology*, **73**, 19–34.

Boast C.W. and Robertson T.M. (1982) A "micro-lysimeter" method for determining evaporation from bare soil: description and laboratory evaluation. *Soil Science Society of America Journal*, **46**, 689–696.

Brutsaert W.H. (1982) *Evaporation into the Atmosphere*, D. Reidel Publishing Company: Dordrecht.

Calder I.R., Hall R.L. and Adlard P.G. (Eds.) (1992) *Growth and Water Use of Forest Plantations: Proceedings of the International Symposium*, Bangalore, 4–7 February, 1991.

Daamen C.C., Simmonds L.P., Wallace J.S., Laryea K.B. and Sivakumar M.V.K. (1993) Use of microlysimeters to measure evaporation from sandy soils. *Agricultural and Forest Meteorology*, **65**, 159–173.

de Bruin H.A.R., Kohsiek W. and van den Hurk B.J.J.M. (1993) A verification of some methods to determine the fluxes of momentum, sensible heat and water vapour using standard deviation and structure parameters of scalar meteorological quantities. *Boundary-Layer Meteorology*, **63**, 231–257.

de Bruin H.A.R., Kohsiek W. and van den Hurk B.J.J.M. (1995) The scintillation method tested over a dry vineyard area. *Boundary-Layer Meteorology*, **76**, 25–40.

de Bruin H.A.R., Nieveen J.P., de Wekker S.F.J. and Heusinkveld B.G. (1996) Large aperture scintillometry over a 4.8 km path for measuring areally-average sensible heat flux; a case study. *Proceedings of the 22nd Conference on Agricultural and Forest Meteorology and 12th Conference on Biometeorology & Aerobiology*, 28 January-2 February, 1996; Atlanta, GA, American Mathematical Society, Boston.

Delta-T devices (1999) *User Manual for the Thetaprobe Soil Moisture Sensor*, Version ML2x-UM-1.21: Cambridge, p. 22.

Delta-T devices (2001) *User Manual for the Profile Probe*, Version PR1-UM-01-2: Cambridge, p. 36.

Dugas W.A. and Bland W.L. (1989) The accuracy of evaporation measurements from small lysimeters. *Agricultural and Forest Meteorology*, **46**, 119–129.

Dugas W.A., Wallace J.S., Allen S.J. and Roberts J.M. (1993) Heat balance, porometer, and deuterium estimates of transpiration from potted trees. *Agricultural and Forest Meteorology*, **64**, 47–62.

Dyer A.J. (1974) A review of flux-profile relationships. *Boundary-Layer Meteorology*, **7**, 363–372.

Field R.T., Fritschen L.J., Kanemasu E.T., Smith E.A., Stewart J.B., Verma S.B. and Kustas W.P. (1992) Calibration, comparison, and correction of net radiation instruments used during FIFE. *Journal of Geophysical Research*, **97**, No.D17, 18681–18695.

Finnigan J.J., Clement R., Malhi Y., Leuning R. and Cleugh H.A. (2003) A re-evaluation of long-term flux measurement techniques – part I. Averaging and coordinate rotation. *Boundary-Layer Meteorology*, **107**, 1–48.

Garratt J.R. (1992) *The Atmospheric Boundary Layer*, Cambridge atmospheric and space science series, Cambridge University Press.

Gash J.H.C. (1986) A note on estimating the effect of a limited fetch on micrometeorological evaporation measurements. *Boundary-Layer Meteorology*, **35**, 409–413.

Gash J.H.C. and Dolman A.J. (2003) Sonic anemometer (co)sine response and flux measurement: I. The potential for cosine error to affect flux measurements. *Agricultural and Forest Meteorology*, **119**, 195–207.

Granier A. (1985) Une nouvelle méthode pour la mesure du flux de sève brute dans le tronc des arbres. *Annales des Sciences Forestières*, **42**, 193–200.

Green A.E., McAneney K.J. and Astill M.S. (1994) Surface-layer scintillation measurements of daytime sensible heat and momentum fluxes. *Boundary-Layer Meteorology*, **68**, 357–373.

Goulden M.L. and Field C.B. (1994) Three methods for monitoring the gas exchange of individual tree canopies: ventilated chamber, sap-flow and Penman-Monteith measurements on evergreen oaks. *Functional Ecology*, **8**, 125–135.

Goulden M.L., Munger J.W., Fan S.-M, Daube B.C. and Wofst S.C. (1996) Measurements of carbon sequestration by long-term eddy covariance: methods and a critical evaluation of accuracy. *Global Change Biology*, **2**, 169–182.

Hillel D. (1998) *Environmental Soil Physics*, Academic press: San Diego.

Horst T.W. (1999) The footprint for estimation of atmosphere-surface exchange fluxes by profile techniques. *Boundary-Layer Meteorology*, **90**, 171–188.

Howell T.A., McCormick R.L. and Phene C.J. (1985) Design and installation of large weighing lysimeters. *Transactions of the American Society of Agricultural Engineers*, **28**(1), 106–112.

Jones, H.G. (1992) *Plants and Microclimate: a Quantitative Approach to Environmental Plant Physiology*, *Second Edition*, Cambridge University Press: Cambridge, p. 428.

Kaimal J.C. and Gaynor J.E. (1991) Another look at sonic thermometry. *Boundary-Layer Meteorology*, **56**, 401–410.

Klay M.C. and Vachaud G. (1992) Seasonal water balance of a sandy soil in Niger cropped with pearl millet, based on profile moisture measurements. *Agricultural Water Management*, **21**, 313–330.

Klocke N.L., Heermann D.F. and Duke H.R. (1985) Measurement of evaporation and transpiration with lysimeters. *Transactions of the American Society of Agricultural Engineers*, **83–2525**, 183–189.

Kohsiek W. (1982) Measuring C_T^2, C_Q^2 and C_{TQ} in the unstable surface layer, and relations to the vertical fluxes of heat and moisture. *Boundary-Layer Meteorology*, **24**, 89–107.

Kustas W.P., Choudhury B.J., Moran M.S., Reginato R.J., Jackson R.D., Gay L.W. and Weaver H.L. (1989) Determination

of sensible heat flux over sparse canopy using thermal infrared data. *Agricultural and Forest Meteorology*, **44**, 197–216.

Laubach J. and McNaughton K.G. (1998) A spectrum-independent procedure for correcting eddy fluxes measured with separated sensors. *Boundary-Layer Meteorology*, **89**, 445–467.

Leuning R. and Judd M.J. (1996) The relative merits of open- and closed-path analysers for measurement of eddy fluxes. *Global Change Biology*, **2**, 241–253.

Liu H., Peters G. and Foken T. (2001) New equations for sonic temperature variances and buoyancy heat flux with an omnidirectional sonic anemometer. *Boundary-Layer Meteorology*, **100**, 459–468.

Lloyd C.R., Culf A.D., Dolman A.J. and Gash J.H.C. (1991) Estimates of sensible heat-flux from observations of temperature – fluctuations. *Boundary-Layer Meteorology*, **57**, 311–322.

Massman W.J. (2000) A simple method for estimating frequency response corrections for eddy covariance systems. *Agricultural and Forest Meteorology*, **104**, 185–198.

Massman W.J. and Lee X. (2002) Eddy covariance flux corrections and uncertainties in long-term studies of carbon and energy exchanges. *Agricultural and Forest Meteorology*, **113**, 121–144.

McGowan M. and Williams J.B. (1980) The water balance of an agricultural catchment: I estimation of evaporation from soil water records. *Journal of Soil Science*, **31**, 217–230.

McMillen R.T. (1988) An eddy correlation technique with extended applicability to non-simple terrain. *Boundary-Layer Meteorology*, **43**, 231–245.

Moncrieff J.B., Massheder J.M., de Bruin H., Elbers J., Friborg T., Heusinkveld B., Kabat P., Scott S., Soegaard H. and Verhoef A. (1997a) A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide. *Journal of Hydrology*, **188–189**, 589–611.

Moncrieff J.B., Valentini R., Greco S., Seufert G. and Ciccioli P. (1997b) Trace gas exchange over terrestrial ecosystems: methods and perspectives in micrometeorology. *Journal of Experimental Botany*, **48**, 1133–1142.

Monteith J.L. (1965) Evaporation and the environment. *Symposia of the Society for Experimental Biology*, **19**, 205–234.

Monteith J.L. and Unsworth M.H. (1990) *Principles of Environmental Physics*, *Second Edition*, Edward Arnold: London, p. 291.

Moore C.J. (1986) Frequency response corrections for eddy correlation systems. *Boundary-Layer Meteorology*, **37**, 17–35.

Panofsky H.A. and Dutton J.A. (1984) *Atmospheric Turbulence; Models and Methods for Engineering Applications*, Wiley & Sons: New York.

Paulson C. (1970) The mathematical representation of wind speed and temperature profiles in the unstable atmospheric surface layer. *Journal of Applied Meteorology*, **9**, 857–861.

Paw U.K.T., Baldocchi D.D., Meyers T.P. and Wilson K.B. (2000) Correction of eddy covariance measurements incorporating both advective effects and density fluxes. *Boundary-Layer Meteorology*, **97**, 487–511.

Perez P.J., Castellvi F., Ibanez M. and Rosell J.I. (1999) Assessment of reliability of bowen ratio method for partitioning fluxes. *Agricultural and Forest Meteorology*, **97**, 141–150.

Raupach M.R. (1992) Drag and drag partition on rough surfaces. *Boundary-Layer Meteorology*, **60**, 375–395.

Roberts J., Cabral O.M.R. and de Aguiar L.F. (1990) Stomatal and boundary-layer conductances in an amazonian terra firme rain forest. *Journal of Applied Ecology*, **27**, 336–353.

Robinson S.M. (1962) Computing wind profile parameters. *Journal of Atmospheric Science*, **19**, 189–190.

Roth K., Schulin R., Flühler H. and Attinger W. (1990) Calibration of time domain reflectometry for water content measurement using a composite dielectric approach. *Water Resources Research*, **26**, 2267–2273.

Rowell D.L. (1994) *Soil Science: Methods & Applications*, Pearson Education Limited: Harlow.

Schmid H.P. (2002) Footprint modelling for vegetation atmosphere exchange studies: a review and perspective. *Agricultural and Forest Meteorology*, **113**, 159–183.

Smith D.M. and Allen S.J. (1996) Measurement of sap flow in plant stems. *Journal of Experimental Botany*, **47**(305), 1833–1844.

Smith K.A. and Mullins C.E. (1991) *Soil Analysis: Physical Methods*, Marcel Dekker: New York.

Steduto P. and Hsiao T.C. (1998) Maize canopies under two soil water regimes – IV. Validity of Bowen ratio energy balance technique for measuring water vapor and carbon dioxide fluxes at 5-min intervals. *Agricultural and Forest Meteorology*, **89**, 215–228.

Steinberg S.L., van Bavel C.H.M. and McFarland M.J. (1989) A gauge to measure mass flow rate of sap in stems and trunks of woody plants. *Journal of the American Society for Horticultural Science*, **114**, 466–472.

Stull R.B. (1988) *An Introduction to Boundary Layer Meteorology*, Atmospheric sciences library, Kluwer Academic Publishers, p. 666.

Tennekes H. (1973) The logarithmic wind profile. *Journal of Atmospheric Science*, **30**, 234–238.

Thom A.S. (1975) Momentum, mass and heat exchange of plant communities. *In Vegetation and the atmosphere*, **1**, 57–109 Monteith J.L. (Ed.), 2 vols. Academic Press London.

Topp G.C., Davis J.L. and Annan A.P. (1980) Electromagnetic determination of soil water content: measurements in coaxial transmission lines. *Water Resources Research*, **16**, 574–582.

Van Bavel C.H.M., Brust K.J. and Stirk G.B. (1968) Hydraulic properties of a clay loam soil and the field measurement of water uptake by roots. II The water balance of the root zone. *Soil Science Society of America Proceedings*, **32**, 317–321.

Van der Molen M.K., Gash J.H.C. and Elbers J. (2004) Sonic anemometer (co)sine response and flux measurements: II The effect of introducing an angle of attack dependent calibration. *Agricultural and Forest Meteorology*, **122**, 95–109 in press.

Verhoef A., de Bruin H.A.R. and Van den Hurk B.J.J.M. (1997a) Some practical notes on the parameter kB^{-1} for sparse canopies. *Journal of Applied Meteorology*, **36**, 560–572.

Verhoef A., McNaughton K.G. and Jacobs A.F.G. (1997b) A parameterization of momentum roughness length and displacement height for a wide range of canopy densities. *Hydrology and Earth Systems Sciences*, **1**, 81–91.

Wallace J.S. and Holwill C.J. (1997) Soil evaporation from tiger-bush in south-west Niger. *Journal of Hydrology*, **188–189**, 426–442.

Webb E.K., Pearman G.L. and Leuning R.L. (1980) Correction of flux measurements for density effects due to heat and water vapour transfer. *Quarterly Journal of Royal Meteorological Society*, **106**, 85–100.

Welles J. and Cohen S. (1996) Canopy structure measurement by gap fraction analysis using commercial instrumentation. *Journal of Experimental Botany*, **47**(302), 1335–1342.

White M.A., Asner G.P., Nemani R.R., Privette J.L. and Running S.J. (2000) Measuring fractional cover and leaf area index in arid ecosystems: digital camera, radiation transmittance, and laser altimetry methods. *Remote Sensing of Environment*, **74**, 45–57.

Wilczak J.M., Oncley S.P. and Stage S.A. (2001) Sonic anemometer tilt correction algorithms. *Boundary-Layer Meteorology*, **99**, 127–150.

Wilson K., Goldstein A., Falge E., Aubinet M., Baldocchi D., Berbigier P., Bernhofer C., Ceulemans R., Dolman H., Field C., *et al.* (2002) Energy balance closure at FLUXNET sites. *Agricultural and Forest Meteorology*, **113**, 223–243.

41: Evaporation Modeling: Potential

RICHARD G ALLEN

Department of Civil Engineering and Department of Biological and Agricultural Engineering, University of Idaho, Kimberly, ID, US

Potential evaporation (E_p) has had a relatively broad range of definition over the past century. It is defined here as the quantity of water evaporated per unit area per unit time from an idealized extensive free-water surface under existing atmospheric conditions. Three primary, common means to estimate E_p have been used during the past century. These are (i) pan evaporation measurement, (ii) an estimate of potential evapotranspiration based on weather data, and (iii) a reference evapotranspiration. Of these three, reference evapotranspiration (ET_{ref}) has the more practical application. "Potential evapotranspiration" (ET_p) had widespread usage from the 1940s through the 1970s, when the term was used to represent a maximum evaporative index from which to derive estimates of actual ET from vegetation. However, there are several major, contrasting definitions for ET_p and several challenges associated with its usage. One of the primary definitions used for ET_p is the rate of evaporation and transpiration from a saturated (free-water) vegetated surface so that the evaporation process occurs at the potential level. Challenges in sustaining a saturated surface and in measuring weather data that are coincident with such a surface make this definition for ET_p theoretically attractive, but practically difficult. Standardized parameterizations of the Penman–Monteith equation are described for calculating ET for grass and alfalfa references. The reference evapotranspiration, despite some shortcomings, can be a consistent and reproducible index for a weather-based potential evaporation.

INTRODUCTION

Evaporation is the process whereby liquid water is converted to water vapor (vaporization) and removed from the evaporating surface (vapor removal). Water evaporates from a variety of surfaces, such as lakes, rivers, pavements, soils, and vegetation. The evaporation of liquid water at the Earth's surface and its consequent flux into the atmosphere is a major component of the hydrologic cycle and an essential ingredient in sustaining biological systems of the earth. Evaporation through plants drives the transpiration process for transport of minerals from soil to plant parts and provides evaporative cooling (*see* **Chapter 42, Transpiration, Volume 1**). Evaporation from soil dries the upper soil profile, thereby impacting strength of the soil surface and infiltration rates during precipitation. Globally, the annual volume of evaporation in essence equals the annual volume of precipitation, with relatively small differences caused by storage changes in ice fields, soil water, and groundwater recharge or extraction. Evaporation has high spatial variability similar to that of precipitation, but is dampened by rainfall runoff, surface energy limitations and availability of water stored in soil and water bodies.

Relatively large quantities of energy are required to change the state of water molecules from liquid to vapor. Solar radiation, and long-wave radiation (*see* **Chapter 39, Surface Radiation Balance, Volume 1**) and, to a lesser extent, heat from beneath the surface and from the lower atmosphere provide this energy. The driving force to remove vapor from the evaporating surface is the gradient between the vapor pressure at the surface and that of the overlying atmosphere. As evaporation proceeds, the surrounding air becomes more humid and the evaporation process will slow down if the humid air is not transferred to the atmosphere. The replacement of the saturated air with drier air is a strong function of wind speed. Hence, solar radiation, air temperature, air humidity, and wind speed are climatological parameters to consider when assessing the evaporation process.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

There is an inverse relation between actual evaporation and evaporative demand, as discussed from various perspectives in (**Chapter 42, Transpiration, Volume 1**; **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**; **Chapter 44, Evaporation from Lakes, Volume 1**; **Chapter 45, Actual Evaporation, Volume 1**). The relation occurs because the evaporation "demand" or "potential" varies with the heat energy stored in, and transported from, the lower atmosphere and the relative dryness (thirst) of the lower atmosphere (in addition to radiation energy), and because atmospheric characteristics depend on the "history" of the air mass as it is influenced by upwind evaporation processes. By definition, potential evaporation is rarely reached over a region, owing to lack of an extensive free-water surface, but is a useful index by which to characterize the atmospheric environment and by which to set limits when quantifying water fluxes.

Definition of Potential Evaporation

Potential evaporation (E_p) has had a broad range of definitions over the past century. E_p is defined here (and in **Chapter 45, Actual Evaporation, Volume 1**) as the quantity of water evaporated per unit area per unit time from an idealized extensive free-water surface under existing atmospheric conditions (Shuttleworth, 1992). This definition can in principle be applied to any surface, including vegetation, having a free-water (saturated) surface. However, owing to the influence of stomatal resistance (*see* **Chapter 42, Transpiration, Volume 1**), the saturated condition for vegetation only occurs during and briefly following rain or sprinkle irrigation events. This creates a problem befalling the use and measurement of E_p (or *potential evapotranspiration* as the process has sometimes been termed when applied to potential evaporation from vegetation).

Impact of Surface Characteristics

Most hydrologic process applications focus on areas having vegetated surfaces. Therefore, it is important that the definition of E_p, if it is used as an index, applies to those surfaces as well as to water bodies. This creates a challenge in defining a standardized E_p index to characterize climatic demand because the potential vapor flux from a saturated surface is influenced by the aerodynamic and radiation properties of the surface. Thus, E_p from an aerodynamically rough surface, such as a forest, will generally be greater than that from an aerodynamically smooth surface, such as short clipped grass (*see* **Chapter 42, Transpiration, Volume 1** and **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**) which is, inturn, more rough than an aerodynamically smooth water surface. E_p from a short clipped grass can also be greater than that from a deep, clear water body that absorbs much of the solar radiation beneath the free-water surface and thus makes it unavailable in real time for conversion to latent heat (i.e. E_p) (*see* **Chapter 44, Evaporation from Lakes, Volume 1**). Consequently, a range of E_p indices are needed to represent the maximum evaporation physically possible for specific surface conditions, even given the same atmospheric conditions.

Feedback to the Atmosphere

In addition to the impact of surface roughness on E_p rate, there are effects of negative feedback between the "demands" of the atmospheric boundary layer and the E_p rate. The negative feedback is caused by the humidification of the boundary layer and reduction of sensible heat content (*see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1** and **Chapter 44, Evaporation from Lakes, Volume 1**). These effects are not immediate over a small stretch of free-water surface, owing to the large storage content of the lower boundary layer for both heat and vapor. However, the effects become pronounced as the fetch of free-water surface increases to distances on the order of thousands of meters. Both of these modifications to the boundary layer (heat and vapor storage) reduce E_p. Therefore, to be strictly true to the definition of E_p given previously, E_p should be estimated (if from weather-based equations) only when the weather data are measured over the same saturated surface from which E_p is occurring. Otherwise, the weather data will not truthfully reflect the feedback that would occur under conditions of E_p. The exception to this is when one is estimating E or evapotranspiration from small stands of vegetation that have different aerodynamic, surface conductance, or wetness conditions to their surroundings. In this case, the E_p and current weather should not be expected to reach an equilibrium.

Another type of feedback concerns the wind velocity profile immediately above vegetation. Generally, as vegetation roughness increases, especially for dense vegetation, the wind velocity profile becomes less "steep" (smaller change in velocity with height) for several meters above the vegetation and the average velocity decreases. Allen and Wright (1997) found wind speed 3 m above 2-m tall corn (i.e. measured at 5 m above ground surface) to be 30% lower than wind speed 3 m above clipped grass. Differences between wind speed from standard weather stations located over short vegetation and wind speed over tall, dense forest can be even greater. Allen and Wright developed roughness-based translation equations to adjust for the impact of roughness on wind speed. This effect should be considered when estimating potential evaporation for specific vegetation.

Because E_p for vegetation only occurs during and for a few hours following wetting events, in principle, one can only calculate or estimate E_p using weather data collected during those times. This is, of course, not very useful for hydrologic process models that are operated

over long, continuous time periods and that need relatively continuous measures of an evaporative index to serve as a boundary condition. Therefore, concessions must be made, since most weather data are collected over nonsaturated terrain or nonsaturated vegetation and, although techniques are now well developed (*see* **Chapter 40, Evaporation Measurement, Volume 1**), evaporation is still challenging and costly to measure.

METHODS TO APPROXIMATE POTENTIAL EVAPORATION

Three primary, common means to estimate E_p have been used during the past century. These are (i) pan evaporation measurement, (ii) an estimate of potential evapotranspiration based on weather data, and (iii) a reference evapotranspiration. Of these three, reference evapotranspiration (ET_{ref}) has the more practical application. The term *evapotranspiration* (ET), developed over the last half-century (Thornthwaite, 1948), is used as a simplifying concession to describe the combined vapor fluxes from soil and vegetation. Some scientists argue that evaporation from soil and evaporation from plants via transpiration both constitute evaporation, so that only the term evaporation is necessary, and the use of the term transpiration is redundant. However, the term ET is useful for efficiency of communication and visualization, since the use of only the term evaporation generally requires qualification as "evaporation from soil" or "evaporation from vegetation leaf surfaces" or "evaporation of transpired water within leaves", or the sum of all these processes. For communication efficiency, the sum of all three of these evaporation components, which constitutes the total flux of vapor from a vegetated surface, is commonly referred to as ET (if one wishes, one can think of the term ET as representing "evaporation total").

Pan Evaporation

Evaporation from an evaporation pan has been used for several centuries as an approximate measure of E_p. However, research over the last several decades points to the need for caution in the use of general pan evaporation data to estimate lake evaporation or evapotranspiration (ET) from land. The standard evaporation pan in much of the world is the USA Class A Pan, which is 1.21 m in diameter and 254 mm deep, constructed of stainless steel, and placed above a 0.15-m tall open timber framework such that the top of the pan is about 0.4 m above the surrounding ground level. Two other commonly used pans are the Russian (Soviet) GGI-3000 (0.3 m^2) pan and the GGI-20 m^2 tank, both placed in the soil with only 0.075 to 0.1 m of rim above the soil surface. Details are given by the World Meteorological Organization (1970). Daily pan evaporation (E_{pan}) is computed from the change in water storage in the pan, inputs of precipitation and water added to maintain an adequate supply.

Owing to differing thermal characteristics between the pan and large water bodies, E_{pan} tends to overestimate the total amount of evaporation and distort the seasonal distribution (*see* **Chapter 44, Evaporation from Lakes, Volume 1**). Thus, on a seasonal basis, E_p usually peaks several months before the peak evaporation of deep lakes. Early studies in the United States in semiarid to arid climates revealed a very significant response in E_{pan} to pan size, but now it is clear that this was primarily the result of a decreasing effect of local advection with increasing size of the water surface area. Pruitt and Doorenbos (1977b) reported almost no difference (4%) between evaporation from a 0.62-m diameter Russian pan and a 5-m diameter Russian pan when both were located in a 5-ha irrigated grass field. On the other hand, at a dry site in Nevada, evaporation from the smaller Russian pan averaged some 1.6 times that of the larger pan, although when corrected for the net heat transfer from soil to the pans, a factor of 1.45 resulted (Hounam, 1973).

Young (1947) presented a good discussion of the problem of local pan environment in relation to estimating lake evaporation. Later studies in India by Ramdas (1957) and studies at Prosser, Washington and in California (Pruitt, 1960, 1966; State of California DWR, 1975, 1979) provided clear evidence that unless the local environment of a pan was taken into account, the estimation of ET (or of lake evaporation) was subject to errors of up to 35%. Pruitt and Doorenbos (1977a) introduced pan factors (k_p) to multiply by E_{pan} for estimating ET from a clipped grass reference surface and evaporation from shallow water bodies. These coefficients take into account the effects on pan evaporation of upwind fetch (both dry and moist), mean relative humidity, and total daily wind (Doorenbos and Pruitt, 1977; Jensen *et al.*, 1990; Shuttleworth, 1992; Allen *et al.*, 1996; FAO, 1998).

Pan maintenance can be challenging. The water level for a Class A pan must be kept within a range of 0.05 to 0.075 m below the top of the pan. If the pan is protected from birds, for example, with 12.5-mm (0.5-inch) mesh screen, k_p should be increased 5 to 10 percent (Stanhill, 1962). Pans made of monel metal, or those older galvanized metal pans, which have lost their original reflectance characteristics, may need a reduction to k_p of up to five percent. In general, the water should be kept clean, although turbidity differences appear to produce little difference in evaporation from Class A pans. There is an obvious need to avoid contamination by oil-related products. Pans are generally inoperable during winter when water can freeze. In some cases, pans are heated; however, the heat addition increases the E_{pan} above ambient conditions. Another complicated situation arises when pans are placed in a small enclosure surrounded by tall crops. The reduced

aerodynamic turbulence over the pan can dramatically reduce the E_{pan} measurement (Jensen *et al.*, 1990).

With the challenges in pan operation and maintenance, and the inability to precisely convert E_{pan} into E_p or even into a reference ET, and due to expense and difficulties in automation of pan measurements for electronic data recording and communication, pans are falling out of common usage. Use of E_{pan} is generally being replaced by the use of reference evaporation or reference ET calculated from weather data.

Potential Evapotranspiration

The term "potential evapotranspiration" (ET_p) had widespread usage from the 1940s through the 1970s (Thornthwaite, 1948; Jensen, 1974), when the term was used to represent a maximum evaporative index from which to derive estimates of actual ET (ET_a) from vegetation (Penman, 1963). However, there are several major, contrasting definitions for ET_p and several challenges associated with its usage. One of the primary definitions used for ET_p is as the rate of evaporation and transpiration from a saturated (free-water) vegetated surface so that the evaporation process occurs at the potential level. In this definition, ET_p is similar to E_p, but applied only to vegetation. As for E_p, the challenge with the saturated surface definition for ET_p is that the magnitude, although depending primarily on atmospheric conditions and surface albedo, varies with the surface characteristics including aerodynamic roughness (*see* **Chapter 42, Transpiration, Volume 1** and **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). Thus, the definition for ET_p should be tied to specific vegetation in regard to height, leaf area, and roughness. This is done using a predictive equation such as the Penman–Monteith (introduced as equation (1) *see* **Chapter 45, Actual Evaporation, Volume 1**) with very small or zero surface resistance and aerodynamic and radiative properties that fit the vegetation. Also, because conditions of saturated vegetation occur only for short periods during and following dewfall, rainfall, or irrigation by sprinkler, it is difficult to obtain the measurements of ET_p for free-water surface conditions by which to develop predictive equations.

A second definition for ET_p is to relax the requirement of a free-water surface (i.e. saturated vegetation surface) and to set ET_p equal to the rate of ET expected from relatively tall, dense, leafy vegetation having dry leaf surface, but that has relatively high soil-water content, is disease and stress free, and is therefore transpiring at a rate governed by nearly maximum leaf and boundary-layer conductances in conjunction with energy availability. The advantage of this definition for ET_p is that it follows more closely an upper bound on ET expected from vegetation on nonrainy days. It is this second definition that was used in papers and equation developments by Penman (1948, 1963) where ET_p was associated with ET measured from a clipped, cool-season grass. A challenge with this definition, however, is again the need to define the specific characteristics and aerodynamic structure of the vegetation representing ET_p. Associated with the definition of the upper bound for ET_p for a specific type of vegetation is the challenge of defining the value to use for the surface (leaf and canopy) conductances, which can be shown to vary with environmental factors of solar radiation, relative humidity at the leaf surface, leaf temperature, soil-water potential and carbon dioxide concentration (*see* **Chapter 42, Transpiration, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**). In some ways, this second definition of ET_p is the best one to use in hydrologic science, with estimates based on the Penman–Monteith or multilayer approaches. However, use of ET_{ref}, as discussed later, may be more consistent and convenient. Also, the feedback between the defined potential surface and its characteristics and the associated weather measurements should be considered. One should not "mix and match" ET_p from tall forest vegetation, for example, with weather data collected over a short grassed surface.

A third definition for ET_p, and perhaps the best one for process modeling, is as the potential ET to expect from any specific type and condition of vegetation or other terrestrial surface under conditions of sufficient soil-water to not inhibit transpiration. This definition, as applied to vegetation, applies even to conditions of low leaf area, for example, during initial plant development, and is therefore synonymous with what is often termed, in agricultural applications, as "potential crop ET" (FAO, 1998). The value for ET_p is strongly governed by the relative leaf area over the surface and leaf conductance properties. Some definitions for this type of ET_p assume a relatively dry soil surface between plants so that the soil evaporation component is small. Other definitions (as in FAO usage FAO, 1998) include a soil evaporation component that can change daily as soil dries. Actual evaporation (or actual ET) is modeled as equal to ET_p until a point when soil-water availability in the effective root-zone falls to a level that can no longer supply water to the plant at the ET_p rate (Hatfield and Allen, 1996).

Reference Evapotranspiration

The definition and use of the term "reference evapotranspiration" (ET_{ref}) was developed in the 1970s (Wright and Jensen, 1972; Pruitt and Doorenbos, 1977) to resolve ambiguities involved in the definition and interpretation of "potential evapotranspiration" as noted in the previous section. The "reference" descriptor points to the use of a specific type of vegetation or specific definition of vegetation properties to represent the evaporative index. Wright and Jensen suggested that maximum ET for nonsaturated conditions may be approximated by ET from a

well-watered reference crop of alfalfa (also called Lucerne) of a height of at least 0.2 m. Doorenbos and Pruitt described their ET_{ref} (termed ET_o) as "*the rate of evapotranspiration from an extensive surface of 8–15 cm tall, green grass cover of uniform height, actively growing, completely shading the ground and not short of water*". Subsequently, this definition specified the grass cover to be a cool-season grass having roughness, density, leaf area, and canopy resistance characteristics similar to perennial ryegrass (*Lolium perenne*) or alta fescue (*Festuca arundinacea* Schreb. "Alta"), since warm-season grass varieties, such as Bermuda (*Cynodon dactylon*), exercise considerable control over transpiration and can have lower ET_{ref} rates.

The benefit of using the reference concept is the ability to readily measure and validate reference ET using living, standardized vegetation. In addition, because stomatal control of the reference surface is intended to approximate that of most agricultural vegetation, ET_{ref} is generally more similar to actual ET than is E_p. An advantage of using ET_{ref} is that weather data are commonly measured above "standardized" weather surfaces that are usually grass or other short growing vegetation. Hence, the predicted ET flux is synchronized with the temperature, humidity, and wind measurements taken over the weather station surface and reflects the impact of feedback mechanisms between the vegetation and overlying boundary layer.

DEFINITION OF REFERENCE *ET*

For general prediction purposes, the complex aerodynamic (turbulent) structures within and above vegetation canopies and the effects of partitioning of net radiation and energy within the canopies can be described in terms of simple resistances (*see* **Chapter 40, Evaporation Measurement, Volume 1**; **Chapter 42, Transpiration, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**). In the most simple form, this is accomplished using the linear "big leaf" model of Monteith (1965, 1985) where two resistances, surface and aerodynamic, are assumed to operate in series between leaf interiors and some reference height above the vegetation. Bulk surface resistance (r_s) is related to the resistance of vapor flow through individual stomatal openings (r_l) and total leaf area (*see* **Chapter 42, Transpiration, Volume 1**). Aerodynamic resistance (r_a) describes the resistance to the random, turbulent transfer of vapor from the vegetation upward to the weather measurement height, and the corresponding vertical transfer of sensible heat away from or toward the vegetation. The r_a term (and to some degree, the r_s term) includes the effects of diffusive resistance through thin molecular layers along leaf surfaces, momentum transfer through pressure forces within the plant canopy, and turbulent transfer among canopy leaves and above the canopy. The r_a is affected by boundary-layer stability (*see* **Chapter 40, Evaporation Measurement, Volume 1**). The location of the r_s and r_a terms as used in the Monteith model are shown in Figure (1) of **Chapter 45, Actual Evaporation, Volume 1**.

Assuming that eddy diffusion transfer factors for latent heat and sensible heat are the same and that differences between transfer factors for momentum and those for heat can be quantified through a simple ratio (*see* **Chapter 40, Evaporation Measurement, Volume 1**), the Penman–Monteith (PM) form of the combination equation (Monteith, 1965) takes the form:

$$\lambda E = \frac{\Delta(R_n - G) + \rho_a c_p (e_s - e_a)/r_a}{\left(\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)\right)} \qquad (1)$$

where R_n and G are net radiation and soil heat-flux densities, Δ is the slope of the saturation vapor pressure–temperature relationship, ρ_a is mean air density, c_p is specific heat of air at constant pressure, $e_s - e_a$ is the vapor pressure deficit of air at the reference (weather measurement) height, and γ is the psychrometric constant. λE in (1) represents evaporation or ET expressed as a latent energy flux density, for example as $W\,m^{-2}$. λE is converted to units of water depth per unit time by dividing by the latent heat of vaporization, λ, and by the density of liquid water.

The Penman–Monteith equation as formulated in equation (1) includes all significant parameters governing energy exchange and the corresponding latent heat-flux (evapotranspiration) from a uniform expanse of vegetation. Most of the parameters in equation (1) can be readily measured or calculated from weather measurements on an hourly or shorter time basis. Equation (1) is generally capable of responding to changes in weather and climate in a manner similar to nonstressed, forage, or annual vegetation, so that it is reasonable to use this equation to represent ET_{ref} conditions.

The ET_{ref} value is typically transformed into an estimate of E_{act} during modeling by applying a cover coefficient, crop coefficient, or crop factor (K_c) that is defined as the ratio of E_{act} to ET_{ref}. The value for K_c when the soil surface is dry is strongly related to the fraction of ground covered by vegetation. K_c is further influenced by the height (roughness) and leaf structure of vegetation as well as surface conductance properties. Although the concept and use of K_c can be unappealing to users who desire a more theoretical and analytical structure to the ET estimate, for example, one that applies equation (1) directly to a specific surface using aerodynamic and surface properties that are defined and calibrated for example, as described in (*see* **Chapter 45, Actual Evaporation, Volume 1**), the consistent and reproducible behavior of the K_c

and the ability to set upper and lower limits and to estimate general values, based only on visual observation of vegetation, is appealing. The specific and direct application of equation (1) to vegetation without K_c and definition of ET_{ref} requires detail of roughness and leaf area properties over time as vegetation develops, the need to account for impacts of feedback between surface and boundary layer, the need to dynamically estimate r_s for the soil component as a function of water content and cover, and the need for a net radiation model specific to the canopy architecture. These needs require relatively complicated, well tested, and data-populated models such as those by Shuttleworth and Wallace (1985), Shuttleworth and Gurney (1990), Dolman (1993), and Huntingford *et al.* (1995). While these models, if sufficiently calibrated, are capable of producing more accurate estimates of ET_a (*see* **Chapter 42, Transpiration, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**) often, hydrologic model applications do not warrant the application intensity or time requirements, and the simple $K_c\,ET_{ref}$ method provides sufficiently accurate estimates. The $K_c\,ET_{ref}$ approach has greatly simplified the complexity and amount of information required to predict ET_a and has enabled the transfer of values for K_c between locations and between climates. This has been a primary reason for the wide acceptance and application of the $K_c\,ET_{ref}$ approach as a working model that can be used until more sophisticated methods become available for direct estimation of actual crop ET. FAO (1998) summarized K_c values for a wide range of agricultural crops and described means for estimating K_c based on visual descriptions of vegetation.

Basis for Reference *ET*

Reference ET is defined as the evapotranspiration from an extensive surface of reference vegetation having a standardized, uniform height and that is actively growing, completely shading the ground, has a dry, but healthy and dense leaf surface, and is not short of water. This definition has typically been applied to two common and standardized reference vegetation types: clipped, cool-season grass and full-cover alfalfa.

Standardized Definitions

It is generally accepted that the grass reference crop is a "cool-season", C-3 type of grass with roughness, density, leaf area, and bulk surface resistance characteristics similar to perennial ryegrass (*Lolium perenne* L.) or alta fescue (*Festuca arundinacea* Schreb. "Alta"). Alfalfa reference ET, typically abbreviated ET_r, was defined by Wright and Jensen (1972) as: "*... ET from well-watered, actively growing alfalfa with 8 in. (20 cm) or more of growth...*" and by Wright (1982) as "*...when the alfalfa crop was well-watered, actively growing, and at least 30-cm tall; so that measured ET was essentially at the maximum expected level for the existing climatic conditions"*. The height of alfalfa (*Medicago sativa* L., vs ranger) in the data set used to develop surface resistance algorithms for the ASCE Penman–Monteith application (Jensen *et al.*, 1990) to ET_r ranged from about 0.15 to 0.80 m in height and averaged 0.47 m (Allen *et al.*, 1989).

Generally, alfalfa ET_r is about 1.1 to 1.4 times that of grass ET, typically abbreviated ET_o, due to the increased roughness and leaf area of alfalfa. The higher value (1.4) represents the ratio of ET_r to ET_o under extremely arid and windy conditions (minimum daytime relative humidity (RH) $<20\%$ and wind speed $>5\,m\,s^{-1}$ (11 mph)) and the lower value (1.1) represents the ratio of ET_r to ET_o under humid, calm conditions (Pereira *et al.*, 1999; Wright *et al.*, 2000). ET_r is sometimes preferred over ET_o because its larger roughness and leafiness cause it to better approximate the upper limit of ET expected from all types of vegetation (Pereira *et al.*, 1999). However, alfalfa does not grow well under some tropical conditions and at elevations above about 2000 m. However, ET_r is now defined by ASCE-EWRI (2005) as a virtual reference crop that can be applied at all locations. Cool-season grasses can be cultivated over a wide range of climates, seasons and elevations, for example, at 4000 m in Bolivia (Garcia *et al.*, 2004).

ET Equations to Define ET_{ref}

Because of the challenges in growing and maintaining living reference vegetation, where the LAI and thus the value for r_s can vary between clippings for grass by nearly a factor of 2, a calibrated ET_{ref} equation is now generally used to represent the "hypothetical" or "virtual" reference. The FAO (Smith *et al.*, 1991, 1996; FAO, 1998) has adopted the PM equation as a standardized definition for ET_o where values for r_s, surface albedo, and aerodynamic roughness are fixed. FAO Irrigation and Drainage Paper No. 56 defined ET_o in terms of the PM equation as the rate of evapotranspiration from "*a hypothetical reference crop with an assumed crop height of 0.12 m, a fixed surface resistance of* $70\,s\,m^{-1}$ *and an albedo of 0.23"*, and where the reference surface "*closely resembles an extensive surface of green grass of uniform height, actively growing, completely shading the ground and with adequate water"*.

Penman–Monteith as a Definition

The Penman–Monteith (PM) equation was introduced in its full form as equation (1). The PM formulation of the combination equation incorporates aerodynamic and surface resistance terms that represent physical characteristics of the particular reference crop. Aerodynamic resistance in equation (1) is generally calculated as

$$r_a \frac{\left(\ln\left(\frac{z_u - d}{z_{om}}\right) - \Psi_m\right)\left(\ln\left(\frac{z_{T,e} - d}{z_{oh}}\right) - \Psi_h\right)}{k^2\,u_z} \qquad (2)$$

where z_u is the measurement height for wind speed, $z_{T,e}$ is the measurement height for temperature and vapor pressure, d is the zero-plane displacement height for the weather measurement surface, z_{om} is the roughness height for momentum transfer, z_{oh} is the roughness height for vapor and sensible heat transfer, k is the von Karman constant (0.41), u_z is the wind speed at the z_u height and Ψ_m and Ψ_h are integrated stability functions for momentum and sensible heat transfer (more theory on stability is given in **Chapter 40, Evaporation Measurement, Volume 1**). Definitions and procedures for calculating stability parameters Ψ_m and Ψ_h are available from Brutsaert (1982), Katul and Parlange (1992) and Allen *et al.* (1996). Normally, Ψ_m and Ψ_h can be assumed zero for ET_{ref} calculations because the reference surface is well watered so that boundary-layer stability is close to neutral or only mildly unstable or stable and values for Ψ_m and Ψ_h are small. This simplifies the calculation process.

Standardized parameterizations of equations (1) and (2) have been proposed by ASCE-EWRI (2005) for both ET_o and ET_r following the format adopted by FAO. When the supporting parameter equations for r_a, ρ_a and λ are reduced and combined into equation (1), the FAO styled, reduced equation used by ASCE-EWRI results

$$ET_{ref} = \frac{0.408\Delta(R_n - G) + \gamma \dfrac{C_n}{T + 273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + C_d u_2)} \quad (3)$$

where ET_{ref} applies to both clipped grass and alfalfa reference surfaces. ET_{ref} has units of mm d^{-1} for 24-h time steps and mm h^{-1} for hourly time steps, R_n and G are in MJ $m^{-2} d^{-1}$ or MJ m^{-2} h^{-1}, T is mean daily or hourly air temperature (°C), u_2 is mean daily or hourly wind speed at 2-m height (m s^{-1}), e_s and e_a are in kPa, Δ and γ are in kPa $°C^{-1}$, and C_n and C_d are coefficients that change with calculation time step, reference type (grass ET_o or alfalfa ET_r), and, in some cases, with time of day. Values for C_n and C_d are given in Table 1. The values for hourly C_d for ET_o by ASCE-EWRI stem from the use of $r_s = 50$ s m^{-1} during daytime and $r_s = 200$ s m^{-1} during nighttime and the 24-h time-step value for C_d for ET_o stems from $r_s = 70$ s m^{-1}. For additional reading on r_s for ET_o, see Allen *et al.* (2005b). For ET_r, the values for hourly C_d stem from the use of $r_s = 30$ s m^{-1} during daytime and $r_s = 200$ s m^{-1} during nighttime and the 24-h time-step value for C_d stems from $r_s = 45$ s m^{-1}. The standardized definitions imply vegetation heights for the ET_o and ET_r surfaces of 0.12 and 0.5 m. R_n can be measured or estimated. If R_n is estimated, FAO (1998) and ASCE-EWRI provide standardized estimation procedures that use a fixed albedo of 0.23. The widely used FAO-56 Penman-Monteith equation (FAO, 1998) uses $C_n = 900$ and $C_d = 0.34$ for daily timesteps. For hourly timesteps, Allen *et al.* (2005b) recommend the use of values shown for ET_o in Table 1 for the FAO method.

The use of the standardized definitions for ET_o and ET_r provide for a consistent index of evapotranspiration that is readily compared between climates, locations, and time periods. The definition, although standardized, is tied to measurable surfaces, so that the predictive equation (3) can be compared to local measurements when desired or necessary.

Limitations of ET_{ref}

Although useful as a standardized climatic evaporation index, ET_o tends to underpredict maximum ET expected from tall vegetation such as trees, brush, and tall grasses like corn and sugar cane when the taller vegetation has unrestricted access to soil-water. Underprediction can be as much as 40% in arid climates and is caused by the close coupling of tall vegetation to the boundary layer. Some of the underprediction is moderated, however, by the fact that temperature and humidity of the lower boundary layer over rough vegetation will be cooler and more humid under conditions of maximum ET, owing to the more efficient scalar transport. Unfortunately, this feedback is generally not taken into account in models when weather data collected over grass are used to predict ET from taller vegetation. The taller alfalfa reference overcomes many of the shortcomings of grass in terms of aerodynamic roughness, but its use is less common in some parts of the globe.

Both definitions for ET_{ref} (clipped grass and alfalfa) assume an extensive, well-watered fetch for the evaporating surface as well as for the weather measurement. In arid settings, air temperature and relative humidity are substantially impacted by absence of an evaporating surface. When

Table 1 Values for C_n and C_d in equation (3) (after ASCE-EWRI, 2004)

	Short reference, ET_o		Tall reference, ET_r		Units for ET_o, ET_r	Units for R_n, G
Calculation time step	C_n	C_d	C_n	C_d		
Daily	900	0.34	1600	0.38	mm d^{-1}	MJ m^{-2} d^{-1}
Hourly during daytime	37	0.24	66	0.25	mm h^{-1}	MJ m^{-2} h^{-1}
Hourly during nighttime	37	0.96	66	1.7	mm h^{-1}	MJ m^{-2} h^{-1}

placed into the reference equation (3), data collected over a dry surface will cause as much as 30% overprediction of ET_{ref}. For further reading, see Allen *et al.* (1983), Ley *et al.* (1996), Jensen *et al.* (1997), FAO (1998), Temesgen *et al.* (1999) and ASCE-EWRI (2005).

One last limitation in the standardized definition for ET_{ref} is that it assumes constant surface resistance and albedo, regardless of environmental conditions or sun angle. When net radiation is measured over the reference surface, sun angle impacts on albedo are taken into account. This poses a challenge when converting the ET_{ref} into actual ET using some type of K_c under conditions where the stomatal behavior of the vegetation being modeled may vary substantially during the day.

Use of ET_{ref} as an Index

ET_o as defined by equation (3) has been applied within a global weather database by the International Water Management Institute that is published in the form of a World Climate Atlas (URL = `http://www.iwmi.cgiar.org/WAtlas/atlas.htm`). This database contains monthly mean weather and ET_o data for the 1961–1990 period and is useful for providing general ET_{ref} inputs for hydrological modeling of river basins and for extracting climate inputs for vegetation modeling. Droogers and Allen (2002) investigated the overall quality and populational behavior of the World Climate Atlas ET_o database. A similar, but independent, compilation of mean monthly weather data including ET_o by equation (3) was published by FAO (1993) as the CLIMWAT database (URL = `http://www.fao.org/ag/agl/aglw/climwat.stm`).

The alfalfa ET_r has been used as a calibration means within the METRIC energy balance model (Tasumi *et al.*, 2005; Allen *et al.*, 2005a) to improve the accuracy of the ET surface derived from satellite images. ET_{ref} is used to interpolate ET maps between satellite image dates within the remote sensing surface energy balance algorithm for land (SEBAL) (Bastiaanssen *et al.*, 1998) and METRIC, and other satellite-based energy balance procedures for deriving actual ET.

PENMAN VERSUS PENMAN–MONTEITH

The Penman–Monteith equation (1) was an extension to the original Penman (1948, 1963) equations where a wind function (w_f) in the aerodynamic term of the numerator was replaced with r_a and the $(1 + r_s/r_a)$ term was added to the denominator. Under free-water surface conditions, r_s tends toward zero, and equation (1) reverts to the Penman equation, with the exception that r_a is used in the numerator rather than the w_f, where generally, $w_f = a + b\,u$, where a and b are empirical coefficients and u is wind speed at some height above the surface (*see* **Chapter 44, Evaporation from Lakes, Volume 1**). In some regards, a distinct advantage of the w_f of the Penman equation is that it has a lower value (parameter a) when u decreases to zero or nearly zero. This is not the case for the classical equations used for r_a in the definition for ET_{ref}, (for example, equation (2) when the instability functions, Ψ, are set to zero), so that r_a tends toward infinity as u decreases toward zero. This behavior for r_a is unrealistic, as under conditions of low wind speed, for example, less than $0.5\,m\,s^{-1}$ at 2-m height, and $R_n - G > 0$, buoyancy forces caused by surface heating will play a significant role in transport of air away from the surface, thereby sustaining some aerodynamic transport. Thus, there should be some maximum value for r_a (conversely a minimum value for u in equation (3)) when the PM equation is applied for ET_{ref}. FAO (1998) and ASCE-EWRI (2005) have recommended limiting wind speed at 2-m height to $0.5\,m\,s^{-1}$ or greater when calculating ET_{ref}.

With the limitation imposed on wind speed in equation (3) for ET_{ref} or with the use of stability correction in equation (2) when used in equation (1) for E_p or ET_p, the Penman–Monteith method is recommended over the Penman, because the explicit definition for r_a via equation (2) allows users to easily modify r_a for changing vegetation roughness. This modification is not very straightforward with the w_f in the Penman equation.

QUALITY OF WEATHER DATA

The accuracy and quality of the E_p, ET_p, or ET_{ref} estimation is only as accurate as the weather data on which it is based. All E and ET equations require weather or climate data that reflect the environment of the area for which ET is estimated. Weather data should be screened before use. This is especially important with electronically collected data, since human oversight and maintenance may be limited. When weather measurements are determined to be faulty, they can be adjusted or corrected using a justifiable and defensible procedure, or the user may elect to replace perceived faulty data with estimates. Simple, visual procedures for qualifying weather data for use in ET estimation were described by Allen (1996), FAO (1998) and ASCE-EWRI (2005).

USE OF SIMPLIFIED EMPIRICAL EQUATIONS

A large number of empirical methods have been developed over the last century by scientists and specialists worldwide to estimate E or ET using a variety of climatic variables. Relationships were often subject to rigorous local calibrations, but proved to have limited global validity. The application of equations (1) and (3) requires data for air

temperature (T), vapor pressure (e_a), net radiation (R_n) or solar radiation (R_s) and wind speed (u). If some of the required weather data are missing or cannot be calculated, one must make a choice. Either equation (1) or (3) can be retained and the missing data estimated using a reliable technique that keys off other data or historical information, or an empirical equation is used that does not call for the missing data. Smith *et al.* (1991), FAO (1998) and ASCE-EWRI (2005) made strong arguments for utilizing the same (theoretically valid) ET_{ref} equation, regardless of data availability, and with estimation of missing data. The basis for the arguments is that one should obtain a more accurate estimate of E or ET using a theoretically correct method that considers all major factors, even when one or more factors are best estimates, than one will find when using an empirical method that essentially "closes its eyes" to many facets of the missing factor. FAO (1998) and ASCE-EWRI proposed general, relatively simple procedures for estimating missing parameters. These include estimating dewpoint temperature from daily minimum air temperature (less some fixed offset that is a function of climatic aridity), estimating R_s as a function of daily maximum and minimum air temperature and extraterrestrial radiation, and estimating missing wind speed using long-term monthly values characteristic of the area. More sophisticated weather generation models are available for estimating missing weather data that preserve some of the correlation between weather parameters and that preserve natural random variation (Richardson and Wright, 1984; Johnson *et al.*, 1996; Nelson, 2004; Meyer *et al.*, 2004).

If empirical methods for ET are to be applied, the four more popular and perhaps dependable methods are the Hargreaves *et al.* (1985), the Priestley and Taylor (1972), and the Makkink (1957) methods. The 1985 Hargreaves equation, which requires only daily maximum and minimum air temperature data along with calculated extraterrestrial radiation, generally provides relatively dependable estimates (Droogers and Allen, 2002; Hargreaves and Allen, 2003). FAO (1998) recommended calibrating the Hargreaves equation against the PM at regional sites, especially if the climate is subhumid to humid. Garcia *et al.* (2004) found the Hargreaves method performed relatively well at two arid locations near 3800-m elevation in Bolivia, but there was worse performance at two more humid locations.

As discussed in (*see* **Chapter 44, Evaporation from Lakes, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**), the Priestley–Taylor method, requiring only R_s and T, defines an equilibrium E or ET for large regions having adequate water supply. This implies it should be used under subhumid to humid conditions. The multiplier for the Priestley–Taylor method generally requires enhancement under arid conditions (Jury and Tanner, 1975; Jones and Kiniry, 1986; Steiner *et al.*, 1991). The Makkink method, also requiring only R_s and T, can be used if humidity data are considered to be of poor quality. Otherwise, one should trust the PM method (equations (1) and (3)) to produce a more accurate estimate, even in humid climates where the vapor pressure deficit ($e_s - e_a$ in the PM equation) is at times small. When humidity data are judged to be poor, one can replace the data with estimates based on daily minimum air temperature or can retreat to the empirical methods.

CONCLUSIONS

The theoretical concepts of potential evaporation, E_p, and potential evapotranspiration, ET_p, make them useful indices and definitions. However, the conditions behind the terms are generally difficult to sustain within hydrologic systems for measurement and are difficult to sustain even for weather data collection, owing to the importance of collecting temperature and humidity data that are coincident to the evaporating surface being modeled. The reference ET surfaces and definitions, ET_o and ET_r, for grass and alfalfa (i.e. short and tall references), are associated with a different set of problems and challenges when used for hydrologic modeling. However, the ET_{ref} is useful as a consistent and reproducible climatic index and is generally synchronized with weather data collected over well-watered grassed surfaces. The collection and use of pan evaporation data is losing popularity as improvements in the ET_{ref} definitions become accepted, and owing to the expense and the difficulty in making accurate automated measurements of evaporation from pans.

FURTHER READING

Monteith J.L. (1995) Accommodation between transpiring vegetation and the convective boundary layer. *Journal of Hydrology*, **166**, 251–263.

REFERENCES

Allen R.G. (1996) Assessing integrity of weather data for use in reference evapotranspiration estimation. *Journal of Irrigation and Drainage Engineering-ASCE*, **122**(2), 97–106.

Allen R.G., Brockway C.E. and Wright J.L. (1983) Weather station siting and consumptive use estimates. *Journal of Water Resources Planning and Management, ASCE*, **109**(2), 134–146.

Allen R.G., Jensen M.E., Wright J.L. and Burman R.D. (1989) Operational estimates of evapotranspiration. *Agronomy Journal*, **81**, 650–662.

Allen R.G., Pruitt W.O., Businger J.A., Fritschen L.J., Jensen M.E. and Quinn F.H. (1996) Evaporation and transpiration. In *ASCE Handbook of Hydrology*, ASCE Manuals and Reports on Engineering Practice No. 28, New York, pp. 125–252.

Allen R.G., Pruitt W.O., Wright J.L., Howell T.A., Ventura F., Snyder R., Itenfisu D., Steduto P., Berengena J., Baselga J., *et al.* (2005b) Standardized surface resistance for hourly calculation of reference ETo by the FAO56 Penman-Monteith method. *Agricultural Water Management*, **73**(3), 17–18.

Allen R.G., Tasumi M., Morse A. and Trezza T. (2005a) A landsat-based energy balance and evapotranspiration model in western US water rights regulation and planning. *Journal of Irrigation and Drainage System*, **19**(2), 10–11.

Allen R.G. and Wright J.L. (1997) Translating wind measurements from weather stations to agricultural crops. *Journal of Hydraulic Engineering-ASCE*, **2**(1), 26–35.

ASCE-EWRI (2005) *The ASCE Standardized Reference Evapotranspiration Equation*, Technical Committee report to the Environmental and Water Resources Institute of the American Society of Civil Engineers from the Task Committee on Standardization of Reference Evapotranspiration, p. 173.

Bastiaanssen W.G.M., Menenti M., Feddes R.A. and Holtslag A.A.M. (1998) A remote sensing surface energy balance algorithm for land (SEBAL): 1. Formulation. *Journal of Hydrology*, **212–213**, 198–212.

Brutsaert W. (1982) *Evaporation into the Atmosphere*, D. Reidel Publishing Company: Dordrecht, p. 300.

Dolman A.J. (1993) A multiple source land surface energy balance model for use in GCMs. *Agricultural and Forest Meteorology*, **65**, 2–45.

Doorenbos J. and Pruitt W.O. (1977) *Crop Water Requirements*, Irrigation and Drainage Paper No. 24, (rev.), FAO: Rome, p. l44.

Droogers P. and Allen R.G. (2002) Estimating reference evapotranspiration under inaccurate data conditions. *Irrigation and Drainage Systems*, **16**, 33–45.

FAO (1993) *CLIMWAT for CROPWAT*, Irrigation and Drainage Paper 49, Food and Agriculture Organization of the United Nations: Rome, p. 113.

FAO (1998) *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*, Irrigation and Drainage Paper 56, Food and Agriculture Organization of the United Nations: Rome, p. 300.

Garcia M., Raes D. and Allen R. (2004) Reference evapotranspiration in the Bolivian highlands (altiplano). *Agricultural and Forest Meteorology*, **125**, 67–82.

Hargreaves G.H. and Allen R.G. (2003) History and evaluation of the Hargreaves evapotranspiration equation. *Journal of Irrigation and Drainage Engineering-ASCE*, **129**(1), 53–63.

Hargreaves G.L., Hargreaves G.H. and Riley J.P. (1985) Agricultural benefits for Senegal river basin. *Journal of Irrigation and Drainage Engineering-ASCE*, **111**, 113–124.

Hatfield J.L. and Allen R.G. (1996) Evapotranspiration estimates under deficient water supplies. *Journal of Irrigation and Drainage Engineering-ASCE*, **122**(5), 301–308.

Hounam C.E. (1973) *Comparison between Pan and Lake Evaporation*, World Meteorological Organization, 354, Technical Note 126.

Huntingford C., Allen S.J. and Harding R.J. (1995) An intercomparison of a single and a dual-source vegetation-atmosphere transfer model applied to transpiration from Sahelian Savannah. *Boundary-Layer Meteorology*, **74**, 397–418.

Jensen M.E. (Ed.) (1974) Consumptive use of water and irrigation water requirements. *Irrigation and Drainage Division Report*, ASCE, p. 215.

Jensen M.E., Burman R.D. and Allen R.G. (1990) *Evapotranspiration and Irrigation Water Requirements*, ASCE Manuals and Reports on Engineering Practice No 70, ASCE, p. 350.

Jensen D.T., Hargreaves G.H., Temesgen B. and Allen R.G. (1997) Computation of ETo under nonideal conditions. *Journal of Irrigation and Drainage Engineering-ASCE*, **123**(5), 394–400.

Johnson G.L., Hanson C.L., Hardegree S.P. and Ballard E.B. (1996) Stochastic weather simulation: overview and analysis of two commonly used models. *Journal of Applied Meteorology*, **35**, 1878–1896.

Jones C.A. and Kiniry J.R. (Eds). (1986) *CERES-Maize: A Simulation Model of Maize Growth and Development*, A&M University Press: College Station.

Jury W.A. and Tanner C.B. (1975) Advection modification of the Priestley and Taylor evapotranspiration formula. *Agronomy Journal*, **67**, 569–572.

Katul G.G. and Parlange M.B. (1992) Estimation of bare soil evaporation using skin temperature measurements. *Journal of Hydrology*, **132**, 91–106.

Ley T.W., Allen R.G. and Hill R.W. (1996) Weather station siting effects on reference evapotranspiration. *Evapotranspiration and Irrigation Scheduling, Proceedings of the International Conference*, ASAE: San Antonio, pp. 727–734.

Makkink G.F. (1957) Testing the Penman formula by means of lysimeters. *Journal of the Institution of Water Engineers*, **11**(3), 277–288.

Meyer C.R., Renschler C. and Vining R.C. (2004) Implementing quality control techniques for random number generators to improve stochastic weather generators: the CLIGEN experience. *13th Conference on Applied Climatology*, Section 5.7, American Meteorological Society, p. 5.

Monteith J.L. (1965) Evaporation and the environment. In *The State and Movement of Water in Living Organisms, XIXth Symposium*, Society for Experimental Biology, Cambridge University Press: Swansea, pp. 205–234.

Monteith J.L. (1985) Evaporation from land surfaces: progress in analysis and prediction since 1948. *Advances in Evapotranspiration, Proceedings of the ASAE Conference on Evapotranspiration*, ASAE, St. Joseph, Michigan: Chicago, pp. 4–12.

Nelson R. (2004) ClimGen – climatic data generator user's manual. *Department of Biological Systems Engineering*, Washington State University: Pullman.

Penman H.L. (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A*, **193**, 120–146.

Penman H.L. (1963) *Vegetation and Hydrology*, Technical Communication No. 53, Commonwealth Bureau of Soils: Harpenden, p. 125.

Pereira L., Perrier A., Allen R.G. and Alves I. (1999) Evapotranspiration: concepts and future trends. *Journal of Irrigation and Drainage Engineering-ASCE*, **125**(2), 45–51.

Priestley C.H.B. and Taylor R.J. (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, **100**, 81–92.

Pruitt W.O. (1960) Relation of consumptive use of water to climate. *Transactions of the American Society of Agricultural Engineers*, **3**(l), 9–l3, l7.

Pruitt W.O. (1966) Empirical method of estimating evapotranspiration using primarily evaporation pans. *Proceedings of Conference on Evapotranspiration and its Role in Water Resources Management*, ASAE: Chicago, December, pp. 57–61.

Pruitt W.O. and Doorenbos J. (1977a) Background and development of methods to predict reference crop evapotranspiration (ETo). *Appendix II, Crop Water Requirements*, Irrigation and Drainage Paper No. 24. (rev.), FAO: Rome, pp. 108–119.

Pruitt W.O. and Doorenbos J. (1977b) Empirical calibration, a requisite for evapotranspiration formulae based on daily or longer mean climatic data? *International Round Table Conference on Evapotranspiration*, ICID: Budapest, p. 20.

Ramdas L.A. (1957) Evaporation and potential evapotranspiration over the Indian sub-continent. *Indian Journal of Agricultural Sciences*, **27**(2), l37–l49.

Richardson C.W. and Wright D.A. (1984) *WGEN: A Model for Generating Daily Weather Variables*, Agricultural Research Service, ARS-8, USDA, p. 83.

Shuttleworth W.J. (1992) Evaporation. In *Handbook of Hydrology*, Maidmentt D.R. (Ed.), McGraw Hill: New York, pp. 4.1–4.53.

Shuttleworth W.J. and Gurney R.J. (1990) The theoretical relationship between foliage temperature and canopy resistance in sparse crops. *Quarterly Journal of the Royal Meteorological Society*, **116**, 497–519.

Shuttleworth W.J. and Wallace J.S. (1985) Evaporation from sparse crops - an energy combination theory. *Quarterly Journal of the Royal Meteorological Society*, **111**, 839–853.

Smith M., Allen R.G., Monteith J.L., Pereira L.S., Perrier A. and Pruitt W.O. (1991) *Report on the Expert Consultation on Procedures for Revision of FAO Guidelines for Prediction of Crop Water Requirements*, Land and Water Development Division, United Nations Food and Agriculture Service: Rome.

Smith M., Allen R.G. and Pereira L.S. (1996) Revised FAO methodology for crop water requirements. *Proceedings International Conference on Evapotranspiration and Irrigation Scheduling*, ASAE: San Antonio, pp. 116–123.

Stanhill G. (1962) The control of field irrigation practice from measurements of evaporation. *Israel Journal of Agricultural Research*, **12**, 51–62.

State of California DWR (1975) *Vegetative Water use in California*, Department of Water Resources Bulletin No. 113-3, p. 104.

State of California DWR (1979) *Evaporation from Water Surfaces in California*, Department of Water Resources Bulletin No. 73–79, p. 163.

Steiner J.L., Howell T.A. and Schneider A.D. (1991) Lysimetric evaluation of daily potential evaporation models for grain sorghum. *Agronomy Journal*, **83**, 240–247.

Tasumi M., Allen R.G., Trezza R. and Wright J.L. (2005) Satellite-based energy balance to assess within-population variance of crop coefficient curves. *Journal of Irrigation and Drainage Engineering-ASCE*, **131**, 94–109.

Temesgen B., Allen R.G. and Jensen D.T. (1999) Adjusting temperature parameters to reflect well-watered conditions. *Journal of Irrigation and Drainage Engineering-ASCE*, **125**(1), 26–33.

Thornthwaite C.W. (1948) An approach toward a rational classification of climate. *Geographical Review*, **38**, 55.

World Meteorological Organization (1970) *Guide to Hydrometeorological Practices*, WMO No. 168.TP.82, WMO: Geneva.

Wright J.L. (1982) New evapotranspiration crop coefficients. *Journal of Irrigation and Drainage Engineering-ASCE*, **108**(2), 57–74.

Wright J.L., Allen R.G., Howell T.A. (2000) Comparison between evapotranspiration references and methods. In *Proceedings of the National Irrigation Symposium*, Evans R.G., Benham B.L. and Trooien T.P. (Eds.), ASAE: Phoenix, November 14–16, 2000, p. 251–259.

Wright J.L. and Jensen M.E. (1972) Peak water requirements of crops in Southern Idaho. *Journal of the Irrigation and Drainage Division, ASCE*, **96**(1), 193–201.

Young A.A. (1947) *Evaporation from Water Surfaces in California*, California Department of Public Works, Division of Water Resources and the U.S. Department of Agriculture: SCS. Bulletin No. 54, p. 68.

42: Transpiration

JOHN ROBERTS

Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

The transpiration process is the uptake of water by plant roots, transport through the plant and evaporation from the leaf through pores called stomata. Evaporation of water from the leaf is determined by atmospheric conditions such as radiation, temperature and humidity deficit but the plant can limit transpiration by partial or complete stomatal closure. Generally stomata open in response to increasing radiation but tend to close with increasing air humidity deficit and reduced availability of soil moisture. Because the stomata have to be open in daylight for the entry of carbon dioxide into the leaf for the photosynthesis process, water loss is an inevitable consequence. Nevertheless transpiration itself has important roles. Nutrients are brought into the plant when water is taken up from the soil and evaporation of transpired water prevents leaf temperature reaching supra-optimal levels. There are numerous ways that transpiration might be measured. These include measurements of soil water changes below vegetation or changes in atmospheric humidity above vegetation. Alternatively measurements can be made on individual plants or leaves. In most circumstances the source of water for transpiration is the soil and principally the surface soil layers where most roots are found. In the future, increased levels of atmospheric CO_2 are expected to reduce transpiration through reduction in stomatal aperture.

WHAT IS TRANSPIRATION?

Transpiration is the process by which water is evaporated from within a plant. Essentially, water is evaporated through small holes (known as *stomata*) in the leaves, and this draws water up through the plant (in microscopic tubes termed *xylem*) from the soil. This "transpiration stream" brings water to the plant to be used in photosynthesis, to produce carbohydrates, and to maintain turgidity (rigidness) in the cells and tissues. However, very little of the water is actually used in photosynthesis and to maintain turgidity. Most of the water sucked up from the soil is evaporated through the stomata. A primary purpose of the stomata is to exchange carbon dioxide and oxygen with the atmosphere in addition to regulating the loss of water from the leaves. The movement of water up from the soil through the plant plays a key role in bringing minerals from the soil into and through the plant. In situations where leaves experience a high radiation loading, leaf temperatures can be critically high. Cooling of the leaf by the dissipation of heat during the evaporation of transpired water is another important role for transpiration.

There is a continuous stream of water from within the leaves of plants down through the plant to the roots and soil. Water molecules bind together and these bonds have substantial strength. There is a continuous column of water from within the leaf drawing up water from the soil. The cohesive forces of water molecules mean that water columns down through the plant can be maintained under significant tensions. In very tall trees, these tensions can be considerable. Suctions equivalent to 5 MPa (50 bars) have commonly been reported for actively transpiring tall trees.

Transpiration from vegetation is usually reported as a depth of water (mm), in the same way that rainfall and evaporation are reported. Transpiration might range from very low or zero in completely water-stressed vegetation, sparse crops, or vegetation in winter. The highest values of transpiration might be up to rates estimated as the potential transpiration rate (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**). In this case, it would be expected that the vegetation completely covered the ground, there was no shortage of soil water and climatic conditions are optimal, for example, high summer or tropical conditions. Transpiration can exceed the potential rate if extra energy is available as advection.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Annual transpiration of global forest types

Vegetation	Location	Annual transpiration (mm)	Reference
Tropical rainforest	Manaus, Brazil (2°57′S: 59° 57′W)	1030	Shuttleworth (1988)
Southern European evergreen oak	Evora, Portugal (38° 32′N: 8° 01′W	207	David *et al.* (2004)
Temperate coniferous forest	Thetford, UK(52° 25′N: 0° 39′E)	352	Gash and Stewart (1977)
Boreal coniferous forest	Saskatchewan, Canada (53° 55′N: 104° 41′W	204	Saugier *et al.* (1997)

In summer conditions in the United Kingdom, maximum transpiration rates of up to 4 mm day^{-1} have been measured routinely in forests although much higher rates ~8 mm day^{-1} have been observed in fast-growing short rotation coppice plantations. In forests with 750 trees ha^{-1}, a transpiration rate of 4 mm day^{-1} would mean that 50 kg of water per day would be lost on average from each tree. In a wheat or barley field with around 250 stems m^{-2}, transpiration loss through each stem would be of the order of 150 gm.

An insight into the range of annual forest transpiration that might be encountered can be achieved by comparing values from various studies carried out in a range of forest types occurring from boreal to tropical regions. The annual transpiration (with associated information) of forest types occurring in different global regions are given in Table 1. As expected the highest annual total is found in tropical rainforest in Brazil. The high annual total is largely a consequence of the evergreen canopy and therefore year-round transpiration with no limitations of solar radiation, air temperature, or available soil moisture. Potential evaporation rates in this area of the Amazon basin would enable much higher transpiration, but the reduction of canopy conductance in response to an increased air vapor pressure deficit (see Figure 1) means that daily transpiration is often around 3.5 mm, barely different from transpiration of Scots pine *(Pinus sylvestris)* measured at Thetford Forest on summer days.

Annual transpiration from the evergreen oak *(Quercus rotundifolia)* woodland in Portugal is 207 mm. Although this estimate does not include losses from ground vegetation beneath the trees, the low transpiration rate is largely a consequence of the sparse open canopy of the woodland. Shortage of water is probably not an issue as the trees had access to groundwater. The annual transpiration from the boreal forest is also low. Although the growing season is short (~140 days), a major constraint on transpiration is probably soil temperature, which will still be cold enough to limit root water uptake, and probably also mineral nutrients throughout the growing season. This was considered as a major factor in producing low stomatal conductances in the jack pine *(Pinus banksiana)* in the boreal forest.

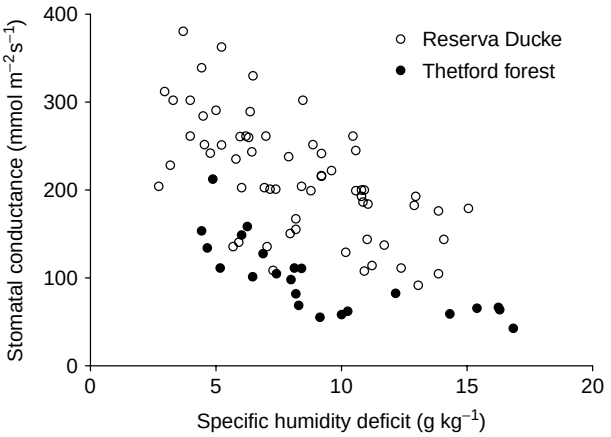


Figure 1 The decline in leaf stomatal conductance with air humidity deficit in *Piptadenia suaveolens*, an upper canopy tree species at the Reserva Florestal Ducke, Manaus, Brazil (unpublished data from John Roberts) and the upper canopy of Scots pine (*Pinus sylvestris* L.) at Thetford Forest, UK (Redrawn from Beadle *et al.*, 1985, *Journal of Applied Ecology* **22**, 557–571, by permission of British Ecological Society)

Transpiration from the temperate coniferous forest is higher than both the southern European and boreal forests. Although not constrained by water stress, daily transpiration is likely to be limited by a strong decline in stomatal conductance with increasing vapor pressure deficit (Figure 1). The annual transpiration of around 325–350 mm year^{-1} shown for Thetford Forest was shown to be very similar for many woodlands (both broadleaf and coniferous) in Europe by Roberts (1983). Roberts identified a number of factors that might contribute to this similar transpiration. Few forests are limited by water stress, and daily transpiration is constrained by probable links between stomatal/surface conductance and air humidity deficit. Furthermore, the presence of understory vegetation below an open tree cover will have a significant role in eliminating tree transpiration differences between dense and open forests. One factor that has been shown to be important in determining transpiration from forests and woodlands is the age of the trees. There is now substantial evidence from studies both on trees and catchments that as trees age their transpiration declines.

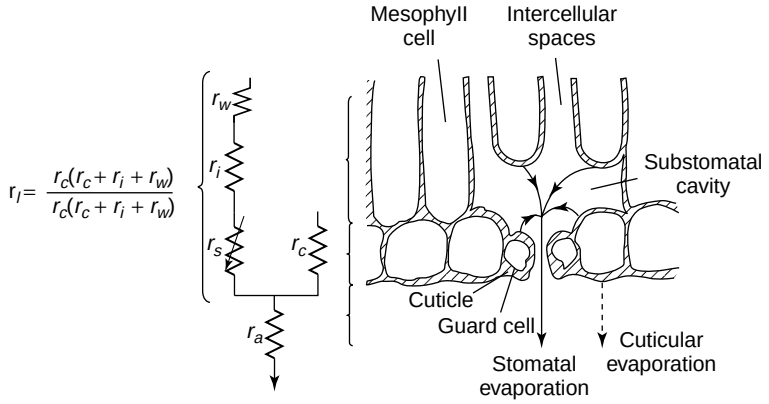


Figure 2 Pathways for water loss from one surface of a leaf showing the boundary layer (r_a), cuticular (r_c), variable stomatal (r_s), intercellular space (r_i), wall (r_w), and leaf (r_l) resistances. The total leaf resistance is the parallel sum r_l for upper and lower surfaces (Redrawn from Jones (1992), Plants and Microclimate, Second Edition, Cambridge University Press)

Practically all of the water lost from vegetation will have been taken from the soil. The amount of water taken up from the soil and used in metabolic processes, for example photosynthesis, is trivial. Some water is lost from plants as transpiration does not come directly from the soil but from storage in the body of the plant. Generally, this stored water is a smaller fraction of the total daily transpiration than water coming directly from the soil. In any case, water lost from storage in the plant tissues will be replaced with water from the soil, and this may occur in the following night, or in the case of some forests, in the following winter.

CONTROLS OF TRANSPIRATION

The evaporation of water from leaf surfaces is influenced by all the factors involved in evaporation from free water surfaces; evaporation is increased as energy inputs increase, and as the vapor pressure gradient from the leaf surface into the atmosphere increases. Evaporation can also be enhanced by increases in wind speed. Nevertheless, there is a particularly important control in the leaf surface. The apertures (stomata) in the leaf surface are able to regulate the loss of water vapor as transpiration. Closure of the stomata also means that the entry of carbon dioxide into the leaf for fixation by photosynthesis will be limited.

A comprehensive discussion of stomata, their structure, occurrence, function, and factors in the control of stomatal opening will be found in Willmer and Fricker (1996). Stomata may occur in small numbers on stems and leaf stalks but are much more common on one or more leaf surface. Stomata occur in more or less equal numbers over the surfaces of grass and cereal leaves, and also in some conifers for example. The most common situation, however, is for stomata to be found on the undersides of leaves, and this is the situation for tropical and temperate broad-leaved trees, and herbaceous plants. Figure 2 shows a cross-section through a stomatal apparatus, with the stomatal opening occurring on the underside of the leaf. Figure 2 shows the resistances met as water vapor is transpired from one surface of the leaf. Evaporation of water from cell walls is regarded by some to be the site of the first resistance (r_w) in the water loss pathway. The transfer resistance within the intercellular spaces (r_i) is followed by the stomatal resistance (r_s), and the boundary-layer resistance (r_b). The pathway of water transfer through the cuticle is very high and is in parallel with the stomatal resistance (r_s). All of these resistances may be expressed as a conductance which is the reciprocal value of the resistance. The stomatal resistance or conductance is variable, depending on the degree of stomatal opening. Figure 2 shows that the whole leaf resistance for one surface of the leaf is r_l, with the total leaf resistance being the parallel sum for the upper and lower surfaces. The dominance of the stomata as the major sites for the transfer of water vapor from the inside of leaves to the outside is reinforced, because the rest of the leaf is covered with the cuticle which has a very high resistance to vapor transfer, unless the leaf surface (and therefore the cuticle) is damaged in some way. The low permeability of the cuticle is reflected in a very high cuticular resistance.

The boundary layer of a leaf refers to the air layer next to the leaf surface, where the surface friction reduces the wind speed in comparison to the bulk airflow. The air layer immediately adjacent to the leaf surface effectively remains static, and the transfer of molecules such as water vapor, occurs by diffusion only. In still air, the low boundary-layer conductance will have a major influence on the flux of water vapor out of a leaf unless the stomata are practically closed. In moving air, on the other hand, the exit of water vapor from the vicinity of the leaf is regulated more by the stomata (Figure 3).

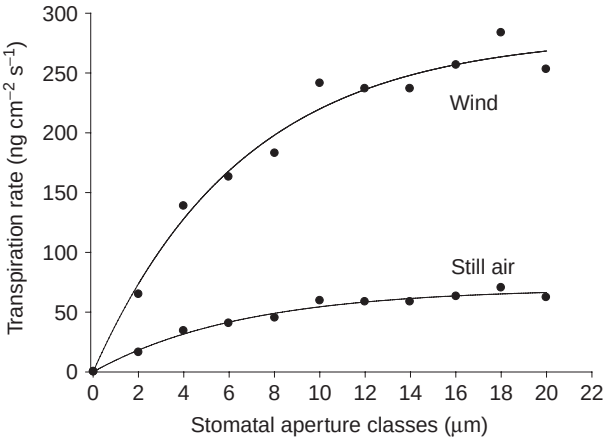


Figure 3 The regulation of transpiration by stomatal aperture in Zebrina pendula leaves under moving and still air conditions (Redrawn from Bange, 1953)

A number of factors, both environmental and within the plant, can influence stomatal aperture. Stomata open in response to increases in solar radiation and show an asymptotic relationship with radiation up to full midday illumination conditions. There is an increase in stomatal opening in response to temperature. Stomatal opening increases up to an optimum value of temperature which is species dependent but can be as much as 30 °C. In many species, particularly trees, stomata close in response to increases in the air atmospheric humidity deficit. The exact mechanism by which plants might sense humidity and reduce stomatal opening remains to be revealed. Nevertheless, strong negative relationships have been observed between stomatal conductance (g_s) and air humidity deficit (D) in a wide variety of vegetation types, particularly those including woody plants (see Figure 1). An important consequence of the negative relationship between g_s and D is that from day to day transpiration remains fairly similar, and never reaches the high values expected from considerations of radiation input. There is a high covariance between D, solar radiation, and air temperature. Therefore on bright, hot days with high D, stomata may be no more open than on a duller, cooler day with smaller D.

It could be postulated that the relationship between g_s and D in plants acts as a form of daily rationing of water, and might delay the development of severe soil water deficits which might lead to a more long-term state of stomatal closure. The reduction in the availability of soil moisture has a fundamental control of stomatal opening, usually operating through the leaf water status and turgidity of the leaf. Plants operate as integrated systems with signals to a leaf coming from all parts of the plant resulting in a net response. As well, the hydraulic signal to the foliage from the roots indicating the development of reduced availability of water in the soil chemical signals can also be transmitted. Stressed roots produce substances that are transported to the guard cells of the stomata via the transpiration stream. There is good evidence the plant hormone, abscisic acid (ABA), is an important hormonal factor which is transported from the roots (Davies and Zhang, 1991). ABA has a strong inhibitory influence on stomatal opening. There is still a need to research particular aspects of ABA and the control of stomata, and therefore transpiration. For example, up until now there has not been a convincing demonstration of the transport of ABA from the roots of mature trees into their canopies, during the development of soil water stress.

Much understanding about the controls of stomatal opening by environmental factors came from detailed studies involving individual plants in controlled conditions. In these studies, the convention was to modify one environmental factor whilst all others were held constant. In real field situations, especially in a complex canopy, the degree to which stomata are open depends on the interplay of a range of environmental conditions. Transpiration from an individual leaf will depend on (1) the degree of stomatal opening (stomatal conductance) determined by environmental conditions, (2) the leaf-boundary conductance which is usually strongly influenced by the size and shape of the leaf, and (3) environmental conditions that force transpiration particularly temperature, available energy, and air vapor pressure deficit. Wind speed around the leaf will have an influence on boundary-layer conductance.

At any one time during the day, the transpiration of a vegetation canopy reflects the environmental influences on the stomata of individual leaves, depending on their position in the canopy. In plant canopies, particularly the more complex ones, for example temperate and tropical forests, there is usually significant systematic variation in leaf stomatal and leaf-boundary conductance down through the canopy. This variation in the conductances coupled with systematic microclimatic variation also down towards the base of the canopy means a major influence on levels of transpiration from different zones in the canopy. Light levels decline with depth in plant canopies. These reduced light levels mean that stomatal opening is reduced, and also that less energy is available to force transpiration. Humidity levels are also usually higher within canopies. This means that the leaf to atmosphere vapor pressure gradient is reduced which limits transpiration. Wind speeds are also lower within canopies. This will mean that leaf boundary-layer conductances might become critically low in canopies where the air is less agitated.

The modification of microclimate within complex canopies means that although the leaf area index (leaf area per unit of ground area, $m^2\,m^{-2}$) might be high, that is, six or more, transpiration in the lower levels of the canopy might be very much reduced. Roberts *et al.*, (1996), working in the tropical rainforest in the central Amazon, Brazil,

examined the interaction of microclimate and leaf conductances in controlling transpiration using a five layer formulation canopy layer and total transpiration estimation routine (CLATTER) of the Monteith–Penman equation (*see* **Chapter 45, Actual Evaporation, Volume 1**). They showed that because of reduced light levels, leaf conductances fall systematically down through the forest canopy. Leaf boundary-layer conductances fell because of the reduced wind speeds towards the forest floor, and humidity deficits also declined systematically towards the forest floor. CLATTER showed that although there is a substantial fraction of the total leaf area index (forest LAI ~6) in the lower layers of the canopy, the fraction of the total forest transpiration from the lower layers is much less than layers in the upper canopy that contain a lower foliage density (Figure 4).

As well as the possession of stomata which can regulate water loss, land plants living in regions of water shortage avoid desiccation, and may possess a large number of structural and life history adaptations that are considered to contribute to conserving water. Some of these adaptations are as follows:

- highly reflective leaf surfaces brought about by wax deposits or reflective hairs;
- stomata sunk into pits in the leaf surfaces;
- reductions in leaf area per mass of plant;
- leaves reduced in size;
- deep roots, or a large mass of root per mass of shoot;
- modifications of stem or root to form water storage organs;
- ability to shed leaves during the driest periods to avoid water deficits;

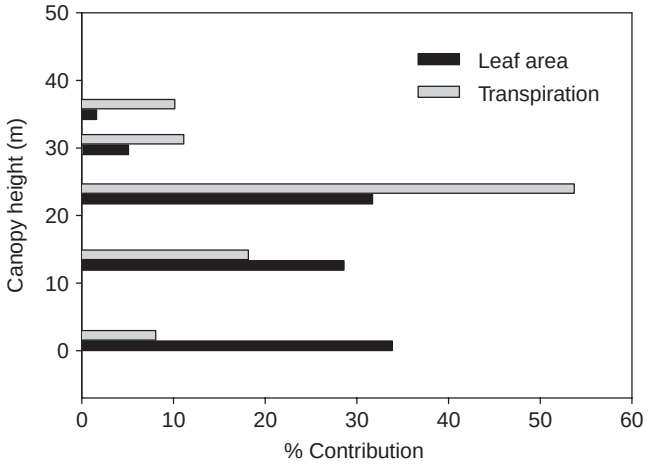


Figure 4 Leaf area in each of five canopy layers as a percentage of total leaf area and transpiration from each layer as a percentage of total transpiration. Reserva Florestal Ducke, Manaus (After Roberts *et al.,* 1996. Reproduced with permission from CEH)

- in dry zones there is a prevalence of annual plants where the plant survives dry periods as a seed, or if a perennial, the plant survives as a subterranean or otherwise much-reduced structure.

WHAT ARE THE ROLES FOR TRANSPIRATION ?

The foliage of plants with a large surface area relative to its volume, constantly exposed to sun and wind, will lose water very rapidly. This does not occur with leaves, because the entire surface is covered with a thin, waterproof layer, the cuticle. Although there may be some microscopic cracks and fissures in the cuticle, largely the cuticle provides a very effective means of restricting the loss of water from the leaf. However, the leaf must be able to take up the gas, carbon dioxide (CO_2), for photosynthesis, and to lose the excess oxygen (O_2) not required for respiration. The leaf must also take up oxygen for respiration when photosynthesis is not taking place, such as during the night and in low sunlight. The stomata in the leaf surfaces control both the entry of CO_2, and the exit of O_2. The diffusion outwards of water vapor is regarded by some as an unnecessary consequence of the stomata being open for the uptake of CO_2. Nevertheless, there are important roles for transpiration and the need for stomata to function as they do in the leaf surface.

Transpiration and Leaf Temperature

A key feature of transpiration for plants is the cooling effect that prevents leaf temperatures from exceeding lethal limits and to be maintained close to the optimum for the functioning of a range of physiological processes which occur inside leaves. Particularly in regions such as deserts, with high radiation inputs, the need for leaves to be cooled is of paramount importance. Figure 5 (Lange, 1959) shows that the temperatures of freely transpiring leaves of the desert cucumber *(Citrullus colocynthis)* are maintained at around or below air temperature, and well below lethal limits. When transpiration is interrupted, leaf temperature rises to beyond critical limits within the next hour.

Evaporation, and especially transpiration, from plant and vegetation surfaces, is particularly important to the land surface energy balance. Therefore, changes in vegetation cover which modify levels of radiant energy used to drive transpiration flux have a large influence on temperatures at the vegetation surface.

Transpiration and Nutrient Uptake

A crucial aspect of transpiration is the associated uptake of certain nutrient ions and their transport and distribution through the plant, to support plant growth and development.

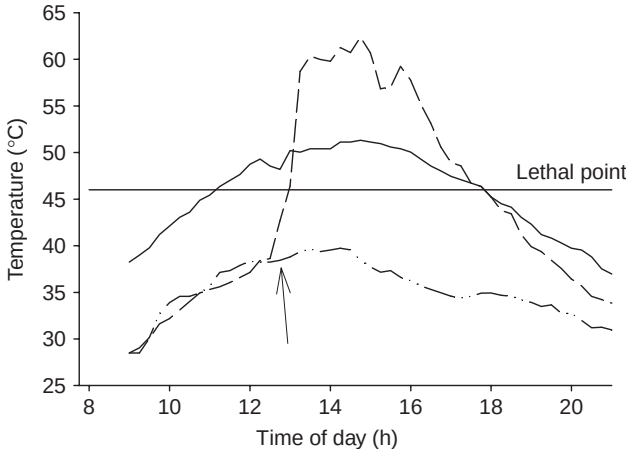


Figure 5 Transpirational cooling in the desert plant (*Citrullus colocynthis*). Graph shows air temperature (___), leaf temperature (-··- ··-) and temperature of another leaf which was cut off (arrow shows time of detachment) to stop transpiration (- - - -). (After Lange, 1959. Reproduced with permission from Elsevier GmbH)

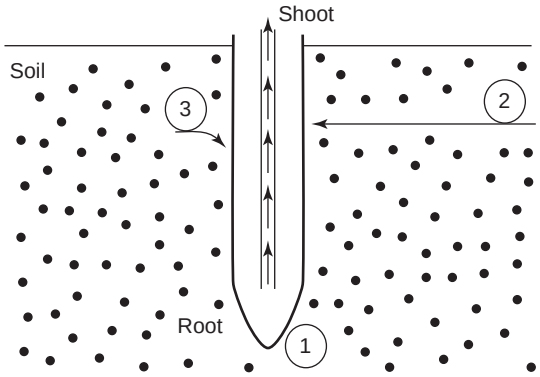


Figure 6 Schematic presentation of the movement of mineral elements to the root of a tree. (1) Root interception: soil volume displaced by root volume. (2) Mass flow: transport of bulk soil solution along the water potential gradient, driven by transpiration. (3) Diffusion nutrient transport along the concentration gradient. • = available nutrients. (After Marschner, 1995)

Barber (1995) discusses three mechanisms by which plants obtain nutrients from soil and the influence of soil moisture conditions and transpiration in these mechanisms. The three methods are shown schematically in Figure 6. All three methods may be in use at the same time but the uptake of certain nutrients may be exclusively by one method.

Root growth can mean that roots encounter unexploited zones of available nutrients, this is regarded as nutrient uptake by interception. Diffusion is the movement of certain nutrient ions along a concentration gradient, usually towards the root surface where concentration is lowest, from the bulk soil where the concentration is highest.

Root interception is used to describe the acquisition of soil nutrients at the interface of the soil and the root. These nutrients are acquired by virtue of the growth of roots in the soil. Because soil moisture status will determine the amount and distribution of root growth, the level of soil moisture will influence nutrient uptake by root interception. Nutrients move in soil in two ways. First, nutrients are transported to the roots by mass flow along with the water taken up as part of transpiration. Therefore, the uptake of nutrients will be directly influenced by the transpiration rate and the nutrient concentrations in the soil from where the water is taken up. Transpiration by the plants has a fundamental role in the mass flow of nutrients through and from the soil, and into the plants via the roots. Mass flow is the process by which plants usually acquire at least three of their most important nutrient requirements, nitrogen, calcium, and magnesium, and considerable nutrient quantities enter the vegetation as part of the water uptake process. It was estimated that $80\,\mathrm{kg\,ha^{-1}\,year^{-1}}$ of nitrogen is taken up by the broadleaf woodland on the Hubbard Brook catchment, USA (Likens and Bormann, 1995). Nutrient uptake will fluctuate with soil water availability. In many circumstances, nutrient availability in the topsoil declines steeply during the growing season because low soil moisture becomes a limiting factor for nutrient delivery to the root surface. The second process by which nutrients move through soil, is by diffusion along a concentration gradient. Rate of ion diffusion is directly related to the water content of the soil (θ). Increases in θ reduce the tortuosity of the diffusion path, and increases the diffusion flux. There is often a linear relation between θ and the diffusion coefficient. Diffusion is particularly important for nutrients such as potassium and phosphorus. Table 2 shows the relative importance of

Table 2 Nutrient demand of a maize crop and estimates of nutrient supply from the soil by root interception, mass flow, and diffusion (After Barber, 1995 by permission of John Wiley & Sons Inc.)

		Estimates of amounts (kg ha^{-1}) supplied by		
	Demand (kg ha^{-1})	Interception	Mass flow	Diffusion
Potassium	195	4	35	156
Nitrogen	190	2	150	38
Phosphorus	40	1	2	37
Magnesium	45	15	100	0

the different mechanisms of nutrient uptake in the case of several key elements.

HOW IS TRANSPIRATION MEASURED?

Because there are a wide range of questions relating to transpiration from plants and vegetation, a wide range of techniques have been used to measure the process. The most appropriate technique to be used depends largely on the temporal and spatial scale over which estimates are needed. There are approaches, because they are capable of measuring total transpiration from stands of vegetation, for example, micrometeorological or soil depletion studies, that are not able to provide detailed information about the contribution of individuals that comprise the vegetation. They are not appropriate for isolated individual plants. Nor can the contribution of different vertical layers in complex vegetation be separated by techniques that measure total vegetation transpiration. Sometimes detailed information is required about the contribution of fragments of the vegetation to total transpiration and the factors that control their transpiration. In this case, approaches such as porometry or sap flow will have to be used.

Soil Water Depletion

Studies of rates of soil water depletion are usually used to estimate transpiration over timescales of at least a few days, and they cannot distinguish water taken up by different species when they are growing in close proximity. Sufficient measurements of soil water content need to be made to account for spatial variation in water storage in the soil. In addition, the amount of drainage needs to be measured, or it must be insignificant. Provided drainage from and recharge to the soil is quantified or insignificant, changes in soil water storage allow evaporation to be calculated when rainfall is absent or measured separately. Soil moisture depletion techniques require repeated *in situ* measurements of soil water content, which may be with a neutron probe (Bell, 1987; Dean *et al.*, 1987), the impedance "ThetaProbe" technique (Gaskin and Miller, 1996), or by using time domain reflectometry (Topp *et al.*, 1980, 1984).

Micrometeorological Methods

A revealing approach to measuring evaporation, where it is possible to discriminate between evaporation from wet canopies and transpiration from dry canopies, exploits some of the micrometeorological methods that are available. These approaches are covered elsewhere in this volume (*see* **Chapter 40, Evaporation Measurement, Volume 1**). Following rain events, there will be a significant contribution to the vapor flux from partially wetted leaves and damp soil and litter surfaces. The values from micrometeorological approaches to determine transpiration flux under these circumstances need to be interpreted with care.

Lysimeters

A lysimeter is a device in which a volume of soil with associated vegetation is isolated hydrologically from the surrounding soil. Drainage is measured or is zero, and in the case of lysimeters that are weighed, changes in water storage are determined by weight difference. The application of weighing lysimeters to studies in mature or semimature trees has been very limited. A single large Douglas fir tree was installed in a weighing lysimeter by Fritschen *et al.* (1973). Reyenga *et al.* (1988) installed a lysimeter in a regenerating eucalyptus forest in New South Wales, Australia. Large potted trees capable of being weighed (up to 500 kg) have proved useful for calibrating other techniques such as isotopic tracers (e.g. Dugas *et al.*, 1993). Another form of lysimeter is the drainage lysimeter, which has been used very commonly in short crops but rarely in forests. Calder (1976) described a drainage lysimeter constructed in a Norway spruce plantation in mid-Wales, UK. The lysimeter was sealed at the base by a soil layer of impermeable clay. Data from the lysimeter associated with nearby net rainfall measurements enabled calculation of three separate years of transpiration loss (Calder, 1977).

The Cut-tree Technique

Excising the bases of plants, even large trees (Ladefoged, 1963) under water and measuring water uptake, can be used to estimate transpiration. However Roberts (1977, 1978), showed that removal of soil and root resistances can improve the leaf water status of cut-trees compared to controls, leading to differences in stomatal conductance and transpiration between normal trees, and those with excised roots. Nevertheless, the tree-cutting technique has proved useful in examining the water relations of mature trees (Roberts, 1977), and the amount of water stored in trees that can contribute to transpiration (Roberts, 1976). The technique has also proved particularly valuable as a means of calibrating other techniques such as isotopic tracers (Waring and Roberts, 1979). Because in the cut-tree technique water is drawn up through the tree in a natural way and conditions around the canopy are not modified, the method offers the best option for the calibration of sap flow techniques.

Sap Flow Techniques

Sap flow techniques provide a means of continuously monitoring rates of sap flow. Information about sapwood cross-sectional area of sampled trees, or the leaf area of the sampled tree in relation to the leaf area of the forest, enables transpiration to be estimated on a land area basis. The range of techniques for measuring sap flow and the limitations of different approaches have been reviewed by Swanson (1994) and Smith and Allen (1996).

HEAT-PULSE VELOCITY

The heat-pulse velocity (HPV) method determines rates of sap flow by determining the velocity of a short pulse of heat removed by the upward-moving sap stream. The technique is only really useful on woody stems, and the depth of sapwood must not be so deep that the sensor probe cannot sample it adequately.

Each set of heat-pulse probes consists of one heater probe and two sensor probes containing miniature thermistors. Typically, four sets of probes are installed at equal distances around the circumference of the stem. The heat-pulse technique is based on a compensation principle; the velocity of sap ascending the stem is determined by correcting the measured velocity of a heat-pulse for the dissipation of heat by conduction through the wood matrix. In practice, this is achieved by installing the sensor probes at unequal distances upstream and downstream of the heater probe. The upstream sensor is usually nearer the heater than the downstream one. Heat-pulse velocity (v_h) is calculated from

$$v_h = \frac{\chi_d - \chi_u}{2t_0} \tag{1}$$

where χ_d and χ_u are the distances between the heater and the upstream and downstream sensors, respectively, and t_0 is the time taken after the heat pulse for the temperature of the two sensors to become equal again.

STEM HEAT BALANCE

The stem heat balance (SHB) method can be used to measure sap flow in both woody and herbaceous stems, and these can be very small in diameter. The approach has been used on branches, small trees, and even roots (Smith *et al.*, 1997). A full description of a SHB gauge is given by Smith and Allen (1996). Heat is applied to the outside of the segment of stem enclosed by the heater, and the sap flow derived from the fluxes of heat into and out of the heated section. Sap flow (F) is related to the different heat losses from the stem section (Swanson, 1994) by

$$F = \frac{Q_h - Q_t - Q_v - Q_s}{C_s \Delta T} \text{g s}^{-1} \tag{2}$$

where Q_h is heater power, Q_v is vertical heat loss, Q_r is radial heat conduction, Q_s is heat storage, C_s is the heat capacity of the sap, and ΔT is the temperature difference between the top and bottom of the heated section. The stem sector heat balance method, as described by Cermák *et al.* (1984), requires stainless steel electrode plates to be inserted in a tree stem. If there is substantial variation in sap flux around a large tree trunk, installations would be made at more than one point.

THERMAL DISSIPATION

Granier (1985, 1987) proposed an alternative sap flow technique. Each probe consists of a pair of needles, which are inserted into the sapwood. The upper needle contains a heating probe and a thermocouple, which is referenced to a second needle inserted in the sapwood lower down in the stem. Continuous heating of the upper needle sets up a temperature difference (ΔT) between the two needles. ΔT is at a maximum when the sap flow is at a minimum and decreases as the sap flow increases. Granier (1985) found that for two conifer species and oak, volumetric sap flux density (u_v, $\text{m}^3\text{m}^{-2}\text{s}^{-1}$) is related to ΔT by the following relationship:

$$u_v = 0.000119\text{Z}^{1.231} \tag{3}$$

where

$$\text{Z} = \frac{\Delta T_0 - \Delta T}{\Delta T} \tag{4}$$

when ΔT_0 is the value of ΔT when there is no sap flow. The mass sap flow rate (F_m) is then

$$F_m = \rho_s u_v A_{sw} \tag{5}$$

where ρ_s is the sap density and A_{sw} is the sapwood cross-sectional area. Granier *et al.* (1990) suggest that the parameters in the equation above are not dependent on wood properties or tree species and that the technique may possibly be used without calibration. However, this possibility needs testing for a wider range of species than has been the case thus far because Granier type gauges are commercially available and now very widely used. A calibration rig in which heat dissipation can be measured in a large beech log while water is passed through at different rates is shown in Figure 7.

Porometers and Infrared Gas Analyzers (IRGAs)

Porometers enable measurements of stomatal conductance, g_s, of individual leaves to be measured *in situ*. An Infrared Gas Analyzer (IRGA) can also be used to determine CO_2 exchange from the leaf as well as g_s. Additional useful information that can be acquired or calculated are leaf transpiration rates, leaf temperatures, and the internal CO_2 concentration of the leaves. g_s determined with a porometer or IRGA gives the most detailed information in both temporal or spatial scales about the environmental and internal controls of g_s, and hence transpiration.

g_s measured with porometers and IRGAs has also been used to estimate transpiration from plant canopies. This involves multiplying g_s by the leaf area index to produce a surface or canopy conductance (g_c) $\cdot$ g_c can be used with a canopy boundary-layer conductance, g_a, to estimate transpiration using the Penman–Monteith equation

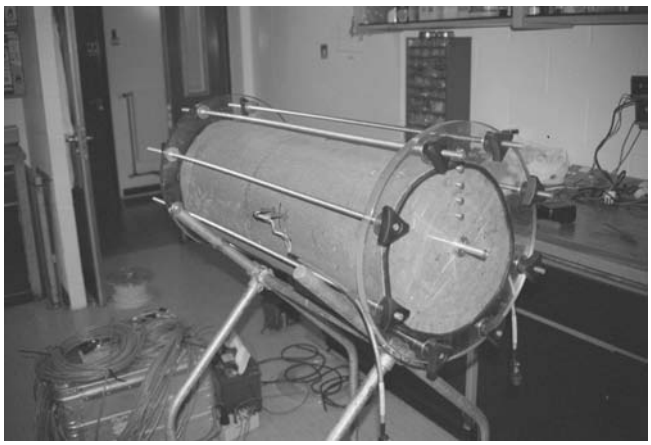


Figure 7 Calibration of the output from a thermal dissipation probe in a large beech (*Fagus sylvatica*) log. In the rig water can be passed through the log at different rates (Photo by John Roberts). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

(*see* **Chapter 45, Actual Evaporation, Volume 1**; Monteith, 1965). This type of approach has been used successfully even in very complex tropical rain forest canopy (Roberts *et al.*, 1993), although detailed information of the vertical distribution of leaf area density and the vertical changes in canopy microclimate are required, which imposes a severe logistical constraint.

Radioactive and Stable Isotope Tracers

A number of tracers have been used to measure transpiration from branches and individual trees. These values can then be scaled up to give stand transpiration. Waring and Roberts (1979) used P^{32} and tritium to measure the transpiration of Scots pine trees. Calder *et al.* (1992) described an approach using deuterium oxide (D_2O) which was injected into eucalyptus trees in plantations. Transpired water was collected in polythene bags tied on to selected branches. From the information of the D_2O injected (M), and the concentration of the transpirate produced over a known time interval ($C\,dt$), transpiration (F) can be calculated.

$$M = F \int_0^\infty C\,dt \tag{6}$$

The time resolution for tracing techniques is relatively low. Transpiration values can only be resolved over a few days. It is, therefore, difficult to use the techniques to understand the influence of short-term environmental fluctuations on transpiration.

SOURCES OF WATER FOR TRANSPIRATION

In the vast majority of cases for land plants, the source of water transpired by plants comes from water stored in the upper layers of soil. Measurements of drying patterns in the few meters of soil below vegetation usually show the largest reduction in soil moisture in the soil surface layer with reduced changes in deeper zones.

Jackson *et al.* (1996) reviewed the pattern of root mass and vertical distribution that had been previously reported for different terrestial biomes. The relationship proposed by Gale and Grigal (1987) was used by Jackson *et al.* to describe the vertical distribution of roots.

$$Y = 1 - \beta^d \tag{7}$$

Y is the cumulative fraction of root mass ($0 < Y < 1$) from the soil surface to a depth of d cm. β is a fitted parameter and has larger values for relatively deeply rooted vegetation. For example, β for temperate coniferous forests is 0.976, but 0.943 for temperate grassland.

The concentration of roots in the soil surface is probably related to a number of factors; the ready availability of plant nutrients from litter deposition and decomposition in the soil surface, and the ease with which surface roots can exploit small frequent inputs of rainfall from storms after soil drying has progressed from a saturated condition. Roots in the soil surface offer the shortest pathway, that is, the lowest resistance for water movement into the plant, and the association of a large population of fine roots that are associated with large structural roots that are in the soil surface for stability against wind blow.

Nevertheless, the small amount of deep roots (constituting a relatively small fraction of total root biomass) have crucial roles in acquiring water for growth and survival during drought. The tropical rainforest in the eastern part of the Amazon basin in Brazil remains evergreen even though the dry season may extend to 4 to 5 months. In this forest Nepstad *et al.* (1994), founds roots to a depth of 18 m. The water stored below 2 m in the soil and available to the trees provided more than 75% of the water extracted from the entire profile during the dry season.

HOW WILL GLOBAL CHANGES INFLUENCE TRANSPIRATION?

Vegetation affects the amount of evaporation from a catchment through transpiration and interception (*see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). Both vary with vegetation type, so a change in vegetation due to a change in climate would have an effect on (catchment) evaporation, with the effect of course depending on the extent of the change in vegetation. The rate of transpiration for a given plant, however, will also be influenced by climate change. Plants absorb CO_2 from the atmosphere as part of the process of photosynthesis, and this CO_2 is absorbed through stomata.

An increasing concentration of CO_2 in the atmosphere has two main (hydrologically relevant) effects on plants, which have been demonstrated through experiments at the leaf and plant scale. These influences are likely to have opposing effects on water use by vegetation. First, net photosynthesis increases with increasing carbon dioxide concentrations in the atmosphere. This is expected to increase plant growth and leaf area index. This should be expected to increase vegetation transpiration (e.g. Long, 1999).

On the other hand, a second effect of CO_2 enrichment is to make stomata smaller, which has the effect of reducing the stomatal conductance to evaporation. The largest reduction in stomatal conductance, as much as 40%, has been found in grasses and herbaceous species, with a lower reduction (~20%) being more typical in woody vegetation (Saxe *et al.*, 1998). The reductions in stomatal conductance lowers the rate of transpiration for a given set of meteorological inputs, and the water use efficiency (WUE: ratio of carbon uptake (i.e. biomass growth) to transpiration) of plants increases. This has frequently been taken to mean that increased CO_2 concentrations will lead to a reduction in catchment scale evaporation, but this is not necessarily so. In fact, many of the few studies that have looked at scales larger than the plant have shown that evaporation per unit area does not decrease as CO_2 concentrations rise, largely because the increased WUE is offset by additional plant growth. Kruijt *et al.* (1999), for example, estimated that increasing CO_2 concentrations would have little net effect on transpiration over forest. In a study using a dynamic vegetation model coupled with a climate model to simulate the effects of both CO_2 increase and associated climate change Betts *et al.* 2000, found that, at the global scale, the effects of CO_2 – enrichment and temperature – increase on plant growth offset reduced stomatal conductance. Also, catchments contain a mix of land uses – including bare soil – which will be affected differently by increasing concentrations of CO_2. The higher proportion of total evaporation derived from intercepted water, the less the direct effect of increasing CO_2 on catchment evaporation. There is clearly considerable uncertainty over the effect of CO_2 enrichment on plant transpiration (Amthor, 1995; Jarvis *et al.*, 1999), and hence catchment evaporation (Field *et al.*, 1995), and it is not possible to draw any general conclusions.

REFERENCES

Amthor J.S. (1995) Terrestrial higher-plant response to increasing atmospheric [CO_2] in relation to the global carbon cycle. *Global Change Biology*, **1**, 243–274.

Bange G.G.J. (1953) On the quantitative explanation of stomatal transpiration. *Acta Botanica Neerlandica*, **2**, 255–297.

Barber S.A. (1995) *Soil Nutrient Bioavailability: A Mechanistic Approach*, *Second Edition*, John Wiley: New York, p. 413.

Beadle C.L., Neilson R.E., Talbot H. and Jarvis P.G. (1985) Stomatal conductance and photosynthesis in a mature Scots pine forest. I. Diurnal, seasonal and spatial variation in shoots. *Journal of Applied Ecology*, **22**, 557–571.

Bell J.P. (1987) *Neutron Probe Practice*, IH Report No. 19, Institute of Hydrology: Wallingford.

Betts R.A., Cox P.M. and Woodward F.I. (2000) Simulated responses of potential vegetation to doubled CO2 climate change and feedbacks on near-surface temperature. *Global Ecology and Biogeography*, **9**, 171–180.

Calder I.R. (1976) The measurement of water losses from a forested area using a 'natural' lysimeter. *Journal of Hydrology*, **30**, 311–325.

Calder I.R. (1977) A model of transpiration and interception loss from a spruce forest in plynlimon, central wales. *Journal of Hydrology*, **33**, 247–275.

Calder I.R., Kariyappa G.S., Srinivasalu N.V. and Srinivasa Murthy K.V. (1992) Deuterium tracing for the estimation of transpiration from trees. 1. Field calibration. *Journal of Hydrology*, **130**, 17–25.

Cermák J., Jeník J., Kucera J. and Zídek V. (1984) Xylem water flow in a crack willow tree (*Salix fragilis* L.) in relation to diurnal changes of environment. *Oecologia*, **64**, 145–151.

David T.S., Ferreira M.I., Cohen S., Pereira J.S. and David J.S. (2004) Constraints on transpiration from an evergreen oak tree in southern Portugal. *Agricultural and Forest Meteorology*, **122**, 193–205.

Davies W.J. and Zhang J. (1991) Root signals and the regulation of growth and development of plants in drying soil. *Annual Review of Plant Physiology and Plant Molecular Biology*, **42**, 55–76.

Dean T.J., Bell J.P. and Baty A.J.B. (1987) Soil moisture measurement with an improved capacitance technique. Part 1: sensor design and performance. *Journal of Hydrology*, **93**, 67–78.

Dugas W.A., Wallace J.S., Allen S.J. and Roberts J.M. (1993) Heat balance, porometer and deuterium measurements of transpiration from eucalyptus and prunus trees. *Agricultural and Forest Meteorology*, **64**, 47–62.

Field C.B., Jackson R.B. and Mooney H.A. (1995) Stomatal responses to increased CO_2: implications from the plant to the global scale. *Plant, Cell and Environment*, **18**, 1214–1225.

Fritschen L.J., Cox L. and Kinerson R.S. (1973) A 28 meter Douglas fir tree in a weighing lysimeter. *Forest Science*, **19**, 43–55.

Gale M.R. and Grigal D.F. (1987) Vertical root distributions of northern tree species in relation to successional status. *Canadian Journal of Forest Research*, **17**, 829–834.

Gash J.H.C. and Stewart J.B. (1977) The evaporation from Thetford forest during 1975. *Journal of Hydrology*, **35**, 385–396.

Gaskin G.J. and Miller J.D. (1996) Measurement of soil water content using a simplified impedance measuring technique. *Journal of Agricultural Engineering Research*, **63**, 153–160.

Granier A. (1985) Une nouvelle méthode pour le mesure de flux de sève brute dans le tronc des arbres. *Annales des Sciences Forestieres*, **42**, 193–200.

Granier A. (1987) Evaluation of transpiration in a Douglas fir stand by means of sap flow measurements. *Tree Physiology*, **3**, 309–320.

Granier A., Bobay V., Gash J.H.C., Gelpe J., Saugier B. and Shuttleworth W.J. (1990) Vapour flux density and transpiration rate comparisons in a stand of maritime pine (*Pinus pinaster* Ait.) in Les Landes forest. *Agricultural and Forest Meteorology*, **51**, 309–319.

Jackson R.B., Canadell J., Ehleringer J.R., Mooney H.A., Sala O.E. and Schulze E.-D (1996) A global analysis of root distributions for terrestrial biomes. *Oecologia*, **108**, 389–411.

Jarvis A.J., Mansfield T.A. and Davies W.J. (1999) Stomatal behaviour, photosynthesis and transpiration under rising CO_2. *Plant, Cell and Environment*, **22**, 639–648.

Jones H.G. (1992) *Plants and Microclimate*, *Second Edition*, Cambridge University Press: Cambridge.

Kruijt B., Barton C., Rey A. and Jarvis P.G. (1999) The sensitivity of stand-scale photosynthesis and transpiration to changes in atmospheric CO_2 concentration and climate. *Hydrology and Earth System Sciences*, **3**, 55–69.

Ladefoged K. (1963) Transpiration of trees in closed stands. *Physiologia Pantarum*, **16**, 378–414.

Lange O.L. (1959) Untersuchungen uber warmehaushalt und hitzeresistenz mauretanischer wusten und savannapflanzen. *Flora (Jena)*, **147**, 595–651.

Likens G.E. and Bormann F.H. (1995) *Biogeochemistry of a Forested Watershed*, *Second Edition*, Springer Verlag: Berlin, p. 159.

Long S.P. (1999) Understanding the impacts of rising CO_2: the contribution of environmental physiology. In *Physiological Plant Ecology*, PressM.C., Scholes J.D. and Barker M.G. (Eds.), Blackwell Science: Oxford, pp. 263–282.

Marschner H. (1995) *Mineral Nutrition of Higher Plants*. Academic Press: London, p. 889.

Monteith J.L. (1965) Evaporation and environment. *Symposium of the Society for Experimental Biology*, **19**, 205–234.

Nepstad D.C., de Carvalho C.R., Davidson E.A., Jipp P.H., Lefebvre P.A., Negreiros G.H., da Silva E.S., Stone T.A., Trumbore S.E. and Vieira S. (1994) The role of deep roots in the hydrological and carbon cycles of amazonian forests and pastures. *Nature*, **372**, 666–669.

Reyenga W., Dunin F.X., Bautovich B.C., Rath C.R. and Hulse L.B. (1988) A weighing lysimeter in a regenerating eucalypt forest: design, construction and performance. *Hydrological Processes*, **2**, 301–314.

Roberts J.M. (1976) An examination of the quantity of water stored in mature *Pinus sylvestris* L. trees. *Journal of Experimental Botany*, **27**, 473–479.

Roberts J.M. (1977) The use of tree cutting techniques in the study of the water relations of mature *pinus sylvestris* L. 1. The technique and survey of the results. *Journal of Experimental Botany*, **28**, 751–767.

Roberts J.M. (1978) The use of tree cutting technique in the study of the water relations of Norway spruce *Picea abies* (L.) Karst. *Journal of Experimental Botany*, **29**, 465–471.

Roberts J.M. (1983) Forest transpiration: a conservative hydrological process? *Journal of Hydrology*, **66**, 133–141.

Roberts J.M., Cabral O.M.R., Fisch G., Molion L.C.B., Moore C.J. and Shuttleworth W.J. (1993) Transpiration from an amazonian rainforest calculated from stomatal conductance measurements. *Agricultural and Forest Meteorology*, **65**, 175–196.

Roberts J.M., Cabral O.M.R., McWilliam A.-L.C., da Costa J.P. and Sá T.D.de A. (1996) An overview of the leaf area index and physiological measurements during ABRACOS. In *Amazonian Deforestation and Climate*, Gash J.H.C., Nobre C.A., Roberts J.M. and Victoria R.L. (Eds.), John Wiley: Chichester, pp. 287–306.

Saugier B., Granier A., Pontailler J.Y., Dufrêne E. and Baldocchi D.D. (1997) Transpiration of a boreal pine forest measured by branch bag, sap flow and micrometeorological methods. *Tree Physiology*, **17**, 511–519.

Saxe H., Ellsworth D.S. and Heath J. (1998) Tree and forest functioning in an enriched CO_2 atmosphere. *New Phytologist*, **139**, 395–436.

Shuttleworth W.J. (1988) Evaporation from Amazonian rainforest. *Proceedings of the Royal Society London*, **B233**, 321–346.

Smith D.M. and Allen S.J. (1996) Measurement of sap flow in plant stems. *Journal of Experimental Botany*, **47**, 1833–1844.

Smith D.M., Jackson N.A. and Roberts J.M. (1997) A new direction in hydraulic lift: can tree roots siphon water downwards? *Agroforestry Forum*, **8**(1), 23–26.

Swanson R.H. (1994) Significant historical developments in thermal methods for measuring sap flow in trees. *Agricultural and Forest Meteorology*, **72**, 113–132.

Topp G.C., Davis J.L. and Annan A.P. (1980) Electromagnetic determination of soil water content; measurement in coaxial transmission lines. *Water Resources Research*, **16**, 574–582.

Topp G.C., Davis J.L., Bailey W.G. and Zebchuk W.D. (1984) The measurement of soil water content using a portable TDR probe. *Canadian Journal of Soil Science*, **64**, 313–321.

Waring R.H. and Roberts J.M. (1979) Estimating water flux through stems of Scots pine with tritiated water and phosphorus-32. *Journal of Experimental Botany*, **30**, 459–471.

Willmer C. and Fricker M. (1996) *Stomata*, *Second Edition*, Chapman & Hall: London, p. 375.

43: Evaporation of Intercepted Rainfall

JORGE SOARES DAVID[1], FERNANDA VALENTE[1] AND JOHN HC GASH[2]

[1] *Instituto Superior de Agronomia, Tapada da Ajuda, Lisboa, Portugal*
[2] *Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK*

When rain falls on to a vegetated surface (gross rainfall), a part is intercepted by the canopy and evaporated directly back into the atmosphere (interception loss). The remainder of the rainfall reaches the ground (net rainfall) either through gaps and by dripping from the canopy (throughfall), or by running down the main stem (stemflow). The canopy storage capacity is the minimum amount of water necessary to completely saturate the canopy surface. Interception loss is conventionally measured as the difference between the incident gross rainfall, and the sum of throughfall and stemflow. It is usually a significant component of the overall evaporation and may play an important role in watershed water balance. The total amount of interception loss depends on the rate of evaporation from the wet canopy, the canopy storage capacity, and the distribution and intensity of rainfall. Interception loss is particularly high in forests due to their high aerodynamic roughnesses. Available models range from empirical relationships to physically based conceptual models. Within the latter, the Rutter-type models are those that have been taken up as the more generally applicable technique. Most of the models rely on the use of the Penman–Monteith equation to estimate the evaporation rate. Current research is largely directed at modeling in horizontally heterogeneous vegetation, where the Penman–Monteith equation may not be valid.

INTRODUCTION

When rain falls on to a vegetated surface (gross rainfall, P_g), a part is intercepted by the canopy and evaporated directly back into the atmosphere. This is termed the *interception loss* (I). The remainder of the rainfall reaches the ground surface either directly through gaps in the canopy and by dripping from the canopy (throughfall, T_f), or by running down the main stem or trunk (stemflow, S_f). Figure 1 shows these main components of the interception process. The actual amount of water that reaches the ground is termed *net precipitation*, and is the sum of throughfall and stemflow. Unless there is fog or cloud interception (*see* **Chapter 38, Fog as a Hydrologic Input, Volume 1**), the net rainfall is less than the incident gross rainfall, which falls on the top of the canopy.

The concept of a canopy storage capacity was first introduced by Horton (1919). Water is retained on the surface of leaves, twigs, and branches by surface tension until the surface tension forces are in balance with the gravitational forces (Leonard, 1967). The canopy then reaches its storage capacity (S) and dripping starts (canopy drainage). S is then defined as the minimum amount of water necessary to completely cover the canopy surface in still air.

The interception process may play an important role both in the water balance of watersheds and in the ecology of the vegetation. Interception occurs with all vegetation types: a forest canopy, understorey or litter layer, or for crops and other short vegetation (Zinke, 1967; Calder, 1990; Ward and Robinson, 2000). However most research, both theoretical and experimental, has been focused on rainfall interception by forests.

The interception loss from forests is usually a significant component (25%–75%) of overall evaporation (Gash and Stewart, 1977; McNaughton and Jarvis, 1983; Shuttleworth, 1988). Expressed as a percentage of gross rainfall, interception loss may vary from 9% in Amazonia (Lloyd *et al.*, 1988) to values as high as 60% in *Picea sitchensis* and *Picea abies* stands in Brittany (Forgeard *et al.*, 1980). This wide variation is mainly due to differences in time distribution patterns of rainfall between regions. The largest

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

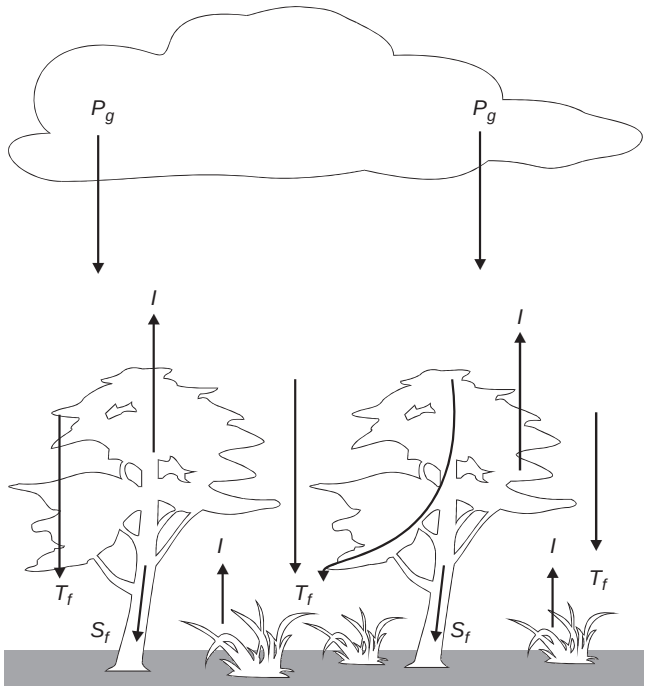


Figure 1 The main components of the rainfall interception process: Gross rainfall (P_g), Interception loss (I), Throughfall (T_f), and Stemflow (S_f)

evaporation losses occur in climates with a well-distributed rainfall of frequent, small storms; rather than in those with a few large ones (Rutter, 1975).

The redistribution of rainfall at the ground surface may also be quite important for the ecology and water relations of the vegetation. Typically, throughfall has a high spatial variability with dripping concentrated mainly at the edges of trees crowns (Eschner, 1967; Shuttleworth, 1988; Lloyd and Marques, 1988; Ward and Robinson, 2000). Additionally, and in spite of being a minor component of net rainfall, stemflow can concentrate a significant amount of water in the soil near the stem base (Eschner, 1967; Rutter, 1975; Herwitz, 1982). The ecological importance of this nonuniform distribution of net rainfall is further increased considering that this incoming water is enriched in nutrients (Chamberlain, 1975; Miller, 1984). The spatial variability of throughfall is particularly acute, and clearly nonrandom, under forests with prominent trees (Herwitz and Slye, 1995), at forest edges (Darnhofer *et al.*, 1989) and under isolated trees (King and Harrison, 1998; Gómez *et al.*, 2002), particularly during rainfall events with high wind speeds.

MEASURING INTERCEPTION

Conventionally, interception loss is measured as the difference between the incident gross rainfall, and the sum of throughfall and stemflow (net rainfall) measured at ground-level. Gross rainfall may be measured above the forest canopy or in a nearby clearing. Measurement of gross rainfall in forest clearings may be prone to errors because of the possible variability of rainfall in space, particularly during convective storms. Gross rainfall is probably best measured above the forest canopy in the same location as the net rainfall measurements. However, measurements of rainfall above the canopy are liable to undercatch as a result of turbulence around the gauge. According to Roberts *et al.* (2004), these errors are minimized if measurements are made just above the canopy.

To overcome the great spatial variability of throughfall, it is usually measured by intensive sampling using an array of funnel or trough gauges, placed randomly beneath the forest canopy. Typically, 20 to 30 funnel gauges would be needed to sample an approximately 1500-m^2 forest plot, with 50 to 90 trees (e.g. Valente *et al.*, 1997). A better long-term statistical sample is obtained by moving the gauges to new random positions after each reading. In heterogeneous forest, such as tropical forests, throughfall variation in space is even greater, and particularly large gauging samples are then needed (Lloyd and Marques, 1988; Bruijnzeel, 1990; Loescher *et al.*, 2002).

Stemflow is usually measured using gutters sealed around the trunk leading to a collecting/measuring device. Stemflow is also very variable from tree to tree. The sampling problems referred to above for throughfall also apply to stemflow, but are not so acute because stemflow is usually a small component of the canopy water balance (e.g. Lloyd *et al.*, 1988; Hutjes *et al.*, 1990; Valente *et al.*, 1997).

Some workers (e.g. Calder and Rosier, 1976; Calder *et al.*, 1986, 1996) prefer to use large plastic sheet gauges to measure the net rainfall. These gauges collect 100% of both throughfall and stemflow over an area of approximately 20 m^2. Although this method overcomes the problem of sampling small-scale variability, it misses any information on the variation of net rainfall in space. Also, while sheet gauges seem to work well in plantation forests, they are problematic when there is a dense understorey, in the case of widely spaced trees (Ward and Robinson, 2000) or in heterogeneous tropical forest (Roberts *et al.*, 2004). Furthermore, it can be difficult to maintain the plastic sheets free from leaks frequently caused by insects and rodents.

As aforementioned, interception loss is conventionally indirectly estimated as the difference between gross rainfall and net rainfall, that is, a small difference between two large numbers. Therefore, both gross rainfall and net rainfall need to be measured with precision to provide an accurate assessment of the interception loss.

Alternative, more direct methods of measuring the evaporation from wet canopies have been attempted, such as the micrometeorological Bowen ratio method (Stewart,

1977) (*see* **Chapter 40, Evaporation Measurement, Volume 1**). Techniques have also been developed to measure the change of water stored on the canopy, either by weighing cut branches (Rutter, 1967) or trees (Teklehaimanot and Jarvis, 1991), by applying straingauges to branches *in situ* (Hancock and Crowther, 1979), or through γ-ray (Calder and Wright, 1986) and microwave (Bouten *et al.*, 1991) attenuation. However, none of these methods have been accepted as a generally applicable technique. More recently, Mizutani *et al.* (1997) and Gash *et al.* (1999) have shown that fast-response ultrasonic anemometers, with good water shedding properties and spike-suppressing software, can function normally during rainfall. The eddy correlation method (*see* **Chapter 40, Evaporation Measurement, Volume 1**) can thus be used to measure evaporation from wet canopies in the turbulent boundary layer above the vegetation. Gash *et al.* (1999) have used this method to measure sensible heat-flux and then derive the evaporation from a wet pine canopy as the residual in the energy balance equation.

FACTORS AFFECTING INTERCEPTION

The total amount of interception loss from a particular vegetation cover depends on the rate of evaporation from the wet canopy as well as on the duration of canopy wetness (Rutter, 1975; Ward and Robinson, 2000). The latter is influenced both by the canopy storage capacity and by the distribution and intensity of rainfall.

Evaporation Rate from Wet Canopies

The most realistic description of evaporation from canopies is given by the Penman–Monteith equation (Monteith, 1965). When the canopy is dry, the flux of water vapor from the canopy surface (transpiration, λE_T) can be expressed (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**, **Chapter 45, Actual Evaporation, Volume 1**) as:

$$\lambda E_T = \frac{sA + \rho\, c_p D/r_a}{s + \gamma(1 + r_c/r_a)} \tag{1}$$

where E_T is the transpiration rate, A the available radiative energy, c_p the specific heat capacity of air at constant pressure, D the vapor pressure deficit, γ the psychrometric constant, s the slope of the saturation vapor pressure curve, λ the latent heat of vaporization of water, and ρ the density of air. r_a is the aerodynamic resistance that constitutes the control of vapor transfer from the foliage surfaces into the atmosphere and r_c is the surface (or canopy) resistance that represents the combined influence of stomata from all leaves.

When the canopy is wetted by rain, evaporation of intercepted rainfall is largely a physical process that does not depend on the functioning of stomata. The evaporation rate from a wet canopy (interception loss, λE_I) may therefore be expressed by equation (1), with $r_c = 0$:

$$\lambda E_I = \frac{sA + \rho\, c_p D/r_a}{s + \gamma} \tag{2}$$

Thus, for a given temperature, the rate of evaporation of intercepted water mainly depends on: the available energy (A), the vapor pressure deficit of surrounding air (D), and the aerodynamic resistance (r_a). The aerodynamic resistance to water vapor transfer is a function of the roughness characteristics of the evaporative surface and of wind speed (Monteith and Unsworth, 1990). With increasing height of vegetation, there is usually an increase in roughness length (z_0) and therefore, a decrease in the aerodynamic resistance (Rutter, 1975).

For many years there was a controversy between some authors who argued that evaporation from wet canopies was a net loss for the water balance and others who supported the view that it was just balanced by a corresponding reduction in transpiration during rainy periods (Ward and Robinson, 2000). These contradictory arguments can be easily clarified by defining a "relative interception rate", dividing equations (2 and 1) (Monteith, 1965; Rutter, 1975):

$$\frac{E_I}{E_T} = 1 + \frac{\gamma}{s + \gamma}\left(\frac{r_c}{r_a}\right) \tag{3}$$

The graph of the relationship between E_I/E_T and r_c/r_a is shown in Figure 2. From this it can be seen that when r_c and r_a are of similar size ($r_c/r_a \approx 1$), as happens with short vegetation, the rate of evaporation of intercepted water will be about the same as the transpiration rate would be in similar weather conditions. However, when r_a is an order of magnitude lower than r_c ($r_c/r_a \approx 10$), as it is for forests (Stewart and Thom, 1973; Thom, 1975; McNaughton and Jarvis, 1983), intercepted water evaporates at 3 to 5 times the transpiration rate (Stewart and Thom, 1973; Rutter,

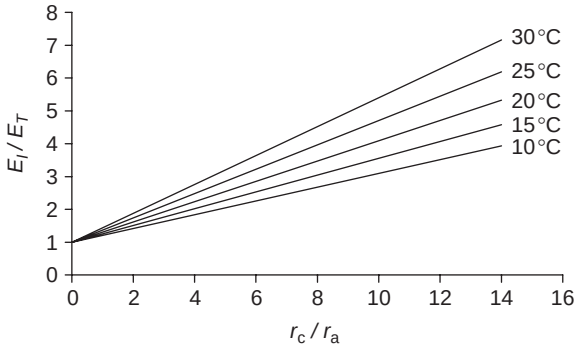


Figure 2 Change of relative interception rate (E_I/E_T) with the ratio r_c/r_a and temperature

1975). The evaporation from wet canopies is thus clearly a net water loss for forests, but not so for short vegetation. This explains the use of the term "interception loss" for forest ecosystems.

The main reason for the high evaporation rate of intercepted rainfall from forests is their high aerodynamic roughnesses and, therefore, correspondingly very low values of r_a. An additional consequence for aerodynamically rough vegetation, such as forests, is that the aerodynamic (right-hand) term of equation (2) is often much larger than the radiation term (Calder, 1990). Under these circumstances, it is the supply of advected energy, via the cooling of the above air mass that is able to support the high rates of evaporation from wet forests (Stewart, 1977; Pearce *et al.*, 1980; Calder, 1990). High evaporation rates are possible even at night when net radiation is close to zero (Pearce *et al.*, 1980; Calder, 1990). Dry-air advection, heat added to the air or removal of water from it may be responsible for the maintenance of a significant air vapor pressure deficit during rainfall events (McNaughton and Jarvis, 1983).

Reviewing the existing literature, Roberts *et al.* (2004) found that the average evaporation rate for wet forest canopies is a very conservative parameter. Values from forests in Great Britain (Gash *et al.*, 1980), in France (Gash *et al.*, 1995), in Portugal (Valente *et al.*, 1997), in Amazonia (Lloyd *et al.*, 1988), in West Africa (Hutjes *et al.*, 1990), and in West Java (Calder *et al.*, 1986) are remarkably similar (varying between 0.17 and 0.34 mm h^{-1}). As such, it would seem that the driving meteorological variables during rainfall usually do not vary much from tropical to temperate latitude forests: radiation is low, vapor pressure deficit is small, and wind speeds are similar.

Canopy Storage Capacity

The canopy storage capacity (S) is usually considered as a constant parameter for a given vegetation cover. However, some authors consider that it may vary with drop size and therefore with rainfall intensity (Calder, 1986; Hall, 1992).

Canopy storage may be evaluated both by indirect or direct methods. Leyton *et al.* (1967) developed the most commonly applied indirect method, whereby S is estimated as the negative intercept of the envelope line between throughfall and gross rainfall, with a preestablished slope (close to 1). By this method, S is considered equal to the total interception loss for those events when the evaporation during rain is negligible.

The alternative direct methods usually rely on weighing after artificial wetting (Aston, 1979; Herwitz, 1985; Teklehaimanot and Jarvis, 1991; Liu, 1998; Llorens and Gallart, 2000) or on γ-ray (Calder and Wright, 1986) and microwave (Bouten *et al.*, 1991; Klaassen *et al.*, 1998) attenuation.

Canopy storage values can range from 0.2 to 3.8 mm depending on species, leaf area index, canopy cover, vegetation structure, and density (Table 1). There is no evidence of a systematic difference between forest and short vegetation. Within forests, S tends to be higher in conifers than in broadleaves. Canopy storage is particularly low in deciduous forests during the leafless (winter) period.

Rainfall Distribution and Intensity

The interception loss is usually greater from intermittent than from continuous rain, when the intensity and total duration are the same (Rutter, 1975). This is due to the significance of the frequency of rewetting cycles, which

Table 1 Typical values of canopy storage capacity (S) for some vegetation types

Vegetation	S(mm)	References
Coniferous		
Pseudotsuga menziesii	2.1–3.8	Rutter (1975); Aussenac and Boulangeat (1980)
Picea spp.	1.0–3.1	Gash *et al.* (1980); Aussenac (1968)
Pinus spp.	0.4–3.0	Valente *et al.* (1997); Aussenac (1968)
Evergreen broadleaves		
Nothofagus spp.	1.0–1.2	Pearce and Rowe (1981); Rowe (1983)
Eucalyptus spp.	0.2–0.8	Valente *et al.* (1997); Aston (1979)
Tropical forest (Amazonia)	0.9–1.4	Lloyd *et al.* (1988); Ubarana (1996)
Deciduous		
Quercus spp.	summer 0.8–1.0 winter 0.3–0.4	Dolman (1987); Rutter (1975)
Carpinus betulus	summer 1.0 winter 0.6	Leyton *et al.* (1967)
Shrubs		
Calluna vulgaris	2.0	Leyton *et al.* (1967)
Herbaceous		
Lolium perenne	0.5–2.8	Marriam (1961), cited by Rutter (1975)
Mixed grasses and legumes	1.0–1.2	Burgy and Pomeroy (1958), cited by Rutter (1975)

seems more important than either the duration or the amount of rainfall (Ward and Robinson, 2000). In some particular situations, the rate of interception loss may be limited by rainfall intensity when the latter is lower than the potential evaporative rate (Rutter, 1975).

It is helpful to consider the evaporation of intercepted water on a rainfall event basis, assuming that each individual storm is preceded by a rainless period long enough to allow the canopy and stems to dry out completely. Making this simplifying assumption, each rainfall event may be divided into three main distinct phases (Gash, 1979): (1) a wetting phase, from the onset of rain until the canopy is saturated; (2) a saturation phase, and (3) a drying phase, lasting from the end of rainfall until the canopy and trunks are completely dry. The total interception loss during the storm can then be considered as the sum of total evaporation loss during phases (1) and (2) plus the evaporation during phase (3) which is equal to the canopy storage (S). From this, it is clear that (1) interception loss of long lasting rainfall events will mainly depend on the rate of evaporation, and (2) interception loss of short-duration events will be mainly determined by the canopy storage. These two extreme situations are typified by the contrasting climatic conditions of New Zealand (Pearce and Rowe, 1981; Rowe, 1983) and tropical rainforests (Lloyd *et al.*, 1988; Herwitz, 1985), respectively. Temperate climatic conditions will lie between these two extremes.

MODELING INTERCEPTION

Interception models allow experimental results to be extrapolated both in time and in space, and they are required as a component of hydrological and meteorological forecasting models. However, models can also provide insight into the functioning of the process.

Despite the pioneering work of Horton (1919), who pointed out that rainfall duration was the key variable, most of the early modeling studies attempted to express the interception loss in relation to gross rainfall, either as a percentage or through simple empirical relationships (Zinke, 1967). However, these relationships are site specific and extrapolation to areas with different rainfall regimes is likely to result in erroneous predictions.

Rutter *et al.* (1971, 1975) were the first to model forest rainfall interception recognizing that the process was primarily driven by evaporation from the wetted canopy. The Rutter model calculates a running water balance of the water stored on the canopy and trunks, relating it to rates of rainfall, drainage, and evaporation. It requires both above canopy rainfall and meteorological data (in short time-steps) as well as four parameters describing the structure of vegetation (among which are the canopy storage capacity and the trunk storage capacity). Trunk storage (S_t) is usually very small compared to canopy storage (S). The evaporation is calculated from the Penman–Monteith equation (2). The main practical problem is the high data requirements of the model. Gash (1979) proposed a simpler storm-based model, known as *Gash's analytical model*. This model incorporates some simple features of the linear regression models, but keeps the fundamental physical background of the Rutter model. The analytical model is less demanding in driving data. It requires the same four structural parameters, as the Rutter model, but applies a single constant value for both the average rainfall intensity and the average evaporation rate during the whole simulation period. This model has been used with considerable success in a wide range of conditions, from temperate forests (Gash *et al.*, 1980; Pearce and Rowe, 1981; Dolman, 1987) to tropical rainforests (Lloyd *et al.*, 1988; Hutjes *et al.*, 1990).

However, when both the Rutter model and Gash's analytical model were used in sparse forests, they were found to overestimate the interception loss (Lankreijer *et al.*, 1993; Gash *et al.*, 1995; Valente *et al.*, 1997). This led to reformulations of both Gash's model (Gash *et al.*, 1995) and the Rutter model (Valente *et al.*, 1997) to take into account the forest sparseness and to improve boundary conditions. In these sparse canopy versions of the models, the wet canopy evaporation is considered linearly dependent on the canopy cover fraction (c). Figure 3 shows the conceptual framework of the sparse forest Rutter model. The reformulated versions of these models are able to simulate well the interception loss both in sparse and closed canopies since, when canopy cover tends to 100%, the sparse model versions reduce to the original ones. Recent applications of the revised Gash's model confirm that model predictions are good in a wide range of conditions, from closed to very sparse canopies (Gash *et al.*, 1995; Valente *et al.*, 1997; Dykes, 1997; Asdak *et al.*, 1998; Carlyle-Moses and Price, 1999; Jackson, 2000). Van Dijk and Bruijnzeel (2001a, 2001b) have further modified the revised version of Gash's model to allow it to be applied to rapidly growing vegetation where the leaf area index is changing through time.

As an alternative to these Rutter-type models, Calder (1986) developed a stochastic model in which the canopy storage varies with drop size, and therefore with rainfall intensity. In a subsequent two-layer version of the stochastic model (Calder, 1996), the size of the primary raindrops is related to rainfall intensity, but the size of the secondary drops, dripping from the upper to the lower levels, is species-dependent. However, at present only the Rutter and Gash models or their derivatives have been taken up as a generally applicable technique (Roberts *et al.*, 2004).

Although the controversy over whether the canopy storage capacity should be treated as a variable or a fixed parameter (Roberts *et al.*, 2004) has not been solved, the above-mentioned models offer a range of reliable modeling tools of different complexity. However, all these models are

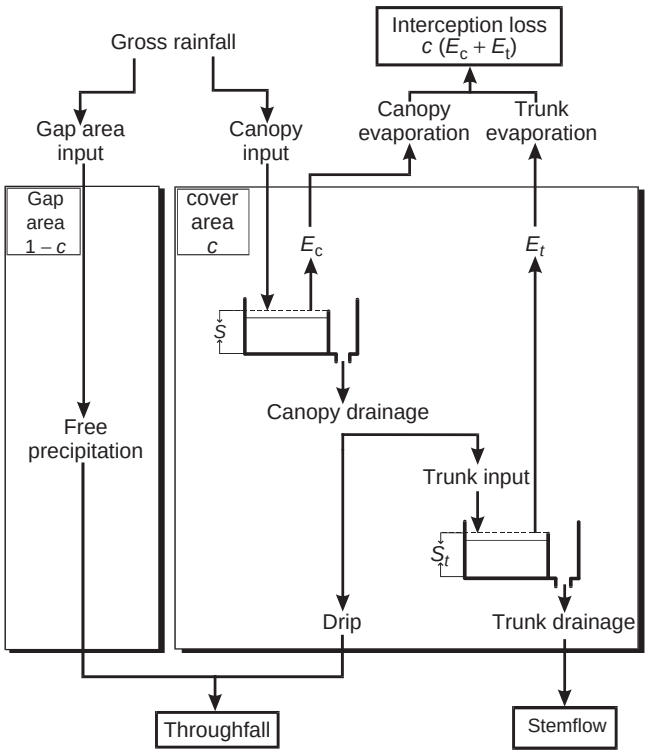


Figure 3 The conceptual framework of the sparse Rutter model (Adapted from Valente *et al.*, 1997)

based on the assumption of horizontally uniform conditions. Such conditions do not always occur and current research is largely directed at modeling interception under these circumstances, where the application of the Penman–Monteith equation may not be valid. For example, small tropical islands and tropical coastal sites have particularly high rates of interception loss, which are poorly simulated by the existing models (Schellekens *et al.*, 1999, 2000; Roberts *et al.*, 2004). Forests with prominent trees and savannah-type ecosystems with isolated trees are another example. In some of these cases, it may also be important to consider the influence of the incident angle of wind-driven rainfall on the high, and nonrandom, spatial variation of throughfall. Three-dimensional models (e.g. Herwitz and Slye, 1995; Xiao *et al.*, 2000), coupled with improved estimates of evaporation, can probably constitute a good basis for the development of new modeling solutions to apply in spatially heterogeneous forests.

REFERENCES

Asdak C., Jarvis P.G. and Van Gardingen P. (1998) Modelling rainfall interception in unlogged and logged forest areas of Central Kalimantan, Indonesia. *Hydrology and Earth System Sciences*, **2**, 211–220.

Aston A.R. (1979) Rainfall interception by eight small trees. *Journal of Hydrology*, **42**, 383–396.

Aussenac G. (1968) Interception des précipitations par le couvert forestier. *Annales des Sciences Forestières*, **25**, 135–156.

Aussenac G. and Boulangeat C. (1980) Interception des précipitations et évapotranspiration réelle dans des peuplements de fenillu (*Fagus sylvatica* L.) et de resineux (*Pseudotsuga menziesii* (Mirb.) Franco). *Annales des Sciences Forestières*, **37**, 91–107.

Bouten W., Swart P.J.F. and de Water E. (1991) Microwave transmission, a new tool in forest hydrological research. *Journal of Hydrology*, **124**, 119–130.

Bruijnzeel L.A. (1990) *Hydrology of Moist Tropical Forests and Effects of Conversion: A State of Knowledge Review*, IHP-UNESCO Humid Tropical Programme: Paris, p. 224.

Calder I.R. (1986) A stochastic model of rainfall interception. *Journal of Hydrology*, **89**, 65–71.

Calder I.R. (1990) *Evaporation in the Uplands*, John Wiley: Chichester, p. 148.

Calder I.R. (1996) Rainfall interception and drop size – development and calibration of the two layer stochastic interception model. *Tree Physiology*, **16**, 727–732.

Calder I.R., Hall R.L., Rosier P.T.W., Bastable H.G. and Prasanna K.T. (1996) Dependence of rainfall interception on drop size: 2. Experimental determination of the wetting functions and two-layer stochastic model parameters for five tropical tree species. *Journal of Hydrology*, **185**, 379–388.

Calder I.R. and Rosier P.T.W. (1976) The design of large plastic sheet net rainfall gauges. *Journal of Hydrology*, **30**, 403–405.

Calder I.R. and Wright I.R. (1986) Gamma ray attenuation studies of interception from sitka spruce: some evidence for an additional transport mechanism. *Water Resources Research*, **22**, 409–417.

Calder I.R., Wright I.R. and Murdiyarso D. (1986) A study of evaporation from tropical rain forest- West Java. *Journal of Hydrology*, **89**, 13–31.

Carlyle-Moses D.E. and Price A.G. (1999) An evaluation of the Gash interception model in a northern hardwood stand. *Journal of Hydrology*, **214**, 103–110.

Chamberlain C.C. (1975) The movement of particles in plant communities. In *Vegetation and the Atmosphere*, Montheith J.L. (Ed.), Vol. 1, Academic Press: London, pp. 155–203.

Darnhofer T., Gatama D., Huxley P. and Akunda E. (1989) The rainfal distribution at a tree/crop interface. In *Meteorology and Agroforestry*, Proceedings of the ICRAF/WMO/UNEP Workshop on Application of Meteorology to Agroforestry systems Planning and Management, Reifsnyder W.E. and Darnhofer T. (Eds.), ICRAF: Nairobi, pp. 87–99.

Dolman A.J. (1987) Summer and winter rainfall interception in an oak forest. Predictions with an analytical and a numerical simulation model. *Journal of Hydrology*, **90**, 1–9.

Dykes A.P. (1997) Rainfall interception from a lowland tropical rainforest in Brunei. *Journal of Hydrology*, **200**, 260–279.

Eschner A.R. (1967) Interception and soil moisture distribution. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 191–200.

Forgeard F., Gloaguen Y.C. and Touffet J. (1980) Interception des précipitations et apport au sol d'élèments mineraux par les eaux de pluie et les pluviolessivats dans une hêtraie atlantique et dans quelques peuplements résineux en Bretagne. *Annales des Sciences Forestières*, **37**, 53–71.

Gash J.H.C. (1979) An analytical model of rainfall interception by forests. *Quarterly Journal of the Royal Meteorological Society*, **105**, 43–55.

Gash J.H.C. and Stewart J.B. (1977) The evaporation from Thetford forest during 1975. *Journal of Hydrology*, **35**, 385–396.

Gash J.H.C., Lloyd C.R. and Lachaud G. (1995) Estimating sparse forest rainfall interception with an analytical model. *Journal of Hydrology*, **170**, 79–86.

Gash J.H.C., Valente F. and David J.S. (1999) Estimates and measurements of evaporation from wet, sparse pine forest in Portugal. *Agricultural and Forest Meteorology*, **94**, 149–158.

Gash J.H.C., Wright I.R. and Lloyd C.R. (1980) Comparative estimates of interception loss from three coniferous forests in Great Britain. *Journal of Hydrology*, **48**, 89–105.

Gómez J.A., Vanderlinden K., Giráldez J.V. and Fereres E. (2002) Rainfall concentration under olive trees. *Agricultural Water Management*, **55**, 53–70.

Hall R.L. (1992) An improved numerical implementation of Calder's stochastic model of rainfall interception – a note. *Journal of Hydrology*, **140**, 389–392.

Hancock N.H. and Crowther J.M. (1979) A technique for the direct measurement of water storage on a forest canopy. *Journal of Hydrology*, **41**, 105–122.

Herwitz S.R. (1982) The redistribution of rainfall by tropical rainforest canopy tree species. *The First National Symposium on Forest Hydrology*, Melbourne 11–13 May, The Institute Of Engineers, National Conference Publication No. 82/6, Australia, pp. 26–29.

Herwitz S.R. (1985) Interception storage capacities of tropical rainforest canopy trees. *Journal of Hydrology*, **77**, 237–252.

Herwitz S.R. and Slye R.E. (1995) Three-dimensional modeling of canopy tree interception of wind-driven rainfall. *Journal of Hydrology*, **168**, 205–226.

Horton R.E. (1919) Rainfall interception. *Monthly Weather Review*, **47**, 603–623.

Hutjes R.W.A., Wierda A. and Veen A.W.L. (1990) Rainfall interception in the tai forest, ivory coast: application of two simulation models to a humid tropical system. *Journal of Hydrology*, **114**, 259–275.

Jackson N.A. (2000) Measured and modelled rainfall interception loss from an agroforestry system in Kenya. *Agricultural and Forest Meteorology*, **100**, 323–336.

King B.P. and Harrison S.J. (1998) Throughfall patterns under an isolated oak tree. *Weather*, **53**, 111–121.

Klaassen W., Bosveld F. and de Water E. (1998) Water storage and evaporation as constituents of rainfall interception. *Journal of Hydrology*, **212–213**, 36–50.

Lankreijer H.J.M., Hendriks M.J. and Klaassen W. (1993) A comparison of models simulating rainfall interception of forests. *Agricultural and Forest Meteorology*, **64**, 187–199.

Leonard R.E. (1967) Mathematical theory of interception. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 131–136.

Leyton L., Reynolds E.R.C. and Thompson F.B. (1967) Rainfall interception in forest and moorland. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 163–178.

Liu S. (1998) Estimation of rainfall storage capacity in the canopies of cypress wetland and slash pine uplands in North-Central Florida. *Journal of Hydrology*, **207**, 32–41.

Llorens P. and Gallart F. (2000) A simplified method for forest water storage capacity measurement. *Journal of Hydrology*, **240**, 131–144.

Lloyd C.R., Gash J.H.C., Shuttleworth W.J. and Marques A.O. (1988) The measurement and modelling of rainfall interception by amazonian rain forest. *Agricultural and Forest Meteorology*, **43**, 277–294.

Lloyd C.R. and Marques A.O. (1988) Spatial variability of throughfall and stemflow measurements in Amazonian rain forest. *Agricultural and Forest Meteorology*, **42**, 63–73.

Loescher H.W., Powers J.S. and Oberbauer S.F. (2002) Spatial variation of throughfall volume in an old-growth tropical wet forest, costa rica. *Journal of Tropical Ecology*, **18**, 397–407.

McNaughton K.G. and Jarvis P.G. (1983) Predicting effects of vegetation change on transpiration and evaporation. In *Water Deficits and Plant Growth*, Kozlowski T.T. (Ed.), Vol. III, Academic Press: New York, pp. 1–47.

Miller H.G. (1984) Water in forests. *Scottish Forestry*, **38**, 165–181.

Mizutani K., Yamanoi K., Ikeda T. and Watanabe T. (1997) Applicability of the eddy correlation method to measure sensible heat transfer to forest under rainfall conditions. *Agricultural and Forest Meteorology*, **86**, 193–203.

Monteith J.L. (1965) Evaporation and environment. In *The State and Movement of Water in Living Organisms*, Fogg G.E. (Ed.), Symposium of the Society for Experimental Biology, Vol. 19, Cambridge University Press: pp. 205–234.

Monteith J.L. and Unsworth M.H. (1990) *Principles of Environmental Physics*, *Second Edtition*, Edward Arnold: London, p. 291.

Pearce A.J. and Rowe L.K. (1981) Rainfall interception in a multi-storied evergreen mixed forest: estimates using Gash's analytical model. *Journal of Hydrology*, **49**, 341–353.

Pearce A.J., Rowe L.K. and Stewart J.B. (1980) Nightime, wet canopy evaporation rates and the water balance of an evergreen mixed forest. *Water Resources Research*, **16**, 955–959.

Roberts J.M., Gash J.H.C., Tani M. and Bruijnzeel L.A. (2004) Controls of evaporation in lowland tropical rainforest. In *Forest-Water-People in the Humid Tropics*, Bonell M., Bruijnzeel L.A. and Kirby C. (Eds.), Cambridge University Press 287–313.

Rowe L.K. (1983) Rainfall interception by an evergreen beech forest, Nelso, New Zealand. *Journal of Hydrology*, **66**, 143–158.

Rutter A.J. (1967) An analysis of evaporation from a stand of Scots Pine. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 403–417.

Rutter A.J. (1975) The hydrological cycle in vegetation. In *Vegetation and the Atmosphere*, Montheith J.L. (Ed.), Vol. 1, Academic Press: London, pp. 111–154.

Rutter A.J., Kershaw K.A., Robins P.C. and Morton A.J. (1971) A predictive model of rainfall interception in forests. I. Derivation of the model from observations in a plantation of corsican pine. *Agricultural Meteorology*, **9**, 367–384.

Rutter A.J., Morton A.J. and Robins P.C. (1975) A predictive model of rainfall interception in forests. II. Generalization of the model and comparison with observations in some coniferous and hardwood stands. *Journal of Applied Ecology*, **12**, 367–380.

Schellekens J., Bruijnzeel L.A., Scatena F.N., Holwerda F. and Bink N.J. (2000) Evaporation from a tropical rainforest, Luquillo Experimental Forest, Puerto Rico. *Water Resources Research*, **36**, 2183–2196.

Schellekens J., Scatena F.N., Bruijnzeel L.A. and Wickel A.J. (1999) Modelling rainfall interception by a lowland tropical rain forest in northeastern Puerto Rico. *Journal of Hydrology*, **225**, 168–184.

Shuttleworth W.J. (1988) Evaporation from Amazonian rainforest. *Proceedings of the Royal Society of London*, **B 233**, 321–346.

Stewart J.B. (1977) Evaporation from the wet canopy of a pine forest. *Water Resources Research*, **13**, 915–921.

Stewart J.B. and Thom A.S. (1973) Energy budgets in pine forest. *Quarterly Journal of the Royal Meteorological Society*, **99**, 154–170.

Teklehaimanot Z. and Jarvis P.G. (1991) Direct measurement of evaporation of intercepted water from forest canopies. *Journal of Applied Ecology*, **28**, 603–618.

Thom A.S. (1975) Momentum, mass and heat exchange of plant communities. In *Vegetation and the Atmosphere*, Montheith J.L. (Ed.), Vol. 1, Academic Press: London, pp. 57–109.

Ubarana V.N. (1996) Observation and modelling of rainfall interception loss in two experimental sites in Amazonian forest. In *Amazonian Deforestation and Climate*, Gash J.H.C., Nobre C.A., Roberts J.M. and Victoria R.L. (Eds.), John Wiley & Sons: Chichester, pp. 151–162.

Valente F., David J.S. and Gash J.H.C. (1997) Modelling interception loss for two sparse eucalypt and pine forests in central Portugal using reformulated Rutter and Gash analytical models. *Journal of Hydrology*, **190**, 141–162.

Van Dijk A.I.J.M. and Bruijnzeel L.A. (2001a) Modelling rainfall interception by vegetation of variable density using an adapted analytical model. Part 1. model description. *Journal of Hydrology*, **247**, 230–238.

Van Dijk A.I.J.M. and Bruijnzeel L.A. (2001b) Modelling rainfall interception by vegetation of variable density using an adapted analytical model. Part 2. model validation for a tropical upland mixed cropping system. *Journal of Hydrology*, **247**, 239–262.

Ward R.C. and Robinson M. (2000) *Principles of Hydrology*, *Fourth Edition*, McGraw-Hill: London, p. 450.

Xiao Q., McPherson E.G., Ustin S.L. and Grismer M.E. (2000) A new approach to modeling tree rainfall interception. *Journal of Geophysical Research*, **105**, 29,173–29,188.

Zinke P.J. (1967) Forest interception studies in the United States. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 137–161.

44: Evaporation from Lakes

JONATHAN W FINCH AND ROBIN L HALL

Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

Lakes are an important part of the hydrological cycle, but quantifying the evaporation rates from them is not a trivial task. The amount of radiant energy captured by a lake is generally the dominant control on the annual evaporation rate. At shorter time periods, the major factors affecting lake evaporation are: the albedo, the heat-storage term of the energy budget, and the atmospheric diffusion processes. The albedo is a function of the solar elevation angle and the proportion of downward diffuse radiation, and can be predicted from empirical relationships. The heat-storage term can decouple the evaporation rates from the net radiation, but can be estimated simply in the case of well-mixed water. However, in the situation where the lake becomes thermally stratified, more complex, hydrodynamic models are required. The three commonly used methods of estimating lake evaporation are: mass transfer, energy balance, and combination equations. There are strengths and limitations to each of these methods.

INTRODUCTION

A lake is defined as a large inland body of water. Globally, lakes cover an area of about 1.5 million km^2 and store approximately 0.8% of the nonfrozen freshwater (Shihklmonaov, 1993). Although this may seem a small percentage of the world's freshwater, lakes are an important part of the hydrological cycle because they act as storage that can have a significant effect on the flows in rivers. They also provide a range of aquatic habitats. From the human perspective they can be vital sources of water supply, in which case they may be natural or manmade. Forecasting the evaporative losses from these water bodies is important for a number of activities, such as reservoir catchment management operations, wetland habitat conservation, and so on. However, quantifying the evaporation from lakes is not a trivial task because there are a number of factors that can affect the evaporation: the climatology and the physiography of the lake and its surroundings. Unlike other land surfaces, the water in a lake is liquid and thus has the potential to transport stored heat within the lake itself and into and out of the lake.

The evaporation rate is controlled by the energy available at the water surface, and the ease with which water vapor diffuses into the atmosphere. This section begins by considering the energy balance of lakes, and the factors affecting it. The factors affecting the processes of diffusion into the atmosphere are then considered. Finally, the three methods most commonly used to estimate the evaporation from lakes are described.

THE ENERGY BALANCE

The amount of radiant energy captured by the lake (net radiation) is generally the dominant control on the annual evaporation rate. The net radiation, Q^*, is the difference between the downward and reflected solar radiation, plus the difference between the downward and upward longwave radiation (*see* **Chapter 39, Surface Radiation Balance, Volume 1**).

Net Short-wave Radiation

The net short-wave radiation, K^*, is that portion of the downward solar radiation captured by the lake, taking into account losses due to reflection. A portion, $K_{\downarrow,d}$, of the solar radiation incident at the top of the atmosphere, K_0, reaches the lake surface as the direct beam; see Figure 1. Some solar radiation also reaches the lake surface as the diffuse component, $K_{\downarrow,f}$, after scattering by atmospheric particles and clouds. Part of the solar radiation reaching

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

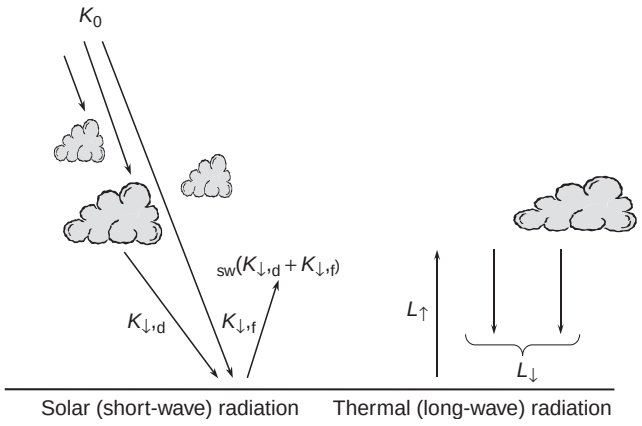


Figure 1 Diagrammatic representation of the components of the net radiation of a lake surface (see text for meaning of symbols). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

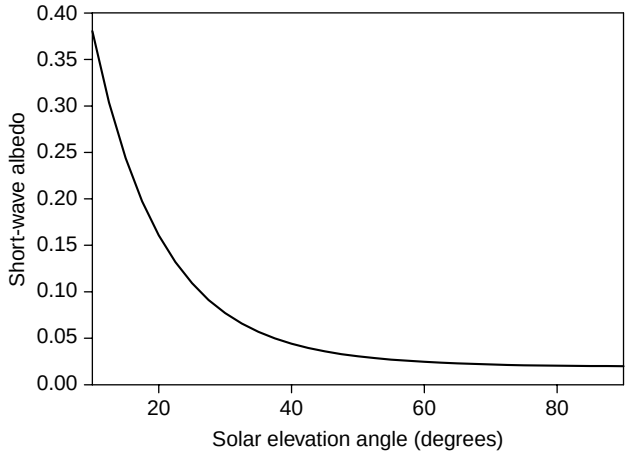


Figure 2 Theoretical variation of the albedo of a water surface with the solar elevation angle

the lake is reflected so that the net short-wave radiation is:

$$Q^* = (1 - \alpha_{sw})(K_{\downarrow,d} + K_{\downarrow,f}) \quad (1)$$

Where α_{sw} is the short-wave albedo, or reflection coefficient, of the downward global solar radiation, that is, the sum of the diffuse and direct beam. The albedo may be affected by a number of factors, such as the turbidity and the reflection coefficient of the lake bottom. There is a considerable literature on the albedo of water, albeit mainly relating to oceanography; Jerlov (1976); Kirk (1994); Mobley (1994) discuss the subject in detail.

The albedo will vary, both over the course of a day and over a year because it is a function of the solar elevation angle. For an infinitely deep body of pure water with no suspended particles and a perfectly smooth surface, the albedo can be calculated from a combination of Snell's law and Fresnel's equation, Figure 2. The albedo has a very low value, about 0.02, at high elevation solar elevation angles. However, as the elevation angle decreases, the albedo steadily increases until, at an angle of 37°, it has doubled from a value of 0.02 to 0.04. At lower elevation angles it rises rapidly. Davies (1972) found that this curve was in good agreement with observations for cloudless skies, that is, when the direct beam can be expected to be significantly greater than the diffuse radiation. However, he noted systematic differences from the theoretical model. At high solar elevation angles, the measured albedo exceeded the model values, which he attributed to the effect of waves on the water surface; whereas at a solar elevation angles less than 30°, the measured values were less than theoretical values, which he attributed to the increasing proportion of diffuse radiation. The minimum albedo, corresponding to high solar elevation angles, will be a function of both the latitude of the site and the time of year. Since the periods of high albedo are associated with low downward solar radiation, the impact of the variations in albedo on the annual evaporation rate is diminished.

For cloudy conditions, the situation is complicated by the fact that a larger proportion of the downward solar radiation is in the form of diffuse radiation, with a tendency to increase the albedo. Thus, the albedo will include a component that is a function of the amount and angular distribution of the downward diffuse solar radiation.

Predicting the albedo of water is thus not a trivial issue. Although physically rigorous numerical models of albedo have been constructed, for example Mobley (1995), they are demanding both computationally and in terms of the input parameters. These models are usually regarded as research tools, and empirical models have tended to be used in practice. Anderson (1954) measured albedo at Lake Hefner, in the southwest of the USA, and produced a series of regression equations relating the albedo to the solar elevation angle and the cloud type, for example, low overcast, high broken, and so on. In practice, this method is difficult to use because the information on the clouds is not readily available. Payne (1972) made measurements of albedo from platforms at two sites (the Sargasso Sea and coastal waters off Massachusetts) to produce a look up table, reproduced graphically in Figure 3, that related the albedo of water to the solar elevation and the transmissivity of the atmosphere (defined as the ratio of the downward solar radiation at the Earth's surface to that at the top of the atmosphere).

Although it is uncertain to what extent Payne's model can be applied to inland waters, the optical characteristics of large, deep lakes are probably not too different to those of oceanic waters. However, for smaller, shallower water lakes, this is not necessarily true because of the possibility of reflectance from the bottom, differences in the waves on the surface, and differences in the amount and type of suspended particles; all of which will tend to increase the albedo.

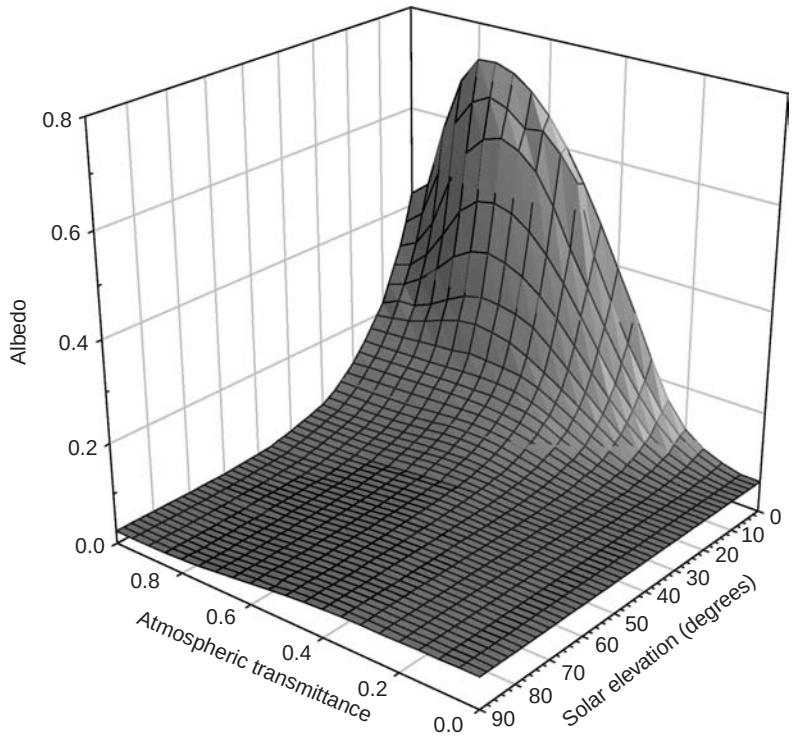


Figure 3 Variation of the albedo of a water surface as a function of solar elevation angle and atmospheric transmittance (Using the data of Payne, 1972)

The surface of the water will rarely be perfectly flat, because of wind-induced waves. The size and orientation of these waves will be a function of the strength and direction of the wind, and the fetch and depth of the lake. The effect of waves on the water surface will be to change the distribution of elevation angles of the incident radiation. However, Richter (1988) failed to detect any correlation between albedo and wind speed in four years of measurements of the radiation balance at a lake in Germany, so it would seem that this factor may not be significant.

Suspended particles in the lake will tend to increase the albedo, mainly by backscattering the radiation that has penetrated into the lake. This effect will be a function of the size and number of the suspended particles that will depend on the type of lake; the phytoplankton abundance in the upper layers will have an effect, as will the amount of suspended sediments. Floating aquatic vegetation can also have an impact on the albedo. Finally, reflectance from the bottom will be a function of the albedo of the bottom materials, the depth and turbidity of the water, and the angle of the incident radiation. Reflectance from the bottom can generally be neglected for lakes over 5 m deep.

Net Long-wave Radiation

There is a significant exchange of radiant energy between the lake surface and the atmosphere in the form of long-wave (thermal) radiation. Net long-wave radiation can be expressed as the difference between the downward radiation, $L_{\downarrow}$, and upward radiation, $L_{\uparrow}$. The downward radiation is related to the temperature and humidity structure of the atmosphere and the cloud cover, because its dominant source is the water vapor molecules in the atmosphere. The reflectivity of the water surface is normally around 0.02; however, it is often taken to have a value of zero, which does not introduce any significant error. The upward radiation emitted from the surface of the lake and can be calculated from the Stefan–Boltzman law:

$$L_{\downarrow} = \varepsilon_a \sigma (T_s + 273.13)^4 \qquad (2)$$

where ε_a is the effective emissivity, σ is the Stefan–Boltzman constant (4.9×10^{-9} MJ m^{-2} K^{-4} d^{-1}) and T_s is the temperature of the water, at the surface, in degrees Celsius.

Both Anderson (1954) and Davies (1972), on the basis of numerous measurements, concluded that the average value of the effective emissivity is 0.97. However, Richter (1988), also on the basis of measurements, found a value of 0.983 was more appropriate. This latter value is supported by more recent work, for example Ogawa *et al.* (2002).

Available Energy

When time-steps of less than a year are of interest, it is generally more convenient to consider the available energy,

A, rather than the net radiation, as this takes into account the change in the heat stored in the lake, W, and the advected energy, A_q, due to inflows and outflows of water.

$$A = Q^* + W + A_q \tag{3}$$

The seasonal distribution of evaporation can be significantly affected by the heat-storage capacity of the lake; which is, to a large extent, determined by its depth. In tropical regions, the temperature of large lakes is generally remarkably uniform, for example, Sene *et al.* (1991). However, at higher latitudes, where there is a strong seasonal variation in the solar radiation input, the stored energy in the lake introduces a time lag with the result that the evaporation rate is decoupled from the net radiation. This is illustrated in Figure 4, which shows the calculated energy balance for two theoretical, well-mixed lakes, at a latitude of 51°36'N in southern England, with water depths of (a) 5 and (b) 15 m. The radiation components were calculated using the model of Finch and Gash (2002), using a daily time-step, for the period 1972 to 2002 inclusive, and then aggregated to 10-day period 30-year averages. The net radiation closely follows the noon solar elevation for both depths, but for a depth of 5 m, there is a lag, of about a month, before the maximum latent heat-flux occurs whilst, for a depth of 15 m, the lag is about two months.

The situation becomes more complicated if a lake becomes thermally stratified. Detailed discussions are given by Imberger and Patterson (1989) and Henderson–Sellers (1984). The temperature dependence of the density of water is a key factor (the maximum density occurs at a temperature of 4 °C). Stratification occurs in large, deep lakes (at mid and high latitudes), and may accentuate the time lag between the net radiation and the evaporation rate. During the early spring, most of the large, temperate lakes exhibit a nearly uniform temperature distribution with depth (homothermal conditions). As the year progresses and the weather warms up, the lake receives heat at an increasingly rapid rate. Initially, the lake remains homothermal because the heat received at the surface layers is transported to deeper layers by wind-induced currents and turbulence. As the rate of heating continues to increase, it begins to exceed the rate of transfer to deeper layers with the result that the temperature of the surface layers increases faster than those of the deeper layers. As the heating continues, a point of inflection develops in the temperature depth profile forming a well-mixed upper layer (the epilimnion) is formed, with relatively intense gradients at its bottom boundary. The plane of maximum temperature gradient is known as the *thermocline*. During the remainder of the heating period, the epilimnion deepens and the thermocline slowly descends into the lake. Once a thermocline has formed, the deeper regions of the lake are relatively uninfluenced by changes in surface conditions.

In the fall, after the lake has attained its maximum heat content, the wind mixing is augmented by convective mixing as a result of surface cooling. The resultant increase in density causes the water to sink and the thermocline moves down rapidly into the deeper layers of the lake, often referred to as turnover. The thermocline continues to move down rapidly as the well-mixed upper layers cool further, until the whole lake again attains homothermal conditions.

A "reverse" stratification can be created in winter, especially in continental climates where the average temperature falls below 4 °C, but the cool layer is much thinner than the epilimnion of summer. Sufficient cooling may permit the lake to freeze over, whilst retaining the temperature of the deeper water in the range of 2–4 °C. If the minimum (winter) temperature of the lake is greater than 4 °C, then there is only one turnover (in the fall).

The net result of the heat storage is that water temperatures are lower than air temperatures during the summer and vice versa during the winter. Thus, the evaporation rates from large deep lakes may be higher in winter than in summer (Blanken *et al.*, 2000).

Change in Heat Storage

It is simple to measure the change in lake heat storage, for example, by deploying thermistors at a range of depths and at a number of locations. However, estimating the change in heat storage using a model can prove a challenge. Predicting the development of thermal stratification requires a comprehensive numerical hydrothermal model. Suitable models are described by Ahsan and Blumberg (1999), Hamilton and Schladow (1997), Hostetler and Bartlein (1990), and Henderson-Sellers (1984). These models are a major topic in themselves and so will not be described here. The problem can be significantly simplified if it is assumed that: the lake is horizontally uniform (so that a one-dimensional model can be used), the heat flux at the bottom of the lake is negligible, the time-step is daily or greater, and the water is well mixed. An analytical solution, using commonly available meteorological data, is provided by the concept of equilibrium temperature. Edinger *et al.* (1968) introduced this concept (and an associated time constant) as the temperature towards which the water temperature is driven by the net heat exchange, that is, when the water is at equilibrium temperature, the net rate of heat exchange is zero. From this, he was able to derive an expression for the temperature of a well-mixed lake as a function of time and water depth, and hence the change in heat stored. Similar approaches have been taken by Keijman (1974), de Bruin (1982), and Finch (2001). The concept was extended by Fraedrich *et al.* (1977) to include the effect of energy advected to a reservoir by inflow and outflow.

Recently, Finch and Gash (2002) have developed a simple finite difference model, where the water temperature is derived by iteration. The model successfully estimated

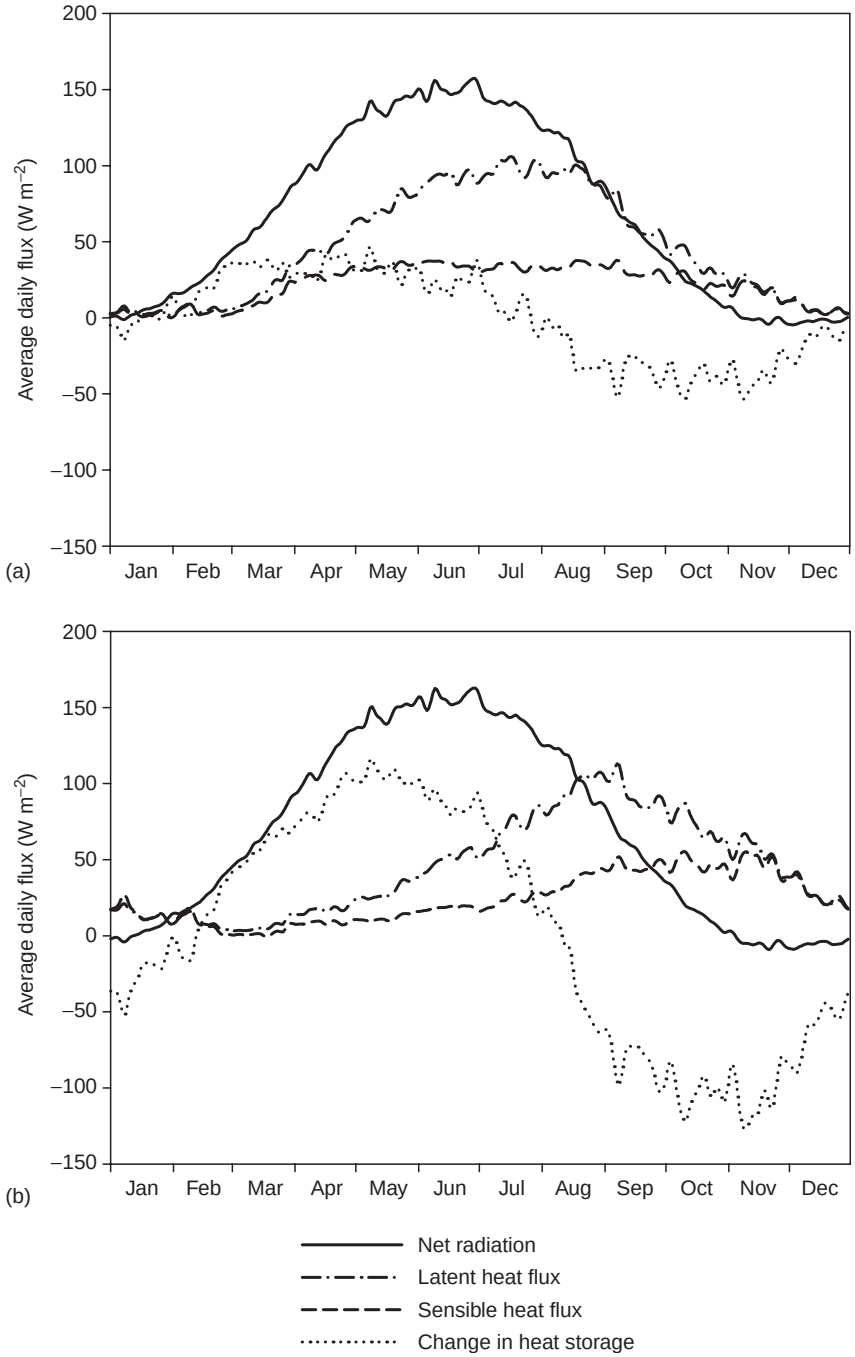


Figure 4 The modeled 28-year average energy-budget components for lake depths of (a) 5 m and (b) 15 m

the annual evaporation rates, from a reservoir with an area of 17.2 ha in southeast England, to an accuracy of 2%, and the monthly evaporation rates to an accuracy of 10%.

Advected Energy

The heat transferred into a lake by precipitation, and inflows and outflows of water may be a significant factor in the energy budget of the whole lake. Inflow includes rivers and land surface run off, bank storage, and seepage from groundwater. Outflow includes rivers, controlled withdrawals (reservoirs), and leakage to groundwater. If the advected energy is a significant component of the energy budget, then it must be estimated. It can be approximated by:

$$A_Q = \rho_w c_w (Q_i T_i - Q_o T_w + P T_p) \qquad (4)$$

where A_Q is the advected energy per unit area of lake, Q_i and Q_o are the rates of inflow and outflow per unit area of lake, P is the rate of precipitation, and T_i, T_w, and T_p are the temperatures (°C) of the inflow, the outflow (i.e. the lake), and the precipitation; ρ_w and c_w are the density and specific heat of water respectively.

Diffusion into the Atmosphere

A number of factors can affect the diffusion processes. With constant wind speed, the evaporation rate is related to the size of the lake surface and to the relative humidity of the air. As a parcel of air moves across a lake, its water vapor content increases, resulting in a decrease in the rate of evaporation, as the layer of air that overlies the lake and, due to its proximity to the water, approaches saturation. The larger the lake, the greater will be the total reduction in the depth of water evaporated. If the lake is large enough, the humidity of the air will be largely independent of the distance it has traveled from the edge and hence the evaporation will be closely related to the amount of energy available. On the other hand, small lakes exert little influence on the temperature or humidity of the overlying air. Some observations indicate, however, that evaporation can be enhanced over large lakes in situations where the effect of increased wind speed more than compensates for the increased humidity of the air associated with the larger distance traveled by the air over water, for example, Venalainen *et al.* (1998). They also noted that evaporation from lakes fringed with forest would be reduced through sheltering: apparently the reduction in turbulence associated with the reduced wind speeds more than compensates for the increased aerodynamic roughness of the forest.

Empirical and theoretical studies have been made of the evaporation from small areas (the oasis effect). As an airstream moves over a homogeneous land cover, the energy exchange with the land surface will tend to reach an equilibrium condition, principally by changes to the humidity of the air immediately above the surface. If the land cover changes, then the energy exchange will also change to establish a new equilibrium. Various studies (e.g. Brutsaert and Yeh, 1969; Brutsaert and Yeh, 1970; Brutsaert and Yu, 1968; and Rao *et al.*, 1974) have shown that, although the effect of a change in surface cover propagates slowly up into the atmosphere, the adjustment of the surface evaporation rate into a moving airstream occurs quickly as the air passes on to a different type of evaporative surface (generally, within 5 to 10 m of the boundary between the two surfaces). This suggests that for lakes, it is the surface characteristics of the lake itself which will dominate. In addition, lakes of a significant size, many kilometers across, can modify the wind direction as well as the strength, for example, Samuelsson and Tjernstrom (2001); Taylor *et al.* (1998).

The preceding discussions apply to a freshwater lake. However, for saline water, the evaporation rate decreases by about 1% for each 1% increase in salinity because of the reduced vapor pressure of the saline water. This effect is normally small enough to be discounted when estimating evaporation rates from "fresh" lakes although methods for quantifying the effect are available, for example, Calder *et al.* (1984); Ali *et al.* (2001), which may be important in semiarid and arid areas.

QUANTIFYING THE EVAPORATION

The three main methods used to estimate the evaporation from lakes are: the bulk transfer; energy balance, and combination equations. The reader is referred to key texts for other methods which are either mainly of historic interest or are rarely used: evaporation pans (Gangopaghaya *et al.*, 1966), mass balance (Harbeck *et al.*, 1954; Lapworth, 1965), and empirical factors (Finch, 2003). Micrometeorological measurement techniques are not included here as they are described elsewhere (*see* **Chapter 40, Evaporation Measurement, Volume 1**).

Bulk (Mass) Transfer

A simple derivation of the bulk transfer equation is given by Sene *et al.* (1991), which has the form

$$E = Cu(e_w^* - e) \tag{5}$$

where E is the evaporation and u is the wind speed, e is the vapor pressure of the air and e_w^* the saturated vapor pressure. The wind speed and the vapor pressures are measured at a reference height above the water surface, typically 2 m. C is the mass-transfer coefficient, which can be thought of as the total drag coefficient; the combination of skin friction and a force resulting from the deceleration of the wind in the direction of flow. Over a uniform surface, C can be calculated from theory as a function of the atmospheric stability and the roughness of the surface which itself is affected by the wind speed (Brutsaert, 1982; see p. 117). This coefficient has often been determined for sea surfaces, although there is considerable scatter in the results (Brutsaert, 1982 see Table 3). For most lakes, the conditions of surface uniformity are not met and it is necessary to make more restrictive assumptions to obtain a theoretical solution to the evaporation and heat transfer equations (Brutsaert, 1982; pp. 167–171). The value of C reflects the transfer characteristics of the particular lake, which are determined by its geometry, plant cover, and the topography, land use and climate of the surrounding land. Moreover, the value of the coefficient is specific for the characteristics of the site used to record the meteorological data; for example, a value derived for wind speed measured at 2 m

will not be correct for use with wind speeds measured at 10 m, even at the same site. Over the years, meteorological data have been inconsistently measured using a variety of different standards, resulting, according to Singh and Xu (1997), in over 100 such evaporation formulae. It is therefore not possible to find a value of C that is applicable to all lakes. Rather, it is best to determine it empirically for a particular lake from the ratio of the mean evaporation rate, measured using a standard method, for example, eddy correlation (*see* **Chapter 40, Evaporation Measurement, Volume 1**) or the energy budget, to the mean vapor pressure gradient. Nevertheless, attempts have been made to produce a generally applicable value. Based on an extensive measurement program on reservoirs in the western USA, Harbeck (1962) suggested an expression for C that incorporated the area of the lake. The transfer equation is

$$E = 2.909 D_w^{-0.05} u_2 (e_w^* - e) \tag{6}$$

where the evaporation rate, E, is in mm day^{-1}, the vapor pressures, e_w^* and e, are in kPa, D_w is the area of the water surface in m^2, and u_2 is the wind speed at 2 m (m s^{-1}) above the water surface. This is suitable for lakes in the range of 50 m $< D_w^{0.5} <$ 100 km that are in a relatively arid environment.

An alternative form for the mass-transfer equation, dating from the nineteenth century, has also been widely used. This takes the form

$$E = f(u)(e_w^* - e) \tag{7}$$

where $f(u)$ is a function of the wind speed that allows for free convection, that is, evaporation when there is no wind. The wind speed function takes the form:

$$f(u) = a + bu \tag{8}$$

where a and b are empirical constants. Sweers (1976) reviewed wind speed functions and concluded that, for a temperate climate, best results were obtained using the wind function of McMillan (1973) adjusted for the area of the lake in relation to the lake studied by McMillan. This function is,

$$f(u) = \left(\frac{5 \times 10^6}{D_w}\right)^{0.05} (3.6 + 2.5u_3) \tag{9}$$

where u_3 is the wind speed measured over the water at 3 m above the surface.

Simon and Mero (1985) gave up trying to use the mass-transfer method to estimate evaporation from Lake Kinneret in Israel because of inconsistent results and large scatter in estimates of the transfer coefficient. In contrast, Sacks *et al.* (1994) found good agreement (generally within 8%) between the energy-budget evaporation and monthly mass-transfer evaporation for a shallow lake in Florida, but larger discrepancies (mean monthly difference of 24%) for a similar but deeper lake, also in Florida. Correcting the mass-transfer coefficient for stability effects (Stauffer, 1991; Harbeck *et al.*, 1958) did not improve matters. Sacks *et al.* suggested that the differences might be a smoothing effect caused by using long-term mean vapor pressure gradients; one of the main problems with this method is that it is sensitive to the errors in the vapor pressure gradient. They also found that using the Harbeck (1962) form for the mass-transfer coefficient produced lower values that resulted in underestimates of the evaporation from the shallow lake by 14% and from the deep lake by 27%. This was in contrast to Sturrock *et al.* (1992) who found that the Harbeck prescription gave higher values than those based upon energy-budget estimates. The reasons for these discrepancies remain unclear.

Energy Budget

The energy-budget method is generally accepted as the most accurate method of measuring the evaporation from a lake (Winter, 1981). As its name implies, the evaporation from a lake is estimated as the energy component required to close the energy budget when all the remaining components of the budget of the lake are known, that is, it is the residual of:

$$\lambda E + c_w(T_w - T_b)E = Q^* - H + W + F_{\text{in}} - F_{\text{out}} + F_p - G \tag{10}$$

where λ is the latent heat of vaporization, c is the specific heat of water, T_w and T_b are the temperature of the evaporated water (generally the water surface) and an arbitrary base temperature respectively, W is the change in heat storage, F_{in} and F_{out} are the heat-fluxes associated with water flows in and out of the lake, F_P is the heat inflow associated with precipitation, and G is the heat conduction occurring between the water and its substrate. All the energy components are in units of energy per unit surface area of the water.

Usually the sensible heat, H, (the amount of energy directly warming the air) cannot be readily determined and it is eliminated through use of the Bowen ratio, β, defined as the ratio between the sensible and latent heat-fluxes (*see* **Chapter 40, Evaporation Measurement, Volume 1**):

$$\beta = \frac{H}{\lambda E} = \frac{\gamma(T_s - T_a)}{(e_s^* - e)} \tag{11}$$

where γ is the psychrometric parameter, T_s and T_a are the temperatures of the water surface and the air at a reference height and e_s^* and e are the saturated vapor pressure of

the air at the water surface temperature and the vapor pressure of the air at the reference height. $H = \beta\lambda E$ can be substituted into the energy balance equation to give the evaporation rate,

$$E = \frac{Q^* + W + F_{\text{in}} - F_{\text{out}} + F_p - G}{\lambda(1+\beta) + c_w(T_s - T_b)} \tag{12}$$

The second term in the denominator represents a correction term for the difference between the temperature of the evaporated water and an arbitrary base temperature.

By suitable selection of averaging period it is sometimes possible to neglect the F_{in}, F_{out}, and G terms. Indeed, it is usually the case that the energy content of a lake is chiefly governed by the exchange of energy through the surface, rather than the inflows, including precipitation, and outflows and the water-substrate interface (Henderson-Sellers, 1986). This would certainly be the case if the volumes of water flowing in and out of the lake are small compared to the volume, or the temperatures are close to the temperature of the lake. Under these conditions, $T_b = T_s$ and the last four terms in the numerator can be neglected. The energy budget is then given by

$$E = \frac{Q^* + W}{\lambda(1+\beta)} \tag{13}$$

This is sometimes referred to as the reduced energy budget equation (Simon and Mero, 1985; Assouline and Mahrer, 1993; dos Reis and Dias, 1998).

The accuracy of the reduced energy balance equation depends upon the timescale and size of the lake. Because of the heat storage, the deeper the lake, the longer the time interval required between measurements of the temperature profile to attain acceptable accuracy in the temperature differences. In the classic study at Lake Hefner (Anderson, 1954), an accuracy of 5% in the evaporation estimate was achieved for periods of a week or more but the accuracy decreased for shorter periods. For a shallow (average depth 0.6 m) lake, Stewart and Rouse (1976) assumed that daily values were sufficiently accurate to use them as a standard against which an alternative method was validated.

The disadvantages of the energy balance method are the large number and frequency of measurements needed and the difficulties inherent in making some of them. Consequently, it is expensive and has not often been used in the more complete form. Exceptions being the Lake Hefner study (Anderson, 1954), the work by Stauffer (1991) on Lake Mendota, Wisnonsin, and Sturrock *et al.* (1992) on Williams Lake, north central Minnesota, and more recently a comparative study of evaporation from two lakes in Florida (Sacks *et al.*, 1994).

For accuracy, measurements of surface and profile water temperatures and the micrometeorology should be made at a representative point or points over the lake. This has often been achieved using an anchored instrumented raft (e.g. Anderson, 1954; Assouline and Mahrer, 1993; Sturrock *et al.*, 1992) but also for ease of maintenance and cost, measurements have been made over land – sometimes with data being used from distant weather stations. Work has been done to determine the effect on the accuracy of the evaporation estimates of using land-based and distant data sources (e.g. Keijman, 1974; Rosenberry *et al.*, 1993; Winter and Rosenberry, 1995).

Estimation of the Bowen ratio, β, requires measurement of air temperature and vapor pressure above the water and the water temperature to derive the saturated vapor pressure at the temperature of the water surface. This is usually achieved using wet and dry bulb thermometers at a reference height on a raft and a thermometer just below the surface of the water.

The change in heat storage, W, per unit surface area is calculated from the following:

$$W = \rho_w c_w h \frac{\Delta T_w}{\Delta t} \tag{14}$$

where ΔT_w is the change in spatially averaged temperature of the lake in time step Δt. For shallow lakes that are well mixed, T_w can be approximated by the surface temperature (Keijman, 1974). This, however, begs the question as to a suitable average value for the surface temperature; in calm conditions and high solar radiation, spatial variation in surface temperature can be large over short timescales. For deep lakes it is necessary to conduct thermal surveys consisting of temperature profiles with depth, ideally measured at a sufficient number of stations to produce a good average. For example, in the exceptional Lake Hefner study, weekly surveys were made at 16 stations and daily at one of two stations (Anderson, 1954), while at Williams Lake, surveys were made fortnightly at 16 stations (Sturrock *et al.*, 1992). Selection of the appropriate time interval, which will depend upon the size of the lake, can result in the value of the change in heat storage being small enough to be neglected.

Estimation of the energy advected in and out of the lake requires that the inflow and outflow are gauged accurately and the water temperature measured. Where inflow or outflow are large relative to the volume of the lake, and cannot be accurately gauged, the energy balance method may become unusable. However, in many lakes the inflow and outflow are relatively small (e.g. Williams Lake, Sturrock *et al.*, 1992). Sturrock *et al.* (1992) calculated the groundwater volumes using the Darcy equation (see Ward and Robinson, 2000, pages 157–159), and used water temperature from wells for inflow, and surface temperature for leakage. The energy advected by rainfall is usually determined from the recorded rainfall and the wet-bulb temperature recorded during rainfall. Sacks *et al.* (1994); Stauffer (1991); Sturrock *et al.* (1992) concluded that for

the lakes that they studied, the advected energy was trivial compared to the other terms; for example, 1% of the radiation terms. However, Sacks *et al.* (1994) and Stauffer (1991) found that the largest of the advected terms was that due to precipitation. In nonnatural or seminatural lakes, for example, reservoirs and cooling ponds, other advective components may be large but easy to measure.

In some circumstances, the heat conduction into the substrate can be significant, Sturrock *et al.* (1992) found that in the summer neglecting it made a 7% difference to the estimated evaporation from Williams Lake (average depth 5.2 m) in Minnesota. Stauffer (1991) states that ignoring sediment heat exchange can be a major source of error in estimation of evaporation and implies that the Lake Hefner results may be flawed through ignoring this component. He used annual sine-wave functions to model the sediment-water heat exchange (Likens and Johnson, 1969).

Combination Equations

Over the last fifty years the most widely used formula to estimate evaporation from water, or vegetation, has been the Penman equation (Penman, 1948) (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**). Its successful application in many different locations is attributable to its physical basis. Linacre (1993) presents a table comparing monthly or annual measured evaporation with Penman estimates for a wide range of lakes from around the world. The median value of the ratio of measured to estimated evaporation is 0.99 with a standard deviation of 0.12.

Penman combined the mass transfer and energy-budget approaches and eliminated the requirement for surface temperature to obtain his expression for the evaporation in mm per day from open water:

$$E = \frac{\Delta A}{\Delta + \gamma} + \frac{\gamma f(u)(e_a^* - e)}{\Delta + \gamma} \tag{15}$$

where A, the available energy, is in units of equivalent depth of water ($\mathrm{mm\,day^{-1}}$), e_a^* and e are the saturation vapor pressure and the actual vapor pressure of the air respectively, at a reference height above the lake surface. Δ and γ are the rate of change of the saturation vapor pressure with temperature and the psychrometric parameter respectively. The psychrometric parameter is calculated as:

$$\gamma = \frac{c_p P}{0.622\lambda} \tag{16}$$

where c_p is the specific heat of the air, P is the atmospheric pressure and λ is the latent heat of vaporization (which is a function of temperature).

The Penman–Monteith equation (Monteith, 1965) (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**) is a more general form of combination equation. Essentially, the evaporation rate is obtained from the simultaneous solution of diffusion equations for heat and water vapor, and the energy balance equation. When applied to a lake it takes the form

$$E = \frac{1}{\lambda}\left[\frac{\Delta A + \rho c_p (e_a^* - e)/r_a}{\Delta + \gamma}\right] \tag{17}$$

where the aerodynamic resistance r_a is the resistance that the water molecules encounter in moving from the water surface to a reference height in the atmosphere and is inversely proportional to the wind speed. Accurate estimates require that the value of the aerodynamic resistance, r_a, accounts for the effects of surface roughness, size of the lake, and atmospheric stability.

When air travels a long distance over a wet surface, it will tend to saturation so that the second term tends to zero. The first term represents the lower limit of evaporation and is referred to as the *equilibrium rate*. However, in practice, equilibrium evaporation is rarely found; because the atmosphere near the surface is never truly homogeneous and, even over oceans, atmospheric humidity deficits develop. Priestley and Taylor (1972) (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**) analyzed data collected over oceans and extensive saturated land surfaces, and found that the evaporation values were fitted using:

$$E = \alpha \frac{\Delta A}{\Delta + \gamma} \tag{18}$$

where the constant α accounts for the evaporation arising from the humidity deficit in addition to the equilibrium term. The equation is now known as the *Priestley–Taylor* equation. Priestley and Taylor found that the average value of α was 1.26 from the data they examined and there has been subsequent corroboration of this value by other studies. de Bruin and Keijman (1979) used the Priestley–Taylor equation to estimate the evaporation from a large shallow lake (Lake Flevo; 460 $\mathrm{km^2}$, mean depth 3 m) in The Netherlands and found very good agreement with daily evaporation measured by the energy-budget and water-budget methods during the summer and early fall with $\alpha = 1.25$. However, they also found diurnal variation in the value of α which they attributed to the variation of the difference between air and water temperatures and suggested that the conditions producing such variation would be expected from many lakes. They also found evidence of seasonality in the value of α, of at least the same magnitude as the diurnal variation in evaporation. This variation is the result of some evaporation occurring when the available energy was zero.

A further simplification is given by Stewart and Rouse (1976), who derived a variation of the Priestley–Taylor equation by using a linear function of downward solar

radiation to replace the net radiation and heat storage. The resulting equation is identical to the formula of Makkink (1957), who used it to estimate the evaporation from well-watered grass and is:

$$E = a\frac{\Delta}{\Delta+\gamma}K_{\downarrow} + b \qquad (19)$$

where $K_{\downarrow}$ is the downward global solar radiation. As with other methods, the disadvantage of the Priestley–Taylor equation is the requirement for measuring the available energy; it is often not possible or too expensive to make adequate measurements of change in the heat stored in the water for a large lake. de Bruin (1978) overcame this difficulty by combining the Penman and Priestley–Taylor equations, thus eliminating the energy term to give the relationship

$$E = \left(\frac{\alpha}{\alpha-1}\right)\left(\frac{\gamma}{\Delta+\gamma}\right) f(u)(e_a^* - e) \qquad (20)$$

This formula requires only measurements of air temperature, humidity deficit, and wind speed at 2 m. de Bruin tested the method by using a form of the wind function given by Sweers (1976) with time-averaged input data measured at the center of Lake Flevo to calculate evaporation for varying time intervals. He found good agreement with estimates from the energy-budget method for intervals of 10 days or more. He also found that the Priestley–Taylor coefficient was not constant for intervals of a day or less.

CONCLUSIONS

The literature on evaporation from lakes tends to be dominated by measurements of specific lakes; which is probably not surprising, given the large number of factors that can affect the evaporation rates. Nevertheless, it is possible to make reliable predictions of the evaporation to an acceptable accuracy for most purposes. Other than limitations in the measurements, there are three major sources of uncertainty.

The first is the albedo of lakes. For conditions of deep, low turbidity water, there is probably sufficient knowledge to give reasonably accurate information from readily available data (such as Payne, 1972). However, when factors, such as bottom reflectance; turbidity and aquatic vegetation, cannot be ignored, the only reliable methods, currently available, are computationally intensive, for example, radiative transfer models, but the effort required to use these methods is seldom justified for operational purposes, in terms of the "value" of the improved accuracy.

The second major source of uncertainty is the heat-storage term of the energy budget. Although it is possible to make measurements to quantify this term, the effort may be significant for many lakes, and there are models available that can quantify this term. The analytical solution provided by the equilibrium temperature model (Edinger *et al.*, 1968) or the numerical solution of Finch and Gash (2002) can be used if the lake water is well mixed. Models are also available for the case when thermal stratification occurs (e.g. Hostetler and Bartlein, 1990; Hamilton and Schladow, 1997). However, these models are essentially one-dimensional and so using them for a three-dimensional situation requires care, or a more complex model.

The final major cause of uncertainty is in the wind function. The literature is dominated by, what are essentially, empirical descriptions of the wind function and it is unclear what the uncertainty is when these functions are applied to locations other than those for which they were determined.

REFERENCES

Ahsan A.K.M.Q. and Blumberg A.F. (1999) Three-dimensional hydrothermal model of Onondaga Lake, New York. *Journal of Hydraulic Engineering-ASCE*, **125**, 912–923.

Ali H., Madramootoo C.A. and Gwad S.A. (2001) Evaporation model of Lake Qaroun as influenced by lake salinity. *Irrigation and Drainage*, **50**, 9–17.

Anderson E.R. (1954) Energy-budget studies. In *Water Loss Investigations: Lake Hefner Studies*, *Technical Report*, Harbeck G.E., Kohler M.A. and Koberg G.E. (Eds.), US Geological Survey, USGS Professional Paper 269 pp. 71–119.

Assouline S. and Mahrer Y. (1993) Evaporation from Lake Kinneret 1 Eddy correlation system measurements and energy budget estimates. *Water Resources Research*, **29**, 901–910.

Blanken P.D., Rouse W.R., Culf A.D., Spence C., Boudreau L.D., Jasper J.N., Kotchubaja R., Verseghy D. and Schertzer W.M. (2000) Eddy covariance measurements of evaporation from a large, deep northern lake. *Water Resources Research*, **36**, 1069–1078.

Brutsaert W. (1982) *Evaporation into the Atmosphere - Theory, History and Applications*, D Reidel Publishing Company: Dordrecht.

Brutsaert W. and Yeh G.T. (1969) Evaporation from an extremely narrow wet strip at ground level. *Journal of Geophysical Research*, **74**, 3431–3433.

Brutsaert W. and Yeh G.T. (1970) Implications of a type of empirical evaporation formula for lakes and pans. *Water Resources Research*, **6**, 1202–1208.

Brutsaert W. and Yu S.L. (1968) Mass transfer aspects of pan evaporation. *Journal of Applied Meteorology*, **7**, 563–566.

Calder I.R. and Neal C. (1984) Evaporation from saline lakes: a combination equation approach. *Hydrological Sciences Bulletin*, **29**, 89–97.

Davies J.A. (1972) Surface albedo and emissivity for Lake Ontario. *Climatological Bulletin*, **12**, 12–22.

de Bruin H.A.R. (1978) A simple model for shallow lake evaporation. *Journal of Applied Meteorology*, **17**, 1132–1134.

de Bruin H.A.R. (1982) Temperature and energy balance of a water reservoir determined from standard weather data of a land station. *Journal of Hydrology*, **59**, 261–274.

de Bruin H.A.R. and Keijman J.Q. (1979) The Priestley-Taylor evaporation model applied to a large shallow lake in the Netherlands. *Journal of Applied Meteorology*, **18**, 898–903.

dos Reis R.J. and Dias N.L. (1998) Multi-season lake evaporation: energy-budget estimates and CRLE model assessment with limited meteorological observations. *Journal of Hydrology*, **208**, 135–147.

Edinger J.E., Duttweiler D.W. and Geyer J.C. (1968) The response of water temperature to meteorological conditions. *Water Resources Research*, **4**, 1137–1143.

Finch J.W. (2001) A comparison between measured and modelled open water evaporation from a reservoir in south-east England. *Hydrological Processes*, **15**, 2771–2778.

Finch J.W. (2003) Empirical factors for estimating open water evaporation from potential evaporation. *Journal of the Chartered Institution of Water and Environmental Management*, **17**, 51–53.

Finch J.W. and Gash J.H.C. (2002) An application of a simple finite difference model for estimating evaporation from open water. *Journal of Hydrology*, **255**, 253–259.

Fraedrich K., Erath B.G. and Weber G. (1977) A simple model for estimating the evaporation from a shallow water reservoir. *Tellus*, **29**, 428–434.

Gangopaghaya M., Harbeck G.E., Nordenson T.J., Omar M.H. and Uryvaev V.A. (1966) *Measurement and Estimation of Evaporation and Evapotranspiration*, World Meteorological Organisation: Technical Note 83.

Hamilton D.P. and Schladow S.G. (1997) Prediction of water quality in lakes and reservoirs. 1. Model description. *Ecological Modelling*, **96**, 91–110.

Harbeck G.E. (1962) *A Practical Field Technique for Measuring Reservoir Evaporation Utilizing Mass-Transfer Theory*, US Geological Survey: USGS Professional Paper 272-E.

Harbeck G.E., Kohler M.A. and Koberg G.E. (1954) *Water-Loss Investigations: Lake Hefner Studies*, US Geological Survey: USGS Professional Paper 269.

Harbeck G.E., Kohler M.A. and Koberg G.E. (1958) *Water-Loss Investigations: Lake Mead Studies*, US Geological Survey, US Geological Survey Professional Paper 298.

Henderson-Sellers B. (1984) *Engineering Limnology*, Pitman Advanced Publishing Program: Boston.

Henderson-Sellers B. (1986) Calculating the surface-energy balance for lake and reservoir modeling – a review. *Reviews of Geophysics*, **24**, 625–649.

Hostetler S.W. and Bartlein P.J. (1990) Simulation of lake evaporation with application to modeling lake level variations of Harvey-Malheur Lake, Oregon. *Water Resources Research*, **26**, 2603–2612.

Imberger J. and Patterson J.C. (1989) Physical limnology. *Advances in Applied Mechanics*, **27**, 303–475.

Jerlov N.G. (1976) *Marine Optics*, Elsevier: Amsterdam.

Keijman J.Q. (1974) The estimation of the energy balance of a lake from simple weather data. *Boundary-Layer Meteorology*, **7**, 399–407.

Kirk J.T.O. (1994) *Light and Photosynthesis in Aquatic Systems*, Cambridge University Press.

Lapworth C.F. (1965) Evaporation from a reservoir near London. *Journal of the Institution of Water Engineers*, **19**, 163–181.

Likens G.E. and Johnson N.M. (1969) Measurement and analysis of the annual heat budget for the sediments in two Wisconsin lakes. *Limnology and Oceanography*, **14**, 115–135.

Linacre E.T. (1993) Data-sparse estimation of lake evaporation, using a simplified Penman equation. *Agricultural and Forest Meteorology*, **64**, 237–256.

Makkink G.F. (1957) Ekzameno de la formulo de Penman. *Netherlands Journal of Agricultural Science*, **5**, 290–305.

McMillan W. (1973) *Cooling from Open Water Surfaces. Final Report: Part 1; Lake Trawsfynydd Cooling Investigation*, NW/SSD/RR/1204/73.

Mobley C.D. (1994) *Light and Water, Radiative Transfer in Natural Waters*, Academic Press: San Diego.

Mobley C.D. (1995) *Hydrolight 3.0 Users' Guide*, SRI International: N00014-94-C-0062.

Monteith J.L. (1965) Evaporation and environment. *19th Symposium of the Society of Experimental Biology*, Cambridge University Press: Cambridge, pp. 205–234.

Ogawa K., Schmugge T., Jacob F. and French A. (2002) Estimation of broadband land surface emissivity from multi-spectral thermal infrared remote sensing. *Agronomie*, **22**, 695–696.

Payne R.E. (1972) Albedo of the sea surface. *Journal of the Atmospheric Sciences*, **29**, 959–970.

Penman H.L. (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London Series A – Mathematical and Physical Sciences*, **A193**, 120–145.

Priestley C.H.B. and Taylor R.J. (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, **100**, 81–92.

Rao K.S., Wyngaard J.C. and Cote O.R. (1974) Local advection of momentum. heat and moisture in micrometeorology. *Boundary-Layer Meteorology*, **7**, 331–348.

Richter D. (1988) Methods for the determination of the albedo and of the longwave outgoing radiation from a free-water surface and of the longwave incoming radiation of the atmosphere. *Zeitschrift fur Meteorologie*, **38**, 219–233.

Rosenberry D.O., Sturrock A.M. and Winter T.C. (1993) Evaluation of the energy budget method of determining evaporation at Williams Lake, Minnesota, using alternative instrumentation and study approaches. *Water Resources Research*, **29**, 2473–2483.

Sacks L.A., Lee T.M. and Radell M.J. (1994) Comparison of energy-budget evaporation losses from two morphometrically different Florida seepage lakes. *Journal of Hydrology*, **156**, 311–334.

Samuelsson P. and Tjernstrom M. (2001) Mesoscale flow modification induced by land-lake surface temperature and roughness differences. *Journal of Geophysical Research-Atmospheres*, **106**, 12419–12435.

Sene K.J., Gash J.H.C. and McNeil D.D. (1991) Evaporation from a tropical lake: comparison of theory with direct measurements. *Journal of Hydrology*, **127**, 193–217.

Shihklmonaov I.A. (1993) World fresh water resources. In *Water in Crisis: A Guide to the World's Fresh Water Resources*, Gleick P.H. (Ed.), Oxford University Press: Oxford, pp. 13–24.

Simon E. and Mero F. (1985) A simplified procedure for the evaluation of the Lake Kinneret evaporation. *Journal of Hydrology*, **78**, 291–304.

Singh V.P. and Xu C. (1997) Evaluation and generalization of 13 mass-transfer equations for determining free water evaporation. *Hydrological Processes*, **11**, 311–323.

Stauffer R.E. (1991) Testing lake energy budget models under varying atmospheric stability conditions. *Journal of Hydrology*, **128**, 115–135.

Stewart R.B. and Rouse W.R. (1976) A simple method for determining the evaporation from shallow lakes and ponds. *Water Resources Research*, **12**, 623–628.

Sturrock A.M., Winter T.C. and Rosenberry D.O. (1992) Energy budget evaporation from Williams Lake – a closed lake in North Central Minnesota. *Water Resources Research*, **28**, 1605–1617.

Sweers H.E. (1976) A nomogram to estimate the heat-exchange coefficient at the air-water interface as a function of wind speed and temperature; a critical survey of some literature. *Journal of Hydrology*, **30**, 375–401.

Taylor C.M., Harding R.J., Pielke R.A., Vidale P.L., Walko R.L. and Pomeroy J.W. (1998) Snow breezes in the boreal forest. *Journal of Geophysical Research-Atmospheres*, **103**, 23087–23101.

Venalainen A., Heikinheimo M. and Tourula T. (1998) Latent heat flux from small sheltered lakes. *Boundary-Layer Meteorology*, **86**, 355–377.

Ward R.C. and Robinson M. (2000) *Principles of Hydrology*, McGraw-Hill Book Company: London, p. 450.

Winter T.C. (1981) Uncertainties in estimating the water balance of lakes. *Water Resources Bulletin*, **17**, 82–115.

Winter T.C. and Rosenberry D.O. (1995) Evaluation of 11 equations for determining evaporation for a small lake in the north central United States. *Water Resources Research*, **31**, 983–993.

45: Actual Evaporation

ALBERTUS JOHANNES DOLMAN

Department of Hydrology and Geo-Environmental Sciences, Faculty of Earth and Life Sciences, Vrije Universiteit, Amsterdam, The Netherlands

Actual evaporation concerns the evaporation from natural surfaces. Globally actual evaporation from the land amounts to around 71 10^3 km^3 $year^{-1}$. Measurement of evaporation occurs by water balance techniques or directly through micrometeorological techniques. Actual evaporation depends on land cover and meteorological conditions. Evaporation can be predicted by various equations, amongst the Penman-Monteith equation which takes into account both the vegetation properties as well as the meteorological conditions. In the case of sparsely vegetated areas soil evaporation and plant evaporation need to be treated separately. Estimating the correct value of surface conductance is a critical issue in applying the Penman-Monteith equation, but models exists. Evaporation from bare soil can be treated as a two stage process, with rapid evaporation at potential rates after rain, and subsequent drying. At larger than catchment scales, feedbacks with the atmospheric boundary layer make it possible to apply evaporation equations based on equilibrium concepts, such as the Priestley-Taylor equation. There has been substantial progress in evaporation modeling and measurements over the last 25 years. This knowledge can be used much more in assessing changes in evaporation of catchments that may result from perturbations in land use and climate.

INTRODUCTION

Evaporation refers to the transfer of moisture from a particular surface to the overlying atmosphere. The physical process of evaporation consists of the exchange of water molecules between a free water surface and the air. The evaporation rate is expressed as the quantity of water evaporated per unit area per unit time from a (water) surface under existing atmospheric conditions. The surface can be any of the following nonexhaustive list: a lake (*see* **Chapter 44, Evaporation from Lakes, Volume 1**), the inside of the plant leaves (*see* **Chapter 42, Transpiration, Volume 1**, the water surface adhering to a soil conglomerate, or the surface of a soil (*see* **Chapter 40, Evaporation Measurement, Volume 1**) or canopy just after rain (*see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**. It is important to distinguish actual evaporation from potential evaporation. Potential evaporation is the quantity of water evaporated per unit area per unit time from an idealized extensive free water surface under existing atmospheric conditions (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**, Shuttleworth, 1992). Actual evaporation is the evaporation that occurs in reality from a natural, nonidealized surface. Globally, evaporation from the land surface to the atmosphere amounts to 71 10^3 km^3 $year^{-1}$. The evaporation from the oceans is much larger, 42 810^3 km^3 $year^{-1}$ (German Advisory Council on Global Change, 1999).

Traditionally, actual evaporation has been estimated as the residual of the catchment water balance (e.g. Bosch and Hewlett, 1982). These studies are very useful to estimate annual evaporation (Zhang *et al.*, 2001, 2004), but suffer from accumulated errors in the residual term that represents evaporation. In contrast, micrometeorological techniques can measure the fluxes of evaporation directly, but these techniques need considerable maintenance, and often are biased towards producing good weather data. However, since 1995, a global network of flux towers (*see* **Chapter 40, Evaporation Measurement, Volume 1**) has emerged that can now be used to estimate evaporation from a variety of land-cover types (Baldocchi *et al.*, 2001).

Zhang *et al.* (2001, 2004) present an elegant analysis of data from 250 catchments worldwide with a "rational function" analysis of the main factors driving evaporation. They

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

suggest that the total evaporation of a catchment can be approximated by considering two factors only, atmospheric demand and water availability. Atmospheric demand is given by the maximum possible evaporation, that is, potential evaporation; water availability can be represented by precipitation. Under dry conditions, potential evaporation exceeds precipitation, and the actual evaporation of a catchment will approach the precipitation amount. Conversely, under wet conditions, water availability exceeds potential evaporation and actual evaporation will approach asymptotically potential evaporation. This approach builds on Budyko (1974) and Milly (1994) who suggested that the long-term water balance of a catchment is determined by the interaction of supply (precipitation) and demand (potential evaporation), mediated by soil moisture storage. Figure 1 shows results of Zhang *et al.* (2001), both from a simple model based on the above considerations for grass and forested catchments, and with catchment data. These results suggest that the scale of catchments such a simple approach works remarkably well, with a linear (1 : 1) relation between rainfall and evaporation for catchments that receive up to 500 mm year^{-1}, and an asymptotic relation (driven by evaporative demand and not by water availability) for catchments receiving more than 500 mm year^{-1}. The difference between forest and grassland originates in the high interception losses of forest that make up a considerable part of the total evaporative loss. Total evaporation from forest saturates at about 1400 mm year, while for grasslands this value is around 900 mm year^{-1}.

These values could in principle be compared to the direct measurements of evaporation such as quoted in Law *et al.* (2002). However, the values of evaporation measured by micrometeorological techniques usually refer to dry-canopy evaporation only (transpiration) and contain little evaporation during wet-canopy evaporation (interception). Consequently, total evaporation values as given in Law *et al.* (2002); Valentini (2003) should be treated with care, when interception losses are not explicitly treated. When this is taken into account, the annual evaporation from Fluxnet data for coniferous forests is 397 ($\pm$31) mm year^{-1}; for mixed evergreen and deciduous forest 386 ($\pm$18) mm year^{-1}; for deciduous broadleaf 512 ($\pm$69) mm year^{-1}; for grassland 494 ($\pm$104) mm year^{-1}, and for crops 666 ($\pm$67) mm year^{-1}. The values for forests are very close to the average of 350 mm that was suggested by Roberts (1983) as a good annual average estimate for transpiration of European forests. This indicates that the FLUXNET values indeed are closer to the annual transpiration rates, than to the total evaporation (transpiration plus interception). Grassland and crops have a higher dry-canopy evaporation than forest, because they are less coupled to the atmosphere and thus show less stomatal control (e.g. Shuttleworth and Calder, 1979).

The Fluxnet data can also be used to identify the main controls on evaporation. Wilson *et al.* (2002) show how the partitioning of energy, as expressed by the ratio of sensible heat to latent heat (Figure 2), the Bowen ratio (β), can be used to group different vegetation types in climate space. The graph shows the position of the individual sites with respect to the magnitude of the latent heat (evaporation) and sensible heat fluxes. In contrast to the

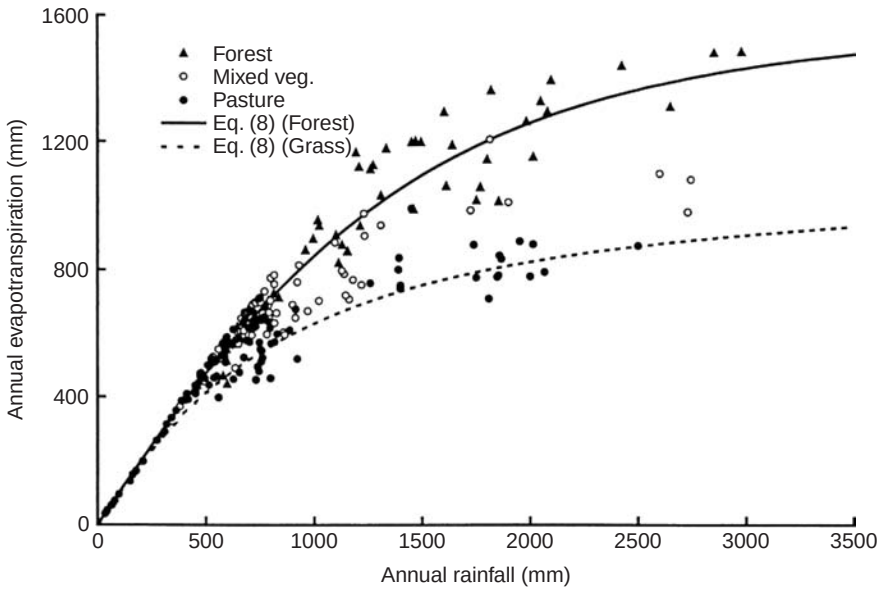


Figure 1 Relationship between annual precipitation and evaporation. The dashed line represents the best fit of a theoretical model for grass-only catchments, the solid line the best fit for forests (Reproduced from Zhang *et al.*, 2001 by permission of American Geophysical Union)

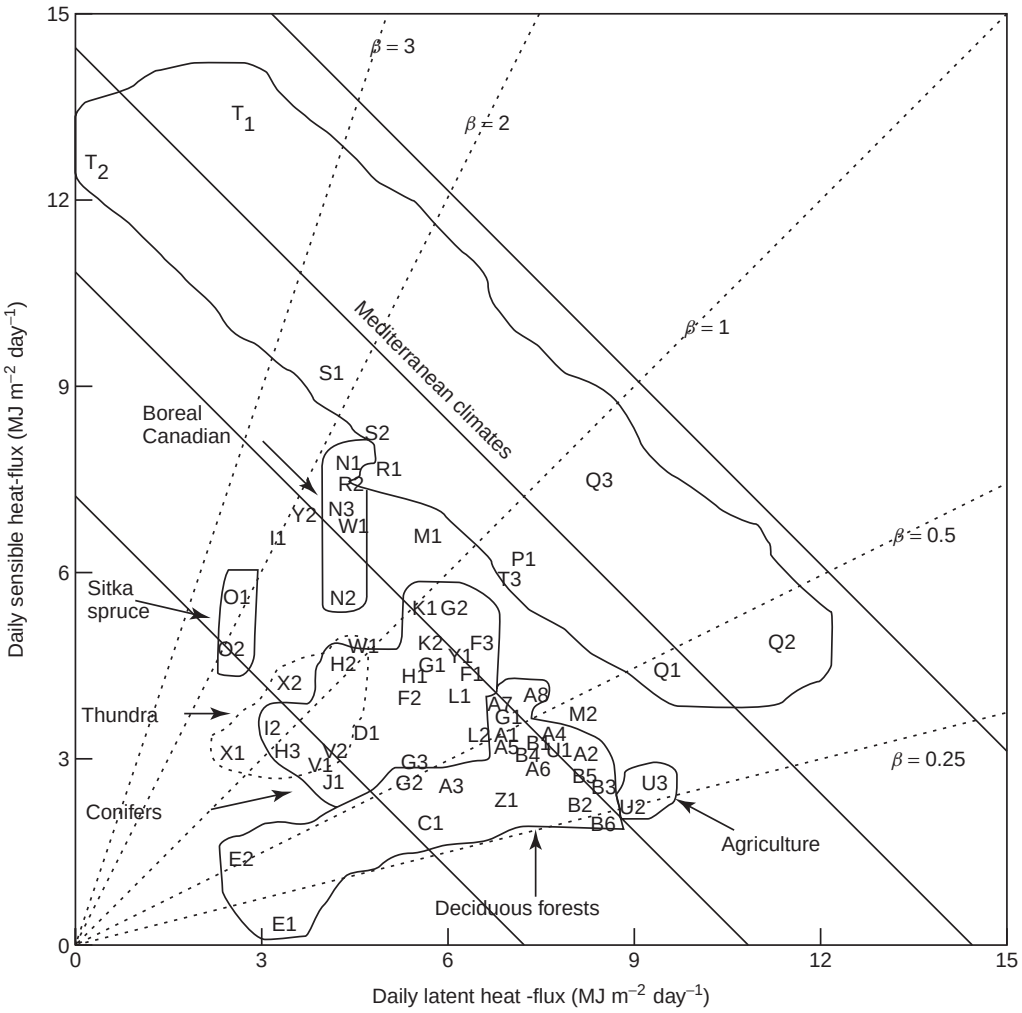


Figure 2 The daily cumulative sensible heat flux *versus* the daily cumulative latent heat flux between days 165 and 235 for the FLUXNET sites analyzed in Wilson *et al.* (2002). The letter and number codes refer to the sites in Table 1 (letter) and year of study at that site (number) of Wilson *et al.* (2002). Also shown are lines of constant Bowen ratio (dashed lines) and lines of constant total turbulent energy fluxes (solid diagonal lines). Enclosed circles denote subjective delineations between different vegetation types and climates (Reproduced from Wilson *et al.* (2002), by permission of American Geophysical Union)

annual rates discussed earlier, these data refer to the growing season only. Also shown are lines of constant available energy (the sum of latent and sensible heat) and lines of constant β-ratio. Moving from the lower left corner of the diagram to the upper right, the available energy increases, whereas moving from the lower right to the upper left, the value of the β-ratio increases. Low evaporation rates are found in Sitka spruce, tundras and boreal forest, high evaporation rates in deciduous forests and agriculture. Most of the forests have low β-ratios implying that most of the energy received is transmitted back to the atmosphere as sensible heat. Deciduous forests tend to have lower β-ratios and higher evaporation rates than coniferous forest. The average β-ratio at tundra sites appears to be close to one.

PREDICTING EVAPORATION

Evaporation at the earth's surface is constrained by the energy available to turn liquid water into its vapor form and by the capability of the surrounding air to transfer moisture away from the saturated surface. At the surface these constraints are best expressed through the energy balance equation and the transfer equations of latent and sensible heat:

$$K_{\downarrow}(1-\varsigma) + L_{\downarrow} - L_{\uparrow} = \lambda E + H + G + \Delta S \quad (1)$$

$$\lambda E = -K_{\nu}\frac{\delta q}{\delta z} \quad (2)$$

$$H = -\rho c_p K_h \frac{\delta T}{\delta z} \tag{3}$$

with $K_\downarrow$ the incoming short-wave radiation ($W\,m^{-2}$), ζ the short-wave albedo, $L_\downarrow$ and $L_\uparrow$ the incoming and outgoing long-wave radiation ($W\,m^{-2}$), λE the evaporation (latent heat flux, $W\,m^{-2}$), λ is the latent heat of evaporation, H the sensible heat flux ($W\,m^{-2}$), G the soil heat flux ($W\,m^{-2}$), ΔS the change in heat storage in biomass and atmosphere ($W\,m^{-2}$), K_v and K_h the transfer coefficients for water vapor and heat respectively, $\delta q/\delta z$ the vertical gradient in specific humidity from the evaporating surface to a reference height above the surface ($g\,kg^{-1}$), and similarly $\delta T/\delta z$ is the vertical gradient in temperature (°K), ρ is the density of air ($kg\,m^{-3}$) and c_p the specific heat of air ($J\,kg^{-1}K^{-1}$).

Penman (1948) was among the first to achieve the crucial combination of the energy balance equation with the transfer equations to arrive at an expression for actual evaporation. His original equation, however, contained an empirical wind function that replaced the transfer coefficients in equations 2 and 3. This wind function was difficult to generalize, and subsequently, Thom and Oliver (1977) provided a more physical basis based on considerations of aerodynamic transfer over rough surfaces by introducing an aerodynamic resistance. To eliminate the surface temperature from the equation combining the bulk discrete versions of equations 1a and 1c, Penman introduced the slope of the saturated specific humidity *versus* temperature curve evaluated at air temperature given by

$$s = \frac{q^*(T_s) - q^*(T_a)}{T_s - T_a} \tag{4}$$

where $q^*(T_s)$ is the saturated specific humidity at surface temperature T_s(°K), and $q^*(T_a)$ is the saturated specific humidity at air temperature T_a. (Penman used the slope of the saturated vapor pressure *versus* temperature curve, but in this paper specific humidity is used throughout.)

Predicting Evaporation from Vegetated Sources

Monteith (1965) took into account the control of vegetation on evaporation and described the transfer of water from the collective saturated surfaces in the plant stomatal cavities in the plant stomata to the air outside the leaves by a canopy scale resistance r_s. The resulting equation, carrying now both an aerodynamic and canopy resistance, is known as the *Penman–Monteith equation* and is arguably still the most advanced resistance model of evaporation used in hydrological practice today (Shuttleworth, 1992). For a mathematically precise and exact definition and brief historical overview of its development, see Raupach (2001). The most common form of the Penman–Monteith (PM) equation reads:

$$\lambda E = \frac{\Delta(Q_* - G) + \dfrac{\rho c_p D}{r_a}}{s + \dfrac{c_p}{\lambda}\left(1 + \dfrac{r_a}{r_s}\right)} \tag{5}$$

with Q_* the net radiation, consisting of the left-hand term of energy balance equation ($W\,m^{-2}$), D the specific humidity deficit (Note the difference between D, the specific humidity deficit and δq, the gradient in specific humidity, in equation 2.) of the air ($g\,kg^{-1}$) equal to q_* (T_a)-q_a, λ the latent heat of vaporization ($J\,kg^{-1}$). Two important variables appear in this equation that replace the transfer coefficients of the transfer equations of heat and moisture, the aerodynamic and surface resistance: r_a and r_s ($s\,m^{-1}$). (Only in the case of a full canopy cover, can the surface resistance be equal to the canopy resistance (e.g. Shuttleworth, 1978). For the derivation of the "big leaf" version of the PM model as presented here, this is not important. When the canopy cover is not full, and there is bare ground directly in contact with the overlying atmosphere, the surface resistance is not equal to the canopy resistance and approaches the reciprocal sum of the canopy and an assumed soil resistance.) A schematic showing the flux of latent heat encountering first the surface resistance and subsequently the aerodynamic resistance r_a, is given in Figure 1. The PM equation assumes that evaporation and sensible heat originate from the same source in the canopy. The main advantage of the Penman–Monteith equation is that the meteorological driving variables, wind speed, specific humidity deficit, and temperature are required only at a single level above the surface, obliterating the need for the notoriously difficult observation of surface values. The main stumbling block for the practical application of this equation is the estimation of the values for aerodynamic and surface resistance. When there is unlimited supply of water, the PM equation can be used to calculate potential evaporation (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1**). When the canopy is wet, the surface resistance equals zero and the PM equation can be used to calculate evaporation of intercepted rainfall (*see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). Although there is a wide range of empirical evaporation equations of which some are particular cases of the PM equation (see Brutsaert, 1982; Shuttleworth, 1992), the use of the PM equation is recommended because of its clear physical interpretation.

Aerodynamic Resistance

The aerodynamic resistance controls the rate of water vapor transfer away from the surface and is intuitively inversely proportional to the wind speed. It can be expressed for conditions of neutral stability as:

$$r_a = \frac{\ln^2\left[\dfrac{(z-d)}{z_0}\right]}{k^2 u(z)} \tag{6}$$

where d is the displacement height (m) and z_0 the aerodynamic roughness length (m), k von Karman's constant ($= 0.41$) and $u(z)$ the wind speed measured at height z ($\mathrm{m\,s^{-1}}$) (the aerodynamic resistance is mathematically the integral with height of the reciprocal of the transfer coefficients of equations 2 and 3 (with $K_v = K_h$) and is obtained by using the logarithmic wind profile). Corrections for non-neutral stability are given by Verma (1989). Typical values for forest are $5\,\mathrm{s\,m^{-1}}$, the values are generally an order of magnitude larger and for grassland approach $50\,\mathrm{s\,m^{-1}}$ (Shuttleworth, 1992; Verma, 1989; Stewart, 1988).

More detailed descriptions of turbulent transport can be used to replace the aerodynamic resistance. Raupach (1989), and subsequently Dolman and Wallace (1991); Finkele *et al.* (2003) applied this Lagrangian theory with success in evaporation models.

Canopy or Surface Resistance

The estimation of the surface resistance is more complicated. The surface resistance is not necessarily the collected reciprocal sum of the individual stomatal resistances ($1/r_s = \Sigma r_{stom}$), although Shuttleworth (1978) argues on theoretical grounds that the difference is negligibly small for most practical applications. Typical unstressed values for surface resistance are $50\,\mathrm{m\,s^{-1}}$. The surface resistance of vegetation shows a marked diurnal and seasonal variation (Dolman *et al.*, 1998) and requires careful estimation. Straightforward sensitivity analysis of the PM equation shows that for low crops, that is, grasslands, the estimation of the radiation part of the PM equation is important and that the estimation of the surface resistance is less important For tall crops such as forest, the estimation of surface resistance is more important as in that case the evaporation is dominated by the humidity deficit and surface resistance. It is thus important to develop adequate models of surface resistance.

In the absence of a full physiological understanding of stomatal behavior, two type of models have been developed. These models have more in common than might appear at first sight. The first is generally referred to as the stress function model and was first developed by Jarvis (1976) to describe leaf level stomatal behavior. It expresses the actual resistance as a minimum resistance (maximum conductance) that is being increased by a number of stress functions that describe the empirically observed collective behavior of the stomata in the form of the canopy resistance (Stewart, 1988) and it is important to note that the version given in equation 7 is the canopy scale version:

$$\frac{1}{r_c} = \frac{1}{r_{min}} f(K_\downarrow) f(D) f(T) f(\theta) \tag{7}$$

where f(i) is the stress function with reference to the variable i, θ the soil moisture deficit, that is, the relative amount of soil moisture compared to field capacity and r_{min} the minimum canopy resistance ($\mathrm{s\,m^{-1}}$) (this equation is often given in the form of conductances where $r_{min} = 1/g_{s,max}$ and g is the conductance). The form of these functions is given by Jarvis (1976) and Stewart (1988). For each vegetation type or application, these functions require derivation from measurements at canopy level or some way of scaling measurements of stomatal resistance to the canopy (Dolman *et al.*, 1991). Maximum canopy conductances for a number of vegetation types are given in Table 1. For comparison, the stomatal conductance is also shown. A slightly wider range of maximum conductances, obtained from eddy correlation data is given in Buchmann and Schulze (1999). Their values range from $7.1\,\mathrm{mm\,s^{-1}}$ for tundra, through $35\,\mathrm{mm\,s^{-1}}$ for temperate grassland to values of 20 to $25\,\mathrm{mm\,s^{-1}}$ for deciduous and evergreen broadleaf forests. Evergreen needleleaf forest have generally lower resistances (e.g. Buchmann and Schulze, 1999). The canopy conductance is to first order, ignoring radiation divergence within the canopy, equal to the amount of leaf area times the stomatal conductance. When radiation absorption in the canopy is taken into account the relationship becomes nonlinear as shown by Dolman *et al.* (1991). It is important to appreciate the difference between values of maximum conductance (minimum resistance) that are observed in the field and given by Kelliher *et al.* (1995) and Buchmann

Table 1 Maximum stomatal (leaf level $g_{s,max}$) and surface conductances (G^3_{max}) for a number of global vegetation types (from Kelliher *et al.,* 1995). Sample standard deviations are also given. The stomatal and surface resistance values are also given

Vegetation type	$g_{s,max}$ ($\mathrm{m\,ms^{-1}}$)	$r_{s,min}$ ($\mathrm{s\,m^{-1}}$)	$G_{s,max}$ ($\mathrm{mm\,s^{-1}}$)	$R_{s,min}$ ($\mathrm{s\,m^{-1}}$)
Temperate grassland	8.0 ± 4.0	125	17.0 ± 4.7	58
Coniferous forest	5.7 ± 2.4	175	21.2 ± 7.1	47
Eucalypt forests	5.3 ± 3.0	189	17.0	58.8
Temperate deciduous forest	4.7 ± 1.7	213	20.7 ± 6.5	48.3
Tropical rainforest	6.1 ± 3.2	164	13.0	76.9
Cereals	11.0	91	32.5 ± 10.9	30.8
Broadleaved herbaceous crops	12.2	82	30.8 ± 10.2	32.5

and Schulze (1999) and those that are obtained by using equation 7 and obtained by mathematically optimizing a maximum conductance against a set of observations. These latter values are theoretical maxima that are never attained, the observed values represent the highest possible value, given the fact that almost always some constraint is reducing the conductance and are thus generally lower than the optimized values.

More recently, a canopy resistance model was developed that is based on the observed linear relation between photosynthesis and conductance at leaf level (Ball *et al.*, 1987). Expressed here as a resistance this relation reads

$$\frac{1}{r_c} = a + \frac{bAh_s}{c_i} \tag{8}$$

where A is the carbon assimilation rate (μmol C m^{-2} s^{-1}) of the leaf, h_s the surface relative humidity just outside the leaf, c_i the internal CO_2 concentration in the stomatal cavities (μmol m^{-3}) and a and b are constants.

This equation needs to be scaled up through the canopy (if the stomatal scale version of equation 7 is used, this would equally apply to the Jarvis model). It is important to stress that the relation between assimilation rate and conductance, though intuitively appealing, is as empirical as the stress function approach in its use of the surface relative humidity to increase the resistance. In fact, this is exemplified in the approach of Leuning (1995) where the relative humidity is replaced by a stress function involving the vapor pressure deficit. In practice, this function has remarkably similar performance to the one used in the Jarvis–Stewart approach. The real appeal of using this A-g_s relation is twofold. First, the calculation of the assimilation rate can be based on purely biogeochemical reasoning. This, in principle, would avoid the necessity of recalibrating the model for each vegetation cover. Second, by linking the water exchange with the exchange of carbon, a more realistic description of the relation between growth and water use is obtained. However, the advantage of using of the A-g_s equation for practical hydrological purposes is still marginal, whereas in contrast, in climate models its use is increasing (Sellers *et al.*, 1992; Jacobs *et al.*, 1994; Cox *et al.*, 1998; Calvet *et al.*, 1998; Ronda *et al.*, 2001).

Evaporation from Bare Soil

Evaporation from bare soil can be a significant component of the water balance, particularly in semiarid environments (Wallace and Holwill, 1997). Soil evaporation can be described as a "two-stage process". The first stage is under the conditions where the available soil moisture is sufficient to meet the atmospheric demand. This happens immediately after rainfall or irrigation events. Soil evaporation under these conditions equals potential evaporation. Typically this stage last 1 to 2 days, although in some cases when evaporative demand is low and the soil contains a high amount of clay, this stage may last for 5 days. In the second stage, the amount of soil moisture has dropped and soil evaporation is no longer only restricted by evaporative demand but also by availability of moisture. In these conditions, the time change of soil moisture can be described as a desorption process with evaporation proportional to the square root of the time since the start of the process:

$$\lambda E_s = \tfrac{1}{2} D (t - t_0)^{-\frac{1}{2}} \tag{9}$$

with λE_s the soil evaporation, t the time and t_0 the time since the start of stage 2, D is a desorptivity (in units of W m^{-2} day^{-1} when evaporation is expressed as a heat flux or in mm day^{-1} when evaporation is expressed as water flux). The desorptivity is assumed constant for a particular soil type. It varies from a value of 2.1 for sandy loam with gravel, to a value of 5 mm day^{-1} for a clay loam soil (Kustas, 2002). Although the two-stage process describes soil evaporation at diurnal timescales, extension of the theory to (sub) hourly timescales is straightforward (Brutsaert and Chen, 1995; Porté-Agel *et al.*, 2000). The determination of the desorptivity coefficient is not without problems, as is the identification of the switch from stage 1 to 2 (Kustas, 2002).

The observed dependence of soil evaporation on available soil moisture suggests that a resistance approach may also be feasible that incorporates a dependence of soil surface resistance on soil moisture. Mahfouf and Noilhan (1991) review several such formulations. These approaches can be divided into so-called α and β approaches. In the α-approach, the saturated humidity in the soil pore space (compare equation 2) is adjusted by a factor α that may be related to soil matrix potential and takes into account that, averaged over a certain depth, the evaporation takes place from a nonsaturated surface. In the β-approach, the humidity in the pore space at the evaporation front is assumed to be saturated, and β is the ratio of an aerodynamic resistance and the sum of the aerodyamic and soil surface resistance. In general, the β-approach appears to yield better results. Relations that try to define the soil resistance as a function of soil moisture (e.g. van de Griend and Owe, 1994) range from linear to exponential (Kustas, 2002). A particular problem in defining these relations is the depth of the soil profile that needs to be considered. Also, approaches that relate the resistance to the depth of the soil surface to an evaporation front suffer from similar problems (Camillo and Gurney, 1986).

Evaporation from Mixed Surfaces

Most natural vegetation, perhaps with the sole exception of tropical rainforests, consists of a mixture of surfaces and soils. Particular pronounced examples can be found in

the semi-arid natural vegetation of savannahs and in the case of developing agricultural crops. In these cases, the sources and sinks of evaporation (the leaves) and sensible heat (mostly the dry soil) are different. A dry soil in such a case transforms the temperature and thereby the specific humidity deficit around the leaves by an extra sensible heat input from the soil. This increases the evaporation from the leaves compared to a situation where the soil releases no additional heat. The reverse, dampening of the evaporation, occurs when the soil is wet and releases latent heat through soil evaporation. Shuttleworth and Wallace (1985) and subsequently Dolman (1993), Huntingford (1995), and Daamen (1997) developed models to take this into account. They take the general form of

$$\lambda E_{\text{tot}} = C_{\text{c}} P M_{\text{c}} + C_{\text{s}} P M_{\text{s}} \tag{10}$$

where the total evaporation λE_{tot} is expressed as a linear sum of the contribution from the canopy and soil. The factors C_c and C_s are functions of the canopy and aerodynamic resistances in the pathways from soil and vegetation to the atmosphere. PM_c and PM_s represent the equivalent Penman–Monteith equations, which would apply to evaporation from a closed canopy and bare soil respectively. The solution to the coupled system involves the definition of a within canopy humidity deficit D_0 (Shuttleworth and Wallace, 1995) that determines the evaporation from the leaves and soil.

Models that go beyond the two-source approach and allow more than two components are described by, for example Verhoef and Allen (2000). Such a situation may occur, for instance, in a savannah where small trees are surrounded by a mixture of grasses and bare soil. Daamen and McNaughton (2000) compare the performance of three types of resistance configurations (Figure 3) to predict evaporation from six land surface types. They conclude that the two-source model (interactive model) is generally superior, but that the differences between the three versions are within $50\,\text{W}\,\text{m}^{-2}$, when the component surface resistances are approximately equal or both very large. When the ratio of the two resistances describing transport from the canopy air to the reference height to that from the component surface to the canopy air, is large, the component evaporation is closely coupled to the canopy air, and an interactive model performs better than either a patch or single surface model (PM model). Such conditions may occur for low overstory canopies (steppe, pasture), sheltered canopies such as orchards, or canopies with large leaf area index but narrow leaves (Daamen and McNaughton, 2000).

Actual Evaporation Based on Crop Factors

It has long been common practice to express actual evaporation as the product of a potential rate (reference crop evaporation, *see* **Chapter 41, Evaporation Modeling: Potential,**

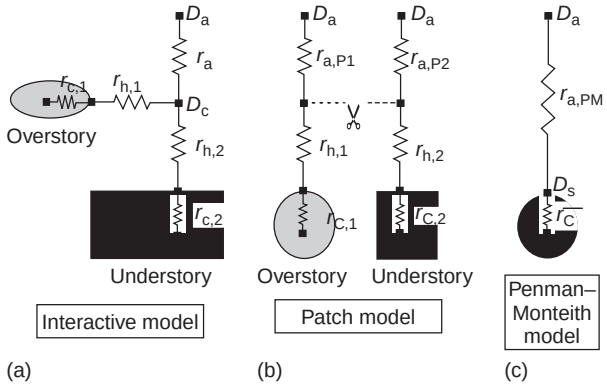


Figure 3 Evaporation (latent heat flux) and sensible heat flux from a vegetated surface to the atmosphere encounters a canopy resistance r_c and subsequently an aerodynamic resistance r_h (between a surface and the within canopy air) and r_a (from the canopy air to the reference height). Shown are three possible configurations of resistance networks to deal with mixtures of canopies and soil. Only in the interactive model (a) is the air ''sensed'' by the canopy dependent on the fluxes from the lower surface. In the two other configurations, they are either separated (b) or aggregated (c) (Reproduced from Daamen and McNaughton (2000), by permission of American Society of Agronomy)

Volume 1) and a reduction factor, which depends on the time of year and vegetation type (Doorenbos and Pruit, 1977). This can be expressed as

$$E_{\text{a}} = kE_{\text{p}} \tag{11}$$

where E_a and E_p are the actual and potential rates of evaporation ($\text{mm}\,\text{day}^{-1}$) and k a reduction or crop factor. It is easily shown (e.g. Wallace, 1995) that the crop factor is not a constant, but a complicated function of the surface (canopy) and aerodynamic properties of the crop and the climate in which they are derived. Wallace (1995) showed the crop factor to be a function of temperature, wind speed, crop height, and cover and rainfall. Although thus in appearance simple, and recently again recommended by the FAO (e.g. Allen *et al.*, 1998), the crop factor approach is conceptually not without its problems.

Actual Evaporation Based on Equilibrium Evaporation Concepts

It is useful at this stage again to define the Bowen ratio $\beta = \text{H}/\lambda\text{E}$. When air moves over a moist surface and gradients of temperature and specific humidity with height are small or become saturated with respect to moisture, Priestley showed that the Bowen ratio should approach a constant value of $\beta = s/\lambda$. Raupach (2001) defines equilibrium evaporation as when the ratio λE/A takes the value $\varepsilon/(1+\varepsilon)$ and $\beta = 1/\varepsilon$ with $\varepsilon = (\lambda/c_{\text{p}})\,\text{dq}_{\text{sat}}/\,\text{dT}$, with A the available energy ($\text{A} = \text{H} + \lambda\text{E}$) the ratio $\text{dq}_*/\,\text{dT}$ contains the

variations in latent and sensible heat content of saturated air and is similar to the more familiar s as used in the simplified analysis above (see also Brutsaert, 1982)) (Brutsaert, 1982). Combining this insight with the Penman–Monteith equation, with the second term above the nominator set to zero and with zero surface or canopy resistance ($r_s = 0$) as appropriate for a moist surface, yields the "equilibrium evaporation" (see also Brutsaert, 1982; Raupach, 2001):

$$\lambda E = \frac{s}{s + \frac{c_p}{\lambda}} Q_* \tag{12}$$

Equilibrium evaporation as defined by these authors refers to the lower limit of evaporation from a moist surface where the specific humidity deficit of the second term in the nominator of the PM equation has become zero as a result of contact of air with a moist surface over a very long fetch.

The second term in the nominator of the PM equation is seen to represent the departure from this equilibrium. Priestley and Taylor (1972) termed this departure from equilibrium evaporation α, and it is easily seen that α is only unity when the specific humidity deficit in the PM equation is zero, in other words when advection is negligible. That this is hardly ever the case proves the fact that most empirical values of α for short crops are of the order 1.2–1.3 (e.g. Brutsaert, 1982). For tall crops, Shuttleworth and Calder (1979) show convincingly that the equilibrium approach is not appropriate because the physiological control of the forest transpiration reduces α below a value of 1 in dry canopy conditions, while in wet canopy conditions large-scale advection and negative sensible heat fluxes (Stewart, 1977) may form an additional supply of energy and force α to be well above the value for short crops. This emphasizes the important point that for tall crops in particular, it is important to estimate dry and wet canopy evaporation separately.

Faith in the equilibrium evaporation concept by hydrologists is largely based on the belief that at larger regional scale, the atmospheric conditions override the surface control. A number of concepts have been derived that use the perceived dampening power of the atmosphere to estimate regional-scale evaporation (e.g. Bouchet, 1963; Morton, 1983). Although McNaughton and Jarvis (1991) show that at larger scale the feedback of the atmospheric boundary layer dampens the effects of surface controls – and thus makes precise estimation of the surface resistance less important – the physical basis of the Bouchet and Morton schemes remains doubtful (see de Bruin, 1983; McNaughton and Spriggs, 1989). Nevertheless, de Bruin showed that the feedback of the increasing atmospheric boundary-layer humidity during the day causes the regional surface conductance to vary less than if the feedback were neglected. When the surface resistance is moderately influenced by soil moisture stress, surface conditions are again important, whereas in the case of a completely dry surface, the boundary layer overrides surface control and a crude estimation of the surface resistance is sufficient. However, the de Bruin (1983) and McNaughton and Spriggs (1989) studies were restricted to a few clear days, with no cloud development or rainfall and require rather detailed observational input from the boundary layer, and clearly have no general practical application as yet. At the annual timescale, there may be promise in extending the complementary relationship along the lines discussed earlier in this paper (Zhang *et al.*, 2001) and incorporate not only the evaporation-potential evaporation feedback, but also the precipitation feedback on evaporation.

CONCLUSIONS

There has been remarkable progress in our understanding of actual evaporation in the last two decades, at least partly due to our increased ability to measure directly the fluxes of latent and sensible heat by micrometeorological techniques (*see* **Chapter 40, Evaporation Measurement, Volume 1**). At the same time, progress in modeling has generated a suite of evaporation models that can now be applied under various circumstances. To a large extent, the choice of a particular model will depend on the data availability and purpose. For simple applications, the crop factor approach may still work best; in hydrological and meteorological models, the more physically and physiologically based models are a more suitable option. These allow predictions to be made under various changing environmental factors, which the more empirical approaches can never realistically attain to predict.

At large scales, the relationships between evaporation, rainfall and available energy appear to be more straightforward than at smaller scales (Wilson *et al.*, 2002) and water use by vegetation in different climates more conservative (Law *et al.*, 2002). This feature can be exploited when predicting the effects of land-cover change on the hydrological balance (e.g. Zhang *et al.*, 2001, 2004). The challenge in the next few years for hydrometeorologists still is to match our micrometeorological estimates with the catchment scale observations. Only then, can process understanding at short time and spatial scales be used to explain the larger integral constraints we observe of catchment scale behavior.

REFERENCES

Allen R.G., Periera L.S., Raes D. and Smith M. (1998) *Crop Evapotranspiration – Guidelines for Computing Crop Water Requirements – FAO Irrigation and Drainage Paper 56*, Food and Agriculture Organization: Rome.

Baldocchi D.D., Falge E., Gu L., Olson R., Hollinger D., Running S., Anthoni P., Bernhofer Ch., Davis K., Fuentes J.,

et al. (2001) FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor and energy flux densities. *Bulletin of the American Meteorological Society*, **82**, 2415–2435.

Ball J.T., Woodrow I.E. and Berry J.A. (1987) A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. In *Progress in Photosynthesis Research: Proceedings of the Seventh International Congress on Photosynthesis*, Providence, Rhode Island, U.S.A., August 10–15, 1986, Biggins J. (Ed.), Vol. 4, Martinus Nijhoff Publishers: 221–224.

Bosch J.M. and Hewlett J.D. (1982) A review of catchement experiments to determine the effect of vegetation changes on water yield and evaporation. *Journal of Hydrology*, **55**, 3–23.

Bouchet R.J. (1963) Evapotranspiration réeale et potentielle, signification climatologique. *International Association of Hydrological Sciences. Proceedings Berkeley Symposium. Publication.*, **62**, 134–152.

Brutsaert W. (1982) *Evaporation into the Atmosphere*, Reidel Publishing Company: Dordrecht, p. 299.

Brutsaert W. and Chen D. (1995) Desorption and the Two Stages of Drying of Natural Tallgrass Prairie *Water Resour.Res.,*, **31**(5), 1305–1313.

Buchmann N. and Schulze E-D. (1999) Net CO2 and H2O fluxes from terrestrial ecosystems. *Global Biogeochemical Cycles*, **13**(3), 751–760.

Budyko M.I. (1974) *Climate and Life*, Academic Press: San Diego.

Calvet J-C., Noilhan J., Roujean J-L., Bessemoulin P., Cabelguenne M., Olioso A. and Wigneron J-P (1998) An interactive vegetation SVAT model tested against data from six contrasting sites. *Agricultural and Forest Meteorology*, **92**, 73–95.

Camillo P.J. and Gurney R.J. (1986) A resistance parameter for bare soil evaporation models. *Soil Science*, **141**, 95–105.

Cox P.M., Huntingford C. and Harding R.J. (1998) A canopy conductance and photosynthesis model for use in a GCM land surface scheme. *Journal of Hydrology*, **212–213**, 79–95.

Daamen C.C. (1997) A two-source model of surface fluxes for millet fieldsin Niger. *Agricultural and Forest Meteorology*, **83**, 205–183.

Daamen C.C. and McNaughton K.G. (2000) Modelling energy fluxes from sparse canopies and understoreys. *Agronomy Journal*, **92**, 837–847.

de Bruin H.A.R. (1983) A model for the Priestley Taylor parameter. *Journal of Applied Meteorology*, **22**, 572–578.

Dolman A.J. (1993) A multiple source land surface energy balance model for use in GCMs. *Agricultural and. Forest Meteorology*, **65**, 2–45.

Dolman A.J., Gash J.H.C., Roberts J.M. and Shuttleworth W.J. (1991) Stomatal and surface conductance of tropical rainforest. *Agricultural and Forest Meteorology*, **54**, 303–318.

Dolman A.J., Moors E.J., Elbers J.A. and Snijders W. (1998) Evaporation and surface conductance of three temperate forests in the Netherlands. *Annals of Forest Science*, **55**, 255–270.

Dolman A.J. and Wallace J.S. (1991) Lagrangian and K-theory approaches in modelling evaporation from sparse canopies. *Quarterly Journal of the Royal Meteorological Society*, **117**, 1325–1340.

Doorenbos J. and Pruit W.O. (1977) *Crop Water Requirements*, Irrigation and drainage paper no. 24, Food and Agriculture Organization: Rome.

Finkele K., Katzfey J.J., Kowalczyk E.A., McGregor J.L., Zhang M. and Raupach M.R. (2003) Modelling of the OASIS energy flux measurements using two canopy concepts. *Boundary Layer Meteorology*, **107**(1), 49–79.

German Advisory Council on Global Change (1999) *World in Transition: Ways Towards Sustainable Management of Freshwater Resources*, Springer Verlag: Berlin. p. 392.

Huntingford C., Allen S.J. and Harding R.L. (1995) An intercomparison of single and dual-source vegetation atmosphere transfer models applied to transpiration from Sahelian savannah. *Boundary-Layer Meteorology*, **74**, 397–418.

Jacobs C.M.J. (1994 Direct impac of atmospheric CO2 enrichment on regional transpiration PhD thesis, University of Wageningen, p. 179.

Jarvis P.G. (1976) The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophical Transactions of the Royal Society of London*, **273B**, 593–610.

Kelliher L.M., Leuning R., Raupach M.R. and Schulze E-D. (1995) Maximum conductances for evaporation from global vegetation types. *Agricultural and Forest Meteorology*, **73**, 1–16.

Kustas W.P. (2002) Evaporation. In *Encyclopaedia of Soil Science: 524 – 530*, Lal R. (Ed). Dekker Publishing: p. 1450, DOI: 10.1081/E-ESS-120001671.

Law B.E., Falge E., Gu L., Baldocchi D.D., Bakwin P., Berbigier P., Davis K., Dolman A.J., Falk M. Fuentes J.D. *et al.* (2002) Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation. *Agricultural and Forest Meteorology* **113**, 97–120.

Leuning R. (1995) A critical appraisal of a combined stomatal–photosynthesis model for C-3 plants. *Plant, Cell Environment*, **18**, 357–364.

Mahfouf J.F. and Noilhan J. (1991) Comparative study of various formulations of evaporation from bare soil using in situ data. *Journal of Applied Meteorology*, **30**, 1354–1365.

McNaughton K.G. and Jarvis P.G. (1991) Effects of spatial scale on stomatal control of transpiration. *Agricultural and Forest Meteorology*, **54**, 279–302.

McNaughton K. and Spriggs T.W. (1989) An evaluation of the Priestley-Taylor equation and the complementary relationship using results from a mixed layer model of the convective boundary layer. In *Estimation of Areal Evapotranspiration*, Black T.A., Spittlehouse D.L., Novak M.D. and Price D.T. (Eds.), IAHS Publication no. 177: Wallingford, pp. 89–104.

Milly P.C.D. (1994) Climate, soil water storage and the average annual water balance. *Water Resources Research*, **30**, 2143–2156.

Monteith (1965) Evaporation and environment. In *The State and Movement of Water in Living Organisms*, Fogg G.E. (Ed.), Cambridge University Press: Cambridge, pp. 205–234.

Morton F.I. (1983) Operational estimates of areal evaporation and their significance to the science and practice of hydrology. *Journal of Hydrology*, **66**, 1–76.

Penman H.L. (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London*, **A193**, 120–146.

Porté-Agel P.M., Cahill A.T. and Gruber A. (2000) Mixture of time scales in evaporation: desorption and self-similarity of energy fluxes. *Agronomy Journal*, **92**,: 832–836.

Priestley C.H.B. and Taylor R.J. (1972) On the assessment of surface heat flux and evaporation using large scale parameters. *Monthly Weather Review*, **100**, 81–92.

Raupach M.R. (1989) A practical Lagrangian method for relating concentrations to source distributions in vegetation canopies. *Quarterly Journal of the Royal Meteorological Society*, **115**, 609–632.

Raupach M.R. (2001) Combination theory and equilibrium evaporation. *Quarterly Journal of the Royal Meteorological Society*, **127**, 1149–1181.

Roberts J.M. (1983) Forest transpiration: a conservative hydrological process? *Journal of Hydrology*, **66**, 133–141.

Ronda R.J., De Bruin H.A.R. and Holtslag A.A.M. (2001) Representation of the canopy conductance in modeling the surface energy budget for low vegetation. *Journal of Applied Meteorology*, **40**, 1431–1444.

Sellers P.J., Berry J.A., Collatz G.J., Field C.B. and Hall F.G. (1992) Canopy reflectance photosynthesis and transpiration III. A reanalysis using enzyme kinetics. And electron transport models of leaf physiology. *Remote Sensing of Environment*, **42**, 187–216.

Shuttleworth W.J. (1978) A simplified one dimensional theoretical description of the vegetation atmosphere interaction. *Boundary-Layer Meteorology*, **14**, 3–17.

Shuttleworth W.J. and Calder I.R. (1979) Has the Priestley-Taylor equation any relevance to forest evaporation? *Journal of Applied Meteorology*, **18**, 638–646.

Shuttleworth W.J. and Wallace J.S. (1985) Evaporation from sparse crops: and energy combination theory. *Quarterly Journal of the Royal Meteorological Society*, **111**, 839–855.

Shuttleworth W.J. (1992) Evaporation. In *Handbook of Hydrology*, Maidmentt D.R. (Ed.), McGraw Hill: New York, pp. 4.1–4.53.

Stewart J.B. (1977) Evaporation from the wet canopy of a pine forest. *Water Resources Research*, **13**, 915–921.

Stewart J.B. (1988) Modelling surface conductance of pine forest. *Agricultural and Forest Meteorology*, **30**, 111–127.

Thom A.S. and Oliver H.R. (1977) On Penman's equation for estimating regional evaporation. *Quarterly Journal of the Royal Meteorological Society*, **193**, 345–357.

van de Griend A.A. and Owe M. (1994) Bare soil surface resistance to evaporation and vapor diffusion under semiarid conditions. *Water Resources Research*, **30**, 181–188.

Valentini R. and Matteuci G. (Eds.) (2003) *Factors Controlling Forest Atmosphere Exchange of Water, Energy and Carbon In Fluxes of Carbon, Water and Energy of European Forests. Ecological Studies*, Vol. 163, Springer Verlag.

Verhoef A. and Allen S.J. (2000) A SVAT scheme describing energy and CO2 fluxes for multi-component vegetation: calibration and test for a Sahelian savannah. *Ecological Modelling*, **127**, 245–267.

Verma S.B. (1989) Aerodynamic resistances to transfers of heat, mass and momentum. In *Estimation of Areal Evapotranspiration*, Black T.A., Spittlehouse D.L., Novak M.D. and Price D.T. (Eds.), IAHS Publ. 177: Wallingford, pp. 13–20.

Wallace J.S. (1995) Calculation evaporation: resistance to factors. *Agricultural and Forest Meteorology*, **73**, 353–366.

Wallace J.S. and Holwill C.J. (1997) Soil evaporation from tiger-bush in south-west Niger. *Journal of Hydrology*, **188 – 189**, 426–442.

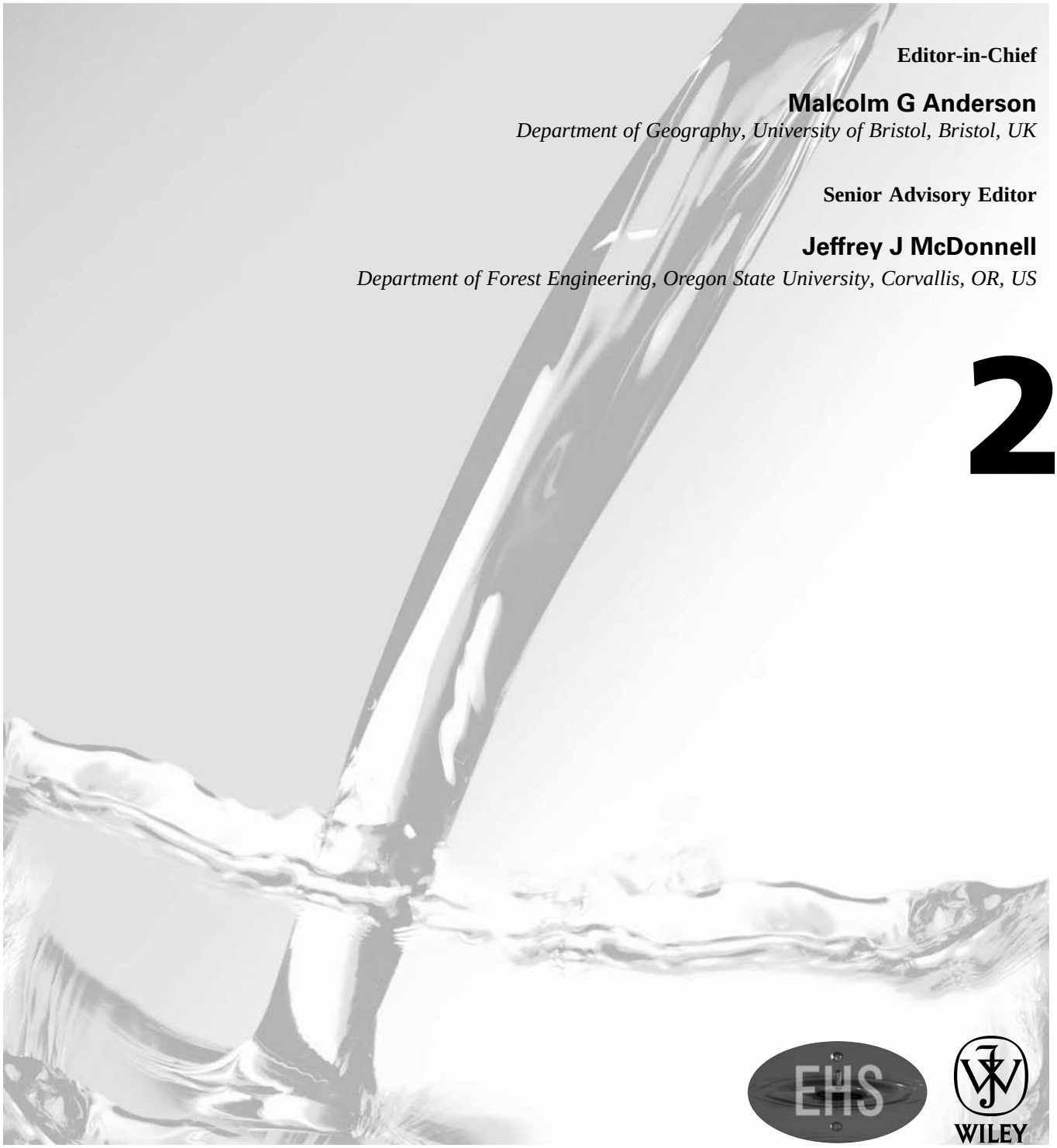
Wilson K.B., Baldocchi DennisD, Aubinet Marc, Berbigier Paul, Bernhofer Christian, Dolman Han, Falge Eva, Field Chris, Goldstein Allen, Granier Andre, *et al.* (2002) Energy partitioning between latent and sensible heat flux during the warm season at FLUXNET sites. *Water Resources Research*, **38**(12), 1294, doi:10.1029/2001wr000989.

Zhang L., Hicke K., Dawes W.R., Chiew F.H.S., Western A.W. and Briggs P.R. (2004) A rational function approach for estimating mean annual evapotranspiration. *Water Resources Research*, **40**, W02502. doi:10.1029/2003WP002710.

Zhang L., Dawes W.R. and Walker G.R. (2001) Response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resources Research*, **37**(3), 701–708.

Encyclopedia of
Hydrological Sciences

Encyclopedia of
Hydrological Sciences



Editor-in-Chief

Malcolm G Anderson
Department of Geography, University of Bristol, Bristol, UK

Senior Advisory Editor

Jeffrey J McDonnell
Department of Forest Engineering, Oregon State University, Corvallis, OR, US

2

WILEY

PART 5

Remote Sensing

46: Principles of Radiative Transfer

MATTHIAS DRUSCH[1] AND SUSANNE CREWELL[2]

[1]*European Centre for Medium-Range Weather Forecasts, Reading, UK*
[2]*Munich University, Munich, Germany*

The article gives an introduction to classical vector radiative transfer theory (RTT). It comprises a brief summary of the fundamental quantities in RTT and the corresponding definitions. Based on these quantities, the transfer equation for radiation will be introduced in its basic form for a plane-parallel, horizontally homogeneous atmosphere. Polarization and the Stokes vector are introduced, which extend the scalar radiative transfer equation to the more general form of the vector radiative transfer equation. Scattering processes and interactions with the surface are discussed within the mathematical framework given in the first part of the paper. As an example, the theoretical concept presented is then applied to passive microwave remote sensing. A solution for the radiative transfer equation, which is commonly used for the retrieval of atmospheric quantities (e.g. water vapor) and land surface properties (e.g. soil moisture), is derived based on approximations and simplifications. More general approaches to solve the radiative transfer equation including multiple scattering are described and discussed in a separate section.

INTRODUCTION

Radiative transfer theory (RTT) describes the interaction between matter and electromagnetic radiation. Electromagnetic waves travel in vacuum and air at the speed of light at $\sim 3 \times 108\,\mathrm{ms}^{-1}$. The electromagnetic spectrum covers gamma rays, X rays, ultraviolet, visible, infrared, microwaves, television signals, and radio waves as shown in Figure 1. Following Planck's law, solar radiation can be described as radiation emitted from a black body at a temperature of about 6000 K. Consequently, the solar component comprises radiation from the Gamma rays to the infrared spectral range. Only ~0.4% of the energy of solar radiation is emitted at wavelengths above 5 μm. The irradiance at the top of the atmosphere (solar constant S_0) is $1366\,\mathrm{Wm}^{-2} \pm 3\,\mathrm{Wm}^{-2}$ depending on the solar-earth distance and natural fluctuations of the sun's activity (Lean and Rind, 1998). Consequently, the energy available for the earth-atmosphere system is:

$$E_{\mathrm{sol},\downarrow} = \pi r_E^2 (1-\alpha) S_0 \tag{1}$$

with the planetary albedo α. This energy is absorbed by the system and reemitted following Stefan–Boltzmann's law:

$$E_{\mathrm{ter},\uparrow} = 4\pi r_E^2 \sigma_B T_E^4 \tag{2}$$

with the blackbody temperature of the earth T_E, the radius of the Earth r_E and the Stefan–Boltzmann constant σ_B. Using equations (1) and (2), a blackbody temperature for the atmosphere-earth system of ~255 K is obtained. At this temperature, a blackbody emits only 0.4% of its radiation at wavelengths below 5 μm (Goody and Yung, 1961). As a consequence, solar and terrestrial radiation can be treated independently. In equations (1) and (2) radiation has been integrated over all wavelengths and angles. In addition, T_E and α are "mean" quantities, which comprise the atmosphere (clouds, gaseous profiles, etc.), ocean (ice, foam, algae, etc.), and land surfaces (vegetation, soil texture, snow, etc.). For a comprehensive description of the interaction of electromagnetic waves and the earth-atmosphere system in hydrological applications and numerical modelling (including climate models, numerical weather prediction (NWP) models, soil-vegetation-atmosphere transfer

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

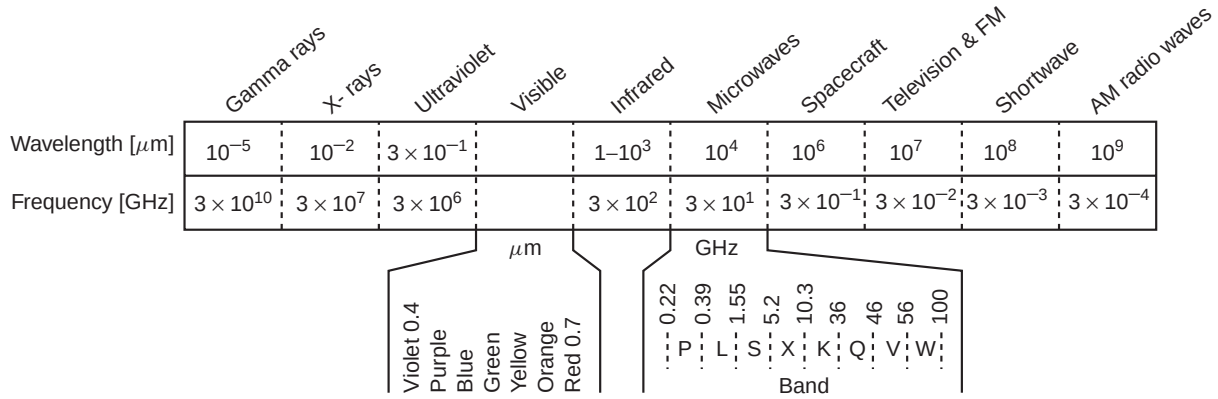


Figure 1 The electromagnetic spectrum (after Liou, 2002; Ulaby *et al.*, 1982)

(SVAT) models, physical and chemical cloud models, vegetation growth models, etc.), the RTT has to be much more complex.

Remote sensing is a second field of research, which is strongly related to RTT. Passive remote sensing techniques make use of the interaction of natural radiation with the variable of interest while active methods employ artificial sources of radiation such as lasers or radio wave sources. Retrieving geophysical parameters from radiation measurements requires RTT unless purely statistical methods are used.

In general, the theoretical and computational work on RTT can be divided in forward (or direct) and backward (or inverse) applications. In forward modelling studies, the geophysical parameters characterizing the earth-atmosphere system and the boundary conditions for the radiation sources are prescribed and the radiation field is computed for the given environment. Numerical climate or NWP modelling are typical examples of forward applications. For the inverse problem, the radiation field is known and it is aimed to retrieve the parameters characterizing the atmosphere and/or the surface. Many of the remote sensing applications have to address the inverse problem. Since analytical solutions for the inverse problem do not exist, a large number of statistical and physical retrieval methods have been developed. Often, variational methods assume a certain set of geophysical parameters and the computed radiances are compared with the observations. In the consecutive iteration steps, the parameters are modified to match the observations taking into account the observation and model errors.

Even if we limit ourselves to forward applications, it is impossible to provide a complete introduction to RTT for the most relevant physical processes at the visible to microwave spectral range within the framework of an article. For the basic laws in RTT, for example, Planck's law, Wien's law, Stefan–Boltzmann law and Kirchhoff's law, the reader is referred to more general textbooks, for example, Salby (1996). The focus of this paper is the classical vector RTT, since it has found extensive applications in research and operations. An overview of radiative transfer theory requires the use of mathematical expressions. The mathematics presented in this article partly exceeds the formalism presented in more general textbooks, for example, Salby (1996). Compared to specific books on radiative transfer (e.g. Chandrasekhar (1960), Liou (2002)) the number of equations is small and the corresponding level of complexity is low. However, the basic equations and theory presented in this article allow the reader to cope with more specific scientific articles addressing radiative transfer, remote sensing, engineering, and biophysical applications. A further reduction of the number of equations would not lead to a better understanding of the topic.

The first two sections comprise a brief summary of the fundamental quantities in RTT and the corresponding definitions. Based on these quantities, the transfer equation for radiation will be introduced in its basic form for a plane parallel, horizontally homogeneous atmosphere. This derivation can be found in many articles and textbooks on remote sensing and radiative transfer (e.g. Ulaby *et al.*, 1982; Liou, 2002; Chandrasekhar, 1960). In the section "Polarization and vector radiative transfer equation", polarization and the Stokes vector will be introduced, which extends the scalar radiative transfer equation to the general form of the vector radiative transfer equation. The sections "Scattering and phase functions" and "Boundary conditions – the surface" address scattering processes and interactions with the surface, respectively. Based on the theoretical framework presented in the first six sections, the radiative transfer equation is simplified and solved for passive microwave applications. More general approaches to solve the radiative transfer equation are presented in the section "Solutions".

DEFINITIONS

In a medium that interacts with radiation at a specific frequency ν, the specific intensity I_ν can vary from point to point and with the direction of propagation at each point. Following Chandrasekhar (1960), the specific (or monochromatic) intensity can be defined as:

$$I_\nu \equiv I_\nu(x, y, z; l, m, n; t) \tag{3}$$

where (x,y,z) and (l,m,n) define the point and the direction cosines and t indicates the temporal dependence of intensity. The specific intensity is given in units of $\mathrm{Wm^{-2}\,sr^{-1}\,Hz^{-1}}$ and is often referred to just as intensity (or radiance) for simplicity. The integration over the entire hemisphere gives the flux density F_ν also called *irradiance*, which is used in energetical studies. Integration over an infinitesimal frequency range $\mathrm{d}\nu$, the directions defining an element of the solid angle $\mathrm{d}\omega$ ($\mathrm{d}\omega = \mathrm{d}\sigma/r^2 = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$) and a time interval $\mathrm{d}t$, leads to the amount of energy dE_ν transported through the cross-section $\mathrm{d}\sigma$ (Figure 2):

$$\mathrm{d}E_\nu = I_\nu \cos\theta\,\mathrm{d}\nu\,\mathrm{d}\sigma\,\mathrm{d}\omega\,\mathrm{d}t \tag{4}$$

If the intensity I_ν at a given point is independent of direction, the radiation field is called *isotropic*. In this case, the relation $F_\nu = \pi I_\nu$ holds. Such a radiation field, which is characterized by constant intensities at each point, is said to be homogeneous. For practical applications in hydrology and atmospheric sciences, the medium (e.g. atmosphere, vegetation canopy, or the soil) that interacts with the radiation is often characterized by stratified plane-parallel layers with invariant physical properties. In a spherical coordinate system where z is the height and θ and ϕ the polar and azimuthal angles, the corresponding specific intensity can be written as:

$$I_\nu \equiv I_\nu(z, \theta, \phi; t) \tag{5}$$

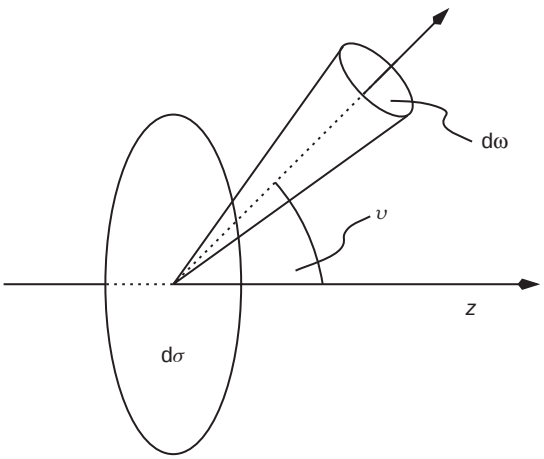


Figure 2 Definition of the specific intensity

SCALAR RADIATIVE TRANSFER EQUATION

Radiation characterized by a specific intensity I_ν passing a volume, defined by its thickness $\mathrm{d}s$ and a unit cross-section, will be weakened by an amount $\mathrm{d}I_\nu$. This extinction is caused by interaction with matter, namely *absorption and/or scattering*. Introducing the mass extinction, absorption, and scattering coefficients σ_e', σ_a', and σ_s' we can write:

$$\sigma_e' = \sigma_a' + \sigma_s' \tag{6}$$

The corresponding volume coefficients σ_e, σ_a, and σ_s can be obtained by multiplication with the density of the medium. Consequently, the amount of radiation absorbed and/or scattered by the matter in a volume $\mathrm{d}\sigma\,\mathrm{d}s$ for a solid angle element $\mathrm{d}\omega$ in a given time interval $\mathrm{d}t$ can be formulated as:

$$\sigma_e\,\mathrm{d}s\,I_\nu\,\mathrm{d}\nu\,\mathrm{d}\sigma\,\mathrm{d}\omega\,\mathrm{d}t \tag{7}$$

On the other hand, radiation intensity can be increased within the volume by contributions from scattering and emission. This source term can be described by introducing the volume emission coefficient j_ν

$$j_\nu\,\mathrm{d}s\,\mathrm{d}\nu\,\mathrm{d}\sigma\,\mathrm{d}\omega\,\mathrm{d}t \tag{8}$$

The change of the radiation intensity propagating along a path $\mathrm{d}s$ caused by extinction (7) and sources (8) expressed in specific intensity is therefore:

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}s}\,\mathrm{d}s\,\mathrm{d}\nu\,\mathrm{d}\sigma\,\mathrm{d}\omega\,\mathrm{d}t = -\sigma_e\,\mathrm{d}s\,I_\nu\,\mathrm{d}\nu\,\mathrm{d}\sigma\,\mathrm{d}\omega\,\mathrm{d}t + j_\nu\,\mathrm{d}s\,\mathrm{d}\nu\,\mathrm{d}\sigma\,\mathrm{d}\omega\,\mathrm{d}t \text{ or}$$

$$\frac{\mathrm{d}I_\nu}{\sigma_e\,\mathrm{d}s} = -I_\nu + \frac{j_\nu}{\sigma_e} \tag{9}$$

The emission coefficient j_ν can be written as the sum of two source terms. Radiation in the direction defined through the angles θ and ϕ is due to scattering processes $j_{\nu s}$ and emission from the medium in local thermodynamic equilibrium, $j_{\nu a,}$. The first term may be expressed as:

$$j_{\nu s}(\theta, \phi) = \sigma_s \frac{1}{4\pi} \int_0^\pi \int_0^{2\pi} p(\theta, \phi; \theta', \phi') I_\nu(\theta', \phi') \sin\theta'\,\mathrm{d}\theta'\,\mathrm{d}\phi' \tag{10}$$

In equation (10) the normalized phase function $p(\theta, \phi, \theta', \phi')$ is introduced. It describes how much radiation is scattered from incidence direction characterized by angles (θ',ϕ') into observation direction given by (θ,ϕ). Perfect scattering in the absence of any absorption requires:

$$\frac{1}{4\pi} \int_0^\pi \int_0^{2\pi} p(\theta, \phi; \theta', \phi')\,\mathrm{d}\theta'\,\mathrm{d}\phi' = 1 \tag{11}$$

Under the assumption of local thermodynamic equilibrium the thermal emission is given through Kirchhoff's law and the Planck function:

$$j_{va} = \sigma_a B_\nu(T) \tag{12}$$

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\frac{1}{e^{h\nu/k_B T} - 1} \tag{13}$$

where h and k_B are the Planck and Boltzmann constants and T the temperature. Combining equations (10) and (12) in (9) gives the scalar radiative transfer equation:

$$\frac{\mathrm{d}I_\nu}{\sigma_e\,\mathrm{d}s} = -I_\nu + \omega_0\frac{1}{4\pi}\int_0^\pi\int_0^{2\pi} p(\theta,\phi;\theta',\phi')I_\nu(\theta',\phi') \sin\theta'\,\mathrm{d}\theta'\,\mathrm{d}\phi' + (1-\omega_0)B_\nu(T) \tag{14}$$

with ω_0 the single scattering albedo defined as:

$$\omega_0 = \frac{\sigma_s}{\sigma_s + \sigma_a} \tag{15}$$

It has to be noted that equation (14) is valid for a specific frequency only. Therefore, the properties of the medium may be frequency dependent as well. RTT in which the radiation directly depends on frequency and where conversion of radiation between different frequencies takes place will not be discussed. Active remote sensing techniques such as, Raman lidar may need other forms of the radiative transfer equation. Methods to solve equation (14) are discussed in the section "Solutions".

POLARIZATION AND VECTOR RADIATIVE TRANSFER EQUATION

Equation (14) describes the radiative transfer for the specific intensity I_ν. Another source of information in remote sensing is the polarization of radiation. In order to model polarization effects due to particle scattering and/or rough surface reflection, we have to extend equation (14) to the full vector equation.

Wave Formalism

Elliptically polarized electromagnetic waves are described by the time-dependent electric field $\boldsymbol{E}$. Following Chandrasekhar (1960), Liou (1992) the field components can be projected to two perpendicular vectors in the plane orthogonal to the direction of propagation:

$$\boldsymbol{E}(t) = \boldsymbol{E}_l(t) + \boldsymbol{E}_r(t) = a_l(t)\cdot\boldsymbol{e}_l + a_r(t)\cdot\boldsymbol{e}_r \tag{16}$$

$\boldsymbol{e}_l$ and $\boldsymbol{e}_r$ are orthogonal unit vectors parallel and perpendicular to the scattering/reflection plane and define the local coordinate system as shown in Figure 3. In this system, the amplitudes and the complex electric field components can be expressed using the circular frequency ω_c, the wave number k, four constants (a_l, a_r, ε_l, ε_r), and the phase lag $\rho = \varepsilon_l - \varepsilon_r$ between both components of the wave:

$$\begin{aligned} a_l &= a_l^0 \sin(\omega_c t - \varepsilon_l) \\ a_r &= a_r^0 \sin(\omega_c t - \varepsilon_r) \\ E_l &= a_l^0 e^{-i\rho} e^{-kz+i\omega_c t} \\ E_r &= a_r^0 e^{-i\rho} e^{-kz+i\omega_c t} \end{aligned} \tag{17}$$

Intensity Formalism

The Stokes vector $\boldsymbol{I} = (I, Q, U, V)$ is defined by the components of the electromagnetic field (e.g. Chandrasekhar, 1960; Liou, 2002). It can be expressed with the real amplitudes, the phase difference and the angles β and χ (Figure 3):

$$\begin{aligned} I &= a_l^2 + a_r^2 = I_l + I_r \\ Q &= a_l^2 - a_r^2 = I_l - I_r = I\cos 2\beta\cos 2\chi \\ U &= 2a_l a_r\cos\rho = I\cos 2\beta\sin 2\chi \\ V &= 2a_l a_r\sin\rho = I\sin 2\beta \end{aligned} \tag{18}$$

The first Stokes parameter gives the total intensity, Q indicates the degree of linear polarization, U describes the plane of polarization and the ellipticity is given by V. For a

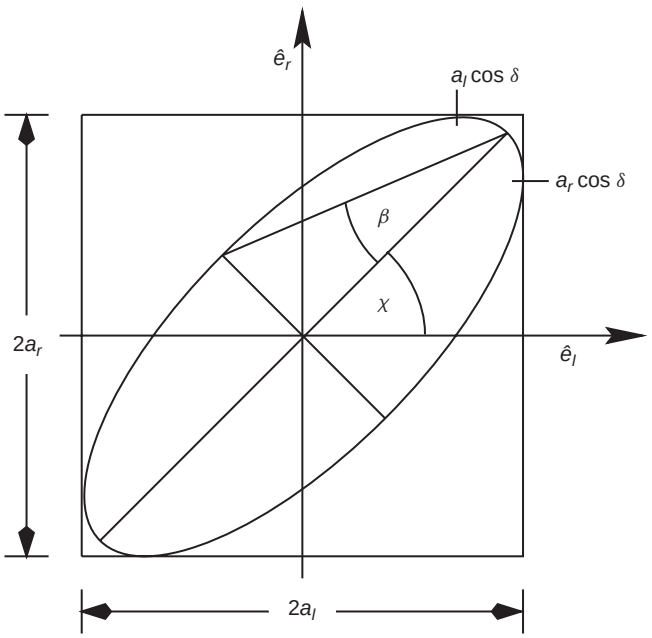


Figure 3 Illustration of an elliptically polarized electromagnetic wave

single wave and fully polarized light $I^2 = Q^2 + U^2 + V^2$ is valid. Natural radiation consists of the superposition of many different waves and contains a mixture of polarized and unpolarized light. In this case, the relation

$$I^2 \geq Q^2 + U^2 + V^2 \tag{19}$$

is valid. The degree of polarization is then given through:

$$P = \frac{\sqrt{Q^2 + U^2 + V^2}}{I} \tag{20}$$

Transformation of Stokes Vectors and Global Coordinates

Equations (18) describe the Stokes vector in local coordinates, which are used to describe single scattering processes. The radiative transfer equation refers to a Cartesian grid for the position and spherical coordinates for directions. Scattering processes change the direction of propagation and therefore a transformation between the local system, in which the scattering process is described, and the global coordinate system for the radiative transfer calculations is required. The geometry is illustrated in Figure 4. Assuming that the incident beam P_1 is scattered at the origin O to P_2, the local coordinate system is defined in the plane $P_1\ OP_2$. Let $\boldsymbol{R}(\Theta)$ be the phase function and Θ the scattering angle in the local system. In order to apply the phase function the Stokes vector of the incident light has to be transformed from the directions to which the polarization refers in global coordinates (plane $P_1 OZ$) to the local coordinates ($P_1 OP_2$). After performing the scattering process, a transformation from $P_1 OP_2$ to $P_2 OZ$ is necessary. The clockwise rotation angles for both rotations are $(-\iota_1)$ and $(\pi - \iota_2)$, respectively. The rotation transformation matrix for the Stokes vector reads:

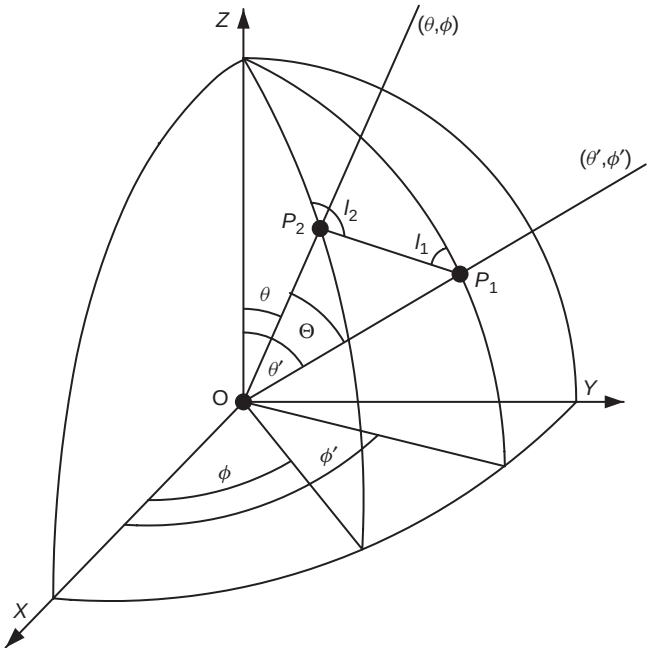


Figure 4 Directions of incoming and outgoing radiation with notation of angles

$$\boldsymbol{L}(\phi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos^2\phi & \sin^2\phi & 0 \\ 0 & \sin^2\phi & \cos^2\phi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{21}$$

The scattered beam is computed from:

$$\boldsymbol{I}(\theta, \phi) = \boldsymbol{L}(\pi - \iota_2)\boldsymbol{R}(\cos\Theta)\boldsymbol{L}(-\iota_1)\boldsymbol{I}(\theta', \phi') \tag{22}$$

The phase function $\boldsymbol{P}(\theta,\phi;\theta',\phi')$ is transformed from the local system to the global system by:

$$\boldsymbol{P}(\theta, \phi; \theta', \phi') = \boldsymbol{L}(\pi - \iota_2)\boldsymbol{R}(\cos\Theta)\boldsymbol{L}(-\iota_1) \tag{23}$$

For a set of discrete angles $(\theta,\phi;\theta',\phi')$, the angles ι_1 and ι_2 can be calculated with $\mu = \cos\theta$ and $\mu' = \cos\theta'$:

$$\cos\iota_1 = \frac{-\mu + \mu'\cos\Theta}{\pm(1 - \cos^2\Theta)^{1/2}(1 - \mu'^2)^{1/2}} \tag{24}$$

and

$$\cos\iota_2 = \frac{-\mu' + \mu\cos\Theta}{\pm(1 - \cos^2\Theta)^{1/2}(1 - \mu^2)^{1/2}} \tag{25}$$

The plus and minus signs in equations (24) and (25) are used for $\pi < \phi - \phi' < 2\pi$ and $0 < \phi - \phi' < \pi$, respectively. The scattering angle Θ, which is given in local coordinates, transforms to the global system through:

$$\begin{aligned} \cos\Theta &= \mu\mu' + (1 - \mu^2)^{1/2}(1 - \mu'^2)^{1/2}\cos(\phi - \phi') \\ &= \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi') \end{aligned} \tag{26}$$

Vector Radiative Transfer Equation

Replacing the intensities in equation (14) by Stokes vectors leads to the Vector Radiative Transfer Equation (VRTE). As a consequence, scattering and extinction coefficients and the phase function become matrices. Following equation (14) the VRTE can be written in the general form:

$$\begin{aligned} \frac{\mathrm{d}^3\boldsymbol{I}(x, y, z, \theta, \phi)}{\frac{1}{\eta}\mathrm{d}x\frac{1}{\delta}\mathrm{d}y\frac{1}{\mu}\mathrm{d}z} &= -\boldsymbol{\sigma_e}(x, y, z, \theta, \phi)\boldsymbol{I}(x, y, z, \theta, \phi) \\ &\quad + \boldsymbol{\sigma_a}(x, y, z, \theta, \phi)\boldsymbol{B}(T(x, y, z)) \\ &\quad + \int_0^{2\pi}\int_0^{\pi} \boldsymbol{P}(x, y, z, \theta, \phi, \theta', \phi') \\ &\quad \boldsymbol{I}(x, y, z, \theta', \phi')\sin\theta'\,\mathrm{d}\theta'\,\mathrm{d}\phi' \end{aligned} \tag{27}$$

In general, the phase matrix $\boldsymbol{P}$ and the extinction matrix σ_e are nondiagonal and couple the four components of the Stokes vector. Therefore, the differential equations do not decouple and have to be solved simultaneously. In the scattering phase matrix $\boldsymbol{P}$, three dimensions for position and space and four for the propagation directions of the incident and scattered radiation have to be considered. Therefore, $\boldsymbol{P}$ cannot be normalized and a volume scattering coefficient cannot be defined. As a consequence, the single scattering albedo ω_0 as defined by equation (15) does not exist.

Another interesting fact to note is that there is no well-defined optical thickness for the four component Stokes vector. The exponential attenuation law (Bouguer–Lambert-Beer's law) for the transmission is not applicable and the transmission along path length ds cannot be calculated using:

$$I(s + \mathrm{d}s) = I(s)e^{-\sigma_e \mathrm{d}s} \tag{28}$$

The differences in the Stokes vector components have to be computed at a series of small enough intervals ds to ensure that the change in intensity is linear in the path length coordinate:

$$\mathrm{d}\boldsymbol{I} = \boldsymbol{\sigma}_e \boldsymbol{I}\, \mathrm{d}s \tag{29}$$

SCATTERING AND PHASE FUNCTIONS

It is one objective of scattering theory to quantify the scattering phase matrix, single scattering albedo, and the scattering cross sections for the individual scatterers, which can comprise molecules, aerosols, ice crystals, leafs, stems, and so on. The size parameter $\alpha_{size} = 2\pi r/\lambda$, which defines the relation between particle radius r and wavelength λ, the complex refraction index and the particle shape are often used to determine the appropriate solution method.

In case of isotropic scattering, the phase function is given by $p(\cos\Theta) = 1$. For many atmospheric applications, Rayleigh scattering is of particular interest. Lord Rayleigh first characterized it in 1871 in the context of the explanation for the blue sky. Although its original application focused on sunlight and atmospheric molecules, it was found that Rayleigh scattering theory could be applied to processes where the particle size is much smaller than the wavelength. As a rule of thumb, α_{size} should be smaller than 0.1. The phase function describing Rayleigh scattering can be formulated as:

$$p(\cos\Theta) = \tfrac{3}{4}(1 + \cos^2\Theta) \tag{30}$$

with the corresponding phase matrix:

$$\boldsymbol{R} = \frac{3}{2}\begin{pmatrix} \cos^2\Theta & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\Theta & 0 \\ 0 & 0 & 0 & \cos\Theta \end{pmatrix} \tag{31}$$

Following Chandrasekhar (1960), the scattered intensity I^{scat} can be written as:

$$I^{\text{scat}} = \frac{128\pi^5}{3\lambda^4}\alpha_p^2 I\, \mathrm{d}\omega \frac{3}{4}\left(1 + \cos^2\Theta\right)\frac{\mathrm{d}\omega'}{4\pi} \tag{32}$$

with α_p being the polarizability. The scattering coefficient per particle can be introduced as:

$$\sigma_{s,p} = \frac{128\pi^5}{3\lambda^4}\alpha_p^2 \tag{33}$$

Figure 5 shows a Rayleigh scattering diagram in polar coordinates and a scattering angle versus scattering phase function representation. The three curves describe I_l, I_r, and $I_l + I_r$ (the scale is not identical in each case). Rayleigh scattering theory cannot be applied if the particle size is not small compared to the wavelength, for example, for interactions of natural light with cloud and rain droplets, aerosols, ice crystals, or vegetation canopies. For particles that are large compared to the wavelength, the geometric optics approach (or ray tracing technique) can be applied (e.g. Macke *et al.*, 1996). For cases with particles sizes, which are comparable in size with the wavelength, exact solutions for single scattering problems exist only for highly symmetrical particle shapes if the surface of a particle is an iso-plane of a coordinate system, in which the vector wave equations can be derived. Under these conditions, the wave equation can be separated into a set of differential equations, which can be solved. Mie (1908) analyzed the problem for spheres in a spherical coordinate system using spherical harmonics. The solution for the scattering matrix and the scattering cross sections are obtained from an infinite series of Mie-coefficients, which are determined from the corresponding Legendre polynomials and spherical Bessel functions (Liou, 1992). A selection of Mie type phase functions is presented in Figure 6. The calculations are based on standard gamma drop size distributions (Hansen and Travis, 1974) characterized by effective radii of 10 μm and 100 μm for cloud water and rain, respectively. The wavelengths referred to in Figure 6 are 0.55 μm, 10 μm, and 1 mm for the visible, infrared, and microwave, respectively. For the visible and infrared wavelengths, the rain and water drops are comparably large and a strong forward peak characterizes the Mie scattering phase functions. In the infrared spectral region, the scattering phase function decreases monotonously with scattering angle. For the visible, a minimum around 105° is obtained. Since the microwave wavelength is much larger than the cloud water droplet, Rayleigh scattering occurs. For nonspherical particles, which are rotationally symmetrical, numerical methods have been developed to extend the applicability of Mie theory (e.g. Mishchenko *et al.*, 1996; Rother and Schmidt, 1996; Mishchenko *et al.*, 2000).

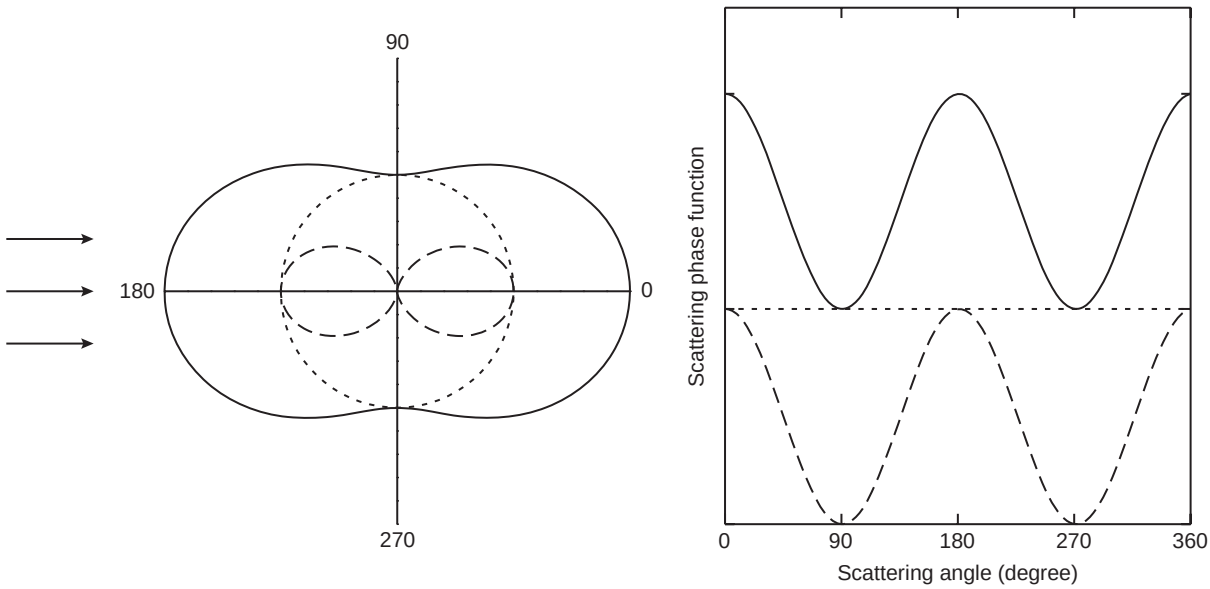


Figure 5 Schematic view of the Rayleigh phase function (after Goody and Yung, 1961)

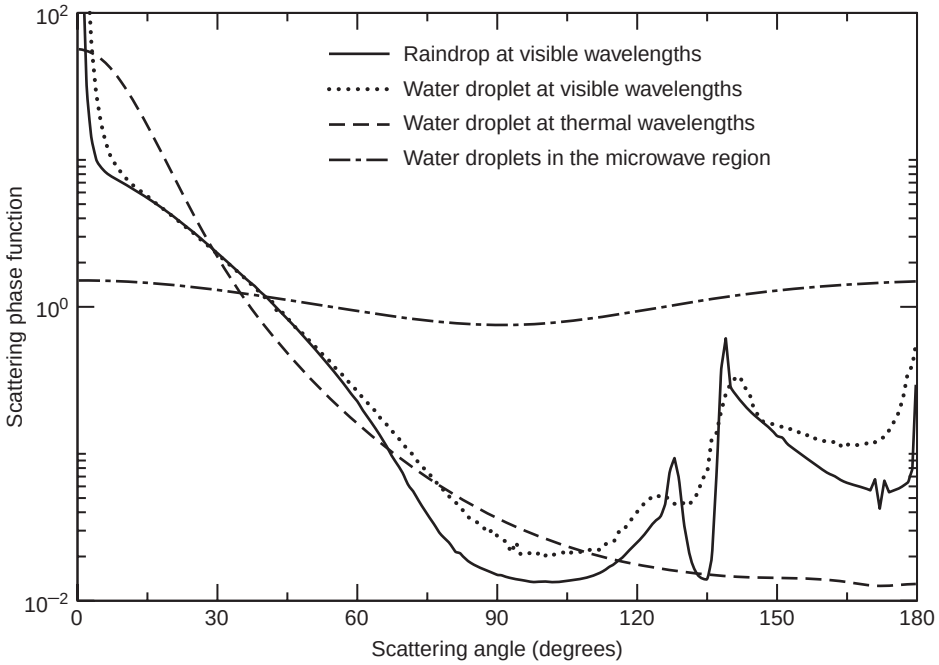


Figure 6 Scattering phase functions for visible (0.55 μm), infrared (10 μm), and microwave (1 mm) wavelengths. Effective radii of 10 μm and 100 μm for cloud water and rain characterize the drop size distributions

In order to model scattering of microwave radiation from vegetation canopies, randomly oriented nonspherical scatterers of finite size can be used. Circular discs are used to model deciduous vegetation (e.g. Tsang *et al.*, 1981). Coniferous vegetation is described by needles (e.g. Eom and Fung, 1986) and stems and branches, which are represented as circular cylinders of finite size (e.g. Cohen *et al.*, 1983; Karam and Fung, 1988). However, it has to be noted that in classical vector radiative transfer theory it is assumed that discrete particles scatter independently. This assumption is valid if the particles are distributed randomly and if the randomness of relative positions is comparable to or larger than the wavelength (Tsang *et al.*, 1985). In dense and/or discontinuous media, which are

characterized by clusters of scattering particles, collective scattering effects have to be included. Numerous models for the different wavelengths exist in the reviewed literature. Tsang *et al.* (1995) provide a brief review on different approaches and an application to passive microwave remote sensing.

BOUNDARY CONDITIONS – THE SURFACE

In general, natural surfaces reflect radiation anisotropically. The reflection at the surface can be described by the bidirectional reflectance function γ. "Bidirectional" refers to the fact that the reflection function depends on the direction of the incident radiation and the direction of the observation or outgoing radiation. The formal representation for γ is (e.g. Mishchenko *et al.*, 1999; Tsang *et al.*, 1985):

$$I(\theta, \phi) = \mu' \gamma(\theta, \phi; \theta', \phi') F \tag{34}$$

with πF the incident flux per unit area perpendicular to the incident beam (indices for polarization and frequency dependency are omitted). $\mu' \gamma_{r,t}(\theta, \phi; \theta', \phi')$ defines the bistatic scattering coefficient $\sigma^o{}_{r,t}$ (r and t indicate the polarization), which satisfies the principle of reciprocity:

$$\sigma^o{}_{r,t}(\theta, \phi; \theta', \phi') = \sigma^o{}_{t,r}(\theta', \phi'; \theta, \phi) \tag{35}$$

The emissivity ε_r and reflectivity Γ_r are given by:

$$\varepsilon_r(\theta', \phi') = 1 - \Gamma_r(\theta', \phi') \tag{36}$$

and

$$\Gamma_r(\theta', \phi') = \frac{1}{4\pi\mu} \int_0^{2\pi} \left(\sigma^o_{rr}(\theta, \phi, \theta', \phi') + \sigma^o_{tr}(\theta, \phi, \theta', \phi') \right) \sin\theta d\theta d\phi. \tag{37}$$

The reflected intensity I^S at polarization r can be expressed as:

$$I^s_r(\theta', \phi') = \frac{1}{4\pi\mu} \int_0^{2\pi} \left(\sigma^o_{rr}(\theta, \phi, \theta', \phi') I_r(\theta, \phi) + \sigma^o_{tr}(\theta, \phi, \theta', \phi') I_t(\theta, \phi) \right) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \tag{38}$$

Since many remote sensing, engineering, and biophysical applications rely on an accurate description of the bidirectional reflection function or the bistatic scattering coefficient the number of models, which relate geophysical surface parameters to both quantities, is large. The concepts and complexities vary depending on the application, for which they were designed. A number of review articles or overviews on the different models and frequency ranges exist, for example, Myneni *et al.* (1989), Goel (1988), Snyder and Wan (1998), Kerr and Wigneron (1995).

APPLICATION TO PASSIVE MICROWAVE RADIOMETRY

Equation (27) is the most general form for vector RTT. For many applications simplifications can be made to obtain analytical solutions. As an example, we will derive a solution for equation (27) for passive microwave applications. The number of geophysical parameters, which can be derived from passive microwave observations, is large and covers a wide range of applications (Ulaby *et al.*, 1982): (1) oceans (surface wind speed, temperature, salinity, and oil spills; sea ice extent and age), (2) meteorology (temperature profile, water vapor profile, liquid water (clouds and rain), and integrated water vapor), and (3) agriculture and hydrology (soil moisture distribution, flood mapping, delineation of freeze-thaw boundaries, snow cover extent, snow water equivalent, snow wetness, and continental ice sheets). The derivation outlined in the following paragraphs has been used as the basis for a number of applications, especially for the land surface (e.g. Drusch *et al.*, 2001; Kerr and Njoku, 1990; Mo *et al.*, 1982).

For plane parallel azimuthally isotropic and horizontally homogeneous atmospheres, the radiances become a function of z and θ. Owing to the azimuthally symmetry, there is no interaction between the first two components and the last two components of the Stokes vector and total intensity and linear polarization can be derived without considering the U and V components of the Stokes vector. From equation 27, one obtains (Tsang *et al.*, 1985):

$$\mu \frac{\mathrm{d}\boldsymbol{I}(z, \theta)}{\mathrm{d}z} = -\boldsymbol{\sigma}_{\mathbf{e}}(z, \theta)\boldsymbol{I}(z, \theta) + \boldsymbol{\sigma}_{\mathbf{a}}(z, \theta)\boldsymbol{B}(T(z)) + \int_0^{\pi} \boldsymbol{P}(z, \theta, \theta')\boldsymbol{I}(z, \theta') \sin\theta' \, \mathrm{d}\theta' \tag{39}$$

which is equivalent to:

$$\cos\theta \frac{d}{\mathrm{d}z} \begin{bmatrix} I_v(z, \theta) \\ I_h(z, \theta) \end{bmatrix} = -\sigma_e(z) \begin{bmatrix} I_v(z, \theta) \\ I_h(z, \theta) \end{bmatrix} + \sigma_a(z) \begin{bmatrix} \frac{B(T)}{2} \\ \frac{B(T)}{2} \end{bmatrix} + \int_0^{\pi} \begin{bmatrix} P_{vv}(z, \theta, \theta') & P_{vh}(z, \theta, \theta') \\ P_{hv}(z, \theta, \theta') & P_{hh}(z, \theta, \theta') \end{bmatrix} \begin{bmatrix} I_v(z, \theta) \\ I_h(z, \theta) \end{bmatrix} \sin\theta' \, \mathrm{d}\theta' \tag{40}$$

The sources of radiation can be summarized as:

$$J(z,\mu)=\frac{\sigma_a}{\sigma_e}\frac{B(T(z))}{2}+\omega_0(z)\int_{-1}^{1}(p_{rv}(z,\mu,\mu')I_v(z,\mu')+p_{rh}(z,\mu,\mu')I_h(z,\mu'))d\mu' \quad (41)$$

with the normalized phase functions p, which are obtained through:

$$p_{r,t}=\frac{1}{\sigma_s}P_{r,t} \quad (42)$$

Introducing the height of the atmosphere z_A and the atmospheric optical depth δ:

$$\delta(z)=\int_z^{z_A}\sigma_e(z')\,\mathrm{d}z'$$
$$\mathrm{d}\delta=-\sigma_e\,\mathrm{d}z \quad (43)$$

the radiative transfer equation for polarized radiation can be written as:

$$\mu\frac{d}{\mathrm{d}\delta}I_r(\delta,\mu)=I_r(\delta,\mu)-J_r(\delta,\mu) \quad (44)$$

which is equivalent to equation (14). A formal solution for this expression can be obtained by separating the upward and downward intensities $I^+(\mu>0)$ and $I^-(\mu<0)$:

$$I_r^+(\delta,\mu)=I_r^+(\delta_A,\mu)\exp\left[-\frac{\delta_A-\delta}{\mu}\right]+\int_\delta^{\delta_A}J_r(\delta',\mu)\exp\left[-\frac{\delta'-\delta}{\mu}\right]\frac{\mathrm{d}\delta'}{\mu} \quad (45)$$

and

$$I_r^-(\delta,\mu)=I_r^-(0,\mu)\exp\left[-\frac{\delta}{|\mu|}\right]+\int_0^{\delta}J_r(\delta',\mu)\exp\left[-\frac{\delta-\delta'}{|\mu|}\right]\frac{\mathrm{d}\delta'}{|\mu|} \quad (46)$$

If we assume a nonscattering atmosphere ($\omega_0=0$ and $\sigma_a/\sigma_e=1$) and apply the Rayleigh–Jeans-Approximation (long wavelength limit, the emission by the earth's atmosphere is directly proportional to the temperature) equations (45) and (46) can be written as:

$$T_{\mathrm{Br}}^+(\delta,\mu)=T_{\mathrm{Br}}^+(\delta_A,\mu)\exp\left[-\frac{\delta_A-\delta}{\mu}\right]+\int_\delta^{\delta_A}T(\delta')\exp\left[-\frac{\delta'-\delta}{\mu}\right]\frac{\mathrm{d}\delta'}{\mu} \quad (47)$$

and

$$T_{\mathrm{Br}}^-(\delta,\mu)=T_{\mathrm{Br}}^-(0,\mu)\exp\left[-\frac{\delta}{|\mu|}\right]+\int_0^{\delta}T(\delta')\exp\left[-\frac{\delta-\delta'}{|\mu|}\right]\frac{\mathrm{d}\delta'}{|\mu|} \quad (48)$$

with T_{Br} being the equivalent Rayleigh–Jeans brightness temperature defined as:

$$I=I_v+I_h=k_B\nu^2c^2(T_{Bv}+T_{Bh}) \quad (49)$$

It should be noted that for several applications it is necessary to calculate the brightness temperatures by inversion of the Planck function to avoid systematic errors. For the sake of simplicity, we focus on the downward radiation at the bottom of the atmosphere ($\delta=\delta_A$, e.g. the total optical depth of the atmosphere), and on the upward radiation at the top of the atmosphere ($\delta=0$) as observed from a satellite. From equation (47) and (48), we obtain:

$$T_{\mathrm{Br}}^+(0,\mu)=T_{\mathrm{Br}}^+(\delta_A,\mu)\exp\left[-\frac{\delta_A}{\mu}\right]+\int_0^{\delta_A}T(\delta)\exp\left[-\frac{\delta}{\mu}\right]\frac{\mathrm{d}\delta}{\mu}$$
$$=T_{\mathrm{Br}}^+(\delta_A,\mu)\exp\left[-\frac{\delta_A}{\mu}\right]+\mathrm{T_{BAup}} \quad (50)$$

and

$$T_{\mathrm{Br}}^-(\delta_A,\mu)=T_{\mathrm{Br}}^-(0,\mu)\exp\left[-\frac{\delta_A}{|\mu|}\right]+\int_0^{\delta_A}T(\delta)\exp\left[-\frac{\delta_A-\delta}{|\mu|}\right]\frac{\mathrm{d}\delta}{|\mu|}$$
$$=T_{\mathrm{BSp}}\exp\left[-\frac{\delta_A}{|\mu|}\right]+T_{\mathrm{Badw}} \quad (51)$$

with T_{BAup}, T_{Badw}, and T_{BSp} being the atmospheric upward, atmospheric downward, and space contribution, respectively. Since scattering in the atmosphere is neglected these contributions are not polarized. The upward radiation at the bottom of the atmosphere/top of the vegetation, $T_{\mathrm{Br}}{}^+(\delta_A,\mu)$, needs further consideration, since this part contains the contribution from the surface. In order to determine $T_{\mathrm{Br}}{}^+(\delta_A,\mu)$ the radiative transfer equation has to be solved for the canopy layer. One way to obtain $T_{\mathrm{Br}}{}^+(\delta_A,\mu)$ is to replace the corresponding atmospheric parameters by the canopy optical depth δ_c and single scattering albedo ω_c in equation (44). Let the phase function describing scattering in the canopy layer being normalized as:

$$\frac{1}{2}\int_{-1}^{1}P(\delta_c,\mu,\mu')\,\mathrm{d}\mu'=1 \quad (52)$$

For vegetation fields with long cylindrical structures the scattering in the forward direction is the dominant part (Chuang *et al.*, 1980). Consequently, the forward scattering can be expressed as the sum of a Dirac- δ-function and a phase function $P^*(\mu,\mu')$, which is not peaked in the forward direction (Joseph *et al.*, 1976):

$$P(\delta_c, \mu, \mu') = 2\alpha\boldsymbol{\delta}(\mu - \mu') + (1-\alpha)P^*(\mu, \mu') \quad (53)$$

It should be noted that it is assumed that the phase function and the single scattering albedo are independent of δ_c in this approach. Different radiative transfer models for vegetation layers are available. An overview can be found in Kerr and Wigneron (1995). However, Joseph *et al.* (1976) showed that equation (44) is invariant in form when the phase function as defined in equation (53) is used. δ_c, ω_c and $P(\delta_c, \mu, \mu')$ have to be replaced by $\delta_c{}^*$, $\omega_c{}^*$ and $P^*(\mu, \mu')$ with:

$$\begin{aligned}\delta_c^* &= (1-\alpha\omega_c)\delta \\ \omega_c^* &= \frac{(1-\alpha)\omega_c}{1-\alpha\omega}\end{aligned} \quad (54)$$

Under the assumption that forward scattering is the dominant process, and since $P^*(\mu, \mu')$ is not well known, this quantity is set to 0. Little calculus yields the upward and downward contributions for the radiation at the top and the bottom of the canopy layer, respectively (Mo *et al.*, 1982):

$$\begin{aligned}T^+_{\mathrm{CBr}}(0,\mu) = T^+_{\mathrm{CBr}}(\delta_c^*,\mu)\exp\left[-\frac{\delta_c^*}{\mu}\right] \\ + (1-\omega_c^*)T_c\left(1-\exp\left[-\frac{\delta_c^*}{\mu}\right]\right)\end{aligned} \quad (55)$$

and

$$\begin{aligned}T^-_{\mathrm{CBr}}(\delta_c^*,\mu) = T^-_{\mathrm{CBr}}(0,\mu)\exp\left[-\frac{\delta_c^*}{\mu}\right] \\ + (1-\omega_c^*)T_c\left(1-\exp\left[-\frac{\delta_c^*}{\mu}\right]\right)\end{aligned} \quad (56)$$

with the canopy temperature T_c. The downward radiation at the top of the canopy $T_{\mathrm{CBr}}{}^-(0,\mu)$ is identical to the downward radiation at the bottom of the atmosphere $T_{\mathrm{Br}}{}^-$ $(\delta_A,\ \mu)$ as given in equation (51). The upward radiation at the bottom of the atmosphere $T_{\mathrm{Br}}{}^+$ $(\delta_A,\ \mu)$ is given by equation (55). Combining (56) with (51) and (55) with (50) gives the upward radiation at the top of the atmosphere $T_{\mathrm{Br}}{}^+(0,\mu)$ and the downward radiation at the bottom of the vegetation layer $T_{\mathrm{CBr}}{}^-(\delta_c{}^*,\mu)$:

$$\begin{aligned}T^-_{\mathrm{CBr}}(\delta_c^*,\mu) = \left(T_{\mathrm{BSp}}\exp\left[-\frac{\delta_A}{|\mu|}\right] + T_{\mathrm{Badw}}\right) \\ \exp\left[-\frac{\delta_c^*}{|\mu|}\right] + (1-\omega_c^*)T_c\left(1-\exp\left[-\frac{\delta_c^*}{|\mu|}\right]\right)\end{aligned} \quad (57)$$

$$\begin{aligned}T^+_{\mathrm{Br}}(0,\mu) = \left(T^+_{\mathrm{CBr}}(\delta_c^*,\mu)\exp\left[-\frac{\delta_c^*}{\mu}\right] + (1-\omega_c^*)\right. \\ \left. T_c\left(1-\exp\left[-\frac{\delta_c^*}{\mu}\right]\right)\right)\exp\left[-\frac{\delta_A}{\mu}\right] + T_{\mathrm{BAup}}\end{aligned} \quad (58)$$

Equations (57) and (58) are coupled through the surface interaction where a portion of the downward radiation at the bottom of the canopy layer is reflected. $T_{\mathrm{CBr}}{}^+(\delta_c{}^*,\mu)$ can be written as:

$$\begin{aligned}T^+_{\mathrm{CBr}}(\delta_c^*,\mu) = (1-\Gamma(\mu))T_S \\ + \frac{1}{2\mu}\int_{-1}^{0}[\sigma^*_{rr}(\mu,\mu')T^-_{\mathrm{CBr}}(\delta_c^*,\mu') \\ + \sigma^*_{rt}(\mu,\mu')T^-_{\mathrm{CBt}}(\delta_c^*,\mu')]\,\mathrm{d}\mu'\end{aligned} \quad (59)$$

with T_S the surface temperature and

$$\Gamma_r(\mu) = \frac{1}{2\mu}\int_0^1 [\sigma^*_{rr}(\mu',\mu) + \sigma^*_{tr}(\mu',\mu)]\,\mathrm{d}\mu' \quad (60)$$

$$\sigma^*_{tr}(\mu',\mu) = \frac{1}{2\pi}\int_0^{2\pi}\sigma^o_{tr}(\mu',\mu,\phi-\phi')\,\mathrm{d}\phi \quad (61)$$

In the case of specular reflection, the bistatic scattering coefficients for cross polarization (subscripts *rt* or *tr*) are 0. For $r = t$, $\sigma*$ is given as:

$$\sigma^*_{rt}(\mu',\mu) = 2\mu\sigma_r^{Spec}\boldsymbol{\delta}(\mu_s - \mu) \quad (62)$$

The reflected radiation can then be written as:

$$T^+_{\mathrm{CBr}}(\delta_c^*,\mu) = (1-\Gamma(\mu))\,T_S + \Gamma(\mu)T^-_{\mathrm{CBr}}(\delta_c^*,\mu) \quad (63)$$

For the computation of $\Gamma(\mu)$, which is based on the dielectric properties of the soil, various methods do exist (e.g. Kerr and Njoku, 1990; Wilheit, 1978). Generally, the reflection of the land surface is not specular. To take into account roughness effects, the specular reflectivities can be modified, for example, according to Choudhury *et al.* (1979) or Wegmüller and Mätzler (1999). A comprehensive and detailed discussion on passive microwave remote sensing of land surfaces can be found in Choudhury *et al.* (1995).

SOLUTIONS

The radiative transfer problem as outlined above is rather complex, and general analytic solutions of the vector radiative transfer equation, which provide the spatial distribution of intensities, do not exist. Depending on the application several approximations may apply and a number of methods to numerically solve the RTT have been developed. Often the approximations cannot resolve the full angular dependence and/or the full polarization information. Furthermore, most solutions are achieved for the one-dimensional problem where only vertical variations are considered and the atmosphere/surface is assumed to be horizontally homogeneous. Additional geometrical complications arise from the sphericity of the Earth's atmosphere. In the case of a nonscattering and horizontally homogeneous atmosphere, the RTT reduces to a relatively simple form (47) and (48), which can easily be solved by numerical integration when the vertical distribution of the absorption coefficient and temperature is known. Because the ratio between the size of the scattering particle and the wavelength determines the scattering efficiency, the neglect of scattering is well justified at microwave wavelengths, except when (large) precipitation drops are present. At short wavelengths, for example, solar radiation, scattering by atmospheric gases needs to be considered. Detailed descriptions of methods to solve the RTT can be found in many text books and publications. Following Lenoble (1985), the different methods have been separated in three groups: (1) exact analytical methods, (2) computational methods, and (3) approximate methods. This section briefly introduces and discusses some of the different concepts. According to Lenoble (1985), the primary usefulness of exact analytical methods is the understanding of the mathematical structure and of the general behavior of solutions of the radiative transfer equation. We focus on computational and approximate methods, which are most common in hydrological and meteorological applications.

Computational methods can take into account the real characteristics of the atmosphere. The accuracy of the results depends on the computer and the computation time. *Monte Carlo* methods simulate the path of individual photons, which experience absorption and scattering according to any given phase matrix within the medium of interest. All radiation variables including polarization can be derived at any point by averaging many photons propagating statistically within the considered domain. Although the computational cost is relatively high, one of the benefits of the Monte Carlo method is the ability to consider complex 3D geometries including broken clouds. The Monte Carlo technique offers a straightforward interpretation of the radiation field, which allows controlling the accuracy by the number of photons considered. Therefore, the Monte Carlo method is often used as a reference to evaluate other models. The Monte Carlo method can be formulated either as a forward or backward problem depending on the application. For example, if remote sensing applications are considered it is useful to consider the photons arriving at the detector and trace their origin backwards. The main disadvantages are (1) the statistical fluctuation in the results and (2) the limitation to optical depths below 100.

In the *Discrete Ordinate Method* (DOM, Chandrasekhar, 1960), the angular dependency of the radiative transfer equation is discretized and the solution consists of a set of first-order differential equations. DISORT (Stamnes *et al.*, 1988) is a numerically implemented solution for vertically inhomogeneous, layered media. It solves the RTT for a scattering, absorbing, and emitting medium with an arbitrarily specified bidirectional reflectivity at the lower boundary. Primary advantages of the DOM are: (1) the solution of the radiative transfer equation can be derived explicitly and the intensity calculations do not depend on the total optical depth of cloud or aerosol layers. (2) Analytic two-stream and four-stream solutions can be derived in closed forms. (3) Computational times are relatively small compared to other techniques. Polarization effects are not included in the calculations.

The Spherical Harmonical Discrete Ordinate Method (SHDOM) calculates nonpolarized monochromatic or spectral band radiative transfer in a one, two, or three-dimensional medium for either collimated solar and/or thermal emission sources of radiation (Evans, 1998). The program package is freely available from **`http://nit.colorado.edu/~evans/shdom.html`**. In contrast to the standard DOM, the angular part of the source function is represented with a spherical harmonic expansion, which allows a more efficient calculation of the scattering integral.

Within the *Successive Order of Scattering* (SOS) method, multiple scattering is treated as a sequence of single scattering events. In order to fulfill this approximation, the medium has to be divided into optically thin layers, which is appropriate in most microwave applications. SOS gives a better understanding of the problem than many of the other methods, because the photon is followed at each scattering process. Again, polarization effects cannot be taken into account. Another disadvantage is the high computational cost for situations with very weak absorption and high optical depths.

The adding, doubling, and matrix operator methods (Plass, 1973) are closely related. The reflection and transmission functions of an initial, optically thin sub layer are obtained by single scattering calculations for all incident and emergent angles at once. This layer is successively doubled until the given optical thickness of the homogeneous layer is reached. Then the homogeneous layers are added to an inhomogenous atmosphere. Polarization can be included in this approach.

Approximate methods include either an approximation of the real atmosphere or of the radiative transfer problem.

The applicability of approximate methods depends primarily on the required accuracy. The *two-stream-approximation* is commonly implemented in numerical weather prediction and climate models to calculate heating rates in the atmosphere and the Earth surface. For this purpose, it is sufficient to derive radiant fluxes rather than intensities, which reduces the computational effort. Since radiative fluxes are angular averaged properties, details of the angular intensity variation can be neglected. The strategy is to introduce an effective angular averaged intensity (stream) and determine an effective scattering angle. It is equivalent to the lowest order ($N = 1$) of the Discrete Ordinate Method. Considering the upward and downward components (two streams) leads to a pair of coupled, first-order differential equations, the *two-stream equations*. The coupling between the differential equations depends on the way the intensity and phase function are approximated (e.g. Liou, 2002). Within a homogeneous medium, the two-stream equations can be solved analytically using standard methods; for anisotropic scattering, a number of methods have been proposed. The two-stream-approximation is most accurate under isotropic conditions when the mean inclination (average cosine) is the same in both directions but works reasonable at the boundaries, too. The *Eddington-approximation* originally developed for studying the radiative equilibrium in stellar atmospheres has a similar application range as the two-stream-approximation. Here the angular dependence is approximated by a polynom resulting in a similar set of equations as for the two-stream approximation. Both approximations can be improved by implementing the Dirac function for a better representation of the forward scattering peak in the phase function and are then referred to as δ-two-stream or δ-Eddington. Differences between the Eddington- and two-stream method can be only due to the choice of the mean inclination, the boundary conditions and the choice of the phase function (Thomas and Stamnes, 1999).

Acknowledgments

The authors would like to thank Dr. H. Czekala (Radiometer Physics GmbH) for many helpful discussions and providing Figures 2, 3, and 4. Prof. A. Macke (Institute for Marine Sciences Kiel) provided Figure 6 and suggestions on the original manuscript. The paper benefited from many comments from Prof. Simmer and Dr. J. Schulz (both Meteorological Institute Bonn), Dr. P. Bauer and Dr. R. Engelen (both ECMWF), and an anonymous reviewer. We would like to thank R. Hine (ECMWF) for his help with the final preparation of the figures.

REFERENCES

Chandrasekhar S. (1960) *Radiative Transfer*, Dover Publications: New York, p. 303.

Choudhury B.J., Kerr Y.H., Njoku E.G. and Pampaloni P. (Eds.) (1995) *Passive Microwave Remote Sensing of Land-Atmosphere Interactions*, VSP Utrecht: The Netherlands, p. 685.

Choudhury B.J., Schmugge T.J., Chang A. and Newton R.W. (1979) Effect of surface roughness on the microwave emission from soils. *Journal of Geophysical Research*, **84**, 5699–5706.

Chuang S.L., Kong J.A. and Tsang L. (1980) Radiative transfer theory for passive microwave remote sensing of a two-layer random medium with cylindrical structures. *Journal of Applied Physics*, **51**, 5588–5593.

Cohen L.D., Haracz R.D., Cohen A. and Acquista C. (1983) Scattering of light from arbitrarily oriented finite cylinders. *Applied Optics*, **23**, 436–441.

Drusch M., Wood E.F. and Jackson T.J. (2001) Vegetative and atmospheric corrections for the soil moisture retrieval from passive microwave remote sensing data: results from the Southern Great Plains hydrology experiment 1997. *Journal of Hydrometeorology*, **2**, 181–192.

Eom H.J. and Fung A.K. (1986) Scattering from a random layer embedded with dielectric needles. *Remote Sensing of Environment*, **19**, 139–149.

Evans K.F. (1998) The spherical harmonic discrete ordinate method for three-dimensional atmospheric radiative transfer. *Journal of the Atmospheric Sciences*, **55**, 429–446.

Goel N.S. (1988) Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data. *Remote Sensing of Environment*, **4**, 1–222.

Goody R.M. and Yung Y.L. (1961) *Atmospheric Radiation*, Oxford University Press: Oxford, p. 519.

Hansen J.E. and Travis L.D. (1974) Light scattering in planetary atmospheres. *Space Science Reviews*, **16**, 527–610.

Joseph J.H., Wiscombe W.J. and Weinman J.A. (1976) The Delta-Eddington approximation for radiative flux transfer. *Journal of the Atmospheric Sciences*, **33**, 2452–2459.

Karam M.A. and Fung A.K. (1988) Electromagnetic scattering from a layer of finite length, randomly oriented, dielectric, circular cylinders over a rough interface with application to vegetation. *International Journal of Remote Sensing*, **9**, 1109–1134.

Kerr Y.H. and Njoku E.G. (1990) A semi-empirical model for interpreting microwave emission from semiarid land surfaces as seen from space. *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 384–393.

Kerr Y.H. and Wigneron J.P. (1995) Vegetation models and observations: a review. *Passive Microwave Remote Sensing of Land-Atmosphere Interactions*, VSP: Utrecht, pp. 317–344.

Lean J. and Rind D. (1998) Climate forcing by changing solar radiation. *Journal of Climate*, **11**, 3069–3094.

Lenoble J. (1985) *Radiative Transfer in Scattering and Absorbing Atmospheres: Standard Computational Procedures*, A. Deepak Publishing: Hampton, p. 300.

Liou K.N. (1992) *Radiation and Cloud Processes in the Atmosphere*, Oxford University Press: Oxford, p. 487.

Liou K.N. (2002) *An Introduction to Atmospheric Radiation*, *International Geophysics Series*, *Second Edition*, Vol. 84, Academic Press: p. 583.

Macke A., Mishchenko M.I. and Cairns B. (1996) The influence of inclusions on light scattering by large ice particles. *Journal of Geophysical Research*, **101**, 23311–23316.

Mie G. (1908) Beiträge zur optik trüber medien, speziell kolloidaler metallösungen. *Annalen der Physik*, **25**, 377–445.

Mishchenko M.I., Dlugach J.M., Yanovitskij E.G. and Zakharova N.T. (1999) Bidirectional reflectance of flat, optically thick particulate layers: an efficient radiative transfer solution and applications to snow and soil surfaces. *Journal of Quantitative Spectroscopy*, **63**, 409–432.

Mishchenko M.I., Hovenier J.W. and Travis L.D. (Eds.) (2000) *Light Scattering by Nonspherical Particles. Theory, Measurements, and Applications*, Academic Press: San Diego.

Mishchenko M.I., Travis L.D. and Mackowski D.W. (1996) T-matrix computations of light scattering by non-spherical particles: a review. *Journal of Quantitative Spectroscopy & Radiative Transfer*, **55**, 535–575.

Mo T., Choudhury B.J., Schmugge T.J., Wang J.R. and Jackson T.J. (1982) A model for microwave emission from vegetation – covered fields. *Journal of Geophysical Research*, **87**, 11229–11237.

Myneni R.B., Ross J. and Asrar G. (1989) A review on the theory of photon transport in leaf canopies. *Agricultural and Forest Meteorology*, **45**, 1–153.

Plass G.N. (1973) Matrix operator theory of radiative transfer: 1. Rayleigh scattering. *Applied Optics*, **12**, 314–329.

Rother T. and Schmidt K. (1996) The discrete mie-formalism for plane-wave scattering on axisymmetric particles. *Journal of Electromagnetic Waves and Applications*, **10**, 273–297.

Salby M.L. (1996) *Fundamentals of Atmospheric Physics*, Academic Press: London.

Snyder W.C. and Wan Z. (1998) BRDF models to predict spectral reflectance and emissivity in the thermal infrared. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 214–225.

Stamnes K., Tsay S.C., Wiscombe W. and Jayaweera K. (1988) A numerically stable algorithm for discrete-ordinate-method radiative transfer in multiple scattering and emitting layered media. *Applied Optics*, **27**, 2502–2509.

Thomas G.E. and Stamnes K. (1999) *Radiative Transfer in the Atmosphere and Ocean*, Dessler A.J., Hughton J.T. and Rycroft M.J. (Eds.), Cambridge University Press: p. 517.

Tsang L., Kong J.A. and Shin R.T. (1985) *Theory of Microwave Remote Sensing*, John Wiley & Sons: New York, p. 613.

Tsang L., Kubacsi M.C. and Kong J.A. (1981) Radiative transfer theory for active remote sensing of a layer of small ellipsoidal scatterers. *Radio Science*, **21**, 771–786.

Ulaby F.T., Moore R.K. and Fung A.K. (1982) *Microwave Remote Sensing – Fundamentals and Radiometry. Volume I*, Artech House: London, p. 456.

Wegmüller U. and Mätzler C. (1999) Rough bare soil reflectivity model. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 1391–1395.

Wilheit T.T. (1978) Radiative transfer in a plane stratified dielectric. *IEEE Transactions on Geoscience Electronics*, **16**, 138–143.

47: Sensor Principles and Remote Sensing Techniques

ANTHONY W ENGLAND

Professor of Electrical Engineering and Computer Science, Professor of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, MI, US

Model-derived estimates of water stored in the upper few meters of soil, in vegetation, and in snow packs are the objectives of a new satellite microwave remote sensing technology. The most mature of the alternatives, this technology will use a combination of longer wavelength satellite microwave brightness observations and Soil Vegetation Atmosphere Transfer (SVAT) models, forced by solar and thermal radiation and by weather, to produce a near-daily global field of stored water. While the near-daily temporal resolution is a unique strength for hydrometeorology, its likely spatial resolution of tens of kilometers is a serious weakness for land-surface hydrology. This article is intended as a review of the concept and the investigations that have contributed to it. It also includes brief surveys of some technologies, like aperture synthesis radiometry and combined active passive sensing, that are being explored to improve spatial resolution.

BACKGROUND

Among Earth science disciplines in the middle of the twentieth century, hydrology was notably constrained by the limitations of *in situ* sampling. The handicap was most severe for hydrometeorology because relevant processes extend over spatial and temporal scales that are difficult to characterize with a limited number of point measurements. The handicap was less severe for land-surface hydrology because much of importance occurs at the spatial scale of the watershed and the temporal scale of weather, scales that are more amenable to point measurements. The point sampling constraint discouraged investigations of linkages between land-surface hydrology and global hydrologic processes as depicted for an arctic terrain in Figure 1. Effectively, the global hydrologic cycle was understood in concept, but experimentally tracking moisture through atmospheric and land-surface elements of the water cycle using presatellite technologies required heroic effort.

Recent convergence of coupled ocean–atmosphere circulation models, satellite remote sensing technologies, and the political imperative to distinguish between human-induced climate change and natural climate variability have stimulated investigations that place land-surface hydrology in the context of the global water cycle (e.g. NRC, 1991; Lettenmaier and Rind, 1992; NRC, 1998; Vörösmarty *et al.*, 2000). Satellite remote sensing technologies have enabled these investigations and, among these technologies, three have become central to satellite sensing of land-surface hydrology. These are (i) model-based estimation of water stored in soil, vegetation, and snow, (ii) satellite sensing of soil moisture, and (iii) higher spatial resolution sensors.

Satellite sensors can be characterized by their spatial, temporal, or spectral resolutions. These attributes become part of the decision space where sensors are matched with applications. Features of the land surface might be heterogeneous at the scale of tens of meters but change occurs slowly relative to a day. For example, satellite sensors for agriculture or land-use mapping typically have spatial resolutions of <100 m but temporal resolutions of weeks (see Table 1). We can estimate the temporal resolution for the sensors in Table 2 from the swath width. If the whole earth is to be covered by an ascending orbital pass, then temporal resolution in days is approximately the circumference of the Earth divided by the product of swath width and number of orbits per day, $\approx 40\,000\ (\mathrm{SW} \times 15)^{-1}$ where swath width (SW) is in kilometers and 15 is the approximate number of orbits completed per day for satellites in

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

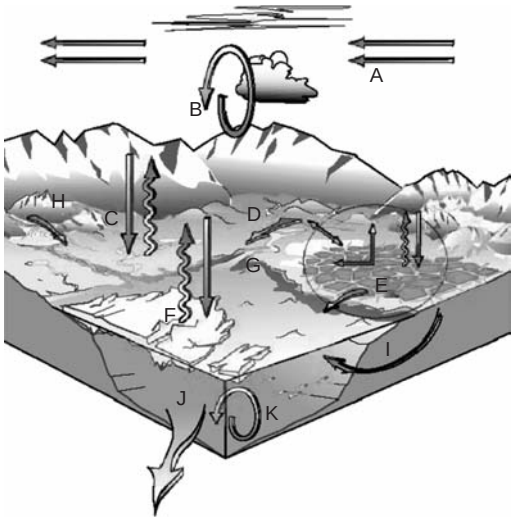


Figure 1 Conceptual model of the hydrologic cycle for the Arctic. Processes include: A – atmospheric boundary fluxes; B – atmospheric dynamics; C – land-surface atmosphere exchanges; D – discharge through well-defined flow networks; E – runoff from poorly organized lowland flow systems; F – sea ice mass balance and dynamics; G – estuarine controls on terrestrial/shelf interactions; H – changes in glacial mass balance and associated runoff; I – direct groundwater discharge to ocean; J – Arctic Ocean dynamics and deep water formation; and K – biological dynamics and oceanic food chains. Hydrometeorology includes process C – land-surface atmosphere exchange. Land-surface hydrology includes processes D – discharge through well-defined flow networks; E – runoff from poorly organized lowland flow systems; and H – changes in glacial mass balance and associated runoff. (From Vörösmarty *et al.*, 2000, Reproduced by permission of the Arctic Consortium of the US (ARCUS)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Low Earth Orbit (LEO). For example, we might estimate that the repeat time for Radarsat and ERS-1 and -2 would be approximately 27 days. In practice this temporal resolution can be shortened by using both ascending and descending passes if power and bandwidth are available, or by relaxing the requirement for global coverage, particularly at the equator. Nevertheless, none of these sensors can be thought of as providing near-daily global coverage.

Hydrometeorology required a different remote sensing strategy. Atmospheric elements of the water cycle have relevant spatial scales near the cell size of atmospheric circulation models, typically tens of kilometers, but temporal scales near a day. Operational sensors in LEO intended for hydrometeorology have had spatial resolutions as poor as 50 km but temporal resolutions of a day at higher latitudes, or twice per day if data from both ascending and descending orbital passes are used. Microwave radiometry has been well suited to high temporal resolution hydrometeorology (see Table 2). Atmospheric temperature profiles are estimated from the 60 and 118 GHz spectral features of oxygen, and water vapor profiles are estimated from the 22 and 183 GHz spectral features of water vapor (e.g. Janssen, 1993). Sea surface winds can be inferred from polarimetric brightness (e.g. Chreny and Raizer, 1998). Near-daily global coverage is readily achieved with wide swaths of conically scanning radiometers (see Figure 2). For example, the Special Sensor Microwave/Imager (SSM/I) with its 1400 km swath width would have a repeat interval of approximately 2 days and, considering both ascending and descending passes, daily global coverage at mid and high latitudes. These applications are aided by the relatively low cost and high reliability of satellite radiometers. An excellent, but somewhat dated, overview of microwave radiometry for environmental remote sensing can be found in the three-volume set by Ulaby *et al.* (1981, 1982, 1986).

Land-surface hydrology is tied to the water cycle and, like hydrometeorology, requires near-daily observations. Under relatively static and spatially homogeneous conditions, the combination of microwave brightness and empirical algorithms yields a reliable estimate of liquid water stored in a shallow soil under a significant vegetation canopy (e.g. Schmugge *et al.*, 1974; Schmugge, 1978; Newton and Rouse, 1980; Wang *et al.*, 1980a; Wang *et al.*, 1980b; Mo *et al.*, 1982; Jackson *et al.*, 1982; Ulaby *et al.*, 1983; Wang *et al.*, 1984; Schmugge *et al.*, 1986; Jackson and Schmugge, 1989; Jackson and O'Neill, 1990; Wang *et al.*, 1990; England *et al.*, 1992; Jackson, 1993; O'Neill *et al.*, 1996; Galantowicz *et al.*, 2000; Crow *et al.*, 2001; Jackson, 2001; Jackson and Lakshmi, 2001; Kerr *et al.*, 2001; Wigneron *et al.*, 2001; Shi *et al.*, 2002; Pellarin

Table 1 Examples of high spatial resolution satellite sensors and their swath widths[a]

Sensor	Description	Spatial resolution (m)	Swath width (km)
Landsat Multispectral Scanner (MSS)	Visible to near-infrared	79	185
Landsat Thematic Mapper (TM)	Visible to thermal infrared	30	185
Radarsat	5.3 GHz SAR	25	100
ERS-1 -2 Active Microwave Instrument (AMI)	5.3 GHz SAR	30	100
Japanese Earth Resources Satellite (JERS-1)	1.275 GHz SAR	18	75
Envisat Advanced SAR (ASAR)	5.3 GHz SAR	30	100

[a]Visible and near-infrared sensors collect data only during sunlit portions of an orbit. Radar sensors typically collect data only during 12–30% of an orbit because of power or data storage limitations. The swath width of radar sensors is often variable as a trade-off with spatial resolution. Listed swath widths correspond to the highest spatial resolution.

Table 2 High temporal resolution satellite microwave radiometers and their swath widths

Sensor[a]	Frequencies	Spatial resolution at lowest frequency (km)	Swath width (km)
Scanning Multichannel Microwave Radiometer (SMMR) (1978–1988)	6.6, 10.7, 18.0, 21.0. 37.0 GHz	~70	800
Special Sensor Microwave/Imager (SSM/I) (1987–present)	19.35, 22.235, 37.00, 85.50 GHz	~55	1400
TRMM Microwave Imager (TMI) (1997–present)	10.7, 19.35, 22.235, 37.00, 85.50 GHz	~40	570
Conically Scanning Microwave Imager/Sounder (CMIS) (to be launched in 2009)	6, 10, 18, 23, 36, 60, 89, 166, 183 GHz	~50	1700
ESA's Soil Moisture Ocean Salinity (SMOS) Mission (to be launched in 2007)	1.4 GHz	~50	800
NASA's Hydrosphere State Mission (HYDROS) (to be launched in 2010)	1.4 GHz active and passive	Radiometry ~40 Radar ~3	1000

[a]Kramer, 1996

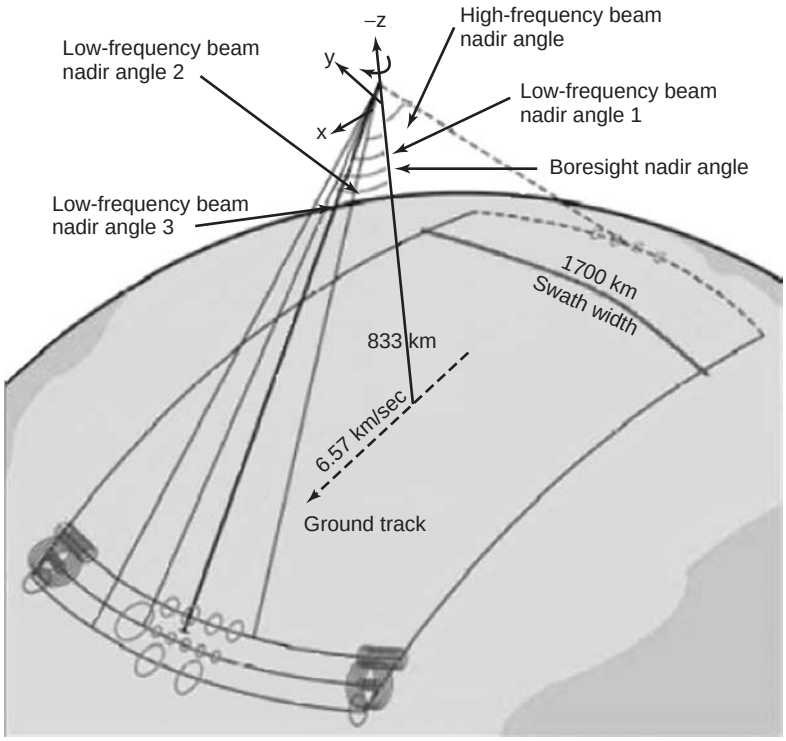


Figure 2 Scan geometry of the National Polar Orbiting Satellite System (NPOESS) Conically Scanning Microwave Imager/Sounder (CMIS) (from **`http://www.ipo.noaa.gov/Technology/cmis_summary.html`**). Like the Special Sensor Microwave/Imager (SSM/I) that preceded it (Hollinger *et al.,* 1987), CMIS scans about its vertical axis to observe swaths on the ground that are at a constant incidence angle —~55° for CMIS and 53° for SSM/I. CMIS is scheduled to be launched in 2009. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

et al., 2003; Hornbuckle and England, 2004). Such combinations of microwave brightness and an empirical algorithm can also be used to estimate the characteristics of a vegetation canopy (Ferrazzoli *et al.*, 1992; Wigneron *et al.*, 1993; Wigneron *et al.*, 1996; Karam, 1997; Njoku and Li, 1999; Ferrazzoli *et al.*, 2000; Owe *et al.*, 2001; Liu *et al.*, 2002; Hornbuckle *et al.*, 2003), whether soils are frozen or thawed (Zuerndorfer *et al.*, 1990; Zuerndorfer and England, 1992; Judge *et al.*, 1997; and Schwank *et al.*, 2004), and Snow Water Equivalent (SWE) (Chang *et al.*, 1982; Hall *et al.*, 1984; Hallikainen and Jolma, 1986; Davis *et al.*, 1987; McFarland *et al.*, 1987; Chang *et al.*, 1991; Foster *et al.*, 1991; Srivastav and Singh, 1991; Chang and Tsang, 1992; Chang *et al.*, 1992; Hallikainen, 1992; Goodison *et al.*, 1993; Koike and Suhama, 1993; Jin, 1997; Pulliainen *et al.*, 1999; Wilson *et al.*, 1999; Derksen *et al.*, 2000; Kharuk *et al.*, 2000; Rosenfeld and Grody, 2000; Macelloni *et al.*, 2001; Pulliainen and Hallikainen,

2001; Derksen *et al.*, 2003; Guo *et al.*, 2003; and Kelly *et al.*, 2003).

Static algorithms have limited value when the moisture content of the upper few centimeters of soil, the depth sensed by microwave techniques, varies significantly throughout the diurnal cycle and rapidly during the initial dry-down after a precipitation event. We usually want to know the moisture content in the root zone or at the freeze/thaw boundary of a permafrost active layer, but the relationship between near-surface water content and the water content in the root zone or deeper is dynamic. Static SWE algorithms fail when snow packs begin to melt and metamorphose, or when snow packs are laterally heterogeneous within the footprint of the sensor. Heterogeneity can be caused by topography or by variations in soil type or vegetation cover. Whatever the cause, stored water should be viewed as a dynamical parameter of a land-surface model. Water stored within the large footprint of a satellite radiometer becomes an area-weighted average of the dynamical processes within the more homogeneous subregions of the footprint.

Technologies that address these challenges – relating near-surface soil moisture to deeper soil moisture, sensing near-surface soil moisture, and achieving acceptable spatial resolutions with satellite sensors – are the themes of this review. The article is organized as (i) *Stored Water as a Modeled Product*, (ii) *Satellite Radiometers for Sensing of Shallow Soil Moisture and Snow Water Equivalent (SWE)*, and (iii) *Improved Spatial Resolution*. Although progress has been made toward retrieval of soil moisture from radar backscatter and from combinations of microwave brightness and radar backscatter, these techniques are not nearly as mature as are retrievals from microwave brightness alone. Only a cursory overview of these approaches is presented because the techniques are in rapid flux and it is unclear how successful they will be.

STORED WATER AS A MODELED PRODUCT

Data from the next generation of satellite radiometers will allow reliable detection of seasonal to decadal changes in the hydrologic cycle as expressed in temporal and spatial patterns of moisture stored in soil and snow. A schematic representation summarizing the concept as it has evolved within the remote sensing land-surface hydrology community is shown in Figure 3. Realization of the approach will yield a model-derived dynamical field of land-surface stored water, that is, the water stored in the upper few meters soil, in vegetation, and in snow. This stored water field and its state govern runoff of precipitation and snow-melt through the saturation and freeze/thaw conditions of soil, and influence the weather through land–atmosphere latent energy exchanges. A temporal record of the stored water field will also serve as a climatic record of changes in the regional moisture regime.

The approach uses the combination of a (i) Soil Vegetation Atmosphere Transfer (SVAT) model for terrain and the corresponding (ii) Radiobrightness (R) model of the microwave brightness of the terrain, to obtain a model-based estimate of stored water. The linked SVAT/R models are forced by weather and downwelling radiance to predict microwave brightness at frequencies that are sensitive to soil moisture. The models are validated and calibrated as

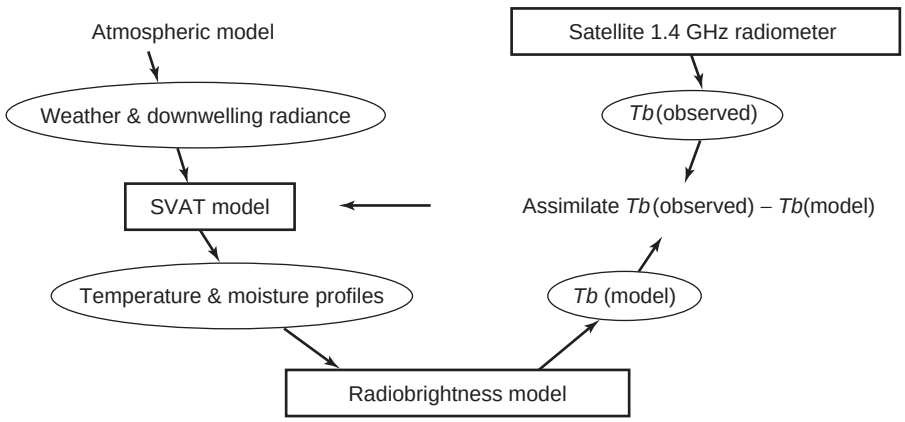


Figure 3 A strategy for remotely monitoring water stored in soil. The Atmospheric Model (or observed weather) provides weather and downwelling short- and long-wavelength radiation as forcing to the SVAT Model. The SVAT Model provides soil and vegetation temperature and moisture profiles to the Radiobrightness Model, which then predicts microwave brightness temperature, *Tb*. At low microwave frequencies, for example 1.4 GHz, a difference between observed brightness from a satellite and predicted brightness represents an error in the model's estimate of the liquid water content of near-surface soil. The error is assimilated by the SVAT Model to improve the current estimate. Over several assimilation cycles, the SVAT model converges to an accurate moisture profile at depths below those sensed by microwave radiometry

forward models, but then inverted to assimilate microwave brightness. Assimilation of an observation that is sensitive to the water content of shallow soils causes modeled soil temperature and moisture profiles to converge to actual profiles (e.g. Entekhabi *et al.*, 1994; Houser *et al.*, 1998; Davis *et al.*, 1995; Galantowicz *et al.*, 1999; Wigneron *et al.*, 1999; Lakshmi, 2000; Reichle *et al.*, 2001; Reichle and McLaughlin, 2001; Walker and Houser, 2001; Burke *et al.*, 2001). For convergence of a model to a realistic state, the SVAT/R model must accurately reflect both the physics of coupled temperature and moisture transport processes, and the emitted radiation field of the soil, vegetation canopy, and snow pack (e.g. Liou and England, 1998b). The evolution of these models has been from empirical models with sufficient free parameters to match observational data, to highly physically based models with ever fewer free parameters that begin to constrain the more parameterized operational models even in the absence of observational data.

SVAT/R models are mature for prairie agricultural hydrology because experimental data from a series of field campaigns in midlatitude prairie were available to test schemes for modeling energy and moisture transport in prairie soils and to calibrate models of prairie terrains. Notable among the field campaigns have been the Konza Prairie observations in 1985 (Schmugge, *et al.*, 1988), the First ISLSCP Field Experiment (FIFE) in 1987 (e.g. Wang *et al.*, 1990), MACHYDRO'90 (Jackson *et al.*, 1994), Washita'92 (Jackson *et al.*, 1995), and the Southern Great Plains Hydrology Experiments – SGP'97 and SGP'99 (Jackson *et al.*, 1999; Jackson *et al.*, 2002; and Guha *et al.*, 2003). There have also been a series of plot-scale experiments to validate modeled linkages between the diurnal, seasonal, and annual characteristics of radiobrightness, and the energy and moisture process in soil vegetation and snow. These experiments have been used in the development of SVAT/R models for prairie agricultural terrains and are beginning to be used in the development of SVAT/R models for arctic tundra. These Radiobrightness Energy Balance Experiments (REBEX) include REBEX-1 in northern prairie during 1992–1993 (Galantowicz and England, 1997), REBEX-3 in arctic tundra in 1994–1995 (Kim and England, 2003), REBEX-4 in northern prairie during the summer of 1995 (Judge *et al.*, 2001), REBEX-5 as part of SGP'97 (Judge *et al.*, 1999), and REBEX-8 in corn during the summer of 2001 (Hornbuckle and England, 2004). There have also been plot-scale empirical examinations of the linkage between radiobrightness and diurnal soil moisture (e.g. Jackson *et al.*, 1997).

(a) The SVAT Model Highly parameterized SVAT models formerly served as the lower boundary for Atmospheric General Circulation Models (AGCMs) (e.g. Trenberth, 1995). Examples of these include the Biosphere Atmosphere Transfer Scheme (BATS) (Dickinson *et al.*, 1986, 1993) and the Simple Biosphere-2 (SiB-2) model (Sellers *et al.*, 1986). Growth of computational power enabled an evolution toward more physically based SVAT models like the Land Surface Model (LSM) (Bonan, 1996) and the Variable Infiltration Capacity (VIC) model (Liang *et al.*, 1994, 1996). Higher fidelity diagnostic SVAT models are used to explore the parameter space of land-surface processes or to provide unique classes of state estimates for other purposes. Reviews and intercomparisons of SVAT models are available in the literature.

The Michigan Land Surface Process (LSP) model is a one-dimensional diagnostic SVAT model that evolved from a simple thermal model for bare soil (England, 1990; England *et al.*, 1992) to one including coupled heat and moisture transport in prairie agricultural terrains. The model provides accurate near-surface moisture gradients throughout the diurnal cycle, through periods of fog, and following precipitation events (Liou and England, 1996, 1998a,b; Liou *et al.*, 1998; Liou *et al.*, 1999; Judge *et al.*, 1999; and Judge *et al.*, 2003). Without the careful coupling of temperature and moisture transport processes, near-surface moisture profiles at the extremes of the diurnal thermal cycle would be unreliable, and because microwave brightness depends upon the near-surface moisture gradient, predicted brightness would be unreliable as well.

Version 3.0 of the LSP model consists of a bilayer vegetation canopy over a 40-node soil. To minimize the number of free parameters, soil constitutive properties are tied to soil texture wherever possible (e.g. de Vries, 1963; Kimball *et al.*, 1976; Lai *et al.*, 1976; Rawls *et al.*, 1991; Rossi and Nimmo, 1994). In theory, each node could delineate a different soil type, but in practice, field data seldom justify more than three soil layers. Node spacing increases exponentially from 2 mm at the surface to span total model thickness (an input parameter), which has been as much as 20 m. The saturation level (water table), also an input parameter, typically lies between 1 and 6 m for prairie sites. Within the saturated zone, moisture flux is assumed to be horizontal and energy flux is assumed to be vertical via heat conduction. Energy and moisture transport in the unsaturated zone are assumed to be vertical and are governed by the following equations (Richards, 1931; Philip and de Vries, 1957; de Vries, 1958);

$$\frac{\partial X_{\mathrm{m}}}{\partial t} = -\nabla \bullet \bar{q}_{\mathrm{m}} \qquad \frac{\partial X_{\mathrm{h}}}{\partial t} = -\nabla \bullet \bar{q}_{\mathrm{h}}$$

$$X_{\mathrm{m}} = \rho_{\mathrm{l}}(\theta_{\mathrm{l}} + \theta_{\mathrm{v}})\; X_{\mathrm{h}} = C_{\mathrm{m}}(T - T_{\mathrm{o}}) + L_{\mathrm{o}}\rho_{\mathrm{l}}\theta_{\mathrm{v}} + \rho_{\mathrm{l}} \int_0^{\theta_{\mathrm{l}}} W \, \mathrm{d}\theta$$

$$\bar{q}_{\mathrm{m}} = -\rho_{\mathrm{l}}(D_T \nabla T + D_\theta \nabla \theta + K\hat{k})$$

$$\bar{q}_{\mathrm{h}} = -\lambda \nabla T + L_o \bar{q}_{\mathrm{v}} + (c_p q_{\mathrm{v}} + c_{\mathrm{l}} q_{\mathrm{l}})(T - T_o) \qquad (1)$$

where X_m and X_h are total moisture and heat contents per unit volume ($kg\,m^{-3}$ and $J\,m^{-3}$)

$\bar{q}_l$, $\bar{q}_v$, and $\bar{q}_m$ are liquid, vapor, and moisture flux densities ($kg\,m^{-2}\,s^{-1}$)

q_h is heat-flux density ($J\,m^{-2}\,s^{-1}$)

ρ_l is the density of liquid water ($kg\,m^{-3}$)

θ_l and θ_v are volumetric liquid water ($m^3\,m^{-3}$) and vapor content (m^3 of precipitable waterper m^3), respectively, and $\theta = \theta_l + \theta_v$

T is absolute temperature (K)

D_T and D_θ are thermal and isothermal moisture (liquid and vapor) diffusivities ($m^2\,K^{-1}\,s^{-1}$)

K is the unsaturated hydraulic conductivity ($m\,s^{-1}$) in the direction of $\bar{q}_m$, $\hat{k}$

C_m and C_d are volumetric heat capacities of moist and dry soils ($J\,m^{-3}\,K^{-1}$)

c_p and c_l are specific heats ($J\,kg^{-1}\,K^{-1}$) of water vapor at constant pressure and of liquid water, respectively

L_o is the latent heat of vaporization ($J\,kg^{-1}$) at the reference temperature, T_o

W is the differential heat of wetting, and

λ is the thermal conductivity of soil ($J\,m^{-1}\,K^{-1}\,s^{-1}$)

Moisture and energy fluxes at the upper boundary are:

$$q_m(0,1) = \rho_l(D_c - E_s - E_{tr} - \text{Runoff})$$
$$q_h(0,1) = R_{ns} - H - L \qquad (2)$$

where q_m (0,1) and q_h (0,1) are moisture and heat-flux densities at the interface between the vegetation and the soil, respectively,

H and L are sensible and latent heat fluxes from the soil, respectively,

D_c is rate of drainage from the canopy ($m\,s^{-1}$), D_c = total precipitation – canopy interception,

E_s is rate of evaporation from the soil ($m\,s^{-1}$),

E_{tr} is the rate of transpiration from the root zone ($m\,s^{-1}$), and

R_{ns} is net radiation (longwave and shortwave) absorbed by the soil ($W\,m^{-2}$).

The lower boundary is assigned to be below the penetration depth of the annual thermal pulse for seasonal studies, or below the penetration depth of a multiyear pulse if the objective is to examine a climate-related sensitivity to a change in forcing. Energy and moisture fluxes at this boundary are set to zero. That is, both soil surface and bottom boundaries are constrained by fluxes – by Neumann boundary conditions.

Temperature and moisture in each layer at time t comprise a state vector that is propagated from time t to time $t + \Delta t$ where Δt is typically 2 min. Elements in the coefficient matrix for the propagation interval are linearized versions of the dynamical relations appropriate for the system's state at time t. The initial state is provided by field measurements or through model spin-up based upon climatic forcing. Propagation can be months for seasonal studies or years for an examination of sensitivity to climate change. Version 3.0 of the prairie LSP model has 1500 lines of Fortran 90 code and propagates one year in less than a minute on a high-end PC running Linux. This performance is achieved by generating look-up tables for the constitutive properties at each expected combination of temperature and moisture content prior to propagating the model rather than recomputing constitutive properties at each propagation interval for each node based upon current state. Performance of the LSP model is illustrated in Figure 4.

(b) The R Model Familiarity with the physics of microwave brightness enables an intuitive expectation for brightness temperatures that warrants a brief tutorial. Primary factors governing the microwave brightness of moist soil are soil temperature and the dielectric properties of water in the form that it is stored in the soil. The dielectric properties of water below ~10 GHz are radically different from those of other common materials in the Earth's environment. The real part of liquid water's complex relative permittivity is ~80 at 1 GHz (see Figure 5) while those of other common materials are less than 10 and generally less than 6. This high relative permittivity results from the coupling between a water molecule's permanent electric moment and the electric field of the passing electromagnetic wave. At 1 GHz, the water molecule physically rotates so that its electric moment is aligned with the wave's electric field. The mechanical energy from this rotation is the source of heat in a microwave oven. As frequency increases, the rotation of the water molecule falls behind the oscillating field of the passing wave. At the upper end of the microwave spectrum, the water molecule does not follow the field and the dielectric properties of water resemble those of most other dielectric media.

Moist soils are a mixture of soil particles, air, free water, water adsorbed to the surface of the soil particles, and ice, if temperatures are below 273 K. The complex relative permittivity of moist soils, $\tilde{\varepsilon}$, is closely approximated by the mixing law,

$$\tilde{\varepsilon}^\alpha = \sum_i \nu_i \tilde{\varepsilon}_i^\alpha \qquad (3)$$

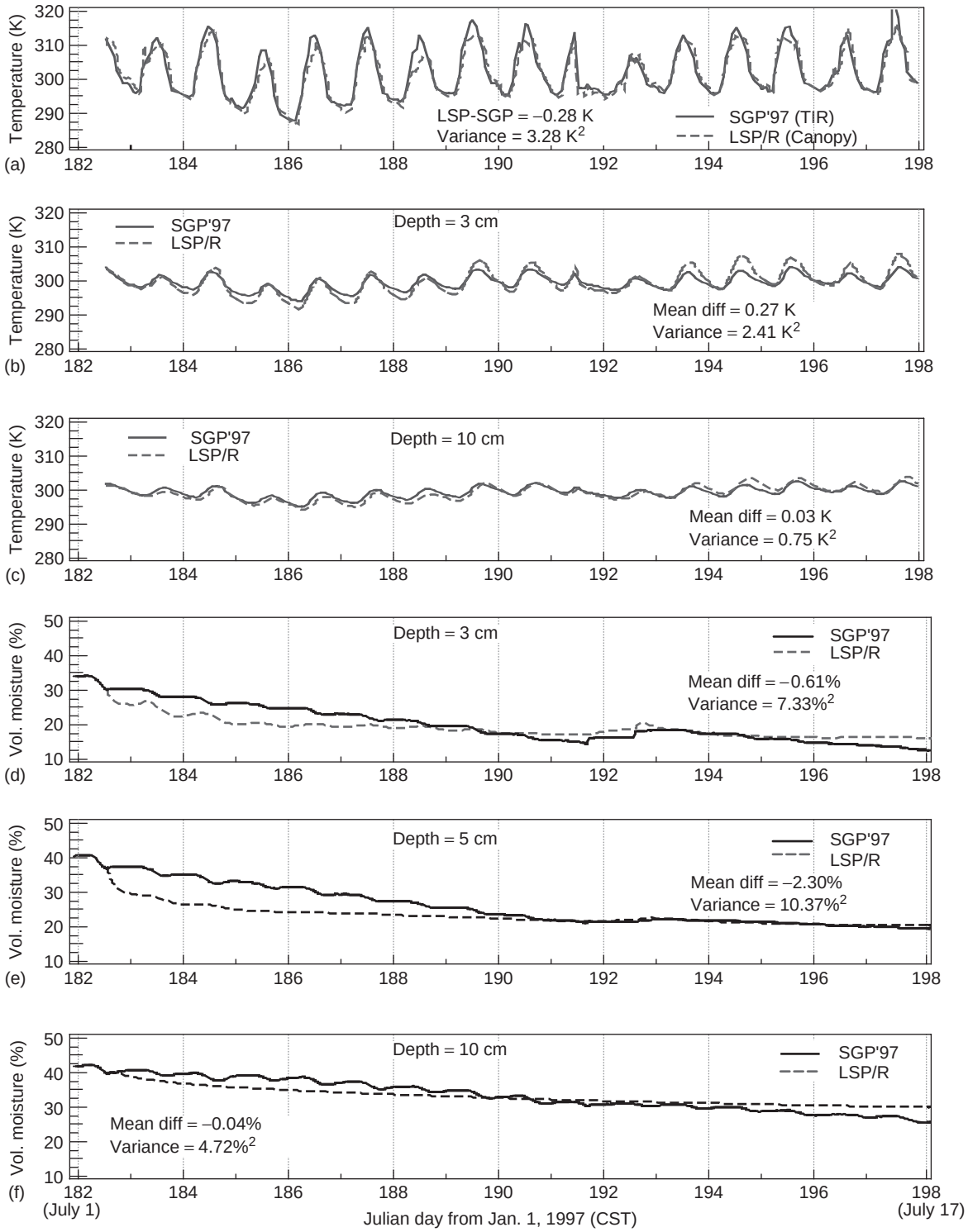


Figure 4 Performance of LSP model during the 1997 Southern Great Plains experiment in Oklahoma (SGP'97). This field contained mixed wheat stubble and short grass. These data were collected in early July. Panel (a) is the measured (solid line) and modeled (dashed line) Thermal Infrared (TIR) temperature or skin temperature of the canopy. Panels (b) and (c) are measured and modeled temperatures at depths of 3 and 10 cm. Panels d–f are measured and modeled volumetric soil moistures at 3, 5, and 10 cm depths, respectively. The errors in soil moisture on days 183–190 are attributable to the soil Moisture Retention Model (MRM). A revised model appears to be more reliable but tests are not completed

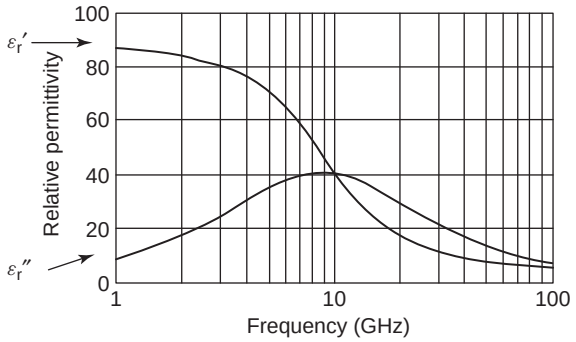


Figure 5 The complex relative permittivity, $\tilde{\varepsilon}_r = \varepsilon_r' - j\varepsilon_r''$, of water at 0 °C (ref). The dielectric properties of water are unique among common materials found in the natural environment. Below the Debye relaxation frequency, ~10 GHz, the real part of the relative permittivity of liquid water can be an order of magnitude greater than those of other materials, and the imaginary part is easily the dominant loss term. Increasing liquid water content in soil or snow means an increasing dielectric contrast between soil or snow and air. This dielectric contrast at the interface reflects upwelling emission within the soil or snow causing darkening of the microwave brightness above the surface. Increasing liquid water content in soil or snow also means increasing opacity so that upwelling emission that escapes the surface originates at shallower depths

where v_i and $\tilde{\varepsilon}_i$ are volume fraction and complex relative permittivity of the i^{th} soil component, respectively, and α is an empirically determined parameter that lies between 0.5 and 1.0. $\alpha = 1.0$ corresponds to a volume-weighted average of complex relative permittivities. The complex index of refraction is $\tilde{n} = \sqrt{\tilde{\varepsilon}}$ so that $\alpha = 0.5$ corresponds to a volume-weighted average of complex indices of refraction. We find that $\alpha = 0.5$ yields an adequate approximation to bulk dielectric properties. Others have found that $\alpha = \sim 0.65$ yields a good fit (Dobson *et al.*, 1985). Theory does not, as yet, guide our choice of α.

The complex index of refraction can be written as $\tilde{n} = n - j\kappa$ where n is the index of refraction and κ is the wave extinction coefficient. Index n governs the wave number, k, of the electromagnetic wave through the relationship $k = k_o$ n in units of rad m^{-1} where k_o is free-space wave number. Note that wave number is spatial frequency (rad m^{-1}) much like $2\pi f$ is angular frequency (rad s^{-1}). Spatial frequency, or wave number, and angular frequency are related through the dispersion relation,

$$\mathrm{k} = \left(\frac{2\pi f}{c}\right) n = \mathrm{k}_o n \qquad \text{rad s}^{-1} \tag{4}$$

where f is frequency, and c is the speed of light in vacuum. The extinction coefficient, κ, governs power loss through the microwave absorption coefficient, $\kappa_a = 2k_o\kappa$ m^{-1}.

A dielectric medium in thermal equilibrium with its extended environment contains electromagnetic radiation obeying Planck's law for spectral radiance (e.g. Bohren and Huffman, 1983). The Rayleigh–Jeans approximation to Planck's law, which is valid at microwave frequencies for all temperatures found in the Earth's natural environment, is

$$B_p^b = \mathrm{k}\left(\frac{fn}{c}\right)^2 \mathrm{T} \qquad \text{W m}^{-2}\ \text{sr}^{-1}\ \text{Hz}^{-1} \tag{5}$$

where B_p^b is Planck spectral radiance of polarization p, k is Boltzmann's constant, and T is temperature in Kelvin of the dielectric and its extended environment. Along any ray element, ds, within the dielectric, Planck brightness, B_p^b, will be incrementally attenuated by $\kappa_a B_p^b$. Thermal equilibrium requires that this loss be balanced by an equal self-emission. This "volume emissive power", $\kappa_a B_p^b$, is an intrinsic property of the dielectric at temperature T.

If the constraint of thermal equilibrium with an extended environment is replaced by the constraint of local thermodynamic equilibrium, then the heuristic differential form of the Radiative Transfer Equation (RTE) in the absence of scattering and refraction is

$$\frac{\mathrm{d}B_p}{\mathrm{d}s} + \kappa_a B_p = \kappa_a B_p^b \qquad \text{W m}^{-3}\ \text{sr}^{-1}\ \text{Hz}^{-1} \tag{6}$$

where B_p is local spectral radiance of polarization, p, $\kappa_a B_p$ is absorption loss in incremental distance ds, and $\kappa_a B_p^b$ is thermal self-emission in incremental distance ds (e.g. Chandrasekhar, 1960). Imposing thermal equilibrium with an extended environment sets $B_p = B_p^b$ and requires the intuitively correct $\mathrm{d}B_p/\mathrm{d}s = 0$. Either equality can be shown to be thermodynamically necessary if the medium is in equilibrium with its extended environment whether or not scattering occurs.

If a nonscattering, nonrefractive medium is nonabsorbing (i.e. $\kappa_a = 0$), equation (6) requires that radiance be invariant along any ray whether or not the medium is in thermal equilibrium with its extended environment. This invariance of B_p is a useful concept where the index of refraction is constant along a ray. For example, the spectral brightness of the Sun is independent of distance from the Sun as observed from anywhere attainable by a ray from the Sun traveling only in the vacuum of space.

The concept of brightness invariance fails when the index of refraction changes as it often does in the earth's environment. Where n changes slowly relative to a wavelength, brightness will change even in the absence of absorption or scattering. Invariance can be reestablished by normalizing spectral brightness by n^2. In practice, spectral brightness is normalized by $\mathrm{k}(fn/c)^2$ to yield brightness temperature,

Tb_p, of polarization p,

$$Tb_p = \frac{B_p}{\mathrm{k}\left(\frac{fn}{c}\right)^2} \qquad \mathrm{K\ sr^{-1}} \tag{7}$$

Steradian is not a true dimension (like radian, it is a ratio), but writing the units of Tb_p as K sr^{-1} reminds us that Tb_p behaves like radiance rather than temperature. Tb_p offers intuitive advantages beyond invariance in the absence of absorption and scattering. The value of B_p^b at 3 GHz for a 300 K black source is 4×10^{-17} W m^{-2} sr^{-1} Hz^{-1}. This awkwardly small number carries little intuitive meaning. In contrast, the brightness temperature, Tb_p, of a 300 K black source is an intuitive 300 K sr^{-1} at all frequencies in any dielectric medium.

Earth's microwave brightness temperature as seen from a satellite, and as shown schematically in Figure 6, is overwhelmingly a product of thermal emission, absorption, and scattering within the earth's environment. The contrast between a warm emitting Earth and the cold extraterrestrial microwave sky gives Earth-viewing microwave radiometry a dynamic range exceeding 150 K. Extraterrestrial microwave sources of significance for our purposes are the isotropic 2.7 K cosmic background at all microwave frequencies, and coronal synchrotron emission from the Sun at microwave noise equivalent temperatures greatly exceeding the Sun's 6000 K blackbody temperature. Solar microwave brightness can saturate or possibly damage a satellite microwave radiometer but is rarely a factor because the Sun occupies only 6×10^{-5} sr of the sky and seldom appears in the beam of an Earth-viewing sensor.

Earth-viewing radiometry is primarily concerned with radiative transfer within horizontally stratified media where the independent variables are height, z (positive upward), and the direction cosine, μ, of the zenith angle. Using $\mathrm{d}s = \mathrm{d}z/\mu$, the RTE for brightness temperature is

$$\mu \frac{\mathrm{d}Tb_p(z,\mu)}{\mathrm{d}z} + \kappa_\mathrm{a}(z) Tb_p(z,\mu) = \kappa_\mathrm{a}(z) T(z) \tag{8}$$

The integral form of equation (8) is

$$\Delta Tb_p(z,\mu) = \int_a^z T(z') \frac{\kappa_\mathrm{a}(z')}{\mu} \mathrm{e}^{-\tau(z',z)/\mu}\, \mathrm{d}z' \tag{9}$$

where $\Delta Tb_p(z,\mu)$ is the contribution at height z in direction μ from emission along the ray between height $z' = a$ to $z' = z$, and $\tau(z',z)$ is the "optical thickness" between z' and z defined

$$\tau(z',z) \equiv \int_{z'}^z \kappa_\mathrm{a}(z'')\, \mathrm{d}z'' \tag{10}$$

Upwelling ($\mu > 0$) and downwelling ($\mu < 0$) brightness temperatures within a layer bounded by $a < z < b$ become, respectively,

$$(\mu > 0) \qquad Tb_p(z,\mu) = Tb_p(a,\mu)\mathrm{e}^{-\tau(a,z)/\mu} + \int_a^z \frac{\kappa_\mathrm{a}(z')}{\mu} T(z') \mathrm{e}^{-\tau(z',z)/\mu}\, \mathrm{d}z' \qquad \mathrm{K\ sr^{-1}} \tag{11a}$$

$$(\mu < 0) \qquad Tb_p(z,\mu) = Tb_p(b,\mu)\mathrm{e}^{-\tau(b,z)/\mu} + \int_b^z \frac{\kappa_\mathrm{a}(z')}{\mu} \mathrm{T}(z') \mathrm{e}^{-\tau(z',z)/\mu}\, \mathrm{d}z' \qquad \mathrm{K\ sr^{-1}} \tag{11b}$$

where $Tb_p(a,\mu)$ in equation (11a) is upwelling brightness temperature entering the layer at $z = a$, and $Tb_p(b,\mu)$ in equation (11b) is downwelling brightness temperature entering the layer at $z = b$. These expressions are equally valid whether the dielectric is an atmosphere, moist soil, a snow pack, or a vegetation canopy to the extent that scattering within these media can be ignored.

If an emitting medium is significantly heterogeneous at the scale of the microwave wavelength, the RTE must be modified to include scattering (e.g. Chandrasekhar, 1960; England, 1974, 1975, 1976; England and Johnson, 1977,

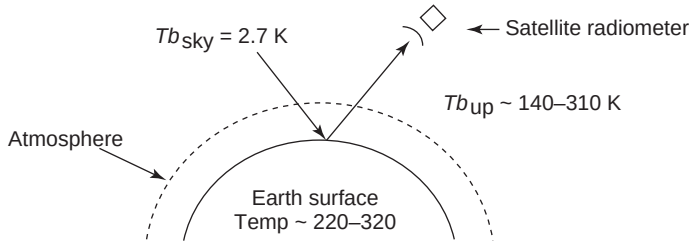


Figure 6 Schematic of microwave brightness as seen by an Earth-viewing satellite. The dynamic range of the Earth's radiometric temperature at 1.4 GHz exceeds 170 K while the dynamic range of the Earth's physical temperature is approximately 100 K. The excess dynamic range of the radiometric temperature is caused by two factors: (i) Dielectric contrasts between the land or water surface and the air, and (ii) the cold sky. If the sky had radiometric temperatures nearer those of the Earth, the upwelling brightness would be less sensitive to dielectric contrasts at the surface. It is often useful to think of a dielectric contrast at the land–atmosphere interface as reflecting cold sky

1978; Tsang and Kong, 2001). For discrete spherical scatterers randomly located at distances large enough that electromagnetic interactions among scatterers can be ignored, the RTE for a horizontally stratified medium becomes

$$\mu \frac{\mathrm{d}Tb_p(z,\mu)}{\kappa_\mathrm{e}\,\mathrm{d}z} + Tb_p(z,\mu) = (1-\omega_\mathrm{o})T(z) + \left(\frac{\omega_\mathrm{o}}{2}\right)\sum_{q=v,h}\int_{-1}^{1} p_{pq}(\mu,\mu')Tb_q(z,\mu')\,\mathrm{d}\mu' \quad (12)$$

where κ_e is total extinction, $\kappa_\mathrm{e} = \kappa_\mathrm{a} + \kappa_\mathrm{s}$, κ_s is scattering loss, ω_o is the single scattering albedo, $\omega_o = \kappa_\mathrm{s}/\kappa_\mathrm{e}$, v and h are vertical and horizontal polarizations, respectively, and $p_{pq}(\mu,\mu')$ is a scattering phase function that accounts for scattering of polarization q from all directions μ' into a brightness temperature of polarization p in direction μ. Numerical solutions of this integral–differential equation typically employ Gaussian quadrature.

Forward solutions of equation 12 are used to understand the effects of scattering, but are too complex to realistically invert. A first-order approximation to equation (12) assumes that the integral term containing the scattering phase function can be ignored. The result is equation (8) with a source that is diminished by the product of $(1-\omega_\mathrm{o})$. This source darkening, or scatter darkening, can be estimated from Raleigh or Mie scattering theory (e.g. Bohren and Huffman, 1983), but is often treated as an empirical parameter to adjust for observed scatter darkening.

The Brightness Temperature of a Vegetation Covered Soil

Temperature and moisture profiles for soil and canopy generated by the SVAT model are used in a Radiobrightness (R) model to compute the expected microwave brightness temperature as sensed immediately above a canopy. The relationship, depicted schematically in Figure 7, is:

$$Tb_p = Tb_p^\mathrm{s} + Tb_p^{\downarrow\mathrm{c}} + Tb_p^{\uparrow\mathrm{c}} + Tb_p^\mathrm{sky} \quad (13)$$

where Tb_p is observed microwave brightness temperature at polarization p (in K sr^{-1}),

Tb_p^s is the brightness temperature of soil attenuated by a canopy,

$Tb_p^{\downarrow\mathrm{c}}$ is downwelling canopy brightness temperature reflected by soil and attenuated by one trip through a canopy,

$Tb_p^{\uparrow\mathrm{c}}$ is an upwelling canopy brightness temperature, and

Tb_p^sky is the sky brightness temperature attenuated by two trips through a canopy and reflected by the soil.

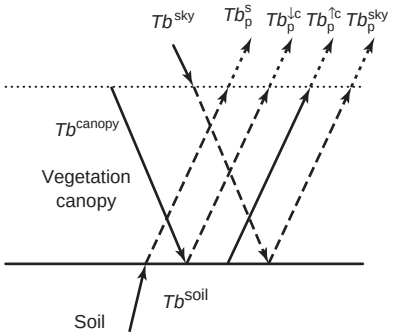


Figure 7 A schematic of contributions to microwave brightness temperature of a vegetation canopy over soil as seen by a satellite sensor. Solid rays indicate sources. Dashed rays indicate absorption. Dotted rays indicate transmission through the atmosphere. Atmospheric absorption and emission is small at low microwave frequencies and is reliably modeled given approximate atmospheric temperature and water vapor profiles. The contributions to brightness temperature above the canopy are, from left to right: unpolarized emission from soil, polarized by transmission through the soil interface, and partially absorbed by the canopy; unpolarized downwelling canopy emission, polarized by reflection from the soil, and partially absorbed by the canopy; unpolarized upwelling canopy emission; and unpolarized downwelling sky brightness temperature, partially absorbed by downward trip through canopy, polarized by reflection at soil interface, and partially absorbed by upward trip through canopy

Downwelling sky brightness temperatures are computed for the atmosphere using equation (11b). A clear atmosphere has brightness temperatures like those shown in Figure 8. Clouds increase these temperatures only at the higher microwave frequencies. Because absorption and emission are inherently linked, this is equivalent to saying that lower frequency microwaves penetrate clouds. Upwelling brightness temperature within soil at $z = 0$ is computed using equation (11a) with $a = -\infty$. The brightness temperature immediately below the soil–canopy interface, $Tb_p(0^-,\mu')$, is dependent upon the temperature profile, $T(z)$, and the dielectric loss profile, $\kappa_\mathrm{a}(z)$, or, equivalently, upon the moisture profile so that recovering soil moisture from an observed brightness temperature requires knowledge or an assumption about these profiles. A common assumption is that temperature and moisture are constant with depth. This assumption can cause errors in 1.4 GHz, h-polarized brightness of 5 to 10 K for moist soils during the heat of an afternoon. A better approach is to use predicted temperature and moisture profiles from an SVAT model.

The upwelling brightness temperature in the moist soil of Figure 7 is partially scattered back into the soil by the wave impedance mismatch at the land–atmosphere interface. The strength of the reflection depends upon the dielectric

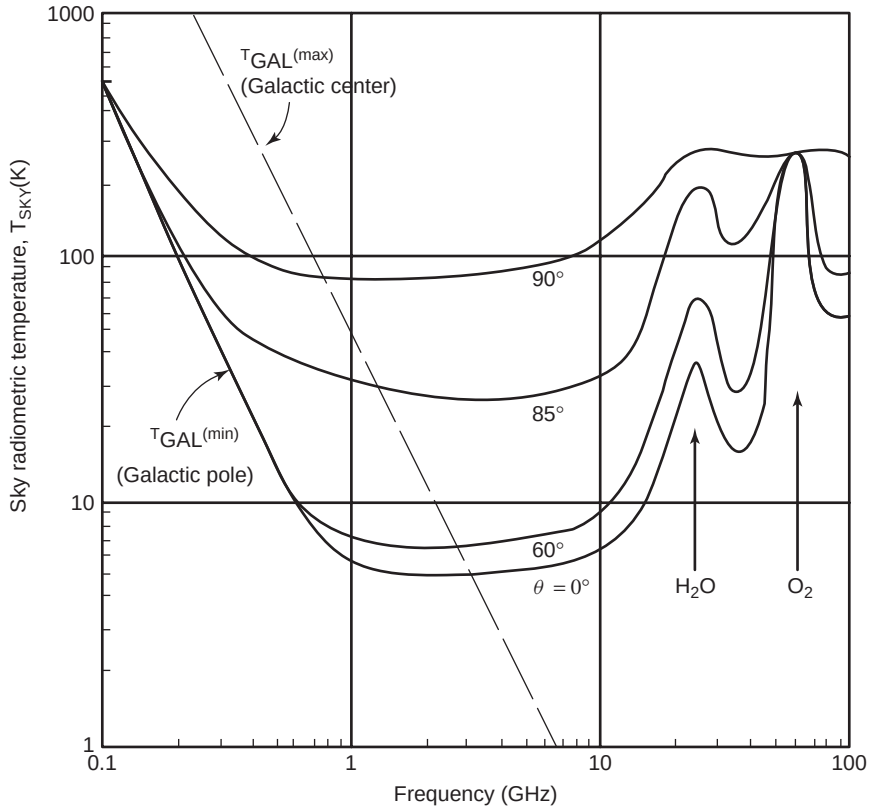


Figure 8 Microwave sky brightness temperature at zenith angles of 0°, 60°, 85°, and 90° for a Standard Atmosphere having a water vapor density of 7.5 g m −3 at sea level. Our galaxy, the Milky Way, is a significant radiometric source at frequencies below 1 GHz. The brightness temperature of a Planck (thermal) source would be a horizontal line on this plot. The −2 slope of the galactic contribution indicates a synchrotron emission process (charges accelerated in magnetic fields). Over a small region around the galactic center, the galactic contribution exceeds 10 K in the 1.4 GHz radiometric window but is negligible by 7 GHz, the next higher common radiometric window. In directions away from the galactic center, the contribution at 1.4 GHz is typically ∼1 K. Emission from the galactic center is a practical consideration only when the sky is being used to calibrate a 1.4 GHz radiometer. Spectral features at 22 GHz and 60 GHz are a water vapor absorption line and the oxygen complex, respectively (Reproduced from Ulaby *et al.*, 1981 by permission of Pearson Education (Addison-Wesley))

contrast between soil and air. If there were no reflection, the emitted brightness temperature would be very close to the physical temperature of the soil at the interface. The large dynamic range in brightness temperature exhibited by soil of varying moisture content is caused by the large variation in reflectivity at the soil–atmosphere interface. This is most easily understood by assuming a quasi-specular soil–vegetation interface. The ratio of reflected to incident wave amplitudes, $\gamma_p(\mu)$, for upwelling radiance in a homogeneous soil is governed by the Fresnel reflection coefficient for polarization p,

$$\gamma_v(\mu) = \left(\frac{\tilde{n}\mu - \mu'}{\tilde{n}\mu + \mu'}\right)$$
$$\gamma_h(\mu) = \left(\frac{\tilde{n}\mu' - \mu}{\tilde{n}\mu' + \mu}\right) \qquad (14)$$

where $\tilde{n}$ is the complex index of refraction of soil at the interface, and μ' and μ are direction cosines of the ray below and above the soil–air interface, respectively. The direction cosines are related by Snell's law,

$$n\sqrt{1-\mu'^2} = \sqrt{1-\mu^2} \qquad (15)$$

The reflectivity for an upwelling brightness temperature at the interface is then

$$R_p(\mu) = \gamma_p(\mu) \cdot \gamma_p^*(\mu) \qquad (16)$$

where $\gamma_p^*(\mu)$ is the complex conjugate of $\gamma_p(\mu)$. If the reflectivity of a quasi–specular interface is $R_p(\mu)$, then its emissivity in direction μ is $e_p(\mu) = (1 - R_p(\mu))$ and the brightness temperature emitted by the soil is

$$Tb_p^s(0^+, \mu) = e_p(\mu) \cdot Tb_p^s(0^-, \mu') \qquad \text{K sr}^{-1} \qquad (17)$$

where $Tb_p^s(0^+, \mu)$ and $Tb_p^s(0^-, \mu')$ refer to the upwelling soil brightness temperature incrementally above and below the interface, respectively. Soil emissivities derived from Fresnel reflection coefficients are good approximations to experimental observations if surfaces are smooth relative to a wavelength and moisture contents vary slowly with depth relative to a wavelength. When surfaces are rough or soils have dried at their immediate surface, Fresnel models yield low estimates of emissivity. Techniques for achieving better estimates require surface roughness statistics if surfaces are not smooth, or temperature and moisture profiles from SVAT models if near-surface temperature or moisture gradients are significant.

Examples of emissivity for a silty loam soil for several percentages of Volumetric Soil Moisture (%VSM) are shown in Figure 9. For a uniform soil half-space temperature of 290 K, the dynamic range of $Tb_h^s(0^+, 1)$ at 1.4 GHz between saturation (~40%VSM) and wilting point (~8%VSM) is ~85 K. For most soils, brightness temperature decreases ~3 K per %VSM increase. For many purposes, knowledge of moisture content to ~4%VSM rms is adequate. Radiometers can be calibrated to better than 2 K and are readily sensitive to 0.5 K. The large dynamic range combined with excellent sensitivity, accuracy, and a minimum of confusing factors in the brightness temperature are reasons that microwave radiometry at frequencies below ~10 GHz is such a powerful tool for sensing soil moisture.

Vegetation canopies are a "confusing factor" of the first order, but ones that are often easily modeled. At lower microwave frequencies, thicknesses of leaves in vegetation canopies are small with respect to a wavelength so that leaves do not scatter but do absorb and emit. This absorption and emission is overwhelmingly by the water within and intercepted upon the leaves. Leaf water can be modeled as uniformly distributed moisture in a homogeneous canopy layer having dielectric properties that are volume-weighted averages of the complex indices of refraction of air, dry leaf material, bound water in leaves, and free water within and upon leaves (England and Galantowicz, 1994). A dual-dispersion model can be used to estimate the dielectric permittivity of water in the leaves (Ulaby and El-Rayes, 1987). The consequent index of refraction of the homogeneous canopy layer, or "cloud" canopy layer, is sufficiently near unity that reflection and refraction at the canopy–air interface can be ignored. Where woody stems, branches, and trunks become a significant fraction of the microwave wavelength, scattering within the canopy will require a second order correction. For a prairie grass canopy, scattering can be ignored at frequencies of 19 GHz and below (England and Galantowicz, 1995). For a corn canopy, scattering

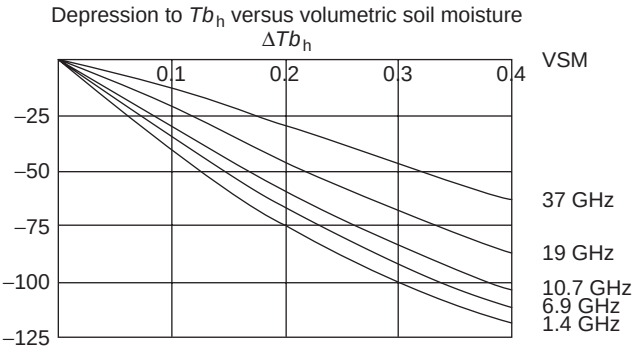


Figure 9 Brightness temperature depression (ΔTb_h) versus Volumetric Soil Moisture (VSM). A silty loam soil having a porosity of 40% is modeled as a uniform half-space having a quasi-specular surface and a temperature just above freezing. The dielectric soil model is from Dobson *et al.* (1985). The depression of the *h*-polarized brightness temperature at a 35° incidence angle is shown as a function of moisture content for 1.4, 6.9, 10.7, 19, and 37 GHz, the five common frequencies of satellite radiometers. The plots are misleading in that they indicate little loss of sensitivity to soil moisture as radiometric frequency increases from 1.4 to 10.7 GHz and only a factor of 2 loss in going to 37 GHz. Three effects not included in this simple model significantly reduce sensitivity to soil moisture with increasing frequency. These are surface roughness, moisture gradients, and vegetation cover, and are discussed in the text. The combination of these effects strongly favor 1.4 GHz for sensing soil moisture

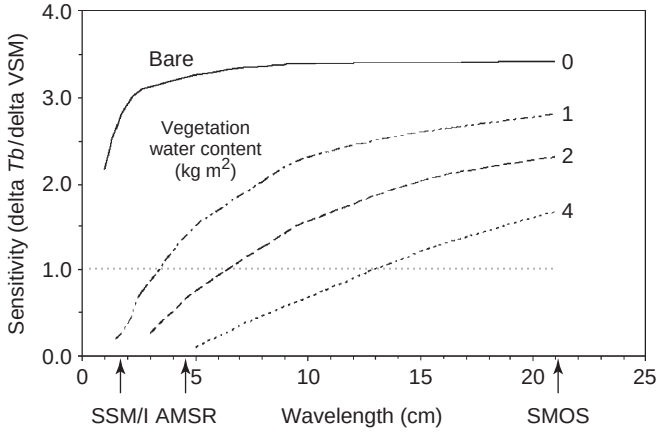


Figure 10 Sensitivity of microwave brightness to moisture in soil under a vegetation canopy. These curves relate the sensitivity of microwave brightness to radiometric wavelength for bare soil (denoted by 0) and for various column densities of liquid water (in kg m^{-2}) in a uniform vegetation canopy. Land-surface modelers typically would like to know volumetric soil moisture (VSM) to something like 4%VSM rms. The usable sensitivity of microwave brightness to soil moisture is ~1 K/%VSM (horizontal dashed line at 1.0 K/%VSM. For example, the figure indicates that microwave brightness at all wavelengths is usably sensitive to moisture in bare soil but, for a canopy moisture column density of 4 kg m^{-2}, about that of mature corn, only the 21-cm wavelength (1.4 GHz) of the SMOS radiometer offers usable sensitivity

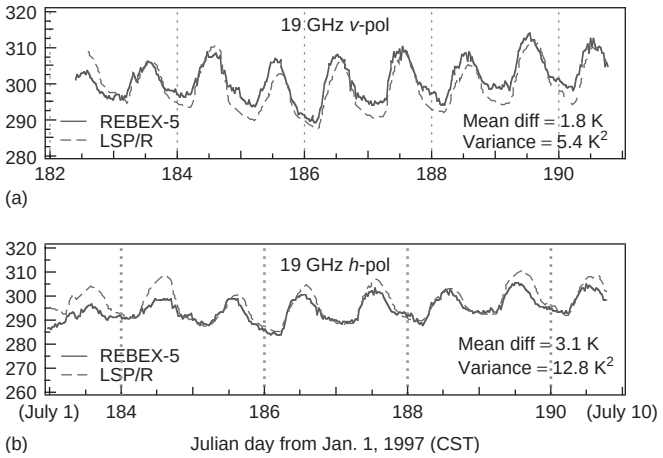


Figure 11 Performance of Radiobrightness model during SGP'97. The field contained mixed wheat stubble and short grass. Panels (a) and (b) compare measured (solid line) and modeled (dashed line) 19 GHz brightness temperatures at the SSM/I's incidence angle of 53°. Panel (a) is vertically polarized brightness temperature and Panel (b) is horizontally polarized brightness temperature

will contribute a few Kelvin error at 1.4 GHz (Hornbuckle *et al.*, 2003).

Propagation within the cloud canopy model is governed by water content expressed as a column density in $kg\,m^{-2}$. Liquid water in and on leaves results in emission and absorption that tends to mask changes in upwelling brightness temperature related to changes in %VSM. The practical limit of useful sensitivity to soil moisture is ~1 K per %VSM. The water column density of short grass might be $1\,kg\,m^{-2}$ and that of mature corn might be $5\,kg\,m^{-2}$. Figure 10 illustrates that a 10 GHz radiometer usefully senses soil moisture through a $\sim 1\,kg\,m^{-2}$ canopy, while for a 1.4 GHz radiometer the limiting canopy column density is nearer $5\,kg\,m^{-2}$. Galactic noise has established the 1.4 GHz protected radio astronomy band as the lowest usable radiometric band. The limits inferred from Figure 10 are consistent both with models and field observations (Hornbuckle and England, 2004). Performance of the R model using these concepts is illustrated in Figure 11.

SATELLITE RADIOMETERS FOR SENSING OF SHALLOW SOIL MOISTURE AND SNOW WATER EQUIVALENT (SWE)

Satellite microwave radiometers that have had, or will have, application to land-surface hydrology appear in Table 2 (e.g. Kramer, 1996). One of the first of these, the Scanning Multichannel Microwave Radiometer (SMMR), was launched on both Seasat and Nimbus-7 spacecrafts in 1978. Seasat failed after only 70 days, but SMMR on Nimbus-7 produced data until it was turned off in July 1988. Figures 5 and 10 suggest that both the 6.6 and 10.7 GHz SMMR channels would have been sensitive to moisture in soil and snow. SMMR's 6.6 GHz data were little used because of their poor quality, but combinations of its 10.7, 18, and 37 GHz data were used in algorithms to classify soils as frozen or thawed, and combinations of 18 and 37 GHz data were used in algorithms to estimate SWE.

In June 1987, the US Defense Meteorological Satellite Program (DMSP) began a series of SSM/I on satellites in LEO. Data from the lowest frequency SSM/I channel, 19 GHz, are sensitive to canopy and soil wetness after rainfall, but their frequency is too high for quantitative estimates of soil moisture where there is any vegetation within the 50-km footprint of the radiometer. Combinations of 19 and 37 GHz SSM/I data are used in algorithms for soil freeze/thaw classification and for SWE.

The Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI) was launched in November 1997. TMI is an evolution of SSM/I but includes 10.7 GHz channels in addition to SSM/I channels. The spatial resolution for the SSM/I channels is twice that of the SSM/I because the instrument is in a lower orbit. Penalties of the lower altitude and inclination are less frequent coverage, observations are not Sun-synchronous, and coverage is limited to $\pm 39°$ of latitude.

Versions of the Advanced Microwave Scanning Radiometer (AMSR) were launched on the National Aeronautics and Space Administration (NASA) Aqua satellite in May 2002 and on the National Space Development Agency (NASDA) ADEOS-2 satellite in December 2002. The ADEOS-2 spacecraft failed in October 2003, but AMSR-E on Aqua continues. The 6.9 GHz channels of AMSR are sensitive to soil moisture through a short grass canopy (Figure 10) if Radio Frequency Interference (RFI), possibly from airports, did not render these channels nearly useless over many land areas. Data from the 10.7 GHz channel appear to be free of RFI over North America but may be experiencing some RFI over Europe.

The National Polar-Orbiting Operational Environmental Satellite System (NPOESS), a cooperative program among the Department of Commerce, (National Oceanic and Atmospheric Administration (NOAA)), the Department of Defense, and NASA, will launch the Conical Scanning Microwave Imager/Sounder (CMIS) in 2008. CMIS will have the AMSR channels plus 60 GHz oxygen- and 183 GHz water vapor channels. It retains the AMSR 2 m aperture and, consequently, its spatial resolution of ~50 km at 6.9 GHz.

The first 1.4 GHz image data from LEO likely will be from ESA's (European Space Agency) Soil Moisture Ocean Salinity (SMOS) mission in 2007 (Kerr *et al.*, 2001). SMOS will be followed by NASA's HYDROS, a 1.4 GHz active/passive mission, in 2010 (see Figure 12) (Entekhabi *et al.*, 2004). The 1.4 GHz brightness temperatures from

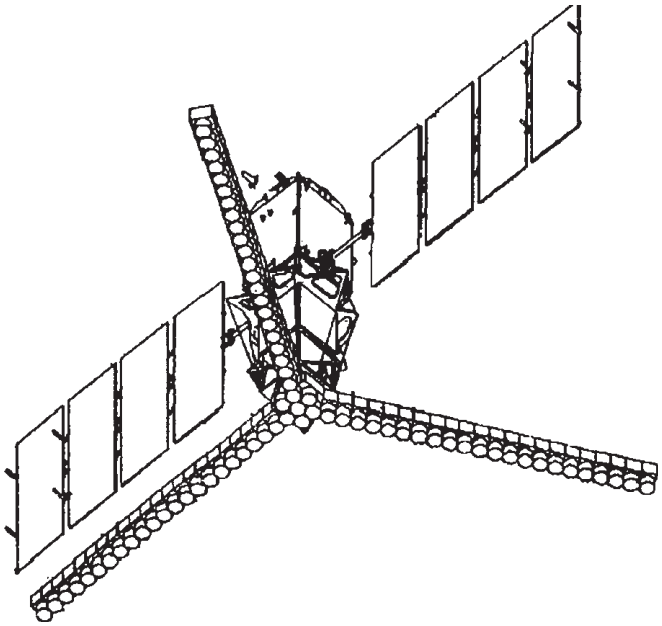


Figure 12 The European Space Agency's (ESA's) Soil Moisture Ocean Salinity (SMOS) satellite. SMOS will be a 1.4 GHz aperture synthesis radiometer in LEO. It is to be launched in 2007 and likely will be the first aperture synthesis radiometer in orbit

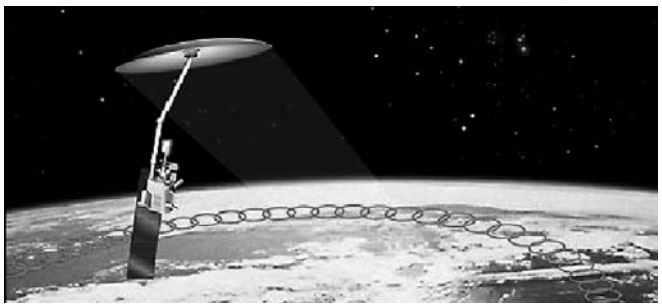


Figure 13 HYDROS is a 1.4 GHz active/passive sensor for mapping soil moisture. HYDROS is a NASA Earth System Science Pathfinder (ESSP) Mission being developed for launch in 2010. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

these satellites will have spatial resolutions of ~40 km and will be sensitive to soil moisture through all but forest canopies. SMOS achieves its spatial resolution with the new aperture synthesis technology – a first for a satellite instrument. NASA's HYDROS (see Figure 13) will be the first active/passive system. The HYDROS strategy is to enhance effective spatial resolution by combining 40 km brightness data with higher spatial resolution radar data to produce a 10-km soil moisture product and a 3-km soil freeze/thaw product.

Fundamentals of a Radiometer

Satellite microwave radiometers are characterized by three primary parameters – spatial resolution or footprint (the 3-dB contour of the antenna beam projected on the ground), uncertainty in noise equivalent temperature or sensitivity (NEΔT), and calibration accuracy (a bias in Kelvins) – and three secondary parameters – antenna radiation efficiency, η_l, antenna beam efficiency, η_a, and receiver noise equivalent temperature, T_{rec}. Even casual users of brightness data should be familiar with the primary parameters. Those who require a quantitative assessment of data quality should understand all six parameters.

An excellent discussion of radiometer fundamentals can be found in Skou (1989). An ideal total power radiometer and its conceptual model are shown schematically in Figures 14(a) and 14(b), respectively. The radiometer is "ideal" because its receiver is linear, gain does not vary with time, and sensitivity is the best achievable for a given

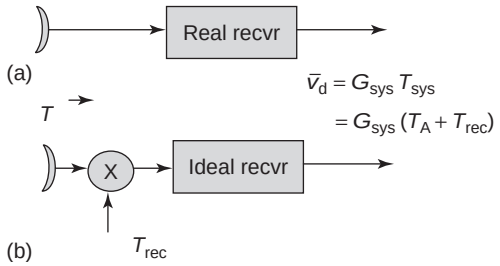


Figure 14 Functional sketch of a radiometer (a) and its model (b). The antenna temperature, T_A, represents the signal and noise delivered to the receiver (Recvr) by the antenna. The ''real'' receiver amplifies T_A and adds its own receiver noise represented by T_{rec}, in the radiometer model. T_{rec} is a fictitious noise added immediately behind the antenna in (b) followed by an ''ideal'' receiver that adds no noise to the output signal. The sum of $T_A + T_{rec} = T_{sys}$ is the system noise. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

system noise and integration period. The radiometer is "total power" because it measures absolute noise equivalent antenna temperatures. In contrast, a Dicke radiometer, typical of the previous generation of radiometers, measures the difference between noise equivalent antenna temperature and noise equivalent reference load temperature. Dicke architecture was used to minimize the effects of nonlinearity and gain instability. Modern radiometers are exceptionally linear and sufficiently stable in gain to be treated as ideal for the few second periods between internal calibrations.

The Antenna

Most radiometer antennas are designed to have narrow, nearly circular beams with side-lobes well below the −13 dB shown from the uniformly illuminated aperture in Figure 15. The directivity of such an antenna, $D_p(\hat{i})$, indicates radiant power of polarization p in direction $\hat{i}$ relative to the radiant power of a hypothetical isotropic radiator emitting equivalent power. Antenna gain is related to directivity by $G_p(\hat{i}) = \eta_l D_p(\hat{i})$ where η_l is radiation efficiency. Radiation efficiencies less than unity are caused by ohmic loss in the antenna. Typical values are ∼0.95 at 1.4 GHz and ∼0.85 at 35 GHz.

A radiometer antenna collects brightness temperatures, $T b_p(\hat{i})$, from all directions to achieve an antenna temperature, T_A,

$$T_A = \int_{4\pi} \frac{D_p(\hat{i})}{4\pi} T b_p(\hat{i})\, d\Omega \qquad K \tag{18}$$

where $d\Omega$ is incremental solid angle in direction $\hat{i}$. If T_o is the physical temperature of an antenna in Kelvin,

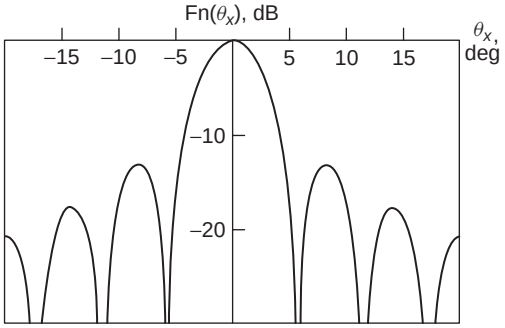


Figure 15 An example of the normalized radiation pattern of a square aperture antenna whose sides are parallel to the x and y axes, and whose normal lies along the z axis. The normalized radiation pattern represents the radiant intensity in a particular direction normalized by maximum radiant intensity. It is measured or modeled at distances from the antenna that are large with respect to any aperture dimension so that the antenna can be considered a point source. The plot is a $y = 0$ cut through the pattern for an aperture whose size to wavelength ratio is 10, and whose radiating field in the aperture plane is uniform in amplitude and phase. The θ_x dimension is angle from the $\hat{z}$ direction measured toward the $\hat{x}$ direction. The maximum lobe, centered on the θ_x, is referred to as the Main lobe. Other lobes are referred to as side-lobes. Because the 1st side-lobe of this antenna is only 13 dB below the Main lobe, it would be considered a poor radiometer antenna. Too much energy would be collected from unwanted directions. Side-lobes of radiometer antennas should be more than 20 dB below the Main lobe. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

then the antenna temperature delivered to a receiver (see Figure 14a) is

$$T'_A = \eta_l T_A + (1 - \eta_l) T_o \qquad \text{or}$$

$$T'_A = \int_{4\pi} \frac{G_p(\hat{i})}{4\pi} T b_p(\hat{i})\, d\Omega + (1 - \eta_l) T_o \qquad K \tag{19}$$

The 3-dB beam width, $\beta_{1/2}$, defines an antenna's angular resolution. The main beam is the solid angle between nulls at $\pm\beta_{null}/2$, and main beam efficiency, η_a, is

$$\eta_a = \frac{\int_{main\ beam} G_p(\hat{i})\, d\Omega}{\int_{4\pi} G_p(\hat{i})\, d\Omega} \tag{20}$$

Communications antennas might have main beam efficiencies less than 0.85, which means that side-lobes account for greater than 15% of the antenna's collecting area. Side-lobes collect energy from undesirable directions and are suppressed in the design of radiometer antennas. Radiometer antennas typically have radiation efficiencies well above 0.90, but achieve this at the cost of some angular resolution.

The projection of the 3-dB beam contour onto the ground defines the footprint of the radiometer. For conically scanning radiometers, like the SSM/I, AMSR, TMI, and CMIS, this footprint is approximately elliptical. The relationship between the radiometer's scanning cone half angle with respect to nadir, θ, and the incidence angle on a spherical Earth, θ', is

$$\sin\theta' = \left(1 + \frac{h}{a_o}\right)\sin\theta \tag{21}$$

where h is satellite altitude above the Earth's surface in km and a_o is the Earth's radius ($a_o = 6400$ km). For example, the SSM/I cone half angle of $\theta = 45°$ and orbit altitude of 833 km results in an incidence angle $\theta' = 53°$ at the Earth's surface. The semimajor axis, X_o, and semiminor axis, Y_o, of the footprint is

$$X_o = \frac{h + a_o(1 - \cos(\theta' - \theta))}{\cos\theta\cos\theta'}\beta_{1/2} \qquad \text{km}$$
$$Y_o = \frac{h + a_o(1 - \cos(\theta' - \theta))}{\cos\theta}\beta_{1/2} \qquad \text{km} \tag{22}$$

where beamwidth, $\beta_{1/2}$, is in radians. The angular resolution for the 19 GHz channel of the SSM/I is 0.0346 radians which corresponds to $X_o = 73$ km and $Y_o = 44$ km. The 19 GHz ground resolution is often quoted as 50 km, which is a slightly optimistic average of the principal axes. The National Snow and Ice Data Center (NSIDC) uses a Backus–Gilbert technique to optimally resample the SSM/I data to a fixed Earth Grid, the Equal Area Scaling Earth–Grid (EASE-Grid), which has a 25-km spacing between grid points. This 25-km data is effectively an oversampling of the original data. The spatial resolution of each resampled point remains $\sim 70 \times 40$ km.

The Receiver

The receiver in Figure 14(a) amplifies the signal from the antenna and adds its own noise. The receiver noise equivalent temperature is modeled in Figure 14(b) as T_{rec} inserted at a point immediately behind the antenna followed by a noise-free receiver. The total noise equivalent temperature, or system noise equivalent temperature, is always defined at the output terminals of the antenna, that is,

$$T_{sys} = T'_A + T_{rec} \qquad \text{K} \tag{23}$$

Voltage samples of a thermal noise signal exhibit a zero-mean Gaussian distribution. The amplitudes of these voltages exhibit a Rayleigh distribution, and the square of the amplitudes, v_d, exhibit an exponential distribution. The square of the amplitude, v_d, is proportional to power or to T_{sys}, so that the mean of v_d is

$$\overline{v_d} = G_{sys}T_{sys} \qquad \text{V} \tag{24}$$

where G_{sys} is the system gain in V K^{-1}. The standard deviation of an exponential distribution is equal to its mean. That is, an estimate of T_{sys} based upon one sample of v_d has uncertainty equal to T_{sys}. Equivalently, the Noise Equivalent uncertainty in Temperature, $NE\Delta T$, for one sample is $NE\Delta T_1 = T_{sys}$ where $NE\Delta T$ is the standard deviation of the expected error in the estimate of T_{sys}. Because thermal noise is a stationary random process, N independent samples of T_{sys} improves the estimate of T_{sys} by $\sqrt{N}$, that is, the $NE\Delta T$ for N samples is

$$NE\Delta T_N = \frac{T_{sys}}{\sqrt{N}} \qquad \text{K} \tag{25}$$

Typical digital radiometers might have system temperatures of $\sim$500 K. Equation (25) requires that $N = 25 \times 10^6$ samples to achieve a desired uncertainty in T_{sys} of 0.1 K. The equivalent for an analog radiometer is $N = \Delta f\tau$ where Δf is the system bandwidth and τ is an integration time. A radiometer whose T_{sys} is 500 K and whose bandwidth is 25 MHz requires an integration time of 1.0 s to achieve an $NE\Delta T$ of 0.1 K.

Equation (25), or its analog equivalent, is an accurate indicator of $NE\Delta T$ if all of the uncertainty results from the statistics of the noise signal. Fluctuations in G_{sys} can add rms contributions that exceed the thermal noise contribution if the physical temperatures of the RF amplifiers are not adequately controlled. This gain instability, caused by fluctuations in temperature among elements within a radiometer, is less of a problem with modern radiometers because their small size and low power consumption allow thermal control to $\sim$20 mK rms. Absolute accuracy always depends upon external calibration.

Calibration

The inherent linearity of microwave radiometers generally permits a two-point calibration as illustrated in Figure 16. Satellite radiometers often use the extraterrestrial sky as the cold load and a heated absorber/emitter as the hot load. Antennas of conically scanning radiometers, like the SSM/I and AMSR, view their calibration loads once per revolution of the antenna. The SSM/I spins at 32 rpm which corresponds to a calibration every 1.9 s.

Calibration is conventionally described by the equations

$$\overline{v_d} = a(T_A + b) \tag{26}$$

where a and b are referred to as gain and offset, respectively. From equations (19), (23), and (24),

$$\overline{v_d} = G_{sys}T_{sys}$$
$$= G_{sys}(T'_A + T_{rec})$$

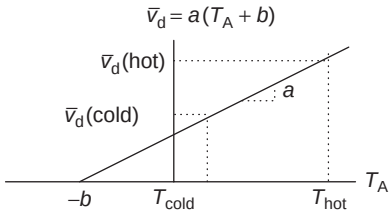


Figure 16 Calibration of a microwave radiometer. The relationship between antenna temperature, T_A, and the average power delivered by a well-designed microwave radiometer (represented by $\bar{v}_d$) is very nearly linear. This linearity allows a two-point calibration, T_{cold} and T_{hot}, to determine gain, 'a', and offset, 'b'. T_{cold} might be the cold sky and T_{hot} might be a black emitter at ambient temperature. Conceptually, 'b' can be thought of as the noise equivalent contribution of the receiver, T_{rec}, and 'a' as the system gain, G_{sys}. In practice, they are not quite equivalent so it is necessary to preserve the distinction between 'a' and G_{sys}, and between 'b' and T_{rec}

$$= G_{sys}(\eta_l T_A + (1-\eta_l)T_o + T_{rec})$$
$$= G_{sys}\eta_l\left(T_A + \left(\frac{1}{\eta_l}\right)((1-\eta_l)T_o + T_{rec})\right) \quad (27)$$

yielding gain and offset,

$$a = G_{sys}\eta_l$$
$$b = \left(\frac{1}{\eta_l}\right)((1-\eta_l)T_o + T_{rec}) \quad (28)$$

The approximation $\eta_l = 1$, a lossless antenna, yields the useful relationships; the slope of the calibration line in Figure 16 is system gain, that is, $a = G_{sys}$, and the $\overline{v_d} = 0$ intercept is the negative of receiver noise equivalent temperature, that is, $b = -T_{rec}$.

IMPROVED SPATIAL RESOLUTION

The section "Stored water as a modeled product" identified stored water as a relevant model-based parameter that could be estimated from a combination of SVAT/R models and assimilated remotely sensed shallow soil moisture. The section "Satellite radiometers for sensing of shallow soil moisture and Snow Water Equivalent (SWE)" identified 1.4 GHz radiometry as the most usefully sensitive single technique for sensing shallow soil moisture. The principal difficulty with their combination is achieving an acceptable spatial resolution. Various criteria are used to define acceptable resolution. The most generous criterion might be the need to resolve a moisture track of a significant precipitation event. Meeting this criterion requires something like a spatial resolution of 10 km. A more aggressive criterion might be to resolve individual agricultural fields so that vegetation canopies within an instrument's footprint would be more nearly homogeneous. Fields in the American Midwest are dominantly a quarter section. This translates to a more demanding spatial resolution of ~1 km. A truly challenging criterion might be to resolve significant spatial heterogeneities in the SWE of an alpine snow pack. Meeting this criterion might require a spatial resolution of 100 m.

Three resolution-enhancing technologies are being pursued. None are likely to yield a 100-m spatial resolution at 1.4 GHz, but the radiometric techniques might achieve 10 km and a combination of radar and radiometry might achieve 1 km. These are disaggregation of brightness temperature observations, aperture synthesis radiometry, and combined active/passive microwave sensing.

Disaggregation of Brightness Temperature Observations

If two distinct terrain types lie within a pixel, the mixed pixel can be tiled with an area-weighted aggregate of SVAT/R models of each terrain. The observed brightness temperature for a pixel is then disaggregated to minimize the sum of squared errors between modeled and observed brightness temperature attributed to each tile. These disaggregated observations are then assimilated into appropriate SVAT/R model. The simplest approach would assume equal confidence in expected brightness temperatures for each tile. A more sophisticated approach would include expected errors in each of the model-based brightness temperatures. Disaggregation of observations has only recently been tried (e.g. Burke *et al.*, 2001; and Burke *et al.*, 2002).

Aperture Synthesis Radiometry

Aperture synthesis radiometry, or Synthetic Thinned Array Radiometry (STAR), is an adaptation of interferometric radio astronomy (e.g. Thompson *et al.*, 1994). Instead of viewing the heavens, STAR radiometers view the Earth either from aircraft or from spacecraft (e.g. LeVine, 1990; LeVine *et al.*, 1994; Camps *et al.*, 1997; Martin-Neira *et al.*, 2002; and Bayle *et al.*, 2002). STAR allows spatial resolutions with sparse arrays that are equivalent to those of filled arrays of a similar circumscribed aperture (see Figure 17), but at a cost of increased signal processing complexity and some loss sensitivity. STAR also offers the flexibility of beam-forming in postprocessing and a reduced aperture flatness requirement. Where the requirement for a real aperture antenna might be a specified shape to 1/20th of a wavelength, the STAR requirement is a more forgiving combination of flatness and knowledge of element location. A hybrid STAR, the 1.4 GHz Electronically Scanned Thinned Array Radiometer (ESTAR) (LeVine *et al.*, 1990), operated successfully on a NASA aircraft between 1990 and

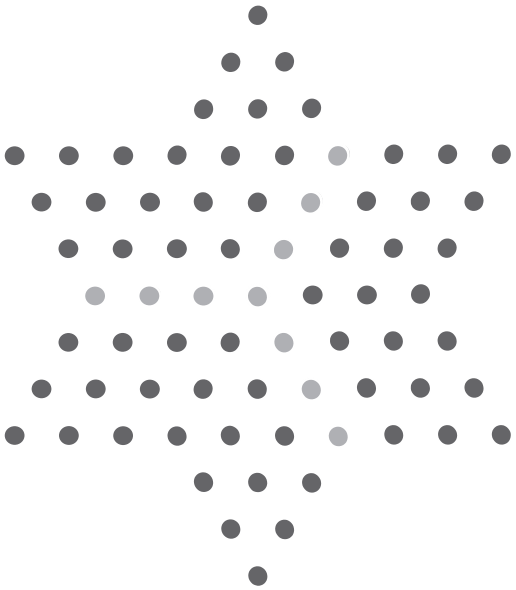


Figure 17 Equivalent filled-aperture antenna. The gray dots represent the actual locations of antenna and receiver elements of an aperture synthesis radiometer, and the combination of gray and black dots represent locations of antenna and receiver elements of an equivalent filled-aperture radiometer. That is, the spatial resolution of this 10-element aperture synthesis radiometer would be equivalent to the spatial resolution of this 73-element real aperture system. The cost is loss of sensitivity in the aperture synthesis system. Whether the cost is acceptable depends upon the application. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2003. ESTAR used a real aperture along-track and aperture synthesis across-track. STAR-Light (see Figure 18), a 1.4 GHz, 2-dimensional STAR radiometer being developed for arctic land-surface hydrology, will achieve a compactness that is compatible with a light aircraft rather than require the 4-engine turboprop aircraft used by previous generations of instruments. ESA's SMOS instrument (see Figure 12) will be a 1.4 GHz, two-dimensional STAR sensor.

STAR implementations become simpler as radiometers become increasingly digital and compact. A Direct Sampling Digital Radiometer (DSDR) avoids mixers by digitally sampling the 1.4 GHz RF signal (Fischman and England, 1999). A correlation DSDR based upon two 1.4 GHz DSDR radiometers achieved analog performance with 3-bit sampling (Fischman *et al.*, 2002). Early DSDRs required 90 dB of RF amplification prior to sampling with commercially available Analog-to-Digital Converters (ADCs). Newer commercial ADCs now require only 70 dB of amplification and some custom ADCs require as little as 45 dB. DSDRs used off-chip ceramic filters but Micro-Electro-Mechanical Systems (MEMS) mechanical filters

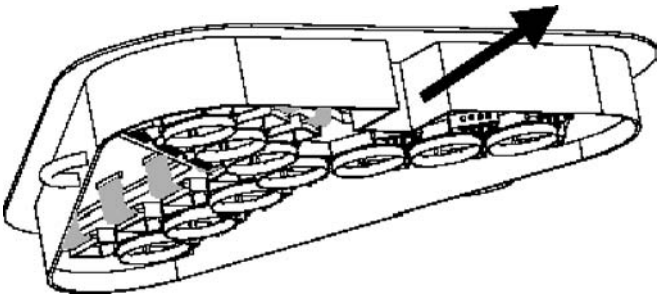


Figure 18 STAR-Light. STAR-Light is a 10 element, 1.4-GHz aperture synthesis radiometer designed to fly on a light aircraft in support of land-surface hydrology in the Arctic. The instrument is approximately 1 m on an edge. The arrow indicates direction of flight. The Office of Polar Programs of the National Science Foundation (NSF) supported development of STAR-Light through Critical Design Review and element prototype fabrication and testing. The prototype elements were used in the Cold Lands Processes Experiment (CLPX) in Colorado during the winter and spring of 2003

have recently operated at 1.4 GHz. We are rapidly approaching the possibility of a complete 1.4 GHz radiometer on a chip with antenna temperature and timing signal in, and digital data out.

Combined Active/Passive Microwave Sensing

Satellite radar has the salient advantage of superior spatial resolution. One hundred meter spatial resolution from LEO with a sensitivity that is appropriate for soil moisture retrieval is within the capability of current technology. In principle, radar and radiometry at the same frequency have similar sensitivities to soil moisture (Du *et al.*, 2000). In practice, scattering from woody elements of a canopy, rough soil surfaces, ice lenses in snow packs, and combinations of these can easily overwhelm the subtle differences caused by variations in soil moisture. The difficulty of soil moisture or SWE retrieval from radar backscatter alone can be appreciated by analogy. Radar sensing through a vegetation canopy and a rough soil surface or through a tree canopy and embedded scatterers in a snow pack is similar to optically viewing objects through frosted glass when the objects are illuminated by a source on the viewer's side of the glass. Light backscattered by the glass tends to overwhelm light reflected by objects behind the glass. Radiometry is like viewing objects through frosted glass when the illumination is on the objects' side of the glass. Where vegetation is relatively sparse, polarimetric radars and multifrequency radars have been used to unravel the contributions of soil moisture or SWE from those of soil surface roughness. As scattering becomes more complex, radar is increasingly limited to the qualitative detection of temporal or spatial changes in soil moisture or SWE. In

these cases, radiometry may play the essential role of calibration of the overall radar scene average. The literature on these investigations is extensive and beyond the scope of this review.

Satellite radar for soil moisture is more difficult to interpret, is of higher cost relative to satellite radiometry, and typically has unacceptably long intervals between repeat coverage. Even with these disadvantages, radar is still the only technology that will achieve the high spatial resolution that some land-surface hydrology demands. The emerging compromise is a combination of radar and radiometry. NASA's HYDROS (see Figure 13) will be the first example of a passive/active microwave system (Entekhabi *et al.*, 2004). The science behind these complementary data is not yet mature, but will be the focus of many investigations over the next decade.

REFERENCES

Bayle F., Wigneron J.-P., Kerr Y.H., Waldteufel P., Anterrieu E., Orlhac J.-C., Chanzy A., Marloie O., Bernardini M., Sobjaerg S., Calvet J.-C., Goutoule J.-M. and Skou N. (2002) Two-dimensional synthetic aperture images over a land surface scene. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 710–714.

Bohren C.F. and Huffman D.R. (1983) *Absorption and Scattering of Light by Small Particles*, Wiley Press: p. 530.

Bonan G. (1996) *A Land Surface Model (LSM Version 1.0) for Ecological, Hydrological, and Atmospheric Studies: Technical Description and User's Guide*, Technical Report NCAR/TN-417+STR, Climate and Global Dynamics Division, NCAR.

Burke E., Bastidas L.A. and Shuttleworth W.J. (2002) Exploring the potential for multipatch soil-moisture retrievals using multiparameter optimization techniques. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 1114–1120.

Burke E., Shuttleworth W., Lee K.-H. and Bastidas L. (2001) Using area-average remotely sensed surface soil moisture in multipatch land data assimilation systems. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 2091–2100.

Camps A., Bara J., Sanahuja I.C. and Torres F. (1997) The processing of hexagonally sampled signals with standard rectangular techniques: application to 2-D large aperture synthesis interferometric radiometers. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 132–139.

Chandrasekhar S. (1960) *Radiative Transfer*, Dover Press: p. 393.

Chang A.T.C., Foster J.L., Hall D.K., Rango A. and Hartline B.K. (1982) Snow water equivalent estimation by microwave radiometry. *Cold Regions Science and Technology*, **5**, 259–267.

Chang A.T.C., Foster J.L., Hall D.K., Robinson D.A., Peiji L. and Meisheng C. (1992) The use of microwave radiometer data for characterizing snow storage in western China. *Annals of Glaciology*, **16**, 26–30.

Chang A.T.C., Foster J.L. and Rango A. (1991) Utilization of surface cover composition to improve the microwave determination of snow water equivalent in a mountain basin. *International Journal of Remote Sensing*, **12**, 2311–2319.

Chang A.T.C. and Tsang L. (1992) A neural network approach to inversion of snow water equivalent from passive microwave measurements. *Nordic Hydrology*, **23**, 173–182.

Chreny I.V. and Raizer V.Y. (1998) *Passive Microwave Remote Sensing of Oceans*, Wiley: p. 195.

Crow W.T., Drusch M. and Wood E.F. (2001) An observation system simulation experiment for the impact of land surface heterogeneity on AMSR-E soil moisture retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1622–1631.

Davis R.E., Dozier J. and Chang A.T.C. (1987) Snow property measurements correlative to microwave emission at 35 GHz. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-25**, 751–757.

Davis D.T., Chen Z., Hwang J.-N., Tsang L. and Njoku E. (1995) Solving inverse problems by Bayesian iterative inversion of a forward model with applications to parameter mapping using SMMR remote sensing data, *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 1182–1193.

Derksen C., Goddison B., Ledrew E. and Walker A. (2003) Combining SMMR and SSM/I data for time series analysis of central North American snow water equivalent. *Journal of Hydrometeorology*, **4**, 304–316.

Derksen C., LeDrew E., Walker A. and Goodison B. (2000) Influence of sensor overpass time on passive microwave-derived snow cover parameters. *Remote Sensing of Environment*, **71**, 297–308.

de Vries D. (1958) Simultaneous transfer of heat and moisture in porous media. *Transactions-American Geophysical Union*, **39**, 909–916.

de Vries D. (1963) *Thermal Properties of Soils, Physics of Plant Environment*, Interscience Press: pp. 210–235.

Dickinson R., Henderson-Sellers A. and Kennedy P. (1993) *Biosphere-Atmosphere Transfer Scheme (BATS) Version 1e as Coupled to the NCAR Community Climate Model*, Technical Report NCAR/TN-387+STR, Climate and Global Dynamics Division, NCAR.

Dickinson R., Henderson-Sellers A., Kennedy P. and Wilson M. (1986) *Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model*, Technical Report NCAR/TN-275+STP, Atmospheric Analysis and Prediction Division, NCAR.

Dobson M., Ulaby F., Hallikainen M. and El-Rayes M. (1985) Microwave dielectric behavior of wet soil – part II: Dielectric mixing models. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-23**, 35–46.

Du Y., Ulaby F.T. and Dobson M.C. (2000) Sensitivity to soil moisture by active and passive microwave sensors. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 105–114.

England A.W. (1974) Thermal microwave emission from a halfspace containing scatterers. *Radio Science*, **9**, 447–454.

England A.W. (1975) Thermal microwave emission from a scattering layer. *Journal of Geophysical Research*, **80**, 4484–4496.

England A.W. (1976) Relative influence upon microwave emissivity of fine-scale stratigraphy, internal scattering and dielectric properties. *Pure and Applied Geophysics*, **114**, 287–299.

England A.W. (1990) Radiobrightness of diurnally heated, freezing soil. *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 464–476.

England A.W. and Galantowicz J.F. (1994) A volume emission model for the radiobrightness of prairie grass, *Proceedings of IGARSS'94*, Pasadena, August 8–12.

England A.W. and Galantowicz J.F. (1995) Observed and modeled radiobrightness of prairie grass in early fall, *Proceedings of IGARSS'95*, Florence, July 10–12.

England A.W. and Johnson G.R. (1977) Microwave brightness spectra of layered media. *Geophysics*, **42**, 514–521.

England A.W. and Johnson G.R. (1978) Spectral gradient of lunar radiobrightness: Heat flow or volume scattering? *Journal of Research, USGS*, **6**, 505–509.

England A.W., Galantowicz J.F. and Schretter M.S. (1992) The radiobrightness thermal inertia measure of soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 132–139.

Entekhabi D., Nakamura H. and Njoku E.G. (1994) Solving the inverse problem for soil moisture and temperature profiles by sequential assimilation of multifrequency remote sensing observations. *IEEE Transactions on Geoscience and Remote Sensing*, **32**, 438–448.

Entekhabi D., Njoku E.G., Houser P., Spencer M., Doiron T., Kim Y., Smith J., Girard R., Belair S., Crow W., *et al.* (2004) The hydrosphere state (Hydros) satellite mission: an earth system pathfinder for global mapping of soil moisture and land freeze/thaw. *IEEE Transactions on Geoscience and Remote Sensing*, **42**, 2184–2195.

Ferrazzoli P., Paloscia S., Pampaloni P., Schiavon G., Solimini D. and Coppo P. (1992) Sensitivity of microwave measurements to vegetation biomass and soil moisture content: a case study. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 781–783.

Ferrazzoli P., Wigneron J.-P., Guerriero L. and Chanzy A. (2000) Multifrequency emission of wheat: modeling and applications. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 2598–2607.

Fischman M. and England A.W. (1999) Sensitivity of a 1.4 GHz direct-sampling digital radiometer. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 2172–2180.

Fischman M.A., England A.W. and Ruf C.S. (2002) How digital correlation affects the fringe washing function in L-Band aperture synthesis radiometry. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 671–679.

Foster J.L., Chang A.T.C., Hall D.K. and Rango A. (1991) Derivation of snow water equivalent in boreal forests using microwave radiometry. *Remote Sensing of Arctic Environments, Arctic*, **44**, 147–152.

Galantowicz J.F. and England A.W. (1997) Seasonal snow-pack radiobrightness interpretation using a SVAT-linked emission model. *Journal of Geophysical Research*, **102**(D18), 21 933–21 946.

Galantowicz J.F., Entekhabi D. and Njoku E.G. (1999) Tests of Sequential Data Assimilation for Retrieving Profile Soil Moisture and Temperature from Observed L-Band Radiobrightness. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 1860–1870.

Galantowicz J.F., Entekhabi D. and Njoku E.G. (2000) Estimation of soil-type heterogeneity effects in the retrieval of soil moisture from radiobrightness. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 312–315.

Goodison B., Barry E. and Walker A.E. (1993) Use of snow cover derived from satellite passive microwave data as an indicator of climate change. *Annals of Glaciology*, **17**, 137–142.

Guha A., Jacobs J.M., Jackson T.J., Cosh M.H., Hsu E.-C. and Judge J. (2003) Soil moisture mapping using ESTAR under dry conditions from the Southern Great Plains Experiment (SGP99). *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 2392–2397.

Guo J.-J., Tsang L., Josberger E.G., Wood A.W., Hwang J.-N. and Lettenmaier D.P. (2003) Mapping the spatial distribution and time evolution of snow water equivalent with passive microwave measurements. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 612–621.

Hall D.K., Foster J.L. and Chang A.T.C. (1984) Nimbus-7 SMMR polarization responses to snow depth in the mid-western U.S.*Nordic Hydrology*, **15**, 1–7.

Hallikainen M.T. (1992) Comparison of algorithms for retrieval of snow water equivalent from Nimbus-7 SMMR data in Finland. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 124–131.

Hallikainen M.T. and Jolma P.A. (1986) Retrieval of the water equivalent of snow cover in Finland by satellite microwave radiometry. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-24**, 855–862.

Hollinger J., Lo R., Poe G., Savage R. and Pierce J. (1987) *Special Sensor Microwave/Imager User's Guide*, Naval Research Laboratory, Washington, DC, pp. 120.

Hornbuckle B.K. and England A.W. (2004) Radiometric sensitivity to soil moisture at 1.4 GHz through a corn crop at maximum biomass. *Water Resources Research*, **40**, W10204, doi:10.1029/2003WR002931, pp. 12.

Hornbuckle B.K., England A.W., DeRoo R.D., Fischman M.A. and Boprie D. (2003) Vegetation Anisotropy at 1.4 GHz. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 2211–2223.

Houser P.R., Shuttleworth W.J., Famiglietti J.S., Gupta H.V., Syed K.H. and Goodrich D.C. (1998) Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resources Research*, **34**(12), 3405–3420, 10.1029/1998WR900001.

Jackson T.J. (1993) Measuring surface soil moisture using passive microwave remote sensing. *Hydrological Processes*, **7**, 139–152.

Jackson T.J. (2001) Multiple resolution analysis of L-band brightness temperature for soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 151–164.

Jackson T.J., Engman E.T., Le Vine D.M., Schmugge T.J., Lang R., Wood E. and Teng W. (1994) Multitemporal passive microwave mapping in MACHYDRO'90. *IEEE Transactions on Geoscience and Remote Sensing*, **32**, 201–206.

Jackson T.J., Gasiewski A.J., Oldak A., Klein M., Njoku E.G., Yevgrafov A., Christiani S. and Bindlish R. (2002) Soil moisture retrieval using the C-band polarimetric scanning radiometer during the Southern Great Plains 1999 Experiment. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 2151–2161.

Jackson T.J. and Lakshmi V. (2001) Introduction to the special issue on large scale passive microwave remote sensing of soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1619–1620.

Jackson T.J., Le Vine D.M., Hsu A.Y., Oldak A., Starks P.J., Swift C.T., Isham J.D. and Haken M. (1999) Soil Moisture mapping at regional scales using microwave radiometry: the Southern Great Plains Hydrology Experiment. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 2136–2151.

Jackson T.J., Le Vine D.M., Swift C.T., Schmugge T.J. and Schiebe F. (1995) Large area mapping of soil moisture using ESTAR passive microwave radiometer in Washita'92. *Remote Sensing of Environment*, **53**, 27–37.

Jackson T.J. and O'Neill P.E. (1990) Attenuation of soil and microwave emission by corn and soybeans at 1.4 and 5 GHz. *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 978–980.

Jackson T.J., O'Neill P.E. and Swift C.T. (1997) Passive microwave observation of diurnal surface soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 1210–1222.

Jackson T.J. and Schmugge T.J. (1989) Passive microwave remote sensing system for soil moisture: some supporting research. *IEEE Transactions on Geoscience and Remote Sensing*, **27**, 225–235.

Jackson T.J., Schmugge T.J. and Wang J.R. (1982) Passive microwave remote sensing of soil moisture under vegetation canopies. *Water Resources Research*, **18**, 1210–1222.

Janssen M.A. (1993) *Atmospheric Remote Sensing by Microwave Radiometry*, Wiley Press: p. 572.

Jin Y.Q. (1997) Snow depth inferred by scattering indices of SSM/I channels in a mesh graph. *International Journal of Remote Sensing*, **18**, 1843–1849.

Judge J., Abriola L.M. and England A.W. (2003) Development and numerical validation of a summertime Land Surface Process and Radiobrightness model. *Advances in Water Resources*, **26/7**, 733–746.

Judge J., England A.W., Crosson W.L., Laymon C.A., Hornbuckle B.K., Boprie D.L., Kim E.J. and Liou Y.A. (1999) A growing season Land Surface Process/Radiobrightness Model for Wheat-Stubble in the Southern Great Plains. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 2152–2158.

Judge J., Galantowicz J.F. and England A.W. (2001) A comparison of ground-based and satellite-borne microwave radiometric observations in the Great Plains. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1686–1696.

Judge J., Galantowicz J.F., England A.W. and Dahl P. (1997) Freeze/thaw classification for prairie soils using SSM/I radiobrightnesses. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 827–832.

Karam M.A. (1997) A pysical model for microwave radiometry of vegetation. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 1045–1058.

Kelly R.E., Chang A.T.C., Tsang L. and Foster J.L. (2003) A prototype AMSR-E global snow area and snow depth algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 230–242.

Kerr Y.H., Waldteufel P., Wigneron J.-P., Martinuzzi J.-M., Font J. and Berger M. (2001) Soil moisture retrieval from space: the soil moisture and ocean salinity (SMOS) mission. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1729–1735.

Kharuk V.I., Ranson K.G., Burenina T.A., Onuchin A.A. and Fedotova Y.V. (2000) Microwave remote sensing as a method for estimating snow reserves in forests of the Western Sayan. *Mapping Sciences & Remote Sensing*, **37**, 172–179.

Kimball B., Jackson R., Reginato R., Nakayama F. and Idso S. (1976) Comparison of field-measured and calculated soil-heat fluxes. *Soil Science Society of America Journal*, **40**, 18–25.

Kim E.J. and England A.W. (2003) A Year-Long Comparison of Plot-Scale and Satellite Footprint-Scale 19 and 37 GHz Brightness of the Alaskan North Slope, *Journal of Geophysical Research*, **108**(D13), 4388, 10.1029/2002JD002393.

Koike T. and Suhama T. (1993) Passive-microwave remote sensing of snow. *Annals of Glaciology*, **18**, 305–308.

Kramer H.J. (1996) *Observations of the Earth and its Environment*, *Third Edition*, Springer Press: p. 960.

Lai S., Tiedje J. and Erickson A. (1976) In situ measurement of gas diffusion coefficients in soils. *Soil Science Society of America Proceedings*, **40**, 3–6.

Lakshmi V. (2000) A simple surface temperature assimilation scheme for use in land surface models. *Water Resources Research*, **36**, 3687–3700.

LeVine D. (1990) The sensitivity of a synthetic aperture radiometer for remote sensing applications from space. *Radio Science*, **25**, 441–453.

LeVine D., Griffis A., Swift C. and Jackson T.J. (1994) ESTAR: a synthetic aperture microwave radiometer for remote sensing applications. *Proceedings of the IEEE*, **82**, 1787–1801.

LeVine D., Kao K., Tanner A., Swift C. and Griffis A. (1990) Initial results in the development of a synthetic aperture radiometer. *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 614–619.

Lettenmaier D.P. and Rind D. (Eds.) (1992) Report of Chapman conference on hydrological aspects of global climate change. *Journal of Geophysical Research*, **97**, 2675–2833.

Liang X., Lettenmaier D.P., Wood E. and Burges S. (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, **99**, 14 415–14 428.

Liang X., Lettenmaier D.P., Wood E. and Burges S. (1996) One-dimensional statistical dynamical representation subgrid spatial variability of precipitation in the 2-layer variable infiltration capacity model. *Journal of Geophysical Research*, **101**, 21 403–21 422.

Liou Y.A. and England A.W. (1996) Annual temperature and radiobrightness signatures for bare soils. *IEEE Transactions on Geoscience and Remote Sensing*, **34**, 981–990.

Liou Y.A. and England A.W. (1998a) A Land Surface Process/Radiobrightness model with coupled heat and moisture transport for freezing soils. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 669–677.

Liou Y.A. and England A.W. (1998b) A Land Surface Process/Radiobrightness model with coupled heat and moisture transport for freezing soils. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 669–677.

Liou Y.A., Galantowicz J.F. and England A.W. (1999) A Land Surface Process/ Radiobrightness model with coupled heat and

moisture transport for prairie grassland. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 1848–1859.

Liou Y.A., Kim E.J. and England A.W. (1998) Radiobrightness of prairie soil and grassland during dry-down simulations. *Radio Science*, **33**, 259–265.

Liu S.-F., Liou Y.-A., Wang W.-J., Wigneron J.-P. and Lee J.-B. (2002) Retrieval of crop biomass and soil moisture from measured 1.4 and 10.65 GHz brightness temperatures. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 1260–1268.

Macelloni G., Paloscia S., Pampaloni P. and Tedesco M. (2001) Microwave emission from dry snow: a comparison of experimental and model results. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 2649–2656.

Martin-Neira M., Ribo S. and Martin-Polegre A.J. (2002) Polarimetric mode of MIRAS. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 1755–1768.

McFarland M.J., Wilke G.D. and Harder P.H. II (1987) Nimbus 7 SMMR investigation of snowpack properties in the northern Great Plains for the winter of 1978–1979. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-25**, 3–46.

Mo T., Choudhury B.J., Schmugge T.J., Wang J.R. and Jackson T.J. (1982) A model for microwave emission from vegetation-covered fields. *Journal of Geophysical Research*, **87**, 11,229–11,237.

Newton R.W. and Rouse J.W. Jr (1980) Microwave radiometer measurements of soil moisture content. *IEEE Transactions on Antennas And Propagation*, **Ap-28**, 680–686.

Njoku E.G. and Li L. (1999) Retrieval of land surface parameters using passive microwave measurements at 6–18 GHz. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 79–93.

NRC (1991) *Opportunities in the Hydrologic Sciences, Committee on Opportunities in the Hydrologic Sciences*, National Research Council, National Academy Press: Washington, p. 348.

NRC (1998) *Global Energy and Water Cycle Experiment (GEWEX) Continental-Scale International Project (GCIP); A Review of Progress and Opportunities*, National Research Council, National Academy Press: Washington, p. 93.

O'Neill P.E., Chauhan N.S. and Jackson T.J. (1996) Use of active and passive microwave remote sensing for soil moisture estimation through corn. *International Journal of Remote Sensing*, **17**, 1851–1865.

Owe M., de Jeu R. and Walker J. (2001) A methodology for surface soil moisture and vegetation optical depth retrieval using the microwave polarization difference index. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1643–1654.

Pellarin T., Calvet J.-C. and Wigneron J.-P. (2003) Surface soil moisture retrieval from L-band radiometry: a global regression study. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 2037–2051.

Philip J. and de Vries D. (1957) Moisture movement in porous materials under temperature gradients. *Transactions-American Geophysical Union*, **38**, 222–232.

Pulliainen J.T., Grandell J. and Hallikainen M.T. (1999) HUT snow emission model and its applicability to snow water equivalent retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 1378–1390.

Pulliainen J.T. and Hallikainen M.T. (2001) Retrieval of regional snow water equivalent from space-borne passive microwave observations. *Remote Sensing of Environment*, **75**, 76–85.

Rawls W.J., Gish T.J. and Brakensiek D.L. (1991) Estimating soil water retention from soil physical properties and characteristics. *Advances in Soil Science*, **16**, 213–233.

Reichle R. and McLaughlin D. (2001) Downscaling of radiobrightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. *Water Resources Research*, **37**, 2353–2364.

Reichle R., McLaughlin D. and Entekhabi D. (2001) Variational data assimilation of microwave radiobrightness observations for land surface hydrologic applications. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1708–1718.

Richards L.A. (1931) Capillary conduction of liquids in porous mediums. *Physics*, **1**, 318.

Rosenfeld S. and Grody N. (2000) Anomalous microwave spectra of snow cover observed from Special Sensor Microwave/Imager measurements. *Journal of Geophysical Research: Atmospheres*, **105**, 14 913–14 925.

Rossi C. and Nimmo J. (1994) Modeling of soil water retention from saturation to oven dryness. *Water Resources Research*, **30**, 701–708.

Schmugge T.J. (1978) Remote sensing of surface soil moisture. *Journal of Applied Meteorology*, **17**, 1549–1557.

Schmugge T.J., Gloersen P., Wilheit T. and Geiger F. (1974) Remote sensing of soil moisture with microwave radiometers. *Journal of Geophysical Research*, **79**, 317–323.

Schmugge T.J., O'Neill P.E. and Wang J.R. (1986) Passive microwave soil moisture research. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-24**, 12–22.

Schmugge T.J., Wang J.R. and Asrar G. (1988) Results from the Push Broom Microwave Radiometer flights over the Konza Prairie in 1985. *IEEE Transactions on Geoscience and Remote Sensing*, **26**, 590–596.

Sellers P., Mintz Y., Sud Y. and Dalcher A. (1986) A simple biosphere model SiB for use within general circulation models. *Journal of Atmospheric Chemistry*, **43**, 505–531.

Shi J., Chen K.S., Li Q., Jackson T.J., O'Neill P.E. and Tsang L. (2002) A parameterized surface reflectivity model and estimation of bare-surface soil moisture with L-band radiometer. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 2674–2686.

Schwank M., Stahli M., Wydler H., Leuenberger J., Mätzler C. and Fluhler H. (2004) Microwave L-band emission of freezing soil. *IEEE Transactions on Geoscience and Remote Sensing*, **42**, 1252–1261.

Skou N. (1989) *Microwave Radiometer Systems: Design and Analysis*, Artech House Press: p. 162.

Srivastav S.K. and Singh R.P. (1991) Microwave radiometry of snow-covered terrains. *International Journal of Remote Sensing*, **12**, 2117–2131.

Thompson A.R., Moran J.M. and Swenson G.W. Jr (1994) *Interferometry and Synthesis in Radio Astronomy*, Krieger Press: p. 534.

Trenberth K.E. (1995) *Climate System Modeling*, Cambridge Press: p. 788.

Tsang L. and Kong J.A. (2001) *Scattering of Electromagnetic Waves: Advanced Topics*, Wiley Press: p. 413.

Ulaby F. and El-Rayes M. (1987) Microwave dielectric spectrum of vegetation-Part II: Dual-dispersion model. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-25**, 550–557.

Ulaby F.T., Moore R.K. and Fung A.K. (1981) *Microwave Remote Sensing: Active and Passive*, Addison-Wesley Press: Vol 1, p. 456.

Ulaby F.T., Moore R.K. and Fung A.K. (1982) *Microwave Remote Sensing: Active and Passive*, Addison-Wesley Press: Vol 2, pp. 457–1064.

Ulaby F.T., Moore R.K. and Fung A.K. (1986) *Microwave Remote Sensing: Active and Passive*, Artech House: Vol 3, pp. 1065–2162.

Ulaby F.T., Razani M. and Dobson M.C. (1983) Effects of vegetation cover on the microwave radiometric sensitivity to soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-21**, 51–61.

Vörösmarty C.J., Hinzman L.D., Peterson B.J., Bromwich D.H., Hamilton L.C., Morison J., Romanovsky V.E., Sturm M. and Webb R.S. (2000) *The Hydrologic Cycle and its Role in Arctic and Global Environmental Change: A Rationale and Strategy for Synthesis Study*, Arctic Research Consortium of the U.S.: Fairbanks, p. 84.

Walker J. and Houser P. (2001) A methodology for initializing soil moisture in a global climate model: Assimilation of near-surface soil moisture observations. *Journal of Geophysical Research*, **106**, 11 761–11 774.

Wang J.R., Newton R.W. and Rouse J.W. Jr (1980a) Passive microwave remote sensing of soil moisture: The effect of tilled row structure. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-18**, 296–302.

Wang J.R., Shiue J.C., Chuang S.L., Shin R.T. and Dombrowski M. (1984) Thermal microwave emission from vegetated fields: A comparison between theory and experiment. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-22**, 143–150.

Wang J.R., Shiue J.C. and McMurtrey J.E. III (1980b) Microwave remote sensing of soil moisture *content* over bare and vegetated fields. *Geophysical Research Letters*, **7**, 801–804.

Wang J.R., Shiue J.C., Schmugge T.J. and Engman E.T. (1990) The L-band PBMR measurements of surface soil moisture in FIFE. *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 906–914.

Wigneron J.-P., Calvet J.-C. and Kerr Y. (1996) Monitoring water interception by crop fields from passive microwave observations. *Agricultural and Forest Meteorology*, **80**, 177–194.

Wigneron J.-P., Laguerre L. and Kerr Y. (2001) A simple parameterization of the L-band microwave emission from rough agricultural soils. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1697–1707.

Wigneron J.-P., Calvet J.-C., Kerr Y., Chanzy A. and Lopes A. (1993) Microwave emission of vegetation: Sensitivity to leaf characteristics. *IEEE Transactions on Geoscience and Remote Sensing*, **31**, 716–726.

Wigneron J.-P., Olioso A., Calvet J.-C. and Bertuzzi P. (1999) Estimating root zone soil moisture from surface soil moisture data and soil-vegetation-atmosphere transfer modeling. *Water Resources Research*, **35**, 3735–3745.

Wilson L.L., Tsang L., Hwang J.N. and Chen C.T. (1999) Mapping snow water equivalent by combining a spatially distributed snow hydrology model with passive microwave remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 690–704.

Zuerndorfer B. and England A.W. (1992) Radiobrightness decision criteria for freeze/thaw boundaries. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 89–102.

Zuerndorfer B., England A.W., Dobson M.C. and Ulaby F.T. (1990) Mapping freeze/thaw boundaries with SMMR data. *Journal of Agriculture and Forest Meteorology*, **52**, 199–225.

48: Ground-based and Airborne Lidar

R MICHAEL HARDESTY AND LISA S DARBY

NOAA Environmental Technology Laboratory, Boulder, CO, US

Lidar remote sensing techniques can be applied to measure a wide variety of atmospheric parameters important in the hydrological sciences, including aerosol distribution, cloud properties, ozone and water vapor concentration, and wind fields. Lidar measurements make use of scattering and extinction of laser radiation in the atmosphere. Instruments are deployed on surface and aircraft platforms to investigate phenomena such as moisture spatial distribution and temporal evolution, ozone sources and transport, boundary-layer height, aerosol distribution, cloud microphysics, and wind structure. New satellite-based instruments enable lidar techniques to be applied on a global scale.

INTRODUCTION

Light Detection And Ranging (Lidar) refers to an active remote sensing technique with significant application in the hydrological sciences for remote probing of atmospheric quantities such as water vapor, cloud and aerosol particles, and winds. Lidar remote sensing involves the transmission and scattering of light to remotely probe the atmosphere. Fundamentally, lidar methodology is directly analogous to radar, differing primarily by the wavelength of the transmitted and received electromagnetic energy (*see* **Chapter 63, Estimation of Precipitation Using Ground-based, Active Microwave Sensors, Volume 2**). Radar systems used for atmospheric observations operate in the radio wave and microwave portions of the electromagnetic spectrum, with wavelengths ranging from about 1 mm to 10 m, while atmospheric lidar systems utilize ultraviolet, visible, and infrared light at wavelengths from approximately 250 nm to 10 μm. The different operating wavelengths result in very different atmospheric scattering and transmission properties. Energy at radar wavelengths is scattered efficiently by rain, clouds, and meter-size atmospheric density inhomogeneities, but is only weakly scattered by atmospheric molecules and aerosol particles. Energy at optical wavelengths, on the other hand, is efficiently scattered by atmospheric aerosol particles and molecules but strongly attenuated by most clouds and heavy precipitation. Consequently, the two instruments are typically deployed in different and sometimes complementary applications. Radars are used to investigate phenomena such as precipitating storms and cloud microphysics, and to measure winds based on the motion of hydrometeors or refractive index inhomogeneities. Lidars are applied to characterize atmospheric density, trace species, and aerosol structure, and to measure wind speeds in optically clear air by observing the movement of molecules and aerosol particles.

In a lidar, a laser transmitter directs a pulse of monochromatic light into the atmosphere. As the light propagates, some of it is scattered by the molecular and aerosol constituents of the atmosphere. Scattering from aerosols (Mie scattering) is elastic (at the same wavelength as the incident beam), while scattering from molecules has both elastic and Raman components. The Raman-scattered light is much weaker than the elastic-scattered portion and is shifted in wavelength by an amount uniquely characteristic of the scattering molecule. If the scatterer is moving toward or away from the lidar due to wind or thermal motion, the wavelength of the scattered radiation is also shifted because of the Doppler Effect. Some of the propagating light may be attenuated by certain species or particles present within the beam that selectively absorb radiation at the transmitted wavelength. Scattering and absorption processes in the atmosphere reduce the intensity of both the forward propagating beam and the backscattered radiation (light scattered directly back toward the source). The lidar receiver collects, detects, and analyzes the backscattered light (the portion of the light scattered directly back to the receiver) as a function of time. By measuring the time elapsed since the laser pulse

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

was transmitted, the range to the portion of atmosphere from which information is received at a given time is determined. The detected backscattered signal contains information on the distribution, concentration, and optical properties of atmospheric constituents, and may also be analyzed to estimate wind velocities and atmospheric turbulence.

SCATTERING AND TRANSMISSION OF LASER RADIATION IN THE ATMOSPHERE

The scattering and transmission properties of light in the atmosphere are fundamental to lidar measurements. The scattering properties of an atmospheric particle depend on its refractive index, shape, size and orientation. In the special case when the particle is much smaller than the wavelength of the incident radiation, as is the case for laser radiation scattered by atmospheric molecules, the scattering process is characterized as *Rayleigh* scattering. In the Rayleigh scattering regime, the energy backscattered by a particle increases proportionally as the inverse of the fourth power of the wavelength. Consequently, lidar systems based on molecular scatter usually operate at short wavelengths, that is, in the visible or ultraviolet spectral regions, to maximize the backscattered signal.

Aerosol particles, the other component of the atmosphere that scatters laser light, result in *Mie* scattering, which applies when the diameter of the scatterers is on the order of or larger than the incident wavelength. Atmospheric aerosol particles that scatter laser light include organic and inorganic particles, such as sulfates and nitrates, dust, soot, smoke, and pollen, as well as liquid water and ice. Although for Mie scattering the relationship between the power backscattered by an ensemble of aerosol particles and the incident wavelength is not simply characterized, typically, the backscattered energy increases roughly proportional to the first or second power of the inverse of the incident wavelength, depending on the refractive index and shape of the scattering particles. In a polluted environment, such as in the vicinity of urban areas, an abundance of large particles often results in a roughly linear relationship between the inverse of the incident wavelength and the backscattered energy, while in more pristine environments such as the free troposphere, the inverse wavelength/backscatter relationship can approach or exceed a square-law relationship. Mie scattering is asymmetric, with more light scattered in the forward direction, and less scattered back toward the lidar.

Figure 1 shows a typical spectrum of elastic backscattered light collected at the lidar receiver for a volume of atmosphere irradiated by a monochromatic laser pulse. Molecular scattering produces the broadband distribution in Figure 1, where the broadening results from the Doppler shifts of the radiation backscattered from molecules moving randomly at their thermal velocities. The width of the molecular velocity distribution in the atmosphere ranges

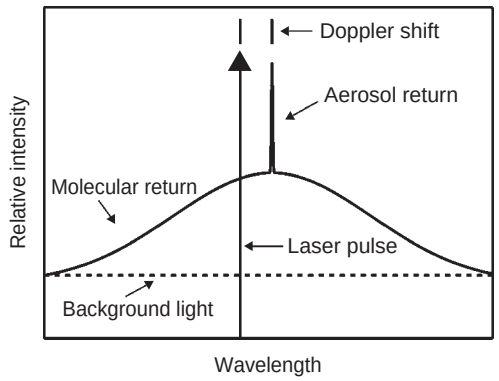


Figure 1 Spectrum of atmospheric elastic backscatter for illumination by a single frequency laser source, showing narrowband aerosol return, wideband molecular return, and Doppler shift of the spectrum resulting from particle motion

from about 320 to 350 $m\,s^{-1}$, scaling as the square root of the temperature. In the center of the spectrum is a much narrower peak resulting from scattering of the light by aerosol particles. Because the thermal velocity of the much larger aerosol particles is very low, the width of the distribution of the aerosol return is determined by the range of velocities of particles moved about by turbulence within the scattering volume. This width is typically on the order of a few $m\,s^{-1}$. Also shown in Figure 1 is an additional broadband distribution due to scattered solar radiation collected at the receiver. In Figure 1 the laser source is shown as monochromatic; if the laser source is not monochromatic, the backscattered signal spectrum is additionally broadened, with the resulting spectrum being the convolution of the spectrum shown in Figure 1 with the spectrum of the laser pulse. As seen in the figure, the entire spectrum of the backscattered radiation is Doppler-shifted in frequency relative to the frequency of the laser pulse because of the mean wind. Because wind velocities are in the order of tens of $m\,s^{-1}$ or less, the mean Doppler shift is less than the width of the molecule-scattered spectrum, but can be much greater than the aerosol-scattered spectrum width.

Both Rayleigh and Mie scattering processes are elastic, where the wavelength of the scattered signal is not altered by the scattering process. Scattering from atmospheric atoms and molecules also includes Raman scatter, which results from a process in which energy is exchanged between the scattered photon and the scattering species. A fundamental characteristic of Raman scatter is the shift of the wavelength of the scattered light from that of the incident light by an amount that is uniquely characteristic of the scattering molecule. As a result, the Raman-scattered spectrum can be analyzed to identify the molecule (e.g., nitrogen or water vapor), determine its number density, and estimate its energy state by comparing the populations in different energy levels. The Raman scattering component

of backscattered light is typically two or three orders of magnitude weaker than the Rayleigh-scattered radiation, however, use of high-quality optical filters to separate the Raman-scattered light enables identification and monitoring of specific atmospheric constituents such as nitrogen and water vapor.

Some of the laser light propagating toward and back from the scattering volume may also be absorbed by specific atmospheric molecules and particles along the propagation path. Different gases and particles typically have distinctive absorption characteristics. The absorption cross sections of aerosol particles usually vary slowly with wavelength, while some molecular absorption spectra may be highly structured, particularly in the infrared portion of the electromagnetic spectrum. The technique known as *differential absorption lidar* (DIAL) takes advantage of this fine structure to measure gaseous species concentrations. In DIAL applications, two closely spaced wavelengths are transmitted, with the wavelengths chosen such that the radiation at one wavelength is more strongly absorbed by the species of interest than at the second wavelength. By comparing the difference in attenuation across a common portion of the transmit/receive path, the concentration of the species of interest may be calculated.

LIDAR SYSTEM CHARACTERISTICS

The basic components that make up a lidar system include the laser transmitter, transmission optics to direct the laser beam into the atmosphere, a telescope to collect the light, and a detector package to convert photons to an electrical signal for analysis and storage. The specific characteristic of each component depends on the application for which the instrument is developed. A primary system characteristic is the laser transmitter wavelength. Rayleigh backscatter and Raman lidar systems designed to estimate atmospheric constituent concentrations and density operate in the visible or ultraviolet region of the spectrum where scattering from molecules is strongest. Lidars for characterizing aerosol properties are more likely to employ lasers transmitting at longer wavelengths, ranging from the visible to the near-infrared or even thermal-infrared regions (i.e., from 0.5 μm to 11 μm wavelengths), where the backscatter from aerosol particles relative to molecular scatter is higher. DIAL system transmitters include two wavelengths, and can incorporate separate lasers or a single tunable laser. Because absorption features of atmospheric gases occur over many regions of the spectrum, DIAL systems operate anywhere from the ultraviolet (e.g., to measure ozone) to the infrared (for water vapor measurements).

The versatility of lidars for observing atmospheric characteristics has resulted in a wide variety of deployment configurations and platforms. Although most lidar systems are still used in research configurations involving operators and intermittent operation, an increasing number of instruments has been demonstrated for stand-alone, unattended operation. Over the past 5 years, autonomous lidar systems that measure aerosol backscatter and depolarization, cloud and aerosol extinction, water vapor and ozone concentrations, and wind shear have been deployed in both short-term and extended observational programs. Lidar systems have also been deployed on aircraft and ships to profile atmospheric properties during a wide variety of field measurement campaigns.

The remainder of this entry focuses on the use of lidars for atmospheric probing. However, lidar methodology is also applicable for a variety of other geophysical activities. Most notably, lidar altimeters employing short pulses have been deployed on aircraft and satellites to map surface topography and coastal bathymetry. Airborne lidars have also been used to investigate sea surface–wave structure, and to locate and characterize fish schools and floating ocean debris.

LIDARS FOR MEASUREMENT OF CLOUD AND AEROSOL PROPERTIES

Aerosol particles suspended in the atmosphere play an important role in a variety of weather and climate processes. For example, increases in anthropogenic aerosols over the past several decades are thought to have mitigated some of the global warming effects of increased greenhouse gases by reflecting solar radiation back to space. Because certain small aerosol particles serve as cloud condensation nuclei, their presence can change the microphysical and radiative properties of clouds, which on large scales can alter the global radiation balance. Also, layers of absorbing aerosols characterized by significant fractions of elemental carbon can alter the thermodynamic structure of the atmosphere, stabilizing or de-stabilizing the atmosphere depending on the height of the absorbing aerosol layer.

Aerosols are air pollutants. Because of the health hazards associated with inhalation of small particles, in 1997, the US Environmental Protection Agency established air quality limits specifying maximum permissible mass concentrations of particles below 2.5 μm in size. Aerosol particles when present in sufficient concentrations also limit visibility, especially in humid regions where the particles can contribute to local smog by increasing their size through deliquescence. The issue of visibility reduction in pristine areas such as national parks has drawn considerable public and research attention in recent years.

Lidar is an excellent tool for investigating the structure and transport of atmospheric aerosol particles in weather and climate research. The signal received from the atmosphere by an aerosol backscatter lidar is expressed in the

single-scattering lidar equation (Collis and Russell, 1976):

$$P_r(R) = P_0 \left(\frac{c\tau}{2}\right) \beta(R) A_r R^{-2} \exp\left[-2 \int_0^R \sigma(r)\, \mathrm{d}r\right] \quad (1)$$

where P_r is the instantaneous received power from range R, P_0 is the transmitted power at time t_0 (the time when the laser pulse is transmitted), c is the velocity of light, τ is the pulse duration, β is the volume backscattering coefficient of the atmosphere, R is the range, A_r is the effective area of the receiver telescope, and σ is the extinction coefficient for the atmosphere through which the beam propagates. Equation (1) is valid for a rectangular pulse and assumes that the volume backscattering coefficient β does not vary over the scale of the pulse. Note that equation (1) refers to the power received by the lidar, not the power detected and analyzed.

Equation (1) is indeterminate because both the backscatter and extinction coefficients are unknown. However, because backscatter lidars frequently operate at wavelengths in the visible or infrared spectral regions where atmospheric absorption is minimal, the absorption term in equation (1) can often be neglected to provide an uncalibrated estimate of the relative backscatter as a function of range. Such measurements provide insights into structure and processes in the atmosphere as shown in the example of Figure 2. This image depicts the aerosol structure observed by a downward-looking, aircraft-mounted, ultraviolet lidar operating at 360 nm wavelength (Alvarez *et al.*, 1998) as it crosses the shoreline of Galveston Bay near Houston, Texas. The aerosol profiles clearly show the large difference between the marine and terrestrial mixed layer depths. Mixed layer depth, indicated by the region with significant aerosol scattering capped by a strong backscatter gradient, is much deeper over the warmer land surface. Because mixed layer depth determines the volume in which air pollutants can be trapped, lidar measurements are helpful in understanding changes in surface pollution concentrations due to dilution processes. Lidar aerosol measurements can also be useful in characterizing entrainment, diffusion, and transport processes that affect local air quality (e.g., Senff *et al.*, 1998).

The lidar observations shown in Figure 2 were obtained with a research lidar operating in the ultraviolet region of the spectrum. Most aerosol backscatter lidars, however, incorporate Nd:YAG lasers transmitting at 1.06 μm wavelength. Nd:YAG lasers are widely used as lidar sources because the technology is mature, the lasers are commercially available from many sources, the atmosphere has good transmission characteristics at 1.06 μm, and the output can be doubled using a nonlinear crystal to produce laser radiation at 532 nm in the visible region of the spectrum.

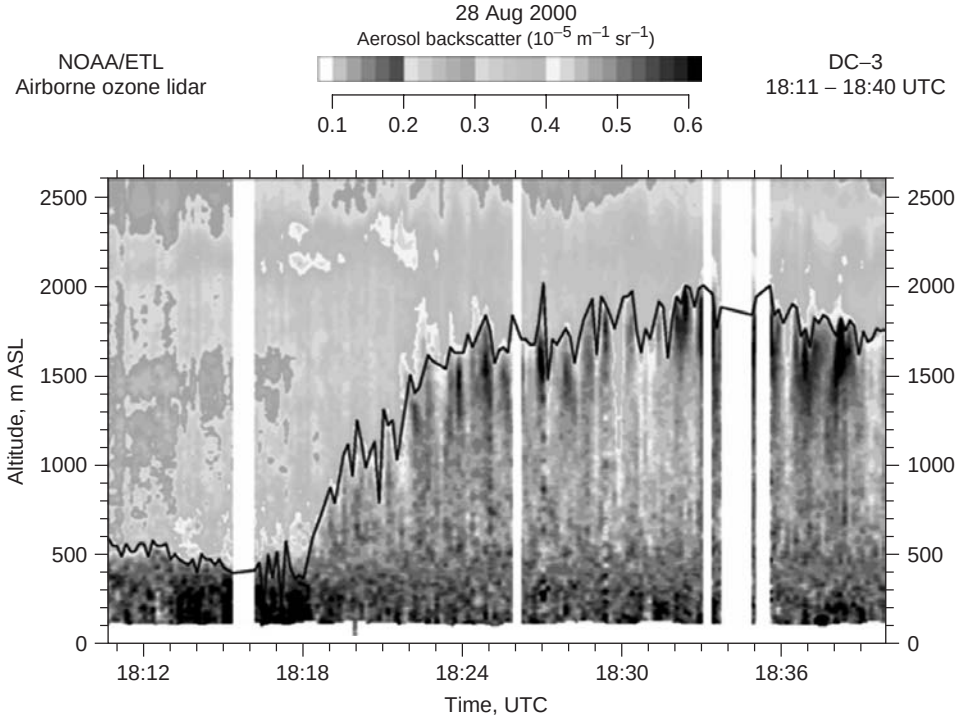


Figure 2 Nadir-pointing airborne backscatter measurements showing strong scattering in the mixed layer, thermal plumes and difference in mixing layer depth over water (left) and land (right) obtained near Houston, Texas (Courtesy C. Senff, University of Colorado). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Some backscatter lidars transmit and receive both 1.06 μm and 532 nm radiation to provide information on wavelength differences in backscattered radiation characteristics.

Although, for simplicity, most backscatter lidars operate in a fixed-pointing mode for observation of the vertical structure of aerosol particles, a scanning capability enables characterization of a 3-dimensional aerosol field. Figure 3 shows a horizontal scan of the aerosol structure just above Lake Michigan as observed by the shore-deployed University of Wisconsin 1.06 μm volume imaging aerosol lidar (Piironen and Eloranta, 1995). The individual fine-scale aerosol structures are well observed by the lidar, which can resolve atmospheric features of about 30 m. The lidar used to obtain the image in Figure 3 is somewhat unique in that the beam can be scanned rapidly in azimuth and elevation, providing very visual images of the three-dimensional evolution of aerosol structure in the internal boundary layer over the lake. Because the 1.06 μm transmitted beam is not eye-safe, even at extended ranges, a human spotter is necessary when the Wisconsin system is operated in a scanning mode. Eye-hazard considerations have led to the development of eye-safe instruments, such as a scanning 1.54 μm aerosol lidar demonstrated at the National Center for Atmospheric Research for tracking clouds of aerosols for Homeland Security and air quality applications (Mayor and Spuler, 2004).

Qualitative backscatter measurements contribute greatly to the understanding of atmospheric structure; however, the importance of understanding the impact aerosol particles have on the earth's radiation requires quantitative estimates of aerosol backscatter and extinction from lidar returns. Because both the backscatter and extinction terms are unknown in equation (1), calibrated estimates of extinction and backscatter profiles are not easily obtained. In many cases, a stable solution for inversion of equation (1) can be obtained by incorporating assumptions on the relationship between $\beta(r)$ and $\sigma(r)$ as well as a boundary value of extinction or backscatter (Klett, 1985). Many inversion techniques used to estimate backscatter and extinction incorporate an estimate of the ratio of extinction to backscatter coefficients $\sigma(r)/\beta(r)$, which is usually referred to as the *lidar ratio*, which has units of sr. The lidar ratio can vary by a factor of three or more; studies aimed at quantifying the lidar ratio for different aerosol regimes have reported values ranging roughly between about 20 sr to more than 70 sr, depending on the type, shape, and chemical composition of the aerosol particles.

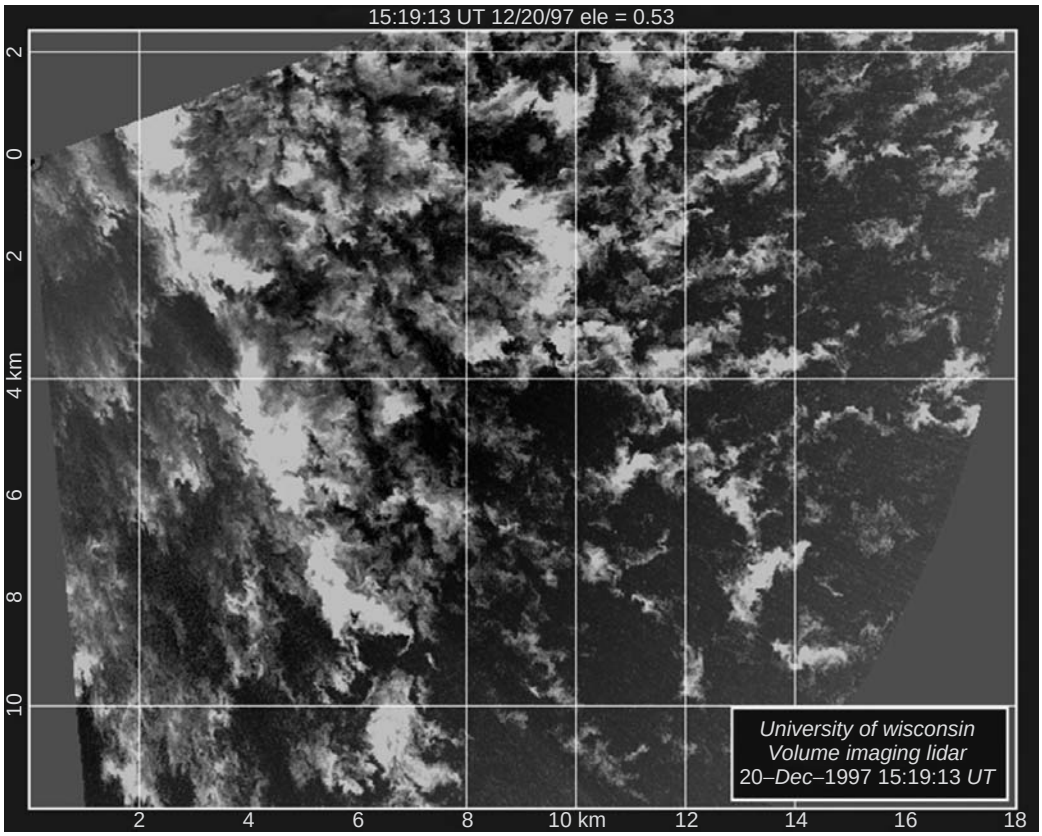


Figure 3 Volume imaging lidar horizontal scans showing regions of high aerosol above Lake Michigan (Courtesy E. Eloranta, University of Wisconsin). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

By employing lidar designs and techniques capable of separating aerosol and molecular-backscattered returns in equation (1), more robust estimates of extinction and backscatter can be calculated. Because molecular scattering and extinction can be accurately specified given information on atmospheric density (i.e., temperature and pressure) available from meteorological analyses, isolating the molecular lidar return enables a direct solution for the range-dependent total extinction of the molecular-backscatter signal. Once the extinction is known, the range-dependent aerosol backscatter coefficient can be computed.

A High Spectral Resolution Lidar (HSRL) design takes advantage of the difference in spectral width between the aerosol and molecular-scattered signals, evident in Figure 1, to separate the two returns. In an HSRL receiver, the collected backscattered signal, which contains both aerosol and molecular-scattered light, is split into two paths. For one path, a narrowband filter removes the aerosol-scattered light prior to detection, while the entire aerosol-plus-molecular scattered light is detected on the other path. By comparing the two signals, the separate aerosol and molecular components of the return are determined, which enables direct solution for the extinction and backscatter coefficients and calculation of the lidar ratio. To achieve the separation of aerosol and molecular-scattered signals, both the transmitter and receiver of a HSRL system must be more complex than those employed in simple backscatter lidars. The transmitted pulse must be spectrally narrow in order to assure that aerosol light does not leak through the receiver aerosol filter, and the receiver must include two receiver paths, each containing very stable filters.

Figure 4 shows estimates of calibrated backscatter and depolarization from a HSRL developed at the University of Wisconsin (Piironen and Eloranta, 1994) for stand-alone

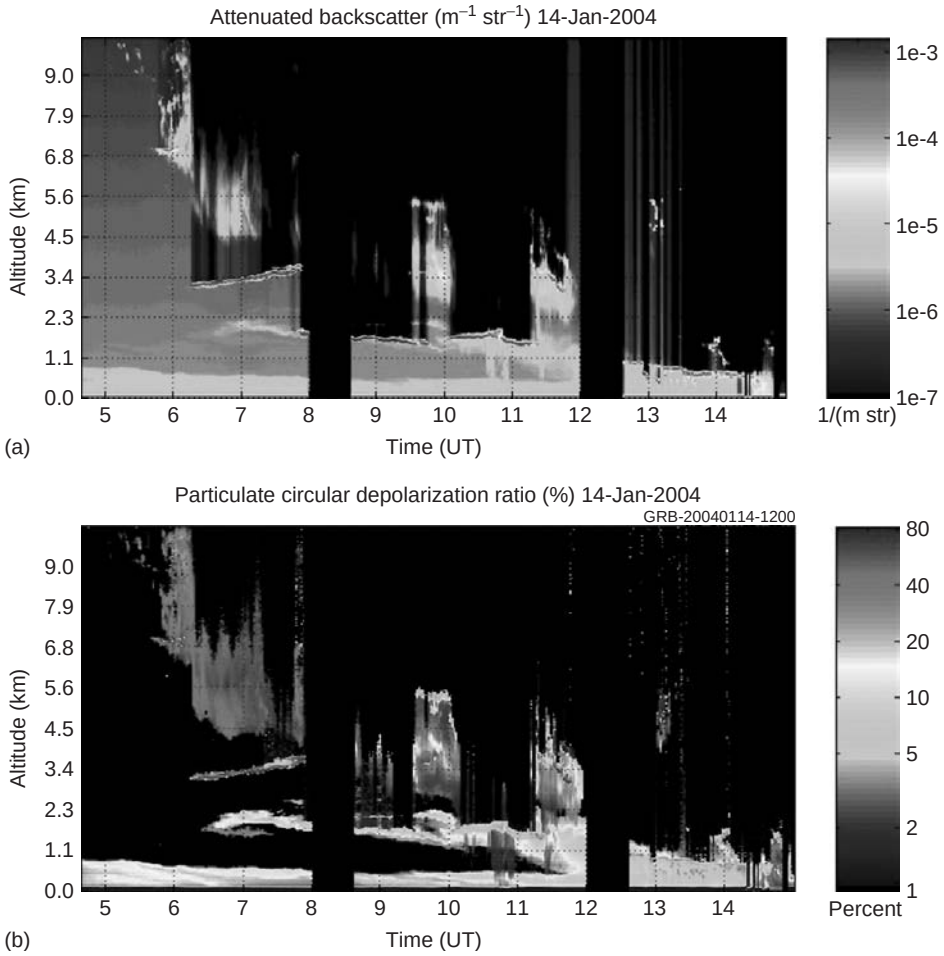


Figure 4 Surface-based high spectral resolution lidar time-height measurements showing calibrated backscatter (a) and depolarization ratio (b) from multiple layers of cloud and aerosol particles over Madison, Wisconsin. Black regions above clouds in (a) are regions of no signal due to extinction of the lidar beam by the clouds (Courtesy E. Eloranta, University of Wisconsin) (*see* **Chapter 65, Estimation of Water Vapor and Clouds Using Microwave Sensors, Volume 2**). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

probing of cloud water and ice and aerosol properties. The Wisconsin lidar incorporates so-called micropulse lidar technology, whose primary characteristics are a low energy, high-pulse repetition rate laser transmitter and photon-counting receiver (Spinhirne *et al.*, 1995). The lidar includes a frequency-doubled Nd:YAG laser transmitting 0.15-mJ, 532-nm pulses at a rate of 4 kHz as the lidar source. A 40-cm diameter telescope collects the backscattered light and directs it into the receiver, where an iodine absorption filter is used to remove the aerosol-backscattered light from the molecular channel. The Wisconsin HSRL produces continuous calibrated profiles of backscatter, extinction (not shown), and depolarization ratio. From the backscatter and extinction information, the optical depth profile is computed, as is seen in Figure 5. Because ice particles introduce more depolarization than water droplets, the depolarization channel provides important information on ice/water phase in clouds (Sassen, 1991). Additional cloud information can be obtained by jointly deploying an HSRL system alongside a mm-wavelength cloud radar, which penetrates extensively into optically thick clouds, and an infrared radiometer. The combination of radar, lidar and radiometer enables differentiation of ice, water, and precipitation, as well as providing information on the size and number concentration of the cloud particles (e.g., Intrieri *et al.*, 2002).

Raman lidar provides another method for separating molecular- from aerosol-backscattered returns to obtain calibrated measurements of cloud and aerosol properties. For characterizing aerosol properties, the receiver includes a channel to measure the combined aerosol and molecular elastic backscatter returns and other channels to measure the Raman-scattered return from nitrogen and water vapor. The Raman channels serve the same function as the molecular channel in an HSRL receiver. Because the Raman-scattered radiation is weak, Raman lidars are characterized by high-power transmitters and large-optic receivers. An example of a Raman lidar is the instrument operating at the Department of Energy's Atmospheric Radiation Measurement Southern Great Plains (SGP) site in northern Oklahoma (Goldsmith *et al.*, 1998). The lidar incorporates a frequency-tripled Nd:YAG lidar transmitter operating at 30 Hz with 350–400 mJ pulses. A 61-cm diameter telescope collects the scattered light. The SGP instrument includes receiver channels at 355 nm (elastic backscattered radiation), 387 nm (nitrogen Raman scatter), and 408 nm (water vapor Raman scatter). The SGP Raman lidar is notable because it has been operating continuously for several years, producing daily records of aerosol backscatter and extinction as well as water vapor concentrations. The instrument is capable of measurements throughout the troposphere at night. During daytime, the strong solar background introduces noise on the Raman channels and limits measurements to a maximum height of roughly 4–6 km above the surface.

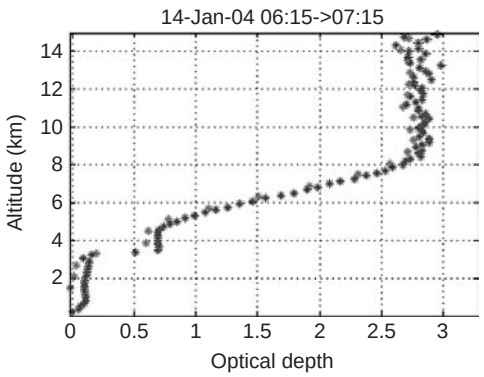


Figure 5 Optical depth profile measured between 6:15 and 7:15 UTC for data shown in Figure 4. Optical depths are computed for 150 m and 600 m vertical averages (Courtesy E. Eloranta, University of Wisconsin). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Measurement of aerosol extinction, backscatter, and depolarization at multiple lidar wavelengths enables estimation of some microphysical properties of the aerosol. Scientists at the Institute for Tropospheric Research in Leipzig, Germany, have deployed a lidar system that transmits at 355, 400, 532, 710, 800, and 1064 nm (Müller *et al.*, 1998). The receiver measures elastic backscatter at each of the six wavelengths, as well as nitrogen Raman backscatter at 387 and 607 nm and water vapor Raman backscatter at 660 nm. Depolarization at 710 nm is also measured. By applying an inversion algorithm specifically designed for the available optical data sets, estimates of effective radius, volume, surface area, number concentration, and refractive index are computed. Results of these lidar inversions were shown to compare well with similar measurements derived from sun photometer observations and direct *in situ* measurements. Although the Leipzig instrument suite is quite complex and the comparison data sets are limited to a few cases, these results provide an indication of the amount of information potentially available from aerosol lidar measurements.

Surface-based, continuously operating aerosol profiling lidars incorporating low pulse energy, high repetition rate, micropulse technology have become widespread. The NASA-sponsored Micropulse Lidar Network (MPLNET) was founded in 2000 to link together similar micropulse lidars for acquisition of long-term observations of aerosol and cloud vertical profiles at several sites around the world (Welton *et al.*, 2001). The MPLNET lidars are colocated with sun photometers to assess aerosol optical properties at different locations and validate satellite retrievals of aerosol optical properties. Another aerosol lidar network, the European Aerosol Research Lidar Network (EARLINET) incorporates 19 different research lidar systems from 11 countries (Matthais *et al.*, 2004). EARLINET aims to provide a quantitative, statistically relevant data set of the vertical

aerosol distribution over Europe, and has also been used to investigate phenomena such as Saharan dust outbreaks, forest fires, and air pollution episodes. EARLINET sites include both aerosol backscatter and Raman lidars to provide calibrated measurements of backscatter and extinction. EARLINET is unique in that quality assurance was carried out among the participating lidar systems based on comparison with a mobile "standard" system and sun photometers. The interest in global aerosol distributions and climatologies has led to the formation or planning of other lidar networks for investigating aerosol distributions in Asia (Murayama *et al.*, 2001), South America, Russia, and the eastern United States.

LIDARS FOR MEASUREMENTS OF ATMOSPHERIC GASES

Lidar systems are also frequently employed to profile gas species concentrations in the atmosphere. In the DIAL technique, laser pulses are transmitted at two wavelengths into the atmosphere along the same atmospheric path. One wavelength (the *online* wavelength) is precisely selected and stabilized to coincide with a spectral absorption feature of the gas of interest. The second wavelength, usually referred to as the *offline* wavelength, is chosen to match the wing of the absorption line or a window region of the spectrum where absorption is minimal. The pulses at the two wavelengths can either be emitted simultaneously using separate laser transmitters, or sequentially. As in an aerosol lidar system, the pulses propagate and are scattered by atmospheric aerosol particles and molecules. Some of the energy in the online wavelength pulse is absorbed by the target gas, with the amount of absorption proportional to the number density of absorbing molecules encountered on the outgoing and returning propagation path. Typically, the online and offline wavelengths are transmitted nearly simultaneously and chosen to be sufficiently close together that differential scattering and extinction by gases other than the gas of interest are negligible. Under these conditions, the range-resolved concentration $\rho(R)$ can be estimated from (Schotland, 1974):

$$\rho(R, R+\Delta R) = \frac{1}{2(K_L - K_W)\Delta R} \times \left[\ln \frac{P_L(R)}{P_L(R+\Delta R)} - \ln \frac{P_W(R)}{P_W(R+\Delta R)}\right] \quad (2)$$

where $\rho(R)$ is the average density of the gas over the range segment from range R to range $R + \Delta R$, P_L and P_W are the optical power received at the online and offline wavelengths respectively, K_L is the gas absorption cross section at the online wavelength and K_W is the gas absorption cross section at the offline wavelength. Equation (2) is derived by applying equation (1) for each of the two wavelengths and forming the difference of the logarithm of $P(R)$ evaluated at R and $R + \Delta R$ for both the online and offline wavelengths.

The DIAL technique has been used to measure a number of atmospheric gases, including water vapor, ozone, ethylene, ammonia, and nitrogen dioxide. The method is most commonly applied to observe water vapor and ozone because of their importance in weather, climate, and air quality processes. DIAL water vapor measurements are usually carried out in the near-infrared at wavelengths below 1 μm, where a large number of absorption features are present. By operating on the strongest absorption lines, measurements of water vapor can be obtained even in the upper troposphere and stratosphere, where water vapor concentrations can be as low as a few ppm by volume. Figure 6 shows an example of water vapor measurements obtained from the NASA Lidar Atmospheric Sensing Experiment (LASE) in an airborne DIAL water vapor system (Ismail *et al.*, 2000) on a flight from Barbados to Wallops Island, Virginia. It can be seen that this system is capable of measuring variations in water vapor that range over about three orders of magnitude between the surface and the tropopause. Figure 6 also shows that features in the moisture field associated with a tropopause fold and frontal lifting are well observed by the lidar. Measurements such

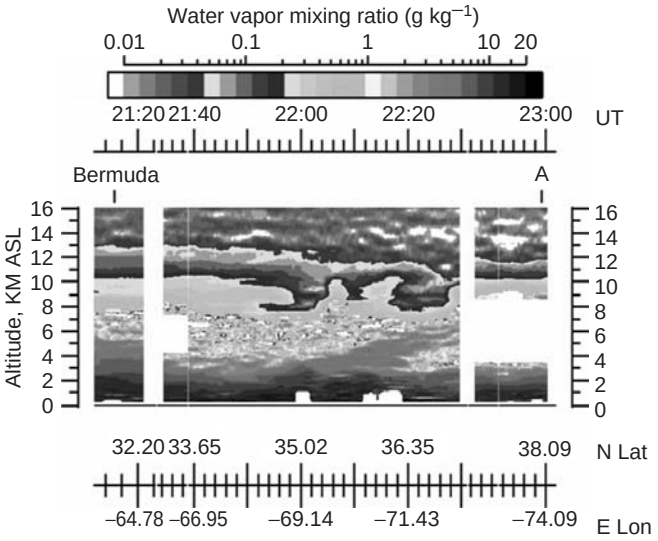


Figure 6 Water vapor measurements obtained by the LASE (Lidar Atmospheric Sensing Experiment) system operating from the NASA ER-2 aircraft on a flight from Bermuda to Wallops Island, Virginia (marked as point A), in July 1996. A tropopause fold containing low stratospheric water vapor can be seen in the top-center of the figure, and the lifting of moisture along a front leading from the boundary layer to cloud formation above 8 km near Wallops Island can also be seen (Reprinted from Browell *et al.*, 2003. ©2003, with permission from Elsevier). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

as those shown in Figure 6 require very precise laser frequency stabilization. The all-solid transmitter in the LASE system is a Ti:sapphire laser pumped by a doubled Nd:YAG laser. The frequency of the Ti:sapphire laser is controlled by injection seeding, using a diode laser that is frequency locked to a water vapor line in the 815-nm region. The online and offline wavelengths are produced by varying the diode laser frequency and are separated by less than 70 pm in wavelength. Another airborne DIAL system, developed and operated at the German Aerospace Center (DLR), operates in the stronger 940 nm region and has been used to measure water vapor in dry regions of the upper troposphere and lower stratosphere (Ehret *et al.*, 1999). DIAL systems such as the NASA and DLR instruments have been applied to observe a variety of important phenomena, including tropospheric/stratospheric exchange, initiation of convection, moisture structure around the dry line, and inflow of moisture into hurricanes.

DIAL methods are also used to measure ozone in the troposphere and stratosphere. The discovery of the ozone hole and subsequent agreements aimed at eliminating chemicals that destroy stratospheric ozone have elevated the importance of ozone measurements. Although global mapping of stratospheric ozone is done from satellites, surface-based lidar instruments at several locations worldwide are being used to study the structure and variability of stratospheric ozone as part of the international Network for the Detection of Stratospheric Change (e.g. Godin-Beekmann *et al.*, 2003). More recently, the importance of lower tropospheric ozone has also been recognized because of its role both as a greenhouse gas and an air pollutant. Tropospheric and stratospheric ozone lidars operate at wavelengths within the Hartley and Huggins ultraviolet absorption bands from about 260 to 355 nm. At these wavelengths Rayleigh scattering dominates, such that much of the atmospheric scattering is from molecules. An example of a tropospheric ozone measurement is presented in Figure 7, which shows ozone measurements made by a downward-looking airborne ozone lidar on a flight directly over the city of Nashville during a high pollution period (Banta *et al.*, 1998). A region of enhanced ozone forms a dome over the city. During the night, ozone above the surface remained as an elevated residual layer and was transported to different locations within the region by elevated wind flows, where it was available for mixing down to the surface on the following day. The lidar that produced the measurements shown in Figure 7 incorporated a KrF excimer laser transmitting at wavelengths of 272, 292, 313, 319,

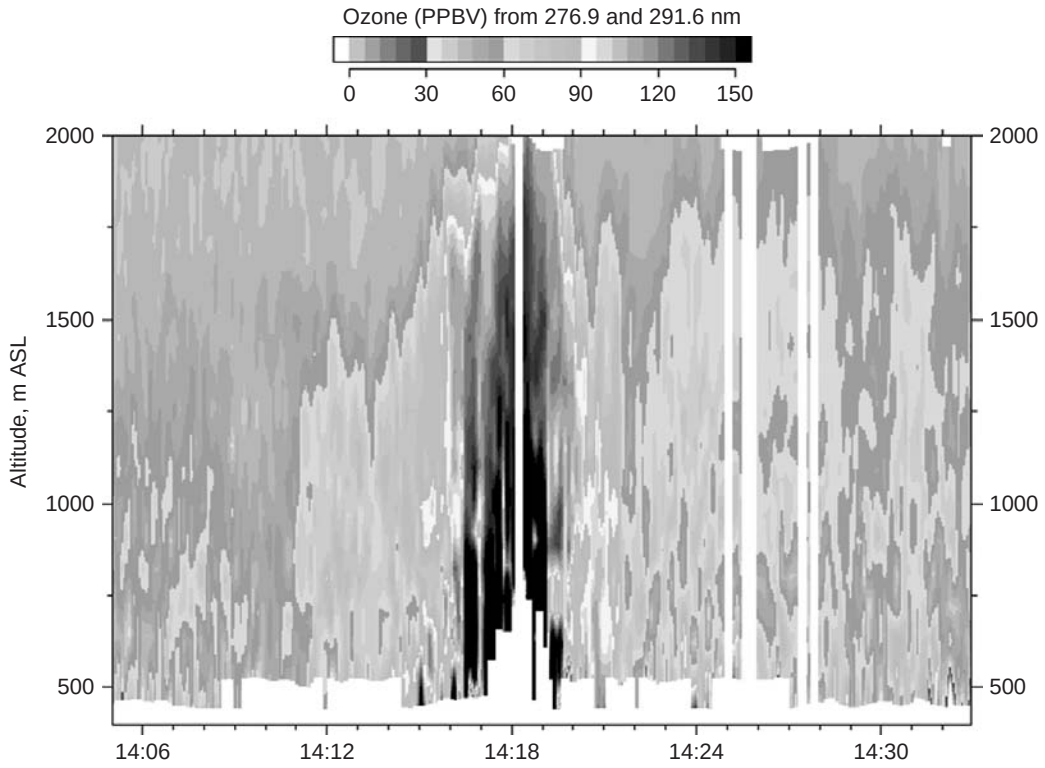


Figure 7 Ozone dome observed by airborne ozone lidar during a flight over Nashville, Tennessee during an air quality stagnation event. The area of high ozone roughly corresponds to the Nashville urban area (Reproduced from Banta *et al.,* 1998 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and 360 nm simultaneously; the light was collected by a 50-cm diameter receiver. The multiple wavelengths each correspond to different ozone absorption cross sections and enable optimal system operation for a range of ozone concentrations.

As noted previously, Raman lidars can also measure atmospheric constituents by detecting the Raman backscattered light from specific molecules. Raman lidar is applied most commonly to measure the water vapor mixing ratio (Whiteman *et al.*, 1992). In a Raman system, mixing ratio is computed by comparing the signal in the water vapor receiver channel with that from a nitrogen channel after correcting for transmission effects. Raman lidars are capable of providing measurements with sufficiently high range and temporal resolution to observe a variety of processes that affect moisture distribution in the atmosphere. Raman lidars operating at night, when solar background light is not present, can profile water vapor well into the upper troposphere. An example of a water vapor measurement is seen in Figure 8, which shows a 24-h water vapor profile generated by the Raman lidar at the SGP site in Oklahoma. Evolution of the water vapor structure is clearly seen. During the early and latter parts of the period, corresponding to daytime measurement, the water vapor concentrations are well mixed. However, during the 12-h period from about 0900 UTC to 2100 UTC, the region of high water vapor appears in a thin layer at about 1 km height and is decoupled from the surface. Measurements such as these are being used to examine and improve cloud and radiative transfer models and to investigate interesting weather phenomena as they occur over the SGP site. Figure 8 also illustrates the effect of sunlight on Raman lidar measurements. Although the daylight measurements (after 12:00 noon UTC) are noticeably noisier, especially in the region above about 4 km, the data in Figure 8 indicate the capability of the instrument to provide high scientific-quality data on a continuous basis.

The potential benefits of continuously profiling water vapor and ozone above the surface have stimulated efforts to produce inexpensive, continuously operating instruments. Although the SGP Raman lidar operates continuously, the size and expense of the system limit the feasibility of deploying multiple systems for extended periods in, for example, mesoscale arrays. Current efforts to produce an economical stand-alone instrument have focused on a micropulse-type lidar system employing off-the-shelf components. Figure 9 shows an example of measurements obtained at night with a stand-alone micropulse DIAL lidar that employs a diode laser transmitter and flared amplifier operating at 823 nm in the infrared (Machol *et al.*, 2004). The lidar observations show good agreement with the radiosonde profile. However, the diode laser plus flared amplifier combination used to produce the measurements of Figure 9 produced insufficient energy to overcome solar background for daytime operation. Although other

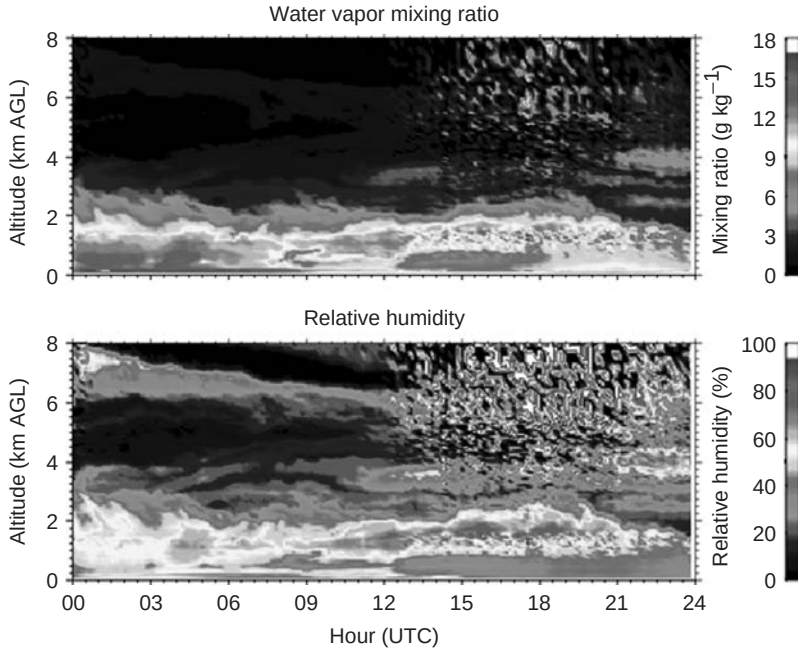


Figure 8 Time-height plot of water vapor obtained with a Raman water vapor lidar at the Atmospheric Radiation Measurement Program Southern Great Plains site in Oklahoma, showing change in water vapor structure and signal intensity during a 24-h period. Increased noise at upper levels during daytime is seen beginning at sunrise (12:00 PM UTC). (Courtesy D. Turner, Battelle). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

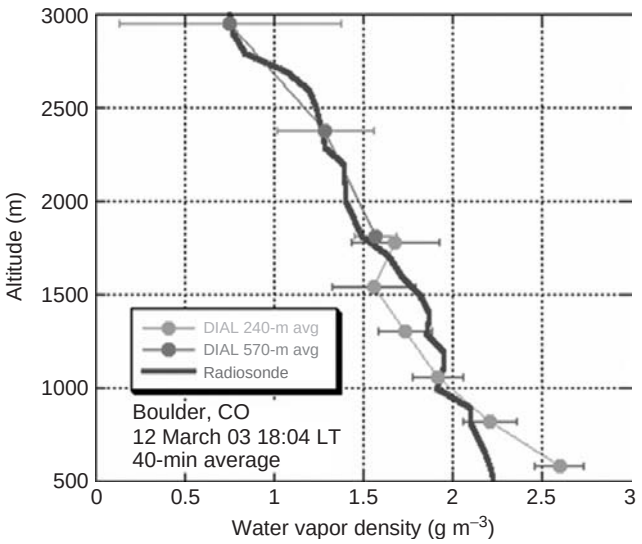


Figure 9 Water vapor profile taken with an autonomous, continuously operating micropulse DIAL system showing good agreement with radiosonde measurements (Courtesy J. Machol, University of Colorado). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

lasers with higher pulse energies and suitable wavelength for stand-alone DIAL are readily available, the challenge remains to identify a laser that is sufficiently economical to enable production and deployment of multiple stand-alone water vapor lidar systems analogous to the MPLNET aerosol network.

LIDARS FOR MEASUREMENT OF WINDS

Lidars applied to measure atmospheric winds make use of the Doppler effect, in which radiation scattered or emitted from a moving particle or molecule is shifted in frequency as a result of the movement of the scatterer. Atmospheric Doppler lidars irradiate a volume of atmosphere with a pulse of very narrow-band laser radiation, then detect the change in frequency of the radiation backscattered from atmospheric aerosol particles or molecules present in the volume. Doppler lidars are well suited for obtaining detailed wind and turbulence observations for a wide variety of applications. A lidar beam can be easily scanned or pointed to characterize wind fields within very confined three-dimensional spaces such as shallow atmospheric boundary layers, narrow canyons, and turbulent structures. Also, the relative compactness of some Doppler lidars makes deployment on aircraft or other moving platforms extremely viable. Doppler lidars deployed on satellites have been proposed, with one instrument under development, to obtain global measurements of atmospheric wind fields. By scanning a lidar beam from an orbiting satellite and analyzing the returns backscattered from clouds, aerosols, and molecules, a satellite-based instrument could provide important wind information from regions that are currently not well observed for assimilation into numerical forecast models.

Frequency analysis in a Doppler lidar receiver is carried out using one of two techniques: *coherent* detection (also known as *heterodyne* detection) or *direct* detection (alternately labeled *incoherent* detection). In a heterodyne receiver, a digitized time series created by mixing laser radiation with the backscattered radiation at the face of a detector is numerically analyzed. Direct detection Doppler lidars employ an interferometer to provide information on the Doppler shift. The interferometer can be implemented in either of two ways: it can serve as a spectrum analyzer to produce a light pattern which contains the information on the frequency content (the so-called *fringe-imaging technique*), or it can used as a frequency discriminator whose output is proportional to the frequency of the input radiation (often called the *edge technique*).

Coherent lidar makes use of a local oscillator (LO) laser with constant amplitude and whose frequency is precisely controlled to be a known frequency offset, typically on the order of tens of MHz, from that of the laser transmitter. The LO beam is made collinear (i.e., mixed) with the backscattered radiation, and directed into an optical detector, which generates, after high-pass filtering, a time-varying electrical signal with amplitude proportional to the amplitude of the backscattered electromagnetic field and frequency equal to the difference between the backscattered field frequency and the LO laser field frequency. This signal is digitized, and spectral analysis applied to estimate its frequency characteristics and compute the Doppler shift. Because the mixing process effectively limits the bandwidth of the receiver, solar background light is not a major source of noise in coherent lidar systems, and performance is not degraded during daytime hours.

Coherent lidars have been used to measure winds for a variety of applications and from an assortment of platforms such as ships and aircraft. Since these lidars operate in the infrared, where aerosol scattering dominates molecular scattering, they require aerosol particles to be present at some level to obtain usable returns. Although clouds also provide excellent lidar targets, most of the more useful applications of coherent lidars involve probing the atmospheric boundary layer or lower troposphere where aerosol content is highest. The capability to scan the narrow lidar beam directly adjacent to terrain, in particular, enables unique, high-resolution probing of wind structure and evolution in complex terrains such as mountains and valleys.

Winds in complex terrain can be complicated to measure and predict because the terrain can force the winds to behave differently at different heights. Also, the evolution of winds over time at different levels can be quite

variable, on account of shallow layers of thermally forced flows. Effects of complex terrain on local winds are often enhanced at night. Doppler lidars measure winds with much more detail in time and space than any standard or even enhanced network of anemometers. In addition to characterizing the winds near the surface, Doppler lidars can observe the wind structure above the surface, typically 1 to 4 km above ground level, depending on the scattering properties of the atmosphere. Two important benefits of applying Doppler lidar for measurements in complex terrain have been (i) increasing the understanding of wind flow by filling in the gaps of standard measurements, both horizontally and vertically, and (ii) providing detailed two- and three-dimensional datasets for use in model evaluations.

An example of the use of Doppler lidar for complex terrain studies can be seen in results from the Vertical Transport and Mixing Experiment (VTMX) in Salt Lake City, Utah. The lidar in this study employed low-elevation-angle conical scans to measure the winds across the Salt Lake City basin. From these scans, the horizontal variability of the winds resulting from the complex terrain was observed, including convergence of winds, canyon outflows, low-level jet (LLJ), and horizontal eddies, as shown in Figure 10. It was apparent from the lidar measurements that the LLJ associated with a notch in the high terrain to the south of the basin was a significant wind feature (Banta *et al.*, 2004). Since the LLJ is elevated, the extensive full-time network of surface measurements that exists in the basin did not detect this feature; the details of its existence were not known previous to the analysis of the VTMX Doppler lidar measurements.

In addition to increasing understanding of the winds in the Salt Lake City basin, lidar measurements were also used to assess the performance of a mesoscale numerical model, the Regional Atmospheric Modeling System (Fast and Darby, 2004; Banta *et al.*, 2004). Issues investigated included the model's ability to predict the horizontal variability of the winds, which includes the LLJ and canyon outflows, the timing of the reversal between up-basin and down-basin flow, and the penetration of the canyon outflows into the basin. A sample comparison between the model output and lidar data is shown in Figure 11.

Doppler lidars can also provide important information on mixing of atmospheric constituents such as pollutants or moisture. Figure 12 shows a Doppler-lidar vertical-plane scan of a gust front during the Southern Oxidants Study (SOS) in Nashville. By analyzing measurements from the Doppler lidar and a colocated ozone profiling lidar, the role of the gust front passage in temporarily redistributing the vertical concentrations of ozone, leading to an increase of ozone at the surface, was apparent (Darby *et al.*, 2002).

Direct detection Doppler lidars are most effective in measuring winds when aerosol particles are not present on account of their capability to estimate the Doppler

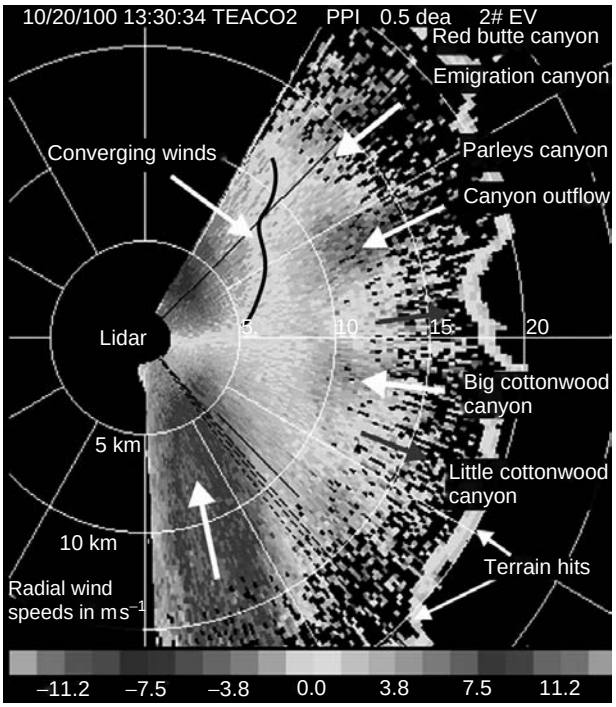


Figure 10 Nearly horizontal azimuth scan (0.5° elevation) from Doppler lidar deployed at Salt Lake City, Utah. Range rings are at intervals of 5 km and velocities on color bar in $m\,s^{-1}$ (positive values toward the lidar). The major canyons are annotated on the figure. Figure shows down-basin jet coexisting with outflows from the major canyons (white arrows), forming a convergence line (Banta *et al.,* 2004. ©2004 AMS). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

shift of radiation scattered by atmospheric molecules. Measurement of Doppler shifts from molecular-scattered radiation is challenging, however, because of the large Doppler-broadened bandwidth of the return (equivalent to approximately $320-350\,m\,s^{-1}$), which is two orders of magnitude greater than the desired measured precision of a few $m\,s^{-1}$ or better. Consequently, large numbers of photons must be detected to reduce the uncertainty in the measurement to an acceptable level. To obtain these high photon counts from molecular-backscattered returns at ranges beyond a few kilometers, some combination of multiple pulse averaging, powerful lasers, and large receiver optics is required.

Molecular-scatter wind measurements have been demonstrated in the visible spectral region at 532 nm wavelength as well as at 355 nm in the near ultraviolet, using both fringe-imaging (Skinner and Hays, 1994) and edge (Gentry *et al.*, 2000) receivers. The ultraviolet region has the dual advantages of enhanced molecular scatter and less restrictive laser eye-safety restrictions. Figure 13 shows a time series of the vertical profile of horizontal wind speeds measured in the troposphere using a molecular-scatter,

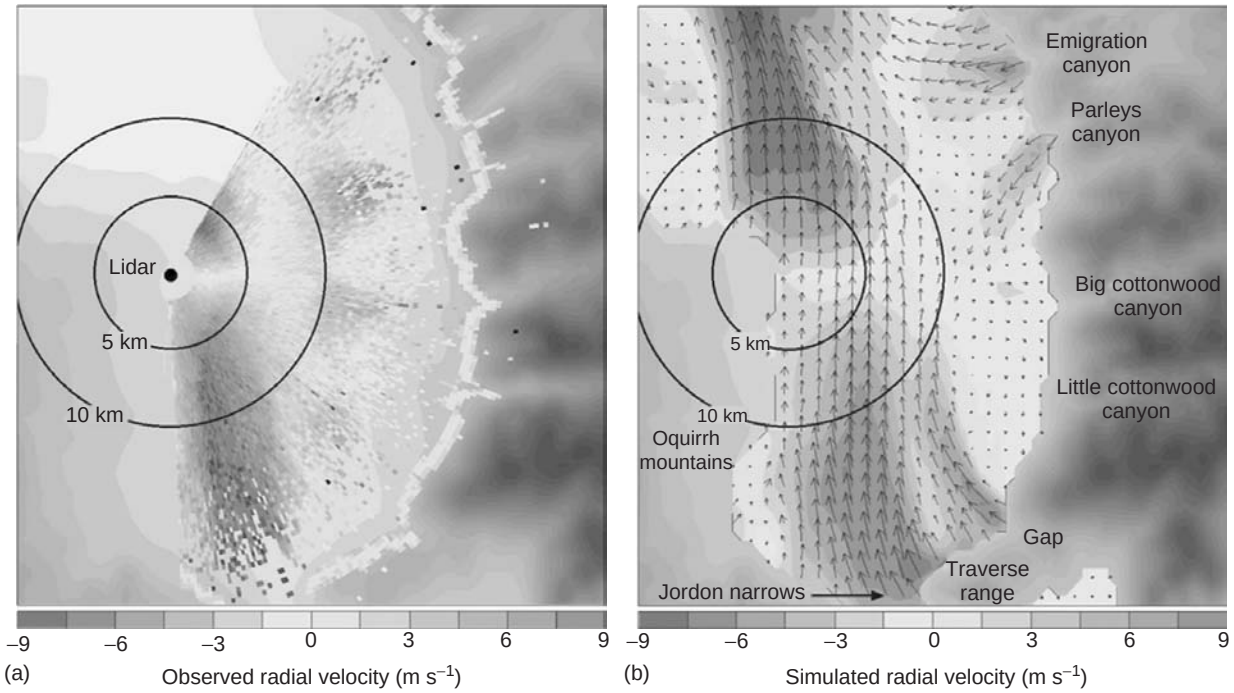


Figure 11 Doppler-lidar data from an azimuth scan at 1/2° elevation as in Figure 10, with radial velocities in m s^{-1} (a). Simulated Doppler-lidar azimuth scan at 1/2° fixed elevation calculated from RAMS model output as described in the text (b). Arrows indicate full horizontal wind vector on the simulated sloping scan surface. Radial velocities indicates the model-derived radial wind component that would be observed by a lidar (Banta *et al.*, 2004. ©2004 AMS). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

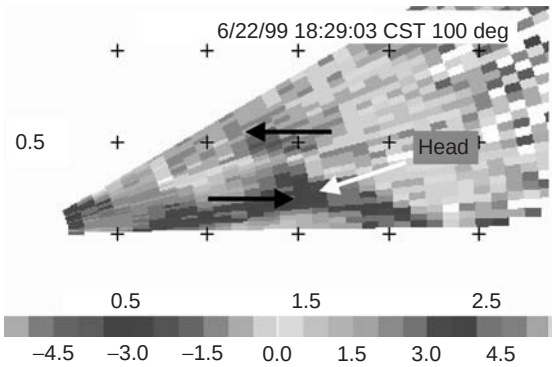


Figure 12 Doppler lidar range–height scan showing a mature gust front several minutes after passing the Doppler lidar. The head of the front is marked with an arrow. The lidar is located at $x = z = 0$ and the tick marks are in 0.5 km increments (Reproduced from Darby *et al.*, 2002 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

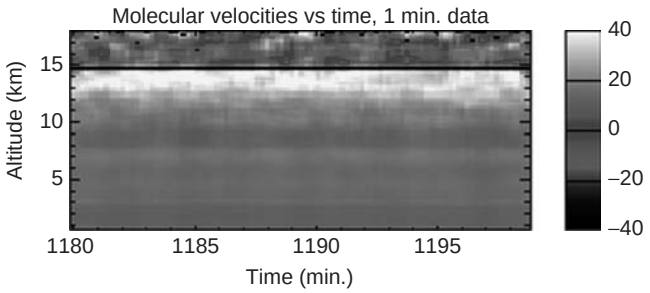


Figure 13 Time series of tropospheric wind speed profiles measured from molecular-backscattered radiation by a −355-nm direct detection lidar over Mauna Loa, Hawaii. Each profile is computed from 1-min averaged returns. (Courtesy M. Dehring, Michigan Aerospace Corporation). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

ground-based, 355-nm wavelength Doppler fringe-imaging lidar at Mauna Loa, Hawaii (Dehring *et al.*, 2003), showing measurements throughout the troposphere. To obtain these measurements, backscattered photons were collected by a 0.5-m receiver aperture, averaged for 1 min, and processed. Figure 14 shows, in more detail, a wind speed profile measured during the period.

FUTURE DIRECTIONS FOR LIDAR REMOTE SENSING

Over the next several years, lidar remote sensing of the environment is expected to expand along several fronts.

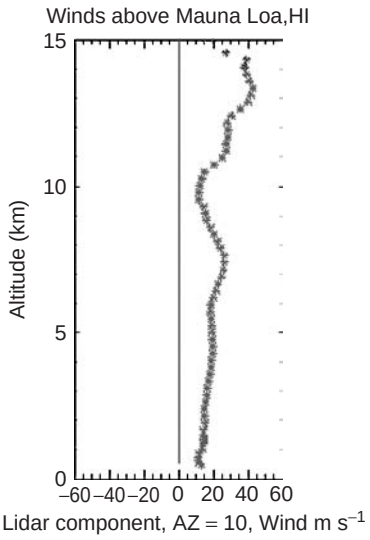


Figure 14 Higher resolution depiction of a wind speed profile measured during observational period shown in Figure 13 (Courtesy M. Dehring, Michigan Aerospace Corporation). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Most notable will be deployments of lidar systems on orbital platforms for global observations of important atmospheric and land surface parameters. The successful launch of the Ice, Cloud, and land Elevation Satellite (ICESat), with its Geophysical Laser Altimeter System (GLAS) illustrated the promise and impact of space-deployed lidar systems. ICESat has provided important measurements of surface ice sheet changes as well as global observations of clouds and aerosol particles. A further step forward will be the deployment of the NASA Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) spacecraft, scheduled for the middle of 2005. CALIPSO, which will fly in formation with the CloudSat spacecraft, will include a three-channel aerosol lidar and passive radiometers to provide key observations of global aerosol and cloud properties, climatologies, and processes. Knowledge gained from CALIPSO will lead to improved characterization of cloud and aerosol radiative and transport processes. Scheduled to follow CALIPSO, in 2007, is the European Space Agency's Atmospheric Dynamics Mission (ADM, Stoffelen *et al.*, 2005), aimed at measuring global wind fields for assimilation into numerical weather forecast models. Data sets from ICESat, CALIPSO, and ADM should provide a strong demonstration of the application of lidar techniques to significantly advance knowledge of important earth system processes.

Improvements in optics technology will also enable new advances in surface-based lidars. Significant research is being focused on improving optical detectors in the eye-safe infrared region beyond 1.5 μm, which will likely pave the way for a new generation of lidars to measure atmospheric aerosol properties, species, and winds. Work also is continuing on development of better and more efficient methods to generate laser radiation in the ultraviolet and infrared spectral regions, and to enable easy tuning of laser wavelengths to probe specific spectral regions. Other technology research is being directed toward making optical components more compact, robust, and economical, which would likely stimulate development of smaller, cheaper, and more portable systems suitable for operation from highway vehicles or unattended aerial vehicles. Ultimately, cheaper instruments that can be produced in quantity should become possible, permitting more extensive deployments of instrument arrays. Such arrays could provide information on, for example, regional distribution of ozone and aerosol particles for air quality forecasting, or, by codeploying water vapor and wind profiling remote sensors, could characterize regional water vapor distribution and moisture transport.

FURTHER READING

Huffaker R.M. and Hardesty R.M. (1996) Remote sensing of atmospheric wind velocities using solid state and CO_2 coherent laser systems. *Proceedings of the IEEE*, **84**, 181–204.

Measures R.M. (1984) *Laser Remote Sensing, Fundamentals and Applications*, Wiley Interscience: New York.

Stephens G.L. (1994) *Remote Sensing of the Lower Atmosphere*, Oxford University Press: New York.

REFERENCES

Alvarez R.J. II, Senff C.J., Hardesty R.M., Parrish D.D., Luke W.T., Watson T.B., Daum P.H. and Gillani N. (1998) Comparisons of airborne lidar measurements of ozone with airborne in situ measurements during the 1995 Southern Oxidants study. *Journal of Geophysical Research*, **103**, 31,155–31,171.

Banta R.M., Darby L.S., Fast J.D., Pinto J.O., Whiteman C.D., Shaw W.J. and Orr B.D. (2004) Nocturnal low-level jet in a mountain basin complex I: evolution and implications to other flow features. *Journal of Applied Meteorology*, **43**, 1348–1365.

Banta R.M., Senff C.J., White A.B., Trainer M., McNider R.T., Valente R.J., Mayor S.D., Alvarez R.J., Hardesty R.M., Parrish D., *et al.* (1998) Daytime buildup and nighttime transport of urban ozone in the boundary layer during a stagnation episode. *Journal of Geophysical Research*, **103**, 22,519–22,544.

Browell E.V., Ismail S. and Grant W.B. (2003) Differential absorption lidar. *Encyclopedia of Atmospheric Science*, Academic Press: London, pp. 1183–1194.

Collis R.T.H. and Russell P.B. (1976) Lidar measurement of particles and gases by elastic backscattering and differential absorption. In *Laser Monitoring of the Atmosphere*, Hinkley E.D. (Ed.), Springer-Verlag: New York, pp. 71–151.

Darby L.S., Banta R.M., Brewer W.A., Neff W.D., Marchbanks R.D., McCarty B.J., Senff C.J., White A.B., Angevine W.M. and Williams E.J. (2002) Vertical variations in O_3 concentrations before and after a gust front passage. *Journal of Geophysical Research*, **107**(D13), 4176, doi:10.1029/2001JD000996.

Dehring M.T., Nardell C.A., Pavlich J.C., Hays P.B. and Dors I. (2003) Performance and comparison of 532-nm and 355-nm groundwinds lidars, in *Lidar Remote Sensing for Industry and Environment Monitoring III*, Singh U.N., Itabe T., Liu Z. (Eds.), *Proceedings of the SPIE*, Vol. 4893, SPIE: pp. 337–347.

Ehret G., Hoinka K.P., Stein J., Fix A., Kiemle C. and Poberaj G. (1999) Low stratospheric water vapor measured by an airborne DIAL. *Journal of Geophysical Research*, **104**, 31351–31360.

Fast J.D. and Darby L.S. (2004) An evaluation of mesoscale model predictions of down-valley and canyon flows and their consequences using Doppler lidar measurements during VTMX 2000. *Journal of Applied Meteorology*, **43**, 420–436.

Gentry B., Chen H. and Li S.X. (2000) Wind Measurements with a 355 nm Molecular Doppler Lidar. *Optics Letters*, **25**, 1231–1233.

Godin-Beekmann S., Porteneuve J. and Garnier A. (2003) Systematic DIAL ozone measurements at Observatoire de Haute-Provence. *Journal of environmental Monitoring*, **5**, 57–67.

Goldsmith J.E.M., Blair F.H., Bisson S.E. and Turner D.D. (1998) Turn-key Raman lidar for profiling atmospheric water vapor, clouds, and aerosols. *Applied Optics*, **37**, 4979–4990.

Intrieri J.M., Shupe M.D., Uttal T. and McCarty B.J. (2002) An annual cycle of Arctic cloud characteristics observed by radar and lidar at SHEBA. *Journal of Geophysical Research*, **107**, 8030, doi:10.1029/2000JC000423.

Ismail S., Browell E.V., Ferrare R.A., Kooi S.A., Clayton M.B., Brackett V.G. and Russell P.B. (2000) LASE measurements of aerosol and water vapor profiles during TARFOX. *Journal of Geophysical Research*, **105**, 9903–9916.

Klett J.D. (1985) Lidar inversion with variable backscatter/extinction ratios. *Applied Optics*, **21**, 1638–1643.

Machol J.L., Ayers T., Schwenz K.T., Koenig K.W., Hardesty R.M., Senff C.J., Krainak M.A., Abshire J.B., Bravo H.E. and Sandberg S.P. (2004) Preliminary measurements with an automated compact differential absorption lidar for the profiling of water vapor. *Applied Optics*, **43**, 3110–3121.

Matthais V., Freudenthaler V., Amodeo A., Balin I., Balis D., Bösenberg J., Chaikovsky A., Chourdakis G., Comeron A., Delaval A., *et al.* (2004) Aerosol lidar intercomparison in the framework of the EARLINET project. 1. Instruments. *Applied Optics*, **43**, 961–976.

Mayor S.D. and Spuler S.M. (2004) Raman-shifted eye-safe aerosol lidar. *Applied Optics*, **43**, 3915–3925.

Müller D., Wandinger U., Althausen D., Mattis I. and Ansmann A. (1998) Retrieval of physical particle properties from lidar observations of extinction and backscatter at multiple wavelengths. *Applied Optics*, **37**, 2260–2263.

Murayama T., Sugimoto N., Uno I., Kinoshita K., Aoki K., Hagiwara N., Liu Z., Matsui I., Sakai T., Shibata T., *et al.* (2001) Ground-based network observation of Asian dust events of April 1998 in east Asia. *Journal of Geophysical Research*, **106**, 18–345–18–359.

Piironen P. and Eloranta E.W. (1994) Demonstration of a high spectral resolution lidar based on an iodine absorption filter. *Optics Letters*, **19**, 234–236.

Piironen A.K. and Eloranta E.W. (1995) Convective boundary layer mean depths, cloud base altitudes, cloud top altitudes, cloud coverages, and cloud shadows obtained from Volume Imaging Lidar data. *Journal of Geophysical Research-Atmospheres*, **100**, 25569–25576.

Sassen K. (1991) The polarization lidar technique for cloud research: a review and current assessment. *Bulletin of the American Meteorological Society*, **72**, 1848–1866.

Schotland R.M. (1974) Errors in the lidar measurement of atmospheric gases by differential absorption. *Journal of Applied Meteorology*, **13**, 71–77.

Senff C.J., Hardesty R.M., Alvarez R.J. II and Mayor S.D. (1998) Airborne lidar characterization of power plant plumes during the 1995 Southern Oxidants Study. *Journal of Geophysical Research*, **103**, 31,173–31,190.

Skinner W.R. and Hays P.B. (1994) Incoherent Doppler lidar for measurement of atmospheric winds. *Proceeding of the SPIE*, **2216**, 383–394.

Spinhirne J.D., Rall J.A.R. and Scott V.S. (1995) Compact eye safe lidar systems. *Reviews of Laser Engineering*, **23**, 112–118.

Stoffelen A., Pailleux J., Kallén E., Vaughan J.M., Isaksen L., Flamant P., Wergen W., Andersson E., Schyberg H., Culoma A., Meynart R., Endemann M. and Engmann P. (2005) The Atmospheric Dynamics Mission for global wind field measurement. *Bulletin of the American Meteorological Society*, **86**, 1–73.

Welton E.J., Campbell J.R., Spinhirne J.D. and Scott V.S. (2001) Global monitoring of clouds and aerosols using a network of micro-pulse lidar systems, In *Lidar Remote Sensing for Industry and Environmental Monitoring*, Singh U.N., Itabe T., Sugimoto N., (Eds.), *Proceedings of the SPIE*, Vol. 4153, SPIE: 151–158.

Whiteman D.N., Melfi S.H. and Ferrare R.A. (1992) Raman lidar system for the measurement of water vapor and aerosols in the Earth's atmosphere. *Applied Optics*, **31**, 3068–3068.

49: Estimation of Surface Insolation

RACHEL T PINKER

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, US

The incoming solar radiation from the sun (insolation) that reaches the Earth's surface (about 50% of that emitted from the sun) determines the exchange of energy between the land and the atmosphere and consequently, controls climate and climate change. Environmental satellites are considered useful tools for providing information on surface radiative fluxes at various temporal and spatial scales, improving estimation of terrestrial water and energy storage. For several years now, observations from satellites have been used to obtain information on the various components of atmospheric and surface radiative fluxes, initially, retrospectively, and more recently, operationally. In this paper, discussed are the needs for information on radiative fluxes in hydrological applications, methodologies for obtaining such information by methods of remote sensing, accuracy of such estimates, and current status of data availability. Examples of applications in hydrological studies and climate research are provided, and links to international activities and future prospects are indicated.

INTRODUCTION

Need for Information

Information on the spatial and temporal distribution of surface radiative fluxes is required for

1. modeling the hydrologic cycle (Rind *et al.*, 1992; Ohmura and Wild, 2002; Sorooshian *et al.*, 2002; Mitchell *et al.*, 2004);
2. representing interactions and feedbacks between the atmosphere and the terrestrial biosphere (Dickinson, 1986; Henderson-Sellers, 1993; Prince and Goward, 1995; Hicke *et al.*, 2002);
3. estimating global terrestrial and oceanic net primary productivity (Platt, 1986; Goward, 1989; Running *et al.*, 1999; Zhao *et al.*, 2005);
4. validating climate models (Garrat *et al.*, 1993; Wild *et al.*, 1995; Wielicki *et al.*, 2002);
5. improving the understanding of transport of heat, moisture, and momentum across the surface–atmosphere interface (Betts *et al.*, 1997; Berbery *et al.*, 1999; Baumgartner and Anderson, 1999; Sui *et al.*, 2002);
6. improving land–atmosphere interaction parameterizations (Chen *et al.*, 1996); and
7. providing information on the dominant forcing functions of the surface energy budgets (Wielicki *et al.*, 1995; Wood *et al.*, 1997; Mitchell *et al.*, 2004; Rodell *et al.*, 2004).

Studies of long-range weather forecasting are performed with the aid of numerical weather prediction and general circulation models (GCMs). Satellite observations are considered the only source of information that can be used for model evaluation on a global scale. Progress in the use of satellites for probing the atmosphere has enabled research to refocus on developing advanced satellite methods in climate research. The Joint Scientific Committee (JSC) of the WCRP endorsed the Global Energy and Water Cycle Experiment (GEWEX) as one of several core activities that should ultimately lead to the prediction of global and regional climate change (Chahine, 1992; Sorooshian, 2003). GEWEX includes radiation, hydrometeorology, and modeling and prediction activities. The objective of the radiation component is to determine atmospheric and surface radiative fluxes at a level of accuracy that will allow the prediction of transient climate variations and long-term climate trends. The WCRP GEWEX Surface Radiation Budget (SRB) project is specially tasked to produce, validate, and assess long-term surface and atmospheric radiative budgets on a global scale (Stackhouse *et al.*, 2000, 2002, 2004).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Specific national research interests are linked to the GEWEX program via continental-scale GEWEX experiments such as the GEWEX Continental-scale International Project GCIP/GAPP over the Unites States, (the GEWEX Americas Prediction Project (GAPP) is a continuation of GCIP), the Large-scale Biosphere-Atmosphere Experiment in Amazonia (LBA), the Baltic Sea Experiment (BALTEX), the GEWEX Asian Monsoon Experiment (GAME) and the African Monsoon Multidisciplinary Analyses (AMMA) project. Methods for deriving surface radiative fluxes have been developed for most of these basin-scale studies (Pereira *et al.*, 1996; Ceballos and Moura, 1997; Hollmann *et al.*, 1999; Raschke *et al.*, 1998; Takamura *et al.*, 2001; Nakajima, 2001; Pinker *et al.*, 2003; Sekiguchi *et al.*, 2003).

Each of these continental-scale regions is characterized by unique climatic conditions, and addresses a wide range of scientific issues. For example, the specific objectives of the GEWEX Continental-scale International Project (GCIP) and the GAPP (a continuation of GCIP) (WCRP-67, 1992; NRC, 1998) are the determination of the time/space variability of the hydrologic cycle and energy budget over the Mississippi River Basin; development and validation of macroscale hydrologic models; and, utilization of existing and future satellite observations for achieving these objectives (Leese, 1994; 1997). A summary of what has been achieved can be found in Mitchell *et al.* (2004) and Roads *et al.* (2003).

Another example is the LBA, whose major objective is to improve understanding of the hydrological cycle of the Amazon region. This region serves as a major modulator of hemispheric climate in part because of the release of latent heat associated with heavy precipitation, which is a major energy source for the atmosphere. Other objectives include monitoring of land use changes and improving estimates of Net Primary Productivity, and soil biogeochemistry of carbon, nutrients, and trace gases (Nobre *et al.*, 2001; Keller *et al.*, 2001; Avissar *et al.*, 2002). The LBA is an excellent example of a multidisciplinary, international research effort, aimed at understanding the climatological, ecological, biogeochemical, and hydrological functioning of Amazonia, its interaction with the Earth system, and its response to land use change.

BALTEX is a long-term multinational project for studying coupled hydrological processes between complicated terrain, sea, ice, and the atmospheric circulation to determine the energy and water budgets of the Baltic Sea and related river basins (Raschke *et al.*, 2002). Major emphasis in BALTEX is given to a thorough understanding of hydrological processes that are to be parameterized in global and limited area models. The goal of the McKenzie GEWEX Study (MAGS) is to improve understanding of the water and energy cycle of the Mackenzie River Basin and provide an improved understanding of cold regions, high latitude hydrological and meteorological processes, snow and ice processes, permafrost, arctic clouds and radiation interactions (Stewart *et al.*, 1998). The GAME aims to improve our understanding of the role of the Asian monsoon in the global climate system and develop methods for long-range forecasting.

Information on surface radiative fluxes is needed to address all of the above issues.

Feasibility

Retrospective Studies

Long-term satellite observations over large spatial scales are now available for implementing inference schemes for radiative fluxes. Under the joint NOAA/NASA PATHFINDER activity (Ohring and Dodge, 1992), uniform, long-term data sets from observations made from numerous satellites, have been prepared into homogeneous time series. Some of these data are processed into reduced resolution, multisatellite, global coverage information, and are known as *International Satellite Cloud Climatology Project* (ISCCP) D1 data (Schiffer and Rossow, 1985). Because of their representation of the diurnal cycle and long-term availability, these data are of particular interest to scientists working on land–atmosphere and ocean–atmosphere interactions, and hydrologic modeling. Parameters derived from these data are based on satellite observations and on ancillary data, as appended to the GEWEX (WCRP-67, 1992) ISCCP D1 product at a nominal resolution of 2.5° (Rossow and Schiffer, 1991, 1999). The satellites that are being used usually have between two to five channels in spectral intervals that are relevant both for inferring the shortwave flux (visible) and for detecting clouds. An example of the channels on selected Geostationary Operational Environmental Satellites (GOES) and the Advanced Very High Resolution Radiometer (AVHRR) on the polar orbiting satellites are given in Table 1. The ISCCP D1 product is an improved version of the ISCCP C1 product (Schiffer and Rossow, 1985). Improvements are related to new cloud screening methodology used for producing the D1 version resulting in better cloud detection over snow cover, in particular, in the Polar Regions.

Table 1a Characteristics of the GOES-8 satellite (After Menzel and Purdom, 1994.)

Channel	Wavelength (μm)	Field of view (km)	Subpoint resol. (km)
1	0.52–0.72	1.0 × 1.0	0.57 × 1.0
2	3.78–4.03	4.0 × 4.0	2.30 × 4.0
3	6.47–7.02	8.0 × 8.0	2.30 × 8.0
4	10.20–11.20	4.0 × 4.0	2.30 × 4.0
5	11.50–12.50	4.0 × 4.0	2.30 × 4.0

Table 1b Characteristics of the AVHRR channels

Channel	Satellites: NOAA-6, 8, 10 (μm)	Satellites: NOAA-7, 9, 11, 12, 14 (μm)	IFOV (millirad)
1	0.580–0.68	0.580–0.68	1.39
2	0.73–1.10	0.73–1.10	1.41
3	3.55–3.93	3.55–3.93	1.51
4	10.50–11.50	10.30–11.30	1.41
5	10.50–11.50	11.50–12.50	1.30

In the last two decades, it has been demonstrated that radiative fluxes could be derived from satellite observations with reasonable accuracy (Raschke and Preuss, 1979; Tarpley, 1979; Pinker *et al.*, 1995, 2003; Frouin and Pinker, 1995; Rossow and Zhang, 1995; Whitlock *et al.*, 1995; Ohmura *et al.*, 1998; Gupta *et al.*, 1997; Zhang and Rossow, 2004). Methods to derive SW fluxes from satellite observations have been implemented by several groups on different spatial and temporal scales. METEOSAT, GOES, GMS and polar orbiting satellites have been all used (Stuhlmann *et al.*, 1990; Pinker and Laszlo, 1992a; Gupta *et al.*, 1999, 2001; Chou, 1994; Brisson *et al.*, 1994; Li *et al.*, 1993; Rossow and Zhang, 1995; Ceballos and Moura, 1997; Ceballos *et al.*, 2004). Valuable experience has been gained from the merging of various global data sets (observations from the TIROS Operational Vertical Sounder (TOVS); snow cover) into the ISCCP data base. Two SW algorithms developed at the University of Maryland (Pinker and Laszlo, 1992a; denoted SRB SW hereafter) and at the NASA Langley Research Center (Gupta *et al.*, 2001) respectively, are currently used at NASA Langley Research Center in support of the WCRP/GEWEX SRB activities. A configuration of satellites, both geostationary and polar orbiting, as illustrated in Figure 1, is needed to cover the entire globe with observations that depict the diurnal cycle. Geostationary satellites cover only part of the globe up to about 50° north and south. The polar orbiting satellites provide coverage at higher latitudes. Figure 2 shows the mean surface radiative fluxes ($W\,m^{-2}$) averaged for the years 1983–2001 for January (a) and July (b) as obtained from the ISCCP D1 data (Rossow and Schiffer, 1991) as

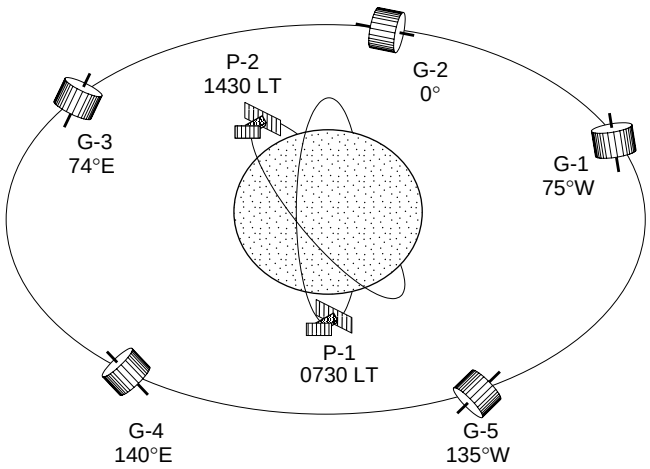


Figure 1 Satellite configuration required for obtaining diurnally resolved surface radiative fluxes includes both geostationary (g) and polar orbiting satellite (morning and afternoon overpass); geostationary satellites cover the globe only up to about 50° north and south

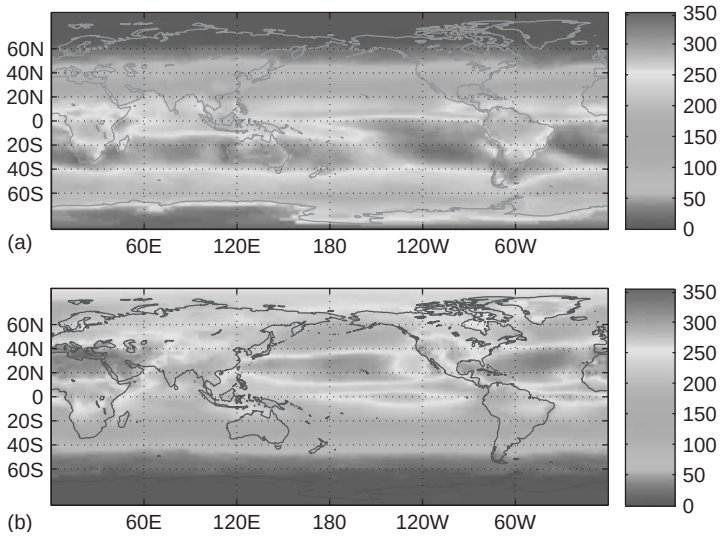


Figure 2 The mean surface radiative fluxes ($W\,m^{-2}$) averaged for the years 1983–2001 for January (a) and July (b) as obtained from the ISCCP D1 data (Rossow and Schiffer, 1991) as inferred from optimally merged geostationary and polar orbiting satellites observations at 2.5° spatial resolution (Zhang *et al.*, 2005). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

inferred from optimally merged geostationary and polar orbiting satellites observations at 2.5° spatial resolution (Zhang *et al.*, 2005).

Real-time Estimates

Operational implementation of retrieval methodologies has been successful and examples will be presented. In support of the World Climate Research Program (WCRP), GEWEX GCIP, and the GAPP, real-time estimates of shortwave radiative fluxes, both at the surface and at the top of the atmosphere, are being produced operationally by the National Oceanic and Atmospheric Administration (NOAA)/National Environmental Satellite Data and Information Service (NESDIS), using observations from the GOES. This is a collaborative effort between NOAA/NESDIS, NOAA/National Centers for Environmental Prediction (NCEP), and the University of Maryland. Inferred shortwave radiative fluxes include total and diffuse quantities (as appropriate), as well as spectral components (e.g. photosynthetically active radiation (PAR)). The interface between the satellite data and inference models has been developed at NOAA/NESDIS (Tarpley *et al.*, 1996). NOAA/NCEP provides information on the state of the atmosphere and surface conditions, as available from analyzed output fields from the Eta model (Rogers *et al.*, 1996). The University of Maryland is involved in model development and modifications (Pinker and Laszlo, 1992a,b; Pinker *et al.*, 2003), sensitivity studies, validation against ground observations, data archiving, and data distribution. The SRB SW model is implemented at NOAA/NESDIS in real time on an hourly basis, for 0.5-degree targets for an area bounded by 66°–126° W longitude and 24°–54° N latitude belts. For each target, at appropriate forecast times, selected data from the NCEP regional forecast model are delivered to the satellite data stream as inputs to the SRB SW model. This approach ensures timely and high quality information input to the satellite inference scheme. In turn, derived radiative fluxes help to diagnose the NCEP forecast model as to its ability to predict correctly radiative fluxes. The website `http://www.atmos.umd.edu/~srb/gcip/webgcip.htm` provides the following hourly, daily, and monthly parameters: shortwave surface downward flux; shortwave surface upward flux; visible surface downward flux (PAR); visible surface upward flux; shortwave top of the atmosphere net flux (down-up). These parameters represent the most needed information on radiative fluxes, and are only a subset of parameters produced by the inference scheme. A comprehensive summary of evaluation of this satellite product against ground truth is presented in Pinker *et al.* (2003). Results from the evaluation of a more recent product in which cloud detection over snow was improved is illustrated in Figure 3. Given is the frequency distribution of differences between daily values as estimated from satellites

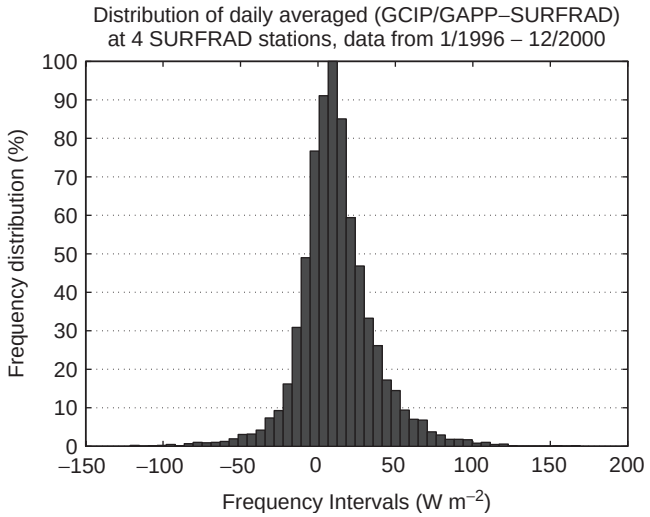


Figure 3 Statistics of errors of the satellite estimates when evaluated against ground observations at four SURFRAD station for the period (1996–2000), daily timescale. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

for the period of 1996–2000 and ground observations at four Surface Radiation Monitoring Network (SURFRAD) stations. As evident, most of the differences are in the range of -50 to $+50\,\mathrm{W\,m^{-2}}$. Attempts to derive surface radiative fluxes over the Amazon region in real time are in progress at the Centro de Previsao de Tempo e Estudos Climaticos (CPTEC)/INPE (Ceballos *et al.*, 2004) and at several European operational centers.

REVIEW OF SELECTED INFERENCE SCHEMES

Methods spanning a wide range of complexities have been developed to derive surface radiative fluxes from satellite observations. Several have been reviewed in Schmetz (1989, 1991, 1993) who critically evaluated sensitivities to input parameters, as well as the physical principles of these methodologies. In reviews that followed, the different parts of the spectrum were discussed independently. The current status of SW retrievals is summarized in Pinker *et al.* (1995) and Whitlock *et al.* (1995). Methods for deriving PAR are described in Frouin and Pinker (1995). A summary of current and future satellite observations of relevance for SRB research is presented in Wielicki *et al.* (1995). In this review, emphasis will be on methods that have consistently provided data sets relevant in climate research.

Physical Principles

In his discussion of the physical principles that allow derivation of SRB from satellite observations, Schmetz (1989) has stressed the importance of the close linear

coupling between SW (0.2–4.0 μm) reflected radiance at the top of the atmosphere (albedo) and the surface irradiance. Cloud extinction (transmittance) and albedo are linearly related since atmospheric constituents do not emit radiation at solar wavelength. There is dependence on solar zenith angle, gaseous and aerosol absorption and scattering, surface reflectivity, and clouds.

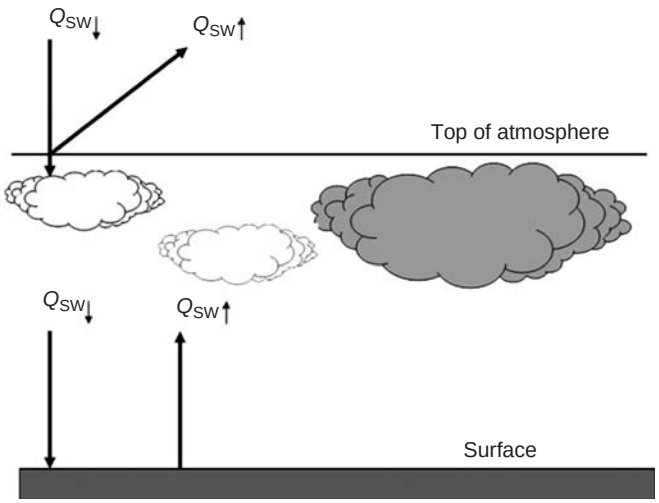


Figure 4 Diagram of various components of shortwave radiative fluxes at Top of the Atmosphere and at the surface. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

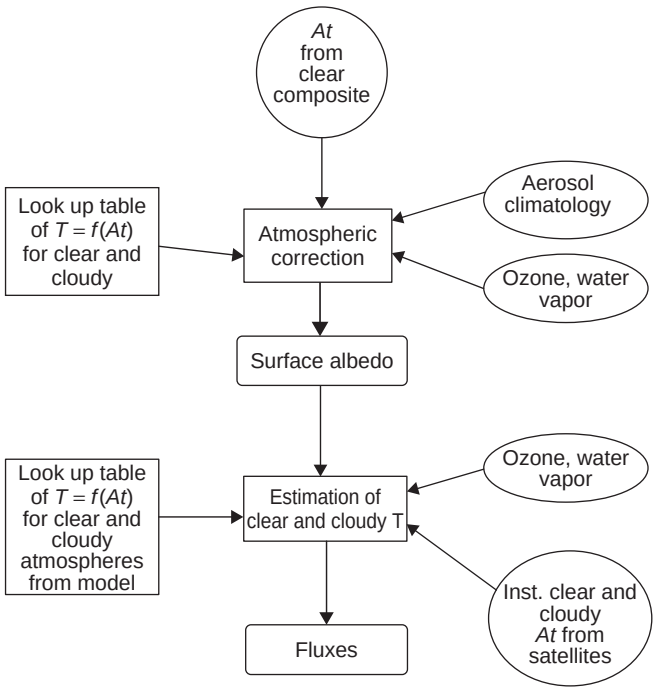


Figure 5 Summary of the processing sequence to infer surface radiative fluxes

Examples of Current Methodologies

In this section, a possible prototype of satellite retrieval methodology of shortwave radiative fluxes at the surface and at the top of the atmosphere (TOA) is illustrated. Existing retrieval methodologies can be viewed as variants of the principles presented here. Several existing methodologies are applicable for the derivation of surface fluxes only. Components of shortwave radiative fluxes are illustrated in Figure 4. Derivation of surface fluxes is possible because for a given atmosphere and surface, TOA albedo and flux transmittance are uniquely related to each other. The discussion in this paragraph will define this relationship. A diagram of the flux retrieval process is presented in Figure 5. TOA downward flux (F_{td}) is calculated from extraterrestrial solar spectrum, accounting for variation in sun–earth distance and position of the sun in the sky relative to local vertical (solar zenith angle). Downward flux at the surface (F_{sd}) is obtained by determining what fraction of F_{td} reaches the surface as radiation is transferred through the atmosphere. This fraction, which is referred to as flux transmittance (T), depends on the composition of the atmosphere (e.g. amount of water vapor and ozone, optical thickness of cloud and aerosol), on length of path the radiation travels through the atmosphere (determined by the solar zenith angle), and to a lesser degree, on the albedo of the surface. Once T is known, surface downward flux is obtained as $F_{sd} = T F_{td}$. The algorithm estimates T from satellite-derived TOA albedo. Once F_{sd} is known, upward flux at the surface (F_{su}) is calculated as $F_{su} = A_s F_{sd}$, where A_s is surface albedo. Similarly, flux reflected to space by the earth–atmosphere system (TOA upward flux, F_{tu}) is obtained from the product of F_{td} and the TOA albedo (A_t), namely, $F_{tu} = A_t F_{td}$. T is determined from a comparison of modeled values of shortwave (0.2–4.0 μm) TOA albedos to shortwave TOA albedo obtained from satellite measurement, and transmittance corresponding to modeled TOA albedo that matches satellite-derived value is selected. For practical reasons, pairs of albedos and transmittances are calculated for atmospheres with a nonreflecting lower boundary. Surface reflection is added in a separate step. Modeled TOA albedos and corresponding transmittances can be calculated spectrally for discrete values of solar zenith angle, amount of water vapor and ozone, aerosol and cloud optical thickness, using a radiative transfer method of preference. Radiative properties of aerosols and clouds can be obtained from independent sources. Absorption by ozone and water vapor needs to be parameterized, for example, as proposed by Ramaswamy and Freidenreich (1992) or others. The albedo-transmittance pairs are made available in a lookup table for the algorithm separately for clear and cloudy atmospheres, and the flux transmittances for clear and cloudy skies are determined by matching the satellite-observed clear and cloudy shortwave TOA albedos, respectively. For a given

solar zenith angle, surface albedo and amount of ozone and water vapor, the matching process involves the adjustment of the aerosol optical depth for clear sky and that of the cloud optical depth for cloudy sky. For example, for GCIP/GAPP, the satellite-observed TOA shortwave albedo is obtained from the visible (0.55–0.75 μm) radiance measured by the imager instrument onboard the GOES satellites through spectral and angular transformations (Zhou *et al.*, 1996). The clear-sky radiance as observed by the satellite is affected both by the surface albedo and the intervening atmosphere. It is quite difficult to separate the affects of each independently. Therefore, at times, information on surface albedo is taken from independent sources or estimated from the "clearest" shortwave TOA radiative flux (or albedo) observed over a number of days (clear-sky composite albedo), and then nominally corrected for Rayleigh scattering, aerosol extinction, and absorption by ozone and water vapor. Next, albedo-transmittance pairs are selected from the lookup table according to the solar zenith angle, water vapor and ozone amount, and are combined with the surface albedo to yield shortwave TOA albedos. One set of pairs is for varying values of aerosol optical depth (clear atmosphere), and the other is for varying values of cloud optical depth (cloudy atmosphere). Finally, the shortwave albedos derived from the instantaneous satellite-observed clear-sky and cloudy-sky radiances are matched with the clear and cloudy sets of albedo-transmittance pairs and clear-sky and cloudy-sky transmittances, and from these, clear-sky and cloudy-sky fluxes are obtained. The clear-sky and cloudy-sky fluxes are then weighted according to the cloud cover (defined as the ratio of number of cloudy pixels to the total number of pixels) to determine the all-sky fluxes. The implementation of the model requires preprocessing of the satellite data and separation of clear and cloudy radiances. The ISCCP D1 and DX data sets provide such information. When original satellite observations are used, there is a need to first separate between clear and cloudy pixels. For the GCIP/GAPP product, this is performed at NOAA/NESDIS (Tarpley *et al.*, 1996). The Clouds and Earth's Radiant Energy System (CERES) instrument on board several Earth Observing System (EOS) satellites from NASA's Earth Science Enterprise (ESE) program provides an opportunity for deriving accurate estimate of SW fluxes (Wielicki *et al.*, 1995). CERES provides a direct measurement of TOA broadband reflectance and augments this measurement with higher resolution retrievals of atmospheric, cloud, and aerosol properties to estimate radiative fluxes. The Surface and Atmospheric Radiation Budget (SARB) component of CERES computes the surface and atmospheric fluxes by iterating retrieved atmospheric properties with the measured TOA fluxes. Intercomparison of older techniques using ISCCP approach described above and the CERES-SARB will occur within the next several years. Progress on derivation of surface albedo and aerosols over land is being made with EOS observations from MODIS and the Multiangle Imaging SpectroRadiometer (MISR) instruments (e.g. Liang *et al.*, 2002; Schaaf *et al.*, 2002).

The various elements of the retrieval algorithms should be tested in a number of different ways. The radiative transfer component needs to be evaluated. A framework for doing so was provided by the Intercomparison of Radiation Codes in Climate Models (ICRCCM) (Fouquart *et al.*, 1991). Surface down-flux estimates can be compared with ground measurements. Several organized activities in the framework of the Satellite Algorithm Intercomparison as sponsored by WCRP and NASA (Whitlock *et al.*, 1995) were conducted with older surface SW measurements, reporting average agreement with ground observations to within $10\,W\,m^{-2}$ on a monthly timescale over a large number of surface sites. Intercomparison with surface sites is ongoing within the GEWEX SRB (Stackhouse *et al.*, 2004) and CERES-SARB programs (Charlock *et al.*, 2000a,b) reporting agreement with ground observations within $10\,W\,m^{-2}$ on a monthly timescale. Numerous independent evaluations were also conducted by algorithm developers or by users (Pinker *et al.*, 2003; Luo *et al.*, 2003). In the GEWEX SRB SW algorithm, the fluxes are calculated in the spectral intervals of 0.2–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7 and 0.7–4.0 μm. Thus, it is possible to obtain fluxes at spectral intervals known to be of significance (e.g. PAR) (Figure 6). This is important, because current GCMs are run in a mode that separates shortwave fluxes at 0.7 μm (Roesch *et al.*, 2002), to allow incorporation of newly derived satellite-based parameters, such as fractional vegetation cover, derived from the Normal Difference Vegetation Index (NDVI), and for

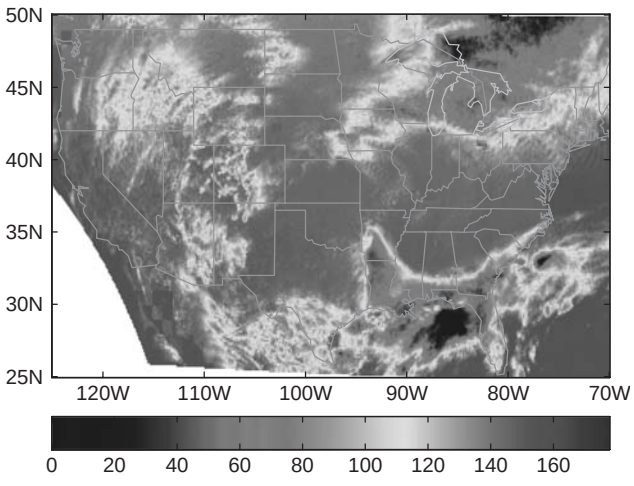


Figure 6 Daily mean surface downward PAR ($W\,m^{-2}$) at 0.5-degree resolution for August 15, 1998. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

improving parameterizations of surface/atmosphere interactions (Gallo and Huang, 1998; Goward and Huemmrich, 1992; Townshend and Justice, 1995; Gutman *et al.*, 1995). Moreover, shortwave fluxes are separated into direct and diffuse components, which is of interest for improved modeling of radiative interaction with vegetation and oceans. Other parameters that are derived include clear-sky and all-sky albedos at the top of the atmosphere and at the surface, and aerosol and cloud optical depths. The shortwave and spectral fluxes are computed separately for clear and all-sky conditions, thus making it possible to derive information on the radiative effects of clouds, known as *cloud-radiative forcing* (Ramanathan *et al.*, 1989).

CURRENT STATUS OF DATA AVAILABILITY

In this section reviewed are products that are readily available, that have been produced for periods of time of interest in climate research, and that have been evaluated against ground observations.

Early GEWEX SRB products are based on ISCCP C1 observations, as provided in:

- First WCRP Surface Radiation Budget Global Data Sets, Shortwave Radiation Parameters March 1985–December 1988, NASA Earth Observing System Distributed Active Archive Center, NASA Langley Research Center, Hampton, VA (Whitlock *et al.*, 1995), at monthly and daily timescales.
- Global l Data Sets Land-Atmosphere Models, ISLSCP Initiative I: 1987–1988, Volume 1–5, NASA Goddard DAAC Science Data Series, include surface shortwave and PAR data, at three hourly intervals at one degree spatial resolution (Seller *et al.*, 1996).
- Global Ecosystem Database, Disk B, National Environmental Satellite, Data, and Information Center National Geophysical Data Center, Boulder, Colorado, November 1997, which provides information at three hourly intervals, monthly averaged on PAR, for the period July 1983 to July 1988 (Kineman, 1997).
- Global l Data Sets Land-Atmosphere Models, ISLSCP Initiative II: consists of a 10-year global data set spanning the years 1987 to 1995 at spatial resolutions of one-quarter to one degree, and available at: **`http://islscp2.sesda.com/ISLSCP2 1/html pages/islscp2 home.html`**

Improved surface SW products using improved ISCCP and other satellite measurement platforms:

- Release 2 of the GEWEX SRB dataset featuring version 2.1 of the SRB SW algorithm using ISCCP DX data mapped to one-degree grids is now available (Stackhouse *et al.*, 2004). The dataset uses water vapor amounts from the NASA Goddard Earth Observing System data assimilation version 1 (GEOS-1). Release 2 spans from July 1983 through October 1995. The dataset also includes fluxes from the thermal infrared (i.e. long wave). An expanded ground measurement database is used, including Baseline Surface Radiation Network (BSRN) measurements, for the assessment of the SW flux products, distributed in space and time, at over 1000 locations.
- ISSCP Flux Data (FD) (Zhang *et al.*, 2004) and University of Maryland GEWEX SRB SW data (Pinker *et al.*, 2005) based on ISSCP D1, for July 1983 through December 2001, at nominal resolution of 280-km. Both are accessible via the internet at respective institutions.
- CERES experiment instruments were launched aboard the Tropical Rainfall Measuring Mission (TRMM) in November 1997 and on the EOS Terra satellite in December 1999 and two additional instruments on the EOS Aqua spacecraft in 2002. Products include both solar-reflected and Earth-emitted radiation from the top of the atmosphere to the Earth's surface. Analyses of the CERES data build upon the foundation laid by previous missions such as the Earth Radiation Budget Experiment (ERBE). The CERES data collected so far demonstrate that the CERES instruments are substantially improved over the ERBE instruments. The CERES data product descriptions are given in the Data Products Catalog that can be accessed on the web at: **`http://asd-www.larc.nasa.gov/DPC/DPC.html.`**
- Real-time GCIP/GAPP data since January 1996 are distributed at: **`http://www.meto.umd.edu/~srb.`** All the input and output parameters, produced in support of GCIP/GAPP (currently, a total of 71), are stored at the University of Maryland. Four types of information are archived: satellite-based information, used to drive the model; auxiliary data used to drive the model; Eta model output products relevant for hydrologic modeling; independently derived satellite products.

EVALUATION OF SATELLITE ESTIMATES USING SURFACE MEASUREMENTS

Evaluation of the SW fluxes is critically dependent upon accurate surface measurements. Historic surface SW measurements were collected under the auspices of the WCRP at the World Radiation Data Center in St. Petersburg, Russia. This database includes sites from all continents and contains measurements in the form of daily and monthly averages. From these sites, the Swiss Federal Institute of Technology derived the Global Energy Balance Archive (GEBA) (Ohmura *et al.*, 1998). These data and data from other historical networks such as the NOAA Climate Monitoring and Diagnostics Laboratory (CMDL) are compiled at the NASA Langley Data Center in the validation

of the GEWEX SRB algorithms. Globally distributed surface measurements over the last 10 years are improving with the advent of the GEWEX BSRN (Ohmura *et al.*, 1998; Whitlock *et al.*, 1995). BSRN is dedicated to providing long-term high quality measurements at locations around the globe critical for satellite validation. New regional-scale networks of surface measurements are also used for validation. Examples of these are observations from the SURFRAD (Hicks *et al.*, 1996; Augustine *et al.*, 2000) and the Atmospheric Radiation Measurement (ARM) observations of shortwave fluxes are obtained from the UnitedStates Southern Great Plain (SGP) central facility operated according to the specifications of the BSRN (Michalsky *et al.*, 1999). Data from special campaigns are also frequently used such as the CERES Program/ ARM Validation Experiment (CAVE) (Rutan *et al.*, 2001), which represents an ongoing campaign that provides calculated radiative fluxes, input data, and validating measurements over the ARM Cloud and Radiation Test bed (CART) site in Oklahoma. In the validation effort, of interest are also the accuracy limits of ground observations as discussed in Shi and Long (2002) and summarized in Table 2 for instantaneous values. More work is needed to better quantify the operational measurement uncertainties (P. Stackhouse and E. Dutton, personal communication, 2003).

Indirect evaluation of the satellite products on global scale against various GCM outputs as well as against ground observations that were extrapolated to global scale is presented in Figure 7. The GCM models were compared to each other under the Atmospheric Model Intercomparison

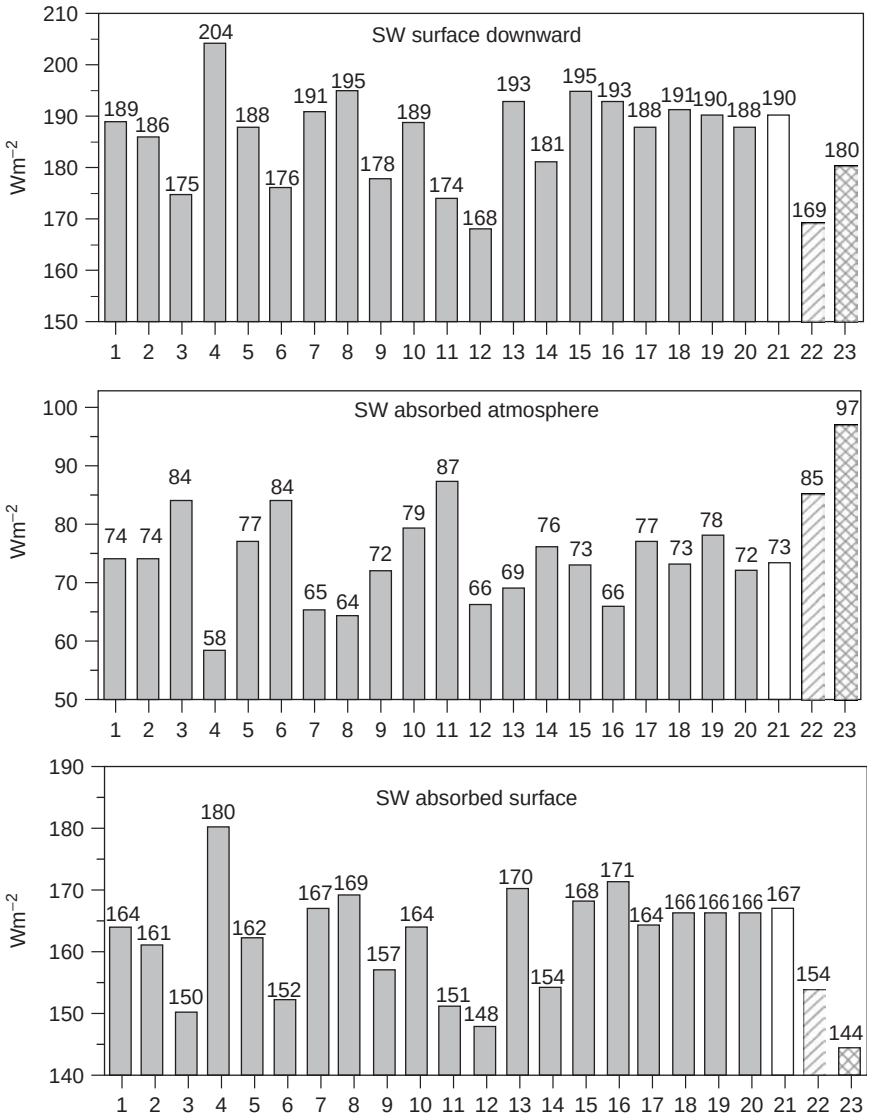


Figure 7 Comparison of global averages of satellite-based estimates of radiative fluxes at the Top of the Atmosphere, the surface and within the atmosphere with those obtained from selected global climate models (After M. Wild, 2005)

Table 2 Operational limits on measured shortwave radiation ($W\,m^{-2}$) (After Shi and Long, 2002)

	Best	Typical	Worst
Diffuse	$4.0 + 1.4$	$9.0 + 3.1$	$12.0 + 3.8$
Direct Normal	$6.3 + 3.3$	$13.3 + 6.3$	$12.0 + 3.8$
Upwelling SW	$11.1 + 2.8$		

Table 3 Global-scale (90S-90N) and large-scale (60S-60N) averages of selected radiative parameters as derived from the twenty year record of satellite observations (Pinker *et al.* 2005)

SW surface ↓	90S-90N	189.55
SW TOA net	90S-90N	240.69
SW surface net	90S-90N	167.26
SW absorbed	90S-90N	73.43
SW TOA albedo	90S-90N	30.61%
SW surf albedo	90S-90N	12.80%
SW surface. ↓	60S-60N	203.72
SW TOA net	60S-60N	263.08
SW surface net	60S-60N	184.76
SW absorbed	60S-60N	78.32
SW TOA albedo	60S-60N	28.49%
SW surf albedo	60S-60N	9.69%

Project (AMIP) (Gates *et al.*, 1999), which was an international effort to determine the systematic climate errors of atmospheric models under realistic conditions using the observed monthly averaged distributions of sea-surface temperature and sea ice as boundary conditions. Twenty models that initially participated in the intercomparison are numbered in Figure 7 from 1 to 20 (CCCM; CCSR; CNRM; COLA; DNM; ECMWF; GISS; GLA; JMA; MGO; MPI; MRI; NCAR; NCEP; PNNL; SUNYA; UGAMP; UIUC; UKMO; YONU); details are provided in Appendix A. Column number 21 represents average results from about 20 years of satellite-derived surface fluxes from ISCCP D1 (Pinker *et al.*, 2005); columns 21 and 22 represent results as obtained by Wild *et al.* (1998) and Gilgen and Ohmura (1999), respectively. In Table 3 presented are selected global and large-scale averages of various radiative parameters as derived from the 20 year satellite record.

APPLICATIONS

LDAS and GLDAS

The importance of land-surface processes for climate and weather modeling has been recognized, and current general circulation and climate models are coupled with Land–surface Models (LSM). Improving weather and seasonal climate prediction requires information on initial states of the atmosphere, oceans, and land. Sparse observations need to be assimilated from various observing platforms into atmospheric initial states via 4-dimensional data assimilation (4DDA) to blend observations with the background fields of an LSM. Global atmospheric 4DDA have been used in operational NWP (Kalnay *et al.*, 1996), yet, errors remain in soil moisture/temperature and surface energy/water fluxes, because of biases in the surface forcing. Under the North American Land Data Assimilation System (NLDAS) project (Mitchell *et al.*, 2004) quality controlled, spatially and temporally consistent, real-time and retrospective forcing data sets were constructed. These data sets were obtained from the best available observations and model output to support LSM activities with the Noah Mosaic (Koster and Suarez, 1992), Variable Infiltration Capacity model (VIC) (Liang *et al.*, 1996) and Sacramento (Burnash *et al.*, 1973) models. They feature hourly temporal resolution and 1/8th degree spatial resolution, and have been evaluated in Luo *et al.* (2003) and Pinker *et al.* (2003). A comparison of radiative forcing from LDAS and Eta Data Assimilation System (EDAS) against ground observations is shown in Figure 8. The studies by Cosgrove *et al.* (2003) and Luo *et al.* (2003) describe the data sources, generation and validation of NLDAS forcing, produced in real-time and retrospectively on the NLDAS grid. This activity is a first real-time operational prototype of a continental-scale uncoupled land 4DDA assimilation executed daily at NCEP. Used are real-time streams of hourly to daily data based on observations of precipitation and insolation fields (from satellites and/or ground observations) that drive four LSMs running in parallel to produce hourly output on a 1/8 degree. It is hoped that such an approach will reduce the errors in the storage of soil moisture and energy that are often present in NWP models and which degrade the accuracy of forecasts.

In parallel to the NLDAS activity, a Global Land Data Assimilation System (GLDAS) is being developed jointly by scientists at NASA's Goddard Space Flight Center (GSFC) and NOAA's National Centers for Environmental Prediction (NCEP) in order to produce such fields (Rodell *et al.*, 2004). Its purpose is similar to that of NLDAS but runs globally at high (0.25°) resolution, and produces results in near-real time (typically within 48 h of the present). In GLDAS, observation-based precipitation and downward radiation and output fields from the best available global coupled atmospheric data assimilation systems are employed as forcing data. The global land-surface fields provided by GLDAS will be used to initialize weather and climate prediction models and will promote various hydrometeorological studies and applications. Currently, GLDAS drives three land-surface models: Mosaic, Noah, and the Community Land Model (CLM). Additional models are slated for future incorporation, including the VIC (Liang *et al.*, 1996) and the Catchment Land–surface Model (Koster *et al.*, 2000).

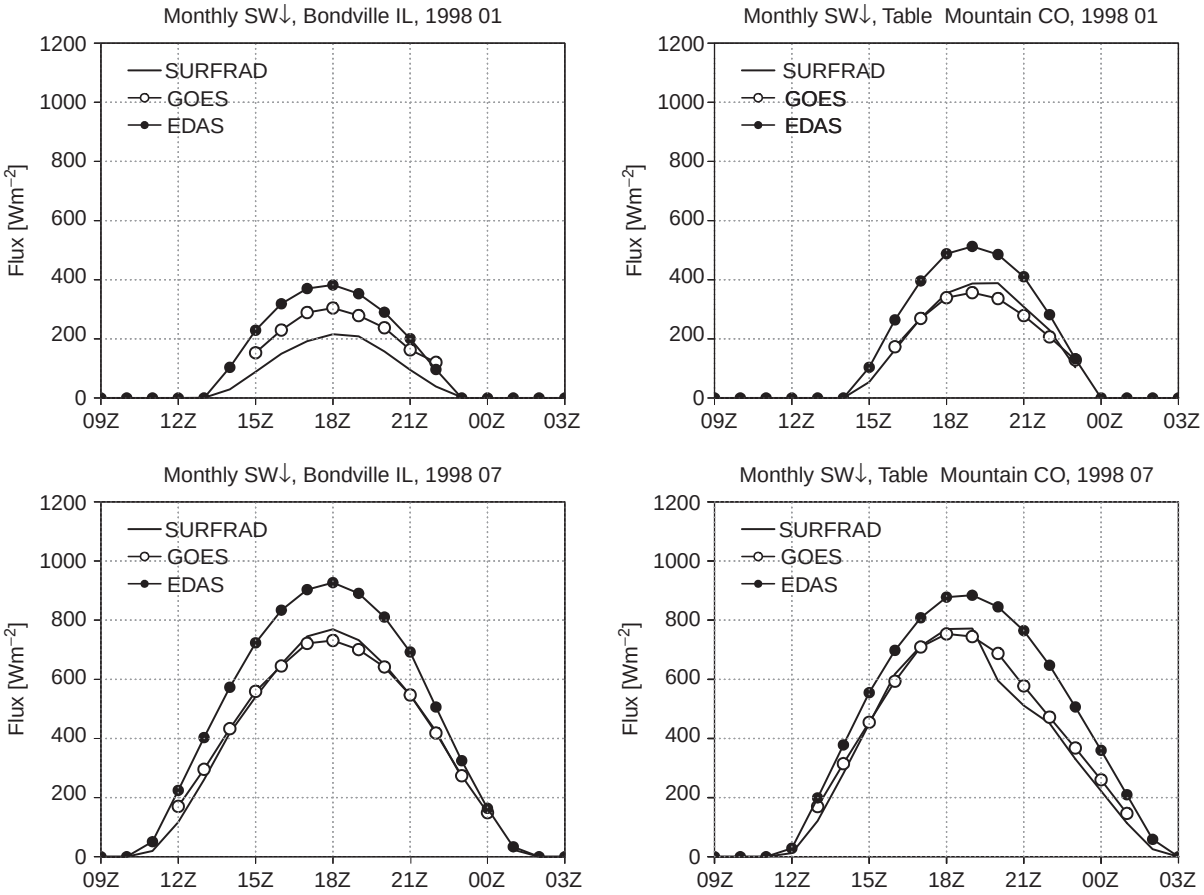


Figure 8 Validation of NLDAS GOES-based (open circles) and EDAS-based (closed circles) shortwave radiation ($W\,m^{-2}$) against SURFRAD ground-based measurements (solid line) at Bondville, IL and Table Mountain, CO for January and July of 1998. Snow cover leads to a wintertime high bias in GOES radiation values at Bondville, IL (upper left) (C.-J. Meng, private communication)

International Satellite Land Surface Climatology Project (ISLSCP) Carbon Cycle Initiative

An important element in predicting climate change is to understand how carbon, energy, and water are exchanged between the atmosphere and the terrestrial biosphere. Climate models require worldwide information about terrestrial changes that are responsible for interactions between the atmosphere and the land surfaces. All activities within the International Satellite Land Surface Climatology Project (ISLSCP) aim at assessing these changes in terms of the physical and biological quantities, which can be related to the exchange of energy and water between the surface and atmosphere. ISLSCP, established in 1983 under the United Nation's Environmental Programme, promotes use of satellite data for the global land-surface data sets needed for climate studies. ISLSCP has played a key role in addressing land-surface processes, developing climate models, experiment design and implementation, and data set development, and as such, has been accepted as a component of the GEWEX program. Its specific objectives are:

1. demonstrate the types of surface and near-surface satellite measurements that are relevant to climate and global change studies;
2. develop and improve algorithms for the interpretation of satellite measurements of land-surface features;
3. develop methods to validate area-averaged quantities derived from satellite measurements for climate simulation models;
4. prepare the groundwork for future operational production of land-surface data sets, which can be directly applied to climate problems.

Initiative I

ISLSCP Initiative I, a pilot project, produced the first global land-cover, hydrometeorology, radiation, and soils data sets regridded to a common 1×1-degree format for 1987–1988. It supported a wide range of uses: weather forecast improvements, hydrological applications, macroscale basin modeling, biogeochemical and carbon tracer models, global carbon flux model comparisons, general circulation models, model validation and comparison,

algorithm development and education. The ISLSCP I data collection is also accessible electronically via the NASA GSFC DAAC.

Initiative II

ISLSCP Initiative II (Hall *et al.*, 2005) consists of a 10-year core global data collection spanning the years 1987 to 1995 at spatial resolutions of one-quarter to one degree. It also includes selected data sets spanning the period, 1982–1999. Included are carbon data sets designed to support global carbon cycling studies. Several WCRP and International Geospehere-Biosphere Program (IGBP) initiatives are leveraged in ISLSCP II, including the GEWEX Global Soil Wetness Project–2 (GSWP-2), the Global Carbon Observing System (GCOS), the NASA's Interdisciplinary Science Projects (IDS), the NASA Seasonal to Interannual Prediction Project (NSIPP), the GEWEX Global Land Atmosphere System Study (GLASS) and others. Many of the data sets are now available electronically at: **`http://islscp2.sesda.com/ISLSCP2_1/html_pages/islscp2_home.html`**

European LDAS Activity (ELDAS)

In the European version of Land Data Assimilation Systems (ELDAS) a fast offline integration scheme is applied to modify NWP model cloud fields based on Meteosat visible and infrared observations. From the updated cloud fields, downward shortwave and longwave radiation at the surface are computed using the NWP radiative transfer model (Meetschen *et al.*, 2004).

Global Soil Wetness Project

Global landmass three-hourly solar forcing for the Global Soil Wetness Project (GSWP-1 and GSWP-2) is being generated at the Center for Ocean-Land-Atmosphere (COLA) Studies. The generation of the surface insolation for the two-year 1987–1988 GSWP-1 is described in Sellers *et al.* (1996). Production and analysis of GSWP-2 (1983–1995) is discussed in Zhao and Dirmeyer (2004).

Oceanic Applications

There is a need for information on accurate global air–sea fluxes for improving and evaluating numerical weather prediction models. As yet, such information is not available (Taylor, 2001). Radiative fluxes absorbed at the ocean surface affect the ocean temperatures and the entire surface energy budget. In the past, latent and sensible heat fluxes have been estimated using bulk formulas (Liu *et al.*, 1979; Geernaert *et al.*, 1987), while more recently, multisensor techniques using microwave radiometers and scatterometers have yielded promising results (Bentamy *et al.*, 2001). As reported in WCRP-112 (WMO/TD-No., 1036) (2000), there are numerous scientific disciplines that require air-sea flux data of heat, water, and momentum. Radiative fluxes are a required input to the estimates of these fluxes. It is also stated that the parameterizations for obtaining the net surface shortwave and long wave fluxes from ship observations are relatively crude and rely on the estimation of cloud amount. For instance, in the studies of Bunker (1976) and Esbensen and Kushnir (1981), shortwave flux parameterization is based on the work of Berlyand (1961) and Budyko (1963). In the study of Hsiung (1986) and Isemer and Hasse (1987), use is made of the parameterization of Reed (1977). In the study of Oberhuber (1988), the parameterization of Zillmann (1972) is used. In the study of Da Silva *et al.* (1994) and Josey *et al.* (1999), a Reed (1977) type approach is used while the study of Lindau (2000) uses the Malevskii *et al.* (1992) approach. In addition to such estimates, number of researchers have also attempted and continue to calculate net fluxes at the ocean surface using radiometers mounted on buoys (e.g. the Tropical Ocean Global Atmosphere Coupled Ocean Atmosphere Response Experiment (TOGA COARE)) (Weller *et al.*, 1996) and are also part of the ongoing Climate Variability and Predictability (CLIVAR) Pan American Climate Studies (PACS), which is a component of the US Global Ocean–Atmosphere–Land System (GOALS) program (**`http://www.ogp.noaa.gov/mpe/cppa/pacs/index.html`**). The overall goals of CLIVAR PACS are to extend the scope and improve the skill of operational seasonal-to-interannual climate prediction over the Americas and the boundary forcing of seasonal-to-interannual climate variations over the Americas. Particular emphasis is placed on warm season rainfall, which is not yet predictable. It is believed that satellite estimates of insolation over the oceans will play an important role in improving the heat flux estimates and will be instrumental in reaching some of the above stated goals. Observations of radiative fluxes over the Atlantic are planned under the Pilot Research Moored Array in the Tropical Atlantic (PIRATA) program. While the evaluation of satellite-based estimates over the oceans has been limited, it is believed that the results over land are good indicators of what can be expected over the oceans, assuming that auxiliary information over the oceans is also available.

FUTURE PROSPECTS

The Coordinated Enhanced Observing Period (CEOP) program is designed to integrate research activities as outlined by the GEWEX Hydrometeorology Panel (GHP). It aims to determine more accurately the water cycle with respect to climate variability and change, as well as providing data to assess the impact of this variability on water resources. Specifically, it plans to carry out intensive regional water and energy cycle experiments with a focus on the five Continental–scale Experiments (CSEs), namely,

BALTEX, GAME, GCIP, LBA, and MAGS. Added to this list was also the Coupling of the Tropical Atmosphere and the Hydrological Cycle (CATCH) region to aid in the study of the Western African Monsoon now known as the *African Monsoon Multidisciplinary Analysis* (AMMA). It is a multidisciplinary project that aims to answer scientific questions related to the understanding of the West African monsoon variability and issues related to prediction and applications. It will require cooperation with several other programs such as the CLIVAR, the emerging Climate and Cryosphere (CliC) project, and the Integrated Global Observing Strategy Partnership (IGOS-P). The Committee on Earth Observation Satellite (CEOS) will contribute to the CEOP satellite data integration and the field campaigns. Satellites are to play a key role in acquiring information under CEOP, in particular, on surface radiative fluxes. New satellite systems will play a key role. These include TERRA, AQUA, ENVISAT, and ADEOS-II, TRMM, Landsat-7, NOAA-K series and other operational satellites. At the same time, numerical weather prediction centers (NCEP, ECMWF, JMA, NASA) have made impressive gains in reanalysis and data assimilation techniques that will benefit CEOP.

SUMMARY

Historical and real-time information on surface and top of the atmosphere radiative fluxes, both on global and regional scales, are now available for the evaluation and improvement of climate models. Such information has already been extensively used, and results have been reported in the open literature. The accuracy of the result is sufficient to detect variability at hourly timescales, as well as to detect major climatic signals, such as the El Nino. Research in this field is now well posed for utilizing new satellite systems that are becoming available, and for integration of ancillary information necessary for the improvement of inference schemes. Many challenges still remain. Clouds are the major modulators of the SRBs. As yet, information on their physical properties and vertical structure is not readily available to evaluate the next stage of possible improvements. Unique situations such as deforestation and biomass burning, volcanic eruptions or pollutant transport, require special attention. For instance, cloud detection techniques that allow to detect the presence of smoke, have to be introduced. There is also a need for improved spectral resolution of the derived fluxes (e.g. PAR and NIR), that are relevant for hydrological and biogeophysical modeling.

Acknowledgments

The work that led to some of the products on surface insolation described in this paper was supported by grants NAGW59634, NAG59916 and NNG04GD65G from NASA and grant NA06GP0404 from OGP NOAA. The help of B. Zhang in the preparation of some of the figures presented here is appreciated. Inputs from K. Mitchell, B. Cosgrove, M. Rodell (NLDAS and GLDAS

Appendix A

1.	CCCM	Canadian Centre for Climate Modelling and Analysis	Victoria, Canada
2.	CCSR	Center for Climate System Research	Tokyo, Japan
3.	CNRM	Centre National de Recherches Météorologiques	Toulouse, France
4.	COLA	Center for Ocean-Land-Atmosphere Studies	Calverton, Maryland
5.	DNM	Department of Numerical Mathematics	Moscow, Russia
6.	ECMWF	European Centre for Medium-Range Weather Forecasts	Reading, England
7.	GISS	Goddard Institute for Space Studies	New York, New York
8.	GLA	Goddard Laboratory for Atmospheres	Greenbelt, Maryland
9.	JMA	Japan Meteorological Agency	Tokyo, Japan
10.	MGO	Main Geophysical Observatory	St. Petersburg, Russia
11.	MPI	Max-Planck-Institut für Meteorologie	Hamburg, Germany
12.	MRI	Meteorological Research Institute	Ibaraki-ken, Japan
13.	NCAR	National Center for Atmospheric Research	Boulder, Colorado
14.	NCEP	National Meteorological Center	Suitland, Maryland
15.	LLNL	Lawrence Livermore National Laboratory	Livermore, California
16.	SUNYA	State University of New York at Albany	Albany, New York
17.	UGAMP	The UK Universities Global Atmospheric Modelling Programme	Reading, England
18.	UIUC	University of Illinois at Urbana Champaign	Urbana, Illinois
19.	UKMO	United Kingdom Meteorological Office	Exeter, UK
20.	YONU	Yonsei University	Seoul, Korea

teams) and P. W. Stackhouse (LaRC GEWEX SRB) are greatly appreciated. The ISCCP D1 data were obtained from the NASA Langley Atmospheric Sciences Data Center (ASDC).

REFERENCES

Augustine J.A., DeLuisi J.J. and Long C.N. (2000) SURFRAD-A national surface radiation budget network for atmospheric research. *Bulletin of the American Meteorological Society*, **81**, 2341–2357.

Avissar R., Silva-Dias P.L., Silva-Dias M.A.F. and Nobre C. (2002) The Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA): insights and future research needs. *Journal of Geophysical Research-Atmospheres*, **107**(D20), 8086, doi:10.1029/2002JD002704.

Baumgartner M.F. and Anderson S.P. (1999) Evaluation of NCEP regional numerical weather prediction model surface fields over the middle Atlantic Bight. *Journal of Geophysical Research-Oceans*, **104**(C8), 18 141–18 158.

Bentamy A., Katsaros K.B., Mestas-Nuñez A.M., Drennan W.M., Forde E.B. and Roquet H. (2001) Satellite estimates of wind speed and latent heat flux over the global oceans. *Journal of Climate*, **16**(4), 636–656.

Berbery E.H., Mitchell K., Benjamin S., Smirnova T., Ritchie H., Hogue R. and Radeva E. (1999) Assessment of surface heat fluxes from regional models. *Journa of Geophysical Research-Atmospheres*, **104**(D16), 19329–19348.

Berlyand T.G. (1961) *Distribution of Solar Radiation on the Continents*, Gidrometeoizdat: Leningrad, pp. 227 (In Russian).

Betts A.K., Ball J.H., Beljaars A.C.J., Miller M.J. and Viterbo P. (1997) Land surface atmosphere interaction: a review based on observational and global modeling perspectives. *Journal of Geophysical Research*, **101**(D3), 7209–7225.

Brisson A., Le Borgne P., Marsouin A. and Moreau T. (1994) Surface irradiance calculated from meteosat sensor data during SOFIA-ASTEX. *International Journal of Remote Sensing*, **15**, 197–203.

Budyko M.I. (Ed.). (1963) *Atlas of Heat Balance of the Earth*, Mezhduvedomstvennyi Geophisichesky Comitet pri Prezidiume Akademii Nauk USSR, Vojekov Main Geophysical Observatory: Moscow, p. 69 (In Russian).

Bunker A.F. (1976) A computation of surface energy flux and annual cycle of the North Atlantic Ocean. *Monthly Weather Review*, **104**, 1122–1140.

Burnash R.J.C., Ferral R.L. and McGuire R.A. (1973) *A Generalized Stream Flow Simulation System: Conceptual Models for Digital Computers*, Joint Federal-State River Forecast Center: Sacramento.

Ceballos J.C., Bottino M.J. and de Souza J.M. (2004) A simplified physical model for assessing solar radiation over Brazil using GOES 8 visible imagery. *Journal of Geophysical Research-Atmospheres*, **109**(D2), Art. No. D02211, 14.

Ceballos J.C. and Moura G.B.A. (1997) Solar radiation assessment using meteosat 4-VIS imagery. *Solar Energy*, **60**, 209–219.

Chahine M.T. (1992) GEWEX is international. *GEWEX News*, **2**(2), 2.

Charlock T.P., Rose F.G., Rutan D.A. and Fu Q. (2000a) Retrievals of the surface and atmospheric radiation budget for January 1998: (a) Validation with collocated observations and (b) Some insights on low latitude atmospheric energetics and circulation. *International Radiation Symposium (IRS-2000)*, St. Petersburg, 24–29 July.

Charlock T.P., Rutan D. and Rose F.G. (2000b) Preliminary CERES retrievals compared with ARM data. *Poster at Tenth ARM Science Team Meeting*, San Antonio, 13–17 March.

Chen F., Mitchell K., Schaake J., Xue Y., Pan H., Koren V., Duan Q. and Betts A. (1996) Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research*, **101**, 7521–7268.

Chou M.-D. (1994) Radiation budgets in the western tropical pacific. *Journal of Climate*, **7**(12), 1958–1971.

Cosgrove B.A., Lohmann D., Mitchell K.E., Houser P.R., Wood E.F., Schaake J.C., Robock A., Sheffield J., Luo L., Duan Q., Pinker R.T., Tarpley J.D., Higgins R.W. and Meng J. (2003) Realtime and retrospective forcing in the North American Land Data Assimilation Systems (NLDAS) project. *Journal of Geophysical Research*, **108**(D22), 8842, doi:10.1029/2002JD003118, 12.

Da Silva A.M., Young C.C. and Levitus S. (1994) *Atlas of Surface Marine Data 1994, NOAA Atlas Nesdis 6 (6 Volumes)*, Available from: U.S. Department Commerce, NODC: Users Services Branch, NOAA/NESDIS E/OC21, Washington.

Dickinson R. (1986) Global climate and its connection to the biosphere. In *Climate-Vegetation Interactions*, Rosenzweig C. and Dickinson R. (Eds.), *Proceedings of a Workshop Held at NASA/Goddard Space Flight Center*: Greenbelt, January 27–29, 1986, pp. 5–8.

Esbensen S.K. and Kushnir Y. (1981) *The Heat Budget of the Global Ocean: An Atlas Based on Estimates from Surface Marine Observations*, Climatic Research Institute Report, No. 29, Oregon State University.

Fouquart Y., Bonnel B. and Ramaswamy V. (1991) Intercomparing shortwave radiation codes for climate studies. *Journal of Geophysical Research*, **96**, 8955–8968.

Frouin R. and Pinker R.T. (1995) Estimating Photosynthetically Active Radiation (PAR) at the earth's surface from satellite observations. *Remote Sensing of Environment*, **51**, 98–107.

Gallo K.P. and Huang A. (1998) Evaluation of the NDVI and sensor zenith angle values associated with the global land AVHRR 1-km data set. *International Journal of Remote Sensing*, **19**, 527–533.

Garrat J.R., Krummel P.B. and Kowalczyk E.A. (1993) The surface energy balance at local and regional scales-a comparison of general circulation model results with observations. *Journal of Climate*, **6**, 1090–1109.

Gates W.L., Boyle J.S., Covey C., Dease C.G., Doutriaux C.M., Drach R.S., Fiorino M., Gleckler P.J., Hnilo J.J., Marlais S.M., *et al.* (1999) An overview of the results of the Atmospheric Model Intercomparison Project (AMIP). *Bulletin of the American Meteorological Society*, **80**, 29–56.

Geernaert G.l, Larsen S.E. and Hansen F. (1987) Measurements of wind stress, heat flux, and turbulence intensity during storm conditions over the North Sea. *Journal of Geophysical Research-Oceans*, **92**(C12), 13 127–13 139.

Gilgen H. and Ohmura A. (1999) The global energy balance archive. *Bulletin of the American Meteorological Society*, **80**(5), 831–850.

Goward S.N. (1989) Satellite bioclimatology. *Journal of Climate*, **2**, 710–720.

Goward S.N. and Huemmrich K.F. (1992) Vegetation canopy PAR absorptance and the normalized difference vegetation index-an assessment using the SAIL model. *Remote Sensing of Environment*, **39**, 119–140.

Gupta S.K., Kratz D.P., Stackhouse P.W. Jr and Wilber A.C. (2001) *The Langley Parameterized Shortwave Algorithm (LPSA) for Surface Radiation Budget Studies (Version 1.0)*, NASA/TP-2001-211272.

Gupta S.K., Ritchey N.A., Wilber A.C., Whitlock C.H., Gibson G.G. and Stackhouse P.W. Jr (1999) A climatology of surface radiation budget derived from satellite data. *Journal of Climate*, **12**(8), 2691–2710, Part 2.

Gupta S.K., Wilber A.C., Ritchey N.A., Whitlock C.H. and Stackhouse P.W. (1997) Comparison of surface radiation fluxes in the NCEP/NCAR reanalysis and the langley 8-year SRB dataset. *First WCRP International Conference on Reanalysis*, Silver Spring, 27–31 October.

Gutman G., Tarpley J.D., Ignatov A. and Olson S. (1995) The enhanced NOAA global land dataset from the advanced very-high resolution radiometer. *Bulletin of the American Meteorological Society*, **76**, 1141–1156.

Hall F.G., Collatz G.G., Los S., Brown de Colstoun E., Landis D. (Eds.) (2005) ISLSCP Initiative II. NASA. DVD/CD-ROM. NASA.

Henderson-Sellers A. (1993) The project for intercomparison of land-surface parameterization schemes. *Bulletin of the American Meteorological Society*, **74**, 1335–1350.

Hicke J.A., Asner G.P., Tucker C., Los S., Birdsey R., Jenkins J.C., Field C. and Randerson J.T. (2002) Trends in North American net primary productivity derived from satellite observations, *Global Biogeochemical Cycles* **16**(2), Art. No. 1019, 22.

Hicks B.B., DeLuisi J.J. and Matt D. (1996) The NOAA Integrated Surface Irradiance Study(ISIS). New surface radiation monitoring network. *Bulletin of the American Meteorological Society*, **77**, 2857–2864.

Hollmann R., Feng J., Leighton H.G., Mueller J. and Stuhlmann R. (1999) ScaRaB as a valuable tool for BALTEX and MAGS: satellite applications for energy budgets and the hydrological cycle. *Advances in Space Research*, **4**(7), 955–958.

Hsiung J. (1986) Mean surface energy fluxes over the global ocean. *Journal of Geophysical Research*, **91**(C9), 10 585–10 606.

Isemer H.–J. and Hasse L. (1987) *The Bunker Climate Atlas of the North Atlantic Ocean*, *Volume 2: Air-Sea Interactions*, Springer-Verlag, Berlin, p. 252.

Josey S.A., Kent E.C. and Taylor P.K. (1999) New insights into the ocean heat budget closure problem from analysis of the SOC air-sea flux climatology. *Journal of Climate*, **12**, 2856–2880.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G., Woollen J., *et al.* (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77**(3), 437–471.

Keller M., Palace M. and Hurtt G. (2001) Biomass estimation in the Tapajos National Forest, brazil-examination of sampling and allometric uncertainties. *Forest Ecology and Management*, **154**, 371–382.

Koster R.D. and Suarez M. (1992) Modeling the land surface boundary in climate models as a composite of independent vegetation stands. *Journal of Geophysical Research-Atmospheres*, **97**(D3), 2697–2715.

Koster R.D., Suarez M.J., Ducharne A., Stieglitz M. and Kumar P. (2000) A catchment-based approach to modeling land surface processes in a general circulation model 1. Model structure. *Journal of Geophysical Research-Atmospheres*, **105**(D20), 24 809–24 822.

Kineman J.J. (Ed.) (1997) *Global Ecosystems Database Disc-B: Including Database, User's Guide, and Dataset Documentation*, USDOC/NOAA National Geophysical Data Center: Boulder, GED: 1B. 640 MB on 1 CDROM.

Leese J.A. (1994) *Implementation Plan for the GEWEX Continental Scale International Project (GCIP). Vol. 3: Strategy Plan for Data Management*, International GEWEX Project Office: Publication Series No. 9, p. 49.

Leese J.A. (1997) *Major Activities Plan for 1998, 1999 and Outlook for 2000 for the GEWEX Continental-Scale International Project (GCIP)*, IGPO Publication Series No. 26, Available from International GEWEX Project Office: Silver Spring.

Li Z., Leighton H.G. and Cess R.D. (1993) Surface net solar radiation estimated from satellite measurements: comparisons with tower observations. *Journal of Climate*, **6**, 1764–1772.

Liang S., Fan H., Chen M., Shuey C.J., Walthall C., Daughtry C., Morisette J., Schaaf C. and Strahler A. (2002) Validating MODIS land surface reflectance and albedo products: methods and preliminary results. *Remote Sensing of Environment*, **83**(1), 149–162.

Liang X., Wood E.F. and Lettenmaier D.P. (1996) Surface soil moisture parameterization of the VIC-2L model: evaluation and modifications. *Global and Planetary Change*, **13**, 195–206.

Lindau R. (2000) *Climate Atlas of the Atlantic Ocean derived from the Comprehensive Ocean Atmosphere Data Set*, Springer Verlag: Berlin, p. 488.

Liu W.T., Katsaros K.B. and Businger J.A. (1979) Bulk parameterization of air-sea exchanges of heat and water-vapor including the molecular constraints at the interface. *Journal of Atmospheric Sciences*, **36**(9), 1722–1735.

Luo L., Robock A., Cosgrove B., Mitchell K., Houser P., Wood E., Schaake J., Lohmann D., Sheffield J., Duan Q., *et al.* (2003) Validation of forcing of the North American Land Data Assimilation System (N-LDAS). *Journal of Geophysical Research-Atmospheres*, **108**(D22), Art. No. 8843, 10.

Malevskii S.P., Girduk G.V. and Egorov B. (1992) *Radiation Balance of the Ocean Surface*, Hydrometeoizdat: Leningrad, p. 148.

Meetschen D., van den Hurk B.J.J.M., Ament F. and Drusch M. (2004) Optimized surface radiation fields derived from meteosat imagery and a regional atmospheric model. *Journal of Hydrometeorology*, **5**(6), 1091–1101.

Menzel W.P. and Purdom J.F. (1994) Introducing GOES-I: the first of a new generation of geostationary operational environmental satellites. *Bulletin of the American Meteorological Society*, **75**, 757–782.

Michalsky J., Dutton E., Rubes M., Nelson D., Stoffel T., Wesley M., Splitt M. and DeLuisi J. (1999) Optimal measurement of surface shortwave irradiance using current instrumentation. *Journal of Atmospheric and Oceanic Technology*, **16**(1), 55–69.

Mitchell K.E., Lohmann D., Houser P.R., Wood E.F., Schaake J.C., Robock A., Cosgrove B.A., Sheffield J., Duan Q.Y., Luo L.F., Higgins R.W., Pinker R.T., Tarpley J.D., Lettenmaier D.P., Marshall C.H., Entin J.K., Pan M., Shi W., Koren V., Meng J., Ramsay B.H. and Bailey A.A. (2004) The multi-institution North American Land Data Assimilation System (NLDAS): utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research-Atmospheres*, **109**(D7), Art. No. D07S90 APR 9 2004, 32.

Nakajima T. (2001) *An Overview of the GAME Radiation Acitivites-2000: Proceedings: The Fifth International Study Conference on GEWEX in Asia and GAME (Vol. 2)*, GAME Publication No. 31 (2), GAME: Nagoya, pp. 363–366.

Nobre C.A., Wickland D. and Kabat P.I. (2001) The large scale biosphere-atmosphere experiment in Amazonia (LBA). *Global Change Newsletter*, **45**, 2–4.

NRC (1998) GCIP: Global Energy and Water Cycle Experiment (GEWEX) continental-scale international project: a review of progress and opportunities. *Report of the Global Energy and Water Cycle Experiment (GEWEX) Panel Climate Research Committee Board on Atmospheric Sciences and Climate Commission on Geosciences*, Environment and Resources, National Research Council, National Academy Press: Washington, p. 93.

Oberhuber J.M. (1988) *An Atlas Based on COADS Data Set: The Budgets of the Heat, Buoyancy and Turbulent Kinetic Energy at the Surface of the Global Ocean*, Rep. Lank-Institut fur Meteorologie: No. 15, Hamburg, p. 100.

Ohmura A., Dutton E.G., Forgan B., Frohlich C., Gilgen H., Hegner H., Heimo A., Konig-Langlo G., McArthur B., Muller G., *et al.* (1998) Baseline Surface Radiation Network (BSRN/WCRP): new precision radiometry for climate research. *Bulletin of the American Meteorological Society*, **79**, 2115–2136.

Ohmura A. and Wild M. (2002) Is the hydrological cycle accelerating? *Science*, **298**(5597), 1345–1346.

Ohring G. and Dodge D.C. (1992) *The NOAA/NASA Pathfinder Program*, Current Problems in Atmospheric Radiation, IRS '92, Deepak Publishing, p. 405.

Pereira E.B., Abreu S.L., Stuhlmann R., Rieland M. and Colle S.M.E. (1996) Survey of the incident solar radiation in Brazil by use of METEOSAT satellite data. *Solar Energy*, **57**(2), 125–132.

Pinker R.T. and Laszlo I. (1992a) Modeling surface solar irradiance for satellite applications on a global scale. *Journal of Applied Meteorology*, **31**, 194–211.

Pinker R.T. and Laszlo I. (1992b) Global distribution of photosynthetically active radiation as observed from satellites. *Journal of Climate*, **5**, 56–65.

Pinker R.T., Laszlo I., Whitlock C.H. and Charlock T.P. (1995) Radiative flux opens new window on climate research. *EOS*, **76**(15), 145.

Pinker R.T., Tarpley J.D., Laszlo I., Mitchell K.E., Houser P.R., Wood E.F., Schaake J.C., Robock A., Lohmann D., Cosgrove B.A., *et al.* (2003) Surface radiation budgets in support of the GEWEX Continental Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research-Atmospheres*, **108**(D22), Art. No. 8844, 18.

Pinker R.T., Zhang B. and Dutton E.G. (2005) Do satellites detect trends in surface solar radiation? *Science*, **308**, 850–854, MS no.: RE1103159/AWM/ATMOS.

Platt T. (1986) Primary production of the ocean water column as a function of surface light intensity: algorithms for remote sensing. *Deep-Sea Research*, **33**, 149–163.

Prince S.D. and Goward S.N. (1995) Global primary production: a remote sensing approach. *Journal of Biogeography*, **22**(4–5), 815–835.

Ramanathan V., Cess R.D., Harrison E.F., Minnis P., Barkstrom B.R., Ahmad E. and Hartmann D. (1989) Cloud-radiative forcing and climate: results from the earth radiation budget experiment. *Science*, **243**, 57–63.

Ramaswamy V. and Freidenreich S.M. (1992) A study of broadband parameterizations of the solar radiative interactions with water vapor and water drops. *Journal of Geophysical Research*, **97**, 11 487–11 512.

Raschke E., Karstens U., Nolte-Holube R., Brandt R., Isemer H.-J., Lohmann D., Lobmeyr M., Rockel B. and Stuhlmann R. (1998) The Baltic sea experiment, BALTEX: a brief overview and some selected results. *Surveys of Geophysics*, **19**, 1–22.

Raschke E., Meywerk J. and Rockel B. (2002) Has the project BALTEX so far met its original objectives? *Boreal Environment Research (BER)*, **7**(3), 175–182.

Raschke E. and Preuss H.J. (1979) The determination of the solar radiation budget at the earth's surface from satellite measurements. *Meteorologische Rundschau*, **32**, 18–28.

Reed R.K. (1977) On estimating insolation over the ocean. *Journal of Physical Oceanography*, **7**, 482–485.

Rind D., Rosenzweig C. and Goldberg G.R. (1992) Modeling the hydrological cycle in assessments of climate change. *Nature*, **358**(6382), 119–122.

Roads J., Lawford R., Bainto E., Berbery H., Fekete B., Gallo K., Grundstein A., Higgins W., Janowiak J., Kanamitsu M., *et al.* (2003) GCIP Water and Energy Budget Synthesis (WEBS). *Journal of Geophysical Research-Atmospheres*, **108**(D16), Art. No. 860, 39.

Rodell M., Houser P.R., Jambor U., Gottschalck J., Mitchell K., Meng C.-J., Arsenault K., Cosgrove B., Radakovich J., Bosilovich M., *et al.* (2004) The global land data assimilation system. *Bulletin of the American Meteorological Society*, **85**(3), 381–394.

Roesch A., Wild M., Pinker R.T. and Ohmura A. (2002) Comparison of spectral surface albedos and their impact on the GCM simulated surface climate. *Journal of Geophysical Research-Atmospheres*, **107**(D14), Art. No. 4221, 18.

Rogers E., Black T.L., Deaven D.G. and DiMego G.J. (1996) Changes to the operational "Early" eta analysis/forecast system

at the National Centers for Environmental Prediction. *Weather and Forecasting*, **11**, 391–413.

Rossow W.B. and Schiffer R.A. (1991) ISCCP cloud data products. *Bulletin of the American Meteorological Society*, **72**(1), 2–20.

Rossow W.B. and Schiffer R.A. (1999) Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, **80**(11), 2261–2287.

Rossow W.B. and Zhang Y.-C. (1995) Calculation of surface and top of the atmosphere radiative fluxes from physical quantities based on ISCCP data sets, 2. Validation and first results. *Journal of Geophysical Research*, **97**, 1167–1197.

Running S.W., Collatz G.J., Washburne J. and Sorooshian S. (1999) Land ecosystems and hydrology. *EOS Science Plan*, **5**, 197–260.

Rutan D.A., Rose F.G., Smith N.M. and Charlock T.P. (2001) Validation data set for CERES surface and atmospheric radiation budget (SARB). *WCRP/GEWEX Newsletter*, **11**(1), 11–12.

Schaaf C.B., Gao F., Strahler A.H., Lucht W., Li X.W., Tsang T., Strugnell N.C., Zhang X.Y., Jin Y.F., Muller J.P., Lewis P., Barnsley M., Hobson P., Disney M., Roberts G., Dunderdale M., Doll C., d'Entremont R.P., Hu B.X., Liang S.L., Privette J.L., Roy D. (2002) First Operational BRDF, Albedo and Nadir Reflectance Products from MODIS. *Remote Sensing of Environment*, **83**, 135–148.

Schiffer R.A. and Rossow W.B. (1985) ISCCP global radiance data set: a new resource for climate research. *Bulletin of the American Meteorological Society*, **66**, 1498–1505.

Schmetz J. (1989) Towards a surface radiation climatology: retrieval of downward irradiance from satellites. *Atmospheric Research*, **23**, 287–321.

Schmetz J. (1991) Retrieval of surface radiation fluxes from satellite data. *Dynamics of Atmospheres and Oceans*, **16**(1–2), 61–72.

Schmetz J. (1993) Relationship between solar net radiative fluxes at the top of the atmosphere and at the surface. *Journal of the Atmospheric Sciences*, **50**(8), 1122–1132.

Sekiguchi M., Nakajima T., Suzuki K., Kawamoto K. and Higurashi A. (2003) *A Study of Atmospheric Radiation Budget in Asia, Proceedings: The Fifth International Study Conference on GEWEX in Asia and GAME (Vol. 3)*, GAME Publication No. 31 (3), GAME: Nagoya, pp. 778–783.

Sellers P.J., Collatz J., Hall F.G., Meeson B.W., Closs J., Corprew F., McManus J., Myers D., Sun K.-J., Dazlich D., *et al.* (1996) The ISLSCP initiative I global datasets: surface boundary conditions and atmospheric forcing for land-atmosphere studies. *Bulletin of the American Meteorological Society*, **77**(9), 1987–2006.

Shi Y. and Long C.N. (2002) Techniques and methods used to determine the best estimate of radiation fluxes at SGP central facility. *12th ARM Science Team Meeting Proceedings*, St. Petersburg, 8–12 April.

Sorooshian S. (2003) GEWEX support reaffirmed at recent WCRP/JSC meeting. *Gewex News*, **13**(2), 2.

Sorooshian S., Whitaker M.P.L. and Hogue T.S. (2002) Regional and global hydrology and water resources issue: the role of international and national programs. *Aquatic Sciences*, **64**(4), 317–327.

Stackhouse P.W. Jr, Cox S.J., Gupta S.K., Chiaachio M. and Mikovitz J.C. (2000) The WCRP/GEWEX surface radiation budget project release 2: an assessment of surface fluxes at 1 degree resolution. *International Radiation Symposium 2000*, St. Petersburg, 24–29 July.

Stackhouse P.W. Jr, Gupta S.K., Cox S.J., Mikovitz J.C. and Chiaachio M. (2002) New results from the NASA/GEWEX surface radiation budget project: evaluating El nino effects at different scales, *11*th *Conference on Atmospheric Radiation*, American Meteorological Society: Ogden, 3–7 June.

Stackhouse P.W. Jr, Gupta S.K., Cox S.J., Mikovitz J.C., Zhang T. and Chiacchio M. (2004) A 12-year surface radiation budget data set. *Gewex News*, **14**(4), 10–12.

Stewart R.E., Leighton H.G., Marsh P., Moore G.W.K., Ritchie H., Rouse W.R., Soulis E.D., Strong G.S., Crawford R.W. and Kochtubajda B. (1998) The Mackenzie GEWEX study: the water and energy cycles of a major North American river basin. *Bulletin of the American Meteorological Society*, **79**, 2665–2684.

Stuhlmann R., Rieland M. and Raschke E. (1990) An improvement of the IGMK model to derive total and diffuse solar radiation at the surface from satellite data. *Journal of Applied Meteorology*, **29**, 586–603.

Sui C.-H., Rienecker M.M., Li X., Lau K.-M., Laszlo I. and Pinker R.T. (2002) The impacts of daily surface forcing in the upper ocean over tropical Pacific: a numerical study. *Journal of Climate*, **16**(4), 756–766.

Takamura T., Okada I., Takeuchi N. and Nakajima T. (2001) Estimation of surface solar radiation from satellite data and its validation using SKYNET data. *Proceedings: The Fifth International Study Conference on GEWEX in Asia and GAME (Vol. 2)*, GAME Publication No. 31 (2), GAME: Nagoya, pp. 536–541.

Tarpley J.D. (1979) Estimating incident solar radiation at the surface from geostationary. *Journal of Applied Meteorology*, **18**, 1172–1181.

Tarpley J.D., Pinker R.T. and Laszlo I. (1996) Experimental GOES shortwave radiation budget for GCIP. *Second International Scientific Conference on the Global Energy and Water Cycle*, Washington, 17–21 June 1996.

Taylor P.K. (2001) *Intercomparison and Validation of Ocean-Atmosphere Energy Flux Fields*, Joint WCRP/SCOR Working group on Air-Sea Fluxes Final Report, WCRP-112, WMO/TD-No. 1036, WMO, p. 306.

Townshend J.G.R. and Justice C.O. (1995) Spatial variability of images and the monitoring of changes in the normalized difference vegetation index. *International Journal of Remote Sensing*, **16**, 2187–2195.

WCRP-67 (1992) *GEWEX Continental Scale International Project*, WMO/TD-No. 461, WMO.

WCRP-112 (WMO/TD-No. 1036) (2000) Intercomparison and validation of ocean-atmosphere energy flux fields by members of the WGASF. In *Final report of the Joint WCRP/SCOR Working Group on Air-Sea Fluxes (SCOR Working Group 110)*, Taylor P.K. (Ed.), World Climate Research Programme/World Meteorological Organization: Geneva.

Weller R.A. and Anderson S.P. (1996) Surface meteorology and air-sea fluxes in the western equatorial Pacific warm pool during

the TOGA coupled ocean-atmosphere response experiment. *Journal of Climate*, **9**, 1959–1990.

Whitlock C.H., Charlock T.P., Staylor W.F., Pinker R.T., Laszlo I., Ohmura A., Gilgen H., Konzelman T., DiPasquale R.C., Moats C.D., *et al.* (1995) First global WCRP shortwave surface radiation budget data set. *Bulletin of the American Meteorological Society*, **76**(6), 1–18.

Wielicki B.A., Cess R.D., King M.D., Randall D.A. and Harrison E.F. (1995) Mission to planet earth-role of clouds and radiation in climate. *Bulletin of the American Meteorological Society*, **76**(11), 2125–2153.

Wielicki B.A., Del Genio A.D., Wong T.M., Chen J.Y., Carlson B.E., Allan R.P., Robertson F., Jacobowitz H., Slingo A., Randall D.A., *et al.* (2002) Changes in tropical clouds and radiation response. *Science*, **296**(5576), U2–U3.

Wild M. (2005) Solar radiation budgets in atmospheric model intercomparisons from a surface perspective. *Geophysical Research Letters*, **32**(7), L07704 10.1029/2005GL022421, 4.

Wild M., Ohmura A., Gilgen H. and Roeckner E. (1995) Validation of GCM simulated radiative fluxes using surface observations. *Journal of Climate*, **8**, 1309–1324.

Wild M., Ohmura A., Gilgen H., Roeckner E., Giorgetta M. and Morcrette J.J. (1998) The disposition of radiative energy in the global climate system: GCM-calculated versus observational estimates. *Climate Dynamics*, **14**(12), 853–869.

Wood E.F., Lettenmaier D.P., Liang X., Nijssen B. and Wetzel S.W. (1997) Hydrological modeling of continental scale basins. *Review of Earth and Planetary Sciences*, **25**, 279–300.

Zhang B., Pinker R.T. and Stackhouse P.W., Jr. (2005) An EOF Iteration approach to obtain homogeneous radiative fluxes from inhomogeneous satellites observations. *Journal of Atmospheric and Oceanic Technology* submitted for publication.

Zhang Y.-C., Rossow W.B., Lacis A.A., Oinas V. and Mishchenko M.I. (2004) Calculation of radiative flux profiles from the surface to top-of-atmosphere based on ISCCP and other global datasets: refinements of the radiative transfer model and the input data. *Journal of Geophysical Research*, **109**, D19105, doi:10.1029/2003JD004457, 27.

Zhao M. and Dirmeyer P. (2004) Pattern and trend analysis of temperature in a set of seasonal ensemble simulations. *Geophysical Research Letters*, **31**, doi:10.1029/2003GL018579, 4.

Zhao M., Heinsch F.A., Nemani R.R. and Running S.W. (2005) *Improvements of the MODIS terrestrial gross and net primary production global data set*. *Remote Sensing of Environment* **95**(2), 164–176.

Zhou L., Laszlo I. and Pinker R.T. (1996) Development of narrow to broadband transformations for GOES 8. *Second International Scientific Conference on the Global Energy and Water Cycle*, Washington, 17–21 June 1996.

Zillmann J.W. (1972) *A Study of Some Aspects of the Radiation and the Heat Budgets of the Southern Hemisphere Oceans*, Bureau of Meteorological Department of the Interior: Canberra.

50: Estimation of the Surface Energy Balance

ZHONGBO SU

International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands

Radiation and latent and sensible heating are among the most important processes in land–atmosphere exchanges. Hence, quantitative understanding and accurate estimation of these fluxes at large scales are imperative for research and applications in areas ranging from numerical weather forecast, climate research, water cycle study, to water resources management, sustainable agricultural production, and ecological conservation. Quantitative remote sensing is probably the only efficient and economically viable technology to provide regional to global radiometric observations of several physical quantities that are relevant to the estimation of these fluxes. This article examines the physical principles underlying the radiation and turbulent heating, along with exploring the possibilities and difficulties for estimation of these fluxes using remote sensing measurements. Past efforts, recent progresses are reviewed, and future research needs are identified. In particular, the Surface Energy Balance System(SEBS), recently developed in Wageningen, is introduced and its strengths and weakness are analyzed in view of future space-borne sensor systems. Several applications derived on the basis of SEBS, including estimation of turbulent heat fluxes, evaporative fraction and actual evaporation, estimation of relative soil moisture, and drought monitoring are discussed.

INTRODUCTION

The international environmental sciences communities have seen a very rapid evolution over the last years. A new global, interdisciplinary science of the Earth System has emerged at the interface of science and policy development. The need to bridge, in a more focused way, the gap between earth sciences and sustainable management of the terrestrial environment is already emerging as an overall priority both in the International Geosphere Biosphere Programme (IGBP) and the World Climate Research Programme (WCRP) (e.g. Kabat, 1999). In this regard, understanding the interactions between terrestrial ecosystems and the atmosphere is essential to address issues such as climate change and environmental degradation. Extensive scientific investigations over the past decades have demonstrated beyond doubt that the biosphere is tightly coupled to its physical environment over a wide range of space and timescales and the atmospheric circulation is strongly influenced by land surface properties and their variability in space and time. Since radiation and turbulent heating are among the most important processes in land–atmosphere exchanges, accurate estimation of these fluxes at large scales is imperative for research and applications in areas ranging from numerical weather forecast, climate research, water cycle study, to water resources management, sustainable agricultural production, and ecological conservation. Such importance may be viewed in the following context. In dealing with climate–land interactions – especially in studies concerning terrestrial water, energy, and carbon cycles, models are often used to assist urgent policy decisions on environmental issues – but we must ask the following questions before the results derived from such models being used for further decision-making processes:

Is the interaction with a dynamic terrestrial biosphere in global models properly understood so that our forecasts of the long-term evolution of climate are reliable?
Are processes at scales from global to local properly linked so that our assessment of the impacts of climate variability on land and water resources is reliable and scenarios on the evolution of the terrestrial biosphere are realistic?

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Are the political decisions based on the current model simulations/predictions/scenarios well founded?

To be able to answer these questions, there is a critical need to understand the feedbacks between the land surface and the atmosphere at various scales (Wood, 1998). As shown by recent works, the land component of the climate system must be viewed as an active participant in the coupled system, rather than a passive recipient of atmospheric forcing (Entekhabi *et al.*, 1999; Koster *et al.*, 2000). In this view, accurate representation of land surface dynamics is more important than it had been thought previously. It has also been suggested that, in the 2001 Intergovernmental Panel on Climate Change (IPCC) report, the role of land surface in modifying the climate has not been adequately considered and the effects of land-cover change on climate may be comparable to, and, perhaps larger than, the effects on climate due to changes in atmospheric composition. The current level of scientific understanding in land–atmosphere interactions has been characterized as "very low" (Pielke, 2001a, b). A recent study by Kalnay and Cai (2003) clearly shows that the impact of land use on climate change has been underestimated. Further, the current parameterizations of land processes in all earth system models are based on simple conceptions that are representative, at best, of homogeneous flat land surfaces, and few observation data sets over complex terrain exist to allow integrated long-term off-line testing that is essential to evaluate land surface parameterizations in climate models. Abilities to estimate radiation and turbulent heat fluxes using radiometric measurements from satellite sensors will help to resolve the above-mentioned uncertainties and difficulties.

Several land–atmosphere exchange processes can be distinguished in terms of exchanges of radiation, heat, water, and carbon, and other gasses. More specifically, we may consider: solar radiation, thermal radiation, sensible heat flux, latent heat flux including phase change, soil heat flux and advection, precipitation, water and vapor transport in soil, as well as biochemical processes in soil and canopy. The driving forces of these processes are solar radiation and wind. Since most of these processes vary continuously in space and time from local to global scale, and from seconds to decades and centuries, their adequate quantification also requires measurements with corresponding spatial-temporal coverage. It is in this aspect that remote sensing can play an essential role. In the following, we will briefly review some progress made in the estimation of turbulent heat fluxes.

A BRIEF HISTORY IN REMOTE SENSING OF TURBULENT HEAT FLUXES

The estimation of atmospheric turbulent fluxes (or evaporation when the latent heat flux is expressed in water depth; in this article, evaporation refers to the rate of the conversion of liquid water to water vapor from water, soil surfaces, and through vegetation) at the land surface has long been recognized as the most important process in the determination of the exchanges of energy and mass among hydrosphere, atmosphere, and biosphere (e.g. Bowen, 1926; Penman, 1948; Monteith, 1965; Priestley and Taylor, 1972; Brutsaert, 1982; Morton, 1983; Menenti, 1984; Famiglietti and Wood, 1994; Sellers *et al.*, 1996; Su and Menenti, 1999). Conventional techniques that employ point measurements to estimate the components of energy balance are only representative of local scales and cannot be extended to large areas because of the heterogeneity of the land surface and the dynamic nature of heat transfer processes. Remote sensing is probably the only technique that can provide representative measurements of several relevant physical parameters at scales from a point to the whole globe. Techniques using remote sensing information to estimate atmospheric turbulent fluxes are therefore essential when dealing with processes that cannot be represented only by point measurements.

In developing remote sensing algorithms for the estimation of atmospheric turbulent fluxes, two basic physical principles, the conservation of energy and turbulent transport, must be considered. The former is the basis of the energy balance approach, and the major challenge is the ability to determine the various physical quantities involved with the required accuracy. The rationale behind the energy balance approach is that evaporation is a change of state of water by demanding a supply of energy for vaporization. The whole problem then reduces to determine all other sources and sinks for energy such as to leave evaporation as the only unknown. The latter recognizes the importance of wind in transporting vapor away from the evaporating surface and, when successful, providing a direct estimate to evaporation. This is often called the *aerodynamic approach* and employs, typically, gradients of wind velocity, temperature, and water vapor density in the near-surface atmosphere where the measurements of these gradients are available. Since the energy budget (i.e., available energy to the surface less the soil heat flux) in the energy balance approach needs to be distributed between sensible heat and latent heat fluxes, which involves, again, the principle of turbulent transport, a complete treatment of both the conservation of energy and turbulent transport processes becomes necessary for developing relevant remote sensing algorithms for these fluxes.

In general, we can distinguish two types of methodologies in remote sensing of turbulent heat fluxes and evaporation: analytical versus (semi-) empirical. The former takes into consideration detailed physical processes at the scale of interest but usually involves complex relationships and requires various input variables, including those that can be observed directly by radiometric measurements

and meteorological variables at a proper reference height. The latter tries to employ empirical relationships and data available chiefly from remote sensing observations. Representative works of the first type include, among others, Jackson *et al.* (1981, 1988) that derived the Crop Water Stress Index (CWSI) by applying the Penman–Monteith equation (Monteith, 1981) to radiometric measurements; Kalma and Jupp (1990), who utilized the dual-source model proposed by Shuttleworth and Wallace (1985), which, by itself, was an extension of the Penman–Monteith approach to take into account the soil and the canopy explicitly; Moran *et al.* (1994) further extended the CWSI approach to partial canopies (the so-called trapezoid-method). More recent work in the same line are those of Chehbouni *et al.* (2001) and Boegh *et al.* (2002), who thoroughly examined the CWSI concept for its strengths and weakness using detailed field measurements collected in Denmark. Other relevant works but of different concepts are those of Chanzy *et al.* (1995) for soil evaporation and Norman *et al.* (1995) using a new type of dual-source model aiming at utilizing directional radiometric measurements. More recently, Kustas and Norman (1999) applied actual soil and vegetation component temperatures to the dual-source model of Norman *et al.* (1995) but did not obtain better results than using only composite temperature without changing the applied Priestley and Taylor (1972) coefficient to a much high value. Mecikalski *et al.* (1999) have applied the Norman *et al.* (1995) approach to continental scale with encouraging results. Efforts related to combined modeling and data assimilation using radiometric measurements have been reported by Castelli *et al.* (1999), Olioso *et al.* (2002), Boni *et al.* (2001a, b), and Caparrini *et al.* (2003, 2004).

Where the (semi) empirical approaches are concerned, the earlier works of Jackson *et al.* (1977) and Seguin and Ittier (1983) are representative and many more recent works still follow the same type of approaches but incorporate more remote sensing variables (e.g. in Nishida *et al.*, 2003a, b, besides using remotely sensed surface radiometric temperature as in Jackson *et al.* (1977), vegetation index is also used).

In addition, Menenti (1993), Carlson *et al.* (1995), Kustas and Norman (1996), and Zhang (1996) have provided excellent reviews to the then up-to-date approaches in remote sensing of turbulent heat fluxes and evaporation.

Similarly, the progress made in Wageningen, the Netherlands, can also be mapped in the analytical *versus* (semi-) empirical fashion. The foundation of the analytical approaches were laid by Menenti (1984) by proposing a two-layer combination equation for a drying soil that was later shown by Menenti (1993) to be able to reduce to the Penman–Monteith combination equation and was also shown by Stanghellini (1987) to be equally valid for a green-house canopy. In a further attempt, Menenti and Choudhury (1993) extended the CWSI concept to the so-called Surface Energy Balance Index (SEBI) approach. While the CWSI was based on surface meteorological scaling, the SEBI concept used Planetary Boundary Layer (PBL) Scaling. However, the parameterization used in SEBI was limited to the then state-of-the-art concepts, namely, the ratio between aerodynamic roughness and thermal dynamic roughness was taken as 10 and the stratification correction was simply taken as 2.9. Application of the SEBI concept to the Aral Sea by Menenti *et al.* (2001) revealed that the parameterization was probably not universal and caused some unexplained scatters in the results. More recently, Su (2002) has proposed the Surface Energy Balance System (SEBS) by extending the SEBI concept with a dynamic model for thermal roughness (Su *et al.*, 2001), the Bulk Atmospheric Similarity (BAS) theory of Brutsaert (1999) for PBL scaling, and the Monin–Obukhov Atmospheric Surface Layer (ASL) similarity for surface layer scaling such that SEBS can be used for both local scaling and regional scaling under all atmospheric stability regimes, thus providing a link for radiometric measurements and atmospheric models at various scales. Using SEBS, Jia *et al.* (2003) have successfully coupled model forecast fields of a large scale Numerical Weather Prediction (NWP) model to radiometric measurements from the Along Track Scanning Radiometer (ATSR) onboard the European Remote Sensing Satellite (ERS-2). Rauwerda *et al.* (2002) have extended SEBS to a parallel-source model and have showed significant improvement in estimated turbulent heat fluxes. On the application side, SEBS has been used to generate daily, monthly, and annual evaporation in a semiarid environment (Li, 2001; Su *et al.*, 2003a) and for drought monitoring (Su *et al.*, 2003b). Another development is that of Jia *et al.* (2001), who have proposed a dual-source model for using component (soil and vegetation) temperatures such as those derivable form ATSR data. The scheme is similar to the Shuttleworth and Wallace (1985) dual-source model but employs the boundary-resistance formulation of Stanghellini (1987) and shows much flexibility in dealing with heterogeneous surfaces.

On the empirical side, the work of Nieuwenhuis *et al.* (1985) was among the earliest attempts but was valid only for single crops. Later, Bastiaanssen (1995) proposed the Surface Energy Balance Algorithm for Land (SEBAL) that required simultaneous presence of absolute dry and wet pixels and has been used for many irrigation studies. More recently, Su *et al.* (1999) have made correction in SEBAL to remedy a theoretical problem and added a scheme to apply NWP fields with an up-scaling and down-scaling approach. In another effort, Roerink *et al.* (2000) proposed a Simplified Surface Energy Balance Index (S-SEBI) by fitting dry and wet cases present in the spatial radiometric data and showed reasonable success for application to semiarid areas.

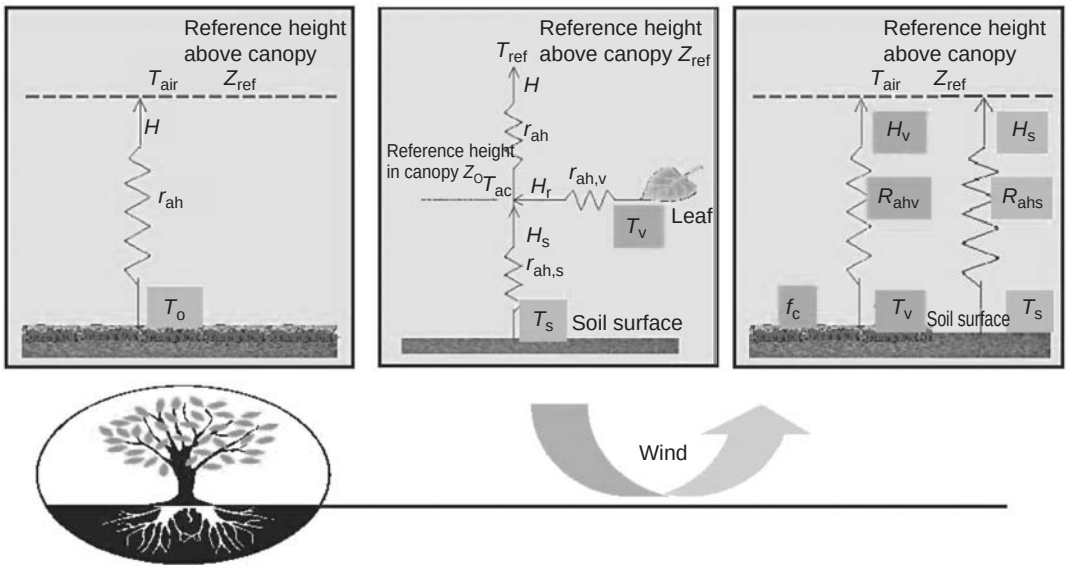


Figure 1 Flux parameterization schemes. Remote sensing of heat fluxes and evaporation – single-source, dual-source, and parallel-source has been shown in the figure. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

In order to give a clear picture of the differences in various schemes encountered in turbulent heat fluxes parameterization, Figure 1 summarizes the single-source, dual-source, and parallel-source schemes. In the following, we will review the difficulties and the relationships between radiometric measurements and atmospheric boundary meteorological variables in estimation of turbulent heat fluxes by means of SEBS (Su, 2002).

THE SURFACE ENERGY BALANCE SYSTEM (SEBS) FOR ESTIMATION OF TURBULENT HEAT FLUXES AND EVAPORATION

SEBS was developed for the estimation of atmospheric turbulent fluxes using satellite earth observation data more coherently. SEBS, as proposed by Su (2002), consists of: a set of tools for the determination of the land surface physical parameters, such as albedo, emissivity, temperature, vegetation coverage, and so on, from spectral reflectance and radiance; an extended model for the determination of the roughness length for heat transfer of Su *et al.* (2001); and a new formulation for the determination of the evaporative fraction on the basis of energy balance at limiting cases. In the application, SEBS requires as inputs three sets of information. The first set consists of land surface albedo, emissivity, temperature, fractional vegetation coverage and leaf area index, and the height of the vegetation (or roughness height). When vegetation information is not explicitly available, the Normalized Difference Vegetation Index (NDVI) is used as a surrogate. These inputs can be derived from remote sensing data in conjunction with other information about the concerned surface. The second set includes air pressure, temperature, humidity, and wind speed at a reference height. The reference height is the measurement height for point application and the height of the PBL for regional application (in this latter case, PBL-averaged meteorological variables are to be used. See later for detailed discussions). This data set can also be variables estimated by a large-scale meteorological model. The third data set includes downward solar radiation and downward longwave radiation that can either be direct measurements, model output, or parameterization.

In SEBS, the friction velocity, the sensible heat flux, and the Obukhov stability length are obtained by solving the system of nonlinear equations. For field measurements performed at a height of a few meters above ground, since the surface fluxes are related to surface variables and variables in the atmospheric surface layer, all calculations use the Monin–Obukhov Similarity (MOS) functions given by Brutsaert (1999). By replacing the MOS stability functions with the Bulk Atmospheric Boundary Layer

Similarity (BAS) functions proposed by Brutsaert (1999), the system equations can be used to relate surface fluxes to surface variables and the mixed layer atmospheric variables provided either by radiosonde data or obtained from atmospheric model fields. The relevant atmospheric boundary layer (ABL) that needs to be considered in different scaling is shown in Figure 2 (*see also* **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**). The determination of the evaporative fraction (the ratio of latent heat flux to the available energy) on the basis of energy balance at limiting cases is carried out, and, finally, the turbulent heat fluxes are determined by utilizing the surface energy balance. Further, by utilizing the conservative characteristics of the evaporative fraction, the daily evaporation can be determined, given the total daily available energy. The different steps involved in SEBS are discussed below.

Determination of Evaporative Fraction Based On Energy Balance at Limiting Cases

The surface energy balance is written in SEBS as

$$R_n = G_0 + H + \lambda E \qquad (1)$$

where R_n is the net radiation, G_0 is the soil heat flux, H is the turbulent sensible heat flux, and λE is the turbulent latent heat flux (λ is the latent heat of vaporization and E is the actual evapotranspiration).

The equation to calculate the net radiation is given by

$$R_n = (1 - \alpha) \cdot R_{swd} + \varepsilon \cdot R_{lwd} - \varepsilon \cdot \sigma \cdot T_0^4 \qquad (2)$$

where α is the albedo, R_{swd} is the downward solar radiation, R_{lwd} is the downward longwave radiation, ε is the emissivity of the surface, σ is the Stefan-Bolzmann constant, and T_0 is the surface radiative temperature measured by a remote sensor (*see also* **Chapter 52, Estimation of Surface Temperature and Surface Emissivity, Volume 2**). α, ε, and T_0 can be derived from remote sensing data from the visible to the thermal infrared spectral range. The simplest of form to calculate the downward solar radiation is $R_{swd} = I_{sc} \cdot e_0 \cdot \cos\theta_z \cdot \exp(-m \cdot \tau)$, where $I_{sc} = 1367\ W \cdot m^{-2}$ is the solar constant, e_0 the eccentricity factor, θ_z the solar zenith angle, m the air mass, and τ the optical thickness. Details on the determination of all the parameters can be found in Iqbal (1983) (*see also* **Chapter 39, Surface**

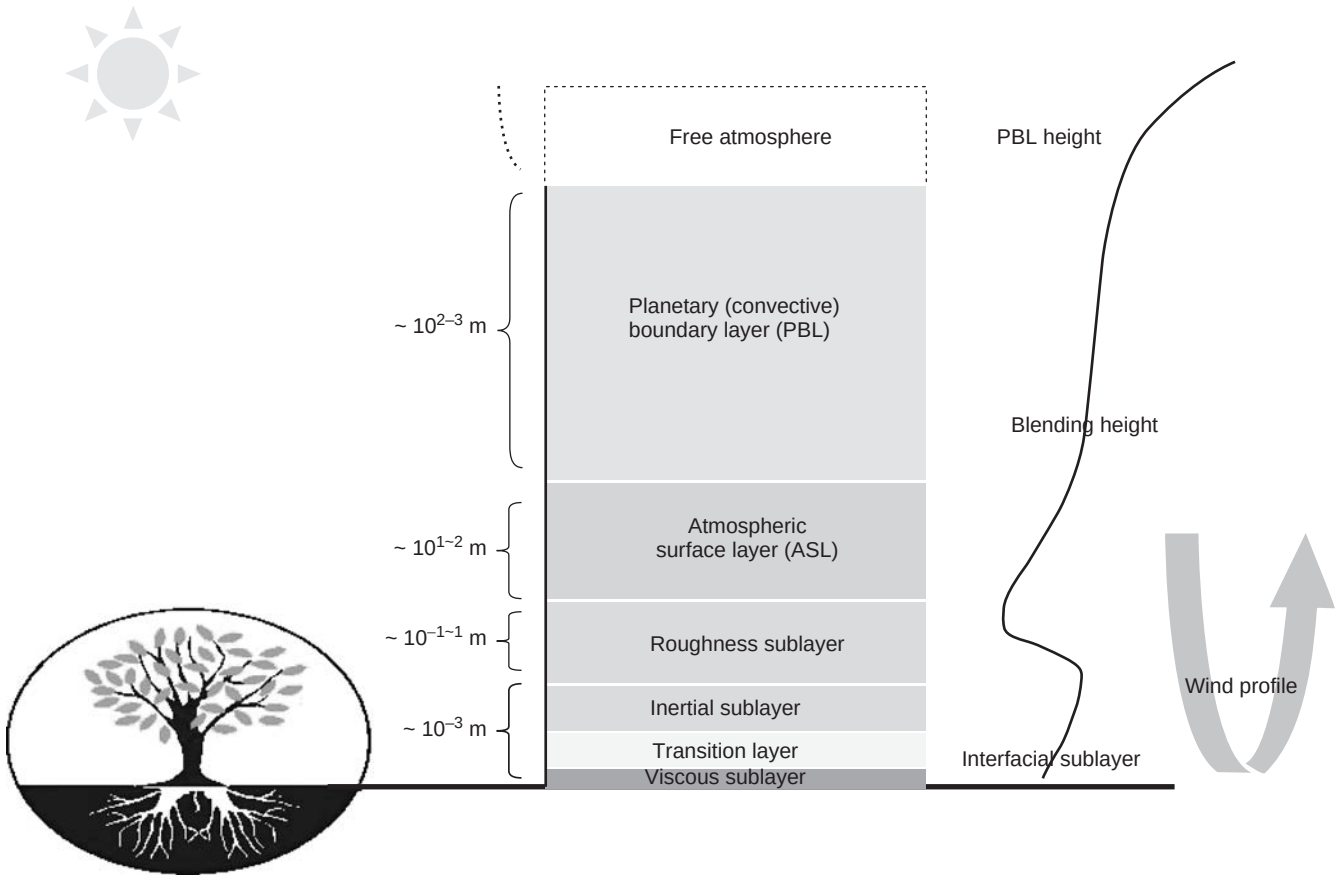


Figure 2 Typical structure of the atmospheric boundary layer. SEBS includes both surface scaling and PBL scaling. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Radiation Balance, Volume 1). The downward longwave radiation R_{lwd} can be calculated as $R_{lwd} = \varepsilon_a \sigma T_a^4$ when there is no measurement, where ε_a is the emissivity of the atmosphere that can be estimated using the Swinbank formula as given by Campbell and Norman (1998) in the form $\varepsilon_a = 9.2 \cdot 10^{-6} \cdot (T_a + 273.15)^2$, with T_a the air temperature at the reference height. Su *et al.* (2001) and Su (2002) have shown that equation (2) provides accurate estimation of the net radiation.

The equation to calculate the soil heat flux is parameterized as

$$G_0 = R_n \cdot [\Gamma_c + (1 - f_c) \cdot (\Gamma_s - \Gamma_c)] \tag{3}$$

in which it is assumed that the ratio of soil heat flux to net radiation is $\Gamma_c = 0.05$ for full vegetation canopy (Monteith, 1973) and $\Gamma_s = 0.315$ for bare soil (Kustas and Daughtry, 1989). An interpolation is then performed between these limiting cases using the fractional canopy coverage, f_c, which can be determined from remote sensing data.

In order to determine the evaporative fraction, energy balance considerations at limiting cases are used. Under the dry limit, the latent heat (or the evaporation) becomes zero due to the limitation of soil moisture, and the sensible heat flux is at its maximum value. From equation (1), it follows,

$$\begin{aligned} \lambda E_{\text{dry}} &= R_n - G_0 - H_{\text{dry}} \equiv 0, \\ \text{or} \quad H_{\text{dry}} &= R_n - G_0 \end{aligned} \tag{4}$$

Under the wet limit, where the evaporation takes place at potential rate, λE_{wet}, (i.e., the evaporation is only limited by the available energy under the given surface and atmospheric conditions), the sensible heat flux takes its minimum value, H_{wet}, that is,

$$\begin{aligned} \lambda E_{\text{wet}} &= R_n - G_0 - H_{\text{wet}}, \\ \text{or} \quad H_{\text{wet}} &= R_n - G_0 - \lambda E_{\text{wet}} \end{aligned} \tag{5}$$

The relative evaporation then can be evaluated as

$$\Lambda_r = \frac{\lambda E}{\lambda E_{\text{wet}}} = 1 - \frac{\lambda E_{\text{wet}} - \lambda E}{\lambda E_{\text{wet}}} \tag{6}$$

By substituting equations (1), (5), and (4) in equation (6) and after some algebra, one obtains

$$\Lambda_r = 1 - \frac{H - H_{\text{wet}}}{H_{\text{dry}} - H_{\text{wet}}} \tag{7}$$

The evaporative fraction is finally given by:

$$\Lambda = \frac{\lambda E}{H + \lambda E} = \frac{\lambda E}{R_n - G} = \frac{\Lambda_r \cdot \lambda E_{\text{wet}}}{R_n - G} \tag{8}$$

Equations (1–8) constitute the basis formulation of SEBS. Further, in SEBS, the actual sensible heat flux H is obtained by solving a set of non-linear equations and is constrained in the range set by the sensible heat flux at the wet limit H_{wet}, and the sensible heat flux at the dry limit H_{dry}.

Determination of Sensible Heat Fluxes H, H_{dry}, H_{wet}

In order to derive the actual sensible heat flux H, the similarity theory is used. In SEBS, distinction is made between the ABL or PBL and the Atmospheric Surface Layer (ASL) similarity. ABL refers to the part of atmosphere that is directly influenced by the presence of the Earth's surface and responds to the surface forcings with a timescale of an hour or less, while ASL refers to usually the bottom 10% of ABL but above the roughness sublayer, that is, the ASL is where turbulent fluxes and stress vary by less than 10% of their magnitude (Stull, 1988; *see also* **Chapter 29, Atmospheric Boundary-Layer Climates and Interactions with the Land Surface, Volume 1**). The roughness sublayer (or the interfacial layer) is the near-surface thin layer of a few centimeters where the molecular transport dominates over turbulent transport (see Figure 2). The thickness of the roughness sublayer is thought to be around 35 times the surface roughness height, or three times the vegetation height (Katul and Parlange, 1992). In ASL, the similarity relationships for the profiles of the mean wind speed, u, and the mean temperature, $\theta_0 - \theta_a$, can be written in integral form as

$$u = \frac{u_*}{k} \left[\ln\left(\frac{z - d_0}{z_{0m}}\right) - \Psi_m\left(\frac{z - d_0}{L}\right) + \Psi_m\left(\frac{z_{0m}}{L}\right) \right] \tag{9}$$

$$\theta_0 - \theta_a = \frac{H}{k u_* \rho C_p} \left[\ln\left(\frac{z - d_0}{z_{0h}}\right) - \Psi_h\left(\frac{z - d_0}{L}\right) + \Psi_h\left(\frac{z_{0h}}{L}\right) \right] \tag{10}$$

where z is the height above the surface, $u_* = (\tau_0/\rho)^{1/2}$ is the friction velocity, τ_0 is the surface shear stress, ρ is the density of air, $k = 0.4$ is the von Karman constant, d_0 is the zero-plane displacement height, z_{0m} is the roughness height for momentum transfer, θ_0 is the potential temperature at the surface, θ_a is the potential air temperature at height z, z_{0h} is the scalar roughness height for heat transfer, ψ_m and ψ_h are the stability correction functions for momentum and sensible heat transfer respectively, L is the Obukhov length defined as

$$L = -\frac{\rho C_p u_*^3 \theta_v}{k g H} \tag{11}$$

where g is the acceleration due to gravity, θ_v is the potential virtual temperature near the surface.

For field measurements performed at a height of a few meters above ground, clearly, since the surface fluxes are related to surface variables and variables in the atmospheric surface layer, all calculations use the MOS stability functions (e.g. Brutsaert, 1999). By replacing the MOS stability functions with the BAS functions proposed by Brutsaert (1999), the system of equations (9–11) can be used to relate surface fluxes to surface variables and the mixed layer atmospheric variables. The criterion proposed by Brutsaert (1999) is used to determine if the MOS or the BAS scaling is appropriate for a given situation. The above functions are valid for unstable conditions only. For stable conditions, the expressions proposed by Beljaars and Holtslag (1991) and evaluated by Van den Hurk and Holtslag (1995) are used for atmospheric surface layer scaling, and the functions proposed by Brutsaert (1982, p. 84) for ABL scaling.

The friction velocity, the sensible heat flux, and the Obukhov stability length are obtained by solving the system of nonlinear equations (9–11) using the method of Broyden (Press *et al.*, 1997). It is important to note that the derivation of the sensible heat flux using equations (9–11) requires only the wind speed and temperature at the reference height as well as the surface temperature, and is independent of other surface energy balance terms.

As stated previously, this derived actual sensible heat flux H is further subjected to constraints in the range set by the sensible heat flux at the wet limit H_{wet}, and the sensible heat flux at the dry limit H_{dry} in SEBS.

The sensible heat flux at the dry limit H_{dry} is given by equation (4) and the sensible heat flux at the wet limit H_{wet} can be derived by combining equation (5) and a combination equation similar to the Penman–Monteith combination equation (Monteith, 1965). Menenti (1984) showed, when the resistance terms are grouped into the bulk internal (or surface) and external (or aerodynamic) resistances, the combination equation can be written in the following form,

$$\lambda E = \frac{\Delta \cdot r_e \cdot (R_n - G_0) + \rho C_p \cdot (e_{\mathrm{sat}} - e)}{r_e \cdot (\gamma + \Delta) + \gamma \cdot r_i} \tag{12}$$

where e and e_{sat} are actual and saturation vapor pressure, respectively; γ is the psychrometric constant, and Δ is the rate of change of saturation vapor pressure with temperature (i.e. $\partial e_{\mathrm{sat}}(T)/\partial T$); r_i is the bulk surface internal resistance, and r_e is the external or aerodynamic resistance. In the above equation, it is assumed that the roughness lengths for heat and vapor transfer are the same (Brutsaert, 1982). It is worthwhile to point out that the Penman–Monteith equation is strictly only valid for vegetation canopy, whereas the definition by means of equation (12) is also valid for soil surface with properly defined bulk internal resistance. The difficulty in using equation (12) to estimate latent heat flux lies in the difficulty to determine the bulk internal resistance r_i, which is strongly regulated by soil water availability. Because the latter is generally not known *a priori*, an alternative is thus proposed in SEBS to avoid the direct use of r_i in estimating λE.

At the wet limit, the internal resistance $r_i \equiv 0$ by definition. Using this property in equation (12) and changing the subscripts correspondingly to reflect the wet-limit condition, the sensible heat flux at the wet limit is obtained as

$$H_{\mathrm{wet}} = \left((R_n - G_0) - \frac{\rho C_p}{r_{ew}} \cdot \frac{e_s - e}{\gamma}\right) \Big/ \left(1 + \frac{\Delta}{\gamma}\right) \tag{13}$$

The external resistance depends also on the Obukhov length, L, which in turn is a function of the friction velocity and sensible heat flux (equation 11). With the friction velocity and the Obukhov length determined by the numerical procedure described previously, the external resistance can be determined from equation (10) as

$$r_e = \frac{1}{k u_*}\left[\ln\left(\frac{z - d_0}{z_{0h}}\right) - \Psi_h\left(\frac{z - d_0}{L}\right) + \Psi_h\left(\frac{z_{0h}}{L}\right)\right] \tag{14}$$

Similarly, the external resistance at the wet limit can be derived as

$$r_{ew} = \frac{1}{k u_*}\left[\ln\left(\frac{z - d_0}{z_{0h}}\right) - \Psi_h\left(\frac{z - d_0}{L_w}\right) + \Psi_h\left(\frac{z_{0h}}{L_w}\right)\right] \tag{15}$$

and the wet-limit stability length can be determined as

$$L_w = -\frac{\rho u_*^3}{k g \cdot 0.61 \cdot (R_n - G_0)/\lambda} \tag{16}$$

Determination of the Roughness Length for Heat Transfer

In the above derivations, the aerodynamic and thermal dynamic roughness parameters need to be known first. When near-surface wind speed and vegetation parameters (height and leaf area index) are available, the within-canopy turbulence model proposed by Massman (1997) can be used to estimate aerodynamic parameters, d_0, the displacement height, and z_{0m}, the roughness height for momentum. This model has been shown by Su *et al.* (2001) to produce reliable estimates of the aerodynamic parameters. If only the height of the vegetation is available, the relationships proposed by Brutsaert (1982) can be used. If a detailed land-use classification is available, the tabulated values of Wieringa (1993) can be used. However, since the aerodynamic parameters depend also on wind speed and wind direction besides the surface characteristics, the latter two approaches should only be used when the first method

cannot be used due to lack of data. When all of the above information is not available or not convenient to use, the aerodynamic parameters can be related to vegetation indices derived from satellite data. However, in this case, care must be taken, because the vegetation indices saturate at higher vegetation densities and the relationships are vegetation-type dependent.

The scalar roughness height for heat transfer, z_{0h}, changes with surface characteristics, atmospheric flow, and thermal dynamic state of the surface (Blümel, 1999; Massman, 1999a). Based on the work of Massman (1999a), a simple roughness model for heat transfer was proposed by Su *et al.* (2001). However, in their model, a functional form to describe the vertical structure of the vegetation canopy is needed in order to calculate the within-canopy wind speed profile extinction coefficient, n_{ec}. For local studies, this information is easily obtained, but for large-scale applications, it is generally impossible to obtain detailed information on the vertical structure of the canopy. In SEBS, n_{ec}, is formulated as a function of the cumulative leaf drag area at the canopy top,

$$n_{ec} = \frac{C_d \cdot LAI}{2u_*^2/u(h)^2} \tag{17}$$

where C_d is the drag coefficient of the foliage elements assumed to take the value of 0.2, LAI is the one-sided leaf area index defined for the total ground area, and $u(h)$ is the horizontal wind speed at the canopy top. The scalar roughness height for heat transfer, z_{0h}, can be derived from

$$z_{0h} = \frac{z_{0m}}{\exp(kB^{-1})} \tag{18}$$

where B^{-1} is the inverse Stanton number, a dimensionless heat transfer coefficient. To estimate the kB^{-1} value, an extended model of Su *et al.* (2001) is proposed as follows,

$$kB^{-1} = \frac{kC_d}{4C_t \frac{u_*}{u(h)}(1 - e^{-n_{ec}/2})} f_c^2 + 2f_c f_s \frac{k \cdot u_*/u(h) \cdot z_{0m}/h}{C_t^*} + kB_s^{-1} f_s^2 \tag{19}$$

where f_c is the fractional canopy coverage and f_s is its compliment. C_t is the heat transfer coefficient of the leaf. For most canopies and environmental conditions, C_t is bounded as $0.005N \le sC_t \le 0.075N$ (N is number of sides of a leaf to participate in heat exchange). The heat transfer coefficient of the soil is given by $C_t^* = \mathrm{Pr}^{-2/3}\,\mathrm{Re}_*^{-1/2}$, where Pr is the Prandtl number, the roughness Reynolds number $\mathrm{Re}_* = h_s u_*/\nu$, with h_s the roughness height of the soil. The kinematic viscosity of the air is given by $\nu = 1.327 \cdot 10^{-5}(p_0/p)(T_a/T_{a0})^{1.81}$(Massman, 1999b), with p and T_a the ambient pressure and temperature, and $p_0 = 101.3\,k\mathrm{Pa}$ and $T_{a0} = 273.15\,\mathrm{K}$. Physically and geometrically, the first term of equation (19) follows the full canopy-only model of Choudhury and Monteith (1988), the third term is that of Brutsaert (1982) for a bare soil surface, while the second term describes the interaction between vegetation and bare soil surface. A quadratic weighting based on the fractional canopy coverage is used to accommodate any situation between the full vegetation and bare soil conditions. For bare soil surface, kB_s^{-1} is calculated according to Brutsaert (1982)

$$kB_s^{-1} = 2.46(\mathrm{Re}_*)^{1/4} - \ln[7.4] \tag{20}$$

The MOS Stability Correction Functions

The MOS stability correction functions for momentum and sensible heat transfer ψ_m and ψ_h, respectively, are defined in the following integrated form

$$\Psi_i(y) = \int_0^y [1 - \phi_i(x)]\frac{\mathrm{d}x}{x} \tag{21}$$

where $y = -(z-d)/L \cdot i$ equals m, or h for momentum and sensible heat transfer, respectively.

The ϕ_i functions are proposed by Brutsaert (1999) as

$$\phi_m(y) = \frac{\left(a + b \cdot y^{m+\frac{1}{3}}\right)}{a + y^m} \tag{22}$$

$$\phi_h(y) = \frac{(c + d \cdot y^n)}{c + y^n} \tag{23}$$

On the basis of data reported by Högström (1988) and Kader and Yaglom (1990), Brutsaert (1999) assigned the constants in equations (22–23) as $a = 0.33$, $b = 0.41$, $m = 1.0$, $c = 0.33$, $d = 0.057$, and $n = 0.78$.

Upon integration of equations (22) and (23) using equation (21), the required MOS stability functions for free convective conditions are obtained.

$$\Psi_m(y) = \ln(a + y) - 3 \cdot b \cdot y^{1/3} + \frac{b \cdot a^{1/3}}{2} \ln\left[\frac{(1+x)^2}{(1 - x + x^2)}\right] + 3^{1/2} \cdot b \cdot a^{1/3} \tan^{-1}\left[\frac{(2 \cdot x - 1)}{3^{1/2}}\right] + \Psi_0, \quad \text{for} \quad y \le b^{-3} \tag{24a}$$

$$\Psi_m(y) = \Psi_m(b^{-3}), \quad \text{for} \quad y > b^{-3} \tag{24b}$$

$$\Psi_h(y) = \left[\frac{(1-d)}{n}\right] \ln\left[\frac{(c + y^n)}{c}\right] \tag{25}$$

where $x = (y/a)^{1/3} \cdot \Psi_0 = (-\ln a + 3^{1/2} \cdot b \cdot a^{1/3} \cdot \pi/6)$ is an integration constant.

Equations (24–25) are extensions to the Businger–Dyer function for unstable conditions. For stable conditions, the expressions proposed by Beljaars and Holtslag (1991) and evaluated by Van den Hurk and Holtslag (1995) can be used. These are given below:

$$\Psi_m(y_s) = -\left[a_s \cdot y_s + b_s\left(y_s - \frac{c_s}{d_s}\right) \cdot \exp(-d_s \cdot y_s) + \frac{b_s \cdot c_s}{d_s}\right] \quad (26)$$

$$\Psi_h(y_s) = -\left[\left(1 + \frac{2a_s}{3}y_s\right)^{1.5} + b_s \cdot \left(y_s - \frac{c_s}{d_s}\right) \cdot \exp(-d_s \cdot y_s) + \left(\frac{b_s \cdot c_s}{d_s} - 1\right)\right] \quad (27)$$

where $y_s = (z-d)/L$, $a_s = 1$, $b_s = 0.667$, $c_s = 5$ and $d_s = 1$.

The Bulk Atmospheric Boundary Layer (ABL) Similarity (BAS) Stability Correction Functions

Under free convective conditions, the outer region of the ABL is well mixed such that the mean profiles of wind and potential temperature are nearly constant with height. This is equivalent to state that the unstable ABL consists of two regions, an inner region where MOS is valid, and a slab outer region where the profiles are constant.

According to Brutsaert (1999), experimental evidence suggests that the height of the ASL, h_{st}, should be scaled with the thickness of the ABL over moderately rough surfaces, but with the surface roughness over very rough terrain. More quantitatively, h_{st}, can be determined in the following ways

$$h_{st} = \alpha_b \cdot h_i \quad (28)$$

where h_i is the height of the ABL and α_b is around 0.10–0.15, or

$$h_{st} = \beta_b \cdot z_0 \quad (29)$$

where β_b is around 100–150
whichever is larger. Setting typical values of $\beta_b/\alpha_b = 10^3$, and $h_i = 10^3\,m$, gives $z_0 = (\alpha_b/\beta_b) \cdot h_i = 1\,m$, which separates very rough from moderate rough terrain.

For moderately rough terrain satisfying $z_0 < (\alpha_b/\beta_b) \cdot h_i$, joining the inner with the outer region such that $u(z-d_0) = u_m$, $\theta_a(z-d_0) = \theta_m$, at $z - d_0 = h_{st}$, (where the subscript m indicates average of the variable in question over the lower half of the mixed layer, which is an advise for practical applications because the entrainment of warmer air into the ABL affects both the potential temperature and the wind profiles in the upper half of the ABL), allows to derive the bulk stability functions

$$B_w = -\ln(\alpha_b) + \Psi_m\left(\frac{\alpha_b \cdot h_i}{L}\right) - \Psi_m\left(\frac{z_0}{L}\right) \quad (30)$$

$$C_w = -\ln(\alpha_b) + \Psi_h\left(\frac{\alpha_b \cdot h_i}{L}\right) - \Psi_h\left(\frac{z_{0h}}{L}\right) \quad (31)$$

Equations (30–31) show that the bulk stability functions depend on both the surface roughness and the height of the ABL for moderately rough terrain.

Similarly, for very rough terrain, that is, $z_0 \geq (\alpha_b/\beta_b) \cdot h_i$, we obtain

$$B_w = -\ln\left(\frac{h_i}{(\beta_b \cdot z_0)}\right) + \Psi_m\left(\frac{\beta_b \cdot z_0}{L}\right) - \Psi_m\left(\frac{z_0}{L}\right) \quad (32)$$

$$C_w = -\ln\left(\frac{h_i}{(\beta_b \cdot z_0)}\right) + \Psi_h\left(\frac{\beta_b \cdot z_0}{L}\right) - \Psi_h\left(\frac{z_{0h}}{L}\right) \quad (33)$$

which state that the bulk stability functions depend on only the surface roughness for very rough terrain.

Finally, for stable conditions, that is, when $h_i/L > 0$, we use (Brutsaert, 1982, equation 4.93, p. 84)

$$B_w = -2.2 \cdot \ln\left(1 + \frac{h_i}{L}\right) \quad (34)$$

$$C_w = -7.6 \cdot \ln\left(1 + \frac{h_i}{L}\right) \quad (35)$$

Both stability correction functions may need to be verified using data from more recent large-scale field experiment and updated accordingly when necessary.

Determination of Turbulent Heat Fluxes and Actual Evaporation

By inverting equation (8), the actual sensible and latent heat fluxes can be obtained as,

$$H = (1-\Lambda) \cdot (R_n - G)$$
$$\lambda E = \Lambda \cdot (R_n - G) \quad (36)$$

When the evaporative fraction is known, the daily evaporation can be determined as

$$E_{\text{daily}} = 8.64 \times 10^7 \times \bigwedge_0^{24} \times \frac{\overline{R_n} - \overline{G_0}}{\lambda \rho_W} \quad (37)$$

where E_{daily} is the actual evaporation on daily basis $(mm \cdot d^{-1}) \cdot \bigwedge_0^{24}$ is the daily average evaporative fraction, which can be approximated by the SEBS estimate since

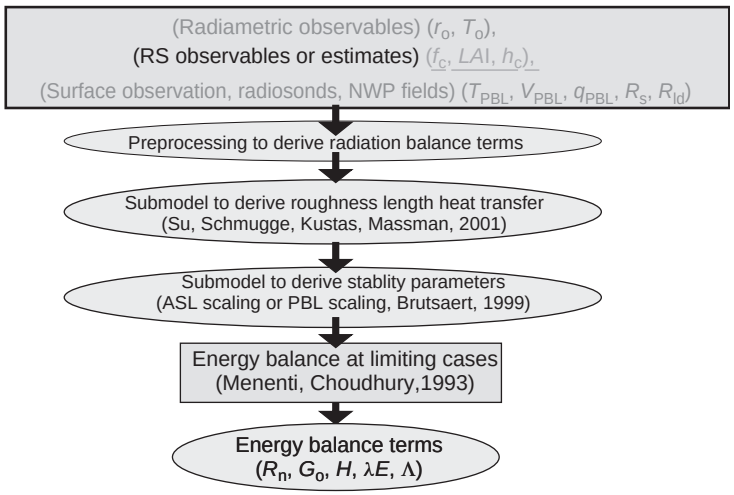


Figure 3 Schematic representation of SEBS. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the evaporative fraction is conservative (Shuttleworth *et al.*, 1989; Sugita and Brutsaert, 1991; Crago, 1996). $\overline{R_n}$ and $\overline{G_0}$ are the daily net radiation flux and soil heat flux, λ is the latent heat of vaporization ($J\,Kg^{-1}$), ρ_w is the density of water ($Kg\,m^{-1}$).

Since the daily soil heat flux is close to zero because the downward flux in daytime and the upward flux at night balance each other approximately, the daily evaporation only depends on the net radiation flux given by

$$\overline{R_n} = (1-\alpha)K_{24}^{\downarrow} + \varepsilon L_{24} \tag{38}$$

where $K_{24}^{\downarrow}$ is the daily incoming global radiation (*see* **Chapter 49, Estimation of Surface Insolation, Volume 2**) and L_{24} is daily net longwave radiation. The daily average albedo, α, and emissivity, ε, can be approximated easily with the same values as used previously in the energy balance equation.

By summing up the corresponding daily evaporation for a certain period, the actual evaporation for that period (i.e., a week, a month, or a year) can be determined. However, errors will occur due to cloud effects. Such effects can be further removed by using the time series processing or by data assimilation procedures. A schematic presentation of SEBS is given in Figure 3 for clarity of the steps involved.

SOME ISSUES IN THE APPLICATION OF SEBS

Difficulty in Determination of Aerodynamic Roughness Height

Aerodynamic roughness height (the displacement height is often related to it) influences greatly the turbulent characteristics near the surface where the heat fluxes originate. Currently, there are several methods that can be used for its determination as shown in Figure 4, including retrievals from wind profiles, which is probably the most accurate method but is limited to the local topographic and canopy structure and varies with wind speed (different fetch) and direction (for heterogeneous terrain). Other methods are based on either vegetation height or land-use

Table 1 Land-use classes in the PELCOM land-use database and associated z_{0m} values

	Land-use class	$z_{0m}(m)$
1	Grass	0.0340
2	Maize	0.4966
3	Potatoes	0.0639
4	Beets	0.0639
5	Cereals	0.4966
6	Other crops	0.0639
7	Greenhouses	0.4066
8	Orchards	0.6065
9	Bulbs	0.0639
10	Deciduous forest	1.2214
11	Coniferous forest	1.2214
12	Heather	0.0408
13	Other open spaces in natural areas	0.0408
14	Bare soil in natural areas	0.0012
15	Freshwater	0.0002
16	Salt water	0.0002
17	Continuous urban area	1.1052
18	Built-up area in rural area	0.5488
19	Deciduous forest in urban area	1.2214
20	Coniferous forest in urban area	1.2214
21	Built-up area with dense forest	1.2214
22	Grass in built-up area	0.0334
23	Bare soil in built-up area	0.0012
24	Main roads and railways	0.0035
25	Buildings in rural areas	0.5488

· Retrievals from wind profiles (Point values)

· z_{0m} = 0.136 *h (Vegetation height)

· Land use + look up table (z_{0m} per class) (e.g. in atmospheric models)

· z_{0m}~ Vegetation index relationships (e.g. z_{0m} = exp(A+B*NDVI))

· Land use + modelling (Hasager, Jensen, 1999)

· LIDAR measurements (e.g. Menenti, Ritchie, 1994)

· Vegetation canopy LIDAR ???

Figure 4 Summary of some methods used to determine the aerodynamic roughness height. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

classes, and assigning each class a nominal value, or using a relationship with vegetation index. One particular promising method is that of Hasager and Jensen (1999) that is able to consider the dynamic flow characteristics and actual land-use characteristics. Menenti and Ritchie (1994) showed that airborne LIDAR measurements of vegetation heights provided reliable determination of the aerodynamic roughness, so that a space-borne LIDAR system when available might be quite promising. An example for the aerodynamic roughness height for each classes of the PELCOM land-use map (Pan-European Land-Use and Land-Cover Monitoring, Mücher *et al.*, 2001) is given in Table 1. Table 2 shows the dependency of the aerodynamic roughness height on wind direction (different land surfaces) and season (summer or winter), and shows the uncertainties that might be encountered when using any of the methods indicated in Figure 4.

At any rate, despite great efforts in the past, the determination of aerodynamic roughness remains a challenging issue for large-scale applications both for remote sensing of surface turbulent heat fluxes and mesoscale and global-scale atmospheric modeling. Figure 5 shows a map of z_{0m} values for the Netherlands with values corresponding to Table 1.

Coupling with Large-Scale Numerical Weather Prediction Models

Owing to its flexibility in scaling between surface and atmospheric variables, SEBS can be easily coupled to the meteorological fields generated by a large-scale atmospheric model. A first attempt has been made by Jia *et al.* (2003) to couple model fields from a NWP model to radiometric measurements from the ATSR onboard the European Remote

Table 2 Effective roughness length (m), representative of a few kilometers of upstream terrain from the Cabauw tower in the Netherlands for 18 wind direction classes (degrees relative to North, after Beljaars and Bosveld, 1997)

Wind direction	0–20	20–40	40–60	60–80	80–100	100–120	120–140	140–160	160–180
Summer	0.06	0.08	0.10	0.15	0.15	0.15	0.11	0.08	0.04
Winter	0.04	0.05	0.05	0.07	0.10	0.12	0.02	0.02	0.02
Wind direction	180–200	200–220	220–240	240–260	260–280	280–300	300–320	320–340	340–360
Summer	0.04	0.04	0.04	0.07	0.06	0.06	0.06	0.05	0.05
Winter	0.03	0.03	0.02	0.04	0.03	0.03	0.04	0.04	0.03

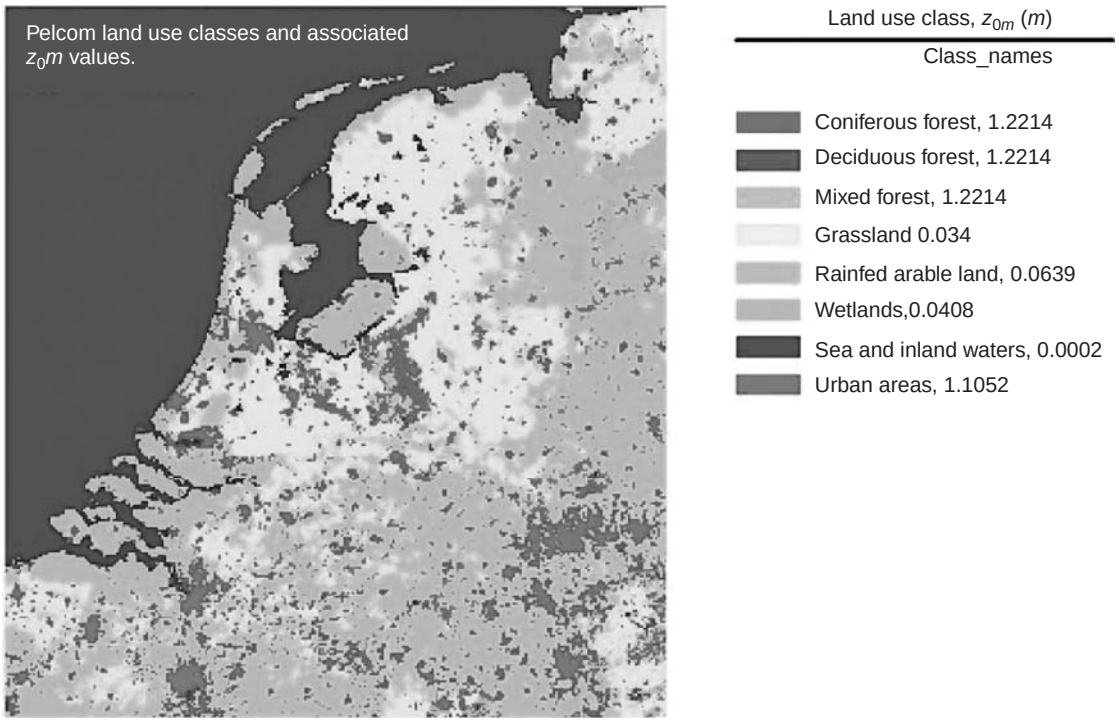


Figure 5 A map of aerodynamic roughness height for the Netherlands based on land-use classes. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Sensing Satellite (ERS-2). In order to validate the effectiveness of the approach, three Large Aperture Scintillometers (LAS) providing line-averaged measurements of sensible heat flux over path lengths from 1 km to 5.2 km are used. The instruments were located far apart at three locations in Spain over different combinations of land cover and hydrological conditions and were operated continuously from April through September 1999. The LAS were installed at sites near Lleida, Badajoz, and Tomelloso (Figure 6). Each scintillometer site is additionally equipped with a shortwave radiometer to measure global radiation. Additional data that are necessary for the processing of LAS data were obtained from weather stations in the near surroundings.

The scintillometer near Lleida is installed in a region with small-scale irrigation at an altitude of 130 m. Main crops are fruit trees (peaches) and alfalfa. The distance between transmitter and receiver is approximately 4690 m. Both receiver and transmitter are installed on a hill, on a tripod at a height of, respectively, 45 m and 39 m above the surface. Meteorological data are obtained from the meteorological station near Juneda (41°33'N 0°49.5'E).

The Badajoz LAS is located in a region with large-scale irrigation (sprinkler irrigation) at an altitude of 168 m. Crops being grown are wheat, corn, alfalfa, lettuce, olives, beans, and tomatoes. The distance between transmitter and receiver is 5250 m ± 200 m. The receiver is installed on top of a hill on a house, 68 m above the surrounding terrain. The transmitter is installed on a water tower, 56 m above the surrounding terrain. Meteorological data are obtained from a station with short grass ground cover some 7 km from the LAS receiver.

The landscape of the Tomelloso site comprises a nearly level, alluvial floodplain to the north and a tilted old alluvial plain to the south. The vegetation consists of nonirrigated vineyards and a few olive orchards. The vines are broadly spaced and trellises are not used. The distance between transmitter and receiver is 1070 m ± 40 m. Both receiver and transmitter are installed on a steel mast at a height of, respectively, 4.15 m and 4.56 m above the surface at an altitude of 670 m. Meteorological data are obtained from a meteorological station near Tomelloso (39°10'29.22 "N, 3°0'2.16"W). Ground cover of the station is bare soil and grass. Longwave and shortwave radiation data, both incoming and outgoing, are available from this meteorological station.

Sensible heat flux H is calculated using SEBS at the ATSR passing time for corresponding days at each experimental site. The meteorological fields of wind, potential temperature, and humidity of air were generated by an advanced NWP model, the Regional Atmospheric Climate Model (RACMO) (Christensen *et al.*, 1996), integrated over the PBL. This implies that a single value of the reference wind, potential temperature, and humidity of air is used for each model grid, that is, 25 km × 25 km, while

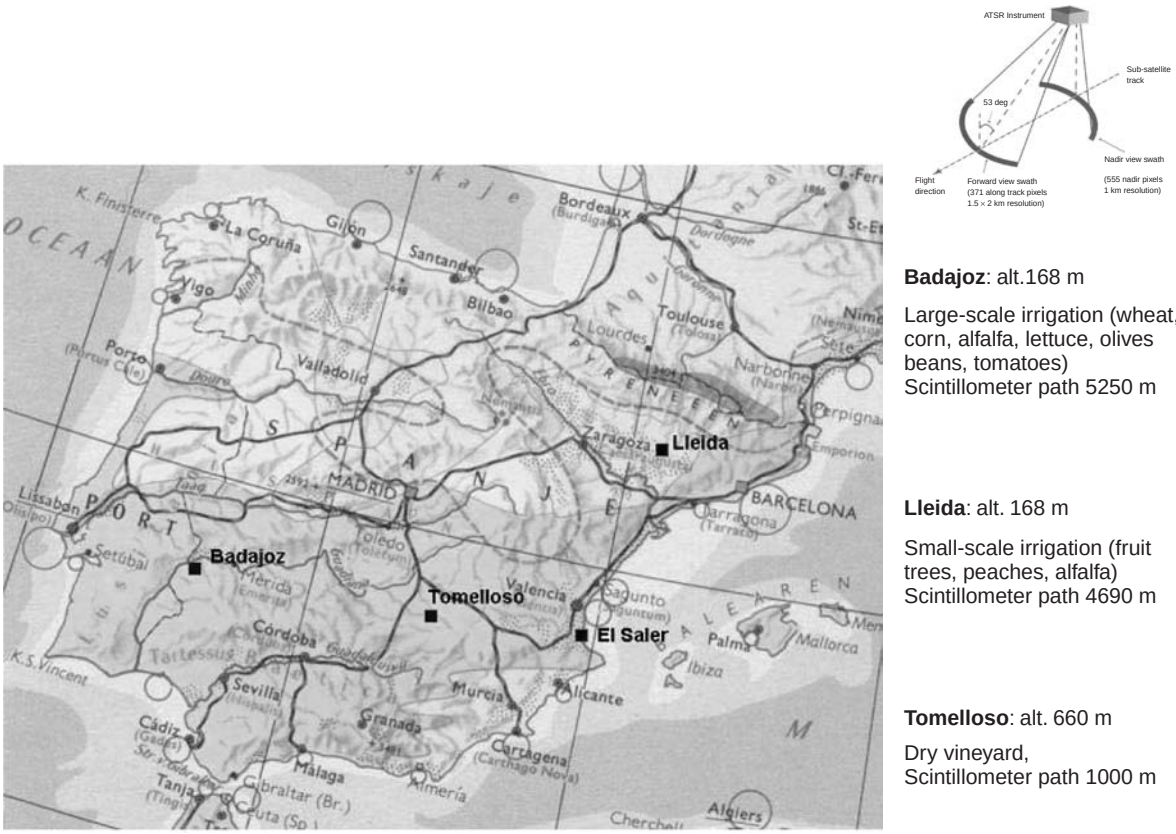


Figure 6 Location of the three study sites in Spain (ATSR data plus meteorological data from nearest stations) and their characteristics. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the observations of vegetation properties have a resolution of 1 km × 1 km and account for sub-grid variability of land cover and hydrological conditions. The path lengths of the LAS measurements correspond to about five, six, and two ATSR pixels for the Lleida, Badajoz, and Tomelloso sites, respectively. The results from SEBS are quite comparable with the measurements as shown in Figure 7. In judging the validation results, it is important to keep the following issues in mind. First, the footprints in LAS measurements and in SEBS estimates are different. While the LAS-measured sensible heat flux is a line average along the path that only captures the contributions of turbulence from upwind direction over the averaging period, the SEBS estimates of sensible heat flux by using ATSR-measured surface variables and the NWP model-derived PBL variables are the average over a larger area, at least at ATSR pixel scale. Second, different temporal scales certainly contribute to the differences for the observed and estimated fluxes. Measurements from LAS are the mean values over 10 minutes, which represent the mixed turbulent characteristics and heat-exchange processes over the observation time, while a satellite only provides the instantaneous observations of surface variables at the satellite passing time. Moreover, if the surface shows large degree

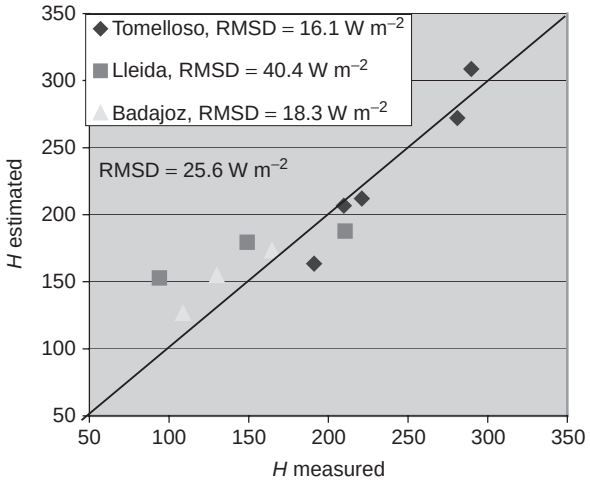


Figure 7 Comparison of sensible heat flux values between SEBS and LAS measurements at the three experimental sites. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of heterogeneity, the determination of adequate pixels used to compare with the LAS-measured flux should be dealt with carefully.

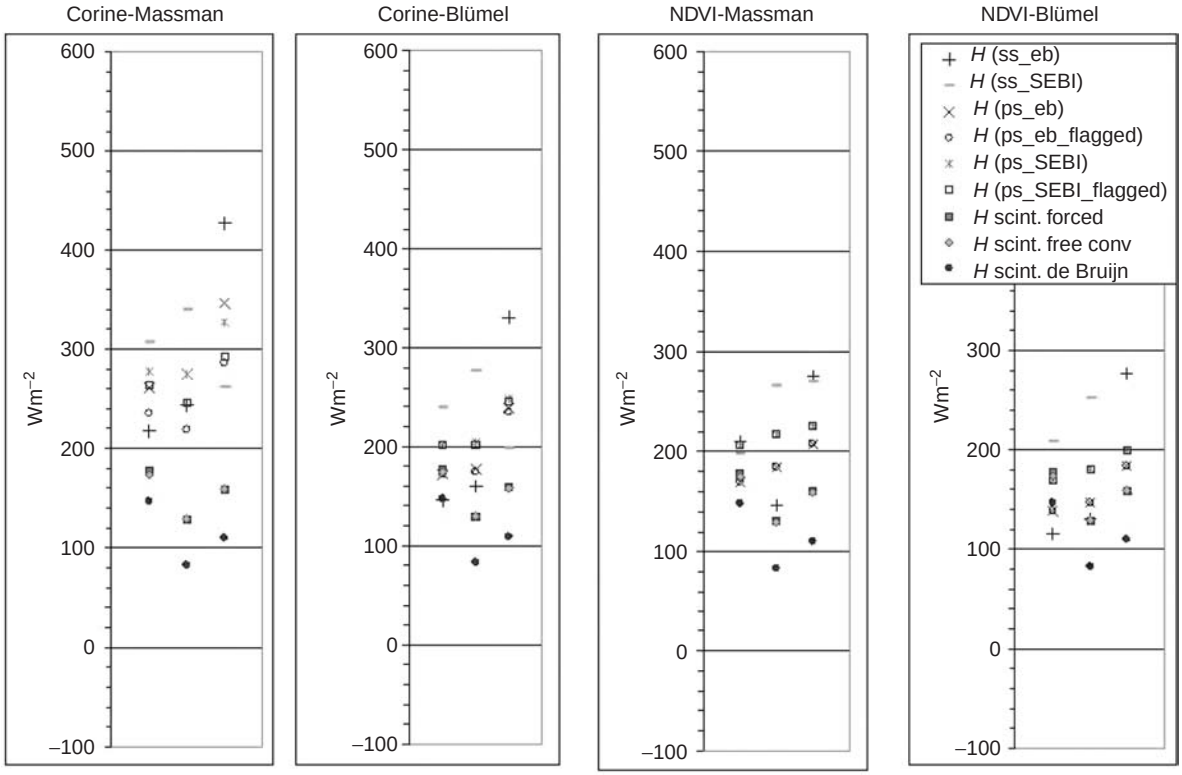


Figure 8 Results derived from both single-source and parallel-source SEBS for Badajoz site (Corine refers to aerodynamic roughness derived from Corine land use map; Massman refers to the thermal dynamic roughness model of Su *et al.*, 2001; Blümel refers to the thermal dynamic roughness model of Blümel (1999); NDVI refers to aerodynamic roughness derived with NDVI – Normalized Difference Vegetation Index; ss – single source; ps – parallel source; eb – sensible heat flux derived from the nonlinear equations; sebi – sensible heat flux derived on the basis of the SEBI concept; flagged – energy limits are applied, that is, SEBS; scint, Forced. scint, Free convection. scint, de Bruin refer to three different ways of estimation of sensible heat fluxes from scintillometer measurements). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Further details about the measurements can be found in Moene and De Bruin (2001); Moene (2001). For details on the scintillometer method, the reader is referred to De Bruin (2002), and more details of the comparison can be found in Jia *et al.* (2003).

Single-source SEBS versus Parallel-source SEBS

Essentially, the parameterization in single-source SEBS and parallel-source SEBS (see Figure 1 for details) are the same except in the case of the parallel-source SEBS, we first set the fractional cover of vegetation as unity and calculate the energy balance and turbulent heat-flux terms, then we do the same for soil by setting the fractional cover of soil as one, and finally we sum up the fluxes terms weighted by their actual fractional coverage.

Despite the simplicity of such a scheme, the results from the parallel-source SEBS are remarkably improved over the single-source SEBS as shown by Rauwerda *et al.* (2002) for three sites in Spain discussed above when the estimated turbulent heat fluxes are compared to scintillometer measurements (see Figures 8, 9 and 10) . The uncertainties caused by the different methods in determination of the roughness height for momentum are also amply demonstrated. It is also helpful to notice the differences in the LAS measurements derived from different processing methods. Further details on the comparison between single-source and parallel-source SEBS and ground measurements are given in Rauwerda *et al.* (2002).

Some SEBS Applications to Water Sources Management

Besides the results presented above, SEBS has been used for several studies for water resources management conducted recently. Examples are estimation of actual evaporation in a semiarid inland basin in Northwest China (Li, 2001; Su *et al.*, 2003a) and drought disaster monitoring (Su *et al.*, 2003b).

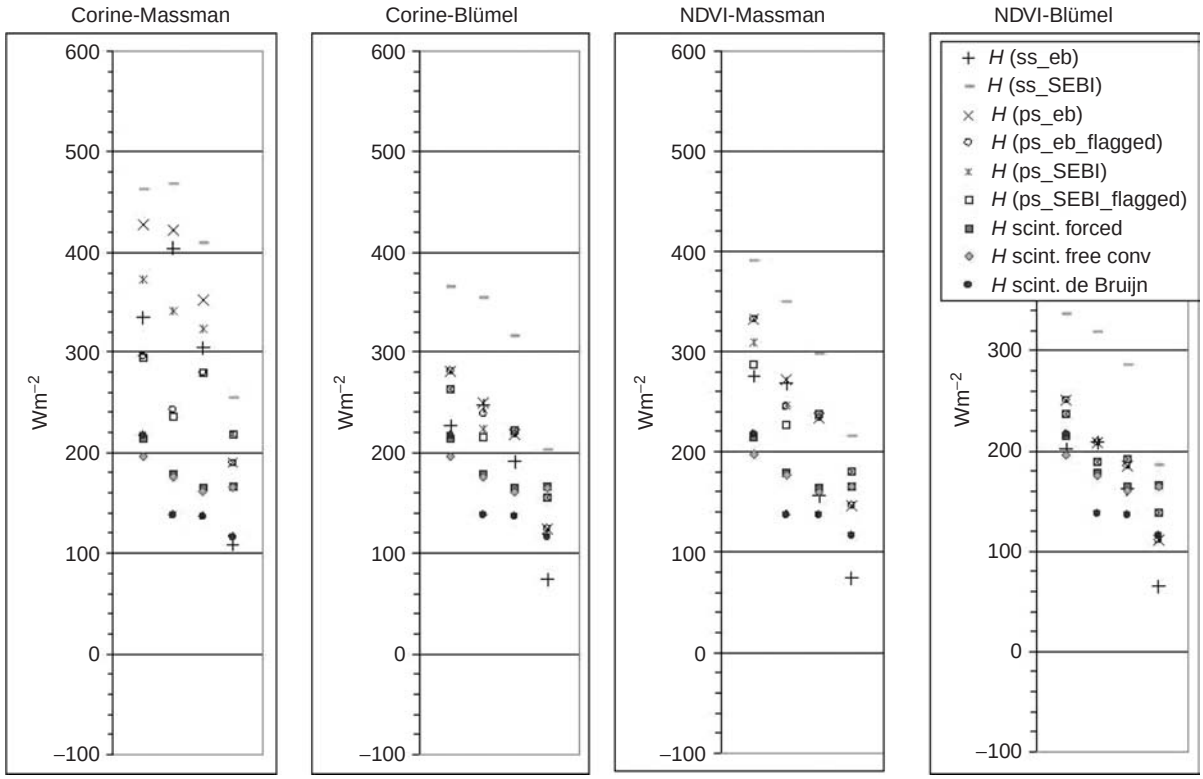


Figure 9 Results derived from both single-source and parallel-source SEBS for Lleida site (explanation of symbols is the same as for in Figure 8). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

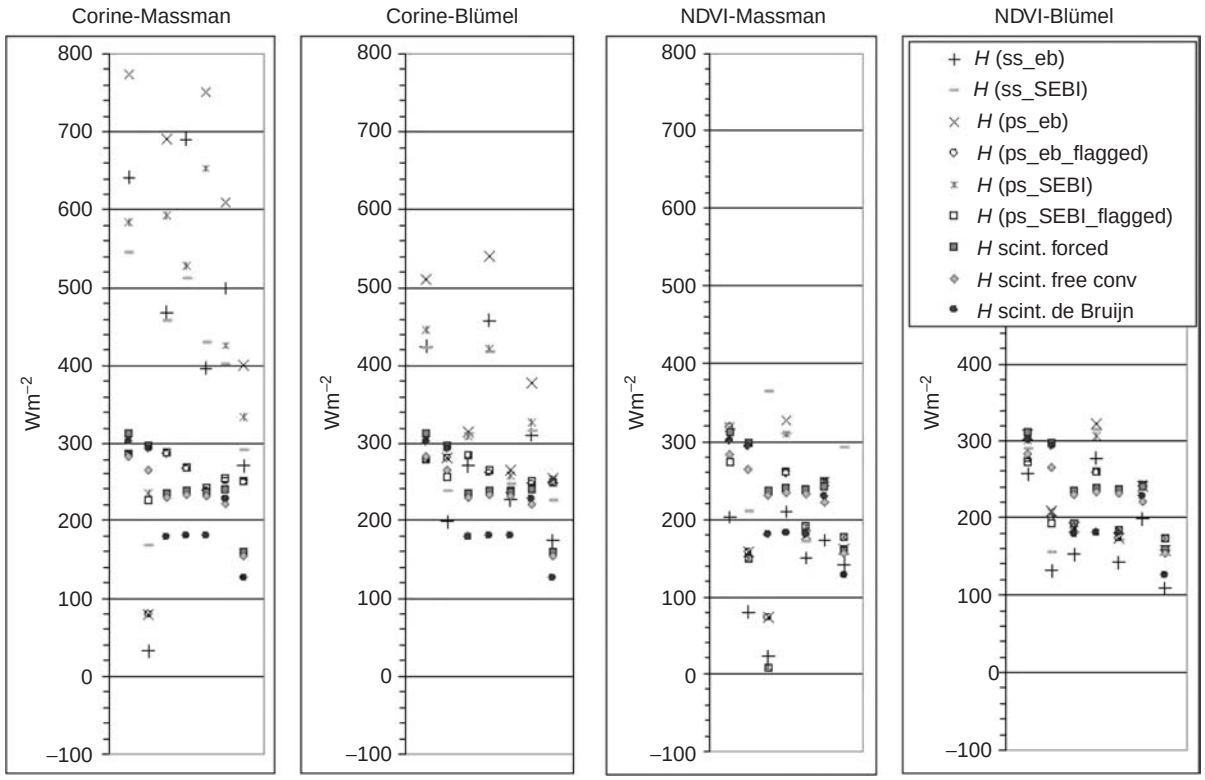


Figure 10 Results derived from both single-source and parallel-source SEBS for Tomelloso site (explanation of symbols is the same as for in Figure 8). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Estimation of Actual Evaporation in the Urumqi River Basin

SEBS is applied to Advanced Very High Resolution Radiometer (AVHRR) data for the estimation of actual evaporation in the Urumqi River Basin located in Xinjiang Uygar Autonomous Region in Northwest China. The Urumqi River Basin has all the features of a continental arid climate. Its diversity in topography offers a variety of land cover and vegetation. As such, the methodology and procedures developed for the Urumqi River Basin can be easily applied to other river basins in Northwest China and elsewhere in the world. Two water balance field experimental stations: Changji Station and Wulapo Station, equipped with routine meteorological instruments, evaporation pans of different sizes (E20, E60, and E5000), and lysimeters with different soils and vegetation types, provided sufficient information for calculation and verification. The size of the AVHRR data being processed for evaporation estimation was set between 86°38'E–88°59'E and 43°07'N-45°02'N, with a total area of about 27 800 km^2. Hundred-and-five scenes in 1995 covering the basin were downloaded from Satellite Active Archive, among which 12 scenes were carefully chosen for evaporation calculation. Meteorological information is derived from the Changji meteorological station including wind speed (m/s), relative humidity, potential air temperature (K), and surface air pressure (Pa), and is scaled to PBL variables using the logarithmic profiles.

Verifications of the derived evaporation are performed by comparing with evaporation from different types of land surface collected in intensive field surveys for water resources assessment of the Urumqi River Basin (Li, 2001). Comparison between the annual evaporation derived from SEBS against the field measurements shows that the estimated values agree very well with the ground data except for the midmountainous area. The reason for the large relative error for the mountainous area could be due to the input of average elevation during the calculation as well as the uncertainty associated with the field information. More details on this case study can be found in Li (2001).

Application of SEBS in Drought Monitoring

On the basis of SEBS, Su *et al.* (2003b) have proposed a method for assessing relative soil moisture with remote sensing data and have theoretically shown that the relative soil moisture in the rooting zone can be derived from relative evaporation. The relationship derived between the relative soil moisture and relative evaporation is confirmed with experimental data collected with lysimeter measurements in an eight-year-old orchard of peach trees (Girona *et al.*, 2002), and with measurements collected in an intensive field experiment in the Global Energy and Water Experiment (GEWEX) Asia Monsoon Experiment in Tibet (GAME/Tibet) in 1998 (Koike, 2000; Ma *et al.*, 2002). Further, it was shown that the proposed theory can be used to define a Drought Severity Index (DSI) for drought monitoring, with the relative evaporation determined with remote sensing data. A demonstration for the latter was performed for North China in 2000 using remote sensing data AVHRR and routine meteorological data obtained from the operational measurement network of the National Meteorological Center of China. In the growing season with low rainfall, the value of DSI was high, and in the reverse case, the DSI value was low, which demonstrates that the DSI is in accordance with the data of rainfall in time and can be used to quantify drought situation both in temporal and spatial scales. Spatially, the derived DSI could identify that the central, the Northern, and then the Southern part of Shanxi province suffer under severe drought conditions in decreasing drought intensity in 2000. Temporarily, low soil water deficit was identified in July and August in most parts of Shanxi province. Higher soil water deficit occurred in April, May, and September 2000 with decreasing degree of intensity. Comparisons between the estimated relative evaporation and the actual measurements of soil moisture confirmed the general validity and robustness of the proposed method, but more extensive validation with field measurements at regional scale is considered necessary to quantify the accuracy of the proposed DSI method.

In all applications, benefits when using radiometric measurements are clearly demonstrated.

Using SEBS in a Data Assimilation System Environment

A valuable first attempt in this direction has been made recently by Wood *et al.* (2003) who applied SEBS to the Southern Great Plains (SGP) region of the United States where the Atmospheric Radiation Measurement (ARM) program is carried out by the US Department of Energy. Energy Balance Bowen Ratio (EBBR) and Soil Water And Temperature System (SWATS) are also available at many ARM/Cloud And Radiation Testbed (CART) stations, and provide measurements of surface latent, sensible, and ground heat fluxes, as well as soil moisture and soil temperature. The surface meteorology used in SEBS is retrieved from the Eta Data Assimilation System (EDAS), while the solar radiation was processed by National Oceanic and Atmospheric Administration (NOAA/NESDIS) as part of the North American Land Data Assimilation System (NLDAS) (Mitchell *et al.*, 1999). The temperature product from the Moderate-Resolution Imaging Spectroradiometer (MODIS) on the Earth Observation Satellite (EOS) Terra was used to estimate the land surface temperature across the domain.

Measurements from the EBBR sites were compared with model-derived latent heat fluxes. For the 40 clear-sky days, an areal average instantaneous flux was calculated across the nine EBBR sites from the collocated SEBS pixels, with observations approximately 20% less than predictions.

Table 3 Statistics of estimated versus observed heat fluxes of the Cotton dataset of Kustas (1990) with different algorithms (Root mean-squared difference between estimated and measured fluxes (Wm^{-2}))

Method & reference	R_n	G_0	H	λE	No. of observations
Observed mean value ($W\,m^{-2}$)	561.74	140.37	116.63	304.74	19
RMSD for SEBS (Su, 2002, Table 1)	22.82	5.42	21.22	29.22	19
RMSD for Two-source model (Kustas and Norman, 1999, Table 3)	19	18	23	42	19
RMSD for Two-source model (Kustas and Norman, 1999, Table 4)	21	13	25	37	19
RMSD for Two-source model (Kustas and Norman, 1999, Table 5)	20	10	36	47	19

Such differences may be attributed to the uncertainties in SEBS, especially in the roughness parameterization as discussed previously, but may similarly be attributed to uncertainties in surface flux measurements, because comparisons among different measurement setups may easily produce discrepancies of up to 25% (e.g. Twine *et al.*, 2000). While there exists a systematic overprediction in the SEBS/MODIS estimates when compared to the EBBR data, the results are encouraging and indicate that the SEBS approach has promise in estimating surface heat flux from space for data assimilation purposes.

Results for SEBS and Other Techniques When Applied to Experimental Data Sets

In order to assess the accuracy of SEBS in comparison with some other techniques, results obtained with the same experimental data sets will be briefly discussed. Three field datasets obtained from flux stations that have been used extensively to benchmark surface energy balance algorithms (e.g. Norman *et al.*, 1995; Zhan *et al.*, 1996; Kustas and Norman, 1999; Su *et al.*, 2001; Su, 2002) will be used as examples. These three datasets are the Cotton dataset of Kustas (1990), the Shrub, and the Grass dataset of Kustas *et al.* (1994).

For the application of SEBS, the aerodynamic parameters are the model estimates as discussed in Su (2003). All other input variables are measured except the downward longwave radiation that is estimated with the Stefan-Bolzmann radiation equation with the measured air temperature at the reference height. The parameterization used in SEBS is meant to be site-independent and, therefore, there is no site-specific adjustment or optimization involved. The details of the application of the other techniques are given in Kustas and Norman (1999) and Zhan *et al.* (1996) with some different site-specific parameterizations involved. The latter has evaluated four different techniques using the same data sets.

Table 3 shows the statistics of estimated versus observed heat fluxes of the Cotton dataset of Kustas (1990) with SEBS and the two-source model of Norman *et al.* (1995) using three different parameterizations (Kustas and Norman, 1999). Given the observed mean values of the net radiation, soil heat flux, sensible heat flux, and latent heat flux, the uncertainties of the estimated values (here expressed as the ratio of the root mean squared difference (RMSD) between the estimated and the observed values) are all within 20% of the mean observed values, except the sensible heat flux for the last row, which reaches 30% of uncertainty. The higher uncertainty in the latter case was due to the use of explicit soil and vegetation component temperatures and might be considered as an indication that the used two-source model needs some modification in order to be able to take advantage of the availability of the component temperatures. It should be noticed that an uncertainty of 20% or less in the estimated turbulent heat fluxes represents the best case that can be achieved given the uncertainties in all the involved parameters and variables in the algorithms as well as the uncertainties in the measured turbulent heat fluxes. More details on the uncertainty analysis can be found in Su (2002).

Tables 4 and 5 display the statistics of estimated versus observed heat fluxes of the Shrub and the Grass dataset of Kustas *et al.* (1994) for SEBS and four different algorithms, including the two-source model of Norman *et al.* (1995). The accuracy of SEBS are comparable to the best case analyzed by Zhan *et al.* (1996), who only reported the statistics for the sensible heat fluxes estimated with the four different algorithms. It is worthwhile to point out that while in SEBS all the available data in the two datasets are used (320 and 281 observations, respectively), the application of the four other algorithms were restricted to well-developed unstable atmospheric surface layer by excluding the more problematic transitions periods and night-time observations (114 and 103 observations, respectively).

Table 4 Statistics of estimated versus observed heat fluxes of the Shrub dataset of Kustas *et al.* (1994) with different algorithms (Root mean-squared difference between estimated and measured fluxes ($W\ m^{-2}$))

Method (Reference)	R_n	G_0	H	λE	No. of observations
SEBS (Su, 2002, Table 2)	35.11	46.29	28.61	82.79	320
Single-source model of Kustas *et al.* (1989) (Zhan *et al.*, 1996, Table 6)	–	–	33	–	114
Single-source model of Troufleau *et al.* (1997) (Zhan *et al.*, 1996, Table 6)	–	–	48	–	114
Dual-source model of Lhomme *et al.* (1994) (Zhan *et al.*, 1996, Table 6)	–	–	52	–	114
Dual-source model of Norman *et al.* (1995) (Zhan *et al.*, 1996, Table 6)	–	–	33	–	114

Table 5 Statistics of estimated versus observed heat fluxes of the Grass dataset of Kustas *et al.* (1994) with different algorithms (Root mean-squared difference between estimated and measured fluxes ($W\ m^{-2}$))

Method (Reference)	R_n	G_0	H	λE	No. of observations
SEBS (Su, 2002, Table 3)	41.26	42.95	36.19	61.34	281
Single-source model of Kustas *et al.* (1989) (Zhan *et al.*, 1996, Table 7)	–	–	41	–	103
Single-source model of Troufleau *et al.* (1997) (Zhan *et al.*, 1996, Table 7)	–	–	33	–	103
Dual-source model of Lhomme *et al.* (1994) (Zhan *et al.*, 1996, Table 7)	–	–	124	–	103
Dual-source model of Norman *et al.* (1995) (Zhan *et al.*, 1996, Table 7)	–	–	34	–	103

From the above evaluations, it might be concluded that SEBS can be applied to different sites and different atmospheric stability situations while maintaining the same parameterizations. This might prove important and critical when application to large scale is desired where no sufficiently detailed information on the site characteristics or the local atmospheric stability regime is available.

DISCUSSION AND CONCLUSIONS

In this article, a brief introduction in estimation of radiation and turbulent heat fluxes using radiometric observations is given. In general, after the derivation of sensible heat flux, it is possible to estimate the latent heat flux as a residual by means of the energy balance equation, as was done in many other studies (e.g. Kalma and Jupp, 1990). However, the associated uncertainty in the derived latent heat flux and consequently in the evaporative fraction is large. This is because the sensible heat flux is, under given surface conditions, determined solely by the surface temperature and the meteorological conditions at the reference height and is not constrained by the available energy. If the surface temperature or the meteorological variables have large uncertainty, a straight propagation of such uncertainty into the resultant latent heat flux and evaporative fraction cannot be avoided. In the SEBS formulation, this uncertainty is limited by consideration of the energy balance at the limiting cases, because the actual sensible heat flux is constrained in the range set by the sensible heat flux at the wet limit derived from a combination equation, and the sensible heat flux at the dry limit set by the available energy. In this aspect, SEBS is similar to previous algorithms that estimate relative evaporation by means of an index (e.g. the CWSI) using a combination equation (Jackson *et al.*, 1981, 1988; Menenti and Choudhury, 1993; Moran *et al.*, 1994). However, none of these previous algorithms incorporated explicitly the formulation of the roughness height for heat transfer, instead they used fixed values. Since the roughness height for heat transfer can vary with geometrical

and environmental variables in several orders of magnitude for different surface types, this ignorance has been shown by several studies to cause great uncertainties in estimation of heat fluxes or evaporation using radiometric temperature measurements (e.g. Verhoef *et al.*, 1997; Su *et al.*, 2001). Although it is possible to calibrate these algorithms such that their estimates reproduce observations at local scale, it would be very difficult to extend them to regional/continental studies by means of satellite observations. These shortcomings are avoided in SEBS. In addition, the application of SEBS does not require any *a priori* knowledge of the actual turbulent heat fluxes, indicating that SEBS is a credible and independent approach. Due to this property, SEBS results may be used for validation and initialization of hydrological, atmospheric, and ecological models that usually require proper partition of the sensible and latent heat flux at different scales. SEBS results can also be used *via* data assimilation in the above models to increase the reliability of the model simulations and predictions. Another valuable approach would be to implement the representation of surface flux exchanges as described in SEBS into a data assimilation system for assimilating remote-sensing data using modern data assimilation techniques. Some initial encouraging attempts have been made recently by Jia *et al.* (2003) and Wood *et al.* (2003) in this direction.

The examples on the use of parallel-source SEBS clearly indicate the benefit to be gained when more detailed measurements of the state of the land surface can be made. Currently, it has been possible to separate several important variables into their components, for example, the temperature of soil and that of vegetation using currently available sensors (e.g. ATSR and AATSR – the Advanced ATSR on the ENVISAT satellite). By means of multidirectional and multispectral to hyperspectral measurements, it would also be possible to derive more detailed component variables, for example, the temperature of soil and that of vegetation separately for sunlit and sun-shade portions using future available sensors. Use of these variables in algorithms will enable much detailed description of the physical processes involved in land–atmosphere interactions.

In summary, it may be stated that advances on improving parameterization and predictability of hydrological and earth system models, in particular, physically based distributed models, will heavily rely on the understanding of physical processes at different scales as well as the ability to obtain distributed physical information, satellite earth observation will prove of paramount importance in the future. Understanding and quantification of radiation and latent and sensible heating through quantitative remote sensing may hold the key to understand the interaction of terrestrial ecosystems and the hydrological cycles from local to global scales, and as such shall meet the challenges in research and applications in areas ranging from numerical weather forecast, climate research, water cycle study, to water resources management, sustainable agricultural production, and ecological conservation.

Acknowledgments

Researches leading to the work presented here were jointly funded by the Dutch Remote Sensing Board (BCRS), the Dutch Ministry of Agriculture, Nature Management, and Fisheries (LNV), and the Royal Netherlands Academy of Science (KNAW). The author has benefited greatly from discussion and debates with many colleagues. In particular, assistance and constructive comments from Li Jia, Jun Wen, Han Rauwerda, Gerbert Roerink, Katja Sintonen and Claire Jacobs (Alterra), Massimo Menenti and Zhao-Liang Li (TRIO/ULP), Bart van den Hurk (KNMI), Jieming Wang and Yaoming Ma (CAREERI/CAS), Henk de Bruin and Arnold Moene (WU), Renhua Zhang and Xiaoming Sun (IGSNRR/CAS), Xiaomei Li and Li Wan (China University of Geosciences), Tom Schmugge and Bill Kustas (USDA/ARS), Bill Massman (USDA/FS), Wilfried Brutsaert (Connell University), Jetse Kalma (The University of Newcastle) and Eric Wood (Princeton University) are gratefully appreciated.

REFERENCES

Bastiaanssen W.G.M. (1995) *Regionalization of Surface Flux Densities and Moisture Indicators in Composite Terrain – A Remote Sensing Approach Under Clear Skies in Mediterranean Climates*, Ph.D. thesis, Wageningen Agricultural University, The Netherlands, p. 273.

Beljaars A.C.M. and Bosveld F.C. (1997) Cabauw data for the validation of land surface parameterization schemes. *Journal of Climate*, **10**, 1172–1193.

Beljaars A.C.M. and Holtslag A.A.M. (1991) Flux parameterization over land surfaces for atmospheric models. *Journal of Applied Meteorology*, **30**, 327–341.

Blümel K. (1999) A simple formula for estimation of the roughness length for heat transfer over partly vegetated surfaces. *Journal of Applied Meteorology*, **38**, 814–829.

Boegh E., Soegaard H. and Thomsen A. (2002) Evaluating evapotranspiration rates and surface conditions using Landsat TM to estimate atmospheric resistance and surface resistance. *Remote Sensing of Environment*, **79**, 329–343.

Boni G.D., Castelli F. and Entekhabi D. (2001b) Sampling strategies and assimilation of ground temperature for the estimation of surface energy fluxes. *IEEE Transactions on Geoscience Remote Sensing*, **39**, 165–172.

Boni G.D., Entekhabi D. and Castelli F. (2001a) Land data assimilation with satellite measurements for the estimation of surface energy balance components and surface control of evaporation. *Water Resources Research*, **37**, 1713–1722.

Bowen I.S. (1926) The ratio of heat losses by conduction and by evaporation from any water surface. *Physical Review*, **27**, 779–787.

Brutsaert W. (1982) *Evaporation into the Atmosphere*, D. Reidel: Dordrecht.

Brutsaert W. (1999) Aspects of bulk atmospheric boundary layer similarity under free-convective conditions. *Reviews of Geophysics*, **37**, 439–451.

Campbell G.S. and Norman J.M. (1998) *An Introduction to Environmental Biophysics*, Springer.

Caparrini F., Castelli F. and Entekhabi D. (2003) Mapping of land-atmosphere heat fluxes and surface parameters with remote sensing data. *Boundary-Layer Meteorology*, **107**, 605–633.

Caparrini F., Castelli F. and Entekhabi D. (2004) Estimation of surface turbulent fluxes through assimilation of radiometric surface temperature sequences. *Journal of Hydrometeorology*, **5**, 145–159.

Carlson T.N., Taconet O., Vidal A., Cillies R.R., Olioso A. and Humes K. (1995) An overview of the workshop on the thermal remote sensing held at La Londe les Maures, France, September 20–24, 1993. *Agricultural and Forest Meteorology*, **77**, 141–151.

Castelli F., Entekhabi D. and Caporali E. (1999) Estimation of surface heat flux and an index of soil moisture using adjoint-state surface energy balance. *Water Resources Research*, **35**, 3115–3125.

Chanzy A., Bruckler L. and Perrier A. (1995) Soil evaporation monitoring: a possible synergism of microwave and infrared remote sensing. *Journal of Hydrology*, **165**, 235–259.

Chehbouni A., Nouvellon Y., Lhomme J.P., Watts C., Boulet G., Kerr Y.H., Moran M.S. and Goodrich D.C. (2001) Estimation of surface sensible heat flux using dual angle observations of radiative surface temperature. *Agricultural and Forest Meteorology*, **108**, 55–65.

Choudhury B.J. and Monteith J.L. (1988) A four layer model for the heat budget of homogeneous land surfaces. *The Quarterly Journal of Royal Meteorological Society*, **114**, 373–398.

Christensen J.H., Christensen O.B., Lopez P., Van Meijgaard E. and Botzet M. (1996) The HIRHAM4 Regional Atmospheric Climate Model, Scientific Report 96-4, Danish Meteorological Institute, Copenhagen, Scientific Report, 51.

Crago R.D. (1996) Comparison of the evaporative fraction and the Priestley-Taylor a for parameterizing daytime evaporation. *Water Resources Research*, **32**(5), 1403–1409.

De Bruin H.A.R. (2002) Renaissance of scintillometry. *Boundary Layer Meteorology*, **105**, 1–4.

Entekhabi D., Asrar G.R., Betts A.K., Beven K.J., Bras R.L., Duffy C.J., Dunne T., Koster R.D., Lettenmaier D.P. McLaughlin D.B., *et al.* (1999) An agenda for land surface hydrology research and a call for the second international hydrological decade. *Bulletin of the American Meteorological Society*, **80**(10), 2043–2058.

Famiglietti J.S. and Wood E.F. (1994) Multiscale modeling of spatially variable water and energy balance processes. *Water Resources Research*, **30**, 3061–3078.

Girona J., Mata M., Fereres E., Goldhamer D.A. and Cohen M. (2002) Evapotranspiration and soil water dynamics of peach trees under water deficits. *Agricultural Water Management*, **54**, 107–122.

Hasager C. and Jensen N.O. (1999) Surface-flux aggregation in heterogeneous terrain. *The Quarterly Journal of Royal Meteorological Society*, **125**, 1–28.

Högström U. (1988) Non-dimensional wind and temperature profiles in the atmospheric surface layer: a re-evaluation. *Boundary-Layer Meteorology*, **42**, 55–78.

Iqbal M. (1983) *An Introduction to Solar Radiation*, Academic Press: Toronto.

Jackson R.D., Idso S.B., Reginato R.J. and Pinter P.J. Jr (1981) Canopy temperature as a crop water stress indicator. *Water Resources Research*, **17**(4), 1133–1138.

Jackson R.D., Kustas W.P. and Choudhury B.J. (1988) A re-examination of the crop water stress index. *Irrigation Science*, **9**, 309–317.

Jackson R.D., Reginato R.J. and Idso S.B. (1977) Wheat canopy temperature: A practical tool for evaluating water requirements. *Water Resources Research*, **13**, 651–656.

Jia L., Menenti M., Su Z., Djepa V., Li Z.-L. and Wang J. (2001) Modeling sensible heat flux using estimates of soil and foliage temperatures: the HEIFE and IMGRASS experiments. In *Remote Sensing and Climate Modeling: Synergies and Limitations*, Beniston M. and Verstraete M. (Eds.), Kluwer Academic Publisher: Dordrecht.

Jia L., Su Z., van den Hurk B., Menenti M., Moene A., De Bruin H.A.R., Yrisarry J.J.B., Ibanez M. and Cuesta A. (2003) Estimation of sensible heat flux using the surface energy balance system (SEBS) and ATSR measurements. *Physics Chemistry of the Earth*, **28**(1–3), 75–88.

Kabat P. (1999) *The Role of Biospheric Feedbacks in the Hydrological Cycle*, The IGBP – BAHC Special Issue, (Global Change Newsletter), IGBP Newsletter, 39.

Kader B.A. and Yaglom A.M. (1990) Mean fields and fluctuation moments in unstably stratified turbulent boundary layers. *Journal of Fluid Mechanics*, **212**, 637–662.

Kalma J.D. and Jupp D.L.B. (1990) Estimating evaporation from pasture using infrared thermometry: evaluation of a one-layer resistance model. *Agricultural and Forest Meteorology*, **51**, 223–246.

Kalnay E. and Cai M. (2003) Impact of urbanization and land-use change on climate. *Nature*, **423**, 528–531.

Katul G.G. and Parlange M.B. (1992) A Penman-Brutsaert model for wet surface evaporation. *Water Resources Research*, **28**(1), 121–126.

Koike T. (2000) The overview of GAME/Tibet. *Proceedings of The Second Session of International Workshop on TIPEX-GAME/Tibet*, Kunming, July 20–22, 2000.

Koster R.D., Suarez M.J. and Heiser M. (2000) Variance and predictability of precipitation at seasonal to interannual timescale. *Journal of Hydrometeorology*, **1**, 26–46.

Kustas W.P. (1990) Estimates of evapotranspiration with a one and two layer model of heat transfer over partial canopy layer. *Journal of Applied Meteorology*, **29**, 704–715.

Kustas W.P., Blanford J.H., Stannard D.I., Daughtry C.S.T., Nichols W.D. and Weltz M.A. (1994) Local energy flux estimates for unstable conditions using variance data in semiarid rangelands. *Water Resources Research*, **30**, 1351–1361.

Kustas W.P. and Daughtry C.S.T. (1989) Estimation of the soil heat flux/net radiation ratio from spectral data. *Agricultural and Forest Meteorology*, **49**, 205–223.

Kustas W.P. and Norman J.M. (1996) Use of remote sensing for evapotranspiration monitoring over land surfaces. *Hydrological Sciences Journal*, **41**(4), 495–516.

Kustas W.P. and Norman J.M. (1999) Evaluation of soil and vegetation heat flux predictions using a simple two-source model with radiometric temperatures for partial canopy cover. *Agricultural and Forest Meteorology*, **94**, 13–29.

Lhomme J.-P., Monteny B. and Amadou M. (1994) Estimating sensible heat flux from radiometric temperature over sparse millet. *Agricultural and Forest Meteorology*, **68**, 77–91.

Li X. (2001) *Estimation of Urumqi River Basin Evaporation with Remote Sensing*, M.Sc. Thesis in Hydrological Engineering, UNESCO-IHE, Delft.

Ma M., Su Z., Li Z.-L., Koike T. and Menenti M. (2002) Determination of regional net radiation and soil heat flux over a heterogeneous landscape of the Tibetan Plateau. *Hydrological Processes*, **16**(15), 2963–2971.

Massman W.J. (1997) An analytical one-dimensional model of momentum transfer by vegetation of arbitrary structure. *Boundary-Layer Meteorology*, **83**, 407–421.

Massman W.J. (1999a) A model study of kB_H^{-1} for vegetated surfaces using 'localized near-field' Lagrangian theory. *Journal of Hydrology*, **223**, 27–43.

Massman W.J. (1999b) Molecular diffusivities of Hg vapor in air, O_2 And N_2 near STP and the kinematic viscosity and the thermal diffusivity of air near STP. *Atmospheric environment*, **33**, 453–457.

Mecikalski J.M., Diak G.R., Anderson M.C. and Norman J.M. (1999) Estimating fluxes on continental scales using remotely-sensed data in an atmosphere-land exchange model. *Journal of Applied Meteorology*, **38**, 1352–1369.

Menenti M. (1984) *Physical Aspects of and Determination of Evaporation in Deserts Applying Remote Sensing Techniques*, Report 10 (special issue), Institute for Land and Water Management Research (ICW): The Netherlands.

Menenti M. (1993) Understanding land surface evapotranspiration with satellite multispectral measurements. *Advances in Space Research*, **13**(5), 89–100.

Menenti M. and Choudhury B.J. (1993) Parameterization of land surface evapotranspiration using a location-dependent potential evapotranspiration and surface temperature range. In *Exchange Processes at the Land Surface for a Range of Space and Time Scales*, Bolle H.J., Feddes R.A. and Kalma J.D. (Eds.), IAHS Publication No. 212: 561–568.

Menenti M., Choudhury B.J. and Di Girolamo N. (2001) Monitoring of actual evaporation in the Aral Basin using AVHRR observations and 4DDA results. In *Advanced Earth Observation – Land Surface Climate*, Su Z. and Jacobs C. (Eds.), Report USP-2, 01–02, Publications of the National Remote Sensing Board (BCRS): pp. 79–83.

Menenti M. and Ritchie J.C. (1994) Estimation of effective aerodynamic roughness of Wulnut Gulch watershed with laser altimeter measurements. *Water Resources Research*, **30**, 1329–1337.

Mitchell K., Houser P., Wood E., Schaake J., Tarpley D., Lettenmaier D., Higgins W., Marshall C., Lohmann D., Ek M., *et al.* (1999) GCIP land data assimilation system (LDAS) project now underway, *GEWEX News*, **9**(4), 3–6.

Moene A.F. (2001) Description of field experiments. In *Advanced Earth Observation – Land Surface Climate*, Su Z. and Jacobs C. (Eds.), Report USP-2, 01–02, Publications of the National Remote Sensing Board (BCRS).

Moene A.F. and De Bruin H.A.R. (2001) Sensible heat flux data derived from the scintillometers. In *Advanced Earth Observation – Land Surface Climate*, Su Z. and Jacobs C. (Eds.), Report USP-2, 01–02, Publications of the National Remote Sensing Board (BCRS).

Monteith J.L. (1965) Evaporation and environment. *Symposia of the Society for Experimental Biology*, **19**, 205–234.

Monteith J.L. (1973) *Principles of Environmental Physics*, Edward Arnold Press.

Monteith J.L. (1981) Evaporation and Surface Temperature. *The Quarterly Journal of Royal Meteorological Society*, **107**, 1–27.

Morton F.I. (1983) Operational estimates of areal evapotranspiration and their significance to the practice of hydrology. *Journal of Hydrology*, **66**, 1–76.

Moran M.S., Clarke T.R., Inoue Y. and Vidal A. (1994) Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index. *Remote Sensing of Environment*, **49**, 246–263.

Mücher C.A., Steinnocher K., Champeaux J.L., Griguolo S., Wester K. and Loudjani P. (2001) *Land Cover Characterization for Environmental Monitoring of Pan-Europe*, Wageningen University and Research Centre, Centre for Geo-information: Wageningen, **http://cgi.girs.wageningen-ur.nl/cgi/projects/eu/pelcom/public/index.htm**.

Nieuwenhuis G.J.A., Schmidt E.A. and Tunnissen H.A.M. (1985) Estimation of regional evapotranspiration of arable crops from thermal infrared images. *International Journal of Remote Sensing Environment*, **6**, 1319–1334.

Nishida K., Nemani R.R., Glassy J.M. and Running S.W. (2003a) Development of an evapotranspiration index from aqua/MODIS for monitoring surface moisture status. *IEEE Transactions on Geoscience Remote Sensing*, **41**(2), 493–501.

Nishida K., Nemani R.R., Running S.W. and Glassy J.M. (2003b) An operational remote sensing algorithm of land surface evaporation, *Journal of Geophysical Research-Atmospheres*, **108**(D9), Art. No. 4270, doi:10.1029/2002JD002062.

Norman J.M., Kustas W.P. and Humes K.S. (1995) A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature. *Agricultural and Forest Meteorology*, **77**, 263–293.

Olioso A., Braud I., Chanzy A., Demarty J., Ducros Y., Gaudu J.-C., Gonzalez-Sosa E., Lewan E., Marloie O., Ottle C., *et al.* (2002) Monitoring energy and mass transfers during the Alpilles-ReSeDA experiment. *Agronomie*, **22**, 597–610.

Penman H.L. (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A*, **193**, 120–146.

Pielke R.A. Sr (2001a) Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. *Reviews of Geophysics*, **39**, 151–177.

Pielke R.A. Sr (2001b) Comments on IPCC report cautiously warns of potentially dramatic climate change impacts. *The Earth Observing System*, **82**, 394–396.

Press W.H., Teukolsky S.A., Vetterling W.T. and Flannery B.P. (1997) *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press.

Priestley C.H.B. and Taylor R.J. (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, **100**(2), 81–92.

Rauwerda J., Roerink G.J. and Su Z. (2002) *Estimation of Evaporative Fractions by the Use of Vegetation and Soil Component Temperatures Determined by Means of Dual-Looking Remote Sensing*, Alterra-report 580, ISSN 1566–7197.

Roerink G.J., Su Z. and Menenti M. (2000) S-SEBI: A simple remote sensing algorithm to estimate the surface energy balance. *Physics Chemistry of the Earth*, **25**, 147–157.

Seguin B. and Ittier B. (1983) Using midday surface temperature to estimate daily evaporation from satellite thermal IR data. *International Journal of Remote Sensing Environment*, **4**, 371–383.

Sellers P.J., Randall D.A., Collatz G.J., Berry J.A., Field C.B., Dazlich D.A., Zhang C., Collelo G.D. and Nounoua L. (1996) A revised land surface parameterization (SiB2) for atmospheric GCMS, part 1: model formulation. *Journal of Climate*, **9**, 676–705.

Shuttleworth W.J., Gurney R.J., Hsu A.Y. and Ormsby J.P. (1989) FIFE: the variation in energy partition at surface flux sites. *IAHS Publication*, **186**, 67–74.

Shuttleworth W.J. and Wallace J.S. (1985) Evaporation from sparse crops – an energy combination theory. *The Quarterly Journal of Royal Meteorological Society*, **111**, 939–955.

Stanghellini C. (1987) *Transpiration of Greenhouse Crops – an Aid to Climate Management*, Ph.D. thesis, Agriculture University, Wageningen.

Stull R.B. (1988) *An Introduction to Boundary Layer Meteorology*, Kluwer Academic Publishers: Dordrecht.

Su Z. (2002) The surface energy balance system (SEBS) for estimation of turbulent heat fluxes. *Hydrol. Earth Sys. Sci.*, **6**(1), 85–99.

Su Z. and Jacobs C. (Eds.) (2001) *Advanced Earth Observation – Land Surface Climate*, Report USP-2, 01–02, Publications of the National Remote Sensing Board (BCRS).

Su Z. and Menenti M. (Eds.) (1999) *Mesoscale Climate Hydrology: The Contribution of the New Observing Systems*, Report USP-2, 99-05, Publications of the National Remote Sensing Board (BCRS).

Su Z., Li X., Zhou Y., Wan L., Wen J. and Sintonen K. (2003a) Estimating areal evaporation from remote sensing. *Paper Presented at the International Geoscience and Remote Sensing Symposium*, 21–25 July 2003, Toulouse (PID20349.pdf).

Su Z., Pelgrum H. and Menenti M. (1999) Aggregation effects of surface heterogeneity in land surface processes. *Hydrology and Earth System Sciences*, **3**(4), 549–563.

Su Z., Schmugge T., Kustas W.P. and Massman W.J. (2001) An evaluation of two models for estimation of the roughness height for heat transfer between the land surface and the atmosphere. *Journal of Applied Meteorology*, **40**(11), 1933–1951.

Su Z., Yacob A., Wen J., Roerink G., He Y., Gao B., Boogaard H. and van Diepen C. (2003b) Assessing relative soil moisture with remote sensing data: theory and experimental validation. *Physics Chemistry of the Earth*, **28**(1–3), 89–101.

Sugita M. and Brutsaert W. (1991) Daily evaporation over a region from lower boundary-layer profiles measured with radiosondes. *Water Resources Research*, **27**(5), 747–752.

Troufleau D., Lhomme J.-P., Monteny B. and Vidal A. (1997) Sensible heat flux and radiometric temperature over sparse Sahelian vegetation I: An experimental analysis of the kB^{-1} parameter. *Journal of Hydrology*, **188–189**(1–4), 815–838.

Twine T.E., Kustas W.P., Norman J.M., Cook D.R., Houser P.R., Meyers T.P., Prueger J.H., Starks P.J. and Wesley M.L. (2000) Correcting eddy-covariance flux underestimates over a grassland. *Agricultural and Forest Meteorology*, **103**, 279–300.

Van den Hurk B.JJ.M and Holtslag A.A.M. (1995) On the bulk parameterization of surface fluxes for various conditions and parameter ranges. *Boundary-Layer Meteorology*, **82**, 199–234.

Verhoef A., de Bruin H.A.R. and van den Hurk B.J.J.M. (1997) Some practical notes on the parameter for sparse vegetation. *Journal of Applied Meteorology*, **36**, 560–572.

Wieringa J. (1993) Representative roughness parameters for homogeneous terrain. *Boundary-Layer Meteorology*, **63**, 323–363.

Wood E.F. (1998) *Hydrologic Measurements and Observations: an Assessment of Needs in Hydrologic Sciences: Taking Stock and Looking Ahead*, National Academy Press, pp. 67–86.

Wood E.F., Su H., McCabe M. and Su Z. (2003) Estimating evaporation from satellite remote sensing. Paper presented at the *International Geoscience and Remote Sensing Symposium*, 21–25 July 2003, Toulouse (PID20004.pdf).

Zhan X., Kustas W.P. and Humes K.S. (1996) An intercomparison study on models of sensible heat flux over partial canopy surfaces with remotely sensed surface temperature. *Remote Sensing of Environment*, **58**, 242–256.

Zhang R.H. (1996) *Remote Sensing Models and Ground Surface Foundation*, The Science Press: Beijing.

51: Spatially Resolved Measurements of Evapotranspiration by Lidar

WILLIAM E EICHINGER

IIHR Hydroscience and Engineering, University of Iowa, Iowa City, IA, US

Spatially resolved estimates of evapotranspiration may be made over an area of about a square kilometer, with relatively fine (25 m) spatial resolution, derived from three-dimensional measurements of water vapor concentration made by a scanning Raman lidar. The method is based upon the Monin–Obukhov similarity theory applied to spatially and temporally averaged data. The method has been applied at several locations to produce maps of the spatial distribution of evapotranspiration rates at regular intervals throughout the day. The estimates of evapotranspiration rates made compare favorably with estimates made using eddy correlation methods. The method is limited in that it assumes that the surface roughness, friction velocity, and Obukhov length are assumed constant over the examined areas. Values for the friction velocity and Obukhov length are required from conventional point instruments. The three-dimensional water vapor concentration and the evaporation maps can be used in a wide variety of ways to study the spatial variations in evapotranspiration caused by changes in soil type and moisture content, canopy type, and topography.

INTRODUCTION

Evapotranspiration is one of the critical variables in both the water and energy balance models of hydrologic systems. These systems as well as the local weather are driven by the way in which energy is partitioned at the earth's surface. This partitioning is determined by the conditions at the soil–plant–atmosphere interface and are often highly spatially variable. Traditional techniques of measuring evapotranspiration rely on point sensors to collect information, which is then assumed to be representative of a much larger area. However, data from individual point sensors near the surface are of limited value. This is due to their relatively small footprint, the necessarily limited number of sensors, and our current inability to extend the values measured at a point (or series of points) to understanding the details of processes that are occurring on larger scales. Part of the problem is that the bulk of the earth's surface is not horizontally homogeneous with respect to topography, soil moisture availability, soil type, or canopy characteristics. More highly resolved information is necessary to separate the contributions of each of these variables. Eddy correlation techniques using fast-response sensors have been successfully used from aircraft to cover large areas, but the usefulness of the data has been called into question for use in mixed canopies where spatially resolved fluxes are desired (for example, Mahrt, 1998). It is possible to obtain spatially resolved evapotranspiration estimates from surface temperature measurements made by aircraft (Neale *et al.*, 2001) or satellite (Moran *et al.*, 1989; Anderson *et al.*, 1997, 2004). These measurements also require supporting meteorological information and require assumptions of local homogeneity and surface roughness characteristics.

The three-dimensional scanning Raman lidar (laser radar) built by Los Alamos National Laboratory (LANL) can provide detailed maps of the water vapor concentration in three dimensions with high spatial (approximately 1.5 m) and temporal resolution (approximately 15 min to scan a field). A method has been developed to estimate spatially resolved evaporative fluxes over the scanned area using this information. A description of the method used to measure the water vapor concentrations is followed by the details of the method used to determine the evapotranspiration rates from the water vapor concentration.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

MEASUREMENT OF WATER VAPOR

There are two methods by which lidars can measure spatially resolved water vapor concentrations. The first uses Raman scattering from nitrogen and water vapor molecules in the atmosphere (Kovalev and Eichinger, 2004). The second is the differential absorption lidar (DIAL) method that uses the difference in absorption between two laser wavelengths, one of which is tuned to an absorption line of the molecule of interest (Measures, 1984).

Raman Lidar Method

Raman lidars are a type of laser radar that use a technique originally pioneered by Cooney *et al.* (1969), Cooney (1970), and Melfi *et al.* (1969). A Raman lidar operates by emitting a pulsed laser beam, usually in the ultraviolet or near ultraviolet, into the atmosphere. Atmospheric gases, such as nitrogen, oxygen, and water vapor absorb this light via the Raman scattering process, causing light of longer wavelengths to be emitted. The amount of the wavelength shift is unique to each molecule. For example, the light scattered by nitrogen is shifted by $2331\,\mathrm{cm}^{-1}$ from the laser wavelength. This enables the measurement of different atmospheric gases by this technique. Figure 1 is a plot of the spectrum of light returning from an emitted 248-nm pulse, showing the peaks from the major atmospheric constituents. Because of the small value of the backscatter cross sections for Raman scattering, the number of photons returning to the lidar is small. Since the probability of Raman scattering is proportional to λ^{-4}, the use of short wavelengths increases the magnitude of the signal. Modern Raman lidars are typically found at ultraviolet wavelengths, particularly at 248 nm (KrF excimer), 266 nm (quadrupled Nd : YAG), 308 nm (XeCl excimer), 351 nm (XeF excimer), and 355 nm (tripled Nd : YAG).

The LANL Raman lidar (shown schematically in Figure 2, photograph in Figure 3) is a typical Raman lidar. In this lidar, the laser is mounted below the telescope. A series of mirrors and lenses is used to expand the beam to make it eye-safe and collinear with the telescope. A forty-five-degree mirror is used to change the optical direction to vertical, allowing the system to make vertical soundings. The scanning mirror enables the system to perform three-dimensional scanning near the earth's surface. At the back of the telescope, the light at the two Raman-shifted wavelengths (from nitrogen and water vapor) is separated and directed to different photomultipliers using a series of dichroic beam splitters. Narrow band interference filters are used to further block unwanted wavelengths of light going to each photomultiplier.

Because of the small cross section for Raman scattering, the number of photons returning to the lidar is small so that photon counting is required to achieve meaningful

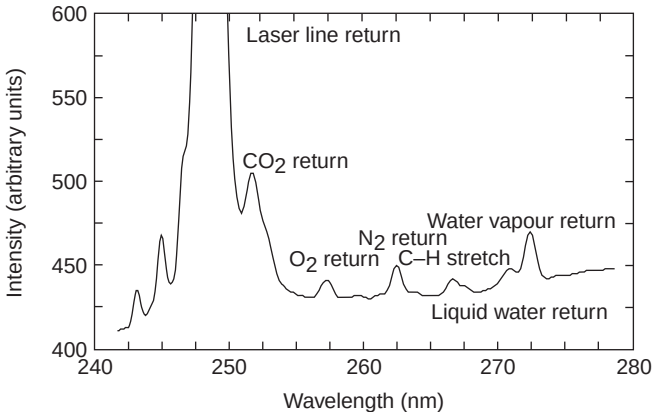


Figure 1 A plot of the spectrum of light returning from a 248 nm Raman lidar showing the peaks from the various atmospheric constituents

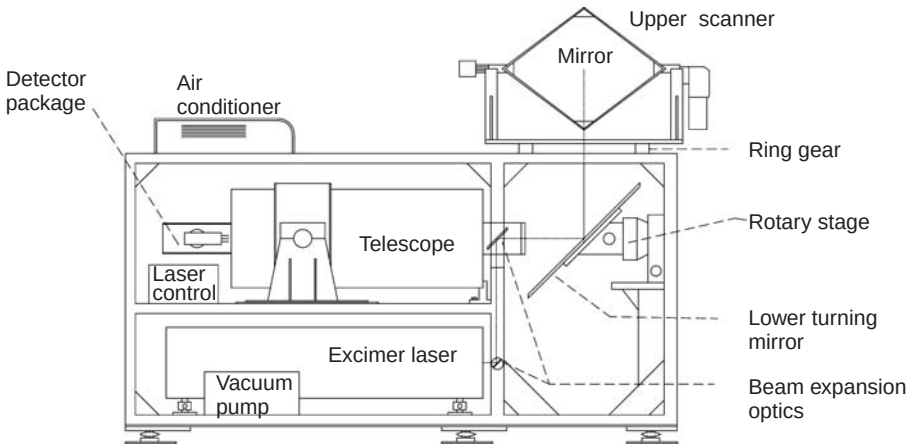


Figure 2 Diagram showing the layout of the Los Alamos Raman lidar. With the exception of the scanning mirror, the arrangement is typical of Raman lidars

Figure 3 Photograph of the LANL Scanning Raman Lidar

signals at long ranges (greater than 1 km). The discrimination of these photons from background light is another issue that must be addressed. To work during the day, many systems operate in the region of the spectrum below about 300 nm where ozone and oxygen strongly absorb sunlight and are thus "blind" to solar photons. Daytime, solar-blind operation for Raman lidars was developed by Renault *et al.* (1980), Cooney *et al.* (1985), and Renault and Capitini (1988). Solar-blind operation requires the use of a laser near 250–260 nm so that the Raman-shifted lines will be below 300 nm. The use of a laser with a wavelength longer than 266 nm will have contamination from sunlight at the Raman-shifted wavelengths. The use of laser wavelengths shorter than 248 nm will be so strongly absorbed by ozone and oxygen absorption at both the emission and Raman-shifted wavelengths that long-range use is not possible. Since wavelengths longer than 300 nm are not strongly absorbed by atmospheric ozone, and molecular scattering is reduced at longer wavelengths, much greater ranges are possible with the use of longer wavelength lasers. However, if the system is expected to operate during the day, the use of an extremely narrow field of view in the receiving telescope is required along with spectrally narrow filters. While several nonsolar blind systems have been built (Cooney, 1983; Ansmann *et al.*, 1992; Goldsmith *et al.*, 1998), there are a multitude of technical issues that make them quite difficult, particularly for use in the boundary layer. Because of the limited amount of returning light, Raman lidar systems tend to use large, powerful lasers and large telescopes so that systems are unusually large and require significant amounts of power. Figure 3 is a photograph of the LANL Scanning Raman lidar mounted on its trailer. The size and power requirements of such systems restrict where they can be used and make them difficult for use in the field.

At least in part because of the limitations of photon counting, most Raman lidars operate in a vertically staring mode. Currently, the two exceptions are the NASA Goddard Space Flight Center (GSFC) Raman lidar (Ferrare *et al.*, 1998), which can scan in a vertical plane, and the LANL Scanning Raman Lidar (Eichinger *et al.*, 1999), which can scan in three dimensions. The GSFC Raman lidar operates primarily in a staring mode along different elevation angles. Operating at various angles to the ground enables the system to achieve higher vertical resolution at low altitude. The typical maximum horizontal range for the LANL lidar, when not using a photon counting mode, is approximately 700 m when scanning, with a corresponding spatial resolution of 1.5 m over that distance. The upper scanning mirror allows three-dimensional scanning of 360 degrees in azimuth and ±22 degrees in elevation.

The backscattered signals from the nitrogen and water vapor Raman scattering are given by the following equations:

$$P_{N_2}(r) = \frac{C_{N_2} E n_{N_2}(r) \sigma_{N_2} \times \exp\left[-\int_0^r [\kappa_t(r', \lambda) + \kappa_t(r', \lambda_{N_2,R})]\, dr'\right]}{r^2}, \tag{1}$$

and

$$P_{H_2O}(r) = \frac{C_{H_2O} E n_{H_2O}(r) \sigma_{H_2O} \times \exp\left[-\int_0^r [\kappa_t(r', \lambda) + \kappa_t(r', \lambda_{H_2O,R})]\, dr'\right]}{r^2} \tag{2}$$

where λ is the laser wavelength, $\lambda_{N_2,R}$ and $\lambda_{H_2O,R}$ are the Raman N_2 and H_20 scattered wavelengths, respectively. Note that the Raman scattered wavelengths in the above equations have different values, $\lambda_{N_2,R}$ for nitrogen and $\lambda_{H_2O,R}$ for water vapor. The functions $P_{N_2}(r)$ and $P_{H_2O}(r)$ are the received signals in the nitrogen and water vapor channels, E is the laser energy per pulse; σ_{N_2} and σ_{H_2O} are the Raman backscatter cross sections for the laser wavelength; r is the distance from the lidar; $n_{N_2}(r)$ and $n_{H_2O}(r)$ are the number density of nitrogen and water molecules at range, r; $\kappa_t(r, \lambda)$, $\kappa_t(r, \lambda_{N_2,R})$, and $\kappa_t(r, \lambda_{H_2O,R})$ are the total extinction coefficients at the laser wavelength, λ, and at the Raman-shifted wavelengths of nitrogen and water vapor molecules; and C_{N_2} and C_{H_2O} are the system coefficients that take into account the effective area of the telescope, the transmission efficiency of the optical train, and the detector quantum efficiency at the Raman-shifted wavelengths.

The principal advantage of the Raman lidar technique lies in having an additional signal from a ubiquitous atmospheric gas (specifically nitrogen or oxygen) in addition to the water vapor signal and the conventional elastic signal. This signal can be used to normalize the signal from water vapor. The backscatter coefficient from a particular

molecule is proportional to the gas density with altitude. The nitrogen density is known or can be calculated from temperature and pressure measurements, which in turn can be obtained from meteorological balloons or climatological data. The extinction coefficients at the emitted and Raman scattered wavelengths in the exponent term of equations (1) and (2), $\kappa_t(h, \lambda)$, $\kappa_t(h, \lambda_{N_2,R})$, and $\kappa_t(h, \lambda_{H_2O,R})$, are nearly the same if the Raman shift is small.

The mixing ratio of any gas is the mass of gas divided by the mass of the dry air in a given volume. A combination of equations (1) and (2) allows determination of the water vapor mixing ratio as a function of distance from the lidar. The mixing ratio can be obtained from the ratio of the signal magnitude in the water vapor channel, $P_{H_2O}(r)$, to the magnitude of the signal in the nitrogen channel, $P_{N_2}(r)$, using the formula (Melfi, 1972),

$$q_w(r) = \frac{P_{H_2O}(r)}{P_{N_2}(r)} \left[\frac{C_{N_2}\sigma_{N_2}n_{N_2}}{C_{H_2O}\sigma_{H_2O}n_{H_2O}fr_{N_2}} \right] \exp\left[\int_0^r [\kappa_t(r', \lambda_{N_2,R}) - \kappa_t(r', \lambda_{H_2O,R})]\, dr' \right] \quad (3)$$

where fr_{N_2} is the fractional N_2 content of the atmosphere (0.78084). Thus, the water vapor mixing ratio at any distance, r, is given by the ratio of the magnitude of the signal in the water vapor channel to the magnitude of the signal in the nitrogen channel, a multiplicative constant (the part in large square brackets), and an exponential correction due to the difference in extinction between the nitrogen-shifted and water vapor–shifted wavelengths. The multiplicative constant can be determined by comparison of the lidar signal with radiosondes (although there are issues with respect to the accuracy of the humidity sensors at low concentrations (Connell and Miller, 1995; Ferrare *et al.*, 1995; Sherlock *et al.*, 1999) or by aiming the lidar horizontally and comparing with calibrated water vapor point sensors at various distances from the lidar.

When the lidar aims along a given line of sight, data are obtained every 1.5 m along that line. By aiming the lidar in a series of different directions, a two or three-dimensional map of the water vapor concentrations can be assembled. Figure 4 is a conceptual drawing showing how the various lidar lines of sight are used to scan the area and produce an image such as the one shown in Figure 5. Figure 5 is a typical scan from the LANL Raman lidar showing the water vapor concentration in a vertical plane above a cornfield in Iowa during the Soil Moisture-Atmospheric Coupling Experiment (SMACEX) (details of SMACEX can be found in Kustas *et al.* (2004)). The dark grey color at the bottom is a result of the attenuation of the laser beam by the ground or canopy (in this case, corn). The change in the lidar signal as it reaches the canopy top enables one to determine the shape and orientation of the surface.

Uncertainty

The uncertainty in the values for the mixing ratio is a strong function of distance from the lidar. Details of the derivation of the uncertainty for Raman lidar water vapor estimates can be found in Eichinger *et al.* (2000). For a mid-range distance (~350 m), the estimated uncertainty is approximately 3.6%. This is consistent with the comparisons of lidar and calibrated references over land surfaces along horizontal paths. The standard deviation between the lidar and capacitance hygrometer data taken at concurrent times and locations was found through regression to be $\pm 0.34\,g\,kg^{-1}$. A large part this uncertainty is due to significant differences between the lidar

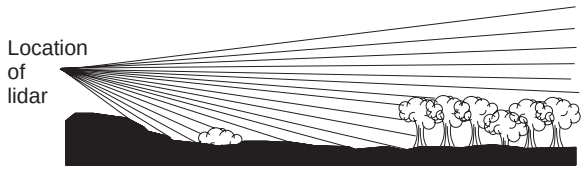


Figure 4 A conceptual drawing showing how different lines of sight from the lidar are combined to map the water vapor concentration in a vertical slice of the atmosphere. The water vapor concentration is determined every 1.5-m along each of the lines shown. The lines of sight in actual practice are 0.07 to 0.25 degrees apart

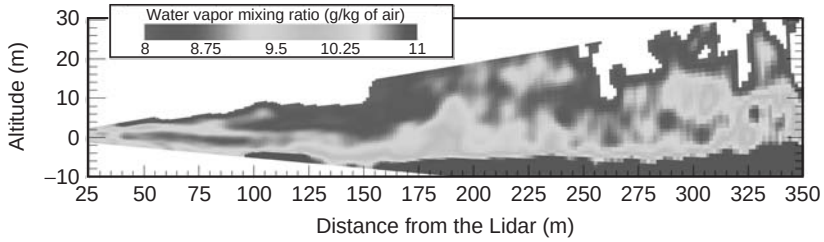


Figure 5 A vertical scan from the Raman lidar showing the water vapor concentrations in a plane perpendicular to the ground over a cornfield in Iowa during SMACEX. Lighter colors represent highest water concentrations and dark grey represent the lowest concentrations. The approximately horizontal dark grey zone at the bottom of the figure is the top of the corn canopy. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and hygrometers that occur in regions where the water vapor concentration changes dramatically. In these regions, the slower time resolution of the capacitance hygrometer "averages" the changes in atmospheric structure as compared to the higher resolution lidar data (Cooper *et al.*, 1996).

Differential Absorption Lidar Method

The choice of a Raman lidar to scan and determine spatial water vapor concentrations is unusual. This type of lidar is most often used for high altitude water vapor measurements using photon counting techniques with relatively poor range resolution. The main competitor to the use of Raman systems is the DIAL method. In this method, two different, but closely spaced, wavelengths of laser light are used. One wavelength is absorbed by the molecule of interest, the other is not. Because the two wavelengths are nearly the same, their atmospheric transmission properties are similar, except for the amount of absorption by the molecule of interest. Thus, any differences in the intensity of the returning light can be used to determine the range resolved concentrations. The water vapor concentration is proportional to the logarithm of the ratio of the lidar returns at the two wavelengths, the distance over which the attenuation occurs (which is the range resolution) and the absorption coefficient of the particular molecular absorption line used. Many of these types of lidars have been or are in use (for example, Higdon *et al.*, 1994; Ehret *et al.*, 1993; Hinkley, 1976). DIAL systems have a number of inherent advantages over Raman lidars. They are generally smaller, lighter, and use considerably less energy than Raman systems. They are especially well suited for use in aircraft because they rely on atmospheric backscatter, which increases with range in a downward looking system, and thus partially compensates for the r^{-2} decrease in signal with range. Because of the strong atmospheric attenuation of UV and near UV light, DIAL systems in the near-infrared are better suited for deep atmospheric sounding. Both systems must average the return from large numbers of laser pulses to obtain data with sufficient statistical significance. However, high energy (250–400 mJ per pulse) excimer lasers with pulse rates of 200 Hz are common. In contrast, the materials commonly used in water vapor DIAL systems (flashlamp or Nd : YAG pumped dyes, alexandrite, and Ti : Sapphire lasers) have inherent limitations on the pulse rates that limit high power systems to 10 to 50 Hz. Thus, Raman systems have the potential for better temporal resolution. While the range resolution of DIAL systems is dependent upon the strength of the absorption feature used, existing systems have range resolutions on the order of 15 to 500 m, which makes them unsuitable for the evapotranspiration technique described here which requires high spatial resolution measurements near the surface. An excellent discussion of the relative technical merits of each type of system can be found in Grant (1991).

DIAL systems have been used to estimate the evaporative flux. A combination of lidar measurements at some point in the middle of the atmospheric boundary layer and measurements of the vertical wind velocity have been used to make an eddy correlation estimate of the evaporative flux. Kiemle *et al.* (1997) used an airborne DIAL system to estimate the entrainment flux at the top of the boundary layer. Senff *et al.* (1994) used a DIAL lidar along with a radar RASS system to obtain the vertical wind velocity. Geiz *et al.* (1999) used a DIAL lidar along with a Doppler lidar to measure the vertical wind velocity. As the height of the measurement increases, the rate at which measurements must be made decreases. The rate at which measurements need to be made must be fast enough to capture all of the frequencies that contribute to the flux (Brutsaert, 1982). In the middle of the boundary layer (300 to 1000 m above the surface), the required data collection rates begin to match those that a DIAL lidar and a RASS system can achieve. It could be argued that, in this situation, DIAL lidars will certainly capture the motion of the large structures that transport the bulk of the water vapor. One of the issues associated with the use of multiple remote sensors, with different divergences, in this way is ensuring that both instruments are measuring the same volume of air.

LIDAR-DERIVED FLUX METHOD

Monin–Obukhov Similarity

The water vapor concentration in the vertical direction can be described using the Monin–Obukhov Similarity Method (MOM) (Brutsaert, 1982). With this theory, the relationship between the water vapor concentration at the surface and that at some height, z, within the inner region of the boundary layer is

$$q_s - q(z) = \frac{E}{L_e k u_* \rho}\left[\ln\left(\frac{z-d_0}{z_{ov}}\right) - \psi_v\left(\frac{z}{L}\right)\right] \qquad (4)$$

where the Obukhov length, L, is defined as:

$$L = -\frac{\rho {u_*}^3}{kg\left[\dfrac{H}{Tc_p} + 0.61E\right]} \qquad (5)$$

where z_{0v} is the roughness length for water vapor; q_s is the surface-specific humidity; T, is the atmospheric temperature; $q(z)$ is the specific humidity at height z; H is the sensible heat flux; E is the latent energy flux (the evapotranspiration rate in Joules); ρ is the density of the air; L_e is the latent heat of evaporation for water; u_* is the friction velocity (Brutsaert, 1982); k is the von

Karman constant (taken as 0.40); d_0 is the displacement height; and g is the acceleration due to gravity. Ψ_v is the Monin–Obukhov stability correction function for water vapor and is calculated for unstable conditions as

$$\psi_v\left(\frac{z}{L}\right) = 2\ln\left[\frac{(1+x^2)}{2}\right]$$

where

$$x = \left(1 - 16\frac{z}{L}\right)^{1/4}$$

and where x_{ov} represents the function x calculated for the value of z_{ov}. The roughness length is a free parameter to be calculated based upon the conditions onsite. The Obukhov length, L, is negative when the atmosphere is unstably stratified. This kind of condition occurs when the soil/canopy is warmer than the air above, causing convective mixing to occur. The value of L is positive when the atmosphere is stably stratified. In this case, the potential temperature increases with altitude, suppressing vertical transport. L is infinite in a neutral atmosphere, when the potential temperature is constant with altitude. This condition normally occurs only in the transition from day to night or night to day but may occur when the winds are exceptionally still.

Heat and momentum fluxes are often determined from measurements of temperature, humidity, and wind speed at two or more heights. These relations are valid in the inner region of the boundary layer where the atmosphere reacts directly with the surface. This region is limited to an area between the roughness sublayer (the region directly above the surface roughness elements) and extending to a height of 5 to 30 m above the canopy top. The concentrations of passive scalars are semilogarithmic with height in this region. The vertical extent of this layer is highly dependent upon the local conditions and wind velocity. The top of this region can be readily *identified by* a departure from the logarithmic profile near the surface. Figure 6 is an example of a water vapor profile with a logarithmic fit showing such a departure at approximately 6 m above the canopy top. The suggestion has been made that the atmosphere is also logarithmic to higher levels and may integrate fluxes over large areas (for example, Brutsaert, 1998).

The method currently used (Eichinger *et al.*, 1993a, 2000) begins by rearranging equation (4) into a linear form

$$q(z) = -Mz' + c \tag{6}$$

where M is the slope of the fitted function ($M = E/(L_e k u_* \rho)$), z' is a reduced height parameter ($z' = \ln(z - d_0) - \Psi_v((z - d_0)/L)$), and c is a regression constant ($c = M\ln(z_o) + q_s$). Measurements of the slope are made on the basis of a least-squares fit to several hundred measurements of water vapor concentration. Having determined M from the slope of the fitted line, the flux is then

$$E = L_e M k u_* \rho \tag{7}$$

where u_* and L are obtained from local measurements using three-dimensional sonic anemometers.

The method is similar to other gradient methods for determining fluxes that are well established (Stull, 1988; Brutsaert, 1982). The lidar method is unique in that it uses

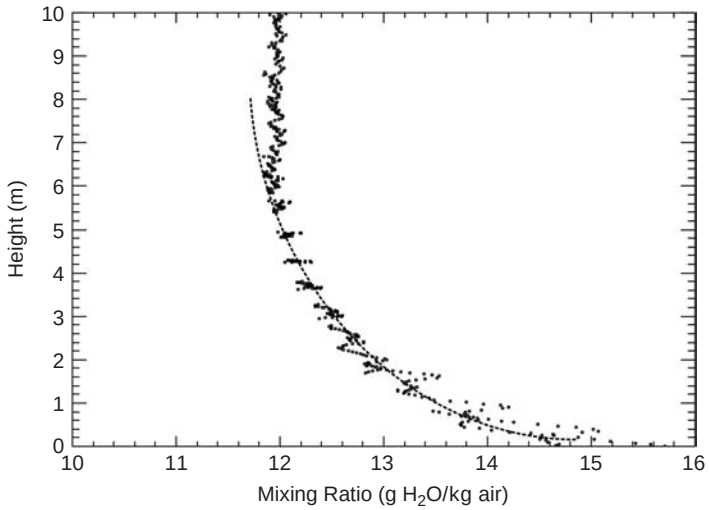


Figure 6 An example of a lidar fitted vertical profile and the data from which it was calculated. The data are from a 25-m section from a vertical scan over salt cedar near the Rio Grande River. The variability in the data about the fitted line is due to the presence of discrete structures. If a large enough area is averaged, the mean value at each elevation converges to a logarithmic profile

a large number of measurements to determine the vertical water vapor gradient. The extension of the method to rough terrain presents issues relating to assumptions of horizontal homogeneity as well as the determination of the canopy top location (with respect to the lidar) and the direction of the normal to the surface.

The flux estimation method used assumes that in some region, which is generally taken to be 25 m in size, but may be any user selected value, the slope of the water vapor concentration in the z direction can be determined from a curve fit using all of the measurements of the water vapor concentration above that region. This assumes horizontal homogeneity inside the region and with the region immediately upwind that the aggregate of the values constitutes a measurement of the average condition over the region, and that the slope in water vapor concentration is the result of conditions inside that region. The effect of these assumptions will be discussed later.

A key capability of the lidar that is useful in estimating fluxes over complex terrain is the ability to determine the location of the canopy top. The lidar is sited so that it overlooks the experimental site and is thus able to determine the location of the canopy top for all of the canopy types. For the case of mixed terrain and canopy, the lidar is used to find the location of the canopy top in the range interval under investigation. Figure 7 is a conceptual drawing of how this is accomplished. The top of the canopy is found either from the abrupt change in the apparent water concentration or from the abrupt change in the elastic lidar signal which is also recorded along each line of sight. The location of the top of the canopy as a function of distance is determined using multiple lines of sight. A linear least-squares fit is made to determine the elevation and slope of the top of the canopy within the range interval under consideration.

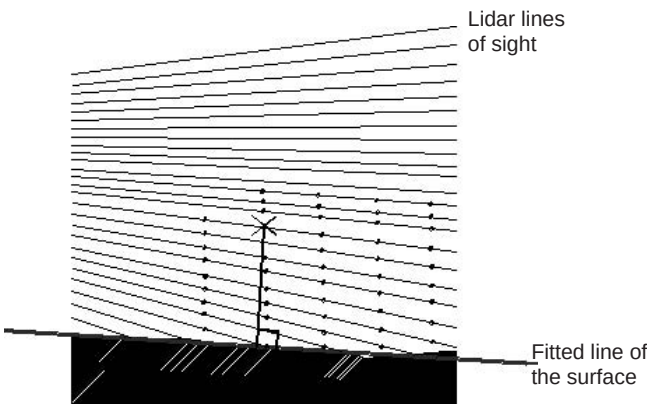


Figure 7 A conceptual drawing of a 25-m region and all of the lidar lines of sight within it. The location of a line approximating the surface is determined and the distance from each measured value to this line along a perpendicular to the line is calculated. All of the measured values of water vapor concentration are used to create vertical profiles similar to that in Figure 6. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

For an individual water vapor measurement, the distance from the measured point to the surface along a line perpendicular to the measured slope and elevation is used as the corrected height above the surface (see Figure 7). This means that the z direction is taken to be the direction perpendicular to the canopy top and not the vertical (gravitational) direction. The reasoning is that, near the surface, the flow of air will be parallel to the local surface and that dispersion of the water vapor released from the surface in the direction perpendicular to the mean flow is most important to the estimation of evapotranspiration (Kaimal and Finnigan, 1994).

For an individual scan, all of the measurements within a region are used to estimate the slope of the single line described by equation (4). While there is spread in the measurements at each height above the ground, the slope can be determined to an accuracy of a few percent. The spread in the measurements is due to the existence of coherent structures containing high and low water vapor concentrations. These structures can be seen in the two-dimensional plot shown in Figure 5. Currently, a measured value of the Obukhov length from point instruments is used to adjust for atmospheric stability. However, in practice, the use of this correction results in a small (usually on the order of 5% or less) change in the estimated flux. A severe limitation of this method is the lack of a u_* measurement for each 25-m region. In the ideal case, we divide the region into surface types and use a measured u_* typical of that region.

When using this method to determine fluxes, the optimum maximum height for which water vapor measurements are included must be determined. This corresponds to the height of the change in slope shown in Figure 6. While the largest possible distance over which the measurements are made leads to the greatest accuracy, measurements too close to the canopy top or so high that they are outside the inner region lead to erroneous estimates of the water concentration gradient. This height varies throughout the day so the method for determination must be dynamic and adjust to the existing conditions. In this analysis, the same height has been used over all of the canopy types, but the possibility exists that different heights would be appropriate over different canopies.

Uncertainty

The fractional uncertainty of the lidar flux measurements can be estimated using

$$\frac{\sigma E}{E} = \left[\left(\frac{\sigma u_*}{u_*} \right)^2 + \left(\frac{\sigma M}{M} \right)^2 + \left(\frac{\sigma \rho}{\rho} \right)^2 + \left(\frac{\sigma q}{q} \right)^2 \right]^{1/2}$$

where σu_*, σM, $\sigma \rho$, and σq are the uncertainties in the u_*, slope, air density, and water vapor concentration measurements respectively (Bevington and Robinson, 1992). The last term on the right is a contribution from a systematic uncertainty (or bias errors) in the lidar measurement of water vapor. While an individual measurement may be uncertain to the 3 to 4% level (a measure of the precision error), the determination of the mean concentration from a number of measurements (a measure of the bias error) is more accurate. This contribution is determined by the calibration error of the instrument and is a function not only of the lidar but also of the instrument(s) used to calibrate the lidar. For this reason, calibration is done with instruments traceable to the U.S. National Institute of Standards and Technology (NIST). As the range increases, the precision of the lidar degrades because of the decreasing signal-to-noise ratio caused by the r^{-2} decrease in the signal (the amount of noise is approximately constant with range), but the mean value of the measurements is maintained. Thus the variation in the data about the fitted line is generally observed to increase with distance, but since the mean value of the measurements is maintained, this should have a minimal effect upon the slope measurements and thus the estimated flux. The bias error in the mean value is taken to be less than 2%.

The value of the slope can be estimated with high certainty due to the large number of measurements used in fitting equation (6). The nominal uncertainty in the least-squares determined value of the slope is 1 to 2%. The air density is obtained from local measurements of temperature and air pressure. The uncertainty in the value of the air density is much less than 2%. The value of the friction velocity, u_*, is normally the primary source of uncertainty. While there are no reported estimates of the accuracy of $u*$ estimates from eddy correlation methods, it is reasonable to assume that the accuracy will be similar to those for eddy correlation flux measurements, which normally range from 5% to 25% (Wilson *et al.*, 2002). The value of the uncertainty of u_* is a function of the uncertainty in the measurements of u_* at a given point but it also contains a contribution from the assumption that a measurement at one point may be applied to a similar surface some distance away (the magnitude of which is highly site-specific). The uncertainty in a measurement of u_* is difficult to assess. While uncertainty estimates based upon the accuracy of the anemometer wind measurements result in estimates on the order of 5%, two anemometers a meter apart in nonideal conditions may have u_* values that differ by as much as 35% (although typical values are generally much less). For a typical measurement of the evaporative flux, the total uncertainty is determined almost totally by the uncertainty in u_* and leads us to estimate an overall uncertainty on the order of 15%. For areas far from u_* measurements, the uncertainty is potentially as much as twice as large.

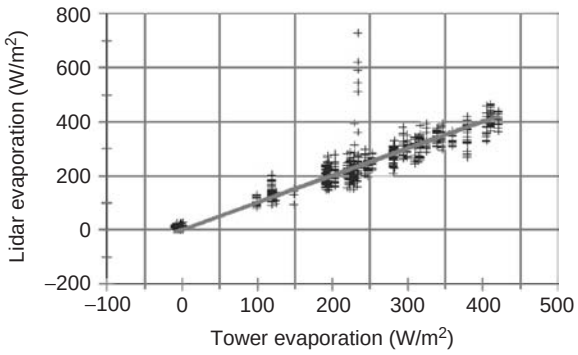


Figure 8 A comparison of eddy correlation evapotranspiration rates over a salt cedar canopy with lidar estimated evapotranspiration rates made in the same 25-m lidar pixel. The agreement is generally good and along the 1 : 1 line. The six point excursion was from an afternoon with exceptionally high winds in which the eddy correlation instruments may have been above the inner region

Figure 8 is a comparison of the latent heat flux estimates from eddy correlation measurements with lidar measurements made in the same 25-m region. These measurements represent half-hour averages. The differences between the two instruments are smaller when more data are available from the lidar for either temporal or spatial averaging. The lidar coordinate system is spherical, so that near the lidar more data are available with which to determine the vertical profile. The smaller the area to be covered, the faster the lidar can revisit the same locations, providing more data for analysis.

The evapotranspiration estimate obtained from each of the 25-m squares can be assembled into a map of the evapotranspiration rates over the scanned area. Figure 9 is a comparison of an evapotranspiration map over salt cedar and a high-resolution thermal infrared image of the same area at the same time. The prevailing wind is from the SW. There is a correlation between the dark grey areas in the thermal infrared image (which are areas of high sensible heat flux, low evapotranspiration) and regions of relatively low evapotranspiration rate (dark greys and blacks). The correlation is not perfect and there is generally a 50 to 75-m offset between the two images in the direction of the wind. As with conventional point instruments, the water vapor concentration in the air above a given point is a function of the emission rate in the upwind direction. As the height above the canopy increases, the farther upwind is the primary source area that determines the concentration. The offset distance seen in this comparison is consistent with footprint calculations for this site and conditions (Cooper *et al.*, 2003).

Limitations of the Method

Despite the successes that have been achieved, there remain significant questions about the method. The method used

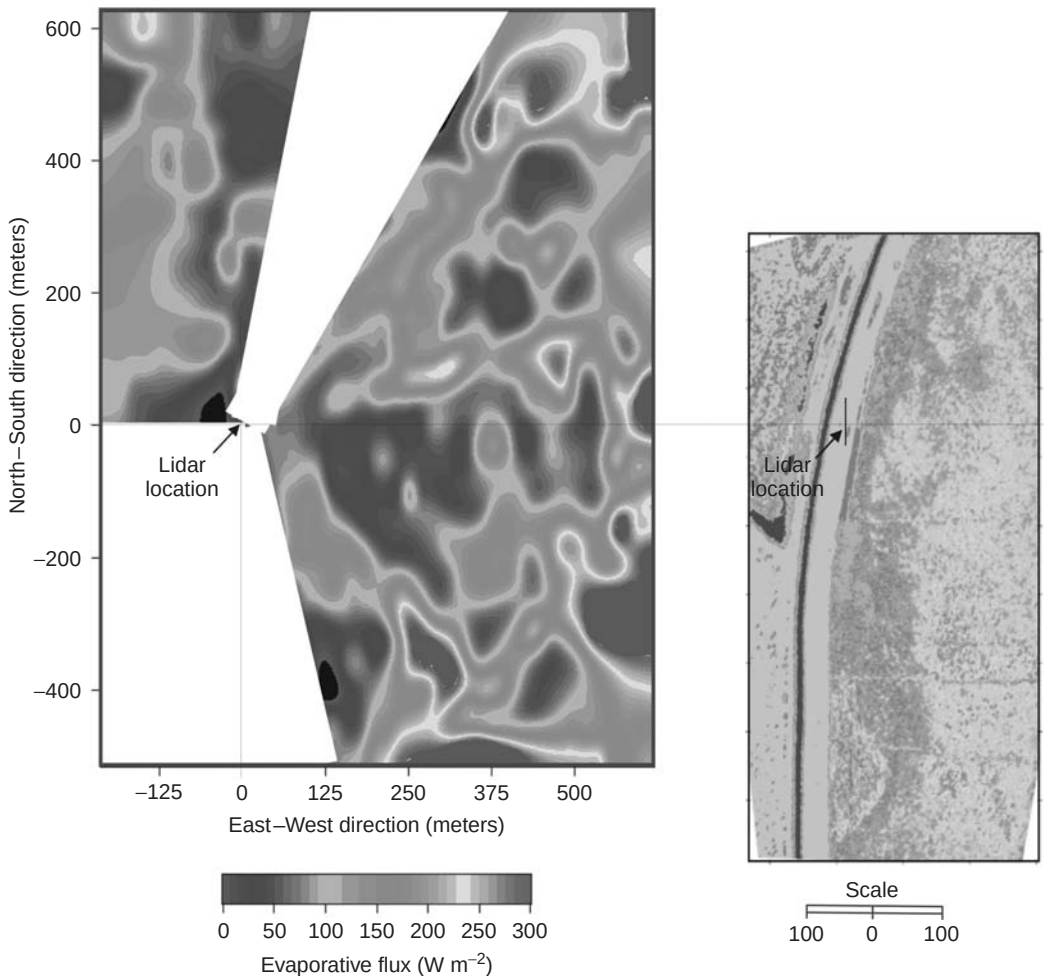


Figure 9 A comparison of an aerial thermal infrared photograph of the field site and a lidar-generated evapotranspiration map. Dark grey areas of the photograph indicate areas that are hot (which should have low evapotranspiration rates) while the lighter areas that are cool, which should have high evapotranspiration rates. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

here to develop maps of the evaporative flux in complex terrain assumes that, in some small region, the slope of the water vapor concentration in the direction perpendicular to the surface is governed by Monin–Obukhov similarity theory. This in turn assumes horizontal homogeneity inside the region and in the region immediately upwind that the aggregate of the values constitutes a measurement of the average condition over the region, and that the slope in water vapor concentration is the result of conditions in that region. Clearly, in transition areas where the canopy type or groundwater availability changes dramatically, these conditions will not apply.

Conditions may occur near areas with sharp transitions in which moist areas upwind alter the water vapor concentration near the canopy top so that it is not logarithmic with height. When this occurs, the methodology described here cannot be used. When the concentrations are not logarithmic with height, the flux estimation method does not always produce evapotranspiration estimates that are unreasonable and thus this condition can be found only by visual examination of the vertical scans. In complex terrain, changes in the canopy lead to changes in the evapotranspiration rate, which lead to changes in the water vapor concentration along the surface. The conservation of water vapor equation can be used to estimate the magnitude of the effect:

$$\frac{\partial q}{\partial t} + \overline{u}\frac{\partial q}{\partial x} + \overline{v}\frac{\partial q}{\partial y} + \overline{w}\frac{\partial q}{\partial z} = -\left[\frac{\partial}{\partial x}(\overline{u'q'}) + \frac{\partial}{\partial y}(\overline{v'q'}) + \frac{\partial}{\partial z}(\overline{w'q'})\right] + \kappa_v \nabla^2 q \quad (8)$$

where u, v, and w are the three components of wind velocity; κ_v is the diffusion coefficient for water molecules in air; and the overline indicates a time average and primed quantities represent fluctuations. Under most circumstances,

the diffusion term (the last term on the right), and terms involving $\overline{v}$, $\overline{v'q'}$, and $\overline{w}$ are small enough to be ignored. Because of advection, changes in evapotranspiration rates may result in flux divergence, particularly in the vertical direction, $\partial\overline{w'q'}/\partial z$. The size of this term can be estimated from the $\overline{u}\partial\overline{q}/\partial x$ and $\partial\overline{q}/\partial t$ terms in the conservation of water vapor equation. It is not uncommon to find horizontal gradients in water vapor concentration on the order of 0.2 g water per kg air in a 25-m by 25-m analysis region downwind of the riparian area. This can lead to a divergence of about 50 W m^{-2} per meter of height above ground. At this time, the effect of advection on the Monin–Obukhov flux method is unknown, but is a subject of current research. We have found that the corrections due to nonstationarity are, in general, small. The changes in water vapor concentration over a 10-min period are on the order of 0.3 to 0.4 g water per kg air or less. This results in a correction of less than 5 W m^{-2}.

Related to the question of advection is the question of the location of the source (also *known as the* footprint) for a measurement at a given height. This is a subject of considerable current interest (for example, LeClerc and Thurtell, 1990; Horst and Weil, 1994; Finn *et al.*, 1996; Horst, 1999). More than two-thirds of the measurements used in any given profile are below 8 m. On more than half of the profiles, the maximum height used is 8 m or less. The greatest curvature in the profile is found at heights less than 4 m, and it is those measurements that play the greatest role in determining the slope of the line. Using the methodology outlined by Horst and Weil (1994), one can estimate the upwind distance contributing to the flux at a given height. For a height of 8 m, the upwind distance, past which less than 20% of the flux is generated, is a factor of approximately 5 to 10 times the measurement height, or about 40 to 80 m. This would tend to indicate that the bulk of the flux in a given 25-m section is generated inside that section and the section immediately upwind. Thus it would be prudent to recognize that the flux locations as given by the methodology are not exact, but rather are somewhat diffuse in the upwind direction. In practice, this has not been an issue in that the estimated fluxes do not show many instances of large, abrupt changes.

Lastly, there remains the question of how well the measurements averaged over distances as short as 25 m and less than a minute in time can represent the average conditions. An individual scan will often show structure near the canopy top. In most cases, this will produce deviations both above and below the average value of the water vapor concentration, but which average to profiles that are logarithmic. Occasionally, there are plumes that contain water vapor concentrations that are significantly higher than normal. In such cases, the profiles are significantly altered and may no longer be logarithmic. At present, when such events are found, the evapotranspiration estimate from that 25-m section is discarded. No analysis method has been found that can incorporate such structures to produce an evapotranspiration estimate. A more detailed analysis of the structure of the water vapor concentrations and fluxes is presented in Cooper *et al.* (2000).

While limitations of the method exist, the amount and type of data provided by the lidar allows one to visually determine what is happening at a particular location that causes the estimates to be anomalous. The existence of visual two-dimensional information that allows one to correct for unusual circumstances is a very powerful asset and offers the potential for improved algorithms which may overcome the deficiencies in the current formulation. At present, these conditions require the intervention of a human analyst to determine the proper method of analysis to be made. However, because of the sheer volume of data that must be processed to use this methodology, it must be highly automated if it is to be truly efficacious.

DISCUSSION

The availability of three-dimensional water vapor concentrations, and two-dimensional evapotranspiration maps and their variation in time provides opportunities to study spatial processes that are difficult or impossible to accomplish with point instruments. For example, evaporation maps are currently being used to examine the relative contributions of near-surface soil moisture and canopy as well as the statistics of the spatial distribution. Similarly, three-dimensional water vapor concentrations have been used to measure the Monin–Obukhov stability correction functions and to estimate the Obukhov length, L.

Evapotranspiration Maps

The evapotranspiration maps can be used to study the spatial variability of evaporation in a number of different ways (Kustas *et al.*, 2004; Eichinger *et al.*, 2004). Figure 10 has examples of evapotranspiration maps made from 08 : 15 to 12 : 15 central daylight time, June 27, 2002 over corn and soybean fields in Iowa during the SMACEX experiment. Each map represents a half-hour average ($\pm$15 min from the given time). The average height of the corn was 1.37 m. The set of maps also includes a high-resolution (1.5 m pixel) multispectral digital image, presented as a false-color composite, of three spectral bands simulating the Landsat Thematic Mapper bands TM2 (green), TM3 (red) and TM4 (near-infrared), at the same scale, showing the canopy condition (Neale and Crowther, 1994; Cai and Neale, 1999). It can be seen that the cornfield canopy (north half of the image) had considerably more biomass than the soybean crop (south half of the image), which shows distinct patterns of low canopy cover and bare soil. While there are recurring correlations between the dark grey areas

in the aerial photograph (which are areas of high canopy cover) and regions of relatively high evapotranspiration rate and, conversely, the correlation is not perfect, nor consistent. This is not totally surprising in that the lidar senses the water vapor content in the atmosphere and is thus sensitive to fluctuations in the wind direction and the intermittent nature of turbulence. Early in the morning on the 27th, there was sufficient near-surface soil moisture so that evapotranspiration was at or near the potential rate and nearly constant over the two fields. By 10 a.m., the soil surface moisture was significantly depleted, but in small areas, it was possible to have evapotranspiration rates higher than the average. By noon, the near-surface water is exhausted and evapotranspiration in the soybean area is appreciably lower than that in the corn. Because the LAI of the soybeans is much less than 1, it is reasonable to assume that the spatial evapotranspiration rates are dominated by soil surface moisture considerations.

Spatial Structure of Evapotranspiration

Evaporative Structure Functions

It has been previously observed that the variance of soil moisture as a function of area follows a power law (Rodriguez-Iturbe *et al.*, 1995; Hu *et al.*, 1997). These investigators found that the spatial correlation remains unchanged with the scale of observation and follows a power law typical of scaling processes. The evapotranspiration maps in Figure 10 can be used to determine if this scaling extends to evapotranspiration and other atmospheric processes that depend on available soil moisture. While it stands to reason that evapotranspiration rates are governed by the availability of soil moisture, there are at least two reasons to expect some modulation. First, there is a great deal of evidence that, at least in the surface layer, the atmosphere acts as an integrator (Brutsaert, 1998). This concept underlies footprint studies in which the water vapor concentration at a given point is a weighted sum of contributions from an upwind region (Lenschow and Stankov, 1986; LeClerc and Thurtell, 1990; Horst and Weil, 1994; Finn *et al.*, 1996; Baldocchi, 1997; Mahrt, 1998; Horst, 1999). The size of the region is a function of the height of the point in question and the properties of the turbulent flow. The second reason is that the plant canopy modifies the evapotranspiration rate by intercepting solar energy, preventing the sun from warming the soil, and by transpiring itself. Separating the effects of canopy and canopy transpiration from the soil contribution has been the subject of a great deal of effort.

Structure functions have been used to examine the spatial variability of discretely measured data sets, most often on one-dimensional data sets. The structure function is calculated as

$$S(lag) = \overline{[E(\vec{r}) - E(\vec{r} - \vec{lag})]^2} \qquad (9)$$

where $\vec{r}$ is the location of a particular evapotranspiration measurement, $\vec{lag}$ is the distance between the two measurements, and the overline denotes averaging of all of the possibilities at each value of $\vec{lag}$. The structure function then provides information on the mean square differences between evapotranspiration measurements separated by various distances. The structure functions for the data sets shown in Figure 10 are typical of the types of data that are seen.

Figure 11 shows the structure functions for the evapotranspiration maps shown in Figure 10 for June 27. The red boxes represent the structure functions for the soybean field and the diamonds represent the corn field. Note that the corn and soybean functions have similar shapes and are offset from each other. The systematic offset is likely caused by the difference in the average evapotranspiration in the two regions. Note that the shapes of the structure functions are not a straight line, which would be anticipated on the basis of the previous work on soil moisture scaling albeit at larger scales. Except for the last 6 to 10 data points on the right side of each plot, each of the values was calculated from several thousand comparisons. The data is smooth and changes systematically. Previous work has indicated a power law profile (Rodriguez-Iturbe *et al.*, 1995; Hu *et al.*, 1997), although over considerably larger areas (~200 m to 20 km) and with considerably sparser range increments. While one could draw a straight line through each of the curves in Figure 11, the data appears to be more complex than what can be explained by a simple power law. Using data from the 1997 Southern Great Plains Experiment, Nykanen and Foufoula–Georgiou (2001) came to a similar conclusion that soil moisture patterns do not exhibit the same scaling patterns at all scales nor under all conditions.

The shapes of the structure functions shown in Figure 11 are more complex than a simple power law. At 8 am, when there is surface moisture available and the canopy cover is considerably sparser, the soybean field structure function is systematically larger than the corn field. The structure functions at 10 a.m. are larger and nearly the same for both crops. By 12 a.m., when the ground has dried, the soybean structure function is systematically lower than that of the corn. At distances longer than 650 m, nearly all decrease in magnitude. While the trends are systematic, these values are determined by fewer and fewer comparisons as the distances get longer, so that an anomalous measurement can have a significant effect. It may be that there is considerably more variability in evapotranspiration rates (and thus more shape and texture to the structure functions) during the periods when dry-down occurs than when the area is either at field capacity or at the wilting point. The slopes of the structure functions at 8 and 10 a.m. are nearly the same at short and medium distances, while the slopes of the two functions at 12 a.m. are smaller and similar to those found

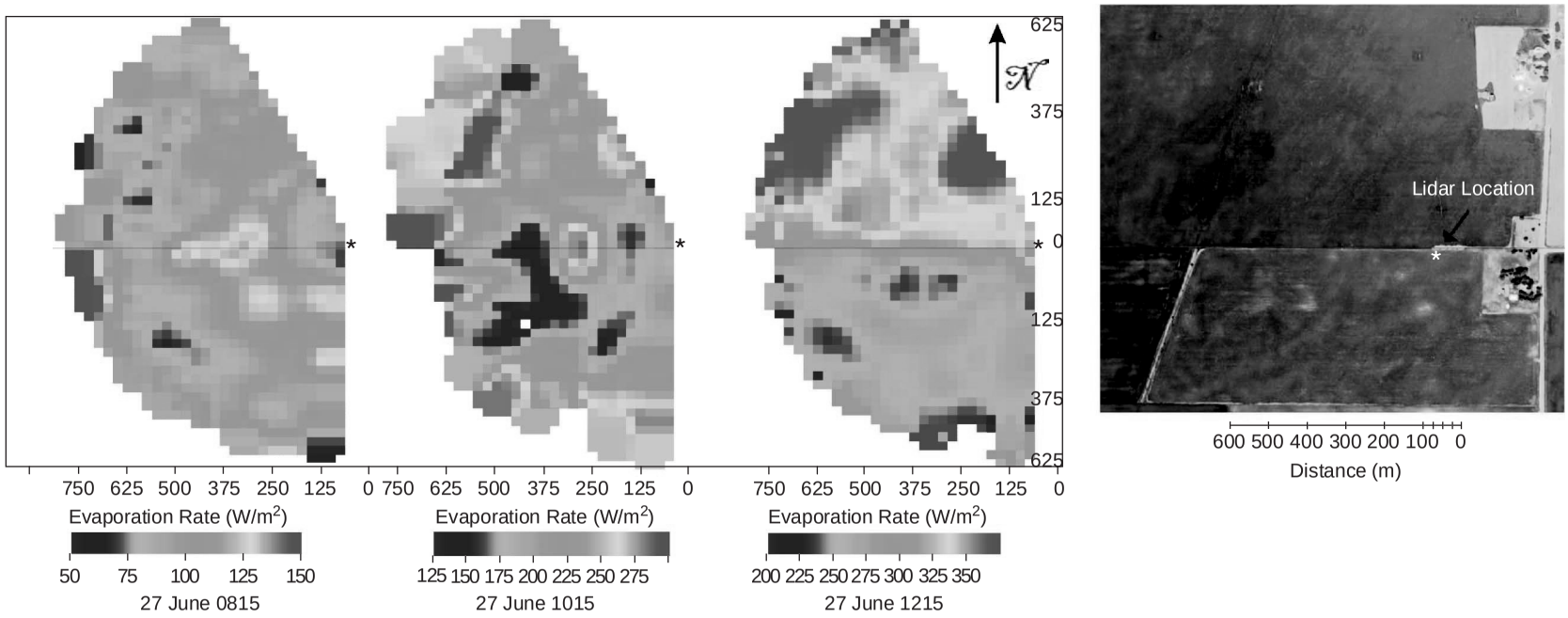


Figure 10 Three evapotranspiration maps of the area around the lidar at various times on June 27th. Light greys indicates areas of higher evapotranspiration and dark greys are lowest. In order to show the variability of the fields, the grey scales are different for each time. Soybeans were planted to the south of the lidar and corn to the north. The dividing line is the fence that can be seen in the photograph just below the lidar location. Also shown is an aerial three-band false-color composite of canopy reflectance at the same scale, for comparison. Dark greys are indicative of greater amounts of biomass. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

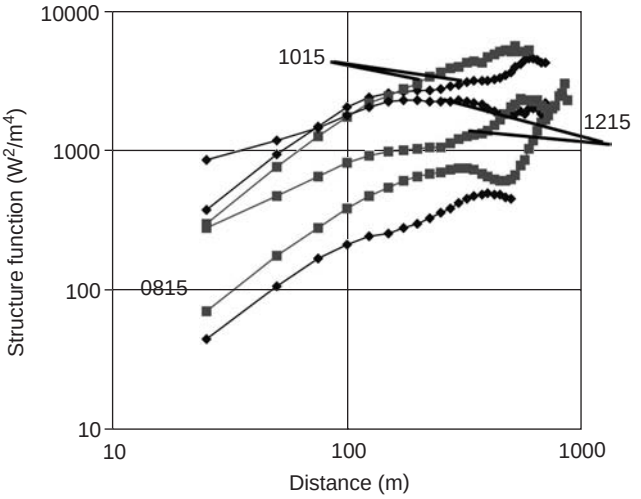


Figure 11 Structure functions for the evapotranspiration maps shown in Figure 7. Diamonds are from the corn and boxes from the soybean fields. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

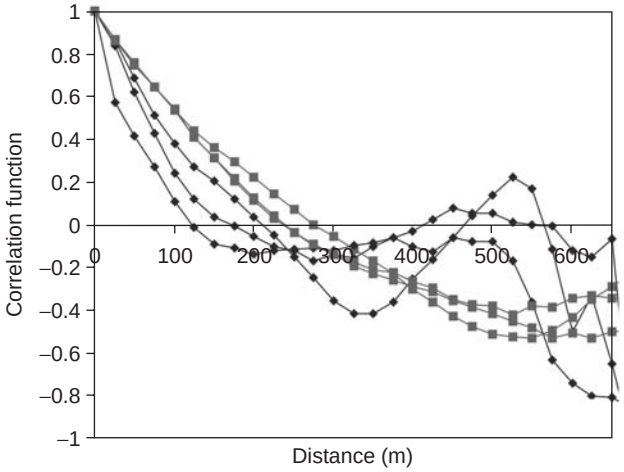


Figure 12 Correlation functions for the evapotranspiration maps shown in Figure 10. Diamonds are from the corn and boxes from the soybean fields. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

when there is little or no surface moisture. Clearly, the availability of near-surface soil moisture affects the details of the structure function.

Evaporative Covariance Functions

The covariance function also provides spatial information. The covariance function, calculated as

$$\mathrm{Cov}(lag) = \frac{\overline{((E(\vec{r}) - E_{\mathrm{ave}}) * (E'(\vec{r} + l\vec{a}g) - E'_{\mathrm{ave}}))^2}}{\sigma_E \sigma_{E'}} \tag{10}$$

where σ_E and $\sigma_{E'}$ are the standard deviations of the two sets of evapotranspiration measurements. The covariance function provides information on whether the deviations from the mean are correlated or anti-correlated at various distances. The distance at which the measurements are uncorrelated (the zero crossing point) provides a measure of the "typical" size of the regions over which evapotranspiration is similar.

Figure 12 shows the covariance functions for the three data sets from June 27. The shapes of the functions are similar, but with significantly different zero crossing points (125 to 275 m). The structure functions for the soybean fields on June 27 are nearly identical and possess a shape similar to those found when there is little or no water available at the soil surface. The cornfield, however, shows a systematic decrease in the location of the zero crossing, 200 m at 8 a.m., 175 m at 10 a.m., and 125 m at noon. While the size of the patches in the evapotranspiration maps seems to increase with time in Figure 10, the size of the differences is getting larger, causing the zero crossing point to decrease. The shapes of the correlation functions after the first zero crossing vary significantly. Since the soybean correlation function does not change with surface moisture availability but that for corn does suggests that the corn is significantly affecting the overall evapotranspiration rate in this dry-down period. In the soybeans with significantly lower canopy coverage, Figure 12 suggests that the availability of near-surface soil moisture does not affect the typical structure size.

Determination of the Obukhov Length

The requirement to measure the friction velocity, u^*, and Obukhov length, L, is a major limitation of the lidar method of evapotranspiration estimation. Efforts are being made to estimate the Obukhov length from measurements of the integral length scale derived from lidar measurements of water vapor. The integral length scale is the integral of a given variables' autocorrelation function in space which can be determined using two-dimensional lidar data. The integral length scale can also be estimated from meteorological variables using the empirical relationship (Cooper *et al.*, 2004):

$$\Lambda = \overline{u}\left[\frac{k(z-d)}{1.25u_*\left(1-\beta_1\frac{z}{L}\right)^a}\right] \tag{11}$$

where β_1 and a are empirically determined constants and k is the von Karman constant. Equation (11) can be simplified by noting that

$$\overline{u}(z) = \frac{u^*}{k}\left[\ln\left(\frac{z-d_0}{z_0}\right) - \psi_m\left(\frac{z-d_0}{L}\right)\right] \tag{12}$$

Substituting and solving for the Obukhov length gives

$$L = \frac{(z-d_0)\beta_1}{1-\left(\dfrac{\Lambda_{q\ \mathrm{lidar}}}{C_1(z-d_0)\left[\ln\left(\dfrac{z-d_0}{z_0}\right)-\Psi_m\left(\dfrac{(z-d_0)}{L}\right)\right]}\right)^a} \tag{13}$$

where C_1 has been substituted for the empirically determined constant 1.25 in equation (11). Values for the constants were determined using a least-squares fit to be $C_1 = 1.15$, $a = 0.25$, and $\beta_1 = 16$. The stability correction function for momentum, Ψ_m, for unstable atmospheres is given by

$$\psi_m\left(\frac{z}{L}\right) = 2\ln\left(\frac{1+x}{2}\right) + \ln\left(\frac{1+x^2}{2}\right) - 2\arctan(x) + \frac{\pi}{2}$$

$$x = \left(1 - 16\frac{z}{L}\right)^{1/4} \tag{14}$$

Equation (16) requires only knowledge of the integral length scale and the height of the measurements to determine L, albeit by an iterative process. Equation (15) can be compared to an empirically derived equation developed by Wilson *et al.* (1981)

$$\Lambda = 0.5\left(1 - 6\frac{(z-d)}{L}\right)^{1/4} \tag{15}$$

A comparison of lidar-derived values of L with those from conventional point instruments is shown in Figure 13. The data used for the estimates comes from water vapor measurements in the Bosque del Apache in New Mexico (Cooper *et al.*, 2004). Note that the bulk of the estimates fall within the ±15% limits. The method also compares favorably with the relationship developed by Wilson *et al.* (1981).

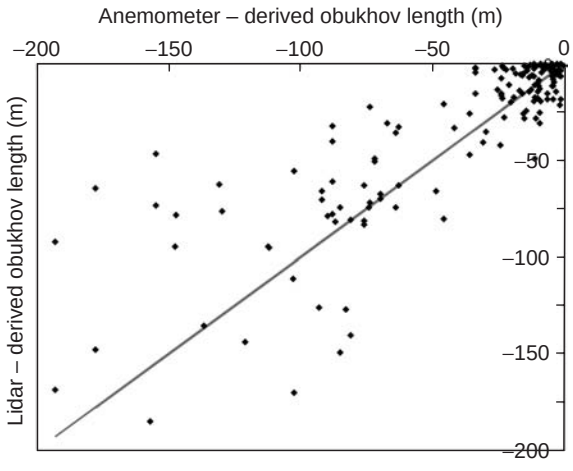


Figure 13 A comparison of lidar-derived Obukhov lengths and eddy covariance derived Obukhov lengths from two field campaigns in New Mexico. A one to one line is shown. The r^2 for the comparison is 0.71

Stability Correction Functions

Measurements of the stability correction functions are historically difficult to make because they require noninterfering measurements of the humidity profile at a range of heights inside the surface layer. A lidar RHI scan makes a large number of measurements at many heights over distances of several hundred meters. The concentrations at all heights in a given range interval are used in the analysis. An advantage of using the lidar is that a relatively large number of data points are obtained, so that one can calculate the degree to which the data fit the model, and thus estimate the uncertainty associated with a least-squares fit to the model. The desire to make accurate measurements of these functions was motivated by the suggestion by Kader and Yaglom (1990) that there are actually three distinct sublayers in convective conditions. An experiment was carried out at the Campbell Tract of the University of California, Davis in August, 1991 over an irrigated bare field (details can be found in Eichinger *et al.*, 1993b). An array of micro-meteorological instruments were used in the study to provide the surface heat and momentum flux values.

The water vapor concentration data from RHI scans (similar to those in Figure 5) were fitted using a least-squares technique to a second-order polynomial in $\ln(z-d_0)$ in a manner similar to Hogstrom (1988):

$$q = q_0 + A[\ln(z-d_0)] + B[\ln(z-d_0)]^2 \tag{16}$$

Derivatives of equation (16) were used to obtain values of dq/dz. Simultaneous values of u_*, L, and E were obtained from the meteorological instruments. These were used to calculate the values of ϕ_v (ζ) (where $\zeta = (z-d_0)/L$) with

$$\phi_v(\zeta) = -\frac{[ku_*(z-d_0)\rho]}{E}\frac{\mathrm{d}q}{\mathrm{d}z} \tag{17}$$

Lidar measurements of the water vapor stability function were made and the results averaged over a wide range of values of ζ. Values were averaged over intervals of 0.01 and the standard deviation of all the values in the interval were calculated. Because of the very large number of measurements (greater than 41 000), an uncertainty can be calculated in each interval of ζ as shown in Figure 14. The ranges of each of the sublayers postulated by Kader and Yaglom (1990) are marked in Figure 14 with a transition section separating each interval.

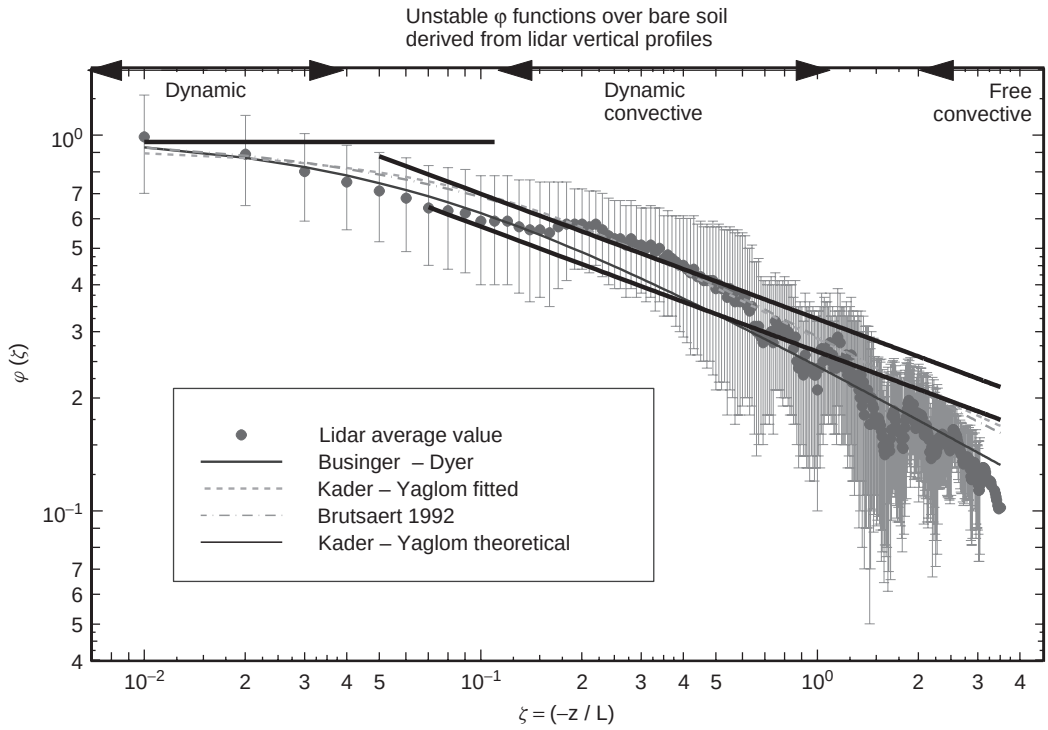


Figure 14 Comparison of lidar-derived stability correction functions to the traditional Businger–Dyer parameterization and the Kader and Yaglom parameterization. The error bars represent the standard deviations of the measurements. The small error bars on the largest values of ζ are indicative of a limited number of measurements and not necessarily decreased uncertainty. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The ϕ_v values are plotted in Figure 14 along with the Businger–Dyer function (Businger *et al.*, 1971; Dyer, 1974), the Kader–Yaglom sublayer models and the smooth Brutsaert function based on the Kader–Yaglom sublayer models. In the dynamic layer, the ϕ_v data clearly show a falloff in comparison to the constant value predicted by the Kader and Yaglom theory though they match the Businger–Dyer parameterization well. In the dynamic-convective sublayer there is a clear and distinct jump in the data which matches the ϕ_h function of Kader–Yaglom remarkably well on average. The Businger–Dyer curve underestimates the mean ϕ_v observations, though it falls within the standard deviation limits. In the free convective sublayer, the data are limited in quantity and range of values; nevertheless, they appear to follow the Businger–Dyer function better than the Kader–Yaglom function, which is remarkable since the Businger–Dyer functions were not based on much data in the free-convective limit. In the region beyond $\zeta = 2$, more measurements are required to better determine all of the scalar and momentum similarity functions.

As noted by Dyer and Bradley (1982), measurements in the free-convective portion of the atmosphere are subject to considerable variability, owing to lack of stationarity and homogeneity despite detrending of data and spatial or temporal averaging of the data. In fact, as pointed out by Dyer and Bradley, the atmosphere has its own statistics and does not necessarily follow average laws at all locations at all times. This is evident in the size of the standard deviations on the data in Figure 14 despite the relative accuracy of a large number of measurements. While the data indicate an improved understanding of the physical processes in the dynamic-convective and free-convective sublayers, given the small differences in the various similarity functions in comparison to the inherent uncertainties, changing from the Businger–Dyer parameterization is, in practice, probably unwarranted.

An additional result from the measurement of the ϕ_v values is the determination of, k, the von Karman constant. This value is normally taken as 0.40 and was used in Figure 14. Using the raw data and fitting the 12 values of ϕ_v below $\zeta = 0.11$ to a second-order polynomial in $\ln(\zeta)$, the y intercept can be determined. The intercept value is the inverse of k, which was found to be 0.37 ± 0.02. Given the scatter in the data, using a number for k other than 0.4 is unwarranted.

SUMMARY

Maps of the spatial distribution of evapotranspiration can be produced using spatial water vapor concentration data

from a scanning Raman lidar. The estimates of evapotranspiration rates from the lidar compare favorably with other estimates made using more conventional methods. The method described allows estimates of the evapotranspiration rate to be made with relatively small (25 m) spatial resolution over an area approaching a square kilometer. Because of the amount of data (25 to 40 Mbytes) and time required to perform the analysis, methods and criteria are currently under development to automate the entire analysis process over all of the azimuthal angles for a given averaging time period. Criteria are being developed to flag and ignore data that do not converge to a logarithmic profile and to not include data above the internal boundary region in the analysis. Automation of the analysis will allow near real time determination of the evapotranspiration estimates.

While there are significant limitations to the method, it remains a relatively direct method of estimating the fluxes in situations where conventional methods fail or when the topography makes it difficult to site instruments or field enough of them to achieve an areal average. A major limitation is related to changes in the topography, and how much of the atmosphere above a given site can be considered to be influenced by the surface and thus used to estimate the flux in that region. An advantage of this method is that the spatial water vapor measurements themselves can be used to determine the regions in which these problems may occur. The large number of water vapor measurements used to determine the slope also makes it possible to determine where the slopes change and thus limit the maximum height of the water vapor measurements used to determine the slope.

Efforts are currently underway to improve the range of the lidar system by increasing the photon efficiency of the system. This will make it possible to scan faster so that more scans can be repeated over a larger area. Efforts are also being made to add the ability to measure spatially resolved temperature using a Raman technique (Nedeljkovic *et al.*, 1993). The addition of temperature will allow determination of the partitioning of solar energy between sensible and latent heat fluxes using similar methods. The ability to estimate these fluxes in a spatial manner will enable progress in a number of fields that examine the role that the canopy plays in partitioning solar energy.

The method used here can provide reliable estimates of the evapotranspiration rate over a relatively large area with relatively fine spatial resolution. The method is a more direct method of estimating the fluxes than most remote sensing techniques that estimate the evapotranspiration rate as a residual. It is, however, limited in that it assumes a single measurement of the friction velocity is representative of the value over a large region. It also provides regular estimates throughout the day as opposed to intermittent satellite or aircraft measurements. This information can be used in a wide variety of ways to study the spatial variations in evapotranspiration caused by changes in soil type and moisture content, canopy type, and topography. Because of the extensive nature of the estimates in space and time, evaluation of the relative contributions of each of these can be determined. The type of measurements provided also provide the opportunity to span the scales between the footprint of point measurements and the kilometer scale measurements made by satellites.

Acknowledgments

The development of the Raman lidar technology and the associated analysis methods have involved a large number of people. In particular, the following people have made significant contributions: Dan Cooper, John Archuleta, Jennifer Nichols, John Prueger, Larry Hipps, Heidi Holder, Bill Kustas, Chris Neale, Rene Van't Land, and Larry Tellier. Without the assistance of Los Alamos National Laboratory, IIHR of the University of Iowa, and the National Soil Tilth Laboratory, this work could never have been done.

FURTHER READING

Ferrare R.A., Melfi S.H., Whiteman D. and Evans K. (1992) Raman lidar measurements of Pinatubo aerosols over southeastern Kansas during November-December 1991. *Geophysical Research Letters*, **19**, 1599–1602.

REFERENCES

Anderson M.C., Norman J.M., Diak G.R. and Kustas W.P. (1997) A two-source time integrated model for estimating surface fluxes for thermal infrared satellite observations. *Remote Sensing of the Environment*, **60**, 195–216.

Anderson M.C., Norman J., Mecikalski J., Torn R., Kustas W. and Basara J. (2004) A multi-scale remote sensing model for disaggregating regional fluxes to micrometeorological scales. *Journal of Hydrometeorology*, **5**, 343–363.

Ansmann A., Riebesell M., Wandinger U., Weitkamp C., Voss E., Lahman W. and Michaelis W. (1992) Combined Raman elastic backscatter lidar for vertical profiling of moisture, aerosol extinction backscatter and lidar ratio. *Applied Physics B*, **55**, 18–28.

Baldocchi D. (1997) Flux footprints within and over forest canopies. *Boundary-Layer Meteorolology*, **85**, 273–292.

Bevington P. and Robinson D. (1992) *Data Reduction and Error Analysis for the Physical Sciences*, *2nd edition*, McGraw-Hill, Inc, New York, 328 pp.

Brutsaert W. (1982) *Evaporation Into the Atmosphere*, D. Reidel: Dordrecht.

Brutsaert W. (1998) Land-surface water vapor and sensible heat flux: spatial variability, homogeneity, and measurement. *Water Resources Research*, **34**, 2433–2442.

Businger J.A., Wyngaard J., Izumi Y. and Bradley F. (1971) Flux-profile relationships in the atmospheric surface layer. *Journal of the Atmospheric Sciences*, **28**, 181–189.

Cai B. and Neale C. (1999) A method for constructing three-dimensional models from airborne digital imagery. *Proceedings of the 17th Biennial Workshop on Color Photography and Videography in Resources Assessment*, American Society of Photogrammetry and Remote Sensing, and Department of Environmental and Resources Sciences, University of Nevada: Reno, May 5–7, 1999.

Connell B.H. and Miller D.R. (1995) An interpretation of radiosonde errors in the atmospheric boundary layer. *Journal of Applied Meteorology*, **34**, 1070–1081.

Cooney J. (1970) Remote measurements of atmospheric water vapor profiles using the Raman component of laser backscatter. *Journal of Applied Meteorology*, **9**, 182–184.

Cooney J. (1983) Uses of Raman scattering for remote sensing of atmospheric properties of meteorological significance. *Optical Engineering*, **22**, 292–301.

Cooney J., Orr J. and Tomasetti C. (1969) Measurements separating the gaseous and aerosol components of laser atmospheric backscatter. *Nature*, **224**, 1098–1099.

Cooney J., Petri K. and Salik A. (1985) Measurements of high resolution atmospheric water vapor profiles by the use of a solar-blind Raman lidar. *Applied Optics*, **24**, 104–108.

Cooper D., Eichinger W., Archuleta J., Hipps L., Kao J., Leclerc M., Neale C. and Prueger J. (2003) Spatial source-area analysis of three-dimensional moisture fields from lidar, eddy covariance, and a footprint model. *Agricultural and Forest Meteorology*, **114**, 213–234.

Cooper D., Eichinger W., Archuleta J., Hipps L., Kao J. and Prueger J. (2004) Lidar derived integral length scales, the Monin-Obukhov length, and spatially resolved latent energy fluxes. *Journal of the Atmospheric Sciences*, (in press).

Cooper D., Eichinger W., Hipps L., Kao J., Reisner J., Smith S., Schaeffer S. and Williams D. (2000) Spatial and temporal properties of water vapor and flux over a riparian canopy. *Agricultural and Forest Meteorolology*, **105**, 161–183.

Cooper D., Eichinger W., Hynes M., Keller C., Lebeda C. and Poling D. (1996) High resolution properties of the equatorial pacific marine atmospheric boundary layer from lidar and radiosonde observations. *Journal of the Atmospheric Sciences*, **53**, 2054–2075.

Dyer A.J. (1974) A review of flux-profile relationships. *Boundary Layer Meteorology*, **7**, 363–372.

Dyer A.J. and Bradley E.F. (1982) An alternative analysis of flux-gradient relationships at the 1976 ITCE. *Boundary Layer Meteorology*, **22**, 3–19.

Ehret G., Kiemle C., Renger W. and Simmet G. (1993) Airborne remote sensing of tropospheric water vapor with a near-infrared differential absorption lidar system. *Applied Optics*, **32**, 4534–4551.

Eichinger W.E., Cooper D., Chen C., Hipps J., Kao J. and Prueger J. (2000) Estimation of spatially distributed latent heat flux over complex terrain from a Raman lidar. *Agricultural and Forest Meteorology*, **105**, 145–159.

Eichinger W., Cooper D., Forman P., Griegos J., Osborne M., Richter D., Tellier L. and Thornton R. (1999) The development of a scanning Raman water vapor lidar for boundary layer and tropospheric observations. *Atmospheric and Oceanic Technology*, **16**, 1753–1766.

Eichinger W., Cooper D., Holtkamp D., Karl R., Moses J., Quick C. and Tiee J. (1993a) Derivation of water vapor fluxes from lidar measurements. *Boundary Layer Meteorology*, **63**, 39–64.

Eichinger W., Cooper D., Parlange M. and Katul G. (1993b) The application of a scanning, Water-Raman lidar as a probe of the atmospheric boundary layer. *IEEE Transactions on Geoscience and Remote Sensing*, **31**, 70–79.

Eichinger W., Cooper D., Hipps L., Kustas W., Neale C. and Prueger J. (2004) Spatial and temporal variation in evapotranspiration using Raman lidar. *Journal of Hydrometeorology*, (in press).

Ferrare R.A., Melfi S.H., Whiteman D. and Evans K. (1995) A comparison of water vapor measurements made by Raman lidar and radiosondes. *Journal of Atmospheric and Oceanic Technology*, **12**, 1177–1195.

Ferrare R.A., Melfi S.H., Whiteman D.N., Evans K.D. and Leifer R. (1998) Raman lidar measurements of aerosol extinction and backscattering, I. methods and comparisons. *Journal of Geophysical Research*, **103**, 19663–19672.

Finn D., Lamb B., LeClerc M. and Horst T. (1996) Experimental evaluation of analytical and lagrangian surface layer flux footprint models. *Boundary Layer Meteorology*, **89**, 283–308.

Geiz A., Ehret G., Schwiesow R., Davis K. and Lenschow D. (1999) Water vapor flux measurements from ground-based vertically pointed water vapor differential absorption and doppler lidars. *Journal of Atmospheric and Oceanic Technology*, **16**, 237–250.

Goldsmith J.E.M., Blair F.H., Bisson S.E. and Turner D. (1998) Turn-key Raman lidar for profiling atmospheric water vapor, clouds, and aerosols. *Applied Optics*, **37**, 4979–4990.

Grant W.B. (1991) Differential absorption and Raman lidar for water vapor profile measurements: a review. *Optical Engineering*, **30**, 40–48.

Higdon N.S., Browell E., Ponsardin P., Grossmann B., Butler C., Chyba T., Mayo M., Allen R., Heuser A.W., Grant W., *et al.* (1994) Airborne differential absorption lidar system for measurements of atmospheric water vapor and aerosols. *Applied Optics*, **33**, 6422–6438.

Hinkley E.D. (Ed.) (1976) *Laser Monitoring of the Atmosphere*, Springer-Verlag: New York.

Hogstrom U. (1988) Non-dimensional wind and temperature profiles in the atmospheric surface layer: a re-evaluation. *Boundary Layer Meteorology*, **42**, 55–78.

Horst T. (1999) The footprint for estimation of atmosphere-surface exchange fluxes by profile techniques. *Boundary Layer Meteorology*, **90**, 171–188.

Horst T. and Weil J. (1994) How far is far enough? The fetch requirements for micrometeorological measurement of surface fluxes. *Journal of Atmospheric and Oceanic Technology*, **11**, 1018–1024.

Hu Z., Islam S. and Cheng Y. (1997) Statistical characterization of remotely sensed soil moisture images. *Remote Sensing of the Environment*, **61**, 310–318.

Kader B.A. and Yaglom M. (1990) Mean fields and fluctuation moments in unstably stratified turbulent boundary layers. *Journal of Fluid Mechanics*, **212**, 637–662.

Kaimal J. and Finnigan J. (1994) *Atmospheric Boundary Layer Flows*, Oxford University Press: New York.

Kiemle C., Ehret G., Geiz A., Davis K., Lenschow D. and Oncley S. (1997) Estimation of boundary layer humidity fluxes and statistics from airborne differential absorption lidar (DIAL). *Journal of Geophysical Research*, **102**, 29189–29204.

Kovalev V. and Eichinger W. (2004) *Elastic Lidar: Theory, Practice and Analysis Methods*, John Wiley & Sons: New York.

Kustas W., Hatfield J. and Prueger J. (2004) The Soil Moisture Atmosphere Coupling Experiment (SMACEX): background, hydrometeorlogical conditions and preliminary findings. *Journal of Hydrometeorology*, (in press).

LeClerc M. and Thurtell G. (1990) Footprint prediction of scalar fluxes using a Markovian analysis. *Boundary Layer Meteorology*, **52**, 247–258.

Lenschow D.H. and Stankov B.B. (1986) Length scales in the convective boundary layer. *Journal of the Atmospheric Sciences*, **43**, 1198–1209.

Mahrt L. (1998) Flux sampling errors for aircraft and towers. *Journal of Atmospheric and Oceanic Technology*, **15**, 416–429.

Measures R.M. (1984) *Laser Remote Sensing*, Wiley Interscience: New York.

Melfi S.H. (1972) Remote measurements of the atmosphere using Raman scattering. *Applied Optics*, **11**, 1605–1611.

Melfi S.H., Lawrence J.D. and McCormick M.P. (1969) Observation of Raman scattering by water vapor in the atmosphere. *Applied Physics Letters*, **15**, 295–297.

Moran S., Jackson R., Raymond L., Gay L. and Salter P. (1989) Mapping surface energy balance components by combining landsat thematic mapper and ground-based meteorological data. *Remote Sensing of Environment*, **30**, 77–87.

Neale C.M.U. and Crowther B. (1994) An airborne multispectral video/radiometer remote sensing system: development and calibration. *Remote Sensing of Environment*, **49**, 187–194.

Neale C., Hipps L.E., Prueger J.H., Kustas W.P., Cooper D.I. and Eichinger W.E. (2001) Spatial mapping of evapotranspiration and energy balance components over riparian vegetation using airborne remote sensing. *International Symposium on Remote Sensing and Hydrology 2000*, IAHS: Publication No.267, pp. 311–315.

Nedeljkovic D., Hauchecorne A. and Chanin M. (1993) Rotational Raman lidar to measure the atmospheric temperature from the ground to 30 km. *IEEE Transactions on Geoscience and Remote Sensing*, **31**, 90–101.

Nykanen D.K. and Foufoula–Georgiou E. (2001) Soil moisture variability and scale–dependency of nonlinear parameterizations in coupled land–atmosphere models. *Advances in Water Resources*, **24**, 1143–1157.

Renault D. and Capitini R. (1988) Boundary-layer water vapor probing with a solar-blind Raman lidar: validations, meteorological observations and prospects. *Journal of Atmospheric and Oceanic Technology*, **5**, 5–15.

Renault D., Pourney C. and Capitini R. (1980) Daytime Raman lidar measurements of water vapor. *Optics Letters*, **5**, 232–235.

Rodriguez-Iturbe I., Vogel G., Rigon R., Entekhabi D., Castelli F. and Rinaldo A. (1995) On the spatial organization of soil moisture fields. *Geophysical Research Letters*, **22**, 2757–2760.

Senff C., Bösenberg J. and Peters G. (1994) Measurement of water vapor flux profiles in the convective boundary layer with lidar and radar-RASS. *Journal of Atmospheric and Oceanic Technology*, **11**, 85–93.

Sherlock V., Hauchecorne A. and Lenoble J. (1999) Methodology for the independent calibration of Raman backscatter water vapor lidar systems. *Applied Optics*, **38**, 5816–5838.

Stull R. (1988) *An Introduction to Boundary Layer Meteorology*, Kluwer Academic Publishers: Boston.

Wilson K., Goldstein A., Falge E., Aubinet M., Baldocchi D., Berbigier P., Bernhofer C., Ceulemans R., Dolman H., Field C., *et al.* (2002) Energy balance closure at FLUXNET. *Agricultural and Forest Meteorology*, **113**, 223–243.

Wilson J.D., Thurtell G.W. and Kidd G.E. (1981) Numerical simulation of particle trajectories in inhomogeneous turbulence, III: comparison of predictions with experimental data for the atmospheric surface layer. *Boundary Layer Meteorology*, **21**, 443–463.

52: Estimation of Surface Temperature and Surface Emissivity

THOMAS J SCHMUGGE

Formerly of: USDA/ARS Hydrology & Remote Sensing Lab, Beltsville, MD, US & Currently at: College of Agriculture, New Mexico State University, Las Cruces, NM, US

Measurements of thermal radiation at infrared wavelengths (7 to 14 μm*) yield much information about the land surface. The primary use of these observations is for surface temperature determination as the emissivity is usually close to one. For this purpose, it is fortuitous that the peak in the thermal emission occurs in an atmospheric transmission window. In addition, there are variations in the emissivity of minerals and soils in the 7–14* μm *region, which can be interpreted for identification purposes. These also produce significant emissivity variations for arid or desert areas, where there are considerable exposed soils and rocks. The emissivity for vegetative canopies has been found to be close to one, with little spectral variation. Applications of the derived surface temperature to study the surface energy balance and to estimate the energy fluxes from the land surface are discussed.*

INTRODUCTION

Measurement of the thermally emitted radiation at various wavelengths from the earth's surface can yield much useful information about parameters such as surface soil moisture and temperature. These two parameters are very important for the study of the land–atmosphere interaction. The partitioning of the net radiation into latent and sensible heating components is determined by the moisture available in the soil for the evapotranspiration process. The magnitude of the surface temperature is the result of this partitioning. If sufficient moisture is available, the surface temperature will be close to that of the air. If not, the surface temperature will rise above that of the air and contribute more strongly to the sensible heat flux.

Climatological observations of surface temperature are also an important indicator of climate change. Jin and Dickinson (2002) have analyzed data from the AVHRR (Advanced Very High Resolution Radiometer) on the National Oceanographic and Atmospheric Administration (NOAA) series of polar-orbiting weather satellites for the period 1982 to 1998 and observed a 0.4 °C increase per decade. This increase is consistent with that found for the global land air temperature and is even somewhat higher. The Moderate Resolution Imaging Spectroradiometer MODIS sensor on National Aeronautic and Space Administration (NASA's) Earth Observation System (EOS) will provide a good tool for extending this time series (Wan *et al.*, 2002).

To estimate surface temperatures, radiation at wavelengths around 10 μm is used because the peak intensity of the thermal emission occurs in this region for terrestrial temperatures (≈300 K), and in addition the atmosphere is relatively transparent and the observed variations in the intensity of radiation are mainly related to surface temperature. However, it must be kept in mind that variations in emissivity of 1% yield brightness temperature variations of 0.3–0.6 K depending on sky radiance. To estimate surface soil moisture, radiation at much longer wavelengths (10's of cm) is used. Changes in the intensity of the emitted radiation at these wavelengths are primarily due to the variation of the emissivity with the moisture content of the soil. This effect results from the large dielectric contrast between water and dry soils. This topic is discussed in a later article (*see* **Chapter 54, Estimation of Surface Soil Moisture Using Microwave Sensors, Volume 2**).

The monitoring of the land-surface fluxes at regional spatial scales is recognized as important for applications such

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

as the modeling of atmospheric behavior, the monitoring of water resources, and the estimation of plant or crop water use (Jackson, *et al.*, 1977). Surface temperature is a key boundary condition for many land-surface atmosphere models, and is a variable that can be measured by satellite remote sensing on a global basis. Indeed, considerable progress has been made in accounting for atmospheric effects and the variation in surface emissivity so that the uncertainty in radiometric surface temperature measurements from satellites is within 1–2 degrees (Norman *et al.*, 1995). This accuracy has been verified with recent results from satellite-borne sensors (Hook *et al.*, 2003; Wan *et al.*, 2002, 2004).

An example of a thermal infrared image from a satellite sensor, the Advanced Spaceborne Thermal Emission and Reflection radiometer (ASTER) (Yamaguchi *et al.*, 1998) onboard NASA's Terra satellite is presented in Figure 1. This figure presents land-surface temperatures derived using ASTER satellite imagery covering an area around the USDA-ARS Grazinglands Research Facility in El Reno, Oklahoma on September 4, 2002. The spatial distribution of land-surface temperature, T_{SURF}, reflects some significant differences in land-cover conditions this time of the year (September), with large areas of bare soil and wheat stubble from harvested winter wheat fields, the light colored areas with temperatures up to 57 °C, and small areas of irrigated crop lands and water bodies, the dark areas with temperature down to 36 °C. Amongst these extremes, are grasslands used for cattle grazing. This type of spatially distributed information is very useful for evaluating spatial patterns of *ET* over large areas, as will be demonstrated later in this article.

When one looks at this image, it is obvious that the large variations in surface brightness temperatures, T_{B}, arise from differences in the surface energy balance for the different surfaces. Recall that T_{B} is a measure of the radiation emitted from the surface, and is directly related to the temperature of the surface and its emissivity. In Figure 1, the hotter fields (white, with $T_{\mathrm{B}} > 50\,^{\circ}\mathrm{C}$) are either bare soil or fields with sparse, newly emerging vegetation. The cooler fields (black, with $T_{\mathrm{B}} \approx 36$–$40\,^{\circ}\mathrm{C}$) are completely vegetated grasslands or forested riparian zones along the streams. These temperature contrasts among fields with different vegetation conditions imply a different partition of the incoming solar energy into latent and sensible heat components. In general, cooler temperatures indicate that there is sufficient moisture available so that most of the incoming energy goes into latent heat or evaporation, while hotter temperatures indicate that most of the incoming energy goes into the sensible or convective heating of the atmosphere. The problem is to quantify these fluxes in terms of the remotely sensed T_{B}. There is a long history on the use of T_{B} to monitor these surface fluxes (e.g. Jackson *et al.*, 1977; Price, 1980; Soer, 1980; Seguin and Itier, 1983) and in this article, we will describe the contributions remotely sensed thermal infrared data can make towards quantifying these fluxes.

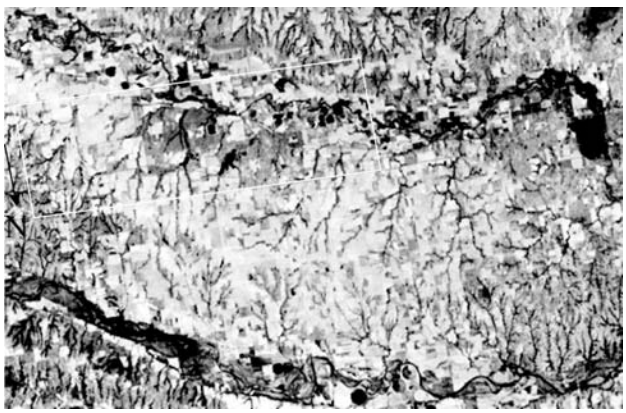


Figure 1 ASTER Thermal Infrared (TIR) imagery for a region in central Oklahoma, just west of Oklahoma City. The data was taken on September 4, 2000 at 17:34 GMT and are for band 13 of ASTER, $\lambda = 10.7\,\mu\mathrm{m}$. The temperatures range from 36 C (black) −57 °C (white). The spatial resolution is 90 m

THERMAL INFRARED RADIATION

The intensity of the thermal radiation from an object is described by the Planck blackbody relationship given as a function of frequency in equation (1a) and as a function of wavelength in equation (1b):

$$L_{BB}(\nu, T) = \frac{2h\nu^3/c^2}{\exp(h\nu/kT) - 1} \tag{1a}$$

$$L_{BB}(\lambda, T) = \frac{2hc^2/\lambda^5}{\exp(hc/\lambda kT) - 1} \tag{1b}$$

where h is Planck's constant ($6.626\mathrm{A}10^{-34}$ joule sec), c is the speed of light and k is Boltzmann's constant ($1.381\mathrm{A}10^{-23}$ joule K^{-1}). The units are $\mathrm{W\,m^{-2}\,sr^{-1}}$ per Hz for equation (1a) or $\mathrm{W\,m^{-2}\,sr^{-1}}$ per meter for equation (1b). Equation (1a) is plotted in Figure 2 for several temperatures. In this figure, the reflected solar radiation is also plotted for albedos of 0.1 and 1 to show that at the wavelengths where the terrestrial thermal radiation peaks, that is, at $\lambda \approx 10\,\mu\mathrm{m}$, the reflected solar radiation is several orders of magnitude weaker. The wavelength for the maximum in the thermal emission curve is given by the Wien displacement law where $\lambda_{\max} T \approx 2.898\mathrm{A}10^{-3}\,\mathrm{mK}$. Thus for a temperature of 300 K, $\lambda_{\max} \approx 9.7\,\mu\mathrm{m}$. The measurements around this wavelength should therefore yield information on surface temperature without any contamination by reflected solar radiation. The cosmic background radiation at $T = 3\,\mathrm{K}$ is also shown for reference.

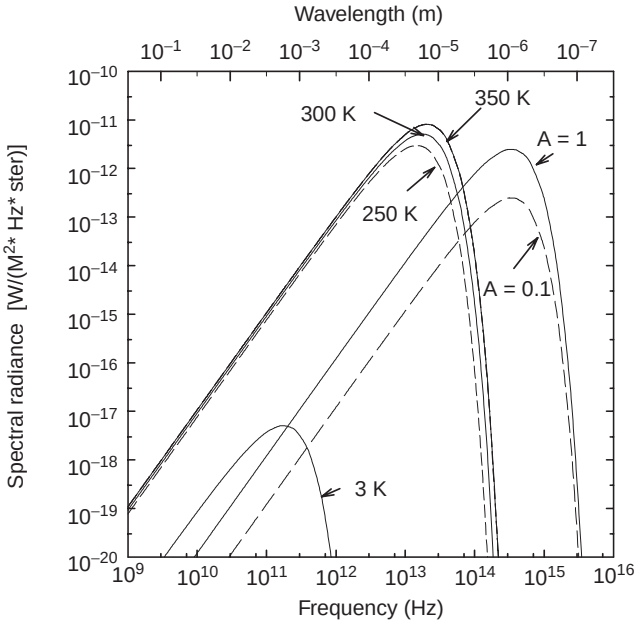


Figure 2 The spectral radiance of a blackbody according to equation (1a) for a typical range of terrestrial temperatures and the cosmic background (3 K). The reflected solar irradiance with indicated values of the albedo (α) is also plotted for comparison

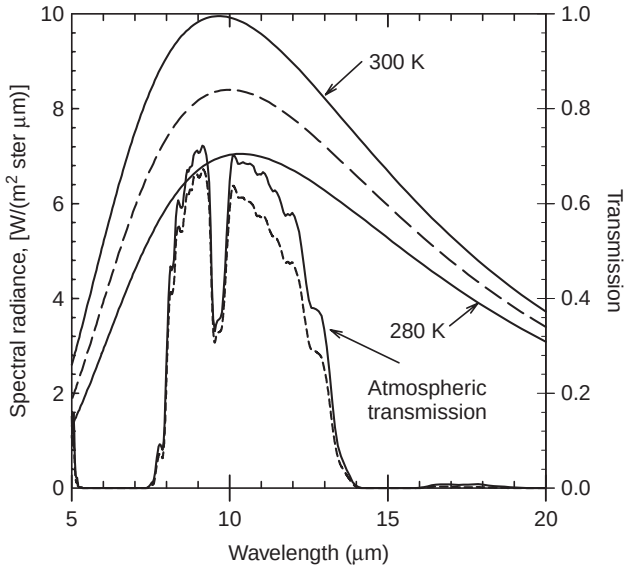


Figure 3 The spectral radiance of a blackbody according to equation (2) for typical terrestrial temperatures and the atmospheric transmission for the United States standard midlatitude summer atmosphere with two moisture levels: the upper curve is for a columnar water content of 2.9 g cm^{-2} and the lower is for 3.5 g cm^{-2}

In the infrared, we usually speak in terms of wavelength, while in the microwave, frequency and wavelength are used interchangeably. This perhaps results from the techniques used to quantify the waves; in the infrared wavelengths can be measured, but not frequency, while in the microwave both frequency and wavelength can be measured. It should be noted that in the infrared, frequency is usually expressed in terms of wave number, that is, $\nu = 1/\lambda$, in cm^{-1}. In this case equation, (1a) is usually represented as

$$L_{BB}(\nu, T) = \frac{\in C_1 \nu^3}{\exp(C_2 \nu / T) - 1} \tag{2}$$

where C_1 is 1.19104×10^{-8} W/(m^2 sr cm^{-4}), C_2 is 1.4388 cmK, and ν is expressed in wave number (cm^{-1}) $\in$ and is the emissivity. For a perfect emitter or blackbody $\in = 1$, and for real surfaces $\in < 1$.

Atmospheric Effects

In Figure 3, we have plotted equation (2) for the wavelength range 5–20 μm at temperatures of 280, 290, and 300 K, that is, near the low range of terrestrial temperatures. At these temperatures, the peak emission occurs in the 8–10 μm range of wavelength. In this figure, we have also plotted the clear sky atmospheric transmission calculated with the Modtran4 path radiance model (Berk *et al.*, 1998; Berk *et al.*, 1999) for the midlatitude summer atmosphere, assuming the radiometer is at satellite altitude. While the atmosphere is relatively transparent in the 8–12 μm range compared to adjacent wavelengths, there is still significant attenuation. As seen in Figure 3, there is only 60–70% transmission with a major dip at about 9.5 μm due to ozone absorption. With the exception of this dip, water vapor is the dominant absorber in the 8–12 μm window. This is mainly due to what is called the *water vapor continuum*, since only in the 8–9 μm range are there any relatively strong absorption lines. Thus, the magnitude of the atmospheric effect will depend on the water vapor content of the intervening atmosphere. The atmospheric transmission for a 20% increase in water vapor is also plotted to indicate this sensitivity to water vapor. This unknown or uncertain atmospheric contribution is one of the problems for the remote sensing of surface temperature at infrared wavelengths. This, of course, is in addition to clouds that will totally obscure the surface at visible and infrared wavelengths.

The radiation values given in Figure 3 are those which would be observed right at the surface; the relation between the radiation, L_i, seen by a radiometer on a satellite or aircraft and the surface temperature is:

$$L_i = [\varepsilon_i \cdot L_{BB_i}(T) + (1 - \varepsilon_i) \cdot L_{\mathrm{atm}_i}(dn)] \cdot \tau_i + L_{\mathrm{atm}_i}(up) \tag{3}$$

where the subscript i indicates the integral of these quantities over the bandwidth for channel i of the radiometer, L_{BBi} is the Planck function given by equation (1), the

L_{atm}'s are the upward and downward components of the atmospheric radiation, and τ is the atmospheric transmission. The values of L_{atm} and τ can be calculated using a model for atmospheric path radiance such as Modtran4, however, it would be preferable to eliminate the atmospheric effects using the multispectral information.

Several approaches have been developed for eliminating atmospheric effects in the estimation of sea surface temperature from space using multichannel thermal data. The technique used with the AVHRR data from the NOAA series of polar-orbiting satellites involves the differential water vapor absorption in the 10–13 μm window with the 11.5–13 μm portion being more strongly absorbed, the so-called *split-window technique* (McClain *et al.*, 1983; Grassl, 1989). However, this assumes that the surface emissivity is constant over this spectral band, which is not the case for land surfaces (Becker, 1987; Grassl, 1989). In order to better understand thermal infrared radiation, we need to study the behavior of the emissivity of terrestrial materials or surfaces.

Owing to the lack of adequate atmospheric profile observations, the development of alternative approaches such as so-called "split-window" methods would be more operationally applicable (e.g. Price, 1984). These split-window methods employ two channels at slightly different wavelengths, λ^1 and λ^2 in equation (4) and (5) to essentially eliminate (using a few approximations) the need for estimating the atmospheric transmission and radiances. This approach has been used quite successfully over oceans where the emissivity is close to one and its spectral variation is small. Over land, this is not the case and the split-window methods are sensitive to uncertainty in the emissivities in the two channels; for example, at a brightness temperature 300 K, a difference, $\{\in\}^1 - \{\in\}^2 \sim 0.01$ can yield an error in land-surface temperature of ~2 K (Price, 1984), which can be corrected for, if the emissivities are known. There has been much recent work on the use of these split-window techniques using the two thermal channels of the (AVHRR) instrument on the NOAA series of meteorological satellites (Wan and Dozier, 1989; Becker and Li, 1990, Kerr *et al.*, 1992, Prata, 1994; Coll and Caselles, 1997). This approach is being used with the multispectral thermal infrared data from the MODIS, onboard the NASA Terra satellite. They use a database of land-surface emissivities based on land cover to correct for emissivity effects (Wan and Dozier, 1996; Wan *et al.*, 2002, 2004) and have been rather successful.

Infrared Emissivity

There has been much use of the spectral variation of infrared emissivities for geological purposes; here we are more interested in the emissivities of soils and vegetation. This discussion is based on papers by Salisbury and D'Aria (1992a,b) which present emissivity spectra in the 8–12 μm range for soils and vegetation. These results were based on laboratory and field measurements made with an interferometer. The reader is referred to a special issue of Remote Sensing of Environment of emissivity (Salisbury, 1992).

To extract the emissivity from the radiance measurements, the assumption is made that the emissivity is unity somewhere in the wave band observed. A blackbody curve is fitted to that portion of the spectra to obtain the temperature. The emissivity is then determined by dividing the observed radiance by the blackbody radiance at that temperature (Kahle and Alley, 1992). However, to do this with field measurements it is necessary to take into account the reflected sky radiation as indicated in Figure 4, where the true and apparent emissivity are plotted as functions of wavelength for a case with significant downwelling radiance. Figure 4a shows the downwelling atmospheric radiance calculated with Modtran4 for a tropical atmosphere with 4.2 g cm^{-2} of water vapor. The upper three curves show surface radiances; the top is for a blackbody at 300 K, the lowest of the three is for the emission from a quartz rich sandy soil, and the middle curve is the upwelling radiance at the surface, that is, the term in brackets in equation (3) which includes the reflected downwelling radiance. When the latter two curves are divided by the blackbody curve, the true and apparent emissivities are obtained as shown in lower portion of the figure. This example shows the importance of measuring the downwelling sky radiance in emittance measurements. Note that for the quartz sand shown here, the emissivity is about 0.8 at $\lambda = 9.2$ μm.

An example of the emissivity variation of soils is given in Figure 5 where spectral variation for three soils from the United States are presented. The soils data are from the Jornada Experimental Range in New Mexico and their spectra were measured at the Jet Propulsion Laboratory. The soils shows the pronounced effect of quartz (SiO_2) (light sand), and gypsum on the emissivity spectra for the two soils with emissivity being less than 0.9 for $\lambda <$ 9.5 μm for quartz or gypsum. Note that in the 10–12 μm range, the emissivity is high ~0.95 and relatively constant. The response functions for the 5 channels of the ASTER radiometer are also shown.

The White Sands National Monument in New Mexico provides an opportunity to obtain data over a surface of relatively homogeneous composition. The White Sands gypsum dunes cover an area of about 20 by 30 km in south central New Mexico. We have a number of good ASTER observations covering the gypsum dunes. The radiances for a 2 by 2 pixel area (180 by 180 m) were analyzed for several dates. The emissivity results are presented in Figure 6 as the filled symbols for eight observations between May 2000 and December 2002. The emissivities were extracted from the atmospherically corrected radiances for the 5 ASTER bands using the

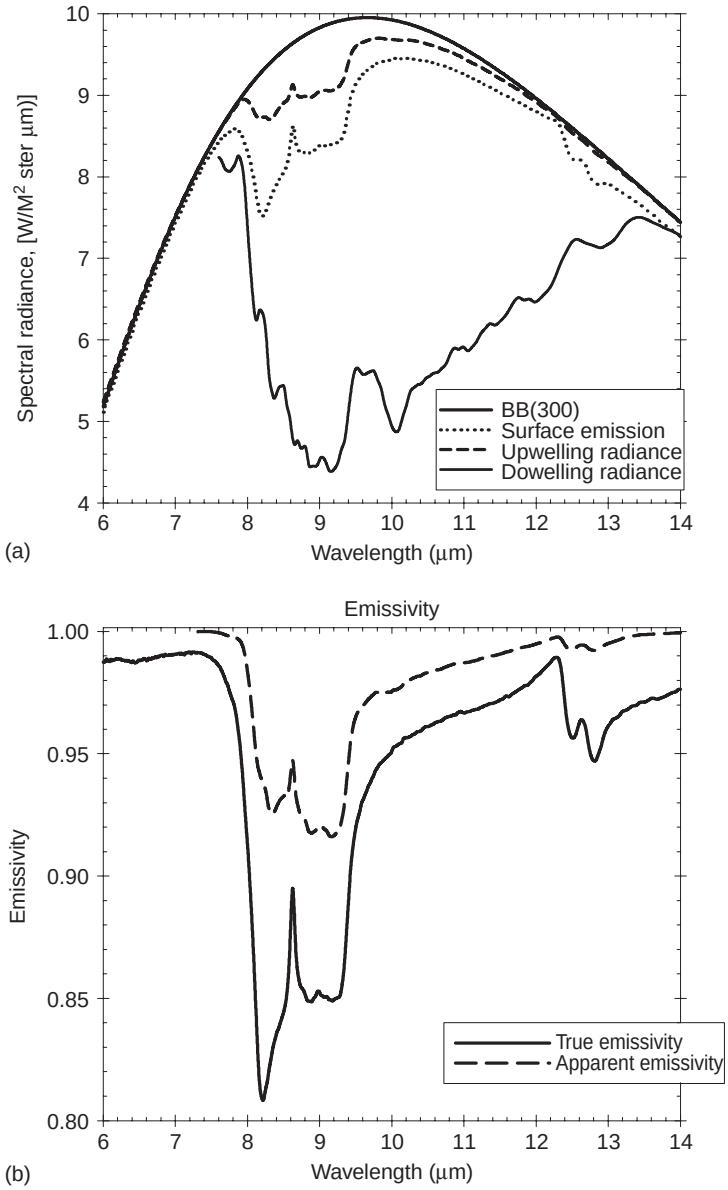


Figure 4 The effect of downwelling atmospheric radiance on apparent emissivity. (a) Radiances; the surface radiance components to equation (3), the bottom curve is the downwelling radiance, the top curve is the blackbody radiance from equation (2), the curve below that is upwelling radiance at the surface, that is, the term in brackets in equation (3) and final curve is the emission from the surface. (b) Emissivity; the true and apparent emissivity

Temperature Emissivity Separation (TES) developed for use with ASTER data (Gillespie *et al.*, 1998). TES relies on an empirical relation between the range of emissivities and the minimum value for the 5 ASTER bands. It works best for areas with significant spectral contrast in the emissivity, for example, arid regions with minimal vegetation cover. The lab results are the open squares and were obtained for each ASTER channel by integrating the product of the ASTER response and gypsum emissivity curve shown in Figure 5. The ASTER results show excellent agreement with the laboratory results and with each other for the center 3 bands. The agreement is particularly clear for the low-emissivity 8.6 μm channel. Bands 10 (8.29 μm) and 14 (11.29 μm) show the biggest differences. Band 10 is the one with the strongest atmospheric effects and the differences may indicate inadequate atmospheric correction. The cause for the band 14 differences is uncertain at the present time.

The open diamond symbols are the results from a CIMEL CE312 (Legrand *et al.*, 2000; Brogniez *et al.*, 2003) radiometer, which has approximately the same 5 spectral bands as ASTER. The shortest wavelength band of the

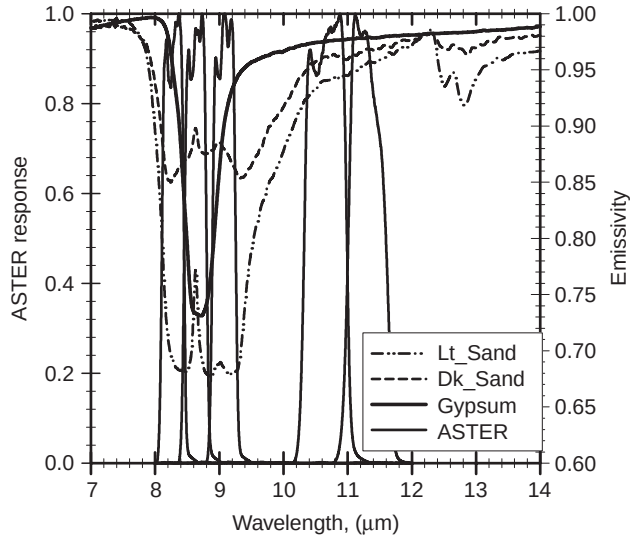


Figure 5 Laboratory measurements of the emissivity for three soils from the United States with the filter functions for the six TIMS channels. The soils are: (i) Brown sand with 0.5% organic carbon, 2% clay, and 96% quartz; (ii) White gypsum dune sand with 0% organic carbon, 0% clay, and 99% gypsum with a trace of quartz; and (iii) Black loam with 6.6% organic carbon, 30% clay, and 56% quartz

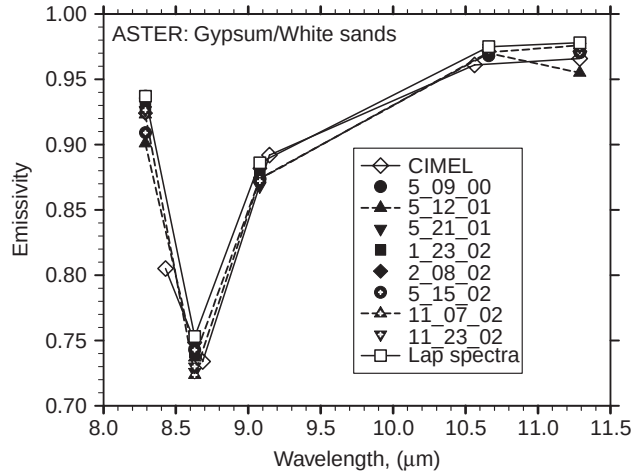


Figure 6 ASTER emissivity results for a gypsum site at the White Sands National Monument in New Mexico. The open symbols are for laboratory (squares) and ground measurements with the CIMEL radiometer (diamonds). The filled symbols are the ASTER results for eight days. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CIMEL is broader than that of the ASTER band, and as a result shows a much lower emissivity than those observed by ASTER band 10. For the other channels, the agreement is very good. CIMEL results for the bare soil at the other sites were also in good agreement with lab measurements. This agreement is an indication validity of TES for surfaces with large emissivity contrasts. The agreement of the ASTER results with both the CIMEL and lab measurements for the gypsum sands is a demonstration that these data can be used to estimate the surface emissivity over large areas. Additional results from the other sites also show good agreement. These results indicate that the TES algorithm appears to work as well with the data from space as it did with the aircraft data presented earlier (Schmugge *et al.*, 2002). This is encouraging for the application of the technique for mapping emissivity over large areas. Examples of this are presented in the papers by Ogawa *et al.* (2002, 2003) where the window emissivity is mapped for a large (400×1500 km) desert area in North Africa and to all of North Africa using a combination of ASTER and MODIS data (Ogawa and Schmugge, 2004).

In addition to minerals and soils Salisbury and D'Aria (1992a), present spectra for a number of terrestrial materials. Of most interest to us, are the results for vegetative components such as green and senescent foliage. For the green foliage, the reflectances are generally less than 5%, emissivity >0.95, while for senescent vegetation the reflectance is generally higher. As expected, the soil litter has low reflectance. When these components are combined to form a vegetative canopy, the results become more complicated because of scattering within the canopy. The net result is simpler because multiple scattering by the leaf surfaces with their low reflectance yields a canopy emissivity very close to one with little spectral variation. This was observed for the prairie grass in FIFE by Palluconi *et al.* (1990) who found an emissivity of 0.99 ± 0.01, with no spectral variation. In Hydrologic and Atmospheric Pilot Experiment-Modelisation du Bilan Hydrique (HAPEX-MOBILHY), Schmugge *et al.* (1991) found little or no spectral variation over a coniferous forest or an oat field and over the Tiger-Bush site in HAPEX-Sahel (Schmugge *et al.*, 1998), using Thermal Infrared Multispectral Scanner (TIMS) data from an aircraft platform. These results for vegetation are in substantial agreement with modeling studies of Norman *et al.* (1990) who derived a value of 0.99 for the emissivity of the grass canopy in FIFE.

SURFACE FLUXES

Energy and Moisture Balance

Evapotranspiration

To understand better how thermal infrared observations can contribute to the determination of the surface fluxes, let us consider the basic energy and moisture balance equations. In the absence of advection or precipitation, the energy balance at the land surface is given by:

$$R_n - G - H - LE = 0 \tag{4}$$

where R_n is the net radiation, G the soil heat flux, H the sensible heat flux, and LE the latent heat or moisture flux into the atmosphere. Here we are treating the heat fluxes (G, H, and LE) away from the surface as being positive. The net radiation is the sum of the incoming and outgoing short and long-wave radiation fluxes:

$$R_n = (1 - \alpha) \cdot R_s + (1 - \varepsilon) \cdot R_{L\downarrow} - \varepsilon \sigma T^4 \tag{5}$$

where α is the surface albedo, R_S is the incoming solar radiation, $R_{L\downarrow}$ is the incoming long-wave radiation, ε is the surface emissivity, and T is the surface temperature in Kelvins.

A later article (*see* **Chapter 50, Estimation of the Surface Energy Balance, Volume 2**) discusses the estimation of the surface evapotranspiration (LE) flux using remotely sensed data and provides an excellent background and introduction to the topic. The Surface Energy Balance System (SEBS) is presented there with examples. This model is based on the Surface Energy Balance Algorithms for Land (SEBaL) model of Bastiaanssen *et al.* (1998a,b). Here we will present an alternate approach for the estimation of LE with remotely sensed data.

This alternate approach considers the soil and vegetation contribution to the total or composite heat fluxes and the soil and vegetation temperatures to the radiometric temperature measurements in a so-called "Two-source" Modeling (TSM) scheme (Norman *et al.*, 1995b). One of the more common ways in estimating ET is to rearrange equation (4) solving for the latent heat flux, LE, as a residual in the energy balance equation for the land surface, namely,

$$LE = R_N - G - H \tag{6}$$

Remote sensing methods for estimating these radiation components of R_N are described in Kustas and Norman (1996). Typically with reliable estimates of solar radiation, differences between remote sensing estimates and observed $R_N - G$ are within 10%.

The largest uncertainty in estimating LE comes from computing H. In resistance form, the relationship between H and the surface-air temperature difference is expressed as

$$H = \rho \cdot C_p \frac{T_0 - T_A}{R_A} \tag{7}$$

where T_0 is the aerodynamic surface temperature (T_0 is the temperature satisfying the traditional expressions for the resistances; see Norman and Becker, 1995), T_A is the near-surface air temperature, D is air density, C_P is the specific heat of air, and R_A is the aerodynamic resistance. Since T_0 cannot be measured, the surface radiometric temperature is often substituted in equation (10) and is frequently rewritten as (e.g. Stewart *et al.*, 1994),

$$H = \rho C_P \frac{T_R(\theta) - T_A}{R_A + R_{EX}} \tag{8}$$

where R_{EX} is the so-called "excess resistance", which attempts to account for the nonequivalence of T_0 and $T_R(\theta)$. The radiometric temperature observations, $T_R(\theta)$, at some viewing angle θ, are converted from satellite brightness temperatures and are an estimate of the land-surface temperature, T_{SURF}. Thus, equations (6)–(8) offer the possibility of mapping surface heat fluxes on a regional scale if R_A and R_{EX} can be estimated appropriately. R_{EX} has been related to the ratio of roughness lengths for momentum, z_{OM}, and heat, z_{OH}, and the friction velocity u_* having the form (e.g. Stewart *et al.*, 1994),

$$R_{EX} = k^{-1} \ln\left(\frac{Z_{OM}}{X_{OH}}\right) u_*^{-1} \tag{9}$$

where $k \approx 0.4$ is von Karman's constant. This is the classical definition which addresses the fact that momentum and heat transport from the roughness elements differ (Brutsaert, 1982), but is just one of several that have been formulated (e.g. McNaughton and Van den Hurk, 1995). There have been numerous efforts in recent years to apply equations (8) and (9) and determine the behavior of R_{EX} or z_{OH} for different surfaces, but no universal relation exists (Kustas and Norman, 1996). Large spatial and temporal variations in the magnitude of z_{OH} have been found. Nevertheless, solving the LE with the approach summarized in equations (6)–(9) is still widely applied.

The TSM scheme (Norman *et al.*, 1995b) allows for equation (7) to be recast into the following expression

$$H = \rho C_P \frac{T_R(\theta) - T_A}{R_R} \tag{10}$$

where R_R is the radiometric-convective resistance given by (Norman *et al.*, 1995b),

$$R_R = \frac{T_R(\theta) - T_A}{\dfrac{T_C - T_A}{R_A} + \dfrac{T_S - T_A}{R_A + R_S}}. \tag{11}$$

T_C is the canopy temperature, T_S is the soil temperature, and R_S is the soil resistance to heat transfer. An estimate of leaf area index or fractional vegetation cover, f_C, is used to estimate T_C and T_S from $T_R(\theta)$,

$$T_R(\theta) \approx (f_C(\theta) T_C^4 + (1 - f_C(\theta)) T_S^4)^{1/4} \tag{12}$$

where $f_C(\theta)$ is the fractional vegetative cover at radiometer viewing angle θ, and R_S is computed from a relatively simple formulation predicting wind speed near the soil

surface (Norman *et al.*, 1995b). With some additional formulations for estimating canopy transpiration, and the dual requirement of energy and radiative balance of the soil and vegetation components, closure in the set of equations is achieved. Through model validation studies, revisions to the original two-source formulations have been made (Kustas and Norman, 1999 and 2000; Kustas *et al.*, 2001, Kustas *et al.*, 2004) which improved the reliability of flux estimation under a wider range of environmental conditions. The modifications include: (i) replacing the commonly used Beer's Law expression for estimating the divergence of the net radiation through the canopy layer with a more physically based algorithm; (ii) adding a simple method to address the effects of clumped vegetation on radiation divergence and wind speed inside the canopy layer, and radiative temperature partitioning between soil and vegetation components; (iii) a scheme for adjusting the magnitude of the Priestley and Taylor (1972) coefficient, α_{pt}, used in estimating canopy transpiration for advective and stressed canopy conditions; and (iv) developing a new soil aerodynamic resistance formulation whose magnitude is a function of both convective (temperature) and mechanical (wind) turbulent transport.

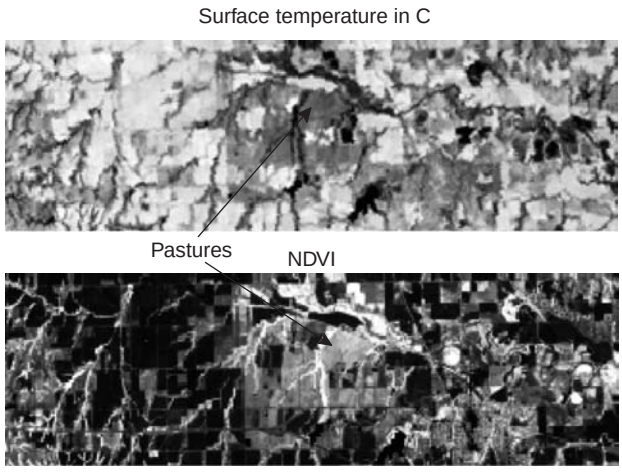


Figure 7 Surface temperatures and *NDVI* values for the portion of Figure 1 outlined by the white box. The area is approximately 30 by 8 km. The temperatures range from 36–57 °C and the *NDVI* values range from −0.1–0.5

Examples of Flux Estimation

An example of an application of the TSM approach for estimating daily *ET* is illustrated in Figures 7 and 8 for the September 04, 2000 ASTER data. Images of $T_R(\theta)$ and Normalized Difference Vegetation Index (*NDVI*) computed from the ASTER red and near-infrared reflectance data and the multispectral thermal infrared data are given in Figure 7 (French *et al.*, 2002; French *et al.*, 2003). The *NDVI* image shows bare soil with values ∼0.0 (light gray) and senescent grazing lands with values ∼0.2 (dark gray). Surface temperature imagery distinguishes bare soils at 55–60 °C (light gray) and from grazing lands at ∼45 °C (medium gray). These remote sensing inputs, in combination with TSM, resulted in 90 m scale estimates of *H* and *LE*. Comparison with ground-based energy balance observations using the Bowen ratio technique show underestimated *H* and *Rn*, but close agreement for *LE*. Assuming constant daytime evaporative fraction, instantaneous values can be integrated to yield daily evaporative flux estimates given in Figure 8. The latter range from about 5 mm day^{-1} for the grazing lands and 8 mm day^{-1} for the water bodies. Many of the low *ET* rates are from fields that are either bare soil or contain wheat stubble from the summer–winter wheat harvests, which generally have the highest $T_R(\theta)$ and $0 \leq$

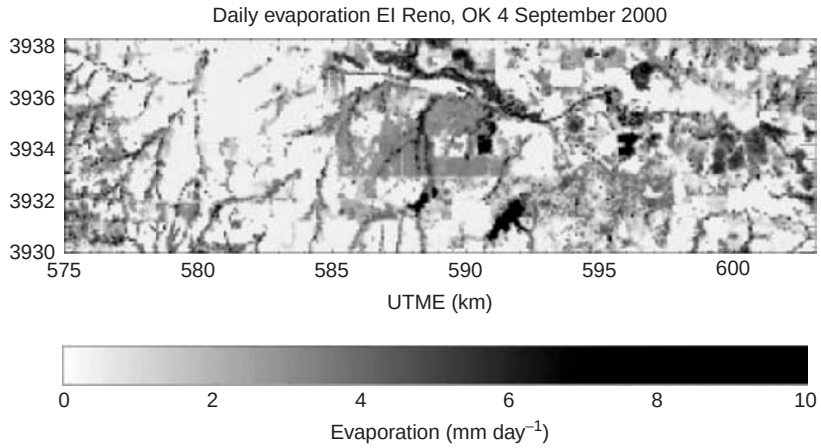


Figure 8 Estimated daily evapotranspiration over the El Reno, Oklahoma study area using ASTER observation at about 11 : 30 am. A drought occurred during the summer of 2000, and by September many of the fields were either plowed or contained senescent vegetation. In this image, 0–1 corresponds to dry soils, 4 to grass lands, and 6–8 to lakes, ponds, and riparian zones

$NDVI \leq 0.1$. Higher ET rates come from grassland sites ($NDVI \geq 0.2$) with the highest rates over irrigated crop fields and riparian areas along streams where $NDVI \geq 0.4$ and water bodies where $NDVI \leq 0$ (Figure 8).

DISCUSSION

In this article, we have reviewed the physics of the thermal radiation emitted from the land surface including the effects of the atmosphere and the surface emissivity. An approach for using the remotely sensed surface temperature to estimate the surface energy fluxes is presented and an example of its application is presented.

FURTHER READING

Price J.C. (1983) Estimating surface temperatures from satellite thermal infrared data. A simple formulation for the atmospheric effect. *Remote Sensing of Environment*, **13**, 353–361.

Shuttleworth W.J. (1991) Evaporation models in hydrology. In *Land Surface Evaporation: Measurement and Parameterization*, Chap. 6 Schmugge, T.J. and Andre, J.-C. (Eds.), Springer-Verlag: New York.

REFERENCES

Bastiaanssen W.G.M., Menenti M., Feddes R.A. and Holtslag A.A.M. (1998a) A remote Sensing surface energy balance algorithm for land (SEBAL) 1. Formulation. *Journal of Hydrology*, **212–213**, 198–212.

Bastiaanssen W.G.M., Pelgrum H., Wang Y., Ma Y., Moreno J.F., Roerink G.J. and van der Wal T. (1998b) A remote sensing surface energy balance algorithm for land (SEBAL) 2. Validation. *Journal of Hydrology*, **212–213**, 213–229.

Becker F. (1987) The impact of spectral emissivity on the measurements of land-surface temperature from a satellite. *International Journal of Remote Sensing*, **8**, 1509–1522.

Becker F. and Li Z.-L. (1990) Towards a local split-window over land surfaces. *International Journal of Remote Sensing*, **11**, 369–393.

Berk, A., Anderson G.P., Acharya P.K., Bernstein L.S., Chetwynd J.H., Matthew M.W., Shettle E.P. and Adler-Golden S.M. (1999) *MODTRAN4 User's Manual*. Air Force Research Laboratory Report, p. 97.

Berk A., Bernstein L.S., Anderson G.P., Acharya P.K., Robertson D.C., Chetwynd J.H. and Adler-Golden S.M. (1998) MODTRAN cloud and multiple scattering upgrade with application to AVIRIS. *Remote Sensing of Environment*, **65**, 367–375.

Brogniez G., Pietras C., Legrand M., Dubuisson P. and Haeffelin M. (2003) A high-accuracy multiwavelength radiometer for in situ measurements in the thermal infrared. Part II: behavior in field experiments. *Journal of Atmospheric and Oceanic Technology*, **20**(7), 1023–1033.

Brutsaert W. (1982) *Evaporation into the Atmosphere, Theory, History and Applications*, D. Reidel: Norwell.

Coll C. and Caselles V. (1997) A split-window algorithm for land-surface temperature from AVHRR data: validation and algorithm comparison. *Journal of Geophysical Research*, **102**(D14), 16697–16712.

French A.N., Schmugge T.J. and Kustas W.P. (2002) Estimating evapotranspiration over El Reno Oklahoma with ASTER imagery. *Agronomie: Agriculture and Environment*, **22**, 105–106.

French A.N., Schmugge T.J., Kustas W.P., Brubaker K.L. and Prueger J. (2003) Surface energy fluxes over El Reno, Oklahoma, using high-resolution remotely sensed data. *Water Resources Research*, **39**(6), 1164, doi:10.1029/2002WR001734.

Gillespie A., Rokugawa S., Matsunaga T., Cothern J.S., Hook S. and Kahle A.B. (1998) A temperature and emissivity separation algorithm for Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) images. *Ieee Transactions on Geoscience and Remote Sensing*, **36**, 1113–1126.

Grassl H. (1989) Extraction of surface temperature from satellite data. In *Applications of Remote Sensing to Agrometeorology*, Toselli F. (Ed.), Proceedings of a Course held at the Joint Research Centre of the Commission of the European Communities in the Framework of the Ispra-Courses, Ispra, Varese, Italy, 6–10 April 1987. Kluwer Academic Publishers: Dordrecht / Boston / London pp. 199–220.

Hook S.J., Prata F.J., Alley R.E., Abtahi A., Richards R.C., Schladow S.G. and Palmarsson S.O. (2003) Retrieval of lake bulk and skin temperatures using Along-Track Scanning Radiometer (ATSR-2) data: a case study using Lake Tahoe, California. *Journal of Atmospheric and Oceanic Technology*, **20**(4), 534–548.

Jackson R.D., Reginato R.J. and Idso S.B. (1977) Wheat canopy temperature: a practical tool for evaluation water requirements. *Water Resources Research*, **13**, 651–656.

Jin M. and Dickinson R.E. (2002) New observational evidence for global warming from satellite. *Geophysical Research Letters*, **29**(10), 1400, doi:10.1029/2001GL013833.

Kahle A.B. and Alley R.E. (1992) Separation of temperature and emittance in remotely sensed radiance measurements. *Remote Sensing of Environment*, **42**, 107–112.

Kerr Y.H., Lagouarde J.-P. and Imbernon J. (1992) Accurate land-surface temperature retrieval from AVHRR Data with use of an improve split-window algorithm. *Remote Sensing of Environment*, **41**, 197–209.

Kustas W.P. and Norman J.M. (2000) Evaluating the effects of sub-pixel heterogeneity on pixel average fluxes. *Remote Sensing of Environment*, **74**, 327–342.

Kustas W.P., Diak G.R. and Norman J.M. (2001) Time difference methods for monitoring regional scale heat fluxes with remote sensing. *Observations and Modeling of the Land-Surface Hydrological Processes*, *American Geophysical Union Water Science and Applications Series 3*, American Geophysical Union, pp. 15–29.

Kustas W.P. and Norman J.M. (1996) Use of remote sensing for evapotranspiration monitoring over land surfaces. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, **41**, 495–516.

Kustas W.P. and Norman J.M. (1999) Evaluation of soil and vegetation heat flux predictions using a simple two-source

model with radiometric temperatures for partial canopy cover. *Agricultural and Forest Meteorology*, **94**, 13–29.

Kustas W.P., Norman J.M., Schmugge T.J. and Anderson M.C. (2004) Mapping surface energy fluxes with radiometric temperature. Book Chapter In *Thermal Remote Sensing in Land Surface Processes*, Quattrochi D.A. and Luvall J.C. (Eds.), Ann Arbor Press.

Legrand M., Pietras C., Brogniez G., Haeffelin M., Abuhassan N.K. and Sicard M. (2000) A high-accuracy multiwavelength radiometer for in situ measurements in the thermal infrared. Part 1: characterization of the instrument. *Journal of Atmospheric and Oceanic Technology*, **71**, 1203–1214.

McClain E.P., Pichel W.G., Walton C.C., Ahmed Z. and Sutton J. (1983) Multichannel improvements to satellite derived global sea surface temperatures. *Advances in Space Research*, **2**, 23–47.

McNaughton K.G. and Van den Hurk B.J.J.M. (1995) A 'Lagrangian' revision of the resistors in the two-layer model for calculating the energy budget of a plant canopy. *Boundary-Layer Meteorology*, **74**, 262–288.

Norman J.M., Chen J. and Goel N. (1990) Thermal emissivity and infrared temperature dependency of plant canopy architecture and view angle. *Proceedings of the IEEE International Geoscience and Remote Sensing of Symposium (IGARSS'90)*, College Park, Vol. III, pp. 1747–1750, 2–24 May 1990.

Norman J.M. and Becker F. (1995) Terminology in thermal infrared remote sensing of natural surfaces. *Remote Sensing Reviews*, **12**, 159–173.

Norman J.M., Divakarla M. and Goel N.S. (1995) Algorithms for extracting information from remote Thermal-IR observations of the Earth's surface. *Remote Sensing of Environment*, **51**, 157–168.

Norman J.M., Kustas W.P. and Humes K.S. (1995b) A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature. *Journal of Agricultural Forest Meteorology*, **77**, 263–293.

Ogawa K. and Schmugge T. (2004) Mapping surface broadband emissivity (8–13.5 :m) of the Sahara Desert using ASTER and MODIS data. *Earth Interactions*, **8**, 1–14, doi: 10.1175/1087-3562.

Ogawa K., Schmugge T., Jacob F. and French A. (2002) Estimation of broadband land-surface window emissivity from multispectral thermal infrared remote sensing. *Agronomie: Agriculture and Environment*, **22**(6), 696–697.

Ogawa K., Schmugge T., Jacob F. and French A. (2003) Estimation of land-surface window (8–12 micrometer) emissivity from multispectral thermal infrared remote sensing – a case study in a part of Sahara Desert. *Geophysical Research Letters*, **30**(2), 1067, doi:10.1029/2002GL016354.

Palluconi F., Kahle A.B., Hoover G. and Conel J.E. (1990) The spectral emissivity of prairie and pasture grasses at Konza Prairie, Kansas. *Symposium on FIFE, American Meteorological Society*, Boston, pp. 77–78.

Prata A.J. (1994) Land-surface temperature derived from the advanced very high-resolution radiometer and the along-track scanning radiometer 2. Experimental results and validation of AVHRR algorithms. *Journal of Geophysical Research*, **99**(D6), 13025–13058.

Price J.C. (1980) The potential of remotely sensed thermal infrared data to infer surface soil moisture and evaporation. *Water Resources Research*, **16**, 787–795.

Price J.C. (1984) Land-surface temperature measurements from the split-window bands of the NOAA 7 advanced very high resolutions radiometer. *Journal of Geophysical Research*, **89**, 7231–7237.

Priestley C.H.B. and Taylor R.J. (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, **100**, 81–92.

Salisbury J.W. (1992) Temperature and emissivity separation. *Remote Sensing of Environment*, **42**(2), iv.

Salisbury J.W. and D'Aria D.M. (1992a) Emissivity of terrestrial materials in the 8–14 :m atmospheric window. *Remote Sensing of Environment*, **42**, 83–106.

Salisbury J.W. and D'Aria D.M. (1992b) Infrared (8–14 :m) remote sensing of soil particle size. *Remote Sensing of Environment*, **42**, 157–165.

Schmugge T.J., Becker F. and Li Z.-L. (1991) Spectral emissivity variations observed in airborne surface temperature measurements. *Remote Sensing of Environment*, **35**, 95–104.

Schmugge T., French A., Ritchie J.C., Rango A. and Pelgrum H. (2002) Temperature and emissivity separation from multispectral thermal infrared observations. *Remote Sensing of Environment*, **79**, 189–198.

Schmugge T., Hook S.J. and Coll C. (1998) Recovering surface temperature and emissivity from thermal infrared multispectral data. *Remote Sensing of Environment*, **65**, 121–131.

Seguin B. and Itier B. (1983) Using midday surface temperature to estimate daily evaporation from satellite thermal-IR data. *International Journal of Remote Sensing*, **4**, 371–383.

Soer G.J.R. (1980) Estimation of regional evapotranspiration and soil moisture conditions using remotely sensed crop surface temperatures. *Remote Sensing of Environment*, **9**, 27–45.

Stewart J.B., Kustas W.P., Humes K.S., Nichols W.D., Moran M.S. and DeBruin H.A.R. (1994) Sensible heat flux – radiometric surface temperature relationship for semiarid areas. *Journal of Applied Meteorology*, **33**, 1110–1117.

Wan Z. and Dozier J. (1989) Land-surface temperature measurement from space: physical principles and inverse modeling. *IEEE Transactions on Geoscience and Remote Sensing*, **27**(3), 268–278.

Wan Z. and Dozier J. (1996) A generalized split-window algorithm for retrieving land-surface temperature for space. *IEEE Transactions on Geoscience and Remote Sensing*, **34**, 892–905.

Wan Z., Zhang Y., Zhang Q. and Li Z.-L. (2002) Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. *Remote Sensing of Environment*, **83**, 163–180.

Wan Z., Zhang Y., Zhang Q. and Li Z.-L. (2004) Quality assessment and validation of the MODIS global land-surface temperature. *International Journal of Remote Sensing*, **25**(1), 261–274.

Yamaguchi Y., Kahle A.B., Tsu H., Kawakami T. and Pniel M. (1998) Overview of Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER). *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 1062–1071.

53: Estimation of Surface Freeze–Thaw States Using Microwave Sensors

KYLE C MCDONALD[1] AND JOHN S KIMBALL[2,3]

[1]Water and Carbon Cycles Group, Science Division, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, US
[2]Flathead Lake Biological Station, Division of Biological Sciences, The University of Montana, Polson, MT, US
[3]Numerical Terradynamic Simulation Group, The University of Montana, Missoula, MT, US

The transition of the landscape between predominantly frozen and nonfrozen conditions in seasonally frozen environments impacts climate, hydrological, ecological, and biogeochemical processes profoundly. Satellite microwave remote sensing is uniquely capable of detecting and monitoring a range of related biophysical processes associated with the measurement of landscape freeze/thaw status. This chapter provides an overview of the development, physical basis, current techniques, and selected hydrological applications of satellite-borne microwave remote sensing of landscape freeze/thaw states for the terrestrial cryosphere. Major landscape hydrological processes embracing the remotely sensed freeze/thaw signal include timing and spatial dynamics of seasonal snowmelt and associated soil thaw, runoff generation, and flooding, ice breakup in large rivers and lakes, and timing and length of vegetation growing seasons and associated productivity and trace gas exchange. This chapter also summarizes the physical principles of microwave sensitivity to landscape freeze/thaw state, recent progress in applying these principles towards satellite remote sensing of freeze/thaw processes over broad regions, and potential for future global monitoring of this significant phenomenon of the global cryosphere.

INTRODUCTION

The terrestrial cryosphere comprises cold areas of Earth's land surface where water is either permanently or seasonally frozen. This includes most regions north of 40 degrees North latitude and most mountainous regions where elevation is greater than 1000 m (Figure 1). The state transition of the land surface between frozen and thawed conditions affects a number of terrestrial processes that cycle between wintertime dormant and summertime active states. These processes occur each year over more than 50 million km^2 of the global cryosphere, affecting surface hydrological activity, meteorological conditions, and ecological trace gas dynamics profoundly (Figure 2). Abrupt near 0 °C, this state transition represents the closest analog to a hydrological and biospheric on/off switch existing in nature. Spatial patterns and timing of landscape freeze/thaw state transitions within the terrestrial cryosphere show substantial variability with measurable impacts to climate, hydrological, ecological, and biogeochemical processes.

In seasonally frozen environments, as early spring air temperatures rise above freezing, the snow pack and surface soil layer reach 0 °C and begin to thaw, resulting in a state change of the included water from solid to liquid. Snowmelt is initiated and runoff and stream discharge are accelerated with this onset of the spring thaw transition period. Ecosystem responses to these changes are equally rapid, with soil respiration and plant photosynthetic activity accelerating with warmer temperatures and the new abundance of liquid water (e.g. Goulden *et al.*, 1998; Black *et al.*, 2000; Jarvis and Linder, 2000). A critical component of the hydrological cycle, seasonal snow cover stores large amounts of freshwater, is a major source of freshwater over wide areas of the midlatitudes, and is the principal source of regional runoff in mountainous areas.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

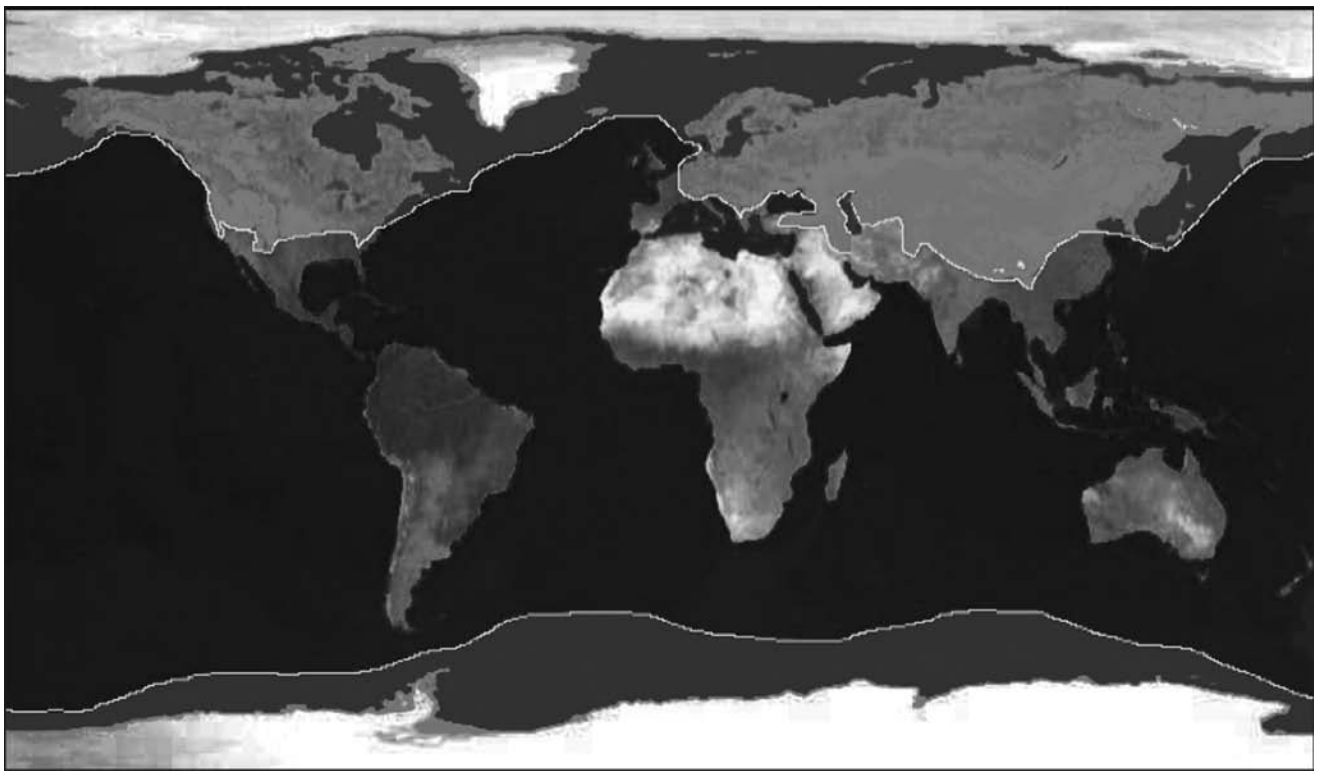


Figure 1 Approximate distribution of frost and ice effected regions. The terrestrial cryosphere comprises land areas shaded light blue or white (north of the yellow line in the northern hemisphere, and Antarctica), and most mountainous regions with elevations over 1000 m. These regions exhibit a 0 °C mean monthly temperature during the coldest month, and approximately 0.25 m of frost penetration one year in ten (after Cline, *et al.*, 1999; based on figure provided courtesy of the US Army Engineer Research and Development Center-Cold Regions Research and Engineering Lab). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The timing of seasonal freeze/thaw transitions generally defines the duration of seasonal snow cover, frozen soils, and the timing of lake and river ice breakup and flooding in the spring (Kimball *et al.*, 2001, 2004a). The seasonal nonfrozen period also bounds the vegetation growing season, whereas annual variability in freeze/thaw timing has a direct impact on net primary production and net CO_2 exchange with the atmosphere (Vaganov *et al.*, 1999; Goulden *et al.*, 1998).

Landscape freeze/thaw state strongly influences the seasonal amplitude and partitioning of surface energy exchange, with major consequences for atmospheric profile development and regional weather patterns (Betts *et al.*, 2001a). At high latitudes, for example, the onset of seasonal thawing in spring is coincident with snowmelt and large decreases in surface albedo from approximately 80% over snow to less than 20% over vegetation, soil, and open water surfaces (Lafleur *et al.*, 1997; Harazono *et al.*, 2003). Over forests, the albedo remains relatively low (~20%) even in winter (Betts and Ball, 1997), but drops below 10% following seasonal snowmelt. Prior to thaw in the early spring, frozen soils, low evaporation rates, and high incident solar radiation and sensible energy loading promote a relatively warm, dry atmospheric boundary layer as indicated by seasonally low relative humidity (Figure 3). The timing of this transition from predominantly frozen, high albedo conditions to thawed, low albedo conditions coincides with an abrupt increase in surface net radiation. With the decrease in albedo, the additional absorption of incident solar radiation causes surface air and soil temperatures to rise dramatically, often 3–5 °C within a few days. The new abundance of liquid water causes a transition of the surface energy balance from a sensible energy dominated system to a latent energy dominated system with associated increases in surface humidity. An increasing trend in relative humidity continues throughout the growing season with the availability of adequate moisture for vegetation canopy growth and evapotranspiration. In the autumn, the land surface is generally warmer than the air, and atmospheric boundary layers are shallow and moist until the onset of freezing temperatures dramatically reduces available moisture, causing an abrupt seasonal decline in relative humidity. Betts *et al.* (1998) found that numerical weather prediction

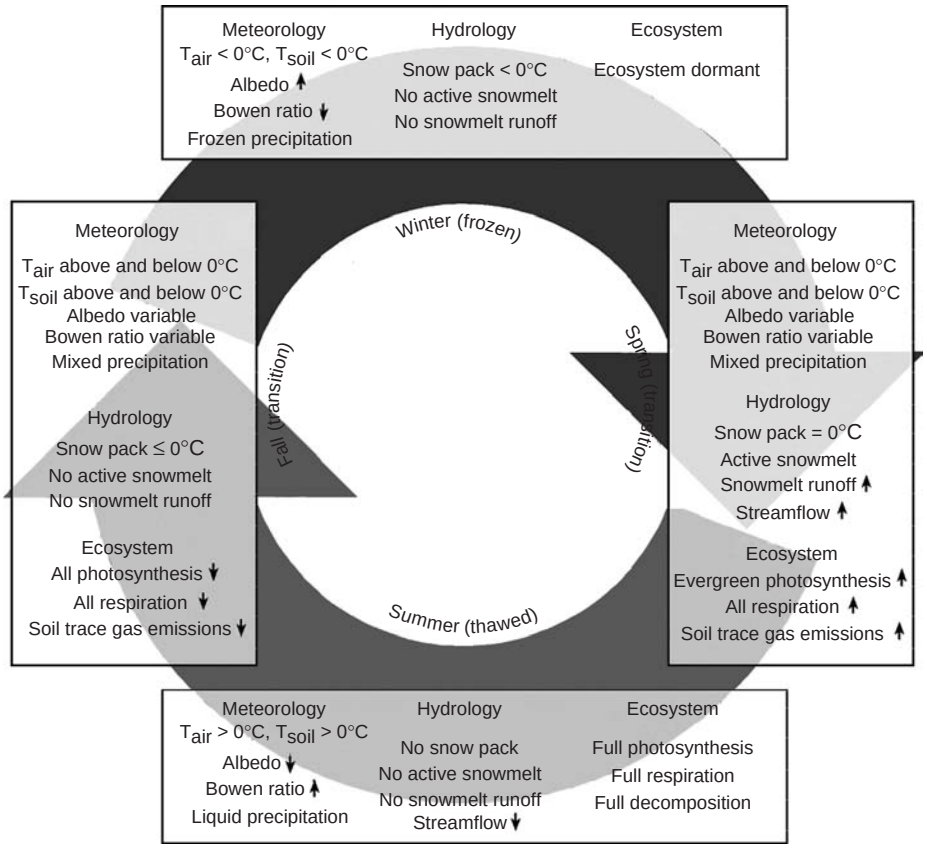


Figure 2 Conceptual diagram showing the general effects of freeze/thaw status and snow on meteorological, hydrological, and ecosystem processes throughout the year (Cline *et al.*, 1999). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

models that do not correctly account for the contrast between frozen and thawed surfaces overestimate springtime latent energy fluxes, leading to forecast errors of up to 5 °C in lower tropospheric temperatures. These errors then propagate throughout the planetary boundary layer and above.

The ability of regional monitoring networks to capture spatial and temporal patterns in landscape freeze/thaw state is severely limited because of the remote locations, large regional extent, and adverse weather conditions that characterize much of the cryosphere. The density of reporting surface weather and hydrological stations in these regions is extremely sparse, on the order of 1 station per million km^2, and has been deteriorating further because of station closures, limited funding, and embargoes on national data (Karl, 1995; Lanfear and Hirsch, 1999). Satellite-borne optical-infrared remote sensing methods provide regional information regarding snow-covered area, surface albedo, and land-surface temperature (LST) that can be used to infer surface freeze/thaw state. For most of the cryosphere, however, frequent cloud cover, low solar elevation angles, shadowing, and reduced illumination restrict regional monitoring from these sensors to relatively coarse 8–16 day temporal composites necessary to mitigate atmospheric effects (Running, 1998; Cihlar *et al.*, 1997). These effects are particularly acute during winter, early spring, and late fall when seasonal freeze/thaw transitions are most likely to occur.

Satellite-borne remote sensing at microwave wavelengths has unique capabilities that allow near real-time monitoring of numerous landscape processes associated with a single measure, freeze/thaw state, without many of the limitations of optical-infrared sensors. These properties include a strong sensitivity to the pronounced contrast in the dielectric constant of frozen and liquid water, the ability of microwaves to penetrate cloud cover, and their independence from solar illumination. Major landscape hydrological processes embracing the remotely sensed freeze/thaw signal include timing and spatial dynamics of seasonal snowmelt and associated soil thaw, runoff generation and flooding, ice breakup in large rivers and lakes, and timing and length of vegetation growing seasons and associated productivity and trace gas exchange. This paper summarizes the physical principles of microwave sensitivity to landscape freeze/thaw state, recent progress in applying these principles toward satellite remote sensing

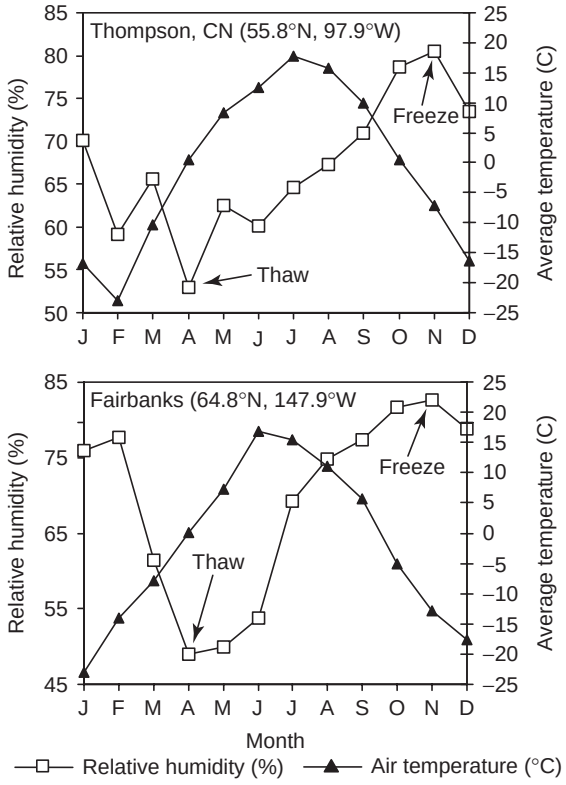


Figure 3 The seasonal cycle of mean monthly relative humidity and air temperature at 2 sites in Northern Canada and Alaska illustrates the large control seasonal surface freeze-thaw dynamics have on boreal climate. Monthly mean relative humidity is at a minimum in April, just prior to spring thaw, because of increased solar irradiance and a general lack of liquid water available for evaporation under frozen conditions. Relative humidity is at a maximum in autumn, when the ground, not yet frozen, is warmer that the air. The consequence is a deep, warm boundary layer in spring, and a very shallow, stratocumulus capped boundary layer in autumn (after Betts *et al.*, 2001b)

of freeze/thaw processes over broad regions, and potential for future global monitoring of this key phenomenon of the global cryosphere.

PRINCIPLES OF MICROWAVE REMOTE SENSING FOR FREEZE/THAW DETECTION

Microwave remote sensing signatures of natural terrain are controlled by the dielectric properties and the structure (shapes, sizes, and spatial arrangement) of the landscape constituents (Elachi, 1987). Radars are active sensor systems that observe a target, providing their own illumination source and measuring the scattered energy returned to the sensor (backscatter). Radiometers are passive sensor systems that observe the target's natural emission (emissivity). Microwave sensors operate at much longer wavelengths than optical sensors, allowing observations day and night throughout the year, regardless of cloud cover, solar zenith angle, and reduced solar illumination that are particularly problematic for optical-infrared remote sensing of the high latitudes. While continuous coverage of surface state conditions is possible, actual coverage is defined by sensor configuration and orbit design.

Freeze/thaw transitions may be spatially and temporally heterogeneous, with the landscape commonly experiencing several thaw and refreeze events in a season, underscoring the need for mapping with combined high spatial resolution and high temporal revisit (Running *et al.*, 1999). Both radiometer and radar measurements have been shown to be sensitive to landscape freeze/thaw state. The effects of vegetation characteristics on these measurements is complex and for the active (radar) case, is influenced by the shapes, sizes, and orientations of the vegetation components to a greater extent than for the passive (radiometer) case. Radars, on the other hand, generally allow measurements with a higher spatial resolution than do radiometers. In general, low spatial resolution, high temporal revisit microwave radiometers (e.g. SSM/I, AMSR-E) and scatterometers (e.g. SeaWinds) are well-suited for quantifying the timing of freeze/thaw transitions across broad regions, whereas high spatial resolution, low-to-moderate revisit Synthetic Aperture Radars (SARs) (e.g. JERS, RADARSAT) are better able to resolve spatial heterogeneity in freeze/thaw transitions in regions of complex topography and land cover. The techniques described here are applicable to both radars and radiometers.

Permittivity and Dielectric Constant

The ability of microwave remote sensing instruments to observe freezing and thawing of a landscape has its origin in the distinct changes of surface dielectric properties that occur as water transitions between solid and liquid phases. A material's permittivity describes how that material responds in the presence of an electromagnetic field (Kraszewski, 1996). As an electromagnetic field interacts with a dielectric material, the resulting displacement of charged particles from their equilibrium positions gives rise to induced dipoles that respond to the applied field. A material's permittivity is a complex quantity (i.e. having both real and imaginary numerical components) expressed as

$$\varepsilon = \varepsilon' - j\varepsilon'' \tag{1}$$

and is often normalized to the permittivity of a vacuum (ε_0), and referred to as the relative permittivity, or the *complex dielectric constant*:

$$\varepsilon_r = \frac{\varepsilon'}{\varepsilon_0} - \frac{j\varepsilon''}{\varepsilon_0} = \varepsilon_r{}' - j\varepsilon_r{}'' \tag{2}$$

The real component of the dielectric constant, $\varepsilon_r{}'$, is related to a material's ability to store electric field energy.

The imaginary component of the dielectric constant, ε_r'', is related to the dissipation or energy loss within the material. At microwave wavelengths, the dominant phenomenon contributing to ε_r'' is the polarization of molecules arising from their orientation with the applied field. The dissipation factor, or *loss tangent,* is defined as the ratio

$$\tan(\delta) = \frac{\varepsilon_r''}{\varepsilon_r'} \tag{3}$$

Consisting of highly polar molecules, liquid water exhibits a dielectric constant that dominates the microwave dielectric response of natural landscapes. Graphed in Figure4, the Debye equation describes the dielectric constant of pure liquid water (Ulaby *et al.*, 1986, pp. 2022–2025). As liquid water freezes, the molecules become bound in a crystalline lattice, impeding the free rotation of the polar molecules and reducing the dielectric constant substantially. Figure 4 compares ε_r of liquid water at 0 °C with ε_r of pure ice across the microwave spectrum.

Brightness Temperature and Backscatter

In general, landscapes of the terrestrial cryosphere consist of a soil substrate that may be covered by some combination of vegetation and seasonal or permanent snow. The sensitivity of radar and radiometer signatures to these landscape features is affected strongly by the sensing wavelength, as well as landscape structure and moisture conditions. The composite remote sensing signature represents a sampling of the aggregate landscape dielectric and structural characteristics, with sensor wavelength having a strong influence on the sensitivity of the remotely sensed signature to the various landscape constituents.

In the microwave region of the electromagnetic spectrum, vegetation canopies may be considered to be weakly scattering media (i.e. media characterized by volume absorption that is much larger than volume scattering), with the first-order approximation to canopy emission and scattering being applicable. In this case, diffuse scattering effects may be ignored and first-order emissivity and scattering model approaches utilized to interpret emissivity and backscatter (Ulaby *et al.*, 1986, vol. II). Figure 5 depicts the first-order contributions to landscape emissivity and backscatter. At higher frequencies, scattering effects of the vegetation are increasingly significant. The dependence of the microwave signatures on vegetation characteristics is complex, with vegetation structure influencing radar signatures to a greater extent than radiometer signatures.

For the passive microwave case, a material's radiometric *brightness temperature*, T_B, is characterized by its emissivity e, as $T_B = e \cdot T$, where T is its physical temperature (K). Emissivity is a unitless variable ranging from 0 for a perfectly nonemitting material, to 1 for a perfect emitter (blackbody) (Ulaby *et al.*, 1986, vol. I). First-order contributions to the landscape brightness temperature (Figure 5a) are (i) emission by the underlying surface that is attenuated

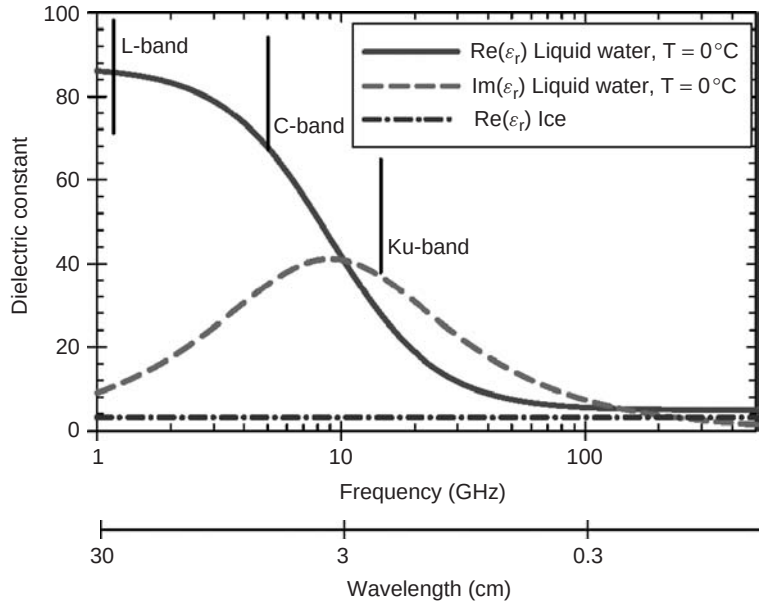


Figure 4 Comparison of the dielectric constants of liquid water at 0 °C and ice shows the contrast across the microwave spectrum. The dielectric constant of water is described by the Debye equation. For pure ice, $\varepsilon_r \approx 3.15 - j\,0.001$ and is independent of temperature and frequency in the microwave region (Ulaby, *et al.*, 1986, p. 848; Kraszewski, 1996). As liquid water possesses a relatively high dielectric constant, its presence or absence and freeze/thaw state strongly influences backscatter and emissivity. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

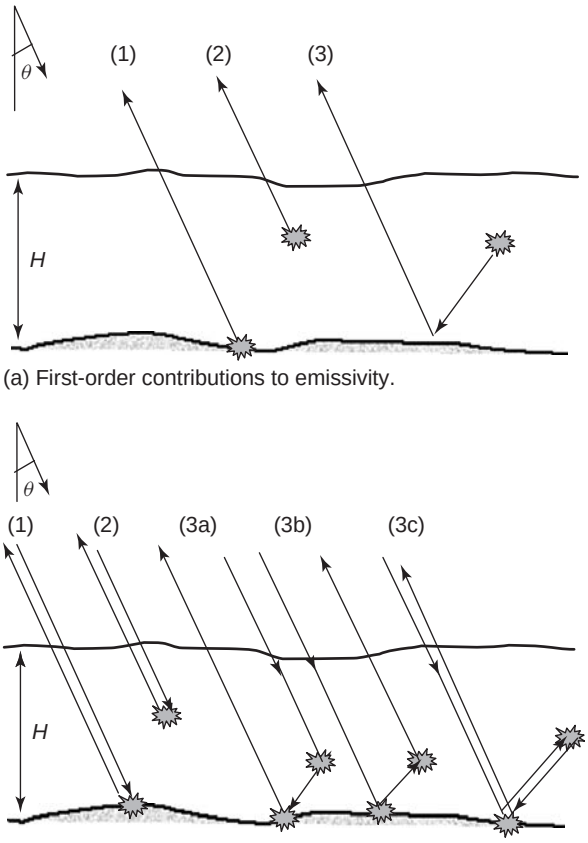


Figure 5 First-order emissivity and scattering contributions to emission (a) and scattering (b) for a land surface covered by a vegetation canopy of height H. The angle θ represents the look angle of the sensing instrument. First-order contributions to landscape emissivity are: (1) emission by the underlying soil and snow that is attenuated by the vegetation, (2) upward-propagating emission by the vegetation itself, and (3) downward-propagating emission by the vegetation that is reflected by the snow-soil surface toward the sensor, attenuated by the vegetation. First-order contributions to backscatter are: (1) backscatter from the underlying snow-soil surface that is attenuated by two-way propagation through the vegetation, (2) backscatter from the vegetation itself, and (3) backscatter arising from multiple scattering between the vegetation and snow-soil surface. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

by upward propagation through the vegetation canopy, (ii) vegetation volume emission propagating in the upward direction, and (iii) vegetation volume emission propagating in the downward direction that is reflected back through the vegetation volume by the underlying surface. These first-order relationships between the landscape physical characteristics and the observed brightness temperature T_{Bp} at polarization p (vertical or horizontal) can be expressed as (Ulaby *et al.*, 1986, vol. II, p. 889):

$$T_{Bp} = T_s e_p \exp(-\tau_c) + T_c(1-\omega)[1-\exp(-\tau_c)] + T_c(1-\omega)[1-\exp(-\tau_c)] r_p \exp(-\tau_c) \quad (4)$$

where T_c and T_s are the physical temperatures (K) of the vegetation canopy and the underlying surface, respectively, τ_c is the vegetation opacity along the slant path defined by the radiometer look angle (θ) and canopy height (H), ω is the single-scattering albedo of the vegetation medium, and r_p is the soil reflectivity at look angle θ. The surface reflectivity is related to its emissivity by $r_p = (1 - e_p)$.

For the active microwave case, the total copolarized backscatter from the landscape at polarization p is the sum of three components (Figure 5b):

$$\sigma^t_{pp} = \sigma^s_{pp} \exp(-2\tau_c) + \sigma^{\text{vol}}_{pp} + \sigma^{\text{int}}_{pp} \quad (5)$$

The first term is the surface backscatter, σ^s_{pp}, modified by the two-way attenuation through a vegetation layer of opacity τ_c along the slant path. The second term represents the backscatter from the vegetation volume, σ^{vol}_{pp}. The third term represents the interaction between the vegetation and soil surface, σ^{int}_{pp}, and includes contributions from terms (3a), (3b) and (3c) of Figure 5(b) (Ulaby *et al.*, 1986 vol II, pp. 860–875; Ulaby *et al.*, 1990; Kuga *et al.*, 1990).

The response of T_{Bp} and σ^t_{pp} to freeze/thaw dynamics is affected by surface roughness, topography, and vegetation and snow cover. The relative contribution of the three terms in each of equations (4) and (5) is influenced strongly by the canopy opacity, τ_c. Opacity, volume scattering, and emission all vary significantly with frequency. For bare surface or sparsely vegetated conditions, the surface terms dominate the received signal, with T_{Bp} and σ^t_{pp} being influenced primarily by contributions from the soil surface and snow cover. Because of the high dielectric constant of liquid water, microwave penetration of the vegetation decreases as biomass moisture levels increase. In general, τ_c increases with increasing frequency, vegetation density, and dielectric constant. High-frequency, short-wavelength energy (e.g. Ku-band) has higher τ_c and does not penetrate as significantly into vegetation canopies as lower-frequency, longer-wavelength energy (e.g. L-band). Hence, higher-frequency sensors are generally less sensitive to properties of surfaces underlying dense vegetation canopies. However, the dependence of T_{Bp} and σ^t_{pp} on snow characteristics is also complex, with snow pack wetness, density, crystal structure, and depth influencing both brightness temperature and backscatter. The effect of snow cover on brightness temperature and backscatter is more significant at higher frequencies because of the increased scattering albedo of the snow pack at short wavelengths relative to longer wavelengths (e.g. Ulaby *et al.*, 1986; Raney, 1998). These phenomena lead to notable differences in the

temporal response of emission and backscatter to landscape freeze/thaw processes with sensor wavelengths (e.g. Way *et al.*, 1994; Way *et al.*, 1997; Frolking *et al.*, 1999; Kimball *et al.*, 2001).

ALGORITHMS

Common approaches employing satellite microwave remote sensing to classifying landscape freeze/thaw state involve temporal change detection schemes applied to time-series radar backscatter or radiometric brightness temperature. The general approach of these techniques is to identify landscape freeze/thaw transition sequences by exploiting the dynamic temporal response of backscatter or brightness temperature to differences in the aggregate landscape dielectric constant that occur as the landscape transitions between predominantly frozen and nonfrozen conditions. These techniques assume that the large changes in dielectric constant occurring between frozen and nonfrozen conditions dominate the corresponding backscatter and brightness temperature temporal dynamics, rather than other potential sources of temporal variability such as changes in canopy structure and biomass or large precipitation events. This assumption is generally valid during periods of seasonal freeze/thaw transitions for most areas of the cryosphere.

Seasonal Threshold Approach

Seasonal threshold approaches examine the time-series progression of the remote sensing signature relative to signatures acquired during a seasonal reference state or states. These techniques are well-suited for application to data with temporally sparse or variable repeat-visit observation intervals and have been applied to ERS and JERS Synthetic Aperture Radar (SAR) imagery (*e.g.* Rignot and Way, 1994; Way *et al.*, 1997; Gamon *et al.*, 2004; Entekhabi *et al.*, 2004). A seasonal scale factor $\Delta(t)$ may be defined for an observation acquired at time t as:

$$\Delta(t) = \frac{\sigma(t) - \sigma_{\mathrm{fr}}}{\sigma_{\mathrm{th}} - \sigma_{\mathrm{fr}}} \tag{6}$$

where $\sigma(t)$ is the measurement acquired at time t, for which a freeze/thaw classification is sought, and σ_{fr} and σ_{th} are measurements (backscatter or brightness temperature) corresponding to the frozen and thaw reference states, respectively. In situations where only a single reference state is available, for example, σ_{fr}, $\Delta(t)$ may be defined as a difference:

$$\Delta(t) = \sigma(t) - \sigma_{\mathrm{fr}} \tag{7}$$

A threshold level T is then defined such that

$$\begin{aligned} \Delta(t) &> T \\ \Delta(t) &\leq T \end{aligned} \tag{8}$$

define the frozen and thawed landscape states, respectively. Whereas equation (6) accounts for differences in the dynamic range of the remote sensing response to freeze/thaw transitions driven by variations in land cover, equation (7) does not scale $\Delta(t)$ to account for the dynamic range in the seasonal response.

The selection of parameter T may be optimized for various land-cover conditions and sensor configurations. In situations where the wintertime microwave signature is not dominated by the snow pack volume, for example, for radar measurements at lower frequencies (e.g. L-band) and where shallow dry snowpacks are common, $\sigma_{\mathrm{fr}} < \sigma_{\mathrm{th}}$ and $\Delta(t) > T$ defines the thawed state. In situations where the wintertime signature is dominated by the snow pack volume, for example, at high frequencies (e.g. Ku-band) and where deep, wet snowpacks are common, the microwave seasonal response to landscape freeze/thaw is more complex, and defining the frozen and thawed landscape state in terms of a single threshold may not adequately represent these processes. In some applications, multiple threshold values may be employed to delineate multiple freeze/thaw events occurring during seasonal transitions.

Moving Window Approach

Moving window techniques classify freeze/thaw transitions based on changes in the radiometric signature relative to the temporally averaged signature computed over a moving window of specified duration. These approaches are useful when applied to temporally consistent data sets consisting of frequent (e.g. daily) observations, and for identifying multiple freeze/thaw transition events. For a measurement $\sigma(t)$ acquired at time t, the difference $\delta(t)$ relative to a moving window mean may be defined as

$$\delta(t) = \sigma(t) - \sigma_{av}(t - L \leq t_0 \leq t - 1) \tag{9}$$

where $\sigma_{av}(t - L \leq t_0 \leq t - 1)$ is the average measurement (backscatter or brightness temperature) acquired over a window of duration L extending over the time interval $(t - L \leq t_0 \leq t - 1)$. The difference $\delta(t)$ may be compared to various thresholds, as in (8), to define the timing of critical freeze/thaw transitions. These approaches have been employed using both NSCAT and SeaWinds scatterometer data for a variety of regions (Frolking *et al.*, 1999; Kimball *et al.*, 2001, 2004a,b; Rawlins *et al.*, 2004). Principal distinctions in the application of equation (9) have been the duration L of the moving window and the selection of thresholds applied to infer transition events.

Temporal Edge Detection Approach

Temporal edge detection techniques classify freeze/thaw transitions by identifying predominant step edges in time-series remote sensing data that correspond to freeze/thaw

transition events. As freeze/thaw events induce large temporal changes in landscape dielectric properties that tend to dominate the seasonal time-series response of the microwave radiometric signatures for the terrestrial cryosphere, edge detection approaches are suitable for identification of these events using time-series microwave remote sensing data. These techniques are based on the application of an optimal edge detector for determining edge transitions in noisy signals (Canny, 1986). The timing of a major freeze/thaw event is determined from the convolution applied to a time series of backscatter or brightness temperature measurements $\sigma(t)$:

$$CNV(t) = \int_{-\infty}^{\infty} f'(x)\sigma(t-x)\,\mathrm{d}x \qquad (10)$$

where $f'(x)$ is the first derivative of a normal (Gaussian) distribution. The occurrence of a step-edge transition is then given by the time when *CNV(t)* is at a local maximum or minimum.

Seasonal transition periods may involve multiple freeze/thaw events. This technique accounts for the occurrence of weak edges, or less pronounced freeze/thaw events, as well as larger seasonal events indicated by strong edges, and can distinguish the frequencies and relative magnitudes of these events. The variance of the normal distribution may be selected to identify step edges with varying dominance, that is, selection of a large variance identifies more predominant step edges, while narrower variances allow identification and discrimination of less pronounced events. This approach has been applied to daily time-series brightness temperatures from the Special Sensor Microwave/Imager (SSM/I) to map primary springtime thaw events annually across the pan-Arctic basin and Alaska (McDonald *et al.*, 2004).

HYDROLOGICAL APPLICATIONS OF SATELLITE FREEZE/THAW DETECTION

The ability of satellite-borne microwave remote sensing instruments to detect freeze/thaw transitions over broad landscapes has been well established. Initial radar studies were carried out from ground-based observations of bare soils and croplands, followed by aircraft campaigns over boreal landscapes, and more recently at regional and continental scales using a variety of satellite-based platforms. Studies using truck-mounted scatterometers have demonstrated the sensitivity of radar backscatter to soil and vegetation frozen and thawed conditions (Wegmuller, 1990; Ulaby *et al.*, 1986). Freeze/thaw transitions in boreal forest landscapes were first observed with imaging radar in a series of L-band (1.1 GHz, 23-cm wavelength) synthetic aperture radar (SAR) images acquired during March 1988 over boreal forests of central Alaska by the NASA/JPL AIRSAR instrument flown onboard a DC-8 aircraft. As air temperatures ranged from unseasonably warm (up to 9 °C) to well below freezing (−8 to −15 °C), a corresponding 4–6 dB decrease in landscape backscatter resulted (Way *et al.*, 1990). An intensive study of the fall freeze transition was carried out for the same region using a temporal sequence (August to November 1991) of European Remote Sensing (ERS-1) satellite C-band (5.2 GHz, 5.7 cm) SAR images processed at a 200-m spatial resolution (Rignot *et al.*, 1994; Rignot and Way, 1994). ERS SAR imagery showed an approximate 3 dB decrease in radar backscatter as the landscape froze, corresponding to *in situ* vegetation, soil, and air-temperature measurements of freezing conditions within mature black spruce, white spruce, and balsam poplar stands. ERS-1 SAR data were also used to assess spring thaw processes within the BOREAS region of central Canada (Way *et al.*, 1997; Gamon *et al.*, 2004). These results were limited to a coarse 3–168 day temporal repeat because of the narrow (100 km) swath width and varying orbital geometry of the ERS SAR sensor. Results showed a two-stage 1–2 dB temporal increase in radar backscatter, relative to midwinter frozen values, corresponding to surface measurements of initial snow pack and soil thaw in March followed by forest canopy thaw in May.

At continental scales, Wismann (2000) mapped spatial and temporal behavior in the surface spring thaw transition across much of Siberia using a temporal change detection analysis of ERS-1 C-band scatterometer data. These data were collected from March to July 1993, processed to a 50-km spatial resolution and an approximate 3-day temporal fidelity. Moderate, 25-km spatial resolution data from the Ku-band NASA scatterometer (NSCAT; 14.0 GHz, 2.1 cm) and SeaWinds-on-QuikSCAT scatterometer (13.4 GHz, 2.2 cm) have been used to assess freeze/thaw cycles over broad boreal, Arctic, and subalpine landscapes of North America (Frolking *et al.*, 1999; Kimball *et al.*, 2001, 2004a,b). Daily backscatter data from these scatterometers show a 5 to 8 dB temporally dynamic response that coincides with landscape springtime freeze/thaw processes as identified by daily snow cover, air, vegetation, and soil temperature measurements from regional surface station networks. The primary thaw transition has been found to be generally coincident with the arrival of maximum surface wetness in spring associated with rising air temperatures, seasonal snowmelt, soil active layer thawing and growing season onset (Kimball *et al.*, 2004a).

Studies using coincident satellite-borne remote sensing and surface network measurements of freeze/thaw processes demonstrate that these seasonal transitions are spatially heterogeneous and that landscapes undergo multiple thaw and refreeze cycles in a season. Figure 6 shows the seasonal progression of daily mean radar backscatter from a mature boreal black spruce forest site in northern Manitoba,

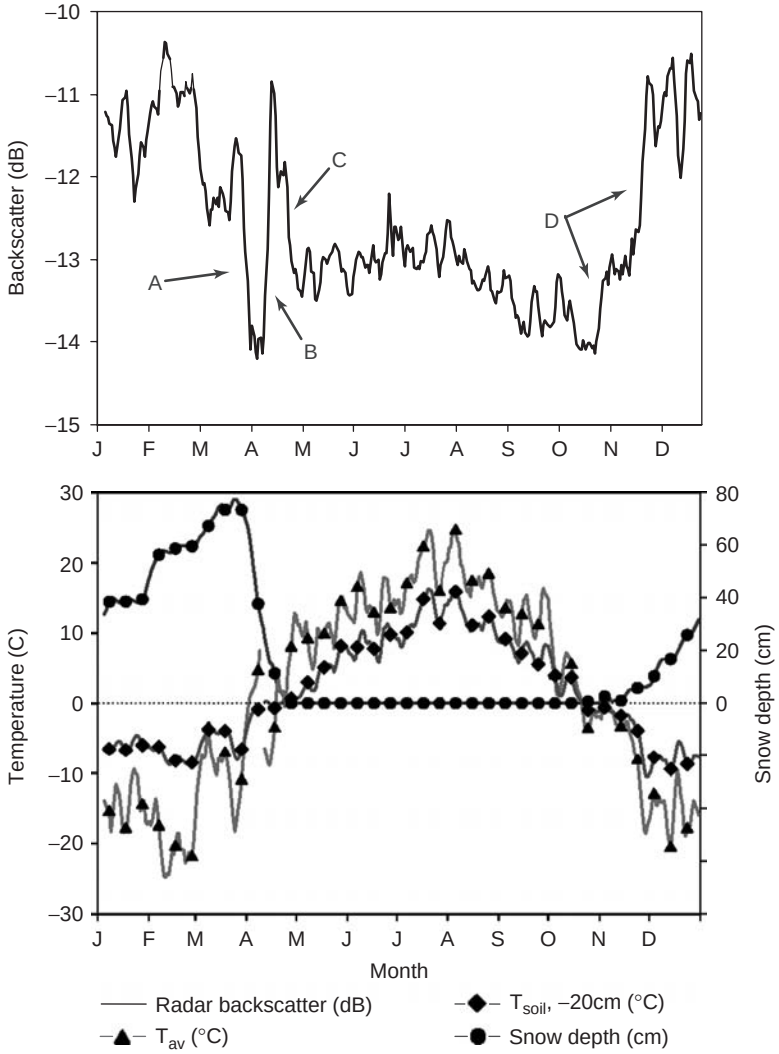


Figure 6 Satellite-borne Ku-band mean daily radar backscatter from the outer beam of the SeaWinds scatterometer flown onboard QuikSCAT measured for a mature boreal black spruce forest in Northern Manitoba, Canada in 2001. Also shown are daily site surface air and soil temperature data from the NSA-OBS Ameriflux site and snow depth data from nearby Thompson, Manitoba, both located within the SeaWinds sensor footprint. These results highlight the large sensitivity of Ku-band radar to landscape seasonal freeze-thaw state transitions as indicated by seasonal changes in snowcover, and air and soil temperatures. Surface state transitions between frozen and thawed conditions are associated with large (1–3 dB) temporal changes in radar backscatter. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Canada measured with the SeaWinds scatterometer. These results demonstrate the pronounced sensitivity of Ku-band radar backscatter to seasonal landscape freeze-thaw state transitions as identified by surface snow cover and air and soil temperature measurements located within the sensor footprint. Surface freeze/thaw state transitions are associated with large (1–3 dB) temporal changes in radar backscatter. A major thaw event in early spring (A) coincides with a 2-dB decrease in radar backscatter, followed by a brief refreezing event (B) corresponding to a 3-dB backscatter increase, final spring thawing (C) and a large backscatter decrease. In the autumn, freezing air and soil temperatures and the arrival of seasonal snow cover coincide with large increases in radar backscatter (D). These time-series backscatter data have been shown to identify spring thaw and autumn freeze processes that coincide with seasonal shifts in boreal ecosystem carbon exchange from a source to sink of atmospheric CO_2. This effectively identifies the seasonal boundaries of the vegetation growing season (Frolking *et al.*, 1999; Kimball *et al.*, 2004b).

Studies utilizing passive microwave radiometry have also demonstrated a capability for distinguishing freeze/thaw cycles with both ground-based and spaceborne systems (England, 1990; Judge *et al.*, 1997; Judge *et al.*, 2001;

Zhang *et al.*, 2003). Data from both the Nimbus 7 Scanning Multichannel Microwave Radiometer (SMMR) and the Special Sensor Microwave/Imager (SSM/I) have been used to identify and monitor snow cover and the freeze/thaw status of prairie soils across the Northern Great Plains region of North America (England, 1990; Zuerndorfer and England, 1992; Judge *et al.*, 1997). Though these sensors have lower spatial resolutions than active microwave sensors, they have been successful at monitoring thaw processes and snow properties in prairie regions with gentle topography, shallow snow, dry soils, and low vegetation densities (e.g. Josberger *et al.*, 1998).

Availability of multiyear global data sets allows investigation of trends and variability in freeze/thaw transitions. SeaWinds-on-QuikSCAT provides consistent global coverage beginning summer, 1999. Figure 7 shows maps of the primary springtime thaw date across the pan-Arctic basin and Alaska as derived from SeaWinds data for years 2000–2003. The multiyear SSM/I time series has also been used to assess seasonal freeze-thaw dynamics at high latitudes and associated linkages to vegetation growing seasons and atmospheric CO_2 concentrations (Mognard *et al.*, 1998; McDonald *et al.*, 2004). These data have proven to be a useful tool for global change monitoring at high latitudes. Analysis of daily time-series SSM/I data for the pan-Arctic basin and Alaska, for example, shows evidence of an approximate 8-day advance in the timing of spring thawing and initiation of seasonal growing seasons at high latitudes from 1988 to 2001, with generally positive impacts to regional vegetation productivity (McDonald *et al.*, 2004). This finding is generally consistent with other observations of a trend towards earlier onset of the growing season and higher vegetation productivity at high latitudes derived from the NOAA AVHRR Pathfinder record (Myneni *et al.*, 1997). Unlike NOAA AVHRR, however, satellite-borne microwave sensors such as SSM/I and SeaWinds provide more consistent, daily observations of the cryosphere without significant signal degradation from atmosphere and illumination effects, allowing improved temporal fidelity and potentially better detection and monitoring of global change.

The global coverage and daily temporal repeat observations of landscape freeze/thaw conditions available from moderate resolution, active/passive microwave sensors such as SeaWinds and the SSM/I have direct application to hydrological monitoring (Running *et al.*, 1999; Kimball *et al.*, 2001; Rawlins *et al.*, 2004). The timing of seasonal ice breakup and the spring flood pulse is the

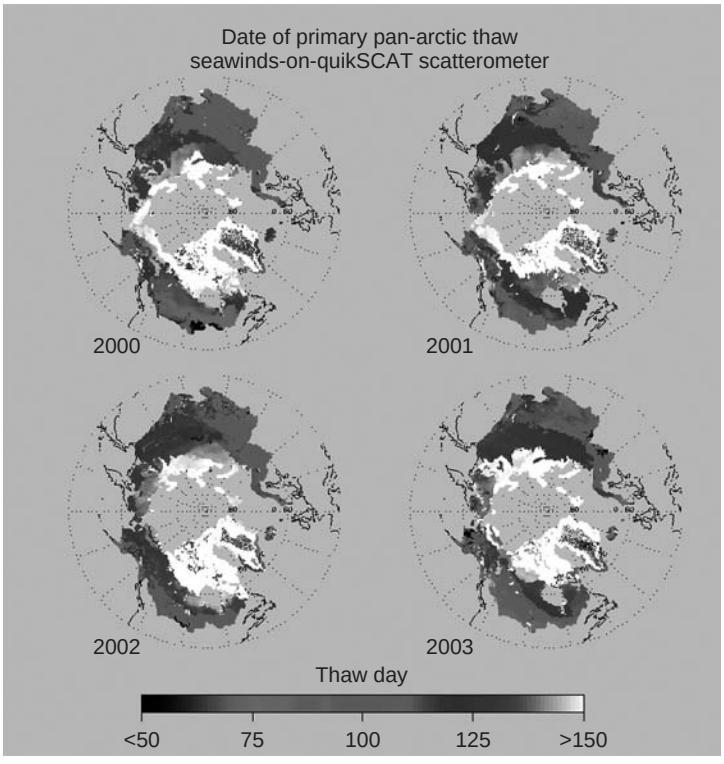


Figure 7 Day of the primary spring thaw event across the pan-arctic drainage basin for 2000–2003. Maps were derived using daily SeaWinds-on-QuikSCAT outer beam data posted to a 25-km grid. The edge detection scheme of equation (10) was applied to derive primary thaw day. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

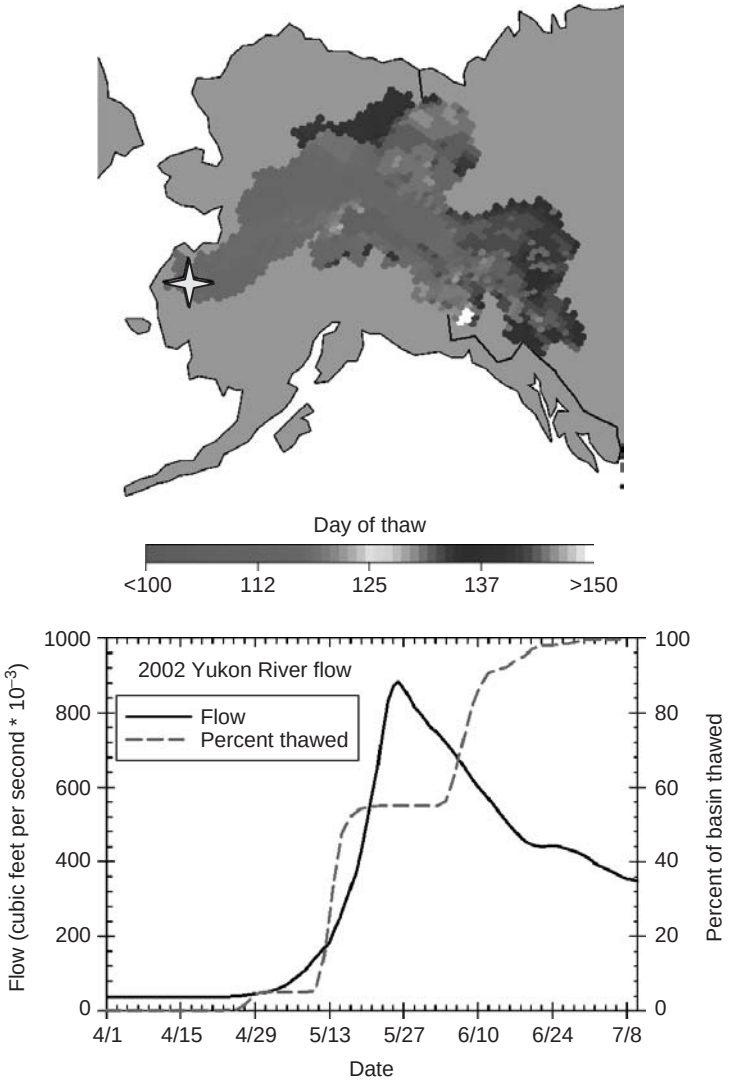


Figure 8 The map at top shows the day of primary thaw during 2002 for the Yukon River basin of Alaska and Yukon Territory. Primary thaw day was derived using the edge detection scheme of equation (10) applied to the SeaWinds-on-QuikSCAT Ku-band scatterometer data. The graph at bottom compares the percent of the Yukon River basin having undergone the primary thaw to daily mean river flow during 2002 measured near the mouth of the Yukon River. The Yukon River drains 831 391 km^2 of boreal NW Canada and Alaska. Yukon River daily flow records were obtained for Pilot Station, Alaska (location indicated by the yellow cross) courtesy of the US Geological Survey. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

dominant annual hydrological event for major rivers of the cryosphere. The rate, location, and extent of thawing and snowmelt in spring have major impacts on runoff and potential for flooding. While stream gages provide accurate measurements of past and current flow conditions, they provide little information on future flows or advanced warning of potential flooding. Figure 8 shows a map of the day of the primary spring thaw event as derived from a temporal classification of SeaWinds daily radar backscatter observations for the Yukon River basin of Alaska and Yukon Territory. The Yukon River drains an area of approximately 840 000 km^2 and is one of the 3 longest rivers in North America. The graph in Figure 8 shows the relationship between stream flow and the spring flood pulse, and the relative proportion of basin area having undergone the primary springtime thaw as determined from the SeaWinds data. Together, the map and associated graph show that for 2002 seasonal thawing of the basin initiated at lower elevations prior to April 30 and progressed to higher elevation areas over an approximate 3-week period. A major thaw event began on May 12, extended over a 2-day period and encompassed approximately 50% of the basin. This event coincided with the timing of major ice breakup and arrival of the primary spring flood pulse for the river. A second

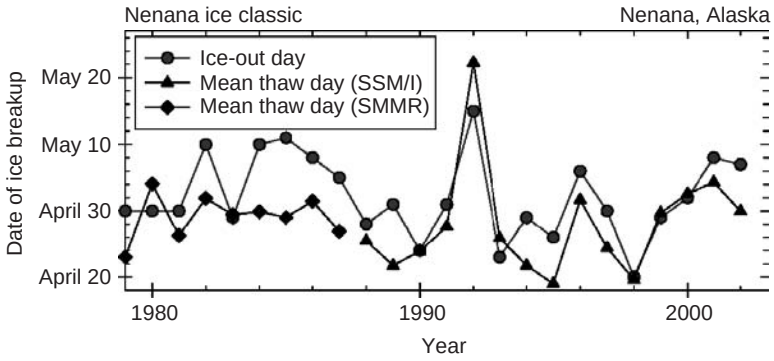


Figure 9 Comparison of mean primary thaw day for the Yukon River basin of Alaska and Yukon Territory with ice breakup day measured on the Tanana River at Nenana, Alaska (http://www.nenanaakiceclassic.com/BreakupLog.htm). Primary thaw day was derived from SSMI and SMMR passive microwave pathfinder data (Armstrong *et al.,* 1994; Armstrong and Brodzik, 1995; Knowles *et al.,* 2002). The relationship between mean thaw day and ice breakup day yields $r^2 = 0.52$ and $P = 7.2 \times 10^{-5}$ (JERS Images Copyright © 1992–1998 NASDA/MITI. Courtesy of the Global Boreal Forest Mapping (GBFM) project). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

major thaw event is also evident, beginning in early June and is associated with thawing of higher elevations. This secondary event does not contribute to the seasonal peak flow, but extends the period of flood recession.

The timing of seasonal ice breakup of major rivers and lakes in spring has important implications for navigation and potential for flooding. Additionally, long-term changes in the timing of these events have been found to be an important indicator of hydrological responses to global change (Magnuson *et al.*, 2000). As the previous example indicated, satellite microwave remote sensing of the timing of regional thawing in spring is closely associated with the timing of the seasonal ice breakup for major rivers. Long-term monitoring of freeze-thaw processes from spaceborne sensors has the potential for enhanced regional monitoring of this important variable. Figure 9 shows the relationship between the timing of ice breakup for the Tanana River at Nenana, Alaska and the average annual spring thaw date for the surrounding Yukon basin as derived from a temporal classification of SMMR and SSM/I time-series T_b observations for the region (McDonald *et al.*, 2004). These results illustrate the strong correspondence between seasonal ice breakup for the river and regional thawing as observed from the passive microwave record. This graphic also highlights the substantial annual variability in both the timing of spring thawing and seasonal ice breakup that characterizes much of the global cryosphere, with the standard deviations in mean thaw day and ice breakup day being 6.7 and 6.2 days, respectively.

While moderate resolution sensors such as SeaWinds and SSM/I provide relatively high temporal fidelity, their capability to distinguish subgrid scale (i.e. $<\sim$25 km resolution) freeze/thaw processes is limited. These limitations can be problematic for distinguishing complex freeze/thaw patterns over landscapes with variable topography and other spatially heterogeneous features. Techniques have been applied to improve spatial resolution of these data with demonstrated success (Early and Long, 2001), though these methods generally sacrifice temporal fidelity in order to enhance spatial resolution. Satellite-borne synthetic aperture radars such as JERS-1, ERS and Radarsat time-series data provide relatively fine scale (e.g. $\leq$200 m) spatial resolution and have been useful for monitoring freeze/thaw transitions, albeit with less temporal fidelity than the moderate resolution sensors. While the relatively coarse temporal fidelity (3–168 day) of these data has limitations for global monitoring, they can provide an accurate means for assessing subgrid scale spatial variability in freeze-thaw processes over complex landscapes. As Figure 10 shows, 100 m resolution JERS-1 L-band SAR data have the capability to accurately distinguish spatial variability in freeze-thaw processes associated with differences in landscape slope, aspect, and vegetation within a complex boreal landscape of central Alaska. While coincident Ku-band daily backscatter time-series data from SeaWinds capture the predominant seasonal freeze/thaw pattern of the 25-km region, these data do not capture the subregional complexity of freeze/thaw spatial patterns associated with variable vegetation canopy conditions and topography.

SUMMARY AND FUTURE DIRECTIONS

We have presented an overview of the development, physical basis, current techniques, and selected hydrological applications of satellite-borne remote sensing of landscape freeze/thaw states. The ability of satellite-based microwave remote sensing to detect freeze/thaw status over broad landscapes has been well established, first through focused ground-based and aircraft observations, and extending to

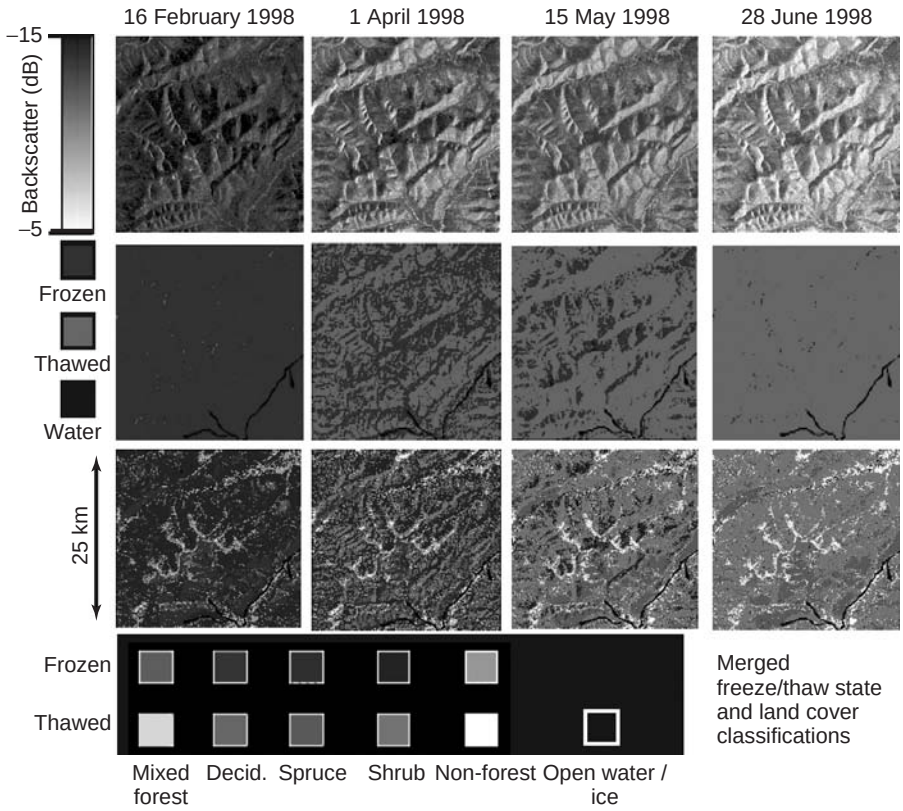


Figure 10 Time series of JERS-1 L-band SAR imagery is applied to assess the spatial complexity in freeze/thaw transition for a boreal landscape. The top series of four images shows backscatter for a 25 km × 25 km region of boreal forest near Fairbanks, Alaska. The middle series shows the corresponding classified freeze/thaw maps. In the bottom series, the freeze/thaw maps have been merged with a land-cover classification map to illustrate the variability of thaw transition with land cover. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

regional and continental scale observations from a variety of satellite-borne active and passive microwave sensors. Application of this information ranges from studies of global change and growing season dynamics to regional monitoring of snowmelt timing and seasonal flooding. Though a variety of current satellite-borne microwave sensors are capable of quantifying landscape freeze-thaw transitions, most are not optimized for this purpose and lack the optimal combination of measurement wavelength, spatial resolution, temporal repeat, and global coverage.

The large range of seasonal and annual variability of freeze/thaw transitions within the cryosphere and the broad ranging impact of this state variable on regional and even global climate, hydrology, and biogeochemistry would suggest that regular, accurate monitoring be a priority in global change research. However, at high latitudes and high elevations the number density of reporting surface weather stations and of hydrological monitoring stations is very low. While optical satellite data suffer from cloud contamination and seasonal solar illumination problems, satellite-borne active and passive microwave remote sensing of landscape freeze/thaw processes may have increased scientific and policy significance for the future. A variety of current satellite-borne microwave sensors are useful for detecting freeze/thaw processes. Development is also under way for a new combined active-passive microwave sensor system with more optimal spectral, spatial, and temporal sampling capabilities for dedicated monitoring of freeze-thaw processes for the global cryosphere. The Hydrosphere State Mission (HYDROS) is currently being developed under NASA's Earth System Science Pathfinder (ESSP) program. This planned satellite mission will provide active and passive L-band observations of the Earth's soil moisture and land surface freeze/thaw conditions for improving our understanding of the linkages between global energy, water, and carbon cycles (Entekhabi *et al.*, 2004). Other missions are also being developed specifically for improved hydrological monitoring of snow and freeze-thaw processes over cold land regions (Cline *et al.*, 1999). Thus regional assessment and monitoring of the freeze/thaw variable will continue, as will the application of this variable for hydrological and global change research.

Acknowledgments

We thank Erika Podest (California Institute of Technology and the University of Dundee, Scotland) for her assistance with the JERS SAR Bonanza Creek data set. We thank Michael Rawlins (University of New Hampshire) for his assistance with the Yukon River catchment data processing. We are grateful to Dr. Donald Cline (National Oceanic and Atmospheric Administration) and Dr. Alan Betts for their helpful input. SeaWinds data were provided by the NASA Ocean Vector Winds Science Team and the Physical Oceanography Distributed Active Archive Center (PO.DAAC) at the Jet Propulsion Laboratory, Pasadena, CA. Nimbus-7 SMMR daily EASE-Grid brightness temperatures and NOAA/NASA Pathfinder SSM/I Level 3 EASE-Grid Brightness Temperatures were obtained from the EOSDIS NSIDC Distributed Active Archive Center (NSIDC DAAC), University of Colorado at Boulder. This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and at the University of Montana, Missoula, under contract to the National Aeronautics and Space Administration.

REFERENCES

Armstrong R.L. and Brodzik M.J. (1995) An earth-gridded SSM/I data set for cryospheric studies and global change monitoring. *Advances in Space Research*, **16**(10), 155–163.

Armstrong R.L., Knowles K.W., Brodzik M.J. and Hardman M.A. (1994) [2000–2002] *DMSP SSM/I Pathfinder Daily EASE-Grid Brightness Temperatures*, National Snow and Ice Data Center, Digital Media and CD-ROM: Boulder.

Betts A.K. and Ball J.H. (1997) Albedo over the boreal forest. *Journal of Geophysical Research*, **102**, 28901–28910.

Betts A.K., Viterbo P., Beljaars A., Pan H.-L., Hong S.-Y., Goulden M. and Wofsy S. (1998) Evaluation of land-surface interaction in ECMWF and NCEP/NCAR reanalysis models over grassland (FIFE) and boreal forest (BOREAS). *Journal of Geophysical Research*, **103**(D18), 23079–23085.

Betts A.K., Viterbo P., Beljaars A.C.M. and van den Hurk B.J.J.M. (2001a) Impact of BOREAS on the ECMWF forecast model. *Journal of Geophysical Research*, **106**(D24), 33593–33604.

Betts A.K., Ball J.H. and McCaughey J.H. (2001b) Near-surface climate in the boreal forest. *Journal of Geophysical Research*, **106**(D24), 33529–33541.

Black T.A., Chen W.J., Barr A.G., Arain M.A., Chen Z., Nesic Z., Hogg E.H., Neumann H.H. and Yang P.C. (2000) Increased carbon sequestration by a boreal deciduous forest in years with a warm spring. *Geophysical Research Letters*, **27**(9), 1271–1274.

Canny J.F. (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**, 679–698.

Cihlar J., Li H., Li Z., Chen J., Pokrant H. and Huang F. (1997) Multitemporal, multichannel AVHRR data sets for land biosphere studies – artifacts and corrections. *Remote Sensing of Environment*, **60**, 35–57.

Cline D., Davis R.E., Edelstein W., Hilland J., McDonald K., Running S., Way J.B. and Van Zyl J. (1999) Cold land processes mission (EX-7) science and technology implementation plan, *Report on the NASA Post-2002 Land Surface Hydrology Planning Workshop*, Irvine, 12–14 April 1999.

Early D.S. and Long D.G. (2001) Image reconstruction and enhanced resolution imaging from irregular samples. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(2), 291–302.

Elachi C. (1987) *Introduction to the Physics and Techniques of Remote Sensing*, John Wiley & Sons: New York, p. 413.

England A.W. (1990) Radiobrightness of diurnally heated, freezing soil. *IEEE Transactions on Geoscience and Remote Sensing*, **28**(4), 464–476.

Entekhabi D., Njoku E., Hauser P., Spencer M., Doiron T., Smith J., Girard R., Belair S., Crow W., Jackson T., *et al.* (2004) The Hydrosphere State (HYDROS) mission concept: an earth system pathfinder for global mapping of soil moisture and land freeze/thaw. *IEEE Transactions on Geoscience and Remote Sensing*, **42**(10), 2184–2195.

Frolking S., McDonald K.C., Kimball J., Way J.B., Zimmermann R. and Running S.W. (1999) Using the space-borne NASA scatterometer (NSCAT) to determine the frozen and thawed seasons of a boreal landscape. *Journal of Geophysical Research*, **104**(D22), 27895–27907.

Gamon J.A., Huemmrich K.F., Peddle D.R., Chen J., Fuentes D., Hall F.G., Kimball J.S., Goetz S., Gu J., McDonald K.C., *et al.* (2004) Remote sensing in BOREAS: lessons learned. *Remote Sensing of Environment*, BOREAS special issue, **89**, 139–162.

Goulden M.L., Wofsy S.C., Harden J.W., Trumbore S.E., Crill P.M., Gower S.T., Fries T., Daube B.C., Fan S.-M., Sutton D.J., *et al.* (1998) Sensitivity of boreal forest carbon balance to soil thaw. *Science*, **279**(9), 214–217.

Harazono Y., Mano M., Miyata A., Zulueta R.C. and Oechel W.C. (2003) Inter-annual carbon dioxide uptake of a wet sedge tundra ecosystem in the arctic. *Tellus*, **55B**, 215–231.

Jarvis P. and Linder S. (2000) Constraints to growth of boreal forests. *Nature*, **405**, 904–905.

Josberger E.G., Mognard N.M., Lind B., Matthews R. and Carroll T. (1998) Snowpack water-equivalent estimates from satellite and aircraft remote-sensing measurements of the red river basin, north-central USA. *Annals of Glaciology*, **26**, 119–124.

Judge J., Galantowicz J.F. and England A.W. (2001) A comparison of ground-based and satellite-borne microwave radiometric observations in the great plains. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(8), 1686–1696.

Judge J., Galantowicz J.F., England A.W. and Dahl P. (1997) Freeze/Thaw classification for prairie soils using SSM/I radiobrightnesses. *IEEE Transactions on Geoscience and Remote Sensing*, **35**(4), 827–832.

Karl T.R. (1995) Long-term climate monitoring by the global climate observing system. *Climatic Change*, **31**(2–4), 185–221.

Kimball J., McDonald K.C., Frolking S. and Running S.W. (2004a) Radar remote sensing of the spring thaw transition across a boreal landscape. *Remote Sensing of Environment*, BOREAS special issue, **89**, 163–175.

Kimball J., McDonald K.C., Running S.W. and Frolking S. (2004b) Satellite radar remote sensing of seasonal growing seasons for boreal and sub-alpine evergreen forests. *Remote Sensing of Environment*, **90**, 243–258.

Kimball J., McDonald K.C., Keyser A.R., Frolking S. and Running S.W. (2001) Application of the NASA scatterometer (NSCAT) for classifying the daily frozen and non-frozen landscape of Alaska. *Remote Sensing of Environment*, **75**, 113–126.

Knowles K., Njoku E., Armstrong R. and Brodzik M.J. (2002) *Nimbus-7 SMMR Pathfinder Daily EASE-Grid Brightness Temperatures*, National Snow and Ice Data Center, Digital Media and CD-ROM: Boulder.

Kraszewski A. (Ed.) (1996) *Microwave Aquametry: Electromagnetic Wave Interaction with Water-Containing Materials*, IEEE Press: Piscataway, p. 484.

Kuga Y., Whitt M.W., McDonald K.C. and Ulaby F.T. (1990) Scattering models for distributed targets. In *Radar Polarimetry for Geoscience Applications*, Ulaby F.T. and Elachi C. (Eds.), Artech House: Dedham.

Lafleur P.M., McCaughey J.H., Joiner D.W., Bartlett P.A. and Jelinski D.E. (1997) Seasonal trends in energy, water, and carbon dioxide fluxes at a northern boreal peatland. *Journal of Geophysical Research*, **102**(D24), 29009–29020.

Lanfear K.L. and Hirsch R.M. (1999) USGS study reveals a decline in long-record streamgages. *EOS*, **80**(50), 605–607.

Magnuson J.J., Robertson D.M., Benson B.J., Wynne R.H., Livingstone D.M., Arai T., Assel T.A., Barry R.G., Card V., Kuusisto E., *et al.* (2000) Historical trends in land and river ice cover in the northern hemisphere. *Science*, **289**, 1743–1746.

McDonald K.C., Kimball J.S., Njoku E., Zimmermann R. and Zhao M. (2004) Variability in springtime thaw in the terrestrial high latitudes: monitoring a major control on the biospheric assimilation of atmospheric CO2 with spaceborne microwave remote sensing. *Earth Interactions*, **8**(20).

Mognard N.M., Baillarin S. and Kerr Y.H. (1998) Monitoring of the boreal regions from 1992 to 1996 with the special sensor microwave imager. *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS)*, **3**, 1268–1270.

Myneni R.B., Keeling C.D., Tucker C.J., Asrar G. and Nemani R.R. (1997) Increased plant growth in the northern high latitudes from 1981–1991. *Nature*, **386**, 698–702.

Raney K.R. (1998) Radar fundamentals: technical perspective. In *Principles and Applications of Imaging Radar*, Vol. 2, Henderson F.M. and Lewis A.J. (Eds.), John Wiley & Sons: New York, pp. 9–130.

Rawlins M.A., McDonald K.C., Frolking S., Lammers R.B., Fahnestock M., Kimball J.S. and Vörösmarty C.J. (2004) Remote sensing of pan-arctic snowpack thaw using the seawinds scatterometer. *Journal of Hydrology*, in press.

Rignot E. and Way J.B. (1994) Monitoring freeze-thaw cycles along north-south Alaskan transects using ERS-1 SAR. *Remote Sensing of Environment*, **49**, 131–137.

Rignot E., Way J.B., McDonald K., Viereck L., Williams C., Adams P., Payne C., Wood W. and Shi J. (1994) Monitoring of environmental conditions in taiga forests using ERS-1 SAR. *Remote Sensing of Environment*, **49**, 145–154.

Running S.W. (1998) A blueprint for improved global change monitoring of the terrestrial biosphere. *Earth Observer*, **10**, 8–13.

Running S., Way J.B., McDonald K.C., Kimball J. and Frolking S. (1999) Radar remote sensing proposed for monitoring freeze-thaw transitions in boreal regions. *American Geophysical Union EOS Newsletter*, **80**(19), 220–221.

Ulaby F.T., Moore R.K. and Fung A.K. (1986) *Microwave Remote Sensing: Active and Passive*, Vols. 1–3, Artech House: Dedham.

Ulaby F.T., Sarabandi K., McDonald K., Whitt M. and Dobson M.C. (1990) Michigan Microwave Canopy Scattering model (MIMICS). *International Journal of Remote Sensing*, **11**(7), 1223–1253.

Vaganov E.A., Hughes M.K., Kirdyanov A.V., Schweingruber F.H. and Silkin P.P. (1999) Influence of snowfall and melt timing on tree growth in subarctic eurasia. *Nature*, **400**, 149–151.

Way J.B., Paris J., Kasischke E., Slaughter C., Viereck L., Christensen N., Dobson M.C., Ulaby F.T., Richards J., Milne A., *et al.* (1990) The effect of changing environmental conditions on microwave signatures of forest ecosystems: preliminary results of the march 1988 Alaskan aircraft SAR experiment. *International Journal of Remote Sensing*, **11**, 1119–1144.

Way J.B., Rignot E., McDonald K., Oren R., Kwok R., Bonan G., Dobson M.C., Viereck L. and Roth J.E. (1994) Evaluating the type and state of Alaska taiga forests with imaging radar for use in ecosystem flux models. *IEEE Transactions on Geoscience and Remote Sensing*, **32**(2), 353–370.

Way J.B., Zimmermann R., Rignot E., McDonald K. and Oren R. (1997) Winter and spring thaw as observed with imaging radar at BOREAS. *Journal of Geophysical Research*, **102**(D24), 29673–29684.

Wegmuller U. (1990) The effect of freezing and thawing on the microwave signatures of bare soil. *Remote Sensing of Environment*, **33**, 123–135.

Wismann V. (2000) Monitoring of seasonal thawing in siberia with ERS scatterometer data. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 1804–1809.

Zhang T., Armstrong R.L. and Smith J. (2003) Investigation of the near-surface soil freeze-thaw cycle in the contiguous United States: algorithm development and validation. *Journal of Geophysical Research*, **108**(D22), 8860–8874.

Zuerndorfer B. and England A.W. (1992) Radiobrightness decision criteria for freeze/thaw boundaries. *IEEE Transactions on Geoscience and Remote Sensing*, **30**(1), 89–102.

54: Estimation of Surface Soil Moisture Using Microwave Sensors

THOMAS J JACKSON

USDA ARS Hydrology and Remote Sensing Lab, Beltsville, MD, US

Soil moisture has been particularly difficult to measure and map using conventional ground-based point sampling. Microwave remote sensing–based soil moisture retrieval has been demonstrated using tower and aircraft instruments. The translation of this approach to satellites and the implementation in hydrologic applications has been limited by both the technology and the satellite systems that were available. Recent developments in both science and associated technologies now make the exploitation of the microwave region for soil moisture mapping feasible. A number of new and potential satellite missions are reviewed and evaluated for soil moisture applications.

INTRODUCTION

Improvements in hydrological modeling can be achieved through observation of the current status of soil moisture. More accurate predictions from these models will lead to better forecasts of floods and other hydrological phenomena. Soil moisture products are now feasible using a new generation of microwave remote sensing satellites. The quality of these products will continue to improve over time as new sensors are launched.

Leese *et al.* (2001) concluded that the optimal approach to monitoring soil moisture would be a combination of model derived estimates using *in situ* measurements and estimates derived from remote sensing data as input data for assimilation (Houser *et al.*, 1998). In this regard, each method produces soil moisture values that are both unique and complementary. These satellite products, combined with existing *in situ* observations, should be exploited in hydrological monitoring, assessment, and prediction.

Remote sensing has been a valuable tool in some aspects of hydrological modeling. Advantages include synoptic coverage, temporal repeat, and in most situations the unrestricted collection and distribution of information.

Microwave sensors are of particular value in hydrology because they respond to the amount of moisture in the soil. There are several methods that have been shown to be capable of providing soil moisture information. Each has unique capabilities (i.e. temporal coverage, spatial resolution) that must be matched with specific types of applications. In the past, options have been limited by available satellite systems. Investigators have demonstrated the potential of these data in hydrologic studies using ground and aircraft systems. However, efforts to use the less than optimal available satellite systems have had very limited success. In recent years, and continuing over the coming decade, a wide range of new and significantly improved satellites will be launched that will offer new opportunities. A description of alternative techniques will be presented along with a review of current and future satellite systems and products.

MICROWAVE REMOTE SENSING OF SURFACE SOIL MOISTURE

Microwave remote sensing provides a direct measurement of the surface soil moisture for a range of vegetation cover conditions within reasonable error bounds. Two basic approaches are used, passive and active. In passive methods, the natural thermal emission of the land surface (or brightness temperature) is measured at microwave wavelengths, using very sensitive detectors. In active methods or radar, a microwave pulse is sent and received. The power

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

of the received signal is compared to that which was sent to determine the backscattering coefficient.

Portions of the microwave region of the electromagnetic spectrum are often referred to as bands, which are identified by a lettering system. Some of the relevant bands that are used in Earth remote sensing are: K (18–27 GHz), X (8–12 GHz), C (4–8 GHz), and L (1–2 GHz). Within these bands, only small ranges exist that are protected for scientific applications, such as radioastronomy and passive sensing of the Earth's surface. There is evidence that there is increasing radio frequency interference (RFI) occurring throughout the world as a result of communications networks.

Microwave sensors operating at very low microwave frequencies (<6 GHz) provide the best soil moisture information. At low frequencies, attenuation and scattering problems associated with the atmosphere and vegetation are less significant, the instruments respond to a deeper soil layer, and have a higher sensitivity to soil water content variations.

Active and passive sensors provide different types of measurements, which are determined by various physical phenomena. However, both types of sensors provide information on the surface reflectivity. The essential relationship needed to utilize microwave remote sensing, is a link between reflectivity and soil moisture. The Fresnel reflection equations are generally used for this purpose (Ulaby *et al.*, 1986). These equations predict the surface reflectivity (r) as a function of dielectric constant (k) and the viewing angle (θ), based on the polarization of the sensor (horizontal-H or vertical-V).

$$r^{\mathrm{H}} = \left| \frac{\cos\theta - \sqrt{k - \sin^2\theta}}{\cos\theta + \sqrt{k - \sin^2\theta}} \right|^2 \tag{1}$$

$$r^{\mathrm{V}} = \left| \frac{k\cos\theta - \sqrt{k - \sin^2\theta}}{k\cos\theta + \sqrt{k - \sin^2\theta}} \right|^2 \tag{2}$$

From the reflectivity, the dielectric constant of the soil can be estimated. The dielectric constant of soil is a composite of the values of its components: air, soil, and water. The basic reason microwave remote sensing is capable of providing soil moisture information is that there is a large difference between the dielectric constants of water (~80) and the other components (~4). On the basis of an estimate of the mixture dielectric constant derived from the Fresnel equations and soil texture information, volumetric soil moisture can be estimated using an inversion of the model. The depth of soil measured is approximately one-fourth the wavelength (wavelength~29.6/frequency (GHz)).

Passive Microwave Techniques

Passive microwave remote sensing utilizes radiometers that measure the natural thermal radio emission within a narrow band centered on a particular frequency. The measurement provided is the brightness temperature in degrees Kelvin, T_{B}, which includes contributions from the atmosphere, reflected sky radiation, and the land surface. Atmospheric contributions are negligible at frequencies <10 GHz, and the following section assumes this to be the case. Galactic and cosmic radiations contribute to sky radiation and have a known value that varies very little in the frequency range used for observations of soil water content ($T_{\mathrm{sky}} \sim 4$ K). The observed brightness temperature of a surface is equal to its emissivity (e_{obs}) multiplied by its physical temperature (T). The observed emissivity is equal to 1 minus the reflectivity, which provides the link to the Fresnel equations and soil moisture for passive microwave remote sensing.

On the basis of the above and neglecting atmospheric contributions, for a specific frequency (f), polarization (p), and viewing angle (θ), the equation for T_{B} is

$$T_{\mathrm{B}_{fp\vartheta}} = \mathrm{e}_{\mathrm{obs}_{fp\vartheta}} T + (1 - \mathrm{e}_{\mathrm{obs}_{fp\vartheta}}) T_{\mathrm{sky}} \tag{3}$$

The second term of equation 3 is about 2 K. A typical range of response for a soil is 60 K; therefore, the reflected sky contribution can be dropped for computational purposes. Equation 3 can be rearranged as follows:

$$\mathrm{e}_{\mathrm{obs}_{fp\vartheta}} = \frac{T_{\mathrm{B}_{fp\vartheta}}}{T} \tag{4}$$

If the physical temperature is estimated independently, the emissivity can be determined from T_{B}. The physical temperature can be estimated using surrogates based on satellite surface temperature, air-temperature observations, or forecast model predictions. As noted above, the derived emissivity is for a specific instrument configuration (frequency, polarization, and viewing angle).

For natural conditions, varying degrees of vegetation type and density are likely to be encountered. The presence of vegetation has a major impact on the microwave measurement. Vegetation reduces the sensitivity of the retrieval algorithm to changes in soil water content by attenuating the soil signal and by adding a microwave emission of its own to the microwave measurement. This attenuation increases with increasing microwave frequency, which is another important reason for using lower frequencies. Attenuation is characterized by the optical depth of the vegetation canopy (Jackson and Schmugge, 1991). A widely used correction for these effects uses information on the vegetation type (parameter b) and the vegetation water content (W_{veg}) as follows:

$$\mathrm{e}_{\mathrm{surf}_{fp\vartheta}} = 1 + [\mathrm{e}_{\mathrm{obs}_{fp\vartheta}} - 1] \exp[-2bW_{\mathrm{veg}} \cos\theta] \tag{5}$$

The corrected emissivity represents that of the soil surface (e_{surf}). Methods have been developed for estimating vegetation type parameter (typically derived from land cover) and vegetation water content using visible/near-infrared remote sensing.

One of the key concepts in passive microwave remote sensing is the effect of frequency on the retrieval capabilities. To illustrate this, sensitivity will be used as defined by the change in T_B for a 1% change in soil moisture. A good system design seeks to maximize sensitivity for a wide range of vegetation conditions. Using simple radiative transfer models (Jackson, 1993), it is possible to simulate the relationship between soil moisture and T_B for different levels of vegetation and frequencies. The slope of these relationships represents the sensitivity. Sensitivity as a function of frequency is presented in Figure 1 for bare soil and a vegetation canopy with a nominal vegetation water content of $1\,kg\,m^{-2}$. Although the contributing depth changes, the sensitivity for bare soil is high and fairly constant. The effect of frequency is clearly seen in the vegetation curve. At high frequencies ($\sim$19 GHz), there is essentially no sensitivity to soil moisture under these vegetation conditions. However, if we could implement a 1 GHz system, there would be good sensitivity over a wide range of vegetation conditions.

Most research and applications involving passive microwave remote sensing of soil moisture have emphasized low frequencies (L-band). In this range, it is possible to develop soil moisture retrievals based on a single H polarization observation (Jackson, 1993). This approach relies on providing ancillary data on temperature, vegetation, land cover, and soils. Other algorithm approaches are described in Njoku *et al.* (2000, 2003).

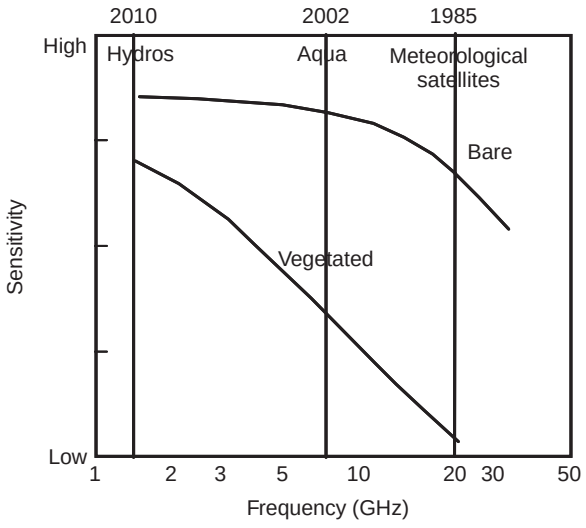


Figure 1 Sensitivity of brightness temperature to soil moisture as a function of frequency

Active Microwave Methods

Active microwave sensors or radars measure the transmitted and received power to yield a variable called the *backscattering coefficient* (σ^o), which can be related to the surface reflectivity (Ulaby *et al.*, 1986). As described previously, the reflectivity can then be used to determine surface soil moisture. For active techniques, the link between the measurement of the backscattering coefficient and the surface reflectivity is a bit more complex than for passive methods. The geometric properties of the soil surface and vegetation have a greater effect on these measurements and simple correction procedures are difficult to develop.

The two most common types of radars are scatterometers and synthetic aperture radars (SAR). Scatterometers are nonimaging radars. In the case of a SAR, the received microwave signal is further processed to produce an image. A SAR uses the phase history of the returned signal to synthesize a very large antenna, and thus significantly improve resolution in the azimuth direction (on the order of 10–100 m even from satellites).

The signals sent and received by a radar system are usually linearly polarized, either horizontal (H) or vertical (V). Transmitter and receiver can each have a different polarization. Possible combinations are HH, VV, HV, and VH. Advanced multipolarization systems can make all of these measurements simultaneously.

Estimating soil water content from radar backscatter is easier where the soil is bare. When there is a vegetation cover, establishing soil water under the canopy is much more difficult. Microwaves transmitted will interact with the vegetation cover. These will be scattered and attenuated by the vegetation. In addition, the energy scattered back towards the sensor is a combination of scattering directly from the canopy and directly from the soil, as well as multiple scattering that results from the signal interacting with both the soil and the canopy. Estimating soil water content under a vegetation canopy is difficult and requires unraveling the contribution of the soil itself from that of vegetation.

The most common approach to estimating soil moisture from backscatter has been linear regression (Brown *et al.*, 1993). This of course does not result in a robust retrieval algorithm. More theoretical approaches have limited applicability and can be difficult to implement. A more promising technique involves semiempirical models. These involve multiple polarization observations and restrictions on the range of applicability. Algorithms incorporating this approach for bare soils are presented in Dubois *et al.* (1995). A robust approach to account for vegetation has not yet been developed and validated.

Table 1 Microwave satellite systems

Satellite Instrument	Period of coverage	Frequency (GHz)	Polarization	Horizontal resolution (km)	Repeat frequency (days)
Passive					
SSM/I	1987–present	19.4, 22.2 37.0, 85.5	H and V	70 to 5 km	1–2
TMI	1998–present	10.7, 19.4, 21.3, 37.0, 85.5	H and V	60 to 6 km	1
AMSR-E (NASA)	2002–present	6.9, 10.7, 18.7, 23.8, 36.5, 89.0	H and V	75 to 7 km	2 – 3
AMSR (Japan)	2002	6.9, 10.7, 18.7, 23.8, 36.5, 89.0	H and V	70 to 6 km	2 – 3
Windsat	2003–present	6.8, 10.7, 18.7, 23.8, 37	H and V (U in some channels)	50 to 10 km	2–3
SMOS	(2007)	1.4	H and V	50 km	2–3
Hydros	(2010)	1.4	H and V (passive) HH, VV (active)	3 to 40 km	2–3
Active					
ERS	1991–present	5.3	VV	30 m	35
Radarsat-1	1995–present	5.3	HH	7 to 100 m	24
ASAR	2002–present	5.3	HH and VV, or HH and HV, or VV and VH	30 to 1000 m	35
Radarsat-2	(2005)	5.3	HH, VV, HV and VH	3 to 100 m	24
Palsar	(2005)	1.27	HH or VV and HV or VH	10 to 100 m	46

LARGE-AREA MULTITEMPORAL AIRCRAFT MAPPING

Key advances in soil moisture remote sensing include not only algorithms and sensors but also the demonstration of the value and potential of the data in a hydrological context. This has been achieved using experimental sensors on aircraft over well-instrumented sites as part of comprehensive field experiments. Many of these have taken place in Oklahoma.

One of the most important of these experiments was Washita'92. Washita'92 was a large-scale study of remote sensing and hydrology, conducted over the Little Washita watershed in southwest Oklahoma (Jackson *et al.*, 1995). Data collection during the experiment included passive and active microwave observations. Data were collected for nine days in June 1992. The watershed was saturated with a great deal of standing water at the outset of the study. During the experiment there was no rainfall, and surface soil water content observations exhibited a dry-down pattern over the period. These fortunate circumstances allowed the examination of post-rainfall hydrologic phenomena over an extended period. Surface soil water content observations were made at sites distributed over the area. Significant variations in the level and rate of change in surface soil water content were noted over areas dominated by different soil textures.

Another unique aspect of Washita'92 was the use of a new aircraft instrument called the *Electronically Scanned Thinned Array Radiometer* (ESTAR). ESTAR represented a breakthrough approach to solving the spatial resolution problem of low frequency passive microwave remote sensing from space. Passive microwave observations were made on eight days. The ESTAR data were processed to produce brightness temperature maps of a 740-km^2 area on each of the eight days. Using the single-channel soil water content algorithm described in Jackson (1993), these data were converted to soil water content images. Gray-scale images for each day are shown in Figure 2. These data exhibited significant spatial and temporal patterns. Spatial patterns were clearly associated with soil textures, and temporal patterns with drainage and evaporative processes. Relationships between the ground sampled soil water content and the brightness temperatures were consistent with previous results.

An aspect of spatial scaling of soil water content sensing was investigated using this data set. All of the soil water content samples collected on a given day were averaged for the study area. This same procedure was used for the brightness temperature, which was then converted to an emissivity estimate by normalizing with the averaged soil temperature data. This results in one pair of emissivity and soil water content values for each of the eight days. When compared with the Fresnel-based predicted relationship, there was very good agreement (standard error of estimate $<5\%$) which indicated that the data interpretation algorithms apply within this region, and that large-scale averaging (740 km^2) does not degrade their predictive ability.

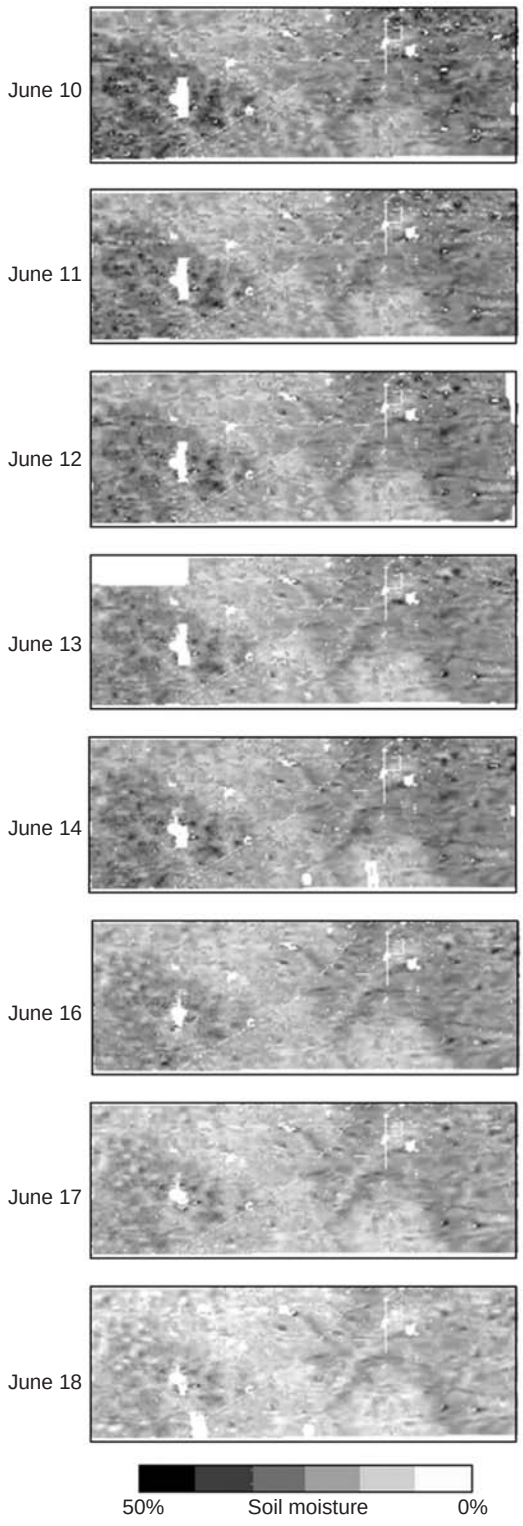


Figure 2 Washita'92 soil moisture

Numerous hydrologic analyses resulted from the Washita'92 soil moisture data sets. These included analysis of scaling (Rodruigez-Iturbe *et al.*, 1995; Dubayah *et al.*, 1997), hydraulic property derivation (Mattikalli *et al.*, 1998a,b), geostatistical relationships (Cosh and Brutsaert, 1999; Lin *et al.*, 1997), and energy balance (Kustas and Jackson, 1999).

Using the model developed in Washita'92, larger and more diverse experiments were conducted in this same region in 1997 and 1999. The 1997 experiments expanded the spatial domain an order of magnitude and the temporal period to one month (Jackson *et al.*, 1999). With these expanded datasets it was possible to further develop concepts established in Washita'92 (Kustas *et al.*, 2001; Kim and Barros, 2002; Crow and Wood 2002) and explore the development of soil moisture data assimilation schemes (Reichle *et al.*, 2001; Margulis *et al.*, 2002). In the 1999 experiment (Jackson *et al.*, 2002), simulators of near future C-band satellite instruments were used. The focus in 1999 was on current and near future satellite instruments.

HYDRAULIC PROPERTIES OF SOILS

One of the most significant findings from the Washita'92 experiment described above was not part of the initial experiment plan. This result involved using the microwave observations to derive soil properties. Soil physical and hydraulic properties are of value for water balance studies and simulations. Although it is possible to obtain these data at the field point scale, the data are not available over large areas, do not take into account spatial variability, and are costly and difficult to measure.

Temporal observations of soil moisture were used in the work of Ahuja *et al.* (1993), Mattikalli *et al.* (1998a), Hollenbeck *et al.* (1996), and Burke *et al.* (1998). These studies constrained the problem by minimizing the role of the surface flux, or by specifying it, in order to estimate the hydraulic properties of the integrated soil profile.

A few studies have employed ground-based passive microwave remote sensing and physical models for the estimation of the saturated hydraulic conductivity (Camillo *et al.*, 1986; Van de Griend and O'Neill, 1986). In these studies, the soil properties at the plot scale were derived through an iterative process of matching remotely sensed T_B with values generated by a microwave simulation model. This approach requires detailed inputs describing the profile soil water content and temperature that will not be available over large areas.

Ahuja *et al.* (1993) hypothesized that the change in soil surface moisture followed a wetting event and two days later could be used to determine the profile average hydraulic conductivity. Using simulations for soils of different textures, they found that a strong relationship existed. This was highest for sandy soils and decreased as clay content increased. In addition, they found that a measurement of the 0–5-cm layer provided as much information as a 0–30-cm observation.

These relationships were affected by evaporation; however, it was more significant for soils with lower hydraulic conductivity and with a shallow depth of observation. Tests using observations showed R^2 values that were generally greater than 0.6. Chen *et al.* (1993) followed up on this approach and utilized a time series of soil moisture to estimate hydraulic properties of macropore soils, which are difficult to be characterized on the basis of readily available soils data.

Hollenbeck *et al.* (1996) explored the use of passive microwave observations, 1.4 GHz, for identifying areas of relatively slow or fast drainage in a region. These differences were interpreted as heterogeneity of the soil hydraulic properties in the region. In this study, the responses were correlated with soil type and geomorphic features. The authors noted that antecedent conditions and the timing of the observations will impact the value of the data in retrieving soil properties.

Mattikalli *et al.* (1998a,b) took the approach presented by Ahuja *et al.* (1993) a step further by utilizing remotely sensed data to provide soil moisture information on a spatial basis. They hypothesized that surface soil moisture changes could be used to identify soil texture classes over an area. An important point that the authors made is that even if this method was not highly accurate for each and every sensor resolution element (pixel), the information provided would significantly improve the definition of the spatial domain of the soil hydraulic properties. In their opinion this was more valuable than a few accurate point estimates.

Mattikalli *et al.* (1998b) used microwave remote sensing data from the Washita'92 experiment described earlier, to obtain spatial and multitemporal soil water content data over the Little Washita watershed in Oklahoma. Analysis of the multitemporal water content maps with the soil maps suggested that the remotely sensed soil moisture and its temporal changes could be used to identify soil texture and to estimate soil hydraulic properties as hypothesized. Validations of the technique indicated that this could be accomplished with acceptable accuracy.

SATELLITE OBSERVING SYSTEMS AND PRODUCTS

Satellite-based sensors offer the advantages of large-area mapping and long-term repetitive coverage. Revisit time can be a critical problem in studies involving rapidly changing conditions, such as surface soil water content. With very wide swaths it is possible to obtain twice daily coverage with a polar orbiting satellite. For most satellites, especially if constant or narrow ranges of the viewing angle are important, the revisit time can be much longer. Optimizing the time and frequency of coverage is a critical problem for studies of soil water content.

Current Passive Systems

Special Satellite Microwave/Imager

Currently, all passive microwave sensors on satellite platforms operate at high frequencies (>6 GHz). Of particular note, due to the longevity of its data record, is the Special Satellite Microwave/Imager (SSM/I) package on the Defense Meteorological Satellite Platforms. These satellites have been in operation since 1987, and provide high frequencies and two polarizations (see Table 1) except 22 GHz (V only).

Interpreting data from the SSM/I to extract surface information requires accounting for atmospheric effects on the measurement (Drusch *et al.*, 2001). When one considers the atmospheric correction, the significance of vegetation attenuation, and the shallow contributing depth of soil for these high frequencies, it becomes apparent that the data are of limited value for estimating soil water content. However, data from the SSM/I can be used under some circumstances, such as in arid and semiarid areas with low amounts of vegetation. Spatial resolution of the SSM/I is very coarse, as shown in Table 1. The SSM/I utilizes conical scanning, which provides measurements at the same viewing angle at all beam positions on a swath. This makes data interpretation more straightforward and simplifies image comparisons. There have been as many as four SSM/I satellites in operation during any given period. Therefore, frequent and even multiple daily passes are typical for most regions of the Earth. Data from the SSM/I are publicly available.

Value-added products from the SSM/I sensors include a wide range of atmospheric and ocean variables. However, for reasons noted above there have been few attempts at generating standard land surface products. NOAA (Basist *et al.*, 1998) utilizes SSM/I data to produce an experimental data product called the Soil Wetness Index (SWI). The key point to recognize with regard to this index is that it is intended to provide information on significantly wet soil conditions (areal extent of flooding), which can be more reliably detected than variations at lower levels of soil moisture. In most cases, these will involve some standing surface water, which will result in a large decrease in brightness temperature.

The SWI uses the difference between the 85 GHz and 19 GHz horizontally polarized brightness temperatures from the SSM/I instrument. These differences are scaled to enhance the extremely wet to flooded soil conditions. Data are composited as temporal five day averages in order to identify flooded areas, which may otherwise be obscured by precipitation. Spatial resolution is 0.3 degrees. For each 0.3-degree pixel, the current SWI is compared to the historic mean and standard deviation (1987–present) as follows

$$\mathrm{SWIZ}_i = \frac{(\mathrm{SWI}_i - \mathrm{Mean}(\mathrm{SWI}_i))}{\mathrm{Sigma}(\mathrm{SWI}_i)}, \qquad (6)$$

where:

i = the ith five-day period: $i = 1, 2, \ldots, 73$
Mean (SWI_i) = the average pixel level SWI for the ith period
Sigma (SWI_i) = the standard deviation of SWI for the ith period.

As a flood index, only the positive anomalies of the SWI are important. These are on a scale of 1 to 3. SWIZ values in the range 1–2 are associated with wet surface conditions; 2–3 are indicative of extremely wet conditions; and >=3 have the potential for flooding. Standard products include daily, five-day composite, and an anomaly product for the continental US and globally.

A few studies have attempted to extract actual soil moisture from SSM/I data (Jackson, 1997; Lakshmi *et al.*, 1997; Vinnikov *et al.*, 1999; Owe *et al.*, 1992). In Jackson (1997), the single channel/ancillary data approach described in Jackson (1993) for L-band, was extended to the higher frequencies of the SSM/I instrument. For the limited validation data set available, the approach performed relatively well with a standard error estimate of 5.3%. It is likely that the success of the approach was related to the limited conditions and light vegetation conditions evaluated.

Tropical Rainfall Measurement Mission Microwave Imager

Another option is the Tropical Rainfall Measurement Mission Microwave Imager (TMI) on the TRMM satellite (Kummerow *et al.*, 1998). It is a five-channel, dual-polarized, passive microwave radiometer. The lowest TMI frequency is 10 GHz (see Table 1), about half that of the SSM/I. The TMI has a higher spatial resolution as compared to the SSM/I. Tropical Rainfall Measurement Mission or TRMM only provides coverage of the tropics, which includes latitudes between 38 °N and 38 °S for the TMI instrument. However, a unique capability of the TMI is its ability to collect data daily, and in many cases more often, within certain latitude ranges. This could facilitate multitemporal and diurnal analyses. Jackson and Hsu (2001), and Wen and Su (2003) demonstrated the potential of using these data to retrieve soil moisture. Bindlish *et al.* (2003) have developed and validated a five-year data set for the Southern United States based upon the TMI data. These studies show the potential of the improved spatial resolution, higher temporal repeat coverage, and lower frequency as compared to the SSM/I. Figure 3 illustrates the difference in the spatial resolution and density of observations between the SSM/I and the TMI instruments. The enhanced information of the TMI results in images with increased spatial detail and structure.

Advanced Microwave Scanning Radiometer

Several new multifrequency passive microwave satellite systems were launched in 2002 and 2003. As opposed to the previously available systems, these offer a lower frequency channel operating at C-band, which should provide a more robust soil moisture measurement. These satellites are the NASA Aqua, Japanese ADEOS-II, and the US interagency Coriolis satellite.

Aqua and ADEOS-II included an instrument called the Advanced Microwave Scanning Radiometer (AMSR). AMSR is multifrequency (Table 1) and includes a 6.9 GHz (C-band) channel with a nominal 60-km spatial resolution. AMSR holds great promise for estimating soil water content in sparsely vegetated regions and is the best source of data in the near term for mapping soil water. On the basis of published results and supporting theory (Njoku and Li, 1999), this instrument should be able to provide information about soil water content in regions of low vegetation cover, less than 1 kg m^{-2} vegetation water content. There are some small differences in the spatial resolution of AMSR and AMSR-E. The most important difference between the two satellites is the time of day of the overpasses. Aqua is 1:30 am and pm local time and ADEOS-II is 10:30 am and 10:30 pm. The orbits result in 2–3 day coverage in the mid-latitudes. Unfortunately, communications with the ADEOS-II satellite were lost in October 2003.

As opposed to previous passive microwave satellite missions, the Aqua project includes soil moisture as a product. The algorithm planned for use with Aqua is a variation of the multiple channel approach described in Njoku and Li (1999), and Njoku *et al.* (2003). Several types of soil moisture products will be produced. These include a daily swath product and a global composite. The swath products include a retrieval of soil moisture for each pixel observed. Results will be composited to a standard grid to generate a global map of surface soil moisture with a nominal spatial resolution of 25 km. After a period of calibration/validation, the soil moisture products should be available on a daily basis.

A somewhat parallel program is being conducted by the Japanese Aerospace Exploration Agency (JAXA) using the AMSR-E and AMSR data. In this program, soil moisture is considered to be a research product. Four algorithms are under consideration (Koike *et al.*, 2000). All algorithms are being evaluated and compared as part of the calibration/validation program. If one of these is considered acceptable, it will be adopted as a standard product algorithm.

Preliminary studies indicate that there is widespread RFI in the C-band channels. Therefore, it is likely that the most useful channels for soil moisture will be those operating at the slightly higher X band.

The Coriolis satellite includes the Windsat instrument, which is a multifrequency passive microwave radiometer system. It is similar to AMSR with some differences in frequencies and more polarization options (Table 1). It is a prototype of one component of the next generation of

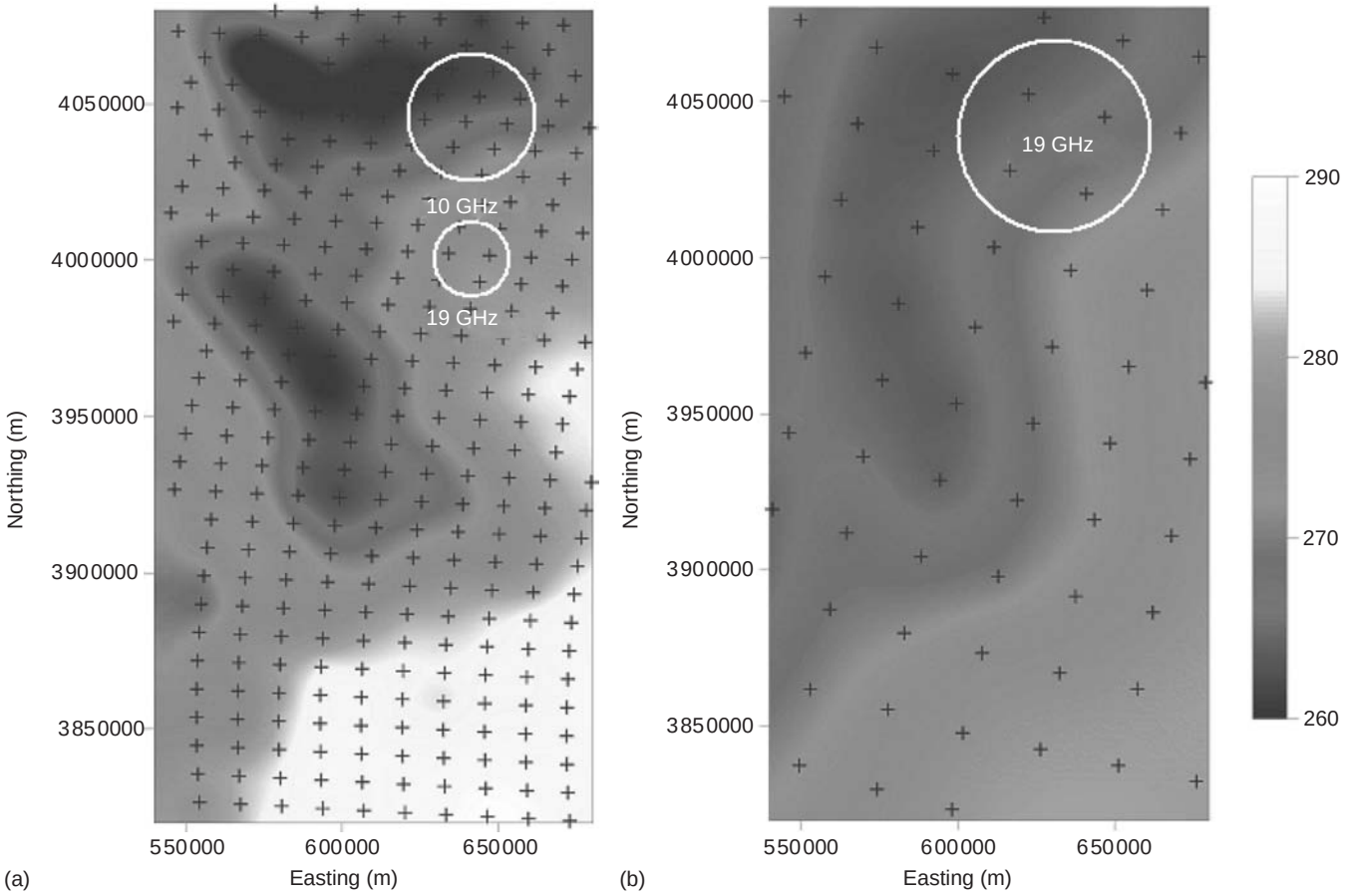


Figure 3 Spatial features of the (a) TMI and (b) SSM/I compared for an area in Oklahoma. The footprint density (+) and size (circles) are shown on interpolated brightness temperature images for one pass. The two TMI footprints shown are approximately 20 km and 40 km, and the SSM/I footprint is 70 km

operational polar orbiting satellites that the United States will implement at the end of this decade. Experience gained using these science missions will provide the basis for future operational products.

Future Passive Systems

Research programs are underway to develop and implement space-based systems with a 1.4 GHz channel that would provide improved global soil moisture information. Toward that goal the European Space Agency (ESA) is developing a sensor system called the *Soil Moisture Ocean Salinity* (*SMOS*) *mission* (Wigneron *et al.*, 2000), and NASA will launch the Hydros mission (Entekhabi *et al.*, 2004).

SMOS will utilize two-dimensional synthetic aperture radiometry at L-band and will provide H and V polarization observations. The data collected during one integration period can be thought of as an image consisting of lines and pixels. A key concept for SMOS is that the incidence angle of each pixel varies on the basis of its position relative to the spacecraft track. For a given data take, it is possible to select a subset of the available pixels within a specific incidence angle range, say 50°. This process can yield a pseudoconical scanned product. It is also possible to select all pixels at another individual incidence angle, which would yield a different subset of observations. By using this idea it is then possible, over selected portions of the swath and data take, to obtain observations of an area on the ground at different incidence angles by using sequential data takes. As described in Wigneron *et al.* (2000), the multiple incidence angle observations of the surface is the key element of the soil moisture retrieval algorithm that is proposed for the mission. The launch of the mission is anticipated in 2007. A minimum temporal coverage of three days is proposed depending on the latitude.

Hydros is an exploratory mission that will provide global measurements of soil moisture and surface freeze/thaw, focusing on regions where these are primary environmental controls. The sampling will be continuous, over the minimum mission duration of three years. Hydros would use an active and passive L-band instrument design for mapping at resolutions ranging from 3 to 40 km depending

upon the product. Hydros data will be merged with data from *in situ* networks and other satellites in a land data assimilation framework, providing value-added end products that will deliver the most comprehensive view yet of Earth's land hydrosphere. Soil moisture will be estimated by Hydros using the radiometer and radar data separately and in combination, taking advantage of the simultaneous, coincident, and complementary nature of the measurements. Both measurement types are sensitive to soil moisture directly. Radiometer measurements currently provide more accurate soil moisture estimates than radar under vegetated conditions because of the maturity of the radiometer-based algorithms compared to those for radar. Radar measurements provide a higher resolution capability and subpixel roughness and vegetation information within the radiometer footprint. The combination of simultaneous radar and radiometer data has the potential to improve the accuracy of soil moisture estimates over heterogeneous surface conditions. Soil moisture will be obtained using both absolute and relative change estimation approaches. Hydros is currently scheduled for 2010.

Current Active Systems

At present, several radar satellites are in orbit (Table 1). The ESA has operated a satellite SAR series called *ERS* since 1991, which provides C-band VV. This satellite includes both a SAR and a scatterometer. Although numerous investigations have been conducted that attempt to utilize ERS-SAR data, few results have been reported in the area of soil moisture estimation. This paucity is due to the limitations of using a single mid-frequency channel-single polarization SAR with an exact repeat cycle of 35 days. With this kind of temporal coverage, the data are of little value in process studies. However, with enough assumptions and effort, it is possible to develop site-specific soil moisture information from the single-channel radar data.

The next satellite in the ERS series is called Envisat and it has C-band Advanced Synthetic Aperture Radar (ASAR) with multiple polarization capabilities. It also offers the option of varying the incidence angle to allow for different viewing angles and more frequent coverage if angle is not a critical parameter in the application. Data from ASAR are just beginning to be made available to investigators.

The Canadian Space Agency operates a C-band satellite SAR called *Radarsat*. Radarsat offers HH polarization and has more flexibility in its data-collection modes. Options include a variable viewing angle and a wide swath (large range of incidence angles). These choices offer more frequent temporal coverage of a particular region of the Earth if angle is not important.

There have been a number of attempts at using the single-channel SAR data to retrieve soil moisture. The general consensus from these studies is that there are too many physical variables that have to be known in order to derive soil moisture (Verhoest *et al.*, 1998; Satalino *et al.*, 2002). These variables include the soil moisture, surface roughness, and vegetation. Site-specific studies and creative selection of data sets have resulted in some degree of success (Altese *et al.*, 1996; Verhoest *et al.*, 1998; Wagner *et al.*, 1999; Le Hegarat-Mascle *et al.*, 2002; Moran *et al.*, 2002; Satalino *et al.*, 2002).

Future Active Systems

Japan will include an L-band SAR called *PALSAR* on an upcoming satellite, the Advanced Land Observing System (ALOS). The lower frequency will offer information that is different from that offered by other satellite SAR systems operating at C-band. PALSAR will have a multipolarization mode as well as varying incidence angles, the advantages of which have been described.

Canada is developing Radarsat-2, which will be similar to Radarsat-1. As with the ASAR, it will be fully polarimetric and consequently, any possible polarization (including, but not limited to HH, VV, HV, and VH) can be generated along with phase information. The satellite will include a mode in which the spatial resolution is 3 m. The expected launch is in 2005.

As noted above, progress in SAR soil moisture mapping has been limited by the available data, single channel. The new generation of multipolarization SAR systems will provide at least one additional measurement. There is hope that more robust soil moisture retrieval methods can be developed as these data become available.

SUMMARY

Soil moisture products based upon satellite microwave remote sensing have become a viable source of information for hydrologic modeling and analysis. New opportunities and improved products will be available within the next five years. Key issues to consider for applications are trade-offs between desired temporal coverage and spatial resolution that are related to active and passive microwave systems.

Currently, passive microwave remote sensing offers more robust and well-validated retrieval algorithms. However, it is limited by spatial resolution at satellite altitudes. Active microwave techniques can provide high spatial resolution data, but suffer from weak soil moisture algorithms. There are many new satellite opportunities within the coming decade that will significantly advance satellite-based soil moisture remote sensing. As these become available, each needs to be evaluated and exploited for hydrological applications.

REFERENCES

Ahuja L.R., Wendroth O. and Nielson D.R. (1993) Relationship between initial drainage of surface soil and average profile saturated conductivity. *Soil Science Society of America Journal*, **57**, 19–25.

Altese E., Bolognani O., Mancini M. and Troch P.A. (1996) Retrieving soil moisture over bare soil from ERS-1 synthetic aperture radar data: sensitivity analysis based on a theoretical scattering model and field data. *Water Resources Research*, **32**, 653–661.

Basist A., Grody N.C., Peterson T.C. and Williams C.N. (1998) Using the special sensor microwave/imager to monitor surface temperatures, wetness, and snow cover. *Journal of Applied Meteorology*, **37**, 888–911.

Bindlish R., Jackson T.J., Wood E., Gao H., Starks P., Bosch D. and Lakshmi V. (2003) Soil moisture estimates from TRMM microwave imager observations over the Southern United States. *Remote Sensing of Environment*, **85**, 507–515.

Brown R.J., Brisco B., Leconte R., Major D.J., Fischer J.A., Reichert G., Korporal K.D., Bullock P.R., Pokrant H. and Culley J. (1993) Potential applications of RADARSAT data to agriculture and hydrology. *Canadian Journal of Remote Sensing*, **19**, 317–329.

Burke E.J., Gurney R.J., Simmonds L.P. and O'Neill P.E. (1998) **U**sing a modeling approach to predict soil hydraulic properties from passive microwave measurements. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 454–462.

Camillo P.J., O'Neill P.E. and Gurney R.J. (1986) Estimating soil hydraulic parameters using passive microwave data. *IEEE Transactions on Geoscience and Remote Sensing*, **24**, 930–936.

Chen C., Thomas D.M., Green R.E. and Wagnet R.J. (1993) Two-domain estimation of hydraulic properties in macropore soils. *Soil Science Society of America Journal*, **57**, 680–686.

Cosh M.H. and Brutsaert W. (1999) Aspects of soil moisture variability in the Washita'92 study region. *Journal of Geophysical Research*, **104**, 19,751–19,757.

Crow W.T. and Wood E.F. (2002) Impact of soil moisture aggregation on surface energy flux prediction during SGP'97. *Geophysical Research Letters*, **29**, 10.1029/2001GL013796.

Drusch M., Wood E.F. and Jackson T.J. (2001) Vegetative and atmospheric corrections for the soil moisture retrieval from passive microwave remote sensing data: results from the southern great plains hydrology experiment 1997. *Journal of Hydrometeorology*, **2**, 181–192.

Dubayah R., Wood E.F. and Lavallee D. (1997) Multiscaling analysis in distributed modeling and remote sensing: an application using soil moisture. In *Scale in Remote Sensing and GIS*, Quatrocchi D.A. and Goodchild M. (Eds.), Lewis Publishers: New York, pp. 93–112.

Dubois P.C., van Zyl J. and Engman E.T. (1995) Measuring soil moisture with imaging radars. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 915–926.

Entekhabi D., Njoku E., Houser P., Spencer M., Doiron T., Smith J., Girard R., Belair S., Crow W. and Jackson T., *et al.* (2004) The hydrosphere state (Hydros) mission: an earth system pathfinder for global mapping of soil moisture and land freeze/thaw (2004). *IEEE Transactions on Geoscience and Remote Sensing*, **42**, 2184–2195.

Hollenbeck K.J., Schmugge T.J., Hornberger G.M. and Wang J.R. (1996) Identifying soil hydraulic heterogeneity by detection of relative change in passive microwave remote sensing observations. *Water Resources Research*, **32**, 139–148.

Houser P.R., Shuttleworth W.J., Famiglietti J.S., Gupta H.V., Syed K. and Goodrich D.C. (1998) Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resources Research*, **34**, 3405–3420.

Jackson T.J. (1993) Measuring surface soil moisture using passive microwave remote sensing. *Hydrological Processes*, **7**, 139–152.

Jackson T.J. (1997) Soil moisture estimation using special satellite microwave/imager satellite data over a grassland region. *Water Resources Research*, **33**, 1475–1484.

Jackson T.J., Gasiewski A.J., Oldak A., Klein M., Njoku E.G., Yevgrafov A., Christiani S. and Bindlish R. (2002) Soil moisture retrieval using the C-band polarimetric scanning radiometer during the southern great plains 1999 experiment. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 2151–2161.

Jackson T.J. and Hsu A.Y. (2001) Soil moisture and TRMM microwave imager relationships in the southern great plains 1999 (SGP99) experiment. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1632–1642.

Jackson T.J., Le Vine D.M., Hsu A.Y., Oldak A., Starks P.J., Swift C.T., Isham J.D. and Haken M. (1999) Soil moisture mapping at regional scales using microwave radiometry: the southern great plains hydrology experiment. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 2136–2151.

Jackson T.J., Le Vine D.M., Swift C.T., Schmugge T.J. and Schiebe F.R. (1995) Large area mapping of soil moisture using the ESTAR passive microwave radiometer in Washita'92. *Remote Sensing of Environment*, **53**, 27–37.

Jackson T.J. and Schmugge T.J. (1991) Vegetation effects on the microwave emission from soils. *Remote Sensing of Environment*, **36**, 203–212.

Kim G. and Barros A.P. (2002) Space-time characterization of soil moisture from passive microwave remotely sensed imagery and ancillary data. *Remote Sensing of Environment*, **81**, 393–403.

Koike T., Njoku E., Jackson T.J. and Paloscia S. (2000) Soil moisture algorithm development and validation for the ADEOS-II/AMSR. *In Proceedings of the International Geoscience and Remote Sensing Symposium, IEEE Catalog No. 00CH37120*, **III**, 1253–1255.

Kummerow C., Barnes W., Kozu T., Shiue J. and Simpson J. (1998) The tropical rainfall measuring mission (TRMM) sensor package. *Journal of Atmospheric and Oceanic Technique*, **15**, 809–817.

Kustas W.P. and Jackson T.J. (1999) The impact on area-averaged heat fluxes using remotely sensed data at different resolutions: a case study using Washita '92 data. *Water Resources Research*, **35**, 1539–1550.

Kustas W.P., Jackson T.J., French A.N. and MacPherson J.I. (2001) Verification of patch- and regional-scale energy balance estimates derived from microwave and optical remote sensing during SGP97. *Journal of Hydrometeorology*, **2**, 254–273.

Lakshmi V., Wood E.F. and Choudhury B.J. (1997) Evaluation of SSM/I satellite data for regional soil moisture estimation

over the red river basin. *Journal of Applied Meteorology*, **36**, 1309–1328.

Leese J., Jackson T., Pitman A. and Dirmeyer P. (2001) GEWEX/BAHC international workshop on soil moisture monitoring, analysis and prediction for hydrometeorological and hydroclimatological applications. *Bulletin of the American Meteorological Society*, **82**, 1423–1430.

Le Hegarat-Mascle S., Zribi M., Alem F., Weisse A. and Loumagne C. (2002) Soil moisture estimation from ERS/SAR data: toward an operational methodology. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 2647–2658.

Lin H.Z., Islam S. and Zong C.Y. (1997) Statistical characterization of remotely sensed soil moisture images. *Remote Sensing of Environment*, **61**, 310–318.

Margulis S.A., McLaughlin D., Entekhabi D. and Dunne S. (2002) Land data assimilation and estimation of soil moisture using measurements from the southern great plains 1997 field experiment. *Water Resources Research*, **38**, 10.1029/2001WR001114.

Mattikalli N.M., Engman E.T., Ahuja L.R. and Jackson T.J. (1998b) Microwave remote sensing of soil moisture for estimation of profile soil properties. *International Journal of Remote Sensing*, **19**, 1751–1767.

Mattikalli N.M., Engman E.T., Jackson T.J. and Ahuja L.R. (1998a) Microwave remote sensing of temporal variations of brightness temperature and near-surface soil water content during a watershed-scale field experiment, and its application to the estimation of soil physical properties. *Water Resources Research*, **34**, 2289–2299.

Moran M.S., Hymer D.C., Qi J. and Kerr Y. (2002) Comparison of ERS-2 SAR and landsat TM imagery for monitoring agricultural crop and soil conditions. *Remote Sensing of Environment*, **79**, 243–252.

Njoku E.G., Jackson T.J., Lakshmi V., Chan T.K. and Nghiem S.V. (2003) Soil moisture retrieval from AMSR-E. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 250–261.

Njoku E., Koike T., Jackson T. and Paloscia S. (2000) Retrieval of soil moisture from AMSR data. In *Microwave Radiometry and Remote Sensing of the Earth's Surface and Atmosphere*, Pampaloni P. and Paloscia S. (Eds.), VSP Publications: The Netherlands, pp. 525–533.

Njoku E.G. and Li L. (1999) Retrieval of land surface parameters using passive microwave measurements at 6 to 18 GHz. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 79–93.

Owe M., Van de Griend A.A. and Chang A.T.C. (1992) Surface moisture and satellite microwave observations in semiarid Southern Africa. *Water Resources Research*, **28**, 829–839.

Reichle R.H., Entekhabi D. and McLaughlin B.B. (2001) Downscaling of radio brightness measurements for soil moisture estimation: a four-dimensional variational data assimilation. *Water Resources Research*, **37**, 2353–2364.

Rodruigez-Iturbe I., Vogel G.K., Rigon R., Entekhabi D., Castelli F. and Rinaldo A. (1995) On the spatial organization of soil moisture fields. *Geophysical Research Letters*, **22**, 2757–2760.

Satalino G., Mattia F., Davidson M.W.J., Le Toan T., Pasquariello G. and Borgeaud M. (2002) On current limits of soil moisture retrieval from ERS-SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 2438–2447.

Ulaby F.T., Moore R.K. and Fung A.K. (1986) *Microwave Remote Sensing: Active and Passive*, Vol. III, From Theory to Application, Artech House: Dedham.

Van de Griend A.A. and O'Neill P.E. (1986) Discrimination of soil hydraulic properties by combining thermal infrared and microwave remote sensing. In *Proceedings of the International Geoscience and Remote Sensing Symposium*, European Space Agency, ESA SD-254, pp. 839–845.

Verhoest N.E.C., Troch P.A., Paniconi C. and De Troch F.P. (1998) Mapping basin scale variable source areas from multitemporal remotely sensed observations of soil moisture behavior. *Water Resources Research*, **34**, 3235–3244.

Vinnikov K.Y., Robock A., Qiu S., Entin J.K., Owe M., Choudhury B.J., Hollinger S.E. and Njoku E.G. (1999) Satellite remote sensing of soil moisture in Illinois, USA. *Journal of Geophysical Research*, **104**, 4145–4168.

Wagner W., Lemoine G. and Rott H. (1999) A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sensing of Environment*, **70**, 191–207.

Wen J. and Su Z. (2003) Determination of land surface temperature and soil moisture from tropical rainfall measuring mission/ microwave imager remote sensing data. *Journal of Geophysical Research*, **108**, 4038 doi:10,1029/2002JD002176.

Wigneron J.P., Waldteufel P., Chanzy A., Calvet J.C. and Kerr Y. (2000) Two-dimensional microwave interferometer retrieval capabilities over land surfaces (SMOS mission). *Remote Sensing of Environment*, **73**, 270–282.

Additional Resources

Tower-based Experiment Data Sets

Between 1979 and 1985 a series of tower (truck-mounted) passive microwave remote sensing experiments were conducted at the Beltsville Agricultural Research Center. Data reports for each of the four years are available at **hydrolab.arsusda.gov/SMBARC/**.

Aircraft-based Experiment Data Sets

Data from the Washita'92 and Washita'94 experiments conducted in Oklahoma are available at **hydrolab.arsusda.gov/washita92/** and **hydrolab.arsusda.gov/washita94/**. The Washita'94 data set link includes access to multifrequency-multipolarization radar data collected in the Shuttle Imaging Radar experiments.

Another source of aircraft based SAR data is the Jet Propulsion Laboratory AIRSAR program **http://airsar.jpl.nasa.gov/index.html**.

Large-scale Experiment Data Sets

Data from recent large-scale field experiments include a wide range of ground, aircraft, and satellite data. The following links provide background and data sets:

Southern Great Plains 1997 Experiment (SGP97)
http://hydrolab.arsusda.gov/sgp97/

http://daac.gsfc.nasa.gov/CAMPAIGN_DOCS/SGP97/sgp97.html
Southern Great Plains 1997 Experiment (SGP99)
http://hydrolab.arsusda.gov/sgp99/
http://daac.gsfc.nasa.gov/CAMPAIGN_DOCS/SGP99/index.shtml
Soil Moisture Experiments 2002 (SMEX02)
http://hydrolab.arsusda.gov/smex02/
http://www.nsidc.org/data/amsr_validation/soil_moisture/smex02/

Passive Microwave Satellite Data Sources

SSM/I
http://nsidc.org/data/daac.html
SMMR
http://nsidc.org/data/daac.html
TMI
http://lake.nascom.nasa.gov/data/dataset/TRMM/
AMSR
http://nsidc.org/data/daac.html

Radar Satellite Data Sources

ERS and Envisat
http://earth.esa.int/services/pg/
http://envisat.esa.int/
Radarsat
http://www.rsi.ca/products/products.htm

55: Estimation of Snow Extent and Snow Properties

DOROTHY K HALL[1], RICHARD EJ KELLY[2], JAMES L FOSTER[1] AND ALFRED TC CHANG[1†]

[1]NASA/Goddard Space Flight Center, Greenbelt, MD, US
[2]University of Maryland, Baltimore, MD, US
[†]Deceased May 26, 2004

Important advances have been made in the measurement of seasonal snow cover since the advent of satellite remote sensing in the mid 1960s. Data from the visible, near-infrared, infrared, and microwave portions of the electromagnetic spectrum have proven useful for measuring different properties of snow. In terms of snow mapping, sensors employing visible and near-infrared wavelengths are now capable of accurately and reliably measuring snow-cover extent with a spatial resolution of up to 250 m on a daily basis, and even higher resolution for less-frequent coverage. Passive-microwave data, available since the 1970s, have been utilized for measuring snow extent, depth and snow-water equivalent (SWE), though at a coarse spatial resolution compared to visible data, while active-microwave sensors such as scatterometers, provide valuable information on snowpack ripening. Capabilities of synthetic-aperture radar (SAR) data for snow-cover studies are still being explored, however, bands on current satellite SAR sensors are not ideal for measuring snow cover. Remote sensing data of snow cover are now well suited for use in hydrologic and general-circulation models. Inclusion of remotely-sensed data significantly enhances our understanding of the Earth's weather and climate, and decadal-scale climate change. Future improvements include refinement of snow-cover extent measurements, minimizing SWE errors, and improving our ability to ingest remote sensing data of snow into models.

INTRODUCTION

Satellite remote-sensing technology has virtually revolutionized the study of snow cover. The high albedo of snow presents a good contrast with most other natural surfaces (except clouds), and therefore is easily detected by many satellite sensors. Weekly snow mapping of the Northern Hemisphere using National Oceanographic and Atmospheric Administration (NOAA) satellite data began in 1966 and continues today in the United States, but with improved resolution and on a daily basis (Matson *et al.*, 1986; Ramsay, 1998; Carroll *et al.*, 2001). In addition, using Earth Observing System (EOS) Moderate Resolution Imaging Spectroradiometer (MODIS) data, beginning in early 2000, global snow cover has been mapped on a daily basis at a spatial resolution of up to 500 m (Hall *et al.*, 2002a).

In addition to the visible/near-infrared data, from MODIS and NOAA satellite sensors, both passive and active microwave data have been useful for mapping snow and determining snow wetness and snow water equivalent (SWE) (Ulaby and Stiles, 1980) since the early 1970s. Using passive microwave data, snow extent, and SWE may be estimated globally on a daily basis since the launch of the Nimbus-7 Scanning Multichannel Microwave Radiometer (SMMR) (Chang *et al.*, 1987), and continuing with the May 2002 launch of the Advanced Microwave Scanning Radiometer (AMSR) (Kelly *et al.*, 2003). Active-microwave sensors, such as from the NASA scatterometer (NSCAT), are especially useful for detecting snowpack ripening (Nghiem and Tsai, 2001).

The geographical extent of snow cover over the Northern Hemisphere varies from a maximum of $\sim 46 \times 10^6$ km^2 in January and February to a minimum of $\sim 4 \times 10^6$ km^2 in

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

August; between 60 and 65% of winter snow cover is found over Eurasia, and most midsummer snow cover is in Greenland (Frei and Robinson, 1999). Numerous studies have shown the importance of accurate measurements of snow and ice parameters as they relate to the Earth's climate and climate change (for example, see Martinelli, 1979; Dewey and Heim, 1981 and 1983; Barry, 1983, 1984, and 1990; Dozier, 1987; Ledley *et al.*, 1999; Foster *et al.*, 1987 and 1996; Serreze *et al.*, 2000; Dozier and Painter, 2004). Measurements have become increasingly sophisticated over time. In addition, as the length of the satellite record increases, it becomes easier to determine trends that have climatic importance.

Three of the most important properties of a snow cover are depth, density, and water equivalent (Pomeroy and Gray, 1995). If the snow depth and density are known, then the SWE may be calculated. SWE is a hydrologically important parameter as it determines the amount of water that will be available as snowmelt.

After introducing snow as a medium, and reviewing its optical and microwave properties, we will show how remote sensing is used to study snow-covered area, SWE, snow wetness and snow albedo, and discuss the parameterization of snow in hydrologic and general circulation models (GCMs).

INTRODUCTION TO SNOW CRYSTALS AND DEVELOPMENT OF A SNOWPACK

Snow is a porous, permeable aggregate of ice grains (Bader, 1962). Snow crystals nucleate and begin their growth in clouds where the temperature is below 0 °C. Snow crystals can occur in a variety of relatively flat hexagonal or six-sided shapes (Figure 1), however, they can also occur as elongated columns and needles. Differences in snow crystals are a result of variations in the temperature and humidity of the atmosphere at the time of their formation and the action of the wind during their descent to the ground (Male, 1980). As a result, there is a myriad of possible shapes that can form depending primarily on the cloud temperature at the time the water vapor freezes.

Freshly fallen snow almost immediately begins to compact and metamorphose, initially preserving the original shape of the snow crystal. The constant jostling and rubbing of crystals against each other causes protuberances to become chipped and broken. As the snow settles under its own weight, melts and refreezes, and is buffeted by the wind, individual crystals are further altered such that after a few days they have little resemblance to their original shape. A seasonal snowpack might have grain radii ranging from 0.1 to 0.5 mm throughout the season.

In the absence of a temperature gradient in the snowpack, snowflakes undergo "destructive metamorphism" and become more rounded over time, with typically slower

Figure 1 This image of a newly fallen hexagonal plate snow crystal was taken with an Electron Scanning Microscope and shows a classic dendritic snow crystal, having a central hexagonal plate, but lacking sharp edges. Because of the air temperature at the time of collection (close to 0 °C), this crystal may have undergone some sublimation (Wergin *et al.,* 1996)

growth rates being more characteristic of the more-rounded crystals. "Constructive metamorphism," when large grains grow at the expense of small grains, occurs when there is a thermal and vapor gradient in the snowpack and snow grains at the base of the pack grow at the expense of smaller grains, hence the crystals develop distinctive shapes (Colbeck, 1982). Faster growth rates give rise to the faceted crystals. Dry snow will metamorphose into large depth-hoar grains when subjected to a strong temperature gradient, and grain growth and dry-snow metamorphism control the movement and redistribution of mass, chemical species, and isotopes in the snowpack (Sturm and Benson, 1997). Depth-hoar crystals may grow to 1 cm in size (Trabant and Benson, 1972).

Seasonal snowpacks develop from a series of winter storms and are often created by various forms of precipitation such as rain and freezing rain. Diurnal melting and refreezing, and wind action are also important in the ultimate development of a snowpack. Thus, seasonal snow cover usually develops a layered structure in which ice layers alternate with layers of a coarser texture (Male, 1980).

FUNDAMENTALS OF THE REMOTE SENSING OF SNOW

Introduction With optical remote sensing, there is a potential to determine the extent and albedo of a snow cover, and some inference as to snow depth can be made based on the ability of a snow cover to cover existing vegetation of known height. Infrared sensors can provide snow

surface temperature, which may be a useful parameter for hydrologic modeling. Microwave measurements have the capability to respond to the bulk properties of a snowpack as well as to variations in other surface and subsurface features because microwaves can penetrate the snowpack and thus provide information on snow depth and SWE when the snowpack is dry. Additionally, the microwave part of the spectrum allows remote observation of snow cover under nearly all weather and lighting conditions. SWE is of critical importance for water resources and hydrologic and GCM modeling. In short, use of optical, infrared, and microwave sensors provides synergy that allows extraction of important snowpack properties for use in models.

Optical Properties The spectral albedo of a surface is the upflux divided by the downflux at a particular wavelength (Warren, 1982). The spectral albedo of fresh snow in the visible region of the spectrum remains high but decreases slowly as snow ages, but in the near-infrared, the spectral albedo of aging snow decreases considerably as compared to fresh snow (O'Brien and Munis, 1975; Warren and Wiscombe, 1980; Wiscombe and Warren, 1980).

The broadband albedo is the reflectance across the reflective part of the solar spectrum. Broadband albedo decreases when grain size increases as the snow ages (Choudhury and Chang, 1979), and melting causes snow grains to grow and bond into clusters (Dozier *et al.*, 1981; Grenfell *et al.*, 1981; Warren, 1982). Snow albedo may decrease by $>25\%$ within just a few days as grain growth proceeds (Nolin and Liang, 2000). For example, Gerland *et al.* (1999) measured a maximum albedo $>90\%$ on Svalbard, Norway, before melt onset, and $\sim 60\%$ after melt had progressed in the spring when the snow was considered old, but still clean. The albedo of a snow cover is also influenced by the albedo of the land cover that it overlies, especially when the snowpack is thin.

Grain size may be estimated using remotely sensed data (Dozier, 1984; Nolin and Dozier, 1993). With the onset of surface melting and associated grain size increase, the near-infrared reflectance decreases dramatically (Warren, 1982) (Figure 2). The near-infrared albedo of snow is very sensitive to snow-grain size while visible albedo is less sensitive to grain size, but is affected by snow impurities. Modeling by Warren and Wiscombe (1980) demonstrates that small but highly absorbing particles can lower the snow albedo in the visible part of the spectrum by 5–15% compared to pure snow. Hansen and Nazarenko (2004) report that anthropogenic soot emissions have reduced snow and ice albedos by 3% in Northern Hemisphere land areas to yield a climate forcing of $+0.3\,\mathrm{W\,m^{-2}}$ in the Northern Hemisphere, thus contributing to global warming.

Though the reflectance of freshly fallen snow is nearly isotropic (Dirmhirn and Eaton, 1975), as snow ages, the specular reflection component increases, especially in the forward direction and with solar zenith angle (SZA) (Salomonson and Marlatt, 1968), and the anisotropic nature of snow reflectance increases with increasing grain size

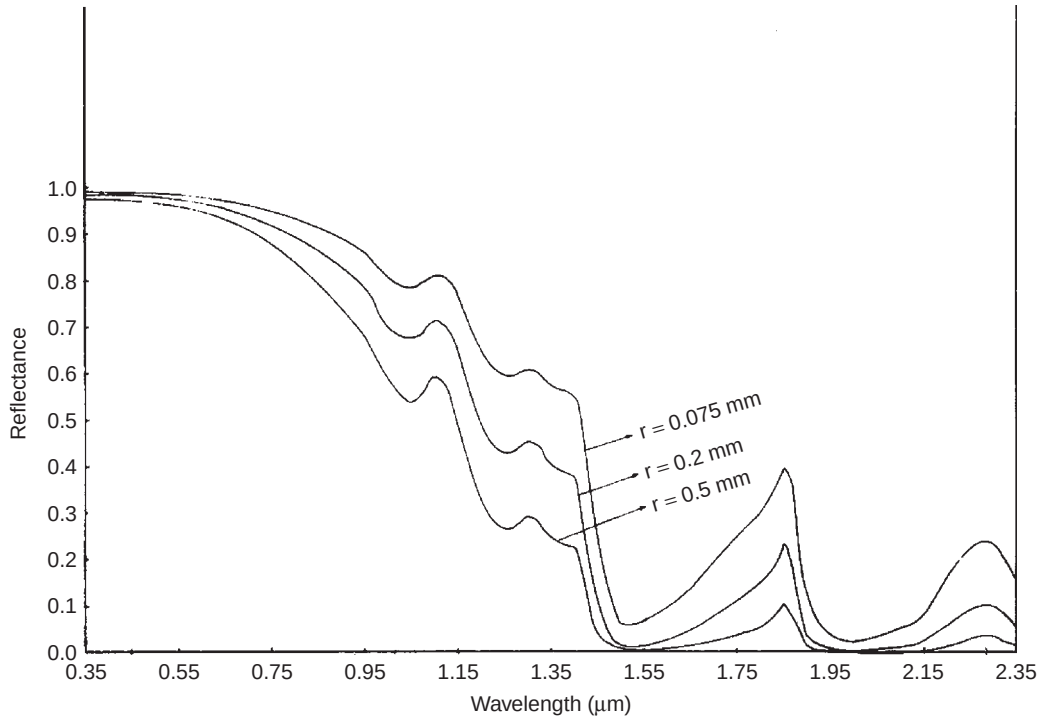


Figure 2 Illustration of the effect of different snow crystals on snow reflectance (From Choudhury and Chang, 1979)

(Steffen, 1987). Effective snow-grain radii typically range in size from ~50 μm for new snow, to 1 mm for wet snow consisting of clusters of ice grains (Warren, 1982).

Snow albedo increases at all wavelengths with the SZA. Additionally, cloud cover normally causes an increase in spectrally integrated snow albedo due to multiple reflections caused by clouds (Grenfell and Maykut, 1977; Warren, 1982).

Microwave Properties In the microwave part of the spectrum (300 to 1 GHz, or 1 mm to 30 cm wavelength), remote sensing can be accomplished either by measuring emitted radiation with a radiometer or by measuring the intensity of the return (in decibels) of a signal sent by a radar.

Microwave emission from a layer of snow over a ground surface consists of contributions from the snow itself and from the underlying ground. Both contributions are governed by the transmission and reflection properties of the air-snow and snow-ground interfaces, and by the absorption/emission and scattering properties of the snow layer.

The dielectric properties of snow at a given microwave frequency are generally dependent on the relative proportion of liquid and solid water in the snow by volume. Even at temperatures <0 °C, liquid-like water exists in thin films surrounding, and bound to, ice crystals (Hobbs, 1974), but is considered to be dry since it contains no "free" liquid water (Leconte *et al.*, 1990). However, snow that contains a large amount of liquid water (>5% by volume) has a high dielectric constant (>35 below 20 GHz) relative to that of dry snow.

Theoretically, the dielectric constant of snow consists of the sum of a real and an imaginary part. Snow is a mixture of air and ice, the dielectric constant of air being 1.0 and ice 3.17 ± 0.07 for frequencies from 1 MHz to well above the microwave region (Evans, 1965). Snow has a dielectric constant between l.2 and 2.0 when the snow densities range from 0.1 to 0.5 g cm^{-3} (Hallikainen and Ulaby, 1986).

If a dry snowpack contains ice and snow layers, specular reflection at the interfaces between layers may occur resulting in strongly enhanced backscatter in the case of active-microwave remote sensing (Mätzler and Schanda, 1984). Or, if the grain sizes of a dry snowpack are large enough relative to the microwave wavelength, volume scattering will occur. Otherwise, the signal is returned mainly from the ground/snow interface.

Longer wavelengths, in general, travel almost unaffected through dry snow. X-band (2.4–3.75 cm, 8.0–12.5 GHz) or lower frequencies (longer wavelengths) are not generally useful for detecting and mapping thin, dry snow because the size of snow particles is much smaller than the size of the wavelength. Thus, at these longer wavelengths, there is little chance for a microwave signal to be attenuated and scattered by the relatively small ice crystals comprising a snowpack (Waite and MacDonald, 1970; Ulaby and Stiles, 1980, 1981). Wavelengths longer than ~10–15 cm are not impeded as they move through most dry seasonal snowpacks (Bernier, 1987).

For snow crystals of a radius >~0.1 mm, scattering dominates emission at higher (>15 GHz) microwave frequencies (Ulaby *et al.*, 1986). Absorption is determined primarily by the imaginary part of the refractive index. In dry snow, the imaginary part is very small, several orders of magnitude smaller than for water (Ulaby and Stiles, 1980).

The backscatter received by a synthetic-aperture radar (SAR) antenna is the sum of surface scattering at the air/snow interface, volume scattering within the snowpack, scattering at the snow/soil interface, and volumetric scattering from the underlying surface (if applicable). Most techniques developed for mapping snow cover using SAR data show promise for mapping only wet snow (see, for example, Rott and Nagler, 1993; Shi *et al.*, 1994). This is because it is difficult to distinguish dry snow from bare ground using SAR data at the X-band and lower frequencies that are currently flown in space. Volume scattering from a shallow, dry snow cover (SWE <20 cm) is undetectable at C-band (5.3 GHz, 5.6 cm), for example, because the backscatter is dominated by soil/snow scattering. Volume-scattering in dry snow results from scattering at dielectric discontinuities created by the differences in electrical properties of ice crystals and air, and by ice lenses and layers. Atmospheric scattering is usually very small and can be neglected (Ulaby and Stiles, 1980; Leconte *et al.*, 1990; Leconte, 1995).

In the case of wet snow (Stiles and Ulaby, 1980; Ulaby and Stiles, 1980; Rott, 1984; Ulaby *et al.*, 1986), when at least one layer of the snowpack (within the penetration depth of the radar signal) becomes wet (4–5% liquid water content), the penetration depth of the radar signal is reduced to about 3–4 cm (or one wavelength at X-band) (Mätzler and Schanda, 1984). Thus, there may be high contrast between snow-free ground and ground covered with wet snow, thus making it possible to distinguish wet and dry land or snow when imaged with C-band SAR from space.

Volume scattering increases with snow-grain size and internal layering, and with amount of snow. Radiation at wavelengths comparable in size to the snow crystal size (about 0.05–3.0 mm, or greater if depth hoar is present) is scattered in a dry snowpack according to Mie scattering theory. (Mie scattering predominates when the particles causing the scattering are larger than the wavelengths of radiation in contact with them.) Currently, only passive microwave sensors operate at these wavelengths from satellites.

In the microwave part of the spectrum, the radiation emitted from a perfect emitter is proportional to its physical temperature, T. However, most real objects emit only a fraction of the radiation that a perfect emitter would emit at its physical temperature. The equivalent temperature of

the microwave radiation thermally emitted by an object is called its *brightness temperature*, T_B, expressed in Kelvins. This fraction defines the emissivity, E, of an object (Chang *et al.*, 1976). In the microwave region,

$$E = \frac{T_B}{T} \tag{1}$$

Microwave emission from a layer of snow over a ground medium consists of two contributions: (i) emission by the snow volume and (ii) emission by the underlying ground. Both contributions are governed by the transmission and reflection properties of the air-snow and snow-ground interfaces, and by the absorption/emission and scattering properties of the snow layer (Stiles *et al.*, 1981), and a myriad of physical parameters that affects the emission (Derksen *et al.*, 2002). As an electromagnetic wave emitted from the underlying surface propagates through a snowpack, it is scattered by the randomly spaced snow particles in all directions. As the snowpack grows deeper, there is more loss of radiation due to scattering, and the emission of the snowpack is reduced, thus lowering the T_B. The deeper the snow, the more crystals are available to scatter the upwelling microwave energy, and, thus, it is possible to estimate the depth and water equivalent of the snow using passive-microwave remote sensing.

Snow grains scatter the electromagnetic radiation incoherently and are assumed to be spherical and randomly spaced within the snowpack. Although most snow particles are generally not spherical in shape, using Mie theory, their optical properties can be simulated as spheres (Chang *et al.*, 1976). We will discuss the scattering power of a variety of different snow crystal shapes and their effect on the microwave emission of a snowpack in a later section.

A wet snowpack radiates like a blackbody at the physical temperature of the snow layer, and is therefore indistinguishable from snow-free soil using microwave remote sensing (Kunzi *et al.*, 1982). The dielectric constants of water, ice, and snow are different enough so that even a little surface melting causes a strong microwave response (Schanda *et al.*, 1983; Foster *et al.*, 1987). The scattering loss decreases drastically with increasing liquid water content (free water) and becomes negligible for values above about 1% (Hallikainen, 1984).

SNOW MAPPING

Early attempts to forecast runoff from the areal extent of snow cover used terrestrial photographs (Potts, 1937). Other observations from aircraft were also undertaken such as locating and mapping snow-cover extent and mapping the location of the snowline. Along with volume, the areal extent of snow cover has been used to predict snowmelt runoff and to forecast floods.

Because of its high albedo, snow was easily observed in the first image obtained from the Television Infrared Operational Satellite-1 (TIROS-1) weather satellite in 1960. Data from meteorological satellites and manned spacecraft were useful in determining snowline elevation, delineating snow boundaries, and observing changes in snow conditions due to rising temperatures and rain-on-snow events (Singer and Popham, 1963), and later, seasonal streamflow could also be estimated (Rango and Salomonson, 1977). Data from Environmental Science Services Administration (ESSA) operational satellites were used as early as the mid-to-late 1960s to determine areal extent of snow cover.

A major step forward in snow mapping came with the advent of the Landsat series of sensors beginning in 1972. Landsat-1 carried a Multispectral Scanner (MSS) sensor with 80-m spatial resolution. With Landsat data came the ability to create detailed basin-scale snow-cover maps on a regular basis when cloudcover permitted. At first, the repeat-pass interval for the Landsat satellite was 18 days, but this was decreased to 16 days with the launch of Landsat-4 in 1982. Landsats-4 and -5 carried a Thematic Mapper (TM) sensor with 30-m resolution, and Landsat-7 carries an Enhanced Thematic Mapper Plus (ETM+) with spatial resolution of 30 m, except in the panchromatic band where the resolution is 15 m. As of this writing, the TM onboard the Landsat-5 satellite is still operating. (Landsat-6 was lost after it failed to reach orbit in 1993.) Though the Landsat series has provided high-quality, scene-based snow maps, the 16- or 18-day repeat-pass interval of the Landsat satellites is not adequate for most snow-mapping requirements, especially during spring snowmelt.

Operational Snow-cover Mapping in the United States The NOAA National Environmental Satellite, Data, and Information Service (NESDIS) began to generate Northern Hemisphere Weekly Snow and Ice Cover analysis charts derived from NOAA's Geostationary Operational Environmental Satellite (GOES) and Polar Orbiting Environmental Satellite (POES) visible satellite imagers in November 1966. Maps were manually constructed, and the spatial resolution of the charts was 190 km. Since 1997, the Interactive Multi-sensor Snow and Ice Mapping System (IMS) has been used by analysts to produce products daily at a spatial resolution of about 25 km, and utilizes a variety of satellite data to generate the maps (Ramsay, 1998). This snow-cover record has been studied carefully (Robinson *et al.*, 1993; Robinson, 1997, 1999) and has been reconstructed following adjustment for inconsistencies that were discovered in the early part of the data set (Robinson and Frei, 2000; Frei *et al.*, 1999). Results show that the Northern Hemisphere annual snow-covered area has decreased (Robinson *et al.*, 1993; Brown and Goodison, 1996; Hughes and Robinson, 1996; Hughes *et al.*, 1996; Armstrong and Brodzik, 1998, 2001; Frei *et al.*, 1999;

Brown, 2000); satellite data show a decrease of about 0.2% per year from 1979–1999 (Armstrong and Brodzik, 2001).

The National Operational Hydrologic Remote Sensing Center (NOHRSC) snow-cover maps, generated by National Weather Service NOHRSC hydrologists, are distributed electronically in near real time to local, state, and federal users during the snow season (Carroll, 1987 and 1995; Cline *et al.*, 1998; Carroll *et al.*, 2001). The NOHRSC 1-km maps are generated primarily from the NOAA polar-orbiting satellites and GOES satellites to develop daily digital maps depicting the areal extent of snow cover for the coterminous United States and Alaska, and parts of southern Canada.

Other Snow Maps Landsat data have been used for measurement of snow-covered area over drainage basins (Rango and Martinec, 1979, 1982). The Landsat TM and ETM+ have been especially useful for measuring snow cover because of the short-wave infrared band – band 6(1.6 μm), which allows snow/cloud discrimination. The reflectance of snow is low and the reflectance of most clouds remains high in that part of the spectrum. Various techniques, ranging from visual interpretation, multispectral image classification, decision trees, change detection, and ratios (Kyle *et al.*, 1978; Bunting and d'Entremont, 1982; Crane and Anderson, 1984; Dozier, 1989; Romanov *et al.*, 2000; Hall *et al.*, 2002a; Romanov and Tarpley, 2003) have been used to map snow cover. Other spectral and threshold tests are also used.

The MODIS was first launched in December 1999 on the Terra spacecraft. MODIS data are now being used to produce daily and eight-day composite (Figure 3) snow-cover products from automated algorithms **`http://modis-snow-ice.gsfc.nasa.gov`** at Goddard Space Flight Center in Greenbelt, Maryland (Hall *et al.*, 2002a). The products are transferred to the National Snow and Ice Data Center (NSIDC) in Boulder, Colorado, where they are archived and distributed via the Earth Observing System Data Gateway (EDG) **`http://nsidc.org`** (Scharfen *et al.*, 2000). The Aqua satellite was launched in 2002 with a second MODIS instrument that enables snow-covered area measurements to be extended farther into the future (Riggs and Hall, 2004).

The MODIS maps provide global, daily coverage at 500-m resolution, and the climate-modeling grid (CMG) maps are available at 0.05 °resolution, which is ~5.6 km at the Equator. The CMG map, designed for climate modelers, provides a global view of the Earth's snow cover in a geographical projection with fractional snow cover reported in each cell. The automated MODIS snow mapping algorithm uses at-satellite reflectances in MODIS bands 4 (0.545–0.565 μm) and 6 (1.628–1.652 μm) to calculate the normalized difference snow index (NDSI) (Hall *et al.*, 2002a). Other threshold tests are also used, including the Normalized Difference Vegetation Index (NDVI) together with the NDSI to improve snow mapping in forests (Klein *et al.*, 1998). Mauer *et al.* (2003) compared MODIS and NOHRSC data for the Columbia River Basin, USA, and found that the maps were comparable, but the MODIS maps generally provided more cloud-free data than did the NOHRSC maps. Other studies have also shown the MODIS snow maps to represent an advance in global snow mapping.

Fractional Snow Cover or Subpixel Snow Mapping Fractional snow cover (FSC) utilizing Landsat and MODIS data has been derived to exploit subpixel information. Much of this work has relied on spectral-mixture modeling (see Nolin *et al.*, 1993; Rosenthal and Dozier, 1996; Vikhamer and Solberg, 2002; Dozier and Painter, 2004), and neural networks (Simpson and McIntire, 2001) but does not provide global coverage. Painter *et al.* (2003) couple

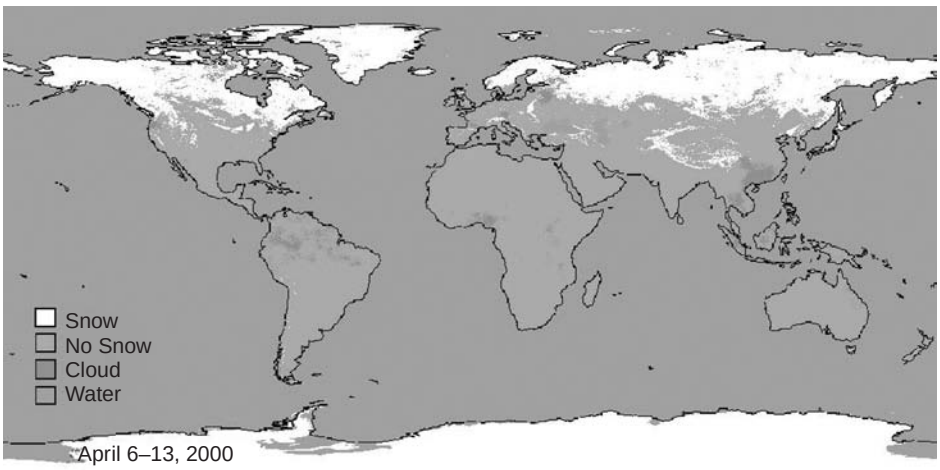


Figure 3 MODIS 8-day composite CMG global snow-cover map. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

spectral-mixture analysis with a radiative transfer model to retrieve subpixel snow-covered area and effective grain size from Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data.

Recently, Salomonson and Appel (2004) have extended the use of the NDSI to provide FSC globally with absolute errors of 0.1 or less over the whole range of FSC from 0 to 100%. In the near future, percent snow cover or fractional snow cover in each pixel (Salomonson and Appel, 2004) will be provided in the 500-m products. Other work using MODIS data (Kaufman *et al.*, 2002) has also developed algorithms to map FSC globally.

The Norwegian Water Resources and Energy Directorate (NVE) and the Tromso Satellite Station (TSS) produce snow maps from Advanced Very High Resolution Radiometer (AVHRR) data using band 2 (0.7–1.1 μm) for snow mapping, and bands 3 (3.6–3.9 μm), 4 (10.3–11.3 μm), and 5 (11.5–12.5 μm) for snow/cloud discrimination (Anderson, 1982; König *et al.*, 2001). An upper limit of 100% snow cover is determined from a glacier or 100% snow-covered region and a lower limit of 0% is determined from water or land areas, and percentage of snow cover is interpolated linearly, thus deriving FSC (König *et al.*, 2001).

Snow Mapping Using Microwave Sensors For dry snow, the naturally emitted microwave radiation of a snowpack is related to several physical properties of the snow as discussed earlier. These properties include the number of snow grains along the emission path (the snow depth in cm), the size of grains (grain radius in mm), and the packing of the grains (volume fraction in % or density in $\mathrm{kg\,m^{-3}}$). Such components control the propagation of radiation, especially at higher frequencies (e.g. 36 GHz) but affect the microwave response less at lower frequencies (e.g. 18 GHz). The brightness temperature difference between 18 GHz and 36 GHz ($T_B18 - T_B36$) is used to minimize the effect of snow temperature on the microwave emissivity. This is the principle that has been used to estimate SWE and snow depth from passive microwave instruments (Chang *et al.*, 1976, 1982; Kunzi *et al.*, 1976, 1982; Goodison and Walker, 1994; Goodison *et al.*, 1986; Grody and Basist, 1996; Foster *et al.*, 1997; Kelly and Chang, 2003). Experiments and applications have shown that:

$$SD = a(T_B18\mathrm{H} - T_B36\mathrm{H}) \qquad (2)$$

where SD is the snow depth in cm, a is constant of 1.59 for SMMR and is derived from radiative transfer experiments (Chang *et al.*, 1982), and $T_B18\mathrm{H}$ and $T_B36\mathrm{H}$ are the horizontally polarized brightness temperatures measured by the spacecraft at 18 and 36 GHz, respectively. If SWE is desired, a is set to a value of 4.8. Research into estimation of SWE and snow depth from passive microwave instruments has used this principle and an example of its application to AMSR-E data is shown in Figure 4 (Kelly *et al.*, 2003).

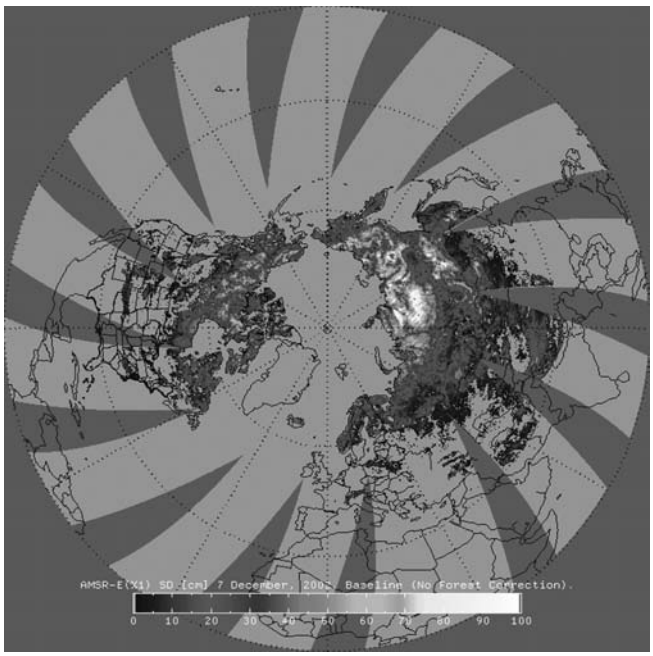


Figure 4 Estimated snow depth for the Northern Hemisphere from AMSR-E data for 7 December 2002 (From R.E.J. Kelly and A.T.C. Chang, unpublished data). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Rott and Nagler (1995) developed a threshold-based algorithm to map the extent of melting snow areas in mountainous regions on glaciers using ERS-1 SAR data. The classification is based on the ratio of the backscatter of the image with wet snow cover *versus* the backscatter of the reference image. They determined a threshold value of −3 dB by comparison with field observations. Using a single-layer backscatter model, seasonal variations as well as day/night changes in snow-covered alpine areas are shown by Nagler and Rott (2000) to be largely due to changes of the liquid water content of the snowpack and the snow surface roughness. They showed the same threshold of −3 dB for identifying wet snow using C-band HH Radarsat SAR and C-band ERS SAR data. Comparison of SAR-derived snow maps with Landsat-derived snow maps showed generally good results, but with a systematic underestimation of the snow extent at the edges of the snowpack using the SAR data.

In alpine regions, Shi *et al.* (1994) used single-pass SAR imagery with polarization to map wet snow cover, finding it difficult to distinguish dry snow cover from bare ground and short vegetation. Further work using single-pass, multifrequency (SIR-C/X-SAR) data (Shi and Dozier, 1997) showed that the coherence between two passes provides a useful measurement and allows development of

algorithms to map both wet and dry snow under specific circumstances.

MODIS snow maps were compared with SSM/I-derived snow maps over the prairie and boreal forest region in western Canada by Bussières *et al.* (2002); generally good correspondence was found in the taiga region in eastern Canada, however, the accuracy of the SSM/I maps was reduced in the fall and spring. As the snow deepens during the winter, snow-grain sizes increase and the temperatures become consistently colder, the ability of the SSM/I to map snow improves, and the agreement between the visible and passive microwave maps improves (Basist *et al.*, 1996). This is because the passive microwave data are not able to effectively map wet snow cover since the penetration of the microwave signal upward through the snowpack is extremely low when the snow is wet.

Snow Albedo

Early attempts to measure snow albedo remotely were conducted from aircraft (Bauer and Dutton, 1962; Hanson and Viebrock, 1964; McFadden and Ragotzkie, 1967; Salomonson and Marlatt, 1968; Dirmhirn and Eaton, 1975). More recently, however, detailed field, aircraft, and satellite studies have been undertaken to derive quantitative measurements of snow reflectance and albedo (for example, see Steffen, 1987; Hall *et al.*, 1989; Duguay and LeDrew, 1992; Winther, 1993; Knap and Oerlemans, 1996; Stroeve *et al.*, 1997; Winther *et al.*, 1999; Greuell *et al.*, 2002).

Some researchers have measured the albedo of snow-covered lands using satellite data on a hemispheric scale (e.g. Kukla and Robinson, 1980; Robock, 1980; Robinson and Kukla, 1985; Robinson *et al.*, 1992; Robinson, 1993). Both Robinson *et al.* (1986) and Scharfen *et al.* (1987) constructed basinwide albedo maps and observed differences in the timing of the melt between years. Robinson and Kukla (1985) used Defense Meteorological Satellite Program (DMSP) imagery (spectral range – 0.4–1.1 μm) to derive a linear relationship between the brightest snow-covered arctic tundra and the darkest snow-covered forest, which were assigned albedos of 0.80 and 0.18, respectively. Scene brightness was then converted to surface albedo by linear interpolation. The surface brightness is a function of the type and density of vegetation and the depth and age of snow (Robinson and Kukla, 1985). The derived "maximum" surface albedo values were useful for climate modeling (Ross and Walsh, 1987).

Surface albedo has also been derived from Landsat MSS and TM data. One approach, based on exact solutions of the radiative transfer equation for upwelling intensity, requires known albedo values derived in each Landsat scene at different points (Mekler and Joseph, 1983). Other approaches rely on a narrowband to broadband conversion to derive albedo (Brest and Goward, 1987; Hall *et al.*, 1989; Duguay and LeDrew, 1992; Knap *et al.*, 1998; Winther *et al.*, 1999). Knap and Reijmer (1998) used Landsat data to derive the Bidirectional Reflectance Distribution Function (BRDF) to describe the complete distribution of the anisotropic reflectance of snow, and Greuell and de Ruyter de Wildt (1999) used BRDF to correct for anisotropic reflectance. The BRDF is the physical property that determines the amount and angular distribution of reflected radiance from a surface (Nicodemus *et al.*, 1977).

AVHRR data have been used to map changes in albedo over the Greenland Ice Sheet during the spring and summer months (Knap and Oerlemans, 1996; Nolin and Stroeve, 1997; Stroeve *et al.*, 1997). While Stroeve *et al.* (1997) found a good correspondence with satellite-derived and surface-measured albedo before snowmelt, after snowmelt began, melt-water ponding on the ice surface precluded accurate comparisons between the satellite-derived and surface-measured albedo.

A near-surface global algorithm has been developed to map snow albedo using MODIS data (Klein and Stroeve, 2002). In deriving albedo, atmospherically corrected MODIS surface reflectances in individual MODIS bands for snow-covered pixels located in nonforested areas are adjusted for anisotropic scattering effects using a discrete ordinates radiative transfer (DISORT) model and snow optical properties. Currently, in the algorithm, snow-covered forests are considered to be Lambertian reflectors. The adjusted spectral albedos are then combined into a broadband albedo measurement using a narrow-to-broadband conversion scheme developed specifically for snow by Shunlin Liang (written communication, 2003) (Liang, 2000; Klein and Stroeve, 2002). A near-global snow albedo product (Figure 5) is available from February 2000 to the present and validation of this product is ongoing.

Snow-water Equivalent

To derive the SWE using passive microwave data, a radiative transfer approach is used in which, for example,

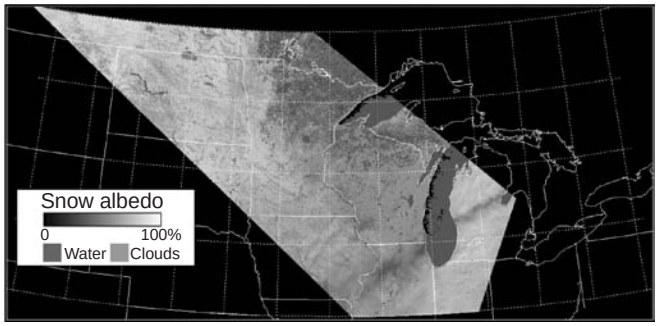


Figure 5 MODIS snow albedo product – north central United States and southern Canada – February 16, 2001. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

an average crystal size of 0.3 mm (radius), a density of 300 kg m^{-1}, and a spherical shape are assumed. It is also assumed that the crystals scatter radiation incoherently and independently of the path length between scattering centers. These quantities are then used in radiative transfer equations to solve the energy transfer through the snowpack (Chang *et al.*, 1976, 1987). Equations 1 and 2 are derived from this work and reasonable results are obtained from its implementation. However, if the crystal radii and snow density differ significantly from the averages and assumptions, then poor SWE values may result. Current efforts are aimed at improving the methods to estimate SWE by incorporating more dynamic parameterizations of these variables.

When viewed with an electron microscope, the detail is so great that individual crystals (Figure 1) can be assigned a specific shape, but the variation between even adjacent crystals can be substantial (Rango *et al.*, 1996; Wergin *et al.*, 2002). While crystal size and effective crystal size (Mätzler, 1997) are strongly related to microwave brightness temperature, it appears from modeling results that the shape of the snow crystal is of little consequence in accounting for the transfer of microwave radiation (at least at 0.81 cm) from the ground through the snowpack (Foster *et al.*, 1999, 2000; Tsang *et al.*, 2000).

Currently, SWE of a dry snowpack can be estimated with passive microwave sensors such as the SSM/I, and the AMSR (see Table 1), which was launched on the Aqua satellite in May of 2002. In Canada, SSM/I data are used to provide operational SWE map products (Figure 6) for the Canadian prairie region.

Forest cover can adversely affect the SWE retrieval accuracy by reducing the characteristic scattering response from snow by suppressing the $T_B18\mathrm{H} - T_B36\mathrm{H}$ signal (Tiuri and Hallikainen, 1981; Hall *et al.*, 1982; Hallikainen, 1984; Kurvonen and Hallikainen, 1997). Foster *et al.* (1997) attempted to correct for this effect by incorporating forest fraction into equation (1) such that:

$$SD = \frac{a(T_B18\mathrm{H} - T_B36\mathrm{H})}{(1-ff)} \qquad (3)$$

where *ff* is the per pixel forest fraction (expressed as a unit percent) that ranges from 0% to 50% (fractions greater than 50% are set to 50%). This approach improved the retrieval accuracy in forested regions.

Snow-grain size is another important parameter that influences the microwave brightness temperature. A model was developed to study the growth of the depth-hoar layer at the base of the snowpack on the Arctic Coastal Plain of Alaska during winter and compared to brightness temperature as derived from the SMMR by Hall *et al.* (1986). Results showed that an approximately 20 K lower T_B was found at inland sites with a comparable snow depth, but a thicker depth-hoar layer than was present at coastal sites. Thus, it is necessary to characterize snow-grain size on a regional basis to enhance the accuracy of snow retrievals using passive microwave data. Using SSM/I data, Mognard and Josberger (2002) modeled seasonal changes in snow grain size using a temperature gradient approach. This information was then used to parameterize the retrieval of snow depth in the northern Great Plains during the 1996–1997 winter season. Taking this approach further, Kelly *et al.* (2003) have recently developed a methodology to estimate snow-grain size and density as the snowpack evolves through the season using SSM/I and simple statistical growth models. These estimated variables are then coupled with a dense media radiative transfer (DMRT) model, described in Tsang and Kong (2001) and Chang *et al.* (2003), to estimate SWE.

As early as 1972, Meier (1972) was able to map the snowline on Mt. Rainier in Washington State in the US, using the 270 K-T_B line and a single channel (19.35 GHz) of an airborne radiometer. Since then, many different algorithms to map snow cover and SWE using passive microwave data have been developed and tested (e.g. Rott *et al.*, 1981; Kunzi *et al.*, 1982; Chang *et al.*, 1982; Goodison, 1989; Goodison and Walker, 1994; Mätzler, 1987; Hallikainen and Jolma, 1992; Grody and Basist, 1996 and Derksen *et al.*, 2002; Pivot *et al.*, 2002; Walker *et al.*, 2002). Some appear to work better under certain conditions than others, and it is now well accepted that there is no algorithm that is ideal globally.

Table 1 Comparison of characteristics of passive microwave sensors (from Foster *et al.*, 2005)

	SMMR	SSM/I	AMSR-E
Platform	Nimbus-7	DMSP F-8, 11, 13	Aqua
Period of Operation	1979–87	1987 to present	2002 to present
Data Acquisition	every other day	daily	daily
Swath Width	780 km	1400 km	1600 km
Frequency (GHz)	18.0 37.0	19.35 37.0	18.7 36.5
Spatial Resolution (km)	60 × 40 (18 GHz) 30 × 20 (37 GHz)	69 × 43 (19.4 GHz) 37 × 29 (37 GHz)	28 × 16 (19.7 GHz) 14 × 8 (36.5 GHz)
Polarization	H & V	H & V	H & V
Orbital Timing (equation Crossing) (for minimum temperature)	midnight	6:00 a.m.	1:30 a.m.

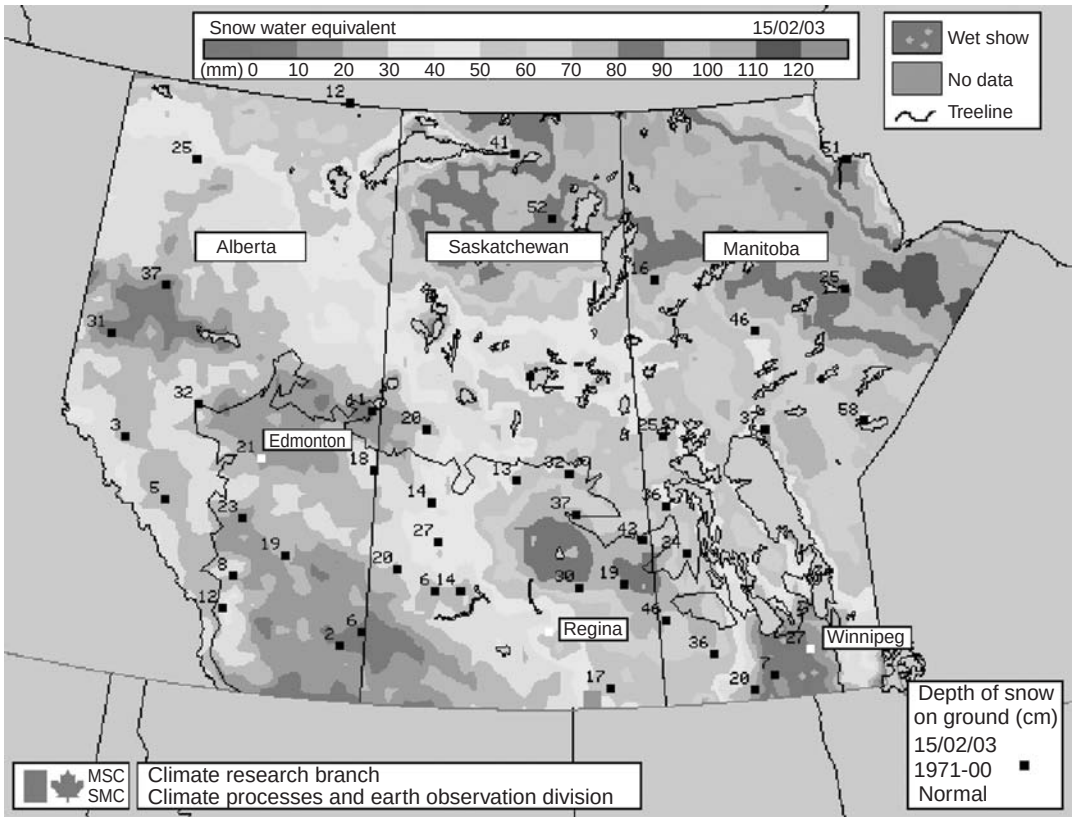


Figure 6 Snow water equivalent (in mm) over the Canadian Prairie region, derived from DMSP SSM/I data for 15 February, 2003. Areas of highest snow water equivalent generally correspond to the areas where the snow cover is deepest (Courtesy of the Climate Research Branch, Meteorological Service of Canada, Environment Canada). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The basic algorithm must be tuned to individual land-cover types (Walker and Goodison, 2000). Walker and Goodison (1993) developed a wet snow indicator using the SSM/I 37 GHz polarization difference for the Canadian Prairies, and Goita *et al.* (2003) developed separate algorithms, both based on the vertically polarized difference index using 18 and 37 GHz data from SSM/I to map SWE in deciduous and coniferous forest types, respectively.

The differences between the day and night brightness temperatures indicate the presence of liquid water in the snowpack. Early in the snow season, the difference is small, indicating the absence of liquid water in the snowpack. As spring approaches, the difference increases, indicating the presence of liquid water during the day, and then the pack refreezes at night. When liquid water does not refreeze at night, the difference again becomes small, and the snowpack is ripe and will soon begin to melt (Josberger *et al.*, 1993). Ramage and Isacks (2002) used SSM/I-derived diurnal-amplitude variations to detect early melt on snow-covered icefields in southeast Alaska, and found that melt timing correlates well with nearby stream hydrographs.

Climate-data-record (CDR) quality data sets of SWE are difficult to derive. The development of long-term CDR quality datasets may be influenced by difficulties related to calibration between sensors. Derksen *et al.* (2003) produced a time series for central North America for the winter season from 1978 through 1999, using both SMMR and SSM/I data. They found evidence of systematic SWE underestimation during the SMMR years, and overestimation of SWE during the SSM/I years. Previous work by Armstrong and Brodzik (2001) also suggested inconsistencies between the SMMR and SSM/I datasets.

Many other researchers have developed algorithms that use multiple sensors to map snow cover and SWE (e.g. Basist *et al.*, 1996; Armstrong and Brodzik, 1998, 2001, 2002; Standley and Barrett, 1999; Tait *et al.*, 2000 and, 2001; Kelly, 2001; Hall *et al.*, 2002b), and in different land covers (Hall *et al.*, 2001). Visible/near-infrared data, with good spatial resolution, provide excellent snow-covered area determination under cloud-free conditions, while passive microwave data provide all-weather day/night snow maps, but with coarser resolution. In addition, problems arise in wet snow and thin, dry snow when using passive microwave data alone.

There are conflicting results showing the sensitivity of radars to SWE. Ulaby *et al.* (1977) found that radar

sensitivity to total SWE (for wet snow) increased in magnitude with increasing frequency (or shorter wavelength) and is almost independent of angle for angles of incidence $>30°$, particularly at higher frequencies. Goodison *et al.* (1980) found little or no relationship between radar return and snowpack properties under either wet or dry snow conditions using L-band airborne SAR. A uniform, low return was found for a given area under both snow-free and snow-covered conditions. However, using X-band, in their study area near Ottawa, Ontario, nonforested areas exhibited higher backscatter when snow cover was present. Areas of ice and dense snow were observed to produce relatively higher radar returns using the X-band SAR (Goodison *et al.*, 1980).

Bernier *et al.* (1999) established a relationship between the backscattering ratios of a winter (snow covered) image and a fall (snow-free) image to estimate the snowpack thermal resistance using the Radarsat SAR. They then estimated the SWE from the thermal resistance and the measured mean density. In related work, Gauthier *et al.* (2001) used Radarsat C-band ScanSAR data to derive SWE in the La Grande River Basin in northern Québec, where they found ScanSAR-derived SWE values to be similar ($\pm 12\%$) to those derived from *in situ* snow measurements in January and March 1999.

SNOW WETNESS

A study of wet snow using C-, L- (1.25 GHz), and P-band (440 MHz) (see Table 2) polarimetric SAR of a mountainous area in Austria (Otzal), Rott *et al.* (1992) showed the importance of surface roughness at C- and L-band frequencies, and the increasing importance of the snow volume contribution with the longer wavelength P-band sensor. Radar polarimetry allows simultaneous measurement of the radar backscatter from a given surface at a number of different polarizations. Using European Remote Sensing Satellite (ERS)-1 images acquired before, during and after the melt period, Koskinen *et al.* (1997) successfully mapped wet snow with C-band SAR in unforested and sparsely forested regions in northern Finland.

Table 2 Band designations, wavelength, and frequency ranges for Earth-sensing radars (Long, 1975)

Band	Wavelength range (cm)	Frequency range (GHz)
Ka-band	0.75–1.1	26.5–40.0
K-band	1.1–1.67	18.0–26.5
Ku-band	1.67–2.4	12.5–18.0
X-band	2.4–3.75	8.0–12.5
C-band	3.75–7.5	4.0–8.0
S-band	7.5–15.0	2.0–4.0
L-band	15.0–30.0	1.0–2.0
P-band	30.0–100.0	0.3–1.0

Extensive aircraft and ground measurements were obtained by the Canada Centre for Remote Sensing (CCRS) over agricultural areas in southern Québec, Canada, in 1988–1990 (Bernier and Fortin, 1998). It was concluded that, at C-band, volume scattering from a shallow dry snowpack (SWE <20 cm) is not detectable. The backscatter using a C-band SAR emanates from the snow/soil interface when the snowpack is dry. Earlier, Mätzler and Schanda (1984) and Mätzler (1987) had concluded that the backscatter change between an unfrozen bare soil and a dry snow cover over unfrozen soil was small – on the order of 1.5 dB at X-band. Rott and Mätzler (1987) found no significant difference between snow-free and dry-snow-covered regions at 10.4 GHz. However, there may be a potential for detecting shallow, dry snow cover with C-band SAR data when the soil beneath the snow is frozen (Bernier and Fortin, 1998).

Though the main scattering contribution from a dry snow cover is from the ground/snow interface, small changes in the snow can be detected using tandem pairs from repeat-pass interferometric synthetic-aperture radar (InSAR) data. Using the coherence measurement of repeat passes, Shi and Dozier (2000) found that both wet and dry snow can be mapped as evidenced by comparison of snow mapped using Landsat imagery. Refraction of microwaves in dry snow was shown to have a significant effect on the interferometric phase difference and a relationship between changes in SWE and the interferometric phase was derived. Using three tandem pairs of InSAR data, Guneriussen *et al.* (2001) found that a snow density of $0.2\,\mathrm{g\,cm^{-2}}$ at 23° incidence angle gives phase wrapping for changes in snow depth of 16.4 cm and equals a SWE of 3.3 cm.

A promising technology for measuring snow cover is scatterometry. A Ku-band (14.6 GHz) scatterometer operated for three months from July to September 1978 on the Seasat satellite, and results show that some of the glacier facies could be mapped using derived backscatter images (Long and Drinkwater, 1994). The NASA Scatterometer (NSCAT) operated on the Advanced Environmental Observation Satellite (ADEOS) from September 1996 to June 1997 and also permitted study of melt zones on Greenland (Long and Drinkwater, 1999). Timing of melt onset was detected by Nghiem *et al.* (2001) on the Greenland Ice Sheet by a sharp decrease in backscatter, and verified with *in situ* measurements, using the SeaWinds scatterometer on the QuikSCAT satellite.

For seasonal snow cover, Nghiem and Tsai (2001) show that NSCAT backscatter patterns reveal boundaries that correspond to various snow classes as defined by Sturm *et al.* (1995). Additionally, they show rapid changes in the backscatter over the northern plains of the United States and the Canadian prairies that led to the major spring 1997 floods in the mid-western United States and southern Canada (Figure 7).

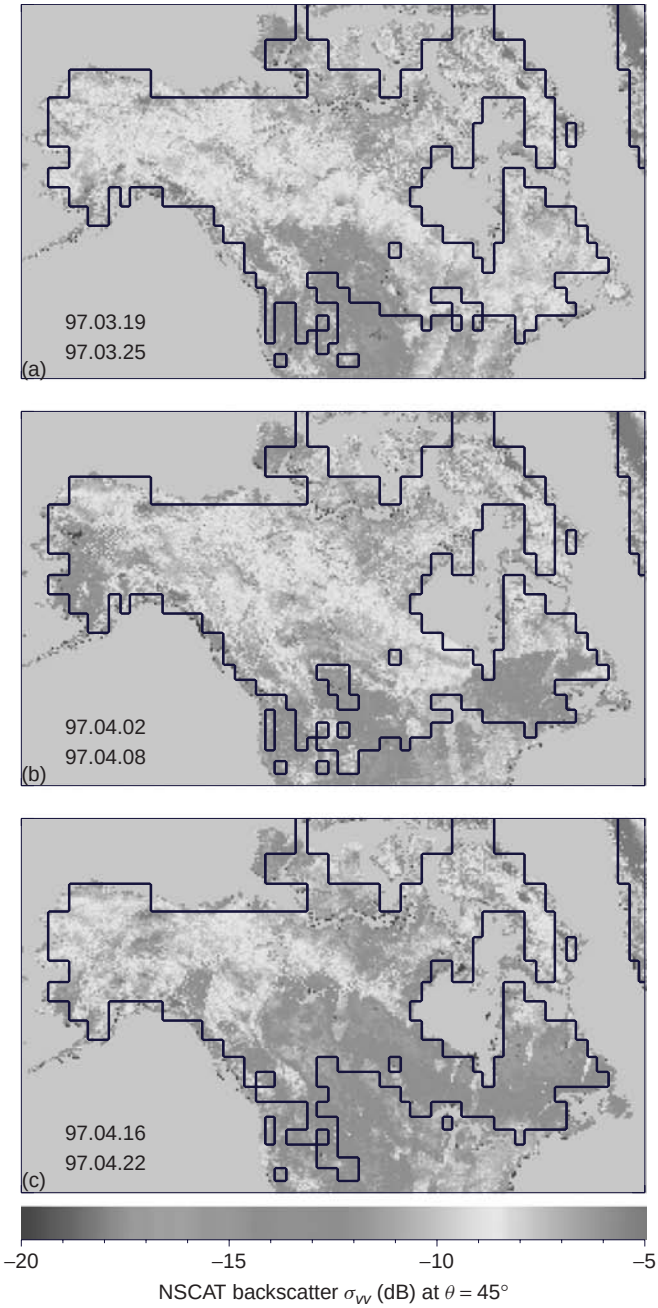


Figure 7 NSCAT backscatter signatures over snow cover corresponding to snow events leading to the 1997 flood in the northern plains of the United States and the Canadian prairies: (a) period of snowmelt from March 19 to March 25, 1997; (b) period of the blizzard from April 2 to April 8, 1997; and (c) period of rapid snow retreat from April 16 to April 22, 1997 (Nghiem *et al.*, 2001). (©2001 IEEE, Courtesy of Son Nghiem, Jet Propulsion Laboratory, Pasadena, CA). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Future Directions and Conclusions

Mapping snow cover areal extent using satellite observations is relatively mature and well validated (see, e.g., Robinson, 1993, 1999; Hughes *et al.*, 1996; Frei *et al.*, 1999; Hall *et al.*, 2002a; Brown *et al.*, 2003; Mauer *et al.*, 2003). Recent global water and climate system studies have begun to examine the link between snow cover areal extent and atmospheric dynamics. For example, Cohen and Entekhabi (1999) investigated the link between early season snow extent in Eurasia and the dynamics of the Siberian high. Saunders *et al.* (2003) show a link between summer snow extent and the winter North Atlantic Oscillation. These studies are important for our understanding of the role of snow in the Earth's hydrological cycle and how it affects human sustainability, especially in regions that are heavily dependent on snowmelt runoff for water supply.

The methodology to map global SWE from remote-sensing instruments is less mature. While microwave remote-sensing observations are helping to advance our ability to effectively characterize water storage in snowpacks, there remain uncertainties about the retrievals from these instruments. Historically, the frequency configurations of space borne active radar systems have produced measurements that are sensitive to the presence or absence of wet snow only and little or no direct information about SWE can be determined from these instruments. Satellite passive microwave measurements now have a 25- year record from which SWE can be estimated. However, with the characteristically large instantaneous fields of view that characterize these instruments (tens of kilometers), the uncertainties associated with SWE estimates are difficult to quantify, and, therefore, are still under investigation. Studies have shown that in noncomplex terrain with low-stand vegetation, reasonable estimates of SWE can be obtained from passive microwave measurements. In other terrain types, however, larger uncertainties persist.

In order to characterize snow water storage, new and improved satellite instrument measurement techniques need to be developed, especially for instruments in the microwave part of the electromagnetic spectrum that are sensitive to snow volume. Ku-band radar measurements are sensitive to SWE and could be developed to resolve fine spatial variations of SWE (tens of meters) through SAR technology. For passive microwave measurements, the relatively low spatial resolution is a key cause of uncertainties in the estimation of SWE. Only through technology improvements in antenna design can instantaneous fields of view be significantly improved, thus increasing the spatial resolution (e.g. Doiron *et al.*, 2004).

With improvements in microwave instrument and measurement techniques of SWE, the uncertainties and errors in SWE estimates will be reduced providing more confidence in our ability to estimate snow water storage throughout the year. If these technology developments come to fruition and new, "more SWE-capable" microwave missions overlap with the current and planned global multidisciplinary instruments, such as the available AMSR-E and the proposed Conical Scanning Microwave Imager/Sounder (CMIS) planned for National Polar-orbiting Operational Environmental Satellite System (NPOESS), the benefits could be great (Kelly *et al.*, 2004). Local scale SWE characterization would be possible; the prospect of combining high spatial resolution-accurate SWE measurements with sophisticated numerical land surface models may then be possible, and is a very exciting one from the water resource management perspective.

Satellite remote sensing has been used to map snow cover for nearly 40 years. Decade-scale CDR-quality records of snow-covered area are already in existence for the Northern Hemisphere (Robinson and Frei, 2000) and are useful in climate models, however, problems exist in developing a CDR for SWE (Armstrong and Brodzik, 2001; Derksen *et al.*, 2003), as discussed above. We can now extend the snow-covered area record using SMMR, SSM/I MODIS, and AMSR data to the global scale, and provide CDR-quality maps of snow-covered area, and continue to study the development of CDR-quality datasets of SWE and snow albedo using visible, near infrared, and passive microwave sensors such as the MODIS and AMSR.

The trend has been toward increasingly automatically processed quantitative maps with error bars provided. Automated processing is necessary so that consistent products can be derived from the observations and long duration data sets might ultimately be available for long-term water-cycle studies. The error estimates associated with snow products are also essential if the products are to be used effectively in combination with catchment-based land surface or climate models. This is because models that require snow-state parameters often require the errors associated with the estimated snow states, especially if data assimilation techniques are used to generate blended products. Whether the snow products are used for initial conditioning or as a forcing variable in a model, or whether the products are used in their own right, the role of remotely sensed observations of snow will continue to be important and is set to play an increasingly important role in climate and hydrological forecasting.

Future sensors will permit automated algorithms to be used to create maps that are consistent with existing maps so that the confidence level of the long-term (~40 year) record is high. The quality allows them to be amenable to comparison with long-term records of other geophysical parameters such as global sea ice, and for input to general circulation models.

Acknowledgments

The authors would like to thank Dr. Andrew G. Klein of Texas A & M University for helpful comments on parts of the manuscript.

REFERENCES

Anderson T. (1982) Operational snow mapping by satellites. *Hydrological Aspects of Alpine and High Mountain Areas, Proceedings of the Exeter Symposium*, July 1982, IAHS Publication 138: pp. 149–154.

Armstrong R.L. and Brodzik M.J. (1998) A comparison of northern hemisphere snow extent derived from passive microwave and visible remote-sensing data. *Proceedings of IGARSS'98, 18th International Geoscience and Remote Sensing Symposium*, 6–10 July 1998, Vol. 3, IEEE: Seattle, pp. 1255–1257.

Armstrong R.L. and Brodzik M.J. (2001) Recent northern hemisphere snow extent: a comparison of data derived from visible and microwave satellite sensors. *Geophysical Research Letters*, **28**(19), 3673–3676.

Armstrong R.L. and Brodzik M.J. (2002) Hemispheric-scale comparison and evaluation of passive-microwave snow algorithms. *Annals of Glaciology*, **34**, 38–44.

Bader, H. (1962) *The Physics and Mechanics of Snow as a Material*, Cold Regions Research and Engineering Laboratory: Hanover, Report II-B, p. 1.

Barry R. (1983) Research on snow and ice. *Reviews of Geophysics and Space Physics*, **21**, 765–776.

Barry R. (1984) Possible CO_2-induced warming effects on the cryosphere. *Climatic Changes on a Yearly to Millennial Basis*, D. Reidel Publishing Company: pp. 571–604.

Barry R. (1990) Evidence of recent changes in global snow and ice cover. *GeoJournal*, **20**(2), 121–127.

Basist A., Garrett D., Ferraro R., Grody N. and Mitchell K. (1996) A comparison between snow cover products derived from visible and microwave satellite observations. *Journal of Applied Meteorology*, **35**(2), 163–177.

Bauer K.G. and Dutton J.A. (1962) Albedo variations measured from an airplane over several types of surface. *Journal of Geophysical Research*, **67**(6), 2367–2376.

Bernier P.Y. (1987) Microwave remote sensing of snowpack properties: potential and limitations. *Nordic Hydrology*, **18**, 1–20.

Bernier M. and Fortin J.-P. (1998) The potential of time series of C-band SAR data to monitor dry and shallow snow cover.

IEEE Transactions on Geoscience and Remote Sensing, **36**(1), 226–243.

Bernier M., Fortin J.-P., Gauthier Y., Gauthier R., Roy R. and Vincent P. (1999) Determination of snow water equivalent using RADARSAT SAR in eastern Canada. *Hydrological Processes*, **13**, 3041–3051.

Brest C.L. and Goward S.N. (1987) Deriving surface albedo measurements from narrow band satellite data. *International Journal of Remote Sensing*, **8**(3), 351–367.

Brown R.D. (2000) Northern hemisphere snow cover variability and change, 1915–1997. *Journal of Climate*, **13**, 2339–2355.

Brown R.D., Brasnett B. and Robinson D. (2003) Gridded North American monthly snow depth and snow water equivalent for GCM evaluation. *Atmosphere-Ocean*, **41**(1), 1–14.

Brown R.D. and Goodison B.E. (1996) Interannual variability in reconstructed Canadian snow cover, 1915–1992. *Journal of Climate*, **9**, 1299–1318.

Bunting J.T. and d'Entremont R.P. (1982) *Improved Cloud Detection Utilizing Defense Meteorological Satellite Program Near Infrared Measurements*, Air Force Geophysics laboratory: Hanscom AFB, AFGL-TR-82-0027, Environmental Research Papers No. 765, p. 91.

Bussières N., De Sève D. and Walker A. (2002) Evaluation of MODIS snow-cover products over Canadian regions. *Proceedings of IGARSS'02*, dates, Toronto, pp. 2302–2304.

Carroll T.R. (1987) Operational airborne measurements of snow water equivalent and soil moisture using terrestrial gamma radiation in the United States. *Large Scale Effects of Seasonal Snow Cover Proceedings of the Vancouver Symposium*, August 1987, IAHS Publication No. 166: pp. 213–223.

Carroll T.R. (1995) Remote sensing of snow in the cold regions. *Proceedings of the First Moderate Resolution Imaging Spectroradiometer (MODIS) Snow and Ice Workshop*, 13–14 September, 1995, NASA Conference Publication 3318: Greenbelt, pp. 3–14.

Carroll T., Cline D., Fall G., Nilsson A., Li L. and Rost A. (2001) NOHRSC operations and the simulation of snow cover properties for the coterminous U.S. *Proceedings of the 69th Western Snow Conference*, Sun Valley, 16–19 April 2001.

Chang A.T.C., Foster J.L. and Hall D.K. (1987) Nimbus-7 SMMR derived global snow cover parameters. *Annals of Glaciology*, **9**, 39–44.

Chang A.T.C., Foster J.L., Hall D.K., Rango A. and Hartline B.K. (1982) Snow water equivalent estimation by microwave radiometry. *Cold Regions Science and Technology*, **5**(3), 259–267.

Chang A.T.C., Gloersen P., Schmugge T.J., Wilheit T. and Zwally H.J. (1976) Microwave emission from snow and glacier ice. *Journal of Glaciology*, **16**(74), 23–39.

Chang A.T.C., Kelly R.E.J., Foster J.L. and Hall D.K. (2003) Estimation of global snow cover using passive microwave data. *Proceedings of the International Society for Optical Engineering (SPIE)*, Hangshou, pp. 381–390, 24–25 October 2002.

Choudhury B.J. and Chang A.T.C. (1979) Two-stream theory of reflectance of snow. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-17**(3), 63–68.

Cline D.W., Bales R.C. and Dozier J. (1998) Estimating the spatial distribution of snow in mountain basins using remote sensing and energy balance modeling. *Water Resources Research*, **34**(5), 1275–1285.

Cohen J. and Entekhabi D. (1999) Eurasian snow cover variability and northern hemisphere climate predictability. *Geophysical Research Letters*, **26**, 345–348.

Colbeck S.C. (1982) An overview of seasonal snow metamorphism, *Reviews of Geophysics and Space Physics*, **20**(1), 45–61.

Crane R.G. and Anderson M.R. (1984) Satellite discrimination of snow/cloud surfaces. *International Journal of Remote Sensing*, **5**(1), 213–223.

Derksen C., Walker A., LeDrew E. and Goodison B. (2002) Time series analysis of passive-microwave-derived central North American snow water equivalent imagery. *Annals of Glaciology*, **34**, 1–7.

Derksen C., Walker A., LeDrew E. and Goodison B. (2003) Combining SMMR and SSM/I data for time series analysis of central North American snow water equivalent. *Journal of Hydrometeorology*, **4**, 304–316.

Dewey K.F. and Heim R. Jr (1981) *Satellite Observations of Variation in Northern Hemisphere Seasonal Snow Cover*, NOAA Technical Report NESS 87: Washington, DC, p. 83.

Dewey K.F. and Heim R. Jr (1983) *Satellite Observations of Variations in Southern Hemisphere Snow Cover*, NOAA Technical Report NESDIS 1: p. 20.

Dirmhirn I. and Eaton F. (1975) Some characteristics of the albedo of snow. *Journal of Applied Meteorology*, **14**, 375–379.

Doiron T.A., Piepmeier J.R., Hilliard L.M., Shirgur B., Kelly R.E.J., Kim E.J. and Cline D. (2004) One dimensional synthetic aperture radiometer using a parabolic cylinder reflector. *Proceedings of 2004 IEEE Aerospace Conference*, 6–13 March, Montana.

Dozier J. (1984) Snow reflectance from Landsat-4 thematic mapper. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-22**(3), 323–328.

Dozier J. (1987) Recent research in snow hydrology. *Reviews of Geophysics*, **25**(2), 153–161.

Dozier J. (1989) Spectral signature of alpine snow cover from the Landsat thematic mapper. *Remote Sensing of Environment*, **28**, 9–22.

Dozier J. and Painter T.H. (2004) Multispectral and hyperspectral remote sensing of alpine snow properties. *Annual Reviews of Earth and Planetary Science*, **32**, 465–494.

Dozier J., Schneider S.R. and McGinnis D.F. Jr (1981) Effect of grain size and snowpack water equivalence on visible and near-infrared satellite observations of snow. *Water Resources Research*, **17**, 1213–1221.

Duguay C. and LeDrew E.F. (1992) Estimating surface reflectance and albedo from Landsat-5 thematic mapper over rugged terrain. *Photogrammetric Engineering and Remote Sensing*, **58**(5), 551–558.

Evans S. (1965) The dielectric properties of ice and snow – a review. *Journal of Glaciology*, **5**, 773–792.

Foster J.L., Barton J.S., Chang A.T.C. and Hall D.K. (2000) Snow crystal orientation effects on the scattering of passive microwave radiation. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(5), 2430–2433.

Foster J.L., Chang A.T.C. and Hall D.K. (1997) Comparison of snow mass estimates from a prototype passive microwave snow

algorithm, a revised algorithm and snow depth climatology. *Remote Sensing of Environment*, **62**, 132–142.

Foster J.L., Chang A.T.C., Hall D.K., Wergin W.P., Erbe E.F. and Barton J. (1999) Effect of snow crystal shape on the scattering of passive microwave radiation. *IEEE Transactions on Geoscience and Remote Sensing*, **37**(2), 1165–1168.

Foster J.L., Hall D.K. and Chang A.T.C. (1987) Remote sensing of snow. *EOS Transactions, American Geophysical Union*, **68**(32), 681–684.

Foster J., Liston G., Koster R., Essery R., Behr H., Dumenil L., Verseghy D., Thompson S., Pollard D. and Cohen J. (1996) Snow cover and snow mass intercomparisons of general circulation models and remotely sensed datasets. *Journal of Climate*, **9**(2), 409–426.

Foster J.L., Sun C., Walker J.P., Kelly R., Chang A., Dong J. and Powell H. (2005) Quantifying the uncertainty in passive microwave snow water equivalent observations, *Remote Sensing of Environment*, **94**, 187–203.

Frei A. and Robinson D.A. (1999) Northern hemisphere snow extent: regional variability 1972–1994. *International Journal of Climatology*, **19**, 1535–1560.

Frei A., Robinson D.A. and Hughes M.G. (1999) North American snow extent: 1900–1994. *International Journal of Climatology*, **19**, 1517–1534.

Gauthier Y., Bernier M., Fortin J.-P., Gauthier R., Roy R. and Vincent P. (2001) Operational determination of snow water equivalent using Radarsat data over a large hydroelectric complex in eastern Canada. *Proceedings of a Symposium on Remote Sensing and Hydrology 2000*, Santa Fe, April 2000, Owe M., Brubaker K., Ritchie J. and Rango A. (Eds.), IAHS Publication No. 267: 343–348.

Gerland S., Winther J.-G., Orbaek J.B., Liston G.E., Oritsland N.A., Blanco A. and Ivanov B. (1999) Physical and optical properties of snow covering Arctic tundra on Svalbard. *Hydrological Processes*, **13**, 2331–2343.

Goita K., Walker A.E. and Goodison B.E. (2003) Algorithm development for the estimation of snow water equivalent in the boreal forest using passive microwave data. *International Journal of Remote Sensing*, **24**(5), 1097–1102.

Goodison B. (1989) Determination of areal snow water equivalent on the Canadian Prairies using passive microwave satellite data. *Proceedings of the IGARSS'89 Symposium*, Vancouver, pp. 1243–1246, July 1989.

Goodison B.E., Rubinstein I., Thirkettle F.W. and Langham E.J. (1986) Determination of snow water equivalent on the Canadian prairies using microwave radiometry. *Modelling Snowmelt-Induced Processes, Proceedings of the Budapest Symposium*, July 1986, IAHS Publication No. 155: pp. 163–173.

Goodison B. and Walker A. (1994) Canadian development and use of snow cover information from passive microwave satellite data. In *Proceedings of the ESA/NASA International Workshop*, Choudhury B.J., Kerr Y.H., Njoku E.G. and Pampaloni P. (Eds.), pp. 245–262.

Goodison B.E., Waterman S.E. and Langham E.J. (1980) Application of synthetic aperture radar data to snow cover monitoring. *Proceedings of the 6th Canadian Symposium on Remote Sensing*, Halifax, pp. 263–271, 21–23 May 1980.

Grenfell T.C. and Maykut G.A. (1977) The optical properties of snow and ice in the arctic basin. *Journal of Glaciology*, **18**, 445–463.

Grenfell T.C., Perovich D.K. and Ogren J.A. (1981) Spectral albedos of an alpine snowpack. *Cold Regions Science and Technology*, **4**, 121–127.

Greuell W. and de Ruyter de Wildt M. (1999) Anisotropic reflectance by melting glacier ice: measurements and parameterizations in Landsat TM bands 2 and 4. *Remote Sensing of Environment*, **70**, 265–277.

Greuell W., Reijmer C.H. and Oerlemans J. (2002) Narrowband-to-broadband albedo conversion for glacier ice and snow based on aircraft and near-surface measurements. *Remote Sensing of Environment*, **82**, 48–63.

Grody N. and Basist A. (1996) Global identification of snowcover using SSM/I measurements. *IEEE Transactions on Geoscience and Remote Sensing*, **34**(1), 237–249.

Guneriussen T., Hogda K.A., Johnsen H. and Lauknes I. (2001) InSAR for estimation of changes in snow water equivalent of dry snow. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(10), 2101–2108.

Hall D.K., Chang A.T.C. and Foster J.L. (1986) Detection of the depth-hoar layer in the snow-pack of the Arctic coastal plain of Alaska, U.S.A., using satellite data. *Journal of Glaciology*, **32**(110), 87–94.

Hall D.K., Chang A.T.C., Foster J.L., Benson C.S. and Kovalick W.M. (1989) Comparison of in situ and satellite-derived reflectances of glacier. *Remote Sensing of Environment*, **28**, 23–31.

Hall D.K., Foster J.L. and Chang A.T.C. (1982) Measurement and modeling of microwave emission from forested snowfields in Michigan. *Nordic Hydrology*, **13**, 129–138.

Hall D.K., Foster J.L., Salomonson V.V., Klein A.G. and Chien J.Y.L. (2001) Development of a technique to assess snow-cover mapping errors from space. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(2), 432–438.

Hall D.K., Riggs G.A., Salomonson V.V., DiGirolamo N.E. and Bayr K.J. (2002a) MODIS snow-cover products. *Remote Sensing of Environment*, **83**, 181–194.

Hall D.K., Kelly R.J.E., Riggs G.A., Chang A.T.C. and Foster J.L. (2002b) Assessment of the relative accuracy of hemispheric-scale snow-cover maps. *Annals of Glaciology*, **34**, 24–30.

Hallikainen M. (1984) Retrieval of snow water equivalent from Nimbus-7 SMMR data: effect of land-cover categories and weather conditions. *IEEE Journal of Oceanic Engineering*, **OE-9**(5), 372–376.

Hallikainen M.T. and Jolma P.A. (1992) Comparison of algorithms for retrieval of snow water equivalent from Nimbus-7 SMMR data in Finland. *IEEE Transactions on Geoscience and Remote Sensing*, **30**(1), 124–131.

Hallikainen M. and Ulaby F.T. (1986) Dielectric and scattering behaviour of snow at microwave frequencies. *Proceedings of the International Geoscience and Remote Sensing Symposium*, Zurich, pp. 87–91, 8–11 September 1986.

Hansen J. and Nazarenko L. (2004) Soot climate forcing via snow and ice albedos. *Proceedings of the National Academy of Sciences*, **101**(2), 423–428.

Hanson K.J. and Viebrock H.J. (1964) Albedo measurements over the northeastern United States. *Monthly Weather Review*, **92**(5), 223–234.

Hobbs P.V. (1974) *Ice Physics*, Clarendon Press: Oxford, p. 837.

Hughes M.G., Frei A. and Robinson D.A. (1996) Historical analysis of North American snow cover extent: merging satellite and station derived snow cover observations. *Proceedings of the 1996 Eastern Snow Conference*, Williamsburg, pp. 21–32.

Hughes M.G. and Robinson D.A. (1996) Historical snow cover variability in the great plains region of the USA: 1910 through to 1993. *International Journal of Climatology*, **16**, 1005–1018.

Josberger E., Campbell W.J., Gloersen P., Chang A.T.C. and Rango A. (1993) Snow conditions and hydrology of the upper Colorado River Basin from satellite passive microwave observations. *Annals of Glaciology*, **17**, 322–326.

Kaufman Y.J., Kleidman R.G., Hall D.K., Martins V.J. and Barton J.S. (2002) Remote sensing of subpixel snow cover using 0.66 and 2.1 μm channels. *Geophysical Research Letters*, **29**(16), doi: 10.1029/2001GL013580.

Kelly R.E.J. (2001) Remote sensing of UK snow covers using multi-sensor satellite imagery. *Proceedings of a Symposium on Remote Sensing and Hydrology 2000*, April 2000, IAHS Publication No. 267: Santa Fe, pp. 72–75.

Kelly R.E.J. and Chang A.T.C. (2003) Development of a passive microwave global snow depth retrieval algorithm for SSM/I and AMSR-E data. *Radio Science*, **38**(4), doi: 10.1029/2002RS002648.

Kelly R.E.J., Chang A.T.C., Foster J.L. and Hall D.K. (2004) Using remote sensing and spatial models to monitor snow depth and snow water equivalent. In *Spatial Modelling of the Terrestrial Environment*, Kelly R.E.J., Drake N.A. and Barr S.L. (Eds.), John Wiley & Sons: pp. 35–58.

Kelly R.E.J., Chang A.T.C., Tsang L. and Foster J.L. (2003) A prototype AMSR-E global snow area and snow depth algorithm. *IEEE Transactions Geoscience and Remote Sensing*, **41**(2), 230–242.

Klein A.G., Hall D.K. and Riggs G.A. (1998) Improving snow-cover mapping in forests through the use of a canopy reflectance model. *Hydrological Processes*, **12**, 1723–1744.

Klein A.G. and Stroeve J. (2002) Development and validation of a snow albedo algorithm for the MODIS instrument. *Annals of Glaciology*, **34**, 45–52.

Knap W.H. and Oerlemans J. (1996) The surface albedo of the Greenland ice sheet: satellite-derived and in situ measurements in the Søndre Strømfjord area during the 1991 melt season. *Journal of Glaciology*, **42**(141), 364–374.

Knap W.H. and Reijmer C.H. (1998) Anisotropy of the reflected radiation field over melting glacier ice: measurements in Landsat TM bands 2 and 4. *Remote Sensing of Environment*, **65**, 93–104.

Knap W.H., Reijmer C.H. and Oerlemans J. (1998) Narrowband to broadband conversion of Landsat TM glacier albedos. *International Journal of Remote Sensing*, **20**(10), 2091–2110.

König M., Winther J.-G. and Isaksson E. (2001) Measuring snow and glacier ice properties from satellite. *Reviews of Geophysics*, **39**(1), 1–27.

Koskinen J.T., Pulliainen J.T. and Hallikainen M.T. (1997) The use of ERS-1 SAR data in snow melt monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 601–610.

Kukla G. and Robinson D.A. (1980) Annual cycle of surface albedo. *Monthly Weather Review*, **108**, 56–68.

Kunzi K.F., Fisher A.D. and Staelin D.H. (1976) Snow and ice surfaces measured by the Nimbus-5 microwave spectrometer. *Journal of Geophysical Research*, **81**, 4965–4980.

Kunzi K.F., Patil S. and Rott H. (1982) Snow-cover parameters retrieved from Nimbus-7 Scanning Multichannel Microwave Radiometer (SMMR) data. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-20**(4), 452–467.

Kurvonen L. and Hallikainen M. (1997) Influence of land-cover category on brightness temperature of snow. *IEEE Transactions on Geoscience and Remote Sensing*, **35**(2), 367–377.

Kyle H.L., Curran R.J., Barnes W.L. and Escoe D. (1978) A cloud physics radiometer. *Third Conference on Atmospheric Radiation*, Davis, pp. 108–109.

Leconte R. (1995) Exploring the behaviour of microwaves in a snowpack using modelling techniques. *Canadian Journal of Remote Sensing*, **22**(1), 23–35.

Leconte R., Carroll T. and Tang P. (1990) Preliminary investigations on monitoring the snow water equivalent using synthetic aperture radar. *Proceedings of the Eastern Snow Conference*, Bangor, pp. 73–86, 7–8 June, 1990.

Ledley T.S., Sundquist E.T., Schwartz S.E., Hall D.K., Fellows J.D. and Killeen T.L. (1999) Climate change and greenhouse gases. *EOS Transactions, American Geophysical Union*, **80**(39), 453–454, 457–458.

Liang S. (2000) Narrow to broadband conversion of land surface albedo I: algorithms. *Remote Sensing of Environment*, **76**, 213–238.

Long D.G. and Drinkwater M.R. (1994) Greenland ice-sheet surface properties observed by the Seasat-A scatterometer at enhanced resolution. *Journal of Glaciology*, **40**(135), 213–230.

Long D.G. and Drinkwater M.R. (1999) Cryosphere applications of NSCAT data. *IEEE Transactions on Geoscience and Remote Sensing*, **37**(3), 1671–1684.

Long M.W. (1975) *Radar Reflectivity of Land and Sea*, D.C. Heath & Company: Lexington.

Male D.H. (1980) The seasonal snowcover. In *Dynamics of Snow and Ice Masses*, Colbeck S. (Ed.), Academic Press: New York, pp. 305–395.

Martinelli M. Jr (1979) Research on snow and ice. *Reviews of Geophysics and Space Physics*, **17**(6), 1253–1288.

Matson M., Roeplewski C.F. and Varnadore M.S. (1986) *An Atlas of Satellite-Derived Northern Hemisphere Snow Cover Frequency*, National Weather Service: Washington, p. 75.

Mätzler C. (1987) Applications on the interactions of microwaves with the natural snow cover. *Remote Sensing Reviews*, **2**, 259–387.

Mätzler C. (1997) Autocorrelation functions of granular media with arrangement of spheres, spherical shells or ellipsoids. *Journal of Applied Physics*, **3**, 1509–1517.

Mätzler C. and Schanda E. (1984) Snow mapping with active microwave sensors. *International Journal of Remote Sensing*, **5**(2), 409–422.

Mauer E.P., Rhoads J.D., Dubayah R.O. and Lettenmaier D.P. (2003) Evaluation of the snow-covered area data product from MODIS. *Hydrological Processes*, **17**, 59–71.

McFadden J.D. and Ragotzkie R.A. (1967) Climatological significance of albedo in central Canada. *Journal of Geophysical Research*, **72**(4), 1135–1143.

Meier M. (1972) Measurement of snow cover using passive microwave radiation. *International Symposia on the Role of Snow and Ice in Hydrology, Proceedings of the Banff Symposium*, Banff, pp. 739–750, September, 1972.

Mekler Y. and Joseph J.H. (1983) Direct determination of surface albedos from satellite imagery. *Journal of Climate and Applied Meteorology*, **22**, 530–558.

Mognard N.M. and Josberger E.G. (2002) Northern Great Plains 1996/97 seasonal evolution of snowpack parameters from satellite passive-microwave measurements. *Annals of Glaciology*, **34**, 15–23.

Nagler T. and Rott H. (2000) Retrieval of wet snow by means of multitemporal SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(2), 754–765.

Nghiem S.V., Steffen K., Kwok R. and Tsai W.Y. (2001) Detection of snowmelt regions on the Greenland ice sheet using diurnal backscatter change. *Journal of Glaciology*, **47**(159), 539–547.

Nghiem S.V. and Tsai W-Y (2001) Global snow cover monitoring with space borne Ku-band scatterometer. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(10), 2118–2134.

Nicodemus F.E., Hsia F.C., Richmond J.J., Ginsberg I.W. and Limperis T. (1977) Geometrical considerations and nomenclature for reflectance, *Technical Report Monograph, 160*, National Bureau of Standards: Gaithersburg.

Nolin A.W., Dozier J. and Mertes L.A.K. (1993) Mapping alpine snow using a spectral mixture modeling technique. *Annals of Glaciology*, **17**, 121–124.

Nolin A.W. and Liang S. (2000) Progress in bidirectional reflectance modeling and applications for surface particulate media: snow and soils. *Remote Sensing Reviews*, **18**, 307–342.

Nolin A.W. and Stroeve J.C. (1997) The changing albedo of the Greenland ice sheet: implications for climate change. *Annals of Glaciology*, **25**, 51–57.

O'Brien H.W. and Munis R.H. (1975) Red and near-infrared spectral reflectance of snow, *Operational Applications of Satellite Snowcover Observations*, a workshop held in South Lake Tahoe, 18–20 August, 1975, NASA SP-391.

Painter T.H., Dozier J., Roberts D.A., Davis R.E. and Green R.O. (2003) Retrieval of subpixel snow-covered area and grain size from imaging spectrometer data. *Remote Sensing of Environment*, **85**, 64–77.

Pivot F.C., Kergomard C. and Duguay C.R. (2002) Use of passive-microwave data to monitor spatial and temporal variations of snow cover at tree line, near Churchill, Manitoba, Canada. *Annals of Glaciology*, **34**, 58–64.

Pomeroy J.W. and Gray D.M. (1995) *Snowcover Accumulation, Relocation and Management*, National Hydrology Research Institute: Environment Canada, NHRI Science Report No. 7, p. 134.

Potts H.L. (1937) A photographic snow survey method of forecasting from photographs. *Transactions of the American Geophysical Union*, South Continental Divide Snow Survey Conference: pp. 658–660.

Ramage J.M. and Isacks B.L. (2002) Determination of melt-onset and refreeze timing on southeast Alaskan icefields using SSM/I diurnal amplitude variations. *Annals of Glaciology*, **34**, 391–398.

Ramsay B. (1998) The interactive multisensor snow and ice mapping system. *Hydrological Processes*, **12**, 1537–1546.

Rango A. and Martinec J. (1979) Application of a snowmelt-runoff model using Landsat data. *Nordic Hydrology*, **10**, 225–238.

Rango A. and Martinec J. (1982) Snow accumulation derived from modified depletion curves of snow coverage. *Symposium on Hydrological Aspects of Alpine and High Mountain Areas*, in Exeter, IAHS Publication No. 138: pp. 83–90.

Rango A. and Salomonson V.V. (1977) Seasonal streamflow estimation in the Himalayan region employing meteorological satellite snow cover observations. *Water Resources Research*, **13**(1), 109–112.

Rango A., Wergin W.P. and Erbe E.F. (1996) Snow crystal imagery uses SEM, Part II metamorphosed snow. *Hydrological Sciences Journal*, **41**, 235–250.

Riggs G.A. and Hall D.K. (2004) Snow mapping with the MODIS Aqua instrument. *Proceedings of the 61st Eastern Snow Conference*, Portland, 9–11 June 2004, pp. 81–84.

Robinson D.A. (1993) Hemispheric snow cover from satellites. *Annals of Glaciology*, **17**, 367–371.

Robinson D.A. (1997) Hemispheric snow cover and surface albedo for model validation. *Annals of Glaciology*, **25**, 241–245.

Robinson D.A. (1999) Northern hemisphere snow cover during the satellite era. *Proceedings of the 5th Conference Polar Meteorology and Oceanography*, American Meteorological Society: Boston, Dallas, pp. 255–260.

Robinson D.A., Dewey K.F. and Heim R.R. (1993) Global snow-cover monitoring: an update. *Bulletin of the American Meteorological Society*, **74**(9), 1689–1696.

Robinson D.A. and Frei A. (2000) Seasonal variability of northern hemisphere snow extent using visible satellite data. *Professional Geographer*, **51**, 307–314.

Robinson D.A. and Kukla G. (1985) Maximum surface albedo of seasonally snow-covered lands in the northern hemisphere. *Journal of Climate and Applied Meteorology*, **24**, 402–411.

Robinson D.A., Scharfen G., Serreze M.C., Kukla G. and Barry R.G. (1986) Snow melt and surface albedo in the arctic basin. *Geophysical Research Letters*, **13**, 945–948.

Robinson D.A., Serreze M.C., Barry R.G., Scharfen G. and Kukla G. (1992) Large-scale patterns and variability of snowmelt and parameterized surface albedo in the arctic basin. *Journal of Climate*, **5**(10), 1109–1119.

Robock A. (1980) The seasonal cycle of snow cover, sea ice and surface albedo. *Monthly Weather Review*, **108**, 267–285.

Romanov P., Gutman G. and Csiszar I. (2000) Automated monitoring of snow cover over North America with multispectral satellite data. *Journal of Applied Meteorology*, **39**, 1866–1880.

Romanov P. and Tarpley D. (2003) Automated monitoring of snow cover over South America using GOES imager

data. *International Journal of Remote Sensing*, **24**(5), 1119–1125.

Rosenthal W. and Dozier J. (1996) Automated mapping of montane snow cover at subpixel resolution from the Landsat thematic mapper. *Water Resources Research*, **32**(1), 115–130.

Ross B. and Walsh J.E. (1987) A comparison of simulated and observed fluctuations in summertime arctic surface albedo. *Journal of Geophysical Research*, **92**(C12), 13115–13125.

Rott H. (1984) The analysis of backscattering properties from SAR data of mountainous regions. *IEEE Journal of Oceanic Engineering*, **OE-0**, 347–355.

Rott H., Davis R.E. and Dozier J. (1992) Polarimetric and multifrequency SAR signatures of wet snow. *Proceedings of the International Geoscience and Remote Sensing Symposium 1992*, Houston, pp. 1658–1660, 26–29 May, 1992.

Rott H., Künzi K. and Patil S. (1981) Temporal and spatial variations of snow properties derived from Nimbus-7 SMMR data. 20th URSI General Assembly, Symposium of Remote Sensing, Washington, August 10–19, 1981.

Rott H. and Mätzler C. (1987) Possibilities and limits of synthetic aperture radar for snow and glacier surveying. *Annals of Glaciology*, **9**, 195–199.

Rott H. and Nagler T. (1993) Capabilities of ERS-1 SAR for snow and glacier monitoring in alpine areas. *Proceedings of the Second ERS-1 Symposium*, 1–6, ESA SP-359.

Rott H. and Nagler T. (1995) Monitoring temporal dynamics of snowmelt with ERS-1 SAR. *Proceedings of IGARSS 95*, Fierenze, pp. 1747–1749, 10–14 July.

Salomonson V.V. and Appel I.L. (2004) Estimating the fractional snow covering using the normalized difference snow index. *Remote Sensing of Environment*, **89**, 351–360.

Salomonson V.V. and Marlatt W.E. (1968) Anisotropic solar reflectance over white sand, snow and stratus clouds. *Journal of Applied Meteorology*, **7**(3), 475–483.

Saunders M.A., Qian B. and Lloyd-Hughes B. (2003) Summer snow extent heralding of the winter North Atlantic Oscillation. *Geophysical Research Letters*, **30**(7), 1378, doi:10.1029/2002GL016832.

Schanda E., Mätzler C. and Künzi K. (1983) Microwave remote sensing of snow cover. *International Journal of Remote Sensing*, **4**(1), 149–158.

Scharfen G., Barry R.G., Robinson D.A., Kukla G. and Serreze M.C. (1987) Large-scale patterns of snow melt on Arctic sea ice mapped from meteorological satellite imagery. *Annals of Glaciology*, **9**, 1–6.

Scharfen G.R., Hall D.K., Khalsa S.J.S., Wolfe J.D., Marquis M.C., Riggs G.A. and McLean B. (2000) Accessing the MODIS snow and ice products at the NSIDC DAAC. *Proceedings of IGARSS'00*, Honolulu, pp. 2059–2061, 23–28 July 2000.

Serreze M.C., Walsh J.E., Chapin F.S. III, Osterkamp T., Dyurgerov M., Romanovsky V., Oechel W.C., Morison J., Zhang T. and Barry R.G. (2000) Observational evidence of recent change in the northern high-latitude environment. *Climate Change*, **46**, 159–207.

Shi J. and Dozier J. (1997) Mapping seasonal snow with SIR-C/X SAR in mountainous regions. *Remote Sensing of Environment*, **59**, 294–307.

Shi J. and Dozier J. (2000) Estimation of snow water equivalence using SIR-C/X SAR, Part I: inferring snow density and subsurface properties. *IEEE Journal of Geoscience and Remote Sensing*, **38**(6), 2465–2474.

Shi J., Dozier J. and Rott H. (1994) Snow mapping in alpine regions with synthetic aperture radar. *IEEE Journal of Geoscience and Remote Sensing*, **32**(1), 152–158.

Simpson J.J. and McIntire T.J. (2001) A recurrent neural network classifier for improved retrievals of areal extent of snow cover. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(10), 2135–2147.

Singer F.S. and Popham R.W. (1963) Non-meteorological observations from weather satellites. *Astronautics and Aerospace Engineering*, **1**(3), 89–92.

Standley A.P. and Barrett E.C. (1999) The use of coincident DMSP SSM/I and OLS satellite data to improve snow cover detection and discrimination. *International Journal of Remote Sensing*, **20**(2), 285–305.

Steffen K. (1987) Bidirectional reflectance of snow at 500–600 nm. In *Large-Scale Effects of Seasonal Snow Cover*, Proceedings of the Vancouver Symposium, August 1987, Goodison B. (Ed.), IAHS: Vancouver, pp. 415–425.

Stiles W.H. and Ulaby F.T. (1980) The active and passive microwave response to snow parameters – 1. Wetness. *Journal of Geophysical Research*, **85**(C2), 1037–1044.

Stiles W.H., Ulaby F.T. and Rango A. (1981) Microwave measurements of snowpack properties. *Nordic Hydrology*, **12**, 143–166.

Stroeve J., Nolin A. and Steffen K. (1997) Comparison of AVHRR-derived and in situ surface albedo over the Greenland Ice sheet. *Remote Sensing of Environment*, **62**, 262–276.

Sturm M. and Benson C.S. (1997) Vapor transport, grain growth and depth hoar development in the subarctic snow. *Journal of Glaciology*, **43**(143), 42–59.

Sturm M., Holmgren J. and Liston G.E. (1995) A seasonal snow cover classification system for local to global applications. *Journal of Climatology*, **8**(5), 1261–1283.

Tait A.B., Barton J. and Hall D.K. (2001) A prototype MODIS-SSM/I snow mapping method. *Proceedings of a Symposium on Remote Sensing and Hydrology 2000*, April 2000, IAHS Publication No. 267: Santa Fe, pp. 139–144.

Tait A.B., Hall D.K., Foster J.L. and Armstrong R.L. (2000) Utilizing multiple datasets for snow-cover mapping. *Remote Sensing of Environment*, **72**, 111–126.

Tiuri M. and Hallikainen M. (1981) Microwave emission characteristics of snow covered earth surfaces measured by the Nimbus-7 satellite. *11th European Microwave Conference*, Amsterdam, September 7–11, 1981.

Trabant D. and Benson C. (1972) Field experiments on the development of depth hoar. *Studies in Mineralogy and Precambrian Geology*, Doe B.R. and Smith K.K. (Eds.), Geological Society of America Memoir 135: pp. 309–322.

Tsang L., Chen C.T., Chang A.T.C., Guo J. and Ding K.H. (2000) Dense media transfer theory based on quasicrystalline approximation with application to passive microwave remote sensing of new snow. *Radio Science*, **35**(3), 731–749.

Tsang L. and Kong J.A. (2001) *Scattering of Electromagnetic Waves: Advanced Topics*, John Wiley & Sons: New York.

Ulaby F.T., Moore R.K. and Fung A.K. (1986) *Microwave Remote Sensing, Active and Passive*, From *Theory to Applications*, Vol. 3, Addison-Wesley Publishing Company: Reading, p. 2162.

Ulaby F.T. and Stiles W.H. (1980) The active and passive microwave response to snow parameters 2. water equivalent of dry snow. *Journal of Geophysical Research*, **85**(C2), 1045–1049.

Ulaby F.T. and Stiles W.H. (1981) Microwave response of snow. *Advanced Space Research*, **1**, 131–149.

Ulaby F.T., Stiles W.H., Dellwig L.F. and Hanson B.C. (1977) Experiments on the radar backscatter of snow. *IEEE Transactions on Geoscience Electronics*, **GE-15**(4), 185–189.

Vikhamer D. and Solberg R. (2002) Subpixel mapping of snow cover in forests by optical remote sensing. *Remote Sensing of Environment*, **84**, 69–82.

Waite W.P. and MacDonald H.C. (1970) Snowfield mapping with K-band radar. *Remote Sensing of Environment*, **1**, 143–150.

Walker A.E. and Goodison B.E. (1993) Discrimination of a wet snow cover using passive-microwave satellite data. *Annals of Glaciology*, **17**, 307–311.

Walker A.E. and Goodison B. (2000) Challenges in determining snow water equivalent over Canada using microwave radiometry. *Proceedings of IGARSS 2000*, Honolulu, 24–28 July 2000.

Walker A.E. and Silis A. (2002) Snow-cover variations over the MacKenzie River Basin, Canada, derived from SSM/I passive-microwave satellite data. *Annals of Glaciology*, **34**, 8–14.

Warren S.G. (1982) Optical properties of snow. *Reviews of Geophysics and Space Physics*, **20**(1), 67–89.

Warren S.G. and Wiscombe W.J. (1980) A model for the spectral albedo of snow, II: snow containing atmospheric aerosols. *Journal of the Atmospheric Sciences*, **37**, 2734–2745.

Wergin W.P., Rango A., Erbe E.F. and Murphy C. (1996) Low temperature SEM of precipitated and metamorphosed snow crystals collected and transported from remote sites. *Journal of Microscopy Society of America*, **2**(3), 99–112.

Wergin W.P., Rango A., Foster J.L., Erbe E.F. and Pooley C. (2002) Irregular snow crystals: structural features as revealed by low temperature scanning electron microscopy. *Scanning*, **24**, 249–256.

Winther J.-G. (1993) Landsat TM-derived and in situ summer reflectance of glaciers in Svalbard. *Polar Research*, **12**(1), 37–55.

Winther J-G., Gerland S., Orbaek J.B., Ivanov B., Blanco A. and Boike J. (1999) *Hydrological Processes*, **13**, 2033–2049.

Wiscombe W.J. and Warren S.G. (1980) A model for the spectral albedo of snow, I. Pure snow. *Journal of the Atmospheric Sciences*, **37**, 2712–2733.

56: Estimation of Glaciers and Sea-ice Extent and their Properties

JULIENNE STROEVE[1], DAVID LONG[2], JOSEFINO C COMISO[3], TED A SCAMBOS[1] AND CHRIS A SHUMAN[3]

[1]National Snow and Ice Data Center, University of Colorado, Boulder, CO, US
[2]Electrical and Computer Engineering Department, Brigham Young University, Provo, UT, US
[3]NASA/Goddard Space Flight Center, Greenbelt, MD, US

Past satellite missions have demonstrated the utility of spaceborne remote sensing in studies of polar ice. Because of their ability to "see" into polar ice without regard to cloud cover or solar lighting, microwave remote sensing instruments have played and will continue to play an increasing role in such studies and in long-term monitoring the polar regions. Microwave sensors can be either active radars which transmit and receive, or receive-only radiometers. Spaceborne microwave sensors such as radiometers and scatterometers offer wide-area, frequent coverage but at lower resolution and a longer historical database. The historic data sets are particularly important since they provide a baseline for studies of global change. Synthetic aperture radar systems possess high resolution capability but have restricted spatial and temporal coverage. Microwave sensors are sensitive to snow and ice structure and are particularly sensitive to freeze/thaw conditions. Optical sensors can also be active (Lidars) or passive. Laser altimeters are an example of an active optical sensor. They provide accurate ice sheet topography information. Passive optical sensing is useful in change detection and infrared ice temperature sensing. In this chapter, essential background in microwave and optical remote sensing of sea-ice, glaciers and ice sheets is provided, along with application examples.

INTRODUCTION

Snow, sea ice, and land ice exert a profound influence not only on the local climates but also on the global climate through their high albedos and thermal and radiative properties. A common feature of climate model projections is that largely due to albedo feedbacks involving snow and sea ice, the effects of greenhouse gas (GHG) loading will be observed first, and will be most pronounced in the polar regions. Thus, long-term, internally consistent datasets are required to detect any climate-change-related trends that may be occurring in these regions.

Remote sensing utilizing instruments spanning the visible through microwave portion of the electromagnetic spectrum have found many applications for mapping sea and land ice. The advantages of satellite remote sensing in mapping snow and ice cover is that large and inaccessible areas such as the polar regions can be studied even under darkness and through fog, cloud, or rain. This latter point is particularly important in the polar regions that are frequently cloud-covered and undergo six months of darkness. This article discusses applications of different satellite sensors to the studies of ice sheets, glaciers, and sea ice. Here, we address passive and active microwave and optical remote sensing as well as infrared observations. We begin the discussion by reviewing the relevant physical properties of sea ice and snow that affect the satellite signal.

PHYSICAL STRUCTURE AND TEMPORAL EVOLUTION OF SEA ICE AND ICE SHEETS

Using remote sensing data is essential to the study of sea ice and ice sheets because of the difficulty in accessing these large and inhospitable remote areas. Remote sensing

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

in the microwave wavelength region is particularly useful because of the ability of microwave instruments to collect data through cloud cover and polar darkness. This section discusses some of the basic physical properties of sea ice, land ice, and snow that affect the satellite signal.

Sea Ice

Sea ice is a complex structure consisting of a variably aged ice matrix with inclusions of brine, air, and solid salts. Scattering and emission of microwave radiation is very sensitive to small variations in the structure and composition of the sea ice. Snow cover on top of the sea ice acts as an insulating blanket for the sea ice since it has a thermal conductivity that is approximately 10 times lower than that of the underlying sea ice and therefore reduces the conductive flux between the ocean and the atmosphere. Snow cover significantly alters the microwave emission of the sea ice as well as the surface albedo and will be discussed later. Next we discuss the cycle of formation and deformation of sea ice. The process can be broken down into four steps: (i) formation, (ii) growth, (iii) deformation, and finally (iv) disintegration.

Formation and Growth

Sea ice tends to form first in shallow waters (i.e. waters of small depths cool faster), in areas of reduced current (i.e. bays and inlets), and also in areas near the mouths of rivers (i.e. areas of low salinity). The greater the depth of high-salinity water, the later (longer) the time until freezing. At the first signs of freezing, an oily appearance occurs on the water caused by the formation of needle-like crystals or plates of about 10 mm, made up of pure ice (no salts). This is called *frazil ice*. As freezing continues, these needle-like crystals increase in number and form a thin layer of slushy ice (or *grease ice*).

The salt in the ocean does not initially become part of the crystals that form, and thus the nearby water becomes saltier. This denser water sinks, and as more crystals form, the salts eventually become frozen to one another in such a way that tiny spaces or pockets remain between groups of crystals. Some of the residual brine will become trapped in those pockets, which are called *brine cells*. Depending on the rate of freezing, the remaining brine will diffuse downward into the water (if freezing is slow), or become trapped in the brine cells (if freezing is rapid). Thus, ice that forms rapidly is saltier than ice that forms slowly with consequent impacts on the ice's material properties.

If the air temperature is very low, a thin sheet of ice will grow quickly in a short span of time, but as the ice becomes thicker (e.g. 8–10 cm), the growth rate slows as the ice thickens further. Since the snow cover acts as a blanket, a few centimeters of snow on top of the ice will drastically slow down the rate of ice thickness growth. Under appropriate conditions, *frost flowers* may form on the ice surface. Frost flowers form directly from water vapor. Salt flowers are caused by the channeling of brine into frost flowers. These intricate crystal forms are important because they can cover large areas and can significantly alter the microwave signature, increasing backscatter and decreasing emission (Rankin *et al.*, 2003).

Deformation

Sea ice changes shape through the movement of winds and currents, by thermal expansion and contraction of the ice, or by contact with converging ice or shorelines. Ice expands rapidly when it first forms and will continue to expand until a critical temperature is reached, after which it contracts slightly: the greater the salinity of the ice, the greater the expansion. The following describes some of the various ice formations caused by deformation:

- **Pressure ridges** form on the ice surface because of thermal expansion or from the pressure exerted on the ice by wind or currents. Pressure ridges can extend downward four to five times as far as they extend upward.
- **Hummocks** are small hills of broken ice that have been forced upwards by pressure. They are less salty than pressure ridges and have a melting point close to 0 °C.
- **Ice floes** are formed by the cracking and breaking of a solid ice layer into separate pieces.
- **Rafting** occurs when two floes are pressed together so that one lies on top of the other.
- **Cracks or Leads** occur when the ice layer breaks open because of the action of winds and/or currents and the floes separate, or because of contraction.

Disintegration

Sea ice disintegrates by:

- *Melting.* Absorption of sun's radiation by snow and ice (after the snow melts off) and conduction of heat from surrounding air, water, and land contribute to the loss of ice due to melting. Once there is open water, such as leads or cracks close to the ice, or water on top of the ice (ponds), melting is accelerated.
- *Conduction.* When warm air is in contact with the liquid water on the ice, heat is transferred from the air to the water. Heat is conducted down into the water, which in turn accelerates the melting of the ice underneath.

Glacial Ice and the Major Ice Sheets

Because of the distinct differences in temporal scale, this section will not be broken down into formation, deformation, and disintegration sections, as was done in the previous

text. It is important to note that glacial ice generally originates as snowfall, which accumulates over hundreds to tens of thousands of years and eventually begins to flow under its own mass. Snow deposition and seasonal melting continually modify the surface. Aeolian processes produce snow drifts and small dunes known as *sastrugi*. Some megadunes in Antarctica are many kilometers in size, with lengths up to 100 km and wavelengths of 5 km. Under appropriate conditions, surface meltwater can percolate down into the snow layer creating ice lenses and other structures that can greatly affect the microwave signature of the ice. Movement of the ice creates crevasses and surface expressions of basal topography greatly modifying the surface of flowing glaciers. Frost flowers and hoar can also form on the ice surface or in the near-surface layers.

On the Greenland ice sheet, a wide range of snow and ice zones are found (Benson, 1962). The boundaries between the facies are sensitive indicators of seasonal climate characteristics (Long and Drinkwater, 1994; Jezek *et al.*, 1994). The dry snow facies are located at the highest elevations of an ice sheet (or glacier), where there is little or no surface melt. The percolation facies, where some melt occurs but meltwater percolates into the snowpack and refreezes, is just below the dry-snow zone in elevation. It is not easily distinguished from the wet snow facies, where complete annual accumulation of snow may experience melt and refreezing. However, at the lower elevations of the wet snow facies there are melt ponds and areas of slush. Still lower is the superimposed ice facies, which represents the zone of refrozen melted snow. Because superimposed ice is created by the melting and refreezing of this year's snow, this zone is still considered part of the ice-sheet accumulation area. The lowest elevation zone is the ice facies, which represents the ablation region. In this region, the annual accumulation of snow and ice is fully melted and ice from the previous years can be removed. Melt is critically important to the Earth's radiation budget, as wet snow can absorb 45% more incoming solar radiation than dry snow (Abdalati and Steffen, 1995). The equilibrium line altitude on a glacier or ice sheet is defined as the boundary between the accumulation area and the ablation area. This line is where the annual mass balance is zero. The firn line is the minimum elevation of firn lying on a glacier surface. Each year's firn line marks a glacier or ice sheet's annual equilibrium line.

Snow metamorphism, and the resulting stratigraphy of the snow and firn (snow that is older than one year), is enhanced by the thermal regime during seasonal change as well as the presence of meltwater. In Greenland, the general evolution in the dry-snow zone closely follows thermal forcing within the upper few meters of the snowpack. Diagenetic changes in the snow and firn can take place over time without melt, but are accelerated with higher temperatures. When summer temperatures in the upper snowpack increase above $-10\,^{\circ}C$, the vapor pressure increases, leading to increased vapor transport and sublimation. The net result is the destruction of original precipitated snow crystals, increased grain sizes, and increased bonding between grains. These changes in snow and firn are common to the spring and summer months in the dry-snow zone at elevations that are too high to experience actual melting.

Further downslope, as summer surface air temperatures increase, diagenesis takes place much more rapidly and dramatically as snow and ice melts (Echelmeyer *et al.*, 1992). Below the dry-snow line lies the percolation zone, where meltwater can percolate downward to form pipes and channels. When meltwater encounters obstructions or snow below the freezing point, it moves laterally to form layers or buried ice lenses and glands (Pfeffer *et al.*, 1990), parallel to the strata. As meteorological conditions vary, the pipes that feed these hard ice layers with melt water may also freeze, resulting in vertically extended ice masses. At lower altitudes where melting is significant, the seasonal snowpack typically becomes isothermal and may melt away to reveal bare glacial ice.

Layering develops in the accumulation zone of ice sheets because of snow precipitation and redistribution patterns and the changes affected by seasonal weather. Such stratigraphy is preserved over time and can be often correlated using passive microwave data over scales of tens or even hundreds of kilometers (Shuman and Alley, 1993). Stratigraphic discontinuities such as hoar layers form because of seasonal changes such as a strong thermal gradient in the near-surface along with low accumulation and wind speed that are characteristic of the austral summer in Greenland. In addition, a discrete boundary in the snow pack often produces a coarse-grained, low-density layer, overlain by a finer-grained, harder layer of higher density, slightly above the previous summer's surface melting. Benson (1962) observed the formation of this sequence under strong negative temperature gradients in mid- to late August during the fall season. In the percolation zone, stratigraphic horizons such as ice firn or ice lenses are seasonal markers. In bare ice zones, stratigraphic markers are effectively removed.

MICROWAVE REMOTE SENSING APPLICATIONS

Remote sensing utilizing instruments in the microwave portion of the electromagnetic spectrum is well suited for large-scale sea ice, ice sheet, and snow studies, including the estimation of the fraction of sea ice, snow accumulation on ice sheets, and snowmelt. Microwave sensors can be divided into two classes: passive (radiometers) and active (radars). Radiometers are passive, receive-only sensors that measure the thermal emission (brightness temperature) of the target in the microwave band. Radars transmit pulses of

microwave energy towards the Earth's surface and measure the reflected energy (Ulaby *et al.*, 1981).

Data from passive microwave instruments now span nearly three decades; active sensors cover somewhat lesser time period. Data from passive microwave radiometers have been available since December 1972 with the launch of the Nimbus-5 Electrically Scanning Microwave Radiometer (ESMR). This early sensor provided the first synoptic overview of the polar ice masses that were not affected by weather or polar darkness for five years. The Nimbus-7 Scanning Multichannel Microwave Radiometer (SMMR) provided the first multifrequency passive microwave observations for almost a decade, from 1978 to 1987. The operational mode Special Sensor Microwave/Imager (SSM/I) onboard a series of Defense Meteorological Satellite Program (DMSP) satellites has continued the satellite passive microwave observations since 1987. Now, with the launch of the Advanced Microwave Scanning Radiometer (AMSR) onboard the Aqua platform, these multidecadal time series will be continued and enhanced through the additional spectral channels, enhanced resolution, and improved instrument performance offered by AMSR.

Active microwave sensors include synthetic aperture radars (SARs), scatterometers, and altimeters (Ulaby *et al.*, 1981). SARs are imaging radars that measure the surface backscatter at fine resolution (Henderson and Lewis, 1998), but have only limited coverage and are typically less well calibrated than scatterometers. Microwave scatterometers were originally designed to measure winds over the ocean but have also been found to be very useful in cryospheric studies (Long *et al.*, 2001). Scatterometers have low spatial resolution, but offer wide-area and frequent coverage. Altimeters are designed to measure surface topography and have very narrow swaths.

The first SAR and scatterometer flew in 1978 aboard Seasat and operated for 3 months. Later active sensors include the European Remote Sensing (ERS) 1 and 2, Active Microwave Instrument (AMI), which included both scatterometer and SAR modes operating at C-band, the Ku-band NASA Scatterometer (NSCAT), and the Ku-band SeaWinds scatterometer series. ERS-1 operated from 1992 through 1996, while ERS-2 operated from 1996 through 2000, with limited current operation to the present. NSCAT operated for 9 months in 1996–1997. The first SeaWinds was launched aboard QuikSCAT in 1999 and continues operation into 2005. The second SeaWinds operated for 10 months aboard the Japanese ADEOS-II spacecraft after its launch in December 2002. Several Space Shuttle-based SAR missions have been flown, including SIR-A, SIR-B, SIR-C, and SRTF. Other orbital SARs include the Japanese JERS-1, the Canadian RADARSAT-1, and the European EnviSAT. Interferometric SAR uses phase information collected from two different antennas on the same spacecraft (SRTF) or from two different orbits (other sensors) to infer surface topography (Ulaby *et al.*, 1981). The RADARSAT Antarctic Mapping Project (RAMP) collected interferometric SAR data and created the first high-resolution topographic map of Antarctica in 1997 (Jezek, 1999; Jezek *et al.*, 2000). A second map was created in 2000. Interferometric SAR data measure ice sheet and glacier velocities (e.g. Kwok and Fahnestock, 1996; Rignot, 1996; Young and Hyland, 2002).

A number of satellite altimeter missions have been flown, including Seasat-A, TOPEX, GEOSAT, ERS-1/2, and JASON-1. These have been used to map ice-sheet topography and study changes in the major ice sheets (e.g. Zwally and Bindschadler, 1983; Zwally *et al.*, 1987; Davis and Zwally, 1993; Bamber, 1994). In the following, we primarily concentrate on radiometer and scatterometer observations. SAR sensing is considered in more detail elsewhere, for example, Henderson and Lewis (1998), Tsatsoulis and Kwok (1998).

Relevant Microwave Electromagnetic Properties of Ice and Snow

Passive microwave sensors observe the radiometric brightness temperature of the surface, which is a function of the physical temperature and the dielectric constant, which is dependent on the wavelength, on the amount of snow and ice, as well as on the structure of the snow and ice (i.e. grain size, shape, etc). The dielectric constant is closely related to the permittivity. Active microwave sensors rely on measurements of the normalized radar backscatter, also termed the *normalized radar cross section* (NRCS) (Ulaby *et al.*, 1981). The backscatter is a function of the permittivity and the roughness of the surface as well as the subsurface structure, with the volume contribution depending on various characteristics of the volume snow and ice properties. The microwave signature of snow and ice are dependent on the backscatter and emission from the surface, emission volume scattering from within the snow and ice, and scattering from layers within the snow and ice.

Since there is a large emissivity difference between ice and open water, sea ice is easily distinguished in the microwave wavelength region depending on the proportion of water and ice in the sensor footprint. The reflectivity of sea ice depends upon the bulk dielectric properties, which is dependent on the distribution of brine, air, and solid salts within the ice. Scattering on the other hand is largely influenced by the surface roughness and the inhomogeneities within the ice, as well as the bulk dielectric constant. Knowledge of small-scale surface roughness across the footprint is essential for the correct interpretation of microwave data. In addition, air pockets, brine inclusions, and depth and type of snow cover in sea ice significantly change the permittivity and hence the backscatter observed by radar instruments.

Dry snow has a low relative permittivity (~1.5–2.5) that increases dramatically in the presence of moisture, resulting in a strong melt signature. Thus, snowmelt can be detected in the microwave wavelength region as liquid water inclusions form around and between snow grains (Colbeck, 1986). Since the permittivity of liquid water is approximately 40 times that of ice crystals (Tiuri *et al.*, 1984), even small amounts of liquid water will affect snow dielectrics. For instance, as moisture bonds to snow grains overlying sea ice, microwave emission is largely due to the snow, rather than the underlying ice (Onstott *et al.*, 1987), and closely resembles a blackbody (Eppler *et al.*, 1992). Microwave brightness temperatures increase rapidly at frequencies greater than 10 GHz during melt onset (Thomas *et al.*, 1985), with greater increases at higher frequencies. As seasonal melt continues, liquid water begins to flow freely around the snow grains (Colbeck, 1986), and microwave emissivities tend to approximate an infinitely thick, wet, snow pack (Onstott *et al.*, 1987). Once ponding and drainage begin, the surface becomes a complex mixture over typical sensor footprint scales (several to tens of km), and typically displays a range of brightness temperatures (Grenfell and Lohanick, 1985).

Rapid changes in passive microwave signatures have been used to detect snowmelt over sea ice (e.g. Anderson, 1997; Smith, 1998) and ice sheets (e.g. Mote and Anderson, 1995; Abdalati and Steffen, 1995; Fahnestock *et al.*, 2002). Meltwater, however, causes passive microwave sea-ice algorithms to significantly underestimate summer ice concentration (e.g. Fetterer and Untersteiner, 1998; Comiso and Kwok, 1996). Partington *et al.* (2003) compared the Arctic sea-ice concentrations from the NASA Team sea-ice algorithm and manual ice charts from the US National Ice Center (NIC) and reported average differences of 23% in early August.

On ice sheets, the freeze-up period marks the end of freshwater input into the ocean and an increase in surface albedo. Over land ice, the diurnal difference in microwave brightness temperatures increases sharply when near-surface air temperatures approach 0 °C (Forster and Baumgras, 2000). Although Winebrenner *et al.* (1994) used active microwave data to identify the continuous melt period, and Parkinson (1992, 1994) used passive microwave data to calculate the time period between melt onset and freeze-up, little work using passive microwave data to distinguish the continuous and advanced melt stages has been done to date (see Fahnestock *et al.*, 2002). Smith (1998) utilized a combination of 19 and 37 GHz vertical brightness temperatures to recognize the start of the freeze-up period. Comiso and Kwok (1996) also showed that freeze-up could be determined on the basis of latitudinal-averaged SSM/I data, but they did not attempt to systematically map spatial and temporal variations.

The change in radar backscatter due to melt as measured by radar sensors is very dramatic and thus is a very powerful tool for mapping melt extent over both sea ice and ice sheets (e.g. Long and Drinkwater, 1999; Wismann, 2000). Nghiem *et al.* (2001) used diurnal variations in backscatter to map melt extent. Figure 1 illustrates two NSCAT backscatter images before and after melt onset.

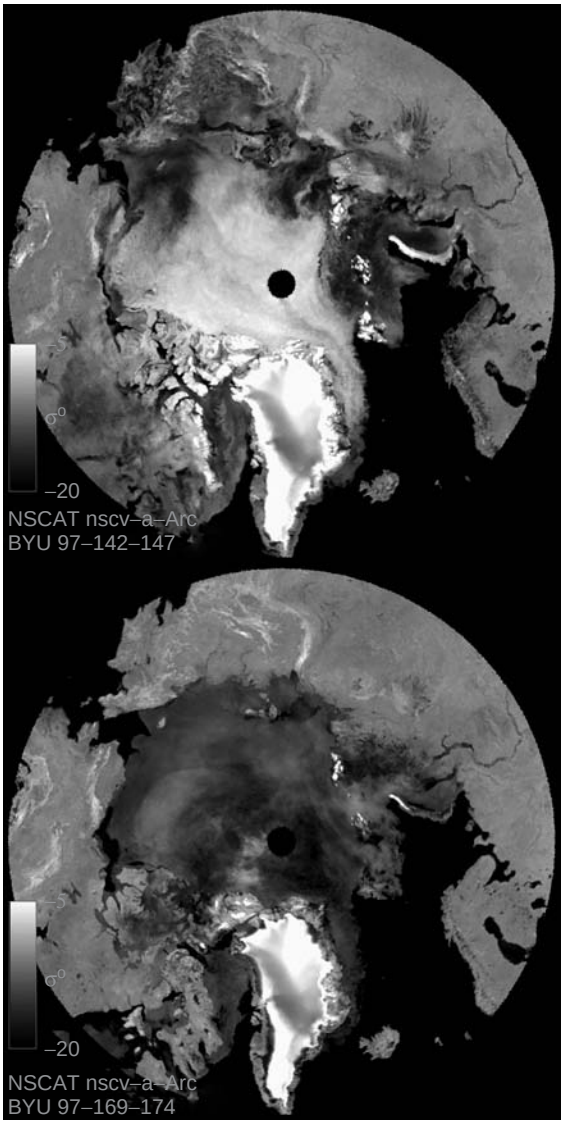


Figure 1 Two NSCAT enhanced-resolution backscatter images. The 13.9 GHz normalized radar cross section at 40° incidence angle (in dB) is shown before (a) and after (b) a major melt event. Note the dramatic change in backscatter over the sea ice and along the periphery of the Greenland ice sheet (from the Scatterometer Climate Record Pathfinder **http://www.scp.byu.edu**). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Electromagnetic Scattering and Emission Models

The relationship between the EM signature of ice and its geophysical characteristics is complicated. While much progress has been made in the methods for estimating geophysical parameters from microwave observations, more remains to be done. A common approach is to use inverse modeling techniques. In inverse modeling, a forward EM model is formulated, which predicts the microwave response as a function of the geophysical properties of the ice. Then, given a measured response, we try to infer or estimate the geophysical parameters that give rise to (or explain) the observed response (see, for example, Remund *et al.*, 2000).

While techniques for modeling the interaction of electromagnetic signals with natural surfaces have advanced significantly, the highly inhomogeneous nature of sea and glacial ice on a variety of spatial scales complicate modeling of snow and ice. Further, multiple combinations of geophysical parameters can give rise to similar microwave responses. An extensive review of ice emission and scattering models and inverse techniques is outside the scope of this article; however, we consider a number of important models (see also, Ulaby *et al.*, 1981).

Early modeling results have been based on simplified geophysics and electromagnetic interaction. As models have developed, the geophysics they incorporate have become more complicated. Drinkwater (1989) and Livingstone and Drinkwater (1991) used simple EM scattering models based on simplified volume backscatter from individual ice crystals, air bubbles and brine pockets, and rough surface scattering. They compared their modeling results with actual observations from both spaceborne and shipborne instruments. By incorporating snow and ice layering, this model has been extended to glacial ice in Greenland (Long and Drinkwater, 1994). However, no existing models adequately incorporate temporal evolution of the ice.

Several models have been developed to compute microwave emission for sea ice, which can be useful to help understand the evolution of passive microwave signatures of the ice pack. One method for the calculation of sea-ice reflectivity is performed using the "many-layer-strong fluctuation theory" approach of Stogryn (1987). This model treats interference between waves reflected and transmitted through the various layer planar interfaces coherently. Microwave emissivities are computed on the basis of Kirchoff's law, which relates emissivity to reflectivity. The model describes the emissivity of isothermal layers of snow and ice. For each type of surface (snow or ice), different parameters are used to describe its structure and constituents. The layer structure is used to calculate the dielectric properties and the emissivity of the sea ice. This modeling approach has been found to work well for sea ice without snow cover. For example, Figure 2 compares

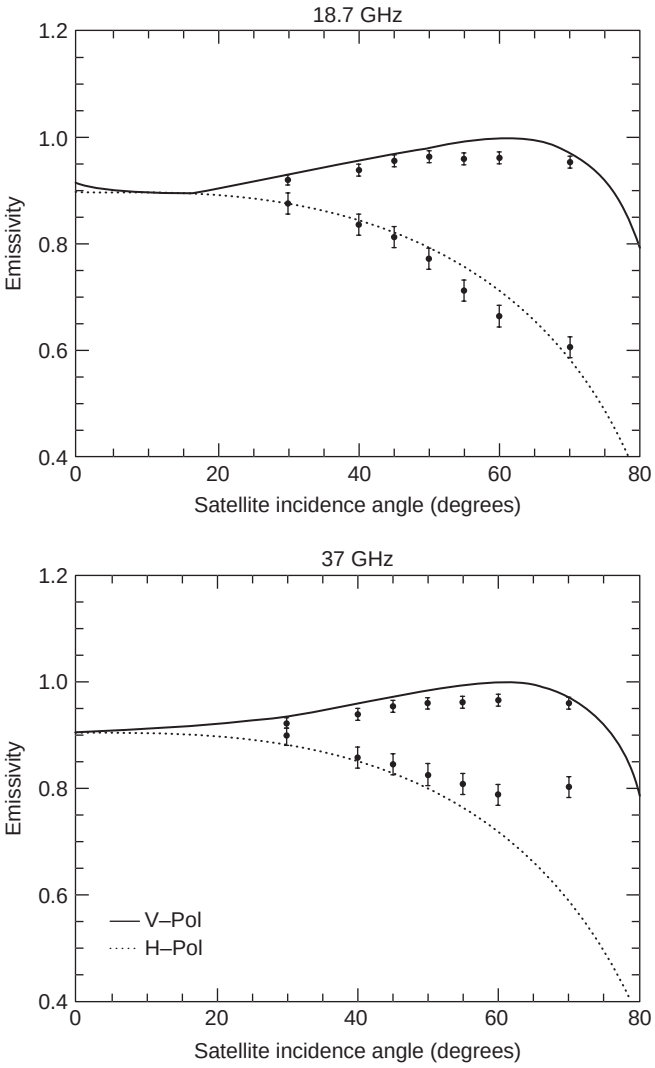


Figure 2 Comparison between modeled and measured emissivities as a function of incidence angle at 18.7 and 37 GHz for a thin gray ice sheet without snow cover

modeled emissivities with observations for a thin (4 cm) gray ice sheet without snow cover using the many-layer-strong fluctuation theory to model the microwave emissivity at 18.7 and 37 GHz. A model specifically designed for the emission of snowpacks has been developed by Wiesmann and Maetzler (1999) and is currently widely used to model the microwave emission of snowpacks.

Nghiem *et al.* (1995) introduced a composite polarimetric model for sea-ice scattering. This model is designed to calculate the effective sea-ice permittivity and the backscattering covariance matrices to compute the polarimetric signature. While emphasizing polarimetric modeling, they provide a good description of the general process of EM modeling of sea ice, including the effects of the ocean/ice interface, surface roughness, temperature, brine inclusions, air bubbles, and solid salt. Frost flowers and dynamic ice

effects such as ridging need further consideration as does the spatial variability of these phenomena relative to the sensor footprint.

A disadvantage of the geophysical modeling approach discussed so far is the significant amount of geophysical information required to compute the EM scattering model. The large number of parameters makes inverting such a comprehensive model difficult, particularly for large footprint sensors such as scatterometers (although the problem is not much simplified for higher-resolution sensors such as SAR.). Complicating this is the (often) wide spatial variation in parameters over the microwave sensor footprint, termed *footprint inhomogeneity*. Each forward model essentially estimates the radiances or backscatter at a single point defined by the set of input conditions. Satellites observe mixtures of such "pure pixels". Relating the satellite measurement to the ice state in practice requires some understanding of how the combinations of various surface types contributes to the set of observations recorded by the satellite sensors. A key component of forward modeling work will therefore require investigation of the contribution of individual snow and ice characteristics to area-averaged sets of radiances comparable to the footprint size of different sensors.

An alternate approach is to reduce the number of model parameters using simplifying assumptions. Swift (1999) uses a very simple model for ice scattering based on just three key parameters based on a bulk parameterization of the backscatter response. While overly simple, Remund *et al.* (2000) successfully used this model to classify ice types in the seas around Antarctica. Similar approaches can be applied to EM modeling of glacial ice. For example, Ashcraft and Long (2005) used a simplified scattering model to study wind-induced roughness on the Greenland ice sheet, which has been remarkably successful.

Application of EM Models to Modeling Ice Sheet Facies

Radar can detect the bare ice zone, the firn line (which does not necessarily coincide with the equilibrium line), the percolation zone, the wet snow zone, and the dry-snow zone. Passive microwave instruments are more limited spatially and can only detect broad changes, primarily the differences between regions of surface melt and dry snow in part because of their significantly coarser resolution. Because of the significant differences between the ice conditions in each facie, different EM modeling approaches have been used traditionally.

Dry-Snow Zone The EM response in the dry-snow zone contains both a surface and volume scattering component. Absorption and scattering losses within the firn are generally small for fine-grained firn and low frequencies (Mätzler and Hüppi, 1989); however, as the ratio of the grain size to wavelength increases, the scattering loss can become significant. While actual firn particles are irregular, they are often modeled as spheres or ellipsoids (Ulaby *et al.*, 1981). Bingham and Drinkwater (2000) propose multilayer scattering and emission models that include the scattering loss.

Until recently, on the basis of the high transmissivity of dry-snow surfaces at microwave frequencies, the air–snow interface roughness has been considered to have only a limited impact on the observed microwave signatures, particularly in the dry-snow zone. Thus, models have primarily addressed the volume scattering from firn crystals (e.g. Long and Drinkwater, 1994). However, Long and Drinkwater (2000) and Ashcraft and Long (2005) showed that an observed anisotropic scattering due to aligned roughness elements in Antarctica and Greenland is caused by the wind-driven roughness in the bedding structure and the annual layering of accumulated and wind-redistributed snow, suggesting links between the firn layer characteristics, wind-induced bedding structures, and the thermodynamic processes linking these physical attributes to the atmospheric conditions on the ice sheet. Understanding this linkage could provide insight into the evolution of the major ice sheets and improve microwave monitoring capabilities; thus, future modeling activities must include this effect.

Percolation Zone Swift *et al.* (1985) considered volume scattering mechanisms in simulations of the backscatter from the percolation zone by using an unbounded halfspace of scatterers. However, it is not necessary to invoke a strong volume scattering layer to simulate the bright levels of backscatter observed at Ku-band. Jezek *et al.* (1993) suggest that dominant stratigraphic features regulate short-wavelength percolation zone scattering. Longer wavelengths such as L-band (21-cm wavelength) penetrate deeper, possibly penetrating the multiple layers found in this glacial regime (Echelmeyer *et al.*, 1992). Such signals are scattered off the ice glands and lenses found up to a few meters beneath the surface. X-band (3-cm wavelength) measurements by Forster *et al.* (1991) indicate that the strongest returns come from annual layers beneath the surface that are shallower than 3 m. Jezek and Gogineni (1992) found that the largest proportion of backscattered returns occurs from an ice layer just beneath the annual snow accumulation. Long and Drinkwater (1994) formulated a simple 2-layer ice/snow model where the primary response from the upper layer is formulated in terms of a nonstratified Rayleigh layer upon an ice horizon of finite thickness, based on Mader (1991) and Jezek *et al.* (1993).

Geophysical Inversions: Sea-ice Concentration and Extent

The goal in this section is to present a few methods to derive the fraction of ice cover and extent from both passive

and active microwave remote sensing. It is important to remember that several other algorithms exist besides the ones described here. Future sea-ice retrieval algorithms will likely focus on blending passive and active methods for more accurate estimates of sea-ice cover.

Passive

Since passive microwave emission is relatively unaffected by the frequent and extensive cloud cover in the polar regions, one of the most important parameters provided by passive microwave data is sea-ice concentration. Sea-ice concentration maps are used to track ice edges, estimate ice extent, ice type, actual ice area, and the amount of open water within the ice pack for both climatological and more pragmatic (i.e. shipping) purposes. The latter is in turn needed to monitor occurrence, impact, and persistence of polynyas, to calculate heat and salinity fluxes between the ocean and the atmosphere in the polar regions, in addition to many other applications.

Several techniques have been developed to obtain ice concentration from passive microwave data (Markus and Cavalieri, 2000; Svendsen *et al.*, 1983; Cavalieri *et al.*, 1984; Swift *et al.*, 1985; Comiso, 1986; Smith, 1996). Steffen *et al.* (1992) and Smith (1996) provide a review of some of these techniques and a comparison of ice concentration from some of the most commonly used sea-ice algorithms. Ice concentration differences between algorithms result from the use of different sets of channels in the algorithms: different responses to changes in atmospheric conditions, surface temperatures, and emissivities; different algorithm tie points (reference brightness temperatures); and from the ways in which the tie points are selected (Smith, 1996). Radiative transfer (RT) modeling suggests that differences may also result from the ways in which the algorithms respond to variations in snowpack conditions such as hoar layers, snow depth/density, ice lenses, flooding at the snow/ice interface, and so on.

Two widely used sea-ice algorithms are the NASA Team (Cavalieri *et al.*, 1984) and the Bootstrap (Comiso, 1986) algorithm. The NASA Team algorithm is based on two ratios, the polarization (PR) and the spectral gradient ratios (GR) defined by:

$$\mathrm{PR} = \frac{[\mathrm{TB}(19\ \mathrm{V}) - \mathrm{TB}(19\ \mathrm{H})]}{[\mathrm{TB}(19\ \mathrm{V}) + \mathrm{TB}(19\ \mathrm{H})]} \quad (1)$$

$$\mathrm{GR} = \frac{[\mathrm{TB}(37\ \mathrm{V}) - \mathrm{TB}(19\ \mathrm{V})]}{[\mathrm{TB}(37\ \mathrm{V}) + \mathrm{TB}(19\ \mathrm{V})]} \quad (2)$$

where TB is the observed brightness temperature at the indicated frequency and polarization. From these two parameters, the first-year ice concentration (CF) and the multiyear ice concentration (CM) are derived, and thetotal ice concentration (CT) is computed as the sum of CF and CM.

The Bootstrap algorithm uses two microwave channels (37 H and 37 V) and selects the slope and offset of the consolidated ice line in the scatter plots of these two channels. An inherent assumption in the algorithm is that there are large regions in the Central Arctic during winter with ice concentrations of 100%. The 37 H versus 37 V scatter plot shows very high correlation for consolidated ice. However, in the seasonal sea-ice regions where first-year ice dominates, as in the waters surrounding the Antarctic, the sole use of the 37 H and 37 V channels does not provide consistent identification of consolidated ice data points. Sometimes, consolidated ice is difficult to distinguish from mixtures of ice and water because of the presence of snow cover, flood, or roughness effects. Under these conditions, the 19 V and 37 H TBs and the Bootstrap algorithm are considerably more consistent and accurate. These channels are thus used primarily in the seasonal ice region, including the Antarctic region.

Comiso *et al.* (1997) compared the NASA Team and the Bootstrap sea-ice algorithms. Results from comparison studies such as this one show significant qualitative and quantitative differences between the NASA Team and Bootstrap algorithms. For example, differences in monthly means as large as 30% were observed in the Southern Hemisphere. Changes were made to the Bootstrap algorithm in order to help reduce the large differences noticed in the Southern Hemisphere. However, large differences also occurred in the Northern Hemisphere. For example, Figure 3 shows a comparison of ice concentrations on September 14, 1998 near the SHEBA ice camp. The image in Figure 3(a) shows ice concentrations from the Bootstrap algorithm, and ice concentrations in (b) are from the NASA Team algorithm. Differences as large as 40% in ice concentrations are observed between the two algorithms. Results from the enhanced NASA Team algorithm (Markus and Cavalieri, 2000), which is used to derive Northern Hemisphere sea-ice concentrations from the AMSR instrument

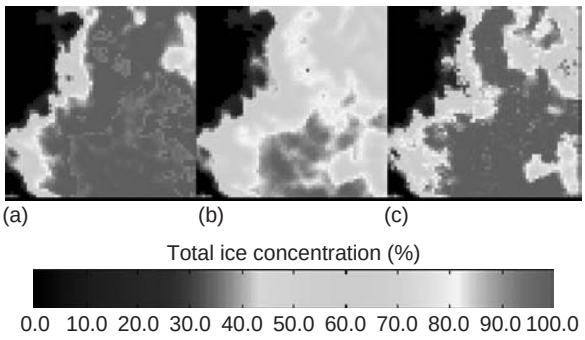


Figure 3 Bootstrap (a), NASA Team (b) and enhanced NASA Team (c) sea-ice concentrations on 14 September, 1998 near the SHEBA ice camp. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

are shown in Figure 3(c). This algorithm appears to correct some of the underestimation of the fraction of ice cover in the NASA team algorithm, but areas of large differences between the Bootstrap ice concentrations are still observed. Time series of ice concentrations near the SHEBA ice camp from September 1997 to September 1998 (not shown) also show large differences among the three algorithms. Many of the differences observed have a stable quality that do not appear to be related to atmospheric effects, but rather as a result of differences in how the algorithms respond to the sea-ice surface characteristics.

Comparative studies with optical satellite imagery, such as Landsat, have not yielded conclusive evaluation of the merit of each technique. However, current work with SAR data may provide new insights about the discrepancies. Some of the observed large discrepancies have led to improvements in the Bootstrap sea-ice algorithm (Comiso *et al.*, 1997), where the Bootstrap algorithm tie points have been adjusted to match ice concentrations derived from SAR during the melt period. However, even though ice concentrations may differ substantially among all the various techniques, they all do tend to be consistent in finding the location of the ice edge.

Active

Sea-ice extent is readily identified with ERS-1/2 scatterometer (ESCAT), using a normalized measure of the isotropy of sea ice relative to ocean (Cavanié and Gohin, 1995; Early and Long, 1997), and with NSCAT and QSCAT, by comparing the response of VV (vertical transmit – vertical receive) and HH (horizontal transmit – horizontal receive) polarizations over ice and water (e.g. Remund and Long, 1999; Haarpainter *et al.*, 2004; Anderson and Long, 2005). Haarpainter *et al.* (2004) suggests that scatterometer polarization ratios can be used to infer sea-ice concentration. The dual-polarization capability of QuikSCAT data can make it more effective in monitoring the decay of sea ice than SAR data (Howell *et al.*, 2005).

The onset of seasonal surface melt and then freeze-up provides a significant contrast to winter, colder conditions over sea ice in the polar regions. Melt onset defines a significant transition in the radiative budget of the ice-covered region, while autumn freeze-up is a key period during which the residual fraction and characteristics of perennial sea ice can be assessed and the initial conditions for winter sea-ice dynamics is established via the distribution and orientation of newly forming ice in leads. Most recently, algorithmic approaches to use scatterometer data to determine melt onset and freeze-up have been examined, including those in Antarctica over a period of several years (Drinkwater and Liu, 2000; Forster *et al.*, 2001; Winebrenner *et al.*, 1998).

Observations and Results

In this section, we provide a brief summary of selected results from the literature. The goal is to provide a starting point for those pursuing research in the field.

Passive Observations of Sea Ice

Sea-ice time series derived from multichannel, passive microwave data are among the longest, continuous, satellite-derived, geophysical records. These climatological data sets extend across nearly three decades. The predominant variability in Arctic sea-ice time series is seasonal, with a typical late winter (March) maximum ice extent $\sim 15 \times 10^6\,\mathrm{km}^2$, compared to a late summer (September) minimum $\sim 6 \times 10^6$ in the Arctic, although the absolute values may vary from study to study because of operational differences and time period sampled (Figure 4). The satellite passive microwave sensor record is augmented by active sensors, though there is some seasonal sensitivity difference between the two sensors, see Figure 5 (Anderson and Long, 2005; Haarpainter *et al.*, 2004).

The first trend analysis based on SMMR data found a slight negative trend in the Arctic sea-ice extent from 1978 to 1987 (Gloersen and Campbell, 1991). The $3.2 \times 10^4\,\mathrm{km}^2\,\mathrm{year}^{-1}$ decrease (2.4% per decade) was found to be statistically significant. Data from the subsequent SSM/I sensors has provided the basis to follow up the SMMR trends. The Johannessen *et al.* (1995) analysis of SMMR and SSM/I records taken separately revealed a greater reduction in the Arctic sea-ice area and extent during the SSM/I period – decreases from 1987 to 1994 were ~4% per decade compared to ~2.5% per decade from 1978 to 1987. However, large interannual variability coupled with the brevity of the individual SMMR and SSM/I records compelled researchers to produce longer time series for more robust trend estimation. Merged SMMR–SSM/I time series have since been produced and analyzed, establishing the trends more robustly (Bjørgo *et al.*, 1997; Cavalieri *et al.*, 1997). The merging of SMMR and SSM/I data involves intercomparison and adjustments (i.e. "intercalibration") based on the 6-week overlap period in 1987, when both types of sensor operated. Two independent analyses of merged SMMR–SSM/I data established the trend in the Arctic ice area and extent (1978–1995) to be about $-3.0 \times 10^5\,\mathrm{km}^2$ per decade, corresponding to ~3% per decade (Bjørgo *et al.*, 1997; Cavalieri *et al.*, 1997; Comiso, 2000; Comiso, 2003).

Parkinson *et al.* (1999) identified the seasonal and geographic patterns of variability and trends in the ice concentration data, 1979–1997, finding winter reductions to be concentrated in the Barents and Greenland seas, with summer reductions more pronounced in the Siberian seas. Maslanik *et al.* (1996) investigated reductions in the Arctic ice cover using SMMR–SSM/I data (1979–1995), with ice reductions in the 1990s found to be most pronounced in

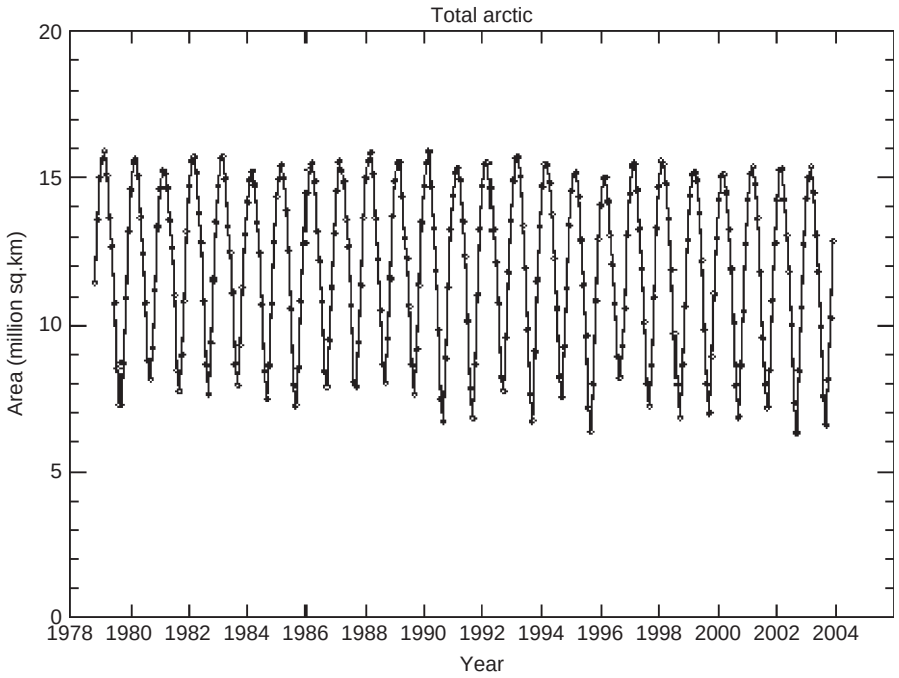


Figure 4 Monthly ice extent for the entire Northern Hemisphere from the NASA Team sea-ice algorithm. Data and figure available from **`http://nsidc.org`**

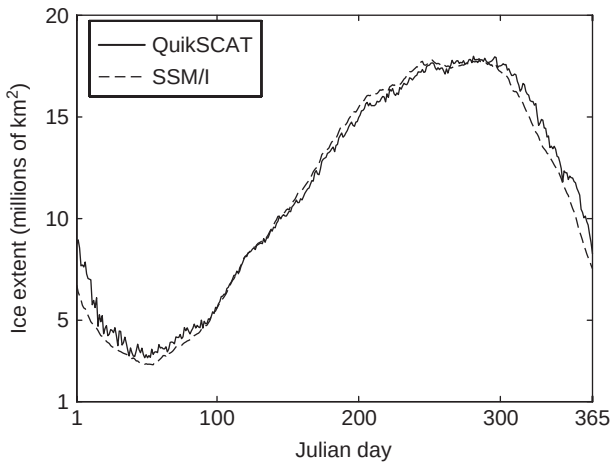


Figure 5 Seasonal Antarctic sea-ice extent in square kilometers derived from active (NSCAT) and passive (SSM/I) data for Antarctica. The differing sensitivities of the active and passive sensing to seasonal melt are apparent (Reproduced from Remund and Long, 1999 by permission of the American Geophysical Union)

the Siberian sector in summer, with record low Arctic ice minima in 1990, 1993, and 1995. Summer 2002 set the 25-year record for minimum ice extent and area (Serreze *et al.*, 2003), with essentially no ice in the Greenland Sea. The situation repeated itself again in 2003 and 2004, strongly reinforcing the downward trend in the Arctic ice cover (Stroeve *et al.*, 2005). Figure 6 shows the time series of monthly ice extent anomalies for the Arctic from November 1978 through September 2002, highlighting the record low ice concentrations observed in September 2002. Although early summer 2003 did not appear to be unusual in terms of ice extent, the September 2003 ice extent nearly reached the 2002 record low, again with no ice in the Greenland Sea. The September trend is $(-0.54 \pm 0.11) \times 10^6\,\mathrm{km}^2$ per decade at a 99% confidence interval (or $-7.7\% \pm 3\%$ per decade), when data through September 2004 is included (Stroeve *et al.*, 2005). These dramatic declines have led to predictions of a seasonally ice-free Arctic Ocean by $\sim$2050.

Antarctic sea ice tells a different story, however, during this time frame. Satellite records of sea ice around Antarctica indicate that the Southern Hemisphere ice cover has increased slightly since the late 1970s, while at the same time that the Arctic sea ice has declined (Figure 7). The upward trend in the Antarctic sea-ice cover is not statistically significant and conflicts with photographic ice core estimates that suggest that the sea-ice cover has receded by about 45% since 1951 (Wolff, 2003). However, detection of long-term change is masked by decadal-scale variability and these decadal fluctuations have produced the apparent short-term increases in ice extent during the satellite era.

Passive Observations of Greenland Melt

Combined, the Greenland and Antarctic ice sheets contain enough water to raise the global mean sea level by

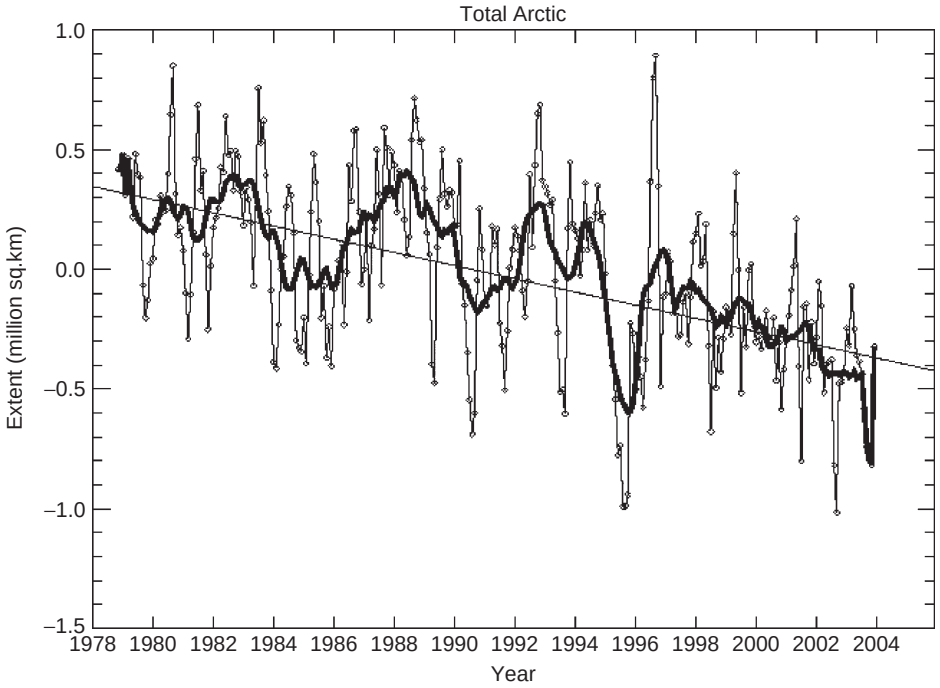


Figure 6 Time series of total ice extent anomalies (departures from the mean) for the Northern Hemisphere from November 1978 through December 2003. A 12-month running mean is included. Image available from the National Snow and Ice Data Center (NSIDC)

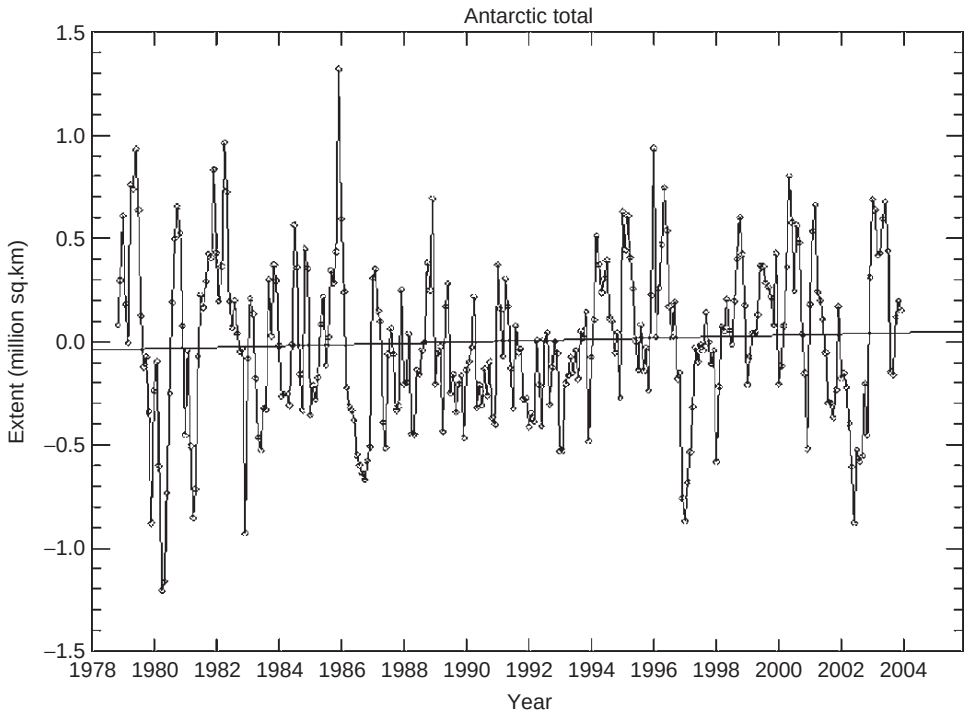


Figure 7 Time series of total ice extent anomalies (departures from the mean) for the Southern Hemisphere from November 1978 through December 2003. Image available from the (NSIDC) (Reprinted with permission from Comiso, J. C. and C. L. Parkinson, Satellite observed changes in the Arctic, Phys. Today 57(8), 38–44, 2004. ©2004 American Institute of Physics)

approximately 70 m, with the Greenland ice sheet comprising around 7 m of that estimate (Warrick *et al.*, 1996). Given detectable ice-sheet elevation changes in Greenland (Krabill *et al.*, 2000, and updates), this area warrants continued monitoring. Microwave remote sensing offers a key tool that can assess broad areas rapidly, for example, melt extent from passive microwave sensors shows an overall increase in total melt area of the Greenland ice sheet since 1978 (Abdalati and Steffen, 2001). Figure 8 shows the interannual variations in mean melt extent from 1979 through 2003. Only the average melt area for the months of June, July, and August of each year are shown. Results show an increase in melt area of 0.62% per year since 1979.

In 2002, the spatial extent of melt on the ice sheet reached a new record, surpassing the previous record spatial melt extent by more than 9% (K. Steffen, personal communication). This is not reflected in Figure 8, however, since only the months of June through August were used to compute the mean melt. In 2002, melt was observed to begin earlier than it had over the past 24 years and extended up to higher elevations, resulting in an overall greater area of melt during 2002, but not necessarily more melt volume. The decline in 1991 Greenland melt extent is thought to be associated with cooling across the region associated with the Mt. Pinatubo eruption that sent aerosols into the atmosphere, which reflected solar radiation and resulted in cooler temperatures and less melt; in other words, a short-term "anomaly" in the upward melt trend (Shuman *et al.*, 2001; Abdalati and Steffen, 1997a).

Active Observations of Glacial Ice, Antarctica

A variety of techniques for observing melt extent have been developed (see bibliography), which are compared in Ashcraft and Long (2004b) for both active and passive measurements. Initial results are encouraging, but further work is required to validate the technique.

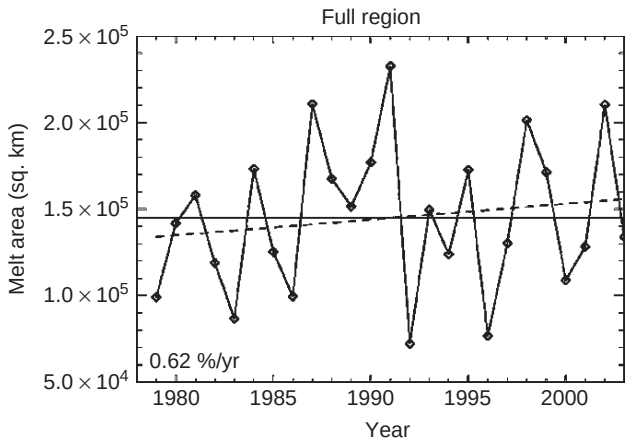


Figure 8 Interannual variations in mean melt extent (average area for June, July, and August). Statistically significant slope is 0.62%/year. Image provided by W. Abdalati from NASA GSFC

The recent collapse of the Larsen-A ice shelf has been conclusively linked to local warming (Scambos *et al.*, 2000, 2003). Surface warming causes additional meltwater formation which saturates the permeable firn and which then fills crevasses in the ice shelves where it can refreeze. Additionally, this flowing meltwater can also enlarge or propagate these large cracks through the thickness of the ice shelf that may be thinning by submarine melting as well. By using microwave sensing, melt on ice shelves can be monitored and shelves of highest risk can be identified. Resolution plays a key role in the effectiveness of this approach. As a demonstration, Fahnestock *et al.*, 2002 and Kunz and Long (2005) have used SSM/I and QuikSCAT data to observe melt onset and duration over a multiyear period. Their results suggest that microwave sensors can be effective for such monitoring.

SAR Applications

Owing to their high resolution, SARs have been used extensively for the study of ice sheets and glaciers. Because of space limitations, we cannot do due justice to the many SAR applications. We are limited to a few introductory references. Some examples include: SAR data has been used to map sea-ice thickness (Kwok *et al.*, 1999b), studies of the Greenland and the Antarctic ice sheet and associated glaciers (Jezek *et al.*, 1993; Fahnestock and Bindschadler, 1993; Fahnestock *et al.*, 1993) such as mapping of ice-sheet topography (Jezek *et al.*, 2000), glacier motion (Kwok and Fahnestock, 1996; Rignot, 1996; Young and Hyland, 2002), estimates of snow accumulation (Munk *et al.*, 2003), and melt onset dates (Kwok *et al.*, 2003), among many, many others! An introductory review of SAR and its applications to ice observation is contained in Henderson and Lewis (1998) with a more extensive summary of SAR applications in the polar regions given by Tsatsoulis and Kwok (1998).

Optical Observation of Ice Sheets and Sea Ice

In this section, we also briefly consider optical sensors, including laser altimetry.

Like microwave sensors, optical sensors can be divided into passive and active sensors. Active optical sensors include laser altimeters such as the ICESat Geoscience Laster Altimeter System (GLAS) (Zwally *et al.*, 2002). Designed for precision ice-sheet elevation measurement, this sensor uses infrared and green lasers to transmit light energy to the surface and to measure the time of flight of the return echo. This is converted to an elevation value based on the orbit trajectory. Passive optical sensors include conventional cameras and precision optical radiometers. The latter can be used for measuring surface skin temperature and surface albedo, though only under cloud-free conditions. Optical ice sensors operate in the visible, near-infrared, and

infrared portions of the electromagnetic spectrum. Here we discuss optical observations of the Earth's polar regions at various frequencies. We first consider passive visible and near-infrared, then infrared observations. Finally, a brief summary of active optical sensing is provided.

Visible and Near-infrared Imaging

The ability of satellite-based visible and near-infrared sensors (VNIR; making measurements in the 400–1600 nm range) to rapidly acquire data over remote regions has led to wide applications in mapping and in the measurement of polar ice. Beginning in 1972, with the launch of the first Landsat (and even sooner in the world of intelligence satellite mapping such as CORONA), images of Antarctica and Greenland were systematically acquired, largely to provide basic mapping data on coastlines, mountain ranges, and ice flow. Over time, with a larger range of sensors available and with a record of past images to review and compare, the use of VNIR images has expanded to include tracking movement and detection of changes, such as velocity mapping, melt-extent mapping, snow cover, and grain-size changes, surface albedo, fracturing and calving, and a host of other processes.

In an outstanding series of volumes, the US Geological Survey has demonstrated the remarkable range of information that may be obtained about glaciers and ice sheets simply by looking at their images (Williams and Ferrigno, 1988–2002). Inmaps produced by US, European, and Australian groups from both Landsat and SPOT (Systeme Probatoire pour l'Observation de la Terre), VNIR sensors revealed the outline and structure of several large ice systems for the first time (e.g. Swithinbank *et al.*, 1988); such work is still important even in the new century in some areas (e.g. Skvarca and De Angelis, 2002). But a further realization of the potential of satellite images came in the late 1980s and early 1990s, when computer technology and data storage systems reached the point where large images could be readily reprocessed and enhanced digitally. Suddenly, a wealth of detail was revealed in the interior of ice sheets, as subtle topographic features almost invisible in the photographic renditions of the satellite data became enhanced by simple digital techniques such as high pass filtering (e.g. Orheim and Lucchitta, 1987; Bindschadler and Vornberger, 1990). The value of vast amounts of weather-satellite image data in the visible and near-infrared range also came to be appreciated, particularly for those sensors with greater radiometric sensitivity (e.g. the NOAA Advanced Very High Resolution Radiometer (AVHRR), see following section on infrared temperature mapping; Comiso, 2000; Hastings and Emery, 1992). This led to the first image maps of the great ice sheets in which the interior topographic detail was shown (e.g. USGS, 1991), rather than the "blank white" areas of earlier maps (AGS, 1971). Since then, VNIR digital image mapping of ice has advanced rapidly, permitting seamless images of both high resolution and wide spatial coverage (e.g. the Moderate Resolution Imaging Spectroradiometer (MODIS) Mosaic of Antarctica (MOA), Bohlander *et al.*, 2004).

As researchers realized the topographic detail present in VNIR satellite images through improved digital processing, they began to strive to include that detail quantitatively in topographic models of the ice via "shape-from-shading" or photoclinometry. In a few cases, multi-image data "stacking" or "data cumulation" has been used to push spatial and spectral resolution of subtle ice features for specific areas beyond what a single image from a given sensor can achieve (Fahnestock *et al.*, 2000). In addition to mapping the extent and surface shape of glacial and sea ice, researchers have also sought to use satellite VNIR images to map important surface properties of snow and ice. Chief among these are albedo, grain size, and surface roughness or backscatter. In these cases, particularly, good radiometry and spectral resolution is favored over spatial detail. Most studies to date have used moderate-spatial-resolution VNIR imagers such as AVHRR, MODIS, and the Multiangle Imaging SpectroRadiometer (MISR) (Stroeve *et al.*, 1997; Nolin *et al.*, 2002), with a few exceptions (Landsat; Bourdelles and Fily, 1993).

Investigators have pushed past the basic objectives of mapping and towards detecting and mapping *changes* in the ice, thus offering the hope of connecting these mapped changes to dynamic or climate causes. Change detection in the coastlines is currently being addressed in a series of new maps from the USGS (Ferrigno *et al.*, 1994), and for glaciers, in large satellite projects designed to monitor the ice-covered regions throughout the planet(the GLIMS project, Bishop *et al.*, 2004). More broadly, recent efforts have assembled long time series of VNIR data over both poles, suitable for a host of change detection and seasonal- or interannual process studies (the Polar Pathfinder AVHRR data set; Maslanik *et al.*, 1997). Huge collections of Antarctic and Greenland Landsat 7 data now exist for the ice-sheet areas, archived by USGS (**`http://edcsns17.cr.usgs.gov/EarthExplorer`**), and MODIS data from both Terra and Aqua are also readily available (**`http://acdisx.gsfc.nasa.gov/data`**).

A major advance in the late 1980s and early 1990s was the development of methods to map ice flow velocity from satellite (and for sea ice, ice drift velocity). Initially, these used standard manual photogrammetric techniques (Lucchitta and Ferguson, 1986), but evolved to use digital correlation to track movement (Bindschadler and Scambos, 1991; Scambos *et al.*, 1992; Emery *et al.*, 1991). As the time series of Landsat, SPOT, and ASTER images has grown, several studies now report changes in glacier ice velocity on the basis of VNIR image pair sequences (e.g. Stearns and Hamilton, 2003; Berthier *et al.*, 2004; Scambos *et al.*, 2004).

Satellite VNIR data will continue to be used to track and monitor the evolution of the polar regions. Further details on recent advances in visible and near-infrared image mapping of glaciers, snow, and sea ice have been published in reviews by Massom (1995), Bindschadler (1998), and Konig (2001).

INFRARED ICE TEMPERATURE OBSERVATION

Surface temperatures have been recorded in parts of the polar regions for as long as a century. The locations of these records are, however, sparsely distributed and do not provide good representation of the large-scale distribution of surface temperatures over ice-covered areas. Because of the vast expanse and general inaccessibility of ice sheets and sea ice, the only practical way to produce spatially detailed maps of surface temperature is through the use of satellite remote sensing. Among the currently proven techniques are those which make use of thermal infrared channels. Thermal infrared is quite useful for this purpose because the emissivity of snow and ice surfaces is uniform and close to unity (Comiso, 1994). The use of passive (or active) microwave systems has been cited as alternatives, (Shuman and Comiso, 2002) but while the microwave systems have some advantages, namely, their ability to penetrate thick clouds, the application is limited because they require the knowledge of the emissivity (or backscatter) of the surface almost on a pixel-by-pixel basis owing to large spatial variability of these parameters at the microwave frequencies. One of the main issues with the use of infrared sensors is that they can provide the desired surface temperatures only during cloud-free conditions. An effective cloud masking technique is thus required, which sometimes is not easy to implement because signatures of clouds can be difficult to distinguish from those of snow and ice-covered areas. While clouds are persistent at times, the frequency of observation by polar orbiting satellite sensors over the Earth's ice-covered regions is very high and with clouds always moving, the chance of getting surface measurements for a particular area is relatively high. Weekly and monthly averages of the data have been generated with practically no gaps over ice-covered areas. Because of the "clear-sky" sampling bias, the resulting maps are regarded as surface temperature maps for these conditions only. However, comparative studies using continuous and cloud-free data recorded by meteorological stations indicate that the biases in the monthly average data are typically small, ranging from 0.5 to 1 K.

The first system that has been used for large-scale studies of surface temperature in the polar region was the Temperature Humidity Infrared Radiometer (THIR) which was onboard the Nimbus-4, Nimbus-5, Nimbus-6, and Nimbus-7 research satellites. The Nimbus-7 system was different from those earlier in the series in that data were digitized on board the spacecraft. Thus, the Nimbus-7 THIR was the only system that provided consistently good quality data that could be used for polar studies (Comiso, 1994). The THIR sensor was a two channel system (6.7 and 11.5 μm) with a resolution of 6.7 km and 20 km for the 11.5 μm and 6.7 μm channel, respectively, and provided reasonably good data from 1978 to 1985. The system, however, did not provide enough channels to allow for accurate cloud masking and atmospheric corrections. The advent of the NOAA AVHRR enabled a significant improvement in the accuracy of the retrieved surface temperature data in terms of spatial and radiometric resolution and the ability to resolve and mask clouds. The sensor has five channels from the visible to infrared wavelength, two of which are window channels in the thermal infrared region. The basic resolution is 1 km at nadir but because of very large storage requirements (considering mass storage technology limitations in the 1980s), only a 5 by 4 km subset, called *GAC* (Global Area Coverage), is automatically stored and archived for global and continuous coverage, while the 1 km data set, called *LAC* (Local Area Coverage), is available only on a limited basis. It is the GAC data that has been used mainly for seasonal and interannual variability studies (e.g. Comiso and Parkinson, 2004). Other systems (e.g. ENVISAT/AATSR and EOS/MODIS) provide improved capabilities and more accurate retrievals, but these data sets began in the late 1990s.

All these infrared sensors measure the temperature of the top layer (usually referred to as *skin depth*) of the ice surface. Comparative analysis of surface temperatures derived from AVHRR data with co-registered and coincident *in situ* measurements confirms that the thermal infrared data captures the same variability in surface temperature as the ground values (Comiso, 2000; Comiso and Parkinson, 2004). Errors associated with the retrieval of surface temperatures have been evaluated and estimated to range from 2 to 3 K (Steffen *et al.*, 1993; Key and Haefliger, 1993; Comiso, 2000; Shuman and Comiso, 2002). The main concern associated with the development and analysis of a long-term temperature record is the relatively short lifespan (typically 5 years) of each AVHRR sensor. The archived composite record is actually a collection of data from several AVHRR sensors with no overlap coverage between sensors. The lack of overlap has made it difficult to compare the performance and compatibility of the different AVHRRs. For lack of a better solution, *in situ* data have been used to improve calibration and ensure consistencies in the derived temperatures from the different sensors.

The AVHRR sensor series data have been used to generate monthly average temperature maps of high latitude regions of the Northern and Southern Hemisphere from August 1981 to the present. The data is gridded at a

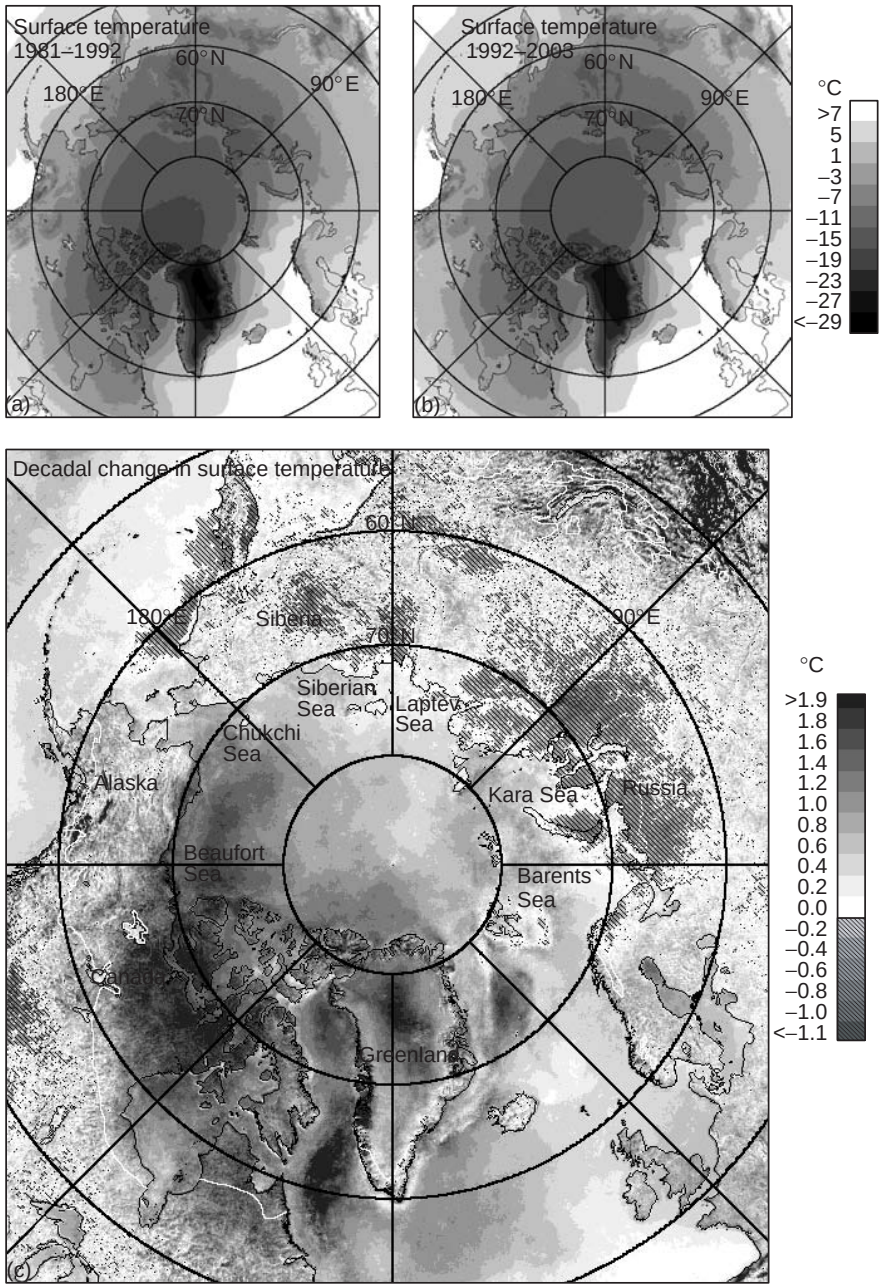


Figure 9 (a) average surface temperature from August 1981 to July 1992; (b) average surface temperature from August 1992 to July 2003; and (c) difference of the first set of data and the second set of data (from Comiso and Parkinson, 2004). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

resolution of about 6 km in polar stereographic format. Compared to other sources of regional surface temperature data (e.g. NCEP or ECMWF), the infrared data provide more realistic representations of the spatial distribution of surface temperatures in areas where there are no *in situ* measurements. For example, the retrieved data are coherent with the expected variations in surface temperature with elevation in Antarctica and in Greenland, whereas models reflect spatial temperature patterns poorly over the ice sheets. Some unique surface features over land and sea ice that are not in other data sources have also been identified.

To illustrate the value of the AVHRR data set, Figure 9(a) shows averages of surface temperature in the Arctic region from August 1981 to July 1992, while Figure 9(b) is a similar average, but for the period from August 1992 to July 2003. The two images appear to be almost identical except for slight shifts in the temperature contours. The

consistency of the images in reproducing distinct surface features is an indication that the cloud masking technique is basically effective, and the data represents mean surface temperatures over these time periods. The difference map shown in Figure 9(c) allows evaluation of decadal changes in the Arctic (i.e. from one 11-year period to the next). Positive changes are represented by warm (light gray shades) colors, while negative changes are represented in cool (black and dark gray shades) colors. The map indicates that the changes are overwhelmingly positive, indicating warming, with the most positive located in the western Arctic (e.g. Beaufort Sea) and North American region. An unexpected result is the detection of large areas of cooling in parts of Russia, which have subsequently been verified using *in situ* data. This indicates that global monitoring is important and that trend studies should not rely solely on measurements from ground stations especially if the latter are sparsely distributed. As a check on the difference map, a trend map has been generated using the same set of data, but with regressions done on a pixel-by-pixel basis. The result was very similar to the difference map shown in Figure 9(c). For the whole area, on the average, the warming trend over the entire Pan–Arctic region during the study period is estimated to be 0.5 K per decade. The maps thus provide the basis to assess the temperature state of different regions of the cryosphere and enable evaluation of vulnerabilities of these regions to melt, ice breakup, permafrost thawing, or iceberg calving. The temperature maps are also useful for specific polar process studies. For example, they have been used to verify that the temperature changes in the Antarctic Peninsula are anomalous and are closely associated with the region (King and Comiso, 2003). Studies of this type also show that the temperature variability is highly correlated with the southern oscillation as well as with the Southern Hemisphere Annular Mode (Kwok and Comiso, 2002; Schneider *et al.*, 2004). In short, these evolving data sets enable a much greater understanding of the spatial and temporal nature of climate variability across the Earth's sea ice and ice sheets.

ICESat (LASER ALTIMETRY)

As the first operational satellite laser altimeter in early 2003, NASA's Ice, Cloud, and land Elevation Satellite (ICESat) has provided precise surface elevation data from 86° N to 86° S latitude (Zwally *et al.*, 2002). The data have been acquired in a discrete series of observational periods and over two repeated track patterns using all three of ICESat's GLAS lasers. Surface elevation data is collected over most of Antarctica and all of Greenland as well as smaller ice caps and glaciers in accordance with the primary mission goal of improving ice-sheet elevation change and ice mass balance assessments (Scambos *et al.*, 2004; Thomas *et al.*, 2004). In addition, ICESat provides precise elevation data across most of the Arctic Ocean and the icy waters surrounding Antarctica, which can be converted into remote sensing–based estimates of sea-ice thickness (Kwok *et al.*, 2004). As the mission enters its third year, particular efforts are on to ensure that the most precise elevation data is derived from the ICESat laser information obtained to date. This section summarizes ICESat's utility for research across the Earth's polar regions.

The GLAS instrument on ICESat is a novel, "active" remote-sensing instrument. Its lasers operate continuously and emit pulses of 1064 and 532 nm (infrared and green light, respectively), at 40 Hz throughout each orbit. The light energy spreads as it travels from the spacecraft and reflects from approximately 70 m "spots" separated by about 170 m along track back to the GLAS detectors. The timing information, after adjustment for travel through the Earth's atmosphere, is then converted into an elevation value for each measurement location. This is based on precise knowledge of where ICESat/GLAS was when the pulse was sent and received as well as where the laser was pointing relative to the stellar background (Brenner *et al.*, 1999).

The resulting elevation values, although subject to refinement as pointing or orbital position information is improved, are capable of centimeter-level precision (Figure 10). This plot shows an ICESat track that has been repeated in late 2003 and early 2004 across Lake Vostok's relatively smooth surface in East Antarctica. ICESat's precision can be visually assessed by comparing the actual elevation profile values to an 11 point ($\sim$1.9 km) running mean derived from them. The differences between the actual and the smoothed elevation profiles has a standard deviation of less than 3 cm for both Laser 2a and Laser 2b over this portion of Lake Vostok. The absolute accuracy of ICESat is more difficult to quantify, given that the ice-sheet surface is not absolutely known and may change with time depending on snow accumulation or ablation as well as ICESat-specific factors such as pulse saturation that are not fully addressed at this time. In addition, ICESat profiles do not usually repeat the same coordinates exactly; therefore any cross track slope over the distance between the profiles, generally less than 100 m, may cause an apparent elevation difference. Despite these uncertainties, ICESat's accuracy is estimated at better than 20 cm for repeat track observation conditions like the ones shown here.

Elevation data of this precision and accuracy has never before been obtained across the polar regions from space. Although compromises have been made to the spatial and temporal density of data acquisition to accommodate laser problems, and reprocessing improvements continue, ICESat is proving that the risks of this novel mission are worth the return. ICESat data is defining a "baseline" shape of the ice sheets and detailing their rapidly changing margins (Scambos *et al.*, 2004; Thomas *et al.*, 2004) as well as

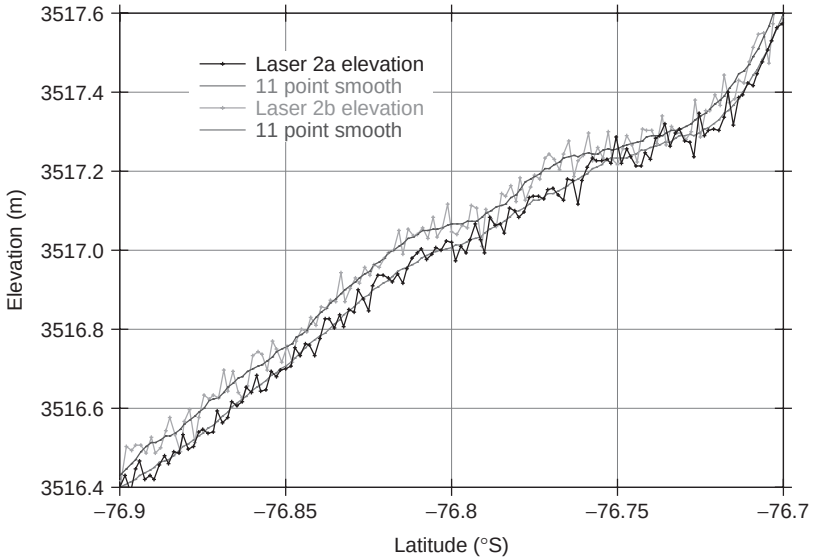


Figure 10 Sample ICESat repeat-track elevation plot over Lake Vostok, East Antarctica. The low slope and relatively low accumulation across this area makes it a useful area for assessment of ICESat precision and accuracy. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

providing extensive data on other ice-sheet phenomena such as ice shelf margins and internal rifts, ice streams, and the broad interior plateaus.

ICESat data over the Arctic Ocean has been used to identify and measure the elevations of the ice pack including thin ice or open water as well as the surface relief of old and first-year ice (Kwok *et al.*, 2004). Uncertainties in the knowledge of sea surface topography or "sea level" across the area require the development of surface references that allow estimation of local freeboard – the height of the air-snow surface above sea level. This can be achieved by using the relative reflectivity signature in the ICESat data of thin ice and open water to identify areas suitable for use as sea-level reference. If snow is absent, laser returns from thin ice in newly opened areas typically have a lower reflectivity than the surrounding ice cover. Returns from open water or very thin ice, however, are frequently saturated and unreliable for direct surface height estimation, so challenges remain. These results are validated by analysis of near-coincident ICESat data and openings in RADARSAT imagery (Kwok *et al.*, 2004).

By determining improved sea-level elevations from open water reference points, an estimate of the sea surface elevation relative to the ellipsoid can be derived. Using this reference surface, the state of the sea-ice pack can be determined relatively consistently, and the distributions of sea-ice thickness can be obtained over spatial scales of about 50 km by 50 km across the Arctic Ocean. Thickness distributions allow the mean sea-ice thickness and the ratio of thicker multiyear ice to thinner ice types to be determined by comparing these estimates at adjacent tracks and crossovers within the grid cell. Variable thickness snow cover can impact the freeboard height conversion to ice thickness. Currently, sea-ice thickness can be estimated using snow-cover climatology and approximate densities.

ICESat's multiple operational periods in 2003 into 2005 thus can generate sea-ice freeboard datasets to allow the study of the interannual as well as seasonal behavior of the freeboard. Maps of sea-ice thickness show similarities to those from satellite radar altimetry, and spatial distributions similar to passive microwave and scatterometer retrievals. The spatial distributions of sea-ice thickness in the Arctic Ocean for the winters (February/March) of 2003 and 2004 show significant variations. In 2004, thicker ice is more compacted adjacent to the Canadian Arctic than it was in 2003, with a larger area of thinner ice in the Beaufort and Chukchi Seas where the summer ice cover has been rapidly decreasing. In 2003, the multiyear ice pack extends farther to the south near the 45° East longitude than is typical. Additional comparisons will be enabled following the launch of ESA's CryoSat in mid-2005.

FURTHER READING

Abdalati W. and Steffen K. (1997b) Snowmelt on the Greenland ice sheet as derived from passive microwave satellite data. *Journal of Climate*, **10**, 165–175.

Ashcraft I.S. and Long D.G. (2004a) Relating microwave backscatter azimuth modulation to surface properties of the Greenland ice sheet. *Journal of Glaciology*, in review.

Drinkwater M.R., Long D.G. and Bingham A.W. (2001) Greenland snow accumulation estimates from scatterometer data. *Journal of Geophysical Research*, PARCA Special Issue, **106**(D24), 33 935–33 950.

Ezraty R. and Cavanié A. (1999) Intercomparison of backscatter maps over Arctic sea ice from NSCAT and the ERS scatterometer. *Journal of Geophysical Research*, **104**(C5), 11 471–11 483.

Ferrigno J., Williams R., Rosanova C., Lucchitta B. and Swithinbank C. (1998) Analysis of coastal change in Marie Byrd Land and Ellsworth Land, West Antarctica using Landsat imagery. *Annals of Glaciology*, **27**, 33–40.

Jezek K.C., Farness K., Carande R., Wu X. and Labelle-Hamer N. (2003) RADARSAT 1 synthetic aperture radar observations of Antarctica: modified Antarctic mapping mission, 2000. *Radio Science*, **38**(4), MAR32.1–MAR32.7.

Johannessen O.M., Shalina E.V. and Miles M.W. (1999) Satellite evidence for an arctic sea ice cover in transformation. *Science*, **286**, 1937–1939.

Kwok R., Cunningham G.F. and Yueh S. (1999a) Area balance of the Arctic Ocean perennial ice zone: October 1996 to April 1997. *Journal of Geophysical Research*, **104**, 25 747–25 759.

Ledroit M., Remy F. and Minster J.F. (1993) Observations of the Antarctic ice sheet with the seasat scatterometer: relation to katabatic-wind intensity and direction. *Journal of Glaciology*, **39**(132), 385–396.

Long D.G., Ballantyne J. and Bertoia C. (2002) Is the number of icebergs really increasing? *EOS, Transactions of the American Geophysical Union*, **83**(42), 469, 474.

Partington K. (1998) Discrimination of glacier facies using multitemporal SAR data. *Journal of Glaciology*, **44**(146), 42–53.

Young N., Hall D. and Hyland G. (1996) Directional anisotrophy of C-band backscatter and orientation of surface microrelief in east Antarctica, *Proceedings of the First Australian ERS Symposium*, pp. 117–127, February 1996.

Zhao Y., Liu A.K. and Long D.G. (2002) Validation of sea ice motion from QuikSCAT with those from SSM/I and Buoy. *IEEE Transactions on Geoscience and Remote Sensing*, **40**(6), 1241–1246.

REFERENCES

Abdalati W. and Steffen K. (1995) Passive microwave-derived snow melt regions on the Greenland ice sheet. *Geophysical Research Letters*, **22**(7), 787–790.

Abdalati W. and Steffen K. (1997a) The Apparent effects of the Mt. Pinatubo eruption on the Greenland ice sheet melt extent. *Geophysical Research Letters*, **24**, 1795–1797.

Abdalati W. and Steffen K. (2001) Greenland ice sheet melt extent. *Journal of Geophysical Research*, **106**, 33 983–33 988.

American Geographical Society of Map Antarctica 1:5 000 000 (1971) New York.

Anderson M. (1997) Determination of a melt-onset date for Arctic sea-ice regions using passive-microwave data. *Annals of Glaciology*, **25**, 382–387.

Anderson H.S. and Long D.G. (2005) Sea ice mapping method for seawinds. *IEEE Transactions on Geoscience and Remote Sensing*, **43**(3), 647–657.

Ashcraft I.S. and Long D.G. (2004b) Comparison of methods for melt detection over Greenland using active and passive microwave measurements. *International Journal of Remote Sensing*, in review.

Ashcraft I.S. and Long D.G. (2005) Observation and characterization of radar backscatter over Greenland. *IEEE Transactions on Geoscience and Remote Sensing*, **43**(2), 237–246.

Bamber J. (1994) A digital elevation model of the Antarctic ice sheet derived from ERS-1 altimeter data and comparison with terrestrial measurements. *Annals of Glaciology*, **20**, 48–54.

Benson C.S. (1962) *Stratigraphic Studies in the Snow and Firn of the Greenland Ice Sheet*, SIPRE Research Report, No. 70, SIPRE.

Berthier E., Raup B. and Scambos T. (2004) New velocity map and mass-balance estimate of Mertz Glacier, East Antarctica, derived from Landsat sequential imagery. *Journal of Glaciology*, **49**(167), 503–511.

Bindschadler R. (1998) Monitoring ice sheet behavior from space. *Reviews of Geophysics*, **36**(1), 79–104.

Bindschadler R. and Scambos T. (1991) Satellite-image-derived velocity field of an Antarctic ice stream. *Science*, **252**, 242–246.

Bindschadler R.A. and Vornberger P.L. (1990) AVHRR imagery reveals. Antarctic ice dynamics. *EOS*, **71**(23), 741–742.

Bingham A.W. and Drinkwater M.R. (2000) Recent changes in the microwave scattering properties of the Antarctic ice sheet. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(4), 1810–1820.

Bishop M., Barry R.G., Bush A.B.G., Copeland L., Dwyer J.L., Fountain A.G., Haeberli W., Hall D.K., Kääb A., Kargel J.S., *et al.* (2004) Global Land Ice Measurements from Space GLIMS): remote sensing and GIS investigations of earth's cryosphere. *Geocarta International*, **19**(2), 57–84.

Bohlander J., Haran T., Scambos T. and Fahnestock M. (2004) A new MODIS-based mosaic of Antarctica: MOA. *EOS, Transactions of the American Geophysical Union*, **85**(47), Fall Meeting Supplement, Abstract C31B–0319 F452.

Bourdelles B. and Fily M. (1993) Snow grain-size determination from Landsat imagery over Terre Adelie, Antarctica. *Annals of Glaciology*, **17**, 86–92.

Bjørgo E., Johannessen O.M. and Miles M.W. (1997) Analysis of merged SMMR-SSMI time series of Arctic and Antarctic sea ice. *Geophysical Research Letters*, **24**, 413–416.

Brenner A.C., Zwally H.J., Bentley C.R., Csathó B.M., Harding D.J., Hofton M.A., Minster J.-B., Roberts L., Saba J.L., Thomas R.H., *et al.* (1999) *Geoscience Laser Altimeter System (GLAS), Algorithm Theoretical Basis Documents*, (Version 3.0), NASA Technical paper, p. 306.

Cavalieri D.J., Gloersen P. and Campbell W.J. (1984) Determination of sea ice parameters with the NIMBUS-7 SMMR. *Journal of Geophysical Research*, **89**(D4), 5,355–5,369.

Cavalieri D.J., Gloersen P., Parkinson C.L., Comiso J.C. and Zwally H.J. (1997) Observed hemispheric asymmetry in global sea ice changes. *Science*, **278**, 1104–1106.

Cavanié A. and Gohin F. (1995) Sea-ice studies with scatterometer. In *Oceanography Applications of Remote Sensing* Ikeda M. and Dobson F. (Eds.), CRC Press, pp. 359–366.

Colbeck S.C. (1986) Classification of seasonal snow cover crystals. *Water Resources Research*, **22**, 59S–70S.

Comiso J.C. (1986) Characteristics of Arctic winter sea ice from satellite multispectral microwave observations. *Journal of Geophysical Research*, **91**(C1), 975–994.

Comiso J.C. (1994) Surface temperatures in the polar regions using Nimbus-7 THIR. *Journal of Geophysical Research*, **99**(C3), 5181–5200.

Comiso J.C. (2000) Variability and trends in Antarctic surface temperatures from in situ and satellite infrared measurements. *Journal of Climate*, **13**, 1674–1696.

Comiso J.C. (2003) Warming Trends in the Arctic. *Journal of Climate*, **16**(21), 3498–3510.

Comiso J.C., Cavalieri D.J., Parkinson C.L. and Gloersen P. (1997) Passive microwave algorithms for sea ice concentration: a comparison of two techniques. *Remote Sensing of Environment*, **60**, 357–384.

Comiso J.C. and Kwok R. (1996) Surface and radiative characteristics of the summer arctic sea ice cover from multi-sensor satellite observations. *Journal of Geophysical Research*, **101**(C12), 28 397–28 416.

Comiso J.C. and Parkinson C.L. (2004) Satellite observed changes in the Arctic. *Physics Today*, **57**(8), 38–44.

Davis C.H. and Zwally H.J. (1993) Geographic and seasonal variations in the surface properties of the ice sheets by satellite radar altimetry. *Journal of Glaciology*, **39**, 687–697.

Drinkwater M.R. (1989) LIMEX'87 ice surface characteristics: implications for C-Band SAR backscatter signatures. *IEEE Transactions on Geoscience and Remote Sensing*, **27**(5), 501–513.

Drinkwater M.R. and Liu X. (2000) Seasonal to interannual variability in Antarctic sea-ice surface melt. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(4), 1827–1842.

Early D.S. and Long D.G. (1997) Azimuth modulation of C-band scatterometer sigma-0 over southern ocean sea ice. *IEEE Transactions on Geoscience and Remote Sensing*, **35**(5), 1201–1209.

Echelmeyer K., Harrison W.D., Clarke T.S. and Benson C. (1992) Surficial glaciology of Jakobshavns Isbrae, West Greenland: Part II. Ablation, accumulation and temperature. *Journal of Glaciology*, **38**(128), 169–181.

Emery W., Fowler C., Hawkins J., Preller R. (1991) Fram Strait satellite image derived ice motions. *Journal of Geophysical Research*, **96**(C3), 4751–4768; Correction: (1991) *Journal of Geophysical Research*, **96**(C5), 8917–8920.

Eppler D.T., Farmer L.D., Lohanick A.L., Anderson M.R., Cavalieri D.J., Comiso J., Gloersen P., Garrity C., Grenfell T.C., Hallikainen M., *et al.* (1992) Passive microwave signatures of sea ice. In *Microwave Remote Sensing of Sea Ice*,Carsey F. (Ed.), AGU: Washington, AGU Geophysical Monograph 68, pp. 41–71.

Fahnestock M.A., Abdalati W. and Shuman C.A. (2002) Long melt seasons on ice shelves of the Antarctic Peninsula: an analysis using satellite-based microwave emission measurements. *Annals of Glaciology*, **34**, 127–133.

Fahnestock M.A. and Bindschadler R.A. (1993) Description of a program for SAR investigation of the Greenland ice sheet and an example of margin change detection using SAR. *Annals of Glaciology*, **39**(131), 119–132.

Fahnestock M., Bindschadler R., Kwok R. and Jezek K. (1993) Greenland ice sheet surface properties and ice dynamics from ERS-1 SAR imagery. *Science*, **262**, 1530–1534.

Fahnestock M., Scambos T., Bindschadler R. and Kvaran G. (2000) A millennium of variable ice flow recorded by the Ross Ice Shelf, Antarctica. *Journal of Glaciology*, **46**(155), 652–664.

Ferrigno J.G., Mullins J., Jerry L., Stapleton J.A., Bindschadler R.A., Scambos T.A., Bellisime L.B., Bowell J.A. and Acosta A.V. (1994) Landsat TM image maps of the Shirase and Siple Coast ice streams, West Antarctica. *Annals of Glaciology*, **20**, 407–412.

Fetterer F. and Untersteiner N. (1998) Observations of melt ponds on Arctic sea ice. *Journal of Geophysical Research*, **103**(C11), 24 821–24 835.

Forster R.R. and Baumgras L.M. (2000) Arctic snowmelt onset dates determined by remote sensing data and streamflow measurements, *Association of American Geographers Annual Meeting*, Pittsburgh, April 4-8.

Forster R.R., Davis C.H., Rand T.W. and Moore R.K. (1991) Snow-stratification investigation on an Antarctic ice stream with an X-band radar system. *Journal of Glaciology*, **37**(127), 323–325.

Forster R.R., Long D.G., Jezek K.C., Drobot S.D. and Anderson M.R. (2001) The onset of Arctic sea-ice snowmelt as detected with passive- and active-microwave remote sensing. *Annals of Glaciology*, **33**, 85–93.

Gloersen P. and Campbell W.J. (1991) Recent variations in Arctic and Antarctic sea-ice covers. *Nature*, **352**, 33–36.

Grenfell T.C. and Lohanick A.W. (1985) Temporal variations of the variations of the microwave signatures of sea ice during the late spring and early summer near Mould Bay, NWT. *Journal of Geophysical Research*, **90**, 5063–5074.

Haarpainter J., Tonboe R.T., Long D.G. and VanWoert M.L. (2004) Automatic detection and validity of the sea ice edge: an application of enhanced resolution quikscat/seawinds data. *IEEE Transactions on Geoscience and Remote Sensing*, **42**(7), 1433–1443.

Hastings D. and Emery W. (1992) The Advanced Very High Resolution Radiometer (AVHRR): a brief reference guide. *Photogrammetric Engineering and Remote Sensing*, **58**(8), 1183–1188.

Henderson F.M. and Lewis A.J. (1998) *Principles and Applications of Imaging Radar: Manual of Remote Sensing*, John Wiley & Sons: New York, Vol. 2.

Howell S.E.L., Yackel J.J., De Abreu R., Geldsetzer T. and Breneman C. (2005) On the utlity of seawinds/QuikSCAT data for the estimation of the thermodynamic state of first-year sea ice. *IEEE Transactions Geoscience and Remote Sensing*, **43**(6), 1338–1350.

Jezek K.C. (1999) Glaciological properties of the Antarctic ice sheet from RADARSAT-1 synthetic aperture radar imagery. *Annals of Glaciology*, **29**, 286–290.

Jezek K.C., Drinkwater M.R., Crawford J.P., Bindschadler R. and Kwok R. (1993) Analysis of Synthetic Aperture Radar data collected over the southwestern Greenland ice sheet. *Journal of Glaciology*, **39**(131), 115–132.

Jezek K.C. and Gogineni S. (1992) Microwave remote sensing of the Greenland ice sheet. *IEEE Geoscience and Remote Sensing Society Newsletter*, **85**, 6–10.

Jezek K.C., Gogineni P. and Shanableh M. (1994) Radar measurements of melt zones on the Greenland ice sheet. *Geophysical Research Letters*, **21**(1), 33–36.

Johannessen O.M., Miles M.W. and Bjørgo E. (1995) The Arctic's shrinking sea ice. *Nature*, **376**, 126–127.

Key J. and Haefliger M. (1993) Arctic ice surface temperature retrieval from AVHRR thermal channels. *Journal of Geophysical Research*, **97**, 5885–5893.

King J.C. and Comiso J.C. (2003) The spatial coherence of interannual temperature variations in the Antarctic Peninsula. *Geophysical Research Letters*, **30**(2), 1040, doi:10.1029/2002GL015580.

Konig M. (2001) Measuring snow and glacier ice properties from satellite. *Reviews of Geophysics*, **39**(1), 1–27.

Krabill W., Abdalati W., Frederick E., Manizade S., Martin C., Sonntag J., Swift R., Thomas R., Wright W. and Yungel J. (2000) Greenland ice sheet: high-elevation balance and peripheral thinning. *Science*, **289**(5478), 428–430.

Kunz L.B. and Long D.G. (2005) Melt detection in Antarctic ice-sheets using spaceborne scatterometers and radiometers. *IEEE Transactions on Geoscience and Remote Sensing*, Submitted to.

Kwok R. and Comiso J.C. (2002) Spatial patterns of variability in Antarctic surface temperature: connections to the southern hemisphere annular mode and the southern oscillation. *Geophysical Research Letters*, **29**(14), 10.1029/2002GL015415.

Kwok R., Cunningham G.F., LaBelle-Hamer N., Holt B. and Rothrock D. (1999b) Ice thickness derived from high-resolution radar imagery. *EOS, Transactions of the American Geophysical Union*, **80**(42), 495.

Kwok R., Cunningham G.F. and Nghiem S.V. (2003) A study of the onset of melt over the Arctic Ocean in RADARSAT synthetic aperture radar data. *Journal of Geophysical Research C: Oceans*, **108**(C11), 27.1–27.13.

Kwok R. and Fahnestock M. (1996) Ice-sheet motion and topography from radar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, **34**(1), 189–200.

Kwok R., Zwally H.J. and Yi D. (2004) ICESat observations of Arctic sea ice: A first look. *Geophysical Research Letters*, **31**, L16401, doi:10.1029/2004GL020309.

Livingstone C.E. and Drinkwater M.R. (1991) Springtime C-Band SAR backscatter signatures of labradore sea marginal ice: measurements versus modeling predictions. *IEEE Transactions on Geoscience and Remote Sensing*, **29**(1), 29–41.

Long D.G. and Drinkwater M.R. (1994) Greenland observed at high resolution by the Seasat-A scatterometer. *Journal of Glaciology*, **32**(2), 213–230.

Long D.G. and Drinkwater M.R. (1999) Cryosphere applications of NSCAT data. *IEEE Transactions on Geoscience and Remote Sensing*, **37**(3), 1671–1684.

Long D.G. and Drinkwater M.R. (2000) Azimuth variation in microwave scatterometer and radiometer data over Antarctica. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(4), 1857–1870.

Long D.G., Drinkwater M.R., Holt B., Saatchi S. and Bertoia C. (2001) Global ice and land climate studies using scatterometer image data. *EOS, Transaction of the American Geophysical Union*, **82**(43), 503, Includes *EOS* Electronic Supplement: `http://www.agu.org/eos_elec/010126e.html`

Lucchitta B.K. and Ferguson H.M. (1986) Antarctica – measuring glacier velocity from satellite images. *Science*, **234**, 1105–1108.

Mader R.E. (1991) *Synthetic Aperture Radar Imaging of Glacial Ice, Technical Report, RSL Report 8291-1, University of Kansas Remote Sensing Laboratory, pp. 111.*

Markus T. and Cavalieri D.J. (2000) An enhancement of the NASA team sea ice algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 1387–1398.

Maslanik J., Fowler C., Key J., Scambos T., Hutchinson T. and Emery W. (1997) AVHRR-based polar pathfinder products for modeling applications. *Annals of Glaciology*, **25**, 388–392.

Maslanik J., Serreze M.C. and Barry R.G. (1996) Recent decreases in Arctic summer ice cover and linkages to atmospheric circulation anomalies. *Geophysical Research Letters*, **23**, 1677–1680.

Massom R. (1995) *Satellite Remote Sensing of Polar Snow and Ice: Present Status and Future Directions*, Belhaven Press in association with the Scott Polar Research Institute, University of Cambridge: London; Lewis Publishers: Boca Raton, 1991.

Mätzler C. and Hüppi R. (1989) Review of signature studies for microwave remote sensing of snowpacks. *Advances in Space Research*, **9**(1), 253–265.

Mote T.L. and Anderson M.R. (1995) Variations in snowpack melt on the Greenland ice sheet based on passive-microwave measurements. *Journal of Glaciology*, **41**, 51–60.

Munk J., Jezek K.C., Forster R.R. and Gogineni S.P. (2003) An accumulation map for the Greenland dry-snow facies derived from spaceborne radar. *Journal of Geophysical Research D: Atmospheres*, **108**(9), ACL8.1–ACL8.12.

Nghiem S.V., Kwok R., Yueh S.H. and Drinkwater M.R. (1995) Polarimetric signatures of sea ice – 1. Theoretical models. *Journal of Geophysical Research*, **100**(C7), 13 665–13 679.

Nghiem S.V., Steffen K., Kwok R. and Tsai W.-Y. (2001) Detection of snowmelt regions on the Greenland ice sheet using diurnal backscatter change. *Journal of Glaciology*, **47**, 539–547.

Nolin A., Fetterer F. and Scambos T. (2002) Surface roughness characterizations of sea ice and ice sheets: case studies with MISR data. *IEEE Transactions on Geoscience and Remote Sensing*, **40**(7), 1605–1615.

Onstott R.G., Grenfell T.C., Maetzler C., Luther C.A. and Svendsen E.A. (1987) Evolution of microwave sea ice signatures during early summer and midsummer in the Marginal Ice Zone. *Journal of Geophysical Research*, **92**, 6825–6835.

Orheim O. and Lucchitta B. (1987) Snow and ice studies by thematic mapper and multispectral scanner images. *Annals of Glaciology*, **9**, 109–118.

Parkinson C.L. (1992) Spatial patterns of increases and decreases in the length of the sea ice season in the north polar region. *Journal of Geophysical Research*, **97**, 14 377–14 388.

Parkinson C.L. (1994) Spatial patterns in the length of the sea ice season in the Southern Ocean, 1979-1986. *Journal of Geophysical Research*, **99**, 16 327–16 339.

Parkinson C.L., Cavalieri D.J., Gloersen P., Zwally H.J. and Comiso J.C. (1999) Arctic sea ice extents, areas and trends, 1978–1996. *Journal of Geophysical Research*, **104**(C9), 20 837–20 856.

Partington K., Flynn T., Lamb D., Bertoia C. and Dedrick K. (2003) Late twentieth century Northern Hemisphere sea-ice record from U.S. National Ice Center ice charts. *Journal of Geophysical Research*, **108**(C11), 3343, doi:10.1029/2002JC001623.

Pfeffer W.T., Illangsekare T.H. and Meier M.F. (1990) Analysis and modeling of melt-water refreezing in dry snow. *Journal of Glaciology*, **36**(123), 238–246.

Rankin A.M., Wolff E.W. and Martin S. (2003) *Frost Flowers – Implications for Tropospheric Chemistry and Ice Core Interpretation*, in press.

Remund Q.P. and Long D.G. (1999) Sea ice extent mapping using Ku-band scatterometer data. *Journal of Geophysical Research*, **104**(C5), 11 515–11 527.

Remund Q.P., Long D.G. and Drinkwater M.R. (2000) An iterative approach to multisensor sea ice classification. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(4), 1843–1856.

Rignot E. (1996) Tidal motion, ice velocity, and melt rate of Petermann-Gletscher, Greenland, measured from radar interferometry. *Journal of Glaciology*, **42**(142), 476–485.

Scambos T., Bohlander J., Shuman C. and Skvarca P. (2004) Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. *Geophysical Research Letters*, **31**, doi:10.1029/2004GL020670.

Scambos T., Dutkiewicz M., Wilson J. and Bindschadler R. (1992) Application of image cross-correlation to the measurement of glacier velocity using satellite image data. *Remote Sensing of Environment*, **42**, 177–186.

Scambos T.A., Hulbe C. and Fahnestock M. (2003) Climate-induced ice shelf disintegration in the Antarctic Peninsula. In *Antarctic Peninsula Climate Variability: Historical and Paleoenvironmental Perspectives*, *Antarctic Research Series, Vol. 79* Domack E., Burnett A., Leventer A., Conley P., Kirby M. and Bindschadler R. (Eds.), American Geophysical Union: Washington, 79–92.

Scambos T.A., Hulbe C., Fahnestock M. and Bohlander J. (2000) The link between climate warming and break-up of ice shelves in the Antarctic peninsula *Journal of Glaciology*, **46**(154), 516–530.

Schneider D.P., Steig E.J. and Comiso J.C. (2004) Recent climate variability in Antarctica from satellite-derived temperature data. *Journal of Climate*, **17**, 1569–1583.

Serreze M.C., Maslanik J.A., Scambos T.A., Fetterer F., Stroeve J., Knowles K., Fowler C., Drobot S., Barry R.G. and Haran T.M. (2003) Record minimum sea ice cover in the Arctic Ocean for summer 2002. *Geophysical Research Letters*, **30**, 1110–1113.

Shuman C.A. and Alley R.B. (1993) Spatial and temporal characterization of hoar formation in central Greenland using SSM/I brightness temperatures. *Geophysical Research Letters*, **20, 23**, 2643–2646.

Shuman C.A. and Comiso J. (2002) In situ and satellite surface temperature records in Antarctica. *Annals of Glaciology*, **34**, 113–120.

Shuman C.A., Steffen K., Box J.E. and Stearns C.R. (2001) A dozen years of temperature observations at Shuman, C., and J.C. Comiso, In situ and satellite surface temperature records in Antarctica. *Annals of Glaciology*, **34**, 113–120;(2002) The summit: Central Greenland automatic weather stations 1987-1999, *Journal of Applied Meteorology*, **40, 4**, 741–752.

Skvarca P. and De Angelis H. (2002) *First Cloud-Free Landsat TM image mosaic of Heilo Patagonico Sur, Southwestern Patagonia, South America*, Direccion Nacional Del Antartico Contribucion No. 535, Instituto Antartico Argentino.

Smith D.M. (1996) Extraction of winter total sea-ice concentration in the Greenland and Barents Seas from SSM/I data. *International Journal of Remote Sensing*, **17**(33), 2625–2646.

Smith D.M. (1998) Observation of perennial Arctic sea ice melt and freeze-up using passive microwave data. *Journal of Geophysical Research*, **103**, 27 753–27 769.

Stearns L. and Hamilton G. (2003) Velocities along Byrd Glacier, East Antarctica, derived from automatic feature tracking. *EOS*, **84**(46), F351.

Steffen K., Bindschadler R., Casassa C., Comiso J., Eppler D., Fetterer F., Hawkins J., Key J., Rothrock D., Thomas R., *et al.* (1993) Snow and ice applications of AVHRR in polar regions. *Annals of Glaciology*, **17**, 1–16.

Steffen K., Cavalieri D.J., Comiso J.C., Germain K.S.t, Gloersen P., Key J. and Rubinstein I. (1992) The estimation of geophysical parameters using passive microwave algorithms. In *Microwave Remote Sensing of Sea Ice*, Carsey F. (Ed.), American Geophysical Union: Washington, Chap. 10, pp. 243–259.

Stogryn A. (1987) Strong fluctuation theory for moist granular media. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-23**(s), 78–83.

Stroeve J.C., Serreze M.C., Fetterer F., Arbetter T., Meier W., Maslanik J. and Knowles K. (2005) Tracking the Arctic's shrinking ice cover: Another extreme September minimum in 2004 Geophys. *Research Letters*, **32**(4), doi:10.1029/2004GL021810.

Stroeve J., Nolin A. and Steffen K. (1997) Comparison of AVHRR-derived and in situ albedo over the Greenland ice sheet. *Remote Sensing of Environment*, **62**(3), 262–276.

Svendsen E., Kloster K., Farrelly B., Johannessen O.M., Johannessen J., Campbell W., Gloersen P., Cavalieri D. and Matzler C. (1983) Norwegian remote sensing experiment: evaluation of the Nimbus-7 SMMR for sea ice research. *Journal of Geophysical Research*, **88**, 2781–2791.

Swift C.T. (1999) Seasat scatterometer observations of sea ice. *IEEE Transactions on Geoscience Remote Sensing*, **37**(2), 716–723.

Swift C.T., Hayes P.S., Herd J.S., Jones W.L. and Delnore V.E. (1985) Airborne microwave measurements of the Southern Greenland ice sheet. *Journal of Geophysical Research*, **90**(B2), 1983–1994.

Swithinbank C., Brunk K. and Sievers J. (1988) A glaciological map of the Filchner-Ronne Ice Shelf, Antarctica. *Annals of Glaciology*, **11**, 150–155.

Thomas R.H., Bindschadler R.A., Cameron R.L., Carsey F.D., Holt B., Hughes T.J., Swithinbank C.W.M., Whillans I.M. and Zwally H.J. (1985) *Satellite Remote Sensing for Ice Sheet Research*, NASA technical memo 86233, NASA Scientific and Technical Information Branch, Washington, p. 32.

Thomas R., Rignot E., Casassa G., Kanagaratnam P., Acuña C., Akins T., Brecher H., Frederick E., Gogineni P., Krabill W.,

et al. (2004) Accelerated sea-level rise from West Antarctica. *Science*, **306, 5694**, 255–258.

Tiuri M.E., Sihvola A.H., Nyfors E.G. and Hallikainen M.T. (1984) The complex dielectric constant of snow at microwave frequencies. *IEEE Journal of Oceanic Engineering*, **OE-9**, 377–382.

Tsatsoulis C. and Kwok R. (Eds.) (1998) *Analysis of SAR Data of the Polar Oceans- Recent Advances*, Springer-Verlag: Berlin.

United Stated Geological Survey Miscellaneous Investigations Map I-2284 (1991) Reston Virginia.

Ulaby F.T., Moore R.K. and Fung A.K. (1981) *Microwave Remote Sensing – Active and Passive*, Addison-Wesley Publishing Co.: Reading, mass., Vols. 1 and 2.

Warrick, R.A., Provost C.L., Meier M.F., Oerlemans J. and Woodworth P.L. (1996) Changes in sea level. In *Climate Change 1995: The Science of Climate Change*, pp. 359–405.

Wiesmann A. and Maetzler C. (1999) Microwave emission model for layered snowpacks. *Remote Sensing of Environment*, **70**, 307–316.

Williams R. and Ferrigno J. (Eds.) (1988–2002) *Satellite Image Atlas of Glaciers of the World*, United States Geological Survey Professional Paper 1386-A through K.

Winebrenner D.P., Long D.G. and Holt B. (1998) Automatable observation of seasonal transitions on arctic sea ice using synthetic aperture radar. In *Recent Advances in the Analysis of SAR for Studies in the Polar Oceans*, Tsatsoulis C. and Kwok R. (Eds.), Springer-Verlager: pp. 129–144.

Winebrenner D.P., Nelson E.D., Colony R. and West R.D. (1994) Observation of melt onset on multiyear Arctic sea ice using the ERS-1 synthetic aperture radar. *Journal of Geophysical Research*, **99**, 22 425–22 441.

Wismann V. (2000) Monitoring of seasonal snowmelt on Greenland with ERS scatterometer data. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(4), 1821–1826.

Wolff E.W. (2003) Whither Antarctic sea ice? *Science*, **302**, 1164.

Young N.W. and Hyland G. (2002) Velocity and strain rates derived from InSAR analysis over the Amery Ice Shelf, East Antarctica. *Annals of Glaciology*, **34**(1), 228–234.

Zwally H.J. and Bindschadler R.A. (1983) Surface elevation contours of Greenland and Antarctic ice sheets. *Journal of Geophysical Research*, **88**(C3), 1589–1596.

Zwally H.J., Major J.A., Brenner A.C. and Bindschadler R.A. (1987) Ice measurements by GEOSAT radar altimetry. *Johns Hopkins APL Technical Digest*, **8**(2), 251–254.

Zwally H.J., Schutz B., Abdalati W., Abshire J., Bentley C., Brenner A., Bufton J., Dezio J., Hancock D., Harding D., *et al.* (2002) ICESat's Laser Measurements of Polar Ice, Atmosphere, Ocean, and Land. *Journal of Geodynamics*, **34**(3-4), 405–445.

57: Land-cover Classification and Change Detection

MATTHEW C HANSEN[1] AND SCOTT J GOETZ[2]

[1]Geographic Information Science Center of Excellence, South Dakota State University, Brookings, SD, US
[2]Woods Hole Research Center, Woods Hole, MA, US

Various multispectral data sets and tools are available for mapping land cover and land-cover change for inputs to hydrologic applications. The best choice is often a function of the specific application. New sensors offer increased capability in mapping spatial detail and temporal variation of land-cover categories. Algorithms have advanced in robustness and include distribution-free methods that are superior to traditional approaches in terms of modeling the multispectral distributions of reference data. Advanced subpixel methods of mapping land cover offer greater thematic coherency and the possibility of using consecutive land-cover characterizations to map change. The success of deriving a meaningful result, such as deriving a relationship between aquatic biotic indices and land-cover change, relies on a high-quality land-cover reference map. By choosing the most appropriate data sources, constructing a defensible land-cover definition set, and employing robust algorithms, analysts allow for the meaningful incorporation of land-cover map information into hydrological studies.

INTRODUCTION

Land cover is defined as the observed biophysical state of the earth's surface, and is largely described by the presence or absence of various vegetation types. Land-cover information is an important parameter for many hydrological studies. Vegetation type and density affect rainfall interception, evapotranspiration, runoff estimation, water quality, and stream flow, among other parameters (Gorte, 2000). The uses of land-cover information in hydrological applications are varied. Broad categories include operational hydrologic modeling, climate modeling, and water quality assessments. Concerning hydrologic modeling, land-cover information can be used to help determine runoff coefficients and to characterize infiltration, erosion, and evapotranspiration for distributed models (Dubayah *et al.*, 2000; Liang *et al.*, 1999). When land-cover conditions change, so do the runoff, soil moisture, and groundwater characteristics of the land. For example, Foley *et al.* (2003) have shown the effects of land-cover change, namely, the increasing dominance of agriculture on water resources across the Mississippi River basin. An advantage of remotely sensed land-cover characterizations is the spatially explicit nature of derived land-cover information. Schumann and Schultz (2000) point out the need to have spatial and temporal detail concerning land-use changes in relation to other river basin characteristics such as soil type. Concerning climate studies, land cover is an important boundary condition in parameterizing soil–vegetation–atmosphere transfer schemes, which can be incorporated into general circulation models. Such models include the Biosphere–Atmosphere Transfer Scheme (BATS) (Dickinson *et al.*, 1986), the Simple Biosphere (SiB) model (Sellers *et al.*, 1986), and the Land–Atmosphere Transfer Scheme (Pauwels and Wood, 1999). Land-cover characteristics govern many variables, including surface roughness for energy transfer, albedo for solar absorption, moisture from canopies and soils, and mechanisms for water runoff (Dickinson, 1995). Land cover is also a critical variable in determining water quality (Nilsson *et al.*, 2003) and for assessing pollution risk and watershed health (Wickham *et al.*, 2000). Remotely sensed data sets offer the opportunity to efficiently quantify various land-cover components over large areas repeatedly through time, enabling researchers to

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

easily incorporate land-cover information into hydrological studies.

This chapter reports methods for characterizing land cover and land-cover change using remotely sensed data sets. In strict terms, it is a primer for technical approaches to land-cover mapping from multispectral data sets. However, using land-cover information as an input to hydrologic applications requires knowing the strengths and limitations of such data sets regarding map accuracy, spatial scale, and thematic content. For example, Pauwels and Wood (2000) report that land-cover classification accuracy had a greater effect on their land–atmosphere model uncertainty than the spatial detail of the land-cover depiction, and that overall uncertainty attributable to land cover was relatively small compared to that of meteorological forcing data (*see* **Chapter 201, Land-Atmosphere Models for Water and Energy Cycle Studies, Volume 5**). An example of varying uncertainty related to thematic content is shown in climate modeling science (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**). Dickinson (1995) advocated the use of subpixel estimates of proportional land cover, as opposed to discrete land-cover classifications, in improving climate model parameterization. Bonan *et al.* (2002) concurred in showing that uncertainty in climate and ecosystem models for such variables as ground temperature and ground evaporation was reduced when continuous representations of vegetation were used in place of discrete biome maps. These studies emphasize the need to understand the interplay of map characteristics and their potential effects on modeling outputs.

This chapter outlines the use of visible (0.4–0.7 μm), near-infrared (0.7–1.3 μm) and middle-infrared (1.3–3 μm) reflected energy, and emitted energy from the thermal infrared (3–14 μm) wavelengths, in mapping land cover. Energy reflected and emitted by the earth in this spectral region is captured by many earth observing satellite sensors. In recent years, significant advances in land-cover mapping using such satellite data have been made. First, improved satellite data sets have become available. Data from new sensors with superior spectral and spatial characteristics have allowed for more accurate depictions of land cover and land-cover change. Algorithms new to remote sensing have been successfully implemented with satellite imagery and found to be both robust and user-friendly. Improvements in thematic depictions including the increased use of subpixel fractional cover estimates have advanced mapping spatial detail and land-cover change over time. This chapter outlines the available data sets, algorithms, thematic outputs, and change detection techniques available to potential users of multispectral remotely sensed data sets. A case study using the addressed techniques for a hydrological application mapping impervious surface and tree cover is also provided.

FINDING APPROPRIATE DATA FOR MAPPING LAND COVER

In attempting to characterize land cover using multispectral remotely sensed data sets, a number of considerations must be taken into account. These considerations lead to the selection and acquisition of data from specific satellite sensors and include requirements concerning the spectral, spatial, and temporal dimensions of the problem. The initial step is to define the thematic content to be mapped. This definition set dictates the level of spatial and spectral detail required for successfully mapping the land-cover theme of interest. For most purposes, definition sets consist of a classification scheme made up of two or more land-cover categories. Less frequently employed are continuous representations of land-cover categories or basic vegetation traits. Instead of discrete land-cover types, these products depict each mapping unit as proportional estimates of a land-cover category or attribute, such as percent vegetative cover.

Whichever approach to characterizing land cover is taken, the analyst must first determine if optical remotely sensed data sets are appropriate for mapping the land-cover variable of interest. Available spectral information captured by a sensor should relate to the absence or presence of the land-cover theme of interest, and this leads ultimately to the selection of a specific data set for inclusion in the study. While multispectral data sets are commonly used for mapping general land-cover categories, researchers should be aware of limitations to using these data. For example, optical data have been found to be of limited value in mapping forested wetlands, and researchers have employed radar data (DeGrandi *et al.*, 2000) or a combination of optical and radar data sources (Townsend and Walsh, 1998) to map flooded forests. Radar data have been shown to be more sensitive to variations in fundamental wetland attributes such as standing water and soil moisture than optical data sets (Ramsey, 1999).

Table 1 shows land-cover features discernible within different regions of the visible and infrared parts of the electromagnetic spectrum. Visible and near-infrared reflected energy have long been used to map the presence and types of vegetation cover as well as nonvegetated surfaces such as bare soils or impervious surfaces (Jensen, 1996). A very useful measure, the Normalized Difference Vegetation Index (NDVI), has been shown to be positively correlated with vegetation greenness (Tucker, 1979). The NDVI is a normalized ratio near-infrared and red reflected energy and takes the following form:

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{RED})}{(\mathrm{NIR} + \mathrm{RED})} \tag{1}$$

The change of vegetation cover through time, or its phenology, is captured using repeated images of derived NDVI. The depiction of phenology with satellite data has

Table 1 Principal spectral divisions and associated utility in observing surface conditions

Blue (400–500 nm): Highly influenced by atmospheric scattering and absorption. Provides penetration into water bodies and strong absorption by green vegetation.
Green (500–600 nm): Positively correlated with presence of healthy vegetation.
Red (600–700 nm): Strongly absorbed by green vegetation. Extinction of red light occurs in complex vegetation canopies. Used to derive Normalized Difference Vegetation Index (NDVI), which correlates strongly with presence of green vegetation. Sensitive to turbidity levels of water.
Near-infrared (700–1100 nm): Responsive to abundance of green vegetation. Used with red reflectance information to create NDVI. Useful in differentiating land from water. Sensitive to leaf geometry (needleleaf/broadleaf) and leaf water content.
Mid-infrared (1100–3000 nm): Used to differentiate clouds, snow and ice. Sensitive to soil moisture levels and related to plant vigor and drought stress. Water is absorbed very strongly. Important in identifying geologic features.
Thermal infrared (3000–30 0 000 nm): Useful in monitoring drought and water stress in plants, allowing for the mapping of drought deciduous woodlands. Also used for land-surface temperature calculations that, in conjunction with NDVI, have been used to measure land-cover change.

significantly enhanced the ability to discriminate many vegetative cover categories (Justice *et al.*, 1985). Visible and near-infrared energy have also proven useful in mapping water bodies and water clarity (Lillesand and Kieffer, 1994; Harrington *et al.*, 1992). Emitted energy in the thermal region has been related to surface temperature (Price, 1984) and has also been used in ratio form with NDVI to predict soil moisture status (Nemani *et al.*, 1993). Studies using both NDVI and surface temperature in concert have shown their utility in documenting both short and long-term environmental variation as well as land degradation (Lambin and Ehrlich, 1996).

The second consideration in defining the land-cover mapping problem is to outline the spatial requirements of the study in terms of areal extent and required spatial detail. For this discussion, areal extents of study can be divided into site level studies ($<100\,km^2$), local scale studies (100 to 10 000 km^2), regional scale studies (10 000 to 1 000 000 km^2), and continental to global scale studies ($>1\,000\,000\,km^2$). The use of specific remotely sensed data sets in land-cover mapping is usually a function of the spatial extent of the area of study. Spatial resolution is generally inversely related to image swath size (ground area captured by the sensor in a single scene), meaning sensors with higher spatial detail cover smaller areas than sensors with lower spatial resolution. So, depicting fine spatial detail over large areas requires the use of multiple images and a higher order of data processing requirements. Cost is often a determining factor in study feasibility and the costs of imagery increase with increasing spatial detail. As a consequence, studies examining detailed landscape features such as individual tree crowns or fine-scale drainage features are going to be limited in areal scale, most probably at the site level, due to the increased costs of the very high-resolution data required for such research. Table 2 shows the various characteristics for selected sensors.

When the spatial extent of the study area is larger than the footprint of desired input imagery, then the mosaicking of images acquired at different times is required. Land-cover characterizations using mosaicked images are performed in one of two ways. Either each image is separately characterized and the resulting products are stitched together, or all images are radiometrically normalized and an algorithm is run on a single mosaiced image. Radiometric normalization is the process by which images of varying illumination geometries and/or atmospheric conditions are made directly comparable. Thus, the use of individual images has significant advantages given that the study area fits within a single orbital swath. In general, less work is involved in preprocessing if the area of study fits within the footprint of a single image.

Third, the temporal scale of the mapping exercise must be defined. Is the study a single-date characterization for use with concurrently collected *in situ* data sets, or are multiple characterizations sought with the goal of performing a land-cover change detection study? In many cases, the simple identification of a land-cover category requires the use of multiple images in order to identify a characteristic vegetation phenological profile. Land-cover mapping using multitemporal image inputs is more commonly performed with coarse resolution sensors that have shorter repeat coverage intervals than fine-resolution sensors. As with areal scale considerations, the required temporal frequency of image capture will govern which sensor's data are most appropriate to the study. Table 2 shows the repeat coverage interval for available optical remote sensors.

A particular advantage of high temporal resolution sensors is the ability to composite data over time. Compositing algorithms preferentially select clear-sky pixels via an algorithm that operates on multitemporal data. Hand-selecting only cloud-free imagery for inputs is highly impractical for large regional and continental-scale study areas. The most commonly used method is a maximum-value compositing approach employing the NDVI (Holben, 1986). The NDVI is a ratio measure of near-infrared to red reflectance and is positively correlated with vegetation greenness. As such, it is a measure used to reduce the inclusion of cloud-contaminated data in time series imagery. Compositing

Table 2 Spectral, spatial, temporal and radiometric characteristics for commonly used multispectral sensors

		Spectral resolution	Spatial resolution (m)	Temporal resolution	Radiometric resolution (bits)	Swath width (km)
Very high spatial resolution	IKONOS	445–516 nm (blue) 506–595 nm (green) 632–698 nm (red) 757–853 nm (nir)	4	pointable	11	11
		450–900 nm (pan)	1			
	Quickbird	450–520 nm (blue) 520–600 nm (green) 630–690 nm (red) 760–900 nm (nir)	2.5	pointable	11	17
		450–900 nm (pan)	0.61			
	SPOT HRG	500–590 nm (green) 610–680 nm (red) 790–890 nm (nir)	10			
		1580–1750 nm (swir)	20	pointable	8	60
		510–730 nm (pan)	5			
	SPOT HRV	500–590 nm (green) 610–680 nm (red) 790–890 nm (nir)	20	pointable	8	60
		510–730 nm (pan)	10			
High spatial resolution	CBERS	450–520 nm (blue) 520–590 nm (green) 630–690 nm (red) 770–890 nm (nir) 510–730 nm (pan)	20	pointable	8	120
	IRS-LISS III	520–590 nm (green) 620–680 nm (red) 770–860 nm (nir) 1.55–1.70 (swir)	23.5	24 days	7	142
	Landsat ETM+	450–520 nm (blue) 520–600 nm (green) 630–690 nm (red) 760–900 nm (nir) 1550–1750 nm (swir) 2080–2350 nm (swir)	30	16 days	8	185
		10 400–12 500 nm (thermal)	60			
		520–900 nm (pan)	15			
Moderate spatial resolution	MERIS	5 bands from 400 to 1050 nm	300	3 days	12	1150
	MODIS	620–670 nm (red) 841–876 nm (nir)	250			
		459–479 nm (blue) 545–563 nm (green) 1230–1250 nm (swir) 1628–1652 nm (swir) 2105–2155 nm (swir)	500	daily	12	2330
		29 other surface/atmosphere/ocean bands	1000			
Coarse spatial resolution	VEGETATION	430–470 nm (blue) 610–680 nm (red) 780–890 nm (nir) 1580–1750 nm (swir)	1000	daily	10	2250
	AVHRR	580–680 nm (red) 725–1100 nm (nir) 3550–3930 nm (thermal) 10 300–11 300 nm (thermal) 11 500–12 500 nm (thermal)	1100	daily	10	2399

reduces the volume of the data, while still retaining the temporal variability of the land-cover signal, given an appropriate compositing period. However, compositing requires significant preprocessing such as the application of an atmospheric correction algorithm in order to create radiometrically normalized images (Cihlar, 2000). Using composited data sets becomes the clearer choice as study areas increase from regional to continental in size. The ability to find individual cloud-free images is less probable as the area of study increases, and compositing becomes a practical solution to reducing cloud contamination. However, compositing can introduce noise as well as reduce spatial fidelity (Justice *et al.*, 1989; Moody and Strahler, 1994) when compared to optimum single-date images.

The aforementioned information domains of spectral, spatial, and temporal characteristics, are shown in Table 2, and are classified by spatial detail into four generic categories: very high resolution (<5 m), high resolution (5–100 m), moderate resolution (100–500 m), and coarse resolution data (>500 m) (*see* **Chapter 47, Sensor Principles and Remote Sensing Techniques, Volume 2**).

INDEPENDENT AND DEPENDENT VARIABLES

The independent variables input to any algorithm are the remotely sensed data. The number of inputs typically equals the number of satellite bands times the number of images (dates). Composites are another source of image inputs representing nominally clear-sky conditions over a fixed period of time and are available only for moderate and coarse resolution data sets. Ancillary data such as digital elevation models (DEMs) can also be used as inputs to mapping exercises. Some algorithms can also ingest categorical information such as soil type as independent variables.

Defining the dependent variable of land-cover type is an important and often overlooked part of classifying land cover. Some classification schemes developed by the modeling and remote sensing communities have significant problems that limit their utility. It is paramount to use a clear set of class labels and definitions for each label when developing a classification legend. These labels and definitions constitute the classification scheme and must be (i) mutually exclusive and (ii) totally exhaustive (Congalton and Green, 1999). It is important to avoid using legends that have holes or overlaps in their classification schemes. This is of particular importance when it comes to validating a land-cover map. If the legend does not hold to these two rules, then it is virtually impossible to validate the map. In addition, hierarchies within the classification scheme should be included to nest classes into categories of increasing/decreasing thematic detail. This affords the analyst to represent various confidence levels for greater- or less-generalized map products. This is useful for showing algorithm limitations by nesting classes in different ways, especially those with low accuracies. Sensitivity to the needs and desires of the user community must also be met by including all classes that can feasibly be mapped. For mapping exercises that produce per-pixel continuous-cover estimates, the same rules for producing a set of mutually exclusive and thematically exhaustive definitions apply.

When mapping using satellite data, a physiognomic-structural vegetation characterization scheme is advocated. First, land cover, defined as the observed biophysical state of the earth's surface, lends itself most unambiguously to a physiognomic definition set (DiGregorio and Jansen, 2000). Second, the signal, being mapped in multispectral and multitemporal space, is highly correlated with vegetation structure and phenology in terms of life form and cover. Third, physiognomic-structural definition sets based on measurable traits such as cover and height allow for validation exercises that can measure these same traits.

SUPERVISED AND UNSUPERVISED CLASSIFIERS

The most general distinction between classifiers is that of supervised versus unsupervised. The former requires training data, a set of land parcels of known cover types that are used to calibrate a classification algorithm. These parcels are identified in the imagery as groups of pixels that are used to develop per class spectral signature statistics. These signatures in turn drive the assignment of the remaining unlabeled pixels in the image. Unsupervised methods, on the other hand, use numerical algorithms to exhaustively cluster the image into spectrally similar groups of pixels. Each pixel in the image is assigned to one of the clusters automatically. These clusters are subsequently labeled by image analysts on the basis of their geographic distribution (Jensen, 1996). In each case, a paramount goal is to minimize the analyst's input. By creating a more repeatable product, the ability to develop baselines for future reference is increased.

The following discussion compares these two fundamental approaches. The first difference is the role of the analyst. A classification is not an automatic procedure. By definition, it relies on an interpreter's skills in labeling either training sites, in the case of supervised classifications, or clusters, in the case of unsupervised classifications. However, the potential to minimize the analyst's role is greatest with the supervised approach. Ideally, the analyst's input for a classification procedure is needed only once. The same training sites can be used repeatedly for successive classifications. For unsupervised approaches, new interpreter input is necessary for each successive clustering, limiting the reproducibility of the map product as compared to the supervised approach. However, the high-quality training needed to perform successive supervised classifications requires significant analysis and data gathering prior to running the algorithm. As a result, many researchers believe

that as the area of study increases, so does the suitability of using an unsupervised classifier because of the difficulty in acquiring reliable training data for larger areas to be used as inputs to a supervised classifier (Cihlar, 2000). Thus, land-cover mapping projects that use moderate and coarse resolution data sets have employed unsupervised techniques more frequently. Exceptions to this include Hansen *et al.* (2000) and Friedl *et al.* (2002), who created global-scale supervised land-cover classifications.

The second difference regards the thematic reference information used in training and labeling for both approaches. For any classification scheme, there are implied spectral boundaries, even if not explicitly stated, which when crossed represent the migration from one class to another. It is the accurate delineation of these boundaries that is sought. For example, if a classification has the goal of mapping tree cover, there is a range of canopy values associated with different tree cover classes. Unsupervised algorithms are based on various clustering regimes around "natural" groupings in multispectral space. In an unsupervised approach, the spectral ranges of these clusters may or may not consistently coincide with the physiognomic-structural limits as represented in the land-cover class legend. The resulting classes yield separable, but not necessarily meaningful classes (Rees, 1990). On the other hand, training data can be delineated that correspond directly to the classes as defined in the land-cover legend. In other words, the supervised approach allows the researcher to delineate training data that correspond to the chosen classification scheme, and does not have to rely on the "natural" spectral groupings to mimic that construct. An example of a "natural" grouping that usually corresponds to a unique multispectral signature is clear water. Few other land-cover classes can be confused with clear water. However, a classification system that breaks up a "natural" grouping along its spectral continuum, such as the example of tree canopy into forest, woodland, and parkland classes, lends itself poorly to a clustering algorithm. The main limitation for the supervised approach is the acquisition of training data, especially for large areas of study. If robust training data exist, a supervised approach is clearly more suitable for characterizing land cover. In the absence of reference data, unsupervised techniques are very useful as a data mining tool that can reveal the underlying distributions in multispectral and map space.

A final consideration is the sensitivity of the algorithm to noise in the data. An algorithm that, in effect, ignores or minimizes the impacts of noise is preferred. Supervised procedures such as classification trees use only the spectral information germane to separating the defined classes. For example, data with cloud contamination will be ignored if other spectral information allows for a separation of the classes in question. The inputs to a clustering algorithm must be preprocessed to eliminate as much noise as possible since all of the data are used in deriving the clusters. If cloudy pixels are left in the inputs of an unsupervised classifier, clusters of this spectral class may well be created. The analyst is left to interpret how to label such classes. Also, some classes, such as bare ground, are highly spectrally variable and an unsupervised classifier will create multiple clusters for what in a given class legend is a single class. Cihlar *et al.* (1998) have shown a way to reduce this problem by creating a parsimonious set of clusters. For many supervised classifiers a simple threshold or two will allow for the characterization of a spectrally variable, yet distinct class such as bare ground (Hansen *et al.*, 1996).

DISTRIBUTION-FREE ALGORITHMS

Parametric classifiers employ assumptions concerning the shape and distribution of decision volumes in multispectral space (Skidmore and Turner, 1988). The assumption of multivariate normality exists for the most used remote sensing classification algorithms, including maximum likelihood and linear discriminant functions. While these classifiers are optimal under conditions of normality, this assumption is rarely valid. As shown in Figure 1, a single class can be represented by a series of spectrally distinct subclasses. Thus, a model of a single data distribution is inappropriate. Distribution-free classifiers, which do not operate under any assumptions of data distributions, can capture the complex intraclass multispectral variability typical of most land-cover types. One of the reasons that normal probability density functions were used to estimate the histograms of training sites originally was that computers were not capable of computing probabilities across all digital counts in the actual histograms (Swain and Davis, 1978). This is not the case today, and classifiers that examine each and every digital number have provided new opportunities to calculate class extent in multispectral space. Advances in algorithm development and computing power have made previously infeasible distribution-free approaches more accessible (Hardin and Thomson, 1992; Paola and Schowengerdt, 1993; Foody, 1995; Ito and Omatu, 1997).

For land cover categories that feature high intraclass spectral variability, unsupervised approaches are superior to heritage supervised parametric algorithms such as maximum likelihood. A useful feature of unsupervised classifiers is the ability to make multiple clusters for a single class. Figure 1 shows a typical problem in characterizing a single land-cover class. An unsupervised algorithm can likely produce clusters that, when aggregated, capture this land-cover type. For a traditional parametric classifier to work, training data for each subcategory must be acquired and distinctly labeled. If all of these spectral subcategories are labeled as a single class, then the assumptions regarding normally distributed data will likely have been violated, leading to suboptimal performance in mapping this or other

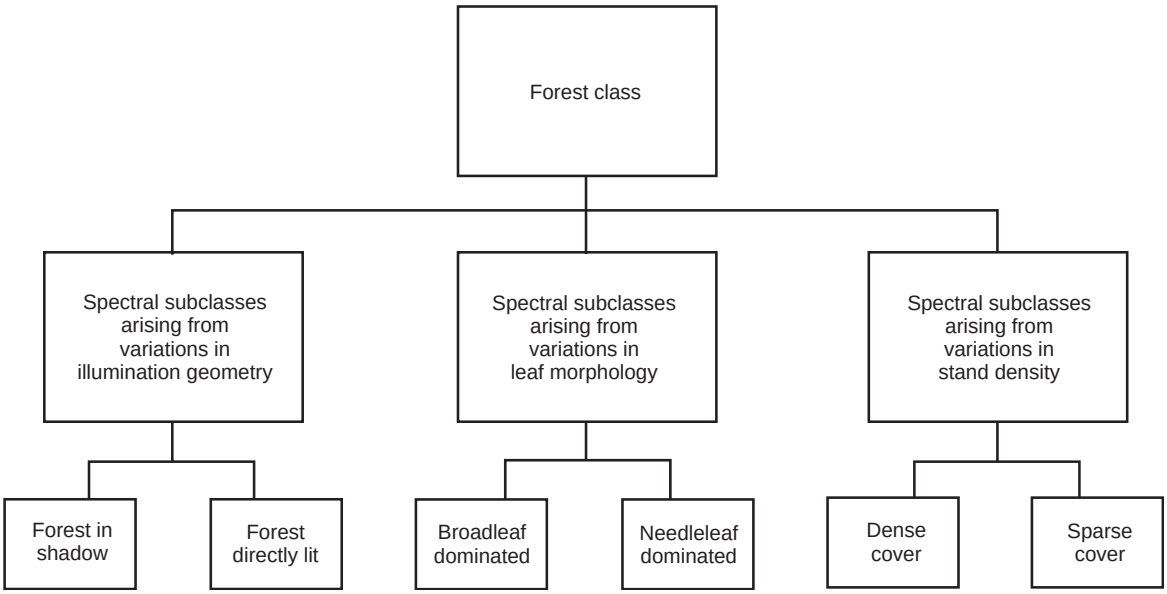


Figure 1 Theoretical intraclass spectral variability of a forest land-cover class, adapted from Campbell (2002)

cover types. However, distribution-free supervised classifiers, such as decision trees and artificial neural networks exist that enforce no distribution assumptions for the input classes. In the training phase, all of the subcategories in Figure 1 can be labeled forest.

A number of comparisons have been made between distribution-free methods and the historically favored parametric techniques. The majority of these comparisons show that no penalty is paid in using distribution-free techniques and that they usually yield higher accuracies. Ince (1987) and Hardin (1994) showed that nearest-neighbor classifiers were superior to parametric methods. Hansen *et al.* (1996) and Friedl and Brodley (1997) found comparable performance between a classification tree approach and a maximum likelihood one. Key *et al.* (1989), Bischof *et al.* (1992) and Gopal *et al.* (1999) tested maximum likelihood versus neural network classifiers and found similar or superior accuracies when using the neural nets.

Three commonly used methods are k-nearest neighbor, decision trees, and neural networks. The generic k-nearest neighbor method assigns each unknown pixel of a satellite image to the cover type of the most spectrally similar k reference training sites using a simple Euclidean or Mahalanobis spectral distance measure (Franco-Lopez *et al.*, 2001). The number k should be large enough to minimize the probability of misclassification and small enough to ensure that the k-nearest neighbors are in fact in close proximity to the candidate pixel. Cross-validation is recommended for finding a suitable value of k (Bishop, 1995). A distance criterion is often added to the algorithm to favorably weight those samples that are closest in spectral space to the candidate pixel.

Figure 2 shows an example of a two-class training set and how different values of k change the decision boundary in a two-dimensional spectral space. As k is increased, the decision boundary is correspondingly generalized. With a k value of 1, each training pixel is used to delineate

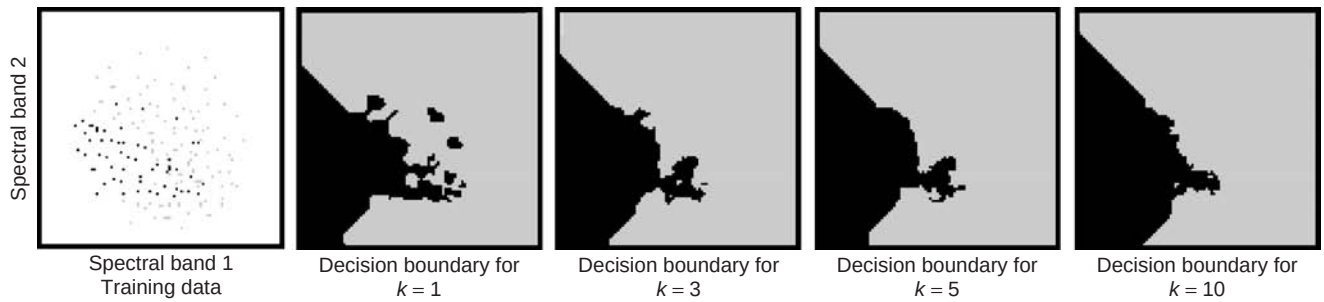


Figure 2 Two-class spectral training data and resultant partitioning of two-dimensional spectral space using a k-nearest neighbor algorithm

the decision boundary. This illustrates not only the great flexibility of distribution-free classifiers in classifying data but also the danger in overfitting the classifier to potential noise or errors in the training data. Choosing the best method to generalize the classifier is often an iterative process and greatly depends on the quality and quantity of reference information.

Decision tree theory (Breiman *et al.*, 1984) has previously been used to classify remotely sensed data sets (Hansen *et al.*, 1996; Friedl and Brodley, 1997; Hansen *et al.*, 2000; Friedl *et al.*, 2002), and offers some advantages over other classification methods. Trees are a hierarchical classifier that predicts class membership by recursively partitioning a data set into more homogeneous subsets. This splitting procedure is followed until a perfect tree (one in which every pixel is discriminated from pixels of other classes, if possible) is created with all pure terminal nodes or until preset conditions are met for terminating the tree's growth. One approach is that of the S-PLUS statistical package (Clark and Pergibon, 1992), which employs a deviance measure to split data into nodes that are more homogeneous with respect to class membership than the parent node. The reduction in deviance (D), is defined as:

$$D = D_s - D_t - D_u \tag{2}$$

where s is the parent node, and t and u are the splits from s. Right and left splits along the digital counts for all data are examined. When D is maximized, the best split has been found, and the data are divided at that digital count and the process repeated on the two new nodes of the tree. The deviance for nodes is calculated from the following:

$$D_i = -2\Sigma n_{ik} \log p_{ik} \tag{3}$$

where n is the number of pixels in class k in node i and p is the probability distribution of class k in node i.

Neural networks are an artificial intelligence technique that transforms feature space into class space and have been employed in many land-cover mapping exercises (Yoshida and Omatu, 1994; Gopal *et al.*, 1999; Kavzoglu and Mather, 2003). Many versions of neural network algorithms exist. A commonly used form is the multilayer perceptron that connects input feature vectors to output class assignments via a series of hidden layers (Atkinson and Tatnall, 1997). The training pattern is fed forward through the network architecture and output labels assigned. Comparisons of the output and input reference labels allow for computing the error, which is then used to adjust the weights of the various connections in an iterative manner to increase the network's accuracy. Increasing the number of hidden layers allows for a more complex network to be derived, but risks losing the ability to successfully generalize feature-label relationships (Foody, 1995). Like the nearest neighbor and classification tree algorithms, overfitting of the model is a potential problem (Abuelgasim *et al.*, 1996).

ALGORITHM COMPARISONS

To illustrate the algorithmic differences between traditional approaches and distribution-free methods, an example data set was analyzed for an area in Colorado, USA. Figure 3 shows a comparison for a set of forest and nonforest training data and examples of how unsupervised clustering, supervised maximum likelihood, and decision tree algorithms vary in partitioning a data set. If classifying only forest and nonforest, the unsupervised clustering approach creates many categories not germane to the task of separating the two land-cover classes. Without training data or a set of *a priori* rules that guide the clustering process, the production of superfluous clusters is often unavoidable. Otherwise, a series of cluster-busting operations must be undertaken to delineate the most accurate decision boundary (Jensen, 1996). The supervised maximum likelihood algorithm, on the other hand, creates a generalized decision boundary related to the assumed normal distributions of the two classes of interest. This creates a rather inflexible model of the actual boundaries between the two multispectral distributions. The distribution-free decision tree algorithm creates what might be called *supervised clusters*, which capture the areas of high and low class membership purity. The strength of this approach is that the clusters or nodes are directly related to the classes of interest, and areas of confusion are well delineated.

The simplest answer as to why superior results have been found using distribution-free classifiers is that a great deal of training data distributions do not conform to the assumptions of multivariate normality. Maximum likelihood classifiers are based on statistics of central tendency and are in many cases poorly suited to delineating class boundaries. Identifying the core area for any given class is, in most cases, the easy part, but correctly characterizing the limits of class memberships is considerably harder. Thus, classifiers that offer the possibility of calculating the edges between classes are most useful. Having a tool that is flexible along these boundaries and conforms to the actual multivariate distributions of the training data often yields superior results.

For distribution-free methods, individual pixels in multispectral space can be identified and included in creating a classification rule set. Thus, the major limitation in using these techniques is the possibility of overfitting the algorithm to the training data. For nearest-neighbor techniques, parameters such as k in the k-nearest neighbor algorithm are used to ensure that isolated and possibly mislabeled reference sites do not adversely affect output map products. The methods of pruning or bagging when using decision tree classifiers (Chan *et al.*, 2001) are also used to

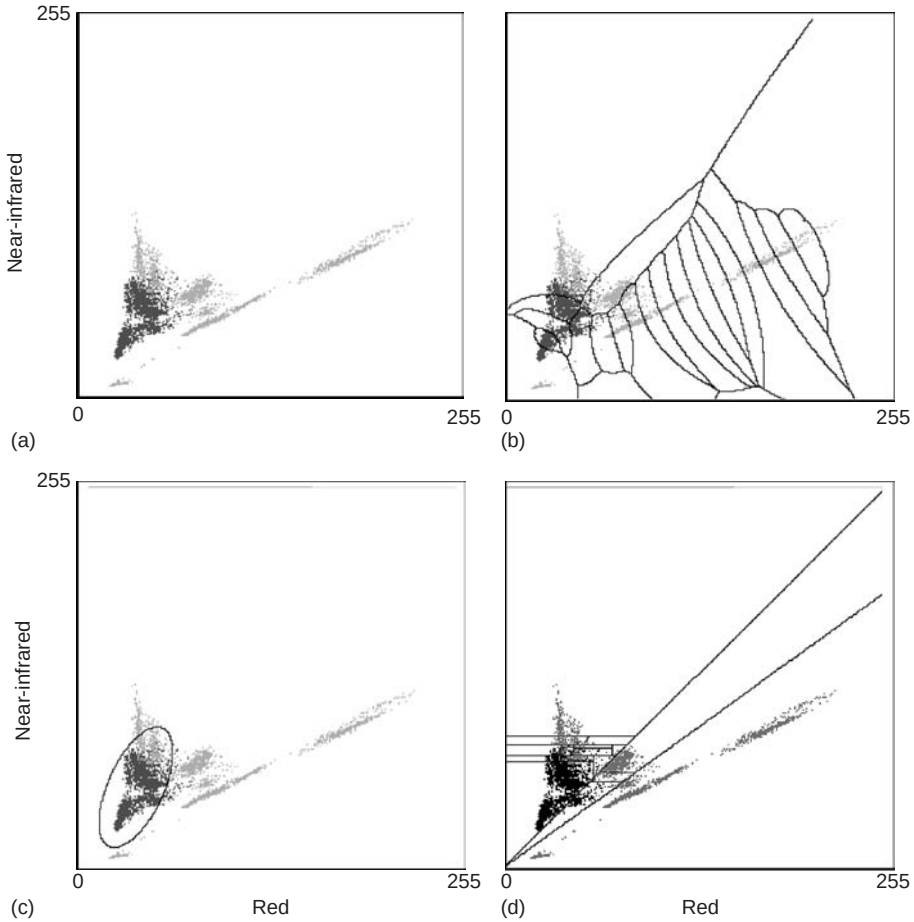


Figure 3 Training data and resultant partitioning of spectral space for data derived using a Landsat Enhanced Thematic Mapper Plus (032/035) image in western Colorado, United States, (a) is the spectral distribution of training data for forest and nonforest classes, where forest consists of broadleaf and needleleaf forests and woodlands of varying illumination conditions and nonforest consists of water, crops, parklands, shrublands and bare ground, (b) is the result of an unsupervised clustering of the data, yielding 20 clusters, with a maximum likelihood rule applied, (c) is the result of a maximum likelihood classifier applied to the training data, and (d) is a decision tree classifier applied to the training data, set to yield 20 terminal nodes

create a generalized output that avoids overfitting. For neural networks, limiting network training times or reducing the number of "hidden" layers (those sets of nodes between the input spectral and output class layers) can help in producing generalized results that best predict unseen cases (Atkinson and Tatnall, 1997).

In choosing which distribution-free methods are the best, there are a few important concerns, foremost of which is performance. Numerous studies have found that there is no single algorithm that is superior in all cases (Curram and Mingers, 1994). For many data sets, there may be a functional equivalency between algorithms with performance being primarily limited by quality of training data. The second concern is ease of use. While algorithms such as nearest neighbor and decision trees have a few, easily understood input parameters, neural networks have a number of obscure and difficult-to-assign input parameters that reduce ease of use (Paola and Schowengerdt, 1993). Another concern is ease of interpretation regarding algorithm outputs. Clearly, establishing the links between the input satellite data and the land-cover assignments is not always easily achieved. The most transparent of the commonly used distribution-free methods is the decision tree. Clear linkages between the splitting rules and class assignments are created, allowing for a ready interpretation from a biophysical standpoint.

CONTINUOUS REPRESENTATIONS OF LAND COVER

Continuous representations of land-cover features are layers in which the percent cover for a given vegetation trait is estimated on a per-pixel basis. These types of maps hold many advantages over land-cover classifications.

Continuously varying depictions of cover provide greater spatial detail than do discrete classifications. Ecotones or highly disturbed and fragmented areas are more accurately depicted by subpixel characterizations that fully exploit the inherent variability found in imagery. This greater thematic detail allows for a better ability to parameterize modeling efforts (DeFries *et al.*, 1995). Unlike classifications, continuous-cover maps allow for the use of consecutive products to measure change. They also allow users to define their own thresholds for land-cover classes. This avoids the common problem of shoehorning one classification legend into another. A major drawback for some methods is the need for precise training data at the subpixel scale. Also, instead of capturing all thematic content in a single map layer, continuous depictions of subpixel cover require data layers for each cover trait of interest.

Much work has been done in calibrating remotely sensed data to fractional vegetation cover. The techniques include (i) using end-member linear mixture modeling (Settle and Drake, 1993; Adams *et al.*, 1995; DeFries *et al.*, 1999); (ii) using fuzzy sets to estimate forest cover (Foody and Cox, 1994); (iii) estimating likelihoods of class membership using logistic regression methods; (iv) empirically estimating percent cover based on calibrating coarse resolution data with high-resolution data (Iverson *et al.*, 1994; Zhu and Evans, 1994; DeFries *et al.*, 1997; Hansen *et al.*, 2002); (v) using high-resolution data to calibrate coarse resolution cover estimates while incorporating the spatial arrangement of the fine-resolution cover estimates (Mayaux and Lambin, 1997); and (vi) employing isolines in scatterplots of red/near-infrared space to map plant density (Jasinski, 1996).

The various approaches can be divided into two general categories, those that rely on exemplar categorical reference data to model the intercategory variation and those that require calibration data along the entire range of mixtures. Examples of the former type include linear mixture models, fuzzy classifiers, and logistic regression approaches. For linear mixture models, a set of spectral endmembers is used to estimate cover fractions for each pixel. The reflectance for any pixel is assumed to be the sum of the reflectances for each endmember within the pixel weighted by the respective proportional covers (DeFries *et al.*, 2000) and takes the following form:

$$R_i = \sum_{j=1}^{Q} r_{ij} x_j + e_i \tag{4}$$

Where R_i is the reflectance in band i, r_{ij} is the reflectance of endmember j in band i, and x_j is the fractional cover of endmember j, e_i is the error term and i is the number of endmembers. The model is further constrained so that the sum of fractional cover estimates sum to 1.

Figure 4 shows an example of a linear mixture model application for monitoring change in coastal wetlands for the Chesapeake Bay area of the Mid-Atlantic United States. For this approach, which employed Landsat Thematic Mapper data, two normalized indices were used to create a transformed spectral space that enabled the labeling of three distinct spectral endmembers: water, vegetation, and soil. The image dates were chosen to account for sources of false change, including phenological and tidal variations. The normalized inputs also accounted for interscene variation in illumination conditions. This enabled the mixture model to detect changes in ponding in coastal marshlands and the derivation of a measure depicting the severity of marshland loss (Kearney *et al.*, 2002).

Limitations to the linear mixture modeling approach include the difficulty in successfully identifying endmembers in multispectral space (Roberts *et al.*, 1998). Also, imputing a linear model upon subpixel, land-cover spectral interactions can often oversimplify the complex spectral relationships of land-cover mixtures, especially at regional and continental scales. It is generally accepted that for small study areas with robust spectral data, usable endmembers can be found for mapping purposes. For larger areas, image stratification can be employed to map regions individually using linear models (Zhu and Evans, 1994). For larger areas where stratification is not possible or not desired, nonlinear models are often preferred.

Fuzzy approaches use classification routines to derive a strength-of-class membership per pixel. The result is a continuous output analogous to posterior probabilities of class membership where the individual class fuzzy membership functions vary from 0 to 1 and sum to 1 for all classes. Fuzzy approaches have been shown to perform comparably to linear mixture models (Foody and Cox, 1994), but have many of the same limitations.

Logistic regression models allow for the use of categorical inputs to create continuous outputs of class likelihoods per pixel. Logistic regressions are a nonlinear transformation of a standard linear regression that employs an *S*-shaped distribution function to constrain estimated output probabilities to between 0 and 1. For remote sensing applications, the dependent variable of class proportion is run against the independent variables of multispectral imagery. The probability of class membership is found by transforming the standard linear regression model with an intercept of a and a slope of b using this function:

$$p = \frac{1}{[1 + \exp(-a - bx)]} \tag{5}$$

where b is now the change in the log odds as a function of x. Figure 5 shows a logistic regression output for mapping intensive mechanized agriculture in an area of Brazil experiencing rapid agro-industrial development. For this method, pixels labeled *crop* and *not crop* are input to the

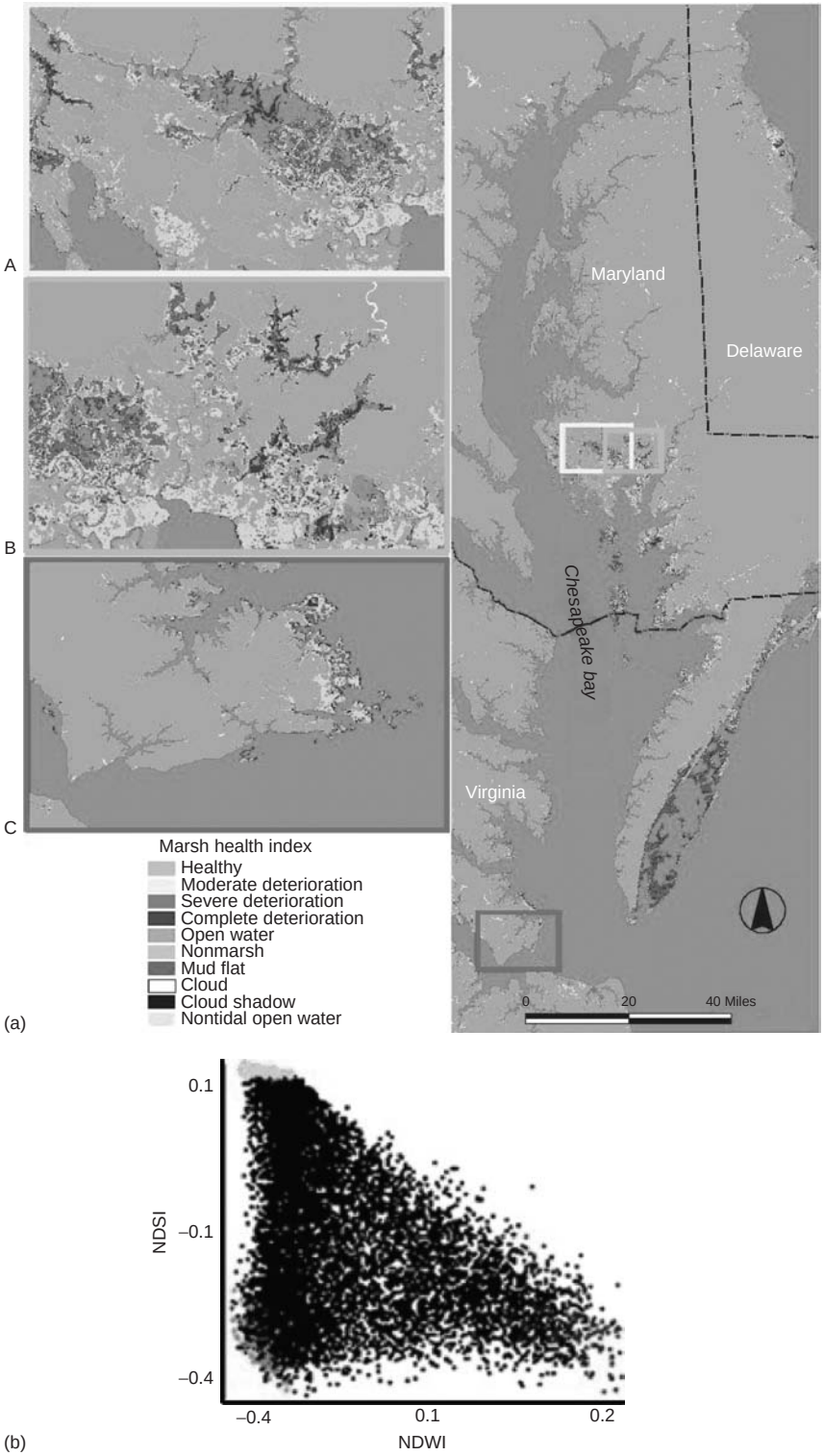


Figure 4 Results of a linear mixture model applied to measure water presence in a coastal marshland environment of the Chesapeake Bay, Mid-Atlantic, United States. (a) Multitemporal application of the model yielded a measure of change in coastal marshlands. (b) Two normalized difference indices, the Normalized Difference Wetness Index ($NDWI = (TM3 - TM5)/(TM3 + TM5)$) and the Normalized Difference Soil Index ($NDSI = (TM5 - TM4)/(TM5 + TM4)$) were used to identify water, vegetation, and soil spectral endmembers as shown (Reproduced from (Kearney *et al.*, 2002) by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

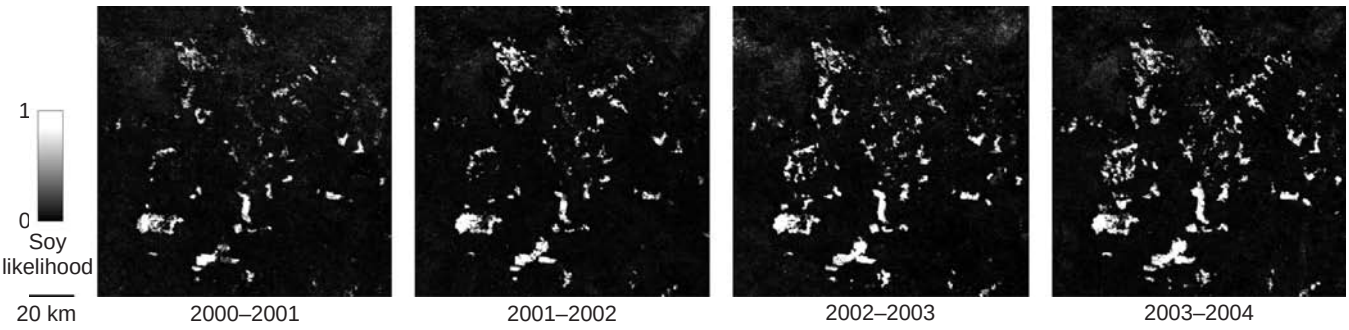


Figure 5 Application of a logistic regression model to map the likelihood of first crop soybean for an area of northeastern Brazil (Tocantins, Maranhao, and Piaui states) experiencing rapid expansion of mechanized agriculture

logistic regression model and a set of regression estimators are derived. These can then be used to derive a likelihood of crop class membership and monitor cropping patterns over time. Logistic regression approaches, like the other aforementioned methods, do not incorporate intermediate cover values in model calibration.

Methods that rely on calibration data at the subpixel level are empirical in nature and include linear techniques such as multivariate linear regression models and nonlinear approaches such as tree-based models. Calibrating empirical models requires significant up-front work using subpixel scale ancillary data sets to create training data over the continuum of cover mixtures. By deriving such a training data set, the true ability to map subpixel cover can be better tested regardless of algorithm choice. However, models that exploit the subpixel information directly in the calibration process should perform better than those that only model subpixel cover variation.

Regression tree models have been used in numerous applications for mapping continuous variables using remotely sensed data (Michaelson *et al.*, 1994; DeFries *et al.*, 1997; Prince and Steininger, 1999; Hansen *et al.*, 2002). The regression tree is a nonlinear tool that recursively splits a continuous training variable into subsets, called *nodes*, that minimize the overall residual sum of squares. The regression tree algorithm takes the form of the following equation:

$$D = \sum_{cases(j)} (y_j - u_{[j]})^2 \qquad (6)$$

where D is the deviance as measured by the corrected sum of squares for a split. This is calculated from all j cases of y and the mean value of those cases, u. The split that minimizes the residual sum of squares is chosen as the best split. For regression trees, the relationship between the independent and dependent variables is not monotonic and nonlinear. This allows for a flexible subsetting of the multispectral image space not feasible with many other methods.

CHANGE DETECTION

Land use and land-cover change includes both the conversion from one land-cover category to another (Riebsame *et al.*, 1994) and the modification, or subtle within-class change, that affects the character of the land cover without changing its overall classification (Coppin *et al.*, 2004). Mapping such changes using remotely sensed data sets offers a unique and powerful tool for measuring land-surface transformations. For hydrological applications, such change can include deforestation mapping within a watershed, or increased ponding within a marshland. The ability to detect land-cover conversions or modifications is a function of the mappability of the classes themselves, the spatial extent of change, and the temporal context in which the change occurs (Singh, 1989). Addressing what cover change dynamics are to be detected is the first order of business. Remote sensing system characteristics, including temporal, spatial, spectral, and radiometric resolutions, dictate which sensors may be of use to a given change detection study. In the temporal domain, change events can vary in terms of persistence, ranging from ephemeral, to interannual, to permanent change. Ephemeral changes are short-term changes in cover such as floods, or seasonal burning in a savanna setting, which do not permanently alter the dominant vegetation cover distribution of the landscape. Interannual changes are variations in land cover largely due to long-term climatic variability such as change in the annual extent of grasslands semiarid regions, or the reduction of woodland canopy cover for an area experiencing long-term drought. Semipermanent/permanent changes are wholesale land-cover conversions and include new construction of impervious surface, deforestation events, or the expansion of agricultural lands. Land-cover modifications, as compared with land-cover conversions, are a form of semipermanent/permanent change within a given land-cover category. This is a more subtle form of change and includes examples such as rangeland degradation due to overgrazing, or forest thinning due to selective logging.

As stated before, the temporal resolution of a sensor is typically inversely related to its spatial resolution. Change events sought on a short-term basis require data from sensors with high temporal revisit rates, which typically have coarse spatial resolutions. Thus, the opportunity to measure change at fine spatial resolutions is governed by a longer interval from time 1 to time 2, as compared with spatially extensive change events. Understanding land-cover change dynamics in the context of sensor capabilities is critical and allows the analyst to construct a defensible land-cover change type definition set. An undeveloped research area is the use of techniques that incorporate multiple sensors in change detection studies. Multiresolution and multitemporal data fusion techniques offer the greatest potential for operational land-cover change monitoring.

The basic principle behind spectral change methods is that variation in spectral signatures between images of different dates indicates land-cover change. However, such variation is not necessarily indicative of land-cover disturbance, and a number of complicating factors must be addressed prior to assigning disturbance using spectral change methods. If data from different sensors are used to find change, then complications arise from differing spectral bandwidths, radiometric, and spatial resolutions. Therefore, it is often recommended to use data from a single sensor. For a given sensing system, spectral variability between dates is often a function of the changes in viewing geometry. Variation in sun-sensor-target geometry can be partially accounted for by obtaining data on anniversary dates from a sensor in a sun-synchronous orbit. In this manner, data are acquired at approximately the same time of day with the same viewing and illumination geometries (Jensen, 1996), minimizing interscene spectral variation unrelated to change. However, for many applications such data are unavailable and these variations must be accounted for by other means, such as the application of an at-sensor or at-surface reflectance correction.

In addition to interscene variations related to sun-sensor-target geometry, atmospheric variability between time periods represents another source of potential error in using spectral change methods. Performing atmospheric corrections for the various effects of water vapor, aerosols, ozone, and Rayleigh scattering is typically not possible unless standard datasets exist that include preprocessed corrections (Vermote *et al.*, 2002). Other procedures to normalize data from multiple images exist, including dark target normalization, histogram matching, and band-to-band regression procedures (Richards, 1993; Teillet and Fedosejevs, 1995; Jensen, 1996). However, many atmospheric effects vary not only between scenes but also within scenes, and this must be considered in evaluating image data for possible use in change detection studies.

Analysts must also recognize the impact of surface environmental factors, wholly unrelated to land-cover change that may alter the spectral character of the surface. Such influences include interscene soil moisture and phenological variation. Apparent land-cover change between dates may simply be the result of images acquired at times of the year when the same vegetation has different spectral properties. For example, mapping deciduous forest in high latitudes requires the acquisition of data from a narrow window of peak cover conditions in July and August. If any of the input images are from other times of the year when the canopy is not exhibiting mature leaf-flush, then spectral change methods will likely lead to false change detection results related to offsets in phenological timing. In most cases, making sure that the input imagery represents common interdate phenological variability is paramount to detecting land-cover conversion events.

There are various approaches to delineating change using remotely sensed data sets. The two general categories of change methods are spectral change and postcharacterization approaches. Spectral change methods use imagery as inputs to detecting land-cover change events, while postcharacterization approaches use derived land-cover map classifications or fractional cover maps as inputs. Commonly used spectral change methods include image algebra, multidate composite, and change vector analyses. Image algebra methods include image ratioing and image differencing, where change pixels are found as outliers in the resultant probability distributions. Since NDVI represents a compression of two bands of spectral information and is a strong correlate with the presence/absence of green vegetation, NDVI image differencing has proven an effective method for identifying land-cover change (Lyon *et al.*, 1998). Identification of change is often performed by thresholding the tails of the differenced image's histogram (Borak *et al.*, 2000).

Multidate composite approaches include algorithms such as Principal Components Analysis (PCA) or unsupervised clustering. If the input images are normalized for variations in environmental factors, then pixels representing change will be captured as unique statistical information. In the case of PCA, the majority of pixels in an image does not change, and as a result they are highly correlated across dates. These pixels represent the majority of the variance in the composite image and are transformed onto the first component, while the change pixels are identified in lesser components (Richards, 1984). Similarly, unsupervised clustering can be used to identify the unique spectral signatures of change pixels in composite images. The major limitation with these methods is the lack of information regarding the labeling of change types.

Change vector analysis incorporates multispectral information to calculate the magnitude and direction of pixel change over time. This approach uses all data layers concurrently and allows for an examination of all change information in multitemporal imagery. As such, it detects both

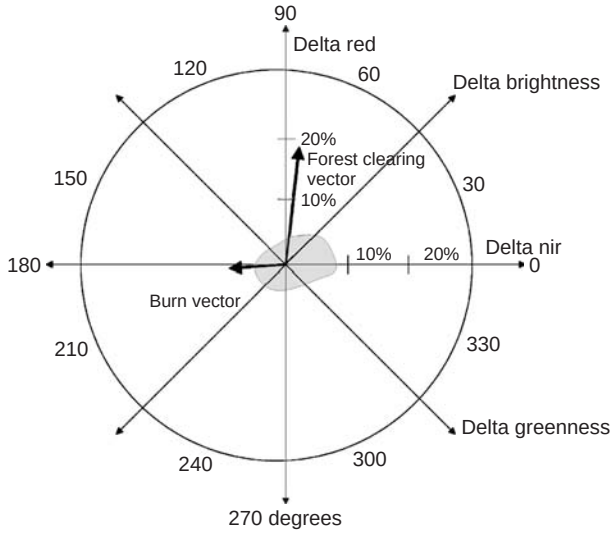


Figure 6 Example of change vector analysis for red and near-infrared spectral bands. Grey area represents time 1 and time 2 interscene spectral variation due to phenological and illumination effects. Vectors extending beyond this area represent actual land-cover change events. In this example, deforestation and burn examples are shown (Reproduced from Zhan *et al.*, 2000 with permission from Taylor & Francis Ltd., `http://www/tandf.co.uk/journals`)

actual land-cover change information as well as interscene background environmental variation. Users of this approach need to account for confounding spurious change classes with actual change classes of interest. Figure 6 shows two change scenarios for the removal of forest canopy cover in change vector space. By accounting for a limit-of-interscene phenological variation for successive images, a threshold of change magnitude and direction can be used to identify specific forest clearing dynamics. For a two band case, as is used in the MODIS change detection algorithm for the 250-m red and near-infrared bands (Zhan *et al.*, 2000), the change vector is calculated as follows:

$$M = \sqrt{(\Delta\rho_{\text{red}})^2 + (\Delta\rho_{\text{nir}})^2} \text{ and } \theta = \arctan\left(\frac{\Delta\rho_{\text{red}}}{\Delta\rho_{\text{nir}}}\right) \quad (7)$$

where M is the change magnitude, θ is the change direction, $\Delta\rho_{\text{red}}$ is the difference in time 1 and time 2 red reflectance and $\Delta\rho_{\text{nir}}$ is the difference in time 1 and time 2 near-infrared reflectance.

Postcharacterization methods compare consecutive land-cover maps to identify change sites. For hard classifications, limitations to the method are a function of the individual time 1 and time 2 classification accuracies. Since change is often a minority occurrence within a given study area, then errors in mapping accuracy confound the ability to delineate actual change. A better approach to detecting change is the use of comparisons of consecutive subpixel

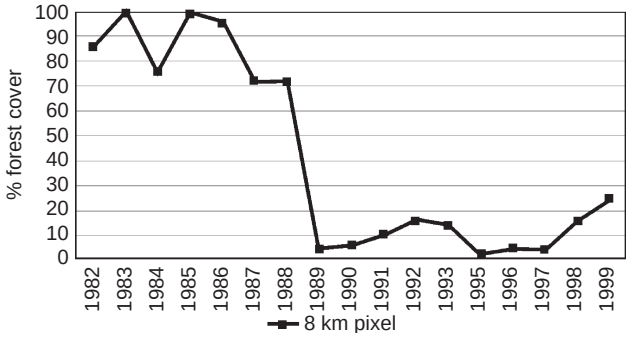


Figure 7 Multitemporal estimates of percent forest cover for an 8-km pixel from the Pathfinder Advanced Very High Resolution Radiometer data set from 1982–1999. Canopy removal due to fire is evident in 1988 within the time series forest cover estimates for this area of Yellowstone National Park, Wyoming, United States (Reproduced from Hansen and DeFries, 2004 by permission of Springer-Verlag)

cover estimates, as shown in Figure 4. Change estimates using fuzzy measures of land cover have been shown to be superior to simple postclassification change detection strategies (Foody, 2001). Since continuous estimates of fractional cover have the same statistical properties of images, approaches such as image differencing can be applied. Figure 7 shows an example of a percent forest cover application that identifies the extensive fire of 1988 in Yellowstone National Park, Wyoming, United States. Here, the biophysical trait of interest, forest cover, is directly characterized and mapped repeatedly through time. Change is found by measuring the ability of the algorithm to map unchanging areas. This measure of algorithm noise is then used with a difference image of consecutive cover estimates to derive a threshold beyond which actual biophysical change in forest area has occurred (Hansen and DeFries, 2004). An advantage of this approach is that the land-cover characterizations can be derived from time-integrated, multitemporal data sets while the change analysis can exploit simple, bitemporal difference image calculations based on the cover estimates.

CASE STUDY–MAPPING LAND COVER FOR ASSESSING STREAM HEALTH IN THE CHESAPEAKE BAY WATERSHED

The links between land cover and water quality, including stream health, have long been known, but not until recently have analyses over large areas emerged as a feasible area of ecohydrological research (Nilsson *et al.*, 2003). Impervious surface areas (ISAs) (like buildings, roads, parking lots) increase the amount of pollutants within and the temperature of runoff reaching streams (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**). Forested riparian buffers, on the other hand, preserve or increase stream

quality by contributing leaf litter, regulating temperature and sunlight, deterring erosion, and resorbing nutrients. In past studies, land cover was developed using aerial photography, but this analysis employs high-resolution satellite imagery and a supervised classification procedure to automatically classify land cover.

The objective in this case study was to utilize fine-resolution land-cover information for a range of small watersheds in suburban Maryland to explore the relationship between stream health and land-cover metrics. The stream health rankings were derived from a combination of physical measurements and macroinvertebrate and fish biological indices developed by state and county collaborators. The land-cover data included ISA and tree cover maps created from very high-resolution satellite data. We used logistic regression models of stream health rankings incorporating the heterogeneous land-cover types and topographic information, and examined the influence of spatial configuration of the landscape variables on predictions of stream health.

Montgomery County, Maryland is part of the Baltimore-Washington metropolitan area, extends over approximately 1300 km^2, and is home to 810 000 people. In this study we analyzed 284 small watersheds (20–47 km^2), which are used as management units (Van Ness *et al.*, 1997). Collaborators from the Montgomery County Department of Environmental Protection (MCDEP) and the Maryland-National Capital Park and Planning Commission (MNCPPC) provided stream health data for the watersheds. They used an Index of Biological Integrity (IBI) for macroinvertebrate species and for fish species based on the ecological assemblages found in reference streams, developed and measured by the Maryland Department of Natural Resources (MD-DNR) (Stribling *et al.*, 1998; Roth *et al.*, 2004). Of the 284 watersheds, 66 were omitted from further analysis due to incomplete assessments, spatial coverage, or excessive cloud cover at the time of satellite image acquisitions. The remaining 218 watersheds included 31 ranked as having excellent stream health, 77 good, 68 fair, and 42 poor (Figure 8).

Space Imaging IKONOS "precision georeferenced" imagery acquired between 6 April 2000 and 23 May 2001 were requisitioned through the NASA (National Aeronautics and Space Administration) Scientific Data Purchase program. The imagery covered a 1313-km^2 area including Montgomery County, at 4-m resolution in four multispectral wavelength bands (blue, green, red, and near-infrared). The images included a mixture of leaf-on (foliage expanded) and leaf-off (foliage absent) imagery acquired between early and late spring. Masks of clouds and cloud shadows for those areas affected were created manually by delineating affected areas in the imagery.

Maps of tree and impervious surface cover (Figure 8) were derived from the imagery using a decision tree classification based on the individual wavelength bands and spectral vegetation indices, to discriminate impervious and tree covered areas. The classification trees were created with S-PLUS statistical software, which uses a univariate algorithm to recursively threshold the training data into homogeneous groupings. The training data for the ISA classification were vector planimetric coverages from the MNCPPC, including building and road footprints. The training data for the tree cover map were forest areas (>60% tree cover) mapped in 1992 using a relatively large minimum mapping unit that captured densely forested areas. Both the training data sets were created from visual interpretation of aerial photography, and were preprocessed for accuracy to account for areas that had changed. The final tree and ISA maps capture changes in the land cover since 1992 and show more detail because of the fine spatial resolution of the imagery. Details on the creation and assessment of the tree cover and impervious surface maps from IKONOS imagery are reported by Goetz *et al.* (2003).

A DEM was created from a topographic map provided by Montgomery County. The vertical and horizontal resolution of the topographic map (<1 foot) was used to create a 4-m resolution DEM, from which a grid of the percent change in slope between the cells was derived. This and a vector hydrology layer provided by the county was used to create 30 m (100 ft) riparian buffers, since this is a common metric used for determining Chesapeake Bay restoration efforts. The tree cover map, the (4 m) ISA map, and the percent change in slope map were intersected with the buffer coverage to calculate for each watershed buffer zone, the proportion of tree cover, ISA, and mean slope (Figure 8). Spatial configuration metrics, such as the distance from the stream to ISA areas and land cover within flow paths, were also calculated and used in the statistical analyses.

Logistic regression models were developed to predict stream health rankings using the land-cover metrics within the watershed and the riparian buffer zones as the independent variables, and the stream health rankings (excellent, good, fair, poor) as the categorical dependent variables. Values in the logistic regression model are chosen using maximum likelihood to predict the probability of a given category (in our case stream health ranking). We allowed the logistic models to include in a stepwise fashion those variables most significant as predictors of stream health. Because ISA and tree cover relate to stream health in different ways, they were both included in the logistic models, but stepwise elimination tests were done to assess their relative importance.

The amount of ISA derived from the fine-resolution map (4 m) varied from 0 to 52%, while the percentage of tree cover varied from 0 to 94% (Figure 8). When the watersheds were grouped by stream health ranking, it was apparent that the average percent ISA increased from excellent to poor rankings, while the average percent tree cover

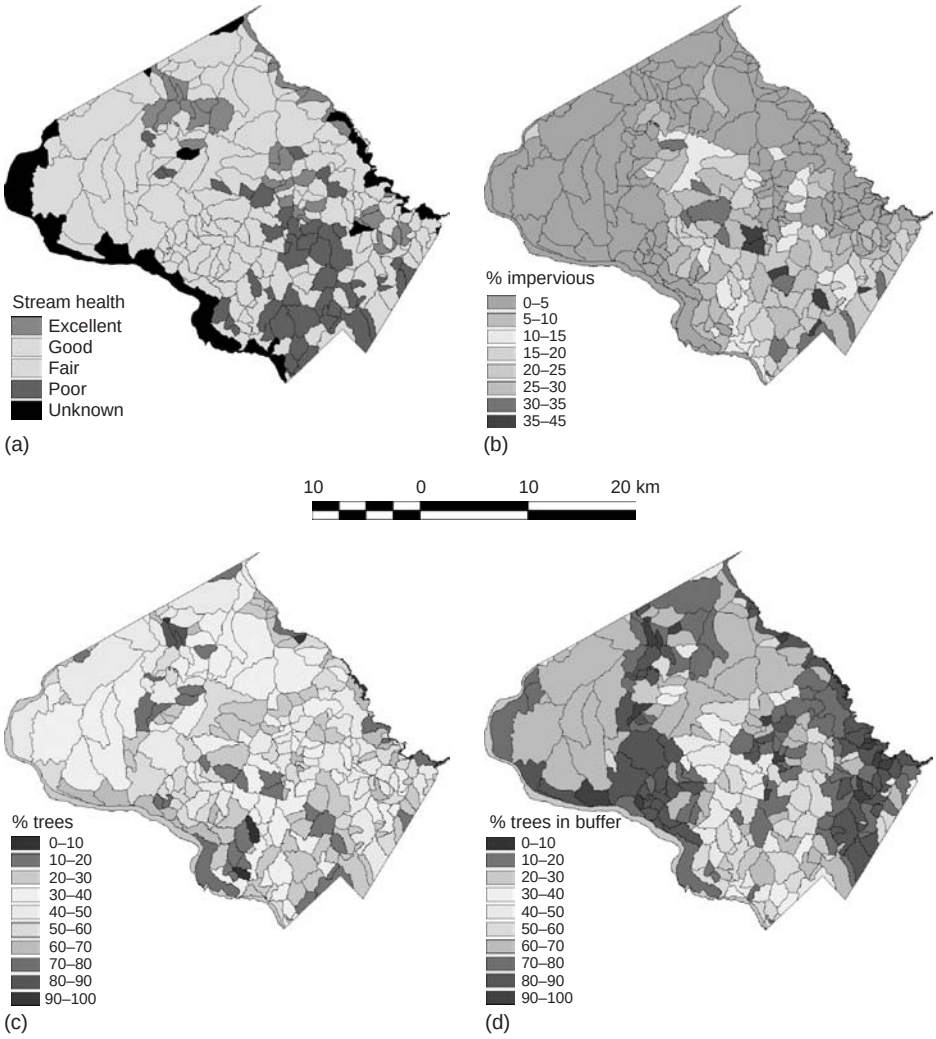


Figure 8 Maps of watershed aggregated values for: (a) stream health rankings, (b) impervious surface cover (%), (c) tree cover (%), (d) riparian buffer tree cover (%). Figures b, c and d were calculated using the IKONOS derived maps of these land-cover variables

Table 3 Small watershed sample size and average statistics by stream health rating category

Stream health rating	N	Area (km^2)	Impervious (%)	Tree cover(%)	Buffered (%)
Excellent	38	272	3.6	50.6	76.8
Good	81	658	4.9	44.6	71.3
Fair	76	451	13.9	37.0	63.2
Poor	50	356	19.5	29.6	56.3

decreased (Table 3). In the riparian buffer zone, tree cover ranged from 0 to 98%, the ISA varied from 0 to 32%, and the mean percentage slope from 3 to 26% (Table 3). The percentage of the buffer occupied by tree decreased and the percentage of ISA in the buffer increased with stream ranking, although a slight but insignificant decrease in percent ISA in the buffer occurred between the excellent and good rankings.

Measurements across scales were correlated, thus not independent estimators, including the percentage ISA in a watershed and in the buffer zone ($r = 0.79$) and the percent tree cover in the watersheds and in the buffer zones ($r = 0.74$). Moreover, the ISA and the tree cover area in the watersheds were inversely correlated ($r = -0.59$); thus as the percent ISA in a watershed increased, the percent tree cover decreased. Logistic regression minimizes these autocorrelation effects, however, and the percent ISA in the watersheds was consistently selected as the primary predictive variable ($r^2 = 0.33$, $p < 0.0001$). This was followed by the percentage tree cover in the buffer zone ($r^2 = 0.35$, $p = 0.04$), and then by other landscape variables such as slope and spatial configuration metrics. Note that r^2 terms from categorical logistic models are considered good fits when they exceed 0.2.

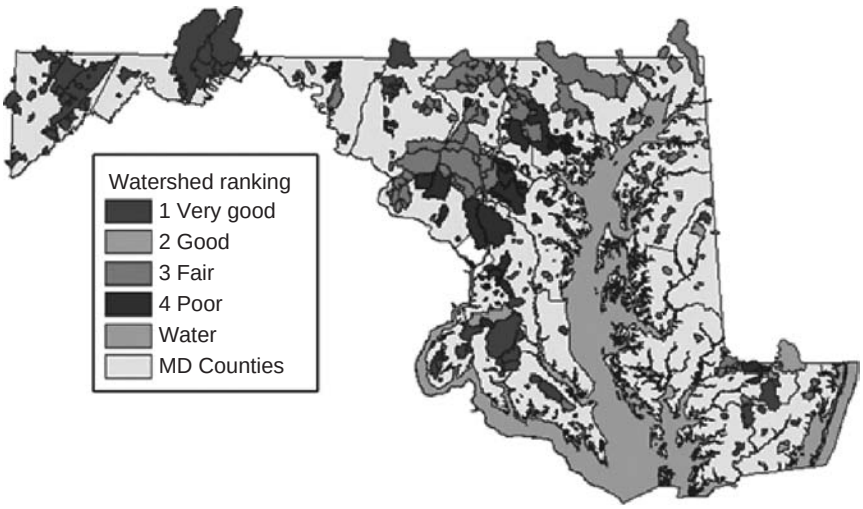


Figure 9 Predicted stream health ratings in some Maryland watersheds based on land-cover characteristics derived from Landsat imagery and the threshold criteria developed using the finer-scale IKONOS analysis results

Interpretation of the results indicates that the processes most affecting stream integrity in these watersheds are those that act primarily at the watershed scale. Several studies have noted that the effect of urbanization on biotic communities is evident between 10% and 15% ISA (Schueler, 1994; Arnold and Gibbons, 1996). The results indicate that watersheds in excellent health averaged less than 8% ISA, watersheds in good health averaged $<$10% ISA, those rated fair averaged $<$20% ISA, and those with a poor health ranking had $>$29% ISA. A few watersheds in this study had greater than 15% ISA but were still rated excellent, indicating that landscape configuration was probably important (the subject of work described by Snyder *et al.*, 2004). These ISA thresholds reflect statistically significant differences between rankings with some precision, as they were calculated from large sample sizes across a wide range of watersheds studied, at fine spatial resolution.

Many management agencies now have access to finer-resolution land-use data, and the analyses indicate that accurate impervious coefficients associated with land-use classes may provide a comparable level of predictive capability as the 4-m product developed and used here. For areas where an existing land-use map is not available, continuous 30-m Landsat land-cover maps of impervious and tree cover, like those recently produced for the Chesapeake Bay watershed (Goetz *et al.*, 2004) and elsewhere (Yang *et al.*, 2003), may be equally adequate. For example, using the continuous Landsat maps and the criteria developed in Montgomery County, the analysis assessed watersheds across the state of Maryland (Figure 9) and identified those that should be targeted for protection (excellent and good rankings) versus those that require restoration (fair or poor rankings). These results demonstrate the utility of impervious surface data for the development of guidelines relevant to stream health, and indicate that minimizing the amount of ISA in a watershed, or mitigating their negative impacts through various management practices, will aid the preservation, targeting, and restoration of stream health.

SUMMARY

Various data sets and tools are available for mapping land cover and land-cover change for inputs to hydrologic applications. The best choice is often a function of the specific application. New sensors offer increased capability in monitoring spatial detail and temporal variation of land-cover categories. Algorithms have advanced in robustness and include distribution-free methods that are superior to traditional approaches in terms of modeling the multispectral distributions of reference data. Advanced subpixel methods of mapping land cover offer greater thematic coherency and the possibility of using consecutive land-cover characterizations to map change. The success of deriving a meaningful result, such as deriving a relationship between aquatic biotic indices and land-cover change, relies on a high-quality land-cover reference map. By choosing the most appropriate data sources, constructing a defensible land-cover definition set, and employing robust algorithms, analysts allow for the meaningful incorporation of land-cover map information into hydrological studies.

REFERENCES

Abuelgasim A., Gopal S., Irons J. and Strahler A. (1996) Classification of ASASmultiangle and multispectral measurements using artificial neural networks. *Remote Sensing of Environment*, **57**, 79–87.

Adams J.B., Sabol D.E., Kapos V., Filho R.A., Roberts D.A., Smith M.O. and Gillespie A.R. (1995) Classification of multispectral images based on fraction endmembers: application to land-cover change in the Brazilian Amazon. *Remote Sensing of Environment*, **52**, 137–154.

Arnold C.L. and Gibbons C.J. (1996) Impervious surface: the emergence of a key urban environmental indicator. *American Planning Association Journal*, **62**, 243–258.

Atkinson P.M. and Tatnall A.R.L. (1997) Neural networks in remote sensing. *International Journal of Remote Sensing*, **18**, 699–709.

Bischof H., Schneider W. and Pinz A.J. (1992) Multispectral classification of Landsat images using neural networks. *IEEE Transactions on Geoscience and RemoteSensing*, **30**, 482–490.

Bishop C.M. (1995) *Neural Networks and Pattern Recognition*, Oxford University Press: Oxford, p. 504.

Bonan G.B., Levis S., Kergoat L. and Oleson K.W. (2002) Landscapes as patches of plant functional types: an integrating concept for climate and ecosystem models. *Global Biogeochemical Cycles*, **16**, 1021–1051.

Borak J.S., Lambin E.F. and Strahler A.S. (2000) The use of temporal metrics for land cover change detection at coarse spatial scales. *International Journal of Remote Sensing*, **21**(6&7), 1415–1432.

Breiman L., Friedman J., Olshen R. and Stone C. (1984) *Classification and Regression Trees*, Wadsworth: Monterey.

Campbell J.B. (2002) *Introduction to Remote Sensing*, Guilford Press: New York, p. 621.

Chan J.C.-W., Huang C. and DeFries R.S. (2001) Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting. *IEEE Transactions of Geoscience and Remote Sensing*, **39**(3), 693–695.

Cihlar J. (2000) Land cover mapping of large areas from satellites: status and research priorities. *International Journal of RemoteSensing*, **21**, 1093–1114.

Cihlar J., Qinghan Xiao, Chen J., Beaubien J., Fung K. and Latifovic R. (1998) Classification by progressive generalization: a new automated methodology for remote sensing multichannel data. *International Journal of Remote Sensing*, **19**, 2685–2704.

Clark L.A. and Pergibon D. (1992) Tree-based models. In *Statistical Models in S*, Hastie T.J. (Ed.), Wadsworth and Brooks: Pacific Grove,.

Congalton R.G. and Green K. (1999) *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers: Boca Raton, p. 137.

Coppin P., Jonckheere I., Nackaerts K., Muys B. and Lambin E. (2004) Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, **9, 25**, 1565–1596.

Curram S.P. and Mingers J. (1994) Neural networks, decision tree induction and discriminant analysis: an empirical comparison. *Journal of the Operational Research Society*, **45**, 440–450.

DeFries R., Field C.B., Fung I., Justice C.O., Matson P.A., Matthews M., Mooney H.A., Potter C.S., Prentice K., Sellers P.J., *et al.* (1995) Mapping the land surface for global atmosphere-biosphere models: toward continuous distributions of vegetation's functional properties. *Journal of Geophysical Research*, **100**, 20 867–20 882.

DeFries R.S., Hansen M., Steininger M., Dubayah R., Sohlberg R. and Townshend J. (1997) Subpixel forest cover in Central Africa from multisensor, multitemporal data. *Remote Sensing of Environment*, **60**, 228–246.

DeFries R., Hansen M. and Townshend J. (2000) Global continuous fields of vegetation characteristics: a linear mixture model applied to multiyear 8 km AVHRR data. *International Journal of Remote Sensing*, **21**, 1389–1414.

DeFries R.S., Townshend J.R.G. and Hansen M.C. (1999) Continuous fields of vegetation characteristics at the global scale at 1-km resolution. *Journal of Geophysical Research*, **104**, 16 911–16 923.

DeGrandi G.F., Mayaux P., Maingreau J.P., Rosenqvist A., Saatchi S. and Simard M. (2000) New perspectives on global ecosystems from wide-area radar mosaics: flooded forest mapping in the tropics. *International Journal of Remote Sensing*, **21**, 1235–1249.

Dickinson R.E. (1995) Land processes in climate models. *Remote Sensing of Environment*, **51**, 27–38.

Dickinson R.E., Henderson-Sellers A., Kennedy P.J. and Wilson M.F. (1986) *Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model*, Technical Note NCAR/TN-275 + STR, National Center for Atmospheric Research: Boulder, p. 69.

DiGregorio A. and Jansen L. (2000) *Land Cover Classification System (LCCS): Classification Concepts and User Manual*, Food and Agricultural Organization of the United Nations: Rome, p. 179.

Dubayah R.O., Wood E.F., Engman E.T., Czajkowski K.P., Zion M. and Rhoads J. (2000) Remote sensing in hydrological models. In *Remote and Sensing in Hydrology and Water Management*, Schultz G.A. and Engman E.T. (Eds.), Springer Verlag: New York, p. 483.

Foley J.A., Kucharik C.J., Donner S.D., Twine T.E. and Coe M.T. (2003) *Land Use, Land Cover and Climate Change Across the Mississippi Basin: Impacts on Land Use and Water Resources*, AGU Chapman Monograph Series, DeFries R., Asner G. and Houghton S. (Eds.), AGU.

Foody G.M. (1995) Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*, **17**, 1317–1340.

Foody G.M. (2001) Monitoring land cover change along the southern limits of the Sahara. *Photogrammetric Engineering and Remote Sensing*, **67**(7), 841–847.

Foody G. and Cox D. (1994) Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. *International Journal of Remote Sensing*, **15**, 619–631.

Franco-Lopez H., Ek A.R. and Bauer M.E. (2001) Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, **77**, 251–274.

Friedl M.A. and Brodley C.E. (1997) Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, **61**, 399–409.

Friedl M.A., McIver D.K., Hodges J.C.F., Zhang X.Y., Muchoney D., Strahler A.H., Woodcock C.E., Gopal S., Schneider A., Cooper A., *et al.* (2002) Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, **83**, 287–302.

Goetz S.J., Jantz C.A., Prince S.D., Smith A.J., Wright R. and Varlyguin D. (2004) Integrated analysis of ecosystem interactions with land use change: the Chesapeake Bay watershed. In *Ecosystem Interactions with Land Use Change*, Asner G.P., DeFries R.S. and Houghton R.A. (Eds.), American Geophysical Union: Washington DC.

Goetz S.J., Wright R.K., Smith A.J., Zinecker E. and Schaub E. (2003) IKONOS imagery for resource management: tree cover, impervious surfaces and riparian buffer analyses in the Mid-Atlantic region. *Remote Sensing of the Environment*, **88**(1/2), 195–208.

Gopal S., Woodcock C. and Strahler A. (1999) Fuzzy ARTMAP classification of global land cover from the 1 degree AVHRR data set. *Remote Sensing of Environment*, **676**, 230–243.

Gorte B.G.H. (2000) Land-use and catchment characteristics. In *Remote and Sensing in Hydrology and Water Management*, Schultz G.A. and Engman E.T. (Eds.), Springer Verlag: New York, p. 483.

Hansen M.C. and DeFries R.S. (2004) Detecting long term global forest change using continuous fields of tree cover maps from 8 km AVHRR data for the years 1982–1999. *Ecosystems*, **7**, 695–722..

Hansen M.C., DeFries R.S., Townshend J.R.G. and Sohlberg R. (2000) Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, **21**, 1331–1364.

Hansen M.C., DeFries R.S., Townshend J.R.G., Sohlberg R., Carroll M. and DiMiceli C. (2002) Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data. *Remote Sensing of Environment*, **83**(1&2), 303–319.

Hansen M., Dubayah R. and DeFries R. (1996) Classification trees: an alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, **17**(5), 1075–1081.

Hardin P.J. (1994) Parametric and nearest-neighbor methods for hybrid classification: a comparison of pixel assignment accuracy. *Photogrammetric Engineering and Remote Sensing*, **60**, 1439–1448.

Hardin P.J. and Thomson C.N. (1992) Fast nearest neighbor classification models for multispectral imagery. *Professional Geographer*, **44**, 191–201.

Harrington J.A. Jr, Schiebe F.R. and Nix J.F. (1992) Remote sensing of Lake Chicot, Arkansas: monitoring suspended sediment, turbidity, and Secchi depth with Landsat MSS data. *Remote Sensing of Environment*, **39**, 15–27.

Holben B.N. (1986) Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing*, **12**, 1147–1163.

Ince F. (1987) Maximum likelihood classification, optimal or problematic? a comparison with the nearest neighbor classification. *International Journal of Remote Sensing*, **8**, 1829–1838.

Ito Y. and Omatu S. (1997) Category classification method using a self-organizing neural network. *International Journal of Remote Sensing*, **18**, 829–845.

Iverson L.R., Cook E.A. and Graham R.L. (1994) Regional forest cover estimation via remote sensing: the calibration center concept. *Landscape Ecology*, **9**, 159–174.

Jasinski M.F. (1996) Estimation of subpixel vegetation density of natural regions using satellite multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **34**, 804–813.

Jensen JohnR. (1996) *Introductory Digital Image Processing: A Remote Sensing Perspective*, *Second Edition*, Prentice-Hall: Upper Saddle River, p. 318.

Justice C.O., Markham B.L., Townshend J.R.G. and Kennard R.L. (1989) Spatial degradation of satellite data. *International Journal of Remote Sensing*, **10**, 1539–1561.

Justice C.O., Townshend J.G.R., Holben B.N. and Tucker C.J. (1985) Analysis of the phenology of global vegetation using meteorological satellite data. *International Journal of Remote Sensing*, **6**, 1271–1318.

Kavzoglu T. and Mather P.M. (2003) The use of backpropagating artificial neural networks in land cover classification. *International Journal of Remote Sensing*, **24**, 4907–4939.

Kearney M.S., Rogers A.S., Townshend J.R.G., Rizzo E., Stutzer D., Stevenson J.C. and Sundborg K. (2002) Landsat imagery shows decline of coastal marshes in Chesapeake and Delaware Bays. *EOS Transactions of the American Geophysical Union*, **16, 83**, 173–178.

Key J., Maslanik J.A. and Schweiger A.J. (1989) Classification of merged AVHRR and SMMR Arctic data with neural networks. *Photogrammetric Engineering and Remote Sensing*, **55**, 1331–1338.

Lambin E.F. and Ehrlich D. (1996) The surface temperature – vegetation index space for land cover and land-cover change analysis. *International Journal of Remote Sensing*, **17**(3), 463–487.

Liang X., Wood E.F. and Lettenmaier D.P. (1999) Modeling ground heat flux in land surface parameterization schemes. *Journal of Geophysical Research*, **104**(D8), 9581–9600.

Lillesand T.M. and Kieffer R.W. (1994) *Remote Sensing and Image Interpretation*, *Third Edition*, John and Wiley and Sons: New York, p. 750.

Lyon J.G., Yuan D., Lunetta R.S. and Elvidge C.D. (1998) A change detection experiment using vegetation indices. *Photogrammetric Engineering and Remote Sensing*, **64**, 143–150.

Mayaux P. and Lambin E.F. (1997) Tropical forest area measured from global land cover classifications: Inverse calibration models based on spatial textures. *Remote Sensing of Environment*, **59**, 29–43.

Michaelson J., Schimel D.S., Friedl M.A., Davis F.W. and Dubayah R.O. (1994) Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science*, **5**, 673–696.

Moody A. and Strahler A.H. (1994) Characteristics of composited AVHRR data and problems in their classification. *International Journal of Remote Sensing*, **15**, 3473–3491.

Nemani R., Pierce L., Running S. and Goward S. (1993) Developing satellite-derived estimates of surface moisture status. *Journal of Applied Meteorology*, **32**(3), 548–557.

Nilsson C., Pizzuto J.E., Moglen G.E., Palmer M.A., Stanley E.H., Bockstael N.E. and Thompson L.C. (2003) Ecological forecasting and the urbanization of stream ecosystems: challenges for economists, hydrologists, geomorphologists, and ecologists. *Ecosystems*, **6**(7), 659–674.

Paola J.D. and Schowengerdt R.A. (1993) *A Review and Analysis of Neural Networks for Classification of Remotely Sensed Multispectral Imagery*, Research Institute for Advanced Computer Science.

Pauwels V.R.N. and Wood E.F. (1999) A soil-vegetation-atmosphere transfer scheme for the modeling of water and energy balance processes in high latitudes, 1. *Model Improvements, Journal of Geophysical Research*, **104**, 27 811–27 822.

Pauwels V.R.N. and Wood E.F. (2000) The importance of classification differences and spatial resolution of land cover data in the uncertainty of model results over boreal ecosystems. *Journal of Hydrometeorology*, **1**, 255–266.

Price J.C. (1984) Land surface temperature measurements from the split window channels of the NOAA 7 advanced very high resolution radiometer. *Journal of Geophysical Research*, **89**, 7231–7237.

Prince S.D. and Steininger M.K. (1999) Biophysical stratification of the Amazon basin. *Global Change Biology*, **5**, 1–22.

Ramsey E.W. III (1999) Radar remote sensing of wetlands. In *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications*, Lunetta R.S. and Elvidge C.D., (Eds.), Ann Arbor Press: Chelsea, p. 318.

Rees W.G. (1990) *Physical Principles of Remote Sensing*, CambridgeUniversity Press: Cambridge, p. 372.

Richards J.A. (1984) Thematic mapping from multispectral image data using the principal components transformation. *Remote Sensing of Environment*, **16**, 35–46.

Richards J.A. (1993) *Remote Sensing Digital Image Analysis*, Springer Verlag: Heidelberg, p. 363.

Riebsame W.E., Parton W.J., Galvin K.A., Burke I.C., Bohren L.C., Young R. and Knop E. (1994) Integrated modeling of land use and cover change. *BioScience*, **44**(5), 350–356.

Roberts D.A., Gardner M., Church R., Ustin S., Scheer G. and Green R.O. (1998) Mapping Chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models. *Remote Sensing of Environment*, **65**, 267–279.

Roth N.E., Southerland M.T., Rogers G.M. and Vølstad J.H. (2004) *Maryland Biological Stream Survey 2000–2004. Volume III: Ecological Assessment of Watersheds Sampled in 2002*. Report CBWP-MANTA-EA-04-1, **http://www.dnr.state.md.us/streams/pubs/ea04-1_data.pdf**, Maryland Department of Natural Resources, Annapolis, p. 318, Accessed in April 2004.

Schueler T.R. (1994) The Importance of Imperviousness. *Watershed Protection Techniques*, **1**(3), 100–111.

Schumann A.H. and Schultz G.A. (2000) Detection of land cover change tendencies and their effect on water management. In *Remote and Sensing in Hydrology and Water Management*, Schultz G.A. and Engman E.T. (Eds.), Springer Verlag: New York, p. 483.

Sellers P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A simple biosphere model (SiB) for use within general circulation models. *Journal of Atmospheric Science*, **43**, 505–531.

Settle J. and Drake N.A. (1993) Linear mixing and the estimation of ground cover proportions. *International Journal of Remote Sensing*, **14**, 1159–1177.

Singh A. (1989) Digital change detection techniques using remotely sensed data. *International Journal of Remote Sensing*, **10**, 989–1003.

Skidmore A.K. and Turner B.J. (1988) Forest mapping accuracies are improved using a supervised nonparametric classifier with SPOT data. *Photogrammetric Engineering and Remote Sensing*, **54**, 1415–1421.

Snyder M., Goetz S.J. and Wright R. (2004) Stream health rankings predicted by satellite-derived land cover metrics: impervious area, forest buffers and landscape configuration. *Journal of the American Water Resources Association*, (in press).

Stribling J.B., Jessup B.K., White J.S., Boward D. and Hurd M. (1998) *Development of a benthic index of biotic integrity for Maryland streams*, Maryland Department of Natural Resources: Annapolis, p. 62.

Swain P.H. and Davis S.M. (Eds.) (1978) *Remote Sensing: The Quantitative Approach*, New York: McGraw-Hill Book Company.

Teillet P.M. and Fedosejevs G. (1995) On the dark target approach to atmospheric correction of remotely-sensed data. *Canadian Journal of Remote Sensing*, **21**, 374–387.

Townsend P.A. and Walsh S.J. (1998) Modeling floodplain inundation using an integrated GIS with radar and optical remote sensing. *Geomorphology*, **21**, 295–312.

Tucker C.J. (1979) Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sening of Environment*, **8**, 127–150.

Van Ness K., Brown K., Haddaway M., Marshall D. and Jordahl D. (1997) *Montgomery County Water Quality Monitoring Program: Stream Monitoring Protocols*, Watershed Management Division, Department of Environmental Protection: Montgomery County.

Vermote E.F., El Saleous N. and Justice C.O. (2002) Atmospheric correction of the MODIS data in the visible to middle infrared: first results. *Remote Sensing of Environment*, **83**(1–2), 97–111.

Wickham J.D., Ritters K.H., O'Neill R.V., Reckhow K.H., Wade T.G. and Jones K.B. (2000) Land Cover as a framework for assessing risk of water pollution. *Journal American Water Resources Association*, **36**, 1417–1422.

Yang L., Xian G., Klaver J.M. and Deal B. (2003) Urban land cover change detection through sub-pixel imperviousness mapping using remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, **69**(9), 1003–1010.

Yoshida T. and Omatu S. (1994) Neural network approach to land cover mapping. *IEEE Transactions on Geoscience and Remote Sensing*, **32**(5), 1103–1108.

Zhan X., DeFries R., Townshend J.R.G., DiMiceli C., Hansen M., Huang C. and Sohlberg R. (2000) The 250 m global land cover change product from the moderate

resolution imaging spectroradiometer of NASA's Earth observing system. *International Journal of Remote Sensing*, **21**, 1433–1460.

Zhu Z. and Evans D.L. (1994) U. S. forest types and predicted percent forest cover from AVHRR data. *Photogrammetric Engineering and Remote Sensing*, **60**, 525–531.

58: Characterizing Forest Canopy Structure and Ground Topography Using Lidar

RALPH DUBAYAH[1], BIRGIT PETERSON[1], JOSHUA RHOADS[1] AND WILLIAM E DIETRICH[2]

[1]*Department of Geography, University of Maryland, College Park, MD, US*
[2]*Department of Earth and Planetary Science, University of California, Berkeley, CA, US*

Remote sensing observations increasingly are used to obtain the detailed information needed about land surface state required for hydrological analyses. However, many of the surface state parameters required for such analyses are related to the vertical structures of surface vegetation and topography, that are often difficult to measure using passive optical and radar remote sensing technologies. A new technology, lidar (light detection and ranging) remote sensing, has proven highly effective for characterizing land surface structure in great detail, including subcanopy topography, canopy height, foliar profile and biomass, among others. In this article, we review the theory and application of lidar remote sensing for characterizing land surface states for hydrological analysis. First we present a brief overview of lidar, and discuss similarities and distinctions between the small-footprint systems commonly used in commercial applications and large-footprint approaches used in research and space-based systems. We next describe common land surface variables that are often used in hydrological analysis and how these are related to vertical structures observable from lidar. Lastly, we present how lidar may be used to recover or parameterize these surface variables; in particular we focus on the observation of forest and topographic structures.

INTRODUCTION

Remote sensing data are used increasingly for the parameterization, initialization, and validation of hydrological models. The creation of consistent, long-term, global and regional scale remote sensing observational records, along with advances in remote sensing technologies, algorithms, and data assimilation methods have greatly advanced our ability to predict accurately hydrologic fluxes. However, many of the land surface state parameters required for such analyses are related to ground topography and the vertical structure of surface vegetation; yet these features are exceptionally difficult to recover using passive optical and radar remote sensing technologies. For example, aerodynamic roughness is generally parameterized as a function of canopy height, a parameter which conventional remote sensing methods have had little success in observing.

Lidar (light detection and ranging) or laser altimetry, in contrast, is a technology with great potential for characterizing the vertical structure of the land surface vegetation and the underlying ground topography because it directly measures such structure (Ackermann, 1999; Dubayah and Drake, 2000; Dubayah *et al.*, 2000; Lefsky *et al.*, 2002;). Lidar uses active laser energy to record the vertical distribution of canopy and ground surface components. From this distribution of energy, an impressive variety of vegetation features can be measured, inferred, or modeled, such as canopy height, foliar profiles, canopy cover, leaf area index (LAI), biomass, canopy volume, basal area, and stem density, among others. Furthermore, even in densely vegetated landscapes, lidar technology can produce "bare earth" topography, which can be constructed and is vastly more detailed than what is commonly available from conventional sources. From these observations, parameterization of variables needed for land surface models is greatly facilitated.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

In this chapter, we explore the potential of lidar remote sensing for characterizing land surface states for hydrological analysis. First we present a brief overview of lidar and laser altimetry. We discuss the distinction between small-footprint lidar commonly used in commercial applications and large-footprint approaches that are at the forefront of airborne and space-based systems. We next describe common land surface variables that are often used in hydrological analysis and how these are related to vertical structure observable from lidar. Lastly, we present how lidar may be used to recover or parameterize these surface variables; in particular, we focus on observations of forest and topographic structure.

LIDAR REMOTE SENSING

Lidar remote sensing is the optical laser analog of radar (light detection and ranging). With most systems, a pulse (or multiple pulses) of laser light is emitted from an airborne or space-based sensor towards a surface where the area irradiated by the laser pulse is referred to as a "footprint". An overview of the physical principles of lidar remote sensing as well as issues surrounding data collection, processing, and interpretation are provided in Wehr and Lohr (1999), Axelsson (1999) and Baltsavias (1999). Lidars that are designed to observe land surface features frequently operate in the near-IR as vegetation is highly reflective in this part of the electromagnetic spectrum. The incident energy is reflected off the surface and collected by a telescope back at the instrument. The roundtrip time of transit of the pulse is recorded from which the range or distance from the instrument to the surface can be calculated. For bare earth targets, for example, unvegetated ground, this distance from the sensor can then be converted to an estimate of topographic elevation relative to the Earth's geoid. For vegetated surfaces, the incident pulse reflects off canopy elements, providing a single range or a series of ranges to one or more portions of the canopy, depending on sensor design and configuration.

Lidar systems used for land surface observation may be distinguished by several characteristics, including footprint size, return signal digitization, pulse rate/scanning pattern, and wavelength.

There are two classes of footprint size in use, "small" and "large". By far, most lidar systems are small-footprint, where the incident beam has a diameter of centimeters to tens of centimeters at the surface. Almost all commercial systems collect such small-footprint data. In contrast, there are a few research systems, most notably the Scanning Lidar Imager of Canopies by Echo Recovery (SLICER, Harding *et al.*, 1994; Blair *et al.*, 1994), the Laser Vegetation Imaging Sensor (LVIS, Blair *et al.*, 1999), and the Geoscience Laser Altimeter System (GLAS, Schutz, 1998), that are large-footprint. With these instruments, footprint diameters are on the order of tens of meters, depending on configuration. For example, LVIS can operate at a variable footprint sizes, from about 5–50 m, whereas GLAS, a space-based lidar, has a large-footprint size of about 120 m.

Pulse rates and scanning patterns also vary with systems. Commercial small-footprint systems commonly operate at frequencies between about 1000 and 30 000 Hz, but have recently gone as high as 100 000 Hz. These instruments have variable scanning patterns (e.g. transect, gridded, or helical). Furthermore, these sensors are not imaging, instead they fire large numbers of pulses in close proximity to sample densely and recreate the surface being flown over. Generally, there is a trade-off between the sampling density required and the scanning width (or swath): high-density sampling will require a high-pulse-rate system and a tighter scanning pattern to reduce spaces between adjacent footprints. In well-vegetated landscapes, filtering algorithms are used to detect the bare earth signal in the data. Typically, significant overlap is needed in successive flight lines to obtain sufficient canopy penetration to characterize the bare earth topography.

The two large-footprint systems in use, LVIS and GLAS, are quite different. LVIS is an airborne scanning/imaging system, laying down adjacent-, along-, and across-track footprints within a 1–2 km wide swath. GLAS is space-based and has a pulse rate of 40 Hz with footprint spacing of 165 m along track, and variable across-track spacing as function of latitude (e.g. 15 km across track at the equator, with track separation decreasing towards the poles).

Return signal digitization is an important feature of lidar systems. Most commercial systems do not fully digitize the return signal, but rather record energy at various points. Common systems in use today are first-, last-, first-and-last-, and multiple-return systems. For example, a system that records first and last returns would provide a range to the surface for the first surface intercepted, say some portion of the canopy top, and the last surface intercepted, somewhere inside the canopy or possibly the ground below the canopy. Other systems, such as LVIS, digitize the entire record of reflected energy, resulting in what is commonly referred to as a *waveform*. The incident pulse fired from the laser is Gaussian in shape. As this Gaussian pulse is reflected off leaves and branches, the returned (non-Gaussian) waveform is fully digitized at some appropriate vertical resolution, for example, every 50 cm.

Small-footprint systems are used primarily to record topography but have also been employed extensively to map forest structure (Nelson *et al.*, 1984; Maclean and Krabill, 1986; Nelson *et al.*, 1988; Naesset, 1997; Nelson, 1997; Magnussen and Boudewyn, 1998; Gaveau and Hill, 2003). Small-footprint systems can be used in low flying airplanes or helicopters to obtain bare earth data density, which in grassy areas, is less than 1 point per square meter

continuously over large areas. Typically small-footprint systems record first and last return, although multireturn systems are becoming commercially available. In either large- or small-footprint systems, filtering algorithms must be used to distinguish true ground surface from vegetation. As mentioned later, much work is under way to develop filtering and classification algorithms (e.g. Barber and Shortridge, 2004; Sithole and Vosselman, 2004).

Figure 1 shows the measurement basis for multiple-return, small-footprint systems. A system designed to record more than one return may register multiple signals from a single tree crown; from a combination of crown material, woody material, and shrub layer, and/or from the ground. These returns from differing elevations can be analyzed statistically (e.g. through cluster analysis) to organize the data into canopy, understory, and ground.

Waveform digitization systems, regardless of footprint size, may improve ground detection because the entire signal is digitized. Furthermore, large-footprint systems, especially imaging systems such as LVIS, do not have the sampling issues present in the crown-detection problem mentioned above, that is, they completely map all portions of the canopy top so that no crowns are missed. However, in both large- and small-footprint systems, the sensitivity of the system is a critical parameter as it determines the minimum area perceivable by the sensor. In the case of large-footprint systems, this means that there must be sufficient leaf material at a particular height so that enough photons are returned above a background noise level. In both large- and small-footprint systems, the beam may penetrate a distance into the canopy before enough material is encountered to produce a strong enough reflected signal.

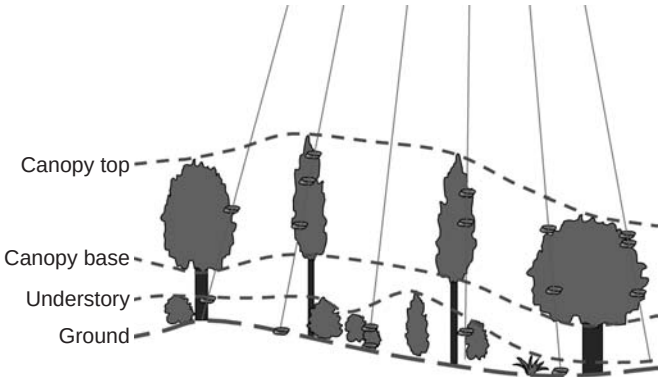


Figure 1 Schematic illustrating the basic principles behind a multiple-return, small-footprint system. The returns may register multiple signals from either a single crown; from a combination of crown, woody or shrub material; and/or from the ground. These returns can be analyzed statistically to organize the data into canopy, understory, and ground. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

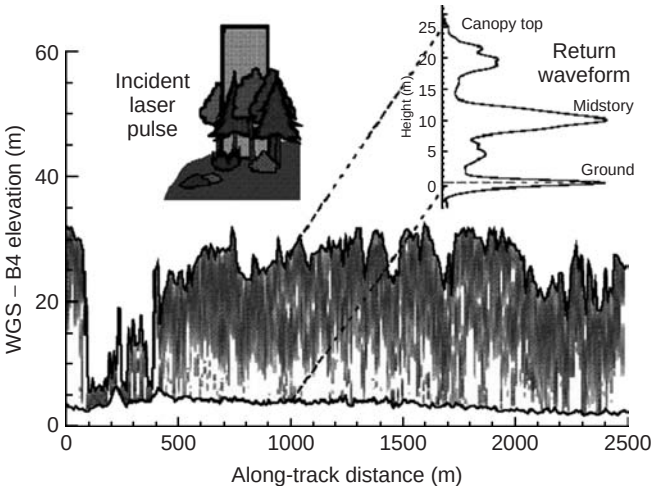


Figure 2 A conceptual drawing depicting the basic principles of large-footprint, waveform-digitizing lidar systems. The incident pulse of laser energy reflects off the canopy material and the ground beneath. The strengths of the amplitudes in the return waveform correspond to the amount of canopy material present at any given height in the canopy. The top of the canopy, the first-return above the background noise threshold, and ground location, the peak of the last return above the noise threshold, are also shown. Canopy height is the difference in elevation between the ground and the canopy top. The profile derived from actual lidar waveform data shows the organization of canopy material both horizontally and vertically; the darker shades correspond to areas with more canopy material, lighter shades to those with less, and clear areas with no canopy, such as in the subcanopy airspaces. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 2 is a conceptual drawing illustrating the basic principles of large-footprint, waveform-digitizing lidar systems, that is, the strengths of the amplitudes in the waveform correspond to the amount of canopy material present at any given height in the canopy. The profile in Figure 2 clearly shows how the waveform data can be aggregated to provide a three-dimensional picture of the horizontal and vertical structure of the forest canopy.

DESIRED LAND SURFACE PARAMETERS FOR HYDROLOGICAL MODELING AND MONITORING

Land surface characteristics are important for hydrological modeling and predicate many of the parameters that control hydrological processes. Virtually every aspect of both the transfer of mass and energy transfer between the surface and the atmosphere, as well as the energy and water balance of the surface alone, is mediated by the characteristics of the land surface and its underlying topography. For example, evapotranspiration and turbulent transfer, short

and longwave energy fluxes, photosynthetic rates, rainfall interception, stemflow, throughflow, and runoff, are all affected by the characteristics of the vegetation and terrain of the area being modeled.

Most land surface structure characteristics are not easy to measure. Field sampling provides accurate data; however, the information gathered through field work is time consuming, expensive to obtain, and covers very limited areas. Remote sensing technology is frequently used to derive forest structure to avoid these issues. Lidar, in particular, is especially suited for the measurements of forest stand characteristics and is able to measure those characteristics from which the hydrological variables are derived. Dubayah *et al.* (2000), Dubayah and Drake (2000), and Lefsky *et al.* (2002) summarize much of this work for large-footprint systems.

Land surface parameters may be derived from lidar in one of three ways. First, some parameters, such as canopy height, may be obtained through direct retrievals from lidar signals. Second, other parameters may be retrieved through modeling, either statistically or using physics-based models such as radiative transfer models. For example, biomass is almost always retrieved through the creation of statistical relationships between ground-based measures of biomass and lidar returns. Third, other parameters may need a combination of data from various sensors, so-called *remote sensing fusion*, for retrieval. Table 1 lists various land surface characteristics that have been successfully obtained from lidar. In this review, we will focus on forest attributes most important for hydrological applications: canopy height, canopy height and foliar profiles, canopy cover and LAI, biomass, height to live crown (HTLC), large tree density, and basal area. In addition, we briefly review the retrieval of topographic structure from lidar.

FOREST STRUCTURE FROM LIDAR

Canopy Height

Canopy height is perhaps the most important land surface characteristic derivable from lidar not only because of its relevance for hydrological modeling, for example, aerodynamic roughness, but also because it is the most difficult to obtain through other remote sensing methods. Canopy height is a relatively direct and straightforward measurement for lidar systems. This is because canopy height is simply the difference between the ground elevation and the top of the canopy. For large-footprint waveform data, the top of the canopy is defined as the height of the first signal above the noise threshold of the waveform. The ground is defined as the peak of the last return (see Figure 2). Gridded height data from scanning large-footprint systems yield a unique view of canopy structure and increase the ability to distinguish between land cover types (Figure 3). Height estimation with small-footprint systems is somewhat more complex and generally requires the creation of interpolated ground and canopy surfaces created from first and last return signals. These two surfaces can then be differenced to calculate the canopy height. With sufficient sampling, individual trees and their heights can be determined (refer to Figure 1). If wall-to-wall sampling by a lidar system, either small- or large-footprint, is not available for an area, then integration of other remote sensing data with more extensive coverage may help extrapolate height information for a larger area (e.g. see Hudak *et al.*, 2002).

Both large- and small-footprint systems retrieve canopy height with errors of a few meters (Nelson *et al.*, 1988; Nilsson, 1996; Magnussen and Boudewyn, 1998; Means *et al.*, 1999; Lefsky *et al.*, 1999a; Hyde *et al.*, In press).

Table 1 Potential of lidar for deriving forest structure characteristics

Land surface characteristic	Lidar derivation
Canopy height	Direct retrieval
Vertical distribution of intercepted surfaces	Direct retrieval
Canopy cover / LAI	Direct and modeled
Canopy height profile	Modeled
Foliar height profile	Modeled
Biomass	Modeled
Height to live crown	Modeled
Large tree density/Basal area	Modeled

Figure 3 Image showing gridded height in Costa Rica in and around a portion of the La Selva Biological Station. The lighter shades indicate taller canopies, blues and greens short. In the image, one can detect a band of vegetation on both sides of the river running through the top portion of the image. The height data improve the identification of land use, especially in conjunction with ancillary data. The northwestern boundary of La Selva is visible where regenerating and secondary forests abut against fields and pasture land. The nearly rectangular area of taller vegetation along the eastern edge of the figure shows a privately owned old-growth forest stand. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Large-footprint systems in particular have been tested in a variety of forest conditions and have successfully recovered canopy height in deciduous, broadleaf forests (Lefsky *et al.*, 1999a), dense conifer forests (Means *et al.*, 1999), and more open conifer forests (Hyde *et al.*, In press). It has been suggested that for small-footprint systems, a combination of sampling density and sensor sensitivity (especially in broadleaf forests) are the dominant factors leading to height recovery biases, whereas for large-footprint systems, sensitivity alone is most important. Insufficient sensitivity can cause the laser energy to penetrate some depth into the canopy before registering a signal, thereby causing an underestimation of the canopy height (Gaveau and Hill, 2003). The sensitivity of the system also determines the likelihood of ground detection. In very closed canopies where closure approaches 99% or greater, systems must have sufficient power to penetrate to the ground with enough energy to provide a clear return pulse from the ground, but not so much that high-amplitude signals from the canopy saturate. Furthermore, height underestimation associated with large-footprint sensors can also be correlated with the location of the tallest trees relative to the footprint center (Hyde *et al.*, In press). Sensitivity may not be as strong at the footprint edges, causing the sensor to register a lower height, especially in sparse forests with conical crowns (e.g. conifers). Some of the additional biases that result from crown shape may be corrected through calibration efforts (Nelson, 1997).

Canopy and Foliar Height Profiles

The vertical distribution of canopy material affects mass and energy at and near the surface. The return waveform of fully digitizing lidar systems records the near-nadir reflections from the vertical distribution of intercepted surfaces (VDICS); indeed, this is the most basic measurement of such systems as shown in Figure 2. The relative strengths of the amplitudes of the return waveform may be used to infer the relative amount of canopy material (leaves and branches) that occur at any height (Harding *et al.*, 2001). However, as lidar energy passes through the canopy, it is attenuated so that there is less energy at lower levels in the canopy. Potentially, therefore, returns from the lower portion of the canopy may be smaller relative to the upper portion, not because of actual differences in canopy material but because of less incident energy. This effect can be adjusted for using the so-called MacArthur–Horn transformation (Means *et al.*, 1999; Lefsky *et al.*, 1999b; Harding *et al.*, 2001). The use of this transformation to increase the relative amplitude of lower portions of the signal is generally limited to canopies that meet relatively strict requirements of horizontal and vertical homogeneity. Once the transformation has been applied, the adjusted waveform is often referred to as *canopy height profile* (CHP), as opposed to the distribution of just the leaf material which is called the *foliar height profile* (FHP), neither one of which is necessarily identical to the VDICS. In retrieval of CHP in canopies that are clumped or relatively open, for example, coniferous stands, use of the MacArthur–Horn transformation is not well established. Derivation of the FHP is more difficult and is dependent on additional information about the canopy, including architectural variation, leaf and branch reflections, leaf orientation, as well as the shading effects mentioned earlier. Drake *et al.* (2002b) derived CHPs for lidar waveform data from Costa Rica and compared these to crown volume profiles calculated from field data. The two sets of profiles, representing different forest types, showed fairly good agreement, as did the metrics derived from both. In another study, Parker *et al.* (2001) demonstrate how lidar-derived CHPs can be used to infer light transmittance through a canopy.

Canopy Cover and LAI

Canopy cover and LAI are frequently used in hydrological modeling in deriving various aspects of the energy and mass balance and are also to infer rainfall interception rates by the canopy. Algorithms for determining canopy cover from large-footprint lidar are based on determining the relative amount of energy in the canopy portion of the return as opposed to the energy in the ground (Figure 4). For example, if the ground return contains 25% of the total energy in the waveform, then a canopy cover of approximately 75% may be assumed. Derivation of canopy

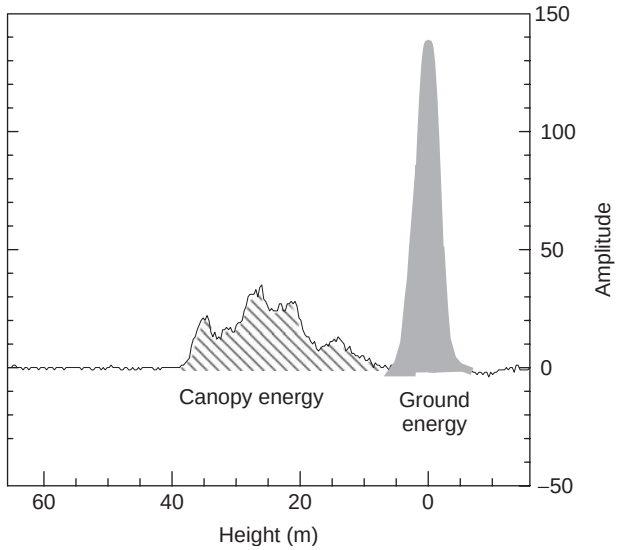


Figure 4 A diagram showing an individual lidar waveform. The waveform consists of a return from the ground (hatched) and from the canopy (grey). Canopy cover can be estimated by determining the ratio of the energy contained in the canopy return to the amount of energy in the ground return. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

cover in this way is dependent on knowing the relative reflectance of the ground and canopy in the wavelength region of the lidar. Retrieval of LAI requires further assumptions (e.g. see Lefsky *et al.*, 1999b).

Biomass

The study of the cycling of water and carbon through forested watersheds requires knowledge of both the current carbon status of the forest, as well as its successional state. Above-ground biomass (AGBM) is about 50% carbon. When this information is combined with successional status gleaned from the distribution of heights in the forest, it provides a powerful initialization of land surface state required for modeling energy, water and mass balances, including carbon.

Canopy height is the simplest lidar metric for predicting AGBM. Tree diameters generally increase as trees get taller, and there have been thousands of studies that have modeled the nonlinear allometric relationships between AGBM and tree diameter and height. Many small-footprint lidar studies have produced accurate predictions of biomass utilizing height alone. Other metrics derived from lidar waveforms may also be used (Lefsky *et al.*, 1999a,b; Drake *et al.*, 2002a; Hyde *et al.*, In press). For example, Drake *et al.* (2002a) used an energy metric, the lidar height of median energy (HOME) to explain almost 90% of variability in AGBM in a tropical forest (Figure 5) and to create a map of biomass (Figure 6). Other methods may use a combination of lidar metrics and other remote sensing and ancillary data. In any case, the prediction of biomass in high-biomass areas, where traditional remote sensing techniques have not fared well, is one of the most spectacular successes of lidar remote sensing.

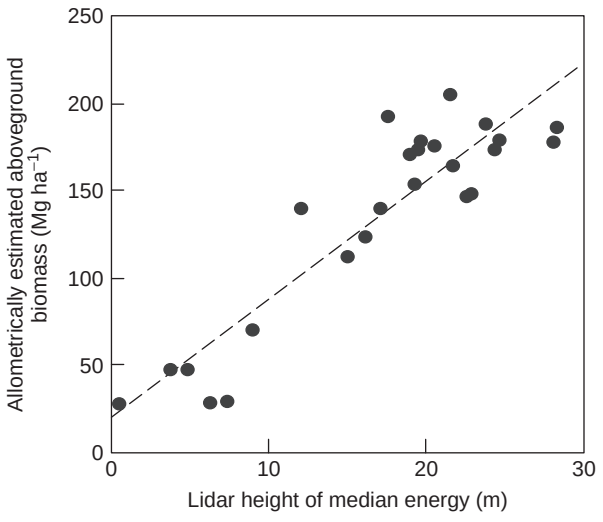


Figure 5 Results of a regression analysis using the LVIS HOME metric to predict AGBM for a series of plots in the rainforest at La Selva Biological Station in Costa Rica (Drake *et al.*, 2002a). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Height to Live Crown

HTLC is the height above the ground where the first living canopy material occurs. When combined with canopy height, it determines canopy depth and volume. It also delineates the subcanopy airspace, which is an important parameter in many land surface models. In theory, the lower edge of the canopy should be observable by lidar, provided it is a relatively sharp boundary, the canopy is relatively uniform, and there is sufficient energy in the waveform at that depth. In practice, this boundary may be difficult to detect, especially in closed forests or forests with a heterogeneous structure. For stands that are more dense and complex, it may be necessary to statistically model HTLC from lidar metrics. Furthermore, without ancillary information either from other remote sensing data or life form knowledge, lidar cannot distinguish between live and dead canopy material. A parameter closely related to HTLC, canopy base height (CBH), has been modeled from a combination of lidar metrics for a variety of forest types in the Sierra Nevada, comparing relatively well, to ground-based estimates ($r^2 = 0.59$) (Peterson *et al.*, 2003). These results are encouraging as they represent one of the few successful efforts to recover CBH from remote sensing. However, background noise, the presence of an understory, slope, as well as the architecture of the canopy itself can limit retrieval of both HTLC and CBH.

Large Tree Density and Basal Area

Tree density and basal area affect such hydrological parameters as aerodynamic roughness and overland flow, in addition to their ecological significance. Because lidar instruments provide measurements of canopy structure, it is possible to identify individual large or emergent tree crowns. Small-footprint systems that have sufficient sampling density can be used to visually identify individual crowns (Figure 7), or image segmentation methodologies may be employed to delimit tree crowns (Moorsdorf *et al.*, 2004). In contrast, large footprints may encompass several crowns, although histograms of canopy heights can be used to infer the frequency of large trees, especially emergent trees. However, other studies have shown good statistical relationships between large-footprint data and a variety of attributes related to tree density, including basal area and stem diameter (Drake *et al.*, 2002a; Lefsky *et al.*, 1999a,b; Means *et al.*, 1999). Drake *et al.* (2003), for example, found that the HOME metric alone accounted for up to 92% of variability of stem diameter and up to 70% of variability in basal area for two different tropical rainforest sites (Figure 8).

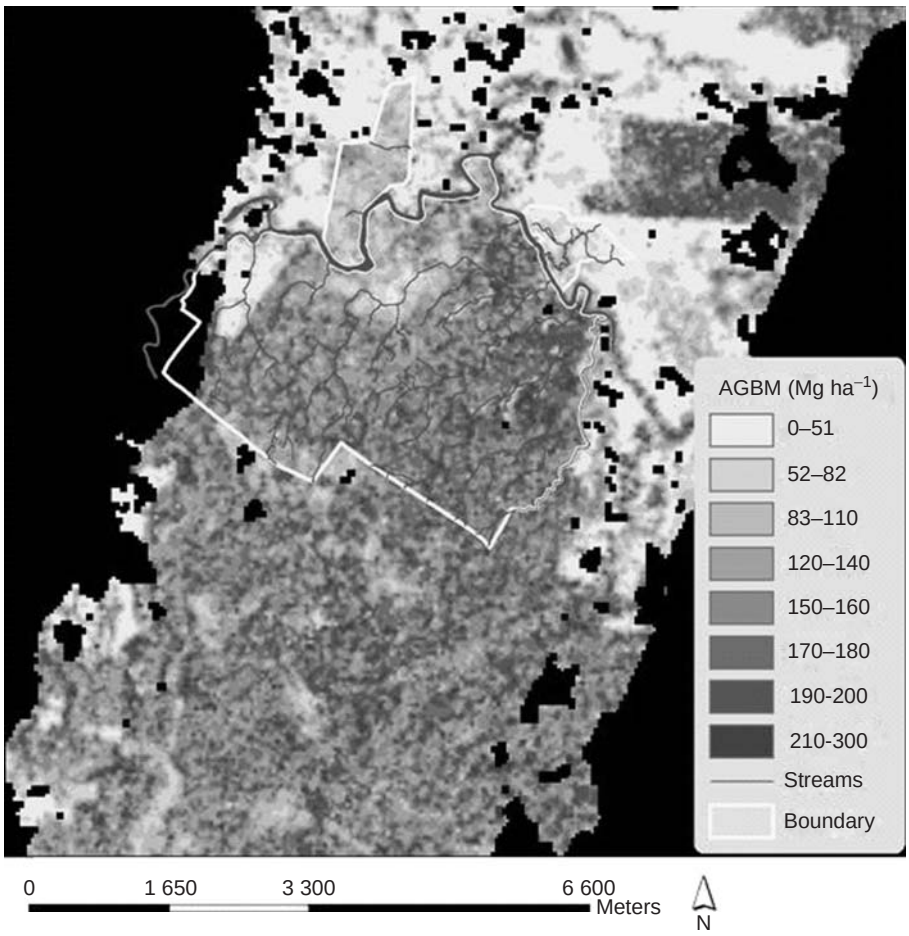


Figure 6 AGBM as derived from the HOME metric calculated from LVIS waveform data (from Drake *et al.*, 2002a). The image shows higher AGBM for the primary forest areas in the southern and eastern parts of La Selva and in Braulio Carrillo National Park to the south. The western parts of La Selva contain secondary and regenerating forests, with markedly lower AGBM values. Much of the land surrounding La Selva is agricultural and therefore also characterized by low AGBM values. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Radiative transfer methods may also be used to recover tree densities within large footprints (e.g. see Ni-Meister *et al.*, 2001). Waveform data can be input into a radiative transfer inversion model to calculate canopy structure characteristics such as stem density and LAI. This process is more complex than relying on statistical models to predict the structural characteristics of interest. However, the use of radiative transfer theory to infer canopy structure from lidar may result in the development of more generally or globally applicable models, thereby reducing the need for the ground-based data required to develop the more common statistical models.

GROUND TOPOGRAPHY FROM LIDAR

Small footprint, fixed-wing or helicopter-based laser swath mapping can provide elevation data of the ground surface over large areas with a spatial resolution of 1 m or better, a

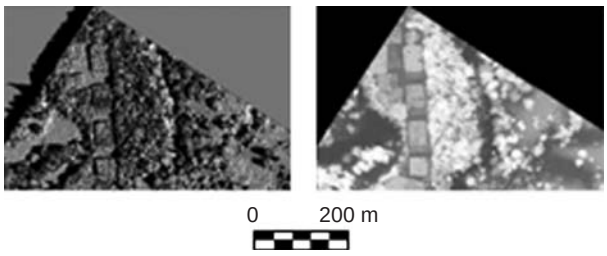


Figure 7 A small-footprint, first-return lidar system (FLIMAP, developed by John E. Chance and Associates, Louisiana) mapped a portion of the La Selva Biological Station in Costa Rica. The samples were gridded to a 1-m resolution. In this image, lighter tones signify higher elevations. Individual tree crowns are clearly recognizable. The rectangular features represent experimental agroforestry plots. A synthetic reflectance image (b) is created by illuminating the DEM on (a), and shows the remarkable canopy structure detail available from small-footprint systems

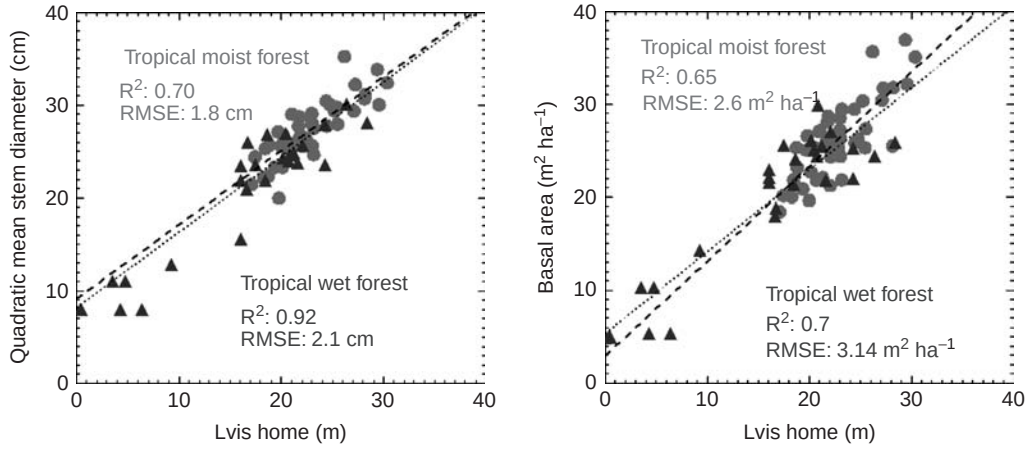


Figure 8 Results of regression analyses comparing lidar-derived HOME to quadratic mean stem diameter and basal area for two different topical rainforest study sites in Panama (circles, dashed line) and Costa Rica (triangles, dotted line) (Drake *et al.*, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

vertical accuracy of 5–10 cm, and a horizontal accuracy of 15–20 cm (Carter *et al.*, 2001). Such unprecedented topographic data promises to revolutionize the studies of hydrology and geomorphology (as well as all other fields that use ground topography). To obtain accurate high-resolution maps from such surveys, however, it is not simply a matter of installing a laser in a plane and flying it around, as considerable care and skill is involved in accurately tracking the plane and its motion (to provide the GPS-based elevation of the laser relative to the ground), calibrating the laser (which must be done for all surveys), and filtering the recovered data to remove vegetation, bridges, buildings, and other features that obscure the topography. Also, if systematically high-resolution data are to be acquired, especially in hilly or mountainous vegetated landscapes, flight lines must have considerable overlap (preferably at least 50%).

The greatest challenge in assembling a high-resolution topographic map from an airborne laser swath survey is the detection of bare earth data in the "point cloud" of all the return data acquired (Figure 9). Much research is under way to develop software that can remove vegetation, buildings, and other features without eliminating critical topographic attributes such as channel banks (e.g. Filin, 2004; Luzum *et al.*, 2004; Brovelli *et al.*, 2004). In areas with variable vegetation cover (e.g. combinations of hardwood, softwood, brush, agricultural fields, etc.), bare earth filtering procedures may be improved by conditioning them by the vegetation type. Included in the return data is the intensity of the signal, and this intensity varies with surface conditions, which may also be used to improve filtering strategies. Once the bare earth data have been generated, the resultant topography will depend on the procedure used for gridding the data. Hence, while high-resolution data fields can now be generated, there is still considerable art to optimizing the analysis of these data to get the most artifact-free realization of the surface topography. Given the challenges of obtaining high-resolution ground data under vegetation, low return-density and errors associated with rugged landscapes,

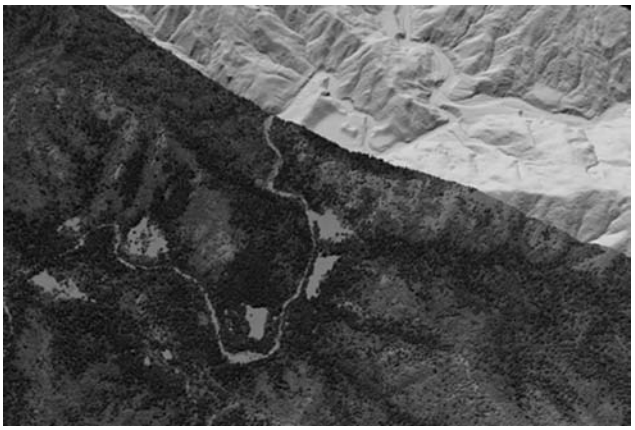


Figure 9 Digital image of the Angelo Reserve on the South Fork Eel River, California derived from data collected by the National Center for Airborne Laser Mapping in 2004 as part of a National Center for Earth-surface Dynamics study. Canopy surface and filtered ''bare earth'' data were computed and the difference in elevation between points was shaded: 1 m or less is light grey, from 1 m to 30 m, the shades are increasingly dark, representing greater differences. This accurately shows the spatial variation in vegetation from brush on the exposed hillslopes to Douglas-fir and Redwoods in the valley bottoms. The narrow brown sinuous line is the South Fork Eel River and adjacent to it our brown colored meadows. In the upper corner, the vegetation has been removed, revealing the bare earth topography in this shaded relief map. The diagonal distance from lower left to upper right corner is about 4 km. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and the still-developing art of data analysis, it is essential that the potential user of such data be actively involved in data evaluation before using it in some project. Major artifacts are easily detectable by simple inspection of topographic maps, shaded relief maps, and plots of bare earth data distribution (ungridded) across the study area.

Of particular relevance to hydrology, high-resolution topographic maps from airborne laser swath maps provide for the first time the actual topography of the channel network over large areas. In nearly all digital elevation models used in hydrologic models, channels are defined as occurring in cells that drain some threshold area, and the size (width and bankfull depth) of the channel is specified by an assumed relationship between channel scale and drainage area (derived from hydraulic geometry observations, i.e. Leopold and Maddock, 1953). Even in the 10-m gridded United States Geological Survey maps, the valleys in which the finest scale channels occur are often not accurately shown (or are simply absent), and the relatively large channels are usually represented by two lines delineating the location of the channel banks. Drainage density affects runoff response (e.g. Beven and Kirby, 1993), and channel topography affects flow routing. Now, with high-resolution airborne laser swath mapping data, the channels on digital terrain maps are actual topographic features, with distinct heads, banks, and topography (Figure 10).

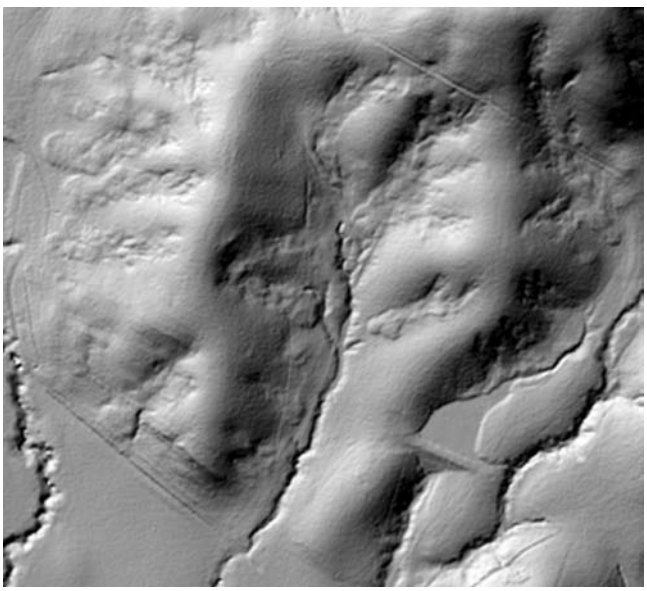


Figure 10 Shaded relief map of a portion of the Napa River watershed revealing landslides, roads, channels, channel heads, and reservoirs. High-resolution airborne laser swath mapping data are for the entire 1100 km^2 are available for download at **http://calm.geo.berkeley.edu/ncalm/index.html** as part of the National Center for Airborne Laser Mapping. Reservoir embankment is about 70-m long

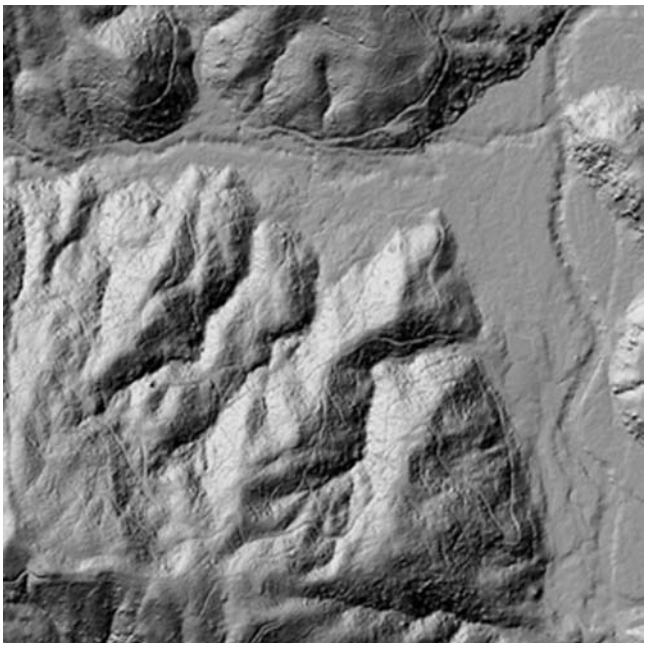


Figure 11 A shaded relief map showing a portion of the South Fork Eel River area outside the Angelo Reserve where a dense network of logging roads and skid trails have been cut across the landscape. Distance from top to bottom of the page is about 1 km. Such road densities may have a long lasting influence on runoff rates

Such high-resolution data detect the topographic presence of dams or even small agricultural reservoirs (Figure 10) and are ideal for accurately quantifying the extent and topography of roads and logging skid trials (Figure 11), which cumulatively in a watershed, may have long-term effects on peak flows (e.g. Jones and Grant, 1996) and sediment loading. A new challenge, however, is that this well-defined road network typically frequently crosses smaller channels via culverts. Lidar does not detect culverts (though the culvert location may be inferred from the adjacent of channel and elevated road bed in some cases) and consequently the roads breakup the channel network. Hence, to create an accurate channel network, either topographic connection across roads must be done by hand in the data, or software must be developed that does this automatically (Barber and Shortridge, 2004).

One of the most widespread applications of lidar in hydrology has been for floodplain mapping and there have been extensive areas overflown for this purpose (especially in Louisiana) (e.g. Pereira and Wicherson, 1999; Barber and Shortridge, 2004). In modeling floods on the River Severn, Cobby *et al.* (2003) found that lidar improved predictions with its more accurate topography because it enabled them to decompose a finite-element mesh to account for vegetation-dependent frictional properties.

Airborne laser swath mapping has enabled the testing and parameterization of theories for soil transport, hillslope evolution, and channel incision (Roering *et al.*, 1999; Stock and Dietrich, 2003). It may eventually lead to greatly improved landscape evolution models. Given the strong dependency of landslides on topography, high-resolution topographic data can greatly improve landslide potential maps (e.g. Dietrich *et al.*, 2001). Such topographic data are proving invaluable in detecting and tracing faults (e.g. Hudnut *et al.*, 2002; Sherrod *et al.*, 2004). These early accomplishments point to a large influence of airborne laser swath mapping on earth sciences.

High-resolution topography presents new challenges and opportunities to watershed runoff models. Should variable grid size models, or models based on irregular triangular networks, be further developed so that the fine-scale topography of critical features, such as channels, can be preserved while coarser scale cells can be used to characterize slowly varying topography, such as floodplains? How can the fine-scale topography be best exploited to improve hydrology modeling? Bates *et al.* (2003), in discussing numerical models of river flooding, have noted: "A newly emergent research area, therefore, is how to integrate such massive data sets with lower-resolution numerical inundation models in an optimum manner that makes maximum use of the information content available". French (2003) has argued that the high-resolution topographic data allows the modeler to focus on the physical mechanisms and parameterization by reducing geometric uncertainty. In the past, numerical models suffered from the crude topographic data available to them; now their use is made problematic because the resolution exceeds their abilities.

This article has focused on small-footprint lidar and the fine-scale topography it can provide. Large-footprint, waveform digitization lidar avoids some of the problems of filtering mentioned above. Such lidars measure topography under the canopy directly, provided there is sufficient energy in the ground return. Hofton *et al.* (2002) showed that LVIS successfully measured subcanopy topography at the La Selva Biological Station in Costa Rica where the canopy cover was often 99% or above. LVIS penetrated the dense canopy to measure the elevation of the surface beneath providing a unique picture of the highly variable terrain present beneath the canopy.

The coming years may produce a confluence of these two technologies, as large-footprint systems decrease their footprint size and small-footprint systems add waveform digitization. Ultimately, the choice of waveform versus multiple-return, and large versus small-footprint sizes will be determined by the particular topographic mapping application.

SUMMARY

Over the last three decades, remote sensing observations have been used to obtain the detailed information about land surface state required for hydrological analyses. During this time, land surface hydrological models have increased in complexity, in spatial and temporal resolution, and in areal extent, placing even greater demands on the accuracy and reliability of the land surface characterizations needed to both parameterize and drive them. Lidar remote sensing provides a new and revolutionary means of obtaining many of the surface descriptors, especially those related to the vertical structure of the canopy and high-resolution topography, which have proven difficult or expensive to derive by other means. The increased use of lidar in hydrological analysis will depend on many factors. First, because lidar is a new technology, its theoretical underpinnings must be further developed by more studies which link observations to physically based models of the radiative transfer of lidar energy through canopies (such as shown by Ni-Meister *et al.*, 2001; Sun and Ranson, 2000; Kotchenova *et al.*, 2004). Second, its efficacy must be further demonstrated across a gradient of environmental conditions and for a broader spectrum of applications. Third, an accompanying evolution of hydrological model structure as well as surface parameterization must be achieved. For example, virtually no model can ingest the vertical structure from lidar directly (because it has not been available). Hurtt *et al.* (2004) have shown how a model that is specifically height-structured could assimilate lidar data directly to initialize land surface states, and greatly extended model predictability. Similarly, model parameterizations need to be rethought. Aerodynamic roughness and infiltration potentially are much better parameterized using the actual profile of canopy material, rather than canopy height and LAI, respectively. Similar problems exist matching the current state of runoff models with the new high-resolution data in which to embed them. Last, the fusion of lidar observations with other, more readily obtained remote sensing data must be explored. Extensive spatial coverage on a regional basis, let alone on a continental or global scale, is probably at least a decade away. In the meantime, the development of algorithms that marry the vertical structure from lidar with a variety of spatially continuous data, such as multi- and hyperspectral, multiangle, and radar observations should be a priority.

Acknowledgments

The authors would like to thank J. Bryan Blair, Michelle Hofton, and Jason Drake for useful inputs included in this chapter, especially pertaining to the deployment and operation of the LVIS instrument and the collection and analysis of lidar data. Dubayah is supported through funding from the NASA Terrestrial Ecology Program. Dietrich

is supported by the National Center for Airborne Laser Mapping (NCALM) and the National Center for Earth-surface Dynamics (NCED).

FURTHER READING

Norheim R.A., Queija V.R. and Haugerud R.A. (2002) Comparison of LIDAR and INSAR DEMs with dense ground control, *Proceedings of the ESRI User Conference*, *San Diego*.

Reutebuch S.E., McGaughey R.J., Anderson H.-E. and Carson W.W. (2003) Accuracy of a high-resolution lidar terrain model under a conifer forest canopy. *Canadian Journal of Remote Sensing*, **29**, 527–535.

Töyrä J., Pietroniro A., Hopkinson C. and Kalbfleisch W. (2003) Assessment of airborne scanning laser altimetry (lidar) in a deltaic wetland environment. *Canadian Journal of Remote Sensing*, **29**, 718–728.

REFERENCES

Ackermann F. (1999) Airborne laser scanning – present status and future expectations. *Journal of Photogrammetry and Remote Sensing*, **54**, 64–67.

Axelsson P. (1999) Processing of laser scanner data – algorithms and applications. *Journal of Photogrammetry and Remote Sensing*, **54**, 138–147.

Baltsavias E.P. (1999) Airborne laser scanning: basic relations and formulas. *Journal of Photogrammetry and Remote Sensing*, **54**, 199–214.

Barber C.P. and Shortridge A.M. (2004) *Light Detection and Ranging (LiDAR) – Derived Elevation Data for Surface Hydrology Applications*, Technical Report WR-1(2004), Institute of Water Research, Michigan State University, East Lansing. p. 11.

Bates P.D., Marks K.J. and Horritt M.S. (2003) Optimal use of high-resolution topographic data in flood inundation models. *Hydrological Processes*, **17**, 537–557.

Beven K. and Kirby M.J. (Eds.) (1993) *Channel Network Hydrology*, John Wiley and Sons: Chichester, p. 319.

Blair J.B., Coyle D.B., Bufton J. and Harding D. (1994) Optimization of an airborne laser altimeter for remote sensing of vegetation and tree canopies. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, Pasadena, CA, pp. 939–941.

Blair J.B., Rabine D.L. and Hofton M.A. (1999) The laser vegetation imaging sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. *Journal of Photogrammetry and Remote Sensing*, **54**, 115–122.

Brovelli M.A., Cannata M. and Longoni U.M. (2004) LiDAR data filtering and DTM interpolation within GRASS. *Transactions in GIS*, **8**, 155–174.

Carter W., Shrestha R., Tuell G., Bloomquist D. and Sartori M. (2001) Airborne laser swath mapping: shines new light on earth topography. *EOS*, **82**, 46,549.

Cobby D.M., Mason D.C., Horritt M.S. and Bates P.D. (2003) Two-dimensional hydraulic flood modelling using a finite-element mesh decomposed according to vegetation and topographic features derived from airborne scanning laser altimetry. *Hydrological Processes*, **17**, 1979–2000.

Dietrich W.E., Bellugi D. and Real de Asua R. (2001) Validation of the shallow landslide model, SHALSTAB, for forest management. In *Land Use and Watersheds: Human Influence on Hydrology and Geomorphology in Urban and Forest Areas*, Wigmosta M.S. and Burges S.J. (Eds.), American Geophysical Union, Water Science and Application: Washington, vol. 2, pp. 195–227.

Drake J.B., Dubayah R.O., Clark D.B., Knox R.G., Blair J.B., Hofton M.A., Chazdon R.L., Weishampel J.F. and Prince S.D. (2002a) Estimation of tropical structural characteristics using large-footprint lidar. *Remote Sensing of Environment*, **79**, 305–319.

Drake J.B., Dubayah R.O., Knox R.G., Clark D.B. and Blair J.B. (2002b) Sensitivity of large-footprint lidar to canopy structure and biomass in a neotropical rainforest. *Remote Sensing of Environment*, **81**, 378–392.

Drake J.B., Knox R.G., Dubayah R.O., Clark D.B., Condit R., Blair J.B. and Hofton M. (2003) Above-ground biomass estimation on closed canopy Neotropical forests using lidar remote sensing: factors affecting the generality of relationships. *Global Ecology and Biogeography*, **12**, 147–159.

Dubayah R.O. and Drake J.B. (2000) Lidar remote sensing for forestry. *Journal of Forestry*, **98**, 44–46.

Dubayah R., Knox R., Hofton M., Blair J.B. and Drake J. (2000) Land surface characterization using lidar remote sensing. In *Spatial Information for Land Use Management*, Hill M.J. and Aspinall R.J. (Eds.), International Publishers Direct: Singapore, pp. 25–38.

Filin S. (2004) Surface classification from airborne laser scanning data. *Computers and Geosciences*, **30**, 1033–1041.

French J.R. (2003) Airborne LiDAR in support of geomorphological and hydraulic modeling. *Earth Surface Processes and Landforms*, **28**, 321–335.

Gaveau D.L.A. and Hill R.A. (2003) Quantifying canopy height underestimation by laser pulse penetration in small-footprint airborne laser scanning data. *Canadian Journal of Remote Sensing*, **29**, 650–657.

Harding D.J., Blair J.B., Garvin J.B. and Lawrence W.T. (1994) Laser altimetry waveform measurement of vegetation canopy structure, *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, Pasadena, CA, pp. 1251–1253.

Harding D.J., Lefsky M.A., Parker G.G. and Blair J.B. (2001) Laser altimeter canopy height profiles – methods and validation for closed-canopy, broadleaf forests. *Remote Sensing of Environment*, **76**, 283–297.

Hofton M.A., Rocchio L.E., Blair J.B. and Dubayah R. (2002) Validation of vegetation canopy lidar sub-canopy topography measurements for a dense tropical forest. *Journal of Geodynamics*, **34**, 491–502.

Hudak A.T., Lefsky M.A., Cohen W.B. and Berterretche M. (2002) Integration of lidar and Landsat ETM+ data for estimating and mapping forest canopy height. *Remote Sensing of Environment*, **82**, 397–417.

Hudnut K.W., Borsa A., Glennie C. and Minster J.B. (2002) High-resolution topography along surface rupture of the 16 October 1999 Hector Mine, California, earthquake (M-w 7.1) from airborne laser swath mapping. *Bulletin of the Seismological Society of America*, **92**, 1570–1576.

Hurtt G.C., Dubayah R., Drake J., Moorcroft P.R., Pacala S.W., Blair J.B. and Fearon M.G. (2004) Beyond potential vegetation: combining lidar data and a height-structured model for carbon studies. *Ecological Applications*, **14**, 873.

Hyde P., Peterson B., Blair J.B., Hofton M., Hunsaker C., Knox R., Walker W. and Dubayah R. (In press). Mapping habitat suitability using waveform lidar: validation of montane forest structures. *Remote Sensing of Environment*.

Jones J.A. and Grant G.E. (1996) Long-term stormflow responses to clearcutting and roads in small and large basins, western Cascades, Oregon. *Water Resources Research*, **32**, 959–974.

Kotchenova S.Y., Song X., Shabanov N.V., Potter C.S., Knyazikhin Y. and Myneni R.B. (2004) Lidar remote sensing for modeling gross primary production of deciduous forests. *Remote Sensing of Environment*, **92**, 158–172.

Lefsky M.A., Cohen W.B., Acker S.A., Parker G.G., Spies T.A. and Harding D. (1999b) Lidar remote sensing of the canopy structure and the biophysical properties of Douglas-fir Western Hemlock forests. *Remote Sensing of Environment*, **70**, 339–361.

Lefsky M.A., Cohen W.B., Parker G.G. and Harding D.J. (2002) Lidar remote sensing for ecosystem studies. *BioScience*, **52**, 19–30.

Lefsky M.A., Harding D., Cohen W.B., Parker G. and Shugart H.H. (1999a) Surface lidar remote sensing of basal area and biomass in deciduous forests of Eastern Maryland, USA. *Remote Sensing of Environment*, **67**, 83–98.

Leopold L.B., Maddock T. Jr (1953) *The Hydraulic Geometry of Stream Channels and Some Physiographic Implications*, U. S. Geological Survey: Prof. Paper p. 252.

Luzum B.J., Slatton K.C. and Shrestha R.L. (2004) Identification and analysis of airborne laser swath mapping data in a novel feature space. *IEEE Geoscience and Remote Sensing Letters*, **1**, 268–271.

Maclean G. and Krabill W.B. (1986) Gross-merchantable timber volume estimation using an airborne lidar system. *Canadian Journal of Remote Sensing*, **12**, 7–18.

Magnussen S. and Boudewyn P. (1998) Derivations of stand heights from airborne laser scanner data with canopy-based quantile estimators. *Canadian Journal of Forest Research*, **28**, 1016–1031.

Means J.E., Acker S.A., Harding D.J., Blair J.B., Lefsky M.A., Cohen W.B., Harmon M.E. and McKee W.A. (1999) Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the Western Cascades of Oregon. *Remote Sensing of Environment*, **67**, 298–308.

Moorsdorf F., Meier E., Kötz B., Itten K.I. and Allgöwer B. (2004) The potential of fuel structure mapping with high resolution small footprint lidar. *Remote Sensing of Environment*, **92**, 353–362.

Naesset E. (1997) Determination of mean tree height of forest stands using airborne laser scanner data. *Remote Sensing of Environment*, **61**, 246–253.

Nelson R. (1997) Modeling forest canopy heights: the effects of canopy shape. *Remote Sensing of Environment*, **60**, 327–334.

Nelson R., Krabill W. and Maclean G. (1984) Determining forest canopy characteristics using airborne laser data. *Remote Sensing of Environment*, **15**, 201–212.

Nelson R., Swift R. and Krabill W. (1988) Using airborne lasers to estimate forest canopy and stand characteristics. *Journal of Forestry*, **86**, 31–38.

Nilsson M. (1996) Estimation of tree heights and stand volume using an airborne lidar system. *Remote Sensing of Environment*, **56**, 1–7.

Ni-Meister W., Jupp D.L.B. and Dubayah R. (2001) Modeling lidar waveforms in heterogeneous and discrete canopies. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1943–1958.

Parker G.G., Lefsky M.A. and Harding D.J. (2001) Light transmittance in forest canopies determined using airborne laser altimetry and in-canopy quantum measurements. *Remote Sensing of Environment*, **76**, 298–309.

Pereira L.M.G. and Wicherson R.J. (1999) Suitability of laser data for deriving geographical information – a case study in the context of management of fluvial zones. *ISPRS Journal of Photogrammetry and Remote Sensing*, **54**, 105–114.

Peterson B., Hyde P., Hofton M., Dubayah R., Fites-Kaufman J., Hunsaker C. and Blair J.B. (2003) Deriving canopy structure for fire modeling from lidar, *Proceedings of the 4th International Workshop on Remote Sensing and GIS Applications to Forest Fire Management*, June 5–7, Ghent.

Roering J.J., Kirchner J.W. and Dietrich W.E. (1999) Evidence for non-linear, diffusive sediment transport on hillslopes and implications for landscape morphology. *Water Resources Research*, **35**, 853–870.

Schutz B.E. (1998) Spaceborne laser altimetry: 2001 and beyond. In *Book of Extended Abstracts, WEGENER-98*, Plag H.P. (Ed.), Norwegian Mapping Authority: Honefoss.

Sherrod B.L., Brocher T.M., Weaver C.S., Bucknam R.C., Blakely R.J., Kelsey H.M., Nelson A.R. and Haugerud R. (2004) Holocene fault scarps near Tacoma, Washington, USA. *Geology*, **32**, 9–12.

Sithole G. and Vosselman G. (2004) Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, **59**, 85–101.

Stock J.D. and Dietrich W.E. (2003) Valley incision by debris flows: evidence of a topographic signature. *Water Resources Research*, **39**, 1089, doi:10.1029/2001WR001057, 25p.

Sun G. and Ranson K.J. (2000) Modeling lidar returns from forest canopies. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 2617–2626.

Wehr A. and Lohr U. (1999) Airborne laser scanning – an introduction and overview. *Journal of Photogrammetry and Remote Sensing*, **54**, 68–82.

59: Estimation of Soil Properties Using Hyperspectral VIS/IR Sensors

ALFREDO R HUETE

Department of Soil, Water and Environmental Science, University of Arizona, Tucson, AZ, US

Knowledge of soil properties and processes are crucial to the understanding of the terrestrial hydrologic cycle and the functioning of terrestrial ecosystems. In this paper, we present the current state and potential of hyperspectral remote sensing techniques for quantitative retrieval of soil properties. Remote sensing is used to detect chemical and physical soil properties either (i) directly from the bare soil pixels, (ii) through advanced spectroscopy methods in mixed "soil-vegetation-litter" pixels, and (iii) by measurements of the overlying vegetated canopy to infer soil properties and moisture status. Optical-geometric properties of soil surfaces reveal information on soil physical features, such as soil structure, crusting, and erosion. We also investigate the use of vegetation water indices to infer soil drying and wetting in the soil root zone. We conclude with a discussion on future needs and directions for remote sensing of soil properties.

INTRODUCTION

Soils are spatially variable, dynamic systems that develop from geologic parent materials under various climate regimes and biotic factors. Knowledge of soil surface properties and processes are crucial to our understanding of the terrestrial hydrologic cycle and the functioning of most terrestrial ecosystems. Relevant soil biotic and abiotic properties include organic matter and nutrient contents, mineralogy, texture and structure, albedo, and water holding capacity. These spatially variable properties form the *pedosphere* in which a complex suite of chemical, physical, and biological processes interacts with the biosphere, lithosphere, and atmosphere. The pedosphere serves as a *geomembrane*, influencing the movement and cycling of water, gases, nutrients, toxic compounds, and salts, with important consequences to water quality, energy balance, ecosystem health, wind and water erosion, and environmental health.

Upwelling reflected electromagnetic radiation in the visible (VIS, 400–700 nm), near-infrared (NIR, 700–1300 nm), and shortwave-infrared (SWIR, 1300–2500 nm) spectral regions provide quantitative information on terrestrial surface conditions of relevance to hydrology. The major factors influencing the interaction of electromagnetic energy with soil surfaces are (i) biogeochemical constituents such as primary and secondary minerals, organic compounds, and salts (Stoner and Baumgardner, 1981; Vitorello and Galvao, 1996); (ii) soil moisture (Bowers and Hanks, 1965; Baumgardner *et al.*, 1985); and (iii) optical-geometric effects involving texture, structure, crusts, and surface area (Irons *et al.*, 1992). The reflectance of a soil is dependent on the presence and amount of all absorbing constituents, including their interactions, for specific texture and structure conditions. Remote sensing of soil surfaces from aircraft and space-based platforms also involve more complex environmental factors such as atmosphere attenuation and scattering, topography and illumination variations, and the presence of litter and vegetation cover (Moran *et al.*, 1997). Early earth-observing satellite sensors provided only limited spectral information for remote sensing studies of soils. More recently, hyperspectral sensors and advanced spectroscopy and noise removal techniques offer much potential for quantitative retrieval of soil properties, for classifying soils, and for studying soil genesis processes and land degradation.

Remote sensing offers an easy method to collect spatial information on soil properties and vegetation canopy optical features, which provide indirect and integrated expressions of below-ground processes that can be used

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

to infer soil properties, as in the cases of geobotany and vegetation stress (Wessman, 1991; Ustin, 1999). In this paper, we discuss the uses of hyperspectral and optical-based remote sensing techniques for quantitative retrieval of soil properties. We then consider the spectral interactions and dynamics of water as it cycles through the soil surface, including the use of vegetation water indices to infer soil drying and wetting in the root zone. We also examine the optical-geometric properties of soil surfaces and their relationship to soil physical features, such as soil structure, crusting, and erosion, and conclude with a discussion on future needs and directions for remote sensing of soil properties.

SOIL BIOGEOCHEMICAL CONSTITUENTS

The most important constituents of the soil matrix are clay minerals, primary and secondary minerals, iron and aluminum sesquioxides, carbonates, gypsum, salts, humic substances, cellulose, and lignin. These chemical parameters affect the nature of the soil reflectance spectrum and are called "chromophores" by Ben-Dor *et al.* (1999). As examples, the type and relative amount of iron oxides in soils cause the red, brown, and yellow colors that are common in soils and used in soil classification (Baumgardner *et al.*, 1985). There are important iron absorption bands at 870 nm (ferric iron) and 1000 nm (ferrous iron) and there is decreasing reflectance in the 500- to 640-nm region with increasing iron contents. The iron mineral hematite has an absorption band extending to the 550-nm region and yields a reddish soil color while the iron mineral goethite gives the soil a yellowish appearance. The presence and type of iron oxides in soils are strongly related to soil weathering processes, climate regime, and the vegetation in which the soils develop. Latz *et al.* (1984) and Palacios-Orueta and Ustin (1998) found significant spectral variations due to iron oxide and organic matter content, which were useful in mapping erosion class and weathering condition.

Other mineral spectra of relevance in soils include kaolinite with a strong absorption feature at ~2200 nm attributed to hydroxyl ion that also yields absorption at 1450 nm. Gibbsite has an absorption feature at 2265 nm (Madeira Netto, 1996) and carbonates, such as calcite, have small absorption features at 2350 and 2360 nm (Ben-Dor and Banin, 1990). Most of these chemical chromophores in soil reflectance spectra are related to vibration of chemical bands and the charge transfer of ions (Ben-Dor *et al.*, 1999).

Soils, however, are generally complex mixtures of numerous mineral and organic constituents with overlapping absorption features making it very difficult to predict mineral and organic abundances based solely on spectral absorption features. Stoner and Baumgardner (1981) measured hundreds of soil samples in the laboratory and visually noted distinct spectral signature shapes, related to the iron oxide, organic matter, and textural attributes of the soils (see Figure 1). Soils devoid of appreciable amounts of iron and organic matter exhibited featureless and "convex" spectral signatures of increasing reflectance with wavelength. Soils with higher organic matter contents displayed "concave" spectral signatures, while soils with low and moderate quantities of iron oxides formed "sigmoidal" spectral signatures with strong absorption features in the blue–green part of the spectrum (Figure 1). The analysis of characteristic spectral signatures, however, yielded only a limited amount of knowledge on soil properties relative to the information content present (Kimes *et al.*, 1993). Huete and Escadafal (1991) found many variants even within one curve type, associated with selective absorptions in the visible spectrum caused by iron oxides.

Soil organic matter (SOM) influences the spectra of soils throughout the VIS to SWIR with numerous absorption features associated with vibration modes of organic functional groups (Henderson *et al.*, 1992; Ben-Dor *et al.*, 1999). The organic matter content in soils is composed of live and dead organisms, rootlets, and litter at varying stages of decomposition, and plays an important role in the water holding capacity, nutrients, and structural properties of a soil (Jenny, 1980). SOM contents range from trace amounts to 80% or more in peats and mucks. The overall behavior of SOM is to decrease reflectance throughout the optical portion of the spectrum, particularly in the VIS-NIR. Increasing amounts of organic matter coat soil particle surfaces and mask out mineralogic spectral features. When SOM exceeds 2%, they effectively mask mineralogic absorption features and may saturate soil particle surfaces, such that it becomes difficult to assess further variations in SOM contents. Dematte *et al.* (2003) showed that

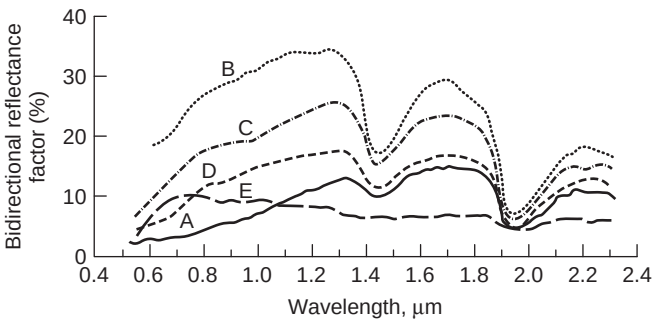


Figure 1 Five basic spectral reflectance curve shapes based on iron and organic matter contents. A – organic dominated soil, B – minimally weathered soil, C – iron-affected soil, D – organic affected soil, and E – iron dominated soil (Reproduced from Stoner and Baumgardner, 1981 by permission of Soil Science Society of America)

the removal of organic matter in soils caused reflectance increases in the VIS-NIR (450 to 1000 nm), with negligible changes from 1250 to 2400 nm. At the same time, other spectral features were observed with the greater exposure of goethite. Similar amounts of organic matter on different textures and particle surface areas will yield different spectral reflectance signatures.

Obukhov and Orlov (1964) and Baumgardner *et al.* (1985) distinguished among "fibric", "hemic", and "sapric" soil reflectance curves describing the stage of leaf litter decomposition and incorporation of plant material into the soil mineral horizon, with sapric being the most decomposed and fibric the least decomposed with recognizable plant material. There are a few spectral features present in undecomposed litter and other nonphotosynthetic vegetation (NPV) parts, associated with cellulose and lignins, including spectral absorptions at 1200 nm, 2175 nm, and 2300 nm with reflectance peaks at 2220 and 1648 nm (Gao, 1996; Asner and Lobell, 2000). In Figure 2, a set of soil and surface litter spectral signatures are shown for a field site in Brasilia National Park, Brazil. The subsoil spectra show a broad and intense iron absorption feature at 870 nm. This iron feature is present in the surface soil, but weaker, due to surface organic matter (5%) partially masking the iron feature. Also, one can see large differences in litter spectra in the SWIR region.

The presence of salts at the surface can be detected with remote sensing data either directly on bare soils, with salt efflorescence and crust, or indirectly through measures of the overlying and impacted vegetation (Mougenout *et al.*, 1993). Both the quantity and mineralogy of salts influence the spectral reflectance of the soil surface, but in manners dependent on soil moisture, color, and roughness (Metternicht and Zinck, 2003). In general, moderately to highly saline areas are easily detected, while low salinity levels and the initial stages of salinization are more difficult to discern. About 7% of the Earth's terrestrial surface and

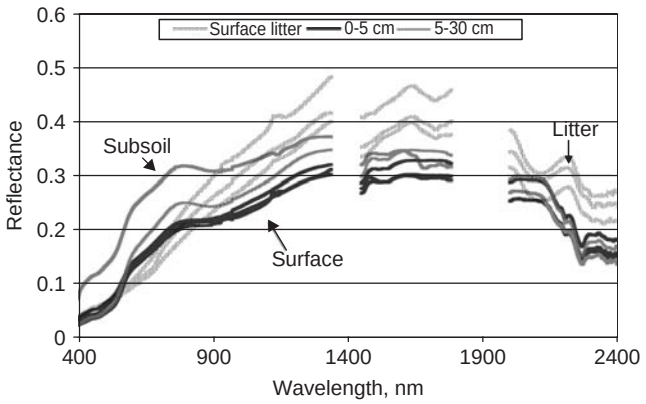


Figure 2 Sample spectral reflectance signatures of surface litter, surface soil (0–5 cm), and subsoil (5–30 cm) for a shrub cerrado site in Brasilia National Park, Brazil

20–30% of the world's irrigated lands are affected by salinity problems (Toth *et al.*, 1991) and soil salinity is a principal cause of soil degradation with negative impacts on water quality, soil structure, soil erosion, and vegetation production (Csillag *et al.*, 1993). Salinization occurs as salts (chlorides, sulfates, carbonates) of sodium, magnesium, or calcium accumulate in the root zone and move upward in the soil and are left at the surface as water evaporates.

For optical remote sensing, the possibility to identify salt-affected soils varies largely with moisture content, salt pureness, and spectral contrast with other spectral features. Salts generally have featureless spectra with no known narrow absorption bands linked specifically to salinity status (Ben-Dor and Banin, 1990; Csillag *et al.*, 1993). Crowley (1991) showed that many saline minerals exhibit diagnostic NIR absorptions mainly attributed to vibrations of hydrogen-bonded structural water, allowing the detection of minor hydrate phases within mineral mixtures dominated by halite. Dehaan and Taylor (2002) found salinized soils in the Murray-Darling Basin of Australia to have distinctive spectral features in the VIS-NIR related to combined water in the hydrated evaporite minerals, allowing discrimination of minerals such as gypsum and polyhalite. A reflectance high (shoulder) at 800 nm was observed in all soils containing halite with a slope reduction between 800 and 1300 nm. They also showed clay mineral absorption depths at 2200 nm (hydroxyl absorption feature) to be reduced as salinization increased in soils.

Surface salinity is a highly dynamic process, causing identification constraints derived from the spatial and temporal variable nature of soil drying (precipitated salts). Salt detection is easier at the end of the dry season, since salts will dissolve during the rainy season, and upon drying, salts tend to concentrate locally, in patchy spots, creating much heterogeneity within a pixel. In contrast to white saline surfaces, pure alkaline soils are usually dark at the surface, because excess sodium causes organic matter to disperse when the soil is moist (Metternicht and Zinck, 2003).

There are numerous remote sensing studies utilizing hyperspectral data to map and monitor salt-affected soils. Csillag *et al.* (1993) used a stepwise principal components analysis on a large set of salt-affected soils and found important spectral sensitivities to salt content in the visible (530–770 nm), NIR (900–1030 nm), and SWIR (2150–2310 and 2330–2400 nm) spectral regions. They also found that coarser bandwidths of 40 nm provided similar accuracies in discerning salinity status, suggesting that broader band sensors such as Landsat TM could provide comparable accuracies (Dwivedi and Rao, 1992). Shi *et al.* (2003) demonstrated the feasibility of detecting physicochemical properties of saline soils in reclaimed coastal areas in Zhejiang Province in China with hyperspectral lab spectra. Ben-Dor *et al.* (2002) have successfully demonstrated the use of advanced spectroscopy techniques in the

quantitative retrieval of soil properties based on their spectral absorption features, employing a "Visible and Near-infrared Analysis" (VNIRA) methodology to predict soil chemical constituents. The VNIRA approach assumes that a concentration of a given constituent is proportional to the linear combination of several absorption features, and is capable of generating empirical models for predicting soil properties using laboratory wet chemistry and spectral measurements on representative sets of "calibration" samples.

The utility of hyperspectral remote sensing of soils at the landscape scale has been demonstrated in numerous studies involving airborne and spaceborne imagery. NASA's Advanced Visible/Infrared Imaging Spectrometer (AVIRIS), a 224 band airborne imaging spectrometer with 10-nm nominal spectral resolution, has been applied to the detection and mapping of minerals, including clays and iron oxides (Clark *et al.*, 1990). Crowley (1993) mapped various hydrated chlorides and sulfates of sodium, potassium, calcium, and magnesium with AVIRIS. Palacios-Orueta and Ustin (1998) successfully applied lab soil spectra to train AVIRIS data in their mapping of organic matter and iron content levels in soils. Using airborne hyperspectral data in northern Israel, Ben-Dor *et al.* (2002) showed the capability of soil surface mapping with VNIRA, deriving quantitative maps of soil properties such as organic matter, clay content, moisture, and soil salinity. They employed similar methods to successfully predict carbonates from the SWIR region (Ben-Dor and Banin, 1990). Dehaan and Taylor (2002) showed that salinized soils could be mapped with airborne hyperspectral HyMap imagery using spectral matching techniques, such as the Spectral-Feature-Fitting (SFF) technique, of field spectra with airborne hyperspectral data. SSF is an absorption feature–based method that matches image spectra to reference spectra using a least-squares technique and continuum-removed data.

The Jet Propulsion Lab (JPL, **http://speclib.jpl.nasa.gov/documents/jpl_desc.htm**) and US Geologic Survey (USGS, **http://speclab.cr.usgs.gov/spectral.lib04/spectral-lib04.html**) maintain digital spectral libraries that provide laboratory measured mineral and organic spectra of a wide range of terrestrial surface constituents found in soils (Grove *et al.*, 2004; Clark *et al.*, 1993). Soil spectral libraries have been used to greatly aid in the characterization of soil properties (Shepherd and Walsh, 2002). Their utility is enhanced with the availability of chemicophysical information, descriptions of spectral variability for a given soil, and information on sampling location and soil genetic factors such as parent material, age, vegetation, topography, and climate.

SOIL MOISTURE

Soil moisture plays a critical role in the evaluation of terrestrial environmental conditions with important influences on hydrologic, pedologic, biogeochemical, ecologic, and atmospheric processes (Goward *et al.*, 2002). Soil moisture acts as an integrator of the occurrence, distribution, and amount of precipitation and its spatial and temporal patterns, in turn, affect the partitioning of surface available energy into sensible and latent heat fluxes, and precipitation into surface runoff, storage, and evaporation. The growth of terrestrial vegetation is largely sustained by the water supply in the soil with the subsoil supplying water to plants and respiring microbial organisms long after the surface has dried out. Despite its importance, there has been limited success in implementing remotely sensed soil moisture observations at the appropriate time and space scales needed for current hydrology, climate, and biogeochemical models (Sellers *et al.*, 1995). Accurate assessments of this variable have been difficult due to complex land cover conditions and extreme soil moisture variability across landscapes, with unknown scale dependencies (Wood, 1997). A fundamental limitation of remote sensing is that only the surface soil moisture which is a few millimeters (optical) or centimeters (microwave) deep is observed and near-surface soil moisture changes more quickly than soil moisture at greater depths, making it very difficult to infer soil wetness below a few centimeters.

The major effect of adsorbed water on soil reflectance is a decrease in reflected energy throughout the shortwave portion of the electromagnetic spectrum (see Figure 3a). Wet soils exhibit distinguishable spectral absorption features in the NIR and SWIR regions, despite the great variability in their signatures. Combination modes of three fundamental vibrations yield water absorption features at 950 nm (very weak), 1200 nm (weak), 1400 nm (strong), and 1900 nm (very strong) (Bowers and Hanks, 1965; Ben-Dor *et al.*, 1999). Generally, strong atmospheric water absorptions at 1400 and 1900 nm restrict the use of these spectral regions, and reflectance measurements at 1500–1730 nm and 2080–2300 nm are more commonly used for observations of surface soil moisture. Jacquemoud *et al.* (1992) found it difficult to model soil spectral reflectance spectra as a function of water content due to saturation at the 1900 nm absorption feature.

Angstrom (1925) attributed the effect of darkening to internal reflections within the water film adsorbed to the soil particle surfaces, where multiple reflections between water and soil particle surface cause increased absorption and decreased reflection,

$$\rho_{\text{wet}} = \frac{\rho_{\text{dry}}}{[n^2(1-\rho_{\text{dry}}) + \rho_{\text{dry}}]} \tag{1}$$

with 'n' as the index of refraction for the adsorbed water ($n = 1.33$ for pure water). The decrease in reflectance is proportional to the thickness of the water film around the soil particles and can be related to the gravimetric water

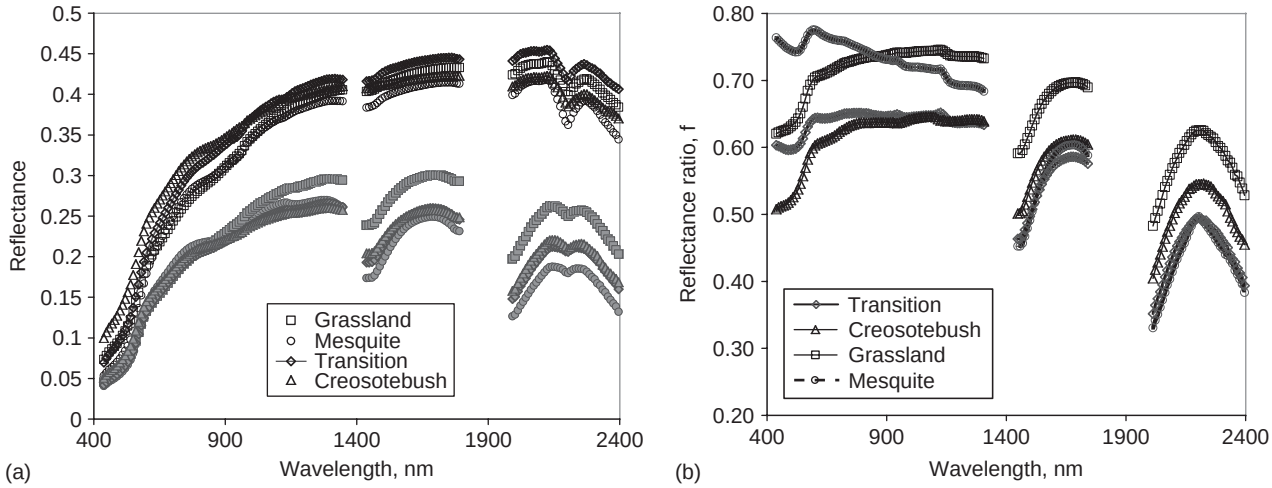


Figure 3 (a) Dry and wet soil spectral signatures for soils at Jornada Experimental Range, New Mexico for different land cover types and (b) the wet-to-dry reflectance ratios

content as well as energy status of the adsorbed water (Idso *et al.*, 1975).

Many controlled studies under laboratory conditions have shown an exponential decrease in reflectance with increasing soil water content (Bedidi *et al.*, 1992; Muller and Décamps, 2001). This nonlinear dependence on moisture varies with soil texture, but was found to be similar across soil types when moisture was expressed as degree of saturation, s,

$$s = \frac{\theta}{(1 - \rho_b/\rho_p)} \tag{2}$$

where θ is the volumetric water content and ρ_b, ρ_p are the bulk and particle densities of the soil respectively (Lobell and Asner, 2002). Lobell and Asner (2002) also applied the absorption depth at 2200 nm, in the more sensitive water-absorbing SWIR region, as a measure of moisture content using the relationship,

$$D = 1 - \frac{\rho_{2200}}{\rho_c} \tag{3}$$

where D is the continuum-removed band, ρ_{2200} is the measured reflectance at 2200 nm, and ρ_c is the value predicted from linear interpolation of reflectance at 2140 and 2260 nm.

Leger *et al.* (1979) studied the interactions of soil water with organic matter and iron oxides and concluded that the interactions among all three components were more important in determining the resulting soil spectra than considering each component individually. Ben-Dor *et al.* (1999) distinguished among the unique spectral influences of hydration water incorporated into the lattice of the mineral (e.g. gypsum), hygroscopic (air dry) water adsorbed tightly as a thin layer onto clays and humic colloids, and free water held in pore space. The absorption effects of hygroscopic water were governed by the relative humidity (vapor pressure deficit) of the atmosphere and, along with pore space water caused soil albedo to change drastically upon wetting.

In field-based measurements, Idso *et al.* (1975) found a linear function of reflectance measurements with the water content of the soil surface. They demonstrated three spatial/temporal stages of drying and were able to distinguish the transition points between the different stages of soil drying with the surface albedo measurements (Idso *et al.*, 1974):

- *Stage 1:* The wet soil surface is at potential evaporation and controlled by atmosphere conditions.
- *Stage 2:* Transition between wet and dry whereby the soil's hydraulic conductivity decreases to the extent that subsurface water is unable to move to the surface fast enough to meet evaporative demand of the atmosphere.
- *Stage 3:* Dry soil surface with low, nearly constant evaporation rate, controlled by the adsorptive forces of the soil.

Under conditions of high evaporative demand, a fairly wet soil may rapidly "dry" at the surface where optical measurements are made. Since the soil drying process is spatially nonuniform, all three stages of drying may contribute to the integrated response of reflected/ emitted energy from remotely sensed measurements. Thus, both linear and nonlinear reflectance variations are present under most soil wetting and drying conditions. The spatial and temporal variability of soil moisture is significant even at fine scales, making *in situ* soil measurements for calibration of remote sensing systems difficult.

Obukhov and Orlov (1964) noted that the spectral curves do not change in appearance with wetting and that the ratio

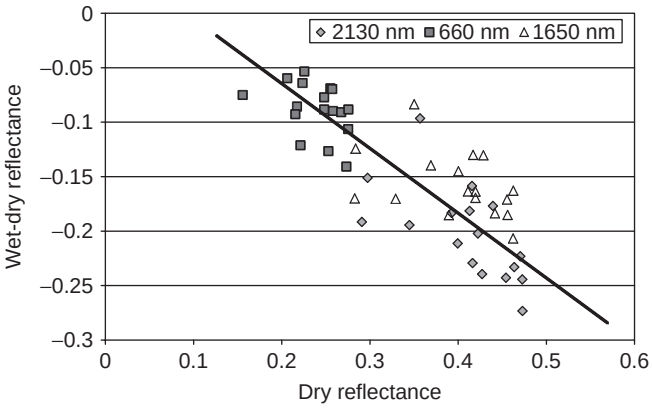


Figure 4 The difference in wet-to-dry soil reflectance plotted against dry soil reflectance for the Argentina and US sites over three wavelengths

of wet soil to dry soil reflectance remains fairly constant in the visible spectrum. The reflectance ratio between wet and dry soils, however, varies with soil type and decreases significantly in the SWIR region (see Figure 3b). The wet soil spectral signatures shown in Figure 3 were 25 to 60% lower in the VIS-NIR (ratios of 0.75 to 0.40) and 60 to 80% lower in the SWIR (Figure 3b). Baumgardner *et al.* (1985) showed that the wet to dry difference was greater in forest soils with low organic matter than in darker, grassland soils with higher organic matter contents. This is also seen in Figure 4, in which there is a linear change in wet–dry reflectance differences such that the brightest soils show the largest decrease in reflectance upon wetting. Knowledge of these changes is important in providing realistic soil albedo values, which are arbitrarily specified in almost all land surface energy and hydrology models.

Many earth-observing sensors have SWIR bands of potential use in surface moisture studies, including Landsat Thematic Mapper bands 5 and 7 (1550–1750 nm, 2080–2350 nm), MODIS (Moderate Resolution Imaging Spectroradiometer) bands 6 and 7 (1628–1652 nm, 2105–2155 nm), and SPOT-VEGETATION (1580–1750 nm), and offer potential soil water indicators. Water absorption in these bands can be expressed as a ratio or in linear combination and then related to soil water content for discrete soil textural classes (Levitt *et al.*, 1990). Huete and Warrick (1990), however, reported limitations in assessing soil moisture content at the surface (0–5 cm) with Thematic Mapper moisture bands, or with various "wetness" indicators under partial vegetation canopies with low (10–20%) vegetative covers.

SOIL PHYSICAL PROPERTIES

Geometric-optical variations of the soil surface yield important textural and morphological information relevant to hydrologic processes such as water infiltration, runoff, and moisture retention. Soil texture is important in modifying the movement of water within a soil, how much water is stored in the soil, and the energy with which it is held. Water will move most rapidly in coarser textured soils and a healthy soil structure promotes infiltration of water into the soil and reduces runoff and erosion. Soil texture and roughness cause significant decreases in reflectance with increasing particle and aggregate size (Baumgardner *et al.*, 1985). Bowers and Hanks (1965), working with kaolinite of different sizes, found an exponential increase in reflectance with decreasing particle size at VIS-NIR wavelengths. As particle size or aggregates become coarser, there are more shadows present and more incident flux is multiply scattered and eventually absorbed in so-called light traps. In general, one observes clayey soils to appear darker than the coarser sand particles due to the aggregated nature of clays and the higher presence of organic matter on clay soils. Ben-Dor *et al.* (1999) noted the important physical chromophores as particle size, sensor viewing angle, and solar incident angle and stated that these influences may change the shape of soil spectra, such as a baseline shift or absorption intensity, but do not alter the position of the diagnostic spectral features.

Structure is an important property of the soil that controls geometric-optical behavior, related to shading and roughness. Structureless soils tend to be the brightest, while healthy-structured soils are darker with well-defined aggregates and peds. Soil crusts influence soil surface roughness, water infiltration, retention, and overland flow. Biological soil crusts, composed of cyanobacteria, fungi, lichens, and mosses are an important surface component in arid and semiarid ecosystems and influence soil stability and local hydrologic patterns both directly and indirectly. They have profound influences on soil spectra through the presence of chlorophyll containing microorganisms (Karnieli *et al.*, 1999).

The physical crust that forms on soil surfaces during rain events also alters soil hydrology by reducing infiltration and increasing runoff. The high energy content of falling raindrops disintegrates soil aggregates and rearranges soil particles and clays within the structural crust of the soil surface. Goldshleger *et al.* (2004) found that clays may either be washed out or be enriched within the soil crust zone dependent on soil texture and mineralogy. They noted that the process of soil structural crusting resulted in distinct spectral reflectance differences. Ben-Dor *et al.* (2004) demonstrated the use of airborne, hyperspectral reflectance measurements of crusted and noncrusted soil surfaces for mapping infiltration rates of loess soils in Israel. Saline crusts vary in structure from smooth to rough, and in color, from white to dark, causing large variations in reflectance spectra (Escadafal *et al.*, 1989). Salt mineralogy has been found to produce distinctive

macromorphological features, such as puffy crusts with the abundance of sodium sulfates, and smoother salt crusts with the presence of chlorides (Eghbalm *et al.*, 1989). Overall, salty crusts are smoother than nonsaline surfaces and have an overall higher reflectance.

Soil surfaces are not Lambertian surfaces and scatter incident radiation anisotropically, and reflectance will vary with direction of illumination and viewing. A fundamental and intrinsic property governing the reflectance behavior of a surface is described by the Bidirectional Reflectance Distribution Function (BRDF), which specifies the behavior of surface reflectance and scattering as a function of view and illumination angles for a given wavelength (λ). The equation describing this function is:

$$\text{BRDF} = L\frac{(\theta_s, \phi_s, \theta_v, \phi_v, \lambda)}{E(\lambda)} \tag{4}$$

with units of sr^{-1}. L is surface leaving radiance in $Wm^{-2}\,sr^{-1}$ and E is the irradiance incident on the surface in Wm^{-2}. θ_s and ϕ_s refer to the solar zenith angle and solar azimuth, respectively, while θ_v and ϕ_v are the sensor view zenith and view azimuth angles, respectively. Often, however, it is defined simply as the bidirectional reflectance factor (BRF $= \pi$BRDF) at a multitude of view zenith and azimuthal angles for a given sun position (Walthall *et al.*, 1985). The BRDF is a physical property of the surface that may be used to derive geometric descriptors of a surface, such as size, shape, and orientation of surface "roughness" elements (Irons *et al.*, 1992). BRDF models are also useful in correcting for sun-soil surface-sensor geometry effects on spectral reflectance (Liang and Strahler, 1994).

Raina *et al.* (1993) were able to incorporate surface terrain and textural characteristics in mapping erosion and salinity classes from Landsat TM imagery. Sensors such as Systeme Probatoire pour l'Observation de la Terre (SPOT) and ASTER greatly improve surface terrain characterization by providing both fine spatial resolution (10-m panchromatic band and 20-m multispectral bands) and stereo-capabilities with pointable sensors, allowing improved mapping of drainage patterns, floodplains, relief, landscape stability, and soil erosion potential, and other structural features.

In summary, the overall complexity of soil spectra prevents a straightforward assessment of reflectance properties by physical models. The spectral response of soils is a product of the interactions between chemical and physical chromophores, and the presence of one chemical constituent, such as organic matter, can mask the presence of other constituents (iron, clays) in the soil matrix. Furthermore, the spectral influence of an absorbing constituent varies with soil physical properties, such as particle size distribution, and the presence of certain chemical constituents, such as salts, can alter a soil's physical properties. This makes it very difficult to model complex mixtures of mineral and organic chromophores with simple linear mixture models. Despite these challenges, soil spectra do carry significant information about many soil properties and for the most part, high spectral resolution data are needed over the spectrally active regions of all chromophores to derive soil properties.

MIXED SOIL AND VEGETATION SPECTRA

A significant problem in extracting soil information from remote sensing imagery is the presence of nonphotosynthetically active vegetation (NPV), litter, and vegetation in pixels. Various methods are used to extract the variability of the soil signal from the vegetation and NPV spectra, including spectral mixture analysis (SMA) and linear unmixing techniques (Smith *et al.*, 1990). The basis of a linear mixture model is that the measured spectral response of a pixel is equal to the weighted sum of multiple reflecting spectral features,

$$d_{ik} = \sum_{j=1}^{n} r_{ij}c_{jk} + \varepsilon \tag{5}$$

where d_{ik} is the measured spectral response of spectral mixture k in waveband i, n is the number of independent reflecting components in the mixture, r_{ij} is the response of component j in waveband i, c_{jk} is the relative contribution of component j in spectral mixture k, and ε is the residual error.

Roberts *et al.* (1993) showed how fine spectral resolution AVIRIS (224 bands) can be used to unmix the soil spectral signal from green vegetation and NPV. Asner and Lobell (2000) developed a SWIR-based, tied endmember approach to the separation of green vegetation, NPV, and soil (see Figure 5). Some shortcomings of linear mixing models are that knowledge of all important endmembers may be required and that significant nonlinear mixing may be present within pixels. Palacios-Orueta *et al.* (1999) utilized a more advanced, hierarchical foreground and background analysis (HFBA), to extract soil information from AVIRIS imagery. The technique employs subsets of foreground and background spectra to emphasize the presence of the spectra of interest and then uses singular value decomposition to extract this signal while simultaneously minimizing the background or undesired spectral variation (Pinzón *et al.*, 1998).

Richardson and Wiegand (1977) and Kauth and Thomas (1976) formulated the "soil-line" concept as a way to separate soil-induced optical variations from those of vegetation. This provided a vegetation measure relative to the wetting and drying behavior of underlying soils and enabled the assessment of surface soil moisture conditions in vegetated

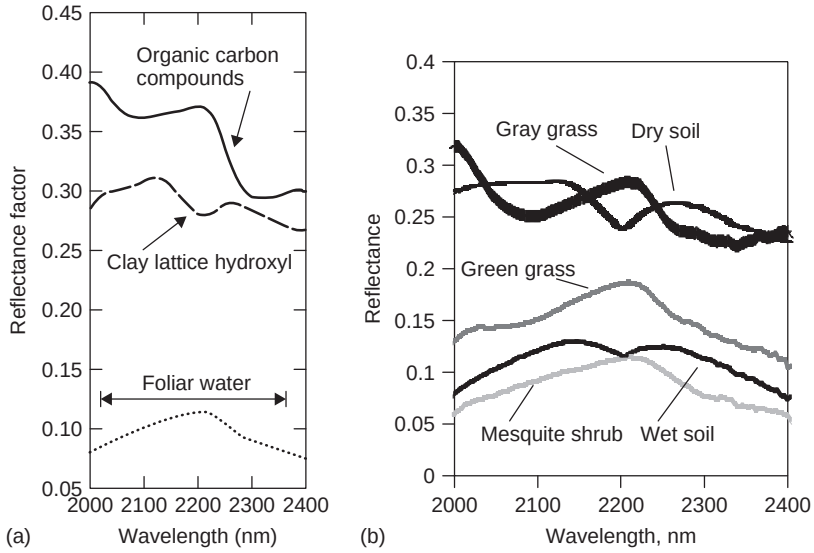


Figure 5 (a) Typical spectra of litter (solid), soil (dashed), and green vegetation (dotted) in the SWIR region (2000–2400 nm) (Reprinted from Asner and Lobell, 2000. ©2000, with permission from Elsevier); (b) Example SWIR spectra from Walnut Gulch Experimental Watershed, Arizona

pixels. When wet, drying, and dry soil spectra are plotted in two-band reflectance space, a "soil-line" is formed with a slope that represents the proportional change in reflectance in each band with soil wetting and drying (Galvao and Vitorello, 1998). In Figure 6, two soil-line derivations are presented, an 'NIR-red' soil line, which is used in the formulation of many vegetation indices, and an "NIR-SWIR" soil line, based on two bands that are increasingly being used to extract land surface moisture information from soil and vegetated surfaces. Overall, soil variations formed soil lines encompassing the dry and wet soil spectral behavior for a range of different soils. The NIR-SWIR soil line has a much lower slope (0.517) than the NIR-red soil line (1.192) due to the greater sensitivity of the SWIR region to soil wetting. The SWIR soil line also shows a greater dispersion of points among the different soils (Figure 6). NIR-red soil-line slopes have been shown to vary greatly over different soil types, ranging from 1.0 to 1.4, with organic-rich peat soils having slopes as high as 3 (Huete *et al.*, 1984; Baret *et al.*, 1993).

In Figure 7, the NIR-SWIR soil line is plotted along with vegetation spectra, using 865 nm and 1640 nm wavelengths. The soil line has a slope that is less than one (0.643), owing to the higher sensitivity of the SWIR band to moisture compared with the NIR band. Although, the soil and litter points are very dispersed, there is a good overall separation between variations attributed to vegetated pixels from those due to soil moisture variations (Figure 7). Variations associated with different vegetation amounts vary along an axis, orthogonal to the soil line, allowing information of soil and vegetation to be extracted that could otherwise not be obtained from any

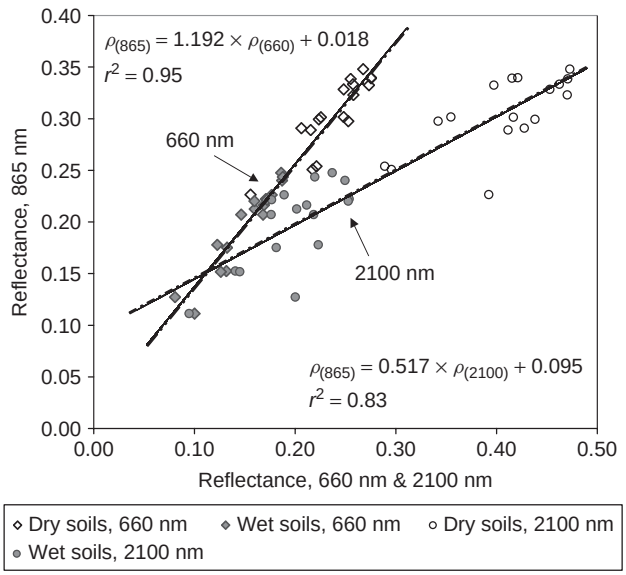


Figure 6 The NIR-Red and NIR-SWIR (2100 nm) ''soil lines'' for the Argentina and US soils from Nacuñan, Chancani, Jornada, and Walnut Gulch

one band by itself. Both increasing vegetation amounts and soil moisture cause the SWIR reflectance to decrease, but in the case of vegetation, NIR reflectance simultaneously increases while in the case of soil moisture, NIR reflectance decreases. Sparsely vegetated points, however, fell within the cloud of soil spectra and could not be resolved from soil. In these cases, the formulation of fine resolution, specific soil lines would improve upon the separation of vegetation and soil moisture reflectance variations.

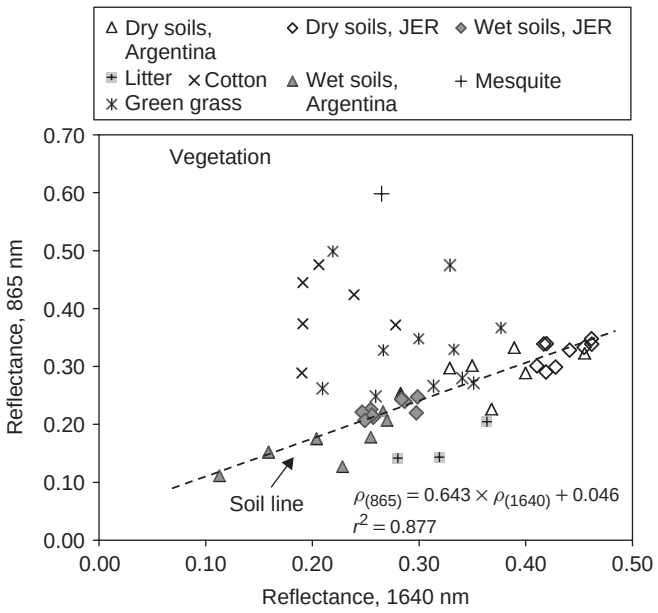


Figure 7 Soil and vegetation spectra plotted in NIR and SWIR (1640 nm) space

Although the presence of vegetation interferes with the reflectance of soil features, one can use vegetation spectra as indirect indicators of soil properties. Wiegand *et al.* (1994) found a high correlation between spectral vegetation index values and soil electrical conductivity, enabling the separation of saline-alkaline soils from nonaffected soils. Zhang *et al.* (1997) used vegetation spectra of halophytic salt-tolerant vegetation (e.g. Cynodon dactilon) as indicators of saline-alkaline affected areas and soil salinity. Ustin (1999) studied the vegetation spectral variations found over Quebec soils with heavy metals and Kooistra *et al.* (2004) utilized hyperspectral reflectance measurements of plant communities as indicators of toxic metal (Ni, Cd, Cu, Zn, and Pb) contamination in Dutch river floodplain soils in the Netherlands. Other studies have focused on the relationships between vegetation indices, rainfall, and soil moisture (Nicholson and Farrar, 1994). They noted significant differences in NDVI-rainfall relationships for various soil types in semiarid Botswana, Sahel, and found a linear relationship between NDVI and rainfall in environments where annual rainfall did not exceed 500 mm year^{-1}, or 50–100 mm per month. Above these limits, a saturation response occurs and NDVI increases with rainfall much more slowly. They further found rain-use efficiency (NDVI/rainfall) to be more a function of the underlying soil than of the vegetation formation.

SOIL-VEGETATION-MOISTURE OPTICS

Knowledge of the water status of a vegetation canopy can provide valuable information on soil moisture condition, vegetation stress, and drought status. Laboratory studies and canopy radiant transfer model simulations have shown that changes in leaf water content have a large effect on reflectances in portions of the NIR and SWIR spectral regions. Important liquid water absorptions are found at 1640 and 2100 nm, with a weaker absorption feature at 1240 nm, which result in negative relationships between reflectances at these wavelengths and leaf water content (Hunt *et al.*, 1989; Gao, 1996). Several studies have shown increases in leaf reflectance associated with plant stress across the SWIR region (Gausman and Allen, 1973; Tucker, 1980), providing useful information to infer soil moisture status in the plant root zone.

The SWIR reflectance values alone, however, are not suitable for retrieving vegetation water content as variations in leaf internal structure and leaf dry matter content as well as canopy geometry, shadowing, and soil surface moisture also influence SWIR reflectance (Cohen, 1991; Ceccato *et al.*, 2001). The assessment of vegetation water content or equivalent water thickness (EWT, g-H_2O/cm^2-leaf area) is significantly improved by the combination of NIR and SWIR bands (Ceccato *et al.*, 2002). For example, the Moisture Stress Index (MSI), calculated as the simple ratio between SWIR (1600 nm) and NIR (820 nm) spectral bands, has been successfully used to derive leaf scale functions of water content (Hunt *et al.*, 1989),

$$\mathrm{MSI} = \frac{\rho_{\mathrm{SWIR}}}{\rho_{\mathrm{NIR}}} \tag{6}$$

Hardisky *et al.* (1983) developed the Normalized Difference Infrared Index (NDII), contrasting the NIR with SWIR (~1650 nm) wavelengths,

$$\mathrm{NDII} = \frac{[\rho_{\mathrm{NIR}} - \rho_{\mathrm{SWIR}}]}{[\rho_{\mathrm{NIR}} + \rho_{\mathrm{SWIR}}]} \tag{7}$$

and found this index to be strongly correlated with canopy water content. Several variants of this index are now used in a wide range of studies using high spectral resolution as well as broadband reflectances. Xiao *et al.* (2004) used the normalized difference between the NIR and SWIR (1580–1750 nm) bands as measures of land surface moisture status with SPOT-4 VEGETATION (VGT) sensor data and named this the Land Surface Water Index (LSWI). Fensholt and Sandholt (2003) formulated the Shortwave Infrared Water Stress Index (SIWSI) from daily MODIS NIR and SWIR (1628–1652 nm) band sensor data,

$$\mathrm{SIWSI} = \frac{[\rho_{\mathrm{SWIR}} - \rho_{\mathrm{NIR}}]}{[\rho_{\mathrm{SWIR}} + \rho_{\mathrm{NIR}}]} \tag{8}$$

They reported high correlations between the SIWSI and soil moisture in the root zone in a Sahel vegetation study

in Senegal. Ceccato *et al.* (2002) proposed a Global Vegetation Moisture Index (GVMI) to retrieve EWT ($g\,m^{-2}$) from SPOT-VGT sensor data at the canopy level,

$$\mathrm{GVMI} = \frac{[(\rho_{\mathrm{NIR}}* + 0.1) - (\rho_{\mathrm{SWIR}} + 0.02)]}{[(\rho_{\mathrm{NIR}}* + 0.1) + (\rho_{\mathrm{SWIR}} + 0.02)]} \quad (9)$$

where $\rho*$ is the reflectance value of a rectified NIR band using apparent, top-of-atmosphere (TOA) reflectances as input.

Whereas the SWIR region responds to both vegetation water content and soil surface moisture, the weaker 1240-nm water-absorbing region has been shown to respond to canopy moisture status only and to be insensitive to surface soil moisture. The Normalized Difference Water Index (NDWI) uses two reflectance bands in the high NIR reflectance plateau of vegetation canopies, at 860 and 1240 nm wavelengths (Gao, 1996),

$$\mathrm{NDWI} = \frac{[\rho_{860\,\mathrm{nm}} - \rho_{1240\,\mathrm{nm}}]}{[\rho_{860\,\mathrm{nm}} + \rho_{1240\,\mathrm{nm}}]} \quad (10)$$

with the weaker, liquid water absorption feature enhanced by the high NIR scattering in the leaf. This formulation was applied to MODIS bands 5 (1230–1250 nm) and 2 (841–876 nm) and was found to be a strong indicator of canopy water content during the growing season in the Sahel (Fensholt and Sandholt, 2003). However, it was found that in dry years the vegetation cover was too dry to provide information on canopy water content, suggesting that a minimum, threshold vegetation amount must be present for the water indices to work. The water band index (WBI $= \rho_{900\,\mathrm{nm}}/\rho_{970\,\mathrm{nm}}$) is also an NIR-based indicator of canopy water content (Peñuelas *et al.*, 1997).

To investigate differences among these vegetation water indices, we compared the NDWI (1240 nm) and the LSWI (1640 and 2130 nm) formulations over a range of green canopy cover conditions with dry and wet surface soil states (see Figure 8). The NDWI primarily responded to the amount of vegetation moisture, as related to green vegetation cover, and was minimally affected by the wetness condition of the soil surface. The LSWI based on 1640 and 2100 nm bands had higher dynamic ranges and were sensitive to both vegetation moisture and surface soil wetness conditions (Figure 8). The wet soil backgrounds for a given amount of vegetation resulted in higher LSWI values, especially with the 2100-nm index. Hence the use of both NDWI and LSWI would allow soil surface moisture influences to be analyzed separately from those due to vegetation moisture status. All three indices appeared very unstable over bare soil conditions with large variations in values, which may restrict their use in more arid regions. The NDWI values also did not become positive until a 60% green canopy cover was present.

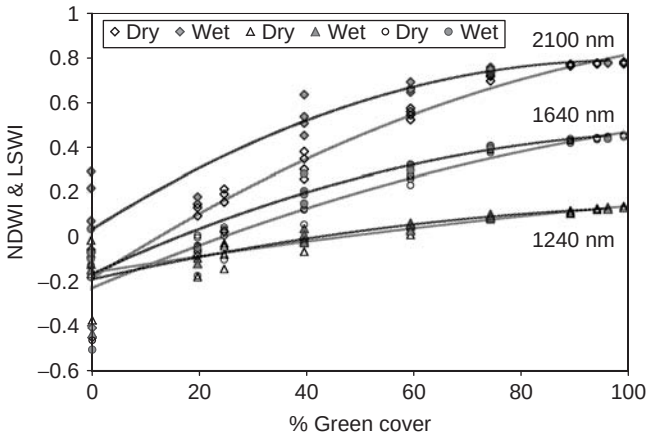


Figure 8 NDWI (1240 nm) and LSWI (1640, 2100 nm) plotted against increasing levels of green cotton cover with dry and wet soil backgrounds

COMBINED MOISTURE AND VEGETATION INDICES

Vegetation water indices employing the 1240-, 1640-, or 2100-nm wavelengths in lieu of the red band used in vegetation indices have recently been used as independent vegetation measures related to canopy moisture condition instead of chlorophyll amount. Although vegetation indices (VIs) have also been correlated with vegetation water content, they are physiologically related to canopy chlorophyll content and absorbed photosynthetically active radiation (Tucker and Sellers, 1986). Thus, VIs generally would depict decreases in plant growth (or senescence) caused by water stress rather than lower water contents. At the Soil Moisture Experiments 2002 (SMEX02) field and satellite campaign, Jackson *et al.* (2004) found the water indices to be superior to VIs in mapping vegetation water content (VWC), as the NDVI was found to saturate, while the water indices continued to show changes in VWC with increasing amounts of green vegetation. Ceccato *et al.* (2002) concluded that VIs were unsuitable for retrieving VWC since relationships between chlorophyll and VWC are specific to each species. Furthermore, decreases in chlorophyll content do not imply a decrease in VWC, and vice versa, a decrease in VWC does not imply a decrease in chlorophyll content.

We combined three vegetation water indices (VWI) with vegetation indices (VI) to ascertain variations specific to each. A combined VI-VWI approach would provide a contextual array of remotely sensed measurements, increasing the measurement domain and reducing the underdetermination problem for assessing soil moisture and vegetation water content (Goward *et al.*, 2002; Zarco-Tejeda *et al.*, 2003). Goward *et al.* (2002) explored this contextual approach for surface moisture studies in the thermal and reflective portions of the spectrum. In Figure 9, the NDWI of soil and vegetation field measurements are plotted with

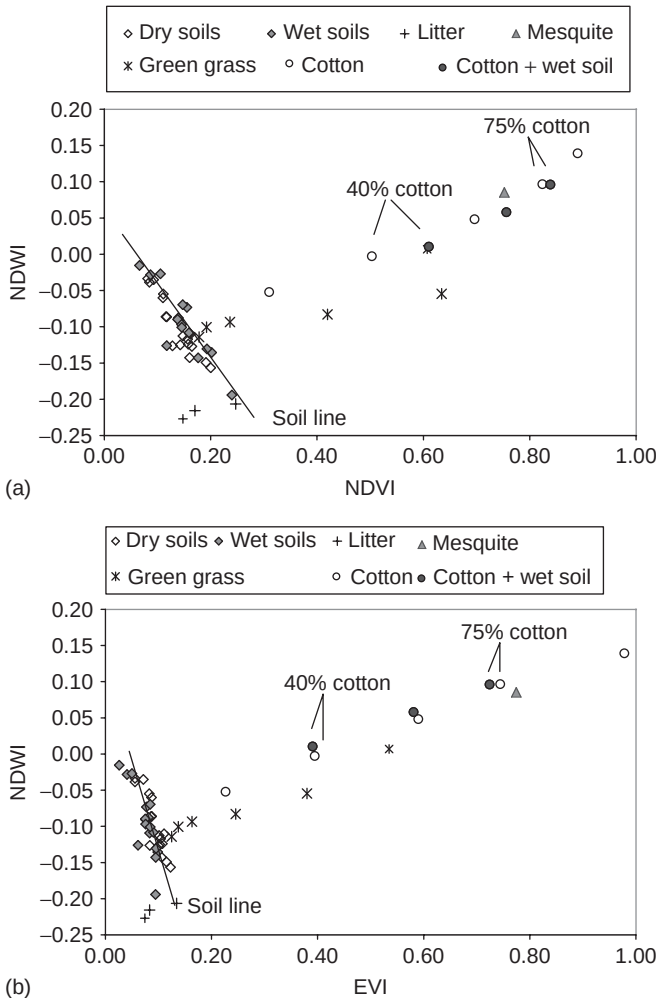


Figure 9 NDWI (1240 nm) plotted against (a) NDVI and (b) EVI for soil and vegetation variations in moisture

the two VIs adopted as standard MODIS products, the normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI),

$$\mathrm{NDVI} = \frac{[\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}]}{[\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}}]} \tag{11}$$

$$\mathrm{EVI} = 2.5\frac{[\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}]}{[1 + \rho_{\mathrm{NIR}} + 6\rho_{\mathrm{Red}} - 7.5\rho_{\mathrm{Blue}}]} \tag{12}$$

There is a strong negative, "soil-line" relationship among the dry and wet soil, as well as surface litter, points in the NDWI-NDVI plot. There is also a strong and positive relationship between NDVI and NDWI for the vegetated points, which increasingly separate from the soil line with increasing amounts of vegetation (Figure 9a). Wetting the soil of the 40% green cotton canopy resulted in higher NDVI values with minimal changes to the NDWI. Although the NDWI was spectrally insensitive to the wet soil, the higher NDVI value would result in error in predicting vegetation cover and associated water fluxes. It is worth noting that both indices in Figure 9(a) are needed for maximum discrimination, and hence detection, of variations in soil and vegetation moisture. The NDWI by itself, would not be able to unambiguously distinguish certain bare soils from 40% vegetation canopies as both resulted in NDWI values close to zero.

As the NDVI was strongly correlated to soil variations, we also analyzed the NDWI-EVI relationships and found a clearer separation of soil and vegetation variance (Figure 9b). The EVI was only weakly correlated with the bare soils and the NDWI and EVI values of the vegetated cases were insensitive to soil surface wetness condition. Thus, changes in NDWI and EVI values could be unambiguously related to vegetation, moisture, and chlorophyll, amounts. Although the NDWI was insensitive to soil wetness condition, it exhibited a considerable range in values for bare soils of varying soil types, (Figure 9a,b). This indicates the need for development a soil-adjusted version of the NDWI.

The equivalent cross plots using LSWI, instead of NDWI, are shown in Figure 10. In the NDVI plot, the bare wet soils deviated significantly from the dry bare soils with the wet soils increasing in both NDVI and LSWI values. The same occurs in the vegetated cotton canopy, where wetting the soil increased values of both NDVI and LSWI. This presents a dilemma since this is the same trajectory of increasing NDVI and LSWI values that vegetated pixels would take with increasing amounts of vegetation, limiting the utility of this combined-index cross plot. As with the NDWI plots, the NDVI-LSWI relationships also showed strong negative correlations for bare soils and litter. The LSWI-EVI plot, on the other hand, more clearly shows the LSWI to respond to moisture variations in both soils and vegetation, but in a distinct manner that separates the moisture signal of soil from that of vegetation. Thus, using both LSWI and EVI, one could distinguish between variations in moisture due to soil from those due to vegetation (Figure 10b).

The unique patterns observed between the two moisture indices, NDWI and LSWI, are shown in Figure 11. One can observe an overall correlation, between the two indices, however, dry soil surfaces are clearly separated (steepest slopes) from surfaces with wet soil and vegetation moisture (lower slopes). Wetting the soil of the 40% cotton canopy influences the LSWI but not the NDWI, as expected. As in the previous figures, there is a large variation in bare soil behavior found in both indices that may hinder the assessment of vegetation moisture in sparser and drier canopies. However, in contrast to the previous Figures 9 and 10, the bare wet soils are now separable from the bare dry soils. The sensitivity of these cross plots to vegetation water content, however, remains to be

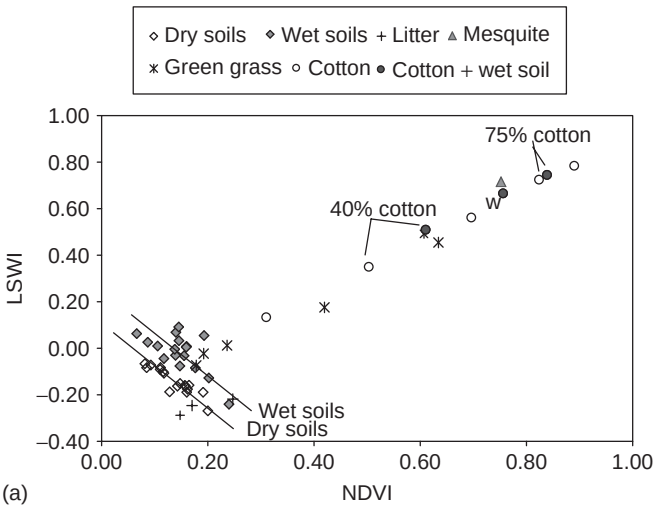


Figure 10 LSWI (2100 nm) plotted against (a) NDVI and (b) EVI for soil and vegetation variations in moisture

tested. These crossplots also do not consider root-zone soil moisture variations and the interactions of soil drying and wetting on chlorophyll versus water content levels in vegetation canopies. The combined use of vegetation indices (VI) and vegetation water indices may yield new insights into the monitoring of soil moisture and separation of moisture signals within a canopy. A combined VI and vegetation water index with thermal or microwave remote sensing approaches would also greatly enhance our understanding of land-surface moisture dynamics.

DISCUSSION

In this paper, we have summarized and also discussed the uses of hyperspectral and optical-based remote sensing techniques for quantitative retrieval of soil properties. Soils are shown to be complex mixtures of numerous absorbing

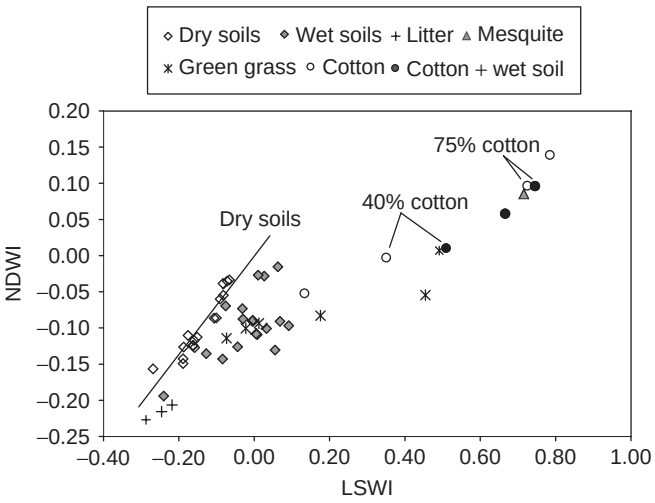


Figure 11 NDWI (1240 nm) plotted with LSWI (2100 nm) for a range of soil and vegetation moisture variations

constituents that require sophisticated methods of analysis to predict and map their abundances. Hyperspectral sensors and advanced spectroscopy and noise removal techniques have greatly improved the feasibility of unambiguously identifying numerous soil and vegetation absorption features, related to mineralogy, organic matter, liquid water, chlorophyll, cellulose, and lignin contents. Hyperspectral remote sensing is used to detect various chemical and physical soil properties either (i) directly from the pure soil pixels, (ii) through advanced spectroscopy methods in mixed "soil-vegetation-NPV" pixels, and (iii) by measurements of the overlying "impacted" vegetated canopy to infer soil properties and moisture status. Soil spectra carry significant information about many soil properties and for the most part, high spectral resolution data are needed over the spectrally active regions of all chemical and physical chromophores to derive soil properties. In most cases, the availability of laboratory or field soil spectra, with physicochemical soil information for calibration purposes, greatly improves upon quantitative retrieval of soil properties.

Although there are strong limitations in the use of remote sensing techniques for soil moisture assessment, optical remote sensing is useful for its spatial coverage and ability to integrate the spatial variability within an area, allowing for scale-dependent studies. Hyperspectral data is increasingly being used to establish relationships between important water absorption features and surface and subsurface soil water status. The water absorption features of both soil and vegetation have been successfully integrated into land-surface-water indices and applied to both hyperspectral and broadband sensor systems. Various water indices have been shown to respond uniquely to soil and vegetation moisture conditions, allowing for the separation of the two sources of variability, and possibly providing a more robust

methodology to assess near-surface, soil moisture conditions. Furthermore, spectral retrieval of both chlorophyll and water contents in vegetation canopies better depict plant physiological status and provide opportunities for improved drought, soil water deficit, and vegetation stress analysis. Much remains to be done, however, in the application of advanced spectroscopy to extract moisture properties of soils and canopies and new models are needed to optimize the water information present in hyperspectral data.

Up until recently, hyperspectral remote sensing was limited to the use of aircraft platforms and field instruments. The first space-borne hyperspectral sensor, Hyperion, was only launched in November 2000 on the Earth Orbiter 1 (EO-1) as part of NASA's New Millennium program. Hyperion is a pushbroom sensor providing 220-, 10-nm bands covering the spectrum from 400 to 2500 nm. Hyperion and AVIRIS imagery have demonstrated their utility in characterizing the ecological variance and complexities of landscapes, including species composition, ecosystem functioning, biogeochemical cycles, and land use, land cover change, and large-scale ecological changes. The success of these applications is providing new impetus for follow-on hyperspectral missions that would provide consistent data in spatial and temporal modes to assess, for example, how soils vary in response to climate and anthropogenic changes.

Combining optical moisture measurements with thermal and microwave measurements in a synergistic mode is also providing powerful new tools for moisture assessments and dynamics at the leaf, soil, and whole ecosystem levels. Various synergy approaches have been used, combining optical and microwave approaches in deriving surface hydrologic fluxes (Kustas *et al.*, 1998; Moran *et al.*, 2002). Microwave data offer improved information on soil, canopy, and landscape roughness useful in soil infiltration, runoff, and erosion models. Thermal measurements provide additional information on plant canopy and soil moisture status, and combined with optical information on soil properties and vegetation amounts, greatly improve hydrologic process studies related to soil moisture. In combination, hyperspectral, thermal, and microwave approaches will provide new insights into complex hydrologic processes, fluxes, and feedbacks among soil moisture, atmosphere moisture, and ecosystem metabolism and development.

REFERENCES

Angstrom A. (1925) The albedo of various surfaces of ground. *Geografiska Annaler*, **7**, 323.

Asner G.P. and Lobell D.B. (2000) A biogeophysical approach for automated SWIR unmixing of soils and vegetation. *Remote Sensing of Environment*, **74**(1), 99–112.

Baret F., Jacquemoud S. and Hanocq J.F. (1993) About the soil line concept in remote sensing. *Advances in Space Research*, **13**(5), 281–284.

Baumgardner M.F., Silva L.F., Biehl L.L. and Stoner E.R. (1985) Reflectance properties of soils. *Advances In Agronomy*, **38**, 1–44.

Bedidi A., Cervelle B., Madeira J. and Pouget M. (1992) Moisture effects on visible spectral characteristics of lateritic soils. *Soil Science Society of America Journal*, **153**, 129–141.

Ben-Dor E. and Banin A. (1990) Near infrared reflectance analysis of carbonate concentration in soils. *Applied Spectroscopy*, **44**, 1064–1069.

Ben-Dor E., Goldshleger N., Braun O., Kindel B., Goetz A.F.H., Bonfil D., Margalit N., Binaymini Y., Karnieli A. and Agassi M. (2004) Monitoring infiltration rates in semiarid soils using airborne hyperspectral technology. *International Journal of Remote Sensing*, **25**(13), 2607–2624.

Ben-Dor E., Irons J.R. and Epema G.F. (1999) Soil reflectance. In *Remote Sensing for the Earth Sciences: Manual of Remote Sensing*, Rencz A.N. (Ed.), Wiley & Sons: New York, pp. 111–188.

Ben-Dor E., Patkin K., Banin A. and Karnieli A. (2002) Mapping of several soil properties using DAIS-7915 hyperspectral scanner data- a case study over clayey soils in Israel. *International Journal of Remote Sensing*, **23**(6), 1043–1062.

Bowers S.A. and Hanks R.J. (1965) Reflection of radiant energy from soils. *Soil Science Society of America Journal*, **2**, 130–138.

Ceccato P., Flasse S., Tarantola S., Jacquemoud S. and Gregoire J.-M. (2001) Detecting vegetation leaf water content using reflectance in the optical domain. *Remote Sensing of Environment*, **77**(1), 22–33.

Ceccato P., Gobron N., Flasse S., Pinty B. and Tarantola S. (2002) Designing a spectral index to estimate vegetation water content from remote sensing data: part 1: theoretical approach. *Remote Sensing of Environment*, **82**(2–3), 188–197.

Clark R.N., King T.V.V., Klejwa M., Swayze G.A. and Vergo N. (1990) High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research*, **95**(B-8), 12653–12680.

Clark R.N., Swayze G.A., Gallagher A.J., King T.V.V. and Calvin W.M. (1993) *The U. S. Geological Survey, Digital Spectral Library: Version 1: 0.2 to 3.0 microns*, Open File Report 93–592, U.S. Geological Survey, p. 1340.

Cohen W.B. (1991) Temporal versus spatial variation in leaf reflectance under changing water stress conditions. *International Journal of Remote Sensing*, **12**, 1865–1876.

Crowley J.K. (1991) Visible and near-infrared (0.4–2.5 small mu, greekm) reflectance spectra of playa evaporite minerals. *Journal of Geophysical Research*, **96**(16), 231–240.

Crowley J.K. (1993) Mapping playa evaporite minerals with AVIRIS data: a first report from death valley, California. *Remote Sensing of Environment*, **44**(2–3), 337–356.

Csillag F., Pasztor L. and Biehl L.L. (1993) Spectral band selection for the characterization of salinity status of soils. *Remote Sensing Environment*, **43**, 231–242.

Dehaan R.L. and Taylor G.R. (2002) Field-derived spectra of salinized soils and vegetation as indicators of irrigation-induced soil salinization. *Remote Sensing of Environment*, **80**(3), 406–417.

Dematte J.A.M., Pereira H.S., Nanni M. R., Cooper M. and Fiorio P.R. (2003) Soil chemical alterations promoted by

fertilizer application assessed by spectral reflectance. *Soil Science*, **168**(10), 730–747.

Dwivedi R.S. and Rao B.R.M. (1992) The selection of the best possible Landsat-TM band combinations for delineating salt-affected soils. *International Journal of Remote Sensing*, **13**, 2051–2058.

Eghbalm M.K., Southard J. and Whittig L.D. (1989) Dynamics of evaporite distribution in soils on a fan-playa transect in the Carizo Plain California. *Soil Science Society of America Journal*, **53**, 898–903.

Escadafal R., Girard M.C. and Courault D. (1989) Munsell soil color and soil reflectance in the visible spectral bands of Landsat MSS and TM data. *Remote Sensing of Environment*, **27**(1), 37–46.

Fensholt R. and Sandholt I. (2003) Derivation of a shortwave infrared water stress index from MODIS near- and shortwave infrared data in a semiarid environment. *Remote Sensing of Environment*, **87**(1), 111–121.

Galvao L.S. and Vitorello I. (1998) Variability of laboratory measured soil lines of soils from southeastern Brazil. *Remote Sensing of Environment*, **63**(2), 166–181.

Gao B.C. (1996) NDWI- a normalized difference water index fro remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, **58**, 257–266.

Gausman H.W. and Allen W.A. (1973) Optical parameters of leaves of 30 plant species. *Plant Physiology*, **52**, 57–62.

Goldshleger N., Ben-Dor E., Benyamini Y. and Agassi M. (2004) Soil reflectance as a tool for assessing physical crust arrangement of four typical soils in Israel. *Soil Science*, **169**(10), 677–687.

Goward S.N., Xue Y., and Czajkowski K.P. (2002) Evaluating land surface moisture conditions from remotely sensed temperature/vegetation index measurements: an exploration with the simplified simple biosphere model. *Remote Sensing of Environment*, **79**, 225–242.

Grove C.I., Hook S.J. and Paylor E.D. II (2004) *Laboratory Reflectance Spectra of 160 Minerals, 0.4 to 2.5 Micrometers*, Jet Propulsion Laboratory Publication.

Hardisky M.A., Klemas V. and Smart R.M. (1983) The influences of soil salinity, growth form, and leaf moisture on the spectral reflectance of spartina alterniflora canopies. *Photogrammetric Engineering and Remote Sensing*, **49**, 77–83.

Henderson T.L., Baumgardner M.F., Franzmeier D.P., Stott D.E. and Coster D.C. (1992) High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal*, **56**, 865–872.

Huete A.R. and Escadafal R. (1991) Assessment of biophysical soil properties through spectral decomposition techniques. *Remote Sensing of Environment*, **35**(2–3), 149–159.

Huete A.R., Post D.F. and Jackson R.D. (1984) Soil spectral effects on 4-space vegetation discrimination. *Remote Sensing of Environment*, **15**(2), 155–165.

Huete A.R. and Warrick A.W. (1990) Assessment of vegetation and soil water regimes in partial canopies with optical remotely sensed data. *Remote Sensing of Environment*, **32**(2–3), 155–167.

Hunt J., Raymond E. and Rock B.N. (1989) Detection of changes in leaf water content using near- and middle-infrared reflectances. *Remote Sensing of Environment*, **30**(1), 43–54.

Idso S.B., Jackson R.D., Reginato R.J., Kimball B.A. and Nakajima F. (1975) The dependence of bare soil albedo on soil water content. *Journal of Applied Meteorology*, **14**, 109–113.

Idso S.B., Reginato R.J., Jackson R.D., Kimball B.A. and Nakayama F.S. (1974) The three stages of drying of a field soil. *Soil Science Society of America Proceedings*, **38**(5), 831–837.

Irons J.R., Campbell G., Norman J.M., Graham D.W. and Kovalick W.M. (1992) Prediction and measurement of soil bidirectional reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 249–260.

Jackson T.J., Chen D., Cosh M., Li F., Anderson M., Walthall C., Doriaswamy P. and Hunt E.R. (2004) Vegetation water content mapping using landsat data derived normalized difference water index for corn and soybeans. *Remote Sensing of Environment*, **92**(4), 475–482.

Jacquemoud S., Baret F. and Hanocq J.F. (1992) Modeling spectral and bidirectional soil reflectance. *Remote Sensing of Environment*, **41**, 123–132.

Jenny H. (1980) *The Soil Resource, Origin and Behavior*, Springer-Verlag: New York.

Karnieli A., Kidron G.J., Glaesser C. and Ben-Dor E. (1999) Spectral characteristics of cyanobacteria soil crust in semiarid environments. *Remote Sensing of Environment*, **69**(1), 67–75.

Kauth R.J. and Thomas G.S. (1976) The tasseled cap. A graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. *2nd International Symposium on Machine Processing of Remotely Sensed Data*, Purdue University, West Lafayette.

Kimes D.S., Irons J.R., Levine E.R. and Horning N.A. (1993) Learning class descriptions from a data base of spectral reflectance of soil samples. *Remote Sensing of Environment*, **43**(2), 161–169.

Kooistra L., Salas E.A.L., Clevers J.G.P.W., Wehrens R., Leuven R.S.E.W., Nienhuis P.H. and Buydens L.M.C. (2004) Exploring field vegetation reflectance as an indicator of soil contamination in river floodplains. *Environmental Pollution*, **127**(2), 281–290.

Kustas W.P., Zhan X. and Schmugge T.J. (1998) Combining optical and microwave remote sensing for mapping energy fluxes in a semiarid watershed. *Remote Sensing of Environment*, **64**(2), 116–131.

Latz K., Weismiller R.A., Van Scoyoc G.E. and Baumgardner M.F. (1984) Characteristic variations in spectral reflectance of selected eroded soils. *Soil Science Society of America Journal*, **48**, 1130–1134.

Leger R.G., Millette J.F. and Chomchan S. (1979) The effects of organic matter, iron oxides and moisture on the color of two agricultural soils in Quebec. *Canadian Journal of Soil Science*, **59**, 191–202.

Levitt D.G., Simpson J.R. and Huete A.R. (1990) Estimates of surface soil water content using linear combinations of spectral wavebands. *Journal of Theoretical And Applied Climatology*, **42**, 245–252.

Liang S. and Strahler A.H. (1994) Retrieval of surface BRDF from multiangle remotely sensed data. *Remote Sensing of Environment*, **50**(1), 18–30.

Lobell D.B. and Asner G.P. (2002) Moisture effects on soil reflectance. *Soil Science Society of America Journal*, **66**, 722–727.

Madeira Netto J.S. (1996) Spectral reflectance properties of soils. *Photo Interpretation*, **34**, 59–70.

Metternicht G.I. and Zinck J.A. (2003) Remote sensing of soil salinity: potentials and constraints. *Remote Sensing of Environment*, **85**(1), 1–20.

Moran M.S., Hymer D.C., Qi J. and Kerr Y. (2002) Comparison of ERS-2 SAR and Landsat TM imagery for monitoring agricultural crop and soil conditions. *Remote Sensing of Environment*, **79**(2–3), 243–252.

Moran M.S., Inoue Y. and Barnes E.M. (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment*, **61**(3), 319–346.

Mougenout B., Pouget M. and Epema G. (1993) Remote sensing of salt-affected soils. *Remote Sensing Reviews*, **7**, 241–259.

Muller E. and Décamps H. (2001) Modeling soil moisture-reflectance. *Remote Sensing of Environment*, **76**, 173–180.

Nicholson S.E. and Farrar T.J. (1994) The influence of soil type on the relationships between NDVI, rainfall, and soil moisture in semiarid Botswana. I. NDVI response to rainfall. *Remote Sensing of Environment*, **50**, 107–120.

Obukhov A.I. and Orlov D.C. (1964) Spectral reflectance of the major soil groups and the possibility of using diffuse reflection in soil investigations. *Soviet Soil Science*, **1**, 174–184.

Palacios-Orueta A., Pinzon J.E., Ustin S.L. and Roberts D.A. (1999) Remote sensing of soils in the Santa Monica Mountains: II. hierarchical foreground and background analysis. *Remote Sensing of Environment*, **68**(2), 138–151.

Palacios-Orueta A. and Ustin S.L. (1998) Remote sensing of soil properties in the Santa Monica Mountains I. spectral analysis. *Remote Sensing of Environment*, **65**(2), 170–183.

Peñuelas J., Piñol J., Ogaya R. and Filella I. (1997) Estimation of plant water concentration by the reflectance Water Index WI (R900/R970). *International Journal of Remote Sensing*, **18**(13), 2869–2875.

Pinzón J.E., Ustin S.L., Castañeda C.M. and Smith M.O. (1998) Investigation of leaf biochemistry by hierarchical foreground/background analysis. *IEEE Transactions on Geoscience and Remote Sensing*, **36**(6), 1913–1927.

Raina P., Joshi D.C. and Kolarkar A.S. (1993) Mapping of soil degradation by using remote sensing on alluvial plain, Rajasthan, India. *Arid Soil Research and Rehabilitation*, **7**, 145–161.

Richardson A.J. and Wiegand C.L. (1977) Distinguishing vegetation from soil background information. *Photogrammetric Engineering and Remote Sensing*, **43**, 1541–1552.

Roberts D., Smith G.M. and Adams M.L. (1993) Green vegetation, non photosynthetic vegetation and soils in AVIRIS data. *Remote Sensing of Environment*, **44**, 255–269.

Sellers P.J., Meeson B.W., Hall F.G., Asrar G., Murphy R. E., Schiffer R.A., Bretherton F.P., Dickinson R.E., Ellingson R.G. and Field C. (1995) Remote sensing of the land surface for studies of global change: models – algorithms – experiments. *Remote Sensing of Environment*, **51**(1), 3–26.

Shepherd K.D. and Walsh M.G. (2002) Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society Of America Journal*, **66**(3), 988–998.

Shi Z., Huang M.X. and Li Y. (2003) Physico-chemical properties and laboratory hyperspectral reflectance of coastal saline soil in Shangyu city of Zhejiang province, China. *Pedosphere*, **13**(3), 193–198.

Smith M.O., Ustin S.L., Adams J.B. and Gillespie A.R. (1990) Vegetation in deserts:I. A regional measure of abundance from multispectral images. *Remote Sensing of Environment*, **31**, 1–26.

Stoner E.R. and Baumgardner M.F. (1981) Characteristic variations in reflectance of surface soils. *Soil Science Society of American Journal*, **45**, 1161–1165.

Toth T., Csillag F., Biehl L.L. and Micheli E. (1991) Characterization of semivegetated salt-affected soils by means of field remote sensing. *Remote Sensing of Environment*, **37**(3), 167–180.

Tucker C.J. (1980) Remote sensing of leaf water content in the near infrared. *Remote Sensing of Environment*, **10**(1), 23–32.

Tucker C.J. and Sellers P.J. (1986) Satellite remote sensing of primary production. *International Journal of Remote Sensing*, **7**(11), 1395–1416.

Ustin S.L. (1999) *Remote Sensing for the Earth Sciences: Manual of Remote Sensing*, Rencz A.N. (Ed.), New York, Wiley & Sons.

Vitorello I. and Galvao L.S. (1996) Spectral properties of geologic materials in the 400 to 2500 nm range: review for applications to mineral exploration and lithologic mapping. *Photo Interpretation*, **34**, 77–99.

Walthall C.L., Norman J.M., Welles J.M., Campbell G. and Blad B.L. (1985) Simple equation to approximate the bi-directional reflectance from vegetative canopies and bare soil surfaces. *Applied Optics*, **24**, 383–387.

Wessman C.A. (1991) Remote sensing of soil processes. *Agriculture, Ecosystems & Environment*, **34**(1–4), 479–493.

Wiegand C., Anderson G., Lingle S. and Escobar D. (1994) Soil salinity effects on crop growth and yield. Illustration of an analysis and mapping methodology for sugarcane. *Journal of Plant Physiology*, **148**, 418–424.

Wood E.F. (1997) Effects of soil moisture aggregation on surface evaporative fluxes. *Journal of Hydrology*, **190**(3–4), 397–412.

Xiao X., Hollinger D., Aber J., Goltz M., Davidosn E.A., Zhang Q. and Moore B. III (2004) Satellite-based modeling of gross primary production in an evergreen needleleaf forest. *Remote Sensing of Environment*, **89**, 519–534.

Zarco-Tejeda P.J., Rueda C.A. and Ustin S. (2003) Water content estimation in vegetation with MODIS reflectance data and model inversion methods. *Remote Sensing of Environment*, **85**, 109–124.

Zhang M., Ustin S., Rejmankova E. and Sanderson E. (1997) Monitoring pacific coast salt marshes using remote sensing. *Ecological Applications*, **7**, 1039–1053.

60: Estimation of River and Water-Body Stage, Width and Gradients Using Radar Altimetry, Interferometric SAR and Laser Altimetry

CHARON BIRKETT[1], DOUG ALSDORF[2] AND DAVID HARDING[3]

[1] *Earth Science Interdisciplinary Center, University of Maryland, College Park, MD, US*
[2] *Department of Geological Sciences, Ohio State University, Columbus, OH, US*
[3] *Geodynamics Branch, NASA/Goddard Space Flight Center, Greenbelt, MD, US*

Major advances in the application of satellite radar altimetry have led the way to a proven ability to determine surface water height, or stage, for some of the largest rivers, wetlands, and lakes around the world. These developments have come at a time when global water resources are of major concern, and there is a much-noted decline in ground-based monitoring networks. Recent interferometric synthetic aperture radar studies have also demonstrated the ability to retrieve stage changes over a wide expanse of inundated floodplain or wetland. While the performance of these radar instruments is continuously being assessed, focus is also on the potential of satellite-based laser altimetry, which may offer smaller observable targets and improved stage accuracies. With these technological capabilities, we have the capacity to determine inland water stage, and to some degree, river width and gradient.

INTRODUCTION

Today, there are many excellent *in situ* hydrologic monitoring networks in operation, measuring lake and river stage (i.e. the height of the water surface), often on an hourly or twice-daily basis, to centimeter accuracy. Some gauges also have the capability to transmit their findings in near-real time. However, gauge sites can be lacking in remote regions and there is often a total absence of recording within wetlands, braided rivers, and floodplains. In many populated regions, the capacity to retrieve stage measurements is also being affected by economic and political factors. These factors can result in explicit data restriction policies as well as delays in information availability. While there is continuing difficulty in obtaining near-real time data from the developing world, recent reports have additionally highlighted the continuing net loss of recording stations for some regions traditionally endowed with dense networks (Lanfear and Hirsch, 1999; Vörösmarty *et al.*, 2001).

The issue of decreasing ground-based observations is a prime concern considering both the human dimension and earth system issues relating to freshwater. For the former, the assessment and management of water resources is a top priority in relation to population growth. The monitoring, prediction, and mitigation of natural hazards such as droughts and floods is also of concern as are transportation and waste processing issues. Regarding studies of the earth system, determination of the surface water balance is critical in understanding both land/ocean and land/atmosphere interactions. The role of rivers, wetlands, and lakes regarding (i) the transport of nutrients, (ii) storage devices and regulators of biogeochemical cycles, (iii) trace gas exchange mechanisms, and (iv) changes on the timing of continental runoff, are all of prime importance.

The effects of climate change contribute additional uncertainty to all of these issues, making the determination of surface water parameters such as river discharge and volume storage even more critical. River discharge is traditionally inferred by the combined knowledge of river-stage and a stage-discharge relationship (referred to as "rating curves", *see* **Chapter 61, Estimation of River Discharge, Volume 2** for details) for a specific location. Lake

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

volume is determined via synergistic knowledge of stage and bathymetry. Ideally, discharge and volume should be acquired via a long-term, global monitoring program with free and rapid dissemination of all information. In this respect, satellite remote sensing is an obvious solution, enhancing the ground-based network information and providing additional data at ungauged locations. While current satellite sensors cannot measure discharge or storage directly, the ability to measure stage is central to both parameters and can be measured, with some limitations, by current technologies. In addition, there is also the potential for river width and gradient to be measured by satellite.

The ability to measure stage using satellite radar altimetry and interferometry has been briefly discussed before in several review papers (Kite and Pietroniro, 1996; Smith, 1997; Mertes, 2002; Mertes *et al.*, 2004). Discussion of river discharge and wetland and lake extent can also be found in **Chapter 61, Estimation of River Discharge, Volume 2**.

SATELLITE RADAR ALTIMETRY

Historical Perspective

Satellite radar altimetry emerged in the early 1960s with objectives of acquiring accurate sea surface topography. By timing the two-way emission and reflection of microwave pulses see section "Principles of Satellite Radar Altimetry", the spatial and temporal variability of sea surface height could be monitored via the utilization of an exact repeat orbit. The technique led to the National Aeronautics and Space Administration's (NASA) GEOS-3 (1975–1978), and the National Oceanographic and Atmospheric Administration's (NOAA) Seasat satellite in 1978. The NOAA Geosat mission followed a decade later and operated successfully between 1986 and 1989.

With multiple ocean, land, and ice science objectives, the European Space Agency (ESA) launched ERS-1 in 1991, and this was followed by ERS-2 (1995) and ENVISAT (2002). An oceanographic theme was continued by NASA and the Centre National d'Etudes Spatiales (CNES) with the launch of TOPEX/POSEIDON (T/P) in 1992 and its follow-on mission, Jason-1, in 2002. The Geosat follow-on program (GFO) is also the US Navy's initiative to develop an operational series of radar altimeter satellites to maintain continuous ocean observation. The first satellite was launched in February 1998 and became operational in November 2000.

Satellite radar altimetry has thus been a prominent technique spanning several decades of operation and application. Although successful, it is important to note that to date no system has been designed and no mission dedicated solely to the monitoring of inland water stage variations.

Principles of Satellite Radar Altimetry

Pulse-limited satellite radar altimetry has been a validated technique for many years now, and a number of excellent published sources describe the fundamentals in great detail (e.g. Ulaby *et al.*, 1981; Chelton *et al.*, 1988, 2001). The basic principles of this nonimaging technique are briefly described here.

Operating from an altitude of 700–1330 km, a continuous stream of microwave pulses (e.g. frequency Ku band 13.6 GHz) is emitted from a nadir-pointing antenna (Figure 1). The resulting radar echo has an intensity (radar backscatter) and shape defined by the surface characteristics (Figure 2). The distance between the antenna and the surface (the altimetric range) is determined via knowledge of the two-way timing difference between pulse emission and echo reception. The spacecraft altitude is estimated from a combination of other technologies (Section "Height (Stage) Construction"). The surface height then, is defined as the difference between orbital altitude and range and is given with respect to a reference datum, a mathematical concept known as a *reference ellipsoid*. If a transformation to a reference system based on mean sea level is required, knowledge of the local geoid must be additionally known.

When referring to inland water bodies, it is this height above the reference ellipsoid, or geoid, that is associated with the traditional ground-based stage measurement of a lake or river gauge. With continuous operation, such stage measurements can be potentially retrieved for any part of an inland water target that sits directly below the satellite overpass. Typically, the instrument footprint size (a function of surface roughness with pulse-limited

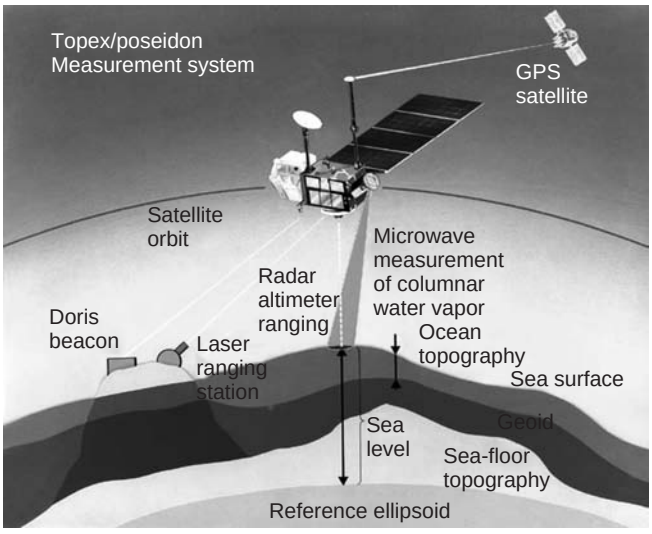


Figure 1 The basic principle of satellite radar altimetry (Image courtesy of the NASA Jet Propulsion Laboratory, California Institute of Technology). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

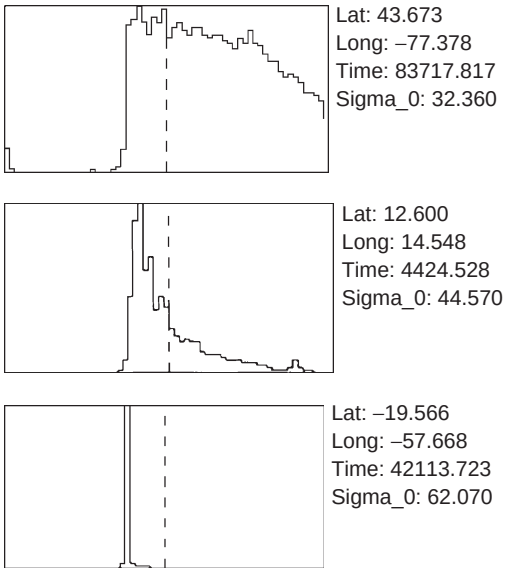


Figure 2 From top to bottom, radar echoes from Lake Ontario, Lake Chad, and the Amazon River during a period of high river stage, demonstrating the effect of differing surface conditions on echo shape and power. Each echo, or waveform, is a representation of power returned as a function of time. Over lakes, waveforms are typically ocean-like, but can become broad-peaked or even narrow-peaked under very calm or icy conditions. Over rivers, floodplains, and wetlands, waveforms are generally narrow-peaked with significant radar backscatter coefficient variability during seasonal inundation

technology) is a few hundred meters to several kilometers wide. Generally, the resulting ground tracks (Figure 3) are repeated to within ± 1 km through each orbit cycle. In reality, there are many limitations to height retrieval but where data is valid, time series of height variations can be constructed during the mission lifetime. If a satellite crosses a river at several locations, then the gradient of a river reach can be estimated if the separation distance between locations is known. Owing to various technological and data-rate limitations (Section "Advantages and Limitations: Accuracy, Resolution, and Target Size"), determination of river width may be more problematic with radar altimetry.

There are complexities to the basic principle outlined here, and instrument design, operation, and data processing may differ between each mission. Over land and inland water, consideration must always be given to the ability to successfully follow (or "track") the rapidly changing fluctuations in elevation and surface complexity. The altimetric range itself is determined from inspection of the echo within a "range window". By continuous inspection of echo power and location, it is the role of the tracking system logic to capture and keep the leading edge, or center of gravity, of each echo within the range window. If the leading edge can be observed, the altimeter is said to be tracking the surface well with surface "lock" being maintained, and the altimetric range can then be derived. Rapid variations in surface roughness and complexity often effect either the ability to retrieve any range value, or the accuracy of the determined range.

The deduced altimetric height must also have a series of corrections applied to take into account various instrument and geophysical factors (Section "Height (Stage) Construction"). It is this postprocessing that allows the reconstruction of the true surface elevation to a sufficient degree of accuracy. In general, ground-processing teams collect and process the telemetered data from the altimeter and auxiliary instruments and redistribute it in raw (Sensor Data Records,SDRs) or processed (Geophysical Data Records, GDRs) form to the user communities. Fast delivery data, with poorer satellite altitude accuracy, may be available within a few hours after satellite overpass. GDR data are often distributed within a month after overpass and contain more precise measurements and corrections. The final stage measurement is associated with a particular time in the mission and is given for a specific geographical location (subsatellite point). It should not, however, be considered as a spot height, rather it is formed via an averaging process that considers all surface heights within the footprint, and the surface heights along a portion of the ground track (to reduce signal to noise). In addition, final stage values will

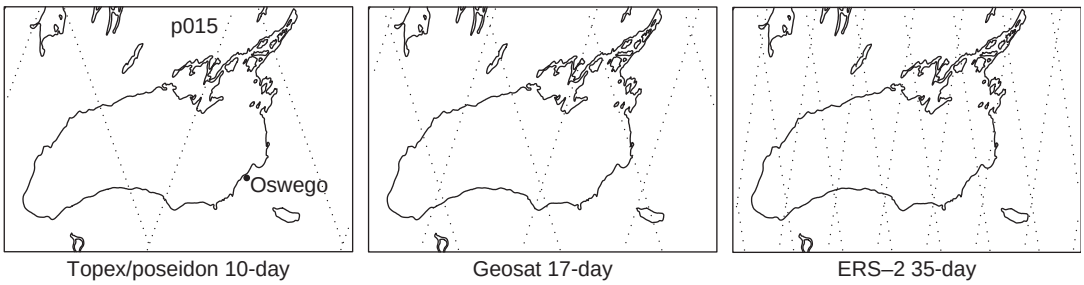


Figure 3 An example of the trade-off between spatial and temporal resolution for differing satellite missions. Dots represent the satellite reference ground track at 1-s intervals. With continuous instrument operation, surface height can be potentially retrieved at any location along these ground tracks. The repeatability of the ground tracks is nominally kept to within ± 1 km

be given at a rate dependant on the delivered mission data. For example, from the T/P GDRs, stage can be constructed every ~580 m along the ground track (10 Hz), while every ~330 m (20 Hz) from the high-resolution ERS datasets.

Height (Stage) Construction

A full discussion of the derivation of altimetric height can be found elsewhere (e.g. Zandbergen, 1990, and various mission data handbooks). A brief description is given here, where relevant to studies of inland water. The height, H, of a non-ocean surface is given by the difference between the satellite orbit altitude, Alt, and the corrected altimeter range, R_{corr}, with a final correction for tidal effects, T.

$$H = (Alt - R_{corr}) - T - [BC] - [G] \quad (1)$$

Both H and Alt are given with respect to a mathematical reference ellipsoid. Tidal effects include corrections for earth, pole, and ocean-loading tides. In general, GDRs and SDRs do not include information pertaining to lake tides or to ocean tidal effects on large rivers. Researchers must seek auxiliary data sets here. Height measurements over the ocean usually include a correction for barometric effects, BC, to take into consideration the response of the sea surface to atmospheric pressure. Although this correction is supplied on some GDRs for some of the largest lakes and inland seas, the correction is not supplied for the smaller lakes. In consideration of the fact that the majority of lakes are small closed systems by comparison with atmospheric pressure systems, BC is often not applied. If heights are to be determined with respect to a mean sea surface, than a correction for the geoid, G, can be applied if such knowledge is available from the GDRs or external sources. The spacecraft altitude, Alt, can be very accurately obtained by a combination of technologies including the Global Positioning System (GPS), Satellite Laser Ranging (SLR), Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS), Precise Range and Range-rate Equipment (PRARE) and by use of the altimetric heights at orbital "cross-over" locations.

The altimetric range itself must be corrected for various instrumental and physical effects. These include an electromagnetic bias (EM-bias), which takes into consideration the fact that surface wave troughs are better reflectors than wave crests. Other corrections include a center of gravity correction to correct for satellite roll and pitch changes due to solar array motion and a significant wave height-attitude correction (SWH/ATT) for combined surface wave status and spacecraft off-nadir pointing effects. Over inland waters, EM-bias and SWH/ATT must certainly be applied to large unfrozen lakes, but the values become more questionable, and may even be defaulted in the GDRs, over the calmer surfaces of rivers and wetlands.

Range must also be corrected for atmospheric effects. The propagation speed of a radar pulse is reduced because of the effects of various gases and water vapor in the Earth's troposphere. The former is fairly predictable, but the latter is highly variable. An onboard microwave radiometer can provide more accurate estimates of the water vapor over large lakes, but for smaller targets and rivers and wetlands, this correction must be sought from a less accurate meteorological model. The radio pulse propagation velocity is also slowed by the free electrons within the Earth's ionosphere and is proportional to the operating frequency of the altimeter instrument. Over inland water, ionospheric corrections from models or derived from the DORIS instrument are utilized.

The precision of the range estimate is dominated by the error associated with estimating the location in the range window of a predefined point on the leading edge of the returned echo (the retracking point). For an ideal ocean-like echo (Brown, 1977), this point corresponds to the mean height in the instrument footprint. Sometimes the altimeter response to large lakes, and nearly always to rivers and wetlands, is not similar to that over the ocean but varies according to surface conditions. For non-ocean-like radar echoes, there may also be a need for postprocessing of the radar echoes, a technique known as *retracking*. In this case, a range correction has to be estimated to take into consideration variable height biases caused by rapid changes in either surface height, or echo shape and intensity.

Technique and Validation

After the construction of the individual 10- or 20-Hz stage measurements, various filters need to be applied to aid the removal of poor or erroneous stage values stemming from geophysical or instrument effects. In particular, the effects of coastlines and islands need to be eliminated (e.g. Birkett, 1995, 1998).

With valid stage data, the user can then explore the potential for river-gradient determination or look to the creation of a time series of stage variations. The latter requires an intercomparison of stage and location (latitude/longitude) data from all repeat tracks (repeated to ± 1 km), either by a nearest-neighbor approach or by interpolation and coalignment of the tracks. Such a "repeat-track technique" is employed to enable time series of relative stage variations to be deduced over a given time period, and for a given section of satellite ground track. In this method, either one repeat pass over the target or a mean reference pass is nominated as a "reference profile" and height differences between this reference pass and successive repeat passes over the target are determined. A time series of elevation variations is then constructed (Birkett, 1995). With some knowledge of the precision of the technique (Section 2.5.1), estimations of resulting stage accuracy can be made and are

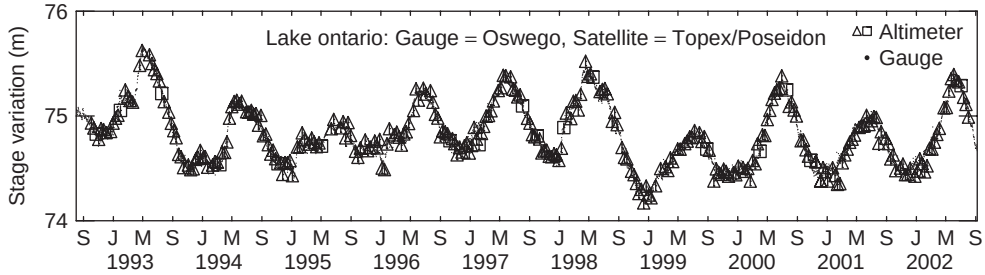


Figure 4 Validation of stage variations over Lake Ontario. Ground-based stage data (dots) are from the Oswego gauge (courtesy of the National Oceanic and Atmospheric Administration/National Ocean Survey). TOPEX/POSEIDON altimeter data from pass015 (Figure 3) are denoted by triangles (NASA radar altimeter) and squares (solid state altimeter). A constant height offset has been applied to the altimetric result. Time steps are in intervals of 4 months (Reproduced from Birkett, 1995 by permission of American Geophysical Union)

further validated by comparisons with ground-based hourly and daily stage data (Figure 4).

Advantages and Limitations: Accuracy, Resolution, and Target Size

The main advantages of this technique include the ability for day/night and all-weather systematic operation with no loss of operation due to cloud coverage. With continuous operation across land surfaces, the instrument behaves like a string of pseudogauges, sampling across lake, inland sea, and wetland surface, and across wide river channels and their floodplains. The presence of vegetation or canopy cover is not a hindrance to this nadir-viewing instrument except at very low inundation levels when dryland conditions are approached. In general, the instruments can observe monthly, seasonal, and interannual variations (Figure 5).

Unlike many gauge networks that operate using a local reference frame, all altimetric height measurements are given with respect to one reference datum, to form a globally consistent, uniform data set that is based on a proven, validated technique. For the most part, the altimeter datasets themselves (DVD, CD-Rom, exabyte, or via ftp) are freely available. There are also several ongoing programs whose objectives are to construct near-real time stage variations for inland waters and place them out in the public domain (Section 2.6). Altimeter instrument design and performance though is not optimized for studies of lakes and rivers. Although there are a number of advantages, the technique is not without its limitations. Two obvious points to note are that the altimetric measurements are constrained by the lifetime of the satellite, and the instruments are nadir-pointing and so do not have a truly global perspective.

Accuracy

The final accuracy of a single altimetric stage value depends on the accuracies of all the parameters and correction terms in equation (1). The satellite altitude and other corrections have root mean square (rms) errors on the order of a few centimeters. The range precision may be poorer, but one way to improve this term is by averaging across the target. In this way, the entire river width, or the full lake crossing extent can be used. The construction of surface height, however, does not take into account the effects of strong or sustained winds, heavy precipitation events, or the presence of lake/river ice. The latter two can hamper the quality and quantity of the range measurement and are sometimes difficult to detect over inland water. Wind setup effects can also effect the mean stage measurement across the lake.

A traditional ground-based gauge offers a height measurement at a defined geographical location, that is, it is essentially a spot height, while an altimeter must rely on the ability to average many echoes along the satellite ground track. The accuracy of a resulting altimetric time series (in comparison with gauge measurements) can vary then according to target size (or crossing extent) and surface roughness. For the largest lakes of the world, validation exercises have shown 3–5 cm rms can be achieved. This degrades though to ten or tens of centimeters rms for smaller lakes (e.g. $\sim$100 km^2) or narrow reservoirs, and holds true for shallow lakes, wetlands, and rivers. This may be considered poor but is clearly acceptable for some science objectives, and is a bonus for those regions around the globe with a total absence of any stage data. With knowledge of the satellite altitude at the 1–2 cm level, the most dominant source of altimetric errors for small lakes and rivers/wetlands is the poor knowledge of the wet tropospheric correction, and the error associated with the extraction of the altimetric range from the narrow-peaked river/wetland echoes.

Resolution and Performance

There is a trade-off between temporal and spatial resolution with each mission (Figure 3), and instrument performance determines the final quantity of retrieved elevation information. The majority of targets will have a single satellite

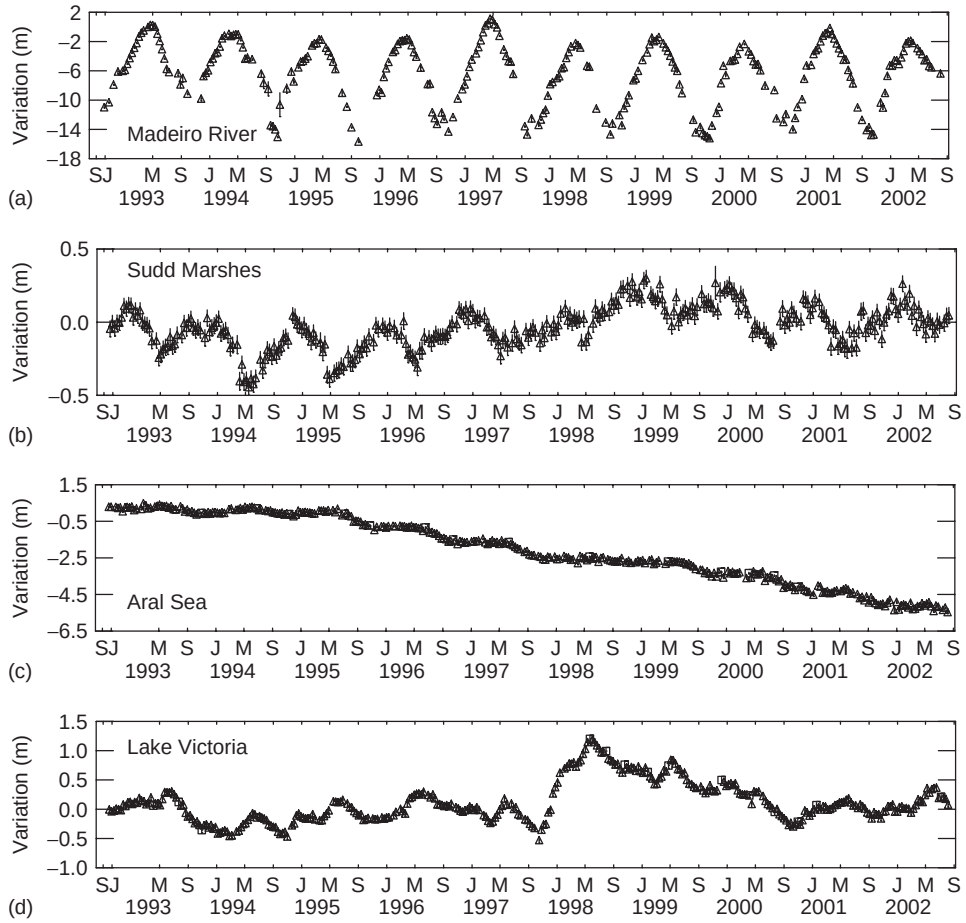


Figure 5 Time series of relative stage variations for a variety of inland water targets; (a) the Madeira River, Brazil, (b) the Sudd Marshes, Sudan, (c) the Aral Sea, and (d) Lake Victoria, Africa. The series was produced using data from the TOPEX/POSEIDON mission archive. Time steps are in intervals of 4 months (Reproduced from Birkett, 1995; Birkett *et al.* 1999, 2002 by permission of American Geophysical Union)

overpass, and thus the temporal resolution or the frequency of the stage measurements will be defined by the satellite repeat period (e.g. one overpass every 10 or 35 days). This is to be compared with the twice-daily, often automated measurement from a gauge. For the largest lakes and inland seas, the satellite temporal resolution will improve depending on the number of differing overpasses and the ability to combine separate time series results.

Spatial resolution is also limited by the narrow field of view. The number of available targets and the location of overpass are predetermined by the satellite orbit that is, its geographical latitude limits and the positioning of its resulting ground tracks. It is thus a case of predetermined target selection depending on the distribution of the inland water basins. Examples can be given for the currently operating ENVISAT and Jason-1 missions. The Mullard Space Science Laboratory Global Lakes Database (Birkett and Mason, 1995) reveals that there are ~1440 large lakes and reservoirs across the globe. With its 35-day repeat period, and ±82° latitude limit, ENVISAT crosses over ~1075 of those listed. With a temporal resolution increased by a factor of three, the number of targets crossed by Jason-1 (10-day repeat, ±66° latitude) is reduced by a factor of three, to only ~360.

Consideration must be given to the performance of the instrument, regarding its ability to track highly varying terrain, and the additional limitations of effective footprint size and GDR data-rate on target size. Surrounding hills, ravines, buildings, riverbanks, lake coastlines, and islands, all offer potential obstacles for the tracking system. The results may be a radar echo too complex to interpret, or the complete loss of surface lock, both effects resulting in the loss of surface height. The effects can be compounded by the presence of multiple water targets within the footprint. The reacquisition of the terrain after loss of surface lock may also take a few seconds, by which time the satellite may have flown past the intended inland water target. The Seasat, Geosat, T/P and Jason-1 radar

altimeters were only designed to track the ocean surface with an assumed sea surface variability and expected wave-height range. The ERS-1 and ERS-2 radar altimeters were designed for both ocean and ice/land measurements and incorporated the ability to switch between two operating modes, ocean- and ice-tracking, to assist with the capture of the radar echoes within the range window (Scott *et al.*, 1994). The ENVISAT radar altimeter, RA-2 has a different tracking system again, with the potential to retain even more land and inland water data.

Target Size

Surface tracking considerations, the effective instrument footprint size, and the averaging of echoes to improve range precision, all place constraints on the minimum size of observable target. The tracking consideration can even be a concern for large targets that have mountainous terrain on the satellite approach path. The footprint size affects the ability to single out a specific river channel or small pool, and determines the ability to separate river channel from inundated floodplain. A final factor is the rate at which stage values are given in the GDR/SDR user datasets. One average height value is given every ~580 m (for T/P, GDR), ~350 m (ERS, ENVISAT), and ~290 m (Jason-1, GDR) along the ground track. Current studies on target size limitation suggest $> 100-300\,\text{km}^2$ (wetlands, lakes) and $>0.5-1$ km river widths, mission dependent. The advent of new tracking techniques and echo interpretation methods, and the potential for increasing the data-rate, may well significantly improve these minimum size limitations.

Applications

The first inland water study was carried out by Brooks (1982) who explored Seasat data over the Canadian lakes for mapping purposes. Rapley *et al.* (1987), Olliver (1987), and Au *et al.* (1989) also used this dataset to perform bathymetric studies on Lake Tanganyika, the Great Lakes, and the Caspian and Black Seas. Guzkowska *et al.* (1990) and Cudlip *et al.* (1992) also utilized these early measurements in several wetland-mapping and river-gradient determination studies. In general, results were hampered by data limitations, technical problems, and the then poor accuracy of the Seasat satellite altitude.

Following the success of the Geosat mission, Guzkowska *et al.* (1990) also presented the first lake level time series for Grand Lac, Cambodia, while Koblinsky *et al.* (1993) explored river-stage accuracies within the Amazon basin. As part of a study dedicated to aridity changes, Birkett (1994) expanded the early Geosat studies to a collection of large ($\geq 100\,\text{km}^2$), closed, and climatically sensitive open lakes. Morris and Gill (1994a) also derived Geosat lake level time series for the Great Lakes and through validation exercise showed the resulting time series were accurate to ~10 cm rms.

The advent of a new satellite mission, TOPEX/POSEIDON (T/P) with excellent knowledge of satellite altitude and a 10-day repeat period, enabled altimetric lake and river stage to be more accurately determined. Morris and Gill (1994b) found rms accuracies of ~3 cm over the Great Lakes when lake tides and wind setup effects were removed. Birkett (1995) examined the contribution of T/P to the continuing study of climatically sensitive lakes, and later (Birkett, 1998) to the study of rivers and wetlands. Dalton and Kite (1995) also examined preliminary T/P data over the Great Lakes, and made low-rate (1 Hz) accuracy comparisons with nearby gauge measurements.

After these early exploratory projects, the application of T/P altimetry to inland water studies became more region-based with science objectives focusing on natural hazards, climate change, and surface water dynamics. Cazenave *et al.* (1997) observed the temporal and spatial variation of the rising and falling waters of the Caspian Sea. Ponchaut and Cazenave (1998) utilized the interannual stage fluctuations of some of the largest lakes in Africa to explore El Niño/Southern Oscillation (ENSO) and recurrent drought connections. The late 1990s saw one of the strongest El Niño's on record in the Pacific and similar reversals of sea surface temperature in the Indian Ocean. Discussion of drought changed to precipitation excess as radar altimetry was used to study the great 1997/1998 flooding events across East Africa (Birkett *et al.*, 1999; Mercier *et al.*, 2002).

The plight of flood or drought and the associated socioeconomic impacts became the focus of several social, remote sensing, and modeling studies concentrating on the water resources of Lake Chad in the Sahel (Birkett, 2000; Sarch and Birkett, 2000). In general, the determination of reservoir storage for irrigation potential is particularly important for all regions with high agricultural demands. With the advent of fast delivery radar altimetry data and validated techniques, operational projects are now underway to assist with the timely delivery of information. One such program (**http://www.pecad.fas.usda.gov/cropexplorer/global_reservoir**) monitors in near-real time the stage variations of nearly a hundred large lakes and reservoirs across the globe using the T/P and Jason-1 datasets.

Temporal fluctuations in river and wetland stage became the focus of two studies in the Amazon basin (Campos *et al.*, 2001; Birkett *et al.*, 2002) and one in the La Plata Basin (Maheu *et al.*, 2003). All three projects examined T/P data at many sites noting the differences of seasonal amplitudes, the time lags of peak periods, and the propagation speeds of the flood waves. The Birkett *et al.* (2002) study (**http://essic8.umd.edu/RALNET-LBA**), part of the Large-scale Biosphere Atmosphere experiment in Amazonia (LBA) project, also utilized the altimetry to reveal the hysteresis characteristics of the Amazon main stem flood

Table 1 Summary of satellite technique capabilities for inland water

Satellite mission	Operation period	Temporal resolution	Spatial resolution	Stage-change accuracy	Minimum target area and width
Radar Altimetry					
ERS-2	1995–2002	35 days	350 m AT	$>$ 9 cm rms	$>$100 km^2 $>$500 m
ENVISAT	Post-2002	35 days	350 m AT	$>$9 cm rms	$>$100 km^2 $>$500 m
T/P	1992–2002	10 days	580 m AT	$>$3 cm rms	$>$300 km^2 $>$1 km
Jason-1	2002–current	10 days	290 m AT	$>$3 cm rms	$>$300 km^2 $>$500 m
InSAR					
SIR-C	October, 2004	1 day	20–90 m PS	$\pm$1.0 cm	0.1 km^2, $\sim$100 m
JERS-1	1992–1998	44 days	20–90 m PS	$\pm$2.4 cm	0.1 km^2, $\sim$100 m
Laser Altimetry					
ICESat	2003–current	91 days	175 m AT	$>$10 cm rms	0.01 km^2, $\sim$100 m

(A) AT = along track (nadir) stage interval, PS = pixel size.
(B) Minimum target area/width for lake or wetland area and river width. For radar and laser altimetry, values given are guidelines based on data-rate and footprint size. Current radar investigations are examining trade-offs between size reduction gains and stage accuracy effects.
(C) Radar and Lidar altimetry can measure stage and stage change over a given time period. The InSAR technique based on double-bounce returns measures only stage change and applies to flooded forest regions.
(D) Stage-change accuracies for radar and laser altimetry are based on comparisons with ground-based gauge data over a given time period. Values can be variable depending on target extent and surface roughness. Values for ERS-2 are based on preliminary validation results in ice tracking mode. Values for ICESat are based on preliminary calibration and validation results.
(E) ICESat temporal resolution: The orbit is a 91-day repeat orbit, but with $\leq 5^\circ$ off-nadir precision pointing capability, repeat profiling is 8 days at 70° lat, or 24 days at the equator.

wave, and to study the temporal and spatial variations of the river gradient.

To summarize, satellite radar altimetry is capable of monitoring stage and stage variations within large lakes, reservoirs, rivers, and wetlands. The current missions are not dedicated to inland water studies, nevertheless, the technique has been successfully utilized within many different interdisciplinary projects and operational programs. It can potentially offer stage measurements at many ungauged locations, but there are limitations to the technique in terms of stage accuracy, target size, temporal, and spatial resolution (see Table 1). Stage accuracy and spatial resolution may be improved for large wetland areas and floodplains by the application of interferometric SAR (Section 3), while lidar altimetry (Section 4) aims to improve both stage accuracy and size limitations for all target types.

INTERFEROMETRIC SYNTHETIC APERTURE RADAR

The ability to only observe stage variations at nadir is a serious limitation of satellite radar altimetry. However, if certain assistance is provided by vegetation, this limitation can be overcome by another satellite-based technique known as *interferometric synthetic aperture radar* (*InSAR*). This extends the capability to a wide image swath and can measure more subtle changes in water level. Across vast wetlands and floodplains this can lead to estimations of significant storage changes. The following sections detail the technological and geomorphic requirements of this methodology as well as its resolutions, advantages, and disadvantages.

Historical Perspective

Interferometric SAR first gained wide recognition with the efforts of Massonnet *et al.* (1993) and Zebker *et al.* (1994) who provided highly detailed maps of ground surface deformation resulting from the Landers earthquake in Southern California. Similarly, Goldstein *et al.* (1993) demonstrated that interferometric processing of SAR imagery could yield velocity maps of flowing glaciers. For the past decade, a tremendous effort has been devoted to this methodology (e.g. review provided by Massonnet and Feigl, 1998), and in fact, the Space Shuttle Radar Topography Mission (SRTM) of February 2000 used interferometric SAR to create the world's first seamless, high-resolution digital elevation model (DEM) (Farr and Kobrick, 2001).

Because SAR systems use an off-nadir viewing geometry, the highly reflective water surface typically causes emitted radar pulses to specularly reflect away from the antenna. Thus, conventional thought suggests that interferometrically measuring water level changes is not possible because of the lack of returned energy. Alsdorf *et al.* (2000, 2001a,b) have now demonstrated that, in the special case of inundated vegetation, radar pulses are returned to the antennae and that it is possible with interferometric processing of the SAR data to produce a centimeter-scale mapping of water level changes throughout inundated vegetation.

Interferometric SAR Principles

To create a map of surface change with interferometry requires two SAR images acquired from nearly identical viewing geometries. The images must be acquired before and after the change phenomenon, for example, an earthquake or floodplain-lake with decreasing storage, thus this approach is called the *repeat-pass* method. After coregistering the images to subpixel accuracy, the complex radar phase and amplitude values at each pixel are subtracted to yield an interferometric phase image with phase values varying between $-\pi$ and $+\pi$. The specific phase value of each pixel is a function of (i) the baseline distance between the two positions of the SAR antennae, (ii) topographic relief, (iii) the degree of correlation between the scattering elements that comprise each pixel (i.e. the "coherence", Li and Goldstein (1990), Zebker and Villasenor (1992)), and (iv) the surface elevation change.

InSAR Baselines

Because of viewing parallax, large baselines are ideal for accurately measuring topography, whereas smaller baselines are desired for mapping changes in surface elevations. To separate these components, either a synthetic interferogram based on a DEM ("two-pass method", Massonnet *et al.*, 1993) or additional SAR interferograms free of displacement phenomenon ("multipass method", Zebker *et al.*, 1994), can be subtracted from the observed phase. Essentially, the observed phase contains both topographic and surface change components. Removal of the topographic component is accomplished by either (i) using a preexisting DEM to construct a model phase based on the known observing geometry or (ii) using additional interferograms with long baselines thus dominated by a topographic signal. Because water surface elevations are constantly changing, the topographic phase cannot be easily extracted using additional SAR images, which suggests that the two-pass method is likely required. The amount of topographic relief captured by one interferometric phase cycle, from $-\pi$ to $+\pi$ is a function of the perpendicular component of the baseline. Short perpendicular components yield more topographic relief per phase cycle than long baselines, thus giving a more reliable estimate of surface change.

The two Amazon demonstrations of this interferometric method discussed in Section 3.5 had perpendicular baselines between satellite acquisitions of +63 m (Shuttle Imaging Radar C-mission,SIR-C, Alsdorf *et al.*, 2000, 2001a) and −118 m (Japanese Earth Resources Satellite, JERS-1, Alsdorf *et al.*, 2001b). These short baselines indicate that 0.5 radians of phase are equivalent to 15 to 20 m of topographic relief, whereas depending on look angle, the same 0.5 radians are also equivalent to about 1 cm of water surface change. Fortunately, in these two cases, the topographic relief across the Amazon floodplain is about 20 m, resulting in the bulk of the interferometric phase signal being related to water level change.

InSAR Coherence

Over land surfaces, where changes in soil moisture, vegetation, and freeze/thaw cycling cause random changes in the structure and dielectric properties of the scattering elements, interferometric coherence typically diminishes with increasing time between SAR acquisitions. Over inundated vegetation, the scattering elements consist of the water surface and vegetation trunks. For short time periods, say 24 h, the water surface is likely to experience a greater change in structure than the trunks of vegetation (i.e. wave action on the water surface). Thus, short-term temporal coherence is probably a function of the stability of the water surface (i.e. dielectric changes in the air-water interface are assumed to be minor). Across seasonal growth periods, however, vegetation changes will also affect coherence. For example, during the 24-h separating SIR-C acquisitions of Alsdorf *et al.* (2000, 2001a), coherence was maintained, and thus it was initially assumed that the heavy Amazon forest canopy prevented wind roughening of the water surface during the 1-day period. Yet given the low vegetation density and large water surface of Lake Balbina (Alsdorf *et al.*, 2001b), it is doubtful that wind and wave action would be similarly subdued for the 44-day temporal baseline of the JERS-1 interferogram. Nevertheless, strong coherence is maintained across the entire JERS-1 image.

Accuracy and Temporal Resolution

The vertical accuracy of the interferometric method is related to coherence and the degree to which the topographic phase can be removed. In cases where the scattering elements of each pixel remain stable, such as a dry desert, and an accurate DEM exists, the resolution of the surface change in the look-direction of the SAR antennae can be less than a centimeter (e.g. Massonnet *et al.*, 1993; Zebker *et al.*, 1994). However, the Amazonian research lacked a DEM, thus based on coherence measurements, vertical accuracies were estimated to vary from ±1.0 cm in the SIR-C case to ±2.4 cm for JERS-1 (see Table 1). Provided that coherence is maintained across the floodplain water body, the spatial accuracy is a standard function of SAR imagery. SAR data typically have pixel sizes between 20 and 90 m, such that many small geomorphic features can be mapped. The number of satellite overpasses varies from 1 day for SIR-C to 44 days for JERS-1. Unfortunately, both of these missions are no longer operating.

Limitations and Advantages

It is must be noted that the InSAR method requires the radar pulse to be returned from the water surface to the antennae, and in the case of the repeat-pass method

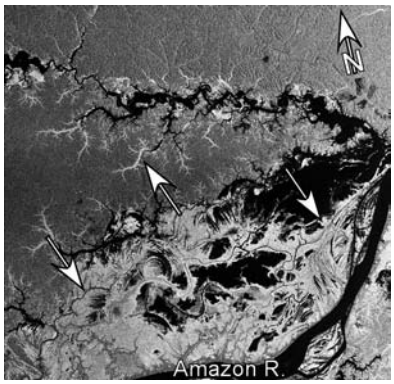
described here, the return does not occur over open water surfaces. Rather, the trunks of vegetation are a demonstrated requirement for measuring changes in water levels. Other features, such as telephone poles, might produce returns, but remain to be proven. Fortunately, the testing ground for this technique, the Amazon floodplain, is about 70% inundated forest (Mertes *et al.*, 1995), suggesting that the geomorphic requirements are met for much of this vast wetland. However, the radar pulse must also penetrate the vegetation canopy in order to reach the underlying water surface. Radar wavelengths of C-band or shorter (~6 cm) do not typically penetrate the dense canopy of the Amazon floodplain at typical off-nadir SAR imaging geometries, whereas L-band (~24 cm) SAR will travel through vegetation cover (e.g. Hess *et al.*, 1995). Research is still required to determine the complete relationship between vegetation habitat, radar wavelength, and look-angle. The strongest advantage of the technique is that it results in a mapping of the centimeter-scale height variations of the water surface across vast wetlands. The spatial resolution of most L-band SAR systems can be used to measure water level changes for ~100-m wide water bodies and greater.

Applications

Although the Amazon main channel contains a small number of stream gauges, there are no gauges on the floodplain itself, opening the question of just how much water is flowing through the vast wetland. A Muskingum model based approach by Richey *et al.* (1989) has provided a significant first step toward estimating the storage volume on the floodplain. However, it assumed that water levels on the floodplain were equivalent to the main channel, whereas Alsdorf (2003) used InSAR data to demonstrate that during recessional flow, water levels are not horizontal and floodplain stage changes are not equivalent to those of the main channel.

Figure 6 provides another view of water level changes using interferometric JERS-1 SAR data collected on April 14 and July 11 of 1996, that is, during inundation of the floodplain. Differential variations in water level changes across the floodplain are notable between floodplain channels (western and eastern most arrows), whereas flooded valley lakes (central arrow) also show water level changes. If flow during inundation mimicked a horizontal water surface with fluctuations equivalent to the main channel, then these interferometric phase values would be uniform across the entire floodplain. Instead, the phase transitions are spatially variable suggesting more complex flow hydraulics.

The great advantage of InSAR methods is the spatial distribution of water level change measurements spread throughout an entire wetland area, whereas the requirement of flooded vegetation to return the off-nadir radar pulse is a strong limitation. As noted in Figure 6, some locations



(a)

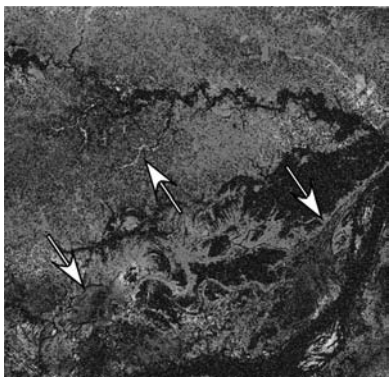

(b)

Figure 6 The Amazon floodplain during inundation. (a) JERS-1 backscatter intensity from April 14, 1996 and (b) interferogram showing water surface changes from April 14 to July 11, 1996. Images are about 70 km × 70 km and are located about 100 km west of Manaus, over the Cabaliana floodplain. In (a), the main Amazon channel as well as some floodplain channels are marked by low amplitudes, whereas most of the Cabaliana floodplain includes bright ''double-bounce'' returns. In (b), the interferometric phase values show sharp transitions within the floodplain (arrows in b). These sharp transitions are indicative of complex flow hydraulics where some portions of the floodplain are filling differently than immediately adjacent regions. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

such as the Amazon floodplain are abundant with flooded forest and with interferometric phase signatures of changing water levels. Such measurements show for the first time, the spatially rapid variation in floodplain water surfaces thus indicating similarly strong variations in wetland flow hydraulics. A challenge ahead is the integration of the high-spatial resolution elevation-change capabilities of InSAR with profile elevation-change data from radar altimetry. Already some promise has been shown by combining T/P data over a large Amazon lake with InSAR data to show hydrologic volumetric loss and gain (Alsdorf *et al.*, 2001b). In the following section, the advantageous along-track spatial resolution of laser altimetry should allow direct

comparisons with InSAR as well as increasing the global coverage of surface water hydraulics.

LASER ALTIMETRY

An alternative altimetric approach to measuring inland water stage can be provided by airborne and spaced-based laser altimetry. Like radar altimetry, the measurement principal is based on precisely measuring the round-trip travel time of a pulse, in this case a brief laser pulse typically 1 to 10 ns in duration. Converting travel time to distance based on the speed of light yields the distance from the instrument to the reflecting target; corrections for atmospheric effects on propagation speed and refraction are required but these are a significantly smaller than atmospheric effects on microwave propagation. Combining the distance measurement with knowledge of the instrument position typically obtained using the Global Positioning System or microwave and/or optical tracking data, and the orientation of the laser pulse, obtained with gyroscopic and/or star tracker instrumentation, yields the horizontal and vertical position of the reflecting target. Laser altimeters have been used extensively from airborne platforms for topographic mapping of ice sheets (e.g. Krabill *et al.*, 2002), land topography (e.g. Haugerud *et al.*, 2003), and forest canopy structure (e.g. Lefsky *et al.*, 2002) and from satellites for global topographic observations of the Moon (Smith *et al.*, 1997), Earth (Garvin *et al.*, 1998), Mars (Smith *et al.*, 1999), and an asteroid (Zuber *et al.*, 2000).

Advantages and Limitations

Although not yet used extensively for observing inland water, characteristics of laser-based altimeter measurement (Bufton, 1989) hold the promise for important lake and river observations. A prime advantage is that the tightly collimated laser pulse enables operation with smaller instrument footprints and thus higher spatial-resolution sampling than traditional radar altimetry (see Table 1).

With submilliradian beam divergence, submeter footprints can be achieved from airborne platforms while footprints as small as tens-of-meters can be achieved from orbital altitudes. With these small-diameter footprints, a laser altimeter is not limited to observing just the sub-satellite nadir point. Laser altimeters can thus acquire data at off-nadir orientations, permitting cross-track scanning to map water surface elevations within a swath. In addition, targeted data acquisitions can be made for water bodies and rivers located on either side of the satellite's ground track. A limitation to off-nadir measurements though is imposed by loss of backscatter signal strength (a function of off-nadir angle and surface roughness, Bufton *et al.*, 1983) caused by specular reflection of the laser pulse from the water surface. Smooth water surfaces under calm wind conditions can thus be difficult to observe at angles that are more than several degrees off-nadir. Also, geolocation of the laser footprints requires accurate knowledge of the laser-pointing vector, at the arc-second level for orbital implementations, requiring greater instrument complexity than a radar altimeter.

An additional benefit of laser altimetry over radar altimetry is the absence of coherent fading. While radar altimetry has to average many echoes, each individual laser pulse provides a unique, elevation measurement with centimeter- to decimeter-level accuracy. High-repetition rate lasers can achieve contiguous footprints along the ground track at orbital velocities, or within scanned swaths at aircraft velocities, providing complete sampling of surface elevations at high-spatial resolution. Because short duration pulses are used, submeter resolution of vertically distinct surfaces such as water and overlying vegetation in a flooded forest can be achieved using the backscatter signal from a single laser pulse, where sufficient gaps in the vegetation allow illumination of the water surface. The water returns correspond to the elevation of the water surface because the near-infrared laser pulses used by most systems negligibly penetrate into the water column.

Operating at optical wavelengths, cloud cover imposes a significant limitation upon laser altimeters, reducing the quantity and accuracy of stage measurements obtained with an orbital system. However, unlike passive optical systems, by using range-gating techniques, laser returns from the Earth's surface can be acquired through thin to moderately dense clouds, up to an optical depth of approximately two. Reduced signal strength and multiple-scattering effects in the presence of clouds do, however, degrade vertical accuracy (Duda *et al.*, 2001; Mahesh *et al.*, 2002).

These characteristics of laser altimeter systems provide the potential to measure the stage of inland water including small water bodies and narrow rivers unresolved by radar altimetry. In addition, small-diameter, contiguous footprints provide a means to determine the bank-to-bank extent of the water surface, even where obscured by vegetation cover, and measure surface gradients induced by flowing water and/or wind.

Applications

To date, laser altimeter measurements of water surfaces in general have focused on calibration of instrumental biases, assessment of vertical accuracy, and characterization of ocean wave structure. Acting essentially as a reference surface while conducting platform attitude maneuvers, lake returns have been used to establish pointing and timing biases for airborne (Vaughn *et al.*, 1996; Ridgway *et al.*, 1997; Hofton *et al.*, 2000) and orbital (Luthcke *et al.*, 2000, 2001, 2002) instruments. After validation exercises over Crowley Lake in Long Valley, California, Hofton *et al.* (2000) achieved a standard deviation from the mean lake level of 2.9 cm for 2500 laser-determined elevations acquired along a 5-km flight path. The gradient induced

in the lake surface by the local gravity potential (i.e. geoid) of 8 cm over 5 km was also observed. By using a radar altimetry mean sea surface model, Luthcke *et al.* (2002) calibrated time-varying instrumental biases for the Shuttle Laser Altimeter (SLA) by minimizing laser range residuals with respect to the open ocean surface. The vertical accuracy of the SLA instrument, described by the rms ocean surface range residuals after calibration, was ~1 m. SLA vertical accuracy, limited by its ranging precision, nonoptimal means to determine Shuttle orbit and attitude, and unmodeled instrumental biases, is an order of magnitude worse than that expected for an optimized orbital laser altimeter system (Luthcke *et al.*, 2002). Hwang *et al.* (2000, 2002) used an airborne scanning system to map three-dimensional ocean wave structure, assess the spectral properties of wave fields, and quantify the distribution of breaking waves. Carter *et al.* (2001) used a similar system to measure the height and slope of sheet flow water in the Florida Everglades with 4 cm vertical precision. In this study, the intensity of the laser backscatter signal was used to identify specular nadir returns from the smooth water surface, thus eliminating returns from overlying grasses.

These initial studies demonstrate some of the potential for laser altimeter measurements of inland surface water characteristics, with prospects for data of high-spatial resolution and centimeter-level precision from future systems designed specifically to address hydrologic monitoring objectives.

FUTURE OUTLOOK

With regards to satellite radar altimetry, the performance of the newly designed RA-2 instrument onboard the ENVISAT satellite will be examined to check on expected tracking continuity (Resti, 1993) over various inland waters. Such a capability may provide surface water height in complex or rugged regions for which previously there was much loss of elevation information. Projects are also underway that are examining various retracking options in lieu of range (and thus) stage determination and accuracy. The future of interferometric SAR for measuring water level changes will continue to rely on vegetation to provide the required scattering centers for returning the radar pulse to the antennae. This also limits the effectiveness of shorter wavelength pulses such as C-band and smaller. In fact, research is still needed to carefully define the vegetation habitats capable of providing this returned echo.

In the near-term, ICESat is providing an opportunity to assess orbital laser altimeter capabilities over water bodies. This NASA Earth Observing System mission was launched in January, 2003 and acquires globally distributed profiles with the Geoscience Laser Altimeter System (GLAS). Observations of ice sheet elevation change, cloud and aerosol heights, land topography, and vegetation characteristics are being made (Zwally *et al.*, 2002). Recording the backscattered infrared laser energy from ~70 m diameter footprints spaced ~175 m along the ground tracks, the predicted elevation accuracy (1 sigma) is 15 cm for flat surfaces. The spacecraft attitude is controlled to point the laser beam to within 35 m (1 sigma) of a specified ground track, enabling selected features to be profiled. Kwok *et al.* (2004) used ICESat data to measure the extent and elevation of open water leads within Arctic sea ice. To assess backscatter strength and water stage accuracy as a function of off-nadir angle, ICESat observations are being acquired for reservoirs and rivers with *in situ* gauge measurements at a variety of angles up to 5°. Methods for automated identification of laser pulses returned from inland water surfaces are being evaluated, and data from flooded forests are being examined to assess waveform retrieval of water surface elevations beneath vegetation. Experience gained from ICESat will thus aid in the design of the next generation of orbital laser altimeters optimized for measurement of inland water stage, extent, and slope.

The application of satellite radar altimetry, laser altimetry, and interferometric SAR to the determination of inland water stage and its variation, has presented the scientific communities with some interesting challenges and thoughts for the future. All instrument types were not specifically designed with inland water applications in mind, but have shown sufficient potential to pave the way for future dedicated missions with modifications to suit the hydrological communities (Vörösmarty *et al.*, 1999; Vörösmarty *et al.*, 2001; Alsdorf *et al.*, 2003; Alsdorf and Lettenmaier, 2003). Almost certainly the ability to acquire further river and lake parameters, such as river width, river discharge, and lake volume would greatly assist both the routine ground monitoring networks and significantly contribute to some of the main science questions (e.g. of the Earth System Science, Land Hydrology, and Water Cycle Divisions within NASA). In this respect, several new groups have emerged under the ESA and NASA auspices (e.g. **`http://www.swa.com/hydrawg`**) to continue to examine the status of current ground-based networks, identify the major science issues and parameter requirements, and promote the advancement of new sensor techniques.

REFERENCES

Alsdorf D.E. (2003) Water storage of the central Amazon floodplain measured with GIS and remote sensing imagery. *Annals of the Association of American Geographers*, **93**(1), 55–66.

Alsdorf D.E., Birkett C.M., Dunne T., Melack J. and Hess L. (2001b) Water level changes in a large Amazon lake measured with spaceborne radar interferometry and altimetry. *Geophysical Research Letters*, **28**(14), 2671–2674.

Alsdorf D.E. and Lettenmaier D.P. (2003) Tracking fresh water from space. *Science*, **301**, 1491–1494.

Alsdorf D., Lettenmaier D., Vörösmarty C., and the NASA Surface Water Working Group. (2003) The need for global, satellite-based observations of terrestrial surface waters. *EOS, Transaction, AGU*, **84**(29), 269–276.

Alsdorf D.E., Melack J.M., Dunne T., Mertes L.A.K., Hess L.L. and Smith L.C. (2000) Interferometric radar measurements of water level changes on the Amazon floodplain. *Nature*, **404**, 174–177.

Alsdorf D.E., Smith L.C. and Melack J.M. (2001a) Amazon floodplain water level changes measured with interferometric SIR-C radar. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1–13.

Au A.Y., Brown R.D. and Welker J.E. (1989) *Analysis of Altimetry over Inland Seas*, NASA Technical Memorandum, 100729, NASA.

Birkett C.M. (1994) Radar altimetry – a new concept in monitoring global lake level changes. *EOS, Transaction, AGU*, **75**(24), 273–275.

Birkett C.M. (1995) The contribution of TOPEX/POSEIDON to the global monitoring of climatically sensitive lakes. *Journal of Geophysical Research*, **100**(C12), 25,179–25,204.

Birkett C.M. (1998) Contribution of the TOPEX NASA radar altimeter to the global monitoring of large rivers and wetlands. *Water Resources Research*, **34**(5), 1223–1239.

Birkett C.M. (2000) Synergistic remote sensing of Lake Chad: variability of basin inundation. *Remote Sensing of the Environment*, **72**, 218–236.

Birkett C.M. and Mason I.M. (1995) A new global lakes database for a remote sensing programme studying climatically sensitive large lakes. *The Journal of Great Lakes Research*, **21**(3), 307–318.

Birkett C.M., Mertes L.A.K., Dunne T., Costa M. and Jasinski J. (2002) Altimetric remote sensing of the Amazon: application of satellite radar altimetry. *JGR*, **107**(D20), 8059, 10.1029/2001JD000609.

Birkett C., Murtugudde R. and Allan T. (1999) Indian Ocean climate event brings floods to East Africa's lakes and the Sudd Marsh. *GRL*, **26**(8), 1031–1034.

Brooks R.L. (1982) *Lake Elevations from Satellite Radar Altimetry from a Validation Area in Canada, Report*, Geoscience Research Corporation: Salisbury.

Brown G.S. (1977) The average impulse response of a rough surface and its applications. *IEEE Transactions on Antennas Propagation*, **AP-25**(1), 67–74.

Bufton J.L. (1989) Laser altimetry measurements from aircraft and spacecraft. *Proceedings of the IEEE*, **77**(3), 463–477.

Bufton J.L., Hoge F.E. and Swift R.N. (1983) Airborne measurements of laser backscatter from the ocean surface. *Applied Optics*, **22**(17), 2603–2618.

Campos I., de O., Mercier F., Maheu C., Cochonneau G., Kosuth P., Blitzkow D. and Cazenave A. (2001) Temporal variations of river basin waters from Topex/Poseidon satellite altimetry: application to the Amazon basin. *Earth and Planetary Sciences*, **333**(10), 633–643.

Carter W., Shrestha R., Tuell G., Bloomquist D. and Sartori M. (2001) Mapping the surface sheet flow water in the Everglades. *International Archives of Photogrammetry and Remote Sensing, XXXIV-3/W4, Proceedings of the ISPRS Workshop Land Surface Mapping and Characterization Using Laser Altimetry*, Annapolis, 22–24 October 2001.

Cazenave A., Bonnefond P., Dominh K. and Schaeffer P. (1997) Caspian Sea level from Topex-Poseidon altimetry: level now falling. *Geophysical Research Letters*, **24**(8), 881–884.

Chelton D.B., Ries J.C., Haines B.J., Fu L.-L. and Callahan P.S. (2001) Satellite altimetry. In *Satellite Altimetry and the Earth Sciences: A Handbook of Techniques and Applications*, Fu L.-L. and Cazenave A. (Eds.), Academic Press: San Diego, pp. 1–131.

Chelton D.B., Walsh E.J. and MacArthur J.L. (1988) 'Nuts and Bolts' of satellite radar altimetry. *Proceedings of the WOCE/NASA Altimeter Algorithm Workshop*, Appendix to U.S. WOCE, Technical Report No.2, Chelton D.B. (Ed.), Corvallis, pp. 1–43.

Cudlip W., Ridley J.K. and Rapley C.G. (1992) The use of satellite radar altimetry for monitoring wetlands. *Proceedings of the 16th Annual Conference of Remote Sensing Society: Remote Sensing and Global Change*, London, pp. 207–216.

Dalton J.A. and Kite G.W. (1995) A first look at using the TOPEX/Poseidon satellite radar altimeter for measuring lake levels. *Proceedings of the International Workshop on the Application of Remote Sensing in Hydrology*, NHRI Symposium 0838-1984, No. 14, Kite G.W., Pietroniro A. and Pultz T.D. (Eds.), Saskatoon pp. 105–112.

Duda D.P., Spinhirne J.D. and Eloranta E.W. (2001) Atmospheric multiple scattering effects on GLAS altimetry – Part I: calculations of single pulse bias. *IEEE Transactions of Geoscience and Remote Sensing*, **39**(1), 92–101.

Farr T. and Kobrick M. (2001) The shuttle radar topography mission. *EOS*, **82**, 47.

Garvin J., Bufton J., Blair J., Harding D., Luthcke S., Frawley J. and Rowlands D. (1998) Observations of the earth's topography from the Shuttle LaserAltimeter (SLA): laser-pulse echo-recovery measurements of terrestrial surfaces. *Physics and Chemistry of the Earth*, **23**, 1053–1068.

Goldstein R.M., Engelhardt H., Kamb B. and Frolich R.M. (1993) Satellite radar interferometry for monitoring ice sheet motion: application to an Antarctic ice stream. *Science*, **262**, 1525–1530.

Guzkowska M.A.J., Rapley C.G., Ridley J.K., Cudlip W., Birkett C.M. and Scott R.F. (1990) *Developments in Inland Water and Land Altimetry*, ESA Contract, CR-7839/88/F/FL.

Haugerud R., Harding D.J., Johnson S.Y., Harless J.L., Weaver C.S. and Sherrod B.L. (2003) High-resolution lidar topography of the Puget Lowland, Washington – a bonanza for earth science. *GSA Today*, **13**(6), 4–10.

Hess L.L., Melack J.M., Filoso S. and Wang Y. (1995) Delineation of inundated area and vegetation along the Amazon floodplain with SIR-C synthetic aperture radar. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 896–904.

Hofton M.A., Blair J.B., Minster J.B., Ridgway J.R., Williams N.P., Bufton J.L. and Rabine D.L. (2000) An airborne scanning laser altimetry survey of Long Valley, California. *International Journal of Remote Sensing*, **21**(12), 2413–2437.

Hwang P.A., Wang D.W., Walsh E.J., Krabill W.B. and Swift R.N. (2000) Airborne measurements of the wavenumber spectra of ocean surface waves. Part I: spectral slope and dimensionless

spectral coefficient. *Journal of Physical Oceanography*, **30**(11), 2753–2767.

Hwang P.A., Wright W., Krabill W.B. and Swift R.N. (2002) Airborne remote sensing of breaking waves. *Remote Sensing of Environment*, **80**, 65–75.

Kite G.W. and Pietroniro A. (1996) Remote sensing applications in hydrological modeling. *Hydrological Sciences Journal*, **41**(4), 563–591.

Koblinsky C.J., Clarke R.T., Brenner A.C. and Frey H. (1993) Measurement of river level variations with satellite altimetry. *Water Resources Research*, **29**(6), 1839–1848.

Krabill W.B., Abdalati W., Frederick E.B., Manizade S.S., Martin C.F., Sonntag J.G., Swift R.N., Thomas R.H. and Yungel J.G. (2002) Aircraft laser altimetry measurement of elevation changes of the greenland ice sheet: technique and accuracy assessment. *Journal of Geodynamics*, **34**(3–4), 357–376.

Kwok R., Zwally H.J. and Yi D. (2004) ICESat observations of Arctic sea ice: a first look. *Geophysical Research Letters*, **31**(16), Article No. L16401.

Lanfear K.J. and Hirsch R.M. (1999) USGS study reveals a decline in long-record stream gauges. *EOS, Transaction, AGU*, **80**(50), 605–607.

Lefsky M.A., Cohen W.B., Parker G.G. and Harding D.J. (2002) Lidar remote sensing for ecosystem studies. *Bioscience*, **52**(1), 19–30.

Li F.K. and Goldstein R.M. (1990) Studies of multibaseline spaceborne interferometric synthetic aperture radars. *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 88–97.

Luthcke S.B., Carabajal C.C. and Rowlands D.D. (2002) Enhanced geolocation of spaceborne laser altimeter surface returns: parameter calibration from the simultaneous reduction of altimeter range and navigation tracking data. *Journal of Geodynamics*, **34**(3–4), 447–475.

Luthcke S.B., Carabajal C.C., Rowlands D.D. and Pavlis D.E. (2001) Improvements in spaceborne laser altimeter data geolocation. *Surveys in Geophysics*, **22**(6), 549–559.

Luthcke S.B., Rowlands D.D., McCarthy J.J., Pavlis D.E. and Stoneking E. (2000) Spaceborne laser-altimeter-pointing bias calibration from range residual analysis. *Journal of Spacecraft and Rockets*, **37**(3), 374–384.

Mahesh A., Spinhirne J.D., Duda D.P. and Eloranta E.W. (2002) Atmospheric multiple scattering effects on GLAS altimetry – Part II: analysis of expected errors in antarctic altitude measurements. *IEEE Transactions on Geoscience and Remote Sensing*, **40**(11), 2353–2362.

Maheu C., Cazenave A. and Mechoso C.R. (2003) Water level fluctuations in the Plata Basin (South America) from Topex/Poseidon satellite altimetry. *Geophysical Research Letters*, **30**(3), 1143–1146.

Massonnet D. and Feigl K.L. (1998) Radar interferometry and its application to changes in the earth's surface. *Reviews of Geophysics*, **36**, 441–500.

Massonnet D., Rossi M., Carmona C., Adragna F., Peltzer G., Feigl K. and Rabaute T. (1993) The displacement field of the Landers earthquake mapped by radar interferometry. *Nature*, **364**, 138–142.

Mercier F., Cazenave A. and Maheu C. (2002) Interannual lake level fluctuations (1993–1999) in Africa from Topex/Poseidon: connections with ocean-atmosphere interactions over the Indian Ocean. *Global and Planetary Changes*, **32**, 141–163.

Mertes L.A.K. (2002) Remote sensing of riverine landscapes. *Freshwater Biology*, **47**, 799–815.

Mertes L.A.K., Daniel D.L., Melack J.M., Nelson B., Martinelli L.A. and Forsberg B.R. (1995) Spatial patterns of hydrology, geomorphology, and vegetation on the floodplain of the Amazon river in Brazil from a remote sensing perspective. *Geomorphology*, **13**, 215–232.

Mertes L.A.K., Dekker A.G., Brakenridge G.R., Birkett C.M. and Létourneau G. (2004) Rivers and Lakes. In *Natural Resources and Environment Manual of Remote Sensing*, Vol. 4, Ustin S.L. and Rencz A. (Eds.), John Wiley & Sons: New York.

Morris C.S. and Gill S.K. (1994a) Variation of Great Lakes water levels derived from Geosat altimetry. *Water Resources Research*, **30**(4), 1009–1017.

Morris C.S. and Gill S.K. (1994b) Evaluation of the TOPEX/POSEIDON altimeter system over the Great Lakes. *Journal of Geophysical Research*, **99**(C12), 24,527–24,539.

Olliver J.G. (1987) An analysis of results from SEASAT altimetry over land and lakes. paper presented at *IAG Symposium, IUGG XIX General Assembly*, International Association of Geodesy: Vancouver.

Ponchaut F. and Cazenave A. (1998) Continental lake level variations from TOPEX/POSEIDON (1993–1996). *Earth and Planetary Sciences*, **326**, 13–20.

Rapley C.G., Guzkowska M.A.J., Cudlip W. and Mason I.M. (1987) *An Exploratory Study of Inland Water and Land Altimetry Using Seasat Data*, ESA Report 6483/85/NL/BI, European Space Agency, Neuilly.

Resti A. (1993) Envisat's radar altimeter: RA-2. *ESA Bulletin-European Space Agency*, **76**, 58–60.

Richey J.E., Mertes L.A.K., Dunne T., Victoria R.L., Forsberg B.R., Tancredi A.C.N.S. and Oliveira E. (1989) Sources and routing of the Amazon river flood wave. *Global Biogeochemical Cycles*, **3**, 191–204.

Ridgway J.R., Minster J.B., Williams N., Bufton J.L. and Krabill W.B. (1997) Airborne laser altimeter survey of Long Valley, California. *Geophysical Journal International*, **131**(2), 267–280.

Sarch M.T. and Birkett C.M. (2000) Fishing and farming at Lake Chad: responses to lake level fluctuations. *The Geographical Journal*, **166**(2), 156–172.

Scott R.F., Baker S.G., Birkett C.M., Cudlip W., Laxon S.W., Mantripp D.R., Mansley J.A., Morley J.G., Rapley C.G., Ridley J.K., *et al.* (1994) A comparison of the performance of the ice and ocean tracking modes of the ERS 1 radar altimeter over non-ocean surfaces. *Geophysical Research Letters*, **21**(7), 553–556.

Smith L.C. (1997) Satellite remote sensing of river inundation area, stage, and discharge: a review. *Hydrological Processes*, **11**, 1427–1439.

Smith D.E., Zuber M.T., Neumann G.A. and Lemoine F.G. (1997) Topography of the moon from the Clementine lidar. *Journal of Geophysical Research – Planets*, **102**(E1), 1591–1611.

Smith D.E., Zuber M.T., Solomon S.C., Phillips R.J., Head J.W., Garvin J.B., Banerdt W.B., Muhleman D.O., Pettengill G.H., Neumann G.A., *et al.* (1999) The global topography

of Mars and implications for surface evolution. *Science*, **284**, 1421–1576.

Ulaby F.T., Moore R.K. and Fung A.K. (1981) Microwave remote sensing fundamentals and radiometry. *Microwave Remote Sensing Active and Passive*, Vol.1, Simonett D.S. (Ed.), Artech House: Boston.

Vaughn C.R., Bufton J.L., Krabill W.B. and Rabine D. (1996) Georeferencing of airborne laser altimeter measurements. *International Journal of Remote Sensing*, **17**(11), 2185–2200.

Vörösmarty C., Askew A., Grabs W., Barry R.G., *et al.* (2001) Global water data: a newly endangered species. *EOS, Transaction, AGU*, **82**(5), 54–58.

Vörösmarty C., Birkett C.M., Dingman L., Lettenmaier D.P., Kim Y., Rodriguez E. and Emmitt G.D. (1999) HYDRA-SAT HYDRological altimetry satellite, NASA Post-2002 land surface Hydrology Mission component for surface water monitoring. *White Paper Report from the NASA Post-2002 Land Surface Hydrology Workshop*, Irvine, April 12–14th.

Zandbergen R.C.A. (1990) *Satellite Altimeter Data Processing: from Theory to Practice*, Delft University Press: Delft.

Zebker H.A., Rosen P.A., Goldstein R.M., Gabriel A. and Werner C.L. (1994) On the derivation of coseismic displacement fields using differential radar interferometry: the landers earthquake. *Journal of Geophysical Research*, **99**, 19617–19634.

Zebker H.A. and Villasenor J. (1992) Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 950–959.

Zuber M.T., Smith D.E., Cheng A.F., Garvin J.B., Aharonson O., Cole T.D., Dunn P.J., Guo Y.P., Lemoine F.G., Neumann G.A., *et al.* (2000) The shape of 433 eros from the NEAR-shoemaker laser rangefinder. *Science*, **289**, 2097–2102.

Zwally H.J., Schutz B., Abdalati W., Abshire J., Bentley C., Brenner A., Bufton J., Dezio J., Hancock D., Harding D., *et al.* (2002) ICESat's laser measurements of polar ice, atmosphere, ocean, and land. *Journal of Geodynamics*, **34**(3–4), 405–445.

61: Estimation of River Discharge

S LAWRENCE DINGMAN[1] AND DAVID M BJERKLIE[2]

[1]*Earth Sciences Department, University of New Hampshire, Durham, NH, US*
[2]*United States Geological Survey, Hartford, CT, US*

River discharge, the volume flow rate through a river cross section, is perhaps the most important single hydrologic quantity. It is a major link in the global hydrologic and geologic cycles and a critical component of climate, and represents the rate at which nature makes water available for human use and management. In the form of floods, it constitutes one of the most destructive natural hazards. Measurement of river discharge thus provides vital information for science and society. Measurement by conventional ground-based methods is expensive and is declining globally. Using various combinations of active and passive imagery and sensors mounted on satellites or aircraft, it is possible to obtain direct quantitative information on several hydraulic variables, including channel configuration and the area, width, elevation, and velocity of the water surface. Computations using conventionally obtained data indicate that this information can be used in various combinations with statistical models and topographic information to generate quantitative time- and location-specific estimates of discharge. Thus, it appears feasible to produce useful river-discharge information via remote sensing, especially for locations that are remote or otherwise difficult to observe conventionally. However, it is likely that this capability will be useful only for relatively large rivers and that, even for these, measurement accuracy will be considerably lower than is possible with ground-based measurements. Given the scientific and societal importance of river-discharge observations, and the present discharge-measurement network and unlikely prospects for its expansion, further development and testing of these approaches, development of concurrent sets of pertinent remotely sensed variables, and a more complete evaluation of the spacing and timing of observations possible with satellite systems is warranted.

INTRODUCTION

This article focuses on the measurement of river discharge via remote sensing, that is, obtaining quantitative time- and location-specific estimates of river discharge using only information provided by ground-, aircraft-, or satellite-mounted sensors not in contact with the water. Thus, it does not include discussion of (i) up-links to satellites from river-based sensors of velocity, width, depth, or water level; or (ii) estimates of river discharge developed from hydrologic models that use remotely sensed information on precipitation, soil moisture, snow, or other hydrologic state variables.

RIVER DISCHARGE

Definition

River discharge, Q, is the volume rate of flow [$L^3\ T^{-1}$] through a stream cross section (Figure 1). It is the product of the cross-sectional area of the flow, A, and the cross-sectional average velocity, U; A is the product of the water-surface width, W, and the cross-sectional average depth, Y. Thus,

$$Q = A \cdot U = W \cdot Y \cdot U \quad (1)$$

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

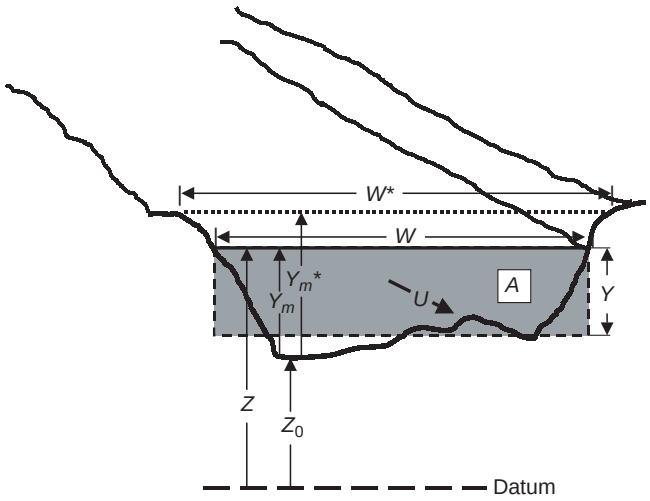


Figure 1 Definitions of quantities involved in river-discharge measurement. $W \equiv$ water-surface width; $Y \equiv$ average depth; $Y_m \equiv$ maximum depth; $U \equiv$ average flow velocity; $A \equiv$ flow cross-sectional area (shaded); $Z \equiv$ water-surface elevation (stage); $Z_0 \equiv$ zero-flow stage. Asterisks denote bank-full quantities

Importance

Global Hydrologic and Geologic Cycles

On the global scale, the time-integrated hydrologic cycle can be depicted as in Figure 2 – the world's oceans receive about 385 000 $km^3\,year^{-1}$ in precipitation and lose 425 000 $km^3\,year^{-1}$ in evaporation, while the continents receive 111 000 $km^3\,year^{-1}$ in precipitation and lose 72 000 $km^3\,year^{-1}$ via evapotranspiration. River discharge is the link that balances the global cycle, returning about 40 000 $km^3\,year^{-1}$ from the continents to the oceans (*see* **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1**).

River discharge is also a major link in the global geologic cycle, delivering some 13.5×10^9 T $year^{-1}$ of particulate material and 3.9×10^9 T $year^{-1}$ dissolved material from the continents to the oceans (Walling and Webb, 1987). The dissolved material is an important determinant of ocean chemistry and constitutes the major source of nutrients for the oceanic food web.

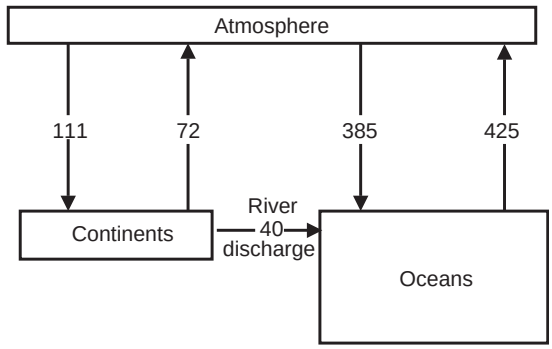


Figure 2 The spatially and temporally aggregated global hydrologic cycle. Fluxes are in 1000 $km^3\,year^{-1}$

Climate

River discharge plays a critical role in regulating global climate. Its effects on sea-surface temperatures and salinities, particularly in the North Atlantic Ocean, drive the global thermohaline circulation that transports heat from low to high latitudes. River inflows also maintain relatively low salinity in the Arctic Ocean, which makes possible the freezing of its surface; the reflection of the sun's energy by this sea ice is an important factor in the Earth's energy balance.

River discharge is also a sensitive indicator of climate change. A simple model can be developed that relates the relative change in discharge, q, ($q \equiv$ ratio of discharge after climate change to discharge before climate change) to the relative change in precipitation, p (Wigley and Jones, 1985):

$$q = \frac{p - (1 - w) \cdot e}{w} \qquad (2)$$

where w is the proportion of precipitation that runs off (a function of local geology and climate) and e is the relative change in evapotranspiration. This model shows that q is always greater than p; for example, with a 10% increase in precipitation ($p = 1.1$) and no change in evapotranspiration ($e = 1$), $q = 1.2$ (20% increase) where $w = 0.5$ (humid regions) and $q = 1.5$ (50% increase) where $w = 0.2$ (semiarid regions). Observations of long-term trends in precipitation and streamflow consistently show this amplification effect (e.g. Karl and Riebsame, 1989).

Water Resources and Flood Hazards

At all spatial scales, from small watersheds to continents, river discharge constitutes all or most of the residual of precipitation minus evapotranspiration, and hence is the rate at which nature makes water available for human use. Water flowing in streams provides a habitat for fish and other organisms of ecological and commercial importance and it is utilized for a wide range of vital purposes, including the transport and treatment of human and industrial wastes, commercial navigation, generation of hydroelectric power, recreation, and esthetic enjoyment. River flow is also the principal source of water withdrawn for human consumption, irrigation, cooling of thermoelectric power plants, and other commercial and industrial uses. Demand for water for all these purposes is growing with population, and roughly 1/3 of the world's peoples currently live under moderate-to-high water stress (Vörösmarty *et al.*, 2000). Long records of river discharge are essential tools for water-resource management because regional water availability is assessed by analyses of the time distribution of discharge in the region.

River flooding is one of the most destructive natural hazards globally. In the United States, flood damages average about $\$4 \times 10^9\,\text{year}^{-1}$ and are increasing rapidly because of the increasing concentration of people and infrastructure in flood-prone areas (van der Vink *et al.*, 2004). Assessment of flood hazards, and of the economic, environmental, and social benefits and costs of various strategies for reducing future flood damages at a riparian location are based on frequency analyses of extreme river discharges at that location. Thus, time-series records of discharge are an essential component of flood-hazard management. Satellite images of current locations of riverine flooding around the globe can be viewed at **www.dartmouth.edu/artsci/geog/floods** (G.R. Brakenridge, Department of Geography, Dartmouth College).

Validation of Hydrologic Models

River discharge is the time- and space-integrated output of the regional hydrologic cycle. And, because river flow is generally concentrated in channels, discharge can in principle be measured with considerably more accuracy and precision than can precipitation, evapotranspiration, or any other component of the hydrologic cycle. For these reasons, measurements of river discharge, in addition to providing direct information about climate, water-resource availability, and flood hazards, are also invaluable for validating the hydrologic models that are the only means of forecasting the effects of land use and climate change on water resources.

Conventional Measurement Techniques

Specialized Methods

Small flows in channels less than a few meters in width can be measured directly (i) volumetrically, (ii) by dilution gaging, and (iii) in weirs or flumes. Volumetric measurement involves diverting the flow and measuring the time required to fill a container; clearly this is feasible only for extremely small flows. In dilution gaging, a known concentration of a conservative tracer is introduced into the flow and the time distribution of its concentration is measured at a downstream location. This technique is suitable for small, highly turbulent streams where complete mixing occurs over short distances (see White, 1978; Dingman, 2002).

In relatively small research watersheds, discharge can be accurately monitored by directing the flow through specially designed weirs or flumes. These devices provide stable relations between water level and discharge, which in many cases can be derived from hydraulic theory (White, 1978; Herschy, 1999a). Generally, however, calibration by volumetric or velocity-area measurement (see "velocity area method") is required over at least some ranges of discharge.

Velocity-Area Method

In streams with widths greater than a few meters, discharge is almost always measured by the velocity-area method (Figure 3). In this method, observations of depth, Y_i, and average velocity, U_i, are made at successive locations, X_i, in a cross section, and discharge is calculated as

$$Q = U_1 \cdot Y_1 \cdot \frac{|X_2 - X_1|}{2} + \sum_{i=2}^{N-1} U_i \cdot Y_i \cdot \frac{|X_{i+1} - X_{i-1}|}{2} + U_N \cdot Y_N \cdot \frac{|X_N - X_{N-1}|}{2} \tag{3}$$

where X_1 and X_N are located at the edges of water at either end of the cross section. Usually $N = 25$ to 30, and the average velocity at each location is estimated from velocity measurements taken at prescribed depths (see Buchanan and Somers (1969), Herschy (1995), or Dingman (2002) for more detailed discussions).

A recent modernization of the velocity-area method uses an Acoustic Doppler Current Profiler (ADCP) to simultaneously measure and integrate the depth and velocity across a channel section, thereby obtaining all of the elements of equation (3) in one pass (Simpson and Oltman, 1992; Morlock, 1996). The ADCP unit is mounted on a boat or raft that traverses the cross section and measures depth via sonar and velocity via the Doppler shift of acoustic energy pulses. This system greatly reduces the time necessary to make a discharge measurement and allows measurements at stages when wading is precluded and at locations lacking stream-spanning structures. However, because this method requires the instrument to have direct water contact, it is not considered a "remote" data-collection system.

The velocity-area method provides quasi-instantaneous measurements of discharge. Continuous records of discharge are developed by relating periodic velocity-area measurements to measured water-surface elevation (stage) to develop a stage-discharge "rating" for the cross section. The stage-discharge rating relation takes the general form

$$Q = a \cdot (Z - Z_0)^b \tag{4}$$

where the coefficients a and b are empirical values characteristic of the specific cross section, and Z_0 is the elevation of zero flow at that location (Rantz *et al.*, 1982; Herschy, 1995). In many cases, the values of a and b change for different ranges of discharge at a given reach.

For the periods between measurements of Q, the stage (Z) is recorded continuously or at frequent intervals and Q is inferred from the rating relation. (Methods of stage measurement are described by Herschy (1999b).) Since the value of Z_0 represents the elevation of zero flow, the term $(Z - Z_0)$ may be viewed as equivalent to the maximum flow depth (Y_m); thus the standard rating curve provides an estimate of discharge from the flow depth. In most situations, the rating relation is subject to change over time due to erosion and/or deposition in the measurement reach,

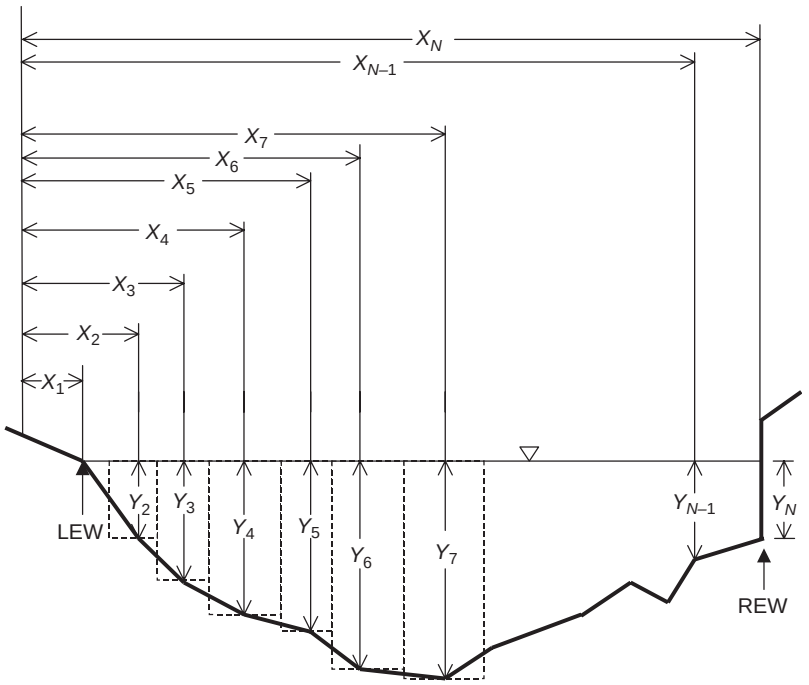


Figure 3 Definitions of terms involved in velocity-area discharge measurement (equation 3). Dashed lines indicate individual subsections, numbered consecutively $i = 1, 2, \ldots, N$. Left and right edges of water (LEW and REW) are defined for an observer facing downstream. (After Buchanan and Somers (1969))

so that periodic velocity-area measurements are required to maintain an accurate rating curve as well as to extend its range.

Most stream gages maintained by the United States Geological Survey (USGS) are now satellite-linked so that "real-time" observations of stage and discharge can be readily accessed via the Internet. This is an important advance in access to river-discharge data, but since the data are collected by conventional methods it is not considered a hydrological application of remote sensing.

Slope-Area Method

The discharge of a recent peak flow in a river reach can be estimated from observations of high-water marks by the slope-area method. This method involves surveying the slope of the water surface, S_s, as revealed by the high-water marks, the reach-averaged wetted perimeter, P, and cross-sectional area, A, beneath the marks. These observations are used to determine the hydraulic radius, R, where $R \equiv A/P$, and to estimate discharge, Q, via the Manning equation:

$$Q = \frac{u \cdot A \cdot R^{2/3} \cdot S_e^{1/2}}{n} \qquad (5)$$

In this equation, u is a unit-conversion factor (=1.00 for S.I. units and 1.49 for English units); the energy slope, S_e, is calculated from the water-surface slope and the reach geometry and generally differs little from S_s; and the reach-resistance factor, n, is estimated by the observer.

Values of n reflect the overall flow resistance of the reach and depend principally on boundary roughness (bed-particle size and bed forms), obstructions, variations in channel geometry, bed and bank vegetation, channel curvature, and slope. The value of n is usually selected based on the observer's judgment guided by tables (e.g. Chow, 1959), photographs (Barnes, 1967; Hicks and Mason, 1991; Coon, 1997), or formulas (Cowan, 1956). The detailed methodology is described by Dalrymple and Benson (1967), Ponce (1989), and Dingman (2002), among others.

Accuracy and Precision

In many situations, the allocation of water resources is governed by laws, treaties, regulations, or contracts. Hence measurement of river discharge often has significant legal, economic, and diplomatic implications, and measurement precision within a few percent is required, particularly when the flows are relatively low.

The precision of a velocity-area measurement depends principally on the regularity of the measured cross section, the number of observations (N), and the method used to determine average velocity. If $20 < N < 30$, accuracies of 5% to 8% can be achieved for individual measurements (i.e. 95% of the measurements will be within 5% to 8% of the true value) (Carter and Anderson, 1963). Precision and accuracy obtainable in slope-area measurements are generally considerably worse, largely because of the uncertainty involved in estimating n in equation (5) and because the

cross section at the time of the flow may be different than the measured cross section due to scour or fill (Kirby, 1987). Herschy (1995) indicated that the typical error range in slope-area estimates is 10% to 20%.

In addition to the factors affecting the precision and accuracy of individual velocity-area measurements, the accuracy of long-term gaging-station records is affected principally by the stability of the measurement cross section and the effects of ice on the stage-discharge relation. The USGS rates measurements at gaging stations as "excellent" (95% of measured values with in 5% of true value), "good" (95% within 10% of true), "fair" (95% within 15% of true), or "poor". Records at most long-term stations are rated "good"; "excellent" measurements can generally be obtained where discharge is measured via calibrated weirs or flumes. Dickinson (1967) concluded that the accuracy of estimates of discharge from a conventionally established rating curve range from about $\pm 12\%$ at the 80% confidence level and about $\pm 20\%$ at the 95% confidence level. Winter (1981) estimated that the uncertainty of long-term average values of streamflow at a gaging station is on the order of 5%.

Cost

The costs of establishing and maintaining a conventional long-term gaging station are considerable. Typical values used by the USGS are \$8000 to \$10 000 to establish a long-term monitoring station and \$6400/year (2003 US$) for operation and maintenance of each station, not including administrative overhead costs (B. Mrazik, USGS, personal communication).

Status of River-discharge Monitoring

Despite the importance of river-discharge information, expansion and maintenance of a comprehensive global river-monitoring network using conventional measurement methodologies face numerous technological, economic, and institutional obstacles (Rodda, 1999). As a result, gaging stations and access to river-discharge information have been declining since the 1980s (Vörösmarty *et al.*, 1999). The International Association of Hydrological Sciences (IAHS, 2001) reports that the density of long-term stream gages is well below WMO-recommended levels in many regions, particularly in Africa, and has declined significantly in the states of the former Soviet Union, Canada, and the United States. The number of stations reporting discharge data to the World Meteorological Association (WMO) Global Runoff Data Center (GRDC) declined by 90% during the period 1995–2000 (IAHS, 2001). Currently, less than 60% of the runoff from the continents is monitored at the point of inflow to the oceans (Fekete *et al.*, 1999) and the distribution of runoff within the continents is even less well monitored.

Hydrographic data obtained from satellites and other remote sources offer the possibility of broad and frequent coverage of river-discharge estimates (Barrett, 1998) and the potential to maintain or even increase the global streamflow-monitoring network. Ultimately, this may be a cost-effective method to obtain the required river-discharge data on a global scale.

REMOTE SENSING OF RIVER DISCHARGE

Introduction and Previous Work

Since discharge cannot be directly measured remotely, it must be determined from predictor variables that can be remotely observed. The feasibility of remote sensing of river discharge thus depends on (i) the accuracy and precision with which predictor variables can be remotely measured, (ii) the uncertainties inherent in models used to convert measurement of predictor variables to discharge estimates, (iii) the spatial and temporal distribution of observations of predictor variables, and (iv) the costs involved in obtaining measurements of predictor variables.

The use of remotely sensed information, including water surface elevation (Z), water-surface velocity (U_s), and water-surface area (A_s), to track changes in river discharge has been shown to be feasible and potentially useful where ground-based data are difficult to obtain (Kuprianov, 1978; Koblinsky *et al.*, 1993; Birkett, 1998; Brakenridge *et al.*, 1994; Brakenridge *et al.*, 1998; Horritt *et al.*, 2001; Birkett *et al.*, 2001). These studies suggest that remotely sensed river hydraulic data could be used to directly estimate the discharge at a specific location, if ground-based discharge measurements are used to develop discharge ratings in conjunction with the remotely observed variable(s).

Estimating discharge in rivers from remotely sensed hydraulic information obtained from aerial and satellite platforms has been explored and summarized by Smith *et al.* (1996), Smith (1997) and Bjerklie *et al.* (2003). Currently deployed satellite-based sensors and other remote data sources can be used to determine channel and water-surface width and water-surface area, water-surface elevation, channel slope, and channel morphology (planform) (Table 1). The surface velocity of rivers can also be observed remotely using various forms of Doppler radar or lidar (Plant and Keller, 1990; Vörösmarty *et al.*, 1999; W. Emmitt, Simpson Weather Associates, personal communication, 2000; Costa *et al.*, 2000; Goldstein *et al.*, 1994; D. Moller, NASA-JPL, personal communication, 2003) mounted on aircraft; however, measuring surface velocity from satellites has not been demonstrated. The key hydrographic variables that cannot be directly measured from aerial and satellite-based sensors are average depth (Y) and average cross-sectional velocity (U). Thus, these variables will need to be estimated from stage and surface velocity, respectively, as will be described later.

Table 1 Summary of capabilities of remote sensing modes potentially usable for obtaining river-discharge estimates

Mode	Platform	Data Type	Resolution	Limitations
Photography	Aircraft, satellites	Surface features including planform, sinuosity, and so on; bank-full and water-surface width. Stereoscopy can provide slope.	Depends on scale and lens and film properties; can be as fine as 1 m.	Cloud cover
Visible and infrared digital imagery	Aircraft, satellites	Surface features including planform, sinuosity, and so on; bank-full and water-surface width.	Depends on sensor, platform, and orbital characteristics. Aerial imagery can be as fine as 1 m; satellite imagery typically 10 m (LandSat 7) or coarser.	Cloud cover
Synthetic-aperture radar (SAR)	Aircraft, satellites, ground vehicles	Surface features including planform, sinuosity, and so on; bank-full and water-surface width. Interferometry can provide slope. Doppler techniques can provide surface velocity.	Space-based SAR 10 to 30 m; aerial SAR 5 m. Velocity from aerial SAR 0.1 m s^{-1} (not verified).	Interpretation can be difficult (e.g. locating edge of water).
Radar altimetry	Aircraft, satellites	Water-surface elevation at discrete points, giving stage and possibly slope.	Typically 0.5 m, but 0.1 m possible from satellites; higher resolution from aircraft.	Footprint size.
Ground-penetrating radar (GPR)	Ground vehicles, cableways, helicopters	Width and depth.	Comparable to conventional water-contact measurements.	Spatial coverage.
Lidar	Aircraft, satellites	Surface velocity, stage, possibly slope.	Elevation to 0.5 m, velocity to 0.1 m s^{-1} (not verified).	Cloud cover.
Topographic maps, DEMs, GIS	None	Static channel dimensions and morphology; ground slope.	Depends on scale.	No time-varying information.

In a proof-of-concept experiment, Costa *et al.*, (2000) measured river depth and width (cross section) using GPR mounted on a cableway, and surface velocity via a microwave Doppler-radar system mounted on a stationary ground vehicle on the river bank. The mean cross-sectional velocity of flow was estimated from the surface-velocity measurements using a single correction factor of 0.85 (Rantz *et al.*, 1982), and enabled the discharge to be computed from equation (1) to an accuracy comparable to conventional ground-based measurements.

The experiment conducted by Costa *et al.* (2000) demonstrates that ground-based remote data-collection systems can replace in-stream handheld measurement methods without significant loss of accuracy. However, this ground-based measurement system is site-specific and therefore would require periodic maintenance visits similar to conventional gaging stations. The Costa *et al.* (2000) experiment also demonstrated that surface-velocity measurements can be reliably used to estimate the mean velocity. Thus in principle, it should be possible to use aerial or satellite measurements of surface velocity to estimate the mean cross-sectional velocity. However, measuring the mean depth from satellites or aerial platforms remains problematic (see discussion below).

Bankside remote-observation systems provide measurement advantages by potentially increasing the frequency of measurements, especially where access is difficult, and by minimizing the risks to observers by reducing the need for direct in-stream work using conventional handheld instruments or ADCP. In rivers that have difficult accessibility and high costs for obtaining ground-based discharge measurements, satellite and aerial platforms can be used to supplement the ground-based network. In addition, because of the potential for relatively frequent global coverage by

satellites, a satellite-based, discharge-measurement system can provide much-needed understanding of the spatial distribution of discharge across the continents.

Satellite or Aircraft-based Measurement of Hydraulic Variables

Water-surface Area and Width

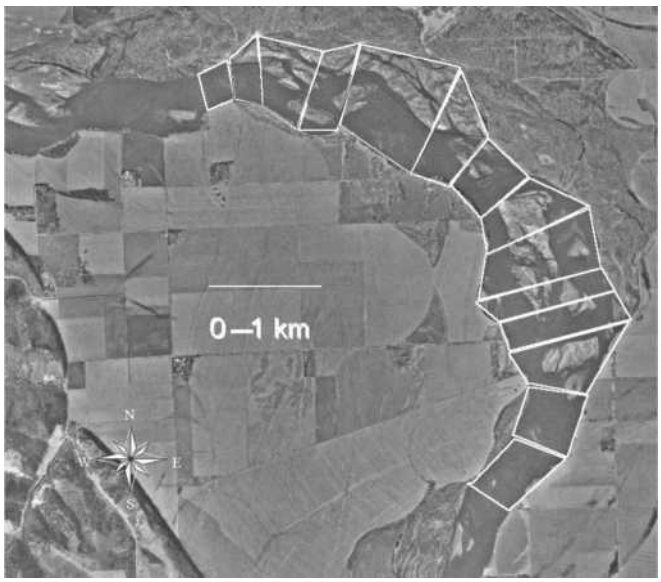
Both bank-full channel water-surface area, A_s^*, and current water-surface area, A_s, of a reach can be measured via a variety of sensors and imagers mounted on satellites and aircraft, including panchromatic and infrared imagers, digital photography, and synthetic-aperture radar (SAR) (Barrett,1998; University of Wisconsin Environmental Remote Sensing Center, 2001). Reach length, L, can also be determined via these sensors, and thus one can compute the reach-average, bank-full width, W^*, and current water-surface width, W, as $W^* = A_s^*/L$ and $W = A_s/L$ respectively. Spatial resolution as fine as 1 or 2 m can be obtained from visible spectrum and near-infrared imagers, and currently deployed SAR imagers have resolutions as fine as 10 m (University of Wisconsin Environmental Remote Sensing Center, 2001). Unlike visible spectrum and infrared imagers, SAR imagers can observe through cloud cover.

The accuracy of measuring surface area and width is only partly a function of the resolution. Measuring along a reach length increases accuracy by averaging errors associated with pixel identification, and also provides a reach-averaged width estimate by averaging local variability in width (Figure 4). Delineating width is also subject to errors associated with vegetation obscuring the water's edge and, in the case of SAR, wet ground, vegetation, wind roughening, and rocks can also confound observation of the water-surface extent. This is a particular problem for the Amazon River, where the water surface extends under forest canopy for a significant part of the year.

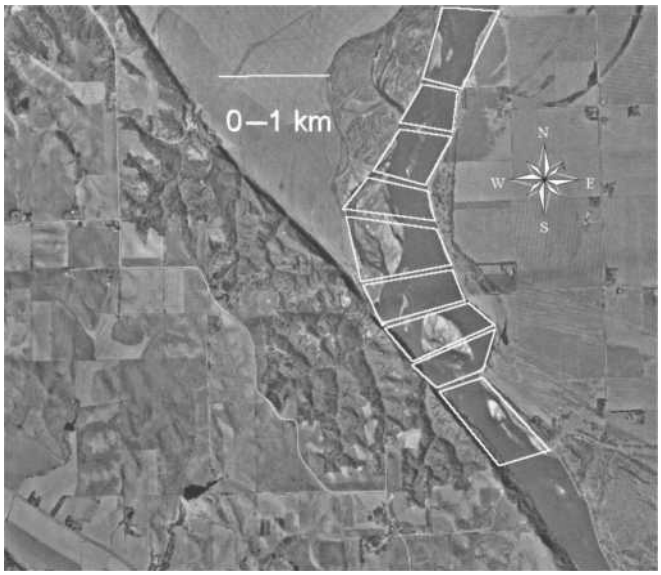
Orbital orientation may also affect the accuracy of satellite observation of water-surface width and area, because the geometric relation between the sensor and the river must be accounted for in image processing, particularly for radar and lidar imagers. This issue could be addressed by linking the processing to geographic information-system (GIS) coverages that contain information about river orientation.

Flood-inundation Area

The water-surface area of flooded and in-bank reaches, used to estimate a mean flow width, can be tracked using a variety of satellite-based sensors including the LandSat MultiSpectral Scanner (MSS), Thematic Mapper (TM), TERRA MODIS, and other visible/infrared spectrum sensors, as well as SAR. MODIS provides daily information, but at relatively coarse resolution (500 m), while TM provides finer spatial resolution but with less frequent overpass; only SAR is capable of observing through clouds.



(a)



(b)

Figure 4 Missouri River near Elk Point South Dakota, showing digitized polygons delineating the maximum channel surface area (Source: 3.75-minute DOQs for Elk Point (a) and Ponca (b) South Dakota, National Digital Orthophoto Program (NDOP))

Measurement of flood-inundation areas from satellite imagery has been summarized by Smith (1997), and such areas are currently measured and tracked globally by G.R. Brakenridge of the Dartmouth Flood observatory (**www.dartmouth.edu/artsci/geog/floods**) using the MODIS sensor mounted on the TERRA satellite. Brakenridge *et al.* (1994) used SAR images from ERS-1 to delineate a flood-inundation area coupled with topographic information to determine water-surface elevations

during the 1993 Mississippi floods. Horritt (2000), Bates and DeRoo (2000), and Horritt *et al.* (2001) have used satellite-mounted SAR imagery to delineate flood boundaries and calibrate river hydraulic models.

Water-Surface Elevation (Stage)

Koblinsky *et al.* (1993) used GEOSAT altimeter data to track elevation changes at several locations in the Amazon River basin with an accuracy on the order of 0.7 m. In that study, the altimeter footprint ranged from 0.2 to 2 km, indicating that the target must be at least that wide to obtain a return unique to the water body. Birkett (1998) measured water-surface elevation changes in several rivers around the globe (including rivers in the Amazon Basin, the Okavango River, the Indus River, and the Congo River) using the TOPEX/POSEIDON altimeter (Figure 5). The accuracy of these measurements ranged from 11 to 60 cm depending on local topography. The theoretical minimum river width that can be observed using the currently deployed TOPEX/POSEIDON altimeter ranges from 0.58 km to 1.16 km (Birkett *et al.*, 2001). However, smaller footprints can be achieved with improved processing algorithms (E. Rodriguez, NASA-JPL, personal communication, 2001). New generation laser altimeters, such as the GLAS altimeter (NASA, 1997) that is deployed on ICESAT, can provide higher resolution and accuracy.

Depth

It is not possible to directly measure water depth using currently available remote sensors. However, it may be possible to obtain useful estimates of average flow depth based on an estimate of bank-full width, W^*, from imagery and successive concurrent observations of water-surface elevation, Z, and water-surface width, W, via altimetry and imagery respectively. This approach is based on the assumption that the channel cross section can be approximated via a power-law relation:

$$Z = Z_0 + 2^b \cdot \left(\frac{Y_m^*}{W^{*b}}\right) \cdot |x|^b \tag{6}$$

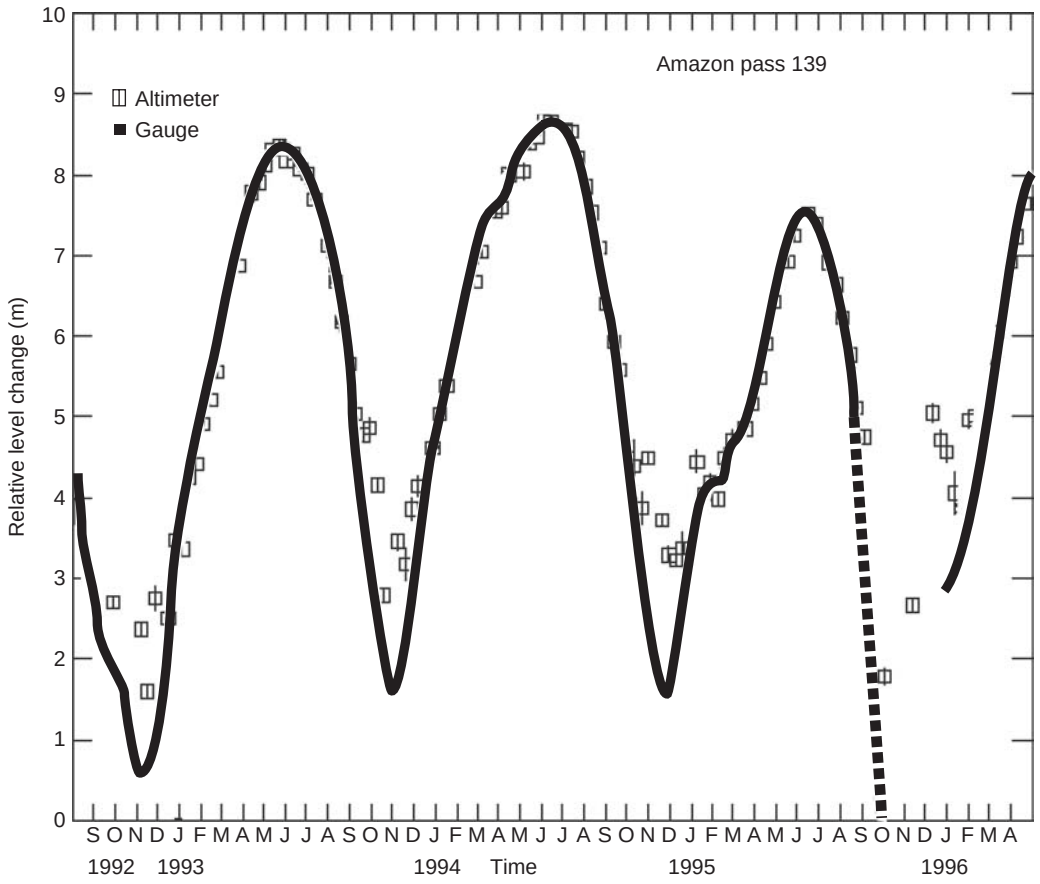


Figure 5 TOPEX/POSEIDON radar-altimetric track for the Amazon River. The time series of relative river water-surface height was derived from two satellite passes over the main branch. Observations are compared against standard river-stage data from a ground-based conventional gauge (Reproduced by permission of C. Birkett from Birkett *et al.*, (2001). *Surface water dynamics in the Amazon Basin: application of satellite radar altimetry.* Presented at 2nd International LBA Scientific Conference, Manaus, Amazonas, Brazil, 7–10 July 2001)

where x is horizontal distance from the channel center, Z_0 is elevation of the lowest (central) point, $Y_m{}^*$ is the bank-full maximum depth, and b is an exponent. Many hydraulic analyses have indicated that a value of $b = 2$ (parabola) closely describes many natural channels (e.g. Leopold *et al.*, 1964), but values as high as $b = 5$ may be appropriate in some cases.

Taking $x = W/2$ and specifying b, the values of Z_0 and $Y_m{}^*/W^*$ can in principle be determined by regression analysis using equation (6) as the regression model. (Alternatively, one could use the model with b unspecified and determine b via nonlinear regression.) Then maximum depth, Y_m, and stage are related as $Y_m = Z - Z_0$, and average depth, Y, is related to maximum depth as

$$Y = \left(\frac{b}{b+1}\right) \cdot Y_m \tag{7}$$

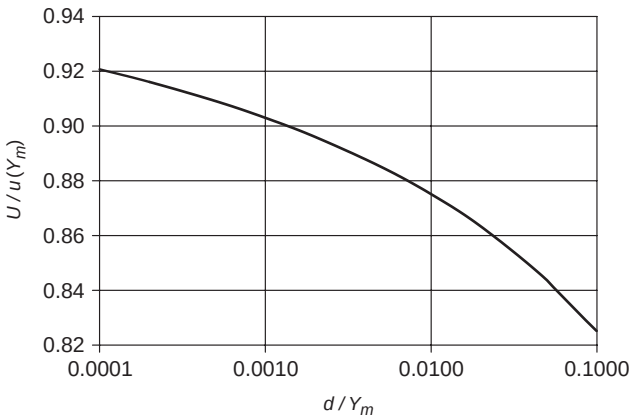


Figure 6 Ratio of average velocity, U, to surface velocity, $u(Y_m)$, as a function of the ratio of bed-material size, d, to flow depth (Y_m) as computed from the Prandtl-von Karman vertical-velocity profile

Velocity

Overall accuracies and observing limitations have not been fully evaluated for remote velocity-measurement techniques. The surface velocity in rivers is potentially measurable from aircraft and satellites with Doppler lidar or radar (Vörösmarty *et al.*, 1999; W. Emmitt, Simpson Weather Associates, personal communication, 2000; Goldstein *et al.*, 1994; D. Moller, NASA-JPL, personal communication, 2003). Wind-generated waves can significantly interfere with the measurement; however, in some circumstances the estimates can be corrected on the basis of the knowledge of the surface wind speed and direction (Kinsman, 1965).

Figure 6 shows the ratio of average vertical velocity to maximum (surface) velocity as a function of the ratio of bed-particle size to flow depth, based on the Prandtl-von Karman vertical-velocity distribution. In most rivers of sufficient size to be remotely observed, the particle size will be small compared to depth, and the average velocity can be estimated as about 0.85 times the surface velocity, as done by Costa *et al.* (2000).

Figure 7 shows estimates of water-surface velocity for a reach of the Missouri River obtained from the aircraft-mounted NASA-JPL AirSAR imager. The estimates were corrected for surface wind using the Bragg correction technique (Bragg, 1913; Kinsman, 1965). The mean channel velocity estimated from this image for those sections of the river flowing towards the sensor was $0.88\,\mathrm{m\,s^{-1}}$ (Moller, NASA, personal communication, 2003). Downstream mean velocity measured by the USGS at a narrower point in the river (with presumably higher velocity) during periods of comparable discharge ranged from 0.96 to $1\,\mathrm{m\,s^{-1}}$. An important limitation of the AirSAR imager is that reliable surface-velocity measurements can only be obtained from river reaches oriented towards the sensor. Thus, although surface velocity can be measured from remote bankside platforms and from aircraft, it is unclear whether this limitation can be overcome to enable the satellite observation of velocity.

Slope

Water-surface slopes over long reaches of the Amazon River have been measured by Mertes *et al.* (1996) and Dunne *et al.* (1998) using altimeters mounted on SEASAT, by Birkett *et al.* (2001) using the TOPEX/POSEIDON altimeter, and by Hendricks *et al.* (2003) using data from the Shuttle Radar Topography Mission (SRTM). While these measurements may be hydraulically meaningful for the Amazon, it is by no means clear that this would be true for smaller rivers because of changes in discharge, hydraulic geometry, and local slopes over the measurement reach.

Measuring hydraulically meaningful water-surface slopes using altimeters would be possible if two water-surface elevations could be obtained at nearly the same time from two closely spaced locations on the river surface. Achieving this measurement objective would require that the altimeter have pointing capability and greater measurement accuracy than the expected 10 to 20-cm accuracy suggested by Birkett (1998). Given a typical river slope in the range of 0.001 to $0.0001\,\mathrm{m\,m^{-1}}$, an observed reach length of 1 to 10 km, and altimeter accuracy of ± 10 to 20 cm, the measured slope could range from negative to several times the actual value. This suggests that slope information obtained from the current generation of altimeters would not provide sufficient spatial resolution to be hydraulically meaningful. Interferometric SAR techniques (Alsdorf *et al.*, 2000; Alsdorf *et al.*, 2001) and laser altimeters may ultimately provide a means to estimate slopes more accurately, but averaging the values obtained from a large sample of water-surface slope measurements in a given reach may be the most meaningful

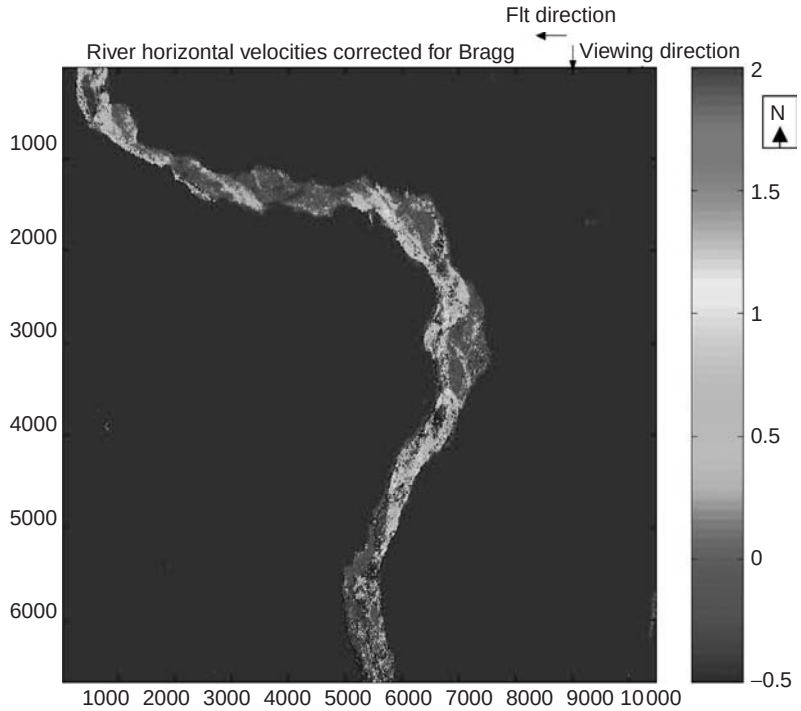


Figure 7 SAR image of the Missouri River near Elk Point South Dakota showing surface-velocity distribution. Scale units are $m\,s^{-1}$ (Reproduced by permission of D. Moller, NASA-JPL). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

remotely sensed slope information that can be considered reliable (Bjerklie *et al.*, 2003).

However, remote measurement of dynamic water-surface slopes may not be necessary. Bjerklie *et al.* (2003) have shown that when estimating discharge with a general slope-resistance equation, the topographic channel slope will provide estimates with accuracy similar to that obtained using a ground-measured dynamic slope over a wide range of rivers. This suggests that a characteristic constant channel slope is hydraulically meaningful for a given river reach.

Channel Morphology

Valley and channel features such as the channel sinuosity and planform, channel slope, meander length, and meander radius of curvature can be observed from a variety of data sources including visible and infrared spectrum images, SAR images, digital elevation models (DEMs), and topographic-map information. Since these features are usually relatively stable over at least decadal time frames, the frequency and timing of observations is not a limiting factor, and therefore high-resolution panchromatic images could be used to measure them when weather conditions permit. As shown in the Smith *et al.* (1996) study discussed in the following section, such information is potentially useful in refining models for remote estimation of discharge.

Remote Estimation of Discharge

Discharge Estimation via Continuity Relation

The previous discussion indicates that there is a possibility that all of the hydraulic elements of equation (1) can be measured or inferred from simultaneous satellite observations. If so, discharge could be calculated directly, with an accuracy dependent on the accuracy and precision of the individual measurements of water-surface width, surface velocity, and stage and of the estimations of mean velocity and mean depth from observations of surface velocity and stage. However, all of these variables, particularly depth, may not be measured or reliably estimated simultaneously. In these situations, general functions that correlate discharge to one or more measured variables would be required.

Discharge Estimation via Single-variable Relations

In principle, it would seem that discharge ratings based on any one of the variables of continuity (width, depth, or velocity) might be developed for a wide range of rivers, because these variables generally increase with increasing discharge (Figure 8). However, because of irregular cross-sectional shape and wide variation in slope, one or more of these variables may change very little or in an unpredictable way with discharge in individual river reaches, so that use of general single-variable ratings would often introduce large estimation errors.

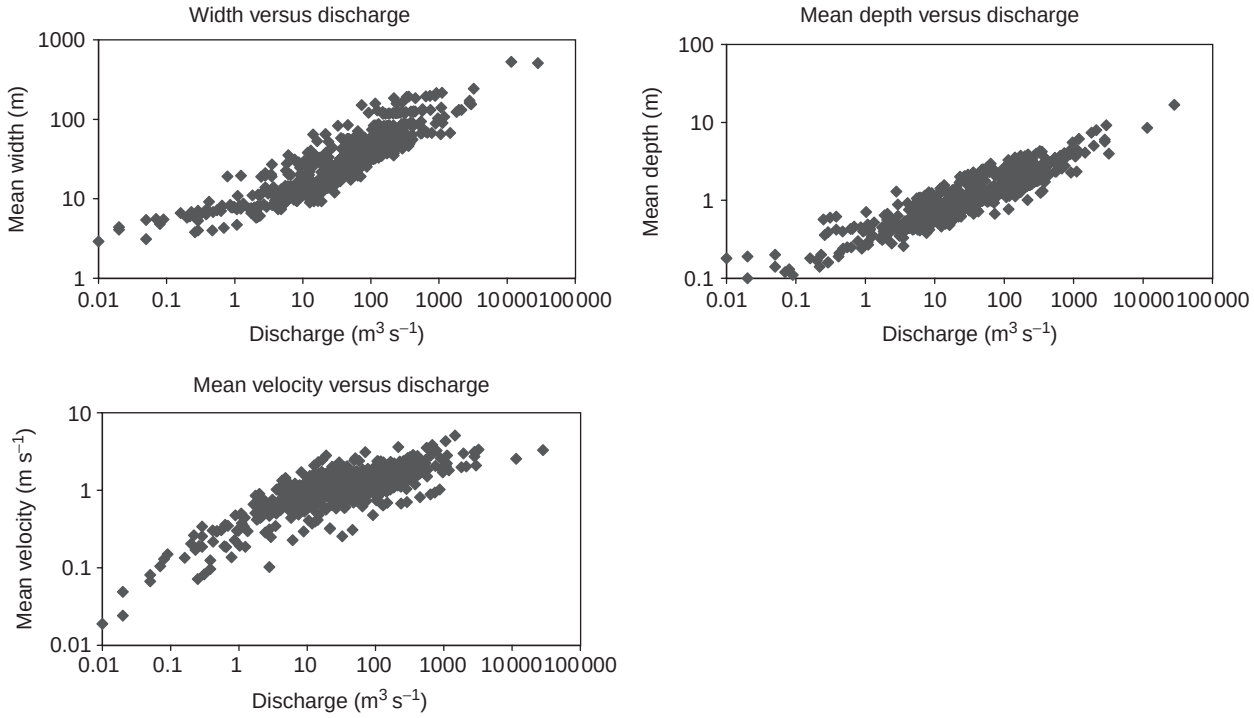


Figure 8 Reach-averaged hydraulic variables versus discharge for 569 velocity-area discharge measurements in 81 rivers

Smith *et al.* (1996) correlated the mean water-surface width, W, estimated from the water-surface area obtained from RADARSAT SAR imagery, with ground-measured discharge in three braided river reaches (lengths on the order of 10 km) to derive power-function discharge ratings. They derived a single best-fit power function for all three rivers, but reported that estimates based on this general rating would be expected to have an accuracy only within a factor of 2. The accuracy of the width-discharge ratings varied between the three rivers, and Smith *et al.* (1996) found that channel sinuosity was a useful additional predictor variable. Figure 9 shows that inclusion of sinuosity, ξ, in a multivariate rating function ($Q = c \cdot W^x \cdot \xi^y$; where c, x, and y are determined by regression) yields a more accurate general predictive equation. This result suggests that multivariate rating functions that include additional hydraulic or morphologic features of a river channel will improve accuracy compared to single-variate functions.

Discharge Estimation via Multivariate Relations

Recent work by Dingman and Sharma (1997) and Bjerklie *et al.* (2003) indicates that statistically derived or general multivariate discharge-estimating equations can be developed and used to estimate discharge for within-bank flows in a wide range of rivers with accuracy comparable to estimates made from slope-area measurements. These equations use the variables of river hydraulics (width, depth, velocity, and slope) in various combinations to estimate discharge. Bjerklie *et al.* (2003) proposed four discharge-estimating equations:

$$Q = k_1 \cdot W^a \cdot Y^{b\cdot} S_c^c \tag{8}$$

$$Q = k_2 \cdot W^d \cdot U^e \cdot S_c^f \tag{9}$$

$$Q = k_3 \cdot W^g \cdot U^h \tag{10}$$

$$Q = k_4 \cdot W^{*i} \cdot Y^{*j} \cdot Y^k \cdot {S_c}^l \tag{11}$$

where S_c is channel slope; W^* and Y^* are the bank-full width and bank-full mean depth respectively; W, Y, and U are as previously defined; and a–l are empirically or theoretically determined exponents. The discharge coefficients k_1 through k_4 reflect components of flow resistance that are not functions of channel adjustments to flow, such as bed material and vegetation (see discussion of equation (5)).

Bjerklie *et al.* (2003) derived values for the coefficients and exponents in equations (8)–(11) via regression analysis. Figure 10 shows discharge predicted from these equations compared with observed discharge for 506 individual flows in 102 rivers with a maximum (bank-full) width

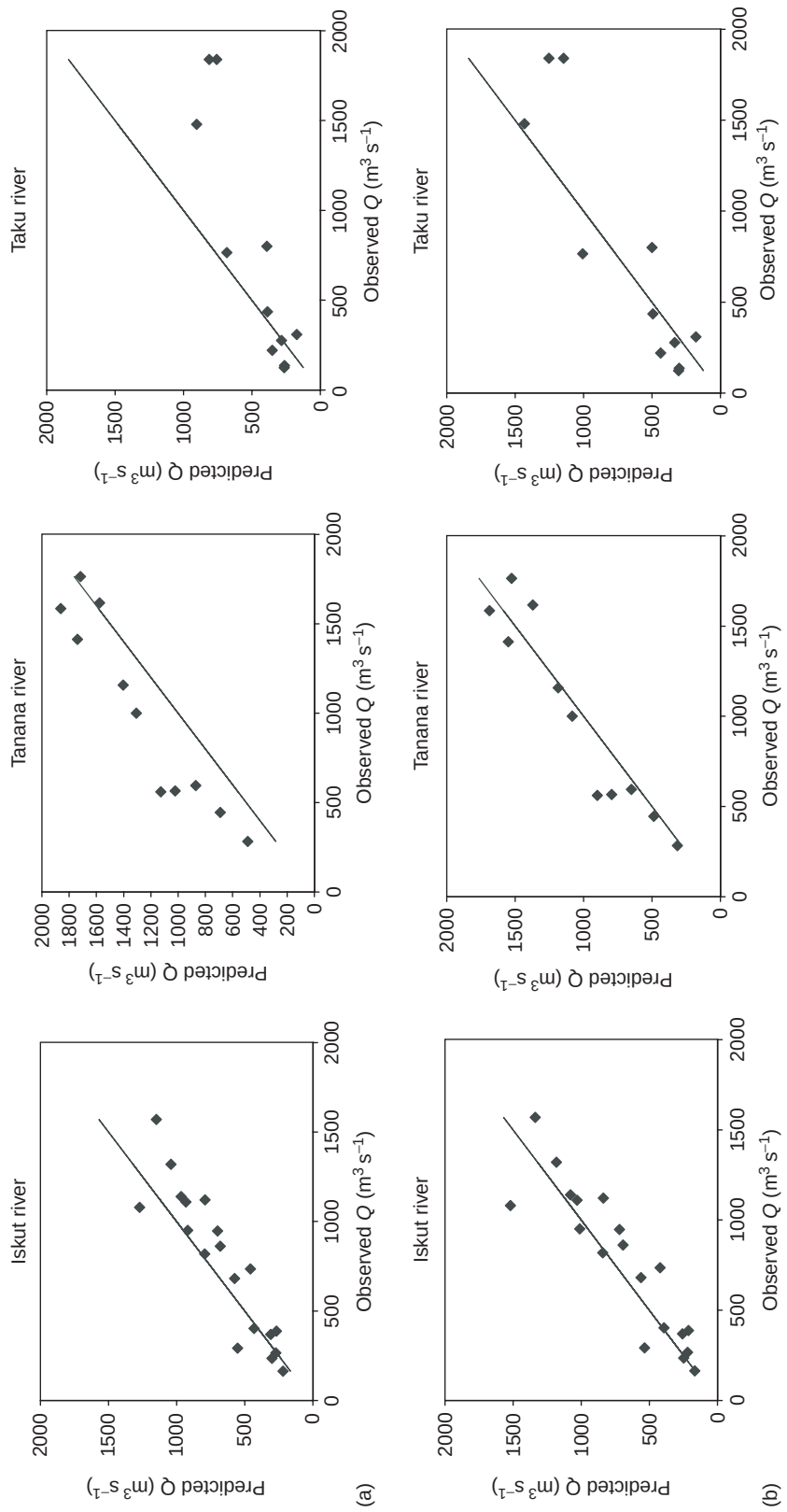


Figure 9 (a) Discharge predicted from remotely observed width versus observed discharge for the Iskut, Taku, and Tanana Rivers, Alaska. Discharge is poorly predicted for all three rivers. (b) Discharge predicted from remotely observed width and sinuosity versus observed discharge for the Iskut, Taku, and Tanana Rivers, Alaska. Discharge predictions cluster more generally around the 1 : 1 line than in (a), indicating that inclusion of sinuosity improves the predictive relation. (Data from Smith *et al.* (1996))

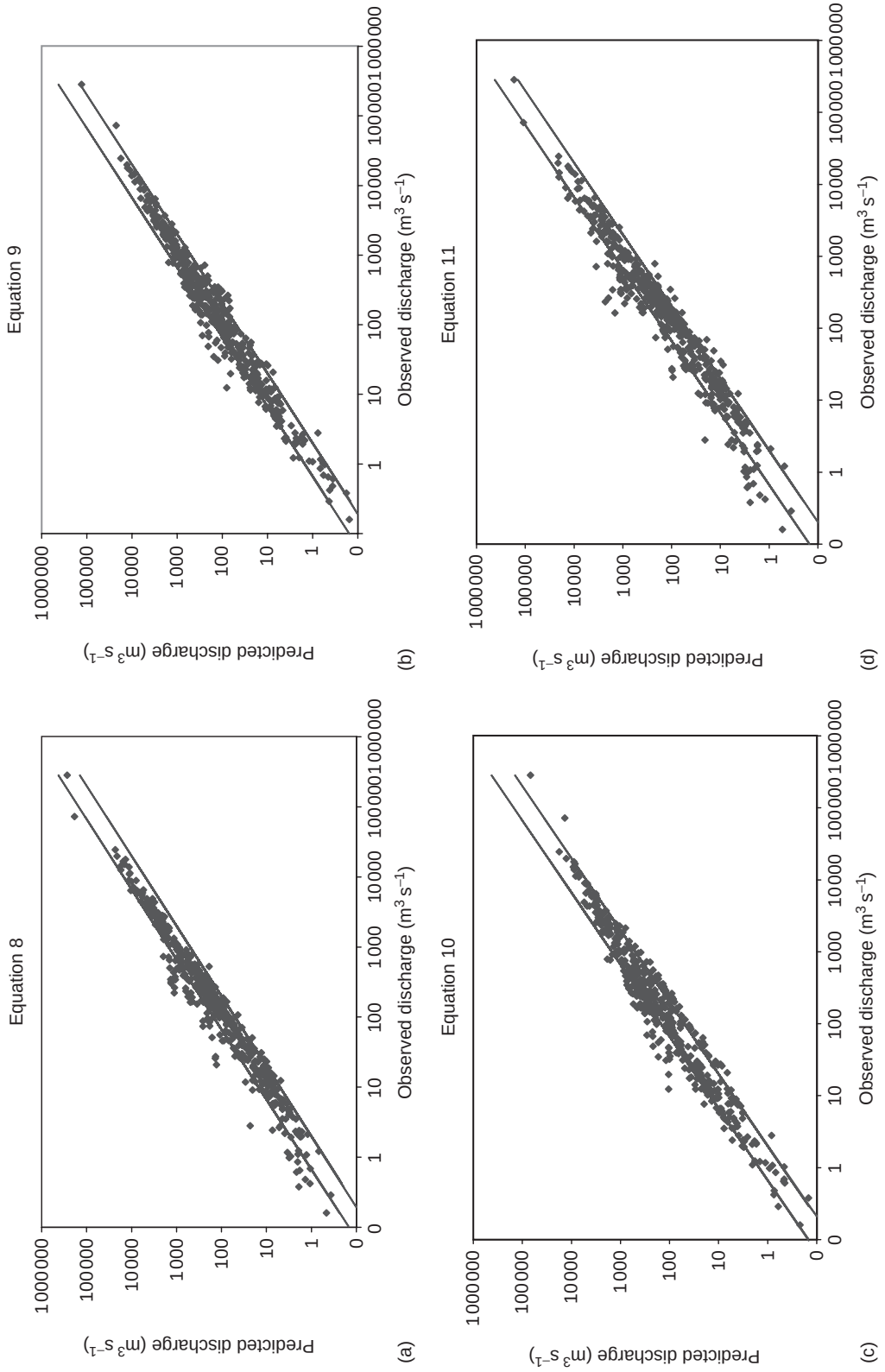


Figure 10 Discharge predicted from regression-derived forms of (a) equation (8); (b) equation (9); (c) equation (10); and (d) equation (11). The upper and lower lines are the plus- and minus-50th-percentile error bars, respectively. (From Bjerklie (2004))

exceeding 30 m, including several of the world's largest rivers (e.g. the Amazon, Yukon, and Mississippi rivers). The accuracy of the four equations varied, but all four produced discharge estimates that were within ±50% of the observed values at least 70% of the time. These results suggest that relations of the form of equations (8)–(11) can provide generally applicable predictions with useful accuracy.

Bjerklie *et al.* (2003) also examined the effects of measurement uncertainty of the hydraulic variables in equations (8)–(11) on discharge-prediction accuracy. They found that prediction error is greatest at lower discharges due to the proportionally larger effect of measurement error and because the equations have larger statistical uncertainty at lower discharges. Thus, generalized prediction equations such as (8)–(11) will tend to be more accurate for larger rivers.

The results of Bjerklie *et al.* (2003) suggest that remotely measured hydraulic variables have the potential to provide useful estimates of discharge for within-bank flows in moderate-sized to large rivers. Equations (8), (9), and (11) require information on the (constant) local channel slope, which can generally be determined from topographic maps and/or DEMs. Use of equations (9) and (10) assumes that surface-velocity measurements can be used to estimate mean velocity. Equations (8) and (11) can be used if the depth of flow can be estimated from a remotely measured stage as discussed earlier (equations 6 and 7), and equation (11) requires information on bank-full width and average depth that may be obtainable from repeated imagery or altimetry at a given reach.

For overbank flows, the relationships analogous to equations (8)–(11) may not exist. Jacobs and Wang (2003) have suggested that the resistance of floodplain vegetation and other obstructions to the flow can be evaluated remotely from ground-cover characteristics in the same manner that surface roughness is estimated for atmospheric modeling. Reliable estimates of floodplain resistance would allow use of traditional resistance equations such as the Manning equation (equation 5 or calibration of prediction equations such as (8)–(11) to specific floodplain characteristics). With this approach, the discharge in inundated areas could also be estimated if topographic information was used to determine floodplain depths, provided areas of stagnant water or nondownstream flow could be identified from surface-velocity observations (Brakenridge *et al.*, 1994).

Discharge Estimation via Remotely Established Rating Relations

As discussed previously, assuming a regular geometric channel shape coupled with an observational history of width and stage provides the basis for converting stage to channel depth (equations 6 and 7). An alternative approach to utilizing stage information to improve prediction accuracy is to predict discharge from remote observations of width, velocity, and slope using equations of the form of (9) and (10), and observe stage remotely via altimetry. One can then develop a stage-discharge rating equation using the observed stage and the model-estimated discharge. Such a rating would average out much of the predictive error and, once the rating was established, enable stage to be the primary discharge-tracking variable.

We have conducted a proof-of-concept test of this approach using data for the USGS gage on the Mississippi River at Thebes, Illinois. In this test we assumed that the ground-measured widths and velocities are representative of measurements made from aircraft- or satellite-mounted sensors and that the ground-measured stages are representative of satellite-altimeter measurements. Rating relations are then established between the observed stage and the discharge estimated from generalized versions of equations (9) and (10). Figure 11 compares these ratings with the USGS rating relation for the Thebes gage. The estimated discharge is taken as the average of the values given by the two equations, and the estimated and observed discharges are compared in Figure 12. The mean relative error of the estimate is 0.01 (+1%) and the standard deviation of the relative error is 0.11 (11%) (i.e. approximately 95% of the estimates are within 22% of the true values).

Frequency of Satellite Observation

As noted earlier, the practical value of remotely sensed observations of river discharge depends on the spatial and temporal frequencies of observation as well as the precision with which predictor variables can be observed and the errors associated with prediction models. Temporal observation frequencies are difficult to generalize because they are functions of the orbital characteristics of satellites, the frequency of signal capture, whether the sensors can see through clouds, and the presence of ice cover. However, general estimates of the spatial frequency of observation can be based on simple quantitative characterizations of the planar aspects of stream networks.

Given a region of area A_D, the drainage density, D, is defined as $D \equiv \Sigma L / A_D$, where ΣL is the total length of streams in the region. From simple geometry, the characteristic (minimum) distance between streams, X, is the inverse of the drainage density. Referring to Figure 13, a random straight-line path crosses adjacent streams at a random angle θ, so the pathwise distance between stream crossings, Δx, is given by

$$\Delta x = \frac{X}{\cos\theta} = \frac{1}{D \cdot \cos\theta} \tag{12}$$

Assuming that all values of θ ($0 \leq \theta \leq \pi/2$) are equally likely and that D has a characteristic fixed value in a region, Sellmann and Dingman (1970) showed that the average

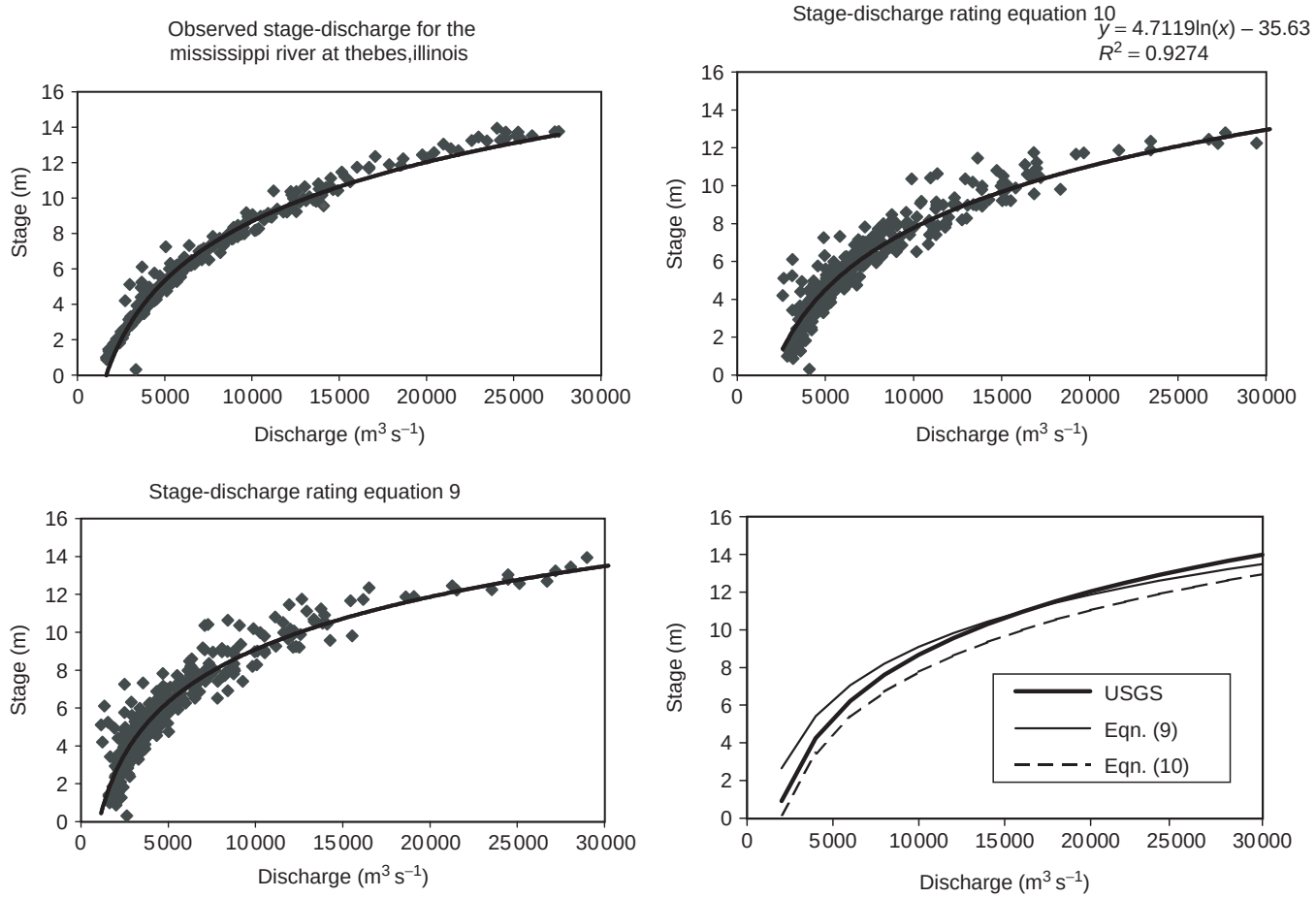


Figure 11 Stage-discharge ratings at the USGS gaging station for the Mississippi River at Thebes, IL, using (a) USGS ground observations of stage and USGS velocity-area measurements of discharge; (b) USGS ground observations of stage and discharge predicted from generalized version of equation (9); (c) USGS ground observations of stage and discharge predicted from generalized version of equation (10); and (d) comparison of rating curves

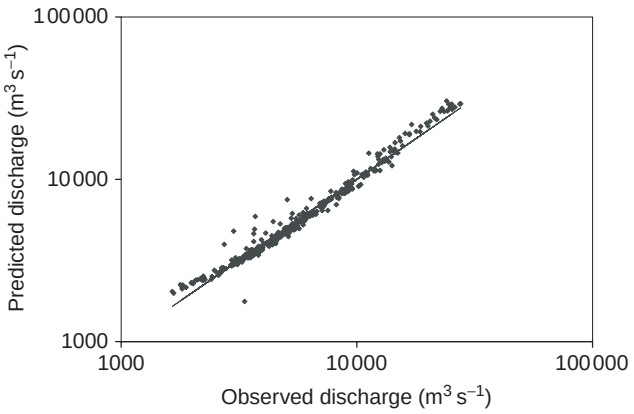


Figure 12 Discharge predicted as average of values predicted from equations (9) and (10) versus velocity-area measurements of discharge at the USGS gaging station for the Mississippi River at Thebes, IL

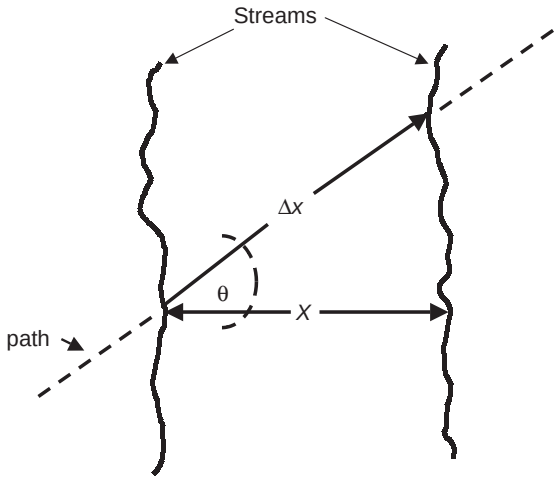


Figure 13 Definitions of terms in equation (12)

number of streams encountered per unit path length, $E(n)$, is given by

$$E(n) = \left(\frac{2}{\pi}\right) \cdot D = 0.637 \cdot D \tag{13}$$

The bifurcation ratio, R_B, in a stream network is the ratio of the number of streams of order ω, $N(\omega)$, to the number of streams of order $\omega + 1$, $N(\omega + 1)$:

$$R_B \equiv \frac{N(\omega)}{N(\omega + 1)} \tag{14}$$

and the length ratio, R_L, is the ratio of the average length of streams of order $\omega + 1$, $L(\omega + 1)$, to the average length of streams of order ω, $L(\omega)$:

$$R_L \equiv \frac{L(\omega + 1)}{L(\omega)} \tag{15}$$

Defining $R \equiv R_B/R_L$, Rodríguez-Iturbe and Rinaldo (1997) showed that the total length of streams in a drainage basin of order Ω, ΣL, is

$$\Sigma L = L(\Omega) \cdot \frac{1 - R^{\Omega}}{1 - R} \tag{16}$$

where $L(\Omega)$ is the length of the main stream. Values of R_B are generally between 3 and 5 and average about 4, and values of R_L are typically $\approx$2 (Rodríguez-Iturbe and Rinaldo, 1997).

Assuming a nearly constant relation between stream size (e.g. width) and stream order in a region, a given remote observing system will be able to extract useful information from streams of order greater than or equal to a "minimum observable order", ω^*. It can be shown that the total length of streams of this or larger order, $\Sigma L(\omega^*)$, is

$$\Sigma L(\omega^*) = L(\Omega) \cdot \frac{1 - R^{\Omega - \omega^* + 1}}{1 - R} \tag{17}$$

Thus in any path in a basin of order Ω, the frequency of encountering a stream large enough to observe, $f(\omega^*)$, is

$$f(\omega^*) = \frac{\Sigma L(\omega^*)}{\Sigma L(\Omega)} = \frac{1 - R^{\Omega - \omega^* + 1}}{1 - R^{\Omega}} \tag{18}$$

Finally, the average number of observable streams encountered per unit path length, $E(n^*)$, is given by the product of equations (13) and (18) as:

$$E(n^*) = 0.637 \cdot f(\omega^*) \cdot D \tag{19}$$

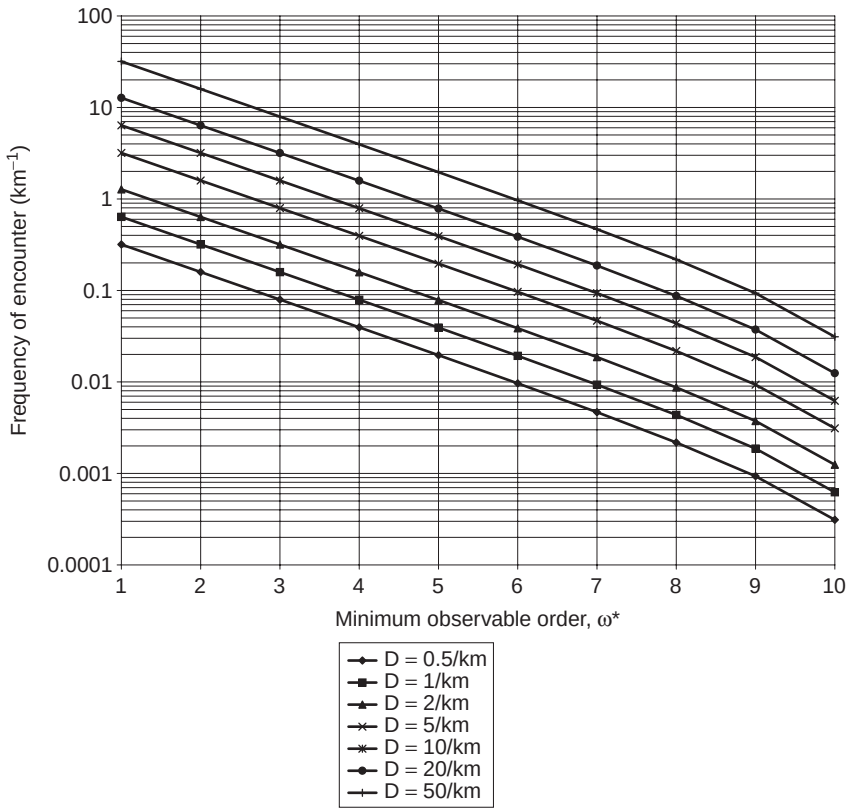


Figure 14 Relations between frequency of stream encounters and minimum observable order, ω^*, as a function of drainage density, D, assuming $R = 2$ and $\Omega = 10$ (equation 19)

Figure 14 shows $E(n^*)$ as a function of D assuming $R = 2$ and $\Omega = 10$. For example, in a basin with $D = 2\,\mathrm{km}^{-1}$ and a system that can observe streams of order $\omega^* = 6$ and larger, we find $f(\omega^*) = 0.030$ and $E(n^*) = 0.039\,\mathrm{km}^{-1}$, or about 4 streams per 100 km of observation path.

Summary and Discussion

Numerous studies have demonstrated the use of remotely sensed hydraulic data in tracking dynamic changes in river-flow conditions. However, few have used remotely sensed information to quantitatively estimate river discharge. One reason for this is that concurrent data sets of remotely sensed width, slope, and either stage or velocity are not generally available at present.

As discussed in this article, analyses using conventional ground-measured data indicate the feasibility of several approaches to developing quantitative discharge estimates using hydraulic data measured from remote platforms. These approaches should be developed further and their accuracy and precision tested as such data sets become more available. Further experience in the remote measurement of river hydraulic variables and the use of hydraulic relations to estimate discharge from these data will probably result in robust estimating methods.

Aircraft- and satellite-based measurement systems specifically designed to monitor rivers and estimate discharge are potentially cost-effective when large areal coverage is required. In the large portions of the world where ground-based measurements are not feasible for political or economic reasons, such systems appear to be the only option for obtaining quantitative river-discharge information.

The needed accuracy and precision of discharge measurements are functions of the intended use, and would be expected to improve over time as large data sets are developed and calibration methods improve. However, because of inherent uncertainties in remote observation and inherent limitations of resolution, it is unlikely that remote measurements of river discharge can replace the need for ground-based measurements for many water-resource assessment needs, particularly in developed countries where ground-based networks already exist. Although the cost of a satellite or aerial measurement system would be relatively high compared to ground-based systems, the large coverage that could be obtained could ultimately contribute significantly to our knowledge of regional and global hydrology and water resources.

Acknowledgments

The authors' research on remote sensing of river discharge has been supported in part by NASA grants NAG5-7601 and NAG5-8683. We thank Charon Birkett (NASA Goddard Space Flight Center) and Delwyn Moller (California Institute of Technology Jet Propulsion Laboratory) for providing information from their current research, and Charles Vörösmarty (Institute for the Study of Earth, Oceans, and Space, University of New Hampshire) for guidance, insight, and ideas that contributed to this article. Comments by Dennis Lettenmaier (Department of Civil Engineering, University of Washington) on earlier drafts of this article were most helpful.

REFERENCES

Alsdorf D.E., Melack J.M., Dunne T., Mertes L.A.K., Hess L.L. and Smith L.C. (2000) Interferometric radar measurements of water level changes on the Amazon River flood plain. *Nature*, **404**, 174–177.

Alsdorf D.E., Smith L.C. and Melack J.M. (2001) Amazon floodplain water level changes measured with interferometric SIR – C radar. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 423–431.

Barnes H.B. (1967) *Roughness characteristics of natural channels*, U.S. Geological Survey: Water-Supply Paper 1849.

Barrett E. (1998) Satellite remote sensing in hydrometry. In *Hydrometry: Principles and Practices*, *Second Edition*, Herschy R.W. (Ed.) John Wiley & Sons: Chichester, pp. 199–224.

Bates P.D. and DeRoo A.P.J. (2000) A simple raster – based model for flood inundation simulation. *Journal of Hydrology*, **236**, 54–77.

Birkett C.M. (1998) Contribution of the TOPEX NASA radar altimeter to the global monitoring of large rivers and wetlands. *Water Resources Research*, **34**, 1223–1239.

Birkett C.M., Mertes L.A.K., Dunne T., Costa M.H. and Jasinski M.J. (2001) Surface water dynamics in the Amazon Basin: application of satellite radar altimetry. *Presented at 2nd International LBA Scientific Conference*, Manaus, 7–10 July 2001.

Bjerklie D.M. (2004) *Development of Hydraulic Relationships for Estimating In – Bank River Discharge Using Remotely Sensed Data*, Ph.D. Thesis, University of new Hampshire.

Bjerklie D.M., Dingman S.L., Vörösmarty C.J., Bolster C.H. and Congalton R.G. (2003) Evaluating the potential for measuring river discharge from space. *Journal of Hydrology*, **278**, 17–38.

Bragg W.L. (1913) The diffraction of short electromagnetic waves by a crystal. *Proceedings of the Cambridge Philosophical Society*, **17**, 43.

Brakenridge G.R., Knox J.C., Paylor E.D. and Magilligan F.J. (1994) Radar remote sensing aids study of the Great Flood of 1993. *Eos, Transactions American Geophysical Union*, **75**, 521–527.

Brakenridge G.R., Tracy B.T. and Knox J.C. (1998) Orbital SAR remote sensing of a river flood wave. *Journal of Remote Sensing*, **19**, 1439–1445.

Buchanan T.J. and Somers W.P. (1969) *Discharge Measurement at Gaging Stations*, U.S. Geological Survey: Techniques of Water-Resources Investigations Book 3 Chapter A8.

Carter R.W. and Anderson I.E. (1963) Accuracy of current meter measurements. *American Society of Civil Engineers Proceedings, Journal of the Hydraulics Division*, **89**, 105–115.

Chow V.T. (1959) *Open-Channel Hydraulics*, McGraw-Hill: New York.

Coon W.F. (1997) *Estimation of Roughness Coefficients for Natural Stream Channels with Vegetated Banks*, U.S. Geological Survey: Water-Supply Paper 2441.

Costa J.E., Spicer K.R., Cheng R.T., Haeni F.P., Melcher N.B. and Thurman E.M. (2000) Measuring stream discharge by non-contact methods: A proof-of-concept experiment. *Geophysical Research Letters*, **27**, 553–556.

Cowan W.L. (1956) Estimating hydraulic roughness coefficients. *Agricultural Engineering*, **37**, 473–475.

Dalrymple T. and Benson M.A. (1967) *Measurement of peak discharge by the slope-area method*, U.S. Geological Survey: Techniques of Water Resources Book 3 Chapter A2.

Dickinson W.T. (1967) *Accuracy of Discharge Determinations*, Colorado State University: Hydrology Papers No. 20.

Dingman S.L. and Sharma K.P. (1997) Statistical development and validation of discharge equations for natural channels. *Journal of Hydrology*, **199**, 13–35.

Dingman S.L. (2002) *Physical Hydrology*, *Second Edition*, Prentice-Hall: Upper Saddle River, NJ.

Dunne T., Mertes L.A.K., Meade R.H., Richey J.E. and Forsberg B.R. (1998) Exchanges of sediment between the floodplain and channel of the Amazon River in Brazil. *Geological Society of America Bulletin*, **110**, 450–467.

Fekete B.M., Vörösmarty C.J. and Grabs W. (1999) *Global Composite Runoff Fields Based on Observed River Discharge and Simulated Water Balance*, Report 22, WMO Global Runoff Data Center.

Goldstein R.M., Barnett T.P. and Zebker H.A. (1994) Remote sensing of ocean currents. *Science*, **246**, 1282–1285.

Hendricks G.A., Alsdorf D.E., Pavelsky T.M. and Sheng Y. (2003) Channel slope from SRTM water-surface elevations in the Amazon Basin. *Poster Presented at Fall 2003 American Geophysical Union Meetings*.

Herschy R.W. (1995) *Streamflow Measurement*, Chapman & Hall: London.

Herschy R.W. (1999a) Flow measurement. In *Hydrometry: Principles and Practices*, *Second Edition*, Herschy R.W. (Ed.) John Wiley & Sons: Chichester, pp. 9–83.

Herschy R.W. (1999b) Hydrometric instruments. In *Hydrometry: Principles and Practices*, *Second Edition*, Herschy R.W. (Ed.) John Wiley & Sons: Chichester, pp. 85–142.

Hicks D.M. and Mason P.J. (1991) *Roughness characteristics of New Zealand rivers*, New Zealand Water Resources Survey, DSIR Marine and Freshwater: Wellington, pp. 329.

Horritt M.S. (2000) Calibration of a two-dimensional finite element flood flow model using satellite radar imagery. *Water Resources Research*, **36**, 3279–3291.

Horritt M.S., Mason D.C. and Luckman A.J. (2001) Flood boundary delineations from synthetic aperture radar imagery using a statistical active contour model. *Journal of Remote Sensing*, **22**, 2489–2507.

IAHS Ad Hoc Group on Global Water Data Sets (2001) Global water data: a newly endangered species. *Eos, Transactions American Geophysical Union*, **82**, 54–58.

Jacobs J.M. and Wang M. (2003) Atmospheric momentum roughness applied to stage-discharge relationships in flood plains. *Journal of Hydrologic Engineering*, **8**, 99–104.

Karl T.R. and Riebsame W.E. (1989) The impact of decadal fluctuations in mean precipitation and temperature on runoff.: a sensitivity study over the United States. *Climatic Change*, **15**, 423–447.

Kinsman B. (1965) *Wind Waves*, Prentice-Hall: Englewood Cliffs.

Kirby W.H. (1987) Linear error analysis of slope-area discharge determinations. *Journal of Hydrology*, **96**, 125–138.

Koblinsky C.J., Clarke R.T., Brenner A.C. and Frey H. (1993) Measurement of river level with satellite altimetry. *Water Resources Research*, **29**, 1839–1848.

Kuprianov V.V. (1978) Aerial methods of measuring river flow. In *Hydrometry: Principles and Practices*, *First Edition*, Herschy R.W. (Ed.) John Wiley & Sons, Chichester, pp. 473–478.

Leopold L.B., Wolman M.G. and Miller J.P. (1964) *Fluvial Processes in Geomorphology*, W.H. Freeman: San Francisco.

Mertes L.A.K., Dunne T. and Martinelli L.A. (1996) Channel-floodplain geomorphology along the Solimes-Amazon River, Brazil. *Geological Society of America Bulletin*, **108**, 1089–1107.

Morlock S.E. (1996) *Evaluation of Acoustic Doppler Current Profiler Measurements of River Discharge*, Water-Resources Investigations Report 95–4218, U.S. Geological Survey.

NASA Earth Observing System GLAS Science Team (1997) *Geoscience laser altimeter system (GLAS) science requirements*, U.S. National Aeronautics and Space Administration ICES-UTA-SPEC-001, GLAS-UTA-REQ-001.

Plant W.J. and Keller W.C. (1990) Evidence of Bragg scattering in microwave Doppler spectra of sea return. *Journal of Geophysical Research*, **95**, 16299–16310.

Ponce V.M. (1989) *Engineering Hydrology*, Prentice-Hall: Englewood Cliffs.

Rantz S.E. *et al.* others (1982) *Measurement and Computation of Streamflow: Volume 1 Measurement of Stage and Discharge*, U.S. Geological Survey: Water Supply Paper 2175, p. 284.

Rodda J.C. (1999) Measuring up to water resources assessment. In *Hydrometry: Principles and Practices*, *Second Edition*, Herschy R.W. (Ed.) John Wiley & Sons, Chichester, pp. 142–160.

Rodríguez-Iturbe I. and Rinaldo A. (1997) *Fractal River Basins: Chance and Self-Organization*, Cambridge University Press: Cambridge.

Sellmann P.V. and Dingman S.L. (1970) Prediction of stream frequency from maps. *Journal of Terramechanics*, **7**, 101–115.

Simpson M.R. and Oltman R.N. (1992) *Discharge-Measurement System Using an Acoustic Doppler Current Profiler with Applications to Large Rivers and Estuaries*, Open-File Report 91–487, U.S. Geological Survey.

Smith L.C. (1997) Satellite remote sensing of river inundation area, stage, and discharge: a review. *Hydrological Processes*, **11**, 1427–1439.

Smith L.C., Isacks B.L., Bloom A.L. and Murray A.B. (1996) Estimation of discharge from three braided rivers using synthetic aperture radar satellite imagery. *Water Resources Research*, **32**, 2021–2034.

University of Wisconsin Environmental Remote Sensing Center (2001) Resources: Satellites **http://www.ersc.wisc.edu/resources/php**, **http://www.ersc.wisc.edu/resources/index.html**.

van der Vink G., *et al.* (2004) Why the United States is becoming more vulnerable to natural disasters. **http://www.agu.org/sci_soc/articles/eisvink.html**.

Vörösmarty C., Birkett C., Dingman S.L., Lettenmaier D.P., Kim Y., Rodriguez E., Emmitt G.D., Plant W. and Wood E.S. (1999) NASA post-2002 land surface hydrology mission component for surface water monitoring, HYDRA - SAT. *Report from the NASA Post 2002 LSHP Planning Workshop*, Irvine, 12–14 April, p. 53.

Vörösmarty C.J., Green P., Salisbury J. and Lammers R.B. (2000) Global water resources: vulnerability from climate change and population growth. *Science*, **289**, 281–288.

Walling D.E. and Webb B.W. (1987) Material transport by the world's rivers: Evolving perspectives. In *Water for the Future: Hydrology in Perspective*, International Association for Scientific Hydrology Publication 164, International Association for Scientific Hydrology.

White K.E. (1978) Dilution methods. In *Hydrometry: Principles and Practices*, *First Edition*, Wiley: Chichester.

Wigley T.M.L. and Jones P.D. (1985) Influence of precipitation changes and direct CO_2 effects on streamflow. *Nature*, **314**, 149–151.

Winter T.C. (1981) Uncertainties in estimating the water balance of lakes. *Water Resources Bulletin*, **17**, 82–115.

62: Estimation of Suspended Sediment and Algae in Water Bodies

JERRY C RITCHIE[1] AND PAUL V ZIMBA[2]

[1] United States Department of Agriculture, Agricultural Research Service Hydrology and Remote Sensing Laboratory, Beltsville, MD, US
[2] United States Department of Agriculture, Agricultural Research Service Catfish Genetics Research Unit, Stoneville, MS, US

Remote sensing techniques can be use to monitor water quality parameters (i.e., suspended sediments-turbidity, chlorophyll, temperature). Optical and thermal sensors on shoreline booms, boats, aircraft, and satellites provide both spatial and temporal information needed to monitor changes in water quality parameters for developing management practices to improve water quality. Recent and planned launches of platforms equipped with improved spectral and spatial resolution sensors should lead to greater use of remote sensing techniques to assess and monitor water quality parameters. Integration of remotely sensed data, GPS, and GIS technologies provides a valuable tool for monitoring and assessing waterways. Remotely sensed data can be used to create a permanent geographically located database to provide a baseline for future comparisons. The integrated use of remotely sensed data, GPS, and GIS will enable consultants and natural resource managers to develop management plans for a variety of natural resource management applications. This review paper identifies current knowledge regarding the assessment of water quality, and identifies promising techniques for high resolution assessment.

INTRODUCTION

Current concerns for the effects of global change and sustainable development of the global environment have overshadowed what many see as the equally important problems associated with the progressive degradation of the global soil resources and the impact of increased sediment and nutrient mobilization on fluvial and aquatic ecosystems (Lal, 1994, 2001; Walling, 1999; Ritchie *et al.*, 2003a). Changes in water quality (*see* **Chapter 91, Water Quality, Volume 3** and **Chapter 92, Water Quality Monitoring, Volume 3**) are the most visible manifestation of this progressive degradation. Water quality is most often defined in terms of physical, chemical, thermal, and/or biological characteristics that affect the properties of water. Good or poor water quality depends on the end use of the water; however, there is general agreement that surface water quality has deteriorated in recent decades. Substances from natural and/or anthropogenic sources have contributed to deterioration of water quality in freshwater and estuarine ecosystems of the world (Dekker *et al.*, 1995). The impacts of these substances include water pollution, reservoir sedimentation, degradation of aquatic habitats, alterations of trophic levels in aquatic food chains, and increased cost of water treatment. Globally, the economic cost of the on-site and off-site impacts of pollutant movement has been estimated at ca. $400 billion per year (Pimentel *et al.*, 1995).

Rapid techniques for monitoring and assessing water quality are critical if surface waters are to be managed to maintain or improve their quality. *In situ* measurements provide accurate information for a point in time and space, but these measurements are difficult, expensive, and often inaccurate for understanding either the spatial or temporal patterns of water quality needed for accurate assessment or management of water bodies (Curran and Nova, 1988; Boyer *et al.*, 2002). These limitations of *in situ* measurement techniques to provide information on the spatial and temporal patterns of water degradation across the landscape restrict our ability to develop cost-effective land management strategies. Such limitations point to the need

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

to investigate other techniques to supplement current techniques for monitoring and quantifying water quality patterns in space and time. Assessment of the magnitude of these problems and the formulation of effective management and control strategies require more reliable data, an improved understanding of the spatial and temporal patterns involved, and faster retrieval of information. Recent reviews have emphasized the potential of remote sensing techniques to monitor water quality (Curran and Novo, 1988; Cracknell, 1999; Ritchie, 2000; Ritchie and Schiebe, 2000; Ritchie *et al.*, 2003b). The purpose of this chapter is to provide an overview of the application of remote sensing techniques and sensors for monitoring and assessing suspended sediments and chlorophyll (algae) components of water quality. These are certainly the most visible manifestations of water quality.

BACKGROUND

Electromagnetic radiation from the Sun is either transmitted into the water or reflected (backscattered) from the water surface. Remote sensing techniques for monitoring water quality are based on the ability to quantitatively measure changes in quality and quantity of the backscattered or emergent radiation characteristics of the surface water (Jerlov, 1976; Kirk, 1983) and to relate these measured changes by empirical or analytical models to the water quality parameter of interest. The optimal sensor/wavelength used to measure a water quality parameter depends on the substance being measured, its concentration, and the sensor characteristics.

Suspended sediments (turbidity), colour (i.e. chlorophylls, carotenoids), chemicals (i.e. nutrients, pesticides, metals), dissolved organic matter (DOM), thermal releases, aquatic organisms including zooplankton, emergent and submergent plants, bacteria, viruses, and oils are key factors affecting water quality. Remote sensing techniques can measure the energy spectra of reflected solar and/or emitted radiation from surface waters caused by suspended sediments, chlorophylls, DOM, oils, aquatic plants, and thermal releases. Most chemicals do not directly change the spectral or thermal characteristics of surface waters so they can only be inferred indirectly from measurements of other water quality parameters (most often chlorophyll) that are affected by these substances. Remote sensing techniques provide spatial and temporal data on surface water parameters thus making it possible to monitor the landscape effectively and efficiently, identifying and quantifying water quality parameters and problems.

Research using remote sensing techniques to measure surface water properties began in the early 1970s. This research used sensors in the laboratory and on boats, airplanes, and satellites to make measurements of the spectral backscattering characteristics of water surfaces. Empirical relationships between spectral properties and *in situ* water quality parameters were established. Ritchie *et al.* (1974) using *in situ* measurements from boats and others (i.e. Klemas *et al.*, 1975; Bowker and Witte, 1975; Johnson, 1976) using aircraft and satellite sensors developed empirical approaches to estimate suspended sediments and chlorophyll from spectral measurements. The general forms of these empirical equations are:

$$Y = a + bX \quad \text{or} \quad Y = ab^X \tag{1}$$

where Y is the remote sensing measurement (i.e. radiance, reflectance, energy) and X is the water quality parameter of interest (i.e. suspended sediments, chlorophyll). a and b are empirically derived coefficients relating measured spectral properties and measured water quality properties. Information about the spectral/optical characteristic of the water quality parameter can often be used to help select the best wavelength(s) or best model in this empirical approach. The empirical nature of these relationships limits their applications to estimating water quality parameters for water bodies with similar properties and conditions. A number of modifications to these linear modeling efforts have been developed; however, all rely on utilizing unique spectral characteristics to estimate the desired component.

These limitations in empirical approaches led to attempts to use analytical approaches based on optical properties of the water and the water quality parameters to develop physically based models of the relationship between spectral and physical characteristics of surface water studied. These physically based reflectance models have been successfully applied for estimating suspended sediment concentrations (Schiebe *et al.*, 1992; Harrington *et al.*, 1992; Dekker *et al.*, 1995). Similar relationships have been developed to estimate phytoplankton chlorophyll in marine and estuarine systems (Gitelson *et al.*, 1986; Vos *et al.*, 1986; Gitelson, 1992; Gitelson *et al.,* 2000; Mittenzwey and Gitelson, 1988; Mittenzwey *et al.*, 1992; Millie *et al.*, 1992; Quibell, 1992; Dekker, 1993; Gitelson *et al.*, 1993a,b, 1994a; Goodin *et al.*, 1993; Matthews and Boxall, 1994; Han *et al.*, 1994; Richardson *et al.*, 1995; Richardson and Zimba, 2002; Rundquist *et al.*, 1995, 1996; Yacobi *et al.*, 1995; Schalles *et al.*, 1997, 1998).

This article will review the use of remote sensing technology to measure suspended sediments and chlorophyll (color). These are the two key pollutants listed by United States Environmental Protection Agency who estimated that approximately 40% of the United States waters do not meet minimum water quality standards (US Environmental Protection Agency, 1998). Similar estimates have been made for waters of other countries. A short discussion of *in situ* sensors for remote measurement is also included since these are essential for measuring water quality parameters related

to colour that cannot be measured with remote sensing techniques and for collecting data to calibrate remote sensing models for determining water quality.

SUSPENDED SEDIMENTS

Suspended sediments (*see* **Chapter 83, Suspended Sediment Transport – Flocculation and Particle Characteristics, Volume 2**) are inorganic particles suspended in water and are the most common pollutant both in terms of weight and volume in surface waters of freshwater systems (Robinson, 1971; Lal, 1994) and are a serious problem in coastal waters (Cracknell, 1999). *In situ* measurements made either by collecting samples and filtering or by using a turbidity meter while providing accurate measurements for a point in time and space cannot account for the high spatial variability in suspended sediments found in most aquatic systems.

Suspended sediments increase the radiance reflected or emergent from surface waters (Figure 1) in the visible and near-infrared proportion of the electromagnetic spectrum (Ritchie *et al.*, 1976). Remote sensing instruments measure the difference in the electromagnetic spectrum in the visible and near-infrared wavelengths from the surface water related to the concentration of suspended sediments. *In situ*, controlled laboratory, aircraft, and satellite measurements have shown that surface water radiance is affected by amount, type, texture, and color of the suspended sediments (Novo *et al.*, 1989), sensor view angle, sun angles (Ritchie *et al.*, 1974, 1975), water depth (Blanchard and Leamer, 1973), and the atmospheric conditions (Curran and Novo, 1988). Spectral sensors on boat, aircraft, and satellite platforms have all been used to study suspended sediment patterns in freshwater and coastal aquatic systems.

Significant relationships between suspended sediments and radiance or reflectance from spectral wavelengths or combinations of wavelengths on satellite and aircraft sensors have been found. Ritchie *et al.* (1974, 1976) using *in situ* measurements found the highest correlations between suspended sediments and radiance or reflectance at wavelengths between 700 and 800 nm. Curran and Novo (1988) in a review of remote sensing of suspended sediments found that the optimum wavelength was related to suspended sediment concentration. Most studies have shown that radiance measured from the water surface increases linearly between 30 and 120 mg l^{-1} of suspended sediments then becomes asymptotic (Whitlock *et al.*, 1982; Novo *et al.*, 1989). The point where the curve becomes asymptotic is related to both the wavelength and the suspended sediment concentration. The greater the wavelength the higher the concentrations of suspended sediments before the curve becomes asymptotic (Figure 2).

This asymptotic relationship between suspended sediments and radiance or reflectance (Ritchie *et al.*, 1976, 1990) means that the amount of reflected radiance tends to saturate as suspended sediment concentrations increase. If the range of suspended sediments is between 0 and 50 mg l^{-1}, reflectance from almost any wavelength will be linearly related to suspended sediment concentrations. As the range of suspended sediments increases from 50 mg l^{-1} to 150 mg l^{-1} or higher, the use of curvilinear relationships becomes necessary. At these higher suspended sediment concentrations, wavelengths between 700 and

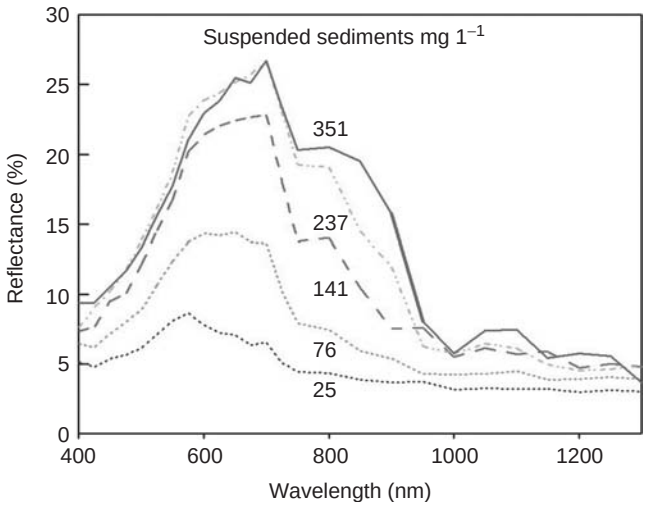


Figure 1 The relationship between reflectance and wavelength as affected by the concentration of suspended sediments (Ritchie *et al.*, 1976). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

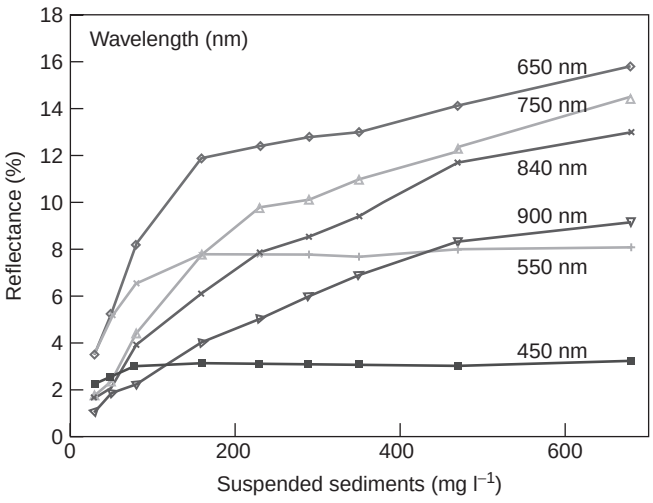


Figure 2 Relationship between reflectance and suspended sediments by wavelength showing nonlinear relation between suspended sediment concentration and reflectance (Adapted from Whitlock *et al.*, 1982). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

900 nm are best for estimating suspended sediment concentrations. Schiebe *et al.* (1992) developed a physically based reflectance model and used it to estimate suspended sediment concentrations in Lake Chicot, Arkansas. Most researchers have concluded that surface suspended sediments can be mapped and monitored in large water bodies using sensors available on current satellites. Figure 3 shows an example of mapping suspended sediments in Lake Chicot, an oxbow lake in Arkansas, USA next to the Mississippi River, using Landsat Thematic Mapper data. Suspended sediments were mapped in ranges defined by a management plan proposed for the lake.

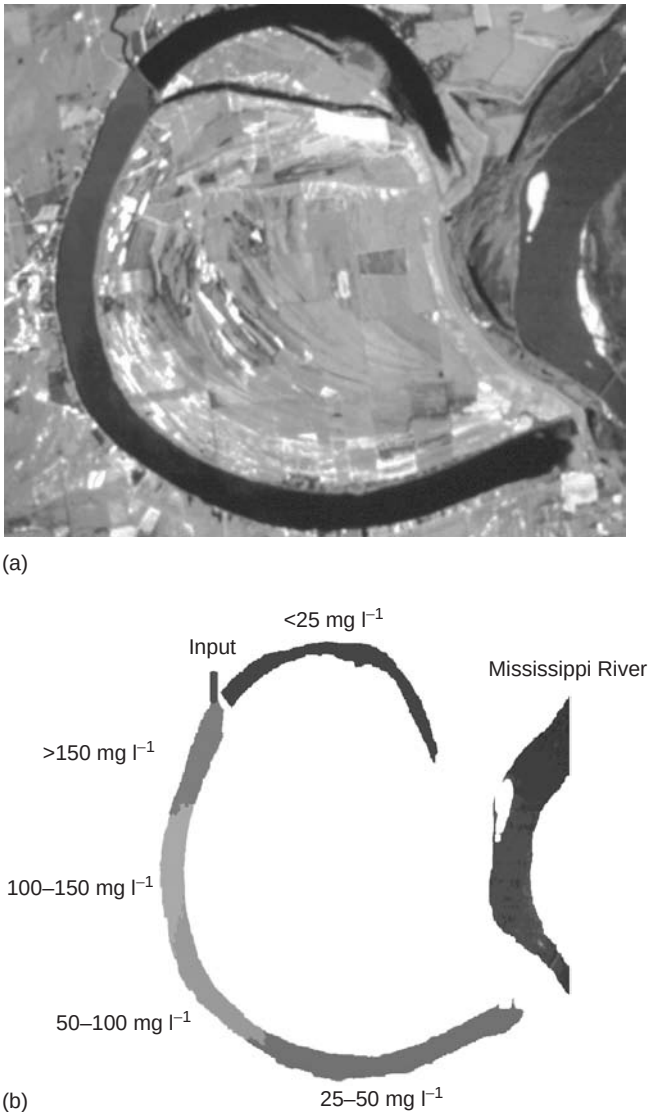


Figure 3 (a) Landsat Thematic Mapper (TM) image of Lake Chicot, Arkansas and (b) a derived image showing categories of suspended sediments mapped in Lake Chicot based on the radiance in the TM image. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Many studies have developed empirical relationships (algorithms) between the concentration of suspended sediments and radiance or reflectance for a specific date and site. Few studies have taken the next step and used these algorithms to estimate suspended sediments for another time or place (Curran and Novo, 1988). Ritchie and Cooper (1988, 1991) showed that an algorithm developed for one year was applicable for several years. Once developed, an algorithm should be applicable until some watershed event changes the quality (i.e. size, color, mineralogy, etc.) of suspended sediment particles delivered to the lake. While most researchers and managers agree that suspended sediments can be mapped with remotely sensed data, the current spatial and temporal resolution of satellite data (Ritchie and Schiebe, 2000; Ritchie *et al.*, 2003b) does not allow the detail mapping of water bodies or measurements in or from streams needed for management decisions. As new satellites come on line with higher resolution spatial and spectral data, greater application of satellite data for monitoring and assessing suspended sediments will be possible.

ALGAE/CHLOROPHYLL

Remote sensing provides a means for estimating chlorophyll (Chl) at local, regional, and even global scales. Several algorithms for remote estimation of chlorophyll concentration in marine, estuarine, and inland waters have been developed (Figure 4a–c). They are all based on reflectance in the red and near-infrared (NIR) range of the spectrum, since other portions of the spectrum are not useful for chlorophyll in productive, coastal, and freshwater ecosystems (Gitelson *et al.*, 1986, 2000; Vos *et al.*, 1986; Gitelson, 1992; Mittenzwey and Gitelson, 1988; Mittenzwey *et al.*, 1992; Millie *et al.*, 1992; Quibell, 1992; Dekker, 1993; Gitelson *et al.*, 1993a,b, 1994a; Goodin *et al.*, 1993; Matthews and Boxall, 1994; Han *et al.*, 1994; Richardson *et al.*, 1995; Richardson and Zimba, 2002; Rundquist *et al.*, 1995, 1996; Schalles *et al.*, 1997, 1998; Yacobi *et al.*, 1995; Zimba and Thomson, 2002). The basic concept of these algorithms is the inclusion of the spectrum range, which shows the maximal sensitivity to changes in chlorophyll concentration and the range with the minimum sensitivity to variation of chlorophyll concentration (Gitelson *et al.*, 1986, 1993a,b, 1994a, 2000; Yacobi *et al.*, 1995). The latter accounts for nonpigmented suspended matter that causes variation in the reflectance as detailed in the previous section of this chapter. The magnitude and position of the reflectance peak near 700 nm (R_{700}) is most sensitive to chlorophyll concentration (Gitelson, 1992), whereas the reflectance at 670 nm (R_{670}) is the least sensitive to changes in algal density, especially for chlorophyll concentrations less than 15 to 20 mg m^{-3} (Figure 4a). For chlorophyll concentrations above 20 mg m^{-3}, changes in the reflectance at

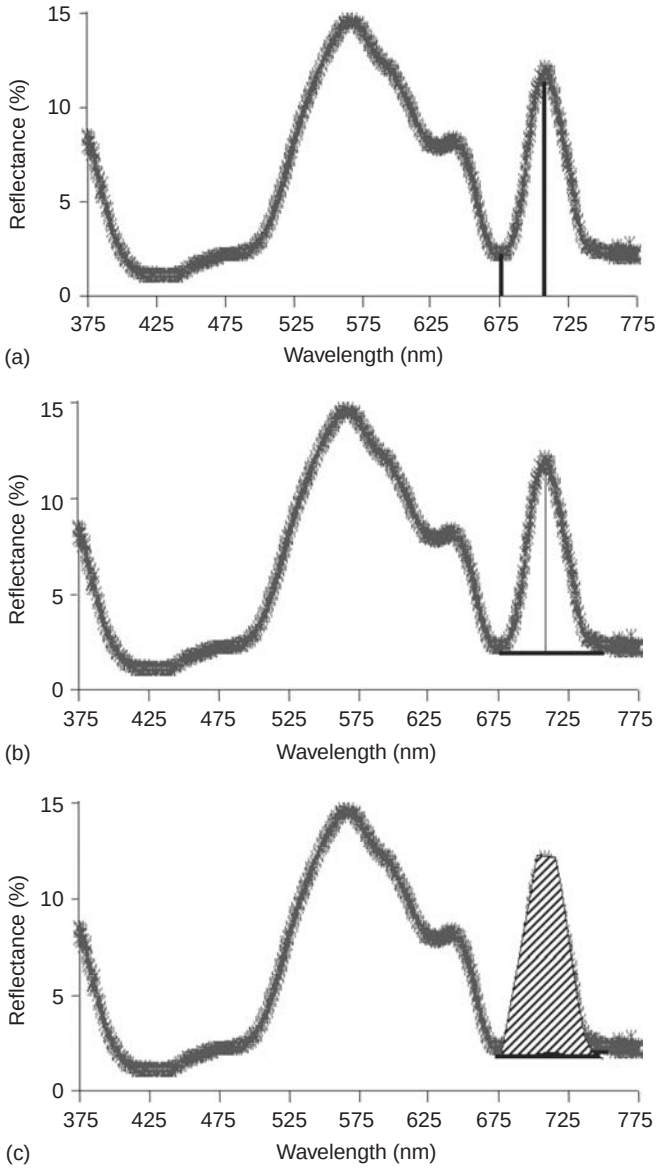


Figure 4 Examples of the three methods commonly used to estimate chlorophyll *a* concentrations- (a) Ratio of reflectance at the peak around 700 nm to reflectance at 670 nm; (Yacobi and Gitelson, 2000) (b) Reflectance height above the baseline between 670 and 800 or 900 nm; (c) Area above the baseline between 670 and 750 nm. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

R_{670} primarily depend on the concentration of nonorganic suspended matter (Gitelson *et al.*, 1993a, 1994b; Dekker, 1993; Yacobi *et al.*, 1995). Reflectance above 750 nm (R_{750}) depends on both organic and nonorganic suspended matter concentrations and is relatively insensitive to algal pigments (Han *et al.*, 1994). The variation of R_{750} is comparatively small because of strong water absorption in the NIR range. The slope of the baseline between 670 and 750 nm depends primarily on scattering by water constituents, and is less dependent on chlorophyll absorption. With variation in the amount of nonorganic and nonpigmented organic suspended matter concentration, the slope of the baseline changes but it has minimal influence on the height and area of the 700-nm peak above the baseline (Figure 4b and c). Therefore, the height of the peak and the area above the baseline between 670 and 750 nm depends mainly on chlorophyll concentration and is most often used as its quantitative measure.

These algorithms work well in systems having low turbidity associated with nonalgal suspended material, such as Lake Kinneret, Israel (Mayo *et al.*, 1995; Yacobi *et al.*, 1995) and Carter Lake, Nebraska (Schalles *et al.*, 1997). These trophically diverse systems illustrate the existence of robust relationships between chlorophyll concentration and variables of reflectance. However, chlorophyll concentrations have to be estimated against a background of high and variable inorganic suspended matter. Measured reflectance spectra in Mississippi catfish ponds typify these variable environments (Figure 5). Turbidity ranged from 40 to 350 FAU (Formazin Attenuation Unit), whereas field chlorophyll samples ranged from 8 to 600 $\mu g\, l^{-1}$. Measured reflectance of algae-laden waters prior and after addition of known clay concentrations (Figure 6a and b) show that for the same chlorophyll concentration, increases in inorganic suspended matter caused an increase in the magnitude of the reflectance peak (R_{700}). Thus, to retrieve accurately measured chlorophyll concentrations against a background of high and variable inorganic suspended matter an improvement of algorithms is required. This is particularly evident when concentrations of suspended inorganic material (montmorillonite clay in this example) exceed 200 $mg\, l^{-1}$ (Figure 6b).

The three algorithms (Figure 4) used for chlorophyll assessment performed poorly in water bodies with very high concentrations of inorganic suspended matter. As previously shown (Figure 6), scattering by inorganic particles affected the magnitude of the peak around 700 nm in the same way as chlorophyll concentration: magnitude of the peak above the baseline between 670 and 750 nm increased with total suspended solids (TSS) increase. Reflectance spectra (Figure 5) showed spectral features of chlorophyll (trough at 676 nm) as well as phycocyanin (trough at 624 nm). The peak around 700 nm corresponds to a minimum in absorption by algae and water and is representative of chlorophyll (Gitelson, 1992). These spectral features are of use for retrieval of chlorophyll and phycocyanin concentration from remotely sensed data. In contrast to productive waters, reflectance around 670 nm varied widely, showing the variation of inorganic suspended matter concentrations encountered.

Gitelson *et al.* (2003) and Dall'Olmo *et al.* (2003) have recently shown that reciprocal reflectance is a proxy of

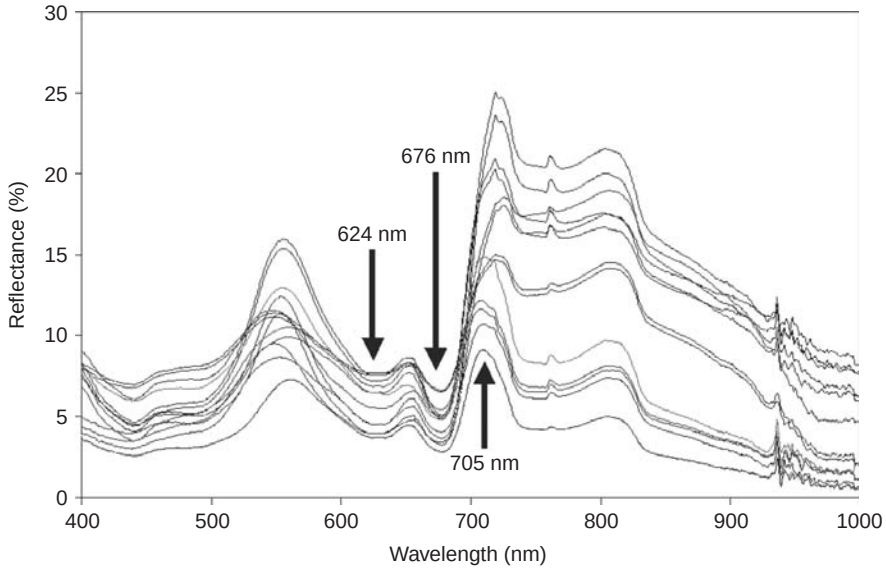


Figure 5 Representative reflectance spectra collected July 2002 from Mississippi channel catfish ponds showing variable suspended solids (705 nm) and algal biomass (chlorophyll *a* trough at 670 nm, phycocyanin trough at 624 nm). Measurements were made using a dual head Ocean Optics SD2000 operated with University of Nebraska CALMIT software (CDAP)

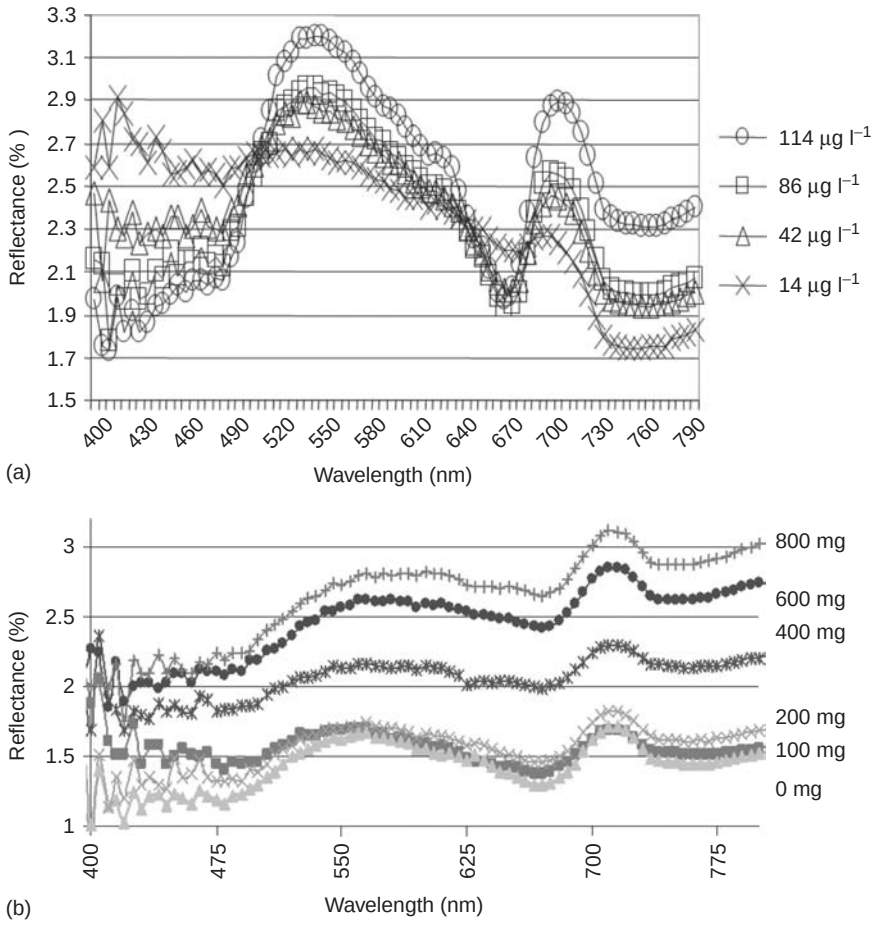


Figure 6 Effect of differing algal biomass (a) and montmorillonite clay additions (b) on reflectance of *Scenedesmus* cultures. Measurements made using a dual head Ocean Optics US2000 spectrometer operated with CDAP (University of Nebraska CALMIT) software. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

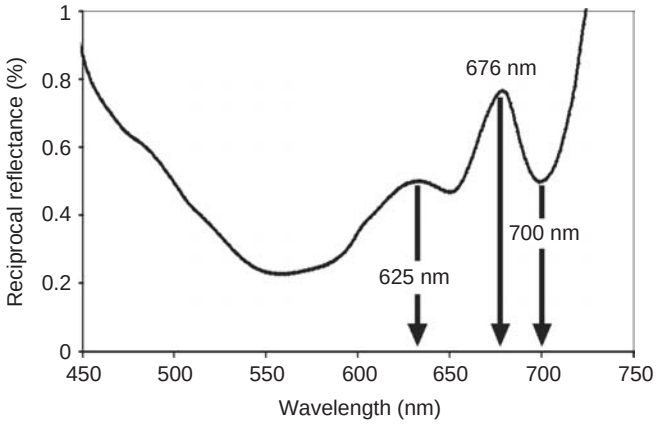


Figure 7 Reciprocal reflectance from Mississippi catfish production ponds sampled in July 2002. Measurements made using a dual head Ocean Optics SD2000 spectral radiometer. Note prominent algal biomass and solids peaks (as in Figure 5)

the ratio a/b, where a and b are absorption and scattering coefficients of the medium. Reciprocal reflectance of these turbid, eutrophic ponds (shown in Figure 7) shows three pronounced spectral features: peak at 625 nm (absorption by phycocyanin), peak at 676 nm (absorption by chlorophyll *a*), and minimum around 700 nm that manifests the minimal absorption by algae and water in this spectral range. Using the spectral features of chlorophyll *a*, Gitelson and colleagues have suggested a new approach for chlorophyll assessment. They subtracted reciprocal reflectance at 700 nm (that is proportional to scattering by all suspended matters) from reciprocal reflectance at 675 nm (that is proportional to absorption by chlorophyll *a* and scattering by all suspended matters as well). As a result, the model $(R_{675})^{-1}$-$(R_{700})^{-1}$ described chlorophyll *a* concentrations better than known algorithms. The resulting image also better fit the field truthed data (Figure 8).

The rationale for this approach is similar to that used in classification of crop health using NDVI (Normalized Difference Vegetation Index) and assessment of CASE 1 waters (i.e. Rouse *et al.*, 1974; Morel and Prieur, 1977). However the inversion method used by Gitelson and coworkers (Gitelson and Kondratyev, 1991; Gitelson *et al.*, 2003a,b; Merzlyak *et al.*, 2003; Dall'Olmo *et al.*, 2003) is an innovative approach to decompose diffusive spectral reflectances and estimate chlorophyll in both terrestrial and aquatic systems having different spectral absorption and backscattering constituents. This technique has the further advantage of removing effects of suspended solids from the chlorophyll estimation and appears to be a unifying concept that will shift methodologies from the left wall as described by Gould (1996).

In systems having highly variable suspended solids and algal biomass, increased temporal coverage is required. Aerial overflights have been effective in monitoring these systems (Millie *et al.*, 1992; Zimba and Thomson, 2002). Millie *et al.* (1992) demonstrated the utility of aerial overflights to identify total algal biomass in ponds. In subsequent analyses, (Millie *et al.*, 1995) were unable to predict algal biomass from aerial overflights collected one month apart. Confounding factors contributing to these models include the sampling frequency relative to the algal division rate, sediment resuspension from abiotic and biotic factors (e.g. wind resuspension, fish feeding), as well as the alteration of pigment composition by algae in response to altered physiological state (Falkowski and Raven, 1997). Zimba and Thomson (2002) used low-altitude remote sensing methods to assess the utility for detecting harmful algal taxa present in 18 catfish ponds. Concurrent field samples were collected weekly for six weeks and the relationship between digital imagery numbers and chlorophyll was stable for over a month period. The carotenoid pigment myxoxanthin, found in coccoid cyanobacteria, was highly correlated with the digital numbers, presence of chlorophyll, as well as the harmful algal species *Microcystis.* Approaches such as digital camera imagery are still in the developmental stage but may provide necessary information for temporally variable systems. Use of cut-band filters to target specific spectral bands of interest will be critical for selectivity.

The ability to discriminate between different algal groups is required for accurate assessment of harmful algal species (Millie *et al.*, 1995; Zimba and Thomson, 2002; Bazzani and Cecchi, 1995). Spectral features unique to these algae must be used to characterize algae in terms other than chlorophyll *a*. Spectral features that have not been exploited include the phycobilin signatures of cyanobacteria, as well as unique carotenoid compositions of phylogenetic groups. For example, Millie *et al.* (1995) identified a unique biomarker (gyrodinium diester) in the bloom forming dinoflagellate *Gymnodinium (Karenia) breve*. This alga causes reoccurring toxic blooms off the Florida west coast, and has expanded in the Gulf of Mexico recently to Alabama, Mississippi, and Texas coast. As an initial identification tool, Millie *et al.* (1995) used fourth derivative analyses of absorbance to discriminate unialgal and mixed assemblages containing this alga. The amount of signal variance explained by such approaches is trivial, but as a first approximation it is a useful concept. Zimba and coworkers (Zimba *et al.*, 2001; Zimba and Thomson, 2002; Zimba and Grimm, 2003) have demonstrated the co-occurrence of the carotenoid myxoxanthin and presence of *Microcystis*, a genus known to produce harmful secondary metabolites, including the off-flavor compound *beta*-cyclocitral and the toxin microcystin. A peak near 504 nm was also identified by Richardson and Zimba (2002) as indicating presence of coccoid cyanobacteria in Florida

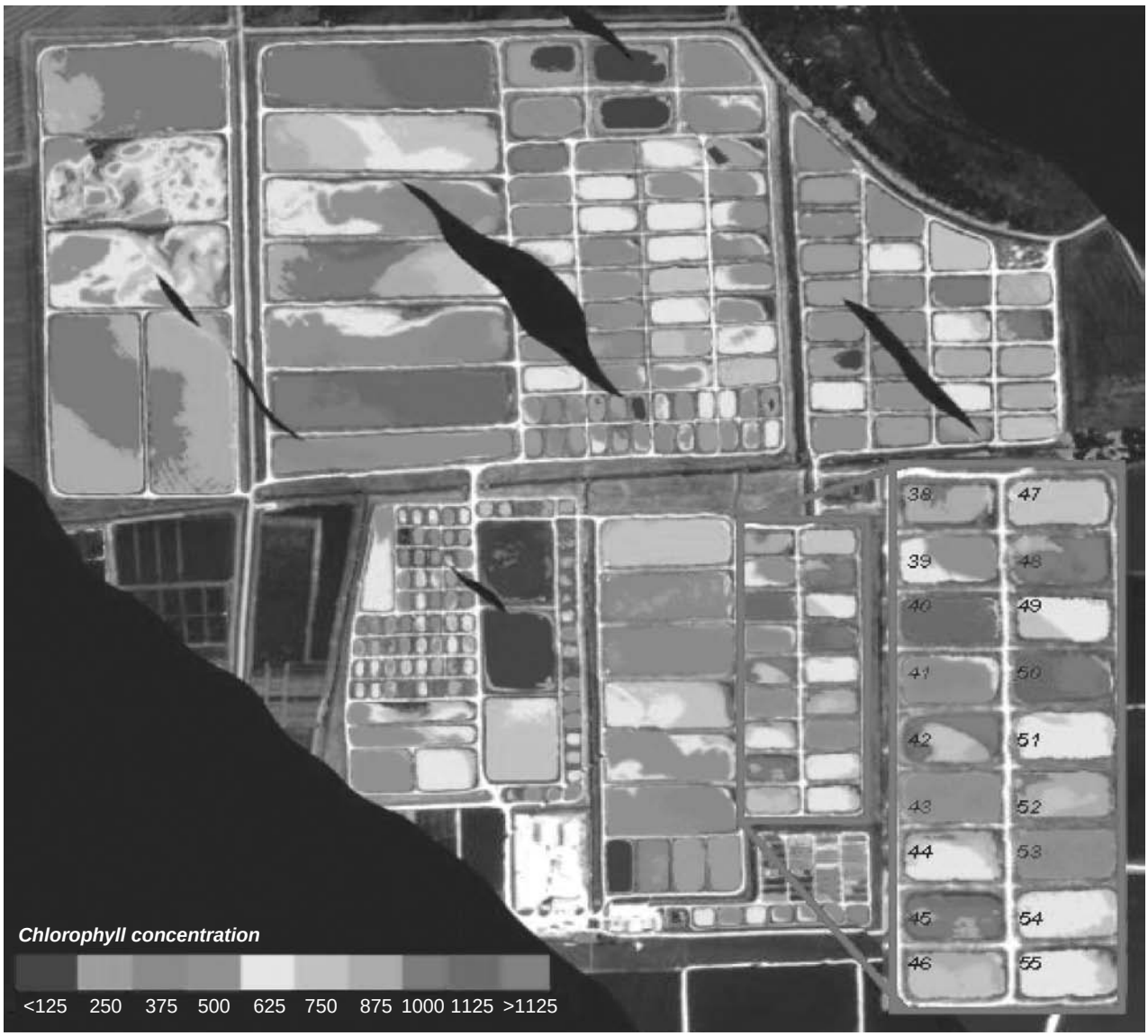


Figure 8 Chlorophyll concentration estimated from Mississippi catfish ponds sampled in July 2002. Imagery captured using an AISA hyperspectral scanner operated by University of Nebraska CALMIT personnel. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Bay, Florida. Phycobilin signatures appear useful in identifying harmful cyanobacterial blooms in catfish production systems in Mississippi.

The development of miniaturized sensor equipment requiring simpler platforms than satellites provides greater opportunities for optimization of methodologies, increased frequency of sampling events, and lower operating costs. Airborne sensor systems are useful to assess optimal bandwidths and thereby build the foundation for incorporation in future satellite designs. Additionally, the resolution of airborne systems (as compared to satellites) may provide greater sensitivity in assessing dynamic features of algal growth, particularly in upwelling areas, riverine plumes, as well as in closely monitoring bloom development within small lakes, reservoirs, and rivers. The rapid development of miniaturized handheld spectral units also will contribute to optimization of assessment methodologies.

CONCLUSIONS

Remote sensing techniques for measuring suspended sediments and algae have improved greatly since the 1970s and presently have many actual and potential applications for assessing water resources and for monitoring water quality. However, limitations in spectral, spatial, and temporal resolution of sensors on current satellites restrict wide scale

application of satellite data for monitoring water quality. New satellites and sensors (hyperspectral, high spatial resolution) already launched or planned to be launched over the next decade may provide the improved spectral and spatial resolution needed to monitor water quality parameters in surface waters from space platforms. However, there may be many cases requiring the use of airborne or even handheld sensors so that temporal and spatial data can be collected on fast developing pollution events (i.e. algae bloom, sediment laden water from storm events).

The integration of remotely sensed data, GPS (Global Positioning Systems), and GIS (Geographic Information Systems) technologies provides a valuable tool for monitoring and assessing surface waters. Remotely sensed data provide a permanent geographically located image database as a baseline for future comparisons. The integrated use of remotely sensed data, GPS, and GIS will enable consultants and natural resource managers to develop management plans for a variety of natural resource management applications. Improvements in software for imagery or reflectance signal processing will also improve the near-real time ability to use these data for management decisions.

Acknowledgments

We thank Anatoly Gitelson for sharing expertise and comments on an earlier draft of this chapter. Mention of proprietary names are necessary to report factually on available data; however, the US Department of Agriculture neither warrants or guarantees the standard of a product and implies no approval of a product to the exclusion of others that may be suitable.

REFERENCES

Bazzani M. and Cecchi G. (1995) Algae and mucillagine monitoring by fluorescence lidar experiments in field. *EARSeL Advances in Remote Sensing*, **3**, 90–101.

Blanchard B.J. and Leamer R.W. (1973) Spectral reflectance of water containing suspended sediment. In *Remote Sensing and Water Resource Management*, Thomson K.P.B., Lane R.K. and Csallay S.C. (Eds.), American Water Resources Association: Urbana, pp. 339–347.

Bowker D.E. and Witte W.G. (1975) An investigation of waters in the lower Chesapeake Bay. *Proceedings of the 9th International Symposium on Remote Sensing of the Environment*, University of Michigan: Ann Arbor, pp. 411–419.

Boyer E.W., Goodale C.L., Jaworski N.A. and Howarth R.W. (2002) Anthropogenic nitrogen sources and relationships to riverine nitrogen export in the northeastern U.S.A. *Biogeochemistry*, **57/58**, 137–169.

Cracknell A.P. (1999) Remote sensing techniques in estuaries and coastal zones-an update. *International Journal of Remote Sensing*, **19**, 485–496.

Curran P.J. and Novo E.M.M. (1988) The relationship between suspended sediment concentration and remotely sensed spectral radiance: a review. *Journal of Coastal Research*, **4**, 351–368.

Dall'Olmo D., Gitelson A.A. and Rundquist D.C. (2003) Towards a unified approach for remote estimation of chlorophyll-a from terrestrial vegetation and turbid productive waters. *Geophysical Research Letters*, **30**, 1938 doi:10.1029/2003GL018065.

Dekker A. (1993) *Detection of the Optical Water Quality Parameters for Eutrophic Waters by High Resolution Remote Sensing*, Ph.D. Thesis, Free University, Amsterdam, p. 212.

Dekker A.G., Malthus T.J. and Hoogenboom H.J. (1995) The remote sensing of inland water quality. In *Advances in Remote Sensing*, Danson F.M. and Plummer S.E. (Eds.), John Wiley & Sons: Chichester, pp. 123–142.

Falkowski P.G. and Raven J.A. (1997) *Aquatic Photosynthesis*, Blackwell Scientific: Malden.

Gitelson A. (1992) The peak near 700 nm on reflectance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. *International Journal of Remote Sensing*, **13**, 3367–3373.

Gitelson A. (1993) The nature of the peak near 700 nm on the radiance spectra and its application for remote estimation of phytoplankton pigments in inland waters. *Optical Engineering and Remote Sensing, SPIE*, **1971**, 170–179.

Gitelson A., Garbuzov G., Szilagyi F., Mittenzwey K.-H., Karnieli A. and Kaiser A. (1993a) Quantitative remote sensing methods for real-time monitoring inland water quality. *International Journal of Remote Sensing*, **14**, 1269–1295.

Gitelson A.A., Gritz U. and Merzlyak N.M. (2003) Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, **160**, 271–282.

Gitelson A.A. and Kondratyev K.Y. (1991) Optical models of mesotrophic and eutrophic water bodies. *International Journal of Remote Sensing*, **12**, 373–385.

Gitelson A., Nikanorov A.M., Sabo G. and Szilagyi F. (1986) Etude de la qualite des eaux de surface par teledetection. *International Association for Hydrological Sciences Publications*, **157**, 111–121.

Gitelson A., Mayo M. and Yacobi Y.Z. (1994a) Signature analysis of reflectance spectra and its application for remote observations of the phytoplankton distribution in Lake Kinneret. *Mesures Physiques et Signatures en Teledetection. ISPRS 6th International Symposium*, Val d'Isere, pp. 277–283.

Gitelson A., Mayo M., Yacobi Y.Z., Parparov A. and Berman T. (1994b) The use of high spectral radiometer data for detection of low chlorophyll concentrations in Lake Kinneret. *Journal of Plankton Research*, **16**, 993–1002.

Gitelson A., Szilagyi F. and Mittenzwey K. (1993b) Improving quantitative remote sensing for monitoring of inland water quality. *Water Resources*, **7**, 1185–1194.

Gitelson A.A., Yacobi Y.Z., Schalles J.F., Rundquist D.C., Han L., Stark R. and Etzion D. (2000) Remote estimation of phytoplankton density in productive waters. *Archives of Hydrobiology Special Issues Advances in Limnology*, **55**, 121–136.

Goodin D.G., Han L., Fraser R.N., Rundquist D.C., Stebbins W.A. and Schalles J.F. (1993) Analysis of suspended solids in water using remotely sensed high resolution derivative

spectra. *Photogrammetric Engineering and Remote Sensing*, **59**, 505–510.

Gould S.J. (1996) *Full House*, Crown Publishers: New York, pp. 244.

Han L., Rundquist D.C., Liu L.L., Fraser R.N. and Schalles J.F. (1994) The spectral responses of algal chlorophyll in water with varying levels of suspended sediment. *International Journal of Remote Sensing*, **15**, 3707–3718.

Harrington J.A. Jr, Schiebe F.R. and Nix J.F. (1992) Remote sensing of Lake Chicot, Arkansas: monitoring suspended sediments, turbidity and secchi depth with Landsat MSS. *Remote Sensing of Environment*, **39**, 15–27.

Jerlov N.G. (1976) *Marine Optics*, Elsevier: Amsterdam.

Johnson R.W. (1976) Quantitative sediment mapping from remotely sensed multispectral data. In *Remote Sensing of Earth Resources*, Shahrokhi F. (Ed.), University of Tennessee: Tullahoma, pp. 565–576.

Kirk J.T.O. (1983) *Light and Photosynthesis in Aquatic Ecosystems*, Cambridge University Press: Cambridge.

Klemas V., Otley M., Philpot W., Wethe C., Rogers R. and Shah N. (1975) Correlation of coastal water turbidity and current circulation with ERTS-1 and Skylab imagery. *Proceedings of the 9th International Symposium on Remote Sensing of the Environment*, University of Michigan: Ann Arbor, pp. 1289–1317.

Lal R. (1994) *Soil Erosion*, Soil and Water Conservation Society: Ankeny.

Lal R. (2001) Soil degradation by erosion. *Land Degradation and Development*, **12**, 519–539.

Matthews A.M. and Boxall S.R. (1994) Novel algorithms for the determination of phytoplankton concentration and maturity. *Proceedings of the Second Thematic Conference on Remote Sensing for Marine and Coastal Environments*, Vol. 1, Environmental Research Institute of Michigan: New Orleans, pp. I173–I180,31 January-2 February 1994.

Mayo M., Gitelson A., Yacobi Y. and Ben-Avraham Z.Z. (1995) Chlorophyll distribution in lake Kinneret determined from Landsat Thematic Mapper data. *International Journal of Remote Sensing*, **16**, 175–182.

Merzlyak M.N., Solovchenko A.E. and Gitelson A.A. (2003) Reflectance spectral features and non-destructive estimation of chlorophyll, carotenoid, and anthocyanin content in apple fruit. *Postharvest Biology and Technology*, **27**, 197–211.

Millie D.F., Baker M.C., Tucker C.S., Vinyard B.T. and Dionigi C.P. (1992) High resolution, airborne remote-sensing of bloom-forming phytoplankton. *Journal of Phycology*, **28**, 281–290.

Millie D.F., Vinyard B.T., Baker M.C. and Tucker C.S. (1995) Testing the temporal and spatial validity of site specific models derived from airborne remote sensing. *Canadian Journal of Fisheries and Aquatic Science*, **52**, 1094–1107.

Mittenzwey K.-H. and Gitelson A. (1988) In-situ monitoring of water quality on the basis of spectral reflectance. *International Revue of Hydrobiology*, **73**, 61–72.

Mittenzwey K.-H., Ullrich S., Gitelson A. and Kondrat'ev K.Y.a (1992) Determination of chlorophyll-a of inland waters on the basis of spectral reflectance. *Limnology and Oceanography*, **37**, 147–149.

Morel A. and Prieur L. (1977) Analysis of variations in ocean color. *Limnology and Oceanography*, **22**, 709–722.

Novo E.M.M., Hansom J.D. and Curran P.J. (1989) The effect of sediment type on the relationship between reflectance and suspended sediment concentration. *International Journal of Remote Sensing*, **10**, 1283–1289.

Pimentel D., Harvey C., Resosudarmo P., Sinclair K., Kurz D., McNair M., Crist S., Shipritz L., Fitton L., Saffouri R., *et al.* (1995) Environmental and economic cost of soil erosion and conservation benefit. *Science*, **267**, 1117–1123.

Quibell G. (1992) Estimation chlorophyll concentrations using upwelling radiance from different freshwater algal genera. *International Journal of Remote Sensing*, **13**, 2611–2621.

Richardson L.L., Buisson D. and Ambrosia V. (1995) Use of remote sensing coupled with algal accessory pigment data to study phytoplankton bloom dynamics in Florida Bay. *Proceedings of the Third Thematic Conference on Remote Sensing for Marine and Coastal Environments*, Vol. II, Environmental Research Institute of Michigan: Ann Arbor, Seattle, pp. 125–134, 18–20 September 1995.

Richardson L.L. and Zimba P.V. (2002) Spatial and temporal patterns of phytoplankton in Florida Bay: utility of algal accessory pigments and remote sensing to assess bloom dynamics. In *South Florida's Linked Ecosystems*, Porter J.W. and Porter K.G. (Eds.), CRC Press: Boca Raton, pp. 461–478.

Ritchie J.C. (2000) Soil erosion. In *Remote Sensing in Hydrology and Water Management*, Schultz G.A. and Engman, E.T. (Eds.), Springer-Verlag: Berlin, pp. 271–286, 349–350.

Ritchie J.C. and Cooper C.M. (1988) Comparison on measured suspended sediment concentrations with suspended sediment concentrations estimated from Landsat MSS data. *International Journal of Remote Sensing*, **9**, 379–387.

Ritchie J.C. and Cooper C.M. (1991) An algorithm for using Landsat MSS for estimating surface suspended sediments. *Water Resources Bulletin*, **27**, 373–379.

Ritchie J.C., Cooper C.M. and Schiebe F.R. (1990) The relationship of MSS and TM digital data with suspended sediments, chlorophyll, and temperature in Moon Lake, Mississippi. *Remote Sensing of Environment*, **33**, 137–148.

Ritchie J.C., McHenry J.R., Schiebe F.R. and Wilson R.B. (1974) The relationship of reflected solar radiation and the concentration of sediment in the surface water of reservoirs. In *Remote Sensing of Earth Resources*, Vol. III, Shahrokhi F. (Ed.), The University of Tennessee Space Institute: Tullahoma, pp. 57–72.

Ritchie J.C. and Schiebe F.R. (2000) Water quality. In *Remote Sensing in Hydrology and Water Management*, Schultz G.A. and Engman E.T. (Eds.), Springer-Verlag: Berlin, pp. 287–303, 351–352.

Ritchie J.C., Schiebe F.R., McHenry J.R., Wilson R.B. and May J. (1975) Sun angle, reflected solar radiation and suspended sediments in north Mississippi reservoirs. In *Remote Sensing of Earth Resources*, Vol. IV, Shahrokhi F. (Ed.), The University of Tennessee Space Institute: Tullahoma, pp. 555–564.

Ritchie J.C., Schiebe F.R. and McHenry J.R. (1976) Remote sensing of suspended sediment in surface water. *Photogrammetric Engineering and Remote Sensing*, **42**, 1539–1545.

Ritchie J.C., Walling D.E. and Peters J. (2003a) Application of geographic information systems and remote sensing for

quantifying patterns of erosion and water quality: introduction. *Hydrological Processes*, **17**, 885–886.

Ritchie J.C., Zimba P.V. and Everitt J.H. (2003b) Remote sensing techniques to assess water quality. *Photogrammetric Engineering and Remote Sensing*, **69**, 695–704.

Robinson A.R. (1971) Sediment: our greatest pollutant. *Journal of Soil and Water Conservation*, **53**, 406–408.

Rouse J.W., Haas R.H., Schell J.A., Deering D.W. and Harlan J.C. (1974) *Monitoring the Vernal Retrogradation of Natural Vegetation*, NASA/GSFC Type III Final Report, NASA/GSFC: Greenbelt.

Rundquist D.C., Han L., Schalles J.F. and Peake J.S. (1996) Remote measurement of algal chlorophyll in surface waters: the case for the first derivative of reflectance near 690 nm. *Photogrammetric Engineering and Remote Sensing*, **62**, 195–200.

Rundquist D.C., Schalles J.F. and Peake J.S. (1995) The response of volume reflectance to manipulated algal concentrations above bright and dark bottoms at various depths in an experimental pool. *Geocarta International*, **10**, 5–14.

Schalles J.F., Gitelson A., Yacobi Y.Z. and Kroenke A.E. (1998) Chlorophyll estimation using whole seasonal, remotely sensed high spectral-resolution data for an eutrophic lake. *Journal of Phycology*, **34**, 383–390.

Schalles J.F., Schiebe F.R., Starks P.J. and Troeger W.W. (1997) Estimation of algal and suspended sediment loads (singly and combined) using hyperspectral sensors and integrated mesocosm experiments. *Proceedings of the Fourth International Conference on Remote Sensing of Marine and Coastal Environments*, Vol. I, Environmental Research Institute of Michigan: Ann Arbor, pp. 247–258, 17–19 March 1997.

Schiebe F.R., Harrington J.A. Jr and Ritchie J.C. (1992) Remote sensing of suspended sediments: the Lake Chicot, Arkansas project. *International Journal of Remote Sensing*, **13**, 1487–1509.

U.S. Environmental Protection Agency Office of Water (1998) *The Quality of Our Nation's Waters: 1996, Executive Summary*, April 1998, Environmental Protection Agency: Washington, Report EPA842-F-99-0003D.

Vos W.L., Donze M. and Bueteveld H. (1986) *On the Reflectance Spectrum of Algae in Water: The Nature of the Peak at 700 nm and its Shift with Varying Concentration*, Communication on Sanitary Engineering and Water Management: Delft, The Netherlands, Technical Report 86, p. 22.

Walling D.E. (1999) Linking land use, erosion and sediment yields in river basins. *Hydrobiologia*, **410**, 223–240.

Whitlock C.H., Kuo C.Y. and LeCroy S.R. (1982) Criteria for the use of regression analysis for remote sensing of sediment and pollutants. *Remote Sensing of Environment*, **12**, 151–168.

Yacobi Y.Z., Gitelson A. and Mayo M. (1995) Remote sensing of chlorophyll in Lake Kinneret using high spectral resolution radiometer and Landsat TM: Spectral features of reflectance and algorithm development. *Journal of Plankton Research*, **17**, 2155–2173.

Yacobi Y.Z. and Gitelson A.A. (2000) Simultaneous remote measurement of chlorophyll and total seston in productive inland waters. *Verhandlungen International Vereinigung Limnologie*, **27**, 2983–2986.

Zimba P.V. and Grimm C.C. (2003) A synoptic survey of musty/muddy odor metabolites and microcystin toxin occurrence and concentration in southeastern USA channel catfish (Ictalurus punctatus Ralfinesque) production ponds. *Aquaculture*, **218**, 81–87.

Zimba P.V. and Thomson S. (2002) Detection of toxin-producing algae by low altitude remote sensing methods. Presented at the 7th *International Conference of Remote Sensing in the Marine and Coastal Environment*, Miami, p. 3–15.

Zimba P.V., Grimm C.C. and Dionigi C.P. (2001) Phytoplankton community structure, biomass, and off flavor: size relationships in Louisiana catfish ponds. *Journal of the World Aquaculture Society*, **32**, 96–104.

63: Estimation of Precipitation Using Ground-based, Active Microwave Sensors

MATTHIAS STEINER

Princeton University, Princeton, NJ, US

Ground-based radar measurements have provided a wealth of information about precipitation systems, how they evolve and organize themselves. Advances in Doppler and polarimetric radar technology, in particular, have yielded unprecedented insight into the dynamics and microphysics of precipitation formation. This paper summarizes the basic components and principals of radar measurements, and discusses how radar can be used to identify different types of precipitation particles as well as recognize nonmeteorological radar echoes, and obtain quantitative precipitation estimates.

INTRODUCTION

Quantitative information about precipitation is of major concern to meteorologists, hydrologic scientists, water resources managers, and environmental legislators. Yet, accurate measurement of precipitation over the relevant space and time scales remains a challenge. Active remote sensors, such as *radar* (*RA*dio *D*etection *A*nd *R*anging), have proven essential in capturing information about precipitation at scales of less than one to several hundred kilometers. For example, modern radar technology enables estimation of precipitation rates and distinction between different types of precipitation. In addition, information about the severity of storms and the associated wind field may be obtained. The quantitative estimation of precipitation, however, is a complex procedure that involves issues of radar hardware and signal processing software design, propagation of electromagnetic waves through the atmosphere and their interaction with precipitation particles, physics of precipitation production, data quality control, rain (and wind field) retrieval, and uncertainty analyses.

The relevant aspects of the radar-based precipitation measurement are highlighted in this article. More detailed information may be obtained from the relatively easy to comprehend textbooks by Battan (1973), Rinehart (1999), and Lhermitte (2002), or the more mathematically comprehensive books by Doviak and Zrnić (1993) and Bringi and Chandrasekar (2001). An excellent review of polarimetric radar measurements is given by Illingworth (2003). The challenges of (operational) radar measurements of precipitation in complex terrain are illustrated by Germann and Joss (2003). The monograph edited by Atlas (1990) provides an overview of the early days of radar meteorology and its evolution over approximately five decades. Skolnik (1988) elaborates on hardware and technical details.

BASIC RADAR PRINCIPLES

The main components of the radar are an active source for emission of electromagnetic radiation (i.e. *transmitter*) and a *receiver* designed to detect and amplify the return signals. The transmitter and receiver are collocated for monostatic radar systems (typical for weather radar) or separated by some distance for bistatic (or multistatic) radar systems. The transmission and reception of electromagnetic signals is facilitated by the *antenna*, a device that is usually combined with a parabolic reflector to focus the energy and guide it away from the radar in a particular direction. Typical gains of focusing the transmitted energy over an isotropic source of radiation are on the order of 20 to 50 dB (i.e. increases of 10^2 to 10^5). The shape and size of the reflector determine the antenna beam pattern. For simplicity, the term radar antenna will refer to both the antenna and the reflector combined. The transmitter is controlled by a modulator that regulates how often a pulse is sent out and provides the correct waveform and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

duration for the transmitted pulse. Most radar transmit a few thousand Watts (1 W = 1 Joule per second) to more than a Megawatt (10^6 W) of power and are capable of detecting returned powers as small as 10^{-14} W or less. If the same antenna is used for transmission and reception (i.e. monostatic radar system), a highly insulated switch is needed to protect the sensitive receiver from the high power output of the transmitter. Computers are used to control the transmission of the radar signals, the direction the radar antenna is pointing, and the processing and display of the received signal.

In the earth's atmosphere, electromagnetic waves travel approximately at the speed of light ($c = 3 \times 10^8\,\mathrm{m\,s^{-1}}$) and are characterized by a frequency f in Hertz (1 Hz = 1 cycle per second) that is related to the radar wavelength λ through

$$f = \frac{c}{\lambda} \tag{1}$$

Typical values of radar wavelengths are compiled in Table 1. There are tradeoffs to consider when determining the appropriate wavelength and size of the radar antenna. For a given wavelength, a larger antenna results in a narrower beam (i.e. better focusing the radiated energy). For a circular parabolic antenna, the *beamwidth* ϑ (in degrees), defined as the angular width where the transmitted power is half the maximum power centered along the antenna pointing axis, is determined by the ratio of radar wavelength λ and antenna diameter d_a (both given in the same units),

$$\vartheta = \frac{180}{\pi}\frac{\lambda}{e_a d_a} \tag{2}$$

where e_a is the antenna aperture efficiency (typically $e_a \approx 0.8$ or less). Desirable beamwidths are on the order of 1 degree or less. A radar operating at a 10-cm wavelength, therefore, requires an antenna with diameter of approximately 10 m to achieve that. Increasing the size of a radar antenna, however, dramatically increases its weight and cost. For practical purposes, therefore, ground-based weather radar, equipped with a steerable antenna that can rotate in azimuth and elevation, employs transmitters with wavelengths in the S, C, or X band. Mobile radar platforms, such as those mounted on trailers, trucks, or airplanes, are typically based on X or K band transmitters, although some C and S band radars are in use also. Longer wavelengths (L, VHF, and UHF bands) are employed in fixed, vertically pointing installations (so-called *wind profilers*) to measure the vertical profile of the horizontal wind speed and direction. Shorter wavelengths (K and W bands) are used for observation of nonprecipitating clouds. A notable exception is the K_u band precipitation radar aboard the Tropical Rainfall Measuring Mission (TRMM) satellite.

Table 1 Standard nomenclature (band designation) for radar wavelengths and frequencies

Band designation	Nominal wavelength	Nominal frequency
HF	10–100 m	3–30 MHz
VHF	1–10 m	30–300 MHz
UHF	0.3–1 m	300–1000 MHz
L	15–30 cm	1–2 GHz
S	7.5–15 cm	2–4 GHz
C	3.75–7.5 cm	4–8 GHz
X	2.5–3.75 cm	8–12 GHz
K_u	1.67–2.5 cm	12–18 GHz
K	1.11–1.67 cm	18–27 GHz
K_a	0.75–1.11 cm	27–40 GHz
W	1–7.5 mm	40–300 GHz

Radar-transmitted signals are described by either a pulse duration τ (if measured in units of time) or a pulse length h (if expressed in units of distance). Weather radars measuring precipitation typically employ simple (i.e. not coded) pulses of durations ranging from 0.2 to 2 μs (1 μs = 10^{-6} s), corresponding to pulse lengths of 30 to 300 m. The pulse duration, together with the beamwidth, provide a measure of the spatial resolution of the radar measurements. The radar sends out signals on a regular basis. The unambiguous distance $r_{\max}$ a pulse can travel to a backscattering target and return to the radar before the next pulse is sent out is determined by the *pulse repetition frequency* (*PRF*)

$$r_{\max} = \frac{c}{2PRF} \tag{3}$$

Typical *PRF* values range from 300 to 3000 Hz, corresponding to unambiguous ranges of 500 and 50 km, respectively. Signals returning to the radar from a distant target after the next pulse has been released cannot be distinguished from a signal more quickly returning from a nearby target and are thus depicted closer to the radar than they actually are. This feature is called *range folding*. Range-folded radar echoes are recognized by their apparent low intensity, radially elongated shape, and reduced vertical extent compared to correctly located storm signatures.

So far we have assumed that the target remains at a fixed distance r from the radar. A target moving relative to the radar causes a shift in frequency of the radar signal, known as the *Doppler effect*. For the radar to detect a frequency shift, the phase ϕ of the returned signal has to be compared to the phase ϕ_0 of the signal sent out, where

$$\phi - \phi_0 = \frac{2r}{\lambda} \tag{4}$$

The change of phase with time from one pulse to the next is given by

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{2}{\lambda}\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{2v}{\lambda} \tag{5}$$

where v is the speed of the target relative to the radar. Note that the radar can only measure the velocity component of

a moving target in the radial direction. The ability of a Doppler radar to detect small phase shifts depends critically on either the stability of the transmitter frequency and phase from one pulse to the next (achieved by coherent transmitters such as klystron tubes) or the capability of the radar to remember the frequency and phase of each pulse sent out so that they can be compared to the returned signal (requires a stable local oscillator).

Radial velocity measurements can be ambiguous. The radar cannot distinguish between a target that has traveled half a wavelength towards the radar within the time from one pulse to the next and a target that has traveled a similar distance away from the radar. The maximum velocity $v_{\max}$ that can be determined unambiguously is called the *Nyquist velocity* and is given by

$$v_{\max} = \frac{\lambda}{4} PRF \tag{6}$$

Velocity folding is often characterized by abrupt changes in the radar-indicated velocity (near the Nyquist velocity) around the aliased region. Doppler radar operations are thus faced with the choice of either maximizing the unambiguous velocity or the unambiguous range measurements. This is known as the *Doppler dilemma* and is expressed by combining equations (3) and (6)

$$v_{\max} r_{\max} = \frac{c\lambda}{8} \tag{7}$$

We can increase $r_{\max}$ and $v_{\max}$ by employing longer wavelengths, which in turn requires bigger radar systems and increased costs, as detailed earlier. If instead the *PRF* is increased, the unambiguous velocity will increase yet the unambiguous range decreases. An acceptable compromise must be determined for each radar system and desired application. Alternatively, emerging techniques based on using multiple *PRF*s or staggered pulse release times may ease the unambiguous velocity problem.

RADAR PARAMETERS AND THEIR METEOROLOGICAL INTERPRETATION

Radar Equation for Distributed Scatterers

Weather echoes obtained by individual radar pulses are constantly changing (fluctuating), because precipitation particles are not homogeneously distributed in space and they rearrange themselves relative to each other and relative to the radar. Thus, echoes from multiple radar pulses are averaged to obtain a more accurate measure of a storm's intensity. The number of independent pulses to be averaged for a good measure of the signal strength depends on how quickly samples decorrelate with time, which is a function of the radar wavelength and scanning, the size distribution of hydrometeors, and the wind shear or turbulence within the radar sampling volume. Decorrelation times are shorter for shorter radar wavelengths and broader hydrometeor size distributions. Typically many tens of radar pulses are averaged for a given point in space.

Assuming a circular parabolic antenna with a Gaussian-shaped beam pattern, it can be shown that the received power P_r backscattered from an ensemble of precipitation particles in a sample volume illuminated by the radar at a distance r is related to the transmitted power P_t through the radar equation

$$P_r = P_t \frac{G^2 \lambda^2 \vartheta^2 h}{1024 \ln(2) \pi^2 r^2} \sum_{\text{Vol}} \sigma_i \tag{8}$$

where G is the antenna gain, λ the radar wavelength, ϑ the angular beamwidth, h the pulse length, and σ_i the backscattering cross-sectional area of particle i. For spherical particles of diameter D, the backscattering cross-sectional area is

$$\sigma = \frac{\pi^5}{\lambda^4} |K|^2 D^6 = \frac{\pi^5}{\lambda^4} \left| \frac{m^2 - 1}{m^2 + 2} \right|^2 D^6 \tag{9}$$

where K is a parameter related to the complex refractive index m. The value of $|K|^2$ depends mainly on the target's composition and to a lesser extent on temperature and radar wavelength. For most commonly used radar wavelengths (S, C, and X band), $|K_w|^2 = 0.93$ for water at temperatures from 0° to 30 °C and $|K_i|^2 = 0.176$ for solid ice particles at temperatures as low as −20 °C. Equation (9) is valid when the size of the sphere is significantly smaller than the radar wavelength, $D < \lambda/10$, which is known as the *Rayleigh approximation*. For spherical particles much larger than the wavelength, $D > 10\lambda$, the backscattering cross-sectional area approaches the geometric area, $\sigma = \pi(D/2)^2$, and optical scattering behavior dominates. The Mie scattering, or resonance region, falls in between the Rayleigh and optical regions. Combining equations (8) and (9) yields the radar equation for spherical particles using the Rayleigh approximation

$$P_r = P_t \frac{\pi^3 G^2 \vartheta^2 h}{1024 \ln(2) \lambda^2} \frac{|K|^2}{r^2} \sum_{\text{Vol}} D_i^6 = C \frac{|K|^2}{r^2} \sum_{\text{Vol}} D_i^6 \tag{10}$$

where the radar-specific parameters have been grouped together to form the "radar constant" C. From (10) it can be seen that radar employing shorter wavelengths are more sensitive than radar based on longer wavelengths. For example, for a given transmitted power, antenna gain, and beamwidth, a 3-cm wavelength radar will receive approximately one order of magnitude more backscattered power from a target than a 10-cm wavelength radar.

Radar Parameters

A variety of radar parameters can be defined, depending on the radar's hardware and signal processing capabilities, as presented below. Parameters based on the received power are discussed first, followed by those that are based on Doppler information. The meteorological relevance of these parameters will be exemplified in a subsequent section.

For raindrops (i.e. spherical particles) and conditions of Rayleigh scattering, a *radar reflectivity factor* Z can be defined as

$$Z = \frac{\lambda^4}{\pi^5 |K_w|^2} \int N(D)\sigma(D)\,\mathrm{d}D = \int N(D)D^6\,\mathrm{d}D \quad (11)$$

where $N(D)$ is the raindrop size distribution and D is the drop diameter with backscattering cross-sectional area $\sigma(D)$ as given in (9). The integration is performed over all diameters present in the radar sample volume. The radar reflectivity factor Z provides a measure of a storm's intensity and is related to the received power P_r through

$$Z_e = P_r \frac{r^2}{C|K_w|^2} \quad (12)$$

obtained by combining equations (10) and (11) and assuming the radar is illuminating raindrops. Because in practice we cannot be sure that the radar is looking at spherical particles or raindrops only, Z is more appropriately termed equivalent radar reflectivity factor Z_e. Z (or Z_e) is given in units of $\mathrm{mm}^6\,\mathrm{m}^{-3}$, but is often expressed in logarithmic terms as $\mathrm{dBZ} = 10\log(Z)$. Small raindrops ($D \leq 1\,\mathrm{mm}$) qualify as spherical particles, whereas bigger raindrops appear on average more elliptical, with a horizontal axis exceeding the corresponding vertical axis. The effect of elliptical drops on the radar parameters is discussed in the following paragraphs.

Electromagnetic waves are polarized (i.e. the coupled electric and magnetic fields occupy well-defined orthogonal planes) perpendicular to the direction of propagation of the wave. The radar may transmit (and receive) either linear or circular polarized waves depending on the hardware. For simplicity, we concentrate on linear polarized electromagnetic waves only and, more specifically, on waves that are either horizontally (subscript h) or vertically (subscript v) polarized. Using the Rayleigh–Gans theory enables calculation of the backscattering cross sections for ellipsoids with minor axis a (vertical) and major axis b (horizontal) as

$$\sigma_{h,v} = \frac{\pi^5}{9\lambda^4} \left| \frac{m^2 - 1}{1 + (m^2 - 1)A_{h,v}} \right|^2 D_{\mathrm{eq}}^6 \quad (13)$$

where D_{eq} is the diameter of a spherical drop of an equivalent volume, and A_h and A_v are geometrical (shape) factors defined as

$$A_v = 1 - 2A_h = \frac{1}{\varepsilon^2} \left\{ 1 - \sqrt{\frac{1 - \varepsilon^2}{\varepsilon^2}} \frac{1}{\sin(\varepsilon)} \right\} \quad (14)$$

with the eccentricity ε of the elliptical cross section given by $\varepsilon^2 = 1 - (a/b)^2$. Note that for $a/b \to 1$ (spherical), the geometric factors become $A_h = A_v \to 1/3$ and the backscattering cross-sectional areas $\sigma_h = \sigma_v \to \sigma$ as shown in (9). On the basis of this analysis, a *differential reflectivity* Z_{DR} (in decibels or dB) can be defined as

$$Z_{\mathrm{DR}} = 10\log\left(\frac{Z_{hh}}{Z_{vv}}\right) = 10\log\left\{ \frac{\int N(D_{\mathrm{eq}}) \left| \frac{m^2 - 1}{1 + (m^2 - 1)A_h} \right|^2 D_{\mathrm{eq}}^6\,\mathrm{d}D}{\int N(D_{\mathrm{eq}}) \left| \frac{m^2 - 1}{1 + (m^2 - 1)A_v} \right|^2 D_{\mathrm{eq}}^6\,\mathrm{d}D} \right\} \quad (15)$$

where Z_{hh} and Z_{vv} are the radar reflectivity factors with horizontal and vertical polarization, respectively. Note, the double subscript of Z denotes the polarization of transmission (first index) and reception (second index) – for example, Z_{hh} is the (co-polar) radar reflectivity factor obtained for both horizontally polarized transmission and reception. The differential reflectivity Z_{DR} provides a reflectivity-weighted measure of the mean shape of particles in the radar sampling volume. The physical particle shape becomes particularly visible for liquid (raindrops) or wet-coated (melting) precipitation particles. As particle density and complex refractive index m decrease, however, the true particle shapes may no longer be revealed; for example, Z_{DR} is approximately zero for dry, low-density ice particles (snow). The measurement of differential reflectivity Z_{DR} typically requires a radar to be capable of switching the polarization of the transmitted wave from pulse to pulse. The latest technological developments now allow for simultaneous transmission (and reception) at different polarizations.

The polarization properties of radar waves reflected by precipitation particles (or other targets) may be altered by the backscattering process from nonspherical or canted particles. It is of interest, therefore, to assess how much of the transmitted signal power is depolarized by backscattering from targets. This has led to consideration of the *linear depolarization ratio* L_{DR} (in dB), defined as the ratio of the received cross-polar signal power (Z_{hv} or Z_{vh}) to the co-polar power (Z_{hh}) either by

$$L_{\mathrm{DR}} = 10\log\left(\frac{Z_{hv}}{Z_{hh}}\right) \quad (16)$$

or

$$L_{\mathrm{DR}} = 10\log\left(\frac{Z_{vh}}{Z_{hh}}\right) \tag{17}$$

depending on the radar hardware. Formulation (16) is appropriate for a radar that cannot switch polarizations from pulse to pulse, while (17) requires this capability. The cross-polar (depolarized) signals are usually much lower than the co-polar signals yielding very small L_{DR} values. Elevated L_{DR} values, indicating a significant signal depolarization, are typically seen for wet tumbling oblate particles, such as melting snowflakes or hail, and ground clutter echoes.

The amount of radar echo fluctuations from pulse to pulse contains meteorologically useful information. The (zero-lagged) correlation between the time series of successive estimates of Z_{hh} and Z_{vv}, termed the *co-polar cross-correlation* $|\rho_{hh,vv}(0)|$, depends on particle shape, variability (e.g. breadth of axis ratio distribution), and orientation (canting). $|\rho_{hh,vv}(0)|$ varies from 0 to 1, with $|\rho_{hh,vv}(0)| = 1$ if all particles were identical. A reduced correlation (< 0.9) results from a mixture of particle types and shapes within the radar pulse volume such as seen in the melting layer, where melting snow aggregates and raindrops coexist. Irregularly shaped hydrometeors (e.g. hail) and particle canting angles having a probability distribution of finite width can cause lower co-polar cross-correlation signatures as well.

Doppler technology provides ground for many more radar parameters. In general, the nth moment of the (reflectivity-weighted) Doppler velocity spectrum is given by

$$\langle v^n \rangle = \frac{\int N(D)D^6 v^n(D)\,\mathrm{d}D}{\int N(D)D^6\,\mathrm{d}D} \tag{18}$$

where $v(D)$ is the size-dependent fall velocity component of particles in the direction of the radar, containing also a radial component of the three-dimensional wind field. Two widely used radar parameters are the *mean Doppler velocity* $\langle v \rangle$, the first moment of the Doppler velocity spectrum, defined as

$$\langle v \rangle = \frac{\int N(D)D^6 v(D)\,\mathrm{d}D}{\int N(D)D^6\,\mathrm{d}D} \tag{19}$$

and the *Doppler spectral width* σ_{vel} defined as

$$\sigma_{\mathrm{vel}} = \sqrt{\langle v^2 \rangle - \langle v \rangle^2} \tag{20}$$

The former indicates the (reflectivity-weighted) mean radial velocity of the hydrometeors in the radar sampling volume, while the latter provides a measure of the associated velocity dispersion and turbulence. In addition, both the mean Doppler velocity and Doppler spectral width contain valuable information for detection of (stationary) ground returns seen by ground-based radar, aiding data quality control procedures.

For Doppler radar that also incorporate polarimetric flexibility, valuable information may be obtained by measuring the difference in phase of the backscattered signals, the so-called *differential phase*

$$\phi_{\mathrm{DP}} = \phi_{hh} - \phi_{vv} \tag{21}$$

where ϕ_{hh} and ϕ_{vv} are the co-polar phase shifts at horizontal and vertical polarization accumulated over the round trip from the radar to the targets and back. At 10-cm wavelength, phase differences between horizontally and vertically polarized signals accumulate primarily as a result of the forward scattering associated with the radar signal propagating through precipitation composed of nonspherical particles. Differences in phase shifts typically grow monotonically with increasing distance from the radar and become apparent especially in intense rainfall. The range derivative of ϕ_{DP} is called the (two-way) *specific differential phase* K_{DP} (in $\mathrm{deg\,km^{-1}}$) and defined by

$$K_{\mathrm{DP}} = \frac{\phi_{\mathrm{DP}}(r_2) - \phi_{\mathrm{DP}}(r_1)}{r_2 - r_1} \tag{22}$$

where $\phi_{\mathrm{DP}}(r_1)$ and $\phi_{\mathrm{DP}}(r_2)$ are the differential phase values at range r_1 and r_2, respectively. K_{DP} shows promise for radar rainfall estimation, because it is approximately proportional to the product of liquid water content times median diameter of the raindrops present in the radar sampling volume (which is closely related to the rainfall intensity). Phase measurements are attractive because they are immune to radar calibration problems and unaffected by signal attenuation. Moreover, they appear little affected by the presence of randomly tumbling hail or partial blocking of the radar beam. However, phase differences between horizontally and vertically polarized radar returns are typically small (much smaller than phase shifts caused by target motions relative to the radar), causing ϕ_{DP} measurements to be noisy. In order to reduce uncertainty, K_{DP} should be estimated from measurements spaced over many (typically tens of) radar resolution volumes in radial direction, depending on the precipitation intensity.

Meteorological Interpretation

The *amplitude, phase*, and *polarization* of backscattered signals received by the radar are key measures containing information on the illuminated targets, such as the *size distribution* of the precipitation particles within the radar sampling volume, particle *shape* (axis ratio) and

orientation, thermodynamic *phase* (solid, liquid, or a mixture of both) and *density*, and particle *fall speed* and *behavior* (spin, wobble). In order to interpret radar signals in meteorologically useful terms, one must consider the radar signal as well as the environment in which it is obtained. For example, environmental conditions (such as temperature, humidity, wind, and turbulence, or storm electric fields) may affect both the propagation of the radar signal and the particles' shape, phase, fall behavior, and orientation.

For all practical purposes, interpretation of radar signals *relies on many assumptions* due to a lack of detailed information about the precipitation particles and environmental conditions. Nonetheless, radar observations have yielded unprecedented insight into storm structure and precipitation formation processes and have assisted in measuring rainfall on large space and time scales. The radar reflectivity factor Z provides a direct measure of the intensity of precipitation (e.g. 10, 25, 40, and 55 dBZ correspond roughly to rainfall rates of 0.1, 1, 10, and 100 mm h^{-1}, respectively; see section on precipitation measurement for details) and has been instrumental in revealing details about the three-dimensional structure and organization of clouds and precipitating storm systems. Radar observations build the basis for characterization of different types of precipitation (convective versus more widespread or stratiform rainfall) and assess their potential to cause severe damage (e.g. heavy rainfall, hail, strong winds, and tornadoes). Moreover, elaborate analyses of the three-dimensional Doppler velocity field may be used to retrieve information about the thermodynamic properties of weather systems. The purpose of this section is to highlight some of the benefits of various polarimetric and Doppler parameters for characterizing meteorological and nonmeteorological radar echoes.

Figures 1 and 3 provide a horizontal, multiparameter depiction of convective storm systems developing near the Colorado–Kansas border. A series of convective storms evolving along a frontal boundary (visible as

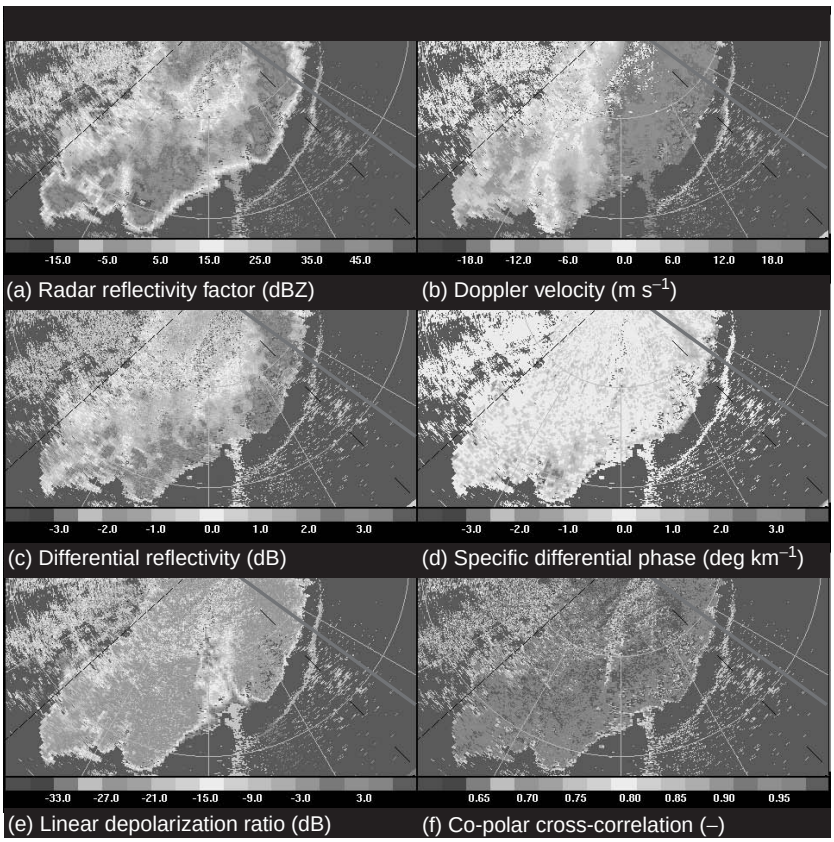


Figure 1 Pseudohorizontal cross section (surveillance scan at 0.5 degree elevation) through a series of convective storms developing at the Colorado–Kansas border, as observed by the National Center for Atmospheric Research (NCAR) multiple-polarization S-Pol radar on 11 June 2000 at 2355 UTC. Shown are (a) radar reflectivity factor Z, (b) radial Doppler velocity $\langle v \rangle$, (c) differential reflectivity Z_{DR}, (d) specific differential phase K_{DP}, (e) linear depolarization ratio L_{DR}, and (f) co-polar cross-correlation $|\rho_{hh,vv}(0)|$. Range rings are 50 km apart, and the red-outlined azimuth indicates the direction of the vertical cross section shown in Figure 2. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

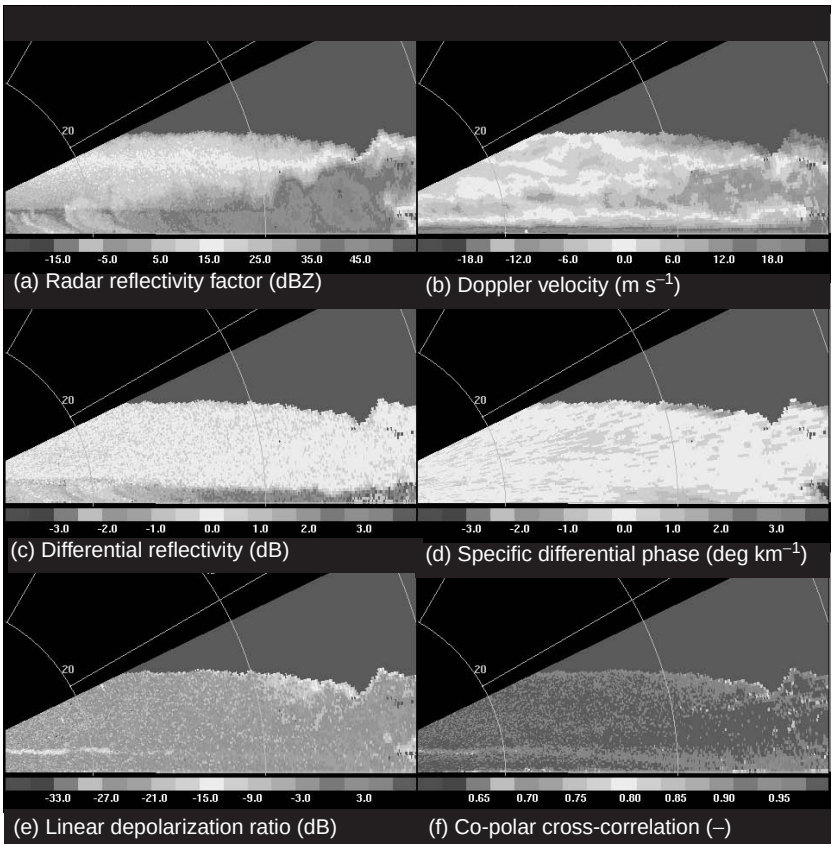


Figure 2 Vertical cross section in the direction of 126 degree azimuth through storms shown in Figure 1. The same parameters are shown as in Figure 1. Range rings are 20 km apart. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

thin low-intensity line ahead of the convective storms) is shown in Figure 1. The corresponding vertical cross section through that storm system is shown in Figure 2. This type of organized convection is also known as *squall line*, exhibiting a leading line of intense and deep convective cells, trailed by more moderate widespread rainfall with a layered (stratified) structure – hence the name stratiform precipitation. The thin horizontal line of enhanced backscatter seen approximately 2.5 km above ground in the vertical cross section (to the left in Figure 2a) marks the so-called *bright band* that indicates the melting layer. This enhancement in radar reflectivity factor is a consequence of the increased likelihood of aggregation of snow particles near the 0 °C level (resulting in larger particles) and the approximately a factor-of-five increase in backscattering intensity (caused by a larger complex refractive index m) as the particles begin to melt and become wet coated. The large snow aggregates will eventually collapse into smaller raindrops upon complete melting. Moreover, because raindrops fall faster than snowflakes, these particles tend to spread out quickly in the vertical decreasing the precipitation particle number density in the radar sampling volume and thus, combined with the reduced particle size, decrease the backscattered signal strength. Figure 2(a) shows fall-streaks within the stratiform portion of precipitation visualizing the vertical shear in the environmental flow, as depicted in Figure 2(b) (note the red colors near the ground indicating flow away from the radar and in green above that flow towards the radar).

The storm depicted in Figures 3 (horizontal cross section) and 4 (vertical cross section) shows a severe, potentially long-lasting convective storm *(supercell)* that may produce heavy rainfall and damaging hail, strong winds, and often tornadoes. This storm was also triggered by a frontal boundary visible ahead of the convective storm. The vertical cross section of this supercell storm (Figure 4a) is markedly different from the vertical cross section through the squall line (Figure 2a). Supercells exhibit a characteristic overhanging intense updraft core (shown in gray indicating radar reflectivity factors in excess of 57 dBZ) where large particles may be suspended and carried aloft (note the storm-relative upsloping flow into that core of high reflectivity in Figure 4b), and potentially get recycled. In this case, precipitation particles grow through collection of supercooled

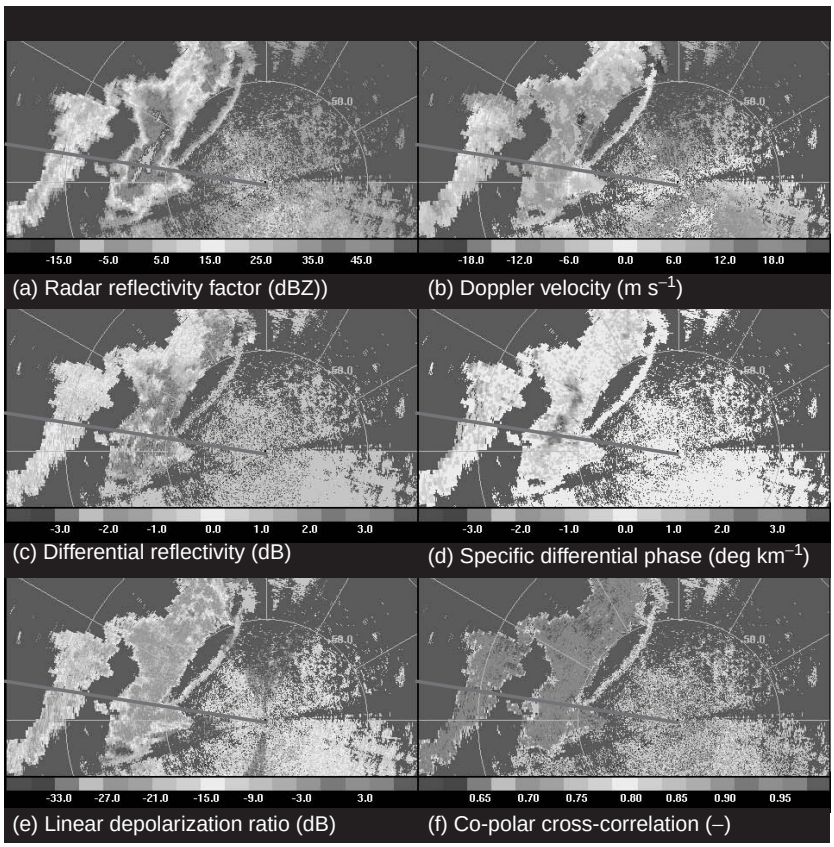


Figure 3 Pseudohorizontal cross section (surveillance scan at 0.5 degree elevation) through convective storms developing at the Colorado–Kansas border, as observed by the NCAR multiple-polarization S-Pol radar on 22 June 2000 at 2315 UTC. The same parameters are shown as in Figure 1. Range rings are 50 km apart, and the red-outlined azimuth indicates the direction of the vertical cross section shown in Figure 4. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

water droplets that freeze onto the already frozen ice particles (a process called *accretion* or *riming*) and thus may grow into high-density graupel or large hailstones. Details about the dynamics and microphysics of such storms may be found in textbooks by Houze (1993) or Doswell (2003).

In contrast to the radar reflectivity factor, the differential reflectivity Z_{DR} is independent of the radar constant (e.g. radar calibration) and is a good measure of the (reflectivity-weighted) mean axis ratio of particles. This parameter is not very sensitive to the particle size distribution. As the density and index of refraction of the particles decrease, the radar reflectivity factor (at any polarization) becomes weaker and less sensitive to shape and orientation, causing the differential reflectivity to approach zero. Therefore, Z_{DR} provides valuable information to distinguish solid (ice) from liquid (rain) particles and aids in quantitative rainfall estimation. The largest values of Z_{DR} (3–5 dB) are associated with individual big raindrops falling from the leading edge of storms or may be found towards the bottom of the melting layer in stratiform precipitation, where a high refractive index is compounded by a large axis ratio of water-coated (wet), melting snowflake aggregates just before collapsing into raindrops.

The big raindrop Z_{DR} signature can be clearly seen in the horizontal cross sections through the squall line (Figure 1c) and the supercell (Figure 3c), associated with regions of high value of radar reflectivity factor. In the vertical cross sections, the distinction between ice (low Z_{DR} values) and rain (moderate to high Z_{DR} values) is very pronounced, and a bright band signature is also recognizable in the stratiform portion of the squall line precipitation (Figure 2c). A core of low Z_{DR} value all the way to the ground is visible in the vertical cross section through the supercell storm (Figure 4c, range 57 km). This low Z_{DR} core associated with high radar reflectivity factor is characteristic for significant hail falling out of the storm. The high Z_{DR} values seen under the overhang, in contrast, are likely produced by smaller hail and graupel that melt and reach the surface as big raindrops. Note also the velocity folding signature at the top of the anvil in Figure 4(b), where the colors go from blue through purple into the red and yellow shades.

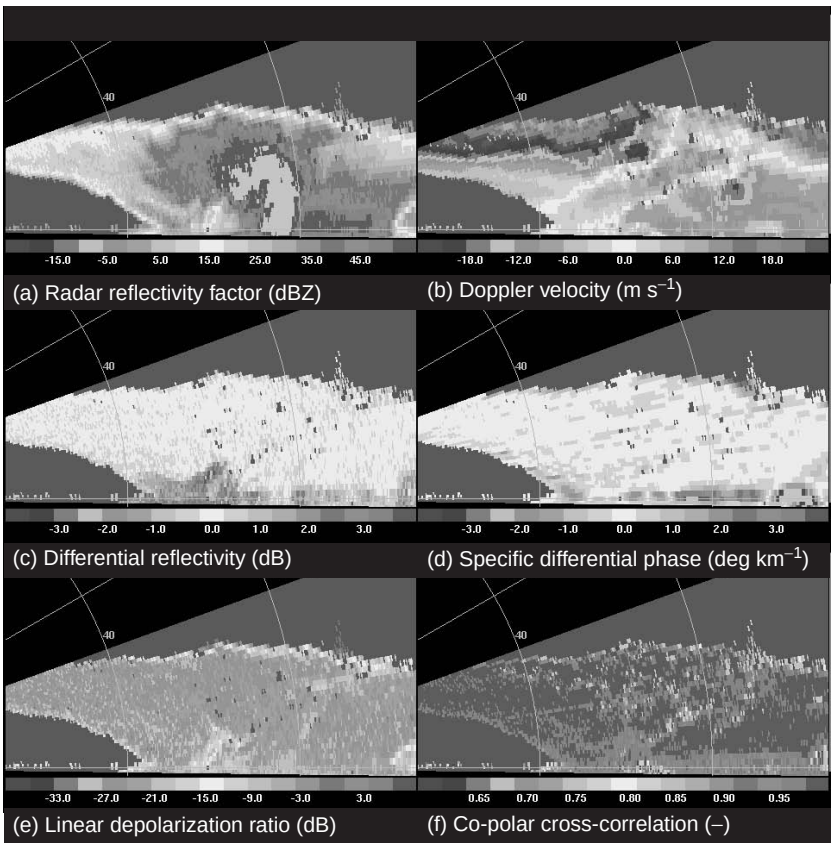


Figure 4 Vertical cross section in the direction of 279 degree azimuth through storms shown in Figure 3. The same parameters are shown as in Figure 1. Range rings are 20 km apart. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Observations and simplified analytical analyses have revealed that increased linear depolarization ratios L_{DR} occur when spheroidal hydrometeors fall with their major or minor axis neither aligned nor orthogonal to the polarization of the transmitted radar signal. Moreover, irregular-shaped hydrometeors and wobbling high-density particles that display a distribution of canting angles tend to cause signal depolarization, especially when they are wet. The lowest measurable signal depolarization is typically on the order of −30 dB or less. Elevated L_{DR} values (−20 dB or higher) are seen for wet ice particles in the melting layer or for large wet hail due to the increased refractive index. The linear depolarization ratio therefore appears to be a good indicator for the fall behavior of particles and particularly for wet (melting) ice particles.

The melting signature of L_{DR} can easily be recognized in the vertical cross sections of the squall line's stratiform precipitation (Figure 2e). The highly elevated L_{DR} values under the overhang of the supercell (Figure 4e) suggest melting of smaller hail and graupel particles (consistent with Z_{DR} features), although a potential contamination from antenna sidelobe effects in the low-reflectivity zone may have to be considered. A melting signature is also visible in the co-polar cross-correlation $|\rho_{hh,vv}(0)|$, for example, associated with the melting layer depicted in the vertical cross section of the squall line (Figure 2f), where partially melted large snow aggregates coexist with smaller raindrops resulting from ice particles that have already completely melted.

Radial Doppler velocity has been widely used to derive vertical profiles of the horizontal wind field, detect shearing winds, storm outflow, and frontal boundaries, and strong rotational signatures such as tornadoes. Doppler spectral width is used less commonly, but may indicate strong turbulence. Analyses based on a single Doppler radar may have to rely on significant assumptions about the homogeneity of the wind field, while data collected by multiple Doppler radar illuminating the same volume in space and time can yield quite detailed information about the three-dimensional wind (and thermodynamic) field. Differential phase and especially the specific differential phase observations have shown some value for rainfall measurement, particularly for higher intensities (see the associated signatures in Figures 1d and 3d, and Figures 2d and 4d, respectively), in the presence of hail within the radar sampling volume, or regions of partial beam blockage.

Table 2 Typical radar parameter values for various types of precipitation particles and nonmeteorological targets as observed at 10-cm wavelength

Particle type	Z_{hh} [dBZ]	Z_{DR} [dB]	$\lvert\rho_{hh,vv}(0)\rvert$	K_{DP} [deg km^{-1}]	L_{DR} [dB]
Drizzle	<25	0	>0.99	0	< −34
Rain	25 to 60	0.5 to 4	>0.97	0 to 10	−34 to −27
Snow, dry low density	<35	0 to 0.5	>0.99	0 to 0.5	< −34
Crystals, dry high density	<25	0 to 5	>0.95	0 to 1	−34 to −25
Snow, wet melting	<45	0 to 3	0.7 to 0.95	0 to 2	−18 to −13
Graupel, dry	40 to 50	−0.5 to 1	>0.99	−0.5 to 0.5	< −30
Graupel, wet	40 to 55	−0.5 to 3	>0.99	−0.5 to 2	−25 to −20
Hail, small <2 cm wet	50 to 60	−0.5 to 0.5	>0.95	−0.5 to 0.5	< −20
Hail, large >2 cm wet	55 to 70	< −0.5 to >3	>0.96	−1 to 1	−15 to −10
Rain and hail	50 to 70	−1 to 1	>0.9	0 to 10	−20 to −10
Ground clutter	10 to 70	−5 to 5	<0.7	0	> −10
Clear air and insects, birds	<25	−3 to 10	<0.5 to 0.8	0	> −10

Besides precipitation echoes, Figures 1 and 3 also show radar returns from clear air and ground targets. For example, ground clutter echoes are recognizable behind the squall line storm (top left corner of panels) by their moderate to high yet noisy reflectivity signature (Figure 1a) and zero radial velocity (Figure 1b). The weak clear air returns from a deep boundary layer depict the approximately south-southeasterly environmental flow field (Figure 3b) ahead of the supercell storm – note the change in color at the radar location from green (towards radar) to yellow (away from radar). There are also some embedded ground clutter echoes recognizable within 25-km range from the radar by their elevated backscatter (Figure 3a) and corresponding zero radial velocity (Figure 3b). Moreover, note the values of differential reflectivity ($Z_{DR} > 4$ dB, shown in gray), linear depolarization ratio ($L_{DR} \geq -10$ dB), and co-polar cross-correlation ($|\rho_{hh,vv}(0)| \ll 0.9$) within the boundary layer that differ markedly from the signatures associated with precipitation echoes. These radar signatures are typical for boundary layer air that is heavily contaminated by insects.

Values that are commonly observed for these radar parameters are compiled in Table 2. This information can be used to design algorithms for precipitation particle type identification and recognition of nonmeteorological radar echoes. Applications thereof range from cloud physics to rainfall estimation and from radar data quality control to hazardous weather detection for aviation safety, and radio communication. In addition, entomology and ornithology may benefit from polarimetric radar measurements.

PRECIPITATION MEASUREMENTS AND ASSOCIATED UNCERTAINTIES

Radar Rainfall Measurement

The radar reflectivity factor Z (mm^6 m^{-3}), defined by (11), is highly correlated with the rainfall rate R (mm h^{-1}), expressed as

$$R = \frac{6\pi}{10^4} \int N(D) D^3 v(D)\, \mathrm{d}D \tag{23}$$

where $v(D)$ represents the vertical fall velocity (m s^{-1}) of a raindrop of size D (mm). Both Z and R depend on the raindrop size distribution $N(D)$ (mm^{-1} m^{-3}), albeit weighted by a different power of the drop diameter. Raindrop spectra can be described by a gamma distribution of the form

$$N(D) = N_0 D^{\mu} \exp(-\Lambda D) \tag{24}$$

where N_0, μ, and Λ are three coefficients that scale the number concentration of drops, curvature (shape), and slope of the drop size distribution. The relationship between the radar reflectivity factor Z and rain rate R is conveniently expressed by a power-law form

$$Z = \alpha R^{\beta} \tag{25}$$

The coefficients α and β depend on the raindrop size distribution and range approximately $100 < \alpha < 1000$ and $1 < \beta < 1.8$. Typical values are $\alpha = 300$ and $\beta = 1.5$. For certain rainfall conditions, these coefficients can be interpreted in terms of the number density of drops and a characteristic drop size. However, for most situations the relationship (25) is merely of a statistical nature – that is, $\langle Z \rangle = \alpha \langle R \rangle^{\beta}$ where the $\langle\rangle$s denote expectations. This is particularly true for empirically derived relationships between radar reflectivity factor and rainfall rate. Typically, α and β vary from storm to storm, but can vary also within storms. Convective rainfall, for example, exhibits significantly different drop size distributions than stratiform rainfall, as reflected by the drop number density (Figure 5c) and mean drop size (Figure 5d). Distinct differences in raindrop spectra have also been found between oceanic and continental precipitation regimes. This great variability in

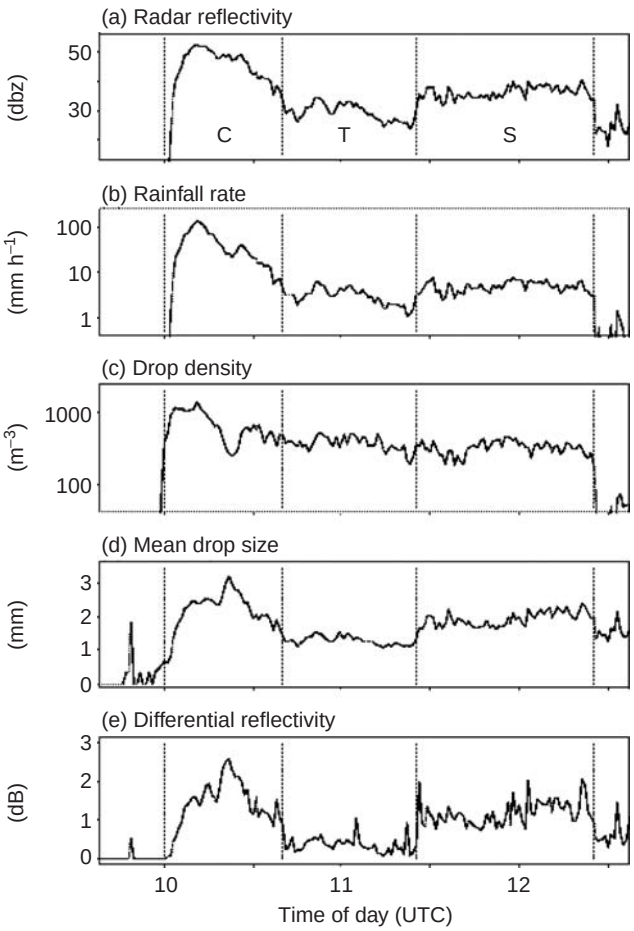


Figure 5 Depiction of a squall line moving over the Goodwin Creek research watershed in northern Mississippi on 27 May 1997. Shown are (a) the radar reflectivity factor, (b) rainfall rate, (c) drop number density, (d) mass-weighted mean drop size, and (e) differential reflectivity, as derived from raindrop size distributions collected by a Joss-Waldvogel disdrometer. This storm was characterized by three distinct phases: a convective leading line (C), a period of stratiform rainfall (S), and a transition (T) from one into the other

drop size distribution resulted in a myriad of $Z-R$ relations reported in the literature (e.g. see Battan, 1973) and may cause appreciable uncertainty (30–50%, sometimes more) in radar-based rain rate estimates. In addition, the conversion of the radar-measured reflectivity factor (a state parameter) into a rainfall rate (a flux parameter) is affected by vertical air motions. For example, an updraft of $2\,\mathrm{m\,s^{-1}}$ may cause a 50% overestimate and, conversely, a downdraft of $2\,\mathrm{m\,s^{-1}}$ may result in a 25% underestimate of the rainfall rate. Because regions of significant up- and downdrafts tend to be spatially and temporally limited, however, such air-motion related uncertainties are often neglected.

For practical purposes, especially operational applications of radar rainfall estimation, a climatological (i.e. default) $Z-R$ relationship is often used to retrieve rain rates from the observed radar reflectivity factor, and the radar-estimated rainfall amounts are subsequently adjusted based on contemporaneous rain gauge observations. This adjustment should take place only after all other corrections (e.g. calibration, attenuation, vertical structure) have been accommodated as necessary or possible. Various methods have been proposed to accomplish this merging of radar and rain gauge information, and it may be done on different space and time scales depending on the application. However, comparably little difference remains among the results based on various adjustment procedures once the mean difference (bias) between the collocated radar rainfall estimates and the gauge observations has been removed. The effect of space and time differences in the radar and gauge observations is very significant on short time scales but diminishes with increasing averaging (i.e. time integration). Under good (i.e. research) conditions, the root-mean-square (*RMS*) differences between gauge-adjusted radar-estimated and gauge-measured rainfall amounts may be as low as 10–20% on a storm-total or daily time scale. The *RMS* differences for operationally obtained radar rainfall estimates, in contrast, may amount to a factor of two or three. *RMS* differences typically increase with increasing spatial and temporal variability of precipitation, and decreasing time integration.

The advent of polarimetric radar observations has enabled new approaches to the radar rainfall measurement. The promise has been that additional radar parameters are obtained that yield information useful for reducing the uncertainty compared to rainfall estimates based on a single parameter (i.e. Z). For example, the combination of radar reflectivity factor Z and differential reflectivity Z_{DR} observations allows for estimation of two coefficients of the raindrop size distribution (24): Z_{DR} provides an estimate of the mean raindrop size in the radar sampling volume (see Figures 5d and 5e), while Z assists in deriving the drop number density. This appears to work particularly well under the assumption of exponential ($\mu = 0$) raindrop size distributions, but has also been used for gamma distributions, although the distribution shape factor μ must be specified (values of μ typically range from 2 to 10). Recent results based on using an additional constraining relationship between the curvature μ and slope Λ of gamma raindrop size distributions demonstrate that pretty reliable raindrop spectra and associated rainfall estimates can be obtained.

Alternatively, the specific differential phase K_{DP} has shown promise for rainfall estimation, especially for high rain rate conditions, when randomly tumbling hail may be present in the radar sampling volume, or under partial beam blockage. A combination of three approaches may be applied, therefore, where rainfall is estimated from radar reflectivity factor Z, differential reflectivity Z_{DR}, and

specific differential phase K_{DP} with increasing weight on the polarimetric radar parameters for increasing precipitation intensity. The respective empirical relationships, however, contain coefficients that may need to be tuned for a given situation, similar to the coefficients in (25). Moreover, the polarimetric radar parameters Z_{DR} and K_{DP} are not fully independent and rely on a number of assumptions about the raindrops, such as their shape with size (axis ratio), orientation and oscillation about the mean state, and fall behavior in a highly dynamic environment. For rainfall estimation purposes, therefore, the added complexity of the measurement and the achievable accuracy may offset the potential benefits of the additional information available. For example, measurement accuracies of $\leq$1 dBZ for the radar reflectivity factor Z, 0.1–0.2 dB for the differential reflectivity Z_{DR}, and $\sim$0.5–1deg km^{-1} for the specific differential phase K_{DP} are needed for rainfall rate estimates to be accurate to approximately 25%. Such accuracies for Z_{DR} and particularly K_{DP} are difficult to achieve. Nonetheless, these parameters may provide valuable information to constrain the rainfall retrieval. Moreover, the polarimetric radar capability pays off in terms of data quality control (e.g. radar calibration, attenuation correction, ground return detection) and identification of precipitation type (including hail discrimination and bright band detection).

Uncertainty of Radar Rainfall Estimates

Variability in raindrop size distribution and vertical air motion cause uncertainty in radar-estimated rainfall amounts, as discussed in the previous section. Although this may be a significant uncertainty, other sources of uncertainty can be much larger. For example, no matter how sophisticated the radar measurements are, they are always affected by the propagation of electromagnetic waves and the *earth's curvature*. Significant gradients in atmospheric humidity, temperature, and pressure may result in nonstandard propagation of electromagnetic waves, causing the radar to "look" at a different altitude above the earth's surface than anticipated. Although the radar signal may intersect ground targets (e.g. trees, buildings, or the terrain) at known locations, *anomalous propagation* may significantly increase ground returns in unexpected places. Furthermore, ground clutter echoes caused by anomalous signal propagation may have an appearance similar to real precipitation echoes, though typically only the lowest elevation scans are contaminated. Automated recognition of such contamination in radar reflectivity data remains a challenge today. One of the great benefits of polarimetric radar information is its value for radar data quality control: nonmeteorological radar echoes (e.g. ground clutter, insects, or birds) are non-Rayleigh scatterers exhibiting signatures significantly different from precipitation (see Figures 1 and 3).

Even under standard propagation conditions, the radar beam widens with increasing distance from the radar (i.e. the radar sampling volume increases) and the center of the beam increasingly rises above the ground, eventually intercepting the melting layer and entering the ice phase of a precipitating cloud system. This may result in *inhomogeneous beam filling* effects or possibly the overshooting of shallow echoes at far ranges. Therefore, the radar information seen aloft must be related to surface rainfall by considering the *vertical structure* of precipitation. A coarse resolution of radar observations at greater ranges prevents obtaining vertical structure information locally though. In this case, one may resort to using a "typical profile" for the anticipated type of precipitation (e.g. convective or stratiform). The radar reflectivity can decrease appreciably with height (up to 8 dB km^{-1}), depending on the type of precipitation. Significant vertical gradients are typically observed in stratiform rainfall, where the reflectivity decreases rapidly with height above the melting layer (i.e. bright band). Weaker gradients are normally seen in deep and intense convective cores, although in a very moist environment effective growth of precipitation particles occurs at low levels that the radar may not capture at farther ranges. Significant evaporation below cloud base can result in negative gradients at lower levels. Accounting for the vertical structure (to the extent possible) when relating radar observations made aloft to surface rainfall is thus very important.

Other concerns relate to the visibility of the radar and the stability of the radar hardware. Major terrain features or obstructions close to the radar can block the radar's view thus creating blind sectors, where no observations can be obtained – at least at lower elevations. Clever placement of the radar is key to guarantee coverage of a desired area although it may be difficult in complex terrain. *Visibility* is of greater importance in site selection than ground clutter concerns – modern signal processing techniques can be used to account for clutter problems, but little can be done to retrieve precipitation echoes in radar-blind sectors. In the latter case, one may have to resort to observations made at some altitude above the ground and relate them to surface rainfall by using information about the precipitation's anticipated vertical structure, as mentioned before.

Stability of the radar hardware is crucial to quantitative radar measurements. Systematic radar calibration offsets can be detected and corrected (e.g. through internal consistency checks using redundant polarimetric radar information), but if the hardware (e.g. transmitted power) varies with time, it is much more difficult. Modern radar systems, therefore, monitor the various components for quality assurance. Well-calibrated radar systems may be accurate to within 1–2 dB and achieve a measurement precision on the order of 1 dB or less. Note that for a typical $Z - R$ relationship, $Z = 300R^{1.5}$, 1.5 dB uncertainty in Z

translates into 1 dB uncertainty in R, which is approximately 25%.

In precipitation-free regions, attenuation due to atmospheric gases (oxygen and water vapor) should be considered. The magnitude of such attenuation accumulated over some distance may be comparable to the accuracy of the radar reflectivity measurements. *Attenuation* of the radar signal as it travels through rain is another source of uncertainty that is particularly relevant for shorter ($\leq$5 cm) wavelengths. Severe signal attenuation (>20 dB) can be caused by intense rainfall (>100 mm h^{-1}) observed at X-band wavelengths, especially when melting high-density ice is present in the radar sample volume. This is demonstrated, for example, in the vertical cross section of radar reflectivity factor (Figure 6b) through a squall line approaching the radar (the horizontal view is depicted in Figure 6a). The corresponding Doppler velocity (Figures 6c and 6d) and spectral width fields (Figures 6e and 6f) are displayed as well. Significant attenuation (and possibly shifts in signal phase) may also occur at the radar itself, when the radome (i.e. the antenna cover) is coated by a variably thick film of water during heavy rainfall. Weaker attenuation (<5 dB) by intervening rainfall may possibly be corrected, but wet radome attenuation is rather difficult to account for. Signal attenuation affects particularly the radar reflectivity factor measurements. Moreover, the difference in attenuation between horizontally and vertically polarized reflectivity observations can be significant at shorter wavelengths (e.g. C and X band) and thus affect the differential reflectivity measurements as well. These are some of the reasons why shorter wavelength radars are less often used for operational rainfall estimation.

If the received radar signals from melting particles or hail, or ground targets, are used for rainfall retrieval, a significant overestimation may result. In contrast, improper conversion of radar signals from the snow phase to rainfall will cause underestimation. (The quantitative measurement

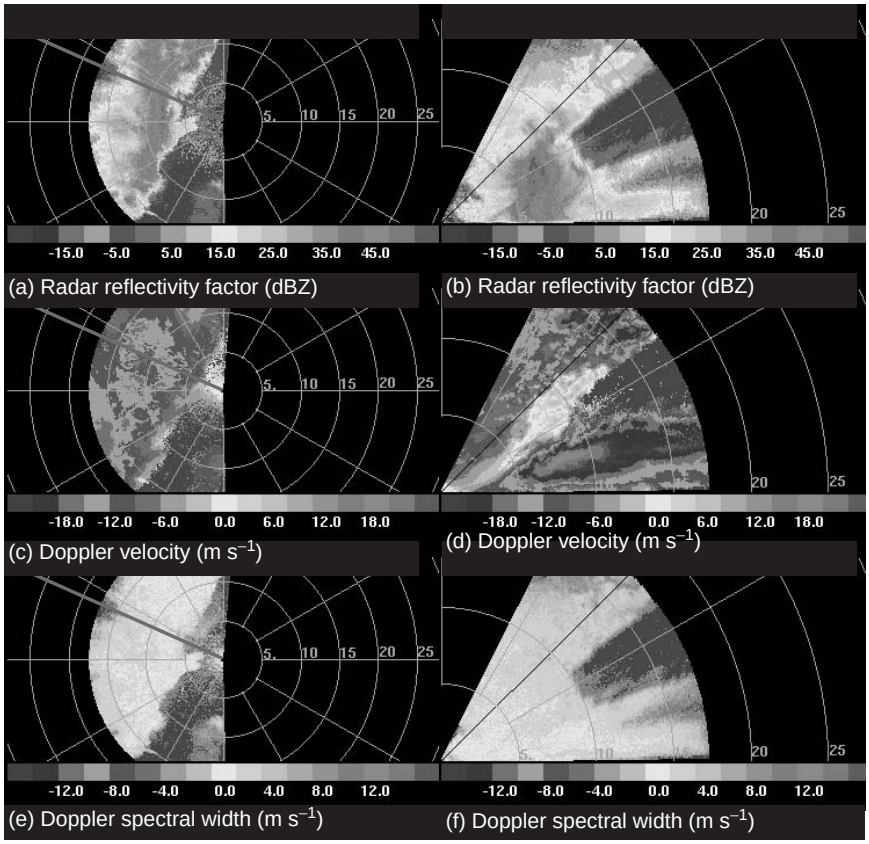


Figure 6 Horizontal (left panels) and vertical (right panels) cross sections through an intense line of convection passing over the Goodwin Creek watershed in northern Mississippi, as observed by the University of Oklahoma Doppler-on-Wheels (DOW) mobile X-band radar on 4 April 2001 at approximately 2314 UTC. Shown are the radar reflectivity factor Z (top panels), radial Doppler velocity $\langle v \rangle$ (middle panels), and Doppler spectral width σ_{vel} (bottom panels). Range rings are 5 km apart, and the red-outlined azimuth in the left panels indicates the direction of the vertical cross section shown in the right panels. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of snowfall is much more complex than that of rainfall, because of the variable density of ice particles and their great range of shapes and sizes.) The polarimetric capabilities of modern radar can alleviate some of these problems through a classification of the dominant precipitation particle type present in the radar sample volume. Other, less frequent problems can arise from nonmeteorological targets such as birds or artificial targets (e.g. chaff dispersed by military actions).

USE OF RADAR FOR HYDROLOGIC APPLICATIONS

Ground-based radar measurements have provided a wealth of information about precipitation systems, how they evolve and organize themselves. Advances in Doppler and polarimetric radar technology have yielded unprecedented insight into the dynamics and microphysics of precipitation formation. However, the technology and knowledge transfer from research to operational applications has been slow. It comes at no surprise, therefore, that most radar-based rainfall estimation still heavily relies on simple $Z-R$ relationships to obtain rainfall rate estimates from reflectivity observations. It is not uncommon for such rain rate estimates, made at spatial and temporal resolutions of $1-2$ km and $5-10$ min, to be uncertain by up to a factor of two or more. Operational radar rainfall estimation is particularly challenging, because only a limited amount of rain gauge information may typically be available in real time for comparison to the radar rainfall amounts.

On the other hand, radar measurements of rainfall have had a substantial impact on hydrologic applications, such as (flash) flood monitoring and warnings. Unlike typical coarse rain gauge networks, radar observations can reveal the significant variability of rainfall over spatial scales of only a few kilometers that is often associated with extreme weather events. The spatial and temporal organization of rainfall, and the motion of storms relative to watersheds play a fundamental role in the *hydrologic response* of a basin. For example, a large fractional coverage of high rainfall rates for durations corresponding to the basin response time will result in maximum river peak discharges. Also, storm motion relative to a catchment may either amplify or dampen a flood wave, depending on whether the storm moves down the basin or in upstream direction, respectively. Algorithms exist for operational identification and tracking of storm cells, and the resulting rainfall centroids and storm motion vectors may be overlaid on topographic and hydrologic information using GIS-based software. Moreover, radar rainfall estimates are ingested to spatially distributed hydrologic models for runoff and river discharge predictions.

Looking ahead, the anticipated upgrade of the U S network of operational weather radars to incorporate polarimetric capability will facilitate adaptation of advanced Doppler and polarimetric technology into the operational arena.

Acknowledgments

The writing of this review was supported by the National Science Foundation (NSF) through Grants EAR-9909696, ATM-9906012 and ATM-0223798, and the National Aeronautics and Space Administration (NASA) Earth Science Enterprise through Grants NAG5-9891 and NAG5-13624. The thoughtful comments and suggestions by Drs. Anthony Illingworth, Edward Brandes, and Remko Uijlenhoet are greatly appreciated.

REFERENCES

Atlas D. (Ed.) (1990) *Radar in Meteorology*, American Meteorological Society: Boston, p. 806.

Battan L.J. (1973) *Radar Observations of the Atmosphere*, University of Chicago Press: p. 324.

Bringi V.N. and Chandrasekar V. (2001) *Polarimetric Doppler Weather Radar*, Cambridge University Press: p. 662.

Doswell C.A. III (Ed.) (2003) *Severe Convective Storms*, *Meteorological Monograph No. 50*, American Meteorological Society: Boston, p. 576.

Doviak R.J. and Zrnić D.S. (1993) *Doppler Radar and Weather Observations*, *Second Edition*, Academic Press: p. 562.

Germann U. and Joss J. (2003) Operational measurement of precipitation in mountainous terrain. In *Weather Radar – Principles and Advanced Applications*, Meischner P.F. (Ed.), Springer: Heidelberg, pp. 52–77.

Houze R.A. Jr (1993) *Cloud Dynamics*, Academic Press: San Diego, p. 573.

Illingworth A.J. (2003) Improved precipitation rates and data quality by using polarimetric measurements. In *Weather Radar – Principles and Advanced Applications*, Meischner P.F. (Ed.), Springer: Heidelberg, pp. 130–166.

Lhermitte R.M. (2002) *Centimeter and Millimeter Wavelength Radars in Meteorology*, Lhermitte Publications: Miami, p. 550.

Rinehart R.E. (1999) *Radar for Meteorologists*, *Third Edition*, Rinehart Publications: Grand Forks, North Dakota, p. 428.

Skolnik M.I. (1988) *Introduction to Radar Systems*, *Second Edition*, 10th Printing, McGraw-Hill International Editions, Electrical Engineering Series: p. 581.

64: Satellite-based Estimation of Precipitation Using Microwave Sensors

GEORGE J HUFFMAN

Science Systems and Applications, Inc. and NASA/Goddard Space Flight Center Laboratory for Atmospheres, Greenbelt, MD, US

Satellite-borne sensors provide critical input to estimating precipitation over the vast majority of the Earth's surface that lacks adequate in situ observing systems. This article reviews the legacy of sensors, satellites, and data archives that have accrued throughout the Space Age, and that largely governs the precipitation estimates that are available to the hydrological community, both currently and over the long term. Tables of publicly available, quasi-global, satellite-based precipitation data sets are introduced, examples are provided, data quality is considered, and future prospects are outlined. One major theme throughout the article is to raise the key issues that users should consider in determining whether a particular set of precipitation estimates is appropriate for their particular application.

INTRODUCTION

Precipitation is one vital input to numerous hydrological studies and applications, yet extensive research is still needed to establish accurate global estimates of the mean and time-dependent record of precipitation (see Figure 1). Precipitation has strong small-scale variability and highly nonnormal statistical behavior, which makes it much more difficult to characterize than, for example, atmospheric temperature. Such attributes require frequent, closely spaced observations to accurately quantify the precipitation pattern, but these requirements cannot be met by surface-based observations over much of the globe, particularly in oceanic, remote, and developing regions. In response, atmospheric scientists have developed a variety of satellite-borne sensors to obtain the requisite information for estimating precipitation on a global basis. As researchers have gained experience with the individual sensors and the performance of single-sensor algorithms for estimating precipitation, they have increasingly moved to using combinations of sensors for improved accuracy, coverage, and resolution. This article focuses on publicly available, quasi-global, operational data sets that are the most likely to be widely useful for hydrological research and applications. None of the current data sets satisfies requirements for both fine-scale space/time detail and a long continuous record. Fortunately, recent and planned developments in the recovery of old data sets and additions to the fleet of satellite sensors promise to enrich the available precipitation database in the early twenty-first century.

SATELLITES, SENSORS, AND DATA

Satellites have been used as platforms for the remote sensing of hydrologically relevant data since the dawn of the Space Age. The first Television-Infrared Observation Satellite (TIROS-1) was launched by the United States in April 1960, providing the first global images of clouds. TIROS already exhibited many characteristics of modern meteorological satellites. Its orbital plane was nearly perpendicular to the Equator (a "polar orbit") and the orbital altitude was relatively low ("low earth orbit" or LEO), causing TIROS to circle the globe about every 90 min. From this vantage point, TIROS could only see a small portion of the globe at any one time, but viewed most of the Earth's surface over the course of 12 h (either on the ascending or descending part of the orbit). TIROS imaged the Earth in visible and infrared (IR) frequency bands that are not

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

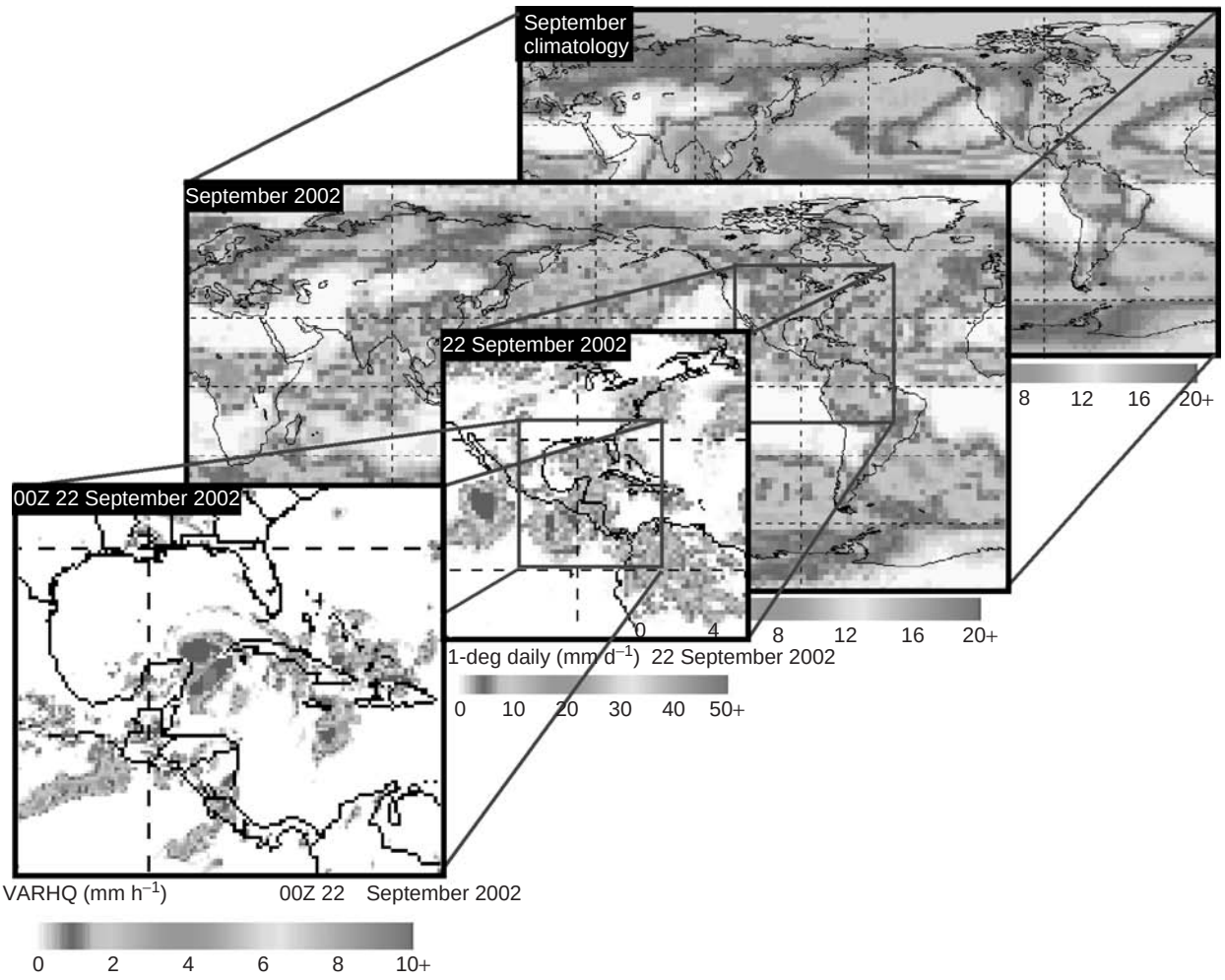


Figure 1 Examples of modern global precipitation estimates at various time resolutions: September Climatology from GPCP Monthly SG, September 2002 from GPCP monthly SG, 22 September 2002 subsetted from GPCP 1DD, and 00Z 22 September 2002 subsetted from TRMM 3B42RT. See Tables 2 and 3 for product details. Hurricane Isadore is located off the tip of the Yucatan peninsula at 00Z 22 September 2002. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

strongly affected by atmospheric effects, other than clouds and dust (so-called "window channels"). Finally, the sensor was "passive", recording the radiant energy that was emitted (IR) or reflected (visible) by the Earth's surface, then modified by a scene-dependent sequence of reflection (for visible), emission, and scattering by atmospheric constituents, including clouds, as the energy upwelled toward the satellite.

Another important milestone for precipitation work came with the launch of the first geostationary Earth orbit (GEO) meteorological satellite in 1974 by the United States. A GEO satellite orbits some 35 400 km above the Equator and takes a day to circle the Earth, so its motion is synchronous with a fixed point on the Earth's surface. Further, notable firsts include the first passive microwave-frequency sensor in 1972, and the first active microwave instrument (i.e. radar) in late 1997 on board the Tropical Rainfall Measuring Mission (TRMM). Interested readers may consult Kidd (2001) for an excellent summary of satellite sensor development related to precipitation. This five-odd decades of work has produced a legacy of sensors, satellites, and data archives that critically governs the precipitation estimates that are available to the hydrological community.

Sensors

Precipitation-related satellite sensor development work has historically focused on just a few channels in the electromagnetic spectrum (*see* **Chapter 47, Sensor Principles and Remote Sensing Techniques, Volume 2**):

IR

The IR channel around 11 microns is a window channel, typically showing the surface, except when clouds are present. Most clouds are optically thick in the IR, meaning that they are very effective at absorbing upwelling radiant energy, which is then reemitted with energy levels characteristic of the temperature of the emitters. This measured energy is termed the *brightness temperature* (T_b). As a result, IR mostly reveals the presence and temperature (and so, implicitly, the height) of clouds. (Typically, temperature is a decreasing function of height in the troposphere, so the IR-estimated cloud temperature can be related to cloud-top height.) For sufficiently large time-space averages, high-level cloudiness is well-correlated to precipitation in the tropics, but the correlation declines dramatically for averages covering less than $2.5° \times 2.5°$ of latitude/longitude and 12 h (Richards and Arkin, 1981; Arkin, 1979). Furthermore, outside the tropics, the occurrence of dense, nonprecipitating cirrus in frontal zones makes the IR-rainfall relationship considerably less reliable for estimating precipitation.

Visible

The visible channel is similar to IR in responding to the surface and cloud tops, but the radiant energy involved is reflected sunlight, rather than emission by the Earth's surface and atmosphere. This reflectance is a complicated function of the viewing and solar illumination angles, and is limited to daylight hours. At present, there are no routinely operational precipitation algorithms that use visible data.

Passive Microwave at ''Emission'' Frequencies

Channels located at passive microwave (PMW) frequencies at and below 37 GHz mostly respond to radiant energy emitted by the Earth's surface, and modified by cloud water and liquid hydrometeors within the intervening atmosphere. Information on the atmosphere's liquid water content is closely related to liquid precipitation at the surface, and is preferable to either IR or the PMW scattering channels (below). The main limitation at these frequencies is that a useful retrieval is only possible when the Earth's surface within the given satellite field of view is entirely covered by water, which has a relatively low and uniform emissivity. Any land or surface ice produces surface emissivities that are higher and exhibit such strong heterogeneities that quantitative precipitation estimates are not currently possible.

Passive Microwave at ''Scattering'' Frequencies

Channels located at PMW frequencies above 37 GHz exhibit emission signals similar to the lower frequency channels over nonprecipitating regions, but once clouds start to develop precipitation-sized frozen hydrometeors, the ice strongly scatters the upwelling radiant energy out of the line of sight of the sensor. This produces a characteristic low T_b embedded in a higher-T_b region. Frozen hydrometeors have a better instantaneous relationship to surface precipitation than IR T_b's, but the relationship is generally not as good as for the liquid hydrometeors sensed by emission channels. Scattering signals are not significantly degraded by land surface in the pixel's field of view, but surface ice (including frost) prevents reliable retrievals whenever such conditions occur.

Passive Microwave at ''Sounding'' Frequencies

The PMW channels chosen to retrieve soundings, that is, to estimate vertical profiles through the atmosphere, saturate in the atmosphere and (mostly) cannot see the surface. Thus, sounding channels do not suffer the same surface effects as the other PMW channels. Sounding channels generally employ frequencies close to water vapor absorption bands and respond to the presence of hydrometeors, making it possible to infer precipitation in snowy situations. Such channels began to enter operational use for estimating precipitation in the early 2000s.

Radar

In contrast to the integrated signal that passive devices sense, active devices such as radar send out short pulses of energy at one or more specified frequencies, creating return signals that have information about conditions at different distances from the remote sensor. Until recently, the power and antenna requirements for radars were not considered economically practical on meteorological satellites. The current state of satellite-borne radar technology only provides a narrow swath of data that is primarily useful for calibrating other sensors' precipitation estimates, for accumulating climatological statistics, and for providing detailed views of specific events that a radar overpass happens to catch.

Satellites

The legacy of satellite types and sensors is the second factor governing precipitation estimation. An unbroken record of IR data from LEO satellites is available from 1979 to the present, principally from the United States National Oceanic and Atmospheric Administration (NOAA) series, with contributions from China and Russia. IR observations from GEO satellites started in earnest in 1981, with nearly full coverage beginning in 1983 from an international constellation provided by Europe, Japan, and the United States The primary shortfall was a persistent gap over the Indian Ocean sector until mid-1998, and various temporary outages for other regions caused by the failure of individual sensors.

A nearly continuous record of PMW scattering and emission channels starts in mid-1987 with the United States Special Sensor Microwave/Imager (SSM/I), which has been carried aboard the Defense Meteorological Satellite Program (DMSP) platforms. Starting in the late 1990s

additional PMW sensors were launched, as well as the TRMM Precipitation Radar (PR).

All of the preceding development work has yielded an unprecedented set of global observations. Five GEO satellites – two Meteorological Satellites (METEOSAT) from Europe, two Geostationary Operational Environmental Satellites (GOES) from the United States, and one Geostationary Meteorological Satellite (GMS) from Japan – provide visible and IR data, as well as other channels. By international agreement, each satellite images the full disk of the Earth as seen from its position for each three-hourly synoptic time (00Z, 03Z, ..., 21Z), together with more-frequent images of operationally chosen sectors. As for the LEO satellites, two to three United States NOAA-series satellites provide visible, IR, and PMW sounding (Advanced Microwave Sounding Unit; AMSU) data. Two to three United States DMSP satellites provide ongoing SSM/I PMW coverage, with additional PMW coverage provided by research satellites, namely, the *TRMM Microwave Imager* (TMI) on TRMM and the Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) on the United States Aqua satellite. Finally, TRMM also hosts the PR. All of the LEO satellites provide near-global coverage from polar orbits, except the TRMM orbit, which only covers the latitude band 36 °N–36 °S for PR and 38 °N–38 °S for the wider-swath TMI.

One subtlety is that most of the LEO satellites are flown in sun-synchronous polar orbits, meaning that they overfly each point on Earth at roughly the same local a.m. and p.m. time each day. Such an orbit introduces a likely bias in estimating precipitation, since the diurnal cycle in precipitation varies with location, synoptic situation, and surface type, and typically is not a simple function of time of day. The alternative is a LEO satellite whose orbit precesses through the day, but this precession usually takes so many days to cover the diurnal cycle that stratifying the observations by time of day is only meaningful for accumulations over months of data. As an example, the TRMM satellite takes 46 days to precess through one day at any one location on the Earth's surface. Regardless of the orbit, it is estimated that eight evenly spaced samples are needed to adequately characterize the daily precipitation (Adams *et al.*, 2002). This can be achieved at any particular site by use of a GEO satellite or by a sufficiently large fleet of LEO satellites. Note that *global* coverage requires a constellation of satellites, whether GEO or LEO. In the case of PMW sensors, which are only available on LEO satellites at present and which have somewhat narrow swaths of data, the "sufficiently large" constellation translates to a requirement for eight satellites.

Data Archives

The third major factor governing remotely sensed estimates of precipitation is access to the data. At present, the bulk of the workhorse satellite observations, namely the PMW and IR, are posted on-line in near-real time or reproduced on tape within a month or so of the observation time. However, before the late 1990s, this was not the case for the IR, and it is still not entirely true for other channels. Instead, long-term operational precipitation algorithms depend on a legacy of subsetted and averaged IR data summaries, and the earlier the era, the less detailed the summary. For example, the Global Precipitation Climatology Project (GPCP) Version 2 monthly Satellite-Gauge (SG) product (Adler *et al.*, 2003) depends on:

- Outgoing Long-wave Radiation (OLR) values computed from LEO IR and then averaged to a $2.5° \times 2.5°$ grid for five-day (pentad) intervals during the period 1979–1985;
- GEO IR T_b data binned into 16 T_b intervals, most of them 5K wide, on a $2.5° \times 2.5°$ grid for five-day (pentad) intervals during the period 1986–1996, with fill-in by 3-T_b-bin histograms of LEO IR data; and
- GEO IR T_b data binned into 24 T_b intervals, most of them 5K wide, on a $1° \times 1°$ grid for individual three-hourly observation times during the period 1997–present, with a parallel grid of LEO IR-based GOES Precipitation Index (GPI; Arkin and Meisner, 1987).

See "Future Prospects" below for notes on possible improvements to the archive issue.

Strengths and Weaknesses

As noted above, GEO IR data exhibit excellent sampling, but are only weakly related to instantaneous precipitation rates and to precipitation outside the tropics. PMW data have an excellent correlation to precipitation over oceans and a good correlation over land, except in regions with frozen land surface, but PMW technology is currently limited to LEO satellites, thereby introducing temporal sampling issues. Radar data have both excellent correlation to precipitation and range-resolved information, but the very narrow data swath, power requirements, and cost severely limit the number of sensors ever likely to be flown at the same time.

In addition, all satellite-based precipitation estimates face some generic problems. Most importantly, the physical problem is under-determined, so that it is clear in advance that simply recording more observations from the same sensor cannot force the error in the estimate to zero. A closely related issue is inhomogeneity in the precipitation scene across an individual pixel, referred to as the beam-filling problem. Such inhomogeneities interact with the nonlinearities in the radiative processes to force the retrieved precipitation value for the pixel away from the true average. Finally, there are dropouts in the various estimates

due to issues in satellite operations, data transmission, and data archiving.

At this juncture, the reader should note that any real system for producing precipitation estimates is subject to weaknesses. Numerical models require computational approximations, most particularly the parameterizations for the precipitation microphysics; experience initialization errors; and can transfer errors from other parts of the model to the precipitation estimates. Surface-based radar suffers calibration uncertainties, beam blockage by surface objects, bright-band uncertainties, and anomalous propagation. Even rain gauges, which are usually considered the gold standard, carry a large list of caveats, led by the fundamental problem of converting point gauge values to areal averages. Other significant gauge problems include losses due to aerodynamics around the gauge orifice, evaporation, side-wetting, splashing, and unrepresentative siting.

This summary of strengths and weaknesses underscores the need for users to question the suitability of any precipitation data set for their particular use. Different sensors, both satellite- and ground-based, observe different aspects of precipitation. We now turn to the estimation schemes, which typically seek to emphasize strengths and minimize the weaknesses of the input data sets that they utilize.

TRANSFORMING OBSERVATIONS INTO ESTIMATES

Each type of sensor mentioned in the previous section has some advantage for estimating precipitation, and a number of algorithms have been developed to exploit that sensor's unique information. Increasingly, researchers have worked to combine observations or precipitation estimates from multiple sensors to achieve better accuracy and/or wider coverage. Of the large number of research algorithms that have been developed, only a few have found operational application on a quasi-global basis. The data sets thus generated are the most likely to be useful to the nonspecialist, and so the discussion here is limited to those archives with

Table 1 Summary of publicly available, quasi-operational, quasi-global precipitation estimates from a single sensor. Where appropriate, the algorithms applied to the individual input data sets are mentioned. The TMI GPROF and PR are also available as separate products from the GDAAC. ''Latency'' gives the typical interval between the end of the observational period and release of the product[a]

Algorithm	Input data	Space/time scales	Areal coverage/start date	Update frequency	Latency	Archive location[b]
GPROF	SSM/I	0.5°/orbit segments	Global – 70 °N–S/July 1987	Month	1 pentad	GDAAC (1)
	SSM/I	0.25°/orbit segments	Global – 70 °N–S/September 2002	3 h	3 h	GDAAC (1)
GPROF (3G68)	TMI	0.5°/h	Global – 37 °N–S/December 1997	Daily	4 days	NASA/GSFC TSDIS (1)
GPI	GEO-IR, LEO-IR	2.5°/pentad	Global – 40 °N–S/January 1986 – March 1997	N/A	N/A	NOAA CPC (1)
	GEO-IR, LEO-IR	1°/3 h	Global – 40 °N–S/October 1996	Month	1 week	NOAA CPC (1)
NESDIS High Frequency	AMSU	0.25°/day 2.5°/pentad, month 1.0°/pentad, month	Global/January 1999	Daily	4 h	NESDIS ORA
NESDIS/FNMOC SI	SSM/I	0.25°/day 1.0°/pentad, month 2.5°/pentad, month	Global/July 1987	Daily	6 h	NESDIS ORA
OPI	LEO-IR	2.5°/day	Global/January 1979	Daily	1 day	NOAA CPC (2)
PR Precip (3G68)	PR	0.5°/h	Global – 37 °N–S/December 1997	Daily	4 days	GDAAC (2)
Wilheit-Chang Statistical	SSM/I	2.5°/month	Global ocean – 60 °N–S/July 1987	Monthly	1 month	NASA/GSFC Code 614.3
Wilheit-Chang Statistical (3A11)	TMI	5°/month	Global ocean – 40 °N–S/January 1998	Monthly	1 week	GDAAC (2)

[a]For expansions, see Appendix 1.
[b]For details, see Appendix 2.

reasonable coverage and continuity. Kidd (2001) points to numerous other examples still in the research stage. For the sake of clarity and compactness, we summarize the publicly available, quasi-global precipitation estimates into: single-sensor satellite data sets (Table 1), combination satellite data sets (Table 2), and combination satellite data sets that incorporate gauge data (Table 3). While space limitations preclude a detailed discussion of each approach, the major themes and issues are outlined in the following subsections. See the Appendix for full expansions of the acronyms used in the tables to conserve space. The reader is referred to **Chapter 35, Rainfall Measurement: Gauges, Volume 1** for a similar discussion of rain gauge observations. Well-calibrated, carefully analyzed rain gauge data sets are critical to the development and routine calibration of remote sensing estimates of precipitation.

PMW-Based Data Sets

The various PMW algorithms may be characterized as predominately statistical, including the National Environmental Satellite Data and Information Service (NESDIS)/Fleet Numerical Meteorological and Oceanographic Center (FNMOC) and Wilheit-Chang algorithms, or physical, such as the Goddard Profiling algorithm (GPROF). In either approach, the under-determined problem is implicitly closed by assuming climatological behavior at critical steps.

As one might expect, various algorithms respond differently to errors and failures in the input data. The 85-GHz channels on the first SSM/I, launched in July 1987, progressively failed during 1989–1991. The Wilheit-Chang results were unaffected because they only depend on the lower frequencies (and correspondingly only provide estimates over oceans), while the NESDIS/FNMOC scheme shifted to a lower-accuracy fall-back algorithm during the failure and the GPROF algorithm exhibited progressively larger errors.

Under normal circumstances, NESDIS/FNMOC is primarily a scattering algorithm, while GPROF uses both emission and scattering signals over ocean and only scattering over land. None of the PMW algorithms is capable of estimating precipitation over frozen/icy surfaces.

One potentially confusing aspect of the PMW algorithms is that multiple algorithms have been developed for each sensor, *and* the same algorithm is sometimes applied to data from different sensors. Another source of confusion is that successive versions of an algorithm can exhibit rather different characteristics. Thus, it is important for users to be clear about which particular sensor-algorithm-version combination is producing the PMW estimates under study.

Table 2 Summary of publicly available, quasi-operational, quasi-global precipitation estimates that are produced by combining input data from several satellite sensors. Many of the input data sets are preprocessed into precipitation estimates. The TCI is also available as a separate Level 2 (satellite swath coordinates) product from the GDAAC. ''Latency'' gives the typical interval between the end of the observational period and release of the product[a]

Algorithm	Input data	Space/time scales	Areal coverage/start date	Update frequency	Latency	Archive location[b]
EURAINSAT/A	SSM/I, GEO-IR	4 km/30 min	Global – 60 °N–S/September 2002	Daily	3 days	Univ. of Birmingham
NOAA CPCP CMORPH	TMI, SSM/I, AMSU, GEO-IR	8 km/h	Global – 60 °N–S/December 2002	1 h	3 h	NOAA CPC (3)
NRL Real Time	SSM/I, GEO-IR	0.25°/h	Global – 40 °N–S/July 2000	Hourly	3 h	NRL Monterey
PERSIANN	TMI, SSM/I, GEO-IR	0.25°/6 h	Global – 50 °N–S/March 2000	6 h	2 days	Univ. of California, Irvine
TCI (3G68)	PR, TMI	0.5°/h	Global – 35 °N–S/December 1997	Daily	4 days	NASA/GSFC TSDIS (1)
TOVS	HIRS, MSU	1°/day	Global/1979	Daily	1 month	NASA/GSFC Code 613
TRMM Real-Time HQ (3B 40RT)	TMI, GPROF-SSM/I	0.25°/3 h	Global – 70 °N–S/29 January 2002	3 h	6 h	NASA/GSFC TSDIS (2)
TRMM Real-Time VAR (3B41RT)	HQ, GEO-IR	0.25°/h	Global – 50 °N–S/29 January 2002	1 h	6 h	NASA/GSFC TSDIS (2)
TRMM Real-Time MPA (3B42RT)	HQ, VAR	0.25°/3 h	Global – 50 °N–S/29 January 2002	3 h	6 h	NASA/GSFC TSDIS (2)

[a]For expansion, see Appendix 1.
[b]For details, see Appendix 2.

Table 3 Summary of publicly available, quasi-operational, quasi-global precipitation estimates that are produced by combining input data from several sensors, including rain gauges. Many of the input data sets are preprocessed into precipitation estimates. ''Latency'' gives the typical interval between the end of the observational period and release of the product[a]

Algorithm	Input data	Space/time scales	Areal coverage/start date	Update frequency	Latency	Archive location[b]
CAMS/OPI	CMAP-OPI, gauge	2.5°/daily	Global/January 1979	Monthly	6 h	NOAA CPC (2)
CMAP	OPI, SSM/I, GPI, MSU, gauge, model	2.5°/monthly	Global/January 1979	Seasonal	3 months	NOAA CPC (2)
FEWS Daily Combination	GPI, FNMOC/NESDIS SSM/I, gauge	10 km/daily	Africa/April 2000 South Asia/April 2001	Daily	6 h	NOAA CPC (2)
GPCP Version 2 SG	*1/79-6/87*, *12/87*: GPCP-OPI, gauge *7/87-present except 12/87*: SSM/I, GEO-, LEO-IR, gauge, TOVS	2.5°/monthly	Global/January 1979	Monthly	3 months	WDC-A
GPCP pentad	OPI, SSM/I, GPI, MSU (*1/79-12/94*), gauge, GPCP SG	2.5°/5-day	Global/January 1979	Seasonal	3 months	WDC-A
GPCP 1DD	SSM/I, GEO-, LEO-IR, TOVS, GPCP SG	1°/daily	Global/October 1996	Monthly	3 months	WDC-A
PREC	CMAP-OPI (1979–1998 for development of oceanic EOFs), gauge	2.5°/monthly (1°,0.5° land)	75 °N – 60°S/January 1948	Monthly	10 days	NOAA CPC (4)
TRMM Plus Other Satellites (3B 42 V.6)	TCI, TMI, SSM/I, GEO-IR, TRMM 3B43	0.25°/3-hourly	Global – 50 °N–S/January 1998	Monthly	1 week	GDAAC (2)
TRMM Plus Other Data (3B 43 V.6)	TCI, TMI, SSM/I, GEO-IR, gauge	0.25°/monthly	Global – 50 °N–S/January 1998	Monthly	1 week	GDAAC (2)

[a]For expansion, see Appendix 1.
[b]For details, see Appendix 2.

Combination Algorithms

Compared to combination data sets that depend only on satellite data (Table 2), approaches that routinely use gauge information in some way (Table 3) should have less bias where the gauges exist. However, users need to be aware of possible nonindependence between the with-gauge combinations and validation data sets built from gauge data. In addition, gauge analyses can introduce local errors into the combination. For example, Nijssen *et al.* (2001) have pointed out that the GPCP Version 2 monthly SG product has a systematic low bias in the Pacific Northwest of the United States, most likely due to preferential gauge siting in valleys in that region of complex terrain.

One important issue in constructing combinations is how to deal with known or suspected biases in the individual input precipitation data sets. For example, the TRMM Multisatellite Precipitation Analysis (MPA) calibrates each of the different PMW estimates to a single standard estimate using histogram matching of coincident precipitation rates. The calibration standard in the real-time MPA is the set of TMI-based GPROF estimates, while in the post-real-time version, the standard is the TRMM Combined Instrument (TCI) data set, which combines TMI and PR data. The IR-based Variable Rainrate Precipitation (VAR; Huffman *et al.*, 2003) algorithm is subsequently calibrated from these adjusted PMW estimates separately for each month and grid box. Similar PMW-IR calibration concepts are applied in the GPCP Monthly SG and GPCP One-Degree Daily (1DD) data sets. The Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP) takes the alternative view that biases are largely unknown and declines to select a standard. Rather, a semiobjectively weighted combination is formed from multiple individual satellite estimates, including the GPI (Xie and Arkin, 1996). The GPCP Pentad data set is a mixture of approaches, since the occurrence of precipitation is computed as in CMAP,

but the amounts are scaled to the GPCP Monthly SG (Xie *et al.*, 2003).

When calibrations are carried out for data fields that are only intermittently nonzero, such as precipitation, it is important that both data sets be averaged to the same spatial grid before calibration. Otherwise, when the finer-scale of the two estimates is averaged to the coarser scale, it will have a higher fractional coverage of nonzero precipitation than the original coarse-grid data set due to averaging zero and nonzero grid boxes around the edges of precipitation systems. At the same time, the averaged fine-scale data will lack the necessary high-end values because of averaging with near-by, lower values. These comments apply both to histogram-matched calibrations between precipitation fields and to histogram-matched precipitation-T_b calibrations for IR schemes.

IR-Based Combination Data Sets

Precipitation estimates based on IR data constitute a special class of combinations because there is no "pure" IR-based precipitation estimate in the sense that one cannot specify a precipitation scene and compute the resulting IR signal. Rather, calibration by some source of precipitation estimates is always necessary between the cloud-top information that the IR provides and the corresponding precipitation. IR techniques that require no additional information use a static calibration. For example, the GPI was calibrated using several months of data from the Global Atmospheric Research Program (GARP) Atlantic Tropical Experiment (GATE) in the summer of 1974 off West Africa. The advantage of static calibration is that the estimate only depends on a single sensor for computing new estimates, which is why the GPI and OLR Precipitation Index (OPI) are included in Table 1. Calibrations that are routinely updated with coincident precipitation data may be termed "dynamic", and require both IR and precipitation information in the calibration step, even if the estimation step only requires IR data. This is the basis for including such combinations in Table 2. One interesting question posed by dynamic calibration is how frequent the updates to the calibration should be. Monthly updates are very stable, but may fail to adequately represent changes in the weather regime, while overpass-by-overpass updates favor sensitivity over stability. This is a matter of current research and diverse practice, so users should be aware of the update interval in data sets that they plan to use.

One popular approach to dynamic IR calibration is histogram matching between coincident sets of precipitation estimates and IR T_b data. This approach assigns the highest precipitation rates to the coldest IR T_b's and successively lower precipitation rates to successively warmer IR T_bs. Figure 2 illustrates the result of such a calculation for (roughly) January 2003 from the VAR technique, which matches IR T_bs to merged, intercalibrated PMW estimates

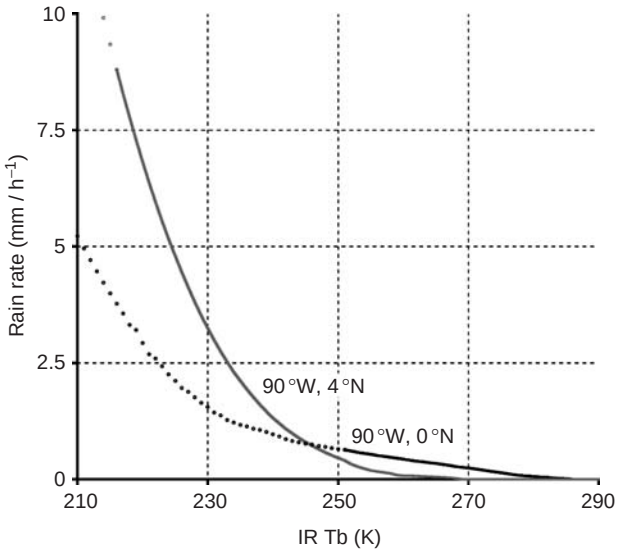


Figure 2 Examples of IR T_b – rain rate calibration curves for instantaneous, 0.25° × 0.25°-latitude/longitude grid estimates, as computed in TRMM 3B41RT for the 30-day accumulation ending 4 February 2003. The red (black) line is for the location 90 °W, 4 °N (90 °W, 0 °N), which is in the core (southern edge) of the ITCZ. The solid curves correspond to the range of data recorded in the accumulation, while the dots correspond to a climatologically based continuation to colder T_bs. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

separately for each month and grid box. Note that the relationship in the heavy rain area of the eastern Pacific Intertropical Convergence Zone (ITCZ) has a lower threshold, and assigns higher precipitation rates to temperatures that are colder than any found in the dry region just 4° of latitude south. Both curves estimate nonzero precipitation at a warmer IR T_b than the GPI threshold of 235K, both because rain tends to occur from relatively warm clouds in that region, compared to the GATE area, and because VAR works with 0.25° × 0.25° latitude/longitude gridbox-average values, while the GPI works with individual 4-km pixels. Note that the VAR threshold over land is frequently colder than the GPI threshold in heavy rain areas.

To summarize, modern combination data sets generally provide the best general-purpose estimate, but users should be aware of possible artifacts and should ask the data set producers questions about the specific performance that they are observing. In such cases, it should be possible to determine which input data set or analysis step is governing the behavior of the final combination.

EXAMPLES

The 23-year average of the GPCP Version 2 monthly SG precipitation estimates is shown in Figure 3. The regional

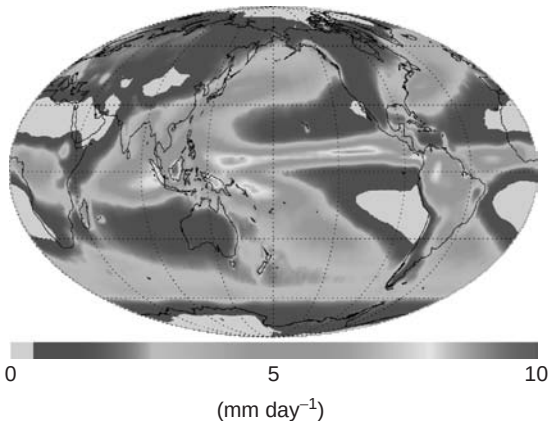


Figure 3 Annual precipitation climatology from 23 years (1979–2001) of GPCP Version 2 Monthly precipitation estimates. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

features are similar to other such climatologies, particularly over land. The goal at this point for precipitation researchers is to continue refining estimates at finer time and space scales, a by-product of which is that the overall climatology should be improved as well. The corresponding time series of the global average appears in Figure 4. The ocean, land, and total averages are shown separately, with thin (thick) lines depicting the month-to-month (running 12-month) global averages. The major dips in the smoothed land time series have a strong negative lead-in-time correlation with the occurrence of El Niño events, likely due to major declines in precipitation in the maritime continent in advance of the El Niño. Some of the year-to-year maxima in the smoothed land time series are related to La Niña events, but there is still a fair amount of interannual variation that seems unrelated to El Niño – Southern Oscillation (ENSO) events. Overall, the interannual variation of the total time series tracks with that of the ocean due to the large fraction of the Earth's surface covered by water. Nonetheless, the combination of an ocean time series with modest positive correlation to ENSO and a land time series with marked negative correlation to ENSO yields a total time series with nearly zero correlation. The overall secular change in the time series of global total is modest, but it is considered premature to trust the absolute calibrations across several important data boundaries in the current version of the data set.

As another example, Hovmöller diagrams for the latitude band 10 °N–10 °S and the period February–December 2002 were constructed for the GPCP Pentad data set, GPCP 1DD, and the then-operational TRMM Real-Time MPA (Figure 5). Given rather different algorithms, as well as time and space resolutions, the agreement is qualitatively impressive. Several eastward-moving (slanting down to the right) Madden–Julian Oscillation events are depicted in the Pacific during the spring, while westward-moving cloud clusters (slanting down to the left) continue unabated in the Atlantic Ocean The weakness of the 2002–2003 El Niño is demonstrated by the lack of significant precipitation in the Equatorial eastern Pacific, although in the climatological ITCZ location north of 10 °N, the precipitation *was* enhanced.

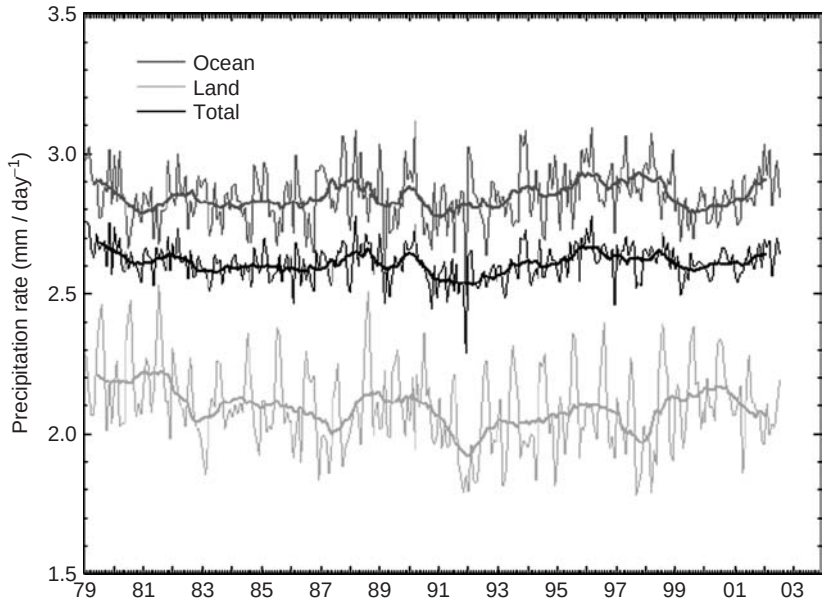


Figure 4 Time series of global-average precipitation for ocean (blue), land (green), and all (black) regions contained in the record of GPCP monthly SG estimates. Thin and thick lines depict original monthly and 12-month-running-mean values, respectively. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

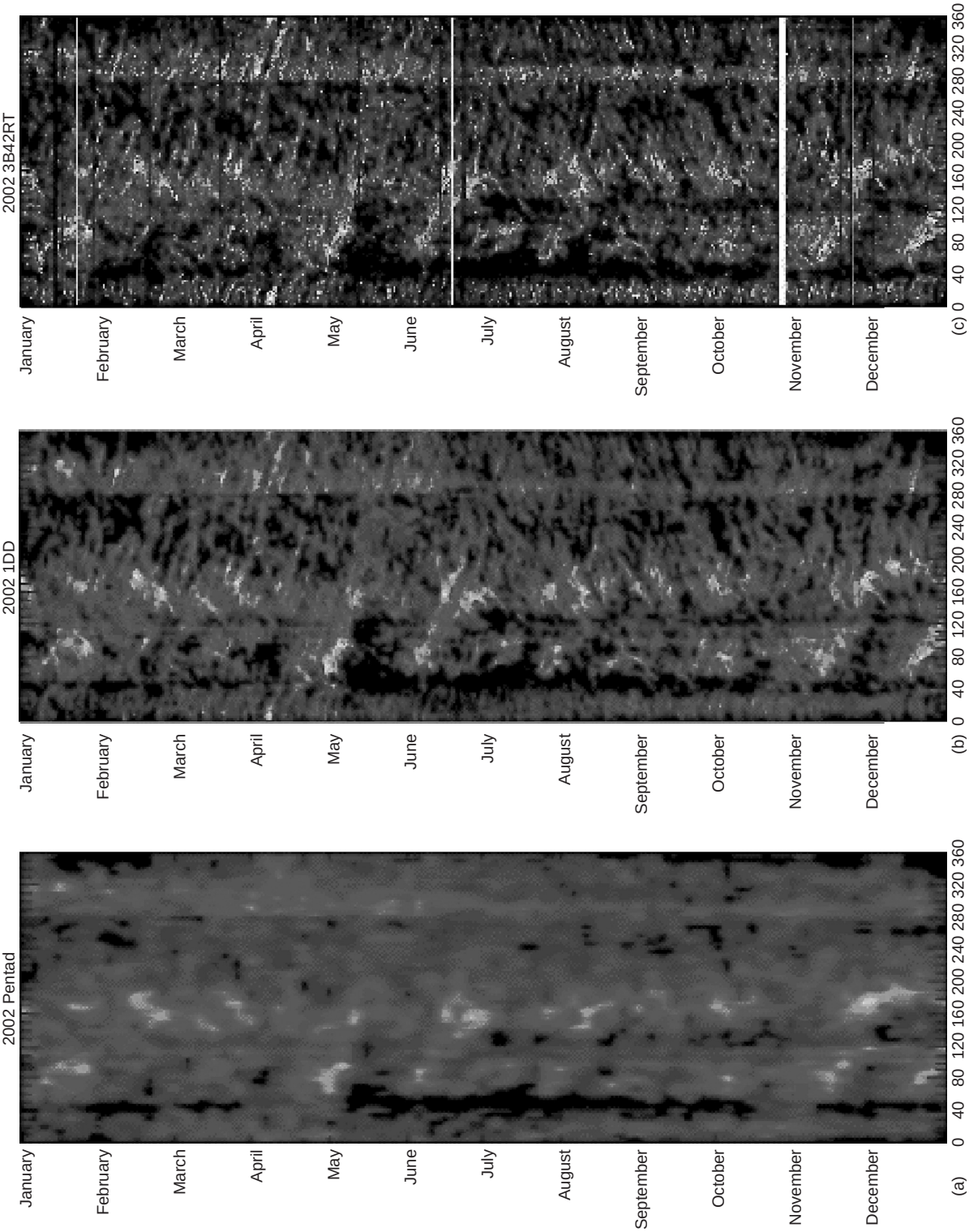


Figure 5 Example Hovmöller diagrams of estimated Equatorial precipitation for the year 2002 from three modern algorithms: GPCP Pentad (a), GPCP 1DD (b), and TRMM 3B42RT (c). In each panel, the data are averaged over the latitude zone 10°N–10°S for each longitude (along the horizontal axis) for each time (along the vertical axis), then regridded to be comparable to the other data sets. Black or white horizontal lines denote missing data. See Tables 2 and 3 for product details. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

At the finest scales, satellite-based estimates facilitate monitoring for extreme events and putting each event in its meteorological context, even when a portion of the overall storm pattern is offshore, beyond the view of radar and gauge coverage. For example, Figure 1 depicts a number of rain bands off the Pacific coast of Central America, even while attention was focused on Hurricane Isadore off the tip of the Yucatan peninsula at the time. These bands provided important antecedent information for the subsequent heavy rains over southern Mexico that appear in the September 2002 field in Figure 1.

EVALUATION

The preceding discussion has already introduced the two major factors that determine the accuracy of precipitation estimates, namely errors due to instrument and algorithm problems, and errors due to sampling. Instrument/algorithm errors are primarily driven by limitations in the retrieval process discussed above, while sampling errors are related to the very small decorrelation distances and times for individual events. Such correlation structure is a direct result of the physics of precipitation, which is generated on the microscale and results in a time series of data values that are nonnegative and only occasionally nonzero (see, for example, Krajewski *et al.*, 2000). One of the major challenges in validating the satellite estimates is accounting for the instrument/algorithm and sampling errors that are latent in the ground truth data. It is extremely rare to have a sufficiently dense, well-calibrated gauge network that covers a reasonably large area, so algorithm validation is usually focused on a few sites that are chosen to span as many climatic regimes as possible. Gebremichael *et al.* (2003) demonstrate the effects of reducing rain gauge density on the estimate of uncertainty in validating precipitation estimates. In regions without adequate gauge coverage, the errors associated with a particular algorithm must be parameterized in terms of the satellite observations themselves (Li *et al.*, 1998; Huffman, 1997). These first studies only parameterize the random part of the error in monthly precipitation estimates, and a comparable representation of bias has not been developed. Furthermore, note that the current generation of error parameterizations does not include whatever diurnal or geographically related biases may be introduced by the sampling pattern in a particular data set. Likewise, none of the current error parameterizations is designed to account for the multiscale nature of errors at submonthly timescales (outlined below).

In the 1990s, a series of intercomparison projects sought to evaluate the then-current precipitation algorithms in a variety of settings. The three Algorithm Intercomparison Projects (AIP-1, -2, and -3; Ebert *et al.*, 1998) and three Precipitation Intercomparison Projects (PIP-1, -2, and -3; Barrett *et al.*, 1994; Smith *et al.*, 1998; Adler *et al.*, 2001 respectively) tested monthly and single-overpass performance in a variety of global locations. It proved to be easy to identify flawed algorithms, but relatively hard to rank the rest. Important results from these studies include: Combination algorithms generally show improved performance compared to single-sensor schemes; PMW estimates are better than IR at short intervals; and all sensors falter in light or cold-season precipitation cases. At present, researchers working under the Coordination Group for Meteorological Satellites International Precipitation Working Group (CGMS-IPWG) are leading a validation effort for the current generation of precipitation algorithms. Most of the algorithms being evaluated are multisensor combination approaches at relatively fine time/space scales (including many of the entries from Tables 2 and 3). Routine daily validation statistics at $0.25° \times 0.25°$ resolution for Australia, the United States, and Western Europe may be accessed through the CGMS-IPWG web site **`http://www.isac.cnr.it/~ipwg/IPWG.html`**.

It is still a matter of research to fully validate precipitation estimates at the finest scales using rain gauges. The error variance is large, mostly due to the lack of accuracy in the satellite estimates, and, less importantly, resulting from uncertainties introduced by approximate time match-ups, navigation uncertainties, and geometric offsets in various satellite footprints, and the point-to-area interpolation uncertainties in the gauge analyses. As well, there can be a seasonal variation in the validation statistics that likely reflects the change in climatic regime between convective activity and more stratiform events. One early study of real-time MPA estimates (Katsanos *et al.*, 2004) employed 12-h accumulations from 73 gauges in the eastern Mediterranean region. The reliability of low and moderate rainrates was good, but high rain rates in the MPA tended to suffer a low bias. Of particular interest to hydrologists, Hossain and Anagnostou (2004) conducted a simulation study of flood prediction driven by satellite estimates. They demonstrated that sparse but high-quality precipitation estimates (having microwave-grade statistics) generated uncertain but relatively unbiased estimates of extreme stream flow, while the use of frequent but lower-quality precipitation estimates (having IR-grade statistics) produced streamflows that had a smaller random error but were biased against large events.

Another approach to assessing precipitation quality at fine scales is to intercompare satellite estimates, for example, PMW-calibrated IR precipitation estimates and the original PMW estimates. Difference maps from this study demonstrate that errors in the histogram-matched IR estimates are not simply a random shuffle of precipitation under a solid cloud canopy. On the one hand, we know that in convective situations, the occurrence of the coldest IR cloud tops lags the peak precipitation rates by as much as an

hour or more. On the other hand, at midlatitudes, the coldest IR cloud tops tend to form clusters ahead of fronts, while the PMW estimates tend to depict linear precipitation features more closely aligned with the frontal zone. Both structures are synoptically consistent patterns, but unfortunately, they do not yield the same precipitation pattern. This result is both a caution for data users and a strong impetus for algorithm developers to seek additional PMW estimates.

It is clear that a great deal of the uncertainty at the fine scales is random, because averaging over progressively larger space/time scales successively reduces the error (see the examples in Serra and McPhaden, 2003). However, the error exhibits multiscale correlations, so averaging eliminates the error more slowly than if the error were a simple random variate at the finest scale. Steiner *et al.* (2003) employ precipitation estimates from ground-based radar data over the central United States to explore the effects of subsampling in time and averaging in time and space. They find power law relationships for the expected inverse dependence of precipitation uncertainty on the time/space averaging area and precipitation rate, as well as for the direct dependence on sampling interval. They note that absolute calibration is not critical for determining the relative uncertainty in a data set of precipitation estimates, but is critical for intercomparing multiple estimates.

At the monthly timescale, the combination data sets provide reasonable behavior, particularly when gauge data are included, although some regions with complex terrain are an exception, as noted above. For example, McCollum *et al.* (2000) and Nicholson *et al.* (2003a,b) show that the individual satellite estimates tend to have a high bias compared to a special dense gauge collection in West Africa, while the GPCP Version 1 monthly SG was relatively unbiased, even though the operational gauge array used in the SG was quite sparse. Subsequent work shows similar results for Version 2. Several Version 5 TRMM products were compared over the Oklahoma Mesonet in Fisher (2004), with particular emphasis on gauge analysis uncertainty and the once-a-day sampling that occurs near TRMM's turning latitudes. Elsewhere, in regions that lack gauges, including the vast expanses of the oceans, the monthly variance is higher, and it is a matter of ongoing research to verify that the biases are small. Although oceans are not the prime focus of hydrologists, the Adler *et al.* (2003) and Bowman *et al.* (2003) studies in the tropical Pacific (GPCP SG against atoll data and TMI and PR against buoy data, respectively) demonstrate the care needed to minimize noise because of spatial and temporal mismatches between sparse gauges and satellite estimates. For all data sources, experience shows that the least reliable monthly estimates are located in polar and high mountain regions, where the precipitation is generally light and frequently frozen, the Earth's surface is frequently frozen, and there tend to be few or no gauge sites due to sparse population.

In the face of the high uncertainty at the fine scales and better statistics for averaged data, it may seem that hydrologists should confine their studies to large basins and multiday, or even monthly periods. However, one should also expect success for algorithms and models that explicitly or implicitly perform some level of averaging, such as a soil wetness model. At the large scales, where random errors are not a factor, biases remain, and in regions for which rain gauge analyses are problematic or absent, it is not yet possible to assess an absolute calibration of the precipitation field. As an alternative, it is sometimes useful to express the precipitation as deviations from the data set's average conditions, even if the record is too short to justify identifying the average as a "climatology". The quantification of biases is a major focal point for the latter part of TRMM and the early phases of the Global Precipitation Measurement (GPM) project.

FUTURE PROSPECTS

The first decade of the twenty-first century is seeing rapid advances in precipitation estimation from space, including more-frequent overpasses by PMW sensors, routine opportunities for intercalibration of precipitation estimates from all satellites, and implementation of channels and algorithms that provide useful information on frozen hydrometeors. Furthermore, the ongoing expansion of computer resources makes it more likely that computationally intensive retrievals and combinations will be routinely performed and posted on the Internet for use, comparison, and possible validation. Although not strictly a satellite issue, current tentative steps to develop regional daily gauge analyses around the globe are key to improving the accuracy of and our understanding of the fine-scale estimates that are being produced. It continues to be the case that many regions lack the gauge density to perform a straightforward daily gauge analysis, particularly when the diversity of definitions for the day's starting hour is taken into account.

On the algorithm side, it is likely that increasingly sophisticated neural networks relating PMW precipitation estimates to IR T_bs could lead to algorithms that exceed the performance of the current deterministic algorithms. The algorithm known as *Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks* (PERSIANN; Sorooshian *et al.*, 2000) is the first such algorithm to be computed quasi-operationally. A second intriguing alternative is morphing. In this approach, the motion of features on successive IR images is used as a first guess for the motion of the realistic precipitation features depicted by successive PMW

images. Once the motion and change of shape of the PMW features is established, they can be intelligently interpolated at a uniform temporal interval. The first morphing estimate was provided by the CPC Morphing algorithm (CMORPH; Joyce *et al.*, 2004). Although neural networks have a longer history, there is a tremendous amount of room for development in both neural networks and morphing.

By the end of the decade, it is expected that the GPM project will have organized the infrastructure, international data agreements, and basic algorithms to acquire, intercalibrate, process, and distribute three-hourly PMW data, including sounding channels for estimating frozen and cold-land precipitation (Adams *et al.*, 2002). One key piece of the GPM operation will be a two-frequency precipitation radar, which will continue routine intercalibration checks with other sensors, as well as building the climatological record that would likely be the first to reveal any systematic changes in the nature of precipitation systems that might arise as part of global change. Moving ahead another decade or so into the future, it is conceivable that PMW sensors will be mounted on GEO satellites, providing the necessary fine time resolution. Until GEO PMW is provided, precipitation estimates will have to depend on GEO IR (plus perhaps other channels) to provide fine time resolution, even if only as an input to a morphing scheme.

Yet another future activity involves the past. Specifically, the nearly complete record of original full-resolution, multichannel GEO and LEO data sets from Europe, Japan, and the United States was almost entirely rescued in the years around 2000. This massive collection of data sets must be accessed, summarized according to modern algorithm requirements, and then processed. It is likely that the first step will be to process the International Satellite Cloud Climatology Project B1 data set (Schiffer and Rossow, 1985), which contains all of the original internationally available satellite data sets, except subsampled by a factor of two in each direction and once every three hours. This set of tasks is quite messy, given the diversity of formats, operational blemishes, and sheer volume involved. However, from the users' perspective, the use of detailed GEO IR back to 1983 and detailed LEO IR back to 1979 would yield significant benefits in estimating precipitation.

CONCLUSION

Satellite estimates constitute the bulk of the data at our disposal for characterizing the global record of precipitation since the late 1970s. The longest records of precipitation have relatively coarse gridding in space and time and suffer from several significant data boundaries, yet they present a relatively coherent picture. Shorter records for more-recent years are available at finer time and space scales. Unfortunately, the IR data that form the bulk of these estimates are only loosely correlated to rainfall at these fine scales. Thus, users are urged to verify that the uncertainty in the fine-scale estimates is compatible with their application. In particular, averaging in space and/or time should be beneficial in improving the error statistics of the resulting estimates.

Looking towards the future, important additions to the observing system are expected, including more PMW satellites and new channels capable of supporting estimates of frozen precipitation and precipitation over icy/frozen surfaces. At the same time, reprocessing of the rescued full historical record should facilitate a much longer record of higher resolution estimates.

Acknowledgments

Thanks are due to Dr. Scott Curtis, who provided the graphics for Figures 3, 4, and 5, and to Mr. David T. Bolvin and a reviewer, who provided valuable perspectives on the

Appendix 1: ACRONYMS USED IN THE TABLES

AMSU	Advanced Microwave Sounding Unit
CAMS	Climate Analysis and Monitoring System (schemes here only use the precip field)
CMAP	CPC Merged Analysis of Precipitation product
CMAP-OPI	CMAP-calibrated OPI
CMORPH	CPC Morphing algorithm
CPC	Climate Prediction Center
FEWS	Famine Early Warning System (United Nations)
FNMOC	Fleet Numerical Meteorological and Oceanographic Center (Navy)
GDAAC	NASA/GSFC Distributed Active Archive Center
GEO	Geostationary Earth Orbit (also, a satellite in GEO)
GEO-IR	IR observed by a GEO satellite
GPCP	Global Precipitation Climatology Project
GPCP-OPI	GPCP V.2 SG-calibrated OPI
GPI	Geostationary Operational Environmental Satellite (GOES) Precipitation Index
GPROF	Goddard Profiling algorithm

(*continued overleaf*)

(*continued*)

HIRS	High-resolution IR Sounder
HQ	High-Quality merged microwave product
IR	Infrared
LEO	Low Earth Orbit (also, a satellite in LEO)
LEO-IR	IR observed by a LEO satellite
MPA	Multisatellite Precipitation Analysis
MSU	Microwave Sounding Unit
NRL	Naval Research Laboratory
NESDIS	National Environmental Satellite Data and Information Service
NASA/GSFC	National Aeronautics and Space Administration/Goddard Space Flight Center
NOAA/NWS	National Oceanic and Atmospheric Administration/National Weather Service
OPI	Outgoing Long-wave Radiation (OLR) Precipitation Index
ORA	Office of Research and Analysis
PERSIANN	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks
PMIR	Passive Microwave/Infrared algorithm
PMM	Probability Matched Method
PR	TRMM Precipitation Radar
PREC	Precipitation Reconstruction
SG	Satellite-Gauge combined precipitation estimate (usually refers to GPCP)
SI	Scattering Index algorithm
SSM/I	Special Sensor Microwave/Imager
TCI	TRMM Combined Instrument precipitation estimate (TSDIS algorithm 2B31)
TMI	TRMM Microwave Imager
TOVS	Television-Infrared Observation Satellite (TIROS) Operational Vertical Sounder
TRMM	Tropical Rainfall Measuring Mission
TSDIS	TRMM Science and Data Information System
VAR	Variable Rainrate precipitation algorithm
V.6	Version 6
WDC-A	World Data Center A (Asheville, North Carolina, USA)
1DD	GPCP One-Degree Daily precipitation product
3B40RT	TRMM Real-Time merged microwave (HQ) precipitation product
3B41RT	TRMM Real-Time TCI-VAR precipitation product
3B42	TRMM Plus Other Satellite precipitation product
3B42RT	TRMM Real-Time VAR-HQ precipitation product
3B43	TRMM Plus Other Data precipitation product
3G68	TRMM supplemental Gridded Hourly Data

Appendix 2: ARCHIVE LOCATIONS REFERENCED IN THE TABLES

GDAAC (1)	**http://daac.gsfc.nasa.gov/hydrology/hd_data.shtml**
GDAAC (2)	**http://lake.nascom.nasa.gov/data/dataset/TRMM/01_Data_Products/02_Gridded/index.html**
NASA/GSFC Code 613	**Joel.Susskind-1@nasa.gov**
NASA/GSFC Code 614.3	**http://gpcp-pspdc.gsfc.nasa.gov**
NASA/GSFC TSDIS (1)	**ftp://trmmopen.gsfc.nasa.gov/pub**
NASA/GSFC TSDIS (2)	**ftp://aeolus.nascom.nasa.gov/pub/merged**
NESDIS ORA	**ftp://ftp.orbit.nesdis.noaa.gov/pub/corp/scsb/rferraro**
NOAA CPC (1)	**http://www.cpc.noaa.gov/products/global_precip/html**
NOAA CPC (2)	**ftp://ftpprd.ncep.noaa.gov/pub/precip**
NOAA CPC (3)	**http://www.cpc.ncep.noaa.gov/products/janowiak/MW-precip_index.html**
NOAA CPC (4)	**ftp://ftpprd.ncep.noaa.gov/pub/precip/50yr/**
NRL Monterey	**ftp://ftp.nrlmry.navy.mil/pub/receive/turk/global_rain** (recent data); **turk@nrlmry.navy.mil** (old data)
Univ. of California, Irvine	**http://hydis8.eng.uci.edu/persiann/index.html**
Univ. of Birmingham	**http://kermit.bham.ac.uk/~kidd/matched/matched.html**
WDC-A	**http://www.ncdc.noaa.gov/oa/wmo/wdcamet-ncdc.html**

content and format of this article. The author is supported under NASA/GSFC Contract NAS5-01070.

FURTHER READING

Herschy R.W. (Ed.) (1998) *Hydrometry: Principles and Practice*, *Second Edition*, ISBN: 0-471-97350-5, John Wiley & Sons, 384 pp.

REFERENCES

Adams W.J., Hwang P., Everett D., Fleming G.M., Bidwell S., Stocker E., Durning J., Woodall C. and Rykowski T. (2002) *Global Precipitation Measurement – Report 8. White Paper*, NASA/TM–22002–211609, Adams W.J. and Smith E.A. (Eds.), NASA Center for AeroSpace Information: 7121 Standard Dr., Hanover, 21076–1320. 34 pp. Also posted at **http://gpm.gsfc.nasa.gov/documents/GPM_Report_8_july02.pdf**.

Adler R.F., Huffman G.J., Chang A., Ferraro R., Xie P., Janowiak J., Rudolf B., Schneider U., Curtis S., Bolvin D., *et al.* (2003) The version 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979-present). *Journal of Hydrometeorology*, **4**, 1147–1167.

Adler R.F., Kidd C., Petty G., Morissey M. and Goodman H.M. (2001) Intercomparison of global precipitation products: The Third Precipitation Intercomparison Project (PIP-3). *Bulletin of the American Meteorological Society*, **82**, 1377–1396.

Arkin P.A. (1979) The relationship between fractional coverage of high cloud and rainfall accumulations during GATE over the B-scale array. *Monthly Weather Review*, **107**, 1382–1387.

Arkin P.A. and Meisner B.N. (1987) The relationship between large-scale convective rainfall and cold cloud over the Western Hemisphere during 1982–1984. *Monthly Weather Review*, **115**, 51–74.

Barrett E.C., Dodge J., Goodman M., Janowiak J. and Smith E. (1994) The first WETNET Precipitation Intercomparison Project: interpretation of results. *Remote Sensing Reviews*, **11**, 303–373.

Bowman K.P., Phillips A.B. and North G.R. (2003) Comparison of TRMM rainfall retrievals with rain gauge data from the TAO/TRITON buoy array. *Geophysical Research Letters*, **30**, 1757, doi:10.1029/2003GL017552.

Ebert E.E. and Manton M.J. (1998) Performance of satellite rainfall estimation algorithms during TOGA COARE. *Journal of Atmospheric Sciences*, **55**, 1537–1557.

Fisher B.L. (2004) Climatological validation of TRMM TMI and PR monthly rain products over Oklahoma. *Journal of Applied Meteorology*, **43**, 519–534.

Gebremichael M., Krajewski W.F., Morrissey M., Langerud D., Huffman G.J. and Adler R. (2003) Error uncertainty analysis of GPCP monthly rainfall products: a data-based simulation study. *Journal of Applied Meteorology*, **42**, 1837–1848.

Hossain F. and Anagnostou E.N. (2004) Assessment of current passive microwave and infra-red based satellite rainfall remote sensing for flood prediction. *Journal of Geophysical Research-Atmospheres*, **109**, D07102 (DOI 10.1029/2003JD003986).

Huffman G.J. (1997) Estimates of root-mean–square random error for finite samples of estimated precipitation. *Journal of Applied Meteorology*, **36**, 1191–1201.

Huffman G.J., Adler R.F., Stocker E.F., Bolvin D.T. and Nelkin E.J. (2003) Analysis of TRMM 3-hourly multi-satellite precipitation estimates computed in both real and post-real time. *Combined Preprints CD-ROM, 83rd AMS Annual Meeting*, Poster P4.11 in: 12th Conf. on Sat. Meteor. and Oceanog., 9-13 Feb. 2003, Long Beach, CA, 6 pp.

Joyce R.J., Janowiak J.E., Arkin P.A. and Xie P. (2004) CMORPH: a method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology*, **5**, 487–503.

Katsanos D., Lagouvardos K., Kotroni V. and Huffman G.J. (2004) Statistical evaluation of MPA-RT high-resolution precipitation estimates from satellite platforms over the central and eastern Mediterranean. *Geophysical Research Letters*, **31**, L06116, doi:10.1029/2003GL019142.

Kidd C.K. (2001) Satellite rainfall climatology: a review. *International Journal of Climatology*, **21**, 1041–1066.

Krajewski W.F., Ciach G.J., McCollum J.R. and Bacotiu C. (2000) Initial validation of the Global Precipitation Climatology Project monthly rainfall over the United States. *Journal of Applied Meteorology*, **39**, 1071–1086.

Li Q., Ferraro R. and Grody N.C. (1998) Detailed analysis of the error associated with the rainfall retrieved by the NOAA/NESDIS SSM/I rainfall algorithm: part I. Tropical oceanic rainfall. *Journal of Applied Meteorology*, **39**, 680–685.

McCollum J.R., Gruber A. and Ba M.B. (2000) Discrepancy between gauges and satellite estimates of rainfall in equatorial Africa. *Journal of Applied Meteorology*, **39**, 666–679.

Nicholson S.E., Some B., McCollum J., Nelkin E., Klotter D., Berte Y., Diallo B.M., Gaye I., Kpabeba G., Ndiaye O., *et al.* (2003a) Validation of TRMM and other rainfall estimates with a high-density gauge data set for West Africa. Part I: validation of GPCC rainfall product and pre-TRMM satellite and blended products. *Journal of Applied Meteorology*, **42**, 1337–1354.

Nicholson S.E., Some B., McCollum J., Nelkin E., Klotter D., Berte Y., Diallo B.M., Gaye I., Kpabeba G., Ndiaye O., *et al.* (2003b) Validation of TRMM and other rainfall estimates with a high-density gauge dataset for West Africa. Part II: validation of TRMM rainfall products. *Journal of Applied Meteorology*, **42**, 1355–1368.

Nijssen B., O'Donnell G.M., Lettenmaier D.P., Lohmann D. and Wood E.F. (2001) Predicting the discharge of global rivers. *Journal of Climate*, **14**, 3307–3323.

Richards F. and Arkin P. (1981) On the relationship between satellite observed cloud cover and precipitation. *Monthly Weather Review*, **109**, 1081–1093.

Schiffer R.A. and Rossow W.B. (1985) ISCCP global radiance data set: a new resource for climate research. *Bulletin of the American Meteorological Society*, **66**, 1498–1505.

Serra Y.L. and McPhaden M.J. (2003) Multiple time- and space-scale comparisons of ATLAS buoy rain gauge measurements with TRMM satellite precipitation measurements. *Journal of Applied Meteorology*, **42**, 1045–1059.

Smith E.A., Lamm J.E., Adler R., Alishouse J., Aonashi K., Barrett E., Bauer P., Berg W., Chang A., Ferraro R., *et al.* (1998) Results of the WetNet PIP-2 project. *Journal of the Atmospheric Sciences*, **55**, 1483–1536.

Sorooshian S., Hsu K.-L., Gao X., Gupta H.V., Imam B. and Braithwaite D. (2000) Evaluation of PERSIANN system satellite–based estimates of tropical rainfall. *Bulletin of the American Meteorological Society*, **81**, 2035–2046.

Steiner M., Bell T.L., Zhang Y. and Wood E.F. (2003) Comparison of two methods for estimating the sampling-related uncertainty of satellite rainfall averages based on a large radar dataset. *Journal of Climate*, **16**, 3759–3778.

Xie P. and Arkin P.A. (1996) Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *Journal of Climate*, **9**, 840–858.

Xie P., Janowiak J.E., Arkin P.A., Adler R.F., Gruber A., Ferraro R., Huffman G.J. and Curtis S. (2003) GPCP pentad precipitation analyses: an experimental data set based on gauge observations and satellite estimates. *Journal of Climate*, **16**, 2197–2214.

65: Estimation of Water Vapor and Clouds Using Microwave Sensors

SUSANNE CREWELL

Meteorological Institute, University of Munich, Theresienstr, 37, 80333 Munich, Germany

Both water vapor and clouds play important roles in the energy and water cycle, and their accurate observation is required. Several techniques to measure water vapor distributions and cloud properties exist with each of them having their specific limitations. Measurements in the infrared and visible spectral range are strongly influenced by clouds and cannot be used to observe water vapor under cloudy conditions. Information about cloud properties is limited to the cloud top and base depending on the use of a space- or ground-based system. Microwave techniques can provide information about water vapor and liquid water also under cloudy conditions, but satellite observations are limited to retrievals above ocean surfaces. Therefore, the combination of different measurements needs to be further pursued in order to gain insight into the three-dimensional distribution of water vapor and cloud properties as well as their temporal variability. Recently, information about the cloud vertical structure including the occurrence of multiple cloud layers and cloud overlap has become available at a few ground-reference sites through sensor synergy including active and passive instruments. In the near future, the first global measurements of this kind will be carried out from satellites.

INTRODUCTION

Only a small fraction of the water in the atmosphere is visible in the form of liquid cloud droplets or cloud ice crystals. Most of the water (more than 99%) is contained in the form of invisible water vapor with a global average of about $25\,\mathrm{kg\,m^{-2}}$ in an atmospheric column. Water vapor is part of a continuous cycle starting with the evaporation and transpiration from the ocean and land surfaces. In the atmosphere, water vapor is transported by complex atmospheric motions leading to a highly variable distribution with fractional amounts of the total mass of air between 0 and 5%. Under certain conditions saturation occurs and water condenses to form clouds. Because the phase transition is connected with the release of latent heat, the transport of water vapor is linked to the energy cycle. Microphysical processes within the clouds eventually lead to the development of precipitation in the form of rain, snow, or hail completing the water cycle with a typical duration of about 10 days. The full hydrological cycle also includes its terrestrial branch (*see* **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1**).

Because of its absorption of solar and infrared radiation at several wavelengths, water vapor is the most important greenhouse gas. Clouds strongly modulate incoming solar and outgoing thermal radiation. Thus, both are key elements in the water and energy budget of the atmosphere and their accurate observation is required for a number of applications ranging from climate studies to numerical weather prediction. The accurate global measurement, modeling, and long-term prediction of water vapor is the primary goal of the Global Energy and Water Cycle Experiment (GEWEX) and Global Water Vapor Project (GVaP). The project demonstrated a large uncertainty in global averaged water vapor column estimates from models *in situ* and satellite measurements, as well as combined data sets with values between 23.4 and $28.7\,\mathrm{kg\,m^{-2}}$ (Randel *et al.*, 1996). For cloud water, no comparable solid database has been developed indicating a much larger uncertainty for estimates of cloud water content.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Traditionally, passive remote sensing methods, which rely on the observation of natural radiation emitted by either the sun or the Earth/atmosphere system, are used for the observation of water vapor and clouds. Owing to the strong interactions of water and clouds with radiation throughout the electromagnetic spectrum, observation techniques exist which make use of absorption and/or scattering of solar radiation, thermal infrared, and microwave radiation. One fundamental problem in remote sensing is that the inversion of the direct measurement (radiances) to the atmospheric state is ambiguous; often no unique solution exists. Consequently, the vertical structure cannot be resolved with high precision. The retrieval, also called *inversion problem*, is often approached by employing simplified regression or neural network models to fit the observable to the variable of interest. More physically based retrieval procedures involve the use of optimal estimation theory. A profound description about geophysical retrieval techniques can be found in Rodgers (2000).

Currently, active techniques that use an artificial source of radiation such as lidar and radar are emerging. These observations can provide a much better vertical resolution than passive methods. The trade off is that global observations from space are difficult to perform because of their high power consumption. Furthermore, they do not provide good spatial coverage, as scanning is only possible to a very limited degree with currently available technology. In principal, only vertical slices through the atmosphere are observed. With CloudSat (Stephens *et al.*, 2002) scheduled for launch in 2005, the first observations from a spaceborne cloud radar will be available.

WATER VAPOR

Traditionally, radiosonde ascents performed a few times per day at about 1000 stations unequally spaced over the globe, have provided the humidity information for atmospheric studies. However, the need for high temporal and spatial resolution, and high-accuracy measurements has pushed forward remote sensing technologies both for ground-based and for spaceborne applications. WMO has set the requirements to an accuracy better than 10% and a vertical resolution of 1–2 km for satellite soundings. Recently, the International H_2O Project (IHOP_2002) field experiment tried to improve the characterization of the four-dimensional (4-D) distribution of water vapor and exploit this information for a better understanding and prediction of convection. The experiment took place over the Southern Great Plains (SGP) of the United States in May/June 2002 and involved a large number of *in situ* and remote sensing instruments. An overview about the experimental setup and some preliminary highlights are given by Weckwerth *et al.* (2004).

A number of different units are used to describe the amount of water vapor in the atmosphere. The *water vapor density* [$\mathrm{kg\,m^{-3}}$] (also known as the *absolute humidity*) describes the water vapor mass per volume. Often the *specific humidity* (ratio of water vapor mass to total mass of air) or *mixing ratio* (ratio of water vapor mass to dry-air mass) are used ($\mathrm{g\,kg^{-1}}$). The *relative humidity* is the fractional amount of vapor in the air relative to the equilibrium vapor pressure of water as measured over a plane liquid surface. Additionally, the *dew point temperature* (K) is used that gives the temperature to which air has to be cooled in order to achieve saturation. To convert between the different humidity measures, knowledge of temperature and pressure is needed. Several remote sensing methods can only derive the total column amount of water vapor which is the vertically integrated water vapor density (IWV) also called *precipitable water* (PW). It is either given in units of mass per unit area ($\mathrm{kg\,m^{-2}}$) or dividing by water density to give the equivalent height of the water column (mm).

Water vapor has distinctive absorption characteristics in all parts of the electromagnetic spectrum relevant in the atmosphere. Thus different techniques – active and passive – exist, which utilize electromagnetic radiation in different spectral regions (Table 1). In the microwave range, the atmosphere is semitransparent and variations in the rotational absorption lines enable the passive remote sensing of water vapor in all weather conditions from different platforms (ground-based, air, and spaceborne). The Global Positioning System (GPS) satellites can be used to estimate water vapor contents from the additional delay of the low-frequency microwave GPS signal caused by atmospheric water vapor (Businger *et al.*, 1996). In the infrared part of the electromagnetic spectrum, thermal radiation emitted by water vapor can be used. However, severe limitations arise when clouds are present. Recently, the infrared technique has been extended to perform high-resolution interferometry to provide water vapor profiles with moderate vertical resolution from recent and future space-based instruments like the Atmospheric Infrared Sounder (AIRS) (Aumann *et al.*, 2003). Solar radiation is absorbed by water vapor at different wavelengths. By directly looking at the sun, a ground-based sun photometer can determine the integrated water content using absorption lines in the near-infrared portion of the solar spectrum. The differential absorption of reflected sunlight measured by satellites can also be used to derive the water vapor column in cloud-free conditions. Currently, LIDAR (Light Detection and Ranging) techniques are emerging, which can yield the water vapor profile with high vertical resolution. However, so far, only one operational water vapor lidar exists worldwide. Detailed descriptions of the different methods are given below.

Table 1 Overview of techniques to observe the integrated water vapor and their limitations

Method	Platform	Parameter	Principle	Strengths and weaknesses
Microwave radiometry	Ground-based	IWV	Water vapor emission at 22.235 GHz rotational line	High temporal resolution time series in all weather conditions except precipitation
		Profile $\Delta z = 0.5-2$ km	Pressure broadening of rotational line	Dense surface networks are too costly
	Spaceborne SSM/I, TMI, AMSU, AMSR	IWV profile $\Delta z = 2-3$ km	As for ground-based	From polar-orbiting satellites with at least daily coverage for all weather conditions, over ocean surfaces only with coarse resolution of about $50 \times 50\,\text{km}^2$
Global positioning System	Ground-based	IWV	Delay of GPS signal caused by water vapor	Inexpensive surface networks for all weather conditions, real-time products have less accuracy
Thermal infrared	Spaceborne GOES, MSG	IWV	Differential emission at 11 and 12 μm (split window)	Good spatial and temporal in cloud-free situations only
	HIRS, GOES	Profile	Channels along 6.7 μm line	Limited accuracy
	AIRS, IASI	Profile	High-resolution spectroscopy	High vertical resolution and accuracy
Sun photometer	Ground-based	IWV	Extinction of direct sunlight at 0.95 μm absorption band	Sun tracking is only possible during daylight and cloud-free conditions
Reflected sunlight	Spaceborne	IVW	Difference of reflectances at 0.95 μm and at close by window wavelengths	High spatial (1×1 km) resolution
	MODIS, MERIS, POLDER			Only in cloud-free situations ore above clouds
Lidar	Ground-based	Profile $\Delta z < 100$ m	Raman backscatter of water vapor	Best during night
			Differential absorption at two channels	Non operational

Microwave Radiometry

In the microwave region of the electromagnetic spectrum, atmospheric emission is due to water vapor, oxygen, and cloud liquid water (Figure 1). Radiative transfer is simplified as scattering by atmospheric constituents (with the exception of rain drops) and can be neglected (*see* **Chapter 46, Principles of Radiative Transfer, Volume 2**). Ground-based, airborne, or spaceborne microwave radiometers measure the thermal emission which is typically given as brightness temperatures (TB) in Kelvin. Water vapor has distinct spectral features at 22.235 and 183 GHz. Since the introduction of satellite remote sensing, dual-channel methods have made use of the rotational line at 22.235 GHz to derive the vertically integrated water vapor over oceans (Grody, 1993). Within this method one channel measures at a frequency close to the line center where water vapor absorption is almost independent of height. The observed brightness in temperature is approximately proportional to IWV. The second channel is located in a window region where water vapor and oxygen absorption are low (e.g. 31.4 GHz). Since microwave emission of clouds increases approximately with the square of the frequency (Figure 1), the second channel can be used to characterize the cloud's contribution to the intensity observed at the first frequency. This combination is extremely powerful as the IWV and the liquid water path (LWP), which is the vertically integrated liquid water density, can be retrieved simultaneously. For the past few decades this principle has been used for ground-based instruments (Westwater, 1978), as well as on the first satellites employing microwave radiometers at similar frequencies. Specifications of the accuracy vary between 0.3 and $1\,\text{kg}\,\text{m}^{-2}$. Sometimes microwave IWV measurements from the ground are used to scale radiosonde or lidar measurements. The experiences of the Atmospheric Radiation Programme (ARM) in respect to scaling are described by Revercomb *et al.* (2003). Limitations of the method are the radiometer uncertainty and the underdetermined retrieval problem as more than one atmospheric state can lead to the same set of brightness

Table 2 Overview of techniques to observe the cloud properties and their limitations. Future space missions are not listed

Method	Platform	Parameter	Principle	Strengths and weaknesses
Microwave radiometry	Ground-based	LWP	Cloud emission proportional to water content and frequency	Most accurate method, high temporal resolution time series
	Spaceborne	LWP	As for ground-based	Over ocean surfaces only with coarse resolution of about $50 \times 50\,km^2$
	SSM/I, TMI, AMSU, AMSR			Inhomogeneous beamfilling in case of precipitation
Cloud radar	Ground-based	Reflectivity and velocity profile $\Delta z < 100\,m$	Backscatter of cloud droplets, Doppler shift gives fall velocity	Vertical structure can be resolved, conversion to liquid (ice) water content is highly uncertain, information about phase
Thermal infrared	Ground-based	Cloud-base temperature	Cloud emission of cloud base	High temporal resolution, atmospheric correction for water vapor emission
	Spaceborne	Cloud-top temperature	Cloud emission of cloud top	High temporal and spatial resolution from geostationary satellite
	GOES, Meteosat, AVHRR, MODIS	Cloud-top pressure		
Reflected sunlight	Spaceborne	Optical depth	Reflectance at non absorbing channel (0.65 μm) and near-infrared absorption	Good global coverage, high spatial $1 \times 1\,km^2$ resolution for MODIS
	MODIS, MERIS, AVHRR	Effective radius LWP		Representative for upper cloud part only
Lidar	Ground-based	Backscatter profile $\Delta z < 100\,m$	Backscatter of cloud droplets	Best during night polarization for phase detection

temperatures. Slight improvements in accuracy can be made by using additional frequencies to further constrain the retrieval problem (Crewell and Löhnert, 2003). From the ground, microwave radiometers can provide unattended measurements of IWV at high temporal resolution during all weather conditions (Figure 2), except during precipitation when water on the antenna/radome emits microwave radiation and disturbs the measurement. Comparisons between different microwave radiometers have shown systematic differences of about $1\,kg\,m^{-2}$ mainly caused by the use of different gas absorption models in the retrieval process. Existing ground-based microwave radiometers have very different designs and specifications and are far more expensive than GPS receivers. However, technological progress has been made within the last few years and long-term time series of IWV now can be gathered and investigated (e.g. Güldner and Spänkuch, 1999).

Because the emission from land surfaces is too high and too variable within the field of view, satellite measurements can provide IWV over the oceans only. At the large microwave wavelengths, diffraction effects of the antenna yield a relatively coarse horizontal resolution at the Earth's surface of about $50 \times 50\,km^2$. By scanning across the satellite track, swaths of about 2000 km can be observed. A comprehensive tutorial about satellite observing techniques in the microwave region can be found at **http://www.meted.ucar.edu/ist/poes2/.** Currently, measurements from the Special Sensor Microwave/Imager (SSM/I), the Advanced Microwave Sounding Unit (AMSU-A), the Tropical Rainfall Measurement Mission (TRMM) Microwave Imager (TMI), and the Advanced Microwave Scanning Radiometer (AMSR) provide good global coverage over the oceans. Real-time water vapor fields for the tropics can be accessed at **http://www.ssd.noaa.gov/PS/TROP/trop-atl.html.**

Water vapor profiles are derived from microwave profilers that measure the atmospheric emission at several frequencies along the wings of pressure-broadened rotational lines. From the ground the 22.235 GHz line is used, whereas from space the strong 183 GHz water vapor line is more suitable (Figure 1). The reason is the high atmospheric opacity at 183 GHz, which prohibits measurements from the ground but which partly eliminates the problem of the unknown surface emission for satellite measurements. The contribution of a certain atmospheric layer to the signal

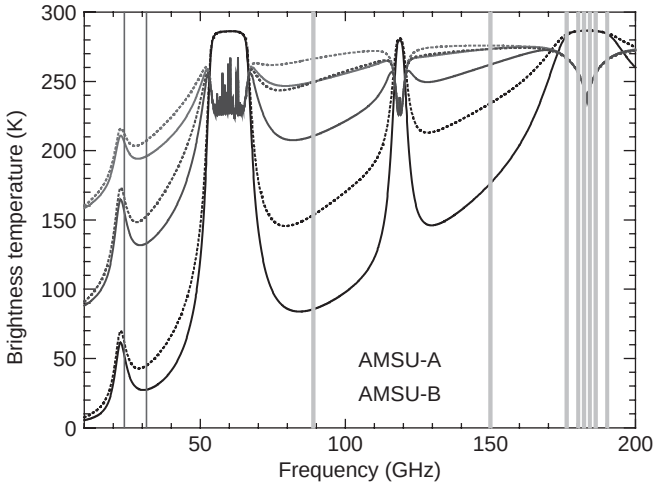


Figure 1 Brightness temperatures in the microwave spectral range simulated for a midlatitude summer atmosphere (IWV = 20 kgm − 2) for ground-based observation (a) and satellite observations over the ocean. As the emission of the ocean surface is polarization dependent, the results are shown for vertical (c) and horizontal (b) polarization. Brightness temperatures were calculated for clear sky (solid) and cloudy (dotted) conditions. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

as observed at the receiver is described by its weighting functions. The reader is referred to Janssen, (1993) for a comprehensive review on microwave remote sensing including details on weighting functions. For ground-based observations the weighting functions at the different frequencies all decrease continuously with height. The satellite weighting functions show marked and well-separated peaks for strongly absorbing frequencies (Figure 3). Channels close to the line center are already saturated in the upper troposphere and do not get contributions from lower layers. For frequencies further away from the line center, atmospheric transmission increases and signals from lower atmospheric layers are observed. The vertical resolution of the ground-based water vapor measurement is estimated to be 0.5–2.0 km; it decreases with altitude and is best close to the surface in the planetary boundary layer (Güldner and Spänkuch, 2001). Measurements made by either AMSU-B or SSM/T-2 can provide low-resolution vertical profiles that are more accurate in the lower troposphere, although measurements over land should be limited to the mid and upper troposphere.

Microwave spectroscopy is also used to observe the water vapor profile in the stratosphere and mesosphere. This is achieved by measuring the inner part (e.g. 1–2 GHz) of the 22 GHz line with high spectral resolution (2 MHz), and consequently deconvoluting the pressure-broadened spectrum. To avoid the complication of the tropospheric contribution obscuring the stratospheric signal, observations are preferably performed where moisture is low, for example, on high mountains or in polar regions.

GPS

Originally designed for navigation purposes, the Global Positioning System can also be used to determine atmospheric water vapor when the position of the GPS satellites and the GPS receiver are precisely known. Since this method has emerged, the number of ground-based GPS receivers organized within networks to derive columnar water vapor is increasing, as well as the use of GPS receivers on Low-earth Orbiting (LEO) satellites for water vapor profiling. An overview about current GPS activities is given in Bengtsson *et al.* (2003).

The time the GPS signal needs to travel from the GPS satellite to the GPS receiver is longer than the time calculated from the distance and the velocity of light in a vacuum. The time difference converted to distance, the so-called delay, is the original GPS measurement from which the water vapor column is derived. The ratio between the speed of light in the atmosphere and the speed of light in a vacuum is the index of refraction. At microwave frequencies the refractivity depends on the free electron density of the ionosphere and a neutral contribution. The ionosphere can cause delays of several tenths of meters but can be corrected using a dual-frequency method, for example, using two GPS signals at 1.2 and 1.5 GHz. The neutral contribution can be further separated into the hydrostatic delay (which can account to 2.3 m but is easily corrected by knowing atmospheric pressure) and the wet delay, the most variable contribution to the GPS signal. Correcting for ionospheric and hydrostatic effects, the wet delay can be determined with good accuracy. In order to take the elevation of the GPS satellite into account, the wet delay is converted to a zenith direction. The zenith wet delay (ZWD) is closely related to the integrated water vapor with a ratio of IWV/ZWD, corresponding approximately to 0.15 (Businger *et al.*, 1996). Further improvements in the accuracy can be obtained by incorporating the effective atmospheric temperature, which can be approximated from the ground-level temperature, in the IWV estimation. Several intercomparison studies have been performed and have demonstrated their potential to measure IWV with a good accuracy of $1-2\,\mathrm{kg\,m^{-2}}$.

As GPS receivers are relatively inexpensive and require low maintenance, they are used in large networks within Europe and the United States (**`http://www.gst.ucar.edu/gpsrg/realtime/`**). Currently, their potential for climate monitoring, atmospheric model evaluation, and data assimilation is being investigated. However, for the latter application, real-time data are necessary. Unfortunately, the exact information about GPS satellite

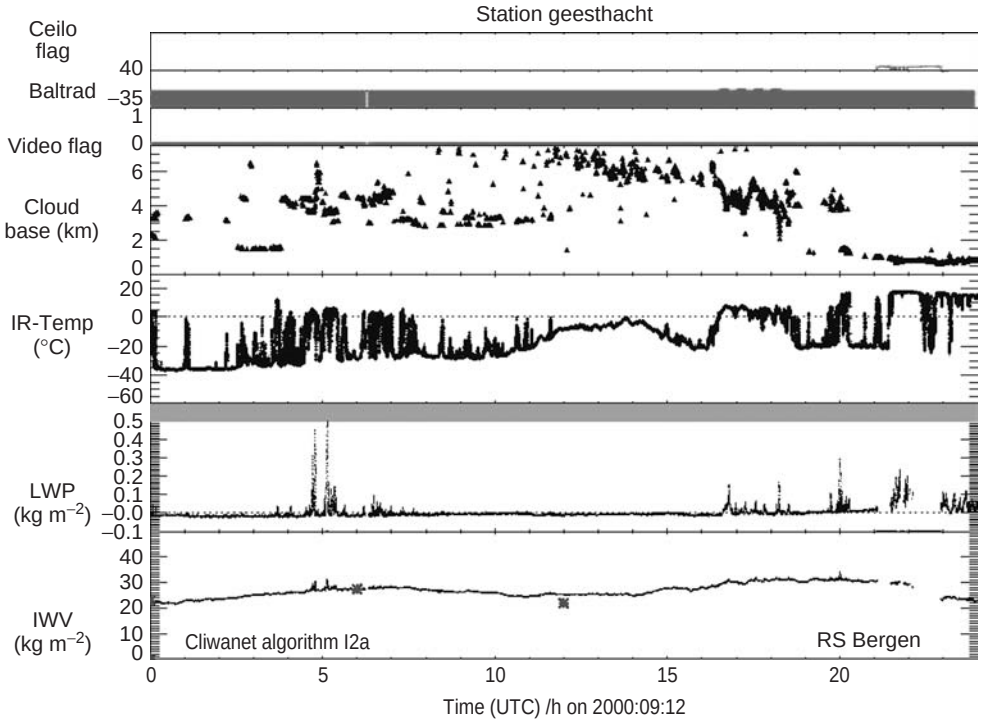


Figure 2 Times series of integrated water vapor (IWV) and liquid water path (LWP) derived from multispectral microwave radiomeasurements, infrared temperature, and cloud-base height derived from lidar ceilometer measured on 12 September 2000 at Geesthacht, Germany. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

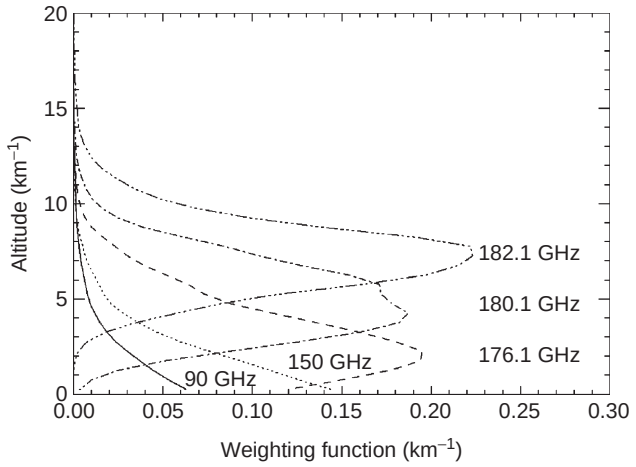


Figure 3 Weighting functions for the frequencies corresponding to the five AMSU-B channels (see Figure 1). The frequencies close to the center of the water vapor line at 183.1 are already saturated in the upper troposphere. The frequencies further away from the line center receive radiation from lower atmospheric levels. Therefore the window frequencies at 90 and 150 GHz also include information from the ground

positions is difficult to provide in real time. Therefore the accuracy for online values (employing predicted orbits) is lower compared to reprocessed data. When high-density network of GPS receivers are used, tomographic techniques can be applied to derive the four-dimensional structure of water vapor in the lower atmosphere (Flores *et al.*, 2001). The same observation principle as with GPS can also be applied to Very Long Base Line Interferometry (VLBI) with large telescopes used in space geodesy.

GPS receivers operated on LEO satellites can determine water vapor profiles via the so-called radio occultation technique. When these satellites rise and set relative to a GPS satellite, the GPS signal frequency changes as a result of the Doppler shift. From the frequency measurement the bending angle of the radio waves can be calculated. The bending angle is determined by the atmospheric refractivity and thus related to the water amount. Encouraging results have been obtained, for example, by the Challenging Minisatellite Payload (CHAMP) satellite (Wickert *et al.*, 2004) and future missions are planned to increase the number of profiles observed per day. One aspect to be clarified, however, is the occurrence of humidity biases especially in the tropics.

Thermal Infrared

Numerous absorption lines and bands exist in the infrared part of the electromagnetic spectrum. The typical temperatures of the Earth/atmosphere system lead to a

maximum emission of infrared radiation at about 10 μm. Exactly in this region – between about 9 and 12 μm – is a so-called window region with high atmospheric transmissivity. What little absorption remains in this spectral region is because of water vapor (and to a minor degree by ozone). However, the signal measured by a satellite in the atmospheric window is mainly dominated by emission from the Earth's surface. Therefore it is possible to simultaneously derive the integrated water vapor and the land surface temperature by the so-called Split-window (SW) technique from measurements at 11 and 12 μm. The method is limited to cloud-free scenes as clouds strongly absorb in this spectral region. The SW algorithm is derived from a perturbation form of the radiative transfer equation (*see* **Chapter 46, Principles of Radiative Transfer, Volume 2**) that is simplified through a parameterization to retrieve bulk layer parameters rather than profile information. This method is applied to the measurements from the Geostationary Operational Environmental Satellite (GOES) and the recently launched Meteosat Second Generation (MSG) satellite and can therefore provide good temporal and spatial coverage. The accuracy of this method has been evaluated in comparison to radiosondes and ranges from about $2.0-7.0\,\mathrm{kg\,m^{-2}}$. Further information about the SW technique and its validation for GOES can be found in Suggs *et al.* (1998).

By using several infrared channels along the wing of the water vapor absorption band (6.7 μm) where each channel receives radiation from a different altitude range, low-resolution water vapor profiles can be derived. For example, GOES has three channels 7.43, 7.02, and 6.51 μm, which give information about low-level, midlevel and upper-level moisture respectively. Near real-time soundings of water vapor and temperature can be accessed for the GOES coverage area at **`http://orbit35i.nesdis.noaa.gov/goes/soundings/skewt/html/skewtus.html`.** Similar infrared channels are used in the High-resolution Infrared Sounder (HIRS) flown as part of the Advanced TIROS Operational Vertical Sounder (ATOVS) package on the polar-orbiting NOAA satellites beginning with NOAA-15. The AMSU instrument described in the previous section is part of the ATOVS packages so that the limitation of the infrared method to cloud-free cases is compensated by the microwave technique. The accuracy of the operational ATOVS humidity soundings is 3 to 6 K in dew point temperature at 2-km vertical resolution (Li *et al.*, 2000).

While the infrared methods described above are based on the observation of a couple of channels, high-resolution interferometry of the infrared spectrum has the potential to retrieve water vapor with a much higher accuracy and vertical resolution, but again in cloud-free conditions only. This sort of instruments resolves the atmospheric emission between approximately 3 and 20 μm with a wavenumber resolution of $0.5\,\mathrm{cm^{-1}}$. The recorded spectra contain a large number of spectral features because of carbon dioxide, ozone, water vapor, and other minor trace gases. Similar to microwave radiometry the radiative transfer equation needs to be inverted to simultaneously retrieve the water vapor and temperature profiles. It should be noted that even with the high number of individual points, the retrieval problem is still ill defined as the channels are strongly correlated. Thus constraints are needed to prevent the inversion from becoming unstable. For the ground-based Atmospheric Emitted Radiance Interferometer (AERI) used by ARM, it was demonstrated that continuous observation of water vapor profiles can be achieved with approximately 5% accuracy in absolute water vapor in the lowest 3 km. A detailed description of the instrument and its validation is given by Feltz *et al.* (2003). The Earth Observing System (EOS) Aqua satellite launched in 2002 has the first high-spectral-resolution infrared sounder for operational weather forecasting onboard. The AIRS (**`http://www-airs.jpl.nasa.gov/index.html`**) performs observations at 2378 channels, while the Infrared Atmospheric Sounding Interferometer (IASI) instrument on the future METOP satellite will provide 8000 channel measurements. IASI will observe temperature and humidity profiles with a vertical resolution of 1.5–2 km and an accuracy of 1 K and 10% respectively to meet the WMO requirements (Lerner *et al.*, 2002).

Sun Photometer

A sun photometer measures the integrated water vapor by the extinction of direct solar irradiance. For this purpose exact tracking is necessary to point the collimator toward the center of the solar disk. A filter wheel rotates in front of the detector to obtain a sequence of measurements at typically 10 different spectral bands. The sun photometer also performs scanning across the sun to detect the presence of thin cirrus clouds that are typically for nonuniform scenes. Scanning is also necessary to derive the prime products: The aerosol size distribution and phase function are determined by inverting sky brightness temperature and incorporating radiative transfer routines. The integrated water vapor is derived from the strong near-infrared absorption band at 0.94 μm. As the strength of the solar transmission is known, and its extinction is described by Bouguer–Lambert–Beer's law (*see* **Chapter 46, Principles of Radiative Transfer, Volume 2**) a simple exponential relation between transmission and IWV can be assumed. This method is limited to cloud-free conditions and daylight. The uncertainty of this method is about 10% and is mainly caused by uncertainties in the knowledge of water vapor spectroscopy. For details see Schmid *et al.* (2001).

Reflected Sunlight from Satellite

From satellites the near-infrared (about 0.95 μm) water vapor absorption band can also be exploited. In contrast to the sun photometer where the direct sunlight provides a reference value, measurements in at least two channels are necessary to derive precipitable water from reflected sun light to account for surface reflectivity. Solar reflectance is measured at one window channel and at a second channel that experiences stronger water vapor absorption. The method can only be applied in regions with a reflective surface, for example, over land and over sun glint regions during daytime only. IWV can be measured with an accuracy of 5–10% from a number of polar-orbiting satellites employing these channels. The Moderate Resolution Imaging Spectrometer (MODIS) onboard the TERRA satellite (in orbit since 1999) uses four channels close to the water vapor absorption region to achieve a higher dynamic range. By using ratios of different channels, the effects of variations in surface reflectivity with wavelength are partially removed. MODIS provides high-resolution (1×1 km) IWV fields during daytime. Details on the remote sensing of water vapor and clouds are given by King *et al.* (1992). In the case of clouds, their high reflectance can be exploited and the water vapor column above the cloud deck can be estimated. The analysis of POLDER (Polarization and Directionality of the Earth Reflectances) measurements showed a mean root mean square (rms) error of 1.8 kg m^{-2} over the ocean and 2.0 kg m^{-2} over land (Albert *et al.*, 2001).

Lidar (Raman Lidar and DIAL)

Two different lidar techniques that provide high vertical resolution profiles of water vapor are emerging: Raman lidar (RL) and Differential Absorption Lidar (DIAL). Both lidar techniques work by emitting pulses of laser light and determining the range to the backscattering object by measuring the travel time of the respective pulse. Vertical resolutions better than 100 m can be achieved. Raman lidar measures the elastic backscatter from molecules and aerosols at the laser wavelength, as well as the Raman backscattered light at the two distinctive wavelengths that represent the Raman (wavelength) shift of nitrogen and water vapor molecules, respectively. The signal in each Raman channel is proportional to the number density of the backscattering molecule. Because nitrogen is uniformly mixed, the mass of nitrogen per kilogram of dry air is constant with altitude. Therefore the ratio between the number densities of water vapor and nitrogen (the ratio between the backscatter signal of both channels) is proportional to the water vapor mixing ratio – the mass of water vapor per mass of dry air (g kg^{-1}). The performance of Raman lidar systems is best during nighttime, in the absence of the daytime solar background. To allow daytime operation, a narrow band interference filter and a narrow field of view are used for the water vapor detection channels. Limiting factors in accuracy are the uncertainties in Raman cross sections and optical transmissions requiring an external calibration. However, this is a single height-independent calibration factor, which is mostly derived from radiosondes or microwave radiometers. The Raman lidar is able to measure upper tropospheric humidity with approximately 5% absolute accuracy. In the lower atmosphere, imperfect alignment can effect the measurement. Most of the presently operating systems cannot – or with strongly reduced performance – be operated during daytime. There is only one system (ARM, Oklahoma) worldwide that can operate unattended all day around the clock (Turner and Goldsmith, 1999).

DIAL measurements are made at two different wavelengths. One wavelength is chosen in a region of high water vapor absorption cross section, whereas at the second wavelength the gaseous absorption should be minimal. If aerosol scattering properties (backscatter and extinction) are the same at both wavelengths, water vapor density can be calculated assuming that the differential water vapor absorption cross section is known. The accuracy of DIAL measurements is specified to be better than 5% in the whole troposphere (Wulfmeyer and Bösenberg, 1998). The main difficulties which were encountered in the application of DIAL to water vapor measurements were associated with the narrow line width of the near-infrared absorption lines, which require tunable laser sources with extremely high frequency stability and narrow bandwidth, and, in particular, with extremely high spectral purity. Daytime measurements of water vapor using DIAL give superior results over RL in terms of resolution (in time and height) and accuracy, and have the advantage of being intrinsically calibrated. RL and DIAL should be regarded as complementary techniques.

Conclusions: Water Vapor

A number of different methods for observing water vapor and its vertical structure exist, each with its own special strengths and weaknesses. From the ground, microwave radiometers and Global Positioning Systems provide unattended all-weather measurements with high temporal resolution but coarse vertical resolution. Under cloud-free conditions high-resolution vertical structures can be observed by lidar. However, the lidar technique is far from the point of being used in operational surface networks. From space the integrated water vapor over the oceans can be observed with good global coverage in all weather conditions using microwave radiometry. Infrared methods from geostationary and polar-orbiting satellites provide a better horizontal resolution also over land but only in cloud-free conditions. Vertical profiles from satellite

soundings in the infrared and microwave regions have a coarse vertical resolution. A major improvement of the vertical resolution is expected by the new generation of infrared spectrometers. Since the methods have complementary benefits, major research is being carried out to integrate the different observations to obtain better products. Some of these approaches will be presented in the Section "Integrated approaches" devoted to sensor synergy.

The specification of accuracy of the different remote sensing methods provides a major challenge. Traditionally, radiosonde soundings are used for this purpose. Differences can be expected because a radiosonde measures the water vapor along its trajectory, for example, at different positions during its ascent. Unfortunately, the quality of standard radiosonde humidity measurements has also been found insufficient to serve as a reference, especially in the upper troposphere. Improvements are possible by employing chilled mirror hygrometers but these lead to much higher costs. The ARM program has performed several water vapor intensive observation periods (IOP) at its central facility in Oklahoma. An overview and initial accomplishments are described by Revercomb *et al.* (2003). The comparison included radio sondes, microwave radiometer, AERI, sun photometer, GPS, Raman lidar, DIAL, and *in situ* sensors. Unfortunately, it is still not possible to constrain the column amounts to better than 2% as biases between the different instruments are too high. More detailed error analysis is required since none of the instruments can be regarded as an absolute reference at any time. In the upper troposphere it was found that Raman lidar and GOES estimates agree much better with each other (10%) than radiosoundings (40% differences) (Soden *et al.*, 2004).

CLOUDS

Owing to a strong influence on atmospheric radiation, clouds are the limiting factor for several methods to derive atmospheric water vapor. On the other hand the strong interaction of clouds and radiation enables remote sensing techniques to infer cloud properties (Table 2). Information about the interior of water clouds, however, is only possible using microwave techniques active as well as passive because clouds are semitransparent in this spectral range. Ice clouds can be probed by optical lidar.

Water clouds consist of a large number of droplets of varying sizes. The number of all droplets within the unit volume is the total number density (m^{-3}). The drop-size distribution (DSD) describes the number density as a function of the droplet radius. Because of the complex microphysical processes within clouds, DSDs are highly variable in time and space. In contrast to raindrops, cloud droplets are perfect spheres. Thus, all cloud microphysical parameters can be calculated from the DSD. For example, the cloud liquid water content (LWC) ($kg\,m^{-3}$) is given by the droplet volume and the density of water. Because the volume of a sphere is proportional to the radius cubed, LWC is also called the *third moment of the DSD*. It comprises one of the most interesting properties of clouds and is the prognostic variable in most numerical weather prediction and climate models to describe clouds, but hardly any observations are available for the validation of the model results. Within these models treatment of ice clouds is even less advanced and prognostic schemes for the ice water content (IWC) are currently under development. For ice particles the observation of their properties becomes more difficult as they can exist in complicated shapes as hexagonal columns and plates, dendrites, needles, and so on. In general, clouds can either occur in a pure liquid or ice phase but also as a mixture of both.

In describing the radiative properties of clouds and their effect on the energy budget of the atmosphere, cloud opacity (also called *optical thickness or optical depth*) is the most important parameter. Cloud opacity τ is a dimensionless measure of the light attenuation caused by scattering and absorption of solar radiation by atmospheric particles. The transmission through the cloud is given by $\exp(-\tau)$, for example, a cloud with an optical depth of 5 has a transmission less than 1%. The effective radius of the cloud is needed for some applications, for example, solar radiative transfer models. It is defined as the third moment of the DSD divided by the second moment, meaning it is a ratio of droplet volume to droplet surface area.

To observe the complete microphysical properties, for example, the DSD and the complex shape of ice or mixed cloud particles, major efforts employing aircraft and balloon measurements are necessary. Traditionally, remote sensing methods have focused on the observation of macroscopic cloud properties such as cloud boundaries and the cloud fraction in a certain area. Currently, a number of synergistic approaches combining active and passive instruments (radar, lidar, and radiometer) are being pushed in an attempt to increase knowledge about cloud microphysics.

Pioneering work has been done within the International Satellite Cloud Climatology Project (ISCCP), which infers cloud properties from operational satellites in geostationary and polar orbit. Established in 1982 as part of the World Climate Research Programme (WCRP), ISCCP collects and analyzes satellite radiance measurements to infer the global distribution of clouds, their properties, and their diurnal, seasonal, and interannual variations from a data set of nearly two decades. The ISCCP products based on harmonized algorithms are cloud cover fraction, optical thickness, cloud-top temperature and its mesoscale variability, cloud-top pressure, and several more. The cloud type

(cumulus, stratus, etc.) is derived from a cloud classification on the basis of optical thickness and cloud-top pressure. A general overview about ISCCP can be found in Rossow and Schiffer (1999). The project web site (Rossow and Duenas, 2004) **http://isccp.giss.nasa.gov/** provides access to data since 1983. In the following section a brief outline about the basic principles in deriving cloud parameters from measurements in the different spectral regions, as well from the ground as from space is given.

Microwave Radiometry

The liquid water path can be estimated from atmospheric emission measurements in the microwave region because in this spectral region the cloud contribution strongly increases with frequency (Fig.1). The standard dual-channel principle has been described above for the determination of precipitable water. For the retrieval of LWP the channel close to the water vapor absorption line corrects for the changing water vapor concentration of the atmosphere. Such observations are, with the exception of expensive and rather limited aircraft measurements, the most accurate method to observe LWP with an estimated accuracy of better than $30\,g\,m^{-2}$. A rough estimate shows that about $10\,g\,m^{-2}$ are caused by the measurement error while the rest can be attributed to the underdetermined retrieval problem. The additional use of the 90 GHz channel can further constrain the problem and improve accuracy to less than $15\,g\,m^{-2}$ (Crewell and Löhnert, 2003).

Satellite measurements (SSM/I, AMSU, TRMM, AMSR) have been available for more than two decades. Their main disadvantage is that reasonable estimates can only be made above the oceans and only with poor (several 10 km) resolution. The high variability of clouds can cause errors in LWP as a result of inhomogeneous beam filling of the satellite field of view, especially in the case of precipitation. Over land the ground-based microwave radiometers can observe LWP with high temporal resolution at one point (Figure 2). First attempts to set up a network of ground-based microwave radiometers have been performed within the BALTEX Cloud Liquid Water Network (Crewell *et al.*, 2002) in north central Europe to evaluate and improve the forecast of clouds in atmospheric models.

Ice clouds are transparent at low microwave frequencies. With increasing frequency the scattering efficiency of ice clouds increases and cirrus clouds reduce the upwelling brightness temperature significantly. Because of the strong water vapor absorption in the submillimetre, the surface is screened from the satellite view. The observed reduction in brightness temperature is closely related to the ice water content and the characteristic particle size. Thus, suggestions for future spaceborne missions have been made (Evans *et al.*, 2002) to observe ice clouds from measurements in the range between 200 and 700 GHz. The frequencies need to be selected carefully as a large number of atmospheric trace gases have emission lines in this region. Currently, the capabilities of this method are being analyzed by investigating the radiative transfer of irregular shaped ice crystals in this frequency range and by aircraft demonstration of the instruments.

Cloud Radar

Radiowave Detection and Ranging (Radar) basically employs the same principle as lidar techniques, but involve radiation in the microwave range. While weather/precipitation radars operate at microwave frequencies below 10 GHz, cloud radars typically transmit radiation at window regions (35 and 95 GHz) to avoid attenuation by water vapor (Figure 1). Because attenuation increases with frequency in the microwave region, cloud radars still suffer from significant attenuation and are limited in their operation range. Ground-based cloud radars typically point vertically and gather time-height series (Figure 4). The vertical resolution of ground-based cloud radars is comparable to lidars with less than 100 m. Future spaceborne radars (Stephens *et al.*, 2002) will have a vertical resolution of about 500 m. Like precipitation radar they mostly operate in the Rayleigh regime, for example, the particle size is small compared to the wavelength. Because cloud droplets ($\sim$10 μm) are much smaller than rain drops ($\sim$1 mm), a higher frequency (shorter wavelength) is used in cloud radars. In the Rayleigh regime the backscatter from cloud droplets is proportional to the sixth moment of the DSD and thus, the derivation of the liquid water content, proportional to the third moment of the DSD, is highly uncertain. Even the presence of one single large drop can dominate the whole measurement. This effect is often evident when clouds start to precipitate only a few drops that evaporate before they can reach the ground (drizzle). In this case the few, but large drops will generate a significant signal and the cloud base would be estimated too low. A detailed description on the use of cloud radar and ceilometer for cloud height estimation is provided by Clothiaux *et al.* (2000). Many relations to convert the radar reflectivity factor Z to the liquid water or ice water content via so-called Z-LWC or Z-IWC relations have been developed. The reader is referred to Sassen and Liao (1996). As indicated above the uncertainty is relatively high. Fox and Illingworth (1997) give an error of 50% for a single LWC retrieval in a nondrizzling stratocumulus cloud.

Liquid water attenuates microwave radiation and can lead to an underestimation of more distant cloud layers. The attenuation is roughly proportional to the liquid water content. As the attenuation increases with frequency, dual wavelength techniques can be used to derive the attenuation and the liquid water content profile, respectively (Gaussiat *et al.*, 2003). Vertical-pointing radars can make use of the Doppler (frequency) shift of backscattered radiation and determine the fall velocity of the cloud particles. While

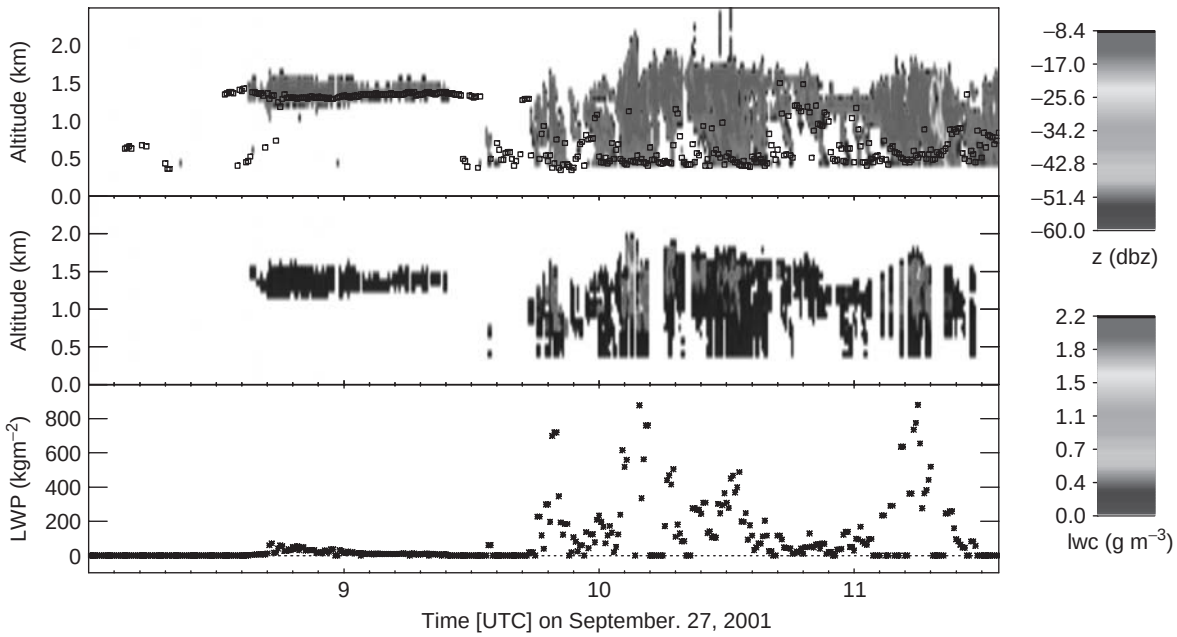


Figure 4 Time series of the radar reflectivity profile measured by a 95 GHz cloud radar on September 27, 2001 at Cabauw, The Netherlands. Squares show the cloud-base height as observed by a lidar ceilometer. The profile of liquid water content was derived from a cloud radar, multifrequency microwave radiometer and auxiliary measurements using the integrated profiling method. The liquid water path is the vertical integral of the profile shown above (courtesy of U. Löhnert). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

cloud droplets do not move vertically, the measurement can identify regions of significant precipitation. Furthermore, snow precipitates much slower than liquid precipitation and the fall velocity can be used as an indicator for precipitation phase. Some radars can also measure polarization quantities. As ice particles cause depolarization due to their irregular shape, this information can be used for phase distinction.

Thermal Infrared

Most water clouds have an emissivity in the infrared spectral region, which is close to one, and thus can be regarded as blackbodies. Therefore, the emitted infrared radiation is directly proportional to the temperature of the cloud boundary via Plancks law. If the atmospheric temperature profile is known, the observed temperature can be converted to cloud height and cloud pressure. The measured radiance may be influenced by water vapor absorption, which however, is low in the atmospheric window between 9.5 and 11.6 μm. In humid conditions the contribution of water vapor between the cloud and the receiver may be as large as a few kelvin. This principle to derive cloud boundary temperature is used from satellites as well as from the ground. Observations from the ground (see Figure 2) show the strong sensitivity of the infrared radiometer; for example, cloud-base temperature can be reliably determined even for clouds with low LWP values (little water) around 4 UTC (in Figure 2). The observations cannot distinguish between water and ice clouds as supercooled clouds can occur down to 233 K, and clouds between 233 and 273 K could be either water, ice, or a mixture.

Very few thin water clouds and many ice clouds have an emissivity significantly below one and as consequence, radiation can penetrate through the cloud, for example, it is semitransparent. Then, a certain part of the measured radiance originates from radiation transmitted through the cloud and the measured signal contains a mixture of both, the radiation behind the cloud and the cloud itself, for example, satellite observations would include the contribution from the warm surface (or a warmer cloud below) leading to a warm bias in the estimated cloud temperature. To avoid this problem observations at several wavelengths are used. Because the wavelength channels are distributed along the wing of a strong absorption band (mostly the CO_2 absorption band centered at 15 μm), the channels have responses from different altitudes. The so-called CO_2 slicing method is applied for instance to MODIS data and derives cloud-top pressure by measuring several channels between 13.3 and 14.2 μm along the strongly absorbing CO_2 band. The more absorbing the band, the more sensitive the channel is to higher clouds. A low cloud will not be visible in measurements from a strongly absorbing channel that is effectively sensitive to higher regions in the atmosphere. If one takes two nearby channels along the band, cloud emissivity can be assumed constant and thus, cancel out if

the ratio of the two channels is taken. The same method is possible using the O_2A-band in the visible spectral range. Cloud-top pressure can be derived with good accuracy from such ratios for mid and high-level clouds. Problems arise in the cold, polar regions.

Reflected Sunlight

Nearly all meteorological satellites employ channels in the visible part of the electromagnetic spectrum where clouds have a high reflectivity, making them so important for the atmospheric energy budget. The measured radiances are mostly converted to reflectances whose values range from 0 to 1. In order to detect the presence of clouds within one satellite pixel, knowledge about the surface reflectivity is required, for example, surface reflectivity maps are generated beforehand under cloud-free conditions. This can be done because changes in surface reflectivity normally are slow (vegetation changes) compared to reflectivity changes due to clouds. The reflectance at a nonabsorbing wavelength, for example, at 0.65 μm, is primarily a function of cloud optical thickness. The reflection function of a near-infrared absorbing band (2.14 μm) is primarily a function of cloud effective radius as small droplets (or ice crystals) reflect more radiation than larger particles. For opaque clouds the simultaneous retrieval of cloud opacity and effective radius from visible and near-infrared reflectivity is well defined (Nakajima and King, 1990). It should be noted, however, that the retrieved effective radius is representative of that portion of the cloud from which the radiation emanates. For an optically thick cloud, this would correspond to the upper portions of the layer. For the retrieval of both optical depth and effective radius, *a priori* information about cloud phase is required, and in the case of ice clouds, assumptions regarding the cross-sectional area for a given crystal size and mass must also be made. Different methods for phase separation are possible. The refractive index of pure water and ice is well separated around 1.7 μm. At this wavelength radiative transfer calculations show that ice crystal reflectivity is approximately constant with wavelength while water cloud reflectivity strongly increases with increasing wavelength. Therefore, high spectral resolution measurements can distinguish cloud phase. Simpler methods compare reflectivities at 1.6 μm and 0.75 μm. Optical depth τ and effective radius r_e can be used to estimate LWP of water clouds by applying the relationship $\mathrm{LWP} = 2\tau\ \mathrm{r}_e/3$ from. If no suitable measurement for deriving the effective radius exists, a value of 10 μm is often assumed. Many more techniques for phase distinction, but also effective radius, liquid water content, and so on exist, which make use of reflected solar radiation and/or infrared brightness temperatures. For example, the retrieval of cloud products from MODIS, which involves channels in the visible and infrared is described in King *et al.* 1992.

Lidar

Cloud particles strongly scatter and absorb the lidar signal at wavelengths in or near the visible part of the spectrum. A water cloud is immediately visible in the measured backscatter profile. Thus, also simple low-power lidar systems – so-called ceilometers – operating unattended (often at airports) can be used to derive the cloud-base height with high accuracy. For this purpose the backscatter signal is converted to the extinction profile, and an extinction threshold is used to estimate cloud-base height. The conversion of the attenuated backscatter signal to extinction is again an underdetermined problem and mostly a constant ratio between lidar backscatter and extinction is assumed.

From the ground the combination of an infrared radiometer and a lidar ceilometer (Figure 2) can provide information about the temperature and the height of the cloud base and therefore gives one point in the temperature profile. The high opacity of water clouds causes the ceilometer signal to be attenuated within the first few hundred meters of the cloud. Therefore higher clouds cannot be detected in the presence of lower water clouds from the ground and vice versa for satellite applications. The lidar backscatter is proportional to the second moment of the DSD. In other words, the backscatter depends on the total cross-sectional area of that the particles in the DSD present to the radiation stream. The cloud radar described above is very sensitive to larger particles (proportional to the sixth moment of the DSD) when the cloud particles are small compared to the radar wavelength (Rayleigh approximation), while the lidar signal is dominated by smaller particles. Therefore the cloud base derived from radar and lidar often does not agree. In the presence of large, precipitating droplets, the radar cloud base (Figure 4) will be lower than the one from the lidar. It should be noted that lidars are more sensitive to thin clouds with small droplets and very often observe clouds that fall under the detection threshold of the radar. It is not uncommon for lidar to observe significantly higher cloud tops in cirrus than cloud radar because of the predominance of small particles near the tops of cirrus layers.

Within the ice clouds the extinction of the lidar signal is much lower. Therefore, the observations have the potential for the derivation of ice cloud properties (particle size, ice–water content). A major difficulty is the correction of uncertainty, given the uncertainty in the extinction-to-backscatter signal. The phase of cloud particles can be detected easily by transmitting linearly polarized light and detecting parallel and orthogonal polarizations in the backscatter. Spherical water drops cause little depolarization, whereas ice particles create depolarization ratios of typically 50%. Information on the derivation of microphysical properties derived from lidar polarization are given by Sassen and Benson (2001).

Conclusions: Clouds

Meteorological satellites have been observing the basic cloud properties, for example, cloud amount, height, and opacity, for several decades by using thermal infrared and reflected solar radiation measurements Long-term products are available by the ISCCP project. Satellite observation of the polar regions is problematic due to the difficulty in separating clouds from the surface. Infrared techniques suffer from the cold temperatures and strong temperature inversions, while the high reflectivity of snow and ice makes clouds difficult to detect from reflected sun light. In general, observations during the night are problematic as only infrared measurements are available. The use of more wavelength channels improves the estimation of cloud phase, particle size, effective radius, and liquid water path. However, this information is mostly limited to the upper part of the cloud. Information about the interior of water clouds is only possible by microwave techniques, as well as active (cloud radar) as passive (radiometer). Owing to the complex microphysical structure, the exact description of the cloud properties can only be gained by sensor combination, which is described below.

INTEGRATED APPROACHES

The description of the various methods used to observe water vapor and clouds given the above discussion reveals the limitations of single instrument techniques. Recently, increased efforts have been made to combine measurements from different instruments. This is true as well for ground-based observation as satellite missions and a combination of both.

In order to observe the atmospheric state as completely as possible, ground-based atmospheric observatories compile a large number of advanced remote sensing and *in situ* sensors. Examples are the Cloud and Radiation Testbeds (CART) of the ARM program in Southern Great Plains, the Tropical Western Pacific, and the North Slope of Alaska, as well as the Coordinated Enhanced Observing Period (CEOP) reference sites like Cabauw (The Netherlands) or Lindenberg (Germany).

On satellites, complementary instruments are operated simultaneously in order to overcome the limitations of single systems. For example, infrared humidity profilers are supplemented by microwave profilers, which have a coarse resolution but provide all weather information. Because measurements of solar radiation have a superior spatial resolution, they can be used to provide subpixel variability of the highly variable cloud field for the other methods. The AIRS sounding suite on NASA's EOS Aqua spacecraft is intended to be used to create global three-dimensional maps of temperature, humidity, and clouds in the Earth's atmosphere with unprecedented accuracy. For that purpose, the AIRS interferometer itself is complemented with a 4-channel visible/near-infrared imaging module, a 15-channel microwave temperature sounder, and a 4-channel microwave humidity sounder (Aumann *et al.*, 2003). To exploit the sensor synergy it is crucial to align the instruments with each other and scan the atmosphere in a synchronized way, as the observation target is highly variable in time and space.

Instead of combining retrieval products derived from single instruments, the direct observables can be integrated in one retrieval system. Here an example for the combination of advanced ground-based measurements including microwave measurements will be given. Both *active* and *passive microwave remote sensing* provide information about microphysical properties of clouds. While the radar provides vertical structure information, microwave radiometers can accurately determine the vertically integrated liquid water content. In the past methods have been developed to combine colocated microwave radiometer and cloud radar measurements to infer LWC and effective radius profiles (Frisch *et al.*, 1998). In principal, such methods use the microwave-derived LWP to scale the LWC derived from the cloud radar reflectivity measurements at each height. However, this involves assumptions about the shape of the DSD and its vertical structure. Multispectral microwave measurements contain additional information on the vertical profiles of temperature, humidity, and to a very limited degree, LWC. Using an Integrated Profiling Technique (IPT) (Löhnert *et al.*, 2004), which applies the optimal estimation theory (e.g. Rodgers 2000), profiles of temperature, humidity, and LWC can be simultaneously retrieved. In contrast to the other methods mentioned above, IPT directly combines 19 microwave brightness temperatures with 95 GHz cloud radar reflectivity profiles, lidar ceilometer cloud base, ground-level measurements of temperature, humidity and pressure, nearest-by radiosonde profiles, and a LWC *a priori* profile obtained from a microphysical cloud model. The advantage of this technique is that the retrieved profiles are physically consistent. This is accomplished by the fact that the retrieved profiles are constrained to the ground-level measurements, they fulfill the condition of saturation within the detected cloud boundaries from ceilometer (base) and cloud radar (top), and their forward-modeled brightness temperatures are constrained to the measured values within their error bars. Momentarily the method is, however, limited to cases of pure water clouds with negligible precipitation (Figure 4). A fundamental requisite in this method is the correct error characterization of the single observations as they determine the degree of how strongly the instruments contribute to the retrieval product. Unrealistic errors in one type of approach might distort completely how other information is used in retrievals. Therefore, much effort needs to be put in the "validation" of error estimates.

Owing to their fine vertical and horizontal resolution, active instruments, in particular, radar and lidar, can be well matched and are uniquely able to provide essential information on the vertical profile. It has already been mentioned that both can provide complementary information about cloud-base height as the lidar is not affected by drizzle. On the other hand, attenuation of the lidar signal is strong and most clouds cannot be penetrated. Therefore cloud-top height derived from cloud radar is more reliable. The complementary nature of lidar and radar not only provides a more complete description of cloud boundaries than either instrument alone but simultaneous lidar and radar retrievals for profiling mean particle size and water content in cirrus clouds are also currently developed (Donovan and van Lammeren, 2001). Ice cloud characteristics can also be derived by combining lidar observations with infrared radiometer observations via the so-called lidar-radiometer (LIRAD) method of ground-based sensors. The method and its application are described in Platt *et al.* (1998).

In addition to the combination of instruments operated on one platform, the synergy between space and ground-based instruments is being pursued. While satellite instruments mostly have good observing capabilities in the middle and upper troposphere, ground-based systems have better resolution in the planetary boundary layer (PBL). Processes in the PBL can change the thermodynamic structure rapidly, which can be well observed in the time series from the ground. Satellite revisit times are less frequent and may be sufficient to detect the development in the upper atmosphere. Clouds block infrared and solar radiation effectively, and the synergy of ground-based and space-borne sensors can provide information above and below the cloud as well as the properties of the cloud boundaries. However, in the simultaneous presence of high and low clouds, information about the midtroposphere can be only gained by employing microwave techniques. An example for this kind of synergy is the blending of ground-based AERI Raman lidar measurements at the ARM CART site with GOES hourly temperature and water vapor retrievals to give high-resolution water vapor profiles (Turner *et al.*, 2000).

CONCLUSIONS

Several remote sensing methods exist to derive water vapor and cloud properties from either the ground or space. However, each of these methods has their limitations and major gaps still exist in the current observing system. Passive sensors generally smooth vertical profiles more than active sensors but provide a better global coverage from space. Infrared and solar techniques can yield good horizontal resolutions but are limited to cloud-free regions. Microwave observations on the other hand can penetrate clouds but have coarse resolutions. In order to overcome the limitations of the different systems, the synergy of different sensors is exploited. For global observations performed from satellites, this is accomplished by combining a complementary suite of instruments on one satellite or arranging complementary satellites in close constellation. In the future, the new generation of sounding interferometers (AIRS, IASI), lidar (CALIPSO), and cloud radar (CloudSat, EarthCare) are expected to provide unprecedented high-quality, high vertical resolution humidity and cloud property measurements from space.

The first spaceborne radar is scheduled to be launched in 2005 within the CloudSat mission (Stephens *et al.*, 2002). It will fly in constellation with four more satellites (Aqua, CALIPSO, PARASOL, and Aura) known as the *A-train*. While Aqua, which has on board MODIS, AIRS, and AMSR, will lead CloudSat by less than 120 s, CALIPSO, which will fly a dual wavelength depolarization lidar, will lag CloudSat only by 15 s. The combined information should give an almost complete picture of the three-dimensional properties of water vapor and clouds. The use of active instruments will for the first time give insight into the vertical cloud structure on the global scale, that is, the occurrence of multilayer clouds, their properties, and their impact on the radiative budget can be investigated.

On the ground, very few, well-equipped reference stations with advanced remote sensing instrumentation exist which are well suited for satellite validation, process studies, and long-term monitoring. To achieve area-wide coverage, much denser networks are needed. Unfortunately, many of the techniques are too costly for this purpose and future work needs to focus on low-cost and low-maintenance operation. Further improvements in observing water vapor and cloud variations are expected to be observed by scanning microwave radiometers and scanning lidars. In order to fully exploit the synergy of the different sensors new algorithms need to be developed. Several retrieval algorithms already exist which combine data from two or three instruments, however, unified retrieval systems that can make use of all available instruments at one site and provide profiles of all meteorological variables and their associated errors are necessary.

It should be noted that most remote sensing methods strongly rely on the accurate knowledge of the radiative transfer in the atmosphere. The need for more accurate models to describe the absorption by the water vapor continuum in the visible, infrared, and microwave range has been the outcome of several studies (Revercomb *et al.*, 2003; Schmid *et al.*, 2001). Potential model improvements will immediately enhance the quality of remote sensing methods.

Acknowledgments

The author would like to thank Dr. Jörg Schulz (DWD) for many helpful discussions and Dr. Ulrich Löhnert for

providing Figure 4. The paper benefited from the comments of many other colleagues and the two reviewers.

REFERENCES

Albert P., Bennartz R. and Fischer J. (2001) Remote sensing of atmospheric water vapor from backscattered sunlight in cloudy atmospheres. *Journal of Atmospheric and Oceanic Technology*, **18**, 865–874.

Aumann H.H., Chahine M.T., Gautier C., Goldberg M.D., Kalnay E., McMillin L.M., Revercomb H., Rosenkranz P.W., Smith W.L., Staelin D.H., *et al.* (2003) AIR S/AMSU/HSB on the AQUA mission: design, science objectives, data products, and processing systems. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 253–264.

Bengtsson L., Robinson G., Anthes R., Aonashi K., Dodson A., Elgered G., Gendt G., Gurney R., Jietai M., Mitchell C. *et al.* (2003) The use of GPS measurements for water vapor determination. *Bulletin of the American Meteorological Society*, **84**, 1249–1258.

Businger S., Chiswell S.R., Bevis M., Duan J., Anthes R.A., Rocken C., Ware R.H., Exner M., VanHove T. and Solheim F.S. (1996) The promise of GPS in atmospheric monitoring. *Bulletin of the American Meteorological Society*, **77**, 5–18.

Clothiaux E.E., Ackerman T.P., Mace G.G., Moran K.P., Marchand R.T., Miller M.A. and Martner B.E. (2000) Objective determination of cloud heights and radar reflectivities using a combination of active remote sensors at the ARM CART sites. *Journal of Applied Meteorology*, **39**, 645–665.

Crewell S., Drusch M., van Meijgaard E. and van Lammeren A.C.A.P. (2002) Cloud observations and modeling within the European BALTEX cloud liquid water network. *Boreal Environmental Research*, **7**, 235–245.

Crewell S. and Löhnert U. (2003) Accuracy of cloud liquid water path from ground-based microwave radiometry. Part II: sensor accuracy and synergy. *Radio Science*, **38**, 8042, doi:10.1029/2002RS002634.

Donovan D.P. and van Lammeren A.C.A.P. (2001) Cloud effective particle size and water content profile retrievals using combined lidar and radar observations. 1. Theory and examples. *Journal of Geophysical Research*, **106**, 27425–27448.

Evans K.F., Walter S.J., Heymsfield A.J. and McFarquhar G.M. (2002) Sub-millimetre-wave cloud ice radiometer: simulations of retrieval algorithm performance. *Journal of Geophysical Research*, **107**, 4028, doi:10.1029/2001JD000709.

Feltz W.F., Smith W.L., Howell H.B., Knuteson R.O., Woolf H. and Revercomb H.E. (2003) Near-continuous profiling of temperature, moisture, and atmospheric stability using the atmospheric emitted radiance interferometer (AERI). *Journal of Applied Meteorology*, **42**(5), 584–597.

Flores A., de Arellano J.V.G., Gradinarsky L.P. and Rius A. (2001) Tomography of the lower troposphere using a small dense network of GPS receivers. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 439–447.

Frisch A.S., Feingold G., Fairall C.W., Uttal T. and Snider J.B. (1998) On cloud radar and microwave measurements of stratus cloud liquid water profiles. *Journal of Geophysical Research*, **103**, 23195–23197.

Fox N.I. and Illingworth A.J. (1997) The potential of a spaceborne cloud radar for the detection of stratocumulus. *Journal of Applied Meteorology*, **36**, 676–687.

Gaussiat N., Sauvageot H. and Illingworth A.J. (2003) Cloud liquid water and ice content retrieval by multi-wavelength radar. *Journal of Atmospheric and Oceanic Technology*, **20**, 1264–1275.

Grody N.C. (1993) Remote sensing of the atmosphere from satellites using microwave radiometry. In *Atmospheric Remote Sensing by Microwave Radiometry*, Janssen M.A. (Ed.), John Wiley & Sons: New York, pp. 259–334.

Güldner J. and Spänkuch D. (1999) Results of year-round remotely sensed integrated water vapor by ground-based microwave radiometry. *Journal of Applied Meteorology*, **38**, 981–988.

Güldner J. and Spänkuch D. (2001) Remote sensing of the thermodynamic state of the atmospheric boundary layer by ground-based microwave radiometry. *Journal of Atmospheric and Oceanic Technology*, **18**, 925–933.

Janssen M.A. (1993) *Atmospheric Remote Sensing by Microwave Radiometry*, John Wiley & Sons: New York, p. 572.

King M.D., Kaufman Y.L., Menzel W.P. and Tanre D. (1992) Remote sensing of cloud, aerosol, and water vapor properties from the moderate resolution imaging spectrometer (MODIS). *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 2–27.

Lerner J.A., Weisz E. and Kirchengast G. (2002) Temperature and humidity retrieval from simulated Infrared Atmospheric Sounding Interferometer (IASI) measurement. *Journal of Geophysical Research*, **107**, 4–1–4–11, 10.1029/2001JD900254.

Li J., Wolf W.W., Menzel W.P., Zhang W.J., Huang H.L. and Achtor T.H. (2000) Global soundings of the atmosphere from ATOVS measurements: the algorithm and validation. *Journal of Applied Meteorology*, **39**, 1248–1268.

Löhnert U., Crewell S. and Simmer C. (2004) An integrated approach towards retrieving physically consistent profiles of temperature, humidity, and cloud liquid water. *Journal of Applied Meteorology*, **43**, 1295–1307.

Nakajima T. and King M.D. (1990) Determination of the optical thickness and effective particle radius of clouds from reflected solar radiation measurements. Part I: theory. *Journal of the Atmospheric Sciences*, **47**, 1878–1893.

Platt C.M.R., Young S.A., Manson P.J., Patterson G.R., Marsden S.C., Austin R.T. and Churnside J.H. (1998) The optical properties of equatorial cirrus from observations in the ARM pilot radiation observation experiment. *Journal of the Atmospheric Sciences*, **55**, 1977–1996.

Randel D.L., Vonder Haar T.H., Ringerud M.A., Stephens G.L., Greenwald T.J. and Combs C.L. (1996) A new global watervapor dataset. *Bulletin of the American Meteorological Society*, **77**, 1233–1246.

Revercomb H.E., Turner D.D., Tobin D.C., Knuteson R.O., Feltz W.F., Barnard J., Bösenberg J., Clough S., Cook D., Ferrare R., *et al.* (2003) The ARM program's water vapor intensive observation periods. *Bulletin of the American Meteorological Society*, **84**, 217–236.

Rodgers C.D. (2000) *Inverse Methods for Atmospheres: Theory and Practice*, World Scientific: Singapore, p. 238.

Rossow W.B. and Duenas E. (2004) The International Satellite Cloud Climatology Project (ISCCP) web site: an online resource for research. *Bulletin of the American Meteorological Society*, **85**, 167–172, doi:10.1175/BAMS-85-2-167.

Rossow W.B. and Schiffer R.A. (1999) Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, **80**, 2261–2287.

Sassen K. and Benson S. (2001) A midlatitude cirrus cloud climatology from the facility for atmospheric remote sensing. Part II: microphysical properties derived from lidar depolarization. *Journal of the Atmospheric Sciences*, **15**, 2103–2112.

Sassen K. and Liao L. (1996) Estimation of cloud content by W-band radar, *Journal of Applied Meteorology*, **35**(6), 932–938.

Schmid B., Michalsky J., Slater D., Barnard J., Halthore R., Liljegren J., Holben B., Eck T., Livingston J., Russell P. *et al.* (2001) Comparison of columnar water-vapor measurements from solar transmittance methods. *Applied Optics*, **40**, 1886–1896.

Soden B.J., Turner D.D., Lesht B.M. and Miloshevich L.M. (2004) An analysis of satellite, radiosonde, and lidar observations of upper tropospheric water vapor from the Atmospheric Radiation Measurement Program. *Journal of Geophysical Research*, **109**(D4), D04105.

Stephens G.L., Vane D.G., Boain R.J., Mace G.G., Sassen K., Wang Z., Illingworth A.J., O'Connor E.J., Rossow W.B., Durden S.L. *et al.* (2002) The CLOUDSAT mission and the A-Train. *Bulletin of the American Meteorological Society*, **83**, 1771–1790.

Suggs R.J., Jedlovec G.J. and Guillory A.R. (1998) Retrieval of geophysical parameters from GOES: evaluation of a split-window technique. *Journal of Applied Meteorology*, **37**, 1205–1227.

Turner D.D., Feltz W.F. and Ferrare R.A. (2000) Continuous water vapor profiles from operational ground-based active and passive remote sensors. *Bulletin of the American Meteorological Society*, **81**, 1301–1317.

Turner D.D. and Goldsmith J.E.M. (1999) Twenty-four-hour Raman lidar water vapor measurements during the Atmospheric Radiation Measurement Program's 1996 and 1997 water vapor intensive observation periods. *Journal of Atmospheric and Oceanic Technology*, **16**, 1062–1076.

Weckwerth T.M., Parsons D.B., Koch S.E., Moore J.A., LeMone M.A., Demoz B.B., Flamant C., Geerts B., Wang J. and Feltz W.F. (2004) An overview of the international H_2O project (IHOP_2002) and some preliminary highlights. *Bulletin of the American Meteorological Society*, **85**, 253–277.

Westwater E. (1978) The accuracy of water vapor and cloud liquid determination by dual-frequency ground-based microwave radiometry. *Radio Science*, **13**, 667–685.

Wickert J., Schmidt T., Beyerle G., Konig R. and Reigber C. (2004) The radio occultation experiment aboard CHAMP: operational data analysis and validation of vertical atmospheric profiles. *Journal of the Meteorological Society of Japan*, **82**, 381–395.

Wulfmeyer V. and Bösenberg J. (1998) Ground-based differential absorption lidar for water-vapor profiling: assessment of accuracy, resolution, and meteorological applications. *Applied Optics*, **37**, 3825–3844.

PART 6

Soils

66: Soil Water Flow at Different Spatial Scales

JAN W HOPMANS AND GERRIT SCHOUPS

Department of Land, Air and Water Resources, University of California, Davis, CA, US

A major challenge in hydrological sciences is the modeling of flow and transport processes and their measurement across a range of spatial or temporal scales. Such needs arise, for example, when watershed processes must be determined from soil hydrological data collected from a limited number of in situ field measurements or analysis of small soil cores in the laboratory. The scaling problem cannot be solved by simple consideration of the differences in space or timescale, for several reasons. First, spatial and temporal variability in soil hydrological properties create uncertainties when changing between scales. Second, flow and transport processes in vadose zone hydrology are highly nonlinear. As a result, vadose zone properties are nonunique and scale-dependent, resulting in effective properties that vary across spatial scales and merely serve as calibration parameters in hydrologic models. Therefore, their estimation for heterogeneous materials can only be accomplished using scale-appropriate measurements and models. We present examples of soil water flow at the pore, local, and regional scales. The inherent complexity of flow in heterogeneous soils, and the need to integrate theory with experiment, requires innovative and multidisciplinary research efforts to overcome limitations imposed by current understanding of scale-dependent soil water flow and transport processes.

INTRODUCTION

The unsaturated zone is bounded by the soil surface at one end and merges with the groundwater in the capillary fringe at the lower end. The distinction between groundwater and the unsaturated zone is usually made within a hydrologic context, emphasizing water as the agent of change of the subsurface and the main driver for transport of chemicals between the atmosphere and groundwater. To emphasize the profound influence of soil chemical and biological processes on water flow and chemical transport, one may generally refer to the unsaturated flow domain as the vadose zone.

The upper part of the vadose zone is the most dynamic, and changes occur at increasingly smaller time and spatial scales when moving from the groundwater towards the soil surface. The most upper part of the vadose zone is subject to fluctuations in water and chemical content by infiltration and leaching, water uptake by plant roots (transpiration), and evaporation from the soil surface. Water is the main ingredient leading to soil formation from the weathering of parent material such as rock or transported deposits, with additional factors, such as climate, vegetation, topography, and parent material, determining soil physical properties. Generally, the soil depth is controlled by the maximum rooting depth (generally within a few meters from the soil surface). However, the vadose zone can extend much deeper than the surficial soil layer and includes unsaturated rock formations and alluvial materials to depths of 100 m or more, determined by hydrologic, topographic, and lithographic characteristics. In the last few decades, research interest in the deeper vadose zones has increased, instigated by the need to sustain quality of groundwater and resources for drinking water and ecological purposes. Scientists are becoming increasingly aware that soil is a critically important component of the earth's biosphere, not only because of its food production function, but also as the safekeeper of local, regional, and global environmental quality. For example, it is believed that management strategies in the unsaturated soil zone will offer the best opportunities for preventing or limiting pollution, or for remediation of ongoing pollution problems. Because chemical residence times in groundwater aquifers can range from years to thousands of years, pollution is often essentially irreversible. Prevention or remediation of soil and groundwater

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

contamination starts, therefore, with proper management of the vadose zone.

Transient isothermal unsaturated soil water flow is generally described by the so-called Richards' equation,

$$\frac{\partial \theta}{\partial t} = \nabla \cdot [K(\theta)\nabla(h_m + z)] + S(t) \qquad (1)$$

which solves for the soil water matric potential (h_m), water content (θ), and water-flux density as a function of time and space, using one, two, or three-dimensional flow models. In equation (1), $S(t)$ represents a sink/source term that is routinely used to describe plant root water uptake, K is the unsaturated hydraulic conductivity tensor (L T^{-1}), and z denotes the gravitational head (L) to be included for the vertical flow component only. The relationship between h_m and θ is determined by the soil water retention function, of which the slope, $C(h_m)$, is the so-called soil water capacity. Boundary conditions must be included to allow for specified soil water potentials and fluxes at all boundaries of the soil domain, whereas user-specified initial conditions and time-varying source/sink terms need be specified. Both the soil water retention and unsaturated hydraulic conductivity relations (in combination referred to as soil hydraulic functions) are highly nonlinear, with both h_m and K varying many orders of magnitude over the water content range of significant water flow.

Many analytical and numerical mechanistic flow models have been developed to solve equation (1), for specific agricultural or environmental applications. In short, the dynamic water flow equation is a combination of the steady state Darcy expression and a mass balance formulation. Using various solution algorithms, the soil region of interest is discretized in finite-size elements that can be one, two, or three-dimensional. Numerical solution requires that mass balance be maintained within each small volume element within the soil domain at all times. Richards equation is typically a highly nonlinear partial differential equation, and is therefore extremely difficult to solve numerically because of the largely nonlinear dependencies of both water content and unsaturated hydraulic conductivity on the soil water matric potential (h_m). Both the soil water retention and unsaturated hydraulic conductivity relationships must be known *a priori* to solve the unsaturated water flow equation. Although soil water retention measurements are time-consuming (Dane and Hopmans, 2002), unsaturated hydraulic conductivity data are even much more difficult to obtain from measurements (Klute and Dirksen, 1986). Functional unsaturated hydraulic conductivity models, based on pore size distribution, pore geometry, and connectivity, require integration of soil water retention models to obtain analytical expressions for the unsaturated hydraulic conductivity. The resulting expressions relate the relative hydraulic conductivity K_r, which is defined as the ratio of the unsaturated hydraulic conductivity K to the saturated hydraulic conductivity K_s, to the effective saturation to yield a macroscopic hydraulic conductivity expression. Pedotransfer function models (*see* **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**) have been developed to estimate soil hydraulic parameters for various functional models (**http://www.usyd.edu.au/su/agric/acpa/software/multistep.htm**).

Soil hydraulic functions that characterize flow and transport processes at larger spatial scales are mostly obtained from relatively small measurement scales. For example, prediction of soil water dynamics at the field scale is routinely derived from the measurement of soil hydraulic properties from laboratory cores, collected from a limited number of sampling sites across large spatial extents, often using large sampling spacings. Typically, the measurement scale for soil hydraulic characterization is in the order of 10 cm, with a sample spacing of 100 m or larger. Soil hydrological parameters obtained from these centimeter-scale measurements are subsequently included in numerical models with a grid or element size ten times as large or larger, with the numerical results extrapolated to field-scale conditions. Because of the high nonlinearity of the soil hydraulic functions, their application across spatial scales is inherently problematic. Specifically, the averaging of processes determined from discrete small-scale samples may not describe the true soil behavior involving larger spatial domains. In addition, the dominant hydrological flow processes may vary between spatial scales, so that potentially different models need to be used to describe water flow at the soil pedon, field scale, or watershed scale. In 1991, the US National Research Council (NRC, (1991)) identified the scaling of dynamic nonlinear behavior of hydrologic processes as one of the priority research areas that offer the greatest expected contribution to a more complete understanding of hydrologic sciences.

Many field experiments have confirmed that soil heterogeneity controls the hydrology of flow and transport, including preferential flow such as through cracks by soil shrinkage (*see* **Chapter 67, Hydrology of Swelling Clay Soils, Volume 2**). Although some hydrological studies successfully applied a deterministic approach, other studies showed the need for either distributed physically based modeling or stochastic modeling at the watershed scale, mostly because deterministic models require an enormous amount of data to accurately represent the multidimensional soil heterogeneity. Alternatively, the conceptual characterization of flow at the large scale may be simplified by modeling the key flow mechanisms for representative elementary areas (REA's) within the larger domain, using REA-appropriate effective hydrological parameters (Blöschl *et al.*, 1995). The need to incorporate the spatial organization of these key properties,

such as the soil hydraulic functions, is now also recognized in soil science. Specifically, we refer to the treatise by Roth *et al.* (1999), outlining a conceptualization of the control of soil heterogeneity and its spatial organization on soil flow and transport processes, using the so-called scaleway approach. In this approach, the three-dimensional field domain is defined by structural and textural elements. The various identified structural elements describe the dominating physical features that affect the hydrological mechanisms operating at the larger spatial scale, whereas the textural patterns within the structural units merely cause perturbations of these main hydrological processes. This hierarchical notation of structure and texture within a hydrological context should not be confused with soil texture and soil structure, although it can be argued that soil properties such as soil structure and texture may have a dominant control on soil hydrology. Thus, characterizing soil hydraulic variability is predicated on identifying those soil hydrological units that cause major differences in soil water regime at the spatial scale of interest. In soil hydrological studies, these structural units across the landscape may be defined by soil map units (Ferguson and Hergert, 1999). The smaller spatial scale level of textural information within the larger structural units can be determined either deterministically or stochastically, for example, through scaling of soil hydraulic properties from laboratory soil cores (Hopmans and Stricker, 1989). The upscaling from the textural to the structural scale level may result in effective, scale-appropriate soil hydraulic functions that may differ in form and parameter values between spatial scales, but serve a similar function in equation (1). Much of the issues associated with spatial scaling of soil hydrological processes were presented in Hopmans *et al.* (2002a). It is also the focus of this treatise to review the current state-of-the-art of solving the unsaturated water flow equation (1) across spatial scales.

SCALE-DEPENDENCY OF SOIL PROPERTIES AND PROCESSES

Upscaling requires integration and aggregation of spatial information into larger spatial units, for example, as in the estimation of an effective field soil water retention or conductivity curves from small-scale laboratory core measurements. As pointed out by Baveye and Boast (1999), Darcy's experiment can in effect be interpreted as yielding an upscaled, effective saturated hydraulic conductivity. In contrast, downscaling is the disaggregation of scale into smaller scales, for example, as in distributed hydrological modeling. Baveye and Boast (1999) discuss the confusion on concepts of scales in vadose zone hydrology. In contrast to the "perceived" clear hierarchy of spatial and temporal scales, they point out that the differentiation between scales is partly arbitrary, and depends, for example, on the scale of measurement and system scale. Moreover, whatever the spatial scale considered, it has its own characteristic dynamics, and should be treated that way, so that focus should be placed on experimentation and measurement methods that are representative for the different scales. Loosely, one may define the microscopic pore scale, the local scale, and the regional scale such as agricultural fields and watersheds. The macroscopic Darcy equation is considered to be valid for the local scale, with a typical size in the range of centimeters to meters. It is the local scale for which equation (1) is considered to be generally valid. It is also this scale, for which the Darcy equation can be derived from the volume averaging of the Stokes equations at the microscopic scale level. Increasingly, innovative measurements and modeling techniques are becoming available to measure and model pore size scale properties and processes. The regional scale typically applies to agricultural fields and watersheds for which the relevant soil hydrological properties become increasingly nonstationary. Yet, it is at these larger spatial scales that solutions are increasingly sought, putting the validity of equation (1) into question.

When increasing the spatial scale, soil properties typically become nonstationary, as evidenced by the delineation of soil map units by a soil survey. As one moves through a sequence of increasing sampling scales, nonstationarities at smaller spatial scales may be eliminated, and soil properties may change from deterministic to random, with the smaller-scale variations filtered out by the larger scale process or aliasing (Kavvas, 1999). The spatial organization and its evolvement across spatial scales can be viewed as continuous with deterministic patterns evolving as the field-of-view changes. At each field-of-view, the large-scale variation can be regarded as deterministic, whereas the smaller scale variations within each main unit can be treated stochastically. As explained earlier, this hierarchy of spatial scales (Cushman, 1990) can be described by structural and textural elements, as defined by the scaleway approach of Roth *et al.* (1999), with the structural elements describing the dominating soil patterns that affect the hydrological mechanisms operating at the *a priori* defined field-of-view. For the purpose of using soil information towards hydrological modeling, one must be careful on focusing on soil properties only, since spatial patterns of soil properties may be different from the functional organization of soil hydrological processes that may also be determined by landscape position and land use. It may be argued that as different flow processes may be dominant at each scale, different mathematical relationships may be required to describe the underpinning physical process at each spatial scale.

According to the scaleway approach, subsequent aggregation of information and the modeling of flow and transport at one specific scale, provides the required information

at the next, larger scale level. For example, if the scale of interest is an agricultural field, one defines the structural elements based on the dominant soil hydrological mechanism that causes the major differences in soil water regime between structural units. Most recently, Becker and Braun (1999) defined these units as hydrotopes or hydrological response units (HRU's) based on differences between vegetation types, shallow groundwater presence, soil type or hillslope. Likewise, in his review on scale issues in hydrological models, Beven (1995) introduced the simple patch model for scale-dependent modeling, with a patch defined as any area of the landscape that has broadly similar hydrological response in terms of the quantities of interest.

SPATIAL SCALING APPROACHES

In soil hydrological studies, soil map units may define the structural units across the landscape (Ferguson and Hergert, 1999), or soil types may need to be regrouped, classifying soils by their hydrologic functioning (Dunn and Lilly, 2001). By combining GIS with fuzzy logic techniques, Zhu and Scott Mackay (2001) generated spatial continuous soil information as input to a distributed watershed model. This provided a gradual transition between HRU's, depending on soil type, position in the landscape, and land use, instead of the discrete and generally coarse resolution of soil maps. This distribution or disaggregation of a watershed in structural units is deterministic (distributed modeling) and their aggregation to the scale of interest may be possible by simple mass conservation principles. The selection of the main HRU's or hydrotopes with their corresponding effective hydrological parameters is determined by calibration, such as was presented by Eckhardt and Arnold (2001) using parameter optimization. Rather than expecting a unique solution for the distributed hydrologic parameters, nonlinearity and measurement uncertainty, Beven (2001) introduced the equifinality concept, indicating that many solutions may be acceptable. Recent optimization algorithms now allow for multiple objective functions for multiparameter distributed watershed modeling, with effective hydrologic parameter values determined by the choice of the most relevant hydrological variables. Increasingly, efficient global optimization algorithm's are developed such as the shuffled complex evolution (SCE) algorithm (Duan *et al.*, 1993) to calibrate for a large number of parameters for distributed watershed models. Madsen (2003) demonstrated the successful application of SCE using multiple objective functions for calibration of a multiparameter distributed watershed model. He showed that if multiple measurement types of different importance are available, a so-called Pareto set of solutions provides trade offs between the different objectives. The current state-of-the-art in calibration of watershed models is presented in Duan *et al.* (2003), and appears ready to be applied to unsaturated water flow modeling (*see* **Chapter 77, Inverse Modeling of Soil Hydraulic Properties, Volume 2**).

Most of the uncertainty in the assessment of water flow in unsaturated soils at the field scale can be attributed to soil spatial variability caused by soil heterogeneity (*see* **Chapter 79, Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport, Volume 2**). The exact nature of the functional dependence of both soil water retention and unsaturated hydraulic conductivity with water content differs among soil types with different particle size compositions, and pore size geometry within a heterogeneous field soil. The scaling approach has been extensively used to characterize soil hydraulic spatial variability, and to develop a standard methodology to assess the variability of soil hydraulic functions and their parameters. The single objective of scaling is to coalesce a set of functional relationships into a single curve using scaling factors that describe the set as a whole (e.g. structural unit). The procedure consists of using scaling factors to relate the hydraulic properties at a given location to the mean properties at an arbitrary reference point. This physically based scaling concept provides for the simultaneous scaling of the soil water retention (Kosugi and Hopmans, 1998) and unsaturated hydraulic conductivity functions (Tuli *et al.*, 2001), leading to scaled mean soil hydraulic functions for each structural unit, to serve as effective soil hydraulic functions (e.g. Mohanty *et al.*, 1997).

Stochastic approaches to upscale soil hydrological processes from the local to the regional scale include Monte-Carlo (MC) analysis and solution of a stochastic form of equation (1). A common assumption in using the stochastic approach is that of ergodicity, that is, the ensemble average is equivalent to the spatial average. Typically, in MC analysis, numerical solutions of equation (1) are repeated using different realizations of spatially variable soil hydrological properties. Realizations are generated using a random number generator from *a priori* knowledge of the statistical distribution of the parameter(s) in questions, including their spatial correlation. Repeated solutions will yield the stochastic information needed for the relevant output variables. An example of this approach was used by Hopmans and Stricker (1989). In this study, scaling factors and soil hydraulic functions were generated for different soil map units and soil horizons to study the soil water regime in a 650-ha watershed, specifically the influence of soil heterogeneity on plant transpiration.

Although MC analysis is conceptually straightforward, it can become computationally very intensive. Another approach is to use a perturbation approximation of equation (1), by representing local soil hydraulic properties as the sum of a deterministic and a stochastic component (Zhang, 2002). The result is that the local Richards equation (1) with deterministic parameter values is replaced by

an upscaled equation with stochastic parameters, that is, the means, variances, and covariances of hydrological parameter values. A disadvantage of this method is that it is mostly restricted to relatively small fluctuations of the stochastic parameters. A third approach assumes that local-scale flow can be simplified so that analytical expressions may be derived that describe the flow statistics at a larger scale. As an example, Chen *et al.* (1994) present a solution for the area-averaged Green and Ampt infiltration model that is applicable for large parameter uncertainties.

The inverse method offers a powerful procedure to estimate flow properties across spatial and temporal scales. As numerical models have become increasingly sophisticated and powerful, inverse methods are applicable to laboratory and field data, no longer limited by the physical dimensions of the soil domain, or type of imposed boundary conditions. Inverse methods might be especially appropriate for estimating regional-scale effective soil hydraulic parameters, from boundary condition measurements. For example, Eching *et al.* (1994) estimated field-representative hydraulic functions using inverse modeling with field drainage flow rate serving as the lower boundary condition for the Richards flow equation applied at the field scale. The application of inverse modeling to estimate soil hydraulic functions for laboratory soil cores has been extensively reviewed by Hopmans *et al.* (2002b). Although the inverse method may suffer from nonuniqueness (e.g. Beven, 2001), the application of inverse methods in general to estimate soil hydraulic functions across spatial scales is very promising, yielding effective hydraulic properties that pertain to the scale of interest.

EXAMPLES OF SOIL WATER FLOW AT DIFFERENT SPATIAL SCALES

The application of the various modeling techniques for unsaturated water flow is presented. First, at the microscale, the measurement of porosity at the pore scale is demonstrated using X-ray computed tomography. Examples are presented for local-scale unsaturated water flow models, whereas the application of using effective soil hydraulic functions is presented at the regional scale for an irrigation water district in California, USA.

Pore Scale

Although equation (1) is not applicable at the pore scale, this example is presented to demonstrate the existence of a Representative Elementary Volume (REV) for porosity, for the first time as it is known (Clausnitzer and Hopmans, 1999), using X-ray computed tomography (CT). Using the three-dimensional spatial distribution of X-ray attenuation as a proxy, porosity measurements for a glass-bead medium were conducted for increasing measurement volumes. X-ray CT measurements were conducted in a random pack of uniform glass beads within a vertical Plexiglas cylinder of 4.76 mm inner diameter. The bead diameter, d_p, was 0.5 mm and the spatial resolution was 18.4 μm, resulting in $(18.4\,\mu m)^3$ voxel volumes (see Figure 1a). In this example, the single structural unit is represented by the glass beads pack, and textural variations are defined by porosity changes at a measurement scale larger than the REV. Starting from the original three-dimensional data set of attenuation values, increasingly larger volumes were extracted, all centered at the same location, beginning with $8 \times 8 \times 8$ voxels and incrementing the cube side length, L, of the averaging volume by 4 voxel lengths (0.0736 mm) in each step. The sequence of porosity calculations with increasing volume size was conducted twice, first with the initial $8 \times 8 \times 8$ averaging volume centered in the air phase, and subsequently with the averaging volume centered in the glass phase. The resulting curves are presented in Figure 1(b), suggesting a REV of about three to five times the bead diameter. Using lattice-Boltzmann modeling, Zhang *et al.* (2000) showed that the REV may depend on the specific soil property measured. In a subsequent study, Clausnitzer and Hopmans (2000) showed that X-ray CT can be used to measure the spatial and temporal distribution of a tracer through this glass-bead medium with a spatial resolution of about 85 microns. Although we have a general good understanding of macroscopic flow and transport, much additional work is needed to describe processes at the scale of pores, thereby improving our understanding of the effects of variations in pore-water velocity and air–water interfaces on flow and transport in unsaturated porous media. Numerical modeling techniques are now becoming available to solve for streamlines and velocities at the pore scale by lattice gas automata (LGA) and the lattice–Boltzmann method.

Local Scale

Most unsaturated water flow and transport models have been developed for applications at the local scale, that is, the laboratory column and field plot scale. Whereas the initial applications were dominantly agricultural, many recent applications are mostly environmental. Also, whereas the earlier models focused on unsaturated water flow only, recent applications require coupling of water flow with models that simulate chemical and biological processes. Although many excellent models are available, we present only a few here as these appear to be applied more often than others. RZWQM (Ahuja *et al.*, 1999) is an integrated physical, biological, and chemical one-dimensional process model, simulating crop growth and movement of water, nutrients, and pesticides over and through the root zone. The model includes a generic crop-growth simulator, estimates soil evaporation and plant transpiration, and links total root water and nutrient extraction to atmospheric demand. The (*see* **Chapter 78, Models**

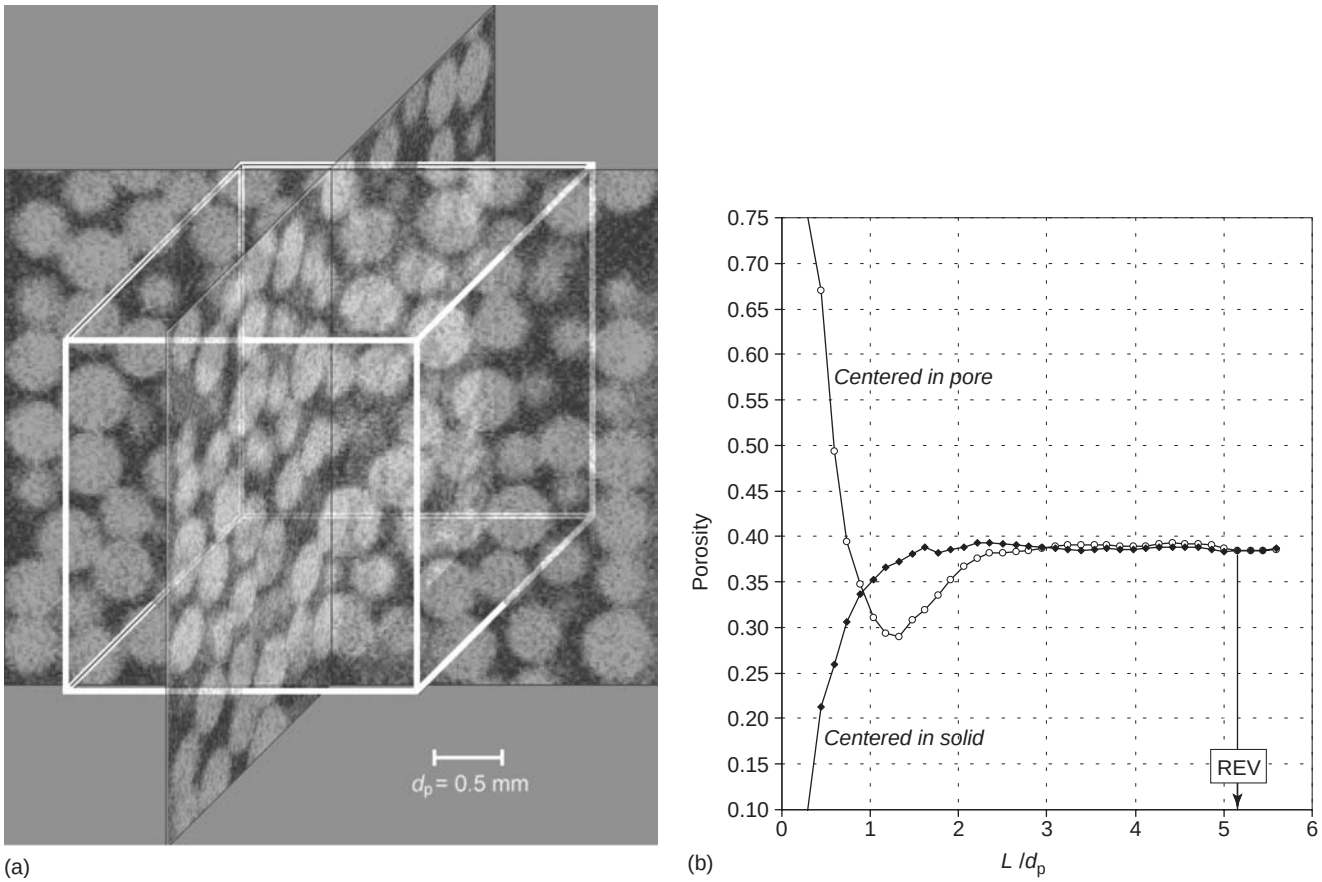


Figure 1 (a) Three-dimensional image of dry glass beads (light grey) and pore space occupied by air (dark grey). Two vertical cross-sections of a glass-bead pack with 2.576 mm side length. (b) Estimated porosity for a cubic domain of increasing size within a glass-bead pack, centered either within the air or the glass phase. Cube side length is expressed as multiple of the bead diameter d_p (0.5 mm) (Reprinted from Clausnitzer and Hopmans (1999). © 1999, with permission from Elsevier)

of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2) SWAP model (Van Dam *et al.*, 1997) combines a one-dimensional water flow and nutrient transport model with a universal crop-growth simulator. Since SWAP has been designed for interactions with surface water and regional drainage, applications are primarily at the field scale. Increased computer capability and development of more efficient computer algorithms has increased the spatial dimension of solutions of equation (1) from one-dimensional to two and three-dimensional. An example of a three-dimensional water flow, coupled with nutrient transport, root water and nutrient uptake, and root growth models was presented by Somma *et al.* (1998).

The application of the aforementioned calibration techniques to a field-scale level was presented in de Vos *et al.* (2002). In this study, four major hydrologic zones with different soil hydraulic functions were identified in a tile-drained field. Soil water matric potential, groundwater level, and piezometric heads, at various locations within the experimental field, field discharge rate and nitrate concentrations were measured during a four-month leaching period. Soil water retention, saturated and unsaturated hydraulic conductivity data for four distinct hydrological soil units were measured from laboratory soil cores. The HYDRUS-2D model (**`http://www.ussl.ars.usda.gov/models/hydrus2d.HTM`**), as presented by Šimunek *et al.* (1999) was used to simulate two-dimensional flow regime and nitrate transport for a 2 by 6 m field plot, matching drainage rates and nitrate concentrations in the drain outlet. Field-effective soil water retention and hydraulic conductivity functions were estimated using an inverse modeling approach, by adjusting the hydraulic parameters that were measured from the laboratory soil cores. The study concluded that effective hydraulic properties were able to describe the average transient soil water behavior for the heterogeneous soil system, as determined from two-dimensional transient water flow modeling. Mohanty *et al.* (1997) demonstrated the application of a two-dimensional local-scale model to describe preferential

flow of a tile-drained field using field-averaged piecewise-continuous hydraulic functions for different soil horizons. Successful applications such as these and the SWAP model demonstrate that local-scale models can be used to spatial scales as large as agricultural fields.

Regional Scale

Soil hydrological information can be fairly easily assigned to local-scale unsaturated water flow models, using small-scale soil properties measurements in the laboratory or field scale. However, in applications at the field to watershed scale, a prohibitively large number of sampling sites are needed to characterize the vadose zone. An alternative approach is to estimate effective values for the hydraulic parameters by inverse modeling. In the presented example study, the inverse approach is applied to the 9700-acres Broadview irrigation district that is located in the San Joaquin Valley, California. The area consists of about 60 tile-drained 160-acre fields. Drains are located at a depth of approximately 2 m with horizontal spacings of 100–200 m. Subsurface drainage flow is measured weekly at 25 sumps, hence each sump collects subsurface drainage from one or more fields. The main crops in the area are cotton, tomatoes, alfalfa, melons, and wheat. Fields are either furrow or sprinkler irrigated. The amount of irrigation water applied to each field is measured every two days from water meters at irrigation turnouts. Rainfall data were obtained from a nearby weather station. A total of 48 shallow groundwater wells are distributed throughout the district from which monthly water table depth readings were available. Topography is nearly flat. The majority of the soils are mapped as vertisols with an average clay content of 50%. The numerical model used in this study was MODHMS (**http://www.modhms.com/modhms.cfm**), which is a MODFLOW-based distributed watershed model. It simulates in an integrated manner evapotranspiration, overland flow, channel flow, and subsurface flow and transport. Three-dimensional variably saturated subsurface flow was simulated with the three-dimensional Richards equation. Input data include spatially distributed crop types, and weekly irrigation and rainfall amounts for each field in the district. Subsurface drains are simulated using a head-dependent function, with drain discharge proportional to head above the drain and the drain conductance, C_d. Separation of crop transpiration and soil evaporation was based on a method developed by Allen *et al.* (1998). Field-specific actual crop transpiration values were determined by a depth-dependent root water uptake distribution, a time-varying crop coefficient, and a water-stress response function. Actual field evaporation was estimated from a wetness function that defines the reduction of soil evaporation with soil surface water content. The Broadview water district was divided into $1\,536\,200 \times 200$ m square cells, using approximately 16 grid cells per field. Vertically, the flow system was divided into four layers with layer thicknesses of 0.9 m, 0.9 m, 1.8 m, and 2.4 m, with a total model thickness of 6 m, resulting in a total of 6144 cells. Adaptive time-stepping was used to solve the variably saturated flow equation. Simulations were done for the 1998-crop year, which started on October 1, 1997, and ended on September 30, 1998. The water flux between the bottom of the model domain at the 6-m depth and the regional aquifer was simulated assuming a linear relation between the head difference of the bottom grid cell and a 10-m deep measured total head in the district, with a proportionality constant defined by the regional conductance, C_b. It was assumed that no lateral flow occurred across the other district boundaries.

Reference soil hydraulic functions $K_{ref}(\theta)$ and $h_{ref}(\theta)$ for clay were estimated based on the neural network analysis (**http://www.ussl.ars.usda.gov/models/rosetta/rosetta.HTM**) of Schaap *et al.* (1998). As part of the calibration, these functions were scaled using a single district-wide scaling factor, α, relating the effective water district hydraulic properties to the reference functions. Calibration parameters included α, C_d, C_b, and the initial head in the unsaturated zone h_{ini}. The latter calibration parameter was added, since it was very sensitive to the initiation of subsurface drainage. These parameters were estimated from weekly district-scale drainage discharge measurements and monthly district-average water table levels using an inverse modeling approach based on the Levenberg–Marquardt algorithm of PEST (**http://www.parameter-estimation.com/html/pest_overview.html**), while minimizing the residuals between simulated and measured drainage discharge and water table levels. Average evapotranspiration (ET), infiltration, drainage rate, and groundwater table levels are presented in Figure 2(a). Optimized district-wide parameter values were: $\alpha = 10.0$, $C_d = 0.08$ m yr^{-1}, $C_b = 3.6 \times 10^{-4}$ m yr^{-1}, and $h_{ini} = -380$ cm. A comparison of measured with optimized drainage rates and groundwater tables is presented in Figure 2(b).

As can be seen in Figure 2(b), the approach of treating the entire district as a homogeneous system in terms of hydraulic properties works surprisingly well, if the interest is only in predicting district-average hydrological conditions. Calibration results were disappointing when comparing drainage data from individual sumps within the district, even though field-specific boundary conditions were used. The optimized district-representative scaling factor was much larger than one, indicating that the district-wide response to irrigation is much faster than that initially estimated from pedotransfer functions on laboratory-scale measurements. This result indicates preferential water flow at this larger scale. However, physical interpretation of the optimized hydraulic functions is difficult since they represent effective properties that are likely not a function

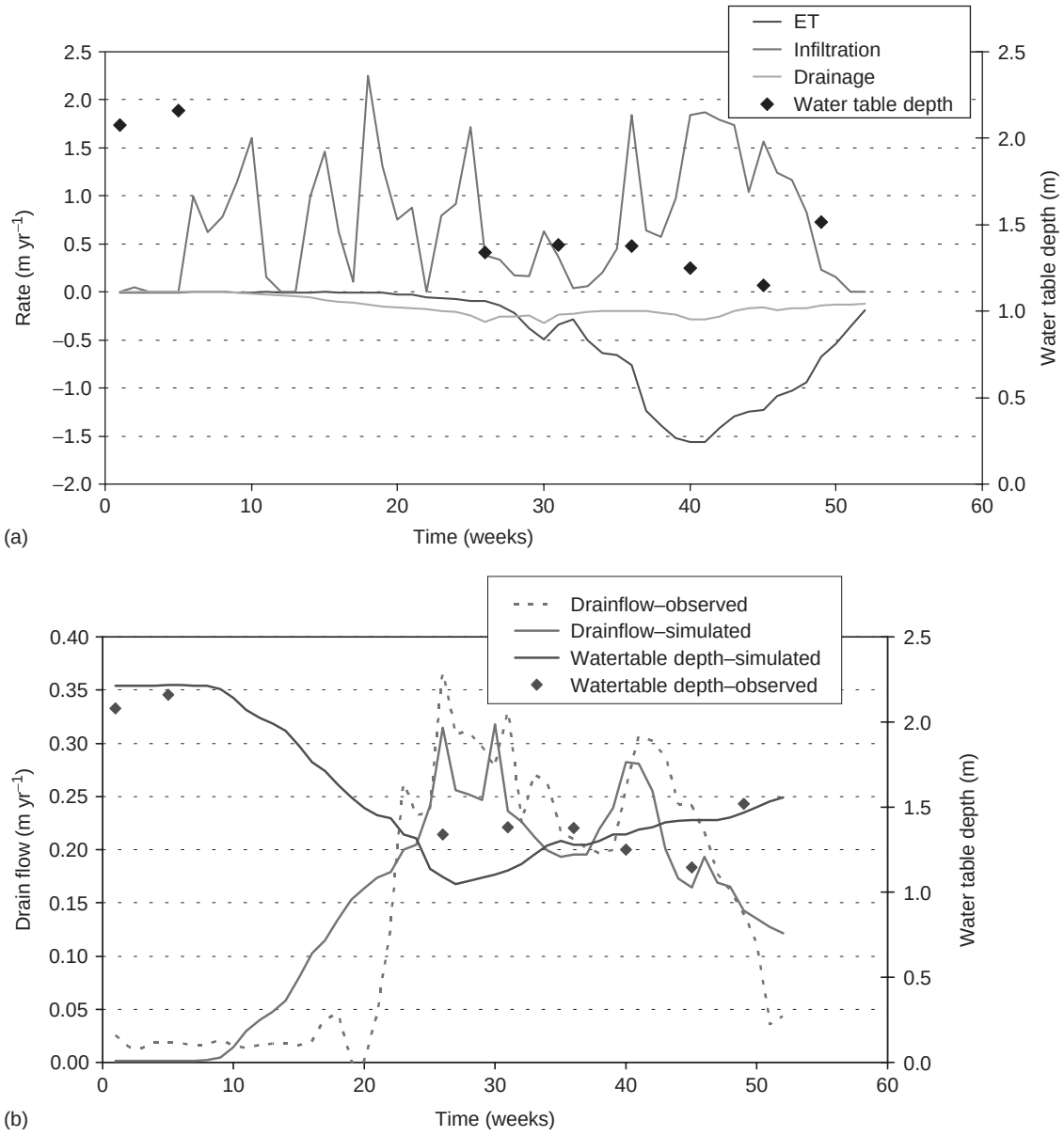


Figure 2 (a) Annual water balance for Broadview Water District, crop year 1998. (b) Comparison of observed and simulated values of district-scale drainage discharge and water table depth after optimization. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of the porous system only, but also depend on the boundary conditions (Blöschl *et al.*, 1995). Therefore, it may be argued that application of the Richards equation at this scale is conceptual rather than physically based. (Beven, 2001). It is expected that the presented calibration approach may be improved by including spatially distributed scaling factors based on the soil map. The selection of using a single scaling factor for the calibration of the entire water district was based on parameter identifiability limitations of the optimization algorithm. Optimization algorithms that can efficiently optimize larger number of parameters, such as SCE, are likely much more appropriate when calibrating distributed scaling factors.

CONCLUDING REMARKS

It is becoming increasingly obvious that there is a pressing need for unsaturated water flow modeling, monitoring and characterization at the regional scale, such as for an agricultural field or watershed. To date, soil hydrologists are quite comfortable in measuring and modeling flow processes at local scales, since physical characterization methods are applicable at the laboratory or at a small field plot scale.

However, the need for hydrological and associated environmental and ecological solutions is becoming increasingly needed at the regional scale. In response, hydrological scientists mostly apply soil hydrological concepts that are considered valid at the local scale and extend those to the regional scale. Whereas this type of approach requires an enormous amount of experimental observations, one may also question whether this is valid.

The scale problem is extremely complex because of the general presence of large spatial and temporal variability of soil hydrological properties and their highly nonlinear nature. Hence, much more developmental work is needed regarding fundamental concepts and measurement technologies to establish appropriate soil hydrological parameters for the description of unsaturated water flow at the larger spatial scales. Simultaneously, the need arises to develop appropriate scale-dependent measurement techniques that allow for model calibration and the estimation of effective soil hydrological properties. Present theory and applications of remote sensing may potentially improve the understanding of large-scale hydrological processes such as runoff, infiltration, and evapotranspiration, including their spatial distribution and scale-dependency. For example, the monitoring of transient soil moisture changes by remote sensing may provide the essential information to estimate upscaled soil hydrological parameters such as needed for the unsaturated hydraulic functions.

Although it is evident that large-scale measurements and modeling is needed, there is also increasing awareness within the scientific community that vadose zone processes are controlled by mechanisms operating at the pore size scale. Improved understanding of these processes will likely be a function of the development and application of innovative noninvasive measurement techniques that operate at the microscopic level. Examples of such techniques are NMR, CT, and laser technologies. Finally, we note that the required improved integration of vadose science in hydrologic models requires interdisciplinary partnerships with surface and groundwater hydrologists, experts in remote sensing, numerical modeling and parameter optimization, water management and policy, experimentalists, and others.

SOFTWARE LINKS

HYDRUS

Software description: Finite element model for simulating the one-dimensional movement of water, heat, and multiple solutes in variably saturated media.

Typical applications: Analysis of water flow and solute transport in soils.

URL: **http://www.ussl.ars.usda.gov/models/hydrus2d.HTM**

Reference: Šimunek J., Šejna M., and van Genuchten M.Th. (1999). *The HYDRUS-2D software package for simulating two-dimensional movement of water, heat, and multiple solutes in variable saturated media.* Version 2.0, IGWMC-TPS-53, International Ground Water Modeling Center, Colorado School of Mines, Golden, Colorado.

MOD-HMS

Software description: MODFLOW-based, fully integrated and comprehensive hydrologic flow and transport modeling system, including 3-D variably saturated subsurface flow, 2-D aerial overland flow, and 1-D channel flow

Typical applications: Integrated analysis of regional-scale flow and transport

URL: **http://www.modhms.com/modhms.cfm**

Reference: HydroGeoLogic Inc. (2001). *MOD-HMS: A Comprehensive MODFLOW-based Hydrologic Modeling System.* Herndon, VA.

ROSETTA

Software description: Neural network pedotransfer functions to estimate unsaturated hydraulic properties from surrogate soil data such as soil texture data and bulk density.

Typical applications: Parameterization of regional-scale vadose zone flow and transport models.

URL: **http://www.ussl.ars.usda.gov/models/rosetta/rosetta.HTM**

Reference: Schaap M.G., Leij F.J., and van Genuchten M.Th. (1998). Neural network analysis for hierarchical prediction of soil–water retention and saturated hydraulic conductivity. *Soil Sci. Soc. Am. J.*, **62**, 847–855.

PEST

Software description: Nonlinear parameter estimation package based on the Levenberg-Marquardt algorithm

Typical applications: Inverse estimation of soil hydraulic parameters using a hydrologic model and observed data

URL: **http://www.parameter-estimation.com/html/pest_overview.html**

Reference: PEST Software (1998). *PEST: Model-Independent Parameter Estimation*. Watermark Computing.

NEURAL MULTISTEP

Software description: Algorithm for prediction of soil–water retention and hydraulic conductivity data.

Typical applications: Soil hydraulic data are required for unsaturated water flow modeling purposes.

URL: **http://www.usyd.edu.au/su/agric/acpa/software/multistep.htm**

Reference: Minasny, B., J.W. Hopmans, T.H. Harter. A.M. Tuli. S.O. Eching and D.A. Denton. 2004. Neural network prediction of soil hydraulic functions for alluvial soils using multistep outflow data. *Soil Science Soc. Amer. J.* 68:417–429.

REFERENCES

Ahuja L.R., Rojas K.W., Hanson J.D., Shaffer M.J. and Ma L. (1999) *Root Zone Water Quality Model: Modeling Management Effects on Water Quality and Crop Production*, Water Resources Publication LLC: Highlands Ranch.

Allen R.G., Pereirea L.S., Raes D. and Smith M. (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*, FAO Irrigation and Drainage Paper 56, FAO: Rome.

Baveye P. and Boast C.W. (1999) Physical scales and spatial predictability of transport processes in the environment. In *Assessment of Non-Point Source Pollution in the Vadose Zone*, Geological Monograph Series 108, Corwin D.L., Loague K. and Ellsworth T.R. (Eds.), American Geophysical Union: Washington, pp. 261–280.

Becker A. and Braun P. (1999) Disaggregation, aggregation and spatial scaling in hydrological modeling. *Journal of Hydrology*, **217**, 239–252.

Beven K. (1995) Linking parameters across scales: subgrid parameterizations and scale dependent hydrological models. *Hydrological Processes*, **9**, 507–525.

Beven K. (2001) How far can we go in distributed hydrological modeling. *Hydrology and Earth System Sciences*, **5**, 1–12.

Blöschl G., Grayson R.B. and Sivapalan M. (1995) On the representative elementary area (REA) concept and its utility for distributed rainfall-runoff modeling. *Hydrological Processes*, **9**, 313–330.

Chen Z.Q., Govindaraju R.S., and Kavvas M.L., (1994) Spatial averaging of unsaturated flow equations under infiltration conditions over areally heterogeneous fields: 1. Development of models. *Water Resources Research*, **30**, 523–533.

Clausnitzer V. and Hopmans J.W. (1999) Estimation of phase-volume fractions from tomographic measurements in two-phase systems. *Advances in Water Resources*, **22**, 577–584.

Clausnitzer C. and Hopmans J.W. (2000) Pore-scale measurements of solute breakthrough using microfocus computed tomography. *Water Resources Research*, **36**, 2067–2079.

Cushman J.H. (1990) An introduction to hierarchical porous media. In *Dynamics of Fluids in Hierarchical Porous Media*, Cushman J.H. (Ed.), Academic Press: San Diego, pp. 1–6.

Dane J.H. and Hopmans J.W. (2002) Soil water retention and storage – introduction. In *Methods of Soil Analysis. Part 4. Physical Methods*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America Book Series: No. 5, pp. 671–674.

De Vos J.A., Raats P.A.C. and Feddes R.A. (2002) Chloride transport in a recently reclaimed Dutch polder. *Journal of Hydrology*, **257**, 59–77.

Duan Q.Y., Gupta H.V. and Sorooshian S. (1993) Shuffled complex evolution approach for effective and efficient global optimization. *Journal of Optimization Theory and Applications*, **76**, 501–521.

Duan Q.Y., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (2003) *Calibration of Watershed Models*, Water Science and Application 6, AGU: Washington.

Dunn S.M. and Lilly A. (2001) Investigating the relationship between a soils classification and the spatial parameters of a conceptual catchment-scale hydrological model. *Journal of Hydrology*, **252**, 157–173.

Eching S.O., Hopmans J.W. and Wallender W.W., (1994) Estimation of in situ unsaturated soil hydraulic functions from scaled cumulative drainage data. *Water Resources Research*, **30**, 2387–2394.

Eckhardt K. and Arnold J.G. (2001) Automatic calibration of a distributed catchment model. *Journal of Hydrology*, **251**, 103–109.

Ferguson R.B. and Hergert G.W. (1999) Sampling and spatial analysis techniques for quantifying soil map unit composition. In *Assessment of Non-Point Source Pollution in the Vadose Zone*, Geological Monograph Series 108, Corwin D.L., Loague K. and Ellsworth T.R., (Eds.), American Geophysical Union: Washington, pp. 79–94.

Hopmans J.W., Nielsen D.R. and Bristow K.L. (2002a) How useful are small-scale soil hydraulic property measurements for large-scale vadose zone modeling. In *Heat and Mass Transfer in the Natural Environment, The Philip Volume*, Geophysical Monograph Series No 129, Smiles D, Raats P.A.C. and Warrick A. (Eds.), AGU: pp. 247–258.

Hopmans J.W., Simunek J., Romano N. and Durner W. (2002b) Simultaneous determination of water transmission and retention properties. Inverse methods. In *Methods of Soil Analysis. Part 4. Physical Methods*, Soil Science Society of America Book Series No. 5, Dane J.H. and Topp G.C. (Eds.), AGU: pp. 963–1008.

Hopmans J.W. and Stricker J.N.M. (1989) Stochastic analysis of soil water regime in a watershed. *Journal of Hydrology*,**105**, 57–84.

Kavvas M.L. (1999) On the coarse-graining of hydrological processes with increasing scales. *Journal of Hydrology*, **217**, 191–202.

Klute A. and Dirksen C. (1986) Hydraulic conductivity and diffusivity: laboratory methods. *Methods of Soil Analysis. Part 1. Physical and Mineralogical Methods*, Agronomy Monograph Number 9, *Second Edition*, ASA: Madison, pp. 687–734.

Kosugi K. and Hopmans J.W. (1998) Scaling water retention curves for soils with lognormal pore size distribution. *Soil Science Society Of America Journal*, **62**, 1496–1505.

Mohanty B.P., Bowman R.S., Hendrikx J.M.H. and van Genuchten M.Th. (1997) New piece-wise continuous hydraulic functions for modeling preferential flow in an intermittent flood irrigated field. *Water Resources Research*, **33**, 2049–2063.

Madsen H. (2003) Parameter estimation in distributed hydrological catchment modeling using automatic calibration with multiple objectives. *Advances in Water Resources*, **26**, 205–216.

National Research Council (1991) *Opportunities in the Hydrologic Sciences*, National Academy Press: Washington.

Roth K., Vogel H.-J. and Kasteel R. (1999) The scaleway: a conceptual framework for upscaling soil properties. In *Modelling of Transport Processes in Soils at Various Scales in Time and Space*, Feyen J. and Wiyo K. (Eds.), International Workshop of EurAgEng's Field of Interest on Soil and Water: 24–26 November, Leuven, Wageningen Pers, Wageningen, pp. 477–490.

Schaap M.G., Leij F.J. and van Genuchten M.Th. (1998) Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity. *Soil Science Society Of America Journal*, **62**, 847–855.

Šimunek J., Šejna M. and van Genuchten M.Th. (1999) *The HYDRUS-2D Software Package for Simulating Two-Dimensional Movement of Water, Heat and Multiple Solutes in Variable Saturated Media*, Version 2.0, IGWMC-TPS-53, International Ground Water Modeling Center, Colorado School of Mines: Golden, Colorado.

Somma F., Clausnitzer V. and Hopmans J.W. (1998) Modeling of transient three-dimensional soil water and solute transport with root growth and water and nutrient uptake. *Plant and Soil*, **202**, 281–293.

Tuli A., Kosugi K. and Hopmans J.W. (2001) Simultaneous scaling of soil water retention and unsaturated hydraulic conductivity functions assuming lognormal pore size distribution. *Advances in Water Resources*, **24**, 677–688.

Van Dam J.C., Huygen J., Wesseling J.G., Feddes R.A., Kabat P., van Walsum R.E.V., Groenendijk P. and van Diepen C.A. (1997) *Theory of SWAP Version 2.0*, SC-DLO, Wageningen Agricultural University: Report 71, Department of Water Resources, The Netherlands.

Zhang D. (2002) *Stochastic Methods for Flow in Porous Media. Coping with Uncertainties*, Academic Press: San Diego.

Zhang D., Zhang R., Chen S. and Soll W.E. (2000) Pore scale study of flow in porous media: scale dependency, REV and statistical REV. *Geophysical Review Letters*, **27**, 1195–1198.

Zhu A.-X. and Scott Mackay D. (2001) Effects of spatial detail of soil information on watershed modeling. *Journal of Hydrology*, **248**, 64–77.

67: Hydrology of Swelling Clay Soils

DAVID SMILES[1] AND PETER A C RAATS[2]

[1] *CSIRO Land and Water, Canberra, Australia*
[2] *Wageningen University and Research Centre, Wageningen, The Netherlands*

Theory of water flow in nonswelling soils has been used for more than 50 years but liquid flow in porous media that change volume with liquid content is not well established in soil science, although it is considered increasingly in chemical and mining engineering and in soil mechanics. Theory of water flow in swelling systems must satisfy material continuity; it must also account for changes in the gravitational potential energy of the system during swelling and for anisotropic stresses that constrain the soil laterally but permit vertical movement. A macroscopic, phenomenological analysis based on Darcy's law provides a useful first approach to the hydrology of such soils and, if we presume that volume change, in the large, is essentially one-dimensional, material coordinates based on the vertical distribution of the solid phase result in a water flow equation analogous to the Richards equation for nonswelling soils. This framework fully accounts for the vertical strain of the solid phase, and solutions to the flow equation are available for a wide range of practically important initial and boundary conditions. The approach has been well tested in clay suspensions and in saturated systems such as mine tailings and sediments. It is also applied in soil mechanics and, here, we apply it to swelling soils. As with the use of the Richards equation in rigid soils, we recognize that complications arise, but the approach remains the most coherent and profitable to support current needs and serve as a point of departure for future research. The use of material coordinates is simple. We discuss some experimental difficulties in using the approach and also consider extension of the approach to more than one dimension.

INTRODUCTION

Pedologists estimate that there are more than 2.60×10^8 ha of swelling clay soils worldwide. They are among the most fertile of soils and are found mostly in developing countries. At the same time, their hydrology is not well understood, and current approaches are generally based on nonswelling soil theory and rarely account for particle motion and volume change. As a result, significant errors in estimates of local water and solute flux can arise because of the neglect of advection of water with the moving solid. Errors can also occur in water balance calculations because swelling can make volume-based estimates of water content ambiguous.

Four major symposia (Kutilek and Sutor, 1976; McGarity *et al.*, 1984; Bouma and Raats, 1984; Kutilek, 1996) and two reviews (Bronswijk, 1991; Fredlund and Rahardjo, 1993) relating to the hydrology of such soils in the last 40 years present a range of views and much experimental information, but no agreed approach to water movement in these systems emerges. Bronswijk (1991) provides much experimental insight into structural behavior and its consequences in these soils as well as an extensive list of references. A conference focused on the mechanical properties of unsaturated soils (Alonso and Delage, 1995) also provides data and recent theory relating to "engineering" soils, as do Fredlund and Rahardjo (1993).

Two distinct approaches are evident. The first deals with the behavior of individual soil structural elements; it envisages rapid water movement through macropores to these elements, which then swell three-dimensionally to produce one-dimensional, vertical profile swelling. This approach concentrates on the relation between three-dimensional volume change of aggregates or clay pastes and one-dimensional soil profile behavior. The second approach is macroscopic; it argues that, in the large, the detail of the volume change of aggregates is irrelevant, provided that the resulting vertical deformation is well described. It presumes that Darcy's law is valid for these soils and it yields

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

an equation analogous to that of Richards (1931), with focus on one-dimensional flow (Philip, 1968a; Kim *et al.*, 1992a,b). The second approach has been used effectively in civil and chemical engineering, but it is not widely used in soil science, and some of its basic concepts remain misunderstood.

Both approaches have been used, but we focus on the second because we think it provides a more generalizable way to develop theory than does the former one and a more coherent basis for future study. It also results in flow equations whose solutions are well understood. We identify some origins of the approach, discuss basic theory and its test and, following that, discuss the extension of these ideas to multidimensional situations. We focus on one-dimensional flow but pay some attention to application of the approach in two and three dimensions.

HISTORICAL BACKGROUND

Nonswelling Soils

One-dimensional flow of water in nonswelling soils is conventionally described by the Richards (1931) equation, which combines the material balance equation for water,

$$\frac{\partial \theta_w}{\partial t} = -\frac{\partial u}{\partial z} \tag{1}$$

with Darcy's law,

$$u = -k(\theta_w)\frac{\partial \Phi}{\partial z} \tag{2}$$

In these equations, t is time, z is the vertical, positive upward coordinate, θ_w is the water volume fraction, u is volume flux of water, $k(\theta_w)$ is the soil hydraulic conductivity, and Φ is the total potential of the water. In a nonswelling soil,

$$\Phi = z + \psi(\theta_w) = z + p_w \tag{3}$$

If Φ and its components are defined as work per unit weight of water, then they have dimensions L and units m H_2O. The hydraulic conductivity, $k(\theta_w)$, then has dimensions LT^{-1} and units m s^{-1}. Defined this way, Φ is identical to the engineer's hydraulic head with z the gravitational component equal to the elevation of the point of concern in the soil relative to a convenient reference height, and $\psi(\theta_w)$ the capillary component that reflects the interaction of the water with the soil solid surfaces and their geometry. The capillary potential, $\psi(\theta_w)$, equals the manometric pressure, p_w, of the water in unsaturated soil and is negative. Buckingham (1907) first enunciated these concepts and also recognized the water content dependence of $\psi(\theta_w)$ and $k(\theta_w)$.

The implementation of the Richards equation requires that $\psi(\theta_w)$ and $k(\theta_w)$ be known. These macroscopic soil properties represent averages over volumes including many hundreds of pores and pore sequences. They are basically unpredictable although, generally, they can be measured without ambiguity at the scale at which the approach is to be applied. When ψ (θ_w) is single valued, the concept of the soil–water diffusivity, $D(\theta_w) = k(\theta_w)$ $(d\psi/d\theta_w)$, can be invoked (Childs and George, 1948).

This approach underpins current water flow theory (Philip, 1969a, 1980; Jury *et al.*, 1991; Marshall *et al.*, 1996). Where it fails, physically based or empirical modeling follows, but these models still retain the basic elements of continuity and flux of the approach leading to the Richards equation. Its pedagogic and practical utility guides our approach to water flow in swelling soils.

Swelling Soils and Other Porous Media

The theory of one-dimensional flow in swelling soils was conceived by Terzaghi (1923) to describe soil consolidation beneath engineering structures. Terzaghi's insight was that the water initially carries an imposed load in saturated soil and the soil compresses as the water redistributes and the load transfers to the soil matrix. He thus associated consolidation with water flow and his approach, like that of Richards, is based on material continuity and Darcy's law. He also related the soil volume to an effective or interparticle stress, and he conceived a spacelike material coordinate defined in terms of the distribution of the solid. It seems that he assumed the deformation to be so small that it produced no significant geometrical change of the region occupied by the solid phase of the soil, and his material coordinates then reduced to a system attached to the solid phase but moving rigidly in space (see Raats and Klute, 1968a). Terzaghi's analysis envisaged imposed loads that greatly exceeded effects of gravity and the displacement of overburden that contributes to the total potential of the water when soil volume changes. He was aware of the water content dependence of the hydraulic conductivity, but he ignored that also, and he solved the resulting linear analogue of the Richards equation using Fourier series.

Civil engineers were uneasy with Terzaghi's material coordinates and his approach was recast in physical space and specifically identified with small-strain behavior. As a result, his classic text (Terzaghi, 1956) disregards many insights of the 1923 work as does Jaeger's text of 1962, where the small-strain linearized approach is maintained and the effect of gravity and overburden overlooked. Biot (1941, 1955) gave a full three-dimensional treatment of the deformation of the solid phase, but, like Terzaghi, in his early work, he restricted his attention to small deformations. The rudimentary material coordinate of Terzaghi and Biot were not rediscovered until the 1960s (McNabb, 1960),

and Gibson *et al.* (1967) extended the approach to include overburden. There was also substantial experimental study of the water potential in loaded saturated and unsaturated swelling soils by Croney and Coleman in the British Road Research Laboratories (Pore Pressure and Suction in Soil, 1961).

Chemical engineers in the USA (e.g. Tiller and Cooper, 1960; Tiller and Shirato, 1964; Leu, 1986) and in Japan (e.g. Atsumi *et al.*, 1973; Shirato *et al.*, 1970, 1986) also developed filtration and sedimentation theory focused on water flow relative to the solid, although explicit combination of continuity and flux to formulate a flow equation such as that of Terzaghi (1923) or Richards (1931) and the importance of gravity and overburden still await general acceptance.

In soil science, Tempany (1917) identified the earliest measurement of volume change in the work of Schübler in the early nineteenth century. Tempany considered shrinkage of pastes as a basis for describing profile behavior, and Haines (1923) developed the shrinkage curve concept in which the volume change accompanying water content change is used to help characterize structural properties of agricultural soils. Woodruff (1936) used anchored rods to measure profile swelling with water content change and Aitchison and Holmes (1953a,b), Yaalon and Kalmar (1973), Bronswijk (1991), Cabidoche and Ozier-Lafontaine (1995), and Coquet (1998) developed that method. Childs (1936) and Nicholson and Childs (1936) applied a diffusion equation to water flow in heavy clay soils.

Spacelike material coordinates to deal with swelling soils were formulated by Raats and Klute (1968a,b, 1969) and by Smiles and Rosenthal (1968), who also tested the approach using bentonite pastes. Contemporaneously, Philip (1968a) analyzed one-dimensional soil volume change accompanying water content change in physical space. Philip's paper deals well with mechanics of swelling but later he abandoned the approach because the mathematics associated with the material approach was analogous to established one-dimensional nonswelling water flow theory, and his methods of solution of the Richards equation (Philip, 1969a) carried over immediately to one-dimensional swelling materials.

We examine the central elements in one-dimensional flow of water in swelling soils and compare them with the more familiar theory of flow in nonswelling soils, we identify limits to the analysis and some issues yet to be dealt with. The focus is one-dimensional because field volume change, in the large, is constrained to the vertical. The approach accepts three-dimensional volume change of soil structural units, but it asserts that if the area of cross section is large enough and if there is no lateral net transfer of material from the control area, then behavior of structural entities beneath it may be described by a representative average vertical displacement. This approach is analogous to the Darcy scale approach to soil water movement in nonswelling soil, which also deals in volume averages and is agnostic about flow detail in individual pores.

In this one-dimensional domain, analysis is based on the material balance equations for the water as well as the solid, because both may be in motion relative to an observer, and on Darcy's law. Both these elements are critical but material balance must be retained even when the applicability of the Darcy equation is uncertain. As with nonswelling soil water flow theory, material properties and the variables required by the approach are defined in terms of the scale of discourse, and demand unambiguous measurement at that scale (Philip, 1973).

THEORY OF FLOW IN SWELLING SYSTEMS

Material Balance

Equations (4) and (5) define material balance for both the water and the soil solid during one-dimensional flow in a swelling system. Equation (6) recognizes that the flux of water occurs both relative to and with the moving soil solid.

$$\frac{\partial \theta_w}{\partial t} = -\frac{\partial F_w}{\partial z} \tag{4}$$

$$\frac{\partial \theta_s}{\partial t} = -\frac{\partial F_s}{\partial z} \tag{5}$$

$$F_w = u + \frac{\theta_w F_s}{\theta_s} = u + \vartheta F_s \tag{6}$$

In these equations, F_w and F_s are the volume flux densities of the water and of the solid relative to the z-coordinate, the volume fractions θ_w and θ_s of the water and the solid are defined per unit area of horizontal cross section of the soil, and the reference volume includes the cracks; $\theta_w + \theta_s \leq 1$, depending on whether or not the soil is water saturated. It is important to note that, u, the Darcy flux of equation (2), is the flux of water relative to the particles in response to the space gradient of the total potential of the water (Gerzevanov, 1937; Biot, 1955; Zaslavsky, 1964; Raats and Klute, 1968a, 1968b). The second term on the right in equation (6) thus describes transfer of water "advected" with the moving solid. The water ratio, ϑ, is defined as:

$$\vartheta = \frac{\theta_w}{\theta_s} = \theta_g \rho_s \tag{7}$$

with θ_g the gravimetric water content (kg.kg^{-1}) and ρ_s the specific gravity of the soil solid. In saturated soils, $\vartheta = e$, the void ratio.

Equations (4), (5) and (6) give (Smiles, 1986)

$$\frac{\partial \vartheta}{\partial t} = -\frac{\partial u}{\partial m} \tag{8}$$

Equation (8) is a material balance equation for the water, based on a "material" coordinate, $m(z, t)$. The spatial time derivatives in equations (4) and (5) and the material time derivative are related by:

$$\left.\frac{\partial}{\partial t}\right|_{\mathrm{m}} = \left.\frac{\partial}{\partial t}\right|_{\mathrm{z}} + \left(\frac{F_{\mathrm{s}}}{\theta_{\mathrm{s}}}\right)\left.\frac{\partial}{\partial z}\right|_{t} \tag{9}$$

The m-coordinate is determined by the distribution of the soil solid, and it accounts for one-dimensional soil displacement that might accompany water content change.

$$\mathrm{d}m(z,t) = \frac{\partial m}{\partial z}\,\mathrm{d}z + \frac{\partial m}{\partial t}\,\mathrm{d}t = \theta_{\mathrm{s}}\,\mathrm{d}z - F_{\mathrm{s}}\,\mathrm{d}t \tag{10}$$

This definition of m satisfies equation (5) for the solid. Integration of equation (10) is simplified if it is based on a surface, $z = 0$, where $F_{\mathrm{s}} = 0$. The soil surface provides such a datum. Then,

$$m = \int_0^z \theta_{\mathrm{s}}\,\mathrm{d}z - \int_0^{\mathrm{t}} F_{\mathrm{s}_{z=0}}\,\mathrm{d}t = \int_0^z \theta_{\mathrm{s}}\,\mathrm{d}z = \int_0^z \left(\frac{\rho}{\rho_{\mathrm{s}}}\right)\mathrm{d}z \tag{11}$$

so m is the cumulative volume of solid, per unit area of cross section, measured from $z = 0$ to the location z and, thus, the cumulative oven dry mass per unit area is measured away from the soil surface to any depth z, and divided by ρ_{s} (Smiles and Rosenthal, 1968). The final integral in equation (11) shows how m is related to the bulk specific gravity ρ.

This formulation is unaffected by crack volume and is indifferent to three-dimensional volume change of aggregates and cracks provided that the reference area of cross section is large enough to ensure that no horizontal net transfer of material occurs. This is implicit in the Raats and Klute (1969) analysis and it is explicitly stated by Philip and Smiles (1969) and by Philip (1969b,c,d). Water content profiles in this formalism are expressed as $\vartheta(m)$ so the integral, $\int_0^m \vartheta\,\mathrm{d}m$, yields the total volume of water in a profile of material length m regardless of the way its physical length might change with water content change (Smiles, 1997).

The m-coordinate, and equation (8), apply to both swelling and nonswelling porous materials, although their use in the latter case is unusual (but see Smiles and Kirby, 1994).

A three-dimensional generalization of equation (8) was derived by Raats and Klute (1968a). It can be specialized to the one-dimensional formulation described above (Raats and Klute, 1969; Raats, 1987) and can also be used at the meso-scale to describe the deformation of individual aggregates (Raats, 1969, 1984, 2002).

Equilibrium and Flow of Water in Saturated Swelling Systems

It is important to reiterate that Darcy's law in swelling systems describes flow of water *relative to* the solid in response to a space gradient of the water potential with the hydraulic conductivity, $k(\theta_{\mathrm{w}})$, a function of θ_{w}. It is also important to discriminate between measures of potential in nonswelling and swelling soils and then between saturated and unsaturated ones.

The Total Potential of Water in Saturated Swelling Systems

We recall that in a nonswelling soil, Φ has a capillary component that can be measured with a manometer, so $\psi(\theta_{\mathrm{w}}) = p_{\mathrm{w}}$. In a swelling system, however, change of water content displaces the whole profile in the gravity field, so we require an extra component of Φ, the overburden potential, Ω. Philip (1969b, 1970) considered this effect from a mechanistic point of view while Croney and Coleman (1961), Groenevelt and Bolt (1972) and Sposito (1975a) considered it from a thermodynamic perspective. Philip's approach builds on that of Buckingham (1907) for nonswelling, unsaturated soil. He argues that Φ represents the work involved in transferring unit weight of water from a reference flat surface of water at atmospheric pressure to a swelling material at some height, z, above that datum. Thus Φ includes work involved in:

1. elevation of the element of water from the datum to the height z (the gravitational potential, z);
2. interaction between the water and the soil solid surfaces and their geometry (the water content–dependent unloaded matric or capillary potential, ψ); and
3. vertical displacement of the wet soil accompanying unit change in water content at z (the overburden potential, Ω).

In a one-dimensional, saturated system that is laterally constrained but free to move vertically, unit change in water content produces unit change in elevation of the soil. This implies that $\mathrm{d}e/\mathrm{d}\vartheta = 1$ if the volume of the system is parameterized by the void ratio e, and Ω equals the engineer's total normal stress, σ, on a horizontal plane at z. Thus,

$$\Phi = z + \psi + \int_{\mathrm{z}}^{T} \gamma\,\mathrm{d}z = z + \psi + \sigma = z + p_{\mathrm{w}} \tag{12}$$

in which ψ is the *unloaded* capillary potential, that is the potential a tensiometer would measure at the existing moisture ratio in the absence of overburden, γ is the wet specific gravity, and $\sigma = \Omega$. We distinguish between σ and Ω to anticipate the situation in unsaturated soils where $\sigma \neq \Omega$. Civil engineers relate p_{w} to σ, and the effective

or interparticle stress, σ', according to the equation (Pore Pressure and Suction in Soil, 1961)

$$\sigma = \sigma' + p_w \tag{13}$$

Equation (13) was formulated and tested by Terzaghi (1923). Comparison with equations (12) and (13) shows that $\sigma' = -\psi$, and, hence, is related to ϑ through $\psi(\vartheta)$. Additional surface loads will increase the overburden but, if they are constant, we may ignore them here without loss of generality.

Figure 1 illustrates and compares components of the total potential of water in swelling and nonswelling systems.

Equation (12) applies to both swelling and nonswelling soils. Thus, water moves through a swelling soil in response to a measurable hydraulic head gradient just as it does in a nonswelling one. The literature is confused on this point. Jury *et al.* (1991), for example, misinterpret it in their Fig. 2.9b and the Glossary of Soil Science (SSSA, 1997) ignores the energetics of soil volume change altogether.

The Potential Gradient in Saturated Swelling Systems

Equations (12) and (13) have been well tested, and equation (12) may be rearranged and differentiated to obtain

$$\frac{\partial \Phi}{\partial z} = \frac{\partial \psi}{\partial z} + (1 - \gamma) = \theta_s \left(\frac{\partial \psi}{\partial m} + (1 - \rho_s) \right) \tag{14}$$

The second equality in equation (14) arises from the first using the definitions of γ, which in the saturated system equals $(\vartheta + \rho_s)/(\vartheta + 1)$, and of m defined by equation (10). The effect of gravity, including the overburden, is represented by the $(1 - \gamma)$ term in z-space, and the $(1 - \rho_s)$ in m-space. In most mineral soils, $\rho_s \approx 2.65$, so these terms are of different sign to that for nonswelling soils, and this gives rise to what Philip (1991) called the *bouleversement*

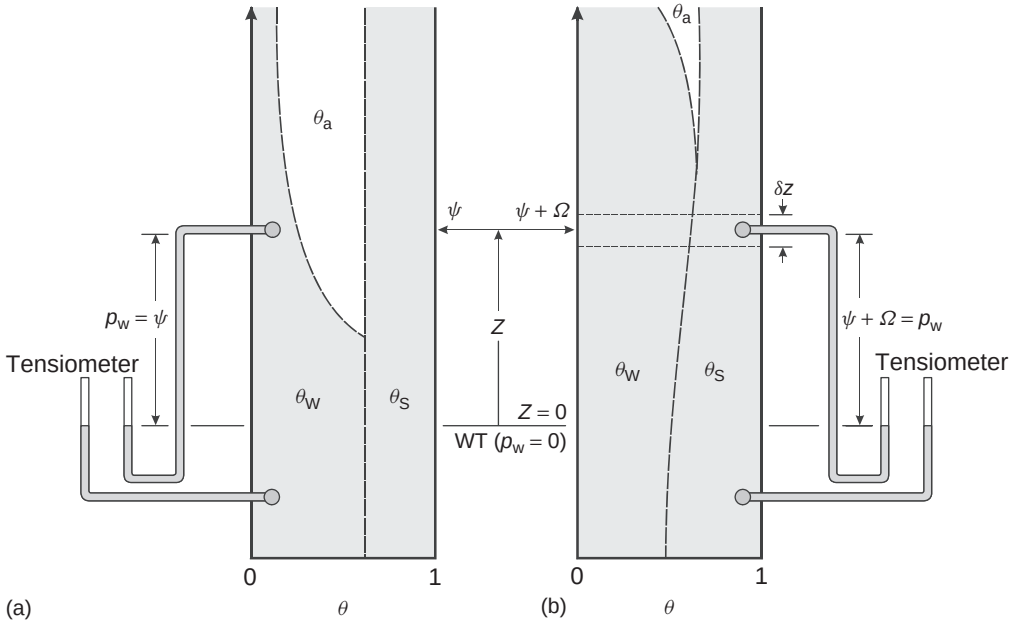


Figure 1 Idealized diagram showing vertical, static equilibrium water content profiles in (a) nonswelling soil and, (b) swelling soil in contact with a free-water surface, WT, at elevation $z = 0$ and at atmospheric pressure and where the manometric pressure $p_w = 0$. The volume fractions of the water, θ_w, of the solid, θ_s, and of the air; θ_a, sum to unity. In swelling soils, θ_s increases with depth towards some maximum density, and θ_w correspondingly decreases. Entry of air above the water table in swelling soil occurs in cracks and then as aggregates desaturate. The diagram also illustrates differences in the components of total potential of the water, Φ, in the two systems. The arrows show the steps in transferring unit amount of water from WT, to the soil at elevation z above it. If the potential is defined per unit weight of water, then z is the gravitational potential of the water relative to WT; ψ is the unloaded ''capillary'' or moisture potential of the water in the soil, and the overburden potential, Ω, arises in swelling soils because of the vertical displacement of the wet profile, δz, that must occur when the element of water is inserted at height z. In the saturated region, unit vertical displacement of the profile above z accompanies insertion of unit amount of water there so Ω is the wet weight of the profile, per unit area above z. In the unsaturated region of the swelling soil, the vertical displacement is less than unity and Ω is reduced by the factor α ($\leq$1), which represents an average value of the shrinking curve in the soil above z. Further detail is presented in the text. The diagram shows that a manometer/tensiometer inserted at elevation z measures $(\psi + \Omega)$ and, at static equilibrium, the manometer/tensiometer water level has the same elevation as the water table, WT, for both swelling and nonswelling soils

in hydrologic behavior of swelling soils. This difference between z- and m-space formulations also gives rise to differences in the flow equations below. It suffices here to note that in physical z-space, if γ is close to unity, the effect of gravity is close to zero. The m-coordinate, because it is tied to the solid distribution, which is also moving in the gravitational field, tends to hide this phenomenon.

Darcy's Law in Saturated Swelling Systems

Substitution of equation (14) in equation (2) gives

$$u = -k(\theta_w)\theta_s\left(\frac{\partial\psi}{\partial m} + (1-\rho_s)\right) = -k_m(\vartheta)\left(\frac{\partial\psi}{\partial m} + (1-\rho_s)\right) \quad (15)$$

and we refer to $k_m(\vartheta)(= k(\theta_w)\theta_s)$ as the "material" hydraulic conductivity.

The Flow Equation in Saturated Swelling Systems

Substitution of equation (15) in equation (8) then yields

$$\frac{\partial\vartheta}{\partial t} = \frac{\partial}{\partial m}\left(k_m(\vartheta)\frac{\partial\psi}{\partial m}\right) + \frac{\partial}{\partial m}(k_m(\vartheta)(1-\rho_s)) \quad (16)$$

Equation (16) is the analogue, in material coordinates, of the Richards equation. It differs only in the definition of k_m and in the modification of the gravity term resulting from the effect of overburden. Its use again requires that $k_m(\vartheta)$ and $\psi(\vartheta)$ be well-defined characteristics of the material. It also requires that $de/d\vartheta = 1$ and ρ_s be known.

Generally, $\psi(\vartheta)$ is hysteretic but, when it is single valued, equation (16) may be written

$$\frac{\partial\vartheta}{\partial t} = \frac{\partial}{\partial m}\left(D_m(\vartheta)\frac{\partial\vartheta}{\partial m}\right) + \frac{\partial}{\partial m}(k_m(\vartheta)(1-\rho_s)) \quad (17)$$

with $D_m(\vartheta)$, the "material" diffusivity or the coefficient of consolidation in soil mechanics (Raats and Klute, 1969; Smiles, 1986) defined by

$$D_m(\vartheta) = k_m(\vartheta)\left(\frac{d\psi}{d\vartheta}\right) \quad (18)$$

The principal differences, which distinguish equilibrium and flow in saturated, swelling systems from equilibrium and flow in nonswelling systems, arise from the overburden effect. Specifically, it

1. reverses (and reduces) the physical effect of gravity in systems where $\rho_s > 1$. Thus, the gravitational potential energy of the wet profile actually increases as water infiltrates (cf equation (14))
2. produces static equilibrium water content profiles that are, in the saturated region, dryer at the bottom than at the top, in contrast to profiles in nonswelling soils (Philip, 1969c; Sposito, 1975a,b) and
3. extends the period of time for which a "gravity free" analysis applies to infiltration in these systems (Smiles, 1974), although the effect is as much due to the relative magnitudes of the average value of $k_m(\vartheta)$ and the square of the sorptivity as it is to the diminution in the effect of gravity (Smiles, 1974).

Equilibrium and Flow of Water in Unsaturated Swelling Systems

The Total Potential of Water in Unsaturated Swelling Systems

In unsaturated swelling systems, unit change in ϑ accompanying addition of water no longer results in unit change in elevation of the wet profile, that is, $de/d\vartheta < 1$. Unit change in the total normal stress, σ, does not then produce unit change in p_w, and a factor $\alpha \leq 1$ is introduced to modify the overburden terms in equations (12) and (13) (cf Croney and Coleman, 1961; Skempton, 1961) so they are written

$$\Phi = z + \psi + \alpha\int_z^T \gamma\, dz = z + \psi + \alpha\sigma = z + p_w \quad (19)$$

and

$$\alpha\sigma = \sigma' + p_w \quad (20)$$

In equation (19), α is the slope of the shrinkage curve (Haines, 1923), and in equation (20) it is the change in p_w with change in σ at constant σ', and, by implication, constant ϑ because $\psi = -\sigma$. In either case, the values of α lie between zero in a nonswelling soil and unity in a saturated swelling one. The "dual" nature of α revealed in equations (19) and (20) is consistent with observations of Croney and Coleman (1961) who suggested that

$$\alpha = \left(\frac{\partial p_w}{\partial\sigma}\right)_\vartheta = \left(\frac{\partial e}{\partial\vartheta}\right)_\sigma \quad (21)$$

Coleman and Croney (1952) and Groenevelt and Bolt (1972) derived the equality between the differentials in equation (21) using thermodynamic arguments (see also Raats, 2002), and Towner (1976) did so using mechanistic ones. However, these differentials are not strictly equal to α, and both Youngs and Towner (1970) and Groenevelt and Bolt (1972) showed that equating of α with the slope of the soil shrinkage curve, as was done by Philip (1969d) following Croney and Coleman, is incorrect. It is now agreed (Philip, 1970) that, while ψ continues to be interpreted as the "unloaded" suction, α in equations (19) and (20) must be an average value of $(\partial e/\partial\vartheta)_\sigma$ or $(\partial p_w/\partial\sigma)_\vartheta$ over the range 0 to P. This perception of α arises from the argument that, while the $\{p_w, \sigma, \vartheta, e\}$ set completely specifies

the state of a swelling system, only three of these variables are independent (Groenevelt and Bolt, 1972; Philip, 1970), so $p_w = p_w(\vartheta, \sigma)$ whence

$$\mathrm{d}p_w = \left(\frac{\partial p_w}{\partial \vartheta}\right)_\sigma \mathrm{d}\vartheta + \left(\frac{\partial p_w}{\partial \sigma}\right)_\vartheta \mathrm{d}\sigma \tag{22}$$

A mathematically practical integration path across the $p_w(\vartheta, \sigma)$ surface up to $\sigma = P$ is

$$\begin{aligned} p_w &= \int_0^\vartheta \left(\frac{\partial p_w}{\partial \vartheta}\right)_{\sigma=0} \mathrm{d}\vartheta + \int_0^P \left(\frac{\partial p_w}{\partial \sigma}\right)_\vartheta \mathrm{d}\sigma \\ &= \psi + \int_0^P \left(\frac{\partial p_w}{\partial \sigma}\right)_\vartheta \mathrm{d}\sigma \end{aligned} \tag{23}$$

Equation (23) can be put in the form (cf equation (19))

$$p_w = \psi + \alpha P \tag{24}$$

with α defined by

$$\begin{aligned} \alpha(\vartheta, P(z)) &= \left(\frac{1}{P(z)}\right) \int_0^P \left(\frac{\partial p_w}{\partial \sigma}\right)_\vartheta \mathrm{d}\sigma \\ &= \left(\frac{1}{P(z)}\right) \int_0^P \left(\frac{\partial e}{\partial \vartheta}\right)_P \mathrm{d}\sigma \end{aligned} \tag{25}$$

where the equality of the differentials in equation (21) has been used. The various definitions of α as a local value and as an average over a defined pressure range have led to ambiguity (see Raats, 2002). Furthermore, substitution of equation (25) in Darcy's law proves impractical, so equation (21) is generally used on the physical grounds (Philip, 1970) that while α is an average of $(\partial e/\partial \vartheta)_P$ over the full range of σ from 0 to P, its variation will be slow. It has, however, been treated as water content–dependent to deal with curvilinearity in the shrinkage curve. We return to this issue later.

The Potential Gradient in Unsaturated Swelling Systems

In this approximation, and if $\vartheta(\psi)$ is single valued, then $\partial\Phi/\partial z$ (from equation (19)) becomes

$$\begin{aligned} \frac{\partial \Phi}{\partial z} &= \frac{\partial}{\partial z}\left(\psi + z + \alpha \int_z^T \gamma \, \mathrm{d}z\right) \\ &= \left(\frac{\mathrm{d}\psi}{\mathrm{d}z}\right) + \left(1 - \gamma\alpha + \left(\frac{\mathrm{d}\alpha}{\mathrm{d}\vartheta}\right) \int_z^T \gamma \, \mathrm{d}z\right) \\ &= \theta_s \left(\frac{\mathrm{d}\psi}{\mathrm{d}\vartheta}\right)\left(\frac{\partial \vartheta}{\partial m}\right) + \left(1 - \gamma\alpha + \frac{1}{\theta_s}\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\vartheta}\right) \int_m^0 \gamma \, \mathrm{d}m\right) \end{aligned} \tag{26}$$

In these equations, the second equality on the right-hand-side is derived by differentiating the first by parts and the third expression uses $\vartheta(\psi)$, and the definition of m, that is, $\partial m/\partial z = \theta_s$ to eliminate z. The leading terms in the second and third equalities are "capillary" terms; the remaining terms represent the effects of gravity and overburden.

Darcy's Law in Unsaturated Swelling Systems

Substitution of the last expression of equations (26) in Darcy's law, in m-space, yields

$$\begin{aligned} u = &-k(\theta_w)\theta_s \left(\frac{\mathrm{d}\psi}{\mathrm{d}\vartheta}\right)\left(\frac{\partial \vartheta}{\partial m}\right) \\ &- k(\theta_w)\left(1 - \gamma\alpha + \frac{1}{\theta_s}\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\vartheta}\right) \int_m^0 \gamma \, \mathrm{d}m\right) \end{aligned} \tag{27}$$

The Flow Equation in Unsaturated Swelling Systems

Substitution of equation (27) in equation (5) then yields the Richards-like equation

$$\frac{\partial \vartheta}{\partial t} = \frac{\partial}{\partial m}\left(D_m(\vartheta)\frac{\partial \vartheta}{\partial m}\right) + \frac{\partial}{\partial m}(k^*(\vartheta)) \tag{28}$$

with $D_m(\vartheta)$ defined by equation (18) and the conductivity-like term $k^*(\vartheta)$ encapsulating the gravity and overburden terms. In many practical cases, α is effectively constant over ranges of water content of concern in the soil. This eliminates the integral term in equations (26) and (27), and $k^*(\vartheta)$ becomes

$$k^*(\vartheta) = \frac{k_m(\vartheta)}{\theta_s}(1 - \gamma\alpha) \tag{29}$$

Further detail of the derivation of equation (28) is illustrated by Philip (1969d) (his equation (32)) and by Kim *et al.* (1992a). Applications of equations (26), (27) and (28) are discussed below. For a saturated system where $\alpha = 1$, equation (28) becomes identical with equation (17).

Extension to Two and Three Dimensions

Three-dimensional Volume Change and Material Properties

One-dimensional volume change accompanying water content change in soil profiles has been measured by Aitchison and Holmes (1953a,b), Bronswijk (1991), Cabidoche and Voltz (1995), Coquet *et al.* (1998), Croney *et al.* (1958), McIntyre *et al.* (1982), Tempany (1917), and Woodruff (1936). These authors focused on the magnitude and significance of profile shrinkage, that is, the value of $(\mathrm{d}e/\mathrm{d}\vartheta)$ averaged over the profile. This profile property has also been related variously to three-dimensional aggregate volume change or to linear shrinkage by some of these authors,

and is supplemented by work such as that of Fox (1964), Berndt and Coughlin (1976), Yule and Ritchie (1980a,b) and Giraldez *et al.* (1983). These studies indicate that three-dimensional volume change of an aggregate is a useful qualitative guide to one-dimensional profile behavior. The coefficient of linear extensibility (COLE), ultimately based on the method of Tempany (1917) also appears useful and the following analysis of multidimensional deformation and one-dimensional flow is relevant.

Multidimensional Deformation and One-dimensional Flow

Drying of a non-rigid soil may start as a combination of purely one-dimensional flow of the aqueous phase and deformation of the solid phase. This changes when cracks appear. Some attempts have been made to cope with such systems by describing them in terms of one-dimensional flow of the aqueous phase and transversely isotropic deformation of the solid phase, that is, a combination of a deformation in the vertical z-direction and an isotropic deformation in the horizontal direction. For one-dimensional deformation (cf equation (10))

$$\frac{\partial m}{\partial z} = \theta_s = \frac{\rho}{\rho_s} \tag{30}$$

and for any transversely isotropic deformation of the soil solid phase, at some point (x, y, z), the deformation in the vertical z-direction is characterized by:

$$\frac{\partial m_z}{\partial z} = \theta_s^n = \left(\frac{\rho}{\rho_s}\right)^n \tag{31}$$

while deformations for any two perpendicular horizontal directions x and y are characterized by:

$$\frac{\partial m_x}{\partial x} = \theta_s^{(1-n)/2} = \left(\frac{\rho}{\rho_s}\right)^{(1-n)/2}$$

$$\text{and} \qquad \frac{\partial m_y}{\partial y} = \theta_s^{(1-n)/2} = \left(\frac{\rho}{\rho_s}\right)^{(1-n)/2} \tag{32}$$

In equation (32), $n = 1$ for purely axial deformation, $n = 1/2$ for balanced axial and lateral deformation, $n = 1/3$ for isotropic deformation, and $n = 0$ for purely lateral deformation.

Furthermore, from equations (31) and (32), it follows that

$$\left(\frac{\partial m_x}{\partial x}\right)\left(\frac{\partial m_y}{\partial y}\right)\left(\frac{\partial m_z}{\partial z}\right) = \theta_s = \frac{\rho}{\rho_s} \tag{33}$$

Disregarding the influence of gravity, so that the gradient of the capillary potential ψ is the only driving force, the flow equation for the water in the vertical z-direction is a nonlinear diffusion equation for the water ratio:

$$\left(\frac{\partial \vartheta}{\partial t}\right)_{m_z} = \frac{\partial}{\partial m_z}\left(\theta_s^{(2n-1)} k(\vartheta) \frac{d\psi}{d\vartheta}\right)\left(\frac{\partial \vartheta}{\partial m_z}\right)_t \tag{34}$$

For purely axial deformation ($n = 1$), equations (31) and (33) reduce to equation (30), equations (32) reduces to $\partial m_x/\partial x = 1$ and $\partial m_y/\partial y = 1$, and the diffusivity in equation (34) reduces to that appearing in equation (28).

The value $n = 1/2$ separates two opposite trends. For $n < 1/2$, swelling causes the factor θ_s^{2n-1} in the diffusivity to increase and shrinking causes it to decrease. For $n > 1/2$, swelling causes the factor θ_s^{2n-1} in the diffusivity to decrease and shrinking causes it to increase.

Raats (1969, 1984, 2002) and also Miller (1975) introduced the expression for the deformation gradient for the class of transversely isotropic deformations in terms of the mass density in an analysis of axial fluid flow in swelling and shrinking rods. The r_s-factor, introduced by Rijniersce (1983, 1984) in a study of physical changes in the sediments of the IJsselmeerpolders, is the reciprocal of the parameter n used above. This r_s-factor has been used extensively to characterize swelling and shrinkage in clay soils (see e.g. Bronswijk, 1991; Bronswijk and Evers-Vermeer, 1990; Garnier *et al.*, 1997a,b; Kim *et al.*, 1999).

DISCUSSION

An approach to water flow in swelling soils that gives rise to theory similar to that of Richards for nonswelling soils has both pedagogic and mathematical benefits (Raats, 2002) and provides access to well-tried technology and to plausible modeling procedures. Measurement problems and difficulties with scale and variability complicate soil measurement and modeling, however.

Scale and Measurement

Fortuitously, space and timescales of laboratory measurement, theory development, and measurement for field application are roughly similar in nonswelling soils. Furthermore, while nonswelling soil structure varies, its range of variability tends to be sufficiently modest that the soil can generally be treated as homogeneous, or at least as systematically heterogeneous at scales of order 0.1 m, and modeling tends to be based on the Richards equation. In swelling soil profiles, however, structural units tend to be much bigger than those in nonswelling soils, and immediate application of laboratory-scale measurement is uncertain except in the case of the finely divided and relatively homogeneous engineering muds and slurries. Characteristically, structural units in self-mulching vertisols vary from a few millimeters at a dry soil surface to units bounded

by slickensides or by tension cracks, which may have characteristic dimensions of order 1 m. Mass movement to produce gilgai features of order 10 m is common. The approach based on Richards equation requires that at some landscape scale, these features may be considered part of a continuum, but the "engineering" required for management will rely on estimates of the material properties of that continuum. These estimates are difficult and conventional devices, in addition to producing uncertain results (e.g. Jarvis and Leeds-Harrison, 1987), reveal heterogeneities, for example, in water content across structural units that are daunting. The increase in scale also increases characteristic equilibration times over and above the effects attributable to the fine particle size that characterizes these soils. For these reasons, values of soil properties averaged over many "point" measurements are required and variation in these averages over extended periods of time would appear to be necessary if systematic field behavior is to emerge. Ultimately, large-scale, remotely sensed data may provide the only practicable way to monitor the behavior of these landscapes. Nevertheless, some effort has been devoted to creating, from detail at one scale, analysis at another. Bronswijk (1991), Garnier *et al.* (1997a) and others, for example, integrate perceptions of three-dimensional behavior of aggregates and cracks to develop an image of one-dimensional profile behavior. The approach remains confined to the beds of relatively small aggregates, and its application to the field is uncertain because, ultimately, focus must remain on formulation of theory at the scale of application (Philip, 1995; Philip and Smiles, 1982). Physical processes at smaller scales are important but rarely permit quantitative prediction of behavior in the large. The "Darcy" field scale has proved most useful in nonswelling soils. It also provides a rational approach to water equilibrium and flow in swelling soils, with the qualification that the material properties that are necessary for the analysis at this scale must exist and be measurable. Common experience suggests that there is often substantial uniformity in the sense that spatially representative water content profiles and their evolution in time can be discerned. The approaches that follow seek to discover and use this uniformity.

The first challenge is to measure representative water contents and their distribution in the landscape so that water balances can be established. Test of generalized theory to deal with equilibrium and flow may then follow.

Material Balance in Swelling Systems

Material balances in one dimension for both the water and the solid are unambiguous if a spacelike coordinate, based on the distribution of the solid, is used to account for the strain of the solid phase (Raats and Klute, 1968a, 1969; Smiles and Rosenthal, 1968; McGarry and Malafant, 1987; Sposito *et al.*, 1976; Smiles, 1997). Such a coordinate is a preferred way to determine the water balance in swelling soils. The fundamental problem is to measure the way the soil volume reflected in the surface elevation is related to the profile water content. Field experiments of Kirby *et al.* (2003) are exemplary, and their data are similar to observations from earlier studies. In these experiments, anchored rods and neutron probes were used to measure volume and water content change in a vertisol with clay (predominantly smectite) content greater than 70%. The soil sequentially supported maize, fallow, and lucerne (alfalfa) over several years. The soil surface rose as much as 200 mm during wetting and fell as much as 140 mm during the lucerne phase, and revealed an average value of $\alpha \leq 0.3$. Errors as great as 20% in the estimation of water storage change arose when swelling was disregarded.

Cabidoche and Ozier-Lafontaine (1995) similarly used a device called *THERESA* to measure the change in thickness during water content change of a mass-based soil layer. It uses a pair of concentric anchored cylinders of small diameter and it revealed behavior similar to that of Kirby *et al.* (2003). Coquet (1998) used a similar device on a less swelling soil.

Both Cabidoche's and Kirby's experimental sets illustrate a very close relationship between the profile water content and the surface elevation, and the use of a profile $\alpha \approx 0.3$ together with surface elevation permitted calculation of the water content within a few percent. These papers also illustrate the difficulty of field work. Data of Kirby *et al.* (2000) are based on averages of five sets of rods; those of Cabidoche and Ozier-Fontaine initially were based on 36 spatially independent stations located according to characteristic spacing of gilgai micro relief under grassland. At another less variable site, only six stations were considered necessary under sugar cane. Coquet (1998) developed his sampling strategy using a geostatistical approach.

Point measurement of water content and bulk density are also difficult. Coquet (1998) illustrated the scale of water content variability when he developed a sampling strategy for his THERESA-like probes. Aitchison and Holmes (1953a) abandoned attempts to measure bulk density in dry profiles while Kirby and Ringrose-Voase (2000) experienced difficulties with very wet rice paddy soils. Jarvis and Leeds-Harrison (1987) discuss some of these difficulties in relation to the neutron probe method and identified particular problems because of cracking about the access tube. Nevertheless, Bavaye (1991) provides an analysis to relate such measurements to the solid-based material coordinate.

These difficulties also affect estimates of solute balance and the consequences carry over to the extreme example of potentially acid sulfate soils. In such soils, with an initial bulk specific gravity as low as 0.75, calculations of water, solute, and organic carbon balance associated with drainage, for example, may err by a factor larger than three if volume change is neglected (Smiles, 1997). Thus, measurement in detail sufficient to fully define the material coordinate in

field soils is critically important. Measured soil bulk density and water content profiles are necessary in mineral soils. If organic matter content is changing or sulfides are oxidized (as may be the case in potentially acid sulfate soils), then the material coordinate may be based on that component of the solid that remains unchanged (Raats, 1998a,b; Smiles, 1997).

Equilibrium and Flow of Water in Swelling Systems

Saturated Systems

For one-dimensional flow in saturated systems, use of the swelling soil analog of the Richards equation is supported by experimental data for filtration and sedimentation of practically important industrial slurries and effluents (Smiles, 1986; Smiles and Kirby, 1992) where the strain is often very large because initial values of ϑ may be as large as 40 and decrease during filtration or centrifugation to values less than five. Smiles (1999) also showed how insights from nonswelling soil water flow theory significantly simplify analysis of centrifugal filtration of particulate suspensions. In addition, Kim *et al.* (1992a) applied the approach to water flow in marine clay and Smiles (1973) used the approach to explain almost 0.5 m of consolidation of estuarine sediment over more than 500 days following construction of a road embankment that was drained at the sediment-embankment interface.

Well-tested methods also permit unambiguous measurements of $k_{\mathrm{m}}(\vartheta)$ (or $D^{*}_{\mathrm{m}}(\vartheta)$) and $\psi(\vartheta)$ in these systems with the qualification that $\psi(\vartheta)$ must be measured in the unloaded state (REFS). Because these systems are saturated, definition of the shrinkage curves $e(\vartheta, \sigma)$ is not needed, although care is required to ensure that sample constraints do not result in experimental artifacts.

Unsaturated Systems

Equilibrium and flow of water in unsaturated swelling soils is more uncertain and it is necessary also to discriminate between laboratory measurements and field studies at a larger scale, but:

1. Philip (1969b) and Sposito (1975a) used equation (19) to explore static equilibrium profiles in these materials and most measured field profiles confirm the Philip and Sposito insights that equilibrium volumetric water contents tend to decrease with depth.
2. Steady vertical flow profiles calculated using equation (27) are also illustrated by Philip (1969d) and Giraldez and Sposito (1978), but neither paper was based on measured material properties.
3. Smiles and Colombera (1975) and Kirby and Smiles (1999) measured transient water content profiles during infiltration from a constant potential water source into beds of natural swelling aggregates that characterize the surface of dry self-mulching clay soils. These tests focused on the extended initial stages of infiltration when the effects of gravity are minimal. Equation (28) was used and transient water content profiles revealed Boltzmann scaling as the approach anticipates.
4. Perroux and Zegelin (1984) described water content profiles associated with constant flux infiltration into self-mulching heavy clay soil at realistic rates, V_0, in both the laboratory and the field. Theory required, and experiments revealed, scaling in terms of $V_0 m$ and $V_0^2 t$. $D(\vartheta)$ was derived in both experimental sets.
5. Kim *et al.* (1992a) and papers identified by Kim *et al.* (1999) use the full equation (28) to analyze behavior of both saturated and unsaturated clays in the laboratory and the field; Giraldez and Sposito (1985) solved it for flux infiltration but did no experiments.
6. The instantaneous profile method for nonswelling soil (Green *et al.*, 1986) was also used by Kirby and Ringrose-Voase (2000) in the field for draining rice soils and the derived material properties were consistent with subsequent predicted behavior.

The instantaneous profile method would appear useful for soil characterization in dryland water management and might be complemented by triaxial and k_o-cell measurement routinely used by civil engineers to measure the coefficient of consolidation. The k_o-cell, in particular, establishes and monitors constraints similar to those in the field during water flow and volume change.

Experimental clarification of a number of issues is still required, however, particularly regarding two- and three-dimensional aspects of flow, although Kim *et al.* (1999), in laboratory studies, marry isotropic behavior of aggregates to essentially one-dimensional behavior of the profile. In the face of the great difficulties involved in field measurement and modeling, a number of other theoretical concerns also remain. Definition of the overburden potential and the assumption of local equilibrium of water potential across aggregates and voids are problems; volume change and definition of material properties necessary to use equations like (28) must also be studied.

The Components of the Total Water Potential

The Overburden Potential

Practically and theoretically, definition of the overburden potential and its components remains a problem. One difficulty is the integration of equation (22) required to define the overburden potential, Ω. This integration requires that we describe work involved in changing the water content of the unloaded soil and then loading it. The "simple, convenient, and useful" path of Philip (1970) that leads to equation (23) requires the soil to dry from saturation water content, under zero load, to some arbitrary

ϑ- value, and then to increase σ from 0 to P at constant ϑ. The first stage in this process involves drainage, the second wetting, because θ_w then increases as e decreases while ϑ remains constant. Groenevelt and Bolt (1972) argued that, from a thermodynamic point of view, hysteresis reflects metastable situations and can be ignored but, within this very robust assertion, it remains difficult experimentally to match the integration path implied in equation (23).

The equality between $(\partial p_w/\partial P)_\vartheta$ and $(\partial e/\partial \vartheta)_P$ expressed in equation (21) also needs investigation. Practically, the former may be easier to measure than the latter, and Miller (1975) suggested that the general theory should be reformulated using the measurable p_w, rather than the more esoteric "unloaded" matric potential, although changes in ϑ would still require measurement of σ as well as p_w. Nevertheless, Croney *et al.* (1958), Talsma and van der Lelij (1976), and Talsma (1977) made successful comparisons of $(\partial p_w/\partial \sigma)_\vartheta$ and $(\partial e/\partial \vartheta)_\sigma$ *in situ*. Such measurements are important but rare and simultaneous measurement of p_w, and the profile $e(\vartheta, \sigma)$ using installations similar to those initiated by Woodruff (1936) are needed if only to establish operationally useful but approximate ways to deal with Ω in unsaturated soils. These issues will be only resolved by laboratory and field experiment as comprehensive and searching as those of Croney *et al.* (1958).

The Unloaded Capillary Potential Characteristic $\psi(\vartheta)$

Within these uncertainties, the unloaded capillary potential characteristic $\psi(\vartheta)$ should be measured on thin samples in exactly the same way as it is in nonswelling soils. The function appears to be hysteretic. In addition, there remain uncertainties about the nature of ψ as a scalar, that have yet to be resolved (Mahony, 1975). This may be of secondary concern practically, compared with continued uncertainty about the relation between $\psi(\vartheta)$ and Ω. As we note above, Miller (1975) discussed the issue and argued that, because only $p_w(= \psi + \Omega)$ permits unambiguous field measurement in relation to hydrostatics and flow, p_w should be the basis for material characterization; his proposition has received little attention.

Capillary Potential Equilibrium Across Aggregates and Voids

Local equilibration of water potential across aggregates and voids is implied in equation (28) but, in flood and furrow irrigated swelling-cracking soils, water often rapidly penetrates the soil profile via cracks, and aggregates may then equilibrate with the "macropore" water quite slowly.

Philip (1968b) extended the Richards equation to absorption into aggregated media where local potential disequilibrium exists, but his approach has not been tested in swelling soils. Bevan and Germann (1981) and Germann and Bevan (1981, 1985) treated flow in the aggregates as Darcian and described flow in the macropores as purely gravitational so the Richards equation was reduced to a kinematic wave equation. Bevan and Germann were not specifically concerned with swelling soils but Ross and Bridge (1984) similarly considered the process in swelling soils as a vertical "infiltration advance" problem with horizontal absorption into aggregates from a rapidly filled crack system. The issues are defined well by Ross and Bridge: they include interaction between precipitation rate and duration and the soil structure associated with the antecedent water content. If absorption of water by aggregates is rapid enough to accept rainfall (or irrigation) and surface cracks close before ponding, then equation (28) will apply. If the soil is dry so cracks are extensive and rainfall or irrigation is intense, then flow down cracks will be rapid and another infiltration model is required. Kim *et al.* (1992b), for example, believed that almost 30% of heavy rainfall moved directly to the water table in a cracked marine sediment; Favre *et al.* (1997) believe that by-pass flow under both rainfall and flood irrigation was only modest; while Kirby and Ringrose-Voase (private communication) working with a rain-fed, self-mulching vertisol, believed that by-pass flow in cracks was relatively minor. Field experiments setting behavioral limits based on rainfall intensity and duration and their interaction with antecedent water content, crack pattern, and the dynamics of surface crack closure are, therefore, desirable.

At the same time, much anecdotal evidence supports the notion that the wetting front, following rainfall on a swelling soil, penetrated about the same distance across an extensive area and, by implication, the water content profiles were fairly uniform. Fawcett (1972), for example, used a penetrometer to show this more detailed sampling of Cabidoche and Ozier-Lafontaine (1995) and Coquet (1998) led to similar expectations, and Wells *et al.* (2003) found that infiltration profiles in cracking soils became increasingly more uniform during successive simulated (heavy) rainfall events separated by drying phases that induced surface cracking.

In recent years, flow of water in structured soils has been incorporated in various comprehensive models such as HYDRUS, MACRO, and SWAP. Discussion of these models is beyond the scope of this article, but the reader is referred to proceedings of two recent symposia for detail (Feyen and Wiyo, 1999; Feddes *et al.*, 2004)

The Hydraulic Conductivity Characteristic $k(\vartheta)$

Focus on α diverted attention from the fact that if we need to concern ourselves with the full family of $e(\vartheta, \sigma)$ curves as illustrated by Groenevelt and Bolt (1972), then $k(\vartheta)$ cannot be unique in an unsaturated soil and should properly be replaced by $k(\vartheta, e)$ (Smiles, 2000). The effect

may be quite important because k scales with the square of a characteristic hydraulic length (Miller and Miller, 1955). This issue had not been identified previously and requires examination (but see Horn *et al.*, 1995). In the interim, it is worth recalling that the k_0-cells routinely used by engineers to measure engineering parameters of consolidating soil cores provide defined lateral constraint but permit free vertical movement (subject to appropriate load) during flow and volume change. These requirements are not always met. Garnier *et al.* (1998), for example, measured infiltration in swelling soil, but their data appear to be vitiated, in part, by artifacts that arose because of inappropriate sample constraints (Kirby and Smiles, 1999).

Subsidiary Issues

There are subsidiary issues, two of which are identified here.

Physical Chemistry of Clays and its Effect on Flow

The structural and flow consequences of the soil physical chemistry are reflected in macroscopically measured $k(\vartheta)$ and $\psi(\vartheta)$ functions (Smiles *et al.*, 1985). In terms of the water flux, the osmotic potential gradient will be important where the reflection coefficient is significant; Barbour *et al.* (1992) provides a recent review.

The Concept of Effective Stress

The effective stress concept is useful in systems that swell, although it leads to debate about the nature of particle–particle interaction in chemical engineering, and it has no meaning in a nonswelling soil. On the other hand, the notions of a measurable capillary potential, ψ, or manometric pressure p_w, permit general analysis of equilibrium and flow of water in and between swelling and nonswelling systems (e.g. Smiles and Kirby, 1994). The capillary potential also permits us to be quite agnostic about detail of microscopic or molecular interactions that are important but are implicit in the macroscopically measured $\psi(\vartheta)$ function.

CONCLUDING REMARKS

Richards' equation provides a well-tested basis for describing water flow in nonswelling soils. It requires that the macroscopic functions $k(\vartheta)$ and $\psi(\vartheta)$ be well-defined. These properties are readily measured using well-established methods at scales that approximate the scale of field application. Use of the Richards equation has been extended to encompass systematic, and to some extent also stochastic, heterogeneity. It can also be used to describe flow in systems where local potential disequilibrium exists.

For one-dimensional water flow in swelling soils, combination of Darcy's law with the continuity equation for water expressed in material coordinates based on the distribution of the soil solid results in an equation of exactly the same form as the Richards equation. This equation also requires that $k_m(\vartheta)$ and $\psi(\vartheta)$ be well-defined and that $e(\vartheta, P)$, which specifies profile swelling behavior, be also known. This approach has been well tested for saturated systems and effectively applied in some unsaturated cases; its general application to swelling soils awaits systematic examination, however, because of the scale and variability of swelling soils and the much greater difficulty in determining representative values of water content, water pressure, and soil density. Nevertheless, examples of successful application of an approach based on the Richards equation are cited.

Finally, we reiterate the importance of the material coordinate in determining material balance of the water, as well as solutes, in soils that change volume with water content change. For most mineral soils, the solid distribution provides such a coordinate. For organic soils or, for example, acid sulfate soils where reaction may destroy part of the matrix, a coordinate based on the distribution of the irreducible mineral fraction is useful. These coordinates are simple to measure and use, and intensive properties, expressed in terms of unit amount of the coordinate, provide an unambiguous basis for profile comparisons in time and physical space.

FURTHER READING

Ahmad A. and Mermut M. (Eds.) (1996) Water relations and water management of vertisols. In *Vertisols and technologies or their management. Developments in soil science*. Elsevier: Amsterdam.

Bronswijk J.J.B. (1992) A general approach to incorporate swelling and shrinkage processes in soil water transport simulation models. *Modeling Geo-Biosphere Processes*, **1**, 253–270.

Bronswijk J.J.B. (1990) Shrinkage geometry of a heavy clay soil at various stresses. *Soil Science Society of America Journal*, **54**, 1500–1502.

Giraldez J.V. and Sposito G. (1983) A general soil volume change equation: II. Effect of load pressure. *Soil Science Society of America Journal*, **47**, 422–425.

Ringrose-Voase A.J., Kirby J.M., Djoyowasito G., Sanidad W.B., Serrano C. and Lando T.M. (2000) Changes to the physical properties of soils puddled for rice during drying. *Soil and Tillage Research*, **56**, 83–104.

Stroosnijder L. and Bolt G.H. (1984) The moisture characteristic of heavy clay soils. In *Proceeding ISSS Symposium on Water and Solute Movement in Heavy Clay Soils*, Bouma J. and Raats P.A.C. (Eds.), Publication 37, International Institute for Land Reclamation and Improvement: Wageningen..

Talsma T. (1974) Moisture profiles in swelling soils. *Australian Journal of Soil Research*, **2**, 71–75.

REFERENCES

Aitchison G.D. and Holmes J.W. (1953a) Aspects of swelling in the soil profile. *Australian Journal of Applied Science*, **4**, 244–259.

Aitchison G.D. and Holmes J.W. (1953b) Seasonal changes of soil moisture in a red-brown earth and a black earth in southern Australia. *Australian Journal of Applied Science*, **4**, 260–273.

Alonso E.E. and Delage P. (Eds.) (1995) Unsaturated soils. *Proceedings of the First International Conference on Unsaturated Soils*, Paris, 6–8 September, 1995.

Atsumi K., Akiyama T. and Miyagawa S. (1973) Expression of thick slurry of titanium dioxide in water. *Journal of Chemical Engineering of Japan*, **6**, 236–240.

Barbour S.L., Fredlund D.G. and Pufahl D. (1992) The osmotic role in the behaviour of swelling clay soils. In *Mechanics of Swelling*, Karalis T.K. (Ed.), Springer-Verlag: Berlin, pp. 97–139.

Bavaye P. (1991) Operational aspects of the mechanics of deforming porous media: theory and application to expansive soils. In *Mechanics of Swelling*, Karalis T.K. (Ed.), Springer-Verlag: London, pp. 79–96.

Berndt R.D. and Coughlin K.J. (1976) Bulk density changes in a cracking clay soil. *Australian Journal of Soil Research*, **15**, 27–37.

Bevan K. and Germann P. (1981) Water flow in soil macropores: 2. A combined flow model. *Journal of Soil Science*, **32**, 15–29.

Biot M.A. (1941) General theory of three-dimensional consolidation. *Journal of Applied Physics*, **12**, 155–164.

Biot M.A. (1955) Theory of elasticity and consolidation for a porous anisotropic solid. *Journal of Applied Physics*, **26**, 182–185.

Bouma J. and Raats P.A.C. (Eds.) (1984) *Proceedings of the ISSS Symposium on Water and Solute Movement in Heavy Clay Soils*, ILRI Publication 37, International Institute for Land Reclamation and Improvement: Wageningen, pp. xii+ 365.

Bronswijk J.J.B. (1991) *Magnitude, Modeling and Significance of Swelling and Shrinkage Processes in Clay Soils*, Doctoral Thesis, Wageningen Agricultural University, Wageningen.

Bronswijk J.J.B. and Evers-Vermeer J.J. (1990) Shrinkage of Dutch clay soil aggregates. *Netherlands Journal of Agricultural Science*, **38**, 175–194.

Buckingham E. (1907) *Studies on the Movement of Soil Moisture*, Bureau of Soils Bulletin No 38, USDA: Washington.

Cabidoche Y.-M. and Ozier-Lafontaine H. (1995) THERESA: I. Matric water content measurements through thickness variations in vertisols. *Agricultural Water Management*, **28**, 133–147.

Cabidoche Y.-M. and Voltz M. (1995) Non-uniform volume and water content changes in swelling clay soil: II. A field study on a vertisol. *European Journal of Soil Science*, **46**, 345–355.

Childs E.C. (1936) The transport of water through heavy clay soils I. *Journal of Agricultural Science*, **26**, 114–127.

Childs E.C. and George N.C. (1948) Soil geometry and soil water equilibria. *Discussions of the Faraday Society*, **3**, 78–85.

Coleman J.D. and Croney D. (1952) *The Estimation of the Vertical Moisture Distribution with Depth in Unsaturated Cohesive Soils*, RN/1709/JDC.DC, DSIR Road Research Laboratory Report, March 1952.

Coquet Y. (1998) In-situ measurement of the vertical linear shrinkage curve of soils. *Soil and Tillage Research*, **46**, 289–299.

Coquet Y., Touma J. and Boivin P. (1998) Comparison of soil linear shrinkage curve from extracted cores and *in situ*. *Australian Journal of Soil Research*, **36**, 765–781.

Croney D. and Coleman J.D. (1961) Pore pressure and suction in soil. *Pore Pressure and Suction in Soil*, Butterworths: London, pp. 31–37.

Croney D., Coleman J.D. and Black W.P.M. (1958) Studies of the movement and distribution of water in soil in relation to highway design and performance, DSIR Road Research Laboratory Report RN/3209/DC.JDC, WPMB, April, 1958.

Favre F., Boivin P. and Woperis M.C.S. (1997) Water movement and soil swelling in a dry, cracked vertisol. *Geoderma*, **78**, 113–123.

Fawcett R.G. (1972) Agronomic and cropping practices. In *Physical Aspects of Swelling Clay Soils*, James B.J.F. (Ed.), *Proceedings of the Symposium Armidale* February 1972, University of New England: Armidale.

Feddes R.A. de Rooij G.H. and van Dam J.C. (2004) *Unsaturated-zone Modeling. Progress, Challenges and Application*, Kluwer, Dordrecht.

Feyen J. and Wiyo K. (1999) *Modeling of Transport Processes in Soils at Various Scales in Space and Time*, Wageningen Pers: Wageningen.

Fox W.E. (1964) A study of bulk-density and water in a swelling soil. *Soil Science*, **98**, 307–316.

Fredlund D.G. and Rahardjo H. (1993) *Soil Mechanics for Unsaturated Soils*, Wiley Interscience: New York.

Garnier P., Angulo-Jaramillo R., DiCarlo D.A., Bauters T.W.J., Darnault C.G.J., Steenhuis T.S., Parlange J.-Y. and Bavaye P. (1998) Dual-energy synchrotron X ray measurements of rapid soil density and water content changes in swelling soils during infiltration. *Water Resources Research*, **34**, 2837–2842.

Garnier P., Perrier E., Angulo-Jaramillo R. and Bavaye P. (1997a) Numerical model of 3-dimensional anisotropic deformation and 1-dimensional water flow in swelling soils. *Soil Science*, **162**, 410–420.

Garnier P., Rieu M., Boivin P., Vauclin M. and Baveye P. (1997b) Determining the hydraulic properties of a swelling soil from a transient evaporation experiment. *Soil Science Society of America Journal*, **61**, 1555–1563.

Germann P. and Bevan K. (1981) Water flow in macropores: 1. An experimental approach. *Journal of Soil Science*, **32**, 1–13.

Germann P. and Bevan K. (1985) Kinematic wave approximation to infiltration into soil with sorbing macropores. *Water Resources Research*, **21**, 990–996.

Gerzevanov N.M. (1937) *The Foundations of the Dynamics of Soils*, *Third Edition*, Stroiizdat: Moscow.

Gibson R.E., England G.L. and Hussey M.J.L. (1967) The theory of one-dimensional consolidation of saturated clays. *Geotechnique*, **17**, 261–273.

Giraldez J.V. and Sposito G. (1978) Moisture profiles during steady vertical flows in swelling soils. *Water Resources Research*, **14**, 314–318.

Giraldez J.V. and Sposito G. (1985) Infiltration in swelling soil. *Water Resources Research*, **21**, 33–44.

Giraldez J.V., Sposito G. and Delgado C. (1983) A general soil volume change equation: I. The two-parameter model. *Soil Science Society of America Journal*, **47**, 419–422.

Green R.E., Ahuja L.R. and Chong S.K. (1986) Hydraulic conductivity, diffusivity, and sorptivity of unsaturated soils: field methods. In *Methods of Soil Analysis. Part 1. Physical and Mineralogical Methods*, *Second Edition*, Klute A. (Ed.), American Society of Agronomy: Madison, pp. 771–796, pp. 123–128.

Groenevelt P.H. and Bolt G.H. (1972) Water retention in soil. *Soil Science*, **113**, 238–245.

Haines W.B. (1923) The volume changes associated with variations of water content in soils. *Journal of Agricultural Science*, **13**, 296–311.

Horn R., Baumgartl T., Grässle W. and Richards B.G. (1995) Stress induced changes of hydraulic properties of soils. In *Unsaturated Soils*, Alonso E.E. and Delage P. (Eds.), *Proceedings of the First International Conference on Unsaturated Soils*, Paris, pp. 123–128, 6–8 September, 1995.

Jaeger J.C. (1962) *Elasticity, Fracture and Flow*, *Second Edition*, Methuen: London.

Jarvis N. and Leeds-Harrison P.B. (1987) Some problems associated with the use of the neutron probe in swelling/shrinking clay soils. *Journal of Soil Science*, **38**, 149–156.

Jury W.A., Gardner W.R. and Gardner W.H. (1991) *Soil Physics*, John Wiley & Sons: New York.

Kim D.J., Angulo-Jaramillo R., Vauclin M., Feyen J. and Choi S.I. (1999) Modelling of soil deformation and water flow in swelling soil. *Geoderma*, **92**, 217–238.

Kim D.J., Diels J. and Feyen J. (1992a) Water movement associated with overburden potential in a shrinking marine clay soil. *Journal of Hydrology*, **133**, 179–200.

Kim D.J., Vereecken H., Feyen J., Vanclooster M. and Stroosnijder L. (1992b) A numerical model of water movement and soil deformation in a ripening marine clay soil. *Modelling Geo-Biosphere Processes*, **1**, 185–203.

Kirby J.M., Bernard A.L., Ringrose-Voase A.R., Young R. and Rose H. (2003) Field swelling, shrinking and water content change in a heavy clay soil. *Australian Journal of Soil Research*, **41**, 963–978.

Kirby J.M. and Ringrose-Voase A.R. (2000) Drying of some Philippine and Indonesian puddle rice soils following surface drainage: numerical analysis using a swelling soil flow model. *Soil and Tillage Research*, **57**, 13–30.

Kirby J.M. and Smiles D.E. (1999) Comments on "Dual-energy synchrotron X ray measurements of rapid soil density and water content change in swelling soils during infiltration" by P Garnier *et al.*. *Water Resources Research*, **35**, 3585–3587.

Kutilek M. (1996) Water relations and water management of vertisols. In *Water in Heavy soils*, Vols. I-III, *Proceeding of ICID/ISSS Symposium*, Bratislava, (1976), Ahmad N. and Mermut M. (Eds.), Elsevier: Amsterdam, pp. 201–230.

Kutilek M. and Sutor J. (Eds.), (1976) *Water in Heavy Soils*, Vols. I-III, *Proceeding of ICID/ISSS Symposium*, Bratislava, (1976), Czechoslovak Scientific & Technical Society: Prague.

Leu K. (1986) Principles of compressible cake filtration. In *Encyclopedia of Fluid Mechanics*, Cheremisinoff N.P. (Ed.), Gulf Publishing Company: Houston, pp. 865–904.

Mahony J.J. (1975) Tensiometer measurements in anisotropically loaded swelling soils. *Soil Science*, **120**, 421–427.

Marshall T.J., Holmes J.W. and Rose C.W. (1996) *Soil Physics*, *Third Edition*, Cambridge University Press: Cambridge.

McGarity J.W., Hoult E.H. and So H.B. (Eds.), (1984) *The Properties and Utilization of Cracking Clay Soils*, *Reviews in Rural Science 5*, University of New England: Armidale.

McGarry D. and Malafant K.W.J. (1987) A cumulative mass coordinate to determine water profile changes in variable volume soil. *Soil Science Society of America Journal*, **51**, 850–854.

McIntyre D.S., Watson C.L. and Loveday J. (1982) Swelling of a clay soil profile under ponding. *Australian Journal of Soil Research*, **20**, 71–79.

McNabb A. (1960) A mathematical treatment of one-dimensional consolidation. *Quarterly of Applied Mathematics*, **XVIII**, 337–347.

Miller E.E. (1975) Physics of swelling and cracking soils. *Journal of Colloid and Interface Science*, **52**, 434–443.

Miller E.E. and Miller R.D. (1955) Theory of capillary flow: I. Practical implications. *Soil Science Society of America Proceedings*, **19**, 267–271.

Nicholson H.H. and Childs E.C. (1936) The transport of water through heavy clay soils II. *Journal of Agricultural Science*, **26**, 128–141.

Perroux K.M. and Zegelin S.J. (1984) Constant flux infiltration in swelling soil. In *The Properties and Utilisation of Cracking Clay Soils*, *Reviews in Rural Science Vol. 5*, McGarity J.W., Hoult E.H. and So H.B. (Eds.), University of New England: Armidale, pp. 150–154.

Philip J.R. (1968a) Kinetics of sorption and volume change in clay-colloid pastes. *Australian Journal of Soil Research*, **6**, 249–267.

Philip J.R. (1968b) The theory if infiltration in aggregated media. *Australian Journal of Soil Research*, **6**, 1–19.

Philip J.R. (1969a) Theory of infiltration. *Advances in Hydroscience*, **5**, 215–296.

Philip J.R. (1969b) Moisture equilibrium in the vertical in swelling soils. I. Basic theory. *Australian Journal of Soil Research*, **7**, 99–120.

Philip J.R. (1969c) Moisture equilibrium in the vertical in swelling soils. II. Applications. *Australian Journal of Soil Research*, **7**, 121–140.

Philip J.R. (1969d) Hydrostatics and hydrodynamics in swelling soils. *Water Resources Research*, **5**, 1070–1077.

Philip J.R. (1970) Hydrostatics and hydrodynamics in swelling soils, Reply to Youngs and Towner. *Water Resources Research*, **6**, 1248–1251.

Philip J.R. (1973) Flow in porous media. In *Theoretical and Applied Mechanics. Proceedings 13th International Congress Theoretical and Applied Mechanics*, Moscow, Becker E. and Mikhailov G.K. (Eds.), Springer-Verlag: Berlin, pp. 279–294.

Philip J.R. (1980) Field heterogeneity: some basic issues. *Water Resources Research*, **16**, 443–448.

Philip J.R. (1991) Flow and volume change in soils and other porous media and in tissues. In *Mechanics of Swelling*, Karalis T.K. (Ed.), Springer-Verlag: London, pp. 3–31.

Philip J.R. (1995) Phenomenological approach to flow and volume change in soils and other media. *Applied Mechanics Reviews*, **48**, 650–658.

Philip J.R. and Smiles D.E. (1969) Kinetics of sorption and volume change in three-component systems. *Australian Journal of Soil Research*, **7**, 1–19.

Philip J.R. and Smiles D.E. (1982) Macroscopic analysis of the behavior of colloidal suspensions. *Advances in Colloid and Interface Science*, **17**, 83–103.

Pore Pressure and Suction in Soil (1961) Butterworths: London.

Raats P.A.C. (1969) Axial fluid flow in swelling and shrinking porous rods. Abstracts *40th Annual Meeting of the Society of Rheology*, pp. 13.

Raats P.A.C. (1984) Mechanics of cracking soils. In *Proceedings of the ISSS Symposium on Water and Solute Movement in Heavy Clay Soils*, Bouma J. and Raats P.A.C. (Eds.), ILRI publication 37, International Institute for Land Reclamation and Improvement: Wageningen, pp. 23–38.

Raats P.A.C. (1987) Applications of the theory of mixtures in soil science. *Mathematical Modelling*, **9**, 849–856.

Raats P.A.C. (1998a) Kinematics of subsidence of soils with a non-conservative solid phase. In *Proceedings on CD-ROM of a symposium on "New concepts and theories in soil physics and their importance for studying changes induced by human activity" at the 16th World Congress of Soil Science*, held 20–26 August 1998 at Montpellier: France.

Raats P.A.C. (1998b) Spatial and material description of some processes in rigid and non-rigid saturated and unsaturated soils. In *Poromechanics. A Tribute to Maurice A. Biot. Proceedings of the Biot Conference on Poromechanics*. Thimus J.-F., Abousleiman Y., Cheng A.H.-D., Coussy O. and Detournay E. (Eds.), held September 14–16, 1998 at Louvain-la-Neuve: Belgium: Balkema: Rotterdam, pp. 135–140.

Raats P.A.C. (2002) Flow of water in rigid and non-rigid, saturated and unsaturated soils. In *Modeling and Mechanics of Granular and Porous Materials*, Capriz G., Ghionna V.N., and Giovine P. (Eds.), in the series *Modeling and Simulation in Science, Engineering and Technology Series*, Bellomo N. (Ed.), Birkhäuser: Basel, pp. 189–218.

Raats P.A.C. and Klute A. (1968a) Transport in soils: the balance of mass. *Soil Science Society of America Proceedings*, **32**, 161–166.

Raats P.A.C. and Klute A. (1968b) Transport in soils: the balance of momentum. *Soil Science Society of America Proceedings*, **32**, 452–456.

Raats P.A.C. and Klute A. (1969) One-dimensional, simultaneous motion of the aqueous phase and the solid phase of saturated and partially saturated porous media. *Soil Science*, **107**, 329–333.

Richards L.A. (1931) Capillary conduction of liquids through porous mediums. *Physics*, **1**, 318–333.

Rijniersce K. (1983) *A Simulation Model for Physical Soil Ripening in the IJsselmeerpolders*, Lelystad, p. 216.

Rijniersce K. (1984) Crack formation in newly reclaimed sediments in the IJsselmeerpolders. In *Proceedings of the ISSS Symposium on Water and Solute Movement in Heavy Clay Soils*, Bouma J. and Raats P.A.C. (Eds.), ILRI publication 37 International Institute for Land Reclamation and Improvement: Wageningen, pp. 59–62.

Ross P.J. and Bridge B.J. (1984) MICCS: a model of infiltration into cracking clay soils. In *The Properties and Utilisation of Cracking Clay Soils*, *Reviews in Rural Science Vol. 5*, McGarity J.W., Hoult E.H. and So H.B. (Eds.), University of New England: Armidale, pp. 155–163.

Shirato M., Kato H., Kobayashi K. and Sakazaki H. (1970) Analysis of settling of thick slurries due to consolidation. *Journal of Chemical Engineering of Japan*, **3**, 98–98.

Shirato M., Murase T., Iwata M. and Kurita T. (1986) Principles of expression and design of membrane compression-type filter press operation. In *Encyclopedia of Fluid Mechanics*, Cheremisinoff N.P. (Ed.), Gulf Publishing Company: Houston, pp. 905–964.

Skempton A.W. (1961) Effective stress in soils, concrete and rocks. *Pore Pressure and Suction in Soil*, Butterworths: London, pp. 4–16.

Smiles D.E. (1973) An examination of settlement data for an embankment on a wet light clay. *Australian Road Research*, **5**, 55–59.

Smiles D.E. (1974) Infiltration into a swelling material. *Soil Science*, **117**, 140–147.

Smiles D.E. (1986) Principles of constant-pressure filtration. In *Encyclopedia of Fluid Mechanics*, Cheremisinoff N.P. (Ed.), Gulf Publishing Company: Houston, pp. 791–824.

Smiles D.E. (1997) Water balance in swelling materials: some comments. *Australian Journal of Soil Research*, **35**, 1143–1152.

Smiles D.E. (1999) Centrifugal filtration in particulate systems. *Chemical Engineering Science*, **54**, 215–224.

Smiles D.E. (2000) Hydrology of swelling soils: a review. *Australian Journal of Soil Research*, **38**, 501–521.

Smiles D.E., Barnes C.J. and Gardner W.R. (1985) Water relations of saturated bentonite: some effects of temperature and salt concentration. *Soil Science Society of America Journal*, **49**, 66–69.

Smiles D.E. and Colombera P.M. (1975) The early stages of infiltration into a swelling soil. In *Heat and Mass Transfer in the Biosphere: 1. Transfer Processes in the Plant Environment*, de Vries D.A. and Afgan N.H. (Eds.), Scripta Book Company: Washington, pp. 77–85.

Smiles D.E. and Kirby J.M. (1992) Water movement and volume change in swelling systems. In *Mechanics of Swelling*. *Proceedings of a NATO Advanced Research Workshop on the mechanics of swelling from clays to living cells and tissues*, Corfu, (1991), Karalis T.K. (Ed.), Springer-Verlag: Berlin, pp. 33–48.

Smiles D.E. and Kirby J.M. (1994) Dewatering of sodium bentonite using a dry plaster-of Paris mould. *Chemical Engineering Science*, **49**, 3711–3717.

Smiles D.E. and Rosenthal M.J. (1968) The movement of water in swelling materials. *Australian Journal of Soil Research*, **6**, 237–248.

Soil Science Society of America (1997) *Glossary of Soil Science Terms*, Soil Science Society of America: Madison, p. 183.

Sposito G. (1975a) A thermodynamic integral equation for the equilibrium profile in swelling soil. *Water Resources Research*, **11**, 499–500.

Sposito G. (1975b) On the differential equation for the equilibrium profile in swelling soil. *Soil Science Society of America Journal*, **39**, 1053–1056.

Sposito G., Giraldez J.V. and Reginato R.J. (1976) The theoretical interpretation of field observations of soil swelling through a material coordinate transformation. *Soil Science Society of America Journal*, **40**, 208–211.

Talsma T. (1977) Measurement of the overburden component of total potential in swelling field soils. *Australian Journal of Soil Research*, **15**, 95–102.

Talsma T. and van der Lelij A. (1976) Infiltration and water movement in an *in situ* swelling soils during prolonged ponding. *Australian Journal of Soil Research*, **14**, 337–349.

Tempany H.A. (1917) The shrinkage of soils. *Journal of Agricultural Science*, **8**, 312–330.

Terzaghi K. (1923) Die berechnung der durchlassigkeitsziffer des tones aus dem verlauf der hydrodynamischen spannungserscheinungen. *Akademie der Wissenschaften in Wein, Sitzungberichte, Mathematisch-Naturewissenschaftliche Klasse*, Part IIa, **132**, 3–4, 125–138.

Terzaghi K. (1956) *Theoretical soil mechanics*, [8th printing] Wiley: New York.

Tiller F.M. and Cooper H.R. (1960) The role of porosity in filtration: IV. Constant pressure filtration. *American Institute of Chemical Engineers Journal*, **6**, 595–595.

Tiller F.M. and Shirato M. (1964) The role of porosity in filtration: VI. New definition of filtration resistance. *American Institute of Chemical Engineers Journal*, **10**, 61–68.

Towner G.D. (1976) A model for swelling soils and its application to the theory of the overburden potential. In *Water in Heavy Soils*, Vol. I, *Proceeding of ICID/ISSS Symposium*, Bratislava, (1976), Kutilek M. and Sutor J. (Eds.), Czechoslovak Scientific & Technical Society: Prague, pp. 2–7.

Wells R.R., DiCarlo D.A., Steenhuis T.S., Parlange J.-Y., Romkens M.J.M. and Prasad S.N. (2003) Infiltration and surface geometry features of a swelling soil following successive simulated rainstorms. *Soil Science Society of America Journal*, **67**, 1344–1351.

Woodruff C.M. (1936) Linear changes in the Shelby loam profile as a function of soil moisture. *Soil Science Society of America Proceedings*, **1**, 65–70.

Yaalon D.H. and Kalmar D. (1973) Vertical movement in an undisturbed soil: continuous measurement of swelling and shrinkage with a sensitive apparatus. *Geoderma*, **8**, 231–240.

Youngs E.G. and Towner G.D. (1970) Comments on "Hydrostatics and hydrodynamics in swelling soils" by J.R. Philip. *Water Resources Research*, **6**, 1246–1247.

Yule D.F. and Ritchie J.T. (1980a) Soil shrinkage relationships of Texas vertisols: I. Small cores. *Soil Science Society of America Journal*, **44**, 1285–1291.

Yule D.F. and Ritchie J.T. (1980b) Soil shrinkage relationships of Texas vertisols: II. Large cores. *Soil Science Society of America Journal*, **44**, 1291–1295.

Zaslavsky D. (1964) Saturated and unsaturated flow equation in an unstable medium. *Soil Science*, **98**, 317–321.

68: Water Movement in Hydrophobic Soils

DR STEFAN H DOERR[1] AND PROFESSOR COEN J RITSEMA[2]

[1] *Geography Department, University of Wales, Swansea, UK*
[2] *Soil Science Center, Wageningen, The Netherlands*

Hydrophobicity (water repellency) can reduce the affinity for soils to water such that infiltration or wetting may be delayed for periods ranging from as little as a few seconds to in excess of weeks. Soil hydrophobicity is thought to be caused primarily by a coating of long-chained hydrophobic organic molecules on individual soil particles. These substances may be released from a range of plants, decaying organic matter, soil fauna, and microorganisms either naturally or during burning. Hydrophobicity tends to be spatially and temporally highly variable, which makes its effects difficult to observe and predict. It is often most prominent after prolonged dry spells and usually disappears after prolonged contact with water. Owing to the cultivation of certain frequently introduced plant species and the increase in wildfires in some regions, hydrophobicity has developed in previously unaffected areas. Amongst its effects are inhibited plant growth, increased overland flow and soil erosion, uneven spatial and vertical wetting patterns, reduced evaporation, and enhanced risk of groundwater contamination associated with enhanced preferential flow. Recent modeling approaches have been successful in predicting the effects of hydrophobicity on soil water movement at plot or small-field scales, however, approaches to predict in detail its hydrological effects at larger scales are yet to be developed.

INTRODUCTION

Soils are commonly considered to wet readily under rainfall or irrigation. This is, however, not always the case as soils can also behave in a hydrophobic (water repellent) manner. Depending on its degree, hydrophobicity can reduce the affinity of soils to water such that they may resist wetting for seconds, hours, days, or even months (e.g. King, 1981; Dekker and Ritsema, 1994; Doerr and Thomas, 2000). Soil hydrophobicity is typically confined to the organically enriched top centimeters or decimeters of the soil and tends to be both spatially and temporally highly variable. It often disappears after prolonged wet periods, but will usually reemerge during drier periods when soil moisture falls below a critical threshold (Dekker *et al.*, 2001).

Early reports of reduced soil wettability date back to 1837, where Waring (as reviewed by Bayliss, 1911) investigated mycelium growth associated with "fairy rings", a term describing a circular area of vigorous plant growth (i.e. grass or crops) on moist soil, surrounded by withered plants covering dry soil. In many cases the fairy ring phenomena was so abundant locally that it materially affected the yield of crops (DeBano, 2000). In 1910, Schreiner and Schorey (p. 9) reported a Californian soil that "could not be wetted, either by man, by rain, irrigation, or the movement of water from the subsoil". Subsequently, studies by Jamison (1946) showed that soil hydrophobicity affected the productivity of citrus orchards in Florida, and nowadays wetting agents are widely used to increase soil wettability, for example, in horticulture and turf grass industries. During the 1990s and early 2000, it became evident that soil hydrophobicity is not restricted to a few locations with very specific environmental conditions. It has now been reported from all continents except Antarctica, from climates ranging from seasonal tropical to subarctic, from land uses including ploughed cropland, grass and shrubland, and a wide range of forest types, and from soils ranging from coarse- to fine-textured (e.g. Wallis and Horne, 1992; Bauters *et al.*, 2000; DeBano, 2000; Doerr *et al.*, 2000b). Its degree can vary considerably from extremely high as observed under Eucalyptus plantations in Portugal (Doerr *et al.*, 1998) to being detectable only with a purpose built micro-infiltrometer as reported from some agricultural soils in Scotland (Hallett and Young, 1999).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Most workers in this field agree with the notion of Wallis *et al.* (1991) that hydrophobicity in soils is the norm rather than the exception, with its degree being variable.

Apart from its often costly implications for plant growth (e.g. York, 1993; Blackwell, 2000), soil erodibility (Shakesby *et al.*, 2000) and other effects described elsewhere (DeBano, 2000; Doerr *et al.*, 2000a) soil hydrophobicity has substantial repercussions for water movement in soils. These include reduced infiltration rates leading to enhanced Hortonian overland flow (cf. **Chapter 111, Rainfall Excess Overland Flow, Volume 3** and **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**), uneven wetting patterns, development of preferential flow (cf. **Chapter 78, Models of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2**), and the accelerated leaching of agrichemicals (e.g. Imeson *et al.*, 1992; Ritsema *et al.*, 1997; Doerr *et al.*, 2003). These soil hydrological effects are discussed in the following sections after a brief introduction to the origin and classification of hydrophobicity. This is then followed by an overview of current approaches to modeling water movement in hydrophobic soils at a range of scales.

ORIGIN AND CLASSIFICATION

The Principles Underlying Hydrophobicity

A wettable (i.e. hydrophilic) surface allows water to spread over it in a continuous film, whereas water on a hydrophobic surface "balls up" into individual droplets. If the surface is a porous medium like sand or soil, water infiltration is inhibited (Figure 1). The affinity or repellency between water and solid surfaces originates from mutual attractive forces (adhesion) and the attraction between the water molecules (cohesion). To understand these forces better, some properties of water are briefly considered here.

A water molecule comprises an oxygen atom with a partial negative charge and two hydrogen atoms with a partial positive charge. The hydrogen and oxygen atom bonds are positioned 105° apart, giving the water molecule a strongly dipolar structure (Parker, 1987). The attraction of these positive and negative ends causes water molecules to form aggregates, held together by hydrogen bonds. Water adheres to most natural surfaces since they consist of positively and negatively charged ions attracting the negative end or the positive ends of a water molecule respectively. The dipole character of water, however, also results in a comparatively strong force counteracting the attraction to charged surfaces. Within a liquid, the net force acting on an individual molecule is zero as it is surrounded by other molecules and their forces. Beyond the surface of a liquid, however, no similar molecules exist to oppose the attraction exerted by the molecules within the liquid. Consequently, the surface molecules experience a net attractive force towards the interior, which promotes the reduction of the surface area of water. Thus, if opposing forces are minimal, liquids will assume a spherical shape (i.e. that of a droplet). To enlarge the surface of a liquid, work is necessary. This work is related to the surface tension (γ) or surface-free energy of the liquid. Most liquids have a surface tension of $20-40\,\mathrm{N\,m^{-1}} \times 10^{-3}$ at 20 °C, but that of water is exceptionally high at $72.75\,\mathrm{N\,m^{-1}} \times 10^{-3}$ (Parker, 1987).

Figure 1 Water droplets resisting infiltration into a sandy dune soil sample from the Netherlands because of extreme hydrophobicity. (photo: E. van den Elsen)

The same principle applies to solid surfaces, although their nature inhibits deformation into a spherical shape. The surface tension of solids therefore leads to lateral forces at the surface. Values for hard solids such as soil minerals range from $500-5000\,\mathrm{N\,m^{-1}} \times 10^{-3}$, increasing with hardness and at its melting point (Zisman, 1964). For water to spread on a solid, the adhesive forces between them must exceed the cohesive forces within the body of water. Thus, surfaces with a surface-free energy $>72.75\,\mathrm{N\,m^{-1}} \times 10^{-3}$ attract water and are therefore hydrophilic. The higher the surface tension of the solid, the stronger is the attraction. All principal soil minerals have a much higher surface-free energy than water and are therefore hydrophilic (Tschapek, 1984), whereas soft organic solids, such as waxes or organic polymers can exhibit γ values below $72.75\,\mathrm{N\,m^{-1}} \times 10^{-3}$ and are thus hydrophobic (Zisman, 1964).

Origin of Hydrophobicity in Soils

In theory, a single layer of hydrophobic molecules can render a hydrophilic mineral surface hydrophobic (Zisman, 1964), although in practice hydrophobic compounds may not be absorbed as uniform monolayers, but as small globules. Thus, an amount equivalent to that of several

monolayers may be required to result in a complete cover on a mineral grain (Ma'shum *et al.*, 1988). This amount is still relatively small. For example, Ma'shum *et al.* (1988) induced severe hydrophobicity in 1000 g of medium-sized sand using only 0.35 g of the hydrophobic compound. Soil hydrophobicity can also be caused by the presence of hydrophobic interstitial matter. If hydrophobic particles are present in the pore spaces of a hydrophilic matrix, the wettability of the composite material is reduced. For example, severe hydrophobicity has been induced by intermixing as little as 2–5% by weight of solid organic matter to hydrophilic sand (McGhie and Posner, 1981). For naturally hydrophobic sand, it has been suggested that slight to moderate levels can be caused by the presence of hydrophobic particles in a soil matrix, but that more extreme hydrophobicity results from a coating on individual soil particles (Bisdom *et al.*, 1993).The biological origin of hydrophobic compounds in soils is undisputed, however, the exact source(s) and chemical nature of these compounds have to date not been fully established. Organic compounds with hydrophobic properties are naturally abundant in the biosphere and may be gradually released into the soil, for example, as roots exudates (Dekker and Ritsema, 1996; Doerr *et al.*, 1998), from soil fauna (Wienhold and Gish, 1991), fungi (Savage *et al.*, 1969) or microbes (Hallett and Young, 1999), or directly from decomposing organic matter (McGhie and Posner, 1981). Depending on vegetation cover present and soil temperatures reached, a rapid release or redistribution of hydrophobic compounds often accompanies the burning of vegetation, leading to a distinct hydrophobic layer at the soil surface or at some depth in the soil profile (DeBano *et al.*, 1976; Doerr *et al.*, 2004) (cf. **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**).

The specific types of compounds suggested to be a cause of hydrophobicity include alkanes (Ma'shum *et al.*, 1988; Roy *et al.*, 1999; Horne and McIntosh, 2000) fatty acids and their salts, and esters (Ma'shum *et al.*, 1988; Franco *et al.*, 2000; Roy *et al.*, 1999) and other related compounds such as phytanes, phytols, and sterols (Franco *et al.*, 2000; Morley *et al.*, 2005). The presence of these compounds, however, does not necessarily lead to the expression of hydrophobicity in soils. It has been demonstrated that comparable amounts of such compounds occur also in wettable soils (Doerr *et al.*, 2005; Morley *et al.*, 2005) and it has been suggested that these compounds only cause hydrophobicity when they form a specific molecular arrangement (Roy and McGill, 2000; Morley *et al.*, 2005).

Classification of Soil Hydrophobicity

Most techniques for measuring and classifying soil hydrophobicity are described in Wallis and Horne (1992), Hallett and Young (1999), Bachmann *et al.* (2000), Letey *et al.* (2000), Wang *et al.* (2000), and Roy and McGill (2002). One of the most common methods, the "Water Drop Penetration Time" (WDPT) test (Van't Woudt, 1959) is referred to later and is therefore briefly described here. It involves placing droplets of distilled water onto a sample surface and recording the time for their complete infiltration. This test broadly determines how long hydrophobicity persists in the contact area of a water droplet. This *persistence* or stability of hydrophobicity is usually somewhat, but not always well, related to the apparent surface tension (i.e. the *severity* of hydrophobicity) of a soil (Dekker and Ritsema, 1994; Doerr, 1998). Hydrophobicity *severity* will determine how strongly a droplet initially balls up when in contact with the soil and can be determined by contact angle measurements (see Letey *et al.*, 2000). Contact angles can, however, change rapidly if the *persistence* of hydrophobicity is low. Perception of what constitutes a high or low degree of hydrophobicity varies widely. To distinguish between hydrophilic and hydrophobic soils, an arbitrary WDPT threshold of 5 s (Bisdom *et al.*, 1993) has been used widely, although considerable effects on water movement at the centimeter-scale have been shown to be caused by lower levels of repellency (Hallett *et al.*, 2004). WDPT values found in the literature are generally presented in categories rather than distinct values and are not always directly comparable because, for practical reasons, this test is often terminated before droplet penetration occurs. In many studies to date, WDPTs exceeding one h have been recorded (= classed as extreme persistence, Bisdom *et al.*, 1993), but *persistence* can in some cases reach levels such that even ponded water may not infiltrate for periods exceeding one month (Doerr and Thomas, 2000). Although the aforementioned laboratory tests allow a general classification of soils according to hydrophobicity *persistence* and *severity*, these measurements do not always relate well to the actual wetting behavior of field soils (Doerr and Thomas, 2000). Thus, Dekker *et al.* (1999) and Doerr *et al.* (2003) recommended the wider use of wetting rate measurements and rainfall simulations.

HYDROPHOBICITY EFFECTS ON SOIL WATER MOVEMENT

The main hydrological impacts of hydrophobicity are: (i) reduced infiltration; (ii) increased overland flow; (iii) spatially localized infiltration and/or percolation, often with fingered flow development; (iv) modifications of the three-dimensional distribution and dynamics of soil moisture; (v) enhanced streamflow responses to rainstorms; and (vi) enhanced total streamflow. Because of enhanced overland flow on, and increased erodibility of, hydrophobic soil, slopewash, and sometimes the formation of rills and gullies may be promoted (see review by Shakesby *et al.*, 2000).

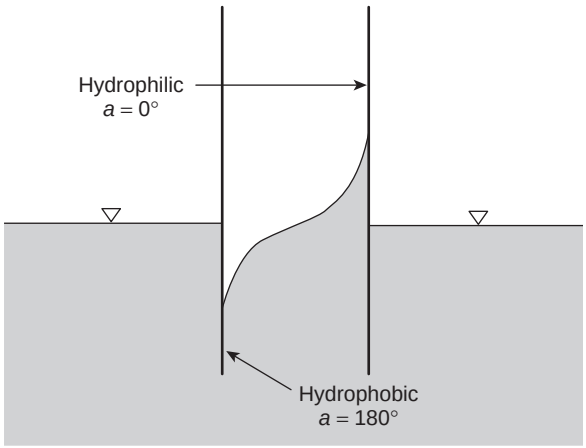


Figure 2 Shape of the meniscus of water between a hydrophilic and a hydrophobic plate with liquid–solid contact angles of 0° and 180° respectively (adapted from Bauters *et al.,* 2000)

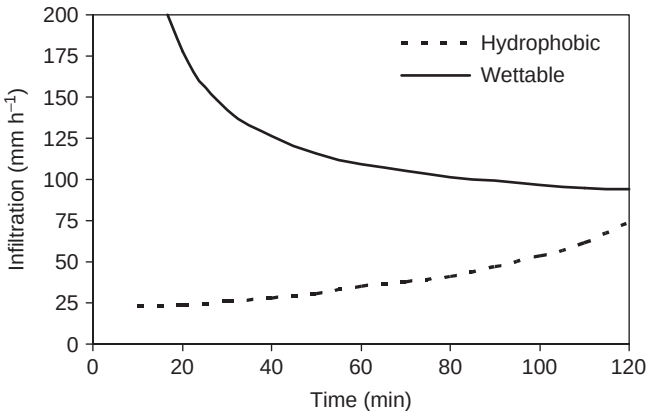


Figure 3 Infiltration rates into hydrophilic and hydrophobic soils (adapted from Letey *et al.,* 1962)

Infiltration and Overland Flow

The most frequently reported impacts of hydrophobicity are those of reduced infiltration into soil (e.g. Van Dam *et al.*, 1990; Imeson *et al.*, 1992; Doerr *et al.*, 2003) and thus increased overland flow (e.g. McGhie and Posner, 1980; Witter *et al.*, 1991; Crockford *et al.*, 1991). For example, the infiltration capacity of a hydrophobic soil was found to be 25 times higher than for a similar soil rendered hydrophilic by heating (DeBano, 1971) and for the first 5 min of measurement a hydrophobic soil was found to exhibit only 1% of the potential infiltration capacity when hydrophilic. Owing to the reduced surface tension of the soil pore walls, a hydrophobic soil matrix will have a *positive* soil water potential. This effectively leads to a capillary depression effect illustrated in Figure 2. Infiltration into a relatively dry hydrophobic soil matrix will therefore only occur if a ponding depth (hydraulic head; entry pressure) sufficient to exceed the positive soil water potential is reached. The necessary ponding depth may decrease with time if hydrophobicity decays during water contact. These factors acting in combination usually result in infiltration rates to *increase* during rainfall, although an initial decrease followed by an increase has also been observed (Clothier *et al.*, 2000). Infiltration theory for wettable soils, however, ascribes infiltration rates to *decrease* over time (Figure 3; Letey *et al.*, 1962; Kirkham and Clothier, 2000). Infiltration behavior of hydrophobic soils has been investigated in detail in a range of laboratory-based studies (Bauters *et al.*, 1998; Carrillo *et al.*, 2000; Wang *et al.*, 2000; Feng *et al.*, 2001). Observed behavior, however, may not always be relevant to field conditions, since some workers (e.g. Bauters *et al.*, 1998; Carrillo *et al.*, 2000; Feng *et al.*, 2001) have used porous media with a permanent (nondecaying) hydrophobicity to avoid the confounding effects of unstable hydrophobicity, although these effects typically do occur under field conditions.

On a hydrophobic soil surface in the field, rainwater will pond and, if rainfall is sufficient and surface detention is exceeded, Hortonian (infiltration-excess) overland flow will occur. The frequency of "gaps" through this layer (such as structural or drying cracks, root holes and burrows, and patches of hydrophilic or less hydrophobic soil) will then determine whether overland flow is widespread or only local (Figure 4a). Under some conditions, for example, where intense soil heating during a wildfire has destroyed hydrophobicity in the top few centimeters of the soil (see DeBano, 1971), a distinct hydrophobic layer underlies a hydrophilic and often highly permeable topsoil or ash layer. Rainfall infiltrating such a topsoil may pond above the hydrophobic layer (Figure 4b) and can subsequently:

1. be stored in the hydrophilic layer and later evaporated or used in transpiration;
2. run off as saturation overland flow when the hydrophilic layer becomes saturated;
3. spread out as "distribution flow" and move vertically downwards as "preferential flow" either through structural or other gaps in the water-hydrophobic layer or as "fingered flow" through vertical cylinders of hydrophilic or less hydrophobic soil as discussed in more detail later;
4. move laterally downslope as throughflow above the hydrophobic layer;
5. enter the matrix of the hydrophobic layer (if ponding depth is sufficient, or hydrophobicity undergoes a phase-change to a hydrophilic condition).

Where overland flow is considered the key effect of hydrophobicity, it is often not clear whether Hortonian or saturation overland flow, or a combination of the two, is involved, particularly if the hydrophilic layer is thin,

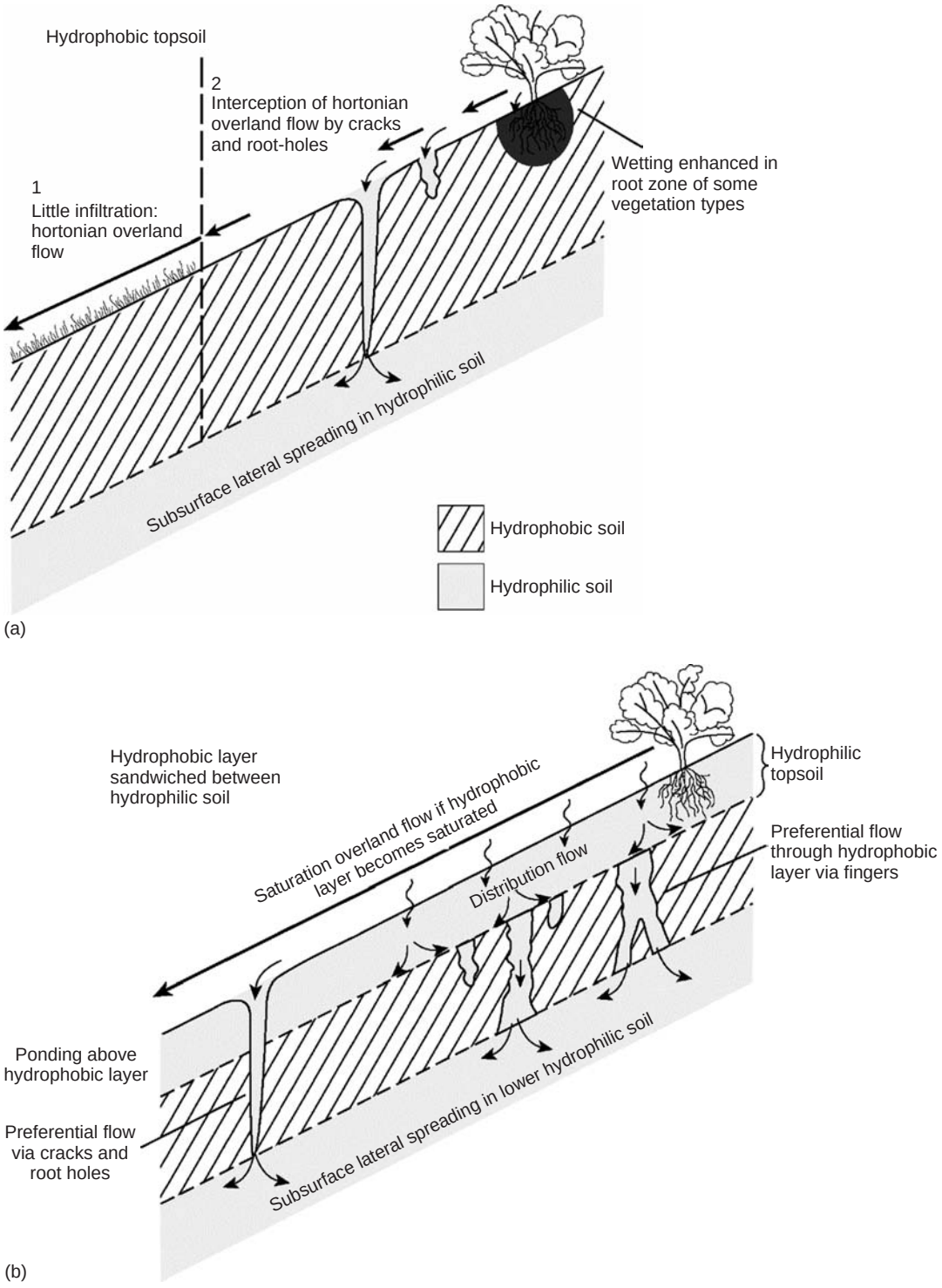


Figure 4 Schematic illustration of possible hydrological responses of soil with (a) a hydrophobic layer located on the surface and (b) a hydrophobic layer sandwiched between hydrophilic soil (adapted from Doerr *et al.,* 2000a)

below the surface, or discontinuous. Thus, in an Australian eucalyptus forest, Burch *et al.* (1989) reported that the development of soil hydrophobicity following drought conditions led to a threefold increase in the rainfall to runoff conversion from an initial 5 to 15%. In a burnt pine forest in South Africa, saturation overland flow promoted by a subsurface hydrophobic layer led to an increase in the stormflow response of 7.5% compared to just 2.2% on

unburnt terrain (Scott and Van Wyk, 1990), and Jungerius and De Jong (1989) attributed the lack of any simple relationship between rainfall and overland flow in Dutch sand dunes to spatial variations in hydrophobicity.

Some studies, particularly those investigating burnt terrain, highlight hydrophobicity as one of several factors enhancing overland flow responses (cf. **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**). Thus, Dyrness (1976) reported a threefold increase in overland flow after a fire in a pine forest in Oregon and Walsh *et al.* (1994) found a 5–25% higher overland flow response on a burnt Portuguese eucalyptus and pine forest compared with the unburnt forest. Rather than invoking only hydrophobicity, postfire increases in overland flow are also attributed to removal of a protective vegetation cover (with resulting increases in rainbeat compaction, inwash of fines into cracks, and rootholes and reduction in infiltration capacity), reduction in soil particle size, erosion of permeable topsoil (cf. **Chapter 26, Weather Patterns and Weather Types, Volume 1**), stone lag development, and organic matter losses (White and Wells, 1982; Imeson *et al.*, 1992; Shakesby *et al.*, 1996). Such increases were demonstrated using rainfall simulation plot experiments carried out in Portugal during dry conditions on unburnt Pinus pinaster slopes and a nearby area that was burnt two years earlier (Walsh *et al.*, 1998). In both areas, most surface soil remained dry and intensely hydrophobic throughout the 1-h 40–46 mm artificial rainstorms. However, whereas in the unburnt soil, overland flow was modest (4%) and most water infiltrated through cracks and rootholes, on the burnt soil overland flow was more substantial (8–20%).

Thus, depending on prefire conditions, postfire increases in overland flow can be associated with:

1. fire inducing or significantly increasing hydrophobicity, thus enhancing overland flow responses until soil hydrophobicity reaches prefire levels;
2. increased effectiveness of preexisting hydrophobicity, where a return to prefire conditions of a prevailing, but less effective soil hydrophobicity would await revegetation and the reformation of a litter mulch;
3. soils that are hydrophilic before and after the fire, but the loss of interception capacity amongst other fire-related changes also enhances overland flow (Zierholz *et al.*, 1995).

Hortonian overland flow (cf. **Chapter 111, Rainfall Excess Overland Flow, Volume 3**) tends to be most pronounced where soils have an uninterrupted hydrophobic layer. Figure 4(a/1) provides an explanation for the contrasts in overland flow response on hydrophobic terrain reported in the literature. Several studies have pointed towards the localization of overland flow on hydrophobic soils. For example, only 25% of the land surface following a fire in southwest Oregon was hydrophobic, resulting in a low impact on infiltration and overland flow (McNabb *et al.*, 1989). Also, Meeuwig (1971) and Imeson *et al.* (1992), working in pine forests in North America and northeast Spain respectively, found that Hortonian overland flow generated on hydrophobic soil around trees tended to infiltrate on adjacent hydrophilic soil around shrubs (Figure 4(a/2)). This pattern was also found using simulated rainstorms in southwest Spain (Cerdà *et al.*, 1998).

The commonly observed temporal variability of hydrophobicity also needs to be considered here. Thus, reductions in infiltration capacity and increases in overland flow can be expected to be most pronounced following prolonged dry periods, when hydrophobicity tends to become most severe. For example, Burch *et al.* (1989) recorded infiltration capacities in an Australian eucalyptus forest of $0.75–1.9\,\mathrm{mm\,h^{-1}}$ when dry, but $7.9–14.0\,\mathrm{mm\,h^{-1}}$ when wet. In many areas, hydrophobicity-linked overland flow can therefore be confined to storm events following dry weather (Sevink *et al.*, 1989; Walsh *et al.*, 1994). Hortonian overland flow responses following dry weather when the soils tend to be most hydrophobic contrast sharply with the muted overland flow following prolonged wet weather when soils are generally hydrophilic (Jungerius and DeJong, 1989; Ferreira *et al.*, 2000). Since few studies address this issue and the understanding of the processes involved in the breakdown of hydrophobicity is poor, it is often not known how long reduced infiltration associated with hydrophobicity persists during wet conditions. For example, in the eucalypt forest in Australia, Crockford *et al.* (1991) found that several weeks of consistently wet weather were required for hydrophobicity to break down and nine hot days were sufficient for its reestablishment. For Portuguese eucalypt stands, Ferreira *et al.* (2000) found that some hydrophobicity still remained after a week of wet weather with 200 mm rainfall, whilst it took up to two months for surface hydrophobicity to become completely reestablished from a wettable soil status (Leighton-Boyce *et al.*, 2005).

Finally, the effects of hydrophobicity on overland flow generation should also be seen in relation to the spatial scale of measurements. Imeson *et al.* (1992) argued that although hydrophobic soils can produce high overland flow rates locally, the effects at slope or catchment scales can more subdued because of the high spatial variations in infiltration. Similar findings were reported by Roberts and Carbon (1971), Pradas *et al.* (1994), and Doerr *et al.* (2003).

Preferential Flow (Including Fingered Flow)

Preferential flow is the concentrated vertical movement of water via preferred pathways through the soil. It may originate for a variety of reasons such as cracks and biopores, textural discontinuities, and unstable wetting fronts, which may result from soil layering, air entrapment, and so on (Ritsema *et al.*, 1993). Although not restricted to

hydrophobic soils (Kung, 1990; Ritsema and Dekker, 1994a), hydrophobicity can be particularly effective at preventing or hindering downward water movement, directing it into structural or textural preferential flow paths (Figure 4 (a/2) and 4 (b)), and creating an unstable irregular wetting front (Figure 5). Consequently, soils may not wet completely with the passage of a wetting front (DeBano, 1971), and water may be channeled via biopores, cracks, and pipes, thereby bypassing the soil matrix (Burch *et al.*, 1989). Root channels and rodent burrows are thought to represent particularly effective bypass routeways through hydrophobic soil (Meeuwig, 1971; Garkaklis *et al.*, 1998). Walsh *et al.* (1995) considered that such bypass routes explained why even large storms produced little overland flow for highly hydrophobic mature pine and eucalyptus forest soils in Portugal.

Where a hydrophobic layer is overlain by hydrophilic soil, infiltrating water tends to pond above the former, which is then directed via lateral flow to preferential flow routeways, leading it through the underlying hydrophobic layer (Figure 4b). This phenomenon has been described for many burnt soils (see review by DeBano, 1981) and has been investigated further by Ritsema *et al.* (1993) and Ritsema and Dekker (1995), who have termed the process *distribution flow*. On a grass-covered sandy dune in Holland, Ritsema *et al.* (1993) used tracers to record distribution flow within a thin (<2.5 cm) hydrophilic, relatively moist humose topsoil, which supplied water via columns of less hydrophobic soil (preferential flow paths) in an otherwise extremely hydrophobic layer to a second hydrophilic zone below 45-cm depth, where the water spread laterally. This has been termed *fingered flow* (Ritsema and Dekker, 1994a,b). The fingers formed only after dry weather when soil moisture levels in the sandy, hydrophobicity-prone layer were below a "critical" value of 4.75% (vol.) (Figure 6). The fingers ranged from 10–50 cm in width (expanding in wetter weather), being the sole means of water transport for several hours during sustained rainfalls. Such fingers have been shown to recur at the same places in successive storms following intervening dry weather, possibly aided by preferential leaching of hydrophobic substances from the finger pathways (Ritsema *et al.*, 1998). In contrast, instead of any fingered flow, a uniform wetting front may develop in a hydrophobic layer if the overlying hydrophilic top-layer is very thick (Van den Bosch *et al.*, 1999). Preferential flow in general is thought to be reinforced by soil water hysteresis between wetting and drying phases, a feature of most hydrophilic soils, but more pronounced in hydrophobic ones (Ritsema *et al.*, 1998).

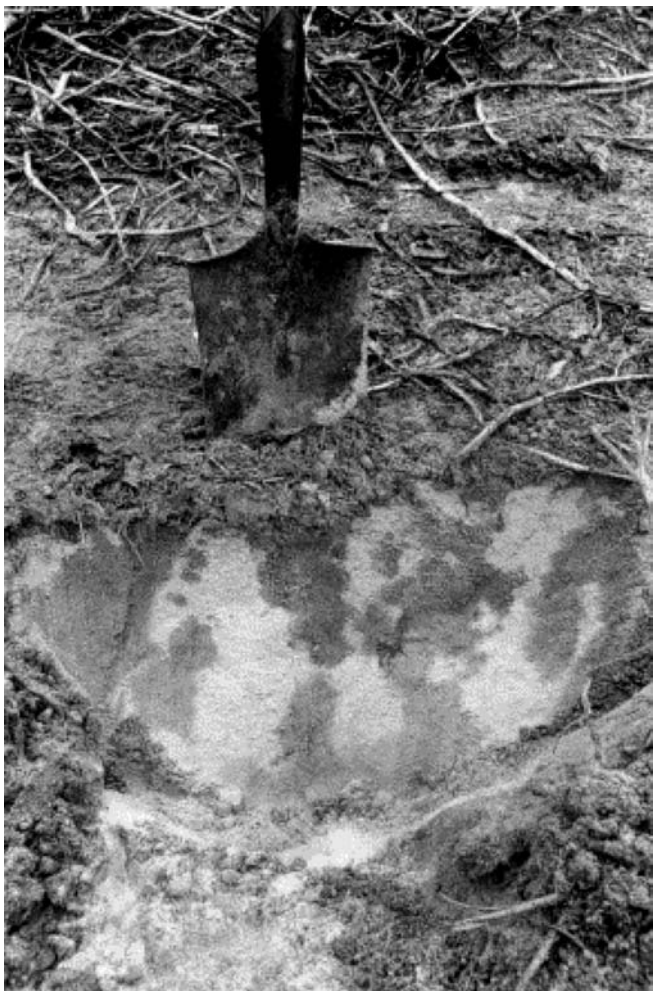


Figure 5 Unstable irregular wetting front and preferential flow pathways in a hydrophobic sandy agricultural soil in Western Australia (photo: S. Doerr)

Effects on the Three-dimensional Distribution and Dynamics of Soil Moisture

As is evident from the previous sections, hydrophobicity influences the three-dimensional distribution and dynamics of soil moisture. Effects will vary with the vertical position of the hydrophobic zone, the frequency of preferential flow routeways through this zone, and the *severity* and *persistence* of its hydrophobicity. For example, a strongly hydrophobic surface layer with preferential flow routeways can lead to dry surface soil and higher soil moisture in the subsoil (e.g. Meeuwig, 1971; Burch *et al.*, 1989; Imeson *et al.*, 1992). In a study in northeast Spain, Imeson *et al.* (1992) described how a hydrophobic surface layer not only trapped water in the B/C horizon, but also prevented evaporation and upward capillary movement of water. It is possible that exudation of hydrophobic compounds is a strategy employed by some plants, fungi, or microorganisms to reduce evaporation, to reduce nutrient leaching in parts of the soil profile, or to suppress the germination and growth of competing vegetation.

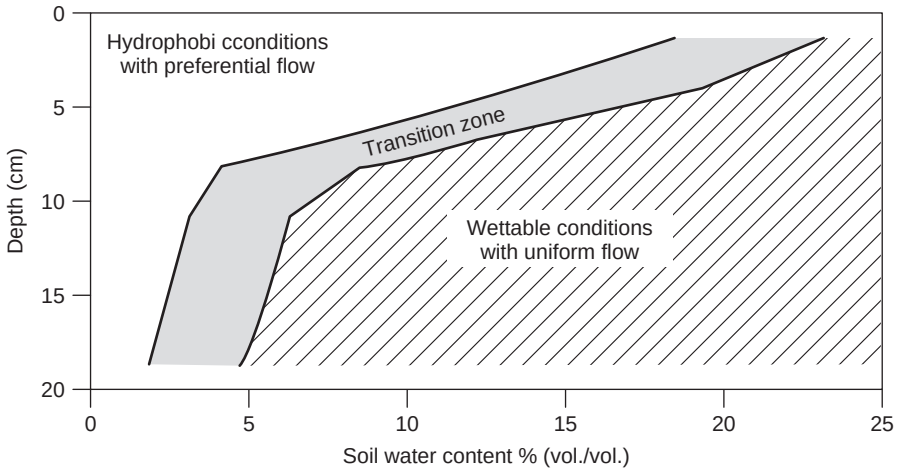


Figure 6 Schematic example of the critical soil moisture threshold changing with depth derived from a sandy soil under permanent grass cover in the Netherlands. Above and to the left of the critical soil water content line, the soil is hydrophobic, which results in the formation of preferential flow paths during infiltration events. At higher soil water contents (below and right of the lower line), the soil is wettable with predominantly uniform flow behavior (modified from Dekker *et al.*, 2001)

Hydrophobicity-induced fingered flow can lead to considerable variations in water content in an initially hydrophobic soil such that zones of very dry soil can be directly adjacent to zones of wet soil. For example, Dekker and Ritsema (1996) found differences in soil moisture of up to 28% (vol.) between closely spaced samples in both clay and sandy soils (see also Figure 5). Such differences not only result in the widely reported poor seed germination and plant growth. Any type of preferential flow path formation can also lead to accelerated leaching of surface-applied agrichemicals and an increased risk of surface and groundwater contamination (Hendrickx *et al.*, 1993; Ritsema *et al.*, 1993).

Effects on Streamflow Generation and Patterns (cf. Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3)

The tendency for fire-induced or enhanced hydrophobicity, combined with vegetation removal, to increase both total streamflow and the magnitude of storm peaks, as well as reducing peakflow response times has been well established (White and Wells, 1982). For example, increases of 800% for streamflow and 450% for catchment runoff efficiency during the first postfire wet season compared to prefire values were attributed in large part to hydrophobicity-enhanced overland flow in a pine forest catchment in Arizona (Campbell *et al.*, 1977). More recent studies have suggested that impacts may be more complex. Scott and Lesch (1997) attributed the lack of streamflow nine years after the afforestation of grassland catchments with *Eucalyptus grandis* and *Pinus patula* in South Africa to an enhanced deep drainage through the hydrophobic soil via preferential flow along the eucalyptus root channels. Instead of promoting overland flow, hydrophobicity associated with afforestation-enhanced water storage at greater depths in the soil, subsequently used in transpiration. Investigating hydrological responses at a range of spatial scales in a catchment dominated by *E. globulus*, Doerr *et al.* (2003) found that the highly hydrophobic soil status associated with dry antecedent conditions enhanced overland flow responses at plot scales (0.12 and 16 m^2) but did not result in greater storm runoff at the catchment scale (0.33 km^2) than was generated by storms with similar rainfall totals and intensities following moist and more wettable soil conditions. Streamflow peak response time, however, was considerably shorter for hydrophobic, compared to more wettable conditions. It was argued that an increased capture of overland flow generated locally under hydrophobic conditions with increasing scale of investigation may be one of the reasons for the limited effects on streamflow magnitude.

Variations in hydrophobicity, its response to fire, and its interaction with other factors were thought by Scott (1993) to be responsible for differences in catchment response to fire between Fynbos vegetation, *Pinus radiata*, and *Eucalyptus festigata* in mountain catchments in South Africa. In the native Fynbos catchments, only a modest increase in total discharge (16%) and no increase in stormflow occurred following both prescribed fire and wildfire. This low response was attributed to only a minor and spatially patchy increase in hydrophobicity, causing little overland flow to reach the base of slopes. The increase in total flow was thought to be due to reduced evapotranspiration. In contrast, in the *E. festigata* and *P. radiata* catchments, the stormflow component increased by 92% and 201% respectively, the latter being attributed to

fire-induced hydrophobicity. For the former, hydrophobicity had been a prefire feature, which became more effective in promoting overland flow following fire.

EVALUATION AND MODELING HYDROLOGICAL EFFECTS OF SOIL HYDROPHOBICITY

Clearly, hydrophobicity can affect soil water movement in a fundamental way such that conventional principles of infiltration and vadose zone water transport may not be applicable. In contrast to wettable soils, which are often dominated by more or less uniform flow conditions, hydrophobic soils exhibit more pronounced preferential flow conditions as outlined earlier. Fundamentally, hydrophobicity induces an extreme form of hysteresis in the water retention characteristic of soils, which during infiltration events might result in wetting front instabilities and the formation of preferential flow paths. These preferential flow paths typically protrude through the entire hydrophobic zone of a soil, which may extend to depths exceeding of 50 cm (Ritsema *et al.*, 2000), but vanish in underlying, more wettable subsoil layers (Figure 4b). Also, preferential flow paths might disappear in case of rising water tables or after an abundance of rainfall. This indicates that soils might shift from a wettable to a hydrophobic state and *vice versa*, causing flow and transport to be either predominately uniform or preferential. The so-called "critical soil water content" as introduced by Dekker and Ritsema (1994) and Ritsema and Dekker (1994b), and modified by Dekker *et al.* (2001), demarcates the soil moisture content range, in which the transition between these two hydrological states of a soil may occur (Figure 6). The critical soil moisture content appears to vary with, for example, soil organic matter content (Dekker *et al.*, 2001) or soil texture (Doerr *et al.*, 2000b). As the complex and dynamically changing state of soils from wettable to hydrophobic and *vice versa* is strongly influenced by the sequence of weather conditions, soil characteristics, and antecedent soil moisture distributions, it is particularly challenging to model and simulate.

The first modeling attempts to simulate flow and transport through hydrophobic porous media were performed by Van Dam *et al.* (1990), using a static two-domain approach in which the area between the fingers behaved inert, and a calibration factor was used to match computed results of flow and transport properties with observed field data. A slightly more advanced model was developed by Steenhuis *et al.* (1994), in which it was assumed that fingers could only form after saturation of the so-called distribution zone. This zone is located at the top of the soil profile and is generally only a few centimeters thick. In case of fingers, solutes move at a velocity corresponding to saturated unit gradient flow, and the fraction of soil participating in fingered flow was made dependent on changes in precipitation.

On the basis of the field observations by Ritsema *et al.* (1993), De Rooij (1995) used a three-region concept in modeling flow and transport in hydrophobic soils. In this approach, the soil is divided into a distribution zone overlying a hydrophobic layer, with a wettable subsoil underneath. The model, which is basically analytical in character, showed that arrival times of solutes at the underlying groundwater level are shorter when the hydrophobic layer increases in thickness.

The initial formation of preferential pathways in hydrophobic soils and their recurrence during subsequent infiltration events have been modeled by Nieber (1996) and Ritsema *et al.* (1998), using two-dimensional numerical modeling approaches and particle tracking routines. Results showed that wetting front instabilities might occur for extremely hydrophobic conditions and that they are characterized by highly hysteretic water retention functions.

The most detailed approach in modeling water movement in hydrophobic soils to date has been made recently by adapting the established SWAP (Soil-Water-Atmosphere-Plant) model (Van Dam *et al.*, 1997; Kroes *et al.*, 1999) to the specific conditions occurring in hydrophobic soils (Ritsema *et al.*, 2005). The SWAP model calculates evapotranspiration, groundwater level, crop growth, storage and transport of water, solute, and heat in soils. Within SWAP, soil water flow is calculated using the Darcy and Richards equations (*see* **Chapter 5, Fundamental Hydrologic Equations, Volume 1**). These equations are numerically solved using the finite difference method. Solute flow is computed using a numerical solution of the convection–dispersion equation. The traditional version of SWAP calculates uniform flow and transport in soils, and thus cannot be applied to hydrophobic soils.

The schematic representation of the preferential flow process as incorporated within the adapted semi-two-dimensional SWAP model by Ritsema *et al.* (2005) is shown in Figure 7. Basic assumptions and rules, which had to be used to make SWAP applicable also to hydrophobic soil conditions, are:

1. Flow switches dynamically from uniform to preferential flow in cases where an intermittent drying event reduces the water content in a layer of the soil profile to below the "critical soil water content". As a consequence, this layer will become hydrophobic, imposing a zero-flux condition at its upper boundary. This allows for the formation of a perched water table above the dry hydrophobic zone.
2. Saturated conditions above the hydrophobic zone may result in lateral flow (depending on layer anisotropy etc.) towards locations where vertical preferential flow paths might form.

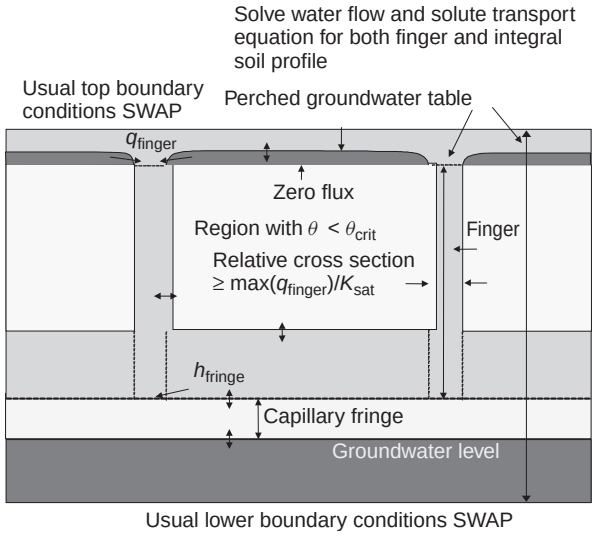


Figure 7 Schematic representation of the preferential flow concept occurring in hydrophobic soils and as incorporated in the modified SWAP model code adapted to hydrophobic conditions

3. The diameters of the preferential flow paths are calculated according to the relationship provided by Selker *et al.* (1996), showing a relationship between the texture of a soil and the flux towards the fingers divided by the saturated soil moisture content of a soil. The number of fingers will depend on the amount and duration of the rainfall. Simulated lateral flow quantities are added dynamically to the fingered flow zone.
4. With sufficiently high fluxes, soil water contents in the preferential flow paths will be near saturation.
5. Lateral diffusion from preferential flow paths towards surrounding dry soil will be slow, due to the extreme hysteresis in the soil water retention characteristic of the hydrophobic material. Hysteresis is accounted for in the adapted SWAP model according to the formulations proposed by Kool *et al.* (1987).
6. Preferential flow paths disappear dynamically either as a consequence of prolonged drying of the soil profile (soil water contents within the fingers drop below the critical level), after extremely wet periods causing many fingers to be formed in the profile, or after a considerable rise of the groundwater table.

Additional model input consists of, amongst others, values of critical soil water contents per layer and anisotropy characteristics of the surface layer. Model output consists of soil water content, pressure head, and solute concentration profiles for both the preferential flow domain and the surrounding immobile matrix.

When tested for a hydrophobic sandy field soil in the Netherlands, computed averages of the water contents within and outside the fingered flow regions obtained with the adapted SWAP code were found to be in much closer agreement with the observed field data than the results generated by the traditional model code. The traditional SWAP model overestimates average soil water contents for the summer period, that is, the period during which hydrophobicity occurred in the experimental field site. In reality, large parts of the soil are then bypassed by infiltrating water as a result of the occurrence of preferential flow paths, resulting in lower average soil water contents per depth, which is predicted by the adapted SWAP model. As a result of the preferential flow process within the experimental field, solutes move more rapidly in the downward direction compared with a situation without hydrophobicity and occurrence of preferential flow paths. This is illustrated in Figure 8, which indicates that arrival times of a conservative tracer (bromide) at 100-cm depth is not only much faster, but total solute loads are also higher compared with a comparable soil without occurrence of water repellency and preferential flow in the respective summer period. The modified SWAP model code has been further tested and applied for untreated and clay-amended hydrophobic soils in southwestern Australia, whereby also effects upon crop productivity have been investigated (Kramers *et al.*, 2005).

So far, the traditional and adapted SWAP model codes have been developed for application on plot and small-field scales, and approaches for modeling hydrophobicity effects at larger scales are yet to be developed. As the spatial and temporal variations in the occurrence, severity, and persistence of hydrophobicity become more difficult to establish with increasing spatial and temporal scales, the uncertainty

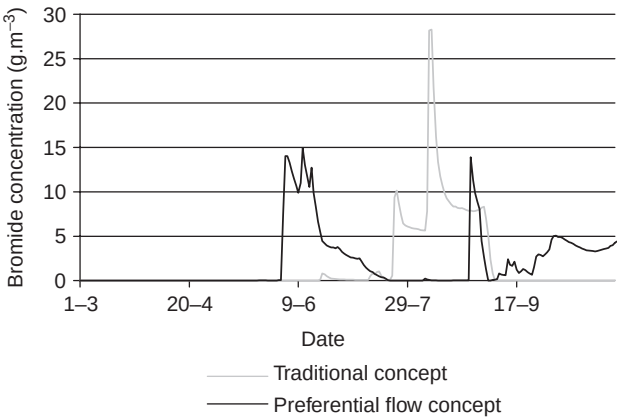


Figure 8 Simulated tracer (bromide) arrival at 100-cm depth for a sandy soil with three different layers after 50 mm of rainfall using the traditional (no preferential flow) and modified (including hydrophobicity-induced preferential flow) SWAP model codes with identical boundary conditions. Accounting for the occurrence of hydrophobicity and preferential flow (modified SWAP code) results in earlier arrival times of a conservative tracer and higher leaching totals

regarding the net-effects of hydrophobicity on soil water movement increases accordingly (Doerr and Moody, 2004). In addition, other hydrologically important factors such as the presence and spatial distribution of macropores (e.g. root channels, animal burrows) become increasingly important, affecting infiltration and water movement in the hydrophobic terrain (Burch *et al.*, 1989), and the variations in the storage capacity of the canopy, litter, and duff layers over larger scales will determine how much water will be delivered to the soil surface during a rainfall event, which in itself is likely to be spatially variable. These spatial and temporal uncertainties pose a considerable problem in evaluating and modeling the effects of hydrophobicity on soil water movement at larger scales (Doerr and Moody, 2004), but need to be tackled in order to upscale existing modeling approaches. Modeling at larger scales would then allow, for example, establishing to what degree hydrophobicity actually contributes to, what are perceived by many as "unusual" hydrological events such as the runoff and erosion events following the 2001 fires in the United States and Australia, or the 2000 and 2002 river floods in Europe.

Acknowledgments

We wish to thank all our colleagues in the field of soil hydrophobicity, whose work or personal communications have helped the preparation of this article. In particular, thanks go to Louis W. Dekker, John A. Moody, Klaas Oostindie, Richard A. Shakesby, and Rory P.D. Walsh for coauthorship on previous work relevant to this article. The preparation of this article was facilitated by a NERC Advanced Fellowship (NER/J/S/2002/00662; SHD).

FURTHER READING

Bauters T.W.J., Steenhuis T.S., DiCarlo D.A., Nieber J.L., Dekker L.W., Ritsema C.J., Parlange J.-Y. and Haverkamp R. (2000) Physics of water repellent soils. *Journal of Hydrology*, **231–232**, 233–243.

Doerr S.H., Shakesby R.A. and Walsh R.P.D. (2000) Soil water repellency: its characteristics, causes and hydrogeomorphological consequences. *Earth Science Reviews*, **5**, 33–65.

Ritsema C.J. and Dekker L.W. (Eds.) (2003) *Soil Water Repellency: Occurrence, Consequences and Amelioration*, Elsevier Science: Amsterdam, p. 352.

Van Dam J.C., Wösten J.H.M. and Nemes A. (1996) Unsaturated soil water movement in hysteretic and water repellent field soils. *Journal of Hydrology*, **184**, 153–173.

REFERENCES

Bachmann J., Horton R., Van der Ploeg R.R. and Woche S. (2000) Modified sessile drop method for assessing initial soil-water contact angle of sandy soil. *Soil Science Society of America Journal*, **64**, 564–567.

Bauters T.W.J., DiCarlo D.A., Steenhuis T.S. and Parlange J.-Y. (1998) Preferential flow in water repellent sands. *Soil Science Society of America Journal*, **62**, 1185–1190.

Bauters T.W.J., DiCarlo D.A., Steenhuis T.S. and Parlange J.-Y. (2000) Soil water content dependent wetting front characteristics in sands. *Journal of Hydrology*, **231–232**, 244–254.

Bayliss J.S. (1911) Observations on *Marasmius oreades* and *Clitocybe gigantea* as parasitic fungi causing "fairy rings". *Journal of Economic Biology*, **6**, 111–132.

Bisdom E.B.A., Dekker L.W. and Schoute J.F.T.h (1993) Water repellency of sieve fractions from sandy soils and relationships with organic material and soil structure. *Geoderma*, **56**, 105–118.

Blackwell P.S. (2000) Management of water repellency in Australia; and risks associated with preferential flow, pesticide concentration and leaching. *Journal of Hydrology*, **231–232**, 384–395.

Burch G.J., Moore I.D. and Burns J. (1989) Soil hydrophobic effects on infiltration and catchment runoff. *Hydrological Processes*, **3**, 211–222.

Campbell R.E., Baker M.B. Jr, Ffolliott P.F., Larson F.R. and Avery C.C. (1977) *Wildfire Effects on a Ponderosa Pine Ecosystem: An Arizona Case Study*, USDA Forest Service: Research Paper RM-191, p. 21.

Carrillo M.L.K., Letey J. and Yates S.R. (2000) Unstable water flow in a layered soil: I. effects of a stable water-repellent layer. *Soil Science Society of America Journal*, **64**, 450–455.

Cerdà A., Schnabel S., Ceballos A. and Gomez-Amelia D. (1998) Soil hydrological response under simulated rainfall in the Dehesa land systems (Extremadura, SW Spain) under drought conditions. *Earth Surface Processes and Landforms*, **23**, 195–209.

Clothier B.E., Vogeler I. and Magesan G.N. (2000) The breakdown of water repellency and solute transport through a hydrophobic soil. *Journal of Hydrology*, **231–232**, 255–264.

Crockford H., Topalidis S. and Richardson D.P. (1991) Water repellency in a dry sclerophyll eucalypt forest – Measurements and processes. *Hydrological Processes*, **5**, 405–420.

DeBano L.F. (1971) The effect of hydrophobic substances on water movement in soil during infiltration. *Soil Science Society of America Proceedings*, **35**, 340–343.

DeBano L.F. (1981) *Water Repellent Soils: A State-of-the-Art*, General Technical Report PS W-46, USDA Forest Service, Berkeley, p. 21.

DeBano L.F. (2000) Water repellency in soils: a historical overview. *Journal of Hydrology*, **231–232**, 4–32.

DeBano L.F., Savage S.M. and Hamilton D.A. (1976) The transfer of heat and hydrophobic substances during burning. *Soil Science Society of America Journal*, **40**, 779–782.

Dekker L.W., Doerr S.H., Oostindie K., Ziogas A.K. and Ritsema C.J. (2001) Water repellency and critical soil water content in a dune sand. *Soil Science Society of America Journal*, **65**, 1667–1674.

Dekker L.W. and Ritsema C.J. (1994) How water moves in a water repellent sandy soil. 1. Potential and actual water repellency. *Water Resources Research*, **30**, 2507–2517.

Dekker L.W. and Ritsema C.J. (1996) Variation in water content and wetting patterns in Dutch water repellent peaty clay and clayey peat soils. *Catena*, **28**, 89–105.

Dekker L.W., Ritsema C.J., Wendroth O., Jarvis N., Oostindie K., Pohl W., Larsson M. and Gaudet J.P. (1999) Moisture distributions and wetting rates of soils at experimental fields in The Netherlands, France, Sweden and Germany. *Journal of Hydrology*, **215**, 4–22.

De Rooij G.H. (1995) A three-region analytical model of solute leaching in a soil with a water repellent top layer. *Water Resources Research*, **31**, 2701–2707.

Doerr S.H. (1998) On standardizing the 'water drop penetration time' and the 'molarity of an ethanol droplet' techniques to classify soil hydrophobicity: a case study using medium textured soils. *Earth Surface Processes and Landforms*, **23**, 663–668.

Doerr S.H., Blake W.H., Shakesby R.A., Stagnitti F., Vuurens S.H., Humphreys G.S. and Wallbrink P. (2004) Heating effects on water repellency in Australian eucalypt forest soils and their value in estimating wildfire soil temperatures. *International Journal of Wildland Fire*, **13**(2), 157–163.

Doerr S.H., Ferreira A.J.D., Walsh R.P.D., Shakesby R.A., Leighton-Boyce G. and Coelho C.O.A. (2003) Soil water repellency as a potential parameter in rainfall-runoff modelling: experimental evidence at point to catchment scales from Portugal. *Hydrological Processes*, **17**, 363–377.

Doerr S.H., Llewellyn C.T., Douglas P., Morley C.P., Mainwaring K.A., Haskins C., Johnsey L., Ritsema C.J., Stagnitti F., Allinson G., Ferreira A.J.D., Keizer J.J., Ziogas A.K. and Diamantis J. (2005) Extraction of compounds associated with water repellency in sandy soils of different origin. *Australian Journal of Soil Research*, **43**(3), in press.

Doerr S.H. and Moody J. (2004) Hydrological impacts of soil water repellency: on spatial and temporal uncertainties. *Hydrological Processes*, **18**, 829–832.

Doerr S.H., Shakesby R.A. and Walsh R.P.D. (1998) Spatial variability of soil hydrophobicity in fire-prone eucalyptus and pine forests, Portugal. *Soil Science*, **163**, 313–324.

Doerr S.H., Shakesby R.A. and Walsh R.P.D. (2000a) Soil water repellency, its characteristics, causes and hydro-geomorphological consequences. *Earth Science Reviews*, **5**, 33–65.

Doerr S.H. and Thomas A.D. (2000) The role of soil moisture in controlling water repellency: new evidence from forest soils in Portugal. *Journal of Hydrology*, **231–232**, 134–147.

Doerr S.H., Walsh R.P.D. and Shakesby R.A. (2000b) Hydrophobicity in soils of the European Atlantic margin: preliminary observations and implications for soil hydrological response. *British Hydrological Society Occasional Paper*, **11**, 211–218.

Dyrness C.T. (1976) *Effect of Wildfire on Soil Wettability in the High Cascades of Oregon*, USDA Forest Service, p. 18, Research Paper PNW-202.

Feng G.L., Letey J. and Wu L. (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal*, **65**, 315–320.

Ferreira A.J.D., Coelho C.O.A., Walsh R.P.D., Shakesby R.A., Ceballos A. and Doerr S.H. (2000) Hydrological implications of soil water-repellency in *Eucalyptus globulus* forests, north-central Portugal. *Journal of Hydrology*, **231–232**, 165–177.

Franco C.M.M., Clarke P.J., Tate M.E. and Oades J.M. (2000) Hydrophobic properties and chemical characterization of natural water repellent materials in Australian sands. *Journal of Hydrology*, **231–232**, 47–58.

Garkaklis M.J., Bradley J.S. and Wooller R.D. (1998) The effects of Woylie (*Bettongia penicillata*) foraging on soil water repellency and water infiltration in heavy textured soils in southwestern Australia. *Australian Journal of Ecology*, **23**, 492–496.

Hallett P.D., Nunan N., Douglas J.T. and Young I.M. (2004) Millimetre-scale variability in soil water sorptivity: scale, surface elevation and subcritical repellency effects. *Soil Science Society of America Journal*, **68**, 352–358.

Hallett P.D. and Young I.M. (1999) Changes to water repellence of soil aggregates caused by substrate-induced microbial activity. *European Journal of Soil Science*, **50**, 35–40.

Hendrickx J.M.H., Dekker L.W. and Boersma O.H. (1993) Unstable wetting fronts in water repellent field soils. *Journal of Environmental Quality*, **22**, 109–118.

Horne D.J. and McIntosh J.C. (2000) Hydrophobic compounds in sands in New Zealand; extraction, characterization and proposed mechanisms for repellency expression. *Journal of Hydrology*, **231–232**, 35–46.

Imeson A.C., Verstraten J.M., van Mulligen E.J. and Sevink J. (1992) The effects of fire and water repellency on infiltration and runoff under Mediterranean type forest. *Catena*, **19**, 345–361.

Jamison V.C. (1946) Resistance to wetting in the surface of sandy soils under citrus trees in central Florida and its effect upon penetration and the efficiency of irrigation. *Soil Science Society of America Proceedings*, **11**, 103–109.

Jungerius P.D. and De Jong J.H. (1989) Variability of water repellence in the dunes along the Dutch coast. *Catena*, **16**, 491–497.

King P.M. (1981) Comparison of methods for measuring severity of water repellence of sandy soils and assessment of some factors that affect its measurement. *Australian Journal of Soil Research*, **19**, 275–285.

Kirkham M.B. and Clothier B.E. (2000) Infiltration into a New Zealand native forest soil. *A Spectrum of Achievement in Agronomy: Women Fellows of the Tri-Societies*, ASA Special Publication No. 62, ASA, pp. 13–26.

Kool J.B., Parker J.C. and van Genuchten M.T. (1987) Parameter estimation for unsaturated flow and transport models: a review. *Journal of Hydrology*, **91**, 255–293.

Kramers G., Van Dam J.C., Ritsema C.J., Stagnitti F., Oostindie K. and Dekker L.W. (2005) A new modelling approach to simulate preferential flow and transport in water repellent porous media: parameter sensitivity, and effects on crop growth and solute leaching. *Australian Journal of Soil Research*, (43), in press.

Kroes J.G., Van Dam J.C., Huygen J. and Vervoort R.W. (1999) *Uder's Guide of SWAP Version 2.0. Simulation of Water Flow, Solute Transport and Plant Growth in the Soil-Water-Atmosphere-Plant Environment*. Technical Document 48, Alterra Green World Research, Wageningen; Report 81,

Department of Water Resources, Wageningen University, p. 127.

Kung K.J.S. (1990) Preferential flow in a sandy vadose zone: field observations. *Geoderma*, **46**, 51–58.

Leighton-Boyce G., Doerr S.H., Shakesby R.A., Walsh R.P.D., Ferreira A.J.D., Boulet A.K. and Coelho C.O.A. (2005) Temporal dynamics of water repellency and soil moisture in eucalypt plantations, Portugal. *Australian Journal of Soil Research*, (43), in press.

Letey J., Carrillo M.L.K. and Pang X.P. (2000) Approaches to characterize the degree of water repellency. *Journal of Hydrology*, **231–232**, 61–65.

Letey J., Osborn J. and Pelishek R.E. (1962) The influence of the water-solid contact angle on water movement in soil. *Bulletin of the International Association of Hydrological Science*, **3**, 75–81.

Ma'shum M., Tate M.E., Jones G.P. and Oades J.M. (1988) Extraction and characterization of water-repellent materials from Australian soils. *Journal of Soil Science*, **39**, 99–110.

McGhie D.A. and Posner A.M. (1980) Water repellence of a heavy-textured Western Australian surface soil. *Australian Journal of Soil Research*, **18**, 309–323.

McGhie D.A. and Posner A.M. (1981) The effect of plant top material on the water repellence of fired sands and water repellent soils. *Australian Journal of Agricultural Research*, **32**, 609–620.

McNabb D.H., Gaweda F. and Fröhlich H.A. (1989) Infiltration, water repellency, and soil moisture content after broadcast burning a forest site in southwest Oregon. *Journal of Soil and Water Conservation*, **44**, 87–90.

Meeuwig R.O. (1971) *Infiltration and Water Repellency in Granitic Soils*, USDA Forest Service: Research Paper INT-111, p. 20.

Morley C.P., Douglas P., Doerr S.H., Mainwaring K., Llewellyn C.T. and Dekker L.W. (2005) Identification of hydrophobic compounds in a sandy soil under permanent grass cover. *Australian Journal of Soil Research*, (43), in press.

Nieber J.L. (1996) Modeling finger development and persistence in initially dry porous media. *Geoderma*, **70**, 207–229.

Parker S.D. (1987) *Encyclopedia of Science and Technology*, McGraw-Hill: New York.

Pradas M., Imeson A.C. and Van Mulligen E. (1994) Comparing the infiltration and runoff characteristics of burnt soils in NE-Catalonia. *Abstracts of the European Society for Soil Conservation Conference on Soil Erosion and Degradation as a Consequence of Forest Fires*, Sala M. and Rubio J.L. (Eds.), Barcelona and Valencia, pp. 24–25.

Ritsema C.J. and Dekker L.W. (1994a) Soil moisture and dry bulk density patterns in bare dune sands. *Journal of Hydrology*, **154**, 107–131.

Ritsema C.J. and Dekker L.W. (1994b) How water moves in a water repellent sandy soil. 2. Dynamics of fingered flow. *Water Resources Research*, **30**, 2519–2531.

Ritsema C.J. and Dekker L.W. (1995) Distribution flow: A general process in the top layer of water repellent soils. *Water Resources Research*, **31**, 1187–1200.

Ritsema C.J., Dekker L.W. and Heijs A.W.J. (1997) Three-dimensional fingered flow patterns in a water repellent sandy field soil. *Soil Science*, **162**, 79–90.

Ritsema C.J., Dekker L.W., Hendrickx J.M.H. and Hamminga W. (1993) Preferential flow mechanism in a water repellent sandy soil. *Water Resources Research*, **29**, 2183–2193.

Ritsema C.J., Dekker L.W., Nieber J.L. and Steenhuis T.S. (1998) Modeling and field evidence of finger formation and finger recurrence in a water repellent sandy soil. *Water Resources Research*, **34**, 555–567.

Ritsema C.J., Van Dam J.C., Oostindie K. and Dekker L.W. (2005) A new modelling approach to simulate preferential flow and transport in water repellent porous media: model structure, input, output and validation. *Australian Journal of Soil Research*, (43), in press.

Roberts F.J. and Carbon B.A. (1971) Water repellence in sandy soils of southwestern Australia: I. Some studies related to field occurrence. *Field Station Record Division Plant Industry CSIRO (Australia)*, **10**, 13–20.

Roy J.L. and McGill W.B. (2000) Flexible conformation in organic matter coatings: an hypothesis about soil water repellency. *Canadian Journal of Soil Science*, **80**, 143–152.

Roy J.L. and McGill W.B. (2002) Assessing soil water repellency using the molarity of ethanol droplet (MED) test. *Soil Science*, **167**, 83–97.

Roy J.L., McGill W.B. and Rawluk M.D. (1999) Petroleum residues as water-repellent substances in weathered nonwettable oil-contaminated soils. *Canadian Journal of Soil Science*, **79**, 367–380.

Savage S.M., Martin J.P. and Letey J. (1969) Contribution of some soil fungi to natural and heat-induced water repellency in sand. *Soil Science Society of America Proceedings*, **33**, 405–409.

Schreiner O. and Shorey E.C. (1910) Chemical nature of soil organic matter. *USDA Bureau Soils Bulletin*, **74**, 2–48.

Scott D.F. (1993) The hydrological effects of fire in South African mountain catchments. *Journal of Hydrology*, **150**, 409–432.

Scott D.F. and Lesch W. (1997) Streamflow responses to afforestation with *Eucalyptus grandis* and *Pinus patula* and to felling in the Mokubulaan experimental catchments, South Africa. *Journal of Hydrology*, **199**, 360–377.

Scott D.F. and Van Wyk D.B. (1990) The effects of wildfire on soil wettability and hydrological behaviour of an afforested catchment. *Journal of Hydrology*, **121**, 239–256.

Selker J.S., Steenhuis T.S. and Parlange J.-Y. (1996) An engineering approach to fingered vadose pollutant transport. *Geoderma*, **70**, 197–206.

Sevink J., Imeson A.C. and Verstraten J.M. (1989) Humus form development and hillslope runoff and the effects of fire and management, under Mediterranean forest in NE Spain. *Catena*, **16**, 461–475.

Shakesby R.A., Boakes D.J., Coelho C.deO.A., Concalves A.J.B. and Walsh R.P.D. (1996) Limiting the soil degradational impacts of wildfire in pine and eucalyptus forests, Portugal: comparison of alternative post-fire management practices. *Applied Geography*, **16**, 337–355.

Shakesby R.A., Doerr S.H. and Walsh R.P.D. (2000) The erosional impact of soil hydrophobicity: current problems and future research directions. *Journal of Hydrology*, **231–232**, 178–191.

Steenhuis T.S., Boll J., Shalit G., Selker J.S. and Merwin I.A. (1994) A simple equation for predicting preferential flow solute concentrations. *Journal of Environmental Quality*, **23**, 1058–1064.

Tschapek M. (1984) Criteria for determining the hydrophilicity-hydrophobicity of soils. *Zeitschrift für Pflanzenernährung und Bodenkunde*, **147**, 137–149.

Van Dam J.C., Hendrickx J.M.H., Van Ommen H.C., Bannink M.H., Van Genuchten M.T.h and Dekker L.W. (1990) Water and solute movement in a coarse-textured water-repellent field soil. *Journal of Hydrology*, **120**, 359–379.

Van Dam J.C., Huygen J., Wesseling J.G., Feddes R.A., Kabat P., Van Walsum P.E.V., Groenendijk P. and Van Diepen C.A. (1997) *Theory of SWAP Version 2.0. Simulation of Water Flow, Solute Transport and Plant Growth in the Soil-Water-Atmosphere-Plant Environment*, Report 71, Department of Water Resources, Wageningen University; Technical Document 45, Alterra Green World Research, Wageningen, p. 167.

Van den Bosch H., Ritsema C.J., Boesten J.J.T.I., Dekker L.W. and Hamminga W. (1999) Simulation of water flow and bromide transport in a water repellent sandy soil using a one-dimensional convection-dispersion model. *Journal of Hydrology*, **215**, 172–187.

Van't Woudt B.D. (1959) Particle coatings affecting the wettability of soils. *Journal of Geophysical Research*, **64**, 263–267.

Wallis M.G. and Horne D.J. (1992) Soil water repellency. In *Advances in Soil Science*, Vol. 20, Stewart B.A. (Ed.), Springer: New York, pp. 91–146.

Wallis M.G., Scotter D.R. and Horne D.J. (1991) An evaluation of the intrinsic sorptivity water repellency index on a range of New Zealand soils. *Australian Journal of Soil Research*, **29**, 353–362.

Walsh R.P.D., Boakes D., Coelho C.deO.A., Goncalves A.J.B., Shakesby R.A. and Thomas A.D. (1994) Impact of fire-induced hydrophobicity and post-fire forest litter on overland flow in northern and central Portugal. *2nd International Conference on Forest Fire Research*, Vol. II, Coimbra, Domingos Xavier Viegas, p. 1149–1159, November 21–24, 1994.

Walsh R.P.D., Coelho C.deO.A., Elmes A., Ferreira A.J.D., Goncalves A.J.B., Shakesby R.A., Ternan J.L. and Williams A.G. (1998) Rainfall simulation plot experiments as a tool in overland flow and soil erosion assessment, north-central Portugal. *GeoÖkoDynamik*, **19**, 139–152.

Walsh R.P.D., Coelho C.deO.A., Shakesby R.A., Ferreira A.D.J. and Thomas A.D. (1995) Post-fire land use and management and runoff responses to rainstorms in northern Portugal. In *Geomorphology and Land Management in a Changing Environment*, McGregor D. and Thompson D. (Eds.), John Wiley & Sons: Chichester, pp. 283–308.

Wang Z., Wu L. and Wu Q.J. (2000) Water-entry value as an alternative indicator of soil water repellency and wettability. *Journal of Hydrology*, **231–232**, 76–83.

White W.D. and Wells S.G. (1982) Forest-fire devegetation and drainage basin adjustments in mountainous terrain. In *Adjustments of the Fluvial System*, *Proceedings of the Tenth Annual Geomorphology Symposia Series*, Rhodes D.D. and Williams G.P. (Eds.), Kendall/Hunt Publishing Company: Binghamton, pp. 199–223, September 21–22, 1979.

Wienhold B.J. and Gish T.J. (1991) Chemical factors contributing to the formation of preferential pathways. In *Proceedings of a National Symposium on Preferential Flow*, Gish T. and Shirmohammadi A. (Eds.), ASAE St. Joseph: Chicago, pp. 100–103, December 16–17, 1991.

Witter J.V., Jungerius P.D. and Ten Harkel M.J. (1991) Modeling water erosion and the impact of water repellency. *Catena*, **18**, 115–124.

York C.A. (1993) A questionnaire survey of dry patch on golf courses in the United Kingdom. *Journal of Sports Turf Research*, **69**, 20–26.

Zierholz C., Hairsine P. and Booker F. (1995) Runoff and soil erosion in bushland, following the Sydney bushfires. *Australian Journal of Soil and Water Conservation*, **8**, 28–37.

Zisman W.A. (1964) Relation of equilibrium contact angle to liquid and solid constitution. In *Contact Angle, Wettability and Adhesion*, *Advances in Chemistry Series, Vol. 43*, Gould R.F. (Ed.), American Chemical Society: Washington, pp. 1–51.

69: Solute Transport in Soil at the Core and Field Scale

MARNIK VANCLOOSTER,[1] MATHIEU JAVAUX[1,2] AND J VANDERBORGHT[2]

[1]Department of Environmental Sciences and Land Use Planning, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[2]Agrosphere Institute, (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

A detailed understanding of the flow and transport processes of chemicals in soils is needed to implement effective and efficient soil and water management strategies and to meet the current challenges of sustainable development. The solute transport process is a key process determining the mass flow of chemical substances once they are released in the soil solution. The transport process is a highly nonlinear and space-time dynamic process for which models are still poorly validated at the larger scale. In this article, we review some recent advances in describing solute transport in soil at the macroscopic core scale, corresponding to the scale of most soil measurement devices, and the pedon scale, corresponding to the scale of a small field. After introducing basic concepts, major problems using the classical Convection Dispersion Equation (CDE) model for describing field-scale solute transport is presented and alternative modeling concepts are introduced.

INTRODUCTION

Soil and groundwater are important natural resources that should be protected from human-induced pollution, requiring a thorough understanding of the fate and transport mechanisms of solutes in soils. Dissolved substances entering the soil from diffuse and point pollution sources, further referred to as solutes, are carried with the water phase through the soil zone, where it is subjected to a range of phase exchange (e.g. sorption) and transformation processes. Along with the boundary conditions at the soil surface, these transport, phase exchange, and transformation processes will determine the ultimate concentrations and fluxes of dissolved substances in soil and groundwater. The modeling and characterization of solute fate and transport are essential for efficient soil and groundwater management and engineering.

The modeling and characterization of solute fate and transport in soils is complicated by the space-time variability of the underlying processes. In addition, any experimental technique is operational at a certain scale, which is not necessarily the scale at which the process can reasonably be described, neither the scale at which a prediction is needed. The expressions of the solute fate and transport are therefore often considered as scale-dependent, and scaling is needed to model and characterize the solute transport processes at the larger spatial scale. The scaling is of utmost importance given the fact that most environmental problems occur at the larger spatiotemporal scales, while most characterization techniques and well-validated modeling concepts are only operational at the lower scale.

This article gives a summary of the processes affecting solute transport at the local core and field-scale and introduces approaches for modeling. The local scale is considered as the laboratory scale, pertaining to the scale where most classical experimental techniques are operational. It is typically the scale of the soil core or the soil lysimeter (Dagan, 1989). The field-scale considered here is the one defined from a pedological point of view, that is, a three-dimensional soil body large enough to permit the study of solute transport for practical purposes. It often corresponds to a poly-pedon and is considered as a homogeneous soil unit in soil maps. After reviewing the basic transport processes that are embedded in classical transport theory, shortcomings of classical modeling approaches for modeling solute transport at the field-scale are illustrated. In

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the last section, alternative approaches are presented which allow a more comprehensive modeling of solute transport in heterogeneous soil.

SOLUTE TRANSPORT: DEFINITIONS, CONCEPTS, AND CLASSICAL MODELLING APPROACHES

Solute, Concentration

Chemical substances can be present in different phases of the soil. The resident chemical concentration [$M L^{-3}$] is defined as the mass storage of the chemical substance per unit volume. The volume averaged total resident concentration C_t^r is the volume-weighed sum of the resident concentrations in the separate phases:

$$C_t^r = \theta C_l^r + C_a^r + \alpha C_g^r + \eta C_n^r \quad (1)$$

where θ [L^3 L^{-3}] is the volume average water content, η is the nonaqueous liquid volume fraction [L^3 L^{-3}], α is the volumetric air content, the subscripts l, a, g, and n define the dissolved, adsorbed, gaseous, and nonaqueous phases respectively. Usually, C_a^r is defined as the product of the soil dry bulk density by the mass density of adsorbed chemicals (mass of chemicals by mass of soil). For inert nonvolatile substances in a nonsaturated soil, composed of only one single liquid phase (i.e. the soil water), the resident fluid concentration of the chemical substance will therefore be $C_l^r = C_t^r/\theta$.

We consider a solute, a chemical substance that dissolves in the liquid soil water phase. The substances may be more or less ionized according to their ionic charge, but may also be present as electrically neutral molecules or complexes of different molecules or ions (de Marsily, 1986). The fate and transport processes of solutes will be influenced by interface phenomena between the soil liquid, solid, and gaseous phases, and the transport phenomena in these phases. However, in this paper, we will focus on "inert" solutes that do not partition between different phases but remain in the water phase. For mobile solutes, flux averaged concentrations are defined as the ratio of solute flux to soil liquid flux through a given area, that is,

$$C_l^f = \frac{J_s}{J_w} \quad (2)$$

where J_s is the area-averaged solute flux vector [$M L^{-2}$ T^{-1}], and J_w is the area-averaged soil water flux vector [$L T^{-1}$]. When water flow in the unsaturated zone is predominantly in the vertical direction and when solutes are applied uniformly over a wide area, the horizontal components of the water flow and solute flux vectors can be neglected. Since this often applies to leaching experiments and to practical applications dealing with diffuse pollution, transport in soils is often treated as a one-dimensional problem. Using a 1-D mass balance equation, both concentration definitions are related through:

$$\frac{\partial C_t^r}{\partial t} + \frac{\partial J_w C_l^f}{\partial z} = 0 \quad (3)$$

Liquid-phase Transport Processes

Convection/Advection

This is the phenomenon where the solutes move with the water. The convective solute flux, $J_{s,c}$ [$M L^{-2}$ T^{-1}] is given by:

$$J_{s,c} = J_w C_l^r \quad (4)$$

where J_w is the area-averaged Darcian water flux vector.

The advance of a solute plume through a soil is characterized by an average velocity, which, in case of inert solute, equals the averaged pore water velocity $\overline{v}$:

$$\overline{v} = \frac{J_w}{\theta_{eff}} \quad (5)$$

where θ_{eff} [L^3 L^{-3}] the effective transport volume. When all the water is accessible for the solutes, θ_{eff} equals the soil volumetric moisture content θ. However, because of the electric repulsion of ions by an oppositely charged solid phase, very slow diffusive mixing between mobile and stagnant pore water, and exclusion of larger molecules from small pores, θ_{eff} may be much smaller than θ.

Diffusion and Dispersion

When solute plume migrates in a soil, it will disperse essentially because of two mechanisms: molecular diffusion and hydrodynamic dispersion. Molecular diffusion is a microscopic process, induced by thermal agitation and molecular collisions, referred to as *Brownian motion*. Brownian particle displacement of a solute added to a stagnant liquid phase leads to a Gaussian distribution of particle locations that results from a large number of independent and zero mean particle displacements (Einstein, 1905). Because of the independence of the particle displacements, the variance of the particle location distribution is proportional to the time: $\sigma_x^2 = 2.D_0.t$, where t [T] is the time and D_0 [L^2 T^{-1}] is the Brownian diffusion coefficient. For sufficiently low concentrations, Brownian motion in a stagnant liquid leads to a diffusive flux J_{dif} [$M L^{-2}$ T^{-1}] that dissipates concentration gradients and can be modeled by Fick's law:

$$J_{dif} = -D_0.\frac{\partial C_l^r}{\partial z} \quad (6)$$

A similar process will occur in a static fluid in a porous medium. The diffusion process in this case will be hindered

by the presence of the solid phase. The cross-sectional area across which diffusion can take place is reduced by a factor that is equal to the volumetric water content, θ. The tortuosity of the pores increases the microscopic distance across which diffusive transport must take place to dissipate macroscopic concentration gradients and reduces the diffusive flux by a tortuosity factor ξ:

$$J_{\text{dif}} = -D_0.\xi.\theta.\frac{\partial C_l^r}{\partial z} \tag{7}$$

Because the tortuosity of the water-filled pore space increases with decreasing soil water content, ξ is a decreasing function of θ.

When water flows through the porous medium, the deviation of the local-scale advection velocities around the mean advection velocity $\overline{v}$ induces hydrodynamic dispersion. At the pore scale, streamline density increases from the pore wall to the pore center. For real and irregularly shaped soil pores, however, this within-pore variability will be more pronounced, as streamline densities change between pore bodies and pore necks. The heterogeneous pore-size distribution with larger flow velocities in larger pores leads to an important additional variability of velocities.

The advection velocity variability generates additional "advective particle displacements", relative to the mean particle displacement $\overline{v}$. Analogous to Brownian motion, the effect of a large number of independent "advective particle displacements" on solute transport can be described for the 1-D case as a Fickian dispersive process:

$$J_{\text{disp}} = -\theta D.\frac{\partial C_l^r}{\partial z} \tag{8}$$

where D [L^2 T^{-1}] is the hydrodynamic dispersion coefficient. Crucial for applying equation (8) is the independence of advective particle displacements. This implies that a solute particle samples a representative set of the advection velocity distribution on its trajectory. The time needed for a particle to sample a representative set of velocities at the macroscopic scale is the characteristic mixing time τ^*. Although the hydrodynamic dispersion has an analogous effect on solute transport as diffusion, that is, it tends to spread out the concentration differences, hydrodynamic dispersion is strongly influenced by the advective velocity $\overline{v}$. The exact relationships between D and $\overline{v}$ can only be obtained from theoretical considerations for simple or hypothetical geometrical pore systems. For small values of $\overline{v}$, chemical diffusion is the dominant process, and D is independent of $\overline{v}$ and reaches a value slightly lower than D_0 (because of matrix tortuosity effects on the diffusion process). In Taylor's dispersion theory for solutes in a cylindrical tube with a laminar flow field, $D = r^2.\overline{v}^2/48.D_0$ (Taylor, 1953), where r [L] is the radius of the tube. In this case, mixing is caused by lateral diffusion so that τ^* does not change with $\overline{v}$. According to the Hagen–Poiseuille relationship, the variance of velocities in a capillar with laminar flow increases quadratically with $\overline{v}$, and so does D. This regime applies for lower flow rates when the mixing caused by local diffusion is faster than the mixing caused by advection of particles into regions with a different particle velocity. For sufficiently high $\overline{v}$, advective mixing due to divergence and convergence of streamlines becomes the dominant mixing process. When the geometry of the water-filled pore space remains constant with $\overline{v}$, that is, in saturated media, and mixing is dominated by advection, the mixing time τ^* is inverse proportional to $\overline{v}$, whereas the variance of velocities increases quadratically with $\overline{v}$. As a consequence, D increases linearly with $\overline{v}$: $D = \lambda.\overline{v}$, with λ [L] being a soil characteristic parameter called hydrodynamic dispersivity. Saffman (1959) derived the theoretical relationship between D and $\overline{v}$ for saturated networks of pores. At the larger scale, for example, the scale of an aquifer, the variance of advection velocities is determined by the spatial variability of the hydraulic conductivity. The relation between D and $\overline{v}$ at the larger scale depends, similarly to the relation at the pore scale, on the dominant mixing process: diffusion or advection.

Formalisms for Modeling Solute Transport

Differential Description of Solute Transport

Averaging the pore-scale transport process over a given soil volume, and assigning the average properties to the centroid of the voxel, results in continuous functions in space of the hydrodynamic properties and state variables. In this continuum approach, differential calculus can be applied to establish mass and momentum balance equations for infinitesimal small soil volume and time increments. For the case of inert solute transport in a macroscopic homogeneous soil, the general continuity equation applies:

$$\frac{\partial C_t^r}{\partial t} = -\frac{\partial Js}{\partial z} \tag{9}$$

Considering in this case that the solutes only resides in the liquid phase (i.e. $C_t^r = \theta.C_l^r$) and decomposing the total solute flux in a convective and dispersive component yields the classical CDE:

$$\theta\frac{\partial C_l^r}{\partial t} + \overline{v}.\frac{\partial C_l^r}{\partial z} - \frac{\partial}{\partial z}\left(\theta D\frac{\partial C_l^r}{\partial z}\right) = 0 \tag{10}$$

This linear partial differential equation is a complete mathematical description of the dynamics of the inert solute transport process over an infinitesimal small space and time increment. The differential equation needs to be integrated over space and time to describe the transport process at integral soil and time scale. Numerous analytical and numerical techniques have been developed to integrate

equation (10) for variable boundary conditions. For an overview, the reader is referred to **Chapter 78, Models of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2**.

The classical CDE equation has been intensively used to model solute transport in saturated porous media (Dagan, 1989). In partially saturated soils, the CDE model is usually considered to be more appropriate to describe transport in repacked or nonstructured soil (Radcliffe *et al.*, 1998; Wilson *et al.*, 1998), but has also successfully been used to describe transport in real structured soils (Bejat *et al.*, 2000; Comegna *et al.*, 1999).

Integral Description of Solute Transport: Transfer Function Theory

In contrast to the differential approach, linear transfer function theory can be used to formulate mathematically the transport process at the integral scale of the considered soil using an integral property: the impulse response function (Jury, 1982; Jury and Roth, 1990). The transfer function formalism is attractive because it allows the convenient description of transport experiments, in particular, column breakthrough experiments. In addition, any linear transport process can be expressed as a transfer function and is a physically equivalent mathematical formalism of the same process. Finally, applying the transfer function directly to integral soil problems often results in effective integral parameterizations, avoiding thereby the parameterization problems often associated with differential continuum modeling approaches.

Using the transfer function theory, the output solute flux leaching from a given integral soil volume with a given transport distance is calculated using a convolution between a soil characteristic impulse response function $f^f(t;z)$ and the time series of the input solute flux:

$$J_s(z,t) = \int_0^t J_s(0,\tau) f^f(t-\tau;z)\,d\tau \qquad (11)$$

where $J_s(z,t)$ is the output solute flux, $J_s(0,\tau)$ the input solute flux, and τ the solute input time. This equation is submitted to three important assumptions (i) the transfer function f^f is measurable, (ii) the transport process is linear and, (iii) the system is stationary, that is, f^f does not depend on the time of the input pulse. For a steady-state water flux, and considering the definition of the solute flux concentration, equations (2) and (11) yield:

$$C_1^f(z,t) = \int_0^t C_1^f(0,\tau) f^f(t-\tau;z)\,d\tau \qquad (12)$$

For steady-state flow, the pulse input response function, $f^f(t;z)$ can be interpreted as a travel time distribution of solute particles that are added at the soil surface to a given depth z in the soil profile.

In equations (11) and (12), a temporal pulse input response function is defined, which links the temporal reaction of the system, that is, the solute flux at a certain depth, to a temporal input function, that is, the solute flux at the soil surface. In analogy, a spatial pulse input response function which links an initial solute mass distribution in the soil profile to the solute mass distribution at a later time can be defined. For steady-state flow and a soil profile with uniform water content, the concentration distribution can be calculated from:

$$C_1^r(z,t) = \int_0^\infty C_1^r(z',0) f^r(z;z';t)\,dz' \qquad (13)$$

where $f^r(z;z';t)$ is the spatial pulse input response function at time t for a concentration Dirac pulse at a depth z' in the soil profile. It should be noted that unlike the temporal pulse input function, the spatial pulse input response function is not a stationary function with depth when the boundary condition at the soil surface influences transport within the soil profile. Nonstationarity in this context refers to the dependence of the spatial pulse input response function on the depth z' where the Dirac pulse is applied.

The drawback of the transfer function model (TFM) representation is that it is not easy to apply to transient flow conditions (Butters and Jury, 1989; Roth *et al.*, 1991). In addition, to reduce errors in the convolution, a good temporal (or spatial) resolution is needed.

LIMITATIONS OF THE CONVECTIVE DISPERSIVE TRANSPORT MODEL

The CDE process model is probably the most popular solute transport model that has been used in soil hydrological research so far. The model considers the soil as a homogeneous matrix in which the pores are well connected, and in which solutes mix perfectly laterally when traveling through the soil. The differential description of the process is given in equation (10) and simplifies for a homogeneous soil profile and steady-state conditions to:

$$\frac{\partial C_1^r}{\partial t} + \bar{v}\frac{\partial C_1^r}{\partial z} - D\frac{\partial^2 C_1^r}{\partial z^2} = 0 \qquad (14)$$

Using transfer function theory, the CDE process in a semiinfinite soil column can be captured as follows:

$$f^f(t;z) = \frac{z}{2\sqrt{\pi . D . t^3}} \exp\left(-\frac{(z-\bar{v}t)^2}{4\,Dt}\right) \qquad (15)$$

A characteristic of the CDE travel time distribution is that the variance of the travel times grows linearly with travel distance z. This is equivalent to the particle location

distribution, which grows linearly with time for a Brownian motion process.

Essential in the derivation of equations(14) and (15) is the consideration that the hydrodynamic dispersion can be described as a diffusion process, that is, on average, all solute particles are subjected to the same forces and the transport time is sufficiently large so that the incremental microscopic particle displacements are no longer statistically correlated. As a corollary, the CDE process cannot be valid for small soil volumes where the travel times are too small as compared to the mixing time, or to describe transport close to interfaces. A comparison between the particle travel time and the mixing time during a transport experiment forms the basis for discrimination between different "transport processes" or "mixing regimes" (Simmons, 1982). If the "mixing" time τ^* is much smaller than the solute travel time, the dispersion of a solute plume can be described as a Fickian concentration gradient-driven process (Dagan, 1989).

Limitations of the CDE in Unsaturated Soils

For stationary flow conditions, D and $\overline{v}$ are independent parameters describing the transport process. In transient conditions, however, the relationship between D and $\overline{v}$ must be taken into account. Experimental evidences show that for transport in homogeneous saturated porous media, D is a monotonous function of $\overline{v}$. In unsaturated media, this relation becomes extremely complicated since the transport volume θ_{eff} changes with the water flux. Therefore, the structure of the water-filled pore space and, hence, the flow field depends on the saturation degree (Flury *et al.*, 1994) so that the variance of local velocities and the mixing time cannot be simply related to the mean advection velocity. As a consequence, no validated theoretical models exist to calculate the relationship between D and $\overline{v}$ for unsaturated soils and the dispersivity λ cannot be considered to be a material constant, that is, independent of θ.

It has been observed that the variance of local advection velocities increase dramatically with increasing flow rate, especially when macropores are activated (e.g. White, 1985; Dyson and White, 1989; Bouma, 1991). On the other hand, a decrease in water saturation may result in larger tortuosity of the solute trajectories, a disconnection of continuous flow paths and a physical nonequilibrium due to a slow diffusive exchange of solutes between mobile and immobile pore regions. Therefore, in some experimental studies (e.g. Corey *et al.*, 1963; De Smedt *et al.*, 1986; Maraqa *et al.*, 1997; Padilla *et al.*, 1999), a higher solute dispersion was found for unsaturated than for saturated flow conditions.

Limitations of the CDE in Structured Media

Flow processes affected by the heterogeneous structural properties of real soils prevailing at different spatial scales contribute to an incomplete mixing of the solutes during the transport in real soils. At the microscopic scale, structural heterogeneity of the pore liquid system is characterized by the complex geometry of the pore liquid phase and irregular distribution of the carrying liquid water phase in the soil. The heterogeneity of the soil pore systems is expressed by the presence of microaggregates, aggregates, and pedons in real soil horizons. Therefore, under such conditions, CDE fails to describe the observed dispersion (e.g. Butters and Jury, 1989; Khan and Jury, 1990). Evidence of non-CDE behavior is obtained when the variance of the solute travel time does not grow linearly with depth, or, similarly when the variance of the solute travel distance does not grow linearly with time.

Scale-variable dispersivity was observed in saturated porous media at the laboratory and the field-scale, as well in homogeneous as in heterogeneous formations (Pickens and Grisak, 1981; Huang *et al.*, 1995; Vanclooster *et al.*, 1995; Pang and Hunt, 2001). Studying solute transport through two undisturbed Spodosols monolithes (0.8 m i.d., 1 m deep), Seuntjens *et al.* (2001) obtained opposite depth scaling behaviors of the dispersivity and related this to the different morphological characteristics of the two pedons.

In the context of transport, the presence of macropores plays a particular role. Macropores are large pores, which form at the macroscopic level an obviously distinguished pore system from the soil matrix pore system. Macropores constitute sometimes a separate and/or a continuous network in which particle velocities may deviate systematically from those in the soil matrix. As a result, the solutes released in the macropore network will be subjected to a preferential flow as compared to flow in the matrix system and will not completely mix with the total pore water volume at short time intervals. Preferential flow through macropores is considered here as a macroscopic process since concentrations in the macropores cannot be easily determined separately from the concentrations in the micropore system.

Besides preferential flow through macropores, preferential flow may as well occur through certain parts of the soil matrix. Fingered flow, for example, will be caused by wetting front instability (De Rooij, 2000) by air entrapment in the matrix (Peck, 1964), by water repellency of the solid phase (Ritsema and Dekker, 1995; Ritsema *et al.*, 1998), or by the increase of the soil hydraulic conductivity with depth (Raats, 1973). Funneled flow will typically develop along the inclined bottom of a fine layer overlaying a coarse sublayer (Kung, 1990; Poletika and Jury, 1994; Quisenberry *et al.*, 1994).

At the field-scale, the soil hydraulic properties such as the soil hydraulic conductivity and water retention curve are variable in space (Jury, 1985; Mallants *et al.*, 1996a). The macroscopic variability of the flow and transport properties at the field-scale will be expressed in a variable flow field,

thereby deviating locally the direction of the macroscopic particle velocity vectors. The mixing of solutes from, for example, low to high conductive zones at these larger scales will only be fully established if again the process is evaluated at sufficiently large transport volumes or transport times (Flühler *et al.*, 1996). A full characterization of the dispersion process at the field-scale would therefore need a full description of the variability of the flow properties. This, however, remains a complicated task, which for the time being is only realizable in a research oriented context.

Van Wesenbeeck and Kachanoski (1995), Kasteel (1997), and Vanderborght *et al.* (1997b) predicted solute dispersion observed in field-scale leaching experiments from the spatial variability of measured soil hydraulic properties. Vanderborght *et al.* (1997b) concluded that the mixing regime could be fairly well reproduced for smaller flow rates if small correlation scales of the hydraulic properties were assumed. For higher flow rates, the dispersion was largely underestimated owing to a large variability of particle velocities at the pore scale when macropores were activated. On the basis of a summary of leaching experiments that were carried out in a range of soils at different scales and for different leaching rates (Vanclooster *et al.*, 1995; Vanderborght *et al.*, 1997a; Jacques *et al.*, 1998, Seuntjens *et al.*, 2001, Vanderborght *et al.*, 2000), Vanderborght *et al.* (2001) linked the solute transport regime to morphological and hydraulic properties. The change of dispersivity with leaching rate was linked to the unsaturated hydraulic conductivity using a multidomain conceptualization of the pore space as was introduced by Steenhuis *et al.* (1990). The mixing regime was further related to soil morphological features, such as vertical tongues, stratification, macropores, and a water repellent layer. In all investigated soils, the hydrodynamic dispersion increased more than linearly with increasing leaching rate confirming that the dispersivity is not an intrinsic soil characteristic for unsaturated flow conditions at this scale of observation.

TRANSPORT MODELS IN HETEROGENEOUS SOILS

Modeling Approach Versus Structural Variability

Given the limitation of the CDE flow concept, alternative transport models have been developed for describing solute transport in real field soils. These alternative models consider the variability of the process due to structural variation in the flow domain. The models can be ranked depending on the degree of explicitness with which the structural variability is considered in the model (see Figure 1).

In the first class of models, the structural variability is embedded in a completely implicit way in the transport models. As a consequence, the parameters of these models need to be derived from model fits to experimental data from tracer experiments. In this class, two asymptotic transport models, which cover the end-points of mixing in the soil, that is, the complete mixing versus no-mixing model, can be grouped (Sposito and Jury, 1988; Flühler *et al.*, 1996). The complete mixing model, that is, the CDE model, implicitly assumes that the spatial scale of the velocity variability or the mixing scale is much smaller than the scale of the transport process. In the opposite no-mixing model, that is, the stream-tube model, it is assumed that the mixing scale is much larger than the transport scale. Also, "generalized" stream-tube models (STM) (e.g., Zhang (2000)), which were introduced to describe intermediate mixing regimes, fall in this class of implicit transport models. In the generalized STMs, the intermediate mixing regime is accounted for by an additional fitting parameter.

In the second class of models, multidomain models, the structure of the flow field is embedded in the model by dividing the porous medium in two (e.g. van Genuchten and Wierenga, 1976; Jarvis *et al.*, 1991a; Chen and Wagenet, 1992; Gerke and van Genuchten, 1993) or more (Steenhuis *et al.*, 1990; Skopp and Gardner, 1992; Gwo *et al.*, 1995;

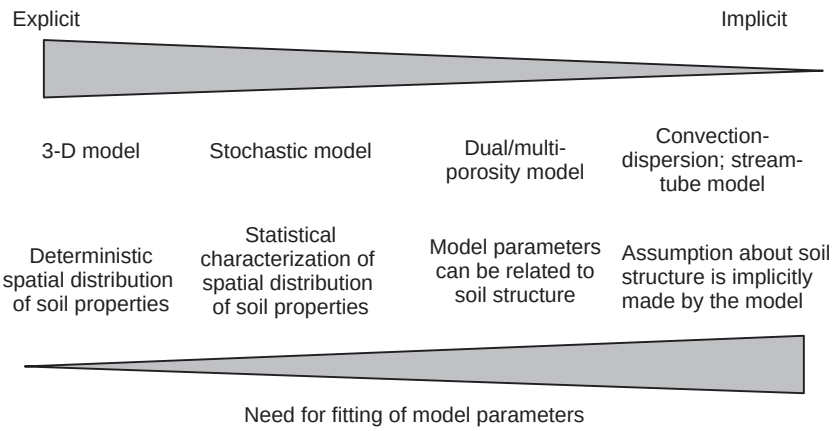


Figure 1 Implementation of structure/heterogeneity in transport models. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Durner and Flühler, 1996) subdomains in which solutes move at different velocities. Between the subdomains solutes are exchanged by diffusion and/or advection. The exchange rate is a fitting parameter that can be related to some extent to the structure and spatial scale of the different flow domains.

In the third class, the structure of the porous medium is characterized in a geostatistical sense. By solving the stochastic-continuum flow and transport equations, the lateral solute mixing and the mixing time are derived from the spatial distribution of macroscopic hydraulic properties (e.g. Russo, 1995a,b).

In the fourth class of models, the structure of the porous medium is explicitly considered and 3-D flow and transport processes, be it at the pore or the macroscopic scale, are solved in a medium with a known structure. When the structure of the medium is identified, the flow and transport processes may be predicted without any fitting of flow and transport parameters (Vogel and Roth, 2003). Kasteel *et al.* (2000) successfully applied this concept at the scale of a small column.

Stream Tube Models (STM)

In a STM, the transport in a 3-D heterogeneous flow field is represented by a set or ensemble of one-dimensional flow lines or "stream tubes" along which a one-dimensional transport process takes place. STMs ignore the lateral exchange of mass between the stream tubes during the transport process, which is characterized in a stream tube by a set of constant local transport parameters (e.g. Jury and Roth, 1990). Flow concentrations at the outlet of the flow domain are obtained by integrating the stream-tube scale local concentration over the set of available stream tubes in the total flow domain, considering the probability function of the local-scale velocities as weighing functions. The effective dispersion generated in this model will be completely determined by the distribution functions of the flow field describing the variability of the advection velocities in the ensemble of stream tubes (Simmons, 1982). STMs have been used for basically two different purposes. In the first type of application, STMs were used just to parameterize the heterogeneity of the flow field between an injection and observation surface. This parameterization was subsequently used to describe the behavior of reactive substances in heterogeneous flow fields using either analytical solutions (e.g. Simmons *et al.*, 1995; Dagan and Cvetkovic, 1996; Cvetkovic and Dagan, 1996) or numerical solutions of the local 1-D transport and reaction equations (e.g. Ginn, 2001; Malmström *et al.*, 2004). The parameterization of the STM is adapted depending on the distance between the injection and observation surface to represent the effect of changing velocities along the stream lines due to local stream line convergence and divergence in a 3-D flow field as well as due to mixing and solute exchange between stream lines (e.g. Cirpka and Kitanidis, 2000). Therefore, this type of application does not make a certain assumption about the 3-D flow field.

In the second application, the STM is assumed to represent the 3-D flow field in the subsurface and is used to predict the solute breakthrough at other depths than the depth at which it is parameterized. The most simple STM considers only piston flow to occur in each stream tube so that the velocity of a particle remains constant along its trajectory. For steady-state flow, this implies the following relation between pulse input response functions or travel time distributions at two different depths (Jury, 1982):

$$\mathrm{f}^{\mathrm{f}}(t;z) = \frac{l}{z}\mathrm{f}^{\mathrm{f}}\left(\frac{tl}{z};l\right) \tag{16}$$

Considering a lognormal distribution for describing the pdf of the local velocity field results in the convective lognormal transport (CLT) model (Jury, 1982). Using transfer function representations, it can be shown that the effective dispersivity in this case will grow linearly with travel depth and travel time. The variance of the solute travel time grows quadratically with depth. The CLT model was successfully used to characterize the solute dispersion in structured soils at the scale of soil columns (Persson and Berndtsson, 1998), undisturbed soil monoliths (Vanclooster *et al.*, 1995; Vanderborght *et al.*, 1997a; Javaux and Vanclooster, 2003b), or at the field-scale (Butters and Jury, 1989; Jacques *et al.*, 1998). Parameters of the STM are obtained by interpretation of transport experiments. The description of the flow process in the stream tubes was improved by considering local-scale variability and dispersion (Bresler and Dagan, 1981; Destouni and Cvetkovic, 1991; Toride and Leij, 1996; Mallants *et al.*, 1996c) and transient conditions (Bresler and Dagan, 1983; Dagan and Bresler, 1983).

However, solute transport in a natural soil cannot always be conceptualized using the two asymptotic models: convective dispersive and stream-tube model. Structural heterogeneity may generate heterogeneous flow, modifying the mixing process when solutes migrate through the soil (White *et al.*, 1986; Snow *et al.*, 1994; Nissen *et al.*, 2000; Seuntjens *et al.*, 2001). In these cases, the solute dispersion process and the variance of the solute travel time with depth, or the variance of the travel depth with time will be an irregular function.

Recently, Zhang (2000) presented a flexible generalized transfer function model (GTF) that encompasses as well the convective stochastic STM as the convective dispersive transport process. In addition, the GTF model can also be used to characterize solute transport processes in heterogeneous soils such as those in which the mean travel time increases nonlinearly with depth and those in which the dispersivity is a scale-dependent function of the travel distance with any power value. The GTF is thereby a comprehensive TFM allowing a formalization of the solute

transport process in heterogeneous soils in a synoptic way. The additional flexibility of the GTF model is at the expense of increasing the number of fitting parameters. However, using the robust estimation techniques developed by Vanderborght *et al.* (1996), Javaux and Vanclooster (2003a) showed that the GTF transport parameters can be estimated in a robust way from classical solute transport experiments.

Multidomain Models

Multidomain models have been developed to describe variability in real soils, in particular, as a result of observed solute transport in soil column studies. In the most simple case, the flow domain is separated in two regions. Examples of these models are the mobile–immobile transport model (MIM; e.g.Coats and Smith, 1964; van Genuchten and Wierenga, 1976) where flow in the immobile region is considered to be negligible and the macropore/micropore flow models (Jarvis *et al.*, 1991a; Chen and Wagenet, 1992; Gerke and van Genuchten, 1993) where advective transport in the micropores is considered. In case of two-region models, the solutes are divided in a mobile and an immobile soil region. Thus:

$$\theta C_{\mathrm{l}}^{\mathrm{r}} = \theta_{\mathrm{m}} C_{\mathrm{m}}^{\mathrm{r}} + \theta_{\mathrm{im}} C_{\mathrm{im}}^{\mathrm{r}} \qquad (17)$$

with θ_{m} is the mobile and $\theta_{\mathrm{im}} = \theta - \theta_{\mathrm{m}}$ the immobile volumetric soil water content, and $C_{\mathrm{m}}^{\mathrm{r}}$ and $C_{\mathrm{im}}^{\mathrm{r}}$ are the mobile and immobile resident concentrations respectively. Lateral mixing of the solutes can be diffusion limited and modeled by a kinetic first-order rate constant, in which the first-order rate constant depends on the effective diffusion and the size of the aggregates in the immobile region. The two-region MIM models were successfully used to describe solute transport in real soils (Nkedi-Kizza *et al.*, 1983; Vanclooster *et al.*, 1992; Mallants *et al.*, 1996b). Other common macroporous models have been successfully used in other studies (Jarvis *et al.*, 1991b).

In contrast to the CDE model, more parameters are needed to apply the two-region model for describing the solute transport process in soil. The increase of parameters definitely increases the flexibility with which the solute transport process can be described. However, this increased flexibility should be carefully balanced with the well-posedness of the identification problem. Vanderborght *et al.* (1997c) for instance, clearly illustrated the uncertainty associated with the identification of MIM parameters from classical breakthrough experiments.

Stochastic-continuum Models

Within a stochastic-continuum approach, the spatial variability and structure of flow properties of the heterogeneous soil are represented by spatial random functions. Similar to a random variable, a spatial random function is characterized by probability density functions. In order to simplify the characterization of the random space function, the second-order stationarity criterion is invoked. Second-order stationarity implies that the mean and variance of the variable are constant in space and that the spatial covariance only depends on the distance between the observations. The stochastic-flow and transport equations, of which the parameters are random space functions, can be solved in an approximate way using a first-order perturbation approach (e.g. Gelhar and Axness, 1983; Yeh *et al.*, 1985; Dagan, 1989; Russo, 1993; Zhang, 2002). The outcome of these first-order solutions is a closed-form relation between the spatial statistics of the flow and transport parameters and those of the flow and transport variables. For transport modeling, the expected concentration or the average concentration at a certain depth in the soil profile is the principal variable of interest. First-order approximate solutions of the stochastic-flow and transport equations revealed that the average concentration may be described by an "effective" convective–dispersive model for transport distances that are much larger than the correlation scale of the hydraulic conductivity. The "effective" or "macrodispersion" coefficient is derived from the spatial variability of the flow parameters. For small travel distances, the macrodispersion was shown to scale linearly with travel distance or time similar to a stream-tube model. However, owing to mixing at larger travel distances, the increase of the macrodispersion with distance flattens of at intermediate travel distances until it reaches a constant value. Another important result of first-order approximate solutions of the stochastic-flow and transport equation is the flow rate dependence of the "macrodispersivity" (Russo, 1995b). Flow rate changes induce different macroscopic velocity fields influencing the mixing process and thus the hydrodynamic dispersivity (e.g. Roth and Hammel, 1996). The change of the macrodispersivity with flow rate or degree of saturation depends on the correlation between the saturated hydraulic and a parameter that quantifies the distance over which capillary forces influence the flow, that is, the capillary length scale (Raats, 1976). For a negative correlation between the capillary length scale and the hydraulic conductivity, the macrodispersivity changes nonmonotonically with degree of saturation and a critical water content exists for which the macrodispersivity is minimal. The flow heterogeneity and macrodispersivity increase when the soil becomes drier or wetter than this critical water content. For a positive correlation, the macrodispersivity increases with decreasing soil water content.

In addition to spatial variability, temporal variability also affects greatly the transport in heterogeneous flow field. However, the solution of the stochastic-flow and transport equations for transient flow conditions is far more complex. Russo *et al.* (1994) carried out numerical experiments of solute transport in generated Gaussian heterogeneous

conductivity fields for transient boundary conditions and conclude to a decrease of the dispersion compared to steady-state conditions. However, these results depend on the parameter cross-correlations. Therefore, the conclusion that transient flow conditions lead to a lower vertical solute spreading or effective dispersivity than in steady-state flow conditions can at present not be generalized. The more so as experimental data that could confirm this conclusion are missing or even contradicting this conclusion (e.g. Vanderborght *et al.*, 2000). Transient boundary conditions may lead additionally to temporal changes of soil physical and physico-chemical properties, such as a change in soil structure and the occurrence of water repellency due to soil drying, which enhance solute spreading. These effects are not considered in modeling approaches that are based on a concept of spatially variable but temporally constant soil properties.

Foussereau *et al.* (2000, 2001) also included the uncertainty of the boundary fluxes in the velocity two-point covariance function on the prediction of expected concentration breakthrough curves at a certain depth. For the cases they considered, that is, an inert tracer and a humid climate without evapotranspiration, they found that the uncertainty of the boundary fluxes dominates the travel time variance of a solute from the surface to a certain observation. However, the importance of the climatic boundary condition uncertainty relative to that of the soil properties depends also on the absolute solute travel time from the surface to a certain depth. The larger this time period, the larger is the time averaging window and the smaller the variability or uncertainty of the averaged rainfall. Therefore, the conclusion that uncertainty of the climatic boundary conditions dominates the prediction of ensemble-averaged concentrations must be questioned for substances that are retarded due to sorption and need a considerably longer time to leach.

Stochastic-continuum approaches were also developed for studying the influence of chemical heterogeneity of the medium on the transport process. For linear reactions (linear adsorption isotherms), the field-scale averaged concentrations can, similarly to the transport of inert tracers, be described using an "effective" convection-dispersive model. But, an important outcome of these studies is that the macrodispersion coefficient for reactive substances depends on the variance and spatial covariance of the local-scale sorption coefficients and their spatial correlation with the hydraulic soil properties (e.g. Bellin *et al.*, 1993; Burr *et al.*, 1994). As a consequence, the macroscale dispersion coefficient that should be used to describe the leaching of a sorbing substance is different from that of an inert tracer. Hence, the use of a model that describes large-scale reactive transport, using a macrodispersion inferred from inert solute transport studies and effective retardation inferred from local-scale batch experiments will fail to describe effective reactive transport. Examples of the failure of the latter approach are illustrated by Kasteel *et al.* (2002) and Javaux (2004).

Although stochastic-continuum models have contributed a lot to our understanding on the effect of heterogeneity on transport and the scale-dependency of effective transport parameters, there are some additional drawbacks that limit their application for real soils. First, the first-order perturbation approach only yields reliable results when the variability of the flow and transport parameters is not too large. This assumption is not always guaranteed in real soils, especially not under unsaturated conditions when the hydraulic conductivity may vary by several orders of magnitude. Second, the stationarity criterion is clearly violated in layered media, such as soils. But, even in a stationary conductivity field, flow may be highly nonstationary as for instance, in case of an infiltration in an initially dry soil. Finally, a reliable characterization of the spatial variability remains a difficult task, which is jeopardized by the limited observations of flow and transport properties in the subsoil.

CONCLUSIONS

For the last decade, solute transport in soil has been studied to be able to cope with the soil management and engineering problem at the larger scale. Understanding the influence of the porous medium heterogeneity and velocity field variability on the plume dispersion is a key issue in this research area. Several modeling approaches (described in Section "Transport models in heterogeneous soils") have been developed to account for an increasing complexity in the description of the medium heterogeneity. In addition, parameter identification approaches have been developed to infer larger-scale solute transport properties from local measurements. Research has also demonstrated the weaknesses of some of the classical modeling approaches, in particular, the convection dispersion model for describing larger-scale solute transport.

However, despite the well-established limitations of the classical modeling approaches and the broad range of alternatives which are presently developed in geosciences, the simple CDE model is still the most popular model used in an engineering context (Vanclooster *et al.*, 2004). The gap between science and engineering in this area can partially be explained by the reductionistic research approach that is often followed when studying soil solute transport processes. Indeed, to understand the effect of flow field variability on transport, geoscientists often built their models on the interpretation of inert solute transport experiments. However, to implement such understanding in engineering models a more holistic approach is needed where the interaction(s) between flow field heterogeneity, local chemical reactivity, and plant uptake are considered.

Recent research on transport of nonlinear reactive substances has shown that such a holistic approach may be

extremely cumbersome, since traditional upscaling procedure may no longer be valid. The integration of plant processes is another point of concern. Although plants are almost always considered in practical problems, the effect of root water uptake on the heterogeneous flow field and the solute transport is essentially unknown. Through numerical-simulations, Russo *et al.* (1998) suggested that crop water uptake intensifies lateral water redistribution in the root zone even more and therefore further reduces the vertical solute spreading as compared with the steady-state flow condition. However, to date, more research is necessary to experimentally characterize and to model the impact of plant water and nutrient uptake on solute transport.

To conclude, the use of the classical CDE model to describe the fate and transport of substances in the subsurface for practical applications should be coupled with the conscience that the dispersion coefficient which is used in this equation and which may have a large impact on the predicted environmental concentrations (e.g. Boesten, 2004) is a parameter that lumps different processes. It depends on static soil properties such as soil heterogeneity and structure, and also on problem specific conditions, such as the scale of the transport process, the boundary conditions, substance properties, and plant root activity. Therefore, the classical representation in textbooks of the hydrodynamic dispersivity being a static soil property is a naïve and incorrect representation of reality.

REFERENCES

Bejat L., Perfect E., Quisenberry V.L., Coyne M.S. and Haszler G.R. (2000) Solute transport as related to soil structure in unsaturated intact soil blocks. *Soil Science Society of America Journal*, **64**, 818–826.

Bellin A., Rinaldo A., Bosma W.J.P., Van der Zee S.E.A.T.M. and Rubin Y. (1993) Linear equilibrium adsorbing solute transport in physically and chemically heterogeneous porous formations. 1. Analytical solutions. *Water Resources Research*, **29**(12), 4019–4030.

Boesten J.J.T.I. (2004) Influence of dispersion length on leaching calculated with PEARL, PELMO and PRZM for FOCUS groundwater scenarios. *Pest Management Science*, **60**(10), 971–980.

Bouma J. (1991) Influence of soil macroporosity on environmental quality. *Advances in Agronomy*, **46**, 1–37.

Bresler E. and Dagan G. (1981) Convective and pore scale dispersive solute transport in unsaturated heterogeneous fields. *Water Resources Research*, **22**, 1531–1536.

Bresler E. and Dagan G. (1983) Unsaturated flow in spatially variable fields. Solute transport models and their application to two fields. *Water Resources Research*, **19**, 421–428.

Burr D.T., Sudicky E.A. and Naff R.L. (1994) Nonreactive and reactive solute transport in three-dimensional heterogeneous porous media: mean displacement, plume spreading, and uncertainty. *Water Resources Research*, **28**, 1517–1530.

Butters G.L. and Jury W.A. (1989) Field scale transport of bromide in an unsaturated soil. 2. Dispersion modeling. *Water Resources Research*, **25**, 1583–1589.

Chen C. and Wagenet R.J. (1992) Simulation of water and chemicals in macropore soils. Part 1. Representation of the equivalent macropore influence and its effect on soil water flow. *Journal of Hydrology*, **130**, 10–126.

Cirpka O. and Kitanidis P. (2000) An advective-dispersive stream tube approach for the transfer of conservative-tracer data to reactive transport. *Water Resources Research*, **36**, 1209–1220.

Coats K.H. and Smith B.D. (1964) Dead end pore volume and dispersion in porous media. *Society of Petroleum Engineers Journal*, **4**, 73–84.

Comegna V., Coppola A. and Sommella A. (1999) Nonreactive solute transport in variously structured soil materials as determined by laboratory-based time domain reflectometry (TDR). *Geoderma*, **92**, 167–184.

Corey J.C., Nielsen D.R. and Biggar J.W. (1963) Miscible displacement in saturated and unsaturated sandstone. *Soil Science Society of America Proceedings*, **27**, 258–262.

Cvetkovic V. and Dagan G. (1996) Reactive transport and immiscible flow in geological media. II. Applications. *Proceedings of the Royal Society of London*, **452**, 303–328.

Dagan G. (1989) *Flow and Transport in Porous Formations*, Springer-Verlag, p. 461.

Dagan G. and Bresler E. (1983) Unsaturated flow in spatially variable fields. Derivation of models of infiltration and redistribution. *Water Resources Research*, **19**, 413–420.

Dagan G. and Cvetkovic V. (1996) Reactive transport and immiscible flow in geological media. I. General theory. *Proceedings of the Royal Society of London Series A*, **452**, 285–301.

de Marsily G. (1986) *Quantitative Hydrogeology. Groundwater for Engineers*, Academic Press, p. 440.

De Rooij G.H. (2000) Modeling fingered flow of water in soil owing to wetting front instability: a review. *Journal of Hydrology*, **231**, 277–294.

De Smedt F., Wauters F. and Sevilla J. (1986) Study of tracer movement through unsaturated sand. *Journal of Hydrology*, **85**, 196–181.

Destouni G. and Cvetkovic V. (1991) Field scale mass arrival of sorptive solute into the groundwater. *Water Resources Research*, **27**, 1315–1325.

Durner W. and Flühler H. (1996) Multi-domain models for pore-size dependent transport of solutes in soils. *Geoderma*, **70**, 281–297.

Dyson J.S. and White R.E. (1989) The effect of irrigation rate on solute transport in soil during steady state water flow. *Journal of Hydrology*, **107**, 19–29.

Einstein A. (1905) Ueber die von der molekularkinetsichen Theorie der Waerme geforderte Bewegung von in ruhenden Fluessigkeiten suspendierten Teilchen. *Annales de Physique*, **17**, 549–560.

Flühler H., Durner W. and Flury M. (1996) Lateral solute mixing processes – a key for understanding field-scale transport of water and solutes. *Geoderma*, **70**, 165–183.

Flury M., Flühler H., Jury W.A. and Leuenberger J. (1994) Susceptibility of soils to preferential flow of water: a field study. *Water Resources Research*, **31**, 2443–2452.

Foussereau X., Graham W., Aakpoji A., Destouni G. and Rao P.S.C. (2000) Stochastic analysis of transport in unsaturated heterogeneous soils under transient flow regimes. *Water Resources Research*, **36**, 911–921.

Foussereau X., Graham W., Aakpoji A., Destouni G. and Rao P.S.C. (2001) Solute transport through a heterogeneous coupled vadose-saturated zone system with temporally random rainfall. *Water Resources Research*, **37**, 1577–1588.

Gelhar L.W. and Axness C. (1983) Three dimensional stochastic analysis of macrodispersion in aquifers. *Water Resources Research*, **19**, 161–190.

Gerke H.H. and van Genuchten M.T.h (1993) A dual porosity model for simulating the preferential movement of water and solutes in structured porous media. *Water Resources Research*, **29**, 305–319.

Ginn T.R. (2001) Stochastic-convective transport with nonlinear reactions and mixing: finite streamtube ensemble formulation for multicomponent reaction systems with intra-streamtube dispersion. *Journal of Contaminant Hydrology*, **47**, 1–28.

Gwo J.P., Jardine P.M., Wilson G.V. and Yeh G.T. (1995) A multiple-pore-region concept to modeling mass transfer in subsurface media. *Journal of Hydrology*, **164**, 217–237.

Huang K., Toride N. and van Genuchten M.T.h (1995) Experimental investigation of solute transport in large, homogeneous and heterogeneous, saturated soil columns. *Transport in Porous Media*, **18**, 283–302.

Jacques D., Kim D.J., Diels J., Vanderborght J., Vereecken H. and Feyen J. (1998) Analysis of steady state chloride transport through two heterogeneous field soils. *Water Resources Research*, **34**, 2539–2550.

Jarvis N., Bergström L. and Messing I. (1991a) Modeling water and solute transport in macroporous soil. I. Model description and sensitivity analysis. *Journal of Soil Science*, **42**, 59–70.

Jarvis N.J., Bergström L. and Dik P.E. (1991b) Modeling water and solute transport in macroporous soil. II. Chloride breakthrough under nonsteady flow. *Journal of Soil Science*, **42**, 71–81.

Javaux M. (2004) Solute transport in a heterogeneous unsaturated subsoil: experiments and modeling, PhD Thesis no. 45, Faculte d'ingenierie biologique, agronomique et environnementale, Universite catholique de Louvain, Louvain-la-Neuve, Belgium.

Javaux M. and Vanclooster M. (2003a) Robust estimation of the generalized solute transfer function parameters. *Soil Science Society of America Journal*, **67**, 81–91.

Javaux M. and Vanclooster M. (2003b) Scale- and rate- dependent solute transport within an unsaturated sandy monolith. *Soil Science Society of America Journal*, **67**, 1334–1343.

Jury W.A. (1982) Simulation of solute transport using a transfer function model. *Water Resources Research*, **18**, 363–368.

Jury W.A. (1985) *Spatial Variability of Soil Physical Parameters in Solute Migration: A Critical Literature Review*, Report EPRI EA-4228, Electric Power Research Institute, Palo Alto.

Jury W.A. and Roth K. (1990) *Transfer Functions and Solute Movement Through Soil*, Birkhäuser Boston: Cambridge.

Kasteel R. (1997) *Solute Transport in an Unsaturated Field Soil: Describing Heterogeneous Flow Fields using Spatial Distribution of Hydraulic Properties*, PhD Thesis no. 12477, Ethan Allen Interiors Inc, Zürich.

Kasteel R., Vogel H.J. and Roth K. (2000) From local hydraulic properties to effective transport in soil. *European Journal of Soil Science*, **51**(1), 81–91.

Kasteel R., Vogel H.J. and Roth K. (2002) Effect of non-linear adsorption on the transport behaviour of brilliant blue in a field soil. *European Journal of Soil Science*, **53**(2), 231–240.

Khan A.U.H. and Jury W.A. (1990) A laboratory study of the dispersion scale effect in column outflow experiments. *Journal of Contaminant Hydrology*, **5**, 119–131.

Kung K.-J.S. (1990) Preferential flow in a sandy vadose zone: 2. Mechanisms and implications. *Geoderma*, **46**, 59–71.

Mallants D., Jacques D., Vanclooster M., Diels J. and Feyen J. (1996c) A stochastic approach to simulate water flow in macroporous soil. *Geoderma*, **70**, 299–324.

Mallants D., Mohanty B., Jacques D. and Feyen J. (1996a) Spatial variability of hydraulic properties in a multi-layered soil. *Soil Science*, **161**, 167–181.

Mallants D., Vanclooster M. and Feyen J. (1996b) A transect study on solute transport in macroporous soil. *Hydrological Processes*, **10**, 55–70.

Malmström M.E., Destouni G. and Martinet P. (2004) Modeling expected solute concentration in randomly heterogeneous flow systems with multicomponent reactions. *Environmental Science and Technology*, **38**, 2673–2679.

Maraqa M.A., Wallace R.B. and Voice T.C. (1997) Effects of degree of water saturation on dispersivity and immobile water in sandy soil columns. *Journal of Contaminant Hydrology*, **25**, 199–218.

Nissen H.H., Moldrup P. and Kachanoski R.G. (2000) Time domain reflectometry measurements of solute transport across a soil layer boundary. *Soil Science Society of America Journal*, **64**, 62–74.

Nkedi-Kizza P., Biggar J.W., Van Genuchten M.T.h, Wierenga P.J., Selim H.M., Davidson J.M. and Nielsen D.R. (1983) Modeling tritium and chloride 36 transport through an aggregated oxisol. *Water Resources Research*, **19**, 691–700.

Padilla I.Y., Yeh T.C.J. and Conklin M.H. (1999) The effect of water content on solute transport in unsaturated porous media. *Water Resources Research*, **35**, 3303–3313.

Pang L. and Hunt B. (2001) Solutions and verification of a scale-dependent dispersion model. *Journal of Contaminant Hydrology*, **53**, 21–39.

Peck A.J. (1964) Moisture profile development and air compression during water uptake by bounded porous bodies: 3. Vertical columns. *Soil Science*, **100**, 44–51.

Persson M. and Berndtsson R. (1998) Estimating transport parameters in an undisturbed soil column using time domain reflectometry and transfer function theory. *Journal of Hydrology*, **205**, 232–247.

Pickens J.F. and Grisak G.E. (1981) Scale-dependent dispersion in stratified granular aquifer. *Water Resources Research*, **17**, 1191–1211.

Poletika N.N. and Jury W.A. (1994) Effects of soil surface management on water flow distribution ans solute dispersion. *Soil Science Society of America Journal*, **58**, 999–1006.

Quisenberry V.L., Philips R.E. and Zeleznik J.M. (1994) Spatial distribution of water and chloride macropore flow in a well structured soil. *Soil Science Society of America Journal*, **58**, 1294–1300.

Raats P.A.C. (1973) Unstable wetting fronts in uniform and nonuniform soils. *Soil Science Society of America Proceedings*, **37**, 681–685.

Raats P.A.C. (1976) Analytic solutions of a simplified equation. *Transactions of the ASAE*, **19**, 683–689.

Radcliffe D.E., Gupte S.M. and Box J.E. (1998) Solute transport at the pedon and polypedon scales. *Nutrient-Cycling-in-Agroecosystems*, **50**, 77–84.

Ritsema C. and Dekker L.W. (1995) Distribution flow: a general process in the top layer of a water repellent soil. *Water Resources Research*, **31**, 1187–2000.

Ritsema C.J., Dekker L.W., Nieber J.L. and Steenhuis T.S. (1998) Modeling and field evidence of finger formation and finger recurrence in a water repellent sandy soil. *Water Resources Research*, **34**, 555–567.

Roth K. and Hammel K. (1996) Transport of conservative chemical through an unsaturated two-dimensional Miller-similar medium with steady state flow. *Water Resources Research*, **32**, 1653–1663.

Roth K., Jury W.A., Flühler H. and Attinger W. (1991) Transport of chloride through an unsaturated field soil. *Water Resources Research*, **10**, 2533–2541.

Russo D. (1993) Stochastic modeling of macrodispersion for solute transport in a heterogeneous unsaturated porous medium. *Water Resources Research*, **29**, 1731–1744.

Russo D. (1995a) On the velocity covariance and transport modeling in heterogeneous anisotropic porous formations. 1. Saturated flow. *Water Resources Research*, **31**, 129–137.

Russo D. (1995b) On the velocity covariance and transport modeling in heterogeneous anisotropic porous formations. 2. Unsaturated flow. *Water Resources Research*, **31**, 139–145.

Russo D., Zaidel J. and Laufer A. (1994) Stochastic analysis of solute transport in partially-saturated heterogeneous soil: I. Numerical experiments. *Water Resources Research*, **30**, 769–779.

Russo D., Zaidel J. and Laufer A. (1998) Numerical analysis of transport in a three-dimensional partially saturated heterogeneous soil. *Water Resources Research*, **34**, 1451–1468.

Saffman P.G. (1959) A theory of dispersion in a porous medium. *Journal of Fluid Mechanics*, **6**, 312–349.

Seuntjens P., Mallants D., Toride N., Cornelis C.h and Geuzens P. (2001) Grid lysimeter study of steady-state chloride transport in two spodosol types using TDR and wick samplers. *Journal of Contaminant Hydrology*, **51**, 13–39.

Simmons C.S. (1982) A stochastic convective transport representation of dispersion in one-dimensional porous media systems. *Water Resources Research*, **18**, 1193–1214.

Simmons C.S., Ginn T.R. and Wood B.D. (1995) Stochastic-convective transport with nonlinear reaction: 1. Mathematical framework. *Water Resources Research*, **31**, 2675–2688.

Skopp J. and Gardner W.R. (1992) Miscible displacement: an interacting flow region model. *Soil Science Society of America Journal*, **56**, 1680–1686.

Snow V.O., Clothier B.E., Scotter D.R. and White R.E. (1994) Solute transport in a layered field soil. Experiments and modeling using the convection-dispersion approach. *Journal of Contaminant Hydrology*, **16**, 339–358.

Sposito G. and Jury W.A. (1988) The lifetime probability density function for solute movement in the subsurface zone. *Journal of Hydrology*, **102**, 503–518.

Steenhuis T.S., Parlange J.Y. and Andreini M.S. (1990) A numerical model for preferential solute movement in structured soils. *Geoderma*, **46**, 193–208.

Taylor G.I. (1953) Dispersion of soluble matter in solvent flowing through a tube. *Proceedings of the Royal Society of London*, **219**, 186–203.

Toride N. and Leij F. (1996) Convective-dispersive stream tube model for field scale solute transport. 1. Moment analysis. *Soil Science Society of America Journal*, **60**, 342–352.

Vanclooster M., Boesten J., Tiktak A., Jarvis N., Kroes J., Clothier B.E. and Green S. (2004) On the use of unsaturated flow and transport models in nutrient and pesticide management. In *Unsaturated Zone Modelling: Progress, Challenges and Applications*, Feddes R., De Rooij G. and Van Dam J. (Eds.), Kluwer Academic Publishers: pp. 331–361.

Vanclooster M., Mallants D., Vanderborght J., Diels J., Van Orshoven J. and Feyen J. (1995) Monitoring solute transport in a multi-layered sandy lysimeter using time domain reflectometry. *Soil Science Society of America Journal*, **59**, 337–344.

Vanclooster M., Vereecken H., Diels J., Huysmans F., Verstraete W. and Feyen J. (1992) Effect of mobile and immobile water in predicting nitrogen leaching from cropped soils. *Modelling Geo-Biosphere Processes*, **1**, 23–40.

Vanderborght J., Gonzalez C., Vanclooster M., Mallants D. and Feyen J. (1997a) Effects of soil type and water flux on solute transport. *Soil Science Society of America Journal*, **61**, 372–390.

Vanderborght J., Jacques D., Mallants D., Tseng P.H. and Feyen J. (1997b) Comparison between field measurements and numerical simulation of steady-state solute transport in a heterogeneous soil profile. *Hydrology and Earth System Sciences*, **4**, 853–871.

Vanderborght J., Mallants D., Vanclooster M. and Feyen J. (1997c) Parameter uncertainty in the mobile-immobile solute transport model. *Journal of Hydrology*, **190**, 75–101.

Vanderborght J., Timmerman A. and Feyen J. (2000) Solute transport for steady-state and transient flow in soils with and without macropores. *Soil Science Society of America Journal*, **64**, 1305–1317.

Vanderborght J., Vanclooster M., Mallants D., Diels J. and Feyen J. (1996) Determining convective lognormal solute transport parameters from resident concentration data. *Soil Science Society of America Journal*, **60**, 1306–1317.

Vanderborght J., Vanclooster M., Timmerman A., Seuntjens P., Mallants D., Kim D.J., Jacques D., Hubrechts L., Gonzalez C., Feyen J., *et al.* (2001) Overview of inert tracer experiment in key Belgian soil types: relation between transport, and soil morphological and hydraulic properties. *Water Resources Research*, **37**, 2873–2888.

van Genuchten M.T.h and Wierenga P.J. (1976) Mass transfer studies in sorbing porous media. I. Analytical solutions. *Soil Science Society of America Journal*, **40**, 473–480.

Van Wesenbeeck I.J. and Kachanoski R.G. (1995) Predicting field-scale solute transport using in-situ measurements of soil hydraulic properties. *Soil Science Society of America Journal*, **59**, 734–742.

Vogel H.-J. and Roth K. (2003) Moving through scales of flow and transport in soils. *Journal of Hydrology*, **272**, 95–106.

White R.E. (1985) The influence of macropores on the transport of dissolved and suspended matter through soil. *Advances in Soil Science*, **3**, 95–121.

White R.E., Dyson J.S., Gerstl Z. and Yaron B. (1986) Leaching of herbicides through undisturbed cores of a structured clay soil. *Soil Science Society of America Journal*, **50**, 277–283.

Wilson G.V., Yunsheng L., Selim H.M., Essington M.E. and Tyler D.D. (1998) Tillage and cover crop effects on saturated and unsaturated transport of fluometuron. *Soil Science Society of America Journal*, **62**, 46–55.

Yeh T.C., Gelhar L.W. and Gutjahr A.L. (1985) Stochastic analysis of unsaturated flow in heterogeneous soils. 1. Statistically isotropic media. *Water Resources Research*, **21**, 477–456.

Zhang R.D. (2000) Generalized transfer function model for solute transport in heterogeneous soils. *Soil Science Society of America Journal*, **64**, 1595–1602.

Zhang D. (2002) Stochastic methods for flow in porous media. *Coping with Uncertainties*, Academic Press: p. 350.

70: Transpiration and Root Water Uptake

PETER DE WILLIGEN[1], MARIUS HEINEN[1] AND MARY BETH KIRKHAM[2]

[1]*Alterra, Wageningen, The Netherlands*
[2]*Department of Agronomy, Kansas State University, Manhattan, KS, US*

Root water uptake is one of the major components of the water balance of a soil. In this article, both the physical as well as the physiological aspects of root water uptake and transpiration are dealt with. The regulation of transpiration by different mechanisms on different levels (e.g. through hormones, adaptation of vessel diameter, root radius, and vertical extension of the root system) is discussed. Attention is paid to the modeling of water uptake by single roots and root systems. The last section deals with the different techniques used to measure root water uptake, for example, lysimeter, sap flow, inverse modeling, and discusses the advantages and disadvantages of the different methods.

INTRODUCTION

Transpiration is defined as the loss of water in the form of vapor through the surface of leaves and other aerial parts of plants (*see also* the definition given in **Chapter 42, Transpiration, Volume 1**). It is due to the evaporative demand of the atmosphere. Plant roots take up all the water needed for transpiration; only a very small part is used for the increase of fresh plant material. Water uptake is one of the major components in the water balance of a soil. On a global scale, precipitation is estimated to amount to 99–115 $Tm^3\ y^{-1}$, evapotranspiration (the sum of evaporation from soil surfaces and open water and transpiration) to 62–75 $Tm^3\ y^{-1}$, and runoff to 37–40 $Tm^3\ y^{-1}$ (Oki, 1999; WW2010, 1999). Transpiration comprises about half of the evapotranspiration, which in the data shown above includes evaporation from bare soil surfaces, lakes, and streams.

Transpiration depends on the type of vegetation and is usually higher than bare soil evaporation. In the experiments of Aase *et al.* (1996) for instance, a lysimeter covered by a lentil crop lost in about 100 days approximately 250 mm more water than a fallow lysimeter. One of the reasons of enhanced water loss by transpiration compared to the bare soil evaporation is that in the latter case a dry surface layer of soil with low water conductivity develops, which protects the soil of any further desiccation. The root system of a plant, however, forms a system of interconnected vessels, which takes up and transports water also from the deeper layers in the root zone.

Transpiration and root water uptake have both physical (see Section "Root water uptake as a physical process") as well as physiological aspects (see Section "Physiological aspects"). The maximum value of transpiration depends on meteorological conditions, which determine the energy balance of the wet leaf surface. Whether or not the leaf surface continues to be wet depends on plant physiological mechanisms, which in their turn depends on the plant–water status and ultimately on root distribution with respect to distribution of soil reserves of water.

ROOT WATER UPTAKE AS A PHYSICAL PROCESS

Water Uptake as a Passive Process

Though regulated by the plant (see Section "Physiological aspects"), the uptake of water is mainly passive. It is often described as a purely physical process, a consequence of gradients in water potential in the path: bulk soil – soil/root interface – root – leaf – atmosphere. To quote Steudle and Peterson (1998): "Under conditions of transpiration, roots do not 'take up water' so much as they allow it to pass through them. In other words water ... moves passively

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

through the roots in response to a water potential gradient set up by transpiration".

Water potential is expressed in units energy per volume ($J\,m^{-3}$), which is equivalent to force per area ($N\,m^{-2}$) (*see also* **Chapter 73, Soil Water Potential Measurement, Volume 2**, and Bolt *et al.*, 1976). In soil physics, it is frequently expressed in the so-called *head equivalent* expressed in m water pressure (*see* **Chapter 67, Hydrology of Swelling Clay Soils, Volume 2**). In the following, we shall use head terminology and units. The total head (H) is the sum of different components; for our purposes the relevant components are the pressure head (h_p), the osmotic head (h_o), and the gravitational head (h_g). Water flow along the path soil – atmosphere has often been described using the attractive analogue of an electrical current flowing through a series of resistances, as was already done in the first half of the last century (Gradmann, 1928; Van den Honert, 1948). So the flow in the path bulk soil – soil/root interface – root xylem can be given as:

$$Q = \frac{H_s - H_{R0}}{\omega_s} = \frac{H_{R0} - H_x}{\omega_{R0}} \qquad (1)$$

where Q is the flow ($m^3\,s^{-1}$), H_S the total head in the bulk soil (m), H_{R0} is the total head at the interface of soil and root (m), H_x the total head in the root xylem (m), ω_S the resistance against flow of the soil ($s\,m^{-2}$), and ω_{R0} the radial resistance against flow of the region between root wall and root xylem ($s\,m^{-2}$). For a given transpiration rate the highest resistance in the path bulk soil – soil/root interface – root is usually that of the (radial) root resistance, provided that the water content of the soil is not too low. The soil resistance increases strongly with decreasing water content. However, up to levels of the pressure head of -10 to -300 m the soil resistance is about one order of magnitude less than the root resistance (see Figure 5, and Section "Modeling").

For transport from the bulk soil to the soil/root interface the osmotic and gravitational head can be ignored: the former because semipermeable membranes are assumed to be absent in the soil, the latter because in this analysis the flow path is taken to be horizontal. However, in the path from the soil/root interface to root xylem, water has to pass through several semipermeable membranes and differences in osmotic head have to be taken into account.

Water Uptake in Relation to Hydraulic and Osmotic Properties

Many studies have shown a nonlinear relation between the flow of water through a plant and the difference in pressure outside and inside the root. This suggests a variable radial root resistance (see Dalton *et al.*, 1975, for references). Gardner (1970) pointed out that the nonlinearity could be explained on the basis of coupled transport of water and ions across root membranes. Dalton *et al.* (1975) and Fiscus (1975) explained the nonlinearity on the basis of data from literature (e.g. Mees and Weatherley, 1957). These and similar data from literature showed that the relation between pressure head difference, Δh_p, and flow, Q, is nonlinear at low flow rates but linear at high flow rates. The relation between Q and Δh_p can be explained taking into account the existence of a semipermeable membrane in the flow path (Dalton *et al.*, 1975; Fiscus, 1975). Later, the description has been extended to two membranes in series (Dainty, 1985; Miller, 1985; Newman, 1976). Recently, Raats (2005) discussed transport through a single membrane and series arrays of membranes.

The nonlinearity at low flow rates, according to the description where the presence of membranes in the flow path is taken into account, arises from the active transport of solutes into the xylem where they manifest themselves osmotically and generate transport of water through the root even when Δh_p is zero. When Δh_p is large, however, the resulting high flow rates dilute these solutes and their effect on Q becomes negligible. A linear relation between Q and Δh_p then results with slope α, the hydraulic conductance of the root system and, when extrapolated, with an intercept on the pressure axis of almost zero. However, the intercept on the pressure axis as found in experiments is considerably larger than what follows from the theory. This still is a matter of much debate. According to Passioura (1988), there is no alternative but to accept the large value of the intercept without explanation and to use the next equation as the most economical way of summarizing the available data:

$$Q = \alpha(\Delta h_p - h_0 - \sigma\,\Delta h_o) \qquad (2)$$

where α is the conductance ($m^2 s^{-1}$) Δh_p (m) is the difference in hydrostatic pressure, h_0 is a parameter, σ is an effective reflection coefficient for weighting any osmotically induced flow, and Δh_o is the difference in osmotic pressure across the roots. In this article *conductivity* (k) is the proportionality coefficient between flux of water and gradient of water potential:

$$q = -k\nabla H$$

In radial coordinates the volume flow for a root of length Δz at any distance from the root midpoint is:

$$Q = -2\pi r k \Delta z \frac{\mathrm{d}H}{\mathrm{d}r}$$

In the case of steady state Q is a constant, so integrating over a finite distance, say from r_x to r_0, yields:

$$Q = -\frac{2\pi k \Delta z (H_0 - H_x)}{\ln r_0 - \ln r_x} = -\alpha(H_0 - H_x), \text{ or}$$
$$\alpha = \frac{2\pi k \Delta z}{\ln r_0 - \ln r_x}$$

Conductance (α) is thus the proportionality coefficient between the flow of water and the difference of water potential at two points of the flow path. Conductance accordingly applies to a larger scale than conductivity. The reciprocal of conductivity and conductance are called *resistivity and resistance respectively*. The reflection coefficient is a unitless number and varies between 0 and 1. If $\sigma = 1$, all solutes are reflected from the membrane and no solute gets across it. If $\sigma = 0$, all solutes can cross the membrane (Kirkham, 2005, p. 221).

It is common to define the "hydraulic resistance", ω, of a root system as $\Delta H/Q$, where ΔH is the difference in total head across the root system. ω equals the reciprocal of α in equation (2) only if $\sigma = 1$ and $h_o = 0$ (with $\Delta H = \Delta h_p - \Delta h_o$). If these conditions are not true, and if we take the true hydraulic resistance to be $1/\alpha$, then ω is only an apparent hydraulic resistance; and variation in it, through time or with Q, need not imply that α varies (Passioura, 1988, p. 255).

Aquaporins

The hydrophobic nature of the lipid bilayer of a membrane presents a considerable barrier to the free movement of water into the cell and between intracellular compartments. However, plasma membranes (around the cell) and tonoplasts (around vacuoles) can be rendered more permeable to water by proteinaceous transmembrane water channels, called *aquaporins*. Aquaporins are a class of water-channel proteins that have been found in nearly all living organisms (Maurel and Chrispeels, 2001). They are highly expressed in plant membranes. They are members of a family of transmembrane channels known as the *major intrinsic protein* (MIP) family (Maurel, 1997).

Water movement through aquaporins can be modulated rapidly. Evidence suggests that these channels may facilitate water movement in drought-stressed tissues and promote the rapid recovery of turgor on watering (Buchanan *et al.*, 2000, p. 1170–1171).

Over the last 12 years, advances at the cellular level have occurred in the understanding of water movement across plant membranes (Tyerman *et al.*, 1999). It has been found that aquaporins can mediate bidirectional water flow. It is generally accepted that they function as narrow pores or channels through which water flows passively down a free energy gradient. Aquaporins consequently enhance the conductivity of the membrane considerably. They can appear in any membrane. But their presence in root membranes is of particular significance in water uptake. Changes in the nature of the potential energy driving water across a tissue may cause changes in the transport pathway, from apoplastic (cell-wall pathway) to cell-to-cell (water-channel pathway), and changes in hydraulic conductivities.

Figure 1 shows the pathways for passive movement of water and solutes across a root cylinder (Steudle and Frensch, 1996). Two parallel pathways (A) and (B) are considered. (A) represents the cell-to-cell path and (B) the

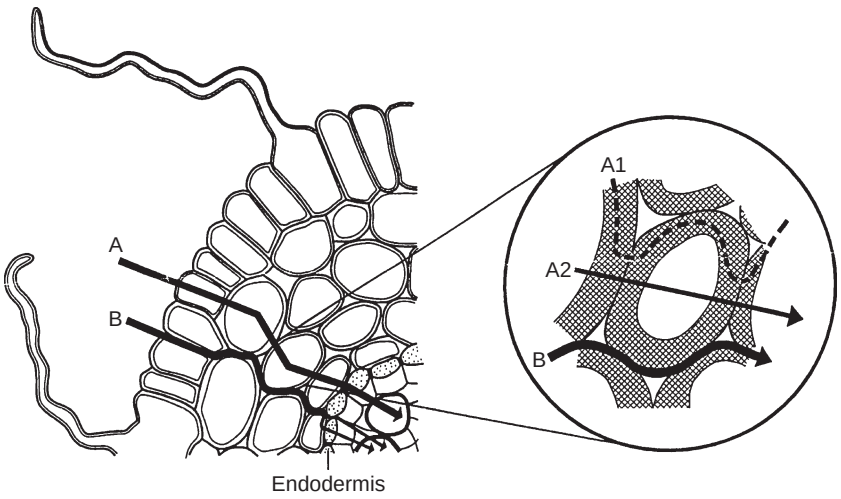


Figure 1 Pathways for passive movement of water and solutes across a root cylinder. Two parallel pathways (A) and (B) are considered. (A) represents the cell-to-cell (protoplastic) path and (B) the apoplastic path. Path (A) is split into symplastic (A1) and transcellular (A2) components. In the symplastic path, flow is restricted to the symplast, that is, water and solutes move from one cell to the next via plasmodesmata. The transcellular path denotes flow across cell membranes. Because components (A1) and (A2) cannot be separated experimentally, the two pathways are summarized as the cell-to-cell path (A). The parallel apoplastic route (B) refers to flow around the protoplasts. Across the endodermis, apoplastic transport is usually considered to be completely blocked by the Casparian band. (From Steudle and Frensch, 1996 by permission of Kluwer Academic Publishers B.V)

apoplastic path. Path (A) is split into symplastic (A1) and transcellular (A2) components. In the symplastic path, flow is restricted to the symplast, that is, water and solutes move from one cell to the next via plasmodesmata. The transcellular path denotes flow across cell membranes where the aquaporins are situated. Because components (A1) and (A2) cannot be separated experimentally, the two pathways are summarized as the cell-to-cell path (A). The parallel apoplastic route (B) refers to flow around protoplasts. Across the endodermis, apoplastic transport (water, solutes) is considered blocked by the Casparian strip. Future work needs to determine how both apoplastic and membrane-bound water flow contribute to the overall water uptake by plant roots.

Hydraulic Redistribution

Roots are movers of water in the soil (Clothier and Green, 1997). One method of movement is through hydraulic redistribution, which occurs when plant roots extract water from a moist soil layer and release it in another part of the root system into a dry soil layer. When transport is upwards the phenomenon is termed *'hydraulic lift'*, whereas the reverse process is called *'downward siphoning'* (see Smith *et al.*, 2004, and the references given there). Kirkham (1983a) presents a physical model based on Darcy's law, which explains the occurrence of hydraulic lift. Hydraulic lift is felt to be important for the survival of plants in arid and semiarid conditions. At night, when stomata are closed and the water potential of the shoot is high because transpiration is much reduced, water at depth can be brought near the soil surface by roots, where it is released into the dry soil, so water is brought to layers where roots are most abundant. Another favorable effect of hydraulic lift is increasing the availability of nutrients in the drier zones. Nutrients are usually most plentiful in the upper layers and their mobility can be considerably enhanced by increased moisture content (Caldwell *et al.*, 1998). The amount of water released during reverse flow is small (Song *et al.*, 2000), but may increase water in soil from the permanent wilting point to the available range.

PHYSIOLOGICAL ASPECTS

Two organs regulate a plant's water status, the leaves and the roots. Here, we consider physiological processes that control transpirational water loss from the leaves and root water uptake. Regulation takes place at different levels, that is, at the level of cells, organs, and assemblages of organs, for example, root systems.

Stomata

Water can be lost through the cuticle or through the stomata. The resistance of movement of water vapor through the cuticle is extremely high, and most water is lost through the stomata when they are open (Waggoner, 1966). The stomata are the main controllers of transpirational water loss and control exchange of carbon dioxide (CO_2) with the atmosphere as well.

A stoma is an opening in the epidermis of leaves and stems, bordered by two guard cells. Figure 2 shows a cross section of a leaf with guard cells and the other cell types in a leaf. The outstanding feature of stomata, the unevenly thickened walls of the guard cells, is related to their changes in shape and the concomitant changes in the size of the stomatal aperture (Esau, 1965, p. 159–160). When turgor increases, the aperture increases in size. Reversed changes occur under decreased turgor. Stomatal guard cells are able to sense environmental signals such as light, carbon dioxide (CO_2) concentration, and humidity, as well as endogenous signals such as plant hormones (Assmann, 1993). Guard cells integrate environmental and endogenous signals, so that the stomatal aperture is neither too small, which would inhibit CO_2 uptake for photosynthesis, nor too large, which would cause excessive water loss.

Environmental Signals

In addition to water availability, light, CO_2, and humidity are the main environmental signals that affect stomatal aperture. Red and blue light stimulate stomatal openings.

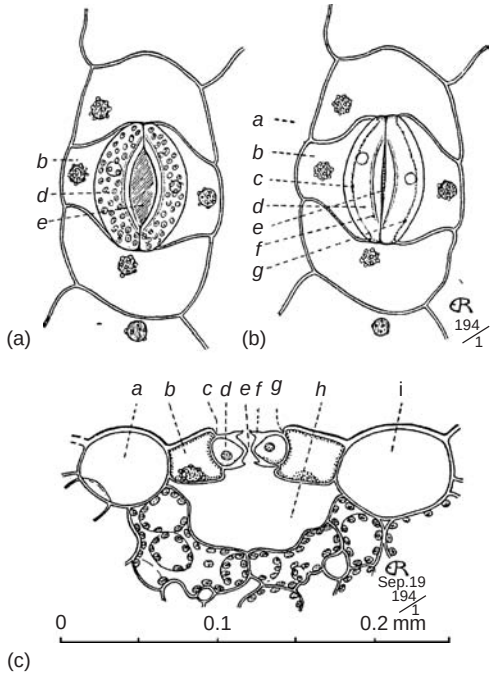


Figure 2 Stoma of *Tradescantia virginiana*. (a) Open stoma viewed from above (b) Closed stoma viewed from above (c) Cross section. *a*, epidermis cell; *b*, subsidiary cell; *c*, joint; *d*, guard cell; *e*, pore; *f*, cuticle; *g* = c; *h*, substomatal cavity (From Reinders, 1957)

Elevated CO_2 typically stimulates stomatal closure, while lowered CO_2 concentrations have the opposite effect. It is the intercellular concentration of CO_2, and not the external concentration, or the concentration within the pore, to which the guard cells are responding. The mechanism of the CO_2 response is unknown.

Reduction in ambient humidity causes stomatal closure, but the reason for the closure is poorly understood. Experiments suggest that transpiration water loss is the important signal. Guard cells respond metabolically to humidity, for example, by losing K^+. Guard cells also may be responding to regulatory metabolites such as abscisic acid (ABA) (see Section "Internal signals"), which are carried in the transpiration stream.

Stomatal closure requires membrane depolarization, which can be driven by anion efflux, and decreases in organic osmotica. Alterations of any of these processes can affect stomatal aperture; such changes are triggered by alterations in the level of activity of two so-called *'second messengers'*, Ca^{2+} and H^+ (pH).

Calcium is an important second messenger. Exogenously applied Ca^{2+} is potent in reducing stomatal aperture (Assmann, 1993). However, there are also data suggesting that Ca^{2+} may play a role in stomatal opening. Also, pH acts as an extracellular signal. As extracellular pH becomes more alkaline, water stress increases, which results in stomatal closure.

Internal Signals

Messengers that are produced within the leaf, or are translocated from a distant site of synthesis include the plant hormones. A hormone, by definition, is a naturally occurring chemical substance formed in some organ of the plant and is carried to another organ or tissue, where it has a specific effect; it operates in small amounts. For 40 years, it has been known that hormones produced in the roots affect shoot–water relations. There are five plant hormones: cytokinins, ethylene, gibberellins, auxins, and abscisic acid. The following paragraphs explain in general terms how hormones affect plant-water relations. For more detailed information, the interested reader can read the reviews for cytokinins, Mok and Mok (2001); for ethylene, Kieber (1997); for gibberellins, Richards *et al.* (2001); for auxins, Leyser (2002); and for abscisic acid, Leung and Giraudat (1998).

Cytokinins are produced in roots and travel to leaves where they promote stomatal opening. Experiments have shown that stomata open more widely than they are under normal conditions when cytokinins are sprayed on plant leaves (Kirkham *et al.*, 1974). The rate of transpiration has been used as a bioassay for cytokinin activity (Luke and Freeman, 1967). Reduced cytokinin production in stressed roots results in increased stomatal resistance (Tal *et al.*, 1970).

Ethylene, the only gaseous hormone, is the simplest organic compound that affects plants. It is active in trace amounts, which makes its detection under nonstressed conditions difficult (Rose, 1985). The effects of ethylene on stomata are contradictory (Kirkham, 1983b). Some experiments show no response of stomata to ethylene, while others have found that ethylene increases stomatal resistance. Essentially, no papers report that ethylene decreases stomatal resistance.

Little information exists concerning the effect of gibberellins on transpiration. Gibberellic acid seems to stimulate transpiration rate (Livnè and Vaadia, 1965) and decrease stomatal resistance (O'Leary and Tarquinio Prisco, 1970).

Depending on the type of auxin and the auxin concentration, either stomatal opening or stomatal closure may occur. Indole-3-acetic acid (IAA) is a natural auxin, and it stimulates stomatal opening at concentrations from 10 μM to 1 μM. With the synthetic auxin 1-NAA (1-napthylacetic acid), maximal stimulation of opening occurs at 5 μM and inhibition of opening occurs at 0.5 μM (Assmann, 1993).

The hormone that has been most widely studied in relation to stomata is abscisic acid because it strongly promotes stomatal closure. ABA is produced in the roots in response to stress, particularly drought stress. It is thought that ABA which is synthesized in roots exposed to dry soil is transported to shoots, where it leads to a reduction in guard cell turgor and stomatal closure (Fry and Huang, 2004, p. 68). Actual measurements of the rate of ABA movement to the shoot, as it is produced under drought stress, are not known.

Even though ABA is important in controlling stomatal aperture, it is probably the balance of hormones that determines the opening of the stomata (Tal and Imber, 1971). Cytokinins decrease stomatal resistance and increase the resistance of the root to absorption of water. The general effect of cytokinins, therefore, is to reduce plant turgor, and their concentration decreases in stressed plants. In contrast, ABA increases stomatal resistance and decreases the resistance of the root to water flow. This hormone increases the plant turgor and its concentration increases in stressed plants. Rapid changes in concentrations of both cytokinins and ABA have been demonstrated in water-stressed plants. The plant undergoes alternating periods of low turgidity during the day, when absorption lags behind water loss by transpiration, and high turgidity at night, when transpiration lags behind water absorption. The periodic changes of water status may induce periodic fluctuations of hormonal concentrations, which influence stomatal and root resistances. The balance of hormones in water-stressed plants may cause the "aftereffect" that is observed in stomata (Fischer *et al.*, 1970). Stomata commonly fail to resume regular opening after a period of wilting of leaves despite rapid recovery of leaf water content. The aftereffect

may be due to an accumulation of an inhibitor like ABA (Livnè and Vaadia, 1972) or a deficiency of a substance that promotes stomatal opening like a cytokinin.

Regulation at the Level of the Root and Roots Systems

Vessel Diameter

The *axial* resistance of a root, that is, the resistance to vertical flow from the root upward through the xylem vessels to the stem, is a function of the vessel diameter. Poiseuille (1799–1869), a French physiologist, discovered the law on velocity of flow of a liquid through a capillary tube. He found that the volume of fluid moving in unit time along a cylinder is proportional to the fourth power of its radius and that the movement depends on the drop in pressure (Kirkham, 2005, p. 216). One of the first important agronomic uses of Poiseuille's law was in a pioneering paper by Passioura (1972) in Australia, where wheat plants face terminal drought. He suggested that when wheat is growing predominantly on stored water, it is an advantage for the plants to have root systems of high axial resistance, so that they will conserve water during early growth and thus have more water available while filling their grain. The xylem of the seminal roots in wheat is dominated by one large metaxylem element (vessel member), the diameter of which (about 50 μm in diameter on average) probably determines the amount of water flowing through the wheat plant, and, thus, indirectly its hydraulic resistance. He suggested that the resistance to flow in the wheat root could be increased in one of two ways: (i) by decreasing the size of the central metaxylem element, or (ii) by reducing the number of seminal roots. He chose to reduce the number of seminal roots. The plants with one seminal root had a larger pressure gradient than the plants with three seminal roots. The single-rooted plants had double the available water at anthesis and produced double the grain yield. Passioura (1972) concluded that it might be possible to conserve water by growing wheat plants with a single seminal root, and it may be possible to breed high root axial resistance into existing cultivars by breeding for smaller vessels. Passioura's classic paper showed that Poiseuille's law can be used to calculate the pressure drop in crops and that this value can be used to breed drought-resistant varieties.

The phloem tissue develops before the xylem tissue (Kirkham, 2005, p. 208). Until the metaxylem elements form, Poiseuille's law cannot be applied and longitudinal (axial) resistance (movement of water in the vertical direction in plants) is large. Taylor and Klepper (1978) discuss axial resistances. Another factor that limits the use of Poiseuille's law is the formation of embolisms in vessels. Air bubbles cause cavitation of the continuous water channels in the vessels. Kirkham (2005, p. 321–327) discusses this problem and how plants maintain water-filled conducting cells.

Richards and Passioura (1981), following the earlier work of Passioura (1972) on the significance of vessel diameter, screened the world's wheat collection (over 1000 accessions) for two factors that control resistance to water movement in the wheat root: number of axes (number of seminal roots) and central metaxylem-vessel diameter. Because there was greater variability in the vessel diameter than in the number of axes, they concluded that the diameter was more important in determining resistance than the number of axes. Accordingly, one should rather breed for change of diameter than for change of number of axes.

Root Radius

Both experimental (Burch, 1979) and theoretical (Gardner, 1960; De Willigen and Van Noordwijk, 1987) studies show that root radius has but a minor influence on root radial resistance and required head gradients in the soil.

Root Growth at Different Soil Depths

A well-developed root system often shows an exponential decrease in root length density with depth (Gerwitz and Page, 1974; Jackson *et al.*, 1996). When soil water is more or less evenly spread over the root zone, distribution of water uptake is determined by distribution of root length density (Sharp and Davies, 1985). In that case most of the extraction of water is from the topsoil. When data from many different experiments were plotted together, it was determined that 40% of the extraction of water is from the first 20% of the root depth, then 33, 20, and 7% from successive depths. This is similar to the rule of thumb that 40% of the water is extracted from the first 30 cm (1 ft), then 30, 20, and 10% from successive 30-cm increments (Gardner, 1983). The percentages are a generalization, and root growth among species and cultivars varies widely. For example, drought-resistant cultivars of wheat have been shown to have more highly branched root systems than drought-sensitive cultivars of wheat (Kirkham and Kanemasu, 1983, p. 490–492).

When there are large differences in water content in the root zone, these patterns mainly determine the distribution of water uptake. Data of Sharp and Davies (1985) depicted in Figure 3 show the distribution of water uptake, root length density, and water content with depth for a well-watered and nonwatered plant. It is clear from this figure that in moist conditions, root length density determines the distribution of water uptake, whereas under dry conditions the distribution of water is more important. Under rain-fed conditions in semiarid regions it is important to have crops that exploit water at all depths in the soil, not just the surface, to make maximum use of water. Two drought-resistant row crops in semiarid areas are grain sorghum and sunflower. A study in Kansas (Rachidi *et al.*, 1993)

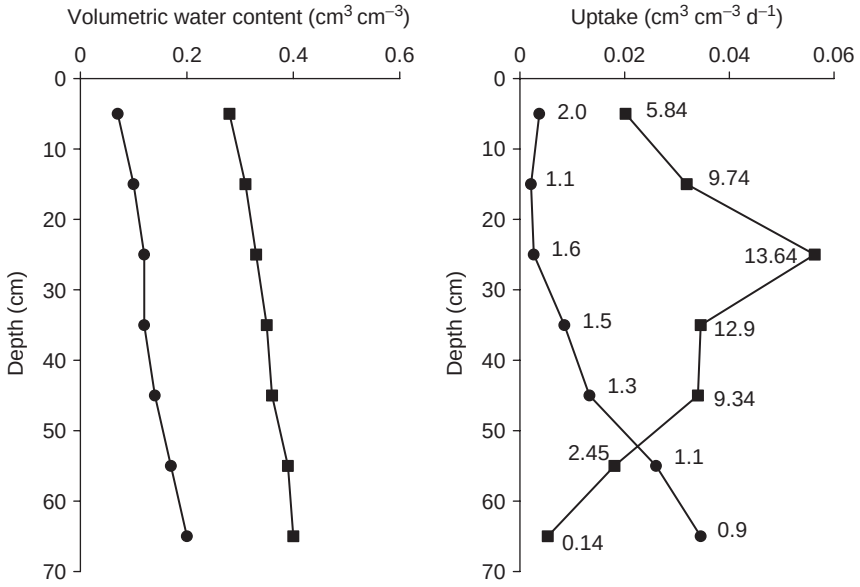


Figure 3 Distribution of water content and water uptake for a maize plant after 18 days, well-watered (■) and unwatered (•). (a) Water content (b) water uptake, the figures at the points give the root length density in cm cm^{-3}. (Data of Sharp and Davies, 1985)

showed that the two crops use water from different depths. Sunflower depleted water to 1 m more depth than sorghum. These results suggest that deep-rooted crops like sunflower might be planted in rotations with shallow-rooted crops to take advantage of water depth.

A split-root experiment with a kiwifruit vine documents the dominant role of surface roots in extracting water under irrigated conditions and demonstrates how irrigation can influence the spatial pattern of root water uptake (Clothier and Green, 1994). Initially, soil water content and water uptake rate were uniformly distributed across the root zone. Next, only one-half of the root zone was irrigated. Following this differential irrigation of the root zone, the sap flow in the "wet" root increased, while the flow in the "dry" root was about halved. Thus, the vine quickly switched its pattern of uptake away from the drier parts of the root zone. Even of greater interest was the vertical pattern of root uptake observed on the wet side. There was a preference of near-surface water uptake. The vine continued to extract water in the densely rooted region surrounding its base, but the shift in uptake to the surface roots on the wet southern side was remarkable. The results showed that greater efficiency in irrigation water use could be obtained by applying small amounts of water more frequently, instead of large, infrequent irrigation. Water from low irrigation would be used rapidly by the active, near-surface roots. This would eliminate the movement of irrigation water into lower regions of the root zone, where draining water passes by inactive roots and is unavailable for plant uptake.

MODELING

Single Root Models

A model by definition is a simplified representation of part of reality; this of course also applies to models on root water uptake. In order to make the problem of transport to and uptake by plant roots manageable, many assumptions and simplifications have to be employed. An obvious and widely used assumption is that of the cylindrical root confined in a soil cylinder, the radius of which is a function of root length density. This leads to the choice of expressing the transport equation in cylindrical coordinates. When tangential and vertical gradients are assumed negligible, it assumes the form:

$$\frac{\partial \theta}{\partial T} = \frac{1}{R}\frac{\partial}{\partial R} D_w R \frac{\partial \theta}{\partial R} \tag{3}$$

where θ is the volumetric water content of the soil (−), T is time (s), and R the radial distance from the center of the root(m). The water diffusivity $D_w(\theta)$ (m^2 s^{-1}) is defined as:

$$D_w = K \frac{\mathrm{d}h_p}{\mathrm{d}\theta} \tag{4}$$

where h_p is the pressure head (m), and K the hydraulic conductivity of the soil (m s^{-1}).

The boundary condition at the outer boundary of the soil cylinder is that of vanishing flux:

$$R = R_1, \;\; D_w \frac{\partial \theta}{\partial R} = 0 \tag{5}$$

where R_1 is the radius of the soil cylinder exploited by a root given by:

$$R_1 = \frac{1}{\sqrt{\pi L_{rv}}} \tag{6}$$

where L_{rv} is the root length density ($m\,m^{-3}$). At the inner boundary – the root surface – the boundary condition is of the form:

$$R = R_0, \;\; D_w \frac{\partial \theta}{\partial R} = \frac{F_1}{2\pi R_0 \Delta z} \tag{7}$$

where R_0 is the root radius (m), Δz the root length (m) and F_1 the uptake rate ($m^3\ s^{-1}$). On the one hand, the uptake rate is linearly related to the difference in hydrostatic and osmotic pressure in analogy to equation (2):

$$F_1 = \alpha\{h_{p,R0} - h_{p,P}\} - \sigma(h_{o,P} - h_{o,R0}) \tag{8}$$

where h_{p,R_0} is the pressure head of the rhizosphere (m), $h_{p,P}$ is the pressure head in the root (m), α is the conductance of the root ($m^2\ s^{-1}$), $h_{o,P}$ is the osmotic head in the root and $h_{o,R0}$ is the osmotic head in the rhizosphere. On the other hand, F_1 is also equal to the transpiration rate per root:

$$F_1 = \pi R_1^2 E = \frac{E}{L_{rv}} \tag{9}$$

where E is the actual transpiration rate ($m\,s^{-1}$), and L_{rv} the root length density ($m\,m^{-3}$).

Both the hydraulic conductivity as well as the pressure head and thus D_w are nonlinear functions of water content, making equation (3), a nonlinear partial differential equation, difficult to solve in general terms. However, approximate solutions can be obtained when assumptions on steady state or steady-rate conditions are applied with or without a linear relation between water content and diffusivity. Though applicable only in a limited range, these have considerably contributed to a better understanding of water transport to plant roots (Gardner, 1960; Passioura, 1980; De Willigen and Van Noordwijk, 1987). A particularly useful partial linearization of (3) is obtained by employing the matrix flux potential Φ ($m^2\ s^{-1}$) (Raats, 1970) defined as:

$$\Phi = \int_{\theta_{ref}}^{\theta} D_w \, d\theta \tag{10}$$

The steady-rate assumption then leads to the next equation giving the distribution of matrix flux potential as a function of radial distance:

$$\Phi = \Phi_{R0} + \frac{F_1}{2\pi \Delta z}\left(\frac{\rho^2}{\rho^2 - 1}\ln r - \frac{r^2 - 1}{2(\rho^2 - 1)}\right) \tag{11}$$

where $\rho = R_1/R_0$ and the dimensionless distance $r = R/R_0$.The distribution of water content around a root can now be calculated relatively simply with equation (11). Figure 4 shows the good agreement between the water content distribution calculated with the steady-rate approximation and with the numerical solution of equation (3).

The single root uptake model can be used to assess the relative importance of the soil resistance with respect to the root resistance. Figure 5 shows the plant and soil resistance as a function of the average pressure head for a root length density of $10^4\ m\,m^{-3}$, and a transpiration of $5 \times 10^{-3}\ m\,d^{-1}$ for a clay, sand, and peat soil in the Netherlands (hydraulic characteristics from Wösten, 1987), and a root conductance of $2.3 \times 10^{-7}\ m.d^{-1}$, the highest value found in literature (De Willigen and Van Noordwijk, 1987). Clearly, even with

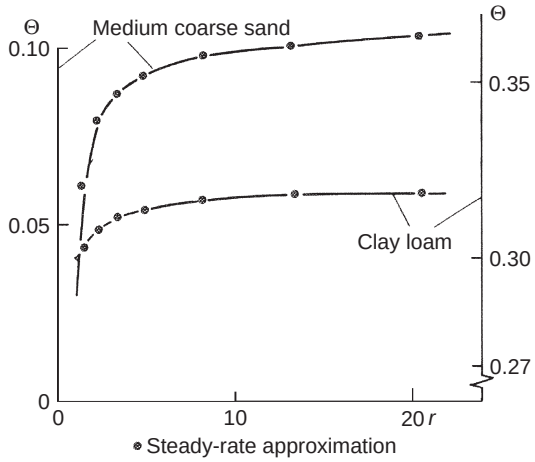


Figure 4 Distribution of the water content around a root taking when the pressure head at the root equals 5 Mpa, calculated with a numerical model (line) and with the steady-rate approximation (points) (equation (11) in the text)

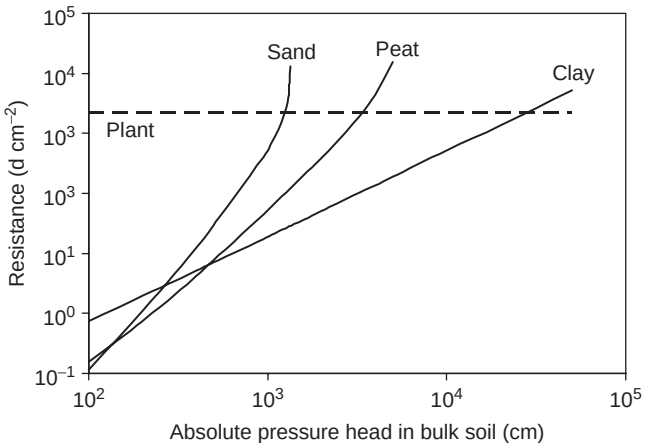


Figure 5 Soil and plant resistance as a function of average pressure head of the soil. Parameters: root radius 0.01 cm, root length 20 cm, root length density 1 cm cm^{-3}, root conductance 23×10^{-6} cm d^{-1}, transpiration 0.5 cm d^{-1}

this high value of the root conductance, the root resistance exceeds the soil resistance for h_p values up to -1000 cm (sand) or even $-30\,000$ cm (clay).

The model can also be used to examine the influence of root length. From a water uptake point of view it appears that doubling the total root length within a layer, for example, increasing the root length density with a factor two, is much less favorable than maintaining the same average root length density and doubling the root length (De Willigen *et al.*, 2000).

Root System Models

One can distinguish two types of uptake models. In the first group where root length density and root water potential do not play a role, uptake is determined by the depth of the root zone and the pressure head or water content of the soil layers (e.g. Feddes *et al.*, 1978; Van Genuchten, 1986; van Keulen and Seligman, 1987). The other is based upon steady state or steady-rate approximations of the single root model (e.g. Nimah and Hanks, 1973; Hillel *et al.*, 1976; De Willigen and Van Noordwijk, 1995).

An example of the first group is the water uptake module as used in the SWATRE model (Feddes *et al.*, 1978). First, a potential water uptake rate is calculated, which is reduced for unfavorable pressure heads in the soil. In its simplest form the potential water extraction rate ($S_{\max}$) is assumed to be equally distributed over the root zone (Z_r):

$$S_{\max} = \frac{E}{Z_r} \tag{12}$$

Later (Bouten, 1992), the potential uptake was distributed according to the relative distribution of the root length density and corrected for the water content distribution. For the ith layer in the soil profile the equation reads:

$$S_{\max,i} = \left[\frac{\frac{\theta_i}{\theta_{s,i}} L_{rv,i} \Delta z_i}{\sum \frac{\theta_i}{\theta_{s,i}} L_{rv,i} \Delta z_i} \right] \frac{E}{\Delta z_i} \tag{13}$$

where $\theta_{s,i}$ is the saturated water content of the layer. The term $\theta_i/\theta_{s,i}$ ensures that water is preferentially taken up from wetter layers. The actual uptake rate (S) in a certain layer is then reduced in the case of nonoptimal pressure head of that layer:

$$S_i = a(h_{p,i}) S_{\max,i} \tag{14}$$

De Willigen and Van Noordwijk (1995) used a water uptake model (WATUP) based on an extension of the single root model treated earlier. Again, for a root system distributed over several soil layers (or compartments), they assumed that within each layer (compartment) the roots are regularly distributed and that the flow of water from the bulk soil to the rhizosphere in each layer equals that from the rhizosphere into the root:

$$[\bar{\Phi}_i - \Phi_{R_0,i}] \frac{\pi \Delta z_i (\rho_i^2 - 1)}{G(\rho_i)} = \alpha (h_{p,R_0,i} - h_{p,P}) \tag{15}$$

where the subscript i denotes the number of the layer and the function $G(\rho_i)$ is given by:

$$G(\rho_i) = \frac{1 - 3\rho_i^2}{4} + \frac{\rho_i^4}{\rho_i^2 - 1} \ln(\rho_i) \tag{16}$$

Also the total uptake should equal the transpiration:

$$\sum_i^N L_{rv,i} \alpha (h_{p,R_0,i} - h_{p,P}) = E_{act} = f(h_{p,P}) E_{pot} \tag{17}$$

This yields $N + 1$ equations with $N + 1$ unknowns: the N values for $h_{p,R0,i}$ and the root pressure head $h_{p,P}$. The equations have to be solved by iteration because of the nonlinear relation between matrix flux potential and pressure head, and the nonlinear nature of $f(h_{p,P})$. For a relative wet soil the major resistance is found in the root (Figure 5), the gradients in the soil are accordingly very small, and $h_{p,P}$ is much smaller (more negative) than $h_{p,R0,i}$. Equation (17) for a given layer can then be approximated as:

$$L_{rv,i} \alpha (h_{p,R_0,i} - h_{p,P}) \approx -L_{rv,i} \alpha h_{p,P} \tag{18}$$

The uptake from a certain layer is thus proportional to the relative root length density in that layer. So for relatively wet conditions, both models yield the same distribution of uptake. Greater differences can be expected for dry conditions. The second model is able to simulate hydraulic redistribution.

Which type of model should be used depends on the conditions. If these are not too extreme the first type is to be preferred, it is simpler and requires fewer parameters. The second type is more flexible and seems therefore a better fit for more extreme circumstances.

MEASURING ROOT WATER UPTAKE

Weighing Lysimeter

Possibly, the most robust way to determine root water uptake is by means of a weighing lysimeter (Howell *et al.*, 1991). The change in weight of a well-defined volume of rooting medium is attributed to transpiration. The measurement can be very precise, say $0.05\,\mathrm{kg\,m^{-2}}$. The volume of rooted medium can range from say, one

liter up to several cubic meters. Instead of weighing the lysimeter, the mass balance method can be used (see Section "Water balance"). A special technique of weighing can be done by the so-called *hanging 'load cells' or hanging weighing scales* (e.g. Ehret *et al.*, 2001). Corrections may be needed for increase in plant fresh weight during the measurement interval. When combined with measurements of changes in water content in the rooting zone (see Section "Water balance") at several depths, differentiation of water uptake with depth can be achieved. For detailed studies, one can consider leaving out a porous rooting medium and letting the plant grow in water on a balance (e.g. the potometric water-budget meter of Flach *et al.*, 1995). The weighing methods are direct, robust, and very precise. However, the conditions in the rooting zone may differ from the surrounding crop, and large weighing lysimeters are expensive and sensitive to wind.

Water Balance

Root water uptake can be determined from a mass balance approach: root water uptake is equal to water input in the system minus water output from the system minus the increase in water storage in the system (e.g. Evett, 2002). This does not require a lysimeter (see Section "Weighing lysimeter"). The changes in water storage in the rooted medium can be followed by means of water content measurements, for example, dielectric methods such as time or frequency domain reflectometry (TDR, FDR), neutron scattering, and ground-penetrating radar (GPR) (*see* **Chapter 72, Measuring Soil Water Content, Volume 2**), or by means of pressure head measurements, for example, tensiometers, electrical resistance blocks, psychrometers. These techniques require calibration and knowledge of the water retention characteristic, which exhibits hysteresis. The water balance method is simple, and differentiation of water uptake with depth can be achieved when sensors are installed at different depths. The disadvantages of this method are that the water output may be difficult to measure (if no lysimeter is used), and that errors (including heterogeneity) in the determination of input, output, and change in storage accumulate in estimate of root water uptake.

Sap Flow

There are techniques to determine the sap flow rate through the stems (or even through individual roots) of plants. Three thermal tracer techniques are known: the heat-pulse method, the heat probe, and the heat balance technique. Smith and Allen (1996) reviewed principles and application of both the heat-pulse and the heat-balance techniques. Another method is the nuclear magnetic resonance (NMR) technique (Van As and Schaafsma, 1984). The advantages of measuring sap flow are: they give a direct indication of water flow through an individual plant; the measurement is rather precise compared to the weighing lysimeter technique (Weibel and De Vos, 1994 give a review of published accuracy tests); the required equipment is commercially available; and the method is nondestructive. However, not all stem diameters can be measured, possible heat-damage of the stem (heat flow) may occur, it requires calibration against another exact method (heat flow), and the sophisticated NMR equipment is heavy, voluminous, and expensive (except, maybe, for proton-NMR).

Inverse Modeling

Simulation models that describe water movement in porous media including a sink term can be used to estimate root water uptake for given conditions. Such an approach can be used when detailed information on root uptake is lacking and other detailed information is available for the other processes described by the model. Using inverse modeling techniques, the sink term (= root water uptake) in the model can be calibrated (e.g. Vogeler *et al.*, 2001; Vrugt *et al.*, 2001; Zuo and Zhang, 2002). The advantage of the inverse modeling technique is that root water uptake can be determined for situations where it is not directly measured. The disadvantages of this method are that it is a very indirect way of determining root water uptake, errors will propagate through errors in model input and model assumptions, and independent calibration is needed of the chosen root water uptake model.

Root Uptake Ability and Plant-water Status

The ability of roots to take up water is expressed by the root hydraulic conductance, and it can be measured by several techniques that have in common the ability to determine the volume of water flowing through a root that is exposed to a known gradient in water potential (e.g. Huang and Nobel, 1994; Fernández *et al.*, 2000).

The plant-water status can be measured by the following techniques: linear variable differential transducer (LVDT) that measures the changes in stem diameter (more flow results in thicker stems); porometer (transient or steady state) that measures stomatal resistance (or conductance) which indicates if stomata are fully open or not; psychrometer technique for measuring the plant (e.g. leaf, stem, root) water potential; and the pressure chamber, an alternative technique to measure plant-water potential. Advantages of the psychrometer technique are that it has a sound theoretical basis, it can be applied to small portions of plant tissue, and it can measure −40 bars or lower. However, this technique requires careful temperature control, and it has relatively long equilibrium times. The advantages of the pressure chamber are that the method is easy (but should be applied with some care) and quick, but it is not suitable

for small plant parts. A disadvantage of the LVDT technique is that it cannot be applied for small stem diameters. A disadvantage of the porometer is that it cannot be used for small leaves. Some general problems associated with measuring plant-water status have been described by, for example, Bennett (1990).

Additional Techniques

Computer assisted tomography applied to x-ray or γ-ray attenuation measurements can be used to show the spatial distribution of water contents in the soil surrounding a single root (e.g. Hamza and Aylmore, 1992; contributions in Anderson and Hopmans, 1994). Another technique to measure water content in the vicinity of roots (e.g. as close as 6 mm) is the dual-probe heat-pulse (DPHP) technique (e.g. Philip and Kluitenberg, 1999; Song *et al.*, 2000). Stable isotopes can be used to determine water uptake and to discriminate water taken up from different depths (Fernández *et al.*, 2000; Bingham *et al.*, 2000).

Split-root techniques can be used to study water movement within root systems or to study the reaction of the plant to highly variable conditions in the rooting zone. For example, such techniques can be used to demonstrate the existence of hydraulic redistribution or to study the effect of osmotic pressure on root water uptake.

Acknowledgment

Thanks are due to Dr. P.A.C. Raats for useful discussions on conductivity, conductance, and effects of build up of osmotic pressure in roots.

FURTHER READING

Bleby T.M., Burgess S.S.O. and Adams M.A. (2004) A validation, comparison and error analysis of two heat-pulse methods for measuring sap flow in *Eucalyptus marginata* saplings. *Functional Plant Biology*, **31**, 645–658.

Boyer J.S. (1969) Measurement of the water status of plants. *Annual Review of Plant Physiology*, **20**, 351–364.

Ferretti D.F., Pendall E., Morgan J.A., Nelson J.A., LeCain D. and Mosier A.R. (2003) Partitioning evapotranspiration fluxes from a Colorado grassland using stable isotopes: Seasonal variations and ecosystem implications of elevated atmospheric CO_2. *Plant and Soil*, **254**, 291–303.

Merta M., Sambale C., Seidler C. and Peschke G. (2001) Suitability of plant physiological methods to estimate the transpiration of agricultural crops. *Journal of Plant Nutrition and Soil Science*, **164**, 43–48.

Smit A.L., Bengough A.G., Engels C., Van Noordwijk M., Pellerin S. and Van De Geijn S.C. (Eds.) (2000) *Root Methods, A Handbook*, Springer-Verlag: Berlin.

Yip D.Y. (2003) Developing a better understanding of the relationship between transpiration and water uptake in plants. *Journal of Science Education and Technology*, **12**, 13–19.

REFERENCES

Aase J.K., Pikul J.L., Prueger J.H. and Hatfield J.L. (1996) Lentil water use and fallow water loss in a semiarid climate. *Agronomy Journal*, **88**, 723–728.

Anderson S.H. and Hopmans J.W. (Eds.) (1994) *Tomography of Soil-Water-Root Processes*, SSSA Special Publication Number 36, Soil Science Society of America: Madison.

Assmann S.M. (1993) Signal transduction in guard cells. *Annual Review of Cell Biology*, **9**, 345–375.

Bennett J.M. (1990) Problems associated with measuring plant water status. *HortScience*, **25**, 1551–1554.

Bingham I.J., Glass A.D.M., Kronzucker H.J., Robinson D. and Scrimgeour C.M. (2000) Isotope techniques. In *Root Methods, A Handbook*, Smit A.L., Bengough A.G., Engels C., Van Noordwijk M., Pellerin S. and Van De Geijn S.C. (Eds.), Springer-Verlag: Berlin, pp. 365–402.

Bolt G.H., Iwata S., Peck A.J., Raats P.A.C., Rode A.A., Vachaud G. and Voronin A.D. (1976) Final report second committee on terminology in soil physics. *ISSS Bulletin*, **49**, 16–22.

Bouten W. (1992) *Monitoring and Modelling Forest Hydrological Processes in Support of Acidification Research*, PhD Thesis, University of Amsterdam.

Buchanan B.B., Gruissem W. and Jones R.L. (2000) *Biochemistry and Molecular Biology of Plants*, American Society of Plant Physiologists: Rockville.

Burch G.J. (1979) Soil and plant resistances to water absorption by plant root systems. *Australian Journal of Agricultural Research*, **30**, 279–292.

Caldwell M.M., Dawson T.E. and Richards J.H. (1998) Hydraulic lift: consequences of water efflux from the roots of plants. *Oecologia*, **113**, 151–161.

Clothier B.E. and Green S.R. (1994) Rootzone processes and the efficient use of irrigation water. *Agricultural Water Management*, **25**, 1–12.

Clothier B.E. and Green S.R. (1997) Roots: the big movers of water and chemical in soil. *Soil Science*, **162**, 534–543.

Dainty J. (1985) Water transport through the root. *Acta Horticultura*, **171**, 21–31.

Dalton F.N., Raats P.A.C. and Gardner W.R. (1975) Simultaneous uptake of water and solutes by plant roots. *Agronomy Journal*, **67**, 334–339.

De Willigen P., Nielsen N.E., Claassen N. and Castrignanò A.M. (2000) Modelling water and nutrient uptake. In *Root Methods*, Smit A.L., Bengough A.G., Engels C., Van Noordwijk M., Pellerin S. and Van de Geijn S.C. (Eds.), Springer-Verlag: Berlin, pp. 510–539.

De Willigen P. and Van Noordwijk M. (1987) *Roots, Plant Production and Nutrient Use Efficiency*, PhD Thesis, Wageningen Agricultural University.

De Willigen P. and Van Noordwijk M. (1995) Model for interactions between water and nutrient uptake. In *Modelling and Parametrization of the Soil-Plant-Atmosphere System, A Comparison of Potato Growth Models*, Kabat P., Marshall B., Van de Broek B.J. and Van Keulen H. (Eds.), Wageningen Press: Wageningen.

Ehret D.L., Lau A., Bittman S., Lin W. and Shelford T. (2001) Automated monitoring of greenhouse crops. *Agronomie*, **21**, 403–414.

Esau K. (1965) *Plant Anatomy*, *Second Edition*, John Wiley & Sons: Chichester.

Evett S.R. (2002) Water and energy balances at soil-plant-atmosphere interfaces. In *Soil Physics Companion*, Warrick A.W. (Ed.), CRC Press: Boca Raton, pp. 127–188.

Feddes R.A., Kowalik P.J. and Zaradny H. (1978) *Simulation of Field Water Use and Crop Yield*, Simulation Monographs, PUDOC: Wageningen.

Fernández J.E., Clothier B.E. and Van Noordwijk M. (2000) Water Uptake. In *Root Methods, A Handbook*, Smit A.L., Bengough A.G., Engels C., Van Noordwijk M., Pellerin S. and Van De Geijn S.C. (Eds.), Springer-Verlag: Berlin, pp. 461–507.

Fischer R.A., Hsiao T.C. and Hagan R.M. (1970) After-effect of water stress on stomatal opening potential. I. Techniques and magnitudes. *Journal of Experimental Botany*, **21**, 371–385.

Fiscus E.L. (1975) The interaction between osmotic- and pressure-induced water flow in plant roots. *Plant Physiology*, **55**, 917–922.

Flach B.M.T., Eller B.M. and Egli A. (1995) Transpiration and water uptake of *Senecio medley-woodii* and *Aloe jucunda* under changing environmental conditions: measurements with a potometric water-budget-meter. *Journal of Experimental Botany*, **46**, 1615–1624.

Fry J. and Huang B. (2004) *Applied Turfgrass Science and Physiology*, Wiley: Hoboken.

Gardner W.R. (1960) Dynamic aspects of water availability to plants. *Soil Science*, **89**, 63–73.

Gardner W.R. (1970) Internal water status and plant response in relation to the external water régime. In *Plant Response to Climatic Factors*, Slatyer R.O. (Ed.), United Nations Educational, Scientific and Cultural Organisation: Paris, pp. 221–225.

Gardner W.R. (1983) Soil properties and efficient water use: an overview. In *Limitations to Efficient Water Use in Crop Production*, Taylor H.M., Jordan W.R. and Sinclair T.R. (Eds.), American Society of agronomy, Crop Science Society of America, Soil Science Society of America: Madison, pp. 45–64.

Gerwitz A. and Page E.R. (1974) An empirical mathematical model to describe plant root systems. *Journal of Applied Ecology*, **11**, 773–781.

Gradmann H. (1928) Untersuchungen über die Wasserverhältnisse des Bodens als Grundlage des Pflanzenwachtums. *I. Jahrbücher für wissenschaftkichen Botanik*, **69**, 1–100.

Hamza M.A. and Aylmore L.G. (1992) Soil solute concentration and water uptake by single lupin and radish plant roots. I. Water extraction and solute accumulation. *Plant and Soil*, **145**, 187–196.

Hillel D., Talpaz H. and Van Keulen H. (1976) A macroscopic-scale model of water uptake by a nonuniform root system and of water and salt movement in the soil profile. *Soil Science*, **121**, 242–255.

Howell T.A., Schneider A.D. and Jensen M.E. (1991) History of lysimeter design and use for evapotranspiration measurements. In *Lysimeters for Evapotranspiration and Environmental Measurements*, *Proceedings of the International Symposium on Lysimetry*, Honolulu Hawaii, Allen R.G., Howell T.A., Pruitt W.O., Walter I.A. and Jensen M.E. (Eds.), American Society of Civil Engineers: New York, pp. 1–9.

Huang B. and Nobel P.S. (1994) Root hydraulic conductivity and its components, with emphasis on dessert succulents. *Agronomy Journal*, **86**, 767–774.

Jackson R.B., Canadell J., Ehleringer J.R., Mooney H.A., Sala O.E. and Sculze E.D. (1996) A global analysis of root distributions for terrestrial biomes. *Oecologia*, **108**, 389–411.

Kieber J.J. (1997) The ethylene response pathway in Arabidopsis. *Annual Review of Plant Physiology and Plant Molecular Biology*, **48**, 277–296.

Kirkham M.B. (1983a) Physical model of water in a split-root system. *Plant and Soil*, **75**, 153–168.

Kirkham M.B. (1983b) Effect of ethephon on the water status of a drought-resistant and a drought-sensitive cultivar of winter wheat. *Zeitschrift Für Pflanzenphysiologie (Journal of Plant Physiology)*, **112**, 103–112.

Kirkham M.B. (2005) *Principles of Soil and Plant Water Relations*, Elsevier: Amsterdam, p. 500.

Kirkham M.B., Gardner W.R. and Gerloff G.C. (1974) Internal water status of kinetin-treated, salt-stressed plants. *Plant Physiology*, **53**, 241–243.

Kirkham M.B. and Kanemasu E.T. (1983) Wheat. In *Crop-Water Relations*, Teare I.D. and Peet M.M. (Eds.), Wiley: New York, pp. 481–520.

Leung J. and Giraudat J. (1998) Abscisic acid signal transduction. *Annual Review of Plant Physiology and Plant Molecular Biology*, **49**, 199–222.

Leyser O. (2002) Molecular genetics of auxin signaling. *Annual Review of Plant Biolology*, **53**, 377–398.

Livnè A. and Vaadia Y. (1965) Stimulation of transpiration rate in barley leaves by kinetin and gibberellic acid. *Physiologia Plantarum*, **18**, 658–664.

Livnè A. and Vaadia Y. (1972) Water deficits and hormone relations. In *Water Deficits and Plant Growth. III. Plant Responses and Control of Water Balance*, Kozlowski T.T. (Ed.), Academic Press: New York, pp. 255–275.

Luke H.H. and Freeman T.E. (1967) Rapid bioassay for phytokinins based on transpiration of excised oat leaves. *Nature*, **215**, 874–875.

Maurel C. (1997) Aquaporins and water permeability of plant membranes. *Annual Review of Plant Physiology and Plant Molecular Biology*, **48**, 399–429.

Maurel C. and Chrispeels M.J. (2001) Aquaporins. A molecular entry into plant water relations. *Plant Physiology*, **125**, 135–138.

Mees G.C. and Weatherley P.E. (1957) The mechanism of water absorption by roots. I. Preliminary studies on the effects of hydrostatic pressure gradients. *Proceedings of the Royal Society of London. Series B*, **147**, 367–380.

Miller D.M. (1985) Studies of root function in Zea mays. IV. Effects of applied pressure on the hydraulic conductivity and volume flow through the excised root. *Plant Physiology*, **77**, 168–174.

Mok D.W.S. and Mok M.C. (2001) Cytokinin metabolism and action. *Annual Review of Plant Physiology and Plant Molecular Biology*, **52**, 89–118.

Newman E.I. (1976) Interaction between osmotic- and pressure-induced water flow in plant roots. *Plant Physiology*, **57**, 738–739.

Nimah A. and Hanks R.J. (1973) Model for estimating soil water, plant, and atmospheric interrelations. I. Description and sensitivity. *Soil Science Society of America Proceedings*, **37**, 522–527.

Oki T. (1999) The global water cycle. In *Global Energy and Water Cycles*, Browning K.A. and Gurney R.J. (Eds.), Cambridge University Press: pp. 10–30.

O'Leary J.W. and Tarquinio Prisco J. (1970) Response of osmotically stressed plants to growth regulators. *Advances of Frontiers in Plant Science*, **25**, 129–139.

Passioura J.B. (1972) The effect of root geometry on the yield of wheat. *Australian Journal of Agricultural Research*, **23**, 745–752.

Passioura J.B. (1980) The transport of water from soil to shoot in wheat seedlings. *Journal of Experimental Botany*, **31**, 333–345.

Passioura J.B. (1988) Water transport in and to roots. *Annual Review of Plant Physiology and Plant Molecular Biology*, **39**, 245–265.

Philip J.R. and Kluitenberg G.J. (1999) Errors of dual thermal probes due to soil heterogeneity across a plane interface. *Soil Science Society of America Journal*, **63**, 1579–1585.

Raats P.A.C. (1970) Steady infiltration from line sources and furrows. *Proceedings of the Soil Science Society of America*, **34**, 709–714.

Raats P.A.C. (2005) Transport across single and series arrays of membranes. In *Proceedings of the IUTAM Symposium on Physicochemical and Electromechanical Interactions in Porous Media*, Huyghe J.M., Raats P.A.C. and Cowin A.C. (Eds.), Kluwer: Dordrecht, pp. 213–218.

Rachidi F., Kirkham M.B., Stone L.R. and Kanemasu E.T. (1993) Soil water depletion by sunflower and sorghum under rainfed conditions. *Agricultural Water Management*, **24**, 49–62.

Reinders E. (1957) *Leerboek De Algemene Plantkunde, Deel I, In-En Uitwendige Morfologie En Voortplanting Van Vaatplanten, Fourth Edition*, Scheltema & Holkema N.V.: Amsterdam.

Richards D.E., King K.E., Ait-ali T. and Harberd N.P. (2001) How gibberellin regulates plant growth and development: a molecular genetic analysis of gibberellin signaling. *Annual Review of Plant Physiology and Plant Molecular Biology*, **52**, 67–88.

Richards R.A. and Passioura J.B. (1981) Seminal root morphology and water use of wheat. II. Genetic variation. *Crop Science*, **21**, 253–255.

Rose É. (1985) *Water Relations of Winter Wheat. I. Genotypic Differences in Ethylene and Abscisic Acid Production During Drought Stress. II. Effect of a Dwarfing Gene on Root Growth, Shoot Growth, and Water Uptake in the Field*. Ph.D. dissertation. Kansas State University, Manhattan, Kansas. Dissertation Abstract No. DA8510242. Kansas State University Library Call No. LD 2668.D5 1985 R67.

Sharp R.E. and Davies W.J. (1985) Root growth and water uptake by maize plants in drying soil. *Journal of Experimental Botany*, **36**, 1441–1456.

Smith D.M. and Allen S.J. (1996) Measurement of sap flow in plant stems. *Journal of Experimental Botany*, **47**, 1833–1844.

Smith M., Burgess S.S.O., Suprayogo D., Lusiana B. and Widianto (2004) Uptake, partitioning and redistribution of water by roots in mixed-species agroecosystems. In *Below-Ground Interactions in Tropical Agroecosystems*, Van Noordwijk M., Cadish G. and Ong C.K. (Eds.), CABI Publishing: Cambridge, pp. 157–170.

Song Y., Kirkham M.B., Ham J.M. and Kluitenberg G.J. (2000) Root-zone hydraulic lift evaluated with the dual-probe heat-pulse technique. *Australian Journal of Soil Research*, **38**, 927–935.

Steudle E. and Frensch J. (1996) Water transport in plants: role of the apoplast. *Plant and Soil*, **187**, 67–79.

Steudle E. and Peterson C.A. (1998) How does water get through roots? *Journal of Experimental Botany*, **49**, 775–788.

Tal M. and Imber D. (1971) Abnormal stomatal behavior and hormonal imbalance in flacca, a wilty mutant of tomato. III. Hormonal effects on the water status in the plants. *Plant Physiology*, **47**, 849–850.

Tal M., Imber D. and Itai C. (1970) Abnormal stomatal behavior and hormonal imbalance in flacca, a wilty mutant of tomato. I. Root effect and kinetin-like activity. *Plant Physiology*, **46**, 367–372.

Taylor H.M. and Klepper B. (1978) The role of rooting characteristics in the supply of water to plants. *Advances of Agronomy*, **30**, 99–128.

Tyerman S.D., Bohnert H.J., Maurel C., Steudle E. and Smith J.A.C. (1999) Plant aquaporins: their molecular biology, biophysics and significance for plant water relations. *Journal of Experimental Botany*, **50**, 1055–1071.

Van As H. and Schaafsma T.J. (1984) Noninvasive measurement of plant water flow by nuclear magnetic resonance. *Biophysical Journal*, **45**, 469–472.

Van den Honert T.H. (1948) Water transport in plants as a catenary process. *Discussions of the Faraday Society*, **3**, 146–153.

Van Genuchten M.Th. (1986) *A Numerical Model for Water and Solute Movement in and Below the Root Zone*, U.S. Salinity Laboratory Research Report, U.S. Salinity Laboratory Research.

van Keulen H. and Seligman N.G. (1987) *Simulation of Water Use, Nitrogen Nutrition and Growth of a Spring Wheat Crop*, PUDOC: Wageningen, Simulation Monographs.

Vogeler I., Green S.R., Scotter D.R. and Clothier B.E. (2001) Measuring and modelling the transport and root uptake of chemicals in the unsaturated zone. *Plant and Soil*, **231**, 161–174.

Vrugt J.A., Hopmans J.W. and Šimůnek J. (2001) Calibration of a two-dimensional root water uptake model. *Soil Science Society of America Journal*, **65**, 1027–1037.

Waggoner P.E. (1966) Decreasing transpiration and the effect upon growth. In *Plant Environment and Efficient Water Use*, Pierre W.H., Kirkham D., Pesek J. and Shaw R. (Eds.), American Society of Agronomy and the Soil Science Society of America: Madison, pp. 49–72.

Weibel F.P. and De Vos J.A. (1994) Transpiration measurements on apple trees with an improved stem heat balance method. *Plant and Soil*, **166**, 203–219.

Wösten J.H.M. (1987) *Beschrijving Van de Waterretentie- En Doorlatendheidskarakteristieken uit de Staringreeks Met Analytische Functies*, Stiboka rapport 2019, Stiboka, Wageningen, p. 53.

WW2010 (1999) Department of Atmospheric Sciences (DAS) at the University of Illinois Urbana-Champaign (UIUC), **`http://ww2010.atmos.uiuc.edu/(Gh)/home.rxml`**.

Zuo Q. and Zhang R.D. (2002) Estimating root-water-uptake using an inverse method. *Soil Science*, **167**, 561–571.

71: Freezing and Thawing Phenomena in Soils

MANFRED STÄHLI

Swiss Federal Research Institute WSL; Water, Soil, and Rock Movements, Birmensdorf, Switzerland

More than one-third of the Earth's land surface is subjected to seasonal or permanent soil frost. The freezing and thawing of the soil may generate severe geotechnical, hydrological, and environmental problems, such as frost heaving, restricted infiltration, and enhanced overland flow. Each soil has its own specific freezing characteristic, primarily depending on texture and the prevailing solute concentration. Fine-textured soils tend to form ice lenses at freezingthat causes the soil to heave. In coarse-textured soil, on the other hand, the ice is more regularly distributed within the matrix. The behavior of the soil is always strongly linked to the prevailing meteorological conditions. The hydraulic conductivity of frozen soils is reduced, similar to very dry soils. Locally, this leads to reduced infiltration and delayed recharge to groundwater during snowmelt. For larger areas, for example, watersheds, the effect of frozen ground on infiltration and runoff is ambiguous, as a number of recent studies showed.

This article summarizes current theories and recent investigations on soil freezing and thawing phenomena with a major emphasis on implications for hydrology.

INTRODUCTION

When temperature in a soil drops below 0 °C, water starts to transform to ice. This phase change is termed as *soil freezing*, resulting in a state of *soil frost.* Approximately 35% of the Earth's land surface is significantly affected by soil freezing and its opposing process, the *thawing* of the soil (Williams and Smith, 1989). This area is referred to as the *periglacial* environment. Whereas in many places of the world the soil freezes seasonally only a few centimeters at the surface or down to 1 or 2 m in the soil, about 20–25% of the land surface (Péwé, 1991) – especially at northern latitudes and high altitudes – remains partially frozen year-round. Here we talk about *permafrost areas* (*see* **Chapter 172, Permafrost Hydrology, Volume 4**).

Substantial environmental, hydrological, and geotechnical problems are related to soil freezing and thawing. Roads and buildings can be damaged by the fact that some soils heave at freezing and collapse during thawing. Precipitation or snowmelt water may pond on a frozen soil surface causing lateral surface runoff and possible enhanced erosion. Water uptake by vegetation may be considerably delayed and reduced because of the presence of soil frost.

This article summarizes current theories and recent investigations on soil freezing and thawing phenomena with a major emphasis on implications for hydrology. Geotechnical aspects related to soil freezing, which have had a key role in soil frost research since the early 1930s, are only superficially considered.

SOIL FREEZING AND THAWING

In a soil the transition from water to ice does not take place at a specific temperature, but over a range of several degrees. This means that below 0 °C, heat loss simultaneously decreases the soil temperature (change in sensible heat) and transforms liquid water to ice (change in latent heat). We can formalize that in the general heat flow equation stating that for a given soil layer of thickness z, all changes in total heat storage during a time period t are balanced by the change in latent heat caused by freezing/thawing and the heat losses/inputs (q_h) to/from the adjacent soil:

$$\frac{\partial(CT)}{\partial t} = L_f \rho_{soil} \frac{\partial \theta_{ice}}{\partial t} + \frac{\partial}{\partial z}(-q_h) \qquad (1)$$

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

where C is the heat capacity, T is the temperature of the soil layer, L_f is the latent heat of freezing ($3.338\ 10^5\ \mathrm{J\,kg^{-1}}$), ρ_{soil} is the soil bulk density and θ_{ice} is the volumetric ice content. This general heat flow equation is complicated by the fact that the ratio between latent and sensible heat loss is not constant, but decreases with decreasing soil temperature. Soil texture and solute concentration determine this partitioning between sensible and latent heat. It is usual to express this soil specific property with the so-called *freezing characteristic curve*, that is, the decrease of the liquid water content below 0 °C (Figure 1). From various laboratory and field experiments it is well known that for clay soils the liquid water content decreases at lower temperatures than for sandy soils, and that even at temperatures below −5 °C a fraction of the total soil water remains unfrozen. Capillary and adsorptive forces of the soil particles that act on the water by reducing its pressure are largely responsible. From thermodynamic principles, we know that in order to maintain equilibrium a change in soil temperature must lead to a corresponding change in the pressure of all water phases (liquid water, vapor, and ice). Thus, there is a unique relationship between a decrease in temperature, $\mathrm{d}T$, and a change in soil-water pressure, $\mathrm{d}p_w$, according to the well-known Clausius–Clapeyron equation:

$$\frac{\mathrm{d}p_w}{\rho_w} - \frac{\mathrm{d}p_{ice}}{\rho_{ice}} = \int_{T_0}^{T} \frac{L_f}{T}\,\mathrm{d}T \qquad (2)$$

where p is pressure (relative to atmospheric pressure), ρ is density and the subscripts w and ice denote liquid water and soil ice. This implies that, as a rough approximation, a decrease in soil temperature by 1 °C lowers the soil-water potential by 1.22 MPa. (Note that equation (2) is strictly valid only if no additional forces act on the ice.) If we now assume that the water retention curve (i.e. the relationship between water pressure and liquid water content; *see* **Chapter 75, Determining Soil Hydraulic Properties,** **Volume 2**) is similar in an unfrozen and a frozen soil, we can explain the differences in freezing characteristic curves of different soil types. Spaans and Baker (1996) demonstrated experimentally that this similarity between freezing/thawing and drying/wetting is justified (Figure 2).

In addition to the soil mineral surfaces, solutes in the soil-water phase also contribute to the freezing point depression through the osmotic potential. In most natural soils, the concentration of dissolved salts is rather weak, lowering the freezing point by less than 0.1 °C. But for ion-rich soils it might be important to account for their concentration and mobility at freezing. Although it is generally known that solutes are excluded from the ice grid, there has been speculation whether they are excluded locally (e.g. in small pockets) or ejected ahead of the ice front. Recent measurements by Spaans and Baker (1997) and Overduin and Young (1997) tend to support local exclusion.

An interesting question to address is where exactly the liquid water, the ice, and the air is located in a partially frozen soil. For sand it is appropriate to adopt the capillary model for explaining the distribution of liquid water, ice, and air in an unsaturated frozen soil. Combining the Laplace surface tension equation, which relates pressures in two phases that meet at a curved interface (*see* **Chapter 73,**

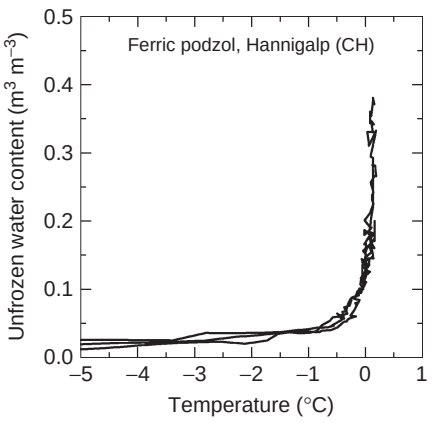


Figure 1 Freezing characteristic curve of a sandy loam (Data provided by M Stähli, WSL)

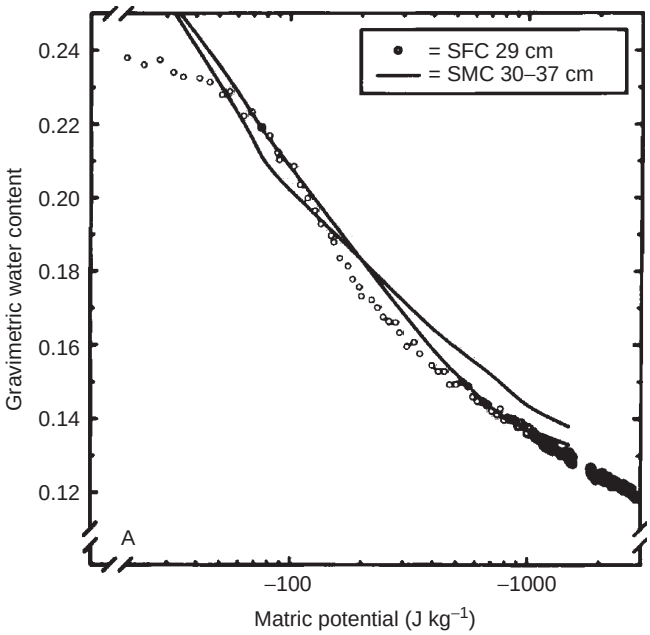
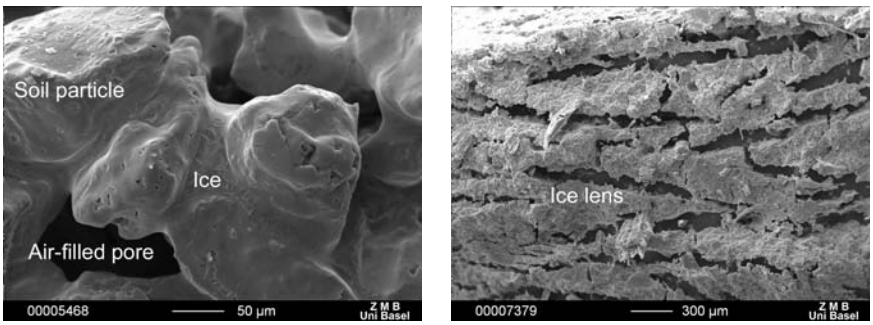


Figure 2 Similarity between freezing/thawing and drying/wetting. Nice agreement of the water retention curve gravimetrically measured in the lab on unfrozen silty loam samples (solid line) and calculated from its freezing characteristic curve (open circles) (Reproduced from Spaans and Baker, 1996; The solid freezing characteristic: its measurement and similarity to the soil moisture characteristic. Soil Science Society of America Journal, 60, page 17 by permission of Soil Science Society of America)



(a) (b)

Figure 3 Low-temperature SEM micrographs of a frozen unsaturated sand (a) and a frozen clay soil (b) (Images provided by D. Mathys, Center of Microscopy, University of Basel)

Soil Water Potential Measurement, Volume 2) with equation (2), provides the pore radius, r_{wi}, at the boundary between unfrozen water and ice:

$$r_{wi} = -\frac{2\sigma_{wi}T_f}{gL_f(T - T_f)} \tag{3}$$

where T_f is the freezing temperature of free water, σ is the surface tension, and g is the gravitational constant. Pores with a radius $< r_{wi}$ are filled with liquid water, whereas pores having a radius $> r_{wi}$ are occupied by ice. The boundary pore radius between ice and air, r_{ia}, can be given by the following expression:

$$r_{ia} = \frac{2\sigma_{ia}}{\rho_w g p_w(\theta_i)} \tag{4}$$

where σ_{ia} is the surface tension between ice and air ($\sim$100 mN m^{-1}; Miller, 1980) and θ_i is the initial water content at freezing. A low-temperature Scanning Electronic Microscope image (Figure 3) illustrates these theoretical considerations.

In a clay soil, unfrozen water surrounds the clay particles as an adsorbed layer, whereas the frozen water accumulates in horizontal bands of pure ice, the so-called *ice lenses* (Figure 3). The mechanisms behind the formation of such ice lenses have probably been the most investigated in frozen soil research during the last century. The theory of Miller (1980) defining the mechanisms for the development of an ice lens and the associated volume expansion, referred to as *frost heave*, is generally accepted as the most plausible model for freezing of clay soils. Air-filled voids in a frozen clay soil are restricted to macropores, such as cracks and root channels.

WATER MOVEMENT IN FROZEN SOILS

As demonstrated above, unfrozen water occupies the smallest pores because it is strongly adsorbed by the soil particles. If we transfer the assumption of similarity between freezing and drying to the hydraulic conductivity (k_w) - water content relationship (*see* **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**), a sharp decrease of k_w has to be expected as freezing sets in and progresses, resulting in a nearly immobile state at some degrees below 0 °C. Burt and Williams (1976) confirmed the validity of this assumption in an ingenious lab experiment. They measured the permeability of different frozen saturated soils at different temperatures closely below 0 °C using a permeameter in which water in two reservoirs was prevented from freezing by dissolved sugar. Their results suggest a decrease in hydraulic conductivity by 3–5 orders of magnitude from an unfrozen to a fully frozen state.

These results (from saturated frozen soils) conflict with a number of field studies and cold-chamber sprinkling experiments clearly demonstrating that under nonsaturated conditions liquid water very well may infiltrate into a frozen soil. There is strong evidence that for this "faster" water flow, the initially air-filled pores play a key role. Similar to low-permeable unfrozen soils, the largest pores, that is, cracks and macropores, become the dominant flow paths for infiltrating water unless they are blocked by ice (Seyfried and Murdock, 1997). The concept of water transfer through frozen unsaturated soil, assuming nearly immobile liquid water film in the smallest pores, and a faster water conducting flow domain in the largest (air-filled) pores, was implemented into a numerical model of soil water and heat transport by Stähli *et al.* (1996).

Whereas for saturated sandy soil it is plausible that water flow dominantly occurs through the liquid water film, it has become clear that for frozen clay soils with segregated ice lenses, additional water transport mechanisms have to be involved. In a classic lab experiment, Miller (1970) demonstrated that ice lenses are not impermeable to water redistribution, but allow water molecules to migrate through the ice, a process referred to *regelation*. The push for this redistribution of H_2O molecules through the ice arises from

pressure differences at both sides of the ice layer. A detailed discussion of this mechanism can be found in Miller (1980).

When describing water movement in frozen soils, what happens at the boundary between frozen and underlying unfrozen soil also needs to be considered. This transition zone is termed the *frozen fringe*. The steep gradient in water pressure difference between the frozen and the unfrozen soil layer induces upwards-directed water redistribution from the subsoil – especially when a shallow groundwater table is present. This uplift of water to the frozen fringe favors the formation of ice lenses and, thereby, forces the soil to heave. *Frost heaving* was earlier believed to be solely a consequence of volume expansion when water transforms to ice ($+9\%_{vol}$). Taber (1930) and Beskow (1935) were the first to show that more important for the initiation of frost heave was the redistribution of water towards the freezing front. Later, Miller (1978) postulated an additional mechanism responsible for the formation of new ice lenses and therefore further promoting frost heave, the so-called *rigid ice model*: Under a temperature gradient, segregated ice may move within the frozen fringe from warmer to colder regions. Soil particles are pushed apart when the ice pressure rises sufficiently, leading to a particle-free zone filled by ice, which is the newly developed ice lens.

INFILTRATION INTO FROZEN SOIL AND ITS SIGNIFICANCE FOR CATCHMENT RUNOFF

At the end of the winter season when the accumulated snow pack melts and releases a large amount of water to the soil surface, it is of particular concern to know whether this water may infiltrate or not. From the previous section, it can be expected that a frozen soil surface markedly reduces or even impedes water infiltration. This has been confirmed in numerous experiments of *local* infiltration into soil columns, plots, or fields. The effect of frozen ground on runoff from larger areas, however, is ambiguous, as a number of recent studies showed.

Local Infiltration into Frozen Soil

The infiltration behavior of frozen agricultural, forest, or alpine soils has been investigated extensively during the last 30 years in the field as well as in cold chambers. Several factors have been demonstrated to influence – directly or indirectly – the infiltration capacity of a frozen soil: soil texture, ice content, initial water content, soil surface management, soil temperature, the presence of a basal ice layer, the overlaying snowpack, and its melt rate, and so on. Many studies have shown concurrently that the total soil-water content at freezing is one of the key-factors governing the infiltration capacity of frozen soils. A dry soil in early winter clearly allows for considerable water infiltration, whereas nearly saturated frozen soils tend to build up a massive ice body and become nearly impermeable. This has been shown in field studies by Granger *et al.* (1984), who made direct *in situ* infiltration measurements at 90 sites in the Canadian Prairies, and Kane and Stein (1983) who used double-ring infiltrometers to determine infiltration rates into variable wet soils on a permafrost-free slope in Alaska. Also, a column study carried out by Stadler *et al.* (2000) in a cold chamber supported this hypothesis: two frozen sandy soil columns having different initial water contents were irrigated with two different dye tracer solutions during two successive 5-h periods. The columns were then sectioned vertically, and the distribution of the dye tracers analyzed. The depth profiles of the dye tracer concentration distribution in the two columns clearly confirmed the effect of the different initial water content.

For agricultural soils the treatment of the soil surface has a significant effect on the frozen soil infiltration, as many experimental studies have shown. Not only might furrows and cracks from tillage or ripping increase the macroporosity and improve water infiltration (Pikul *et al.*, 1990) but the presence of residue and stubble can also reduce the frost depth and thereby indirectly favor infiltration.

On sloped locations, inhibited soil infiltration may produce lateral surface or near-surface runoff associated with the risk for erosion. The partitioning between vertical soil percolation and lateral runoff in relation to the soil frost conditions and snowmelt release has recently been investigated on a boreal transect in northern Sweden (Stähli *et al.*, 2001), as well as in two studies in the Swiss Alps (Stadler *et al.*, 1996; Bayard *et al.*, 2005). The latter two studies nicely demonstrated the increase of lateral near-surface runoff with the presence of soil frost compared to unfrozen situations. The mechanisms for this enhanced lateral runoff appeared conspicuously in a dye tracer experiment by Stähli *et al.* (2004): An examination of vertical profiles, excavated in an early stage of the snowmelt, showed that the stained meltwater was concentrated in the uppermost 20 cm, demonstrating the impeding effect of the frozen zone (Figure 4). In addition, and probably even more important, was a basal ice layer that built up on the soil surface.

Large-scale Effects of Frozen Soil on the Hydrology

While the local influence of frozen soil on water infiltration is undisputed, it is much more difficult to prove its impact for catchment runoff or groundwater recharge at a larger scale. Most probably, this arises from the large spatial heterogeneity in frost occurrence and depth, as well as from reinfiltration of laterally displaced meltwater at lower areas. The heterogeneity of the frost distribution already becomes significant for water fluxes when we look at the runoff from single agricultural fields. Baker and Spaans (1997) provided an illustrative example on the Rosemont experimental field

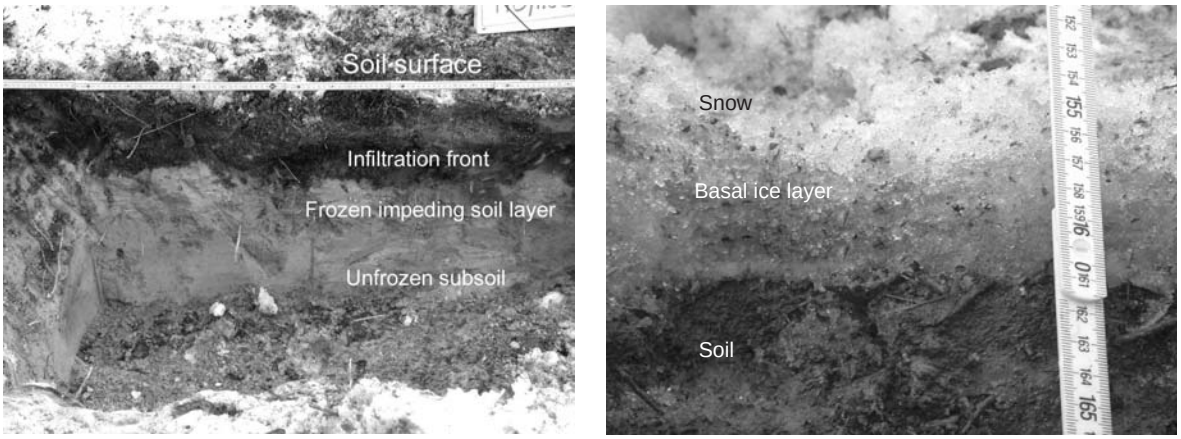


Figure 4 (a) Excavated vertical profile showing inhibited snowmelt infiltration into a frozen sandy alpine soil visualized with a dye tracer (dark color). The frost depth at this stage was approximately 45 cm. (b) Basal ice layer of a few centimeters built up at the base of the snowpack above frozen soil generating considerable lateral surface runoff (Images provided by D. Bayard, EPF Lausanne, and M. Stähli, WSL)

in Minnesota. During the spring melt, they first observed ponds of meltwater in local depressions of the field that suddenly drained in a short time although the soil was still considerably frozen. In a spatial examination of the frost distribution across the field, they found some single unfrozen locations in the depressions where the water obviously had bypassed the frozen soil layer. A closer analysis of the spatial distribution of soil frost across this relatively flat arable field (Cherkauer *et al.*, 2004) suggested that land use, soil moisture content, and soil properties were more important for the spatial pattern of soil frost than the thickness of the snowcover.

However, at a larger scale (e.g. watersheds) and especially in alpine areas with a more pronounced topography, the variation of the snowcover with its insulating effect becomes more dominant for the formation of soil frost. Unfortunately, studies on frost-depth distribution at the landscape scale are scarce.

Analysis of long-term records of frost depth and runoff from small forested catchments in Vermont (Shanley and Chalmers, 1999) and northern Sweden (Lindström *et al.*, 2002) did not support the hypotheses of faster runoff or increased peak flow with increasing frost depth. Also, Lindström *et al.* (2002) found that their hydrological runoff model, which did not account for soil freezing processes, did not simulate a systematic bias compared with runoff measurements in springs with deep soil frost. And finally, a study by Cherkauer and Lettenmaier (1999) demonstrated that runoff predictions for two subcatchments in the upper Mississippi River basin were not significantly improved by introducing a frozen soil algorithm that realistically simulated the frost-depth development.

All these published studies tend to imply that – in contrast to general belief, probably arising from local scale observations – catchment runoff is not significantly altered by the presence of a frozen soil.

FROST EFFECT ON OTHER SURFACE PROCESSES

Apart from soil infiltration and heaving, other important land surface processes are also affected by soil frost. It appears that during recent years, these processes – in the context of a changing climate – rather than the soil freezing and thawing mechanism itself have been addressed with more emphasis by the research community.

It has been recognized that water uptake by trees can be severely reduced by the presence of soil frost. This is of special concern for permafrost regions and boreal forest. Mellander *et al.* (2004) showed in a field experiment in northern Sweden that transpiration (*see* **Chapter 70, Transpiration and Root Water Uptake, Volume 2**) of 70-year old Scots pine trees on a deeply frozen soil plot was delayed by more than 1 month compared with a nonfrozen soil plot. The primary reason for inhibited water transport was not the lack of water availability, but rather the low soil temperature. In a similar study in New Hampshire in the United States, Tierney *et al.* (2001) observed larger fine root mortality on plots with deep soil frost, presumably because of mechanical damage from the pore ice, which is probably an additional reason for reduced water uptake.

Gas emission from the soil to the atmosphere seems to be strongly affected by frozen ground. Recent nitrous oxide emission measurements by Teepe *et al.* (2001) and van Bochove *et al.* (2001) have shown that accumulating gas is trapped below a frozen soil layer. As soon as the frozen layer melts, the trapped gases may escape to the atmosphere resulting in a considerable peak of nitrous oxide emission.

Thawing soils above a frozen subsoil have been shown to be most susceptible to surface erosion and rill or gully development (e.g. Oygarden, 2003). The high-surface saturation due to the low permeability of the frozen subsoil may temporally reduce the strength, increasing the potential for surface erosion. Surface residue cover seems to have a great impact in reducing frozen soil erosion (Cruse *et al.*, 2001).

NUMERICAL MODELS OF SOIL FREEZING AND FROZEN SOIL INFILTRATION

During the last 20 years, quite a large number of mesoscale meteorological models (*see* **Chapter 32, Models of Global and Regional Climate, Volume 1**), watershed runoff models (*see* **Chapter 132, Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change, Volume 3**), and erosion models (*see* **Chapter 82, Erosion Prediction and Modeling, Volume 2**) have been extended to mimic the presence of soil frost and its effect on surface processes (e.g. Bathurst and Cooley, 1996; Flanagan *et al.*, 2001; Mölders *et al.*, 2003).

But only a few numerical models have been developed with the principle goal of providing a detailed description of the combined water, heat, and solute fluxes in a freezing soil. The two models that probably have gained widest recognition are the Simultaneous Heat and Water (SHAW) Model (Flerchinger and Saxton, 1989) from the USDA Agricultural Research Service, and the Coupled heat and mass transfer (COUP) model for soil-plant-atmosphere system (former name: the SOIL-model; Jansson and Halldin, 1979; Jansson and Karlberg, 2001) developed at the Swedish University of Agricultural Sciences and KTH Stockholm. Both models simulate vertical water and heat fluxes in a layered (frozen or unfrozen) soil profile following the basic concept of the hydrodynamic model originally suggested by Harlan (1973). Boundary conditions, such as surface energy balance, snowpack, evapotranspiration, groundwater, or deep percolation are treated in detail by both models using standard meteorological data as the driving inputs. The models have been used for agricultural applications, as well as for studying the effect of winter conditions (possibly changing with a warming climate) on runoff and solute transport. The SHAW model and its user manual can be downloaded from **http://www.nwrc.ars.usda.gov/Models/SHAW/.**

The COUPModel includes a user-friendly interface and a comprehensive help manual. It can be downloaded from **http://www.lwr.kth.se/Coup.htm.**

For hydrological applications at the landscape scale, Cherkauer and Lettenmaier (2003) introduced a snow and frozen soil algorithm into the Variable Infiltration Capacity (VIC) macroscale hydrologic model (Liang *et al.*, 1994), allowing spatial variation in soil ice content. The code of the modified VIC macroscale hydrologic model is available at **http://danpatch.ecn.purdue.edu/~cherkaue/software.htm.**

Other numerical models of water and heat transport in a partly frozen soil have been developed by Zhao *et al.* (1997) in Canada, Gusev and Nasonova (2002) in Russia, and Ippisch (2002) in Germany. However, to the author's knowledge these models are not available for public use.

Acknowledgments

I would like to gratefully acknowledge the comments of Per–Erik Jansson, KTH Stockholm and John Baker, University of Minnesota, as well as the assistance of Melissa Swartz and Brian Mc Ardell in proofreading the manuscript. Financial support was provided by Swiss Federal Research Institute WSL.

FURTHER READING

Iskandar I.K., Wright E.A., Radke J., Sharratt B.S., Groenevelt P.H. and Hinzman L.D. (1997) *International Symposium on Physics, Chemistry, and Ecology of Seasonally Frozen Soils*, Fairbanks, June 10–12 1997, CRREL Special Report 97-10, p. 574.

Lundin L.-C. (1990) Hydraulic properties in an operational model of frozen soil. *Journal of Hydrology*, **118**, 289–310.

SOFTWARE LINKS TEXTBOX

Name of the software	Main author of the software	Institute	URL-site
CoupModel	Per – Erik Jansson	KTH, Stockholm, Sweden	**http://www.lwr.kth.se/Coup.htm**
SHAW	Gerald Flerchinger	Northwest Watershed Research Center, Boise, ID, USA	**http://www.nwrc.ars.usda.gov/Models/SHAW/.**
VIC macroscale hydrologic model	Keith Cherkauer	Purdue University, West Lafayette, IN, USA	**http://danpatch.ecn.purdue.edu/~cherkaue/software.htm**

REFERENCES

Anderson E.A. (1976) *A Point Energy and Mass Balance of a Snow Cover*, NOAA Technical Report No. NWS 19, United States Department of Commerce, National Oceanic, and Atmosphere Administration, National Weather Service, Silver Spring, MD, p. 150.

Baker J.M. and Spaans E.J.A. (1997) Mechanics of meltwater movement in seasonally frozen soil. In *International Symposium on Physics, Chemistry, and Ecology of Seasonally Frozen Soils*, Iskandar I.K., Wright E.A., Radke J., Sharratt B.S., Groenevelt P.H. and Hinzman L.D. (Eds.), Fairbanks: Alaska, June 10–12, 1997, pp. 31–36.

Bathurst J.C. and Cooley K.R. (1996) Use of the SHE hydrological modelling system to investigate basin response to snowmelt at Reynolds Creek. *Idaho Journal of Hydrology*, **175**(1–4), 181–211.

Bayard D., Stähli M., Parriaux A. and Flühler H. (2005) The influence of seasonally frozen soil on the snowmelt runoff at two Alpine sites in southern Switzerland. *Journal of Hydrology*, (in press).

Beskow G. (1935) *Tjälbildning Och Tjällyftningen Med Särskild Hänsyn Till Vägar Och järnvägars*. Årsbok 26, (1932) No 3, Sveriges Geologiska Undersökning: Stockholm.

Burt T.P. and Williams P.J. (1976) Hydraulic conductivity in frozen soils. *Earth Surface Processes*, **1**(3), 349–360.

Cherkauer K.A., Baker J.M. and Lettenmaier D.P. (2004) Field observations of the spatial distribution of snow and frozen soil. *Water Resources Research*, (in press).

Cherkauer K.A. and Lettenmaier D.P. (1999) Hydrologic effects of frozen soils in the upper Mississippi River basin. *Journal of Geophysical Research*, **104**(D16), 19599–19610.

Cherkauer K.A. and Lettenmaier D.P. (2003) Simulation of spatial variability in snow and frozen soil. *Journal of Geophysical Research*, **108**(D22), 8858, doi:10.1029/2003JD003575.

Cruse R.M., Mier R. and Mize C.W. (2001) Surface residue effects on erosion of thawing soils. *Soil Science Society of America Journal*, **65**, 178–184.

Flanagan D.C., Ascough J.C. II, Nearing M.A. and Laflen J.M. (2001) The Water Erosion Prediction Project (WEPP) model. In *Landscape Erosion and Evolution Modeling*, Harmon R.S. and Doe W.W. III (Eds.) Kluwer Academic Publishers: Norwell, p. 51.

Flerchinger G.N. and Saxton K.E. (1989) Simultaneous heat and water model of a freezing snow-residue-soil system, I. Theory and development. *Transaction of the ASAE. American Society of Agricultural Engineers*, **32**, 565–571.

Granger R.J., Gray D.M. and Dyck G.E. (1984) Snowmelt infiltration to frozen Prairie soils. *Canadian Journal of Earth Sciences*, **21**, 669–677.

Gusev Y.M. and Nasonova O.N. (2002) The simulation of heat and water exchange at the land-atmosphere interface for the boreal grassland by the land-surface model SWAP. *Hydrological Processes*, **16**, 1893–1919.

Harlan R.L. (1973) Analysis of coupled heat-fluid transport in partially frozen soil. *Water Resources Research*, **9**, 1314–1323.

Ippisch O. (2002) *Coupled Transport in Natural Porous Media*, Dissertation. Rupertus Carola University of Heidelberg, p. 144, **http://www.ub.uni-heidelberg.de/archiv/1872**.

Jansson P.-E. and Halldin S. (1979) Model for the annual water and heat flow in a layered soil. In *Comparison of forest and energy exchange models, Int. Soc. Ecol. Modelling*, Halldin S. (Ed.), Copenhagen: Denmark, pp. 145–163.

Jansson P.-E. and Karlberg L. (2001) Coupled heat and mass transfer model for soil-plant-atmosphere systems. TRITA-AMI Report, 30 87, ISSN 1400-1306, KTH: Stockholm.

Kane D.L. and Stein J. (1983) Water movement into seasonally frozen soils. *Water Resources Research*, **19**, 1547–1557.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, **99**, 14415–14428.

Lindström G., Bishop K. and Ottosson-Löfvenius M. (2002) Soil frost and runoff at Svartberget, Northern Sweden – measurements and model analysis. *Hydrological Processes*, **16**, 3379–3392.

Mellander P.-E., Bishop K. and Lundmark T. (2004) The influence of soil temperature on transpiration: a plot scale manipulation in a young Scots pine stand. *Forest Ecology and Management*, **195**, 15–28.

Miller R.D. (1970) Ice sandwich: Functional semipermeable membrane. *Science*, **169**, 584–585.

Miller R.D. (1978) Frost heaving in non-colloidal soils. *Proceedings of the Third International Conference on Permafrost*, Vol. 1, N.R.C.C: Edmonton, Ottawa, pp. 708–713.

Miller R.D. (1980) Freezing phenomena in soils. In *Application of Soil Physics*, Hillel D. (Ed.), Academic Press: pp. 254–299.

Mölders N., Haferkorn U., Döring J. and Kramm G. (2003) Long-term numerical investigations on the water budget quantities predicted by the hydro-thermodynamic soil vegetation scheme (HTSVS) – Part I: description of the model and impact of long-wave radiation, roots, snow, and soil frost. *Meteorology and Atmospheric Physics*, **84**(1–2), 137–156.

Overduin, P.P. and Young K.L. (1997) The effect of freezing on soil moisture and nutrient distribution at Levison-Lessing Lake, Taymyr Peninsula, Siberia. In *International Symposium on Physics, Chemistry, and Ecology of Seasonally Frozen Soils*, Iskandar I.K. *et al.* (Ed.), U.S. Army Cold Regions Research and Engineering Laboratory (CRREL): Hanover, NH, USA Fairbanks, June 10–12 1997, pp. 327–333.

Oygarden L. (2003) Rill and gully development during an extreme winter runoff event in Norway. *Catena*, **50**, 217–242.

Péwé T.L. (1991) Permafrost. In *The Heritage of Engineering Geology: The First Hundred Years*, *Centennial Special Vol. 3*, Kiersch G.A. (Ed.), Geological Society of America: Boulder, pp. 277–298.

Pikul J.L. Jr, Zuzel J.F. and Ramig R.E. (1990) Effect of tillage induced macroporosity on water infiltration. *Soil and Tillage Research*, **17**, 153–165.

Seyfried M.S. and Murdock M.D. (1997) Use of air permeability to estimate infiltrability of frozen soil. *Journal of Hydrology*, **202**, 95–107.

Shanley J.B. and Chalmers A. (1999) The effect of frozen soil on snowmelt runoff at Sleepers Reiver, Vermont. *Hydrological Processes*, **13**, 1843–1857.

Spaans E.J.A. and Baker J.M. (1996) The soil freezing characteristic: Its measurement and similarity to the soil

moisture characteristic. *Soil Science Society of America Journal*, **60**, 13–19.

Spaans E.J.A. and Baker J.M. (1997) Examining the use of time domain reflectometry in frozen soil. In *International Symposium on Physics, Chemistry, and Ecology of Seasonally Frozen Soils*, Iskandar I.K. *et al.* (Ed.), U.S. Army Cold Regions Research and Engineering Laboratory (CRREL): Hanover, NH, USA Fairbanks, June 10–12 1997, pp. 565–570.

Stadler D., Stähli M., Aeby P. and Flühler H. (2000) Dye tracing and image analysis for quantifying water infiltration into frozen soils. *Soil Science Society of America Journal*, **64**, 505–516.

Stadler D., Wunderli H., Auckenthaler A., Flühler H. and Bründl M. (1996) Measurement of frost-induced snowmelt runoff in a forest soil. *Hydrological Processes*, **10**, 1293–1304.

Stähli M., Bayard D., Wydler H. and Flühler H. (2004) Snowmelt infiltration into alpine soils visualized by dye tracer technique. *Arctic Antarctic and Alpine Research*, **36**(1), 128–135.

Stähli M., Jansson P.-E. and Lundin C. (1996) Preferential water flow in a frozen soil – a two-domain model approach. *Hydrological Processes*, **10**, 1305–1316.

Stähli M., Nyberg L., Mellander P.-E., Jansson P.-E. and Bishop K. (2001) Soil frost effects on soil and runoff dynamics along a boreal transect: 2. Simulations. *Hydrological Processes*, **15**, 927–941.

Taber S. (1930) The mechanics of frost heaving. *Journal of Geology*, **38**(4), 303–317.

Teepe R., Brumme R. and Beese F. (2001) Nitrous oxide emissions from soil during freezing and thawing periods. *Soil Biology and Biochemistry*, **33**, 1269–1275.

Tierney G.L., Fahey T.J., Groffman P.M., Hardy J.P., Fitzhugh R.D. and Driscoll C.T. (2001) Soil freezing alters fine root dynamics in a northern hardwood forest. *Biochemistry*, **56**, 175–190.

van Bochove E., Theriault G., Rochette P., Jones H.G. and Pomeroy J.W. (2001) Thick ice layers in snow and frozen soil affecting gas emissions from agricultural soils during winter. *Journal of Geophysical Research*, **106**(D19), 23061–23071.

Williams P.J. and Smith M.W. (1989) The frozen earth. *Fundamentals of Geocryology*, Cambridge University Press, p. 306.

Zhao L.T., Gray D.M. and Male D.H. (1997) Numerical analysis of simultaneous heat and mass transfer during infiltration into frozen ground. *Journal of Hydrology*, **200**(1–4), 345–363.

72: Measuring Soil Water Content

G CLARKE TOPP[1] AND TY PA FERRÉ[2]

[1] *Agriculture and Agri-Food Canada, Ottawa, ON, Canada*
[2] *University of Arizona, Tucson, AZ, US*

Major advances in the measurement of soil water content have arisen from electromagnetic (EM) methods that have developed rapidly in the last 20 years. Estimates of water content from EM measurements make use of the large relative permittivity of water compared to other soil components. Time domain reflectometry (TDR) and capacitance approaches use "probes" that convey signal into the soil and thus can measure principally the upper one-meter depth. Ground penetrating radar (GPR) using noninvasive, transmitting, and receiving antennae possesses the capability to measure to even greater depths without causing soil disturbance. Remote radar and passive microwave methods, operating generally above 1 GHz, derive their information from within a few centimeters of the ground surface. Thermogravimetric and neutron moderation continue as viable long-standing methods but are being used less as these methods are not amenable to data-logging. The variety of instruments has increased the surface and near-surface soil water measurement capabilities. Now it is possible for hydrologists to make informed choices among methods, and it is important to do so to optimize their study results.

INTRODUCTION

The measurement of soil water content has undergone revolutionary advancements in the last 20 years. From having gravimetric sampling and neutron moderation as the primary field methods in the early 1980s, we now have numerous options, such as time domain reflectometry (TDR), capacitance (and impedance) devices, ground penetrating radar (GPR), airborne/satellite active radar, and passive microwave methods. These five newer methods are all based on electromagnetic (EM) measurements. The development of improved understanding of microwave interaction in soil was part of the basis for these advances, but could only be brought into practice with the accompanying development of instrumental capability for effective measurement at frequencies above 10 MHz and improved data handling capabilities for large data sets. All of the EM methods make use of the high relative permittivity (dielectric constant) of the water (80) in soil compared with the permittivities of the other soil components, which range from one for air to three to five for typical soil solids. Owing to this contrast, methods that measure the bulk dielectric permittivity of soil are effective for the measurement of volumetric water content.

The diversity of EM methods provides a great range of measurement opportunities at or near the soil surface. Capacitance and TDR methods allow for alteration of the configuration of their embedded probes to meet specific measurement requirements within the soil. GPR methods generate a signal whose interaction with the soil is detected with a dedicated receiving antenna and interpreted for estimation of water content. GPR, especially when used within boreholes, can be used to profile the volumetric water content to tens of meters depth. Remote radar from airborne or satellite platforms uses transmitter and receiver to obtain a reflection from the soil surface giving an estimation of water content. Passive microwave utilizes radiation emanating from the soil as a measure of the water content of the top few centimeters. Standard non-EM methods complement these EM methods. Neutron moderation techniques operate primarily below 20 cm depth, and the gravimetric method may be used from wherever appropriate samples can be removed for later laboratory analysis. In total, the hydrologist has available a variety of water content measurement possibilities from which to choose for specific applications. This article reviews the basis and capability of each method with a view to offering assistance when making choices among these water content measurement methods.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Several detailed procedural reviews of these methods are available for further reference (e.g. Noborio, 2001; Topp and Ferré, 2001; Hendrickx *et al.*, 2002; Huisman *et al.*, 2003a; Jackson *et al.*, 1996; Gardiner *et al.*, 2001; Topp and Ferré, 2002; Robinson *et al.*, 2003; Jones *et al.*, 2002; Wraith, 2003).

THE WAVE EQUATION FRAMEWORK FOR EM METHODS IN SOIL

As the majority of current water content measurement methods rely on the measurement of EM properties, a brief review of EM wave propagation follows as a framework for comparing these methods. Specifically, for in-field use, there are five well-developed EM techniques operating between frequencies of 10 MHz and 10 GHz. These rapidly developing methods for measurement of soil water content have their bases in an understanding of the propagation of EM waves in space or on a transmission line. The equation describing the transmission of an EM wave along a transmission line has the following form in the general case of a conductive medium (Kraus, 1984):

$$V(z) = V_1 e^{\alpha z} e^{j(\omega t + \beta z)} + V_2 e^{-\alpha z} e^{j(\omega t - \beta z)} \qquad (1)$$

where V is the voltage between the transmission line conductors, z is the propagation direction, V_1 and V_2 are constants, t is time, ω is the angular frequency, α is the attenuation coefficient, and β is the phase constant. In soils, the quantities α and β are dependent upon the water content and electrical conductivity of the soil and, therefore, serve as a useful indirect measurement of these properties. Other useful parameters for characterizing soil physical properties can be derived from equation (1), such as phase velocity, v, characteristic impedance, Z_0, and voltage reflection coefficient, ρ. Incorporating the electromagnetic properties of soils gives the following equations that form the basis of the EM methods presented here. The velocity of propagation of an EM wave is defined as (Von Hippel, 1954):

$$v = \frac{c}{\sqrt{\frac{\varepsilon_r'}{2}\left[1 + \sqrt{1 + \left(\frac{\varepsilon_r'' + \sigma_0/\omega\varepsilon_0}{\varepsilon_r'}\right)^2}\right]}} \approx \frac{c}{\sqrt{\varepsilon_{ra}}} \qquad (2)$$

where c is the velocity of propagation of light in free space, ε_r' and ε_r'' are the real and imaginary components of the relative permittivity or dielectric constant, respectively, σ_0 is the dc electrical conductivity, ε_0 is the permittivity of free space, and ε_{ra} is the apparent relative permittivity.

Although the soil electrical conductivity has offered some complication for EM methods, it can also be estimated from some measurements, such as TDR. This provides an extra characterization of the hydrological regime, which can be useful in monitoring processes such as soluble salt transport. The right-most approximation in equation (2) recognizes that conductivity is present in soil, but in the methods to be presented in the following section, the effects of conductivity are generally assumed negligible. The characteristic impedance and reflection coefficient are defined as:

$$Z_0 = 120\pi \sqrt{\frac{1}{\varepsilon_r' - j\left(\varepsilon_r'' + \frac{\sigma_0}{\omega\varepsilon_0}\right)}} \approx \frac{120\pi}{\sqrt{\varepsilon_{ra}}} \qquad (3)$$

where $j = \sqrt{-1}$, and

$$\rho = \frac{Z_L - Z_0}{Z_L + Z_0} \qquad (4)$$

where Z_L is the load impedance (of the soil, in our case). Although the reflection coefficient is presented here for a transmission line, it applies equally to reflections from boundaries in media for the radar backscatter and ground penetrating radar methods.

The Electrical Properties of Water – Pure and in Soil

The unique electrical properties of water, both pure and in soil, form the basis of indirect electromagnetic water content measurements in soil. The relative dielectric permittivity of water is generally more than an order of magnitude larger than that of other soil components. As a result, the bulk dielectric permittivity of a soil was found to be a function of the volumetric water content, with only a slight dependence on the volume fraction of solids and electrical conductivity (Topp *et al.*, 1980). In each of the five EM methods presented here, a measurement of the relative permittivity of the soil is used to infer the volumetric water content. In Figure 1, the frequency bands of these instruments have been superimposed on the relative permittivity of pure water as a function of frequency. The figure shows that for those methods that operate at frequencies below 1 GHz, ε_r' of pure water is constant over their entire operating range. The operating assumption underlying the inference of water content from dielectric measurements on soils is that the majority of the water in soil retains a relative permittivity very similar to that of pure water. The figure also shows that those methods that operate at lower frequencies also avoid the frequency band at which ε_r'' increases to a maximum. This imaginary part of the dielectric permittivity describes the dielectric loss that arises due to mechanisms of absorption such as molecular rotations, molecular vibrations, or electronic transitions. The dc electrical conductivity, σ_0, (see equation 2) adds to dielectric losses to

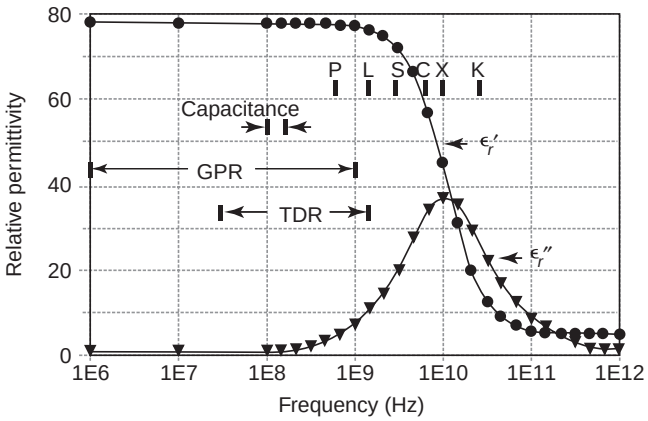


Figure 1 The real, ε_r', and imaginary, ε_r'', components of relative permittivity of water at 25 °C as a function of frequency. The operating frequencies of microwave sensors are designated by the letter band identifiers. Frequency range of other methods are also given (Reproduced from Topp and Ferré (2002) by permission of Soil Science Society of America)

produce greater attenuation of an EM signal travelling in a soil. For most nonclay, nonsalinized soils, these conductivity and dielectric loss effects are negligible, allowing for accurate measurement of the volumetric water content with EM methods operating between 30 MHz and 1 GHz.

There have been numerous attempts to establish the nature of the relationship between ε_r' and the volumetric water content to use as a basis for calibrations. Later improvements or refinements to the earlier empirical relationships have used dielectric mixing formulae that require more prior knowledge of the soil, such as density, texture, and/or organic content. The dielectric mixing formulae and related experimental work have shown a robust linear relationship between the volumetric water content, θ_v, and $\sqrt{\varepsilon_r'}$ is applicable for a wide range of soils (Topp and Reynolds, 1998):

$$\theta_v = 0.115\sqrt{\varepsilon_r'} - 0.176 \qquad (5)$$

This relationship deviates by less than 0.01 $m^3\,m^{-3}$ from the polynomial equation introduced by Topp *et al.* (1980) over the range 0.05–0.45 $m^3\,m^{-3}$. Equation (5) may be used for methods that measure relative permittivity (dielectric constant) as an approximation of ε_r'. We prefer to use ε_{ra} rather than ε_r' in equation (5), because it indicates clearly that there is an implicit assumption that electrical conductivity and/or dielectric loss influences are ignored.

TWO INTRUSIVE METHODS

The TDR and capacitance (and impedance) methods depend on probes to guide the signal through the soil being measured. Both methods have made use of parallel rods inserted into the soil from the surface to serve as the waveguide. Capacitance devices are also configured where both capacitor electrodes are on a single rod. In this case, the electrodes are as a sequence of parallel circumferential rings spaced along plastic pipes in pairs. The pipes are installed in the soil, usually inside another concentric plastic pipe. The EM fields fringing through the plastic interact with the soil. Similarly, TDR probes have been designed to measure through plastic sheaths.

TDR

Over the past 30 years, TDR has been used to measure water content at many scales and under a broad range of conditions (Topp and Reynolds, 1998; Robinson *et al.*, 2003) and has become a standard method of water content measurement second only to the thermogravimetric method. Most TDR instruments launch a fast rise voltage step (rise time < 200 ps) along a transmission line buried in the soil or medium of interest. The voltage pulse propagates as a planar EM wave, travelling in the soil and guided by the conductors. This waveguide is generally referred to as the TDR probe (Figure 2). As discussed below, these probes may have a variety of configurations. The properties of the soil that govern the propagation of the TDR pulse are described collectively by the propagation constant of the soil from which velocity of propagation (v) is derived (Ferré and Topp, 2002). The velocity, v, is entered into equations (2) and (5) to give the volumetric water content of the soil through which the EM wave has propagated.

The configuration of the waveguide or probe determines the extent and shape of the measured soil sample (Robinson *et al.*, 2003). The basic elements are conductive components, often parallel metallic rods, which act as waveguides and the soil material in which the wave or signal propagates. Currently, the most common soil probes are of the parallel rod type. The balanced pair transmission line, consisting of two parallel rods, rapidly became the probe of choice for use in field measurements (Figure 2). These usually varied in length, depending on the measurement requirement, from 0.1 to 1.0 m, and with rod separations from 0.01 to 0.1 m. In multipronged probes, one prong or wire is centrally located and variable numbers of prongs are located circumferentially around the central wire. These configurations, even with only two outer prongs, emulate a coaxial transmission line and result in a marginally improved TDR reflection over those from the balanced pair configuration. The extra rod(s), however, make for greater installation difficulty and associated soil disturbance compared with a parallel rod-pair. An important design parameter for accurate water content determination is the probe length. The minimum practical probe length for standard equipment is 10 cm. The upper limit on length of probe is largely determined by electrical conductivity, clay content, and maximum water content expected. The cross-sectional dimensions of TDR

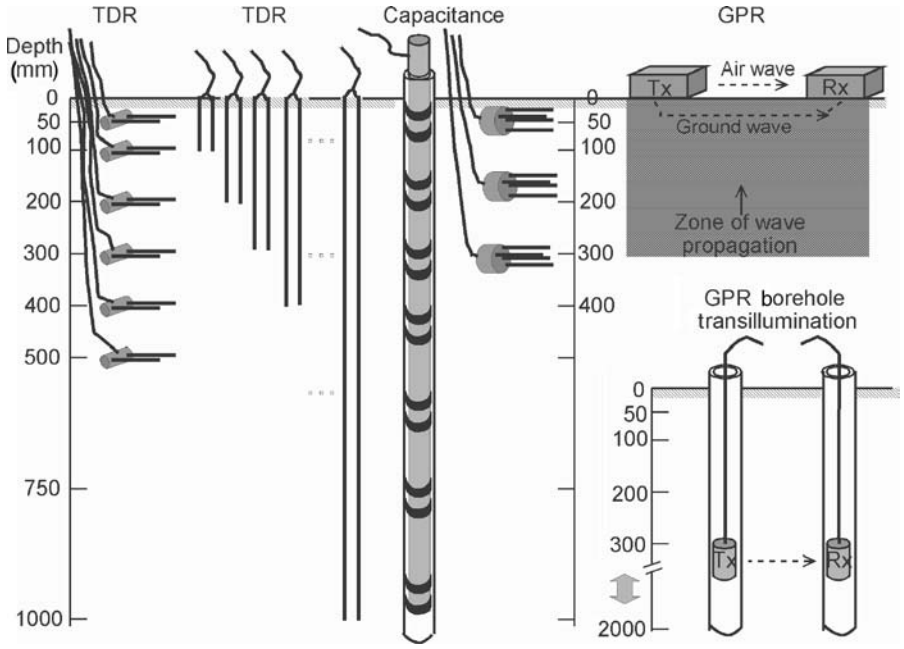


Figure 2 Schematic representations of two configurations each of TDR, capacitance, and GPR methods

probes, that is, the rod diameters and their separation, are at least as important as length considerations. Ideally, the EM energy should propagate in the soil well away from the regions that may be disturbed by inserting the rods. Ferré and Topp (2002) and Robinson *et al.* (2003) have provided more specific criteria affecting probe performance and the selection of probes. Detailed procedures for TDR measurement have been given by Ferré and Topp (2002).

An important application of TDR relates to the measurement of soil water content profiles or the vertical distributions to 1.5 m depth. Three general approaches for water content profiling are now available. Each offers certain advantages along with some limitations. Probes of sequentially increasing depths, inserted from the surface and near to each other, provide the total water to each depth (Figure 2). From these, one compares adjacent depths to get the water content in each depth interval. Secondly, individual TDR probes can be inserted into the sidewall of an excavation (Figure 2). Measurements from probes at a variety of depths give an estimate of the water content distribution with depth. Thirdly, the MoisturePoint TDR instrument (Environmental Sensors, Inc.) uses electronic segmentation of a single probe to isolate each depth increment for TDR measurement. Thus, selected depth intervals can be isolated during the TDR measurement giving the water content of each interval along the probe. Which method is best is dependent on the purpose for which data will be applied and the amount of allowable site disturbance. Excavation for horizontal installations causes the most disturbance.

The measurement and monitoring of soil water content have been advanced significantly by data acquisition and signal multiplexing capability. Although the Tektronix model 1502 B or C cable tester coupled with custom software and a PC have been foundational TDR instruments, numerous other instruments that have been designed specifically for soil measurement are now available (Ferré and Topp, 2002). Most commercial instruments perform analyses of the TDR traces internally and display or record the interpreted water contents only.

Capacitance Measurements

As the name implies, capacitance devices determine the apparent capacitance of a probe placed in or near a soil (Gardiner *et al.*, 1998; Paltineanu and Starr, 1997; Robinson *et al.*, 1998; Starr and Paltineanu, 1998). The capacitance probe (along with the soil) forms part of an inductance-capacitance (LC) resonant circuit with a specific frequency that depends upon the water content of the soil near the probe. The probe capacitance, C, is expressed as (Starr and Paltineanu, 2002):

$$C = g\varepsilon_{ra} \tag{6}$$

where g is the geometrical constant for the electrode configuration. With appropriate choice of inductance, L, it is possible to establish the resonant frequency for soil probes in the frequency range of 100 to 150 MHz (Figure 1) where salt and other conductivity factors are minimized. The oscillation frequency, F, is an inverse square root function of capacitance and a function of water content:

$$F = \left(2\pi\sqrt{LC}\right)^{-1} = f(\theta_v) \tag{7}$$

In principle, inserting equation (6) into equation (7) leads directly to a calibration equation for capacitance probes having a similar basis as equation (5). In practice, however, such a calibration approach has not proven robust and empirical calibrations are recommended for each soil. There are some instruments that operate on a very similar basis, although the manufacturers may claim to measure impedance or some other parameter. In effect, these are capacitance instruments since the capacitance contribution to impedance is the quantity most affected by water content.

Capacitance instruments can be adapted to use different electrode configurations (Figure 2). They generally fall into two categories, two or more parallel rods constructed to be pushed into, or buried in, the soil, or one or more pairs of cylindrical metal rings separated by a nonconducting plastic ring and mounted on a cylindrical support that is inserted into a previously installed PVC access pipe. The zone of influence or sensor sampling volume for rod-type capacitance probes is similar to that of a TDR probe using the same rods as a probe. That is, it is integrated along the length of the rods, and largely contained between the electrode rods. These probes are best suited for surface or burying near surface, provided soil disturbance by installation is minimized. Capacitance probes configured as one or more pairs of cylindrical metal rings are well suited for discrete depth interval measurements in the soil profile. The zone of influence of the electromagnetic field in the metal ring probes includes the plastic separating the rings, the PVC access pipe, and the surrounding soil that is in a fringe field of the capacitor. As a result, the sample area is limited to the region immediately adjacent to the access tubes.

Semipermanently installed capacitance probes placed in access tubes can be automated for real-time measurements over large areas, with probe readings essentially unaffected by cable lengths up to 500 m. The access pipe must be carefully installed so as not to introduce air-gaps along the pipe or disrupt soil structure near the pipe. Capacitance probes configured with multiple capacitance sensors are reported to be highly sensitive and robust for field-scale research. Starr and Paltineanu (2002) have given specific procedures for calibration and use of capacitance probes.

THREE NONINTRUSIVE METHODS

Three methods operating at frequencies >10 MHz, such as GPR, active microwave, and passive microwave methods, measure EM waves after they have propagated through and/or reflected from the soil. In each case, the detected wave contains the desired water content information. An electromagnetic induction technique, primarily aimed at measuring electrical conductivity, has been used to infer water content from electrical conductivity.

GPR

GPR is a geophysical method that uses electromagnetic signals that propagate as waves to map subsurface structure. GPR instruments are quite simple in concept, consisting of a transmitter, a receiver, and a data recording and display device. The details of making such devices in reliable, easy-to-use form are complex, and the instrumentation side of GPR has only recently come of age for effective application in soil. Even a few years ago, measurements discussed here were not practical because of instrumentation limitations. Electromagnetic fields are strongly absorbed by soils and rocks. In some situations, the fields are absorbed so rapidly that penetration into the ground can be very limited. Wave penetration is required to provide information at depths other than at the surface. GPR penetration is determined by the electrical conductivity of the soil, water content antenna frequency, and the presence of strong reflectors. Typical wave penetration would be 1 to 3 m in silty sands and nonsaline clays while low frequency GPR waves can penetrate tens of meters in dry, nonsaline sandy soil. As a result, GPR response is highly site-specific and testing at a specific location before use of GPR is highly recommended if substantial depth penetration is required. GPR is used for estimation of soil water content in several ways (Davis and Annan, 2002; Huisman and Bouten, 2003; Huisman *et al.*, 2003a), as discussed below.

For surface-launched and subsurface reflections and scattering, a GPR transmitter and receiver are placed on the ground surface to obtain good energy coupling into the subsurface and to minimize any effects of above-ground features. As the transmitter and receiver are moved over the ground surface, a GPR section is collected that images reflections from interfaces in the soil or scattering from localized objects. As with TDR, the primary information used for water content estimates is the velocity of the signal. Equations (2) and (5) can be used to relate this velocity to the volumetric water content. Unlike TDR, the EM wave transmitted from a GPR signal travels unguided through the medium. Therefore, the traversed path length must be defined in order to translate the recorded pulse travel times to velocities. Either discreet buried targets, such as buried pipes, or flat surfaces, such as a water table, can be used if their depth is known independently. The resulting water content from this method is a linear weighted average of the water contents along the travel path of the wave (Huisman *et al.*, 2003a).

Surface-launched direct wave arrivals make use of the fact that there are radar signals that travel both through the air and through the ground between the transmitter and the receiver (Figure 2), where travel path equals antenna separation (Huisman and Bouten, 2003). The physical process underlying the signal transmission requires that both the transmitter and the receiver be placed close to the ground surface. The signals travel through the air and

through the ground with the intrinsic velocities of the two materials. The ground signal will always arrive later in time. Measuring the differential travel time yields an estimate of the ground wave velocity since the source-receiver separation and the velocity in air are known. Using the ground wave velocity and combining equations (2) and (5), one can then infer the volumetric water content. Requiring only one transmitter and one receiver makes this method conceptually straightforward and offers the possibility for a system to be towed over the ground and make continuous recordings of the air-wave and ground-wave arrivals. In practice, however, the direct ground wave may not be always readily visible on the record, particularly in more electrically conductive soils. Although the transmitter and receiver must be in close contact with the soil, they must also be separated sufficiently from each other to give separation of air and ground wave arrivals (Huisman *et al.*, 2003b). The depth of the ground-wave measurement is a complex function of wavelength, soil properties, and antenna separation. The depth of wave propagation uncertainty gives ambiguity to the water content and requires more research investigation.

Air-launched surface reflection methods place the transmitter and the receiver at some distance above the ground using a wheeled vehicle or a low-flying air platform (Figure 3) (Redman *et al.*, 2002; Huisman *et al.*, 2003a). In these methods, the air-soil reflection coefficient is measured remotely, requiring no physical contact with the ground. The lack of physical contact leads to the attractive possibility of moving the GPR system rapidly over large areas to generate large-scale maps of soil water content conditions. The reflection coefficient from the air-ground boundary is represented by equation (4). To measure the reflection coefficient to the required high degree of accuracy, one must be able to measure the amplitude of the return signal. Usually, the system is calibrated by measuring the reflected signal from a "perfect" reflector such as a metal plate of sufficient areal extent to cover the measurement footprint. Extracting water content information involves substitution of equation (3) into equation (4) for the air-soil interface to obtain ε_{ra} of the soil. Equation (5) is then used to infer the volumetric water content. The operating frequency must be high enough, typically greater than 10 MHz, to eliminate σ_0 as a dominant factor in equation (3). The resulting water content measurement represents a nonlinear averaging over a depth that is dependent on the operating frequency and the water content, but heavily weighted to the near-surface conditions.

Surface-launched multioffset subsurface reflections use one or more transmitters and/or multiple receivers to measure the GPR response from a ground-based system. The array of units is moved over the ground and measures the signals as they travel over a multiplicity of travel paths at every measurement point. Essentially, this is equivalent to common midpoint, CMP, or wide-angle reflection and refraction, WARR, soundings commonly used in seismic geophysical sounding. The advantage of this approach is that it does not require any controlled information about the subsurface to be available to extract a radar velocity versus depth; it only requires the presence of reflecting

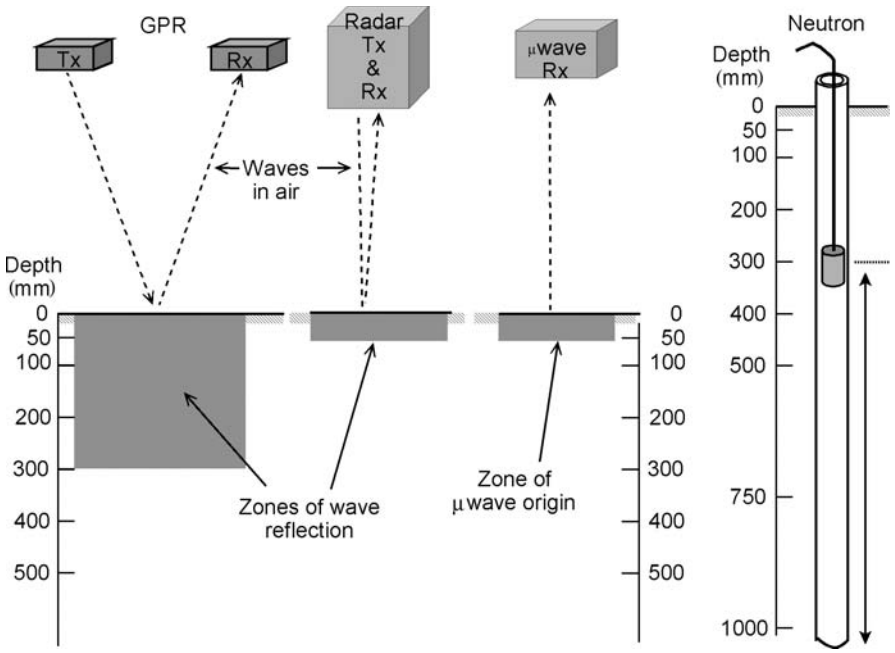


Figure 3 Schematic representations of air launched surface reflection GPR, active microwave remote sensing, passive microwave remote sensing, and neutron themalization methods

horizons or objects. The multiplicity of travel paths down and back from the reflecting horizon can be used to estimate both velocity and depth of the horizon. Measurement accuracies of better than 5 to 10% in volumetric soil water content are practical. This method is attractive and has been used in experimental groundwater investigations on a larger scale. The limitations of this method are that one must have detectable stratigraphy and adequate GPR signal penetration. In addition, multitransmitter and/or multireceiver GPR systems are not usually in standard commercial existence. The complexity and costs are factors that will likely preclude rapid advance of this technique except for research investigations. Use of a single channel instrument at successive separations is an inexpensive way to develop the technique before moving to the added cost of the multiple receiver system. The penalty is the time required to carry out each CMP sounding in the field. The advantage of this method is that it gives both reflecting horizon depth and water content as independent measurements from the data. One can envisage that a system like this could be towed over the ground to map water content versus depth over large areas in a fairly efficient manner. This approach to measuring soil water content is attractive, but the technological cost of the system will be high and the practicality of automated data reduction still needs to be demonstrated.

Borehole transillumination is a means whereby a transmitter or a receiver or both are placed into the ground and, thus, is, in reality, an intrusive measurement (Figure 2) (Parkin *et al.*, 2000; Ferré *et al.*, 2003). The physical measurement being made is the transit time for the signal between the transmitter and the receiver. Access pipes for transmitter/receiver may be vertical or horizontal. The method is fairly simple in concept in that one has known geometrical path lengths and one only has to measure the travel time to infer a velocity that can be associated with the particular path travelled. By making measurements over multiple paths and using tomographic reconstruction techniques, one can generate images of velocity in the subsurface. From this velocity image, one can then infer the water content distribution over the same area, using equations (2) and (5). The signal amplitude can also be measured and then the signal attenuation in the soil can be determined. This, combined with the velocity, permits the electrical conductivity to be determined. This technique has been demonstrated to work well and can be very reliable for making water content measurements (Ferré *et al.*, 2003).

The requirement for semipermanent access pipe(s) makes this technique much more of a site-specific investigation tool and may be used as a calibration or a reference point for area reconnaissance techniques. The minimum allowable separation of boreholes for ground penetrating radar transillumination is defined by the need to achieve far field EM transmission conditions. As a rule of thumb, this distance is approximately equal to the length of the dipole antennae used (e.g. 1 m for 100 MHz antennae). The maximum separation is limited by signal attenuation. As with surface ground penetrating radar applications, signal attenuation increases with increasing bulk electrical conductivity, increasing antennae frequency, and increased effects of dielectric relaxation. Unlike surface-based radar reflection techniques, borehole transillumination only makes use of the first arriving energy and only requires a one-way travel path through the medium. As a result, achievable borehole separations are typically at least twice the maximum depth of investigation using surface-based radar approaches. As an illustrative example, a 100 MHz signal can be propagated at least 2–3 m in most nonclay soils. However, it is recommended to estimate or measure the dielectric and electrical properties of site soils before installing borehole radar access tubes.

A number of commercial companies manufacture GPR instruments, primarily focused on geological, geotechnical, and engineering applications. All of the instruments are primarily designed for use in the reflection-sounding mode where the transmitter and receiver are placed in a single package. These are satisfactory for the air-launched surface reflection and the surface-launched reflection and scattering methods. The remaining GPR methods given here require separation of transmitter and receiver antennas, which is only offered by three commercial vendors (Davis and Annan, 2002).

Active Microwave Remote Sensing

Active microwave remote sensing (radar) is similar in principle to air-launched surface reflection GPR that can be mounted on either airborne or space-borne platforms. Synthetic Aperture Radars (SARs) in the 1 to 10 GHz frequency range are most commonly used for soil water content estimation (Boisvert *et al.*, 1996; Jackson, 2004). A microwave pulse is generated by a transmitter, and this signal is directed towards the ground target via a transmitting antenna (Figure 3). The polarization (generally horizontal (H) or vertical (V)) of the transmitted wave can be chosen by employing different antennae. The backscattered signal travels back to the radar's receiving antenna. A receiver then amplifies the signal, which is processed electronically to provide the transmission time, signal amplitude, and signal phase. The transmission time to and from the target is used to determine the range or distance to the target. The relative permittivity of the soil surface is the primary factor determining the strength of the signal returned to the sensor from the target. The frequencies used for soil water content estimation correspond to wavelengths in air of 3 to 30 cm. In wet soil, the wavelengths are reduced by a factor of approximately 5. At 5.3 GHz (C-Band and 6 cm wavelength in air), plants, plant residue, and surface roughness become major factors causing scattering of the

transmitted wave and, thus, obscuring a pure reflection and preventing the direct application of a reflection coefficient, equation (3), for determining the volumetric water content (Jackson *et al.*, 1997; McNairn *et al.*, 1996). Hence, empirical calibrations are most often used. Some semiempirical calibrations have been developed that make use of polarization data to estimate crop and roughness contributions and to refine estimates of soil permittivity. More research must be undertaken to develop models to account for vegetation effects, but using current approaches, soil water measurements from SAR are limited to surfaces without significant vegetation cover.

An important factor is the depth of soil sensed by the radar reflection. An approximate depth estimate is the wavelength in soil, which varies inversely with the water content. The 1 to 10 cm range is that which is measured by current radar systems. A high proportion of the reflection originates within the upper few centimeters, regardless of the water content. The lateral resolution of radar is at best 30 m and the lateral resolution of most methods used for calibrating radar has been of the 2 to 5 cm range. Hence, reference data for calibrating radar remains a challenge in that reference methods, usually TDR and gravimetric sampling, do not offer adequately fine vertical resolution and or sufficient lateral coverage. Nevertheless, successful correlation models for water content have been developed offering promise for further development and use of radar imagery for soil water mapping (McNairn *et al.*, 2002). In the context of soil water detection, airborne SAR images are currently used in research to simulate imagery of future space-borne systems.

Passive Microwave Remote Sensing

In the passive microwave method, the ground surface is the source of the EM signal. This signal is detected and interpreted for estimation of water content. This method measures the natural thermal emission of the land surface using very sensitive detectors that are tuned to specific frequency bands in the microwave region (0.3 to 30 GHz, Figure 1) (Jackson, 1993, 2002, 2004). Owing to the uncontrolled source strength, it is more complicated to infer the soil water content from these measurements than from active radar. Interpretations are made as follows: (i) normalize brightness temperature to emissivity; (ii) remove effects of vegetation; (iii) account for surface roughness effects; (iv) relate emissivity to relative permittivity; and, finally, (v) convert relative permittivity to water content using an equation similar to equation (5). The relationship between relative permittivity and water content was developed for passive microwaves by assuming that the emitting surface is a planar. This permitted the use of Fresnel reflection equations that predict the surface microwave reflectivity as a function of relative permittivity and the viewing angle. Invoking the relationship between reflectivity and emissivity, and assuming that effects of conductivity are negligible, one gets a relationship between relative permittivity and the measured emissivity. The presence of vegetation reduces the sensitivity to soil water content by attenuating the signal originating from the soil and by adding extraneous microwave emissions to the microwave signal. Because this attenuation increases as frequency increases, lower frequencies are commonly used. Semiempirical approaches have been used to correct emissivity for both vegetation and surface roughness.

The EM methods presented to this point have used gravimetric water content measurement as their primary reference for calibration. But, over the past decade, TDR has become an alternative reference because TDR is easier to use and provides equal or better quality data compared with gravimetric methods. This has not been the case for passive microwave measurements, where the development of calibrations followed the empirical path and developed separate equations for different frequencies and different soil textures based solely on gravimetric measurements. Curves for several combinations of these parameters as given by Jackson (2002) are shown in Figure 4. For comparison, equation (5) for TDR is also given (Figure 4). These passive microwave calibrations show close agreement with equation (5), particularly at lower frequency and for loamy textured soils. As shown by other researchers, there is increasing deviation from equation (5) with increasing clay content (Wraith and Or, 1999). As suggested above and in Figure 1, there is also a systematic difference between the TDR-measured dielectric permittivity and that measured by passive microwave methods as the frequency of operation increases beyond the TDR bandwidth.

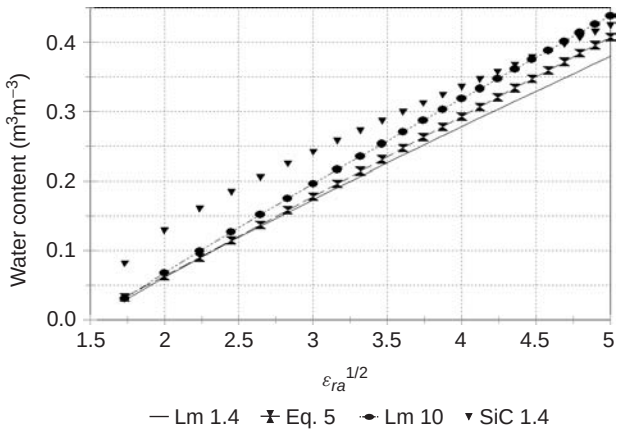


Figure 4 Water content to relative permittivity relationship from TDR (equation 5) and from passive microwave at 1.4 and 10 GHz for loam (Lm) and silty clay (SiC) soils (Reproduced from Topp and Ferré (2001) by permission of MFPA Weimar)

To date, passive microwave systems for soil water content have only been employed on aircraft platforms as prototypes of future space-borne instruments. Depth resolution limits for passive microwave satellites are similar to active radar but the lateral resolution is broader than for active radar, at greater than 50 km from current and future satellites systems.

Electromagnetic Induction

The bulk electrical conductivity (EC) of a medium is also strongly dependent on the volumetric water content. Because this property is relatively simple to measure, EC-based methods have long been used to infer water content. However, the EC is also sensitive to other soil properties, including the temperature, the clay content, and the salinity. As a result, EC-based water content measurements are best viewed as qualitative measures of water content. Electromagnetic induction (EMI) uses two antennae to transmit and receive electromagnetic signals through the subsurface. EMI instruments are not designed to measure the velocity of propagation of these signals as described above. Rather, they employ lower frequency signals and primarily measure the signal loss to determine the electrical conductivity of the medium and to measure different properties. The volumetric water content can then be inferred through relationships with the measured bulk electrical conductivity of the soil, but the calibration process is not straightforward and is both site- and temperature-restrictive. EMI has been used with success for intraseasonal comparisons in quite different geological and hydrological settings (Kachanoski *et al.*, 1988; Sheets and Hendrickx, 1995).

TWO LONG-STANDING METHODS, OPERATING ON NON-EM BASES

The thermogravimetric method has often been considered the only direct measure of water content as it determines the actual amount of water in a soil sample through evaporative drying of the sample. This method has long been considered the standard against which all other methods have been evaluated and against which procedures have been calibrated.

Thermogravimetric Water Content

In the gravimetric method, the water initially present in a sample is determined by recording the loss of mass in response to heating of the sample. Samples can be heated in a variety of ways including: incandescent heating to a controlled temperature to achieve a constant mass; microwave heating to a constant mass, and vacuum distillation of water to a desired vapor pressure level. Ovens using incandescent heaters and convective heat transfer are most common and a standard for mineral soils is based on heating at 105 °C until the mass of soil is constant (Topp and Ferré, 2002). Samples must be removed from the site of measurement for the gravimetric method. This method may be used for laboratory experiments and in field situations. The sample removal requirement makes this a method that is destructive of the site and not amenable to repetitive measurement or data-logging.

Although thermogravimetry is considered a standard method, it has a significant accuracy limitation that is often not recognized. The gravimetric method results in water content on a mass basis, which can be estimated to a high degree of accuracy. All other methods here included make estimates of water content on a volume basis. Conversion from mass to volume basis requires an estimate of soil bulk density, which can seldom be estimated to better than 5% accuracy. Hence, using the gravimetric method as a standard has at least this level of uncertainty associated with comparisons to other methods measuring on a volume basis.

Neutron Thermalization or Neutron Moderation

These instruments include a radioactive source of high-energy, (fast) neutrons. When these epithermal neutrons collide with atoms in the soil, they lose energy, becoming thermalized. Given that hydrogen nuclei are similar in size and mass to neutrons, they serve particularly well in thermalizing fast neutrons. Thus, a measure of the quantity of thermalized neutrons returning to a detector on the probe within a given time period gives a good measure of the saturation of hydrogen atoms in the soil. That is, the measured concentration of thermal neutrons changes mainly with changes in the hydrogen content of the surrounding material. On the timescales of normal interest in water management, changes in H content occur mainly due to changes in soil water content. Therefore, the concentration of thermal neutrons surrounding a neutron source can be precisely related to soil volumetric water content at the time. The source and detector are commonly placed together in a cylindrical probe that is lowered into the soil via a preinstalled access pipe (Figure 3). The volume of measurement is approximately a sphere with radius of about 0.15 m in a wet soil and up to 0.5 m at 0.02 m^3 m^{-3}, limiting measurements to 0.3 m depth and below.

In addition to hydrogen, atoms of oxygen, carbon, and nitrogen, as well as other soil components, are somewhat effective in the neutron thermalization. This requires that neutron probes be calibrated for each soil. There are well-documented calibration and use procedures (Hignett and Evett, 2002) from which it is possible to achieve accurate water content profiles to depths not easily attained with most other methods. As a result of the large and variable measurement volume, the method is inappropriate where detailed vertical definition is required or where sharp water

content gradients exist. Due to regulations concerning the use of radioactive sources, the method is not suitable for automatic, unattended measurement and is, therefore, labor intensive.

ON CHOOSING SOIL WATER INSTRUMENTATION

A number of methods of measuring soil water content have been described. While the wide variety of methods available is a benefit to hydrologists, it also embodies a responsibility to choose appropriately the instrumentation that will "do the job" most efficaciously. The variety of hydrological research and applications requiring soil water content is vast and precludes a prescriptive evaluation of the methods for all uses. We offer, instead, some considerations that should be included when initiating projects involving soil water measurement.

Accuracy and Precision Requirements

Although very important, accuracy is often not considered when choosing instrumentation but, instead, one adopts the assumption that more accuracy is always better. In addressing questions of accuracy, it is important to identify what will limit the accuracy of the final data to be used. For example, in environments where spatial variability contributes large uncertainties, more useful data may result from more replications in space using less accurate, more rapid methods. Furthermore, it is imperative to evaluate whether it is important to have absolute accuracy or relative accuracy. That is, most instruments provide better measures of the magnitude of change of water content than absolute water content. This is especially true of methods that require calibration for soil and vegetation factors such as capacitance probes, neutron moderation, active remote radar, and passive microwave. For GPR, there have not been enough studies to document its relative performance. TDR shows the least difference between relative and absolute accuracy, unless clay content is high and temperatures change dramatically (Wraith and Or, 1999). Relative accuracy is often limited by precision or repeatability of measurement. Improved precision results from those methods where fewer factors change between repeat measurements. Included here would be TDR using semipermanently installed probes, capacitance at installed sites, and neutron moderation.

The limit for accuracy for any method is determined by the accuracy of the reference method used to form the calibration equation. Traditionally, the thermogravimetric method has been the reference standard that has a practical accuracy of about 5% of the range (i.e., $\sim\pm 0.02\,m^3\,m^{-3}$), being limited by the accuracy of the bulk density as already stated. TDR has been shown to have a similar level of accuracy and has become widely used as a reference method for calibration of other methods, in particular GPR and active remote radar.

Sample Geometry

Each of the methods presented performs measurements over vastly different geometries of samples, making choices between instruments on this basis impractical. Sample geometry, however, can be an important consideration in at least two situations. Where two methods are to be used in complementary working relationship, it is important to optimize the complementarity by choosing appropriate sample arrangements. For example, TDR probes can be designed with a variety of configurations. Often, TDR is increasingly used to spot-check or monitor within transects of GPR and remote radar. In both cases, it is imperative that the TDR used for ground truthing measures throughout the same increment of soil, that is, tens of millimeters deep for remote radar and 100 s of millimeters for the GPR equivalent.

The geometry of the sampled soil should be representative of the hydrological process under study. In the case of intrusive devices (capacitance and TDR), the installation of probes and/or access pipes causes disturbance to the soil. Probe size and configuration must optimize the measurement in soil not affected by the installation and the installation must not affect the flow of water. This is particularly important for capacitance approaches where only the fringing EM field is used, which lies within the zone disturbed by the access pipes.

One of the most difficult regions of the soil to measure is at the soil surface, a most important hydrological interface. The factors affecting soil water content are usually variable in time and space as a result of vegetation, rainfall, traffic, and evaporation, all interacting to change the soil and water properties at the soil surface. TDR offers the best capability for accurate measurement in the 10s of mm depth, but one must choose the probe location and orientation carefully to assure that the EM field remains within the soil for measurement. Most existing TDR probes cannot be left in place at the surface as they would interfere with infiltration and evaporation. From 50 mm and deeper, horizontally oriented probes inserted from a small soil pit have worked successfully (Topp *et al.*, 2000). Improved measurements at the soil surface would be possible from TDR probes specifically designed for that measurement requirement. At the larger scale, surface-launched direct wave GPR may become an important method for monitoring near-surface water contents.

Site Replication

When it comes to field measurement in heterogeneous media, more is generally better. But, diminishing returns

and cost usually limit the number of replicates that can be collected. The nature and scale of variability generally determine the acceptable degree of replication. A number of geostatistical guides and handbooks are available to assist this choice (e.g. Warrick and Van Es, 2002). One will either need to characterize the site variability or to rely on previously obtained characterizations of the site variability to design optimal sampling protocols. For example, to have representative measure of the root zone soil water dynamics in a maize crop with 0.75 m row spacing, one would ideally choose cross-row lateral replication spacing and number in relation to the 0.75 m dimension. Many factors are not as conveniently periodic but should be recognized in developing a sampling strategy. Oftentimes, the number of measurement sites and the spatial replication at a site are restricted by instrumental or operational factors, such as allowable lengths of cables or land use traffic. Such restrictions cause the measurement site to be a surrogate sample and increase the requirement that such sites are adequately representative of the overall landscape and hydrological processes. All available site information becomes very helpful in the site selection process.

CONCLUDING REMARKS

The advancements in electromagnetic methods provide the hydrological sciences with a variety of choices for measurement of soil water content. These EM methods allow greater emphasis on soil water measurement at and near the soil surface, which is one of the most dynamic interfaces in the hydrological cycle. Capacitance, GPR, and TDR methods offer data-logging options to enable the recording of both spatial and temporal dynamics of the changes in soil water content. As these methods advance, it will be increasingly important that hydrologists make informed decisions regarding the most appropriate water content measurement method(s) for specific applications.

REFERENCES

Boisvert J.B., Pultz T.J., Brown R.J. and Brisco B. (1996) Potential of synthetic aperture radar for large scale soil moisture monitoring. *Canadian Journal of Remote Sensing*, **22**, 2–13.

Davis J.L. and Annan A.P. (2002) 3.1.3.5 Ground penetrating radar to measure soil water content. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 446–463, 534–545.

Ferré P.A. (Ty) and Topp G.C. (2002) 3.1.3.4 Time domain reflectometry. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 434–446, 534–545.

Ferré T.P.A., von Glinski G. and Ferré L.A. (2003) Monitoring the maximum depth of drainage in response to pumping using borehole ground penetrating radar. *Vadose Zone Journal*, **2**, 511–518.

Gardiner C.M.K., Dean T.J. and Cooper J.D. (1998) Soil water content measurement with a high frequency capacitance sensor. *Journal Agricultural Engineering Research*, **71**, 395–403.

Gardiner C.M.K., Robinson D.A., Blyth K. and Cooper J.D. (2001) Soil water content measurement. In *Soil and Environmental Analysis: Physical Methods*, *Second Edition*, Smith K. and Mullins C.E. (Eds.), Marcell Dekker: New York.

Hendrickx J.M.H., Wraith J.M., Corwin D.L. and Kachanoski R.G. (2002) Solute content and concentration. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 1253–1322.

Hignette C. and Evett S.R. (2002) Neutron thermalization. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 501–521, 534–545.

Huisman J.A. and Bouten W. (2003) Accuracy and reproducibility of measuring soil water content with the ground wave of ground penetrating radar. *Journal of Environmental and Engineering Geophysics*, **8**, 65–73.

Huisman J.A., Hubbard S.S., Redmond J.D. and Annan A.P. (2003a) Measuring soil water content with ground penetrating radar: a review. *Vadose Zone Journal*, **2**, 476–491.

Huisman J.A., Snepvangers J.J.J.C, Bouten W. and Heuvenlink G.B.M. (2003b) Monitoring temporal development of spatial soil water content variation: comparison of ground penetrating radar and time domain reflectometry. *Vadose Zone Journal*, **2**, 519–529.

Jackson T.J. (1993) Measuring surface soil moisture using passive microwave remote sensing. *Hydrological Processes*, **7**, 139–152.

Jackson T.J. (2002) 3.1.3.8 Passive microwave remote sensing methods. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 488–497, 534–545.

Jackson T.J. (2005) Hydrological application of remote sensing: surface states- surface soil moisture (active and passive). In *Encyclopedia of Hydrological Sciences*, Anderson M.G. (Ed.), John Wiley.

Jackson T.J., McNairn H., Weltz M.A., Brisco B. and Brown R. (1997) First order surface roughness correction of active microwave observations for estimating soil moisture. *IEEE Transactions on Geoscience Remote Sensing*, **35**, 1065–1069.

Jackson T.J., Schmugge T. and Engman E.T. (1996) Remote sensing applications to hydrology: soil moisture. *Hydrological Sciences Journal*, **41**, 517–530.

Jones S.B., Wraith J.M. and Or D. (2002) Time domain reflectometry (TDR) measurement principles and applications. HP today scientific briefing. *Hydrological Processes*, **16**, 141–153.

Kachanoski R.G., Gregorich E.G. and van Wesenbeeck I.J. (1988) Estimating spatial variations of soil water content using

non-contacting electromagnetic inductive methods. *Canadian Journal of Soil Science*, **68**, 715–722.

Kraus J.D. (1984) *Electromagnetics*, *Third Edition*, McGraw-Hill: New York, pp. 775.

McNairn H., Boisvert J.B., Major D.J., Qwyn Q.H.J., Brown R.J. and Smith A.M. (1996) Identification of agricultural tillage practices from C-band radar backscatter. *Canadian Journal of Remote Sensing*, **22**, 154–162.

McNairn H., Pultz T.J. and Boisvert J.B. (2002) 3.1.3.7 Active microwave remote sensing methods. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 475–488, 534–545.

Noborio K. (2001) Measurement of soil water content and electrical conductivity by time domain reflectometry: a review. *Computers and Electronics in Agriculture*, **31**, 213–237.

Paltineanu I.C. and Starr J.L. (1997) Real-time soil water dynamics using multisensor capacitance probes: laboratory calibration. *Soil Science Society of America Journal*, **61**, 1576–1585.

Parkin G., Redman D., von Berotldi P. and Zhang Z. (2000) Measurement of soil water content below a waste water trench using ground penetrating radar. *Water Resources Research*, **36**, 2147–2154.

Redman J.D., Davis J.L., Galagedara L.W. and Parkin G.W. (2002) Field studies of GPR air launched surface reflectivity measurements of soil water contents. *Proceedings of Ninth Conference on Ground Penetrating Radar*, SPIE 4758: 156–161.

Robinson D.A., Gardner C.M.K., Evans J., Cooper J.D., Hodnett M.G. and Bell J.P. (1998) The dielectric calibration of capacitance probes for soil hydrology using an oscillation frequency response mode. *Hydrology and Earth System Sciences*, **2**, 83–92.

Robinson D.A., Jones S.B., Wraith J.M., Or D. and Friedman S.P. (2003) A review of advances in dielectric and electrical conductivity measurements in soils using time domain reflectometry. *Vadose Zone Journal*, **2**, 444–475.

Sheets K.R. and Hendrickx J.M.H. (1995) Noninvasive soil water content measurement using electromagnetic induction. *Water Resources Research*, **31**, 2401–2409.

Starr J.L. and Paltineanu I.C. (1998) Real-time soil water dynamics over large areas using multisensor capacitance probes and monitoring system. *Soil and Tillage Research*, **47**, 43–49.

Starr J.L. and Paltineanu I.C. (2002) 3.1.3.6 Capacitance devices. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 463–474, 534–545.

Topp G.C., Davis J.L. and Annan A.P. (1980) Electromagnetic determination of soil water content: measurements in coaxial transmission lines. *Water Resources Research*, **16**, 574–582.

Topp G.C., Dow B., Edwards M., Gregorich E.G., Curnoe W.E. and Cook F.J. (2000) Oxygen measurements in the root zone facilitated by TDR. *Canadian Journal of Soil Science*, **80**, 33–41.

Topp G.C. and Ferré P.A. (Ty) (2001) Electromagnetic wave measurements of soil water content: A state-of-the-art. In *Proceedings of Fourth International Conference on Electromagnetic Wave Interaction with Water and Moist Substances*, Hübner C. (Ed.), MFPA an der Bauhaus-Universität Weimar: Amalienstr. 13, Weimar, pp. 327–335.

Topp G.C. and Ferré P.A. (Ty) (2002) 3.1 Water content. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 417–545.

Topp G.C. and Reynolds W.D. (1998) Time domain reflectometry: A seminal technique for measuring mass and energy in soil. *Soil and Tillage Research*, **47**, 125–132.

Von Hippel A.R. (1954) Theory. In *Dielectric Materials and Applications*, Von Hippel A.R. (Ed.) MIT Press: Cambridge, and John Wiley & Sons: New York, pp. 3–46.

Warrick A.W. and Van Es H.M. (2002) Chapter 1 Soil sampling and statistical procedures. In *Methods of Soil Analysis Part 4 Physical Methods*, *SSSA Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 1–200.

Wraith J.M. (2003) Measuring solutes and salinity using time domain reflectometry. In *Encyclopedia of Water Science*, Stewart B.A. and Howell T.A. (Eds.), Dekker Publications: New York.

Wraith J.M. and Or D. (1999) Temperature effects on soil bulk dielectric permittivity measured by time domain reflectometry: experimental evidence and hypothesis development. *Water Resources Research*, **35**, 361–369.

73: Soil Water Potential Measurement

WOLFGANG DURNER[1] AND DANI OR[2]

[1]Institute of Geoecology, Department of Soil Physics, Braunschweig Technical University, Braunschweig, Germany
[2]Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, US

Soil water potential controls the dynamics of water in soils. This article describes methods to measure soil water potential. Available methods differ with respect to the moisture range where they are applicable. In the moist range, the pore water pressure of the soil is directly measured with tensiometers. In the intermediate moisture range, reference porous media are suitable, where the water saturation changes with water potential. Gypsum blocks, granular matrix sensors, heat dissipation matric potential sensors, and filter paper are used. If embedded in the soil and in hydraulic equilibrium with it, soil water potential is obtained by a calibration relationship from a water content dependent measurement in the reference porous medium. In relatively dry soils, water potential is derived from vapor pressure measurements by thermocouple psychrometry. This article presents the theoretical background of the various measurement techniques and illustrates the construction principles of measuring instruments. Problems and pitfalls of water potential measurements, in particular with tensiometers, are discussed. A list of manufacturers of water potential measurement instruments is given at the end of the article and a reference list provides material for further background information.

INTRODUCTION

Soil water potential reflects the energy state of water in porous media and thus drives movement of water. Its proper measurement is fundamental for the analysis of water dynamics in the vadose zone. The energy state of soil water can be expressed as energy per mass of water ($J\,kg^{-1}$), as energy per volume of water ($J\,m^{-3}$), which is equivalent to pressure (Pa), or as energy per unit weight of water ($J\ (9810\,kg\,m\,s^{-2})^{-1}$). The latter leads to the expression of the water potential as pressure head, with units of water column length (m), and is commonly used in vadose zone hydrology.

The lack of single measurement technology covering the entire energy range of interest, from moist to dry conditions present a challenge to the measurement of soil water potential. Tensiometry, which is an accurate and widely used technique to determine soil water (matric) potential, requires extensive maintenance and is restricted to relatively wet conditions. Instruments for water potential measurement measure beyond the tensiometric measurement range are heat dissipation sensors (HDS), gypsum blocks, granular matrix sensors, or filter paper. They all depend on equilibration of a reference porous medium with the surrounding soil and require individual calibration to infer soil water potential. At the dry end of the moisture range, thermocouple psychrometers, which use equilibrium water vapor pressure (relative humidity) in soil air, are often used to infer soil water potential.

Measurement range, accuracy, repeatability, response time, and spatial resolution of specific sensors are important considerations to their potential applications and in the analysis of soil water measurements (Or and Wraith, 2002). Our discussion in this contribution focuses on principles and applications of tensiometry, because this technology is the most widely used water potential measurement technology in practice. The techniques which extend the measurement range toward drier conditions will be presented subsequently, focusing first on the more classic reference porous media such as gypsum blocks and granular porous media, and then discussing the recent developments in HDSs. Finally, the principles of vapor pressure–based methods and thermocouple psychrometry along with available instruments will be described.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

GENERAL MEASUREMENT PRINCIPLES

The measurement of soil water potential requires hydraulic equilibrium between soil water in the porous medium around an instrument and the measurement device. For thermocouple psychrometers, this contact is provided by the vapor phase, and for all other instruments by the fluid phase. The tensiometer porous cup is placed in contact with the soil at the location of interest. The pore water outside the cup and the liquid in the interior of the tensiometer are bridged through water-filled fine pores of the cup. Following pressure equilibration, the liquid pressure in the tensiometer is directly measured by a manometer. This measurement technique is straightforward, but the pressure range is limited by the air entry pressure of the porous cup material. It is further restricted to pressures where the fluid is able to exert a pressure force, which in practice is stable only at positive absolute fluid pressures.

Measurement of water potential at drier conditions rely on reference porous media such as gypsum blocks, granular blocks of filter paper brought into close contact with the surrounding soil. With varying matric potential of the soil, the water saturation of these reference porous media vary in a characteristic manner, dependent on the particular pore-size distribution of the reference porous material. The sensor's water content is measured and related to matric potential by an individual calibration relationship. Indirect measurement methods are accordingly based on water content sensor technology, as described in Topp and Ferré (2005), and use weight, electrical, thermal, or dielectric properties of the reference porous medium to infer matric potential. Hysteresis and slow response introduce difficulties as discussed in the following sections.

Finally, the measurement of soil water potential in very dry conditions is based on equilibrium between liquid soil water and water vapor in the ambient soil atmosphere. The drier the soil the fewer water molecules escape into the ambient atmosphere, resulting in a lower relative humidity. The thermocouple psychrometer measures the dew-point temperature in a headspace of a closed chamber with gas-permeable walls, the value is related to relative humidity and soil water potential at the same ambient conditions.

Total water potential is composed of gravitational, matric, pneumatic, overburden, thermal, and osmotic components. For a fixed measurement point, we often disregard thermal and gravitational components (Durner and Lipsius, 2005). In a rigid porous medium, the solid phase does not contribute to the pressure of the water phase, and the overburden potential component can be disregarded. In the absence of semipermeable membrane, that is, when dissolved ions are free to move between the surrounding soil and the measurement cell, the osmotic component of the total potential will not affect the measurement (this is typical for all instruments referred to in this paper, except the psychrometer where vapor–gas interface represents an osmotic barrier). Therefore, psychrometers provide a combined measurement of osmotic and matric potentials. Tensiometers measure the sum of the pneumatic and matric potential, because tensiometer cups are impermeable for the gas phase. Reference porous media (HDSs, gypsum blocks, granular matrix sensors, filter paper) measure the matric potential only, because they are permeable for the gas phase in their operation range and the pneumatic potential does not contribute to the signal.

Any of the measuring instruments referred to thereafter influence the state of the signal to be measured. The measurement characteristic of an instrument is determined by the relation between the sensitivity and the conductance of the instrument. The sensitivity expresses the ratio between the change of the signal (pressure change Δp) with respect to the amount of water ΔV that is required to flow from the surrounding soil to (or from) the instrument (Richards, 1949),

$$S = \frac{\Delta p}{\Delta V} \tag{1}$$

The conductance is dependent on the area of exchange between instrument and soil, and on the flow resistance of the instrument. In the case of tensiometers, this resistance is determined by the thickness and the saturated conductivity of the porous membrane. In the case of equivalence porous media, it is dependent on the volume and porosity of the sensor and its unsaturated conductivity. In general terms, the measurement characteristic of modern advanced tensiometers is very good and almost constant across the measurement range due to the high sensitivity of the pressure transducers and sufficient conductance of the porous cups. The sensitivity of equivalent porous media (EPM) is lower, since by principle the water content of the sensor must change. Furthermore, the conductance varies with suction, since the flow rate is determined by the unsaturated conductivity of the EPM.

TENSIOMETRY

Tensiometers for matric potential determination along with TDR for water content measurement (Topp and Ferré, 2002; Topp and Ferré, 2005) are the most widely used measurement technologies for soil hydraulic properties in the field. Tensiometers are particularly suited for the measurement of soil water energy status at high water contents where transport processes are relatively rapid (e.g., solute transport from the vadose zone down to groundwater). Owing to their limited measurement range, tensiometers are unsuited for drier conditions where soil water becomes limiting for plant growth. Under relatively moist soil conditions, tensiometric measurements are robust, reliable, and accurate, with a considerable body of theoretical and practical knowledge gained by soil physicists,

hydrologists, and other practitioners. However, tensiometers require maintenance and their use at frozen conditions is problematic.

In the following, we discuss principles of tensiometric measurements, tensiometer design, and measurement range, and show recent developments. A treatise on all aspects of practical applicability of tensiometers, including a recent list of vendors of tensiometer systems and components (limited to US market) was provided by Young and Sisson (2002).

Measurement Principle

A tensiometer is used to measure the soil water energy state, which in unsaturated soils is dominated by the matric component of the potential (resulting from combined effect of capillary and adsorptive forces exerted by the soil matrix). A water-filled reservoir, connected to a pressure-sensing element (pressure gauge or transducer), is brought into contact to the soil through a rigid porous membrane, which is permeable for water and impermeable for air. The membrane is in most cases made of a porous ceramic cup, which is brought into contact with the soil by inserting the tensiometer into a predrilled access hole with a diameter at the position of the cup slightly smaller than the porous cup. Alternatively, the tensiometer can be inserted to a larger predrilled hole, where a slurry of nonswelling fine material or sieved local soil is filled in, which is used to provide hydraulic contact between the cup's surface and the soil. When the matric potential of the soil is lower (more negative) than in the tensiometer, water from the tensiometer is attracted by the surrounding porous media, thereby creating suction sensed by the pressure gauge. Water flows until equilibrium is reached and the suction inside the tensiometer equals the soil matric potential. When the soil is wetted, flow may occur in the reverse direction.

Partial contact between the cup and the surrounding material is in principle sufficient, since contact area does not affect the equilibrium pressure, but rather it affects the rate at which equilibrium is attained. Similarly, if material properties near the cup are altered by slight compression of the surrounding soil or by difference of hydraulic properties of the contact material, it does not affect the measured soil water potential value, as energy status is continuous across different porous materials in hydraulic equilibrium.

The porous cup consists of a rigid, fully wettable material with pores of radius r_{max}, through which the spatially discrete arrangement of the water phase of the porous medium (outside the membrane) is connected with the continuous water phase inside the membrane. The fluid pressure outside and inside the membrane equilibrates, and is measured by a pressure sensor. In most cases, the membrane has the form of a porous ring or a porous cup at the scale of few centimeters, and pure water is used as a pressure transducing fluid. To operate properly, the pores of the membrane must remain impermeable for the nonwetting fluid (soil, air), that is, fully water saturated. Pore openings of the membranes are in the range of micrometers, which is about 6000-fold larger than the size of a hydrated Na molecule (Young and Sisson, 2002), thus making them permeable to dissolved ions. Hence, the pressure measured inside the porous cup is not affected by the osmotic potential component.

In most hydrological applications, it is not the absolute pressure of soil water that is of interest, but rather the capillary pressure, which is defined as pressure difference between the water and the gas phase in the soil

$$p_w - p_a = p_c + \frac{1}{L}\frac{\partial \theta}{\partial t} \tag{2}$$

where p_w is the macroscopic pressure in the water phase, p_a is the pressure in the air phase, p_c is the capillary pressure, L is a nonnegative material coefficient, and t indicates the time. Under hydrostatic conditions, the capillary pressure is equal to the matric potential (Hassanizadeh *et al.*, 2002). This equality is implicit in most vadose zone hydrology applications, even at transient conditions. Pressure sensors used for tensiometers, therefore, typically are differential, measuring the difference between the pressure in the tensiometer cell and ambient air pressure, which is negative in unsaturated soils. The absolute value of this pressure difference is called *suction* or *tension*. For isothermal conditions and negligible effect of soil overburden pressure, this suction represents the matric and pneumatic potentials.

For most practical purposes, the contribution of the pneumatic potential in the unsaturated zone is zero, that is, gas pressure in the vicinity of a tensiometer cup is equal to ambient atmospheric pressure. However, there are cases where during rainfall, a layer of air may become sandwiched between a sharp wetting front and the groundwater, and air pressure will increase substantially. Also, in soils that are very wet, the local pressure of enclosed air is not identical to ambient air pressure as long as the gas-phase continuity is restricted. If the tensiometric pressure in such cases is interpreted as matric potential, it will be a misconception and this contributes to a class of phenomena called *dynamic effects* (Schultze *et al.*, 1999; Durner and Lipsius, 2005).

Measurement Range

The measurement range of tensiometers is restricted by the air entry pressure of the porous membrane, p_e. The capillary rise equation relates this pressure difference to a pore size

$$r = \frac{(2\sigma \cos \varphi)}{p_e} \tag{3}$$

where r is the effective pore radius (L), σ is the surface tension between fluid and gas phase (MT^{-2}), φ is the contact angle between the fluid and the cup material, and p_e is the air entry pressure ($ML^{-1}T^{-2}$). To remain fully water saturated up to a pressure difference of 1 bar, membranes of perfectly wettable material ($\gamma = 0°$) must have pore diameters smaller than 3 μm. Young and Sisson (2002) list the properties of various materials used in practice. Mostly, porous ceramic materials are used as membranes, with maximum pore diameters of about 2 μm.

A second constraint to tensiometric measurement range is imposed by the spontaneous boiling or cavitation of water inside the tensiometer at subatmospheric pressures. As the energy state of water in the porous medium becomes more negative and the liquid water pressure inside the tensiometer approaches the vapor pressure $p_v(T)$ at ambient temperature, T, spontaneous evaporation (boiling) of water occurs. However, thermodynamic equilibrium can be inhibited for considerable periods. This metastable state associated with the delay in boiling (defervescence) enables considerable extension of tensiometric measurement range. The theoretical limit of suction is given by the energy needed for cavitation in a pure water phase, which needs a calculated tensile strength of 140 MPa (Fisher, 1948; Zheng *et al.*, 1991). Practical limits of 140 kPa (microtensiometer model #4, Nardeux, St. Avertin, France; Tamari *et al.*, 1993) and −400 kPa (microtensiometer model T5, UMS, Munich; G. von Unold, UMS Corp. Munich, personal communication), and −1400 kPa (tensiometer prototype; A.A. Diene, *Development For High Suction Tensiometers Analyzed in Laboratory Lysimeters in the Study of Drainage*. MSc. Thesis, COPPE/UFRJ, Rio de Janeiro, Brasil) have been reported. If the stress on water surpasses the boiling inhibition, a sudden pressure jump occurs in the tensiometer cell to the pressure of the water boiling point. Requirements for a stable defervescence are the absence of boiling germs, absence of air bubbles or dissolved gases, a strongly hydrophilic ceramic surface, a low-enough air entry value of the porous ceramic, and surfaces that are all hydrophilic and polished, even pressure transducer and ceramic. To reach this, a clean and precise manufacturing is necessary. As a further technique, microorganisms are applied, which apparently "bind" water by adhesive forces to all surfaces, and take up the rest of dissolved gases (G. von Unold, UMS Corp., personal communication). In such a system, the defervescence is reversible and reproducible.

Tensiometer Designs

Historically, the concept of soil water potential in unsaturated soil was introduced by Buckingham (1907) from a theoretical perspective. The most widely cited first descriptions of tensiometers are attributed to Gardner *et al.* (1922) and Richards (1928), but recently Or (2001) has shown that already Livingston (1908, 1918) had proposed the use of a tensiometer, consisting of a porous water-filled cup and a pressure manometer. In these early descriptions, a liquid-filled porous cup is connected to a U-manometer, filled with water or mercury (Figure 1).

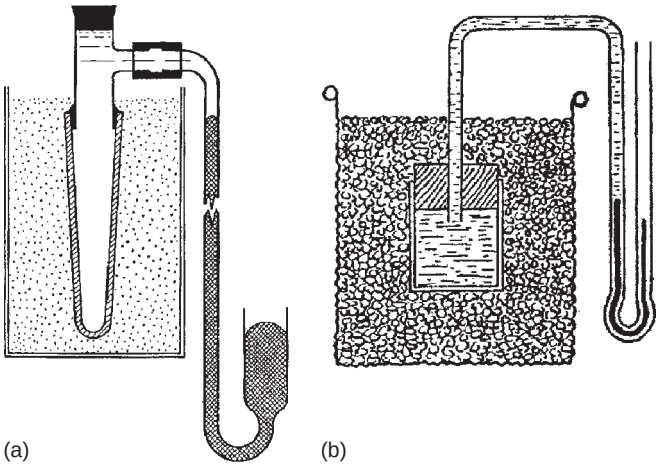


Figure 1 (a) Richards (1928) *tensiometer design*. (b) Haines (1927) *design attributed to* Livingston (1918). (Reprinted from Or (2001) with permission of Soil Science Society of America)

Up to the 1980s, tensiometers were constructed on the basis of the design shown in Figure 1, using mercury manometers. These instruments have advantages with respect to the accuracy of the readings, but suffer from inherently sluggish response times owing to their low sensitivity.

Mechanical Tensiometer

Owing to environmental concerns and their low sensitivity, mercury manometers were replaced by mechanical pressure gauges and pressure transducer tensiometers in laboratory experiments (Vachaud and Thony, 1971) and field investigations (Bianchi, 1962; Watson, 1967). Problems are manometer hysteresis, temperature effects, and costs.

Tensiometers Without a Permanent Manometer (Septum Tensiometers)

In an attempt to reduce costs, Marthaler *et al.* (1983) proposed a mobile pressure transducer system. This type of tensiometers is in wide use nowadays. Tensiometer tubes without a manometer are installed in the soil, filled with deaerated water, and plugged with septum stoppers. A mobile electronic pressure transducer manometer is connected to a syringe needle, which is stitched through the septum to measure the pressure in the tensiometer. Insertion of the syringe needle into a completely water-filled system exerts a positive pressure and can significantly alter interior pressure, which in turn drives a small amount of water out of the ceramic cup. To avoid this, Marthaler *et al.* (1983) recommend leaving a 2-cm high air bubble

at the upper end of the tensiometer. Inserting the needle into the air pocket below the septum leads to a moderate increase in pressure, which again leads to a leakage of water into the soil, resulting in a new equilibrium pressure. Marthaler *et al.* (1983) estimated complete reequilibration to take from 2 to 10 min. In practice, this type of tensiometer needs careful and skilled operators. Otherwise, problems with clogged needles, air leakage while boring through the septum, air leakage by perforation of septums, and temperature effects (Warrick *et al.*, 1998) make readings unreliable. Furthermore, the existence of an air bubble leads to a dramatic worsening of the instrument's sensitivity at high suctions, as will be shown later.

Generally, these tensiometers may not be suitable for precise measurements.

Pressure Transducer Tensiometers

In pressure transducer tensiometers, a metal membrane of a permanently connected pressure transducer is deformed by a pressure change. The electrical response can, for example, be caused by a silicon crystal, onto which a piezoresistive circuit is fused (Tandeske, 1991). Pressure transducer tensiometers have a far higher sensitivity than U-manometer tensiometer, and are nowadays widely used. Further, they allow automated logging of the tensiometer signal in almost arbitrary temporal resolution, which enables us to measure the fast changes of the pressure signals in soils during transient flow processes, thus enhancing our process understanding of water flow in soils. Following traditional tensiometer designs, the early pressure transducer tensiometers had the manometer installed at the soil surface as the tensiometer tip is transmitted by a water column to a manometer located at the soil surface. This makes them very sensitive to temperature fluctuations, as will be discussed below. Furthermore, their use at larger depths is impractical.

Advanced Tensiometers

The most advanced way to measure pressures in tensiometer cells is by electronic pressure transducers that are embedded in the instrument directly at the ceramic cup. Advanced pressure transducer-equipped tensiometers have features that enhance measurement stability and control (Sisson *et al.*, 2002) such as: (1) integrated amplifier and signal conditioning for the electric output, (2) temperature-compensation and proximity to the porous cup, (3) filling status indicator, (4) external *in situ* refilling, and (5) an integrated temperature sensor in the porous cup for additional control of temperature effects. Figure 2 shows the design drawing of such tensiometers. Figure 3 shows two examples of commercially available modern tensiometers.

Self-filling Tensiometers

Recently, a self-filling tensiometer has become commercially available (TS1, UMS, Munich, Germany). After drying periods, when the soil is rewetted by rain, an integrated

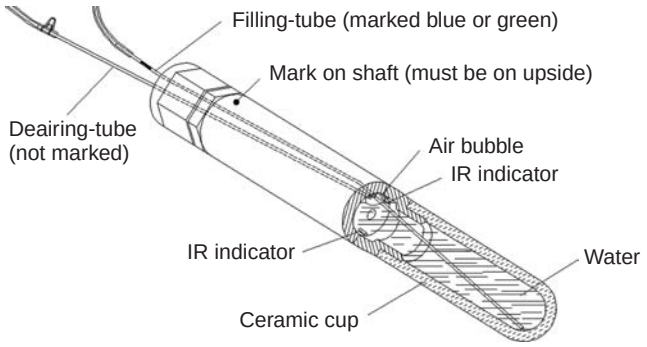


Figure 2 Schematic of tensiometer head of a modern pressure transducer tensiometer. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

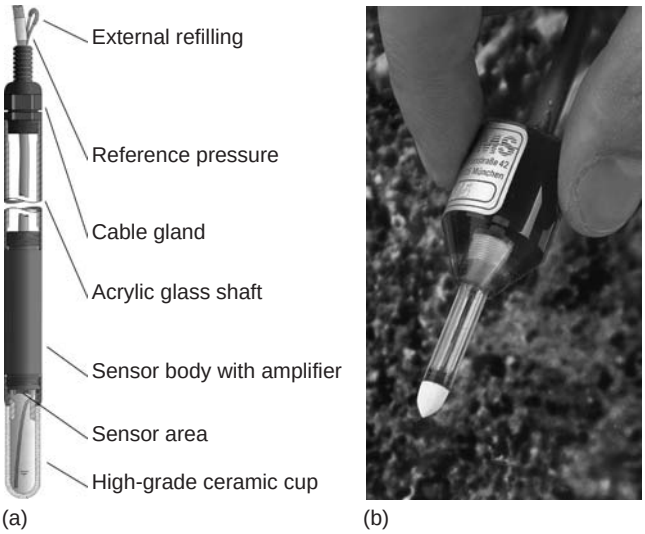


Figure 3 Advanced pressure transducer tensiometers. (a) Pressure transducer tensiometer T8 by UMS. (b) Mini tensiometer T5 by UMS (Reproduced by permission of UMS Inc., Munich, Germany, 2004). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

micropump, controlled by an onboard controller, starts to suck the water from the soil into the tensiometer until the whole interior is water filled again. The rate of water intake can be concurrently used to estimate the unsaturated conductivity of the surrounding soil.

Osmotic Tensiometers

Various designs attempt to extend the range of tensiometers by filling the tensiometer with solutions made up of large molecules (see Figure 4). Peck and Rabbidge (1969) used polyethylene glycol with a molecular weight of 20 000 to extend the measurement range to −15 bar. They stressed the importance of temperature correction in this system, since thermal transients give rise to expansion of both

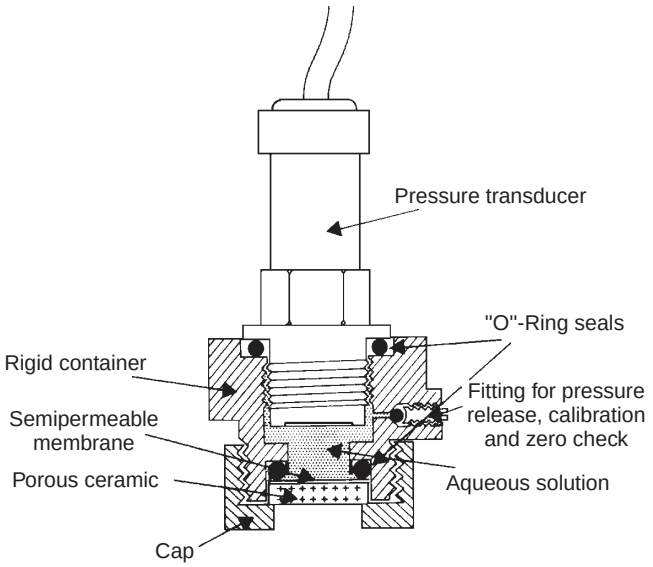


Figure 4 Schematic of an osmotic tensiometer. (Adapted from Peck and Rabbidge (1969) with permission of Soil Science Society of America)

the solution and the tensiometer cell, giving "spurious" readings. Bocking and Fredlund (1979) investigated the practicability of the osmotic tensiometer and concluded that many difficulties stand in the way of a routine application, amongst them leaking and drifting, slow response time, and temperature dependence. Biesheuvel *et al.* (1999) simulated the response of the osmotic tensiometer, incorporating membrane permeability, instrumental sensitivity, differential temperature change, and osmotic pressure – temperature dependency. They found reasonable accuracy when comparing their model with results obtained from a prototype that was exposed in a climatic chamber in pure water. However, Gee and Ward (1999) conclude that despite recent attempts with improved membrane materials and sensor configurations, to date the problems of sensor drift and leakages of the osmotic solutions remain unsolved, and no commercially osmotic tensiometers are presently available.

Measurement Practice

Good measurement practice with tensiometers requires knowledge on proper installation and maintenance, and interpretation of the signals. The specific measurement characteristic and the practical functioning of a tensiometer will depend on a variety of factors. Among those is the type of pressure sensor, the size and type of the porous cup material, the amount of water and air inside the tensiometer, and the contact between the tensiometer and the surrounding soil.

Soil-tensiometer Contact

For any instruments for soil water potential measurement, the hydraulic contact between the tensiometer cup and the soil is essential for a proper operation. If the tensiometer cup is pressed in a predrilled hole that is slightly narrower than the diameter of the cup itself, this contact is generally sufficient. In stony soils, this practice might, however, not be feasible. In these cases, either a slurry of local sieved fine soil or else a silty contact material can be used to provide contact in a predrilled access hole of larger diameter. Sand cannot be used as contact material, since the hydraulic contact at lower potentials gets essentially lost owing to the steep decrease of sand's unsaturated conductivity. Clay and loam material is also not recommended, since it hampers the fast equilibration at high potentials because of its low saturated conductivity, and, further, is likely to shrink when the soil is drying, thus creating an isolating crevice between the soil and the tensiometer cup.

Response Times

The response time of a tensiometer is determined by the sensitivity S of the reading and the resistance R of the cup/soil system against water flow (Flühler and Roth, 2004)

$$t_r = \frac{R}{S} \tag{4}$$

where the sensitivity S is defined by equation (1), and the resistance of the cup is given by

$$R = \frac{1}{k_{cup}} = \frac{d_{membrane}}{A_{cup} \cdot k_{membrane}} \tag{5}$$

In (4) and (5), t_r is the response time (T), S is the sensitivity of the reading with respect to volumetric water exchange between soil and instrument ($M L^{-4} T^{-2}$), R is the hydraulic resistance of the porous cup ($L^4 T M^{-1}$), V is the volume of Water (L^3), k_{cup} is the hydraulic conductance of the cup, $d_{membrane}$ (L) is the thickness of the cup, A_{cup} (L^2) is surface the of the cup, and $k_{membrane}$ ($L^2 T M^{-1}$) is the hydraulic conductivity of the cup material. In the early water tensiometers and mercury tensiometers, the response times were slow, even in wet soil, because the sensitivity S was low. Advanced tensiometers, which have very small water reservoirs and highly sensitive pressure transducers nowadays have response times in the range of seconds, if measurements are done at wet soil conditions. A further improvement in the response characteristic of tensiometers results from minimizing the flow resistance of the porous cups. This is achieved by "high flow" material, where the bulk of the porous cup is made of a material with large pores, and just a skin on the outside of the tensiometer has pore sizes in the required range of small pores. In relatively dry soil, however, the response time of the instrument is primarily affected by the resistance of the

soil to unsaturated water flow around the tensiometer cup. This leads to response times in the range of minutes in the intermediate moisture range, up to days at the dry range.

Temperature Influence on Readings

Temperature changes and the formation of gas bubbles pose particular problems for tensiometric measurements. Temperature changes may cause spurious readings, because they affect the calibration of the pressure sensor, the expansion of the tensiometer shaft and the cup material, and the expansion of water. The combined effect is not easy to correct. Air-temperature fluctuations in the exposed head space of the tensiometer can cause large air-pressure changes, leading to misinterpretation of the soil water potential (Butters and Cardon, 1998). Warrick *et al.* (1998) showed in the limiting case of zero conductivity around the tensiometer cup that a cyclical variation of temperature of $35 \pm 15\,^{\circ}\mathrm{C}$ leads to variations in water pressure inside the cup of $\pm$70-cm water head. While cup impedance was found to be a negligible factor, their analysis suggested that conductivity of the soil immediately around the cup is the main factor governing temperature-induced pressure fluctuations inside the cup.

Frost

Measuring matric potentials at temperatures below $0\,^{\circ}\mathrm{C}$ is a challenge, since water in the tensiometer, which acts as pressure-transmitting agent freezes and the tensiometer fails. In some instances, the use of a glycol-water mixture or the combination of water in the measurement cell with oil above it toward the surface has been proposed. Finally, for larger depths ahead of the freezing front, it would be possible to obtain measurements with an advanced tensiometer (no continuous water column to the surface).

Formation of Gas Bubbles

If a tensiometer reading approaches its minimum pressure, gas bubbles (air or water vapor) may come into play, which reduce the sensitivity of the system. A large gas bubble in the system essentially deactivates the reaction. Gas entry can be caused by (i) exceeding the air entry value of the membrane, (ii) aggregation of small air bubbles, which result from dissolved air in the water that has been filled into the tensiometer, (iii) air that diffuses through the porous membrane, in particular over long measurement times, and (iv) by reaching the boiling pressure of the soil water. To reduce the effects of gas bubble formation within the tensiometer during a long-term operation, air stripper tensiometers have been introduced using a stripper tube in the porous cup to continually remove air (Miller and Salehzadeh, 1993). The tube contains droplets of water ("wet vacuum") to minimize net diffusion of water molecules into and out of the tube.

Tensiometer Performance in Drying Soils

When a soil is drying, the response times of tensiometers can become much longer, in particular, if their sensitivity is not high. In general, the decrease in performance is not caused by the conductance of the porous cup, but rather by the decreasing conductance of the surrounding soil, which becomes the limiting factor. The deterioration of the response characteristic in drying soils is, in particular, dramatic in cases where there is an air bubble in the head space of a tensiometer, as recommended for stitch tensiometers, since the expansion of the air bubble causes the sensitivity of the instrument to continuously decrease with increasing tension.

This is illustrated in Figure 5, where the interplay between the matric potential and the tensiometer response in a drying soil is modeled. The simulated domain is 2-D in cylindrical coordinates, with a depth of 100 cm and a radius of 75 cm. The tensiometer has a length of 30 cm and is placed 25 cm in the soil, with a porous cup of 50 mm length and 12.5 mm radius at its end. The tensiometer is originally filled with air-free water, but has an air bubble of 1-cm length at the top. The soil is a silty loam with a homogeneous initial capillary pressure of -10 kPa. Drying takes place with a transpiration rate of 5 mm day^{-1} through root water uptake, with roots being equally distributed along the upper 50 cm. Figure 5(b) shows how the tensiometer cup short-circuits pressure differences along the drying front in the surrounding soil matrix. With ongoing drying of the soil, the expansion of the air bubble in the tensiometer causes water to leave the tensiometer, but this effect is negligible up to a pressure of -50 kPa at about 12 days. After that, the unsaturated water flow in the soil surrounding the tensiometer is no longer high enough to keep the tensiometer reading in equilibrium with matric potential at some distance from the instrument. From then on, the tensiometer acts more and more as a local irrigation device, and the true pressure deviates far from the measured pressure. This analysis confirms experimental results by Reece (1996) who compared HDS measurements with tensiometric measurements and found tensiometers to be less responsive and measuring less negative potentials in the range from -70 to -90 kPa.

The important point is that there is no indication from the tensiometer reading itself about this deviation, except from the observation of an increasing length of the air bubble. The extent of these kinds of deviations in a specific practical situation depends on factors such as soil type (in particular, unsaturated conductivity), rate of transpiration, and root distribution.

The systematic deviation between true soil state and tensiometer reading can be amended to a large extent by the use of advanced tensiometers. This is illustrated in Figure 6, which shows typical tensiometer readings during

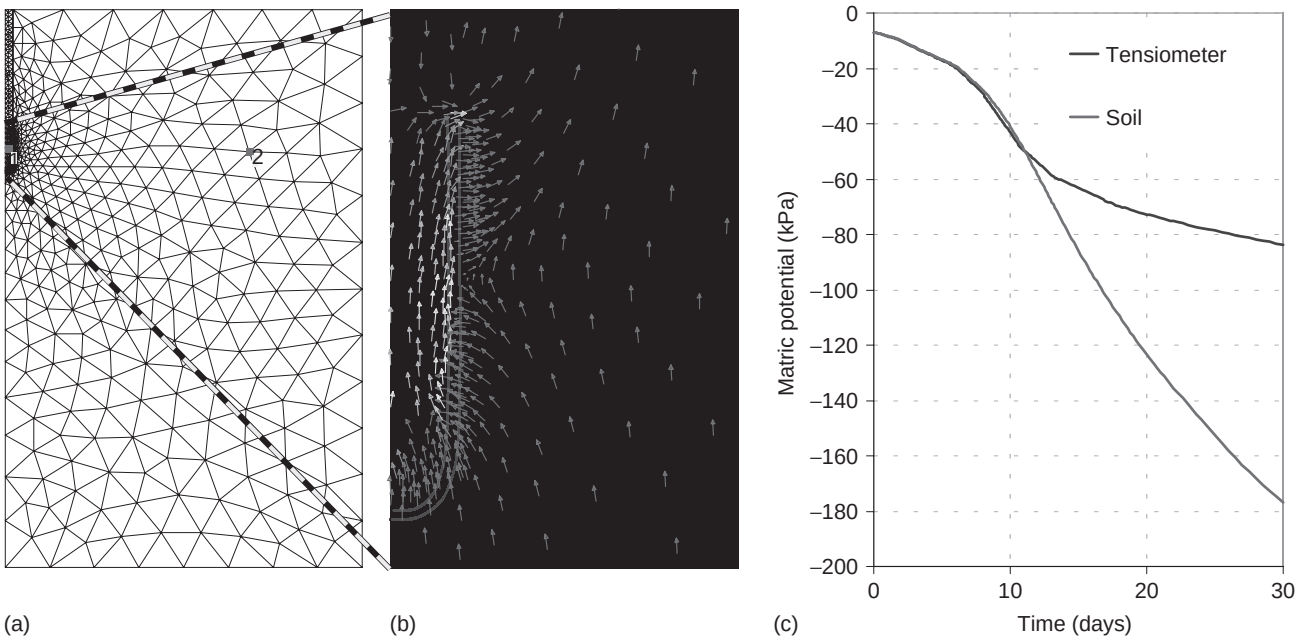


Figure 5 Hydraulic functioning of a vertical tensiometer with a small gas bubble. (a) Simulated domain. (b) Flow through the porous cup at $t = 30$ days. (c) True pressure head in the soil (grey, dashed) and apparent pressure by tensiometer readings (black). See text for further explanations. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

a vegetation period. The accordance of the readings of the instruments that are placed in parallel at 20 cm and 40 cm depth is excellent. Second, we see the influence of root water extraction in the form of daily oscillations. In the early period, when the roots are just in the upper soil, only the tensiometers at 20-cm depth show this influence. At later time, the roots' influence can be seen at the depths of 40, 60, and finally also at 80 cm. Upon close inspection, one sees that the two tensiometers at 40-cm depth are not equally affected by the roots, which indicates local differences in root activity in the direct vicinity of the two tensiometers.

Since defervescence cannot be avoided on the long range, boiling occurs in the tensiometer once its measurement range is reached. For the uppermost tensiometers, this is the case around May 27. Then, the build up of the vapor phase inside the membrane leads to water flowing out of the tensiometer, wetting the surrounding soil. The pressure inside the tensiometer will remain at the level of the temperature-dependent vapor pressure until all water inside the cup has evaporated. The pores of the membrane remain water filled (Figure 6).

When all water inside the porous cup has been absorbed by the soil, the water release from the porous membrane to the surrounding soil ceases and water potential starts to drop quickly. Soon the air entry point of the largest pores in the membrane is reached and some pores become conductive to air. This leads to ventilation of the cup, and the pressure inside the tensiometer cup rises steeply until it reaches atmospheric pressure. For the depicted tensiometers at 20-cm depth, this is the case on June 7. After that, the instrument can only be used if the cup is again filled with degassed water.

GYPSUM BLOCKS AND OTHER REFERENCE POROUS MEDIA

Measurement of soil matric potential in the intermediate pressure range between tensiometric (0 to −90 kPa) and field psychrometers (500 to 50 000 kPa) poses experimental challenge. Limitations in obtaining direct measurements resulted in reliance on alternative methods based on measuring water content or water content–related properties in well-characterized reference porous media that are in hydrostatic equilibrium with the surrounding soil. Typically, these methods require calibration to infer matric potential from the sensor response. Wettable porous media with pore-size distribution that leads to continuous and repeatable desaturation of water in the matric potential range of interest may be used for this purpose. The requirements for the porous materials to equilibrate with soil is a major obstacle for using these sensors in processes with high temporal dynamics. In particular, a mismatch between the pore-size distributions of the sensor and the soil can lead to hydraulic decoupling of the sensors, which is mostly relevant to coarse-textured soils (Wraith and Or, 2001).

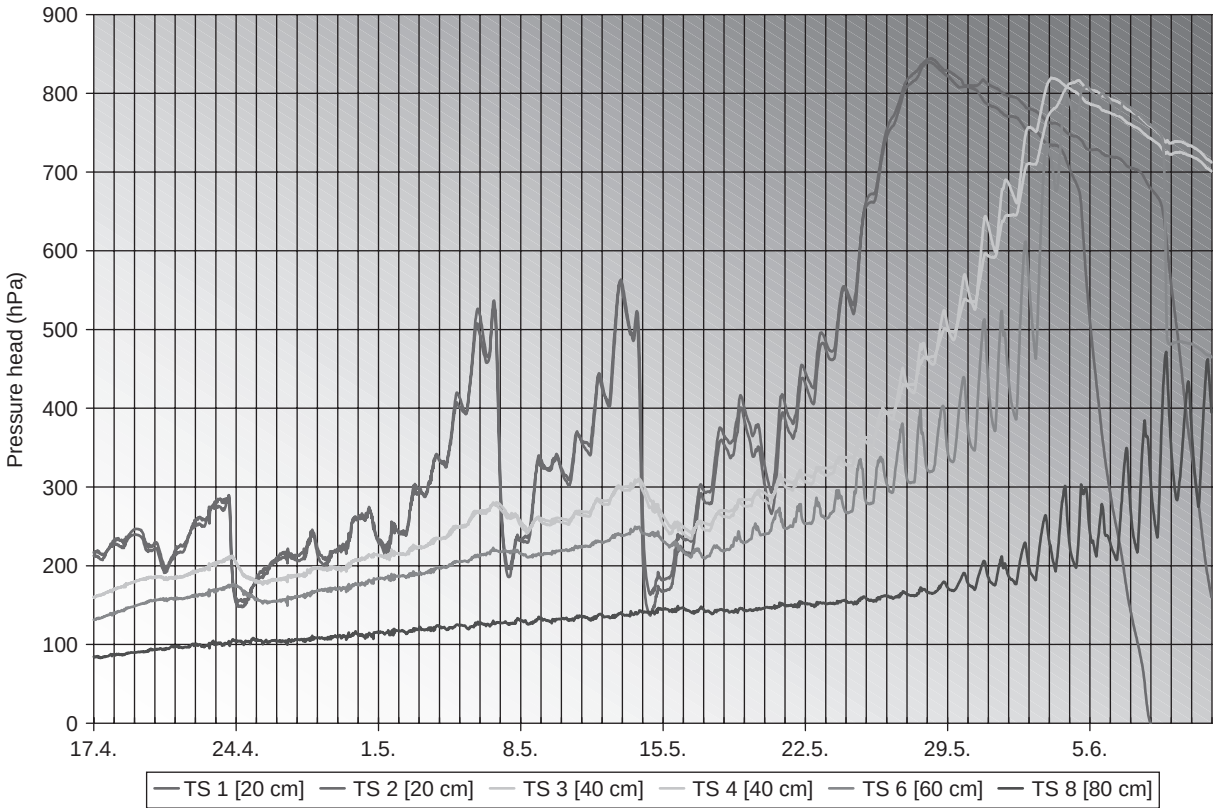


Figure 6 Typical tensiometer readings from a barley field during a growing period, interrupted by irrigation events (May 6, May 13) and some rain events (e.g., April 23, May 20). Two replicate tensiometers were installed at 20-cm depth (TS1, TS2) and 40-cm depth (TS3, TS4); further tensiometers were installed at depths of 60 and 80 cm (TS6 and TS8, resp.). The measurements show daily oscillations, caused by root water uptake. Roots extend in the early period to the upper tensiometers only, but then grow continuously deeper. TS1 and TS2 reach their measurement range limit on May 27. Then, water inside the tensiometer cup starts boiling and passes through the porous cup of the tensiometer, irrigating the surrounding soil. The decrease of the depicted head, which is the difference between atmospheric pressure p_a and the vapor pressure p_v inside the tensiometer, is due to an increase of p_v, caused by an increase in soil temperature (data not shown). From June 4 onwards, the tensiometer cup of TS1 starts leaking and the pressure difference drops to zero within a few days. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Classic porous reference materials are blocks of gypsum, fiberglass, or nylon. Recently, granular matrix sensors have become available, which consist of gypsum wafers embedded in granular matrix (Scanlon *et al.*, 2002). Measurements of water content are obtained either by electrical resistance, heat dissipation (see next section), or by measuring the electromagnetic properties of the reference porous medium. Gypsum blocks with embedded electrodes to measure the electric resistance have been used for more than 60 years in agricultural applications (Buoyoucos and Mick, 1940). The electrical conductivity in these blocks is buffered by the ion strength of the saturated gypsum solution ($\sim$0.2 S m^{-1}). This is contrary to nylon and fiberglass blocks, where the electrical conductivity is dependent also on the ionic strength of the soil solution. The problems of these sensors are a limited temporal stability, a long response time, appreciable hysteresis, and the need for individual calibration. Furthermore, the measurements are temperature dependent and must be corrected accordingly. Handling and disadvantages of these probes are well known and documented (Scanlon *et al.*, 2002), and it appears that not much development has taken place during the last decade.

More recently, activities have been directed toward methods where the water content of the reference porous medium is obtained from its dielectric permittivity (Or and Wraith, 1999; Wraith and Or, 2001). In 1996, Delta-T Devices introduced the so-called "equitensiometer", where a capacitance probe is embedded in a porous matrix. Noborio *et al.* (1999) embedded a portion of a TDR probe in porous gypsum, with the remainder of the electrode rods surrounded by soil. The signal travel times through the two sequential media were separately analyzed to obtain their respective apparent dielectric constant. The electric permittivity of the porous gypsum was calibrated against water content in the pressure plate apparatus. Or and Wraith (1999) introduced a TDR-based sensor, where disks of

porous ceramic and plastic, having a variety of pore sizes from 120 to 0.6 μm, were stacked within a coaxial cage. The pack was brought in in hydraulic equilibrium with the surrounding soil and its water content measured by TDR. The relationship between integral water content and water potential of the sensor's porous matrix was initially calibrated and subsequently used to infer the matric head of the surrounding soil. Both the studies of Or and Wraith (1999) and Noborio *et al.* (1999) identified a need for porous materials having a wide range of pore sizes. Or and Wraith (1999) reported a trade-off between the sensor's matric potential range and its sensitivity to changes in the surrounding soil. To provide a flexible alternative to probes constructed using porous disks, Wraith and Or (2001) proposed that previously characterized porous media having similar particle and/or pore-size distributions as those of soils or other media under investigation, may be used as the porous matrix for TDR-based water potential sensors.

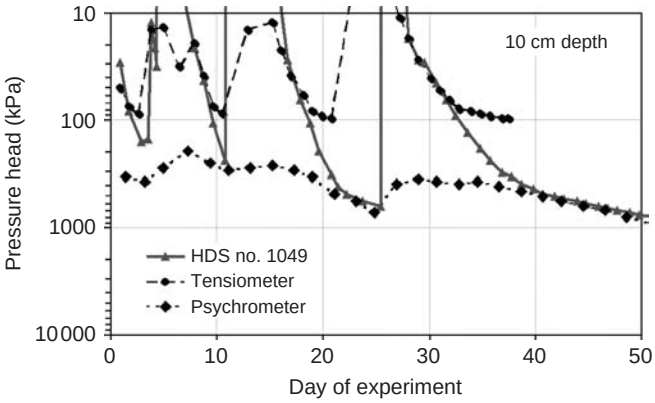


Figure 7 Changes in water potential measured by heat dissipation sensor, tensiometer, and psychrometer during a laboratory experiment [Redrawn from C.F. Reece, *Soil Sci. Soc. Am. J.*, **60**, 1022–1028 (1996)]. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

HEAT DISSIPATION MATRIC POTENTIAL SENSORS

Similar to gypsum blocks and granular matrix sensors, HDSs infer water potential by measuring a water content–dependent property in a porous medium, which is in hydraulic equilibrium with the surrounding soil. Fredlund (1992) compiled a brief historical review of HDSs showing that already 75 years ago Shaw and Baver (1939) demonstrated that soil water potential could be inferred from the rate of head dissipating into the soil. A HDS consists of a heating element and thermocouple embedded within a porous ceramic matrix (Phene *et al.*, 1971). The measurement is based on the rate of temperature rise from a line heat source, embedded in the cylindrical porous ceramic. The rate of heat dissipation in a porous medium is dependent on specific heat capacity, thermal conductivity, and density. Since these properties are greatly affected by the water content, heat dissipation can be related by a calibration relation to water potential.

Reece (1996) presented a thorough analysis of HDSs and concluded that the range of measurable matric potentials lies between −10 and −1000 kPa. In a comparison with tensiometers and psychrometers, he found that at no time did HDS response appear to lag tensiometer or psychrometer measurements under wetting or drying cycles (Figure 7). A significant advantage of HDSs is their insensitivity to solution salt content, in contrast to electric conductivity–based sensors. Additionally, sensors are relatively inexpensive (<$100 per sensor, Flint *et al.*, 2002). Often, variations in heat transfer properties between heater and ceramic of different sensors necessitates individual calibration. Flint *et al.* (2002) developed a normalization procedure that simplified calibration and presented temperature correction, using sensors from three sources and different calibration methods.

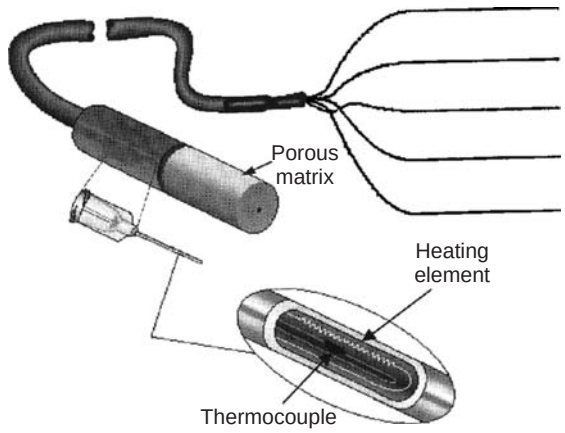


Figure 8 Schematic of heat dissipation sensor CSI 229 (Reproduced with courtesy of Campbell Scientific Inc., Logan, UT, USA)

The normalized calibration resulted in a mean absolute error of 23% over a matric potential range of −0.01 to −35 MPa.

Commercial availability of sensors is listed in Flint *et al.* (2002) and Scanlon *et al.* (2002). Sensors available from Campbell Scientific, Inc. (CSI, Logan, UT) measure matric potential in the range of approximately −0.01 to −100 MPa with sensitivity proportional to matric potential value. Resolution is approximately 1 kPa at matric potentials greater than −0.1 MPa and is capable of responding to changes in the matric potential of dry soils (see Figure 8).

VAPOR PRESSURE–BASED METHODS AND THERMOCOUPLE PSYCHROMETRY

Psychrometric measurements are based on equilibrium between liquid soil water and water vapor in the ambient

soil atmosphere. Water potential in the air phase is related to relative humidity, *RH*, through the Kelvin equation (Or and Wraith, 2002)

$$RH = \frac{e}{e_0} = \exp\left(\frac{M_w g h}{RT}\right) \quad (6)$$

where e is water vapor pressure, e_0 is saturated vapor pressure at the same temperature, M_w is the molecular weight of water ($0.0018\,\mathrm{kg\,mol^{-1}}$), g is the gravitational acceleration ($9.81\,\mathrm{m\,s^{-2}}$), h is the water potential per unit weight (m), R is the ideal gas constant ($8.31\,\mathrm{J\,K^{-1}\,mol^{-1}}$), and T is the absolute temperature (K). The relative humidity of the air can be determined from the dew-point temperature, using a chilled mirror. In water activity measurement devices (Decagon Inc., Pullman), water potential is measured by equilibrating the liquid-phase water of a sample with the vapor-phase water in the headspace of a closed chamber. A thermoelectric (Peltier) cooler controls the mirror temperature. A beam of infrared light is directed onto the mirror and reflected back to a photodetector, which detects the change in reflectance when condensation occurs on the mirror. A thermocouple attached to the mirror accurately measures the dew-point temperature.

Psychrometers measure the difference between a dry bulb and wet-bulb temperature. The dry bulb is at the temperature of the surrounding soils, the wet bulb at the temperature of an evaporating surface. The lower the humidity, the higher will be the rate of evaporation from the wet bulb, and thus the temperature depression below ambient. Since air is an effective diffusion barrier for most solutes, the corresponding water potential includes the osmotic and the matric potential. Rearranging (6) and taking a log-transformation leads to

$$h = \frac{RT}{M_w g} ln\left(\frac{e}{e_0}\right) \quad (7)$$

For the range of e/e_0 near 1, which is usually encountered in soils in humid climates, equation (7) can be simplified to

$$h = \frac{RT}{M_w g}\left(\frac{e}{e_0} - 1\right) = 0.471 \cdot 10^6\ T\left(\frac{e}{e_0} - 1\right) \quad (8)$$

for h in m.

A thermocouple psychrometer consists of a fine-wire chromel-constantan or other bimetallic thermocouple. A thermocouple is a double junction of two dissimilar metals. When the two junctions are subject to different temperatures, they generate a voltage difference (Seebeck effect). Conversely, when an electrical current is applied, the junction is heated or cooled, depending on the direction of the current (Peltier effect). For typical soil use, one junction of the thermocouple is suspended in a thin-walled porous ceramic or stainless screen cup buried in the soil, while another is embedded in an insulated plug to measure the ambient temperature at the same location (Figure 9). By an electrical current, the suspended thermocouple is cooled below the dew point until water condenses on the junction. The cooling current then stops, and as water evaporates, it draws heat from the junction, depressing it below the temperature of the surrounding air until it attains wet-bulb temperature. The difference in temperatures between the wet and dry bulb is related to the relative humidity by the psychrometer equation

$$\frac{e}{e_0} = 1 - \left(\frac{s + \gamma}{e_0}\right)\Delta T \quad (9)$$

where s is the slope of the saturation water vapor pressure curve ($s = \mathrm{d}e_0/\mathrm{d}T$), γ is the psychrometric constant ($\sim 0.067\,\mathrm{kPa\,K^{-1}}$ at 20°C), and ΔT is the temperature difference (K) between the dry and wet bulb.

An accurate determination of the temperature difference plays a critical role in psychrometric water potential determination. For water potential measurements to be accurate to $\sim 10^5$ kPa, temperature difference measurements need to be accurate to 0.005 °C. Psychrometers are therefore highly susceptible to thermal gradient effects and do not perform well at shallow soil depths. The necessity of equilibrium of different phases further causes a relatively slow response time. When the osmotic potential is negligible, the soil water potential measured by a psychrometer is nearly equal to the soil matric potential. In principle, soil psychrometers may be buried in a soil and left for long periods, although corrosion is a problem in some environments.

PotentiaMeter (model WP4, Decagon, WA) is a new device for rapid measurement of water potential using the chilled-mirror dew-point technique (described above) (see Figure 10). The vapor pressure in the headspace above a thermally equilibrated soil sample is computed as the saturation vapor pressure at dew-point temperature; with known sample temperature the water potential in the sample is determined as discussed in psychrometric measurements. Since both dew-point and sample surface temperatures are simultaneously measured, the need for complete thermal equilibrium is eliminated, and measurement time is reduced to less than five minutes. Readings are provided directly in MPa with accuracy for water potential range of 0 to -10 ± 0.1 MPa, and from 0 to -60 MPa $\pm 1\%$.

CONCLUDING REMARKS

In contrast with soil water content sensors, no single matric potential sensor is currently capable of covering the entire range of interest (10^{-1}–10^2 kPa tensiometers; 10^1–10^5 kPa reference porous media; and 10^5–10^8 kPa thermocouple psychrometers). The gap in measurement capabilities of these two important soil attributes is accentuated

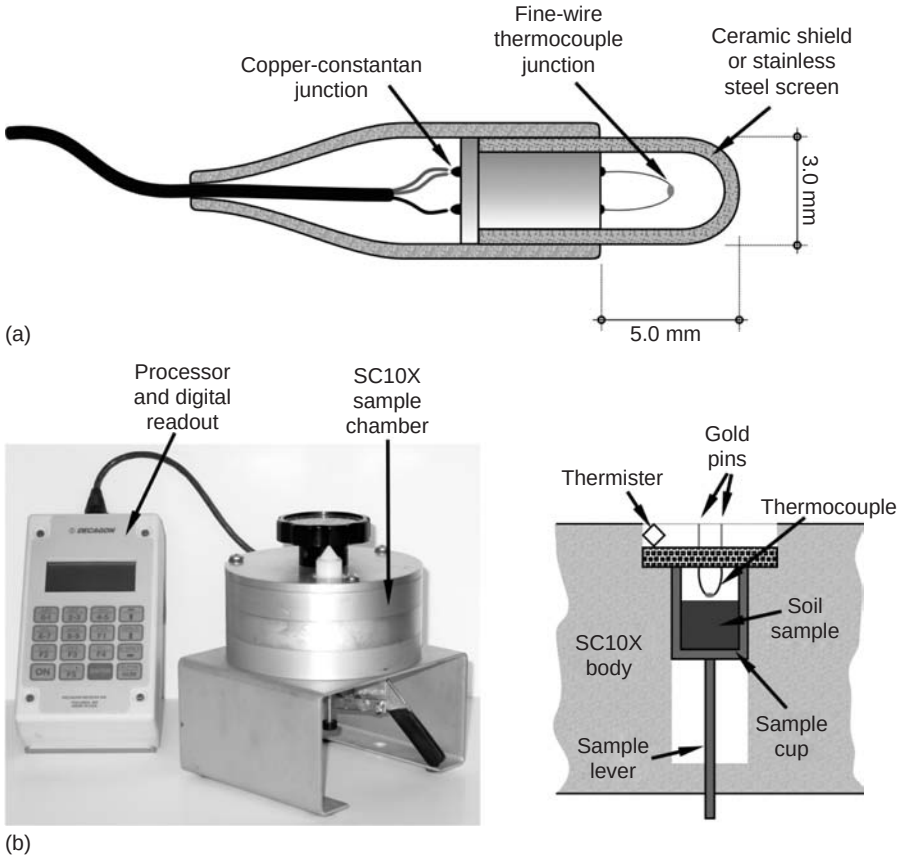


Figure 9 Field thermocouple psychrometer with porous ceramic 229 (Reproduced with courtesy of Wescor Inc., Logan, UT, USA). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 10 WP4 Dewpoint PotentiaMeter (Reproduced with courtesy of Decagon Devices Inc., Pullman, WA, USA). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

by rapid developments in water content detection using electromagnetic methods (TDR and other noninvasive microwave methods). Some progress has been made in the reliability and ease of use of tensiometers, which have evolved and are now considered among the most accurate and reliable soil hydration status measurements. Their limited measurement range, however, remains a severe limitation for field and laboratory studies.

Indirect methods, especially those based on hydraulic equilibration with reference porous media remain limited by technical challenges. It is to be expected that the combination of different porous materials in EPM, and the use of TDR or the use of HDSs to determine their composite water content, will bring certain progress. If the problem of a fully reproducible packing technique can be solved, the method proposed by Wraith and Or (2001) appears to be very promising. For the present, we concur with Smiles (2001), who claims that it "remains a challenge that there are still no good methods to measure suction in the field over any length of time, and comments on the osmotic tensiometer remain pertinent".

Finally, while water content measurements methods are rapidly evolving toward noninvasive estimation of water contents over large areas and large soil volumes (ground-penetrating radar, electrical tomography, passive microwave radiation), the methods for soil water energy state remain restricted to point measurements.

MANUFACTURERS

A broad overview over suppliers of soil water content sensing is given by the SOWACS web page (**http://www.sowacs.com/suppliers/index.html**). Instrumentation for water potential measurement can be found at

- **Adcon's Electrotensiometer** (**http://www.adcon.com/**) has a standard ceramic cup, but the signal from the vacuum sensor above the water-filled tube is conditioned to be compatible with Adcon's data acquisition units.
- **Campbell Scientific Inc**. manufactures dataloggers, data acquisition systems, and measurement and control products, and the heat dissipation matric water potential sensor.
- **Decagon Devices Inc**. (**http://www.decagon.com/**) produces and sells water activity meters.
- **Delta-T Devices Ltd** (**http://www.delta-t.co.uk/**) offer a range of electronic, pressure transducer tensiometers, including miniature and rugged-use models. Typical usage is in multiple arrays, automatically recorded by a field data logger.
- **Earth Systems Solutions** (**http://www.earthsystemssolutions.com/**) are the American distributors for various SDEC (French) tensiometers including accessories and electronic transducers for continuous logging.
- **Irrometer** (**http://www.irrometer.com/**). Original American Suppliers since 1951. They also produce the Watermark sensor.
- **SDEC's Tensionics** (**http://www.sdec-france.com/us/index.html**). A French company that also makes a capacitance sensor. They also have a "Tensiometer" (electronic readout unit), which is designed for use as a portable gauge for use with tensiometers.
- **SoilMoisture Equipment Corporation** (**http://www.soilmoisture.com/**) supply tensiometers as well as many other devices for monitoring soil and plant water potential Suppliers index.
- **UMS Inc. Munic**, Germany, (**http://www.ums-muc.de/**), produce a variety of high-quality tensiometers, including the new self-filling type TS1.

FURTHER READING

Dane J.H. and Topp G.C. (2002) Methods of soil analysis. *Physical Methods*, Part 4, SSSA Book Series No. 5, Soil Science Society of America: Madison.

REFERENCES

Bianchi W.B. (1962) Measuring soil moisture tension changes. *Agricultural Engineering*, **43**, 393–399.

Biesheuvel P.M., Raangs R. and Verweij H. (1999) Response of the osmotic tensiometer to varying temperatures: modeling and experimental validation. *Soil Science Society of America Journal*, **63**, 1571–1579.

Bocking K.A. and Fredlund D.G. (1979) Use of the osmotic tensiometer to measure negative pore water pressure. *Geotechnical Testing Journal*, **2**, 3–10.

Buckingham E. (1907) Studies in the movement of soil moisture. *Soils Bureau Bulletin*, Vol. 38, U.S. Department of Agriculture.

Buoyoucos G.J. and Mick A.H. (1940) *An Electrical Resistance Method for the Continuous Measurement of Soil Moisture Under Field Conditions*, Technical Bulletin, 172, Michigan Agricultural Experiment Station, East Lansing.

Butters G.L. and Cardon G.E. (1998) Temperature effects on air-pocket tensiometers. *Soil Science*, **163**, 677–685.

Durner W. and Lipsius K. (2005) Determining soil hydraulic properties. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Fisher J.C. (1948) The fracture of liquids. *Journal of Applied Physics*, **19**, 1062–1067.

Flint A.L., Campbell G.S., Ellett K.M. and Calissendorff C. (2002) Calibration and temperature correction of heat dissipation matric potential sensors. *Soil Science Society of America Journal*, **66**, 1439–1445.

Flühler H. and Roth K. (2004) Physik der ungesättigten Zone. Lecture notes, Institute of Terrestrial Ecology, Swiss Federal Institute of Technology Zurich, Switzerland, and Institute of Environmental Physics, University of Heidelberg Germany.

Fredlund D.G. (1992) Background, theory, and research related to the use of thermal conductivity sensors for matric suction measurement. *Advances in Measurement of Soil Physical Properties: Bringing Theory into Practice*, SSSA Special Publication 30, Soil Science Society of America: Madison, pp. 249–261.

Gardner W., Israelsen O.W., Edlefsen N.E. and Clyde D. (1922) The capillary potential function and its relation to irrigation practice, (Abstract). *Physical Review*, **20**, 196.

Gee G.W. and Ward A.L. (1999) Innovations in two-phase measurements of soil hydraulic properties. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), Riverside, 22–24 October 1997, pp. 241–270.

Haines W.B. (1927) Studies in the physical properties of soils: IV. A further contribution on the theory of capillary phenomena in soil. *Journal of Agricultural Research (Cambridge)*, **17**, 264–290.

Hassanizadeh S.M., Celia M.A. and Dahle H.K. (2002) Dynamic effect in the capillary pressure–saturation relationship and its impacts on unsaturated flow. *Vadose Zone Journal*, **1**, 38–57.

Livingston B.E. (1908) A method for controlling plant moisture. *Plant World*, **11**, 39–40.

Livingston B.E. (1918) Porous clay cones for the auto–irrigation of potted plants. *Plant World*, **21**, 202–208.

Marthaler H.P., Vogelsanger W., Richard F. and Wierenga P.J. (1983) A pressure transducer for field tensiometers. *Soil Science Society of America Journal*, **24**, 624–627.

Miller E.E. and Salehzadeh A. (1993) Stripper for bubble-free tensiometry. *Soil Science Society of America Journal*, **57**, 1470–1473.

Noborio K., Horton R. and Tan C.S. (1999) Time domain reflectometry probe for simultaneous measurement of soil matric potential and water content. *Soil Science Society of America Journal*, **63**, 1500–1505.

Or D. (2001) Who invented the tensiometer? *Soil Science Society of America Journal*, **65**, 1–3.

Or D. and Wraith J.M. (1999) A new TDR-based soil matric potential sensor. *Water Resources Research*, **35**, 3399–3407.

Or D. and Wraith J.M. (2002) Soil water content and water potential relationships. In *Soil Physics Companion*, Warrick A. (Ed.), CRC Press: Boca Raton, pp. 49–84.

Peck A.J. and Rabbidge R.M. (1969) Design and performance of an somotic tensiometer for measuring capillary potential. *Soil Science Society of America Proceedings*, **33**, 196–202.

Phene C.J., Hoffman G.J. and Rawlins S.L. (1971) Measuring soil matric potential in situ by sensing heat dissipation within a porous body: I. Theory and sensor construction. *Soil Science Society of America Proceedings*, **35**, 27–33.

Reece C.F. (1996) Evaluation of a line heat dissipation sensor for measuring soil matric potential. *Soil Science Society of America Journal*, **60**, 1022–1028.

Richards L.A. (1928) The usefulness of capillary potential to soil moisture and plant investigators. *Journal of Agricultural Research (Cambridge)*, **37**, 719–742.

Richards L.A. (1949) Methods of measuring soil moisture tension. *Soil Science*, **68**, 95–112.

Scanlon B.R., Andraski B.J. and Bilskie J. (2002) Miscellaneous methods for measuring matric or water potential. In *Methods of Soil Analysis, Part 4, Physical Methods*, SSSA Book Series No. 5, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 643–670.

Schultze B., Ippisch O., Huwe B. and Durner W. (1999) Dynamic nonequilibrium in unsaturated water flow. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.T.h, Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 877–892.

Shaw B. and Baver L.D. (1939) An electrothermal method for following moisture changes of the soil in situ. *Soil Science Society of America Proceedings*, **4**, 78–83.

Sisson J.B., Gee G.W., Hubbell J.M., Bratton W.L., Ritter J.C., Ward A.L. and Caldwell T.G. (2002) Advances in tensiometry for long–term monitoring of soil water pressures. *Vadose Zone Journal*, **1**, 310–315.

Smiles D. (2001) Book review on Characterization and measurement of the hydraulic properties of unsaturated porous media. *Proceeding of an International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, Vols. 1 and 2, Riverside, 22–24 October 1997; *Soil Science*, **166**, 150–152.

Tamari S., Gaudu J.C. and Simmoneau T. (1993) Tensiometric measurement with metastable state of water under tensions. *Soil Science*, **156**, 149–155.

Tandeske D. (1991) *Pressure Sensors: Selection and Application*, Marcel Dekker: New York.

Topp G.C. and Ferré T.P.A. (2005) Measuring soil water content. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Topp G.C. and Ferré P.A. (2002) 3.1 Water Content. In *Methods of Soil Analysis, Part 4, Physical Methods*, Dane J.H. and Topp G.C. (Eds.), SSSA Book Series No. 5, Soil Science Society of America: Madison, pp. 417–421.

Vachaud G. and Thony J. (1971) Hysteresis during infiltration and redistribution in a soil column at different initial water contents. *Water Resources Research*, **7**, 111–120.

Warrick A.W., Wierenga P.J., Young M.H. and Musil S.A. (1998) Diurnal fluctuations of tensiometric readings due to surface temperature changes. *Water Resources Research*, **34**, 2863–2870, doi:10.1029/98WR02095.

Watson K.K. (1967) A recording field tensiometer with rapid response characteristics. *Water Resources Research*, **5**, 33–39.

Wraith J.M. and Or D. (2001) Soil water characteristic determination from concurrent water content measurements in reference porous media. *Soil Science Society of America Journal*, **65**, 1659–1666.

Young M.H. and Sisson J.B. (2002) 3.2.2 Tensiometry. In *Methods of Soil Analysis, Part 4, Physical Methods*, Dane J.H. and Topp G.C. (Eds.), SSSA Book Series No. 5, Soil Science Society of America: Madison, pp. 575–608.

Zheng Q., Durben D.J., Wolf G.H. and Angell C.A. (1991) Liquids at large negative pressures: water at the homogeneous nucleation limit. *Science*, **254**, 829–832.

74: Soil Hydraulic Properties

WOLFGANG DURNER[1] AND HANNES FLÜHLER[2]

[1] *Institute of Geoecology, Department of Soil Physics, Braunschweig Technical University, Langer Kamp 19c, D-38106 Braunschweig, Germany*
[2] *Institute of Terrestrial Ecology, Swiss Federal Technical University at Zürich, Schlieren, Switzerland*

The accurate knowledge of hydraulic properties for unsaturated soils is critical in addressing problems in a variety of disciplines such as hydrology, ecology, environmental sciences, soil science, and agriculture. The purpose of this paper is to review the characterization of unsaturated soil hydraulic properties for their applicability in models simulating unsaturated water transport. In a theoretical section, we present the fundamentals for the definitions of the hydraulic properties in the framework of the continuum theory as well as provide the common parameterizations of the hydraulic functions. The characterization and parameterization of hysteresis and the phenomenon of dynamic effects in hydraulic properties are addressed in subsequent sections. Finally we discuss issues related to the handling of spatial and temporal variability, including geostatistical characterization, upscaling, and scale dependency of effective properties. The review closes with a summary on limits and opportunities for modeling water transport with Richards' equation.

INTRODUCTION

The accurate characterization of the hydraulic properties for unsaturated soils is critical in addressing many problems in hydrology, ecology, environmental sciences, soil science, agriculture, and other disciplines. Knowledge of the hydraulic properties is required in nearly all basic and applied aspects of soil, water, nutrient, and salinity management research (van Genuchten *et al.*, 1999). This chapter describes how hydraulic functions are defined in the framework of the continuum theory by averaging the microscopic phase properties (see Section "What are soil hydraulic properties?"). The hydraulic properties must be parameterized in closed-form expressions for their later application in simulation models. In Section "Parameterization of hydraulic functions", we discuss common parameterizations and illustrate in particular some problems and pitfalls associated with the widely used van Genuchten/Mualem model (van Genuchten, 1980). It is common knowledge that the hydraulic properties of soils are hysteretic in nature, which is especially significant with regard to transient water and solute transport. Nevertheless, there still exists very little standard representation of hysteresis in numerical models and it continues to remain a poorly modeled phenomenon. Section "Hysteresis" briefly reviews the current state of knowledge with respect to the parameterization and incorporation of hysteresis into numerical models.

Frequent observations on water flow phenomena in unsaturated soils under transient conditions show that the concept of uniquely occurring soil hydraulic properties must be questioned, since the relationship between matric potential and water content may depend on boundary conditions. Section "Dynamic effects" reviews the history of empirical evidence for these so-called *dynamic effects* and presents approaches on how to deal with them in numerical modeling. The final part of the paper, Section "Variability", discusses the impact of soil variability on the definition and measurement of effective hydraulic properties for large-scale applications. We show how geostatistical methods in combination with the concept of "scaling" are used to handle the variability of unsaturated hydraulic properties and we further illustrate the difference between mean and effective hydraulic conductivity in an example.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

WHAT ARE SOIL HYDRAULIC PROPERTIES?

Unsaturated soils contain two fluid phases, namely, soil water and soil air, and an immobile phase, the soil matrix. In an universal sense, "soil hydraulic properties" describe macroscopic relations between the chemical potential, the phase concentration, and the transmission behavior of water and gases in soil. These relations depend on a multitude of factors, including temperature, pore space geometry, surface properties of the soil matrix, chemical composition of the soil solution, and properties of the wetting and nonwetting fluids that occupy the complementary parts of the pore space. Two-phase flow systems have historically been described using fluid pressures and volumetric saturations as primary variables. The capillary pressure, defined as the difference between fluid phase pressures under equilibrium conditions, is related algebraically to the saturation by a nonlinear relationship, which is usually highly hysteretic. Fluxes of water and air are related to the potential gradients of the respective phases. The coefficient of proportionality between flux and gradient involves the relative permeability, which is a nonlinear function of saturation. Macroscopic mass balance equations, augmented by these constitutive relationships, represent the state of the art in mathematical descriptions of two-phase flow in porous media (Held and Celia, 2001).

Representative Elementary Volume and Measurement Window

On a microscopic scale, water content is either one (in the water phase) or zero (else). To come to a reasonable definition of water content, we must average the phase density over a certain soil volume. A practical choice is to choose the smallest volume that contains all structural elements of the pore system in sufficient abundance. We call this minimum volume *representative elementary volume*, REV (Hubbert, 1956; Bear, 1972). In fine-textured nonstructured soils, such as fine sands or loess, volumes of $1\,cm^3$ or less might be sufficiently large to yield a reliable average for soil porosity. Buchter *et al.* (1994) estimated the REV of the porosity of a Rendzina with >50% stones to be $1.5\,dm^3$. In soils with shrinkage cracks, such as vertisols, the REV is as large as the whole soil profile. Since large structures will often extend preferentially into one spatial direction, it follows that the size of the REV will vary in different spatial directions. In general, the REV for repacked samples, which contain the soil fraction with particle diameters less than 2 mm, will be much smaller than the REV of undisturbed samples, since stones, worm channels, roots, cracks, and lenses are large structures and must be contained in an REV in sufficient numbers. The question as to how well and accurate we can measure hydraulic properties will depend on the relation of the REV scale to the size of the measuring window.

If we extend the REV concept to characterize structural properties of natural soil, we see that no single REV exists. Enlarging the averaging volume will lead again and again to the inclusion of new structural elements of larger size (Figure 1). If, on the other hand, we extend the sampling region keeping the sample volume constant, we would observe that the average variability increases continually. This shows that observed variability of a measured soil hydraulic property will depend on the size of the measuring window, on the size of structural elements of the system, and on the extent of the region of sampling. In other words: the question whether a property is "homogeneous" or "heterogeneous" depends on one hand on the extent and the structure of the system under consideration, and on the other hand, on the averaging volume. This implies that an adequate sampling volume must be adapted to the state property to be measured, and also to the system to be described.

Continuum Approach, Darcy's Law, Richards' Equation

Averaging the local phase fractions or phase properties in an REV, and mapping the resulting value to the central

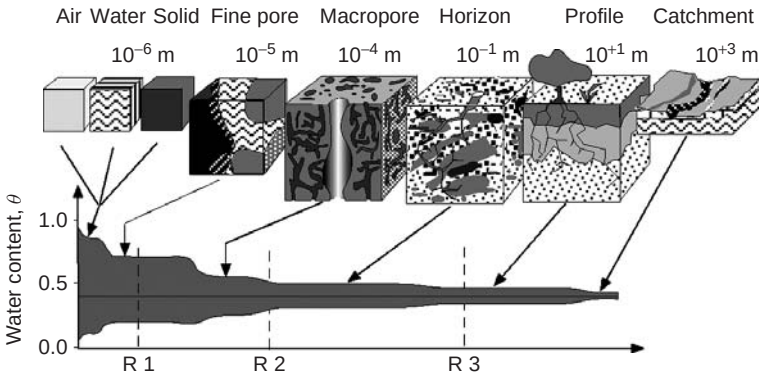


Figure 1 Scale dependence of water content variability (Courtesy of H. Flühler and K. Roth). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

point of the averaging volume leads to the definition of spatially continuous variables such as water content, air content, soil density, porosity, or water potential. The state variable "water content" thus represents the hydraulic condition of the soil at any given time and location as a result of hydraulic processes, which in turn are governed by soil hydraulic properties. This "continuum approach" (Cushman, 1984) enables us to relate in a mass balance the temporal changes of water content at a point to spatial gradients of the water flux, here noted in a one-dimensional form

$$\frac{\partial \theta}{\partial t} = -\frac{\partial q}{\partial z} + s \tag{1}$$

where θ is volumetric water content [$L^3\,L^{-3}$], q is volumetric water flux [$L^3\,L^{-2}\,T^{-1}$], t is time [T], z is a spatial coordinate [L], and s is a source/sink term [T^{-1}]. The water flux is driven by a gradient of the water pressure, p_w,

$$q = -K^*(\theta)\frac{\partial (p_w - \rho g)}{\partial z} \tag{2}$$

where K^* is the hydraulic conductivity [$L^3\,T\,M^{-2}$] and p_w is water pressure [$M\,L^{-1}\,T^{-2}$], ρ is the density of water [$M\,L^{-3}$], g is the gravitational acceleration [$L\,T^{-2}$]. Water pressure is an expression of water potential in units of energy per volume (Durner and Or, 2005; *see* **Chapter 73, Soil Water Potential Measurement, Volume 2**). The total water potential, ψ_{tot}, is affected by the position in the gravity field (gravitational potential, ψ_g), by forces due to water sorption on soil particle surfaces and capillary forces in the soil pore system (matric potential, ψ_m), by dissolved substances in the soil solution (osmotic potential, ψ_o), by forces due to soil swelling and overburden (overburden potential, ψ_Ω), and by forces that result from air pressure acting on water–air interfaces (pneumatic potential ψ_p). If a pore system is rigid, nonswelling, isotropic, and osmotic gradients and flow resistance of the nonwetting fluid are negligible, isothermal soil water movement is governed just by gradients in the matric, and the gravitational potential (Jury *et al.*, 1991; Kutilek and Nielsen, 1994).

In soil physics, it is customary to express the water potentials in units of energy per weight, that is, as heads [L]. Neglecting the pneumatic, osmotic, and overburden components of the water potential, and considering that the gradient of the gravitational potential is equal to -1 (the direction of the spatial coordinate taken positive downward) leads to

$$q = -K(\theta)\left(\frac{\partial h}{\partial z} - 1\right) \tag{3}$$

where h is the pressure head [L] and $K(\theta)$ is the hydraulic conductivity [$L\,T^{-1}$]. Equation (3) expresses the Darcy–Buckingham law (Darcy, 1856; Buckingham, 1907) and defines the saturated/unsaturated hydraulic conductivity. It implies that the hydraulic conductivity is a system property that can be determined only by an inverse approach, that is, by matching its value to be consistent with an observed dynamic system behavior (Durner and Lipsius, 2005; *see* **Chapter 75, Determining Soil Hydraulic Properties, Volume 2**). For saturated water flow, K will not depend on h, and equation (3) is linear. This is contrary to the unsaturated case, and makes it simple to invert the equation analytically by simple rearrangement.

Inserting equation (3) in equation (1), and replacing the dependency of K on θ by h leads to the pressure-form of the one-dimensional Richards' equation

$$C(h)\frac{\partial h}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} - 1\right)\right] + s \tag{4}$$

where *C(h)* is the specific water capacity [L^{-1}], defined by the change of water content with pressure head, $C = \partial\theta/\partial h$. Equation (4) was first derived by Richards (1931) and is the fundamental model for describing water flow in the unsaturated zone on the macroscopic scale (i.e. the scale where we usually measure soil hydraulic properties). The model is completed by appropriate initial and boundary conditions, and knowledge of the coefficients *C(h)* and *K(h)*. Since C and K are nonlinearly dependent on h, solution of equation (4) requires generally numerical methods. The neglecting of the pneumatic potential implies that equation (4) is valid only if gradients in the pressure of the nonwetting fluid are negligible, or – in other words – if air is free to move in the soil at any system state. This is not always the case, neither in nature nor in measurement experiments. Equation (4) is frequently used as a process model for water transport at much larger scales. The coefficients C and K are then used as *effective* properties, which have the same names as for the local scale. However, their values are no longer consistent with the local, REV-based definition, as shown in Section "Variability".

In analogy to a diffusion process, water transport can also be described by

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[D(\theta)\left(\frac{\partial \theta}{\partial z}\right)\right] - \frac{\partial K(\theta)}{\partial z} + s \tag{5}$$

where D is the water diffusivity [$L^2\,T^{-1}$]. For horizontal water transport and without sinks or sources, equation (5) is formally equal to Fick's second law of diffusion. Comparison of equation (4) and (5) shows that the water diffusivity is given by the ratio of conductivity and specific water capacity,

$$D(\theta) = \frac{K(\theta)}{C(\theta)} \tag{6}$$

The coefficient D is highly water content-dependent and expresses a propagation velocity for water content changes.

It is highest for low damping (low capacity) and high hydraulic conductivity, which is the case for conditions near water saturation. Equation (5) is less frequently applied than the pressure head form of Richards' equation, (4), because its use is restricted to purely unsaturated soils, and, for vertical transport, the knowledge of the hydraulic conductivity function is also required. Therefore, our discussion will be restricted to the water retention curve (WRC) and the conductivity curve.

Retention Curve and Hydraulic Conductivity Curve

In the remainder of this chapter, we will use the term "soil hydraulic properties" for the constitutive relationships $\theta(h)$ and $K(\theta)$, or $K(h)$, as defined by equations (3) and (4) in the framework of the continuum approach. The relationship $\theta(h)$ is called *water retention curve, WRC* (Also, "water retention characteristic", "soil moisture characterisitic", "capillary pressure – saturation relationship", "water characterisitic curve", "water content – matric potential curve", "pF curve"). In principle, it can be determined by monitoring simultaneously the state variables h and θ at an identical point in space during a hydraulic process that changes the systems state. The dependence of the hydraulic conductivity, K, on water content is called *hydraulic conductivity curve*. It can be determined from simultaneous flux and head gradient measurements by inverting equation (4), or more simply, equation (3).

A multitude of empirical investigations lead to general expectations about the shapes of the relationships $\theta(h)$, $K(h)$ and $K(\theta)$ for different soils (Carsel and Parrish, 1988; Schaap, 2005; *see* **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**). Figure 2 shows examples of hydraulic properties for a sand, a silt, and a clay. The depicted functions have been derived by a neural network prediction with the program Rosetta (Schaap *et al.*, 2001). They are based on many measurements documented in the UNSODA soil properties database (Nemes *et al.*, 2001) and follow the parameterization of van Genuchten/Mualem (see Section "Parameterization of hydraulic functions"). It is important to note that these "typical" shapes of hydraulic properties are largely based on measurements on sieved, repacked samples. Compared to undisturbed soils, significant differences must be expected, in particular in the range near saturation, because of the structural pore system (Durner, 1994). For an illustration, see Figure 1 in Durner and Lipsius (2005); *see* **Chapter 75, Determining Soil Hydraulic Properties, Volume 2**.

PARAMETERIZATION OF HYDRAULIC FUNCTIONS

Water Retention Functions

For use in simulation models of unsaturated water transport, the constitutive relationships must be expressed in a continuous way over the whole moisture range, from dryness to saturation. This requires interpolation and smoothing of measured data, and it is common to express these interpolations by parametric functions. Numerous expressions have been proposed to describe the water retention characteristic (Brooks and Corey, 1964; King, 1965; Brutsaert, 1967; Visser, 1968; Taylor and Luthin, 1969; Laliberte, 1969; Farrel and Larson, 1972; Ahuja and Swarzendruber, 1972; Rogowski, 1972; Campbell, 1974; Su and Brooks, 1975; D'Hollander, 1979; Simmons *et al.*, 1979; van Genuchten, 1980; Kovacs, 1981; Russo, 1988; Kosugi, 1996; Assouline *et al.*, 1998). All these curves describe a continuous change of water content from a maximum value, θ_s, which is called *saturated water content*, toward a minimum value, θ_r, which

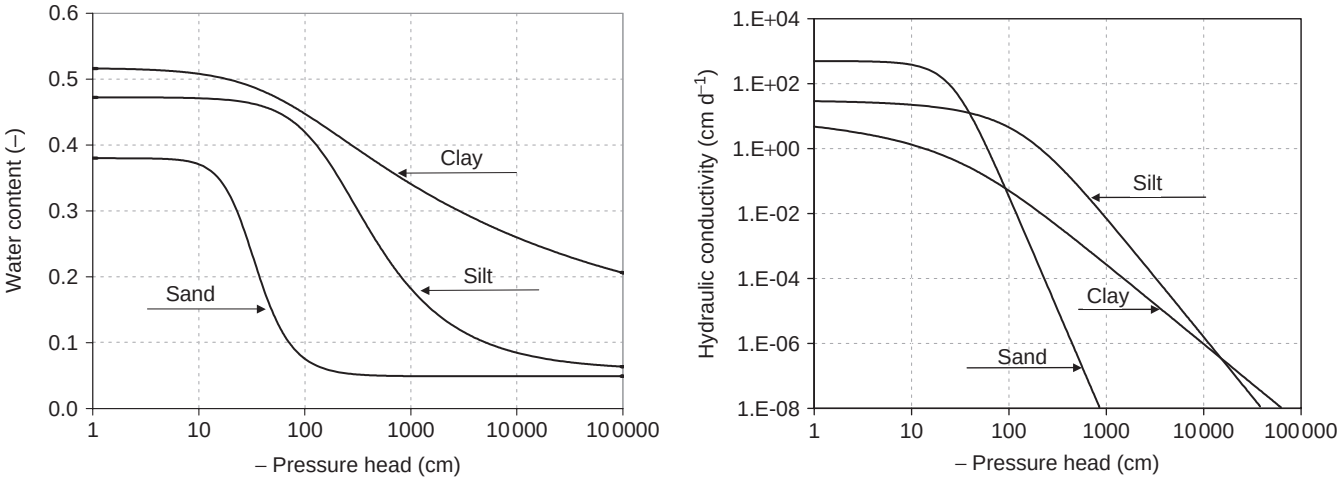


Figure 2 Typical hydraulic properties of differently textured soils (a) Water retention curves. (b) Unsaturated conductivity curves

is called *residual water content*. The transition from full to partial saturation takes place at some characteristic pressure head (often called *bubbling pressure* or *air entry pressure*) that is related to a characteristic width of the largest pores in the porous medium, and thereby depends on soil texture and structure. The slope of the WRC at pressures lower than the air entry pressure depends on the width of the pore-size distribution. Thus, all parametric expressions need, in general, at least four adjustable parameters, and differences in their ability to describe observed data are generally small (Buchan and Grewal, 1990).

Empirical studies (van Genuchten and Nielsen, 1985) showed that out of the functions listed above, the expression proposed by van Genuchten (1980) is particularly suited,

$$S_e = (1 + (\alpha|h|)^n)^{-m} \tag{7}$$

where $S_e = (\theta - \theta_r)/(\theta_s - \theta_r)$ is a scaled water content, called *effective saturation*, and $\alpha > 0$ [L^{-1}], $n > 1$ [$-$] and $m > 0$ [$-$] are empirical curve shape parameters. The inverse of α is related to the air entry pressure head, n is related to the width of the pore-size distribution between saturation and air entry pressure, and the product nm is related to the width of the pore-size distribution between air entry pressure and dryness. In practical use, the van Genuchten equation (7) is mostly applied with the constraint $m + 1/n = i$ $(i = 1, 2, \ldots)$. This reduces the flexibility of the function, but has the advantage that it yields closed-form expressions for the hydraulic conductivity functions when inserted into the conductivity prediction models of Mualem (1976), Burdine (1953), or Childs and Collis-George (1950). This makes it particularly easy to use the hydraulic relationships in numerical models (Luckner *et al.*, 1989), and thus the constrained van Genuchten/Mualem model has become *de facto* a standard in numerical modeling of water transport in unsaturated porous media.

When hydraulic properties are expressed by parametric functions, the process of measuring hydraulic properties is reduced to the process of estimating function parameters. In doing this, it is of crucial importance to be aware that there is no *a priori* evidence that the selected empirical relationship is indeed suited for a given soil. In particular, any extension beyond the moisture range covered by the actual measurement method is (sometimes reasonable) speculation. Accordingly, the determination of optimal model parameters should be combined with a measure that tests the adequacy of the chosen functional model approach (Finsterle, 1999). Unfortunately, this is rather seldom done in practice, because it requires sufficient measurements over the range of interest to be statistically significant.

The approximate nature of the van Genuchten function and other low-parameterized functions as models for the effective pore-size distribution becomes evident in the asymptotic behavior toward saturation and dryness. From physical reasoning, finite values for both the water capacity function and the slope of the conductivity function are to be expected over the whole moisture range, but for values of $n < 2$, the slope of the van Genuchten/Mualem conductivity function becomes infinity close to saturation. Constraining the n parameter to values greater than 2 is impractical since empirical investigations for all textural soil groups, except for sands, yield optimal fits for $n < 2$ (Carsel and Parrish, 1988). Also, toward dryness, the residual water content must be regarded as a pure empirical fitting parameter, and analyses of both, water retention data (Rossi and Nimmo, 1994) and conductivity data, (Tuller and Or, 2001) show that the usual functions are not capable in describing the measured values in the dry range appropriately. On the basis of thermodynamic reasoning, the true equilibrium water contents at pressure head ranges beyond -10^5 cm are clearly smaller than those given by the usual parameterizations, approaching oven dryness at pressure heads of $h = -10^7$ cm. For the practice of simulating water transport in the liquid phase, however, the use of residual water concept appears to yield reasonable results.

The above-cited parametric expressions for WRCs describe measured retention data well if the soils do not exhibit a distinct secondary pore system. To describe naturally structured soils, more flexible expressions have been introduced. These are either constructed by superposition of basic shape functions (Othmer *et al.*, 1991; Durner, 1992; Rossi and Nimmo, 1994) or by piecewise combinations of local shape functions, such as splines or Hermitian basis functions (Erh, 1972; Rossi and Nimmo, 1994; Mohanty *et al.*, 1997; Kastanek and Nielsen, 2001; Prunty and Casey, 2002; Bitterlich *et al.*, 2004). To determine these types of functions accurately, the demand on the precision and range of measurement methods is high (Zurmühl and Durner, 1998).

The consequences of using oversimplified hydraulic functions in practical applications depends on the extent and nature of the deviations between true and parameterized curves, on the range and type of applied boundary conditions, and on the purpose of the measurements. In general, even small changes in hydraulic descriptions can cause significant changes in hydraulic behavior, which is a prerequisite for the identification of hydraulic properties from transient-flow experiments (Vogel *et al.*, 2001; Lambot *et al.*, 2004). But for soil classification purposes, such as the estimate of plant available water, simplified descriptions may perfectly serve its purpose. For use in models of water and solute transport, spatial variability of soil hydraulic functions may be so large that seemingly small errors in the representation of the properties at a point scale appear irrelevant. Deviations of systematic nature, however, can have systematic and significant consequences for the predicted hydraulic conductivity curves

(Durner, 1994) and thus lead to a different effective system behavior, regardless of the superposed local variability. For very demanding applications, for example, tests of conductivity prediction models, investigations on dynamic flow behavior, or derivations of pedotransfer functions (Schaap, 2005; *see* **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**), systematic biases induced by oversimplified functions are intolerable since they lead to misinterpretations of comparisons between observations and models.

Hydraulic Conductivity Functions

Similar to the water retention characteristic, the dependence of the hydraulic conductivity on water content, water saturation, or pressure head, is generally expressed by a simple parametric expression. Common to all expressions is the scaling of a shape function, the so-called *relative conductivity function* $K_r(\theta)$, to the saturated conductivity, K_s

$$K(\theta) = K_s \cdot K_r(\theta) \tag{8}$$

Parametric expressions for the relative conductivity function are less manifold, as compared to the water retention function, because the conductivities are often coupled to the retention model. Simple empirical expressions for the $K(\theta)$ and $K(h)$ relation have been proposed, amongst others, by Gardner (1958), Brooks and Corey (1964), King (1965), and Rijtema (1965). Most frequently used are the exponential model (Gardner, 1958)

$$K(\theta) = K_s \cdot \mathrm{e}^{\alpha h} \tag{9}$$

and the power function (e.g. Averjanov, 1950; Irmay, 1954, Brooks and Corey, 1964; Ahuja and Swarzendruber, 1972),

$$K(\theta) = K_s \cdot S_e^{\beta} \tag{10}$$

where α and β are empirical coefficients. The combination of statistical pore-bundle conductivity models, for example, of Burdine (1953) or Mualem (1976), with the retention function of van Genuchten (1980) leads to closed-form expressions, which are sometimes used as empirical "stand-alone" expressions of the conductivity curve. The van Genuchten/Mualem model (van Genuchten, 1980) is most frequently used,

$$K_r(S_e) = S_e^l \left[1 - \left(1 - S_e^{1/m}\right)^m\right]^2 \tag{11}$$

where the parameter l [−] is called *tortuosity factor*, and m is related to the retention curve parameter n in equation (7) by $m = 1 - 1/n$. When applying the van Genuchten/Mualem conductivity model (11), the user should be aware of a model artifact. For small values of the parameter n, an unrealistic steep decrease of the predicted conductivity near saturation occurs, because the function considers no finite air entry pressure. Modifications of the van Genuchten/Mualem conductivity function have been proposed that avoid this artifact (Vogel and Cislerova, 1988; Vogel *et al.*, 1999, 2001).

In analogy to the WRC, piecewise combined functions are used to express the conductivity function with enhanced flexibility (e.g. Poulsen *et al.*, 2002; Schwartz and Evett, 2002; Bitterlich *et al.*, 2004). Recently, analytical expressions for retention and conductivity functions have been proposed which are not fully empirical, but depend on hydrodynamic considerations about the topology of the pore space of porous media. Notable are the functions of Assouline *et al.* (1998), Assouline (2001), Tuller and Or (Tuller *et al.*, 1999; Or and Tuller, 1999; Tuller and Or, 2001), and Zhu and Mohanty (2003).

HYSTERESIS

The dependencies of θ and K on the pressure head h are hysteretic. This implies that the relationships, as shown in Figure 2, are valid only for a certain history of wetting and drying, and for one particular saturation or desaturation path. Most measurement experiments therefore rely on a uniquely defined initial hydraulic state, from which the system is changed continually in one direction. Drainage experiments in the laboratory start usually from a maximum achievable saturation. Rewetting the soil from complete dryness will yield different relationships. Further, drying and wetting cycles from intermediate states between full saturation and oven dryness will again give different relationships. Hysteresis of the water retention characteristic is schematized as shown in Figure 3. Drainage of a full

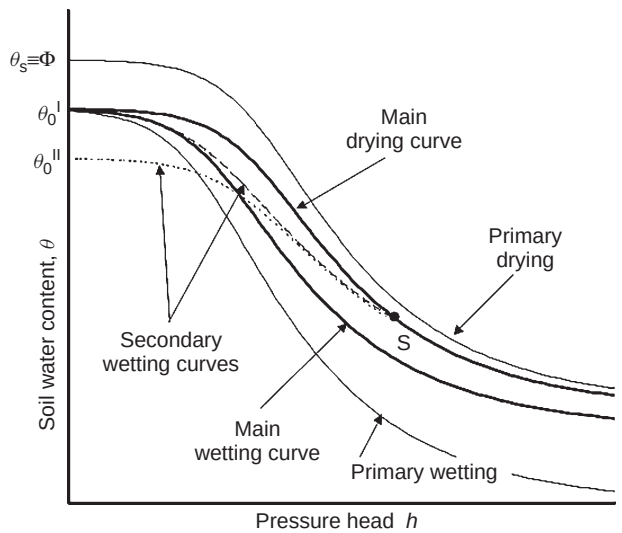


Figure 3 Schematization of hysteresis of the water retention characteristic

saturated medium yields the *primary drying curve*, imbibition from oven dryness yields the *primary wetting curve*, drainage from satiation yields the *main drying curve*, imbibition from the residual water content yields the *main wetting curve*, and any changes between imbibition and drainage at intermediate pressure values yield *secondary scanning curves*.

Hysteresis has been recognized as an important component of soil water redistribution since the early work of Haines (1930). The maximum water saturation of soils in the unsaturated zone under natural conditions is always significantly lower than the porosity, due to air entrapment, with typical saturation values of 80 to 90% of pore space (Klute, 1986; Dane and Hopmans, 2002). Full saturation of the pore space can be obtained only by specific experimental manipulations (slow wetting from the bottom, flushing with carbon dioxide, flushing with de-aired water, and application of vacuum). To distinguish the "saturated water content" from the typical maximum saturation in the field, Hillel (1980) proposed the term "satiation" for the latter.

The hysteresis of $\theta(h)$ leads to hysteresis of $K(h)$, since conductivity is controlled by the water-filled pore space. The hysteresis of the $K(\theta)$ relation of soils is generally regarded as negligible (Mualem, 1986; Kool and Parker, 1987). Ignoring hysteresis in the K-θ-h relations may be partly based on an assumption that its effects are negligible, partly on the absence of a well-validated hysteretic model that is easy to calibrate (Lenhard *et al.*, 1991), and partly on the difficulties of performing measurement experiments that are suitable to identify parameters of hysteresis models. Neglecting hysteresis leads to significant errors in the water redistribution under transient boundary conditions (Dane and Wierenga, 1975; Gillham *et al.*, 1979, Hoa *et al.*, 1977; Viaene *et al.*, 1994; Lehmann *et al.*, 1999; Si and Kachanoski, 2000). Russo *et al.* (1989) and Mitchell and Mayer (1998) investigated the influence of hysteresis on solute transport and found that the magnitude of the deviations between hysteretic and nonhysteretic simulations was not a simple function of single parameters, but rather depended on the combined values of many or all of the hydraulic parameters. Hysteresis is also seen as one of the major obstacles for comparing different measurement methods for the $K(\theta)$ (Stolte *et al.*, 1994; Basile *et al.*, 2003).

Hysteresis has been incorporated into simulations of unsaturated flow since Rubin (1965). Jaynes (1992) showed that soil water redistribution is retarded by hysteresis, and its magnitude depends on the rate at which the soil-water content varies. This supports the consensus that highly transient soil water conditions enhance hysteresis effects (Topp *et al.*, 1967). Hysteresis is typically assumed to be caused by the inkbottle effect, but other processes such as contact angle hysteresis, shrink-swell effects, or entrapped air play a significant role (Hillel, 1980). Contact angle hysteresis is affected by numerous factors, like surface roughness, chemical heterogeneity, fluid dynamics, particle shape, and gas adsorption, and all these relations depend on the composition of the soil solution (Dury *et al.*, 1998; Henry *et al.*, 2001). Therefore, hysteresis effects on soil-water redistribution may result from a single process, or interactions between different processes, and *a treatise on the role of individual factors controlling hysteresis is rather speculative* (Kutilek and Nielsen, 1994, p. 73).

Considerable effort has been put into the analysis and description of hysteretic soil hydraulic properties. This has led to numerous models for describing hysteresis in $\theta(h)$. Both, empirical models (Scott *et al.*, 1983; Hogarth *et al.*, 1988) and physically based models (Poulovassilis and Childs, 1971; Parlange, 1976; Mualem, 1974, 1984; Poulovassilis and Kargas, 2000) have been proposed. Kool and Parker (1987) combined the Scott hysteresis model with the retention function of van Genuchten (1980), which leads to an easily usable and consistent set of constitutive relationships (Luckner *et al.*, 1989). These models provide a simple means for determining scanning curves from a limited amount of data, such as the main wetting and drying hysteresis curves. The models of Parlange (1976) and Mualem (1984) need only one branch of the loop to predict all scanning curves. However, when comparing different models of hysteresis using 10 measured scanning curves, Viaene *et al.* (1994) concluded that the best models were the conceptual models needing two branches for calibration. Jaynes (1992) showed with simulation studies that none of the models were consistently better than the others. Other investigations found the empirical model of Kool and Parker (1987) to be only approximately valid (Schultze *et al.*, 1999; Basile *et al.*, 2003). Incorporating a physically based hysteresis model in numerical codes is demanding, since bookkeeping of the wetting and drying history at any point in the soil must be provided.

DYNAMIC EFFECTS

Hydraulic properties of porous media are commonly assumed to be of "static" nature, that is, to depend only on the size distribution and geometrical arrangement of pores in the porous medium and on the wetting/drying history (hysteresis), provided the solid matrix is rigid, fluid properties are constant, and external conditions do not change with time. This implies that the relationship between water content and pressure head during monotonous draining or imbibition processes is not affected by the dynamics of the water flow. As a consequence, $\theta(h)$ can be expressed by a unique characteristic, which can be determined by any available measuring technique (Durner and Lipsius, 2005; *see* **Chapter 75, Determining Soil Hydraulic Properties, Volume 2**). However, investigations by various authors

indicate that hydraulic properties may depend not only on the wetting and drying history, but also on the dynamics of water flow. This implies that hydraulic functions, which are measured under static equilibrium conditions, are not applicable to simulate water flow under transient conditions.

Davidson *et al.* (1966) found in imbibition and drainage studies that equilibrium water contents were affected by the dynamics of the wetting history. More water was removed from samples by one single large pressure decrease than by a sequence of small decreases. In contrast, more water was sorbed by the soils when a series of small pressure steps was applied in the imbibition process. Topp *et al.* (1967) compared water retention characteristics of vertical sand columns under static equilibrium conditions, under steady state flow, and under transient-flow conditions. They found that for transient experiments, at a given potential, water contents at drainage were significantly higher than water contents determined under static equilibrium or steady state conditions. Smiles *et al.* (1971) performed experiments where a series of imbibition/drainage cycles were applied to horizontal sand columns by imposing stepwise changes in the water pressure at one column end. They found that the retention characteristic depended on the size of the imposed pressure. Vachaud *et al.* (1972) confirmed these results for vertical soil columns. Stauffer (1977) investigated drainage in vertical columns packed with sand. Under dynamic flow conditions, he measured higher water contents at a given pressure as compared to the static $\theta(h)$ relationship. Similar to Smiles *et al.* (1971), gas phase pressure was additionally measured and no deviation from atmospheric air pressure was observed. Plagge (1991) and Plagge *et al.* (1999) confirmed these findings with evaporation experiments on silty soils. Tensiometric pressures and water contents were measured with microtensiometers and TDR (time domain reflector) probes, which were installed at multiple depth levels in soil columns. Dependent on the distance from the boundary of the soil column, the locally measured retention curves differed considerably. The differences were systematic and reproducible, and could not be attributed to faults in packing technique or to measurement errors. Lennartz (1992) investigated dynamic nonequilibrium systematically by performing evaporation experiments on repacked soil samples of four different substrates. He found significant differences between static and dynamic $\theta(h)$ characteristics, but could not identify a unique and simple tendency.

Wildenschild *et al.* (2001) investigated the flow-rate dependence of unsaturated hydraulic properties for disturbed soils in short laboratory columns by performing one-step and multistep outflow experiments. Retention characteristics were obtained from measured tensiometric pressures in the soil column and the average water content. For a sandy soil, they found pressure head differences of 10 to 15 cm (for a given saturation) and water content differences of up to 7% (for a given potential) between the slowest and the fastest outflow experiments. At a given pressure head, more water was retained with greater applied pressure steps. Conversely, Constantz (1993) reported higher water contents in slow multistep drainage experiments as compared to fast one-step drainage experiments. He attributes this to variations in pore water salt concentrations that induce differences in the pore water surface tension. Simunek *et al.* (2001) performed upward infiltration experiments and observed ongoing infiltration despite apparent equilibrium in the pressure heads. Also, during a 5-day redistribution phase where no flux across the boundaries occurred, they found an increase in tension in the whole soil sample by $\Delta h = 50\,\text{cm}$. All the above-mentioned experiments indicate that the dynamic effect depends on the size of the pressure changes, being larger for big changes. Further, there is some indication that dynamic effects are larger for soils with a wide pore-size distribution. Without relating the observed phenomena to specific processes, we may categorize them in a general way as "dynamic nonequilibrium" (Schultze *et al.*, 1999). According to Klute (1986, p. 660), *the reasons for and implication of this observation continue to be a subject of investigation*.

On the basis of a theoretical framework for multiphase flow in porous media, Hassanizadeh and Gray (1993) and Hassanizadeh *et al.* (2002) postulated the existence of a dynamic component in unsaturated water flow. Their analysis yields an approximate capillary pressure equation with a dynamic term, which depends linearly on the rate of change in water saturation

$$p_a - p_w = p_c - \frac{1}{L}\frac{\partial S}{\partial t} \qquad (12)$$

where p_w [$\mathrm{M\,L^{-1}\,T^{-2}}$] is the macroscopic pressure in the water phase, p_a [$\mathrm{M\,L^{-1}\,T^{-2}}$] is the pressure in the air phase, p_c [$\mathrm{M\,L^{-1}\,T^{-2}}$] is the capillary pressure, L [$\mathrm{T\,L\,M^{-1}}$] is a nonnegative material coefficient, $S = \theta/\theta_s$ [$-$] expresses the soil water saturation, and t [T] indicates the time. Equation 12 states that the pressure difference between air and water pressure is larger than the (equilibrium) capillary pressure under drainage conditions, and smaller when imbibition occurs.

Macroscopic simulation of water transport including dynamic effects is still in its infancy. Stauffer (1977) simulated the dynamic process with Richard's equation by using a dynamic retention characteristic, where the difference to the static characteristic depended on the rate of change of the local water content with time. Schultze *et al.* (1999) showed that a two-phase model explains much of the observed dynamic flow behavior in a soil column. Simunek *et al.* (2001) simulated water transport with a dual-continuum model (Gerke and van Genuchten, 1993), and found good agreement with measurements, when

interpreting measured tensiometric measurements as being representative for the interaggregate pore space.

Summarizing, evaluations of transient-flow experiments, thermodynamic reasoning, pore-network modeling (Held and Celia, 2001), and continuum percolation theory (Hunt, 2004) all lead to the conclusion that hydraulic relations are not only dependent on the wetting/drying history and the actual system state, but also on the rate of change of the system state. This is a problem, because hydraulic functions are primarily used in model applications under transient conditions, whereas most measurement experiments are based on equilibrium conditions (Durner and Lipsius, 2005; *see* **Chapter 75, Determining Soil Hydraulic Properties, Volume 2**). However, it is currently not clear as to what degree dynamic effects will affect water flow on time and spatial scales that are relevant for most applications. Despite its early notion, the problem of dynamic nonequilibrium during water flow in soils has not been adequately treated yet. Reasons for this are the high requirements with respect to the measurement, and even more the historic inability to tackle the theoretical problem. Without the availability of fast computers and fast measurement techniques, quantitative comparisons of observed transient-flow process variables with simulation results obtained by models of increasing complexity, such as Richards' equation, the coupled two-phase flow model, and the pore-network models, were not possible. Since dynamic nonequilibrium in water flow is closely related to hysteresis, a satisfying solution of the hysteresis problem also depends on the progress of this issue. In part, the distinction between hysteresis and dynamic nonequilibrium in water flow can be seen as a matter of the time scale, since a true equilibrium distribution of the water phase (which actually may require enormous time) will cause much of the apparently "static" hysteresis to disappear.

VARIABILITY

Soils are heterogeneous from the pore to the geologic scale. We have already noted that homogeneity in soils does not exist unless we refer to the concept of the REV. If the scale of observation is less than the REV, the soil is heterogeneous (Figure 1). Spatial variability of soils critically affects the effective capacity and transmission properties of water, solutes and gases, and controls by the formation of micro-niches and gradients of intensive soil properties during flow processes and the ecological functioning of the vadose zone. Reviews by Warrick and Nielsen (1980), Peck (1983) and Jury (1985) have shown that in soils, water flow and transport properties are the most variable. Jury *et al.* (1991), Kutilek and Nielsen (1994), Warrick (1998), Mulla and McBratney (2002), and van Es (2002) have reviewed techniques to handle spatial variability and soil heterogeneity.

Solute transport in soils is critically dependent on the variability of soil water fluxes. Increased travel time variance results in increased probability of short travel times, that is, preferential flow and transport. Soil water flux variability depends on the soil moisture status (Roth, 1995), and can be enhanced as compared to the variability of the soil moisture or hydraulic conductivity parameters. Owing to the nonlinearity of unsaturated water transport, spatial variability leads to scale dependence of hydraulic properties. Investigation of these questions is amongst the most critical topics in contemporary soil physical and hydrologic research (Sposito, 1998; Roth *et al.*, 1999). Scale issues are treated in this Encyclopedia in **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2** (Hopmans and Schoups, 2005, this issue).

Basically, there are two ways to tackle properties of heterogeneous soils. One way lies in the investigation of frequency distributions of local properties and in the characterization of their spatial structure. The other way is to search for effective transport processes and parameters, allowing for the application of overall initial and boundary conditions for a quasi-homogeneous system. Unfortunately, there is no general upscaling rule for hydraulic properties of unsaturated soils, and even the question about the existence of appropriate process descriptions with effective hydraulic properties at larger scales is not resolved yet. This is of fundamental importance when comparing the results of laboratory and field experiments.

Spatial and Temporal Variability

Since measurement scales are often smaller than the REV scale, soil hydraulic properties are often extremely variable in space and do not conform to conventional statistical assumptions. The structure of variability may be random, correlated, periodic, or in any combination, and may also be scale dependent (van Es, 2002). Spatial variability of hydraulic properties has been documented in many articles (e.g. Nielsen *et al.*, 1973; Simmons *et al.*, 1979, Russo and Bouton, 1992; Istok *et al.*, 1994; Jarvis and Messing, 1995; Mohanty *et al.*, 1994; Shouse *et al.*, 1995; Russo *et al.*, 1997; Shouse and Mohanty, 1998). Water contents in the field are mostly found to be normally distributed with medium variation (coefficient of variation, $0.15 < cv < 0.5$; Jury *et al.*, 1991; Warrick, 1998), with increasing variability when the soils dry out (Shouse *et al.*, 1995). Field investigations of hydraulic conductivities yield in most cases lognormal distributions, with variances of $\log(K)$ in the order of 1 (Dagan, 1984; Butters *et al.*, 1989; Heuvelman and McInnes, 1997; Holt *et al.*, 2002).

Spatial variability is overlaid by temporal variability. The soil pore system is affected by seasonal changes, by diurnal cycles of temperature and corresponding energy gradients, by plant and root growth, by irrigation cycles, and by freezing and thawing. Water transport in soil is

generally driven by weather-related processes or irrigation, which have high short-term dynamics. Soil properties that affect hydraulic behavior, aeration, erosion and runoff, and aggregation are strongly affected by processes such as soil wetting and drying, freezing and thawing, and weather-related behavior of soil organisms (van Es, 2002).

Cultural influences cause spatial and temporal extrinsic variability. Drainage, tillage, vehicle traffic, overburden pressure, plant cover, and soil amendments may dramatically alter mean behavior and variability patterns. For example, tillage tends to homogenize soils spatially, but may cause greater temporal variability from loosening and subsequent settling and recompaction (Mapa *et al.*, 1986; Starr, 1990, van Es *et al.*, 1999).

Geostatistical Characterization of Random Variability

Variations of soil properties are not completely disordered over the field and contain systematic and random components. To describe spatial correlation structures, tools of geostatistics, based on the theory of regionalized variables, have been developed (Matheron, 1963; Journel and Huijbregts, 1978) and successfully applied in soil science (e.g. Warrick and Nielsen, 1980). On the basis of proper statistical information, stochastic random fields can be created and the effective behavior of these "virtual realities" can be investigated.

In stochastic models, it is often assumed that the correlation lengths of different unsaturated parameters are the same (Holt *et al.*, 2002). Horizontal correlation scales for hydraulic properties were reported to be in the order of 6 m to 20–30 m (Ciollaro and Romano, 1995; Shouse *et al.*, 1995), whereas vertical correlation scales are in the order of 0.1 to 0.3 m (Robin *et al.*, 1991; Wierenga *et al.*, 1991). Heuvelman and McInnes (1997) investigated spatial variability of water fluxes on a scale of centimeters and found water flux densities to be normally distributed in a sandy topsoil, becoming lognormally distributed in a loamy subsoil. Jury (1985) and Shouse *et al.* (1995) noted that the range of spatial influence may be related to the scale of the sampled area, and that the commonly observed high nugget variance of soil hydraulic properties may be more closely related to measurement error than to high frequency, small-scale variations.

A particular problem of vadose zone spatial variability results from the fact that the typical travel distance is small compared with the distance required for a solute plume to reach asymptotic, Fickian behavior (Jury and Flühler, 1992). This means that hydraulic properties on the scale of interest are not randomly distributed. Hence, application of geostatistical concepts to derive asymptotic large-scale behavior is of limited value. This is further aggravated by the fact that proper three-dimensional geostatistical characterization of hydraulic properties needs such a high number of measurements that it cannot be obtained in practice. It appears that a reliable prediction of solute transport in natural fields requires the quasi-deterministic characterization of the largest spatial system structures, which are often on the same scale as the scale of interest (Vogel and Roth, 2003). This cannot be obtained by point measurements. Therefore, the further development of hydrogeophysical methods, possibly coupled with remote sensing technologies, and the linking of these data with functional hydraulic properties, for example, by stochastic data fusion (Yeh and Simunek, 2001), will play a crucial role in future developments.

Influence of Measurement Errors on Spatial Data Analysis

The analysis of spatial variability depends on measurements that are assumed to have little or no experimental or calibration error. Unfortunately, this is rarely the case. Many estimated hydraulic properties are likely to contain systematic error, or bias, in particular, if they are not measured directly. Although most studies carefully document instrumental procedures, the magnitude of errors in hydraulic property estimates and their impact on spatial statistics determined from field data are rarely evaluated. Ciollaro and Romano (1995), for example, used inverse modeling of evaporation experiments to derive the hydraulic properties of a 135-m transect, and found a discrepancy between optimized saturated conductivity and directly measured saturated conductivity by a factor of 10. Shouse *et al.* (1995) conclude that more information is needed on experimental error, its causes, and remedies.

If nonlinear inversion models, whether analytical or numerical, are used to infer property values from observed system responses and boundary conditions, random error in the observations (observation error) can lead to spatially correlated, systematic error, or bias, in the derived property value. Spatial bias may also result when the inversion model (e.g. governing equations, boundary conditions, initial conditions, constitutive models, etc.) is inadequate (inversion model error) (Kempthorne and Allmaras, 1986). Holt *et al.* (2002) used a Monte Carlo approach to explore the potential impact of observation and inversion model errors on the spatial statistics of field-estimated unsaturated hydraulic properties. For this analysis they simulated tension infiltrometer measurements in a series of idealized realities, each consisting of spatially correlated random property fields. They showed that estimated hydraulic properties are strongly biased even by small, simple observations, and inversion model errors. Error in spatial statistics was more than an order-of-magnitude, and artificial cross-correlation between measured properties occurred. Unfortunately, there are no unique indicators of bias, as property values may appear reasonable and spatial statistics may look realistic.

Errors in the spatial statistics of hydraulic properties cause critical stochastic model assumptions to be violated, limiting the usable parameter space for model predictions. Even where critical assumptions are valid, stochastic model predictions show significant error, and the magnitude and pattern of error changes with the true property means, the flow conditions, and the type of measurement error. Experimental technologies need to be developed that reduce experimental error and increase precision.

Scaling of Hydraulic Functions

A particular challenge for handling spatially variable unsaturated hydraulic properties lies in the fact that not just distributions of single parameters, like K_s, must be characterized, but the variability of the whole constitutive relationships, $K(\theta)$ and $\theta(h)$. A way to handle this is using the concept of similar media scaling. The objective of scaling is to coalesce a set of hydraulic relationships into a single reference curve using scaling factors that describe the set as a whole. This method was introduced by Miller and Miller (1956), and its concepts and limitations were reiterated and assessed later by Miller (1980), Tillotson and Nielsen (1984), and Sposito and Jury (1985). Simmons *et al.* (1979) extended similar media scaling theory to characterize the spatial variability of field water retention measurements. Jury (1985) indicated that an obvious limitation of scaling theories is that the errors involved in measuring the properties used to calculate the scale factors are carried along as a part of the scale factor sample variance. Reviews and application examples of scaling soil hydraulic functions are given by Hillel and Elrick (1990), Kutilek and Nielsen (1994), and Roth (1995).

Mean and Effective Soil Hydraulic Properties

Since the scale of interest is in general larger than a local REV, upscaling of hydraulic properties is required to get a proper description on the scale of interest. When nonlinear properties have to be averaged, the mean property is in the general case not identical to the "effective" property. An effective property represents a homogeneous medium with a macroscopic behavior that is quasi-identical to the heterogeneous system. For the unsaturated case, the differences between mean and effective properties are dependent on the spatial variability of the local properties, but furthermore also on the boundary conditions, as will be illustrated for the case of steady state water flow into a moderately heterogeneous soil.

Figure 4 shows a simulated domain of a weakly heterogeneous soil, composed of irregular lenses of sand and silt forming a 2-m deep soil body (Flühler and Roth, 2004). The structures have a wide extension in the horizontal and a small extension in the vertical scale. Water movement takes place from top to bottom. The water flux q through the upper boundary is continuous and constant. At the lower boundary there is contact to the groundwater table, that is, $h = 0$. The depicted structure is assumed to be periodic, that is, water flowing out through a vertical boundary is assumed to flow in at the opposite boundary. The hydraulic properties of the sand are given by the van Genuchten/Mualem coefficients $\alpha = 0.02\,\mathrm{cm}^{-1}$, $n = 4.0$, $\theta_s = 0.3$, $\theta_r = 0$, $K_s = 10^{-4}\,\mathrm{m\,s}^{-1}$, and those of the silt by $\alpha = 0.005\,\mathrm{cm}^{-1}$, $n = 1.33$, $\theta_s = 0.4$, $\theta_r = 0$, $K_s = 10^{-5}\,\mathrm{m\,s}^{-1}$. The situation shown in Figure 4 is the stationary flow field corresponding to a weak rain, with a flux density at the upper boundary of $1.8 \times 10^{-7}\,\mathrm{m\,s}^{-1} = 1.56\,\mathrm{cm\,d}^{-1}$. At this low infiltration rate, the silt in the upper 50-cm region, far from the groundwater table, has a much higher conductivity than the sand, and water flows convergent through the silt lenses. In this stationary situation,

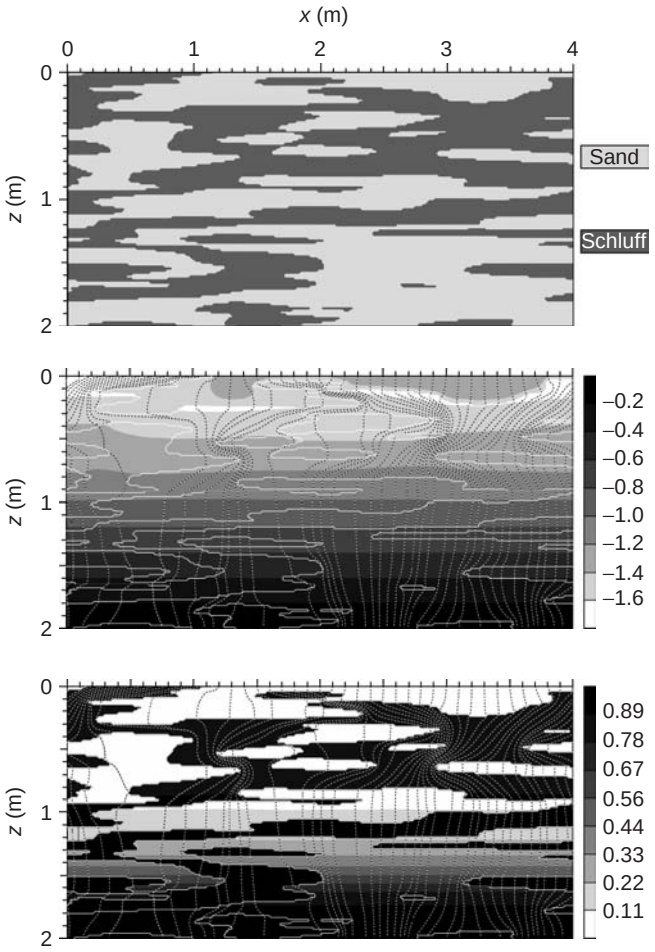


Figure 4 Hydraulic system properties in a soil with heterogeneous layers at a low constant infiltration. (a) Material distribution. (b) Pressure head distribution at a very low infiltration, $q = 1.56\,\mathrm{cm\,d}^{-1}$. (c) water saturation. The border sand-silt is indicated by the white drawn lines, the streamlines are shown by dotted white lines. (Courtesy of H. Flühler and K. Roth)

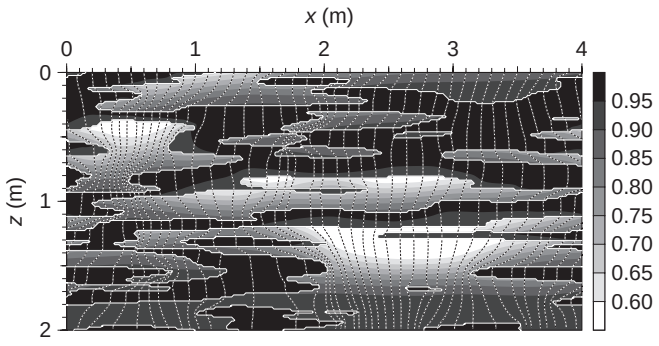


Figure 5 Water saturation during steady state infiltration with $q = 5.4 \times 10^{-5}\,m\,s^{-1} = 47\,cm\,d^{-1}$ into the heterogeneous soil of Figure 4 (Courtesy of H. Flühler and K. Roth)

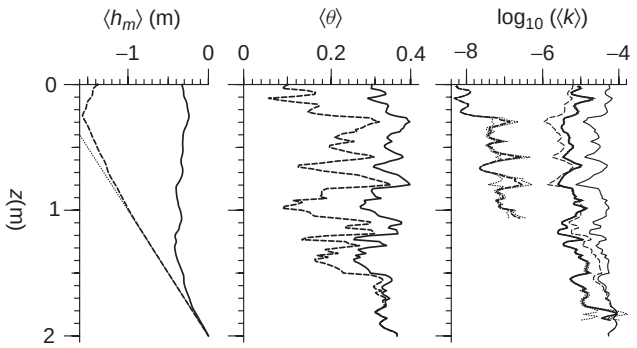


Figure 6 Mean and effective hydraulic properties of the heterogeneous soil of Figure 5 for a flux density of $1.8 \times 10^{-7}\,m\,s^{-1}$ (dashed lines) and $5.4 \times 10^{-6}\,m\,s^{-1}$ (drawn lines). (a) mean hydraulic head $\langle h_m \rangle$ [m]; the dotted line represents the hydraulic head at hydrostatic equilibrium. (b) Mean volumetric water content $\langle \theta_w \rangle$. (c) Mean hydraulic conductivity, $< k >$ (thin lines) and effective conductivity k_{eff} (thick lines), k in $[m\,s^{-1}]$. The thin dotted lines show uncertainty bands of one standard deviation of the estimation of k_{eff}, Values with an estimation uncertainty which is greater than the value itself are omitted (Courtesy of H. Flühler and K. Roth)

the water saturation reflects the substrate changes very sharply, whereas the lateral distribution of the pressure head is much smoother, in particular, close to the groundwater table. As a fundamental consequence, single point measurements of matric potential are more representative for the average pressure than point measurements of water contents for the average water content. At about 1.2-m depth, the two materials have essentially the same conductivity, leading to a uniform flow of water. Close to the groundwater table, the silt lenses are obstacles to water flow, and water flows preferentially in the sand matrix. Figure 5 shows the water saturation in the same profile, corresponding to a stationary flow with an increased infiltration rate of $q = 5.4 \times 10^{-6}\,m\,s^{-1}$. Now, the silt lenses are obstacles to water flow in the whole profile.

Figure 6 shows the corresponding depth profiles of the mean hydraulic properties $\langle h \rangle$, $\langle \theta \rangle$, and $\log_{10}\langle K \rangle$, obtained by horizontally averaging the local values of the heterogeneous soil. The dashed lines indicate the situation for the low water flux density, the drawn lines for the high flux density. It is apparent that the average water content, $\langle \theta \rangle$, varies much more than the average water potential, $\langle h \rangle$, which reflects a typical gravitational flow situation for the high flux rate (Figure 6, left, drawn line) and a situation close to hydrostatic equilibrium conditions for the low flow rate (Figure 6, left, dashed line). Figure 6 (right) depicts the mean (thin lines) and effective hydraulic conductivity (thick lines) for the corresponding one-dimensional water flux. We observe that even for the case of quasi-gravitational unit-gradient flow, the mean hydraulic conductivity, $\log_{10}\langle K \rangle$, in the upper part of the profile is higher than the infiltration rate, since the local water flux is strongly deflected from the straight vertical movement.

The *effective hydraulic conductivity* for the flow situation is obtained by dividing the macroscopic water flux density by the gradient of the mean water potential. For the high flux density, the resulting effective hydraulic conductivity is in the range of the infiltration rate, which is about 5 times smaller than the mean hydraulic conductivity. For the low infiltration rate, the difference between mean and effective conductivity is in the order of two magnitudes.

CONCLUDING REMARKS

Enormous advances have been made during the last few decades regarding our understanding and our ability to model flow and transport processes in the vadose zone (Simunek, 2005; *see* **Chapter 78, Models of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2**; van Genuchten and Simunek, 2004). The current scientific perspective is that the Richards' equation provides the most appropriate tool on which any description of water transport in the vadose zone can be based. Other simplified descriptions, such as bucket models, are not considered as a viable alternative (Vanclooster *et al.*, 2004). More comprehensive descriptions, such as the use of two-phase flow models (for water and air) appear superfluous for standard uses. The validity of the Richards' equation for larger scales is a matter of ongoing debate. From a theoretical standpoint and based on stochastic hydrology a Richards'-type equation for water flow appears to retain its validity for larger scales, yet with different constitutive relationships, which are dependent both on the stochastic distribution of local soil hydraulic properties and also on the boundary conditions (Harter and Hopmans, 2004).

For the characterization of soil hydraulic properties the most critical problem lies in the derivation of effective properties for large-scale applications, such as for hill slopes,

catchments, or entire regions. Nevertheless, the description of hydraulic functions on the measurement scale, needs improvement as well, both in the moisture range near saturation (for solute transport and infiltration), and also toward dryness. The latter is of particular importance for simulating evaporation and transpiration under conditions where the conductivity of the soil becomes a limiting factor. Furthermore, a valid description of hysteresis in the intermediate moisture range is required for transient-flow conditions. The development of accurate, fast and reliable measurement methods, a further assessment of "dynamic effects", and the description of the hydraulic properties for shrinking and swelling soils (Smiles and Raats, 2005; *see* **Chapter 67, Hydrology of Swelling Clay Soils, Volume 2**) remain major challenges to soil scientists and hydrologists. Some important issues with respect to the future use of Richards' equation are (i) a suitable integration of boundary conditions on large scales (integration in landscape models), (ii) the adequate representation of soil hydraulic functions on large scales, (iii) hysteresis, (iv) spatial and temporal variability, (v) structured soils and macropores, and (vi) unstable wetting fronts as well as factors dealing with soil hydrophobicity.

REFERENCES

Ahuja L.R. and Swarzendruber D. (1972) An improved form of the soil-water diffusivity function. *Soil Science Society of America Proceedings*, **36**, 9–14.

Assouline S. (2001) A model for soil relative hydraulic conductivity based on the water retention characteristic curve. *Water Resources Research*, **37**, 265–271.

Assouline S., Tessier D. and Bruand A. (1998) A conceptual model of the soil water retention curve. *Water Resources Research*, **34**, 223–231.

Averjanov S.F. (1950) About permeability of subsurface soils in case of incomplete saturation. *English Collection*, **7**, 19–21.

Basile A., Ciollaro G. and Coppola A. (2003) Hysteresis in soil water characteristics as a key to interpreting comparisons of laboratory and field measured hydraulic properties. *Water Resources Research*, **39**, 1355, doi:10.1029/2003WR002432.

Bear J. (1972) *Dynamics of Fluids in Porous Media*, American Elsevier: New York.

Bitterlich S., Durner W., Iden S.C. and Knabner P. (2004) Inverse estimation of the unsaturated soil hydraulic properties from column outflow experiments using free-form parameterizations. *Vadose Zone Journal*, **3**, 971–981.

Brooks R.H. and Corey A.T. (1964) *Hydraulic Properties of Porous Media*, Hydrology Paper 3, Colorado State University: Fort Collins, p. 27.

Brutsaert W. (1967) Some methods for calculating unsaturated permeability. *Transactions of the American Society of Agricultural Engineers*, **10**, 400–404.

Buchan G.D. and Grewal K.S. (1990) The power-function model for the soil moisture characteristic. *Journal of Soil Science*, **41**(1), 111–117.

Buchter B., Hinz C. and Flühler H. (1994) Sample size for determination of coarse fragment content in a stony soil. *Geoderma*, **63**, 265–275.

Buckingham E. (1907) Studies on the movement of soil moisture. *US Department of Agriculture Bureau of Soils Bulletin* No. 38, USDA, Washington, DC.

Burdine N.T. (1953) Relative permeability calculations from pore-size distribution data. *Transactions of The American Institute of Mining, Metallurgical and Petroleum Engineers*, **198**, 71–78.

Butters G.L., Jury W.A. and Ernst F.F. (1989) Field scale transport of bromide in an unsaturated soil, 1, Experimental methodology and results. *Water Resources Research*, **25**, 1575–1581.

Campbell G.S. (1974) A simple method for determining unsaturated conductivity from moisture retention data. *Soil Science*, **117**, 311–314.

Carsel R.F. and Parrish R.S. (1988) Developing joint probability distributions of soil water retention characteristics. *Water Resources Research*, **24**, 755–769.

Childs E.C. and Collis-George G.N. (1950) The permeability of porous materials. *Proceedings of the Royal Society of London Series A*, **201**, 392–405.

Ciollaro G. and Romano N. (1995) Spatial variability of the soil hydraulic properties of a volcanic soil. *Geoderma*, **65**, 263–282.

Constantz J. (1993) Confirmation of rate-dependent behaviour in water retention during drainage in nonswelling porous materials. *Water Resources Research*, **29**, 1331–1334.

Cushman J.H. (1984) On unifying the concepts of scale, instrumentation, and stochastics in development of multiphase transport theory. *Water Resources Research*, **20**, 1668–1678.

Dagan G. (1984) Solute transport in heterogeneous porous formations. *Journal of Fluid Mechanics*, **145**, 151–177.

Dane J.H. and Hopmans J.W. (2002) Water retention and storage: laboratory. In *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 675–720.

Dane J.H. and Wierenga P.J. (1975) Effect of hysteresis o the prediction of infiltration, redistribution and drainage of water in a layered soil. *Journal of Hydrology*, **25**, 229–242.

Darcy H. (1856) *Les Fontains Publiques de la Ville de Dijon*. Dalmont: Paris.

Davidson J.M., Nielsen D.R. and Biggar J.W. (1966) The dependence of soil water uptake and release upon the applied pressure increment. *Soil Science Society of America Proceedings*, **30**, 298–303.

D'Hollander E.H. (1979) Estimation of the pore size distribution from the moisture characteristic. *Water Resources Research*, **15**(1), 107–112.

Durner W. (1992) Predicting the unsaturated hydraulic conductivity using multi-porosity water retention curves. In *Proceedings of the International Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, van Genuchten M.Th., Leij F.J. and Lund L.J. (Eds.), University of California: Riverside, pp. 185–201.

Durner W. (1994) Hydraulic conductivity estimation for soils with heterogeneous pore structure. *Water Resources Research*, **30**, 211–223.

Durner W. and Lipsius K. (2005) Determining hydraulic functions. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Durner W. and Or D. (2005) Soil water potential measurement. *Encyclopedia of Hydrological Sciences*, this issue. John Wiley & Sons.

Dury O., Fischer U. and Schulin R. (1998) Dependence of hydraulic and pneumatic characteristics of soils on a dissolved organic compound. *Journal of Contaminant Hydrology*, **33**, 39–57.

Erh K.T. (1972) Application of spline functions to soil science. *Soil Science*, **114**, 333–338.

Farrel D.A. and Larson W.E. (1972) Modeling the pore structure of porous media. *Water Resources Research*, **8**(3), 699–706.

Finsterle S. (1999) *iTOUGH2 User's Guide*, Report LBNL-40040, Lawrence Berkeley National Laboratory, Berkeley.

Flühler H. and Roth K. (2004) *Physik der ungesättigten Zone*, Lecture notes, Institute of Terrestrial Ecology, Swiss Federal Institute of Technology: Zurich, Institute of Environmental Physics, University of Heidelberg.

Gardner W.R. (1958) Some steady state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science*, **85**, 228–232.

Gerke H.H. and van Genuchten M.Th. (1993) A dual-porosity model for simulating the preferential movement of water and solutes in structured porous media. *Water Resources Research*, **29**, 305–319.

Gillham R.W., Klute A. and Heermann D.F. (1979) Measurement and numerical simulation of hysteretic flow in a heterogeneous porous medium. *Soil Science Society of America Journal*, **43**, 1061–1067.

Haines W. (1930) Studies in the physical properties of soil: V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science*, **20**, 97–116.

Harter T. and Hopmans J.W. (2004) Role of vadose-zone flow processes in regional-scale hydrology: review, opportunities and challenges. In *Unsaturated-Zone Modeling*, Wageningen UR Frontis Series, Feddes R.A., de Rooij G.H. and van Dam J.C. (Eds.), Kluwer Academic Publishers: Dordrecht, pp. 179–210.

Hassanizadeh S.M., Celia M.A. and Dahle H.K. (2002) Dynamic effect in the capillary pressure–saturation relationship and its impacts on unsaturated flow. *Vadose Zone Journal*, **1**, 38–57.

Hassanizadeh S.M. and Gray W.G. (1993) Thermodynamic basis of capillary pressure in porous media. *Water Resources Research*, **29**, 3389–3405.

Held R.J. and Celia M.A. (2001) Modeling support of functional relationships between capillary pressure, saturation, interfacial area and common lines. *Advances in Water Resources*, **24**, 325–343.

Henry E.J., Smith J.E. and Warrick A.W. (2001) Surfactant effects on unsaturated flow in porous media with hysteresis: horizontal column experiments and numerical modeling. *Journal of Hydrology*, **245**, 73–88.

Heuvelman W.J. and McInnes K.J. (1997) Spatial variability of water fluxes in soil: a field study. *Soil Science Society of America Journal*, **61**, 1037–1041.

Hillel D. (1980) *Fundamentals of Soil Physics*, Academic Press: New York, p. 413.

Hillel D. and Elrick D.E. (Eds.) (1990) *Scaling in Soil Physics. Principles and Applications*, SSSA Special Publication No. 25, Soil Science Society of America: Madison.

Hoa N.T., Gaudu R. and Thirriot C. (1977) Influence of the hysteresis effect on transient flows in saturated–unsaturated porous media. *Water Resources Research*, **13**(6), 992–996.

Hogarth W., Hopmans J., Parlange J.Y. and Haverkamp R. (1988) Application of a simple soil-water hysteresis model. *Journal of Hydrology*, **98**, 21–29.

Holt R.M., Wilson J.L. and Glass R.J. (2002) Spatial bias in field-estimated unsaturated hydraulic properties. *Water Resources Research*, **38**, 1311–1336.

Hopmans J.W. and Schoups G.H. (2005) Soil water flow at different scales. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Hubbert M.K. (1956) Darcy's law and the field equations of the flow of underground fluids. *AIME Petroleum Transaction*, **207**, 222–239.

Hunt A.G. (2004) Continuum percolation theory for pressure-saturation characteristics of fractal soils: extension to non-equilibrium. *Advances in Water Resources*, **27**, 245–257.

Irmay S. (1954) On the hydraulic conductivity of unsaturated soils. *EOS Transaction. AGU*, **35**, 463–467.

Istok J.D., Rautman C.A., Flint L.E. and Flint A.L. (1994) Spatial variability of hydrologic properties in a volcanic tuff. *Ground Water*, **32**, 751–760.

Jarvis N.J. and Messing I. (1995) Near saturated hydraulic conductivity in soils of contrasting texture measured by tension infiltrometers. *Soil Science Society of America Journal*, **59**, 27–34.

Jaynes D.B. (1992) Estimating hysteresis in the soil water retention function. In *Proceedings of the International Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, van Genuchten M.Th., Leij F.J. and Lund L.J. (Eds.), University of California: Riverside, pp. 219–232.

Journel A.G. and Huijbregts C.hJ. (1978) *Mining Geostatistics*, Academic Press: London.

Jury W.A. (1985) *Spatial Variability of Soil Physical Parameters in Solute Migration: A Critical Literature Review*, EPRI Topical Report E4228, Electric Power Research Institute, Palo Alto.

Jury W.A. and Flühler H. (1992) Transport of chemicals through soil: mechanisms, models, and field applications. *Advances in Agronomy*, **47**, 141–201.

Jury W.A., Gardner W.R. and Gardner W.H. (1991) *Soil Physics*, Wiley: New York, p. 328.

Kastanek F.J. and Nielsen D.R. (2001) Description of soil water characteristics using cubic spline interpolation. *Soil Science Society of America Journal*, **65**, 279–283.

Kempthorne O. and Allmaras R.R. (1986) Errors and variability of observations. In *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, *Second Edition*, Klute A. (Ed.), American Society of Agronomy: Madison, pp. 1–31.

King L.G. (1965) Description of soil characteristics for partially saturated flow. *Soil Science Society of America Proceedings*, **29**, 359–362.

Klute A. (1986) Water retention: laboratory methods. In *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, *SecondEdition*, Klute A. (Ed.), American Society of Agronomy: Madison, pp. 635–662.

Kool J.B. and Parker J.C. (1987) Development and evaluation of closed-form expressions for hysteretic soil hydraulic properties. *Water Resources Research*, **23**(1), 105–114.

Kosugi K. (1996) Lognormal distribution model for unsaturated soil hydraulic properties. *Water Resources Research*, **32**, 2697–2703.

Kovacs G. (1981) *Seepage Hydraulics*, Elsevier: Amsterdam.

Kutilek M. and Nielsen D.R. (1994) *Soil Hydrology*, Catena Verlag: Cremlingen–Destedt, p. 370.

Laliberte G.E. (1969) A mathematical function for describing capillary pressure–desaturation data. *Bulletin of the International Association of Scientific Hydrology*, **14**(2), 131–149.

Lambot S., Hupet F., Javaux M. and Vanclooster M. (2004) Laboratory evaluation of a hydrodynamic inverse modeling method based on water content data. *Water Resources Research*, **40**, W03506, doi:10.1029/2003WR002641.

Lehmann P., Stauffer F., Hinz C., Dury O. and Flühler H. (1999) Effect of hysteresis on water flow in a sand column with a fluctuating capillary fringe. *Journal of Contaminant Hydrology*, **33**, 81–100.

Lenhard R.J., Parker J.C. and Kaluarachchi J.J. (1991) Comparing simulated and experimental hysteretic two-phase transient fluid flow phenomena. *Water Resources Research*, **27**, 2113–2124.

Lennartz F. (1992) Einfluß von instationären Fließzuständen auf die Wassergehalt–Wasserpotentialbeziehung, Ph. D. Thesis, *Schriftenreihe des Instituts für Wasserwirtschaft und Landschaftsökologie der Chrisitan–Albrechts–Universität Kiel*, Kiel, p. 123.

Luckner L., van Genuchten M.Th. and Nielsen D.R. (1989) A consistent set of parametric models for the two-phase flow of immiscible fluids in the subsurface. *Water Resources Research*, **25**, 2187–2193.

Mapa R.B., Green R.E. and Santo L. (1986) Temporal variability of soil hydraulic properties with wetting and drying subsequent to tillage. *Soil Science Society of America Journal*, **50**, 1133–1138.

Matheron G. (1963) Principles of Geostatistics. *Economic Geology*, **58**, 1246–1266.

Miller E.E. (1980) Similitude and scaling of soil-water phenomena. In *Applications of Soil Physics*, Hillel D. (Ed.), Academic Press: New York, pp. 300–318.

Miller E.E. and Miller R.D. (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics*, **27**, 324–332.

Mitchell R.J. and Mayer A.S. (1998) The significance of hysteresis in modeling solute transport in unsaturated porous media. *Soil Science Society of America Journal*, **62**, 1506–1512.

Mohanty B.P., Ankeny M.D., Horton R. and Kanwar R.S. (1994) Spatial analysis of hydraulic conductivity measured using disc infiltrometers. *Water Resources Research*, **30**, 2489–2498.

Mohanty B.P., Bowman R.S., Hendrickx J.M.H. and van Genuchten M.Th. (1997) New piecewise-continuous hydraulic functions for modeling preferential flow in an intermittent-flood-irrigated field. *Water Resources Research*, **33**(9), 2049–2206.

Mualem Y. (1974) A conceptual model of hysteresis. *Water Resources Research*, **10**(3), 514–520.

Mualem Y. (1976) A new model of predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 513–522.

Mualem Y. (1984) A modified dependent-domain theory of hysteresis. *Soil Science*, **137**(5), 283–291.

Mualem Y. (1986) Hydraulic conductivity of unsaturated soils, predictions and formulas. In *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, *Second Edition*, Klute A. (Ed.), American Society of Agronomy: Madison, pp. 799–823.

Mulla D.J. and McBratney A.B. (2002) Soil spatial variability. In *Soil Physics Companion*, Warrick A.W. (Ed.), CRC Press: Boca Raton, Chap. 9.

Nemes A., Schaap M.G., Leij F.J. and Wösten J.H.M. (2001) Description of the unsaturated soil hydraulic database UNSODA version 2.0. *Journal of Hydrology*, **251**, 51–162.

Nielsen D.R., Biggar J.W. and Erh K.T. (1973) Spatial variability of field-measured soil-water properties. *Hilgardia*, **42**, 215–259.

Or D. and Tuller M. (1999) Liquid retention and interfacial area in variably saturated porous media: upscaling from single-pore to sample-scale model. *Water Resources Research*, **35**, 3591–3606.

Othmer H., Diekkrüger B. and Kutilek M. (1991) Bimodal porosity and unsaturated hydraulic conductivity. *Soil Science*, **152**, 139–150.

Parlange J.Y. (1976) Capillary hysteresis and the relationship between drying and wetting curves. *Water Resources Research*, **12**, 224–228.

Peck A.J. (1983) Field variability of soil physical processes. In *Advances in Irrigation*, Hillel D.I. (Ed.), Academic Press: New York, Vol. 2.

Plagge R. (1991) *Bestimmung Der Ungesättigten Hydraulischen Leitfähigkeit im Boden*, Bodenökologie und Bodengenese, Ph. D. Thesis, *Schriftenreihe des FG Bodenkunde und Regionale Bodenkunde am Institut für Ökologie*, TU Berlin, Heft 3, Berlin, p. 152.

Plagge R., Haeupl P. and Renger M. (1999) Transient effects on the hydraulic properties of porous media. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 905–912.

Poulovassilis A. and Childs E.E. (1971) The hysteresis of pore water: the non-independence of domains. *Soil Science*, **112**, 301–312.

Poulovassilis A. and Kargas G. (2000) A note on calculating hysteretic behaviour. *Soil Science Society of America Journal*, **64**, 1947–1950.

Poulsen T.G., Moldrup P., Iverson B.V. and Jacobsen O.H. (2002) Three-region campbell model for unsaturated hydraulic conductivity in undisturbed soils. *Soil Science Society of America Journal*, **66**, 744–752.

Prunty L. and Casey F.X.M. (2002) Soil water retention curve description using a flexible smooth function. *Vadose Zone Journal*, **1**, 179–185.

Richards L.A. (1931) Capillary conduction of liquids through porous media. *Physics*, **1**, 318–333.

Rijtema P.E. (1965) *An Analysis of Actual Evapotranspiration*, Agricultural Research Report, No 659, Pudoc, Wageningen.

Robin M.J.L., Sudicky E.A., Gillham R.W. and Kachanoski R.G. (1991) Spatial variability of strontium distribution coefficients and their correlations with hydraulic conductivity in the Canadian forces base borden aquifer. *Water Resources Research*, **27**, 2619–2632.

Rogowski A.S. (1972) Estimation of the soil moisture characteristic and hydraulic conductivity: comparison of models. *Soil Science*, **114**, 423–429.

Rossi C. and Nimmo J.R. (1994) Modeling of soil water retention from saturation to oven dryness. *Water Resources Research*, **30**, 701–708.

Roth K. (1995) Steady state flow in an unsaturated, two-dimensional macroscopically homogenous miller-similar medium. *Water Resources Research*, **31**, 2127–2140.

Roth K., Vogel H.-J. and Kasteel R. (1999) The scaleway: a conceptual framework for upscaling soil properties. In *Modelling of Transport Processes in Soils at Various Scales in Time and Space*, Feyen J. and Wiyo K. (Eds.), *International Workshop of EurAgEng's Field of Interest on Soil and Water, Leuven*, Wageningen Pers: Wageningen, pp. 477–490, 24–26 November.

Rubin J. (1965) Numerical method for analyzing hysteresis-affected, post-infiltration redistribution of soil moisture. *Soil Science Society of America Proceedings*, **31**, 13–27.

Russo D. (1988) Determining soil hydraulic properties by parameter estimation: on the selection of a model for the hydraulic properties. *Water Resources Research*, **24**, 453–459.

Russo D. and Bouton M. (1992) Statistical analysis of spatial variability in unsaturated flow parameters. *Water Resources Research*, **28**, 1911–1925.

Russo D., Jury W.A. and Butters G.L. (1989) Numerical analysis of solute transport during transient irrigation: 1. The effect of hysteresis and profile heterogeneity. *Water Resources Research*, **25**(10), 2109–2118.

Russo D., Russo I. and Laufer A. (1997) On the spatial variability of parameters of the unsaturated hydraulic conductivity. *Water Resources Research*, **33**, 947–956.

Schaap M.G. (2005) Models for indirect estimation of soil hydraulic properties. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Schaap M.G., Leij F.J. and van Genuchten M.Th. (2001) Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, **251**, 163–176.

Schultze B., Ippisch O., Huwe B. and Durner W. (1999) Dynamic nonequilibrium in unsaturated water flow. In *Proceedings of the International Workshop on Characterization und Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 877–892.

Schwartz R.C. and Evett S.R. (2002) Estimating hydraulic properties of a fine-textured soil using a disc infiltrometer. *Soil Science Society of America Journal*, **66**, 1409–1423.

Scott P.S., Farquhar G.J. and Kouwen N. (1983) *Hysteretic Effects on Net Infiltration. Advances in Infiltration*, American Society of Agricultural Engineers Publication: St. Josephs, pp. 11–83, 163–170.

Shouse P.J. and Mohanty B.P. (1998) Scaling of near-saturated hydraulic conductivity measured using disc infiltrometers. *Water Resources Research*, **34**, 1195–1205.

Shouse P.J., Russell W.B., Burden D.S., Selim H.M., Sisson J.B. and Van Genuchten M.Th. (1995) Spatial variability of soil water retention functions in a silt loam soil. *Soil Science*, **159**, 1–12.

Si B.C. and Kachanoski R.G. (2000) Unified solution for infiltration and drainage with hysteresis: theory and field test. *Soil Science Society of America Journal*, **64**, 30–36.

Simmons C.S., Nielsen D.R. and Biggar J.W. (1979) Scaling of field–measured soil–water properties, I. methodology, II. hydraulic conductivity and flux. *Hilgardia*, **50**, 1–25.

Simunek J. (2005) Models of water flow and solute transport in the unsaturated zone. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Simunek J., Wendroth O., Wypler N. and van Genuchten M.Th. (2001) Nonequilibrium water flow characterized from an upward infiltration experiment. *European Journal of Soil Science*, **52**, 13–24.

Smiles D.E. and Raats (2005) Hydrology of heavy clay soils. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Smiles D.E., Vachaud G. and Vauclin M. (1971) A test of the uniqueness of the soil moisture characteristic during transient, nonhysteretic flow of water in rigid. *Soil Science Society of America Journal*, **35**, 534–539.

Sposito G. (1998) *Scale Dependence and Scale Invariance in Hydrology*, Cambridge University Press: New York, p. 423.

Sposito G. and Jury W.A. (1985) Inspectional analysis in the theory of water flow through unsaturated soil. *Soil Science Society of America Journal*, **49**, 791–798.

Starr J.L. (1990) Spatial and temporal variation of ponded infiltration. *Soil Science Society of America Journal*, **54**, 629–636.

Stauffer F. (1977) *Einfluß der Kapillaren Zone auf Instationäre Drainagevorgänge*, Ph. D. Thesis, ETH, Zürich.

Stolte J., Freijer J.I., Bouten W., Dirksen C., Halbertsma J.M., Van Dam J.C., Van den Berg J.A., Veerman G.J. and Wösten J.H.M. (1994) Comparison of six methods to determine unsaturated soil hydraulic conductivity. *Soil Science Society of America Journal*, **58**, 1596–1603.

Su C. and Brooks R.H. (1975) Soil hydraulic properties from infiltration tests. *Watershed Management Proceedings, Irrigation and Drainage Division*, American Society of Chemical Engineers: Logan Utah, pp. 516–542, August 11–13.

Taylor G.S. and Luthin J.N. (1969) Computer methods for transient analysis of water table aquifers. *Water Resources Research*, **5**(1), 144–152.

Tillotson P.M. and Nielsen D.R. (1984) Scale factors in soil science. *Soil Science Society of America Journal*, **48**, 953–959.

Topp G.C., Klute A. and Peters D.B. (1967) Comparison of water content–pressure head data obtained by equilibrium, steady state, and unsteady–state methods. *Soil Science Society of America Proceedings*, **31**, 312–314.

Tuller M. and Or D. (2001) Hydraulic conductivity of variably saturated porous media: film and corner flow in angular pore space. *Water Resources Research*, **31**(5), 1257–1276.

Tuller M., Or D. and Dudley L.M. (1999) Adsorption and capillary condensation in porous media -liquid retention and interfacial configurations in angular pores. *Water Resources Research*, **35**, 1949–1964.

Vachaud G., Vauclin M. and Wakil M. (1972) A study of the uniqueness of the soil moisture characteristic during desorption by vertical drainage. *Soil Science Society of America Proceedings*, **36**, 531–532.

Vanclooster M., Boesten J., Tiktak A., Jarvis N., Kroes J.G., Munoz-Carpena R., Clothier B.E. and Green S.R. (2004) On the use of unsaturated flow ad transport models in nutrient and pesticide management. In *Unsaturated-zone Modeling*, Wageningen UR Frontis Series, Feddes R.A., de Rooij G.H. and van Dam J.C. (Eds.), Kluwer Academic Publishers: Dordrecht, pp. 331–362.

van Es H.M. (2002) Soil variability. In *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: ISBN 0-89118-810-X, pp. 1–11.

van Es H.M., Ogden C.B., Hill R.L., Schindelbeck R.R. and Tsegaye T. (1999) Integrated assessment of space, time, and management-related variability of soil hydraulic properties. *Soil Science Society of America Journal*, **63**, 1599–1607.

van Genuchten M.Th. (1980) A closed–form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

van Genuchten M.Th. and Nielsen D.R. (1985) On describing and predicting the hydraulic properties of unsaturated soils. *Annales De Geophysique*, **3**, 615–628.

van Genuchten M.Th. and Simunek J. (2004) Integrated modelling of vadose-zone flow and transport processes. In *Unsaturated-zone Modeling*, Wageningen UR Frontis Series, Feddes R.A., de Rooij G.H. and van Dam J.C. (Eds.), Kluwer Academic Publishers: Dordrecht, pp. 37–72.

van Genuchten M.Th., Leij F.J. and Wu L. (1999) Characterization and measurement of the hydraulic properties of unsaturated porous media. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 1–12.

Viaene P., Vereecken H., Diels J. and Feyen J. (1994) A statistical analysis of six hysteresis models for the moisture characteristic. *Soil Science*, **157**, 345–355.

Visser W.C. (1968) An empirical expression for the desorption curve, in water in the unsaturated zone. Rijtema P.E. and Wassink H. (Eds.), *Proceedings of the Wageningen Symposium, IASH/AIHS*, UNESCO: Paris, Vol. 1, pp. 329–335.

Vogel T. and Cislerova M. (1988) On the reliability of unsaturated hydraulic conductivity calculated from the moisture retention curve. *Transport in Porous Media*, **3**, 1–15.

Vogel T., Nakhaei M. and Cislerova M. (1999) Description of soil hydraulic properties near saturation from the point of view of inverse modeling. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., *et al.* (Eds.), University of California: Riverside, pp. 693–704.

Vogel H.-J. and Roth K. (2003) Moving through scales of flow and transport in soil. *Journal of Hydrology*, **272**, 95–106.

Vogel T., van Genuchten M.Th. and Cislerova M. (2001) Effect of the shape of the soil hydraulic functions near saturation on variably-saturated flow predictions. *Advances in Water Resources*, **24**, 133–144.

Warrick A.W. (1998) Spatial variability. In *Environmental Soil Physics*, Hillel D. (Ed.), Academic Press: San Diego, pp. 655–676.

Warrick A.W. and Nielsen D.R. (1980) Spatial variability of soil physical properties in the field. In *Applications of Soil Physics*, Hillel D. (Ed.), Academic Press: New York, pp. 319–344.

Wierenga P.J., Hills R.G. and Hudson D.B. (1991) The Las Cruces trench site: characterization, experimental results, and one-dimensional flow predictions. *Water Resources Research*, **27**, 2695–2705.

Wildenschild D., Hopmans J.W. and Simunek J. (2001) Flow rate dependence of soil hydraulic characteristics. *Soil Science of Society of America Journal*, **65**, 35–48.

Yeh T.-C. and Simunek J. (2001) Stochastic fusion of information for characterizing and monitoring the vadose zone. *Vadose Zone Journal*, **1**, 207–221.

Zhu J. and Mohanty B.P. (2003) Effective hydraulic parameters for steady state vertical flow in heterogeneous soils. *Water Resources Research*, **39**, 1227.

Zurmühl T. and Durner W. (1998) Determination of parameters for bimodal hydraulic functions by inverse modeling. *Soil Science Society of America Journal*, **62**, 874–880.

75: Determining Soil Hydraulic Properties

WOLFGANG DURNER AND KAI LIPSIUS

Institute of Geoecology, Department of Soil Physics, Braunschweig Technical University, Braunschweig, Germany

Hydraulic properties are required for modeling water and solute transport in unsaturated soils. The bottleneck for the successful application of numerical simulation models lays usually in their parameter estimation requirements. Methods to determine hydraulic properties can be classified into indirect and direct approaches. Indirect methods encompass the estimation of hydraulic properties by pedotransfer functions from more easily measured soil properties, and the prediction of the unsaturated hydraulic conductivity function from the water retention curve (WRC). In direct methods, observations of flow attributes from laboratory or field experiments are evaluated. This article reviews common methods to estimate the hydraulic conductivity function from the water retention characteristic and various direct measurement techniques in the laboratory and the field. We conclude with an outlook on contemporary developments in measurement techniques, stressing the key role of inverse modeling of experiments to derive optimum hydraulic properties and the importance of a future combination of noninvasive measurement techniques with inverse modeling by stochastic data fusion.

INTRODUCTION

A proper characterization of water flow processes is needed in nearly all basic and applied aspects of soil, water, nutrient, and salinity management research (van Genuchten *et al.*, 1999b). Water flow in soils is typically described with the Richards equation (Richards, 1931)

$$C(h)\frac{\partial h}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} - 1\right)\right] + s \qquad (1)$$

where t is time [T], z is a spatial coordinate [L], positive downward, h is the matric potential, expressed as pressure head [L], $C(h)$ is the specific water capacity [L^{-1}], defined by the change of the volumetric water content θ [L^3 L^{-3}], with pressure head, $C = \partial\theta/\partial h$, $K(h)$ is the unsaturated hydraulic conductivity [L T^{-1}], and s is a source/sink term [T^{-1}]. The model is completed by appropriate initial and boundary conditions. Since C and K are nonlinearly dependent on h, its solution generally requires numerical methods. Equation (1) is derived from the combination of the Darcy-Buckingham equation and continuity considerations in the framework of the continuum theory, and is valid for the measurement scale (see Durner and Flühler, 2005; **Chapter 74, Soil Hydraulic Properties, Volume 2**). It is also frequently used as process model for water transport at much larger scales. The coefficients C and K are then used as *effective* properties, which have the same names as for the local scale, but their values are no longer necessarily consistent with the local definition. Determining hydraulic properties is the process of deriving the constitutive relationships $\theta(h)$ and $K(\theta)$, or $K(h)$, as used in equation (1). The relationship $\theta(h)$ is called (water retention curve) *WRC* . The dependence of the hydraulic conductivity, $K(h)$, on water content or pressure head is called '*hydraulic conductivity curve*'.

This article reviews the indirect and direct methods for the determination of soil hydraulic properties. Direct methods are based on flow experiments in the field or with soil samples in the laboratory. They rely on observations of flow attributes, such as water potential, water content or water flux density. These measurements are far from simple. Water content measurement is treated in this encyclopedia in **Chapter 72, Measuring Soil Water Content, Volume 2** (Topp and Ferré, 2005), soil water potential measurement is treated in **Chapter 73, Soil Water Potential Measurement, Volume 2** (Durner and Or, 2005). Water flux density can be determined at system boundaries,

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

using scales or burettes in the laboratory. For measuring fluxes *in situ*, there are no reliable and accurate methods. Accordingly, *in situ* flux measurements are not used up to date for the determination of hydraulic properties (Dirksen, 1999b).

Indirect methods are used to estimate hydraulic properties from more easily measured data, using regression or neural network algorithms. In particular, the unsaturated conductivity function is seldom measured, but commonly estimated from the WRC and matched to a single measured conductivity value. Because of the shortcomings of direct measurement procedures, indirect estimation methods are gaining popularity. All estimation procedures, however, need the results of direct measurements as benchmarks for validation. Furthermore, reliable and efficient experimental procedures are critical to improve the understanding of flow and transport processes in variably saturated media, regardless of the advances in the formulation of indirect methods.

A review on measurement methods needs to address aspects of sensor technology, instrumentation, experimental design, techniques to evaluate observed data, scale issues, parameterization of hydraulic functions, parameter estimation techniques, and uncertainty estimation and propagation. Covering all these issues in the required depth would be far beyond this contribution. So the focus is on principles of indirect estimation procedures (Section "Indirect estimation of hydraulic functions"), on direct laboratory measurement methods (Section "Laboratory methods"), and on field methods (Section "Field methods"). Topics that are closely related to soil hydraulic measurements are treated in the accompanying contributions in this encyclopedia. These are, in particular, soil water potential measurement (Durner and Or (2005), **Chapter 73, Soil Water Potential Measurement, Volume 2**), water content measurement (Topp and Ferrè (2005), **Chapter 72, Measuring Soil Water Content, Volume 2**), estimation of hydraulic properties by pedotransfer functions (Schaap, 2005, **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**), scale issues (Hopmans and Schoups, 2004, **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**), recent developments in parameter estimation techniques (Vrugt and Dane, 2005, **Chapter 77, Inverse Modeling of Soil Hydraulic Properties, Volume 2**), and uncertainty propagation in hydrologic models (Brown and Heuvelink, 2005, **Chapter 79, Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport, Volume 2**). The definition of hydraulic functions in the framework of the continuum theory, the parameterization of the hydraulic functions including hysteresis, and the issue of scale dependence of hydraulic properties are treated by Durner and Flühler (2005)).

Methods and techniques to measure soil hydraulic properties are described in a series of monographs. The reference with the widest distribution is the classic monograph "Methods in Soil Physics – Physical Methods" of the American Society of Agronomy. The previous edition (Klute, 1986) has recently been revised (Dane and Topp, 2002), and covers almost all practical measurement techniques including soil sampling, uncertainty, and sensor technology. Hydraulic measurement methods are further discussed in some textbooks, for example, Kutilek and Nielsen (1994), Dirksen (1999a), and Flühler and Roth (2004). The scientific state of knowledge on direct, inverse, and indirect measurement methods at the time of the millennium is documented on 1600 pages in the "Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media" (van Genuchten *et al.*, 1999a). Direct measurement methods cover about a fourth of the volumes.

The Purpose of Hydraulic Measurements

Methods to determine hydraulic properties differ with respect to their accuracy, measurement range, difficulty of implementation, and demand for time and capital. Before selecting a specific measurement method, the purpose of the measurements must be manifest. Purposes may be classified into three groups.

(i) Soil hydraulic properties are often needed for a hydraulic classification of soils, in a similar manner as the particle size distribution is used for a textural classification. Knowledge of basic hydraulic properties, such as field capacity or plant available water content, is useful for a variety of purposes, where soil moisture storage, soil wetness (affecting oxygen supply for plant roots and redox state of soil), surface runoff, susceptibility to soil erosion, and other large-scale properties of soils are of interest. Indirect methods are often appropriate for this group (Schaap, 2005, **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**).

(ii) Today's prevalent demand for hydraulic properties is their use in numerical simulation of water transport by Richards' equation. Estimation of water recharge through the vadose zone for water balance calculations are a classic application, but more important is now their use for agricultural, ecological, and environmental purposes, such as irrigation control, fertilizer management, and contaminant fate modeling. The focus of subsurface models of water and solute transport has increasingly been shifted toward environmental research, with the primary concern on the subsurface fate and transport of various substances, such as nutrients, pesticides, pathogens, pharmaceuticals, viruses, bacteria, colloids, and toxic trace elements (Simunek, 2005, **Chapter 78, Models**

of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2). The crucial bottleneck for the successful application of these models is their parameter estimation requirements.

(iii) Finally, accurate measurement of soil hydraulic properties is required for a further improvement in the basic understanding of soil hydraulic processes, that is, in order to test and improve the process knowledge we have. Examples are the understanding and assessment of nonequilibrium phenomena in water flow, further progress in describing hysteresis in soil water flow, and the question up to which scale the Richards equation process model provides an appropriate effective process representation for unsaturated water transport (Durner and Flühler, 2005).

Whereas the demand on accuracy, resolution, precision, and reliability of the measurements is moderate for the first group, it is higher for the second and third. For the second group, we are particularly faced with scale considerations. For the third group, the precision, reliability, and validity of measurements are of utmost importance, in order to avoid misconceptions.

The Challenge of Determining Soil Hydraulic Properties

Determining soil hydraulic properties is demanding for a variety of reasons. Soils are porous media with a three-dimensional arrangement of interconnected voids that form a highly complex pore system. The topology of this system shows, in general, a hierarchical arrangement, with spatial and temporal variability on a multitude of scales. The microscopic properties of the pore system determine the macroscopic hydraulic behavior. A complete understanding of water flow in soils requires a thorough understanding of processes on scales much smaller than the usual measurement scale and the ability to express effective hydraulic properties at scales much larger than the measurement scale (Durner and Flühler, 2005; Hopmans and Schoups, 2004, **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**).

A specific problem in the determination of soil hydraulic properties lays in the fact that quality control and validation of measurement results is extremely difficult. Contrary to soil chemical analysis, there is virtually no possibility for reliable interlaboratory comparisons (Dirksen, 1999b). The reasons for this are: (i) as opposed to consolidated porous media, the soil pore system is not a stable structure. Just those parts of the pore system, which control the water transmission near saturation, are most fragile, and there is always a danger that the measurement process itself changes the system (Ghezzehei and Or, 2003). Therefore, repetitive measurements on the same soil sample by different laboratories are impractical. The sampling process itself often causes the most severe disturbance, when an "undisturbed" soil sample is isolated from the natural embedding. (ii) Soils exhibit considerable temporal variability (Mapa *et al.*, 1986; Ahuja *et al.*, 1998; Leij *et al.*, 2002). Thus, measurements at the same site may yield different results if applied at different times (van Es *et al.*, 1999). (iii) Soil is a living organism and the pore system is affected by a variety of interacting biological, chemical, and physical processes. Matrix surface properties are variegated and may change depending on physical, chemical, and biological factors, thereby changing the macroscopic hydraulic behaviour. (iv) Spatial variability of hydraulic properties, finally, is probably the biggest problem (Nielsen *et al.*, 1973, 1986). Different measurement methods use different sample volumes and sample numbers, dictated by standard procedures and the apparatus available for the various methods. This implies that, in a comparison of measurement methods, considerable uncertainty about the result of the comparison will always be induced by natural variability (Stolte *et al.*, 1994; Munoz-Carpena *et al.*, 2002). For example, the determination of field-saturated conductivity, K_{fs}, may vary two or more orders of magnitude among different field and laboratory methods.

Testing measurement methods on synthetic soil-like porous media is not a solution to this dilemma. The pore system properties of repacked soil samples are often very different from the properties of an undisturbed soil (Torquato, 2001). It is notable that especially uniform fine sand, being the favorite material used for research purposes in soil physics, has properties that are quite untypical for a structured natural soil. Since the early observations of Kozeny (1927) on particle segregation during packing, the problem of constructing a synthetic porous medium in a fully reproducible manner, with pore system properties comparable to a natural soil, has remained unresolved (Lebron and Robinson, 2003).

An overview on various field and laboratory methods for determining unsaturated hydraulic properties shows that many techniques have been proposed, but most of them are limited to relatively narrow ranges of water potential (h) or water content (θ). Figure 1 illustrates this, showing results of five different measurement methods to determine the hydraulic conductivity. The existing experimental procedures all have their own unique advantages and limitations (Gee and Ward, 1999), and selecting the most appropriate measurement for a specific task is usually not trivial.

Classification of Methods

Determining hydraulic properties of soil encompasses direct measurements and indirect estimation methods. Because of the shortcomings of direct measurement procedures, *indirect estimation methods* are gaining popularity. Computers offer the possibility to generate indirect estimates using

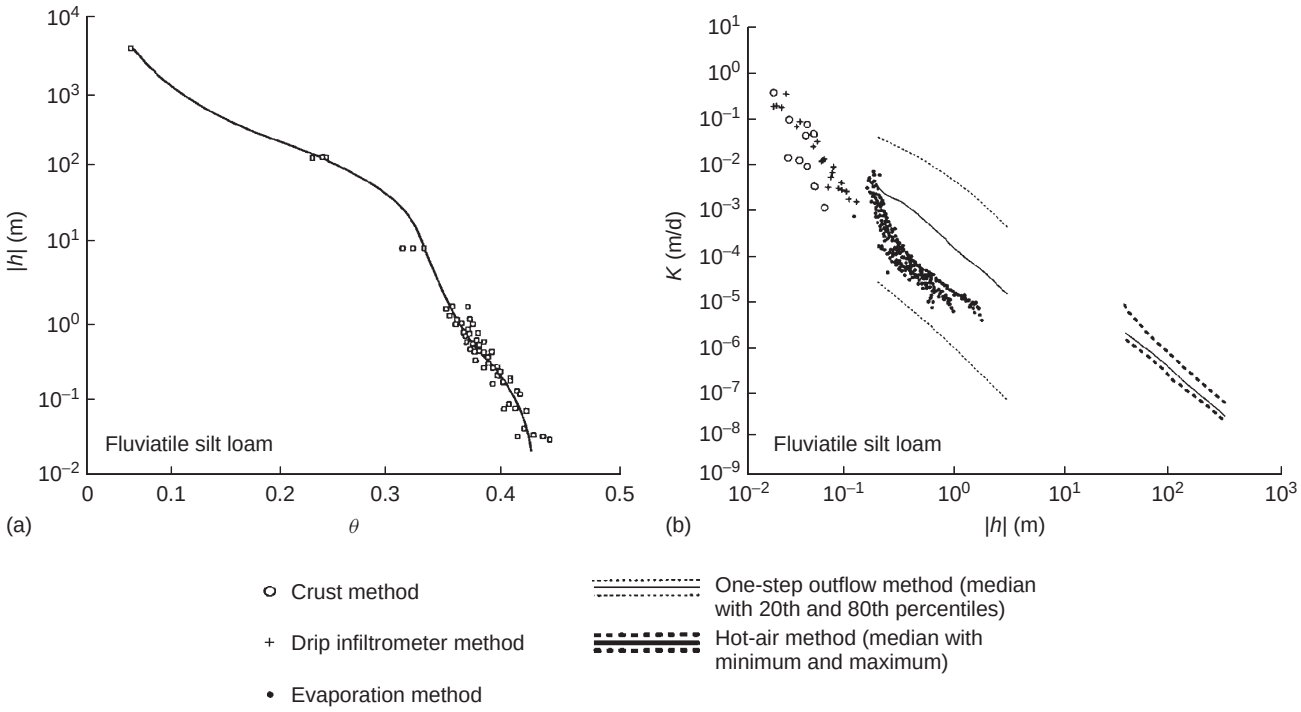


Figure 1 Hydraulic properties of a fluviatile silt loam. (a) Water retention curve measurements. (b) Unsaturated conductivity measurements, determined by five different methods (Reproduced from Stolte *et al.,* 1994 by permission of Soil Science Society of America)

regression or neural network algorithms. In practice, unsaturated conductivity is seldom measured, but estimated from the saturated conductivity (or preferably an other matching point) and the WRC. The principles of indirect estimation procedures are outlined in Section "Indirect estimation of hydraulic functions".

All estimation procedures need the results of *direct measurements* as benchmarks for validation of their results. Reliable and efficient experimental procedures are critical to improve the understanding of flow and transport processes in variably saturated media, regardless of the advances in the formulation of indirect methods. Experiments for measuring hydraulic properties are based on hydrostatic, steady state, or transient flow conditions. The methods can be grouped in those that aim at measuring (i) water retentivity, (ii) saturated conductivity, (iii) unsaturated conductivity or water diffusivity, and (iv) simultaneously retentivity and conductivity. We may further distinguish between *laboratory methods* and *field methods*. The reason for treating laboratory (Section "Laboratory methods") and field experiments (Section "Field methods") in this review in separate sections is less motivated by the different scale, but by the fact that in the laboratory a much better control of boundary conditions, fluxes across boundaries, and integral water content measurements can be achieved.

Finally, methods differ in how the observations of flow attributes, such as pressure heads, water contents or fluxes are evaluated. In the early stages of soil physics, methods have been developed for determining the WRC or the saturated conductivity (Gardner, 1986). Determination of the water retention characteristic can be done directly, by pairing water content and water potential measurements in the laboratory or in the field. The determination of saturated conductivity is achieved by a closed-form inversion of the flow equation. Analytical solutions for the unsaturated conductivity can be obtained in the laboratory by a series of consecutive unit-gradient experiments, where a constant flux or pressure head is applied at the top of a sample, and a corresponding suction at the bottom (Dirksen, 1991, 1999a, 1999b). These methods are conceptually straightforward and easy to implement. Their main disadvantage is that they take a long time, and are therefore tedious and expensive. In order to achieve hydraulic equilibrium, the sample sizes must be kept small and sizes are often below the representative elementary volume, REV (Durner and Flühler, 2005). Accordingly, the resulting properties are generally highly variable.

Measurement of unsaturated hydraulic conductivity and water diffusivity poses greater obstacles. The methods are based on solving the inverse problem, where a model of the flow process is optimized to match observations (Russo *et al.*, 1991; Hopmans *et al.*, 2002; Vrugt and Dane, 2005, **Chapter 77, Inverse Modeling of Soil Hydraulic**

Properties, Volume 2). For some simplified cases, solving the inverse problem is accomplished by closed-form solutions, such as the determination of diffusivity from a quasi-analytical solution of the Richards' equation. In most other cases, solution of the inverse problem can only be achieved by fitting numerically simulated data to observations. This requires the use of nonlinear parameter estimation techniques. In doing this, the next logical step is to simultaneously estimate retentivity and conductivity parameters by inverse modeling.

METHODS TO DETERMINE SOIL HYDRAULIC FUNCTIONS

Indirect Estimation of Hydraulic Functions

Estimation by Pedotransfer Functions

In pedotransfer functions, the water retention and conductivity functions are derived from more easily measured soil properties, such as soil texture, bulk density, and organic matter content. Methods to derive soil hydraulic properties indirectly include multiple regression, classification, and neural network predictions (van Genuchten *et al.*, 1999a). Comparisons of indirectly determined hydraulic properties to directly measured properties are manifold and remain a topic of ongoing research (Schaap, 2005, **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2**). In general, it is found that the accuracy of pedotransfer estimates of hydraulic functions for characterizing field average is comparable to simple direct measurements, if spatial variability is considered. Depending on the desired use of these properties, indirect methods have evolved to a point where they provide reliable answers for many problems. However, further improvement of indirect methods hinges on experimental data, which must be obtained with direct procedures.

Conductivity Estimation by Statistical Estimation Methods

The most frequently practiced way to determine the unsaturated hydraulic conductivity function is by estimating its shape from the WRC. The equilibrium WRC can be regarded as an effective statistical cumulative distribution of equivalent pore sizes of a porous medium. From that, attempts to predict the conductivity curve based on capillary models of the pore space by application of Poiseuille's law have been early set (Kozeny, 1927; Purcell, 1949; Childs and Collis-George, 1950; Burdine, 1953; Fatt, 1956; Wyllie and Gardner, 1958; Marshall, 1958; Millington and Quirk, 1961; Mualem, 1976; Alexander and Skaggs, 1984). Capillary pore-bundle models are now routinely applied in an implicit manner, because they yield closed-form equations for the hydraulic conductivity function if combined with specific retention functions. Notable are the models Brooks-Corey/Burdine (Brooks and Corey, 1964), van Genuchten/Mualem (van Genuchten, 1980;), or Russo/Gardner (Russo, 1988). The model of van Genuchten/Mualem is nowadays the most frequently used. It is given by

$$S_e = (1 + (\alpha|h|)^n)^{-m} \tag{2}$$

where $S_e = (\theta - \theta_r)/(\theta_s - \theta_r)$ is a scaled water content, called "*effective saturation*", and $\alpha > 0$ [L^{-1}], $n > 1$ [$-$] and $m > 0$ [$-$] are empirical curve shape parameters. If m is related to n by $m = 1 - 1/n$, the corresponding conductivity function is given by

$$K_r(S_e) = S_e^l \left[1 - (1 - S_e^{1/m})^m\right]^2 \tag{3}$$

where the parameter l [$-$] is called "tortuosity factor". When applying the van Genuchten/Mualem conductivity model (3), the user should be aware of a model artifact for small values of the parameter n, which is discussed in Durner and Flühler (2005).

Despite the fact that these models often rely on oversimplified representations of the medium's pore space as a bundle of cylindrical capillaries, there is much empirical evidence that capillary models yield reasonable shapes for the conductivity curve in a limited range of matric potentials. However, the prediction of the absolute level of conductivities was never successful (Jackson *et al.*, 1965). Hence, "matching factors" are needed to match the unsaturated conductivity curve with observed conductivity values. A matching factor is used to scale a predicted value of the relative conductivity curve, K_r (θ), to a directly measured conductivity, K_{ref} (θ_{ref}), at a reference water content, θ_{ref}.

$$K(\theta) = \frac{K_{ref}(\theta_{ref})}{K_r(\theta_{ref})} K_r(\theta) \tag{4}$$

In practice, most often the saturated conductivity, K_s, is used as matching factor, because measurement of the saturated conductivity is relatively easy, both with respect to the experimental procedure and with respect to the inversion the Darcy-Buckingham flux equation. However, already as early as in the 1960s, the notorious instability of matching the predicted conductivity function to a measured value at saturation was evident (Jackson *et al.*, 1965). This is due to the extreme dependency of the conductivity near saturation on pore structure (Durner, 1994), which causes an enormous variability of the saturated conductivity in natural structured soils and makes the mathematical prediction near saturation unstable. The current recommendation is to use an unsaturated reference measurement, which leads to a more reliable description of the overall conductivity function (Hopmans *et al.*, 2002). This, however, leaves the

question open as to how the shape of the conductivity function close to full saturation should be described.

Also, the usual predictive models tend to fail at low matric potentials (Nimmo, 1999), with deviations up to five orders of magnitude (e.g. Kosugi, 1999). This was often attributed to the moisture dependence of pore tortuosity and connectivity, and was accounted for by introducing a tortuosity/connectivity term θ^l. Contrary to theoretical considerations of, for example, Mualem (1976), the analysis by Hoffmann-Riem *et al.* (1999) shows that the tortuosity factor l must be regarded as a (second) purely empirical matching factor, with a high variation from soil to soil. This is confirmed by the study of Schaap *et al.* (1999) who found an overall averaged optimum value of $l = -1$ for the van Genuchten/Mualem model when testing soils of the UNSODA database. If interpreted in the original model theory of Mualem (1976), this value would be "unphysical". Tuller and Or (2001) argue that the failure of predicted conductivity functions in the low moisture range is due to the lack of consideration of flow in thin liquid films and in crevices and corners of angular pores. In a series of papers (Tuller *et al.*, 1999; Or and Tuller, 1999; Tuller and Or, 2001) they developed a thermodynamically based conductivity estimation model that improves the predictions at low saturations significantly.

Estimation of Constitutive Relationships by Pore-network Models

Computational pore-scale network models describe two-phase porous media flow systems by resolving individual interfaces at the pore scale, and tracking these interfaces through the pore network (Celia *et al.*, 1995). Coupled with volume-averaging techniques, these models can reproduce relationships between macroscopically measurable variables like capillary pressure, water content, and relative permeability. Over the last decade, this approach has taken a tremendous development. This is due to rapid advances in instrumental techniques for a noninvasive three-dimensional characterization of the pore spaces and for flow and transport processes (Blunt and Hilpert, 2001). In connection to this, the use of fast computers allows to use sophisticated image analysis tools and tomography to reconstruct three-dimensional statistically equivalent pore-scale network models (for an overview, see Blunt and Hilpert, 2001). Also, simulation of water and air distribution and transport in reconstructed porous networks on the microscale, for example, by Lattice-Boltzmann techniques (Krafczyk and Rank, 1995), can be used to infer macroscopic hydraulic properties. On a macroscopic level, Liu and Bodvarsson (2003) modeled effective constitutive relationships for fracture networks and for dual-permeability media. Currently, the use of these techniques is in a research state, and aims more toward understanding various flow phenomena than on practical determination of hydraulic properties.

Laboratory Methods

Laboratory methods have the advantage of being conducted in a controlled environment. On the experimental side, significant progress has been achieved through the use of improved computer technology that lead to automation of experimental devices with flexible control of boundary conditions and precise data acquisition at high temporal resolution (Durner *et al.*, 1999a). Dirksen (1999b) states that measurements of hydraulic properties should be made in the laboratory, unless there are overriding reasons to perform them *in situ*, such as the presence of strongly layered soil profile, large unstable structural elements, and an abundance of stones. Vogel and Roth (2001) put forward that direct measurements of the hydraulic functions are only feasible on the scale of core samples treatable in the laboratory. Laboratory methods are, however, subject to the limitation that some disturbance is introduced in manipulating the sample, even if "undisturbed" soil cores are used. In addition, the measurements may be affected by hydraulic effects not present in the field (Munoz-Carpena *et al.*, 2002). Santos *et al.* (1999) pointed out the importance of *in situ* methods to obtain reference values against which laboratory methods should be tested.

Direct laboratory measurement methods can be grouped into static equilibrium methods to determine water retentivity (Section "Hydrostatic equilibrium methods"), steady-state flux methods to determine conductivity (Section "Steady-state laboratory methods"), and transient methods to determine unsaturated conductivity or diffusivity (Section "Transient laboratory methods"). Transient methods are increasingly replacing the conventional equilibration and steady-state methods, because they can be applied to larger soil samples and speed up the experiments considerably. Traditionally, transient methods focus on determining the unsaturated hydraulic diffusivity or conductivity by measuring water fluxes at the system boundary. However, simultaneous measurements of water content and pressure in a soil sample during computer-controlled transient experiments are probably the best and quickest method to derive, also directly, the water retention function (Schultze, 1998; Zurmühl, 1998).

Hydrostatic Equilibrium Methods

Hydrostatic equilibrium methods can be used to determine the WRC. There are several standard methods. For references to the primary literature and for details in the application of the methods, the reader is referred to Kutilek and Nielsen (1994) and Dane and Topp (2002). Nimmo (2002) compiled a table with useful guidelines for the selection of the most appropriate method for a given purpose, considering aspects of sample size, measurement range, difficulty, duration, and cost.

- *Long column method*: A long homogeneous column is put in contact with a water reservoir at its lower end.

Dependent on the initial state of the column, saturated or dry, an imbibition process or a drainage process will take place. After equilibration, the column is sliced and the water contents are determined. Problems of this method lie in the extreme long time demand necessary for equilibration, the limited pressure range, and in the difficulty of obtaining a homogeneous soil column. Advantage of the method is an almost arbitrary fine resolution with respect to $\theta(h)$.

- *Hanging water column*: A sample is placed on a fine sand bed or a porous plate. Suction is applied via a hanging water column, with a length smaller than the corresponding air-entry pressure of the sand, or via a controlled vacuum, which is applied to the space below the porous plate. After the sample has reached hydraulic equilibrium, its water content is determined gravimetrically. Advantages of this classic method are its relatively low cost and technical operating expense, which allows measuring many samples simultaneously.
- *Suction table*: the method is similar to the hanging water column, but instead of a water column, a technically controlled vacuum can also be applied to the water phase. The range of the method is limited by the atmospheric pressure.
- *Pressure plate extractor*: A sample is placed on a porous plate inside a pressure container. Air pressure is applied to the container, which induces displacement of water toward and through the porous plate to the free atmosphere. After reaching equilibrium, the water content is determined gravimetrically. This method can be applied in combination with the Hanging water column method, where it yields values for the dry range of the $\theta(h)$.
- *Pressure cell* (*Tempe cell*): A ring sample is mounted on a porous plate and connected to a top, through which the gas phase pressure can be regulated. Increasing the air pressure induces displacement of water through the porous plate, which can be measured in a burette that is connected to the outlet.
- *Centrifuge method*: An initially saturated sample is placed in a centrifuge and spinned with a specific velocity. After the centrifugal and capillary forces reach equilibrium the saturation can be determined. The method is well suited to determine residual water saturation. It has also been applied to measure hydraulic conductivity. The pressure range depends on the power of the centrifuge. Problems are costs and the mechanical destruction of the sample under high acceleration.
- *Controlled liquid volume*: In this method, defined points of $\theta(h)$ are determined by adding or extracting controlled volumes of liquid. Water then redistributes within the sample until the pressure is equalized. The main advantage of this method is that it achieves faster equilibration. However, during water redistribution part of the sample will be draining, and the other part will be wetting, and therefore hysteresis may affect the overall distribution.

The above listed methods are in most cases used in a series of equilibration steps to desaturate a soil from full saturation, thus yielding a drying curve. Common to all hydrostatic equilibrium methods is the considerable time requirement, in particular, if saturations at low potentials are to be determined. Since equilibrium time grows with the square of the length of the sample (Miller, 1980), equilibration times of large samples become prohibitively long for high suctions. Nonequilibration is in particular a problem for sands, where higher tensions can only be reached by equilibration via the gas phase (Gee *et al.*, 2002). For the dry range of a retention curve, there are not many alternatives. In the literature, the freezing method is described (Spaans and Baker, 2002; Bittelli *et al.*, 2003), which makes use of the thermodynamic equivalence between freezing and drying. Recently, vapor pressure measurements of soil with known water contents have been discussed (Nimmo and Winfield, 2002).

The usual way to evaluate equilibrium measurements is to pair the pressure at the center of the sample with the average water content of the whole sample. This linearization results in a retention characteristic that is smoothed around the air-entry point, as compared to the true point characteristic, in particular, for coarse porous media and for large samples. Liu and Dane (1995) and Jalbert and Dane (2001) provide correction procedures for this problem.

Steady-State Laboratory Methods

Steady-state methods aim at obtaining the hydraulic conductivity at a particular saturation, by inversion of Darcy's law. The procedure is based on the fundamental assumption that the rate of flow is proportional to the pressure gradient and the permeability constant is a property of the porous medium (Klinkenberg, 1941). We will restrict our discussion to methods that use water as a fluid. For practical recommendations and details on influences of temperature, ion composition, and other factors affecting fluid properties, see Dane and Topp (2002).

Steady-state in water flow is reached by applying either constant head or constant flux boundary conditions to the top or the bottom of soil samples. A flux boundary condition can be realized by drip irrigation (mostly through needles), by spraying (Dirksen and Matula, 1994), or by pumping water via a porous plate onto the top of the sample. None of these procedures is trivial. At the lower boundary, suction can be applied through a porous plate (sintered glass, metal, ceramic, or nylon). It is generally advisable to check the pressure head inside the sample by tensiometric measurements (Dirksen, 1991). For methods that aim at the saturated conductivity, K_s [L T^{-1}], it is important

that the flow resistance of the experimental device, in particular the membrane, is kept smaller than the resistance of the soil. Since K_s is extremely dependent on continuous macropores along the flow direction, any change of soil structure (compaction, sealing of water entry surface), and any bypass of water (gap between column casing and soil) must be avoided.

- *Constant head permeameter*: Water is applied by means of a mariotte device to the top of a soil sample at a constant head. Water leaving the sample at the bottom is also kept at a constant head, either at zero pressure by free drainage through a suspension mesh to the atmosphere, or to a positive pressure via an overflow vessel. The experiment can equally be performed with the flow direction of water from the bottom to the top. K_s is calculated by rearranging the Darcy equation. Limitations of the method are related to small or inadequate sample size, soil disturbance during core collection, and possible short-circuit flow through macropores along the core wall. On the other hand the method is simple, inexpensive, and convenient. Despite potential limitations, the method remains as one of the most popular means for measuring K_s and is often used as a benchmark for evaluating other methods (Reynolds *et al.*, 2000).
- *Crust method*: The crust method (Bouma *et al.*, 1983) determines hydraulic conductivity, *K(h)*, by measuring the water flux density into an unsaturated soil sample. Unsaturated conditions are obtained by supplying water through a crust made of sand and special cement with a lower hydraulic conductivity than the soil sample. In a typical setup, a saturated soil sample is placed on a long sand column to attain nearly gravitational flow in the sample. Water is supplied to the crust by ponding, using a Mariotte burette. A tensiometer is placed in the soil few centimeters below the top of the sample. The flux density is derived from the discharge from the burette. If only gravitational flow is assumed and steady-state is reached (e.g. 24 h no change in pressure head), the flux density, q [$\mathrm{L\,T^{-1}}$], equals the unsaturated hydraulic conductivity. To obtain information about the whole range of *K(h)*, several crusts would be needed. This makes the crust method very tedious. The crust method is also used in the field.
- *Drip Infiltrometer Method*: The drip infiltrometer method (Dirksen, 1991) determines the hydraulic conductivity by infiltration. Typically, the soil sample is placed on a sand box with a hanging water column. Water is applied from a reservoir through hypodermic needles on top of a sample. Tensiometers are placed in the soil at different depths. The pressure head gradient ($\mathrm{d}h/\mathrm{d}z$) in the sample can be controlled by adjusting the height of a hanging water column with overflow, attached to the sand box. Measurements start after steady-state is reached. By measuring the flux density (q) and the pressure heads (h), the unsaturated hydraulic conductivity *K(h)* is calculated for each compartment between two tensiometers with Darcy's law. Again it is very laborious to measure the whole range of *K(h)*. Obtaining values for low flux densities is limited by the application of a constant flux rate and very long times to attain steady-state.
- *Atomized Water Spray Method*: Delivery of water uniformly to the soil surface can also be accomplished with a controlled atomized spray. Dirksen and Matula (1994) constructed and tested a water spray system that delivers water accurately to the surface of a 20-cm diameter soil column at rates ranging from 10^{-4} to 10^{-7} cm $\mathrm{s^{-1}}$. The advantage of this method over more conventional laboratory methods is that large samples can be tested and it requires no ceramic plates and therefore does not have a contact resistance problem. However, the system is relatively cumbersome and does require thermal control for optimal results.
- *Steady-state evaporation method*: Fujimaki and Inoue (2003) introduced a method that determines hydraulic conductivity from a water content profile under steady-state upward flow. Steady-state was induced by setting both constant meteorological conditions and a constant inflow rate from the bottom, with evaporation demand being higher than the supply rate. After steady-state is reached, the soil column is sectioned to measure the water content profile. Isothermal water vapor flux is evaluated via inverse optimization. An independently determined $\theta(h)$ is also required. Unlike most other methods, this method is not limited to the wet range/tensiometer range ($h > -700$ cm).
- *Disc Infiltrometer method*: In the disc infiltrometer method (Simunek *et al.*, 1999), water is applied at negative pressure using a tension disc (see Section "Field methods") that is placed on top of the sample. The soil profile is instrumented with both Time domain reflectometers (TDR) (at one location) and tensiometers (at several locations). The measured data are analyzed using Wooding's (1986) analytical solution. Wooding's analysis requires steady-state infiltration rates at different supply pressure heads. Simunek *et al.* (1999) compared analysis by numerical inversion against results of Wooding's solution. Unsaturated hydraulic conductivities corresponded well, but K_s was overestimated.
- *Heat Pipe Method*: Very low unsaturated *K(h)* measurements (10^{-7}–10^{-12} cm $\mathrm{s^{-1}}$) have been reported by Globus and Gee (1995) using a heat pipe method. The method uses a partially wetted sample that is sealed and equilibrated with an applied thermal gradient. In the heat pipe method, equilibrium is attained when liquid water flow from the cool end equals water vapor flow from the warm end, and the water content profile

has become stable. After a given period, ranging from days to several weeks, the soil is removed from the column and sampled for water content. Water diffusivity values for the soil are determined from the measured water content gradients. *K(h)* is subsequently computed from an independently measured $\theta(h)$ for the soil. The method requires temperature control and is time consuming. *K(h)* is actually a lumped parameter, since it represents a combined estimate of the vapor and liquid conductivity (Gee and Ward, 1999).

Transient Laboratory Methods

With decreasing pressure head, steady-state methods loose their attractiveness, because constant fluxes and unit gradients are hard to obtain. Transient laboratory techniques do not employ equilibrium steps and are therefore much more rapid. Traditional transient laboratory methods, such as outflow or evaporation experiments, however, show relatively little sensitivity to the hydraulic conductivity at near-saturated conditions. Thus it is advisable to determine the hydraulic conductivity in the wet range with steady-state experiments, whereas transient conditions are used for the drier range (Wendroth and Simunek, 1999).

In the last few decades, several transient methods have been proposed for characterizing soil hydraulic properties, including one-step outflow, multistep outflow, evaporation, downward, and upward infiltration methods (UIMs). Each of these methods uses the change in column weight versus time to infer hydraulic properties. If simultaneous determination of retention and conductivity is sought, additional monitoring of matric potential changes at one or more positions in the samples, or independently determined $\theta(h)$ are required. Otherwise, only diffusivity can be determined. All transient methods assume uniform one-dimensional flux according to Richards' equation in a homogeneous porous medium, and apply constant or changing heads or constant or changing fluxes to the top or to the bottom of a soil sample. For unsaturated experiments, where matric potential is negative, a porous plate or a membrane with appropriate pore sizes is required to apply the pressure. In the following, the principles of the various methods are shortly described. For details in the experimental and evaluation procedures, the reader is referred to Kutilek and Nielsen (1994) and Dane and Topp (2002).

- *Falling Head Permeameter*: This method is used to determine K_s for samples with low conductivity. The upper surface of a soil sample is connected to a water-filled burette, and the lower face to a constant boundary pressure. When water is allowed to flow, the water level falls, and the pressure difference between the upper and lower boundary decreases exponentially. Since the pressure difference is large as compared to the length of the soil sample and the diameter of the burette is usually chosen to be considerably smaller than the diameter of the soil sample, a high hydraulic gradient is applied and the flux can be monitored with high precision. This speeds up the measurement process for soils with low hydraulic conductivity, as compared to the constant head method. Other shortcomings of the constant head permeameter method, however, cannot be overcome. The conductivity is calculated by

$$K_s = \frac{1}{(t_2 - t_1)} \frac{A_B L_S}{A_S} \ln\left(\frac{h_2}{h_1}\right) \tag{5}$$

where A_B and A_S are the areas of the burette and the soil, respectively, L_S = height of the soil, and h_1 and h_2 are the head differences between the upper and lower boundary at times t_2 and t_1, respectively.

- *One-step and Multistep Outflow method*: The one-step method was originally introduced by Gardner (1958) to determine the diffusivity, $D(\theta)$. Saturated soil samples on top of a ceramic plate are exposed to a high air pressure on top of the sample or to a sudden reduction in water pressure below the ceramic plate. The pressure change induces unsaturated flow in the soil sample, with the ceramic plate remaining saturated. The cumulative outflow of water is recorded in a burette or by a scale. The historic reasons for favoring the *one-step* method (Doering, 1965) lie in the possibility to apply semianalytical solutions of the flow equation to identify the hydraulic diffusivity function, provided the conductivity curve follows an exponential function (Gardner, 1958; Gardner, 1962). Since the mid-80s, following the pioneering work of Kool and coworkers (Kool *et al.*, 1985a,b; Kool and Parker, 1987) outflow methods are now evaluated by inverse modeling to determine the soil hydraulic functions $\theta(h)$ and *K(h)* simultaneously (Hopmans *et al.*, 2002). Because *one-step* experiments often yielded nonunique inverse solutions, investigators recommended to include additional $\theta(h)$ data (van Dam *et al.*, 1992; Bohne *et al.*, 1993), or tensiometric measurements in the object function (Toorman *et al.*, 1992; Eching and Hopmans, 1993). A major point of critique with respect to the one-step method has always been that the quick change of the boundary condition does not represent natural conditions (van Dam *et al.*, 1992, 1994), and the sensitivity of the method can be low if, as a result of a large pressure drop, a very thin drained soil layer next to the porous plate controls the total flow rate (van Dam *et al.*, 1992). For these reasons, the multistep method was suggested, where the pressure at the boundary is changed in several small steps (van Dam *et al.*, 1990). Eching *et al.* (1994), van Dam *et al.* (1994), Crescimanno and Iovino (1995), and Zurmühl (1996) demonstrated for the van Genuchten/Mualem model that the multistep method is superior to the one-step method because the hydraulic parameters are less correlated.

The outflow methods are constrained to the wet moisture range by the air-entry pressure of the porous plate.

- *Continuous Inflow/Outflow method*: Even with several small pressure changes as applied in the multistep outflow method, the pressure changes at the soil's boundary still occur shock-wise, which – at least for the drainage process – never occurs under natural conditions. As an alternative, Durner *et al.* (1999b) suggested a continuous smooth change of the boundary pressure, which showed to yield good results. A soil sample is placed on a membrane plate in contact with a pressure-controlled reservoir. The reservoir headspace is controlled by air pressure manipulation (Butters and Duchateau, 2002) to initiate water flow in the soil sample. Theoretically, both wetting and draining $\theta(h)$ and *K(h)*, up to the air-entry value of the membrane plate, including saturation can be measured. Butters and Duchateau (2002) used a combination of direct Darcian analysis and numerical inversion of Richards' equation for estimation of the hydraulic properties. The direct analysis provides K_s and θ_s that can be used to anchor these parameters in the inverse analysis. As a limitation, the direct K estimate applies to fairly short samples and is susceptible to tensiometer errors, especially of systematic nature. Durner (1994) recommended to directly and independently measure $K(h)$ in the range close to saturation where the inverse estimation from outflow experiments is insensitive.
- *Wind's Evaporation Method*: Similar to the outflow methods, Wind's evaporation method (Wind, 1968) aims at determining simultaneously $\theta(h)$ and *K(h)*. Tensiometers are installed at regular depth intervals in a saturated soil sample. The sample is placed on a balance and its surface is exposed to free or forced evaporation. Weight and pressure heads are recorded periodically. After completing the experiment, the volumetric water content of the sample is determined, which allows to recalculate the total water contents during the experiment. The water contents in the compartments around the tensiometers at different times are estimated from the measured pressure heads and the total water contents by fitting a parametric model of water retention characteristic to the observations. The subsequent evaluation is identical to the instantaneous profile experiment (see the Section on "Field Methods") by calculating the flux density from the decrease of the water contents of the compartments, and the pressure head gradient from the pressure heads of two adjacent compartments. Finally, the unsaturated hydraulic conductivity can be calculated with Darcy's law. The program package METRONIA (Halbertsma and Veerman, 1994; **`http://dino.wiz.uni-kassel.de/model_db/mdb/metronia.html`**) is available for this analysis. Bertuzzi *et al.* (1999) found that small uncertainties in tensiometer data greatly influence the hydraulic conductivity determined under wet conditions, but $\theta(h)$ estimates are relatively robust. Also, thermal effects on hydraulic conductivity estimates were far from negligible. Wendroth and Simunek (1999) reported that Wind's method (1968) method does not yield valid estimates of $\theta(h)$ and *K(h)* for layered soil samples. Finally the evaporation method fails in the near-saturation range where the hydraulic conductivity is highest, leading to very small hydraulic gradients that cannot be determined with sufficient accuracy (Wendroth and Simunek, 1999).
- *Hot-air Method*: The hot-air method (Arya *et al.*, 1975) determines the diffusivity function $D(\theta)$. With this method, hot air is blown over the surface of an initially uniformly wet soil column. Before the water content at the bottom of the sample starts decreasing, evaporation is stopped and the sample is cut into small layers. The water content of each layer is determined gravimetrically. The diffusivity $D(\theta) = K(\theta)/C(\theta)$, defined by the ratio of hydraulic conductivity, K, and the specific water capacity, C, at a water content, θ, can be calculated from

$$D(\theta_x) = \frac{1}{2t} \left.\frac{\mathrm{d}x}{\mathrm{d}\theta}\right|_x \int_{\theta_x}^{\theta_i} x \ \mathrm{d}\theta \qquad (6)$$

  where t = time [T], θ_i = initial volumetric water content [−], θ_x = volumetric water content [−] at distance x [L] from the evaporating surface. The method is known to be greatly affected by thermal effects.
- *Other Evaporation Methods*: Gabele and Hoch (1999) proposed to adapt the evaporation rate to the K-range of interest by either dividing the experiment into two or three sections, each starting with a different evaporation rate after a zero-flux period of several hours or continuous variation of the evaporation rate in response to measured state variables. Simunek *et al.* (1998) simultaneously estimated the van Genuchten parameters for $\theta(h)$ and *K(h)* from an evaporation experiment with parameter optimization techniques. Some uncertainty exists in the soil hydraulic parameters caused by high correlation between θ_r and n. Pressure heads measured close to the soil surface were found to be more valuable for the parameter estimation technique than those measured at lower locations. Tensiometer readings at one position are already sufficient to guarantee precise estimation of the soil hydraulic characteristics within the range of measurements (Simunek *et al.*, 1998). Extrapolation beyond this range involved a high level of uncertainty. Romano and Santini (1999) extended the analysis to compare various expressions for soil hydraulic properties. As for Wind's method, accuracy of hydraulic conductivity estimates is limited mainly due

to (i) near-zero hydraulic gradients close to saturation and (ii) extremely steep h-gradients at low conductivities.

- *Combined Outflow/Evaporation Method*: To reduce the duration of the evaporation experiment and to obtain the retention data in the low suction range, an outflow experiment can be conducted before the initiation of evaporation. After attaining near-equilibrium at negative pressure, the bottom boundary is sealed, the soil surface uncovered, and the evaporation rate at a controlled constant temperature measured. Fujimaki and Inoue (2003) determined $\theta(h)$ by curve-fitting of the equilibrium outflow and psychrometric data obtained from soil samples after evaporation, while the $K(\theta)$ was estimated inversely using cumulative evaporation amounts and the final water content profile.
- *Sorptivity Method*: This and the following methods combine laboratory experiments and parameter estimation for determining soil hydraulic properties in the wetting direction, rather than in the drying direction, as obtained through outflow methods. The sorptivity method (Dirksen, 1974) determines the soil water diffusivity function $D(\theta)$ by means of a series of one-dimensional absorption experiments in soil of constant initial water content θ_0. The sorptivity S [$\mathrm{L\,T^{-1/2}}$] is defined as proportionality constant between the cumulative amount of freely supplied water, I [L], infiltrating into an initially unsaturated horizontal soil column, and the square root of time, $S = I/\sqrt{t}$. Infiltration is controlled by mechanically supplying water to the absorption interface proportional to $t^{1/2}$. After supplying water for a certain time, the final water content of the top layer θ_1 is determined gravimetrically. The diffusivity can be calculated from

$$D(\theta_1) = \frac{\pi S^2}{4(\theta_1 - \theta_0)^2} \left[\frac{(\theta_1 - \theta_0)}{(1+\gamma)\log_e} \times \frac{\mathrm{d}}{\mathrm{d}\theta_1}(\log S^2) - \frac{1-\gamma}{1+\gamma} \right] \quad (7)$$

  where D = diffusivity [$\mathrm{L^2\ T^{-1}}$], θ = volumetric water content, S = sorptivity [$\mathrm{L\,T^{1/2}}$], and γ = weighting parameter [−]. For details, see Klute and Dirksen (1986) and Clothier and Scotter (2002).
- *Water Absorption into a Horizontal Column*: A horizontally placed soil column is connected to a water reservoir, causing absorption of water. The advance of the wetting front with time and the amount of water infiltrated are recorded. From these data, Shao and Horton (1998) estimated the van Genuchten parameters n and α by an approximate analytical solution. Additional information about K_s is needed. Wang *et al.* (2002) used analytical solutions of horizontal infiltration to estimate Brooks and Corey parameters. Input data needed are infiltration rate, cumulative infiltration, and distance of the wetting front with infiltration time.
- *Upward Infiltration Method*: In the upward infiltration method, UIM (Hudson *et al.*, 1996), a constant flux of water is imposed at the bottom of the soil sample, and pressure heads are measured inside the sample using tensiometers. To maximize information for the inverse analysis, Young *et al.* (2002) suggested to initiate infiltration by a certain tension at the bottom of the sample, using a Mariotte system, rather than by imposing a boundary flux. For tension infiltration, the soil controls the total amount of water being taken up, thus providing additional information and control for the numerical inversion. The method is most useful in the wet range of the retention and conductivity functions. Simunek *et al.* (2000) suggested a method based on modification of the UIM to capture nonequilibrium flow parameters. The observed nonequilibrium behavior could be described with the dual-porosity and dual-permeability models, but a unique set of soil hydraulic parameters could not be identified.
- *Downward Infiltration Method*: Water is supplied at the top of a soil sample at a prescribed head (usually ponding) or at a prescribed rate. Van Genuchten parameters can be estimated by numerical inversion of Richards' equation. Measurements needed for the inversion method include: soil water tension versus time at one distance from the soil surface, the initial water content, and a final steady-state water content behind the wetting front. The objective function used for parameter optimization is constructed from two parts, one from the transient tension versus time curve, and another from the steady-state water content data (Zou *et al.*, 2001).

Field Methods

Laboratory experiments have the advantage of being comparatively quick and precise, but they often lead to soil-physical properties that are not representative for the field. Direct *in situ* measurement of hydraulic and retention properties still provides perhaps the most reliable, and often, the only means for determining hydraulic properties, despite their high cost and extreme time demands (Tseng and Jury, 1993). Two distinct types of field procedures are in common use: internal drainage (ID) flux methods (Section "Internal drainage method") and infiltration methods. Infiltration methods can be further distinguished in steady water application flux methods (Section "Steady flow infiltration methods"), ponded infiltration methods (Section "Pressure ring infiltration"), tension infiltrometry (Section "Tension disc infiltration") and infiltration from wells or bore holes (Section "Infiltration from wells and bore holes").

Field methods have the advantage of dealing with soil in natural conditions. However, small-scale heterogeneity in

soil conditions may introduce large variation in measured values (Munoz-Carpena *et al.*, 2002). Since most field measurements are confined to produce a single measurement at a single field location, adequate evaluation of field hydraulic and chemical transport properties requires a large number of measurements.

Errors in the evaluation of unsaturated hydraulic conductivity under field conditions may arise from a number of sources. In the instantaneous profile experiment, measurements of pressure head and water content are uncertain owing to spatial variability, which leads to significant uncertainty by error propagation (Flühler *et al.*, 1976). Tension disc and ring infiltrometer also have limitations; most of them being associated with the simplifying assumptions of the analysis used to infer soil hydraulic properties from water and solute flow measurements. Infiltration rates during the measurement time can be variable, both increasing and decreasing (Logsdon, 1997). Factors affecting temporal variability of hydraulic properties during infiltration experiments are hydrophobicity and swelling (Roth *et al.*, 1999; Angulo-Jaramillo *et al.*, 2000). These limitations lead in general to uncertainty, whether or not there will be major advances in *in situ* determination of *K(h)*. In an extensive review on published field measurement data, Jury (1985) found extremely high variation of hydraulic conductivity parameters, which he attributed partly as apparent variability, caused by fitting oversimplified functions to the data. Tseng and Jury (1993) point out that since the true physical state of a soil can never be known in any field study, absolute comparisons of properties estimated from measurements, and thus actually occurring in the field, can never be made. Hence, method comparisons must always be relative, and absolute statements upon the accuracy of any method are not possible.

Internal Drainage Method

The internal drainage method, ID, or instantaneous profile method is regarded as a reference method to measure *in situ* unsaturated hydraulic properties for both homogeneous and layered soils (Hillel *et al.*, 1972). The method was first suggested by Richards *et al.* (1956) on the basis of simultaneous monitoring of water retention and flux in laboratory soil columns. Later, the method was adapted for practical field use (Watson, 1966). Simplified variants have been developed which require fewer *in situ* measurements but depend on flow approximations (Nielsen *et al.*, 1973; Ahuja *et al.*, 1988). Their use is mainly limited by high demands on equipment and time.

In the ID, large rings (about 2 m in diameter) are inserted into soil with a certain height left above ground. Within each ring, tensiometers and TDR are installed via an access tube. Water is ponded on the soil surface until the soil is wetted beyond the maximum depth for which the determinations are desired. The groundwater table should be sufficiently below this depth to obtain maximum possible unsaturated drainage. To minimize evaporation, temperature fluctuation, and protect the surface from rainfall during subsequent drainage, the soil surface inside the rings is covered. Both θ and h are monitored for weeks. Measurements are taken with decreasing frequency. Evaluation is performed using instantaneous profile data analysis, which is based on the Darcian analysis of transient soil water content and hydraulic head profiles (Watson, 1966). Zhang *et al.* (2003) presented an improved analysis of the data from drainage experiments using inverse modeling, which uses nonlinear regression methods to estimate hydraulic parameters. All van Genuchten hydraulic parameters could be estimated uniquely when both water content and pressure head data were used.

Errors in K values obtained during the early stages of drainage are primarily due to errors in determining the hydraulic gradient, while at later times, errors in water content measurement are more serious. Using typical measurement errors, the calculated uncertainty of K in the moist range is in the order of 40%, and in the dry range up to more than 100% (Flühler *et al.*, 1976). The failure of even one tensiometer, which is not uncommon in the field, will significantly affect the calculation of hydraulic properties. Because θ and h determination is needed over a long period of time, the ID is time- and equipment-intensive, and thus costly, especially if several sites must be monitored to estimate spatial variability.

Steady Flow Infiltration Methods

Infiltration rates effectively integrate properties of the porous media underneath the infiltrometer, including the influence of local-scale heterogeneity, different soil structure, and textural irregularities, preferential pathways, layering, and anisotropy. Hence, infiltration rates provide a good way for estimating the effective near-saturated soil hydraulic properties (Mertens *et al.*, 2002). Steady flux is achieved either by uniform application of water by sprinkler or trickler, or by creating a low permeability crust at the soil surface (Hillel and Gardner, 1970). Measurements taken under steady-state flow conditions are easier and less prone to errors (porous cup equilibration of tensiometers) than those taken under transient conditions. It is also possible to repeat pressure head, water content, and flow rate measurements at steady-state, whereas for transient conditions, the entire experiment must be repeated. However, considerable effort and time are required for the steady flux method to be of practical use in the field. The crust method, while accurate, cannot be used when the soil is wet and the method can be very costly to run, since up to 100 days may be required to complete the measurement of *K(h)* over the required potential range. In all steady flow methods only wetting is considered.

Recent innovations in steady flow infiltration methods include inverse estimation of soil hydraulic parameters in

combination with multipurpose probes that couple TDR and tensiometry used under constant flux infiltration (Gee and Ward, 1999). Si and Kachanoski (2000) used multipurpose TDR probes coupled with a series of constant-rate infiltration experiments to estimate the effective one-dimensional field-average hydraulic properties $K(\theta)$ and $\theta(h)$. The TDR probes are installed vertically to measure the rate of change of local soil water storage (q) along the probe during constant-rate water application. The values of q are equal to local soil water flux, and assuming unit gradient, are set equal to K at the steady-state θ and h measured at long times. The measured values of K, θ and h from different water application rates are combined to obtain average hydraulic functions.

- *Line Source Method*: Zhang *et al.* (2000a,b) estimated hydraulic properties by means of multipurpose TDR probes and existing quasi-analytical, steady-state solutions for infiltration from a surface line source. They installed 50 nests of multipurpose TDR probes with a between-nest spacing of 0.15 m, indicating the high costs connected with this method. Inverse procedures are used to estimate the inverse macroscopic capillary length scale, α, and K_s, from h, θ, and conservative ionic tracer travel time (T). Measurements of only h or only T will not give unique estimates of α and K_s. Combining measurements of θ with h and/or T gives unique estimates of α and K_s.
- *Point Source Method*: Al-Jabri *et al.* (2002) used a much simpler approach: The hydraulic properties were determined by applying three discharge rates higher than the infiltration rate. This produces circular saturated areas at the soil surface beneath each emitter that reach constant size depending on the irrigation rate. The radius, r, of the areas is measured. The method required only two days to collect data. Once steady-state occurs, Wooding's (1986) solution can be applied. K_s is yielded graphically as the intercept of a plot of discharge rate versus $1/r$. The mean K_s found were larger than with tension discs. Error sources are different radii for the same infiltration rate, and noncircular areas. The method does not offer a high level of control, but is very cheap. A single supply tube can allow for application of the point source method to multiple locations. Evaluation of the setup and procedure with natural field conditions, such as tillage practices is required for further establishment.

Pressure Ring Infiltration

Ring infiltrometers are probably the most widely used device for measuring field infiltration rates (Wu *et al.*, 1997). The technique is useful in the estimation of the *in situ* field-saturated hydraulic conductivity, K_{fs}, and matrix flux potential Φ_m. Water is supplied to the soil surface at a positive pressure head h_0 either by a Mariotte bottle allowing a wide range of h_0, or by a small capillary tube also acting as a measuring burette. Water can be infiltrated at a constant or at a continuous falling head. A double ring infiltrometer can be used to minimize lateral flow during the experiment. Infiltration from the outer ring before and during the measurement should guarantee one-dimensional downward flow from the inner ring. The analysis is performed only with infiltration from the inner ring, and is essentially the same as for single ring infiltrometers.

Various techniques based on either transient or steady-state water flow approaches have been used to infer soil hydraulic properties from ponded ring infiltration tests (Angulo-Jaramillo *et al.*, 2000). Basically the flow from a ring infiltrometer set at a positive pressure head h_0 is controlled by K_{fs} which accounts for the gravity effect and by Φ_m, the matrix flux potential for saturated conditions, which is defined as

$$\Phi_m = \int_{h_u}^{0} K(u)\,\mathrm{d}u \tag{8}$$

Constant head conditions have traditionally been used, because constant head devices are easy to maintain experimentally and because the analysis is relatively simple (Mertens *et al.*, 2002). It is possible to show (Elrick and Reynolds, 1992) that the steady-state flow rate out of a ring infiltrometer is given by

$$q_{0\infty} = K_{fs}\left(1 + \frac{H}{\pi r_d G}\right) + \frac{\Phi_m}{\pi r_d G}, \tag{9}$$

where G is a shape parameter, and r_d is the radius of the ring. Applying successively to the same ring two positive hydraulic heads H_1 and H_2 allows solving simultaneously the resulting two equations for K_{fs} and Φ_m. In highly heterogeneous and low permeable soils, the method can cause a large percentage of invalid (i.e. negative) and unrealistic K_{fs} and Φ_m. Here the one-dimensional steady-state saturated flow analysis can be replaced by a transient analysis using nonlinear least-squares inversion procedures.

Another possibility to analyze infiltration experiments is to use falling head methods, where a ring is inserted a known distance into the soil, and attached to a mariotte reservoir. Infiltration from the ring is allowed to come to steady-state under a constant ponded head, after which the head is allowed to fall and the head is measured as a function of time (Parkin *et al.*, 1999). A numerical inversion procedure is used to calculate the soil hydraulic properties. Measurements of early-time infiltration under both constant and falling head can better characterize the hydraulic properties of low permeability media.

Advances have been made in using ring infiltrometers and TDR to measure hydraulic properties in unsaturated soils (Parkin *et al.*, 1995). Estimates for hydraulic conductivity can be achieved for heads greater than −60 cm, which

is the range with most effect on ponded infiltration. However, generally ponded infiltration measurements are not sensitive for estimating unsaturated soil hydraulic conductivity (Bagarello *et al.*, 2000).

Tension Disc Infiltration

A tension disc infiltrometer (TI) is a constant head infiltrometer that can operate at either a positive or a negative head and thus it can be used for determination of both saturated and unsaturated hydraulic properties. TI have become very popular devices for the *in situ* estimation of soil surface hydraulic properties (Angulo-Jaramillo *et al.*, 2000). Because of their portability, disc and ring infiltrometers are useful instruments to investigate statistical distributions of hydraulic conductivities. While ponded infiltration is used to determine the saturated hydraulic conductivity, tension infiltrometry also provides an opportunity to estimate unsaturated hydraulic properties. TI associated with conservative tracers was also used for inferring parameters describing the water-borne transport of chemicals and other parameters such as mobile/immobile water content fraction or exchange coefficient (Clothier *et al.*, 1992; Jaynes *et al.*, 1995). A number of challenges still remain unresolved for both theory and practice for tension disc infiltrometers. They include questions on how to consider and characterize saturated-unsaturated preferential flow or preferential transport processes, and the problem of contact resistance between the tension disc membrane and the soil.

During tension disc infiltration, a reservoir tower provides the water supply, and a bubble tower with a moveable air-entry tube imposes the pressure head h_o of the water at the cloth base. For measurements under tension, intimate hydraulic contact between the soil surface and the source of water is essential. This is generally achieved by pouring at the surface a layer of sand that should be made as thin as possible. The contact sand can introduce flow impedance effects at the high infiltration rates associated with ponded conditions. The same is true for the porous membrane of the infiltrometer (Mohanty *et al.*, 1998). For every imposed value h_o, the cumulative infiltration, $I(t)$, is recorded either by noting the water level drop in the reservoir tower or by using pressure transducers (e.g. Ankeny *et al.*, 1988). Additionally, initial and final water content are determined. Water flow from a tension infiltrometer disk to the underlying soil follows a three-dimensional flow process. Various techniques have been developed to infer hydraulic properties from measurements of either transient flow rates, or steady ones, that emanate from a disc (Wang *et al.*, 1998). The methods of analysis of cumulative infiltration are based either on quasi-analytical solutions or on inverse parameter estimation techniques, depending on whether the soil profile is homogeneous or not (Angulo-Jaramillo *et al.*, 2000). Vandervaere *et al.* (2000) found for 7 largely different soils that the lateral capillary flow (S^2) dominates the gravity flow (K) during a disk infiltrometer experiment. Thus hydraulic conductivity must be considered as a variable playing a minor role, and hence difficult to determine accurately. Although performing infiltration experiments at the same location for several tensions can affect infiltration rates, TI causes relatively little disturbance of the soil macrostructure. Still the tension disc method yielded lower K_s values under high-permeability conditions relative to other methods, because the ring may be too small to adequately sample cracking clay loam soil (Reynolds *et al.*, 2000).

- *Steady-State Analytical Solutions* Wooding's equation (1986) equation is at the heart of most of these analyses. It approximates the steady infiltration rate, $q_{0\infty}$ from a disc as

$$q_{0\infty} = K_0 + \frac{4\Phi_0}{\pi r_d} \tag{10}$$

where Φ_0 is the matrix flux potential for unsaturated conditions defined by

$$\Phi_0 = \int_{\theta_n}^{\theta_0} D(u)\,du = \int_{h_n}^{h_0} K(u)\,du, \theta_0 \geq \theta_n, h_n \leq h_0 \tag{11}$$

In these equations, K_0 is the hydraulic conductivity at the imposed pressure head h_0, and $D(u)$ is the capillary diffusivity [L^2T^{-1}]. The subscript 0 refers to the condition imposed at the supply surface of the disc, and the subscript n denotes the antecedent condition prevailing in the soil before the infiltration takes place. Early-time infiltration data can be used to estimate the sorptivity (White and Sully, 1987) and, consequently, the matrix flux potential. Equation (10) can be solved for K_0 and Φ_0 by using either multiple radii (MR)(at the same value h_0) or multiple head (for a given disc radius) procedures. Drawback of the MR method lies in the fact that the two (or more) disk experiments must be performed at different locations, which introduces complications from the short-distance spatial variability of soil properties (Wang *et al.*, 1998).

Large variations or discrepancies are reported in the literature when comparing Wooding's method results with other methods. One possible source of error is its limitations for relatively small disk sizes. Several studies have shown that Wooding's approach will overestimate the soil hydraulic conductivity if steady-state infiltration is not reached. Discrepancies between tension infiltrometer and other methods in practice, however, are caused probably more by variability within each method such as soil heterogeneity or simplification of the hydraulic conductivity function to an exponential expression, than by inherent limitations of the steady-state solutions (Wang *et al.*, 1998).

- *Transient State: Quasi-analytical Solutions* Restrictions in the use of Wooding's equation, together with the fact that much useful information is lost by ignoring the

transient stage and large time savings can be achieved (Logsdon, 1997) have strengthened the need for a transient three-directional infiltration equation for disc infiltrometers. The most recent expressions (Warrick, 1992; Haverkamp *et al.*, 1994; Zhang, 1998) have in common the following two-term form of the cumulative infiltration equation:

$$I(t) = C_1\sqrt{t} + C_2 t \tag{12}$$

but they differ by the expressions of the coefficients C_1 and C_2. Haverkamp *et al.* (1994), for example, posed the following physically based expressions valid for times not approaching steady-state:

$$C_1 = S_0$$

$$C_2 = K_n + \frac{1}{3}(2-\beta)(K_0 - K_n) + \frac{\gamma}{r_d(\theta_0 - \theta_n)}S_0{}^2 \tag{13}$$

where β is a parameter depending on the capillary diffusivity function. It lies in the interval [0, 1], and γ is a constant approximately equal to 0.75.

Vandervaere *et al.* (2000) proposed several new methods for the analysis of tension disk infiltrometer tests and the determination of sorptivity and hydraulic conductivity. The single test method uses one disk radius and one value of pressure head. The multiradii method uses two or more disk radii and one value of pressure head. The multiple sorptivity methods use one or multiple disk radii and two or more values of pressure head. Method appropriateness was shown to be highly dependent on the ratio of K and S. A good experimental strategy consists of choosing the method of analysis on the basis of the dominant flow: vertical capillary flow, vertical gravity flow, or lateral capillary flow.

- *Transient State: Inverse Estimation* An alternative to direct estimates of soil hydraulic properties is the use of inverse methods when *in situ* conditions differ strongly from assumptions required for the use of semianalytical solutions of the flow equation, that is, nonuniform water content distribution, and multilayered soils. From an analysis of numerically generated data, Simunek and van Genuchten (1996) concluded that the cumulative infiltration curve by itself does not contain enough information to provide a unique inverse solution. Schwartz and Evett (2002) reported that inverse optimizations, which included in the objective function both, water retention and cumulative infiltration, led to excellent fits of this data for high initial volumetric water content. Simunek and van Genuchten (1997) put forward that the best practical scenario is to estimate the hydraulic parameters from the cumulative infiltration curve measured at several consecutive tensions applied to the soil surface, in conjunction with the knowledge of the initial and final water content.

Infiltration from Wells and Bore Holes

Tension disc and ring infiltrometers are constrained to the near-surface and are easily impacted by microtopography. These limitations can be partly overcome by infiltration from bore holes with or without cone penetrometers. Combinations of TDR and permeameters with cone penetrometer technology show promise for minimally intrusive, cost-effective measurements of $K(h)$ and $\theta(h)$ to greater depths (Gee and Ward, 1999).

Guelph Permeameter(GP) The Guelph Permeameter (GP) is a constant head well permeameter consisting of a mariotte bottle that maintains a constant water level inside a hole augered into unsaturated soil. Flow from this permeameter is assumed to reach steady-state after a transient state during which the soil saturated bulb and the wetting zone increase in size by migrating quasi-spherically from the infiltration surface. At steady-state, the saturated bulb remains essentially constant in size, while the wetting front continues to increase (Reynolds *et al.*, 1985). The analysis requires measurements from two different water levels in the same well. Munoz-Carpena *et al.* (2002) provided methods for inverse optimization for parameter identification. They found that GP gives lower K_s estimates compared to other methods. A multiplying factor of 2 to 3 has been proposed to account for air entrapment in the field soil.

Cone Permeameter The cone permeameter is fitted into a hole, where a soil core of smaller diameter was removed with a sampler. Water is injected into the soil through a screen, and the progress of the wetting front is measured with two tensiometer rings positioned above the screen. Tests are conducted with two sequentially applied pressure heads of different magnitudes. After the supply valve is closed, tensiometers monitor the redistribution of water in the soil profile. It is possible to evaluate wetting only, or wetting and drying simultaneously, via analysis of different parts of the experiment, that is, during water application and during redistribution. Cumulative inflow and pressure head readings are analyzed with inverse optimization. Final moisture content information improves estimates of unknown hydraulic parameters (Gribb, 1996). Soil densification caused by pushing the cone permeameter to the testing depth resulted in lower values of θ_s and K_s (Kodesova *et al.*, 1999).

Use of Tensiometer Response to Measure Soil Hydraulic Conductivity

The use of a ceramic cup tensiometer to measure soil hydraulic conductivity was proposed as early as 1937 by Richards *et al.* (1937). Recently, Hayashi *et al.* (1997) developed a method to determine the hydraulic conductivity of unsaturated soils from the response of a tensiometer to an artificially induced perturbation of pressure. The response

time of tensiometers is controlled by soil hydraulic conductivity when the flow resistance of the porous cup is sufficiently small. Advantages are the fast procedure (1 h) and the use of readily available equipment. Tests must be repeated over time to determine the *in situ* relationship between matric potential and hydraulic conductivity. The estimated conductivity is sensitive to the disturbance of the soil in close proximity to the porous cup. It is also sensitive to changes in the matric potential during a response test. Timlin and Pachepsky (1998) also developed a method to determine unsaturated hydraulic conductivity using tensiometers. They measured water flux into a tensiometer after a suction is applied to its inside. The reduced pressure in the tensiometer causes water to flow into the tensiometer from the soil leading to a decrease of the volume of air, and subsequently pressure increases. The unsaturated conductivity parameters are determined by using a two-dimensional finite element soil model coupled with inverse optimization. Shortcomings of the method are clogging of the tensiometer pores by fine soil material, which leads to a change of conductivity of the ceramic cup over time, and possible poor contact between tensiometer and soil.

Integrated Determination by Inverse Modeling

By the very nature of the unsaturated flow processes, the classic direct measurements of hydraulic properties remain extremely demanding with respect to experimental procedures, requiring, for example, excessive equilibrium times, small sample sizes, strictly stationary flow or pressure conditions, and perfect contact between porous membranes and measurement sensors. At the same time, they generally yield only limited information about the hydraulic properties, for example, one point on a moisture retention curve. In order to come to a faster and more integrated way to characterize hydraulic properties of soils, inverse simulation of transient flow processes on the laboratory and on the field scale will play the key role in future developments of measuring techniques. The highly dynamic and strongly nonlinear water flow in the unsaturated zone is clearly too difficult to represent in analytical or semiempirical approaches (Kutilek and Nielsen, 1994; Feddes, 1995). With the progress in computer simulation, an intensive further development of transient methods that aim on the simultaneous determination of the hydraulic properties over a wide moisture range by inverse simulation is currently underway (Vrugt and Dane, 2005, **Chapter 77, Inverse Modeling of Soil Hydraulic Properties, Volume 2**). Specifically, the inclusion of the measurement instrument itself and its interaction with the surrounding soil environment in the inverse analysis bears a tremendous potential to improve the precision of hydraulic properties estimates. Furthermore, developments in combined instrumentation, where controlled or uncontrolled water fluxes are measured together with water content and water potential simultaneously by the same instrument, will become more important.

Still, in the attempt to estimate soil hydraulic properties it appears that we are faced with an analogue to the Heisenberg uncertainty principle: The more precise the hydraulic properties are obtained for a point in space or an isolated soil sample, the less relevant this precision might be for characterizing the flux at the field scale. In the range where the process model is valid, we can now achieve an almost perfect history matching by inverse modeling of laboratory experiments, which allows characterizing the hydraulic properties of soil samples to an exceptional high degree of accuracy (Bitterlich *et al.*, 2004). This kind of measurement evaluation further allows to expand the limits of our conceptual understanding of the flow process (e.g. two-phase flow effects, hysteresis model inadequacies, dynamic effects). However, flow and transport processes in vadose zone hydrology may be controlled by structures, which have an extent similar to the scale of the system of interest (Vogel and Roth, 2003). This means that without identifying these structures, our efforts on local precision are futile. As hydraulic properties cannot be directly measured on larger scales, we depend on proxy variables. If structure is identified at the scale of interest, it might be sufficient to determine the hydraulic properties of the structural elements in an approximate manner. Stochastic fusion of a variety of moisture-related instrumental information by noninvasive sensoring methods (electrical resistivity tomography, passive microwave, ground-penetrating radar, and others) and its inverse evaluation (Yeh and Simunek, 2001) appears to be a promising direction.

While keeping in mind that the general interest clearly focuses on effective large-scale properties, it must be stressed that only a thorough understanding of the underlying point-scale processes allows to judge on appropriate ways to upscale local properties. For processes of dissipative nature, effective macroscale properties obtained just by averaging will be a valid description (e.g. Schmalz *et al.*, 2003). However, other processes must be known (and measured) in spatial and/or temporal detail, because neglecting them will lead to systematic misconceptions. For example, small changes in water content in the pressure range near saturation, as frequently encountered in natural soil samples, indicate a secondary pore structure, which controls preferential flow of water on the scale of interest. Neglecting this by fitting oversimplified models for the constitutive relationships will lead to a systematic misconception and systematic errors in modeling water and solute transport (Mahmood and Hubbard, 2003).

It appears that in the development of measurement techniques for hydraulic properties characterization there is currently a threefold transition:

- There is a transition from "classic" methods, which depend on equilibrium, steady-state flow, or strict boundary conditions, toward generalized nonequilibrium methods. This represents a change in paradigm from measurements that are based on high experimental and low calculus requirements towards measurements with great freedom in experimental boundary conditions, but high requirements with respect to sensor response (in particular, with respect to temporal resolution and signal accuracy) and with respect to a sophisticated and robust model inversion.
- There is also a transition from point measurement to large-scale measurements, based on a combination of new hydrogeophysical techniques that are applicable on large areas such as satellite-based ground-penetrating radar. The focus must be to identify the largest structures in the materials on the respective scale of interest because these determine the system behavior. Together with flow observations on large areas (drained land) and regional estimates of evapotranspiration, these observations can be evaluated by inverse modeling in order to search for the existence of effective processes and properties on the scale of interest (Feddes *et al.*, 1993).
- Finally, there is an increasing recognition that due to the overwhelming complexity of natural flow processes, the idea to measure "true" and "universal" hydraulic properties on a large scale must be replaced by object-dependent measurement methods and properties, where the target variable of interest (e.g. irrigation demand, mean flow to the groundwater, structure of flow field for solute transport) decides upon the most appropriate manner to evaluate the data (Abbaspour *et al.*, 2001).

CONCLUDING REMARKS

Our knowledge of basic vadose zone flow and transport processes has advanced considerably over the last several decades. Unfortunately, the ability to determine process parameters has not kept pace with the ability to put the processes into numeric models. Although various field and laboratory methods for determining unsaturated hydraulic properties have been proposed, there remains a glaring lack of standardized procedures. Despite enormous investments of time and money made by soil scientists, hydrologists, and others, improvements over several decades in direct methods to measure hydraulic properties of field soils have been rather marginally achieved. Specialized equipment to measure the unsaturated hydraulic properties is generally expensive because of a relatively small market. Therefore, a disproportionate amount of time and money is still being used to conduct routine experimental work. The difficulty to solve Richards' equation for transient flux conditions hindered for a long time the rigorous evaluation of experiments and the development of experiments that are designed to maximize the information content with respect to the properties of interest (Vrugt and Bouten, 2001). Furthermore, not much of the knowledge gained during the last decade on measurement methods is being used in management practices. Thus, narrowing the gap between the state of the art and the state of the practice is one of the fundamental issues required in order to solve problems such as contamination of the vadose zone and groundwater.

On the positive side, experimental procedures are now greatly benefiting from the use of flexible inverse estimation procedures, improved numerical models, more powerful computers, and increased automation in data collection efforts. For many years, measuring and monitoring methods have lagged behind numerical analyses of water flow and solute transport through the vadose zone. However, recent advances in electronic components, renewed interest in development of monitoring methods, and the infusion of geophysical methods into vadose zone hydrology have begun to address this imbalance. Future developments in instrumental field techniques and in improved inverse modeling techniques must direct toward a point where available tools will be so well adapted to the needs of soil physicists and vadose zone hydrologists, that they are also easy to use in practice. Essential in that is that inverse modeling techniques must not only be robust and thus lead to accurate results, but also must obligatorily indicate the adequacy of the underlying model assumptions and the uncertainty of the resulting estimates. Because of the increasing speed of computers, there is a rapid development in this area (e.g. Vrugt *et al.*, 2003), and the use of Monte Carlo based stochastic methods to estimate the uncertainties, appears very promising.

Probably the most challenging task with respect to determining hydraulic properties of soils is the coupling of soil mechanics, soil chemistry, and possibly soil biology, with soil hydraulics. Since there is an enormous sensitivity of parameters, such as the saturated hydraulic conductivity, on these factors, this is not merely an academic question. Despite the long-lasting knowledge upon the importance of the various feedbacks (e.g. Nielsen *et al.*, 1986), not much progress has been achieved in this field during the last decades. Substantial progress will depend on the ability of scientists to bridge the traditional disciplinary boundaries (US Department of Energy, 2001). To date, we still suffer from the *'inability to integrate simultaneously the most relevant physical, chemical, and biological processes in a unified theoretical framework'* (Nielsen *et al.*, 1986).

REFERENCES

Abbaspour K.C., Schulin R. and van Genuchten M.Th. (2001) Estimating unsaturated soil hydraulic parameters using ant colony optimization. *Advances in Water Resources*, **24**, 827–841.

Ahuja L.R., Fiedler F., Dunn G.H., Benjamin J.G. and Garrison A. (1998) Changes in soil water retention curves due to tillage and natural reconsolidation. *Soil Science Society of America Journal*, **62**, 1228–1233.

Ahuja L.R., Ross J.D., Bruce R.R. and Cassel D.K. (1988) Determining unsaturated hydraulic conductivity from tensiometric data alone. *Soil Science Society of America Journal*, **52**, 27–34.

Alexander L. and Skaggs R.W. (1984) Predicting unsaturated hydraulic conductivity from soil texture. *Journal of Irrigation and Drainage Engineering-Asce*, **113**, 184–197.

Al-Jabri S.A., Horton R. and Jaynes D.B. (2002) A point source method for rapid simultaneous estimation of soil hydraulic and chemical transport properties. *Soil Science Society of America Journal*, **66**, 12–18.

Angulo-Jaramillo R., Vandervaere J.P., Roulier S., Thony J.L., Gaudet J.P. and Vauclin M. (2000) Field measurement of soil surface hydraulic properties by disc and ring infiltrometers- a review and recent developments. *Soil and Tillage Research*, **55**, 1–29.

Ankeny M.D., Kaspar T.C. and Horton R. (1988) Design for an automated infiltrometer. *Soil Science Society of America Journal*, **52**, 893–896.

Arya L.M., Farrel D.A. and Blake G.R. (1975) A field study of water depletion patterns in presence of growing soybean roots. I. Determination of hydraulic properties of the soil. *Soil Science Society of America Journal*, **39**, 424–430.

Bagarello V., Iovino M. and Tusa G. (2000) Factors affecting measurement of the near-saturated soil hydraulic conductivity. *Soil Science Society of America Journal*, **64**, 1203–1210.

Bertuzzi P., Mohrath D., Bruckler L., Gaudu J.C. and Bourlet M. (1999) Wind's evaporation method: experimental equipment and error analysis. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 323–328.

Bittelli M., Flury M. and Campbell G.S. (2003) A thermo-dielectric analyser to measure the freezing and moisture characteristic of porous media. *Water Resources Research*, **39**, 1041–1050.

Bitterlich S., Durner W., Iden S.C. and Knabner P. (2004) Inverse estimation of the unsaturated soil hydraulic properties from column outflow experiments using free-form parameterizations. *Vadose Zone Journal*, **3**, 971–981.

Blunt M.E. and Hilpert M. (2001) Pore scale modeling. *Advances in Water Resources*, **24**, 231–478.

Bohne K., Roth C., Leij F.J. and van Genuchten M.Th. (1993) Rapid method for estimating the unsaturated hydraulic conductivity from infiltration measurements. *Soil Science*, **155**, 237–244.

Bouma J., Belmans C., Dekker L.W. and Jeurissen W.J.M. (1983) Assessing the suitability of soils with macropores for subsurface liquid waste disposal. *Journal of Environmental Quality*, **12**, 305–311.

Brooks R.H. and Corey A.T. (1964) *Hydraulic Properties of Porous Media*, Hydrology Paper Vol. 3, Colorado State University: Fort Collins, pp. 1–27.

Brown J. and Heuvelink G.B.M. (2005) Assessing uncertainty propagation through physically-based models of soil water flow and solute transport. *Encyclopaedia of Hydrological Sciences*, John Wiley & Sons.

Burdine N.T. (1953) Relative permeability calculation size distribution data. *Transactions of the American Institute of Mining, Metallurgical and Petroleum Engineers*, **198**, 71–78.

Butters G.L. and Duchateau P. (2002) Continuous flow method for rapid measurement of soil hydraulic properties: I. Experimental considerations. *Vadose Zone Journal*, **1**, 239–251.

Celia M.A., Reeves P.C. and Ferrand L.A. (1995) Recent advances in pore-scale models for multiphase flow in porous media, U.S. Nat. Rep. Int. Union Geophys. 1991–1994. *Reviews of Geophysics*, **33**, 1049–1057.

Childs E.C. and Collis-George G.N. (1950) The permeability of porous materials. *Proceedings of the Royal Society of London. Series A*, **201**, 392–405.

Clothier B.E., Kirkham M.B. and McLean J.E. (1992) In situ measurement of the effective transport volume for solute moving through soil. *Soil Science Society of America Journal*, **56**, 733–746.

Clothier B. and Scotter D. (2002) Unsaturated water transmission parameters obtained from infiltration. In *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society America: Madison, pp. 879–898.

Crescimanno G. and Iovino M. (1995) Parameter estimation by inverse method based on one-step and multi-step outflow experiments. *Geoderma*, **68**, 257–277.

Dane J.H. and Topp G.C. (2002) *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series No. 5*, ISBN 0-89118-810-X, Soil Science Society of America: Madison, p. 1692.

Dirksen C. (1974) Field test of soil water flux meters. *Transaction of ASAE*, **17**, 1038–1942.

Dirksen C. (1991) Unsaturated hydraulic conductivity. In *Soil Analysis Physical Methods*, Smith K.A. and Mullins C.E. (Eds.), Marcel Dekker: New York, pp. 209–270.

Dirksen C. (1999a) *Soil Physics Measurements*, Catena Verlag: Reiskirchen.

Dirksen C. (1999b) Direct conductivity measurements for evaluating approximate and indirect determinations. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., *et al.* (Eds.), University of California, Riverside, pp. 271–278.

Dirksen C. and Matula S. (1994) Automatic atomized water spray assembly for soil hydraulic conductivity measurements. *Soil Science Society of America Journal*, **58**, 319–325.

Doering E.J. (1965) Soil water diffusivity by the one-step method. *Soil Science*, **99**, 322–326.

Durner W. (1994) Hydraulic conductivity estimation for soils with heterogeneous pore structure. *Water Resources Research*, **30**, 211–223.

Durner W. and Flühler H. (2005) Soil hydraulic properties, *Encyclopaedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Durner W. and Or D. (2005) Soil water potential measurement, *Encyclopaedia of Hydrological Sciences*, This issue, John Wiley & Sons.

Durner W., Priesack E., Vogel H.-J. and Zurmühl T. (1999a) Determination of parameters for flexible hydraulic functions by inverse modeling. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 817–829.

Durner W., Schultze B. and Zurmühl T. (1999b) State-of-the-art in inverse modeling of inflow/outflow experiments. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 661–681.

Eching S.O. and Hopmans J.W. (1993) Optimization of hydraulic functions from transient outflow and soil water pressure data. *Soil Science Society of America Journal*, **57**, 1167–1175.

Eching S.O., Hopmans J.W. and Wendroth O. (1994) Unsaturated hydraulic conductivity from transient multistep outflow and soil water pressure data. *Soil Science Society of America Journal*, **58**, 687–695.

Elrick D.E. and Reynolds W.D. (1992) Methods for analyzing constant-head well permeameter data. *Soil Science Society of America Journal*, **56**, 320–321.

Fatt I. (1956) The network model of porous media. *Transaction of the American Institute of Mining Metall. Pet. Engineering*, **207**, 144–181.

Feddes R.A. (1995) *Space and Time Scale Variability and Interdependencies in Hydrological Processes*, Cambridge University Press: Cambridge.

Feddes R.A., Menenti M., Kabat P. and Bastiaanssen W.G.M. (1993) Is large-scale inverse modelling of unsaturated flow with areal average evaporation and surface soil moisture as estimated from remote sensing feasible? *Journal of Hydrology*, **143**, 125–152.

Flühler H., Ardakani M.S. and Stolzy L.H. (1976) Error propagation in determining hydraulic conductivities from successive water content and pressure head profiles. *Soil Science Society of America Journal*, **40**, 830–836.

Flühler H., and Roth K. (2004) *Physik der ungesättigten Zone*, Lecture notes, Institute of Terrestrial Ecology, Swiss Federal Institute of Technology, Zurich and Institute of Environmental Physics, University of Heidelberg.

Fujimaki H. and Inoue M. (2003) A flux-controlled steady-state evaporation method for determining unsaturated hydraulic conductivity at low matric pressure head values. *Soil Science*, **168**, 385–395.

Gabele T. and Hoch R. (1999) Soil-dependent flux boundary conditions for wind's evaporation method. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 313–322.

Gardner W.R. (1958) Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science*, **85**, 228–232.

Gardner W.R. (1962) Note on the separation and solution of diffusion type equations. *Soil Science Society of America Proceedings*, **26**, 404.

Gardner W.H. (1986) Early soil physics into the mid-20th century. *Advances in Soil Science*, **4**, 1–101.

Gee G.W. and Ward A. (1999) Innovations in two-phase measurements of hydraulic properties. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., *et al.* (Eds.), University of California, Riverside, pp. 241–270.

Gee G.W., Ward A.L., Zhang Z.F., Campbell G.S. and Mathison J. (2002) The influence of hydraulic nonequilibrium on pressure plate data. *Vadose Zone Journal*, **1**, 172–178.

Ghezzehei T.A. and Or D. (2003) Stress-induced volume reduction of isolated pores in wet soil. *Water Resources Research*, **39**(3), 1067, doi:10.1029/2001WR001137.

Globus A.M. and Gee G.W. (1995) A method to estimate moisture diffusivity and hydraulic conductivity of moderately dry soil. *Soil Science Society of America Journal*, **59**, 684–689.

Gribb M. (1996) Parameter estimation for determining hydraulic properties of a fine sand from transient flow measurements. *Water Resources Research*, **32**, 1965–1974.

Halbertsma J.M. and Veerman G.J. (1994) *A New Calculation Procedure and A Simple Set-up for the Evaporation Method to Determine Soil Hydraulic Functions*, Report 88, SC-DLO, Wageningen.

Haverkamp R., Ross P.J., Smettem K.R.J. and Parlange J.Y. (1994) Three dimensional analysis of infiltration from the disc infiltrometer. Part 2. Physically based infiltration equation. *Water Resources Research*, **30**, 2931–2935.

Hayashi M., van der Kamp G. and Rudolph D.L. (1997) Use of tensiometer response time to determine hydraulic conductivity of unsaturated soil. *Soil Science*, **162**, 566–575.

Hillel D. and Gardner W.R. (1970) Measurements of unsaturated conductivity and diffusivity by infiltration through an impeding layer. *Soil Science*, **109**, 149–153.

Hillel D., Krentos V.D. and Stylianou Y. (1972) Procedure and test of an internal drainage method for measuring soil hydraulic characteristics. *In Situ, Soil Science*, **114**, 395–400.

Hoffmann-Riem H., van Genuchten M.Th. and Flühler H. (1999) General model for the hydraulic conductivity of unsaturated soils, In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F. and Wu L. (Eds.), University of California, Riverside, pp. 31–42.

Hopmans J.W. and Schoups G.H. (2004) Soil water flow at different scales, *Encyclopedia of Hydrological Sciences*, This issue, John Wiley & Sons.

Hopmans J.W., Šimunek J., Romano N. and Durner W. (2002) Simultaneous determination of water transmission and retention properties – Inverse methods. In *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 963–1008.

Hudson D.B., Wierenga P.J. and Hills R.G. (1996) Unsaturated hydraulic properties from upward flow into soil cores. *Soil Science Society of America Journal*, **60**, 388–396.

Jackson R.D., Reginato R.J. and van Bavel C.H.M. (1965) Comparison of measured and calculated hydraulic conductivities of unsaturated soils. *Water Resources Research*, **1**(3), 375–380.

Jalbert M. and Dane J.H. (2001) Correcting laboratory retention curves for hydrostatic fluid distributions. *Soil Science Society of America Journal*, **65**, 648–654.

Jaynes D.B., Logsdon S.D. and Horton R. (1995) Field method for measuring mobile/immobile water content and solute transfer rate coefficient. *Soil Science Society of America Journal*, **59**, 352–356.

Jury W.A. (1985) *Spatial Variability of Soil Physical Parameters in Solute Migration: A Critical Literature Review*, EPRI Topical Report E4228, Electric Power Research Institute. Palo Alto.

Klinkenberg L.J. (1941) The permeability of porous media to liquids and gases. *Drilling and Production Practice*, American Petroleum Institute: New York, pp. 200–213.

Klute A. (1986) Water retention: laboratory methods, In *Methods of Soil Analysis, Part 1*, *Second Edition*, Klute A. (Ed.), Agronomy Monogr. 9, ASA and SSSA, Madison, pp. 635–662.

Klute A. and Dirksen C. (1986) Hydraulic conductivity and diffusivity: laboratory methods. In *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, Klute A. (Ed.), American Society of Agronomy, Soil Science Society of America (publisher): Madison, pp. 687–734.

Kodesova R., Ordway S.E., Gribb M. and Simunek J. (1999) Estimation of soil hydraulic properties with the cone permeameter: field studies. *Soil Science*, **164**, 527–541.

Kool J.B. and Parker J.C. (1987) Development and evaluation of closed-form expressions for hysteretic soil hydraulic properties. *Water Resources Research*, **23**(1), 105–114.

Kool J.B., Parker J.C. and Van Genuchten M.Th. (1985a) Determination of soil hydraulic properties from one-step outflow experiments by parameter estimation: I theory and numerical studies. *Soil Science Society of America Journal*, **49**, 1348–1354.

Kool J.B., Parker J.C. and van Genuchten M.Th. (1985b) ONESTEP: a nonlinear parameter estimation program for evaluating soil hydraulic properties from one-step outflow experiments. *Virginia Agricultural Experimental Station Bulletin*, **85**, 43.

Kosugi K. (1999) Lognormal distribution model for soil hydraulic properties, In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F. and Wu L. (Eds.), University of California, Riverside, pp. 19–30.

Kozeny, J. (1927) Über kapillare Leitung des Wassers im Boden (Aufstieg, Versickerung und Anwendung auf die Bewässerung). *Sitzungsbericht der Akademie der Wissenschaften zu Wien*, **136**, 271–306.

Krafczyk M. and Rank E. (1995) A parallelized lattice-gas solver for transient Navier-Stokes-flow: implementation and simulation results. *International Journal for Numerical Methods in Engineering*, **38**, 1243–1258.

Kutilek M. and Nielsen D. (1994) *Soil Hydrology*, Catena Verlag: Cremlingen–Destedt, p. 370.

Lebron I. and Robinson D.A. (2003) Particle size segregation during hand packing of coarse granular materials and impacts on local pore-scale structure. *Vadose Zone Journal*, **2**, 330–337.

Leij F.J., Ghezzehei T.A. and Or D. (2002) Analytical models for soil pore-size distribution after tillage. *Soil Science Society of America Journal*, **66**, 1104–1114.

Liu H.H. and Bodvarsson G.S. (2003) Upscaling of constitutive relations in unsaturated heterogeneous tuff matrix. *Journal of Hydrology*, **276**, 198–209.

Liu H.H. and Dane J.H. (1995) Improved computational procedure for retention relations of immiscible fluids using pressure cells. *Soil Science Society of America Journal*, **59**, 1520–1524.

Logsdon S.D. (1997) Transient variation in the infiltration rate during measurement with tension infiltrometers. *Soil Science*, **162**, 233–241.

Mahmood R. and Hubbard K.G. (2003) Simulating sensitivity of soil moisture and evapotranspiration under heterogeneous soils and land uses. *Journal of Hydrology*, **280**, 72–90.

Mapa R.B., Green R.E. and Santo L. (1986) Temporal variability of soil hydraulic properties with wetting and drying subsequent to tillage. *Soil Science Society of America Journal*, **50**, 1133–1138.

Marshall T.J. (1958) A relation between permeability and size distribution of pores. *Journal of Soil Science*, **9**, 1–8.

Mertens J., Jacques D., Vanderborght J. and Feyen J. (2002) Characterisation of the field-saturated hydraulic conductivity on a hillslope: In situ single ring pressure infiltrometer measurements. *Journal of Hydrology*, **263**, 217–229.

Miller E.E. (1980) Similitude and scaling of soil-water phenomena. In *Applications of Soil Physics*, Hillel D. (Ed.), Academic Press: New York, pp. 300–318.

Millington R.J. and Quirk J.P. (1961) Permeability of porous solids. *Transactions of the Faraday Society*, **57**, 1200–1206.

Mohanty B.P., Skaggs T.H. and van Genuchten M.Th. (1998) Impact of saturated hydraulic conductivity on the prediction of tile flow. *Soil Science Society of America Journal*, **62**, 1522–1529.

Mualem Y. (1976) A new model of predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 513–522.

Munoz-Carpena R., Regalado C.M., Alvarez-Benedi J. and Bartoli F. (2002) Field evaluation of the new philip-dunne permeameter for measuring saturated hydraulic conductivity. *Soil Science*, **167**(1), 9–24.

Nielsen D.R., Biggar J.W. and Erh K.T. (1973) Spatial variability of field-measured soil-water properties. *Hilgardia*, **42**, 215–259.

Nielsen D.R., van Genuchten M.T. and Biggar J.W. (1986) Water flow and solute transport processes in the unsaturated zone. *Water Resources Research*, **22**, 89–108.

Nimmo J.R. (1999) Predicting soil-water retention and hydraulic conductivity from textural and structural information. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 923–930.

Nimmo J.R. (2002) Guidelines for method selection [water retention and storage]. In *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series*

No. 5, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 717–720.

Nimmo J.R. and Winfield K.A. (2002) Miscellaneous methods [water retention and storage]. In *Methods of Soil Analysis, Part 4, Physical Methods*, *Soil Science Society of America Book Series No. 5*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 710–714.

Or D. and Tuller M. (1999) Liquid retention and interfacial area in variably saturated porous media: upscaling from single-pore to sample-scale model. *Water Resources Research*, **35**, 3591–3606.

Parkin G.W., Elrick D.E., Kachanoski R.G. and Gibson R.G. (1995) Unsaturated hydraulic conductivity measured by TDR under a rainfall simulator. *Water Resources Research*, **31**, 447–454.

Parkin G.W., Elrick D.E. and Reynolds D.W. (1999) Recent advantages in using ring infiltrometers and TDR to measure hydraulic properties of unsaturated soils. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 553–562.

Purcell W.R. (1949) Capillary pressures – their measurement using mercury and the calculation of permeability therefrom. *Transactions of the American Institute of Mining And Metallurgical Pet. Engineers*, **186**, 39–48.

Reynolds W.D., Bowman B.T., Brunke R.R., Durry C.F. and Tan C.S. (2000) Comparison of tension infiltrometer, pressure infiltrometer, and soil core estimates of saturated hydraulic conductivity. *Soil Science Society of America Journal*, **64**, 478–484.

Reynolds W.D., Elrick D.E. and Clothier B.E. (1985) The constant head well permeameter: effect of unsaturated flow. *Soil Science*, **139**, 172–180.

Richards L.A. (1931) Capillary conduction of liquids through porous media. *Physics*, **1**, 318–333.

Richards L.A., Gardner W.R. and Ogata G. (1956) Physical processes determining water loss from soil. *Soil Science Society of America Proceedings*, **20**, 310–314.

Richards L.A., Russell M.B. and Neal O.R. (1937) Further developments on apparatus for field moisture studies. *Soil Science Society of America Proceedings*, **2**, 55–64.

Romano N. and Santini A. (1999) Determining soil hydraulic functions from evaporation experiments by a parameter estimation approach: Experimental verifications and numerical studies. *Water Resources Research*, **35**, 3343–3359.

Roth K., Vogel H.-J. and Kasteel R. (1999) The scaleway: a conceptual framework for upscaling soil properties. In *Modelling of Transport Processes in Soils at Various Scales in Time and Space*, Feyen J. and Wiyo K. (Eds.), *International Workshop of EurAgEng's Field of Interest on Soil and Water*, Leuven: Wageningen Pers, pp. 477–490, 24–26 November.

Russo D. (1988) Determining soil hydraulic properties by parameter estimation: on the selection of a model for the hydraulic properties. *Water Resources Research*, **24**, 453–459.

Russo D., Bresler E., Shani U. and Parker J.C. (1991) Analyses of infiltration events in relation to determining soil hydraulic properties by inverse problem methodology. *Water Resources Research*, **27**, 1361–1373.

Santos M.J., Goncalves M.C. and Pereira L.S. (1999) Determining the unsaturated soil hydraulic conductivity in the entire suction range using a two-step method. In *Proceeding of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 303–312.

Schaap M.G. (2005) Models for indirect estimation of soil hydraulic properties, *Encyclopedia of Hydrological Sciences*, This issue. John Wiley & Sons.

Schaap M.G., Leij F.J. and van Genuchten M.Th. (1999) A bootstrap – neural network approach to predict soil hydraulic parameters, In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., *et al.* (Eds.), University of California, Riverside, pp. 1237–1248.

Schmalz B., Lennartz B. and van Genuchten M.T. (2003) Analysis of water flow in a large sand tank. *Soil Science*, **168**, 3–14.

Schultze B. (1998) *Optimierung der Messung bodenhydraulischer Eigenschaften durch inverse Simulation von Ausfluß- und Rückflußexperimenten an Bodensäulen*, Ph.D. Thesis, Department of Hydrology, University of Bayreuth, Germany.

Schwartz R.C. and Evett S.R. (2002) Estimating hydraulic properties of a fine-textured soil using a disc infiltrometer. *Soil Science Society of America Journal*, **66**, 1409–1423.

Shao M. and Horton R. (1998) Integral method for estimating soil hydraulic properties. *Soil Science Society of America Journal*, **62**, 585–592.

Si B.C. and Kachanoski R.G. (2000) Unified solution for infiltration and drainage with hysteresis: Theory and field test. *Soil Science Society of America Journal*, **64**, 30–36.

Simunek J. (2005) Models of water flow and solute transport in the unsaturated zone, *Encyclopedia of Hydrological Sciences*, This issue, John Wiley & Sons.

Simunek J. and van Genuchten M.Th. (1996) Estimating unsaturated soil hydraulic properties from tension disc infiltrometer data by numerical inversion. *Water Resources Research*, **32**, 2683–2696.

Simunek J. and van Genuchten M.Th. (1997) Estimating unsaturated soil hydraulic properties from multiple tension disc infiltrometer data. *Soil Science*, **162**, 383–398.

Simunek J., Wendroth O. and van Genuchten M.Th. (1998) Parameter estimation analysis of the evaporation method for determining soil hydraulic properties. *Soil Science Society of America Journal*, **62**, 894–905.

Simunek J., Wendroth O. and van Genuchten M.Th. (1999) Estimating unsaturated soil hydraulic properties from laboratory tension disc infiltrometer experiments. *Water Resources Research*, **35**, 2965–2979.

Simunek J., Wendroth O., Wypler N. and van Genuchten M.Th. (2000) Non-equilibrium water flow characterized by means of upward infiltration experiments. *European Journal of Soil Science*, **52**, 13–24.

Spaans E.J.A. and Baker J.M. (2002) Determination of soil moisture characteristic by freezing. In *Methods of Soil Analysis*, *Third Edition,* Dane J. and Topp G.C. (Eds.), American Society of Agron, pp. 704–709.

Stolte J., Freijer J.I., Bouten W., Dirksen C., Halbertsma J.M., Van Dam J.C., Van den Berg J.A., Veerman G.J. and Wösten J.H.M.

(1994) Comparison of six methods to determine unsaturated soil hydraulic conductivity. *Soil Science Society of America Journal*, **58**, 1596–1603.

Timlin D. and Pachepsky Y. (1998) Measurement of unsaturated soil hydraulic conductivities using a ceramic cup tensiometer. *Soil Science*, **163**, 625–635.

Toorman A.F., Wierenga P.J. and Hills R.G. (1992) Parameter estimation of hydraulic properties from one-step outflow data. *Water Resources Research*, **28**, 3021–3028.

Topp G.C. and Ferré P.A. (2005) Soil water flow at different spatial scales. *Encyclopedia of Hydrological Sciences*, This issue, John Wiley & Sons.

Torquato S. (2001) Random heterogeneous materials: microstructure and macroscopic properties. *Interdisciplinary Applied Mathematics*, Vol. 16, Springer: New York.

Tseng P.H. and Jury W.A. (1993) Simulation of field measurement of hydraulic conductivity in unsaturated heterogeneous soil. *Water Resources Research*, **29**, 2087–2099.

Tuller M. and Or D. (2001) Hydraulic conductivity of variably saturated porous media: film and corner flow in angular pore space. *Water Resources Research*, **31**(5), 1257–1276.

Tuller M., Or D. and Dudley L.M. (1999) Adsorption and capillary condensation in porous media -liquid retention and interfacial configurations in angular pores. *Water Resources Research*, **35**, 1949–1964.

U.S. Department of Energy (2001) *A National Roadmap for Vadose Zone Science & Technology Understanding, Monitoring, and Predicting Contaminant Fate and Transport in the Unsaturated Zone*, Revision 0.0. DOE/ID-10871. Available on-line at **`http://www.inel.gov/vadosezone`**, (verified 12 April 2004).

van Dam J.C., Stricker J.N.M. and Droogers P. (1990) *From One-step to Multi-step: Determination of Soil Hydraulic Functions by Outflow Experiments*, Report 7, Agricultural University, Wageningen.

van Dam J.C., Stricker J.N.M. and Droogers P. (1994) Inverse method to determine soil hydraulic functions from multistep outflow experiments. *Soil Science Society of America Journal*, **58**, 647–652.

van Dam J.C., Stricker J.N.M. and Verhoef A. (1992) An evaluation of the one-step outflow method. In *Proceedings of the International Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, van Genuchten M.Th., Leij F.J. and Lund L.J. (Eds.), University of California: Riverside, pp. 633–644.

Vandervaere J.P., Vauclin M. and Elrick D.E. (2000) Transient flow from tension infiltrometers: II. Four methods to determine sorptivity and conductivity. *Soil Science Society of America Journal*, **64**, 1272–1284.

van Es H.M., Ogden C.B., Hill R.L., Schindelbeck R.R. and Tsegaye T. (1999) Integrated assessment of space, time, and management-related variability of soil hydraulic properties. *Soil Science Society of America Journal*, **63**, 1599–1607.

van Genuchten M.Th. (1980) A closed–form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

van Genuchten M.Th., Leij F.J. and Wu L. (1999a) *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, University of California: Riverside.

van Genuchten M.Th., Leij F.J. and Wu L. (1999b) Characterization and measurement of the hydraulic properties of unsaturated porous media, In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., *et al.* (Eds.), University of California: Riverside, pp. 1–12.

Vogel H.-J. and Roth K. (2001) Quantitative morphology and network representation of soil pore structure. *Advances in Water Resources*, **24**, 233–242.

Vogel H.-J. and Roth K. (2003) Moving through scales of flow and transport in soil. *Journal of Hydrology*, **272**, 95–106.

Vrugt J.A. and Bouten W. (2001) Information content of data for identifying soil hydraulic parameters from outflow experiments. *Soil Science Society of America Journal*, **65**, 19–27.

Vrugt J.A. and Dane J.H. (2005) Inverse modeling of soil hydraulic properties.i. *Encyclopedia of Hydrological Sciences*, this issue, John Wiley & Sons.

Vrugt J.A., Gupta H.V., Bastidas L.A., Bouten W. and Sorooshian S. (2003) Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, **39**(8), 1214, doi:10.1029/2002WR001746.

Wang Q., Horton R. and Shao M. (2002) Horizontal infiltration method for determining Brooks-Corey model parameters. *Soil Science Society of America Journal*, **66**, 1733–1739.

Wang D., Yates S.R., Lowery B. and van Genuchten M.Th. (1998) Estimating soil hydraulic properties using tension infiltrometers with varying disk size. *Soil Science*, **163**, 256–361.

Warrick A.W. (1992) Models for disc permeameters. *Water Resources Research*, **28**, 1319–1327.

Watson K.K. (1966) An instantaneous profile method for determining the hydraulic conductivity of unsaturated porous materials. *Water Resources Research*, **2**, 709–715.

Wendroth O. and Simunek J. (1999) Soil hydraulic properties determined from evaporation and tension infiltration experiments and their use for modeling field moisture status. In *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 737–748.

White I. and Sully M.J. (1987) Macroscopic and microscopic capillary length and times scales from field infiltration. *Water Resources Research*, **23**, 1514–1522.

Wind G.P. (1968) Capillary conductivity data estimated by a simple method. In *Water in the Unsaturated Zone*, *Proceedings of the Wageningen Symposium*, Vol. 1, June 1966, Rijtema P.E. and Wassink H. (Eds.), International Association of Scientific Hydrology, Gent/Brugge/Paris, pp. 181–191.

Wooding R.A. (1986) Steady Infiltration from a shallow circular pond. *Water Resources Research*, **4**, 1259–1273.

Wu L., Pan L., Roberson M.J. and Shouse P.J. (1997) Numerical evaluation of ring-infiltrometers under various soil conditions. *Soil Science*, **162**, 771–777.

Wyllie M.R.J. and Gardner G.H.F. (1958) The generalized Kozeny-Carman equation: Part 2 – a novel approach to problems of fluid flow, *World Oil*, **146**, 210–228.

Yeh T.-C. and Simunek J. (2001) Stochastic fusion of information for characterizing and monitoring the vadose zone. *Vadose Zone Journal*, **1**, 207–221.

Young M.H., Karagunduz A., Simunek J. and Pennell K.D. (2002) A modified upward infiltration method for characterizing soil hydraulic properties. *Soil Science Society of America Journal*, **66**, 57–64.

Zhang R. (1998) Estimating soil hydraulic conductivity and capillary length from the disc infiltrometer. *Soil Science Society of America Journal*, **62**, 1513–1521.

Zhang Z.F., Kachanoski R.G., Parkin G.W. and Si B. (2000a) Measuring hydraulic properties using a line source: I analytical expressions. *Soil Science Society of America Journal*, **64**, 1554–1562.

Zhang Z.F., Kachanoski R.G., Parkin G.W. and Si B. (2000b) Measuring hydraulic properties using a line source: II field test. *Soil Science Society of America Journal*, **64**, 1563–1569.

Zhang Z.F., Ward A.L. and Glee G.W. (2003) Estimating soil hydraulic parameters of a field drainage experiment using inverse techniques. *Vadose Zone Journal*, **2**, 201–211.

Zou Z.Y., Young M.H., Li Z. and Wierenga P.J. (2001) Estimation of depth averaged unsaturated soil hydraulic properties from infiltration experiments. *Journal of Hydrology*, **242**, 26–42.

Zurmühl T. (1996) Evaluation of different boundary conditions for independent determination of hydraulic parameters using outflow methods. In *Parameter Identification and Inverse Problems in Hydrology, Geology and Ecology*, Vol. 23, DuChateau P. and Gottlieb J. (Eds.), Water Science and Technology Library, Kluwer: Dordrecht, pp. 165–184.

Zurmühl T. (1998) Capability of convection dispersion transport models to predict transient water and solute movement in undisturbed soil columns. *Journal of Contaminant Hydrology*, **30**, 99–126.

76: Models for Indirect Estimation of Soil Hydraulic Properties

MARCEL G SCHAAP

Department of Environmental Sciences, University of California, Riverside, CA, US

This article describes methods for the indirect estimation of soil hydraulic properties, such as soil water retention and hydraulic conductivity characteristics. Indirect methods can be classified into semiphysical methods that are based on mechanical assumptions regarding particle and pore arrangements, and statistical methods that are known as pedotransfer functions. Both classes are described and evaluated on their merits, and some characteristic examples are given. A list of applicable software is described at the end of the article and an extensive reference list provides sufficient material for further background information.

INTRODUCTION

Qualitative knowledge about soil hydraulic properties such as water retention and hydraulic conductivity has historically been an important factor for assessing the suitability of land for agriculture, settlement, or trafficability. In modern agricultural and engineering practices, varying degrees of quantitative detail about soil hydraulic properties are needed for determining the soil water holding capacity, infiltration, percolation, and runoff rates, or for quantifying the transport of pollutants in soil. Although automation and computer technology have certainly advanced the ease with which hydraulic properties can be measured (cf. Dane and Topp, 2002, *see* **Chapter 75, Determining Soil Hydraulic Properties, Volume 2** by Dürner), their determination has by no means become easy. Many measurement methods are still labor-intensive and expensive. In addition, soil hydraulic properties are often subject to considerable spatial and sometimes temporal variability, making measurements less representative. Measurements of hydraulic properties for regional, continental, or global scales are virtually impossible.

The expense and difficulty of performing soil hydraulic measurements in laboratory or field are often used as arguments for developing indirect methods for estimating soil hydraulic properties using widely available surrogate data. The general method is to define physical relations or find statistical correlations between predictors such as soil texture and the soil hydraulic properties. After the modeling exercise, the model can be tested on independent data, or be applied for common use. In the past decades, literally hundreds of such studies have been performed. This review considers two main classes of indirect methods, semiphysical and empirical approaches, and briefly discusses common modeling concepts. Reviews with slightly different perspectives concerning indirect models can be found in Rawls *et al.* (1991) and Wösten *et al.* (2001). Most model development is data-driven and requires soil databases in which both predictors and measured hydraulic properties are present. This review will therefore also discuss some pertinent databases and software packages.

SEMIPHYSICAL MODELS

Semiphysical methods recognize the shape similarity between the cumulative particle-size distribution and the water retention characteristic. Although none of the models can predict hydraulic properties from first principles, they do offer valuable conceptual insights into the physical relations between the texture distribution and the pore-size distribution. A drawback of these methods is that they often require very detailed particle-size distributions, making them almost as difficult to apply as direct measurements. Three model types can be discerned in the literature.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

The Arya–Paris Model

The Arya and Paris (1981) model uses information from k particle-size classes to estimate k pairs of water contents and pressure heads. Each class is thought of to consist of n_k spherical particles of mean radius R_k. The pore volume of this class is associated with a cylindrical pore of radius r_k. The Arya–Paris model assumes that the bulk density is the same for each particle class. The pore volume, V_k, for each particle-size class is then

$$V_k = \left(\frac{W_k}{\rho_s}\right)e \tag{1}$$

where W_k is the total mass in particle class k and ρ_s is the particle density. The water content follows by summing the V_k of all particle-size classes $1 \ldots k$, starting with the class with the smallest R_k. The void ratio, e, is given by

$$e = \frac{(\rho_s - \rho_b)}{\rho_b} \tag{2}$$

The radius of the pore belonging to class k is given by (Arya and Paris, 1981):

$$r_k = \frac{R_k\sqrt{2en_k^{(1-\alpha_{ap})}}}{3} \tag{3}$$

where n_k is the number of spherical particles of radius R_k required to fill the mass in the particle-size class. The corresponding pressure head is subsequently calculated with the capillary law. The empirical parameter α_{ap} accounts for non-spherical soil particles and should be equal or greater than 1. Arya and Paris (1981) initially determined that α_{ap} ranged between 1.31 and 1.43 for five texturally different soils, but later it was found that α_{ap} varied between 1.02 and 2.97 (Arya *et al.*, 1982; Schuh *et al.*, 1988; Mishra *et al.*, 1989). Arya *et al.* (1999a) modified the original Arya–Paris concept by allowing α_{ap} to vary according to particle size and extended the model to unsaturated conductivity (Arya *et al.*, 1999b). To our knowledge, this model extension has not yet been tested on independent data.

Fractal Approaches

Several studies used fractal concepts to develop indirect methods. Fractal patterns reveal themselves by exhibiting power-law scaling relations between the observed quantity and the measurement scale, R. Employing a somewhat similar logic as the Arya–Paris model, fractal behavior may be found on a particle number basis (e.g. Tyler and Wheatcraft, 1989) as

$$N \sim R^{-D_N} \tag{4}$$

where N is the number of particles greater than a measurement scale R, and D_N is the fractal dimension (ranging between 0 and 3). A fractal particle-size distribution thus exhibits a straight line in a simple log N versus log R plot. Tyler and Wheatcraft (1989) argued that the parameter α_{ap} in the Arya and Paris (1981) equation should be equal to $D_N - 1$. Tyler and Wheatcraft (1992), however, indicated that "soils that show fractal scaling are a rather small subset of the soils commonly encountered in the field". Indeed, using more than 1100 soil data sets, Tietje and Tapkenhinrichs (1993) demonstrated that the fractal approach of Tyler and Wheatcraft (1989) was not more accurate than the original Arya and Paris (1981) concept with $\alpha_{ap} = 1.38$. Other fractal-based approaches were proposed by Rieu and Sposito (1991a, b) and Bird *et al.* (2000) who included soil structure into the fractal concept.

The Haverkamp and Parlange Model

Haverkamp and Parlange (1986) built a semiphysical model around the simple assumption that the pore radius r is linearly related to the particle radius R

$$r = \gamma R \tag{5}$$

where γ is a factor of proportionality, requiring that pores of different sizes have the same shape. Combined with the capillary law that links pore size to capillary pressure, it follows that a retention characteristic can be derived from a cumulative particle-size distribution. Haverkamp and Parlange (1986) developed a somewhat complex model that included hysteresis phenomena to estimate Brooks and Corey (1964) retention parameters by curve fitting the van Genuchten (1980) equation to the cumulative particle-size distribution. This approach was later simplified and tested by Schaap and Bouten (1996), who showed that the model can successfully be applied to sandy soils. Bouraoui *et al.* (1999) showed that a modified and simplified Haverkamp and Parlange concept could be applied to a large part of the textural triangle. However, beyond these references, the Haverkamp and Parlange model has attracted little following in the literature.

EMPIRICAL MODELS

Empirical methods, often called *pedotransfer functions* (PTFs, Bouma and van Lanen, 1987), generally focus on practical applicability and often use more or less simple statistical models to estimate hydraulic properties. Contrary to most physically based models, empirical methods often require limited – but easily accessible – input data such as sand, silt, or clay percentages and porosity, although more elaborate combinations of input data are also possible (cf. Rawls *et al.*, 1991; Wösten *et al.*, 2001). There are many

different PTFs, and it is impractical to describe all these in detail. We will limit ourselves to describing some modeling concepts while giving some examples of characteristic PTFs. The concepts distinguished in this review are, class-based methods, point-based, and parametric approaches.

Class PTFs

Class PTFs provide hydraulic properties for particular soil classes. The advantage of class PTFs is their simplicity (essentially, they consist of lookup tables) and modest requirements regarding input data. Only class information is necessary, thus enabling estimates of hydraulic properties from qualitative field data. Class PTFs for the 12 USDA textural classes were reported by Clapp and Hornberger (1978), Rawls *et al.* (1982), Carsel and Parrish (1988), Rawls and Brakensiek (1985), and Schaap and Leij (1998), among others. Wösten *et al.* (1999) provided average van Genuchten (1980) retention- and unsaturated conductivity parameters for 12 soil classes based on the FAO textural classification but also made a further distinction for subsoils, topsoils, and organic soils. The resulting class average parameters were used in conjunction with a GIS system to estimate the available water content on a scale of 1 : 1 000 000 for most of the European Union. Other large-scale applications can be found in Kern (1995) and Imam *et al.* (1999). A drawback of class PTFs is that discrete changes of hydraulic properties occur between two adjacent classes (e.g. loam to sandy loam, or topsoil to subsoil). Such changes may not always be realistic, especially for small-scale applications. Instead, point-based or parametric PTFs provide continuously varying estimates that may be more useful in such cases.

Point PTFs

Point PTFs use simple linear expressions to estimate individual water retention or conductivity points (i.e. water content-pressure or conductivity-pressure points) from texture and other soils data. Point PTFs are arguably the most precise of PTFs because they estimate the hydraulic points directly, without relying on parameterized forms of the hydraulic characteristics such as done by parametric PTFs (cf. Schaap and Bouten, 1996; Minasny *et al.*, 1999 for some comparisons). A drawback of point PTFs is that they are limited in making estimates at specific pressure heads. However, if necessary, parameterized results can be obtained by fitting appropriate retention equations to the point estimates (e.g. Saxton *et al.*, 1986). In most cases, separate regression equations are used for each pressure. Tietje and Tapkenhinrichs (1993) noted that this causes some point PTFs to exhibit unrealistically increasing water contents with stronger suctions. Examples of point PTFs can be found in Rawls *et al.* (1982), Rawls *et al.* (1983), Ahuja *et al.* (1985), Rajkai and Varallyay (1992), Thomasson and Carter (1992), Bristow *et al.* (1999), Minasny *et al.* (1999), and Renger *et al.* (1999).

Parametric PTFs

Parametric PTFs estimate parameters of retention or conductivity equations, such as the Brooks and Corey (1984), van Genuchten (1980), and the Mualem (1976) equations. Contrary to class PTFs, these models estimate hydraulic parameters that vary continuously with input data. Contrary to point PTFs, parametric PTFs can provide hydraulic properties at arbitrary capillary pressures. It is impossible to describe all of these approaches; we will therefore give only a few well-known examples.

Brakensiek *et al.* (1984), Rawls and Brakensiek (1985), and Rawls *et al.* (1992) presented parametric PTFs that estimated Brooks and Corey (1964) parameters and saturated hydraulic conductivity from porosity, ϕ, and sand and clay percentages (S and C, respectively). In these PTFs, the saturated water content is set equal to the porosity while the other Brooks-Corey parameters and the saturated hydraulic conductivity are related to S, C, and ϕ, using the polynomial

$$\begin{aligned} p = {} & a_1 + a_2 S + a_3 C + a_4 \phi + a_5 S^2 + a_6 C^2 + a_7 \phi^2 \\ & + a_8 S\phi + a_9 C\phi + a_{10} S^2 C + a_{11} S^2 \phi + a_{12} C^2 \phi \\ & + a_{13} S C^2 + a_{14} C\phi^2 + a_{15} S^2 \phi^2 + a_{16} C^2 \phi^2 \end{aligned} \quad (6)$$

where p is a Brooks–Corey parameter or the saturated hydraulic conductivity and a_i are model coefficients.

Vereecken *et al.*, (1989, 1990) provided expressions for water retention and unsaturated hydraulic conductivity for 182 Belgian soil horizons. The water retention was described with a modified van Genuchten equation and the unsaturated hydraulic conductivity was described with the Gardner (1958) equation. Vereecken *et al.* (1989, 1990) provided a number of ways to estimate the seven hydraulic parameters in these equations. One of the approaches was fitting the parameters $a_1 \ldots a_6$ in the equation

$$\begin{aligned} f(p) = {} & a_1 + a_2 f(C) + a_3 f(Si) + a_4 f(S) \\ & + a_5 f(OM) + a_6 f(\rho_b) \end{aligned} \quad (7)$$

where p is the hydraulic parameter being estimated, S: Sand, C: clay, Si: silt, ρ_b: bulk density, OM: organic matter, and $f()$ indicates that a transformation (e.g. logarithms, exponents, etc.) may be applied to the parameter in ellipses.

Artificial neural networks form a special class of PTFs and were introduced by Pachepsky *et al.* (1996), Schaap and Bouten (1996), and Tamari *et al.* (1996). Neural networks are sometimes described as "universal function

approximators" that can "learn" to approximate any continuous (nonlinear) function to any desired degree of accuracy (cf. Haykin, 1994). An advantage of neural networks as compared to regression PTFs is that they require no *a priori* model concept (e.g. linear or exponential functions). This property makes ANNs well suited to build empirical PTFs. However, the method also results in black-box models in which the exact relations between predictors and hydraulic properties are difficult to determine.

Mixed results have been obtained with neural network PTFs. Schaap and Bouten (1996) and Schaap *et al.* (1998) showed that neural networks made estimates with significantly smaller errors than more traditional approaches. Pachepsky *et al.* (1996) found that neural networks perform better than multiple linear regressions when used as point PTFs, but that the two methods produced comparable results when used as parametric PTFs. Tamari *et al.* (1996) reported that neural networks were not better than multiple linear regressions if the uncertainty in the data was large. Minasny *et al.* (1999) found that nonlinear regression reached a similar performance as a neural network approach.

DATABASES AND SOFTWARE

Several public databases have been compiled that suit the development and testing of models for the indirect estimation of hydraulic properties. The oldest database is probably UNSODA (Nemes *et al.*, 2001 and available at: **www.ussl.ars.usda.gov/models/unsoda.htm**). This world wide database contains data about laboratory and field hydraulic and other relevant soil characteristics for 790 soils. The HYPRES database contains similar soils data as UNSODA, but for European soils. HYPRES was described in Wösten *et al.* (1999) and is available at **www.macaulay.ac.uk/hypres/index.html**.

Several software packages for estimating soil hydraulic properties are available on the internet. Soil Water Characteristics from Texture (SWCT) is based on Saxton *et al.* (1986) and estimates wilting point, field capacity, and available water content. SWCT is part of the Soil Plant Atmosphere Water (SPAW) Field and Pond Hydrology package and is targeted at farmers and resource managers interested in water and nutrient budgeting in soil and ponds, and is available at **http://www.bsyse.wsu.edu/saxton/spaw/**.

SOILPAR is a program developed by M. Donatelli and M. Acutis at the Research Institute for Industrial Crops (ISCI), Bolongna, Italy. The program implements 10-point PTFs and four parametric PTFs for water retention and hydraulic conductivity, and can be downloaded at **http://www.sipeaa.it/ASP/ASP2/SOILPAR.asp**. The program provides a wide range of plotting and data analysis functions.

Rosetta is a Windows-based program that implements artificial neural network PTFs published by Schaap *et al.* (1998), Schaap and Leij (1998), and Schaap and Leij (2001), and is available from **http://www.ussl.ars.usda.gov/models/rosetta/rosetta.HTM**. The program implements five parametric PTFs for the estimation of water retention saturated and unsaturated hydraulic conductivity. Rosetta uses a hierarchical approach to maximize the accuracy of the PTF estimates given a particular data availability.

Minasny and McBratney (2002) developed the Neuropack software package (**http://www.usyd.edu.au/su/agric/acpa/neuropack/neuropack.htm**). Unlike the other software packages that implement existing PTFS, Neuropack is primarily intended to develop neural network-based PTFs. Optimization results can be saved and or be tested using independent soils data.

CONCLUSION

Indirect methods are valuable assets in many soils applications because they allow estimation of soil hydraulic properties where none exist or where direct measurements would be prohibitive. The many different models that exist are mostly geared toward particular soils or datasets (Schaap and Leij, 1998). A number of common PTFs were tested for general applicability in Tietje and Tapkenhinrichs (1993), Kern (1995), Tietje and Hennings (1996), Schaap *et al.* (1998), Minasny *et al.* (1999), and Wösten *et al.* (2001). However, these studies showed that no clearly superior PTF exists that works well in all applications. When possible, it is therefore advisable to test several PTFs for accuracy before blindly trusting the results.

Some important issues exist that are worth considering within the context of indirect estimations. Most indirect methods are based on hydraulic data from laboratory measurements, yet are usually applied to field situations. Laboratory and field measurements do not necessarily yield similar results, so it is possible that estimates of indirect methods are biased. Another issue is that most indirect methods were developed using data from specific parts of the world, particularly for soils from temperate climates. With the exception of a few studies (e.g. Epebinu and Nwadialo, 1993; Tomasella and Hodnett, 1997; Bristow *et al.*, 1999), hydraulic data and corresponding indirect methods about tropical soils are a virtual *terra incognita*. Obtaining comprehensive soil and hydraulic data for regions beyond those now represented in regional and international databases will probably help improve the general applicability of indirect methods and aid the management of local natural and agricultural resources.

Acknowledgment

M.G. Schaap was supported, in part, by the SAHRA science and technology center under a grant from NSF (EAR-9876800).

REFERENCES

Arya L.M., Leij F.J., Shouse P.J. and van Genuchten M.Th. (1999a) Relationship between the Hydraulic conductivity function and the particle-size distribution. *Soil Science Society of America Journal*, **63**, 1063–1070.

Arya L.M., Leij F.J., van Genuchten M.Th. and Shouse P.J. (1999a) Scaling parameter to predict the soil water characteristic from particle-size distribution data. *Soil Science Society of America Journal*, **63**, 510–519.

Ahuja L.R., Naney J.W. and Williams R.D. (1985) Estimating soil water characteristics from simpler properties and limited data. *Soil Science Society of America Journal*, **49**, 1100–1105.

Arya L.M. and Paris J.F. (1981) A physico-empirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data. *Soil Science Society of America Journal*, **45**, 218–1227.

Arya L.M., Richter J.C. and Davidson S.A. (1982) *A Comparison of Soil Moisture Characteristic Predicted by the Arya-Paris Model with Laboratory-Measured Data*, AgRISTARS Technical Report SM-L1-04247, JSC-17820. NASA/Johnson Space Center: Houston.

Bird N.R.A., Perrier E. and Rieu M. (2000) The water retention function for a model of soil structure with pore and solid fractal distributions. *European Journal of Soil Science*, **51**, 55–63.

Bouraoui F., Haverkamp R., Zammit C. and Parlange J.-Y. (1999) Physically-based pedotransfer function for estimating water retention curve shape parameters. In *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 947–958.

Bouma J. and van Lanen J.A.J. (1987) Transfer functions and threshold values: From soil characteristics to land qualities. In *Quantified Land Evaluation. International Institute Aerospace*, Beek K.J., Burrough P.A. and McCormack D.E. (Ed.), Survey Earth Science, ITC Publication No. 6: Enschede, pp. 106–110.

Brakensiek D.L., Rawls W.J. and Stephenson G.R., (1984) *Modifying SCS Hydrologic Soil Groups and Curve Numbers for Rangeland Soils*, ASAE Paper No. PNR-84-203, St. Joseph.

Bristow K.L., Smettem K.R.J., Ross P.J., Ford E.J., Roth C.H. and Verburg K. (1999) Obtaining hydraulic properties for soil water balance models: some pedotransfer functions for tropical Australia. In *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 1103–1120.

Brooks R.H. and Corey A.T. (1964) *Hydraulic Properties of Porous Media*, Hydrol. paper 3, Colorado State University: Fort Collins.

Carsel R.F. and Parrish R.S. (1988) Developing joint probability distributions of soil water retention characteristics. *Water Resources Research*, **24**, 755–769.

Clapp R.B. and Hornberger G.M. (1978) Empirical equations for some soil hydraulic properties. *Water Resources Research*, **14**, 601–604.

Dane J.H. and Topp G.C. (2002) *Methods of Soil Analysis, Part 4: Physical Methods*, Soil Science of America, Inc.: Madison Wisconsin.

Epebinu O. and Nwadialo B. (1993) Predicting soil water availability from texture and organic matter content for Nigerian soils. *Communications in Soil Science and Plant Analysis*, **24**, 633–640.

Gardner W.R. (1958) Some steady state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science*, **85**, 228–232.

Haverkamp R. and Parlange J.Y. (1986) Predicting the water-retention curve from particle size distribution: 1. Sandy soils without organic matter. *Soil Science*, **142**, 325–339.

Haykin S. (1994) *Neural Networks, a Comprehensive Foundation*, *First Edition*, Macmillan College Publishing Company: New York.

Imam B., Sorooshian S., Mayr T., Schaap M.G., Wösten H. and Scholes B. (1999) *Comparison of Pedotransfer Functions to Compute Water Holding Capacity Using the van Genuchten Model in Inorganic Soils*. Report to the IGBP-DIS soils data tasks, IGBP-DIS Working Paper #22, IGBP-DIS office, CNRM, 42 avenue G. Coriolis, 31057: Toulouse Cedex.

Kern J.S. (1995) Evaluation of soil water retention models based on basic soil physical properties. *Soil Science Society of America Journal*, **59**, 1134–1141.

Minasny B. and McBratney A.B. (2002) The neuro-m method for fitting neural network parametric pedotransfer functions. *Soil Science Society of America Journal*, **66**, 352–362.

Minasny B., McBratney A.B. and Bristow K.L. (1999) Comparison of different approaches to the development of pedotransfer functions of water retention curves. *Geoderma*, **93**, 225–253.

Mishra S., Parker J.C. and Singhal N. (1989) Estimation of soil hydraulic properties and their uncertainty from particle size distribution data. *Journal of Hydrology*, **108**, 1–18.

Mualem Y. (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 513–522.

Nemes A., Schaap M.G., Leij F.J. and Wösten J.H.M. (2001) Description of the unsaturated soil hydraulic database UNSODA version 2.0. *Journal of Hydrology*, **251**, 151–162.

Pachepsky YaA, Timlin D. and Varallyay G. (1996) Artificial neural networks to estimate soil water retention from easily measurable data. *Soil Science Society of America Journal*, **60**, 727–733.

Rajkai K. and Varallyay G. (1992) Estimating soil water retention from simpler properties by regression techniques. In *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, van Genuchten M.Th, Leij F.J. and Lund L.J. (Eds.), *Proceedings of the International Workshop*, Riverside, 11–13 October 1989, University of California: Riverside, pp. 417–426.

Rawls W.J., Ahuja L.R. and Brakensiek D.L. (1992) Estimating soil hydraulic properties from soils data. In *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, van Genuchten M.Th, Leij F.J. and Lund L.J. (Eds.), *Proceedings*

of the International Workshop, Riverside, 11–13 October 1989, University of California: Riverside, pp. 329–340.

Rawls W.J. and Brakensiek D.L. (1985) Prediction of soil water properties for hydrologic modeling. In *Watershed Management in the Eighties*, Jones E.B. and Ward T.J. (Eds.), *Proceedings of the Irrigation and Drainage Division*, 30 April - 1 May 1985, ASCE: Denver, pp. 293–299.

Rawls W.J., Brakensiek D.L. and Saxton K.E. (1982) Estimation of soil water properties. *Transactions of the ASAE*, **25**, 1316–1320.

Rawls W.J., Brakensiek D.L. and Soni B. (1983) Agricultural management effects on soil water processes, Part I: soil water retention and Green and Ampt infiltration parameters. *Transactions of the ASAE*, **26**, 1747–1752.

Rawls W.J., Gish T.J. and Brakensiek D.L. (1991) Estimating soil water retention from soil physical properties and characteristics. In *Advances in Soil Science*, Stewart B.A. (Ed.), Springer-Verlag: New York.

Renger M., Stoffregen H., Klocke J., Facklam M., Wessolek G., Roth C.H. and Plagge R. (1999) Autoregressive procedure to predict hydraulic conductivity: measured and predicted results. In *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.Th., Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 1037–1046.

Rieu M. and Sposito G. (1991a) Fractal fragmentation, soil porosity and soil water properties: I. Theory. *Soil Science Society of America Journal*, **55**, 1231–1238.

Rieu M. and Sposito G. (1991b) Fractal fragmentation, soil porosity and soil water properties: II. Applications. *Soil Science Society of America Journal*, **55**, 1239–1244.

Saxton K.E., Rawls W.J., Romberger J.S. and Papendick R.I. (1986) Estimating generalized soil-water characteristics from texture. *Soil Science Society of America Journal*, **50**, 1031–1036.

Schaap M.G. and Bouten W. (1996) Modeling water retention curves of sandy soils using neural networks. *Water Resources Research*, **32**, 3033–3040.

Schaap M.G. and Leij F.J. (1998) Database related accuracy and uncertainty of pedotransfer functions. *Soil Science*, **163**, 765–779.

Schaap M.G., Leij F.J. and van Genuchten M.Th (1998) Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity. *Soil Science Society of America Journal*, **62**, 847–855.

Schaap M.G., Leij F.J. and van Genuchten M.Th (2001) Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, **251**, 163–176.

Schuh W.M., Cline R.L. and Sweeney M.D. (1988) Comparison of a laboratory procedure and a textural model for predicting in situ soil water retention. *Soil Science Society of America Journal*, **52**, 1218–1227.

Tamari S., Wösten J.H.M. and Ruiz-Suárez J.C. (1996) Testing an artificial neural network for predicting soil hydraulic conductivity. *Soil Science Society of America Journal*, **60**, 1732–1741.

Thomasson A.J. and Carter A.D. (1992) Current and future uses of the UK soil water retention dataset. In *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, van Genuchten M.Th., Leij F.J. and Lund L.J. (Eds.), *Proceedings of the International Workshop*, Riverside, 11–13 October 1989, University of California: Riverside, pp. 355–358.

Tietje O. and Hennings V. (1996) Accuracy of the saturated hydraulic conductivity prediction by pedo-transfer functions compared to the variability within FAO textural classes. *Geoderma*, **69**, 71–84.

Tietje O. and Tapkenhinrichs M. (1993) Evaluation of pedotransfer functions. *Soil Science Society of America Journal*, **57**, 1088–1095.

Tomasella J. and Hodnett M.G. (1997) Estimating unsaturated hydraulic conductivity of Brazilian soils using soil-water retention data. *Soil Science*, **162**, 703–712.

Tyler S.W. and Wheatcraft S.W. (1989) Application of fractal mathematics to soil water retention estimation. *Soil Science Society of America Journal*, **53**, 987–996.

Tyler S.W. and Wheatcraft S.W. (1992) Fractal scaling of soil particle-size distributions: analysis and limitations. *Soil Science Society of America Journal*, **56**, 362–369.

van Genuchten M.Th. (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

Vereecken H., Maes J. and Feyen J. (1990) Estimating unsaturated hydraulic conductivity from easily measured soil properties. *Soil Science*, **149**, 1–12.

Vereecken H., Maes J., Feyen J. and Darius P. (1989) Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. *Soil Science*, **148**, 389–403.

Wösten J.H.M., Lilly A., Nemes A. and Le Bas C. (1999) Development and use of a database of hydraulic properties of European soils. *Geoderma*, **90**, 169–185.

Wösten J.H.M., Pachepsky Y.A. and Rawls W.J. (2001) Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology*, **251**, 123–150.

77: Inverse Modeling of Soil Hydraulic Properties

JASPER A VRUGT[1] AND JACOB H DANE[2]

[1]Formerly of: Institute for Biodiversity and Ecosystem Dynamics – Physical Geography, University of Amsterdam, Amsterdam, The Netherlands & Currently at: Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, US
[2]Department of Agronomy and Soils, Auburn University, Auburn, AL, US

Numerical-simulation models of water flow and solute transport in the vadose zone are important tools in environmental research. The accuracy of predictions with these models heavily relies on accurate estimates of the unsaturated soil hydraulic parameters. Initial attempts on the estimation of the soil hydraulic parameters mainly focused on the use of relatively simple static or steady-state flow experiments. While these laboratory methods have the advantage of being relatively simple to implement, they are typically time consuming and require restrictive initial and boundary conditions to satisfy the assumptions of the corresponding analytical solutions. Significant advances in computational capabilities in the 1980s have stimulated research on the use of Inverse Modelling (IM) for the estimation of soil hydraulic properties. Parameter estimation using IM accommodates much more flexible experimental conditions than required for the static or steady-state flow methods, thereby enabling a rapid characterization of the hydraulic properties of the considered soil domain. This article explores the historical development of current perspectives on the identification of a single "best" set of soil hydraulic parameters using IM (thereby effectively neglecting the influence of possible sources of uncertainty on the final parameter estimates), and discusses alternative (Bayesian and multiple-criteria) parameter estimation strategies which can be used to quantify the uncertainty (probabilistic and multiobjective) associated with the inversely estimated soil hydraulic properties. Throughout this article, we use a classical transient laboratory outflow experiment to further illustrate the various aspects of the IM procedure for estimating the soil hydraulic parameters.

INTRODUCTION

An adequate hydrological description of water flow and contaminant transport in the vadose zone heavily relies on accurate estimates of the soil water retention and hydraulic conductivity function (hereafter referred to as soil hydraulic properties). To enable an accurate soil physical characterization, methodologies need to be developed which allow a rapid, reliable, and cost-effective estimation of the hydraulic properties of the considered soil domain, including its spatial variability. Most of the early work reporting on the estimation of the hydraulic properties of unsaturated soils focused on the use of relatively simple static or steady-state flow experiments. These static or steady-state flow experiments have the advantage of being relatively simple to implement. However, these methods are typically time consuming and require restrictive initial and boundary conditions to satisfy the assumptions of the corresponding analytical solutions.

Significant advances in computational capabilities in the 1980s have stimulated research on the use of Inverse Modelling (IM) for the estimation of soil hydraulic properties. While other applications of IM exist, such as retrieving the initial conditions of an initial/boundary-value differential equation (Press *et al.*, 1992) or reconstructing a contaminant source history (Neupauer *et al.*, 2000), this article primarily focuses on the inverse problem of parameter estimation applied to soil hydrology.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Parameter estimation using IM accommodates more flexible experimental conditions than required for the static or steady-state flow methods, thereby enabling a rapid characterization of the hydraulic properties of the considered soil domain. We share the recent opinion voiced by Hopmans and Šimůnek (1999) that parameter estimation using IM has tremendous potential in characterizing flow and transport processes in the vadose zone, while simultaneously presenting us with an additional tool to better understand the controlling mechanisms. However, when using an IM approach, the soil hydraulic properties can no longer be estimated by direct inversion, as applicable for the steady-state and static experiments (with a few exceptions, such as the instantaneous profile method for which both direct and indirect approaches exist), but only meaningfully inferred using an iterative solution of the flow process, thereby placing a heavier demand on computational resources. In this iterative process, repeated numerical simulations of the governing transient flow equations are performed, thereby minimizing the difference between the observed and model predicted flow variables. Nowadays, IM is a frequently employed approach within the hydrologic science community to estimate effective parameters, at scales ranging from laboratory samples to regional applications. While, IM has been applied successfully mostly to the parameter estimation of soil water retention and unsaturated hydraulic conductivity functions, recent progress is also reported in the description of water uptake by plant roots (Musters *et al.*, 1999; Vrugt *et al.*, 2001a,b; Hupet *et al.*, 2002) and estimation of soil thermal properties and water fluxes (Hopmans *et al.*, 2002). In this article, we shall focus our attention almost exclusively on the estimation of soil hydraulic properties using IM. Where necessary, reference is made to other articles in this Encyclopedia, whose reading will improve the understanding of the experiments and mathematical concepts underlying the IM procedure.

THE PROCESS OF PARAMETER ESTIMATION

The process of IM is closely related to parameter estimation or model calibration. Parameter estimation or model calibration is a common problem in many areas of process modelling, both in on-line applications, such as real time optimization, and in off-line applications, such as the modelling of reaction kinetics and phase equilibrium. The goal is to determine values of model parameters that provide the best fit to measured data, generally based on some type of least squares or maximum likelihood criterion. Usually, this requires the solution of a nonlinear and frequently nonconvex optimization problem.

Among the first to suggest the application of computer models to estimate soil hydraulic parameters were Whisler and Whatson (1968), who report on estimating the saturated hydraulic conductivity of a draining soil by matching observed and simulated drainage. Since their early endeavors, the application of IM to the estimation of unsaturated flow and transport properties has significantly evolved. In the field of vadose zone hydrology, the parameter-estimation technique usually involves the indirect estimation of the soil water retention and unsaturated soil hydraulic conductivity characteristics by repeated numerical solution of the Richards' equation (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**):

$$C(h_{\mathrm{m}})\frac{\partial h_{\mathrm{m}}}{\partial t} = \nabla \cdot [K(h_{\mathrm{m}})\nabla(h_{\mathrm{m}} + z)] \qquad (1)$$

where $C(h_{\mathrm{m}})$ is the so-called soil water capacity, representing the slope of the soil water retention curve, h_{m} (L) denotes the soil water matric head, t (T^{-1}) represents time, $\boldsymbol{K}$ (LT^{-1}) is the unsaturated soil hydraulic conductivity tensor, and z (L) is the depth included for the vertical flow component only.

In the IM procedure, both the soil water retention and unsaturated soil hydraulic conductivity relationship are described by an analytical model with yet unknown parameter values. Please refer to **Chapter 76, Models for Indirect Estimation of Soil Hydraulic Properties, Volume 2** for information about parametric expressions for the soil hydraulic properties. A transient experiment is performed under controlled conditions with prescribed initial and boundary conditions (*see* **Chapter 75, Determining Soil Hydraulic Properties, Volume 2**). During the experiment, one or more flow-controlled variables are measured. An extensive review on the measurement of flow- and transport processes across a range of spatial and temporal scales is given in **Chapter 72, Measuring Soil Water Content, Volume 2**. Subsequently, the parameters in the hydraulic functions are estimated by minimizing a user-specified measure, containing the difference between the observed and model-predicted flow variables.

To further illustrate the IM procedure, consider Figure 1, which presents a scatter-plot of measured (X,y) data. Also included are three different lines, corresponding to the fixed model structure $\hat{y} = \beta X$, with different values for the parameter β. The aim of the IM procedure is now to estimate that value of the parameter β that provides the best fit to the measured data. Note that, in this simple linear case, analytical solutions for estimating β exist. A more complex problem arises when the fitting model is represented by one or more differential equations, generally containing more than a single unknown parameter value. A typical example of such a problem involves the solution of the Richards' equation augmented with parametric forms of the soil water retention and unsaturated soil hydraulic conductivity function. In this particular case, the parameters can no longer be estimated by visually inspecting the agreement between observations and model predictions, and the search efficiency of computer-based solution algorithms is needed

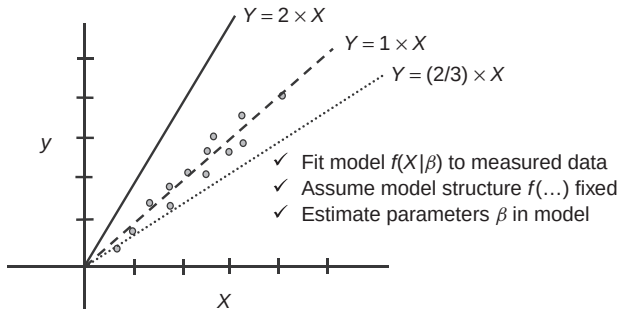


Figure 1 Scatter plot of measured (X, y) data. The three different lines, correspond to the fixed model structure $\hat{y} = \beta X$, with different values for the parameter β. The goal of the IM procedure is to find that value of β which generates the best fit to the measured data. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

to determine the unknown soil hydraulic parameters. The resulting iterative solution of the Richards' equation is in contrast to the direct inversion techniques, which are applicable when using linear models, or the static or steady-state flow experiments for which analytical solutions exist.

In summary, the aim of the IM procedure is to identify the parameter values, such that the difference between the observed and model predicted flow variables is minimized, while accounting for uncertainties in the measured input–output time series (data uncertainty) and uncertainties in the structural ability of the Richards–hydraulic model, to simulate the transient flow process (model uncertainty).

HISTORICAL BACKGROUND

During the last two decades, a great deal of research has been devoted to explore the applicability and suitability of the inverse approach for the identification of soil hydraulic properties. That research has focused primarily on five issues (i) the type of transient experiment and kind of prescribed initial and boundary conditions suited to yield a reliable characterization of the soil hydraulic properties (van Dam *et al.*, 1992, 1994; Ciollaro and Romano, 1995; Santini *et al.*, 1995; Šimůnek and van Genuchten, 1996, 1997; Šimůnek *et al.*, 1998b; Romano and Santini, 1999; Durner *et al.*, 1999; Wildenschild *et al.*, 2001), (ii) the determination of the appropriate quantity and most informative kind of observational data (Zachmann *et al.*, 1981; Kool *et al.*, 1985; Parker *et al.*, 1985; Kool and Parker, 1988; Valiantzas and Kerkides, 1990; Toorman *et al.*, 1992; Eching and Hopmans, 1993; Eching *et al.*, 1994; Durner *et al.*, 1999; Vrugt *et al.*, 2002; among others), (iii) the selection of an appropriate model of the soil hydraulic properties (Zachmann *et al.*, 1982; Russo, 1988; Zurmühl and Durner, 1998), (iv) the adoption and development of Bayesian and multiple-criteria techniques to quantify the uncertainty associated with the inversely estimated parameters (Kool and Parker, 1988; Hollenbeck and Jensen, 1998; Vrugt and Bouten, 2002; Vrugt *et al.*, 2003a), and (v) the construction and weighting of multiple sources of information in an objective function (van Dam *et al.*, 1994; Clausnitzer and Hopmans, 1995; Hollenbeck and Jensen, 1998; Vrugt and Bouten, 2002).

With respect to issue (i), various investigations have reported the usefulness of the evaporation method, multistep outflow (MSO), ponded infiltration, tension infiltrometer experiments, multistep soil water extraction, and the cone penetrometer method. For more information about these different methods and the measurement of flow- and transport processes across a range of spatial and temporal scales in general, please refer to **Chapter 72, Measuring Soil Water Content, Volume 2**. With regard to the information content of the measurements (issue ii), work reported by Kool *et al.* (1985) and Kool and Parker (1988) suggested that the transient experiments should cover a wide range in water contents (θ), and preferably include tensiometer measurements within the soil sample. The benefit of using tensiometer data, simultaneously with outflow measurements in the IM procedure is intuitively clear, as the optimized soil water retention curve is then forced to match the observed $\theta(h_m)$ data (Eching and Hopmans, 1993). Additionally, in the case of outflow experiments, van Dam *et al.* (1994) demonstrated that more reliable parameter estimates can be obtained by incrementally increasing the pneumatic pressure in several steps, instead of using a single pressure increment. From all transient experiments, the evaporation and an MSO method are most popular for inversely estimating the water retention and unsaturated soil hydraulic conductivity characteristics.

Throughout the remaining part of this article, we use, therefore, a classical transient laboratory outflow experiment to illustrate the various aspects of the IM procedure for estimating the soil hydraulic parameters. Observed outflow measurements combined with soil water pressure head data within the soil core were taken from Wildenschild *et al.* (2001). We revisited the data for the sandy (Lincoln) soil, consisting of an MSO experiment with 5 consecutive pressure steps (0-25-35-62-80-100 mbar). Figure 2 presents the observed outflow data, as well as the measured soil water pressure heads within the soil sample. The soil hydraulic properties (issue iii) are described with the closed-form parametric expressions of Mualem-van Genuchten (Mualem, 1976; van Genuchten, 1980; MVG). The unsaturated soil hydraulic parameters (θ_r, α, n, K_s, and l; where θ_r is the residual water content, α and n are curve fitting parameters for the water retention curve, K_s is the saturated hydraulic conductivity, and l is a fitting parameter for the hydraulic conductivity function) were inversely estimated using the HYDRUS-1D model (Šimůnek *et al.*,

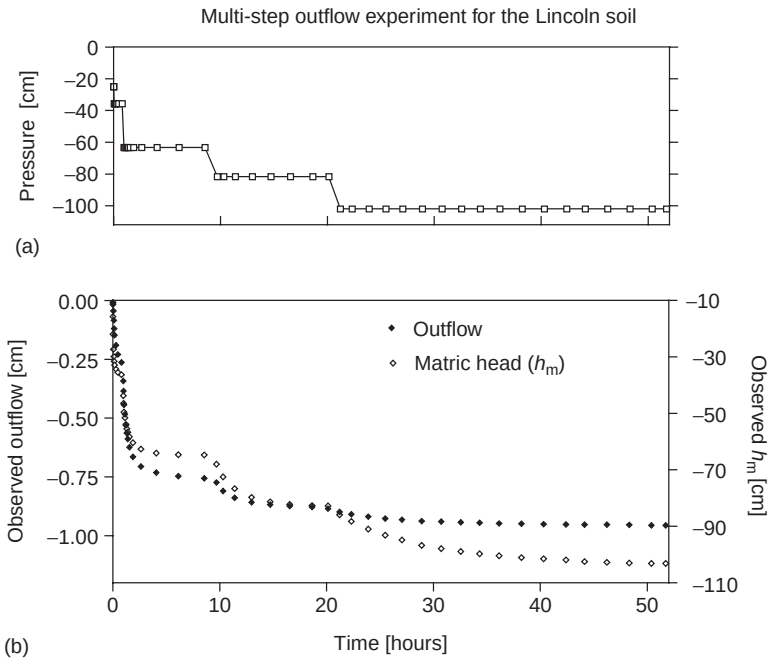


Figure 2 Example data of a transient outflow experiment which is extensively used throughout this article: (a) augmented pressure steps during the MSO experiment for the sandy (Lincoln) soil, (b) observed cumulative outflow and pressure head values as time develops within the soil sample during the transient experiment

Table 1 Prior parameter ranges of the MVG parameters for the Lincoln soil used throughout this article

Parameter	Unit	Min	Max
θ_r	[cm^3 cm^{-3}]	0.00	0.25
α	[cm^{-1}]	$4.5 \cdot 10^{-2}$	0.05
N	[−]	1.50	6.00
K_s	[cm h^{-1}]	0.25	3.00
L	[−]	−3.00	3.00

1998a). This model numerically solves the one-dimensional form of the Richards's equation, depicted in equation (1), using a Galerkin-type linear finite element scheme. The saturated water content (θ_s) of the Lincoln soil was fixed at its measured value of 0.37 m^3 m^{-3}. The prior ranges for each of the unsaturated soil hydraulic parameters are defined in Table 1. For more information about the experimental setup and soil, please refer to Wildenschild *et al.* (2001).

Although laboratory experiments have the advantage of being quick and precise, they do not necessarily lead to soil physical properties that can be considered representative for a field situation. Among the first to apply the IM approach to a field situation were Dane and Hruska (1983) who estimated soil physical parameters using transient drainage data. The application of IM to estimate soil hydraulic properties across spatial scales is very promising, yielding effective hydraulic properties that pertain to the scale of interest (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**). Sections "Confronting parameter uncertainty" and "Advances in inverse modelling" of this article extensively address issues of how to infer parameter uncertainty from IM studies (issue iv), and how to weigh multiple sources of information in one objective function (issue v). Finally, in Section "Closing comments", we briefly discuss the limitations of current IM/parameter estimation strategies and point to a more generic framework to account for input, output, parameter, and model structural uncertainty.

MATHEMATICAL DEVELOPMENT

To facilitate the description of the parameter estimation or IM procedure, let us write the combined Richards-hydraulic model f as

$$\hat{y} = f(X|\beta) + e, \tag{2}$$

where $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ is an $m \times 1$ vector of model predictions, $X = (X_{11}, X_{12}, \ldots, X_{mr})$ is an $m \times r$ matrix of input variables, $\beta = (\beta_1, \beta_2, \ldots, \beta_r)$ denotes the vector of r unknown parameters, and e is a vector of statistically independent errors with zero expectation and constant variance σ^2.

We assume that

$$\beta \in B \subseteq \Re^r \tag{3}$$

where $\Re^r$ denotes the r-dimensional Euclidean space. If B is not the entire domain space $\Re^r$, the problem is said to

be constrained. Given observed output values of $\boldsymbol{y}$, let us denote the residual errors from the predictions by

$$E(\beta) = G[\hat{y}(\beta)] - G[\boldsymbol{y}] = \{e_1(\beta), e_2(\beta), \ldots, e_m(\beta)\} \tag{4}$$

where the function $G(\cdot)$ allows for linear or nonlinear transformations of the simulated and observed data. The IM approach now relies on the estimation of the set of parameters β such that the measure E, commonly called *the objective function*, is in some sense forced to be as close to zero as possible. The development of a measure $E(\beta)$ that mathematically measures the "size" of $E(\beta)$ is typically based on assumptions regarding the distributions of the measurement errors presented in the data. By far the most popular measure for E is the simple least square (SLS) or maximum likelihood estimator, appropriate when the measurement errors are believed to be homoscedastic and uncorrelated. The SLS is expressed as

$$\underset{\beta}{\text{minimize}} \quad \text{SLS} = \sum_{j=1}^{m} e_j(\beta)^2 \tag{5}$$

Under these circumstances, Hollenbeck and Jensen (1998) stretched the importance of model adequacy before sound statements can be made about the final parameter estimates and their uncertainty. Model adequacy is determined from

$$p_{\text{adeq}} = 1 - Q\left(\frac{\text{SLS}}{\sigma_T^2}, m - r\right) \tag{6}$$

where σ_T denotes the error deviation of the measurements and $Q(\cdot)$ is the χ^2 cumulative distribution with $(m - r)$ degrees of freedom. Having homoscedastic and uncorrelated error residuals, the adequacy test gives us a measure of how well the optimized model fits the observations relative to their measurement precision. Using the definition of equation (6), models are adequate if $p_{\text{adeq}} > 0.5$, since then SLS $< m\sigma_T^2$. The validity and robustness of the standard SLS estimator has, however, to be questioned when significant outliers are present in the measurements (Finsterle and Najita, 1998). For these cases, more complex estimators can be used, which explicitly account for correlated and heteroscedastic error residuals (Sorooshian and Dracup, 1980).

FREQUENTIST VERSUS BAYESIAN APPROACH

While the Frequentist or classical statistics considers the model parameters β in equation (2) to be fixed but unknown, the Bayesian statistics treat them as probabilistic variables having a joint posterior probability density function (pdf), which captures the probabilistic beliefs about the parameters β in light of the observed data $\boldsymbol{y}$. The advantage of using Bayesian estimators rather than simple least-squares estimators for the problem of finding unsaturated soil hydraulic parameters has recently been discussed by Hollenbeck and Jensen (1998). The posterior pdf $p(\beta|\boldsymbol{y})$ is proportional to the product of the likelihood function and the prior pdf. The prior pdf with probability density (or mass) function $p(\beta)$ summarizes information about β before any data are collected. This prior information usually consists of realistic lower and upper bounds on each of the parameters, thereby defining the feasible parameter space, B, and imposing a uniform (noninformative) prior distribution on this rectangle.

Assuming that the residuals are mutually independent, each having the exponential power density, the likelihood of a parameter set β for describing the observed data $\boldsymbol{y}$ can be computed using Box and Tiao (1973)

$$L(\beta|\boldsymbol{y}, \gamma) = \left[\frac{\omega(\gamma)}{\sigma}\right]^m \exp\left[-c(\gamma)\sum_{j=1}^{m}\left|\frac{e_j(\beta)}{\sigma}\right|^{(2/d(1+\gamma))}\right] \tag{7}$$

where

$$\omega(\gamma) = \frac{\{\Gamma[3(1+\gamma)/2]\}^{1/2}}{(1+\gamma)\{\Gamma[(1+\gamma)/2]\}^{3/2}}$$
$$c(\gamma) = \left\{\frac{\Gamma[3(1+\gamma)/2]}{\Gamma[(1+\gamma)/2]}\right\}^{(1/(1+\gamma))} \tag{8}$$

in which Γ denotes the gamma operator. The parameter γ specifies the error model of the residuals. The residuals are assumed normally distributed when $\gamma = 0$, double exponential when $\gamma = 1$, and tend to a uniform distribution as $\gamma \to -1$. Assuming a noninformative prior of the form $p(\beta, \sigma|\gamma) \propto \sigma^{-1}$, Box and Tiao (1973) showed that the influence of σ can be integrated out of the likelihood function in equation (7), leading to the following form of the posterior density of β,

$$p(\beta|\boldsymbol{y}, \gamma) \propto [M(\beta)]^{-m(1+\gamma)/2} \tag{9}$$

where

$$M(\beta) = \sum_{j=1}^{m}\left|e_j(\beta)\right|^{2/(1+\gamma)} \tag{10}$$

For $\gamma = 0$, notice that M is simply the familiar "sum of squared residuals" function previously denoted in equation (5). For more information about the Bayesian inference scheme, please refer to Box and Tiao (1973).

MANUAL VERSUS AUTOMATIC SOLUTION ALGORITHMS

In the process of parameter estimation or model calibration, the hydrologist adjusts the values of the model parameters such that the model is able to closely match the behavior of the real system it is intended to represent. In its most elementary form, this calibration is performed by manually adjusting the parameters while visually inspecting the agreement between observations and model predictions (Dane and Hruska, 1983; Janssen and Heuberger, 1995). In this approach, the "closeness" of the model with the measurements is typically evaluated in terms of several subjective visual measures, and a semiintuitive trial-and-error process is used to perform the parameter adjustments (Boyle *et al.*, 2000). Because of the subjectivity and time-consuming nature of manual trial-and-error calibration, there has been a great deal of research into the development of automatic methods for calibration of hydrologic models (e.g., Gupta and Sorooshian, 1994). Automatic methods seek to take advantage of the speed and power of digital computers, while being objective and relatively easy to implement.

Automatic solution algorithms that have been developed in the past to solve the nonlinear SLS optimization problem, stated in equation (5), may be classified as local search methodologies, when seeking for systematic improvement of the objective function using an iterative search starting from a single arbitrary initial point in the parameter space or as global search methods in which multiple concurrent searches from different starting points are conducted within the parameter space.

One of the simplest local-search optimization methods, which is commonly used in the field of soil hydrology, is a Gauss–Newton type of derivative-based search

$$\beta^{(k+1)} = \beta^{(k)} - H(\beta^{(k)})^{-1}(\nabla f(\beta^{(k)}))^T \qquad (11)$$

where $\beta^{(k+1)}$ is the updated parameter set and $\nabla f(\beta^{(k)})$ and $H(\beta^{(k)})$ denote the gradient and Hessian matrix respectively evaluated at $\beta = \beta^{(k)}$

$$\nabla f(\beta) = \left(\frac{\partial f}{\partial \beta_1}(\beta) \ldots, \frac{\partial f}{\partial \beta_r}(\beta)\right)$$
$$H(\beta) = \left(\frac{\partial^2 f}{\partial \beta_i \partial \beta_j}(\beta)\right)_{i,j} \qquad (12)$$

From an initial first guess of the parameters $\beta^{(0)}$, a sequence of parameter sets, $\{\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(k+1)}\}$, is generated that is intended to converge to the global minimum of $E(\beta)$ in the parameter space. If $f(X|\beta)$ depends linearly on each parameter $\beta_j (j = 1, 2, \ldots, r)$, the minimization problem stated in equation (6) reduces to a Linear Regression problem for which analytical solutions exist. Doherty (1994) and Clausnitzer and Hopmans (1995) presented general local optimization codes, which can be coupled with any parameter estimation problem.

The derivative-based search defined in equations (11) and (12) will evolve towards the global minimum in the parameter space in situations where the objective function exhibits a topographical convex shape in the entire parameter domain. However, practical experience with hydrologic models suggests that the objective function seldom satisfies these restrictive conditions. To illustrate this problem, consider Figure 3 (taken from Vrugt and Bouten, 2002), which presents posterior marginal distributions (histograms) for two hydraulic parameters in the global minimum of the parameter space, derived using measured outflow dynamics from a transient one-step outflow experiment. If the objective function would exhibit a convex shape in the entire parameter domain, the histograms for each of the soil hydraulic parameters would exhibit a clear Gaussian distribution with a single mode. However, the large number of

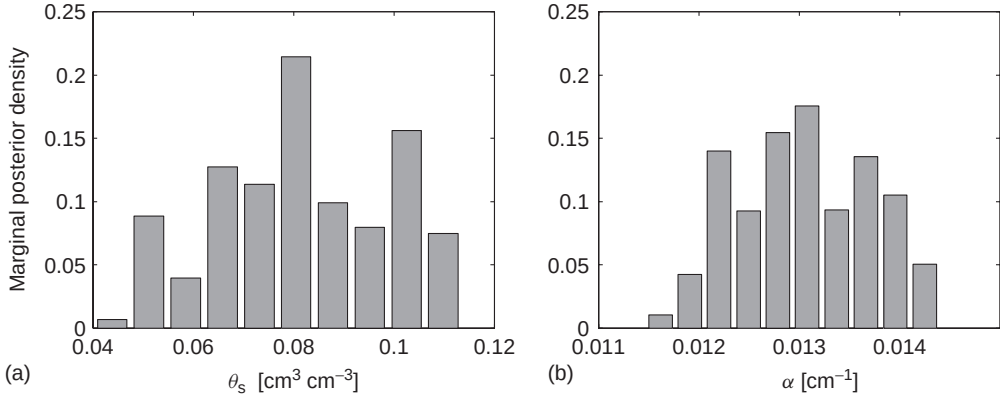


Figure 3 Marginal posterior probability density distributions for the soil hydraulic parameters (a) θ_r, and (b) α, in the MVG model using observed outflow dynamics during a transient one-step outflow experiment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

different modes (local minima) for each of the parameters depicted in Figure 3 is the most probable reason for the numerous reports in the literature of the inability to find a unique set of hydraulic parameter values using observed outflow dynamics from MSO experiments (Kool *et al.*, 1985; Parker *et al.*, 1985; Toorman *et al.*, 1992; van Dam *et al.*, 1992; among others). As the local gradient-based search algorithms are not designed to handle the peculiarities of the response surface illustrated in Figure 3, they will terminate their search prematurely with their final solution essentially being dependent on the starting point in the parameter space. Another emerging problem, reported by Hopmans *et al.* (2002), is that some of the hydraulic parameters are typically highly correlated, further lowering the chance of getting a single unique solution with local-search methodologies.

The existence of nonuniqueness problems with local-search methodologies has led soil hydrologists to argue that there is not sufficient information in the conducted measurements to enable a unique characterization of the soil hydraulic properties. So even with laboratory experiments, where a soil can be manipulated with great flexibility, nonuniqueness problems are often experienced (van Dam *et al.*, 1992). Seemingly, there was a widespread conviction that the best way to solve the nonuniqueness problem would be to add more and better measurements (Eching and Hopmans, 1993; van Dam *et al.*, 1994). However, research into data requirements has led to the understanding that the information content of the data is far more important than the amount of data used for the identification of the model parameters (Sorooshian *et al.*, 1983; Gupta and Sorooshian, 1985; Yapo *et al.*, 1996; Vrugt *et al.*, 2002). Vrugt *et al.* (2003a) argued, therefore, that it is not the information in the measurements that is lacking to obtain a unique set of parameters, but the fact that the classical local-search optimization procedures utilized in soil hydrology are typically not powerful enough to solve the problems illustrated in Figure 3. Their arguments on uniqueness of parameters are not based on the convergence problems of the applied optimization methods, but on the shape of the posterior marginal distributions of the parameters. Hence, closer inspection of Figure 3 demonstrates that for each of the two parameters there is a single optimum with highest posterior probability in the global minimum. Indeed, although not shown here, the multivariate probability distribution of the parameters confirms that these regions with highest posterior probability for each parameter coincide. In other words, the uniqueness of parameter estimates is inferred from the shape of the univariate posterior probability distributions of the parameters. In particular, ill-posedness of an inverse problem due to data-noninformativeness issues, as demonstrated by Zijlstra and Dane (1996), would result in large uncertainties associated with the final parameter estimates. Thus, an optimization algorithm that successfully identifies the multivariate joint probability distribution of the parameters in the vicinity of the global minimum has desirable properties, since it not only provides useful information about the interactions of the parameters in the full parameter space, but also provides useful information about the quality of the data. Hence, ultimately, data quality still determines the reliability of the final parameter estimates.

In the development of suitable automatic calibration approaches, we must consider the fact that the hydrologic model optimization problem suffers from the existence of multiple optima in the parameter space (with both small and large domains of attraction), discontinuous first derivatives, and curving multidimensional ridges. These considerations inspired Duan *et al.* (1992), in a related field (surface hydrology) to develop a powerful, robust, and efficient global-optimization procedure, entitled, the Shuffled Complex Evolution (SCE-UA) global optimization algorithm, developed at the University of Arizona. By merging the strengths of the Downhill Simplex procedure (Nelder and Mead, 1965) with the concepts of controlled random search, systematic evolution of points in the direction of global improvement, competitive evolution (Holland, 1975), and complex shuffling, the SCE-UA algorithm represents a synthesis of the best features of several optimization strategies. The strength and reliability of the SCE-UA global optimization algorithm have since been independently tested and proven by numerous researchers, and the algorithm is now extensively used worldwide (e.g. Duan *et al.*, 1992, 1993; Sorooshian *et al.*, 1993; Luce and Cundy, 1994; Tanakamaru, 1995; Gan and Biftu, 1996; Kuczera, 1997; Hogue *et al.*, 2000; Boyle *et al.*, 2000; among many others). Application of the SCE-UA algorithm to the inverse estimation of soil hydraulic parameters has, to the authors' knowledge, been limited to the work by Vrugt and Bouten (2002) and Mertens *et al.* (2005). Other global optimization methods that have been used for the inverse estimation of soil hydraulic properties are the Annealing-simplex methods (Pan and Wu, 1998), Genetic algorithms (Takeshita, 1999; Vrugt *et al.*, 2001a), Grid sampling strategies (Abbaspour *et al.*, 1997), and Ant-colony methods (Abbaspour *et al.*, 2001). Notwithstanding their successful application, as compared with the SCE-UA algorithm, these global search algorithms are computationally very demanding, requiring a substantial number of model evaluations to converge to the global solution. With the current available SCE-UA algorithm, one can now have confidence that the global minimum of a predefined objective function is found. For more information about the current state-of-the-art in calibration in surface hydrology, the reader is referred to the work presented in Duan *et al.* (2003).

To demonstrate the features and applicability of the SCE-UA optimization procedure for estimating the soil hydraulic properties, consider Figure 4. It shows the evolution of the

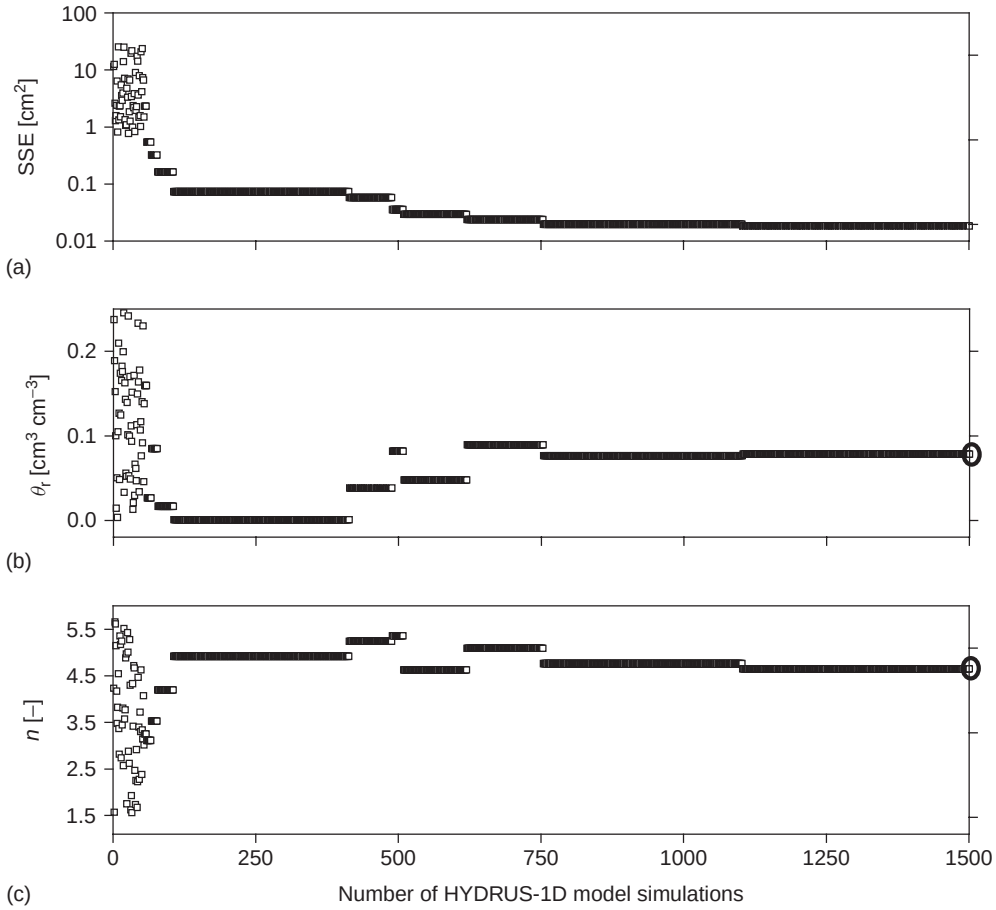


Figure 4 Evolution of the SLS objective function value and the best parameter values found so far with the SCE-UA algorithm as a function of the number of HYDRUS-1D model evaluations: (a) SLS-measure, (b) θ_r, and (c) α parameter. The circles indicate the best parameter values found with a traditional local-search methodology

best parameters as a function of the number of HYDRUS-1D model simulations for the MSO outflow experiment of the Lincoln soil. The objective function was the SLS estimator, defined in equation (5), containing only the residuals of the observed and predicted outflow. For the sake of brevity, we only display the results for the parameters θ_r and n. The circles indicate the best parameter values found with a traditional local-search methodology (Levenberg-Marquardt) using repeated optimizations with different starting points. The initial population of parameter samples used with the SCE-UA algorithm reflects the initial parameter uncertainty. After processing information about the shape of the response surface with the optimization algorithm, the best parameter values jump back and forth through the feasible parameter space. Finally, after approximately one thousand HYDRUS-1D model evaluations, the algorithm settles down in the most promising region of the parameter space (also referred to as global minimum), and the "best" parameter set, which minimizes the SLS criterion within this region is identified.

CONFRONTING PARAMETER UNCERTAINTY

Parameter estimates obtained using IM are generally error-prone because the data used for their estimation contain measurement errors and because the Richard's type of model never perfectly represents the actual flow and transport processes. Confidence intervals on the estimated hydraulic model parameters can be used to express the degree of uncertainty in the estimated values. Durner *et al.* (1999) recently pointed out that the improved interpretation of parameter uncertainty can yield valuable information to enable a better judgment of the limits of our theoretical understanding of unsaturated water flow in soils.

Traditional First-Order Approximation

A frequently employed approach, which is particularly popular in the area of vadose zone hydrology, is to obtain confidence intervals of parameters by utilizing a first-order approximation to the model f, previously defined in equation (2), near its minimum (see Carrera and Neumann,

1986 in groundwater hydrology, Kool and Parker, 1988 in unsaturated soil water flow, and Kuczera and Parent, 1998 in rainfall-runoff modelling). The estimated posterior distribution of β is then expressed as

$$p(\beta|\boldsymbol{y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\beta-\beta_{\text{opt}})^T\ J^T\ J(\beta-\beta_{\text{opt}})\right] \quad (13)$$

where J is the Jacobian or sensitivity matrix evaluated at $\beta = \beta_{\text{opt}}$. All relevant inferences about β can be made from the knowledge that the posterior distribution of β is the multivariate normal distribution $N_p\ [\beta_{\text{opt}}, \sigma^2\sqrt{(J^T J)^{-1}}]$. Hence, the marginal posterior distribution of β_i is the multivariate normal distribution, $N_i\ [\beta_{i,\text{opt}}, \sigma^2 \sum_{ii}]$, where $\sum_{ii}$ is the ith diagonal element of $\sqrt{(J^T J)^{-1}}$ (Carrera and Neumann, 1986; Kool and Parker, 1988).

If the hydrologic model is linear (or very nearly linear) in its parameters, the posterior probability region estimated by equation 13 will give a good approximation of the actual parameter uncertainty. However, for nonlinear models (e.g., soil hydrological models), with strong parameter interdependence, this approximation can be quite poor (Kuczera and Parent, 1998; Vrugt and Bouten, 2002). Besides exhibiting strong and nonlinear parameter interdependence, the surface of the posterior parameter distribution $p(\beta|\boldsymbol{y})$ can deviate significantly from the multinormal distribution. It may also have multiple local optima and discontinuous derivatives (Duan *et al.*, 1992). Moreover, the ellipsoid region, defined by equation (13) may represent a very poor approximation of parameter uncertainty, as for example, in the case of a strongly hyperbolic banana-shaped curvature in the $p(\beta|\boldsymbol{y})$ surface.

Uniform Grid Sampling – Response Surface Analysis

Alternatively, a robust but computationally intensive method for the calculation of confidence intervals, contrasting the classical first-order approach, involves the generation of contour plots (Toorman *et al.*, 1992; Gribb, 1996; Romano and Santini, 1999; Vrugt *et al.*, 2001c). It requires exhaustive discretizing of the parameter space and computing the objective function for each grid point, which is a rather primitive approach. Recently, Abbaspour *et al.* (1997, 1999) used a similar kind of sampling strategy, entitled the Sequential Uncertainty Fitting algorithm (SUFI) for estimating subsurface flow and transport parameters. However, as simple as this approach might be, it requires massive computing resources for highly dimensioned parameter spaces. For example, if we wish to sample, on average, a parameter within a hypercube with resolution equal to one tenth of the parameter range in a 6-parameter space, we need 10^6 model runs. Moreover, failure to maintain an adequate sampling density may result in undersampling or too course sampling in certain regions of the parameter space. Consequently, this may lead to errors in the computed parameter confidence intervals.

A simplified two-dimensional approach using this uniform grid sampling strategy is often used in soil hydrology to study the shape of the objective function in the vicinity of the global optimum. These contour plots or response surfaces are obtained by contouring the objective function, previously defined in equation (5), for two selected parameters, while keeping the additional hydraulic parameters constant at their most optimal values. These response surfaces reveal the occurrence of local minima, the presence of a well-defined global minimum, the parameter sensitivity, and parameter correlations. If response surfaces do not display a well-defined global minimum in the two-dimensional parameter planes, the conventional inverse parameter estimation technique may certainly be expected to be unsuccessful in a multidimensional plane. Excellent examples of the use of response surfaces are presented in Toorman *et al.* (1992) and Šimůnek *et al.* (1998b). However, by limiting the number of parameters within a single response surface analysis, the behavior of the objective function in the different parameter planes can only suggest how the objective function might behave in the full parameter space. For example, local minima of the objective function could exist, but not appear in the cross-sectional planes (Šimůnek and van Genuchten, 1996). In the next section, we will describe a method that is especially designed to quantify the behavior of the objective function in the full parameter space, thereby simultaneously generating a description of the multivariate joint interaction between the soil hydraulic parameters.

To illustrate the usefulness and insights response surfaces can offer with respect to parameter identifiability, we generated contour plots of the SLS objective function in the vicinity of the most-likely parameter value using the outflow data of the Lincoln soil. The response surfaces were calculated on a rectangular grid. Each parameter domain, was discretized into 4 equidistant discrete points, with domain ranges as specified in Table 1, resulting in 1600 grid points for each response surface. Interpolation between points was done using MATLAB (MATLAB, 1999). The four, selected response surfaces, presented in Figure 5 are a representative set of the total of 10 possible response surfaces. The location of the optimum in the parameter space, previously identified using the SCE-UA algorithm, is indicated with a circle in each of the Figures. Most of the 10 response surfaces showed relatively well-defined global minima and no additional local minima. However, the shape of the contour lines in Figures 5b and d reveal the existence of a positive correlation structure between the parameters $\theta_r - n$, and $K_s - l$ in the HYDRUS-1D model. In other words, if θ_r or K_s is perturbed positively, then n or l will also be perturbed positively (Toorman *et al.*, 1992). So, response surface analyses reveal very useful information about parameter correlation and parameter identifiability or

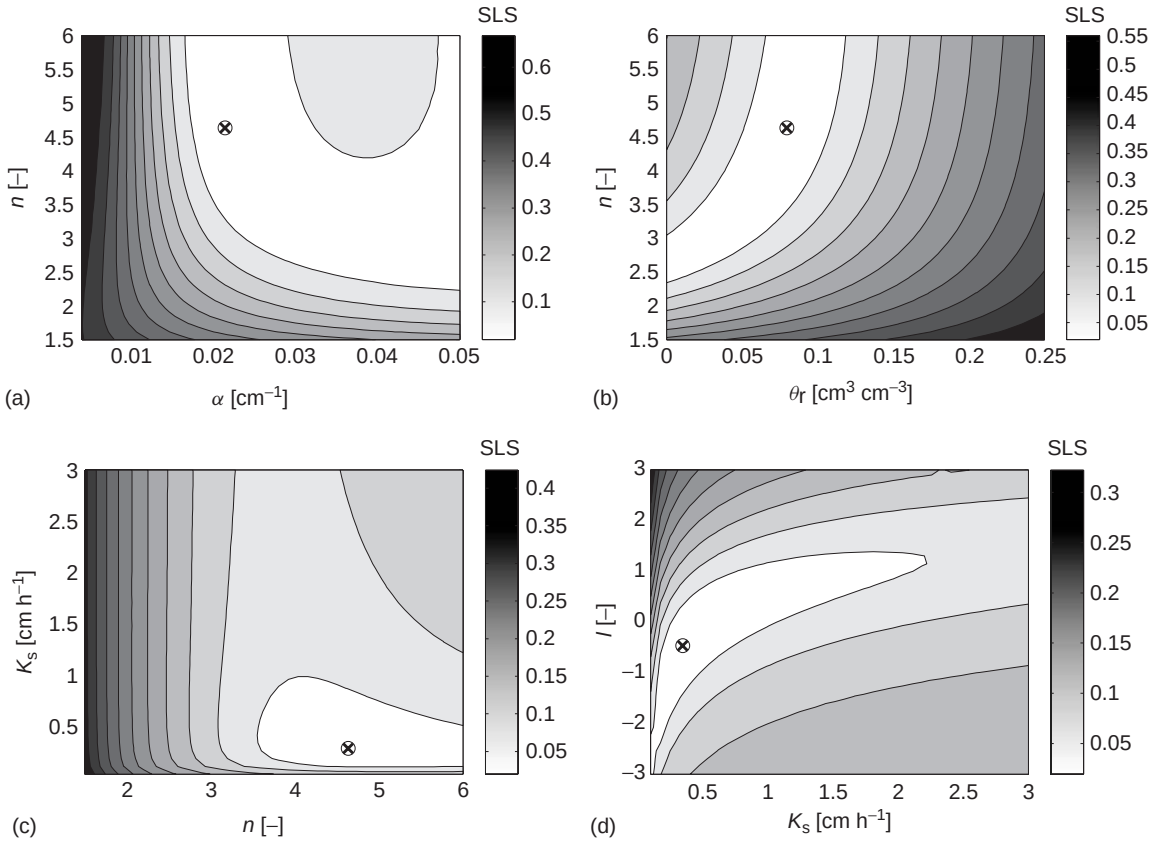


Figure 5 Contours of the square root of the SLS objective function, defined in equation (5) in the (a) $\alpha - n$, (b), $\theta_r - n$, (c), $n - K_s$, and (d) $K_s - l$ sliced parameter planes using the outflow data of the Lincoln soil. The SCE-UA identified ''best'' parameter values are indicated with circles in each of the graphs

sensitivity in the vicinity of the global optimum. However, as pointed out before, the method is computationally very demanding, as for each response surface a large number of simulations are needed, and it is graphically limited to two-dimensions.

Monte Carlo Sampling of Posterior Distribution

The Markov Chain Monte Carlo (MCMC) method for assessing parameter confidence intervals in nonlinear models is based on the idea that instead of explicitly computing the probability distribution, $p(\beta|\boldsymbol{y})$, as done in equation (13), it is sufficient to approximate the form of the density by drawing a large random sample from $p(\beta|\boldsymbol{y})$ (Mosegaard and Tarantola, 1995). Diagnostic measures of central tendency and dispersion of the posterior distribution can be estimated by computing the mean and standard deviation of the sample. This directly leads to the question of how to efficiently sample from $p(\beta|\boldsymbol{y})$. To address this question, Vrugt *et al.* (2003b) have developed a new algorithm, which merges the sampling strategy of the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970) with the strengths and efficiency of the SCE-UA population evolution method.

The goal of the original SCE-UA algorithm (Duan *et al.*, 1992) was to find a single best parameter set in the feasible space. The SCE-UA begins with a random sample of points distributed throughout the (bounded) feasible parameter space, and uses an adaptation of the Downhill Simplex search strategy to continuously evolve the population toward better solutions in the search space, progressively relinquishing occupation of regions with lower posterior probability. This genetic drift, where the members of the population drift towards a single location in the parameter space (i.e., the mode of $p(\beta|\boldsymbol{y})$), is typical of many evolutionary search algorithms. By replacing the adapted Downhill Simplex strategy with a MH strategy, the tendency of the algorithm to collapse into the relatively small region containing the "best" parameter set is avoided. The new algorithm, entitled the Shuffled Complex Evolution Metropolis (SCEM-UA) global optimization algorithm was developed in collaboration between the University of Arizona and the University of Amsterdam, and merges the strengths of the MH algorithm, controlled random search, competitive evolution, and complex shuffling. The stochastic nature of the MH annealing scheme avoids the tendency of the SCE-UA algorithm to collapse into a single region of

attraction (local minima), while the information exchange (shuffling) between parallel sequences allows the search to be biased in favor of better regions of the parameter space. The SCEM-UA is therefore able to infer ***both*** the most-likely parameter set and its underlying posterior probability distribution within a single optimization run.

In summary, the SCEM-UA algorithm begins with an initial population of points (parameter sets) randomly distributed throughout the feasible parameter space. For each parameter set, the posterior density is computed using a Bayesian inference scheme, such as the one presented in Section "Frequentist versus bayesian approach" of this article. The population is partitioned into complexes, and in each complex a parallel sequence is launched from the point that exhibits the highest posterior density. A new candidate point, in each sequence, is generated using a multivariate normal distribution, either centered around the current draw of the sequence or the mean of the points in complex, augmented with the covariance structure induced between the points in complex. The Metropolis-annealing (Metropolis *et al.*, 1953) criterion is used to test whether the candidate point should be added to the current sequence. Subsequently, the new candidate point randomly replaces an existing member of the complex. Finally, after a certain number of iterations, new complexes are formed through a process of shuffling. This series of operations results in a robust MCMC sampler that conducts a robust and efficient search of the parameter space. Diagnostic measures of parameter interaction and parameter identifiability can be estimated using the joint multivariate posterior distribution of the model parameters as inferred with the SCEM-UA algorithm. For more information about the SCEM-UA algorithm, please refer to Vrugt *et al.* (2003b).

The applicability and power of the SCEM-UA algorithm for estimating the unsaturated soil hydraulic parameters (θ_r, α, n, K_s, and l) from transient MSO experiments will now be illustrated. The stationary posterior distribution, corresponding to the density criterion specified in equation (9), was estimated with the SCEM-UA algorithm using observed outflow data for the Lincoln soil in combination with 10 000 HYDRUS-1D model evaluations. The most important results of this analysis are summarized in Figures 6, 7, 8, and 9.

Figure 6 presents the marginal posterior probability distributions for the parameters θ_r, α, n, K_s, and l in the MVG model. These histograms were derived from the SCEM-UA samples that were generated after convergence had been achieved to a stationary posterior distribution. The values

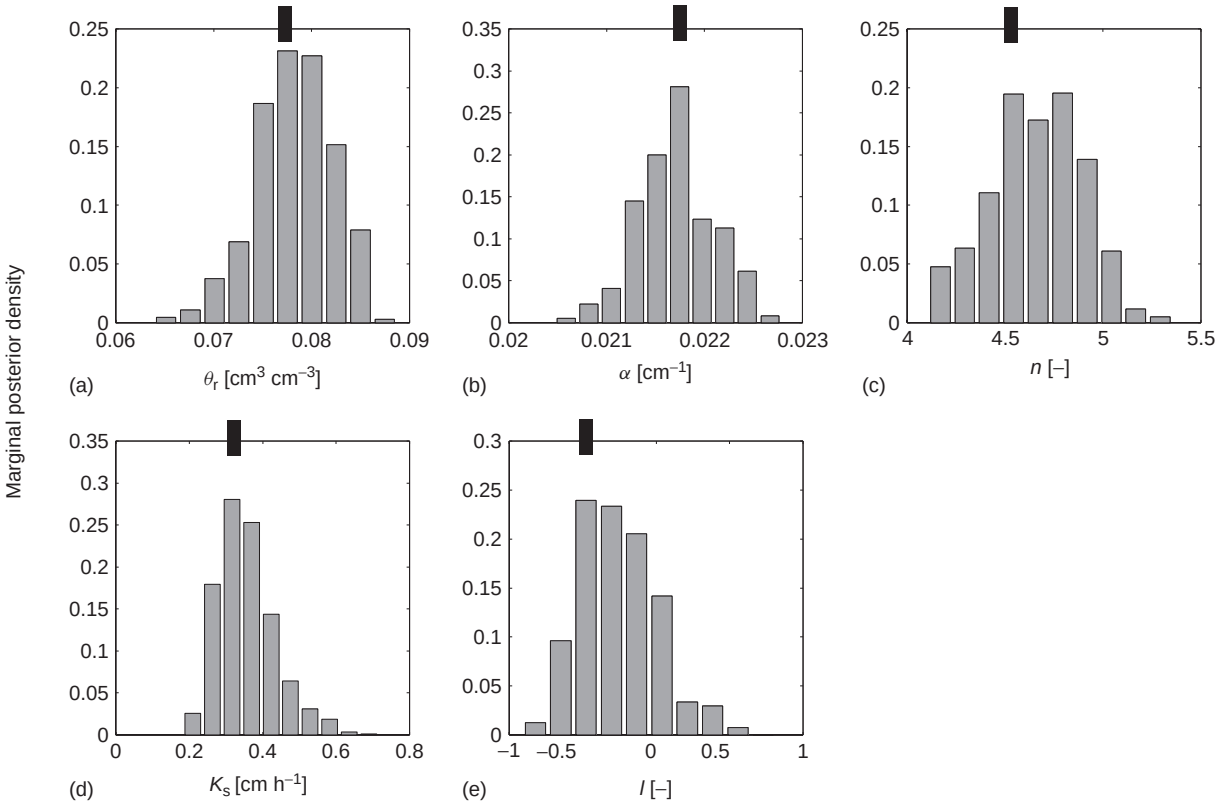


Figure 6 Marginal posterior probability distributions of the MVG parameters: (a) θ_r, (b) α, (c) n, (d) K_s, and (e), l. The SCE-UA estimated SLS parameter values are separately indicated with the squared symbols on top of each of the graphs. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

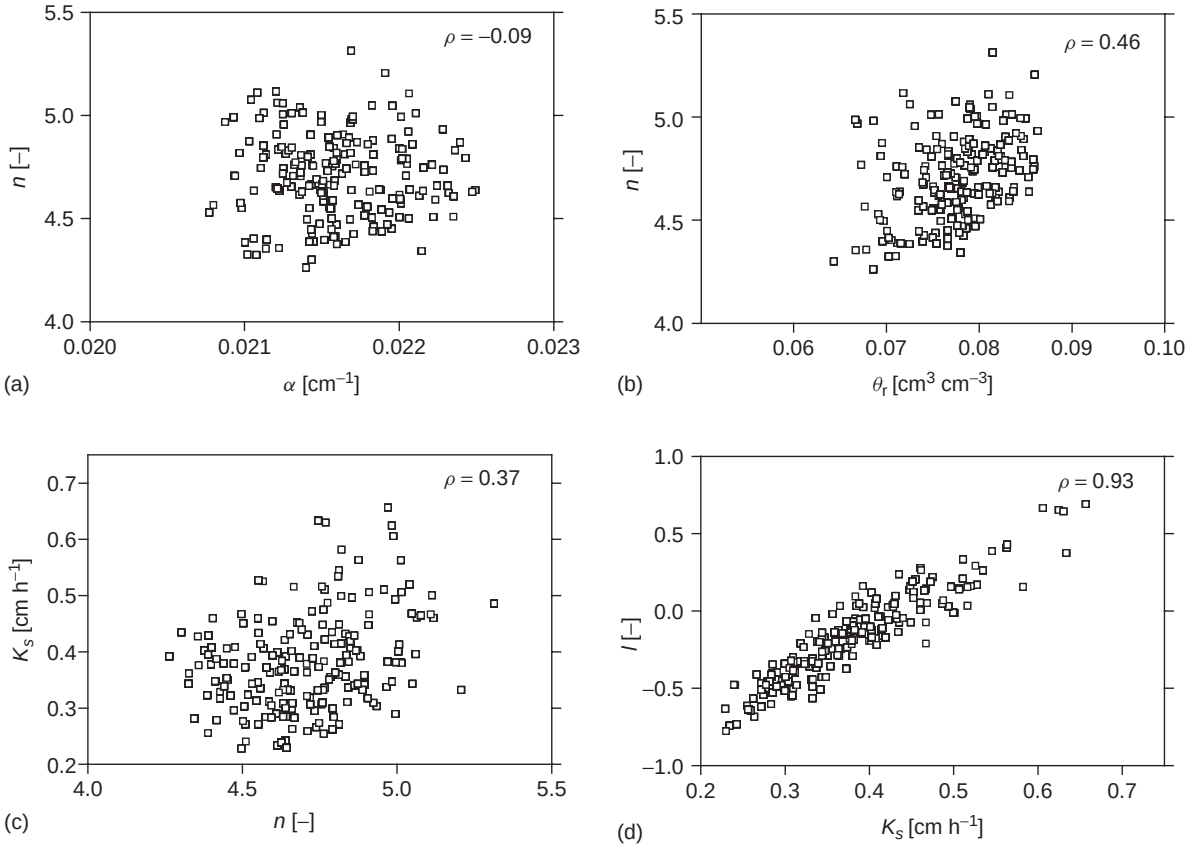


Figure 7 Scatter plots of 4000 combinations of (a) $\alpha - n$, (b), $\theta_r - n$, (c), $n - K_s$, and (d) $K_s - l$ parameters, sampled for the MSO experiment using the SCEM-UA algorithm

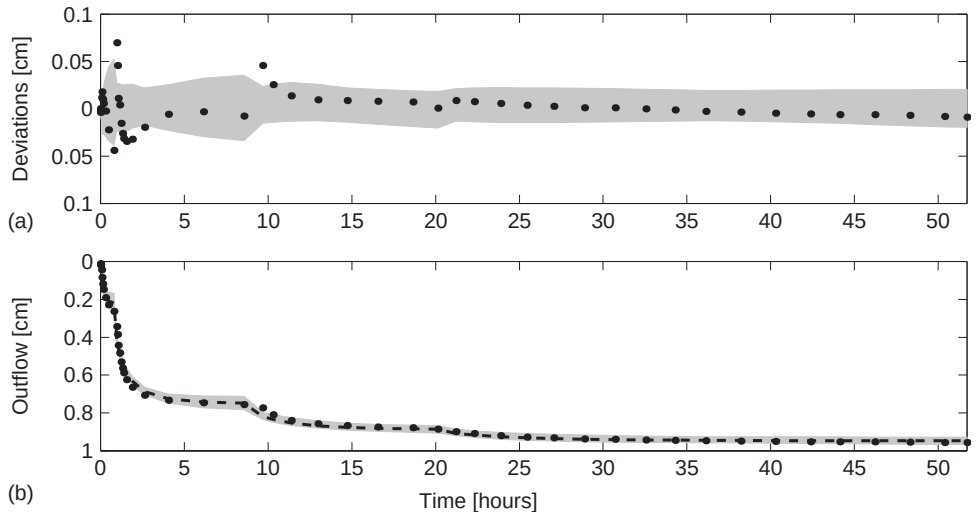


Figure 8 (a) Uncertainties in the simulated outflow associated with the most probable parameter set of the MVG model. The light-gray region denotes the residual parameter uncertainty. (b) Outflow prediction uncertainty ranges of the HYDRUS-1D model associated with the uncertainty in the parameter estimates. The black dots denote the observed outflow measurements of the Lincoln soil. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

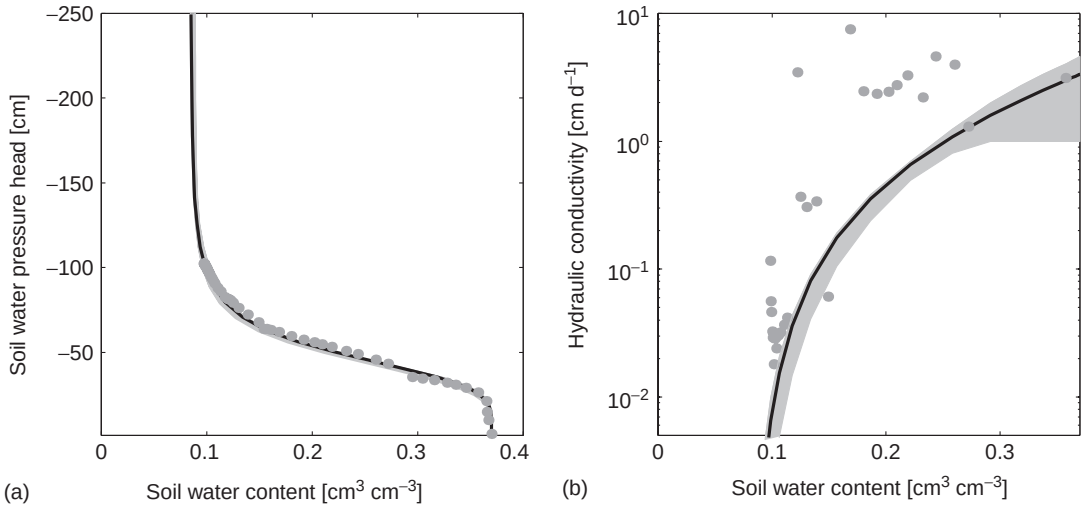


Figure 9 Prediction uncertainty intervals for the retention and hydraulic conductivity curves associated with the uncertainty in the parameter estimates for the MSO outflow experiment using the MVG parametric model. The black dots corresponds to the directly estimated retention and hydraulic conductivity points, whereas the dark lines resulted from the most-likely parameter sets. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of the most-likely parameter estimates, previously identified with the SCE-UA algorithm, are separately indicated with squared symbols in each of the graphs. While the parameters θ_r (Figure 6a), α (Figure 6b), are reasonable well described by the multinormal distribution, as outlined in Section "Traditional first-order approximation", the remaining unsaturated soil hydraulic parameters n, K_s, and l are better described with bimodal or lognormal distributions respectively. Clearly, the SCEM-UA algorithm has many desirable features as it not only correctly infers the most-likely parameter set but also generates useful information about the shape of the response surface in the vicinity of the optimum, without having to do too many simulations.

To allow comparison between the SCEM-UA sampled parameter space and the response surfaces depicted in Figure 5, consider the scatter plots (Figure 7) of the sampled $\alpha - n$ (a), $\theta_r - n$ (b), $K_s - n$ (c), and $K_s - l$ (d) parameter space after convergence had been achieved to a limiting distribution. In general, the results of the classical response surface analysis presented in Figure 5, and the scatter plots presented in Figure 7 are conclusive. Indeed, a similar shape induced in the objective function, and thus, in the correlation structure between the parameters, is found in the vicinity of the optimum.

To verify the "correctness" of the Richards-hydraulic model, consider the residuals from the most probable parameter set found with the SCE-UA algorithm and the prediction uncertainty intervals of the HYDRUS-1D simulated outflows associated with the posterior parameter estimates (Figure 8; light-gray region). The black dots correspond to the measured outflow data of the Lincoln soil, whereas the most optimal parameter set is indicated with the dashed line. Note that the uncertainty ranges generally bracket the observed outflows during almost the entire period. However, the prediction uncertainty associated with the posterior parameter estimates does not include the observations and display bias (systematic error) at the beginning of the outflow experiment.

Finally, Figure 9 depicts how the posterior uncertainty associated with the soil hydraulic parameters for the MSO experiment for the Lincoln soil translates into uncertainty associated with the retention and unsaturated soil hydraulic conductivity curves. The directly estimated retention and hydraulic conductivity points are indicated with the black symbols. Note that the fit to the directly estimated hydraulic conductivity points is quite poor. This suggests that future attempts to improve the combined Richards-soil hydraulic model would be most productive if focused directly on the validity of the model assumptions.

ADVANCES IN INVERSE MODELLING

While considerable progress has been made in the application of automated optimization algorithms, the emphasis of most of this research has been on the estimation of a single optimal set of model parameters, thereby effectively neglecting the influence of possible sources of uncertainty on the final parameter estimates. Uncertainty is an unavoidable element of any modelling exercise in hydrology. Recently, however, there is an increasing awareness, that hydrologic model identification and evaluation procedures should explicitly include an estimate of the uncertainties associated with the model parameters and underlying model structure (Kuczera and Parent, 1998; Bates and Campbell,

2001; Thyer *et al.*, 2002; Vrugt *et al.*, 2003b). Recent investigations have therefore been directed towards the development of strategies that explicitly account for model structural uncertainty, input, state, parameter, and output uncertainty. Bayesian, pseudo-Bayesian, multiple-criteria, and recursive model identification strategies are currently receiving considerable attention to investigate their usefulness for quantifying the various sources of uncertainty associated with the application of hydrologic models. For instance, significant advances in computational capabilities now allow the use of MCMC samplers that are especially designed to converge to an ensemble of parameter sets, each associated with a certain likelihood, rather than a single "best" estimate, in the context of Bayesian model identification techniques.

The goal should therefore be to develop an IM strategy that can explicitly account for model structural, input and parameter uncertainty, can deal with multiple sources and types of information, and can simultaneously provide probabilistic estimates of the uncertainty associated with the model predictions. This is especially true now the interests have begun to switch to larger-scale applications of the IM approach in soil and surface hydrology, for which different types of data at a variety of temporal and spatial scales are available. For instance, many of the latest hydrologic watershed or land-surface models simulate several output fluxes (e.g., water, energy, chemical constituents, etc.) for which measurement data are available, and all these data must be correctly utilized to ensure proper model calibration (Beven and Kirkby, 1979; De Grosbois *et al.*, 1988; Kuczera, 1982, 1983; Woolhiser *et al.*, 1990; Kuczera and Mroczkowski, 1998; Gupta *et al.*, 1998). Madsen (2003) recently demonstrated the successful application of the SCE-UA algorithm to calibrate a multiparameter, semidistributed, watershed model, thereby yielding effective hydrologic parameter values at the support of interest.

One strategy to explicitly recognize the multiobjective nature of the calibration problem is to define several optimization criteria (objective functions) that measure different (complementary) aspects of the system behavior and to use a multicriteria optimization method to identify the set of nondominated, efficient, or Pareto optimal solutions. The Pareto solutions represent trade-offs among the different incommensurable and often conflicting objectives, having the property that moving from one solution to another, results in the improvement of one objective while causing deterioration in one or more others. To date, to the authors' knowledge, no applications of this approach are yet found in the area of IM of soil hydraulic properties.

The classical, single-objective optimization approach operates under the central assumption that a single objective function is able to properly extract all of the information contained in the time series of observations. However, practical experience with the calibration of hydrological models suggests that the magnitude of structural error in the model for some portions of the model response may, in general, be equivalent to or even substantially larger than the measurement error and that these structural or model errors do not necessarily have any inherent probabilistic property that can be exploited in the construction of an objective function (Gupta *et al.*, 1998). Due to the presence of these structural inadequacies in the hydrologic model, any single (scalar) objective function, no matter how carefully chosen, is inadequate to properly measure all of the characteristics of the observed data deemed to be important.

These considerations imply the design of an IM or calibration strategy that has the ability to simultaneously incorporate several competing objectives. A strategy that can address this challenge is multiobjective optimization, which has its roots in late-nineteenth century welfare economics, in the work of Edgeworth (1881), and can be stated as

$$\min_{\beta \in B} \; E(\beta) = \begin{bmatrix} E_1(\beta) \\ E_2(\beta) \\ \vdots \\ E_M(\beta) \end{bmatrix} \tag{14}$$

where $E_i(\beta)$ is the ith of M objective functions. The solution to this problem will, in general, no longer be a single "best" parameter set but will consist of a Pareto set $\boldsymbol{P}(B)$ of solutions in the feasible parameter space B corresponding to various trade-offs among the objectives. The Pareto set of solutions defines the minimum uncertainty in the parameters that can be achieved without stating a subjective relative preference for minimizing one specific component of $E(\beta)$ at the expense of another. Figure 10 illustrates the Pareto solution set for a simple problem where the aim is to simultaneously minimize two objectives (E_1,E_2) with respect to two parameters (β_1,β_2). The individual points C and D minimize objectives E_1 and E_2, respectively, whereas the solid line joining C and D represents the theoretical Pareto set of solutions. The black dots (Figure 10a) indicate an initial set of parameter estimates, while the number in subscript (Figure 10b) denotes their corresponding Pareto rank. Moving along the line from C to D results in the improvement of E_2 while successively causing deterioration in E_1. The points falling on the line CD represent trade-offs between the objectives and are called *nondominated, noninferior*, or *efficient solutions*. Put simply, the feasible parameter space can be partitioned into "good" or Pareto solutions and "bad" or "inferior" solutions. In the absence of additional information, it is impossible to distinguish any of the Pareto solutions (rank 1 points) as being objectively better than any of the other Pareto solutions. Furthermore, every member of the Pareto set will match some characteristic of the observed data better than any other member of the Pareto set, but the trade-off will be that some other characteristic of the observed data will not be as well-matched (Yapo

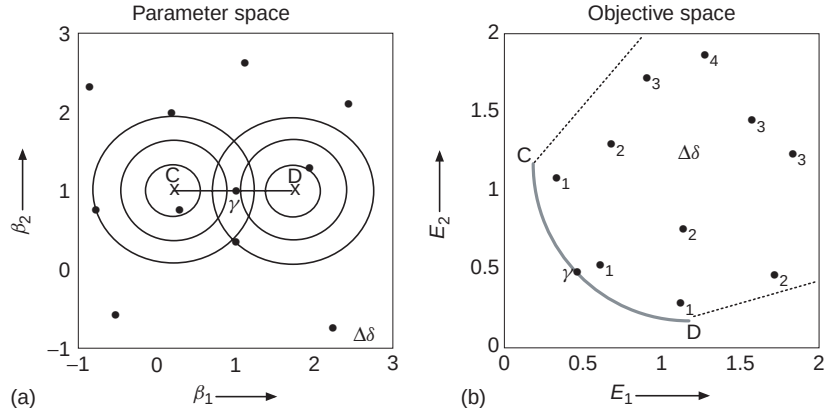


Figure 10 Illustration of the concept of Pareto optimality for a problem having two parameters (β_1, β_2) and two-criteria (E_1, E_2), (a) in the parameter, and (b) objective space. The points C and D indicate the solutions that minimize each of the criteria E_1 and E_2. The thick line joining C and D corresponds to the Pareto set of solutions; γ is an element of the solution set, which is superior in the multicriteria sense to any other point in δ

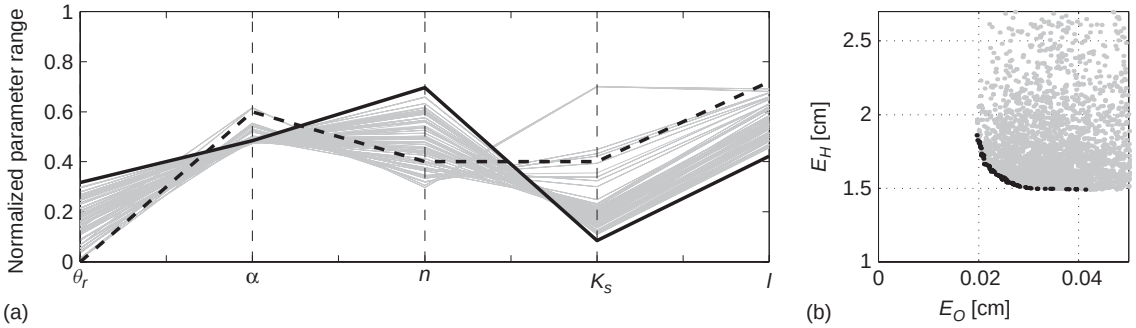


Figure 11 (a) normalized parameter plots for each of the unsaturated soil hydraulic parameters using a two-criteria $\{E_O, E_H\}$ calibration with the MOSCEM-UA algorithm. Each line across the graph denotes a single parameter set: gray – Pareto solution set, solid and dashed black lines are single criterion solutions of E_O and E_H respectively; (b) two-dimensional projection of the objective space. The Pareto solution set is indicated with black dots. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

et al., 1998). Because of errors in the model structure (and other possible sources), it is usually not possible to find a single point β at which all of the criteria have their minima. Note that this multiobjective equivalence of parameter sets is different from the probabilistic representation of parameter uncertainty, estimated using the SCEM-UA algorithm.

From another perspective, the multicriteria approach offers a way forward since it circumvents the problem of specifying user-subjective weights to different sets of observations, like outflow and tensiometer measurements. Improper selection of these weights can influence not only the confidence intervals of the parameters, but also the location of the minimum of the objective function (Hollenbeck and Jensen, 1998).

Finally, we will consider the estimation of unsaturated soil hydraulic properties using a Multiobjective parameter optimization approach. Again, the observed outflow and tensiometer readings during the MSO experiment of the Lincoln soil were used. A pair of Root Mean Square Error (RMSE) objective functions were computed. First, E_O was computed to measure the ability of the HYDRUS-1D model to simulate the observed outflow measurements. Second, E_H was computed to measure the ability of the model to simulate the observed tensiometer data within the soil sample during the outflow experiment. The Pareto optimal solution space for the two-criteria problem was estimated using a population size of 250 points and 10 000 calculations using the recently developed Multiobjective Shuffled Complex Evolution Metropolis (MOSCEM-UA) algorithm (Vrugt *et al.*, 2003c). This algorithm employs a similar evolution strategy as the SCEM-UA algorithm, with the small adaptation that the former algorithm uses the concepts of Pareto dominance rather than direct, single-objective function evaluations to evolve the initial populations of points towards a set of solutions stemming from a stable distribution (Pareto set). The results of this multicriteria $\{E_O, E_H\}$ calibration are summarized in Figure 11.

Figure 11a presents normalized parameter plots for each of the unsaturated soil hydraulic parameters (θ_r, α, n, K_s, and l) in the HYDRUS-1D model. The 5 parameters are listed along the x-axis, while the y-axis corresponds to the parameter values scaled according to their prior uncertainty ranges defined in Table 1 to yield normalized ranges. Each line across the graph represents one parameter set. The solid and dashed lines correspond to the single-objective solutions of E_O and E_H, respectively, obtained by separately fitting both criteria using the SCE-UA algorithm. The gray lines denote members of the Pareto set of solutions. The small plot at the right-hand side (Figure 11b) depicts two-dimensional projections of the bicriterion trade-off surface represented by the Pareto set of solutions. It clearly illustrates that the MOSCEM-UA algorithm has generated a fairly uniform approximation of the Pareto set, thereby containing the single criterion solutions at the extreme ends of the Pareto frontier (see Figure 11a). Because of errors in the model structure, it is not possible to find one single set of parameters that simultaneously minimizes both criteria. Instead, it is common to have sets of solutions with the property that moving from one to another results in improvement of the ability to fit the observed outflow while causing deterioration in fitting the soil water pressure head data. There is considerable uncertainty associated with the saturated hydraulic conductivity (K_s) and the parameter n, which primarily determines the slope of the retention curve in the intermediate water-content range. This indicates that the combined Richards'-hydraulic model structure is in need of further improvements. Although not further pursued here, the use of decoupled hydraulic functions might alleviate some of the errors in the model structure and further reduce the correlation among the model parameters.

CLOSING COMMENTS

Major weakness of all parameter estimation approaches include their underlying treatment of the uncertainty in the input–output representation of the model as being primarily (and explicitly) attributed to uncertainty in the parameter estimates, without explicit treatment of the input, output, and model structural uncertainties. Hence, uncertainties in the modelling procedure stem not only from uncertainties in the parameter estimates, but also from measurement errors associated with the system input (forcing) and output, and from model structural errors arising from the aggregation of spatially distributed real-world processes into a mathematical model. Not properly accounting for these errors during parameter estimation can result in model simulations and their associated prediction uncertainty bounds, which do not consistently represent and bracket the measured system behavior. This is usually evidenced by residuals, which exhibit considerable variations in bias (nonstationarity), variance (heteroscedasticity), and correlation structures under different hydrologic conditions (see, for instance, Figure 8a at the beginning of the outflow experiment). Several contributions to the hydrologic literature have, therefore, brought into question the continued usefulness of the classical paradigm for estimating model parameters (Beven and Binley, 1992; Gupta *et al.*, 1998), especially when multiple (and often conflicting) sources of information are available for parameter estimation (see, for instance, the trade-off in the fitting of the outflow and soil water pressure heads in Figure 11).

Some interesting methods for addressing these problems, particularly in the context of estimating reasonable confidence bounds on the model parameters and simulations, have recently begun to appear in the hydrologic model identification literature. One attempt to more completely treat input, output, parameter, and model structural errors is the Simultaneous Optimization and Data Assimilation (SODA) framework of Vrugt *et al.* (2005). This method merges the strengths of the search efficiency and explorative capabilities of the SCEM-UA algorithm with the power and computational efficiency of the Ensemble Kalman Filter (Evensen, 1994) to simultaneously estimate model parameters and state variables. The ability of SODA to explicitly deal with input, output, parameter, and model structural errors, results in improved estimates of parameter and model prediction uncertainty ranges.

SOFTWARE LINKS

Hydrus-1D

- Software description: finite element model for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media.
- Typical application: analysis of water flow and solute transport in soils
- URL: `http://www.pc-progress.cz/Fr_Hydrus1D.htm`
- Reference: Šimůnek, J., M. Šejna, and M. Th. van Genuchten. (1998b). The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. Version 2.0, *IGWMC – TPS – 70,* International Ground Water Modeling Center, Colorado School of Mines, Golden, Colorado, 202pp.

SCEM-UA and MOSCEM-UA Optimization codes

- Software description: computerized algorithms for solving single and multiobjective optimization problems. The software is available in the MATLAB and C-environment. Any computer coded model can be coupled to these algorithms.
- Typical applications: estimation of soil hydraulic parameters, calibration of watershed and land-surface models. Recent applications involve the estimation of bird-migration behavior, the determination of water content

profiles using measured TDR-waveforms, the cycling of nitrogen in forested ecosystems, and the interaction between dissolved organic matter and heavy metals.
- URL: `http://www.science.uva.nl/ibed/research/Research_Fields/cbpg/software/`
- Reference: Vrugt, J.A., Gupta, H.V., Bouten, W. and Sorooshian, S. (2003b). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water. Resour. Res.*, **39**, 1201, doi:10.1029/2002WR001642. Vrugt, J.A., H.V. Gupta. L. Bastidas, W. Bouten, and S. Sorooshian. (2003c). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water. Resour. Res.*, 1214, doi:10.1029/2002WR001746.

Other Optimization Packages

- NLFIT – A windows-based program for model calibration and identification. Developed by George Kuczera
  URL: `http://www.eng.newcastle.edu.au/~cegak/`
- PEST – A general-purpose parameter estimation utility. Developed by John Doherty
  URL: `http://www.parameter-estimation.com/html/pest_overview.html`

Acknowledgments

The Earth Life Sciences and Research Council (ALW) supported the investigations of the first author with financial aid from the Netherlands Organization for Scientific Research (NWO).

FURTHER READING

Beven K.J. and Binley A.M. (1992) The future of distributed models: Model calibration and uncertainty prediction. *Hydrology Process*, **6**, 279–298.

Vrugt J.A. and Bouten W. (2001) Information content of data for identifying soil hydraulic parameters from outflow experiments. *Soil Science Society of America Journal*, **65**, 19–27.

REFERENCES

Abbaspour K.C., Schulin R. and van Genuchten M.T.h (2001) Estimating unsaturated soil hydraulic parameters using ant-colony optimization. *Advances in Water Resources*, **24**, 827–841.

Abbaspour K.C., van Genuchten M.T.h, Schulin R. and Schläppi E. (1997) A sequential uncertainty domain inverse procedure for estimating subsurface flow and transport parameters. *Water Resources Research*, **33**, 1879–1892.

Abbaspour K.C., Sonnleitner M.A. and Schulin R. (1999) Uncertainty in estimation of soil hydraulic parameters by inverse modeling: Example lysimeter experiments. *Soil Science Society of America Journal*, **63**, 501–509.

Bates B.C. and Campbell E.P. (2001) A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modelling. *Water Resources Research*, **37**, 937–947.

Beven K.J. and Kirkby M. (1979) A physically based variable contributing area model of basin hydrology. *Hydrological Science Bulletin*, **24**, 43–69.

Box G.E.P. and Tiao G.C. (1973) *Bayesian Inference in Statistical Analyses*, Addison-Wesley Publication: Reading.

Boyle D.P., Gupta H.V. and Sorooshian S. (2000) Toward improved calibration of hydrological models: Combining the strengths of manual and automatic methods. *Water Resources Research*, **36**, 3663–3674.

Carrera J. and Neumann S.P. (1986) Estimation of aquifer parameters under transient and steady state conditions. 2. Uniqueness, stability and solution algorithms. *Water Resources Research*, **22**, 211–227.

Ciollaro G. and Romano N. (1995) Spatial variability of the soil hydraulic properties of a volcanic soil. *Geoderma*, **65**, 263–282.

Clausnitzer V., and Hopmans J.W. (1995) Non-linear parameter estimation: LM_OPT. General-purpose optimization code based on the Levenberg-Marquardt algorithm. *Land, Air and Water Resoures Paper no. 100032*, University of California, Davis.

Dane J.H. and Hruska S. (1983) In-situ determination of soil hydraulic properties during drainage. *Soil Science Society of America Journal*, **47**, 619–624.

De Grosbois E., Hooper R.P. and Christopherson N. (1988) A multi-signal automatic calibration methodology for hydrochemical models: A case study of the Birkenes model. *Water Resources Research*, **24**, 1299–1307.

Doherty J. (1994) *PEST. Watermark Computing*, Corinda, p. 122.

Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (2003) Calibration of watershed models. *Water Science and Application 6*, AGU: Washington.

Duan Q., Gupta V.K. and Sorooshian S. (1993) Shuffled complex evolution approach for effective and efficient global minimization. *Journal of optimization theory and applications*, **76**, 501–521.

Duan Q., Sorooshian S. and Gupta V.K. (1992) Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, **28**, 1015–1031.

Durner W., Schultze E.B. and Zurmühl T. (1999) State-of-the-art in inverse modelling of inflow/outflow experiments. In *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media, Proceedings of International Workshop*, van Genuchten M.T.h, Leij F.J. and Wu L. (Eds.), Riverside, pp. 661–681, October 22–24.

Eching S.O. and Hopmans J.W. (1993) Optimization of hydraulic functions from transient outflow and soil water pressure data. *Soil Science Society of America Journal*, **57**, 1167–1175.

Eching S.O., Hopmans J.W. and Wendroth O. (1994) Unsaturated hydraulic conductivity from transient multistep outflow and soil water pressure data. *Soil Science Society of America Journal*, **58**, 687–695.

Edgeworth F.Y. (1881) *Mathematical Physics*, C. Kegan Paul and Co.: London.

Gan T.Y. and Biftu G.F. (1996) Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions, and model structure. *Water Resources Research*, **32**, 3513–3524.

Evensen G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast statistics. *Journal of Geophysical Research*, **99**, 10143–10162.

Finsterle S. and Najita J. (1998) Robust estimation of hydrogeologic model parameters. *Water Resources Research*, **34**, 2939–2947.

Gribb M.M. (1996) Parameter estimation for determining hydraulic properties of a fine sand from transient flow measurements. *Water Resources Research*, **32**, 1965–1974.

Gupta H.V., Sorooshian S. and Yapo P.O. (1998) Towards improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, **34**, 751–763.

Gupta V.K., and Sorooshian S. (1994) Calibration of conceptual hydrologic models: Past, present and future. Invited paper in: (Eds), *Trends in Hydrology, Research Trends*, Council of Scientific Research Integration: Trivandrum, pp. 329–346.

Gupta V.K. and Sorooshian S. (1985) Uniqueness and observability of conceptual rainfall-runoff model parameters. The percolation process examined. *Water Resources Research*, **19**, 269–276.

Hastings W.K. (1970) Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, **57**, 97–109.

Hogue T.S., Sorooshian S., Gupta H.V., Holz A. and Braatz D. (2000) A multistep automatic calibration scheme for river forecasting models. *Journal of Hydrometeorology*, **1**, 524–542.

Holland J. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press: Ann Arbor.

Hollenbeck K.J. and Jensen K.H. (1998) Maximum likelihood estimation of unsaturated soil hydraulic parameters. *Journal of Hydrology*, **210**, 192–205.

Hopmans J.W. and Šimůnek J. (1999) Review of inverse estimation of soil hydraulic properties. In *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.T.h, Leij F.J. and Wu L. (Eds.), University of California: Riverside, pp. 643–659.

Hopmans J.W., Šimůnek J., Romano N. and Durner W. (2002) Simultaneous determination of water transmission and retention properties – Inverse methods. In *Methods of Soil Analysis. Part 4. Physical Methods*, SSSA Book Ser. 5, Dane J.H. and Topp G.C. (Eds.), SSSA: Madison, pp. 963–1008.

Janssen P.H.M. and Heuberger P.S.C. (1995) Calibration of process-oriented models. *Ecological Modelling*, **83**, 55–66.

Kool J.B. and Parker J.C. (1988) Analysis of the inverse problem for transient unsaturated flow. *Water Resources Research*, **24**, 817–830.

Kool J.B., Parker J.C. and van Genuchten M.T.h (1985) Determining soil hydraulic properties for one-step outflow experiments by parameter estimation. I. Theory and numerical studies. *Soil Science Society of America Journal*, **49**, 1348–1354.

Kuczera G. and Parent E. (1998) Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. *Journal of Hydrology*, **211**, 69–85.

Kuczera G. (1982) On the relationship between the reliability of parameter estimates and hydrologic time series used in calibration. *Water Resources Research*, **18**, 146–154.

Kuczera G. (1983) Improved parameter inference in catchment models: 2. combining different kinds of hydrologic data and testing their compatibility. *Water Resources Research*, **19**, 1163–1172.

Kuczera G. (1997) Efficient subspace probabilistic parameter optimization for catchment models. *Water Resources Research*, **33**, 177–185.

Kuczera G. and Mroczkowski M. (1998) Assessment of hydrological parameter uncertainty and the worth of multiresponse data. *Water Resources Research*, **34**, 1481–1489.

Luce C.H. and Cundy T.W. (1994) Parameter identification for a runoff model for forest roads. *Water Resources Research*, **30**, 1057–1069.

Madsen H. (2003) Parameter estimation in distributed hydrologic catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources*, **26**, 205–216.

MATLAB, Version 5.3. (1999) The Mathworks, Natick, MA, USA.

Mertens J., Madsen H., Kristensen M., Jacques D. and Feyen J. (2005) Sensitivity of soil parameters in unsaturated zone modeling and the relation between effective, laboratory and in-situ estimates. *Hydrological Processes*, in press, DOI:10.1002.hyp.5591.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953) Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1091.

Mosegaard K. and Tarantola A. (1995) Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research-Solid Earth*, **100**(B7), 12431–12447.

Mualem Y.A. (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 513–522.

Musters P.A.D. and Bouten W. (1999) Assessing rooting depths of an Austrian Pine stand by inverse modeling soil water content maps. *Water Resources Research*, **35**, 3041–3048.

Nelder J.A. and Mead R. (1965) A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Neupauer R.M., Borchers B. and Wilson J.L. (2000) Comparison of inverse methods for reconstructing the release history of a groundwater contamination source. *Water Resources Research*, **36**, 2469–2476.

Pan L.H. and Wu L.S. (1998) A hybrid global optimization method for inverse estimation of hydraulic parameters: Annealing-simplex method. *Water Resources Research*, **34**, 2261–2269.

Parker J.C., Kool J.B. and van Genuchten M.T.h (1985) Determining soil hydraulic properties from one-step outflow experiments by parameter estimation. II. Experimental studies. *Soil Science Society of America Journal*, **49**, 1354–1359.

Press W.H., Teukolsky S.A., Vetterling W.T. and Flannery B.P. (1992) *Numerical recipes in Fortran: The art of scientific computing*, *Second Edition*, Cambridge University Press: New York.

Romano N. and Santini A. (1999) Determining soil hydraulic functions from evaporation experiments by a parameter

estimation approach: Experimental verifications and numerical studies. *Water Resources Research*, **35**, 3343–3359.

Russo D. (1988) Determining soil hydraulic properties by parameter estimation: On the selection of a model for the hydraulic properties. *Water Resources Research*, **24**, 453–459.

Santini A., Romano N., Ciollaro G. and Comegna V. (1995) Evaluation of a laboratory inverse method for determining unsaturated hydraulic properties of a soil under different tillage practices. *Soil Science*, **160**, 340–351.

Šimůnek J. and van Genuchten M.T.h (1996) Estimating unsaturated soil hydraulic properties from tension disc infiltrometer data by numerical inversion. *Water Resources Research*, **32**, 2683–2696.

Šimůnek J. and van Genuchten M.T.h (1997) Estimating unsaturated soil hydraulic properties from multiple tension disc infiltrometer data. *Soil Science*, **162**, 383–398.

Šimůnek J., Šejna M. and van Genuchten M.T.h (1998a) The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. Version 2.0, *IGWMC – TPS – 70*, International Ground Water Modeling Center, Colorado School of Mines: Golden, Colorado, p. 202.

Šimůnek J., Wendroth O. and van Genuchten M.T.h (1998b) A parameter estimation analysis of the evaporation method for determining soil hydraulic properties. *Soil Science Society of America Journal*, **62**, 894–905.

Sorooshian S. and Dracup J.A. (1980) Stochastic parameter-estimation procedures for hydrologic rainfall-runoff models – correlated and heteroscedastic error cases. *Water Resources Research*, **16**, 430–442.

Sorooshian S., Duan Q. and Gupta V.K. (1993) Calibration of rainfall-runoff models: Application of global optimisation to the Sacramento Soil Moisture accounting model. *Water Resources Research*, **29**, 1185–1194.

Sorooshian S., Gupta V.K. and Fulton J.L. (1983) Evaluation of maximum-likelihood parameter estimation techniques for conceptual rainfall-runoff models – influence of calibration data variability and length on model credibility. *Water Resources Research*, **19**, 251–259.

Takeshita Y. (1999) Parameter estimation of unsaturated soil hydraulic properties from transient outflow experiments using genetic algorithms. In *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media. Proceedings of International Workshop*, *Proc. Int. Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, van Genuchten M.T.h, Leij F.J. and Wu L. (Eds.), University of California, Riverside, CA, October 22–24.

Tanakamaru H. (1995) Parameter estimation for the tank model using global optimisation. *Transactions of the Japanese society of irrigation, drainage and reclamation engineering*, **178**, 103–112.

Thyer M., Kuczera G. and Wang Q.J. (2002) Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation. *Journal of Hydrology*, **265**, 246–257.

Toorman A.F., Wierenga P.J. and Hills R.G. (1992) Parameter estimation of soil hydraulic properties from one-step outflow data. *Water Resources Research*, **28**, 3021–3028.

Valiantzas J.D. and Kerkides D.G. (1990) A simple iterative method for the simultaneous determination of soil hydraulic properties from one-step outflow data. *Water Resources Research*, **26**, 143–152.

van Dam J.C., Stricker J.N.M. and Droogers P. (1992) Evaluation of the inverse method for determining soil hydraulic functions from one-step outflow experiments. *Soil Science Society of America Journal*, **56**, 1042–1050.

van Dam J.C., Stricker J.N.M. and Droogers P. (1994) Inverse method for determining soil hydraulic functions from multistep outflow experiments. *Soil Science Society of America Journal*, **58**, 647–652.

van Genuchten M.T.h (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

Vrugt J.A. and Bouten W. (2002) Validity of first-order approximations to describe parameter uncertainty in soil hydrologic models. *Soil Science Society of America Journal*, **66**, 1740–1751.

Vrugt J.A., Bouten W., Gupta H.V. and Hopmans J.W. (2003a) Toward improved identifiability of soil hydraulic parameters: On the selection of a suitable parametric model. *Vadose Zone Journal*, **2**, 98–113.

Vrugt J.A., Bouten W., Gupta H.V. and Sorooshian S. (2002) Toward improved identifiability of hydrologic model parameters: The information content of experimental data. *Water Resources Research*, **38**, 1312. doi:10.1029/2001WR001118.

Vrugt J.A., Gupta H.V., Bouten W. and Sorooshian S. (2003b) A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, **39**, 1201. doi:10.1029/2002WR001642.

Vrugt J.A., Gupta H.V., Bastidas L., Bouten W. and Sorooshian S. (2003c) Effective and efficient algorithm for multi-objective optimization of hydrologic models. *Water Resources Research*, **39**, 1214. doi:10.1029/2002WR001746.

Vrugt J.A., Hopmans J.W. and Šimůnek J. (2001b) Calibration of a two-dimensional root water uptake model. *Soil Science Society of America Journal*, **65**, 1027–1037.

Vrugt J.A., van Wijk M.T., Hopmans J.W. and Šimůnek J. (2001a) One, two, and three-dimensional root water uptake functions for transient modelling. *Water Resources Research*, **37**, 2457–2470.

Vrugt J.A., Diks C.G.H., Gupta H.V., Bouten W. and Verstraten J.M. (2005) Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, **41**, W01017.

Whisler F.D. and Whatson K.K. (1968) One-dimensional gravity drainage of uniform columns of porous materials. *Journal of Hydrology*, **6**, 277–296.

Wildenschild D., Hopmans J.W. and Šimůnek J. (2001) Flow rate dependence of soil hydraulic characteristics. *Soil Science Society of America Journal*, **65**, 35–48.

Woolhiser D.A., Smith R.E. and Goodrich D.C. (1990) A kinematic runoff and erosion model: documentation and user manual. *ARS-77*, U.S. Department of Agriculture, Tucson, Arizona, pp. 130.

Yapo P., Gupta H.V. and Sorooshian S. (1996) Calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, **181**, 23–48.

Yapo P.O., Gupta H.V. and Sorooshian S. (1998) Multi-objective global optimisation for hydrologic models, J. *Hydrology*, **204**, 83–97.

Zachmann D.W., Duchateau P.C. and Klute A. (1981) The calibration of Richards flow equation for a draining column by parameter identification. *Soil Science Society of America Journal*, **45**, 1012–1015.

Zachmann D.W., Duchateau P.C. and Klute A. (1982) Simultaneous approximation of water capacity and soil hydraulic conductivity by parameter identification. *Soil Science*, **134**, 157–163.

Zijlstra J., and Dane J.H. (1996) Identification of hydraulic parameters in layered soils based on a quasi-Newton method. *Journal of Hydrology*, **181**, 233–250.

Zurmühl T., and Durner W. (1998) Determination of parameters for bimodal hydraulic functions by inverse modelling. *Soil Science Society of America Journal*, **62**, 874–880.

78: Models of Water Flow and Solute Transport in the Unsaturated Zone

JIRKA ŠIMŮNEK

Department of Environmental Sciences, University of California, Riverside, CA, US

A large number of models for simulating water flow and solute transport in the unsaturated zone are now increasingly being used for a wide range of applications in both research and management. Modeling approaches range from relatively simple analytical and semianalytical solutions, to complex numerical codes. While analytical and semianalytical solutions are still popular for some applications, the ever-increasing power of personal computers and the development of more accurate and numerically stable solution techniques have motivated much wider use of numerical codes in recent decades. The wide use of numerical models is also significantly enhanced by their availability in both the public and commercial domains, and by the development of sophisticated graphics-based interfaces that can tremendously simplify their use.

In this paper I focus mainly on numerical models, give a brief history of their development, and discuss some of the more often used numerical techniques including relatively efficient matrix solvers that now are available for multidimensional models. Names and web addresses of some of the more popularly used numerical codes simulating vadose zone processes are also provided. Finally, some typical problems in which the numerical codes have been applied are identified.

INTRODUCTION

Many models of varying degree of complexity and dimensionality have been developed during the past several decades to quantify the basic physical and chemical processes affecting water flow and pollutant transport in the unsaturated zone. Computer models based on analytical and numerical solutions of the flow and solute transport equations are now increasingly being used for a wide range of applications in research and management of natural subsurface systems. Modeling approaches range from relatively simple analytical and semianalytical models to more complex numerical codes that permit consideration of a large number of simultaneous nonlinear processes. Whereas analytical and semianalytical solutions are still more popular for most relatively simple applications, the ever-increasing power of personal computers and the development of more accurate and numerically stable solution techniques have given rise to the much wider use of numerical models in recent decades. The wide use of numerical models has also been significantly enhanced by their availability in both public and commercial domains, and by the development of sophisticated graphics-based interfaces that tremendously simplify their use.

Analytical, semianalytical, and numerical models are usually based on the following three governing equations for water flow, solute transport, and heat movement, respectively:

$$\frac{\partial \theta(h)}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] - S \qquad (1)$$

$$\frac{\partial \theta Rc}{\partial t} = \frac{\partial}{\partial z}\left[\theta D\left(\frac{\partial c}{\partial z}\right) - qc\right] - \phi \qquad (2)$$

$$\frac{\partial C(\theta)T}{\partial t} = \frac{\partial}{\partial z}\left[\lambda(\theta)\left(\frac{\partial T}{\partial z}\right) - C_{\mathrm{w}}qT\right] \qquad (3)$$

Suitable simplifications (mostly for analytical approaches) or extensions thereof (e.g. for two- and three-dimensional systems) are also employed. In equation (1), often referred to as the Richards equation, z is the vertical coordinate positive upwards, t is time, h is the pressure head, θ is

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the water content, S is a sink term representing root water uptake or some other sources or sinks, and $K(h)$ is the unsaturated hydraulic conductivity function, often given as the product of the relative hydraulic conductivity, K_r, and the saturated hydraulic conductivity, K_s. In equation (2), known as the *convection-dispersion equation* (CDE), c is the solution concentration, R is the retardation factor that accounts for adsorption, D is the dispersion coefficient accounting for both molecular diffusion and hydrodynamic dispersion, q is the volumetric fluid flux density, and ϕ is a sink/source term that accounts for various zero- and first-order or other reactions. In equation (3), T is temperature, λ is the apparent thermal conductivity, and C and C_w are the volumetric heat capacities of the soil and the liquid phase, respectively. Solutions of the Richards equation (1) require knowledge of the unsaturated soil hydraulic functions, that is, the soil water retention curve, $\theta(h)$, describing the relationship between the water content θ and the pressure head h, and the unsaturated hydraulic conductivity function, $K(h)$, defining the hydraulic conductivity K as a function of h or θ. While under certain conditions (i.e. for linear sorption, a concentration-independent sink term ϕ, and a steady flow field) equations (2) and (3) are linear equations, equation (1) is generally highly nonlinear because of the nonlinearity of the soil hydraulic properties. Consequently, many analytical solutions have been derived in the past for equations (2) and (3) and these analytical solutions are now widely used for analyzing solute and heat transport under steady-state conditions. Although a large number of analytical solutions of (1) exist, they can generally be applied only to drastically simplified problems. The majority of applications for water flow in the vadose zone requires a numerical solution of the Richards equation.

ANALYTICAL MODELS

Analytical methods represent a classical mathematical approach to solve differential equations, leading to an exact solution for a particular problem. Analytical solutions are usually obtained by applying various transformations (e.g. Laplace or Fourier transformations) to the governing equations, invoking a separation of variables, or the Green's function approach (e.g. Leij *et al.*, 2000). Analytical models usually result in an explicit equation that, for example, states that concentration (or the pressure head, water content, or temperature) is equal to a certain value at particular time and location. One can therefore evaluate a particular variable directly without time-stepping, typical of numerical methods. Many analytical solutions lead to relatively complicated formulations that include infinite series and/or integrals that need to be evaluated numerically, which suggests some ambiguity in the often-claimed advantage of exactness of analytical methods over numerical techniques. On the other hand, using analytical solutions one can often more easily evaluate interrelationships among parameters, and get better insight into how various processes control the basic flow and transport processes (e.g. using dimensionless variables and parameters). Analytical solutions are often also used to check the correctness and accuracy of numerical models, although numerical models can equally well be used to check the correctness of the complex analytical solutions.

Analytical solutions can usually be derived only for simplified transport systems involving linearized governing equations, homogeneous soils, simplified geometries of the transport domain, and constant or highly simplified initial and boundary conditions. Unfortunately, analytical solutions for more complex situations, such as for transient water flow or nonequilibrium solute transport with nonlinear reactions, are generally not available and/or cannot be derived, in which case numerical models must be employed.

Solute Transport

Numerous analytical solutions of equations (2) and (3), or their two- and three-dimensional equivalents, have been developed in the last 40 years and are now widely used for predictive purposes and/or analyzing laboratory and/or field observed concentration distributions. The majority of these solutions pertains to equations (2) and (3) assuming constant water content, θ, and flux, q, values (i.e. for steady-state water flow conditions). Since equation (2) and (3) have the same form, analytical solutions derived for solute transport can often also be used immediately for many heat transport problems, and *vice versa*.

Some of the more popular one- and multidimensional analytical transport models have been CXTFIT (Parker and van Genuchten, 1984), AT123D (Yeh, 1981), and 3DADE (Leij and Bradford, 1994). A large number of analytical models for one-, two-, and three-dimensional solute transport problems were recently incorporated into the comprehensive software package STANMOD (STudio of ANalytical MODels) (Šimůnek *et al.*, 1999b) (**`http://www.ussl.ars.usda.gov/models/stanmod/stanmod.HTM`**). This Windows-based computer software package includes not only programs for evaluating analytical solutions for equilibrium convective-dispersive solute transport (i.e. the CFITM of van Genuchten (1980) for one-dimensional transport and 3DADE for three-dimensional problems), but also additional programs that solve more complex problems. For example, it also incorporates the CFITIM (van Genuchten, 1981) and N3DADE (Leij and Toride, 1997) programs for nonequilibrium convective-dispersive transport (i.e. the two-region mobile-immobile model for physical nonequilibrium and the two-site sorption model for chemical nonequilibrium) for one- and multidimensional formulations, respectively. STANMOD also includes CXTFIT2 (Toride *et al.*, 1995), an updated

version of CXTFIT to solve both direct and inverse problems for three different one-dimensional transport models: (i) the conventional CDE (1); (ii) the chemical and physical nonequilibrium CDEs; and (iii) a stochastic stream tube model based upon the local scale equilibrium or nonequilibrium CDE. All three models consider linear adsorption, as well as the zero- and first-order decay/source terms. In addition, STANMOD includes the CHAIN code of van Genuchten (1985) for analyzing the convective-dispersive transport of up to four solutes involved in sequential first-order decay reactions. Examples are the migration of radionuclides, in which the chain members form first-order decay reactions, and the simultaneous movement of various interacting nitrogen or organic chemicals. The latest version of STANMOD also includes the screening model of Jury *et al.* (1983) for describing transport and volatilization of soil-applied volatile organic chemicals.

Water Flow

The highly nonlinear Richards equation can be solved analytically only for a very limited number of cases involving homogeneous soils, simplified initial and boundary conditions, and relatively simple constitutive relationships describing the unsaturated soil hydraulic properties. The Richards equation for this purpose needs to be first linearized. This can be accomplished using several mathematical transformations, with the most common transformation being the Kirchhoff integral transformation:

$$\Phi = \int_{-\infty}^{h} K(h)\,\mathrm{d}h \tag{4}$$

Linearization of the Richards equation often also involves using an exponential law relating hydraulic conductivity and the pressure head (Gardner, 1958):

$$K(h) = K_s \exp(\alpha h) \tag{5}$$

where α is an empirical coefficient.

Development of analytical and semianalytical solutions of the unsaturated flow equations has been geared mostly towards infiltration problems (Wooding, 1968; Philip, 1969, 1992). While some analytical solutions are widely used, for example, for evaluating tension disc experiments (Wooding, 1968), designing irrigation systems (Philip, 1992), or studying flow around buried objects (Warrick and Knight, 2002), many others appear to be only of academic interest.

NUMERICAL METHODS

Numerical methods are superior to analytical methods in terms of being able to solve practical problems. They allow users to design complicated geometries that reflect complex natural geologic and hydrologic conditions, control parameters in space and time, prescribe realistic initial and boundary conditions, and to implement nonlinear constitutive relationships. Numerical methods usually subdivide the time and spatial coordinates into smaller pieces, such as finite differences, finite elements, and/or finite volumes, and reformulate the continuous form of governing partial differential equations in terms of a system of algebraic equations. In order to obtain solutions at certain times, numerical methods generally require intermediate simulations (time-stepping) between the initial condition and the points in time for which the solution is needed. The following two sections review the history of development of various numerical techniques used in vadose zone flow and transport models. The review is based in part on an earlier review by van Genuchten and Šimůnek (1996). After reviewing various numerical techniques I will also discuss the need for efficient matrix solvers for 2D and 3D models, and provide a list of some of the most often used flow/transport models.

Numerical Solution of Richards Equation

A variety of numerical methods may be used to solve the variably saturated flow equation. The popularity of numerical methods stems from the fact that the Richards equation can be solved analytically only for a very limited number of cases. Even so, the highly nonlinear nature of the Richards equation also hampered the development of computationally efficient numerical methods that are stable under all conditions, particularly for infiltration in very dry soils. Stable numerical solutions still require relatively fine discretizations of both the time and space domains, often resulting in excessive CPU and simulation times, especially for two- and three-dimensional problems and/or problems involving highly transient boundary conditions.

Early applications of numerical methods for solving variably saturated flow problems generally used classical finite differences. Integrated finite difference, finite volumes, and especially, finite element methods (Neuman, 1973; Huyakorn *et al.*, 1986; Šimůnek *et al.*, 1999a) became increasingly popular in the seventies and thereafter. While finite difference methods today are used in a majority of one-dimensional models, finite volume methods and/or finite element methods coupled with mass lumping of the mass balance term are usually used in two- and three-dimensional models. Finite element methods used with unstructured triangular and tetrahedral elements allow for a more precise description of complex transport domains compared to finite differences.

The Richards equation can be formulated in three different ways. A mixed formulation arises when both the water content and the pressure head variables appear simultaneously in the governing equation, such as in equation (1). The h-based formulation is obtained when the time derivative of the water content ($\partial\theta/\partial t$) on the left side of

(1) is rewritten using the soil water capacity C as follows: $C\partial h/\partial t$, where C is defined as the slope of the retention curve, that is, $\mathrm{d}\theta/\mathrm{d}h$. The θ-based formulation is obtained when the product of the hydraulic conductivity $K(h)$ and the pressure head gradient ($\partial h/\partial z$) on the right side of (1) is replaced with the product of the water diffusivity $D_{\mathrm{w}}(\theta)$ and the water content gradient, $\partial\theta/\partial z$. The θ-formulation allows for very efficient numerical solutions, even for infiltration into initially dry soils. The use of this formulation is, however, straightforward only for homogeneous and unsaturated soils. This is because, the driving force for water movement is not the water content gradient, but the pressure head gradient; water in heterogeneous soils hence does not necessarily flow from locations with a higher water content to locations with a lower water content. Special provisions must be taken for such heterogeneous systems (Hills *et al.*, 1989), a likely reason why θ-formulations are rarely used in numerical models.

Celia *et al.* (1990) suggested that numerical solutions based on the standard h-based formulation of the Richards equation often yield poor results, characterized by relatively large mass balance errors and incorrect predictions of the pressure head distributions in the soil profile. They solved the mixed formulation of the Richards equation using a modified Picard iteration scheme that possesses mass-conserving properties for both finite element and finite difference spatial approximations. Milly (1985) earlier presented two mass-conservative schemes for computing nodal values of the water capacity in the h-based formulation to force global mass balance. Several highly efficient numerical schemes based on different types of pressure head transformations were presented recently by Hills *et al.* (1989), Ross (1990), and Kirkland *et al.* (1992). Hills *et al.* (1989) showed that the θ-based form of the Richards equation can yield fast and accurate solutions for infiltration into very dry heterogeneous soil profiles. However, the θ-based numerical scheme cannot be used for soils having saturated regions. Kirkland *et al.* (1992) expanded the work of Hills by combining the θ-based and h-based models to yield a transformation method applicable also to variably saturated systems. Their approach involved a new dependent variable, being a linear function of the pressure head and the water content in the saturated and unsaturated zones, respectively. Additional transformations of the Richards equation, with the common goal of decreasing its nonlinearity and increasing the efficiency of the numerical solution, were reviewed by Williams *et al.* (2000). Many of these transformations were, however, used mostly by the authors themselves and have not reached wide acceptance. Most popularly used vadose zone flow models presently utilize the mixed formulation of Celia *et al.* (1990); these models include SWAP (van Dam *et al.*, 1997) and HYDRUS (Šimůnek *et al.*, 1998, 1999a).

Time and space discretization of the Richards equation generally leads to a nonlinear system of algebraic equations. These equations are most often linearized and solved using the Newton–Raphson or Picard iteration methods. Picard iteration is widely used in unsaturated zone models because of its ease of implementation, and because this method preserves symmetry of the final system of equations. The Newton–Raphson iteration procedure is more complex and results in nonsymmetric matrices, but often achieves a faster rate of convergence and can be more robust than Picard iteration for certain types of problems (Paniconi and Putti, 1994).

Numerical Solution of the Transport Equation

A large number of methods are also available to numerically solve the convection-dispersion solute transport equation. These methods may be broadly classified into three groups: (i) Eulerian, (ii) Lagrangian, and (iii) mixed Lagrangian–Eulerian methods. In the Eulerian approach, the transport equation is discretized by means of a usual finite difference or finite element method using a fixed grid system. For the Lagrangian approach, the mesh moves along with the flow or remains fixed in a deforming coordinate system. A two-step procedure is followed for a mixed Lagrangian-Eulerian approach. First, convective transport is considered using a Lagrangian approach in which Lagrangian concentrations are estimated from particle trajectories. Subsequently, all other processes including sinks and sources are modeled with an Eulerian approach using any finite element or finite differences method, leading to the final concentrations.

Standard finite difference and Galerkin-type finite element methods belong to the first group of Eulerian methods. Finite differences and finite elements methods provided the early tools for solving solute transport problems and still are the most popular methods being used. Numerical experiments have shown that both methods give good results for transport in which dispersion is a relatively dominant process (e.g. as indicated by the grid Peclet number, $v\Delta x/D$, where v is pore-water velocity and Δx the grid size). However, both methods can lead to significant numerical oscillations and/or artificial dispersion for convection-dominated transport problems. Still, a relatively simple method may be used to prevent or limit numerical oscillations. By selecting an appropriate combination of relatively small space and time steps, it is possible to virtually eliminate all oscillations. Perrochet and Berod (1993) developed criteria for minimizing or eliminating numerical oscillations based on a "performance index". They concluded that all oscillations should be eliminated when the performance index, defined as the product of the local Peclet and Courant ($v\Delta t/\Delta x$) numbers, is less than 2. When small oscillations in the solution can be tolerated, the performance index can be easily increased to about 5 or 10 (Perrochet and Berod, 1993).

One alternative is the use of upwind finite difference methods. This method virtually eliminates numerical oscillations, even for purely convective transport. The disadvantage of this method is that it may create significant and often unacceptable numerical dispersion. Similarly, upstream weighting has been proposed for finite elements. The method consists of using different weighting functions for terms having spatial derivatives than for other terms in the transport equation. This approach places greater weight on the upstream nodes within a particular element.

While Lagrangian methods (or method of characteristics) virtually eliminate problems with numerical oscillations, they may introduce other problems, notably artificial dispersion and nonconservative solutions. Lagrangian methods are also relatively difficult to implement in two and three dimensions. Instabilities resulting from inappropriate spatial discretizations may occur during longer simulations because of deformation of the stream function. Furthermore, nonrealistic distortions of the results may occur when modeling the transport of solutes that are subjected to different sorption/exchange or precipitation reactions.

The Eulerian and Lagrangian approaches can also be combined to the mixed Eulerian–Lagrangian method. Because of the different mathematical nature of the diffusive (parabolic) and convective (hyperbolic) terms in the convection-dispersion equation, the transport equation is best decomposed into a mixed problem consisting of a purely convective problem, followed by a pure diffusion-only problem. Methods based on this approach are called *operator-splitting* or *splitting-up methods*. Convective transport is then solved with the Lagrangian approach, while all other terms of the transport equation are solved using Eulerian methods.

Still other solutions exist, such as the use of a so-called "random walk" process or a combination of analytical and numerical techniques. In a random walk approach, solute transport is modeled using a large number of particles. Displacement of each particle during each time step is given by a certain distance, this being the sum of two velocity contributions – a deterministic and stochastic contribution. Studies with this type of method indicate that it may be necessary to use many thousands of particles in order to obtain relatively precise results. An example of a combination of analytical and numerical techniques is given by Sudicky (1989) who modeled solute transport using Laplace transforms with respect to time, and Galerkin finite elements for the spatial domain. The use of Laplace transforms avoids the need for intermediate simulations (time-stepping) between the initial condition and the points in time for which solutions are needed, while also less stringent requirements are needed for the spatial discretization. This combination of analytical and numerical techniques, unfortunately, has however, one important limitation. Since Laplace transforms eliminate time as an independent variable in the governing transport equation, all coefficients such as water content, flow velocity, and retardation factors, must be independent of time. This means that combination methods can only solve solute transport problems during steady-state water flow, and hence are inappropriate for transient variably saturated flow situations typical of most field problems.

Many of the above methods for numerically solving transport equation were developed primarily for saturated conditions, for which coarse spatial discretizations and large flow velocities often produce large Peclet numbers, which may lead to significant numerical oscillations. Flow velocities in the vadose zone are usually much smaller. Also, the nonlinearity of the Richards equation generally requires numerical solutions on much finer spatial grids than in groundwater studies. Consequently, the Peclet numbers are significantly smaller for vadose zone applications than in groundwater flow studies, thereby allowing the adoption of more oscillation-prone methods. The majority of vadose zone models therefore can use relatively standard finite element or finite difference methods, which, if combined with upwind or upstream weighting or proper self-adjusting or other time step scheme, should eliminate most or all numerical oscillations.

Matrix Solvers

Discretization and subsequent linearization (as needed) of the governing partial differential equations for water flow and solute transport leads to a system of linear equations

$$[\mathbf{A}]\{\mathbf{x}\} = \{\mathbf{b}\} \tag{6}$$

in which $\{\boldsymbol{x}\}$ is an unknown solution vector, $\{\boldsymbol{b}\}$ is the known right-hand-side vector of the matrix equation, and $[\boldsymbol{A}]$ is a sparse banded matrix, which is symmetric for water flow if the modified Picard procedure is used, but asymmetric for water flow if the Newton–Raphson method is used. Matrix $[\boldsymbol{A}]$ is generally asymmetric for solute transport, unless convection is not considered in the formulation. Matrix $[\boldsymbol{A}]$ is tridiagonal for one-dimensional applications and thus can be solved very quickly and efficiently using simple Gaussian elimination. For higher-dimensional application, matrix $[\boldsymbol{A}]$ is a sparse matrix with a number of lines equal to the number of nodes in the spatial discretization scheme. The number of nodes is on the order of tens of thousands for a typical two-dimensional application, and on the order of millions for a three-dimensional application. Since this matrix needs to be inverted many times during a typical numerical run (at each time step for solute transport, and additionally also at each iteration for unsaturated water flow), the need for very efficient solvers cannot be underestimated.

Traditionally, the matrix equations have been solved using direct methods, such as Gaussian elimination or LU decomposition. These methods usually take advantage of the banded nature of the coefficient matrices and, in the case of water flow, of the symmetric properties of the matrix. Direct solution methods have several disadvantages as compared to iterative methods. For example, they require a fixed number of operations (depending upon the size of the matrix), which increases approximately by the square of the number of nodes. Iterative methods, on the other hand, require a variable number of repeated steps, with the number increasing at a much smaller rate (about 1.5) with the size of a problem (Mendoza *et al.*, 1991). A similar reduction also holds for the memory requirement since iterative methods do not require one to store nonzero matrix elements. Memory requirements, therefore, increase at a much smaller rate with the size of the problem when iterative solvers are used. This memory requirement is associated with the need to minimize the size of the band (i.e. the largest distance between two neighboring nodal numbers) of matrix [***A***] for direct methods. While minimization of the matrix band is trivial for finite difference methods, it can be rather complex when unstructured triangular finite element meshes are used. Round-off errors also represent less of a problem for iterative methods as compared with direct methods. This is because round-off errors in iterative methods are self-correcting. Finally, for time-dependent problems, a reasonable approximation of the solution (i.e. the solution at the previous time step) exists for iterative methods, but not for direct methods. In general, direct methods are more appropriate for relatively small problems and for finite difference codes, while iterative methods are more suitable for larger problems and codes using unstructured finite element grids.

While many iterative methods have been used in the past for handling large sparse matrix equations, a variety of increasingly powerful preconditioned accelerated iterative methods, such as the preconditioned conjugate gradient method (PCG), are now becoming available also. Since the system of linear equations resulting from discretization of the solute transport equation is nonsymmetrical, it is necessary to either formulate the transport problem in such a way that it leads to a symmetric matrix, or to use an extension of PCG for nonsymmetrical matrices, such as ORTHOMIN (generalized conjugate residual method) (Mendoza *et al.*, 1991), GMRES (generalized minimal residual method), biconjugate gradients, TFQMR (transpose-free quasi-minimal residual algorithm), CGSTAB (conjugate gradient stabilized method), and conjugate gradient squared procedures. Both the preconditioned conjugate gradient and ORTHOMIN methods consist of two essential parts: initial preconditioning, and iterative solution with either conjugate gradient, CGSTAB, or ORTHOMIN acceleration (Mendoza *et al.*, 1991). Incomplete lower-upper (ILU) factorization can be used as preconditioning of matrix [***A***], which is then factorized into lower and upper triangular matrices by partial Gaussian elimination. The preconditioned matrix is subsequently inverted repeatedly using updated estimates to provide a new approximation of the solution.

Available Model for Unsaturated Zone

Most of the early models developed for studying processes in the near-surface environment focused mainly on variably saturated water flow. They were used primarily in agricultural research for the purpose of optimizing moisture conditions to increase crop production. This focus has increasingly shifted to environmental research, with the primary concern now being the subsurface fate and transport of various agricultural and other contaminants, such as pesticides, nutrients, pathogens, pharmaceuticals, viruses, bacteria, colloids, toxic trace elements, and/or fumigants, and also the evaluation of water recharge through the vadose zone. While the earlier models solved the governing equations (1) through (3) for relatively simplified system-independent boundary conditions (i.e. specified pressure heads or fluxes, and free drainage), models developed recently can cope with much more complex system-dependent boundary conditions evaluating surface flow and energy balances and accounting for the simultaneous movement of water, vapor, and heat. Examples are DAISY (Hansen *et al.*, 1990), TOUGH2 (Pruess, 1991), SHAW (Flerchinger *et al.*, 1996), SWAP (van Dam *et al.*, 1997), HYDRUS-1D (Šimůnek *et al.*, 1998), UNSATH (Fayer, 2000), and COUP (Jansson and Karlberg, 2001). Several models now account also for the extremely nonlinear processes associated with the freezing and thawing cycle (e.g. DAISY, SHAW, and COUP).

Models have recently also become increasingly sophisticated in terms of the type and complexity of solute transport processes that can be simulated. Transport models are no longer being limited to solutes undergoing relatively simple chemical reactions such as linear sorption and first-order decay, but now consider also a variety of nonlinear sorption and exchange processes, physical and chemical nonequilibrium transport, volatilization, gas diffusion, colloid attachment/detachment, decay chain reactions, and many other processes (e.g. the HYDRUS-1D and -2D codes of Šimůnek *et al.*, 1998, 1999a). For example, the general formulation of the transport equations in the HYDRUS codes allows one to simulate not only nonadsorbing or linearly sorbing chemicals but also a variety of other contaminants, such as viruses (Schijven and Šimůnek, 2002), colloids (Bradford *et al.*, 2002), cadmium (Seuntjens *et al.*, 2001), and hormones (Casey *et al.*, 2003), or chemicals involved in the sequential biodegradation of chlorinated

aliphatic hydrocarbons (Schaerlaekens *et al.*, 1999; Casey and Šimůnek, 2001).

Options to simulate carbon and nitrogen cycles are also becoming a standard feature of many environmental models, such as DAISY, LEACHN (Hutson and Wagenet, 1992), RZWQM (Ahuja and Hebson, 1992), and COUP. These models typically distribute organic matter, carbon, and organic and mineral nitrogen over multiple computational pools, while allowing organic matter to be decomposed by multiple microbial biomass populations. They can account for most of the major reaction pathways, such as mineralization-immobilization of crop residues, manure and other organic wastes, mineralization of the soil humus fractions, interpool transfer of carbon and nitrogen, nitrification (ammonium to nitrate-N), denitrification (leading to the production of N_2 and N_2O), volatilization loss of ammonia (NH_3), production and consumption of methane (CH_4) and carbon dioxide (CO_2), changes in the carbon nitrogen ratio of organic matter, and microbial biomass growth and death (Ahuja and Hebson, 1992).

Efforts are also on to couple physical flow and transport models with geochemical models to simulate even more complex reactions, such as surface complexation, precipitation/dissolution, cation exchange, and/or biological reactions (e.g. Ahuja and Hebson, 1992; Šimůnek and Suarez, 1994; Šimůnek and Valocchi, 2002; Jacques *et al.*, 2002). Models considering these chemical reactions, including the ability to simulate the transport of multiple chemical species and carbon dioxide, are required for studying water management practices and irrigation techniques under arid and semiarid conditions, evaluation of water suitability for irrigation, and reclamation of sodic soils (Šimůnek and Suarez, 1997).

Another active area of research involves attempts to extend existing models that simulate uniform flow to situations where nonequilibrium and/or preferential flow occurs. Examples of this are the MACRO (Jarvis, 1994) and HYDRUS-1D (Šimůnek *et al.*, 2003) models. Possible approaches for simulating preferential flow differ in terms of their underlying assumptions and complexity. They range from relatively simplistic models to more complex, physically based, dual-porosity, dual-permeability, and multiregion-type models. A relatively simple dual-porosity flow model results when the Richards equation is combined with composite (double-hump type) equations for the hydraulic properties to account for both soil textural (matrix) and soil structural (fractures, macropores, peds) effects on flow. A more complex dual-porosity, mobile-immobile water flow model results when the Richards or kinematic wave equations are used for flow in the fractures, and immobile water is assumed to exist in the matrix. Even more complex are various dual-permeability models such as the formulations of Gerke and van Genuchten (1993) and Pruess (1991), or the kinematic wave approach as used in the MACRO model of Jarvis (1994). These formulations all assume that water is mobile in both the matrix and fracture domains, while invoking terms that account for the exchange of water and solutes between the matrix and the fractures.

A large number of models are now available for simulating processes in the vadose zone. Some of these models are in the public domain, such as MACRO, SWAP, UNSATH (Fayer, 2000), VS2DI (Healy, 1990), and HYDRUS-1D (Šimůnek *et al.*, 1998), while others are in the commercial domain, such as HYDRUS-2D (Šimůnek *et al.*, 1999a). These models vary widely in terms of their complexity, sophistication, and ease of use. Although some models are still being run under the DOS operating system, with associated difficulties of preparing input files and interpreting tabulated outputs, many others, especially those in the commercial domain, are supported by sophisticated graphics-based interfaces that tremendously simplify their use (Šimůnek *et al.*, 1998, 1999a).

Table 1 gives a summary of some of the more widely used numerical models for simulating variably saturated water flow and solute transport in soils. This table also provides Internet addresses and references where additional information about each model can be found. With the exception of HYDRUS-2D, TOUGH2, and VS2DTI, all models given in Table 1 are one-dimensional models, perhaps reflecting the fact that the majority of applications for unsaturated zone models is still only one dimensional.

CONCLUSIONS

Much of the research in the field of soil science has focused in recent decades upon understanding the fundamentals of variably saturated water flow and pollutant fate and transport processes. As society continues its rapid technological development, the types of pollution problems and chemicals posing significant environmental threats, have become increasingly complex. Problems such as the transport of pesticides, colloids, bacteria, viruses, pharmaceuticals, reproductive hormones, nutrients, and toxic trace elements, carbon sequestration, and bioremediation of organic contaminants, all require a thorough understanding and coupling of multiple hydrogeological, geochemical, and microbiological processes. It is the continually increasing speed and power of modern computers that will enable such models to become convenient tools for analysis of complex geochemical systems. Although more and more complex models are being constantly developed, currently available models are still relatively specialized and no single model is presently available that can describe the multiple problems and chemicals mentioned above. Development of numerical models capable of describing unstable and/or preferential flow, as well as models coupled with sophisticated geochemical models capable of describing both instantaneous

Table 1 Some of the widely used numerical models for simulating variably saturated water flow and solute transport in soils

Model name	Internet address, reference, interesting/special features
COUP	**http://www.lwr.kth.se/Vara**%20 Datorprogram/CoupModel/ (Jansson and Karlberg, 2001) Carbon/nitrogen cycle, thawing/freezing cycle, coupled water, vapor, and heat transport, crop growth
DAISY	**http://www.dina.dk/~daisy/** (Hansen *et al.*, 1990) Carbon/nitrogen cycle, crop growth, management practices, pesticide processes
HYDRUS-1D	**http://www.hydrus2d.com** (Šimůnek *et al.*, 1998) Multiple soil hydraulic functions, neural network-based pedotransfer functions, nonlinear nonequilibrium solute transport, mobile-immobile and two-site sorption concepts, chain reactions, volatilization, inverse option, intuitive sophisticated graphical interface
HYDRUS-2D	**http://www.hydrus2d.com** (Šimůnek *et al.*, 1999a) Two-dimensional, multiple soil hydraulic functions, neural network-based pedotransfer functions, nonlinear nonequilibrium solute transport, mobile-immobile and two-site sorption concepts, chain reactions, inverse option, unstructured triangular finite element meshes, intuitive sophisticated graphical interface
MACRO	**http://www.mv.slu.se/BGF/Macrohtm/macro.htm** (Jarvis, 1994) Preferential flow using kinematic wave equation, snow accumulation, pesticide transport
RZWQM	**http://gpsr.ars.usda.gov/products/rzwqm.htm** (Ahuja and Hebson, 1992) Complex modular program, crop growth, chemical equilibrium module, management practices, pesticide processes
SHAW	**http://www.nwrc.ars.usda.gov/models/shaw** (Flerchinger *et al.*, 1996) Thawing/freezing cycle, coupled water, vapor, and heat transport, multispecies plant canopy
SWAP	**http://www.swap.alterra.nl/** (van Dam *et al.*, 1997) A three-level drainage system at regional scale, crop growth
SWIM	**http://www.clw.csiro.au/products/swim** (Verburg *et al.*, 1996) Bypass flow, flexible description of hydraulic properties, hyperbolic sine transformation of the pressure head
TOUGH2	**http://www-esd.lbl.gov/TOUGH2/** (Pruess, 1991) Multidimensional multiphase fluid and heat flow, dual permeability
UNSATH	**http://hydrology.pnl.gov/resources/unsath/unsath_download.asp** (Fayer, 2000) Coupled water, vapor, and heat transport, no solute transport
VS2DTI	**http://water.usgs.gov/software/vs2di.html** (Healy, 1990) Two-dimensional, finite differences

and kinetic chemical and biological reactions will undoubtedly remain a focus of research in the near future.

The accuracy of the obtained predictions depend to a large extent upon the accuracy of available model input parameters and upon proper conceptualization of soil heterogeneity and other system complexities, such as the possible presence of nonequilibrium flow and transport, including preferential flow. Processes are often described and their parameters measured on a much smaller scale than those for which the model predictions are being sought. Consequently, many model parameters often need to be calibrated so that they reflect the bulk behavior of the heterogeneous system, in which case they can be used for larger scale predictions. New measuring techniques that provide model parameters on the scale at which predictions are made are badly needed for successful applications of unsaturated flow and transport models in a predictive mode at the larger scale.

One may expect that unsaturated zone flow and transport models will be used increasingly for integrating fundamental knowledge about the vadose zone to yield tools for developing cost-effective, yet technically sound strategies for resource management and pollution remediation and prevention. Unsaturated zone transport models are indispensable tools for analyzing complex environmental pollution problems, and for developing practical management strategies. Models can help guide field observations by identifying which parameters and processes control system behavior. Following Steefel and Van Cappellen (1998) and Šimůnek and Valocchi (2002), several specific key ways in which unsaturated flow and transport models can be used are identified below:

1. Physical, chemical, and biological processes are often studied in isolation either in the laboratory or in the field under controlled conditions. Mathematical models can be used to investigate the impacts of multiple coupled biogeochemical reactions and other interactions in the presence of complex flow fields and spatial heterogeneity. These models also enable extrapolation to environmentally relevant temporal and spatial scales.
2. Numerical transport models provide a useful tool for interpreting experimental results. Models can help understand qualitative and quantitative trends and relationships present in the data. Properly applied modeling to interpret results of field experiments can lead to more effective quantitative understanding of underlying biogeochemical processes.
3. One of the most powerful applications of numerical flow and transport models is conducting sensitivity

analyses. Such analyses permit a systematic evaluation of the impact of model parameters (physical, chemical, and/or biological), initial conditions, and boundary conditions upon the model output. The results of a sensitivity analysis provide insight into the relative importance of individual processes and reactions within a complex biogeochemical system. Results can help one identify the most important parameters and processes, and thereby provide guidance in allocation of resources for laboratory and field investigations.

4. Numerical flow and transport models are tools for integrating all of our knowledge obtained from simulations, sensitivity analyses, and laboratory and field experimentation. This integration will often lead to more coherent and rigorous conceptual models for the underlying coupled flow, transport, and reactions processes.

REFERENCES

Ahuja L.R. and Hebson C. (1992) *Root Zone Water Quality Model*, GPSR Technical Report No. 2, USDA, ARS, Fort Collins.

Bradford S.A., Yates S.R., Bettehar M. and Šimůnek J. (2002) Physical factors affecting the transport and fate of colloids in saturated porous media. *Water Resources Research*, **38**(12), 1327, doi:10.1029/2002WR001340, 63.1–63.12.

Casey F.X.M., Larsen G.L., Hakk H. and Šimůnek J. (2003) Fate and transport of 17β-Estradiol in soil-water systems. *Environmental Science and Technology*, **37**(11), 2400–2409.

Casey F.X.M. and Šimůnek J. (2001) Inverse analyses of the transport of chlorinated hydrocarbons subject to sequential transformation reactions. *Journal of Environmental Quality*, **30**(4), 1354–1360.

Celia M.A., Bououtas E.T. and Zarba R.L. (1990) A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research*, **26**, 1483–1496.

Fayer M.J. (2000) *UNSAT-H Version 3.0: Unsaturated Soil Water and Heat Flow Model*. Theory, User Manual, and Examples. Pacific Northwest National Laboratory 13249.

Flerchinger G.N., Hanson C.L. and Wight J.R. (1996) Modeling evapotranspiration and surface energy budgets across a watershed. *Water Resources Research*, **32**, 2539–2548.

Gardner W.R. (1958) Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science*, **85**, 228–232.

Gerke H.H. and van Genuchten M.T.h (1993) A dual-porosity model for simulating the preferential movement of water and solutes in structured porous media. *Water Resources Research*, **29**, 305–319.

Hansen S., Jensen H.E., Nielsen N.E. and Svendsen, H. (1990) *DAISY: Soil Plant Atmosphere System Model*, NPO Report No. A 10, The National Agency for Environmental Protection, Copenhagen, pp. 272.

Healy R.W. (1990) *Simulation of Solute Transport in Variably Saturated Porous Media with Supplemental Information on Modifications to the U.S. Geological Survey's Computer Program VS2DI*, Water-Resources Investigation Report 90–4025, U.S. Geological Survey, p. 125.

Hills R.G., Hudson D.B., Porro I. and Wierenga P.J. (1989) Modeling one-dimensional infiltration into very dry soils, 1. Model development and evaluation. *Water Resources Research*, **25**(6), 1259–1269.

Hutson J.L. and Wagenet R.J. (1992) *LEACHM:Leaching Estimation and Chemistry Model*, Research Series no. 92-3, Cornell University: Ithaca.

Huyakorn P.S., Springer E.P., Guvanasen V. and Wadsworth T.D. (1986) A three-dimensional finite-element model for simulating water flow in variably saturated porous media. *Water Resources Research*, **22**, 1790–1808.

Jacques D., Šimůnek J., Mallants D. and van Genuchten M.T.h (2002) Multicomponent transport model for variably-saturated porous media: application to the transport of heavy metals in soils. In *Computational Methods in Water Resources*, *Developments in Water Science 47*, Hassanizadeh S.M., Schotting R.J., Gray W.G. and Pinder G.F. (Eds.), *XIVth International Conference on Computational Methods in Water Resources*, June 23–28 2002, Elsevier: Delft, pp. 555–562.

Jansson P.-E. and Karlberg L. (2001) *Coupled Heat and Mass Transfer Model for Soil-Plant-Atmosphere Systems*, Royal Institute of Technology, Department of Civil and Environmental Engineering: Stockholm, p. 325.

Jarvis N.J. (1994) *The MACRO Model (Version 3.1), Technical Description and Sample Simulations*, Reports and Dissertations 19, Department Soil Science Swedish University of Agricultural Sciences, Uppsala, p. 51.

Jury W.A., Spencer W.F. and Farmer W.J. (1983) Behavior assessment model for trace organics in soil: I. Model description. *Journal of Environmental Quality*, **12**(4), 558–564.

Kirkland M.R., Hills R.G. and Wierenga P.J. (1992) Algorithms for solving Richards' equation for variably saturated soils. *Water Resources Research*, **28**(8), 2049–2058.

Leij F.J. and Bradford S.A. (1994) *3DADE: A Computer Program for Evaluating Three-Dimensional Equilibrium Solute Transport in Porous Media*, Research Report No. 134, U. S. Salinity Laboratory, USDA, ARS, Riverside.

Leij F.J., Priesack E. and Schaap M.G. (2000) Solute transport modeled with green's functions with application to persistent solute sources. *Journal of Contaminant Hydrology*, **41**, 155–173.

Leij F.J. and Toride N. (1997) *N3DADE: A Computer Program for Evaluating Nonequilibrium Three-Dimensional Equilibrium Solute Transport in Porous Media*, Research Report No. 143, U. S. Salinity Laboratory, USDA, ARS, Riverside.

Mendoza C.A., Therrien R. and Sudicky E.A. (1991) *ORTHOFEM User's Guide*, Version 1.02. Waterloo Centre for Groundwater Research, University of Waterloo: Waterloo, Ontario.

Milly P.C.D. (1985) A mass-conservative procedure for time-stepping in models of unsaturated flow. *Advances in Water Resources*, **8**, 32–36.

Neuman S.P. (1973) Saturated-unsaturated seepage by finite elements. *Proceedings of the ASCE, Journal of Hydraulic Division*, **99**, 2233–2250.

Paniconi C. and Putti M. (1994) A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems. *Water Resources Research*, **30**(12), 3357–3374.

Parker J.C. and van Genuchten M.T.h (1984) *Determining Transport Parameters from Laboratory and Field Tracer Experiments*, Bulletin 84-3, Va Agricultural Experiment Station: Blaacksburg.

Perrochet P. and Berod D. (1993) Stability of the standard Crank-Nicholson-Galerkin scheme applied to the diffusion-convection equation: some new insights. *Water Resources Research*, **29**(9), 3291–3297.

Philip J.R. (1969) Theory of infiltration. *Advances in Hydroscience*, **5**, 215–296.

Philip J.R. (1992) What happens near a quasi-linear point source? *Water Resources Research*, **28**, 47–52.

Pruess K. (1991) *Tough2 – A General-Purpose Numerical Simulator for Multiphase Fluid and Heat Flow*, Report LBL-29400, Lawrence Berkeley Laboratory, Berkeley.

Ross P.J. (1990) Efficient numerical methods for infiltration using Richards' equation. *Water Resources Research*, **26**(2), 279–290.

Schaerlaekens J., Mallants D., Šimůnek J., van Genuchten M.T.h and Feyen J. (1999) Numerical simulation of transport and sequential biodegradation of chlorinated aliphatic hydrocarbons using CHAIN_2D. *Journal of Hydrological Processes*, **13**(17), 2847–2859.

Schijven J. and Šimůnek J. (2002) Kinetic modeling of virus transport at field scale. *Journal of Contaminant Hydrology*, **55**(1–2), 113–135.

Seuntjens P., Tirez K., Šimůnek J., van Genuchten M.T.h, Cornelis C. and Geuzens P. (2001) Aging effects on cadmium transport in undisturbed contaminated sandy soil columns. *Journal of Environmental Quality*, **30**, 1040–1050.

Šimůnek J., Jarvis N.J., van Genuchten M.T.h and Gärdenäs A. (2003) Nonequilibrium and preferential flow and transport in the vadose zone: review and case study. *Journal of Hydrology*, **272**, 14–35.

Šimůnek J. and Suarez D.L. (1994) Two-dimensional transport model for variably saturated porous media with major ion chemistry. *Water Resources Research*, **30**(4), 1115–1133.

Šimůnek J. and Suarez D.L. (1997) Sodic soil reclamation using multicomponent transport modeling. *Journal of Irrigation and Drainage Engineering-ASCE*, **123**(5), 367–376.

Šimůnek J., Šejna M. and van Genuchten M.T.h (1998) *The HYDRUS-1D Software Package for Simulating the One-Dimensional Movement of Water, Heat, and Multiple Solutes in Variably-Saturated Media*, Version 2.0, IGWMC – TPS-70, International Ground Water Modeling Center, Colorado School of Mines: Golden, p. 202.

Šimůnek J., Šejna M. and van Genuchten M.T.h (1999a) *The HYDRUS-2D Software Package for Simulating Two-Dimensional Movement of Water, Heat, and Multiple Solutes in Variably Saturated Media*, Version 2.0, IGWMC – TPS-53, International Ground Water Modeling Center, Colorado School of Mines: Golden, p. 251.

Šimůnek J., van Genuchten M.T.h, Šejna M., Toride N. and Leij F.J. (1999b) *The STANMOD Computer Software for Evaluating Solute Transport in Porous Media Using Analytical Solutions of Convection-Dispersion Equation*, Versions 1.0 and 2.0, IGWMC – TPS-71, International Ground Water Modeling Center, Colorado School of Mines: Golden, p. 32.

Šimůnek J. and Valocchi A.J. (2002) Geochemical transport. In *Methods of Soil Analysis, Part 1, Physical Methods*, *Third edition*, Chap. 6.9, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America, Madison, pp. 1511–1536.

Steefel C.I. and Van Cappellen P. (1998) Reactive transport modeling of natural systems. *Journal of Hydrology*, **209**, 1–7.

Sudicky E.A. (1989) The Laplace transform Galerkin technique: a time-continuous finite element theory and application to mass transport in groundwater. *Water Resources Research*, **25**(8), 1833–1846.

Toride N., Leij F.J. and van Genuchten M.T.h (1995) *The CXTFIT Code for Estimating Transport Parameters from Laboratory or Field Tracer Experiments*, Version 2.0, Research Report No. 137, U. S. Salinity Laboratory, USDA, ARS, Riverside.

van Dam J.C., Huygen J., Wesseling J.G., Feddes R.A., Kabat P., van Valsum P.E.V., Groenendijk P. and van Diepen C.A. (1997) *Theory of SWAP, Version 2.0. Simulation of Water Flow, Solute Transport and Plant Growth in the Soil- Water-Atmosphere- Plant Environment*, Department Water Resources, WAU, Report 71, DLO Winand Staring Centre, Technical Document 45, Wageningen.

van Genuchten M.T.h (1980) *Determining Transport Parameters from Solute Displacement Experiments*, Research Report No. 118, U. S. Salinity Laboratory, USDA, ARS, Riverside.

van Genuchten M.T.h (1981) *Non-Equilibrium Transport Parameters from Miscible Displacement Experiments*, Research Report No. 119, U. S. Salinity Laboratory, USDA, ARS, Riverside.

van Genuchten M.T.h (1985) Convective-dispersive transport of solutes involved in sequential first-order decay reactions. *Computers & Geosciences*, **11**(2), 129–147.

van Genuchten M.T.h and Šimůnek J. (1996) Evaluation of pollutant transport in the unsaturated zone. In *Regional Approaches to Water Pollution in the Environment*, *NATO ASI Series: 2. Environment*, Rijtema P.E. and Eliáš V. (Eds.), Kluwer: Dordrecht, 139–172.

Verburg K., Ross P.J. and Bristow K.L. (1996) *SWIMv2.1 User Manual*, Divisional Report 130, CSIRO.

Warrick A.W. and Knight J.H. (2002) Two-dimensional unsaturated flow through a circular inclusion. *Water Resources Research*, **38**(7), 18.1–18.6, 10.1029/2001WR001041.

Williams G.A., Miller C.T. and Kelley C.T. (2000) Transformation approaches for simulating flow in variably saturated porous media. *Water Resources Research*, **36**(4), 923–934.

Wooding R.A. (1968) Steady infiltration from large shallow circular pond. *Water Resources Research*, **4**, 1259–1273.

Yeh G.T. (1981) *Analytical Transient One-, Two-, and Three-Dimensional Simulation of Waste Transport in the Aquifer System*, ORNL-5602, Oak Ridge National Laboratory: Oak Ridge.

79: Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport

JAMES D BROWN[1] AND GERARD BM HEUVELINK[2]

[1] *Universiteit van Amsterdam, Nieuwe Achtergracht, Amsterdam, The Netherlands*
[2] *Wageningen University, Wageningen, The Netherlands*

Soil hydrological models are inherently imperfect because they abstract and simplify "real" hydrological patterns and processes. Indeed, an important aim of modeling is to establish the simplest description possible for adequately addressing a particular problem. Also, models are frequently based on input data that are known to be inadequate for some practical purpose. Thus, uncertainties in model outputs originate from uncertainties in input data, which include measurement and interpolation errors, and in models that include conceptual, logical, mathematical, and computational errors. Understanding the causes and consequences of uncertainty in soil hydrological modeling is useful for: (i) establishing the utility of data and models as decision-support tools; (ii) directing resources towards improving data and models, and (iii) seeking alternative ways of managing soils when the opportunities for accurate modeling are limited. This chapter focuses on statistical methods for assessing uncertainty in soil data and models, propagating uncertainties through models, and assessing the contribution of different sources of uncertainty to the overall uncertainties in model predictions. In addition, it explores the impacts of scale, and changes between scales, on the outcomes of an uncertainty analysis. It concentrates on physically based models of soil water flow and solute transport, and provides numerous examples from the literature here.

INTRODUCTION

Soils are an elementary yet highly complex component of many natural, modified, and managed ecosystems with physical and hydrological properties that vary both systematically through space and time and in a seemingly erratic way (Webster, 2000). Soil properties are a product of underlying physical, chemical, and biological processes and, in turn, affect these underlying processes. The relationships between soil properties and processes are usually complex and often involve nonlinear interactions and scale-dependencies that are difficult to explain or to predict (Addiscott and Tuck, 2001). These complexities are evident when modeling soil water flows and solute transport. Thus, while hydrological data and models have continued to improve over recent years, they are rarely certain or "error-free". Rather, the combined effects of unpredictable variations in soil properties and simplified representations of complex hydrological processes lead to errors in model outputs (e.g. Zhang *et al.*, 1993; Leenhardt, 1995; Dillah and Protopapas, 2000; Dubus and Brown, 2002). These errors may be sufficiently large to result in poor decisions about the exploitation and management of soils. For example, poor estimates of soil hydraulic properties may encourage excess irrigation, leading to soil erosion and salinization. While it is generally accepted that soil data and models are not "error-free", these errors may be difficult to quantify in practice. Indeed, the quantification of error implies that the "true" character of soils is known (i.e., error is a specific departure from "reality"). In the absence of such confidence, we are uncertain about the "true" properties and processes that characterize soils. Understanding how these uncertainties affect model predictions is important for: (i) establishing the utility of data and models

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

as decision-support tools; (ii) directing resources towards improving data and models; and (iii) seeking alternative ways of managing soils when the opportunities for accurate modeling are limited.

Uncertainties in soil data combine with uncertainties in hydrological models and lead to uncertainties in model predictions. This article focuses on uncertainty propagation through physically based hydrological models and, specifically, models describing soil water flow and solute transport through the unsaturated zone. Solute transport is important because many practical applications require knowledge about the distribution of chemicals in soils (e.g. soil acidification, heavy metal pollution, nutrient availability, salinization, and nitrate leaching), and much of the research on uncertainty propagation in soil hydrology has focused on solute transport (e.g. McKone, 1996; Foussereau *et al.*, 2001; Seuntjens *et al.*, 2002; Sohrabi *et al.*, 2002; Vachaud and Chen, 2002a; De Vries *et al.*, 2003). While there are longstanding literatures on uncertainty propagation in other areas of hydrology, including groundwater hydrology (e.g. Christensen and Cooley, 1999; Wang and McTernan, 2002) and surface hydrology (e.g. Beven and Binley, 1992; Brazier *et al.*, 2001; Beven and Freer, 2001), these are reviewed elsewhere (*see* articles **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**, **Chapter 77, Inverse Modeling of Soil Hydraulic Properties, Volume 2**, and **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**), and are only referred to in discussing methods for uncertainty propagation.

Uncertainties in soil data are discussed below under "Uncertainties in Model Inputs", while uncertainties in models are discussed in a later section. In the section "Uncertainty Propagation", the statistical procedures available for propagating uncertainties in model inputs and models through to uncertainties in model outputs are reviewed. Techniques for evaluating the contribution of different sources of uncertainty to the overall uncertainties in model predictions are discussed in the section "Evaluating the Contribution of Different Uncertainties". Finally, the section on "Scale and Uncertainty" considers the impacts of scale, and changes between scales, on the outcomes of an uncertainty analysis. First, however, it is useful to consider the nature of uncertainty and how it might be quantified.

WHAT IS UNCERTAINTY AND HOW CAN WE QUANTIFY IT?

Uncertainty is an expression of confidence about what we "know", both as individuals and communities, and is, therefore, subjective. Different people can reach different conclusions about how uncertain something is, based on their own personal experiences and world-view, as well as the amount and quality of information available to them (Cooke, 1991). It is not an inherent property of the environment (Quantum Theory can be ignored here), but may be encouraged in people by some aspects of the environment. For example, the environment may appear more complex than our abstractions and simplifications imply (e.g. kinetic processes in pesticide sorption), or too variable for us to capture (e.g. infiltration rates in Mediterranean soils), too large and interconnected (open) for us to observe everything at once (e.g. global weathering of minerals), or too small to observe at practical scales (e.g. soil pore volume in the field), too opaque for observation (e.g. hydraulic conductivity), or because we do not have the capacity to observe it (e.g. soil matrix over large areas). Uncertainty differs from ignorance, because ignorance involves a lack of awareness about our imperfect knowledge (Smithson, 1989). It also differs from error, because this involves a specific departure from "reality" (Heuvelink, 1998a).

Representing Uncertainty with Probability Distributions

Ideally, our representations of reality would be perfectly accurate or "error-free". In practice, however, they are not, and we are aware of this (otherwise we would be ignorant). Rather, we are uncertain about the errors in our representations. However, we may be able to specify some boundaries for our uncertainty, which would allow us to explore its impacts on the outcomes of a decision-making exercise (e.g. a hydrological model). For example, the nitrate concentration in a sample of soil water may be measured as $68.6\,g\,m^{-3}$. The "true" value remains unknown, but control experiments in a laboratory suggest that the measuring instrument is unbiased and has a standard deviation of $5\,g\,m^{-3}$. If measurement error is the only source of error in the sampled value and the laboratory measurements are relevant to the current situation, we know that the "true" value (t) lies within $58.8\,g\,m^{-3} < t < 78.4\,g\,m^{-3}$ 95% of the time. Thus, we can express our lack of confidence (uncertainty) about the "true" nitrate concentration with a probability distribution of possible nitrate concentrations (Heuvelink, 1998a). A probability distribution function (pdf) is characterized by its shape (e.g. Gaussian, exponential, and uniform) and by its parameter values. One important parameter is the variance, as this denotes the average magnitude of uncertainty in the variable of interest.

In order to represent uncertainty with a pdf, it is necessary to specify the "domain" of the probability model, to identify the parameters of the model, and to estimate the values of those parameters. In principle, a number of subjective decisions must be introduced at each stage in order to uniquely and completely define the pdf for a given variable (see Cooke, 1991). In practice, however, many of these "decisions" are assumed implicitly. It is, therefore, instructive to consider the range of conditions

and assumptions required. The "domain" of the probability model includes a set of conditions that describe: (i) the times for which the model is valid; and (ii) the locations or areas for which the model is valid. For example, the probability model may apply for a time frame of one week (e.g. a period of constant calibration for a measuring instrument) and for a spatial domain of one land use parcel. The second set of conditions govern the application of probabilities within this domain, and include: (i) the pattern of uncertainties in time; (ii) the pattern of uncertainties in space; (iii) the relationship between the size of the uncertainties and the size of the measured variable; and (iv) any restrictions on sizes and patterns imposed by other variables ("cross-correlation"). The patterns of uncertainty in time and space are important because the impacts of correlated error (or bias) may differ substantially from those associated with random error in environmental research. Similarly, when many different variables are used in a model, "cross-correlation" is important because extreme values in one variable may coincide with extreme values in another. For example, the cadmium, zinc, and lead concentrations in a polluted soil are strongly correlated, and, hence, the uncertainties associated with spatial interpolation of these properties are also correlated (Leenaers *et al.*, 1990). Examples of pdf*s* for uncertain environmental variables are shown in Figure 1. Figure 1(a) shows a pdf for a continuous numerical variable while Figure 1(b) shows a pdf for a discrete numerical variable. Figure 1(c) shows a joint pdf for two uncertain numerical variables, where the deviation from a circular shape denotes statistical dependence or "cross-correlation" between the variables. Figure 1(d) shows a semivariogram model, which describes the magnitude of variation as a function of distance (Goovaerts, 1997).

The conditions and parameters that specify a probability model may be determined through a data-driven approach based on "validation" or (geo)-statistical estimation or a people-driven approach based on expert elicitation. When more accurate data are readily available or observations of one variable can be used to diagnose uncertainties in another, expert opinions on the uncertainties in data and models may converge. In contrast, when a "people-driven

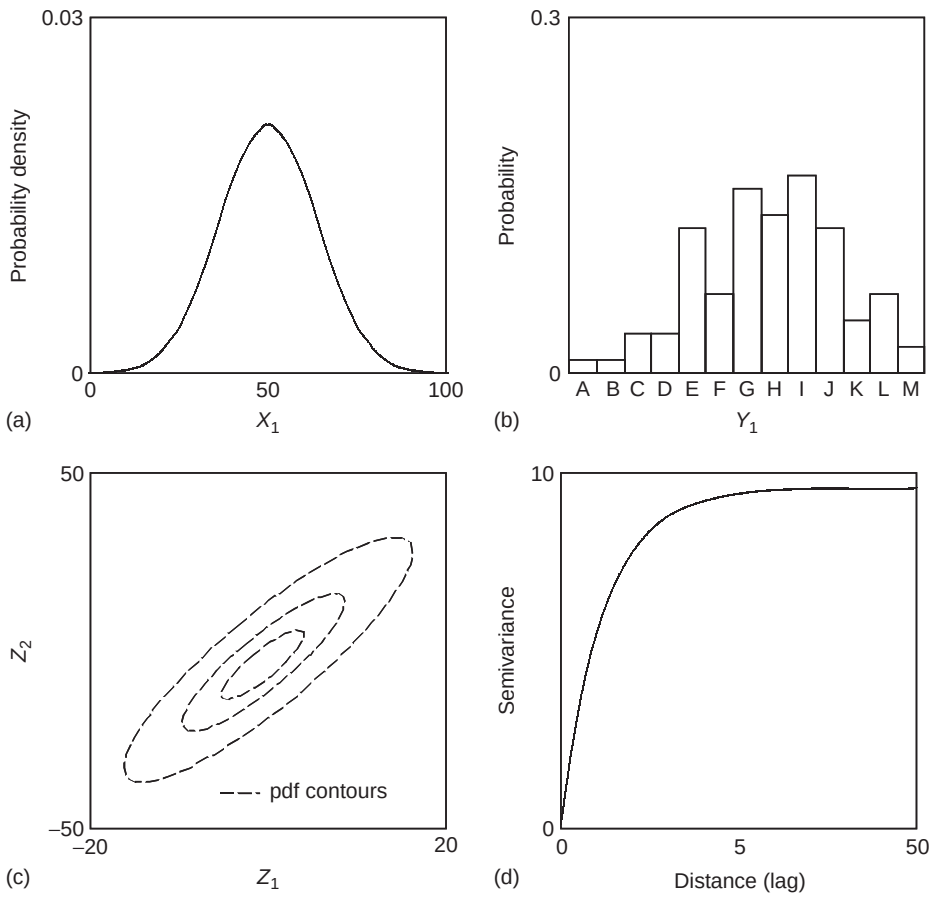


Figure 1 Example probability models for (a) a continuous numerical variable, (b) a categorical variable, and (c) two statistically dependent (cross-correlated) continuous numerical variables. Inset (d) shows the spatial patterns of uncertainty (autocorrelation) for a continuous numerical variable in the form of a ''variogram''

approach" is required, estimates of uncertainty may differ substantially between individuals, and also between groups of scientists.

UNCERTAINTIES IN MODEL INPUTS

As indicated above, the conditions and parameters that define a pdf may be estimated through a data-driven approach or through expert elicitation ("people-driven" approach). Data-driven approaches are considered first.

Data-driven Approach

Inputs to soil hydrological models include the "forcing inputs" required to generate flow (e.g. meteorological data), the boundaries in which flow occurs (positions of the air-soil and groundwater interfaces) and the characteristics of the soil within these boundaries. Many inputs are defined through measurements in the field or in the laboratory, which introduces measurement uncertainty. While soil properties vary continuously through space and time, measurements always occupy a limited number of space-time points (even remote sensing). When exhaustive inputs are required but only partial observations are available, they must be interpolated, which leads to interpolation uncertainty. Interpolation uncertainty generally increases with sample distance and soil variability, but also depends on the interpolation algorithm employed (Goovaerts, 2001). Measurement uncertainties can be estimated through comparisons with more accurate data, through laboratory testing of measurement instruments, or through repeat measurement with the same instrument (assuming constant environmental conditions). While the former provides an indication of accuracy or "bias", the latter two approaches only indicate precision.

In order to interpolate soil data, and to assess the uncertainties associated with space-time interpolation, some assumptions must be made about the behavior of soils at unmeasured times and locations. A common approach to sampling soil properties involves separating the field site into "homogeneous" units, sampling these units, and calculating a within-unit sample mean ("best estimate") and variance (uncertainty) (Voltz and Webster, 1990). This approach forms the basis for so-called "pedotransfer functions", which allow soil properties to be estimated at arbitrary locations within a "homogeneous" sample unit (Schaap and Leij, 1998; Wösten *et al.*, 2001; Minasny and McBratney, 2002). In practice, however, it may not be possible, or appropriate, to separate soil properties into "homogeneous" units, but to assume instead that soil properties vary continuously in space and time. Alternative statistical techniques, such as time-series analysis and geostatistics, are available to interpolate continuous data from partial measurements and to estimate the uncertainties associated with this interpolation (e.g. Angulo *et al.*, 1998; Goovaerts, 1997, 2001). In soil hydrology, geostatistics has been widely used to estimate interpolation uncertainty (e.g. *see* Goovaerts, 2001 for a review), both in spatial applications and in space-time interpolations (e.g. Kyriakidis and Journel, 1999; Snepvangers *et al.*, 2003). Spatial dependence between samples can be modeled with the sample semivariogram, and this information can be used for optimal prediction of soil properties at unmeasured locations through kriging (Goovaerts, 1997). The kriging variance provides an explicit measure of interpolation uncertainty.

Where soil characteristics are grouped into classes, categorical data may be used to describe soil properties for hydrological studies. In an uncertain categorical distribution, each location has a single "true" outcome (as far as the categories are identifiable and sufficient) but the precise outcome remains unknown. The probability of each outcome can be described with a discrete pdf for that location. A discrete pdf is simply a table listing each possible outcome and its associated probability. In practice, however, the distribution at one location may be sensitively dependent upon the distributions at surrounding locations, because classification errors are often statistically correlated (and, thus, dependent) in space and time. Capturing this dependence is important, but not straightforward, because multivariate discrete pdfs are characterized by a large number of parameters. In recent years, geostatistical techniques have been developed and applied for handling spatial autocorrelation in uncertain categorical data (e.g. Bierkens and Burrough, 1993; Finke *et al.*, 1999; Kyriakidis and Dungan, 2001), but identifying realistic pdfs remains inherently difficult in practice. Often, categorical data are not used directly in hydrological models, but, rather, continuous variables are related to categories of environmental variables using statistical models such as "pedotransfer" functions (see above). Here, uncertainties in the continuous variable become dependent upon (correlated with) those in the categorical variable. These uncertainties must be represented by a joint pdf when the parameters are used together in a hydrological model, otherwise the propagated uncertainties may become unrealistic.

People-driven Approach

Estimating the parameters of a probability model will always require "expert" judgment, because data must be processed and interpreted by people. However, in the absence of more accurate data, expert judgment may be the only means of estimating these parameters. So-called "expert elicitation", which formalizes the processes of estimating probability models through expert judgment (e.g. Cooke, 1991; Kaplan, 1992), has not been widely used in soil hydrology, but has been used successfully in other areas of environmental research (e.g. Morgan *et al.*, 2001). Here,

the probability models were found to be highly sensitive to the questions posed in eliciting a parameter value and to the people estimating those values (e.g. Morgan and Henrion, 1990). In principle, therefore, expert elicitation should aim to canvass a range of informed opinion about uncertainties in data, but in practice, this may not be possible, and excessive optimism or pessimism will not simply be confined to individuals. This is important because uncertainty is inherently a social construct, and will vary between people regardless of the focus of an argument (but not independently from it). For example, in a survey of soil scientists, Heuvelink and Bierkens (1992) found that respondents were overly optimistic about the predictive power of general-purpose soil maps. Nevertheless, it will often be necessary to quantify the uncertainties associated with input data for soil hydrological models through a "people-driven" approach (e.g. Keller *et al.*, 2002; Kroeze *et al.*, 2003).

UNCERTAINTIES IN MODELS

Environmental models are inherently imperfect because they abstract and simplify real patterns and processes that are themselves imperfectly known and understood. In physically based modeling, it is important to distinguish between the predictive performance of a model and its ability to explain environmental phenomena (Beven, 2001). Indeed, it is widely acknowledged that the ability of a model to predict environmental patterns satisfactorily does not mean that its explanations of these patterns are also satisfactory (e.g. Oreskes *et al.*, 1994; Rykiel, 1996; Anderson and Bates, 2001). This chapter focuses on the predictive power of models, but it is important to acknowledge that models may perform well for bad reasons, and explanatory uncertainty should, therefore, be investigated where possible (i.e., representing structural uncertainty as statistical noise is not sufficient).

Model uncertainties are "case-dependent" because hydrological models do not perform consistently for all applications involving soil water flow and solute transport. The definition of a "case" is important here, because experience is often used to "validate" models and to estimate the uncertainties associated with model parameters. In essence, a "case" refers to the circumstances in which a model performs consistently without needing to modify its underlying structure and parameter values. In accepting this case-dependence, it follows that model predictions can only be assessed through some form of comparison with direct observations or experience and, specifically, a comparison with observations or experience from the same "case" (which need not imply the same times or locations). Since observations are themselves uncertain, departures between predicted and observed outputs cannot simply be related to uncertainties in model predictions, but must also include uncertainties in empirical observations (Heuvelink and Pebesma, 1999).

Components of Model Uncertainty

Model uncertainty includes uncertainties in the structure of the model (conceptual or logical uncertainties), uncertainties in model parameters, and uncertainties in the solution of the model (Addiscott *et al.*, 1995). For example, a soil hydrological model may ignore macropore flow (structural uncertainty), may use uncertain estimates of hydraulic conductivity (parameter uncertainty), and may solve a set of continuous partial differential equations using a discrete numerical scheme (solution uncertainty). Uncertainties in model structure may originate from a perceived lack of knowledge about the real processes operating or a belief that the model intentionally abstracts and simplifies known processes. Model parameters are not inherently uncertain because they do not refer to real, measurable quantities, but are empirical quantities that allow general models to be applied to specific cases. For example, varying the friction coefficient in a hydraulic model allows the same model structure to be applied under different soil conditions, but the friction coefficient accounts for "surface roughness" at finer scales (among other things, in simple models) and cannot be measured at large spatial scales. In practice, therefore, it is difficult to define a single, optimal, set of parameters *a priori*. Moreover, nonuniqueness or "equifinality" of parameter values is common in environmental modeling because model structures only approximate reality. This leads to uncertainties in model predictions and explanations, and is particularly important where models are complex (many degrees of freedom) and observations are limited (few degrees of constraint).

If models cannot be identified uniquely, they must be represented by a probability distribution of possible models, each with a certain chance of performing well. Given uncertainties in model inputs and models, an uncertainty analysis aims to identify how these uncertainties "propagate" to model outputs (the forward problem). In practice, however, as model uncertainties are difficult to estimate *a priori*, it may be useful to compare the results of the "forward problem" to empirical observations. If these observations allow some of the original models to be rejected as improbable, the original assessment of model uncertainty should be improved (the inverse problem). When uncertainties in model inputs are known, solving the inverse problem allows model uncertainty to be identified explicitly, but only for that "case". However, model uncertainties cannot be disaggregated further into structural uncertainty and parameter uncertainty, because model parameter values do not refer to real quantities and cannot, therefore, be delimited by physical arguments or by comparisons with field observations.

Inverse Modeling

As initial estimates of model uncertainty may be no more than informed guesses, it is useful to update these estimates by comparing model predictions with empirical observations (see "Uncertainty Propagation" also). So-called "inverse modeling" has been widely used in uncertainty schemes over recent years (e.g. Abbaspour *et al.*, 1999, 2000; Schmied *et al.*, 2000; Beven and Freer, 2001; Vrugt *et al.*, 2003, and articles **Chapter 77, Inverse Modeling of Soil Hydraulic Properties, Volume 2** and **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**). Most of these schemes use Bayes' theorem, as this allows a "prior" distribution of uncertain input and parameter values to change in response to the amount and perceived (weighted) value of information available (Frenc and Smith, 1997; Bernado and Smith, 2001). The prior is sampled to provide a range of possible models for simulating the same problem, and uncertainties are propagated to model outputs by implementing the sample (running the models) and recording the results. This is known as *Monte Carlo simulation* because different possible models are sampled randomly from the prior distribution (Hammersley and Handscomb, 1979). Once the sample has been implemented, the results are compared with empirical observations and either accepted as possible or rejected as improbable (Beven and Binley, 1992). If some results are rejected, the corresponding sample can be eliminated from the prior and the resulting, calibrated distribution (posterior) used to reassess predictive uncertainty.

Two approaches are available for sampling the prior, namely: (1) a predefined probability sample from the entire distribution, which may consist of a simple random sample leading to standard Monte Carlo (e.g. the Generalised Likelihood Uncertainty Estimation (GLUE) approach of Beven and Binley, 1992), stratified random sampling, involving "Latin Hypercube" techniques (e.g. Stein, 1987), fully stratified sampling, involving the "Sectioning Method" (Addiscott and Wagenet, 1985), or (2) an interactive sample (optimization) based on a random walk approach (Markov Chain Monte Carlo or MCMC). Standard Monte Carlo and MCMC are not fundamentally different in principle. However, in practice, MCMC is more efficient because it attempts to minimize the sample required to "adequately" represent the posterior. It can, therefore, be distinguished from predefined probability sampling using arguments of acceptable risk; that is, the risk of taking too few samples from the prior to adequately determine the (posterior distribution of) model uncertainties. An important disadvantage of standard Monte Carlo, and Monte Carlo in general, is the time taken to perform an uncertainty analysis, and other types of propagation tools are available for simple models (see below). An important disadvantage of MCMC is that its predictions may be systematically biased, as the sampling algorithm may become caught in localized pockets of good performance without accommodating the full range of equifinality implied by the prior distribution (see Page *et al.*, 2003). The risk of bias can be reduced by employing a variant of standard MCMC, known as "*shuffled complex evolution*", which evolves different areas of the parameter space simultaneously (Vrugt *et al.*, 2003).

Another important application of inverse modeling in soil hydrology is for reducing "errors" in model predictions without an uncertainty analysis. Kalman filtering (e.g. Cahill *et al.*, 1999) is an example of this approach. Kalman filtering is used in dynamic modeling to update state variables and parameter values through time as new information becomes available, but has not, in general, been used to evolve different plausible parameter values (equifinality) within an uncertainty framework (although see Wendroth *et al.*, 1999).

Structural Uncertainty

To date, uncertainty analyses of environmental models have typically focused on model inputs and parameter values, as evidenced by the range of schemes available for input uncertainty propagation and for evaluating plausible parameter values (e.g. Janssen *et al.*, 1994; Bennett *et al.*, 1998; Clausnitzer *et al.*, 1998; Duke *et al.*, 1998; Haan *et al.*, 1998; Heuvelink, 1998a; Hanson, 1999; Beven and Freer, 2001; Dillah and Protopapas, 2000; Christiaens and Feyen, 2001, 2002; Vrugt *et al.*, 2003). The focus here partly reflects a consensus that model inputs and parameters are an important source of uncertainty in simulation predictions. However, it also reflects the relative ease with which uncertainty in model inputs and parameters can be quantified in comparison to assessments of structural uncertainty in models. In practice, structural uncertainty may be more important than parameter uncertainty in evaluating model performance, but such uncertainties are difficult to assess explicitly or to separate from other uncertainties during the calibration process (Beven and Binley, 1992). For example, in a study of the Uhlirska Catchment, Czech Republic, Blazkova *et al.* (2002) found that uncertainties in the water table depths predicted by TOPMODEL could be attributed to uncertainties in the topographic data or to structural uncertainties in the model, for which further interpretation was difficult. However, Mackay and Robinson (2000) successfully evaluated the internal contradictions or "semantic errors" (one form of structural problem) between hydrological submodels in predicting water table depths with TOPMODEL.

As indicated above, uncertainties in model structure may originate from a perceived lack of knowledge about the real processes operating, or a belief that the model abstracts and simplifies known processes. Both of these are relevant in modeling soil water flows and solute transport. At an elementary level, structural uncertainty need not exist

in hydraulic modeling, because Newton's laws (the general) cannot be improved, nor expressed more accurately in practice (Navier–Stokes equations). Rather, it is the simplification of these equations, their discretization in space and time, and their numerical solution that lead to uncertainty in specific cases. In this context, some types, scales or directions of flow may be deemed "less important" than others and either removed from the numerical model or their effects on the mean (important) flow incorporated through one or more empirical parameters. In practice, however, the distinction between "important" and "less important" flow requires "expert" judgment in specific cases, for which uncertainty (as well as ignorance) may be substantial. For example, while an empirical diffusion/dispersion parameter may be sufficient for describing the "mean flow" of water through a control volume (e.g. Darcy-Richards), it may not be appropriate for describing the "mean transport of solutes" within the fluid, as solute transport is sensitive to the "local" distribution of fluid velocities. Moreover, Newton's laws only resolve transfers of energy, matter, and momentum through the environment. They do not resolve the processes that lead to changes in the storage and transfer of these elementary units. For example, they cannot predict changes in rainfall inputs or in the hydraulic properties of the soil caused by vegetation growth. In practice, we are usually interested in the processes, as well as the transfers, because dominant process controls change through time and space, and, thus, Newton's laws cannot predict the future with some arbitrary degree of accuracy. In this context, structural uncertainty arises because our best representations of environmental processes are deemed insufficient.

In principle, the impacts of structural uncertainty can (and should) be evaluated by exploring different process formulations (explanatory uncertainty) or, less ideally, by adding correlated noise to model structures. In practice, however, identifying alternative process formulations or appropriate levels and patterns of noise is not straightforward.

UNCERTAINTY PROPAGATION

The combination of uncertainties in model structure, inputs, parameter values, and solution leads to a distribution of "models" for any given case, which expresses our lack of confidence about a "correct" model for that case. This leads to uncertainty in model outputs, as predictions will vary according to model inputs, structure, parameter values, and solution method (Figure 2). When uncertainties in input data and models lead to uncertainties in model output, the original uncertainties are said to have "propagated" through the modeling system. The Monte Carlo method, described above, is one (very useful) approach to this problem, but not the only one.

The problem of uncertainty propagation can be formulated generically as follows (Heuvelink, 1998a). Let y be the output of a model g that incorporates any number of

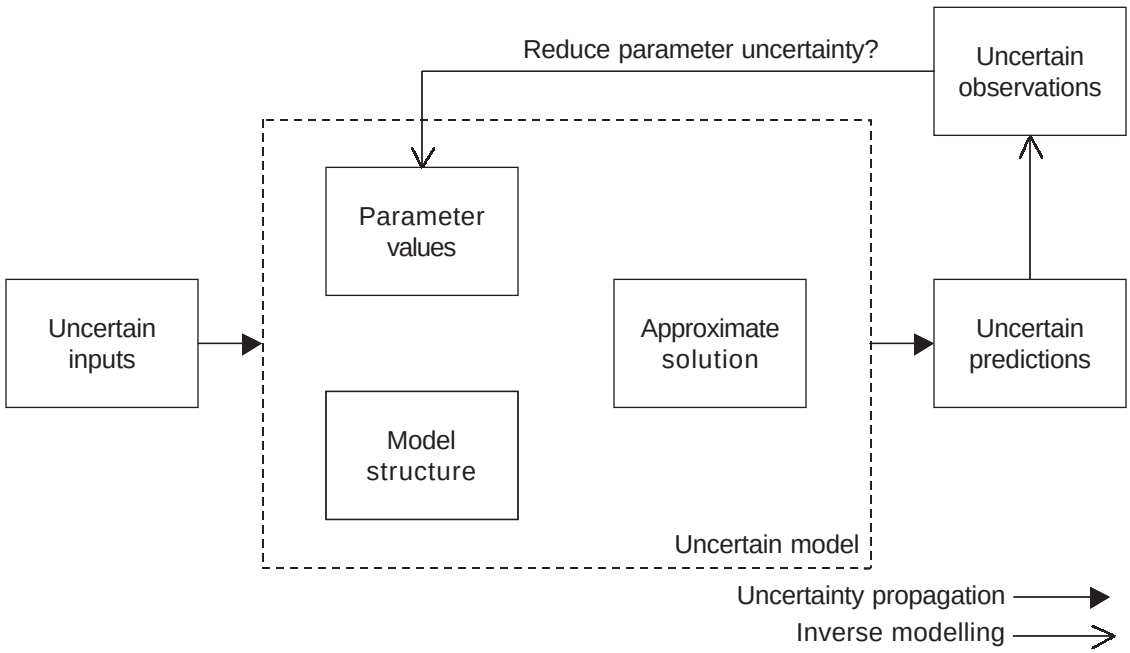


Figure 2 Uncertainties in model inputs combine with uncertainties in the model (parameters, structure, and solution) and propagate through the model leading to uncertainties in model predictions. Model parameter values may be calibrated against uncertain observations through ''inverse modeling''. Dependencies are not shown

certain elements, but m uncertain elements x_i:

$$y = g(x_1, \ldots, x_m) \quad (1)$$

The elements x_i represent input uncertainties as well as model uncertainties. Hence, they might refer to uncertainties in soil organic matter or porosity, but could also refer to uncertainties in van Genuchten parameters or in model structure. In terms of the latter, x_i could be a residual noise superimposed upon a deterministic model or a binary random variable that distinguishes between two alternative model structures. The aim here is to determine the uncertainty in the output y, given the operation g and the uncertainties in the inputs x_i. The output y will have a probability distribution, the variance of which is a measure for the propagated uncertainty in y. When g is linear and all uncertainties x_i are quantitative and quantified, the variance of y can be derived analytically. However, linear models are rare in soil hydrology, and it is, therefore, useful to consider numerical methods for solving equation (1). Two methods are considered below and, for simplicity, it is assumed that the uncertainties x_i are real numbers and are unbiased, although generalizations to the biased case are relatively straightforward, even if the quantification of bias is not.

Taylor Series Method

The Taylor Series Method (TSM), also known as *first-order analysis*, approximates g with a truncated Taylor series (Figure 3). It greatly simplifies the process of propagating uncertainty through a soil hydrological model. However, it also introduces an approximation error, which is proportional to the divergence of g from a linear form. The TSM yields analytical expressions for the mean and variance of the model output (Heuvelink, 1998a). The variance expression contains the correlations and standard deviations of the uncertain inputs and model parameters, as well as the mathematical derivatives of the model (which are assumed to exist). These derivatives reflect the sensitivity of the model to changes in each of the inputs and parameters. Rosenblueth (1975) proposed an alternative, but similar, approach to TSM, which allows the moments (e.g. mean and variance) of a function to be estimated from 2^m model outputs, evaluated at all 2^m corners of a "hypercube" representing the input space for m model inputs. The need to evaluate all 2^m outputs is a significant restriction for complex models, but can be reduced to 2^m outputs by using points on the diameters of a hypersphere, rather than the corners of the inscribed hypercube (Christian and Baecher, 2002).

The TSM has been widely used to evaluate uncertainty propagation in soil hydrological models, and particularly those involving solute transport (Kuczera, 1988; Bennett *et al.*, 1998; Tiktak *et al.*, 1999; Diaz-Diaz and Loague, 2000).

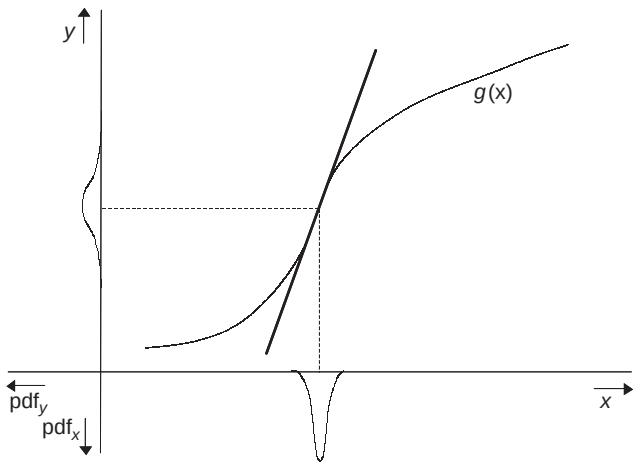


Figure 3 Uncertainty propagation with the TSM for a single input x. The curved line represents the function g, which is approximated by a linear function (dark solid line). The linear function has the same value and gradient as g when evaluated at the mean of x. After linearization, pdf$_y$ is easily computed from pdf$_x$. In this example, pdf$_y$ is much wider than pdf$_x$ because g is steep near the mean of x

Monte Carlo Method

The Monte Carlo method adopts an entirely different approach to TSM, as it retains the original model g, but randomly samples from the joint distribution of possible input and model uncertainty for g (Hammersley and Handscomb, 1979). The Monte Carlo method involves:

Repeat N times a, b:

a. Generate a set of realizations of the uncertain inputs and model parameters, structure, and solution.
b. For this set of realizations, compute and store the model output.

Compute and store sample statistics from the N outputs.

A representative sample from the joint distribution of uncertain inputs and model parameters, structure, and solution can be obtained using an appropriate pseudorandom number generator and a sufficiently large sample size (Ross, 1990; Van Niel and Laffan, 2003). The accuracy of the Monte Carlo method is inversely proportional to the square root of the number of runs N and, therefore, increases gradually with N. As such, the method is computationally expensive, but can reach an arbitrarily level of accuracy, unlike TSM or Rosenblueth's method. Of course, the time taken to perform an uncertainty analysis must influence the choice of propagation tool (Heuvelink, 2002). However, the Monte Carlo method is generic, invokes fewer assumptions, and requires less user-input than other propagation tools. Moreover, the accuracy of the method can

be fixed in advance according to the level of risk associated with a decision. The standard Monte Carlo method can be adapted to reduce the number of samples (model simulations) required for a given level of accuracy in the computed uncertainty. Examples of modified sampling schemes include Latin Hypercube sampling (e.g. Rossel *et al.*, 2001; Christiaens and Feyen, 2002; Sohrabi *et al.*, 2002; Minasny and McBratney, 2002) and MCMC (e.g. Soulsby *et al.*, 2003; Vrugt *et al.*, 2003; see also the discussion in "Uncertainties in Models"). The Monte Carlo method has been widely used to propagate uncertainties through soil hydrological models and has become increasingly popular over recent years as the computational overheads of performing such an analysis have declined. Examples in soil hydrology can be found in Petach *et al.* (1991); Bennett *et al.* (1998); Duke *et al.* (1998); Hansen *et al.* (1999); Kros *et al.* (1999); Dillah and Protopapas (2000); Thorsen *et al.* (2001); Keller *et al.* (2002); De Vries *et al.* (2003). For distributed modeling of soil water flows and solute transport, the sampling algorithm must accommodate spatial and temporal dependencies in model inputs and parameters (De Roo *et al.*, 1992; Endreny and Wood, 2001). This can be achieved through sequential simulation of distributed input or parameter values (e.g. Goovaerts, 1997).

EVALUATING THE CONTRIBUTION OF DIFFERENT UNCERTAINTIES

Prioritizing the Key Sources of Uncertainty

As indicated above, soil data and models are inherently imperfect because they abstract and simplify "real" soil properties and processes. However, it is neither possible nor desirable to eliminate all of the uncertainties associated with data and models, because resources are always limited and must be used effectively (Van Rompaey and Govers, 2002). Rather, an uncertainty analysis should aim to focus on those inputs and model entities that are likely to contribute most to the uncertainties associated with model predictions. The contribution of a specific, uncertain variable to the overall uncertainty in model predictions will depend upon the sensitivity of the model to that variable and the uncertainties associated with it. In general, models of soil water flow and solute transport will be highly sensitive to the specification of physical parameters, such as bulk density and hydraulic conductivity (e.g. Schaap and Leij, 1998; Abbaspour *et al.*, 2000; Christiaens and Feyen, 2001; Lenhart *et al.*, 2002; Vachaud and Chen, 2002b). As a calibration parameter in soil hydrological models, hydraulic conductivity may also be associated with the largest uncertainties (e.g. Rong and Wang, 2000; Schaap *et al.*, 2001; Foussereau *et al.*, 2001; Minasny and McBratney, 2002), although other parameters may be more important in specific cases (e.g. Seuntjens *et al.*, 2002; Dubus and Brown, 2002).

Evaluating the contribution of different sources of uncertainty to the overall uncertainties in model predictions is important for: (i) understanding where the greatest sources of uncertainty reside and, therefore; (ii) directing efforts towards these sources. For example, De Vries *et al.* (2003) found that denitrification in the soil column was the main source of uncertainty in predicting nitrogen inflow to groundwater in Dutch agricultural soils and suggest that improvements in field-based observations and process studies would lead to the greatest reductions in these uncertainties. Of course, too normative an emphasis on reducing uncertainty is inappropriate when the outcomes of an uncertainty analysis are highly sensitive to the assumptions made in performing that analysis (e.g. model uncertainty). Similarly, it is not appropriate to separately assess the contribution of model structure and parameter uncertainties because: (i) model parameters are empirical quantities; they are not inherently uncertain; and (ii) separate assessment is doubtful, as model uncertainties can only be assessed by comparing model predictions with independent observations. However, it should be possible to assess the contribution of different input uncertainties and the total model uncertainty to the overall uncertainties in model outputs.

In most cases, it is neither conceptually appropriate nor theoretically feasible to resolve every source of uncertainty in model predictions, but it is also important to consider practical arguments when addressing specific sources of uncertainty in models. In particular, there is a need for pragmatism when improving model inputs that do not contribute significantly to the overall uncertainties in model outputs. For example, Loague *et al.* (1989) suggest that improvements in a model of pesticide leaching should focus on soil organic carbon rather than bulk density, as model uncertainties were more sensitive to organic carbon content than bulk density. Equally, attempts to reduce uncertainty in model outputs must be balanced against the range of inputs for which uncertainties are poorly defined or remain unknown and against the practical benefits of improving model inputs or structures. For example, high quality input data are of little use when the models themselves are poor or inherently uncertain. Similarly, improvements in models should be justified against the need for additional complexity, the accuracy of new concepts, and their relevance, as well as the resources required to implement them (Van Rompaey and Govers, 2002).

Partitioning Property

Under the fairly strict assumptions that model and input uncertainties are mutually uncorrelated, only quantifiable uncertainties are involved, g is continuously differentiable with respect to all uncertain inputs, and the TSM approximation error is relatively small, the following holds

(Heuvelink, 1998a):

$$\mathrm{Var}(y) \approx \sum_{i=1}^{m} \mathrm{Var}(x_i) \cdot \left(\frac{\partial g}{\partial x_i}\right)^2 \qquad (2)$$

where the derivative of g with respect to each of the inputs x_i is evaluated about its mean. That is, the variance of y is the sum of the products of the variances from each uncertain input and the squared derivative of g with respect to that particular input. This partitioning property allows the user to determine which sources of uncertainty are the main contributors to uncertainty in model predictions. It also allows an assessment of how the output variance will decline for a given reduction in the variance from one or more inputs. Clearly, the output variance will mainly improve from a reduction of the input variance that contributes most to the uncertainties in model output (other factors being equal). However, this may not correspond to the input with the largest variance, because the hydrological model will display different sensitivities to each uncertain input. Similarly, the contributions of the input uncertainties, as well as the magnitude of the output uncertainty, will depend upon the specific "case" considered and the types of output predicted. While it is instructive to consider equation (2), its assumptions of near-linearity in g and zero correlation between input uncertainties are rarely met in practice. Rather, for most practical applications in soil hydrology, where these assumptions are deemed unacceptable, more advanced analysis techniques, often based on Monte Carlo simulation, may be employed (*see* Jansen *et al.*, 1994; Jansen, 1999; Chan *et al.*, 2000).

Examples from soil hydrology where the contributions of different sources of uncertainty were compared include Zhang *et al.* (1993); Finke *et al.* (1996); Piggott and Cawlfield (1996); Tiktak *et al.* (1999); Barlund and Tattari (2001); Dubus and Brown (2002); Keller *et al.* (2002); Sohrabi *et al.* (2002); De Vries *et al.* (2003).

SCALE AND UNCERTAINTY

Issues of scale have received increasing attention in soil hydrology over recent years (e.g. Blöschl and Sivapalan, 1995; De Vries *et al.*, 1998; Bierkens *et al.*, 2000). This partly reflects a consensus that hydrological patterns and processes are strongly scale-dependent (Blöschl and Sivapalan, 1995) and that the implications of scale, and changes between scales, have not been fully acknowledged in hydrological modeling (Beven, 1995, 2001). It also reflects an increasing demand for "policy-relevant" research in hydrology (Beven, 2000; Clifford, 2002), where there is a need to integrate social and physical perspectives on environmental problems that traverse a range of political and geographic scales (e.g. Dumanski *et al.*, 1998; Kros *et al.*, 1999; Kros *et al.*, 2002).

The need to operate at a range of scales, or to change between scales, introduces uncertainty because the dominant patterns and process controls may not be known at all scales, or incorporated practically in models at the scales of interest (e.g. Heuvelink and Pebesma, 1999; Vachaud and Chen, 2002b; Hennings, 2002). For example, while macropore and preferential flow are dominant at the "pedon scale", they remain important contributory processes to flow patterns at "field", "catchment", and "regional" scales (Heuvelink, 1998b; Heuvelink and Pebesma, 1999; Zehe *et al.*, 2001), but are difficult to parameterize from general-purpose soil survey data and, hence, to incorporate in large-scale models of soil water flow (Simmonds and Nortcliff, 1998).

Scale Dependence of Model Inputs

When model inputs, or the observations used to test model outputs, are defined with a different control volume or "support" from that required by the model, these data must be aggregated or disaggregated, or the model must be redefined at an appropriate scale (Heuvelink and Pebesma, 1999; Bierkens *et al.*, 2000). The basic elements of "support" include the domain chosen to represent a real entity, together with the size (resolution), shape, and orientation of the "building blocks" or space-time units used to discretize that entity (Webster and Oliver, 1990). An important consequence of representing model inputs as stochastic quantities is that changes in support will affect uncertain quantities (e.g. a variance) more than deterministic ones (e.g. an average). Thus, aggregating or disaggregating data may greatly affect the probability distribution for those data, and, particularly, its width, without significantly affecting the mean value (Heuvelink and Pebesma, 1999; Heuvelink, 2002). For example, in a study of soil acidification with the SMART2 model, Kros *et al.* (2002) found that predicted concentrations of aluminum and nitrate in the soil solution were highly dependent upon the support size of model entities. In practice, space-time aggregation should lead to a reduction in uncertainty and to an increase in the spatial autocorrelation of model inputs because much of the variability at finer scales is lost and, thus, disappears as a source of uncertainty (e.g. Heuvelink and Pebesma, 1999). By contrast, space-time disaggregation will lead to an increase in uncertainty and to a reduction in the spatial autocorrelation of model inputs because the attribute variability is increased at finer scales.

Scale Dependence of Models

Uncertainties in model structure (dominant process controls) and parameter values are also sensitive to changes in scale. However, unlike aggregation or disaggregation of data, model parameters cannot be transformed using

the original quantities alone because they do not, in general, refer to real, measurable things. Rather, upscaling or downscaling of model parameters can only be achieved by (re)-calibrating the model at another scale, and may require adaptation of the original model because functional relationships are typically nonlinear and process controls usually change with scale (Addiscott and Tuck, 2001). More fundamentally, upscaling or downscaling of model parameters may hide underlying problems with model structure if process controls are not simultaneously evaluated for their relevance and sufficiency at other scales (i.e., uncertainty in explanation).

CONCLUSIONS

A mismatch between the complexity of soil patterns and processes and our ability to capture them adequately for some practical purpose leads to uncertainty in the predictions of soil hydrological models. These uncertainties originate from a lack of confidence in model inputs, which include measurement and interpolation errors, and in models that include conceptual, logical, and computational errors. Uncertainties in model inputs and models can be described with probability distribution functions, for which a number of conditions and parameters must be estimated *a priori*. In principal, this can be achieved with a "data-driven" approach for model inputs. Where observations are lacking, a "people-driven" approach is required, but the resulting estimates may introduce bias into the uncertainty analysis. For estimating uncertainties in model parameters, a data-driven approach cannot be used because model parameters do not, in general, refer to real, measurable quantities. Here, a "people-driven" approach is useful for making an initial assessment of parameter uncertainties, which may be updated later through inverse modeling. In practice, uncertainties in model concepts (structural or explanatory uncertainties) are difficult to estimate *a priori* or to isolate through inverse modeling, but may ultimately determine the utility of model predictions.

Uncertainties in model inputs and models combine and propagate through a hydrological model, leading to uncertainties in model outputs, which can be quantified using a range of statistical techniques. In soil hydrology, the TSM and Monte Carlo Simulation (MCS) have been widely used for uncertainty propagation. In recent years, MCS has largely replaced the TSM, as it is generic, invokes fewer assumptions, and requires less user-input than other propagation tools. Moreover, continued improvements in computing hardware and sampling techniques (e.g. MCMC) have allowed MCS to be applied to increasingly complex models, including spatially and temporally distributed models of soil water flow and solute transport. Nevertheless, uncertainty analyses remain complicated when uncertain inputs are "autocorrelated" in space or time or "cross-correlated" between variables. More fundamentally, autocorrelation and cross-correlation will introduce more degrees of freedom than can be constrained uniquely through comparisons of model predictions and independent observations (i.e., inverse modeling), leading to the characteristic equifinality of hydrological models, but also implying sensitivity to the specific range of observations available. Indeed, while it has long been recognized that mathematical models should not be more complicated than specific applications require, recent research in soil hydrology has consistently reiterated the need to balance model complexity against the "testability" of model predictions. It has also argued for improvements in field observations as a key source of information for reducing uncertainties in soil hydrological models. In this context, it is helpful to identify those variables that contribute most to uncertainties in model outputs, because resources are always limited and must be used effectively.

In evaluating the contribution of different sources of uncertainty to the overall uncertainties in model predictions, it is instructive to consider the partitioning property for a multivariate distribution. In practice, however, more advanced analysis techniques based on Monte Carlo simulation are necessary for investigating the sources of uncertainties in complex hydrological models. While analysis of variance (ANOVA) techniques should assist in targeting resources towards those inputs and parameters that contribute most to uncertainties in model outputs, they are ultimately constrained by the interdependence of model inputs, structure, and parameters following calibration and the lack of physical reasoning attached to parameter uncertainties and, hence, the difficulties in separating model parameter and structural uncertainties. When evaluating predictive uncertainties, it is, therefore, important to distinguish between the predictive performance of a model and its ability to explain environmental phenomena. Indeed, estimates of uncertainty in model predictions will be unreliable if the explanations upon which they are based are inappropriate. In principal, this implies a detailed analysis of the "contingencies of place" associated with a particular modeling study and careful inspection of any modeling assumptions introduced within this context. In practice, however, this assumes a depth and breadth of expertise, including a detailed knowledge of mathematical modeling and field techniques, that is rarely, if ever, available (Blöschl, 2001). Thus, while decision-makers, and those affected by decisions, would benefit most from improved access to uncertainty tools, they currently benefit least because of the inherent limitations of existing tools and the practical disadvantages of performing such analyses. These issues must be addressed if an important aim of developing uncertainty tools is to encourage more widespread criticism of data and models in soil hydrology.

Given that uncertainty analyses are not benign instruments, with the capacity to both encourage unreasonable decisions and impede reasonable ones, there is a need to balance the complexity of an uncertainty analysis against the expertise of the user and the risks associated with bad decisions. While MCS dramatically simplifies the problem of propagating uncertainties through soil hydrological models, there are many problems of equal or greater complexity that remain. These include assessments of uncertainty in space-time categorical data, analyses of statistical dependence within and between uncertain inputs and parameters (autocorrelation and cross-correlation), assessments of structural uncertainty in models, and analyses of the scale-dependence of model input and model uncertainties. In evaluating modeling uncertainties, the importance of accounting for different modeling scales and changes between scales can hardly be over-emphasized because the consequence of using wrong (combinations of) support is to invalidate the uncertainty analysis. In soil hydrology, and in environmental science more generally, the need for interdisciplinary, "policy-relevant" research will only increase this problem in future. Here, the benefits of performing an uncertainty analysis are clear (e.g. Beven, 2000), but the challenges of balancing physical and statistical realism against the need for pragmatism in extending their application are considerable.

REFERENCES

Abbaspour K.C., Kasteel R. and Schulin R. (2000) Inverse parameter estimation in a layered unsaturated field soil. *Soil Science*, **165**, 109–123.

Abbaspour K.C., Sonnleitner M.A. and Schulin R. (1999) Uncertainty in estimation of soil hydraulic parameters by inverse modeling: example lysimeter experiments. *Soil Science Society of America Journal*, **63**, 501–509.

Addiscott T.M. and Wagenet R.J. (1985) A simple method for combining soil properties that show variability. *Soil Science Society of America Journal*, **49**, 1365–1369.

Addiscott T.M., Smith J. and Bradbury N. (1995) Critical evaluation of models and their parameters. *Journal of Environmental Quality*, **24**, 803–807.

Addiscott T.M. and Tuck G. (2001) Non-linearity and error in modelling soil processes. *European Journal of Soil Science*, **52**, 129–138.

Anderson M.G. and Bates P.D. (2001) *Model Validation: Perspectives in Hydrological Science*, John Wiley & Sons: Chichester.

Angulo J.M., Gonzalez-Manteiga W., Febrero-Bande M. and Alonso F.J. (1998) Semi-parametric statistical approaches for space-time process prediction. *Environmental and Ecological Statistics*, **5**, 297–316.

Barlund I. and Tattari S. (2001) Ranking of parameters on the basis of their contribution to model uncertainty. *Ecological Modelling*, **142**, 11–23.

Bennett D.H., James A.L., McKone T.E. and Oldenburg C.M. (1998) On uncertainty in remediation analysis: variance propagation from subsurface transport to exposure modeling. *Reliability Engineering and System Safety*, **62**, 117–129.

Bernado J.M. and Smith A.F.M. (2001) *Bayesian Theory*, John Wiley & Sons: Chichester.

Beven K. (1995) Linking parameters across scales: subgrid parameterizations and scale-dependent hydrological models. *Hydrological Processes*, **9**, 251–290.

Beven K. (2000) On model uncertainty, risk and decision making. *Hydrological Processes*, **14**, 2605–2606.

Beven K. (2001) On explanatory depth and predictive power. *Hydrological Processes*, **15**, 3069–3072.

Beven K. and Binley A. (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K. and Freer J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, **249**, 11–29.

Bierkens M.F.P. and Burrough P.A. (1993) The indicator approach to categorical soil data. I. Theory. *Journal of Soil Science*, **44**, 361–368.

Bierkens M.F.P., Finke P.A. and De Willigen P. (2000) *Upscaling and Downscaling Methods for Environmental Research*, Kluwer: Dordrecht.

Blazkova S., Beven K., Tacheci P. and Kulasova A. (2002) Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): The death of TOPMODEL? *Water Resources Research*, **38**, 1257.

Blöschl G. (2001) Scaling in hydrology. *Hydrological Processes*, **15**(4), 709–711.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modeling - a review. *Hydrological Processes*, **9**, 251–290.

Brazier R.E., Beven K., Anthony S.G. and Rowan J.S. (2001) Implications of model uncertainty for the mapping of hillslope-scale soil erosion predictions. *Earth Surface Processes and Landforms*, **26**, 1333–1352.

Cahill A.T., Ungaro F., Parlange M.B., Mata M. and Nielsen D.R. (1999) Combined spatial and Kalman filter estimation of optimal soil hydraulic properties. *Water Resources Research*, **35**, 1079–1088.

Chan K., Saltelli A. and Tarantola S. (2000) Winding stairs: a sampling tool to compute sensitivity indices. *Statistics and Computing*, **10**, 187–196.

Christensen S. and Cooley R.L. (1999) Evaluation of prediction intervals for expressing uncertainties in groundwater flow model predictions. *Water Resources Research*, **35**, 2627–2639.

Christiaens K. and Feyen J. (2001) Analysis of uncertainties associated with different methods to determine soil hydraulic properties and their propagation in the distributed hydrological MIKE SHE model. *Journal of Hydrology*, **246**, 63–81.

Christiaens K. and Feyen J. (2002) Constraining soil hydraulic parameter and output uncertainty of the distributed hydrological MIKE SHE model using the GLUE framework. *Hydrological Processes*, **16**, 373–391.

Christian J.T. and Baecher G.B. (2002) The point-estimate method with large numbers of variables. *International Journal*

for Numerical and Analytical Methods in Geomechanics, **26**, 1515–1529.

Clausnitzer V., Hopmans J.W. and Starr J.L. (1998) Parameter uncertainty analysis of common infiltration models. *Soil Science Society of America Journal*, **62**, 1477–1487.

Clifford N.J. (2002) Hydrology: the changing paradigm. *Progress in Physical Geography*, **26**, 290–301.

Cooke R.M. (1991) *Experts in Uncertainty. Opinion and Subjective Probability in Science*, Oxford University Press.

De Roo A.P.J., Hazelhoff L. and Heuvelink G.B.M. (1992) Estimating the effects of spatial variability of infiltration on the output of a distributed runoff and soil erosion model using Monte Carlo methods. *Hydrological Processes*, **6**, 127–143.

De Vries W., Kros J., Groenenberg J.E., Reinds G.J. and Van Der Salm C. (1998) The use of upscaling procedures in the application of soil acidification models at different spatial scales. *Nutrient Cycling in Agroecosystems*, **50**, 225–238.

De Vries W., Kros J., Oenema O. and De Klein J. (2003) Uncertainties in the fate of nitrogen II: a quantitative assessment of the uncertainties in major nitrogen fluxes in the Netherlands. *Nutrient Cycling in Agroecosystems*, **66**, 71–102.

Diaz-Diaz R. and Loague K. (2000) Regional-scale leaching assessments for Tenerife: effect of data uncertainties. *Journal of Environmental Quality*, **29**, 835–847.

Dillah D.D. and Protopapas A.L. (2000) Uncertainty propagation in layered unsaturated soils. *Transport in Porous Media*, **38**, 273–290.

Dubus I.G. and Brown C.D. (2002) Sensitivity and first-step uncertainty analyses for the preferential flow model MACRO. *Journal of Environmental Quality*, **31**, 227–240.

Duke L.D., Rong Y. and Harmon T.C. (1998) Parameter-induced uncertainty in modeling vadose zone transport of VOCs. *Journal of Environmental Engineering-ASCE*, **124**, 441–448.

Dumanski J., Pettapiece W.W. and McGregor R.J. (1998) Relevance of scale dependent approaches for integrating biophysical and socio-economic information and development of agroecological indicators. *Nutrient Cycling in Agroecosystems*, **50**, 13–22.

Endreny T.A. and Wood E.F. (2001) Representing elevation uncertainty in runoff modelling and flowpath mapping. *Hydrological Processes*, **15**, 2223–2236.

Finke P.A., Wladis D., Kros J., Pebesma E.J. and Reinds G.J. (1999) Quantification and simulation of errors in categorical data for uncertainty analysis of soil acidification modelling. *Geoderma*, **93**, 177–194.

Finke P.A., Wösten J.H.M. and Jansen M.J.W. (1996) Effects of uncertainty in major input variables on simulated functional soil behaviour. *Hydrological Processes*, **10**, 661–669.

Foussereau X., Graham W.D., Akpoji G.A., Destouni G. and Rao P.S.C. (2001) Solute transport through a heterogeneous coupled vadose-saturated zone system with temporally random rainfall. *Water Resources Research*, **37**, 1577–1588.

Frenc S. and Smith J.Q. (1997) *The Practice of Bayesian Analysis*, Arnold: 152–171.

Goovaerts P. (1997) *Geostatistics for Natural Resources Evaluation*, Oxford University Press.

Goovaerts P. (2001) Geostatistical modelling of uncertainty in soil science. *Geoderma*, **103**, 3–26.

Haan C.T., Storm D.E., Al-Issa T., Prabhu S., Sabbagh G.J. and Edwards D.R. (1998) Effect of parameter distributions on uncertainty analysis of hydrologic models. *Transactions of the ASAE*, **41**, 65–70.

Hammersley J.M. and Handscomb D.C. (1979) *Monte Carlo Methods*, Chapman & Hall: London.

Hansen S., Thorsen M., Pebesma E.J., Kleeschulte S. and Svendsen H. (1999) Uncertainty in simulated nitrate leaching due to uncertainty in input data. A case study. *Soil Use and Management*, **15**, 167–175.

Hanson K.M. (1999) A framework for assessing uncertainties in simulation predictions. *Physica D*, **133**, 179–188.

Hennings V. (2002) Accuracy of coarse-scale land quality maps as a function of the upscaling procedure used for soil data. *Geoderma*, **107**, 177–196.

Heuvelink G.B.M. (1998a) *Error Propagation in Environmental Modelling with GIS*, Taylor & Francis: London.

Heuvelink G.B.M. (1998b) Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient Cycling in Agroecosystems*, **50**, 255–264.

Heuvelink G.B.M. (2002) Analysing uncertainty propagation in GIS: why is it not that simple? In *Uncertainty in Remote Sensing and GIS*, FoodyG.M. and Atkinson P.M. (Eds.), Wiley: Chichester, pp. 155–165.

Heuvelink G.B.M. and Bierkens M.F.P. (1992) Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma*, **55**, 1–15.

Heuvelink G.B.M. and Pebesma E.J. (1999) Spatial aggregation and soil process modelling. *Geoderma*, **89**, 47–65.

Jansen M.J.W. (1999) Analysis of variance designs for model output. *Computer Physics Communications*, **117**, 35–43.

Jansen M.J.W., Rossing W.A.H. and Daamen R.A. (1994) Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. In *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, Grasman J. and Van Straten G. (Eds.), Kluwer: Dordrecht, pp. 334–343.

Janssen P.H.M., Heuberger P.S.C. and Klepper O. (1994) UNCSAM: a tool for automating sensitivity and uncertainty analysis. *Environmental Software*, **9**, 1–11.

Kaplan S. (1992) 'Expert information' versus 'expert opinions': another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Reliability Engineering and System Safety*, **35**, 61–72.

Keller A., Abbaspour K.C. and Schulin R. (2002) Assessment of uncertainty and risk in modeling regional heavy-metal accumulation in agricultural soils. *Journal of Environmental Quality*, **31**, 175–187.

Kroeze C., Aerts R., Van Breemen N., Van Dam D., Van Der Hoek K., Hofschreuder P., Hoosbeek M., De Klein J., Kros J. and Van Oene H. and (2003) Uncertainties in the fate of nitrogen I: an overview of sources of uncertainty illustrated with a Dutch case study. *Nutrient Cycling in Agroecosystems*, **66**, 43–69.

Kros J., Mol-Dijkstra J.P. and Pebesma E.J. (2002) Assessment of the prediction error in a large-scale application of a dynamic soil acidification model. *Stochastic Environmental Research and Risk Assessment*, **16**, 279–306.

Kros J., Pebesma E.J., Reinds G.J. and Finke P.A. (1999) Uncertainty assessment in modelling soil acidification at the European scale: a case study. *Journal of Environmental Quality*, **28**, 366–377.

Kuczera G. (1988) On the validity of first-order prediction limits for conceptual hydrologic models. *Journal of Hydrology*, **103**, 229–247.

Kyriakidis P.C. and Dungan J.L. (2001) A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. *Environmental and Ecological Statistics*, **8**, 311–330.

Kyriakidis P. and Journel A.G. (1999) Geostatistical space-time models: a review. *Mathematical Geology*, **31**, 651–684.

Leenaers H., Okx J.P. and Burrough P.A. (1990) Comparison of spatial prediction methods for mapping floodplain soil pollution. *Catena*, **17**, 535–550.

Leenhardt D. (1995) Errors in the estimation of soil water properties and their propagation through a hydrological model. *Soil Use and Management*, **11**, 15–21.

Lenhart T., Eckhardt K., Fohrer N. and Frede H.G. (2002) Comparison of two different approaches of sensitivity analysis. *Physics and Chemistry of the Earth*, **27**, 645–654.

Loague K., Yost R.S., Green R.E. and Liang T.C. (1989) Uncertainty in pesticide leaching assessment in Hawaii. *Journal of Contaminant Hydrology*, **4**, 139–161.

Mackay D.S. and Robinson V.B. (2000) A multiple criteria decision support system for testing integrated environmental models. *Fuzzy Sets and Systems*, **113**, 53–67.

McKone T.E. (1996) Alternative modeling approaches for contaminant fate in soils: uncertainty, variability, and reliability. *Reliability Engineering and System Safety*, **54**, 165–181.

Minasny B. and McBratney A.B. (2002) Uncertainty analysis for pedotransfer functions. *European Journal of Soil Science*, **53**, 417–429.

Morgan M.G. and Henrion M. (1990) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York.

Morgan M.G., Pitelka L.F. and Shevliakova E. (2001) Elicitation of expert judgments of climate change impacts on forest ecosystems. *Climatic Change*, **49**, 279–307.

Oreskes N., Shraderfrechette K. and Belitz K. (1994) Verification, validation, and confirmation of numerical-models in the earth-sciences. *Science*, **263**, 641–646.

Page T., Beven K.J., Freer J. and Jenkins A. (2003) Investigating the uncertainty in predicting responses to atmospheric deposition using the model of acidification of groundwater in catchments (MAGIC) within a generalised likelihood uncertainty estimation (GLUE) framework. *Water Air and Soil Pollution*, **142**, 71–94.

Petach M.C., Wagenet R.J. and DeGloria S.D. (1991) Regional water flow and pesticide leaching using simulations with spatially distributed data. *Geoderma*, **48**, 245–269.

Piggott J.H. and Cawlfield J.D. (1996) Probabilistic sensitivity analysis for one-dimensional contaminant transport in the vadose zone. *Journal of Contaminant Hydrology*, **24**, 97–115.

Rong Y. and Wang R.F. (2000) Monte Carlo vadose tone model for soil remedial criteria. *Soil and Sediment Contamination*, **9**, 593–610.

Rosenblueth E. (1975) Point estimates for probability moments. *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 3812–3814.

Ross S.M. (1990) *A Course in Simulation*, MacMillan: New York.

Rossel R.A.V., Goovaerts P. and McBratney A.B. (2001) Assessment of the production and economic risks of site-specific liming using geostatistical uncertainty modelling. *Environmetrics*, **12**, 699–711.

Rykiel E.J. (1996) Testing ecological models: the meaning of validation. *Ecological Modelling*, **90**, 229–244.

Schaap M.G. and Leij F.J. (1998) Database-related accuracy and uncertainty of pedotransfer functions. *Soil Science*, **163**, 765–779.

Schaap M.G., Leij F.J. and Van Genuchten M.T. (2001) ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, **251**, 163–176.

Schmied B., Abbaspour K. and Schulin R. (2000) Inverse estimation of parameters in a nitrogen model using field data. *Soil Science Society of America Journal*, **64**, 533–542.

Seuntjens P., Mallants D., Simunek J., Patyn J. and Jacques D. (2002) Sensitivity analysis of physical and chemical properties affecting field-scale cadmium transport in a heterogeneous soil profile. *Journal of Hydrology*, **264**, 185–200.

Simmonds L.P. and Nortcliff S. (1998) Small scale variability in the flow of water and solutes, and implications for lysimeter studies of solute leaching. *Nutrient Cycling in Agroecosystems*, **50**, 65–75.

Smithson M. (1989) *Ignorance and Uncertainty: Emerging Paradigms*, Springer-Verlag: New York.

Snepvangers J.J.J.C, Heuvelink G.B.M. and Huisman J.A. (2003) Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma*, **112**, 253–271.

Sohrabi T.M., Shirmohammadi A. and Montas H. (2002) Uncertainty in nonpoint source pollution models and associated risks. *Environmental Forensics*, **3**, 179–189.

Soulsby C., Petry J., Brewer M.J., Dunn S.M., Ott B. and Malcolm I.A. (2003) Identifying and assessing uncertainty in hydrological pathways: a novel approach to end member mixing in a Scottish agricultural catchment. *Journal of Hydrology*, **274**, 109–128.

Stein M.L. (1987) Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, **29**, 143–151.

Thorsen M., Refsgaard J.C., Hansen S., Pebesma E.J., Jensen J.B. and Kleeschulte S. (2001) Assessment of uncertainty in simulation of nitrate leaching to aquifers at catchment scale. *Journal of Hydrology*, **242**, 210–227.

Tiktak A., Leijnse A. and Vissenberg H. (1999) Uncertainty in a regional-scale assessment of cadmium accumulation in the Netherlands. *Journal of Environmental Quality*, **28**, 461–470.

Vachaud G. and Chen T. (2002a) Sensitivity of computed values of water balance and nitrate leaching to within soil class variability of transport parameters. *Journal of Hydrology*, **264**, 87–100.

Vachaud G. and Chen T. (2002b) Sensitivity of a large-scale hydrologic model to quality of input data obtained at different

scales; distributed versus stochastic non-distributed modelling. *Journal of Hydrology*, **264**, 101–112.

Van Niel K. and Laffan S.W. (2003) Gambling with randomness: the use of pseudo-random number generators in GIS. *International Journal of GIS*, **17**, 49–68.

Van Rompaey A.J.J. and Govers G. (2002) Data quality and model complexity for regional scale soil erosion prediction. *International Journal of GIS*, **16**, 663–680.

Voltz M. and Webster R. (1990) A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Science*, **41**, 473–490.

Vrugt J.A., Bouten W., Gupta H.V. and Hopmans J.W. (2003) Toward improved identifiability of soil hydraulic parameters: on the selection of a suitable parametric model. *Vadose Zone Journal*, **2**, 98–113.

Wang T.A. and McTernan W.F. (2002) The development and application of a multilevel decision analysis model for the remediation of contaminated groundwater under uncertainty. *Journal of Environmental Management*, **64**, 221–235.

Webster R. (2000) Is soil variation random? *Geoderma*, **97**, 149–163.

Webster R. and Oliver M.A. (1990) *Statistical Methods in Soil and Land Resource Survey*, University Press: Oxford.

Wendroth O., Rogasik H., Koszinski S., Ritsema C.J., Dekker L.W. and Nielsen D.R. (1999) State-space prediction of field-scale soil water content time series in a sandy loam. *Soil and Tillage Research*, **50**, 85–93.

Wösten J.H.M., Pachepsky Y.A. and Rawls W.J. (2001) Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic properties. *Journal of Hydrology*, **251**, 123–150.

Zehe E., Maurer T., Ihringer J. and Plate E. (2001) Modeling water flow and mass transport in a loess catchment. *Physics and Chemistry of the Earth Part B-Hydrology Oceans and Atmosphere*, **26**, 487–507.

Zhang H., Haan C.T. and Nofziger D.L. (1993) An approach to estimating uncertainties in modeling transport of solutes through soils. *Journal of Contaminant Hydrology*, **12**, 35–50.

PART 7

Erosion and Sedimentation

80: Erosion and Sediment Transport by Water on Hillslopes

ANTHONY JOHN PARSONS

Department of Geography, University of Leicester, Leicester, UK

Much of the terminology and many of the concepts within the field of erosion and sediment transport by water on hillslopes, derive from the literature of agricultural engineering. Of particular importance has been the distinction between rills and interrill areas. One definition of the former is that they are channels small enough to be removed by ploughing; gullies, by contrast, are not. The distinction of gullies, rills, and interrill areas is artificial. None the less, it provides a convenient framework for an examination of the processes of hillslope erosion and sediment transport.

In interrill areas, the dominant mechanism of sediment detachment is that of raindrop impact. Sediment detached by raindrops may be split into that which is transported away from the location of detachment in splash droplets (rainsplash), and that which is simply dislodged by the impact of raindrops but which either remains at, or falls back to, the site of detachment. The former is relatively easy to measure; the latter is not, but may be quantitatively much more important. The rate of detachment is a function of the rainfall energy at the soil surface, so that where vegetation intercepts some of the energy of the rainfall, or a layer of surface water exists, some of the energy of the falling rain will be dissipated. The rate of detachment is also affected by surface gradient. Whereas detachment in interrill areas is due to the energy of falling raindrops, sediment transport is mainly controlled by flow energy. Several authors have attempted to apply transport-capacity equations developed for alluvial rivers even though the hydraulic conditions in shallow overland flow are very different from those in rivers. However, sediment transport by interrill flow needs to consider not only the capacity of the flow, but also its competence.

As threads of interrill flow become deeper and faster, a threshold is reached beyond which significant flow detachment begins to take place. Once this occurs the flow begins to erode definable channels, or rills. As with interrill overland flow, the transport capacity of rill flow has typically been estimated using equations taken from the literature developed for alluvial rivers. However, sediment transport by rill flow is equally determined by the competence of the flow. This is particularly the case for stony soils.

Gullies have been relatively neglected in the agricultural literature, so that, whereas there is considerable qualitative literature on gully growth and development, quantitative information is limited. However, gullies may account for between 10% and 94% of total soil loss on hillslopes.

Modelling of hillslope erosion and sediment transport has been undertaken by both geomorphologists and agricultural engineers. There is a shared history, inasmuch as, through time, both demonstrate increasingly explicit representation of processes in their modelling as understanding of processes has increased and increased computing power has become available.

On hillslopes, some of the highest rates of erosion and sediment transport are attained on bare, agricultural fields. Certainly, in such areas rates of erosion are of significant concern. Hence, it is not surprising that much of the research into rates and controls of hillslope erosion and sediment transport has been undertaken by agricultural engineers, and that much of the terminology and many of the concepts within this subject area derive from

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the literature of agricultural engineering. Of particular importance into research on erosion and sediment transport on agricultural land has been the distinction between rills and interrill areas. One definition of the former is that they are channels small enough to be removed by ploughing (Soil Conservation Society of America, 1982). Gullies, by contrast, are not. Consequently, the existence of gullies precludes mechanized agriculture, a result of which is that they have been studied to a lesser degree than their significance for hillslope erosion may warrant. The distinction of gullies, rills, and interrill areas is artificial, both in agricultural fields and on hillslopes covered by natural or seminatural vegetation, because the former evolve from the latter. None the less, it provides a convenient framework for an examination of the processes of hillslope erosion and sediment transport.

INTERRILL EROSION AND SEDIMENT TRANSPORT

In interrill areas, the dominant mechanism of sediment detachment is that of raindrop impact. The kinetic energy of falling rain ranges from 1000 to 100 000 times that of shallow interrill flow (Ekern, 1953). Ghadiri and Payne (1981) calculated the maximum impact stress caused by a falling raindrop to be 2–6 MPa. However, Nearing *et al.* (1987) argued that the nonrigid, nonhomogeneous nature of soil surfaces reduced the impact pressure. These authors obtained maximum impact stresses between 190 and 290 kPa from a 5.6-mm diameter raindrop. Experiments by Young and Wiersma (1973) showed that when soils were protected from 89% of the energy of rainfall without reduction in the application of water rate, soil loss decreased by between 90 and 94% for the three soils they tested. Sediment detached by raindrops may be split into that which is transported away from the location of detachment in splash droplets (rainsplash), and that which is simply dislodged by the impact of raindrops, but which either remains at, or falls back to, the site of detachment. The former is relatively easy to measure, and several methods have been devised for its measurement (e.g. Ellison, 1944; Morgan, 1978; Parsons *et al.*, 1991. See also, Poesen and Torri, 1988; Torri and Poesen, 1988). Rates of rainsplash have been shown to be directly proportional to rainfall energy (Watung *et al.*, 1996) and intensity (Park *et al.*, 1983) and inversely proportional to soil shear strength (Cruse and Larson, 1977), such that general equations for rainsplash have the form

$$R = aE^b \tag{1}$$

and

$$R = cI^d \tag{2}$$

where R is rainsplash rate (Kg m^{-2} s^{-1}), a and c are soil erodibility coefficients, E is rainfall energy (J m^{-2} s^{-1}), I is rainfall intensity (mm s^{-1}), and b and d are exponents, the latter generally assumed to equal two (Sharma *et al.*, 1993). Equations (1) and (2) have been linked through relationships between rainfall energy and intensity, such as that given by Brandt (1990), where

$$E = 8.95 + 8.44 \log I \tag{3}$$

However, the assumption that there is a simple relationship between rainfall energy and intensity is questionable (see Parsons and Gadian, 2000; Salles *et al.*, 2002; van Dijk *et al.*, 2002). Consequently, rainfall kinetic energy is better calculated as the sum of the kinetic energy of individual raindrops (Sharma *et al.*, 1995), such that

$$E = k \sum_{i=1}^{N} d_i^3 v_i^2 n_i \tag{4}$$

where N is the number of drop diameter classes, n_i the number of drops in class I, d_i the mean diameter of drops in class I (mm), v_i is the impact velocity of drops of diameter d_i (m s^{-1}).

Cruse and Larson's experiments indicated a relationship between detachment D (g) and soil shear strength s (g cm^{-2}) of

$$(D^*10^4)^{0.5} = 6.4337 - 0.092s + 0.0004s^2 \tag{5}$$

Because shear strength decreases with the percentage of sand in a soil and increases with the percentage of silt and clay, detachment increases with the sand content of soils (Govers, 1991). Similarly, soil shear strength is affected by moisture content, such that detachment varies with moisture content (Truman and Bradford, 1990).

Dislodgement is difficult to measure, but laboratory experiments by Schultz *et al.* (1985) indicated that dislodged sediment may amount to around ten times that found in raindrop splash. This study remains unique in attempting to quantify dislodgement, despite its apparent importance to interrill erosion. Consequently, there remains a paucity of knowledge into rates of total detachment, as opposed to rates of rainsplash, (which are frequently, but incorrectly, equated with total detachment), that may be experienced under a range of rainfall and soil surface conditions.

Importantly, the rate of detachment is a function of the rainfall energy at the soil surface, so that where vegetation intercepts some of the energy of the rainfall, or a layer of surface water exists, some of the energy of the falling rain will be dissipated and be unavailable to cause sediment detachment. Wainwright *et al.* (1999) demonstrated that rainfall energy under creosotebush was reduced to 70% of that outwith the canopy during simulated rainstorms, whereas rainfall intensity was only reduced to 90%.

These authors argued that the difference in effective kinetic energy (that possessed by individual raindrops sufficient to cause soil detachment) was even greater (amounting to only 55% of that outwith the canopy), and that such differences in effective kinetic energy beneath vegetation canopies is responsible for the development of mounds beneath desert shrubs by the process of differential splash (Parsons *et al.*, 1992). In contrast, Morgan (1985) found an inverse relationship between rainfall energy and detachment beneath corn under simulated rainstorms of between 40 and 100 mm/hr, but a similar inverse relationship at only the lowest of these intensities under soybean, and a positive relationship between detachment and rainfall energy at the higher intensities. The argument is complex because interception by vegetation not only reduces the energy of falling raindrops available to detach soil particles, but also reduces their compaction effects. Thus, runoff is also likely to be less where rainfall is intercepted as found by Roy and Jarrett (1991), who investigated the effects of coarse surface fragments on interrill erosion. For a layer of water at the ground surface, Torri *et al.* (1987) demonstrated an exponential decrease in the detachment rate with increasing depth of surface runoff that could be described by the equation

$$A = A_0 e^{-qh} \tag{6}$$

where A is the rate of detachment ($m^3\,s^{-1}\,m^{-1}$), A_0 is the rate of detachment of at zero water depth, h is the depth of the water layer (m), and q is an empirical constant that varies with soil surface condition.

Finally, consideration needs to be given to the effect of surface gradient on rates of detachment. Poesen (1985) suggested that, under some circumstances, detachment varies with gradient so that there may exist a positive relationship between the rate of detachment and gradient. A more detailed examination of this phenomenon by Torri and Poesen (1992) demonstrated that this relationship results from a reduction of resistance to detachment with gradient that varies with its cohesion and angle of friction. Hence, the effect is more significant for some soils than others. For soils with a low (25°) angle of friction, the detachment rate doubles with an increase in gradient from 3° to 19°. In contrast, for soils with a high angle of friction (55°), there is only a 10% increase in detachment over the same range of gradient.

Whereas detachment in interrill areas is due to the energy of falling raindrops, sediment transport is mainly controlled by flow energy. Although splash droplets do transport some sediment, in most situations the amount is small in comparison. Kotarba (1980), however, argued that for the steppe zone of Mongolia, splash transport is the dominant form of sediment transport. On its granitic terrain, infiltration rates are high and overland flow is spatially of relatively little importance. In contrast, all rainfall causes some splash transport. Poesen (1985) proposed a sediment transport model for splash as follows:

$$qs = \frac{(KE)\cos a}{R(BD)}(0.301\sin a + 0.019(D_{50})^{-2.22})(1 - e^{-2.42\sin a}) \tag{7}$$

where qs is net downslope splash transport ($m^3\,m^{-1}\,a^{-1}$), KE is rainfall kinetic energy ($J\,m^{-2}\,a^{-1}$), R is resistance to detachment ($J\,kg^{-1}$), BD is bulk density ($kg\,m^{-3}$), a is slope (degrees) and D_{50} is median grain size (m).

For interrill flow, as with all sediment-transporting media, research has focused on the amount of sediment that can be transported – the sediment transport capacity, and the size of sediment that can be transported – the sediment transport competence. Several authors have attempted to apply transport-capacity equations developed for alluvial rivers (see the discussion in Ferro, 1998), and such derivations have been used in process-based models of soil erosion (e.g. Nearing *et al.*, 1989) even though the hydraulic conditions in shallow overland flow are very different from those in rivers. Everaert (1991) developed a range of equations for transport capacity TC ($g\,cm^{-1}\,s^{-1}$) of interrill flow for particles of different sizes D (μm), based upon Bagnold's (1980) modified stream power Ω ($g^{1.5}\,s^{-4.5}\,cm^{-0.67}$) that have the general form

$$TC = a\Omega^b D^c \tag{8}$$

where a is a constant and b and c are exponents that vary with particle size. Everaert noted that these equations were developed from experiments performed on regular beds, whereas interrill flow typically takes place on irregular surfaces. However, he argued that because stream power takes account of the effect of irregular beds on flow velocity, it is a suitable basis for interrill transport-capacity equations. Agarwal and Dickinson (1991) studied the effects of particle size, slope, and discharge on transport capacity of overland flow and undertook multiple regression to produce

$$q_s = Cq^{b1}S^{b2}D_{50}^{b3} \tag{9}$$

where q_s is sediment transport capacity ($Kg\,m^{-1}\,s^{-1}$), q is unit discharge, S is slope ($m\,m^{-1}$), and D_{50} is median particle size (mm).

Abrahams *et al.* (2001) proposed that sediment-transport capacity for turbulent interrill overland flow could be given by

$$\phi = a\theta^{1.5}\left(1 - \frac{\theta_c}{\theta}\right)^{3.4}\left(\frac{u}{u_*}\right)^c\left(\frac{w_i}{u_*}\right)^{-0.5} \tag{10}$$

where ϕ is the dimensionless sediment transport capacity, θ is the dimensionless shear stress, θ_c its critical value, u is

flow velocity, u_* is shear velocity, w_i is the inertial settling velocity (all $m\,s^{-1}$) of the sediment, and a is given by

$$\log a = \frac{-0.42C_r}{D_r^{0.2}} \tag{11}$$

where C_r is the concentration and D_r the size of the roughness elements (m), and c is given by

$$c = \frac{1 + 0.42C_r}{D_r^{0.2}} \tag{12}$$

However, much interrill overland flow is not fully turbulent, so that the validity of this equation for interrill flow, in general, may be questionable.

Wainwright and Parsons (1998) demonstrated the strong dependence of such equations on a knowledge of the hydraulics of flow. Such knowledge is seldom available. Even, as in their case, where a good estimate of the flow hydraulics is available, the differing relationships of these equations to depth of flow means that small errors in the estimation of this quantity can lead to markedly different predictions from the various equations.

The fundamental problem for understanding sediment transport in interrill flow is that sediment entrainment is achieved by the energy of falling raindrops (Young and Wiersma, 1973), whereas transport is mainly due to flow energy. Both Agarwal and Dickinson (1991) and Abrahams *et al.* (2001) sought to avoid the issue of sediment entrainment by delivering sediment to the flow from a hopper such that their analysis was concerned only with the ability of the flow to transport the sediment supplied to it. Parsons *et al.* (1998) argued that sediment transport by interrill flow needs to consider not only the capacity of the flow, but also its competence. Individual particles will be detached by raindrops and then transported a finite distance before coming to rest. Both the detachability and the transport distance of the particle are a function of its size. These authors undertook laboratory experiments and developed a general equation to predict travel distances of particles ranging from 3 mm to 10 mm in diameter under rain with intensities up to $138\,mm\,hr^{-1}$ falling onto shallow flow up to 5 mm deep. Their equation is

$$ML = 0.525RE^{2.35}FE^{0.981} \tag{13}$$

where M is particle mass (g), L is travel distance/unit time ($cm\,min^{-1}$), RE is rainfall energy ($J\,m^{-2}\,s^{-1}$), and FE is flow energy ($J\,m^{-2}\,s^{-1}$). Parsons and Stromberg (1998), using the same data set, developed a relationship for L against particle diameter D (cm). Their relationship

$$L = 9970D^{-3.44} \tag{14}$$

accorded with research undertaken by Church and Hassan (1992), showing that travel distance declines with particle size more rapidly than might be expected. They showed that this sharp decline in travel distance was due mainly to the fact that the probability of coming to rest increases rapidly with increasing particle size. In contrast, the probably of entrainment is much more weakly related to particle size. Such a relationship is consistent with the observation that sediment transported in interrill flow is finer than that detached by rainfall (Parsons *et al.*, 1991). Parsons *et al.* (2004) applied equation (13) to particles 80 μm in diameter and obtained good agreement between measured and predicted travel distances.

Since sediment detachment and transport are both dependent on the depth of interrill flow, the former inversely and the latter directly so, the two are not independent. Abrahams *et al.* (1991) investigated the downslope pattern of sediment transport on a large runoff plot in southern Arizona, and showed that initially sediment transport increased with distance from the top of the plot, but that it subsequently decreased. These authors attributed this pattern to downslope changes in the spatial pattern of interrill flow. However, an alternative interpretation that is consistent with the results of Parsons *et al.* (1998) is that sediment transport by interrill flow decreases downslope because of the reduction in detachment by raindrops with increasing flow depth (equation 6), and a less than compensatory increase in the effect of flow energy (equation 13).

RILL EROSION AND SEDIMENT TRANSPORT

As threads of interrill flow become deeper and faster, a threshold is reached beyond which significant flow detachment begins to take place. Once this occurs, the flow begins to erode definable channels or rills, on the hillslope. The threshold for erosion by flow and the formation of rills has been investigated by Govers (1985), Govers and Rauws (1986), and Bryan and Poesen (1989). These authors collectively came to the conclusion that a shear velocity of around $3–3.5\,cm\,s^{-1}$ was necessary for rills to develop. Rauws and Govers (1988) developed a relationship between the critical grain shear velocity and soil cohesion whereby

$$u_{*_{cr}} = 0.89 + 0.56C' \tag{15}$$

where $u_{*_{cr}}$ is critical grain shear velocity ($cm\,s^{-1}$) and C' is soil cohesion (kPa). However, as Nearing (1991) has pointed out, although the shear stress exerted by shallow flow around such a threshold is of the order of a few pascal, the typical shear strength of soils is several kPa. Furthermore, Nearing (1994) conducted laboratory experiments that demonstrated that, for similar mean shear stress, sediment entrainment by laminar flow is negligible compared to that for turbulent flow. These experiments

provide empirical support for the argument that sediment entrainment by shallow flow, and hence rill initiation, is only possible where the flow is turbulent. The important characteristic of turbulent flow is that, although mean shear stress may be only a few pascal, local and instantaneous shear stress can be two or more orders of magnitude higher, and hence sufficient to exceed soil shear strength. Parsons and Wainwright (in press) developed this argument further, noting that, whereas Nearing (1991) proposed that both turbulent-burst shear stresses and soil shear strength are normally distributed, the available evidence indicates that log-normal distributions are more characteristic (see Foster *et al.*, 1984; Parsons and Wainwright, in press). In addition, they drew attention to the study by Dunne and Aubrey (1986) which showed that, although incision may occur in response to soil detachment by flow, the formation of rills is inhibited by soil detachment by rainfall. Consequently, they proposed a threshold between rilled and unrilled surfaces that depended on three parameters: flow energy, rainfall energy, and soil-strength variability (Figure 1).

As with interrill overland flow, the transport capacity of rill flow has typically been estimated using equations taken from the literature developed for alluvial rivers. Govers (1992) undertook an analysis of the performance of these equations for shallow rill flow, and concluded that none performed well over the range of conditions he tested. He argued that previous positive results for some equations (e.g. the conclusion by Alonso *et al.* (1981) that Yalin's (1963) equation gave the best results, and Moore and Burch's (1986) successful application of unit stream power theory) were obtained because the experiments used in the evaluations were over a very limited range of conditions (e.g. in experiments by Alonso *et al.*, the slope never exceeded $0.07\,\mathrm{m\,m^{-1}}$, and all data used by Moore and Burch were collected on slopes of less than $0.05\,\mathrm{m\,m^{-1}}$),

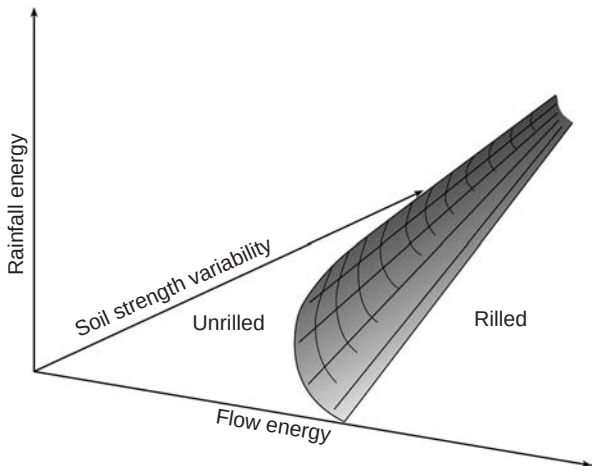


Figure 1 Parameters controlling the onset of rilling (after Parsons and Wainwright, submitter)

and that the empirical basis of all existing equations means that they are unlikely to be good predictors outside the domain for which they were developed. Nearing *et al.* (1991) demonstrated relationships between detachment by shallow flow and flow depth, slope, and size of aggregates. They showed that the detachment rate was not a unique function of either shear stress or stream power, and argued that if either variable were the correct one for determining detachment rates, then these rates would be equally sensitive to slope as to flow depth. However, their experimental results showed a greater sensitivity to slope than to depth. Brown *et al.* (1989) identified reduced rates of rill erosion as a result of incorporation of crop residue into the soil. They interpreted this result as caused by the stabilizing effect of crop residue on sand-sized aggregates.

Sediment transport by rill flow is equally determined by the competence of the flow. This is particularly the case for stony soils. Poesen (1987) undertook an analysis of sediment transport in rills in the loess region of central Belgium, and concluded that coarse rock fragments were more readily removed in shallow rill flow (implying a small ratio of flow depth to rock-fragment size) than might be anticipated from the Shields criterion, for which he proposed a critical value of 0.015 compared to the value of 0.05 used for flows on gentle gradients and for which the ratio of flow depth to sediment size is large. However, Abrahams *et al.* (1988) argued that higher values of the Shields criterion are needed to initiate movement of particles as the ratio of fragment size to depth decreases (particularly in cases where the ratio is <1).

GULLY EROSION AND SEDIMENT TRANSPORT

Thomas and Welch (1988) argued that the Universal Soil Loss Equation often underestimates erosion on agricultural land because it fails to take account of erosion from ephemeral gullies. More recent models for predicting soil erosion (e.g. EUROSEM, Morgan *et al.*, 1998), likewise, deal with rill and interrill processes but fail to model gully erosion. However, Foster (1986) suggested that gully erosion may be of equal importance to the sum of rill and interrill erosion. Osborn and Simanton (1989) noted that, whereas there is considerable qualitative literature on gully growth and development, quantitative information is limited. However, a more recent review by Poesen *et al.* (2003) presents data from a range of environments showing that gullies account for between 10% and 94% of total soil loss. Poesen and Govers (1990) distinguished between ephemeral gullies and gullies associated with banks. The former are continuous, temporary channels that may be capable of being obliterated by cultivation, but which recur in the same place during subsequent runoff events. Ephemeral gullies result from concentrated flow erosion,

that is, sediment detachment and transport are functions of flow intensity (Poesen and Govers, 1990). Begin and Schumm (1979) developed a relationship for the critical flow shear stress τ_{cr} whereby

$$\tau_{cr} = (c\gamma)A^{rf}S \tag{16}$$

where A is drainage area (ha), S is slope ($m\,m^{-1}$), rf is an exponent, c a constant and γ the weight of water per unit volume. Inasmuch as A and S together define threshold conditions for gully initiation, locations with and without gullies may be defined in terms of a discriminant function of the form

$$S = aA^{-b} \tag{17}$$

Vandaele *et al.* (1996) and Vandekerckhove *et al.* (2000) present a range of such functions for a number of sites and land-use conditions across Europe. Poesen and Govers subdivided ephemeral gullies according to their width–depth ratio (WDR): those that have a WDR ≤ 1, and those with WDR $\gg 1$. The latter can generally be removed by tillage, whereas the former cannot. Watson *et al.* (1986) developed a model to predict ephemeral gully width W where

$$W = 2.66Q_p^{0.396}n^{0.387}S^{-0.16}\tau^{-0.024} \tag{18}$$

where Q_p is peak flow discharge ($m^3\,s^{-1}$), n is Manning's roughness coefficient and τ_{cr} is critical flow shear stress (Pa). Nachtergaele *et al.* (2002) presented a simpler equation for the width of gullies in cropland where

$$W = 2.51Q^{0.412} \tag{19}$$

The similarity of the exponents in these equations around 0.4 also agrees with that obtained by Sidorchuk (1999). For ephemeral gullies in calcareous loess in central Belgium, Poesen *et al.* (1998) related detachment rate DR ($kg\,m^{-2}\,s^{-1}$) to concentrated flow shear stress τ ($N\,m^{-2}$), such that

$$DR = 0.75\tau - 0.08 \tag{20}$$

Gullies associated with banks form where a flow line crosses a steep bank in the landscape (often artificial). These gullies cannot be obliterated by tillage and are permanent. Poesen (1989) has shown that, unlike ephemeral gully erosion, flow intensity is not a dominant factor explaining erosion from gullies associated with banks. Local factors such as piping, mass movement and biotic influences play a significant role in the development of these gullies, and hence the amount of erosion that takes place in them. Gomez *et al.* (2003) used aerial photography to develop a relationship between sediment production and gully area over a period of 49 years in the Te Weraroa catchment in New Zealand. The relationship they obtained was

$$y = 460 + 2750x + 160x^2 \tag{21}$$

where y is sediment production ($m^3\,yr^{-1}$) and x is gully area (ha).

MODELING HILLSLOPE EROSION AND SEDIMENT TRANSPORT

Two main groups have undertaken modeling of hillslope erosion and sediment transport: geomorphologists and agricultural engineers. The latter group is concerned with soil erosion, particularly on agricultural land. The former group is concerned both with short-term modeling of erosion and sediment transport, where their work overlaps and can be considered with that of the agricultural engineers, and with long-term hillslope and landform evolution. There is a shared history, inasmuch as, through time, both demonstrate increasingly explicit representation of processes in their modeling as understanding of processes has increased and increased computing power has become available.

Modeling long-term hillslope and landform evolution can be dated at least as far back as the work of Fisher (1866), though the work of the likes of Bakker and Le Heux (1946, 1947, 1950), and Scheidegger (1961) and Hirano (1966, 1968, 1972, 1975, 1976). However, all such models are, at best, only weakly founded upon intuitive generalizations about the operation of hillslope processes (Parsons, 1988, p. 154). Kirkby (1971) argued that a general equation for the transport capacity C of hillslope processes that applies to all processes of entrainment and transport by water could be presented such that

$$C = f(a) \cdot \left(-\frac{\partial y}{\partial x}\right)^n \tag{22}$$

where a is the area drained per unit contour length, and n is a constant exponent describing the influence of increasing gradient. If it is assumed that erosion takes place at transport capacity and base level is fixed, then characteristic forms will evolve for each $f(a)$ and n. More recent long-term modelling of erosion has tended to consider three-dimensional landscape evolution and to adopt a numerical-simulation approach. Armstrong (1976) developed a similar transport-capacity driven model for landscape evolution, but allowed the rate of sediment removal also to be constrained by the rate of weathering. Willgoose *et al.* (1991a, b) developed the SIBERIA model in which

$$q_s = q_{sf} + q_{sd} \tag{23}$$

where q_s is sediment transport rate per unit width, q_{sf} is fluvial sediment transport, and q_{sd} is diffusive sediment transport (all $m^3 s^{-1} m^{-1}$ width). q_{sf} is given by

$$q_{sf} = \beta_1 q^{m_1} S^{n_1} \tag{24}$$

and q_{sd} by

$$q_{sd} = DS \tag{25}$$

where q is discharge per unit width ($m^3 s^{-1} m^{-1}$ width), S is slope ($m m^{-1}$), D is diffusivity ($m^3 s^{-1} m^{-1}$ width), and β_1, m_1 and n_1 are calibrated parameters. These, and other similar models share with Kirkby (1971) the representation of erosion processes and sediment transport as functions of slope length and/or gradient.

Soil erosion models have evolved to make explicit use of many of the erosion and sediment transport equations presented above. Early models, such as the USLE (Wischmeier and Smith, 1978), employed simple regression techniques to establish relationships between erosion rates and parameters of rainfall (intensity), soil (erodibility), slope (length and gradient), crop type and management practice, but without explicit representation of individual processes. More recent models (Nearing *et al.*, 1989; Morgan *et al.*, 1998) have taken explicit account of both interrill and rill erosion. In EUROSEM (Morgan *et al.*, 1998), for example, interrill detachment DR ($m^3 s^{-1} m^{-1}$) is modeled using the equation

$$DR = \frac{k}{\rho_s} KE e^{-qh} \tag{26}$$

which is equivalent to merging equations (1) and (6), with explicit determination of A_0, where k is an index of soil detachability ($g J^{-1}$), ρ is particle density ($kg m^{-3}$), and KE is rainfall kinetic energy ($J m^{-2}$), and interrill sediment transport capacity is modeling using the equation based on those developed by Everaert (1991).

Although members of this latest generation of soil erosion models make use of recent developments in understanding of erosion and sediment transport, they fail to include gully erosion (primarily because, as noted above, these models are for soil erosion on agricultural land, and the existence of gullies precludes mechanized agriculture), which tends to be studied separately (e.g. Sidorchuk, 1998; Woodward, 1999). However, to develop a full understanding of rates of erosion in different settings, gully erosion must be included. Furthermore, these models separate erosion and sediment transport in rills and interrill areas, whereas, in reality, there are dynamic changes during an erosion event such that rills grow in interrill areas. Favis-Mortlock (1998) developed a model to simulate the initiation of rills on previously unrilled surfaces. While this model is grounded in an explicit representation of erosion and transport processes, there remains some way to go, before a full understanding of the process of rill initiation is developed. The next step in modeling erosion and sediment transport processes on hillslopes needs to be the development of fully dynamic models that permit the growth of both rills and gullies.

REFERENCES

Abrahams A.D., Parsons A.J. and Luk S.-H. (1991) The effect of spatial variability in overland flow on the downslope pattern of soil loss on a semiarid hillslope, southern Arizona. *Catena*, **18**, 255–270.

Abrahams A.D., Li G., Krishnan C. and Atkinson J.F. (2001) A sediment transport equation for interrill overland flow. *Earth Surface Processes and Landforms*, **26**, 1443–1459.

Abrahams A.D., Luk S.-H. and Parsons A.J. (1988) Threshold relations for the transport of sediment by overland flow on desert hillslopes. *Earth Surface Processes and Landforms*, **13**, 407–419.

Agarwal A. and Dickinson W.T. (1991) Effect of texture, flow and slope on interrill sediment transport. *Transactions of the American Society of Agricultural Engineers*, **34**, 1726–1731.

Alonso C.V., Neibling W.H. and Foster G.H. (1981) Estimating sediment transport capacity in watershed modelling. *Transactions of the American Society of Agricultural Engineers*, **24**, 1211–1226.

Armstrong A.C. (1976) A three-dimensional simulation of slope forms. *Zeitschrift für Geomorphologie*, Supplementband, **25**, 20–28.

Bagnold R.A. (1980) An empirical correlation of the bedload transport rates in flumes and natural rivers. *Proceedings of the Royal Society, Series A*, **372**, 453–473.

Bakker J.P. and Le Heux J.W.N. (1946) Projective geometric treatment of O. Lehmann's theory of the transformation of steep mountain slopes. *Koninklijke Nederlandsche Akademie van Wetenshschappen. Series B*, **49**, 533–547.

Bakker J.P. and Le Heux J.W.N. (1947) Theory on central rectilinear recession of slopes. *Koninklijke Nederlandsche Akademie van Wetenshschappen Series B*, **50**, 959–966, 1154–1162.

Bakker J.P. and Le Heux J.W.N. (1950) Theory on central rectilinear recession of slopes. *Koninklijke Nederlandsche Akademie van Wetenshschappen Series B*, **53**, 1073–1084, 1364–1374.

Begin Z.B. and Schumm S.A. (1979) Instability of alluvial valley floors: a method for its assessment. *Transactions of the American Society of Agricultural Engineers*, **22**, 347–350.

Brandt J. (1990) Simulation of the size distribution and erosivity of raindrops and throughfall drops. *Earth Surface Processes and Landforms*, **15**, 687–698.

Brown L.C., Foster G.R. and Beasley D.B. (1989) Rill erosion as affected by incorporation of crop residue and seasonal consolidation. *Transactions of the American Society of Agricultural Engineers*, **32**, 1967–1978.

Bryan R.B. and Poesen J. (1989) Laboratory experiments on the influence of slope length on runoff, percolation and rill development. *Earth Surface Processes and Landforms*, **14**, 211–231.

Church M. and Hassan M.A. (1992) Size and distance of travel of unconstrained clasts on a streambed. *Water Resources Research*, **28**, 299–303.

Cruse R.M. and Larson W.E. (1977) Effect of soil shear strength on soil detachment due to raindrop impact. *Soil Science Society of America Journal*, **41**, 777–781.

Dunne T. and Aubrey B.F. (1986) Evaluation of Horton's theory of sheetwash and rill erosion on the basis of field experiments. In *Hillslope Processes*, Abrahams A.D. (Ed.), Allen and Unwin: Boston, pp. 31–53.

Ekern P.C. (1953) Problems of raindrop impact erosion. *Agricultural Engineering*, **34**, 23–25, 28.

Ellison W. (1944) Studies of raindrop erosion. *Agricultural Engineering*, **25**, 131–136, 181–182, 306.

Everaert W. (1991) Empirical relations for the sediment transport capacity of interrill flows. *Earth Surface Processes and Landforms*, **16**, 513–532.

Favis-Mortlock D.T. (1998) A self-organising dynamic systems approach to the simulation of rill initiation and development on hillslopes. *Computers and Geosciences*, **24**, 353–372.

Ferro V. (1998) Evaluating overland flow sediment transport capacity. *Hydrological Processes*, **12**, 1895–1910.

Fisher O. (1866) On the disintegration of a chalk cliff. *Geological Magazine*, **3**, 354–356.

Foster G.R., Huggins L.F. and Meyer L.D. (1984) A laboratory study of rill hydraulics: II shear stress relationships. *Transactions of the American Society of Agricultural Engineers*, **27**, 797–804.

Foster G.R., National Research Council, Board on Agriculture (1986) Understanding ephemeral gully erosion. *Soil Conservation: Assessing the National Research Inventory*, National Academy Press: Washington, pp. 90–118.

Ghadiri H. and Payne D. (1981) Raindrop impact stress. *Journal of Soil Science*, **32**, 41–49.

Gomez B., Banbury K., Marden M., Trrustrum N.A., Peacock D.H. and Hoskin P.J. (2003) Gully erosion and sediment production: Te Weraroa stream, New Zealand. *Water Resources Research*, **39**(7), 1187, doi:10.1029/2002WR001342.

Govers G. (1985) Selectivity and transport capacity of thin flows in relation to rill erosion. *Catena*, **12**, 35–49.

Govers G. (1991) Spatial and temporal variations in splash detachment: a field study. *Catena*, Supplement **20**, 15–24.

Govers G. (1992) Evaluation of transporting capacity formulae for overland flow. In *Overland Flow: Hydraulics and Erosion Mechanics*, Parsons A.J. and Abrahams A.D. (Eds.), UCL Press: London, pp. 243–273.

Govers G. and Rauws G. (1986) Transport capacity of overland flow on plane and irregular beds. *Earth Surface Processes and Landforms*, **11**, 515–524.

Hirano M. (1966) A study of a mathematical model of slope development. *Geographical Review of Japan*, **45**, 606–617.

Hirano M. (1968) A mathematical model of slope development. *Journal of Geosciences Osaka City University*, **11**, 13–52.

Hirano M. (1972) Theory on graded slopes. *Geographical Review of Japan*, **45**, 703–716.

Hirano M. (1975) Simulation of development process of interfluvial slopes with reference to graded form. *Journal of Geology*, **83**, 113–123.

Hirano M. (1976) Mathematical model and the concept of equilibrium in connection with slope shear ratio. *Zeitschrift für Geomorphologie*, Supplementband, **25**, 50–71.

Kirkby M.J. (1971) Hillslope process-response models based on the continuity equation. *Institute of British Geographers Special Publication*, **3**, 15–30.

Kotarba A. (1980) Splash transport in the steppe zone of Mongolia. *Zeitschrift für Geomorphologie*, Supplementband, **35**, 92–102.

Moore I.D. and Burch G.J. (1986) Sediment transport capacity of sheet and rill flow: application of unit stream power theory. *Water Resources Research*, **22**, 1350–1360.

Morgan R.P.C. (1978) Field studies of rainsplash erosion. *Earth Surface Processes*, **3**, 295–299.

Morgan R.P.C. (1985) Effect of corn and soybean canopy on soil detachment by rainfall. *Transactions of the American Society of Agricultural Engineers*, **28**, 1135–1140.

Morgan R.P.C., Quinton J.N., Smith R.E., Govers G., Poesen J.W.A., Auerswald K., Chisci G., Torri D. and Styczen M.E. (1998) The European soil erosion model (EUROSEM): a dynamic approach to predicting sediment transport from fields and small catchments. *Earth Surface Processes and Landforms*, **23**, 527–544.

Nachtergaele J., Poesen J., Sidorchuk A. and Torri D. (2002) Prediction of concentrated flow width in ephemeral gully channels. *Hydrological Processes*, **16**, 1935–1953.

Nearing M.A. (1991) A probabilistic model of soil detachment by shallow turbulent flow. *Transactions of the American Society of Agricultural Engineers*, **34**, 81–85.

Nearing M.A. (1994) Detachment of soil by flowing water under turbulent and laminar conditions. *Soil Science Society of America Journal*, **58**, 1612–1614.

Nearing M.A., Bradford J.M. and Holtz R.D. (1987) Measurement of water impact pressures on soil surfaces. *Soil Science Society of America Journal*, **51**, 1302–1306.

Nearing M.A., Bradford J.M. and Parker S.C. (1991) Soil detachment by shallow flow at low slopes. *Soil Science Society of America Journal*, **55**, 339–344.

Nearing M.A., Foster G.R., Lane L.J. and Finkner S.C. (1989) A process-based soil erosion model for USDA-water erosion prediction technology. *Transactions of the American Society of Agricultural Engineers*, **32**, 1587–1593.

Osborn H.B. and Simanton J.R. (1989) Gullies and sediment yield. *Rangelands*, **11**, 51–56.

Park S.W., Mitchell J.K. and Bubenzer G.D. (1983) Rainfall characteristics and their relation to splash erosion. *Transactions of the American Society of Agricultural Engineers*, **26**, 795–804.

Parsons A.J. (1988) *Hillslope Form*, Routledge: London.

Parsons A.J., Abrahams A.D. and Luk S.-H. (1991) Size characteristics of sediment in interrill overland flow on a semiarid hillslope, southern Arizona. *Earth Surface Processes and Landforms*, **16**, 143–152.

Parsons A.J., Abrahams A.D. and Simanton J.R. (1992) Microtopography and soil-surface materials on semi-arid piedmont hillslopes, southern Arizona. *Journal of Arid Environments*, **22**, 107–115.

Parsons A.J. and Gadian A.M. (2000) Uncertainty in modelling the detachment of soil by rainfall. *Earth Surface Processes and Landforms*, **25**, 723–728.

Parsons A.J. and Stromberg S.G.L. (1998) Experimental analysis of size and distance of travel of unconstrained particles in interrill flow. *Water Resources Research*, **34**, 2377–2381.

Parsons A.J., Stromberg S.G.L. and Greener M. (1998) Sediment-transport competence of rain-impacted interrill overland flow. *Earth Surface Processes and Landforms*, **23**, 365–375.

Parsons A.J. and Wainwright J. Depth distribution of interrill overland flow and the formation of rills. *Hydrological Processes*, (in press).

Parsons A.J., Wainwright J., Powell D.M., Kaduk J. and Brazier R.E. (2004) A conceptual model for determining soil erosion by water. *Earth Surface Processes and Landforms*, **29**, 1203–1302.

Poesen J. (1985) An improved splash transport model. *Zeitschrift für Geomorphologie*, **29**, 193–211.

Poesen J. (1987) Transport of rock fragments by rill flow – a field study. *Catena*, Supplement, **8**, 35–54.

Poesen J. (1989) Conditions for gully formation in the Belgian loam belt and some ways to control them. *Soil Technology Series*, **1**, 39–52.

Poesen J. and Govers G. (1990) Gully erosion in the loam belt of Belgium: typology and control measures. In *Soil Erosion on Agricultural Land*, Boardman J., Dearing J.A. and Foster I.D.L. (Eds.), Wiley: Chichester, pp. 513–530.

Poesen J. and Torri D. (1988) The effect of cup size on splash detachment and transport measurements. Part I field measurements. In *Geomorphic Processes in Environments with Strong Seasonal Contrasts*, *Volume 1: Hillslope Processes, Catena Supplement*, Vol. 12, Imeson A. and Sala M. (Eds.), pp. 111–126.

Poesen J., Nachtergaele J., Verstraeten G. and Valentin C. (2003) Gully erosion and environmental change: importance and research needs. *Catena*, **50**, 91–133.

Poesen J., Vandaele K. and van Wesemael B. (1998) Gully erosion: importance and model implications. In *Modelling Soil Erosion by Water*, *NATO ASI Series 155*, Boardman J. and Favis-Mortlock D. (Eds.), Springer-Verlag: Berlin, pp. 285–311.

Rauws G. and Govers G. (1988) Hydraulic and soil mechanical aspects of rill generation on agricultural soils. *Journal of Soil Science*, **39**, 111–124.

Roy B.L. and Jarrett A.R. (1991) The role of coarse fragments and surface compaction in reducing interrill erosion. *Transactions of the American Society of Agricultural Engineers*, **34**, 149–154.

Salles C., Poesen J. and Sempere-Torres D. (2002) Kinetic energy of rain and its functional relationship with intensity. *Journal of Hydrology*, **257**, 256–270.

Sharma P.P., Gupta S.C. and Foster G.R. (1993) Predicting soil detachment by raindrops. *Soil Science Society of America Journal*, **57**, 674–680.

Sharma P.P., Gupta S.C. and Foster G.R. (1995) Raindrop-induced soil detachment and sediment transport from interrill areas. *Soil Science Society of America Journal*, **59**, 727–734.

Scheidegger A.E. (1961) Mathematical models of slope development. *Bulletin of the Geological Society of America*, **72**, 37–50.

Schultz J.P., Jarrett A.R. and Hoover J.R. (1985) Detachment and splash of a cohesive soil by rainfall. *Transactions of the American Society of Agricultural Engineers*, **28**, 1878–1884.

Sidorchuk A. (1998) A dynamic model of gully erosion. In *Modelling Soil Erosion by Water*, *NATO ASI Series 155*, Boardman J. and Favis-Mortlock D. (Eds.), Springer-Verlag: Berlin, pp. 451–460.

Sidorchuk A. (1999) Dynamic and static models of gully erosion. *Catena*, **37**, 401–414.

Soil Conservation Society of America (1982) *Resource Conservation Glossary*, Soil Conservation Society of America: Ankeny.

Thomas A.W. and Welch R. (1988) Measurement of ephemeral gully erosion. *Transactions of the American Society of Agricultural Engineers*, **31**, 1723–1728.

Torri D. and Poesen J. (1988) The effect of cup size on splash detachment and transport measurements. Part II theoretic approach. In *Geomorphic Processes in Environments with Strong Seasonal Contrasts*, *Volume 1: Hillslope Processes, Catena Supplement*, Vol. 12, Imeson A. and Sala M. (Eds.), pp. 127–137.

Torri D. and Poesen J. (1992) The effect of soil surface slope on raindrop detachment. *Catena*, **19**, 561–578.

Torri D., Sfalanga M. and Del Sette M. (1987) Splash detachment: runoff depth and soil cohesion. *Catena*, **14**, 149–155.

Truman C.C. and Bradford J.M. (1990) Effect of antecedent soil moisture on splash detachment under simulated rainfall. *Soil Science*, **150**, 787–798.

Vandaele K., Poesen J., Govers G. and van Wesemael B. (1996) Geomorphic threshold conditions for ephemeral gully incision. *Geomorphology*, **16**, 161–173.

Vandekerckhove L., Poesen J., Oostwoud Wijdenes D., Nachtergaele J., Kosmas C., Roxo M.J. and de Figeiredo T. (2000) Thresholds for gully initiation and sedimentation in Mediterranean Europe. *Earth Surface Processes and Landforms*, **25**, 1201–1220.

van Dijk AIJM, Bruijnzeel L.A. and Rosewell C.J. (2002) Rainfall intensity – kinetic energy relationships: a critical literature appraisal. *Journal of Hydrology*, **261**, 1–23.

Wainwright J. and Parsons A.J. (1998) Sensitivity of sediment-transport equations to errors in hydraulic models of overland flow. In *Modelling Soil Erosion by Water*, *NATO ASI Series 155*, Boardman J. and Favis-Mortlock D. (Eds.), Springer-Verlag: Berlin, pp. 271–284.

Wainwright J., Parsons A.J. and Abrahams A.D. (1999) Rainfall energy under creosotebush. *Journal of Arid Environments*, **43**, 111–120.

Watson D.A., Laflen J.M. and Franti T.G. (1986) *Estimating Ephemeral Gully Erosion*, Paper No. 86–2020, American Society of Agricultural Engineers: St. Joseph.

Watung R.L., Sutherland R.A. and El_Swaify S.A. (1996) Influence of rainfall energy flux density and antecedent soil moisture content on splash transport and aggregate enrichment ratios for a Hawaiian oxisol. *Soil Technology*, **9**, 251–272.

Willgoose G.R., Bras R.L. and Rodriguez-Iturbe I. (1991a) A physically based coupled network growth and hillslope

evolution model: 1. Theory. *Water Resources Research*, **27**, 1671–1684.

Willgoose G.R., Bras R.L. and Rodriguez-Iturbe I. (1991b) A physically based coupled network growth and hillslope evolution model: 2. Applications. *Water Resources Research*, **27**, 1685–1696.

Wischmeier W.H. and Smith D.D. (1978) *Predicting Rainfall Erosion Losses*, USDA Agricultural Handbook No. 537.

Woodward D.E. (1999) Method to predict cropland ephemeral gully erosion. *Catena*, **37**, 393–399.

Yalin M.S. (1963) An expression for bed-load transportation. *Proceedings of the American Society of Civil Engineers, Journal of the Hydraulics Division*, **89**, 221–248.

Young R.A. and Wiersma J.L. (1973) The role of rainfall impact in soil detachment and transport. *Water Resources Research*, **9**, 1629–1636.

81: Erosion Monitoring

BENT HASHOLT

University of Copenhagen, Copenhagen, Denmark

First the need for and relevance or erosion monitoring is stated, and then definitions, terms and classifications of the topic are described. The monitoring is treated for active erosive agents under natural conditions and at different scales. Monitoring in controlled systems and in laboratory experiments are treated separately. Special emphasis is on problem encountered in monitoring programs. References on key instrumentation are provided.

THE NEED FOR AND THE MEANS OF EROSION MONITORING

Erosion has a central role in global material cycling; our whole perception of the formation of the surface of the earth is based on knowledge of land forming processes and geological processes, such as orogenesis and the ensuing denudation of the mountain ridges. Monitoring of erosion is fundamental for establishing the rates of these processes. It is therefore of primary significance for the evolution of basic research in geosciences. Scientific subjects such as geomorphology, geochemistry, and geology need to monitor erosion processes in order to understand mechanisms and to extrapolate in space and time. Erosion monitoring is therefore one of the basic issues in quantitative interpretation of basic research in these fields (Loughran, 1989). It is of ultimate importance to be able to quantify the outcome of different processes in order to determine their relative significance. Within recent decades, modeling of the global material cycles has become increasingly important in order to assess the growing impact of man. To produce models that provide relevant and reliable results, it is important that all major processes are included, the mechanisms known and their relative importance quantified. Monitoring is therefore essential to provide knowledge of processes and to provide data for model testing and validation within basic research in the geosciences.

Erosion monitoring is of similar importance in fields of applied science, such as agricultural engineering and construction engineering. In these cases, there are direct practical needs for the results of the monitoring. The growth of the population of the Earth leads to utilization of former virgin land and often the more intensive use of already cultivated areas. Sustainable rates of erosion must be established for different types of land and different types of crops and agricultural practice. Other human activities that reshape the surface of the Earth include urbanization and large-scale construction works. Urbanization in developing countries might lead to severe erosion problems that need to be monitored, so that they can be taken into account in city planning. Engineers working in the field of river training and regulation have long recognized the need for monitoring, and much progress in the understanding of erosion and transport processes originate from this field. A specialist aspect of construction work is the building of large dams for power production, irrigation and flood control, and often all purposes are combined. The lifetime of these reservoirs is very dependant on erosion rates in their catchments, but also the effect of the dams on erosion and sedimentation in river-reaches below the dam might need to be monitored in order to cope with environmental problems, for example, scour at installations close to the river and/or damage to coral reefs due to flushing of reservoirs.

To summarize, erosion monitoring is of critical importance both for the further development of basic earth sciences and for applied sciences dealing with construction works of vital importance for sustainable human life.

DEFINITIONS, TERMS, AND CLASSIFICATION OF EROSION MONITORING

Erosion is the removal of material (as chemical compounds or in particulate form) from surfaces caused by one or more active agents. Active agents are: wind (sometimes including the content of snow or particulate matter), rain,

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

running water, sediment-laden water, mass movements, and glaciers. A complete review of erosion monitoring should therefore include all the agents mentioned However, most monitoring literature tends to concentrate on water and wind, probably because these are most often encountered when dealing with practical erosion problems in populated areas.

Different physical processes are active for the different agents; therefore, there is an interaction between monitoring and perception of process. Looking at erosion on a microscale, the monitoring could focus on the very moment that a chemical compound or a particle is removed from the surface; it could focus on the void left by the particle or the velocity and mass of the removed matter, that is, the transport. From a monitoring point of view, it is therefore important to distinguish the process of erosion, the forms caused by removal of material, or the transported matter released by the process. However, such a distinction is not always possible and must partly be considered as a scale issue. From the microscale to the global scale, erosion takes many different forms and could therefore be classified according to both agent and scale. Often monitoring embraces a cascade of processes and only considers the end result, for example, monitoring the sediment output from a plot, without considering the several phases of erosion and sedimentation occurring within the plot, before the sediment reaches the outlet of the plot.

Very often erosion is monitored by measuring the transport of material released by erosion. A problem related to measurements of transport is that the transport very often consists of different phases, for example, fluids and particles, which have to be monitored separately, for example, by filtering of water samples. Another classification of monitoring might be based on the methodology employed, but this approach is difficult because of conflicts with scale and process issues. An important precondition for the monitoring of erosion is the ability to identify types of processes and the related morphological forms, for example, drop marks, rills, and gullies. There is a clear need for standardized definitions of erosion patterns and their dimensions. An example of an attempt to standardize the recording and mapping of erosion features is found in Mollenhauer (1995) and Thorne (1998).

In the following section, the description of erosion monitoring is based on erosive agent and on types of erosion from small scale to larger scale, assuming that the practical use of an encyclopaedia will be to find examples of how to monitor different types of erosion, either for pure practical reasons or for experimental purposes.

EROSION BY WATER

Rain

The energy of falling raindrops can cause erosion, termed *splash erosion*, which can be separated into the detachment of particles and the associated saltation of the particle caused by the collision impact. Erosion can be measured as material removed from the surface by comparison of images obtained before and after the erosion event; the procedure is facilitated by image-analyzing software. Mass transfer can be determined using image analysis on records assembled using high-speed cameras. Tracers, for example colored particles, can be used to measure travel distances and particle sorting. Often different types of trap are used to collect the material moved by splash erosion, for example, splash boards or plates that are placed near the erosion surface and weighed before and after an erosion event. In other cases, funnels are located flush with the surface and the material is collected in a bottle below the funnel. In other cases, splash cups may be used. Here a sample of soil in the center of the cup is exposed to the rainfall and the detached material is collected in the surrounding tray (see Figure 1). The size of the splash cup influences the result and must be taken into account in correction of the result, (Poesen and Torri, 1988). Rainfall intensity and related energy are the most important causes of erosion, however, wind speed and direction relative to slope, and the aspect of the eroded surface, play an important role in influencing the amount of erosion; to explain differences in the results, it is important to include such measurements in the monitoring programe.

Running Water

Running water on the surface will occur either because of saturation of the soil with water (saturation overland flow) or because the infiltration capacity is exceeded by the rainfall intensity (Hortonian overland flow). Water on the surface is critical for the occurrence of erosion; however, it is rather difficult to observe this kind of flow and even more difficult to monitor it. Very often it is important just to

Figure 1 Splash cup. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 2 Overland flow sampler. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 3 Gerlach trough. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

know if overland flow has occurred, and an overland flow sampler can be used (Figure 2). The sampler can be placed in minor depressions in the landscape but are often placed at "worst-case" locations, that is, places where overland flow will occur each time there is overland flow. The sampler will then indicate whether erosion by running water has been possible, but to quantify the amount, it is necessary to install several samplers. The small size of the sampling device reflects the need to place it in small depressions in the terrain, with the requirement that it should be easy to install and should not interfere with the management of agricultural fields. The flanges at the intake section assist water intake, prevent undercutting by running water, and define the width of the intake area (Hasholt, 1998). The time sequence of overland flow can be determined by installing a pressure transducer in the collecting bottle. Most eroded location spots commonly coincide with the concave parts of slopes with small depressions, protorills or previous pathways of running water.

The detachment of single particles can be observed either by the naked eye or using cameras. Often such observations are carried out in flumes where the bed is covered with homogenous material to determine critical bed shear stress or critical velocity. In nature, erosion by running water might be represented by linear features and particle separation on the surface, leaving streaks of sand and coarser particles. This type of erosion is termed *sheet erosion* or *interrill erosion. In situ* it can be observed using analysis of high-speed camera images. Often the erosion is determined using repeated surveys of the surface. The lowering of the surface is the net erosion after the erosion event. An alternative technique is to measure the lowering of the surface relative to markers placed in the surface, for example, erosion pins. The number of pins and their spacing determines the accuracy of the monitoring. The pins must not be too large and create their own scour patterns; on the other hand, they should be large enough to resist the erosive forces. Erosion might also be observed relative to natural objects in the field, for example, tree roots and rock outcrops. In urban areas, the erosion can be related to foundations of houses or fences. Due to its large area and its small vertical change, sheet erosion is difficult to monitor for shorter time durations or single events. An alternative approach involves the use of tracers; this will be described later in relation to erosion from slopes. Special problems occur on cultivated fields, because the tillage operations may allow only a small time "window" for operation of samplers that have to be removed before ploughing or harrowing. Larger samplers employing the same trapping principle have been constructed by Gerlach (1966). A gutter of approximately 50-cm width (Figure 3), is buried on the slope to be investigated, and several gutters are often placed *en echelon* on the slope. It is then important that the volume of the collection bottles for the lower troughs should match the larger runoff values. Several problems have been observed, in particular when freezing occurs, the lip can be displaced so that bypass and undercutting might occur, and the pits for the collecting bottles can be filled with water, unless they are properly drained. It is also difficult to determine the contributing area exactly. However, useful information about the occurrence and frequency of erosion events can be obtained. The samplers described above and other trap-type samplers are capable of monitoring interrill erosion and to indicate whether erosion by running water has occurred. Rain splash can impact through the film of running water, so that soil detachment and transport occur simultaneously.

Rill erosion is characterized by recognizable micro "stream channels", as a certain critical shear velocity must

be surpassed in order to initiate rills, the amount of surface runoff, and therefore, the contributing area needs to be greater than for sheet erosion. The slope length needed to initiate rills is usually more than 10 m. To monitor rill erosion in nature, one has to be able to find places of occurrence and to identify the form. Risk maps might be used to assess the areal extent of rill erosion, but field observations are often needed to test the predictions from the risk maps. Once the location of rills is satisfactorily determined, a measuring program can be initiated. If the dimensions of a single rill are measured in selected cross-sections, the eroded volume for a reach of the rill can be determined by multiplying the cross-sectional area by half the length of the rill between consecutive cross-sections. Assuming that a correct volume can be determined by measurements of cross-sections every 50 cm, it was found that the rill volume can be determined within ± 5 % using a spacing of *circa* 5 m between cross-sections. To monitor the erosion from a whole rill system, a transect must be laid out along the contours and the cross-section of the intersecting rills must be measured. Counting of rills and measurement of representative rills might also be applied. Large-scale aerial photos or satellite images can be applied to locate rill systems, and, to a certain level, to also measure the lengths of rills. However, minor rills might only be observed by careful fieldwork. Major effort has to be spent on representative evaluation and calculation to obtain accurate observations.

Gullies are relatively permanent steep-sided water courses that experience ephemeral flows during rain storms; they are often characterized by a head cut and various steps or knick-points along their course. They can be monitored in the same way as for rills; however, their larger dimensions might call for use of theodolites and optical distance meters. For the location of gully areas, aerial photos and satellite images might be more useful. Measurements of the volumes of both rills and gullies will not include sediment transported from the interrill and intergully areas, and the overall erosion therefore tends to be underestimated. Theoretically, this erosion could be monitored if the flow of water and sediment in rills and gullies were recorded during erosion events. However, as these forms originate during the erosion event, choosing the right location for monitoring equipment is nearly impossible. Furthermore, because of the ephemeral character of the flow, very frequent sampling is needed. Because of the large range of particle sizes transported, the sampler also has to be capable of dealing with this. Another problem, particularly in the case of rills, is that farmers till the soil as soon as the soil is dry enough after the winter, when erosion often occurs. This leaves a very small "time window" for monitoring of rill systems in cultivated areas. Monitoring of erosion on hillslopes must include measurement of the deposition of material at the lower part of the slopes; otherwise the net erosion is overestimated. On its way to the permanent drainage system, the eroded material, when passing across the banks of streams, might cause bank erosion, which can be monitored using surveying methods or erosion pins. Erosion associated with channel, bed, and bank erosion can be monitored using surveying methods. However, this approach cannot account for rapid changes. A method for continuous registration of bank erosion (PEEP) has therefore been developed by Lawler (1992).

An evaluation of sediment mobilization within a catchment must in principle be undertaken at all erosion localities in the catchment, for example, hillslopes and streams. However, such a task is huge and therefore unrealistic, and the catchment is commonly divided into smaller representative areas. Such sampling areas are chosen to represent different scales, from single field to slope units and first order catchments nested inside the study catchment.

Another method used to describe the areal extent of erosion and deposition is measuring the redistribution of selected tracers (Zapata, 2002). Caesium-137 is the most commonly used tracer in soil erosion measurements. This radionuclide was produced from the 1950s to the 1970s by atmospheric testing of nuclear weapons. After it was distributed globally in the stratosphere, it was subsequently deposited on the Earth's surface by rain and as dry fallout. Regional variations in fallout as a result of precipitation-amount occur, but within small areas a reasonably uniform spatial distribution can be assumed. After reaching the surface, the isotope is quickly and strongly adsorbed to clay and fine silt particles in the soil, and an absence of particles capable of adsorbing the isotope is a potential limitation for the application of the method. Since the method was pioneered in the United States by Ritchie and McHenry (1975), it has been applied successfully in many regions. A thorough description is given in Walling and Quine (1990). The radiocaesium content of soil cores collected on a grid system varying in density from 10×10 m to 20×20 m is determined and the spatial pattern of caesium-137 inventories is established. Figure (4) (Walling and Quine, 1990) illustrates the principles of the method. An area that has no erosion or deposition and has not been cultivated is commonly used as reference site, most of the isotope being found near the surface. The presence of the radiocaesium at depth is assumed to be caused by biological activity, for example, earthworms. Cultivated land is characterized by a more even distribution down to the plough depth. At sites with no erosion, the inventory is the same as at the reference site. At eroded locations, the inventory is less and at deposition sites the inventory is larger than at the reference site. The radiocaesium inventory can be calibrated to estimate erosion and deposition rates by comparison with samples collected from erosion plots, where the erosion rate is known (see below). Point estimates of soil loss may differ by as much as an order of magnitude;

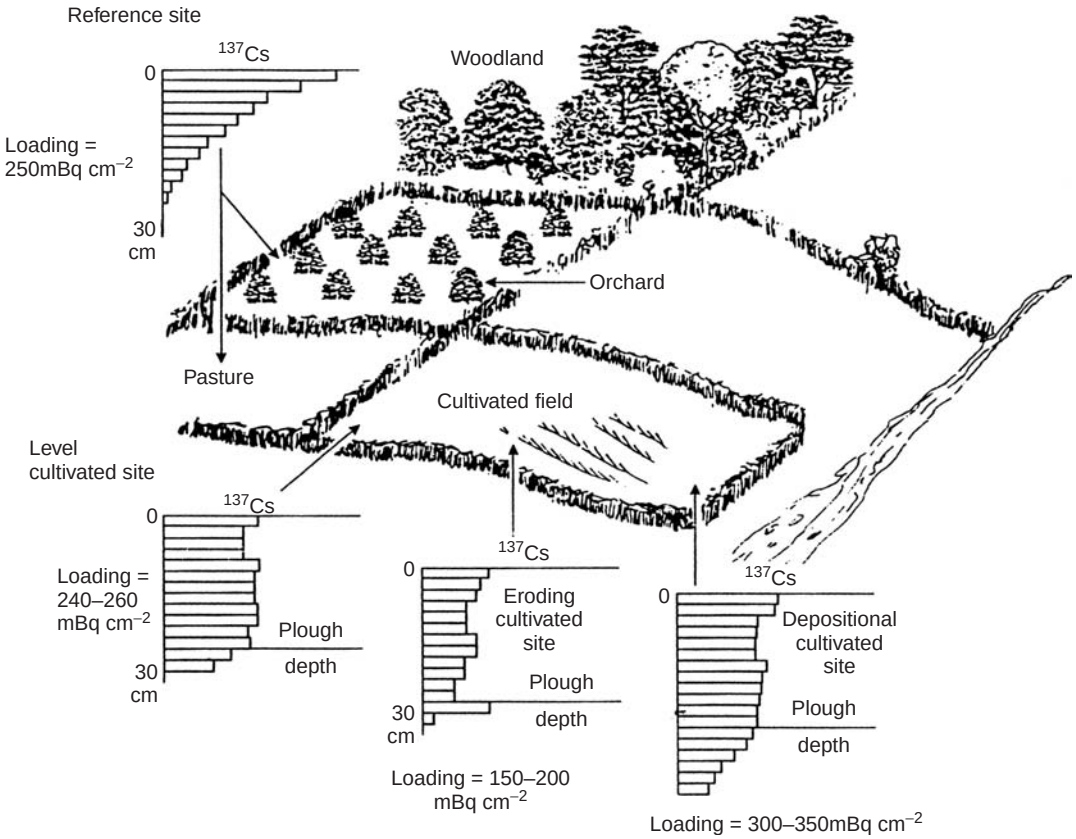


Figure 4 Schematic representation of the effect of erosion and deposition upon the loading and profile distribution of caesium-137. (Reproduced from Walling and Quine, 1990, by permission of John Wiley & Sons)

possible errors include uneven mixing of the radionuclide in the cultivated layer and changes in land use since fallout. Preferential adsorption of the isotope during the processes of erosion, transport, and deposition may be another source of error. The method accounts for the sum of all erosion processes, whereas other measurements, for example, erosion plots, may represent only the influence of interrill and rill erosion.

When delivered to the drainage system, the eroded material can be monitored while passing a gauging station located at a cross-section near the catchment outlet. Eroded material is carried by the water either in solution or as particles. In both cases, the concentration has to be determined, either by chemical analysis, by filtering, or by centrifuging. The transport rate is calculated as the flow of water multiplied by the sediment or solute concentration.

Water flow can be monitored using weirs and current meters, or electromagnetic sensors, (Hershey, 1999). Water samples used to measure suspended sediment concentrations should be collected using isokinetic samplers so that particle separation is avoided, (Hershey, 1999). Sampling must represent the actual distribution of the particulate matter, and an average concentration can be obtained using depth-integrated sampling, and standardized methods are described in ISO manual 4363. Several investigations have demonstrated that very frequent sampling is necessary to obtain accurate and reliable results in the case of erosion events caused by storms. Several types of automatic water sampler have been constructed to overcome this problem. Water is drawn into an evacuated sampling bottle or pumped using different types of pumps, for example, peristaltic pumps as in the ISCO sampler.

The main problems using automatic water samplers are finding of a representative location for the intake and the need to regulate the flow velocity in the intake to avoid particle separation. Indirect measurements for determining the concentration of specific ions and of particles have been employed, but all methods need calibration against standard analysis. Successful application of transmissometers that measure the obscuration or attenuation of white or infrared light in a measuring gap due to the presence of particles has, for example, been reported by Walling (1977) using Partech equipment. Other sensors measure the optical backscattering caused by the presence of particles. In the case of very high concentrations, measurements of the attenuation of γ-radiation have been employed. The location of the sensors can be a problem, together with the fouling of the sensors. Sensors must be located where the

stream is flowing freely, without influence from upstream point sources; in some cases they must be located within a stream concentrator, Hasholt (1991). If the sensors are located in the stream, false results may occur because of drifting debris collecting on the sensor. In this case, self-cleaning, for example swing mountings, can be used. To minimize fouling, at least weekly cleaning may be necessary. Some sensors wipe themselves clean at regular intervals or keep the measuring gap protected until the measuring time. A particular problem is the monitoring of bed load together with the unsampled zone below the intake of the samplers mentioned above. In principle, bed load can be monitored using different types of traps. Pit traps are buried flush with the riverbed; it is, however, very difficult to keep the pit-sampler flush with the mobile bed surface. Another type of sampler, the so-called pressure-difference sampler, has a smaller intake section that is gradually widened so that the streamlines inside the sampler diverge causing velocity to drop and the moving sediment to settle. After a period of time, the sampler is brought to the surface and the collected sediment is weighed and the transport rate is calculated from the time of exposure of the sampler on the riverbed. Another method involves repeated surveys of moving bed-forms; the velocity of the bed-form may be found and multiplied with the average height of the bed-form along its full length. Tracers can also be used, for example, magnetic particles, or particles dyed with fluorescent colours that can be counted under ultraviolet light. In the case of bed load consisting of larger particles, pebble size, built-in radio transmitters may be traced using special antennas (Ergenzinger *et al.*, 1989). Continuous recording of bed load has been carried out using pressure transducers built into the riverbed, or underwater microphones that record the noise of the impact of the transported particles. A thorough review of bed load monitoring methods is presented in Hubbel (1964) and new techniques are described in Kuhnle and Derrow (1994).

The results from gauging stations are often considered evenly distributed over the catchment area, however, if the catchment is large, for example, greater than 10 km^2, it may be very difficult to relate the transport at the gauging station to erosion events or certain types of erosion directly, because of delivery delays and sedimentation within the catchment. Lakes and reservoirs in catchments may be used to measure sedimentation. The volume of the reservoir can be surveyed at successive occasions, and the difference between survey is then the sedimentation volume caused by net upstream erosion in the catchment. To determine the erosion by weight, the density of the deposited sediment must be determined by analyzing representative cores. The volume differences can be determined quite rapidly using echosounders and a combined positioning system based on GPS. Before the results are applied, they must be corrected for bypass loss based on knowledge of the trap-efficiency of the lake or the reservoir (Brune, 1953).

MASS MOVEMENTS

Water plays a role in debris flows and mass movements, mostly by increasing mass beyond a critical point or by providing lubrication and the presence of slip surfaces. Large masses of material may be eroded and moved by these processes. Monitoring is mainly carried out using surveying techniques, as discussed above. Mass movements can be monitored using laser distance meters or transmitting GPS location devices. Because of the episodic occurrence of mass movements, it is very difficult to establish monitoring programs, and monitoring is therefore often *ad hoc* and carried out at places where these movements interfere with road works or other technical installations. Very few, if any, investigations of mass movements in whole catchments exist (Temple and Rapp, 1972; Selby, 1993).

EROSION BY WIND

Erosion by wind can be monitored using the same surveying methods that are applied for water erosion. An indication of the efficacy of wind erosion may be obtained by grain size analysis of eroded surfaces and surfaces in depositional areas. Often, wind erosion on fields can be estimated by measuring the depth of sediment deposited against hedges and shelter belts, although fine material can be carried very long distances. From personal experience, for example, in Ghana during the Harmattan, dust from source areas 1000 km away can be caught in traps. As a result of shifting wind directions and simultaneous changes in the length of fetch, it is very difficult to relate wind erosion to specific areas.

On site, different types of trap are applied for monitoring the transport caused by the erosion. The horizontal pit type is buried level with the eroding surface, the catch of sediment varying with the length of the sampler along the wind direction (Borsy, 1972). It is very difficult to estimate the contributing area because of shifts in the direction of the wind. Other types of samplers are trays placed vertically above the surface; they are operated by the pressure-difference principle, like many river bed load samplers. To keep the orifice orientated into the wind, the sampler can be mounted with bearings and a vane so it can follow the shifting wind directions. Several types are described by Bagnold (1941), and Nickling and McKennea Neuman (1997), for example. Such equipment can also be made continuously-recording by mounting the collecting tray on a recording balance. Other automatic monitoring devices are based on acoustics, and record the noise created by the impact of the moving particles on a membrane

with a microphone behind, for example, the saltiphone developed at the Agricultural University, Wageningen, in the Netherlands (Spaan and van den Abeele, 1991). Calibration against other traps is, however, needed.

EROSION BY ICE

This kind of erosion is rarely included in normal monitoring programs, probably because it acts over longer timescales and is not active close to populated areas. Recently, this type of erosion has become of interest because of the growing interest in global material budgets and the development of hydropower plants in glaciated areas. The effect of glacial erosion can be monitored using surveying methods to measure the volume of the landforms created by the moving ice. Subglacial erosion has been monitored at Bondhusbreen, Norway (Østrem, 1975), where eroded sediment was collected in a sedimentation chamber close to the sole of an active glacier. This chamber acts as a sediment trap, which is emptied from time to time. From the same area, information about eroded volume can be obtained from cores taken in a proglacial lake where the thickness of varves can be measured. Glacial erosion is often quantified by establishing a gauging station on a proglacial river, just outside the terminus of the glacier. Such investigations are described by Gurnell and Clark (1987), and Hasholt (1992).

MONITORING OF EROSION IN CONTROLLED SYSTEMS

Measuring erosion in nature and under natural climatic conditions is difficult because of the random location of erosion sites, and the episodic nature of the climatic conditions leading to erosion. The need for information about the effect of changes in single parameters such as soil type, crops and tillage practices has led to the development of controlled installations, where factors determining erosion can be manipulated and the resulting erosion measured accurately.

A major problem concerning measurement of erosion under natural conditions involves the determination of the contributing area. The controlled system is therefore most often a bounded area, a so-called plot. The plot edges are made of stable material, sheets of metal, plastic, or similar material that does not leak or rust. The edges should be embedded deep enough to avoid loosening by freeze thaw or drying of the soil, and so that no undercutting by running water can take place.

The edges should extend well above the surface to avoid overtopping by running water and significant in-splash; normally a height of 15–20 cm is chosen. The size of such plots ranges from 0.5 m^2 up to about 400 m^2. The

Figure 5 Erosion plot with rainfall simulator. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

choice of plot size depends on the processes to be studied; splash and interrill erosion can be studied using the smaller plots, but for studies of rill erosion, plots must be at least 10 m long. Normally the plots (Figure 5) are rectangular with the longest side parallel to the slope direction. In order to evaluate the effect of slope gradient, plots of same area and length may be established on slopes of varying steepness, representing typical conditions in the surrounding catchment. A standard plot with a length of 22 m, a width of 1.8 m, and a slope of 5° has often been used in soil erosion studies all over the world. At the downslope end of the plot, the eroded sediment is trapped in a gutter or collecting trough covered with a lid to prevent direct entry of rainfall. From the gutter, the water and sediment is led to a measuring and collecting system, which can be designed in many different ways. The construction of the gutter and the measuring and collecting system is critical to the ability of the system to provide accurate results for the time resolution required. In many cases, the gutter bottom has a gentle slope or is horizontal, causing the deposition of the larger and heavier fractions of the eroded material. The gutter therefore has to be emptied separately after a storm, and it is impossible to measure the true sediment concentration distribution during the storm. To facilitate this, the gutter must be constructed so that its slope exceeds that of the plot. From the gutter, runoff and sediment are channeled into collecting tanks. The number and construction of the collecting tanks depends on the size of the plot and the volume of runoff. In the case of large plots and runoff volumes, the overflow from the first tank passes into another tank through a divisor which splits the flow and passes one part as a sample into the next tank. Examples of divisors are the Coshocton wheel (see Figure 6) and the Geib multislot divisor. The sediment in the water can be allowed to settle or a flocculating agent

Figure 6 Coshocton wheel. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 7 Sediment sampler funnel in collecting position. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

can be added. The clear water is drawn off and the volume of remaining sediment is measured, dried, and weighed. If the volume is too large, only a sample of known volume is weighed and the total weight in the tank is found by multiplication by the sample value. Where a divisor is used, the volume from the measuring tank must be multiplied by the division factor. The volume of water and sediment may also be measured using flumes or weirs. Often, an H-flume is used because it is nonsilting and unlikely to become blocked by debris. Thompson type triangular weirs, with different angles, are also often used. Flow can also be registered by counting the rotations of the Coshocton wheel or the number of tips with a tipping-bucket device.

If the objective of the sampling is to determine the time variation of the sediment concentration in the runoff from the erosion plot, the methods above are insufficient. If observers are present, manual in-stream sampling is an option, assuming that no sedimentation take place in the collecting system before the sampling point. If manual sampling is not possible, automation is needed. Several automatic samplers are available. However, an important problem is that high concentrations and large particle sizes can clog sampler tubes, and it may also be difficult to obtain velocities large enough in the sampling tubes to prevent separation of particle sizes. Another possibility is the application of transmissometers or backscatter sensors. Practical experiences have, however, proved that these sensors are very difficult to install in the system without causing disturbance of flow, resulting in particle separation. To overcome these problems, an automatic sampler without pumps has been constructed (Hasholt, 1998). The water and sediment mixture is passed through a moveable funnel. When not sampling, flow is passed directly to a main collecting tank; when sampling, a solenoid places the moveable funnel above a fixed funnel that directs the water and sediment mixture directly into a sampling bottle placed on a conveyor belt (Figure 7). After sampling over a predetermined time interval, the moveable funnel swings back, activated by a spring. The collecting tank is placed on a balance capable of weighing 1500 kg with a resolution of ± 100 g. Simultaneously the level of the water in the collecting tank is measured with an ultrasonic distance meter with a resolution of ± 1 mm. Theoretically, the flow mass can be calculated by the simultaneous solution of two equations with two unknowns, knowing the density of water and of sediment. Because the accuracy of volume determination is an order of magnitude less than that of weight, the overall accuracy is not good enough for a calculation of the flow for time intervals shorter than hours. Instead, the increase in weight is used as a measure of flow rate; the calculated flow rate is used to trigger flow-proportional sampling and to calculate the opening time of the moveable funnel, so that for a given rate of flow the sampling bottles (250 ml) are filled approximately to 0.75 of their volume. After sampling, the conveyor belt moves the bottle away and places a new bottle underneath the fixed funnel. The number of bottles on the conveyor belt is 149 on a 1×1 m table, (Figure 8). The sampler can be operated using both 220 Volt AC and 12 Volts DC, so that it can be used at remote locations with no access to mains power. Further automation was employed to monitor the erosion; sampling triggered a video camera overlooking the plot, so that the initiation of rills and melting of snow cover could be observed. To give enough slope to avoid sedimentation in the system, the vertical interval for the installation needed to be approximately 1 m. Further details of this installation are presented in UNESCO Technical Documents in Hydrology No. 60 (cf. Walling and Summer, 2002).

In many installations, the collecting tanks have to be placed in pits or cellars up to 3 m below the intake gutter, at the downslope end of the plot. In case of the presence

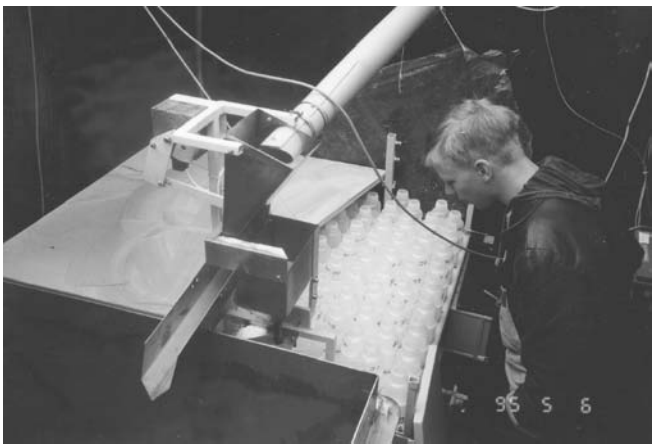


Figure 8 Sediment sampler with collecting trough and conveyor belt. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of shallow groundwater, the construction of cellars may seriously disturb the natural groundwater flow and prevent the formation of saturated overland flow that can occur frequently in areas experiencing high rainfall of moderate intensities. This is believed to be the explanation of cases of very low erosion rates recorded on some Danish plots. To overcome this problem, a watertight steel cellar was built in an area where the groundwater during winter time was only 50–100 cm below the surface. To prevent the cellar from being lifted by buoyancies of up to 9 tonnes, the bottom of the cellar was extended with flanges, so that the weight of soil on these flanges was able to keep the cellar in position.

U.S.D.A. Agricultural Handbook No. 224 provides full details of the equipment, and manufacturing instructions for divisors and the operation of erosion plots. The construction, maintenance, and running of such plot installations is very expensive, plots are therefore mainly employed at permanent research stations to study factors affecting erosion, for example, tillage practices and crop types, because the conditions on each plot can be controlled. Usually, experiments are carried out to assess the influence of one or more factors on the erosion. Assuming the aim is to study the effect of slope steepness; all other factors in the experimental design are kept constant, while the erosion from a range of slopes of different steepness is measured. This is kept within the range expected in the application area. The actual magnitude of the factors to be held constant also has to be decided, for example, slope length and soil type. Because measurements are subject to error, no single measurement of soil loss from a plot can be considered as the absolutely correct value. Errors can be assessed in respect of variability. Several replicate experiments are required to determine mean values and coefficients of variation. From a review of field and laboratory experiments, Luk (1981) found values from 10 to 40% for the coefficient of variation, while extreme values of soil loss ranged from 40–120%. If more variables are involved, the number of replicates needed increase dramatically. Several plots should be investigated simultaneously, with the treatment and factors of the individual plots randomized to fulfill statistical demands. A thorough discussion of experimental design is presented in Morgan (1995).

The measurement of natural erosion or erosion on plots under natural climatic conditions can be tedious, and it can be nearly impossible to obtain homogenous experimental conditions. Therefore, many experiments on plots are carried out using artificial rainfall Figure 5 or runoff. The obvious benefit is that it is possible to choose the starting conditions and, in particular, the time and duration of the application of the rain and/or overland flow. The crucial point is if the simulator is able to produce the right characteristics of the rainfall, for example, drop-size distribution and fall velocities. Rainfall simulators are classified according to the drop-formers used, either tubing tips or nozzles. Reviews of simulators produced by different researchers have been published by Bubenzer (1979) and Hudson (1971). Recently, a rainfall simulator to cover different sizes of plots has been designed by Bazzofi *et al.* (1980).

LABORATORY EXPERIMENTS

It is tempting to carry out controlled, quantitative experiments in laboratories, however, it is very difficult, if not impossible, to construct a scaled-down version of field conditions, because scale equivalence cannot be maintained in soil particles and raindrops without changing their basic properties or behaviour. Laboratory measurements must therefore be treated as representing true-to-scale field simulations. To initiate erosion, artificial rainfall and/or runoff have to be applied. It is important that the simulator produces a correct drop-size distribution, fall velocities, and intensities. Most laboratories have insufficient heights for drops to reach terminal velocities during their fall, so that their kinetic energy is too low. This could be overcome using pressurized water, but that influences drop-size and intensity. Studies of splash erosion have been carried out by Høgh-Schmidt and Brogaard (1975) in elevator shafts where the terminal velocity could be obtained to study erosion from splash cups. To evaluate soil erodibility, a Kamphorst rainfall simulator can be placed on different types of soil. The fall height is too small, but the energy of the artificial rainfall is considered close to natural conditions (Kamphorst, 1987). Small installations can be useful to evaluate the relative effect of different soil conditions, but they are not able to deal with processes acting beyond their scale. Different experimental set-ups are found at individual research institutes. At Silsoe College, UK, the set-up comprises a metal soil tray, 1.0 m wide, 0.5 m deep, and

2.5 m long. A mesh, or a geotextile membrane, is placed below the soil layer to allow infiltration. Runoff can be supplied at the upstream end of the tray. Slope steepness can be adjusted and a rainfall simulator can be positioned over the tray. At the Experimental Geomorphology Laboratory, University of Leuven, Belgium, the flume is 0.4 m wide and 4 m long. It is fed by water at its upper end and slope can be adjusted. The smaller the installation, the more serious are the edge effects. However, these can partly be reduced by using only the central parts of the plot for measurement. At the Centre for Geomorphology in Caen, France, a flume with a length of 20.4 m, width 1.4 m, and a depth of 0.7 m has been constructed for studies of rill erosion. Another long flume has been constructed at the Soil Erosion Laboratory, University of Toronto, Canada; the flume consists of 10 flexible modules, each 2.45 m long, 0.85 m wide, and 0.31 m deep. The elements can be tilted to produce varying profile shapes. The rainfall simulator is placed 8 m above the flume (Bryan and Poesen, 1989).

Laboratory studies on wind erosion have mainly been carried out in wind tunnels. Tunnels can either be of the open-circuit type or the closed-circuit type. Scale problems occur, because it can be difficult to reproduce a wind profile similar to nature, as the height in the wind tunnel is too small. It is also difficult to control the humidity of the air, which is an important factor determining the critical wind velocity. Principles of wind tunnel design are described by Pope and Harper (1966).

In laboratory experiments, the natural packing of soil, soil moisture level and groundwater are difficult factors to reproduce. The installation of vegetation in laboratory erosion experiments has also created difficulties. Therefore, a tendency in recent years has been a combination of laboratory experiments and rainfall simulators on plots with natural vegetation and soil conditions.

FUTURE MONITORING OF SOIL EROSION

Experiments to elucidate erosion processes will continue to develop improved algorithms for erosion models that, over time, may replace the costly hardware experimental set-ups. In particular, there is a need for research on erosion related to frozen soil and winter conditions.

In cases where experiments are carried out using experimental facilities, the need for automation is less, because the experiments using rainfall simulators can allow the experiments to be carried out within normal working hours. Automation, however, will be needed on remote locations, for example, after cultivation of virgin land with unknown soil and microclimate. The principles behind the sampler mentioned above (Walling and Summer, 2002), may be useful in construction of a commercial sampler that can cope with the large range of particle sizes observed during erosion events.

The development of sensors that do not disturb the flow, for example, arranged around the tubes transporting water and sediment from the erosion plot, is an important requirement for accurate studies of processes.

New, fast image analysis programs combined with digital video cameras may provide new methods for monitoring erosion by quantifying the so far subjective visual observations carried out by human observers.

REFERENCES

Bagnold R.A. (1941) *The Physics of Blown Sand and Desert Dunes*, Chapman & Hall: London.

Bazzofi P., Torri D. and Zanchi C. (1980) Nota 1: simulatore di pioggia. *Anali Institutio Sperimentale per lo Studio e la Difesa del Suelo*, **11**, 129–140.

Bernard C. (Eds.) (1990) *Landslides*, Vol. 3, Balkema: Rotterdam.

Borsy Z. (1972) Studies on wind erosion in the windblown sand areas of Hungary. *Acta Geographica Debrecina*, **10**, 123–132.

Brune G.M. (1953) Trap efficiency of reservoirs. *Transaction of the American Geophysical Union*, **34**(3), 407–418.

Bryan R.B. and Poesen J. (1989) Laboratory experiments on the influence of slope length on runoff, percolation and rill development. *Earth Surface Processes and Landforms*, **14**, 211–231.

Bubenzer G.D. (1979) Rainfall characteristics important for simulation. *Proceedings, Rainfall Simulator Workshop*, Tucson, Arizona USDA-SEA Agricultural Reviews and Manuals ARM-W-10: 22–34.

Ergenzinger P.K., Schmidt K-H and Busskamp R. (1989) The pebble transmitter system (PETS): a technique for studying coarse material erosion, transport and deposition. *Zeitschrift fur Geomorphologie N.F.*, **33/4**, 503–508.

Gerlach T. (1966) *Współczesby Rozwój Stoków w Dorzeczu Górnego Grajcaka*, Prace Geograf. IG PAN 52, french summary.

Gurnell A. and Clark M.J. (1987) *Glacio-Fluvial Sediment Transfer, An Alpine Perspective*, Wiley: Chichester.

Hasholt B. (1991) *Influence of Erosion on the Transport of Suspended Sediment and Phosphorus*, International Association of Scientific Hydrology Publication No. 203: 329–338.

Hasholt B. (1992) Sediment transport in a proglacial valley, sermilik, east greenland. *Geografisk Tidsskrift*, **75**, 30–39.

Hasholt B. (1998) *Assessment of Erosion and Some Implications for Model Validation*, International Association of Scientific Hydrology Publication. No. 249: 249–260.

Hershey R.W. (1999) *Hydrometry: Principles and Practices*, Wiley: Chichester.

Høgh-Schmidt K. and Brogaard S. (1975) The energy of raindrops. *Geografisk Tidsskrift*, **75**, 24–29.

Hubbel D.W. (1964) *Apparatus and Techniques for Measuring Bed Load*, U.S. Geological Survey Water Supply Paper 1748, Washington.

Hudson N. (1971) *Soil Conservation*, Batsford: London.

International Standard ISO 4363–1977 Methods for measurements of suspended sediment. Switzerland.

Kamphorst A. (1987) A small rainfall simulator for the determination of soil erodibility. *Netherlands Journal of Agricultural Sciences*, **35**, 407–415.

Kuhnle R.A. and Derrow R.W. II (1994) Using the sedbed monitor to measure bed load. *Proceedings of Fundamentals and Advancements in Hydraulic Measurements and Experimentation*, ASCE: 129–158.

Lawler D.M. (1992) Design and installation of a novel automatic erosion monitoring system. *Earth Surface Processes and Landforms*, **17**, 455–463.

Loughran R.J. (1989) The measurement of soil erosion. *Progress in Physical Geography*, **13**(2), 216–233.

Luk S.H. (1981) Variability of rainwash erosion within small sample areas. *Proceedings, Twelfth Binghampton Geomorphology Symposium*, Allen and Unwin: London, 243–268.

Mollenhauer K. (Ed.) (1995) *Bodenerosion Durch Wasser - Kartieranleitung Zur Erfassung Aktueller Erosionsformen*, Deutscher Verband für Wasserwirtschaft und Kulturbau Merkblätter zur Wasserwirtschaft 239/1996: ISBN 3-89554-045-5.

Morgan R.P.C. (1995) *Soil Erosion & Conservation*, Longman: Essex.

Morris G.L. and Fan J. (Eds.) (1998) *Reservoir Sedimentation Handbook*, McGraw-Hill: New York.

Nickling W.G. and McKennea Neuman C. (1997) Wind tunnel evaluation of a wedge-shaped aeolean sediment trap. *Geomorphology*, **18**, 333–345.

Østrem G. (1975) *Sediment Transport in Glacial Meltwater Streams*, Society of Economic Paleontologists and Mineralogists Special Publication No. 23: 101–122.

Poesen J. and Torri D. (1988) The effect of cup size on splash detachment and transport measurements. Part I: field measurements. *Catena Supplement*, **12**, 113–126.

Pope A. and Harper J.J. (1966) *Low-Speed Wind Tunnel Testing*, Wiley: Chichester.

Ritchie J.C. and McHenry J.R. (1975) Fallout Cs-137: a tool in conservation research. *Journal of Soil and Water Conservation*, **30**, 283–286.

Selby M.J. (1993) *Hillslope Materials and Process*, *Second Edition*, Oxford University Press: Oxford.

Spaan W.P. and van den Abeele G.D. (1991) Wind borne particle measurements with acoustic sensors. *Soil Technology*, **4**, 51–63.

Temple P.H. and Rapp A. (1972) Landslides in the mgeta area, western uluguru mountains, Tanzania. *Geografiska Annaler*, **54A**, 157–193.

Thorne C.R. (1998) *Stream Reconnaissance Handbook*, Wiley: Chichester.

United States Department of Agriculture (1979) *Field Manual for Research in Agricultural Hydrology*, USDA Agricultural Handbook 224: Washington.

Walling D.E. (1977) Assessing the accuracy of suspended rating curves for a small basin. *Water Resources Research*, **13**, 531–538.

Walling D.E. and Quine T.A. (1990) Use of caesium-137 to investigate patterns and rates of soil erosion on arable fields. In *Soil Erosion in Agricultural Land*, Boardman J., Foster I.D.L. and Dearing J.A. (Eds.), Wiley: Chichester, pp. 33–53.

Walling D.E. and Summer W. (Eds.) (2002) *IHP-VI Technical Documents in Hydrology*, Vol. 60, UNESCO: Vienna.

Zapata F. (Ed.) (2002) *Handbook for the Assessment of Soil Erosion and Sedimentation Using Environmental Radionuclides*, Klüver: Dortrecht.

82: Erosion Prediction and Modeling

MARK NEARING, KEN RENARD AND MARY NICHOLS

Southwest Watershed Research Center, Tucson, AZ, US

Models of soil erosion by water historically have been primarily either empirically-based or process-based (sometime referred to as physically-based). The first models of soil erosion were empirically-based. The prime example of the empirically-based model is the Universal Soil Loss Equation (USLE). More recent models have been based on equations that describe the physical, biological, and/or chemical processes that cause or affect soil erosion. It is important to understand that both the process-based and the empirically-based model possess a major empirical component, in the sense that the constitutive equations use parameters based on experimental data. This chapter details some of the important elements of soil erosion models and factors that influence their use.

INTRODUCTION

Soil erosion models play an important role both in meeting practical needs of soil conservation goals and in advancing the scientific understanding of soil erosion processes. They are used to help land managers choose practices to reduce erosion rates. Erosion prediction models are used for assessment and inventory work to track temporal changes in erosion rates over large areas. Erosion models are also used for engineering purposes, such as predicting rates of sediment loading to artificial reservoirs. Increasingly, governments are using erosion models and their results as a basis for regulating conservation programs. Models are used wherever the costs or time involved in making soil erosion measurements are prohibitive.

Erosion models play at least two important roles with respect to the science of soil erosion. First, erosion models are necessarily process integrators. Most often our knowledge of erosion mechanisms from experimental data is limited in scope and scale. Information may sometimes be misleading in terms of the overall effects on large integrated systems where many processes act interdependently. If individual processes that are well described from erosion experiments are correctly integrated via a process-based model, the result may be used to study model predictions and assess the behavior of the integrated system. Secondly, erosion models also help us to focus our research efforts. They help us to see where the gaps in knowledge exist and where to best focus our efforts to increase our overall erosion prediction capabilities.

In selecting or designing an erosion model, a decision must be made as to whether the model is to be used for on-site, off-site concerns, or both. On-site concerns are generally associated with degradation or thinning of the soil profile in the field, which may result in crop productivity loss. Conservationists refer to this process as *soil loss*, referring to the net loss of soil over only the portion of the field that experiences net loss over the long term. Areas of soil loss end where net deposition begins. Off-site concerns, on the other hand, are associated with the sediment that leaves the field, which we term here *sediment yield*. In this case, we are not necessarily concerned with the soil loss, or for that matter the amount of sediment deposited prior to leaving the field, although estimation of both of these may be used to estimate sediment yields. Ideally, a model will compute soil loss, deposition, and sediment yield, and thus have the capability to address both on-site and off-site issues.

Models of soil erosion by water historically have been primarily either empirically based or process-based (sometimes referred to as physically based). The first models of soil erosion were empirically based. The prime example of the empirically based model is the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1960, 1965, 1978). More recent models have been based on equations that describe the physical, biological, and/or chemical

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

processes that cause or affect soil erosion. It is important to understand that both the process-based and the empirically based models possess a major empirical component, in the sense that the constitutive equations use parameters based on experimental data.

CHOOSING AND USING AN APPROPRIATE MODEL

Choosing how to manage land, from the practical perspective, is often a matter of choosing between an array of potential options. Often, therefore, what we need to know is not necessarily the exact erosion rate for a particular option to a high level of accuracy, but rather we want to know how the various options stack up against one another. We may certainly be interested to have a general quantitative idea of the erosion rate, but for purposes of land management, this is not critical. Choosing which model to use then becomes a matter of (i) what type of information we would like to know, and (ii) what information (data) we have for the particular site of application. We may know, for example, that the USLE provides only estimates of average annual soil loss on the portion of the field that experiences a net loss of soil. If we have an interest in off-site impacts, then we probably want to choose a process-based model that will provide estimates of the sediment leaving the hillslope or watershed. If we have an interest in obtaining other auxiliary information about our choice of management strategy, such as soil moisture or crop yields, we might also decide to use a process-base that provides such information. On the other hand, if data are limited for the situation to be modeled, then a simple empirical model might be the only viable option.

Making broad-scale erosion surveys in order to understand the scope of the erosion problem over a region and to track changes in erosion over time can be done with many models. A statistical sampling scheme to take random points over the area of interest can be made, and the erosion model then applied to each point (USDA, 1996). In this case, we are not so concerned about the individual prediction for each point of application, but rather the ability of the model to predict overall averages of soil loss in a quantitatively accurate manner. While we know that none of the erosion models will necessarily predict erosion for a particular site to the quantitative level of accuracy we would like to see for survey assessment purposes (Nearing, 2000), many models do predict the averages for treatments quite effectively (Risse *et al.*, 1993; Rapp, 1994; Zhang *et al.*, 1996). The issues related to choosing the correct model are related to the information desired and the available data.

Conservation compliance, governmental policy making, and regulation of land user's actions follow the same guidelines as for applications of erosion assessment and conservation planning. The information desired and data availability are again the keys to choice of model. In this case, however, the argument is often given, most often by the farmer who is being regulated, that if we know that there are uncertainties in the erosion predictions for individual applications, how can we be sure that his field is being evaluated accurately. The answer is, of course, that we cannot be sure. If the model predicts that the farmer's field is eroding at a rate in excess of what our society's policy indicates to be acceptable, the model could well be wrong for this particular field. This problem is really no different than that faced by insurance companies as they set rates for insurance coverage. My home may be more secure from the possibility of fire than my neighbor's home because I am more careful than my neighbor. But unless my home falls in a different category (e.g. better smoke-alarm protection), I will not have much luck in going to my insurance company and asking for a lower payment rate. Likewise, if I am the farmer, I cannot expect to give a coherent argument for lower soil loss than the model predicts unless I conduct some practice, such as reduced tillage or buffers, which arguably reduces erosion.

Complexity and uncertainty are key issues relative to the development, understanding, and use of erosion models for conservation purposes. They are inevitable considerations because of the many complex interactions inherent in the erosional system as well as the enormous inherent variability in measured erosion data. These issues do not, however, prevent us from using models effectively for conservation planning. In fact, the scientific evidence indicates that choice of models, which implies choice of model complexity, is more a matter of the type of information desired and the quality and amount of data available for the specific application. If our goal is to know to a high level of accuracy the erosion rate on a particular area of ungauged land, we cannot rely upon the models. Natural variability is too great, and uncertainty in predictions is too high (see Nearing *et al.*, 1999; Nearing, 2000). For appropriate and common uses, such as those discussed above, models can be effective conservation tools.

THE UNIVERSAL SOIL LOSS EQUATION

Undoubtedly the prime example of an empirically based model is the USLE, which was developed in the United States during the 1950s and 1960s (Wischmeier and Smith, 1960, 1965). This equation has been adapted, modified, expanded, and used for conservation purposes throughout the world (e.g. Schwertmann *et al.*, 1990; Larionov, 1993).

The USLE was originally based on statistical analyses of more than 10 000 plot-years of data collected from natural runoff plots located at 49 erosion research stations in the United States, with data from additional runoff plots and experimental rainfall simulator studies incorporated into the final version published in 1978 (Wischmeier and Smith,

1978). The large database upon which the model is based is certainly the principal reason for its success as the most used erosion model in the world, but its simplicity of form is also important:

$$A = RKLSCP \quad (1)$$

where A ($t\,ha^{-1}\,yr^{-1}$) is average annual soil loss over the area of hillslope that experiences net loss, R ($MJ\,mm\,hr^{-1}\,ha^{-1}\,yr^{-1}$) is rainfall erosivity, K ($t\,hr\,MJ^{-1}\,mm^{-1}$) is soil erodibility, L (unitless ratio) is the slope length factor, S (unitless ratio) is the slope steepness factor, C (unitless ratio) is the cropping factor, and P (unitless ratio) is the conservation practices factor. Terminology is important here. Note first that the USLE predicts soil loss (see discussion above) and not sediment yield. Second, the word erosivity is used to denote the driving force in the erosion process (i.e. rainfall in this case) while the term erodibility is used to note the soil resistance term. These two terms are not interchangeable. Third, the model predicts average annual soil loss: it was not intended to predict soil loss for storms or for individual years. Conservationists often describe the predictions as long term, while from the geomorphic perspective the predictions would be referred to as medium term (Govers, 1996).

The units of the USLE appear a bit daunting as written (equation (1)), but become somewhat more clear with explanation. The units were originally written, and are still used in the United States, as Imperial, but conversion to metric is generally straightforward (Foster *et al.*, 1981a). The key to understanding the dimensional units lies with the definition of rainfall erosivity and the concept of the unit plot. Wischmeier (1959) found for the plot data that the erosive power of the rain was statistically best related to the total storm energy multiplied with the maximum 30-min storm intensity. Thus, we have the energy term (MJ) multiplied by the intensity term ($mm\,hr^{-1}$) in the units of R, both of which are calculated as totals per hectare and per year. The unit plot was defined as a standard of 9% slope, 22.13 m length (most of the early erosion plots were 1.83 m (6 ft) wide. A length of 22.13 m (72.6 ft) and a width of 1.83 m (6 ft) resulted in a total area of 1/100 of an acre. Prior to the days of calculators and computers, this was obviously a convenient value for computational purposes), and left fallow (cultivated for weed control). The K value was defined as A/R for the unit plot. In other words, erodibility was the soil loss per unit value of erosivity on the standard plot. The remaining terms, L, S, C, and P are ratios of soil loss for the experimental plot to that of the unit plot. For example, the C value for a particular cropped plot is the ratio of soil loss on the cropped plot to the value for the fallow plot, other factors held constant.

The USLE reduced a complex system to a quite simple one for purposes of erosion prediction. There are many complex interactions within the erosional system which are not, and cannot be, represented within the USLE. On the other hand, for the purposes of general conservation planning and assessment, the USLE has been, and still can be, used with success.

THE REVISED USLE: RUSLE1 AND RUSLE2

The USLE was upgraded to the Revised Universal Soil Loss Equation (RUSLE1) during the 1990s (Renard *et al.*, 1997) and evolved to the current RUSLE1.06c released in mid 2003 (USDA-ARS-NSL, 2003). RUSLE1 is land-use independent and applies to any land use having exposed mineral soil and Hortonian overland flow. RUSLE2 was also released in mid 2003, and it too is land-use independent (USDA-ARS-NSL, 2003).

RUSLE1 and RUSLE2 are hybrid models that combine index and process-based equations. RUSLE2 expands on the hybrid model structure and uses a different mathematical integration than does the USLE and RUSLE1. Both are computer based, and have routines for calculating time-variable soil erodibility, plant growth, residue management, residue decomposition, and soil surface roughness as a function of physical and biological processes.

RUSLE1 and RUSLE2 compute deposition on concave slopes, for vegetative strips, in terrace channels, and in sediment basins using process-based equations for transport capacity and deposition. RUSLE2 also computes an enrichment ratio for the sediment leaving the end of the slope. The model uses 10 yr–24 hr precipitation amounts to compute runoff. Enrichment ratio is the ratio of specific surface area of the sediment to specific surface area of the soil subject to erosion, and is useful for assessments of chemical carrying capacity of sediment.

RUSLE1 and 2 both have new relationships for L and S factors which include ratios of rill and interrill erosion, and additional P factors for rangelands and subsurface drainage, among other improvements. Within the United States, RUSLE2 uses updated values for rainfall erosivity (R), based on weather data collected from 1960 to 1999. RUSLE1 and 2 have the advantage of being based on the same extensive database as is the USLE, with some of the advantages of process-based computations for time-varying environmental effects on the erosional system. They still have the limitations, however, in model structures that allow only for limited interactions and interrelationships between the basic multiplicative factors of the USLE (equation (1)).

REPRESENTATIONS OF RILL AND INTERRILL PROCESSES

Differentiating between rill and interrill erosional areas is a useful, if somewhat arbitrary, division between dominant processes of erosion on a hillslope surface. In the

original description of the processes, Meyer *et al.* (1975) differentiated between areas of the hillslope dominated by shallow sheet flow and raindrop impact and those of small concentrated flow channels, which they termed rills. The concept is useful in terms of mathematical descriptions of erosion, and serves as a basis for many process-based erosion simulation models. The concept is also useful in terms of focusing experimental research on the two primary sources of eroded soil. The separation of the two primary sediment sources facilitates the mathematical modeling of nonpoint source pollutants in surface runoff. However, the concept is somewhat arbitrary because it implies a clear delineation between dominant processes on a given area, where, in fact, overlap occurs. Flow depths on a hillslope would be more correctly described in terms of frequency distributions of depth where processes tend more towards rill or interrill depending on the flow depth (Lewis *et al.*, 1990). Nevertheless, the introduction of the concept of rill versus interrill sediment source areas is the cornerstone of current erosion research and development of process-based erosion prediction technology. It is this subdivision of the erosion process which opened the "black box" which was employed earlier by empirically based erosion models.

Rills are conceived as being the primary mechanism of sediment transport in the downslope direction. Depths of flow in rills is considered to be relatively large (normally on the order of cm) compared to average broad-sheet flow depths (on the order of mm). Detachment of soil in rills is primarily by scour, whereas the principal mechanism of detachment in interrill areas is by raindrop splash. Models of rill and interrill erosion generally treat interrill areas as being sediment feeds for rills. The rills then act to transport the sediment generated in the interrill areas, as well as the soil detached by scour in the rills, down the slope.

PROCESS-BASED MODELS

Various process-based erosion models have been developed in the last 10 years including EUROSEM in Europe (Morgan *et al.*, 1998), the GUEST model in Australia (Misra and Rose, 1996), and the WEPP model in the United States (Flanagan and Nearing, 1995). We will focus here on the example of the WEPP model, largely because it is the technology most familiar to the author.

The Sediment Continuity Equation

Process-based (also termed physically based) erosion models attempt to address soil erosion on a relatively fundamental level using mass balance differential equations for describing sediment continuity on a land surface. The fundamental equation for mass balance of sediment in a single direction on a hillslope profile is given as:

$$\frac{\partial(cq)}{\partial x} + \frac{\partial(ch)}{\partial t} + S = 0 \qquad (2)$$

where c (kg m^{-3}) is sediment concentration, q (m^2 s^{-1}) is unit discharge of runoff, h (m) is depth of flow, x (m) is distance in the direction of flow, t (s) is time, and S [kg (m^{-2} s^{-1})] is the source/sink term for sediment generation. Equation (2) is exact. It is the starting point for development of physically based models. The differences in various erosion models are primarily: (i) whether the partial differential with respect to time is included, and (ii) differing representations of the source/sink term, S. If the partial differential term with respect to time is dropped then the equation is solved for the steady state, whereas the representation of the full partial equation represents a fully dynamic model. The source/sink term for sediment, S, is generally the greatest source of differences in soil erosion models. It is this term which may contain elements for soil detachment, transport-capacity terms, and sediment deposition functions. It is through the source/sink term of the equation that empirical relationships and parameters are introduced.

The sediment continuity equation in physically based models is normally written in terms of a single flow direction, x. The equation could be written and solved for the x and y directions to describe sediment continuity on a two-dimensional surface. To date, however, the approach taken to describe sediment continuity on two-dimensional surfaces has been to use the unidirectional equation with the x-direction being the direction of water flow at a given point on the landscape surface. Modeling of erosion on watersheds in current process-based erosion models generally involves dividing the watershed area into overland flow elements and channel elements. The overland flow elements are typically either rectangular, representing hillslopes adjacent to channel elements, or they are squares within a pattern of a grid which overlays the watershed. In both cases rill and interrill erosion processes are described in the overland flow elements, and sediment generated from those overland flow elements is considered to be delivered to the channel elements to be transported through the channel network. In some cases, sediment from an overland flow element may be routed to and potentially through another overland flow element before reaching a channel element.

In this chapter, we will address the application of the sediment continuity equation to the routing of sediment within overland flow elements. In doing so, we will focus on two of the many existing models to exemplify the concepts introduced. The Water Erosion Prediction Project (WEPP) Hillslope Profile Model (Lane and Nearing, 1989; Nearing *et al.*, 1989; Flanagan and Nearing, 1995) derives from a family of models developed by Foster (see Foster,

1982), and shares common descriptions of erosion with CREAMS (Foster *et al.*, 1981b). WEPP is a steady-state model which is intended to be used at the field planning level in much the same way as the USLE is currently used for conservation planning. As such, the model places a strong emphasis on the effects of soil and plant management practices on erosion. It is a continuous simulation model which operates on a daily time-step. The Hairsine and Rose model (Hairsine, 1988; Hairsine and Rose, 1992a, b) is a single-event dynamic model. This model derives from previous models discussed by Rose *et al.* (1983) and Rose (1985).

Forms of the Sediment Continuity Equation

For the case of steady state conditions, and using the concepts of rill and interrill erosion, the sediment continuity equation (2) may be rewritten as it is in the WEPP model as

$$\frac{\mathrm{d}G}{\mathrm{dx}} = D_r + D_i \tag{3}$$

where G [kg $(\mathrm{m}^{-1}\,\mathrm{s}^{-1})$] is sediment load per unit width in the flow (equal to cq in equation 2), D_r [kg $(\mathrm{m}^{-2}\,\mathrm{s}^{-1})$] is net rill erosion rate per unit area of rill bottom, and D_i [kg $(\mathrm{m}^{-2}\,\mathrm{s}^{-1})$] is interrill sediment delivery to the rill (as with rill erosion, expressed on a per unit rill area basis). For a given set of conditions, the interrill sediment delivery may be calculated and set as a constant in equation (3). For the case of net detachment in a rill, the D_r term will be positive, indicating a net increase in sediment load with downslope distance. For the case of deposition, the D_r term is negative.

In the WEPP model the sediment continuity equation is applied within the rills, which are described hydraulically as small rectangular shaped channels. This approach contrasts with most other erosion models, such as CREAMS (Foster *et al.*, 1981b), KINEROS (Smith *et al.*, 1995), and the model of Rose *et al.* (1983), which use uniform flow hydraulics to describe detachment of soil and transport of sediment by flowing water. The recent model of Hairsine and Rose (1992b), however, also uses rill hydraulics for describing rill erosion processes.

In formulating equation (3) from equation (2), several major assumptions and decisions regarding the representation of erosion have been made. In dropping the dynamic term, one must be able to establish a representative steady-state erosion rate and erosion time period which will provide a good estimate of the overall erosion rate for a storm. It has also been decided in formulating equation (3) that the rill and interrill formulation is appropriate and will provide a reliable framework for making erosion predictions. The fact that D_r and D_i represent "net", rather than instantaneous terms, is also important. For the interrill case, D_i is an estimate of the amount of sediment delivered to the rill from interrill areas. It does not explicitly account for the individual processes of splash detachment, deposition of splashed materials on interrill areas, and transport of the splashed materials in the shallow interrill flow. For the rill case, D_r represents a net movement of soil to the flow from the bed. This implies physically that detached sediment, once in the flow of the rill, will be transported downslope in the rill flow until an area of net deposition is reached, whereby the sediment may fall out and rest on the bed. The net rill detachment rate, D_r, is a function of four primary factors: (i) the amount of sediment in the flow, (ii) the hydrodynamics of the flow, (iii) the resistance of the soil to detachment by flow, and (iv) ground surface cover. The mathematical representation of each of these factors will be addressed below.

The Hairsine and Rose erosion model describes erosion as a balance of several instantaneous processes, rather than net detachment or deposition in rill or interrill areas. In that model, net detachment or deposition rate is conceived as a balance between several processes which occur simultaneously, those being: (i) the movement of sediment particles which are in the flow to the bed, (ii) the movement of previously detached sediment into the flow, and (iii) the detachment of soil particles from the bulk soil mass. The model assigns a separate term for each of these individual processes. The movement of sediment particles which are in the flow to the bed is "deposition", the movement of previously detached sediment into the flow is "reentrainment", and the detachment of soil particles from the bulk soil mass is "entrainment". Hairsine and Rose introduced entrainment and reentrainment terms for both rill and interrill erosion. Hairsine and Rose's model (from Hairsine, 1988) uses a sediment continuity equation of the form

$$\partial \frac{(c_i q)}{\partial x + \partial} \frac{(c_i h)}{\partial t} = e_i + e_{di} + r_i + r_{ri} - d_i \tag{4}$$

where e_i is entrainment by rainfall, e_{di} is reentrainment by rainfall, r_i is entrainment by surface water flow, r_{ri} is reentrainment by surface water flow, d_i is the continuous deposition term, and the subscript i indicates particle settling velocity class of the sediment. Net rates of detachment, deposition, as well as sediment transport capacity are implicit concepts embodied in this type of representation.

Because the concepts of transport capacity, T_c, and detachment rate capacity, D_{rc}, are not introduced *a priori*, Rose (1985) argues that the model of Rose *et al.* (1983) (which is a predecessor to Hairsine, 1988) is conceptually simpler than the model of Foster and Meyer (1972) (which is a predecessor to WEPP). On the other hand, as noted by Yu (2003), both of the models result in similar patterns of erosion behavior for similar conditions.

USING THE MODELS

The erosion portion of erosion prediction technology is actually only a small portion of the functioning whole. In order to be useful, a model must have the auxiliary components that describe the state of the system during the rainstorm that causes the erosion. The WEPP computer model includes seven major components, including climate, infiltration, water balance, plant growth and residue decomposition, surface runoff, erosion, and channel routing for watersheds. The climate component of the model, CLIGEN (Nicks, 1985), generates daily precipitation, daily maximum and minimum temperature, and daily solar radiation based on a statistical representation of weather data at a particular location. The climate model has been tested for erosion and well parameterized for the United States (Baffaut *et al.*, 1996). The infiltration component of the hillslope model is based on the Green and Ampt equation, as modified by Mein and Larson (1973), with the ponding time calculation for an unsteady rainfall (Chu, 1978). The water balance and percolation component of the profile model is based on the water balance component of the Simulator for Water Resources in Rural Basins (SWRRB) (Williams and Nicks, 1985; Arnold *et al.*, 1990), with some modifications for improving estimation of percolation and soil evaporation parameters. The plant growth component of the model simulates plant growth and residue decomposition for cropland and rangeland conditions. The residue and root decomposition model simulates decomposition of surface residue (both standing and flat), buried residue, and roots for the annual crops specified in the WEPP User Requirements (Flanagan and Livingston, 1995) plus perennial crops of alfalfa and grasses. Surface runoff is calculated using a kinematic wave equation. Flow is partitioned into broad-sheet flow for interrill erosion calculations and concentrated flow for rill erosion calculations. The erosion component of the model uses a steady-state sediment continuity equation that calculates net values of detachment or deposition rates along the hillslope (Nearing *et al.*, 1989). The erosion process is divided into rill and interrill components where the interrill areas act as sediment feeds to the rills, or small channel flows. The model is applicable to hillslopes and small watersheds.

Because the model is based on all of the processes described above, and more, it is possible with WEPP to have an enormous array of possible system interactions represented in the simulations. Just to name a few examples, slope length and steepness effects are functions of soil consolidation, surface sealing, ground residue cover, canopy cover, soil water content, crop type, and many other factors. Ground residue cover is a function of biomass production rates, tillage implement types, residue type, soil moisture, temperature and solar radiation, previous rainfall, and many other factors. Rill erosion rates are a function of soil surface roughness, ground cover, consolidation of the soil, soil physical and chemical properties, organic matter, roots, interrill erosion rates, slope, and runoff rates, among other factors. The lists continue *ad infinitum*. These are interactions which are simply not possible to represent with an empirical model. WEPP is a complex model in this sense.

The disadvantage of the process-based model is also the complexity of the model. Data requirements are huge, and with every new data element comes the opportunity to introduce uncertainty, as a first order error analysis would clearly indicate. Model structure interactions are also enormous in number, and with every structural interaction comes the opportunity for error, as well. In a sense, the goal in using the process-based model is to capture the advantages of the complexity of model interactions, while gaining the accuracy and dependability associated with the simpler empirically based model. This can be done, and was done with the WEPP model, using a combination of detailed sensitivity analyses and calibration of the model to the large database of natural runoff plot information used to develop the USLE and RUSLE. Without the tie between model and database, and without knowledge of the sensitive input variables so as to know where to focus efforts, turning a complex model such as WEPP into a useful conservation tool would not be possible. Thus in a sense, even though WEPP routines are process-based descriptors of various components of the erosional system, ultimately the model must be empirically based on the same type of data as was used to develop the USLE and RUSLE, along with additional experimental data collected specifically for WEPP.

REFERENCES

Arnold J.G., Williams J.R., Nicks A.D. and Sammons N.B. (1990) *SWRRB: A Basin Scale Simulation Model for Soil and Water Resource Management*, Texas A&M University Press: College Station, p. 142.

Baffaut C., Nearing M.A. and Nicks A.D. (1996) Impact of climate parameters on soil erosion using CLIGEN and WEPP. *Transaction of the American Society of Agricultural Engineering*, **39**, 447–457.

Chu S.T. (1978) Infiltration during an unsteady rain. *Water Resources Research*, **14**(3), 461–466.

Flanagan D.C. and Livingston S.J. (1995) *USDA-Water Erosion Prediction Project: WEPP User Summary*, NSERL Report No. 11, USDA-ARS National Soil Erosion Research Laboratory: West Lafayette.

Flanagan D.C. and Nearing M.A. (1995) *USDA-Water Erosion Prediction Project: Hillslope Profile and Watershed Model Documentation*, NSERL Report No. 10,USDA-ARS National Soil Erosion Research Laboratory: West Lafayette.

Foster G.R. (1982) Modeling the erosion process. In *Hydrologic Modeling of Small Watersheds*, Haan C.T. (Ed.) ASAE Monography No. 5: St. Joseph, pp. 297–380.

Foster G.R., Lane L.J., Nowlin J.D., Laflen J.M. and Young R.A. (1981b) Estimating erosion and sediment yield on field-sized areas. *Transaction of the American Society of Agricultural Engineering*, **24**(5), 1253–1262.

Foster G.R., McCool D.K., Renard K.G. and Moldenhauer W.C. (1981a) Conversion of the universal soil loss equation to SI metric units. *Journal of Soil and Water Conservation*, **36**, 355–359.

Foster G.R. and Meyer L.D. (1972) A closed-form soil erosion equation for upland areas. In *Sedimentation (Einstein)*, Shen H.W. (Ed.), Colorado State University: Ft. Collins.

Govers G. (1996) *Soil Erosion Process Research: A State of the Art*, Academie voor Wetenschappen, Letteren en Schone Kunsten van Belgie. Klasse der Wetenschappen: Jaargang 58, Nr. 1.

Hairsine P.B. (1988) *A Physically-Based Model of the Erosion of Cohesive Soils*, Ph.D. Dissertation, Griffith University, Brisbane.

Hairsine P.B. and Rose C.W. (1992a) Modeling water erosion due to overland flow using physical principles: I. sheet flow. *Water Resources Research*, **28**(1), 237–243.

Hairsine P.B. and Rose C.W. (1992b) Modeling water erosion due to overland flow using physical principles: 2. Rill flow. *Water Resources Research*, **28**(1), 245–250.

Lane L.J. and Nearing M.A. (Ed.) (1989) *USDA-Water Erosion Prediction Project: Hillslope Profile Model Documentation*, NSERL Report No. 2, USDA-ARS National Soil Erosion Research Laboratory: West Lafayette.

Larionov G.A. (1993) *Erosion and Wind Blown Soil*, Moscow State University Press: Moscow, p. 200.

Lewis S.M., Barfield B.J. and Storm D.E. (1990) Probability distributions for rill density and flow. *Winter Meeting ASAE*, Chicago, Paper No. 90-2558.

Mein R.G. and Larson C.L. (1973) Modeling infiltration during a steady rain. *Water Resources Research*, **9**(2), 384–394.

Meyer L.D., Foster G.R. and Romkens M.J.M. (1975) Source of soil eroded by water from upland slopes. *Present and Prospective Technology for Predicting Sediment Yields and Sources*, ARS-S-40, USDA, Agricultural Research Service: pp. 177–189.

Misra R.K. and Rose C.W. (1996) Application and sensitivity analysis of process-based erosion model GUEST. *European Journal of Soil Science*, **47**, 593–604.

Morgan R.P.C., Quinton J.N., Smith R.E., Govers G., Poesen J.W.A., Auerswald K., Chisci G., Torri D. and Styczen M.E. (1998) The European Soil erosion Model (EUROSEM): a dynamic approach for predicting sediment transport from fields and small catchments. *Earth Surface Processes and Landforms*, **23**, 527–544.

Nearing M.A. (2000) Evaluating soil erosion models using measured plot data: accounting for variability in the data. *Earth Surface Processes and Landforms*, **25**, 1035–1043.

Nearing M.A., Foster G.R., Lane L.J. and Finkner S.C. (1989) A process- based soil erosion model for USDA-water erosion prediction technology. *Transaction of the American Society of Agricultural Engineering*, **32**(5), 1587–1593.

Nearing M.A., Govers G. and Norton L.D. (1999) Variability in soil erosion data from replicated plots. *Soil Science Society Of America Journal*, **63**(6), 1829–1835.

Nicks A.D. (1985) Generation of climate data. In *Proceedings of the Natural Resources Modeling Symposium*, DeCoursey D.G. (Ed.), Pingree Park, October 16-21, 1983, USDA-ARS: ARS-30.

Rapp J.F. (1994) *Error Assessment of the Revised Universal Soil Loss Equation Using Natural Runoff Plot Data*, M.S. Thesis, School of Renewable Natural Resources, University of Arizona, Tucson; Renard K.G., Foster G.R., Weesies G.A., McCool D.K. and Yoder D.C. (1997) *Predicting Soil Erosion by Water – a Guide to Conservation Planning with the Revised universal soil loss equation (RUSLE)*, *Agricultural Handbook No. 703*, US Government Printing Office: Washington.

Risse L.M., Nearing M.A., Nicks A.D. and Laflen J.M. (1993) Assessment of error in the universal soil loss equation. *Soil Science Society of America Journal*, **57**, 825–833.

Rose C.W. (1985) Developments in soil erosion and deposition models. *Advances in Soil Science*, Vol. 2, Springer-Verlag: New York.

Rose C.W., Williams J.R., Sander G.C. and Barry D.A. (1983) A mathematical model of soil erosion and deposition processes: I. theory for a plane land element. *Soil Science Society of America Journal*, **47**, 991–995.

Schwertmann U., Vogl W. and Kainz M. (1990) *Bodenerosion Durch Wasser*, Eugen Ulmer GmbH and Company: Stuttgart, p. 64.

Smith R.E., Goodrich D.C., Woolhiser D.A. and Unkrich C.A. (1995) KINEROS: a KINematic runoff and EROSion model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publication: Highlands Ranch.

U.S. Department of Agriculture, Agricultural Research Service, National Sediment Laboratory (USDA-ARS-NSL) (2003) *RUSLE1.06c and RUSLE2*, Internet site:**www.sedlab.olemiss.edu/rusle**; U.S. Department of Agriculture (1996) *Summary Report 1992 National Resource Inventory*, US Government Printing Office: Washington.

Williams J.R. and Nicks A.D. (1985) SWRRB, a simulator for water resources in rural basins: an overview. *Proceedings of the Natural Resources Modeling Symposium*, DeCoursey D.G. (Ed.), USDA-ARS ARS-30: Pingree Park, October 16-21, 1983, pp. 17–22.

Wischmeier W.H. (1959) A rainfall erosion index for a universal soil loss equation. *Soil Science Society of America Proceedings*, **23**, 322–326.

Wischmeier W.H. and Smith D.D. (1960) A universal soil-loss equation to guide conservation farm planning. *Transactions of the 7th International Congress Soil Science*, Brussels, Belgium 418–425.

Wischmeier W.H. and Smith D.D. (1965) *Predicting Rainfall Erosion Losses in the Eastern U.S. – A Guide to Conservation Planning*, *Agricultural Handbook No. 282*, US Government Printing Office: Washington.

Wischmeier W.H. and Smith D.D. (1978) *Predicting Rainfall Erosion Losses. A Guide to Conservation Planning*, *Agriculture Handbook No. 537*, USDA-SEA, US Government Printing Office: Washington, p. 58.

Yu B. (2003) A unified framework for water erosion and deposition equations. *Soil Science Society of America Journal*, **67**(1), 251–257.

Zhang X.C., Nearing M.A., Risse L.M. and McGregor K.C. (1996) Evaluation of runoff and soil loss predictions using natural runoff plot data. *Transaction of the American Society of Agricultural Engineering*, **39**(3), 855–863.

83: Suspended Sediment Transport – Flocculation and Particle Characteristics

IAN G DROPPO

National Water Research Institute, Environment Canada, Burlington, ON, Canada

In the history of sediment transport research within hydrological sciences, there has possibly been no greater paradigm shift than that which has been provided by the identification of flocculation as a dominant mechanism operating within, and mediating, cohesive suspended sediment transport. Flocculation (the process of aggregating smaller particles together to form larger particles) significantly modifies the hydrodynamic properties of the sediment by altering their effective size, shape, density, porosity, and composition. Such changes have a dramatic impact on the downward flux of sediments and as such influence the transportation and fate of suspended sediments. The traditional assumption of suspended sediment being inorganic individual grains for the quantification and modeling of sediment transport is likely to result in erroneous results and management decisions. Flocs settle faster than their individual constituent particles (but orders of magnitude slower than that predicted by Stokes' equation for a similar size), and, for large flocs, can have porosity close to 100% and a corresponding density close to that of water. Flocs are composed of significant living organic components (primarily bacteria) which produce an extracellular polymeric substance (fibrils) which glues particles together giving them greater strength than an electrochemically bound floc, and a pseudo plastic nature. Because of this active living component, flocs should be thought of as micro ecosystems (composed of a matrix of water, inorganic, and organic particles) with autonomous and interactive physical, chemical, and biological functions or behaviors operating within the floc matrix. A better understanding of how floc structure influences floc behavior will help to refine our knowledge and management of suspended sediment transport in a variety of aquatic ecosystems.

INTRODUCTION

In the transport of suspended sediment, particles, and their interactions with the terrestrial and aquatic environment, have both theoretical and practical relevance within hydrological sciences (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**, **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**, **Chapter 81, Erosion Monitoring, Volume 2**, **Chapter 82, Erosion Prediction and Modeling, Volume 2**, **Chapter 84, Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands, Volume 2**, **Chapter 88, Reservoir Sedimentation, Volume 2**, **Chapter 85, Sediment Yields and Sediment Budgets, Volume 2**, **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**, **Chapter 87, Sediment Yield Prediction and Modeling, Volume 2**, **Chapter 89, On the Worldwide Riverine Transport of Sediment – Associated Contaminants to the Ocean, Volume 2**, **Chapter 90, Lake Sediments as Records of Past Catchment Response, Volume 2**, **Chapter 111, Rainfall Excess Overland Flow, Volume 3** and **Chapter 140, Transport of Sediments, Volume 4**). Physical, chemical, and biological processes at play with or within particles exert an important controlling influence on sediment transport. Not only is the sediment transport itself dynamic due

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

to hydrologic, hydraulic, and sedimentological variables but so too are the particles themselves. The unstable nature of suspended cohesive sediment particles, relative to their surrounding environment, result in the majority of these fine-grained sediments (silts and clays) existing primarily as flocculated or aggregated particles (Bale and Morris, 1987; Asper, 1987; van Leussen, 1988; Walling and Moorehead, 1989; Burban *et al.*, 1990; Azetsu-Scott and Johnson, 1992; Droppo and Ongley, 1992, 1994; Phillips and Walling, 1995, 1999; Droppo *et al.*, 1997, 1998a, 2000). The flocculation process can be defined simply as the aggregation of smaller inorganic and organic particles by various complex physical, chemical, and biological means to form larger composite particles (Droppo, 2001). Flocculation has the effect of altering the downward flux of sediment during transport due to changes in the effective particle, size, shape, density, porosity, and composition. A flocculated particle is in continuous interaction with its aquatic surroundings; as the medium in which it is transported provides the floc with further building materials, energy, nutrients, and chemicals for biological change, chemical reactions, and morphological development.

Flocs are generally considered to form within the water column or on the surface of the bed (Droppo, 2001), while aggregates are generally considered to form outside of, and be transported to, the aquatic system as water stable soil aggregates (Wall *et al.*, 1978). The differentiation of these two sediment particle forms is difficult due to their similar optical properties (particularly when the two flocculate together to form hybrid particles) (Droppo and Stone, 1994; Droppo and Ongley, 1994), and as such, this contribution uses the terminology floc/flocculation and aggregate/aggregation interchangeably. Given the complex composition of flocs and the complicated physical, chemical, and biological interactions within flocs and within the systems where they occur, Droppo *et al.* (1997) define a flocculated particle as "a microecosystem (composed of a matrix of water, inorganic, and organic particles) with autonomous and interactive physical, chemical, and biological functions or behaviors operating within the floc matrix". That is to say, these are not static physical entities transported within the water column, but rather, are dynamic particles with a significant "living" component mediating the physical, chemical, and biological behavior of the sediment and the water column as well.

While the awareness of flocculation as an important factor influencing sediment transport is increasing, the lack of a standard method for the assessment of floc structure and size, in conjunction with no standard equation to characterize the interaction and formation of flocculated particles, means that a poor understanding of how floc structure influences sediment transport prevails. This contribution provides a description of particle characteristics (composition, size, shape, porosity, density) and how these play a deterministic role in sediment transport behavior. The contribution represents an overview of research carried out on the structure and behavior of cohesive fine-grained sediments in river and lake systems. The contribution does not attempt to characterize a specific site's sediment characteristics, but rather, focuses on the general characteristics of sediment and its influence on behavior.

Floc Structural and Behavioral Issues

The process of flocculation increases the effective particle size by orders of magnitude over the absolute particle sizes and, as such, also changes the effective particle shape, density, porosity, and composition of the characteristic particle (suspended or bed sediment) within a system (Gibbs, 1985; Li and Ganczarczyk, 1987; Nicholas and Walling, 1996; Droppo *et al.*, 1997, 1998b; Phillips and Walling, 1999; Droppo, 2001). As floc form is in a continuous state of change due to the complex interactions of the physical, chemical, and biological factors controlling flocculation (Droppo, 2001), floc settling velocity and, therefore, sediment transport, will also be controlled by these factors in conjunction with the turbulence of flow. As such, floc form (size, density, porosity, shape, composition) controls, to a large extent, floc behavior (transport/settling) and will be discussed below. All floc sizing, density, porosity, shape, and settling velocity data which follows were measured following the computer image analysis methods of Droppo *et al.* (1997).

Particle (Floc) Morphology and Size

Figure 1 illustrates the complex structure of cohesive sediment flocs at a gross scale (Figure 1b) and the complexity of the internal floc structure at a fine scale (Figure 1c). Figure 1(a) illustrates the primary particles which compose the floc sample of Figure 1(b) (imaged after whole floc sample was sonicated to break up flocs and diluted). It is evident that flocs are very different from their individual primary particle components as they are irregular, porous, heterogeneous, composite structures. Flocs are composed of an active biological component (primarily bacteria, although at times other organisms can be incorporated), a nonviable biological component (e.g. detritus, extracellular polymeric substances), inorganic particles (e.g. clay particles), and water held within or flowing through pores (Ongley *et al.*, 1981; Peart and Walling, 1982; Walling and Kane, 1984; Droppo and Ongley, 1992, 1994; Ongley *et al.*, 1992; Leppard, 1992; Woodward and Walling, 1992; Walling and Woodward, 1993; Liss *et al.*, 1996; Droppo *et al.*, 1997; Droppo, 2001). The arrows in Figure 1(c) point out examples of extracellular polymeric fibrils produced by bacteria; a very important part of the floc discussed below. The dark regions represent clay platelets. All of these components are

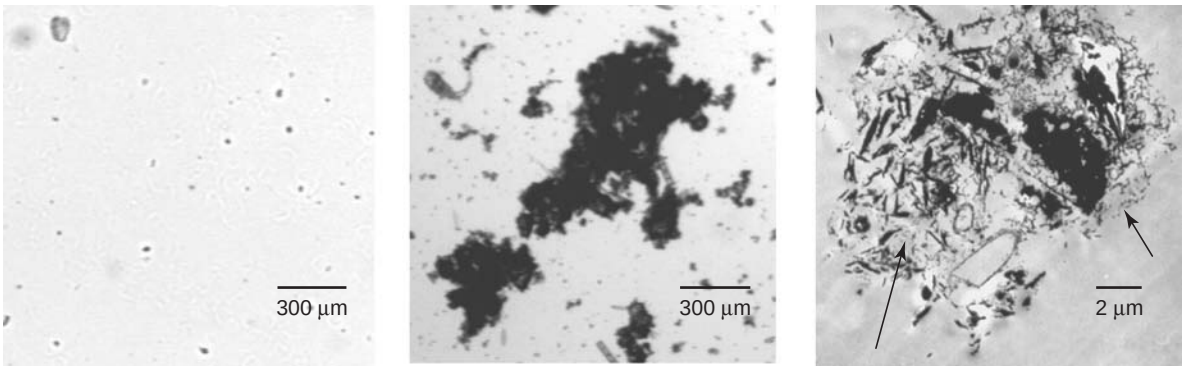


Figure 1 Micrographs illustrating the complexity of flocs from a river environment. (a) Primary particles (imaged with a conventional optical microscope), (b) natural floc particles (imaged with a conventional optical microscope), and (c) transmission electron micrograph of internal matrix of a floc (arrows denote examples of extracellular polymeric fibrils produced by bacteria with the matrix of the floc. Dense dark regions are clay particles)

linked together in the matrix by various physical, chemical, and biological mechanisms (Droppo, 2001). While electrochemical flocculation exists, it is believed that the biological bonding/bridging, primarily through extracellular polymeric substances (EPS) (Figure 1c), is the main component promoting floc building and stability (Liss *et al.*, 1996). EPS is manifested as micro fibrils (5–20 nm in diameter) or fibril bundles which has a "sticky" consistency (Leppard, 1992) often promoting flocs with pseudo plastic nature (Liss *et al.*, 1996). It is also this material which results in the development of micro pores within the floc (Figure 1c). These micro pores exhibit significant surface tension which results in bound water being an integral part of the overall floc structure. This has concomitant effects on the floc density and therefore sediment transport as described below.

Figure 2 provides a conceptual diagram which is specific to how floc structure can influence sediment transport. It is by no means a complete comprehensive evaluation of the process of flocculation (the process is too complex and still not fully understood to provide such a model), but rather provides the rationale for considering flocs to be micro ecosystems unto themselves which can actively mediate sediment transport through the dynamic process of flocculation. In Figure 2, the floc is broken down into its primary components (water, inorganic particles, organic particles, and pores) with substantive structural characteristics stated. The individual components influence the transport of the sediment by mediating factors such as particle size, density, water content, cohesiveness, and stability. Obviously each of these components operates interactively with one another and cannot, in reality, be isolated. It is the interaction of these components and the resultant structures which can modify a floc's transport behavior with concomitant effects on the overall transport of suspended sediment. From Figure 1 and 2, it is apparent that cohesive suspended sediment should be viewed as a collection of dynamic flocs and not simply as the traditionally viewed collection of discrete individual grains of sediment. For a comprehensive discussion on floc structure, the reader is referred to Liss *et al.* (1996) and Droppo (2001).

The dramatic influence of flocculation on *in situ* particle size is shown in the ternary plot of Figure 3. Particle distributions were classed by percent sand, silt, and clay size for both their natural (flocculated) and sonicated (primary) distributions. The floc distribution classes were primarily within the sand and silt size classes even though their constituent material was primarily clay size as illustrated by the primary particle data points on the plot. The difference between the primary particles and flocs is so extreme that the points lie on the edge of the ternary plot. Such a high degree of flocculation has been observed before for multiple river samples (Petticrew and Droppo, 2000).

The Effect of Floc Size on Sediment Transport

As flocculation significantly alters the effective size and structure of particles, it will also have an influence on settling velocity and, therefore, sediment transport. While it is accepted that flocs will settle at different rates than individual constituent particles (Hawley, 1982; Ozturgut and Lavelle, 1984; Gibbs, 1985; Li and Ganczarczyk, 1987; Burban *et al.*, 1990; Azetsu-Scott and Johnson, 1992; van Leussen and Cornelisse, 1993), the theories of particle/floc settling are still centered around the balance of forces of a single particle settling under gravity in a viscous fluid (Tambo and Watanabe, 1979; Li and Ganczarczyk, 1987; Fennessy *et al.*, 1994). The balance of forces is based on a gravitational force acting downwards, a buoyant force acting upwards and a drag force acting upwards. The resulting equation which balances these forces is given in

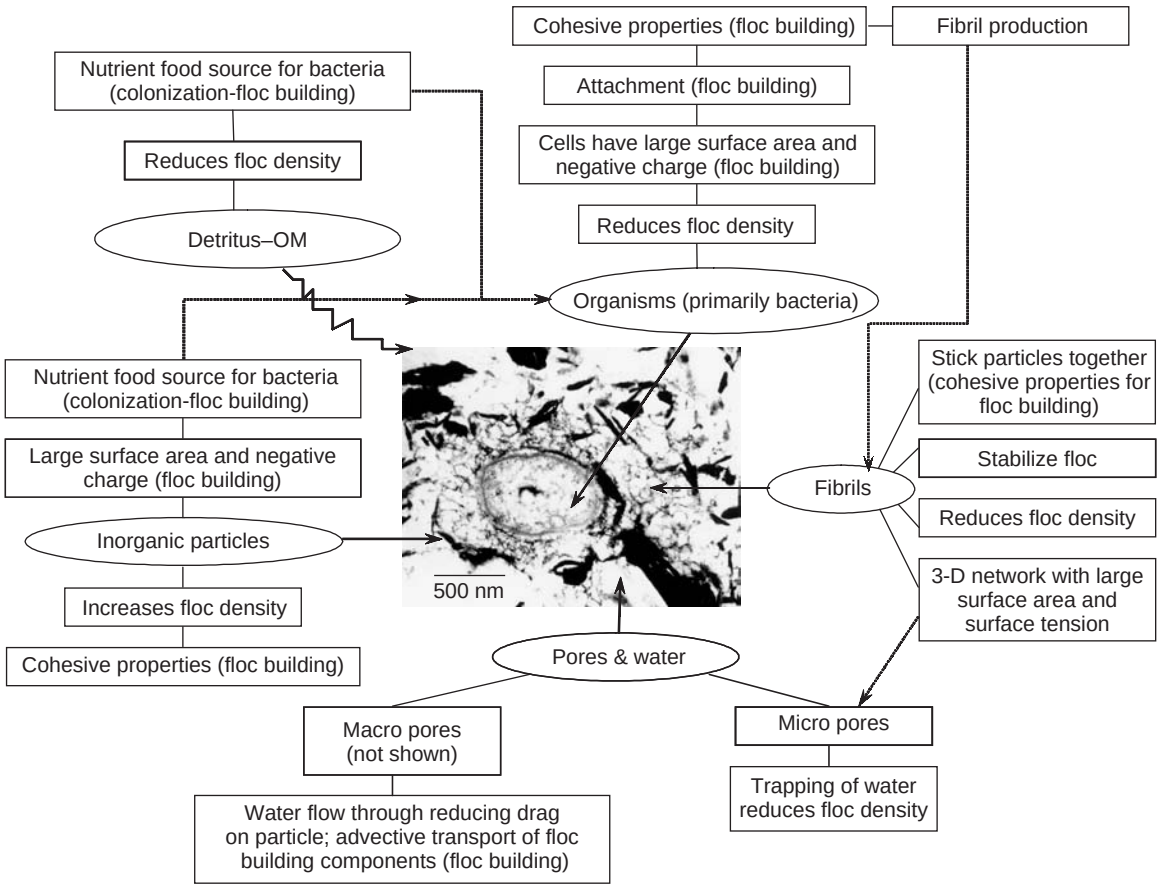


Figure 2 Conceptual diagram of the complex linkage of the floc components and their behavior/functions within the floc which influence suspended sediment transport. Jagged line implies not differentiable within micrograph but present

equation (1).

$$\frac{1}{6}\pi d^3(\rho_s - \rho_w)g = \frac{1}{2}\frac{24}{R_e}\rho_w\overline{\omega}^2\frac{\pi d^2}{4} \quad (1)$$

$$R_e = \frac{\overline{\omega}\rho_w d}{\mu} \quad (2)$$

Where: $\overline{\omega}$ = settling velocity
d = particle diameter
ρ_s = density of the solid particle
ρ_w = density of the water
μ = dynamic viscosity (kinematic viscosity $\times\rho_w$)
g = acceleration due to gravity
R_e = Reynolds Number (see equation 2)

On the left-hand side of equation (1) are the forces acting downwards on a particle [mass $1/6\pi d^3(\rho_s - \rho_w)$; and acceleration due to gravity (g)]. On the right-hand side of equation (1) are the forces acting upwards on a particle (the coefficient of drag ($Cd = 24/R_e$); shear effects ($Cd\rho_w\overline{\omega}^2$) and particle area ($\pi d^2/4$).

By solving equation (1) for $\overline{\omega}$, the Stokes' equation is derived (equation 3).

$$\overline{\omega} = \frac{1}{18}d^2(\rho_s - \rho_w)\frac{g}{\mu} \quad (3)$$

Stokes' Law is based on the settling of single impermeable spherical particles in a laminar region ($R_e < 0.2$) (Allen, 1990), and is therefore not ideal for the determination of floc settling velocity, density, or porosity due to a floc's general irregular shape, porous, and heterogeneous nature (Figures 1b and 1c). Settling in the laminar region may or may not be met, given differences in particle morphology and environmental conditions. In addition, Stokes' Law requires knowledge of the particle density which is generally assumed to be 2.65 g cm^{-3} (the density of quartz particles) for solid particles (ρ_s). However, the wet floc density (ρ_f) is known to be substantially less than this, often close to the density of water (Droppo, 2001). Density and porosity effects will be discussed below. Stokes' Law, or a modification thereof, is nevertheless often used to determine one or more of settling velocity, wet density, and porosity of individual flocs (Tambo and Watanabe, 1979; Li and

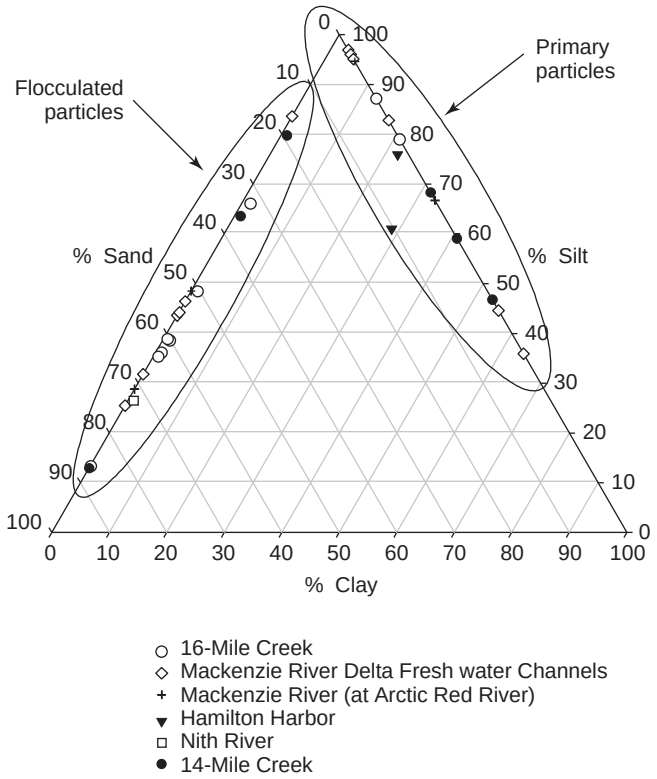


Figure 3 Ternary plot showing the relative proportions of sand, silt, and clay sizes for both original floc distributions and their primary distributions after undergoing sonication. The plot shows the significant influence that flocculation has on particle size

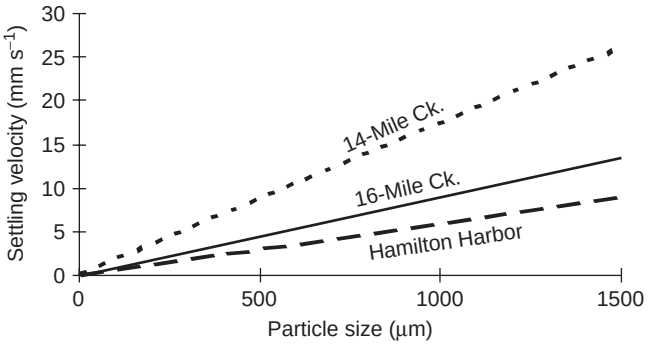


Figure 4 Regression lines for three floc samples. All regression lines are significant at $p = 0.05$. The plot shows that site-specific variables will influence floc settling velocity and transport

Ganczarczyk, 1987; Fennessy *et al.*, 1994), and does provide an indication of how these variables are related to floc size. Modifications of Stokes' Law, to improve floc settling velocity predictions, have generally relied on the use of shape factors in the calculation of drag coefficients as a correction for the influence of flocculation (Klemetson, 1985; Allen, 1990). Shape effects on floc settling behavior are also discussed below.

It has been well documented that as floc size increases so too does its settling velocity (Hawley, 1982; Li and Ganczarczyk, 1987; Burban *et al.*, 1990; Lick *et al.*, 1992; Eisma *et al.*, 1997; Hill *et al.*, 1998; Curran *et al.*, 2002). The relationship of floc size to settling is, however, very different from that predicted by the Stokes' equation. The Stokes' equation predicts that settling velocity will be proportional to the diameter of the particle squared (equation 3); however, multiple samples from multiple environments have shown that this relationship is in fact linear (Droppo *et al.*, 2002). This difference is owing to floc structure being dramatically different from the Stokes' assumed solid spherical particles (Figure 1b and 1c). Example linear regression lines are shown in Figure 4 for three different samples (two river and one lacustrine). As is characteristic from multiple samples, the lacustrine flocs settle slower for similar sizes, which is likely to be related to compositional and formational (shear) differences. Given the lower turbulence of lake environments, lacustrine flocs are often shown to be less dense and of a more open matrix than river flocs which were formed in a higher shear environment resulting in more compact flocs.

If Stokes' equation were used for flocs, it could overestimate settling velocity by orders of magnitude (e.g. for a 500-μm particle, Stokes predicts a settling velocity of 225 mm s^{-1} (assuming a density of 2.65 g cm^{-3}), while a measured 14-Mile Creek floc of similar size only has a settling velocity of 9 mm s^{-1} (actual density estimated to be approximately 1.1 g cm^{-3})). If such calculated settling estimates were used for sediment transport prediction, it could seriously underestimate the transport distance of particles and associated contaminants. On the other hand if only the primary particle size was used for prediction (as often they are), an overestimate of sediment transport would be predicted with concomitant management errors owing to the smaller size of the primary particles relative to the floc particles. The assumption of particles settling within the Stokes' region ($R_e < 0.2$) is also not always met for flocs as Droppo *et al.* (2002) found that the largest fluvial flocs often settle well out of this zone. Often these larger flocs represent the majority of sediment volume in transport (Droppo and Ongley, 1994).

There is consistently a high degree of variation for the relationship of floc size to settling velocity (e.g. r^2 values for the regression lines plotted in Figure 4 ranged from 0.4 to 0.6). This low explanation of variance is typical and is attributed to the morphological and compositional differences between flocs within a sample site. Flocs with similar measured diameters may have very different, shape, porosity, density, and composition and as such will settle with different rates. While it is evident that floc size is an important factor in influencing settling velocity, the other

characteristics of density, porosity, and shape (see next section) all interact resulting in a particle settling velocity which is well below that predicted by Stokes' equation.

The Effect of Floc Shape on Sediment Transport

Sediment transport will also be influenced by the shape of the particles. While cumulatively (from a sediment load perspective) this factor may not noticeably influence the overall flux of sediment in a river, it is important to understand its influence on individual particle behavior. The shape of a particle is known to influence settling velocity due to resistance effects against flow (fluid drag forces) (Richards, 1982; Li and Ganczarczyk, 1987). Particles which have the same density and diameter of an equivalent sphere but have a different shape exhibit different settling rates (Krumbein, 1942; Lerman, 1979; Richards, 1982; Ozturgut and Lavelle, 1984). The shape of a floc is influenced by its source, composition, and by the flow field in which it is transported and/or eroded in. While it is difficult to experimentally determine the effect of floc shape on settling velocity due to the difficulty in making or maintaining realistic flocs with realistic mass and density in the laboratory, it has been found that elongated flocs settle slower than spherical flocs in quiescent settling columns (Droppo *et al.*, 1998b). This relationship is likely to be related to elongated flocs preferentially settling with their long-axis parallel to the direction of settling (Figure 5), thus promoting more fluid drag on the particle and a slower settling velocity in theory. In later work, Droppo *et al.* (2002), however, found the opposite relationship which would suggest that floc shape may have a minimal influence on settling velocity. Plotting multiple data sets of aspect ratio to settling velocity showed no significant relationship. While it is likely that size is a much better predictor of settling velocity than shape, Droppo *et al.* (2002) found that generally a power function of the form $S_f = ad^m$ could be used to explain the relationship between shape factor and floc size (although the fit was generally poorly defined), where: S_f was the shape factor defined in equation (4), d was diameter and a and m are parameters which are dependent on the type of particle being investigated and the condition under which the flocs were formed (a and m are derived empirically). As a floc increases in size, generally it becomes more convoluted in shape (Droppo *et al.*, 2002) and more elongated (de Boer, 1997; de Boer *et al.*, 2000). In reality, for turbulent environments, flocs, if able to settle, will due so in a tumbling fashion. Such a motion in settling would further suggest that the influence of individual floc shape on settling velocity is limited and is likely of limited use for the prediction of sediment transport. For quiescent environments such as lakes and backwater areas of rivers, the settling and shape results above may be more characteristic.

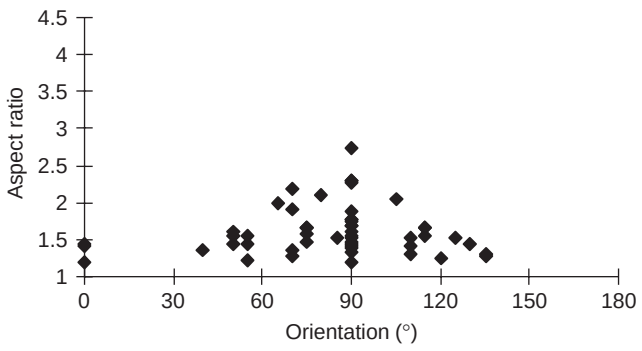


Figure 5 Example of the relationship between floc settling orientation and aspect ratio (floc length/floc width) for a lacustrine sediment (Hamilton Harbour). Note the bell shape of the distribution suggesting a predominance of long-axis settling

$$S_f = \frac{4\pi \times \text{Area}}{\text{Perimeter}^2} \quad (4)$$

($S_f = 1$ for a perfect circle, successively lower factors represent a more convoluted floc, and close to 0 = approaching a line).

The Effect of Floc Density and Porosity on Sediment Transport

While floc density will influence the downward flux of sediments (equation 1), no standard relationship exists between floc density and floc size or settling velocity. This is due to the complexities of the floc matrix and high variability within a natural system in terms of fluid shear and floc morphology. The density of a floc is influenced by its composition (inorganic and organic particles, EPS, water content, etc.) and its porosity (pore size and structure) and has typically been derived from equation (3) (ρ_f is substituted for ρ_s, where ρ_f is the wet density of the floc).

It has been shown constantly, regardless of the environment (e.g. marine, estuarine, wastewater, fluvial), that floc density decreases as floc size increases (Hawley, 1982; Gibbs, 1985; Andreadakis, 1993; Fennessy *et al.*, 1994; Droppo *et al.*, 2002). This is shown in Figure 6 for multiple river and lake samples. The variation between the lines is reflective of the different factors and relative importance of each, in influencing floc growth and structure between environments and sample times (Li and Ganczarczyk, 1989; 1990). Petticrew and Droppo (2000) and Droppo *et al.*, (1998b) found that flocs derived from the bed of two different rivers settled faster than those formed within the water column. This was attributed to density differences between the two floc populations with the suspended flocs being of a more open matrix, lower density structure with characteristic lower settling velocity as compared to the more dense bed-derived flocs.

Lagvankar and Gemmel (1968) suggest that for very large flocs there is less dependency of density on size since

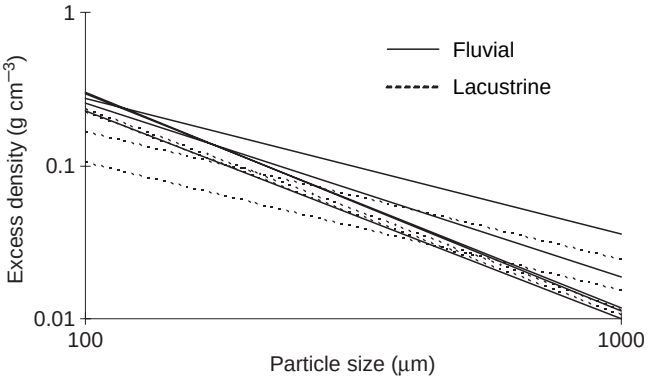


Figure 6 Examples of the inverse relationship of floc size to floc density for five river and four lake samples. Variations in slope are related to structural floc differences driven by site and time specific variables

the density becomes, in effect, constant for large flocs. This trend is shown in Figure 7 where for the larger flocs the particle density approaches the density of water. Given this relationship, and the fact that the range in floc densities is small (density typically ranges from 1.0 to 1.4 g cm^{-3} with the majority of flocs below 1.1 g cm^{-3}; Figures 6 and 7 and Riley, 1970; Krone, 1972; Shiozawa *et al.*, 1985) and close to the density of water (over a wide range of floc sizes), it is unlikely that density plays a strong role in particle settling and transport relative to the effect of floc size. This has been discussed in Droppo (2001) and Droppo *et al.* (2002).

The floc characteristics of size, shape, and density are highly influenced by floc porosity, which is developed through the heterogeneous arrangement of various floc constituents and dictated by such factors as the flow field (energy), biological colonization and metabolic production of sticky fibrils, and the trapping of additional inorganic or organic particles via surficial or internal physical (collision), chemical (electrochemical) or biological means. Once formed, the pores may provide an active micro environment which may support the uptake and/or transformation of various chemicals/nutrients. Pores may further influence electrochemical and/or diffusional gradients or advective-driven chemical/nutrient removal (Sherman, 1953; Li and Ganczarczyk, 1988; Logan and Hunt, 1987, 1988). All of these may in turn influence the hydrodynamics of a particle during transport through structural and density modifications (Droppo, 2001).

Porosity cannot be measured directly, due to the three-dimensional and tortuous nature of flocculated sediment and as such it is generally estimated using a mass balance of respective densities (equation 5). Porosity influences settling velocity indirectly by being one of the primary factors controlling density. As a floc grows, the amount of open void space (pores) increases simply due to the nature of the contact points between particles forming the matrix. With an increase in void space (porosity), there will be a proportional increase in water content. As such, the greater the porosity, the greater the water content, the lower the density, and, in theory, the slower its settling velocity (assuming constant floc size and no flow through the floc). Figure 7 shows that as a floc increases in size the porosity increase, approaching 100% which corresponds to the density of water. This relationship has been shown by numerous researchers (Tambo and Watanabe, 1979; Logan and Hunt, 1987; Li and Ganczarczyk, 1987, 1988).

$$\varepsilon = \frac{(\rho_s - \rho_f)}{(\rho_s - \rho_w)} \tag{5}$$

Where ε = floc porosity

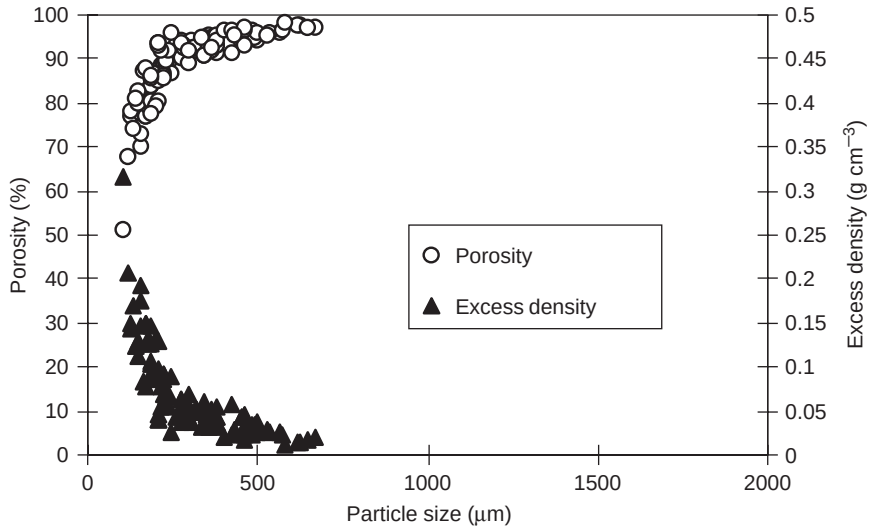


Figure 7 Example of the relationship between calculated floc porosity and density for a river suspended sediment sample

For the data in Figure 7, a value of $1.65\,g\,cm^{-3}$ was used to represent the dry density of the flocs (ρ_s) in equation (5) as the organic content of the flocs will reduce the density substantially below that of the traditionally used $2.65\,g\,cm^{-3}$ (Droppo *et al.*, 1997). This lower value is higher than those stated in the literature for microbial dominated flocs (Li and Ganczarczyk, 1987 (assumed $1.45\,g\,cm^{-3}$), Andreadakis, 1993 (assumed $1.34\,g\,cm^{-3}$)) given the mix of inorganic and organic particles in aquatic suspended sediments.

Porosity directly influences floc density and settling as individual large pores may potentially act as passages for the flow-through of water during settling. Such movement of water through the floc will, in theory, increase the settling velocity of the floc due to a reduction in drag (Zahid and Ganczarczyk, 1990). Thousands of low-resolution observations of freshwater flocs have revealed very few macro pores which may be capable of "flow-through". Multiple high resolution observations of floc pores (e.g. with electron microscopy) such as those in Figure 1(c) and 2 reveal the prevalence of EPS fibrils which may in fact fill much of the grossly observed pores with micro pores. These micro pores will have a high capacity for the retention of water with concomitant effects on density. As such the majority of water within freshwater flocs is believed to be bound and not free flowing. As porosity is strongly correlated to floc density, it is likely that porosity's influence on settling velocity is minimal relative to floc size (Droppo *et al.*, 2002)

The Effect of Floc Strength on Sediment Transport

Given that a floc is exposed to a variety of forces during its formation and transport, floc strength, or its ability to withstand shear, is a critical characteristic which will dictate the transport and fate of suspended sediment. The strength of a floc is dependent on its physical structure and composition and on the external forces operating on the floc. A discussion of the physical forces (stress relationships) acting on a particle is beyond the scope of this contribution. The reader is referred to Tambo and Hozumi (1979) and Partheniades (1986) for a discussion on this aspect. However, it is worth considering the potential scenarios or "life cycles" of a floc within a turbulent regime as related to floc strength.

As described above, flocculation influences the downward flux of sediment by principally increasing the particle size. While the floc may settle faster than individual particles (flow dependent), its ability to settle out of suspension at a given flow is dependent on its strength and its ability to withstand the higher shear stresses imparted onto the floc by the near bottom zone. As a floc approaches this near bottom zone, if the critical shear stress for floc break up is lower than the bed shear stress then the flocculated particle will break up and be resuspended potentially to go through the flocculation process again and attempted settling further down flow. If the floc enters a zone where its strength is greater than the bed shear stress, then it may settle out of suspension (Partheniades, 1986) forming a surficial fine-grained laminae (SFGL) (Droppo and Stone, 1994). This vertical and horizontal loop in floc transport has been called "floc recycling" by Droppo *et al.* (1998b).

Within river systems, this SFGL may in itself undergo a form of recycling as entrainment of the deposited flocculated material is dependent on the critical bed shear stress for erosion being surpassed. Once the flocs are resuspended, they then re-enter the floc recycling process only to at some point (shear dependent) redeposit in short-term storage as a new layer of SFGL or longer-term storage as over bank sediment or reservoir/lake/oceanic deposits. This linkage of floc and SFGL recycling is illustrated in Figure 8. Floc strength is therefore a critical characteristic which will dictate its transport history.

CONCLUSIONS

Suspended sediment particles are often implicitly or explicitly assumed to be single grain objects of inorganic composition. This is motivated by numerous constraints/biases (e.g. an individual's discipline, focus of research, laboratory/field capabilities and the scale of interest) as well as a lack of a standard method for sampling, observing, and measuring floc/aggregate particles. In addition, the lack of a standard defining equation, which can explain the formation and interaction of flocs/aggregates with their surrounding aquatic environment, has furthered this assumption. It is, however, well documented that cohesive suspended sediment is preferentially transported as flocculated material. Flocs are very different from their constituent primary particles as they are complex, heterogeneous, porous structures, composed of inorganic particles, organic particles (viable and nonviable), and water. Flocculation, because of its ability to alter the structure (size, shape, composition, density, and porosity) of suspended sediments has significant implications for the transport and fate of suspended sediment in aquatic systems. The study of floc characteristics and their influence on sediment transport is complicated by the dynamic nature of flocculation. Flocs are in a continual state of change due to the flow dynamics of the system and because of the complex interactions between the principle components of the floc (water, inorganic particles, organic particles, and pores). The significant "living" component of the floc (primarily bacteria) plays a strong role in mediating the structure, and therefore transport of flocs due to their secretion of extracellular polymeric fibrils (EPS). This material promotes floc building by its sticky nature, and provides the floc with a plastic/elastic nature which results in a relatively strong structure. In this review, the transport of sediment is analogous to particle

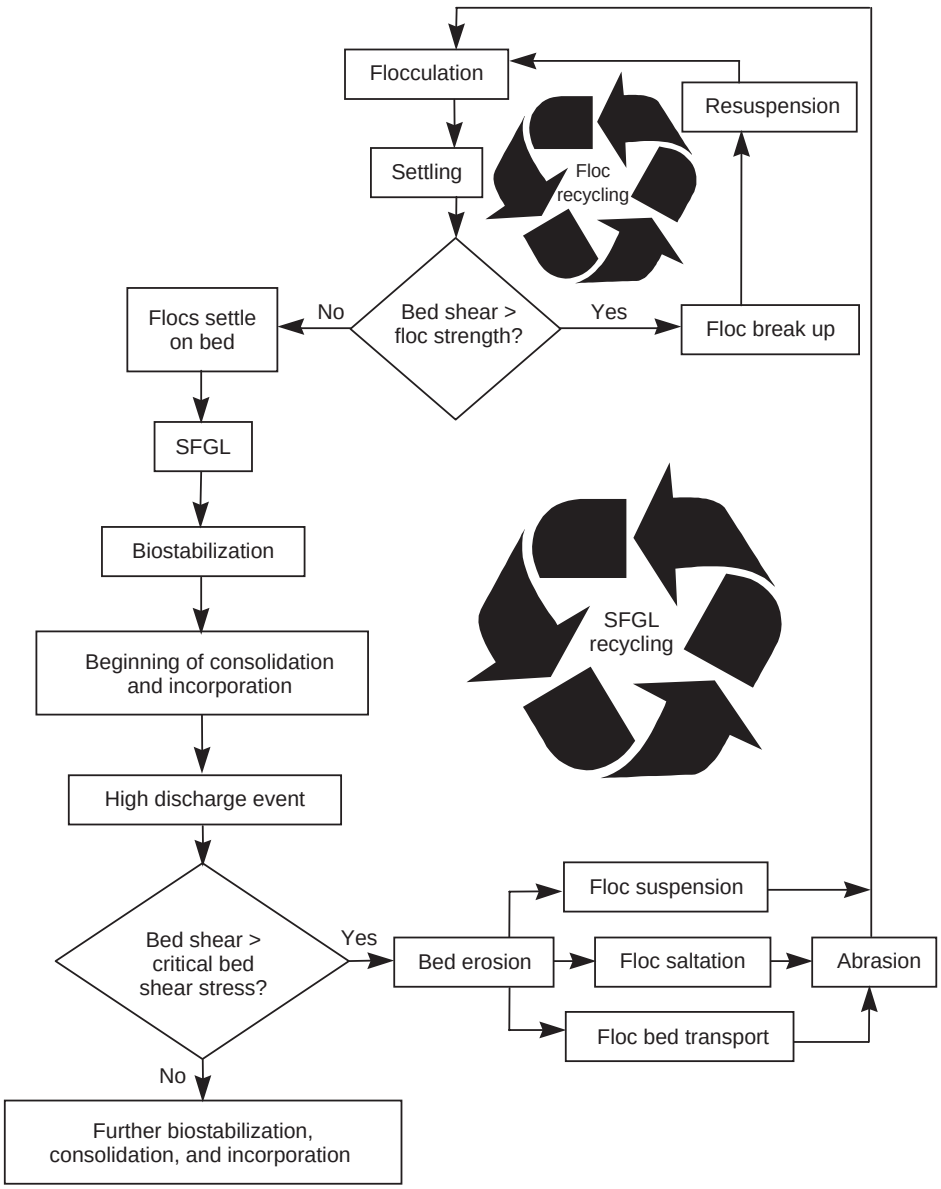


Figure 8 Conceptual model of SFGL and floc recycling loops linked for a river system (Reprinted from Droppo *et al.*, 2001. © 2001, with permission from Elsevier)

settling velocity. A positive linear relationship between floc settling velocity and size is standard. Generally, there is a high degree of variability in this relationship which is related to differences in floc structure (i.e shape, composition, density, porosity) over a narrow size range. Floc density decreases as floc size increases with densities of very large flocs (i.e. $>500\,\mu m$) approaching the density of water. Conversely, the porosity of a floc increases with floc size, approaching 100% (corresponding to the density of water). Floc size is the dominant factor influencing sediment transport relative to floc density, porosity, and shape. This is due to porosity being directly related to density and the range of floc densities being relatively small (1.0 to $1.4\,g\,cm^{-3}$) and close to that of the water. As such, a change in floc density or porosity does not have a significant effect on settling velocity. Floc shape was found to have minimal effect on the settling velocity of the flocs and is believed to be even less significant in influencing sediment transport in turbulent environments. Given the significant influence that flocculation has on the transport of suspended sediment, it is important that the structure and behavior (transport/settling) of flocculated particles be known. A better understanding of flocculated particle characteristics will help to refine our knowledge of cohesive suspended sediment transport within a variety of aquatic ecosystems.

Acknowledgments

The author would like to thank C. Jaskot and M. West for their assistance with image analysis.

FURTHER READING

Buffle J. and van Leeuwen H.P. (1992) *Environmental Particles*, Vol. 1 and 2 Lewis Publishers: Boca Raton.

Burt N., Parker R. and Watts J. (1997) *Cohesive Sediments*, John Wiley & Sons: Chichester.

Droppo I.G. (2000) Filtration in particle size analysis. *Encyclopedia of Analytical Chemistry*, Meyers R. A. (Ed.), John Wiley & Sons: Chichester, pp. 5397–5413.

Droppo I.G., Leppard G.G., Liss S.N. and Milligan T.G. (Eds.) (2005) *Flocculation in Natural and Engineered Environmental Systems*. CRC Press, Boca Raton.

Mehta A.J. (1993) *Nearshore and Estuarine Cohesive Sediment Transport*, American Geophysical Union: Washington.

Syvitski J.P.M. (1991) *Principles, Methods, and Applications of Particle Size Analysis*, Cambridge University Press: Cambridge.

Wotton R.S. (1990) *The Biology of Particles in Aquatic Systems*, CRC Press: Boca Raton.

REFERENCES

Allen T. (1990) *Particle Size Measurement*, Chapman & Hall: London.

Andreadakis A.D. (1993) Physical and chemical properties of activated sludge floc. *Water Research*, **27**, 1707–1714.

Asper V.L. (1987) Measuring the flux and sinking speed of marine snow aggregates. *Deep-Sea Research*, **34**, 1–17.

Azetsu-Scott K. and Johnson B.D. (1992) Measuring physical characteristics of particles: a new method of simultaneous measurement of size, settling velocity and density of constituent matter. *Deep-Sea Research*, **39**, 1057–1066.

Bale A.J. and Morris A.W. (1987) *In situ* measurement of particle size in estuarine waters. *Estuarine Coastal Shelf Science*, **24**, 253–263.

Burban P.-Y., Xu Y.-J., McNeil J. and Lick W. (1990) Settling speeds of flocs in fresh water and seawater. *Journal of Geophysical Research*, **95**(C10), 18213–18220.

Curran K.J., Hill P.S. and Milligan T.G. (2002) Fine-grained suspended sediment dynamics in the Eel river flood plume. *Continental Shelf Research*, **22**, 2537–2550.

de Boer D.H. (1997) An evaluation of fractal dimensions to quantify changes in the morphology of fluvial suspended sediment particles during baseflow conditions. *Hydrological Processes*, **11**, 415–426.

de Boer D.H., Stone M. and Levesque L.M.J. (2000) Fractal dimensions of individual flocs and floc populations in streams. *Hydrological Processes*, **14**, 653–667.

Droppo I.G. (2001) Rethinking what constitutes suspended sediment. *Hydrological Processes*, **15**, 1551–1564.

Droppo I.G., Jeffries D., Jaskot C. and Backus S. (1998a) Freshwater flocculation in cold regions: a case study from the Mackenzie river delta, Northwest territories, Canada. *Arctic*, **51**, 155–164.

Droppo I.G., Lau Y.L. and Mitchell C. (2001) The effect of depositional history on contaminated bed sediment stability. *Science of the Total Environment*, **266**, 7–13.

Droppo I.G., Leppard G.G., Flannigan D.T. and Liss S.N. (1997) The freshwater floc: a functional relationship of water and organic and inorganic floc constituents affecting suspended sediment properties. *Water, Air and Soil Pollution*, **99**, 43–53.

Droppo I.G. and Ongley E.D. (1992) The state of suspended sediment in the freshwater fluvial environment: a method of analysis. *Water Research*, **26**, 65–72.

Droppo I.G. and Ongley E.D. (1994) Flocculation of suspended sediment in rivers of southeastern Canada. *Water Research*, **28**, 1799–1809.

Droppo I.G. and Stone M. (1994) In-channel surficial fine-grained sediment laminae (Part I): physical characteristics and formational processes. *Hydrological Processes*, **8**, 101–111.

Droppo, I.G., Walling, D.E. and Ongley, E.D. (1998b) Suspended sediment structure: implications for sediment and contaminant transport modelling. In *Modelling Soil Erosion, Sediment Transport and Closely Related Hydrological Processes*, Summer W., Klaghofer E. and Zhang W. (Eds.), IAHS Publication No. 249, IAHS: Wallingford, pp. 437–444.

Droppo I.G., Walling D.E., Ongley E.D. (2000) The influence of floc size, density and porosity on sediment and contaminant transport. In *The Role of Erosion and Sediment Transport in Nutrient and Contaminant Transfer*, Stone M. (Ed.), IAHS Publication No. 263, IAHS: Wallingford, pp. 141–147.

Droppo I.G., Walling D.E. and Ongley E.D. (2002) Suspended sediment structure: implications for sediment transport/yield modelling. In *Erosion and Sediment Transport/Yield Modelling*, Summer W. and Walling D.E. (Eds.), IHP-VI – Technical Document in Hydrology, No. 60, UNESCO: Paris, pp. 205–228.

Eisma D., Dyer K.R. and van Leussen W. (1997) The *in-situ* determination of the settling velocities of suspended fine-grained sediment – a review. In *Cohesive Sediments*, Burt N., Parker R. and Watts J. (Eds.), John Wiley & Sons: Chichester, pp. 17–44.

Fennessy M.J., Dyer K.R. and Huntley D.A. (1994) INSSEV: an instrument to measure the size and settling velocity of flocs *in situ*. *Marine Geology*, **117**, 107–117.

Gibbs R.J. (1985) Estuarine flocs: their size, settling velocity and density. *Journal of Geophysical Research*, **90**(C2), 3249–3251.

Hawley N. (1982) Settling velocity distribution of natural aggregates. *Journal of Geophysical Research*, **87**(C12), 9489–9498.

Hill P.S., Syvitski J.P., Cowan E.A. and Powell R.D. (1998) *In situ* observations of floc settling velocity in Glacier Bay, Alaska. *Marine Geology*, **145**, 85–94.

Klemetson S.L. (1985) Factors affecting stream transport of combined sewer overflow sediments. *Journal Water Pollution Control Federation*, **57**, 390–397.

Krone R.B. (1972) *A Field Study of Flocculation as a Factor in Estuarial Shoaling Processes*, U.S. Corps of Engineers, Committee on Tidal Hydraulics, Technical Bulletin, 19, p. 62.

Krumbein W.C. (1942) Settling-velocity and flume-behavior of non-spherical particles. *American Geophysical Union Transactions*, **23**, 621–632.

Lagvankar A.L. and Gemmel R.S. (1968) A size-density relationship for flocs. *Journal of American Water Works Association*, **60**, 1040–1046.

Leppard G.G. (1992) Evaluation of electron microscopic techniques for the description of aquatic colloids. In *Environmental Particles* Vol. 1, Buffle J. and van Leeuwen H.P. (Eds.), Lewis Publishers: Boca Raton: pp. 231–289.

Lerman A. (1979) *Geochemical Processes: Water and Sediment Environments*, John Wiley & Sons: New York.

Li D.-H. and Ganczarczyk J. (1987) Stroboscopic determination of settling velocity, size and porosity of activated sludge flocs. *Water Research*, **21**, 257–262.

Li D.-H. and Ganczarczyk J. (1988) Flow-through activated sludge flocs. *Water Research*, **22**, 789–792.

Li D.-H. and Ganczarczyk J. (1989) Fractal geometry of particle aggregates generated in water and wastewater treatment processes. *Environmental Science and Technology*, **23**, 1385–1389.

Li D.-H. and Ganczarczyk J. (1990) Structure of activated sludge flocs. *Biotechnology and Bioengineering*, **35**, 57–65.

Lick W., Lick J. and Ziegler C.K. (1992) Flocculation and its effect on the vertical transport of fine-grained sediments. *Hydrobiologia*, **235/236**, 1–16.

Liss S.N., Droppo I.G., Flannigan D. and Leppard G.G. (1996) Floc architecture in wastewater and natural riverine systems. *Environmental Science and Technology*, **30**, 680–686.

Logan B.E. and Hunt J.R. (1987) Advantages to microbes of growth in permeable aggregates in marine systems. *Limnology and Oceanography*, **32**, 1034–1048.

Logan B.E. and Hunt J.R. (1988) Bioflocculation as a microbial response to substrate limitations. *Biotechnology and Bioengineering*, **31**, 91–101.

Nicholas A.P. and Walling D.E. (1996) The significance of particle aggregation in the overbank deposition of suspended sediment on river floodplains. *Journal of Hydrology*, **186**, 275–293.

Ongley E.D., Bynoe M.C. and Percival J.B. (1981) Physical and geochemical characteristics of suspended solids, Wilton Creek, Ontario. *Canadian Journal of Earth Sciences*, **18**, 1365–1379.

Ongley E.D., Krishnappan B.G., Droppo I.G., Rao S.S. and Maguire R.J. (1992) Cohesive sediment transport: Emerging issues for toxic chemical management. *Hydrobiologia*, **235/236**, 177–187.

Ozturgut E. and Lavelle J.W. (1984) New method of wet density and settling velocity determination for wastewater effluent. *Environmental Science and Technology*, **18**, 947–952.

Partheniades E.P. (1986) Turbidity and cohesive sediment dynamics. In *Marine Interface Ecohydrodynamics*, Partheniades E. and Nihoal J.C.J. (Eds.), Elsevier Science Publishers: Amsterdam, pp. 515–550.

Peart M.R. and Walling D.E. (1982) Particle-size characteristics of fluvial suspended sediment. In *Recent Developments in the Explanation and Prediction of Erosion and Sediment Yield*, Walling E.D. (Ed.), IAHS Publication No. 137, IAHS: Wallingford, pp. 397–407.

Petticrew E.L. and Droppo I.G. (2000) The morphology and settling characteristics of fine-grained sediment from a selection of Canadian rivers. *Contributions to IAP-V by Canadian Experts*, IHP-V – Technical Documents in Hydrology, No. 33, UNESCO: Paris, pp. 111–126.

Phillips J.M. and Walling D.E. (1995) Measurement *in situ* of the effective particle-size characteristics of fluvial suspended sediment by means of a field-portable laser backscatter probe: some preliminary results. *Marine and Freshwater Research*, **46**, 349–357.

Phillips J.M. and Walling D.E. (1999) The particle-size characteristics of fine-grained channel deposits in the river Exe Basin. *Hydrological Processes*, **13**, 1–19.

Richards K. (1982) *Rivers: Form and Process in Alluvial Channels*, Methuen and Company: New York.

Riley G.A. (1970) Particulate matter in sea water. *Advances in Marine Biology*, **8**, 1–118.

Sherman I. (1953) Flocculant structure of sediment suspended in Lake Mead. *American Geophysical Union. Transactions.*, **34**, 394–406.

Shiozawa T., K. K.awana, Hoshika A. and Tanimoto T. (1985) Suspended matter and bottom sediment in the Seto Island sea. *Bulletin on Coastal Oceanography*, **22**, 149–156.

Tambo N. and Hozumi H. (1979) Physical characteristics of flocs – II. Strength of floc. *Water Research*, **13**, 13.

Tambo N. and Watanabe Y. (1979) Physical characteristics of flocs – 1. The floc density function and aluminum floc. *Water Research*, **13**, 13.

van Leussen W. (1988) Aggregation of particles, settling velocity of mud flocs: a review. In *Physical Processes in Estuaries*, Dronkers J., van Leussen W. (Eds.), Springer-Verlag: New York, 347–403.

van Leussen W. and Cornelisse J.M. (1993) The determination of the sizes and settling velocities of estuarine flocs by an underwater video system. *Netherlands Journal of Sea Research*, **31**, 31.

Wall G.J., Wilding L.P. and Smeck N.E. (1978) Physical, chemical and mineralogical properties of fluvial unconsolidated bottom sediments in northwestern Ohio. *Journal of Environmental Quality*, **7**, 7.

Walling D.E. and Kane P. (1984) Suspended sediment properties and their geomorphological significance. In *Catchment Experiments in Fluvial Geomorphology*, Burt T.P. and Walling D.E. (Eds.), Geobooks: Norwich, pp. 311–344.

Walling D.E. and Moorehead P.W. (1989) The particle-size characteristics of fluvial suspended sediment: an overview. *Hydrobiologia*, **176/177**, 125–149.

Walling D.E. and Woodward J.C. (1993) Use of a field-based water elutriation system for monitoring the *in situ* particle-size characteristics of fluvial suspended sediment. *Water Research*, **27**, 1413–1421.

Woodward J.C. and Walling D.E. (1992) A field sampling method for obtaining representative samples of composite fluvial suspended sediment particles for SEM analysis. *Journal of Sedimentary Petrology*, **62**, 62.

Zahid W.M. and Ganczarczyk J.J. (1990) Suspended solids in biological filter effluents. *Water Research*, **24**, 24.

84: Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands

HANS MIDDELKOOP

Department of Physical Geography, Utrecht University, The Netherlands

Floodplain sedimentation plays a keyrole in the sediment budget of river basins, but it also relates to river management, flood protection, civil engineering, and ecological rehabilitation of rivers. Understanding of the process, therefore, is essential both from a geomorphologic and river management points of view. This paper provides a review of recent research on floodplain sedimentation, focusing on field-based methods to reconstruct and document spatial patterns of overbank sedimentation, as well on mathematical modeling approaches. Examples of floodplain sedimentation over various timescales are given for the lower Rhine River in the Netherlands. The final section synthesizes the processes, their controls, and the resulting patterns and amounts of floodplain deposition.

INTRODUCTION

A floodplain comprises the area next to a river channel that is inundated once during a given recurrence interval. Floodplains are complex assemblages of landforms built up by channel bars and bed forms, lateral accretion scroll bars and natural levees, point bars overlaid by overbank deposits, old channels, backswamps, and crevasse splays (Figure 1). Many lowland floodplains undergo regular flooding, causing deposition of fine material from suspension during inundation. In spite of the infrequency or irregularity of flooding, extensive bodies of overbank fines may accumulate by vertical aggradation over a period of decades to centuries. At this timescale, only a small fraction of the total alluvium in a river valley is transported by the river; the bulk is stored in floodplains.

The earliest studies on floodplain sediments and fluvial architecture were concerned merely with lateral accretion, whereas relatively little attention was paid to overbank deposition rates (Marriott, 1992). Estimates of overbank deposition rates initially relied on visually distinct soil profile characteristics such as color or texture to discern flood deposits, whereas analysis of deposition processes focused on crevasse spay development or analysis of deposits after rare, extreme flood events (Simm, 1995). However, little progress was made in documenting spatial patterns of overbank deposition resulting from the more frequent and moderate-magnitude floods. This may be due to the difficulties associated with sampling small quantities of spatially and temporally highly variable overbank deposition over large floodplain areas. Another practical limitation is that sampling surveys are difficult to plan because of infrequent or unpredictable inundation of the area of interest, while equipment and fieldwork are faced with potentially hazardous flood conditions.

In recent years, the study of floodplain deposition has gained interest due to the increased awareness of the geological, geomorphological, environmental, and societal significance of the process. Floodplain deposition is now known as a key component in the sediment budgets of catchments, since floodplains may form significant sinks for suspended sediment transport through a river system (Walling *et al.*, 1998a; Walling, 1983) (*see* **Chapter 85, Sediment Yields and Sediment Budgets,**

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

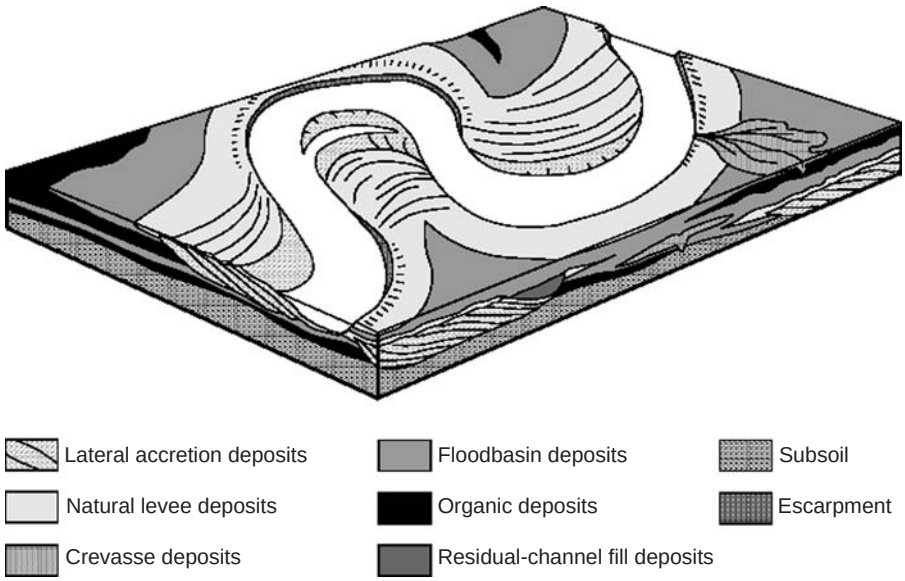


Figure 1 Block diagram of a meandering river. River flow is to the left

Volume 2). On the geological timescale, alluvial deposits are volumetrically important sediment bodies in the rock record. Furthermore, suspended sediment plays an important role of in the transport of particulate-bound pollutants and contaminants through fluvial systems, and their accumulation in floodplain environments (Salomons and Förstner, 1984; Foster and Charlesworth, 1996). In addition to their functioning as sediment sinks in catchments budgets, floodplain deposits provide aggregated information about sediment delivery and changes in sediment yield from the upstream catchments, associated with changes in climate and land use (Hudson, 2003) (*see* **Chapter 85, Sediment Yields and Sediment Budgets, Volume 2** and **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**).

Understanding of floodplain processes also relates to river management, flood protection, civil engineering, and river rehabilitation. For centuries, floodplains have been areas of preferred and most intense human settlement, due to the proximity of water for transport and supply, and abundant fertile land for agriculture. Floodplains have become densely populated areas with extensive urban and industrial areas and intensive farming. Recent years have witnessed major river floods causing extensive floodplain inundation with severe human and economic losses. This has led to a growing concern over the frequency and magnitude of river floods and floodplain inundation and their relationship to intensified floodplain occupation, river canalization, and the potential implications of climate change (Knox, 1993). The development of appropriate flood management strategies has raised the need of a sound understanding of the hydraulic characteristics of overbank flows, and the ability to quantify and predict the relationships between river discharge, floodplain topography and the resulting patterns of overbank flow, and deposition of sediments and associated pollutants. Finally, river rehabilitation projects that involve landscaping measures such as reopening of secondary channels, lowering the floodplain surface, removing minor embankments, or the introduction of floodplain forests (Silva *et al.*, 2000) will lead to major changes of the floodplain topography and vegetation, which in turn may considerably affect overbank sedimentation rates (Middelkoop and Van der Perk, 1998).

These issues demand knowledge and understanding of the process of overbank deposition, the quantification of the spatial and temporal variability of contemporary deposition, as well of the main controls of these related to flood magnitude, sediment load in the river, and floodplain topography. Given the possible effects of climate change and changes in land use and management on river flow and catchment sediment budgets, there is a need to place information on contemporary suspended sediment fluxes and overbank deposition rates into a longer-term context to identify long-term budgets and trends (Owens *et al.*, 1999a).

Overbank Deposition

Overbank deposition of suspended sediment upon floodplains occurs during periods of high discharge when fine-grained material, supplied to the fluvial system by catchment runoff, is transported out of channel over the floodplain and deposited in regions of reduced flow competence (Nicholas and Walling, 1997b). Mean annual overbank sedimentation rates are controlled by factors such as flood

frequency and magnitude, concentration and settling characteristics of the transported suspended sediment, and floodplain topography. Contemporary floodplain sedimentation rates reported in the literature typically range between 0.5 and 20 mm $year^{-1}$, with maximum values observed during extreme flood events up to several centimeters (e.g., Simm, 1995; Terry *et al.*, 2002). Deposition rates show considerable spatial variation not only between different floodplains but also within individual floodplain reaches. Observed *spatial patterns* of overbank deposition are associated with local differences in flooding duration, floodplain topography, and distance to the river channel, and have been related to the influence of floodplain topography and surface roughness upon overbank flow patterns and flow competence (Middelkoop and Asselman, 1998; Middelkoop and Van der Perk, 1998). *Temporal* variations in floodplain sedimentation rates are generally related to hydrologic changes and to changes in land use and soil conservation practices in the upstream areas.

Although deposition rates are often low and in some cases almost imperceptible, substantial amounts may be deposited when large areas are involved (Walling and Owens, 2003). Consequently, because of overbank deposition, floodplains may trap a considerable part of the suspended sediment load transported through a river system to the catchment outlet (*see* **Chapter 85, Sediment Yields and Sediment Budgets, Volume 2** and **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**). The resulting conveyance losses of suspended sediments may be in the order of 10% to more than 40% of the annual load entering the reach (e.g. Lambert and Walling, 1986; Walling *et al.*, 2003). In contrast with channel-bed storage where the residence time of sediments is usually in the order of a year, sediment deposited on the floodplain may remain stored for decades to centuries before reentering the river system due to overbank erosion. This role as sink for sediments may complicate the interpretation of downstream sediment yields in terms of upstream sediment budgets of drainage basins, particularly in larger drainage basins (Walling *et al.*, 1998a; Asselman *et al.*, 2003).

Contaminant Deposition

Many contaminants, including heavy metals and pesticides, as well as nutrients are transported through the fluvial system in association with fine sediment (Salomons and Förstner, 1984; Foster and Charlesworth, 1996) (*see* **Chapter 89, On the Worldwide Riverine Transport of Sediment – Associated Contaminants to the Ocean, Volume 2**). Consequently, sediment-associated contaminants will also be deposited on river floodplains during overbank flooding. In polluted river basins particularly, because of the release of large volumes of wastes in the nineteenth and twentieth centuries from mining and industrial activities, floodplain soils may have become heavily contaminated by pollutants, owing to the deposition of contaminated sediments over many years.

Initial studies on sediment-associated river pollution tended to focus on the analysis of suspended sediment in the river, but increasingly it is being recognized that floodplain deposits should also be considered when analyzing the fate of contaminants and the temporal and spatial patterns of contaminant distribution within river systems (Salomons and Förstner, 1984; Leenaers, 1989; Owens *et al.*, 2001; Walling *et al.*, 2003). Because of overbank sedimentation, floodplains function as sinks in sediment-associated contaminant budgets of catchment contaminants (Walling and Owens, 2003; Owens *et al.*, 2001). This can result in a reduction of the nutrient or contaminant flux at the catchment outlet, such that the contaminant flux measured at the catchment outlet may significantly underestimate the total mass of the contaminants mobilized within the catchment. Conversely, contaminant tracers contained in floodplain soils may allow determination of sediment sources by means of fingerprinting (He and Owens, 1995; Collins *et al.*, 1997; Owens *et al.*, 1999a), and may help estimate and understand contaminant budgets of river catchments (e.g. Walling and Owens, 2003).

Within the floodplain environment, the accumulation of contaminants may lead to severely enhanced levels of contamination within the floodplain ecosystem and food webs. Nature rehabilitation projects that involve excavation of floodplain deposits may be confronted with high sanitation costs of the contaminated sediments (Thonon *et al.*, 2002), while geochemical mapping of floodplains may become problematic because of large spatial variations in contamination, associated with the history of pollutant release into the river system as well as spatial patterns of past overbank sedimentation rates (Bradley, 1995).

As a consequence of the role of floodplains as a sink for contaminants, contaminated overbank deposits present a potential source for future contamination of the river system, when these are mobilized because of bank erosion (e.g., Bradley and Cox, 1990; Bradley, 1995; Walling *et al.*, 2003). While mine wastes and other point sources are important primary supplies of metals into the river system, significant quantities of metals are cycled through the river system as a result of floodplain deposition and the subsequent erosion of contaminated floodplain deposits. This causes heavily polluted floodplains to be considered as "chemical time bombs". This cycling of metals may continue for many years, even if direct inputs to the river were to cease. Understanding the rate and scale of this continued cycling of metals in river catchments relies on the analysis of the interaction between floodplain and the channel, and on the processes of sedimentation and erosion involved.

Outline

The present paper focuses on the deposition of overbank sediments and associated particulate-bound heavy metals outside the channel of meandering lowland rivers. It first gives a review on methods of documenting floodplain sedimentation rates and patterns in the medium-term (10^1-10^2 years) and event-based timescale, as well as concepts applied for modeling the process of overbank deposition. Subsequently, examples of floodplain sedimentation over various timescales are shown for the lower Rhine distributaries in the Netherlands. This is followed by a more general overview of the main mechanisms and controls of the sedimentation process. Finally, an outlook is given on research issues for improving our present understanding of floodplain sedimentation and its geomorphologic and ecologic significance.

QUANTIFICATION OF MEDIUM-TERM AMOUNTS AND PATTERNS OF FLOODPLAIN SEDIMENTATION

Introduction

Investigation of floodplain sedimentation at medium-term (10–100 years) timescales is carried out to reconstruct changes in overbank sedimentation rates and sediment sources in historic time (Walling *et al.*, 1999; Owens *et al.*, 1999a), and to determine the role of floodplains storage as sink in the suspended sediment budget of river systems. Within the floodplain area, these reconstructions are needed to understand the relationships between average deposition rates and distance to the main channel, relative elevation, and other aspects of floodplain topography.

Estimates of sedimentation rates averaged over a number of years rely on determining the date of one or more particular levels in the sediment layer, and calculating the average rate of deposition of the material overlying these levels. Methods to determine the age of levels within the profiles may be based on (Simm, 1995): (i) Historic data and map comparison (e.g., Wolfenden and Lewin, 1977; Magilligan, 1985), and impressively long records from China (Xu, 2003), (ii) previous benchmark studies (e.g. Leopold, 1973) (iii) buried soils, roads, or artifacts (e.g. Costa, 1975), (iv) dendrochronology of floodplain vegetation (Nakamura and Kikuchi, 1996), (v) radiocarbon dating (e.g. Alexander and Prior, 1971; Baker *et al.*, 1985; Brown, 1987), (vi) using fallout radionuclides such as ^{10}Be, ^{210}Pb and ^{137}Cs as tracers (e.g. Campbell *et al.*, 1982, 1988; Osterkamp, 1989; Walling *et al.*, 1986b; Walling *et al.*, 1989; Walling and Bradley, 1989; Walling *et al.*, 1992; Walling and He, 1993, 1994a, 1994b; Simm, 1995, He and Walling, 1996), and (vii) relating trace metal concentrations in the sediment profile to the known history of mining activities in the upstream catchment (e.g. Magilligan, 1985; Lewin and Macklin, 1987; Knox, 1987, 1989; Leenaers, 1989; Bogen *et al.*, 1992). Measurements of the depth distribution of fallout radionuclides and heavy metals within a sediment core collected from a floodplain provide an effective means to determine past sedimentation rates. The basis of using these tracers to establish a chronology of sediment cores is that their input into the environment is constant, or shows an *a priori* known variation in time. Differences in total amounts of tracers, or in the depth at which characteristic changes in tracer concentrations within the soil profiles occur, are a measure of deposition rates.

Fallout Radionuclides

The basis of using fallout radionuclides in sediment budget investigations is that their fallout input can be assumed to be effectively uniform over a relatively small area, and that the fallout is rapidly and firmly fixed by the surface soil. Subsequent redistribution of the particle-bound fallout in the environment will therefore reflect erosion and deposition of sediment particles. Measurements of the spatial distribution of radionuclides in floodplain soils can therefore be used to investigate spatial patterns of sediment deposition. As sorption of fallout radionuclides by sediment particles generally is associated with cation exchange at the surface of sediment particles, and hence depends on their grain size, comparison of fallout concentrations in different sediment samples should be done for similar grain size fractions taken from each sample (Walling and Woodward, 1992; He and Owens, 1995; Walling *et al.*, 1998a). Since different source areas of sediments deposited on floodplains may contain different amounts of radionuclides, these can also be used as "fingerprints" to reconstruct the sediment source areas (e.g. He and Owens, 1995).

^{210}Pb

The unstable isotope ^{210}Pb is one of the products of the ^{238}Uranium decay series, with a half-life of 22.3 years. ^{210}Pb in surface sediment has two sources: *in situ* production and atmospheric deposition, both of which can be considered constant through time. For age determination it is essential to separate the ^{210}Pb content of a sample into these two components. Uranium-238 is present in rocks and soils and decays to ^{226}Radium (half-life 1622 years), which in turn decays to ^{222}Radon. This is an inert gas (half-life 3.8 days) that decays through a series of short-lived processes to ^{210}Pb. The ^{210}Pb produced by the *in situ* decay of ^{226}Ra is called "supported", and will be in equilibrium with ^{226}Ra. Part of the ^{222}Radon gas, however, escapes to the atmosphere, where it decays to ^{210}Pb. The atmospherically produced ^{210}Pb is subsequently deposited into surface waters through precipitation and dry deposition, where it is readily adsorbed by suspended sediment. This atmospheric deposition is called the *unsupported* or *excess* lead, as

it is not supported by decay of its *in situ* parent. In the upper part of a sediment core, the unsupported lead is recognized as an excess activity over the supported activity, expressed in $Bq\,kg^{-1}$. After a period of about 100 years (i.e. 5 half-lifetimes), this excess ^{210}Pb has become too small to be determined, thereby confining the applicability of the method to this period. As supported ^{210}Pb will be in equilibrium with ^{226}Ra, the amount of unsupported ^{210}Pb can be determined by measuring both total ^{210}Pb and ^{226}Ra, and subtracting the supported component. Another method is to estimate the supported ^{210}Pb from samples in deep, old parts of the sediment core, where the supported component has decayed. At sites with constant sedimentation rates (such as floodplains), the intensity profile of the excess ^{210}Pb will show an exponential decrease with depth, the decrease rate of which is determined by the deposition rate (Appleby and Oldfield, 1978; Dominik *et al.*, 1981, 1984). Alternatively, the sedimentation rate can be determined from the total excess ^{210}Pb activity over an entire sediment core (Walling and He, 1994a).

^{137}Cs

^{137}Cs is an artificial fallout radionuclide with a half-life of 30.2 years that has been introduced into the atmosphere by nuclear tests in the 1950s and 1960s. It circulated globally in the stratosphere, and fallout was associated with precipitation. The temporal pattern of this fallout is well documented and is assumed relatively uniform over relatively small areas. Fallout was first detected in the early 1950s; it reached a peak in 1963, and declined after the Nuclear Test Ban Treaty to low levels by the mid-1970s. Since the 1980s the ^{137}Cs fallout has become negligible, with exception of areas of Europe and adjacent regions affected by a short-term fallout due to the Chernobyl nuclear power plant incident of 1986. The deposition patterns of the Chernobyl fallout show considerable variation even at the local scale, associated with the pathway of the radioactive cloud and local rainstorms (De Cort *et al.*, 1998; He and Owens, 1995). With time, ^{137}Cs may migrate slightly down the profile, due to diffusive and convective processes and bioturbation (He and Owens, 1995; Walling and He, 1997b). The time-variable ^{137}Cs inputs can be detected from sediment cores in deposition zones such as floodplains, allowing the establishment of the chronology of the sediments and thereby estimating sedimentation rates over the past 40 years. To determine the deposition rates, three approaches can be followed (Walling and He, 1997b): The first is to compare the shape of the ^{137}Cs depth profiles with that in a reference site that does not experience flooding (Figure 2). The degree of vertical stretching of the ^{137}Cs profile is a measure of the deposition rate. The second approach uses the depth at which significant ^{137}Cs concentrations are first recorded to estimate the position of the surface in the mid-1950s, and the position of the levels with peak activity to estimate the depth of the 1963 and 1986 surfaces. However, in both these cases, it is necessary to take account of the potential for postdepositional redistribution of ^{137}Cs in the sediment profile, since this influences the depth of ^{137}Cs peaks within the profile. Furthermore, this approach requires a high-resolution sampling of a sediment core and subsequent detection of ^{137}Cs activities. An alternative and third method therefore is to measure the total ^{137}Cs inventory for single whole core obtained from floodplain sites and subsequently comparing this with the total ^{137}Cs amounts in sediment cores obtained at nearby undisturbed and nonaccumulating reference sites above the floodplain level. The excess ^{137}Cs in the floodplain core compared to the reference is a measure of the floodplain sedimentation rate (Walling and Bradley, 1989; Walling *et al.*, 1992, 1998a; Walling and He, 1994a, 1994b, 1997b; Nicholas and Walling, 1997b). In this case, it is important to retrieve the complete ^{137}Cs profile, but the advantage is that only a single determination of ^{137}Cs is required at each sample location. Using bulk ^{137}Cs inventories of entire cores is believed to yield more representative estimates of spatial variability in overbank sedimentation than if only a small number of sectioned cores are analyzed (Walling *et al.*, 1998a). With increasing sensitivity of modern γ-ray measurement devices, however, it has become possible to analyze ^{137}Cs activities over an entire sediment core at centimeter resolution. An example is the PHAROS high-resolution radiometric core analyzer that allows measuring radionuclides in a drill core to determine sedimentation by the peak and onset of ^{137}Cs (Rigollet and De Meijer, 2002; Van Wijngaarden *et al.*, 2002).

Because of their different age ranges represented by ^{210}Pb and ^{137}Cs, ca. 100 years and ca. 40 years respectively, combining these tracers may allow determining changes in historic sedimentation rates (Walling and He, 1994b; Owens *et al.*, 1999a; Owens and Walling, 2002) (Figure 2). Radiometric dating of floodplain sediment has been applied in many river systems across Europe, with many examples from the United Kingdom, but also from other rivers within and outside Europe (e.g. He and Owens, 1995; Nicholas and Walling, 1997b; Siggers *et al.*, 1999; Owens *et al.*, 1999a; Stam, 1999; Saxena *et al.*, 2002).

Heavy Metals

In most catchments with intensive mining and industrial activity, large amounts of pollutants including heavy metals are released into the fluvial system. Because of rapid and firm adsorption to sediment particles, heavy metals will be transported in association with sediments through the fluvial system, and will also be deposited on floodplains during overbank flow. In many fluvial systems metal contamination has been present for more than a century, during which time the metal concentrations in the river sediment have varied greatly, depending on the intensity and type of

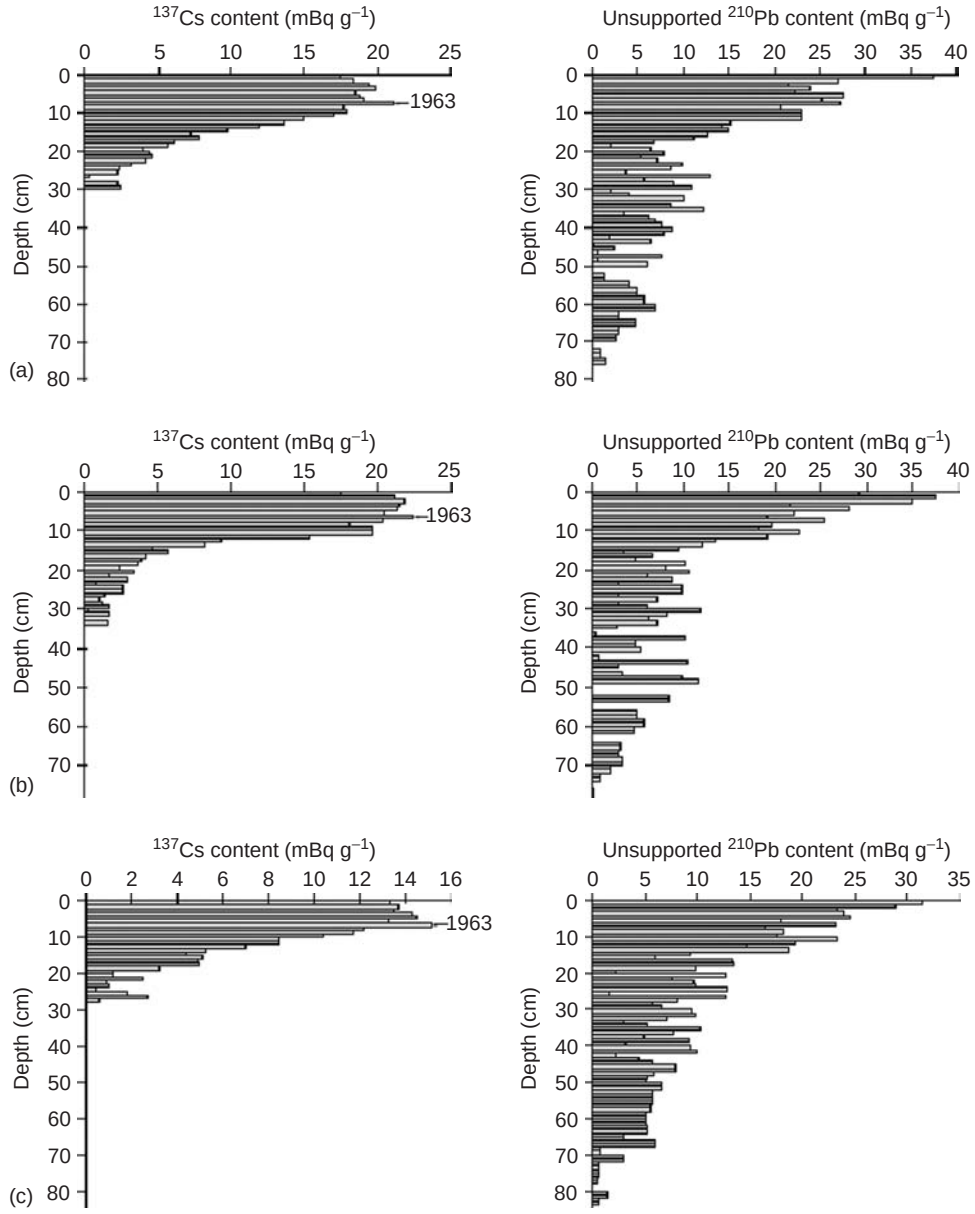


Figure 2 The depth distributions of ^{137}Cs and unsupported ^{210}Pb concentrations in the floodplain cores collected from the three sites along the Teviot and Tweed rivers, Scotland (Reproduced from Owens and Walling, 2002 by permission of John Wiley & Sons Ltd)

mining and industrial activity (e.g. Swennen *et al.*, 1994; Taylor, 1996; Owens *et al.*, 1999a; Middelkoop, 2000; Owens and Walling, 2003). Heavy metals can be used as tracers to estimate average floodplain sedimentation rates by comparing spatial differences among different concentration profiles in the floodplain soils, or by comparing concentration profiles with the chronology of the metal pollution of the river sediment, which has been reconstructed from historic or other independent sources. Because of their preferential bonding to finer sediment fractions and organic matter, metal concentrations may demonstrate large variability within floodplains due to varying grain size and organic matter contents of the sediment (Leenaers *et al.*, 1998; Middelkoop, 2000). These differences must be isolated first from the variations in pollution due to upstream pollution of the catchment to enable establishment of the sediment chronology (Owens *et al.*, 1999a; Middelkoop, 2002a; Owens and Walling, 2003).

1D Sedimentation Models

Attempts to use heavy metal profiles in floodplain sediment for establishing chronologies by only considering

characteristic changes or peaks in the concentration profile may encounter some major limitations. Firstly, a high sampling resolution is required to enable precise determination of the depth of peaks or of other characteristic changes in contaminant concentration profiles. Secondly, there has to be a detailed and accurate record of historic river pollution. Finally, postdepositional redistribution of the contaminants by chemical, physical, or biological processes may result in changes in the vertical profiles, which lead to inaccurate assignments of ages to profile depths (Hudson-Edwards *et al.*, 1998). Therefore, instead of directly assigning ages to characteristic horizons in a floodplain profile, depth–age relationships may be determined using 1D models that replicate the development of a floodplain soil profile.

Middelkoop (2002a) determined average sedimentation rates along the Rhine from heavy metal profiles in floodplain soils by means of inverse modeling. This was done with a one-dimensional model of sedimentation and postdepositional mixing that simulates the development of a metal profile in a floodplain soil using the annual amounts of sediment deposition and concentrations of Cu, Pb, and Zn in the deposited sediment as input. For this purpose, the trends in metal pollution of the Rhine river sediment deposited on the floodplain during the past century were reconstructed independently from dated sediment cores from small ponds within the floodplain. Estimation of the average sedimentation rate during the development of a floodplain metal profile was done by calibration of the model against the observed metal profile in the floodplain soil, thereby varying the input sedimentation rate. The advantage of this method is that it uses information over an entire profile, while it requires only a limited number of samples, accounts for postdepositional redistribution of the metals, and provides quantitative estimates of the precision of the obtained sedimentation rates.

In a study on the Powder River, Montana, USA, Moody and Troutman (2000) used a simple one-dimensional model of floodplain growth by vertical aggradation to demonstrate the monotonic growth of floodplain, as deposition rate of sediment during overbank flooding for the duration that the threshold discharge corresponding to the floodplain height was exceeded. Because of increasing floodplain elevation the annual duration of deposition gradually decreases, the growth rate of the floodplain decreases over the years. Net annual deposition rates were estimated by comparing floodplain growth curves with empirical growth curves based on dendrochronology and direct field measurements.

QUANTIFICATION OF SHORT-TERM AMOUNTS AND PATTERNS OF FLOODPLAIN SEDIMENTATION

Studying shorter-term (e.g. annual, flood event) rates and patterns of floodplain deposition is essential to understand the relationships between sedimentation, flood characteristics (magnitude and sediment concentrations in the main channel), and floodplain properties and site characteristics (local elevation, distance to the main channel, topographical elements). Over the past decennia, different methods have been employed for measuring and quantifying shorter-term rates. A recent review of these approaches is given by Steiger *et al.* (2003) and the most generally used ones are summarized here.

Estimation of Average Deposition from Conveyance Losses

From observations of suspended sediment transport along a river reach, conveyance losses over the reach under consideration can be determined. These can provide estimates of average sedimentation rates along river reaches between gauging sites during individual flood events or over longer periods of time (e.g. Lambert and Walling, 1986; Gretener and Strömquist, 1987; Walling *et al.*, 1986a; Walling and Bradley, 1989). The reliability of observations of suspended sediment transport is heavily dependent upon the temporal frequency of transport observations, particularly during a flood event when concentrations are high and change rapidly in time, and upon the degree to which the observations represent sediment transport over the entire river channel cross section (e.g. Walling, 1983; Walling and Webb, 1987). Furthermore, suspended sediment input through bank erosion or from tributaries in between the upstream and downstream sample stations, and varying within-channel storage of fine sediments may lead to inaccurate estimates of conveyance loss due to floodplain deposition. Comparison of estimated conveyance losses along a stretch of the River Culm, Devon, UK with ^{137}Cs measurements and sediment trap measurements indicated that the conveyance loss approach only provides tentative estimates of average sedimentation rates (Walling *et al.*, 1986a).

Post-event Field Surveys

When no sampling or monitoring devices have been installed prior to a flood event, deposition patterns resulting from the flood must be documented by post-event field surveys onto the horizontal distribution of the deposition using aerial photographs or multispectral remote sensing images. The deposition depth is determined using existing marker horizons, or using techniques to analyze subsurface sedimentary structures. Such approaches are particularly suitable in mapping sandy deposits. Discrimination of freshly deposited sediment may be based upon burial of natural marker horizons such as the vegetation growing on the pre-event ground surface, or upon differences in colour or texture between the deposited and pre-event surface sediments. This method has been used for more than 60 years, initially in US surveys, but later

also reported from European rivers (e.g. Mansfield, 1939; Wolman and Eiler, 1958; McKee *et al.*, 1967; Kesel *et al.*, 1974; Brown, 1983; Gomez *et al.*, 1995, 1997, 1998; Ten Brinke *et al.*, 1998; Wyźga, 1999). Ten Brinke *et al.* (1998) determined thickness of sand splays deposited during two high-magnitude floods along the lower Rhine by field measurements, and subsequently extended these observations over the entire bank of the lower Rhine distributaries using aerial photographs to estimate the total deposited sand volume. Magilligan *et al.* (1998) did the same for overbank deposits after an extreme flood on the Upper Mississippi River using aerial survey and Landsat-5 TM satellite data. Mertes (1994) calculated a mass balance of the sediment transported along floodplain transects on the Amazon River from patterns of sediment concentration on a Landsat TM satellite image, from which he estimated amounts of floodplain deposition as well as order-of-magnitude calculations of vertical deposition rates.

Analysis of subsurface annual layers of sandy overbank deposits and crevasse splays can be made using ground-penetrating radar or electromagnetic induction. Ferguson and Brierley (1999) discriminated well-sorted sandy beds and thicker silty-sand beds in river levees using ground-penetrating radar profiles run across the levee and adjacent geomorphic units. Bristow *et al.* (1999) used the same equipment to survey the depth of crevasse splays. Kitchen *et al.* (1996) coupled electromagnetic induction, ground conductivity sensing, and locational information derived using a global positioning system to map sand deposition depth after major floods.

Artificial Marker Horizons

Instead of depending upon naturally occurring marker horizons, artificial markers spread over the floodplain in advance of a flood can locate the pre-event surface after deposition has occurred. Examples of marker types are based on colour differences of red sand, brick dust, coal dust, or white clay, mostly used in coastal marshes and wetlands (Steiger *et al.*, 2003). This approach has the advantage that the pre-event surface is clearly marked and that the marker has negligible influence on the hydraulic resistance of the surface for which sedimentation rates are estimated. Nevertheless, the method may fail when the marker horizon is eroded or redistributed. Also, difficulties may arise when trying to remove the freshly deposited sediment from the underlying marked sediment without contaminating the new deposits with the underlying pre-event deposits.

Erosion Pins

Erosion pins are long, thin metal pins inserted into river banks so that repeated measurement of the exposed length of each pin provides an estimate of the local bank erosion rate. Where net deposition of sediment occurs, repeated measurements of pin burial have shown an accurate approach to estimate depths of sedimentation, particularly on plane surfaces within wetlands (e.g. Ranwell, 1964). Automated continuous measurement of erosion–deposition can be made using photoelectronic erosion pins (PEEP) introduced by Lawler (1991). Erosion pins are inexpensive tools that can be used in large numbers. Although they do not disturb the sediment, the pins may have an important impact on local flow patterns and therefore, on local sedimentation rates. Erosion pins may also be unstable in noncohesive sediment or when exposed to floating matter such as woody debris. Consequently, erosion pins are most likely to generate accurate results in lower energy environments (Steiger *et al.*, 2003).

Sediment Traps

Numerous recent studies to quantify spatial patterns of overbank sedimentation resulting from individual floods have employed sediment traps placed directly on the floodplain surface in advance of a flood event. (e.g. Mansikkaniemi, 1985; Gretener and Strömquist, 1987; Lambert and Walling, 1987; Walling and Bradley, 1989; Asselman and Middelkoop, 1995; Simm, 1995; Middelkoop and Asselman, 1998; Steiger *et al.*, 2001a,b). Mansikkaniemi (1985) pioneered with different sediment samplers of about 0.5×0.5 m in size, including simple plywood boards with 5–7-cm long bristles and rubber sheets with a rough surface. Sedimentation results showed no significant variation between the different types of traps. Over the past years, various trap types have been applied such as, fire-clay roof tiles (approximate size: 20×30 cm) (Brunet *et al.*, 1994; Steiger *et al.*, 2001b; Steiger and Gurnell, 2003), 15-cm diameter plastic circles with roughened upper surface to reduce sediment wash-off, used within a bottomland hardwood wetland (Kleiss, 1996; Wardrop and Brooks, 1998), simple flat smooth plates or sheets of plastic (area 0.06–0.08 m^2) (Pinay *et al.*, 1995; Dezzeo *et al.*, 2000), and carpet squares (20×20 cm) stitched onto aluminum frames (Walker, 1995).

Most studies employed sediment traps consisting of 1–2-cm-long tufts of artificial grass attached to a pliable base varying in size between 0.01 and 0.25 m and fixed onto the ground by steel pins. After the flood the sediment is removed using a high-pressure cleaner (Asselman and Middelkoop, 1995; Steiger *et al.*, 2001a), or through careful brushing after drying. This type of sediment traps is an adequate method for sampling sediment deposited by flowing water, particularly where sites are subject to repeated inundation, because (Steiger *et al.*, 2003) (i) the surface roughness of the mats reduces problems of sediment removal by floodwaters or rainfall, (ii) they are easily manipulated in the field, (iii) their pliable base permits installation on irregular surfaces and slopes, (iv) they can be

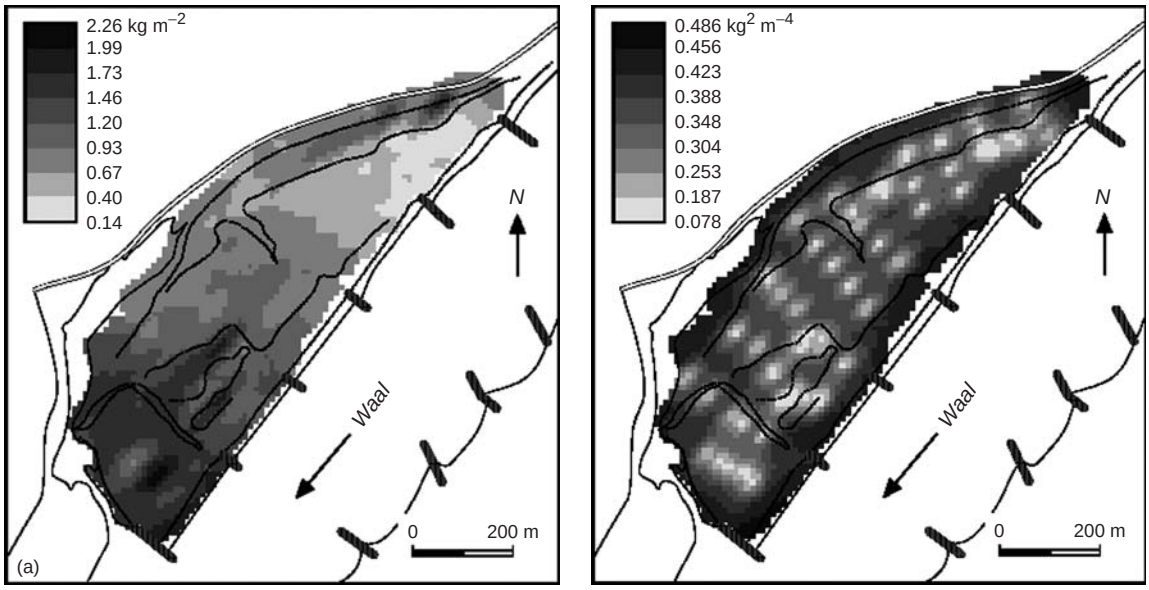


Figure 3 Interpolation result of sediment deposition on a floodplain section along the Wall River, The Netherlands, sampled with sediment traps. (a) interpolated deposition. (b) estimation variances of interpolation. The position of the sediment traps is marked by the low variances

securely attached to the ground and so can resist relatively high local shear stresses, (v) they can withstand repeated flooding and laboratory processing, (vi) they allow to sample a wide range of deposition amounts (10^{-1}–10^{4} kg m^{-2}), and (vii) it is possible to fully recover the deposited sediment to determine the amount of sediment deposited and to support a range of other analyses.

Sampling Design and Interpolation

To document the deposition patterns, traps have been distributed over the floodplain using a variety of sampling designs. Instead of placing single traps to collect "representative" samples of sedimentation rates, traps can be arranged in clusters to determine at-the-site variation and to calculate average deposition rates from clusters of replicate traps (cf. Asselman and Middelkoop, 1995; Steiger *et al.*, 2001b; Steiger and Gurnell, 2003). The sampling layout may be along transects perpendicular to the main channel or along the *a priori* expected largest gradient in deposition. Other sampling schemes may include a stratified random sampling based on different morphological, topographical, and land use settings (e.g. Walling and Bradley, 1989; Simm, 1995; Steiger and Gurnell, 2003), or sampling in a semiregular grid, consisting of several transects oriented in the direction perpendicular to the main stream, with sampling intervals increasing away from the main channel from about 1 to about 100 m in the distal parts (Asselman and Middelkoop, 1995). Differences in sediment accumulation over short distances can be determined by including clusters of several traps within the sample layout.

To calculate the total amount of deposited sediment, and to determine the full 2D pattern of the sediment deposition, the trap measurements can be interpolated using different techniques. An adequate method for this purpose is kriging, which is a form of weighted local averaging where the interpolation weights are determined from variograms (Burrough, 1986). To cope with anisotropy of spatial variability, separate variograms can be determined for sediment deposition in the direction parallel and perpendicular to the main flow direction (Asselman and Middelkoop, 1995; Middelkoop and Asselman, 1998). When considerable trends exist in the sampled deposition, interpolation should be carried out using "universal kriging", while "block kriging" produces interpolations results as a raster to match the format and spatial support of raster-based model output. This enables calibration and validation of model results, where the cell-by-cell comparison of observed and calculated deposition can be weighed according to the estimated interpolation variances of the observed data (Middelkoop and Asselman, 1998; Middelkoop and Van der Perk, 1998) (Figure 3).

MODELING FLOODPLAIN SEDIMENTATION

Introduction

The spatial variability of observed sediment deposition on floodplains can usually be correlated with floodplain topography, flooding frequency, and distance to the main river channel. However, these relationships are essentially empirical, and they do not provide quantitative predictive schemes that link hydraulic and sedimentological variables

on a conceptual basis. Predicting sedimentation rates in a quantitative way using empirical relations may become unsuccessful when large floodplain areas with irregular topography are concerned, and becomes even more difficult when important boundary conditions of sedimentation change, that is, river discharge and suspended sediment concentrations change. For those cases, a process-based or deterministic model is needed.

Large-scale 1D Sedimentation Model

To assess the impact of large-scale landscaping measures along a 100-km floodplain reaches of the lower Rhine distributaries in the Netherlands on the sediment budget of the lower Rhine, Narinesingh *et al.* (1999) and Asselman and Van Wijngaarden (2002) developed a 1D floodplain sedimentation model that simulates sedimentation as a function of hydraulic conditions and different silt-sized sediment fractions. The modeling approach is based on a procedure to estimate sedimentation in settling tanks, depending on the trapping efficiency of the tank. This efficiency is a function of the ratio between the tank area and the discharge through it, which can be regarded as a measure of the residence time of the water and the sediment particles in the tank. The equation for the sediment trapping efficiency (E) reads:

$$E = 1 - \mathrm{e}^{\left(-w_\mathrm{s}\frac{A}{Q}\right)} \tag{1}$$

where w_s is the settling velocity of the suspended sediment particles, A is the surface area of the tank and Q is the discharge through the basin. Sedimentation is computed as: $S = Q_\mathrm{s}E$, with S is the sedimentation in kg/ day, Q_s is the suspended sediment load transported into the tank in kg/day. In the application of this method to river floodplains, the river was schematized as a main channel with a number of floodplain sections in connection (Figure 4). Within a floodplain section, different stream "tubes" were schematized, each representing an individual settling basin with an inlet at the upstream end and an outlet at the downstream end. The model was calibrated using sediment loss along an entire river branch at varying discharges, and by comparing the predicted sediment accumulation with the observed sediment deposition during a major flood event.

Models of Lateral Sediment Transfer

During floods, flow is fast and deep in the main channel and shallow and slow over the floodplain. The strong lateral interaction between the two flows results in a transfer of momentum as the flow is decreased in the channel and increased over the floodplain. This is evident in turbulent eddies along the interface between the main channel and the floodplain. Furthermore, suspended sediment concentrations are higher in the main channel due to its greater transporting capacity. Consequently, the interaction between the channel and the floodplain leads to lateral transfer of sediment.

Allen (1985), James (1985), and Pizzuto (1987) considered the lateral transfer and deposition of sediment particles onto the floodplain by a diffusion analogy. Pizzuto (1987) derived a mathematical model for the lateral transfer of suspended sediment in cross section from the main channel to the floodplain under conditions of steady river flow, with no current directed perpendicular to the channel, and with constant diffusivity coefficients along distance from the main channel. His model predicts, for a particular grain size, an exponential decrease in sediment deposition H with increasing distance from the main channel as:

$$H = \left(\frac{\mathrm{V}_\mathrm{s}^2}{e_z}\right) Z_0 t \left[\left(\frac{\sin h(Gn)e^{-G})}{\cosh(G)} + e^{-Gn}\right)\right] \tag{2}$$

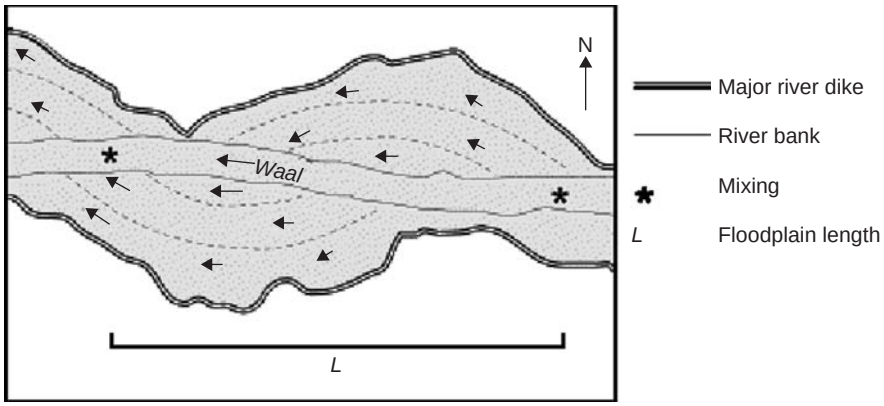


Figure 4 Embanked river floodplains of the Waal, each schematized as three parallel settling reservoirs. (Reprinted from Asselman and Van Wijngaarden, 2002. © 2002, with permission from Elsevier)

where H = deposited sediment thickness after a time t; t = time; V_s = settling velocity of the sediment, calculated using Stokes' Law; e_z = vertical sediment diffusivity; Z_0 = suspended sediment concentration; n = distance across the floodplain as a proportion of the total floodplain width W; G = a dimensionless variable that relates to the floodplain width W, sediment settling velocity and the vertical and horizontal diffusivities. G is calculated using $G = W V_s/(e_y\, e_z)^{0.5}$, where e_y = horizontal sediment diffusivity.

According to the model, the thickness of sediment deposition drops rapidly within a short distance from the main channel for floods of short duration (0.5 h). As modeled flood duration increases, the curve of deposition becomes increasingly concave upward, and reaches a constant shape after more than 10 h of modeling. Pizzuto (1987) assumed that this represented an equilibrium sediment concentration, and used the patterns resulting from the equilibrium state for comparison with field observations of thickness of overbank sediment. The model concept fairly well predicted the distribution of fine sediments, but sand-sized material was actually carried farther across the floodplain than the profile simulated by the steady-state model, unless unrealistically large diffusion coefficients were invoked (Pizzuto, 1987; Marriott, 1996). James (1985) used a numerical model to simulate the transfer of sediments from the channel to the adjacent floodplain in which he combined a diffusion analogy with an element of convection. The model was run using different grain sizes, resulting in sediment concentrations dominated by sand-sized grains near the channel, while silt-sized grains occur in gradually decreasing concentrations across the whole width of the floodplain.

In the past two decennia, different numerical models for simulating floodplain deposition of meandering channels, and avulsion-controlled alluvial architecture have been developed, all of which consider overbank deposition as one of the processes in the long-term ($\sim 10^2$–10^3 years) development of floodplains. Examples are the model of Allen (1978), who assumed deposition to be uniform over the entire floodplain, the model of meander migration and floodplain deposition of Howard (1992, 1996), and the alluvial stratigraphy model of Bridge and Leeder (1979), who adopted an empirical approach to relate overbank deposition rates to the distance from the main channel.

Howard (1992, 1996) developed a heuristic relationship for long-term overbank sedimentation as a function of local elevation and distance from the main channel in a model that simulates river meandering, as:

$$r = \left[\nu + \mu\ \mathrm{e}^{\left(\frac{-D}{\lambda}\right)} \right] \mathrm{e}^{[-\gamma (E_{\mathrm{act}} - E_{\mathrm{bed}})^W]} \qquad (3)$$

where r = long-term average overbank sedimentation, ν = a position-independent deposition rate of fine sediment, μ = overbank deposition rate of coarser sediment by overbank diffusion, λ = a characteristic diffusion/advection length scale, D = distance to the nearest channel, E_{act} = local floodplain elevation, E_{bed} = mean bed elevation, γ = elevation decay rate, and W = exponent, set to 2.0.

In the three-dimensional alluvial stratigraphy model version of Bridge and Leeder (1979) (Bridge and Mackey, 1993; Mackey and Bridge, 1995), overbank aggradation rate decreases exponentially with distance from the channel, according to:

$$r_z = a_0\ \mathrm{e}^{\left(\frac{-bz_c}{z_m}\right)} \qquad (4)$$

where r_z = time-averaged deposition at distance z from the channel-belt edge, a_0 = channel-belt aggradation rate at zero distance from the channel, z_c = cross-valley distance from the channel-belt edge, z_m = maximum floodplain distance from the channel-belt edge, and b = aggradation coefficient, defining the rate of decrease of overbank aggradation with distance from the active channel belt. The exponent b must be derived from field observations. Estimates of b vary between about 0.35 based on medium-term ($\sim$100 years) reconstructions to values of 5–10 based on single-flood observations (Törnqvist and Bridge, 2002).

Although the results have been supported by both laboratory measurements and field observations (Marriott, 1992, 1996), the ability of these models to predict and explain sediment deposition on natural floodplains has not been accurately established. It is difficult to apply such models when the channel sinuosity is high, when the flow over a floodplain is not uniform, or when this flow pattern changes under varying river discharge. In fact, the model of Pizzuto (1987) reduces the process of sediment deposition to a simple function of distance to the main channel, suspended sediment concentrations, particle fall velocity, and a diffusion constant. As these models all tended to examine floodplain sedimentation in cross section, they all predict an exponential decrease in deposition with increasing distance from the main channel, which is the expected pattern along straight river channels (Nicholas and Walling, 1997a,b).

Two-dimensional Mathematical Modeling of Overbank Flow and Sediment Deposition

The complex topography of natural floodplains substantially controls the local floodplain inundation frequency, as well as overbank flow patterns and flow velocities (Lewin and Hughes, 1980; Simm, 1995; Nicholas and Walling, 1997a, Nicholas and McLelland, 1999). Consequently, the overbank flow is generally not parallel to the main channel but has a component perpendicular to the main channel. In this situation, sediment transport over the floodplain occurs mostly by convection, and the spatial distribution of sediment deposition will be controlled by the water flow pattern. Therefore, to simulate sediment transfer and deposition, a model must be capable of predicting detailed

flow patterns and relating these to sediment transport and deposition processes. This is done in a two-stage approach with coupled models of water flow and sediment transport and deposition. Once a detailed model of the hydraulic flow pattern is obtained, the sedimentation process is determined by using the hydraulic pattern as a basis.

Water Flow

Over the past decennium, various two-dimensional hydraulic models have been developed for simulating patterns of floodwater flow, and which serve as a basis for calculating overbank deposition on natural floodplains (*see* **Chapter 141, Computer Modeling of Overbank Flows, Volume 4**). Recent overviews and evaluations of these models are reported by Hervouet and Van Haren (1996), Horritt and Bates (2002), and Nicholas and Mitchell (2003). Different approaches have been developed to model flow over the river channel and floodplain, varying from simple diffusion-wave approaches to models that solve the full set of Navier–Stokes equations of continuity of mass and momentum. Using these equations, the model calculates water depth h and flow velocity in x and y directions. The latter type of models considers two-dimensional shallow-water equations (Sain-Venant), in which depth-average values of the Navier–Stokes equations are used (Hervouet and Van Haren, 1996). An example of this type is the 2D finite-volume *Hydro2de* model (Beffa and Connell, 2001; Nicholas and Mitchell, 2003; Sweet *et al.*, 2003) that was applied to explore the relations between floodplain topography, floodwater hydraulics, and sedimentation patterns in British lowland floodplains. This model considers the depth-averaged Navier–Stokes equations of conservation of mass and momentum for shallow-water flow over a DEM.

The conservation of mass and momentum can be written as

$$\frac{\partial h}{\partial t} + \frac{\partial q}{\partial x} + \frac{\partial r}{\partial y} = 0 \tag{5}$$

$$\frac{\partial q}{\partial t} + \frac{\partial (q^2/h)}{\partial x} + \frac{\partial (qr/h)}{\partial y} + \frac{g}{2}\frac{\partial (h^2)}{\partial x} + gh\frac{\partial z}{\partial x} - \frac{1}{\rho}\frac{\partial (h\tau_{xy})}{\partial y} - \frac{1}{\rho}\frac{\partial (h\tau_{xx})}{\partial x} + \frac{\tau_{bx}}{\rho} = 0 \tag{6}$$

$$\frac{\partial q}{\partial t} + \frac{\partial (r^2/h)}{\partial y} + \frac{\partial (qr/h)}{\partial x} + \frac{g}{2}\frac{\partial (h^2)}{\partial y} + gh\frac{\partial z}{\partial y} - \frac{1}{\rho}\frac{\partial (h\tau_{yx})}{\partial x} - \frac{1}{\rho}\frac{\partial (h\tau_{yy})}{\partial y} + \frac{\tau_{by}}{\rho} = 0 \tag{7}$$

where h is flow depth, q and r are unit discharge in x and y directions, respectively, t is time, z is bed elevation, g is acceleration due to gravity, ρ is the water density and τ_{xx}, τ_{yy}, τ_{xy}, τ_{yx} are turbulent stresses and τ_{bx} and τ_{by} are bed shear stresses. Assuming turbulent normal stresses to be negligible, shear stresses are modeled using the Boussinesq approximation:

$$\tau_{xy} = \tau_{yx} = \rho \nu_t \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \tag{8}$$

where $u = q/h$, $v = r/h$, $\nu_t = ku^*h$, with u^* is the friction velocity and k is a constant of which the value is assumed to equal 0.067 for two-dimensional open-channel flows. Details of implementation of the shallow-water equations in the model and numerical solutions for these models are given in Connell *et al.* (2001), Hervouet and Van Haren (1996), and Beffa and Connell (2001).

A relatively simple method is the concept of diffusion wave as used by, for example, Nicholas and Walling (1997a) and the LISFLOOD model (Bates and De Roo 2000). In their models, water flow is described in terms of continuity of mass and momentum equations discretized over a grid of cells, assuming that the flow is simply a function of the free surface height difference between those cells. The model only represents mass transfer from the channel to the floodplain and neglects transfer of momentum. The model assumes that the flow between two cells is a function of the free surface height difference between those cells, and the equations read (Horritt and Bates, 2002):

$$\frac{\mathrm{d}h^{i,j}}{\mathrm{d}t} = \frac{Q_x^{i-1,j} - Q_x^{i,j} + Q_y^{i,j-1} - Q_y^{i,j}}{\Delta x \Delta y} \tag{9}$$

$$Q_x^{i,j} = \frac{h_{\text{flow}}^{5/3}}{n} \left(\frac{h^{i-1,j} - h^{i,j}}{\Delta x} \right)^{1/2} \Delta y \tag{10}$$

where $h_{i,j}$ is the water-free surface height at cell (i, j), Δx and Δy are the cell dimensions, n is the Manning's friction coefficient for the floodplain, and Q_x and Q_y describe the volumetric flow rates between floodplain cells. Q_y is defined analogously in equation (10). The flow depth h_{flow} is defined as the difference between the highest water surface in the two cells and the highest floodplain surface. Although this approach does not accurately represent diffuse wave propagation on the floodplain, the model is computationally simple and results have been shown to be consistent with models that faithfully consider the diffusive wave equation.

A distinction can be made between finite element and raster-based (finite difference) approaches to represent the spatial layout of the model. Finite element models operate on a mesh of nodes connected in the form of triangular elements. This allows using a finer mesh within the main channel, while the mesh is coarser over the floodplain where flow is spatially more constant. Floodplain topography is often sampled onto the mesh using nearest

neighbor interpolation from a digital terrain model (DTM), while the elevations of the channel and bank nodes are usually obtained from channel surveys (e.g. Horritt and Bates, 2002). An example of finite element models is the TELEMAC-2D model (Galland *et al.*, 1991). In raster-based models, the hydraulic equations are discretized over a grid of square or curved cells, usually based on gridded DTMs.

These models have been applied to natural river floodplains of a range of sizes, using different spatial resolutions of the model mesh or raster grid to represent the floodplain topography. Studies in 5–30-km-long floodplain reaches (e.g. Bates *et al.*, 1996; Hervouet and Van Haren, 1996; Connell *et al.*, 1998) typically were conducted using low-resolution models (10–100 m spacing of grid nodes) that represent the overall valley floor topography and channel planform, but have a poor resolution to represent small-scale characteristics of the floodplain surface topography, such as small embankments, depressions, drainage ditches, or abandoned channels. Since these topographic features are important controls of overbank inundation sequences and sediment transfer over the floodplain (Nicholas and Walling, 1997a; Middelkoop and Van der Perk, 1998), finer grids were applied in models covering floodplain sections of a few square kilometers, while recent studies (e.g. Nicholas and Mitchell, 2003) used increasingly finer resolutions (1 to 2.5 m) to represent floodplain topography in high detail.

Hydraulic roughness of the floodplain surface is an important control of flow patterns. Two types of roughness may be distinguished in hydraulic modeling (i) form-scale roughness associated with channel and floodplain topography, which can be represented when using an accordingly fine model grid; and (ii) subgrid-scale roughness, which is usually parameterized by a friction coefficient (Manning's n, Nikuradse k or Chézy coefficient). This friction coefficient is spatially variable and usually has to be determined by calibration. To reduce the number of parameters to be determined by calibration, several studies used one Manning's n value for the channel and one for the floodplain (e.g. Horritt and Bates, 2002; Nicholas and Mitchell, 2003), although Hesselink *et al.* (2003) used four different roughness values associated with different land use types on the floodplain.

Floodplain vegetation exerts considerable resistance to the overbank flow, and hydraulic roughness of vegetation is increasingly being considered in modeling studies (e.g. Fischer-Antze *et al.*, 2001; Stoesser *et al.*, 2003; Nicholas and McLelland, 2004). This is represented by a drag force exerted by a vegetative element (usually a cylindrical stem) on the flow, while the total resistance depends on vegetation height, average number of stems per unit surface area, and average stem diameter.

Boundary conditions of these models are the upstream discharge and an imposed downstream water surface elevation. Laterally, the models are closed. The models are calibrated against observed area of inundation or stage-discharge relations (e.g. Horritt and Bates, 2002). Simulations of flood events occur by dynamically simulating an entire flood wave. Alternatively, the hydraulic equations may be solved for a number of different stationary upstream discharge conditions, resulting in a set of stationary flow surfaces (Nicholas and Walling, 1997a, 1998; Middelkoop and Van der Perk, 1998). This latter approach neglects differences in discharge–stage relationships between rising and falling stages of a flood, as well as backwater ponding. The advantage is that it allows the determination of average conditions based on long-term frequency distributions of discharges.

The recent more complex models based on highly detailed DEMs allow simulating the transition between critical and subcritical flow, as well as inundation of previously dry surfaces (Nicholas and Mitchell, 2003; Hesselink *et al.*, 2003). Still, varying the spatial resolution of the calculation mesh discretization may affect the model results considerably in a complex way, and therefore may be considered at least as important as a typical calibration parameter (Hardy *et al.*, 1998).

Sediment Deposition

Sediment transport and deposition models of overbank fines are essentially based on horizontal convection-diffusion fluxes of sediment along with the overbank flow, in combination with settling velocity rates of the conveyed particles. Amounts of floodplain sedimentation S_i over a certain interval of time at each point i in the model grid are calculated using the following equation (Nicholas and Walling, 1997a; Middelkoop and Van der Perk, 1998; Sweet *et al.*, 2003):

$$S_i = \sum_{Q=Q\mathrm{bf}}^{Q\,\mathrm{max}} T_Q c_{Q,i} D_{Q,i} \tag{11}$$

where T_Q = the duration that discharge Q occurs, $c_{Q,i}$ = the suspended sediment concentration over this period of time at location i, D_{Qi} = the rate of deposition per unit of sediment concentration for this discharge at location i. The summation is done over all discharges occurring between bankfull (Q_{bf}) and the maximum discharge (Q_{max}) occurring at the site.

S_i can be determined for a flood event by subdividing the flood wave into a number of discrete discharge situations Q with associated sediment concentrations in the river, $c_{Q,r}$, with duration T_Q. Similarly, average annual deposition can be calculated by deriving T_Q for all Q from flow-duration curves, and by estimating suspended sediment concentrations $c_{Q,r}$ associated with Q from rating curves (Middelkoop and Van der Perk, 1998). The sediment deposition term $D_{Q,i}$ for discharge Q at position i can be approximated by assuming a linear dependency on

sediment concentration and particle settling velocity V_s: $D = kc_i V_s$, where k is an empirical coefficient that controls the proportion of sediment that is deposited.

Instead of using an empirical coefficient k, both depositional (D) or erosional (E) fluxes can be calculated, using the Krone and Partheniades formulae, respectively:

$$D = V_s c\left[1 - \left(\frac{u^*}{u^*_d}\right)^2\right] + c\left[\frac{UhpE}{\Delta x}\right] \quad (12)$$

$$E = M\left[\left(\frac{u^*}{u^*_e}\right)^2 - 1\right] \quad (13)$$

where u^* = critical or friction velocity at the bottom, u^*_d and u^*_e are critical shear velocities for settling and erosion respectively, and M = the Partheniades constant. The second term in the Krone equation (12) represents deposition due to sediment trapping by vegetation (Sweet *et al.*, 2003), with U = flow velocity; h = flow depth, p = proportion of the depth occupied by vegetation, E = trapping efficiency by vegetation, and Δx = grid spacing.

Concentration patterns of suspended sediment in the floodwater c_Q resulting from convection, dispersion, and deposition of suspended sediment are predicted by solving the depth-averaged mass balance equation for transport by convection and diffusion in the two horizontal dimensions for each node in the model grid (Nicholas and Walling, 1996, 1997a,b; Cancino and Neves, 1999), written in conservative form as:

$$\frac{\partial c}{\partial t} + uh\frac{\partial c}{\partial x} + vh\frac{\partial c}{\partial y} + D = \frac{\partial}{\partial x}\left(\varepsilon h\frac{\partial c}{\partial x}\right) + \frac{\partial}{\partial y}\left(\varepsilon h\frac{\partial c}{\partial y}\right) \quad (14)$$

where u and v are flow velocity components in x and y directions, h = water depth, c = depth-integrated suspended sediment concentration, ε is a horizontal mixing coefficient and D is the net rate of deposition. Under steady-state conditions, $\partial c/\partial t = 0$. Horizontal mixing coefficients can be determined using $\varepsilon = \lambda h u^*$, where λ is a constant (assigned a value of 0.17 by Nicholas and Walling (1997b)) and u^* = shear velocity given by $u^* = (ghS_f)^{1/2}$, where g = acceleration due to gravity and S_f = water surface slope. These equations must be solved in an iterative way, as D depends on the local sediment concentration.

In ponded areas, an additional amount of sediment deposition from trapped sediment occurs. Assuming that all suspended sediment in ponded areas settles, this can be calculated using (Nicholas and Walling, 1997b; Middelkoop and Van der Perk, 1998):

$$D_{p,i} = h_{p,i} c_i \quad (15)$$

where $D_{p,i}$ = deposition at location i resulting from settling of trapped sediment; $h_{p,i}$ = depth of ponded water at location i; c_i = sediment concentration in the ponded water. This amount is added to the total deposition.

The sediment characteristics that control deposition (and erosion) rates depend on grain size distribution of the suspension load. Consequently, the total suspension load must be subdivided into j sediment grain size fractions, each with different values for sediment concentrations in the water (c_j), settling velocities (V_{sj}), mixing coefficients (ε_j), and critical shear velocities for deposition and erosion (u^*_j), which results in j deposition rates D_{Qj}, that are added proportionally to the relative frequency of each sediment fraction j in the total sediment load to obtain the total deposition rate D_Q (e.g. Nicholas and Walling, 1997b). In this respect, a distinction must be made between the so-called *ultimate* grain size, which relates to the distribution of the discrete primary particles of the suspended sediment, and the *effective* grain size. Because of flocculation, settling velocity (V_s) cannot be determined from the ultimate grain size using Stokes' law, but it must be derived from estimated effective sizes of the sediment flocs (Peart and Walling, 1982; Nicholas and Walling, 1996, 1997b) (*see* **Chapter 83, Suspended Sediment Transport – Flocculation and Particle Characteristics, Volume 2**), or it must be measured *in situ* during overbank flooding (Thonon and Van der Perk, 2003).

Several model parameters are not *a priori* known, and must be determined by model calibration. This is also the case for parameters that have been determined at the point scale, but whose values may considerably change at the modeling resolution over large floodplain areas (Middelkoop and Van der Perk, 1998). Calibration and validation of sedimentation models is usually done by comparing predicted patterns of floodplain deposition with medium-term floodplain deposition rates reconstructed from ^{137}Cs measurements or heavy-metal concentrations in floodplain soil profiles (Nicholas and Walling, 1997a,b; Simm *et al.*, 1997; Sweet *et al.*, 2003), or with observed sediment deposition after a flood event (Nicholas and Walling, 1997a,b; Middelkoop and Van der Perk, 1998; Middelkoop, 2002b).

Sedimentation modeling becomes computationally expensive because the equations that determine the concentration pattern must be solved in an iterative way, while several model parameters must be determined by calibration. Alternatives that have been applied to reduce model complexity include (i) omitting the diffusion terms in the sediment transport equations for situations in which convective transport is presumed the dominant transporting mechanism to reduce the number of model parameters (Middelkoop and Van der Perk, 1998), (ii) neglecting erosion or reentrainment of sediment, or (iii) considering the

mass balance as a one-dimensional case, thereby allowing to solve the decline in sediment concentration by a noniterative means (Sweet *et al.*, 2003). A last simplification of these modeling approaches is that they do not consider the changes in topography due to deposition.

The choice for the modeling concept used and the number of grain size classes considered should be guided by the objectives of the study, the spatial scale considered, and the available data. A lack of reliable reference data at the required scale, particularly concerning small-scale turbulent flow patterns, sediment settling velocities or effective grain size distributions, or effective shear stresses for deposition may not justify the development and use of a complex, physically detailed model to simulate overbank deposition. Instead, a simpler sedimentation model may have a higher predictive capability, when it considers the dominant processes and whose model parameters can be more likely calibrated (Anderson *et al.*, 1996; Van der Perk, 1997; Middelkoop and Van der Perk, 1998). Improvements of sedimentation models in the future should aim at determination of *in situ* effective grain sizes, inherent settling rates and reentrainment of sediment, and a better treatment of interactions between floodplain vegetation and hydraulics and sedimentation.

Three-dimensional CFD Modeling of Floodplain Processes

In recent years, computational fluid dynamics (CFD) modeling of three-dimensional flow hydraulics and sediment transport has increasingly been applied to fluvial and estuarine systems. This modeling approach is indispensable to study three-dimensional aspects of flow, such as complex flow patterns in natural river channel bends, strong vertical gradients in flow velocity over rough areas, or stratified tidal flow patterns in estuaries (e.g. Cancino and Neves, 1999; Fischer-Antze *et al.*, 2001; Rodriguez *et al.*, 2004; Stoesser *et al.*, 2003; Lesser *et al.*, 2004). These models basically solve the three-dimensional Navier-Stokes equations and consider turbulence of flow, and often consider both suspended sediment transport and bottom sediment transport. Examples that have been applied to simulate channel processes are the TELEMAC-3D model (**`http://www.hrwallingford.co.uk/software/telemac.html`**), or the Delft3D model (**`http://www.wldelft.nl/soft/d3d/intro/index.html`**) (Lesser *et al.*, 2004). Still, few studies using CFD modeling have been undertaken for floodplain areas. Apart from the computational requirements, there are several reasons for this. (i) When compared to channel hydraulics, floodplain flow, and sedimentation processes can often be well predicted by two-dimensional approaches; (ii) it is difficult to specify the model boundary conditions; (iii) CFD modeling demands a good estimation of flow resistance structures of vegetation; and (iv) it is difficult to collect precise field measurements of 3D flow patterns and sediment conveyance over an inundated floodplain, necessary to validate the model.

Nicholas and McLelland (2004) simulated 3D overbank flow patterns on a 1-km-wide section of the River Culm (UK) natural floodplain. They showed the occurrence of complex 3D overbank flow patterns that are stage-dependent and topographically driven. Stoesser *et al.* (2003) were able to predict overbank flow velocities on vegetated floodplains using a 3D numerical model. Both studies indicated that compared to traditional 2D models, the 3D CDF approach may considerably reduce the modeling uncertainty associated with calibration of flow resistance and provide vertical flow profiles, in spite of the above-mentioned difficulties.

OVERBANK SEDIMENTATION ON THE EMBANKED FLOODPLAIN OF THE LOWER RHINE

Introduction

The lower Rhine River in the Netherlands (Figure 5) has been embanked by artificial levees since late Middle Ages. In subsequent centuries, people have increasingly affected the main channels and the floodplain by cutting-off meanders, modifying the bifurcations of the Rhine distributaries in the eighteenth century, and narrowing the main channel to a standard width, and fixing the banks

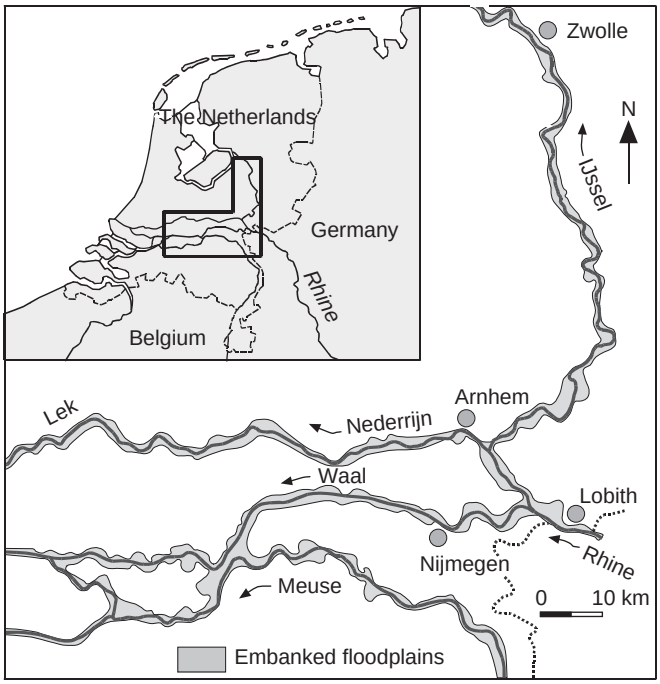


Figure 5 The embanked floodplains of the lower Rhine and Meuse in the Netherlands

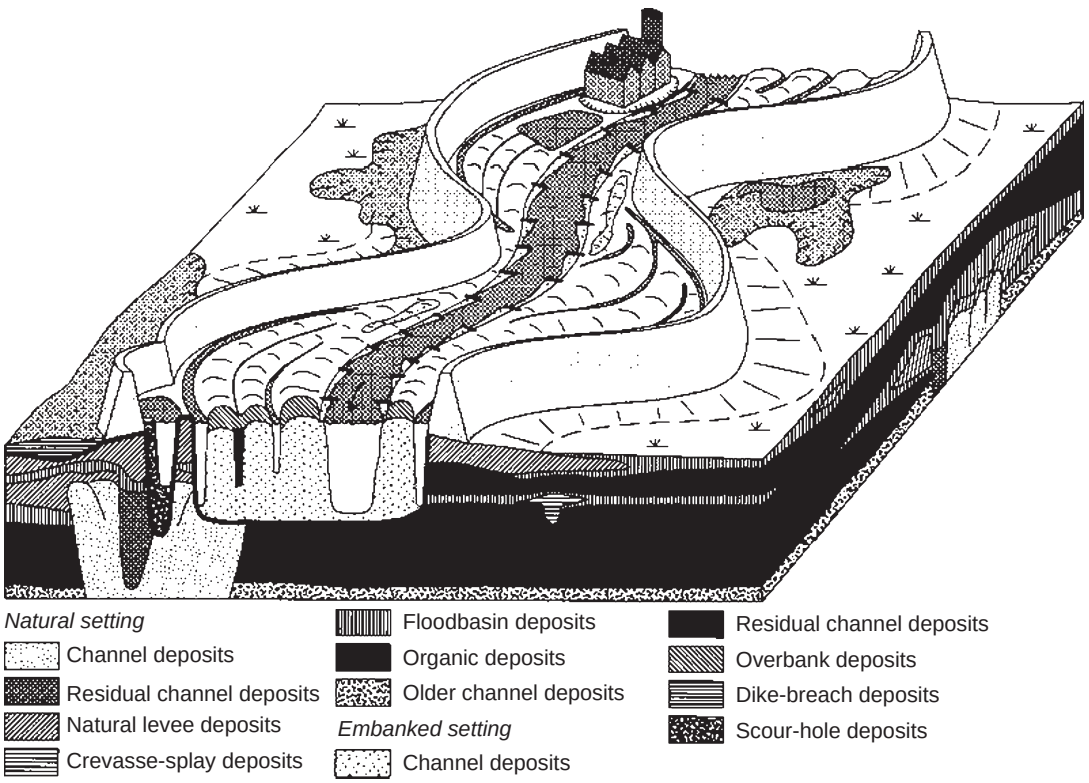


Figure 6 Architectural elements of the embanked floodplain of the lower Rhine (*Source*: Hesselink, 2002)

with a regular array of groins in the late nineteenth century. The floodplains have been extensively used for agriculture, and several floodplain sections have been protected from minor floods by small embankments ("minor dikes"). To date, the sedimentological structure of the embanked Rhine-Meuse floodplains consists of ridges of sandy channel-bed sediments overlaid by a 1–2 m thick body of overbank sediments, intersected with silted-up swales and residual channels (Figure 6). Using different methods, overbank deposition on the embanked floodplain of the lower Rhine was analyzed over various timescales: historic (centuries), the past century, and contemporary deposition rates.

Average Sedimentation Rates During the Past Centuries

Before 1850, the Rhine distributaries had low-sinuous meandering channels that slowly migrated in a downstream direction as the river dikes prevented lateral expansion of the meanders. Groins placed along the inner bends artificially enhanced lateral accretion, resulting in land reclamation from the river. The resulting growth of newly formed floodplain sections was mapped in detail, for example, to delineate new accretions of disputed ownership and to determine tax rates (Hesselink, 2002). Using old river maps made since the sixteenth century, the age of different floodplain sections and thus of the beginning of overbank sedimentation was determined (Middelkoop, 1997). From these the average deposition rates of the overbank sediments were determined. Figure 7 shows average overbank deposition rates, for floodplains of different ages, varying between 5 and 15 mm year^{-1} for the past 100–300 years. The decreasing trend of average sedimentation rates with increasing floodplain age indicates that deposition rates decrease

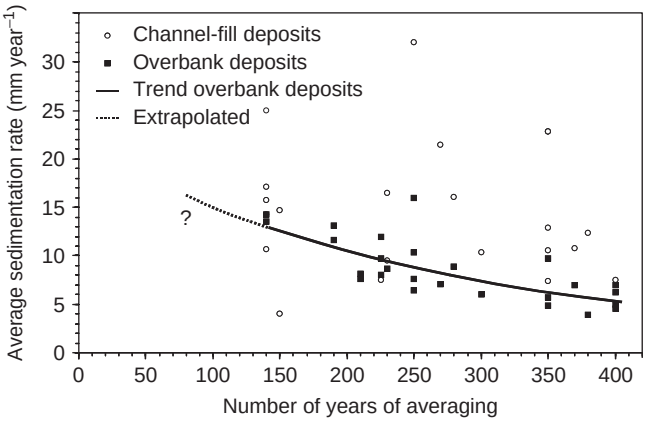


Figure 7 Estimated average floodplain sedimentation rates for varying floodplain ages, based on historic data

with time, as the floodplain elevation increases due to sedimentation.

Sedimentation Rates over the Past Decennia

Because of the heavy metal pollution of the river Rhine during the past century, the embanked floodplains of the lower river Rhine in the Netherlands contain large amounts of heavy metals. Deposition of heavy metals has occurred since the end of the nineteenth century. Periods of maximum pollution occurred in the 1930s and 1960s, when Cu, Pb, and Zn concentrations were about 6 to 10 times higher than background values (Figure 8) (Middelkoop, 2000).

Depending on local sedimentation rates and changing pollution trends in the past, the metal pollution varies greatly between different floodplain sections, as well as vertically within the floodplain soil profiles (Figures 9 and 10). Maximum metal concentrations in floodplain soils vary from 30 to 130 mg kg^{-1} for Cu, from 70 to 490 mg kg^{-1} for Pb, and from 170 to 1450 mg kg^{-1} for Zn. Although the investigated profiles of the lower Rhine floodplains reflect similar pollution trends, several differences between the profiles can be observed:

- Total metal contents and maximum metal concentrations in the Waal floodplain are generally higher than in the Nederrijn-Lek and IJssel floodplains;
- Low floodplain sections along the Waal that are not bordered by a minor river dike (e.g. Klompenwaard and Variksche Plaat) are contaminated by heavy metals to a great depth, usually more than 1 m below surface. Here, the metal concentrations in the upper parts of the profiles are considerably lower than the concentrations at several dm depth;
- The metal pollution is much greater in the natural levees and at a short distance from the main channel, than in the distal parts or behind a minor dike. These differences are demonstrated in the profiles of Huissen, Amerongen, and Vorchten. Metal accumulations in depressions are slightly greater than at nearby elevated sites.

The metal profiles measured in these soil profiles were used as a tracer to estimate average floodplain sedimentation rates over the past 150 years by comparing the concentration profiles in the floodplain soils with the chronology of the metal pollution of the river Rhine sediment. Reconstruction of sedimentation rates was done with a 1D simulation model of floodplain sedimentation with contaminated sediments, from which deposition rates were determined using an inverse modeling calibration procedure (Middelkoop, 2002a).

The model results for the heavy metal profiles are summarized in Table 1. Average sedimentation rates over the past century have varied between 0.18 and 11.6 mm year^{-1}. Estimation errors of the sedimentation rates were of the order of 5% (for high sedimentation rates) to 70% (for the lowest sedimentation rate) of the mean value. Lowest sedimentation rates (0.2 to 1 mm year^{-1}) were found along the rivers Nederrijn-Lek and IJssel, at sites behind a minor dike or natural levee. Near the channel banks, sedimentation rates are 1 to 3 mm year^{-1}. Along the river Waal, the lowest values were obtained at sites behind a minor dike (1.2 to 3 mm year^{-1}). Sedimentation rates for sites on natural levees or close to the river channel vary between 3 and 7 mm year^{-1}. The highest sedimentation rates (>10 mm year^{-1}) have occurred at the low floodplains Variksche Plaat and Klompenwaard. Thus, sedimentation rates on the Waal floodplain have been generally higher than those in the Nederrijn-Lek and IJssel

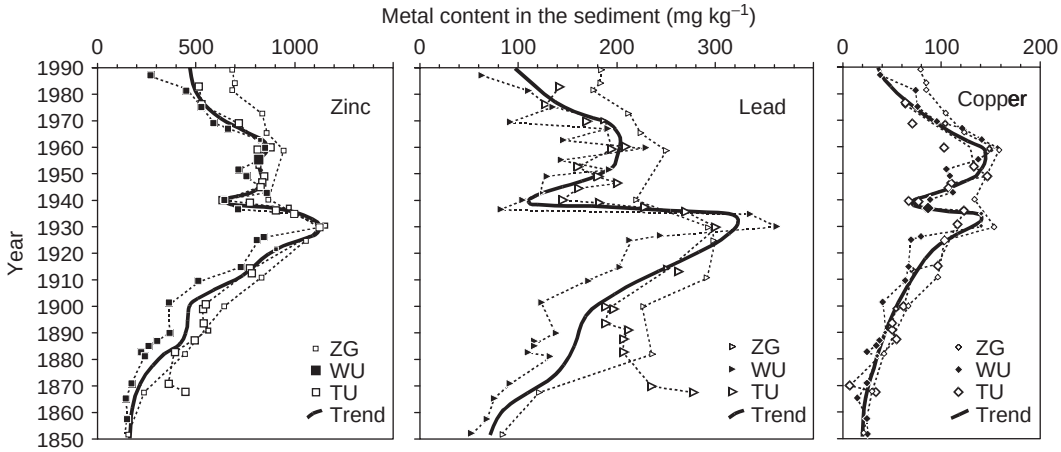


Figure 8 The temporal trend in pollution of the lower River Rhine over the past 150 years, reconstructed on the basis of metal concentrations in sediments from small ponds within the floodplain area

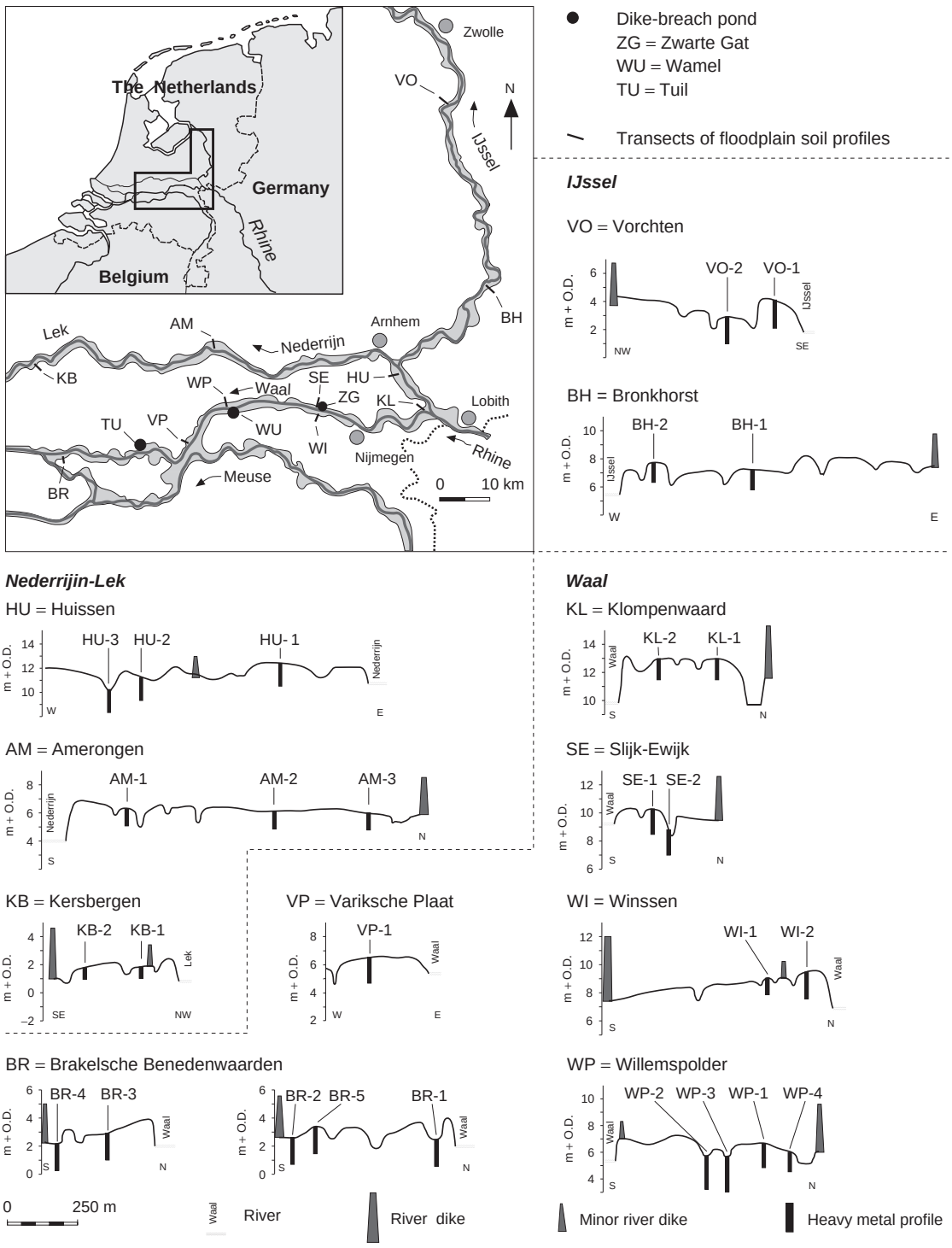


Figure 9 Location of the heavy metal profiles in the floodplain of the lower Rhine branches

floodplains. Within a floodplain section, sedimentation rates decrease with increasing distance from the river channel, but may be higher in local depressions, while a minor dike may reduce the sedimentation rate by a factor 2 to 3.

Contemporary Sedimentation Rates

Deposition During Flood Events

Contemporary floodplain sedimentation rates were studied using sediment traps placed during two low-magnitude

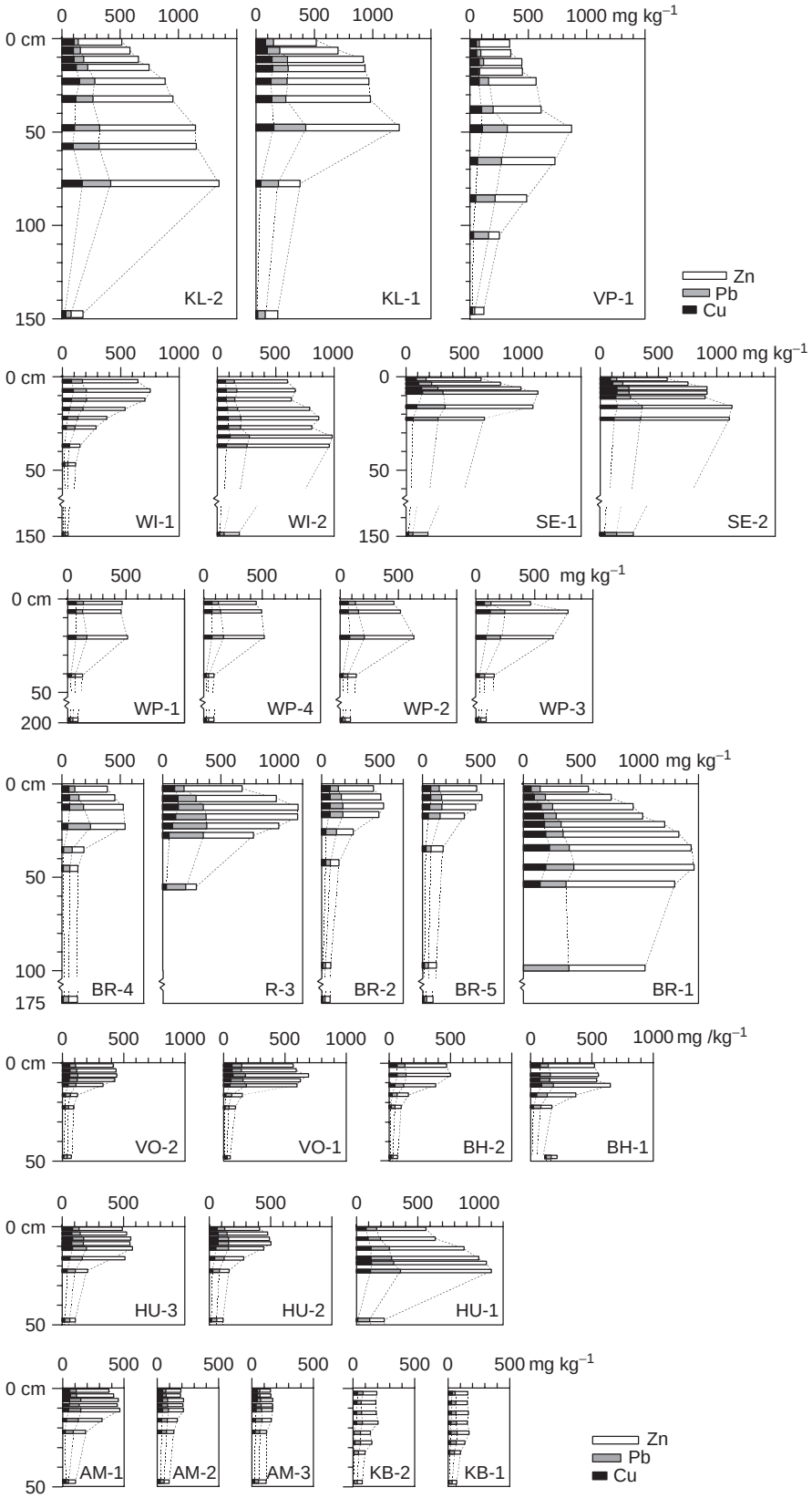


Figure 10 Heavy metal profiles in the Lower Rhine floodplains

Table 1 Site characteristics of the investigated soil profiles and reconstructed sedimentation rates

	Code	Site description	Avg. flood time	Dist. from channel bank	Sedimentation rate[a]	
			(days year^{-1})	(m)	mm year^{-1}	kg m^{-2} year^{-1}
1	AM-1	Natural levee	1.5	150	0.99	1.20
2	AM-2	Central part of floodplain	1.5	550	0.35	0.42
3	AM-3	Central part of floodplain	1.5	790	0.18	0.22
4	KB-2	Central part of floodplain, behind minor dike	1.0	60	0.27	0.32
5	KB-3	Central part of floodplain	1.0	170	0.19	0.23
6	HU-1	Natural levee	3.8	170	3.30	3.96
7	HU-2	Behind minor dike	2.5	400	0.95	1.14
8	HU-3	Behind minor dike, depression	2.5	430	1.23	1.48
9	BH-1	Central part of floodplain	6	330	1.12	1.34
10	BH-2	Natural levee	6	20	0.73	0.88
11	VO-1	Natural levee	8	50	1.04	1.25
12	VO-2	Central part of floodplain	8	190	0.64	0.77
13	KL-1	Central part of low floodplain	8	260	7.26	8.71
14	KL-2	Part of low floodplain	8	50	11.55	13.86
15	WI-4	Behind minor dike, elevated area	1.8	220	2.50	3.00
16	WI-5	Natural levee	2.5	20	5.0	6.00
17	SE-1	Behind natural levee, elevated area	7	50	3.0	3.60
18	SE-2	Behind natural levee, depression	7	180	3.15	3.78
19	WP-1	Central part of floodplain, behind minor dike	2.2	670	1.86	2.23
20	WP-2	Central part of floodplain, behind minor dike	2.2	250	2.52	3.02
21	WP-3	Depression, behind minor dike	2.2	360	2.75	3.30
22	WP-4	Depression, behind minor dike	2.2	610	1.85	2.22
23	VP-1	Central part of low floodplain	5.4	380	10.45	12.54
24	BR-1	Depression behind natural levee	4.4	10	7.00	8.40
25	BR-2	Central part of floodplain, behind minor dike	<1	480	1.38	1.66
26	BR-3	Central part of floodplain, behind minor dike	1.8	210	3.50	4.20
27	BR-4	Central part of floodplain, behind minor dike	<1	360	1.93	2.32
28	BR-5	Central part of floodplain, behind minor dike	<1	480	1.20	1.44

[a]Conversion from mm to kg m^{-2} has been based on a soil bulk density of 1.2 kg dm^{-3}.

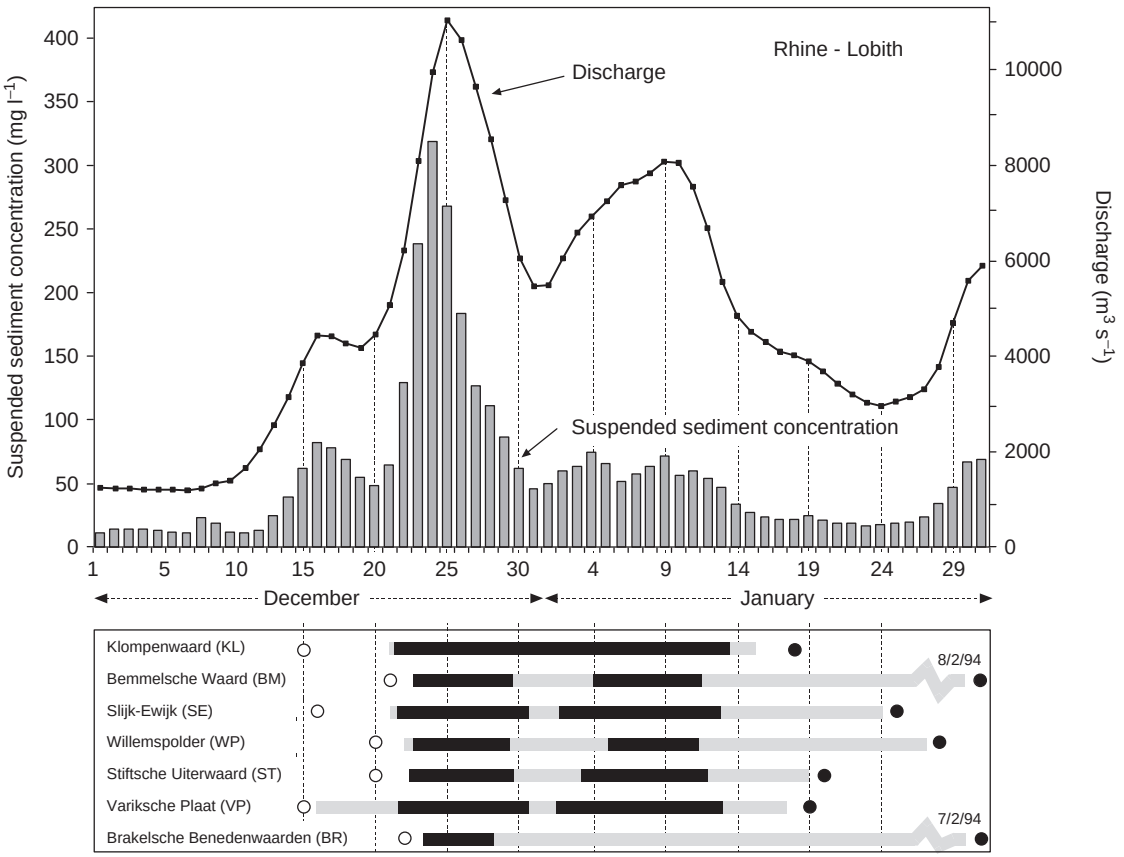


Figure 11 Rhine discharge and suspended sediment load of the Rhine at the German-Dutch border during the flood of December 1993. Below the graph it is for each investigated floodplain section: the date the sediment traps were placed (open circle); the period of inundation (light grey bar); the period of sediment conveyance over the entire floodplain (black bar); the date the sediment trap was collected (black circle)

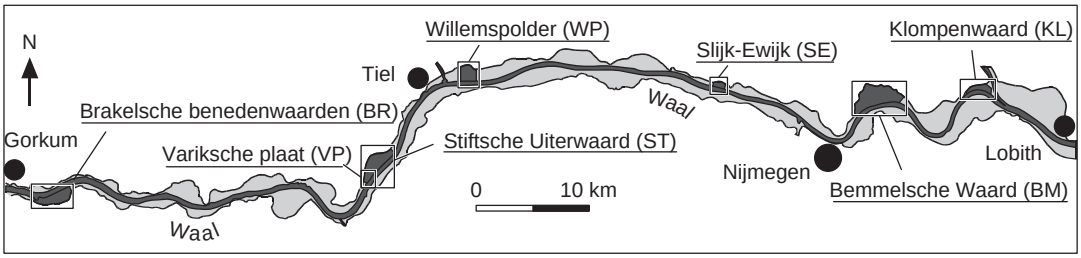


Figure 12 Location of the sampled floodplain sections during the December 1993 flood

events (April 1994 and January 1993) and two high-magnitude events (December 1993 (Figure 11), January 1994). To document spatial patterns of the deposition as well as total deposition per section, the sampling results have been interpolated using universal block kriging to a 10×10-m grid (Middelkoop and Asselman, 1998).

The deposition during the 1993 flood was investigated in detail using about 800 sediment traps placed in different sections (Figure 12). Figures 13 and 14 show typical examples of deposition patterns that resulted from the flood. During the flood, two types of overbank sediment were formed: sandy bed load material and silt- and clay-sized wash-load material. Deposition of sand was highly variable, and formed locally as elongated sand sheets along the natural levees. Sediment deposition generally decreased with increasing distance from the river channel. In sections with a natural levee, such as the Klompenwaard, deposition amounts dropped rapidly with distance from the natural levee. At a distance of about 50 to 100 m from the levee, the deposition amounts level out to a constant

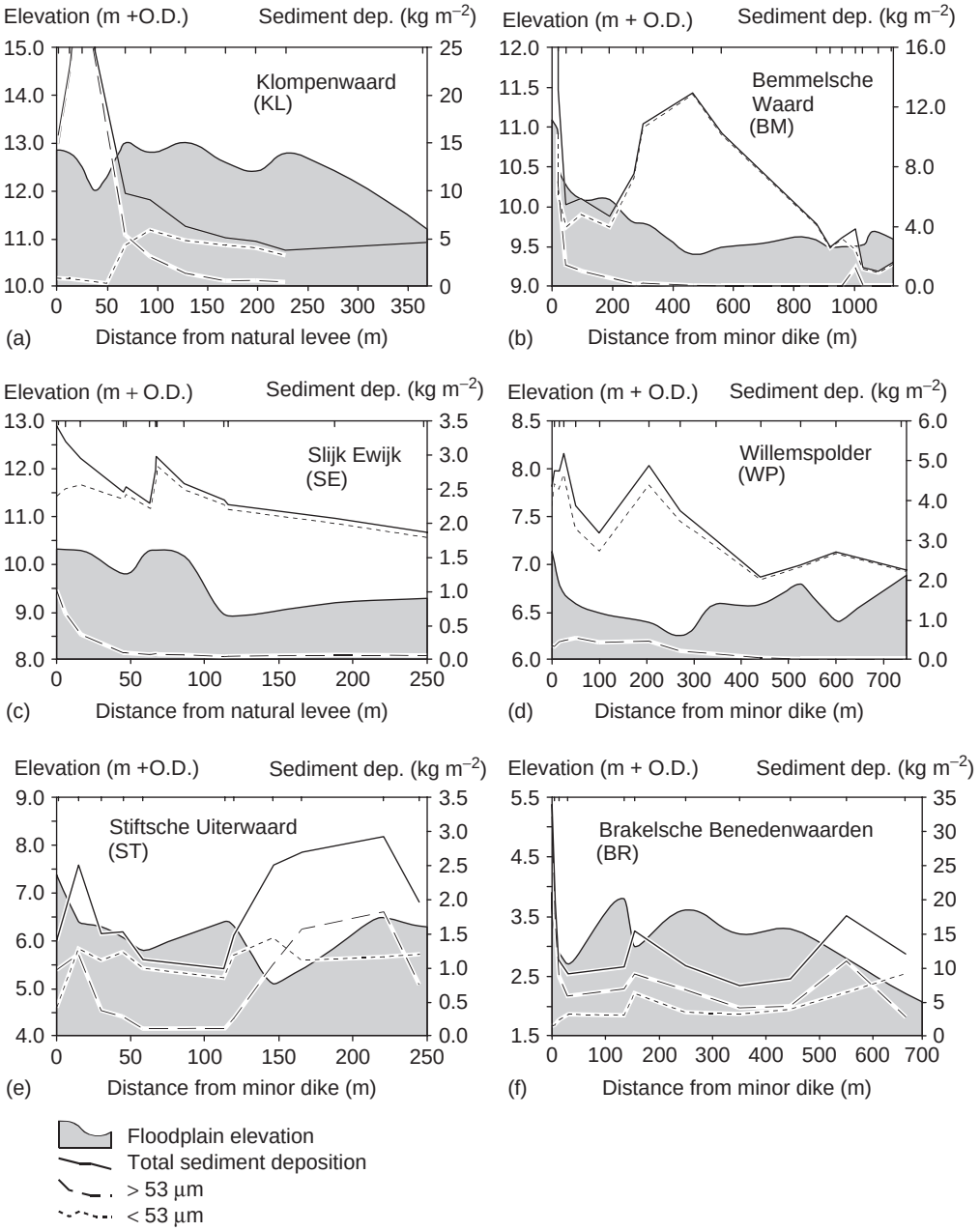


Figure 13 Floodplain elevation and overbank deposition in December 1993 on transects across different sections of the Waal floodplain. Total deposition, as well as sand and silt/clay-sized sediment deposition are shown

value. This trend mainly depends on the sand deposition, as the amounts of silt and clay remained almost constant with varying distance from the main channel when little relief is present. Percentages of sand decreased from about 90% at the levee to 15–5% in the central parts of the floodplain. Deposition amounts in depression were 50–100% larger than on more elevated locations, but varied within residual channels, depending on the prevailing current velocities. Minor embankments bordering several floodplain sections, such as the Bemmelsche Waard strongly influence the deposition pattern. Maximum sedimentation occurred at the location where the water flows over the minor river dike, while deposition amounts generally decreased with distance from the minor dike. After recession of the flood, an estimated additional amount on the order of 100–200 g m^{-2} sediment had settled from the ponded water behind the minor embankment. Local elevation differences are of minor importance when a minor dike isolates the floodplain from the main channel.

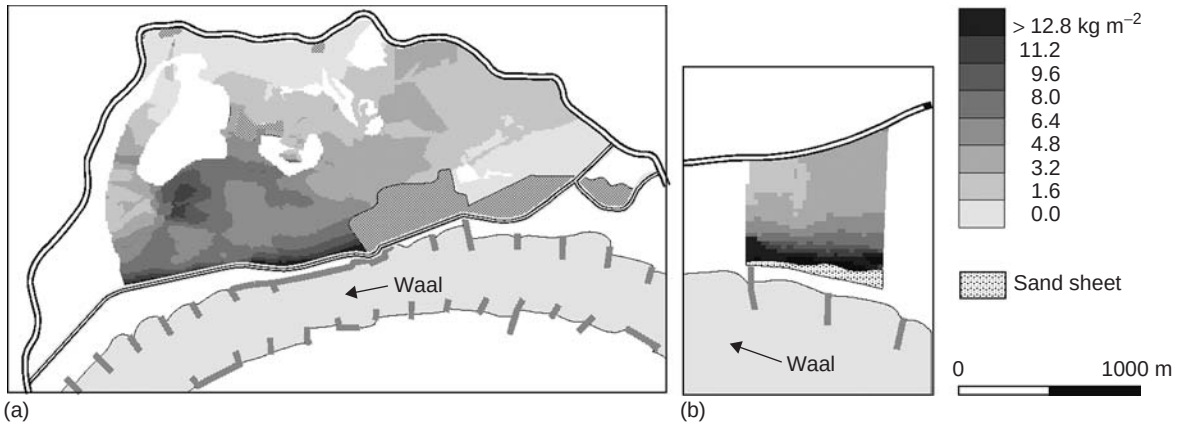


Figure 14 Floodplain deposition in December 1993 in the Bemmelsche Waard (a) and Klompenwaard (b)

Table 2 Overbank deposition during the 1993 flood on different sections of the Waal floodplain

Floodplain section	Sediment deposition		Inundation time (days)	Deposition rate	
	Total kg m^{-2}	$<53\,\mu$m kg m^{-2}		Total kg m^{-2} day^{-1}	$<53\,\mu$m kg m^{-2} day^{-1}
KL	6.61	3.98	23	0.29	0.17
BM	3.86	3.74	15	0.26	0.25
SE	2.24	2.20	20	0.11	0.11
WP	2.58	2.50	12	0.21	0.21
ST[a]	3.57	1.20	15	0.24	0.08
VP	2.60	2.15	21	0.12	0.10
BR[a]	6.03	1.50	8	0.75	0.19

[a] = *without sediment from local source.*

For each floodplain section the average deposition of sand, and of silt and clay-sized material were determined separately (Table 2). Average deposition of silt and clay-sized overbank fines varied between 1.20 and 4.0 kg m^{-2} in the investigated sections. During the December 1993 flood the estimated total accumulation of suspended sediment on the entire Waal floodplain was 0.24 Mton, which is about 7.7% of the average annual load of suspended sediment transported by the river Rhine at Lobith. During the flood, the total trapping efficiency of the Waal floodplain for suspended sediment was about 19%.

Ten Brinke *et al.* (1998) independently quantified the deposition of sand on the channel banks of the lower Rhine distributaries during the same high-magnitude floods. Sand was deposited as sheets or oblong crevasse-like spays intermittently all along the entire branches, with average thickness of 4–7 cm and covering 1–4% of the floodplain area. The deposition pattern showed that convection had been the dominant transporting mechanism, while substantial deposition had occurred along the inner bends. Ten Brinke *et al.* (1998) mention helical current as an important mechanism, while locally stronger currents from the main channel onto the floodplain over lower parts of a levee also contributed to the overbank deposition. The total volume was on the order of 250 000 m^3, equivalent to an overall transfer of 1 m^3 channel-bed sediment per running meter bank onto the floodplain.

Influence of Flood Magnitude

Figures 15 and 16 show the deposition of sand and overbank fines on the Variksche Plaat floodplain section resulting from floods of different magnitudes (Asselman and Middelkoop, 1998). Deposition was determined on four parts within this section with different elevation and, hence, different inundation times per flood. Figure 16(a) shows total deposition per flood for each inundation zone within the floodplain and Figure 16(b) shows the corresponding deposition rates per day of inundation. During the low-magnitude floods (April 1994 and January 1993), total deposition was the highest in those parts that experienced deposition for a longer time. However, during the high-magnitude floods (December 1993 and January 1995), less sediment was deposited on the lower parts. Sedimentation (deposition amount per day of inundation, Figure 16(b)) is less efficient in low-lying areas than higher parts, in spite of the longer inundation. Also, when compared with

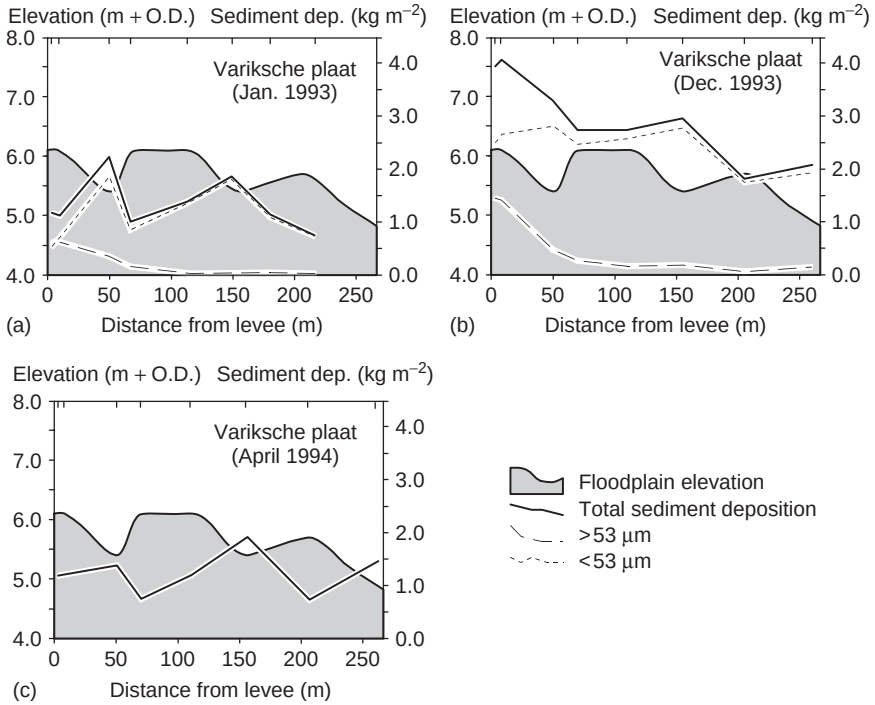


Figure 15 Floodplain elevation and overbank deposition on transects across the Variksche Plaat (VP) section of the Waal floodplain during different floods. Total deposition, as well as sand and silt/clay-sized sediment deposition are shown

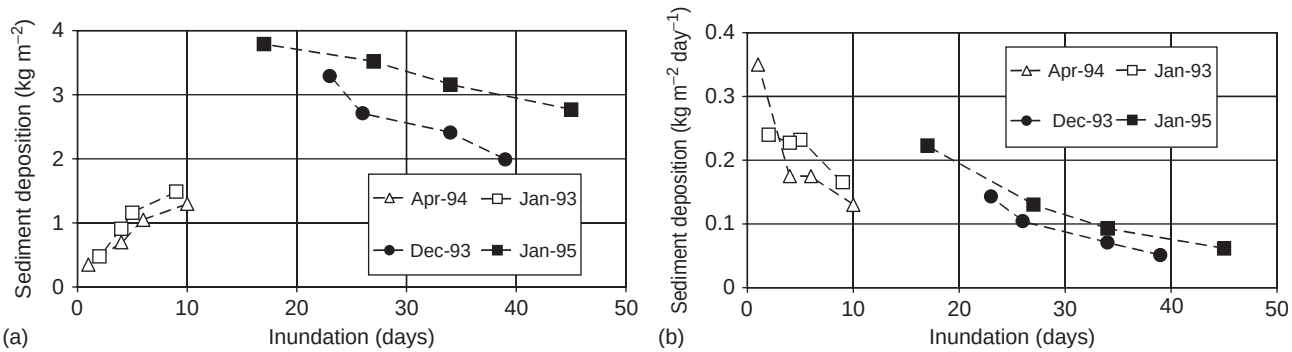


Figure 16 Overbank deposition on different parts of the Variksche Plaat (Waal floodplain) during four flood events of different magnitudes. Deposition was determined on four parts within this section with different elevation and, hence, different inundation times per flood. (a) shows total deposition per flood for each inundation zone within the floodplain, (b) shows the accordingly deposition rates per day of inundation

the amount of sediment transported through the river, the amount of sediment deposition increases less than proportionally for higher floods. The results demonstrate that *total sediment deposition* on this floodplain section increases with flood magnitude with the increase of sand deposition being considerably larger than for silt and clay. However, sediment-trapping efficiency in such a low-lying floodplain decreases at high discharge because flow velocities become too high for the sediment to settle. *Patterns* of sediment deposition also change with varying shape and magnitude of the flood wave. During minor floods, large accumulations of fine sediments occur in low areas and depressions. High-magnitude floods result in a pattern dominated by sand deposition along the main channel.

Annual Average Sediment Deposition Rates

As only a limited number of floods could be sampled, average annual sediment deposition was determined for two sections of the Waal floodplain using a sedimentation model, SEDIFLUX. This model uses water flow patterns calculated by a 2D hydrodynamic model as input. (Middelkoop and Van der Perk, 1998). As an example,

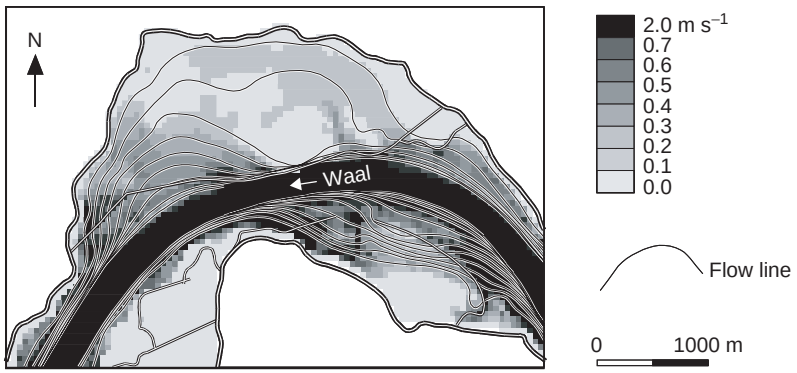


Figure 17 Flow pattern over the Bemmelsche Waard during maximum stage of overbank flooding

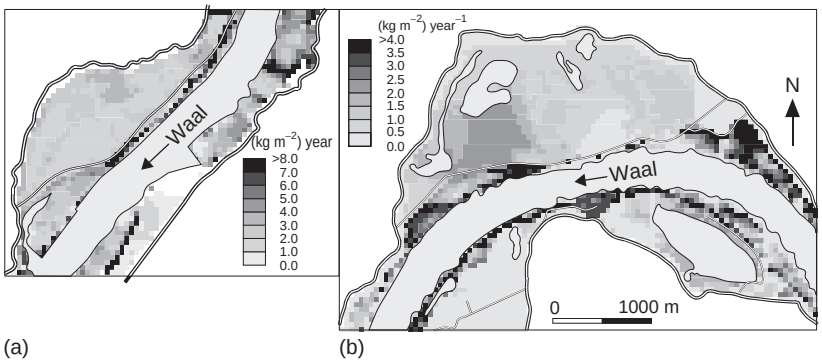


Figure 18 Annual average deposition rates on the Stiftsche Uiterwaard(a) and Bemmelsche Waard (b), calculated using the SEDIFLUX model (Middelkoop, 2002)

Figure 17 shows a typical flow pattern that occurs during inundation of the Bemmelsche Waard, which is bordered by a minor dike. Using this raster-based model, sediment deposition was calculated for a set of steady-state flow conditions resulting from a stepwise increasing river discharge, and using standardized concentrations of the suspension load in the main river channel. By multiplying the calculated deposition rates obtained for these different flow conditions in the river with the duration of each situation and with the associated suspended sediment concentration in the river, the total deposition during the high-magnitude, December 1993 flood event was obtained. The SEDIFLUX model was calibrated by varying the critical shear stress for settling and average settling velocity of the sediment particles. Calibration and validation was done by comparing calculated sediment deposition for the 1993 flood with raster maps of observed deposition, interpolated using universal block kriging. Subsequently, annual average overbank deposition was calculated using the average flow-duration curve of the Rhine and a rating curve to determine the corresponding suspended sediment concentrations.

The model results (Figure 18) show that annual average sedimentation rates in the Bemmelsche Waard are between 0.9 and 1.1 $kg\,m^{-2}\,year^{-1}$. In the zone behind the lowest part of the minor dike, the sedimentation rate is much higher: 2.2 to 2.4 $kg\,m^{-2}\,year^{-1}$. Sedimentation rates are lowest in the distal floodplain parts near the river dike: 0.6 to 0.7 $kg\,m^{-2}\,year^{-1}$. In the area between the minor dike and the river channel, the average sedimentation rate is more than 4 $kg\,m^{-2}\,year^{-1}$. The sediment deposition caused by ponding behind the minor dike may contribute about 25 to 50% to the total deposition during a minor flood. During high-magnitude floods, this contribution is less than 10%. Average annual sediment deposition within the Stiftsche Uiterwaard (Figure 18(a)) is generally larger than that within the Bemmelsche Waard (Figure 18(b)). Behind the natural levee, the sedimentation rate varies from 1.5 to 2.5 $kg\,m^{-2}\,year^{-1}$. In the central area, this is about 1.5 to 2.0 $kg\,m^{-2}\,year^{-1}$. In the upstream part of the Variksche Plaat, maximum average sedimentation rates are about 3.0 to 3.5 $kg\,m^{-2}\,year^{-1}$. In the downstream part, less sedimentation occurs (1.5 to 3.0 $kg\,m^{-2}\,year^{-1}$). In parts of this downstream area, however, sedimentation also results from back-ponding in depressions and it already occurs during lower discharges than those considered by the model.

Effective Discharge for Sedimentation

The local sediment deposition rate that occurs at a certain river discharge for a standard suspended sediment

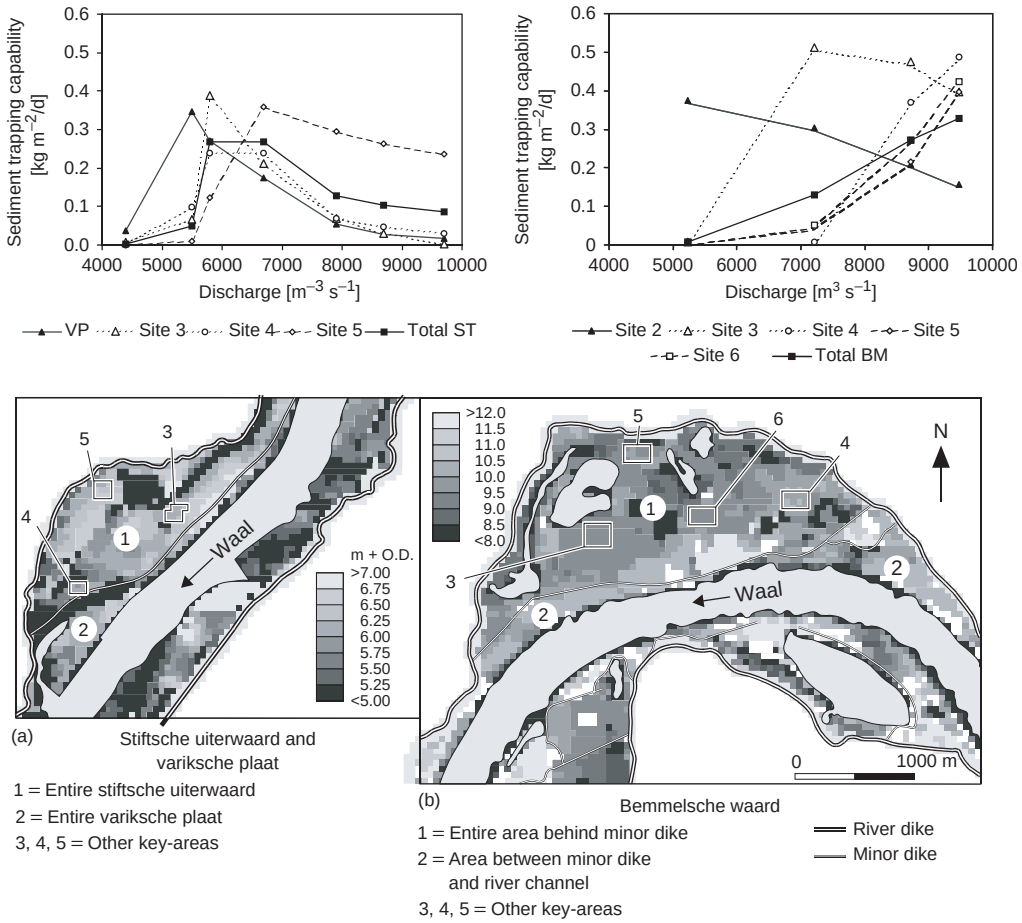


Figure 19 (a) and (b): Sediment trapping capability curves for different parts of the Bemmelsche Waard and Stiftsche Uiterwaard along the Waal river. The locations of the key areas are shown in the figure below the graphs

concentration in the river determines the sediment trapping capability (STC) of a floodplain section (Middelkoop *et al.*, 2002). The deposition rates plotted against river discharge makes up an STC curve that describes how deposition due to sediment trapping varies with river discharge, regardless of the frequency of occurrence of discharge or suspended sediment concentrations in the river. Thus, differences in STC curves are determined only by floodplain topography. STC curves were determined using the SEDIFLUX model after plotting for a site of interest, the calculated deposition rates for the standard sediment concentration in the river against increasing river discharge. The STC curves for the Bemmelsche Waard and Stiftsche Uiterwaard (Figure 19) show that the relative amounts of sediment deposited from the overbank flow on the floodplain are spatially variable, and that they also vary with discharge. After multiplication of these results with the suspended sediment concentrations associated with the varying discharge, the effective discharge for sedimentation is obtained, that is, the discharge at which in the long-term maximum sedimentation occurs (Asselman, 1999). Relatively low discharges are most effective for overbank deposition on the lower parts (i.e. Variksche Plaat, and between the main channel and minor embankment of the Bemmelsche Waard, sites VP and BM-1 in Figure 19). In contrast, high discharges contribute more to the long-term deposition in the distal floodplain parts, particularly behind the high minor dike of the Bemmelsche Waard (Figure 19b, site BM-5). These results confirm the observation that deposition amounts on the Variksche Plaat did not proportionally increase with flood magnitude.

Heavy Metal Deposition

Along with the overbank fines, particulate-bound heavy metals are transported through the Rhine and accumulate on the floodplain. Average heavy metal concentrations in Rhine sediment measured during the December 1993 flood at Lobith (German-Dutch border) were about 1.6 mg kg^{-1} for Cd, 64 mg kg^{-1} for Cu, 90 mg kg^{-1} for Pb and 400 mg kg^{-1} for Zn. These concentrations were lower than those during the preceding period of low discharge. By determining the metal concentrations in the overbank

deposits collected using sediment traps during the same flood, the spatial patterns of heavy metal deposition on the floodplain were analyzed. The results for two sections are discussed here.

In the Klompenwaard, metal concentrations in the overbank deposits strongly increase with increasing distance from the main channel (Figure 20). This trend coincides with increasing percentages of clay (% $<2\,\mu$m) and organic matter. This relationship clearly illustrates the preferential bonding of the trace metals to clay particles and organic matter. However, because of the large deposition along the channel banks, metal deposition was considerably higher at short distances from the river (about 5.0 g Zn per sq.m) than in the distal parts of the floodplain (about 1.5 g Zn per sq. m). A different pattern was found in the Bemmelsche Waard (Figure 21), which is bordered by a minor dike, where high metal concentrations occurred in the fine sediment deposited in the central parts (up to 5.0 g Zn per sq. m). Because of the large sediment deposition, metal deposition in the zone behind the overflow section of the minor dike is highest. These results demonstrate that the patterns of metal deposition during a flood event are strongly controlled by the patterns of sediment deposition, as well as by the *ultimate* grain sizes and organic matter contents in the deposited sediment. Both show major gradients within the floodplain, associated with the flow patterns from which settling of the sediment occurred.

Sensitivity to Future Changes in Upstream Catchment Conditions

Changes in climate and land use in the upstream basin may affect deposition rates on the floodplains of the lower river reaches. In a sensitivity study for the lower Rhine floodplains, the Rhine discharge regime and the suspended sediment rating curves were adapted according to different scenarios of climate change and land use development (i.e. reduction in arable land) in the Rhine basin (Asselman*et al.*, 2003), Table 3. These were used as inputs for the SEDIFLUX model to determine annual sedimentation rates under changed climatic and land use conditions. In this study, surface erosion due to rainfall and snowmelt was assumed the primary source of suspended load in the Rhine. The model results indicate that a reduction in the area of arable land leads to a considerable reduction in sediment supply from the upstream basin, resulting in a considerable reduction in annual sediment load in the lower Rhine. Consequently, floodplain deposition rates decrease by about 45%. Climate change and the associated higher winter precipitation and more intense summer rainfall is expected to cause an increased sediment yield due to rainfall erosion and an increase in peak flows of the Rhine. As a consequence, larger proportions of the annual sediment load are transported at higher discharges, when the sediment trapping efficiency of the floodplain becomes lower due to the higher current velocities. This causes only slightly increased floodplain sedimentation rates. The effects of reduced areas of sediment production, intensified erosion, and larger peak flows due to land use and climate change partly counterbalance one another. Depending on the areas of land use change involved and the assumed climate change, this may lead to a net decrease of floodplain deposition rates or (in case of the upper climate scenario) increased deposition rates (Table 3).

PATTERNS, PROCESSES, AND CONTROLS OF FLOODPLAIN SEDIMENTATION

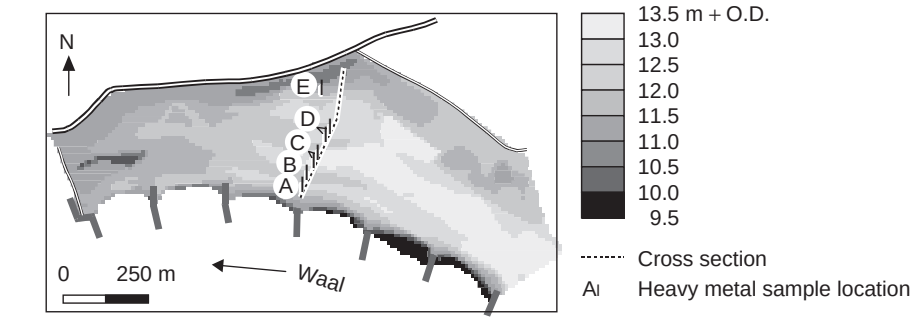
Deposition Patterns

The average annual amount of overbank deposition on a floodplain is primarily determined by the average amount of sediment that is carried over the floodplain annually and the deposition rate from the overbank flow. The first factor depends on the average yearly floodplain inundation time controlled by the hydrological regime of the river, channel geometry, and floodplain elevation, on the amount of suspended sediment in the river, and on the flow patterns over the floodplain. Deposition rates on the floodplain are determined by the local sediment concentrations in the overbank flow, grain size characteristics and settling velocity of the entrained particles, and the hydraulic patterns of the overbank flow.

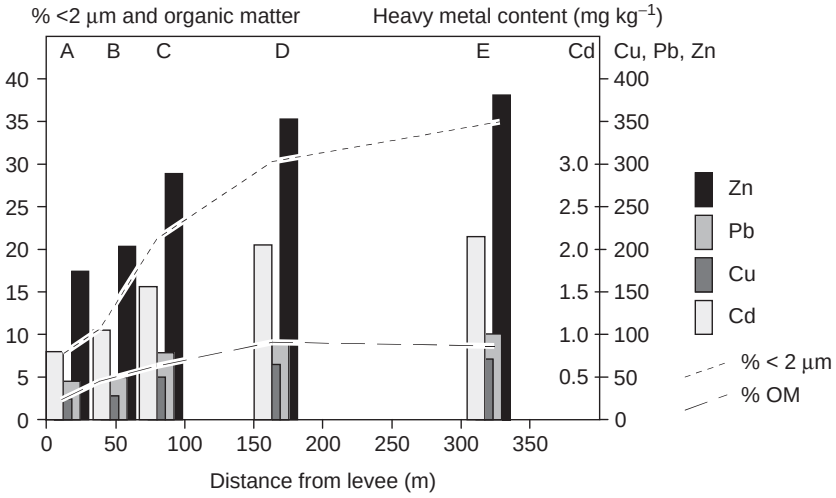
Overbank Flow

In the course of a flood event, complex flow patterns of the floodwaters occur over the floodplain, which may considerably change with the rising and falling floodwater stages. The following phases can be distinguished:

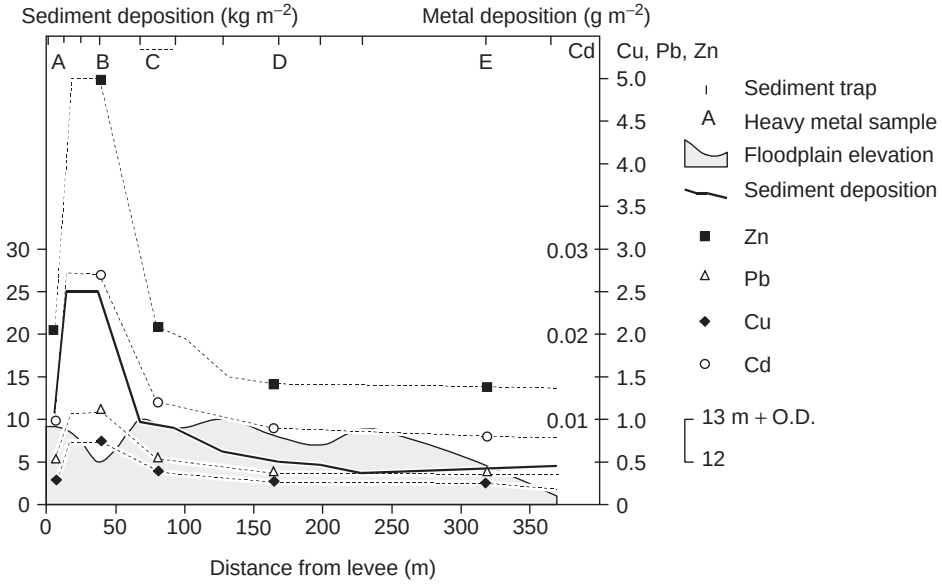
1. During the early stages of flooding, inundation patterns are largely the product of passive backwater ponding in drainage ditches and other low-lying areas near the river channel, as the levees still prevent the floodwaters to enter the floodplain.
2. With increasing discharge, inundation progresses via depressions of chutes, old channels, and ditches, from which the adjacent floodplain starts to inundate. Once out of bank, floodwaters are prevented from returning to the main channel by the natural levees, and they are diverted in a downstream direction over the floodplain, and concentrate in depressions or ditches along which they return to the main channel further downstream. Inundated areas become interconnected so that a number of distinct flow regions arise. Floodplain sections that are separated by minor embankments from the main channel do not inundate yet.
3. As the stage advances further, the flow overtops the natural levees almost entirely, while the lower sections of minor embankments are overtopped as well. The



(a) Floodplain elevation and sampling cross section

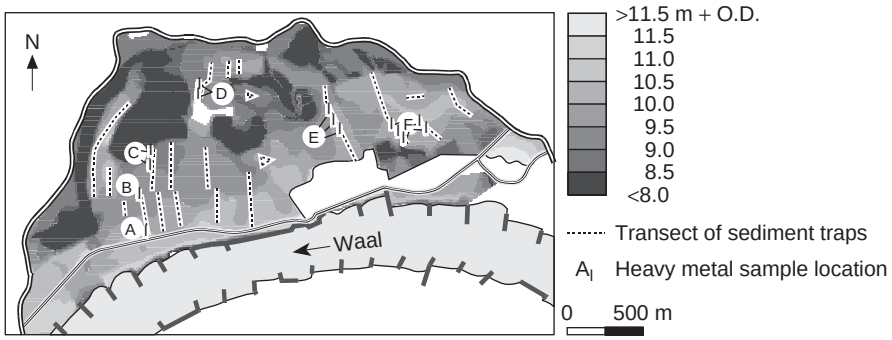


(b) Cross section; sediment composition and heavy metal content

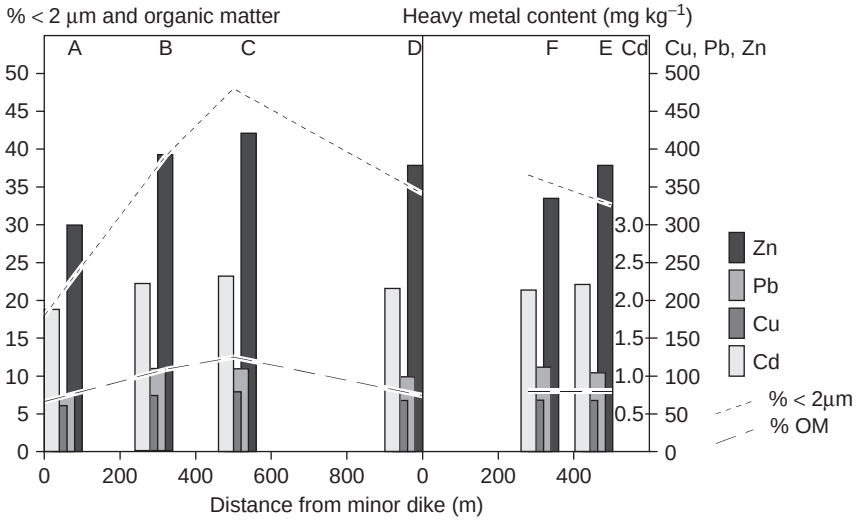


(c) Cross section; floodplain elevation and deposition of sediment and heavy metals

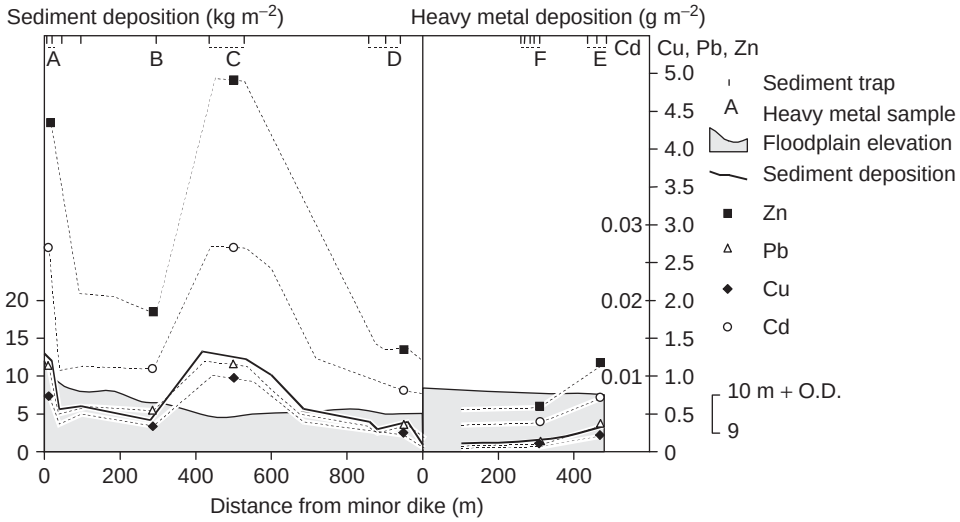
Figure 20 Heavy metal concentrations and deposition on the Klompenwaard



(a) Floodplain elevation and sampling cross sections



(b) Cross sections; sediment composition and heavy metal content



(c) Cross sections; floodplain elevation and deposition of sediment and heavy metals

Figure 21 Heavy metal concentrations and deposition on the Bemmelsche Waard

Table 3 Changes in sediment supply due to rainfall erosion in the Rhine basin (range of values downstream of the Alpine area), annual sediment load in the Rhine at the Dutch-German border, and overbank deposition on Stiftsche Uiterwaard (Waal) floodplain section (*Source*: Asselman *et al.*, 2003). CP = central projection: autonomous reduction in arable land without climate change; CPC idem with low, central, and upper estimates of climate change

	CP	CPC-low	CPC-central	CPC-up
dT (°C)	0	+1	+2	+4
dP summer (%)	0	−0.9	−1.9	−3.7
dP winter (%)	0	6.1	12.6	24.3
Area agriculture (% range)	−30 to −55	−30 to −60	−30 to −60	−30 to −60
Sediment supply (% range)	−13 to −33	−18 to −43	−15 to −42	−35 to +35
Sediment load in Lower Rhine (%)	−17	−19	−13	9
Deposition on Waal floodplain (%)	−47	−36	−8	73

emergent areas between flow paths over the floodplain shrink, while sediment is conveyed over the inundated area. Embankments of roads or other minor dykes may strongly affect the flow patterns over the floodplain by obstructing or diverting the water flow.

4. At the highest flood stages, floodwater flow extends across the entire floodplain conveying large amounts of sediment. Flow directions are influenced less by local topographic structures and more by the overall geometry of the valley floor. However, flow velocities are strongly reduced in areas behind minor embankments.
5. After the recession of the flood, the inundated water drains through depressions and ditches to the main channel. Still, water may remain ponded for many days behind embankments or in enclosed depressions within the floodplain.

Sediment Deposition

The hydraulic patterns that develop during overbank flooding determine the spatial distribution of sediment deposition by controlling the frequency and duration of inundation, the amount of suspended sediment in the floodwaters, and the local current velocities. Overbank sediments deposited during a flood can be subdivided into two major components: (i) sand – originating from the channel bed – laid down as splays on the top and landward slope of the natural levee, and (ii) clay and silt-sized sediments – wash-load – deposited over the remaining part of the floodplain.

Sand The lee-side of natural and artificial levees are the zones where most deposition of sand may occur, although fine sand is present across the entire width of the floodplain. These sand deposits may refer to the influence of the zone with turbulent eddies where the channel and floodplain flows mix (Marriott, 1992, 1996). Large sand deposition occurs where the water flow from the channel enters the floodplain, since at this point sediment concentrations are high, while flow velocities are drastically reduced within a few meters' distance (Ten Brinke *et al.*, 1998; Middelkoop and Asselman, 1998). Sand deposition during the highest flood stages may particularly occur on the inside of meander bends, not only due to helicoidal flow, but also as a result of a general flow direction in the down-valley direction that tends to cut off meander bends. Sand deposition on levees is highly variable along the channel, depending on local elevation and overbank flow velocities. Sand splays therefore occur as a few cm to several dm thick natural levee deposits intermittently occurring along the entire riverbank. Where a levee fails, sand deposition may develop into local crevasse splays (Gomez *et al.*, 1997).

Silt and Clay-sized Sediment The settling of the finer sediment from the floodwater continues as it is conveyed over the floodplain. The deposition rate depends on the flow depth, flow velocity, local sediment concentration, and effective grain size of the sediment. Progressive sedimentation from the suspension results in a longitudinal decline in sediment concentration along the flow path and, hence, in decreasing deposition. The rate of the decline depends on the effective settling velocities as well as on the water flow velocities, and therefore, is not necessarily linear or exponential. Behind natural levees a steep depositional gradient occurs away from the channel edge, mainly due to sand deposition (Middelkoop and Asselman, 1998). Also, the mean particle size of the flood deposits decreases with increasing distance to the main channel, with a rapid fall within the first few tens of meters from the channel bank (Marriott, 1992, 1996; He and Walling, 1996; Walling *et al.*, 1998b). Farther away from the levee, the decline becomes more gradual and is largely determined by deposition of finer sediments. Lowest deposition rates occur within the distal parts of floodplains where the floodwater has become depleted from sediment due to settling in the proximal parts. Where a return flow from the floodplain to the main stream occurs, local suspended sediment concentrations are low due to settling in the upstream area, while flow velocities increase, preventing the remaining fines from settling. Consequently, deposition rates are low in spite of the proximity of the main channel. Enhanced deposition locally occurs where the flow velocity drops, due to lateral divergence or increasing water depth in depressions. Deposition may also be promoted in the lee of upstanding dense vegetation, reducing floodwater velocities and trapping sediment.

However, where flow velocities are high, such as on low floodplain parts adjacent to the main channel or where flow concentrates, turbulent eddies in the floodwater may hinder settling, reducing local deposition rates. Back-ponding results in enhanced deposition. This occurs in early flood stages of frequent low-magnitude floods, and is due to retention-ponding after the recession of large-magnitude floods. Minor dikes reduce the interaction between the floodplain and the flow in the river channel, causing lower current velocities over the floodplain, while only the fine suspended fraction present in the upper part of the water column will enter the floodplain. Although a minor dike decreases the effective time of sediment transport onto the floodplain, sedimentation rates per day of inundation are relatively high because the lower flow velocities, longer residence times, and retention-ponding lead to a more efficient sedimentation.

Thus, within the channel belt, suspended sediment transport processes are dominated by longitudinal and convective currents that promote high concentrations of suspended sediments and high rates of overbank deposition. The deposition pattern over the floodplain confirms this convective transport. Outside the channel belt, convective currents perpendicular to the main channel are weak, and suspended sediment is therefore strongly influenced by diffusive mechanisms. Here, patterns of sediment concentrations and deposition rates are controlled by both local and more distant floodplain topography, which determine the point in the channel from which suspended sediment is supplied, and the transport conditions in the intervening regions.

Relation to Flood Magnitude

In low floodplain areas in particular, the amount of sediment deposition may increase less than proportionally with the flood duration or magnitude. This effect can be, in the first place, attributed to decreasing sediment concentrations in the river in the course of a flood due to sediment depletion in the upstream basin (cf. Walling, 1974, 1978; Wood, 1976; Asselman, 1997). In addition, sedimentation on low floodplains becomes inefficient when the discharge is very high. In those situations, the sediment trapping capability (STC) of the floodplain is low: water flow velocities are so high that most sediment is conveyed over the floodplain without settling. In contrast, sedimentation rates behind minor embankments are high even during high flood stages due to the major reduction of flow velocities behind the embankment, rendering deposition efficient. Consequently, the effective discharge for floodplain sedimentation (calculated as the discharge where the product of the associated STC, suspended sediment concentration in the floodwater, and average frequency of occurrence reaches the maximum value) is spatially distributed. In low-lying floodplain sections, the more frequent moderate floods are in the long term more effective for floodplain sedimentation than high-magnitude floods. High-magnitude events are more effective for floodplain sedimentation in floodplain sections behind minor embankments, in spite of their lower frequency of occurrence (Middelkoop, 1997; Asselman and Middelkoop, 1998; Middelkoop *et al.*, 2003).

Medium-term Deposition Rates

Average floodplain sedimentation rates reconstructed using ^{137}Cs or heavy-metal profiles show spatially varying deposition rates where the patterns are similar to those observed at the flood event, and thus also reflect the processes and controls described earlier (Bradley, 1995; Owens *et al.*, 1999b; Van Wijngaarden *et al.*, 2002; Middelkoop, 2000). As sediment deposition increases the floodplain elevation over time, the overbank flooding frequency gradually reduces. When the deposition per flood or per day of flooding remains constant, this results in a decreasing growth rate of the floodplain (Moody and Troutman, 2000). Estimates of average contemporary floodplain sedimentation rates therefore, may expected to be lower than deposition rates decennia ago. For example, reconstructed average sedimentation rates on the lower Rhine floodplains over the past century are higher than estimates of contemporary deposition rates (Middelkoop, 1997). In other cases, however, sedimentation rates have been fairly constant over the past century (Owens *et al.*, 1999a).

Regional Controls of Sediment Deposition Rates

While within-reach patterns of sedimentation rates reflect small-scale topographic controls on flow characteristics and sedimentation, between-reach variations in total sediment storage are largely a product of downstream changes in cross-valley floor morphology and flood frequency (Sweet *et al.*, 2003). Valley width, through its control on flood hydraulics, has a strong influence on the amount of sediment deposited on the floodplain. In narrow valleys, floodwaters are laterally constricted, increasing flow velocity and depth, thereby favoring sediment transport. Conversely, in wide valleys, floodwaters are shallow and slow, promoting the deposition of sediment (Lecce, 1997). Also, downstream variations in specific stream power are seen as having an important control (Magilligan, 1992). Graf (1983) suggested that deposition occurs where specific stream power decreases in the downstream direction. Rumsby (2000) found that because of the changes in channel shape and capacity, associated with channel incision over centuries, overbank sedimentation rates on islands have varied considerably over the past centuries. Nevertheless, estimated *volumes* of sedimentation may remain more or less constant and consistent in time.

Floodplain aggradation may nonetheless show continuing nondecreasing rates with no change in bed height, when flood magnitudes or sediment availability tend to increase.

This may occur because of (i) a change in partitioning between in a multichannel system, or the abandonment and silting of secondary channels, causing flood stages in the river to increase; (ii) a reduction in channel width; or (iii) external factors including climate change or land use changes leading to higher peak flows and larger sediment yield from the upstream catchment (Brown, 1996).

Another control of the amounts of floodplain deposition may be the varying suspension load in the river in the course of a flood event. Because of extreme hysteresis effects in the discharge–sediment load relationship in the upper Mississippi, and after comparing the amounts of floodplain deposition after recent high-magnitude floods, Benedetti (2003) concluded that the magnitude and timing of suspended sediment concentration peaks are also important controls on deposition, because they determine the availability of suspended sediment for transport to the floodplain during peak flood stages.

These results demonstrate that caution is needed when attempting to identify causal linkages between changes in vertical floodplain sedimentation rates and catchment land cover and/or climate factors. Therefore, it is important to take into account the local geomorphological setting of the sampling site (Rumsby, 2000).

Sedimentation Rates and Conveyance Losses

A wide range of floodplain sedimentation rates and estimated conveyance losses due to overbank sedimentation has been reported from various rivers around the world, and obtained using different methods (see Table 4). Deposition rates may vary from about 0.1 mm year^{-1} to several centimeters per year. Highest floodplain sedimentation rates are reported from humid tropical regions, monsoon areas, and regions with high sediment yields due to massive upland erosion and deforestation. Examples include estimates of floodplain sedimentation in the Yamuna river basin, India, where Saxena *et al.* (2002) determined deposition rates between 24.8 and 59.9 mm year^{-1}. Terry *et al.* (2002) found extremely large sedimentation rates in a tropical island (Fiji) river system characterized by biannual cyclone-induced floods. Average rates up to 3.2 cm year^{-1} over the past ca. 45 years were reconstructed, which are due to excessively high suspended sediment concentrations in the river (200–500 g l^{-1}) that are attained during the biggest floods.

Overbank deposition may result in significant conveyance losses from the total suspension load in a river. Asselman and Van Wijngaarden (2002) estimated that the average annual sediment accumulation on the embanked floodplain of the river Rhine's lower distributaries in the Netherlands equals 1.72 kg m^{-2} year^{-1-1}, which is equivalent to 0.394 Mton year^{-1}. This is about 13% of the total annual suspended sediment load transported into The Netherlands at Lobith. Middelkoop and Asselman (1998) estimated that during the high-magnitude flood (70-year recurrence time) of the Rhine in 1993, about 0.24 Mton of overbank fines was deposited on the Waal floodplain, which is about 7.7% of the annual sediment load in the Waal and 19% of the sediment transported through the Waal during this flood. Reconstructed conveyance losses due to sediment storage by floodplain sedimentation along UK rivers (Culm, Ouse) during the past decennia vary from 10–40% (Lambert and Walling, 1987), to about 40–50% (Walling *et al.*, 1998a), and up to 50–60% (Sweet *et al.*, 2003) of the sediment delivered to these rivers). When compared to within-channel storage of fine sediments ($< 150\,\mu$m), floodplain trapping efficiency is higher, and causes a net loss from the river system. Because of the large residence time within the floodplain, these conveyance losses due to floodplain sedimentation thus constitute a major component of the sediment budget of the rivers (Owens *et al.*, 1999b).

Floodplain Deposition of Heavy Metals

In addition to suspended sediment, overbank floodplain deposits may be used to investigate sediment-associated contaminant sources and transport dynamics in river basins. While the contamination includes nutrients, trace metals, organic pollutants such as pesticides, Polychlorinated Biphenyls (PCBs), and Polycyclic Aromatic Hydrocarbons (PAHs), attention has been focused on heavy metals because of their relatively long pollution history with often huge releases, and their tendency to remain well bonded to sediment within the fluvial system. The amount and spatial patterns of heavy metal deposition on floodplains depend on a sequence of controls, including the amounts of pollutants released in the upstream basin over the past centuries, metal concentrations in the suspension load and preferential bonding of pollutants to finer particles and organic matter, and the patterns of floodplain deposition and the associated variation in grain sizes.

Contamination of the fluvial system originates from industrial and urban areas and associated point sources, from the diffuse release from industrial wastes, or from the reworking of historically contaminated floodplain deposits along the river. Often, a pronounced decrease in concentrations occurs in a downstream direction due to input of sediments from uncontaminated tributaries (Leenaers and Rang, 1989). Nevertheless, increased concentrations deposited further downstream indicate that remobilization of contaminants stored in upstream floodplains has occurred due to bank erosion, and was subsequently deposited on the downstream floodplain (Dennis *et al.*, 2003). The temporal trends of heavy-metal pollution resulting from nineteenth and twentieth century mining and industrial activities are reflected by variations in metal content (i.e. increased levels of Cd, Cu, Pb, and Zn) in floodplain soil profiles (Swennen *et al.*, 1994; Marron, 1986; Stam, 1999; Middelkoop, 2000; Walling and Owens, 2003). In the upper part of the

Table 4 Methods and vertical aggradation rates of floodplains (rates expressed in units used in the original paper)

Reference	River	Basin area (km^2)	Method	Time period	Sedimentation rate
Lambert and Walling (1987)	Culm, UK	276	Astro-tuf sediment traps (artificial grass)	1 year	0.49 mm $year^{-1}$
Brown (1987)	Severn, UK	10 000	Soil profile	10 000 (pre-settlement)	1.4 mm $year^{-1}$
Simm (1995)	Culm, UK	276	Astro-tuf sediment traps (artificial grass)	mean of 14 flood events	0.1–8.1 mm $year^{-1}$
Nicholas and Walling (1997a)	Culm, UK	274	Astro-tuf sediment traps, modeling, ^{137}Cs	flood events; 40 years	1–6 mm a^{-1}
Collins *et al.* (1997)	Severn, Exe, UK	4325 and 600	^{137}Cs	40–50 years	6.0–9.1 mm $year^{-1}$
Walling *et al.* (1998)	Ouse, Wharfe, UK	3315 and 818	^{137}Cs	40 years	0.010–0.554 g cm^{-2} $year^{-1}$
Walling and He (1998)	UK rivers	276–3713	^{137}Cs	40 years	0.08–0.16 g cm^{-2} $year^{-1}$ 0.7–1.6 mm $year^{-1}$
Walling and He (1997a)	Culm, UK	274	^{210}Pb, ^{137}Cs	40–100 years	0.4–9.5 mm $year^{-1}$
Owens *et al.* (1999a)	Ouse, UK	3520	^{210}Pb, ^{137}Cs	30–100 years	0.11–1.04 g cm^{-2} $year^{-1}$
Owens *et al.* (1999b)	Tweed, Scotland	4390	^{137}Cs	30–40 years	0.16–2.18 kg m^{-2} $year^{-1}$ 0.13–2.2 mm $year^{-1}$
Rumsby (2000)	Tyne, UK	1918	Soil profiles, heavy metals	400 years	12.5–73.6 mm $year^{-1}$ 547–1123 m^3 km^{-1} $year^{-1}$
Sweet *et al.* (2003)	Culm, UK	274	Modeling, ^{137}Cs	40 yr	$<$0.5 g cm^{-2} $year^{-1}$, locally 0.5–2 g cm^{-2} $year^{-1}$
Asselman and Middelkoop (1995)	Rhine, The Netherlands	165 000	Artificial grass sediment traps	Jan. 1993 flood event	mean = 0.82 mm

(*continued overleaf*)

Table 4 (*continued*)

Reference	River	Basin area (km^2)	Method	Time period	Sedimentation rate
Asselman and Middelkoop (1995)	Meuse, The Netherlands	33 000	Artificial grass sediment traps	Jan. 1993 flood event	mean = 0.47 mm
Middelkoop and Asselman (1998)	Rhine, The Netherlands	165 000	Artificial grass sediment traps	Dec. 1993 flood event	1.2–4.0 kg m^{-2} (clay-silt fraction)
Middelkoop and Asselman (1998)	Meuse, The Netherlands	33 000	Artificial grass sediment traps	Dec. 1993 flood event	1.0–2.0 kg m^{-2} 2.0 (clay-silt fraction)
Stam (1999)	Geul, The Netherlands	350	^{210}Pb, ^{137}Cs, heavy metals	since ca. 1800 A.D.	10–20 mm $year^{-1}$
Lecce (1997)	Blue river, Wisconsin – USA	208	Lithology		
Gomez *et al.* (1995)	Mississippi	308 000	Post-flood survey	1993 flood event	< 2–200 mm
Kesel *et al.* (1974)	Mississippi, USA	??	Thickness of flood deposits	1973 flood	5–840 mm
Magilligan (1985)	Galena, USA	526	Soil profiles	60 years	7.5 mm $year^{-1}$
Saxena *et al.* 2002	Yamuna, India	40.2% of Ganges	^{210}Pb, ^{137}Cs	50 years	25–60 mm $year^{-1}$
Terry *et al.* (2002)	Wainimala, Fiji	ca. 1000	^{137}Cs	45 years	up to 32 mm $year^{-1}$
Ormerod (1998)	Ironbark Creek, Australia	??	^{137}Cs	50 yr	2.6–27.6 mm $year^{-1}$
Campbell *et al.* (1982)	Maluna Creek, Australia	??	^{137}Cs	15 yr	5.7 mm $year^{-1}$
Goodbred and Kuehl (1998)	Brahmaputra, Bangladesh	??	^{210}Pb, ^{137}Cs	50–100 years	1.6 mm $year^{-1}$
Mertes (1994)	Amazon, Brazil	??	Measurement of flood deposit	recent annual floods	270–420 mm $year^{-1}$

profiles, there is a general trend of decreasing concentrations, reflecting changes in the nature of industrial activity in the last decennia and compliance with legislation and directives aimed at reducing the discharge of contaminants to water courses. Conversely, analysis of contaminants that have accumulated over the past centuries in floodplains or in floodplain lakes is an important means of assessing the degree of pollution in drainage areas (Middelkoop, 2000; Swennen *et al.*, 1998; Hutchinson, 2003).

The absolute concentrations associated with floodplain and channel-bed sediments are usually lower than corresponding values in the suspended sediment, which primarily reflects differences in particle size composition of the different sediments (Owens *et al.*, 2001). Furthermore, because of dilution with cleaner eroded sediment, metal concentrations in suspended sediment are lower during periods of increased discharge when overbank deposition occurs than during periods of low flow (Middelkoop *et al.*, 2002).

Mobile rivers with floodplains show a complex pattern of sediment pollution (Wolfenden and Lewin, 1977). Metal values reflect the timing of deposition of lateral accretions and point bar sediments in relation to the period of mining activity. Metal concentrations in low-energy environments are considerably higher than in the coarse point bars, complicating the age–contamination relationship in floodplain areas. Within the floodplains, metal contamination of floodplain sediments is spatially highly variable on account of the variation in deposition amounts and grain sizes of overbank sediments (Leenaers and Rang, 1989; Bradley, 1995; Middelkoop, 2000). Spatial variations in metal *concentrations* in overbank sediment deposited during a flood are influenced by differences in the clay and organic matter contents in the sediment, due to the strong preferential bonding of heavy metals to these components. Consequently, the highest metal concentrations generally occur in the finer sediments deposited in the distal parts of the floodplain and in ponded areas after settling of the finest sediments, while metal concentrations in the sand sheets are low. The patterns of metal *deposition*, however, are generally determined by the amounts of deposited sediment, which in turn are controlled by the flow patterns during overbank flooding. In the long term, the largest amounts of metals will accumulate nearby the natural levee due to the large sediment deposition, and in low-lying areas that inundate often and where large amounts of fine sediments are deposited.

Estimates of contemporary metal deposition as well as estimated deposition over the past century demonstrate that in polluted catchments, conveyance losses of particulate-associated contaminants due to floodplain deposition can be significant. Examples of metal supply onto floodplains are the deposition of 5–10 kg Cd, 75–150 kg Cu, 250–300 kg Pb, and 1000–1500 kg Zn per km^2 on the lower Rhine-Meuse floodplain measured during a high-magnitude event (Middelkoop, 2000), and a contemporary annual supply per square kilometer on the Derwent River floodplain (UK) equal to 10 kg Cd, 50 kg Cu, 1800 kg Pb, and 500 Zn, estimated by Bradley (1995) from the amounts of metals in floodplain soils. Walling and Owens (2003) estimated for polluted rivers in Yorkshire, UK, that up to 26–47% of the total flux of contaminants through the main channel has been stored on the floodplain. This is, in some cases, more than the conveyance losses of suspended sediment.

Given the amounts of sediments trapped by floodplain sedimentation and concentrations of contaminants in the sediments, floodplains may constitute important sinks for contaminants (Owens *et al.*, 2001). While the residence time of polluted sediment within channels may be in the order of less than a year, most of the sediment and pollutants stored in floodplains may have a residence time between 10^1 and 10^3 years. Where lateral channel migration is important, the pollutants will be reintroduced back into the river channel. This may make the floodplain a net source of contaminants even if primary inputs have ceased (Dennis *et al.*, 2003). If the frequency of high-magnitude floods increases in the forthcoming decades due to climate change as predicted by climate models, it is probable that historically mined areas will experience significant flood-related contamination as a result of enhanced channel bank erosion and subsequent reworking of metal-rich floodplain material (Macklin, 1996). Therefore, determination of the magnitudes and rates of contaminant accumulation on the floodplains, and the residence time of the sediment-associated contaminant storage on floodplains is essential for understanding, monitoring, and predicting the fate and release of sediment-associated contaminants in river systems.

Impact of Climate Change and Land Use and River Management

Two primary controls of floodplain sedimentation rates, that is, discharge and sediment yield in the upstream basin, depend on changes in climate and land use in the upstream catchment. Climate change firstly may affect the flooding frequency of a floodplain, as many climate models predict an increase of peak flows for temperate climate regions (Arnell, 1992; Middelkoop *et al.*, 2001; Asselman *et al.*, 2003). Furthermore, intensified summer precipitation and higher runoff will cause increased sediment yield in the upstream basin due to an increase in surface and bank erosion. Also, river training works and reservoirs may strongly affect the amounts of sediment transferred to downstream floodplains.

Owens and Walling (2002) found changes in sediment sources within a Scottish catchment, associated with changes in climate and land use over the past century. The extremely long record of sedimentation history of the lower Yellow river (Xu, 2003) demonstrated on a century scale that large-scale medieval deforestation of loess areas, followed by a period of increased precipitation has

increased sedimentation rates in the lower sedimentation zone from 2–4 mm year^{-1} to about 20 mm year^{-1}. The complete embankment of the river has further increased sedimentation rates on the remaining active floodplain to over 80 mm year^{-1} in recent years. Conversely, average sedimentation rates along the Upper Mississippi, Iowa, and Wisconsin, USA, over the past 50 years are in the order of about 100 mm year^{-1}, which is less than rates that prevailed in the twentieth century when agricultural land disturbance was at maximum, but they are an order of magnitude larger than long-term average rates for the Holocene (Benedetti, 2003). Magilligan (1985) demonstrated for a catchment in the United States that soil conservation measures may result in an overall reduction in bankfull channel capacity and bankfull width. While sediment yield and deposition rates were reduced, sediment transport relative to sediment storage has become more important. Valley width appeared a major control in determining the areas of transport and deposition along the river course. Stam (2002) showed that average floodplain sedimentation rates along the Geul river (The Netherlands) over 40–60 year time slices varied between 2 and 12 mm year^{-1} and attributed the higher sedimentation rates to more abundant low-magnitude floods that occurred due to deforestation and mining activities in the upstream catchment. Wyźga (2001) demonstrated that river training works and artificial bank protection have led to a dramatic reduction of the potential of the floodplains in the upper Vistula river basin to store sediment. Because of bank protection, the river vertically incises into its riverbed and the frequency of overbank flooding has reduced, while during floods, overbank flow velocities on the narrow floodplains are mostly too high for sediment deposition.

All these studies indicate the sensitivity of floodplain deposition rates to changes in climate, land use, and river training. These controls have not only changed in the past, but will do so in the forthcoming century as well, due to global warming, socioeconomic changes in agricultural land use (both causing intensified erosion by excessive land use and reduced sediment yield due to land conservation measures), and as a consequence of future river management.

CONCLUSIONS AND PROSPECTS

Floodplain sedimentation processes can be considered as a four-dimensional evolution process: spatially varying vertical accretion over time. Understanding floodplain development therefore requires not only analysis of two- or three-dimensional processes of channel-floodplain interaction, and of the spatially variable hydraulic patterns over the floodplain, but also the geomorphological development over time, and the controls of these. Consequently, identifying and understanding the factors that determine the complex discharge–sediment relationships are key steps in understanding the evolution of a floodplain over time and interpreting floodplain sedimentation as paleoenvironmental indicators.

The available data to document this evolution, however, do not fully represent these four dimensions at once, but are sparse in time and space and difficult to acquire (Anderson *et al.*, 1996). Analysis of floodplain evolution therefore relies on integrating time series of discharge and sediment transport from rivers, analysis of contemporary two-dimensional hydraulic patterns using remote sensing and sampling of the spatially variable overbank deposition, and reconstructions of past sedimentation rates, and relating these to the controls of the processes involved. While sediment traps have proven adequate tools for documenting spatial patterns of overbank deposition during single flood events, using radionuclides and heavy metals as tracers provides spatially distributed estimates of long-term sedimentation rates, avoiding representativeness problems associated with event-based sampling. Use of two- or three-dimensional models of floodplain hydraulics and sediment deposition is essential for quantifying the role of controls on the sedimentation processes, and for relating deposition at the event scale to longer-term floodplain development. However, this raises the demand for data to establish process models, to parameterize, and validate them.

Recent research has shown tremendous progress in documenting and quantifying floodplain sedimentation processes and their controls, and its role in sediment budgets of catchments.

Still, to improve the ability to predict changes in floodplain sedimentation rates in response to changes in climate and land use, to regulatory measures in upstream catchments, and to landscaping measures for flood prevention and river restoration, many research themes deserve further attention. Some key issues that are worth mentioning may be the following:

- Improvement of hydraulic modeling relies on a better representation of floodplain geometry (e.g. using high-resolution laser altimetry data; *see* **Chapter 58, Characterizing Forest Canopy Structure and Ground Topography Using Lidar, Volume 2**) and hydraulic processes, in particular, the role of hydraulic roughness of floodplain vegetation and the representation of turbulence. Improvement of sedimentation models requires additional information that allows parameter estimation, such as *in situ* measurement of effective grain size and aggregate settling velocity distributions under varying hydraulic conditions. Model validation requires field-scale observations of 2D or 3D water flow patterns and measurements of 2D patterns of sediment deposition resulting from various flood events of different magnitudes.

- In addition to the geomorphological implications, overbank deposition of sediment and contaminants may have considerable ecological impact. Therefore, the role of floodplain sedimentation as a control in spatial patterns of riparian ecology should be further explored.
- Improved understanding of catchment budgets requires better insight in the sources of the overbank deposits using various types of chemical and radiometric fingerprinting techniques.
- Finally, to fully understand the role of floodplains as sinks or storages of sediment and associated pollutants over a time period of 10^2–10^3 years, improved and quantified estimates of residence times of sediment and pollutants within floodplain areas are required. This demands research in which erosion and deposition processes within the floodplain area are considered together, which may be carried out by geomorphological reconstructions, contemporary process studies, and modeling exercises.

REFERENCES

Alexander C.S. and Prior J.C. (1971) Holocene sedimentation rates in overbank deposits in the black bottom of the lower Ohio river, southern Illinois. *American Journal of Science*, **270**, 361–372.

Allen J.R.L. (1978) Studies in fluviatile sedimentation: an elementary geometrical model for the connectedness of avulsion related to channel sandbodies. *Sedimentary Geology*, **24**, 253–267.

Allen J.R.L. (1985) *Principles of Physical Sedimentology*, Allen & Unwin: London.

Anderson M.G., Bates P.D. and Walling D.E. (1996) The general context for floodplain process research. In *Floodplain Processes*, Andersen M.G., Walling D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 1–13.

Appleby P.G. and Oldfield F. (1978) The calculation of lead-210 dates assuming a constant rate of supply of unsupported ^{210}Pb to the sediment. *Catena*, **5**, 1–8.

Arnell N.W. (1992) Factors controlling the effects of climate change on river flow regimes in a humid temperate environment. *Journal of Hydrology*, **132**, 321–277.

Asselman N.E.M. (1997) *Suspended Sediment in the River Rhine*, PhD thesis Department of Physical Geography, Netherlands Geographical Studies 234, Faculty of Geographical Sciences, Utrecht University: Utrecht.

Asselman N.E.M. (1999) Grain-size trends used to assess the effective discharge for floodplain sedimentation, river Waal, The Netherlands. *Journal of Sedimentary Research*, **69**, 51–61.

Asselman N.E.M. and Middelkoop H. (1995) Floodplain sedimentation: quantities, patterns and processes. *Earth Surface Processes and Landforms*, **20**, 481–499.

Asselman N.E.M. and Middelkoop H. (1998) Temporal variability of contemporary floodplain sedimentation in the Rhine-Meuse delta, The Netherlands. *Earth Surface Processes and Landforms*, **23**, 595–609.

Asselman N.E.M., Middelkoop H. and Van Dijk P.M. (2003) The impact of climate change on soil erosion, transport and deposition of suspended sediment in the river Rhine. *Hydrological Processes*, **17**, 3225–3244.

Asselman N.E.M. and Van Wijngaarden M. (2002) Development and application of a 1D floodplain sedimentation model for the river Rhine in The Netherlands. *Journal of Hydrology*, **268**, 127–142.

Baker V.C., Pickup G. and Polach H.A. (1985) Radiocarbon dating of flood deposits, Katherine Gorge, Northern Territory, Australia. *Geology*, **13**, 344–347.

Bates P.D., Anderson M.G., Price D.A., Hardy R.J. and Smith C.N. (1996) Analysis and development of hydraulic models for floodplain flows. In *Floodplain Processes*, Andersen M.G., Walling D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 215–254.

Bates P.D. and De Roo A.P.J. (2000) A simple raster-based model for floodplain inundation simulation. *Journal of Hydrology*, **236**, 54–77.

Beffa C. and Connell R.J. (2001) Two-dimensional flood plain flow – Part 1: model description. *Journal of Hydraulic Engineering*, **6**, 397–405.

Benedetti M.M. (2003) Controls on overbank deposition in the upper Mississippi River. *Geomorphology*, **56**, 271–290.

Bogen J., Bölviken B. and Ottesen R.T. (1992) Environmental studies in western Europe using overbank sediment. In *Proceedings of the Oslo Symposium, August 1992: Erosion and Sediment Transport Monitoring Programmes in River Basins*, Bogen J., Walling D.E. and Day T.J. (Eds.), IAHS Publication 210, IAHS, pp. 317–325.

Bradley S.B. (1995) Long-term dispersal of metals in mineralized catchments by fluvial processes. In *Sediment and Water Quality in River Catchments*, Foster I.D.L., Gurnell A.M. and Webb B.W. (Eds.), John Wiley & Sons: Chichester, pp. 161–177.

Bradley S.B. and Cox J.J. (1990) The significance of the floodplain in the cycling of metals in the river Derwent catchment, UK. *Science of the Total Environment*, **97/98**, 441–454.

Bridge J.S. and Leeder M.R. (1979) A simulation model of alluvial stratigraphy. *Sedimentology*, **26**, 617–644.

Bridge J.S. and Mackey S.D. (1993) A revised alluvial stratigraphy model. In *Alluvial Sedimentation*, Marzo M. and Puidefábregas C. (Eds.), Spec. Publ. Int. Assoc. Sedimentol. 17, Spec. Publ. Int. Assoc. Sedimentol. pp. 319–336.

Bristow C.S., Skelly R.L. and Ethridge F.G. (1999) Crevasse splays from the rapidly aggrading, sandy-bed, braided Niobrara River, Nebraska: effect of base-level rise. *Sedimentology*, **46**, 1029–1047.

Brown A.G. (1983) An analysis of overbank deposits of a flood at Blandford Forum, Dorset, England. *Reviews in Geomorphology*, **32**, 95–99.

Brown A.G. (1987) Holocene floodplain sedimentation and channel response of the lower river Severn, United Kingdom. *Zeitschrift für Geomorphologie N.F.*, **31**, 293–310.

Brown A.G. (1996) Floodplain palaeoenvironments. In *Floodplain Processes*, Andersen M.G., Walling D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 95–138.

Brunet R.C., Pinay G., Gazelle F. and Roques L. (1994) Role of the floodplain and riparian zone in suspended matter and

nitrogen retention in the Adour river: south-west France. *Regulated Rivers: Research and Management*, **11**, 357–369.

Burrough P.A. (1986) *Principles of Geographical Information Systems for Land Resources Assessment*, *Monograph on Soil and Resources Survey No. 12*, Clarendon Press: Oxford.

Campbell B.L., Loughran R.J. and Elliot G.L. (1982) Caesium-137 as an indicator of geomorphic processes in a drainage basin system. *Australian Geographical Studies*, **20**, 49–64.

Campbell B.L., Loughran R.J., Elliott G.L. and Shelly D.J. (1988) Mapping drainage basin sediment sources using caesium-137. In *Proceedings of the Albuquerque Symposium, August 1986, Drainage Basin Sediment Delivery*, Hadley R.F. (Ed.), IAHS Publication 159, IAHS pp. 437–446.

Cancino L. and Neves R. (1999) Hydrodynamic and sediment suspension modelling in estuarine systems. Part I: description of the numerical models. *Journal of Marine Systems*, **22**, 105–116.

Collins A.L., Walling D.E. and Leeks G.J.L. (1997) Fingerprinting the origin of fluvial suspended sediment in larger river basins: combining assessment of spatial provenance and source type. *Geografiska Annaler*, **79A**(4), 239–254.

Connell R.J., Beffa C. and Painter D.J. (1998) Comparison of observations by floodplain residents with results from a two-dimensional flood plain model: Waiho river, New Zealand. *Journal of Hydrology*, **37**, 55–79.

Connell R.J., Beffa C. and Painter D.J. (2001) Two-dimensional flood plain flow – Part 2: model validation. *Journal of Hydraulic Engineering*, **6**, 406–415.

Costa J.E. (1975) The effects of agriculture on erosion and sedimentation in Piedmont Province. *Bulletin of the Geological Society of America*, **86**, 1281–1286.

De Cort M., Dubois G., Fridman S.D., Izrael Y.A., Janssens A., Jones A.R., Kelly G.N., Kvasnikova E.V., Matveenko I.I., Nazarov I.M., *et al.* (1998) *Atlas of Caesium Deposition on Europe after the Chernobyl Accident*, Report EUR 16733, European Commission, Luxembourg.

Dennis I.A., Macklin M.G., Coulthard T.J. and Brewer P.A. (2003) The impact of the October-November 2000 floods on contaminant metal dispersal in the river swale catchments, North Yorkshire, UK. *Hydrological Processes*, **17**, 1641–1657.

Dezzeo N., Herrera R., Escalante G. and Chacon N. (2000) Deposition of sediments during a flood event on seasonally flooded forests of the lower Orinoco River and two of its back-water tributaries. *Biogeochemistry*, **49**, 241–257.

Dominik J., Mangini A. and Müller G. (1981) Determination of recent deposition rates in lake constance with radioisotopic methods. *Sedimentology*, **28**, 653–677.

Dominik J., Mangini A. and Prosi F. (1984) Sedimentation rate variations and anthropogenic metal fluxes into lake constance sediments. *Environmental Geology*, **5**, 151–157.

Ferguson R.J. and Brierley G.J. (1999) Levee morphology and sedimentology along the lower Tuross river, south-eastern Australia. *Sedimentology*, **46**, 627–648.

Fischer-Antze T., Stoesser T., Bates P.D. and Olsen N.R.B. (2001) 3D numerical modeling of open-channel flow with submerged vegetation. *Journal of Hydraulic Research*, **39**, 303–310.

Foster I.D.L. and Charlesworth S.M. (1996) Heavy metals in the hydrological cycle: trends and explanation. *Hydrological Processes*, **10**, 227–261.

Galland J.C., Goutal N. and Hervouet J.-M. (1991) TELEMAC – a new numerical-model for solving shallow-water equations. *Advances in Water Resources*, **14**, 138–148.

Gomez B., Eden D.N., Peakock D.H. and Pinkney E.J. (1998) Floodplain construction by rapid vertical accretion: Waipaoa river, New Zealand. *Earth Surface Processes and Landforms*, **23**, 405–413.

Gomez B., Mertes L.A.K., Phillips J.D., Magilligan F.J. and James L.A. (1995) Sediment characteristics of an extreme flood: 1993 upper Mississippi river valley. *Geology*, **23**, 963–966.

Gomez B., Phillips J.D., Magilligan F.J. and James L.A. (1997) Floodplain sedimentation and sensitivity: summer 1993 flood, upper Mississippi river valley. *Earth Surface Processes and Landforms*, **22**, 923–936.

Goodbred S.L. Jr and Kuehl S.A. (1998) Floodplain processes in the Bengal basin and the storage of Ganges-Brahmaputra river sediment: an accretion study using ^{137}Cs and 210Pb geochronology. *Sedimentary Geology*, **121**, 239–258.

Graf W.L. (1983) Downstream changes in stream power in the Henry Mountains, Utah. *Annals of the Association of American Geographers*, **73**, 373–387.

Gretener B. and Strömquist L. (1987) Overbank sedimentation rates of fine grained sediments. A study of the recent deposition in the lower river Fyrisan. *Geografiska Annaler*, **69A**, 139–146.

Hardy R.J., Bates P.B. and Anderson M.G. (1998) The importance of spatial resolution in hydraulic models for floodplain environments. *Journal of Hydrology*, **216**, 124–136.

He Q. and Owens P. (1995) Determination of suspended sediment provenance using Caesium-137, unsupported Lead-210 and Radium-226: a numerical mixing model approach. In *Sediment and Water Quality in River Catchments*, Foster I.D.L., Gurnell A.M. and Webb B. (Eds.), John Wiley & Sons: Chichester, pp. 207–227.

He Q. and Walling D.E. (1996) Use of fallout Pb-210 measurements to investigate longer-term rates and patterns of overbank sediment deposition on the floodplains of lowland rivers. *Earth Surface Processes and Landforms*, **21**, 141–154.

Hervouet J.-M. and Van Haren L. (1996) Recent advances in numerical methods for fluid flows. In *Floodplain Processes*, Andersen M.G., Walling D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 183–214.

Hesselink A.W. (2002) *History makes a River. Morphological Changes and Human Interference in the River Rhine, The Netherlands*. PhD thesis Department of Physical Geography. Netherlands Geographical Studies 292. Utrecht University: Utrecht.

Hesselink A.W., Stelling G.S., Kwadijk J.C.J. and Middelkoop H. (2003) Inundation of a Dutch river polder, sensitivity analysis of a physically based inundation model using historic data. *Water Resources Research*, **39**, 1234.

Horritt M.S. and Bates P.D. (2002) Evaluation of 1D and 2D numerical models for predicting river flood inundation. *Journal of Hydrology*, **268**, 87–99.

Howard A.D. (1992) Modelling channel migration and floodplain development in meandering streams. In *Lowland Floodplain Rivers: Geomorphological Perspectives*, Carling P.A. and Petts G.E. (Eds.), John Wiley & Sons: Chichester, pp. 1–42.

Howard A.D. (1996) Modelling channel evolution and floodplain morphology. In *Floodplain Processes*, Andersen M.G., Walling

D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 15–62.

Hudson P.F. (2003) Floodplains: environment and process – editorial. *Geomorphology*, **56**, 223–224.

Hudson-Edwards K.A., Macklin M.G., Curtis C.D. and Vaughan D.J. (1998) Chemical remobilization of contaminant metals within floodplain sediments in an incising river system: implications for dating and chemostratigraphy. *Earth Surface Processes and Landforms*, **23**, 671–684.

Hutchinson S.M. (2003) Environmental archives of heavy metal pollution or contaminated land? A case study of former water powered industrial sites in South Yorkshire, UK. *Journal de Physique IV*, **107**, 645–648.

James C.S. (1985) Sediment transfer to overbank sections. *Journal of Hydraulic Research*, **23**, 435–452.

Kesel R.H., Dunne K.C., Mc Donald R.C., Allison K.R. and Spicer B.E. (1974) Lateral erosion of overbank deposition on the Mississippi river in Louisiana, caused by 1973 flooding. *Geology*, **2**, 461–464.

Kitchen N.R., Sudduth K.A. and Drummond S.T. (1996) Mapping of sand deposition from 1993 midwest floods with electromagnetic induction measurements. *Journal of Soil and Water Conservation*, **51**, 336–340.

Kleiss B.A. (1996) Sediment retention in a bottomland hardwood wetland in eastern Arkansas. *Wetlands*, **16**, 321–333.

Knox J.C. (1987) Historical valley floor sedimentation in the upper Mississippi valley. *Annals of the Association of American Geographers*, **77**, 224–244.

Knox J.C. (1989) Long- and short-term episodic storage and removal of sediment in watersheds of southwestern Wisconsin and northwestern Illinois. In *Proceedings of the Baltimore symposium, May 1989, Sediment and the Environment*, Hadley R.F. and Ongley E.D. (Eds.), IAHS Publication 184, IAHS, pp. 157–164.

Knox J.C. (1993) Large increases in flood magnitude in response to modest changes in climate. *Nature*, **361**(6411), 430–432.

Lambert C.P. and Walling D.E. (1986) Suspended sediment storage in river channels: a case study of the river Exe, Devon, UK. In *Drainage Basin Sediment Delivery*, Hadley R.F. (Ed.), IAHS Publication 159, IAHS, pp. 263–276.

Lambert C.P. and Walling D.E. (1987) Floodplain sedimentation: a preliminary investigation of contemporary deposition within the lower reaches of the river Culm, Devon, UK. *Geografiska Annaler*, **69A**, 393–404.

Lawler D.M. (1991) A new technique for the automatic monitoring of erosion and deposition rates. *Water Resources Research*, **27**, 2125–2128.

Lecce S.A. (1997) Spatial patterns of historical overbank sedimentation and floodplain evolution: Blue river, Wisconsin. *Geomorphology*, **18**, 265–277.

Leenaers H. (1989) *The Dispersal of Metal Mining Wastes in the Catchment of the River Geul (Belgium – The Netherlands)*, Thesis, Department of Physical Geography, Utrecht University: Utrecht.

Leenaers H. and Rang M.C. (1989) Metal dispersal in the fluvial system of the river Geul: the role of discharge, distance to the source, and floodplain geometry. In *Proceedings of the Baltimore Symposium, May 1989, Sediment and the Environment*, Hadley R.F. and Ongley E.D. (Eds.), IAHS Publication 184, IAHS, pp. 47–55.

Leenaers H., Schouten C.J. and Rang M.C. (1998) Variability of the metal content of flood deposits. *Environmental Geology and Water Sciences*, **11**, 95–106.

Leopold L.B. (1973) River channel change with time: an example. *Geological Society of America Bulletin*, **84**, 1845–1860.

Lesser G.R., Roelvink J.A., Van Kestera J.A.T.M. and Stelling G.S. (2004) Development and validation of a three-dimensional morphological model. *Coastal Engineering*, **51**, 883–915.

Lewin J. and Hughes D. (1980) Welsh floodplain studies II. Application of a qualitative inundation model. *Journal of Hydrology*, **46**, 35–49.

Lewin J. and Macklin M.G. (1987) Metal mining and floodplain sedimentation in Britain. In *International Geomorphology*, Part I, Gardiner V. (Ed.), John Wiley & Sons: Chichester, pp. 1009–1027.

Mackey S.D. and Bridge J.S. (1995) Three-dimensional model of alluvial stratigraphy: theory and application. *Journal of Sedimentary Research*, **B65**, 7–31.

Macklin M.G. (1996) Fluxes and storage of sediment-associated heavy metals in floodplain systems: assessment and river basin management issues at a time of rapid environmental change. In *Floodplain Processes*, Andersen M.G., Walling D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 441–460.

Magilligan F.J. (1985) Historical floodplain sedimentation in the Galena river basin, Wisconsin and Illinois. *Annals of the Association of American Geographers*, **75**, 583–594.

Magilligan F.J. (1992) Thresholds and the spatial variability of flood power during extreme floods. In *Geomorphic Systems*, Phillips J.D. and Renwick W.H. (Eds.), Elsevier: Amsterdam, pp. 373–390.

Magilligan F.J., Phillips J.D., James L.A. and Gomez B. (1998) Geomorphic and sedimentological controls on the effectiveness of an extreme flood. *Journal of Geology*, **106**, 87–95.

Mansfield G.R. (1939) Flood deposits of the Ohio river, January-February 1937, a study of sedimentation. In *Floods of the Ohio and Mississippi Rivers*, Grover N.C. (Ed.), Water Supply Paper, USGS 838, USGS, pp. 693–736, January-February 1937.

Mansikkaniemi H. (1985) Sedimentation and water quality in the flood basin of the river Kyronjoki in Finland. *Fennia*, **163**, 155–194.

Marriott S.B. (1992) Textural analysis and modeling of a flood deposit – river Severn, UK. *Earth Surface Processes and Landforms*, **17**, 687–697.

Marriott S.B. (1996) Analysis and modeling of overbank deposits. In *Floodplain Processes*, Andersen M.G., Walling D.E. and Bates P.B. (Eds.), John Wiley & Sons: Chichester, pp. 63–93.

Marron D.C. (1986) Floodplain storage of metal-contaminated sediments downstream of a gold mine at Lead, South Dakota. In *Chemical Quality of Water in the Hydrological Cycle. Chelsea Michigan, USA*, Averett R.C. and McKnight D.M. (Eds.), Lewis Publishers: pp. 193–210.

McKee E.D., Crosby E.J. and Berryhill H.L. (1967) Flood deposits, Bijou Creek, Colorado, June 1965. *Journal of Sedimentary Petrology*, **37**, 829–851.

Mertes L.A.K. (1994) Rates of flood-plain sedimentation on the central Amazon river. *Geology*, **22**, 171–174.

Middelkoop H. (1997) *Embanked Floodplains in the Netherlands. Geomorphological Evolution Over Various Time Scales*. PhD thesis Department of Physical Geography. Netherlands Geographical Studies 224. Utrecht University: Utrecht.

Middelkoop H. (2000) Heavy-metal pollution of the river Rhine and Meuse floodplains in The Netherlands. *Netherlands Journal of Geosciences*, **79**, 411–428.

Middelkoop H. (2002a) Reconstructing floodplain sedimentation rates from heavy metal profiles by inverse modelling. *Hydrological Processes*, **16**, 47–64.

Middelkoop H. (2002b) Application of remote sensing and gis-based modelling in the analysis of floodplain sedimentation. In *Application of Geographic Information Systems and Remote Sensing in River Studies*, Leuven R.S.E.W., Poudevigne I. and Teeuw R.M. (Eds.), Backhuys Publishers: Leiden, pp. 95–117.

Middelkoop H. and Asselman N.E.M. (1998) Spatial variability of floodplain sedimentation at the event scale in the Rhine-Meuse delta, The Netherlands. *Earth Surface Processes and Landforms*, **23**, 561–573.

Middelkoop H., Daamen K., Gellens D., Grabs W., Kwadijk J.C.J., Lang H., Parmet B.W.A.H., Schädler B., Schulla J. and Wilke K. (2001) Impact of climate change on hydrological regimes and water resources management in the Rhine basin. *Climatic Change*, **49**, 105–128.

Middelkoop H., Thonon I. and Van der Perk M. (2002) Effective discharge for heavy metal deposition on the lower river Rhine flood plain. *Proceedings of International Symposium held at Alice Springs, Australia, September 2002, The Structure, Function and Management Implications of Fluvial Sedimentary Systems*, IAHS Publication 276, IAHS, pp. 151–159.

Middelkoop H. and Van der Perk M. (1998) Modelling spatial patterns of overbank sedimentation on embanked floodplains. *Geografiska Annaler*, **80A**, 95–109.

Moody J.A. and Troutman B.M. (2000) Quantitative model of the growth of floodplains by vertical accretion. *Earth Surface Processes and Landforms*, **25**, 115–133.

Nakamura F. and Kikuchi S. (1996) Some methodological developments in the analysis of sediment transport processes using age distribution of floodplain deposits. *Geomorphology*, **16**, 139–145.

Narinesingh P., Klaassen G.J. and Ludikhuize D. (1999) Floodplain sedimentation along extended river reaches. *Journal of Hydraulic Research*, **37**, 827–845.

Nicholas A.P. and Mitchell C.A. (2003) Numerical simulation of overbank processes in topographically complex floodplain environments. *Hydrological Processes*, **17**, 727–724.

Nicholas A.P. and McLelland S.J. (1999) Hydrodynamics of a floodplain recirculation zone investigated by field monitoring and numerical simulation. In *Floodplains: Interdisciplinary Approaches*, Marriott S.B. and Alexander J. (Eds.), Special Publication 163, Geological Society: London, pp. 15–26.

Nicholas A.P. and McLelland S.J. (2004) Computational fluid dynamics modelling of three-dimensional processes on natural river floodplains. *Journal of Hydraulic Research*, **42**, 131–143.

Nicholas A.P. and Walling D.E. (1996) The significance of particle aggregation in the overbank deposition of suspended sediment on river floodplains. *Journal of Hydrology*, **186**, 275–293.

Nicholas A.P. and Walling D.E. (1997a) Modelling flood hydraulics and overbank deposition on river floodplains. *Earth Surface Processes and Landforms*, **22**, 59–77.

Nicholas A.P. and Walling D.E. (1997b) Investigating spatial patterns of medium-term overbank sedimentation on floodplains: a combined numerical modelling and radiocaesium-based approach. *Geomorphology*, **19**, 133–150.

Nicholas A.P. and Walling D.E. (1998) Numerical modelling of floodplain hydraulics and suspended sediment transport and deposition. *Hydrological Processes*, **12**, 1339–1355.

Ormerod L.M. (1998) Estimating sedimentation rates and sources in a partially urbanised catchment using caesium-137. *Hydrological Processes*, **12**, 1009–1020.

Osterkamp W.R. (1989) Sediment storage and movement on the Southern high plains of Texas as indicators by beryllium-ten. In *Proceedings of the Baltimore Symposium, May 1983, Sediment and the Environment*, Hadley R.F. and Ongley E.D. (Eds.), IAHS Publication 184, IAHS pp. 173–182.

Owens P.N. and Walling D.E. (2002) Changes in sediment sources and floodplain deposition rates in the catchment of the river Tweed, Scotland, over the last 100 years: the impact of climate and land use change. *Earth Surface Processes and Landforms*, **27**, 403–423.

Owens P.N. and Walling D.E. (2003) Temporal changes in the metal and phosphorus content of suspended sediment transported by Yorkshire rivers, U.K. over the last 100 years, as recorded by overbank floodplain deposits. *Hydrobiologia*, **494**, 185–191.

Owens P.N., Walling D.E. and Leeks G.J.L. (1999a) Use of floodplain sediment cores to investigate recent historical changes in overbank sedimentation rates and sediment sources in the catchment of the River Ouse, Yorkshire, UK. *Catena*, **36**, 21–47.

Owens P.N., Walling D.E. and Leeks G.J.L. (1999b) Deposition and storage of fine-grained sediment within the main channel system of the river Tweed, Scotland. *Earth Surface Processes and Landforms*, **24**, 1061–1076.

Owens P.N., Walling D.E., Carton J., Meharg A.A., Wright J. and Leeks G.J.L. (2001) Downstream changes in the transport and storage of sediment-associated contaminants (P, Cr and PCBs) in agricultural and industrialized drainage basins. *The Science of the Total Environment*, **266**, 177–186.

Peart M.R. and Walling D.E. (1982) Particle size characteristics of fluvial suspended sediment. Recent developments in the explanation and prediction of erosion and sediment yield. *Proceedings of the Exeter Symposium, July 1982*, IAHS Publication 137, IAHS, pp. 397–407.

Pinay G., Ruffinoni C. and Fabre A. (1995) Nitrogen cycling in two riparian forest soils under different geomorphic conditions. *Biogeochemistry*, **30**, 9–29.

Pizzuto J.E. (1987) Sediment diffusion during overbank flows. *Sedimentology*, **34**, 301–317.

Ranwell D.S. (1964) Spartina salt marshes in Southern England, II. Rate and seasonal pattern of sediment accretion. *Journal of Ecology*, **52**, 79–94.

Rigollet C. and De Meijer R.J. (2002) PHAROS: pluri-detector, high-resolution, analyser of radiometric properties of soil. *Nuclear Instruments and Methods*, **A 488**, 642–653.

Rodriguez J.F., Bombardelli F.A., Garcia M.H., Frothingham K.M., Rhoads B.L. and Abad J.D. (2004) High-resolution numerical simulation of flow through a highly sinuous river reach. *Water Resources Management*, **18**, 177–199.

Rumsby B. (2000) Vertical accretion rates in fluvial systems: a comparison of volumetric and depth-based estimates. *Earth Surface Processes and Landforms*, **25**, 617–631.

Salomons W. and Förstner U. (1984) *Metals in the Hydrocycle*, Springer-Verlag: Berlin.

Saxena D.P., Joos P., Van Grieken R. and Subramanian V. (2002) Sedimentation rate of the floodplain sediments of the Yamuna river basin (tributary of the river Ganges, India) by using Pb-210 and Cs-137 techniques. *Journal of Radioanalytical and Nuclear Chemistry*, **251**, 399–408.

Siggers G.B., Bates P.D., Anderson M.G., Walling D.E. and He Q. (1999) A preliminary investigation of the integration of modelled floodplain hydraulics with estimates of overbank floodplain sedimentation derived from Pb-210 and Cs-137 measurements. *Earth Surface Processes and Landforms*, **24**, 211–231.

Silva W., Klijn F. and Dijkman J. (2000) *Ruimte voor Rijntakken. Wat het Onderzoek Ons Geleerd Heeft*, (Room for the Rhine distributaries. What we have learned from the research), RIZA/WL|Delft Hydraulics: Arnhem/Delft.

Simm D.J. (1995) The rates and patterns of overbank deposition on a lowland floodplain. In *Sediment and Water Quality in River Catchments*, Foster I.D.L., Gurnell A.M. and Webb B.W. (Eds.), John Wiley & Sons: Chichester, pp. 247–264.

Simm D.J., Walling D.E., Bates P.D. and Anderson M.G. (1997) The potential application of finite element modelling of floodplain inundation to predict patterns of overbank deposition. *Hydrological Sciences Journal*, **42**, 859–875.

Stam M.H. (1999) The dating of fluvial deposits with heavy metals, ^{210}Pb and ^{137}Cs in the Geul catchment (The Netherlands). *Physics and Chemistry of the Earth*, **B 24**, 155–160.

Stam M.H. (2002) Effects of land-use and precipitation changes on floodplain sedimentation in the nineteenth and twentieth centuries (Geul River, The Netherlands). *Special Publication International Association of Sedimentology*, **32**, 251–267.

Steiger J. and Gurnell A.M. (2003) Spatial hydrogeomorphological influences on sediment and nutrient deposition in riparian zones: observations from the Garonne River, France. *Geomorphology*, **49**, 1–23.

Steiger J., Gurnell A.M. and Goodson J.M. (2003) Quantifying and characterizing contemporary riparian sedimentation. *River Research and Applications*, **19**, 335–352.

Steiger J., Gurnell A.M. and Petts G.E. (2001a) Sediment deposition along the channel margins of a reach of the middle River Severn, UK. *Regulated Rivers-Research and Management*, **17**, 443–460.

Steiger J., Gurnell A.M., Ergenzinger P. and Stelder D. (2001b) Sedimentation in the riparian zone of an incising river. *Earth Surface Processes and Landforms*, **26**, 91–108.

Stoesser T., Wilson C.A.M.E., Bates P.D. and Dittrich A. (2003) Application of a 3D numerical model to a river with vegetated floodplains. *Journal of Hydroinformatics*, **5**, 99–112.

Sweet R.J., Nicholas A.P., Walling D.E. and Fang X. (2003) Morphological controls on medium-term sedimentation rates on British lowland river floodplains. *Hydrobiologia*, **494**, 177–183.

Swennen R., Van der Sluys J., Hindel R. and Brusselmans A. (1998) Geochemistry of overbank and high-order stream sediments in Belgium and Luxemburg: a way to assess environmental pollution. *Journal of Geochemical Exploration*, **62**, 67–79.

Swennen R., Van Keer I. and De Vos W. (1994) Heavy metal contamination in overbank sediments of the Geul river (East Belgium): its relation to former Pb-Zn mining activities. *Environmental Geology*, **24**, 12–21.

Taylor M.P. (1996) The variability of heavy metals in floodplain sediments: a case study from mid Wales. *Catena*, **28**, 71–87.

Ten Brinke W.B.M., Schoor M.M., Sorber A.M. and Berendsen H.J.A. (1998) Overbank sand deposition in relation to transport volumes during large-magnitude floods in the Dutch sand-bed Rhine river system. *Earth Surface Processes and Landforms*, **23**, 809–824.

Terry J.P., Garimella S. and Kostaschuk R.A. (2002) Rates of floodplain accretion in a tropical island river system impacted by cyclones and large floods. *Geomorphology*, **42**, 171–182.

Thonon I. and Van der Perk M. (2003) Measuring suspended sediment characteristics using a LISST-ST in an embanked flood plain of the river Rhine. *Proceedings of the Oslo Workshop, June 2002, Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, IAHS Publication 283, IAHS, pp. 37–44.

Thonon I., Van der Perk M. and Middelkoop H. (2002) Sediment-associated heavy metal pollution on the flood plains of the Rhine-Meuse delta, The Netherlands. In *Proceedings of International Symposium held at Alice Springs, Australia, September 2002, The Structure, Function and Management Implications of Fluvial Sedimentary Systems*, Poster Report Booklet, Alice Springs, pp. 44–47.

Törnqvist T.E. and Bridge J.S. (2002) Spatial variability of overbank aggradation rate and its influence on avulsion frequency. *Sedimentology*, **49**, 891–905.

Van der Perk M. (1997) Effect of model structure on the accuracy and uncertainty of results from water quality models. *Hydrological Processes*, **11**, 227–239.

Van Wijngaarden M., Rigollet C. and De Meijer R.J. (2002) Assessment of historic flood plain sedimentation rates along the river Meuse in The Netherlands using ^{137}Cs dating with PHAROS. *Proceedings of an International Symposium held at Alice Springs, Australia, September 2002, The Structure, Function and Management Implications of Fluvial Sedimentary Systems*, IAHS Publication 276, IAHS, pp. 389–397.

Walker I. (1995) Sedimentation in the inundation forest flanking the central Amazonian blackwater stream Rio Tarumã Mirím (Manaus, Amazones State). *Amazonica*, **13**, 237–243.

Walling D.E. (1974) Suspended sediment and solute yields from a small catchment prior to urbanization. In *Fluvial Processes in Instrumented Watersheds*, Gregory K.J. and Walling D.E. (Eds.), Institute of British Geographers Special Publication 6, Institute of British Geographers, pp. 169–192.

Walling D.E. (1978) Suspended sediment and solute response characteristics of the river Exe, Devon, England. In *Research in Fluvial Systems*, Davidson-Arnott R.G.D. and Nickling W. (Eds.), Geoabstracts: Norwich, pp. 169–197.

Walling D.E. (1983) The sediment delivery problem. *Journal of Hydrology*, **65**, 209–237.

Walling D.E. and Bradley S.B. (1989) Some applications of caesium-137 measurements in the study of fluvial erosion, transport and deposition. *Proceedings of the Jerusalem Workshop, Erosion, Transport and Deposition Processes*, IAHS Publication 189, IAHS, pp. 179–203.

Walling D.E., Bradley S.B. and Lambert C.P. (1986a) Conveyance losses of suspended sediment within a flood plain system. In *Drainage Basin Sediment Delivery*, Hadley R.F. (Ed.), IAHS Publication 159, IAHS pp. 119–132.

Walling D.E., Bradley S.B. and Wilkinson C.J. (1986b) A caesium-137 approach to the investigation of sediment delivery from a small agricultural drainage basin in Devon, UK. In *Drainage Basin Sediment Delivery*, Hadley R.F. (Ed.), IAHS Publication 159, IAHS, pp. 423–435.

Walling D.E. and He Q. (1993) Use of caesium-137 as a tracer in the study of rates and patterns of floodplain sedimentation. *Proceedings of the Yokohama Symposium, July 1993, Tracers in Hydrology*, IAHS Publication 215, IAHS, pp. 319–328.

Walling D.E. and He Q. (1994a) Rates of overbank sedimentation on the flood plains of several British rivers during the past 100 years. *Proceedings of the Canberra Symposium, December 1994, Variability in Stream Erosion and Sediment Transport*, IAHS Publication 224, IAHS, pp. 203–210.

Walling D.E. and He Q. (1994b) Changing rates of overbank sedimentation on the floodplains of British rivers during the past 100 years. In *Fluvial Processes and Environmental Change*, Brown A.G. and Quine T.A. (Eds.), John Wiley & Sons: Chichester, pp. 207–229.

Walling D.E. and He Q. (1997a) Investigating spatial patterns of overbank sedimentation on river floodplains. *Water Air and Soil Pollution*, **99**, 9–20.

Walling D.E. and He Q. (1997b) Use of fallout Cs-137 in investigations of overbank sediment deposition on river floodplains. *Catena*, **29**, 263–282.

Walling D.E. and He Q. (1998) The spatial variability of overbank sedimentation on river floodplains. *Geomorphology*, **24**, 209–223.

Walling D.E. and Owens P.N. (2003) The role of overbank floodplain sedimentation in catchment contaminant budgets. *Hydrobiologia*, **494**, 83–91.

Walling D.E., Owens P.N., Carter J., Leeks G.J.L., Lewis S., Meharg A.A. and Wright J. (2003) Storage of sediment-associated nutrients and contaminants in river channel and floodplain systems. *Applied Geochemistry*, **18**, 195–220.

Walling D.E., Owens P.N. and Leeks G.J.L. (1998a) The role of channel and floodplain storage in the suspended sediment budget of the river Ouse, Yorkshire, UK. *Geomorphology*, **22**, 225–242.

Walling D.E., Owens P.N. and Leeks G.J.L. (1998b) The characteristics of overbank deposits associated with a major flood event in the catchment of the river Ouse, Yorkshire, UK. *Catena*, **32**, 309–331.

Walling D.E., Owens P.N. and Leeks G.J.L. (1999) The role of channel and floodplain storage in the suspended sediment budget of the river Ouse, Yorkshire, UK. *Geomorphology*, **22**, 225–242.

Walling D.E., Quine T.A. and He Q. (1992) Investigating contemporary rates of floodplain sedimentation. In *Lowland Floodplain Rivers: Geomorphological Perspectives*, Carling P.A. and Petts G.E. (Eds.), John Wiley & Sons: Chichester, pp. 165–184.

Walling D.E., Rowan J.S. and Bradley S.B. (1989) Sediment-associated transport and redistribution of Chernobyl fallout radio-nuclides. In *Proceedings of the Baltimore Symposium, May 1989, Sediment and the Environment*, Hadley R.F. and Ongley E.D. (Eds.), IAHS Publication 184, IAHS, pp. 37–45.

Walling D.E. and Webb B.W. (1987) Suspended load in gravel-bed rivers: UK experience. In *Sediment Transport in Gravel-Bed Rivers*, Thorne C.R., Bathurst J.C. and Hey R.D. (Eds.), John Wiley & Sons: Chichester, pp. 691–732.

Walling D.E. and Woodward J.C. (1992) Use of radiometric fingerprints to derive information on suspended sediment sources. In *Proceedings of the Oslo symposium, August 1992, Erosion and Sediment Transport Monitoring Programmes in River Basins*, Bogen J., Walling D.E. and Day T.J. (Eds.), IAHS Publication 210, IAHS pp. 153–164.

Wardrop D.H. and Brooks R.P. (1998) The occurrence and impact of sedimentation in central Pennsylvania wetlands. *Environmental Monitoring and Assessment*, **51**, 119–130.

Wolfenden P.J. and Lewin J. (1977) Distribution of metal pollutants in floodplain sediments. *Catena*, **4**, 309–317.

Wolman M.G. and Eiler J.P. (1958) Reconnaissance study of erosion and deposition produced by the flood of August 1955 in Connecticut. *American Geophysical Union Transactions*, **39**, 1–14.

Wood P.A. (1976) Controls of variation in suspended sediment concentration in the river Rother, West Sussex, England. *Sedimentology*, **24**, 437–445.

Wyźga B. (1999) Estimating mean flow velocity in channel and floodplain areas and its use for explaining the pattern of overbank deposition and floodplain retention. *Geomorphology*, **28**, 281–297.

Wyźga B. (2001) Impact of the channelization-induced incision of the Skawa and Wisloka rivers, southern Poland, on the conditions of overbank deposition. *Regulated Rivers–Research and Management*, **17**, 85–100.

Xu J. (2003) Sedimentation rates in the lower Yellow river over the past 2300 years as influenced by human activities and climate change. *Hydrological Processes*, **17**, 3359–3371.

85: Sediment Yields and Sediment Budgets

DESMOND E WALLING

Department of Geography, University of Exeter, Exeter, UK

Measurements of sediment transport have been undertaken on many rivers throughout the world. Most of the available data relate to suspended sediment loads and the results provide a wealth of information on the variation of suspended sediment yields in both space and time. This contribution reviews current knowledge regarding land–ocean sediment transfer and sediment fluxes to the oceans, global patterns of sediment yield and their controls, and temporal variability of sediment yields in response to both natural controls and human activity and environmental change. In order to understand the sediment response of a drainage basin, it is important to take account of the complex linkages between sediment mobilisation and sediment output, and particularly the role of both short- and long-term storage. The sediment budget provides a useful conceptual framework for this purpose and current understanding of the structure of catchment sediment budgets is reviewed.

SEDIMENT YIELDS

The Context

Measurements of sediment transport or load (*see* **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**) have been undertaken on many rivers throughout the world, either within the framework of national monitoring programmes or linked to specific research investigations or water resource assessment studies. The results provide a wealth of information on the sediment loads carried by rivers in different areas of the globe and their variation in both space and time. In reviewing this information, it is useful to distinguish between different components of the sediment or particulate load, and more particularly the suspended sediment load and the bed load (*see* **Chapter 140, Transport of Sediments, Volume 4**). Much less information is available for bed load transport, since this is far more difficult to measure, particularly in large rivers. However, in most environments, the suspended load dominates the sediment load and will provide a reasonable estimate of the total particulate flux. When comparing the sediment loads of different rivers, attention commonly focuses on the annual load, or more particularly the mean annual load, since this provides a useful index of the amount of sediment mobilized from the catchment surface or the mass of sediment entering a reservoir or lake or discharged to the ocean. The mean annual load of a river is frequently referred to as its sediment yield ($t\,year^{-1}$). Since the magnitude of the sediment yield from a catchment or river basin will be strongly influenced by its size, it is convenient to express the sediment yield in terms of the sediment yield per unit area (i.e. $t\,km^{-2}\,year^{-1}$) when making comparisons between different catchments. The sediment yield per unit area is termed the specific sediment yield, to distinguish it from the gross sediment yield, but in practice the term sediment yield is frequently used to denote the specific sediment yield. As indicated above, the suspended sediment load dominates the total particulate load of most rivers and much of the existing work on sediment yields has focused on suspended sediment yields.

Some Basic Facts

The current availability of suspended sediment yield data for the world's rivers affords a reasonable basis for establishing the global range of specific suspended sediment yields. Taking basins $>100\,km^2$ in size, in order to avoid local anomalies, documented global minima for mean annual specific suspended sediment yield lie around $1\,t\,km^{-2}\,year^{-1}$. Branski (1975), for example, reports values of $<1.0\,t\,km^{-2}\,year^{-1}$ for several rivers in northern Poland and Branski and Banasik (1996) indicate that specific sediment yields are less than $5\,t\,km^{-2}\,year^{-1}$ over much of

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

lowland Poland. Dedkov and Mozzherin (1984) cite similar low values ($<1.0\,t\,km^{-2}\,year^{-1}$) for rivers in several areas of the former Soviet Union, including the Kola Peninsula, the Yenesei and Dneiper basins, and the catchment of the Sea of Azov. In most cases, these basins are characterized by very low relief, the presence of lakes and marshes, dense vegetation or ground cover and surface materials that are either resistant to erosion (e.g. crystalline rocks) or which afford only a limited source of fine sediment for transport as suspended sediment (e.g. coarse glacial deposits). All these characteristics limit sediment mobilization and transport.

Maximum documented mean annual specific suspended sediment yields exceed $10\,000\,t\,km^{-2}\,year^{-1}$ and Table 1 lists a number of rivers for which such extreme values have been reported. The highest value listed in Table 1 is a mean annual specific suspended sediment yield of $53\,500\,t\,km^{-2}\,year^{-1}$ for the Huangfuchuan River ($3199\,km^2$) in the People's Republic of China. This river is a tributary of the Middle reaches of the Yellow River (Huanghe), which drain the highly eroded gullied loess region. Here, the semiarid climate with high intensity rainfall in summer, the lack of vegetation cover, the highly erodible loess, and the densely dissected and gullied terrain combine to produce some of the highest rates of soil loss in the world. The dense gully networks and the high suspended sediment concentrations (hyperconcentrations), which may exceed $1000\,kg\,m^{-3}$ and which limit sediment deposition, also combine to produce a highly efficient sediment delivery system, such that most of the sediment mobilized by erosion is transported to the catchment outlet. A range of factors reflecting the topography, the erosivity of the local hydrometeorological regime, the erodibility of the terrain, tectonic instability, and surface disturbance by human activity can be invoked to account for the other examples of high suspended sediment yields noted in Table 1. The semiarid climate is again a key causal factor in the Kenyan example, but the impact of severe overgrazing is also important. For Taiwan, Java, and New Guinea, the steep mountainous relief, tectonic activity, high rainfall totals, and intense agricultural activity are key controls. In North Island, New Zealand, soft sedimentary rocks and tectonic activity account for the high specific sediment yields of rivers such as the Waiapu, and in South Island, New Zealand, the steep relief, tectonic uplift, and the very high rainfall (up to $9000\,mm\,year^{-1}$) are again key controls.

Table 1 Maximum reported values of mean annual specific suspended sediment yield for world rivers

Country Area	River sediment yield	Drainage (km^2)	Mean annual ($t\,km^{-2}\,year^{-1}$)
China	Huangfuchuan	3199	53 500
	Dali	187	21 700
China (Taiwan)	Erjenhsi	350	28 911
	Tsengwen	1000	28 000
Kenya	Perkerra	1310	19 520
Java	Cilutung	600	12 000
	Cikeruh	250	11 200
New Guinea	Aure	4360	11 126
North Island New Zealand	Waiapu	1378	19 970
	Waingaromia	175	17 340
	Hikuwai	307	13 890
South Island New Zealand	Hokitika	352	17 070
	Cleddau	155	13 300
	Haast	1020	12 736

Based on data presented by Walling and Webb (1996).

Although suspended sediment will, by definition, comprise only particles of fine sediment, information on the grain size composition of the suspended sediment transported by a number of rivers in different areas of the world reported by Walling and Moorehead (1989) and presented in Figure 1 serves to emphasize the potential range of size distributions that may exist. Thus, for example, the situation in the Barwon River in Australia, where the majority of the suspended sediment is clay-sized and $<0.002\,mm$ in diameter may be contrasted with the situation in the Huangfu River, a tributary of the Middle Yellow River in China, where up to 60% of the suspended sediment load is sand-sized (i.e. $>0.062\,mm$ in diameter) and clay-sized particles comprise only a small proportion of the total load. Such contrasts in the grain size composition of the suspended sediment load will have important implications for sediment chemistry and for the sediment-associated transport of nutrients and contaminants (cf. Horowitz, 1991), since the specific surface area of the sediment ($m^2\,g^{-1}$), which exerts a fundamental control on its capacity to adsorb nutrients and contaminants, increases markedly with decreasing particle size. Typical values for clay (e.g. $20–800\,m^2\,g^{-1}$) are, for example, several orders of magnitude greater than those for silt and sand. The contrasts in the grain size composition of suspended sediment evident in Figure 1 reflect both the granulometry of the source materials within a catchment and their degree of weathering, and the physiographic conditions. In the latter case, both low relief and low drainage density could be expected to promote the preferential deposition of coarser particles during transfer of mobilized sediment to the channel network, thereby increasing the proportion of finer particles. The relatively coarse nature of the loess deposits underlying the catchment of the Huangfu River, is clearly reflected by the grain size composition of the sediment, whereas the very fine suspended sediment found in the Barwon river reflects, at least in part, the very low relief of its catchment area.

The limited availability of reliable information regarding annual bed load transport by the world's rivers precludes a detailed assessment of the relative magnitude of the two load components, but some indication of the likely situation is provided by the data assembled by Dedkov

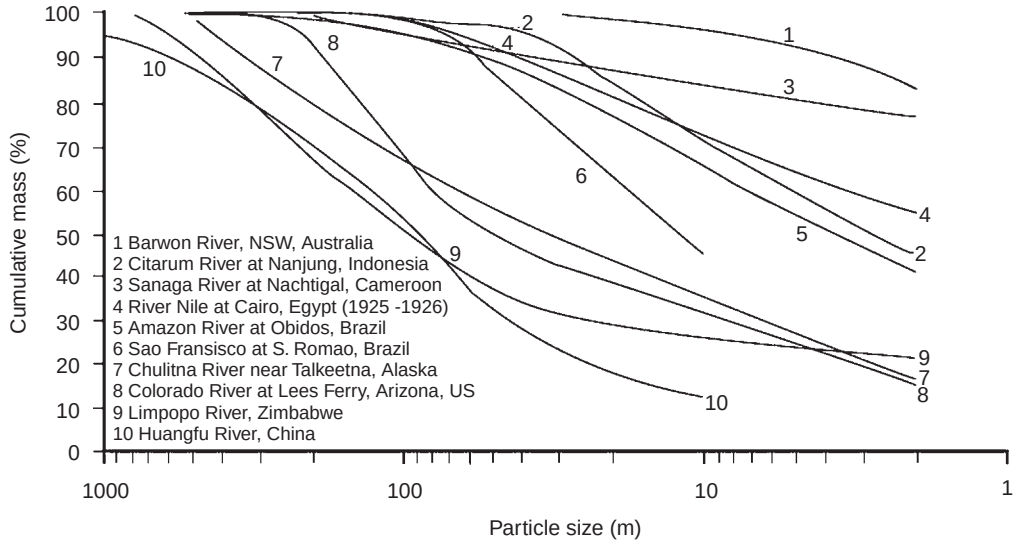


Figure 1 The grain size composition of suspended sediment transported by a range of world rivers (based on data presented by Walling and Moorehead, 1989)

and Mozzherin (1984) for a number of the worlds larger ($>5000\,km^2$) rivers, although the majority of these were located in the former Soviet Union. The estimates of bed load discharge were derived using a variety of methods, including both direct measurements and assessments of reservoir siltation and the use of empirical formulae, either for estimating bed load transport or for estimating the relative magnitude of the bed load component from measurements of the grain size composition of the suspended load and the bed material. These data indicated that bed load was on average ca. 8% of the suspended sediment load for "Plains" rivers and ca. 23% for "Mountain" rivers. Bed load was shown to reach its maximum importance, relative to the suspended sediment load, in glacierized mountains, for which an average value of 36% was reported.

Sediment Flux to the Oceans

Linked to the magnitude considerations outlined above, it is also useful to consider the total sediment flux from the land surface of the globe to the oceans. In the absence of information on bed load fluxes for most rivers, and accepting that the suspended load will dominate the global flux, attention will focus on suspended sediment loads. Since the 1950s, a number of workers have attempted to establish the magnitude of this flux (see Table 2), but it must be recognized that any such attempt faces several important problems. Leaving aside the question of the reliability of sediment load data, which represents a major source of uncertainty (see Walling and Webb, 1981), the problems include the lack of information on suspended sediment loads for many areas of the world, particularly large areas of Africa and South America, the location of measuring stations above the tidal limit, such that the measured loads may overestimate the flux to the ocean, where significant deposition occurs in the lower reaches of a river basin, and the temporal variation of annual sediment loads. In the latter context, the high interannual variability of annual sediment loads means that a short period of record may not provide a reliable estimate of the mean annual flux. The coefficient of variation of the annual suspended sediment load from a river basin will commonly exceed that of the annual water discharge and values in excess of 100%

Table 2 Existing estimates of total suspended sediment transport to the oceans

Author	Estimated mean annual load (10^9 t)
Fournier (1960)	51.1
Kuenen (1950)	32.5
Gilluly (1955)	31.7
Jansen and Painter (1974)	26.7
Pechinov (1959)	24.2
Syvitski (2003)	24.0
Schumm (1963)	20.5
Milliman and Syvitski (1992)	20.0
Holeman (1968)	18.3
Goldberg (1976)	18.6
USSR National Committee for the IHD (1974)	15.7
Sundborg (1973), Walling and Webb (1983)	15.0
Ludwig and Probst (1996)	14.8
Ludwig and Probst (1996)	14.7
Milliman and Meade (1983)	13.5
Lopatin (1952)	12.7
Mackenzie and Garrels (1966)	8.3

are not uncommon (Walling and Kleo, 1979). A lengthy period of record (e.g. 20 years) is therefore needed to characterize reliably the mean annual suspended sediment load of a river. Furthermore, it must be recognized that the records of suspended sediment flux available for many rivers cannot be regarded as stationary, since sediment yields are likely to be sensitive to climate change, changing land use and the construction of reservoirs. The latter will trap sediment (*see* **Chapter 88, Reservoir Sedimentation, Volume 2**) and therefore reduce land–ocean transfers. The global land–ocean flux cannot therefore be viewed as a static quantity, but rather one that is continually changing. As a result, use of data from different periods can introduce further uncertainty into the final result. The potential significance of the latter problem is well demonstrated by the sediment load data available for the Yellow River in China. Before 1985, the mean annual suspended sediment load for Lijin on the lower Yellow River averaged about 1×10^9 t, but this load declined markedly after 1985, as a result of climate change (lower annual rainfall), increased water abstraction for both irrigation and industrial use and the implementation of soil conservation and sediment control programmes in the upstream catchment, so that since 1985 it has averaged less than 0.5×10^9 t and in 1997 it fell to a record low of 0.03×10^9 t (Yang, 1998; Wang, 2003).

Table 2 presents some existing estimates of the total suspended sediment flux to the oceans. The values evidence considerable variability and range between 8.3 and 51.1 billion tons. Much of the variability reflects the limited data available to early workers and the need to extrapolate these data to large areas of the globe with no measurements. More recent studies have been able to take advantage of the increasing number of rivers for which sediment load data are available and values in the range 15 to 20×10^9 t year^{-1} have been most frequently cited. However, although data are now available for an increasing number of the major rivers of the world, which in turn account for a substantial proportion of the total land surface draining to the oceans, uncertainties still exist regarding the sediment flux from the smaller river basins draining to the oceans. Milliman and Syvitski (1992) have, for example, estimated that, whereas the world's 100 largest rivers drain ca. 60% of the land surface draining to the oceans, data from more than 10 000 rivers would be required to produce a definitive value for the total land–ocean flux. Any attempt to extrapolate existing data from large rivers to these smaller river basins faces substantial problems, since the specific sediment yields of smaller river basins are typically much greater than those of large basins. Furthermore, many smaller rivers drain areas of high relief, which are commonly characterized by high specific sediment yields. In particular, Milliman and Syvitski (1992) highlighted the need for more information on the sediment yields of the islands of Pacific Asia, where steep topography, tectonic instability, and land use pressure combine to produce very high sediment yields.

The most rigorous attempt to estimate the global suspended sediment flux to the oceans available to date is probably the work of Milliman and Meade (1983), which was further refined by Milliman and Syvitski (1992). Based on summation of the available river load data and extrapolation of this information to ungauged areas, Milliman and Meade (1983) produced an estimate of the mean annual suspended sediment flux to the oceans of 13.5×10^9 t year^{-1}. This value excluded sediment deposited in major reservoirs which would previously have reached the oceans and Milliman and Meade (1983) suggested that it should be increased to ca. 14 billion t year^{-1}, to represent the "natural" flux. A reassessment of the values used by Milliman and Meade (1983) when extrapolating to ungauged areas led Milliman and Syvitski (1992) to suggest that yields from many areas had been underestimated and that the total flux was likely to be ca. 20 billion t year^{-1}. Further refinement of the extrapolation to ungauged areas undertaken by Syvitski and Vorosmarty using digital elevation data for the global landmass and reported by Syvitski (2003) has increased the estimate to 24×10^9 t year^{-1}. Whereas the estimates produced by Milliman and Meade (1983), Milliman and Syvitski (1992) and Syvitski (2003) were based on summation of available load data and extrapolation of these data to ungauged areas, a different approach was adopted by Ludwig and Probst (1996), who used a database comprising sediment yield data for 58 rivers and information on climate and land surface characteristics for the land surface of the globe, extracted from available data sets at a resolution of $0.5° \times 0.5°$ latitude/longitude, to develop simple predictive models relating specific sediment yield to the product of several variables, including mean annual runoff, average basin slope, an index of the magnitude and seasonality of the annual precipitation and basin lithology. Both a single model for the entire globe and separate models for four different climate zones were developed and these were used to estimate the mean annual specific sediment yield for the total land surface of the globe and for the individual climate zones. The products of these specific sediment yields and the respective exoreic continental areas provided estimates of the total suspended sediment flux to the oceans of 14.8×10^9 t year^{-1} using the global model and 14.7×10^9 t year^{-1} using the models for the individual climate zones. These values are less than those proposed by Milliman and Syvitski (1992) and Syvitski (2003), but might arguably be expected to underestimate the total flux, since a number of river basins with high specific sediment yields were excluded from the global database.

In both the approaches outlined above, and indeed for most of the estimates listed in Table 1, at least some of the specific sediment yield data employed in deriving the

estimate of land–ocean flux will have been influenced by trapping of sediment in reservoirs and will thus underestimate what might be seen as the "natural" flux. In this respect, the estimates presented in Table 1 could be seen as underestimates. However, it is important to recognize that any attempt to establish the "natural" flux should also take account of the effects of land clearance, agricultural activity, and related activities in increasing sediment fluxes. It is possible that these opposing effects could be of a similar magnitude and thus cancel each out, but it would seem likely that the increased sediment flux associated with accelerated erosion caused by human activity would exceed any reduction caused by reservoir trapping. A wide range of estimates of the magnitude of the increase in the contemporary land–ocean sediment flux relative to the "pristine" situation prior to significant human impact have appeared in the literature. For example, Panin (2004) used evidence from past geological periods to estimate that contemporary sediment transport to the oceans is about 1.2 times the natural or pristine level. However, other authors have based their estimates on comparison of sediment yields from areas impacted by human activity with those from areas less influenced by anthropogenic impacts to suggest greater increases. Milliman and Syvitski (1992) and McLennan (1993) suggested a twofold increase, whereas Farnsworth and Milliman (2003) estimated that human activity may be directly or indirectly responsible for 80–90% of the delivery of fluvial sediment to the coastal ocean and thus that the flux has increased between five and tenfold.

Turning to the opposite effect, namely, the trapping of sediment in reservoirs, Vorosmarty *et al.* (2003) have undertaken a detailed analysis of the global distribution of reservoirs and their role in intercepting sediment transport. They have estimated that more than 40% of the water discharge of the world's rivers is currently intercepted by large ($\geq 0.5\,km^3$ maximum storage capacity) impoundments and that as much as 16% of the contemporary sediment flux from the land to the oceans is currently being trapped behind such dams. If the impact of the >44 000 smaller registered reservoirs is also considered, the proportion trapped increases to 28%, representing $4–5 \times 10^9\,t\,year^{-1}$. This trapping is clearly focused within those river basins of the world where the flow is regulated by dams and in these regulated basins the proportion of the total sediment flux that is trapped is as high as 53%. However, these values take no account of the ca. 8 00 000 smaller unregistered impoundments which can be expected to further increase the proportion of the global land–ocean sediment flux that is now trapped by dams. It is also important to recognize that the growth of dam construction on the world's rivers has been relatively recent. Vorosmarty *et al.* (2003) suggest that for the first half of the twentieth century, sediment retention in regulated basins was relatively modest, accounting for less than 5% of the global flux. However, after 1950, with the great expansion of dam building worldwide, the value tripled to 15% by 1968 and doubled again to nearly 30% by 1985 when it essentially stabilized. The values reported above suggest that, whilst significant, the trapping of sediment by reservoirs is likely to be significantly less than any increase in the land–ocean sediment flux caused by land use activities and that the latter can be expected to have caused a substantial increase in the contemporary land-ocean sediment flux relative to the "pristine" condition.

The Global Pattern of Sediment Yield and Its Controls

The literature provides examples of a number of attempts to generate maps of the global pattern of suspended sediment yield. Such attempts have inevitably faced a number of problems relating to the lack of data for many areas of the world, the extrapolation or interpolation procedures used, the nonstationary nature of sediment load data in many areas of the world, and the scale-dependent nature of sediment yield data. In the latter context, it is important to recognize that whilst information on the sediment yield at the outlets of major river basins may be a key requirement for assessing sediment flux to the oceans, the associated values of specific suspended sediment yield provide no information on the spatial variability of sediment yield within the upstream basin and indeed serve to mask such variability. Furthermore, the magnitude of specific suspended sediment yields has also been shown by many workers to be influenced by drainage area (e.g. Walling, 1983; Milliman and Syvitski, 1992) so that a map depicting the specific sediment yields of drainage basins of one size (e.g. $10^3\,km^2$) would present a different range of values from a map based on the sediment yields of much larger basins (e.g. $10^5\,km^2$).

Although the early attempts to produce maps of the global pattern of sediment yield by Fournier (1960) and by the Soviet scientist Lopatin (see Strakhov, 1967) have been frequently cited in attempts to describe the global denudation system, these maps were necessarily based on very limited data (96 rivers in the case of Fournier and 60 rivers for Lopatin) and must therefore be viewed as unreliable (see Walling and Webb, 1983). More recent improvements in data availability have permitted the production of more meaningful maps, although these still possess many uncertainties due to the lack of sediment yield data for many areas of the world. Those concerned primarily with fluxes to the oceans (i.e. Milliman and Meade, 1983) are, as noted above, of limited value for establishing the detail of the global pattern involved and in some cases only those areas of the globe for which data were available were mapped (e.g. Jansson, 1988). However, more explicit

attempts to map these patterns have been documented by Walling and Webb (1983, 1987), Walling (1987), Dedkov and Mozzherin (1984), and Lvovich *et al.* (1991). The maps produced by Walling and Webb (1983) and by Lvovich *et al.* (1991) are presented as Figures 2 and 3. Both relate to intermediate sized drainage basins and, although the procedures used to derive the maps were somewhat different (the former map is based primarily on spatial interpolation of available river load data, whereas the latter is based on extrapolation of available data using regional relationships between specific sediment yield and mean annual runoff), the patterns depicted evidence clear similarities. Maximum sediment yields are associated with the loess region of China and the young Cenozoic mountain areas around the Pacific margin. High values are also found in other mountain areas and in regions with Mediterranean and semiarid climates and in the seasonally humid tropics. Low values are associated with desert regions and with the low relief, glaciated regions of the Canadian Shield and northern Eurasia. In both cases, these maps must be seen as only tentative representations of the detailed global patterns involved, and further work is required to extend the sediment load databases employed, to assess and take account of the reliability of those data, to standardize the resulting estimates of sediment yield to a common drainage basin area, and to refine the interpolation and extrapolation procedures employed, particularly for those areas where data are totally absent. Recent advances in the development of detailed global databases for land surface and climate and hydrological characteristics (e.g. Vorosmarty *et al.*, 2000), coupled with GIS techniques, clearly offer potential to produce a new generation of maps.

Attempts to account for the broad trends encompassed by maps such as Figures 2 and 3 and the data used in their development have involved a number of different approaches. Whilst not aiming explicitly to identify specific controls, the work of Dedkov and Mozzherin (1984) is worthy of mention as one of the few attempts to develop a global zonation of sediment yield based on morphoclimatic zones. These authors assembled sediment load data from more than 3000 measuring stations on world rivers and firstly subdivided these data into *Plains* rivers and *Mountain* rivers. Specific suspended sediment yields associated with the latter were on average more than three times greater than those for the former. They subsequently attempted to characterize each of the morphoclimatic zones within these two broad categories by typical suspended sediment yields. Yield values were given for both small ($<5000\,km^2$) and large ($>5000\,km^2$) rivers, in order to take account of the scale problem identified above. The results of this analysis presented in Figure 4 are highly dependent on the representativeness of the sediment yield data available for the different morphoclimatic zones, since the values depicted are simply mean values for the available data. Figure 4 should therefore not be viewed as a definitive representation of the global zonation of suspended sediment yields, but it affords a useful indication of the patterns involved. In the case of plains rivers,

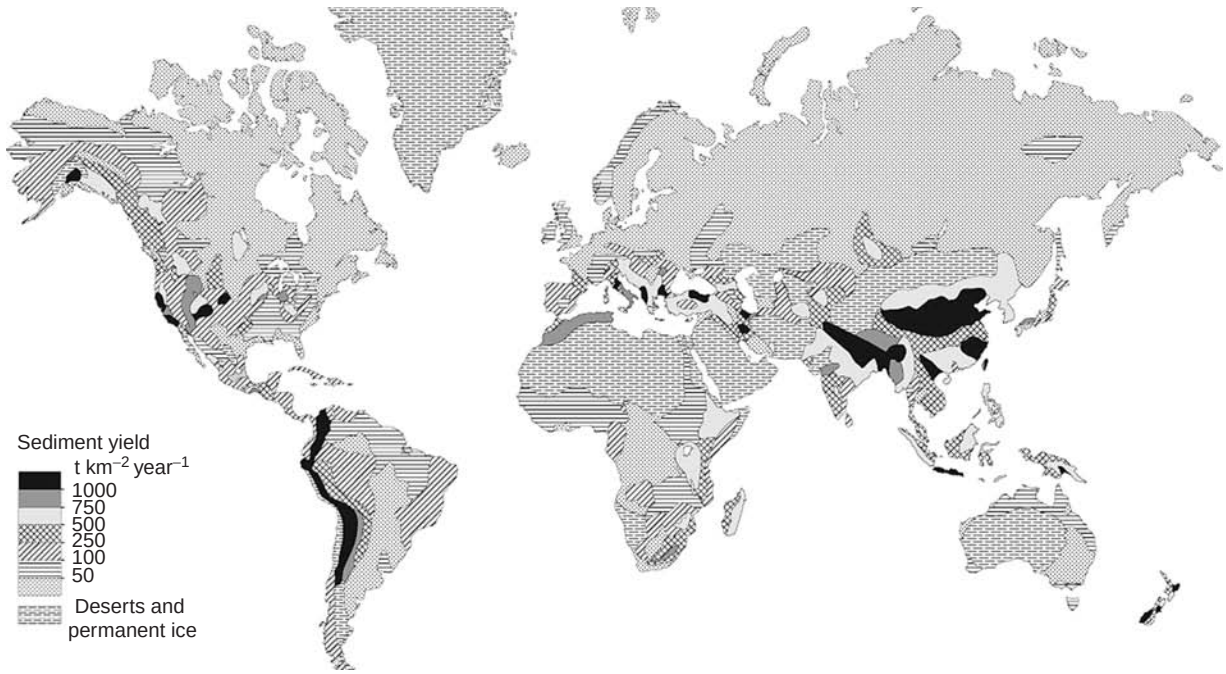


Figure 2 The world map of suspended sediment yield (Reproduced from Walling and Webb, 1983 by permission of John Wiley & Sons Ltd)

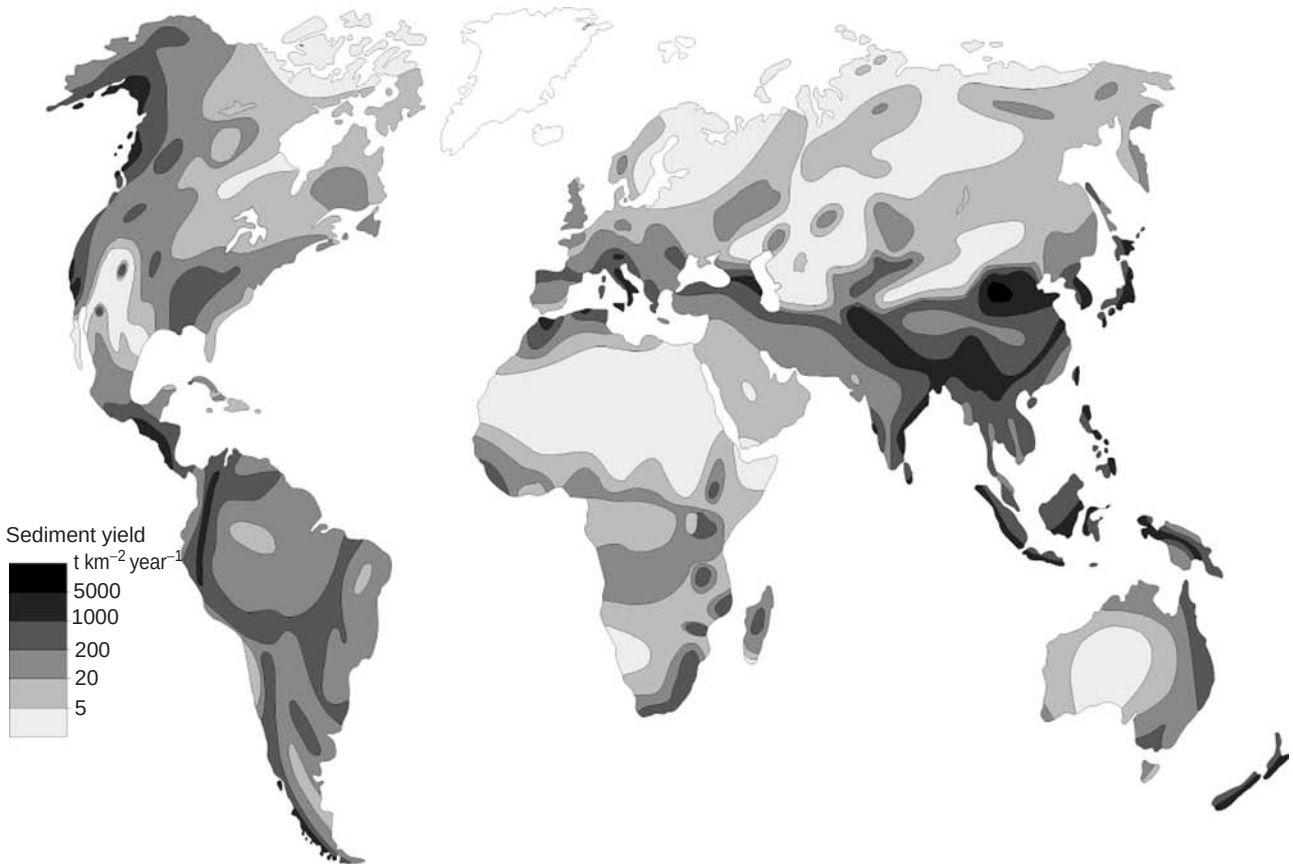


Figure 3 The world map of suspended sediment yield produced by Lvovich *et al.* (1991)

it emphasizes the relatively low specific suspended sediment yields encountered in the temperate and equatorial belts and the existence of much higher values in subtropical and tropical regions. A similar pattern is shown by mountain rivers, but in this case the rivers in the glacial zone produce the highest specific suspended sediment yields.

Attempts to identify the key controls responsible for the global patterns evident in Figures 2 and 3 and in the global zonation documented in Figure 4 have frequently focused on climatic controls and, more particularly, the relationship between specific suspended sediment yield and mean annual precipitation and runoff and measures of precipitation seasonality (e.g. Langbein and Schumm, 1958; Fournier, 1960; Douglas, 1967; Wilson, 1969; Walling and Webb, 1983; Jansson, 1988). However, the availability of larger and more representative databases has demonstrated that there is no simple and well-defined relationship between specific sediment yield and annual rainfall and runoff (cf. Walling and Webb, 1983). The relationships between mean annual specific sediment yield and effective precipitation (defined as the annual precipitation required to generate the given annual runoff, assuming an annual mean temperature of 50°F) and mean annual runoff proposed by Langbein and Schumm (1958) and Douglas (1967), respectively, are, nevertheless, presented in Figure 5, since they have been widely cited in the literature. The relationship reported by Langbein and Schumm (1958) has been referred to as the "Langbein–Schumm Rule". This "rule", which provides a convincing explanation of the trend shown in Figure 5(a), highlights the interaction of changes in erosive energy and vegetation cover density associated with changing annual precipitation. Maximum suspended sediment yields occur in areas with an annual effective precipitation of approximately 300 mm (i.e. semiarid areas). In areas with increased precipitation, the vegetation cover density increases, causing reduced erosion and sediment yield. In drier areas, the available energy for water erosion is also limited, resulting in reduced sediment yields. The relationship reported by Douglas (1967) (Figure 5a) evidences a similar peak for semiarid areas (i.e. for a mean annual runoff of ca. 60 mm), but in this case sediment yields increase again when mean annual runoff exceeds ca. 500 mm, suggesting that the increased erosive energy overrides the effect of any increase in vegetation cover density for higher values of mean annual runoff.

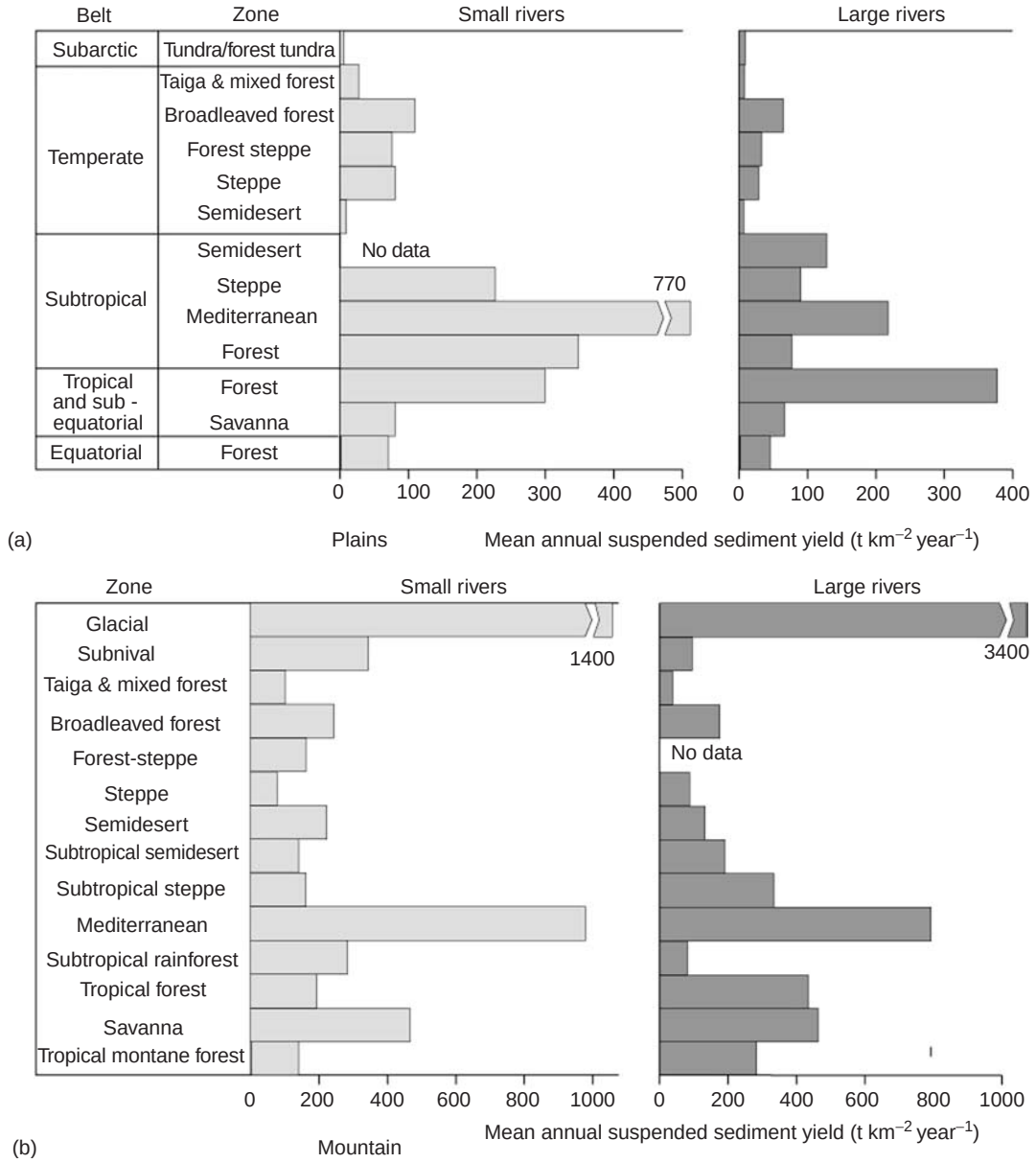


Figure 4 The global zonation of suspended sediment yield proposed by Dedkov and Mozzherin (1984)

The lack of a clear relationship between specific suspended sediment yield and annual precipitation or runoff demonstrated by Walling and Webb (1983) reflects the important role of many other factors in controlling global variations in sediment yield. Pinet and Souriau (1988), Milliman and Syvitski (1992), and Summerfield and Hulton (1994) have, for example, emphasized the role of relief in accounting for global variations in sediment yield. Based on a dataset comprising the suspended sediment loads of 33 major world rivers, Summerfield and Hulton (1994) demonstrated that the highest degree of statistical explanation was provided by variables reflecting catchment relief, as represented by the basin relief and relief ratio. Whereas basin relief accounted for 64% of the total variance of the dataset, mean annual runoff accounted for only 20% of the variance. Lvovich *et al.* (1991) further demonstrated the important role of variables other than annual precipitation or runoff by analyzing a very much larger database of sediment loads for world rivers. They found a positive log/log relationship between specific sediment yield and mean annual runoff, but showed that the intercept of this relationship varied markedly between different regions, thereby masking any general global relationship.

A useful demonstration of the complex set of controls influencing the global pattern of suspended sediment yield is provided by the work of Ludwig and Probst (1996). These

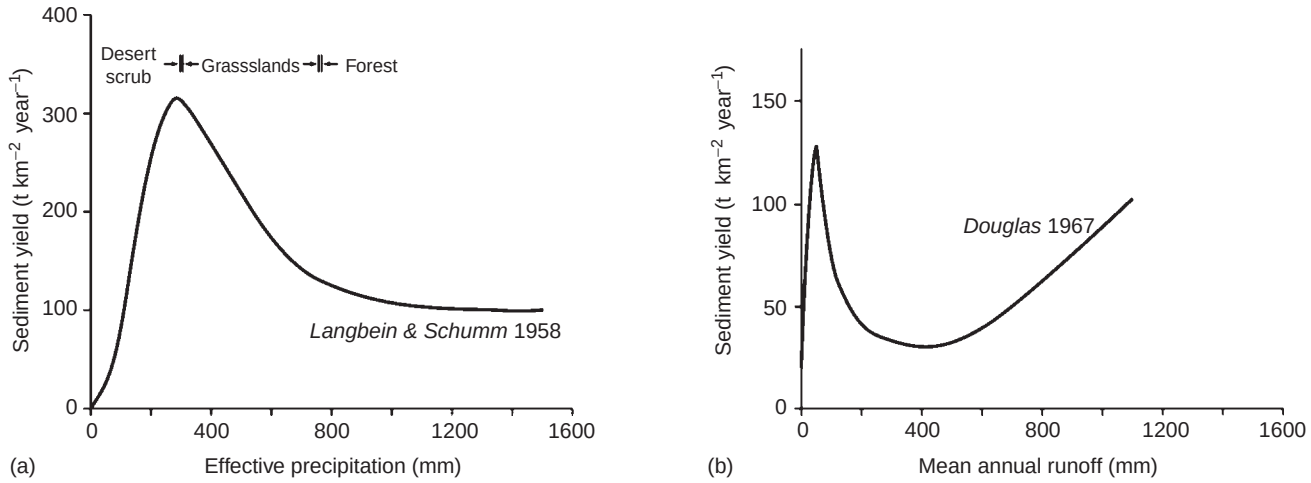


Figure 5 The relationships between mean annual suspended sediment yield and mean annual effective precipitation and mean annual runoff proposed by Langbein and Schumm (1958) and Douglas (1967) (Reproduced from Langbein and Schumm (1958) by permission of American Geophysical Union)

authors assembled specific sediment yield data for 60 major world river basins which accounted for about 50% of the total exoreic continental area and were representative of the major ecosystems of the land surface of the globe. Environmental data, including variables such as mean annual precipitation and runoff, the seasonal variability of precipitation, basin slope, lithology, biomass density, and percentage cultivated area were extracted from a range of global data sets at a grid point resolution of $0.5° \times 0.5°$ longitude/latitude and area-weighted values were obtained for each river basin. The significant correlations between the values of specific suspended sediment yield for the 60 river basins and the different environmental variables are summarized in Figure 6. Figure 6 indicates that mean annual runoff has the strongest (positive) correlation with specific sediment yield and the next strongest correlations, which again are all positive, are with mean annual precipitation, precipitation variability, biomass density, lithology (erodibility) and basin slope. These results are, however, sensitive to multicollinearity or interdependence between the variables, so that the positive relationship between mean annual suspended sediment yield and biomass density, which runs counter to what might be intuitively expected in terms of the influence of vegetation cover density on erosion rates, probably reflects the close association between biomass density and mean annual runoff. This and other significant intercorrelations between the independent variables are also shown on Figure 6. There is clearly scope to extend the analysis reported by Ludwig and Probst (1996) to include more river basins and to incorporate more rigorous multivariate analysis, in order to identify the "real" effect of the individual variables. Furthermore, since most of the river basins represented were relatively large (between $5903 \times 10^3\,km^2$ and $9 \times 10^3\,km^2$) there is again scope to include smaller basins which are more representative of global variability in the key controlling variables. However, the analysis does serve to emphasize the multiple controls on the patterns evident in Figures 2 and 3.

Further evidence of the relative importance of controls such as mean annual rainfall or runoff and basin geology can be gleaned from the large number of regional and national studies documented in the literature, which have attempted to develop predictive relationships. The results from two such studies undertaken in New Zealand (Hicks *et al.*, 1996) and the Maghreb region of North Africa (Heusch and Milliés-Lacroix, 1971) are shown in Figure 7(a) and 7(b). The results from New Zealand clearly demonstrate the role of mean annual rainfall and rock type in controlling suspended sediment yields in that country, since variation in either of these controls can cause sediment yields to vary over three orders of magnitude. A similar, albeit slightly less marked trend, is evident for the Maghreb region of North Africa. In their study of the spatial variability of suspended sediment yields in New Zealand, Hicks *et al.* (1996) suggest that the influence of mean annual rainfall and rock type, with the latter reflecting both lithology and tectonic instability, explains the broad pattern of sediment yields, and that other variables such as catchment slope and secondary climatic parameters that index the seasonality operate to cause further local variation within the broad pattern produced by variations in mean annual precipitation and rock type. In some areas, land use will also exert a key control over the pattern of suspended sediment yield, since, if other controls are held constant, the contrasts between, for example, forest areas and areas with intensive cultivation could be reflected by differences in sediment yield involving several orders of magnitude. This situation is well illustrated by the general analysis of land

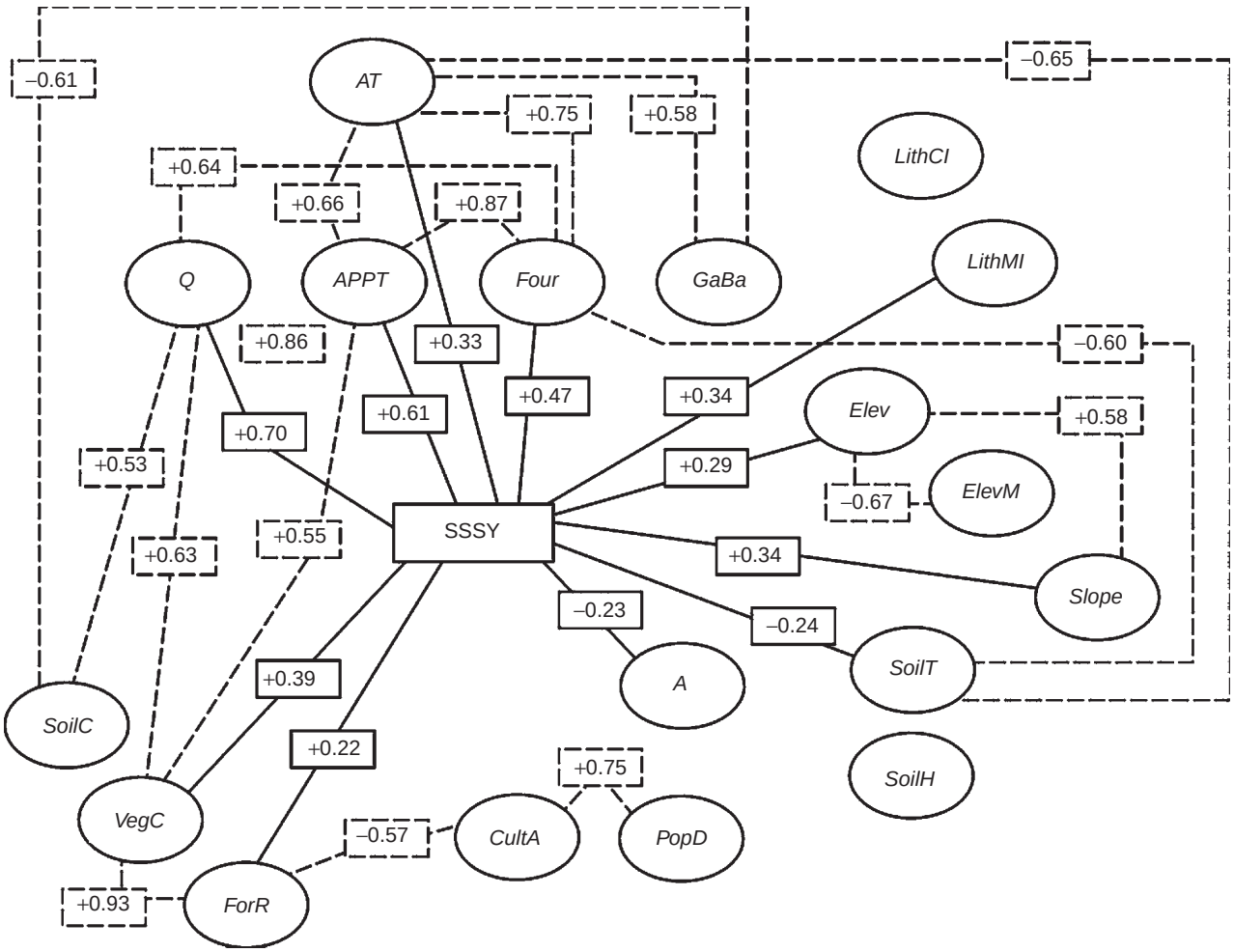


Figure 6 The correlation (r values) between mean annual specific suspended sediment yield (SSSY) and a range of catchment characteristics for 60 major world river basins (the solid lines show correlations significant at $P < 0.1$). Correlations between the catchment characteristics are shown by dashed lines, with only correlations with r values < -0.5 and $> +0.5$ depicted. *Q* is mean annual runoff depth; *AT* and *APPT* are the mean annual temperature and mean annual precipitation, respectively; *Four* is a modified form of the Fournier index of intra-annual variability of precipitation: *GaBa* is an aridity index based on monthly precipitation and temperature data; *LithMl* and *LithCl* are measures of the erodibility of the dominant basin lithology, with respect to mechanical and chemical erosion, respectively; *Elev* is the mean modal basin elevation and *ElevM* is the maximum basin elevation; *Slope* is the average basin slope; *SoilT* is a measure of soil erodibility based on soil texture; *SoilH* is the average soil depth; *A* is the basin area; *SoilC* is the average organic carbon content of the soil; *VegC* is the average biomass density; *ForR* is the proportion of forest in the basin; *CultA* is the proportion of cultivated land in the basin and *PopD* is the mean population density in the basin (Based on Ludwig and Probst, 1996, with permission from IAHS Press)

use effects on sediment yields in Kenya reported by Dunne (1979), which is presented as Figure 7(c). In this region, contrasts in land use can be seen to induce a two- to three-fold variation in specific sediment yield, for a given value of mean annual runoff.

The Role of Drainage Basin Area

As indicated above, there is a general expectation that specific suspended sediment yield will decrease as catchment area increases (see Walling, 1983; Milliman and Syvitski, 1992). This inverse trend is commonly accounted for in terms of the increased opportunity for deposition and storage as travel distances increase and the sediment is transported into areas with reduced slope gradients and well-developed floodplains. Thus the probability that a sediment particle will be deposited increases with increasing transport distance and therefore catchment area. This effect is frequently paralleled by a general decrease in specific sediment yield from downstream areas, due to reduced relief and slope gradients and reduced precipitation and runoff, such that the average sediment yield per unit area also decreases as catchment area increases. Figure 8(a)

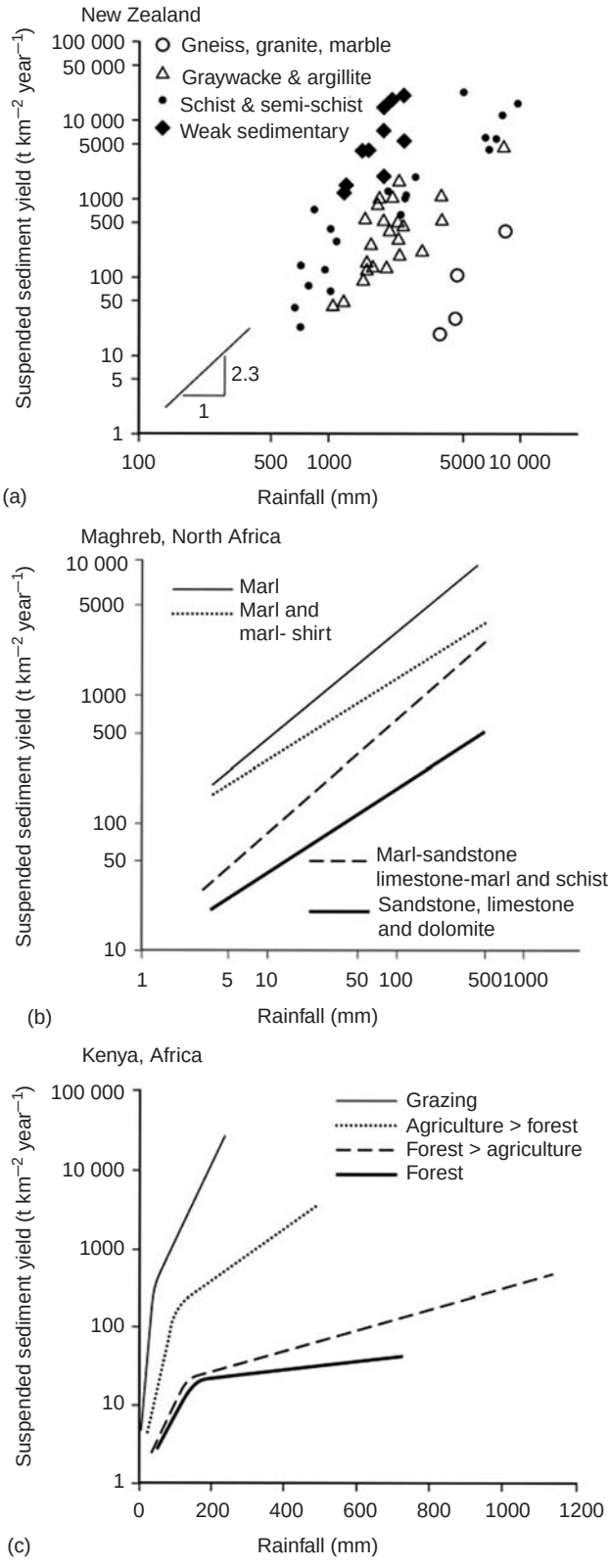


Figure 7 Regional relationships between mean annual suspended sediment yield and controlling factors reported (a) by Hicks *et al.* (1996), (b) by Heusch and Milliés-Lacroix (1971), and (c) by Dunne (1979)

presents a selection of the typical inverse relationships between specific suspended sediment yield and catchment area that have been widely reported in the literature.

The simple inverse relationship between specific sediment yield and catchment area has, however, been questioned by a number of recent studies. Church and Slaymaker (1989), for example, suggest that in British Columbia, Canada, specific sediment yield frequently increase downstream up to a catchment area of c. 3×10^4 km^2, as a result of remobilization of Quaternary sediments stored in the valley and channel systems. It is also relatively easy to conceive of situations where deviations for the "standard" inverse relationship could be accounted for by specific local conditions. For example, sediment yields in headwater areas characterized by resistant rocks and good vegetation cover could be much lower than in downstream areas developed on softer, more erodible rocks and characterized by more intensive land use, and in such circumstances specific sediment yields could increase downstream.

A more fundamental questioning of the traditional view is, however, provided by Dedkov and Mozzherin (1992) and Dedkov (2004), who have suggested that river systems will be characterized by either positive or negative relationships between specific sediment yield and catchment area, according to the relative importance of channel and slope erosion. Where slope erosion (i.e. sheet, rill, and gully erosion) represents the dominant sediment source, most of the erosion will be concentrated in the headwater areas and conveyance losses associated with transfer of the sediment through the basin will result in an inverse relationship between specific sediment yield and basin area. However, where channel erosion is dominant, as for example in forested areas with dense vegetation cover, which protects the surrounding slopes, erosion rates are likely to increase downstream in response to increasing entrainment and transport of sediment associated with increased water discharge. Specific sediment yield will therefore increase downstream, resulting in a positive relationship between specific sediment yield and basin area. Examples of both situations are presented in Figure 8(b) which plots the relationship between the downstream change in total suspended sediment yield (i.e. downstream/upstream) and the equivalent change in catchment area, for a number of river basins where there are several measuring stations along the main river. For those basins with limited cultivation (i.e. where channel erosion is likely to be dominant), the total sediment load increases more rapidly than the catchment area, whereas the reverse situation exists for intensively cultivated catchments (where sheet, rill and gully erosion are likely to be dominant). Dedkov and Mozzherin (1992) and Dedkov (2004) based their conclusions on a study of sediment yield data from a range of morphoclimatic zones and suggested that the positive relationship, indicative of the dominance of channel erosion, was typical of forested areas

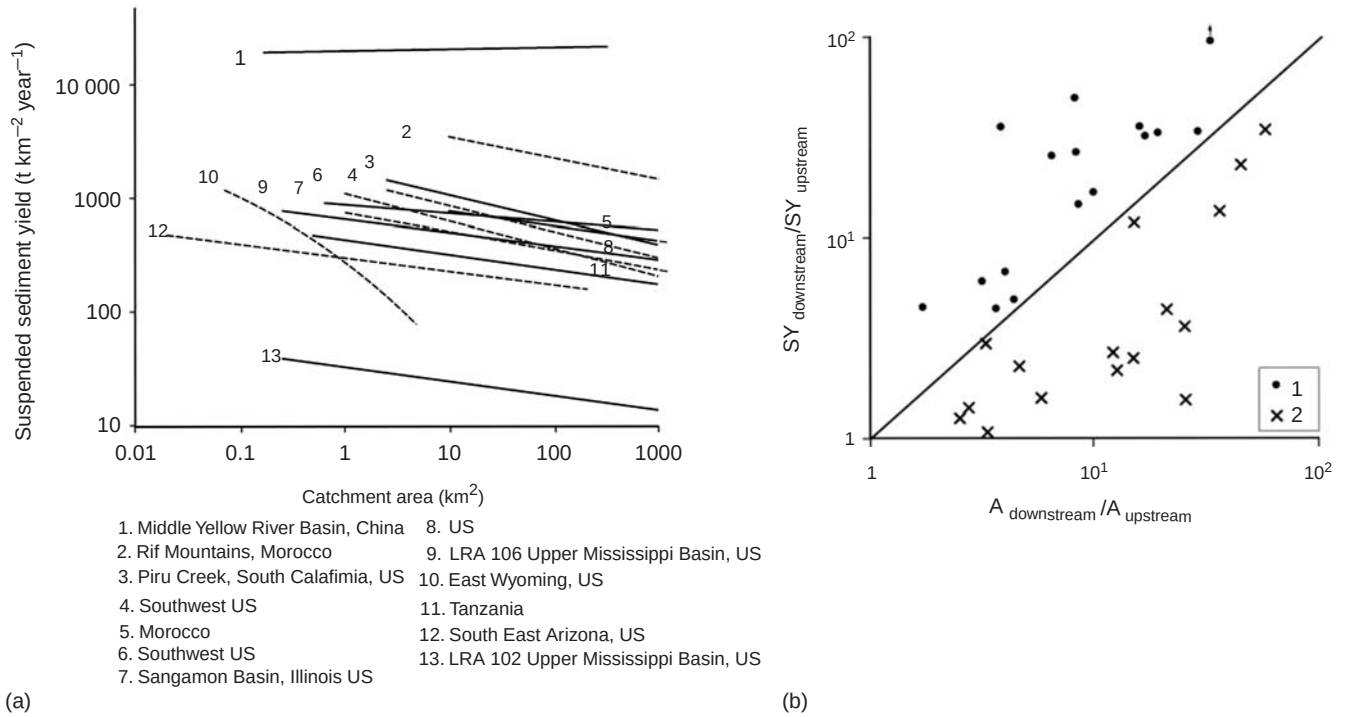


Figure 8 Relationships between mean annual specific suspended sediment yield and catchment area reported by Walling (1983) and the relationship between the downstream increase in total suspended sediment yield and the downstream increase in catchment area for catchments where cultivated land is limited (1) and where it predominates (2) (Based on Dedkov, 2004, with permission from IAHS Press)

with an undisturbed vegetation, whereas the inverse relationship is more characteristic of natural zones with poor vegetation cover and strong surface erosion (e.g. glacial, subnival, and semiarid zones) and of areas disturbed by human activity (e.g. agriculture). The traditional emphasis on an inverse relationship between specific sediment yield and catchment area may therefore reflect the preponderance of results derived from areas such as the United States, which are heavily impacted by human activity, rather than a more universal principle.

Temporal Variability in Suspended Sediment Yields

The suspended sediment loads of most rivers will exhibit considerable variation through time, in response, for example, to individual flood events, discharge magnitude, and the overall seasonal regime. Meybeck *et al.* (2003) assembled records of daily suspended sediment loads from 60 monitoring stations in different areas of the globe and used the ratio of the mean daily specific flux to the median daily specific flux to characterize such variability. The values obtained ranged from 1–2 in the case of large river basins such as the Loire, Mississippi, Rio Grande, and Seine, to >50 for the Eel, the Walla Walla and the Matanuska in the United States and the Red Deer in Canada. The daily load data were also used to establish the percentage of time needed to transport 50% of the total sediment flux and the values obtained ranged from <16.5% for the Middle Rhine, Mississippi, St Lawrence, and Somme, to <0.4% for the Bermejo and Elvira in Bolivia and the Walla Walla in the United States. These results indicate that temporal variability decreases with increasing basin size and that the most variable flux regimes are associated with small- to medium-sized basins in steep mountainous terrain. Such analysis emphasizes the short-term variability of suspended sediment transport, but this should be seen as reflecting the transport regime, rather than the longer-term pattern of sediment yield, where the variability of annual sediment yields is of greater interest.

In considering the temporal variability of annual suspended sediment fluxes or yields, it is necessary to consider both the natural variability and their response to anthropogenic influences, such as land use change and reservoir construction. As indicated above, the coefficient of variation of the annual sediment yields from a given river basin commonly exceeds that of the annual runoff or discharge and values >100% are not uncommon. Walling and Kleo (1979) report a study in which the coefficient of variation was calculated for the records of annual sediment yield available for more than 250 rivers located in

different parts of the world. The analysis was limited by the short period of record available for many of the stations, but the results suggest that a coefficient of variation of 70% is typical of annual sediment yield data, although the values ranged between 20% and 214%. No clear relationship was found between the coefficient of variation and mean annual precipitation, although there was some evidence to suggest that CV values were higher in semi-arid areas. For many rivers, interannual variability will reflect an essentially regular pattern of variability about the mean, whereas in others low frequency, high magnitude events, such as those associated with hurricanes or cyclones, can represent a key component of the variability. Figure 9 presents two examples of such situations, highlighted by the work of Meade *et al.* (1990). In the case of the 8700 km^2 Juniata River in Pennsylvania, USA (Figure 9a), the extreme flood associated with Hurricane Agnes in 1972 transported more sediment in a few days than would normally be transported in three or four years and the annual sediment yield for 1972 is very considerably higher than in any other year during the 33 year record. Another example is provided in Figure 9(b) for the Eel River in California (8100 km^2), where an intense Pacific storm, occurring in December, 1964 (1965 water year) transported more sediment in three days than would normally be transported in 10 average years and introduced a major perturbation into the time series of annual suspended sediment yields. Although it falls outside the direct evidence provided by river monitoring, another even more striking example of the potential for extreme events to distort the pattern of annual sediment yield from a river basin is provided by a study reported by Milliman *et al.* (1996) that used offshore sediment deposits to reconstruct the sediment discharges associated with Jøkulhaup floods from the Alsek River, which drains a 28 000 km^2 basin in Alaska and Yukon Territory. Based on the offshore sediment deposits, the average sediment yield from the Alsek River over the past 3000–4000 years was estimated to be $6–8 \times 10^6$ t year^{-1}. However the loads associated with individual Jøkulhaup events were estimated to be more than two orders of magnitude higher at 2.5×10^9 t year^{-1}. In such events, the sediment load transported by this relatively small river would be equivalent to ca. 10% of the current mean annual land–ocean sediment flux.

Land clearance, land use change and other human activities, including reservoir construction, can also result in major perturbations in longer-term records of suspended sediment flux. The potential impact of land clearance and associated population growth and use change on suspended sediment yields is clearly demonstrated by an analysis of reservoir sedimentation data for a number of reservoir catchments in Southeast Asia reported by Abernethy (1990) and presented in Figure 10. In this case, annual rates of increase of sediment yield were estimated to lie within the range 2.48–6.02% year^{-1} and Abernethy (1990) suggested that these increases closely paralleled the rate of population increase in the catchment, although the ratio of the rate of increase of sediment yield to that for population was significantly greater than unity and about 1.7 : 1 (Figure 10). Based on this evidence, he suggested that annual sediment yields in some developing countries could be expected to double in about 20 years. Whilst this example provides

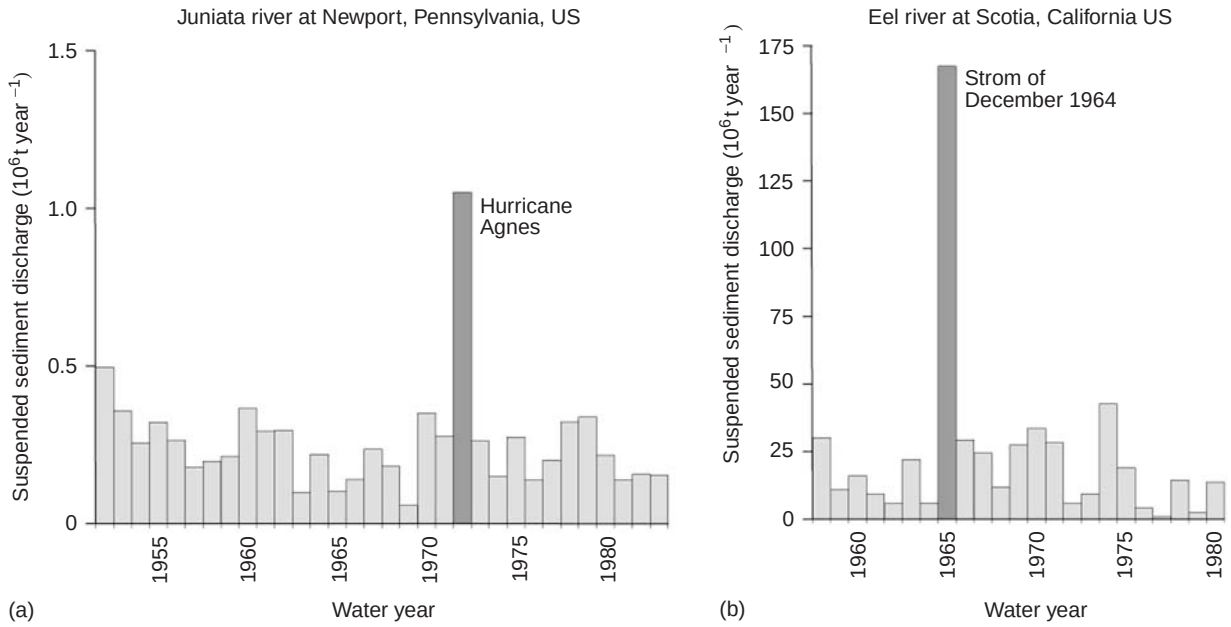


Figure 9 The impact of extreme events on annual suspended sediment yields (Based on Meade *et al.*, 1990)

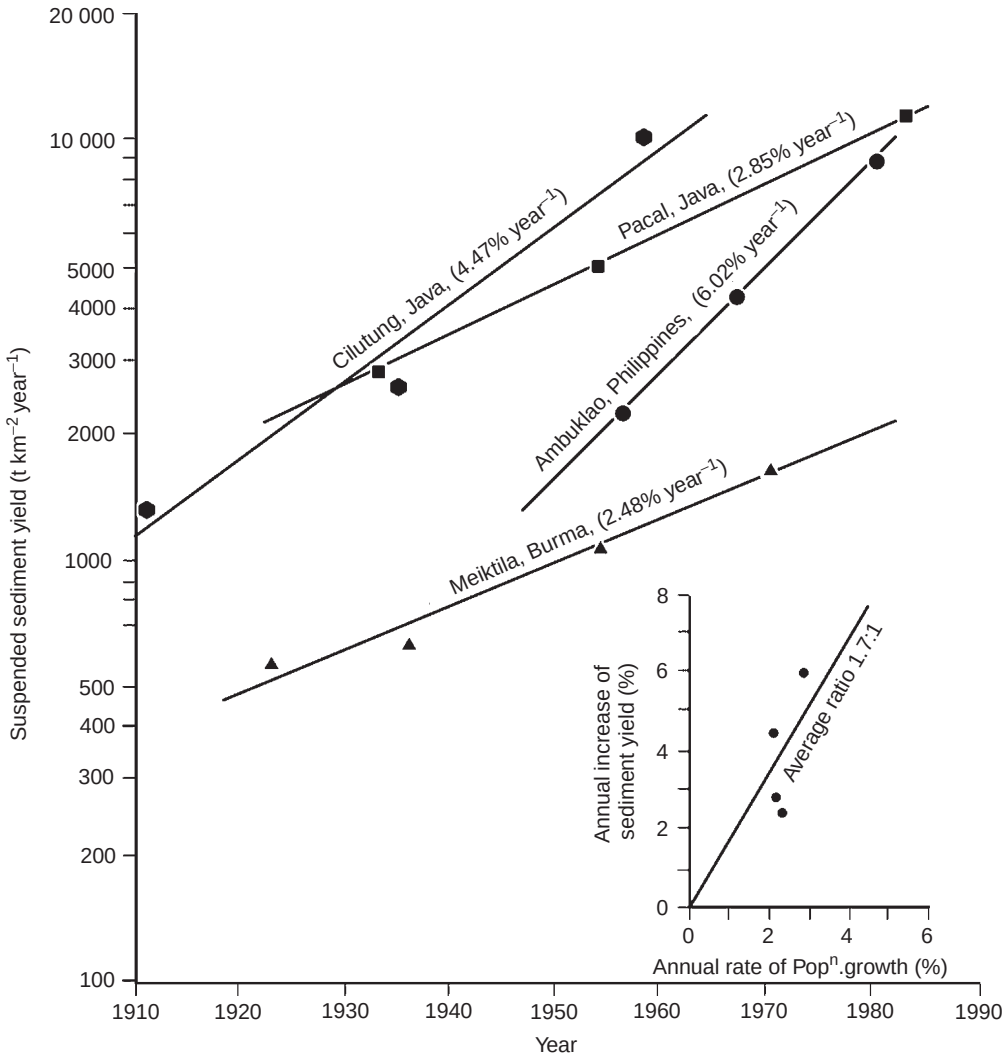


Figure 10 Increases in sediment yield during the twentieth century in selected reservoir catchments in Southeast Asia. (Figure drawn by Dr C. Abernethy)

clear evidence of an increase in sediment yield caused by land clearance and population growth, it is important to note that such trends are absent from other larger rivers where they might have been expected. One example of this situation is provided by the 14 028 km^2 Chao Phraya basin draining the highlands of northern Thailand, for which Alford (1992) found no evidence of an increase in sediment yield during the period extending from the late 1950s to the mid-1980s, despite substantial deforestation and extensive land clearance by swidden agriculture within the catchment. A similar situation is provided by the Upper Yangtze River at Yichang, China, which drains an area of 1005×10^3 km^2 in central and western China (cf. Dai and Tan, 1996; Walling and Fang, 2003). Although the population of this area has more than doubled since the early 1950s and there has been an major expansion of agricultural activity within the basin, the time series of annual sediment yields for the period 1950–1991 is essentially stationary and shows no evidence of change. Walling (2000) has suggested that the absence of a significant change in sediment yield from these and other large river basins could reflect the influence of "buffering", whereby any increases in sediment mobilization are attenuated by storage within the fluvial system.

An analysis of available longer-term records of annual sediment yield from more than 140 rivers in different areas of the world reported by Walling and Fang (2003), nevertheless, provides clear evidence of significant temporal trends in the sediment yields of many river basins, as a result of human activity. In some cases, such as the River Danube in Central Europe (Figure 11a), reservoir

construction has caused a significant reduction in the annual sediment flux. In this case, the annual suspended sediment load has decreased by about 70% over the period extending from the mid-1960s to the 1990s, although the annual runoff has remained essentially stable. The progressive flattening of the double mass plots presented in Figure 11(a) suggests that the decrease in the sediment flux has continued to intensify through to the mid-1990s. Examples of rivers where annual sediment yields show a significant increase in recent decades, despite annual runoff totals remaining

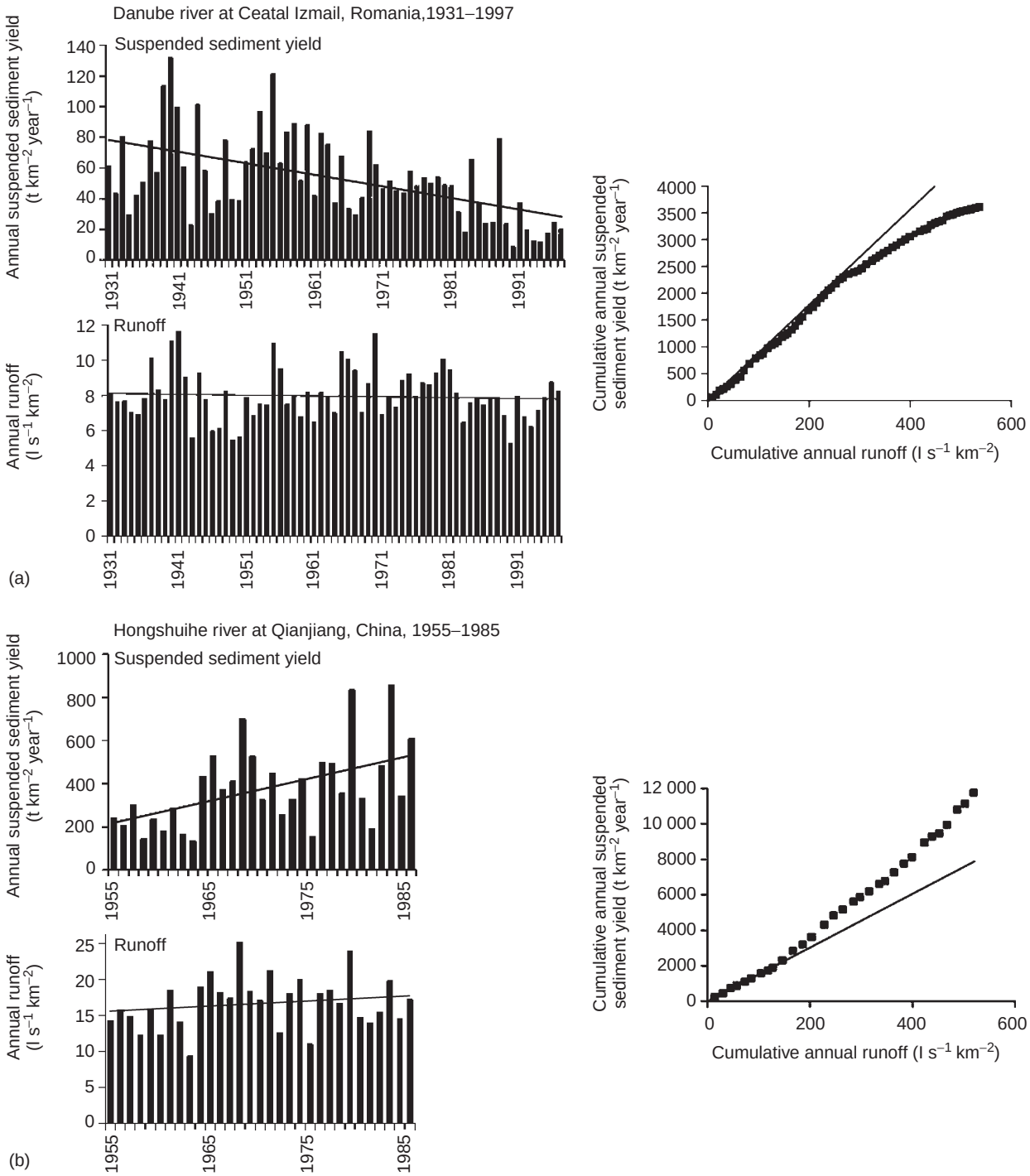


Figure 11 Recent trends in annual suspended sediment yield and runoff for the River Danube, Romania, the Hongshuihe River, China, and the Kolyma River in Western Siberia (Based on data presented by Walling and Fang, 2003)

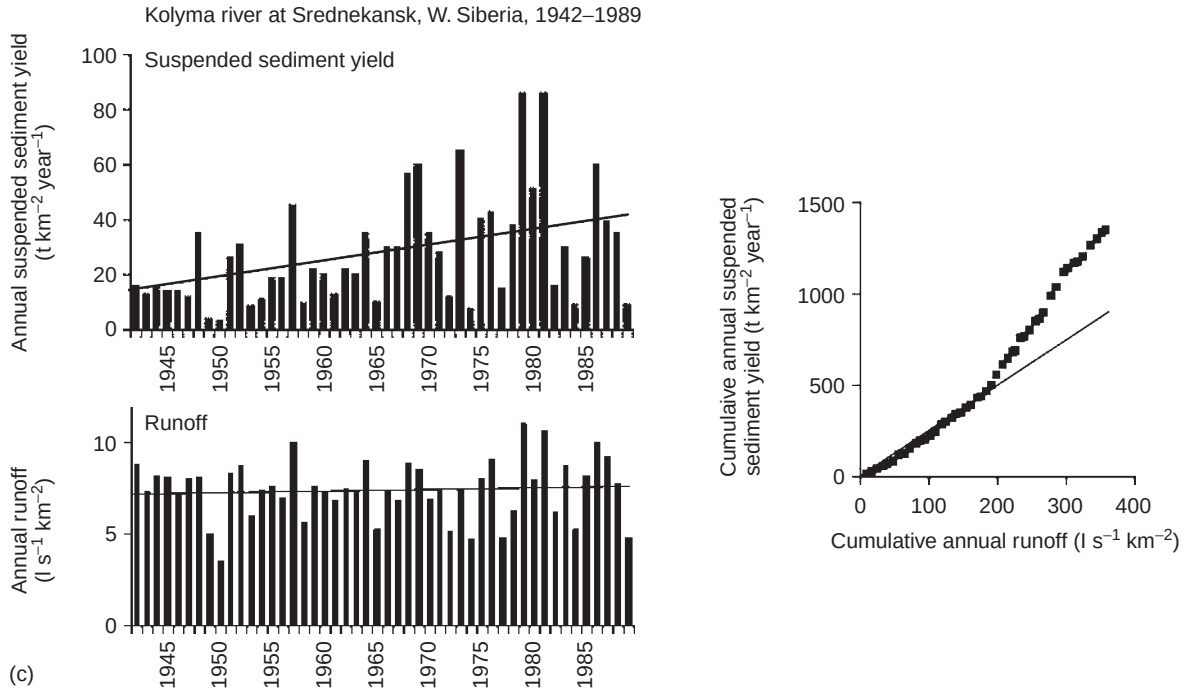


Figure 11 (*continued*)

essentially stationary, are shown in Figures 11(b) and 11(c). The Hongshuihe River, a tributary of the Pearl River in Guangxi Province, China, which drains a catchment of 1 28 165 km^2, shown in Figure 11(b) provides clear evidence of increasing sediment yields, with loads increasing by ca. 75% over the period of record. In this case, population growth can be linked to both an expansion of the area of agricultural land and intensification of land use activities over the period of record, which in turn have resulted in increased rates of soil loss and sediment yield. The double mass plot suggests that the increase in sediment flux commenced in the mid-1960s and its progressive steepening further suggests that the impact of land use intensification has continued through to the present. The other example shown in Figure 11(c) relates to the Kolyma River, which drains a 99 400 km^2 catchment in western Siberia. The runoff record for this river again shows no significant trend over the 50 year period, but the sediment yields show a significant increase and the double mass plot indicates that sediment yields have more than doubled since the mid-1960s. This increase has been attributed to widespread gold mining in the catchment, which disturbs the catchment surface and greatly increases sediment mobilization by erosion (cf. Bobrovitskaya, 1996). The specific suspended sediment yield of this basin under "natural" conditions was relatively low (ca. 20 t km^{-2} year^{-1}) and the disturbance caused by mining activity has caused a substantial relative increase in the total sediment load.

Other rivers can be expected to show evidence of changes in both runoff and sediment yield and Walling and Fang (2003) cite the example of the Dnestr River in the Ukraine (850 km^2) where both runoff and sediment yield have shown a significant upward trend since the 1960 s in response to forest clearance and the expansion of agriculture as well as a shift to wetter conditions. Conversely, the Yellow River at Sanmenxia (667 948 km^2), the gauging station located close to where the river leaves the loess region and enters the North China Plain, shows a statistically significant decline in both runoff (ca. 13%) and sediment yield (ca. 50%) since the 1960s, in response to drier conditions, increased water abstraction and a comprehensive soil and water conservation and sediment control programme in the loess region, the main source of the high sediment load transported by the river. In the case of the sediment load, the shift to drier conditions accounted for ca. 50% of the decrease and the remaining ca. 50% of the decrease was attributed to the impact of soil and water conservation and sediment control works. Land use change can result in reduced sediment yields as well as the more commonly cited increase.

SEDIMENT BUDGETS

The Context

Although the sediment yield from a river basin will reflect the rates of sediment mobilization or soil loss within the

basin, it is important to recognize that only a proportion, and probably only a small proportion, of the mobilized sediment will reach the catchment outlet. Much of the sediment mobilized within the basin will be deposited and stored at the foot of slopes, in depressions and valley floors, in channels, on floodplains and in backswamps, as well as in ponds, wetlands, lakes, and reservoirs. The inverse relationships between specific sediment yield and catchment area presented in Figure 8(a) have already been identified as reflecting the increasing significance of conveyance losses and sediment storage as catchment area increases. Such conveyance losses and sediment storage must clearly be taken into account when attempting to use measurements of sediment yield to provide an indication of rates of soil loss or soil degradation within a drainage basin or to predict the impact of climate or land use change or other causes of changing catchment response on sediment yields. Equally, where there is a need to develop a sediment management strategy for a catchment, aimed at reducing downstream sediment fluxes, it is necessary to recognize that reducing erosion rates may not result in an equivalent reduction in sediment flux at the catchment outlet, due to storage effects. Furthermore, it is possible that a reduction in sediment input to the channel system could result in increased transport capacity and therefore remobilization of stored sediment, such that the sediment flux at the catchment outlet remains essentially unchanged. An integrated view of sediment mobilization, transfer and output is necessary when interpreting sediment yields in terms of sediment mobilization and delivery within the upstream basin and assessing the potential for reducing the sediment output from a river basin. In addition, it will frequently be necessary to recognize that the sediment output from a catchment can be derived from different sources, for example, erosion of the catchment surface by sheet and rill erosion, gully erosion, channel erosion, and mass movements transporting sediment directly to the channel system. The different pathways followed by sediment mobilized from the different sources are likely to offer different opportunities for sediment storage and associated flux attenuation, such that control of certain sources may have a greater impact on the downstream flux than control of others.

The Sediment Delivery Ratio

Early attempts to take account of the importance of storage in influencing the relationship between sediment mobilization and sediment output from a catchment focused on the simple concept of a sediment delivery ratio (SDR), which expresses the relationship between gross erosion within a drainage basin and the sediment output from that basin (see e.g. Glymph, 1954; Maner, 1958; Walling, 1983). Since the SDR will, by definition, be <1, it implicitly takes account of storage. The magnitude of the SDR for a particular basin will be influenced by a wide range of catchment characteristics, including the nature and location of the sediment sources, relief, and slope characteristics, the drainage pattern and channel conditions, vegetation cover, land use, and soil texture. A number of authors have attempted to establish predictive relationships that can be used to estimate the SDR for a drainage basin, and an inverse relationship between SDR and catchment area has been widely proposed (cf ASCE, 1975; Walling, 1983), such that the SDR for a small $0.1\,km^2$ basin might be as high as 0.5, whereas it could decrease to about 0.1 for a $1000\,km^2$ basin. However, many uncertainties surround any attempt to establish the SDR for a catchment, not least the difficulty of establishing the magnitude of the gross erosion, which must be compared with the measured sediment yield, in order to calculate the SDR. In many studies, the gross erosion has been quantified by using the Universal Soil Loss Equation (USLE) or a similar predictive tool to estimate sheet and rill erosion and correcting (increasing) this value to take account of additional contributions from channel and gully erosion. Uncertainties associated with the resulting values of gross erosion will be propagated through to the estimates of SDR. In addition, the SDR must be seen as a lumped black box concept, which fails to take account of both spatial and temporal variability in sediment delivery (see Walling, 1983). Attempts to provide a more explicit representation of storage, as well as contributions from different sources and different transfer pathways, have resulted in the concept of a sediment budget. This links sources, sinks, and output by a mass balance equation, namely,

$$\text{Output} = \text{Input(from sources)} \pm \Delta\text{Storage} \qquad (1)$$

and, equally importantly, provides a schematic representation of the transfer of sediment between sources and sinks and the catchment outlet (see Figures 12 and 13).

Sediment Budgets

Although it has been suggested that the concept of a sediment budget can be traced back to the early work of G K Gilbert, when investigating the impact of mining activity on the Sacramento River in California (cf. Gilbert, 1917), the credit for this concept is commonly accorded to Jackli (1957) and Rapp (1960), who were amongst the first to attempt to document rates of sediment mobilization and transfer within small catchments. Their work was extended by many subsequent workers, and particularly by Dietrich and Dunne (1978), who produced a comprehensive representation of the sediment budget of the $16.2\,km^2$ Rock Creek catchment in the Oregon Coast Range. Interestingly, all three of these studies were undertaken in mountain environments and their perspective was primarily geomorphological, aimed at understanding rates

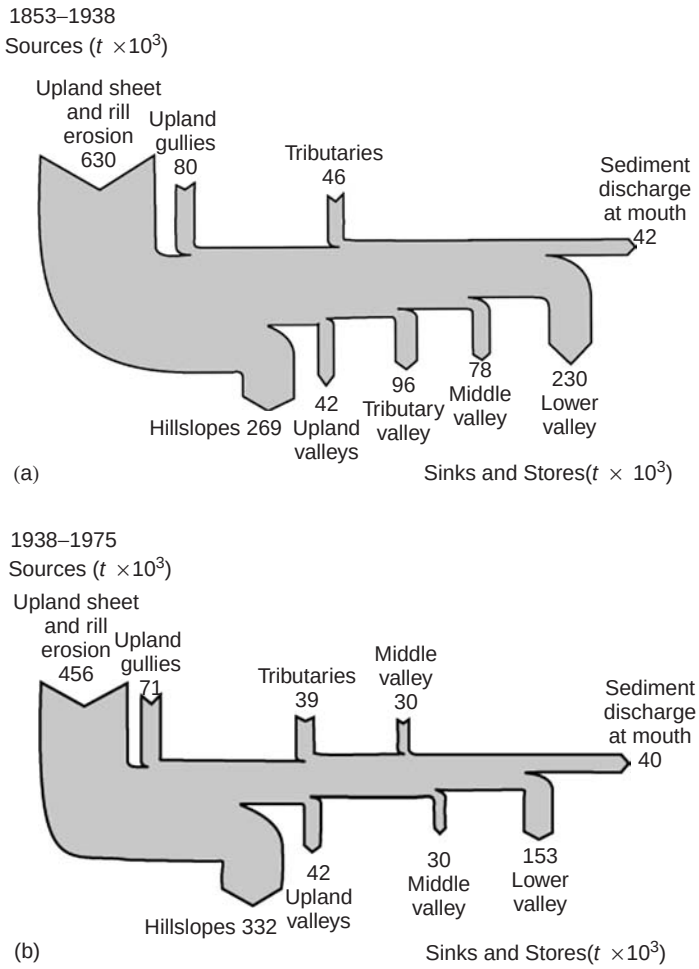


Figure 12 The sediment budgets for Coon Creek, Wisconsin, USA, for the periods 1853–1938 and 1938–1975 produced by Trimble (1983). The fluxes indicated represent mean annual values

of landscape development. As such they included chemical weathering and solute transport as well as the mobilization, transfer, and storage of both coarse (bed load) and fine (suspended load) clastic sediments. Arguably, it was the work of Meade and Trimble (e.g. Meade and Trimble, 1974; Meade, 1982; Trimble, 1976, 1983) and others in demonstrating the important role of alluvial storage within the channel and floodplain systems of drainage basins in eastern and central United States, that focused attention on the important role of such storage in influencing suspended sediment yields and promoted the wider application of the sediment budget concept to hydrological investigations of erosion and sediment delivery in drainage basins.

Figure 12 presents results from the classic work of Trimble (1983) in establishing a sediment budget for the 360 km^2 basin of Coon Creek, Wisconsin, USA, for both the period of intensive agriculture and severe soil erosion that followed the expansion of agriculture into the area and the subsequent period when soil conservation measures were introduced. A key feature of the sediment budget for the first period (1853–1938) is the relatively small proportion of the sediment mobilized within the catchment by erosion that reached the catchment outlet (i.e. ca. 5%). Large amounts of sediment were stored in colluvial deposits, upland, and tributary valleys and the main valley and did not reach the catchment outlet. The sediment budget for the subsequent period indicates that, although the introduction of soil conservation measures reduced erosion rates from the slopes of the catchments by ca. 25%, the sediment output was little changed due to the increased efficiency of sediment transfer through the channel network and remobilization of sediment stored within the middle valley, during the phase of accelerated soil loss. This example clearly demonstrates the significant improvement in understanding of the linkage between erosion and sediment mobilization, sediment delivery, and conveyance losses and sediment yield afforded by the sediment budget when compared with the sediment delivery ratio concept.

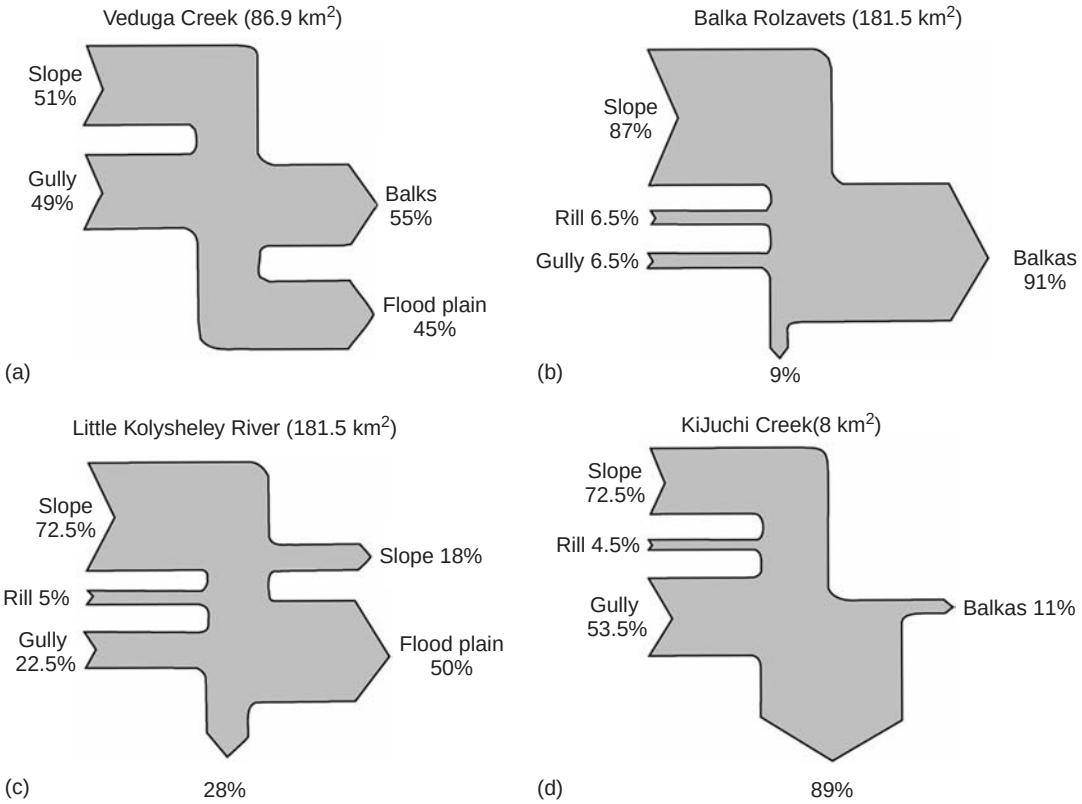


Figure 13 The sediment budgets for four drainage basins on the Russian Plain, documented by Golosov *et al.* (1992)

Furthermore, it usefully emphasizes why an understanding of the overall sediment budget of a river basin is essential when assessing the likely impact of changes in land use on sediment yields and the potential for implementing soil conservation and other sediment control measures to reduce downstream sediment yields. In this case, a reduction in rates of soil loss in the upstream areas by ca. 25% and a ca. 25% increase in the amount of sediment trapped in colluvial stores on or below the slopes resulted in a negligible decrease in sediment yield at the catchment outlet.

Although the greatest utility of the sediment budget concept is arguably at the basinwide scale, it can also be applied to individual components of the overall sediment budget. Figure 13 focuses on the sediment budget structure of small tributary basins on the Russian Plain, a region which is heavily impacted by land use activities and soil erosion. In these examples, based on the work of Golosov *et al.* (1992) the proportion of the sediment mobilized within the catchments that reached the basin outlet ranged between zero and 89%. This substantial range emphasizes the considerable diversity in sediment budget structure to be found within even a relatively homogeneous region, as well as potential contrasts in the sensitivity of the basins to changing land use or soil conservation measures. In the former case, changes in land use and erosion rates would be unlikely to exert a significant influence on the downstream sediment yield, whereas in the latter case sediment yields could be expected to be highly sensitive to such changes. Sediment budgets have also been constructed for floodplain systems, to assess their role in trapping and storing sediment moving through the main channel system of a drainage basin. Work on the catchments of the River Ouse (3315 km^2) and the River Wharfe (818 km^2) in Yorkshire, UK, reported by Walling *et al.* (1999a), for example, demonstrated that 30–40% of the sediment delivered to the main channel system was deposited on the adjacent river floodplains, and did not reach the catchment outlets (see Figure 14).

Establishing Sediment Budgets

Although the sediment budget concept affords an invaluable basis for interpreting the linkages between erosion and sediment mobilization, sediment delivery, and conveyance losses and sediment yield, it introduces something of a dilemma in that it is difficult to establish a detailed sediment budget for a drainage basin, and particularly for anything other than a relatively small basin. It is clearly difficult to obtain a detailed assessment of rates of sediment mobilization and deposition over a large heterogeneous

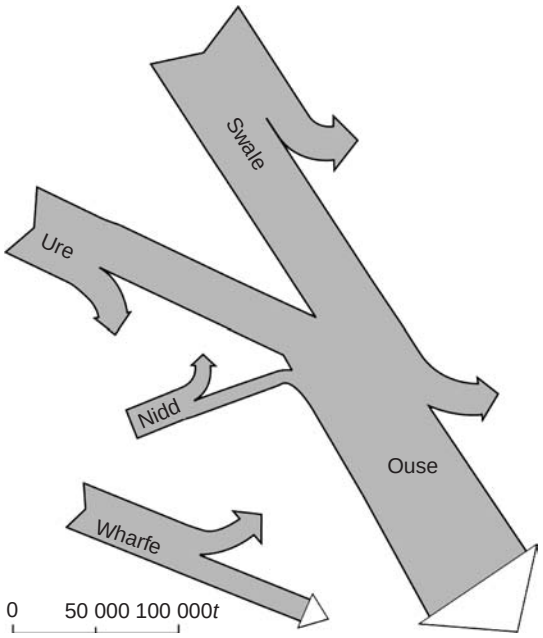


Figure 14 The role of overbank floodplain sedimentation in the sediment budget of the main channel systems of the River Ouse and its major tributaries and the River Wharfe, Yorkshire, UK (based on Walling *et al.*, 1998)

catchment and even studies of the conveyance losses associated with floodplain deposition face significant problems in assessing depositional fluxes along extensive floodplain reaches, where deposition rates can be expected to be highly variable spatially. Furthermore, temporal (e.g. interannual) variability in sediment generation and transfer introduces additional problems in establishing a representative budget. Thus the utility of the concept could be seen more in terms of providing a conceptual framework than as a practical tool. However, recent advances in sediment source fingerprinting afford a means of obtaining a spatially integrated assessment of sediment sources within a drainage basin (e.g. Walling and Woodward, 1995; Walling *et al.*, 1999b) and the use of environmental radionuclides as sediment tracers provides a means of assessing erosion and deposition rates both on the slopes of a catchment and on river floodplains (see Owens *et al.*, 1997; Walling and He, 1997; Walling, 1998, 2002; Zapata, 2002). Walling *et al.* (2001) have attempted to combine these novel approaches with more traditional measurement techniques, to provide an integrated framework for establishing catchment sediment budgets (cf. Walling *et al.*, 2001, 2002), although this approach is primarily suited to smaller drainage basins.

REFERENCES

Abernethy C. (1990) The use of river and reservoir sediment data for the study of regional soil erosion rates and trends. *Paper Presented at the International Symposium on Water Erosion*, Sedimentation and Resource Conservation, Dehradun, India, October, 1990.

Alford D. (1992) Streamflow and sediment transport from mountain watersheds of the Chao Phraya basin, northern Thailand: a reconnaissance study. *Mountain Research and Development*, **12**, 257–268.

ASCE (1975) *Sedimentation Engineering*, ASCE: New York.

Branski J. (1975) Ocena denudacji dorzecza Wisley na podstawie wynikow pomiarow rumowiska unoszonego. *Prace Instytutu Meteorologii I Gospodarki Wodnej*, **6**, 1–58.

Branski J. and Banasik K. (1996) Sediment yields and denudation rates in Poland. In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 133–138.

Bobrovitskaya N.N. (1996) Long-term variations in mean erosion and sediment yield from rivers of the former Soviet Union. In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 407–413.

Church M. and Slaymaker H.O. (1989) Disequilibrium of holocene sediment yield in glaciated British Columbia. *Nature*, **337**, 452–454.

Dai D. and Tan Y. (1996) Soil erosion and sediment yield in the upper Yangtze basin. In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 191–203.

Dedkov A.P. (2004) The relationship between sediment yield and drainage basin area. In *Sediment Transfer through the Fluvial System*, Golosov V., Belyaev V. and Walling D.E. (Eds.), IAHS Publication No. 288, IAHS Press: Wallingford, pp. 197–204.

Dedkov A.P. and Mozzherin V.T. (1984) *Eroziya I Stok Nanosov Na Zemle*, Izdatelstvo Kazanskogo Universiteta.

Dedkov A.P. and Mozzherin V.T. (1992) Erosion and sediment yield in mountain areas of the world. In *Erosion, Debris Flows and Environment in Mountain Regions*, Walling D.E., Davies T.R. and Hasholt B. (Eds.), IAHS Publication No. 209, IAHS Press: Wallingford, pp. 29–36.

Dietrich W.B. and Dunne T. (1978) Sediment budget for a small catchment in mountainous terrain. *Zeitschrift Für Geomorphologie*, **29**, 191–206.

Dunne T. (1979) Sediment yield and land use in tropical catchments. *Journal of Hydrology*, **42**, 281–300.

Douglas I. (1967) Man, vegetation and the sediment yields of rivers. *Nature*, **215**, 925–928.

Farnsworth K.L. and Milliman J.D. (2003) Effects of climatic and anthropogenic change on small mountainous rivers: the Salinas river example. *Global and Planetary Change*, **39**, 53–64.

Fournier F. (1960) *Climat et Erosion*, PUF: Paris.

Gilbert G.K. (1917) Hydraulic mining debris in the Sierra Nevada. *U.S. Geological Survey Professional Paper*, **105**, 154.

Glymph L.M. (1954) *Studies of Sediment Yields from Watersheds*, IAHS Publication No. 36, IAHS Press: pp. 173–191.

Goldberg E.D. (1976) *The Health of the Oceans*, UNESCO: Paris.

Golosov V.N., Ivanova N.N., Litvin L.F. and Sidorchuk A.Yu. (1992) Sediment budgets of river catchments and river channel

aaggradation on the Russian plain. *Geomorphology (Moscow)*, **4**, 62–71.

Gilluly J. (1955) Geologic contrasts between continents and ocean basins. *Geological Society America Special Paper*, **62**, 7–18.

Heusch B. and Milliés-Lacroix A. (1971) Une méthode pour estimer l'écoulement et l'érosion dans un bassin. Application au Maghreb, *Mines et Geologie* (Rabat) No. 33.

Hicks D.M., Hill J. and Shankar U. (1996) Variation of suspended sediment yields around New Zealand: the relative importance of rainfall and geology. In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 149–156.

Holeman J.N. (1968) The sediment yield of major rivers of the world. *Water Resources Research*, **4**, 737–747.

Horowitz A. (1991) *A Primer on Sediment-Trace Element Chemistry*, Lewis: Michigan, pp. 134.

Jackli H. (1957) *Gegenwartsgelogische Bundnerischen Rheingebietes: Ein Beitrag Zur Exogenen Dynamik Alpiner Gebirgslandschaften*, Swiss Geotechnical Commission, Geotechnical Series 36, Kummerley and Frei: Berlin.

Jansen J.H.L. and Painter R.B. (1974) Predicting sediment yield from climate and topography. *Journal of Hydrology*, **21**, 371–380.

Jansson M.B. (1988) A global survey of sediment yield. *Geografiska Annaler Series A*, **70**, 81–98.

Kuenen P.H. (1950) *Marine Geology*, Wiley: Chichester.

Langbein W.B. and Schumm S.A. (1958) Yield of sediment in relation to mean annual precipitation. *Transaction of American Geophysical Union*, **39**, 1076–1084.

Lopatin G.C. (1952) Detritus in the rivers of the USSR. *Zap. Vses. Geogr. Obsch.*, Vol. 14, Geografgiz Publication: Moscow.

Ludwig W. and Probst J.-L. (1996) A global modelling of the climatic, morphological and lithological control of river sediment discharges to the oceans. In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 407–413.

Lvovich M.I., Karasik G.Ya., Btratseva N.L., Medvedeva G.P. and Maleshko A.V. (1991) *Contemporary Intensity of the World Land Intracontinental Erosion*, USSR Academy of Sciences: Moscow.

MacKenzie F.T. and Garrels R.M. (1966) Chemical mass balance between rivers and oceans. *American Journal of Science*, **264**, 507–524.

Maner S.B. (1958) Factors affecting sediment delivery rates in the Red Hills physiographic area. *Transaction of the American Geophysical Union*, **39**, 669–675.

Meade R.H. (1982) Sources, sinks and storage of river sediment in the Atlantic drainage of the United States. *Journal of Geology*, **90**, 235–252.

Meade R.H. and Trimble S.W. (1974) Changes in sediment loads of rivers in the Atlantic drainage of the United States since1900. *Effects of Man on the Interface of the Hydrological Cycle with the Physical Environment*, IAHS Publication No. 113, IAHS Press: Wallingford, pp. 99–104.

Meade R.H., Yuzyk T.R. and Day T.J. (1990) Movement and storage of sediment in rivers of the United States and Canada. *The Geology of North America: Surface Water Hydrology*, Vol. 1, Geological Society of America: Boulder, pp. 255–280.

Meybeck M., Laroche L., Durr H.H. and Syvitski J.P.M. (2003) Global variability of daily total suspended solids and their fluxes in rivers. *Global and Planetary Change*, **39**, 65–93.

McLennan S.M. (1993) Weathering and global denudation. *Journal of Geology*, **101**, 295–303.

Milliman J.H. and Meade R.H. (1983) World-wide delivery of river sediment to the oceans. *Journal of Geology*, **91**, 1–21.

Milliman J.D., Snow J., Jaeger J. and Nittrouer C.A. (1996) Catastrophic discharge of fluvial sediment to the ocean: evidence of Jokulhlaups events in the Alsek sea valley, southeast Alaska (USA). In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 407–413.

Milliman J.H. and Syvitski J.P.M. (1992) Geomorphic/tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *Journal of Geology*, **100**, 325–344.

Owens P.N., Walling D.E., He Q., Shanahan J. and Foster I.D.L. (1997) The use of caesium-137 measurements to establish a sediment budget for the river start catchment, Devon, UK. *Hydrological Sciences Journal*, **42**, 405–423.

Panin A. (2004) Land-ocean sediment transfers in palaeotimes, and implications for present-day natural fluvial fluxes. In *Sediment Transfer through the Fluvial System*, Golosov V., Belyaev V. and Walling D.E. (Eds.), IAHS Publication No. 288, IAHS Press: Wallingford, pp. 197–204.

Pechinov D. (1959) Vodna eroziya I to'rd ottok. *Priroda*, **8**, 49–52.

Pinet P. and Souriau M. (1988) Continental erosion and large-scale relief. *Tectonics*, **7**, 563–582.

Rapp A. (1960) Recent development of mountain slopes in Karkevagge and surroundings, northern Scandinavia. *Geografiska Annaler*, **42A**, 71–200.

Schumm S.A. (1963) The disparity between present rates of erosion and orogeny. *U.S. Geological Survey Professional Paper*, **454H**, 13.

Strakhov N.M. (1967) *Principles of Lithogenesis*, Oliver and Boyd: London.

Summerfield M.A. and Hulton N.J. (1994) Natural controls on fluvial denudation rates in major world drainage basins. *Journal of Geophysical Research*, **99**, 13 871–13 883.

Sundborg A. (1973) Significance of fluvial processes and sedimentation. *Fluvial Processes and Sedimentation*, (Proceedings of Hydrology Symposium University of Alberta, Edmonton), Canada National Research Council: pp. 1–10.

Syvitski J.P.M. (2003) Supply and flux of sediment along hydrological pathways: research for the 21st century. *Global and Planetary Change*, **39**, 1–11.

Trimble S.W. (1976) Sedimentation in Coon Creek Valley, Wisconsin. *Proceedings of the Third Federal Inter-Agency Sedimentation Conference*, US Water Resources Council: Washington, pp. 5–100, 5–112.

Trimble S.W. (1983) A sediment budget for Coon Creek basin in the driftless area, Wisconsin, 1853–1977. *American Journal of Science*, **283**, 454–474.

USSR National Committee for the IHD (1974) *Mirovoi Vodnyi Balans I Vodnye Resursy Zemli*, Gidrometeoizdat: Leningrad.

Vorosmarty C.J., Fekete B.M., Meybeck M. and Lammers R.B. (2000) A simulated topological network representing the global system of rivers at 30-minute spatial resolution (STN-30). *Global Biogeochemical Cycles*, **14**, 599–621.

Vorosmarty C.J., Meybeck M., Fekete B., Sharma K., Green P. and Syvitski J.P.M. (2003) Anthropogenic sediment retention: major global impact from registered river impoundments. *Global and Planetary Change*, **39**, 169–190.

Walling D.E. (1983) The sediment delivery problem. *Journal of Hydrology*, **65**, 209–237.

Walling D.E. (1987) Rainfall, runoff and erosion of the land: a global view. In *Energetics of Physical Environment*, Gregory K.J. (Ed.), Wiley: Chichester, pp. 89–117.

Walling D.E. (1998) Opportunities for using environmental radionuclides in the study of watershed sediment budgets. *Proceedings of the International Symposium on Comprehensive Watershed Management*, Beijing, pp. 1–16, September 1998.

Walling D.E. (2000) Linking land use, erosion and sediment yields in river basins. *Hydrobiologia*, **410**, 223–240.

Walling D.E. (2002) Recent advances in the use of environmental radionuclides in soil erosion investigations. *Nuclear Techniques in Integrated Plant Nutrient, Water and Soil Management*, IAEA: IAEA C&S Papers Series 11/C, Vienna, pp. 279–301.

Walling D.E., Collins A.L., Sichingabula H.M. and Leeks G.J.L. (2001) Integrated assessment of catchment sediment budgets. *Land Degradation and Development*, **12**, 387–415.

Walling D.E. and Fang D. (2003) Recent trends in the suspended sediment loads of the world's rivers. *Global and Planetary Change*, **39**, 111–126.

Walling D.E. and He Q. (1997) Use of fallout caesium-137 in investigations of overbank sedimentation on river floodplains. *Catena*, **29**, 263–282.

Walling D.E. and Kleo A.H.A. (1979) Sediment yields of rivers in areas of low precipitation: a global view. *The Hydrology of Areas of Low Precipitation*, IAHS Publication No. 128, IAHS Press: Wallingford, pp. 3–19.

Walling D.E. and Moorehead P.M. (1989) The particle size characteristics of fluvial suspended sediment. *Hydrobiologia*, **176/177**, 125–149.

Walling D.E., Owens P.N. and Leeks G.J.L. (1998) The role of channel and floodplain storage in the suspended sediment budget of the River Ouse, Yorkshire, UK. *Geomorpholgy*, **22**, 225–242.

Walling D.E., Owens P.N. and Leeks G.J.L. (1999a) Rates of contemporary overbank sedimentation and sediment storage on the floodplains of the main channel systems of the Yorkshire Ouse and the River Tweed, UK. *Hydrological Processes*, **13**, 993–1009.

Walling D.E., Owens P.N. and Leeks G.J.L. (1999b) Fingerprinting suspended sediment sources in the catchment of the River Ouse, Yorkshire, UK. *Hydrological Processes*, **13**, 955–975.

Walling D.E., Russell M.A., Hodgkinson R.A. and Zhang Y. (2002) Establishing sediment budgets for two small lowland agricultural catchments in the UK. *Catena*, **47**, 323–353.

Walling D.E. and Webb B.W. (1981) The reliability of suspended sediment load data. *Erosion and Sediment Transport Measurement*, IAHS Publication No. 133, IAHS Press: Wallingford, pp. 79–88.

Walling D.E. and Webb B.W. (1983) Patterns of sediment yield. In *Background to Palaeohydrology*, Gregory K.J. (Ed.), Wiley: Chichester, pp. 69–100.

Walling D.E. and Webb B.W. (1987) Material transport by the world's rivers: evolving perspectives. In *Water for the Future: Hydrology in Perspective*, Rodda J.C. and Matalas N.C. (Eds.), IAHS Publication No. 164, IAHS Press: Wallingford, pp. 313–329.

Walling D.E. and Webb B.W. (1996) Erosion and sediment yield: a global overview. In *Erosion and Sediment Yield: Global and Regional Perspectives*, Walling D.E. and Webb B.W. (Eds.), IAHS Publication No. 236, IAHS Press: Wallingford, pp. 1–19.

Walling D.E. and Woodward J.C. (1995) Tracing sources of suspended sediment in river basins. A case study of the river Culm, Devon, UK. *Marine and Freshwater Research*, **46**, 327–336.

Wang Z.-Y. (2003) Sedimentation in the yellow river and estuary: management strategies. *From Watershed Slopes to Coastal Areas; Sedimentation Processes at Different Scales*, Pre-print papers of UNESCO/ICCORES Workshop, Venice, December, 2003, Session 3 Paper 1.

Wilson L. (1969) Les relations entre les processus géomorphologi ques et le climat moderne comme méthode de palaeoclimatologie. *Revue de Geographie Physique et de Geologie Dynamique Ser 2*, **11**, 303–314.

Yang Z.-S. (1998) Yellow river's water and sediment discharge decreasing steadily. *The Earth Observing System*, **79**, 589, 592.

Zapata F. (Ed.) (2002) *Handbook for the Assessment of Soil Erosion and Sedimentation Using Environmental Radionuclides*, Kluwer: Dordrecht.

86: Measuring Sediment Loads, Yields, and Source Tracing

GRAHAM LEEKS

Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

The sediment load of rivers is, for both practical purposes and field research, usually divided into the finer suspended load and coarser bed load. A range of instruments can be used, depending on the mode of sediment transport. Measurements of the bed load can be made using fixed and portable traps, samplers and tagging methods, and electronic detection systems. Suspended sediment load is usually measured with both manual and automatic samplers. Continuous measurements are made possible by using turbidity as a surrogate measure of suspended sediment concentrations. An advantage in the use of the manual and automatic samplers is to provide samples which can be subjected to further analysis (e.g. of particle size distributions or chemistry). One of the opportunities this can provide is to identify up-catchment sediment sources, including channel, gully, sheet/rill, and surface erosion under different land uses. Source "fingerprints" are identified from combinations of sediment properties. Multivariate analysis and mixing models can then be used to quantify the relative contributions from individual sources.

INTRODUCTION

The physical dynamics of sediments on catchment surfaces and within river channels, including erosion, transport, and deposition, is one of the fundamental drivers of the movement of materials (physical, chemical, and biological) through river basins. From sediment transport theory, and laboratory flume experiments, it is anticipated that the sediment load in rivers could be divided into three components or modes of transport – the suspended load, saltation load, and bed (or traction) load. In practice, for field research in rivers and streams, it is not possible to measure the saltation load separately from the bed- and suspended load. A fourth mode of transport, floatation load, can also be observed. This is mainly organic debris which may vary in scale from small leaf and invertebrate fragments up to whole trees. A consensus on the methodologies for measurement of the floatation load in rivers has not been developed, and quantification has been rare. Most of the following descriptions are based upon the bed load and suspended load modes of transport. However, it is also worth making a brief reference to another overlapping classification of sediment load, which is sometimes applied. This is a two-fold division based upon a differentiation between sediment source and size characteristics. This is between the fine material from catchment surfaces known as the *wash load* and coarser channel-derived *bed-material load*. The former is transported in suspension, whereas the latter is transported both in suspension and as bed load.

The rest of this article focuses on methods which have been commonly used in river sediment load, monitoring, and source tracing, in terms of manual and automated systems of measurement and analysis, with accompanying references which describe the methodologies in more detail or provide examples of case study applications.

BED-LOAD MEASUREMENT

The bed load or traction load comprises sediment that moves whilst remaining in contact with the channel bed. This is mostly in the size ranges between sands, gravels, cobbles, and boulders. Bed-load monitoring can present greater difficulty, in terms of determining appropriate sampling strategies, relative to suspended load. The movement of the bed load tends to be more intermittent and erratic than the suspended load. Critical flow thresholds for movement

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

are higher than for suspended load and are complicated by particle shape, the range in particle sizes, and packing of the bed material. The relationships with water discharge rates and its distribution through the stream channel cross section are also more complicated than is generally the case for suspended load. In some cases, pulsing effects have been observed (e.g. Bathurst *et al.*, 1985; Reid *et al.*, 1985; Tacconi and Billi, 1987). Hubbell (1987) noted that the bed-load transport rate can vary in a cyclic manner even during steady flow conditions. This was observed whether pronounced bedforms(such as dunes) were present or absent.

Assessment of the volume of sediment which has been transported, its weight, and size distribution are most conveniently carried out after flood events. Therefore, in many cases, traps are used to catch the total bed load generated by a flood. Traps of this sort can consist of excavated chambers lined with concrete or simple flow barriers formed with timber, steel girders, boulders, concrete, or gabion baskets. In each case, the purpose is to slow the flow of water and sediment sufficiently for the coarse sediment to be deposited. By assessing catchment yields over several years and dividing by catchment area, comparisons of annual loss with other catchments can be made. Links to catchment management including land use cover and practices have been inferred. Examples of the use of this type of measurement and data are described in Leeks and Marks (1997) and Newson (1980) in relation to forest and grassland paired catchment experiments in mid Wales, UK (see Figure 3).

More sophisticated versions of bed-load traps have also been developed. Reid *et al.* (1985) used pressure pads to monitor continuously the accumulation of bed load in a trap in Turkey Brook, UK (see also Sear *et al.*, 2000 and Habersack *et al.*, 2001). In the United States, a conveyor belt system was installed beneath a slot excavated in the bed of the East Fork River (Leopold and Emmett, 1976; Leopold, 1992). Bed-load material falling in to the slots was then conveyed to the bank side for analysis. A vortex sampler has also been developed in the United States which takes bed-load material and part of the water discharged from a flow-gauging structure to a work pit where the amounts of bed load can be continuously monitored (Klingeman and Emmett, 1982). A further example of a vortex trap has been installed on Virginio Creek in Tuscany, Italy. Details of the design and the data obtained are provided by Tacconi and Billi (1987).

In addition to permanent structures built into the stream bed, a number of portable samplers are also available. These can be lowered into the channel to capture and filter out the bed-load material over short sampling periods. Portable samplers can be used in detailed investigations of the critical flow thresholds for bed-load movement and of changes in bed-load sediment discharges and associated particle properties (e.g. size distributions), within high flow events. Such instruments include the VUV sampler developed by Novak (1957) and the sampler developed by Helley and Smith (1971) shown in Figure 1.

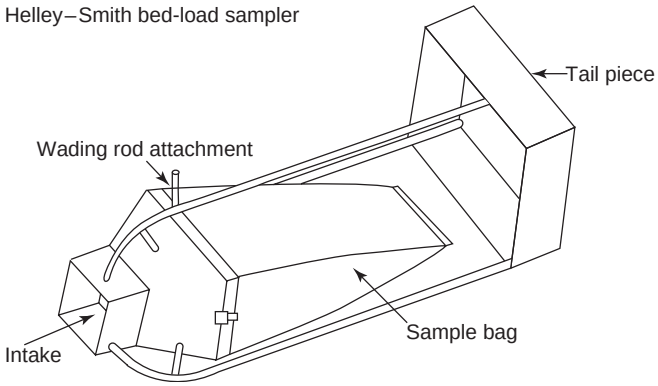


Figure 1 Diagram showing a modified Helley–Smith sampler used for sampling of bed load

Portable samplers can be used at a number of points across the channel to produce sediment discharge values representative of the cross section. Obtaining detailed data through the full cross section can be very time consuming and unless the stage levels remain static, this may be of limited value. Bed-load movement usually occurs in a number of flow threads across the channel. Where these are limited in number, for example, either side of a protruding boulder, sampling may be restricted to a much smaller number of points along the cross section. Detailed cross-sectional information obtained from many verticals can therefore be used to reduce the number of verticals which are repeatedly sampled and the time required to obtain sediment data to represent the cross section.

Another important consideration is that investigations of the efficiency of the Helley–Smith sampler indicate that it over-samples by a factor of 1.5 (Hubbell *et al.*, 1985). Pitlick (1988) highlighted the potential for additional scour as the sampler is placed in the channel because of its high hydraulic efficiency. Sediment catch efficiency may also vary across the flow range. Bunte and Abt (2003) found higher rates of transport during lower flows, and lower transport rates at high flows, using Helley–Smith samplers in four gravel-bedded mountain streams relative to the rates measured using bed-load traps. A variety of modifications have been made to Helley–Smith samplers, which may reduce over-sampling (Kuhnle, 1992) and increase stability on the bed (Figure 4) (Bathurst *et al.*, 1985).

There are also a number of methodologies for indirect measurement of bed-load movement. These include acoustic detectors, which sense the sounds produced as particles collide in the flow and with the river bed, to give indications of particle mobilization thresholds (Ergenzinger and De Jong, 2003) and variations in rates of transport over time (e.g. Bogen and Møen, 2003). It is also possible to

label sediment particles by attaching magnetic materials or coils, radio transmitters, or luminescent coatings to particles and to detect transport using electromagnetic, video, or photographic detection systems. In addition to documenting thresholds of movement and variations in transport through flow events, these methods can also provide useful data on distances of travel. Manipulation of the data collected using these novel methodologies to derive estimates of overall rates of bed-load transport has proved both promising and very difficult. There have been some successes (e.g. Sear *et al.*, 2003), but further work on both the technical and numerical methodologies is required.

SUSPENDED SEDIMENT LOAD MEASUREMENT

The suspended load comprises sediment which is buoyed up by the flow of water, varying in size mainly from clay through to fine sand. Although bed load is predominantly inorganic, suspended sediment can be almost entirely organic (e.g. down stream of a point discharge of effluent). In most instances, however, it is a mixture of both inorganic and organic material. Droppo (2001) emphasizes that suspended sediment is not made up of discrete particles, but is a "composite matrix of water, inorganic and organic particles" forming a "complex microecosystem". The fundamentals of quantifying instantaneous suspended sediment load or flux are through the combination of measurements of both sediment concentration and water discharge. The approaches used to monitor suspended concentration range from instantaneous spot sampling to continuous turbidity recording.

Manual Sampling

The simplest technique used to provide a sample for laboratory analysis of sediment concentration is to take a "gulp" sample (i.e. fill a wide-necked container directly from a stream). However, if sand-sized particles are likely to be in suspension, then, to prevent under- or over-representation, an isokinetic sampler should be used (Edwards and Glysson, 1999). The manual instantaneous point sampling method is most effective where suspended sediment is distributed uniformly through the cross section (as in some upland streams which have a rough bed and high turbulence during high flows), or where it has been possible to establish a relationship between the spot value and concentrations within the overall cross section.

Where there is variation in concentrations within the channel cross section, depth -integrating samplers, which are lowered slowly through a number of vertical profiles, can permit collection of more representative samples. At high flows within a large river cross section, vertical and horizontal velocity gradients will exist. Variations in suspended sediment concentration are directly related to velocity (Ingram *et al.*, 1991). In order to take account of such variations, depth-integrated sampling at several vertical sections across a channel is recommended. Detailed depth-integrated and point sampling of six North American rivers by Horowitz *et al.* (1990), revealed that although there was considerable variation in suspended sediment concentration in the sections studied, this was mostly accounted for by the $>63\,\mu m$ fraction.

The most widely used manual suspended-load samplers are the depth-integrating suite of instruments developed by the US Geological Survey (USGS) (Guy and Norman, 1970). These samplers have the advantage of being highly portable and can be used quickly at a large number of sites during periods of high flow rates, either in handheld versions or by use on cableways. The use of cableways and larger, heavier samplers are particularly useful as flow depth, river width, and velocities make wading less convenient and/or hazardous. The main disadvantages in the use of these and other manual samplers are the requirements for dedicated and very flexible field teams. In addition, a large number of samples are usually generated, which require analysis for concentration using vacuum filtration techniques. However, these techniques also produce physical samples which can then be subjected to other types of analyses (e.g. for geochemical properties). This will be further considered in Section "Investigating suspended sediment properties and sediment source tracing".

Automatic Sampling

Equipment that is designed to automatically extract water from the stream and place it in a sequence of containers can reduce the labor input in the field. Those systems which incorporate stage triggers or are driven by information from sensors through programmed loggers can initiate sampling at preset flow, water quality, or sediment concentration thresholds (e.g. Lewis and Eads, 2001). Series of samples linked to patterns of variation in the river (e.g. rates of rise or fall in river stage) may also be produced. These methods are particularly useful in upland (Leeks, 1992) and urbanized catchments, which may respond rapidly to rainfall (see Figure 2 and Old *et al.*, 2003), or in extended monitoring networks, where it is difficult to deploy other instruments and staff resources to many sites with sufficient speed to catch peaks in flow and sediment (Evans *et al.*, 1997). Although automatic samplers can reduce response times, there remains the need for simple, but time-consuming laboratory analysis to derive sediment concentration data. The efficiency of the autosampler in extracting a representative water sample through its inlet is a further source of uncertainty. Comparisons between methods of obtaining river water samples and their effect on suspended sediment concentration have been carried out (Evans *et al.*, 1997; Tai *et al.*, 1991 and Horowitz *et al.*,

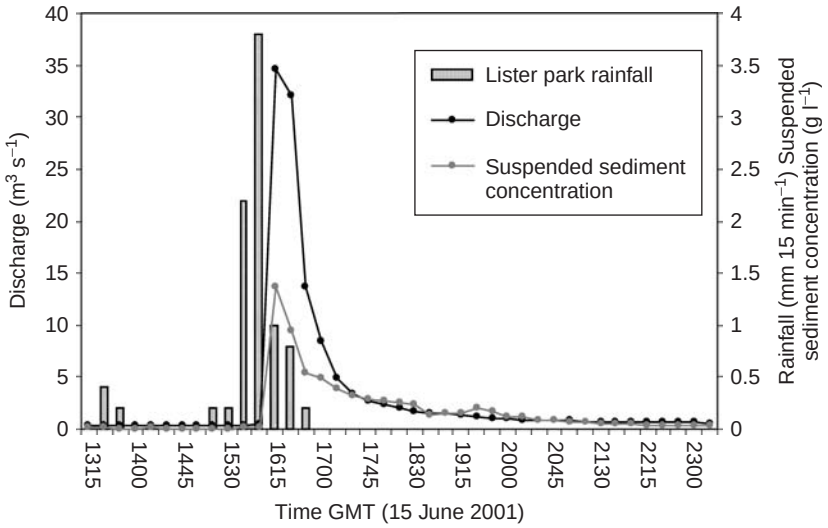


Figure 2 Rainfall, river flow and suspended sediment concentration time series in an highly urbanized catchment, at a downstream river monitoring station, Bradford Beck, UK (after Old *et al.*, 2004). This diagram illustrates the rapid suspended sediment response in the Beck river to a short and intense storm rainfall event

Figure 3 Emptying of a bed-load trap at the CEH upland experimental catchments, Plynlimon, Mid Wales, UK. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 4 Use of a modified Helley–Smith sampler on a mountain stream (Roaring River, Colorado). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

1989). Such comparative work is of considerable value in giving additional confidence in the quality and appropriate use of field sediment load data. For example, Evans *et al.*, 1997 carried out several field tests on the River Swale, UK. In a sequence of 10 successive automatic samples taken every 3 min in parallel with a manual USGS sampler, it was found that sample sediment concentrations were initially more than 35% higher in the automatic sampler (despite a complete purging cycle before sampling). The second samples showed concentrations within 2% of each other, whilst the following eight automatic samples all exhibited concentrations 10–14% lower than those from the USGS sampler.

It is also possible to programme samplers to fill single containers with multiple samples taken at regular intervals (or weighted by flow or stage) to represent sediment concentrations over extended periods, that is, time-integrated samples (see Section "Investigating suspended sediment properties and sediment source tracing").

Turbidity Monitoring

An alternative to the collection of large numbers of bulk samples is the use of turbidity monitors (Eden, 1965 and Thorn and Burt, 1975) to provide instantaneous point

measurements. Experience from a range of studies (e.g. Gippel, 1995; Foster *et al.*, 1992, Glysson and Gray, 2002; Gray *et al.*, 2002) using turbidity as a surrogate for suspended sediment concentration, and monitoring this property along with water discharge at high frequencies, often provides the most practical means of estimating suspended sediment fluxes in rivers. The basic sensor technologies involve the beaming of visible or infrared light along single or dual light paths through water sediment mixtures directly towards photoelectric sensors (absorptiometric methods) or by measuring the scatter of light away from a light path (nephelometric methods).

Instruments based upon nephelometric or absorptiometric principles can be used to provide semicontinuous in-river turbidity measurements. These systems have also been used as laboratory-based instruments to speed up analysis (Ward and Chikwanha, 1980). They do not, however, provide suspended load concentration values which are representative of the whole water flow cross section and it must be emphasized that the turbidity of waters is not only a function of suspended sediment concentration. For this reason, cross-calibration with automated samplers and isokinetic USGS depth-integrating samplers is recommended (Figure 5). Wass and Leeks (1999) found that point samples, represented by turbidity sensors, consistently underestimated the load by between 2% and 12% for two rivers in northern England (the Ure and Aire respectively), relative to the values obtained using the depth-integrated method as described by Guy and Norman (1970).

Scatter in the relationship between turbidity and suspended sediment concentration is also partly a function of measurement error. Sensors are sensitive to changes in temperature, aging of light sources, fouling by debris and algae, variations in incident light, and vandalism. A more detailed account of some these problems can be found in Wass *et al.* (1997), Gippel (1995) and Old *et al.* (2002). The errors introduced by such problems are difficult to quantify, but calibration checks are carried out to monitor any instrument drift. Instrument drift and total failures are not uncommon, so regular calibration against a standard calibration in Formazin Turbidity Units (FTU) is recommended. In the past, Formazin and Fullers Earth have been used. However, there are possible Health and Safety issues involved in use of Formazin. The US Environmental Protection Agency has now endorsed an alternative method using a suspension of polymer beads.

Turbidity has a linear relationship with suspended sediment concentration only when the particle properties are constant (Foster *et al.*, 1992). For example, if the particle size distribution or mineralogy of the suspension changes, then so does the turbidity. Across natural catchments, variations in geology, soils, and land cover (and effluent inputs) mean that the sediment delivered to a river is usually heterogeneous, and represents a mixture of materials derived from different sources (Walling *et al.*, 1999). By comparing the suspended sediment concentrations of many samples with the measured turbidity, an assessment of the impact of this heterogeneity on the accuracy of load calculations can be made. Schoellhamer and Wright (2003) maintain that the suitability of optical sensors should be evaluated on a site-specific basis. When used for continuous stream monitoring, there is a need for regular cleaning of sensor heads and calibration under a variety of flow conditions. Some turbidity sensors incorporate self cleansing cycles (e.g. mechanical wipers which clean sensor lenses). It may also be necessary to establish laboratory analysis protocols to exclude (or include) larger particles (e.g. those above 6 mm diameter), which can have significant impact upon calculations of concentrations.

Figure 5 Use of a handheld depth-integrating USGS suspended-load sampler at a downstream flow gauging station during a high flow event (following a test release from Llyn Clywedog Reservoir, Mid Wales, UK). Please note: when rivers are hazardous and/or there is significant cross-sectional variation in sediment concentrations, cable systems or safe bridging points are used. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

As mentioned earlier, maintaining a good turbidity record can be difficult because of a range of factors, including sensor blockage and drift. Where instantaneous turbidity peaks are known to be spurious, manual manipulation and

interpolation are needed to correct or infill the data. In the event of larger sections of data being lost, a rating curve can be developed between sampled suspended sediment concentrations and flow for adjacent storm events. This relationship is then used to develop a synthetic suspended sediment record for the missing sections as described in Wass and Leeks (1999). However in all cases, it is recommended that the original data (including raw outputs from turbidity monitors) should be archived, in parallel with modified and derived data, in the event that subsequent revisions are required.

More advanced technologies have also been used in more limited applications, such as Acoustic Doppler Velocimeters (ADV) as described by Nicholas (2003). Laser technologies have also been used (Agrawal and Pottsmith, 2001; Melis and Topping, 2003), whilst at very high concentrations in which optical sensors do not work well, nuclear techniques have been applied (e.g. at concentrations above $5000\,\mathrm{mg\,l^{-1}}$).

Estimating Suspended Sediment Yields

To calculate the longer-term flux, needed to establish the sediment yield from a catchment, the derived record of suspended sediment concentration is combined with the flow and summed over the desired period of record as described by the formula below:

$$L = \sum_{i=1}^{n} (C_i Q_i) \tag{1}$$

Where L is the suspended sediment load [kg], C_i is the suspended sediment concentration [$\mathrm{g\,l^{-1}}$], and Q is the flow [$\mathrm{m^3\,s^{-1}}$] summed over the period of record, n.

Estimates of the longer-term (e.g. annual) amount of suspended load can be derived using equation (1) from the continuous records provided by turbidity sensors (e.g. Truhlar, 1978; Brabben, 1981; Grobler and Weaver, 1981; Wass and Leeks, 1999 and Old *et al.*, 2002). A second method is to use a rating curve technique relating sediment concentrations, derived from spot or depth-integrated sampling, to water discharge. The rating curve, derived by regression analysis, is combined with the flow series or a flow duration curve to calculate the quantity of suspended sediment transported during given periods of time. A detailed review of the errors inherent in the rating technique has been presented by Walling (1977). Errors in estimates can be reduced by partitioning data in relation to other factors such as season and hysteretic effects. Furthermore, if data are log-transformed to define rating relationships, a bias correction methodology (as described by Ferguson, 1986) is desirable when back-transforming data.

Both methods of estimating suspended sediment load are generally less successful in upland streams than in downstream reaches for several reasons. Seasonal changes in the composition of suspended sediment are more marked in some upland channels, making estimation of suspended sediment concentration from turbidity measurement less reliable. Concentrations of suspended matter tend to be lowest in rural upland streams and therefore more sensitive sensors are required. In the case of the rating curve technique, correlation between water discharge rate and sediment concentration is usually much poorer in smaller headwater subcatchments than in larger catchments.

It is possible to combine suspended and bed-load sediment methods to estimate total sediment yield or to use other techniques which directly measure total loads. In small catchments, devices such as the Coshocton Wheel samplers, which are designed to obtain a composite proportional samples of runoff, have been used to measure runoff and sediment outputs. This sampler collects a composited representative sample of runoff from each storm event at sites where flow is measured using weir structures. The sample can be used to determine average concentrations and total transport of the water and suspended sediment in runoff events. Proportionality is maintained by a revolving vane and movable outfall that collect a consistent fraction of all runoff above a threshold or base flow level (Bonta, 2002; Bonta and Pierson, 2003). Surveys of alluvial fans and deltaic deposits at the mouths of rivers and repeated topographic surveys and dating of sediments in lakes and reservoirs (e.g. Foster and Lees, 1999) have also been used to estimate total sediment yields from upstream catchments.

INVESTIGATING SUSPENDED SEDIMENT PROPERTIES AND SEDIMENT SOURCE TRACING

As indicated previously, the main focus of measuring suspended sediment transport in rivers is commonly establishing the instantaneous load or the longer-term sediment yield, by obtaining measurements of sediment concentration that can be combined with water discharge data to estimate the sediment flux. In some instances, however, information on the physical and chemical properties of the sediment, rather than the suspended sediment concentration, may be required. These situations could include investigations of sediment-associated nutrient and contaminant fluxes (e.g. Walling *et al.*, 2001) and sediment source tracing investigations (e.g. Walling and Woodward, 1995). In this case, there is a need to collect a representative sample of the suspended sediment transported by a river for further physical or chemical analysis. In some situations, the small amounts of sediment that can be recovered from the small samples collected by a depth-integrating suspended sediment sampler or an automatic sampler may meet this requirement, but more commonly a larger sample will be required. In upland catchments, for example, where sediment concentrations are relatively low, a 500-ml sample may only yield less

than 100 mg of sediment. If standard sampling equipment is used, this may need to be constructed of special materials to avoid contamination of the sediment sampler prior to analysis. Where larger samples are required, pump samplers may be used to collect bulk water samples in special containers and the sediment can be recovered from these samples by centrifugation or settling. An alternative approach involves the use of a portable, continuous-flow centrifuge, with water being pumped directly from the river into the centrifuge which recovers the fine sediment (e.g. Owens *et al.*, 2001; Ongley and Thomas, 1989). In both cases, however, the sample will effectively be an instantaneous sample of the sediment flux and numerous samples may be required to document the temporal variability of the suspended sediment properties and its controls.

The time-integrating trap sampler described by Phillips *et al.* (2000) affords a means of collecting a more temporally representative sample of suspended sediment. This device comprises a cylindrical chamber, with a small diameter inlet and outlet tube, which is orientated into the flow and deployed in the river for an extended period (e.g. several weeks). A representative sample of the suspended sediment transported by the river or stream during the period of deployment collects within the chamber and this can be recovered by emptying the chamber. In small streams the trap sampler can be fixed above the bed using metal supports driven into the bed, but in larger rivers the chamber may need to be suspended from a fixed structure within the channel or at the edge of the channel. Although this type of time-integrating trap sampler affords an effective means of overcoming problems associated with the temporal representativeness of instantaneous samples, it is clearly important to recognize that a sample withdrawn from a single point within the channel may not be representative of the overall suspended sediment flux.

Measurements of the particle size distribution will frequently be made on sediment samples collected for subsequent analysis, and a wide range of devices and methods have been used for such analysis. It is, however, important to recognize that most of such analyses provide information on the *absolute* grain-size composition of the sample, which reflects the size distribution of the discrete mineral grains after removal of the organic component of the sample and chemical or ultrasonic dispersion. It is now well known that the suspended sediment particles transported by a river will commonly comprise a substantial proportion of composite particles (i.e. flocs or aggregates), which are larger and differ in both density and shape from the discrete particles of which they are formed (*see* **Chapter 83, Suspended Sediment Transport – Flocculation and Particle Characteristics, Volume 2**). The grain-size distribution of the actual particles transported by the river may therefore be significantly different from the absolute grain-size composition of the constituent mineral particles, and this is commonly referred to as the *effective* or *in situ* size distribution. Since the latter size distribution will change immediately after the sample is withdrawn from the river, it is best measured *in situ*. Laser probes provide a means of undertaking such measurements (e.g. Phillips and Walling, 1995a; Melis and Topping, 2003). However, results presented by Phillips and Walling (1995b) suggest that a meaningful indication of the *in situ* grain-size distribution can be obtained from samples returned to the laboratory for measurement using a laser granulometer, without further processing, provided care is taken in agitating and resuspending the sediment.

Another context in which suspended sediment sampling is undertaken is for sediment source tracing or "fingerprinting". Although the traditional focus of sediment measurement programmes has been in quantifying the *amount* of sediment transported by a river, there is an increasing need to obtain information on the *source* of that sediment. Information on source could relate to the relative contributions of different parts of a catchment to the suspended sediment flux at the catchment outlet, but more commonly the emphasis will be placed on quantifying the relative importance of different source types within a catchment. Source types could include channel erosion, gully erosion, and sheet/rill, or surface erosion of the slopes of the catchment under different land use (e.g. cultivated land or pasture / rangeland). Information on source type is of particular importance for the design of effective sediment control or management strategies, since it is necessary to target the control measures at those sources which are of most importance for producing sediment. Thus, for example, it could prove of little value to implement a programme for controlling channel erosion in a catchment if the main sediment source was erosion of the catchment surface.

In many environments where specific sediment yields are relatively low, it may prove difficult to identify the key sources of sediment within a catchment by field observation. Furthermore, it is important to recognize the key importance of slope-channel connectivity in influencing the importance of a source to the sediment flux at the catchment outlet. Thus, for example, areas of severely eroding slopes within a catchment could be of limited importance as a sediment source if they were not well connected to the channel network and most of the mobilized sediment was deposited at the bottom of the slope or in ephemeral valleys, prior to reaching the permanent channel system. A means of establishing the source of the sediment actually transported by the river could therefore prove of great value in many situations. Source "fingerprinting" procedures provide such a means. In essence, this approach is founded on a comparison of the "fingerprints" of transported sediment with those of potential sources and involves, firstly, the selection of one or more physical or chemical properties which clearly differentiate potential source materials and, secondly, the

comparison of measurements of the same property obtained from suspended sediment with equivalent values for potential sources, in order to identify the likely source of that sediment. By using composite fingerprints involving a variety of fingerprint properties, multivariate statistical techniques to test source discrimination, and quantitative mixing (or unmixing) models, the fingerprinting approach can provide reliable quantitative information on the sources of the sediment transported by a river or stream, and their relative contributions.

Further details of the source fingerprinting approach can be found in reports of its application provided by Walling and Woodward (1995), Walling *et al.* (1999), and Russell *et al.* (2001) for both small catchments and larger river basins in the United Kingdom, by Collins *et al.* (2001) for a catchment in Zambia, as well as by others. Following the outline of the methodology provided previously, the first stage of a source fingerprinting investigation commonly involves collecting representative samples of the various potential sources in a catchment or river basin (i.e. source types or spatial units), analyzing those samples for a range of fingerprint properties, and selecting a set of properties which together clearly discriminate the potential sources. This set of properties is used as the composite fingerprint in the subsequent fingerprinting exercise. Statistical procedures, such as the Kruskal–Wallis test and multiple discriminant analysis, provide a means of confirming which properties are best able to assist in discriminating between the sources and selecting the optimum set for use as the composite fingerprint. Once the composite fingerprint has been identified, it is necessary to collect samples of the suspended sediment transported by the river and to analyze these for the properties comprising the composite fingerprint. To facilitate direct comparison between the properties of the potential sources and the suspended sediment samples, only the $<63\,\mu m$ fraction of both groups of samples is commonly analyzed.

Once the information on the fingerprint property values has been obtained for both the suspended sediment sample and the potential sources, a linear mixing model can be used to establish the relative contributions of the individual sources to the sediment sample. Further corrections for contrasts in particle size composition and organic matter content between the sediment and the source materials can be introduced at this stage. It is important to recognize that the results obtained from the mixing model will relate solely to the particular sediment sample used in the mixing model and it will generally be necessary to analyze a substantial number of suspended sediment samples for use with the mixing model, in order to obtain representative information on the relative importance of the different potential sources. Their contributions can be expected to vary both within and between individual storm events and seasonally. A load-weighted average is frequently used to estimate the overall relative contributions of the individual sources. Where a time-integrated sampler, such as that described by Phillips *et al.* (2000) and mentioned previously, is used to collect the suspended sediment samples, these samples should prove more representative of the overall sediment flux.

A wide range of sediment properties have been successfully used as fingerprint properties, although fallout radionuclides, including ^{137}Cs and excess ^{210}Pb, have proved particularly useful for discriminating between surface and channel sources and between cultivated and non-cultivated surface sources. However, until more generic guidelines become available, it will rarely be possible to define the set of fingerprint properties to be used without empirical tests, and the use of statistical procedures to identify properties that afford good discrimination between potential sources and to select the optimum set of properties as the composite fingerprint currently remains a key component of the methodology (Collins and Walling, 2002).

Acknowledgments

Thanks are extended to Gareth Old for commenting on the text and Prof. D.E. Walling for providing many useful suggestions, particularly with regard to the Section "Investigating suspended sediment properties and sediment source tracing".

REFERENCES

Agrawal Y. and Pottsmith C. (2001) Laser sensors for monitoring sediments: capabilities and limitations. *Federal Interagency Sedimentation, Proceedings of the Seventh Conference*, Reno, III–144 to III–151.

Bathurst J.C., Leeks G.J.L. and Newson M.D. (1985) Field measurements for hydraulic and geomorphological studies of sediment transport- the special problems of mountain streams. *Proceedings of the International Symposium on Measuring Techniques in Hydraulic Research*, IAHR Section on Hydraulics Instrumentation: Delft, pp. 137–151, 22–24 April.

Bogen J. and Møen K. (2003) Bed load measurements with a new passive acoustic sensor. In *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: pp. 181–192.

Bonta J.V. (2002) Modification and performance of the Coshocton wheel with the modified drop-box weir. *Journal of Soil and Water Conservation*, **57**(6), 364–373.

Bonta J.V. and Pierson F.B. (2003) Design, measurement, and sampling with drop-box weirs. *Applied Engineering in Agriculture*, **19**(6), 689–700.

Brabben T.E. (1981) Use of turbidity monitors to assess sediment yields in East Java, Indonesia. *Erosion and Sediment Transport Measurement*, International Association of Hydrological Sciences: Publication No. 133, pp. 105–113.

Bunte K. and Abt S.R. (2003) Sampler size and sampling time affect bed load transport rates and particle sizes measured

with bedload traps in gravel-bed streams. In *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: pp. 126–133.

Collins A.L. and Walling D.E. (2002) Selecting fingerprint properties for discriminating potential suspended sediment sources in river basins. *Journal of Hydrology*, **261**, 218–244.

Collins A.L., Walling D.E., Sichingabula H. and Leeks G.J.L. (2001) Using ^{137}Cs measurements to quantify soil erosion and sediment redistribution for areas under different land use in the Upper Kaleya River basin, Southern Zambia. *Geoderma*, **104**(3–4), 193–214.

Droppo I. (2001) Rethinking what constitutes suspended sediment. *Hydrological Processes*, **15**, 1551–1564.

Eden G.E. (1965) The measurement of turbidity in water. *Proceedings of the Society for Water Treatment and Examination*, **14**, 27–41.

Edwards T.K. and Glysson G.D. (1999) *Field Methods for Measurement of Fluvial Sediment*, US Geological Survey: Book 3, Chapter C2, p. 89.

Ergenzinger P. and De Jong C. (2003) Perspectives on bed-load measurement. *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: 113–125.

Evans J.G., Wass P.D. and Hodgson P. (1997) Integrated continuous water quality monitoring for the LOIS rivers programme. *Science of the Total Environment*, **194/195**, 111–118.

Ferguson R.I. (1986) River loads underestimated by rating curves. *Water Resources Research*, **22**, 74–76.

Foster I.D.L. and Lees J.A. (1999) Changing headwater suspended sediment yields in LOIS catchments over the last century: a paleolimnological approach. *Hydrological Processes*, **13**, 1137–1153.

Foster I.D.L., Millington R. and Grew R.G. (1992) The impact of particle size controls on stream turbidity measurement; some implications for suspended sediment yield estimation. *Erosion and Sediment Transport Monitoring Programmes in river Basins*, Vol. 210, IAHS Publisher: pp. 51–62.

Gippel C.J. (1995) Potential of turbidity monitoring for measuring the transport of suspended solids in streams. *Hydrological Processes*, **9**, 83–97.

Glysson G.D. and Gray J.R. (2002) *Turbidity and Other Sediment Surrogates Workshop*, April 30–May 2, 2002, Reno, Nevada, `http://water.usgs.gov/osw/techniques/turbidity.html`.

Gray J.R., Glysson, G.D. and Mueller, D.S. (2002) Comparability and accuracy of fluvial-sediment data – a view from the U.S. geological survey. *Proceedings of the ASCE Specialty Conference, Hydraulic Measurements and Experimental Methods Conference*, Estes Park, Colorado, July 28-August 1, 2002, `http://Water.Usgs.Gov/Osw/Techniques/Asce.Pdf`.

Grobler D.C. and Weaver A.V.B. (1981) Continuous measurement of suspended sediment in rivers by means of a double beam turbidity meter. *Erosion and Sediment Transport and Measurement*, International Association of Hydrological Sciences: Publication No. 133, pp. 97–103.

Guy H.P. and Norman V.W. (1970) Field methods for measurement of fluvial sediment. *U.S. Geological Survey Techniques of Water Resources Investigations*, Book 3, Chap. C2, U.S. Geological Survey, p. 59.

Habersack H., Nachtenebel P.N. and Laronne J.B. (2001) The continuous measurement of bed load flux in a large alpine gravel bed river. *Journal of Hydraulic Research*, **39**, 125–133.

Helley E.J. and Smith W. (1971) *Development and Calibration of a Pressure-difference Bedload Sampler*, *United States Geological Survey Open-File Report*: Menlo Park, p. 18.

Horowitz A.J., Elrick K.A. and Hooper R.C. (1989) Comparison of instrumental dewatering methods for the separation and concentration of suspended sediment for subsequent trace element analysis. *Hydrological Processes*, **3**, 163–184.

Horowitz A.J., Rinella A.R., Lamothe P., Miller T.L., Edwards T.K., Roche R.L. and Rickert D.A. (1990) Variations in suspended sediment and associated trace element concentrations in selected riverine cross sections. *Environmental Science and Technology*, **24**, 1313–1320.

Hubbell D.W. (1987) Bed load sampling and analysis. In *Sediment Transport in Gravel Bed Rivers*, Thorne C.R., Bathurst J.C. and Hey R.D. (Eds.), John Wiley and Sons: Chichester, pp. 89–119.

Hubbell D.W., Stevens H.H., Skinner J.V. and Beverage J.P. (1985) New approach to calibrating bedload samplers. *Proceedings of the America of Society Civil Engineering, Journal of Hydraulics Engineering*, **111**(4), 677–694.

Ingram J.J., Abt R.A. and Richardson E.V. (1991) Sediment discharge computation using point sampled suspended sediment data. *Journal of Hydraulic Engineering-ASCE*, **117**, 758–773.

Klingeman P.C. and Emmett W.W. (1982) Gravel bedload transport processes. In *Gravel-bed Rivers. Fluvial Processes, Engineering and Management*, Hey R.D., Bathurst J.C. and Thorne C.R. (Eds.), John Wiley and Sons: Chichester, pp. 141–179.

Kuhnle R.A. (1992) Fractional transport rates of bedload on Goodwin Creek. In *Dynamics of Gravel Bed Rivers*, Billi P., Hey R.D., Thorne C.R. and Tacconi P. (Eds.), John Wiley, Chichester, pp. 641–660, pp. 141–155.

Leeks G.J.L. (1992) Impact of plantation forestry on sediment transport. In *Dynamics of Gravel bed Rivers*, Billi P., Hey R.D., Thorne C.R. and Tacconi P. (Eds.), John Wiley, Chichester, pp. 641–660.

Leeks G.J.L. and Marks S.D. (1997) Dynamics of river sediments in forested headwater streams: Plynlimon. *Hydrology and Earth Systems*, **1**, 483–498.

Leopold L.B. (1992) Sediment size that determines channel morphology. In *Dynamics of Gravel bed Rivers*, Billi P., Hey R.D., Thorne C.R. and Tacconi P. (Eds.), John Wiley: Chichester, pp. 297–311.

Leopold L.B. and Emmett W.W. (1976) Bedload measurements, East Fork River, Wyoming. *Proceedings of the National Academy of Sciences of the United States of America*, **73**, 1000–1004.

Lewis J. and Eads R.E. (2001) Turbidity threshold sampling for suspended sediment load estimation. *Federal Interagency Sedimentation, Proceedings of the Seventh Conference*, Reno, III–110 to III–117.

Melis T.S. and Topping D.J. (2003) Testing laser based sensors for continuous in situ monitoring of suspended sediment in the Colorado River, Arizona. In *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological*

Advances, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: pp. 21–27.

Newson M.D. (1980) The erosion of drainage ditches and its effects on bed-load yields in Mid Wales. *Earth Surface Processes and Landforms*, **5**, 275–290.

Nicholas A.P. (2003) Modelling and monitor flow and suspended sediment transport in lowland river floodplain environments. In *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: pp. 45–54.

Novak P. (1957) 'Bed-load' meters; development of a new type and determination of their efficiency with the aid of scale models. *Transactions of the 7th General Meeting of the International Association for Hydraulic Research*, IAHR: Delft.

Old G.H., Leeks G.J.L., Packman J.C.P., Smith B.P.G., Goodwin T., Guymer I., Hewitt N., Holmes M., Shepherd W. and Young A. (2002) *Fine Sediment Dynamics in Urban Systems*, R and D Technical Report P2-153/TR, Environment Agency (North East Region), p. 116.

Old G.H., Leeks G.J.L., Packman J.C., Smith B.P.G., Lewis S., Hewitt E.J., Holmes M. and Young A. (2003) The impact of a convectional summer rainfall event on river flow and fine sediment transport in a highly urbanised catchment: Bradford, West Yorkshire. *Science of the Total Environment*, **314**, 475–494.

Old G.H., Leeks G.J.L., Packman J.C.P., Stokes N., Williams N.D., Smith B.P.G., Hewitt E.J. and Lewis S. (2004) Dynamics of sediment-associated metals in a highly urbanised catchment: Bradford, West Yorkshire. *Water and Environmental Management Journal*, **18**, 11–16.

Ongley E.D. and Thomas R.L. (1989) Dewatering suspended solids by continuous-flow centrifugation: practical considerations. *Hydrological Processes*, **3**, 255–260.

Owens P.N., Walling D.E., Carton J., Meharg A.A., Wright J. and Leeks G.J.L. (2001) Downstream changes in the transport and storage of sediment-associated contaminants (P, Cr and PCBs) in agricultural and industrialised drainage basins. *Science of the Total Environment*, **266**, 177–186.

Phillips J.M., Russell M.A. and Walling D.E. (2000) Time-integrated sampling of fluvial suspended sediment: a simple methodology for small catchments. *Hydrological Processes*, **14**, 2589–2602.

Phillips J.M. and Walling D.E. (1995a) Measurement *in situ* of the effective particle size characteristics of fluvial suspended sediment by means of a field-portable laser backscatter probe: some preliminary results. *Marine and Freshwater Research*, **46**, 349–357.

Phillips J.M. and Walling D.E. (1995b) Assessment of the effects of sample collection, storage and resuspension on the representativeness of measurements of the effective particle size distribution of fluvial suspended sediment. *Water Research*, **29**, 2498–2508.

Pitlick J. (1988) Variability of bed load measurement. *Water Resources Research*, **24**, 173–177.

Reid I., Frostick L.E. and Layman J.T. (1985) The incidence and nature of bedload transport during flood flows in coarse-grained alluvial channels. *Earth Surface Processes and Landforms*, **10**, 33–44.

Russell M.A., Walling D.E. and Hodkinson R. (2001) Suspended sediment sources in two small lowland agricultural catchments in the UK. *Journal of Hydrology*, **252**, 1–24.

Schoellhamer D.H. and Wright S.A. (2003) Continuous measurement of suspended-sediment discharge in rivers by use of optical backscatterance sensors. In *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: pp. 28–36.

Sear D.A., Damon W., Booker D.J. and Anderson D. (2000) A load cell based continuous recording bed load trap. *Earth Surface Processes and Landforms*, **25**, 659–673.

Sear D.A., Lee M.W.E., Carling P.A., Oakey R.J. and Collins M.B. (2003) An assessment of the accuracy of the Spatial Integration Method (SIM) for estimating coarse bed load transport in gravel-bedded streams using tracers. In *Erosion and Sediment Transport Measurement in Rivers: Technological and Methodological Advances*, Bogen J., Fergus T. and Walling D.E. (Eds.), Vol. 283, IAHS: pp. 164–171.

Tacconi P. and Billi P. (1987) Bed load transport measurements by the Vortex tube trap on Virginio Creek, Italy. In *Sediment Transport in Gravel Bed Rivers*, Thorne C.R., Bathurst J.C. and Hey R.D. (Eds.), John Wiley and Sons: Chichester, pp. 583–616.

Tai D.Y., Jennings M.E., White K.D. and Garcia L.A. (1991) *Evaluation of a Modified Automatic Sampler for the Collection of Water Samples for Analysis of Trace Organic Compounds or Suspended Sediment*, U. S. Geological Survey open file report 91–469, U. S. Geological Survey.

Thorn M.F.C. and Burt T.N. (1975) *Transport of suspended sediment in the tidal River Crouch*, Report No. INT 148, Hydraulics Research Station, Wallingford.

Truhlar J.F. (1978) Determining suspended sediment loads from turbidity records. *Hydrological Sciences Bulletin*, **23**, 409–417.

Walling D.E. (1977) Assessing the accuracy of suspended sediment rating curves for a small basin. *Water Resources Research*, **13**, 531–538.

Walling D.E., Owens P.N. and Leeks G.J.L. (1999) Fingerprinting suspended sediment sources in the catchment of the River Ouse, Yorkshire, UK. *Hydrological Processes*, **13**, 955–975.

Walling D.E., Russell M.A. and Webb B.W. (2001) Controls on the nutrient content of suspended sediment transported by British rivers. *The Science of the Total Environment*, **266**, 113–123.

Walling D.E. and Woodward J.C. (1995) Tracing sources of suspended sediment in river basins. a case study of the River Culm, Devon, UK. *Marine and Freshwater Research*, **46**, 327–336.

Ward P.R.B. and Chikwanha R. (1980) Laboratory measurement of sediment by turbidity. *Journal of the Hydraulic Division of the American Society of Civil Engineers*, **HY6**, 1041–1054.

Wass P.D. and Leeks G.J.L. (1999) Suspended sediment fluxes in the Humber catchment, UK. *Hydrological Processes*, **13**, 935–953.

Wass P.D., Marks S.D., Finch J.W., Leeks G.J.L. and Ingram J.K. (1997) Monitoring and preliminary interpretation of in-river turbidity and remote sensed imagery for suspended sediment transport studies in the Humber catchment. *Science of the Total Environment*, **194/195**, 263–283.

87: Sediment Yield Prediction and Modeling

SUE WHITE

Institute of Water and Environment, Cranfield University, Silsoe, Bedfordshire, UK

The estimation of sediment yield – the amount of sediment reaching or passing a point of interest in a given period of time – is an inherently complex and uncertain science as it is the integrated outcome of erosion, sediment transport, and deposition processes through an often complex landscape. Thus, sediment yield can vary greatly from year to year with variation in climate, land use, and land management. A range of estimation techniques have been developed ranging from simple empirical relationships between sediment yield and river flow or basin characteristics, through methods to scale soil erosion estimates or measurements using delivery ratios, to physically based distributed models. Descriptions of the various approaches, together with advantages and shortcomings of these methods, are given. For prediction of future sediment yields, the uncertainties are even greater as climate and land-use pressures are not easily defined for hypothetical situations. Here, recommendations are made to incorporate uncertainty into the prediction and decision-making process.

INTRODUCTION

Sediment yield can be defined as the amount of sediment reaching or passing a point of interest in a given period of time. Sediment yield estimates are normally given as tons per year or kilograms per year. Specific sediment yield, as with specific flow, brings the catchment area into the equation and here yields are defined as tons per square kilometer per year or kilograms per hectare per year. By definition, in estimating sediment yield, we are taking a snapshot of a part of the sediment cycle. With any modeling exercise, we are trying to take a look at a particular point of interest to see how the variability in the sediment cycle upstream of this point presents itself in terms of sediment delivery over time. This also differentiates between sediment transport modeling, where we are looking solely at in-stream processes, and sediment yield modeling, where we are attempting to make an integrated assessment of both supply and transport. Traditionally, sediment yield estimates were required as part of the feasibility study for a proposed dam or barrage. In these cases, an average annual sediment yield was often considered to be sufficient. Even this apparently simple output has proved to be difficult to achieve, and a number of well-publicized cases have seen reservoirs losing capacity at a rate far exceeding that estimated before construction of the dam (Table 1). Today, sediment yield estimates are also required as part of our assessment of sediment budgets and delivery of sediments and contaminants to the estuarine or ocean system.

A large part of the failure to achieve reasonable estimates of average annual sediment yield lies in the practice of extrapolating relationships derived from field data with no consideration of how appropriate they might be for future conditions. The impact of changing land use on sediment yield has been extensively documented (e.g. Lu and Higgitt, 1998; Harden, 1993; García-Ruiz *et al.*, 1995). Similarly, there are frequent reported changes in sediment deposition or yield related to climate and flow variability (e.g. Valero-Garcés *et al.*, 1999; Bathurst *et al.*, 1996). Yet other papers detail both influences (e.g. Xu, 2003; Asselman *et al.*, 2003; Lang *et al.*, 2003). Reservoir catchments in developing countries are particularly prone to act as concentrators of human activity postconstruction. Such activity will mean dramatic changes to land use and, potentially, to water as well as sediment yields. Other factors have played their part in the frequent underestimation of sediment yield. Another major contributor to error is the use of sediment concentration – water discharge relationships based on sparse data, often monitored at low river flow rates. We will see later on that extreme nonlinearity and nonstationarity in these relationships make their use a high-risk strategy.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Predicted and actual sediment yields for reservoirs

Reservoir	Sedimentation rate $m^3 \times 10^6$ $year^{-1}$		Source
	Preconstruction estimate	Postconstruction actual	
Karangkates, East Java	0.33	2.04	Brabben (1982)
Wlingi, East Java	0.38	1.42	Fish (1983)
Bhakra, Punjab, India	28.4	41.6	Patnaik (1975)
Panchet, Bihar, India	2.5	11.8	Patnaik (1975)
Tungabhadra, Karnataka, India	12.1	50.6	Patnaik (1975)
Nizam Sagar, Andhra Pradesh, India	0.66	10.8	Patnaik (1975)
Ukai, Gujarat, India	9.2	26.8	Patnaik (1975)
Magat, Luzon, Philippines	5.5	11.0	Wooldridge (1986)
Kamburu, Kenya	0.3	2.3	Wooldridge (1984)
Ambuklao, Luzon, Philippines			Abernethy, (1984, Personal Communication).
1956–1967	2.72	2.45	
1967–1980	2.72	4.98	

The processes involved in sediment supply, transport, and deposition are generally not linearly dependent on the causal factors, nor do they relate to one another in a linear manner. Rather, at any scale from the field to a large river basin, there will be a cycle of supply-transport-deposition (Figure 1), which will repeat in both space and time. Thus, sediments that are deposited at a point in a river basin may become part of the supply chain for sites further downstream or may remain deposited for long time periods. This sediment erosion-transport-deposition cycle is summarized in the concept of a sediment delivery ratio (SDR) (Walling, 1983; Walling, 1988).

There may be large time lags built into the sediment supply-transport-deposition system as is the case for man-made structures such as dams or estuarine barrages, although natural barriers, such as hedgerows or topographic depressions may also store sediments for periods of months to centuries. Sediment release from such stores may happen in extreme or catastrophic events, or may be as a result of human intervention. Similar sediment release occurs as a result of mass movements or river bank collapses, which may be associated with very wet conditions, or occurrences such as earthquakes. Such sudden high inputs of sediment would mean that supply is likely to exceed the capacity of water to move these sediments either locally or further through the system. Thus, in our modeling, we need to consider the concepts of supply and transport limitations on sediment yield as well as including the capability to model extreme events.

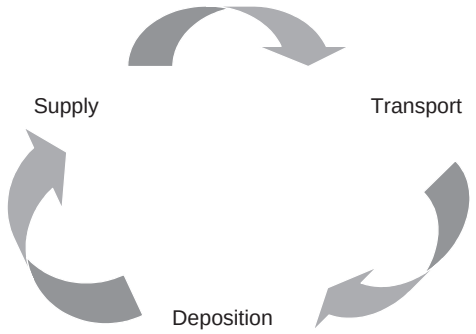


Figure 1 The sediment cycle. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

As has been demonstrated in previous articles (*see* **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**; **Chapter 81, Erosion Monitoring, Volume 2**; **Chapter 82, Erosion Prediction and Modeling, Volume 2**; **Chapter 84, Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands, Volume 2**; **Chapter 88, Reservoir Sedimentation, Volume 2**; **Chapter 85, Sediment Yields and Sediment Budgets, Volume 2**; **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**), soil erosion, sediment transport, and deposition processes are highly variable in both time and space. Monitoring is therefore complex and expensive. Modeling provides an alternative approach to improved estimation and understanding of sediment movement and delivery and, in addition, can be used to investigate possible future land-use and climate scenarios. However, it should be stated early in this discussion that sediment yield modeling is not a complete substitute for monitoring. Any mathematical representation of a complex system requires validation against real-world data. Once validated, a model can be a powerful tool for investigating various possible natural and/or anthropogenic scenarios to provide a range of possible sediment yield futures, that is, sediment yield prediction. Validation should not simply be a comparison of predicted and measured sediment yields,

but should also evaluate the adequacy of the model to represent processes. This could be done in a variety of ways depending on the complexity of the model, from considering sediment yield estimates at various subbasin scales, to comparing model performance in wet versus dry conditions (here a cross-calibration exercise might be useful), to comparing the shapes of sediment flux curves and sediment size distributions at detailed timescales and validating their behavior with respect to water discharge curves (e.g. comparing times of sediment and water flow peaks).

Recently, studies have become much more concerned with the dynamic nature of sediment yield, both within river systems and at sink points such as natural or artificial water and sediment storage bodies. For rivers and estuaries, sediment yield dynamics are important in terms of ecosystem health and for planning and management of dredging for navigation or flood protection. Sediments may also have contaminants bound to them and any assessment of contaminant delivery may also need an estimate of sediment yield. For impounded water bodies, such information can be used to manage the impoundment in a way that minimizes sediment trapping or water quality problems, and this maximizes the useful life of the structure. Thus, the interest in and demand for sediment yield prediction and modeling has grown rapidly.

For most rivers in the world, the amount of sediment transported by water is limited by the amount of sediment available for transport, rather than the capacity of the river to carry sediment. The most extreme sediment concentrations are seen in rivers in China, where the combination of fine-grained soil, erosive rainfall, and overland flow and extensive land disturbance for agricultural production produce high rates of sediment supply to the river system. In other locations, sediment supply limitations occur both long-term and at the event level as demonstrated through the hysteretic nature of sediment concentration versus water discharge loops. Any apparent relationship between instantaneous sediment concentration and water flow, or between sediment load and water discharge at the event level, is because both supply and transport mechanisms involve many of the same causal factors. An example is precipitation, which both erodes soil and contributes to river flow: more intense rainfall has more erosive power and is likely to result in more rapid and higher peak river flows.

Sediment sources have conventionally been thought of as fields, hillsides, forests, construction sites, and so on: a range of spatially distributed sources around the river basin whose capacity to supply sediment varies with both climate inputs and the way in which these areas are managed. It is here that much of the monitoring and research effort has been focused. However, many recent studies have shown that riverbanks and roads also play a very important role in supplying sediment to the river (Walling and Woodward, 1995; Walling *et al.*, 1999; Carter *et al.*, 2003; Gruszowski *et al.*, 2003), and, in some cases, this input dominates over other sediment sources. Activation of these sources may be infrequent and the material supplied from riverbank sources may have been previously deposited by the river system – the timescale of sediment yield prediction therefore needs to be long. The linkage between river morphology and sediment yield is often not addressed in sediment yield modeling and yet may prove to be fundamental to ensuring better yield assessment or prediction.

There is a further complication in the task of sediment yield modeling and this is because sediments occur in a range of shapes, sizes, and densities. In general, this variability is determined by the soils of the river basin upstream of a point of interest, although there are actions taken by humans that can change particle structure and/or the accessibility of parts of the sediment size range to transport. Thus, when we cut down forests, plough up rangeland, dump mine waste, or lay concrete over a new urban area, we are changing the balance between supply and transport that nature would find for itself, and in so doing we also change the proportions of different sediment particles available for transport. Human activity also adds a range of particulate matter to the equation through industrial and sewage discharges to rivers and to the atmosphere, through automobile exhausts and through a range of other activities. The role of flocculation in altering sediment transport characteristics has been discussed in **Chapter 83, Suspended Sediment Transport – Flocculation and Particle Characteristics, Volume 2**. A good sediment yield model should be capable of representing or reproducing the effects of such complexity, either explicitly through mathematical relationships or implicitly through process understanding.

APPROACHES TO SEDIMENT YIELD MODELING

Historically, there have been a number of attempts to estimate sediment yield using modeling, but these can be broken down into four main groups:

1. Empirical models based on broad catchment and climate descriptors, where sediment yield equations are derived from known basin characteristics.
2. Soil erosion and sediment delivery approaches in which measured or estimated soil erosion rates are factored by an SDR, which is often based on catchment characteristics.
3. Physically based and/or distributed basin modeling approaches in which movement of water and soil is estimated in a distributed way throughout a catchment.
4. Models relating sediment concentration or load to river flow where measured sediment concentration data is related to river flow characteristics.

1. Empirical models based on broad catchment and climate descriptors

Some of the earliest attempts to model basin sediment yield were those based on relationships developed between basin landscape, soil and vegetation characteristics, and basin precipitation. Notable amongst these simple models are those developed by Langbein and Schumm (1958) who related sediment yield to effective rainfall in small basins, Fournier (1960) who related rainfall to sediment yield for a series of 78 river basins around the world, and Fleming (1969) who used data from 240 basins worldwide to produce a range of formulae relating sediment load to catchment area, mean annual flow discharge, and land use. Similar thinking is seen in the soil erosion versus average annual precipitation relationship proposed by Hudson (1971), where erosion rates with natural vegetative cover peak at around 750-mm annual precipitation, although greatly increased erosion rates are possible at higher rainfalls once natural vegetative cover is disturbed. However, such models do not account for variability within a basin or changes that may take place. In addition, relationships change around the globe as precipitation patterns, soil types, topography, land use and management, and underlying geology change. A method to address such shortcomings was introduced by Al Kadhimi (1980) where an attempt was made to improve on the Fournier and Fleming models by including factors describing the catchment and its rainfall. A range of global sediment yield maps have been produced using the same general concepts as described above (Walling and Webb, 1983; Fournier, 1960; Strakhov, 1967). These show areas of high sediment yield in southeast Asia, west Africa, the western side of South America, New Zealand, and some Mediterranean areas.

More sophisticated versions of this approach include regression models in which explanatory variables may include water discharge, precipitation, parameters representing vegetative cover, soil type, soil moisture, and so on. These parameters may be spatially distributed and/or weighted. In a study in Malawi White *et al.*, (1988) reported explanation of some 69% of variability in sediment yield from four small basins using precipitation, flow at the basin outlet, total event rainfall (mm), total event runoff (mm), peak 30-minute rainfall intensity ($mm\,h^{-1}$), and runoff coefficient as explanatory variables. Figure 2 shows predicted versus observed event sediment yields for the four basins. The four basins were similar in all aspects apart from their land use and management, resulting in sediment yields differing by 100 times between the lowest (forested (Mphezo) and complete land-use plan (Bvumbwe)) and highest (traditional agriculture (Mindawo I)) sediment yields. An intermediate basin (physical conservation only (Mindawo II)) had annual sediment yields 10 times higher than the forested site and 10 times less than the traditional agricultural site.

Recent work has investigated the use of neural networks rather than multiple regression models to fit the best relationships between causal and resultant values. For the same Malawian data set, Abrahart and White (2001) developed a neural network model which outperformed the multiple linear regression technique, and also offered transferability to other sites. Figure 3 gives the same predicted to observed relationships for the neural network model as seen above for the multiple linear regression approach.

However, there are still major shortcomings with such empirical approaches:

- Such generalized modeling or mapping cannot be used to give reliable estimates of sediment yield for engineering purposes.

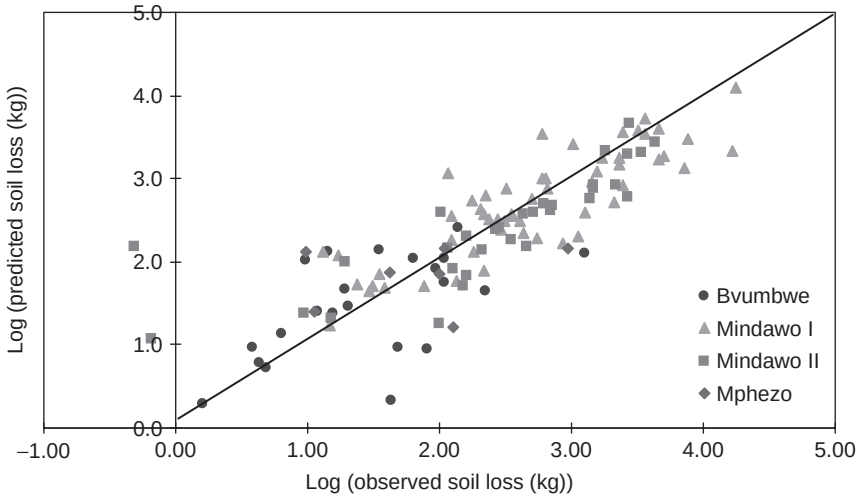


Figure 2 Predicted and observed sediment yield for runoff events in four small basins in Malawi derived from multiple linear regression relationship. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

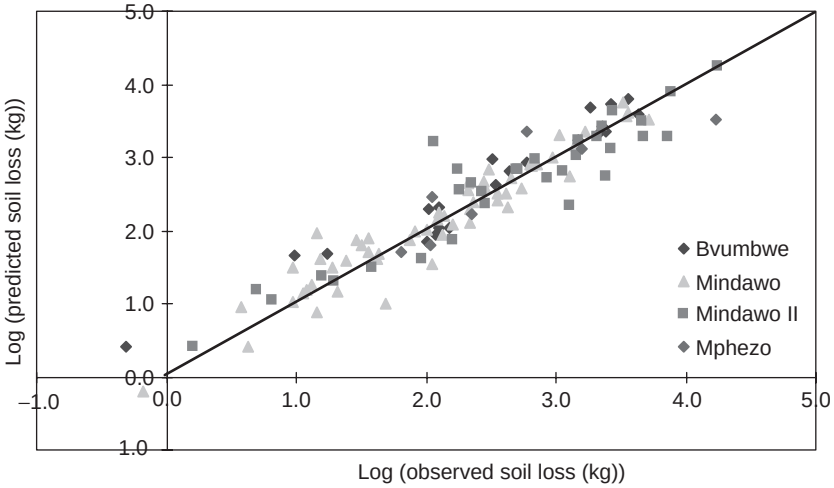


Figure 3 Predicted and observed sediment yield for runoff events in four small basins in Malawi derived from a neural network model using the same inputs as for multiple linear regression relationship. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

- Where these approaches have been tested on a range of basins, the fact that monitored sediment yield for similar catchments can vary by orders of magnitude means that there is little hope of accurate yield assessment.
- The presence of man-made structures such as dams, or natural lakes or swamps, also complicates the application of such general models and means that beyond a broad estimate of potential yield under natural conditions their usefulness is limited.
- These approaches provide limited capability for modeling changed situations.

2. Soil erosion and sediment delivery approaches

As suggested by the title, two parallel strands of research contributed to these approaches. The first was the development of empirical soil erosion models such as the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1978) or the Soil Loss Estimator for Southern Africa (SLEMSA) (Elwell, 1980). These were succeeded by modifications such as RUSLE (Revised Universal Soil Loss Equation) (Renard *et al.*, 1997) and MUSLE (Modified Universal Soil Loss Equation) (Williams, 1975) and subsequently by more physically based models such as WEPP (Water Erosion Prediction Project) (Flanagan *et al.*, 2001), KINEROS2 (Kinematic Runoff and Erosion 2 Model) (Smith *et al.*, 1995) and EUROSEM (European Soil Erosion Model) (Morgan *et al.*, 1998). Such models allow estimation of actual or potential annual soil loss – or sediment yield – at the field scale. By summing up soil loss estimates across space to cover a complete river basin, an estimate of total soil loss could be made. However, not all of this eroded soil will be supplied to the river network: some will be trapped for very long periods behind obstacles, some will be double accounted as previously deposited sediment is reeroded. Thus was born the concept of the SDR, which relates the amount of sediment leaving a river basin to the total eroded sediment supply within it. There is a wide range of formulae for the SDR, most of them relating SDR to basin area, basin relief, basin shape, annual runoff, stream network characteristics, and gully density (e.g. Maner, 1958; Roehl, 1962; Williams and Berndt, 1972; Williams, 1977; Mutchler and Bowie, 1975; Mou and Meng, 1980). Many authors include in their SDR investigations the formula suggested by ASCE (American Society of Civil Engineers) (1975):

$$\mathrm{SDR} = kA^n$$

where A = basin area, and k and n are numerical constants for a particular basin. Reported values of n vary from -0.01 to -0.7, with some of this range being explained by regional differences and some by differences in sediment characteristics. Intuitively, one would expect an inverse relationship between SDR and basin area, as the smaller, steeper headwaters provide more efficient transfer of material than larger, flatter basins.

Application of the erosion + SDR methodology was considerably facilitated by the development of Geographical Information Systems (GIS), which allow automated application of a range of erosion models for a regular grid covering an entire river basin. A major shortcoming was the definition of a lumped SDR value for an entire river basin, meaning that in practice one normally required monitored sediment yield in order to calculate the SDR as the ratio of total yield to total erosion.

Ferro and Minacapilli (1995) addressed the problem of the lumped SDR approach by introducing a spatially distributed SDR at the level of morphological units.

Their SEDD model uses an empirical travel time concept, based on slope length and steepness, to investigate the effectiveness of sediment transport through morphological units. The model assumes that rivers can transport all supplied sediment. Mashriqui and Cruise (1997) have developed a methodology based on "grouped response units" defined by topography and soil characteristics. Kothyari *et al.* (1994) used a similar approach, but here the basin was subdivided into time–area segments for routing of sediment overland. Other approaches have continued to use USLE but have calculated delivery to the stream network as a function of slope, distance to the stream, or aspect (e.g. Lea, 1991).

The WaTEM/SEDEM (Van Oost *et al.*, 2000; Van Rompaey *et al.*, 2001) models provide a slightly different approach using a grid-based model to estimate annual soil erosion using RUSLE and then calculating annual transport capacity for individual cells. The models allow calculation of annual sediment delivery to rivers from hillslopes. Eroded sediment is routed across the land surface to rivers in a topographically driven manner, restricted by the annual transport capacity of flow. Original claims that the SEDEM (Sediment delivery model) model could be applied to basins up to 10 000 km^2 have since been questioned (Verstraeten *et al.*, 2003). However, SEDEM has been used to look at the impact of land-use change in the Czech Republic (Van Rompaey *et al.*, 2003), for defining catchment management priorities in Belgium (Verstraeten *et al.*, 2002), and for looking at the erosion and sediment delivery implications of historic and possible future land-use changes in Belgium (Van Rompaey *et al.*, 2002).

Although many combinations of erosion and sediment delivery modeling are available, they all still require calibration and thus cannot be transferred to other catchments and environments. Other shortcomings include:

- USLE (and its derivatives) and SLEMSA are not easily applied to nonagricultural land uses or to areas outside the range of original development and application.
- The approach ignores the fact that fields are often not the dominant sediment sources – gullies, mass movements, avalanches, and river banks are all major contributors to sediment supply, and one or a combination of them often far exceed field-scale sheet and rill erosion as modeled by soil erosion models.
- The error in the sediment supply term can be compensated for in the sediment delivery term where data exist to determine the SDR empirically. More often, SDR is calculated as a function of catchment characteristics such as slope length and steepness, and no allowance is made for other landscape complexity.

3. Physically based and/or distributed basin modeling approaches

A different modeling approach is taken by Arnold *et al.* (1995), in which soil–water balance models generate boundary conditions for a sediment generation and routing model called ROTO (Routing outputs to outlets). This model was developed for use at the river basin scale where capability to represent change in land use and agricultural management is required.

As discussed in the article on erosion modeling (*see* **Chapter 82, Erosion Prediction and Modeling, Volume 2**), models such as the EUROSEM (Morgan *et al.*, 1998) and WEPP (Flanagan *et al.*, 2001) were an attempt to bring more physically realistic components into erosion modeling at a small catchment scale. Again, these are dealing with erosion from fields rather than sediment yield in its totality. In order to incorporate the ability to model supply, and transport of sediment into larger catchment-scale models, it is necessary to have a realistic representation of water movement through the catchment with correctly defined components of total river flow moving to the river by their correct routes, that is, models should have the ability to model surface, through flow and groundwater contributions to streams in a spatially distributed and dynamic manner. This demands a higher level of both complexity in the model structure and sophistication in model validation than that used for straightforward hydrological modeling. Models that have the ability to model surface flow as a separate component include those that are predominantly physically based (e.g. SHE-TRAN, Ewen *et al.*, 2000; Wicks, 1988; Wicks and Bathurst, 1996; Areal Nonpoint Source Watershed Environmental Response Simulation (ANSWERS), Bouraoui and Dillaha, 1996; Bouraoui and Dillaha, 2000) and those that are more conceptual in approach (e.g. Soil and Water Assessment Tool (SWAT), Neitsch *et al.*, 2001; Arnold *et al.*, 1998; a development from the CREAMS (Chemicals Runoff and Erosion from Agricultural Management Systems) approach discussed in the erosion modeling article (*see* **Chapter 82, Erosion Prediction and Modeling, Volume 2**), Williams and Arnold, 1997) and AGNPS (Agricultural Nonpoint Source) (Young *et al.*, 1989). Both approaches may explicitly include components for modeling of erosion and transfer of sediment; however, even in the most advanced physically based models, there are still some aspects of sediment behavior that are represented by empirical relationships.

More recently, a range of erosion and overland sediment transport approaches based on overland flow transport capacity have been developed. Notable amongst these is the Limburg Soil Erosion Model, LISEM (De Roo *et al.*, 1996a,b; Jetten and De Roo, 2001). LISEM is a physically based single-event erosion model that works on a raster-grid basis. It has been tested, with varying success, in a range of

environments (e.g. Hessel *et al.*, 2003; Jetten *et al.*, 2003) with limitations sometimes due to unmet data requirements and sometimes due to physical limitations in the model equations or grid size used. LISEM calculates infiltration using a finite difference version of the Richard's equation and then calculates the transport capacity of the resulting overland flow using stream power approaches. The model is very sensitive to the parameters controlling infiltration (saturated conductivity and initial moisture content) and to the slope angle, where the stream power concept is only valid for slopes of up to 20%.

None of the currently available distributed modeling approaches at river basin scale includes all potential sediment source mechanisms in a physically realistic manner. It is indeed arguable that if such complete physically based models were developed, available data would not be sufficient to provide a full spatial and temporal representation of all causal factors. Possible future developments in this area include the use of remote sensing and Internet databases to address some of the spatiotemporal data acquisition issues implicit in physically based modeling. In their defence, these models do provide the possibility of investigating future climate or land-use scenarios (Bathurst *et al.*, 1996; Bathurst *et al.*, 2002), and if used with care and thoroughly validated should prove to be robust.

- Although these models are of use in modeling the dynamics of sediment and associated contaminants, they are very data hungry and still rely on empirical relationships for many of the sediment detachment and transfer processes.
- Some of these large basin models use USLE to estimate erosion rates, and thus the same caveats apply as detailed above.
- These models do not provide quick answers. Setting up, calibrating, and modeling a river basin model will typically require a time input of several man-months.
- However, such a modeling approach does provide a way to investigate possible future scenarios in climate or land use and management, and can include repeated application to account for uncertainty.

4. Models relating sediment concentration or load to river flow

This is the most commonly used approach for sediment yield assessment in practice, and assumes that river flow is the dominant factor in sediment yield, rather than sediment supply, or that these two controlling factors vary in parallel. It is important to remember that the concentration of the finest material in transport, the wash load, is explicitly defined as being independent of water discharge, although here the relationship between supply and transport means that peaks in wash load are often seen at times of high flow. A relationship between suspended bed material sediment concentration or sediment load and water discharge – a sediment-rating curve (American Society of Civil Engineers. Task Committee for the Preparation of the Manual on Sedimentation, 1975) – can be derived. Such rating curves always demonstrate a wide scatter of points in part due to hysteresis effects at the event scale. These occur because of sediment exhaustion as sediment supplies are transported down the river system or sediment "bunching" as material that is mobilized at the subcatchment scale in spatially limited events is then moved through the system in events affecting larger areas. Such an effect was reported in the Philippines (White, 1995) where localized thunderstorm events prepared sediments at the subcatchment scale, but did not cause high transporting capacity (flow) throughout the river system. Once larger cyclone events occurred, these previously mobilized sediments were rapidly transported through the system.

Increases in sediment concentration for a particular water flow rate have also been seen for the River Tees in the United Kingdom, following severe flooding in the winter of 2000–2001. Here, a rating curve between mean event sediment concentration and peak event flow has been derived for the period before the flood event. This showed relatively little scatter and had performed well over a range of river discharges. After the flood events, when a considerable amount of sediment was released from riverbanks as a result of bank collapse, the relationship between sediment concentration and peak flow moved upwards on the graph (Figure 4). The two lines are significantly different by a factor of up to 2 in mean sediment concentration for a particular size of event. This new relationship persisted over several months while sediment stocks in the river were exhausted. By March 2003, the rating curve had still not completely returned to, but was approaching, the pre-flood line. The Tees river is unusual in having an estuarine barrage close to its outlet to the North Sea. Before this barrage was built, a feasibility assessment had estimated sediment yield to the barrage at 35 630 t year^{-1}. In the water year 2000–2001, monitored sediment yield was in excess of 85 000 t, and yield per unit flow (tons per millimeter of specific flow) had increased from 78 in the initial feasibility study to 111 in the post-flood period. In this particular situation, the increased sediment yield has little influence on the viability of the barrage impoundment, as high sediment loads occur with high river flows, which have the ability both to transport material through and out of the system and to scour out previously deposited sediments. For the Tees, 80% of sediment yield is delivered by the top 10% of the river flows, thus emphasizing the need for both data and models that address such events.

Extreme examples of high sediment yield events are seen in mountain basins where events that may occur only once in a decade or longer can dominate sediment yield patterns

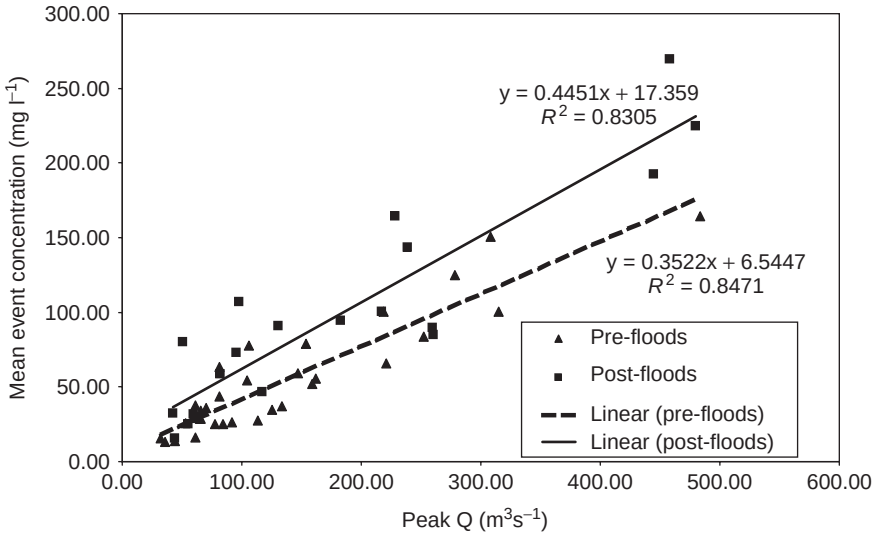


Figure 4 Sediment-rating curves for the Tees river, UK, pre- and post-floods in winter 2000–2001

for years to come (Alvera and García-Ruiz, 2000; White and García-Ruiz, 1998; White *et al.*, 1997b). A single event in a 30-ha basin in the high Pyrenees transported more sediment than the total seen in the following 5 years, in part because of sediment exhaustion following the event, in part because of the extreme nature of the event and the high flows associated with it (White *et al.*, 1997a). Hence, extreme care is required in interpreting data and extrapolating relationships derived for such areas.

- Although these approaches are based on real sediment flux data, they still require a large amount of data gathering to give realistic estimates of a long-term average annual sediment yield.
- It is important that data gathering is carried out across a range of flow and sediment supply conditions. Special attention should be paid to extreme, high flow, and storm events.
- These approaches are based on what has happened in the past rather than what may happen in the future. Understanding of sediment supply and transport processes is therefore required to extrapolate to unmonitored future climate and/or land-use scenarios.
- Models can be transferred only to situations that have been seen before in the data set. The large difference between apparently similar basins in Malawi demonstrates the scale of error that can be made for relatively minor land management changes.

SEDIMENT YIELD PREDICTION

Prediction is by definition a forward-looking activity and thus has a large number of unknowns and uncertainties attached to it. These undefined aspects have rarely been addressed in conventional feasibility studies for dams, where one may be predicting sediment yields forward for the next 100 or more years. Most prediction of this type has been done assuming stability in the driving processes, clearly not a reasonable assumption in today's world where climate change is widely accepted to be influencing river flows and hydrological behavior of river basins. Given the major changes in land use that have been seen worldwide over the last century, such an assumption becomes completely unrealistic and unsupportable. **Chapter 88, Reservoir Sedimentation, Volume 2** suggests a changed approach to reservoir design and management, based on life-cycle concepts (Palmieri *et al.*, 2003) rather than the design life approach traditionally used. In the latter, sedimentation was accepted as given, and reservoirs were designed to fulfil their purpose over a limited number of years, often based on very poor preconstruction sediment yield estimates (see Table 1). One of the outcomes of this lack of sediment yield estimation capability is the annual reduction in water storage capacity in reservoirs, now being seen internationally (World Commission on Dams, 2000).

In a feasibility study for a proposed new dam in the Philippines, (White, 1987; White, 1993) sediment yield estimates were made using an uncertainty approach in which model inputs such as the slope and intercept of a sediment-rating curve, the water level in the reservoir at the time of sediment delivery, and the frequency and impact of cyclone events were varied over reasonable ranges. A daily time-step model based on different statistical flow distributions and rating curves for wet and dry season conditions was used as the basis for sediment yield estimation. Cyclone events, which fall outside the standard frequency distribution patterns, were then superimposed on more "normal" flow patterns as additional intermittent

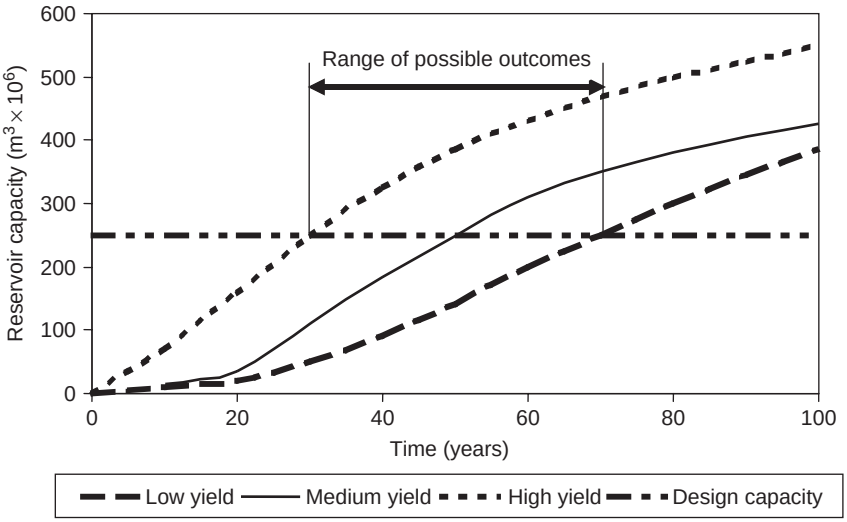


Figure 5 Idealized approach to prediction of a range of sediment yield rates and their relation to viable reservoir life

sediment pulses. Sediment inputs to the reservoir were then deposited in locations related to current water level, and sediment deposits were thus built up in a spatially varying manner. Failure of this reservoir was determined by sedimentation threatening a water transfer tunnel located along the reservoir profile. Parameters within this modeling framework were then subject to variation, providing a range of predictions of effective reservoir life (Figure 5). These different scenarios could then be tested against the required economically operational life of the reservoir. Such an approach was novel, as previous studies only provided an average annual sediment yield with no definition of deposition locations or possible changes in sediment supply regime or river flows. By investigating the range of possible outcomes for the proposed reservoir, a much more robust assessment of viability could be made.

Since this study, other authors have looked at the inclusion of uncertainty in sediment yield prediction. Work in Europe, using SHE-TRAN, has looked at model application to future land-use and climate scenarios in order to assess a range of possible outcomes, including sediment yield (e.g. Bathurst *et al.*, 1996; Bathurst *et al.*, 2002) using a varying model parameterization to investigate the range of outcomes. Tattari and Barlund (2001) used their ICECREAM model, a field–scale mathematical simulation model for prediction of water transport, soil and nutrient losses, to investigate the uncertainty in model output arising from parameter uncertainty. The uncertainty analysis was carried out using a Monte Carlo simulation package. A similar Monte Carlo approach was used by Salas and Shin (1999) to look at error in reservoir sedimentation rates in Colorado. They concluded that annual streamflow and sediment load were the two most important factors determining sedimentation rate, emphasizing again the need for robust sediment yield modeling and a realistic consideration of error. Moving on still further, Tate and Farquharson (2000) used varying sediment yield and flow inputs to the Tarbela dam to study different management scenarios and economic consequences. Such uncertainty in sediment yield estimation will continue to exist and should always be taken into account when assessing the viability, operation, or future development of an impoundment or other sediment delivery zone.

RECOMMENDATIONS

To conclude, a number of recommendations can be made about the way in which sediment yield is modeled, in order that a robust estimate or prediction can be made:

1. Understand what is controlling sediment yield in the basin of interest.
2. Be realistic about how things – land-use, land management, climate – have changed and how they are likely to change.
3. Never make a single number prediction of sediment yield – at least include some error bars or a distribution of likely yields.
4. All estimates require some sediment yield data – for model development, calibration, and validation.
5. Available data must be put into a longer term context.
6. Extreme events must be included in the assessment as they often control sediment yield over the long term.

FURTHER READING

Córdova J.R. and González M. (1997) Sediment yield estimation in small watersheds based on streamflow and suspended sediment discharge measurements. *Soil Technology*, **11**, 57–65.

Dickinson A., Amphlett M.B. and Bolton P. (1990) *Sediment Discharge Measurements Magat Catchment: Summary Report 1986–1988*, Hydraulics Research Report No. OD 122, Hydraulics Research, Wallingford.

Flanagan D.C. and Nearing M.A. (Eds.) (1995) *USDA Water Erosion Prediction Project. Hillslope Profile and Watershed Model Documentation*, USDA-ARS-NSERL Report No. 10, NSERL, West Lafayette.

Nagy H.M., Watanabe K. and Hirano M. (2002) Prediction of sediment load concentration in rivers using artificial neural network model. *Journal of Hydraulic Engineering*, **128**, 588–395.

Peronne J. (1999) Sediment yield prediction using AGNPS. *Journal of Soil and Water Conservation*, **54**(1), 415–419.

Tawfik M., Ibrahim A. and Fahmy H. (1997) Hysteresis sensitive neural network for modeling rating curves. *Journal of Computing in Civil Engineering*, **11**(3), 206–211.

Wass P.D. and Leeks G.J.L. (1999) Suspended sediment fluxes in the Humber catchment, UK. *Hydrological Processes*, **13**(7), 935–953.

Williams J.R. and Berndt H.D. (1977) Sediment yield prediction based on watershed hydrology. *Transactions of ASAE*, **20**(6), 1100–1104.

SWCS (1995) *User Guide: Revised Universal Soil Loss Equation. Version 1.04*, Soil and Water Conservation Society, p. 145.

REFERENCES

Abrahart R.J. and White S.M. (2001) Modelling sediment transfer in Malawi: comparing back-propagation neural network solutions against a multiple linear regression benchmark using small data sets. *Physics and Chemistry of the Earth*, **26**(1), 19–24.

Al Kadhimi A.M.H. (1980) *Land use, Water Yield and Soil Erosion: Simulation of Cause and Effect*, PhD thesis, Department of Civil Engineering, University of Strathclyde, Glasgow.

Alvera B. and García-Ruiz J.M. (2000) Variability of sediment yield from a high mountain catchment, central Spanish Pyrenees. *Arctic Antarctic and Alpine Research*, **32**(4), 478–484.

American Society of Civil Engineers. Task Committee for the Preparation of the Manual on Sedimentation (1975) In *Sedimentation Engineering*, Vanoni V.A. (Ed), ASCE: New York.

Arnold J.G., Srinivasan R., Muttiah R.S. and Williams J.R. (1998) Large area hydrologic modeling and assessment part I: model development. *Journal of American Water Resources Association*, **34**(1), 73–89.

Arnold J.G., Williams J.R. and Maidment D.R. (1995) Continuous-time water and sediment routing model for large basins. *Journal of Hydraulic Engineering ASCE*, **121**(2), 171–183.

Asselman N.E.M., Middelkoop H. and van Dijk P.M. (2003) The impact of change in climate and land use on soil erosion, transport and deposition of sediment in the River Rhine. *Hydrological Processes*, **17**, 3225–3244.

Bathurst J.C., Kilsby C.G. and White S.M. (1996) Modelling the impacts of climate and land use change on basin hydrology and soil erosion in Mediterranean Europe. In *Mediterranean Desertification and Land Use*, Thornes J.B. and Brandt C.J. (Eds.), Wiley.

Bathurst J., Sheffield J., Vicente C., White S.M. and Romano N. (2002) Modelling large basin hydrology and sediment yield with sparse data: the Agri basin, southern Italy. In *Mediterranean Desertification: A Mosaic of Processes and Responses. Part 2: Regional Studies*, Geeson N.A., Brandt C.J. and Thornes J.B. (Eds.), John Wiley & Sons.

Bouraoui F. and Dillaha T.A. (1996) ANSWERS-2000: runoff and sediment transport model. *Journal of Environmental Engineering, ASCE*, **122**, 493–502.

Bouraoui F. and Dillaha T.A. (2000) ANSWERS-2000: non-point-source nutrient planning model. *Journal of Environmental Engineering, ASCE*, **126**, 1045–1055.

Brabben, T.E., 1982 Sedimentation *Survey, Karangkates Reservoir, East Java, Indonesia. Field data and Reservoir Survey*, Hydraulics Research Report No. OD 50, Hydraulics Research, Wallingford, June 1982.

Carter J., Owens P.N., Walling D.E. and Leeks G.J.L. (2003) Fingerprinting suspended sediment sources in a large urban river system. *Science of the Total Environment*, **314**, 513–534.

De Roo A.P.J., Offermans R.J.E. and Cremers N.H.D.T. (1996b) LISEM a single-event physically based hydrological and soil erosion model for drainage basins. II. Sensitivity analysis, validation and application. *Hydrological Processes*, **10**, 1119–1126.

De Roo A.P.J., Wesseling C.G. and Ritsema C.J. (1996a) LISEM a single-event physically based hydrological and soil erosion model for drainage basins. I. Theory, input and output. *Hydrological Processes*, **10**, 1107–1117.

Elwell H.A. (1980) A soil loss estimation technique for Southern Africa. *Proceedings of Conservation'80*, *Silsoe Research Institute*: Bedford.

Ewen J., Parkin G. and O'Connell P.E. (2000) SHETRAN: distributed river basin flow and transport modeling system. *Proceedings of the American Society of Civil Engineers, Journal of Hydrologic Engineering*, **5**(3), 250–258.

Ferro V. and Minacapilli M. (1995) Sediment delivery processes at basin scale. *Hydrological Sciences Journal*, **40**(6), 703–717.

Fish I.L. (1983) *Reservoir Sedimentation Study, Wlingi, East Java, Indonesia*, Hydraulics Research Report No. OD 59, Hydraulics Research, Wallingford.

Flanagan D.C., Ascough J.C., Nearing M.A. and Laflen J.M. (2001) The Water Erosion Prediction Project (WEPP) model. In *Landscape Erosion and Evolution Modelling*, Harmon R.S. and Doe W.W. (Eds.), Kluwer Academic: New York, pp. 477–516.

Fleming G. (1969) Design curves for suspended sediment load estimation. *Proceedings of Institution of Civil Engineers*, **43**, 1–9.

Fournier F. (1960) *Climat et Erosion, la Relation Entre L'erosion du sol par l'eau et les Precipitations Atmospheriques*, Presses Universitaires de France: Paris.

García-Ruiz J.M., Lasanta T., Martí C., González C., White S.M., Ortigosa L. and Ruiz Flaño P. (1995) Changes in runoff and erosion as a consequence of land-use changes in the central

Spanish Pyrenees. *Physics and Chemistry of the Earth*, **20**(3/4), 301–307.

Gruszowski K.E., Foster I.D.L., Les J.A. and Charlesworth S.M. (2003) Sediment sources and transport pathways in a rural catchment, Herefordshire, UK. *Hydrological Processes*, **17**, 2665–2681.

Harden C.P. (1993) Land-use, soil-erosion and reservoir sedimentation in an Andean drainage-basin in Ecuador. *Mountain Research and Development*, **13**(2), 177–184.

Hessel R., Jetten V., Baoyuan L., Yan Z. and Stolte J. (2003) Calibration of the LISEM model for a small loess plateau catchment. *Catena*, **54**, 235–254.

Hudson N. (1971) *Soil Conservation*, Batsford Academic.

Jetten V. and De Roo A.P.J., (2001) Spatial analysis of erosion conservation measures with LISEM. In: *Landscape Evolution and Erosion Modelling*, Chap. 14, Harmon R. and Doe W.W. (Eds.), Kluwer Academic Plenum: New York, pp. 429–445.

Jetten V., Govers G. and Hessel R. (2003) Erosion models: quality of spatial predictions. *Hydrological Processes*, **17**, 887–900.

Kothyari U.C., Tiwari A.K. and Singh R. (1994) Prediction of sediment yield. *Journal of Irrigation and Drainage Engineering ASCE*, **120**(6), 1122–1131.

Lang A., Bork H.-R., Mackel R., Preston N., Wunderlich J. and Dikau R. (2003) Changes in sediment flux and storage within a fluvial system: some examples from the Rhine catchment. *Hydrological Processes*, **17**, 3321–3334.

Langbein W.B. and Schumm S.A. (1958) Yield of sediment in relation to mean annual precipitation. *Transactions, American Geographical Union*, **39**, 1076–1084.

Lea N.J. (1991) An aspect driven kinematic routing algorithm. *Proceedings of the Workshop on the Hydraulic and Erosion Mechanics of Overland Flow*, University of Keele, University College Press: London.

Lu X.X. and Higgitt D.L. (1998) Recent changes in sediment yield in the Upper Yangtze, China. *Environmental Management*, **22**(5), 697–709.

Maner S.B. (1958) Factors affecting sediment delivery rates in the Red Hills physiographic area. *Transactions, American Geographical Union*, **39**, 669–675.

Mashriqui H.S. and Cruise J.F. (1997) Sediment yield modelling by grouped response units. *Journal of Water Resources Planning and Management ASCE*, **123**(2), 95–104.

Morgan R.P.C., Quinton J.N., Smith R.E., Govers G., Poesen J.W.A., Auerswald K., Chisci G., Torri D. and Styzcen M.E. (1998) The European Soil Erosion Model (EUROSEM): a dynamic approach for predicting sediment transport from fields and small catchments. *Earth Surface Processes and Landforms*, **23**(6), 527–544.

Mou J. and Meng Q. (1980) *Sediment Delivery Ratio as Used in the Computation of Watershed Sediment Yield*, Chinese Society of Hydraulic Engineering: Beijing.

Mutchler C.K. and Bowie A.J. (1975) Effect of land use on sediment ratios. *Proceedings of the Third Federal Inter-Agency Sedimentation Conference*, US Water Resources Council: Washington, pp. 1, 11–12.

Neitsch S.L., Arnold J.G., Kiniry J.R. and Williams J.R. (2001) *Soil and Water Assessment Tool User's Manual, Version 2000*, Grassland, Soil and Water Research Laboratory, USDA Agricultural Research Service: Temple, `ftp://ftp.brc.tamus.edu/pub/swat/doc/swat2000.pdf`

Palmieri A., Shah F., Annandale G.W. and Dinar A. (2003) *Reservoir Conservation: The RESCON Approach*, The World Bank: Washington.

Patnaik N. (1975) Soil erosion: a menace to the nation. *Indian Farming*, **24**(11), 7–10.

Renard K.G., Foster G.R., Weesies G.A., McCool D.K. and Yoder D.C. (1997) Predicting soil erosion by water – a guide to conservation planning with the revised universal soil loss equation (RUSLE). *Agricultural Handbook No. 703*, US Government Printing Office: Washington.

Roehl J.W. (1962) *Sediment Source Areas, Delivery Ratios and Influencing Morphological Factors*, IAHS Publication No. 59, IAHS, pp. 202–213.

Salas M. and Shin H.S. (1999) Uncertainty analysis of reservoir sedimentation. *Journal of Hydraulic Engineering ASCE*, **125**(4), 339–350.

Smith R.E., Goodrich D.C. and Quinton J.N. (1995) Dynamic, distributed simulation of watershed erosion: the KINEROS2 and EUROSEM models. *Journal of Soil and Water Conservation*, **50**(5), 517–520.

Strakhov N.M. (1967) *Principles of Lithogenesis*, Vol. 1, Oliver and Boyd: Edinburgh.

Tate E.L. and Farquharson F.A.K. (2000) Simulating reservoir management under the threat of sedimentation: the case of Tarbela dam on the Indus. *Water Resources Management*, **14**, 191–208.

Tattari S. and Barlund I. (2001) The concept of sensitivity in sediment yield modelling. *Physics and Chemistry of the Earth Part B*, **26**(1), 27–31.

Valero-Garcés B.L., Navas A., Machin J. and Walling D. (1999) Sediment sources and siltation in mountain reservoirs: a case study from the central Spanish Pyrenees. *Geomorphology*, **28**(1–2), 23–41.

Van Oost K., Govers G. and Desmet P.J.J. (2000) Evaluating the effects of landscape structure on soil erosion by water and tillage. *Landscape Ecology*, **15**(6), 579–591.

Van Rompaey A.J.J., Govers G. and Puttemans C. (2002) Modelling land use changes and their impact on soil erosion and sediment supply to rivers. *Earth Surface Processes and Landforms*, **27**, 481–494.

Van Rompaey A., Krasa J., Dostal T. and Govers G. (2003) Modelling sediment supply to rivers and reservoirs in Eastern Europe during and after the collectivisation period. *Hydrobiologia*, **494**, 169–176.

Van Rompaey A.J.J., Verstraeten G., Van Oost K., Govers G. and Poesen J. (2001) Modelling mean annual sediment yield using a distributed approach. *Earth Surface Processes and Landforms*, **26**, 1221–1236.

Verstraeten G., Van Oost K., Van Rompaey A., Poesen J. and Govers G. (2002) Evaluating an integrated approach to catchment management to reduce soil loss and sediment pollution through modelling. *Soil Use and Management*, **19**, 386–394.

Verstraeten G., Van Rompaey A., Poesen J., Van Oost J. and Govers G. (2003) Evaluating the impact of watershed

management scenarios on changes to sediment delivery to rivers? *Hydrobiologia*, **494**, 153–158.

Walling D.E. (1983) The sediment delivery problem. *Journal of Hydrology*, **65**, 209–237.

Walling D.E. (1988) Erosion and sediment yield research – some recent perspectives. *Journal of Hydrology*, **100**, 113–141.

Walling D.E., Owens P.N. and Leeks G.J.L. (1999) Fingerprinting suspended sediment sources in the catchment of the River Ouse, Yorkshire, UK. *Hydrological Processes*, **13**, 955–975.

Walling D.E. and Webb B. (1983) Patterns of sediment yield. In *Background to Paleohydrology*, Gregory K.J. (Ed.), John Wiley & Sons: Chichester.

Walling D.E. and Woodward J.C. (1995) Tracing sources of suspended sediment in river basins: a case study of the River Culm, Devon, UK. *Journal of Marine and Freshwater Research*, **46**, 327–336.

White S.M. (1987) *Casecnan Sedimentation Study*, Vols. I and II, HR Report EX1596. HR Wallingford: Oxon, UK.

White S.M. (1993) *Sediment Yield Estimation from Limited Data Sets: A Philippines Case Study*, PhD thesis, Exeter University, Exeter.

White S.M. (1995) Soil erosion and sediment yield in the Philippines. In *Sediment and Water Quality in River Catchments*, Foster I.D.L., Gurnell A.M. and Webb B.W. (Eds.), Wiley.

White S., Coe R. and Nice S. (1988) *Multiple Regression in Runoff and Erosion Studies: A Case Study from Malawi*, Hydraulics Research Report No. OD/TN 37, Hydraulics Research, Wallingford.

White S.M. and García-Ruiz J.M. (1998) Extreme erosional events and their role in mountain areas of northern Spain. *Ambio*, **XXVII**(4), 300–305.

White S.M., García Ruiz J.M., Martí C., Alvera B. and Del Barrio G. (1997a) Sediment transport in a high mountain catchment, central Spanish Pyrenees. *Physics and Chemistry of the Earth*, **22**(3–4), 377–180.

White S.M., García Ruiz J.M., Martí C., Valero B., Errea M.P. and Gómez Villar A. (1997b) The Biescas campsite disaster and its spatial context. *Hydrological Processes*, **11**(14), 1797–1812.

Wicks J.M. (1988) *Physically-Based Mathematical Modelling of Catchment Sediment Yield*, PhD thesis, University of Newcastle Upon Tyne, Tyne.

Wicks J.M. and Bathurst J. (1996) SHESED: a physically-based, distributed erosion and sediment yield component for the SHE hydrological modelling system. *Journal of Hydrology*, **175**, 213–238.

Williams J.R. (1975) *Sediment Yield Prediction with Universal Equation Using Runoff Energy Factor*, United States Department of Agriculture-Agricultural Research Service, ARS-S-40.

Williams J.R. (1977) *Sediment Delivery Ratios Determined with Sediment and Runoff Models*, IAHS Publication No. 122, IAHS, pp. 168–179.

Williams J.R. and Arnold J.G. (1997) A system of erosion-sediment yield models. *Soil Technology*, **11**, 43–55.

Williams J.R. and Berndt H.D. (1972) Sediment yield computed with universal equation. *Journal of Hydraulics Division, ASCE*, **98**, 2087–2098.

Wischmeier W.H. and Smith D.D. (1978) Predicting rainfall erosion loss: a guide to conservation planning. *Agricultural Handbook No. 537*, U.S. Department of Agriculture: Washington.

Wooldridge R. (1984) *Sedimentation in Reservoirs: Tana River Basin, Kenya, III – Analysis of Hydrographic Surveys of Three Reservoirs in June-July 1983*, Hydraulics Research Report No. OD 61, Hydraulics Research, Wallingford.

Wooldridge R. (1986) *Sedimentation in Reservoirs: Magat Reservoir, Cagayan Valley, Luzon, Philippines. 1984 Reservoir Survey and Data Analysis*, Hydraulics Research Report No. OD 69, Hydraulics Research, Wallingford.

World Commission on Dams (2000) *Dams and Development: A New Framework for Decision-Making*, Earthscan Publications.

Xu J. (2003) Sedimentation rates in the lower Yellow River over the past 2300 years as influenced by human activities and climate change. *Hydrological Processes*, **17**, 3359–3371.

Young R.A., Onstad C.A., Bosch D.D. and Anderson W.P. (1989) AGNPS: a nonpoint-source pollution model for evaluating agricultural watersheds. *Journal of Soil & Water Conservation*, **44**, 168–173.

88: Reservoir Sedimentation

GEORGE W ANNANDALE

Engineering and Hydrosystems Inc., Denver, CO, US

The storage loss due to reservoir sedimentation has been found to exceed the storage added worldwide by the mid 1990s. Recovery of this lost storage is estimated to range between US$ 10 billion and US$ 20 billion per year, not accounting for the growth in world population. Reservoir sedimentation does not only impact water supply and power supply reliability, but also has significant impacts on the environment, including downstream river degradation, and impacts on recreation, flood management, infrastructure, and the economy. The article illustrates how sediment is trapped in reservoirs, provides empirical techniques for calculating the volume of sediment that can be trapped by reservoirs, outlines sediment management techniques, which includes flushing, dredging, dry excavation, hydrosuction, and bypassing, and introduces the life-cycle management approach. This approach to reservoir sedimentation management is in contrast to the conventional design life approach, which is inappropriate for the design, construction, and management of dams and their reservoirs. Numerous references for further reading are listed.

INTRODUCTION

Reservoir sedimentation occurs when sediment conveyed by river flow into a reservoir settles. Sediment transported by rivers originates from riverbeds and banks, and from eroding surfaces in catchments. This sediment is carried in suspension, as bedload, or as washload as long as the sediment transport capacity of the water flowing in the river is high enough to convey it. When a river flows into a reservoir, the flow velocity of the water reduces, which leads to a reduction in sediment transport capacity and concurrent deposition of the sediment. Large particles settle first, followed by smaller particles further downstream in the reservoir. The larger particles form a delta at the upstream end of the reservoir, and the finer particles settle downstream closer to the dam. The water closer to the dam could become clearer as sediment settles out, but in some cases remains turbid. The level of turbidity in the quiescent water closer to the dam depends on the physical and chemical properties of the colloidal-sized suspended particles and the physical and chemical properties of the water. The upper, mild slope of the delta is *known as* the *topset* slope, and the steeper slope at the end of the delta as the frontset slope (Figure 1).

Sediment can also be conveyed into reservoirs by means of density currents, although the occurrence of density currents is not common (Figure 2). The density of these currents is higher than that of the water in a reservoir, resulting from high sediment concentrations in the inflowing water and temperature differences between the inflowing water and the water in the reservoir. The presence of density currents can often be detected by floating debris that remains at the point of subsidence at the upstream end of the reservoir. The higher density of a density current causes it to dive below the surface of the water in the reservoir and move along its bed towards the dam. A density current can move up the dam wall and recirculate in the reservoir. One of the management techniques used to manage sediment deposition in reservoirs that are subject to density currents is to install low-level venting gates that are opened during such events. By timing the opening of the gates correctly, it is possible to vent the high sediment load carried by a density current and discharge it downstream of the dam. Density current venting makes it possible to reduce the volume of sediment depositing in a reservoir, although this is not common.

As sediment is deposited in reservoirs, it reduces the useful capacity of these facilities, which, in turn, reduces the reliability of water and power supply, especially during

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

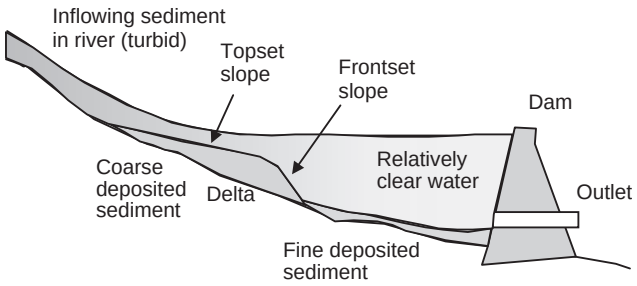


Figure 1 Delta formation resulting from deposited sediment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

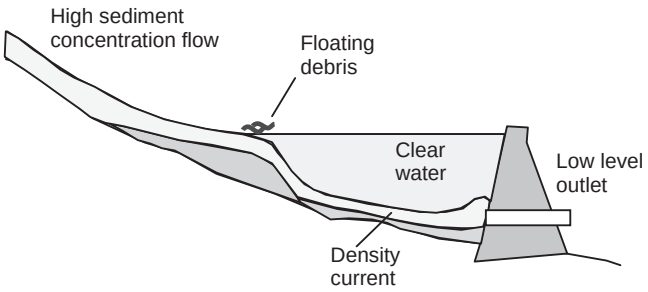


Figure 2 Density current flowing into a reservoir. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

periods of drought. It is, therefore, important to consider the worldwide reduction in water storage due to reservoir sedimentation, consider its impacts, and stress the importance of reservoir sustainability and reservoir sedimentation management.

WORLDWIDE STORAGE REDUCTION RESULTING FROM RESERVOIR SEDIMENTATION

A comparison between the reservoir storage added internationally each year due to construction of new dams and the volume lost due to sedimentation is shown in Figure 3. This figure shows that the incremental storage added annually has been reducing since about 1970 and that the cumulative storage lost to reservoir sedimentation exceeded the incremental addition of storage in about 1995. If this trend is maintained, the loss in reservoir storage will continue to exceed its addition, leading to a continued reduction in the reliability of water supply as the world population grows. A clear need for sustainable management of reservoirs and water resource infrastructure exists. Reservoirs are valuable community resources requiring sustainable management and protection.

The International Commission on Large Dams (ICOLD) (1998) estimates the total current storage volume contained in man-made reservoirs worldwide amounts to about

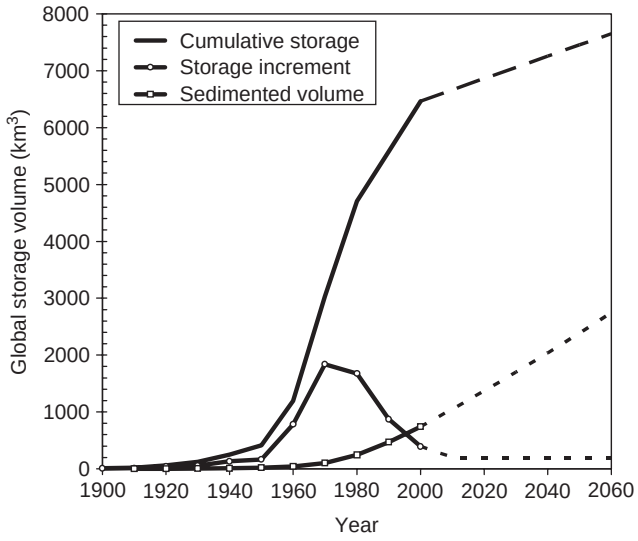


Figure 3 Cumulative reservoir storage constructed, incremental storage constructed, and storage lost due to reservoir sedimentation (Morris, 2003)

7000 km^3. This includes storage for water supply, irrigation, power generation, and flood protection. Estimates of average reservoir storage loss due to sediment deposition throughout the world vary, but are generally considered to range between 0.5 and 1% annually (White, 2001). This means that between 35 km^3 and 70 km^3 of reservoir storage is lost every year due to reservoir sedimentation. In order to maintain the reservoir storage that currently exists in the world, it is necessary to invest significant sums of money to develop new water supply infrastructure. Should the water supply infrastructure be replaced by other man-made reservoirs, it is estimated that the investment required to replace the lost storage capacity would range between US$ 10 billion and US$ 20 billion annually (2003 value), a substantial investment. This estimate of required investment only covers replacement of lost storage, and does not include creation of additional storage to maintain and improve reliable water supply to a growing world population.

The rate of reservoir sedimentation varies internationally, and is dependent on regional factors. The largest losses in storage due to reservoir sedimentation generally occur in China, followed by North America, the rest of Asia and the Pacific Rim, Europe, the Middle East, Africa, and South America (Figure 4).

IMPACTS

The consequences resulting from reservoir sedimentation are varied and include reduction in the reliability of water supply and impacts on the environment, infrastructure,

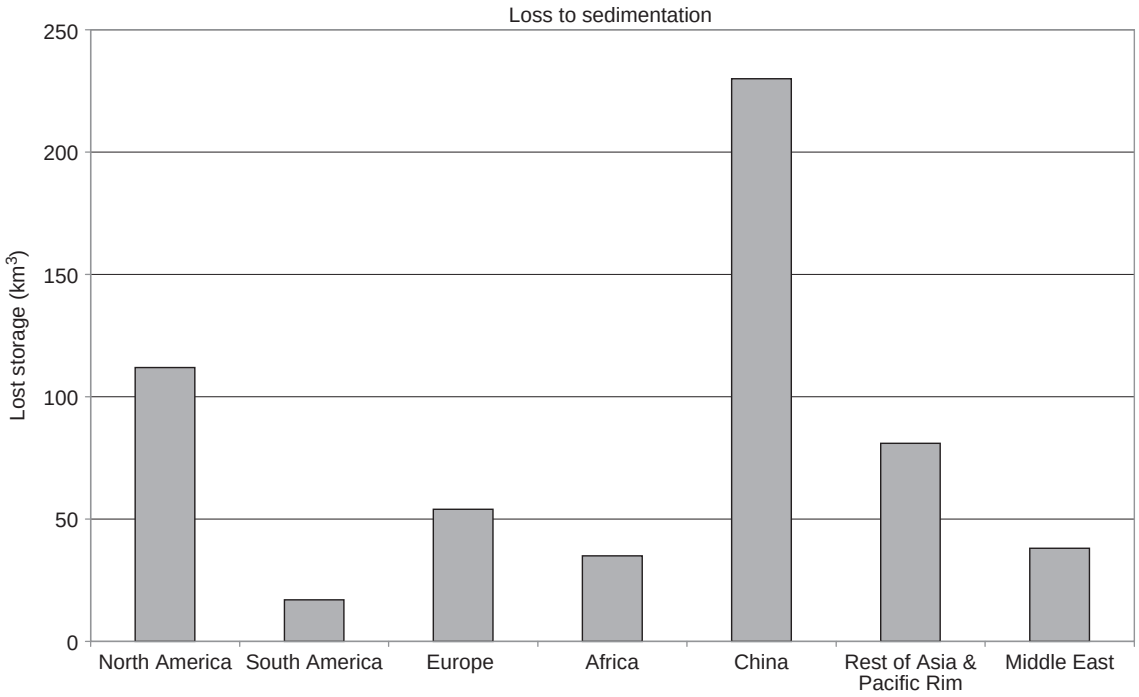


Figure 4 Reservoir storage lost due to sedimentation (data source: White, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

economy, life, and property. Although reservoir sedimentation leads to multiple negative impacts, it can also, in some cases, lead to more favorable environmental, social, and economic conditions as indicated further on in this article.

Impact on Water Supply Reliability

The effects of reservoir sedimentation on the reliability of water supply are not obvious during the early stages of reservoir sedimentation. Figures 5 and 6 relate the reliability of water supply to reservoir size and draft from a reservoir. The reservoir size is expressed in dimensionless form as the ratio between reservoir volume and the mean annual runoff (MAR), and dimensionless draft is expressed as the ratio between draft and MAR. Draft in this article is defined as the volume of water that a user wishes to withdraw from a reservoir. Figure 5 illustrates how the reliability of water supply varies with draft and reservoir volume for a river with an annual coefficient of variation (C_v) of 0.3 flowing into the reservoir, and Figure 6 shows it for an annual coefficient of variation of 1.2. C_v is a dimensionless variable that is defined as the standard deviation divided by the mean. Therefore, the annual coefficient of variation for river flow is the standard deviation of annual flows divided by the mean annual flow in a river. Small values of C_v (say 0.3) are indicative of river flows that do not vary much from year to year. A value of 1.2 is characteristic of annual river flows that can

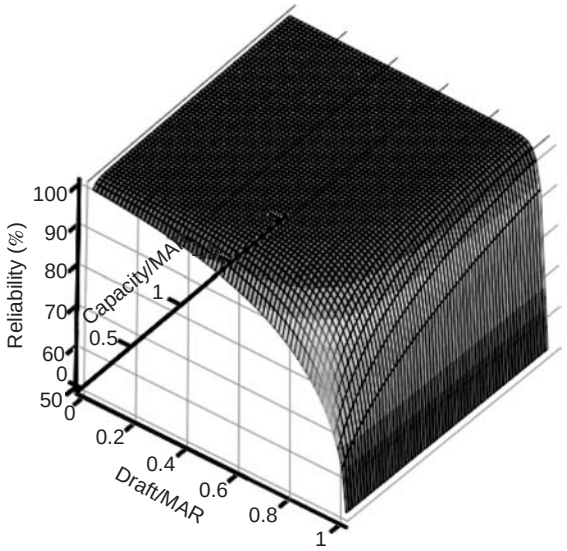


Figure 5 Reliability of water supply as a function of dimensionless reservoir capacity and dimensionless draft for an annual coefficient of variation of inflow equal to 0.3 (temperate regions). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

vary quite considerably on an annual basis, and represent flow in rivers located in arid and semiarid regions.

These two figures indicate the sensitivity of water supply reliability to storage capacity (volume) and draft, and

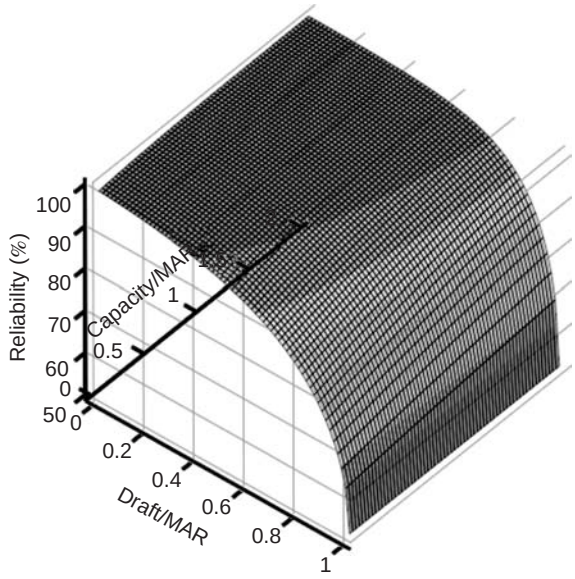


Figure 6 Reliability of water supply as a function of dimensionless reservoir capacity and draft for a coefficient of variation of inflow equal to 1.2 (arid/semiarid regions). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

provide insight into the impact of reservoir sedimentation on the reliability of water supply to growing communities. Such impacts are more noticeable in fast-growing communities located in arid and semiarid regions. The trend in both cases indicates a sudden drop in water supply reliability as the dimensionless draft increases to values that are on the order of 90% and higher. However, the change is more abrupt when the draft becomes very high in the case of a reservoir in a temperate climate (Figure 5) than what it is in the case of one in a semiarid region (Figure 6). For example, in the case of storage reduction due to reservoir sedimentation in temperate climates (Figure 5), if the reservoir capacity reduces by 50% by changing from twice the MAR to one time the MAR, and the draft concurrently changes from 50% of the MAR to, say, 80% of the MAR, the change in reliability of water supply is insignificant (it changes from a theoretical value of close to 100% to 99.9%). However, in the case of a reservoir in a semiarid region (Figure 6), the same change in storage due to reservoir sedimentation and the same growth in water demand will have a significant impact on the reliability of water supply (it changes from 98.7% to 88.5%).

When assessing the impact of reservoir sedimentation, it is therefore important to take account of the hydrologic characteristics of inflow, specifically its mean and variability (expressed in terms of C_v), and the changes in capacity and draft, which can be expressed in terms of the MAR. A set percentage reduction in storage due to reservoir sedimentation combined with a concurrent set increase in percentage draft does not have the same impact on reliability of water supply in temperate and semiarid regions. The adverse impact of storage loss due to reservoir sedimentation on water supply reliability in semiarid regions with high sediment yield and fast-growing populations is greater than in temperate regions.

Environmental Impacts

The environmental impacts of reservoir sedimentation, either positive or negative, occur within upstream and downstream of reservoirs. The delta region of sedimentation in a reservoir (in the upstream reaches) can change the morphology of the inflowing rivers, leading to marshy environments.

Within a reservoir, continued sedimentation can lead to changes in the properties of the lake bed, which in turn leads to changes in the food chain and reproductive environments for fish and other aquatic species. However, the environmental impacts from reservoir sedimentation are often most noticeable in the river reaches downstream of dams.

As sediment deposits in a reservoir upstream of a dam, the amount of sediment that is available for transport in the river reach downstream of the dam is reduced. The water discharged downstream of the dam, therefore, has a greater ability to entrain sediment from the riverbed and banks, which generally leads to degradation of downstream river reaches. In cases where the riverbed material is easily eroded, such degradation can lead to deepening of the river and general degradation of the riverbed and banks. In cases where the river bed contains gravel, cobbles, and boulders that are resistant to the erosive power of the discharged water, the river often widens as the preferred source of sediment shifts to the river banks, which might be more vulnerable to erosion.

Additionally, degradation of riverbeds downstream of dams often leads to deterioration in spawning habitat for fish. For example, some fish like trout require pea-sized gravel to lay their eggs. Reservoir sedimentation results in a reduction in the amount of pea-sized gravel in the river reach downstream of a dam, and consequent poorer conditions for spawning. However, some consequences of reservoir sedimentation can have a positive impact on the environment.

Examples of positive effects of reservoir sedimentation on the environment include the effects on Chesapeake Bay and the positive effect on trout fishing downstream of dams in the upper portions of the South Platte River in Colorado. In the case of Chesapeake Bay, reservoir sedimentation in upstream reservoirs prevents sedimentation of the bay. Trout fishing streams in Colorado are ranked, with Gold Medal trout fishing streams considered prime fishing spots. The Gold Medal trout fishing streams in the upper reaches of the South Platte River are generally located downstream

of dams. The reason for this is that the sediment generated in these Rocky Mountain catchments are generally very young, granitic sediments with rough sharp edges. By depositing these sediments upstream of the dams by the process of reservoir sedimentation provides improvement to the habitat for fish downstream of dams.

Hydropower

Hydropower plants are generally used to provide peaking power, contrary to thermal power plants that are more effective in providing base load power. Peaking power demand is the power that develops in a power grid due to sudden increases in the need for power. Hydropower plants have the flexibility to provide in this sudden increase in power demand. However, some hydropower plants do provide base load, which is the power that is used on average in a power grid.

Hydropower plants that mainly supply base load require storage to ensure enough water for long term power generation and can be adversely affected by storage loss due to reservoir sedimentation. This is not necessarily the case for power plants that are used to generate peaking power. In the latter case, less water is usually needed, and the impact of storage loss is less important.

However, sedimentation in reservoirs could result in sediment being discharged through the turbines should the deposited sediment move closer to the forebay of the hydroelectric plant. When sediment is discharged through turbines, it leads to erosion of the turbine blades, and deterioration.

Recreational Impacts

Reservoir sedimentation adversely impacts boating and fishing on lakes, but can have a positive impact on other sports, like duck hunting. As the volume of sediment deposited in a reservoir or lake increases, it decreases its depth, which can adversely affect boating. The changes in bed sediments can also lead to changes in the aquatic habitat of lakes, which affects the food chain and the type and abundance of fish.

Flood Management

The impact of reservoir sedimentation on flood management is twofold. If a reservoir is specifically constructed to reduce flood peaks downstream of the dam, reservoir sedimentation adversely impacts the effectiveness of such measures once the reservoir volume has been significantly reduced.

Reservoir sedimentation can also adversely impact flooding in reaches upstream of the reservoir. During delta formation, some of the sediment is deposited upstream of the normal pool level in a reservoir. This changes the fluvial geomorphology of the river in the reaches upstream of the dam, which generally results in higher flood levels than what have been experienced prior to reservoir sedimentation.

Infrastructure

Infrastructure built in rivers, like intakes, and bridges, are affected by reservoir sedimentation. Intakes are adversely affected in river reaches upstream of reservoirs when the river aggrades due to reservoir sedimentation. When the riverbed aggrades enough, it is possible for the sediment to block the intakes and prevent water from flowing into it. Degradation of the riverbed and banks downstream of a dam can also adversely affect intakes. As the riverbed elevation decreases, intakes on the river banks can be left high and dry.

In the case of bridges, the effects are different on the upstream and downstream sides of the reservoir and dam. On the upstream side of a reservoir subject to sedimentation, aggradation of the riverbed can lead to a reduction in the size of the bridge opening, which in turn could lead to the bridge being overtopped by flood waters on a more frequent basis than prior to the impacts of reservoir sedimentation taking effect. On the downstream side of a dam, degradation of the riverbed can in turn lead to excessive scour at bridges. As the riverbed elevation decreases, the foundations of bridge piers and abutments can be undermined and lead to failure.

Economic

The economic impacts of reservoir sedimentation can be significant. As reservoir sedimentation reduces the reliability of water and power supply and impacts the effectiveness of flood management, it affects public health, economic production, property value, and the general standard of living. For example, the value of property on lakeshores is adversely impacted if sedimentation moves the shoreline away from the property and results in aquatic vegetation taking up the water front. Property values are also affected if floods occur on a more regular basis due to the effects of reservoir sedimentation, than what has previously been experienced. When the reliability of water and power supply reduces, the effectiveness of industrial production is reduced, which can have significant effects on the economy of a region if economic multiplier effects are taken into account.

PREDICTION OF RESERVOIR SEDIMENTATION

Good planning of water resource development requires, amongst other things, the ability to predict the anticipated

effects of reservoir sedimentation. This is done by estimating the sediment yield from a river and estimating the amount of this sediment that will be deposited in a reservoir. Once it is known how much sediment could potentially deposit in a planned reservoir, it is then possible to assess its impacts and decide how it could be managed.

Estimating sediment yield for catchments is presented in other articles in this encyclopedia and is not repeated here. However, once it is known how much sediment could flow into a reservoir, the amount that could deposit can be estimated by making use of empirical relationships like the Brune (1953) and Churchill (1948) curves. The Brune curve (Figure 7) relates the percentage of sediment that is likely to be retained in a reservoir to the ratio of the reservoir capacity to the MAR. This figure indicates that small reservoirs (in terms of their capacity relative to the MAR) can trap large proportions of the sediment flowing into them. For example, the graph indicates that approximately 80 to close to 100% of the sediment flowing into a reservoir can be trapped if the reservoir capacity is only about 10% as large as the MAR.

The version of the Churchill curve presented in Figure 8 was modified by Roberts (1982) to relate the percentage of sediment that can pass through a reservoir to a dimensionless index, known as the "*sedimentation index*". The sedimentation index is a function of the reservoir capacity, discharge, and reservoir length, and is made dimensionless by multiplying it with the acceleration due to gravity (g). As long as the units that are used when inserting the different variables into the equation are consistent, the index will be dimensionless and can be used to enter the graph on the abscissa. Once it is known that reservoir sedimentation could be a potential problem in the long-term operation of a reservoir, management techniques should be considered.

DENSITY OF DEPOSITED SEDIMENT

As sediment is deposited in reservoirs, its density depends on the type of sediment. In the case of coarse, noncohesive sediment, the water between the individual sediment particles escape relatively easily as the sediment deposits and the grains essentially touch each other. As additional sediment is deposited on top of that already in place, the increase in sediment density over time is not that significant. However, if silt- and clay-sized material is deposited, it initially forms a very porous mass with a significant volume of water within its matrix. The initial dry density of deposited sediment consisting principally of silt and clay is often lower and at times only slightly higher than that of water. Typically, the initial dry density of deposited fine sediment can range from 950 kg m^{-3} to 1100 kg m^{-3}. After about 50 years, the density of such fine deposits gradually increases up to a maximum value of about 1300 kg m^{-3}. An empirical method to predict the density of deposited sediment and how it might change with time can be found in Lane and Koelzer (1943).

RESERVOIR SEDIMENTATION MANAGEMENT

Reservoirs are valuable community resources and should be managed as such by making use of a life-cycle management approach (Palmieri *et al.*, 2003). The life-cycle approach to sustainable development and management of reservoirs

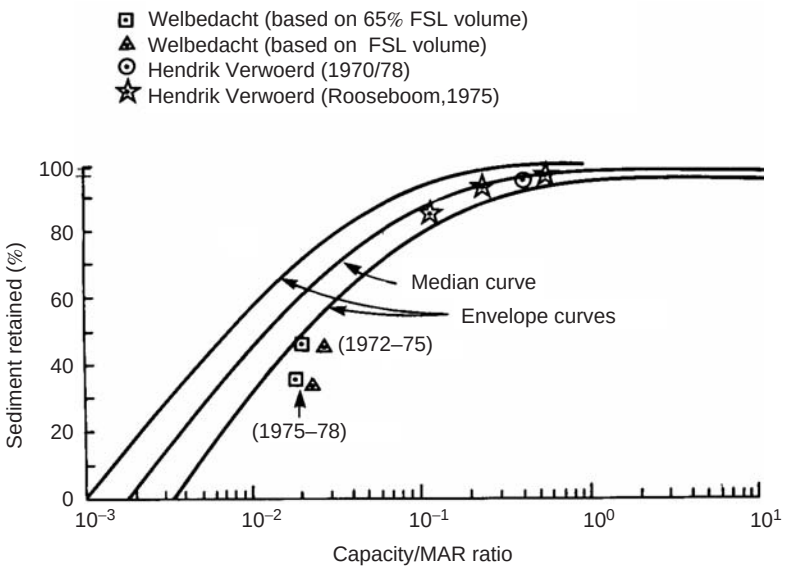


Figure 7 Brune (1953) Curve for determining sediment trap efficiency of reservoirs

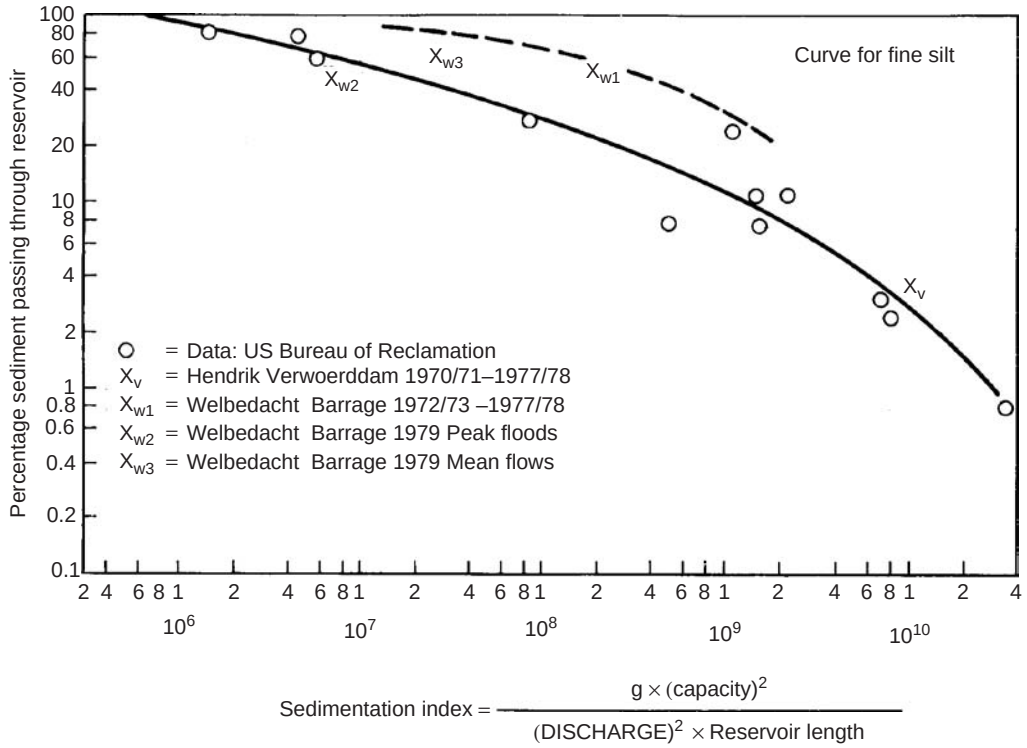


Figure 8 Churchill (1948) curve, as modified by Roberts (1982) for determining sediment trap efficiency of reservoirs

was developed because of the unique nature of these facilities. Reservoirs, contrary to most other infrastructure, are not easily refurbished once they have lost their utility due to storage loss resulting from reservoir sedimentation. Whereas other infrastructure, such as roads or buildings, can be refurbished once they have reached the end of their design life, making their continued use possible, the utility of reservoirs subject to sedimentation can generally not be recovered as easily. The approach to designing, constructing, and operating reservoirs can, therefore, not follow the conventional "design life" approach generally implemented by engineers, but adoption of a life-cycle management approach is required.

There is a marked difference between the conventional design life approach that pays scant attention to sustainability and regards environmental effects as residuals that may be recognized and are not accounted for as project liabilities, and the life-cycle management approach that focuses on sustainability. The process of life-cycle management contains many of the same elements as the design life approach but arranged in a circular fashion, indicating perpetual use of the infrastructure. Consequently, the opportunity exists to incorporate changing environmental and societal concerns that are often associated with direct impacts of a dam. Operation and maintenance, as well as sediment management, are conducted in a way that will encourage sustainable use. Figure 9 shows the differences between the design life approach and the life-cycle management approach.

Principal differences between the two approaches are that the design life approach is linear, and no allowance is made for care of the dam and reservoir at the end of its "life", with residual concerns pertaining to a variety of safety, social, and environmental issues remaining. On the other hand, the life-cycle management approach, with its circular progression, is principally aimed at continued, indefinite use of the facility with decommissioning only an option if no other management options for perpetual use are available. Economic analysis, which incorporates the cost of dam decommissioning, usually indicates that it is more economical to manage the dam and reservoir in a sustainable manner than to abandon it by decommissioning once it is completely filled with sediment (Palmieri *et al.*, 2003).

In order to implement the life-cycle management approach, it is preferred to consider operation and maintenance issues for sustainable management and use during the planning and design phases of the facility. This ensures that sediment management can be implemented in the most economical way. In existing facilities, where such provisions were not made during the original planning, design, and construction of the dam and reservoir, it is often possible to make modifications to the facility to implement sediment management. Several techniques are available to remove or prevent sediment from accumulating

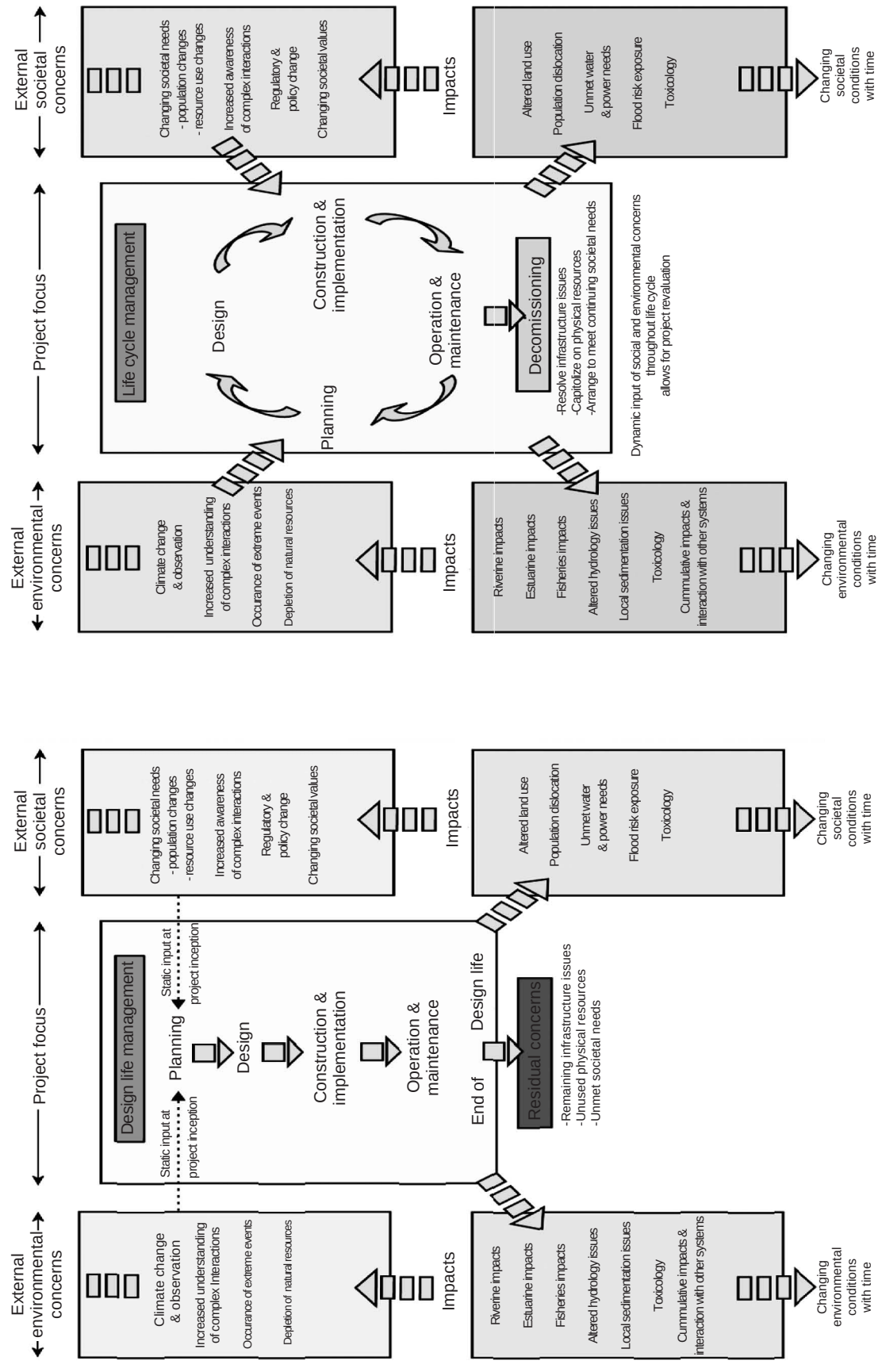


Figure 9 Comparison of the design life and life-cycle management approaches (Palmieri *et al.*, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

in a reservoir (Morris and Fan, 1997). The ones commonly used are: (i) flushing, (ii) sluicing, (iii) density current venting, (iv) mechanical removal (traditional dredging, hydrosuction removal system (HSRS), trucking) and bypassing. Catchment management to reduce sediment yield is an ideal that should be considered.

Flushing

Flushing is a technique whereby the flow velocities in a reservoir are increased to such an extent that deposited sediments are remobilized and transported through low-level outlets in the dam. *Sluicing* and *density current* venting also use the flow of the water, but their purpose is to pass a substantial portion of the incoming sediment load through the reservoir before the sediment particles can settle.

Dredging

Traditional dredging mechanically removes sediment from under the water by making use of suction pumps and pipelines. Disposal may be downstream in the river or at an off-site location.

Hydro-suction

The *HSRS* is a variation of conventional suction pipe dredging that uses the hydraulic head available at the dam as the energy for dredging instead of pumps powered by electricity or diesel (Hotchkiss and Huang, 1995). As such, where there is sufficient head available and the reservoir length is relatively short (usually not more than about 3000 m), the operating costs are substantially lower than those of traditional dredging.

Dry Excavation

Trucking (also *known as dry excavation*) requires the lowering of the reservoir water level during the dry season when the reduced river flows can be adequately controlled without interference with the excavation works. The sediments are excavated and transported for disposal using traditional earth moving equipment.

Bypassing

Bypassing is a technique that is used to prevent, or at least reduce, the amount of sediment entering a reservoir. This can be accomplished by constructing off-channel storage reservoirs or by creating channels or tunnels that bypasses sediment flows around a reservoir under high flow conditions.

FURTHER READING

Abdelhadi M.L. (1995) Environmental and socio-economic impacts of erosion and sedimentation in North African countries. *6th International Symposium on River Sedimentation*, Central Board of Irrigation and Power: New Delhi, pp. 1141–1152.

Annandale G.W. (1987) *Reservoir Sedimentation*, Elsevier Science Publishers: New York.

Annandale G.W. (1996) Prediction of sediment distribution in a dry reservoir: a stochastic approach. *Proceedings of the 6th Federal Interagency Sedimentation Conference*, Las Vegas, pp. 1–85–1–92.

Axelsson V. (1992) X-ray and radiographic analysis of sediment cores from the Cachi reservoir. In *Sedimentological Studies in the Cachi Reservoir, Costa Rica*, Jansson M.B. and Rodriguez A. (Eds.), UNGI Report No. 81, Department of Physical Geography, UPPSALA University: Sweden.

Bechtel Construction, Incorporation (1987) *Evaluation of the Effect of Reducted Spill and Recommended Flushing Flows below Rock Creek and Cresta Dams in the North Fork of the Feather River*, Report to Pacific Gas and Electric Company: San Francisco.

Bechtel Construction, Incorporation (1990) *Flushing Flow Evaluation: The North Fork of the Feather River below Poe Dam*, Report to Pacific Gas and Electric Company: San Francisco.

Bell H.S. (1942) *Stratified Flows in Reservoirs and its Use in Preventing Silting*, USDA, Miscellaneous Publication No. 491, USGPO: Washington.

Blanton J.O. (1982) *Procedures for Monitoring Reservoir Sedimentation*, US Bureau of Reclamation: Denver.

Borland W.M. (1971) Reservoir Sedimentation. In *River Mechanics*, Chap. 29, Shen H.W. (Ed.), Water Resources Publications: Fort Collins.

Borland W.M. and Miller C.R. (1958) Distribution of sediment in large reservoirs. *Journal of Hydraulics Division, ASCE*, **84**(HY2), pp. 1–18.

Bouchard J.P. (1995) Mud investigations and management, simulation of reservoir sediments. *Proceedings of the 6th International Symposium on River Sedimentation*, Central Board of Irrigation and Power: New Delhi.

Bouchard J.P., Cordelle M., Labadie G. and Lorin J. (1989) Numerical simulation of mud erosion in reservoirs by floods, application to reservoirs in the Durance river. *Technical Session B*. IAHR, 23rd Congress, Ottawa.

Bu N., Su F. and Shang R. (1980) Sediment problems in Liujiaxia and Yangouxia reservoirs, *Proceedings of International Symposium River Sedimentation*. Beijing, pp. 737–752.

Bukler J.H., Skelly T.M., Luepke M.J. and Wilken G.A. (1988) Case study: Teh lake springfield sediment removal project. *Lake and Reservoir Management*, **4**(1), 143–152.

Bureau of Reclamation (1948) *Lake Mead Density Current Investigations, 1937–1948*, Boulder City.

Burns M. and McArthur R. (1996) Sediment deposition in Jennings Randolph reservoir, Maryland and West Virginia, *Proceedings of the 6th Federal Interagency Sedimentation Conference*. Las Vegas, pp. 10.16–10.21.

Chang H.H. (1996) Reservoir erosion and sedimentation for model calibration. *6th Federal Interagency Sedimentation Conference*, Las Vegas.

Chang H.H., Harrison L.L., Lee W. and Tu S. (1995) Fluvial modeling for sediment-pass-through operations of reservoirs. *Proceedings of Hydraulics Division Meeting, San Antonio*, ASCE: New York.

Chen J. and Zhao K. (1992) Sediment management in Nanqin reservoir. *International Journal of Sediment Research*, **7**(3), 71–84.

Churchill M.A. (1948) Discussion of "Analysis and Use of Reservoir Sedimentation Data". by L.G. Gottschalk, *Proceedings of the Federal Interagency Sedimentation Conf*. Denver, Colorado, pp. 139–140.

Collier M., Webb R.H. and Schmidt J.C. (1995) *Dams and Rivers: A Primer on the Downstream Effects of Dams*, Circular 1126, USGS: Denver.

Crowder B.M. (1987) Economic costs of reservoir sedimentation: a regional approach to estimating cropland erosion damages. *Journal of Soil and Water Conservation*, **42**(3), 194–197.

Dendy F.E., Allen P.B. and Piest R.F. (1979) Sedimentation. *Field Manual for Research in Agricultural Hydrology*, *Agricultural Handbook, No. 224*, Chap. 4, USDA: Superintendent of Documents, Washington.

Dendy F.W., Champion W.A. and Wilson R.B. (1973) Reservoir sedimentation surveys in the United States. In *Man-Made Lakes: their Problems and Environmental Effects*, *Geophysical Monograph No. 17*, Ackermann W.C., White G.F. and Worthington E.B. (Eds.), American Geophysical Union: Washington.

Du G. and Zhang Z. (1989) *The Erosion of Cohesive Sediments in Reservoirs*, Selected Scientific Papers, I.W.H.R., 29, pp. 152–163.

Duguennois, H. (1956) *New Methods of Sediment Control in Reservoirs*, Water Power, May, pp. 174–180.

Eakin H.M. and Brown C.B. (1939) *Silting of Reservoirs*, Technical Bulletin No. 524, USDA: Washington.

Eftekharzadeh S. and Laursen E.M. (1990) A new method for removing sediment from reservoirs. *Hydro Review*, **9**(1), 80–84.

Fan J. (1985) Methods of preserving reservoir capacity. In *Methods of Computing Sedimentation in Lakes and Reservoirs*, Bruk S. (Ed.), UNESCO: Paris, pp. 65–164.

Fan J. (1986) Turbid density currents in reservoirs. *Water International*, **11**(3), 107–116.

Fan J. (1991) Density currents in reservoirs. *Workshop on Management of Reservoir Sedimentation*, New Delhi.

Fan J. (1995) An overview of preserving reservoir storage capacity. *Proceedings 1995 International Workshop on Reservoir Sedimentation at San Francisco*, Federal Regulatory Commission: Washington.

Fan J. and Jiang R. (1980) On methods for the desiltation of reservoirs, Com. 15. *International Seminar of Experts on Reservoir Desiltation*, Tunis.

Fan J. and Morris G.L. (1992a) Reservoir sedimentation I: delta and density current deposits. *Journal of Hydraulic Engineering-ASCE*, **118**(3), 354–369.

Fan J. and Morris G.L. (1992b) Reservoir sedimentation II: reservoir desiltation and longterm storage capacity. *Journal of Hydraulic Engineering-ASCE*, **118**(3), 370–384.

Fang D. and Cao S. (1996) An experimental study on scour funnel in front of a sediment flushing outlet of a reservoir, *Proceedings of 6th Federal Interagency Sedimentation Conference*, Las Vegas, pp. 1.78–1.84.

Ford D.E. (1990) Reservoir transport processes. In *Reservoir Limnology*, Thornton K.W., Kimmel B.L. and Payne F.E. (Eds.), John Wiley & Sons: New York.

Ford D.E. and Johnson M.C. (1983) *An Assessment of Reservoir Density Currents and Inflow Processes*, Technical Report E-83-7, US Army Corps of Engineers, Vicksburg.

Frenette M. and Julien P.Y. (1986) Advances in predicting reservoir sedimentation. *Proceedings of the 3rd International Symposium River Sedimentation*, University of Mississippi: Oxford.

Gogus M. and Yalcinkaya F. (1992) Reservoir sedimentation in Turkey. *5th International Symposium River Sedimentation*, Karlsruhe.

Harrison L.L., Lee W.H. and Tu S. (1995) Sediment pass-through, an alternative to reservoir dredging. *Waterpower '95*, ASCE: New York, pp. 2236–2245.

Hesse L.W. and Newcomb B.A. (1982) Effects of flushing spencer hydro on water quality, fish and insect Fauna in the Niobrara River, Nebraska, North American. *Journal of Fisheries Management*, **2**, 45–52.

Hotchkiss R.H. and Xi H. (1995) Designing a hydrosuction sediment removal system. *6th International Symposium river sedimentation, Central Board of irrigation and power*, New Delhi, pp. 165–174.

Hwang J.S. (1985) The study and planning of reservoir desilting in Taiwan. *Water International*, **10**(1), 7–13.

James W.F. and Kennedy R.H. (1987) Patterns of sedimentation in DeGray lake, Arkansas. In *DeGray Lake Symposium*, Kennedy R.H. and Nix J. (Eds.), US Army Corps of Engineers, Waterways Experiment Station: Vicksburg.

Jansson M.B. (1992a) Suspended sediment inflow to the Cachi reservoir, Costa Rico. In *Sedimentological Studies in the Cachi Reservoir, Costa Rico*, UNGI Report No. 81, Jansson M.B. and Rodriguez A. (Eds.), Department of Physical Geography, Uppsala, University: Sweden.

Jansson M.B. (1992b) Suspended sediment outflow from the Cachi reservoir during flushing in 1990. In *Sedimentological Studies in the Cachi Reservoir, Costa Rico*, UNGI Report No. 81, Jansson M.B. and Rodiguez A. (Eds.), Department of Physical Geography, Uppsala, University: Sweden.

Jowett I. (1984) Sedimentation in New Zealand hydroelectric schemes. *Water International*, **9**, 172–176.

Kereselidze N.B., Kutavaya V.I. and Tsagareli Y.A., (1985) Silting and flushing mountain reservoirs, exemplified by the Rioni series of *hydroelectric stations*, *Hydrotechnical Construction (English translation of Gidrotekhnicheskie Stroitel'stvo)* **19**(9), Sept. 1985, pp. 514–520.

Lai J.S. and Shen H.W. (1996) Flushing sediment through reservoirs. *Journal of Hydraulic Research*, **34**(2), 237–255.

Lara J.M. (1962) *Revision of Procedures to Compute sediment Distribution in Large Reservoirs*, US Bureau of Reclamation: Denver.

Lara J.M. (1973) A unique sediment depositional pattern. *Man-Made Lakes: their Problems and Environmental Effects*, *Geophysical Monograph 17* American Geophysical Union: Washington.

Lara J.M. and Pemberton E.L. (1963) Initial unit weight of deposited sediment. *Proceedings of Federal Interagency Sedimentation Conference*, USDA-ARS Miscellaneous Publication No. 970, USDA-ARS: pp. 818–845.

Ligon F.K., Dietrich W.E. and Trush W.J. (1995) Downstream ecological effects of dams. *Bioscience*, **45**(3), 183–192.

Lijklema L., Aalderink R.H., Blom G. and AnDuin eH.S. (1994) Sediment transport in shallow lakes: two case studies related to eutrophication. In *Transport and Transformation of Contaminants near the Sediment-Water Interface*, DePinto J.V., Lick W. and Paul J.F. (Eds.), Lewis Publishers: Boca Raton.

Lin B., Dou G., Xie J., Dai D., Chen J., Tang R. and Zhang R. (1989) On some key sedimentation problems of Three Gorges Project (TGP). *International Journal of Sediment Research*, **4**(1), 57–74.

Lin B., Dou G., Xie J., Dai D., Chen J., Tang R. and Zhang R. (1993) On some sedimentation problems of Three Gorges Project (TGP) in the light of recent findings. In *Notes on Sediment Management in Reservoirs; National and International Perspectives*, Fan S.S. and Morris G.L. (Eds.), Federal Energy Regulatory Commission: Washington.

Long Y. and Zhang Q. (1981) Sediment regulation problems in sanmenxia reservoir. *Water Supply and Management*, **5**(4/5), 351–360.

Lu Y. (1995) Three gorges project on the Yangtze river. *Annual Meeting of ICOLD Executive Committee*, Oslo.

Mahmood K. (1987) *Reservoir Sedimentation: Impact, Extent, Mitigation*, World Bank Technical Report No. 71, Washington.

McHenry J.R. (1974) Reservoir sedimentation. *Water Resources Bulletin*, **10**(2), 329–337.

Meadowcroft I.C., Berress R. and Reeve C.E. (1992) Numerical modeling of reservoir sedimentation. In *Water Resources and Reservoir Engineering*, Parr N.M., Charles J.A. and Walker S. (Eds.), Thomas Telford: London.

Morris G.L. (1995a) Reservoirs and the sustainable development of water resources. *15th Annual USCOLD Lecture Series*, USCOLD: Denver.

Morris G.L. (1995b) Reservoir sedimentation and sustainable development in India: problem scope and remedial strategies. *6th International Symposium River Sedimentation*, Central Board of Irrigation and Power: New Delhi, pp. 53–61.

Morris G.L. (1996) Design analysis of sediment-excluding off-stream reservoir, Rio Fajardo, Puerto Rico. *6th Federal Interagency Sedimentation Conference*, Las Vegas, pp. V.78–V.84.

Morris G.L., Colón R., Laura R. and Anderson G.T. (1992) BRASS modeling of Loíza reservoir Puerto Rico, for sediment management operations. *Water Forum '92*, ASCE: New York.

Morris G.L. and Hu G. (1992) HEC-6 modeling of sediment management in Loiza reservoir, Puerto Rico. In *Hydraulic Engineering, Water Forum '92*, Jennings M. and Bhowmik N. (Eds.), ASCE: New York, pp. 630–635.

Nizery A. and Bonnin J. (1953) Observations systématiques de courants de densité dans une retenue hydro-electrique. *Proceedings of Minnesota International Hydraulics Convention*, pp. 369–386.

Noll J.J. (1953) *The Silting of Caonillas Reservoir, Puerto Rico*, USDA, Soil Conservation Service: Washington.

Northwest Institute of Hydraulic Research and Administrative Bureau of Yeyu River (1972) Preliminary experience of the Heisonglin reservoir management. *Proceedings of Symposium Reservoir Sediment Measurement and Research*, Sanmenxia, pp. 169–183.

Okada T. and Baba K. (1982) *Sediment Release Plan at Saluma Reservoir*, 14th ICOLD, Q.54, R.4., pp. 41–64.

Orvis C.J. (1989) *Elephant Butt Reservoir 1988 Sedimentation Survey*, Bureau of Reclamation: Denver.

Otsubo K. and Muraoka K. (1988) Critical shear stress of cohesive bottom sediments. *ASCE Journal of Hydraulic Engineering*, **114**(10), 1241–1256.

Pan X. (1990) *A Summary on Erosion and Sedimentation in Sanmenxia Reservoir and its Downstream Reaches on the Yellow River*, Yellow River Commission: Beijing.

Pansic N., Austin R.J. and Finis M. (1995) Sediment management for dam decommissioning. *15th Annual USCOLD Lecture Series*, USCOLD: Denver, pp. 29–46.

Parhami F. (1986) Sediment control methods in Sefid-rud reservoir. *Proceedings of the 3rd International Symposium River Sedimentation*, University of Mississippi: Oxford, pp. 1047–1055.

Partheniades E. (1962) *A Study of Erosion and Deposition of Cohesive Soils in Salt Water*, Ph.D. Dissertation, University of California, Berkeley.

Partheniades E. (1965) Erosion and deposition of cohesive soils. *Journal of Hydraulics Division ASCE*, **91**(HY1), 105–138.

Paul T.C. and Dhillon G.S. (1988) Sluice dimensioning for desilting reservoirs. *Water Power and Dam Construction*, **40**(5), May 1988, pp. 40–44.

Planning Associates (1991) *Rock Creek-Cresta Dredging Project*, :Plumas County EIR No. 50, Quincy.

PRASA (Puerto Rico Aqueduct and Sewer Authority) (1995) *Supplementary Preliminary Environment Impact Statement for the Carraizo Reservoir Dredging Project*, San Juan.

Qian N. (1982) Reservoir sedimentation and slope stability; technical and environmental effects. *Proceedings of 14th ICOLD Congress*, Río de Janeiro, 639–690.

Ramey M.P. and Beck S.M. (1990) *Flushing Flow Study, North Fork of the Feather River Below Poe Dam*, Pacific Gas and Electric Company: San Francisco.

Ramírez C. and Rodríguez A. (1992) History of the Cachí reservoir. In *Sedimentological Studies on the Cachí Reservoir, Costa Rica*, UNGI Report No. 81 Jansson M.B. and Rodríguez A. (Eds.), Department of Physical Geography, Uppsala University: Sweden.

Reiser D.W., Ramey M.P. Beck S., Lambert R. and Geary R.E. (1989) Flushing flow recommendations for maintenance of Salmonid spawning gravels in a steep, regulated stream. *Regulated Rivers: Research and Management*, **3**(1/4), 267–275.

Ritchie J.C., Cooper C.M. and McHenry R. (1986) Sediment accumulation rates in lakes and reservoirs in the Mississippi river

valley. *3rd International Symposium on River Sedimentation*, University of Mississippi: Oxford.

Rooseboom A. (1992) *Sediment Transport in Rivers and Reservoirs: A South African Perspective*, Report to Water Research Commission of South Africa, by SigmaBeta Consulting Engineers: Stellenbosch.

Sanmenxia Reservoir Hydrologic Experimental Station and IWHR (1962) *Report on Density Currents Observed During 1961 in Sanmenxia Reservoir*, Beijing.

Scheuerlein H. (1995) Downstream effects of dam construction and reservoir operation. *6th International Symposium on River Sedimentation*, Central Board of Irrigation and Power: New Delhi, pp. 1101–1108.

Shahin M.M.A. (1993) An overview of reservoir sedimentation in some African river basins. *Sediment Problems: Strategies for Monitoring, Prediction and Control*, IAHS Publication No. 217, IAHS: Wallingford, pp. 93–100.

Simons R.K. and Simons D.B. (1992) Sediment problems associated with dam removal. *ASCE National Hydraulics Conference*, Muskegon River, Nashville.

Stall J.B. and Lee M.T. (1980) Reservoir sedimentation and its causes in Illinois. *Water Resources Bulletin*, **16**(5), 874–880.

Strand R.I. and Pemberton E.L. (1987) Reservoir sedimentation. *Design of Small Dams*, US Bureau of Reclamation: Denver.

Summer W., Stritzinger W. and Zhang W. (1994) The impact of run-of-river hydropower plants on the temporal suspended sediment transport behavior. *Variability in Stream Erosion and Sediment Transport*, IAHS Publication No. 224, IAHS: Wallingford, pp. 411–419.

Sundborg A. and Jansson M.B. (1992) Present and future conditions of reservoir sedimentation. In *Sedimentological Studies on the Cachí Reservoir, Costa Rica*, UNGI Report No. 81, Jansson M.B. and Rodríguez A. (Eds.), Department of Physical Geography, Uppsala University: Sweden.

Sundborg A. (1992b) Sedimentation in the Cachí reservoir illustrated by mathematical modeling. In *Sedimentological Studies in the Cachí Reservoir, Costa Rica*, UNGI Report No. 81, Jansson M.B. and Rodríguez A. (Eds.), Department of Physical Geography, Uppsala University: Sweden.

Sundborg A. (1992a) Erosion processes in the Cachí reservoir during the flushing period in 1990. In *Sedimentological Studies in the Cachí Reservoir, Costa Rica*, UNGI Report No. 81 Jansson M.A. and Rodríguez A. (Eds.), Department of Physical Geography, Uppsala University: Sweden.

Tang R. and Lin W. (1987) A study on sedimentation problems of the Gezhouba project. *International Journal of Sediment Research*, **1**, 69–101.

Tejwani K.G. (1984) Reservoir sedimentation in India: its causes, control and future course action. *Water International*, **9**(4), 150–154.

Thevenin J. (1960) La sedimentation des barrages-reservoirs en Algerie et les moyens mis en oeuvre pour preserver les capacities. *Annales de l'Institute Technique du Batiment et des Travaux*, No 156.

Tolouie E. (1989) *Reservoir Sedimentation and De-siltation*, M.Phil. Thesis, University of Birmingham.

Tolouie E. (1993) *Reservoir Sedimentation and De-siltation*, Ph.D. Thesis, University of Birmingham.

Trimble S.W. (1981) Changes in sediment storage in coon Creek basin, Driftless area, Wisconsin, 1853–1975. *Science*, **214**(9), 181–183.

US Army Corps of Engineers (1989) *Sedimentation Investigations of Rivers and Reservoirs*, Engineering Manual 1110-2-4000, Washington.

Urlapov G.A. (1977) Irrigation reservoir that silts up insignificantly. *Soviet Hydrology Selected Papers*, **16**(3), 256–258.

Van Den Wall Bake G.W. (1986) The siltation and erosion survey in Zimbabwe: drainage basin sediment delivery. *Drainage Basin Sediment Delivery*, IAHS Publication No. 159, IAHS: Wallingford, pp. 69–80.

Varma C.V.J., Rao A.R.G. and Natarajan S., (1992) Sedimentation of Indian reservoirs. An Assessment. *Proceedings of the 5th International Symposium River Sedimentation*, Karlsruhe.

Wada A. (1995) Japan's experience in reservoir sediment management. *International Reservoir Sedimentation Workshop*, Federal Energy Regulatory Commission: San Francisco, Washington.

Wang W.C., Tsai C.T., Hsu S.K. and Hsieh C.D. (1995) Evaluation of alternatives for reservoir sediment removal: a case study. In *15th Annual USCOLD Lecture Series*, USCOLD: Denver, pp. 381–387.

Webb R.M.T. and Gómez- Gómez F. (1996) *Sedimentation Survey of Lago Dos Bocas, Puerto Rico, August 1994* , Water-Resources Investigations Report 95–4214, USGS, San Juan.

Webb R.M.T. and Soler-Lopez L.R. (1997) *Sedimentation History of Lago Loiza 1953–1994*, USGS: San Juan.

Xia M. (1984) Sediment regulation in reservoirs on heavily sediment-laden rivers. *Yellow River*, (5), 3–8.

Xia M. and Ren A. (1980) Methods of sluicing sediment from Heisonglin reservoir and its utilization downstream. *Proceedings of International Symposium River Sedimentation*, Vol. 2, Guanghua Press: Beijing.

Xia M. (1986) New desilting method used in Heisonglin reservoir: lateral erosion on floodplain. In *Collected Research Papers*, Vol. 2, Part 2, Northwest Institute of Hydrological Research: Yangling.

Xia M. (1989) Lateral erosion, a storage recovery technique for silted-up reservoirs. *Proceedings of the 4th International Symposium River Sedimentation*, Beijing, pp. 1143–1149.

Zhang Q. and Long Y. (1980) Sediment problems of Sanmenxia reservoir. *Proceedings of the International Symposium River Sedimentation*, Vol. 2, Guanghua Press: Beijing, pp. 707–716.

Zhou Z. (1993) Remarks on reservoir sedimentation in China. In *Notes on Sediment Management in Reservoirs: National and International Perspectives*, Fan S.S. and Morris G. (Eds.), Federal Energy Regulatory Commission: Washington, pp. 153–160.

REFERENCES

Brune G.M. (1953) Trap efficiency of reservoirs. *Transactions-American Geophysical Union*, **34**(3), 407–418.

Churchill M.A. (1948) Analysis and use of reservoir sedimentation data. *Proceedings of the Federal Interagency*

Sedimentation Conference, US Bureau of Reclamation: Denver.

Hotchkiss R.H. and Huang X. (1995) Hydrosuction sediment-removal systems (HSRS): principles and field test. *Journal of Hydraulic Engineering*, **121**, 479–489.

ICOLD (International Commission on Large Dams) (1998) *World Register of Dams*, CD-ROM.

Lane E.W. and Koelzer V.A. (1943) Density of sediments deposited in reservoirs. *A Study of Methods Used in Measurement and Analysis of Sediment Loads in Streams*, Report No. 9. Hydraulic Lab, University of Iowa.

Morris, G.L. (2003) Reservoir sedimentation management: worldwide status and prospects. *Proceedings of the Challenges to the Sedimentation Management for Reservoir Sustainability*, Third World Water Forum, Otsu.

Morris G.L. and Fan J. (1997) *Reservoir Sedimentation Handbook: Design and Management of Dams, Reservoirs and Watersheds for Sustainable Use*, McGraw Hill: New York.

Palmieri A., Shah F., Annandale G.W. and Dinar A. (2003) *Reservoir Conservation: The RESCON Approach*, The World Bank, 1818 H Street, Washington.

Roberts C.P.R. (1982) Quoted in Rooseboom, A, *Sediment Transport, HYDRO 82*, University of Pretoria: Pretoria.

White R. (2001) *Evacuation of Sediment from Reservoirs*, Thomas Telford: London.

89: On the Worldwide Riverine Transport of Sediment – Associated Contaminants to the Ocean

PETER W SWARZENSKI[1] AND PAMELA L CAMPBELL[2]

[1] *United States Geological Survey, St Petersburg, FL, US*
[2] *ETI, St Petersburg, FL, US*

The continuous and persistent weathering of the continents and the ensuing transport of both eroded and anthropogenic products by rivers account for almost all the dissolved solutes, particulates, and sediment-associated contaminants delivered to the ocean. Transported within this riverine load are essential nutrients (e.g. P, N, C), (oxy)hydroxides of Fe/Mn that act as highly efficient riverine scavengers, an ever-changing suite of man-made organic and inorganic constituents, as well as a wide variety of natural solutes (e.g. Si, Ca, Mg, Sr). The environmental behavior of these river-borne constituents is controlled by dynamic phase partitioning between operationally defined dissolved, colloidal, and particulate pools control. As most trace elements are bound onto river particulates, their fate is defined foremost by the nonconservative nature of such solids as they move along the particle-size continuum or are exchanged and remobilized in river bed and floodplain sediments. This paper describes the global riverine transport of sediment-associated contaminants to the ocean.

INTRODUCTION

One of the original scientific programs on worldwide river quality was initiated in the mid-1950s by the Association for Scientific Hydrology (Durum *et al.*, 1960). Livingston (1963) effectively mined these data compilations and created the first truly global-scale treatise on river chemistry. Subsequent to this work, large-scale geochemical research efforts were soon initiated for the first time on some major, "undisturbed" river systems, for example, the Amazon (Gibbs, 1973; Stallard, 1980), the Zaire-Congo (Eisma, 1978), and the Mekong rivers (Carbonnel and Meybeck, 1975) (Figure 1). These early investigations also spurred the creation and implementation of comprehensive river monitoring networks (i.e. US Geological Survey's NASQAN: Hooper *et al.*, 2001), as well as other multiyear river research programs (i.e. SCOPE-CARBON: Degens and Kempe, 1982). The worldwide freshwater quality monitoring network – GEMS-WATER – launched in the late 1970s by UNEP, WHO, UNESCO, and WMO established a global-scale database on the quality of rivers. From such reports, a reliable assessment of an entire continent's river water quality could be compiled and utilized for the first time. To validate such very large datasets, up-to-date internal guidelines on collection and analytical methods are not only essential, but also directly define such a program's success. For example, there has been an almost systematic decrease in "average" riverine trace-metal concentrations over the past 2 to 3 decades, as collection and analytical methods continually become refined (Horowitz *et al.*, 2001). It has been increasingly accepted that a comprehensive evaluation of river quality must also include a study of particulates and colloids as many trace elements, radionuclides, and organic compounds are strongly bound to such solid phases; hence their fate is tied directly to the nonconservative nature of these particulates (Trefry *et al.*, 1986) in river systems.

A river's combined load of suspended particulate matter and dissolved salts represents the single most important input to the world's ocean. This riverine load is the product of persistent chemical and physical weathering of the continents, as well as more recent anthropogenic

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

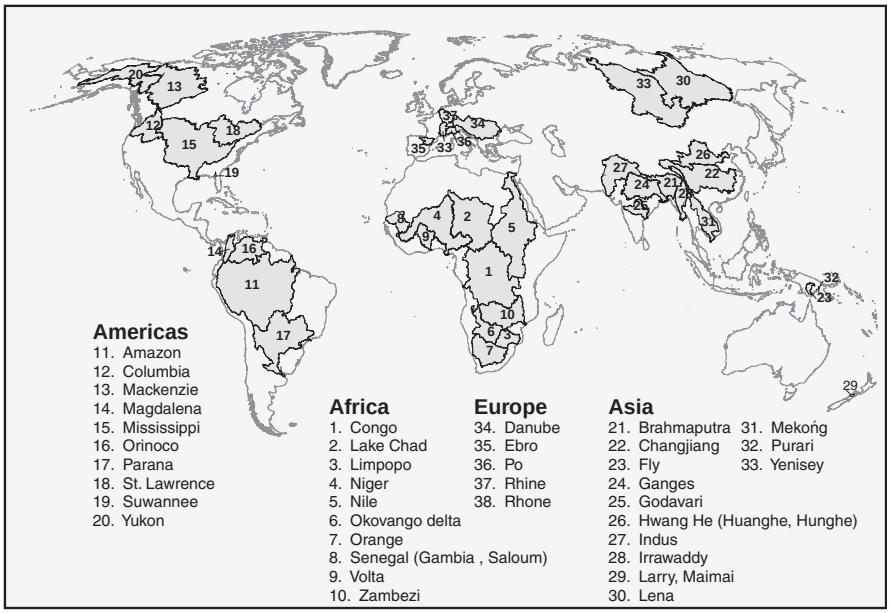


Figure 1 A map showing select watersheds and rivers cited in the text. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

additions in the form of organic and inorganic constituents. Natural weathering processes are modulated by complex geographic, climatic, geologic, and biologic forces, yet reliable long-term riverine datasets are beginning to provide at least a reliable first-order estimate of present and future trends in this fluvial input (White and Blum, 1995; *see* **Chapter 80, Erosion and Sediment Transport by Water on Hillslopes, Volume 2**). Being able to accurately predict future river trends and contaminant discharge estimates to the ocean is critical to form a reliable assessment of anthropogenic imprints on such riverine discharge cycles. This chapter summarizes the worldwide transport of riverine particulates, dissolved solids, and sediment-associated contaminants on a continent-to-continent basis. With the notable exception of just a few relatively unpolluted river systems (e.g. Amazon, Yukon), today's river loads are typically heavily impacted by man and require careful monitoring and management.

Dissolved Loads

Water discharge from a river system can be simply expressed as an integration of net precipitation, evapotranspiration, and bidirectional groundwater exchange within a watershed (Schlosser *et al.*, 2002; Vörösmarty and Fekete, 2002). The world's 10 largest rivers (Table 1) account for close to 40% of the total fluvial water discharge entering the ocean (Figure 2). For example, the Amazon River alone contributes close to 20% of the world's freshwater

Table 1 Average discharge of the top 10 major river systems to the oceans (after Meade, 1996)

Average suspended-sediment discharge (10^6 t y^{-1})[a]		Average water discharge (10^9 m^3 y^{-1})[b]	
Amazon	1200	Amazon	6300
Huanghe	1100	Zaire	1250
Ganges/Brahmaputra	1000	Orinoco	1200
Changjiang	480	Ganges/Brahmaputra	970
Irrawaddy	260	Changjiang	900
Magdalena	220	Yenisey	630
Mississippi	210	Mississippi	530
Godavari	170	Lena	510
Hunghe	160	Mekong	470
Mekong	160	Parana	470

[a]1990 estimates.
[b]1980 estimates.

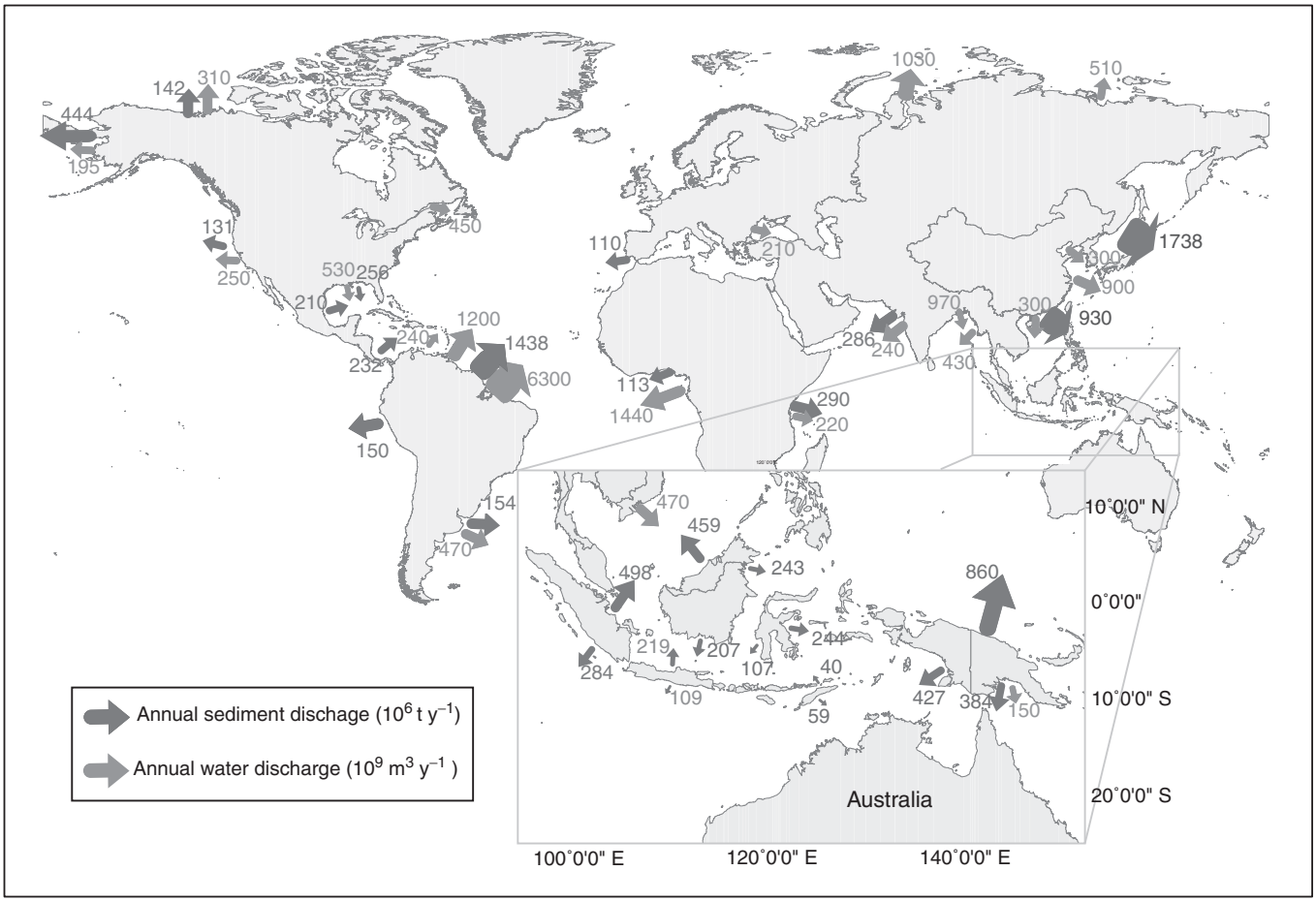


Figure 2 A compilation of annual water and sediment discharged from select large drainage basins of the world (after Milliman and Meade, 1983; Milliman and Syvitski, 1992; Milliman *et al.*, 1999). Arrow width is an approximation of relative discharge. Note that discharge numbers refer to average annual inputs. Arrows do not necessarily indicate the direction of flow. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

discharge, a value more than the combined total of the next seven largest rivers (Milliman and Meade, 1983; Meade, 1996). Material flux estimates derived from such tremendous discharge values are therefore inherently large and globally significant. Tropical regions with high rainfall, such as southern Asia, Oceania, and northern South America, contribute well over 60% of the global riverine freshwater discharge (Milliman *et al.*, 1999; Dai and Trenberth, 2002). The African continent, in contrast, discharges very little river water to the sea, with the notable exception of the Zaire–Congo and Niger Rivers.

While the geochemical composition of the ocean can be considered globally quite homogenous, continental surface waters are, however, a very unique expression of the particular terrain they traverse (Table 2). This riverine signature is defined by a combination of weathering processes, aeolian, and anthropogenic inputs, as well as the internal cycling of organic matter, and may include: (i) the prevalence of easily weathered minerals (e.g. calcite, gypsum, halite), (ii) short- and long-term climate oscillations (i.e. temperature and rainfall), (iii) the proximity to oceanic aerosols at the land-sea interface, and (iv) terrestrial primary productivity as a source of carbon and other nutrients. Global river water quality is, as a consequence, intrinsically complex and heterogeneous (Gibbs, 1977; White and Blum, 1995). In smaller watersheds (i.e. $<10\,km^2$), the concentrations of dissolved materials generally span 2–3 orders of magnitude, as opposed to 1–2 in larger watersheds (Shiller and Boyle, 1987). However, most of this variability can be found in just a few constituents, for example, sodium, chloride, sulfate, and total suspended matter (Meybeck and Helmer, 1989), which are also most affected by anthropogenic activities such as pollution (Berner and Berner, 1996).

Particulate Loads

Slightly less than half (~40%) of the global particulate load (Table 3) is carried by highly turbid rivers (total

Table 2 Average composition of the world's river systems, per continent[a] (after Holland, 1978)

Continent	HCO_3^-	SO_4^{2+}	Cl^-	NO_3^-	Ca^{2+}	Mg^{2+}	Na^+	K^+	Fe	SiO_2	Sum
North America	68	20	8	1	21	5	9	1.4	0.16	9	142
South America	31	4.8	4.9	0.7	7.2	1.5	4	2	1.4	11.9	69
Europe	95	24	6.9	3.7	31.1	5.6	5.4	1.7	0.8	7.5	182
Asia	79	8.4	8.7	0.7	18.4	5.6	9.3	nd	0.01	11.7	142
Africa	43	13.5	12.1	0.8	12.5	3.8	11	nd	1.3	23.2	121
Australia	31.6	2.6	10	0.05	3.9	2.7	2.9	1.4	0.3	3.9	59
World	58.4	11.2	7.8	1	15	4.1	6.3	2.3	0.67	13.1	120
anions[b]	0.96	0.23	0.22	0.02	nd	nd	nd	nd	Nd	nd	1.43
cations[b]	nd	nd	nd	nd	0.75	0.34	0.34	0.06	Nd	nd	1.43

nd = no data.
[a]Livingston (1963); concentration in ppm.
[b]Milli-equivalents of strongly ionized components.

Table 3 Estimated drainage basin area, sediment yield, and sediment discharge rates per continent (after Milliman and Meade, 1983). Note that Oceania describes large Pacific Islands; the desert regions (11.40 km^2) of northern Africa, the Saudi Arabian Peninsula and western Australia are assumed to contribute no annual sediment discharge

Continent	Drainage basin area (10^6 km^2)	Sediment yield ($t\,km^{-2}\,yr^{-1}$)	Sediment discharge (10^6 $t\,yr^{-1}$)
Total	183.43	1671	13 526
Oceania	3.00	1000	3000
Asia	16.88	380	6349
South America	17.90	97	1788
North and Central America	17.50	84	1462
Europe	4.61	50	230
Africa	15.34	35	530
Australia	2.20	28	62

suspended matter concentrations exceed 1500 $mg\,L^{-1}$), yet these rivers represent only 2% of the world's freshwater discharge (Meybeck, 1982). These very turbid rivers (i.e. Huanghe, Orange, Indus) are located principally in arid and semiarid regions. Rivers that drain mountainous regions (i.e. Mackenzie, Irrawaddy, Mekong, Ganges-Brahmaputra, Changjiang) carry an additional ~40% of the particulate load to the sea. The slow-flowing lowland rivers (i.e. Amazon, Zambezi, Paraná, Zaire-Congo) carry the remaining 20% of riverine particulates to the ocean, but constitute more than 50% of the global riverine water discharge (Meade, 1996).

Much has been written on the intricate relation among continental erosion, sediment yield, climate, and topography (White and Blum, 1995). For example, in the late 1950s, Langebein and Schumm (1958) established a positive correlation between sediment yield in many smaller watersheds and effective rainfall. However, as can be expected, such a simple relation between these two parameters cannot be maintained on a worldwide basis. Global erosion rates and sediment yields clearly depend on many interwoven factors such as, total precipitation, the seasonality of this rainfall, temperature, vegetative cover, topography, and the nature of surficial sediments and underlying bedrock (*see* **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2**). These watershed characteristics often control the riverine transport of solutes and sediment-associated contaminants (Shiller and Boyle, 1987; Chabaux *et al.*, 2003). In their comprehensive review, Milliman and Meade (1983) report that as drainage basin areas increase by an order of magnitude, sediment yield typically decreases sevenfold. This global negative correlation between drainage basin size and sediment yield reflects the inability of smaller, higher energy drainage basins to store sediments in their lower reaches as they discharge to the ocean (Meade *et al.*, 1985). As a consequence, it recently has been shown that the input from these smaller mountainous streams and rivers to the sea may have a much greater influence on the worldwide delivery of riverine particulates to the ocean than previously thought (Milliman and Meade, 1983; Milliman *et al.*, 1999; Holmes *et al.*, 2002; Syvitski, 2002).

In many watersheds, erosion rates have been substantially impacted since settlement by man. Further, the recent surge in global dam and reservoir construction (Figure 3), has effectively reduced the sediment load of many large river systems (GESAMP, 1993). In contrast, conversion of native grasslands and forests to croplands can increase

soil erosion as well as sediment loads. Sediment yields reported in Table 3 are inherently different than terrestrial erosion rates in that sediment is eroded typically at a rate much greater than what can be carried away by rivers. This difference is due at least in part to a river's capacity to store sediment downstream, either within the thalweg during low flow conditions, or as prograding deltaic sediments at the land–sea margin (Meade *et al.*, 1985). As a consequence, larger watersheds generally produce much smaller and less variable sediment yields (Milliman and Meade, 1983; White and Blum, 1995; Canfield, 1997). While riverine sediment yield is a reasonable estimate of the off-continent flux of sediment, only a small component is transported out of an estuarine system and off the continental shelf into the open ocean. Thus, estimates of worldwide riverine sediment discharge represent a source of sediments to the coastal, rather than the pelagic ocean (Milliman and Meade, 1983). The many well-known physicochemical reactions (i.e. surface complexation, dissolution, flocculation, ion exchange, sorption, (co)-precipitation, electron transfer, and biological uptake) that occur in an estuary ultimately regulate oceanic constituent distributions in what Turekian (1977) eloquently termed the "*great particle conspiracy*" (Figure 4).

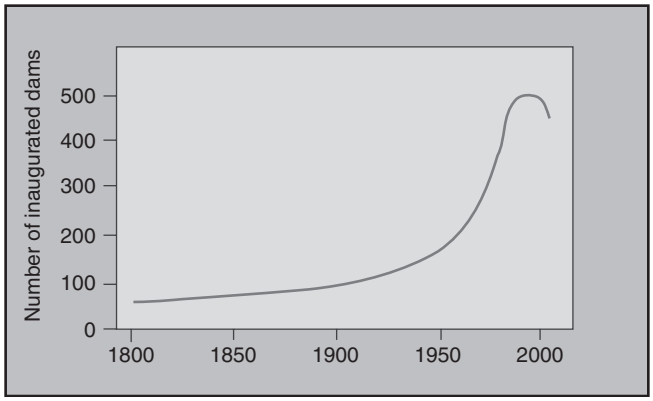


Figure 3 Global dam construction (modified from GESAMP, 1993). In terms of a ratio of the number of dams inaugurated relative to a continents population, North America leads the world (22.5 dams per million people), while Africa has the fewest new dams per population (1.4 dams per million people). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The mineralogy of riverine particulate matter is largely a function of grain size. Clays constitute most particles that are less than about 6 μm, whereas coarser particles consist dominantly of quartz, feldspar, and carbonate fragments. The most prevalent clay minerals in riverine particulates are illite, smectite, chlorite, and kaolinite, and are present in major world rivers on the order of 45–60%,

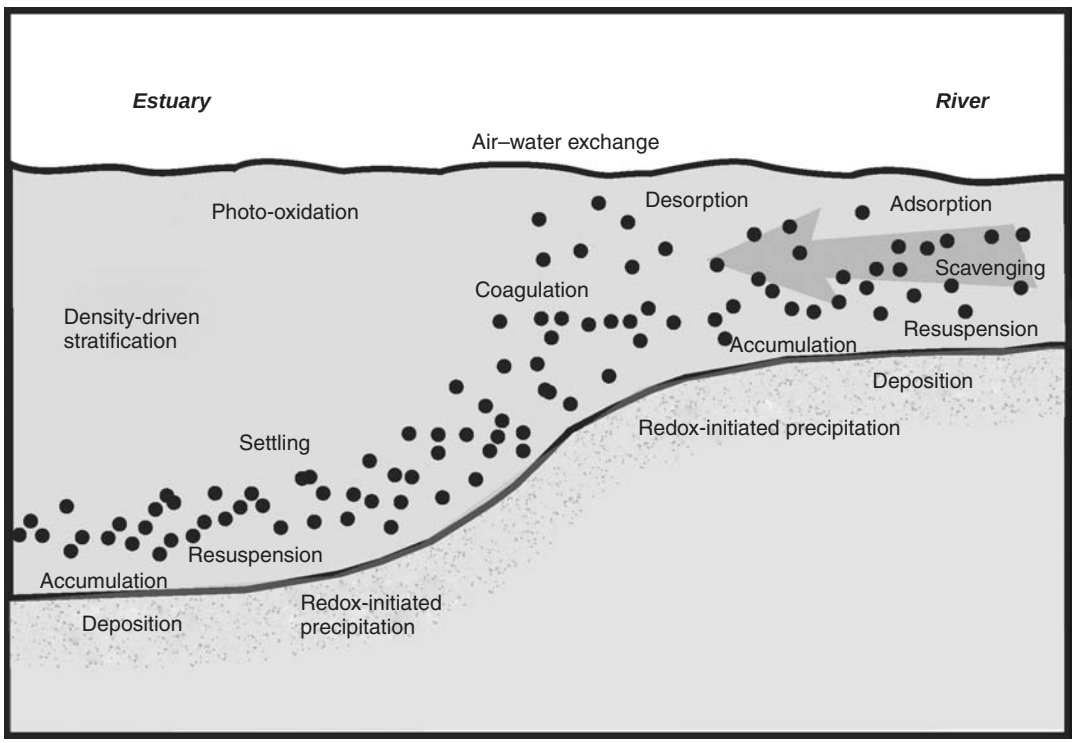


Figure 4 Idealized cartoon depicting some of the physical and biogeochemical processes/reactions that may affect trace elements during riverine transport and estuarine mixing. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

0–50%, 30–40%, and 0–100% respectively (Irion, 1991). Regardless of latitude (i.e. Yukon versus Amazon), rivers that drain young mountain ranges typically are rich in illite and chlorite (Gibbs, 1977). In contrast, temperate river systems such as the Mississippi River, as well as volcanic-borne rivers of Oceania (i.e. Fly River), are usually rich in smectite minerals.

River-borne sediments can conceptually be classified according to their origin, that is, lithogenic, authigenic, cosmogenic, and biogenic (Goldberg, 1954). However, it is usually difficult to chemically isolate such fractions. A more practical classification scheme divides sediments into their respective chemical components or phases: carbonates, clays, silicates, metal oxides/hydroxides, sulfides, phosphates, and organic matter (Jenne, 1977; Gibbs, 1973; Gibbs, 1977; Poulton and Raiswell, 2000). Most sediments are a complex combination of these various phases and may consist of multiple layers or components. These chemically distinct sediment components yield important information about the processes that control element partitioning between the dissolved and particulate phases (Förstner *et al.*, 1989).

ROLE OF PARTICULATES IN THE TRANSPORT OF RIVERINE CONTAMINANTS

It has been widely recognized that the reactive surface sites of riverine particulates, including colloids, are the major sequestering agent of dissolved trace elements and organic compounds in the aquatic environment (Goldberg, 1954; Turekian, 1977; Goldberg *et al.*, 1988; Zhang and Liu, 2002). Mechanisms involved in this dynamic dissolved-colloid-particulate partitioning (Figure 5) are complex and still not well understood. Particulate material, *per se*, consists of various phases that result from natural chemical alteration (e.g. adsorption, flocculation, complexation, or redox reactions) of weathered parent material during riverine transport (Martin and Meybeck, 1979; *see* **Chapter 83, Suspended Sediment Transport – Flocculation and Particle Characteristics, Volume 2**). Both macro- and microorganisms also may be important in waters that have low concentrations of inorganic suspended material, due to the active and/or passive incorporation of dissolved elements into soft tissues, tests, and fecal material (Dzombak and Morel, 1990). Fine-grained particles and colloids generally occur as aggregates of intermeshed or layered Fe/Mn oxides as well as organic and carbonate surface coatings (Jenne, 1977; Singh and Subranmanian, 1984; Hsi and Langmuir, 1985; Stumm, 1992). These reactive surface sites may then act as sources or sinks for dissolved elements in the aquatic environment, depending both on the chemical reactivity (i.e. lability) and on the residence time of a particular system (i.e. kinetics) (Stumm and Morgan, 1981; Sulzberger *et al.*, 1989).

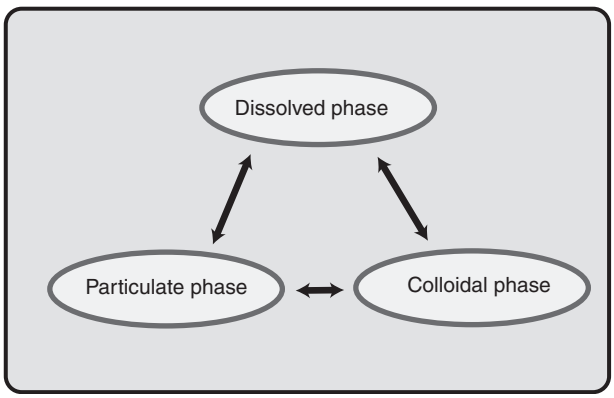


Figure 5 Phase partitioning between operationally defined dissolved (<0.45 μm), colloidal (0.45–0.001 μm), and particulate (>0.45 μm) pools. It is generally recognized that each of these pools exists within a dynamic and highly nonconservative particle size continuum (from (Swarzenski *et al.,* 1995) with permission from Elsevier). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The affinity of reactive elements for particle surface sites controls their behavior and fate in rivers. It is useful to look at the reversible phase associations of elements to infer processes of removal for a wide variety of compounds. The most important geochemical components are the iron and manganese (oxy) hydroxides, organic coatings, and particles, carbonates, sulfides, phosphates, and detrital clays and silicates (Stumm and Morgan, 1981). The influence of these mineral components is variable and usually related to physicochemical conditions, such as, pH, Eh, conductivity, alkalinity, and temperature. River-influenced coastal areas such as estuaries are well-known sinks for many terrestrially derived elements (Sholkovitz, 1976; Swarzenski *et al.*, 2003), and many reactive pollutants tend to accumulate in these environments (Förstner and Wittman, 1979; Skei and Paus, 1979; Benoit *et al.*, 1994). Widely fluctuating chemical conditions (i.e. conductivity, pH, dissolved oxygen) and biological blooms can cause a remobilization or repartitioning of these phase-species associations (Sholkovitz *et al.*, 1994). In sediments, one phase commonly does not have a large binding capacity, and a variety of phases may exist simultaneously, depending on season and flow conditions (Sharp *et al.*, 1982). The following text describes the role of various mineral components in riverine constituent transport.

Fe and Mn Oxides

In rivers, the (oxy)hydroxides of iron and manganese are among the most important particulate components in that they typically control metal ion sorption kinetics and thus metal behavior (cf. Singh and Subranmanian, 1984). These

oxides have been described as the major carrier for many reactive elements in river water (Gibbs, 1973; 1977; Benoit *et al.*, 1994). The amorphous nature of Fe and Mn oxides present in natural waters promotes active surface sorption and complexation of various ions.

Iron and Mn oxides are typically present as coatings on clay particles (Davis and Kent, 1990; Dzombak and Morel, 1990). Because of the extremely insoluble nature of the trivalent iron oxides and fast oxidization–precipitation kinetics, an amorphous iron oxide exists as the thermodynamically favorable phase (Zinder *et al.*, 1986). As a group, manganese oxides usually are even more amorphous in character than iron oxides (Balistrieri and Murray, 1986). This amorphous nature is most likely due to the formation of nonstoichiometric Mn oxides with variable oxidation states (+2, +3, +4) and also greater isomorphic substitution as a result of slower precipitation kinetics (Jenne, 1977). The surface area of these poorly defined iron and manganese oxides can be very large (up to $600\,m^2\,g^{-1}$; Dzombak and Morel, 1990), and thus are a tremendous source of potential sorption sites. The great abundance of reactive surface sites enables these oxides to be highly efficient scavengers for radionuclides and trace elements (Singh and Subranmanian, 1984) during riverine transport.

Two general types of scavenging mechanisms may occur at the oxide surface: specific (chemical) and nonspecific (physical) adsorption (Davis and Kent, 1990; Stumm, 1992). Briefly, nonspecific adsorption describes the process by which the sorbed ion retains its solvation shell and forms an outer-sphere complex with surface functional groups such as oxides (Davis and Kent, 1990; Dzombak and Morel, 1990). Adsorption of these weakly bound ions is more dependent on the electrostatic attraction between a charged oxide surface and an oppositely charged ion. Conversely, specific adsorption describes the coordinative displacement of H^+ ions bound to the surface oxygen by cations; hence, this process is dominated by covalent bonding rather than electrostatic attraction. This latter mechanism of strongly bound metal ion adsorption will yield inner-sphere complexes at the oxide surface (Waite *et al.*, 1994). Either sorption mechanism will yield a cation-oxide complex that can behave very differently across chemical gradients in natural waters (Mayer *et al.*, 1984). For example, one would expect an outer- rather than an inner-sphere cation-oxide complex to be susceptible to ionic strength variations. An observed linear decrease in uranium K_d's (distribution coefficients) across a salinity gradient on the Amazon River shelf led Swarzenski *et al.* (1995) to conclude that U was most likely associated in a weakly bound, reactive outer-sphere cation-oxide complex. Once a cation is adsorbed to an oxide surface, which is usually a very rapid process, incorporation into the oxide lattice may occur either through molecular diffusion or precipitation. Some cations can be physically trapped in this manner into the lattice framework where they remain relatively immobile (e.g. ^{137}Cs, ^{7}Be).

Dissolution of Fe and Mn oxides can readily occur in response to a pH change, a decrease in the redox potential (Eh), or an increase in the concentration of complexing agents (Moore *et al.*, 1979). The aqueous manganese cycle, from the dissolved ion through its removal to sediments, is illustrated in Figure 6. This sequence, which in many ways is analogous to the redox chemistry of Fe, is very important in the recycling of redox-sensitive and particle-reactive elements (Broecker and Peng, 1982) across redox boundaries. Balistrieri and Murray (1986) indicate that the cycling of manganese from the dissolved to particulate state may be responsible for large-scale removal of reactive elements in the ocean. Indeed, Fe/Mn hydrous oxides are closely tied to the behavior of trace metals, rare earth elements, and radionuclides in many rivers (Gibbs, 1977; Plater *et al.*, 1992; Sholkovitz *et al.*, 1994), estuaries (Sholkovitz, 1993; Swarzenski *et al.*, 1995), anoxic basins (Swarzenski *et al.*, 1999), and interstitial waters (Burdige, 1993).

As a class, the manganese oxides appear to be more efficient scavengers of select dissolved elements than iron oxides (Balistrieri and Murray, 1986; Sholkovitz *et al.*, 1994; Swarzenski *et al.*, 1999). The sorptive properties of these oxides are related to their unique structural and colloidal characteristics, high surface charge, and cation adsorption capacity over a narrow pH range (Means *et al.*, 1978a, b). For example, Carpenter *et al.* (1984) found ~92% of ^{210}Pb associated with the manganese oxide phase, and less than 4% with the organic phase.

Organic and Inorganic Carbon

Dissolved and particulate riverine carbon may be both organic and inorganic in nature. The various forms of riverine carbon may have different sources, characteristics, and aqueous residence times (Table 4). Through various phase-partitioning processes, carbon is responsible for the transport and delivery of CO_2, as well as other carbon-associated contaminants and nutrients to the world's ocean (Meybeck, 1982; 1993).

In rivers that exhibit a natural pH range, dissolved inorganic carbon (DIC) is found mainly as the bicarbonate ion, HCO_3^-. Riverine DIC originates entirely from atmospheric and soil CO_2 in noncarbonate environments (i.e. metamorphic, volcanic, plutonic, most sandstone, and shale), while ~50% of riverine DIC in carbonate environments results from the weathering of bedrock (Esser and Kohlmaier, 1991). Riverine (particulate inorganic carbon) PIC is derived from the mechanical erosion of sedimentary rock, which results in the transfer of carbonate material from their source to the oceans.

Riverine organic carbon exists within a continuous size spectrum that extends in size from monomers and colloids to large macroparticles and aggregates. Two major pools of

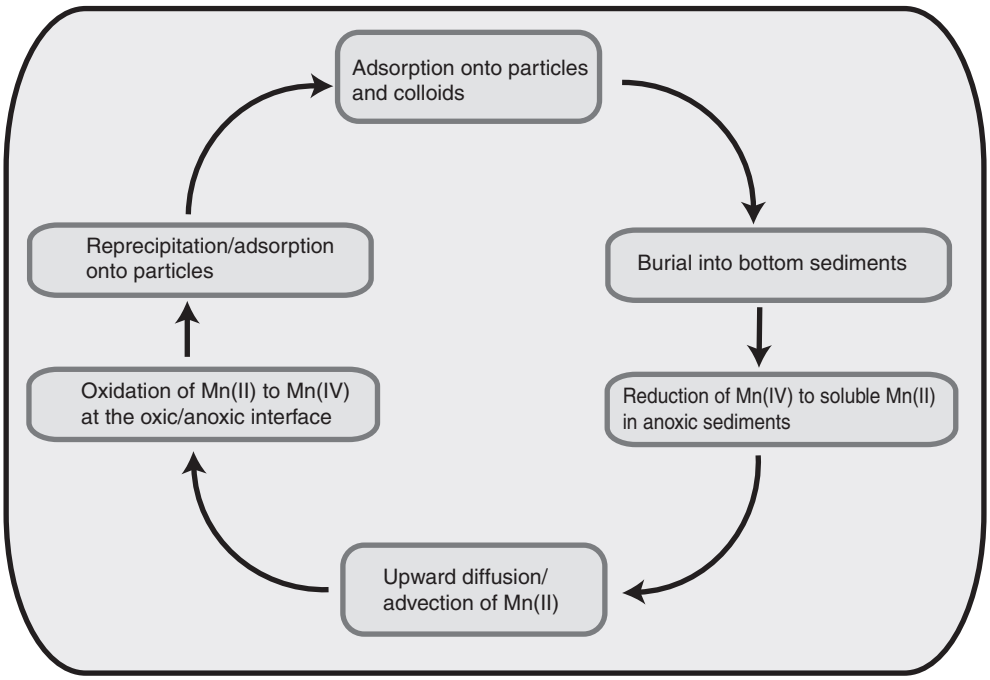


Figure 6 Idealized redox transformations of the manganese cycle (Reproduced from Stumm and Morgan, 1981, by permission of John Wiley). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Table 4 Various forms of riverine carbon and their natural sources (adapted from Meybeck, 1993)

Phase components			Natural origin
total carbon	dissolved carbon	DIC	carbonate mineral weathering atmospheric and soil CO_2
		DOC	soil and plant leaching
	particulate carbon	POC	soil erosion sedimentary rocks autochthonous sources
		PIC	autochthonous sources sedimentary rocks

DIC = dissolved inorganic carbon; DOC = dissolved organic carbon; POC = particulate organic carbon; PIC = particulate inorganic carbon.

organic carbon can be separated simply by an operational definition; organic carbon retained, for example, on a 0.4–0.7 μm filter is labeled (*particulate organic carbon*), *POC* and the carbon transferred through the filter membrane is termed (*dissolved organic carbon*). *DOC* In their review of world riverine carbon fluxes to the sea, Ittekot and Laane (1991) thus confirm that the transport of carbon in rivers, expressed as a ratio of DOC/POC, is inversely related to the concentration of suspended particulate matter (Figure 7).

The dominant components of organic carbon in natural water are operationally defined fulvic and humic acids, which together constitute humic substances (~50%), and hydrophilic acids (~30%). Simple compounds such as carbohydrates and carboxylic acids comprise the remaining 20% (Thurman, 1985). The relative abundance of these operationally defined organic components can shift dramatically from one river to the next. For example, humic substances in highly organic rivers such as the Suwannee River can exceed 90% of the entire organic carbon content. The predominant functional groups of humic substances are carboxylic acids and phenols, yet many other groups (e.g. N, S) may also strongly affect metal scavenging, depending on the particular environmental setting.

DOC concentrations in rivers range from more than $20\,\mathrm{mg\,L^{-1}}$ in some polluted or tropical rivers to less than $1\,\mathrm{mg\,L^{-1}}$ in alpine streams (Thurman, 1985). In pristine rivers, DOC is derived mainly from soil leaching with the absolute amount of exported DOC linked to the river-basin drainage patterns. For example, Moore (1989) compared two stream basins in Westland, New Zealand, and determined that the poorly drained Larry River exported $78.1\,\mathrm{g\,C\,m^{-2}\,yr^{-1}}$ (a global maximum), while the well-drained Maimai basin exported only $6.8\,\mathrm{g\,C\,m^{-2}\,yr^{-1}}$.

In addition to the strong influence riverine DOC has on the global carbon cycle, riverine organic carbon also can affect nutrient and contaminant availability as well as constituent transport (Meybeck, 1993; Onstad *et al.*, 2000). Disaggregation/aggregation reactions involving the solid

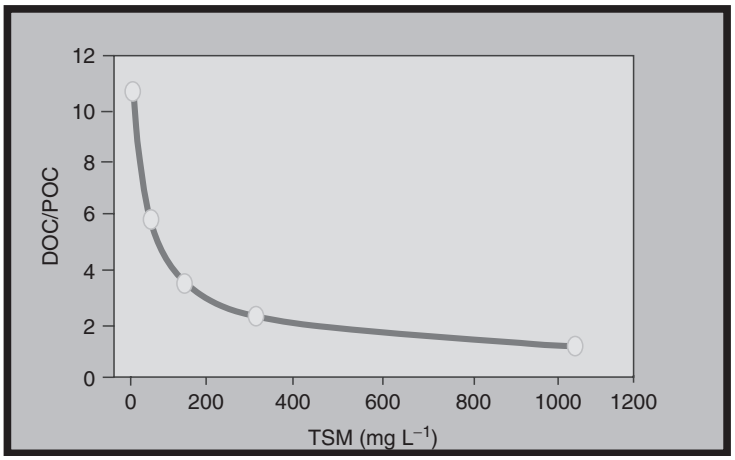


Figure 7 Relation between the ratio of dissolved and particulate organic carbon and total suspended matter from different world rivers (after Ittekot and Laane, 1991). This ratio changes with increasing concentrations of total suspended sediments. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

phase may facilitate a redistribution of contaminants into the water column that can lead to a concomitant increase in potential toxicity. Certain components of DOC (i.e. humic and fulvic acids) have cation binding and ion-charge properties that partition trace elements from the solid to the dissolved phase.

Organic carbon has a net negative surface charge that enables it to interact readily with cations, resulting in complex associations by ion exchange, adsorption, and chelation (Duinker, 1980). Organic carbon is able to modify the solubility, redox potential, and precipitation behavior of many cations, organometallic complexes, and organic contaminants (Thurman, 1985). Such transformations due to the presence of organic carbon may cause an increased solubility and mobilization in such elements as Cu, As, Pb, Hg, Cd, Co, Ag, Ni, Sn, Al, U, and Th (Santschi *et al.*, 1988) and possibly the transuranics (Olsen *et al.*, 1982). Actual mechanisms are still largely unknown, but as a class, organic material is believed to control cation adsorption onto marine particulate matter (Balistrieri and Murray, 1986). Clay minerals also strongly adsorb humic material that can then interact with metal oxides and hydroxides and greatly affect surface adsorption characteristics.

Cation adsorption onto organic matter is not well understood because of the inherent physical and chemical complexity of these organic macromolecules. Most likely, as with Fe/Mn hydrous oxides, surface functional groups of organic carbon can form either inner- or outer-sphere complexes. These sorption reactions (specific and nonspecific) commonly involve a transfer of protons, and are therefore extremely pH sensitive. A lowering in pH can frequently initiate the reduction and subsequent dissolution of Fe/Mn oxides. The nonreductive dissolution of Fe(III) or Mn(IV) can be further enhanced if the H^+ bonds to the oxygen on the oxide surface, thereby weakening the critical Fe- or Mn oxygen bond (Waite, 1990). A similar mechanism of nonreductive dissolution can be invoked for organic matter in which ligands complex Fe on the oxide surface and form soluble Fe-ligand complexes (Stone and Morgan, 1984).

Decomposition of organics within the sediment column may lead to changes in the distribution of sediment-associated contaminants, possibly altering them into more readily available or mobile phases. Flow, currents, and biological conditions also can alter the phase association of trace elements. For example, Presley *et al.* (1980) found that low flow conditions in the Mississippi River initiated an increase in particulate organic matter that could cause a change in particulate-trace element interactions. Complexing agents, such as EDTA, and natural organics also are known to promote the formation of strong complexes with certain radionuclides and can reduce the binding capacity of the particulates for the radionuclides (Means *et al.*, 1978b). In their review, Förstner and Wittman (1979) indicate that sedimentary organics play a vital role in the remobilization of trace and transuranium elements. Laboratory experiments have shown Cu, Co, Mn, Ni, and Zn can all be solubilized from carbonate and sulfide phases by interaction with organic matter (Rashid and Leonard, 1973).

Carbonates, Sulfides, and Refractory Minerals

While hydrous Fe/Mn oxides and organics are certainly the most important particulate components in terms of regulating contaminant behavior in rivers, the role of carbonates, sulfides, and refractory minerals may become significant in certain river systems. For example, under anaerobic conditions, specific bacteria are able to utilize sulfate as their terminal electron acceptor, and derive energy from the reduction of sulfate to sulfide. In the presence of sulfide, ferrous iron is rapidly removed as FeS and, due to

the fine size of the particles, adsorbs onto larger sediment grains (Berner, 1971; Burdige, 1993). This FeS coating gives anoxic sediments their characteristic black color.

Reactive trace elements may also be removed from river water by forming insoluble sulfide compounds (e.g. FeS, PbS, ZnS). This removal process, whether it be a pure sulfide precipitation, solid sulfide formation, adsorption onto FeS, or a combination of these, is often not well-known (Oakley *et al.*, 1981). Coprecipitation with iron sulfides appears to be a sink for trace elements in anoxic settings (Swarzenski *et al.*, 1999); however, direct evidence is sparse (Presley *et al.*, 1980; Förstner and Wittman, 1979). Reduction of Fe/Mn oxides in the sediment column may release substantial quantities of adsorbed trace elements that may quickly adsorb onto FeS and pyrite (Shannon and White, 1991).

The sulfide component also may be an important source of elements to the water column. Oxygen can be introduced into anoxic sediments through various physical or biological processes. This oxygenation may then cause the oxidation of sulfide minerals, and result in the subsequent release of any associated trace elements. This process may be important, especially in intermittently anoxic river sediments, where seasonal remobilization and mixing can produce ephemeral oxidizing conditions within surficial sediments.

Carbonate minerals also are potential scavengers of trace elements, and are present in varying degrees in most rivers and estuaries. Carbonate precipitation in estuarine and coastal sediments may be important in areas of high sedimentation, which limits the diffusional reequilibration of pore water to the overlying waters (Jenne, 1977). Coprecipitation of zinc and cadmium with the carbonate fraction was prevalent in Rhine River sediments (Förstner *et al.*, 1989). Under anoxic conditions, manganese carbonate precipitation may control interstitial water manganese concentrations (Förstner and Wittman, 1979).

The detrital or refractory component is a dominant phase for many elements during riverine transport. Gibbs (1973; 1977) looked at Amazon and Yukon River sediments and found the predominant components of Cu, Cr, and Co were in the crystalline lattices of detrital minerals. This phase is considered to be geochemically almost nonreactive, and unavailable for further desorption/adsorption reactions under normal conditions. Trace elements associated with this phase are assumed to be present in the weathered material, and should therefore have concentrations close to crustal abundances. The primary importance of detrital mineral phases, such as clays and silicates, is to provide adsorption sites for reactive phases or nucleation centers for flocculation and precipitation reactions (Duinker, 1980).

Chemical Extractions

Analytical techniques utilizing selective chemical extractants have been widely used to understand and describe the nature of trace element associations in sediments (cf. Gad and LeRiche, 1966; Chao, 1972; Malo, 1977; Guy *et al.*, 1978; Tessier *et al.*, 1979; Chester *et al.*, 1988). These techniques permit for a gross quantification of operationally defined sediment phases and the trace elements associated with them (Tessier *et al.*, 1979; Luoma and Bryan, 1981). Studies dealing with trace element remobilization (Skei and Paus, 1979; Presley *et al.*, 1980), potential bioavailability (Luoma and Jenne, 1976; Luoma and Davis, 1983), fate of anthropogenic riverine inputs (Slavek *et al.*, 1982; Van Valin and Morse, 1982; Raoux *et al.*, 1999), and transport and removal mechanisms of trace elements (Gibbs, 1973; 1977; Oakley *et al.*, 1981; Poulton and Raiswell, 2000) have all used selective extraction schemes. Figures 8 and 9 illustrate the partitioning of Fe, Mn and Zn, and Ni, respectively, into three operationally defined sediment phases in four large rivers (Poulton and Raiswell, 2000). In this extraction scheme, the dithionite-soluble fraction targets oxide surfaces while the HCl-soluble fraction isolates both sheet silicates as well as some carbonates. In general, multiple extraction methods should, however, be applied with much caution and may easily be subject to misinterpretation.

Radionuclide Associations with Particulate Components

Most chemical extraction studies have been developed for a suite of trace elements, including Cu, Pb, Zn, Cd, Co, and Ni. The application of such extraction schemes to radionuclides also provides a useful tool to assess the environmental fate and behavior of such elements that are potentially toxic and recalcitrant (Harmon and Ivanovich, 1993).

The role of different sediment components for radionuclides in riverine sediments and suspended particles still needs to be better understood (Chabaux *et al.*, 2003; Swarzenski *et al.*, 2004). In a recent study of uranium series disequilibrium within river sediments, Plater *et al.* (1992) reported that over 70% of all sediment-bound U exists in a chemically inert residual component. In an investigation of the behavior of uranium on the Amazon shelf, McKee *et al.* (1987) utilized a two-step $NH_2OH^{\bullet}$ HCl-Na citrate and HF-$HClO_4$ leach to separate U into a reactive ferric oxyhydroxide and a residual component. Their data suggest vastly different partitioning of U within these two fractions. At a distal offshore station (relative to the river mouth), ~80% of the uranium was bound onto an operationally defined residual component, while at a proximal station, 60% of the uranium was bound to Fe-oxyhydroxides. Results like those of McKee *et al.* (1987) and Means *et al.* (1978b) demonstrate that differing environmental conditions; that is, organic rich, reducing sediments, high biological activity, may impose differing phase associations. The ability of certain radionuclides

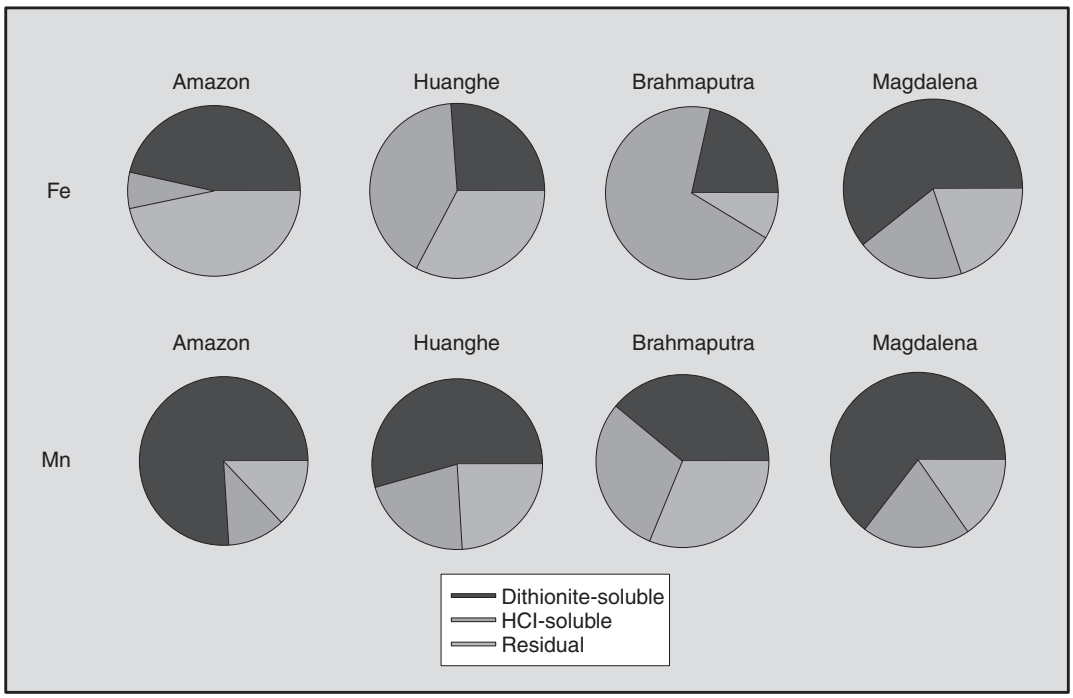


Figure 8 Fe (wt%) and Mn (ppm) partitioning into three operationally defined mineral components for the Amazon, Huanghe, Brahmaputra, and Magdalena rivers (data from Poulton and Raiswell, 2000). A dithionite extraction is expected to target elements associated with Fe hydroxide surfaces; HCl extraction will remove both sheet Si minerals as well as hydrolysable organic matter. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

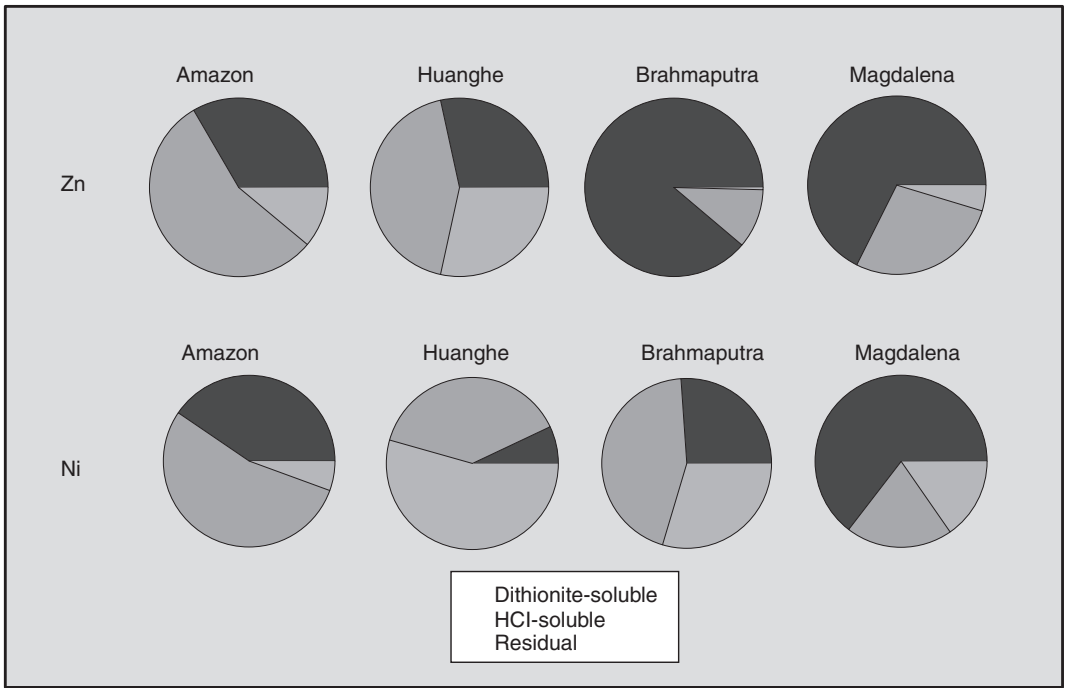


Figure 9 Zn (ppm) and Ni (ppm) partitioning into three operationally defined mineral components for the Amazon, Huanghe, Brahmaputra, and Magdalena rivers (data from Poulton and Raiswell, 2000). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

to associate with different phases in varying environments may lead to problems in the use of radioactive elements for geochronological studies, chemical tracers, or any description of the geochemistry of the element (e.g. Benoit *et al.*, 1994). Geochemical processes, such as, particulate removal, diffusion, and biological activity, all affect the equilibrium of the natural radioactive decay series elements. A thorough treatise on riverine transport of select U-Th series isotopes is presented in Chabaux *et al.* (2003).

Riverine Flux of Nitrogen and Phosphorus

Alterations to the riverine flux of nitrogen and phosphorus have not been confined exclusively to the twentieth century (Meybeck, 1982). However, over the last 3–4 decades there has been a tremendous interest in more comprehensively understanding the local and regional transport processes of N and P at the land–sea margins (Nixon *et al.*, 1996). The impact of enhanced N and P inputs can trigger microbial and planktonic water column blooms that affect suspended matter concentrations and light availability necessary for photosynthesis. Prolonged blooms may alter local benthic and water column communities and can eventually lead to ecosystem-wide degradation (McIsaac *et al.*, 2001). For example, in the Gulf of Mexico adjacent to the Mississippi River, low oxygen concentrations (i.e. hypoxia) in a stratified water column occur frequently during summer months.

In nonimpacted river catchments, the physicochemical weathering of soils and biological uptake/regeneration, contribute to the mobilization and subsequent transport of N and P in river systems (Alexander *et al.*, 2000). Altered river catchments (i.e. deforestation, increased land drainage, impoundments, industry and agriculture) greatly increase the natural terrestrial flux of N and P to the sea. Except along active upwelling coastlines, rivers usually are considered the major estuarine contributor of N and P for both impacted and pristine watersheds, and this flux is a function of a river's sediment and water discharge budget. There are typically numerous N and P species in a river that may reside either in a dissolved (nitrate, nitrite, ammonium, organic N, organic P and inorganic P) or particle-bound (organic N, organic P and inorganic P) pool. In general, particulate species tend to dominate the overall riverine load of P more so than N, although in heavily impacted river systems, enhanced concentrations of nitrate and dissolved inorganic P species may also be observed (Downing *et al.*, 1993). It has been estimated that recent anthropogenic perturbations have increased global riverine P fluxes by as much as a factor of 3; riverine N fluxes have similarly increased across North America (3x) and Europe (6x) (Howarth *et al.*, 2002). Table 5 shows estimates ($Mt\,yr^{-1}$) for the global riverine flux of N and P into the ocean (after Tappin, 2002).

Table 5 $^{-}$Estimates ($Mt\,yr^{-1}$) for the global riverine flux of N and P (after Tappin, 2002)

Nitrogen		Phosphorus	
		DOP	0.6
DON	10	PIP	12
DIN	10–20	POP	8
PN	21	TDP	2
TDN	7	TPP	20
Total N	48–58	Total P	22.6

DON = dissolved organic N; DIN = dissolved inorganic N, PN = particulate N, TDN = total dissolved N; DOP = dissolved organic P, PIP = particulate inorganic P, POP = particulate organic P, TDP = total dissolved P and TPP = total particulate P.

CONTINENTAL DISCHARGES

It has been widely shown (cf. Trefry *et al.*, 1986; Swarzenski *et al.*, 1995; Horowitz *et al.*, 2001) that the vast majority (up to 90%) of trace elements carried downstream in a river are associated with particulates and colloids. The fate of inorganic and organic constituents during riverine transport, regardless of continent, is thus directly tied to the nonconservative fate of these solids. For example, riverine particles and colloids undergo constant aggregation/disaggregation reactions, they also can be seasonally stored in the lower reaches of slow-flowing rivers, and can be subsequently remobilized from riverbed and riverbank sediments during high discharge (Walling *et al.*, 2003). Table 6 summarizes sediment-associated trace-metal concentrations from some major world rivers.

Europe

Despite being relatively diminutive in size, the rivers of Europe are historically well documented and often well researched. For example, much more has been written on the Rhine or Danube Rivers than the comparatively sized rivers of Asia or the Americas. Well over 100 European rivers empty into varied coastal receiving basins located in the Barents, White, Norwegian, North, Baltic, Mediterranean, Black, Caspian, and Adriatic Seas, as well as the Atlantic Ocean. Combined, these European rivers contribute approximately 7.4% ($2800\,km^3\,yr^{-1}$) to the world's total freshwater discharge of $37\,700\,km^3\,yr^{-1}$ (Milliman, 2001), yet only about 2% of the total annual sediment discharge. Although none of the European rivers rank among the top 10 rivers of the world as calculated using drainage basin, water discharge, or length, only the Rhone, Po, and Danube Rivers have annual sediment loads that exceed 10 million tons (Milliman and Meade, 1983).

Table 6 Concentrations ($\mu g\, g^{-1}$) of particulate trace metals in some major world river systems

River	Co	Cr	Cu	Fe	Mn	Ni	Pb	Zn
Amazon[d]	41	193	266	5.55	1033	105.1	nd	nd
Danube[d]	13	100	89	5.50	600	40	178	91
Fly[a]	20	83	33	nd	721	34	13	108
Ganges[d]	14	71	30	3.70	1000	80	nd	163
Huanghe[f]	16	69	35	nd	636	57	nd	88
Mississippi[b]	21	77	38	4.43	1260	50.5	39	193
Niger[d]	40	150	60	0.92	650	120	nd	nd
Orinoco[e]	nd	84	61	7.40	588	34	nd	77
St. Lawrence[c]	14	85	40	5.07	1311	nd	nd	342
Zaire–Congo[d]	30	211	100	7.10	1400	54	220	300

[a]Carey *et al.* (2002).
[b]Presley *et al.*(1980); Trefry *et al.* (1986).
[c]Yeats and Brewer (1982).
[d]Martin and Meybeck (1979).
[e]Eisma (1978).
[f]Poulton and Raiswell (2000).
nd = no data.

European rivers, although geographically confined to a small continent, exhibit a nearly global representation of rivers that can range from arctic to subtropical, low to high discharge, low to high suspended and dissolved loads, and low to high degree of human impact (Cotté-Krief *et al.*, 2000). Today, the rivers of Europe are among the most industrialized and populated, yet some of these rivers (e.g. the Po and Ebro Rivers of Greece) have undoubtedly been impacted by human activities as far back as 2,500–3,000 years ago. More recently, many rivers of Europe have been affected by urbanization, deforestation, agriculture, dam construction, mining, and industry (GESAMP, 1993; Milliman, 2001).

Africa

Several large rivers drain the African continent and rank very high among the world's largest rivers in terms of their respective drainage basins and discharge characteristics (Milliman and Meade, 1983). In terms of combined annual sediment discharge, African rivers contribute about 4% of the total worldwide estimate of $13\,526 \times 10^6\, t\, yr^{-1}$. Similar to other continents, the drainage basins of large rivers within Africa may be divided into coastal (peripheral) and inland systems. The upper Niger River, Lake Chad, and the Okavango Delta are examples of inland drainage basins. Prominent examples of coastal drainage systems include the Zaire–Congo, lower Niger, Orange, Senegal, Gambia, Volta, and Saloum, which all flow into the Atlantic Ocean. The Zambesi and Limpopo flow into the southern Indian Ocean and the Nile discharges into the Mediterranean Sea.

Africa is a well-weathered continent where the topography has been largely reduced by mechanical and chemical denudation, and the soil profiles are commonly lateritic, and consequently rich in Fe–Mn–Al hydroxides (Martin, 1988). As a result, dissolved and particulate loads are generally much lower than in other large rivers of the world (Milliman and Meade, 1983; Dupré *et al.*, 1996). Most African rivers are seasonal in discharge, yet this cyclicity has been modified by the construction of numerous dams and reservoirs (Frater, 1987). Such alterations to the natural flow of a river affect both the suspended sediment and the total dissolved solids load of a river. The Nile represents an excellent example of a river that has been almost entirely modified within the last 30–40 years by the construction of the Aswan Dam, irrigation, and extensive industry (GESAMP, 1993). The Niger, Zambesi, and Orange are further examples of African rivers that have, to varying degrees, been harnessed by man's activities. The particulate loads discharged by African rivers into the surrounding oceans are quite low. In fact, many large African rivers first empty into inland swamps or lakes before making their way to the coast, and these lakes can efficiently lower and store dissolved and particulate materials of a river (Milliman and Meade, 1983).

Asia

The largest Asian rivers can be grouped into those that drain the Himalayas (i.e. Ganges, Brahmaputra, Indus, Irrawaddy) and those that drain mainland China (i.e. Huanghe, Changjiang). Combined, these rivers contribute almost 50% of the total annual sediment discharged to the ocean (Zhang, 1995). Most of this discharge is the result of a short and energetic rainy season (monsoon-driven) and extreme topographic relief within the drainage basins (Sarin *et al.*, 2002; Meade, 1996). While the Himalayan rivers exhibit pronounced local temporal and spatial deviations (Ramesh *et al.*, 2002), collectively, their chemistry closely reflects the world average river water composition

as reported by Meybeck (1982). In contrast to the industrial rivers of Europe and North America, trace element concentrations associated with suspended particles from the two largest Chinese rivers, today, remain quite similar to background soil levels within the respective watersheds (Zhang, 1995). In the Huanghe River, 90% of the sediment load is derived from easily weathered loess deposits (Poulton and Raiswell, 2000). Although many of the Asian rivers carry a large sediment load (e.g. Ganges-Brahmaputra), as much as half of the annual sediment flux can be sequestered in their low-lying floodplains and thus, not directly delivered to the ocean. This has important implications for the off-continent riverine transport of sediment-associated contaminants (Sarin *et al.*, 2002; Zhang and Liu, 2002).

Oceania

The importance of rivers and streams draining the high-standing islands in the East Indies in global river–ocean mass-balance estimates has recently become clear. Although these islands only account for roughly 2% of the world's landmass, they contribute as much as 30% of the global sediment entering the sea (Milliman and Meade, 1983; Milliman and Syvitski, 1992; Milliman *et al.*, 1999). This disproportionately large riverine sediment load delivered from small watersheds is the result of steep topographic relief, heavy precipitation, and relatively young bedrock that is easily weathered. Together, Asia and Oceania contribute up to 75% of the riverine sediments discharged into the ocean (Milliman and Syvitski, 1992), therefore dominating the oceanic particulate flux. While many of these rivers are still relatively pristine (e.g. Purari River of Papua New Guinea), there are some notable exceptions where large mining operations, and dense urban centers introduce a wide range of inorganic and organic contaminants. From a study on riverine fluxes from several high-standing islands of Oceania, Carey *et al.* (2002) suggest that the sediment-associated trace-element concentrations approximate upper-continental crustal values. This confirms that the riverine particulates of Oceania are generally less weathered during transport than those carried by large river systems, such as the Amazon.

South America

Well over two-third's of the South American continent is drained by only six large rivers that collectively contribute about $8000\,km^3\,yr^{-1}$ of fresh water to the Atlantic Ocean. One single river, the Amazon, is responsible for almost 60% of this discharge estimate. About 13% of the world suspended-sediment budget is derived from just three South American rivers: the Amazon, the Orinoco and the Paraná. Although the South American continent extends over almost 70 degrees of latitude, almost 80% lies within the humid tropics. Of the world's major river systems, the Amazon River has few rivals. It is the second longest river in the world after the Nile, and its drainage basin ($6.3 \times 10^6\,km^2$) produces the world's largest freshwater ($6300 \times 10^9\,m^3\,yr^{-1}$) and sediment ($1{,}000-1{,}300 \times 10^6\,t\,yr^{-1}$) discharge (Meade, 1996). As in some of the large Asian rivers, large quantities of fine-grained sediment are stored and resuspended seasonally in the lower Amazon (Meade *et al.*, 1985), effectively modulating the annual discharge of sediment-bound riverine constituents. Nonetheless, the Amazon represents a tremendous source of dissolved, colloidal, and particulate riverine constituents to the Atlantic Ocean (Swarzenski *et al.*, 1995).

North America

Although the discharge and sediment loads of the major North America rivers have been reasonably well-quantified since the 1950s, the large rivers of Alaska and Canada that discharge into the Arctic Ocean have only just recently been studied in a rigorous manner (Dai and Trenberth, 2002; Holmes *et al.*, 2002; Syvitski, 2002; Álvarez *et al.*, 2003). In general, the rivers of North America that drain either the eastern or western seaboard have small sediment loads. Many of these rivers are either dammed (e.g. Columbia), or they flow first through a network of lakes and reservoirs that efficiently trap sediments (i.e. St. Lawrence). Rivers that discharge into the Gulf of Mexico carry the greatest sediment load. Of these, the Mississippi is by far the largest, both in terms of water and sediment discharge, although Meade (1995) has shown a roughly 50% decrease in this sediment load from 1700 to 1985 because of changing agricultural practices and dam construction. With the exception of Li and Sr, the majority (~70 to 90%) of trace metals are transported downstream in association with suspended sediments (Trefry *et al.*, 1986; Horowitz *et al.*, 2001). Telltale signs of upriver mining and industry as well as the distinct climatic control on watershed weathering can be clearly seen in the particulate loads of select North American rivers (Shiller, 1993; White and Blum, 1995; Canfield, 1997; Horowitz *et al.*, 2001).

Future Trends in River Research

One of the most difficult tasks for river scientists and managers alike is to more accurately relate recent anthropogenic influences in watersheds to the coastal delivery of sediments and sediment-associated contaminants. From detailed studies, rivers are now well known to effectively store their sediment load across many different timescales that can range from daily events to 1000- yr climatic cycles. Such storage and remobilization within a river also may integrate geologic and geochemical signatures that can further obscure

the relation between upland soil erosion and river-sediment transport. By developing better watershed erosion estimates and riverine particulate budgets that include storage and remobilization terms, river-discharge estimates and predictive models can be produced that are both reliable and informative for sediment-associated contaminant inventories and off-continent flux estimates.

REFERENCES

Alexander R.B., Smith R.A. and Schwartz G.E. (2000) Effect of stream channel size on delivery of nitrogen to the Gulf of Mexico. *Nature*, **403**, 758–761.

Álvarez M., Ríos A.F. and Pérez F.F. (2003) Transports and budgets of total inorganic carbon in the subpolar and temperate North Atlantic. *Global Biogeochemical Cycles*, **17**, 2-1–2-21.

Balistrieri L.S. and Murray J.W. (1986) The surface chemistry of sediments from the Panama basin: The influence of Mn oxides on metal adsorption. *Geochimica Cosmochimica Acta*, **50**, 2235–2243.

Benoit G., Oktay-Marshall S.D., Cantu A., Hood E.M., Coleman C.H., Corapcioglu M.O. and Santschi P.H. (1994) Partitioning of Cu, Pb, Ag, Zn, Fe, Al, and Mn between filter-retained particles, colloids, and solution in six Texas estuaries. *Marine Chemistry*, **45**, 307–336.

Berner R.A. (1971) *Principles of Chemical Sedimentology*, McGraw-Hill Book Company: New York, p. 240.

Berner E.K. and Berner R.A. (1996) *Global Environment: Water Air and Geochemical Cycles*, Prentice Hall: New Jersey, p. 376.

Broecker W.S. and Peng T.-H. (1982) *Tracers in the Sea*, Lamont-Doherty Geological Observatory: p. 690.

Burdige D.J. (1993) The biogeochemistry of manganese and iron reduction in marine sediments. *Earth Science Reviews*, **35**, 249–284.

Canfield D.E. (1997) The geochemistry of river particulates from the continental USA: Major elements. *Geochimica Cosmochimica Acta*, **61**, 3349–3367.

Carbonnel J.P. and Meybeck M. (1975) Quality variation of the Mekong River at Phnom Penh, Cambodia, and chemical transport in the Mekong Basin. *Journal of Hydrology*, **27**, 249–265.

Carey A.E., Nezat C.A., Lyons W.B., Hicks D.M. and Owen J.S. (2002) Trace metal fluxes to the ocean: The importance of high-standing oceanic islands. *Geophysical Research Letters*, **29**, 141–144.

Carpenter R., Peterson M.L., Bennett J.T. and Somayajulu B.L.K. (1984) Mixing and cycling of uranium, thorium and Pb-210 in Puget sound sediments. *Geochimica et Cosmochimica Acta*, **48**, 1949–1964.

Chabaux F., Riotte J. and Dequincey O. (2003) U-Th-Ra fractionation during weathering and river transport. *Reviews in Mineralogy and Geochemistry*, **52**, 533–576.

Chao T.T. (1972) Selective dissolution of manganese oxides from soils and sediments with acidified hydroxylamine hydrochloride. *Soil Science Society of America*, **36**, 764–768.

Chester R., Thomas A., Lin F.J., Basaham A.S. and Jacinto G. (1988) The solid state speciation of copper in surface water particulates and oceanic sediments. *Marine Chemistry*, **24**, 261–292.

Cotté-Krief M.H., Guieu C., Thomas A.J. and Martin J.M. (2000) Sources of Cd, Cu, Ni and Zn in Portuguese coastal waters. *Marine Chemistry*, **71**, 199–214.

Dai A. and Trenberth K.E. (2002) Estimates of freshwater discharge from continents: Latitudinal and seasonal variations. *Journal of Hydrometeorology*, **3**, 660–687.

Davis J.A. and Kent D.B. (1990) Surface complexation modeling in aqueous geochemistry. *Reviews in Mineralogy*, **23**, 177–260.

Degens E.T. and Kempe S. (1982) Riverine carbon-an outlook. In *Transport of Carbon and Minerals in Major World Rivers*, Degens E.T. (Ed.) Pt. 1. Vol. 52, Mitteilung des Geologischer. – Paläontologisher Institutue, University Hamburg: SCOPE/UNEP Sonderbund, pp. 757–764.

Downing J.P., Meybeck M., Orr J.C., Twilley R.R. and Scharpenseel H.W. (1993) Land and water interface zones. *Water, Air and Soil Pollution*, **70**, 123–137.

Duinker J.C. (1980) Suspended matter in estuaries: Adsorption and desorption processes. In *Chemistry and Biogeochemistry of Estuaries*, Olausson E. and Cato I. (Eds.), John Wiley & Sons, Ltd: pp. 121–151.

Dupré B., Gaillardet J., Rousseau D. and Allegré C. (1996) Major and trace elements of river-borne material: The Congo Basin. *Geochimica Cosmochimica Acta*, **60**, 1301–1321.

Durum W.H., Heidel G. and Tison L.J. (1960) Worldwide runoff of dissolved solids. *International Association for Hydrological Science*, **51**, 618–628.

Dzombak D.A. and Morel F.M.M. (1990) *Surface Complexation Modeling: Hydrous Ferric Oxide*, Wiley.

Eisma D. (1978) Geobiochemcial investigations in the Zaire River, estuary and plume. *Netherland Journal of Sea Research*, **12**, 255–420.

Esser G. and Kohlmaier G.H. (1991) Modeling terrestrial sources of N, P and S and organic carbon to rivers. In *Biogeochemistry of Major World Rivers*, Vol. 42, Degens E.T., Kempe S. and Richey J.E. (Eds.), SCOPE, John Wiley & Sons: pp. 297–322.

Förstner U. and Wittman G.T.W. (1979) *Metal Pollution in the Aquatic Environment*, Springer-Verlag: Berlin, p. 486.

Förstner U., Ahlf W. and Calmano W. (1989) Studies on the transfer of heavy metals between sedimentary phases with a multi-chamber device: Combined effects of salinity and redox variation. *Marine Chemistry*, **28**, 145–158.

Frater A. (1987) *Great Rivers of the World*, Little Brown and Co.: p. 160.

Gad M.A. and LeRiche H.H. (1966) A method for separating the detrital and non-detrital fractions of trace elements in reduced sediments. *Geochimica et Cosmochimica Acta*, **30**, 841–846.

GESAMP (1993) Anthropogenic influences on sediment discharge to the coastal zone and environmental; consequences, *UNESCO-IOC*, Vol. 52, pp. 1–67.

Gibbs R.J. (1973) Mechanisms of trace-metal transport in rivers. *Science*, **18**, 71–73.

Gibbs R.J. (1977) Transport phases of transition metals in the Amazon and Yukon Rivers. *Geological Society of America Bulletin*, **88**, 829–843.

Goldberg E.D. (1954) Marine geochemistry I: Chemical scavengers of the sea. *Journal of Geology*, **62**, 249–265.

Goldberg E.D., Koide M., Bertine K., Hodge V., Stallard M., Martincic D., Mikac N., Brancia M. and Abaychi J.K. (1988) Marine geochemistry 2: Scavenging redux. *Applied Geochemistry*, **3**, 61–571.

Guy R.D., Chakrabarti C.L. and McBain D.C. (1978) An evaluation of extraction techniques for the fractionations of copper and lead in model sediment systems. *Water Research*, **12**, 21–24.

Harmon R.S. and Ivanovich M. (1993) *Uranium-Series Disequilibrium: Applications to Earth, Marine, and Environmental Sciences*, *Second Edition* Ivanovich M. and Harmon R.S. (Eds.), Oxford Science Publications: Clarendon, pp 910.

Holland H.D. (1978) *Chemistry of the Atmosphere and the Oceans*, John Wiley & Sons: New York.

Holmes R.M., McClelland J.W., Peterson B.J., Shiklomanov I.A., Shiklomanov A.I., Zhulidov A.V., Gordeev V.V. and Bobrovitskaya N.N. (2002) A circumpolar perspective on fluvial sediment flux to the Arctic Ocean. *Global Biogeochemical Cycles*, **16**, 451–4514.

Hooper R.P., Aulenbach B.T. and Kelly V.J. (2001) The national stream quality accounting network: A flux based approach to monitoring the water quality of large rivers. *Hydrological Processes*, **15**, 1089–1106.

Horowitz A.J., Elrick K.A. and Smith J.J. (2001) Annual suspended sediment and trace element fluxes in the Mississippi, Colorado, and Rio Grande drainage basins. *Hydrological Processes*, **15**, 1169–1207.

Howarth R.W., Sharpley A. and Walker D. (2002) Sources of nutrient pollution to coastal waters in the United States: Implications for achieving coastal water goals. *Estuaries*, **25**, 656–676.

Hsi C.D. and Langmuir D. (1985) Adsorption of uranyl onto ferric oxyhydroxides: Applications of the surface complexation site-binding model. *Geochimica et Cosmochimica Acta*, **49**, 1931–1941.

Irion G. (1991) Minerals in rivers. In *Biogeochemistry of Major World Rivers*, Vol. 12, Degens E.T., Kempe S. and Richey J. (Eds.), SCOPE, John Wiley & Sons: pp. 265–281.

Ittekot V. and Laane R.W.P.M. (1991) Fate of riverine particulate organic matter. In *Biogeochemistry of Major World Rivers*, Vol. 42, Degens E.T., Kempe S. and Richey J. (Eds.), SCOPE, John Wiley & Sons: pp. 233–243.

Jenne E.A. (1977) Trace element sorption by sediments and soils – sites and processes. In *Molybdenum in the Environment*, Chappel W.R. and Peterson K. (Eds.), Marcel Dekker, Inc.: New York, pp. 425–553.

Langebein W.B. and Schumm S.A. (1958) Yield of sediment in relation to mean annual precipitation. *Transactions-American Geophysical Union*, **39**, 1076–1084.

Livingston D.A. (1963) Chemical composition of rivers and lakes. G. Data of geochemistry. *United States Geological Survey Professional Papers*, **440G**, 1–64.

Luoma S.N. and Jenne E.A. (1976) Estimating bioavailability of sediment-bound trace metals with chemical extractants. *Trace Substances and Environmental Health*, **10**, 343–351.

Luoma S.N. and Bryan G.W. (1981) A statistical assessment of the form of trace metals in oxidized estuarine sediments employing chemical extractions. *Science of the Total Environment*, **17**, 165–196.

Luoma S.N. and Davis J.A. (1983) Requirements for modeling trace metal partitioning in oxidized estuarine sediments. *Marine Chemistry*, **12**, 159–181.

Malo B.A. (1977) Partial extraction of metals from aquatic sediments. *Environmental Science and Technology*, **11**, 277–282.

Martin J.M. and Meybeck M. (1979) Elemental mass-balance of material carried by major world rivers. *Marine Chemistry*, **7**, 173–206.

Martin O. (1988) Flux of particulate inorganic matter through the Niger River into the Atlantic Ocean. *Netherland Journal of Sea Research*, **22**, 91–97.

Mayer L.M., Schick L.L. and Chang C.A. (1984) Incorporation of trivalent chromium into riverine and estuarine colloidal material. *Geochimica et Cosmochimica Acta*, **48**, 1717–1722.

McIsaac G.F., David M.B., Gertner G.Z. and Goolsby D.A. (2001) Nitrate flux in the Mississippi River. *Nature*, **414**, 166–167.

McKee B.A., DeMaster D.J. and Nittrouer C.A. (1987) Uranium geochemistry on the Amazon Shelf: Evidence for uranium release from bottom sediments. *Geochimica et Cosmochimica Acta*, **51**, 2779–2786.

Meade R.H. (1995) Contaminants in the Mississippi River, 1987–1992. *United States Geological Survey Circular*, Vol. 1133, 1–140.

Meade R.H. (1996) River-sediment inputs to major deltas. In *Sea-level Rise and Coastal Subsidence: Causes Consequences and Strategies*, Milliman J.D. and Haq B.U. (Eds.), Kluwer Academic Publishers: Dordrecht, pp. 63–85.

Meade R.H., Dunne T., Richey J.E., Santos U.D.M. and Salati E. (1985) Storage and remobilization of suspended sediment in the lower Amazon River of Brazil. *Science*, **228**, 488–490.

Means J.L., Crerar D.A., Borsik M.P. and Duguid J.O. (1978a) Radionuclide adsorption by manganese oxides and implications for radioactive waste disposal. *Nature*, **274**, 44–47.

Means J.L., Crerar D.A., Borsik M.P. and Duguid J.O. (1978b) Adsorption of Co and selected actinides by Mn and Fe oxides in soils and sediments. *Geochimica et Cosmochimica Acta*, **42**, 1763–1773.

Meybeck M. (1982) Carbon, nitrogen and phosphorus transport by world rivers. *American Journal of Science*, **282**, 401–450.

Meybeck M. (1993) Riverine transport of atmospheric carbon: Sources, global typology and budget. *Water, Air and Soil Pollution*, **70**, 443–463.

Meybeck M. and Helmer R. (1989) The quality of rivers: From pristine stage to global pollution. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **75**, 283–309.

Milliman J.D. (2001) Delivery and fate of fluvial water sediment to the sea: A marine geologist's view of European rivers. *Scientia Marina*, **65**, 121–131.

Milliman J.D. and Meade R.H. (1983) Worldwide delivery of river sediment to the oceans. *Journal of Geology*, **91**, 1–21.

Milliman J.D. and Syvitski J.P.M. (1992) Geomorphic/tectonic control of sediment discharge to the ocean: The importance of small mountainous rivers. *Journal of Geology*, **100**, 525–544.

Milliman J.D., Fransworth K.L. and Albertin C.S. (1999) Flux and fate of fluvial sediments leaving large islands in the East Indies. *Journal of Sea Research*, **41**, 97–107.

Moore R.M., Burton J.D., Williams P.J., Le B. and Young M.L. (1979) The behavior of dissolved organic material, iron and manganese in estuarine mixing. *Geochimica et Cosmochimica Acta*, **43**, 919–926.

Moore T.R. (1989) Dynamics of dissolved organic carbon in forested and disturbed catchments, Westland, New Zealand 1. Maimai. *Water Resources Research*, **25**(6), 1321–1330.

Nixon S.W., Ammerman J.W., Atkinson L.P., Berounsky V.M., Billen G., Boicourt W.C., Boynton W.R., Church T.M., Ditoro D.M., Elmgren R., et al (1996) The fate of nitrogen and phosphorus at the land-sea margin of the North Atlantic Ocean. *Biogeochemistry*, **35**, 141–180.

Oakley S.M., Nelson P.O. and Williamson K.J. (1981) Model of trace-metal partitioning in marine sediments. *Environmental Science and Technology*, **15**, 474–480.

Olsen C.R., Cutshall N.H. and Larsen I.L. (1982) Pollutant-particle associations and dynamics in coastal marine environments: A review. *Marine Chemistry*, **11**, 501–533.

Onstad G.D., Canfield D.E., Quay P.D. and Hedges J. (2000) Sources of particulate organic matter in rivers from the continental USA: Lignin phenol and stable carbon isotope compositions. *Geochimica et Cosmochimica Acta*, **64**, 3539–3546.

Plater A.J., Ivanovich M. and Dugdale R.E. (1992) Uranium series disequilibrium in river sediments and waters: The significance of anomalous activity ratios. *Applied Geochemistry*, **7**, 101–110.

Poulton S.W. and Raiswell R. (2000) Solid phase associations, ocean fluxes and the anthropogenic perturbation of transition metals in world river particles. *Marine Chemistry*, **72**, 17–31.

Presley B.J., Trefry J.H. and Shokes R.F. (1980) Heavy metal inputs to Mississippi delta sediments. *Water, Air, and Soil Pollution*, **13**, 481–494.

Ramesh R., Purvaja R., Ramesh S. and James R.A. (2002) Historical pollution trends in coastal environments of India. *Environmental Monitoring and Assessment*, **79**, 151–176.

Raoux C., Bayona J.M., Miquel J.-C., Teyssie J.-T., Fowler S.W. and Albaigés D. (1999) Particulate fluxes of aliphatic and aromatic hydrocarbons in near-shore waters to the Northwestern Mediterranean Sea, and the effect of continental runoff. *Estuarine, Coastal and Shelf Science*, **48**, 605–616.

Rashid M.A. and Leonard J.D. (1973) Modifications in the solubility and precipitation behavior of various metals as a result of their interaction with sedimentary humic acid. *Chemical Geology*, **11**, 89–97.

Santschi P.H., Bajo C., Mantavani M., Orciuolo D., Cranston R.E. and Bruno J. (1988) Uranium in pore waters from North Atlantic (GME and Southern Nares Abyssal Plain) sediments. *Nature*, **331**, 155–157.

Sarin M.M., Sudheer A.K. and Balakrishna S. (2002) Significance of riverine carbon transport: A case study of a large tropical river, Godavari (India). *Science in China*, **45**, 97–108.

Schlosser P., Newton R., Ekwurzel B., Khatiwala S. and Mortlock R. (2002) Decrease of river runoff in the upper waters of the Eurasian Basin, Arctic Ocean, between 1991 and 1996: Evidence from δ^{18}O data. *Geophysical Research Letters*, **29**, 31–35.

Shannon R.D. and White J.R. (1991) The selectivity of a sequential extraction procedure for the determination of iron oxyhydroxides and iron sulfides in lake sediments. *Biogeochemistry*, **14**, 193–208.

Sharp J.H., Culberson C.H. and Church T.M. (1982) The chemistry of the Delaware estuary. General considerations. *Limnology and Oceanography*, **27**, 1015–1028.

Shiller A.M. (1993) Comparison of nutrient and trace element distributions in the delta and shelf outflow regions of the Mississippi and Atchafalaya Rivers. *Estuaries*, **16**, 541–546.

Shiller A.M. and Boyle E.A. (1987) Dissolved vanadium in rivers and estuaries. *Earth and Planetary Science Letters*, **86**, 214–224.

Sholkovitz E.R. (1976) Flocculation of dissolved organic and inorganic matter during mixing of river water and seawater. *Geochimica et Cosmochimica Acta*, **40**, 831–845.

Sholkovitz E.R. (1993) The geochemistry of rare earth elements in the Amazon River estuary. *Geochimica et Cosmochimica Acta*, **57**, 2181–2190.

Sholkovitz E.R., Landing W.M. and Lewis B.L. (1994) Ocean particle chemistry: The fractionation of rare earth elements between suspended particles and seawater. *Geochimica et Cosmochimica Acta*, **58**, 1567–1579.

Singh S.K. and Subranmanian V. (1984) Hydrous Fe and Mn oxides – scavengers of heavy metals in the aquatic environment. *CRC Critical Reviews*, **14**(1), 33–90. No.

Skei J. and Paus P.E. (1979) Surface metal enrichment and partitioning of metals in a dated sediment core from a Norwegian fjord. *Geochimica et Cosmochimica Acta*, **43**, 239–246.

Slavek J., Wold J. and Pickering W.F. (1982) Selective extraction of metal ions associated with humic acids. *Talanta*, **29**, 743–749.

Stone A.T. and Morgan J.J. (1984) Reduction and dissolution of manganese(III) and manganese(IV) oxides by organics. 1. Reaction with Hydroquinone. *Environment Science and Technology*, **18**, 450–456.

Stallard R.F. (1980) *Major Element Geochemistry of the Amazon River system*, Vol. 28, Woods Hole Oceanographic Institute No. 80, MIR-WHOI: 1–362.

Stumm W. (1992) *Chemistry of the Solid-Water Interface*, John Wiley & Sons: p. 428.

Stumm W. and Morgan J.J. (1981) *Aquatic Chemistry: An Introduction Emphasizing Chemical Equilibria in Natural Waters*, John Wiley & Sons.

Sulzberger B., Suter D., Siffert C., Banwart S. and Stumm W. (1989) Dissolution of Fe(III)(hydr)oxides in natural waters; Laboratory Assessment on the kinetics controlled by surface coordination. *Marine Chemistry*, **28**, 127–144.

Swarzenski P.W., McKee B.A. and Booth J.G. (1995) Uranium geochemistry on the Amazon Shelf: Chemical phase partitioning and cycling across a salinity gradient. *Geochimica et Cosmochimica Acta*, **59**, 7–18.

Swarzenski P.W., McKee B.A., Sørensen K. and Todd J.F. (1999) ^{210}Pb and ^{210}Po, manganese and iron cycling across the O_2/H_2S interface of a permanently stratified Fjord: Framvaren, Norway. *Marine Chemistry*, **67**, 199–217.

Swarzenski P.W., Porcelli D., Andersson P.S. and Smoak J.M. (2003) The behavior of U- and Th-series nuclides

in the estuarine environment. *Reviews in Mineralogy and Geochemistry*, **52**, 577–606.

Swarzenski P.W., Campbell P.L., Porcelli D. and McKee B.A. (2004) The estuarine geochemistry and isotope systematics of ^{234,238}U, in the Fly and Amazon Rivers. *Continental Shelf Research*, **24**, 2357–2372.

Syvitski J.P.M. (2002) Sediment discharge variability in Arctic rivers: Implications for a warmer future. *Polar Research*, **21**(2), 323–330.

Tappin A.D. (2002) An examination of the fluxes of N and P in temperate and tropical estuaries: Current estimates and uncertainties. *Estuarine, Coastal and Shelf Science*, **55**, 885–901.

Tessier A., Campbell P.G.C. and Bisson M. (1979) Sequential extraction procedure for the speciation of particulate trace metals. *Analytical Chemistry*, **51**, 844–851.

Thurman E.M. (1985) *Organic Geochemistry of Natural Waters*, Kluwer Academic Publishers Group: p. 496.

Trefry J.H., Nelson T.A., Trocine R.P., Metz S. and Vetter T.W. (1986) Trace metal fluxes through the Mississippi River delta system. *Rapport et Procs-verbaux des Reunion Commission internationale pour l'exploration scientifique de la Mer*, **186**, 277–288.

Turekian K.K. (1977) The fate of metals in the oceans. *Geochimica et Cosmochimica Acta*, **41**, 1139.

Waite T.D. (1990) Photo-redox processes at the mineral-water interface. *Reviews in Mineralogy*, **23**, 559–603.

Waite T.D., Davis J.A., Payne T.E., Waychunas G.A. and Xu N. (1994) Uranium(VI) adsorption to ferrihydrite: Application of a surface complexation model. *Geochimica et Cosmochimica Acta*, **58**, 5465–5478.

Walling D.E., Owens P.N., Carter J., Leeks G.J.L., Lewis S., Meharg A.A. and Wright J. (2003) Storage and sediment-associated nutrients and contaminants in river channel and floodplain systems. *Applied Geochemistry*, **18**, 195–220.

White A. and Blum A.E. (1995) Effects of climate on chemical weathering in watersheds. *Geochimica et Cosmochimica Acta*, **59**, 1729–1747.

Van Valin R. and Morse J. (1982) An investigation of methods commonly used for the selective removal and characterization of trace metals in sediments. *Marine Chemistry*, **11**, 535–564.

Vörösmarty C.J. and Fekete B.M. (2002) Global system of rivers: Its role in organizing continental land mass and defining land-to-ocean linkages. *Global Biogeochemical Cycles*, **14**, 599–621.

Yeats P.A. and Brewer J.M. (1982) Discharge of metals from the St. Lawrence River. *Canadian Journal of Earth Science*, **19**, 982–992.

Zhang J. (1995) Geochemistry of trace metals from Chinese river/estuary systems: An overview. *Estuarine, Coastal Shelf Science*, **41**, 631–658.

Zhang J. and Liu C.L. (2002) Riverine composition and estuarine geochemistry of particulate metals in China – Weathering features, anthropogenic impact and chemical fluxes. *Estuarine, Coastal and Shelf Science*, **54**, 1051–1070.

Zinder B., Furrer G. and Stumm W. (1986) The coordination chemistry of weathering: II. Dissolution of Fe(III) oxides. *Geochimica Cosmochimica Acta*, **50**, 1861–1869.

90: Lake Sediments as Records of Past Catchment Response

JOHN DEARING

Department of Geography, University of Liverpool, Liverpool, UK

The sediments that accumulate at the beds of lakes are natural archives of information about catchment processes that extend back far further than instrument records. This article reviews the kinds of questions that long records can address, and summarises the technical and analytical approaches taken. Records of river discharge, sediment delivery, and soil erosion obtained in this way over decades and centuries are increasingly used to understand the behaviour of fluvial systems, especially the combined effects of human activity and climate. Case-studies now available for different parts of the world can be used to provide the temporal perspective to the current levels of fluvial and erosional processes in human-dominated landscapes, and together show the strong dependence of sediment delivery on catchment size.

LAKE SEDIMENTS AS SUITES OF ERODED SOIL

"Lake sediments as suites of eroded soil" was a phrase coined by John Mackereth, a British scientist working at the Freshwater Biological Association in the English Lake District, in the 1960s. He was the first scientist to articulate the potential for lake sediments to provide a unique source of information about past hydrological and sediment responses in the surrounding catchment. His 1966 paper describes the physical and geochemical properties of 6-m-long cores sampled from different lakes in the Lake District and a number of arguments for catchment-specific responses in weathering processes and soil erosion through the Holocene (Mackereth 1966). The ensuing 35 years has seen a steady rise in the number of studies engaged in reconstructing hydrological responses to climate and human activities, and today there are few zones of the world where lake sediments have not contributed to our understanding of the rate, causes and trajectory of soil erosion, flooding, or mass movements.

Drainage basins are complex cascading systems where responses in the fluvial and sediment systems to human actions or meteorological events may reverberate over large distances with significant time lags. The effect of these responses has the potential to alter agroecosystem sustainability, flood regimes, hazard risk, and even to feedback into global climates through the modification of oceanic biogeochemical cycles. The impact on global water budgets and sediment flux through future human activities, especially adaptive activities responding to rising population numbers and climate change, now demands information about response characteristics of hydro-geomorphic systems across a wide range of spatial and temporal scales. In particular, unraveling the interacting roles of climate and human impact on sediment flux may require information about process-response mechanisms that take place over long timescales (cf. Oldfield and Dearing, 2003). The following describes the scientific questions that lake sediment studies can address before reviewing lake sediment studies of past catchment responses.

SCIENTIFIC QUESTIONS

Measuring flows of water and sediment with instruments is the conventional method for monitoring hydro-geomorphic processes over time. Large numbers of gauging station records exist for river discharge but not normally for longer than a few decades. One recent review of global rivers (Walling and Fang, 2003) shows that suspended sediment records normally extend back only as far as the 1950s. Historical reconstructions that combine land use records, field monitoring, and modeling provide detailed insights

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

into catchment dynamics (e.g. Trimble, 1999), but are rare and restricted to small basins. Unfortunately, short-term monitoring records, however detailed, often represent an obstacle to the assessment of impacts of climate and human activities on hydro-geomorphic processes that operate over longer timescales. Long-term perspectives gained from using lake sediments to reconstruct hydrological processes offer the opportunity to address a number of important questions for catchments worldwide:

Are processes at the present time responding to natural climate variability or have they been modified by human activities? This is especially important to answer as we move into a period of global warming and, in many parts of the world, increased pressure on landscapes for farming and urbanization.

How large has the impact of human activities been on processes compared with the preimpact natural levels? If we want to identify catchment management or soil conservation strategies, it helps to know how far the system has moved from its "natural" state.

How has the coupling of subsystems within a catchment changed through time? The degree and extent to which subsystems in major drainage basins are coupled is often only poorly known. Major issues about the effects of land use on downstream locations, such as where sediment in reservoirs originates, or what is the contribution of surface soil to a river's total suspended load, are important questions to answer.

How are all these long-term changes a function of the size of the catchment and its subsystems? Impact–response relationships in catchments are partly a function of catchment size, with sediment storage modulating the record of sediment transport through a fluvial system: we need to have more information on what these relationships are.

How do we test predictive models that simulate future changes over decades and centuries? Reconstructed hydrological records open up the possibility of testing and validating simulation models over longer periods than the last 50–70 years provided by monitored records.

METHODS AND TECHNIQUES

The strength of lake sediment studies lies in the accumulation of particles that are derived from the surrounding catchment, allowing the coupled lake-catchment system to be used as a fundamental unit of study (Oldfield, 1977). Careful examination of the visual, microscopic, or measured sediment properties allows past environmental processes and conditions to be reconstructed, an indirect method that uses the sediment components as environmental indicators or *proxies* of past conditions. Lakes and reservoirs cover $\sim 3.1 \times 10^6\,km^2$ or 2.3% of the land surface and drain ~20% of the land surface. Freshwater bodies vary enormously in their size and form, from recently created small ponds and farm dams less than 1 ha in area and a couple of meters deep, to the 30 million year old Lake Baikal, in Siberia, which at over 1600 m deep contains the world's largest volume of freshwater. The large variations in lake and catchment size have led to the development of many different sampling and analytical procedures, covered by several reviews (Dearing, 1986; Lowe and Walker, 1997; Roberts, 1998; Smol, 2002). Crucial to lake sediment reconstructions of hydrological processes are sediment chronology and analytical measurements used as proxy data.

Chronology

Typically, around 5–15 m of mud has accumulated at the bottom of lakes during the Holocene, the geological period covering the past 11 000 years. In some lakes, the mud is deposited with a seasonal cycle that gives rise to visible layers known as *laminations* which afford a means of dating through counting the seasonal layers from the present to find the sediment age for any particular depth. But in the majority of sedimentary archives, dating relies on analytical procedures such as ^{14}C dating of organic remains (~40 000–300 year BP), optically stimulated luminescence dating of quartz or feldspar grains (~1 000 000–100 year BP) and ^{210}Pb dating that gives sediment ages over the past 100–150 years. Other dating techniques utilize pollution markers like the radioactive fall-out of ^{137}Cs and ^{241}Am from the Chernobyl accident (1986) and early bomb tests (1963), or other markers in the sediment that have a known and often local date, such as pollen linked to the introduction of an exotic plant species or tephra layers from known volcanic eruptions. Dating is the most important aspect of lake sediment studies but also the most problematic (Lowe and Walker, 1997; Roberts, 1998). Old carbon, hardwater effects, insufficient bleaching, postdepositional mobility of ^{137}Cs are all potential problems that need to be excluded or avoided by careful site and sample selection.

Proxy Measurements

The reconstruction of past environments is never perfect, but the ingenuity of paleoecologists has led to a surprisingly large number of environmental states and processes that may be successfully reconstructed (Lowe and Walker, 1997). Several hydrological processes may be reconstructed (Table 1), and climate, vegetation, and soils may also be inferred through a combination of paleoenvironmental analyses. Sediment accumulation rates in dated cores, particularly annually laminated sediments, can provide an estimate of sediment load to the lake bed, though sediment focusing may mean that basinwide accumulations derived from multiple cores are needed for greatest accuracy. Particle-size, and some magnetic properties dependent on particle-size, can be used as a proxy for flood intensity, especially in sediment cores sampled from zones of the lake bed influenced by inflowing rivers and streams. Most

Table 1 Examples of hydrological processes recorded in lake sediment

Response	Proxy evidence
Flood events	Stratigraphy, particle-size, magnetic susceptibility
Water balance	
Salinity	Diatoms
Lake levels	Stratigraphy
Sediment flux	
Accumulation rates	^{14}C, ^{210}Pb, ^{137}Cs, laminations, biostratigraphy
Sediment properties/sources	
Minerogenic particulates	LOI, geochemistry, particle-size, mineral magnetics
Organic particulates	LOI, organic chemistry, isotopes
Nutrient flux	
Phosphorus	Diatoms, geochemistry
pH	Diatoms

measurements used for reconstructing hydro-geomorphic processes require some degree of calibration with known or monitored changes. In some studies, the proxy in modern sediments is calibrated against modern conditions or monitored records to produce a statistical *transfer function* that may then be applied to earlier sediments; in others the sediment property is matched to material in the catchment to give a fingerprint of a particular type, zone, or horizon of soil, and substrate.

LAKE SEDIMENT RECORDS OF CATCHMENT PROCESSES

Holocene Trends

Small lakes draining small-medium catchments ($<10^3$ km^2) provide the largest number of reconstructed hydrological records. These records give information about long-term trends (10^2–10^3 year) and short-term responses (10^0 year) in the sediment flux record, changes in sediment source and often information about river discharge and links between forcings and responses. Long proxy records of sediment flux (measured as sediment load to the lake) found in previously glaciated parts of Central and Northern Europe and North America (Dearing, 1994) tend to show high pre-Holocene rates declining to minimum values during the early to mid-Holocene, suggesting a shift from high energy-high sediment supply to low energy-low sediment supply conditions (Figure 1a and b). In contrast, mid-to-late-Holocene trends appear to differ according to the level of human activity in the drainage basin. In North America, the rise in sediment flux is much later than in Europe and corresponds to the introduction of European agricultural methods within the past 350 years. Older mid-Holocene peaks in lake sediment accumulation in North America are rare and may actually represent artifacts of uncalibrated radiocarbon dating (Webb and Webb, 1988). In Central and Northwest Europe, later peaks in sediment load often register Bronze Age, Iron Age, Medieval, and Modern phases of farming, and these are comparable to the history of pollen evidence (% nonarboreal pollen – NAP) for open land (Figure 1b) and erosion forcings deduced from geomorphology, archaeology, and environmental history (Figure 1c). Edwards and Whittington's (2000) review of 50, largely upland, sites in Britain and Ireland shows that there is pollen evidence for anthropogenic impact at virtually all sites where accelerated sedimentation occurs. The start of accelerated accumulation varies from site to site with dates clustered around 5295–4970, 4530–4235, and 2980–2810 BP. At lowland sites too, for example, at Llangorse Lake, in South Wales, sediment accumulation rates rise ~4–13-fold following major deforestation phases after ~5000 BP and ~4-fold after the mid-nineteenth century (Jones *et al.*, 1985). At Blelham Tarn, in the English Lake District, the recent rises in sedimentation rate (~10-fold since the 1950–1960s) are positively correlated with high levels of grazing in the catchment (van der Post *et al.*, 1997). Erosion records in Central America show the initial impact of maize cultivation by indigenous agriculturalists ~4–5000 BP at both lowland (Goman and Byrne, 1998) and high altitude (Bradbury, 2000) locations in Mexico, followed by a second phase of erosion (between 2600 and 1600 BP) in the lowlands linked to drier conditions (Goman and Byrne, 1998). The study by O'Hara *et al.* (1993) at Lake Pátzcuaro in the Michoacán Highlands identified erosion early as 3640–2500 BP well before the intense erosion of the Preclassic/early Classic periods where erosion rates were double those under preimpact conditions. In contrast to the lowlands, catchment stability was related to reduced agriculture brought on by drought conditions. Accelerated erosion in the later Postclassic period (850–350 BP), predating European contact, increased average rates to 5–6 times the preimpact level. At some sites, both response and recovery trends are observed in the erosion record, as in the history of Mayan agriculture in Guatemala (Deevey *et al.*, 1979; Curtis *et al.*, 1998). Records from Southeast Asia are rare, but evidence from two small lake-catchments in subtropical Yunnan, China, show sedimentation rates 13–15 times above the predisturbance levels (Whitmore *et al.*, 1994). In New Zealand, lake sediment studies (Page and Trustrum, 1997; Eden and Page, 1998) show mean accumulation rates increasing by ~5-fold under European pasture compared with rates under Polynesian occupation. Polynesian deforestation led to a landscape dominated by bracken fern/Coriara vegetation, but mean accumulation rates were less than twice those under pre-Polynesian indigenous forest. The studies illustrate the complex interactions between

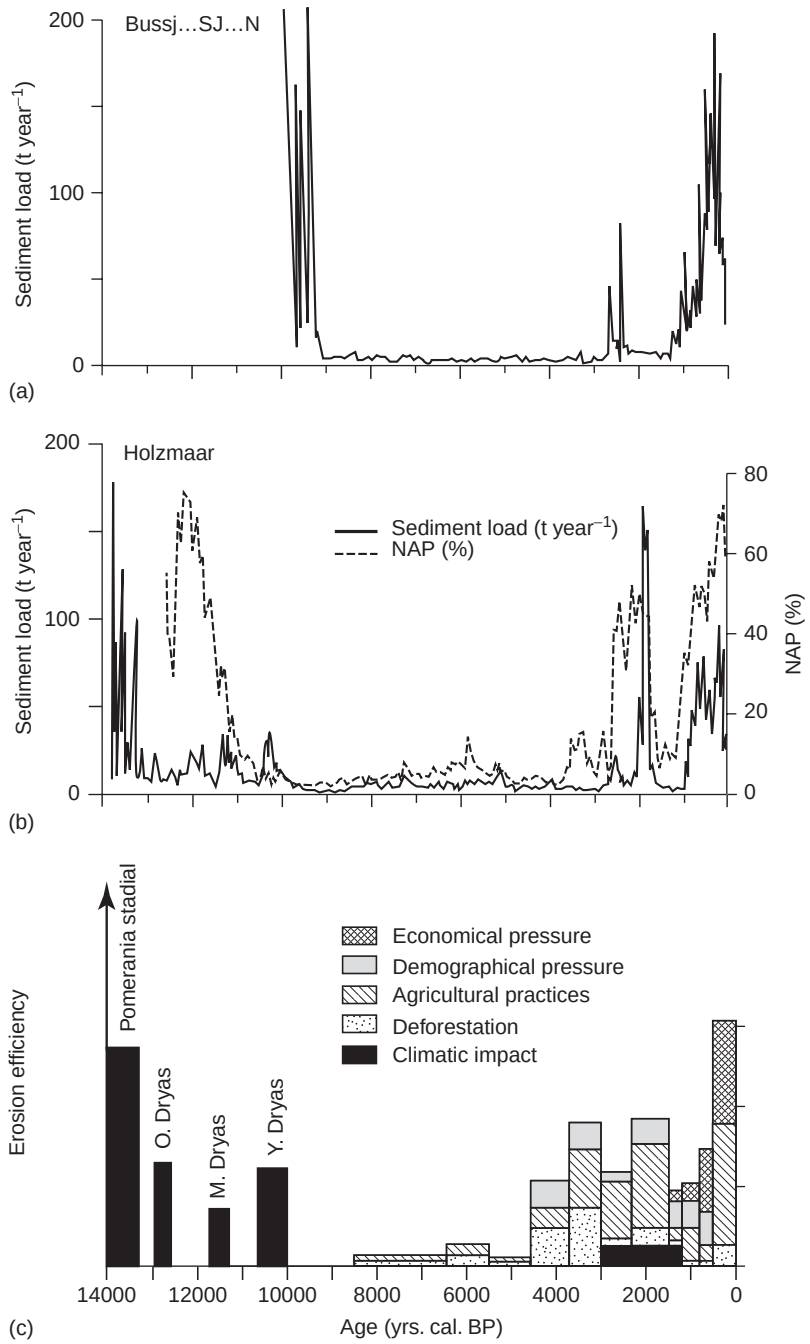


Figure 1 Holocene sediment flux in Northwest Europe: (a) Bussjössjön, southern Sweden (Reproduced from Dearing *et al.*, 1990, by permission of John Wiley & Sons Ltd); (b) Holzmaar, Germany, showing also nonarboreal pollen proxy for open land (Reprinted from Zolitschka 1998, © 1998, with permission from Elsevier); (c) history of erosion forcings deduced from geomorphological investigations and environmental history (Reproduced from Van Vliet–Lanoë *et al.*, 1992, by permission of Oxbow Press, Oxford)

land use, earthquakes, and high-magnitude storm events that trigger land slides.

Recent Records

Detailed records of recent sediment accumulation rates period exist for many sites, usually based on a ^{210}Pb chronology covering the last 100 years, often show strong correlations between sediment accumulation and land use (Figure 2). At Frain's Lake, Michigan, (Davis, 1976) the sedimentation record shows how the levels rise ~30-fold following deforestation and stabilize at a far lower level

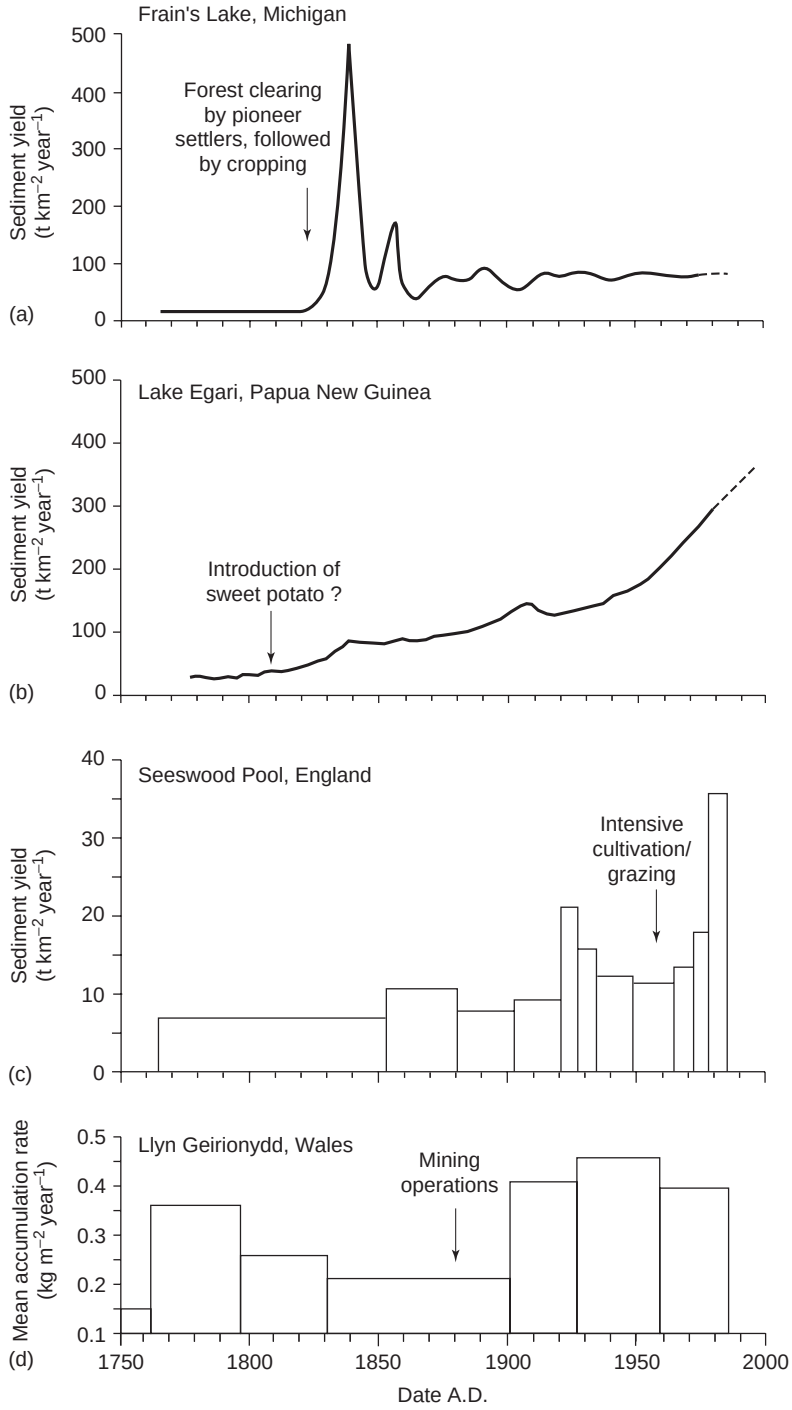


Figure 2 Recent responses to human impact in small catchments: (a) Frain's Lake, Michigan (Reproduced from Davis, 1976, by permission of Cambridge Unviersity Press); (b) Lake Egari, Papua New Guinea (Reproduced by permission of Oldfield *et al.*, 1985); (c) Seeswood Pool, England (Reproduced from Dearing and Foster, 1987, by permission of John Wiley & Sons Ltd); (d) Llyn Geirionydd, Wales (Reproduced from Dearing, 1992, by permission of John wiley & Sons Ltd)

than recorded for the initial clearance phase, but still higher than preclearance levels (Figure 2a). In the Papua New Guinean Highlands, Oldfield *et al.* (1985) show that sediment flux continues to rise into the late twentieth century following clearance and the introduction of sweet potato in the nineteenth century (Figure 2b). This suggests that sediment supply continues to increase at the present time and long-term agricultural sustainability may

be difficult to achieve. Changes in agricultural methods in lowland England since 1950 towards more intensive cultivation, and especially more intense cattle grazing, appear to have been the main cause of higher rates of sediment accumulation at Seeswood Pool (Figure 2c) (Dearing and Foster, 1987). In South America, Cisternas *et al.* (2001) argue that 15-fold rises in sediment accumulation rates in a lake in Central Chile, since the 1800s, are associated not only with losses of forest but also with replacement of the native forest by pine plantations. The erosional impacts of land-use change other than deforestation and farming are also clearly recorded, as seen in Wales (Figure 2d) where there is a long history of mining and quarrying (Dearing *et al.*, 1981; Dearing, 1992). In upland Scotland and Ireland, initiation of peat erosion is recorded in sediments between A.D. 1500 and 1700 and was caused either by climate change during the Little Ice Age or by an increase in the intensity of burning (Stevenson *et al.*, 1990). In the Canadian prairies, sedimentation rates reached a maximum in the 1950–1960s (De Boer, 1997) and lake sediment records seem to confirm the effectiveness of recent soil conservation schemes in reducing erosion rates.

Large Catchments and Sediment Storage

There are far fewer lakes, and consequently lake sediment records, representing large catchments (10^3–10^6 km^2). At the lower end of this range, the sediment records are familiar. The stepped rise of sediment flux in Europe, broadly paralleling human activities, is seen in the Upper Doubs River (Bichet *et al.*, 1999) in the French Jura. In North and South America, the impact of European agricultural methods in large catchments is also clear. There is, however, a tendency for records from lakes draining large catchments to exhibit lower levels of sediment flux variability. Colman *et al.* (2000) show that sedimentation rates in southern Lake Michigan have risen by 2–11-fold in the period 1840-modern compared with mid-Holocene levels, which compares with the ~30-fold increase in the smaller Frain's Lake in Michigan. In the tropics and subtropics there are few case studies to draw upon, but increases in recent sedimentation rates in the large subtropical Lake Erhai, Yunnan, China, are only 1.3 times the predisturbance values (Whitmore *et al.*, 1994), despite a mountainous catchment and well-documented deforestation and irrigated farming practices extending back at least 2000 years (Elvin *et al.*, 2002). Lakes with smaller catchments in Yunnan show records with much higher variability. Change in sediment flux is most suppressed in the largest drainage basins. At Lake Baikal, Siberia, fed by the River Selenga, rises in accumulation rate over the last 100 years are only ~2-fold and comparisons with older ^{14}C dated sequences suggest that the mean modern accumulation rates are very close to long-term averages (Appleby *et al.*, 1998), even though the main period of industrial activity and population expansion has been in the last 75 years (Flower, 1998). Estimates for the Black Sea basin suggest modern rates are lower or equal to the mean Quaternary rate. Estimates of sediment yield for the drainage basin of the Caspian Sea suggest modern rates match the long-term Quaternary average in the south basin but are higher by up to 38-fold in the north basin. Modern rates for the Aral Sea basin are also higher than long-term averages by up to 200-fold, but the sedimentation rate data from both the Aral and north Caspian Sea are strongly affected by water level changes and a significant aeolian dust input (Einsele and Hinderer, 1997).

At the global scale, "modern" monitored sediment loads are normally positively related to drainage basin area, while area-specific sediment yields are inversely related (e.g. Milliman and Syvitski, 1992). Similarly, lake sediment accumulation rates have also been found to correlate positively with catchment : lake area ratio (Blais *et al.*, 1998), while area-specific sediment yields derived from lake sediment records are inversely related to the catchment: lake area ratio (Dearing and Foster, 1993). Barlow and Thompson (2000) show that 66% of the variance in mean sediment flux estimates derived from UK lakes and reservoirs can be explained by the log of catchment area. What these sediment trends with drainage basin area do not describe, however, is the influence of spatial scale on the variability of sediment fluxes through time at any given location. An analysis of the minimum and maximum sediment accumulation rates in lakes representing 7 orders of magnitude of catchment sizes (10^{-1} to 10^6 km^2) shows that catchment size seems to be a major control on the variability of sediment delivery over long timescales (Figure 3), with the largest ratios of impact/preimpact rates found in small catchments. Despite long histories of climate and human impact, the variability of fluvial sediment transport in catchments $>10^4$ km^2 may be very low as a result of large storage on floodplains insensitive to hydrological changes (Dearing and Jones, 2003).

Calibration of Lake Records

There are some studies that attempt explicitly to use the lake sediment record as an extension to monitored records of contemporary hydro-geomorphic processes. Studies in two catchments in Midland England (Merevale Lake and Seeswood Pool: Figure 2c) showed that lake sediment-derived estimates of sediment flux are not only comparable with those derived from stream monitoring (Foster *et al.*, 1985; Dearing and Foster, 1987), but could be used as the basis for reconstructing past sediment budgets (Foster *et al.*, 1990). Direct calibration of sediment properties to meteorological or hydrological data is becoming more common where high temporal resolution in sediment properties exists (e.g. Cooper and O'Sullivan, 1998; Wohlfarth

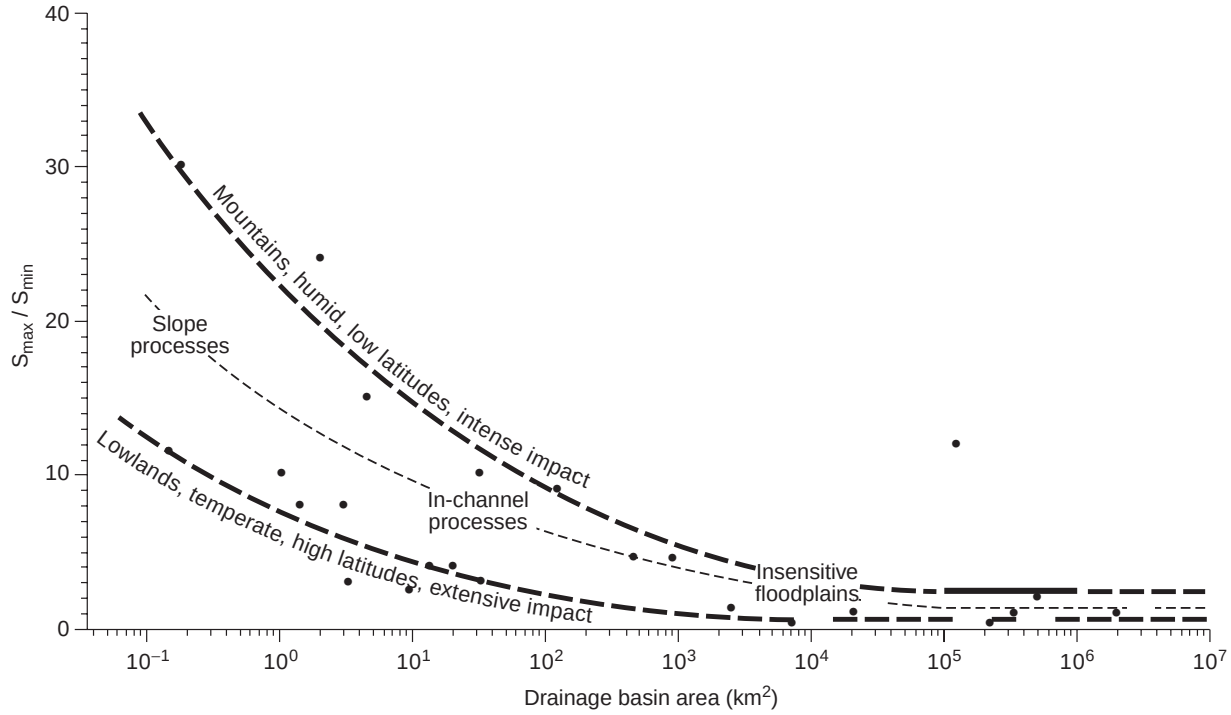


Figure 3 The relative magnitude of change between minimum, baseline, or preimpact sediment flux and maximum sediment flux (dimensionless ratio S_{max}/S_{min}) plotted against drainage basin size (km^2) from published case studies of late-Holocene sediment accumulation rates. Also shown are apparent controls on upper and lower limits, and the likely trend of increasing decoupling between slope and channel, and increasing sediment storage, as basin area of basin increases. Horizontal line (10^5–10^6 km^2) shows maximum range of estimated long-term sediment yields for large Asian rivers. The data suggest that although sediment flux at a site, and at a point in time, is controlled by the local combination of climate, land use, topography, elevation, and tectonic factors, the range of long-term variability is constrained by catchment area. The available data suggest that the upper limit of the distribution is associated with mountainous landscapes, humid climate (especially in low latitudes), and intense human impact. The lower limit is associated with lowlands, temperate climate, and extensive human impacts (Reprinted from Dearing and Jones, 2003, © 2003, with permission from Elsevier)

et al., 1998; Zolitschka, 1998; Foster *et al.* 2003). Dearing and Flower (1982) demonstrated that variations in the magnetic susceptibility of material in Lough Neagh sediment traps closely followed variations in the preceding monthly precipitation. They argued that high susceptibility was linked to the transport of relatively large particles of primary titanomagnetite in flood events, thus providing a process link between hydrological processes and the lake sediment susceptibility record. Thus Holocene lake sediment records of magnetic susceptibility that increase towards the present (Figure 4), probably reflect an increasing frequency and/or magnitude of flooding caused by climate or human-induced changes in runoff and increased sediment supply; an interpretation which is strongly supported by correlations with pollen evidence for open landscapes. If the same hydrodynamic explanation is applied to records with long-term trends in magnetic susceptibility decreasing to the present, the implications are that sediment availability and the efficiency of fluvial processes in many high latitude and upland sites have declined during the Holocene and may have been insensitive to climate change.

Changes in sediment source over time may indicate the spatial location of catchment disturbance. The Holocene record of sediment load at Bussjössjön (Figure 1) in southern Sweden (Dearing *et al.*, 1990), broadly similar to the Holzmaar record, is complemented by magnetic-based tracing of sediment sources (Dearing, 1999, 2000). The results suggest that sediment has been derived largely from channel sources, despite modern process monitoring that measures high rates of surface soil movement on surrounding arable fields. Similarly, in their study of sediment delivery in the Southern Tablelands of Southeast Australia, Wasson *et al.* (1998) use historical surveys and reservoir sedimentation rates in the Jerrabomberra Creek catchment to show that initial disturbance in the late nineteenth century, largely through sheep grazing, increased sediment yields 900-fold, which then reduced to 6-fold after the 1940s. Monitoring and tracer studies in the same landscape (e.g. Olley *et al.*, 1993) showed that channel incision and gullying were the

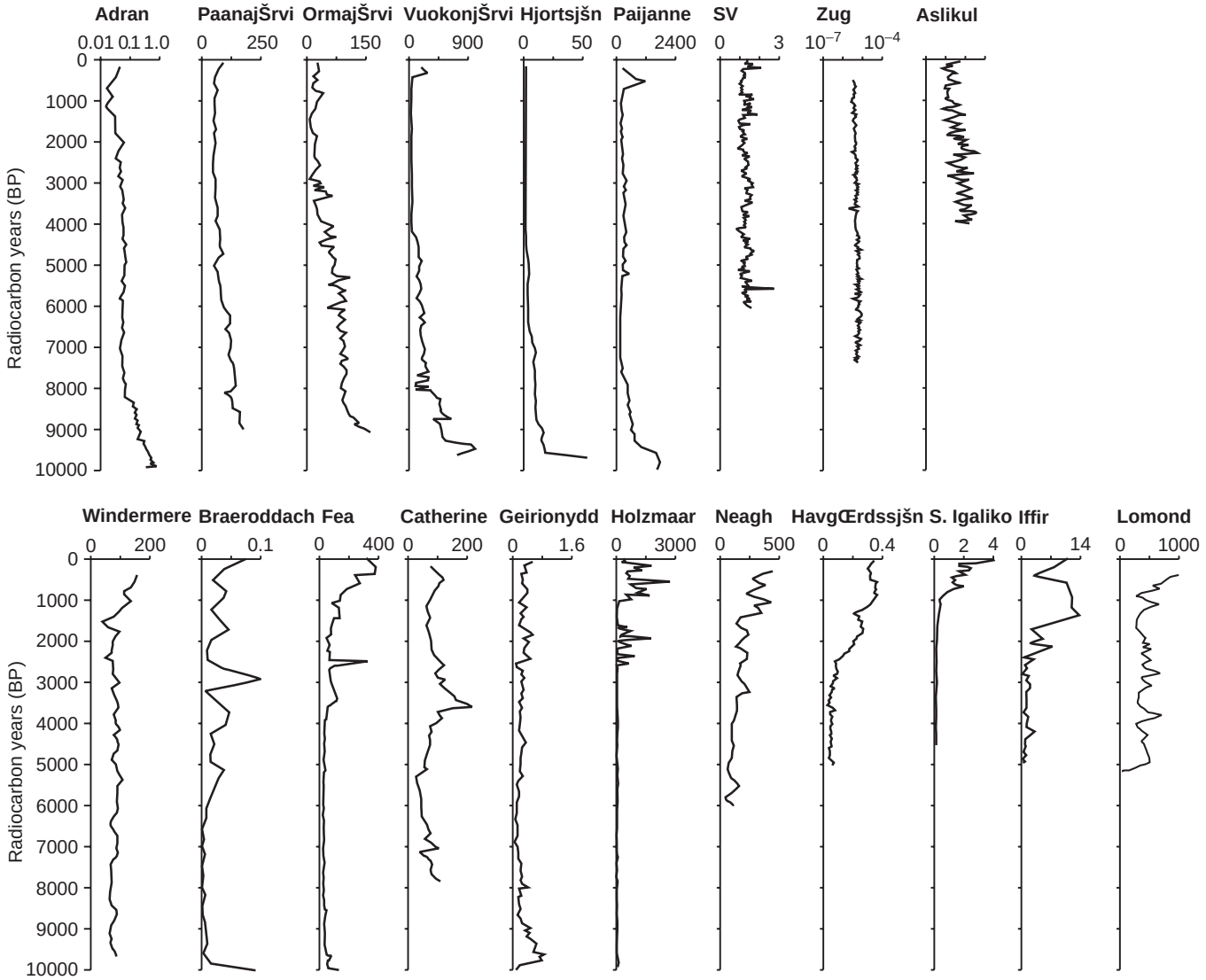


Figure 4 Holocene lake sediment records of magnetic susceptibility, used to infer river discharge. (a) Decreasing or stationary trends; (b) increasing trends from mid-to-late Holocene to the present. Sediment dating variously derived from radiocarbon, paleomagnetic secular variation and late-glacial/Holocene boundaries, with approximate timescale of 0–10 000 BP (Reproduced from Dearing, 1999, by permission of Cambridge University Press)

dominant processes of both sediment production and downstream sediment delivery.

Unraveling Natural Variability and Human Impact

In a few studies, high resolution and multiproxy analyses of sediment cores provide the means to unravel the interactions between climate and human impact. For example, the laminated sediment record from Holzmaar, Germany, (Zolitschka, 1998) shows major rises in minerogenic sediment flux dated to the periods of the Iron Age, the Roman Empire, the High Middle Ages and the seventeenth to eighteenth centuries (Figure 1b). The record from lake Bussjössjön, southern Sweden, (Dearing *et al.*, 1990), similar except for the lack of a sedimentation spike ~2000 BP attests to the fact there was no Roman occupation (Figure 1a). The overall similarity in these records, except where human impact differs, strongly indicates the overriding importance of land use. Estimated sediment yields at Holzmaar rose from $\sim 1.5\,\mathrm{t\,km^{-2}\ year^{-1}}$ before major human impact to a maximum of $49\,\mathrm{t\,km^{-2}\ year^{-1}}$ in the seventeenth and eighteenth centuries, an overall maximum increase of 33-fold over prehuman impact values. A mathematical analysis of the accumulation rate data for the period 3000–9000 cal. BP, prior to major human impact, suggests that it may be a time period across Central and Northwest Europe when the hydro-geomorphic system was in

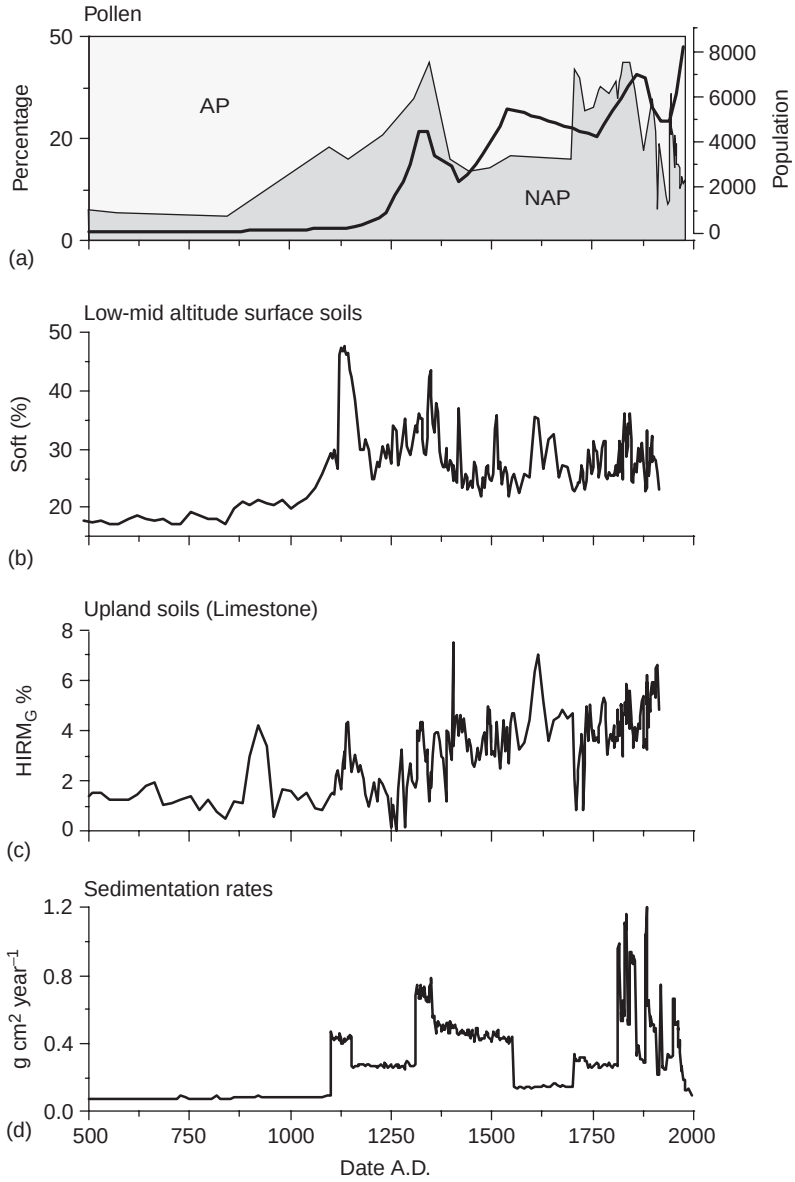


Figure 5 Erosion and land use history (~A.D. 500–2000) of the Petit Lac d' Annecy catchment, Haute Savoie, France, based on lake sediment analyses of sediment flux and sources, and historical documents: (a) arboreal (AP) and nonarboreal (NAP) pollen curves (shaded) as measures of forest cover and land openness, with human population (solid line); (b) soft magnetic remanence proxy for lowland soil; (c) high magnetic remanence proxy for upland soil; (d) sediment accumulation rate as proxy for sediment flux (Reproduced from Dearing *et al.*, 2001, With kind permission of Springer Science & Business Media)

a state of self-organized criticality where the impact of external forcings on sediment flux from the catchment was subordinate to the internal organization of processes (Dearing and Zolitschka, 1999).

At another site in Europe, Lac d'Annecy, France, the Late-Holocene record of sediment flux (Figure 5d) is also similar to Holzmaar, Bussjössjön, and other records in the region, such as the Jura (Bichet *et al.*, 1999), but with a strong erosional pulse triggered by deforestation (% nonarboreal pollen – NAP) and population expansion (Figure 5a) into high valleys during the monastic period in the tenth to thirteenth centuries. Here, sediment-source tracing allows disaggregation of the total sediment flux into specific sediment sources that provides a basis for reconstructing the degree of slope-channel coupling. Interestingly, the early deforestation phase caused not only a rise in low-mid-altitude soil erosion (Figure 5b), but also apparently triggered long-term instability in the high valleys, as shown by the continued upward trend in the signature for upland soil (Figure 5c). The Petit Lac d'Annecy catchment

has excellent land use records, particularly for population, forest clearance, and agricultural land use (Crook *et al.*, 2002), giving the opportunity to define the timescales over which different types of interaction between climate and land use occur. Recent studies have therefore used land use records in combination with a calibrated magnetic proxy for flooding to model the responses to meteorological events and land use (Foster *et al.*, 2003). Climate fluctuations at Annecy have controlled fluxes of total suspended sediment fluxes on short timescales 10^0-10^2 year, whereas phases of land use change have driven the major shifts in the sediment system over timescales of 10^2-10^3 year. It is the intermediate timescale of several decades to a few centuries that has seen the strongest interactions between a fluctuating climate and subtle shifts in land use. Importantly, this timescale dominates global projections and, at least for this subalpine landscape, may be the period in which sediment responses are the most complex to understand and model. The use of climate model outputs as drivers for modeling future sediment responses requires downscaling to spatial scales of the "human landscape" represented by small-medium drainage basins. In this sense, 200-year long reservoir sediment records in Midland England have already been successfully used (Wilby *et al.*, 1997) to model the impact of climatic change on sediment yields under different land uses, using a weather pattern–based approach.

SUMMARY

- Holocene proxy records of sediment yield point to the dominant role of human activities in controlling erosion rates, rather than climate directly. The impact of early agriculturalists before 2000 BP is clear in records from Europe, South America, New Zealand, and China. Later rates of sediment yield are normally higher, especially where European agricultural techniques were introduced within the past 400 years. In many catchments, the current levels of sediment loss are at least 10 times greater than the preimpact state. Sediment source data are less widely available, but where they exist tend to show that sediment is often derived from nonsurface sources. This suggests that catchment disturbance has, in general, caused rates of runoff and stream channel erosion to rise faster than delivery rates of surface soil.
- Analysis of the variability of sediment flux records suggests that the magnitude of changes in sediment transport following catchment disturbance increases as catchment size becomes smaller. This means that surface soil erosion is less marked in lake records of large catchments simply because of greater storage, and not necessarily because of less erosion at the field scale. This also implies that rates of change in sediment accumulation on floodplains increase with basin size. Thus, floodplain and lake sediment records within the same drainage basin may provide contrasting records of fluvial responses. Lake sediments, dominated by the fines in the system, are more likely to preserve the temporal shifts in sediment supply often linked to land use change and more efficient slope-channel coupling. In contrast, alluvial stratigraphies may be more sensitive to fluvial competence, driven by meteorological events, both to entrain and transport coarser particles and to rework existing sediments within the channel and valley floor.
- In terms of global transport of sediment to the coast, long-term perspectives suggest that disturbance of small catchments that feed directly to the coast will deliver relatively more sediment than from large catchments where there is significant sediment storage on floodplains and in large reservoirs. Future climate shifts or widespread land-use changes may result in only minor or even undetectable relative changes in the sediment flux at the outlets of large basins and in near-shore zones. This has implications for the stability of coastlines and the transfer of inorganic and organic sediment to the continental shelf.
- In the future, we may expect that lake sediment records will help form the basis for classifying the sensitivity of modern fluvial and sediment systems to future impacts. For example, systems that have evolved into complex self-organized states under low levels of disturbance may be relatively more susceptible to dramatic shifts in climate and land use than systems already conditioned by long histories of human impact. Recent mathematical and cellular automaton models of long-term erosion (Coulthard *et al.*, 2000; Tucker and Slingerland, 1997) also suggest that sediment delivery over timescales of 10^1-10^2 years is essentially a highly nonlinear product of land-cover change and high-magnitude rainfall events. There are now the exciting prospects of using lake records to test these models, thereby providing a framework of study across a very wide range of spatial (10^0-10^6 m^2) and temporal (10^0-10^3 year) scales.

REFERENCES

Appleby P.G., Flower R.J., Mackay A.W. and Rose N.L. (1998) Paleolimnological assessment of recent environmental change in Lake Baikal: sediment chronology. *Journal of Paleolimnology*, **20**, 119–133.

Barlow D.N. and Thompson R. (2000) Holocene sediment erosion in Britain calculated from lake-basin studies. In *Tracers in Geomorphology*, Chap. 24, Foster I.D.L. (Ed.), Wiley: Chichester, pp. 455–472.

Bichet V., Campy M., Buoncristiani J.-F., Digiovanni C., Meybeck M. and Richard H. (1999) Variations in sediment

yield from the Upper Doubs River carbonate watershed (Jura, France) since the late-glacial period. *Quaternary Research*, **51**, 267–279.

Blais J.M., France R.L., Kimpe L.E. and Cornette R.J. (1998) Climatic changes in northwestern Ontario have had a greater effect on erosion and sediment accumulation than logging and fire: evidence from Pb-210 chronology in lake sediments. *Biogeochemistry*, **43**(3), 235–252.

Bradbury J.P. (2000) Limnologic history of Lago de Patzcuaro, Michoacan, Mexico for the past 48,000 years: impacts of climate and man. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **163**, 69–95.

Cisternas M., Araneda A., MartÌnez P. and Pèrez S. (2001) Effects of historical land use on sediment yield from a lacustrine watershed in Central Chile. *Earth Surface Processes and Landforms*, **26**, 63–76.

Colman S.M., King J.W., Jones G.A., Reynolds R.L. and Bothner M.H. (2000) Holocene and recent sediment accumulation rates in southern Lake Michigan. *Quaternary Science Reviews*, **19**, 1563–1580.

Cooper M.C. and O'Sullivan P.E. (1998) The laminated sediments of Loch Ness, Scotland: preliminary report on the construction of a chronology of sedimentation and its potential use in assessing Holocene climatic variability. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **140**, 23–31.

Coulthard T.J., Kirkby M.J. and Macklin M.G. (2000) Modelling geomorphic response to environmental change in an upland catchment. *Hydrological Processes*, **14**, 2031–2045.

Crook D.S., Siddle D.J., Jones R.T., Dearing J.A., Foster G.C. and Thompson R. (2002) Forestry and flooding in the Annecy Petit Lac catchment, Haute-Savoie, 1730–2000. *Environmental History*, **8**, 403–428.

Curtis J.H., Brenner M., Hodell D.A., Balser R.A., Islebe G.A. and Hooghiemstra H. (1998) A multi-proxy study of Holocene environmental change in the Maya lowlands of Peten, Guatemala. *Journal of Paleolimnology*, **19**, 139–159.

Davis M.B. (1976) Erosion rates and land use history in southern Michigan. *Environmental Conservation*, **3**, 139–148.

De Boer D.H. (1997) Changing contributions of suspended sediment sources in small basins resulting from European settlement on the Canadian prairies. *Earth Surface Processes and Landforms*, **22**, 623–639.

Dearing J.A. (1986) Core correlation and total sediment influx. In *Handbook of Holocene Palaeoecology and Palaeohydrology*, Berglund B.E. (Ed.), John Wiley: Chichester, pp. 247–272.

Dearing J.A. (1992) Recent sediment yields and sources in the Llyn Geirionydd lake-catchment, N. Wales. *Earth Surface Processes and Landforms*, **17**, 1–22.

Dearing J.A. (1994) Reconstructing the history of soil erosion. In *The Changing Global Environment*, Roberts N. (Ed.), Blackwell: pp. 242–261.

Dearing J.A. (1999) Holocene environmental change from magnetic proxies in lake sediments. In *Quaternary Climates, Environments and Magnetism*, Chap. 7, Maher B.A. and Thompson R. (Eds.), Cambridge University Press: Cambridge, pp. 231–278, 335.

Dearing J.A. (2000) Natural magnetic tracers in fluvial geomorphology. In *Tracers in Geomorphology*, Chap. 3, Foster I.D.L. (Ed.), Wiley: Chichester, pp. 57–82.

Dearing J.A., Alström K., Bergman A., Regnell J. and Sandgren P. (1990) Recent and long-term records of soil erosion from southern Sweden. In *Soil Erosion on Agricultural Land*, Boardman J., Foster I.D.L. and Dearing J.A. (Eds.), John Wiley & Sons: London, pp. 173–191.

Dearing J.A., Elner J.K. and Happey-Wood C.M. (1981) Recent sediment flux and erosional processes in a Welsh upland lake catchment based on magnetic susceptibility measurements. *Quaternary Research*, **16**, 356–372.

Dearing J.A. and Flower R.J. (1982) The magnetic susceptibility of sedimenting material trapped in Lough Neagh, Northern Ireland, and its erosional significance. *Limnology and Oceanography*, **27**, 969–975.

Dearing J.A. and Foster I.D.L. (1987) Limnic sediments used to reconstruct sediment yields and sources in the English Midlands since 1765. In *International Geomorphology 1986 Part I*, Gardiner V. (Ed.), John Wiley: pp. 853–868.

Dearing J.A. and Foster I.D.L. (1993) Lake sediments and geomorphological processes: some thoughts. In *Geomorphology and Sedimentology of Lakes and Reservoirs*, McManus J. and Duck R.W. (Eds.), John Wiley: Chichester, pp. 5–14.

Dearing J.A., Hu Y., Doody P., James P.A. and Brauer A. (2001) Preliminary reconstruction of sediment-source linkages for the past 6000 years at the Petit Lac d'Annecy, France based on mineral magnetic data. *Journal of Paleolimnology*, **25**, 245–258.

Dearing J.A. and Jones R.T. (2003) Coupling temporal and spatial dimensions of global sediment flux through lake and marine sediment records. *Global and Planetary Change*, **39**, 147–168.

Dearing J.A. and Zolitschka B. (1999) System dynamics and environmental change: an exploratory study of Holocene lake sediments at Holzmaar, Germany. *Holocene*, **9**, 531–540.

Deevey E.S., Rice D.S., Rice P.M., Vaughan H.H., Brenner M. and Flannery M.S. (1979) Mayan urbanism: impact on a tropical karst environment. *Science*, **206**, 298–306.

Eden D.N. and Page M.J. (1998) Palaeoclimatic implications of a storm erosion record from late Holocene lake sediments, New Zealand. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **139**, 37–58.

Edwards K.J. and Whittington G. (2000) Lake sediments, erosion and landscape change during the Holocene in Britain and Ireland. *Catena*, **42**, 143–173;Einsele G. and Hinderer M. (1997) Terrestrial sediment yield and the lifetimes of reservoirs, lakes, and larger basins. *Geologische Rundschau*, **86**(2), 288–310.

Elvin M., Crook D.S., Jones R.T. and Dearing J.A. (2002) The impact of clearance and irrigation on the environment in the Lake Erhai Catchment from the ninth to the nineteenth century. *Journal of Southeast Asian Studies*, **23**, 1–60.

Flower R.J. (1998) Paleolimnology and recent environmental change in Lake Baikal: an introduction and overview of interrelated concurrent studies. *Journal of Paleolimnology*, **20**, 107–117.

Foster G.C., Dearing J.A., Jones R.T., Crook D.C., Siddle D.S., Appleby P.G., Thompson R., Nicholson J. and Loizeaux J.-L. (2003) Meteorological and land use controls on geomorphic and fluvial processes in the pre-Alpine environment: an integrated lake-catchment study at the Petit Lac d'Annecy. *Hydrological Processes*, **17**, 3287–3305.

Foster I.D.L., Dearing J.A., Simpson A., Carter A.D. and Appleby P.G. (1985) Lake catchment based studies of erosion and denudation in the Merevale catchment, Warwickshire, U.K.*Earth Surface Processes and Landforms*, **10**, 45–68.

Foster I.D.L., Grew R. and Dearing J.A. (1990) Magnitude and frequency of sediment transport in agricultural catchments: a paired lake-catchment study in Midland England. In *Soil Erosion on Agricultural Land*, Boardman J., Foster I.D.L. and Dearing J.A. (Eds.), John Wiley: Chichester, p. 687.

Goman M. and Byrne R. (1998) A 5000-year record of agriculture and tropical forest clearance in the Tuxtlas, Veracruz, Mexico. *The Holocene*, **8**, 83–89.

Jones R., Benson-Evans K. and Chambers F.M. (1985) Human influence upon sedimentation in Llangorse Lake, Wales. *Earth Surface Processes and Landforms*, **10**, 227–235.

Lowe J.J. and Walker M.J.C. (1997) *Reconstructing Quaternary Environments*, *Second Edition*, Longman: London.

Mackereth F.J.H. (1966) Some chemical observations on post-glacial lake sediments. *Philosophical Transactions of the Royal Society of London Series B*, **250**, 165–213.

Milliman J.D. and Syvitski J.P.M. (1992) Geomorphic/tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *Journal of Geology*, **100**, 525–544.

O'Hara S.L., Street-Perrott F.A. and Burt T.P. (1993) Accelerated soil erosion around a Mexican highland lake caused by prehispanic agriculture. *Nature*, **362**, 48–51.

Oldfield F. (1977) Lakes and their drainage basins as units of sediment-based ecological study. *Progress in Physical Geography*, **1**, 460–504.

Oldfield F., Appleby P.G. and Worsley A.T. (1985) Evidence from lake sediments for recent erosion rates in the Highlands of Papua New Guinea. In *Environmental Change and Tropical Geomorphology*, Douglas I. and Spencer E. (Eds.), Allen & Unwin: London, pp. 185–195.

Oldfield F. and Dearing J.A. (2003) The role of human activities in past environmental change. In *Paleoclimate, Global Change and the Future, IGBP Synthesis Book Series*, Chap. 7, Alverson K., Bradley R. and Pedersen T. (Eds.), Springer Verlag: pp. 143–162, 220.

Olley J.M., Murray A.S., Mackenzie D.H. and Edwards K. (1993) Identifying sediment sources in a gullied catchment using natural and anthropogenic radioactivity. *Water Resources Research*, **29**, 1037–1043.

Page M.J. and Trustrum N.A. (1997) A late Holocene lake sediment record of the erosion response to land use change in a steepland catchment, New Zealand. *Zeitschrift fur Geomorphologie*, **41**(3), 369–392.

Roberts N. (1998) *The Holocene. An Environmental History*, *Second Edition*, Blackwell: Oxford, p. 316.

Smol J.P. (2002) *Pollution of Lakes and Rivers*, Arnold: London, p. 280.

Stevenson A.C., Jones V.J. and Battarbee R.W. (1990) The cause of peat erosion: a paleolimnological approach. *The New Phytologist*, **114**, 727–735.

Trimble S.W. (1999) Decreased rates of alluvial sediment storage in the Coon Creek Basin, Wisconsin, 1975–1993. *Science*, **285**, 1244–1246.

Tucker G.E. and Slingerland R. (1997) Drainage basin responses to climate change. *Water Resources Research*, **33**, 2031–2047.

Van der Post K.D., Oldfield F., Haworth E.Y., Crooks P.R.J. and Appleby P.G. (1997) A record of accelerated erosion in the recent sediments of Blelham Tarn in the English Lake district. *Journal of Paleolimnology*, **18**, 103–120.

Van Vliet-Lanoë B., Helluin M., Pellerin J. and Valadas B. (1992) In *Past and Present Soil Erosion*, Bell M. and Boardman J. (Eds.), Oxbow Books Oxbow Monograph (Oxford) 22, pp. 101–114.

Walling D.E. and Fang D. (2003) Recent trends in the suspended sediment loads of the world's rivers. *Global and Planetary Change*, **39**, 111–126.

Wasson R.J., Mazari R.K., Starr B. and Clifton G. (1998) The recent history of erosion and sedimentation on the Southern Tablelands of southeastern Australia: sediment flux dominated by channel incision. *Geomorphology*, **24**, 291–308.

Webb R.S. and Webb T. III (1988) Rates of sediment accumulation in pollen cores from small lakes and moires of eastern North America. *Quaternary Research*, **30**, 284–297.

Whitmore T.J., Brenner M., Engstrom D.R. and Song X. (1994) Accelerated soil erosion in watersheds of Yunnan Province, China. *Journal of Soil and Water Conservation*, **49**, 67–72.

Wilby R.L., Dalgleish H.Y. and Foster I.D.L. (1997) The impact of weather patterns on historic and contemporary catchment sediment yields. *Earth Surface Processes and Landforms*, **22**, 353–363.

Wohlfarth B., Holmquist B.V., Cato I. and Linderson H. (1998) The climatic significance of clastic varves in the Ångermanälven Estuary, northern Sweden, AD 1860 to 1950. *The Holocene*, **8**, 521–534.

Zolitschka B. (1998) A 14,000 year sediment yield record from western Germany based on annually laminated lake sediments. *Geomorphology*, **22**, 1–17.

Encyclopedia of
Hydrological Sciences

Encyclopedia of
Hydrological Sciences

Editor-in-Chief

Malcolm G Anderson
Department of Geography, University of Bristol, Bristol, UK

Senior Advisory Editor

Jeffrey J McDonnell
Department of Forest Engineering, Oregon State University, Corvallis, OR, US

3

WILEY

PART 8

Water Quality and Biogeochemistry

91: Water Quality

MICHEL MEYBECK[1], NORMAN E PETERS[2] AND DEBORAH V CHAPMAN[3]

[1]Sisyphe/CNRS, University of Paris, Paris, France
[2]United States Geological Survey, Georgia Water Science Center, Atlanta, GA, US
[3]Environmental Research Institute and Department of Zoology, Ecology and Plant Science, University College Cork, Cork, Ireland

Water quality is the term used to describe the chemical, physical, and biological characteristics of water. Furthermore, the physical, chemical, and biological characteristics of a water body, "its water quality", determine the suitability of that water for a particular value, for example, potability, ecosystem status, agriculture, industry, recreation. Water-quality issues have become a rapidly evolving component of the environmental sciences primarily due to the increasing demand on water resources and amenity value and the intricate linkage between water-quality use and ecosystem health. Water quality varies markedly in time and space. Episodic temporal water-quality variations can occur in minutes whereas, periodic variations can occur on short timescales from diurnal, associated with variations in light and temperature, to seasonal and longer associated with climatic variations. Spatial variations also vary markedly from millimeters and centimeters in soil profiles to very large-scale variations (>1000 km), for example, associated with water residence times, lithology, landscape position, land use, and bioclimatic zone. In the absence of the effects of human activities, water quality is primarily controlled by climate (precipitation and temperature) and geology (lithology, geomorphology, soil). The water quality of a given volume of water is derived from the sum of effects of mixing and interactions from all upstream sources including atmospheric deposition (precipitation and dry deposition), soil water (matrix and macropore), groundwater, springs, wetlands, irrigation ditches and canals, streams, ponds, lakes, reservoirs, rivers, and estuaries. In most parts of the world, human activities have now caused multiple and complex changes in background water quality.

INTRODUCTION

Water quality is a rapidly evolving environmental science discipline that began more than 150 years ago with a few basic physical (e.g. temperature), chemical (e.g. ammonia, dissolved oxygen), and microbiological (e.g. fecal coliform) variables. Now, water quality integrates dozens of analyses, *in situ* or in the laboratory, on water, particulates, and aquatic organisms and has evolved towards a global understanding of the aquatic environment including the description of habitat and biotic communities. Water-quality studies are conducted by all types of hydroscientists and on all types of continental waters including precipitation, soil water (matrix and macropore), groundwater, springs, wetlands, irrigation ditches and canals, streams, ponds, lakes, reservoirs, rivers, and estuaries. The currently rapid evolution of the water-quality discipline is attributed to the intricate relation among humans and the ecosystem or environmental needs and wants for water (short and long-term), the associated water-quality requirements for the prescribed use and the availability of water with an appropriate quality, and the decreasing availability of water having acceptable water quality.

Water-quality research and development has been driven for many years by the need for (i) scientific understanding of the aquatic environment, (ii) qualification of water for human uses, and (iii) management of land, water, and biological resources. More recently, a fourth objective also has driven water-quality research and development: the fluxes of dissolved and particulate material, through rivers and groundwater, including organic and inorganic components to lakes and reservoirs (Vollenweider, 1968)

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

and from the land to the oceans as part of the Earth System functioning (Livingstone, 1963; Garrels *et al.*, 1973; Berner and Berner, 1987; Vörösmarty and Meybeck, 2004), and is responsible for the degradation of the highly productive and recreationally attractive estuarine and coastal zone (Rabalais and Turner, 2001).

Water quality varies in time and space and its assessment largely depends on the state of scientific knowledge, the financial and technical means available to support the assessment, and the level of "water-literacy" of the community/society. Despite natural linkages between water quality and the lithosphere, water-quality characteristics are increasingly dependent on effects of human activities. Human activities create multiple and competing pressures on land and water resources through agriculture, urbanization, mining, industry, energy production, tourism, water supply, and transportation (water and land). These activities generally degrade water quality while being very demanding for specific water-quality criteria.

This article focuses primarily on surface water chemistry, and river chemistry in particular. Characteristics of water quality are presented in the first section ("Introduction"). General origins, pathways, and controls of water-quality variables are presented in the Section "Water-quality characteristics" and the lithological controls on river chemistry are presented in the Section "Natural river water quality: origins, pathways, and controls of river chemistry". The spatial and temporal distributions of water quality in river basins are the foci for Sections "Spatial variability of major ion concentrations in streams and rivers" and "Temporal variations of river water quality" respectively. Finally, the Section on "Riverine fluxes" focuses on riverine fluxes.

For additional information about water quality, read the articles on monitoring (**Chapter 92, Water Quality Monitoring, Volume 3**), acidification (**Chapter 95, Acidic Deposition: Sources and Effects, Volume 3**), effects of human activities on water quality (**Chapter 93, Effects of Human Activities on Water Quality, Volume 3**), point and nonpoint sources of pollution (**Chapter 94, Point and NonPoint Source Pollution, Volume 3**), nutrient cycling (**Chapter 96, Nutrient Cycling, Volume 3**), urbanization (**Chapter 97, Urban Water Quality, Volume 3**), pathogens (**Chapter 98, Pathogens, Volume 3**), trace elements, salinization (**Chapter 185, Integrated Land and Water Resources Management, Volume 5**), and modeling (**Chapter 100, Water Quality Modeling, Volume 3**). Also, please see Hem (1989), Hem *et al.* (1990), Meybeck (1986), Berner and Berner (1996), and Drever (1988, 2003) for surface water geochemistry; Chapman (1996), Neal *et al.* (2000) and Meybeck (2002) for multiple approaches for local water-quality assessment; Meade (1995), Chapman (1996), Hooper and Kelly (2001), and Kimstach *et al.* (1998) for regional water-quality monitoring and assessments; Meybeck *et al.* (1989), Meybeck (2003) for global assessment; Livingstone (1963), Meybeck and Ragu (1996), and Global Environmental Monitoring System GEMS Water Program (GEMS, 2004) for global databases including hundreds of references on water-quality variables; Chapman (1996), Zhulidov *et al.* (2000), Timmerman *et al.* (2004), and MHSPE (1995) for water-quality surveys and their critics; Phillips *et al.* (1999) for river-flux calculations loads; Helsel and Hirsch (1993) for statistical methods for analyzing water-quality data; Peters (1996) for trend analysis; and Horowitz (1995) for the chemistry of suspended material.

WATER-QUALITY CHARACTERISTICS

Water quality incorporates information primarily from three scientific disciplines (physics, chemistry, and biology) to provide a holistic characterization of a freshwater body.

Physical and Chemical Aspects of Water Quality

The physical component consists of physical measurements that describe the aquatic habitat and its time and space variability. Some physical characteristics that effect biota and chemistry include, but are not limited to, the distribution and characteristics (scattering) of light in a water body, usually below the air-water interface; the color of the water; the water temperature ($T°$) and its distribution temporally and spatially in the water body; specific conductance (K_{SC}), the ability of the water to transmit an electrical current at 25 °C that is primarily determined by the total concentration of charged solutes dissolved in the water; the amount and distribution of particle sizes of suspended material and bed material; the density of the water; the hydrologic characteristics including water discharge and the circulation or hydrodynamics; the slope of the water surface, the spatial interrelation of water flow, and characteristics of the bed material, biota, and geomorphology.

The chemical data comprises analyses of water either filtered or raw (unfiltered), particulates, biota tissues or whole organisms; for organic and inorganic components, elemental isotope content, metals, and radionuclides. For solids, the chemical contents are usually expressed as mass per mass (e.g. $g\,kg^{-1}$, $\mu g\,g^{-1}$), and for "dissolved" (i.e. conventionally after filtration on 0.45 μm, which may still include part of the colloidal fraction) or suspended components, the contents typically are reported as mass per unit volume of water (e.g. $mg\,l^{-1}$ or ppm, $\mu g\,l^{-1}$ or ppb), or for solutes, as moles of solute per unit volume of water ($mol\,l^{-1}$, $\mu mol\,l^{-1}$) or equivalents of charge per unit volume of water ($eq\,l^{-1}$, $meq\,l^{-1}$, $\mu eq\,l^{-1}$). Most natural freshwaters, in the absence of human impacts, have a relatively narrow range of few characteristics that either

reflect the chemistry, such as specific conductance (K_{SC}), or control the chemistry, such as the acid–base status as measured by pH (minus the logarithm of the hydrogen ion concentration) and alkalinity (capacity of water to neutralize acid), and the oxidation and reduction conditions or redox as typically measured by ion pairs and the dissolved oxygen (DO) concentration. In addition, biogeochemical cycles are affected by the affinity of the elements and compounds to adsorb or bind to each other and to other compounds and solids, and by their ability to dissolve in water.

For natural or background conditions, the chemical composition of waters and waterborne particulates is characterized by dozens of elements and compounds including:

Major dissolved ions: cations: Ca^{2+}, Mg^{2+}, Na^{+}, and K^{+}; anions: Cl^{-}, SO_4^{2-}, HCO_3^{-}, and CO_3^{2-}; generally reported in $mg\,l^{-1}$ in ion or elementary form (e.g. mg NO_3^{-} l^{-1} or mg NO_3-N l^{-1})

Minor dissolved elements: Fe, Al, Mn, Ba, and Sr, reported in $\mu g\,l^{-1}$;

Trace dissolved elements: As, B, Cd, Cr, Cu, Hg, Pb, Se, Sn, Sb, Zn, and many others that generally are reported as $\mu g\,l^{-1}$ or $ng\,l^{-1}$;

Natural organic compounds: amino-acids, lipids, humic and fulvic acids, and natural hydrocarbons, generally in trace amounts and dissolved organic carbon (DOC in $mg\,l^{-1}$);

Synthetic organic compounds: several hundreds of man-made products (*xenobiotics*), generally toxic, must be controlled particularly for drinking waters including herbicides, pesticides, solvents, plasticizers, fire retardants, hydrocarbons, and polyaromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs). Some of them are very Persistant Organic Pollutants (POPs) and can be transported in aerosols over long distance, or attached to suspended matter.

Nutrients: used by aquatic plants, including nitrogen (N) species (NO_3^{-}; NO_2^{-}; NH_4^{+}); dissolved and particulate organic N (DON and PON); Kjeldahl N (N_K); total N (TN); phosphorus (P) species orthophosphates (PO_4^{3-}); dissolved organic P (DOP); particulate P (PP); total P (TP), and dissolved silica (H_4SiO_4, generally reported as SiO_2).

Miscellaneous: total dissolved solids (TDS); sum of cations $\sum^{+}$ and anions $\sum^{-}$ (expressed in $meq\,l^{-1}$; they should be equal in balanced analyses); suspended solids (SS), typically expressed as total suspended solids concentration (TSS), and deposited sediments in lakes, reservoirs, floodplain, and sometimes river beds. SS are often considered as a key media for the monitoring of nutrients (N, P), particulate organic carbon (POC; carbon linked to carbonate minerals or PIC both expressed in % of particulate matter), minor and trace particulate elements (Al, Fe, Mn, Si, As, Cd, Cu, Hg, Pb, Sb, Zn) and these are commonly analyzed and reported as content per dry weight ($\mu g\,g^{-1}$, $ng\,g^{-1}$); chlorophyll A and total pigments ($\mu g\,l^{-1}$) as a measure of active and total phytoplankton biomass.

Biological and Microbiological Aspects of Water Quality

All aquatic organisms, whether bacteria, invertebrates, vertebrates, vascular plants, or planktonic algae, have preferred environmental conditions in which they thrive. Many, however, can also be found within certain tolerance limits of these conditions. The basic parameters to which biota are sensitive are physical (e.g. light, $T°$, DO, turbulence, sediment grain size) and chemical (e.g. acidity, nutrients, toxic chemicals). Aquatic organisms need a source of energy – for photosynthetic organisms this means adequate light and nutrients and for others it means a source of suitable food, ranging from organic detritus to plant material to other invertebrates or vertebrates. Thus, directly, or indirectly through the food chain, aquatic organisms are dependent on the physical and chemical aspects of water quality.

In turn, the presence of aquatic organisms influences water quality – photosynthesis by planktonic algae and macrophytes can increase oxygen concentrations to levels of supersaturation at times of high light intensity, and the respiratory action of bacteria decomposing dead and dying plankton or macrophytes can use up the oxygen in the water thus resulting in a daily O_2 cycle whose intensity is a measure of the aquatic production. A living plankton is a component of the total suspended matter, and decaying plants and plankton contribute organic matter particles to both suspended and deposited sediments with a ratio algal organic carbon/total pigments between 30 and $40\,g\,g^{-1}$ for plankton. During photosynthesis (or production P), phytoplankton uptakes dissolve inorganic nitrogen, NO_3^{-} and NH_4^{+}, and phosphorus (PO_4^{3-}) according to the Redfield molar ratio C_{106}:N_{16}:P_1. As organisms decay, they also release nutrients back into solution and the seasonal cycles of phytoplankton species and/or macrophytes can greatly influence the seasonal fluctuations in lake nutrient concentrations. Under less productive conditions, the oligotrophic state, diatom algae are generally dominant particularly in spring blooms of the temperate and cold regions. In addition to N and P nutrients, these algae also use dissolved silica. As a result of seasonal algal uptakes marked cycles of H_4SiO_4, NO_3^{-}, NH_4^{+}, and PO_4^{3-} are observed in lakes and reservoirs and in some highly productive (eutrophic) rivers. The bacterial decomposition (or respiration, R) of the organic matter produced in the aquatic system leads to oxygen depletion. When the P/R ratio <1 there is a net O_2 depletion, as in lake bottom waters in the absence of vertical mixing, or in turbid macrotidal estuaries. This depletion may also be caused by the decomposition of organic wastes, for example, from urban sewage, thus resulting in an oxygen sag-curve downstream of untreated and partially treated effluents. Thus the traditional perception of water quality

cannot be separated from the aquatic organisms present in the water body and this has led, in recent decades, to a more holistic approach to the study of water quality to include some, or all, biological components of aquatic systems.

The interaction between suitable habitats, that is, environmental conditions, and suitable food supplies leads to communities of organisms that typically thrive together in natural water bodies. These communities include bacteria using dissolved organic carbon, primary producers (plants and plankton), herbivorous invertebrates, detritivores, fish, and even mammals (the tropic pyramid). Any impact on one of these groups can cause effects on the dependent groups, for example, through loss of food supply, alteration of environmental conditions (physical or chemical), and altered predator–prey relationships. Ultimately, the natural balance of the whole aquatic system may be altered, and with it the physical, chemical, and biological water quality.

Aquatic organisms are more and more used to describe the water quality in its wider definition (Chapman, 1996). Certain individuals or combinations of organisms can be associated with narrow ranges of water quality, particularly the extremes of clean and polluted water. Thus, the organisms themselves provide a monitoring tool, whereby their presence or absence can be indicative of certain classes of water quality (*see* **Chapter 92, Water Quality Monitoring, Volume 3**). Some of those aquatic organisms are chosen for their limited mobility (e.g. benthos); they integrate at one given station the effects of human influences on their environment (chemical and physical) over the period of their lifetime that can vary from months to decades. Other organisms have a great mobility, such as fish, or may even travel along an entire aquatic system from headwaters to ocean, and are chosen for their overall integration of both physical and chemical degradation (e.g. salmon). Some aquatic organisms also accumulate toxic compounds and can serve as indicators of contamination through biomagnification of the signal. Organisms also demonstrate physiological or morphological effects caused by the presence of anthropogenic emissions (*see* **Chapter 92, Water Quality Monitoring, Volume 3** and MHSPE (1995, vol 32)).

Historical interest in water quality has been largely based on potential human health impacts and, even today, the presence of pathogenic organisms in water, including some bacteria, enteric viruses, and parasites, is a major concern. In many developing parts of the world, it is the only aspect of water quality to which resources are directed for monitoring and management. *Escherichia coli* (E. coli, a primary indicator of fecal contamination) and fecal coliforms are usually found in low concentrations (10^0 to 10^3 counts per 100 ml) in the absence of human impacts, and therefore, are the most common indicators of fecal contamination. Fecal pathogens can cause a wide variety of intestinal diseases, many of which result in diarrhea; the World Health Organization (WHO) considers them as the most important water-quality issue globally, particularly in the developing world. Once added to water by disposal of human excreta, particularly domestic sewage, coliforms decay very rapidly at ambient temperature (by one to two orders of magnitude within 24 h). In highly contaminated water, total coliform may exceed 10^5 or 10^6 coli $100\,ml^{-1}$. WHO recommends that water containing any *E. coli* should not be used for drinking (WHO, 2004). Also, a monthly geometric mean for *E. coli* of <130 coli $100\,ml^{-1}$ is recommended for contact recreation, such as swimming and boating.

More recent human health concerns relate to the presence of naturally occurring planktonic organisms known as *cyanobacteria* or blue–green algae. These organisms thrive in highly eutrophic water bodies with limited water circulation, such as ditches, shallow lakes, and ponds and the surface layers of stratified lakes in the summer months. Some species produce toxins and release them into the water as the cells decay during the end-of-season collapse of the population. The toxins are hard to detect, but the presence of the species themselves can be used to signal caution that the water may not be of "good quality" for human as for domestic animals (Chorus and Bartram, 1999).

Water-quality issues can be assessed by specific variables (Table 1). Depending on the availability of human and financial resources, their monitoring can focus on first order variables or include also the second order variables. Simple to complex combinations of water-quality variables referred to as a *water-quality index*, are used to indicate the status of a freshwater body with respect to some particular problem or issue. For example, pH and alkalinity, a measure of the capacity of a water body to neutralize acid, are used to indicate the susceptibility of a water body to acidification, and various nutrient species and chlorophyll A concentrations are used to determine the tropic status of a water body. Sodium and calcium are combined in a Sodium Adsorption Ratio, which is used in irrigation to assess the risk of salinization Hem (1989).

NATURAL RIVER WATER QUALITY: ORIGINS, PATHWAYS, AND CONTROLS OF RIVER CHEMISTRY

The natural origins and pathways of riverine materials are multiple and vary from headwaters to estuaries and are controlled by many physical, chemical, and biochemical processes (Figure 1 left part). The fate of river-borne material eventually depends on the chemical composition of the dissolved or particulate phases, the source of the component, the hydrologic pathways, the transformation of the component or other interactions or reactions with materials

Table 1 Major water quality issues. Physical and chemical monitoring variables, their relevance to aquatic biota (0 to +++ scale), the origin of pressures, and limitations of uses (**bold**: very important)

Issues	1st priority variables	2nd priority variables	Aquatic Biota	Origins of stresses	Use limitations
Thermal pollution	Temperature[a]		++	**D, F, G_1, G_2**, Z	G, I
Salinization	Electrical conductivity[a]	Major ions	++	**B_3**, D, **K, X**, Z	A, **B_3**, C, D, I
Acidification	pH[a], Alkalinity	Major ions; DOC	++	**D, E, K, G_2**	C, **D, I**
Oxygen balance	Dissolved oxygen[a]	BOD, COD; DOC, POC	+++	**B_2, J, D**	**C**, D_2, H, **I**
Chemical contamination					**A**, B, **C**, D_2, H, **I**
NH_4^+	NH_4^+		+	**B_2, J**	
NO_3^-	NO_3^-			**B_1, J**	
Metals	Cd, Cu, Hg, Pb, Zn; Al	As, Cr, Ni, Se, Sb...	++	**D**, E, J, **K**	
POPs	total PCBs; total PAH	Individual compounds	+	**D, J**	
Pesticides	total pesticides	Individual compounds	+	**B**, J	
Endocrine disrupters		Individual compounds	?	**B_2, D, J**	
Microbial contamination	Total coli; *E. coli*	Streptococci...		**B_2, J**	**A, B, C**, D_2, **H**, I
Eutrophication	Total P and N; chloro A	C, N, P, Si species, algal counts	++	**B, B_2**, D, I, **J**, W	C, E, G, **H, I**
Water-related diseases	Specific surveys (parasites, insects larvae...)			**B_2, J**, X, Y, W, Z	**A, C**, D_2, **H**
Habitat destruction	Specific surveys		+++	B,**E, F, G_3, J**, X	
Radionuclide contamination	Total α, β, γ activity	Individual radionuclides		D, **G**, J	**A, B, C**, D, **H, I**

[a]possible *in situ* and continuous measurement
Human activities: A = traditional agriculture and rural settlement, B = Intensive agriculture (B_1 = fertilized crops, B_2 = cattle feedlots, B_3 = irrigated agriculture), C = Drinking water, D = Industries (D_1 = nuclear industries), D_2 = food industries, E = Transportation (ground, air, fluvial), F = Damming, G = Energy production (G_1 = nuclear power plant, G_2 = coal and fuel power plant, G_3 = hydropower), H = Recreation, I = Fisheries, J = Urbanization, K = Mining.
Global change: W = Biodiversity alteration, X = Hydrological balance, Y = Sea level rise, Z = Global warming.

during transport, the residence time and interaction of the various hydrological pathways resulting in mixing, and biogeochemical processes. Atmospheric deposition (Figure 1, #4) is an important source of Na^+, Cl^-, and SO_4^{2-}, originating from ocean aerosols (Figure 1, #1), with the highest deposition occurring closest to the source, and somewhat important sources of Ca^{2+}, Mg^{2+}, and SO_4^{2-}. Dust originates from Aeolian erosion (Figure 1, #3) and volcanism and terrestrial vegetation (Figure 1, #2). Decaying terrestrial and aquatic vegetation (Figure 1, #2), which previously fixed atmospheric CO_2 and N_2, are the main source of C and N species in surface waters. Soil leaching provides DOC and DON. Soil mechanical erosion delivers POC, PON, PP, and SS. Mineral weathering provides major ions and silica (Figure 1, #10) through different water pathways including groundwater inputs (Figure 1, #11); the highest streamwater concentrations are associated with the highest groundwater proportion in surface waters during the dry season. Saline deep aquifers and hydrothermal water can yield high concentrations of SO_4^{2-}, Cl^-, and K^+ in some volcanic and/or tectonically active regions, as in the East African Rift or in New Zealand (Figure 1, #12). Erosion or resuspension of particulates resulting from stream and riverbed incision also may be important.

The fate of SS in aquatic systems is characterized by a succession of deposition and remobilization within a river channel dependent on variations in the velocity of streamflow and through different filters that characterize the river system. The filters include hillslopes and wetlands located in headwaters (Figure 1, #5), lakes (Figure 1, #6), alluvial plains and related wetlands (Figure 1, #9), and estuaries. Each filter retains particulate matter and associated chemicals (PIC, POC, P, PN, PP, and POPs), except during major floods when some of these materials may be remobilized. The in-stream/channel retention (erosion and deposition) of SS varies markedly between high and low flows and is highly variable during individual hydrological events. However, riverbed storage generally is not permanent on medium timescales (2–10 years), unless cementation occurs in the channel sediment (e.g. the Huang He river channel in China).

Several chemical and biochemical processes occurring in terrestrial vegetation, soils, and waterscape filters, such as wetlands, lakes, and reservoirs control

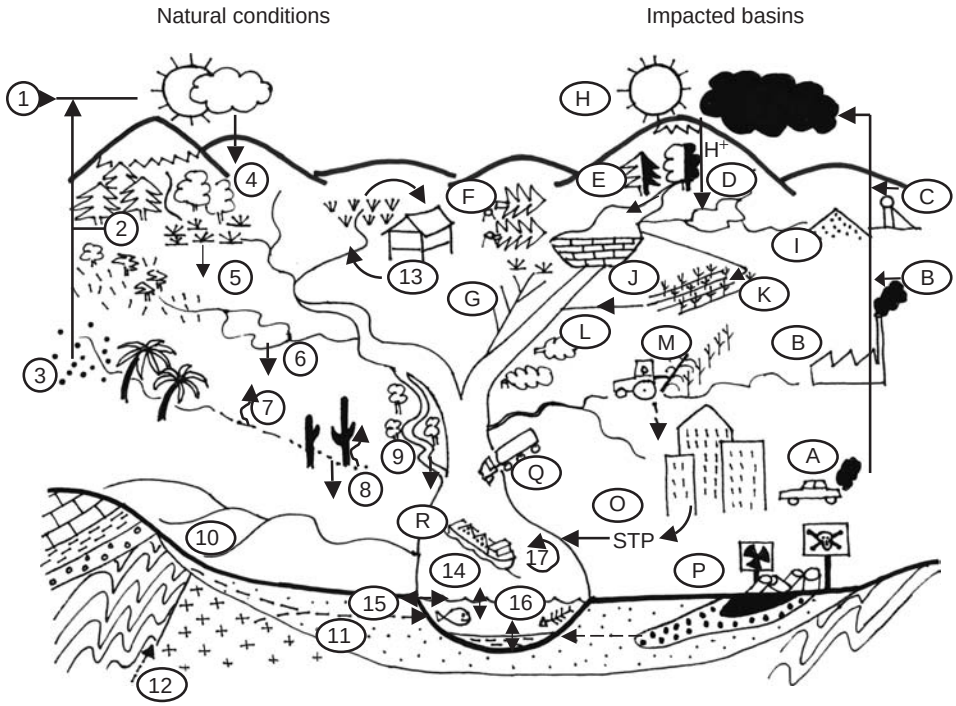


Figure 1 Sources, pathways, and main processes regulating water quality in natural conditions and impacted basins. Atmospheric fallout (4) originating from oceanic inputs (1), vegetation emissions (2), and Aeolian erosion (3); retention and transformation in wetlands (5) and lakes (6); evaporation leading to salinization (7) and precipitation in soils and endorheic basins (8); retention and exchange with floodplain (9); chemical weathering and mechanical erosion of various rock types (10); direct inputs of groundwater (11); hypothermal inputs (12); closed N, P cycles in traditional agriculture (13); exchange between surface waters and atmosphere (14), groundwater (15), and sediments (16); internal cycling of carbon and nutrients in aquatic food webs (17). Contaminated and/or acidified atmospheric deposition (D) due to urban and traffic (A); industrial sources and mining/smelting sources (C); forest die back (E) and forest cutting (F); wetland draining (G); climate change (H); mining (I) and industrial (B) wastewaters; river fragmentation through damming (J); enhanced evaporation after irrigation (K); use of fertilizers and pesticides in agriculture (K, M); river course channelization and floodplain isolation (L); release of treated and untreated urban sewage (O); leak of dangerous chemicals from waste dumps (P); accidental spills (Q) and chronic leaks (R); enhanced eutrophication (17)

major ions, nutrients, organic carbon, and particulates (Table 2). For a more detailed discussion of nutrient cycling in freshwater, *see* **Chapter 96, Nutrient Cycling, Volume 3**.

Evaporation concentrate solutes in surface waters (Figure 1, #7) and ultimately in dry and semiarid regions (e.g. Central Asia, Altiplan, Great Basin, and Lake Eyre Basin) causes the water to become saturated with respect to various minerals that precipitate typically in the order of calcite, magnesium calcite, gypsum, and anhydrite. The precipitated minerals occur in soils, floodplains, and depressions where water ponds after floods (Figure 1, #8).

In impacted basins, new sources and pathways of materials occur leading to accelerated fluxes and/or increased retention (e.g. in impoundments). Physical changes due to climate change and land use change are also important for water quality. These impacts are schematically described in Figure 1 (right part) and in Table 2, and fully developed in **Chapter 95, Acidic Deposition: Sources and Effects, Volume 3**, **Chapter 93, Effects of Human Activities on Water Quality, Volume 3** and **Chapter 94, Point and NonPoint Source Pollution, Volume 3**.

SPATIAL VARIABILITY OF MAJOR ION CONCENTRATIONS IN STREAMS AND RIVERS

Global Scale Variability of Water Chemistry

At the global scale under natural conditions, major ion concentrations of streams are regulated by chemical weathering for more than 90% of river basin area and for more than 90% of river water flow followed by atmospheric deposition and evaporation (Gibbs, 1970). In some highly dilute waters and for some constituents, the storage and cycling of elements by vegetation may be another important control on streamwater chemistry (Meybeck, 2003).

Table 2 Principal sources, sinks, and controls of main chemical descriptors of rivers under natural conditions. For an explanation of the numbers in column 1, see Figure 2

Figure 2 Schematic	Processes	SS	POC	H_4SiO_4	Ca^{2+} Mg^{2+}	Na^+	K^+	Cl^-	SO_4^{2-}	HCO_3^- (DIC)	TN	PO_4^{3-}	DOC	O_2
					Sources									
4	Atmospheric inputs	+			+	++	+	++	+		+			
2	Terrestrial vegetation uptake		+++							++	+++		+++	
5, 10	Soil (erosion, weathering, leaching)	+++	+++	+++	+++	++	++	+	++	+++	+++	+++	+++	
11	Groundwaters			++	+++	++	+		+	++		+		
12	Hydrothermal			+		+	+	+	+					
	River bed incision	++												
					Controls									
2, 5, 6, 9, 16, 17	Biogeochemical cycles	+	++	++	+		+		+	++	++	++	++	
8	Chemical precipitation			++	+++				+	+++				
16, 17	P/R balance in aquatic systems		+++										+++	+++
					Sinks									
5, 6, 8, 9	Particulate retention (permanent)	+++	+++								+			
16	River bed retention (short term)	+			+					+				
2	Terrestrial vegetation storage		+				+		+		+			
8	Soil storage	+	+	+	+	+	+	++	+	+	+			

+++: essential process; ++ important in some occasions; + may occur. SS: suspended solids; POC and DOC: particulate and dissolved organic carbon.

Chemical weathering is controlled by soil mineralogy, rainfall amount, and partitioning into groundwater recharge and surface runoff, temperature, and terrestrial vegetation (Drever, 2003). The relative abundance of minerals in the soil and surficial rocks and their sensitivity to weathering are major factors controlling the composition of a basin's surface water.

Surface water can be further classified into nearly a dozen of principal chemical types on the basis of major ion concentrations (Meybeck, 2003). These include a few classes that are dominated by atmospheric deposition and evaporation/precipitation (Table 3).

Atmospheric deposition may affect river water quality. In the first 100 km of coastlines where rainfall is an important contributor to surface water, the ocean inputs of Na^+, Cl^-, Mg^{2+}, K^+, and SO_4^{2-} may be important and may even dominate relative to the contribution of chemical weathering. These constituents are present not only in rainfall, but in oceanic aerosols that accumulate on vegetation and land surfaces and are washed off by the rainfall. For the coastal ribbon the primary source of Na^+, Cl^-, Mg^{2+} is almost always atmospheric deposition, particularly if the ionic content in local rock types is negligible or low and, consequently, the concentration ratios among these constituents are similar to that of seawater (Table 3(b), The Cusson River).

Except in evaporitic sedimentary rocks, chlorine is generally very rare at the Earth's surface. Chloride ion (Cl^-), the dominant form of chlorine in surface water, is generally attributed to ocean-derived atmospheric deposition in the absence of human activity. In turn, the related ions can be determined on the basis of ionic ratios of the solute to Cl^-, as occurs in seawater (e.g. $Na^+/Cl^- = 0.88\,eq\,eq^{-1}$). However, Cl^- is concentrated and contributed by many human activities and natural levels of Cl^- are often exceeded by anthropogenic inputs, as for Na^+, K^+, and SO_4^{2-}.

Table 3 Natural stream water chemistry. (a) common variations associated with bedrock type of moderately undisturbed basins in France, (b) uncommon controls from around the World (full references in Meybeck 1986, 2003)

Basin bedrock type	H_4SiO_4 ($mg\,l^{-1}$)	Σ^+ ($\mu eq\,L^{-1}$)	Ca^{2+} ($mg\,l^{-1}$)	Mg^{2+} ($mg\,l^{-1}$)	Na^+ ($mg\,l^{-1}$)	K^+ ($mg\,l^{-1}$)	Cl^- ($mg\,l^{-1}$)	SO_4^{2-} ($mg\,l^{-1}$)	HCO_3^- ($mg\,l^{-1}$)	Chemical type ($eq\,L^{-1}$)
(a) Weathering dominance: examples on monolithologic basins										
Alkaline granites, gneiss, micaschists	8.4	130	0.4	0.2	1.9	0.3	0.0	1.9	5.5	Na^+–HCO_3^-
Quartz sands & sandstones	10.2	170	1.0	0.4	1.8	0.3	0.0	3.3	6.2	Na^+–HCO_3^-
Trachy-andesites	11.4	220	3.2	1.5	4.0	1.6	0.0	0.5	29.9	Na^+–Ca^{2+}–HCO_3^-
Calco-alkaline granites, gneiss, micaschists	6.0	300	3.2	0.9	1.2	0.5	0.0	2.2	15.5	Ca^{2+}–HCO_3^-
Arkose sands	12.0	400	3.8	1.7	1.1	0.8	0.0	3.8	19.5	Ca^{2+}–HCO_3^-
Shales	5.4	500	6.0	1.8	0.9	0.3	0.0	6.0	22.9	Ca^{2+}–HCO_3^-
Basalts	12.0	500	4.2	2.3	2.0	0.5	0.0	0.5	29.9	Ca^{2+}–Mg^{2+}–HCO_3^-
Serpentinite	13.6	1500	11.4	10.9	2.4	0.6	0.0	5.0	85.1	Mg^{2+}–HCO_3^-
Flyshs	3.0	2200	34.8	5.1	0.8	0.4	0.0	2.1	131.0	Ca^{2+}–HCO_3^-
Marls	5.4	3000	49.9	5.1	1.7	0.6	0.0	2.9	179.0	Ca^{2+}–HCO_3^-
Limestones	3.6	4500	85.7	1.6	0.6	0.7	0.0	5.4	267.0	Ca^{2+}–HCO_3^-
Chalks	12.0	4500	85.7	1.4	2.3	0.9	0.0	5.4	267.0	Ca^{2+}–HCO_3^-
Coal shists	9.0	5000	30.1	33.4	16.1	2.0	0.0	132.0	137.0	Mg^{2+}–SO_42-
Dolomite	4.0	5900	63.8	33.0	0.1	2.3	0.0	24.1	329.0	Ca^{2+}–Mg^{2+}–HCO_3^-
Gypsum marl	9.6	22000	339.0	58.8	1.0	1.7	0.0	878.0	228.0	Ca^{2+}–SO_4^{2-}
(b) Other chemical controls: atmospheric inputs, vegetation storage, evaporation, hydrothermalism										
Rio Negro tributary (1)	4.5	18.1	0.0	0.0	0.2	0.1	–	–	–	Na^+–DOC
Rivière de l'est (2)	22.4	1255	6.1	5.8	9.1	3.1	3.2	1.6	68.7	Mg^{2+}–HCO_3^-
Cusson (3)	15.1	1463	4.0	3.8	20.7	2.0	35.8	8.7	16.7	Na^+–Cl^-
Lufira (4)	18.6	3830	38.5	17.6	10.2	0.8	10.3	18.1	198.0	Ca^{2+}–HCO_3^-
Semliki (5)	12.8	8736	11.2	38.4	79.5	60.0	45.2	97.9	330.0	Na^+–HCO_3^-
Tedzhen (10)		13240	79.2	41.1	131.0	6.7	144	223.0	243.0	Ca^{2+}–SO_4^{2-}
Saoura (6)	–	26150	122.0	53.4	356.0	7.2	582	348.0	153.0	Na^+–Cl^-
Redwater (7)	7.0	40700	87.0	121.2	599.0	9.6	14.4	1410.0	663.0	Na^+–SO_4^{2-}
Tokaanu (8)	286.0	41600	120.0	15.2	760.0	50.6	1336	116.0	88.8	Na^+–Cl^-
Salt (9)	1.2	312,000	606.0	68.2	6340.0	12.2	9924	1390.0	190.0	Na^+–Cl^-

(a) analyses corrected for rain inputs assuming Cl^- to be of oceanic fallout origin. (b) uncorrected analyses from various climatic zones and lithology: (1) pure quartz sand, Brasil; (2) basalt (La Réunion); (3) arkose sand with important oceanic fallout (Landes, France); (4) with limestone outcrops Congo basin; (5) outlet of Lake Edward, hydrothermal inputs (Uganda); (6) evaporated endorheic river (Morocco); (7) pyritic shales (Montana); (8) hydrothermal stream (New. Zealand); (9) draining salt deposits (N.W.T., Canada); and (10) Central Asia.

In basins where chemical weathering is very low (see Rio Negro tributaries, Table 3(b)), terrestrial vegetation is a major store and regulates the release of elements derived from atmospheric deposition, including Ca^{2+}, K^+, SO_4^{2-}, NO_3^-, NH_4^+, and Cl^-.

Examples of average chemical concentrations of streams draining monolithologic basins are given in Table 3(a) for 15 rock types from a set of 200 small temperate and relatively undisturbed basins in France. A second set of analyses is from heterogeneous lithologic basins (Table 3(b)). Differences in the solute concentrations among these rivers are primarily due to the bedrock composition except for H_4SiO_4, which also increases with average annual temperature for a given rock type. Note that except for H_4SiO_4, the solute concentrations for the wet tropics such (e.g. Rivière de l'Est, basalt, La Réunion or Lufira, limestone, Congo basin) are very similar to those in the temperate regions for the same rock type. For the Cusson River in the coastal French Landes (quartz and feldspar sand deposits), more than 80% of major ions are derived from the atmospheric deposition of ocean aerosols, and consequently, the Na^+ to Cl^- ratio ($0.87\,eq\,eq^{-1}$) is approximately that of seawater (0.88). If the atmospheric deposition of marine-derived constituents (i.e. all Na^+, all Cl^-, most Mg^{2+}, and most SO_4^{2-}) were subtracted from the Cusson River analyses, the cation sum would decrease from $1463\,\mu eq\,l^{-1}$ to $250–300\,\mu eq\,l^{-1}$.

Waters draining basins containing evaporite minerals, particularly halite (NaCl), are difficult to differentiate from waters subjected to high evaporation in arid and semiarid regions. Both can be extremely mineralized. For example, where rainfall greatly exceeds potential evaporation, the Salt River in northern Canada ($\Sigma^+ = 312\,meq\,l^{-1}$) (Table 3(b), #9) and where evaporation exceeds rainfall by

about 10 times (runoff $q = 14\,\mathrm{mm\,y^{-1}}$), the Saoura River (South Morocco, $\Sigma^+ = 26.1\,\mathrm{meq\,l^{-1}}$, Table 3(b), #6) and the Tedzhen, a river in Central Asia ($\Sigma^+ = 13.2\,\mathrm{meq\,l^{-1}}$, Table 3(b), #10).

The hydrothermal and/or deep-water discharge to surface water can effect surface water solute concentrations in some volcanic regions (e.g. Iceland, New Zealand, Japan, Kamchatka), active mountain building regions (e.g. Himalayas, Caucasus), and in rift zones (e.g. Semliki River of East Africa, Table 3(b), #5). H_4SiO_4 concentrations also can be very high (e.g. the Tokaanu in New Zealand, Table 3(b), #8).

Spatial Distribution of Water Chemistry within River Basins

The statistical analysis of mean chemistry at stations that are implicitly assumed to be independent from each other is a first step in many spatial analyses. Mapping of water chemistry is common for groundwater, but is more difficult for surface water within river basins because: (i) river water quality integrates the water quality of contributing waters upstream, that is, from the entire drainage basin, (ii) differences in the drainage area of each station reflects, to some extent, the residence time of water in the basin that, in turn, a effects weathering-derived solute concentrations, and (iii) the upstream–downstream interdependency of nested stations.

For river basins with many water-quality monitoring stations, concentration distributions can be represented and compared using a probability scale. *Very common* concentrations are represented by concentrations between C_{25} and C_{75} percentiles, *common* concentrations between $C_{10}-C_{25}$ and $C_{75}-C_{90}$, *uncommon* concentrations C_1-C_{10} and $C_{90}-C_{99}$, and *rare* concentrations between $C_{0.1}-C_1$ and $C_{99}-C_{99.9}$ (Meybeck, 1996). These qualifications apply to any basin scale and are useful for comparing results among regions. Examples of within-basin spatial distributions are shown in Figure 2 for SO_4^{2-} and pH for some regions in relatively undisturbed basins, from the very dilute central Amazon waters to the mineralized Mackenzie River basin.

Longitudinal and Transversal Profiles of River Chemistry

Most mapping is realized on major stream orders through longitudinal profiles by simple interpolation between stations, particularly on the river main stem, such as for the Mississippi River (Meade, 1995). In river basins that are heterogeneous with respect to climate or lithology, water

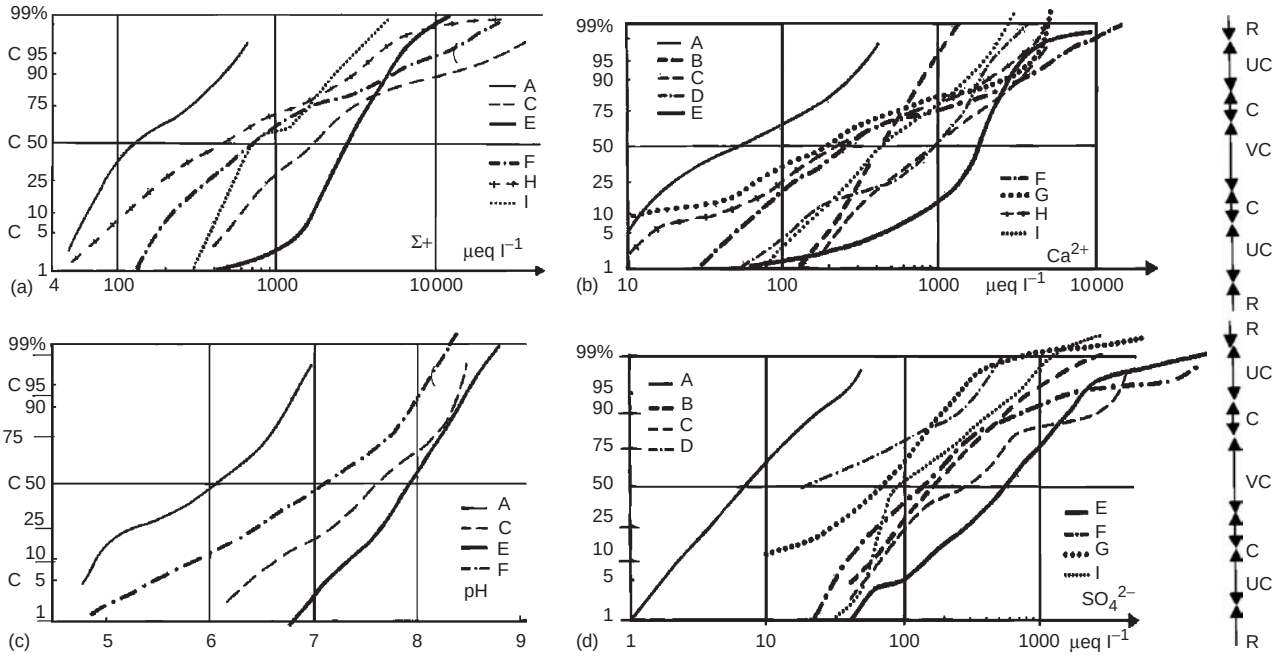


Figure 2 Cumulative distribution of ion chemistry in rivers from selected relatively undisturbed regions. (a) cation sum Σ^+ (μeq L^{-1}); (b) Ca^{2+} (μeq L^{-1}); (c) pH; (d) SO_4^{2-} (μeqL^{-1}). A: central and lower Amazon ($n = 40$); B: Japan ($n = 225$); C: Andean tributaries of Amazon ($n = 42$); D: Thailand ($n = 31$); E Mackenzie river basin ($n = 100$); F monolithologic French streams ($n = 250$); G ''temperate'' stream model; H: Monolithologic miscellaneous streams ($n = 75$); I: major world rivers ($n = 60$). For distributions of K^+, Na^+, HCO_3^-, and SiO_2: see Meybeck (1994). For more detailed information about the individual data series, see Meybeck, 2003 and Meybeck and Ragu, 1996 and references therein. VC = ''Very common'' concentrations: between 25% and 75% of distributions; C = ''common'': 10 to 25% and 75 to 90%; UC = ''uncommon'': 1 to 10% and 90% to 99%; R = ''rare'': 0.1 to 1% and 99% to 99.9%

quality can vary longitudinally, and, occasionally, across the river section as a result of poor mixing of tributary flow into the river.

For example, the TDS of the Arkansas River, a major tributary of the Mississippi River varies longitudinally (Stoner, 1984). The river headwaters in Colorado are somewhat mineralized with an average TDS of 273 mg l^{-1}. Downstream in Kansas, TDS increases to 3670 mg l^{-1}, then decreases to 880–1100 mg l^{-1} in Oklahoma and increases again to 3000 mg l^{-1} in Arkansas. During low flow at some stations, TDS and Cl^- may exceed 35 000 and 250 mg l^{-1} (the drinking water limit) respectively. The cause of the salinity variations is mostly natural due to drainage of underlying Permian salt deposits, common in this basin but rarely found at the global scale, but evaporation in reservoirs also may be important.

Under certain conditions, river water quality varies laterally across the river section, such as downstream of confluences where the water quality of each stream differs markedly. Two examples have been particularly well described: (i) in Brazil, the confluence of the Rio Negro "Black water" (organic rich, extremely dilute; Table 3(b), #1) with the Rio Solimoes "White waters" (sediment rich, dilute) that forms the Amazon River downstream of Manaus; (ii) in Canada, the triple confluence of the Upper St Lawrence River (i.e. the Lake Ontario outlet) that is depleted in DOC and is medium dilute, the Ottawa River that is organic rich and dilute, and streams draining the southern Laurentian Mountains, that have high SS and are medium dilute (Centre Saint Laurent, 1996). In each case, complete mixing does not occur until hundreds of kilometers downstream of the confluence despite very high-stream velocity (>1 m s^{-1}).

In basins impacted by human activities, the spatial distribution of water quality is generally more variable and is often affected by stepwise degradation of the aquatic habitat, downstream of point discharges of pollutants, downstream of nonpoint or diffuse sources of pollutants atmospheric deposition, and mobilization of the agrochemicals applied to fields, or as a result of water engineering (e.g. dams, diversions, irrigation returns).

TEMPORAL VARIATIONS OF RIVER WATER QUALITY

Temporal variations in river systems are specific for each water-quality variable and provide important information on the origins, pathways, and controls of that variable, particularly when evaluated with respect to changes in climate, hydrology, and land use or human activities in the basin. The variations range from sub-daily to secular. Although some patterns are associated with major changes either due to climate or human activities, frequently periodic patterns occur within each year associated with seasonality for which the cycles are called "*regimes*", and daily, called *diurnal* (24 h) and *nychtemeral* cycles; *nychtemeral* cycles are more specifically linked to the alternating light conditions from day and night that controls aquatic primary production. In most rivers, important water-quality variations are observed during floods due to the mixing of surface runoff, subsurface runoff, and groundwater discharge – for streams – and/or, for the large and heterogeneous basins, due to the mixing of major tributaries. In impacted basins during high water stages, point sources of contaminants are diluted, whereas soil erosion (e.g. TP) and leaching (e.g. NO_3^-) may result in higher concentrations. Initial steps in analyzing streamwater concentration variations include assessing the concentration relation with discharge and evaluating concentrations patterns during individual floods.

Long-term (>10 years), fixed-interval (weekly to bimonthly) monitoring provides information for assessing seasonal variations of river water quality. Several types of seasonal chemical regimes have been described from these data including (i) the influence of seasonal river flow regime; (ii) seasonal biogeochemical cycles; and (iii) seasonal anthropogenic affects from various types of human activities, such as reservoir operation, agriculture, forestry, mining, industry, and tourism (Table 1). At least 10 years of water-quality indicators are needed to detect trends from interannual variations by appropriate statistical tests (Peters, 1996 and Hooper and Kelly, 2001). The trends of riverine concentrations are generally linked to the magnitude of human activities and related implementation of waste management and pollution abatements.

Concentrations Variations with River Discharge

Constituent concentrations (C_i) typically vary with water discharge (Q). The use of specific discharge $\bar{q}$ (l s^{-1} km^{-2}) also allows for the easy comparisons of stations with very different basin size; normalizing q to long-term average q and concentrations C to median values C_{50} aids the comparison of different chemical variables from trace elements to major constituents (Figure 3). Most major ions and dissolved silica concentrations are higher in groundwater than in soil water or surface water due to a longer residence time of the groundwater that provides more time for concentrations to increase as a result of weathering. For streams where the source of these constituents is primarily groundwater, concentrations moderately decrease (partial dilution) with increasing discharge (Figure 3, #A). Near-perfect dilution (i.e. a quasi-constant flux) (Figure 3, #B) occurs downstream of point sources, such as saline springs and wastewater outfalls (e.g. NH_4^+, PO_4^{3-}). Clockwise hysteresis (concentrations increase more rapidly with increasing discharge on the rising limb of a storm hydrograph than decrease during recession) often occurs for DOC concentrations (Figure 3, #C1) because of rapid leaching or flushing and subsequent depletion of the DOC from the

upper soil layers during stormflow. SS concentrations in lowland basins typically show the same pattern because of the scouring of previously deposited channel sediment (Figure 3, #C_2). In contrast, in mountainous regions on a double logarithmic scale, the SS concentration is often linearly related to discharge (Figure 3, #D). Most particulate nutrients, pollutants, and metals, follow patterns of C_2 or D. POC, expressed in percent of SS, typically decreases with increasing SS in most rivers, but the concentrations as $mg\,l^{-1}$ increase moderately (Figure 3, #E) with regards to the SS pattern.

A few other patterns occur including: (i) apparent plateau at high flows (Figure 3, #F) resulting from seasonal nutrient uptake at lake outlets or stream denitrification of NO_3^- at summer low flows, a common pattern in some western European rivers; (ii) a dilution/resuspension pattern (Figure 3, #G) occurs for TP concentrations downstream of urban wastewater outfalls; the dilution of urban sewage PO_4^{3-} is followed by a resuspension of P-rich sediment, an example of the inverted behavior for two chemical species that are often analyzed together in unfiltered samples; (iii) a dilution/hysteresis pattern (Figure 3, #H), either clockwise or counterclockwise occurs in heterogeneous basins because of the different timing of tributary discharges to a river; they are often associated with variable concentrations because of basin heterogeneity, some combination of processes, or processes yet to be identified; (iv) near constant concentrations (Figure 3, #I) can be observed in groundwater dominated streams and are also observed for Cl^- when derived primarily from atmospheric deposition; and (v) temporally, sometimes seasonally, variable rare patterns with maximum or minimum concentrations lagging climatic or hydrologic causes of variation.

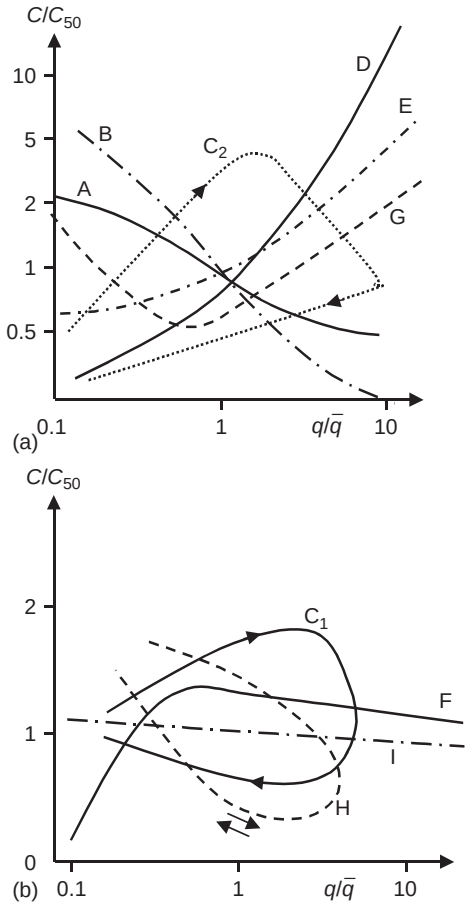


Figure 3 Concentration patterns (C. normalized to median concentrations, C_{50}) with respect to specific discharges ($\bar{q}$, normalized to the annual average q). A, B: dilution; C_2, D, E: soil erosion and leaching; F, G: complex variations; C_1, C_2, H: hysteresis cycles; and I: constant

Common and Rare Events

Water quality can change rapidly under natural conditions. Some variations are systematic and may occur each year. For example, during snow melt in the northeastern United States and in Scandinavia, acid anions (NO_3^-, SO_4^{2-}) and H^+ ion increase in streamwater resulting in episodic stream acidification, which has adverse affects on biota, such as macroinvertebrates and fish (*see* **Chapter 95, Acidic Deposition: Sources and Effects, Volume 3**). Other natural accidental events, including extreme droughts, hurricanes, landslides, forest fires, and volcanic eruptions can result in major changes in water quality, with concentration variations at least one order of magnitude higher than under normal conditions. These events typically are rare on the human timescale (>100 year return period), and are generally not captured during routine surveys. The duration of rare events may be short, but their impact on water quality may be enormous and last for (forest fires)hundreds of years. After the mount St Helen eruption (Oregon, USA), the local river TSS concentrations and yields increased by two orders of magnitude and the sediment discharged by the Columbia River increased from 10 to 40 $Mt\,y^{-1}$ in the years after the eruption, although most of the sediment (140 Mt) remained in the Cowlitz River, a Columbia River tributary (Meade and Parker, 1985).

Water-quality Trends and Interannual Variations

Year-to-year or interannual variations of water quality are caused by climatic and related hydrological variations (e.g. water discharge for rivers) and by changes of human activities in river basins. During wet years, material derived from point sources are diluted more than during dry years, and major ions, NH_4^+ or PO_4^{3-} concentrations are generally lower (Figure 3, dilution patterns #A, B, H). In contrast, average concentrations of particulate-bound components (Figure 3, patterns #C_1, C_2, D, E, and G) are much higher than during dry years. Particulate-bound components include POPs and total (unfiltered water) metal

concentrations, POC, and TP. Interannual variations of eutrophication indicators such as total pigment, chlorophyll A, and daily cycles of pH and DO are caused by insulation and water temperature variations.

RIVERINE FLUXES

The mass output (or flux) from a given upstream basin is used to evaluate material balances (outputs minus inputs), and more recently, to assess the status of degradation or improvement of stream reaches to assist land and water resources managers. For example, water-quality standards in the United States have targeted the total maximum daily load (TMDL) of contaminants for stream reaches (US EPA, 2004).

A mass flux (F, mass per time) combines concentration (C, mass per volume) of a solute, solid, or compound with discharge or streamflow (Q, volume per time). The mass flux is commonly determined for major ions, suspended matter (SPM), toxic substances, nutrients, carbon species, and organic matter, and several commonly used synthetic organic compounds, such as PAHs, PCBs, herbicides, and pesticides. The flux typically is determined at stream-gauging stations at which streamflow is continuously monitored because the computation requires discharge, and sometimes the use of specific sampling instrumentation and at higher frequencies than available from synoptic or routine manual sampling. Theoretically, F should be derived from the continuous measurements between t_1 and t_2 of both Q and C of the targeted dissolved or particulate constituent (i):

$$F_i(t) = \int_{t_1}^{t_2} C_i(t)\, Q(t)\, \mathrm{d}t \qquad (1)$$

Because C_i is rarely continuously measured or estimated (e.g. through a relation with a surrogate variable that is continuously measured like electrical conductivity), it is necessary to rely on a set of discrete water-quality (chemistry) analyses obtained from n fixed samples taken at time t_j coupled with continuous Q data, sometimes only with discrete Q_j measured during the water-quality sampling. Different approaches to flux calculation can be used and many of them incorporate concentration–discharge relationships (Phillips *et al.*, 1999).

CONCLUSION

During the last 150 years, measurement of water quality has evolved from few physical and chemical determinations to the analysis of the concentrations of hundreds of chemical compounds conducted in many types of media from dissolved to various particulate phases and to the consideration of aquatic biota and habitat. Also, while water quality is still considered a "subjective standard for human usage" for drinking water and multiple other uses, such as agriculture and industry, it is increasingly complemented by *aquatic environmental quality* that aims to be an "objective attribute for classification" (Boon and Howell, 1997).

Under natural conditions, water quality can vary markedly in time and space with most chemical descriptors varying by more than 1–2 orders of magnitude. In addition, human activities can markedly affect spatial and temporal variations of water quality, that is, chemical concentrations, the nature of the habitat, and modification of the aquatic communities. Metrics of water quality and reference scales also are evolving rapidly reflecting human and water relations at a given location and time for a given culture, even if some water-quality criteria (e.g. drinking water) are now universally used.

The continuous developments of our scientific and technical knowledge on water quality and of new issues and the growing linkage between water quality and social sciences (e.g. predictive models of carbon and nutrients for the twenty-first century combining Global Change and local pressures scenarios) have transformed water-quality studies into an autonomous multidisciplinary hydroscience.

REFERENCES

Berner E.K. and Berner R.A. (1987) *The Global Water Cycle, Geochemistry and Environment*, Prentice Hall: Englewood Cliffs, p. 397.

Berner E.K. and Berner R.A. (1996) *Global Environment, Water, Air and Geochemical Cycles*, Prentice Hall: Englewoods Cliff.

Boon P.J. and Howell D.L. (Eds.) (1997). *Freshwater Quality: Defining the Indefinable*? Scottish Natural Heritage/HMSO Publications, Edinburgh.

Centre Saint Laurent (1996) *Rapport-Synthèse sur l'état du Saint Laurent. Vol.I. L'écosystème du Saint Laurent*, Environment Canada and Editions Multimondes: Montréa.

Chapman D.V. (Ed.) (1996) *Water Quality Assessments: A Guide to the Use of Biota Sediments and Water in Environmental Monitoring*, *Second Edition*, Chapman & Hall: London.

Chorus I. and Bartram J. (Eds.) (1999) *Toxic Cyanobacteria in Water. A Guide to Their Public Health Consequences, Monitoring and Management*, E & FN Spon: London.

Drever J.I. (1988) *The Geochemistry of Natural Waters*, *Second Edition*, Prentice-Hall: New York.

Drever J.I. (Ed.) (2003) Surface and groundwater, weathering and soils. In *Treatise on Geochemistry*, Vol. 5, Holland H.D. and Turekian K.K. (Eds.), Elsevier: Amsterdam.

Garrels R.M., Mackenzie F.T. and Hunt C. (1973) *Chemical Cycles and the Global Environment*, William Kaufman: Los Altos, p. 306.

GEMS (2004) United Nations Environment Programme GEMS/WATER. `http://www.gemswater.org/index.html` [18 October 2004].

Gibbs R. (1970) Mechanism controlling world water chemistry. *Science*, **170**, 1088–1090.

Helsel D.R. and Hirsch R.M. (1993) *Statistical Methods in Water Resources*, *Studies in Environmental Science 49*, Elsevier: New York.

Hem J.D. (1989) *Study and Interpretation of the Chemical Characteristics of Natural Water*, U.S. Geological Survey Water Supply Paper, 2254, USGS: Reston.

Hem J.D., Demayo A. and Smith R.A. (1990) Hydrogeochemistry of rivers and lakes. In *The Geology of North America*, *Volume 0–1. Surface Water Hydrology*, Wolman M.G. and Riggs H.C. (Eds.), Geological Society of America, pp. 189–231.

Hooper R.P. and Kelly V.J. (Eds.) (2001) Water quality of large US rivers: results from the US Geological Survey's national stream quality accounting network. *Hydrological Processes*, Special issue, **15**(7), 1085–1393.

Horowitz A.J. (1995) *The Use of Suspended Sediment and Associate Trace Element in Water Quality Studies*, International Associates of Hydrological Sciences: Special Publication, 4.

Kimstach V., Meybeck M. and Baroudy E. (Eds.) (1998) *A Water Quality Assessment of the Former Soviet Union*, E&FN Spon: London.

Livingstone D.A. (1963) Chemical composition of rivers and lakes chapter G. *Data of Geochemistry*, U.S. Geological Survey Professional Paper 440 G, USGS: Reston, pp. G1–G64.

Meade R.H. (Ed.) (1995) *Contaminants in the Mississippi River 1987-1992*, U.S. Geological Survey Circular 1133, USGS: Reston.

Meade R.H. and Parker R.S. (1985) Sediment in rivers of the United States. In *National Water Summary, 1984*, Water Supply Paper 2275, USGS: Reston, pp. 40–60.

Meybeck M. (1986) Composition chimique naturelle des ruisseaux non pollués en France. *Sciences Geologiques Bulletin*, **39**, 3–77.

Meybeck M. (1994) Origin and variable composition of present day riverborne material. In *Material Fluxes on the Surface of the Earth. Studies in Geophysics*, Ed. Board on Earth Sciences and Resources-National Research Council, National Academic Press: Washington, pp. 61–73.

Meybeck M. (1996) River water quality, global ranges tine and space variabilities. *Verhandlung Internationale Vereiningung Limnologie*, **26**, 81–96.

Meybeck M. (2002) Riverine quality at the anthropocene: propositions for global space and time analysis, illustrated by the Seine River. *Aquatic Sciences*, **64**, 376–393.

Meybeck M. (2003) Global occurrence of major elements in rivers. In *Treatise on Geochemistry*, *Volume 5, Surface and Ground Water, Weathering and Soils*, Holland H.D., Turekian K.K. and Drever J.I. (Eds.), Elsevier: Pergamon, pp. 207–224.

Meybeck M., Chapman D. and Helmer R. (Eds.) (1989) *Global Fresh Water Quality : A First Assessment*, Basil Blackwell: Oxford.

Meybeck M. and Ragu A. (1996) *River Discharges to the Oceans. An Assessment of Suspended Solids, Major Ions, and Nutrients*, Environment Information and Assessment Report UNEP: Nairobi.

MHSPE (1995) Monitoring water quality in the future. *Chemical Monitoring, vol 2 Toxicity Parameters, vol 3 Biomonitoring, vol 4 Monitoring Strategies for Complex Mixtures, vol 5 Organizational Aspects*, Vol. 1, Minister of Housing Spatial Planning Development: Zoetermeer.

Neal C., House W.A., Leeks G.J.L., Whitton B.A., Williams R.J. (Eds.) (2000) Water quality of UK rivers entering in the North Sea (LOIS). *Science of the Total Environment*, **251–252**, 1–703.

Peters N.E. (Eds.) (1996) Trends in water quality. *Hydrological Processes*, Special issue, **10**, 127–356.

Phillips J.M., Webb B.W., Walling D.E. and Leeks G.J.L. (1999) Estimating the suspended sediment load of rivers in the LOIS study area using in frequent samples. *Hydrological Processes*, **13**, 1035–1050.

Rabalais N.N. and Turner R.E. (2001) Hypoxia in the Northern Gulf of Mexico : description, causes and change. In *Coastal Hypoxia. Consequences for Living Resources and Ecosystems*, Rabalais N.N. and Turner R.E. (Eds.), Caostal and Estuarine Studies 58, American Geophysical Union, pp. 1–36.

Stoner J.D. (1984) Dissolved solids in the Arkansas river basin. *National Water Summary, 1984*, US Geological Survey: Water Supply, Paper 2275, pp. 741–778.

Timmerman J.G., Behrens H.W.A., Bernardini F., Daler D., Ross P., van Ruiten K.J.M. and Ward R.C. (Eds.) (2004) *Monitoring Taylor Made (IV)*, RIZA: Lelystad.

US Environmental Protection Agency (EPA) (2004) **`http://www.epa.gov/owow/tmdl/overviewfs.html`** [18 October 2004].

Vollenweider R.A. (1968) *Scientific Fundamentals of the Eutrophication of Lakes and Flowing Waters*, Technical Report DA5/SCI/68.27, OCDE, Paris, p. 250.

Vörösmarty C.J. and Meybeck M. (2004) Responses of continental aquatic systems at the global scale : new paradigms, new methods. In *Vegetation, Water, Humans and the Climate : a New Perspective on an Interactive System*, *IGBP Synthesis Series*, Chap. D4, Kabat P., Claussen M., Dirmeyer P.A., Gash J.H.C., de Guenni L.B., Meybeck M., Vörösmarty C.J., Hutjes R.W.A. and Lütkemeier S. (Eds.), Springer Verlag: Berlin, pp. 375–413.

WHO (2004) *Guidelines for Drinking-Water Quality*, vol. 1, *Third edition*, WHO: Geneva, 366 pp.

Zhulidov A.V., Khlobystov V.V., Robarts R.D. and Pavlov D.F. (2000) Critical analysis of water-quality monitoring in the Russian Federation and Former Soviet Union. *Canadian Journal of Fisheries Aquatic Sciences*, **57**, 1932–1939.

92: Water Quality Monitoring

DEBORAH V CHAPMAN[1], MICHEL MEYBECK[2] AND NORMAN E PETERS[3]

[1] *Environmental Research Institute and Department of Zoology, Ecology and Plant Science, University College Cork, Cork, Ireland*
[2] *Sisyphe/CNRS, University of Paris, Paris, France*
[3] *U.S. Geological Survey, Georgia Water Science Center, Atlanta, GA, US*

Water quality monitoring is the process of gathering data that describes the physical, chemical, and biological condition of a water body. This chapter presents an overview of the processes involved in water quality monitoring and illustrates some of the approaches used with examples of existing monitoring programs. Over the last century, improved understanding of water quality, combined with advances in measurement and monitoring technology have increased the possibility of measuring hundreds of different variables in surface and groundwaters. However, efficient use of resources for monitoring depends on careful selection of objectives for the water quality monitoring program and on targeting variables and monitoring methods that address those objectives. Recent appreciation of the close link between the physical and chemical condition of a water body and its biological component has led to the incorporation of biological approaches in many large-scale programs to assess surface water quality. Today, monitoring programs take many forms, from measurement of a few specific variables to establish trends or effectiveness of remediation measures, to sophisticated evaluation of toxic impacts of wastewaters, to relatively simple determination of the state of the aquatic environment using citizen participation. Confidence in all data gathered and its resultant usefulness for management and policy development is essential; this can only be achieved by a careful program of quality assurance that extends from sample collection in the field, to analysis in the laboratory, to data handling and manipulation.

INTRODUCTION

Effective management of rivers, lakes, and groundwaters must be based on an understanding of their physical, chemical, and biological characteristics, that is, the water quality. This understanding can only be achieved through the collection and interpretation of appropriate information. The raw data that provides this information is obtained through the process of monitoring or data gathering. Monitoring is an activity that can take many forms, as illustrated in this chapter, but essentially it consists of the systematic collection of data over temporal or spatial scales. Human impacts on water quality and quantity (*see* **Chapter 93, Effects of Human Activities on Water Quality, Volume 3** and also **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5** and **Chapter 188, Land Use and Water Quality, Volume 5** for examples of specific impacts) have resulted in more complex management problems and an associated demand for relevant information that will help support management decisions. In addition, policy makers increasingly have to consider regional and even global scales in policy design. This expansion in the need for information is posing challenges for those responsible for monitoring, both in terms of meeting diverse needs within a limited resource base and in designing cost-effective methods and approaches that provide data that will be useful and meaningful.

The term "monitoring" is used loosely to include all activities that involve the gathering of data by making measurements, whether in the field situation or during a process such as water treatment. In practice, many monitoring programs are set up to fulfill a particular goal aimed either at determining whether water is fit for a specific use or whether water quality has deteriorated

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

because of human activities. Examples of more precise monitoring goals are:

- Operational monitoring (also known as *surveillance*) – checking compliance with set criteria such as maximum allowable concentrations, water quality standards or discharge consents
- Surveys – limited duration data gathering exercises to determine the status of the water body at that given time (for example, in response to an accidental pollution incident)
- Background monitoring – collection of baseline data to indicate reference, background or unperturbed conditions for comparison with other sites, or for comparison with the same sites when studying trends or recovery/remediation effects (often used to determine impacts of future or remote activities)
- Trend monitoring – long-term data gathering specifically aimed at showing environmental change over time
- Flux monitoring – gathering of data at defined boundaries to determine the flux of specific materials from one environment, or water body, to another

The diversity of reasons for monitoring and of methods used is too great to discuss in detail in this chapter – thus it presents a guide to the principal activities involved, together with an overview of the approaches taken to monitoring water quality in different geographical regions and with different objectives. Some selected examples of water quality monitoring programs from around the world are also included. Examples have been chosen to illustrate some practical approaches that are in more widespread use and from well-established programs that have been in operation for at least a number of years.

To ensure that monitoring activities result in useful and credible information, it is necessary to choose appropriate methods at all stages of the monitoring process – from field to laboratory, and even to data interpretation, storage, and handling – and to apply rigorous quality checks to all these stages. Reliable and accurate water quality data are the foundation of the information that is needed for aquatic resource management in conjunction with the appropriate hydrological information. The raw data, in themselves, may be of little value unless some process of interpretation and presentation is applied to make them accessible and understandable to the users, namely, water resource managers, policy makers, and the public. This interpretation can take the form of a simple graph showing fluctuations over time to a complete assessment of the environmental situation. Thus, water quality monitoring should not be seen as an activity on its own, but as part of the overall water quality assessment process. This process begins with the setting of objectives, continues through the data gathering exercise and concludes with interpretation and assessment.

HISTORICAL DEVELOPMENT OF WATER QUALITY MONITORING

The concept and understanding of water quality has been developing gradually over more than one hundred years and is likely to continue to evolve because of (i) increasing demand for water qualified by the type of use, (ii) increasing human demands on continental water resources, (iii) improving knowledge and perceptions about the aquatic environment, and (iv) developing technology, from field measurements and sampling, and laboratory analyses to data management, analysis, and communications.

In the 1850s, the first water quality analyses (e.g. Seine and Thames rivers) were based on a few physical and chemical variables, such as T°, pH, DO, K_{SC}, Cl, NH_4^+, and NO_3^-. The longest record of water quality for these variables exceeds 100 years for the Rhine River. Just before World War I, the first systematic geochemical studies of rivers were conducted on United States' and Canadian rivers, and routinely included the concentrations of major ions: Ca^{2+}, Mg^{2+}, Na^+, K^+, (usually measured as Na^+ plus K^+), Cl^-, SO_4^{2-}, NO_3^-, H_4SiO_4, and HCO_3^-, and sometimes other variables including NH_4^+, PO_4^{3-}, iron (as Fe_2O_3), aluminum (as Al_2O_3), and salinity. Although some sewage contamination studies were conducted as early as 1900, the first routine surveys of microbiological contamination indicators, including total fecal coliform counts, were conducted during the 1930s at public water supply intakes of major cities. Around the same time, the first routine measurements of riverine suspended particulate matter (SPM) were conducted to determine baseline fluxes prior to damming (e.g. USA, former USSR, China).

During the 1950s and 1960s, atmospheric nuclear bomb testing created a need for artificial radionuclide and plutonium surveys. These surveys generally were independent of other types of water quality monitoring and typically were not accessible to scientists or to the general public. Meanwhile, eutrophication issues and the development of automated nutrient analyses accelerated the measurements of N species and other nutrients including PO_4^{3-}, total P, and H_4SiO_4. Since the middle of the 1970s, organic carbon analyzers have enabled the analyses of total, dissolved, and particulate organic carbon concentrations (TOC, DOC, and POC, respectively), which have been gradually replacing biochemical oxygen demand (BOD) and chemical oxygen demand (COD) measurements. Investigations of chemical components of water quality also have benefited greatly, since the 1970s, from technological advances in instrumentation and analytical techniques such as atomic absorption, ion chromatography, inductively coupled plasma spectrometry, gas chromatography, and mass spectrometry. Analytical improvements have been mostly driven by the growing demand for analyses of a growing list of contaminants such as polychlorinated biphenyls (PCBs), industrial products similar to the DDT insecticide, polyaromatic hydrocarbons

(PAHs), phthalate plasticizers and other persistent organic pollutants (POPs). POPs are highly toxic and accumulate in soils, sediments, and biological tissues because they have a high affinity for particulate matter, are highly soluble in fat tissues, and have a low solubility in water. Most of these compounds are not natural and typically occur at very low concentrations in the environment. The first routine surveys of sediment quality were also conducted in the 1970s. The combination of chemical (e.g. C, N, P, metals, POPs) and biological analyses (e.g. pigments, diatom assemblages) of sediment cores permitted the reconstruction of past water quality issues (e.g. salinization, eutrophication, oxygenation, acidification) over decades and more, opening another rapidly growing-scientific field in water quality, namely, the study of sediment archives.

Since the 1960s, water quality monitoring has also included biota, especially benthic species, photosynthetic pigment concentrations (as an indicator of primary productivity or algal biomass), algal species counts, and chemical analyses of contaminants in biota (e.g. freshwater mussels) or in specific tissues of the biota (e.g. fish muscle) (see following text).

The total number of variables that should be considered in a comprehensive water quality monitoring program, if all regulations and water quality approaches were strictly followed, would now exceed 500. Because of financial constraints, and because human activities which affect water quality vary spatially and temporally, most monitoring programs rarely include more than 100 variables for individual samples, even at the most well-equipped monitoring laboratories. In less developed countries, where water quality monitoring exists, the number of variables rarely exceeds 20, including major ions, some nutrients, total suspended solids (TSS), pH, T°, and conductivity. The choice of variables, and the laboratory-based analytical methods associated with those variables, is often governed by the resources (financial, technical, and human) available for the monitoring program. Where resources are not limited, the analytical techniques should be selected for which standardized methods and appropriate quality assurance programs (see following text) exist, bearing in mind any specified or anticipated detection limits. Electronic information sources, such as the National Environmental Methods Index (NEMI) (NWQMC, 2004a), can assist with such selection.

Efforts to obtain useful information with limited expenditure and/or resources have driven the development of water quality indicators (physical, chemical, and biological – see following text). An indicator is a measured variable, or combination of variables, that can be related to particular environmental conditions and is representative of those specific conditions, for example, the Secchi depth (see following text) as an indicator of eutrophication. It can also be a specific aquatic organism that is associated with a defined range of water quality variables (see following text).

MONITORING PROGRAM DESIGN

The setting of objectives should be the first activity in the steps required to design and implement a water quality monitoring program (Figure 1). Water quality monitoring can be very demanding on the resources of any organization, sometimes requiring considerable personnel, and/or advanced technical facilities. Expenditure on such resources must be justifiable by the ultimate provision of useful information for management or policy needs. In order to ensure that resources are not wasted, it is important to define clearly the objectives of any monitoring program in terms of its expected outputs. The objectives should take into consideration the technical capability of the monitoring organization and should be achievable within the resources available to the operators of the program. The more focused and well-defined the objectives, the more likely it is that resources will not be wasted and that expectations will be met (e.g. NWQMC, 2004b).

Water quality objectives, the related monitoring variables, and the type of water quality surveys and studies can vary markedly for the same water body among hydrological disciplines depending on the purpose of the scientific investigation (e.g. process studies made by geochemists, biogeochemists, and some hydrologists) or requirement for compliance with water quality criteria as

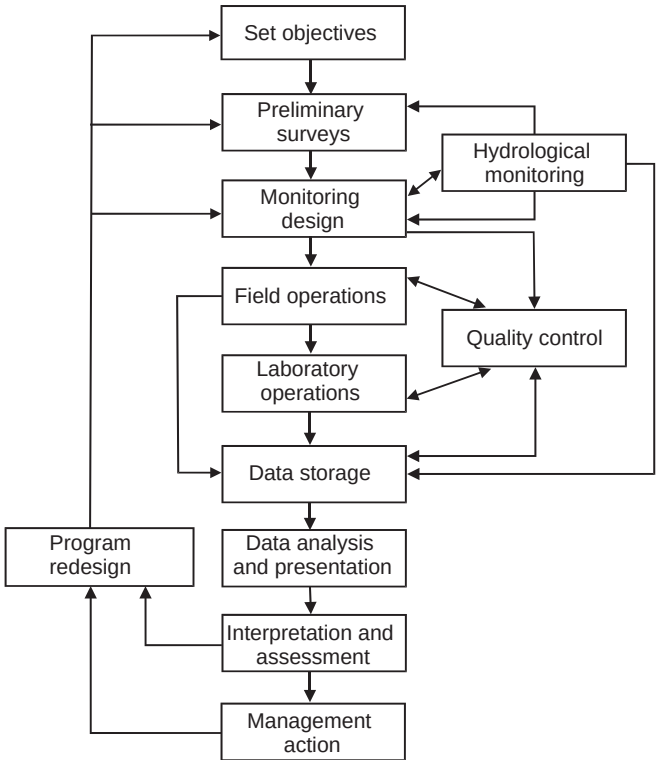


Figure 1 Stages of the water quality monitoring process

Table 1 Typical monitoring objectives and associated activities undertaken by different hydrological science disciplines

Disciplines	Objectives	Variables	Type of study, survey, or monitoring program
Environmental chemists	Detect and quantify trace amounts of any contaminants including degradation products	Trace elements, synthetic organic compounds, new compound on the market (e.g. pesticides) and/or synthesized by chemical and pharmaceutical industries	In-depth inventories; surveys of ambient backgrounds; sediment archives
Geochemists	Chemical composition with regards to natural drivers (lithology, climate...); inputs to oceans	Ca^{++}, Mg^{++}, Na^{+}, K^{+}; Cl^{-}, SO_4^{2-}, HCO_3^{-}; dissolved/colloidal/particulate trace elements; chemical speciations	Process studies (stream scale); fluxes (basin scale)
Biogeochemists	Cycles and transformations of nutrients (dissolved and particulate organic/inorganic)	Carbon, nitrogen, phosphorus, and silica species; micronutrients	Links with algal production and bacterial activities
Ecologists	Physicochemical quality; occurrence and levels of contaminants; pigment levels	pH; T° TSS; color; dissolved oxygen; total dissolved solids; nutrients; ammonia contaminants; endocrine disruptors; chlorophyll A	Overall quality; general mapping; seasonal regime extreme events; bioaccumulations; biomagnifications
Limnologists	Ecological functioning; conservation; potential for lake uses; paleo-records of water quality	Dissolved O_2; nutrients; pH; conductivity; pigments; contaminants	Process studies (fine time, space, and scales); vertical profiles; seasonal variations; sediment mapping; sediment archives
Hydrologists	River fluxes of dissolved and particulate matter; tracers of water pathways	Major ions; SPM; natural isotopic tracers (e.g. $^{13}HCO_3^{-}$, $^{14}HCO_3^{-}$, $^{2}H_2O$, $^{3}H_2O$, $^{34}SO_4^{2-}$ ...); artificial radionuclides	Flood cycles; seasonal regimes; yearly fluxes; trends; stream order; structure of water quality
Sanitary engineers	Water-borne and water-related diseases; indicators of pollution	biochemical oxygen demand and chemical oxygen demand; microbial contamination indicators (fecal and total coli, streptococci); nitrate; endocrine disruptors; radionuclides	Water quality at intakes; continuous surveillance; detection of threshold quality (accidents)
Water manager	Comparison with water use criteria; pollution sources inventories	Most of above variables with growing focus on microcontaminants such as pesticides and endocrine disruptors	Contaminant mapping and budgets; longitudinal profiles; long-term surveillance; emergency survey
Community/citizen volunteer monitor	Trends in water quality in relation to water use/human impact; protection of biodiversity	Simple field measurements of physical/chemical variables; identification and enumeration of organisms	Trends; spatial surveys; species inventories

might be evaluated by sanitary engineers or water managers (Table 1).

Well-defined objectives aid the efficient design of all stages of the monitoring program (Table 2), from the selection of measurement variables to site selection, to choice of field and laboratory methods, to data interpretation and presentation techniques (Spooner and Mallard, 2003). The simplest monitoring programs may have only one objective; an example would be to check that water abstracted from a well is fit for human consumption according to the World Health Organization (WHO) guidelines for drinking water quality (WHO, 2004). In this case there is a specific monitoring site (the point of abstraction from the well) and the variables to be measured, together with their recommended frequency, are defined by the need to check whether guidelines concentrations are met for specific variables. The output of the program is an assessment of whether, on each sampling occasion, the water in the well meets the concentrations defined in the guidelines and is considered fit to drink.

Table 2 Principal activities associated with each stage of a monitoring program

Stage of monitoring program	Principal activities
Setting objectives	Consider water uses, legislation, guidelines, standards, and so on
	Consider economic and technological constraints
	Define expected outcomes of monitoring program
	Establish data quality objectives
Preliminary survey	Survey literature and databases for existing physical, chemical, biological, or hydrological data, information on methods, and so on.
	Test field and laboratory methods if necessary
	Carry out special survey to evaluate potential sites and/or methods for long-term use
	Establish measurement quality objectives
	Evaluate technical and financial resources required
Program design	Decide spatial and temporal sampling regime
	Select monitoring media (water, sediments, biota), variables (physical, chemical, biological), field sites, sampling frequency, specific methods, and equipment
	Produce final program design, including guidelines for technical personnel, standard operating procedures, field record sheets, laboratory record sheets
Implementation	Field operations: Collect biota, sediments, and water samples for laboratory analysis, take *in situ* physicochemical measurements and carry out on-site biological and chemical analyses, record field data
	Hydrological measurements: collect information on flow, velocity, water level, and so on.
	Laboratory operations: Preparation of sample bottles and addition of pretreatment chemicals, analysis of samples, recording of results
Quality control	Checking accuracy of field and laboratory methods, for example, with sample blanks and duplicate samples
	Checking in-house analytical techniques with sample blanks and spiked samples
	Participation in interlaboratory quality assurance exercises
	Regular checking for suspect data in databases
Data manipulation	Data storage: Transferring results from field and laboratory operations into database
	Analysis of data: Application of statistical methods, for example, correlations, trend analysis
	Assimilation and presentation of data: Tables of results, data summaries, graphs

Source: Adapted from Chapman (1996).

The increasing need over the last few decades to manage and protect all aspects of water resources has led to a greater understanding of the functioning of water bodies and of their interactions with other components of the hydrological cycle. The demand for management of freshwater resources in an integrated way, such as that required under the European Union Water Framework Directive (EU, 2000) can lead to the need for monitoring programs with multiple objectives. It is particularly important that monitoring programs with multiple objectives are reviewed periodically to evaluate whether all the objectives are still valid and necessary and whether the program is achieving the objectives. There should be regular feedback from the users of the outputs of the monitoring program to the designers and operators in order to ensure that the program remains appropriate and cost-effective (see Figure 1).

Just as monitoring programs should be designed to meet specific objectives, they should also be designed specifically to suit the type of water body. This means that at least a fundamental understanding of the hydrology of the specific water body is necessary. Where this information does not already exist, it may have to be obtained through a preliminary survey. Typical hydrological information required before designing a monitoring program is summarized for the three main types of water body in Table 3. Knowledge of hydrological features is essential for selecting the appropriate sites for sampling. For example, it is important to know the direction and velocity of groundwater flow when selecting sites for boreholes with the aim of studying pollutant dispersion. In river waters, variations in contaminant concentrations are closely related to discharge and depend on whether the contaminant is in the dissolved state or is associated with particulate matter (*see* **Chapter 91, Water Quality, Volume 3**). Discharge in rivers is highly variable depending on rainfall, the size, and other physical and geological characteristics of the catchment and the resultant runoff, varying over minutes to hours to days to months to seasons, and for some climate impacts, to years or decades. Flux calculations (*see* **Chapter 86, Measuring Sediment Loads, Yields, and Source Tracing, Volume 2** and **Chapter 91, Water Quality, Volume 3**) depend on the river discharge at the time of sampling. Thus, wherever possible, it is recommended that measurements are made at frequent intervals or even continuously. In lakes and reservoirs, short residence times might influence the frequency of monitoring, whereas turbulence and mixing

Table 3 Hydrological information and key measurements required for increasing complexity of water quality monitoring programs

Level of complexity of monitoring program	Rivers	Lakes/reservoirs	Groundwaters
All programs	Map of catchment plus:	Depth at samples site(s); water residence time; thermal regime plus	Type of aquifer; direction of groundwater flow plus:
Low ▼ ▼	Water level at time of sampling	Lake level at time of sampling, depth of thermocline if present	Piezometric level
▼ ▼ ▼	Discharge at time of sampling	As above plus vertical profiles of temp. and O_2 at time of sampling	Piezometric level between sampling, aquifer map
High	Continuous measurement of discharge	As above plus rate of water inflow and patterns of water movement within lake	Full knowledge of groundwater hydrodynamics

Modified from Meybeck *et al.* (1996).

might influence the sampling station locations and depths from which the samples are collected. For an explanation of the hydrology of lakes and reservoirs, *see* **Chapter 108, Lake Ecosystems (Stratification and Seasonal Mixing Processes, Pelagic and Benthic Coupling), Volume 3** and **Chapter 109, Reservoirs, Volume 3**.

The necessity for preliminary information on chemical and ecological interactions depends on the nature of the intended water quality monitoring program. If a biological monitoring approach based on the presence or absence of certain species is to be used (see following text), it may be necessary to determine first of all whether the water body has any natural chemical characteristics that might affect the presence or absence of species to be included in the monitoring program. Similarly, a preliminary survey of the sources and nature of chemical discharges into a water body can assist in the targeted selection of chemical compounds to include in a biological or ecotoxicological monitoring program.

Many of the detailed aspects of program design, such as frequency of sample collection, number of samples, number of replicate samples, and so on, are important to the successful application of certain data analysis methods and statistical techniques. Thus it is important to consider, and even to select, the type of data analysis methods that will be applied to the results obtained before proceeding any further with the monitoring program design (Dixon and Chiswell, 1996). Finally, using all available information on the water body, including that from published sources, previous programs, or specific preliminary surveys, the monitoring program can be designed in detail. The design should consider and specify all the following activities:

- sites to be sampled, that is, location specified by grid reference,
- sample details, for example, precise depth at which each sample is to be taken,
- frequency of sampling,
- variables to be measured in the field, including hydrological variables,
- variables to be measured in samples returned to the laboratory,
- requirements for replicate sampling,
- methods for field analysis,
- methods and equipment for field sample collection and storage,
- methods for laboratory analysis of samples, including any instructions for storage of samples,
- recording procedures and storage methods for results,
- quality control procedures related to both field, laboratory, and data handling operations.

Time taken in specifying and documenting the precise details of each step of the monitoring program can be justified by the resultant confidence in the reliability of the data generated. Over the last decade, much progress has also been made in testing and standardizing aquatic monitoring methods, particularly through the International Organization for Standardization (ISO) (ISO, 2004a).

MONITORING IN THE FIELD

There are four important aspects to the field-based activities within a monitoring program:

- selection of sites at which samples should be taken,
- frequency with which samples should be taken,
- the type of samples to take,
- the methods to be used for sampling.

Selection of sampling locations depends on an understanding of the water body and the objective of the program.

It may also have to take into consideration the practicality of the method, the accessibility of the sites and intended use of the data obtained (e.g. statistical techniques, association with other grid-referenced data, anticipated links with other monitoring activities, etc.).

Programs with the objective of determining baseline or background water quality must be based on sample sites that are distanced from any human influence or, at the very least, be unaffected by the emissions for which they form the baseline or background measurements. Programs monitoring for impacts of particular emissions should include sites close to the emission and, if possible, distributed spatially over the anticipated range of effect. Sampling sites for operational monitoring programs are sometimes defined by the national or international water quality standards or guidelines against which the water quality is being monitored, or by the operational license controlling the activity. Typically, such sites are at the point of water abstraction or discharge. Water quality status monitoring requires sites to be geographically dispersed to cover the area of interest and to include all relevant water bodies, for example, nationally, regionally, or globally. For practical reasons it may be necessary to select one site to be reasonably representative of a large water body, although this assumption should be tested first with a set of synoptic samples collected over a wider area within the water body. Careful choice of such monitoring sites to avoid unusual influences, such as wastewater discharges and outfalls, is essential.

Sampling frequency must be chosen to give adequate data for interpretation of the expected changes in water quality without imposing too great a demand on resources. Long-term trends can be studied by means of infrequent sampling at regular intervals, such as once a year, provided the samples are always taken at the same time of the year to eliminate changes due to seasonal variations in water quality. Sampling to determine fluxes and loads may require high frequency sampling, continuous monitoring techniques or targeted sampling in relation to periods of high flow (see Robertson and Roerish, 1999 for a discussion of the applicability of different water quality sampling strategies). The degree of variability in physical, chemical, and biological aspects of water quality may have to be determined by a preliminary study before deciding on the most appropriate sampling frequency. Groundwaters usually show little variability over time, unless very shallow or influenced by tidal fluctuations, whereas most rivers are highly variable, depending on rainfall and runoff, which in turn is related to climate and season.

The *sampling methods* used depend on the medium to be sampled (water, sediment or biota) and the variables to be measured in the sample (e.g. USGS, 2004a). Chemical and microbiological analyses require samples to be collected without the risk of contamination by any substance or organism that will subsequently be measured in the sample. This can involve special cleaning or sterilization procedures for sampling equipment and sample storage bottles, as well as care in the way the samples are taken. For example, water samples for subsequent analysis of dissolved or particulate metal concentrations should not come into contact with metals, that is, sampling apparatus and sample bottles should be constructed of polyethylene plastic, Teflon, or glass with no metal parts.

Discrete samples of water, sediment, or biota are only representative of the point of sampling at the time that the sample was collected. Hydrological influences, such as flow in rivers and groundwaters or stratification in lakes, may require that either a number of samples are taken at different locations at the same time or that a time- or depth-integrated sample be obtained. Integrated samples can be taken either by mixing a series of discrete samples or by the use of a special sampling technique, for example, a hosepipe sampler, pumping mechanism, or isokinetic sampler (Figure 2). Time-integrated samples are produced by combining discrete water samples from the same site at regular time intervals or by continuously withdrawing water into a large container over a fixed time. This approach is only suitable where the variable being studied does not change biologically or chemically during the time interval between the first and the last sample being taken and mixed together. The precise choice of method should, where possible, be governed by the monitoring objectives and the associated measurement quality objectives.

SAMPLING DIFFERENT MONITORING MEDIA

Depending on the objectives of the monitoring program, it may be appropriate to take water samples, living organisms, sediments, or a combination of these media (see above). For certain types of biological or sediment samples, the same or similar collection methods may be used as for water samples. Specialized approaches may require their own methods and specific precautions to ensure that samples are not contaminated. Some of the more common methods are described here (Table 4, Figure 2).

Water Samples

Discrete water samples, sometimes known as *grab samples*, such as those commonly used for chemical, nutrient, and microbiological analysis, are taken by a field operator and transported to the laboratory. There are several methods available for collecting discrete samples, and the choice depends on the subsequent sample handling or on the need for depth (vertically from top to bottom of a site on the water body) and width (horizontally across the water body) integrated samples (e.g. USGS, 2004a). The simplest method involves submerging a jug or bucket –

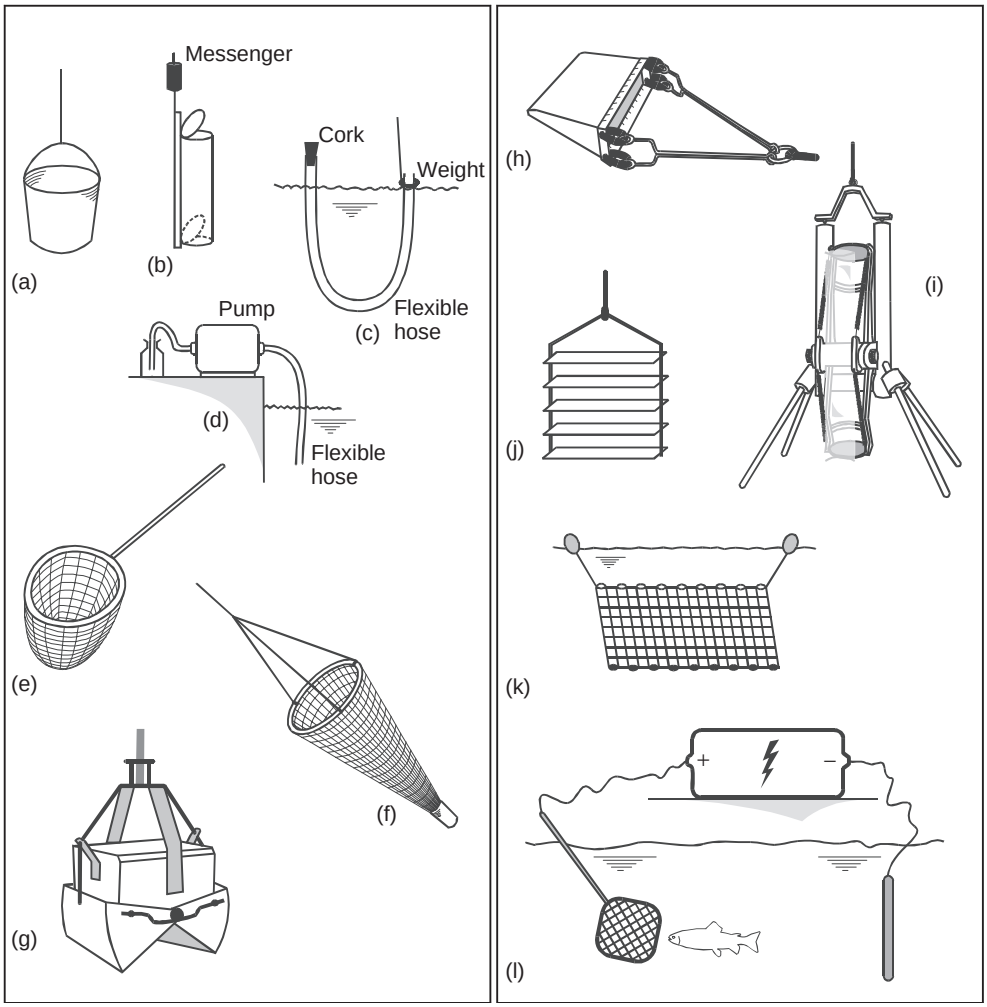


Figure 2 A selection of methods and equipment for collecting discrete or depth- and width-integrated water, sediment, and biota samples. (a) Simple bucket or beaker on a length of rope for collecting surface samples; (b) surface operated bottle sampler for grab samples at depth; (c) flexible pipe for integrating a depth sample in still waters; (d) mechanical or electrical pump for lifting volumes of water from specific depths; (e) fine mesh net on a rigid pole for collecting nonmotile organisms in shallow waters, for example, benthic invertebrates; (f) fine mesh net on a calibrated rope for collecting open water samples of organisms, for example, plankton; (g) mechanically operated grab for collected samples of deposited sediment and associated organisms in deep waters; (h) dredge-type sampler for collecting surface living organisms from shallow or deep waters; (i) remotely operated corer for collecting deposited sediments with minimal disturbance; (j) artificial substrate (e.g. flat plates suspended on a rope) for collecting attached organisms in all types of water body; (k) net or trap for collection of organisms such as fish; (l) electro-fishing equipment for near-quantitative sampling of fish

open face downwards – to the required depth and then slowly turning it upright so that it fills at the sampling depth (Figure 3). More complex methods involve remotely triggered devices, such as the Friedinger and Ruttner samplers (see Table 4, Figure 2), that open at depth to contain a vertical column of water of known volume. For certain advanced analytical measurements, or for variables occurring at very low concentrations, discrete samples are essential, combined with special pretreatment or storage conditions (e.g. addition of fixative chemicals, storage in the cold and dark). Depending on the subsequent analyses to be performed on the sample obtained (e.g. trace metals, microbiological), it may be necessary to take special precautions to avoid contamination of the water sample or the collection of an unrepresentative sample. Such precautions include facing the opening of the collection vessel into the water current, wearing sterile gloves, rinsing the sample bottle several times in the water body to be sampled and discarding the water (away from the collection point) prior to filling, avoiding sampling close to stream or lake beds so as not to disturb fine sediments, and careful cleaning of samplers between sampling trips.

Some water quality variables can be measured *in situ* with portable or fixed submersible probes on site or by

Table 4 Comparison of methods and equipment for sampling water, suspended particles, and aquatic organisms (see Figure 2)

Sampler/sampling mechanism	Figure no.	Type of sample	Most suitable habitats	Advantages	Disadvantages
Jug, bucket, beaker, or bottle	2a	Water, suspended sediment/particles inc. plankton and microorganisms	Lakes and rivers	Cheap, simple, quantitative.	Surface or subsurface samples only.
Bottle samplers (e.g. Friedinger, Van Dorn, Ruttner)	2b	Water, suspended sediment/particles inc. plankton and microorganisms	Still or slow flowing waters, groundwaters	Quantitative. Enables samples to be collected from discrete depths.	Expensive unless manufactured ''in-house''.
Hosepipe sampler	2c	Water, suspended sediment/particles inc. plankton and microorganisms	Still waters	Integrates sample from surface to depth. Quantitative. Cheap and simple to use.	Small volume of sample.
Water pump	2d	Water, suspended sediment/particles inc. plankton and microorganisms	Lakes and rivers, groundwaters	Quantitative if calibrated. Rapid collection of large volume samples. Integrated depth sampling possible.	Expensive. Requires power supply.
Isokinetic sampler		Water and suspended sediment	Flowing waters	Enables flow-related depth-integrated samples	Expensive
Collection by hand		Macrophytes, attached or clinging organisms	River and lake margins, shallow waters, stony substrates	Cheap – no equipment necessary.	Qualitative only. Specific organisms only collected.
Hand net on pole (c. 500 μm mesh)	2e	Benthic invertebrates	Shallow river beds, lake shores	Cheap, simple.	Semiquantitative. Mobile organisms may avoid net.
Plankton net	2f	Phytoplankton and/or zooplankton depending on mesh size	Open waters, mainly lakes	Cheap and simple. Large volume or integrated samples possible.	Qualitative only (unless calibrated with a flowmeter). Selective according to mesh size. Possible damage to organisms.

(*continued overleaf*)

Table 4 (*continued*)

Sampler/sampling mechanism	Figure no.	Type of sample	Most suitable habitats	Advantages	Disadvantages
Grab (e.g. Ekman, Peterson, Van Veen)	2g	Sediments, benthic invertebrates living in or on the sediment. Macrophytes and attached organisms	Sandy or silty sediments, weed zones	Quantitative sample. Minimum disturbance to sample.	Expensive. Requires winch for lowering and raising.
Dredge type (e.g. Surber sampler)	2h	Mainly surface living benthic invertebrates	Bottom sediments of lakes and rivers	Semiquantitative or qualitative analysis depending on sampler	Expensive. Mobile organisms avoid sampler. Natural spatial orientation of organisms disturbed.
Corer (e.g. Jenkins)	2j	Sediments, microorganisms and benthic invertebrates living in sediment	Fine sediments, usually in lakes	Discrete, quantitative samples	Expensive. Small quantity of sample.
Artificial substrates (e.g. glass slides plastic baskets)	2k	Epiphytic algae, attached invertebrate species, benthic invertebrates	Open waters of rivers and lakes, weed zones, bottom substrates	Semiquantitative. Cheap	May not be truly representative of natural communities. Positioning in water body important for successful use.
Poisons (e.g. rotenone)		Fish	Small ponds or river stretches	Total collection of fish species in area sampled	Destructive technique
Fish net/trap	2l	Fish	Open waters, river stretches, lakes	Cheap. Nondestructive.	Selective. Qualitative unless mark recapture techniques used.
Electro-fishing	2m	Fish	Rivers and lake shores	Semiquantitative. Nondestructive.	Selective according to current used and fish size. Expensive. Potential safety risk.

Modified from Friedrich *et al.* (1996).

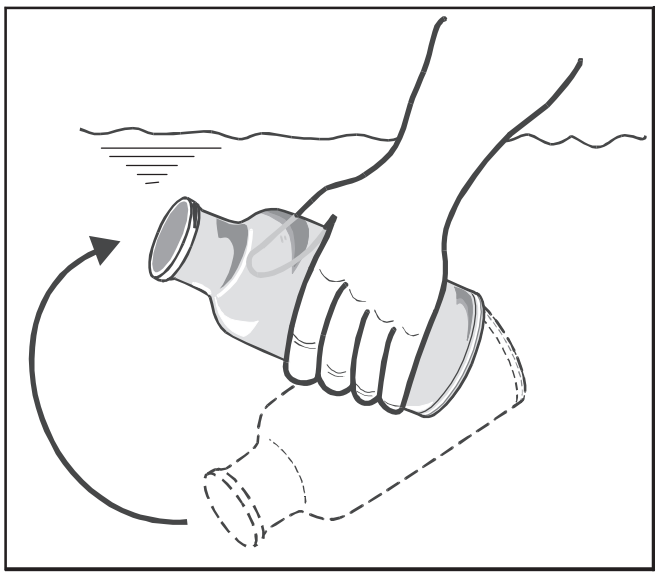


Figure 3 Taking a surface water sample by hand – the bottle is submerged with the opening facing downwards and then tilted up to allow it to fill at the required depth. The bottle should be rinsed several times in the water body before taking the sample and, if used in a river, the opening should be faced upstream. Care should be taken to avoid disturbance of sediment

diverting a flow of water through a monitoring instrument situated very close to the water body. Fixed monitoring instruments provide real-time data on changes occurring within the water body. The outputs can be either stored onto a data logging system and periodically downloaded or transmitted telemetrically to a central data gathering facility. Such systems are currently well developed for measurement of parameters such as velocity, temperature, oxygen, conductivity, pH, total organic carbon, turbidity, and fluorescence (Glasgow *et al.*, 2004). These types of measurements provide an essential early warning mechanism of changes in water quality at sites of water abstraction for municipal or industrial use. Permanent hydrometric stations provide real-time information of discharge that is essential for flood control planning and efficient operation of hydropower generation facilities in small catchments (e.g. NRC, 2004; USGS, 2004b).

Automated sample collection systems are available that divert water from the water body to a unit which fills sample bottles at preset time intervals or when remotely triggered to do so. Such sample collection mechanisms are particularly useful in remote locations where much time and effort may be spent traveling to and from the sampling site, especially if the required sampling frequency is high such as when estimates are being made of rainfall runoff or when loads are being calculated. The range of variables that can be determined in such samples is sometimes limited to those that will not be affected by the time delay between automatic collection and subsequent analysis in the laboratory, although some sampling systems enable the addition of chemical fixatives *in situ* and can refrigerate the samples until collection and analysis. Some automated sample collection systems can also be coupled to automatic analyzers capable of making a limited range of measurements within defined limits of accuracy (Glasgow *et al.*, 2004). All such remote and automated equipments do, nevertheless, require occasional maintenance and could also be subject to undesirable human interference.

Sediment

Sediments and particulate matter can be responsible for transporting and storing contaminants and nutrients in rivers and lakes. Monitoring programs involving sediment samples may be designed to determine, for example, the transport of contaminants in suspended sediment, storage of nutrients in surficial deposited sediment or the historical record of accumulation of contaminants over time. These three objectives require different approaches to the collection of the samples for analysis. The concentration of suspended sediments in rivers is closely related to discharge because of turbulent resuspension, especially during floods; thus it is essential that the discharge is known or measured at the time of sampling. Suspended sediment samples are usually collected in rivers by taking water samples using one of the methods mentioned above. The water samples can then be filtered or centrifuged to concentrate the sediment from the known volume of water sampled. In lakes, suspended sediment concentrations are determined from known volume water samples or using sediment traps. A simple sediment trap consists of a tube with a conical base ending in a collection vessel; the trap is usually suspended in the water column for a period of time ranging from hours to weeks. The tube collects all suspended material that is settling in the water column and can provide an estimate of sedimentation rates over time.

Sampling methods can affect data interpretation; for example, SPM is sampled by horizontal isokinetic samplers (to sample in such a way that the water-sediment mixture moves with no change in velocity as it leaves the ambient flow and enters the sampler intake) at several verticals in a cross-section of a river. SPM differs from TSS in a river sample because TSS samples are collected at one point on the river surface. Although, TSS may be very different from SPM for the same site at the same time, TSS are routinely used as a substitute for SPM to determine particulates fluxes because they are easier and much cheaper to obtain.

Samples of surficial deposited sediments in lakes or rivers can be collected by grabbing a volume of sediment with a scoop by hand (where the water is sufficiently shallow) or, in deeper waters, by using a remotely triggered grab sampler (Figure 2). Undisturbed samples that can be used to study the profile of grain size or chemical content with

depth in the sediment (e.g. for historical studies) can be obtained using a coring device (Elliott and Tullett, 1978) (Figure 2). A full discussion of the use of particulate material in monitoring programs is available in Thomas and Meybeck (1996).

Biological Samples

There are many ways in which biological samples are incorporated in water quality monitoring programs (Table 5) and the choice of sampling method is dictated by the biological approach being used. Most methods are either (i) passive, in which organisms are collected from their own environment for enumeration or analysis or (ii) active, in which organisms are deliberately placed into a particular environment and then removed for analysis at later intervals. Examples of some biological sampling methods are given in Table 4. Microbiological monitoring to determine whether water is of adequate quality for human consumption is the most widely performed water quality monitoring activity; it is carried out on surface and groundwaters, as well as treated water, prior to distribution to domestic users. Microbiological monitoring may also be included in monitoring programs designed to detect the impact of wastewater discharges, runoff from agricultural activity or the quality status of water resources. Samples must be collected, handled, and processed with special care to prevent contamination (see, for example, Myers and Wilde, 2003 for a detailed description of methods). Another common biological monitoring approach involves identifying and counting the invertebrates present at specific sites in a river and calculating a water quality index or pollution score (see Section on "Examples of monitoring programs"). This approach is based on the principle that certain organisms have particular environmental preferences or tolerances and can act as indicators of defined quality ranges (sometimes broadly correlated with physical and chemical quality indicators) (see following text). The index or score system usually prescribes the sampling and analysis method – the most common method of collecting samples is with a net, which provides a semiquantitative sample. Some of the more widely accepted methods have now been standardized by the ISO and by relevant national authorities. Planktonic organisms, suspended in the water column, can also be sampled by filtration, centrifugation, or sedimentation of a known volume of water sample collected by one of the water sampling methods described above and in Table 4. Particular groups of organisms can be collected by encouraging them to adhere to, or live on or in artificial substrates, such as glass or plastic plates with flat surfaces or baskets of stones (see Figure 2) (Chapman and Jackson, 1996).

Some groups of relatively immobile aquatic organisms provide useful media for studying the long-term pollution or recovery of water bodies with contaminants such as heavy metals or organic chemicals. These substances can be difficult to measure at low concentrations in the water, but are accumulated within the tissues of the organisms at concentrations higher than that found in the water itself. The higher concentrations can often be analyzed with less sophisticated equipment and thus enable detection in situations that might otherwise have been limited by available analytical equipment. Some organisms, such as those that feed on fine organic particles, can assimilate both dissolved and particulate contaminants, while others may specifically reflect the presence of substances dissolved in the water, or of accumulation in the aquatic food chain. When designing a monitoring program aimed at studying concentrations of contaminants in organisms it is important that the natural or background concentrations of the contaminant of interest, *in the specific tissues of the monitored organism*, should be known or determined in advance of the sampling program. An alternative approach is to use active monitoring, where organisms from a single batch or reference site, and with known background concentrations of the contaminant of interest, are placed in the water body to be monitored. Subsamples are then taken at time intervals for analysis.

EXAMPLES OF MONITORING PROGRAMS

The complexity of a monitoring program can range from regular measurements of one or two variables at a single site over a long time period to many physical, chemical, and biological variables at numerous sites and at varying time intervals, depending on the objectives. Every monitoring program should be designed to answer specific questions, that is, to generate data that can be used for research, management, or policy development. For this reason, there are many examples of different types of monitoring programs (Table 1) at local, national, regional, and even global scales (see Dixon and Chiswell, 1996 for a review of monitoring program design). Table 6 presents a selection of current examples of monitoring programs from around the world. More common approaches to water quality monitoring are highlighted in this section by reference to some specific examples.

Programs Based on Physical and Chemical Measurements

Some physical measurements can be very simple and inexpensive to perform, requiring relatively little training or expertise, and for these reasons they are incorporated in most monitoring programs (irrespective of whether they will yield useful information or not!). Methods range from the simple turbidity or transparency tube (e.g. Kent State University, 2004a) and Secchi disc (see following text) to handheld digital meters and *in situ* continuous monitoring

Table 5 Uses of biological methods for different types of water quality monitoring programs and their relative advantages and disadvantages

	Ecological methods		Microbiological methods	Physiological and biochemical methods	Bioassays and toxicity tests	Chemical analysis of biota	Histological and morphological methods
	Indicator species[a]	Community studies[b]					
Principal organisms used	Invertebrates, plants, and algae	Invertebrates	Bacteria and viruses	Invertebrates, algae, fish	Invertebrates, fish	Fish, shellfish, plants	Fish, invertebrates
Types of monitoring	Water quality status, impact surveys, trend monitoring	Impact surveys, trend monitoring	Operational surveillance, impact surveys	Early warning monitoring, impact surveys	Operational surveillance, early warning monitoring, impact surveys	Impact surveys, trend monitoring	Impact surveys, early warning monitoring
Types of pollution or other human impacts detected	Organic matter pollution, nutrient enrichment, acidification	Organic matter pollution, toxic wastes, nutrient enrichment	Human and animal fecal waste, organic matter pollution	Organic matter pollution, nutrient enrichment, toxic wastes	Toxic wastes, pesticide pollution, organic matter pollution	Toxic wastes, pesticide pollution, human health risks (toxic contaminants)	Toxic wastes, organic matter pollution, pesticide pollution
Advantages	Simple to perform. Relatively cheap. No special equipment needed. Trained biologist/ecologist may be necessary	Simple to perform. Relatively cheap. No special equipment needed. Minimal biological expertise required	Can indicate risk to human health. Simple to perform. Relatively cheap. Very little special equipment required	Usually very sensitive. Commercial kits available for some methods. Cheap and expensive options. Some methods allow continuous monitoring	Usually simple to perform. Minimal equipment requirements. Fast results. Relatively cheap. Some continuous monitoring possible	Can indicate risk to human health. Requires less advanced equipment than for the chemical analysis of water samples	Some methods very sensitive. Simple and complex methods available. Cheap or expensive options
Disadvantages	Some methods only applicable within limited geographical area. Knowledge of taxonomy required. Results can be influenced by natural changes in aquatic environment	Relevance to aquatic systems not always tested. Results can be influenced by natural changes in aquatic environment	Organisms can be easily transported in water and thus it may be difficult to relate positive results to a specific pollution source	Specialized knowledge and techniques required for some methods	Laboratory-based tests not always indicative of field conditions	Analytical equipment and well trained personnel necessary. Generally expensive	Specialized knowledge required. Some special equipment needed for certain methods

[a]For example, biotic indices.
[b]For example, diversity or similarity indices. Modified from Friedrich *et al.* (1996).

Table 6 Examples of different types of water quality monitoring programs

Program type or name	Organizing agencies	Main aims or objectives	Key approaches	Further information
National water quality assessment program	Water Science Agencies (e.g. US Geological Survey), Environmental Protection Agencies, National Water Authorities	National water quality status and trends for rivers	Annual survey using water quality index based on benthic invertebrates	e.g. **http://water.usgs.gov/nawqa/** **http://www.epa.ie/Water/**
Great North American Secchi Dip-In	US EPA/North American Lake Management Society/Kent State University	National trends in eutrophication status of lakes and reservoirs	Citizen involvement; Annual measurement of transparency using a Secchi disc	**http://dipin.kent.edu**
Volunteer Lake Monitoring/Volunteer Stream Monitoring	US EPA Office of Water	National water quality status of lakes and of streams and rivers	Citizen involvement. Physicochemical and biological measurements at regular (e.g. monthly) intervals	**http://www.epa.gov/volunteer/** **http://www.ccmn.ca/english/**
GEMS/Water	United Nations Environment Programme (UNEP) and others	Assessment of global water quality and trends for rivers and lakes	Physical and chemical measurements at regular intervals. Results submitted to central database.	**http://www.gemswater.org/**
Lake recovery following controls on nutrient inputs	Lake Commissions and Environment Agencies	Trends in nutrients and trophic status	Nutrient concentrations and chlorophyll a	Commission internationale pour la protection des eaux du Léman, Lausanne
Suitability for withdrawal and treatment for public supply	All public water supply authorities in European Union (EU) countries	Compliance with quality criteria for water abstraction for public supply	Physical and chemical measurements as directed by EU Directives	**http://europa.eu.int/eur-lex/en/index.html**
Suitability for use in industrial or food processing	Individual process plants	Assessment of compliance with quality criteria of industrial process	Continuous online measurement of TOC, NH_4, pH	Numerous commercial examples
Early warning of contaminant upstream of water abstraction point	e.g. International Commission for the Protection of the Rhine	Warning of the presence of toxic chemicals or substances	Continuous bioassay using fish and *Daphnia* sp	**http://www.iksr.org/**

instruments measuring, for example, temperature, oxygen, conductivity (as an indicator of total dissolved solids) or pH.

An indication of the trophic state of lakes, reservoirs, and ponds can be obtained from the measurement of transparency. Transparency is the ability of light to pass through the water column and is affected by particulate and dissolved material present in the water. It can be measured optically with a light meter or approximated by using a Secchi Disc (a flat round disc painted in black and white quarter segments). The depth at which the quartered disc disappears and reappears from sight when lowered and then raised in its horizontal position, is known as the *Secchi depth*. In eutrophic lakes, phytoplankton densities in the upper water layers can be very high, leading to reduced Secchi depth. Thus, measurements of Secchi depth provide a simple, inexpensive indicator of spatial or temporal variations in phytoplankton population density (and other suspended material) in lakes when biological expertise to identify characteristic species associated with nutrient enrichment is lacking. This simple technique is being used for a long-term monitoring program aimed at examining spatial differences and trends in the transparency of lakes throughout North America. Monitoring is performed on a designated day each year by volunteers and the results are submitted to a central database via the Internet. The program known as the *Great North American Secchi Dip-In* is coordinated by Kent State University, with the sponsorship of the United States Environmental Protection Agency (US EPA) and the North American Lake Management Society and commenced in 1994. Data are now available on more than 5900 water bodies (Kent State University, 2004b)

Carefully selected chemical measurements can often be used to indicate the impact of certain human activities on water quality. Typical examples are phosphorus, which is usually associated with sewage and domestic wastewater discharges, and nitrates, which are frequently associated with agricultural activities such as the use of inorganic fertilizers. Regular measurements of total phosphorus in Lake Geneva from the 1950s to the present time indicate the improvement in water quality resulting from efforts to control phosphorus in wastewater discharges (Figure 4). Monitoring water quality to determine the presence of, and/or recovery from, wastewater discharges can be based either on one or more selected chemicals that are known to occur in the effluent, such as a specific organic compound, or a general indicator measurement such as total organic carbon that can be automated and monitored continuously. For most regular monitoring activities, water samples are taken to the laboratory for analysis, but portable instruments exist for selected chemical measurements, such as nitrates. Although these instruments often have a narrow range of measurement and limited precision, they can be useful for emergency water quality monitoring, for obtaining a rough indication of a likely problem requiring more detailed investigation or for occasions where large distances need to be covered between sites without recourse to an analytical laboratory.

In some countries where technical resources are limited, water quality monitoring in river stretches affected by numerous industrial effluents or multiple sources of pollution is based on easily performed physical and chemical, and sometimes simple biological, measurements. The results from each site are compiled into a water quality index (usually based on a numerical score, e.g. 1–100) that provides a simple indication to policy makers of whether conditions are improving or getting worse. The Oregon Water Quality Index is just one of many examples – it is based on eight water quality variables (temperature, dissolved oxygen, BOD, pH, ammonia + nitrate nitrogen,

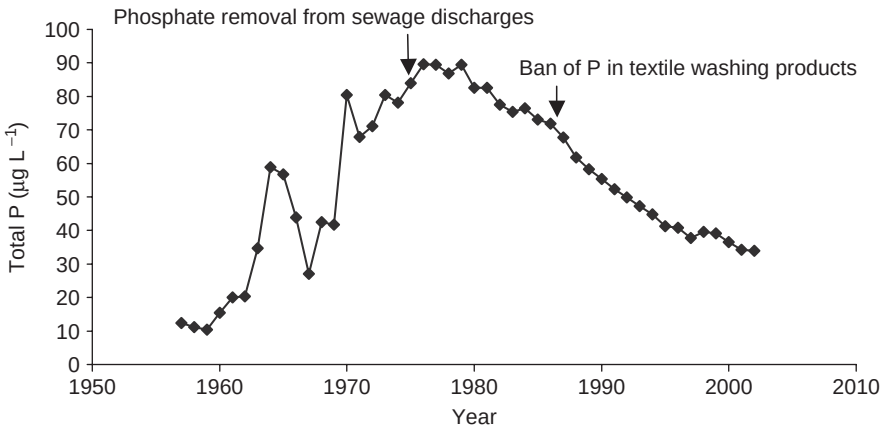


Figure 4 Decreasing concentrations of mean total phosphorus in Lake Geneva, reflecting improvements in water quality because of controls applied to wastewater discharges between 1973 and 2000 (Data from Commission Internationale pour la Protection des Eaux du Léman contre la pollution (CIPEL)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

total phosphorus, total solids, and fecal coliforms) and has been in use since the 1970s, although it has been updated recently to reflect recent understanding of water quality (Cude, 2001).

Biological Monitoring of Water Quality

Any change in water quality (whether natural or human-induced) usually affects, to some degree, the aquatic biota associated with the water body. Such effects, ranging from the death of sensitive species, to changes in the aquatic community structure, to metabolic changes or reproductive disorders and the accumulation of toxic substances in cells and tissues, have been exploited in many ways as a means of monitoring water quality or the "health" of a water body. Biological monitoring and assessment is now widely accepted as an effective means of determining the response of an aquatic system to human interference and different methods and approaches are being routinely included in many state, national, and regional assessments of water quality. In addition to water quality criteria based on physical and chemical measurements, biological quality criteria are now being built into regulatory standards to monitor point and nonpoint sources of pollutants into surface waters (Davis and Simon, 1995).

Even a quick visual or qualitative assessment of the biota present in a water body can give a rough evaluation of water quality or trophic status, especially if carried out by an experienced aquatic biologist. Such approaches range from simply identifying the common macrophytes present (if any), to lifting stones and examining them for specific invertebrate species, to taking semiquantitative benthic invertebrate samples with a net on a long pole. Such basic monitoring approaches can be developed into numerical indices that rank water quality, such as the many diversity, biotic, or similarity indices that are based on mathematical theory (Washington, 1984). One such example is the use of indicator species of macrophytes to assess the trophic status of rivers in the United Kingdom that are subject to wastewater discharges from sewage treatment works (Dawson *et al.*, 2000).

Benthic invertebrates are particularly suitable for monitoring water quality impaired by the presence of biodegradable organic matter because many species are sensitive to depletion of dissolved oxygen and they are relatively immobile. There are numerous examples of monitoring programs based on the presence or absence of sensitive benthic species, known as *indicator species*, and each has been (or should be) tailored to meet national or regional objectives. In general, the method is applicable to a wider geographical area when the score is based on the family level of identification as in the Biological Monitoring Working Party Score (BMWP) (ISO-BMWP, 1979).

Indictor species can even be used for rapid, on-site bioassessment of the effects of biodegradable organic wastes in rivers. Typically, benthic invertebrates are collected in shallow stretches of the rivers by kicking up the substrate and collecting dislodged invertebrates in a net on a long pole (Figure 2). The relative abundance of easily identifiable indicator species is noted whilst still in the field. The presence or absence of certain species, together with their relative abundance, enables the sample to be assigned an index. In Ireland, the sample is assigned a Q value ranging from Q1 to Q5 – the higher the Q value, the better the species community diversity and the better the water quality. Where additional observations are made, such as dissolved oxygen measurements, the presence of silt, and so on, the Q value can be assigned more accurately to correspond with four major water quality classes (McGarrigle *et al.*, 2002):

Class A Unpolluted (Q5, Q4–5, Q4); Class B Slightly polluted (Q3–4); Class C Moderately polluted (Q3, Q2–3); Class D Seriously polluted (Q2, Q1–2, Q1). This approach forms the basis of the Irish national river water quality monitoring program where designated sites are sampled at the same time of the summer each year. The results provide the data for the annual reports of water quality that identify trends in impacts from agriculture and other sources of biodegradable organic waste, such as sewage discharges, and the food processing and dairy industries (McGarrigle *et al.*, 2002; Clabby *et al.*, 2003; EPA, 2004). The simple classification enables policy makers and the public to see whether river water quality is improving or not (Figure 5). The cost-effectiveness of such approaches makes them particularly attractive for use in developing countries, but it is essential that the method is validated in the region of interest before being put into widespread use (e.g. Henne *et al.*, 2002).

Recent developments in the use of biological monitoring include the setting of biological criteria for water bodies (equivalent to water quality criteria based on physical and chemical standards) based on a variety of numerical indices (e.g. diversity indices (Washington, 1984)), indicator species, fish stock assessments, and so on. An example of the development of such indices has been presented by Yoder and Rankin (1995). The approach has been endorsed by the USEPA (US EPA, 1999a; Hall *et al.*, 2000) as a means of describing reference conditions of aquatic communities inhabiting water bodies that have been designated for specific uses. Their use can aid the detection of impacts on water bodies, indicate whether a water body is meeting the quality criteria for its designated uses and indicate whether additional monitoring is necessary (Simon, 2000). The success of these criteria is based on the widely accepted premise that biological data are a better predictor of environmental impacts than chemical or toxicological data, especially when other information on the nature of the impact is lacking (Simon, 2000).

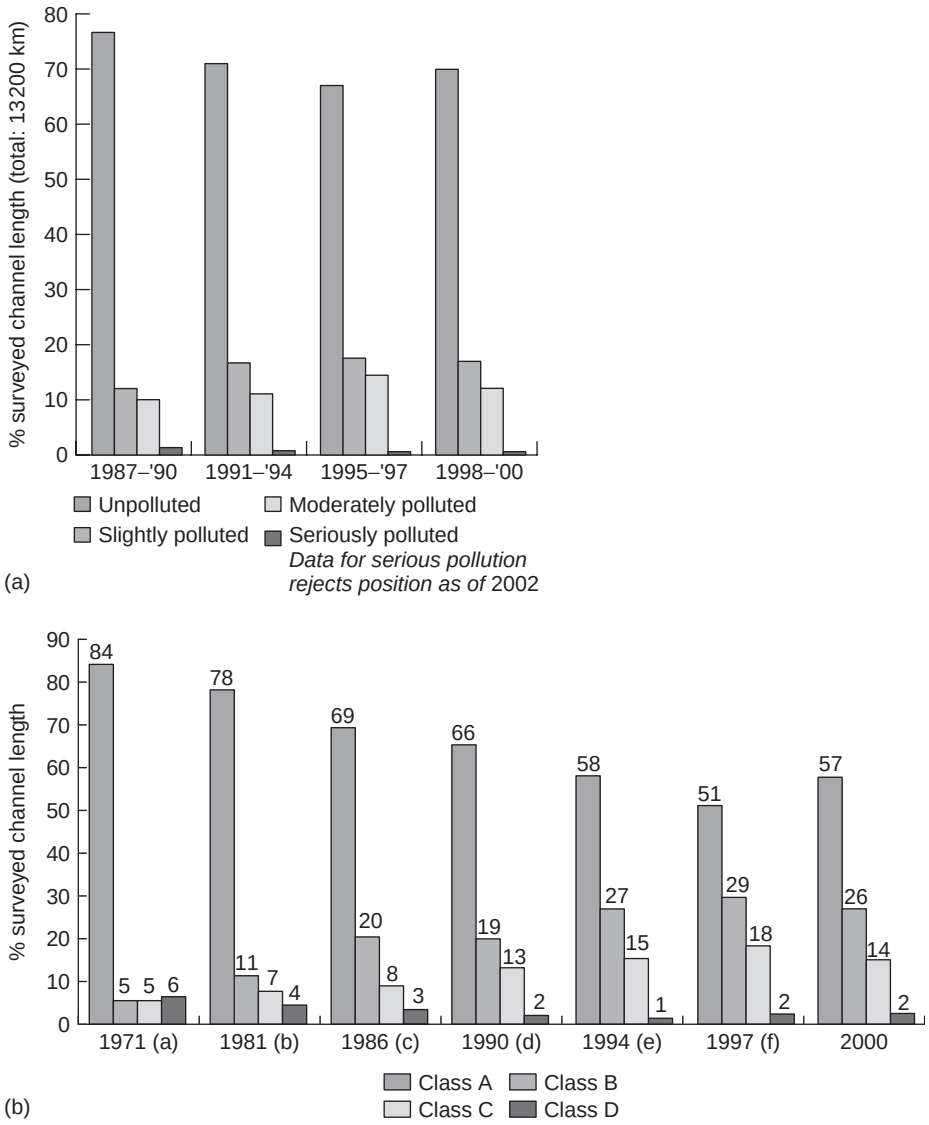


Figure 5 Changes in river water quality in Ireland as indicated by a biological assessment procedure. From McGarrigle *et al.*, 2002 (with permission from the Environmental Protection Agency) (**`http://www.epa.ie/NewsCentre/ReportsPublications/IrelandsEnvironment2004/`**). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Monitoring Toxic Pollution

Deliberate or accidental discharges of toxic compounds into water bodies upset the natural water quality, present a risk to the aquatic ecosystem, and may pose a threat to human health if the water is to be extracted for municipal use. Monitoring designed to establish the spatial and temporal extent of the discharge may take the form of a limited duration survey if the discharge is short-lived, but could also involve routine evaluation of the water quality if the discharge is continuous or long-term (as in some industrial discharges). Measuring chemical compounds can be expensive and require technically advanced equipment. Thus, when the discharge has known toxic potential, or when a combination of compounds could lead to toxic impacts, it can be simpler (and sometimes cheaper) to use a routine bioassay technique. Many toxicity and bioassay methods have been standardized (see ISO, 2004a) and some, such as the *Salmonella* mutagenicity assay (Ames test), have been incorporated into local or regional water quality monitoring programs, for example, São Paulo State, Brazil (Umbuzeiro *et al.*, 2001). Changes in the activity of fish in response to adverse environmental conditions, that is, a change in water quality, have been incorporated into biological early warning systems that monitor water quality upstream of water intake points on highly industrialized rivers. The river water is diverted

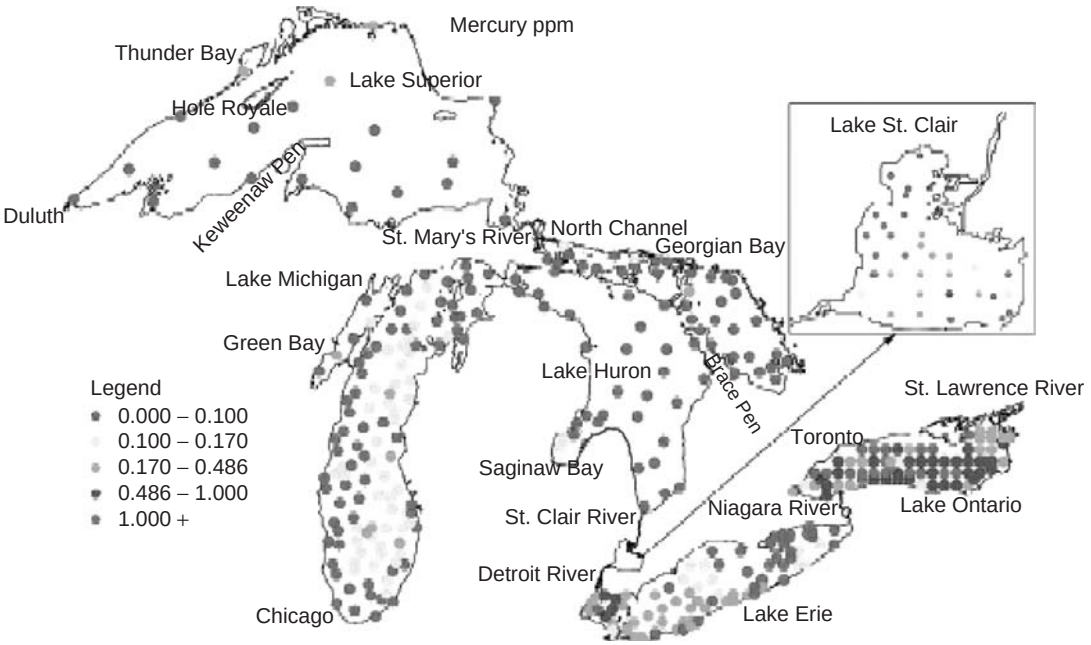


Figure 6 Total mercury concentrations in surficial sediments of the Great lakes (1997–2000). Spatial and temporal trends in surface water and sediment contamination in the Laurentian Great Lakes. *Environmental Pollution*, **129**, 131–144. (Reprinted from Marvin *et al.*, 2004. © 2004, with permission from Elsevier). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

through tanks containing live fish and detectors register the levels of their activity.

Potentially toxic contaminants released to the environment from various human activities are transported with atmospheric deposition, runoff or direct discharge to surface water bodies. Persistent substances can accumulate in sediments to concentrations above those in the water itself and which can be more easily detected. Deposited sediments can also be used to illustrate historical changes in the concentration of persistent compounds or metals in the environment (e.g. Van Metre *et al.*, 1998). Sediments have been used in the long-term and spatial monitoring of contaminants such as mercury in the Great lakes of North America and Canada (Figure 6) (Marvin *et al.*, 2004)

Community and Citizen Monitoring Programs

The growing demand for information about water quality from policy makers and the public is stretching the resources of many government agencies and local authorities, and they can no longer keep up with the demand for monitoring data. One approach to meeting the demand for information while minimizing costs, is to involve local people or other interested citizens in monitoring activities. Such volunteer programs, that range from simple activities such as the Great North American Secchi Dip-In (see above) to detailed lake and stream monitoring (e.g. US EPA Office of Water; Canadian Community Monitoring Network) are proving to be extremely valuable for generating water quality information and even for getting local people involved in the management of their water resources. Training programs and instruction manuals for citizens are available from many State agencies (e.g. US EPA, 1997, 2004; CCMN, 2004).

DATA HANDLING AND REPORTING

Data handling and reporting are the final steps of the water quality assessment process (see Figure 1). In the past, there was a tendency to archive data sets generated from monitoring activities with very little scrutiny or interpretation of the data. As a result, the data generated were not transformed into useful information and rarely served the purpose of management or policy generation or refinement (Ward *et al.*, 1986). Data collection is the principal objective of all monitoring activity and in order to make the monitoring effort worthwhile, the data generated must be assessed or synthesized to provide meaningful information for management, policy guidance, or public use. Advances in computer-based data storage and handling facilities have not only made it easier to share data and to analyze it and present it in many different ways, but they have also created a stronger need to ensure the quality, compatibility, and comparability of data made freely available, especially through the Internet. Developments in online accessibility to databases, combined with a need for public accountability, has led to many international, national, or state agencies

providing access to raw and/or synthesized data through the Internet (e.g. EEA, 2004; USGS 2004e). All data made available for public use must be reliable (see above regarding quality of data).

Modern monitoring techniques often result in data that is stored directly onto the electronic media from field or analytical equipment, using personal digital assistants (PDAs), field computers, data loggers, telemetric transfer mechanisms or cabled links. In addition, the widespread availability of inexpensive, handheld Global Positioning System (GPS) units has led to greater use of Geographic Information Systems (GIS) in data storage and presentation. Nevertheless, some information such as that recorded in the field by a field operator may have to be entered into a database manually and it is therefore essential that quality assurance procedures are applied as rigorously to data storage and handling activities as to any other step in the complete monitoring process (see following text). Links should be maintained between all relevant data that could have a bearing on the interpretation of any particular data set so that misinterpretation is avoided. For example, it is important that discharge data are available in association with water quality data for rivers; this might necessitate combining data from a hydrometric database with that from a different water quality database.

QUALITY ASSURANCE AND CONTROL

The data generated from a monitoring program must be of high quality, reliable, credible and, as far as possible, compatible with data generated from similar monitoring programs. This should be achieved through a rigorous process of quality assurance applied at all stages of the monitoring process, from sample collection in the field to data storage. It has been suggested by Meybeck *et al.* (1996) that approximately 10–20% of the total financial, technical, and personnel resources available in any monitoring program should be devoted to quality assurance procedures. For data compatibility with other monitoring programs, it is useful to adopt standardized methods, such as those published by the International Organization for Standardization (ISO) (ISO, 2004a), the German Industrial Standards (DIN – Deutsche Industrie-Normen), US EPA or other national standards where available. If accreditation and validation by a third party is not appropriate (perhaps because of the expense involved) quality assurance can be implemented by reference to international standards such as ISO 9000 or ISO 14000 (ISO, 2004b).

Whereas quality control may be very familiar to those working in an analytical laboratory, it is much less so to workers engaged in field work or nonchemical monitoring methods. Quality control for field sampling and measurements can be achieved by thorough training and the provision of detailed written operation procedures or manuals. Such procedures should also specify the appropriate storage and handling for the analysis that is to follow, ensuring that no deterioration or changes in composition can take place that might influence the eventual analytical results. Field note books or record sheets should always be used to record all necessary information in the field, such as the time, date, and location of sample collection, depth, method used, measurements taken, deviations from standard procedures, and so on. These notebooks can help variations from expected analytical results to be checked for possible explanations, such as unforeseen environmental influences, deviation from standard methodology or sampling site, and so on. From the point of sampling, all samples should be logged with a unique identification code that accompanies the sample through its handling and analysis in the laboratory to the final output and storage of results. In this way it is possible to trace the history of each sample including, for example, whether the sample was split for different analyses, diluted, concentrated, and so on. All equipment, whether field or laboratory, should be properly maintained and periodically calibrated (where appropriate) in accordance with the manufacturers' recommendations. Records of these activities should also be kept.

Analytical Quality Control

Analytical quality control is well established in many laboratories involved in water quality monitoring programs and demonstrates that the laboratory is producing data of adequate precision, accuracy, and sensitivity. It is based on a system of traceability and feedback; for this purpose a laboratory logbook should be maintained for all analytical procedures to which samples are submitted. There are two main aspects to analytical quality control: internal quality control and external quality control. The former involves the choice of method appropriate to the objectives of the monitoring program, and the validation of the method. Validation includes determining (i) the linearity of the calibration, (ii) the limit of detection, (iii) within-day or day to day precision, and (iv) the accuracy of the method. Typical methods for checking validity include:

- the inclusion of a blank sample, for example, distilled water, in a batch of analyses,
- duplicate analyses carried out on the same sample, and
- the use of certified reference materials.

The use of reference materials in routine analysis and the production of Shewhart charts enable a continuous check to be kept on the precision and accuracy of the technique (Briggs, 1996). Where problems are encountered, they should be addressed.

External quality control is particularly important where different laboratories contribute results to a single monitoring program and for which comparability between the

data generated by the participating laboratories is important (e.g. USGS, 2004c). External quality control also provides a method by which a laboratory can have its own accuracy checked independently (e.g. USGS, 2004d). Samples with known and unknown concentrations of the relevant variables are distributed to participating laboratories from a single reference laboratory. Each laboratory analyses the samples and reports its results to the reference laboratory. The deviation of each laboratory from the target value of each individual analysis is reported to the participating laboratory together with comments on whether the accuracy of the results submitted is satisfactory or not (Briggs, 1996; USGS, 2003). If the results show poor accuracy, measures should be taken by the laboratory to improve performance and the level of confidence associated with the results from that laboratory should be associated with any data made available for comparison with results from other laboratories. Laboratories providing data to national or regional databases may be required to take part in national accreditation programs (e.g. NWQMC, 2004b,c). As a general rule about 10% of all analyses should include external quality assurance samples.

Data Checking and Validation

The full process of quality control in a monitoring program should extend to the handling and storage of the data. Results should be scrutinized as soon as they are generated and before they are stored in a database or entered into reports. Inexplicable results should either be omitted from reports or flagged and, where possible, a degree of confidence assigned. The sudden or unexpected occurrence of unusually high or low values should automatically initiate checks on all stages of the sample handling and processing from collection in the field to final analysis in the laboratory. Events that could have resulted in unusual results, such as flood conditions in a river, change in the method of sampling, different analytical procedure, and so on, should have been recorded in field and laboratory notebooks. Where no explanation can be found, equipment and reagents should be checked and instruments should be recalibrated. Errors can be introduced in the manual copying of instrumental outputs or readings and their transcription from one notebook to another or onto a computer through keyboard entry. Many modern instruments will connect directly to the computer and transfer data electronically, thus reducing the risk of transcription errors. Databases for storing monitoring data can be arranged to highlight results that do not fall within expected ranges and this acts as another check on their validity.

CONCLUSIONS

Monitoring programs to define water quality occur at local, national, regional, and even global scales. These programs take many forms depending on the specific program objectives, varying from highly specific measurements associated with regulatory compliance to broad, multidisciplinary programs aimed at defining environmental quality status. The methods and approaches that are currently in widespread use vary from simple physical measurements to determination of trace concentrations of complex organic chemicals, from identifying the presence or absence of key indicator organisms to measuring physiological and biochemical processes in organisms, and from annual sampling events to determine long-term trends to intensive or continuous discharge-linked measurements to determine fluxes. Modern requirements to make monitoring data and their interpretation available to diverse users, including policy makers, other monitoring programs and often the general public, has placed an even greater emphasis on the need for quality assurance of monitoring data. Quality assurance techniques need to be applied to all of the steps involved in the gathering of monitoring data – from field sample collection, to laboratory analysis, to data handling and storage. This overview has highlighted some of the diverse approaches currently taken to monitoring water quality and the key steps involved in the monitoring processes.

REFERENCES

Briggs R. (1996) Analytical quality assurance. In *Water Quality Monitoring: A Practical Guide to the Design and Implementation of Fresh Water Quality Studies and Monitoring Programmes*, Bartram J. and Ballance R. (Eds.), E & FN Spon: London, pp. 215–236.

CCMN (2004) Canadian Community Monitoring Network home page. **http://www.ccmn.ca/english/** [26 October 2004]

Chapman D.V. (1996) Water-quality monitoring. In *Water Resources. Environmental Planning, Management and Development*, Biswas A.K. (Ed.), McGraw-Hill: New York, pp. 209–248.

Chapman D.V. and Jackson J. (1996) Biological monitoring. In *Water Quality Monitoring. A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes*, Bartram J. and Ballance R. (Eds.), E & FN Spon: London, pp. 263–302.

Clabby K.J., Lucey J. and McGarrigle M.L. (2003) *Interim Report on the Biological Survey of River Quality. Results of the 2002 Investigations*, Environmental Protection Agency: Wexford. [available at **http://www.epa.ie/Water/**]

Cude C.G. (2001) Oregon Water Quality Index: A tool for evaluating water quality management effectiveness. *Journal of the American Water Resources Association*, **37**(1), 125–137.

Davis W.S. and Simon T.P. (1995) *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*, Lewis Press: Boca Raton.

Dawson F.H., Newman J.R., Gravelle M.J., Rouen K.J. and Henville P. (2000) Assessment of trophic status of rivers using macrophytes. *Evaluation of the mean trophic rank*. R&D Technical Report E39, Environment Agency, UK, 178 pp.

Dixon W. and Chiswell M. (1996) Review of aquatic monitoring programme design. *Water Research*, **30**(9), 1935–1948.

EEA (2004) European Environment Agency data service. **http://dataservice.eea.eu.int/dataservice/** [26 October 2004]

Elliott J.M. and Tullett P.A. (1978) *A Bibliography of Samplers for Benthic Invertebrates*, Occasional Publication No. 4, Freshwater Biological Association: Ambleside.

EPA (2004) *Irelands Environment 2004*. **http://www.epa.ie/NewsCentre/ReportsPublications/IrelandsEnvironment2004/** [28 October 2004]

EU (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. *Official Journal of the European Communities*, **L327**, 1–72.

Friedrich G., Chapman D. and Beim A. (1996) The use of biological material. In *Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring*, *Second Edition*, Chapman D. (Ed.), E&FN Spon: London, pp. 175–242.

Glasgow H.B., Burkeholder J.M., Reed R.E., Lewitus A.J. and Kleinman J.E. (2004) Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies. *Journal of Experimental Marine Biology and Ecology*, **300**, 409–448.

Hall R.K., Wolinsky G.A., Husby P., Harrington J., Spindler P., Vargas K. and Smith G. (2000) Status of aquatic bioassessment in US EPA region IX. *Environmental Monitoring and Assessment*, **64**, 17–30.

Henne L.J., Schneider D.W. and Martinez R. (2002) Rapid assessment of organic pollution in a west-central Mexican river using a family-level biotic index. *Journal of Environmental Planning and Management*, **45**(5), 613–632.

ISO (2004a) home page **http://www.iso.org/iso/en/ISOOnline.frontpage** [26 October 2004]

ISO (2004b) ISO 9000 and ISO 14000 – In Brief **http://www.iso.org/iso/en/iso9000-14000/index.html** [26 October 2004]

ISO-BMWP (1979) *Assessment of the Biological Quality of Rivers by a Macroinvertebrate Score*, ISO/TC147/SC5/WG6/N5, International Organization for Standardization: Geneva.

Kent State University (2004a) **http://dipin.kent.edu/Transparency_Tube.htm** [22 October 2004]

Kent State University (2004b) **http://dipin.kent.edu** [22 October 2004]

Marvin C., Painter S., Williams D., Richardson V., Rossmann R. and Van Hoof P. (2004) Spatial and temporal trends in surface water and sediment contamination in the Laurentian Great Lakes. *Environmental Pollution*, **129**, 131–144.

McGarrigle M.L., Bowman J.J., Clabby K., Lucey J.J., Cunningham P., MacCárthaigh M., Keegan M., Cantrell B., Lehane M., Clenaghan C., *et al.* (2002) *Water Quality in Ireland 1998–2000*, Environmental Protection Agency: Wexford. [available at **http://www.epa.ie/Water/**]

Meybeck M., Kimstach V. and Helmer R. (1996) Strategies for water quality assessment. In Chapman D. (Ed.), *Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring*, *Second Edition*, E&FN Spon: London, pp. 23–57.

Myers D.N. and Wilde F.D. (Eds.) (2003) *Biological Indicators: U.S. Geological Survey Techniques of Water-Resources Investigations*, *Third Edition*, book 9, Chap. A7, accessed 19th June 2004, from **http://pubs.water.usgs.gov/twri9A/**

NRC (National Research Council) (2004) *Assessing the National Streamflow Information Program*, Committee on Review of the USGS National Streamflow Information Program, National Academies Press: Washington, p. 176.

NWQMC (2004a) National Water Quality Monitoring Council, National Environmental Methods Index (NEMI). **http://wi.water.usgs.gov/methods/about/publications/nemi_fs2.pdf** [21 December 2004]

NWQMC (2004b) National Water Quality Monitoring Council. **http://water.usgs.gov/wicp/acwi/monitoring/index.html** [21 December 2004]

NWQMC (2004c) National Water Quality Monitoring Council, Accreditation of Laboratory and Field Activities for Water-Quality Monitoring. **http://wi.water.usgs.gov/methods/about/publications/accred_fs.pdf** [21 December 2004]

Robertson D.M. and Roerish E.D. (1999) The influence of various water quality sampling strategies on load estimates for small streams. *Water Resources Research*, **35**(12), 3747–3759.

Simon T.P. (2000) The use of biological criteria as a tool for water resource management. *Environmental Science and Policy*, **3**, S43–S49.

Spooner C.S. and Mallard G.E. (2003) Identifying monitoring objectives. *Water Resources Impact*, **5**(5), 11–13. **http://water.usgs.gov/wicp/acwi/monitoring/pubs/0309impact**.pdf [21 December 2004]

Thomas R. and Meybeck M. (1996) The use of particulate material. In *Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring*, Chapman D. (Ed.), *Second Edition*, E&FN Spon: London, pp. 127–174.

Umbuzeiro G.A., Roubicek D.A., Sanchez P.S. and Sato M.Z. (2001) The Salmonella mutagenicity assay in a surface water quality monitoring program based on a 20-year survey. *Mutation Research*, **491**, 119–126.

US EPA (1997) *Volunteer Stream Monitoring: A Methods Manual*, EPA 841-B-97-003, United States Environmental Protection Agency. Office of Water: Washington. **http://www.epa.gov/volunteer/** [26 October 2004]

US EPA (1999a) *Biological Criteria: National Program Guidance for Surface Waters Fact Sheet*, United States Environmental Protection Agency. Office of Water: Washington.

US EPA (2004) *Volunteer Lake Monitoring*, EPA 440-4-91-002, United States Environmental Protection Agency. Office of Water: Washington. **http://www.epa.gov/volunteer/** [26 October 2004]

USGS (2003) *Results of the U.S. Geological Survey's Analytical Evaluation Program for Standard Reference Samples Distributed in March 2003*, U.S. Geological Survey Open-File Report 03–261, U.S. Geological Survey, Washington, p. 114.

USGS (2004a) *Water-Quality Information – Field Procedures*. **http://water.usgs.gov/owq/Fieldprocedures.html** [8 October 2004]

USGS (2004b) *A New Evaluation of the USGS Streamguaging Network*, A Report to Congress **http://water.usgs.gov/streamgaging/** [8 October 2004]

USGS (2004c) *National Water Quality Laboratory, Laboratory Performance Evaluation Studies*. **http://nwql.usgs.gov/Public/Performance/publiclabpe.html** [8 October 2004]

USGS (2004d) *National Water Quality Laboratory, Laboratory Audits, External Audits*. **http://nwql.usgs.gov/Public/Performance/publiclabaudit.html** [8 October 2004]

USGS (2004e) *Water-Quality Data for the Nation*. **http://waterdata.usgs.gov/nwis/qw** [26 October 2004]

Van Metre P.C., Mahler B.J. and Callender E. (1998) Trends in Organochlorine and Radionuclide concentrations in the upper Rio Grande based on sediment core analyses from Elephant Butte Reservoir, New Mexico. *International Journal of Sediment Research*, **13**(4), 1–11.

Ward R.C., Loftis J.C. and McBride G.B. (1986) The data-rich but information-poor syndrome in water-quality monitoring. *Environmental Management*, **10**(3), 291–297.

Washington H.G. (1984) Diversity, biotic and similarity indices. A review with special relevance to aquatic systems. *Water Research*, **18**(6), 653–659.

WHO (2004) *Guidelines for Drinking-Water Quality. Volume 1. Recommendations*, *Third Edition*, World Health Organization: Geneva, p. 515.

Yoder C.O. and Rankin E.T. (1995) Biological criteria development and implementation in Ohio. In *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*, Davis W.S. and Simon T.P. (Eds.), Lewis Press: Boca Raton, pp. 109–144.

93: Effects of Human Activities on Water Quality

NORMAN E PETERS[1], MICHEL MEYBECK[2] AND DEBORAH V CHAPMAN[3]

[1]U.S. Geological Survey, Georgia Water Science Center, Atlanta, GA, US
[2]Sisyphe/CNRS, University of Paris, Paris, France
[3]Environmental Research Institute and Department of Zoology, Ecology and Plant Science, University College Cork, Cork, Ireland

Water quality comprises the physical, chemical, and biological characteristics of a water body. The water body acquires these characteristics from a suite of complex interactions among the water, atmosphere, soils, and lithology. Human activities affect both water quality and quantity. Human activities change land use and land cover, which changes the water balance and usually changes the relative importance of processes that control water quality. Furthermore, most human activities generate waste ranging from gases to concentrated radioactive wastes. Although each issue can be subdivided into a myriad of individual processes or activities, the primary water-quality issues affected by human activities include organic material, trace elements (heavy metals), acidic atmospheric deposition and runoff, salinization, nutrients (primarily nitrogen and phosphorus), pathogenic agents including bacterial pathogens, enteric viruses, and protozoans, suspended sediment, oil and grease, synthetic organic compounds, thermal pollution, exotic and invasive species, pesticides and herbicides, and radioactivity. In addition to the various issues, each human activity has a potential cyclical and cascading effect on water quality and quantity along hydrologic pathways. The degradation of water quality in one part of a watershed can have negative effects on users downstream; the timescale of effects is determined by the residence time of that substance along various hydrological pathways. An extremely important factor is that substances added to the atmosphere, land, and water generally have relatively long timescales for removal or cleanup. The nature of the substance, including its affinity for adhering to soil and its ability to be transformed, affect the mobility and the timescale for the removal of the substance and its effects on water quality, for example, biota.

INTRODUCTION

Human activities alter the natural characteristics of air, land, and water, which subsequently affect water quality. Prior to and immediately after World War II, the main water contamination problems in developed countries were fecal and organic pollution from untreated human waste and the by-products of early industries, particularly in urban areas. Through improved waste treatment and disposal during the past few decades, developed countries have addressed this problem resulting in water-quality improvements. Most of this has been accomplished through pollution laws and pollution-control technologies, particularly with respect to point sources such as factories and sewage treatment plants. Prior to wastes being treated, the management philosophy was essentially that *dilution is the solution to pollution*, which still persists in many areas of the world.

Human activities can affect water quality directly and indirectly. Direct effects are those that change water quality through the addition of some chemical constituent, physical characteristic, or biological component. The discharge of wastewater to a stream directly affects the stream chemistry, the application and leaching of fertilizer affects groundwater chemistry that can affect surface water by groundwater discharge, and the combustion of fossil-fuel, including coal, oil, and wood plus forest fires (biomass burning), affects air quality. The air quality affects precipitation, chemistry, and the water quality of the receiving water bodies. Increasing the temperature of a stream or lake by discharging heated water from the cooling tower of a power plant is an example of a direct change in physical characteristic.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Biological systems are directly affected by exotic and invasive species, for example, the introduction of nonindigenous aquatic species (animals or plants), such as zebra, mussels, or Brazilian waterweed, to streams and lakes. Indirect effects include alteration of the landscape (construction, mining, and farming), which affects hydrological pathways that change the rates at which water interacts with the environment and flushes out materials from that landscape.

An extreme example is the landscape alteration accompanying urbanization where pavement and building construction produce impervious surfaces that cause rainfall or snowmelt to rapidly runoff through concrete drains and pavement-lined stream channels. The runoff washes off chemicals and biota, such as bacteria and feces, which accumulated prior to the hydrologic event. In the natural state or prior to making the surfaces impervious, more of the rainfall or snowmelt infiltrates the soil and replenishes soil moisture, which had been transported to the atmosphere by evapotranspiration through grasses, shrubs and trees, and recharges groundwater. The groundwater travels along a much longer hydrological pathway to the stream than that occurs when water runs off an impervious surface.

Water-quality degradation is one of the most persistent, and in most cases, visible signs of human effects on the natural environment. As we continue to learn more about the complex interactions among the physical, chemical, and biological components of aquatic ecosystems, and as we have begun to clean up some of the oldest and worst contaminant problems, we find that our efforts are often offset by a misunderstanding or an incomplete knowledge of the processes and a myriad of new contaminants that have unknown effects on aquatic ecosystems and may degrade our potable water. Many of the dominant natural factors and the primary human activities affecting water quality are shown in Figure 1. The primary water-quality issues, their causes and effects, are listed in Table 1. The World Resources Institute (WRI)

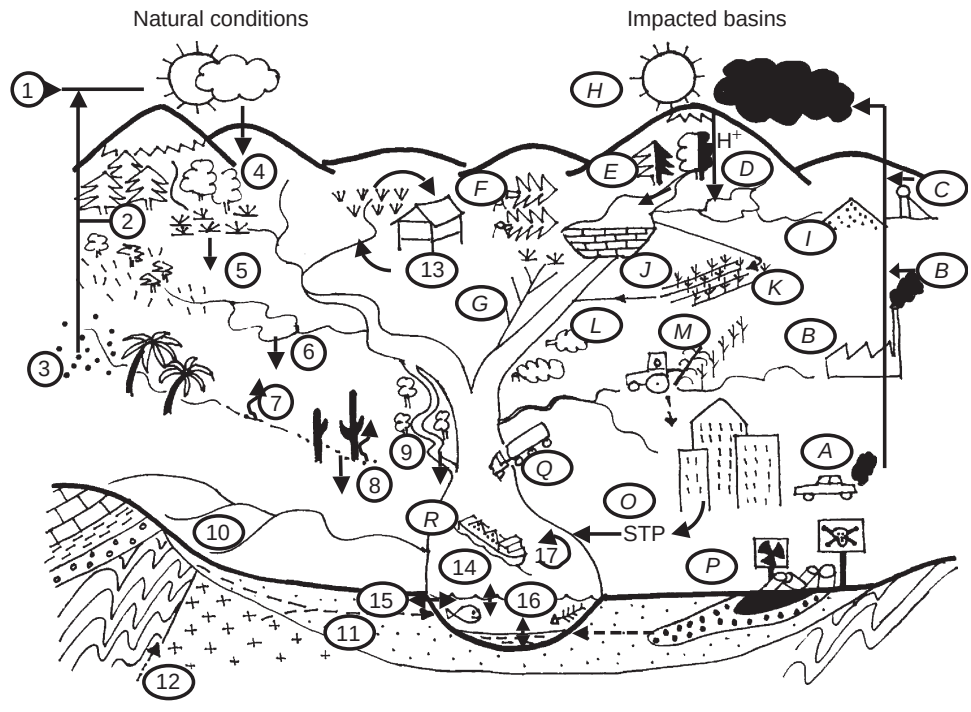


Figure 1 Sources, pathways, and main processes regulating water quality in natural conditions and impacted basins. Atmospheric fallout (4) originating from oceanic inputs (1), vegetation emissions (2) and eolian erosion (3); retention and transformation in wetlands (5) and lakes (6); evaporation leading to salinization (7) and precipitation in soils and endorheic basins (8); retention and exchange with floodplain (9); chemical weathering and mechanical erosion of various rock types (10); direct inputs of groundwater (11); hypothermal inputs (12); closed N, P cycles in traditional agriculture (13); exchange between surface waters and atmosphere (14), groundwater (15) and sediments (16); internal cycling of carbon and nutrients in aquatic food webs (17). Contaminated and/or acidified atmospheric deposition (D) due to urban and traffic, (A), industrial sources and mining/smelting sources (C); forest die back (E) and forest cutting (F); wetland draining (G); climate change (H); mining (I) and industrial (B) wastewaters; river fragmentation through damming (J); enhanced evaporation after irrigation (K); use of fertilizers and pesticides in agriculture (K, M); river course channelization and floodplain isolation (L); release of treated and untreated urban sewage (O); leak of dangerous chemicals from waste dumps (P); accidental spills (Q) and chronic leaks (R); enhanced eutrophication (17)

Table 1 Common contaminants and their sources and effects on water quality

Contaminant	Primary source(s)	Source/contaminant examples	Some primary effects
Organic material	Domestic sewage, industrial wastewater, and agricultural activities	Urban outfalls of untreated or treated sewage, pulp and paper plant discharges, farmyard and feedlot runoff	Decomposition of the organic material uses oxygen from the water, which stresses or in the extreme kills aquatic biota
Trace elements (heavy metals)	Industry and mining	Improper disposal or leakage of silver salts from film processing, atmospheric deposition of fly ash from fossil-fuel combustion, leakage, or flooding of tallings ponds,	Persist in vegetation, sediments, and wetlands, and accumulate in biota. Toxic to humans and aquatic organisms.
Acidic atmospheric deposition and runoff	Fossil-fuel combustion and mining runoff	Automobiles and power plants, and coal mine areas	Acidification of streams and lakes, which stresses or is toxic to aquatic organisms, and mobilizes metals
Salinization	Leaching and precipitation of salts from soils by irrigation, movement of saltwater from over pumping groundwater in coastal areas or reduced streamflow from overuse of freshwater upstream	Groundwater pumping in coastal areas, crop or horticulture irrigation in arid or semiarid areas	Salt accumulation poisons the soils so that plants will not grow, intrusion of saltwater into freshwater causes the freshwater to become nonpotable
Nutrients (N and P)	Runoff from farm land, septic tank leakage and sewage disposal, atmospheric deposition of N	Runoff of inorganic fertilizer or manure from cropland and animal feedlots, sewage effluent discharges to streams, migration of nutrients in septic tank leachate through groundwater to streams or lakes	Nutrients stimulate growth of aquatic plants causing eutrophication (algal blooms), change in species composition, decrease in water transparency, the excess plant material decomposes causing oxygen depletion, and so on. High nitrate concentrations in drinking water can cause methemoglobinemia (blue baby syndrome)
Pathogenic agents including bacterial pathogens, enteric viruses, protozoans	Sewage and urban stormwater mobilizing human and animal feces	Bacteria – *E. coli, Salmonella, Shigella* Viruses – Hepatitis A, Rotavirus Protozoa – *Giardia, Cryptosporidium*	Infectious diseases are spread through contaminated drinking water leading to diarrhea and intestinal diseases and parasites and in the extreme causing mortality

(*continued overleaf*)

Table 1 (*continued*)

Contaminant	Primary source(s)	Source/contaminant examples	Some primary effects
Suspended sediment (+ and −)	Soil erosion, land clearing for construction or agriculture in watersheds, damming	Erosion and transport from cultivated crop land, housing construction, runoff from impervious areas causing high streamflow and bank collapse, sediment starved rivers due to damming, reservoir sedimentation	Sediment entrainment affects habitat of aquatic organisms, sediment deposition fills lakes and fouls water intakes and filters, sediment-bound contaminants can affect human and aquatic ecosystem health, removal of sediment in reservoirs increases anoxia and generates greenhouse gases and also starves rivers of sediment for maintaining habitat and fertilizing floodplains
Oil and grease	Stormwater runoff, leakage from fuel storage and distribution facilities, industrial and urban waste	Leakage from barges or tankers, physical clogging of drains causes sewage to overflow	Affects respiratory system of aquatic organisms; affects feathers, preventing birds from flying; sewage overflow can be toxic to humans and aquatic organisms when ingested
Synthetic organic compounds	Untreated and treated sewage, septic tank leachate, urban storm runoff, runoff from animal pastures and feedlots	Prescription and nonprescription drugs (including antibiotics, antibacterial compounds), caffeine and derivatives, plasticizers, fire retardants, solvents, plant, and animal steroids, PAH, Fuels, and so on.	Some pharmaceuticals are known to disrupt the endocrine cycle of humans and aquatic organisms, affecting reproduction; effects vary by compound
Thermal pollution	Power generation, industrial cooling	Discharge of heated water in power plant and industrial (e.g. brewing) cooling towers to streams or lakes	Affects metabolic rates of biota and can change species composition of aquatic organisms. Affects oxygen levels and decomposition rates of organic material.
Exotic and invasive species	Transportation of water and other earth resources	Discharge of ballast water from ships, use of invested crops or other resources	Change in food webs and associated biological communities. Changes in biogeochemical cycling
Pesticides and herbicides	Agricultural use, urban use	Seasonal applications to cropland and runoff during the spring, runoff from suburban and urban areas, herbicide use along highways, domestic sewage	Toxic to aquatic biota
Radioactivity	Nuclear industry, military	Atmospheric fallout from bomb testing, leaks from nuclear power plants, migration, or leaks from waste facilities	Carcinogenic to humans and biota (terrestrial and aquatic)

provides a wide range of summary information regarding human activities on a global scale, and several of their maps have been included in this article to highlight the geographical distribution of various characteristics (WRI, 2004). Many of the topics in Table 1 will be discussed later and also topics of other articles of the encyclopedia of hydrological sciences, for example, acidification, **Chapter 95, Acidic Deposition: Sources and Effects, Volume 3**; point and nonpoint source pollution, **Chapter 94, Point and NonPoint Source Pollution, Volume 3**; nutrient cycling, **Chapter 96, Nutrient Cycling, Volume 3**; urban water quality, **Chapter 97, Urban Water Quality, Volume 3**; pathogens, **Chapter 98, Pathogens, Volume 3**; and salinization, **Chapter 99, Salinization, Volume 3**. The relative global impacts of these issues on specific water bodies (streams, lakes, reservoirs, and groundwater) are listed in Table 2. Also, water quality can affect human health, which is beyond the scope of this article (For more details of water-quality effects on human health, see Barzilay *et al.*, 1999).

When considering the impact of direct and indirect human effects on water quality, it is important to note that the effects are cyclical and cascading along hydrologic pathways in a watershed context. Hydrologic pathways are routes along which water moves from the time it is received as precipitation (e.g. rain and snow) until it is delivered to the most downstream point in a watershed, and the drainage area is defined by the downstream point to which flow converges. The degradation of water quality in upstream parts of a watershed can have negative effects on downstream users, and because there is generally a continuum of human activities and users throughout a watershed, the degradation effects cascade through the watershed. Cyclical effects include the artificial movement of water upstream, such as groundwater abstraction for irrigated agriculture. Each pathway has a residence time associated with transport and mixing. Consequently, everyone lives downstream of the effects of some human activity, and in the extreme, contaminants can originate outside a watershed and be deposited into it from the atmosphere. Also, substances added to the atmosphere, land, and water generally have relatively long timescales for removal or cleanup.

LANDSCAPE ALTERATION

Humans have altered the landscape in most areas of the world to satisfy basic needs in providing water, food, shelter, fuel for cooking and heating, and related services. Humans have also surpassed needs and have been currently very aggressive in providing for wants, which appear to be limited only by technology and imagination. The most intense landscape alteration has occurred in lowlands in fertile floodplains, where agricultural production is the highest, typically in small or localized areas for human settlement and mining.

As human population has become more concentrated or dense, the demand for resources to supply needs and wants has increased and so has the production efficiency. Whereas hunting and gathering were the main methods of food supply to many relatively sparse indigenous populations historically, higher population densities can only be supported by farming, which includes alteration of the landscape for crop production and pasturing for livestock. In addition, building construction requires basic earth resources ranging from grasses and wood to soils and rock.

Table 2 The relative water-quality issues of different water bodies, estimated at the global scale. The issues and related effects vary widely at the regional and local scales and the degradation generally reflects the economic development and land use. For a more complete discussion about the sources and effects of each issue, see Chapter 1 of Chapman (1996)

	Water body		
Issue	Rivers	Lakes/reservoirs	Groundwater
Pathogens	xxx	x[a]	x
Suspended solids	xx	x	na
Organic material[b]	xxx	x	x
Eutrophication[c]	x	xxx	na
Nitrate	x	o	xxx
Salinization	x	x	xxx
Trace elements	xx	xx	xx[d]
Synthetic organic compounds	xxx	xx	xxx[d]
Acidification	x	xx	o
Modification of hydrological regimes[e]	xx	x	x

xxx, severe or global deterioration
xx, moderate deterioration
x, occasional or regional deterioration
o, rare deterioration
na, not applicable

[a]mostly in small shallow water bodies.
[b]other than resulting from aquatic primary production.
[c]algae and macrophytes.
[d]from landfills and mine tailings.
[e]water diversion, damming, overpumping, and so on.

Farming, lumbering, and mining supply these resources. Transportation has expanded from simple road building for the movement of humans and animals to trains, automobiles, and airplanes having a much higher demand for a complex array of earth resources including manufacturing materials and fuel (e.g. petroleum products). Mining provides geologic materials, which are altered by various processes to forms that can be used for manufacturing. In some cases, the geologic materials are used directly for construction such as sand and gravel, or a specific element or compound is extracted from the ore deposit leaving most of the geologic material behind as waste. Also, water supply and the use of water for navigation, manufacturing, and power generation requires alteration of river channels and the building of dams.

Deforestation

Deforestation, or the removal of a forest, typically occurs in humid climates, such as tropical or temperate zones, where rainfall is abundant and vegetation can easily grow (Figure 2). The impact of deforestation can have both indirect and direct effects on water quality.

The density of leaves and needles that can be maintained on a plant is a function of environmental conditions, mainly the amount of available nutrients, water, and solar radiation (sun energy), and the relative number of competing plants in the area. The leaves and needles, as well as branches and whole plants, which die and fall on the ground, are broken down by chemical reactions and insects, and in the process, release nutrients to the soil. This biogeochemical cycle also yields carbon dioxide, which when combined with water produces a mildly acidic solution, carbonic acid. The carbonic acid in soil waters and groundwater attacks rocks and minerals, breaking them down into more fundamental and smaller particles and releasing solutes.

Figure 2 Clearcut logging in New Zealand. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Typically, larger plants (trees compared to shrubs or grasses) transpire more water per unit area. Consequently, deforestation (tree or forest removal) has a large effect on water partitioning by decreasing transpiration and decreasing the permeability of the surface soils caused by an extensive root structure, and subsequently, increasing the runoff and water yield. The trees also reduce the erosion impact of rain drops by changing the pathway of the rain through the canopy along branches and down stems as stemflow and through the leaves/needles and branches either to shrubs or soil and litter as throughfall. The trees lower groundwater levels, and after deforestation, groundwater levels rise and waterlogging occurs, particularly, in and around lowland areas.

Another effect of deforestation is changing the thermal regime of the landscape and rivers. Forests provide shade, which is critical for the development of a particular aquatic biota. For example, fish, such as trout, are sensitive to increasing water temperatures and related decreases in dissolved oxygen concentrations resulting from tree removal from riparian zones.

The direct water-quality effect is a rapid turnover of nutrients and increased erosion. Erosion mobilizes both mineral soil and organic material, which can affect aquatic food webs by destroying the habitat and changing the balance of nutrients in the streams. The deforestation of riparian areas also reduces the food supply for macroinvertebrates and fish, thus affecting the biological quality. Higher water tables within the basin can also have a large effect on soil characteristics and stream chemistry, particularly in areas where the groundwater chemistry is enriched in salts because of weathering. For example, the removal of forests in eastern Australia, such as in the Murray–Darling basin, caused the water tables to rise to the surface of the soils. The highly saline groundwater contaminated the soils and the streams.

Deforestation has been extensive during the past 8000 years (Figure 3). Tropical forests, such as the Amazon basin, are currently being rapidly logged, burnt, or otherwise degraded, losing from 11 000 to 19 000 km^2 annually during the 1990s (Houghton *et al.*, 2000). For more detailed information about land use and land-cover effects on runoff from deforestation, see **Chapter 119, Land Use and Landcover Effects on Runoff Processes: Forest Harvesting and Road Construction, Volume 3**.

Agriculture

Food production is determined by the soil fertility, which was historically augmented with the application of organic matter and manure (Figure 4), but is now being increasingly augmented by a myriad of inorganic fertilizers to increase plant growth rates, and chemicals for the control

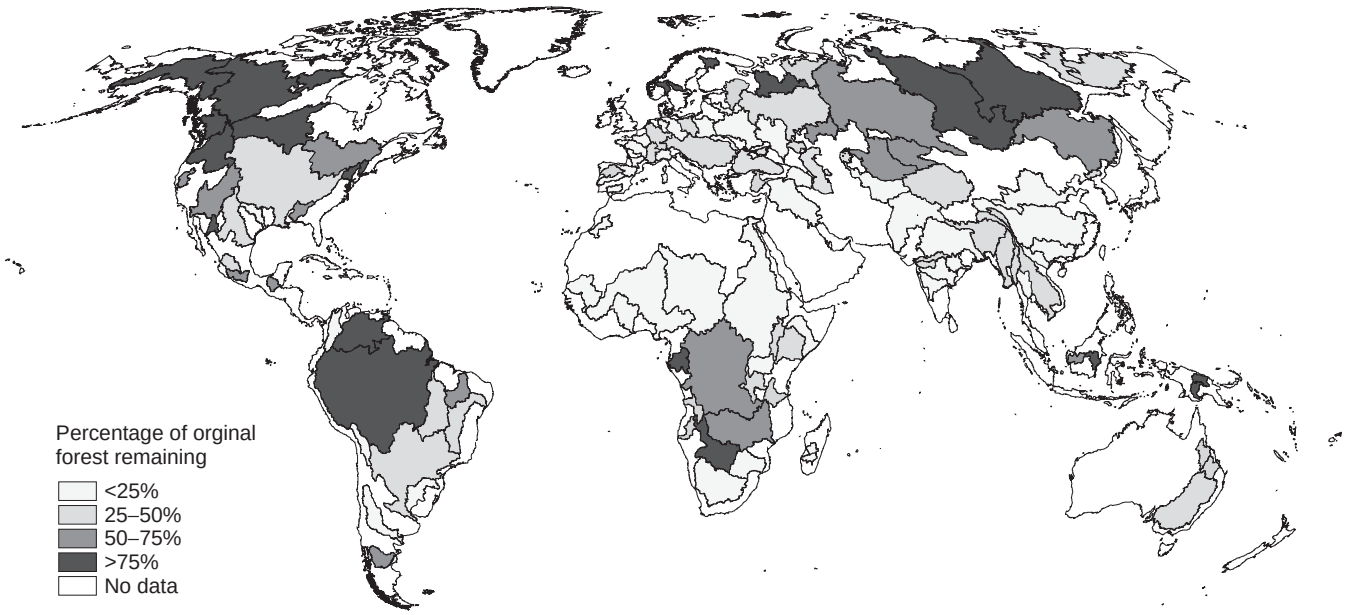


Figure 3 Remaining original forest cover by basin (Taken from **http://earthtrends.wri.org**/ ©2003 WRI by permission of World Resources Institute; Billington *et al.*, 1996; WRI *et al.*, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 4 Manure application (Courtesy of Kevin D. Richards, US Geological Survey). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

(a)

(b)

Figure 5 (a) Aerial pesticide application (Courtesy of W.D. Mullins, US Geological Survey); (b) Tractor application of pesticide (Courtesy of C.G. Crawford, US Geological Survey). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of weeds (herbicides) and insects (insecticides or pesticides) (Figure 5). Lowland areas, including floodplains and riparian zones, in general are the areas targeted for agricultural production. In many cases, the high water table of lowland areas requires drainage and in the extreme, wetland removal. The act of tilling the land and planting crops or maintaining pastures originally altered the natural vegetation, and consequently, caused a change in water partitioning. Plants use soil water and groundwater. The soil water and groundwater is typically drawn through the roots and then transpired as water vapor through the leaves or needles by the process called *evapotranspiration*, which is driven by sunlight and photosynthesis. The water transported through the plant also carries nutrients and releases carbon dioxide to the atmosphere. In the process, the plants develop roots, which change the porosity and infiltration rate of the soil, and these roots extract nutrients from the soil, some of which are products of the chemical breakdown of rocks and minerals or chemical weathering.

Figure 6 Cattle grazing with direct access to a stream in a pasture near Hamilton, New Zealand. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Raising stock likewise has some major effects on water quality. Direct impacts result from overgrazing and allowing livestock direct access to surface water (Figure 6). The treatment and the disposal of animal wastes and the increasing use of pharmaceuticals in animal production alters the chemical composition of receiving waters. Direct water-quality impacts also result from poor storage and treatment of manure slurry from concentrated animal production and feedlots. For more detailed information about land use and land-cover effects on runoff from agriculture, see **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**.

The removal of native vegetation for farming increases the erosion potential of the land, and the transport of sediment is largely determined by the intensity and the duration of major hydrological events, which provides the energy for suspending or otherwise mobilizing solids. Technology and developments in agricultural management practices have changed markedly and these affect the potential mobility of contaminants. The development of larger farm vehicles and machinery require larger fields, which reduce the number of hedgerows and encourage the entrainment of sediment and other contaminants. Ploughing a field along the contours rather than up and down the slope reduces erosion, but the timing of the ploughing is critical with respect to major hydrological events. Likewise, the timing and the rate of application of manure or inorganic fertilizers and other agrochemicals as detailed below has a major effect on groundwater and surface water quality.

Agrochemicals

The demand for food has increased proportionally with the increase in human population. Technology has made some large contributions to food production efficiency, but not without negative effects on water quality. The major agrochemicals affecting water quality are fertilizers, pesticides, and herbicides. Global population increased from 2.5 billion people in 1950 to more than 6 billion people in 2004. The related food production increases and the resulting effects on water quality have been spatially disproportional because of the geographic distribution of the population and hydroclimatic conditions in the high population areas.

Historically, some of the modern-day farming practices such as application of manure and liming, and crop rotations with legumes were used. Ancient civilizations, such as the Egyptians (~3000 B.C.), recognized the importance of flood deposits in maintaining soil productivity. The Chinese have used organic manures to fertilize crops for the past 3000 years. The Greeks recorded the use of city sewage on vegetable crops and olive groves and a canal system for delivering the sewage to fields about 5000 years ago. Furthermore, the Greeks recorded the effects of green manure crops, particularly legumes, on crops. Both the Greeks and the Romans recorded the use of marl and plant ashes for soil enrichment and the acidity of lowland soils was adjusted using ashes or lime. Mineral fertilizers were not entirely unknown in past civilizations, as the Greeks and Romans fertilized plants with saltpeter, potassium nitrate. However, the extensive use of inorganic fertilizers did not begin until the twentieth century.

Two main factors contributed to the accelerated use of inorganic fertilizers beginning in the early 1900s. The first factor was the rapid commercialization and progressively cheaper synthetic production of ammonium, called *the Haber–Bosch process*. The second factor was the discovery of a couple of large sedimentary rock deposits of hydroxyapatite, a calcium phosphate mineral, which when treated with sulfuric acid yielded a highly soluble form called *superphospate*. The accelerated use of inorganic fertilizers began following World War II in most developing countries, for example, in Europe and North America, and a couple of decades later in many developing countries (Figure 7). Crop production (t ha^{-1}) followed, but was disproportionate to the increased use of inorganic fertilizers, increasing only by a factor of 3 from 1950 to 2000, while inorganic nitrogen fertilizer application increased by a factor of 23 and phosphorus application increased by a factor of 8.

The timing of the agrochemical application; the rate at which it is applied; the chemical, physical, and hydrologic characteristics of the watershed; and the hydrologic conditions at the time of application affect the rate at which the chemical can leach into groundwater and/or runoff into surface water. Needless to say, the more agrochemical added, the higher the probability that leaching and runoff of the agrochemical will occur. The continued use and the

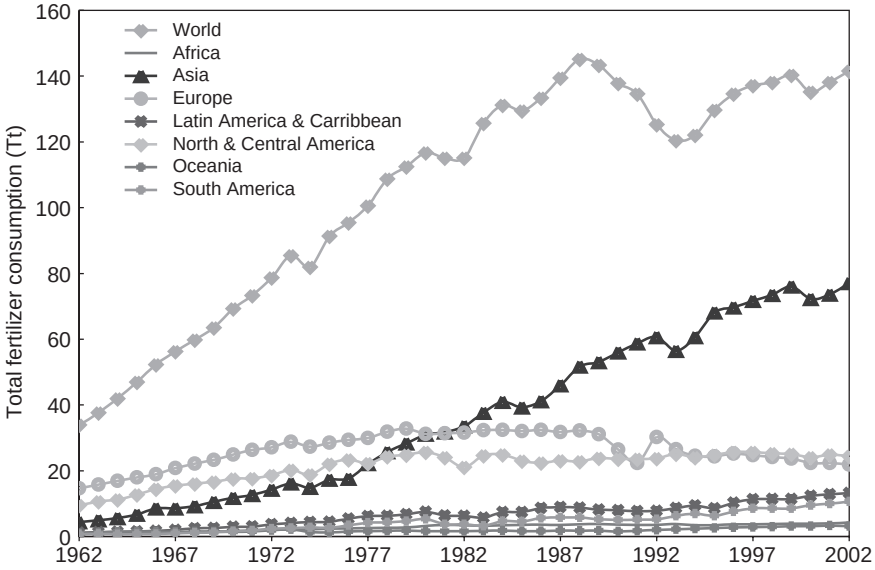


Figure 7 Annual fertilizer consumption of the World from 1962 to 2002 (FAOSTAT data, 2005). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

high application rates of fertilizer have resulted in a high probability of nitrate contamination of shallow groundwater (Burkart and Stoner, 2001). Furthermore, direct runoff and discharge of contaminated groundwater in agricultural areas having high fertilizer application rates increases the eutrophic status of receiving waters. Inorganic and organic nutrient concentrations increase, causing aquatic plant growth to increase. In the extreme, the occurrence of algal blooms has become more frequent and increased plant growth can result in hypoxia (oxygen depletion) because of plant decay. The high use of inorganic nitrogen fertilizers in the midwestern United States and subsequent transport in the Mississippi River system is considered responsible for the persistence of hypoxia in an increasingly larger area of the Gulf of Mexico (Burkart and James, 1999). Hypoxia is detrimental to the survival of bottom dwelling organisms, such as shellfish and bottom feeding fish.

In general, surface water phosphorous concentrations have decreased significantly downstream of urban areas since the 1980s, primarily because of the improvements in wastewater treatment and the decreasing use of phosphorous detergents, which affects urban wastewater (point source) and not agricultural runoff (diffuse or nonpoint source). However, phosphorous concentrations are still much higher than the background and remain a problem in most regions of the developed countries, for example, Europe and North America. The number of sewage treatment plants has also increased, which has reduced the amount of ammonium, which is toxic to fish. However, the sewage treatment process converts ammonium to nitrate, which is released into waterways and affects the nutrient balance. Dissolved nitrate in drinking water can also harm human health. For a more detailed discussion about nutrient cycling, see **Chapter 94, Point and NonPoint Source Pollution, Volume 3**.

Pesticides and herbicides are also leached or mobilized with fine sediment and have different water-quality effects than the more mobile inorganic fertilizers. Many of the first pesticides and herbicides were extremely toxic and persistent in the environment. For example, lindane, DDT, and dioxin, which were applied several decades ago, are still measurable today. The symbiotic and cumulative environmental effects of these compounds, even at low concentrations, are still not known.

Irrigated Lands

In contrast to the effects of deforestation for establishing agriculture in humid areas, establishing agriculture in semiarid and arid zones requires large additions of water to sustain the vegetation (Figure 8). Diverted surface water or abstracted groundwater is used for irrigation in semiarid and arid zones and each affects water quality. As discussed in the section on deforestation, a large percentage of available water is transpired to the atmosphere by vegetation. The low supply and high demand on water in arid and semiarid areas accentuate water-quality degradation issues; the poorer the quality the less water is available from an already depleted resource. Poorly designed and implemented irrigation schemes can cause waterlogging, salinization, and alkalization of soils.

Salinization results from concentrating solutes to excess in soils and surface water by relatively increasing the

Figure 8 Irrigation (Courtesy of T.R. Maret, US Geological Survey). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

loss of water due to evaporation and evapotranspiration with respect to supply. The result is that water added by irrigation is evaporated at a relatively high rate, leaving behind brine, which ultimately creates a soluble residue in the soil upon complete evaporation. The residue poisons the soil and prevents vegetation from growing even if sufficient water was available. The solution to potential salinization for fertile agricultural areas is to routinely flush the soil, which dissolves the salts and transports them in drainage water to nearby streams and lakes. Naturally, unless flushing continues throughout the entire system, the salts will continue to accumulate in a downstream water body. In some circumstances, the salinization is seasonal and a combination of snowmelt or seasonally high rainfall may result in an effective flushing of the salt from an entire watershed. For a more complete discussion about salinization processes and effects on water quality, see **Chapter 99, Salinization, Volume 3**.

Mining

Archaeological discoveries have shown that mining, the process of extracting useful rocks and minerals from the surface of the Earth, was conducted by humans as early as 40 000 years ago from a mine shaft in Switzerland (Gregory, 2001). The mining of metals, such as gold, silver, copper, and lead date to approximately 4000 B.C. and quarried stone was used for pyramid construction in Egypt in 2600 B.C. (Gregory, 2001). The disaggregation of the lithosphere varies by the mining technique, which is determined by the type of rock or mineral being mined. Some common mining practices include, (i) open pit (or strip) mining, in which the soil and rock (overburden) above a seam of coal, mineral, or rock is removed and the exposed material is extracted (Figure 9); (ii) hydraulic mining, in which a

Figure 9 Open pit or strip mine for building stone near Pirenopólis, Brazil. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

powerful jet of water is used to dislodge minerals present in unconsolidated ground or coal from a seam; (iii) placer mining, in which the mineral is separated by dredging and washing and typically occurs in alluvial deposits below or adjacent to streams; (iv) hardrock mining, in which picks and shovels, rock drills, and explosives create shafts and tunnels following the ore body or vein; and (v) solution mining, in which the mineral is dissolved and mobilized using a solvent, typically hot water. In each case, the earth materials are disturbed, affecting hydrologic pathways and exposing new mineral surfaces to the air or water. Each of these can change the natural characteristics of weathering and release contaminants. The valuable rock or mineral is surrounded by gangue (uneconomic material) that needs to be separated in a concentrating process and the processed gangue, called *tailings*, is typically discarded in piles and ponds. Sulfide minerals are commonly found in the rocks adjacent to coal deposits, and when they are exposed to the air and water, they oxidize and hydrolyze producing sulfuric acid. The sulfuric acid can acidify streams and lakes producing a pH in the range from 2 to 3 that is sufficiently acidic to mobilize metals, which are deleterious to aquatic biota. Also, the breaching of tailings ponds during flood conditions poses a major environmental threat by releasing high concentrations of contaminants and sediment to streams and lakes.

Urbanization

Land-use change affects freshwater quality (Falkenmark *et al.*, 1999). The geographic distribution of cities, roads, mining, agricultural land, and natural areas within a watershed influences infiltration properties, transpiration rates, and runoff patterns, affecting water quantity and quality.

Figure 10 Imperviousness produced by a mixture of houses, commercial buildings, parking lots, sidewalks, and roads in a residential area of Auckland, New Zealand. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Increasing imperviousness of an area increases the volume and discharge (flow rate) of runoff of receiving streams and impacts the water quality and biodiversity of freshwater systems. Urbanized areas have more impervious areas than nonurbanized areas and also have more (or at least concentrated) sewage and industrial pollution (Figure 10).

Construction, such as is prevalent for highways and buildings, generally results in the alteration of the landscape and the movement of earth materials, which may generate high sediment concentrations and loads in rivers. Construction alters the partitioning of water as discussed previously when forests are cleared, but roofs, parking lots, and roads stop water from infiltrating the soil and convey the runoff into ditches, drainage channels, and retention basins. The net result of the increased imperviousness is a rapid washoff of contaminants that accumulate on the surfaces and an increased velocity and volume of runoff, which increases bank erosion along the waterway. Materials are excavated and relocated, altering the existing hydraulic characteristics.

The potential deleterious human activity effects on water quality are amongst the highest for urban areas. Also, the most highly urbanized watersheds generally are in coastal areas (Figure 11). Of the 6.2 billion people in the world in 2002, approximately 47% live in urban areas, which are also the areas undergoing the most rapid change in population (WRI, 1996). It is projected that 75% of the global human population will live in urban areas by 2030 and that three of four urban dwellers will live in a megacity, that is, having a population of more than 10 million people (WRI, 1996). For more detailed information about land use and land-cover effects on runoff from urbanization and suburban development, see **Chapter 117, Land Use and Land Cover Effects on**

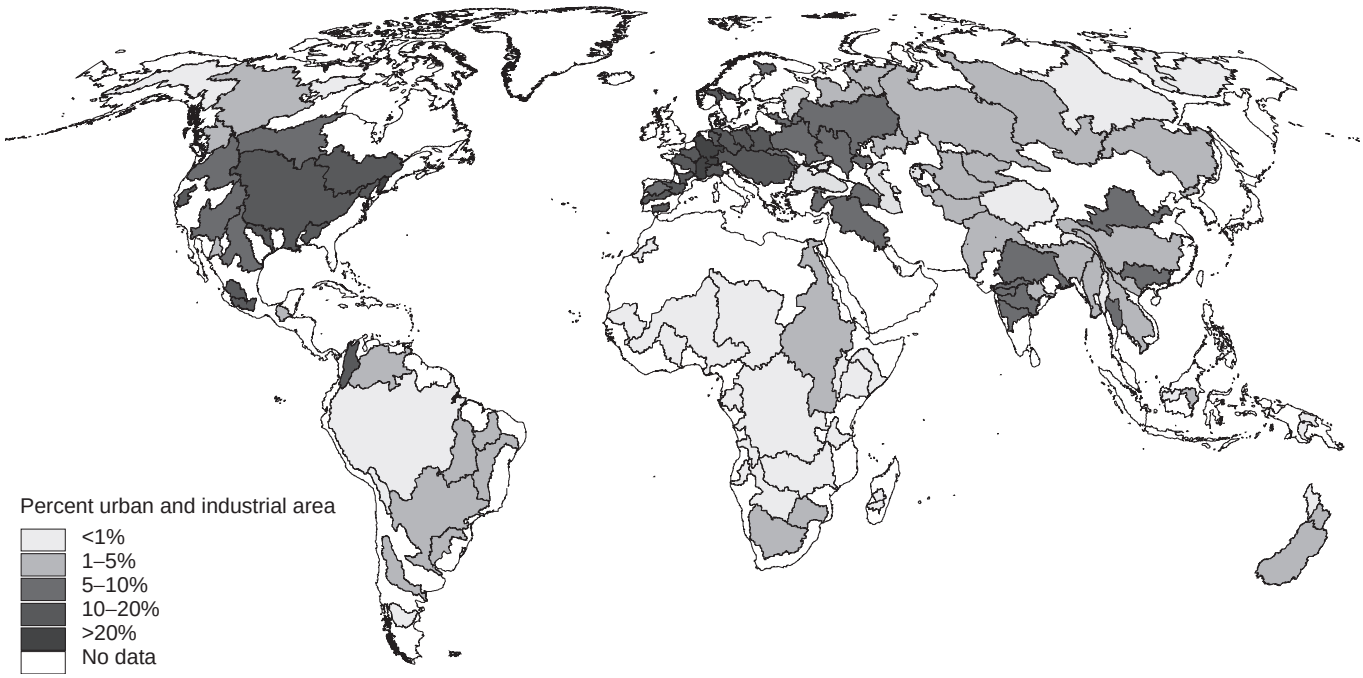


Figure 11 Global distribution of urban and industrial areas by basin (Taken from **http://earthtrends.wri.org**/ ©2003 WRI by permission of World Resources Institute; NOAA, 1998; WRI *et al.*, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Runoff Processes: Urban and Suburban Development, Volume 3.

Dams and River Channel Alteration

Dams have been built for thousands of years. They have been built to manage floods, to provide drinking water and water for industry and farming, and to provide electricity. In addition, the natural channels of many major rivers have been managed (channelized) primarily by dredging to maintain deep water for ship, boat, and barge traffic, and by altering the bank material to reduce erosion, which generally has been effective for low to moderate flows, but not during major floods. During 2000, approximately 50% of the world's rivers had at least one large dam, and the flow in more than 37% of the major river basins has been altered (WRI, 2004; Figure 12). About 50% of the world's large dams were built primarily to provide irrigation, which accounts for about 40% of irrigated farmlands worldwide. During 2000, large dams generated about 20% of electricity; 33% of countries rely on hydropower to supply more than 50% of their electricity.

Dams and their management change the natural flow of rivers, affecting the timing and the amount of supply of water and sediment downstream. Under natural conditions, the ecology of riparian zones, naturally very fertile areas, was determined by flood flows and associated deposition of sediment during overbank discharges. Major damage due to flooding has decreased considerably in most major river basins of the world. Dams have served the public well by controlling river flow for the protection of life and property. Downstream water supply from dams, particularly large dams, typically is derived from reservoir bottom water, which has distinctly different thermal and chemical characteristics from the background river affecting nutrient cycling. Consequently, these characteristics have a deleterious effect on natural aquatic biological communities downstream and the terrestrial biota, which are linked to the aquatic biota through food chains. Dams cause a change in biological communities and a reduction in biodiversity. Furthermore, dams influence carbon cycling. Reservoirs are sites for carbon burial, that is, in the sediments, and the anoxic conditions in the sediments and bottom waters of the reservoir produce methane from decomposing biomass. Emissions of methane and carbon dioxide from tropical reservoirs have been shown to be of similar order of magnitude as fossil-fuel greenhouse gas emissions. For more detailed information about floodplain sedimentation and reservoir sedimentation, see **Chapter 84, Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands, Volume 2** and **Chapter 88, Reservoir Sedimentation, Volume 2**, respectively.

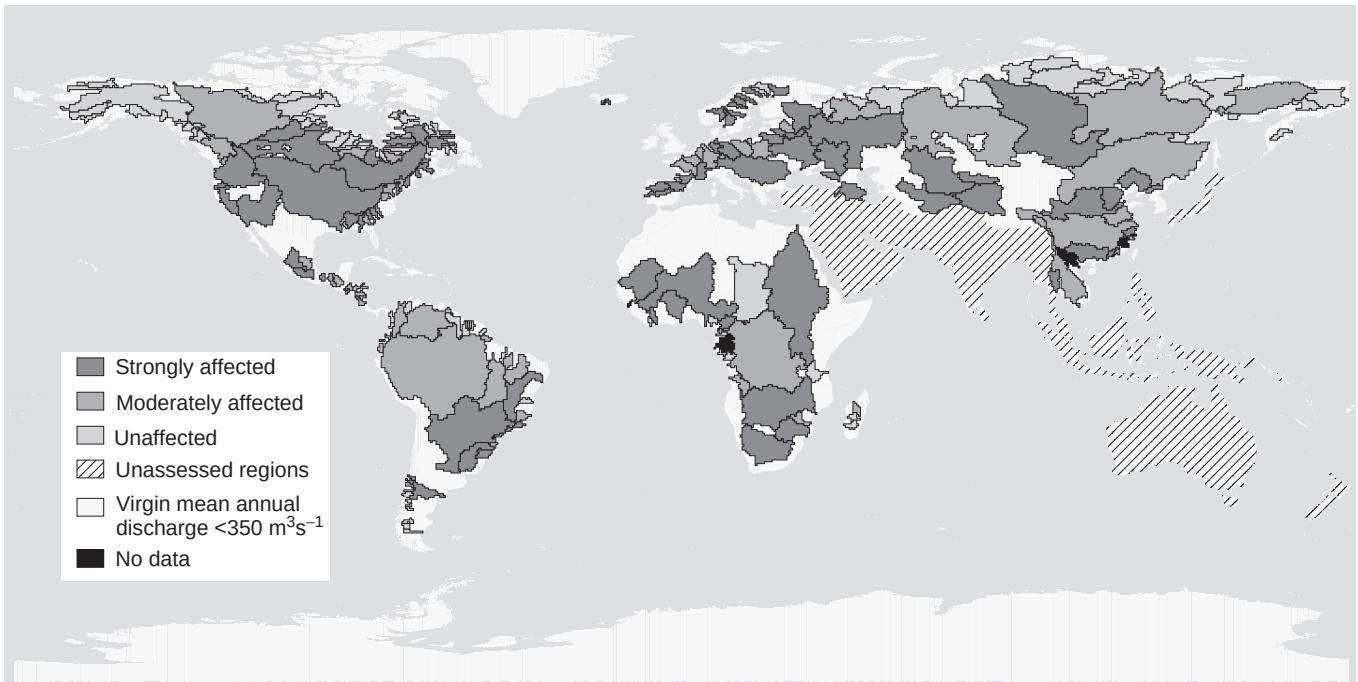


Figure 12 Degree of river fragmentation and flow regulation by basin (Taken from **http://earthtrends.wri.org**/ ©2003 WRI by permission of World Resources Institute; Dynesius and Nilsson, 1994; Revenga *et al.*, 2000). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

OTHER DIRECT EFFECTS ON WATER QUALITY

Water pollution spans a wide range of chemical, physical, and microbial factors, but over the years, the balance of major pollutants has shifted markedly in most industrialized countries. Unfortunately, a new suite of contaminants from intensive agriculture and development activities in watersheds has kept the cleanup from being complete. In general, national water cleanup programs have been effective in reducing point sources of contamination and controlling the dumping of wastes from pipes. However, cleanup programs have *not* been effective in reducing "nonpoint" pollutants such as nutrients, sediments, and toxic compounds that come in runoff from agriculture, urban and suburban stormwater, mining, and oil and gas operations. In developing countries, sewage is still being discharged directly to rivers, streams, and coastal waters without any waste treatment (WRI, 1996). In developing countries, the more traditional pollution problems, such as sewage, in combination with new contaminants, such as pesticides and insecticides, heavily degrade water quality, particularly near urban industrial centers and intensive agricultural areas.

Surface-water bodies, such as streams, lakes, and reservoirs, are extremely susceptible to direct discharges of liquid and solid waste. Very dilute waters, typically in headwater areas, are susceptible to impacts from atmospheric deposition (acid rain). However, surface-water bodies are not the only ones that suffer from pollution. Groundwater, which is a critical source of drinking water and irrigation water for people in semiarid and arid regions, is also affected. The major causes of groundwater pollution are leaching of pollutants from agriculture, industry, and untreated sewage, and saltwater intrusion caused by over pumping.

Once the pollutants enter an aquifer, the environmental damage can be severe and long lasting, partly because of the long time needed to flush pollutants from the aquifer (Table 3). Because groundwater is primarily used for drinking water, pollution from untreated sewage, intensive agriculture, solid waste disposal, and industry can cause serious human health problems. People in rural areas primarily rely on shallow groundwater supply for drinking water, and shallow groundwater is more susceptible than deeper groundwater to contamination because of improper human waste disposal and agrichemical application. Although information to assess the status of groundwater resources is largely lacking globally, groundwater contamination from fertilizers, pesticides, industrial effluents, sewage, and hydrocarbons is occurring in many parts of the world.

As noted for surface water, nitrate pollution is one of the most serious threats to groundwater. In general, the risk of agrochemical pollution of groundwater supplies is directly related to the amount of agrochemicals applied to the land, the characteristics of the individual chemical, and the permeability of the soil. Furthermore, although untreated wastewater (point source) was a primary source of nitrate in surface waters 40 years ago, the increased use of inorganic fertilizers in agriculture, coupled with improved urban and suburban wastewater treatment, has resulted in a shift to diffuse or nonpoint sources, and especially agriculture, as the most important source of nitrate in surface waters and groundwater (Table 4). The trends of nitrate concentrations in shallow groundwater are similar to that of fertilizers used (Figure 7). For example, 50% of groundwater samples in a heavily fertilized region of northern China contain nitrate levels above the safe limit for drinking water ($10\,mg\,l^{-1}$). In the United States, where groundwater supplies drinking water for more than 50% of the population, high nitrate concentrations are widespread in shallow groundwater aquifers in agricultural areas and groundwater nitrate pollution in Europe is similarly widespread (Burkart and Stoner, 2001).

Although the map in Figure 3 is of the areal extent of deforestation, it largely represents the distribution of watersheds containing intensive agriculture. Watersheds with intensive agricultural development are likely to experience water-quality degradation from pesticide and nutrient runoff and increased sediment loads. The intensive agricultural land is concentrated in Europe, India, eastern China, southeastern Asia, and the midwestern United States, with smaller concentrations in Argentina, Australia, and Central America. Africa is striking in its lack of intensively cropped land, with the exception of small patches along the Mediterranean coast and in South Africa.

Surface-water quality has improved in most developed countries since about 1985, but nitrate and pesticide contamination remain persistent problems. Data on water quality in other regions of the world are sparse, but water quality appears to be degraded in almost all regions with intensive agriculture and rapid urbanization. Unfortunately, little information is available to evaluate the extent to which chemical contamination has impaired the health of freshwater ecosystems. However, incidents of algal blooms and eutrophication (the enrichment of water by nutrients, especially compounds of nitrogen and phosphorus, which will accelerate the growth of algae and higher forms of plant life) are widespread in freshwater systems all over the world – an indicator that these systems are affected by water pollution (Figure 13). In addition, the massive loss of wetlands at a global level has greatly impaired the capacity of freshwater systems to filter and purify water. Groundwater quality suffers from many of the same pollution problems as surface waters and faces the additional challenge of being very difficult to restore once the underlying aquifer is contaminated because the residence times are much longer and variable than for surface

Table 3 Relations among water-quality issues, causes, controlling factors, and spatial and temporal scales of degradation and recovery (Peters and Meybeck, 2000)

			Timescale		Major controlling factors	
Major causes/issues	Major related issues[a]	Space scale	Contamination[b]	Cleanup[c]	Biophysical	Human
Population	Pathogens	Local	<1 year	<1 year		Density & treatment
	Eutrophication*	Regional	<1–10 year	1–100 year		Treatment
	Micropollutants	Regional	<1–10 year	1–100 year		Miscellaneous
Water management[d]	Eutrophication*	Regional	<1 year	10–>100 year		Flow
	Salinization	Regional	10–100 year	10–>100 year		Water balance
	Parasites	Regional	1–10 year	>100 year		Hydrology
Land management	Pesticides	Local-regional	<1 year	1–100 year		Agrochemicals
	Nutrients (NO_3^-)	Local-regional	10–100 year	>10 year		Fertilizer
	Suspended solids*	Local-regional	<1–10 year	10–100 year		Construction/clearing
	Physical changes	Local	<1–10 year	>100 year		Cultivation
Atmospheric transport	Acidification*	Regional	>10 year	10 year		Cities, melting & fossil-fuel emissions
	Micropollutants	Regional	>10 year	1–100 year		Industry
	Radionuclides	Regional–global	<1 year	≫100 year		
Concentrated pollutant sources:						
Megacities	Pathogens	Local	<1 year			Population & treatment
	Micropollutants	Local-regional	<1 year			Types of mines
Mines	Salinization	Local-regional	10–100 year			
	Metals	Local-regional	<1 year			Waste management
Nuclear industry	Radionuclides	Local-Global	<1 year			
Global climate change	Salinization	Global	>10 year	>100 year	Temperature and precipitation	Fossil-fuel emissions and Greenhouse gases
Natural ecological conditions	Parasites*	Regional	Permanent	Permanent	Climate and hydrology	
Natural geochemical conditions	Salts	Regional	Permanent	Permanent	Climate and lithology	
	Fluoride**	Local-regional			Lithology	
	Arsenic**, metals**	Local-regional				

[a] *is relevant primarily to surface water and ** is relevant primarily to groundwater.
[b] Space scales: local – <10 000 km^2; regional – 10^5 to 10^6 km^2; and global – 10^7 to 10^8 km^2.
[c] Lag between cause and effect.
[d] Longest timescale is for groundwater, followed by lakes, and shortest for rivers and streams.

Table 4 Annual nutrient discharge (nitrogen and phosphorus) to surface waters in the United States, in 10^3 Mt (modified from Carpenter *et al.*, 1998)

Source	Nitrogen	Phosphorus
Nonpoint sources:		
Croplands	3204	615
Pastures	292	95
Rangelands	778	242
Forests	1035	495
Rural lands	659	170
Miscellaneous	695	68
Total nonpoint discharges	6663	1658
Total point sources	1495	330
Total (point + nonpoint)	8158	2015
Nonpoint, % of total	82%	84%

Figure 13 Eutrophication (Courtesy of I.R. Waite, US Geological Survey). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

waters and the hydrological pathways are generally not known.

Waste Generation, Treatment, and Disposal – Air, Land, and Water

Wastes are generated from most human activities and their disposal depends on the form and the treatment, which is often a function of the population density. In rural areas, sewage treatment may not be a viable option and solid and liquid wastes are disposed of on the land or through septic systems. In urban areas, wastewater systems are sewered. Sewage can be classified into three types: domestic sewage carries used water from houses and apartments and is called *sanitary sewage*; industrial sewage is used water from manufacturing or chemical processes; and storm sewage, or storm water, is runoff from precipitation that is collected in a system of pipes or open channels. Sewage directly affects water quality. Sanitary sewage contains a variety of dissolved and suspended impurities, including organic material and disease-causing microbes. Industrial sewage typically contains specific and identifiable chemical compounds. Storm sewage carries dissolved and suspended materials, which are washed from the land surface. In addition to the solid and liquid wastes, contaminants can be dispersed to the atmosphere either because of combustion or the exposure of waste to the atmosphere and the volatility of the individual compound. For example, biomass burning, which occurs on a large scale in the Amazon basin, not only affects the composition of the atmosphere (Crutzen and Andreae, 1990), but the large emissions (gases, aerosols, and particles) also affect climate (Levine *et al.*, 1995).

Human, agricultural, commercial, manufacturing, and industrial wastes occur in solid, liquid, and gaseous forms and each can affect water quality. The characteristics of the waste and the method used for disposal will have varying effects on contaminant transport and transformation, and consequently, on the effect that a given contaminant in the waste will have on water quality. For example, the combustion of fossil fuels (coal and oil) produces gaseous and aerosols that have resulted in major human and environmental health problems through the exposure to heavy metals on particles, for example, fly ash, and the downwind transport and transformation of sulfur dioxide and nitrous oxides resulting in an acidic atmospheric deposition. The high nitrogen content in pollution can fertilize terrestrial vegetations such as the forests in the Sierra Nevada, California downwind of Los Angeles. In contrast, high pollutant emissions and atmospheric deposition have been attributed to forest decline and dieback in some areas, such as the "Black Triangle" consisting of southwestern Poland, eastern Germany, and northern Czech Republic (Figure 14). The high smog levels in cities, which have deleterious effects on the human respiratory system, also results from the emissions of noxious gaseous, which when washed out of the atmosphere in precipitation, can deplete nutrients from the soils and degrade the quality of surface and groundwater. Historical changes in waste disposal or treatment practices have largely resulted from scientific evidence of cause and effect relations. These changes also have affected the transformation and transport of contaminants.

For acid rain, although smog effects on human health were suspected as early as the late 1800s, the deleterious effects on lakes and rivers were only realized when fisheries and other aquatic biota were affected, and the scientific investigations suggested a causal linkage among pH, alkalinity (the capacity of a water to neutralize acidity), and fish decline or deformation. The particles emitted

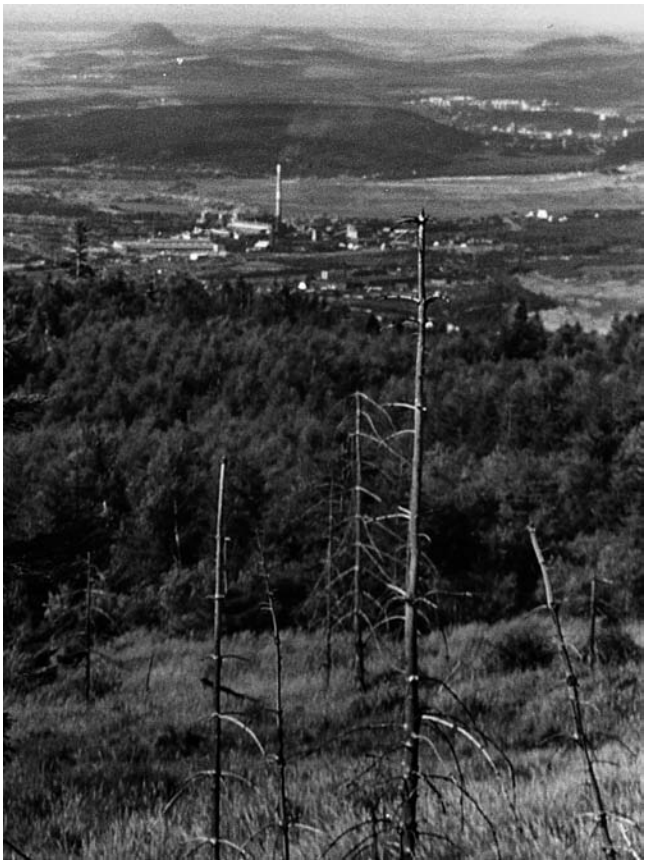


Figure 14 Dead trees in the foreground and power plant with plume in the background, Krusne Hóry, Czech Republic. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

during the combustion of a fossil fuel, particularly coal, are extremely basic, and when combined with rainfall that contains sulfuric and nitric acid produced from the gases emitted by the fossil-fuel combustion, neutralize the acid rain. However, it was first recognized that the particles needed to be removed because of human health concerns related to the heavy metal content. Scrubbers were installed on the smoke stacks of power-generation stations to remove the particulates, and that resulted in concentrated gaseous emissions, which produced very acidic rainfall near the stack. The next solution was to build taller smoke stacks to put the emissions higher into the atmosphere where they could be diluted and transported further downwind. This improved the air and the precipitation quality near the smoke stack, but transported the contaminants to the land and water farther downwind of the stack, evolving from a local to a regional problem. The land and water eventually showed signs of stress from accumulated loadings. Finally, additional technological solutions were applied to remove the acid-producing gases from the stacks, and in the case of coal, to wash or otherwise remove the acid-gas–producing compounds from the coal before combustion. For a more comprehensive discussion about freshwater acidification, see **Chapter 95, Acidic Deposition: Sources and Effects, Volume 3**.

Miscellaneous

There is a growing awareness that increased water used by humans does not only reduce the amount of water available for industrial and agricultural development, but has a deleterious effect on aquatic ecosystems and their dependent species. Human activities have severely affected the condition of freshwater ecosystems, to a point where many freshwater species are facing rapid population declines or extinction. The introduction of exotic or invasive species, such as the zebra mussel, water hyacinth, and hydrilla in the United States, also is stressing ecosystems, because many of these species have no predators to keep their reproduction and growth in check. Consequently, the invasive or exotic species out-compete the indigenous biota and have caused major changes in biological communities and related food webs (Mooney and Hobbs, 2000).

In addition, biota and humans, in particular, are susceptible to various pathogenic agents in freshwater, including enteric viruses, protozoa, bacteria, and parasites. From 1% to 25% of a given warm-blooded animal group is *infected* and given survival times of a few days to several years. The potential for contamination of surface water by a pathogenic agent is high depending on the safeguards and wastewater and solid waste treatment processes used (Geldreich, 1999). Receiving waters of urban centers and highly concentrated animal operations are particularly susceptible to pathogens.

REFERENCES

Barzilay J.I., Weinberg W.G. and Eley J.W. (1999) *The Water We Drink: Water Quality And Its Effects On Health*, Rutgers University Press: New Brunswick, p. 180.

Billington C., Kapos V., Edwards M., Blyth S. and Iremonger S. (1996) *Estimated original forest cover map – a first attempt and the world forest map*. United Nations Environment Program, World Conservation Monitoring Center: Cambridge, UK.

Burkart M.R. and James D.E. (1999) Agricultural-nitrogen contributions to hypoxia in the Gulf of Mexico. *Journal of Environmental Quality*, **28**(3), 850–859.

Burkart M.R. and Stoner J.D. (2001) Nitrogen in groundwater associated with agricultural systems. In *Nitrogen in the Environment*, Follett R.F. and Hatfield J.L. (Eds.), Elsevier: pp. 123–145.

Carpenter S.R., Caraco N.F., Correll D.L., Howarth R.W., Sharpley A.N. and Smith V.H. (1998) Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, **8**(3), 559–568.

Chapman D.V. (Ed.) (1996) *Water Quality Assessments: A Guide to the Use of Biota Sediments and Water in Environmental Monitoring*, *Second Edition*, Chapman and Hall: London, p. 626.

Crutzen P. and Andreae M. (1990) Biomass burning in the tropics: impact on atmospheric chemistry and biogeochemical cycles. *Science*, **250**, 1669–1677.

Dynesius M. and Nilsson C. (1994) Fragmentation and flow regulation of river systems in the northern third of the World. *Science*, **266**, 753–762.

Falkenmark M., Andersson L., Castensson R. and Sundblad K. (Eds.) (1999) *Water, A Reflection of Land Use – Options for Counteracting Land and Water Mismanagement*, NFR, Swedish Natural Science Research Council: Stockholm, p. 128.

FAOSTAT data (2005) FAOSTAT Database Query. **http://apps.fao.org/faostat/form?collection=Fertilizers&Domain=Means&servlet=1&hasbulk=0&version=ext&language=EN** [last accessed January 2005]

Geldreich E.E. (1999) Pathogenic agents in freshwater resources. *Hydrological Processes*, **10**(2), 315–333.

Gregory C.E. (2001) *A Concise History of Mining*, A.A. Balkema: Exton, p. 197.

Houghton R.A., Skole D.L., Nobre C.A., Hackler J.L., Lawrence K.T. and Chomentowski W.H. (2000) Annual fluxes of carbon from deforestation and regrowth in the Brazilian Amazon. *Nature*, **403**, 301–304.

Levine J.S., Cofer W.R., Cahoon D.R. and Winstead E.L. (1995) Biomass burning: a driver for global change. *Environmental Science and Technology*, **29**(3), 120A–125A.

Mooney H.A. and Hobbs R.J. (2000) *Invasive Species in a Changing World*, Island Press.

National Oceanic and Atmospheric Administration (NOAA) (1998) *Stable Lights and Radiance Calibrated Lights of the World CD-Rom: Nighttime Lights of the World*. NOAA-National Geophysical Data Center: Boulder, Colorado, US. Available online at: **http://julius.ngdc.noaa.gov:8080/production/html/BIOMASS/night.html**.

Peters N.E. and Meybeck M. (2000) Water-quality degradation effects on freshwater availability: impacts of human activities. *Water International*, **25**(2), 185–193.

Revenga C., Brunner J., Henninger N., Kassem K. and Payne R. (2000) *Pilot Analysis of Global Ecosystems: Freshwater Systems*, World Resources Institute: Washington.

WRI (2004) World Resources Institute. **http://www.wri.org/** [last accessed January 2005]

WRI, IUCN, IWMI, and Ramsar Convention Bureau (2003) *Watersheds of the World CD-Rom*, WRI: Washington, IUCN-The World Conservation Union: Gland and Cambridge.

WRI, UNEP, UNDP, & World Bank (1996) *World Resources 1996-97*, Oxford University Press: New York, p. 365.

94: Point and NonPoint Source Pollution

KEITH LOAGUE[1] AND DENNIS L CORWIN[2]

[1]Department of Geological and Environmental Sciences, Stanford University, Stanford, CA, US
[2]USDA-ARS, George E. Brown, Jr. Salinity Laboratory, Riverside, CA, US

The information age has ushered in a global awareness of complex environmental problems that do not respect political or physical boundaries: climatic change, ozone layer depletion, deforestation, desertification, and pollution from point and nonpoint sources. Among these global environmental problems, point and nonpoint source pollution represent a perfect example of a complex multidisciplinary problem that exists over multiple scales with tremendous spatial and temporal complexity. A point source of pollution discharges to the environment from an identifiable location, whereas a nonpoint source of pollution enters the environment from a widespread area. The ability to accurately assess present and future point and nonpoint source pollution impacts on ecosystems ranging from local to global scales provides a powerful tool for environmental stewardship and guiding future human activities.

INTRODUCTION

The objective of this chapter is to introduce the subject of point and nonpoint source pollution, the entirety of which could easily fill several volumes. The point and nonpoint source problems are defined and important related legislation is identified in the following two subsections. Monitoring and modeling (with consideration for uncertainty) are discussed, relative to characterizing the impacts from point and nonpoint source pollution, in the section on assessment. Excerpts from three case studies are presented in the example section; two of the case studies are for nonpoint source pollution, the third case study is for point source pollution. It should be pointed out that the chapter is focused, with the exception of one of the case studies, on the United States.

Point versus Nonpoint Source Pollution

Point source pollutants, in contrast to nonpoint source pollutants, are associated, as the name suggests, with a point location such as toxic-waste spill site (see Figure 1). As such, point source pollutants are, compared to nonpoint source pollutants, characteristically (i) easier to control, (ii) more readily identifiable and measurable, and (iii) generally more toxic. Nonpoint sources of pollution (see Figure 1) are the consequence of agricultural activities (e.g. irrigation and drainage, applications of pesticides and fertilizers, runoff and erosion); urban and industrial runoff; erosion associated with construction; mining and forest harvesting activities; pesticide and fertilizer applications for parks, lawns, roadways, and golf courses; road salt runoff; atmospheric deposition; livestock waste; and hydrologic modification (e.g. dams, diversions, channelization, over pumping of groundwater, siltation). Point sources include hazardous spills, underground storage tanks, storage piles of chemicals, mine-waste ponds, deep-well waste disposal, industrial or municipal waste outfalls, runoff, and leachate from municipal and hazardous waste dumpsites, and septic tanks. Compared to point source pollution, nonpoint source pollution is more difficult, related to monitoring and enforcement of mitigating controls, due to the heterogeneity of soil and water systems at large scales. Characteristically, nonpoint source pollutants (i) are difficult or impossible to trace to a source, (ii) enter the environment over an extensive area and sporadic timeframe, (iii) are related (at least in part) to certain uncontrollable meteorological events and existing geographic/geomorphologic conditions, (iv) have the potential for maintaining a relatively long active presence on the global ecosystem, and (v) may result in long-term, chronic (and endocrine) effects on human health and soil-aquatic degradation.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

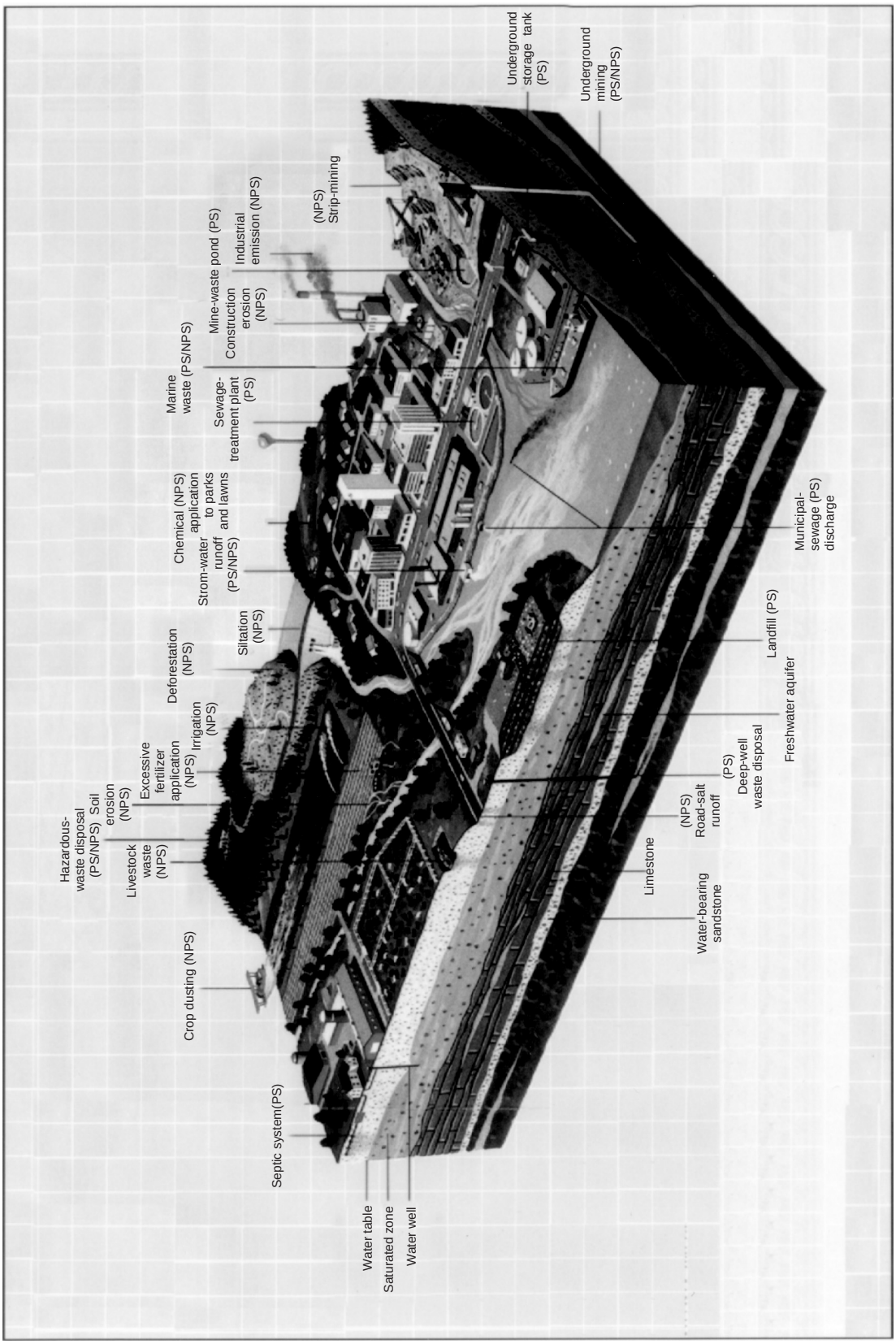


Figure 1 Examples of point and nonpoint source pollution (Reproduced from Corwin *et al.* (1999) by permission of American Geophysical Union. Adapted from a map published by National Geographic Society, 1993). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Nonpoint source pollution was generally not recognized until the mid-1960s. Initially, nonpoint source pollution was associated entirely with pollution from storm water and runoff. Subsequently, nonpoint source pollution has expanded to encompass all forms of diffuse pollutants. Nonpoint pollutants are defined as "contaminants of [air, and] surface and subsurface soil and water resources that are diffuse in nature and cannot be traced to a point location" (Corwin and Wagenet, 1996). It is important to note the statutory definitions of point and nonpoint source pollution (see Novotny and Olem, 1994). In essence, point sources of pollution "were originally defined as pollutants that enter the transport routes at discrete identifiable locations and that can usually be measured", while nonpoint source pollution was "everything else". Table 1 provides a list of statutory point and nonpoint sources of pollution compiled by Novotny and Olem (1994).

There can be a fine line between point and nonpoint source pollution. The distinction depends entirely upon the scale of interest. One person's point source pollution can easily be another person's nonpoint source pollution. For example, a single quarter-section (65 ha), in a watershed of hundreds of thousands of hectares, might be considered a point source of nitrogen fertilizer, and, yet, within the area of interest, the fertilizer could be viewed as a nonpoint source because of the broadcasting of millions of granules of the active ingredient. The scale of reference ultimately determines whether a pollutant is viewed as coming from a point or nonpoint source.

Historically, point source pollutants have received the greatest attention, both publicly and scientifically, because of the conspicuous severity of their impacts at a localized point (e.g. Love Canal (Mercer *et al.*, 1983) and Woburn (Harr, 1995)). However, over recent years, public, political, and scientific attention has shifted more and more toward pollutants that are widespread. This shift reflects an awareness of the scope and potential impact of the nonpoint source pollution problem (see Corwin and Loague, 1996; Corwin *et al.*, 1999).

Table 1 Statutory point and nonpoint sources of pollution (after Novotny and Olem, 1994)

Statutory Point Sources
• Municipal and industrial wastewater effluents
• Runoff and leachate from solid waste disposal sites
• Runoff and infiltrated water from concentrated animal feeding operations
• Runoff from industrial sites not connected to storm sewers
• Storm sewer outfalls in urban centers with populations of more than 100 000
• Combined sewer overflows
• Leachate from solid waste disposal sites
• Runoff and drainage water from active mines, both surface and underground, and from oil fields
• Other sources, such as discharges from vessels, damaged storage tanks, and storage piles of chemicals
• Runoff from construction sites that are larger than 2 ha
Statutory Nonpoint Sources
• Return flow from irrigated agriculture
• Other agricultural and silvicultural runoff and infiltration from sources other than confined concentrated animal operations
• Unconfined pastures of animals and runoff from range land
• Urban runoff from sewered communities with a population of less than 100 000 not causing a significant water quality problem
• Urban runoff from unsewered areas
• Runoff from small and/or scattered (less than 2 ha) construction sites
• Septic tank surfacing in areas of failing septic tank systems and leaching of septic tanks effluents
• Wet and dry atmospheric deposition over a water surface (including acid rainfall)
• Flow from abandoned mines (surface and underground), including inactive roads, tailing, and spoil piles
• Activities on land than generate wastes and contaminants, such as:
• Deforestation and logging
• Wetland drainage and conversion
• Channeling of streams, building of levees, dams causeways, and flow-diversion facilities on navigable waters
• Construction and development of land
• Interurban transportation
• Military training, maneuvers, and exercises
• Mass outdoor recreation

A Bit of the Legal History, for the United States

Up until the 1800s, the rural environment remained largely pristine in contrast to the filth of the urban areas as exemplified by historic centers such as ancient Rome, and, in the Middle Ages, London and Paris. By the mid-1800s, the sewage and runoff of the urban areas that polluted the surface waters became associated with waterborne disease. The concern for public health led to the first awareness of the repercussions of the environmental degradation and the close association of humankind's well being to environmental quality, thereby heralding the first environmental activist period. By the twentieth century, sewage systems and sewage treatment plants came into being.

Public interest in the environment through the first half of the twentieth century was almost negligible primarily because the epidemics associated with the Middle Ages had been controlled or eliminated, and because the use of chemical insecticides and fertilizers did not occur to an appreciable extent until the late 1950s. Once human-made chemicals, such as DDT (1,1,1-trichloro-2,2-bis-(4-chloropheny)-ethane)), were introduced in the mid-1950s

and the dumping of outflows from post-World War II factories continued unabated, pollution of soil and water resources rapidly increased. The book *Silent Spring* by Rachel Carson in 1962, which revealed the spread and potential danger of toxic human-made chemicals to the unwitting public, is heralded as initiating the second environmental activist period.

Freeze (2000) astutely points out that environmental policy and perspective over the last half century have been driven by the prevailing socio-politico-economic climate. The environmental pendulum, as Freeze refers to it, has swung from underkill to overkill and from concern that breeds action to disillusionment that breeds reaction. Table 2 provides a timetable of some of the significant environmental and statutory events that have shaped the last six decades and an overview of the pendulum swing from the environmental perspective to a retrospective from the New Right.

The positive result of the environmental activism of the 1960s and the 1970s was the formation of the Environmental Protection Agency (EPA) in 1970 and the Clean Water Act (CWA) of 1972 (Public Law 92–500). It should be pointed out, however, that the CWA called for "zero discharge" into surface waters, which, in actuality, shifted the onus of pollution from surface water to soil and groundwater. Prior to the CWA, the Refuse Act of 1899 and the Water Pollution Control Act of 1948 (Public Law 80–845) were used to control water pollution from industrial sources, but these archaic statutes were generally ineffective in controlling pollution.

The CWA was intended to be a comprehensive water quality program. For example, under the CWA, each State was to establish water quality standards based upon a total maximum daily load (TMDL), which is the amount of a pollutant that a body of water can handle from all sources. Once a TMDL is established, it is used as a basis for putting limits on the amount of that pollutant that can be discharged into the system. In reality, most of the emphasis of the CWA was placed on controlling point source rather than nonpoint source pollution (Novotny and Olem, 1994). The CWA established three crucial tasks (Novotny and Olem, 1994): (i) the regulation of point source discharge, (ii) the regulation of oil spills and other hazardous substances, and (iii) financial assistance for wastewater treatment plant construction. The most significant contribution of the CWA was the formulation of a means of enforcement through the National Pollution Discharge Elimination System (NPDES). This system has served as the basic mechanism for enforcing the implementation of pollution abatement of point sources and some nonpoint sources legally classified as point sources (Novotny and Olem, 1994).

The Water Quality Act (Clean Water Act) of 1987 (Public Law 100–24) provided three sections (Section 319, 402, and 404) that are significant to the control of nonpoint source pollution. Of these, Section 319 most directly applies to nonpoint pollution, requiring each State to prepare a State Nonpoint Source Assessment Report, and provides the incentive of matching Federal funds to encourage States to develop and implement management programs.

Table 2 Timetable of environmental and statutory events (after Freeze, 2000)

Decade	Environmental perspective	Environmental events	US Federal response	Prevalent environmental concepts	Influential books	New right retrospective
1940s	Chemical Revolution				*A Sand County Almanac* Leopold (1949)	Chemical Revolution
1950s	Age of Carelessness			Throwaway society		Age of Economic Prosperity
1960s	Age of Awakening		EPA (established, 1970)	Conservation, wilderness protection	*Silent Spring* Carson (1962)	Age of Social Upheaval
1970s	Age of Awareness and Action	Earth Day Love Canal	CWA, SDWA, TSCA, RCRA, PP list, NOR	Limits to growth	*The Closing Circle* Commoner (1971)	Age of Overreaction
1980s	Age of Disillusion	Three mi Island	CERCLA, HSWA, SARA	Soft energy paths		Age of Vindication
1990s	Age of Reaction	Earth Summit	PPA in Rio	Sustainable yield	*Earth in the Balance* Gore (1992)	Age of Reason

CERCLA = Comprehensive Environmental Response, Compensation, and Liability Act; CWA = Clean Water Act; EPA = Environmental Protection Agency; HSWA = Hazardous and Solid Waste Amendments (to RCRA); NOR = National Organics Reconnaissance; PPA = Pollution Prevention Act; PP list = Priority Pollutant List; RCRA = Resource Conservation and Recovery Act; SARA = Superfund Amendments Reauthorization Act; SDWA = Safe Drinking Water Act; TSCA = Toxic Substances Control Act.

Section 319 establishes a regulatory link between nonpoint source pollution and groundwater quality; specifically requiring States to develop best management practices (BMPs) that consider the impact of management on both surface water and groundwater quality. Examples of BMPs include grassy waterways, buffer strips, reduced application rates for pesticides and fertilizers, and the proper storage and disposal of animal wastes.

Besides the CWA, there are a plethora of Federal statutory laws concerning the environment, floodplain management, the US Department of Agriculture, and mining that affect point and nonpoint source pollution and water quality (e.g. National Environmental Policy Act; Federal Environmental Pesticide Control Act; Rare and Endangered Species Act; Safe Drinking Water Act; Federal Insecticide, Fungicide, and Rodenticide Act; Toxic Substance Control Act; The Wild and Scenic River Act; Coastal Zone Management Act; Flood Control Act (and amendments); National Flood Insurance Programs; Flood Disaster Protection Act; Rural Development Act; The Food Security Act; Surface Mining Control and Reclamation Act; Federal Land Policy and Management Act). Novotny and Olem (1994) provide an excellent discussion of the laws, regulations, and policies related to point and nonpoint sources of pollution. Table 3 shows the relative importance of pollutants with respect to their source. Domenico and Schwartz (1990) provide a list of the EPA priority pollutants.

ASSESSMENT OF POINT AND NONPOINT SOURCE POLLUTION

Assessment involves the determination of change of some constituent over time and space. This change can be measured in either real time or simulated with a model. Real-time measurements reflect the activities of the past, whereas simulations can provide useful glimpses into the future. Both means of assessment are valuable. Related to real-time measurement and monitoring, the distribution and trends of pesticides in the atmosphere (Majewski and Capel, 1995), groundwater (Barbash and Resek, 1996), surface water (Larson *et al.*, 1997), and fluvial sediments and aquatic biota (Nowell *et al.*, 1999) have been carefully assessed. For more than 10 years, the US Geological Survey's (USGS) *National Water Quality Assessment* (NAWQA) program (see **`http://water.usgs.gov/nawqa`**) has played a key role in the assessment of nonpoint source contamination from various pollutants to both surface water and groundwater, with studies in more than 50 major river basins and aquifers (Gilliom, 2001). An example of the importance of real-time measurement is the field

Table 3 Relative importance of pollutant concentration in soil-water systems (after Peirce *et al.,* 1998)

Nonpoint source	Suspended solids and sediment	BOD	Nutrients	Toxic metals	Trace elements	Pesticides	Pathogens	Salinity/ TDS
Urban storm runoff	M	L-M	L	H	M	L	H	M
Construction	H	N	L	N-L	N-L	N	N	N
Highway de-icing	N	N	N	N	N	N	N	H
Instream hydrologic modification	H	N	N	N-H	N-H	N	N	N-H
Noncoal mining	H	N	N	M-H	M-H	N	N	M-H
Agriculture								
• Nonirrigated crop production	H	M	H	N-L	N-L	H	N-L	N
• Irrigated crop production	L	L-M	H	N-L	H	M-H	N	H
• Pasture and range	L-M	L-M	H	N	H-L	N	N-L	N-L
• Animal production	M	H	M	N-L	N-L	N-L	L-H	N-L
Forestry								
• Growing	N	N	L	N	N	L	N-L	N
• Harvesting	M-H	L-M	L-M	N	N	L	N	N
Residuals management	N-L	L-H	L-M	L-H	N-H	N	L-H	N-H
On-site sewage disposal	L	M	H	L-M	L-M	L	H	N
Instream sludge accumulation	H	H	M-H	L-H	L-H	M-H	L	N
Direct precipitation	N	N	N-M	L	L	L	N-L	N
Air pollution fallout	M	L	L-M	L-H	N-M	L-M	N-L	N
Natural background	L-M	L-H	M	N-M	N-M	N	N-L	N-H

N = Negligible; L = Low; M = Moderate; H = High; TDS = Total Dissolved Solids.

characterization of the April 1977 accidental spill of approximately 1900 L of the pesticide EDB (ethylene dibromide) within approximately 20 m of a well that provided drinking water to the village of Kunia on the Hawaiian island of Oahu. The point source area around this spill would eventually become a *Superfund* site.

Modeling

A distinct advantage of simulation is that it can be used to alter the occurrence or future impact of detrimental conditions. Obviously, simulation cannot replace *real* data. It should also be pointed out that field observations and simulation are not mutually exclusive entities. Simulation requires considerable information to establish (i) the problem of interest (e.g. climatic data, land-cover characteristics, spatial distributions of near-surface soil-hydraulic properties) and (ii) test (in some cases) the model's performance (e.g. spatial and temporal distribution of chemical concentrations). The use of mathematical models for assessing point and nonpoint sources of pollution, especially after ground truth comparisons, fills in information gaps, identifies critical areas and chemicals for future monitoring, and provides the *what if* capability needed for both regulation and remediation.

By definition, a mathematical model integrates existing knowledge into a framework of rules, equations, and relationships for the purpose of quantifying how a system behaves. As long as a model is applied over the range of conditions from which it was developed, it serves as a useful tool for prognostication. When sufficient information exists to characterize point and nonpoint pollution at a given time, a model can be calibrated (e.g. adjusting model parameter values within a reasonable range so that simulated concentrations closely match observed concentrations) and subsequently employed to make predictions (Loague and Green, 1991). When sufficient information does not exist, a model can still be used in a *what if* mode to address questions related to potential impacts (Abrams and Loague, 2000). Models can range in complexity from the simplest empirical equation to complex sets of partial-differential equations that are only solvable with numerical approximation techniques.

Addiscott and Wagenet (1985) present a categorization for models on the basis of a conceptual approach, distinguishing between deterministic and stochastic, and mechanistic and functional. The key distinction between deterministic and stochastic models is, according to Addiscott and Wagenet (1985), that deterministic models "presume that a system or process operates such that the occurrence of a given set of events leads to a uniquely definable outcome" while stochastic models "presuppose the outcome to be uncertain". Stochastic models consider the statistical credibility of both input conditions and model predictions, whereas deterministic models ignore any uncertainties in their formulation. The second level of model distinction is between mechanistic and functional models. As described by Addiscott and Wagenet (1985), "mechanistic is taken [here] to imply that the model incorporates the most fundamental mechanisms of the process, as presently understood", whereas the term functional is used for "models that incorporate simplified treatments of solute and water flow and make no claim to fundamentality, but do, thereby, require less input data and computer expertise for use".

Three categories of, for example, deterministic models are: (i) regression models, (ii) overlay and/or index models, and (iii) transient-state solute transport models. Regression models generally use multiple linear regression techniques to relate various causative factors. Overlay and/or index models, which can be property or process based, compute an index of pollutant mobility. Transient-state, process-based models are capable of simulating the movement of a pollutant in a dynamic flow system. Transient-state, process-based models describe some or all of the processes involved in, for example, solute transport (e.g. advection, dispersion, diffusion, and retardation). Corwin *et al.* (1997) summarize several geographical information system (GIS) based nonpoint source pollution modeling efforts.

Uncertainty

There can be considerable uncertainty in modeling point and nonpoint source pollution. So, why model? In general, there are two idealized uses for simulation. The first is the prediction (or forecasting) of future events based upon a calibrated and validated model. The second use is the development of concepts for the design of future experiments to improve the understanding of processes. There are three sources of error inherent to modeling: (i) model error, (ii) input error, and (iii) parameter error. Model error results in the inability of a model to simulate the given process, even with the correct input and parameter estimates. Input error is the result of errors in the source terms (e.g. soil-water recharge, chemical application rates). Input error can arise from measurement, juxtaposition, and/or synchronization errors. Parameter error has two possible connotations. For models requiring calibration, parameter error usually is the result of model parameters that are highly interdependent and nonunique. For models with physically based parameters, parameter error results from an inability to represent aerial distributions on the basis of a limited number of point measurements. The aggregation of model error, input error, and parameter error is the total (or simulation) error. For multiple-process and comprehensive model simulation, error is complicated further by the propagation of error between model components.

In general, the methods for characterizing uncertainty can be grouped into three categories (Loague and Corwin, 1996): (i) first-order analysis, (ii) sensitivity analysis, and

(iii) Monte Carlo analysis. First-order analysis is a simple technique for quantifying the propagation of uncertainty from input parameter to the model output. The first-order approximation of functionally related variables is obtained by truncating a Taylor-series expansion (about the mean) for the function after the first two terms. Sensitivity analysis is used to measure the impact that changing one factor has on another. The sensitivity of a model's output to a given input parameter is the partial derivative of the dependent variable with respect to the parameter. Monte Carlo analysis is a stochastic technique of characterizing the uncertainty in complex hydrologic response model simulations. The Monte Carlo method considers each model input parameter to be a random variable with a probability density function (PDF). Monte Carlo simulations are based upon a large number of realizations, from every input parameter distribution, created through sampling the different PDFs with a random number generator. A separate hydrologic response simulation is made for each parameter realization. The number of possible simulations, based upon all the combinations of parameter realizations, is infinite; therefore, a finite number of cases (e.g. several hundred) are usually investigated. Estimates of the average simulated hydrologic response, and the associated uncertainty, are made from the combined outputs of the simulations (i.e. the total ensemble of the different realizations). Loague and Corwin (1996) provide examples of first-order uncertainty analysis, sensitivity analysis, and Monte Carlo simulation.

THREE EXAMPLES

There are thousands of documented cases of point and nonpoint source pollution, far too many to list here. Three examples (from the first author's work) that combine field observations and modeling are briefly discussed in this section. The second example (II) is a local-scale legacy assessment of *point source* pollution from three Manufactured Gas Plants. The first (I) and third (III) examples are both for *nonpoint source* pollution. Example I is a regional-scale legacy assessment associated with agrochemical use in California's Central Valley. Example III is a regional-scale assessment focused on the future management of forested areas on the Canary Island of Tenerife. Assessments of the type shown in Examples I, II, and III can be useful for regulatory decisions and remediation efforts. It is worth pointing out that both examples I and II were undertaken related to major civil actions.

Example I: Nonpoint Source Pollution, the Fresno Case Study

Between the late 1950s and the time of its statewide cancellation in August of 1977, there was widespread use of DBCP (1,2-dibromo-3-chloropropane) throughout the San Joaquin Valley in California. More than two decades after its cancellation, DBCP-contaminated groundwater persisted as a problem in the San Joaquin Valley. The objective of the Fresno Case study (Loague *et al.*, 1998a,b) was to address, from a simulation perspective supported by field observations, if *label recommended* NPS applications were likely to be the principal source of the DBCP groundwater contamination in Fresno County.

The San Joaquin Valley, at the southern end of California's Central Valley, extends in a southeasterly direction for approximately 400 km from just south of Sacramento. East-central Fresno County, situated between the San Joaquin River to the north and the Kings River to the south, is the largest agricultural county in the valley. The spatial distribution of DBCP use in the study area, between 1960 and 1977 (see Figure 2a), was estimated using land-cover maps for different years.

The numerical model used for the 1D simulations (Loague *et al.*, 1998a) of dissolved phase DBCP concentration profiles in the unsaturated zone was PRZM-2 (Mullins *et al.*, 1993). The potential fate and transport of DBCP between the surface and the water table for multiple NPS applications, related to different and changing land use between 1960 and 1977, was quantitatively estimated for 1172 elements for a 35-year period. The aggregate of the DBCP concentrations loaded to the water table for each grid element make up the annual loading files for the 3D saturated transient transport simulations. The numerical models used for the 3D simulations (Loague *et al.*, 1998b) of saturated subsurface fluid flow and DBCP transport are MODFLOW (McDonald and Harbaugh, 1988) and MT3D (Zheng, 1992), respectively. Recharge to the water table for the saturated simulations was estimated as the residual (precipitation plus irrigation minus evapotranspiration) in the PRZM-2 water-balance simulations. The area focused upon for the saturated simulations was represented by a 3D finite-difference grid made up of 76 440 elements (i.e. 2184 1 km^2 surface elements with 35 layers).

The 13-step approach used by Loague *et al.* (1998a,b) to simulate the impact of multiple DBCP applications under changing land use and groundwater pumping/recharge in the Fresno study area are summarized in Table 4. Figure 2(b) shows a 1999 snapshot of the simulated DBCP loading water table from the Fresno Case study. The simulation results from the Fresno Case study lead to the following general comments (Loague *et al.*, 1998a,b):

- The areas most likely to facilitate DBCP leaching through the entire unsaturated soil profile were targeted.
- The first appearance of DBCP above the detectable limit at the water table was simulated as most likely occurring between 1961 and 1965.
- The estimated DBCP concentrations reaching the saturated subsurface exceed the maximum contaminant level (MCL) at several locations at different times. The

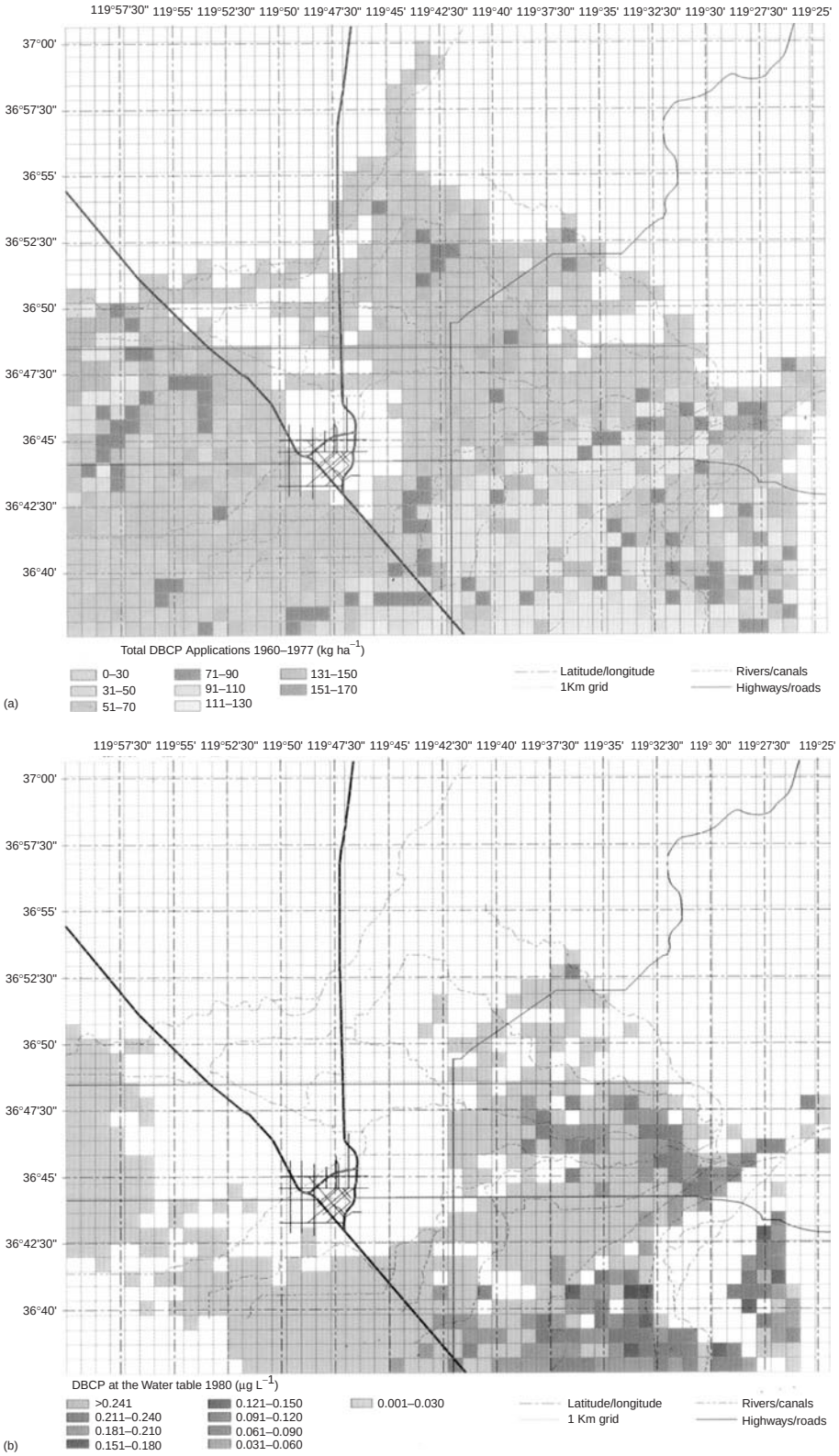


Figure 2 Components of the Fresno Case study example (i.e. Example I) of nonpoint source pollution (After Loague *et al.*, (1998a) © 1998, with permission from Elsevier). (a) Aggregate of the estimated DBCP applications. (b) Simulated 1990 DBCP loading at the water table (Note, this snapshot is taken from a continuous simulation animation). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Table 4 Steps used in the Fresno Case study (after Loague *et al.*, 1998a,b)

1. Approximate the climatic history (1950–1994).
2. Approximate the distribution of soils.
3. Approximate the land-cover history (1958–1994).
4. Approximate the average irrigation history (1960–1994).
5. Approximate the water table depth history (1960–1994).
6. Simulate (with PRZM-2) transient unsaturated fluid flow and DBCP transport (1960–1994).
7. Abstract the DBCP concentration at the water table for each element (1960–1994).
8. Approximate the geology.
9. Approximate the distribution of saturated hydraulic conductivity.
10. Approximate the recharge history (1960–1994).
11. Approximate the pumping history (1960–1994).
12. Simulate (with MODFLOW) groundwater flow (1960–1994).
13. Simulate (with MT3D) saturated transient DBCP transport (1960–1994).

Table 5 Steps used in the simulations of three former manufacture gas plants (after Abrams and Loague, 2000)

1. Approximate the climatic history (MGP inception* – 2000).
2. Approximate the soil properties for each site.
3. Approximate the operational history of each MGP.
4. Simulate (with LEACHM) unsaturated fluid flow and dissolved naphthalene transport for subsurface sources (MGP inception – 2000).
5. Abstract the dissolved naphthalene concentration at the water table for each source (MGP inception – 2000).
6. Approximate the geology.
7. Approximate the distribution of saturated hydraulic conductivity of each site.
8. Simulate (with MODFLOW) groundwater flow for each site (MGP inception – 2000).
9. Simulate (with MT3D) saturated dissolved naphthalene transport for each site (MGP inception – 2000).

first appearance above the MCL was between 1965 and 1970 (note, by 1990, the concentrations are below the MCL).

- Relative to the size of the study area, the extent and duration of the estimated DBCP contamination was small.
- DBCP concentrations are a function of the spatial and temporal variations in (i) the application rates, (ii) the application frequency, (iii) the unsaturated profile thickness, (iv) the soil-hydraulic properties, and (v) the near-surface sorption.
- The DBCP plume evolves (grows and retracts) with time owing to the loading rates at the water table.

Example II: Point Source Pollution, Three Manufactured Gas Plants

From the early 1800s to about 1950, manufactured gas plants (MGPs) were operated in the United States to produce gas from coal or oil. By the time they began to close down, when the distribution of natural gas became more economical, approximately 2000 MGPs had been constructed throughout the USA. The contamination legacy resulting from the production of manufactured gas poses a significant and ongoing groundwater quality problem. The disposal of by-products from the manufacturing process at MGPs contributed to the contamination of MGP sites. Although coal tar was perhaps the most abundant manufacturing by-product, it was not always considered waste as it was sold for multiple uses (e.g. dyes, explosives, pesticides, pharmaceutical preparations, pipe coatings, plastics, road tar, roofing pitch, and wood preservatives). Because of its commercial value, coal tar was usually stored in holding tanks. After a MGP was decommissioned, the coal tar was often left in the holding tanks. Fifty years later, many of these holding tanks remained as *point sources* of contamination.

The purpose of the study reported by Abrams and Loague (2000) was to address, in a *what if* (forensic) mode, using all available data, if contamination emanating from the holding tanks at three former MGP sites (located in Indiana), could cause undesirable impacts to groundwater quality. The nine-step approach used by Abrams and Loague (2000) to simulate the subsurface transport of dissolved naphthalene, under variable precipitation for ~150 years, is summarized in Table 5. The numerical model used for the 1D unsaturated naphthalene transport simulations in this study was a slightly modified version of LEACHM (Hutson and Wagenet, 1992). The characterization, through simulation, of the temporally variable naphthalene loading histories at the water table was used for the input to subsequent 3D simulations of saturated subsurface fluid flow with MODFLOW (McDonald and Harbaugh, 1988) and solute transport with MT3D (Zheng, 1992).

The simulation results reported by Abrams and Loague (2000) indicate that accidental releases of dissolved naphthalene, from the holding tanks at the three former MGP sites, most likely resulted in severe, negative groundwater quality impacts. Figure 3 illustrates, for one of the three sites, components of effort reported by Abrams and Loague (2000). The assumptions made by Abrams and Loague (2000) in conducting these simulations had little impact on the overall conclusion that leaks from the holding tanks can reasonably be expected to lead to groundwater contamination. The sensitivity analyses performed by Abrams and Loague (2000) support this conclusion. It is interesting to note that the approaches used in Examples I (see Table 4)

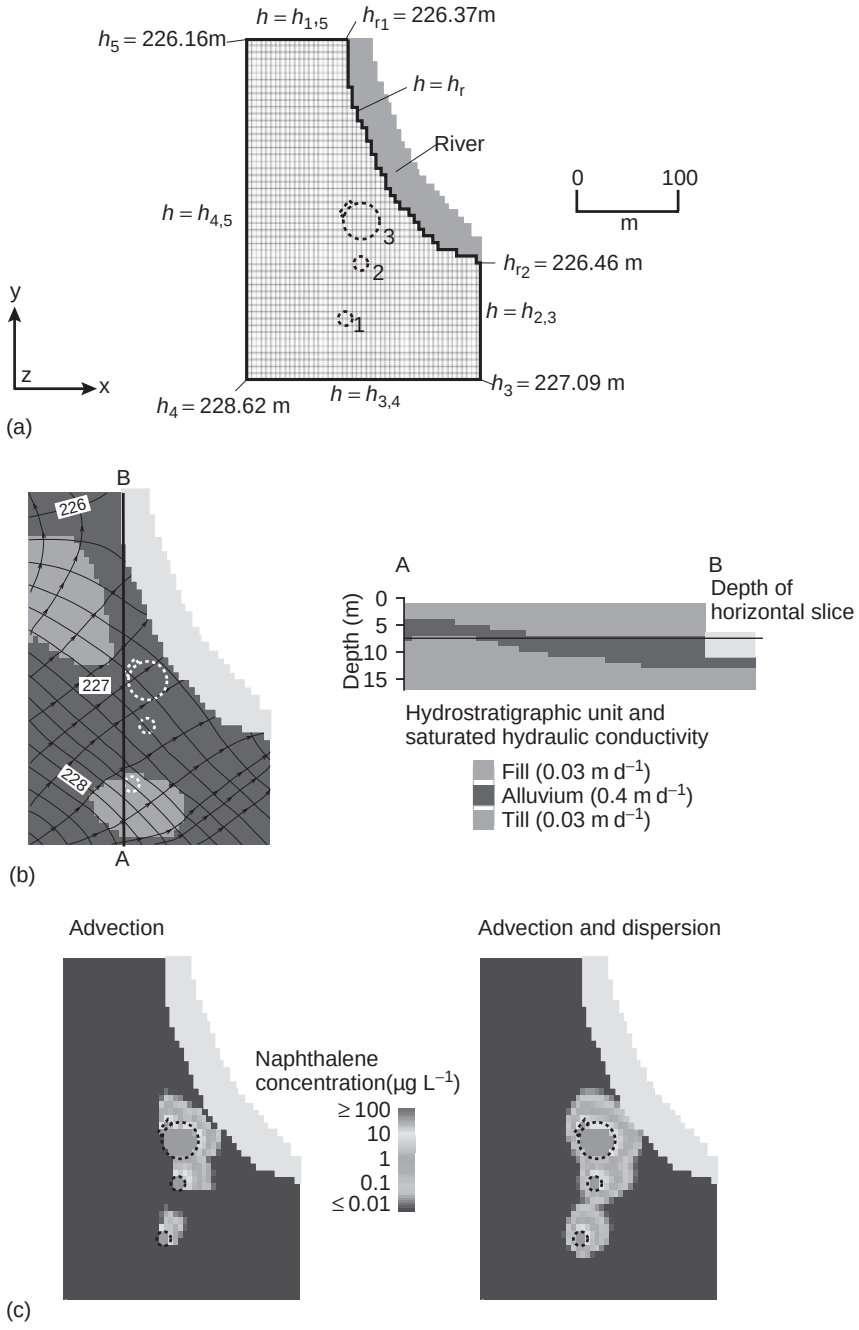


Figure 3 Components of the manufactured gas plant example (i.e. Example II) of point source pollution (Reproduced from Abrams and Loague (2000), by permission of Springer). (a) Site map showing finite-difference grid and boundary-value problem. (b) Site map and vertical cross section showing the spatial distribution of hydrostratigraphic units and saturated hydraulic conductivity values. (c) Snapshots from two saturated-zone solute transport simulations (Note, these snapshots are taken from continuous simulation animations). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and II (see Table 5) are very similar. The major differences between the nonpoint and point source pollution examples are the scale of the two problems and the characterization of the sources (i.e. dispersed and well defined).

Example III: Nonpoint Source Pollution, the Forested Areas of Tenerife

It is widely known that the use of pesticides in modern agriculture can result in groundwater contamination.

The impact of pesticide use in forestry has received considerably less attention. Pesticide use in forest management falls into two main categories. The first category consists of applying herbicides, soil insecticides, and fumigants for site preparation before reforestation, using herbicides to control undesired vegetation during initial tree growth, and applying insecticides and fungicides to prevent and control sporadic outbreaks of pests. The second category of pesticide use in forest management is the application of herbicides to clear firebreaks and road edges. Pesticide applications in forest management began during the late 1950s and early 1960s, with chlorinated insecticides, such as DDT. In many cases, organochlorines are still manufactured today for use in less developed countries.

The focus of the study reported by Diaz–Diaz and League (2001) was to estimate the potential for groundwater contamination on the Canary Island of Tenerife resulting from the use of pesticides for forest management purposes. In Tenerife, there are currently no guidelines for the use of pesticides in forested areas despite the ongoing use of some chemicals. Diaz–Diaz and League (2001) used an index-based model to rank the leaching potential of 50 pesticides that are or could be used for forest management in Tenerife forests. Once the pesticides having the greatest leaching potential were identified, regional-scale groundwater vulnerability assessments with consideration for uncertainty were generated using soil, climatic, and chemical information in a GIS framework for all of the pine forest areas.

On the basis of the leaching potential ranking, Diaz–Diaz and League (2001) suggest that, for the pine forest areas of Tenerife (see Figure 4a), the potential leachers are picloram (potassium salt of 4-amino-3,5,6-trichloro-2-pyridinecarboxylic acid), tebuthiuron (*N*-(5-(1,1-dimethyl)-1,3,4-thiadiazol-2-yl)-*N*,*N*′-dimethylurea), carbofuran (2,3-dihydro-2,2-dimethyl-7-benzofuranyl-*n*-methylcarbamate), triclopyr (triethylamine salt of 3,5,6-trichloro-2-pyridinyloxyacetic acid), and hexazinone (3-cyclohexyl-6-(dimethylamino)-1-methyl-1,3,5-triazine-2,4-(1*H*,3*H*)-dione). Following the identification of the five potential leachers, Diaz–Diaz and League (2001) prepared regional-scale groundwater vulnerability assessments in a GIS-driven format, for the pine forest areas of Tenerife. Two *hot spots* were identified on the northern side of the study area, their locations corresponding to soils classified as Inceptisols and high recharge rates. The uncertainties in the leaching assessments, related to the uncertainties in the soil, recharge, and pesticide data, were shown to be significant. Figure 4(b) illustrates the distribution of soils and recharge across the pine forest areas of Tenerife. Figure 4(c) shows the regional-scale groundwater vulnerability assessments, and their uncertainty, for picloram and carbofuran.

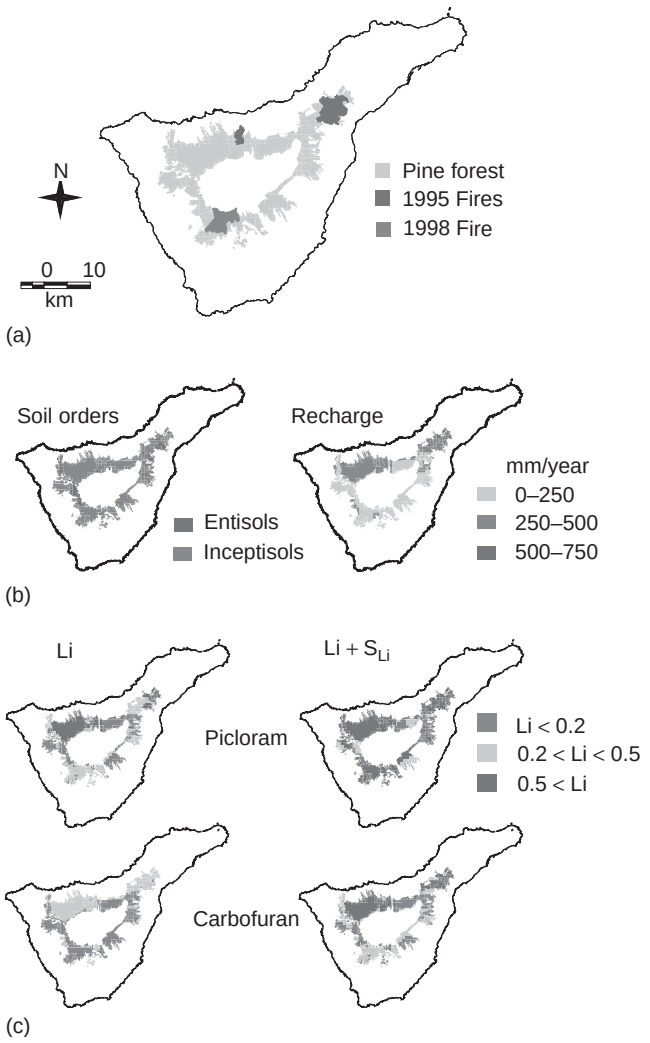


Figure 4 Components of the Tenerife example (i.e. Example III) of nonpoint source pollution (Reproduced from Diaz–Diaz and League (2001), by permission of Alliance Communications). (a) Distribution of pine forest areas, showing the locations of fires in 1995 and 1998. (b) Distribution of soil orders (left) and average recharge rates (right) for the pine forest areas in (a). (c) Distribution of groundwater vulnerability estimates (left) and groundwater vulnerability estimates with consideration of uncertainty (right) for two pesticides [notes: the leaching index (Li) values range between zero and one, the larger the value, the more likely it is that the pesticide will leach; the uncertainty is represented by one standard deviation (S) in Li]. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

EPILOGUE AND FOOD FOR THOUGHT

This article only provides an introduction to the huge subject of point and nonpoint source pollution. The value of real-time measurements for assessing point and nonpoint sources of pollution and the heroic contributions of the NAWQA program are stressed. The use (and types) of

mathematical models as well as the characterization of uncertainties associated with modeling point and nonpoint source problems are covered, illustrated by three examples.

There has been no effort in the allotted space here to address the exposure/human-health outcome question as it relates to point and nonpoint source pollution. Unquestionably, point and nonpoint sources of pollution pose a potentially serious long-term health threat. The EPA, CDC (Center for Disease Control and Prevention), NIH (National Institutes for Health), and the USGS (and their counterparts in Canada and Europe) have each focused considerable energy on developing the linkages between environmental indicators (contaminants), exposure routes, and human-health outcomes. Nevertheless, there is a feeling amongst some scientists and policymakers that the perceived severity of the threats resulting from point and nonpoint source pollution, relative to human health, must be better documented. This view is clearly expressed in Al Freeze's (2000) book *The Environmental Pendulum*. Anyone interested in this controversy should investigate further.

The reader interested in point and nonpoint source pollution may also want to read the following chapters: **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**; **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**; **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**; **Chapter 69, Solute Transport in Soil at the Core and Field Scale, Volume 2**; **Chapter 78, Models of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2**; **Chapter 79, Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport, Volume 2**; **Chapter 91, Water Quality, Volume 3**; **Chapter 100, Water Quality Modeling, Volume 3**; **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**; **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**; **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**; **Chapter 153, Groundwater Pollution and Remediation, Volume 4** and **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**.

REFERENCES

Abrams R.H. and Loague K. (2000) Legacies from three manufactured gas plants: groundwater quality impacts. *Hydrogeology Journal*, **8**, 594–607.

Addiscott T.M. and Wagenet R.J. (1985) Concepts of solute leaching in soils: a review of modelling approaches. *Journal of Soil Science*, **36**, 411–424.

Barbash J.E. and Resek E.A. (1996) *Pesticides in Ground Water: Distribution, Trends, and Governing Factors*, Ann Arbor Press: Chelsea.

Carson R. (1962) *Silent Spring*, Houghton Mifflin: Boston.

Commoner B. (1971) *The Closing Circle*, Knoph: New York.

Corwin D.L. and Loague K. (Eds.) (1996) *Applications of GIS to the Modeling of Non-Point Source Pollutants in the Vadose Zone*, Special Publication 48, Soil Science Society of America: Madison.

Corwin D.L., Loague K., and Ellsworth T.R. (Eds.) (1999) *Assessment of Non-Point Source Pollution in the Vadose Zone*, Geophysical Monograph 108, AGU Press: Washington.

Corwin D.L., Vaughan P.J. and Loague K. (1997) Modeling nonpoint source pollutants in the vadose zone with GIS. *Environmental Science and Technology*, **31**, 2157–2175.

Corwin D.L. and Wagenet R.J. (1996) Applications of GIS to the modeling of non-point source pollutants in the vadose zone: a conference overview. *Journal of Environmental Quality*, **25**, 403–411.

Diaz–Diaz R. and Loague K. (2001) Assessing the potential for pesticide leaching for the pine forest areas of Tenerife. *Environmental Toxicology and Chemistry*, **20**, 1958–1967.

Domenico P.A. and Schwartz F.W. (1990) *Physical and Chemical Hydrogeology*, John Wiley & Sons: New York.

Freeze R.A. (2000) *The Environmental Pendulum: A Quest for the Truth About Toxic Chemical, Human Health and Environmental Protection*, University of California Press: Berkeley.

Gilliom R.J. (2001) Pesticides in the hydrologic system – what do we know and what's next? *Hydrological Processes*, **15**, 3197–3201.

Gore A. (1992) *Earth in the Balance: Ecology and the Human Spirit*, Houghton Mifflin: New York.

Harr J. (1995) *A Civil Action*, Vintage Books: New York.

Hutson J.L. and Wagenet R.J. (1992) *LEACHM: A Process-Based Model of Water and Solute Movement, Transformations, Plant Uptake, and Chemical Reactions in the Unsaturated Zone (Version 3)*, *Research Series 92–3*, Department of Soil, Crop, and Atmospheric Sciences, New York State College of Agriculture and Life Sciences, Cornell University: Ithaca.

Larson S.J., Capel P.D. and Majewski S.S. (1997) *Pesticides in Surface Waters: Distribution, Trends, and Governing Factors*, Ann Arbor Press: Chelsea.

Leopold A. (1949) *A Sand County Almanac*, Oxford University Press: New York.

Loague K., Abrams R.H., Davis S.N., Nguyen A. and Stewart I.T. (1998b) A case study simulation of DBCP groundwater contamination in Fresno County, California: 2. Transport in the saturated subsurface. *Journal of Contaminant Hydrology*, **29**, 137–163.

Loague K. and Corwin D.L. (1996) Uncertainty in regional-scale assessments of non-point source pollutants. In *Applications of GIS to the Modeling of Non-Point Source Pollutants in the Vadose Zone*, Corwin D.L. and Loague K. (Eds.), Special Publication 48, Soil Science Society of America: Madison, pp. 131–152.

Loague K. and Green R.E. (1991) Statistical and graphical methods for evaluating solute transport models: overview and application. *Journal of Contaminant Hydrology*, **7**, 51–73.

Loague K., Lloyd D., Nguyen A., Davis S.N. and Abrams R.H. (1998a) A case study simulation of DBCP groundwater contamination in Fresno County, California: 1. Leaching through the unsaturated subsurface. *Journal of Contaminant Hydrology*, **29**, 109–136.

Majewski S.S. and Capel P.D. (1995) *Pesticides in the Atmosphere: Distribution, Trends, and Governing Factors*, Ann Arbor Press: Chelsea.

McDonald M.G. and Harbaugh A.W. (1988) *A Modular Three-Dimensional Finite-Difference Groundwater Flow Model*, Scientific Software Group: Washington.

Mercer J.W., Silka L.R. and Faust C.R. (1983) Modeling groundwater flow at Love Canal, New York. *Journal of Environmental Engineering*, **109**, 924–942.

Mullins J.A., Carsel R.F., Scarbough J.E. and Ivery A.M. (1993) *PRZM-2, A Model for Predicting Pesticide Fate in the Crop Root and Unsaturated Soil Zones: Users Manual for Release 2*, EPA-600/R-93/046, USEPA Environmental Research Laboratory: Athens.

Novotny V. and Olem H. (1994) *Water Quality – Prevention, Identification, and Management of Diffuse Pollution*, Van Nostrand Reinhold: New York.

Nowell L.H., Capel P.D. and Dileanis P.D. (1999) *Pesticides in Stream Sediments and Aquatic Biota: Distributions, Trends, and Governing Factors*, CRC Press: Boca Raton.

Peirce J.J., Weiner R.F. and Vesilind P.A. (1998) *Environmental Pollution and Control*, *Fourth Edition*, Butterworth-Heinemann: Boston.

Zheng C. (1992) *MT3D: A Modular Three-Dimensional Model for Simulation of Advection, Dispersion, and Chemical Reactions of Contaminants in Groundwater Systems (Version 1.5)*, S.S. Papadopulos and Associates: Bethesda.

95: Acidic Deposition: Sources and Effects

CHARLES T DRISCOLL[1], KATHY FALLON-LAMBERT[2], AND LIMIN CHEN[1]

[1]Civil and Environmental Engineering Department, Syracuse University, Syracuse, NY, US
[2]Ecologic: Analysis and Communications, Quechee, VT, US

Acidic deposition delivers acids and acidifying compounds to the Earth's surface, which are then transported through soil, vegetation, and surface waters and, in turn, set off a cascade of adverse ecological effects. Acidic deposition has altered forest soil by accelerating the leaching of available base cations, enhancing the accumulation of sulfur and nitrogen, and increasing the concentration of dissolved inorganic aluminum in soil waters. Soils that are compromised by acidic deposition are less able to neutralize additional amounts of acidic deposition, and provide poorer growing conditions for plants. Acidic deposition has impaired the surface water quality by lowering pH, decreasing acid-neutralizing capacity (ANC), and increasing concentrations of dissolved inorganic aluminum. These changes have reduced the species diversity and abundance of aquatic life. Regulatory controls initiated in Europe and North America over the last three decades have decreased emissions of sulfur dioxide and to a lesser extent nitrogen oxides. Emission reductions have resulted in widespread decreases in concentrations of sulfate in surface waters, with some waters showing an increase in ANC. Given the loss of acid-neutralizing base cations and the accumulation of sulfur and nitrogen in soil, many ecosystems have become more sensitive to additional acidic deposition and recovery will likely be delayed. Long-term research shows that deeper emissions cuts will lead to greater and faster recovery from acidic deposition in impacted regions.

ACIDIC DEPOSITION

Acidic atmospheric deposition, popularly referred to as *acid rain*, is the transfer of strong acids and acid-forming substances from the atmosphere to the Earth's surface. Acidic deposition is comprised of sulfuric and nitric acids, and ammonium derived from atmospheric emissions of sulfur dioxide, nitrogen oxides, and ammonia, respectively. These compounds are emitted by the burning of fossil fuels and by agricultural activities. Once such compounds enter an ecosystem, they can acidify soil and surface waters and bring about a series of ecological changes. The term acidic deposition encompasses all forms in which these compounds are deposited to the Earth, including gases, particles, rain, snow, clouds, and fog (see Box 1; *see* **Chapter 28, Clouds and Precipitation, Volume 1**; **Chapter 30, Topographic Effects on Precipitation, Volume 1**). Acidic deposition was first reported in the United Kingdom in the latter half of the nineteenth century (Gorham, 1989). Ecological effects were first documented in Scandinavia in the 1960s with the link between acidic deposition, surface water acidification, and loss of fisheries (Gorham, 1989). Atmospheric deposition of sulfate, nitrate, and ammonium are elevated in eastern North America, Europe, and large portions of Asia (Rodhe *et al.*, 1995).

Electric utilities account for the greatest proportion of anthropogenic sulfur dioxide emissions in Europe and North America. For example in the United States in 2002, the major sources of sulfur dioxide emissions were electric utilities (67%), industrial combustion (15%), and industrial processes (9%; EPA 2004). Transportation sources – including cars, trucks, and non-road vehicles (i.e. construction equipment) – accounted for more than 50% of anthropogenic nitrogen oxide emissions in the United States. Other major sources of nitrogen oxides include electric utilities (22%) and industrial combustion (14%). Ammonia emission are derived largely from livestock waste and fertilized soil (83% of total ammonia; Driscoll *et al.*, 2003). Motor vehicles and industrial processes also contribute to ammonia emissions.

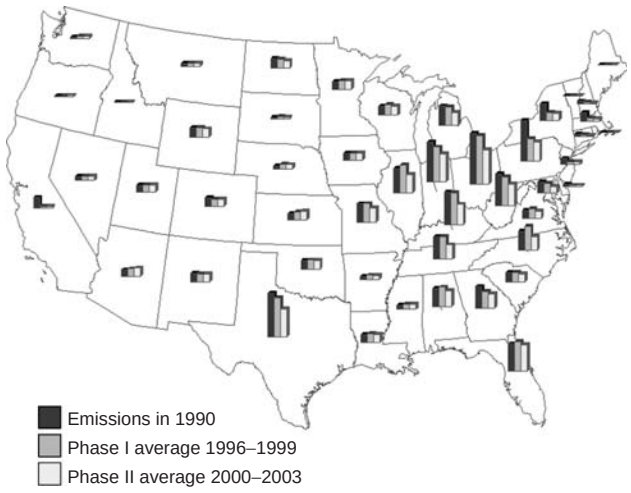
Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Box 1 How is acidic deposition monitored?

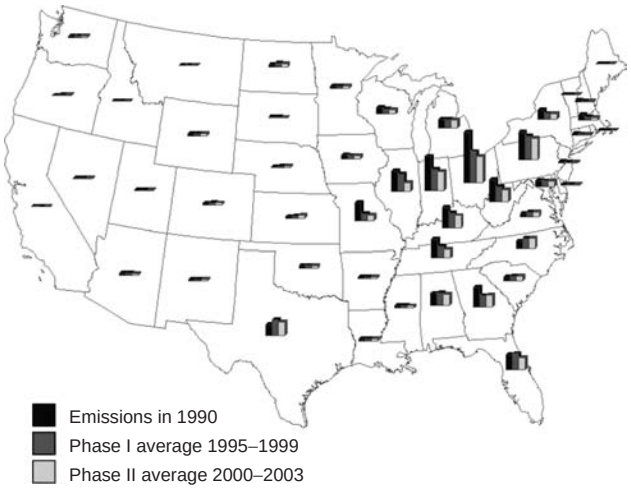
Acidic deposition occurs in three forms: wet deposition, which falls as rain, snow, sleet, and hail; dry deposition, which includes particles, gases, and vapor; and cloud or fog deposition which occurs at high altitudes and coastal areas. In the United States, wet deposition has been monitored at more than 200 sites, by both independent researchers and the interagency National Atmospheric Deposition Program/National Trends Network (**http://nadp.sws.uiuc.edu/**). Dry deposition is monitored at 70 sites in the United States by the U.S. Environmental Protection Agency Clean Air Status and Trends Network (**http://www.epa.gov/castnet/**), and at 13 other sites by the National Oceanic and Atmospheric Administration AIRMON-dry Network. Cloud and fog deposition has been monitored for limited periods at selected high-elevation sites, largely by independent researchers. Dry and cloud deposition patterns are extremely variable over space and time, making it difficult to characterize patterns. Therefore, even though cloud and dry deposition comprise a significant proportion of total deposition, this report primarily presents general patterns and trends of wet deposition. Some researchers also measure bulk deposition, in which deposition is collected in an open collector. Bulk deposition is greater than wet deposition because it includes some dry deposition.

An airshed or source area is an area where "significant portions of emissions result in deposition of air pollutants to a region" (**www.epa.gov**). In North America, emissions of sulfur dioxide are highest in the midwestern United States (hereafter the Midwest), with seven states in the Ohio River Valley accounting for 51% of the total sulfur dioxide emissions in the United States in 2002 (Figure 1a). Five of these states are also amongst the highest emitters of nitrogen oxides (Figure 1b). Moreover, the Midwest is a significant source of atmospheric ammonia. In addition to regional pollution sources, local emissions of sulfur dioxide and nitrogen oxides from electric utilities and motor vehicles have significant impacts on local air quality. Analysis of continental air currents shows that a multistate region, including the Midwest comprise the source area for sulfur dioxide, nitrogen oxide, and ammonium emissions that are transported downwind to acid-sensitive areas of eastern North America (Butler *et al.*, 2001d).

There have been significant efforts to reduce emissions of acidic and acidifying substances in North America and Europe over the past three decades. Although regulatory controls have decreased emissions, levels remain high compared to background conditions. Importantly, emissions and deposition of base cations (i.e. elements such as calcium and magnesium that help counteract acidic deposition) have declined substantially since the early 1960s with the enactment of pollution controls to reduce fine particulate matter (Hedin *et al.*, 1994).



(a) Scale: Largest bar equals 500 000 tons of emissions in Ohio, 1990



(b) Scale: Largest bar equals 2.2 million tons of emissions in Ohio, 1990

Figure 1 State by state emissions of sulfur dioxide (a) and nitrogen oxides (b) in the United States. Values are shown for three periods: 1990, after Phase I (1995–1999), and after Phase II (2000–2003) of control of utility emissions in response to 1990 Amendments of the Clean Air Act (after US EPA, 2004). Note the bars are scales to 1990 emissions for Ohio (2 million metric tons for sulfur dioxide and 454 000 tons for nitrogen oxides). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Total sulfur dioxide emissions in the United States peaked in 1973 at approximately 29 million metric tons annually. The 1970 and 1990 Amendments of the Clean Air Act (CAAA) led to a 52% decrease in sulfur dioxide emissions nationwide, to approximately 13.9 million metric tons in 2002. The multistate source area for the eastern North America has shown substantial decreases in sulfur dioxide emissions over this period (Figure 2). The 1990 CAAA set a cap of 14 million metric tons of total annual sulfur dioxide emissions to be achieved by 2010. The cap

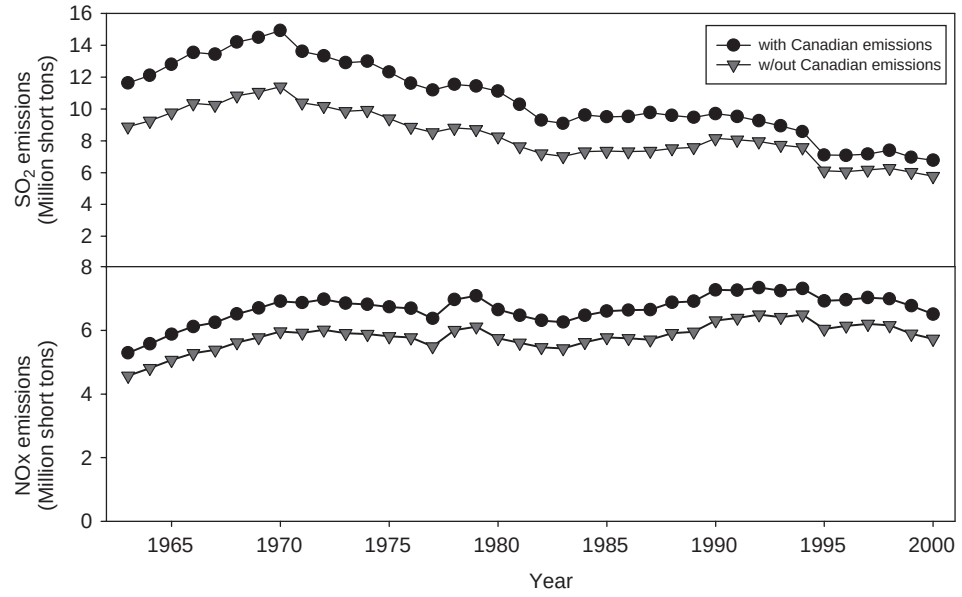


Figure 2 Annual emissions of sulfur dioxide and nitrogen oxides for the source area of the Hubbard Brook Experimental Forest. The source area was determined by 24-h back trajectory analysis. Shown are emissions from both United States and Canadian sources. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

on electric utilities is set at 8.9 million metric tons and the cap on industrial sources is 5.6 million metric tons to be reached by 2010.

Nitrogen oxide emissions in the United States have increased over the past decades, peaking at nearly 22.7 million metric tons in 1990. From 1990 to 2002, nitrogen oxide emissions have decreased by 12%. The 1990 CAAA calls for an additional reduction that will result in the emission of 1.8 million fewer tons of nitrogen oxide than the level that would have occurred without the CAAA. However, no cap on total annual emissions of nitrogen oxides was set. Nevertheless, it is expected that nitrogen emissions will decrease gradually in the future because of a variety of federal and state emission control programs. In contrast to sulfur dioxide, the multistate source area for the eastern North America has shown little change in nitrogen dioxide emissions since the early 1970s although some decrease has been evident in recent years (Figure 2).

Ammonia emissions play an important role in the acidification of soil and surface waters. Deposition of ammonium accounts for approximately 30% of the total nitrogen deposition measured in eastern North America and has not changed appreciably over the past 30 years. Trends in ammonia emissions in the United States are consistent with this pattern and have shown little change over the past 10 years.

European efforts to reduce emissions of air pollutants have been brought together in a series of protocols under the United Nations Economic Council for Europe (UN/ECE) Convention on Long Range Transboundary Air Pollution (LRTAP; Sundqvist *et al.*, 2002; Ferrier *et al.*, 2001). The first binding protocol was the 1985 Protocol on the Reduction of Sulfur Emissions that was intended to reduce sulfur dioxide emissions by at least 30% by 1993 compared to the 1980 levels. This was expanded in 1994 with the Protocol on Further Reductions of Sulfur, with the objective of decreasing sulfur dioxide emissions by 80% by 2010 from 1980 values. In 1988, the Protocol concerning Control of Emissions of Nitrogen Oxides stabilized nitrogen oxide emissions. Finally, the 1999 Protocol to Abate Acidification, Eutrophication, and Ground-level Ozone established national caps for multiple air pollutants, including sulfur dioxide, nitrogen oxides, and ammonia, with attainment expected by 2010 (Kurz *et al.*, 2001). These actions have resulted in marked reductions in emissions of sulfur dioxide (65%) and nitrogen oxides (30%) from 1990 to 2002. Ammonia emissions have remained essentially constant over the same period. Once implemented, sulfur dioxide emissions will decrease by at least 63%, nitrogen oxide emissions by 41%, and ammonia emissions by 17% relative to 1990 values.

An important framework for emission reductions and ecosystem recovery is "critical loads." Critical loads are based on the idea that emission control strategies should be effect-driven. A critical load is the total deposition of a pollutant to an ecosystem below which significant harmful ecological effects do not occur (Nilsson and Grennfelt, 1988). Following the UN/ECE Convention on LRTAP, maps of critical loads were developed for Europe (Posch *et al.*, 1995). These maps have been revised (Posch and Hettelingh, 2001) and the resulting critical loads were used

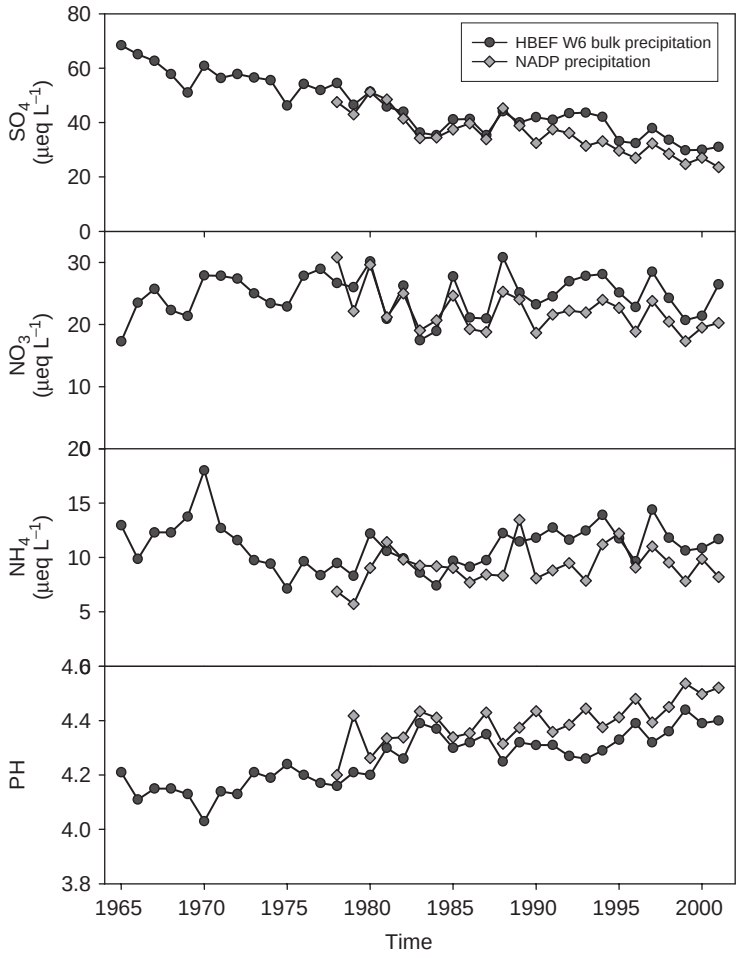


Figure 3 Annual volume-weighted sulfate, nitrate, and ammonium concentrations and pH in bulk and wet deposition at the Hubbard Brook Experimental Forest, New Hampshire, 1963–2000. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

to regulate emissions through the 1999 Protocol. Critical loads have only been applied to a limited extent in the United States.

In contrast to North America and Europe, emissions in Asia have been increasing and are expected to continue to increase in the coming decades (Galloway, 1995; Klimont *et al.*, 2001).

Trends in acidic deposition mirror emission trends in the source area. For example, over the past 30 years in eastern North America, sulfate deposition has declined, but nitrogen and ammonium deposition have remained relatively stable (Figure 3). Decreases in precipitation sulfate have coincided with increases in pH. The Hubbard Brook Experimental Forest in New Hampshire has one of the longest continuous records of precipitation chemistry (see Box 2). Long-term data from the Hubbard Brook show declining concentrations of sulfate in bulk deposition since the mid-1960s and wet deposition since the late 1970s (see Figure 3). On the basis of these long-term data, there is a strong positive correlation that exists between sulfur dioxide emissions in the source area and sulfate concentrations in precipitation at Hubbard Brook

Box 2 The Hubbard Brook Experimental Forest

The Hubbard Brook Experimental Forest is a long-term ecological research site established by the United States Department of Agriculture Forest Service in the White Mountains of New Hampshire to investigate the structure and function of forest and aquatic ecosystems, and their response to disturbance (Likens and Bormann, 1995; Groffman *et al.*, 2004). Hubbard Brook was the site where acidic deposition was first reported in North America (Likens *et al.*, 1972). Hubbard Brook receives elevated inputs of acidic deposition and the forest ecosystem is very sensitive to these inputs. There have been long-term measurements and studies of acidic deposition and its effects on forests and streams at Hubbard Brook (Likens *et al.*, 1996; Driscoll *et al.*, 2001).

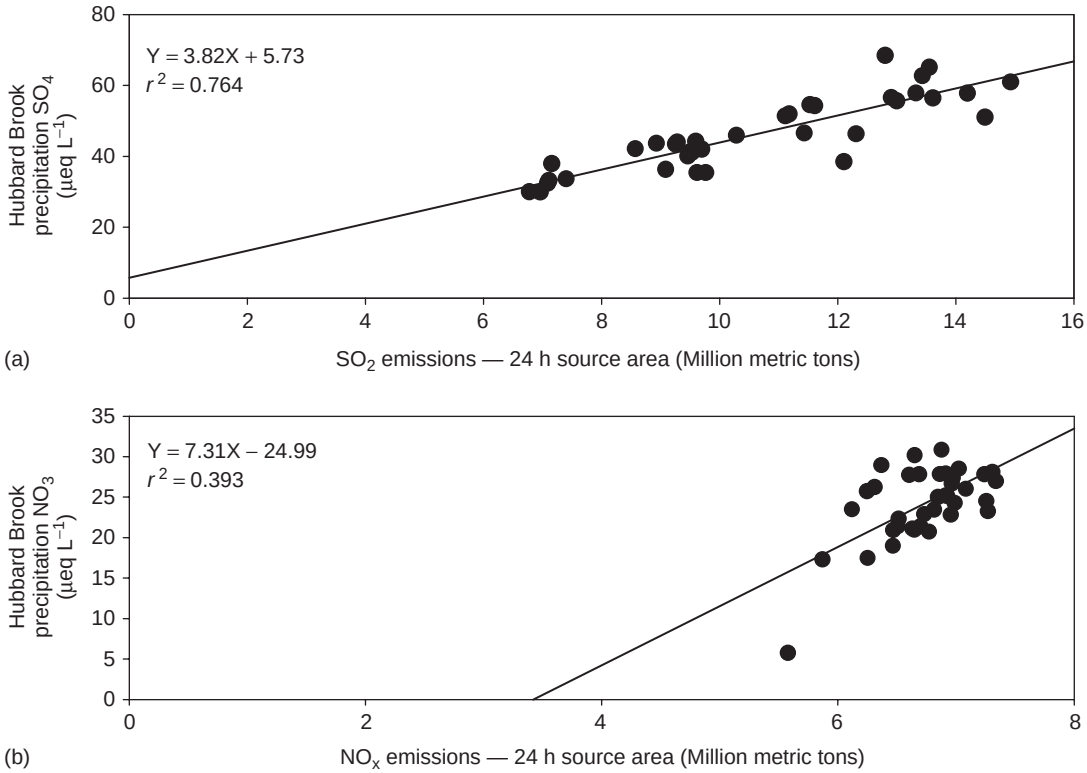


Figure 4 Relationships between sulfur dioxide and nitrogen oxide emissions for the source area of the Hubbard Brook Experimental Forest (see Figure 2) and annual volume-weighted concentrations of sulfate and nitrate in bulk deposition

(Figure 4). It is now expected that the sulfate concentration of atmospheric deposition will decrease in a direct linear response to decreases of sulfur dioxide emissions in the source area.

The relationship between sulfur dioxide emissions and wet sulfate deposition extends throughout the eastern United States. The portion of the eastern United States with high wet deposition of sulfate decreased markedly over the interval 1989–1991 to 2001–2003 (see Figure 5). These reductions in wet sulfate deposition are consistent with the emissions reductions called for in the 1990 CAAA.

In contrast to sulfate trends in wet deposition, concentrations of nitrate or ammonium at Hubbard Brook have not shown large changes since 1963 (see Figure 3). There is a relationship between nitrate concentrations in bulk deposition at Hubbard Brook and nitrogen oxide emissions in the source area (Butler *et al.*, 2003), but the relationship is much weaker than observed for sulfate (Figure 4). This weak relationship is due to the fact that nitrogen oxide emissions and nitrate in bulk deposition have not changed much since measurements were initiated in 1963. Patterns of wet deposition of nitrogen at Hubbard Brook are consistent with the picture across the entire eastern United States, which shows limited change over the last several years (see Figure 6).

EFFECTS OF ACIDIC DEPOSITION ON ECOSYSTEMS

Acidic deposition alters soils, stresses forest vegetation, acidifies lakes and streams, and harms fish and other aquatic life. These effects can alter important ecosystem services such as forest productivity and water quality (*see* **Chapter 193, Markets for Watershed Services, Volume 5**). Decades of acidic deposition have also made many ecosystems more sensitive to continuing pollution. Moreover, the same pollutants that cause acidic deposition contribute to a wide array of other important environmental issues at local, regional, and global scales (see Table 1).

Effects of Acidic Deposition on Forest Ecosystems

Until recently, understanding the effects of acidic deposition on soils was limited. However, current research has shown that acidic deposition has chemically altered forest soils with serious consequences for acid-sensitive ecosystems. Soils compromised by acidic deposition lose their ability to neutralize continuing inputs of strong acids, provide poorer growing conditions for plants, and extend the time needed for ecosystems to recover from acidic deposition. Acidic deposition has altered and continues to alter base-poor forest soils in three important ways. Acidic

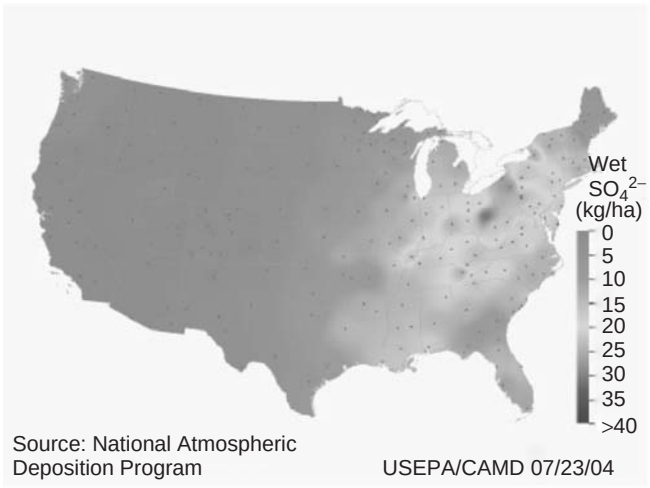


Figure 5 Annual sulfate in wet deposition in the eastern United States for 1989–1991 and 2001–2003 (data were obtained from the National Atmospheric Deposition Program). Note that there have been marked reductions in wet sulfate deposition in response to emission controls. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

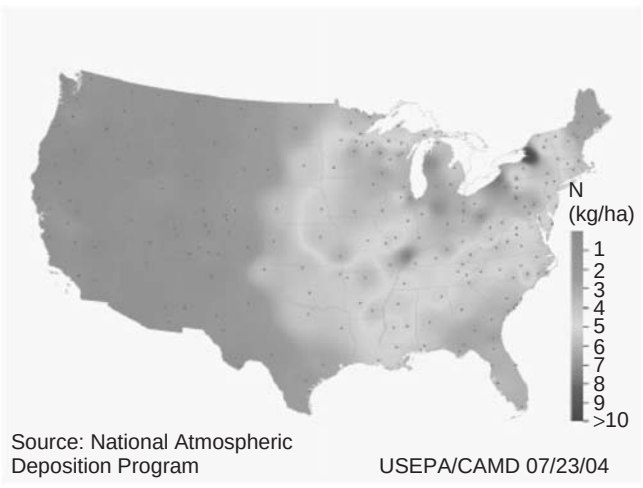


Figure 6 Annual inorganic nitrogen (ammonium plus nitrate) deposited in wet precipitation in the eastern United States for 1989–1991 and 2001–2003 (data were obtained from the National Atmospheric Deposition Program). Note that there have been limited changes in wet deposition of inorganic nitrogen. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

deposition depletes available calcium and other nutrient cations (e.g. magnesium, potassium) from soil, facilitates the mobilization of dissolved inorganic aluminum into soil water, and increases the accumulation of sulfur and nitrogen in soil.

The cycling of calcium and other nutrient cations in forest ecosystems involves the inputs and losses of these materials (Figure 7; *see* **Chapter 69, Solute Transport in Soil at the Core and Field Scale, Volume 2**). For most forest ecosystems, the supply of calcium and other nutrient cations largely occurs by weathering (i.e. the breakdown of rocks and minerals in soil). Calcium and other nutrient cations may also enter forests by atmospheric deposition, although this pathway is generally much smaller than weathering. Losses largely occur by vegetation uptake and drainage waters. An important pool of ecosystem calcium and nutrient cations is the soil-available pool, or the soil cation exchange complex. Plants are generally able to utilize this source of nutrients. Forest ecosystems that are naturally sensitive to acidic deposition are generally characterized by low rates of weathering and generally low quantities of available base cations (i.e. calcium, magnesium, sodium, potassium). Under conditions of elevated inputs of acidic deposition and subsequent transport of sulfate and nitrate in drainage waters, nutrient cations will be displaced from available pools and leached from soil (Ruess and Johnson, 1986). This condition is not problematic for areas with high weathering rates and high

Table 1 The links between sulfur dioxide and nitrogen oxide emissions, acidic deposition, and a range of environmental issues

Problem	Linkage to acid deposition	Reference
Coastal eutrophication	Atmospheric deposition adds nitrogen to coastal waters.	Paerl *et al.*, 2002
Mercury	Deposition of sulfate enhances methylation of mercury. Surface water acidification increases mercury accumulation in fish.	Branfireun *et al.*, 1999; Driscoll *et al.*, 1994
Visibility	Sulfate aerosols diminish visibility and views.	Malm *et al.*, 1994
Climate change	Sulfate, nitrate, and ammonium aerosols may offset global warming in the short-term, but nitrous oxide is a potent greenhouse gas.	Moore *et al.*, 1997 (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**)
Tropospheric ozone	Emissions of nitrogen oxides contribute to the formation of ozone.	NAPAP (1998)
Airborne particulate matter	Emissions of sulfur dioxide, nitrogen oxides, and ammonia contribute to airborne particulate matter and associated health effects.	`http://www.epa.gov/air/urbanair/pm/index.html`
Corrosion and damage to structures and monuments	Acidic substances enhance corrosion	Sherwood and Lipfert, 1990

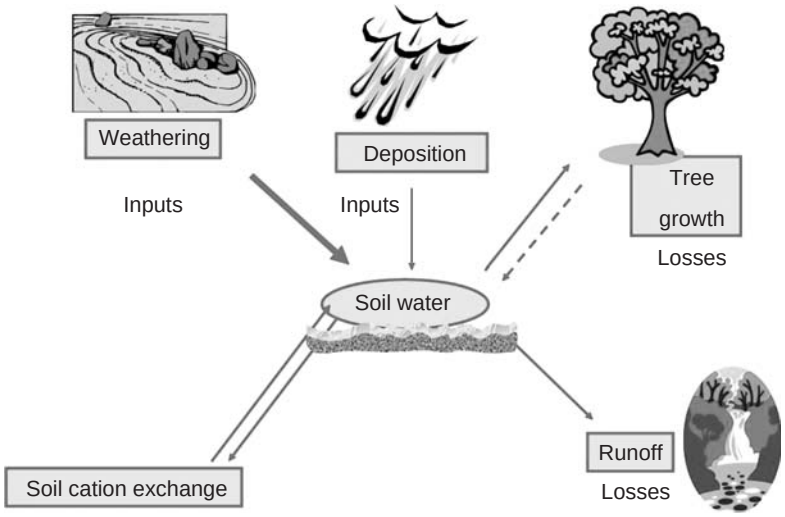


Figure 7 Conceptual diagram illustrating calcium cycle in forest watersheds. Inputs of calcium include weathering and atmospheric deposition; of these weathering is usually the greatest. Losses of calcium include tree accumulation and stream runoff. Under conditions of elevated acidic deposition stream losses increase, potentially depleting available calcium from the ecosystem, particularly from the soil exchange complex. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

pools of available nutrient cations. However, over the past century, acidic deposition has accelerated the loss of large amounts of available calcium and magnesium from the soil in acid-sensitive areas (Likens *et al.*, 1996; Kirchner and Lydersen, 1995; Huntington *et al.*, 2000). Depletion occurs when base cations are displaced from the soil by acidic deposition at a rate faster than they can be replenished by the slow breakdown of rocks or the deposition of base cations from the atmosphere. This depletion of base cations fundamentally alters soil processes, compromises the nutrition of some trees, and hinders the capacity for sensitive soils to recover from inputs of acidic deposition.

Dissolved inorganic aluminum is often released from soil to soil water, vegetation, lakes, and streams in forested regions with high acidic deposition, low stores of available calcium, high soil acidity, and limited watershed retention of atmospheric inputs of sulfate and/or nitrate (Cronan and Schofield, 1990). High concentrations of dissolved inorganic aluminum can be toxic to plants, fish, and other organisms. Concentrations of dissolved inorganic aluminum in streams and lakes in acid-sensitive regions

receiving high inputs of acidic atmosphere are often above levels considered toxic to fish and much greater than concentrations observed in forested watersheds with low inputs of acidic deposition (Driscoll *et al.*, 1988).

Acidic deposition results in the accumulation of sulfur and nitrogen in forest soils. As sulfate is released from the soil in response to decreases in emissions and atmospheric deposition of sulfur, it is transported to adjacent streams and lakes (Likens *et al.*, 2000). The recovery of surface waters in response to emission controls has therefore been delayed and will not be complete until the sulfate left by a long legacy of acidic deposition is released from the soil.

Similarly, nitrogen has accumulated in soil beyond the amount needed by the forest and appears now to be leaching into surface waters in Europe and North America (Dise and Wright, 1995; Aber *et al.*, 2003). Forests typically require more nitrogen for growth than is available in the soil. However, several recent studies suggest that in some areas, nitrogen levels are above what forests can use and retain. This condition is referred to as "nitrogen saturation" (Aber *et al.*, 1989; Aber *et al.*, 1998). Note that the levels at which atmospheric nitrogen deposition can result in elevated leaching losses of nitrate from forest watersheds appear to be higher in Europe ($9-25\,\mathrm{kg\,N\,ha^{-1}\,year^{-1}}$) than eastern North America ($7-8\,\mathrm{kg\,N\,ha^{-1}\,year^{-1}}$). The reason for this difference is not evident, but may be due to a greater fraction of atmospheric nitrogen deposition occurring as ammonium in Europe; ammonium inputs are more readily retained in watersheds than nitrate. Alternatively, this difference may be due to highly managed forests in Europe and greater nitrogen retention compared to eastern North America.

Acidic Deposition Stress to Trees

Although it is difficult to separate the effects of air pollution from other stresses, recent research shows that acidic deposition appears to have contributed to the decline of red spruce trees throughout the eastern North America and sugar maple trees in central and western Pennsylvania in the United States. Symptoms of tree decline include poor crown condition, reduced tree growth, and unusually high levels of tree mortality.

Red spruce and sugar maple are the tree species that have been most intensively researched, and therefore, they provide instructive case studies of the effects of acidic deposition on trees. Red spruce and sugar maple research has shown that acidic deposition has both direct and indirect effects on trees. In acid-impacted forests, acidic deposition harms trees directly by leaching calcium from the leaves and needles (i.e. foliage) of trees, rendering them more susceptible to winter injury. Acidic deposition can also affect trees indirectly by changing the underlying soil chemistry. In acid-sensitive soils, acidic deposition depletes available nutrient cations, such as calcium and magnesium, which are important plant nutrients that are necessary to maintain the health and vigor of trees. The depletion of nutrient cations also leads to soil acidification that increases the availability of aluminum to the trees' roots, thereby impairing the ability of trees to obtain necessary nutrients from the soil.

Red Spruce

Acidic deposition appears to be the major cause of red spruce decline at high elevations in North America. Since the 1960s, more than half of large canopy trees in the Adirondack Mountains of New York and the Green Mountains of Vermont and approximately one-quarter of large canopy trees in the White Mountains of New Hampshire have died. Significant growth declines and winter injury to red spruce have been observed throughout its range, suggesting that damage from acidic deposition is likely widespread (DeHayes *et al.*, 1999).

Recent research indicates that the decline of red spruce is linked to the leaching of calcium from cell membranes in spruce needles by acidic deposition (DeHayes *et al.*, 1999). The loss of calcium renders the needles more susceptible to freezing damage, thereby reducing the tolerance of trees to low temperatures and increasing the occurrence of winter injury and subsequent tree damage or death. In addition, elevated aluminum concentrations in the soil, resulting from soil acidification, limits the ability of red spruce to take up water and nutrients through its roots. This limitation can lead to nutrient deficiencies that can lower a tree's tolerance to environmental stress and cause decline.

Sugar Maple

The decline of sugar maples has been studied in the eastern United States since the 1950s and there is growing evidence that sugar maple decline is linked to acidic deposition. Extensive mortality among sugar maples in Pennsylvania appears to result from deficiencies of base cations, coupled with other stresses such as insect defoliation or drought. Sugar maples are most prone to die on sites where base cation concentrations in soil or foliage are lowest (Horsley *et al.*, 2000). Data from many acid-sensitive regions link the loss of soil calcium and magnesium with the leaching of these base cations by acidic deposition. Low levels of base cations can cause a nutrient imbalance and reduce a tree's stress tolerance. As such, acidic deposition is a predisposing factor in sugar maple decline. Under these conditions, the likelihood increases that stresses such as insect infestation and drought will cause the dieback of a tree's crown or kill a tree.

Finally, there may be adverse effects on other tree species. For example, one might speculate that hardwood species such as white ash and basswood, which prefer rich sites high in nutrient cations, may experience problems in areas where nutrient cations have been depleted by acidic deposition. However, additional research is needed

Box 3 What is ANC?

Acid-neutralizing capacity, or ANC, is the ability of water from a lake or stream to neutralize strong acid (Stumm and Morgan, 1996). ANC is an important measure of the impacts of acidic deposition as well as an indicator of chemical recovery from acidic deposition. Surface waters with ANC values below 0 μeq L^{-1} during base flow conditions are considered chronically acidic. Waters with ANC values ranging from 0 to 50 μeq L^{-1} are susceptible to episodic acidification. Waters with ANC values greater than 50 μeq L^{-1} are less sensitive to acidic deposition. The capacity of a watershed to prevent decreases in ANC and resist the effects of acidic deposition depends on many factors, including climate, soil conditions, surficial and bedrock geology, and land-use history.

to completely assess the response of these tree species to acidic deposition.

Effects of Acidic Deposition on Aquatic Ecosystems

Acidic deposition degrades water quality by lowering pH levels (i.e. increasing acidity), decreasing acid-neutralizing capacity (ANC; see Box 3), and increasing dissolved inorganic aluminum concentrations. While sulfate concentrations in lakes and streams have decreased over the last 20 years, they remain high compared to background conditions ($<20\,\mu eq\,L^{-1}$; Sullivan, 1991).

An important characteristic influencing the acid–base status of surface waters is the supply of naturally occurring organic solutes. These materials include naturally occurring organic acids that decrease the ANC of waters and mobilize aluminum from soil through complexation reactions. Watersheds with an abundance of wetlands typically have high concentrations of dissolved organic carbon, and associated surface waters can be naturally acidic due to organic acids. However, in contrast to watersheds that are acidified by acidic deposition, waters that are acidic due to organic acids have high concentrations of aluminum which is largely complexed with organic solutes and therefore less toxic to aquatic biota (Driscoll *et al.*, 1980, 1988).

Acidification of surface waters due to elevated inputs of acidic deposition have been reported in many acid-sensitive areas receiving elevated inputs of acidic deposition, including Great Britain, Nordic countries, northern, central, and eastern Europe (Evans *et al.*, 2001), southwestern China (Seip *et al.*, 1995), southeastern Canada (Jeffries, 1991), the northeastern United States (Driscoll, 1991) the upper Midwest (Cook and Jager, 1991), and the Appalachian mountain region of the United States (Elwood, 1991). Large portions of the high-elevation western United States are also potentially sensitive to acidic deposition (Fenn *et al.*, 2003); however, atmospheric deposition to this region is relatively low. Concern over effects of acidic deposition in the mountain regions of western United States may be overshadowed by potential effects of elevated nitrogen deposition, including eutrophication of naturally nitrogen-limited lakes.

One of the most highly impacted areas in North America is the Adirondack region of New York. A comprehensive survey of Adirondack lakes greater than 0.2 ha in surface area was conducted between 1984 and 1987 to obtain detailed information on the acid–base status of waters in this region (Kretser *et al.*, 1989). Of the 1469 lakes surveyed, 24% had summer pH values below 5.0. Also, 27% of the lakes surveyed were chronically acidic (i.e. ANC $< 0\,\mu eq\,L^{-1}$) and an additional 21% were susceptible to episodic acidification (i.e. ANC between 0 and $50\,\mu eq\,L^{-1}$; see Box 4). Note that 54% of these acid-sensitive lakes (733 lakes) are characterized by relatively low concentrations of dissolved organic carbon (i.e. $<6\,mg\,C\,L^{-1}$). The chemical composition of these lakes suggests that their acidity was largely derived from inputs of sulfate associated with acidic deposition (Driscoll *et al.*, 2003). In contrast, 46% of the lakes are characterized by high concentrations of dissolved organic carbon (i.e. $>6\,mg\,C\,L^{-1}$) and naturally occurring organic acids. These lakes are probably naturally acidic. While the contribution of naturally occurring acidity is greater in these lakes, sulfate was the dominant anion; the acidity of these lakes has been clearly enhanced by acidic deposition.

Decreases in pH and elevated concentrations of dissolved inorganic aluminum have reduced the species diversity and

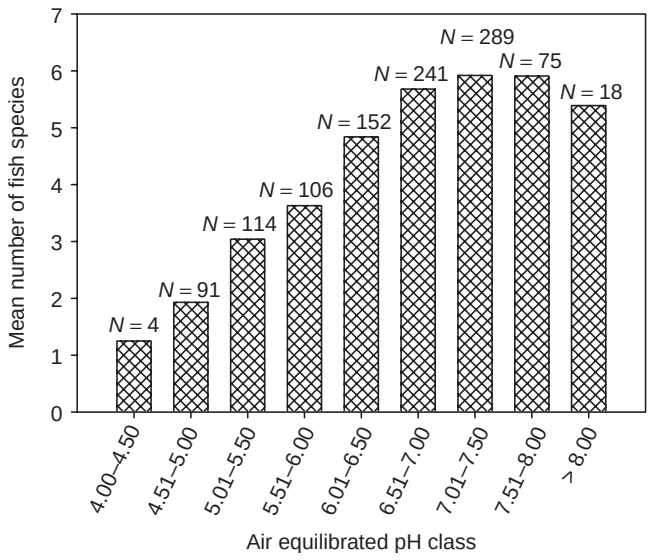


Figure 8 The mean number of fish species for pH classes from 4.0 to 8.0 in lakes in the Adirondack region of New York. *N* represents the number of lakes in each pH class (After Driscoll *et al.* 2001. © American Institute of Biological Sciences)

Table 2 Biological effects of surface water acidification (After Baker *et al.*, 1990)

pH decrease	General biological effects
6.5–6.0	Small decrease in species richness of phytoplankton, zooplankton, and benthic invertebrate communities resulting from the loss of a few highly acid-sensitive species, but no measurable change in total community abundance or production
	Some adverse effects (decreased reproductive success) may occur for highly acid-sensitive species (e.g. fathead minnow, striped bass)
6.0–5.5	Loss of sensitive species of minnow and dace, such as blacknose dace and fathead minnow; in some waters decreased reproductive success of lake trout and walleye, which are important sport fish species in some areas
	Visual accumulations of filamentous green algae in the littoral zone of many lakes, in some streams
	Distinct decrease in the species richness and change in species composition of the phytoplankton, zooplankton, and benthic invertebrate communities, although little if any change in total community biomass or production
5.5–5.0	Loss of several important sport fish species, including lake trout, walleye, rainbow trout, and smallmouth bass; as well as additional nongame species such as creek chub
	Further increase in the extent and abundance of filamentous green algae in lake littoral areas and streams
	Continued shift in the species composition and decline in species richness of the phytoplankton, periphyton, zooplankton, and benthic invertebrate communities; decrease in the total abundance and biomass of benthic invertebrates and zooplankton may occur in some waters
	Loss of several additional invertebrate species common in oligotrophic waters, including *Daphnia galeata mendotae, Diaphanosoma leuchtenbergianum*, and *Asplanchna priodonta*; all snails, most species of clams, and many species of mayflies, stoneflies, and other benthic invertebrates Inhibition of nitrification
5.0–4.5	Loss of most fish species, including most important sport fish species such as brook trout and Atlantic salmon; few fish species able to survive and reproduce below pH 4.5 (e.g. central mud minnow, yellow perch, and in some waters, largemouth bass)
	Measurable decline in the whole-system rates of decomposition of some forms of organic matter, potentially resulting in decreased rates of nutrient cycling
	Substantial decrease in the number of species of zooplankton and benthic invertebrates and further decline in the species richness of the phytoplankton and periphyton communities; measurable decrease in the total community biomass of zooplankton and benthic invertebrates in most waters
	Loss of zooplankton species such as *Tropocyclops prasinus mexicanus, Leptodora kindtii*, and *Conochilis unicornis*; and benthic invertebrate species, including all clams and many insects and crustaceans
	Reproductive failure of some acid-sensitive species of amphibians such as spotted salamanders Jefferson salamanders, and the leopard frog

abundance of aquatic life in many streams and lakes in acid-sensitive areas (Table 2). Fish have received the most attention to date, but entire food webs are often adversely affected (Baker *et al.*, 1990).

Decreases in pH and increases in dissolved inorganic aluminum concentrations have diminished the species diversity and abundance of plankton, invertebrates, and fish in acid-impacted surface waters. For example, in the Adirondacks, a significant positive relationship exists between the pH and ANC levels in lakes and the number of fish species present in those lakes (see Figure 8). The Adirondack Lakes Survey showed that 24% of lakes (i.e. 346) in this region do not support fish. These lakes had consistently lower pH and ANC and higher concentrations of dissolved inorganic aluminum than lakes that contained one or more species of fish. Experimental studies and field observations demonstrate that even acid-tolerant fish species such as brook trout have been eliminated from some waters.

Although chronically high acid levels stress aquatic life, acid episodes are particularly harmful because abrupt, large changes in water chemistry allow fish few areas of refuge (see Box 4). High concentrations of dissolved inorganic aluminum are directly toxic to fish and pulses of dissolved inorganic aluminum during acid episodes are a primary cause of fish mortality (Baker *et al.*, 1996; Van Sickle *et al.*, 1996). High acidity and dissolved inorganic aluminum levels disrupt the salt and water balance in a fish's blood, causing red blood cells to rupture and blood viscosity to increase (MacAvoy and Bulger, 1995). Studies show that the viscous blood strains the fish's heart, resulting in a lethal heart attack.

ECOSYSTEM RECOVERY

Recovery from acidic deposition involves decreases in emissions resulting from regulatory controls that in turn lead to reductions in acidic deposition and allow *chemical recovery*. The chemical recovery process is characterized by decreases in concentrations of sulfate, nitrate, and dissolved inorganic aluminum in soils and surface waters. If sufficient, these reductions will eventually lead to increased pH and ANC, as well as higher concentrations of base cations in water and on the soil exchange complex. As

Box 4 Seasonal and episodic acidification

Seasonal acidification is the periodic increase in acidity and the corresponding decrease in pH and ANC in streams and lakes. Episodic acidification is caused by the sudden pulse of acids due to spring snowmelt (*see* **Chapter 114, Snowmelt Runoff Generation, Volume 3**) and large rain events in the spring and fall. Increases in nitrate are important to the occurrence of acid episodes and tend to occur when trees are dormant, therefore using less nitrogen. Short-term increases in acid inputs to surface waters can reach levels that are lethal to fish and other aquatic organisms.

chemical conditions improve, the potential for the second phase of ecosystem recovery, biological recovery, is greatly enhanced.

An analysis of the scientific literature suggests that the following five thresholds can serve as indicators of chemical recovery. If chemical conditions in an ecosystem are above these thresholds, it is unlikely that the ecosystem has been substantially impaired by acidic deposition. Conversely, if chemical conditions are below these thresholds, there is a high likelihood that the ecosystem is vulnerable to acidic deposition (see Table 3).

The time required for chemical recovery varies widely among ecosystems, and is primarily a function of:

- the historic loading rate of sulfur and nitrogen oxides;
- the rate and magnitude of decreases in acidic deposition;
- the extent to which available base cations such as calcium have been depleted from soil;
- the extent to which sulfur and nitrogen have accumulated in the soil and the rate at which they are released as deposition declines;
- the weathering rate of the soil and underlying rock and the associated supply of base cations to the ecosystem; and
- the rate of atmospheric deposition of base cations.

Table 3 Indicators of chemical recovery from acidic deposition

Forest ecosystems
• Soil base saturation of 20% or higher (i.e. the percent of available cations in the soil that are bases)
• Calcium to aluminum molar ratio in the soil solution of 1.0 or greater
Aquatic ecosystems
• Stream and lake pH of 6.0 or higher (except where pH is lower under background conditions)
• Stream or lake ANC of 50 μeq L^{-1} or higher
• Stream or lake concentrations of dissolved inorganic aluminum less than 2 μmol L^{-1}.

As chemical conditions in soils and surface waters improve, *biological recovery* is enhanced. Biological recovery is likely to occur in stages, since not all organisms can recover at the same rate and may vary in their sensitivity to acidic deposition. The current understanding of the response of biological species to improvements in chemical conditions is incomplete, but research suggests that stream macroinvertebrates may recover relatively rapidly (i.e. within 3 years), while lake zooplankton may need a decade or more to fully reestablish. Fish populations in streams and lakes should recover in 5–10 years following the recovery of the macroinvertebrates and zooplankton that serve as food sources (Gunn and Mills, 1998). It is possible that, with improved chemical conditions and the return of other members of the aquatic food web, the stocking of streams and lakes could help accelerate the recovery of fish.

Terrestrial recovery is even more difficult to project than aquatic recovery. Given the life span of trees and the delay in the response of soil to decreases in acidic deposition, it is reasonable to suggest that decades will be required for affected trees on sensitive sites to recover once chemical conditions in the soil are restored. Overall, the timing and extent of chemical and biological recovery depend on how soon and how significantly emissions that cause acidic deposition are reduced. Moreover, human influences (i.e. land disturbance, introduction of exotic or invasive species) in addition to acidic deposition, can delay biological recovery after chemical recovery has occurred.

Long-term stream data from Hubbard Brook reveal a number of long-term trends that are consistent with trends in lakes and streams across Europe and eastern North America (Stoddard *et al.*, 1999; Evans *et al.*, 2001; Stoddard *et al.*, 2003; see Figure 9). Specifically, the concentration of sulfate in streams at Hubbard Brook declined 32% between 1963 and 2000. The pH of streams subsequently increased from 4.8 to 5.0. Although this represents an important improvement in water quality, streams at Hubbard Brook remain acidic compared to background conditions, when stream pH was estimated to be approximately 6.0. Moreover, the ANC at Hubbard Brook – a biologically important measure of a lake or stream's susceptibility to acid inputs – has remained acidic (ANC $< 0\,\mu$eq L^{-1}).

Trends in surface water chemistry in Europe (Evans *et al.*, 2001) and eastern North America (Stoddard *et al.*, 1999) indicate that recovery of aquatic ecosystems impacted by acidic deposition is occurring over a large geographic scale since the early 1980s. Some regions are showing rather marked recovery, while others exhibit low or nonexistent increases in ANC. On the basis of long-term monitoring, virtually all surface waters impacted by acidic deposition in Europe and eastern North America exhibit decreases in sulfate concentrations. This pattern is consistent with decreases in emissions of sulfur dioxide and atmospheric sulfate deposition. The exception to

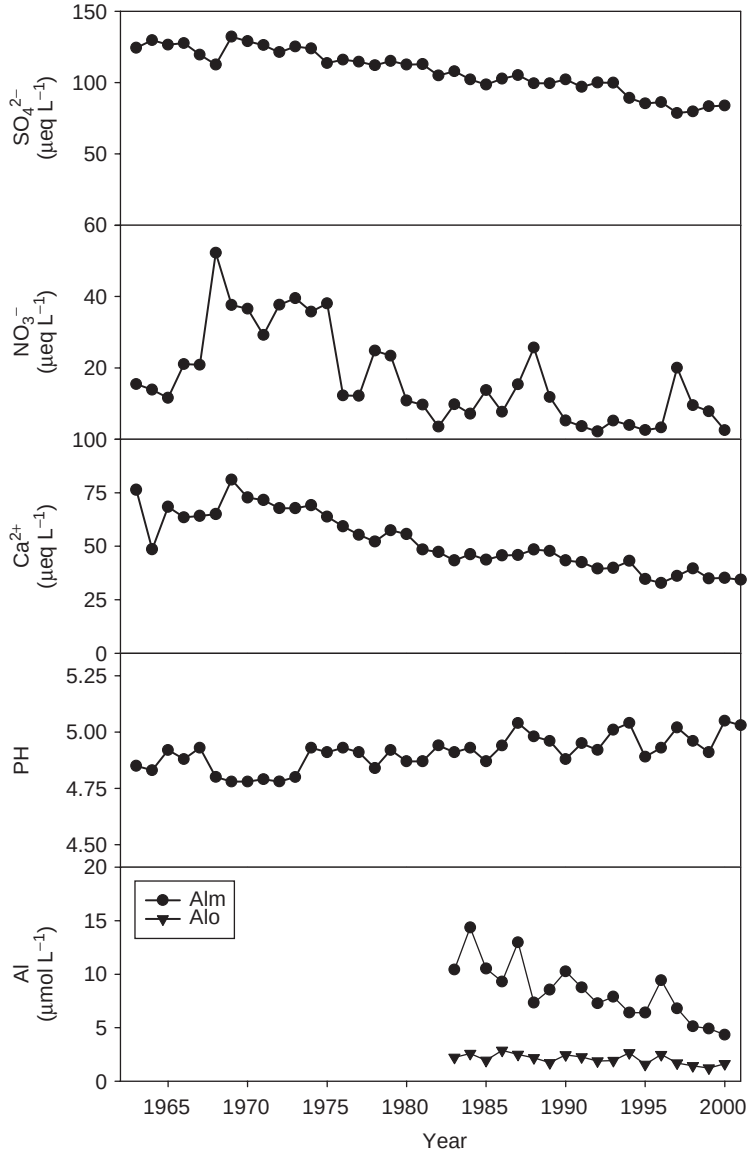


Figure 9 Annual volume-weighted stream water sulfate, nitrate, calcium concentrations, pH, and concentrations of total (Alm) and organic dissolved aluminum (Alo) at the reference watershed of the Hubbard Brook Experimental Forest from 1963 to 2000. Note that dissolved inorganic aluminum is the difference between total and organic dissolved aluminum

this pattern is streams in unglaciated Virginia. Watersheds in this region exhibit strong adsorption of atmospheric sulfate deposition by highly weathered soils. In Europe, the most marked decreases in surface water sulfate have occurred in the Czech Republic and Slovakia, regions that have experienced historically very high rates of atmospheric sulfate deposition. Almost more than half of the surface waters monitored in Europe show an increase in ANC (Evans *et al.*, 2001). The rate of ANC increase in Europe is relatively high. This pattern is in part due to the relatively high rates of sulfate decreases, but also that decreases in base cations only account for about half of the decreases in sulfate plus nitrate, allowing for relatively large rates of ANC increases. In contrast, in the United States only three regions show statistically significant increases in ANC-lakes in the Adirondacks and upper Midwest and streams in the Northern Appalachian Plateau (Stoddard *et al.*, 2003). In the United States, decreases in the sum of base cations closely correspond to decreases in sulfate plus nitrate, limiting rates of ANC increase.

Three factors have limited the recovery in chemical water quality at Hubbard Brook and other watersheds in acid-sensitive regions that have received elevated inputs of acidic deposition. First, levels of acid-neutralizing base cations in surface waters have decreased markedly because of the depletion of available base cations from the soil

and, to a lesser extent, a reduction in atmospheric inputs of base cations. Second, as forest ecosystems mature, their requirement for nitrogen decreases (Aber *et al.*, 1989; Aber *et al.*,1998). As a result, forested watersheds with limited disturbance that extracts nitrogen (e.g. tree harvesting, fire, agriculture) are expected to exhibit increasing losses of nitrate as forests develop. Finally, sulfur has accumulated in the soil under previous conditions of high atmospheric sulfur deposition and is now being released to surface water as sulfate, even though sulfate deposition has decreased.

While there is considerable information about the response of surface waters to decreases in acidic deposition under chronic conditions, much less is known about how episodic acidification responds to these changes. Laudon and Hemond (2002) reported decreases in episodic acidification following decreases in atmospheric sulfur deposition in northern Sweden. Unfortunately, comparable data sets have not been developed for other regions.

An alternative to recovery from controls on emissions of acidic or acid-forming substances is mitigation. Mitigation (or base addition or liming) involves the application of basic materials directly to surface waters or watersheds to neutralize strong acid inputs (Olem *et al.*, 1991). The most common material for mitigation is calcium carbonate (or limestone), although other materials have been effectively used. Mitigation has been practiced in Europe and North America to treat the effects of acidic deposition with some success. Direct application to lakes has been shown to neutralize acidity and allow for the survival of fish and other sensitive aquatic biota. This approach is less successful in lakes with short hydraulic residence times and to recover a reproducing fish population (Driscoll *et al.*, 1996). Alternatively, watershed treatment has been shown to be successful over the longer term by improving the base status of soil and allowing for a reproducing fish population. Note, mitigation is not an attractive alternative to source control. It might be implemented in areas that exhibit severe depletion of exchangeable nutrient cations and/or have biological species that are endangered because of acidic deposition. Acidic deposition impacts ecosystems in remote and wilderness areas that are difficult or inappropriate to treat by base addition. While no negative short-term effects of base treatment have been noted, long-term studies of the ecological response to base treatment have not been conducted.

RECOVERY OF ACID-SENSITIVE ECOSYSTEMS WITH FUTURE DECREASES IN EMISSIONS

To date, emissions targets set in the United States and Europe have been met or exceeded. There are widespread decreases in surface water concentrations of sulfate and some waters are showing increases in ANC. Nevertheless, data suggest that these targets may not be sufficient to achieve the full recovery of sensitive ecosystems. In order to evaluate the extent to which historic and future emissions reductions will facilitate ecosystem recovery from acidic deposition, it is necessary to use acidification models to project the future relationship between emissions, deposition, and chemical recovery of acid-sensitive forest watersheds (see Box 5).

Box 5 Acidification models

Scientists have developed computer models that depict the physical, chemical, and biological processes within forest watersheds. Watershed acidification models can be used as research and management tools to investigate factors responsible for the historical acidification of soil and water, as well as the ecosystem response to anticipated future changes in acidic deposition. In order to effectively predict the pH, ANC, and aluminum concentrations in streams, all major chemicals must be accurately simulated (e.g. sulfate, nitrate, calcium, magnesium). The acidification model PnET-BGC was used for this assessment because it has been rigorously tested at Hubbard Brook and other sites in the northeastern United States, and it allows the user of the model to consider the ecosystem response to multiple chemicals simultaneously. Other frequently used acidification models include Model of Acidification of Groundwater in Catchment (MAGIC) (Cosby *et al.*, 2001), and the nutrient cycling model NuCM (Lui *et al.*, 1992).

We used the model PnET-BGC (Gbondo-Tugbawa *et al.*, 2001) to compare current emissions reductions required by the 1990 CAAA with an additional 55% and 75% cut in emissions of sulfur dioxide and 20% and 30% decreases in nitrogen oxides by 2010 (Figure 10). These scenarios are based on the electric utility emission reductions embodied in bills recently introduced to the US Congress. PnET-BGC considered changes in sulfur dioxide and nitrogen oxide emissions. It was assumed that base cation and ammonium deposition and climate would remain unchanged.

According to the results of the computer model, the 1990 CAAA will have a positive effect on stream concentrations of sulfate at Hubbard Brook, but will not facilitate appreciable progress toward chemical recovery of key indicators of acidification stress, such as pH or ANC. With moderate reductions in emissions (i.e. sulfur dioxide 55%, nitrogen oxides 20%) beyond the requirements of the 1990 CAAA, measurable chemical improvements occur. However, none of the five indicators reach the threshold needed to support complete biological recovery at Hubbard Brook by 2050 (see Table 3). More aggressive reductions in emissions (i.e. sulfur dioxide 75%, nitrogen oxides 30%) beyond the 1990 CAAA hastens and promotes more significant improvements in chemical conditions. For example, under this scenario streams in watersheds similar to Hubbard Brook would change from acidic to nonacidic in roughly

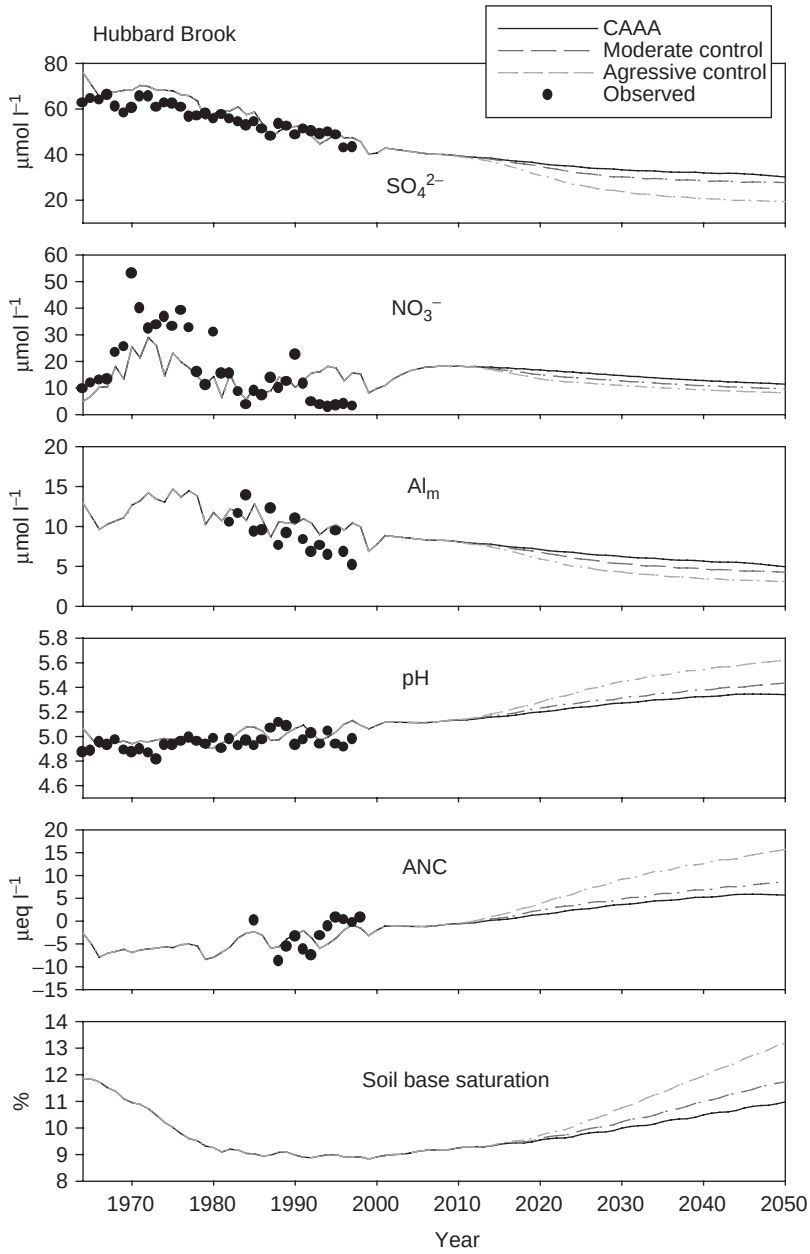


Figure 10 Time series of predictions with the acidification model PnET-BGC of changes in stream chemistry at Hubbard Brook to changes in past and potential future emissions of sulfur dioxide and nitrogen oxides, including the 1990 Amendments of the Clean Air Act and moderate and aggressive emission control scenarios. Shown are model-predicted stream concentrations of sulfate, nitrate, acid-neutralizing capacity, pH and dissolved inorganic aluminum, and soil percent base saturation. Measured values are indicated for comparison. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

20 to 25 years. By 2050, the concentration of aluminum and the base cation content of the soil in these watersheds would begin to approach recovery thresholds or preindustrial levels.

The model results suggest that full implementation of the 1990 CAAA will not bring about substantial improvements in chemical recovery at Hubbard Brook. The results further demonstrate that the process of recovery will be slow, particularly for sensitive systems such as Hubbard Brook. Similar analyses have been conducted at regional scales using PnET-BGC for the Adirondack region of New York (Chen and Driscoll, 2005) and northern New England (Chen and Driscoll in press), with similar results obtained. Other analyses have been conducted to evaluate the response of watersheds in Canada and Europe to future emission reductions (Wright, 2003).

In sum, acidic deposition is a pervasive problem that has had a greater impact on soils, terrestrial vegetation, surface waters, and aquatic biota than previously projected. Although abatement strategies in Europe and North America have had positive effects, emissions remain high compared to background conditions. Given the accumulation of acids and loss of buffering capacity in the soil, many areas are now more sensitive to acidic deposition and have developed an inertia that will delay recovery. Nevertheless, calculations from computer models show that deeper emission cuts will lead to greater and faster recovery from acidic deposition.

Acknowledgment

This article was developed as part of the Hubbard Brook Research Foundation Science-Links program, funded by Jessie B. Cox Charitable Trust, Davis Conservation Foundation, Geraldine R. Dodge Foundation, McCabe Environmental Fund, Merck Family Fund, John Merck Fund, Harold Whitworth Pierce Charitable Trust, the Sudbury Foundation, and the Switzer Environmental Leadership Fund of the New Hampshire Charitable Foundation. Support was also provided by the W. M. Keck Foundation and the National Science Foundation. We are indebted to Gene Likens for the use of long-term precipitation and steam chemistry data at Hubbard Brook. This is a contribution of the Hubbard Brook Ecosystem Study.

REFERENCES

Aber J.D., Goodale C.L., Ollinger S.V., Smith M.-L., Magill A.H., Martin M.E. and Stoddard J.L. (2003) Is nitrogen deposition altering the nitrogen status of northeastern forests? *BioScience*, **53**, 375–390.

Aber J., McDowell W., Nadelhoffer K., Magill A., Berntson G., Kamakea M., McNulty S., Currie W., Rustad L. and Fernandez I. (1998) Nitrogen saturation in temperate forest ecosystems: hypotheses revisited. *BioScience*, **48**, 921–934.

Aber J.D., Nadelhoffer K.J., Steudler P. and Melillo J.M. (1989) Nitrogen saturation in northern forest ecosystems. *BioScience*, **39**, 378–386.

Baker J.P., Gherini S.A., Christensen S.W., Driscoll C.T., Gallagher J., Munson R.K. and Newton R.M. (1990) *Adirondack Lake Survey: An Interpretive Analysis of Fish Communities and Water Chemistry, 1984-87*, Adirondack Lakes Survey Corporation: Ray Brook.

Baker J.P., Van Sickle J., Gagen C.J., DeWalle D.R., DeWalle D.R., Sharpe W.F., Carline R.F., Baldigo B.P., Murdoch P.S., Bath D.W. and Kretser P.J. Jr (1996) Episodic acidification of small streams in the northeastern United States: effects on fish populations. *Ecological Applications*, **6**, 422–437.

Branfireun B.A., Roulet N.T., Kelly C.A. and Rudd J.W.M. (1999) In situ sulphate stimulation of mercury methylation in a boreal peatland: toward a link between acid rain and methyl mercury contamination in remote environments. *Global Biogeochemical Cycles*, **13**, 743–750.

Butler T.J., Likens G.E. and Stunder B.J. (2001) Regional-scale impacts of phase I of the clean air act amendments: the relationship between emissions and concentrations, both wet and dry. *Atmospheric Environment*, **35**, 1015–1028.

Butler T.J., Likens G.E., Vermeylen F.M. and Stunder B.J.B. (2003) The relation between NO_x emissions and precipitation NO_3^- in the eastern USA. *Atmospheric Environment*, **37**, 2093–2104.

Chen L. and Driscoll C.T. (2005) Regional assessment of the response of the acid-base status of lake-watersheds in the Adirondack region of New York to changes in atmospheric deposition using PnET-BGC. *Environmental Science and Technology*, **39**, 787–794.

Chen L. and Driscoll C.T. Regional application of an integrated biogeochemical model to northern New England and Maine. *Ecological Applications* (in press).

Cook R.B. and Jager H.I. (1991) Upper midwest. In *Acidic Deposition and Aquatic Ecosystems: Regional Case Studies*, Charles D.F. (Ed.), Springer-Verlag: New York, pp. 421–466.

Cosby B.J., Ferrier R.C., Jenkins A. and Wright R.F. (2001) Modelling the effects of acid deposition: refinements, adjustments and inclusion of nitrogen dynamics in the MAGIC model. *Hydrology and Earth System Sciences*, **5**, 499–518.

Cronan C.S. and Schofield C.L. (1990) Relationships between aqueous aluminum and acidic deposition in forested watersheds of North America and Northern Europe. *Environmental Science and Technology*, **24**, 1100–1105.

DeHayes D.H., Schaberg P.G., Hawley G.J. and Strimbeck G.R. (1999) Acid rain impacts calcium nutrition and forest health. *BioScience*, **49**, 789–800.

Dise N.B. and Wright R.F. (1995) Nitrogen leaching from European forests in relation to nitrogen deposition. *Forest Ecology and Management*, **71**, 153–161.

Driscoll C.T. (1991) Northeast overview. In *Acidic Deposition and Aquatic Ecosystems: Regional Case Studies*, Charles D.F. (Ed.), Springer-Verlag: New York, pp. 129–132.

Driscoll C.T., Baker J.P., Bisogni J.J. and Schofield C.L. (1980) Effect of aluminum speciation on fish in dilute acidified waters. *Nature*, **284**, 161–164.

Driscoll C.T., Cirmo C.P., Fahey T.J., Blette V.L., Bukaveckas P.A., Burns D.J., Gubala C.P., Leopold D.J., Newton R.M., Raynal D.J., *et al.* (1996) The Experimental Watershed Liming Study (EWLS): Comparison of lake and watershed neutralization strategies. *Biogeochemistry*, **32**, 143–174.

Driscoll C.T., Driscoll K.M., Mitchell M.J. and Raynal D.J. (2003) Effects of acidic deposition on forest and aquatic ecosystems in New York State. *Environmental Pollution*, **123**, 327–336.

Driscoll C.T., Johnson N.M., Likens G.E. and Feller M.C. (1988) The effects of acidic deposition on stream water chemistry: a comparison between Hubbard Brook, New Hampshire and Jamieson Creek, British Columbia. *Water Resources Research*, **24**, 195–200.

Driscoll C.T., Lawrence G.B., Bulger A.J., Butler T.J., Cronan C.S., Eagar C., Lambert K.F., Likens G.E., Stoddard J.L. and Weathers K.C. (2001) Acidic deposition in

the northeastern US: sources and inputs, ecosystems effects, and management strategies. *BioScience*, **51**, 180–198.

Driscoll C.T., Yan C., Schofield C.L., Munson R. and Holsapple J. (1994) The mercury cycle and fish in the Adirondack lakes. *Environmental Science and Technology*, **28**, 136A–143A.

Elwood J.W. (1991) Southeast overview. In *Acidic Deposition and Aquatic Ecosystems: Regional Case Studies*, Charles D.F. (Ed.), Springer-Verlag: New York, pp. 291–295.

Evans C.D., Cullen J.M., Alewell C., Kopacek J., Marchetto A., Moldan F., Prechtel A., Rogora M., Vesely J. and Wright R.F. (2001) Recovery from acidification in European surface waters. *Hydrology and Earth System Sciences*, **5**, 283–298.

Fenn M.E., Baron J.S., Allen E.B., Rueth H.M., Nydick K.R., Geiser L., Bowman W.D., Sickman J.O., Meixner T., Johnson D.W., *et al.* (2003) Ecological effects of nitrogen deposition in the western United States. *BioScience*, **53**, 404–420.

Ferrier R.C., Jenkins A., Wright R.F., Schopp W. and Barth H. (2001) Assessment of recovery of European surface waters from acidification 1970–2000: an introduction to the special issue. *Hydrology and Earth System Sciences*, **5**, 274–282.

Galloway J.N. (1995) Acid deposition: perspectives in time and space. *Water Air and Soil Pollution*, **85**, 15–24.

Gbondo-Tugbawa S.S., Driscoll C.T., Aber J.D. and Likens G.E. (2001) Validation of an integrated biogeochemical model (PnET-BGC) at a northern hardwood forest ecosystem. *Water Resources Research*, **37**, 1057–1070.

Gorham E. (1989) Atmospheric deposition to lakes and its ecological effects: a retrospective and prospective view of research. *Canadian Journal of Fisheries and Aquatic Sciences*, **1643**, 63–85.

Groffman P.M., Driscoll C.T., Likens G.E., Fahey T.J., Holmes R.T., Eagar C. and Aber J.D. (2004) Nor gloom of night: a new conceptual model for the Hubbard Brook Ecosystem Study. *BioScience*, **54**, 139–148.

Gunn J.M. and Mills K.H. (1998) The potential for restoration of acid-damaged lake trout lakes. *Restoration Ecology*, **6**, 390–397.

Hedin L.O., Granat L., Likens G.E., Buishand T.A., Galloway J.N., Butler T.J. and Rodhe H. (1994) Steep declines in atmospheric base cations in regions of Europe and North America. *Nature*, **367**, 351–354.

Horsley S.B., Long R.P., Bailey S.W. and Hall T.J. (2000) Factors associated with the decline of sugar maple on the Allegheny Plateau. *Canadian Journal of Forest Research*, **30**, 1365–1378.

Huntington T.G., Hooper R.P., Johnson C.E., Aulenbach B.T., Cappellato R. and Blum A.E. (2000) Calcium depletion in a southeastern United States forest ecosystem. *Soil Science Society of America Journal*, **64**, 1845–1858.

Jeffries D.S. (1991) Southeastern Canada: an overview of the effect of acidic deposition on aquatic resources. In *Acidic Deposition and Aquatic Ecosystems: Regional Case Studies*, Charles D.F. (Ed.), Springer-Verlag: New York, pp. 273–289.

Kirchner J.W. and Lydersen E. (1995) Base cation depletion and potential long-term acidification of Norwegian catchments. *Environmental Science and Technology*, **29**, 1953–1960.

Klimont Z., Cofala J., Schopp W., Amann M., Streets D.G., Ichikawa Y. and Fujita S. (2001) Projections of SO2, NOx, NH3 and VOC emissions in East Asia up to 2030. *Water Air and Soil Pollution*, **130**, 193–198.

Kretser W., Gallagher J. and Nicolette J. (1989) *Adirondack Lakes Study. 1984–1987. An Evaluation of Fish Communities and Water Chemistry*, Adirondacks Lakes Survey Corporation: Ray Brook.

Kurz D., Rihm B., Alveteg M. and Sverdrup H. (2001) Steady-state and dynamic assessment of forest soil acidification in Switzerland. *Water Air and Soil Pollution*, **130**, 1217–1222.

Laudon H. and Hemond H.F. (2002) Recovery of streams from episodic acidification in Northern Sweden. *Environmental Science and Technology*, **36**, 921–928.

Likens G.E. and Bormann F.H. (1995) *Biogeochemistry of a Forested Ecosystem*, *Second Edition*, Springer-Verlag: New York.

Likens G.E., Bormann F.H. and Johnson N.M. (1972) Acid rain. *Environment*, **14**, 33–40.

Likens G.E., Butler T.J. and Buso D.C. (2000) Long- and short-term changes in sulfate deposition: effects of the 1990 Clean Air Act Amendments. *Biogeochemistry*, **52**, 1–11.

Likens G.E., Driscoll C.T. and Buso D.C. (1996) Long-term effects of acid rain: response and recovery of a forested ecosystem. *Science*, **272**, 244–246.

Lui S., Munson R., Johnson D.W., Gherini S., Summers K., Hudson R., Wilkinson K. and Pitelka L.F. (1992) The nutrient cycling model (NuCM): overview and application. In *Atmospheric Deposition and Forest Nutrient Cycling: A Synthesis of the Integrated Forest Study*, Johnson D.W. and Lindberg S.E. (Eds.), Springer-Verlag: New York, pp. 583–609.

MacAvoy S.E. and Bulger A.J. (1995) Survival of brook trout (*Salvelinus fontinalis*) embryos and fry in streams of different acid sensitivity in Shenandoah National Park, USA. *Water Air and Soil Pollution*, **85**, 439–444.

Malm W.C., Sisler J.F., Huffman D., Eldred R.A. and Cahill T.A. (1994) Spatial and seasonal trends in particle concentration and optical extinction in the United States. *Journal of Geophysical Research*, **99**, 1347–1370.

Moore M.V., Pace M.L., Mather J.R., Murdoch P.S., Howarth R.W., Folt C.L., Chen C.Y., Hemond H.F., Flebbe P.A. and Driscoll C.T. (1997) Potential effects of climate change on freshwater ecosystems of the New England/Mid-Atlantic region. *Hydrological Processes*, **11**, 925–947.

National Acid Precipitation Assessment Program [NAPAP] (1998) *NAPAP Biennial Report to Congress. An Integrated Assessment*. National Acid Precipitation Program. Washington.

Nilsson J. and Grennfelt P. (1988) *Critical Loads for Sulphur and Nitrogen*, The Nordic Council of Ministers NORD: Copenhagen, p. 15.

Olem H., Schreiber R.K., Brocksen R.W. and Porcella D.B. (1991) *International Lake and Watershed Liming Practices*, The Terrene Institute: Washington.

Paerl H.W., Dennis R.L. and Whitall D.R. (2002) Atmospheric deposition of nitrogen: implications for nutrient over-enrichment of coastal waters. *Estuaries*, **25**, 677–693.

Posch M., de Vries W. and Hettelingh J.-P. (1995) Critical loads of sulphur and nitrogen. In *Calculation and mapping of critical thresholds in Europe: Status Report 1995*, Posch M.,

de Smet P.A.M., Hettelingh J.-P. and Downing R.J. (Eds.), National Institute of Public Health and the Environment: Bilthoven, pp. 31–41.

Posch M. and Hettelingh J.-P. (2001) From Critical Loads to Dynamic Modelling. In *Modelling and Mapping of Critical Thresholds in Europe, CCE Status Report 2001*, Posch M., de Smet P.A.M., Hettelingh J.-P. and Downing R.J. (Eds.), National Institute for Public Health and the Environment: Bilthoven.

Rodhe H., Langner J., Gallardo L. and Kjellstrom E. (1995) Global scale transport of acidifying pollutants. *Water Air and Soil Pollution*, **85**, 37–50.

Ruess J.O. and Johnson D.W. (1986) *Acidic Deposition and the Acidification of Soils and Waters*, *Ecological Studies, Vol. 59*, Springer-Verlag: New York.

Seip H.M., Dianwu Z., Jiling X., Dawei Z., Larssen T., Bohan L. and Vogt R.D. (1995) Acidic deposition and its effects in southwestern China. *Water Air and Soil Pollution*, **85**, 2301–2306.

Sherwood S.L. and Lipfert F. (1990) *Distribution of Materials Potentially at Risk from Acidic Deposition*, Report 21, National Acid Precipitation Assessment Program, Acidic Deposition: State of Science and Technology.

Stoddard J.L., Jeffries D.S., Lukewille A., Clair T.A., Dillon P.J., Driscoll C.T., Forsius M., Johannessen M., Kahl J.S., Kellogg J.H., *et al.* (1999) Regional trends in aquatic recovery from acidification in North America and Europe. *Nature*, **401**, 575–578.

Stoddard J.L., Kahl J.S., Deviney F.A., DeWalle D.R., Driscoll C.T., Herlihy A.T., Kellogg J.H., Murdoch P.S., Webb J.R. and Webster K.E. (2003) *Response of surface water chemistry to the Clean Air Act Amendments of 1990*, US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory: Research Triangle Park.

Stumm W. and Morgan J.J. (1996) *Aquatic Chemistry, Chemical Equilibria and Rates in Natural Waters*, *Third Edition*, John Wiley & Sons: New York, p. 1022.

Sullivan T.J. (1991) Long-term temporal trends in surface water chemistry. In *Acidic Deposition and Aquatic Ecosystems: Regional Case Studies*, Charles D.F. (Ed.), Springer-Verlag: New York., pp. 615–639.

Sundqvist G., Letell M. and Lidskog R. (2002) Science and policy in air pollution abatement strategies. *Environmental Science and Policy*, **230**, 1–10.

United States Environmental Protection Agency [US EPA] (2004) *2003 Progress Report. EPA Acid Rain Program*, EPA 430-R-04-009, Clean Air Markets Division, U.S. Environmental Protection Agency, (**`www.epa.gov/airmarkets`**).

Van Sickle J., Baker J.P., Simonin H.A., Baldigo B.P., Kretser W.A. and Sharpe W.F. (1996) Episodic acidification of small streams in the northeastern United States: fish mortality in field bioassays. *Ecological Applications*, **6**, 408–421.

Wright R.F. (2003) Predicting recovery of acidified freshwaters in Europe and Canada. *Hydrology and Earth System Sciences*, **7**, 429–430.

96: Nutrient Cycling

ANTHONY C EDWARDS[1] AND ROBERT G WETZEL[2†]

[1]Formerly of: Catchment Management Group, Macaulay Institute, Aberdeen, UK & Currently at: Nether Backhill, Aberdeen, UK
[2]Department of Environmental Sciences and Engineering, The University of North Carolina, Chapel Hill, NC, US
[†]Deceased April 18, 2005

Nutrient concentrations within aquatic ecosystems have increased markedly during the last fifty years due to a variety of causes. Significant changes to the rates of nutrient cycling, particularly of nitrogen and phosphorus, have influenced their biological availability on a global scale. Three broad groupings of processes that directly influence nutrient cycling within freshwaters are described. Retention modifies the availability of nutrients for transport, while selectivity between individual chemical species results in preferential uptake and/or transport, and finally transformation processes alter the physical state or reactivity of particular nutrients. Differences in spatial and temporal dynamics of delivery and transport mechanisms among individual nutrients have caused variable relative changes in nitrogen and phosphorus concentrations and fluxes. These composite dynamic factors make it difficult to couple causal relationships between nutrient sources and their impacts.

INTRODUCTION

Human activity has developed within the geographical constraints imposed by the physical landscape with the result that intensive agriculture and high urban densities tend to be concentrated toward the flatter, floodplain coastal regions. Where demand for land has been historically high, such as Western Europe, few lowland areas exist that have not been extensively exploited. These developments have often been accompanied by the straightening of watercourses and many other impacts, such as abstraction, on water resources (*see* **Chapter 91, Water Quality, Volume 3**; **Chapter 191, Environmental Flows: Managing Hydrological Environments, Volume 5**). In some situations, this exploitation has been so extreme that rivers that were once constantly flowing are now dry for extended periods or flow is maintained only through effluent discharges. Anthropogenic pollution, together with modifications of the environment and climate, are now so pervasive that no aquatic system of the biosphere can be considered as truly pristine or unaltered in some manner (*see* **Chapter 93, Effects of Human Activities on Water Quality, Volume 3**). A current estimate of the global riverine flux of total nitrogen is suggested to be double that of the preindustrial era (Green *et al.*, 2004). It is within this background of multiple and complex stresses that nutrient cycling in aquatic systems is now considered.

An improved prediction of the impact of nutrient enrichment on ecosystem community structure and function requires an understanding of the mechanisms that are responsible for nutrient enrichment and subsequent cycling (*see* **Chapter 90, Lake Sediments as Records of Past Catchment Response, Volume 2**). Increased concentration and flux of individual nutrients have been most apparent during the last five decades although it is likely that systems were suffering from nutrient enrichment well before this (de Jonge *et al.*, 2002). For example, paleolimnological studies of a closed lake of volcanic origin in Italy demonstrated the marked effects of alterations to the drainage basin area by the construction of a road alongside the lake in early historical times (Hutchinson *et al.*, 1970). Increased eutrophication in the sedimentary records coincided precisely with the major disturbance of the drainage basin by the construction of a Roman road, the Via Cassia, about 171 B.C. when a large increase in inputs of phosphorus (P) and alkaline earths, particularly calcium, resulted and the lake

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Examples of biotic or abiotic processes that influence nutrient availability and transport within terrestrial and aquatic systems

	Terrestrial systems		Aquatic systems	
	Biotic	Abiotic	Biotic	Abiotic
Retention	Immobilization of N and P in organic matter	Sorption processes involving both organic and inorganic surfaces	Microbial retention totally dominates	Deposition of suspended material onto floodplain
Selectivity	Preferential plant uptake of particular solutes	Particle-size separation during erosion events	Increasing recalcitrance of DOM by selective utilization of labile compounds	Preferential exchange of NH_4^+ and PO_4^{3-} by stream bed sediment
Transformation	Nitrification and the generation of the mobile nitrate anion	Increased crystalinity of amorphous Fe/Al minerals during drying cycles	Denitrification and utilization of inorganic N Turnover rates are inversely proportional to organism size	Redox cycles involving Fe/Mn minerals and release/uptake of associated P under reducing conditions

rapidly became eutrophic, causing the deposition of much organic matter.

Anthropogenic sources of nutrients vary widely in their composition and mode of delivery to aquatic systems. Actual responses will depend greatly upon the type and individual properties of receiving waters. Direct inputs of industrial and domestic effluent (*see* **Chapter 94, Point and NonPoint Source Pollution, Volume 3**) contribute significant quantities of both nitrogen (N) and P and continue to be responsible for many local and regional-scale impacts (e.g. Smith *et al.*, 2003). Many intensive agricultural production systems are currently experiencing a surplus of nutrients as a consequence of excessive use of fertilizers and the inefficient utilization and recycling of the resulting animal waste products (Steele, 1995; Smaling *et al.*, 1999). Even remote "pristine" environments do not escape being impacted by the growing global atmospheric N cycle and the possibility of long-range N transport thought to be responsible for the nutrient imbalances implicated by the N saturation hypothesis (Galloway, 1998; Galloway and Cowling, 2002). Much of the recent increase in atmospheric N is directly attributable to emissions from automobiles and animal waste products which means that nutrient enrichment and the resulting environmental impacts are a shared human responsibility (*see* **Chapter 184, Global River Carbon Biogeochemistry, Volume 5**).

Various physically, chemically, and biologically mediated processes act to modify profoundly nutrient concentrations and fluxes. Establishing the combined significance of these processes hampers attempts to understand the fundamentals of "nutrient cycling" or measure the extent of impact that various human driven perturbations have on aquatic systems. Ecosystem level expression of these processes occurs through three quantifiable responses. The first involves *retention* which influences the availability of nutrients for transport; second is the *selectivity* shown by individual process for a particular chemical species leading to preferential uptake and/or transport; and finally there are *transformation* processes that result in a physical change in either state or reactivity of a particular chemical species. These responses will be referred to in the sections that follow and together contribute to a sequence of interlinked mechanisms that operate to buffer downstream nutrient losses by introducing selective and seasonal lag times. Table 1 provides examples of processes that may operate within terrestrial and aquatic systems.

Although nutrient retention is virtually always only temporary, timescales can vary from minutes to thousands of years depending upon local circumstances. The net result is a damping of system response and a modification of the timing, chemical form, and quantity of chemicals available either for transport or uptake.

NUTRIENT CYCLING WITHIN TERRESTRIAL AND AQUATIC SYSTEMS

Here nutrient cycling is considered in a broad context that recognizes the importance of a "terrestrial component" as the ultimate source for most of the water and many of the nutrients that enter surface and groundwater. Meyer *et al.* (1988) suggested three reasons a greater understanding of elemental cycling within streams could help improve understanding of ecosystem behavior. These are as follows:

1. nutrients regulate ecological processes in streams, especially when a nutrient is limiting;
2. nutrients link terrestrial and aquatic ecosystems; and
3. stream processes alter the timing, magnitude, and form of elemental fluxes and thereby nutrient availability to downstream communities.

Terrestrial and aquatic ecosystems are therefore linked through the essentially unidirectional movement of water and substances that occur in response to hydrological events across the drainage basin. Strong relationships between particular attributes of terrestrial ecosystems (such as land cover) and the chemical composition of drainage water have been regularly reported. As the downgradient receivers of terrestrial land drainage, most aquatic systems are especially sensitive and therefore highly responsive to changes that affect either runoff quality or quantity. For this reason, water that drains from terrestrial land often provides a useful indicator of perturbation. Regular sampling and analysis can usefully be employed as a diagnostic integrator and indicator of wider ecosystem damage (e.g. Aber *et al.*, 2003).

The routing and residence time of water within individual system components are influenced by geomorphological features of the landscape. The proportion of incident precipitation that ultimately contributes to surface or groundwater flow varies widely among ecosystems and also seasonally. For example, only about one-third of the incoming precipitation of many forested drainage basins leaves as runoff and groundwater seepage to streams and rivers during active growth seasons (*see* **Chapter 186, Water and Forests, Volume 5**). Changes to land management such as the removal of existing vegetation, greatly increase the amount and accelerate the flow rate of runoff (e.g. Likens and Bormann, 1995). Vegetation type and soil composition influence not only the amount of runoff but also the composition and quantity of allochthonous organic matter and nutrients that enter streams and lakes. The duration of contact with soil and associated microbiota influence the content of dissolved salts and organic products conveyed in the water.

Although it might be considered that once in the drainage network most nutrients move unidirectionally to the nearest large lake or ocean, this is not always necessarily the case. The composition of surface waters reflects the influence of terrestrial and in-stream processes and therefore represents the net balance between processes that either consume or generate a particular solute. For example, inorganic forms of N may be utilized as a substrate for in-stream denitrification. A recent example from Hubbard Brook demonstrated this situation, where in-stream demand for nitrate appeared to outstrip nitrification, resulting in a reduction of downstream nitrate export (Bernhardt *et al.*, 2002). During periods of high precipitation, runoff volumes may in some cases exceed the storage capacity of river channels resulting in the inundation of floodplains, often for long periods of time. As a consequence of the reduction of energy caused by overspilling the channel, suspended solids (SS), organic matter, and many associated nutrients can be returned to the flooded land and result in highly productive adjacent habitats (*see* **Chapter 84, Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands, Volume 2**). Channel engineering and progressive urbanization of the river plain increases the risk of flooding, while growth of macrophytes can reduce in-stream flows resulting in localized sedimentation (e.g. Haslam, 1978; Dodds and Biggs, 2002). Rivers act as effective redistributors of nutrients and sediment from upstream to downstream locations which may include standing waters.

Because of their geographical situation, estuaries are considered to be especially sensitive to nutrient enrichment. Conley (2000) recently reviewed data for a geographically wide range of riverine inputs to estuaries and highlighted a number of underlying issues that relate to nutrient cycling. It was demonstrated that N and P loading has increased by 6–50 and 18–180 times, respectively, compared to "pristine" conditions. Interestingly, this enrichment demonstrates that not only have quantities increased over time but because the rate of change has differed between nutrients, the ratio of N to P delivered to these estuaries has narrowed over time which may have important implications for the timing and nature of nutrient limitations. This conclusion is supported by other investigations, which have also shown that chemical composition and bioavailability has also altered. Under "pristine" conditions, it has been argued, for example, that a selective preference and utilization of inorganic N forms will result in their depletion and a relative increase in the leaching of organic forms of N for primary forests in temperate South America (Hedin *et al.*, 1998; Perakis and Hedin, 2002). Organic forms of both N and P can also contribute significantly to total losses derived from agriculturally managed systems (Howarth *et al.*, 1996).

PHOSPHORUS AND NITROGEN CYCLES IN FRESHWATER ECOSYSTEMS

Ecological interest in P stems from its major role in biological metabolism and the relatively small amounts of P available in the hydrosphere. In comparison to other major nutritional and structural components required by biota (carbon, hydrogen, nitrogen, oxygen, and sulfur), P is the least abundant, and most commonly is the first element to limit biological productivity of fresh waters. Nitrogen is also a major constituent of cellular protoplasm of organisms and a major nutrient that affects and sometimes limits the productivity of fresh waters.

Nutrients enter the food chain initially through uptake (autotrophs or heterotrophs), the efficiency of which requires the three conditions:

1. they should be present at concentrations that are sustained during biologically active uptake periods above some ecological threshold value (usually defined by a membrane transport system);
2. they should be in a labile (readily bioavailable) chemical form;
3. they should be present in pools which are physically accessible to the biota.

For many natural ecosystems, one or both of these conditions (together with many other physical/chemical limiting factors) are not met and result in suboptimal conditions for growth. The global nutrient enrichment of aquatic ecosystems means that one type of limitation that has restricted productivity levels below the system potential maximum, in many cases, has been removed. For example, the role of elemental imbalances that occur between trophic levels has been reviewed recently (Frost *et al.*, 2002) for benthic systems their potential to affect population dynamics, trophic interactions, and gross transfer efficiencies has been demonstrated.

Types of Phosphorus

Much quantitative data exist on the seasonal and spatial distribution of P in streams and lakes, as well as the loading rates to recipient waters from drainage basins (Wetzel, 2001). Orthophosphate (PO_4^{3-}) is the only directly utilizable form of soluble inorganic P. Phosphate is extremely reactive and interacts with many cations (e.g. Fe and Ca) to form, especially under oxidizing conditions, relatively insoluble compounds that precipitate out of the water. The immediate bioavailability of phosphate is also reduced by its adsorption to inorganic colloids and particulate compounds (e.g. clays, carbonates, and hydroxides).

Much of the organic P resides in living and senescent plant organic matter or is associated with organic compounds immobilized with soil particles (Wetzel, 1999). Mineralization rates are masked by this large pool, from which a small but critical quantity of labile organic P is rapidly recycled among the soil, plants, and microbes. The organic P transformations are controlled by a combination of P concentrations in solution and biological activities, the most important of which are microbial alterations of redox and bonding to particles. As the concentrations or biological availability of the inorganic pool decreases, much of the P resides in organic compounds and polyphosphates, usually resulting in an increase in the relative rate of recycling between the inorganic and organic P pools.

In aquatic ecosystems, P availability depends to a great extent on its delivery from the drainage basin. Both physical and biotic processes in lakes, streams, and wetlands tend toward concentration of P in inorganic and organic pools of the sediments. Exchange rates between various deposits of P and the interstitial water of the sediments depend on local adsorption and diffusion coefficients and their alteration by enzyme-mediated reactions of the microbiota (Wetzel, 2001). In organic-rich sediments not illuminated by sunlight, high sediment demand for oxygen results from microbial respiratory metabolism, which may be exacerbated by slow rates of diffusion from the overlying water and inorganic elements such as Fe(II) that accumulate in reduced form. On a molar ratio of P liberated per unit of organic carbon mineralized by bacteria, the availability of alternate electron acceptor compounds such as sulfate largely control mineralization rates and release of sorbed and bonded P from sediments.

Algal photosynthesis on sediment surfaces, as well as oxygen release from the roots of emergent aquatic plants of wetlands into the rhizosphere, markedly alters redox conditions of the sediments on a distinct diurnal basis (Roden and Wetzel, 1996; Wetzel, 1999, 2001; Reddy *et al.*, 2005). These rapid (minutes) shifts in the depth of the oxidized microzone, by as much as a factor of 100, and redox conditions in the sediments, in turn, form a highly dynamic sediment system that rapidly alter both mineralization rates of organic P pools and inactivation of inorganic P by oxidative reactions (Carlton and Wetzel, 1988). Of the several fractions of organic P, a low molecular weight fraction cycles rapidly (hours) with soluble forms. Dissolved and colloidal organic P fractions released from microbiota are cycled much more slowly, particularly if sorbed to inorganic and detrital organic particles.

Distribution and Balance of Phosphorus

The range of total P in fresh waters is large, from $<5\,\mu g\,L^{-1}$ in very unproductive (oligotrophic) waters to $>100\,\mu g\,L^{-1}$ in highly eutrophic waters. Most uncontaminated fresh waters contain 10–50 μg total P/L. Concentrations of total and soluble P tend to increase markedly in the lower strata (hypolimnia) of productive lakes that are thermally stratified, largely from the release of soluble P from the stores of accumulated sediments.

Dissolved P concentrations in rivers are generally higher than in lakes, and often increase markedly following rainfall and snowmelt events on land in the drainage basin. At low concentrations in streams, P can be limiting, and microbial productivity often increases with greater loading of P. In addition to biological uptake, P abundance is influenced also by abiotic adsorption and desorption reactions with organic and inorganic particles. Where dissolved organic P compounds are less chemically active they can be exported downstream for considerable distances. Most recycling of P (>95%) is associated with microbiota attached to particles.

The P content of groundwater is generally low (average ~20 μg P/L) even in areas where soils contain relatively

large quantities of P. The low P content is a result of the relatively insoluble nature of phosphate-containing minerals and efficient scavenging of soluble P by biota and soil particles. However, certain situations can arise where groundwater experiences high concentrations of P as a consequence of localized contamination arising from effluents derived from septic systems or agricultural feedlots (Maule and Fonstad, 2000).

P uptake and kinetics in numerous temperate lakes show that turnover times are extremely rapid (5–100 min) during summer periods of high demand and low loading inputs, and efficient uptake systems can reduce P concentrations to well below analytical detection limits (Hudson *et al.*, 2000). During winter periods, turnover rates slow by as much as 2 orders of magnitude. P turnover rates are generally faster under more oligotrophic conditions of lower P availability. Sedimentation of particles results in constant losses of P from the upper strata of water (*see* **Chapter 88, Reservoir Sedimentation, Volume 2**). As a result, fresh supplies of P must enter the ecosystems in order to maintain or increase productivity.

Phosphorus enters fresh waters from atmospheric precipitation and from groundwater and surface runoff. Small amounts of P in atmospheric precipitation and particulate material fallout originates from fine particles of soil and rock, particularly from agriculture, from living, and dead organisms (particularly as volatile compounds released from plants), and from natural fires and the burning of fossil fuels. The loading rates of phosphorus in surface waters vary greatly with patterns of land use, geology and morphology of the drainage basin, soil productivity, human activities, pollution, and other factors (Table 2). When P is added to unproductive fresh waters, either experimentally or as a result of human activities, a common response is a rapid (days) increase in algal and bacterial productivity. Because of numerous sites of rapid sequestering and losses, inputs must be more or less continuous, however, in order to maintain a higher level of productivity. For this reason priority is often given to the management of P inputs.

Numerous mass balance models have been developed to predict, on the basis of P loading and hydraulic retention times, the anticipated responses of algal biomass and less commonly productivity. The models yield a reasonably accurate estimate of permissible P loading needed to achieve a certain level of lake productivity if loading is reduced. In certain shallow lakes with greater than average turbulence, large littoral areas, and small anaerobic hypolimnia, reduced productivity does not always occur as rapidly in response to decreased P loading as predicted from the models. In these lakes, P release from sediment sources ("internal loading") is much greater than common values (10–30% of total loading) for deeper, more intensely thermally stratified lakes.

Types of Nitrogen

The nitrogen cycle is a complex biochemical process in which N in various forms is altered by nitrogen fixation, assimilation, and reduction of nitrate to N_2 by denitrification. The nitrogen cycle of fresh waters is essentially microbial. Dominant forms of nitrogen in fresh waters include (i) dissolved molecular N_2 and N_2O, (ii) ammonium nitrogen (NH_4^+), (iii) nitrite (NO_2^-), (iv) nitrate (NO_3^-), and (v) various organic compounds, such as amino acids, amines, nucleotides, proteins, and humic compounds of modest N content.

Table 2 General ranges of primary productivity of phytoplankton and related characteristics of lakes of different trophic categories (From Wetzel (2001) after numerous sources cited therein)

Drainage basin type	Total dissolved inorganic nitrogen losses (kg N km^{-2} $year^{-1}$)	Total phosphorus losses (kg P km^{-2} $year^{-1}$)
Undisturbed temperate forest	ca. 200	ca. 2
Undisturbed boreal forest	90–160	3–9
Cleared forest, igneous watershed	–	ca. 5
Cleared forest, sedimentary watershed	–	ca. 11
Cleared forest, volcanic watershed	–	72
Pasture, low intensity	100–1000	8–20
Pasture, high intensity	2000–25 000	–
Arable, cereals	4000–6000	–
Arable, cash crops	4000–<10 000	–
Arable, intensive (UK, USA, Netherlands)	–	7–190
Mixed upland (Northern UK)	530–630	ca. 34
Groundwater leachates	427–638	35–93
Urban runoff	ca. 1000	ca. 100
General soil productivity		
Low	<500	<20
Medium	500–2500	20–50
High	>2500	>50

Distribution and Balance of Nitrogen

The nitrogen cycle consists of a balance between N inputs to and nitrogen losses from aquatic ecosystems. Sources of N include (i) N contained in particulate "dry fallout" and in precipitation, (ii) N_2 fixation both in the water and the sediments, and (iii) inputs of N from surface and groundwater drainage. Reduced forms of N are a major constituent of the effluent from sewage treatment works and livestock wastes. Biological fixation of N_2 by soil bacteria can be a major source of N. In lakes and streams, N_2 fixation by heterotrophic bacteria and certain cyanobacteria is quantitatively less significant, except under certain conditions of severe depletion of combined inorganic N compounds. N_2 fixation by cyanobacteria is usually much greater than fixation by heterotrophic bacteria, and in particular, in wetlands surrounding or adjacent to surface waters can add significant amounts of combined N to freshwater ecosystems.

Ammonia is generated by heterotrophic bacteria as the primary nitrogenous end product of decomposition of proteins and other nitrogenous organic compounds. Ammonia concentrations are usually low as it is readily assimilated by plants, strongly retained by sediment, and oxidized in nitrification to NO_2^- and NO_3^-. Under anaerobic conditions of productive waters, as in the lower strata of a thermally stratified lake, large concentrations of NH_4–N can occur. Nitrate is assimilated and aminated into organic nitrogenous compounds within organisms. During metabolism of these organisms, and at death, much of the N is liberated as ammonia or organic nitrogenous compounds that are degraded to ammonia and related compounds.

Dissolved organic nitrogen (DON), largely as complex amino nitrogen compounds in polypeptides and humic substances, often constitutes over 50% of the total soluble N in fresh waters having both autochthonous and allochthonous origins (Egeberg *et al.*, 1999). Organic carbon-to-nitrogen ratios (C : N) indicate an approximate state of resistance of complex mixtures of organic compounds to decomposition, because proteolytic metabolism by fungi and bacteria removes proportionally more N than C. Wider C : N ratios commonly occur in residual organic compounds which are more resistant to decomposition. Organic materials from terrestrial and wetland sources commonly have C : N ratios from 45 : 1 to 50 : 1, and contain many humic compounds from plant structural tissues of low N content. Changes in the protein-carbohydrate-lipid ratios alter the particulate C : N ratio as a result of P or N limitation.

Increased loading of both inorganic and organic N to rivers and lakes frequently results from agricultural activities, effluent from sewage treatment works (STW), intensive livestock enterprises, and anthropogenic atmospheric pollution (Table 2). In unproductive oligotrophic lakes, P availability is often the principal limitation nutrient for plant growth. As P loading to fresh waters increases and they become more productive, P requirements become saturated with high P availability. Nitrogen availability then becomes more important as a limiting nutrient (Maberly *et al.*, 2003).

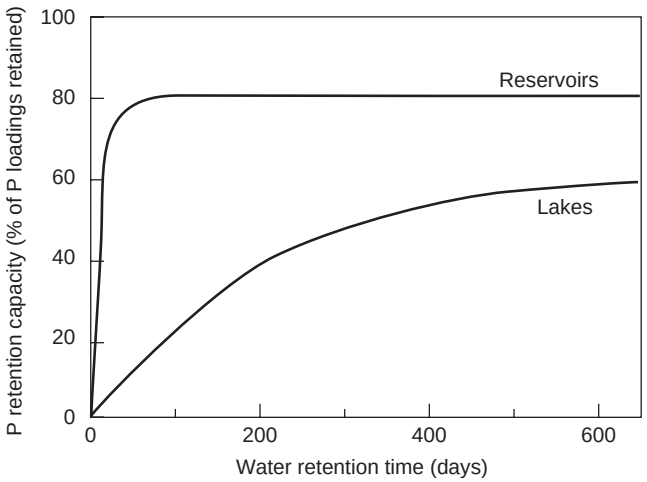


Figure 1 Phosphorus retention capacity of lakes and reservoirs as a percentage of total phosphorus loading retained in relation to water retention times. (Modified from Straškraba, 1996)

Retention capacities for nutrients can differ considerably between stratified natural lakes and reservoirs (*see* **Chapter 109, Reservoirs, Volume 3**), largely as a function of marked differences in water retention times (e.g. Figure 1). Although P retention times are similar when comparing water retention times of a year or longer, the water retention times of lakes are much longer (1–7 years on the average) than those of reservoirs (often <100 days and highly variable).

In running waters, N is used repeatedly and released back to the water as it passes downstream. The rate of utilization and release depends upon physical and biological retentiveness, largely by the microbiota attached to the streambed. The average distance a nutrient molecule travels downstream during one cycle through the water, biota, and substrata compartments is the *spiraling length*. Nutrient limitations to biotic productivity are not common in small rivers and streams, where nutrients are efficiently retained and recycled. As rivers become larger, nutrient retention is less and nutrient limitations can become less significant.

The processes of N cycling in the water of streams and rivers are similar to those of lakes and are influenced to a large extent by bacterial, fungal, and other microbial metabolism. Once nitrogen is loaded to a stream or river, the attached bacteria, fungi, and algae are the primary organisms controlling the spatial and temporal variations within the water. The processes of nitrification and denitrification often function simultaneously and reciprocally in running water sediments, where many microzones of steep redox

gradients occur in the hyporheic zone of the streambed. The highly dynamic longitudinal pattern of N_2O outgassing from streams draining agricultural areas has been demonstrated by Reay *et al.* (2003).

RELATIONSHIP BETWEEN DRAINAGE BASIN FEATURES AND WATER CHEMISTRY

Land Cover and Terrestrial Process

Concentrations and the resulting flux of dissolved and particulate material vary greatly both spatially between locations and temporally at individual sites (Johnes and Butterfield, 2002). While differences in in-stream utilization can account for some of the observed variability, much can be traced backward and explained by differences in the geology, soils, vegetation cover, and land use existing within the contributing terrestrial catchment area (*see* **Chapter 115, Landscape Element Contributions to Storm Runoff, Volume 3**; **Chapter 202, Use of Climate Information in Water Resources Management, Volume 5**). The significance of any single attribute may well vary seasonally. A study by Johnson and Gage (1997), for example, demonstrated that drainage basin factors accounted for much of the observed variation in total dissolved solids; however, during autumn, geological factors and shared influence of geology/landscape structure plus land use exerted more influence than with land use only. It is reasonable to expect that a relationship should exist between certain catchment attributes (through their combined influence upon key transformation processes), the composition of drainage water, and consequently the structure and functioning of aquatic systems they maintain.

The important role which transformation processes have upon nutrient retention through their influence upon the chemical form and fluxes of N is demonstrated in Table 3. In this example an upland seminatural, highly "organic" soil is compared to a lowland "mineral" agricultural soil. Although each has a similar soil N capital of $\sim$5000 t ha^{-1}, which in the case of the organic soil has collected over thousands of years, these two situations represent the extreme of conditions. The principle factor that ultimately governs N composition of drainage waters is the balance, which exists among immobilization, mineralization, and nitrification processes. Although both terrestrial systems appear to display a net retention of N, the much greater inputs and associated losses of nitrate that occur from the agricultural site results from the dominance of nitrification.

It is possible to develop these general relationships in two ways. The first considers the attributes of the total hydrologically contributing land area, while the second approach requires an increasing level of sophistication where relationships are developed using a greater awareness of spatial partitioning. The first situation using the "whole" contributing area assumes that areas having a similar range of attributes all contribute equally, irrespective of their distance from the water course, whereas the latter approach implies some degree of spatial weighting. Johnson and Gage (1997) compared the strength of relationships for individual determinants obtained with either the whole drainage basin area or a spatially weighted definition, such as just the attributes of a 100-m buffer strip immediately adjacent to the river. While river nitrate concentrations were better related to the whole catchment, total phosphorus (TP) and SS were best described during the summer with land use in the buffer zone. Observations of this nature elucidate how catchments operate while also demonstrating the independent and temporal nature of transport mechanisms for individual substances.

The Terrestrial–Aquatic Interface

It has been reasonably argued that the area immediately adjacent to a river, which encompasses both riparian and hyporheic zones, represents the interface and site of connectivity between terrestrial and aquatic systems. The transient interface between surface and groundwater is an important point linking lateral nutrient fluxes

Table 3 A comparison of N capital, transformation processes, and fluxes of N between extensively managed seminatural upland and intensive low land agricultural systems

Attribute	Upland seminatural	Lowland agriculture
Soil N capital and chemical form	$\sim$5000 t ha^{-1}, reduced (organic and ammonium)	$\sim$5000 t ha^{-1}, reduced (organic and ammonium) and nitrate
Sources of N input	Atmosphere, biological fixation	Atmosphere, biological fixation, fertilizer/manure
Dominant transformation processes	Immobilization	Mineralization and nitrification
Stream water composition	Organic N and ammonium $<$1 mg N L^{-1}	Nitrate $\sim$7 mg N L^{-1}
Annual N flux	$\sim$2 kg N ha^{-1}	$>$25 kg N ha^{-1}

from terrestrial systems and the longitudinal processes in streams (Dahm *et al.*, 1998). A constant exchange of water and solutes occurs between the hyporheic zone and the river (*see* **Chapter 113, Hyporheic Exchange Flows, Volume 3**). Typical lengths of hyporheic zone range from centimeters to meters (Harvey and Wagner, 2000) and where this interaction leads to an improved supply of substrate, this zone contributes significantly to in-stream metabolism and recycling. Important processes include redox cycles and denitrification. Recent reviews have summarized the importance of the biological and physicochemical processes (Pusch *et al.*, 1998; Wetzel, 2001) operating within it that can retain and transform (Table 1) nutrients, ultimately modifying delivery. For these reasons, this zone, in particular, receives special attention often being the focus of attempts to "manage" nutrient transport from the wider landscape. Hedin *et al.* (1998) highlight the importance of the soil–stream interface as an area with high potential for removing dissolved N by denitrification and suggest that the process is limited by the availability of oxidizable carbon via shallow flow paths. Support for management policies that concentrate on "this last line of defense" in the protection of water quality is debatable, although the critical role of the riparian corridor to overall river health and function is undeniable.

NUTRIENT TRANSPORT AND DELIVERY TO AQUATIC ECOSYSTEMS

Sources of Nutrients

Convention usually distinguishes between "point" and "diffuse" nutrient sources (*see* **Chapter 94, Point and Non-Point Source Pollution, Volume 3**). While this distinction may well represent an oversimplification of the actual situation, it does enable some basic differences between the attributes of the dominant delivery mechanisms to be made (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**; **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**). Classical examples of point sources (such as STW effluent) tend to have a constant and concentrated composition relative to the receiving waters with a semicontinuous delivery that may be independent of catchment hydrological processes. Importantly, the delivery of nutrients from point source origins can completely overwhelm naturally occurring short-term or seasonally associated dynamics of in-stream N and P cycles. In contrast to point sources, which are comparatively easy to locate and quantify accurately, diffuse sources usually have poorly defined and variable points of delivery. These basic differences are now considered for diffuse and point sources separately, and various specific case study examples can be found in Haygarth and Jarvis (2002).

Diffuse Sources

The significance of waterborne transport means that the physical linkage or degree of "hydrological connectivity" that exists between terrestrial and aquatic systems is an important feature that helps define nutrient delivery from diffuse sources. River composition and downstream nutrient transport vary considerably between steady and storm flow hydrological conditions. The former is most likely to be dominated by water originating from deeper soil horizons and groundwater, while the latter will be typified by episodic, high-energy events reflecting the increasing proportion of runoff derived from near-surface soil horizons and overland flow. Some of the implications arising from these differences between these two states are summarized in Table 4.

Point Sources

Despite ongoing and extensive regulation, nutrient inputs from point sources continue to have pronounced impacts on aquatic ecosystems (Haggard *et al.*, 2001). The often high concentrations of reactive P in STW effluent usually results in an intense zone of physicochemical and biological activity, tending to favor adsorption downstream of point sources (*see* **Chapter 97, Urban Water Quality, Volume 3**). This situation has been well demonstrated by Owens and Walling (2002) where the TP content of SS varied from $<2000\,\mu g\,g^{-1}$ for headwater areas located upstream of the main urban and industrialized regions to concentrations in excess of $7000\,\mu g\,g^{-1}$ at downstream sites. Phosphorus inputs from point sources may also widen the inorganic to organic P ratio from <2 in the headwaters to >4 in the lower reaches.

Regional-scale Relationship Between Point and Diffuse Sources

An example of how P export from a catchment having both urban and rural land uses can be separated on the basis of source, transport, and potential availability has been summarized by Dorioz *et al.* (1998). Export regimes were defined using a unique set of hydrological conditions (Table 5), TP inputs, in-stream TP retention, and bioavailability. The first hydrological condition experienced an extended period of stable and low flows where surface runoff was negligible and nutrient supply dominated by those originating from point sources. The overall system balance was toward nutrient accumulation through sedimentation and the selective adsorption of P. Conditions 2 and 3 (Table 5) are typified by increasing amounts of rainfall occurring against a background of stable low flows, but some overland flow may be generated especially from low-permeability urban areas. The system balance moved

Table 4 A comparison of the changing conditions likely to affect nutrient cycling during periods of steady and changing flow conditions

Attribute	Reducing/steady river flow	Increasing/event based river flow
Proportion of time	Long duration giving rise to regular flow conditions	Episodic, short-term dynamic flow events
Energy conditions	Low–medium	High
Hydrological connectivity/routing	Deeper longer residence time soil drainage and groundwater	Increasing proportion of rapid near-surface/overland flow
Nutrient transport	Primarily as soluble forms	Tendency for dominance of particulate associated forms
Seasonal occurrence	Summer and winter periods where soils are frozen	Spring and autumn
Biological production potential	High, as a result of clear water and greater light penetration during summer months	Reduced, as a result of poor light transmission and physical damage (e.g. abrasion) and low temperatures
In-stream processes and the balance between export or import conditions	Sediment deposition, nutrient retention, shorter spiraling distances, biotic uptake, and transformations	Sediment resuspension, nutrient export, longer spiraling distances
Drainage water composition	Solute dominated with composition being reflective of geochemical and in-stream biological processes	Particulate dominated with composition being more reflective of terrestrial biogeochemical processes

Table 5 Four reoccurring hydrological conditions in relation to P transport and catchment export dynamics (Dorioz *et al.*, 1998)

	Flow conditions	Surface runoff	Dominant P source	P form	System P balance
1	LF, constant, or decreasing flow for >7 days	None	Point sources	Soluble & bioavailable	In > Ex
2	LF, small storm event, so that 7-day flow greater than preceding weekly average	Urban	Point sources, and in-stream stores	Particulate & highly bioavailable	Ex > In
3	HF, substantial rainfall follows at least 7-day period of constant or decreasing flow	Urban	Diffuse urban sources and all stored P is exported	Highly bioavailable, high conc., and fluxes	Ex > In
4	HF, prolonged periods of rainfall	Urban and agricultural land	Diffuse agricultural sources	Conc. are low but fluxes high	In = Ex

Note: LF – low flow, HF – high flow, In – input, Ex – Export.

toward a net export of the previously accumulated in-stream nutrients. The final condition occurred where rainfall was prolonged and stream flow was increased to the extent where there was little possibility of retention and imports essentially equaled exports, but some storage might be possible if the system floods and sedimentation occurred over the floodplain.

Contrasting the Sources and Behavior of N and P: Catchment-scale Implications

The previous sections have highlighted differences between nutrient sources and some of the implications these might have for their delivery to and retention within river systems (*see* **Chapter 69, Solute Transport in Soil at the Core and Field Scale, Volume 2**). The widescale and well-documented increase in surface and groundwater nitrate concentrations provides an excellent example with which to further develop this theme and demonstrate the strong linkage that can develop between the products of "terrestrial" process and the composition of surface waters (*see* **Chapter 188, Land Use and Water Quality, Volume 5**). Transformation of DON via NH_4^+ to NO_3^- by mineralization and nitrification processes (Table 1) is sensitive to certain soil properties that include pH and texture

in addition to a constant supply of N and C substrates. Particularly favorable conditions for nitrification arise after soil cultivation, which combined with nutrient amendments and periodic liming means that there is often a strong relationship between the area of cultivated agricultural land and NO_3^- concentration. This situation is demonstrated in Table 6, which shows the strong positive relationship that can exist between averaged NO_3^- concentrations and the proportion of agricultural land (which ranged from 0 to >90% of the total catchment area) for 57 rivers located throughout Scotland and sampled at their lowest nontidal point. Phosphate (and NH_4^+) concentrations correlated much better to human population densities (which ranged from essentially <1 to >1000 capita/km^2).

The actual data are plotted in Figure 2 where some of the scatter for NO_3^- may be attributable to and explained by the very general nature of the land cover classification (arable plus all grassland). The relationship between PO_4^{3-} and agricultural land was poor (Table 6) but much better correlated to the census derived population density within each catchment, which suggests a predominantly point (assuming a reticulated STW system) source for P (Figure 2b). Together, these catchments account for over half the land area and half the population of Scotland. The N : P ratio (NO_3^--N and PO_4^{3-}-P) varied widely, but only at population densities >100 people/km^2 did it fall below 10; below this population density, the ratio was always greater than 10, and on occasions reached 100.

Transport of N and P

The nutrient flux that occurs from terrestrial ecosystems displays a high degree of spatial and temporal variability with the principle transport mechanism differing fundamentally between N and P (*see* **Chapter 89, On the Worldwide Riverine Transport of Sediment – Associated Contaminants to the Ocean, Volume 2**). The short-term retention and therefore capacity for downstream transportation of nutrients is the product of the attributes of delivery (chemical form, quantity, and timing) coupled with those related to the efficiency of various in-stream processes. Nutrient dependent limitations on productivity develop when biological demand exceeds local rates of supply. The annual cycle of increasing productivity often coincides with conditions of reduced summer discharge, and three properties of the local environment become central to the continued resupply of nutrients:

1. the buffering capacity and degree to which the immediate local environment has become saturated for individual nutrient species;
2. the quantity and biological availability of nutrients supplied from "external" sources; and
3. the local efficiency with which nutrients are being recycled.

Nutrients are transported either fully dissolved or in some association with SS (*see* **Chapter 140, Transport of Sediments, Volume 4**). The extent of partitioning between

Table 6 Matrix of correlation coefficients (r) between various nutrient and SS concentrations for 57 river water samples collected throughout Scotland on a monthly basis and averaged over a single year. (Samples were collected and analyzed by the Scottish Environment Protection Agency)

	Agricultural land use	Population density
SS	0.46***	0.34*
NH_4^+	0.15^{ns}	0.52***
NO_2^-	0.23^{ns}	0.48***
NO_3^-	0.77***	0.31*
PO_4^{3-}	0.28*	0.61***

Note: ns – not significant, $*P < 0.05$, $***P < 0.001$.

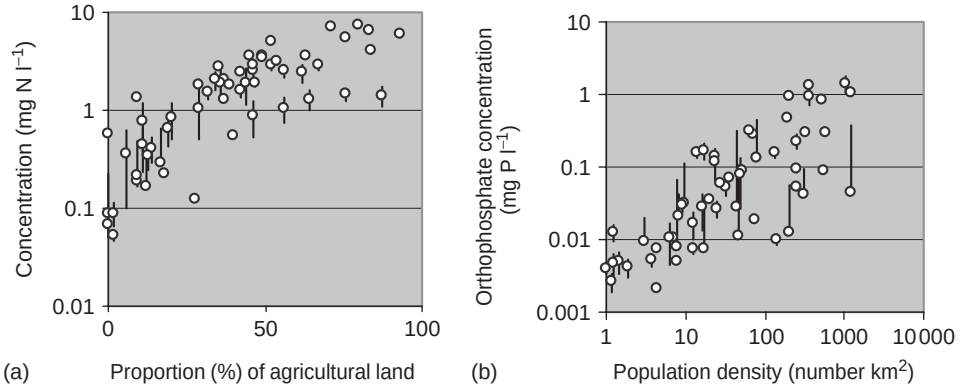


Figure 2 Annually averaged (a) NO_3^--N and (b) PO_4^{3-}-P concentrations (with SD) for samples collected at approximately monthly time intervals over a single year by the Scottish Environment Protection Agency and plotted against the proportion of agricultural land (a) or population density (b) in each catchment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

these two phases is a function of the solute involved, physicochemical properties of the sediment, and the prevailing environmental conditions. Physical form has a direct implication for nutrient cycling through its influence upon "accessibility" for uptake, which occurs primarily from solution, whilst also influencing general mobility and residence times. Changes in environmental circumstances, such as redox status, discharge, or large point source inputs may alter the state of dynamic equilibrium existing between solid and solution phases.

Together, these factors combine to profoundly influence nutrient cycles, through the contrasting behavior and properties of NO_3^- and PO_4^{3-} in surface waters. Importantly, these fundamental differences help explain the tendency for PO_4^{3-} to be widely considered as the nutrient most likely to be limiting in fresh waters. The greater reactivity of PO_4^{3-} results in a high selectivity for sorption onto solid phases and therefore the maintenance of low soluble P concentrations. This situation is the opposite to that which exists for NO_3^-, which generally has a poor affinity for solid phases and explains why NO_3^- concentrations are commonly 2 orders of magnitude greater than those of PO_4^{3-} (Table 7).

These differences in physicochemical behavior also means that while concentrations of PO_4^{3-} should be subjected to strong (internal) buffering from sorbed P, the reverse would be true for NO_3^- where concentrations reflect a combination of terrestrial (external) sources and the balance between in-stream utilization and production. An example of the changing conditions for low and high flows is shown in Table 7. Although NO_3^- concentrations remained similar to those for PO_4^{3-}, TP and SS increased by ~10, 50, and 290 times respectively at the high flow. The resulting change in instantaneous load was only 1 order of magnitude for NO_3^- compared to more than 3 orders of magnitude for P and SS and demonstrates the tendency for most P to be mobilized during episodic events often associated with SS. In situations without significant point sources, it might be expected that, during periods of extended steady flow, conditions will favor soluble forms of N and P leading to a widening of the N : P ratio compared to storm flow, where high SS and the associated P causes a narrowing of the N : P ratio. For the example illustrated in Table 7 the NO_3-N : TP ratio narrows from 762 at low flow to 17.6 at high flow. These values are largely for transport, and do not connotate biological availability, because many chemical states of the transported materials are relatively recalcitrant to biotic utilization.

Timing of N and P Delivery with Biological Requirement

The simple comparison made in Table 7 can be extrapolated to provide insight into the differing transport mechanisms that operate at the catchment scale. The effective delivery of nutrients from "diffuse sources" that occurs during episodic storm events introduces a strong temporal component to the dynamics of annual transport of N and P by rivers. High rates of nutrient flux tend to be associated with periods of greatest storm frequency, although it is unlikely that these high flow conditions are particularly favorable for productivity and uptake. This potential mismatch between the timing of nutrient supply and demand means that system retention times become important. With increasing retention times, as in lakes and reservoirs, a greater proportion of these episodically delivered nutrients will remain potentially available for biological uptake.

Attempts to fully understand the dynamics of nutrient availability requires a good understanding of the physicochemical differences in the reactivation and transport mechanisms that together modify the timing of delivery and composition of nutrient fluxes to drainage waters. The seasonality of biological uptake means that only a proportion of the total annual flux of nutrients constitutes the "biologically significant" fraction. With increasing catchment size and overall complexity, it becomes difficult to separate and quantify the precise contribution that individual sources make and therefore also where the most effective remediation strategies should be targeted. An extreme example of this situation has regularly been described for large, strongly retentive standing water bodies (also large slow flowing rivers), where "internal" cycling of existing nutrients is well capable of sustaining biological activity over long time periods. The size of an internal nutrient supply will influence the effectiveness of management strategies that focus only on reducing fresh inputs of nutrients.

By way of contrast to nutrient loading, the highest nutrient concentrations of impacted waters often occur under conditions of low flow where groundwater sources may dominate, dilution of point sources is minimal, and efficient recycling of existing nutrients in lakes becomes important. These different situations have lead to confusion

Table 7 Comparison between concentrations ($mg\,L^{-1}$) and instantaneous flux ($mg\,s^{-1}$) of nutrients between low and high stream discharge of a small second-order stream (Bronie Burn) draining agricultural land in NE Scotland

	Flow	SS		NO_3^--N		TP		PO_4^{3-}-P	
	($L\,s^{-1}$)	($mg\,L^{-1}$)	($mg\,s^{-1}$)	($mg\,L^{-1}$)	($mg\,s^{-1}$)	($mg\,L^{-1}$)	($mg\,s^{-1}$)	($mg\,L^{-1}$)	($mg\,s^{-1}$)
Low flow	2	4.64	9.28	12.2	24.4	0.016	0.032	0.016	0.032
High flow	72	1335	96 120	13.2	947.5	0.75	54	0.17	12.2

as to whether a nutrient loading or concentration should be considered when determining impact assessments. As a result, targeting of the most appropriate remediation actions in order to minimize impacts is difficult.

BIOLOGICAL EFFECTS OF NUTRIENT ENRICHMENT

Introduction

The productivity and internal metabolism of aquatic ecosystems are driven by energy acquired by photosynthesis from solar radiation. Inland waters receive organic products of photosynthesis directly from their aquatic flora and indirectly from the drainage basins as particulate and dissolved organic matter from terrestrial and wetland plants imported by stream water, storm runoff, groundwater, and the atmosphere (*see* **Chapter 102, Trophic Dynamics, Volume 3**).

Light is attenuated exponentially with increasing depth in water. The ultraviolet wavelengths are absorbed strongly by water and dissolved organic compounds in the water. The blue portion (400–500 nm) of the visible spectrum penetrates most deeply in relatively clear waters. Infrared wavelengths are absorbed rapidly by the molecular structure of the water molecules and much of this energy is dissipated as heat. Because water is less dense as temperature increases above 4 °C, less dense warmer water floats upon the more dense cooler water. The density stratification separates the lake into a warmer, less dense stratum (epilimnion) that overlies a cooler, denser zone (hypolimnion). The interface zone (metalimnion) between these strata is a region of rapid thermal discontinuity, often several degrees change per meter.

The resulting density stratification affects not only the thermal structure and water mass stratification, but also the hydrodynamics of lakes and reservoirs. Patterns of density-induced stratification influence physical and chemical properties and cycle both spatially within the water body and seasonally. These properties structure aquatic habitats and have marked effects on all chemical cycles, metabolic rates, and the population dynamics of organisms and their productivities (reviewed in detail in Wetzel, 2001).

In large lakes and reservoirs, the volume of water with sufficient light to support photosynthesis (euphotic zone) is small in relation to the total volume of water containing stored dissolved oxygen that can be used for heterotrophic respiration. In small and relatively shallow lakes, however, the proportion of water supporting photosynthesis is large relative to total lake volume, and respiratory demands can exhaust oxygen dissolved in the lower strata of the lake. The process of eutrophication exacerbates the exhaustion of dissolved oxygen in large portions of lake ecosystems. Eutrophication is the increased loading of organic matter to an aquatic ecosystem. Often in lakes and reservoirs, eutrophication results in increased photosynthetic production of organic matter in response to excessive loading of nutrients, particularly phosphorus, to the water from external sources (*see* **Chapter 108, Lake Ecosystems (Stratification and Seasonal Mixing Processes, Pelagic and Benthic Coupling), Volume 3**).

In flowing waters where individual nutrient retention times may be extremely short and sustained high concentrations are not typical, aquatic biota must display high uptake and utilization efficiencies in an almost opportunistic approach to nutrient acquisition. Periods of greatest nutrient loading are usually associated with storm events where physical conditions are not always conducive to active growth. It is likely that under these circumstances, periods of low flow and concentrations of individual nutrients become more important (Dodds *et al.*, 1996). The complexity of full ecosystem responses to nutrient additions has been demonstrated experimentally by Peterson *et al.* (2001). The first year of N additions resulted in a large increase in primary productivity, which was followed during the second year by secondary grazers, which resulted in a return to clear water conditions. Such studies have important implications with respect to the capacities for stream ecosystems to store and retain nutrient loads received from the catchment areas.

Succession of Eutrophication

Increased inorganic nutrient loading, particularly of P and combined N species, is fundamental to initial eutrophication and to the maintenance of high, sustained productivity by the phytoplankton communities. Low rates of productivity in oligotrophic lakes are maintained to a large extent by low inputs of organic matter, concomitant with slow rates of decomposition. Oxidizing hypolimnetic conditions are thus maintained and result in low rates of nutrient release from the sediments in a cyclical causal system. Dissolved organic compounds from both internal and external sources are usually low, which results in limited availability of organic micronutrients and reduced complexing capabilities for essential inorganic micronutrients.

Under eutrophic conditions, the loading rates of nutrients, especially P and combined N, are relatively high. As rates of photosynthetic productivity and organic loading to lower strata increase, nutrients are released from sediments into anoxic hypolimnia, increasing recycling rates. Phytoplanktonic productivity of eutrophic lakes and reservoirs increases markedly. In increasingly eutrophic lakes, dense algal communities reduce light penetration and thereby compress the depth of the trophogenic zone. Some partial circumvention of the effects of self-shading can be achieved by buoyant species (e.g. gas-vacuolated cyanobacteria) or even flagellated species. Light limitations caused by self-shading set an upper boundary to the phytoplanktonic photosynthetic productivity, beyond which further

increases are not possible, regardless of increased nutrient availability. Further increases in photosynthetic productivity are possible only by extending the length of the growing season (e.g. in the tropics) or by increasing turbulence and frequency of exposure to available light, as in artificially mixed sewage lagoons.

Impacts of Nutrient Enrichment

Nutrient enrichment can modify both biomass productivity and community structure (see Harper, 1991; Moss, 1998) often being associated with a decrease in species biodiversity. Rates of algal production increase, as nutrient limitation of the phytoplankton of infertile waters are increasingly relieved by sustained nutrient inputs. As the densities of the phytoplankton community progressively reduce the available light and thickness of the trophogenic zone, productivity per unit surface area is usually lower under these hypereutrophic conditions than under less productive conditions where a thicker trophogenic zone exists.

Certain characteristic phytoplankton associations occur repeatedly in lakes of increasing nutrient enrichment. Although variance is high, there is a general tendency to shift from a dominance of diatoms, dinoglagellates, chrysophytes, and cryptophytes in oligotrophic waters to a dominance of cyanobacteria, euglenophytes, and diatoms and green algae among eutrophic waters (Table 8).

The ranges of primary productivity of phytoplankton commonly associated with oligotrophy and eutrophy are large (Table 8). The primary productivity of phytoplankton is commonly used as a major criterion for determining trophic state of a lake or reservoir. This criterion, however, assumes that dissolved and particulate organic matter inputs from littoral and external terrestrial sources are small relative to those of the phytoplankton. In many cases on a global basis, littoral productivity and allochthonous organic inputs are much larger than inputs produced by phytoplankton productivity (Wetzel, 1990a). In such cases, phytoplankton productivity alone is a poor estimator of the trophic state of a lake or reservoir. The primary productivity of phytoplankton must be viewed as a variably important contributor to overall aquatic metabolism (Table 9).

Remediation via Nutrient Reduction

Eutrophication is a process leading to increased biological productivity and decreased basin volume from excessive addition of dissolved and particulate inorganic and organic materials to lakes and reservoirs. Although much of the earlier emphasis on eutrophication was on the basis of increased growth of phytoplankton algae, a broader context would include the rapid alterations to the basin morphometry caused by siltation. Increased littoral areas lead to rapid increases in the development of aquatic macrophytes of high productivity and accelerated internal loadings of organic matter. As lakes become shallower, the exchange of nutrients that are remineralized in the sediments within the productive trophogenic zone becomes more likely.

Enhanced P availability is most frequently the cause of enhanced productivity in lakes and reservoirs. Although other elements, particularly N, can become the dominant limiting nutrient for phytoplankton growth in certain waters when P supplies are sufficient, on a global basis P is the foremost nutrient limiting to autochthonous photosynthesis on a sustained long-term basis. Lake and reservoir restoration efforts have, as a result, been directed largely toward reducing the loading of P, and to a lesser extent N, to the surface waters by advanced wastewater treatment, diversion, land management, or reducing the P load in wastewater by restricting the P content of detergents. If releases of sediment P stores become significant, in part stimulated by low redox potentials that have resulted from organic matter loading, production, and deposition, attempts are often made to control the availability or recycling of nutrients by physical (e.g. aeration) or chemical (e.g. P inactivation, sediment oxidation) methods within the lake or reservoir (Cooke *et al.*, 1993; Eiseltová, 1994; Wetzel, 1990b, 2001).

Lake and reservoir management and restoration have focused on problems particularly associated with excessive nutrient loading and poor land management. Prevention of nutrient pollution and eutrophication is clearly the most prudent long-term solution. Reductions in nutrient loadings often require evaluation of diffuse and point sources within the drainage basin and a systematic multifaceted program of reduction and control. Once the loadings of nutrients and water are known, a nutrient budget can be used to evaluate the dynamics of annual inputs and outputs of substances. The budget permits the estimation of changes in algal biomass that could result from increases or decreases in nutrient loadings, water residence time, or increased mean depth as may occur from sediment removal.

Phosphorus release from sediments can be very high in lakes with anoxic hypolimnia. This release can delay the predicted reduction in phytoplankton biomass following reductions in external nutrient loading. The internal loading can be estimated and incorporated into nutrient budget data and models.

Nine common methods are used, singly or in combination, to reduce nutrient availability or suitable habitat for photosynthesis (Klapper, 1991; Cooke *et al.*, 1993).

1. Nutrient removal by advanced treatment and land management
2. Nutrient diversion of external nutrient loadings away from receiving rivers, reservoirs, or lakes
3. Hypolimnetic withdrawal of nutrient-enriched hypolimnetic water by siphoning, pumping, or deepwater discharge

Table 8 General ranges of primary productivity of phytoplankton and related characteristics of lakes of different trophic categories[a]

Trophic type	Mean primary productivity ($mg\,C\,m^{-2}\,day^{-1}$)[b]	Phyto plankton density ($cm^3\,m^{-3}$)	Phyto plankton biomass ($mg\,C\,m^{-3}$)	Chloro phyll a ($mg\,m^{-3}$)	Dominant phyto plankton	Light extinction coefficients (m^{-1})	Total organic carbon ($mg\,L^{-1}$)	Total P ($\mu g\,L^{-1}$)	Total N ($\mu g\,L^{-1}$)	Total inorganic solids ($mg\,L^{-1}$)
Ultraoligotrophic	<50	<1	<50	0.01–0.5		0.03–0.8		<1–5	<1–250	2–15
Oligotrophic	50–300		20–100	0.3–3	Chrysophyceae, Cryptophyveae	0.05–1.0	<1–3			
Oligomesotrophic		1–3			Dinophyceae, Dinophyceae, Bacillariophyceae			5–10	250–600	10–200
Mesotrophic	250–1000		100–300	2–15		0.1–2.0	<1–5			
Mesoeutrophic		3–5						10–30	500–1100	100–500
Eutrophic	>1000		>300	10–500	Bacillariophyceae, Cyanobacteria	0.5–4.0	5–30			
Hypereutrophic		>10			Chlorophyceae, Euglenophyceae			30–>5000	500–>15 000	400–60 000
Dystrophic	<50–500		<50–200	0.1–10		1.0–4.0	3–30	<1–10	<1–500	5–200

[a]Modified from Wetzel (2001), after many authors and sources.
[b]Referring to approximately net primary productivity, such as measured by the ^{14}C methods.

Table 9 Comparative characteristics of phytoplankton and their productivity among river, reservoir, and natural lake ecosystems[a]

Property	Rivers	Reservoirs	Natural lakes
Phytoplankton diversity	Very low in low-order streams; increasing in high-order large rivers	Low in riverine zones; increasing in lacustrine zone	High diversity in oligotrophic lakes, decreasing in eutrophic lakes; similar in tropical and temperate lakes
Phytoplankton biomass	Very low in low-order streams; increasing in large rivers, although often light-limited	Moderately high in response to high nutrient supply in riverine sections; less in lacustrine zone where nutrients often limiting	Highly variable (to 5 orders of magnitude) in temperate lakes seasonally; much less variable in tropical lakes
Phytoplankton productivity	Low, often light-limited by advective flows and turbidity	Highest in transitional zone; reduced in riverine zone by light availability and often in lacustrine zone by nutrients; marked horizontal gradients; volumetric productivity (P_{max}) decreases from headwaters to dam; areal productivity relatively constant horizontally	Low in comparison to littoral productivity; increasing with moderate nutrient loading; declining at very high nutrient loading; seasonal and vertical gradients predominate; small horizontal gradients; light and inorganic nutrient limitations predominate

[a]Modified from Wetzel (2001) derived from numerous sources cited therein.

4. Dilution and flushing of a reservoir by a source of nutrient-poor water
5. Phosphorus precipitation and inactivation due to adsorption or precipitation after the addition of aluminum compounds, or natural clay and carbonate particles
6. Sediment oxidation by the additions of an alternative electron acceptor, such as nitrate to delay the reduction of iron and release of P
7. Aeration of the anoxic hypolimnia of lakes and reservoirs
8. Sediment removal in shallow eutrophic lakes to remove nutrient accumulations and increase mean depth
9. Food-web manipulations of zooplankton and fish and their efficiencies of herbivory on algae, animal predation, and nutrient recycling.

River Management and Restoration

From a geographical context, two general approaches are employed to reduce the flux of nutrients deriving from terrestrial sources. Attention has been placed upon management options that result in the interception and retention of nutrients within the riparian zone or through the use of more general blanketed catchmentwide actions that favor measures aimed at reducing the nutrient surplus of agricultural systems (e.g. Öborn *et al.*, 2003). Successful action measures require the simultaneous and integrated management of both aspects in order to deliver the best options for long-term restoration of rivers (*see* **Chapter 185, Integrated Land and Water Resources Management, Volume 5**).

The Riparian Zone

A fundamental tenet of both river management and restoration is to protect or restore the riparian floodplain areas of streams and rivers. In many cases, streams and rivers have been channelized by straightening and deepening the channel, particularly in lowland agricultural areas to increase drainage. Accompanying the loss of stream length and meandering is a loss of riffles and pools, a loss of riparian floodplains and wetlands, and a loss of riparian vegetation (*see* **Chapter 107, Natural and Constructed Wetlands, Volume 3**). As a result, hydraulic head is increased when energy dissipation is reduced. Increased flows reduce habitat diversity, particularly in the sediments and adjacent wetland, with a catastrophic reduction in habitat diversity and biodiversity (e.g. Gore, 1985; Petersen *et al.*, 1992; Eiseltová and Biggs, 1995). Loss of riparian floodplains and wetlands decreases water tables, increases rates of water runoff, enhances nutrient losses from adjacent land, and increases scouring and sediment yield to the channel. Water retention capacities are greatly reduced and as a result, during high precipitation events, flooding is very common and much more severe that would be the case if portions of the riparian and floodplain environments were retained.

Several means of river management and restoration measures include:

1. allowing the river channel to migrate laterally with periodic inundation of the floodplain above river bank capacity;
2. encouraging and developing buffer strips of plant and microbial communities along each side of the river channel to function in bank stabilization and as a metabolic filter for effective sequestering and retention of nutrients;
3. encouraging the development of natural wetlands, particularly at juncture points of tributaries to the main channel; and

4. reducing of slope gradient at the edge of the channel to encourage channel meandering and floodplain regeneration.

The Wider Catchment Area

Unfortunately, while it is possible to effectively reduce the flux of nutrients derived from terrestrial sources, it is difficult to envisage a general situation in the near future where these losses approach levels that existed prior to the recent (last 50 years) phase of agricultural intensification. The reasons for this lack of response relate to the range of retention and transformation processes described earlier, together with the accumulated increased nutrient capital of many terrestrial and aquatic ecosystems and the extent of "chronic" background enrichment experienced by many surface and groundwaters. This rather pessimistic observation means it is even more important to agree to the prioritization and introduction of appropriate catchmentwide management objectives. A good discussion of this rapidly developing area is available at USEPA (2003a) and Water Framework Directive (2000).

The often very site-specific nature of many nutrient driven biological impact means that a wide range of catchment-scale management options exist that might prove to be effective. These may tackle either the direct input of nutrients or indirectly through modifying the hydrological transport pathways and connectivity to river systems. Some of the options are listed here.

1. Reduce sources of rapid storm runoff from industrial/urban areas by increasing infiltration and short-term storage.
2. Reduce nutrient capital of soil-plant systems, through improved efficiency of nutrient use and/or change of land use.
3. Attempt to minimize production of soil nitrate.
4. Develop natural barriers (e.g. hedgerows) to dissipate energy and generation of surface runoff and soil erosion.
5. Follow and enforce strict best management agricultural practices.

MODELING

Models aid in the identification and quantification of nutrient sources both spatially and temporally across an individual catchment (*see* **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**; **Chapter 100, Water Quality Modeling, Volume 3**). The complexity and range of issues involved in the modeling of nutrient supply, transport, and impact has meant that a variety of approaches have been adopted (see Haygarth and Jarvis, 2002). As with most modeling exercises, the appropriate choice depends greatly upon the amount of temporal and spatial resolution that is considered necessary together with the availability of reliable data for parameterization. Often the choice has to be made between those models that are complex, data hungry, process based, and single element specific, and those that have both poor temporal and spatial resolution, but that are easily parameterized and often employ simple regression-type relationships. While the former provide a high degree of spatial and temporal understanding, they are not necessarily robust or easily used among different geographical situations. Examples of the second approach include loss coefficients defined for particular landcover/soil-type/climatic combinations (*see* **Chapter 15, Digital Elevation Model Analysis and Geographic Information Systems, Volume 1**) using relationships developed between certain catchment level attributes and the N and P losses.

A variety of process-based models is available, which, because of the fundamental differences between N and P, tend to emphasize biologically mediated aspects of the N cycle as compared to those physicochemical process that are important for P. This coupling helps to explain why many of the process-based models tend to be nutrient specific. A review of current modeling efforts can be found at Benchmark Models (2003), where the objective is to establish a set of criteria to assess the appropriateness of integrated models for use in the implementation of Water Framework Directive. Other approaches prefer to focus upon the transport and delivery aspects of nutrient transfer to aquatic systems, and therefore include a strong underlying hydrological component.

One particularly important modeling goal is the accurate apportioning of nutrient sources that contribute to both total and biologically significant losses. As the size/area of a catchment increases it becomes difficult to resolve individual sources because of the compounding effects of damping and retention. Source apportioning is therefore an important component of the effective targeting of nutrient management. One recent approach that should be transferable between river systems and uses the natural hierarchical stream order structure combined with a hydrological and biogeochemical model has been described for the River Danube by Garnier *et al.* (2002).

Models that improve the system-level understanding of increased nutrient supply and impacts together with the linked attempts to better define system threshold or target concentrations have been developed. The most successful and widely used models were developed and applied to standing waters. Being able to identify the key role that physical factors such as water depth and turnover time play within broad classes of standing waters has been central to this success. The lack of a similar understanding and an ascribed series of generalized physical criteria are perhaps the reasons for the much poorer prediction capacities for

flowing waters. River systems function as both a conduit for nutrient transport as well having the potential to be impacted in their own right.

THE FUTURE EMPHASIS

The contribution individual potential sources of nutrients make at both local and whole catchment scale requires further quantification. The accumulated impacts of nutrient enrichment on ecosystems immediately down gradient and ultimately coastal/marine systems remains a major research objective. Central to this requirement is the development of target or threshold values for individual nutrients which must reflect their relationship to realistically set conditions for "background" states. The use of ecoregions is one way forward and this approach can be investigated further at USEPA (2003b). A similar level of understanding of the relationship that exists between nutrient supply and biological impact that has been established for standing waters is required for flowing systems. Unfortunately, terrestrial nutrient cycles are inevitably "leaky", which is particularly so for NO_3, while point sources are often major contributors of P. Even when apparently achievable objectives have been decided upon, the buffering mechanisms operating within soil-plant nutrient cycles introduce a strong hysteresis into the recovery. An example of the difficulties that this can introduce has become evident for the Baltic Sea where responses to management action plans were slower than anticipated (Grimvall *et al.*, 2000).

For many river situations, it is not always the total nutrient loss that determines a potential impact; the matching of periods of nutrient supply with those of biological demand needs much greater consideration. To achieve this requires a greater understanding of the particular aquatic system attributes that help to modify retention times of individual nutrients. A recent example of this situation has been described by Edwards *et al.* (2003) for a eutrophic estuarine system. An estimated 70% of the terrestrially derived nitrate passes through the estuarine system of the River Ythan, NE Scotland, during the biologically dormant period when algal growth is at a minimum.

The central role that hydrologically mediated transport plays in the delivery of N and P means that changes in either the frequency of storm events or quantity of rainfall can have very significant implications for nutrient cycling. Periods of low summer flows can be associated with poor dilution of point sources and high river concentrations (*see* **Chapter 34, Climate Change – Past, Present and Future, Volume 1**; **Chapter 189, Land Use and Water Resources Under a Changing Climate, Volume 5**). The possibility that climate change might be altering rainfall-runoff patterns requires consideration. The tremendous seasonal and annual variability in climate, which influences nutrient supply and retention means that many aspects relating to eutrophication, especially for the shorter residence time riverine environments, remain poorly understood.

FURTHER READING

The following sections provide an overview of the essential attributes and factors that are responsible for regulating nutrient cycling in both standing and flowing water bodies. In such a short text it is difficult to cover all topics or eventualities adequately, therefore the following books are recommended to provide a more comprehensive overview in general (Moss, 1998; Harris, 1999; Lewis, 2000; Wetzel, 2001; Smith, 2003) and specific (Brezonik, 1993; Stumm and Morgan, 1996) issues.

Brezonik P.L. (1993) *Chemical Kinetics and Process Dynamics in Aquatic Systems*, Lewis Publishers: Boca Raton.

Harris G.P. (1999) Comparison on the biogeochemistry of lakes and estuaries: Ecosystem processes, functional groups, hysteresis effects and interactions between macro- and microbiology. *Journal of Marine and Freshwater Research*, **50**, 791–811.

Lewis W.M. (2000) Basis for the protection and management of tropical lakes. *Lakes and Reservoirs: Research and Management*, **5**, 38–48.

Pieterse N.M., Bleuten W. and Jørgensen S.E. (2003) Contribution of point sources and diffuse sources to nitrogen and phosphorus loads in lowland river tributaries. *Journal of Hydrology*, **271**, 213–225.

Smith V.H. (2003) Eutrophication of freshwater and coastal marine ecosystems – A global problem. *Environmental Science and Pollution Research*, **10**, 126–139.

Stumm W. and Morgan J.J. (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*, *Third Edition*, John Wiley & Sons: New York, p. 1040.

REFERENCES

Aber J.D., Goodale C.L., Ollinger S.V., Smith M.L., Magill A.H., Martin M.E., Hallett R.A. and Stoddard J.L. (2003) Is nitrogen deposition altering the status of northeastern forests? *BioScience*, **53**, 375–389.

Benchmark Models (2003) *Benchmark Models for the Water Framework Directive*, **http://www.environment.fi/default.asp?contentid=79639&lan=EN**, (accessed 3/12/2004).

Bernhardt E.S., Hall R.O. and Likens G.E. (2002) Whole-system estimates of nitrification and nitrate uptake in streams of the Hubbard Brook Experimental Forest. *Ecosystems*, **5**, 419–430.

Carlton R.G. and Wetzel R.G. (1988) Phosphorus flux from lake sediments: Effects of epipelic algal oxygen production. *Limnology and Oceanography*, **33**, 562–570.

Conley D.J. (2000) Biogeochemical nutrient cycles and nutrient management strategies. *Hydrobiologia*, **410**, 87–96.

Cooke G.D., Welch E.B., Peterson S.A. and Newroth P.R. (1993) *Restoration and Management of Lakes and Reservoirs*, *Second Edition*, Lewis Publishers: Chelsea.

Dahm C.N., Grimm N.B., Marmonier P., Valett H.M. and Vervier P. (1998) Nutrient dynamics at the interface between surface waters and groundwaters. *Freshwater Biology*, **40**, 427–451.

de Jonge V.N., Elliott M. and Orive E. (2002) Causes, historical development, effects and future challenges of a common environmental problem: Eutrophication. *Hydrobiologia*, **475/476**, 1–19.

Dodds W.K. and Biggs B.J.F. (2002) Water velocity attenuation by stream periphyton and macrophytes in relation to growth form and architecture. *Journal of the North American Benthological Society*, **21**, 2–15.

Dodds W.K., Hutson R.E., Eichem A.C., Evans M.A., Gutter D.A., Fritz K.M. and Gray L. (1996) The relationship of floods, drying, flow and light to primary production and producer biomass in a prairie stream. *Hydrobiologia*, **333**, 151–159.

Dorioz J.M., Cassell E.A., Orand A. and Eisenman K.G. (1998) Phosphorus storage, transport and export dynamics in the Foron River watershed. *Hydrological Processes*, **12**, 285–309.

Edwards A.C., Sinclair A.H. and Domburg P. (2003) Identification, designation and formulation of an action plan for a nitrate vulnerable zone: A case study of the Ythan catchment, NE Scotland. *European Journal of Agronomy*, **20**, 165–172.

Egeberg P.K., Eikenes M. and Gjessing E.T. (1999) Organic nitrogen distribution in NOM size classes. *Environment International*, **25**(2–3), 225–236.

Eiseltová M. (1994) *Restoration of Lake Ecosystems: A Holistic Approach*, Vol. 32, Publications of the International Waterflow and Wetlands Research Bureau: Gloucester, p. 182.

Eiseltová M. and Biggs J. (1995) *Restoration of Stream Ecosystems: An Integrated Catchment Approach*, Vol. 37, Publications of the International Waterflow and Wetlands Research Bureau: Gloucester, p. 170.

Frost P.C., Stelzer R.S., Lamberti G.A. and Elser J.J. (2002) Ecological stoichiometry of trophic interactions in the benthos: Understanding the role of C : N : P ratios in lentic and lotic habitats. *Journal of the North American Benthological Society*, **21**, 515–528.

Galloway J.N. (1998) The global nitrogen cycle: Changes and consequences. *Environmental Pollution*, **102**(S1), 15–24.

Galloway J.N. and Cowling E.B. (2002) Reactive nitrogen and the world: 200 years of change. *Ambio*, **31**, 64–71.

Garnier J., Billen G., Hannon E., Fonbonne S., Videnina Y. and Soulie M. (2002) Modelling the transfer and retention of nutrients in the drainage network of the Danube River. *Estuarine, Coastal and Shelf Science*, **54**, 285–308.

Gore J.A. (1985) *Restoration of Rivers and Streams*, Butterworths: Boston, p. 320.

Green P.A., Vörösmarty C.J., Meybeck M., Galloway J.N., Peterson B.J. and Boyer E.W. (2004) Pre-industrial and contemporary fluxes of nitrogen through rivers: A global assessment based on typology. *Biogeochemistry*, **68**, 71–105.

Grimvall A., Stålnacke P. and Tonderski A. (2000) Time scales of nutrient losses from land to sea – A European perspective. *Ecological Engineering*, **14**, 363–371.

Haggard B.E., Storm D.E. and Stanley E.H. (2001) Effect of a point source input on stream nutrient retention. *Journal of the American Water Resources Association*, **37**, 1291–1299.

Harper D. (1991) *Eutrophication of Freshwaters: Principles, Problems and Restoration*, Chapman & Hall: London, p. 256.

Harvey J.W. and Wagner B.J. (2000) Quantifying hydrologic interactions between streams and their subsurface hyporheic zones. *In Stream and Ground waters*, Jones J.B. and Mulholland P.J. (Eds.), Academic Press: pp. 4–44.

Haslam S.M. (1978) *River Plants: The Macrophytic Vegetation of Watercourses*, Cambridge University Press: Cambridge, p. 396.

Haygarth P.M. and Jarvis S.C. (Eds) (2002) *Agriculture, Hydrology and Water Quality*. CAB International: Wallingford, p. 502.

Hedin L.O., von Fischer J.C., Ostrom N.E., Kennedy B.P., Brown M.G. and Robertson G.P. (1998) Thermodynamic constraints on nitrogen transformations and other biogeochemical processes at soil-stream interfaces. *Ecology*, **79**, 684–703.

Howarth R.W., Billen G., Swaney D., Townsend A., Jaworski N., Lajtha K., Downing J.A., Elmgren R., Caraco N., Jordan T., *et al.* (1996) Regional nitrogen budgets and riverine N and P fluxes for the drainages to the North Atlantic Ocean: Natural and human influences. *Biogeochemistry*, **35**, 75–139.

Hudson J.J., Taylor W.D. and Schindler D.W. (2000) Phosphate concentrations in lakes. *Nature*, **406**, 54–56.

Hutchinson G.E., Bonatti E., Cowgill U.M., Goulden C.E., Levanthal E.A., Mallett M.E., Margaritora F., Patrick R., Racek A., Roback S.A., *et al.* (1970) Ianula: An account of the history and development of the Lago di Monterosi, Latium, Italy. *Transactions of the American Philosophical Society, N. S.*, **60**(4), 178.

Johnes P.J. and Butterfield D. (2002) Landscape, regional and global estimates of nitrogen flux from land to sea: Errors and uncertainties. *Biogeochemistry*, **57–58**, 429–476.

Johnson L.B. and Gage S.H. (1997) Landscape approaches to the analysis of aquatic ecosystems. *Freshwater Biology*, **37**, 113–132.

Klapper H. (1991) *Control of Eutrophication in Inland Waters*, Ellis Horwood: New York, p. 337.

Likens G.E. and Bormann E.H. (1995) *Biogeochemistry of a Forested Ecosystem*, *Second Edition*, Springer-Verlag: New York.

Maberly S.C., King L., Gibson C.E., May L., Jones R.I., Dent M.M. and Jordan C. (2003) Linking nutrient limitation and water chemistry in upland lakes to catchment characteristics. *Hydrobiologia*, **506–509**, 83–91.

Maule C.P. and Fonstad T.A. (2000) Impacts of cattle penning on groundwater quality beneath feedlots. *Canadian Agricultural Engineering*, **42**(2), 87–93.

Meyer J.L., McDowell W.H., Bott T.L., Elwood J.W., Ishizaki C., Melack J.M., Peckarsky B.L., Peterson B.J. and Rublee P.A. (1988) Elemental dynamics in streams. *Journal of the North American Benthological Society*, **7**, 410–432.

Moss B. (1998) *Ecology of Fresh Waters: Man and Medium, Past to Future*, *Third Edition*, Blackwell Science Ltd: p. 560.

Öborn I., Edwards A.C., Witter E., Oenema O., Ivarsson K., Withers P.J.A., Nilsson S.I. and Richert Stinzing A. (2003) Element balances as a tool for sustainable nutrient management: A critical appraisal of their merits and limitations within an agronomic and environmental context. *European Journal of Agronomy*, **20**, 211–225.

Owens P.N. and Walling D.E. (2002) The phosphorus content of fluvial sediment in rural and industrialized river basins. *Water Research*, **36**, 685–701.

Perakis S.S. and Hedin L.O. (2002) Nitrogen loss from unpolluted South American forests mainly via dissolved organic compounds. *Nature*, **415**, 416–419.

Petersen R.C., Petersen L.B.M. and Lacoursiére J. (1992) A building-block model for stream restoration. In *River Conservation and Management*, Boon P.J., Calow P. and Petts G.E. (Eds.), John Wiley and Sons: Chichester, pp. 293–309.

Peterson B.J., Wollheim W.M., Mulholland P.J., Webster J.R., Meyer J.L., Tank J.L., Marti E., Boowden W.B., Valett H.M., Hershey A.E., *et al.* (2001) Control of nitrogen export from watersheds by headwater streams. *Science*, **292**, 86–90.

Pusch M., Fiebig D., Brettar I., Eisenmann H., Ellis B.K., Kaplan L.A., Lock M.A., Naegeli M.W. and Traunspurger W. (1998) The role of micro-organisms in the ecological connectivity of running waters. *Freshwater Biology*, **40**, 453–495.

Reay D.S., Smith K.A. and Edwards A.C. (2003) Nitrous oxide emission from agricultural drainage waters. *Global Change Biology*, **9**(2), 195–203.

Reddy K.R., Wetzel R.G. and Kadlec R. (2005) Biogeochemistry of phosphorus in wetlands. In *Phosphorus: Agriculture and the Environment*, Sims J.T. and Sharpley A.N. (Eds.), Soil Science Society of America: (In press).

Roden E.E. and Wetzel R.G. (1996) Organic carbon oxidation and suppression of methane production by microbial Fe(III) oxide reduction in vegetated and unvegetated freshwater wetland sediments. *Limnology and Oceanography*, **41**, 1733–1748.

Smaling E.M.A., Oenema O. and Fresco L.O. (Eds) (1999) *Nutrient Disequilibria in Agroecosystems: Concepts and Case Studies*. CAB International: Wallingford, p. 336.

Smith S.V., Swaney D.P., Talaue-McManus L., Bartley J.D., Sandhei P.T., McLaughlin C.J., Dupra V.C., Crossland C.J., Buddemeier R.W., Maxwell B.A., *et al.* (2003) Humans, hydrology, and the distribution of inorganic nutrient loading to the ocean. *BioScience*, **53**, 235–245.

Steele K. (Ed) (1995) *Animal Waste and the Land-Water Interface*. Lewis Publishers: Boca Raton, p. 589.

Straškraba M. (1996) Ecotechnological methods for managing non-point source pollution in watersheds, lakes and reservoirs. *Water Science and Technology*, **33**, 73–80.

USEPA (2003a) Pollution Prevention **http://www.epa.gov/agriculture/apol.html**, (accessed 3/12/2004).

USEPA (2003b) **http://www.epa.gov/waterscience/criteria/nutrient/ecoregions/rivers/** and **http://www.epa.gov/waterscience/criteria/nutrient/ecoregions/lakes/**, (both accessed 3/12/2004).

Water Framework Directive (2000) **http://europa.eu.int/comm/environment/water/water-framework/index_en.html**, (accessed 3/12/2004).

Wetzel R.G. (1990a) Land-water interfaces: Metabolic and limnological regulators. *Verhandungen Internationale Vereinigung für Theoretische und Angewandte Limnologie*, **24**, 6–24.

Wetzel R.G. (1990b) Reservoir ecosystems: Conclusions and speculations. In *Reservoir Limnology: Ecological Perspectives*, Thornton K.W., Kimmel B.L. and Payne F.E. (Eds.), Wiley-Interscience: New York, pp. 227–238.

Wetzel R.G. (1999) Organic phosphorus mineralization in soils and sediments. In *Phosphorus Biogeochemistry in Subtropical Ecosystems*, Reddy K.R., O'Connor G.A. and Schelske C.L. (Eds.), Lewis Publishers: Boca Raton, pp. 225–245.

Wetzel R.G. (2001) *Limnology: Lake and River Ecosystems*, *Third Edition*, Academic Press: San Diego, p. 1006.

97: Urban Water Quality

J BRYAN ELLIS,[1] JIRI MARSALEK[2] AND BERNARD CHOCAT[3]

[1] *Urban Pollution Research Centre, Middlesex University, Enfield, UK*
[2] *National Water Research Institute, Burlington, ON, Canada*
[3] *INSA de Lyon, URGC Hydrologie Urbaine, Villeurbanne Cedex, Lyon, France*

Steady growth of population, due to overall population increases and continuing migration from rural to urban areas, creates enormous demands and stresses on urban waters with respect to water supply, drainage, flood protection, wastewater management, and beneficial uses of receiving waters and groundwater. Urban water issues are therefore in the forefront of water management priorities in practically all regions of the world, though often for broadly varying reasons. Key issues of urban water management are discussed in this article, which focuses on the evolution of urban drainage infrastructure, characterization of urban drainage, urban runoff impacts on receiving waters, urban drainage management, water, and wastewater reuse and future perspectives and priorities. The discussion focuses on the collection and transport of urban effluents (sewer systems), characterization of urban drainage provided by combined or storm sewers (flows and their quality), impacts of urban drainage effluents on receiving waters, and groundwater and impact mitigation by integrated urban drainage management with the emphasis placed on the management of surface runoff and water/wastewater reuse. While the progress in integrated engineering science, watershed-based management and new water technologies is impressive, the challenges of maintaining and improving urban water services, particularly in low-income countries, are formidable and may be further exacerbated by demographic, social, and climate change.

INTRODUCTION

Flood protection, drainage, and sanitation have always ranked highly in the needs of most societies, and, even in early civilizations, cities such as Ur and Babylon of the second century millennium B.C. Mesopotamian Empire possessed sophisticated wastewater collection and stormwater drainage systems, the remains of which can still be found. Significant advances in urban drainage technology were introduced during the period of the Roman Empire with roadway drainage, underground conduits, and sewer networks primarily intended for flood mitigation and the drainage of lowlands. The collection of rainwater for household and public use was also considered important especially given that domestic water consumption during the Roman period reached very high levels of 300 to 500 L per person per day.

Sanitation practices deteriorated after the decline of the Roman Empire with surface drains and streets being used in the Middle Ages as the only means of conveyance and disposal of all kinds of water-borne wastes. Water consumption declined to less than 15 L per head per day with already polluted urban waters being abstracted for further use in paper, fabric, and leather industries. Stormwater and foul sewage streams thus became indiscriminately mixed, becoming so noxious that they had to be covered and turned into sewers, giving rise to the birth of the "combined" sewer principle. The first beginnings of modern urban drainage practices were intended for stormwater control and were initiated in European cities during the nineteenth century, particularly following the numerous epidemics of typhoid and cholera in Europe and the United States in the 1830–1870 period. Inlets, gutters, and sewers replaced open street channels in Paris during 1810 to 1839 under the efforts of the engineers Bruneseau and Emmery, and, in 1843, the first comprehensively planned sewerage system for a major city was undertaken in Hamburg, Germany. The London sewer system designed by Bazalgette was introduced between 1859

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

and 1865, and, in the United States, main urban drainage systems were introduced in Chicago (1850s) and New York (1880). A good review of the history of urban drainage in the United States is given in Burian *et al.* (2000) with information on recent urban pollution control strategies provided in the chapter by Loagne and Lorin (*see* **Chapter 94, Point and NonPoint Source Pollution, Volume 3**).

The perspective of urban drainage also changed from a design standpoint during the late nineteenth century. Intuitive reasoning about conversion of rainfall into runoff led to the emergence of the Rational Method through the work of Mulvaney (1851), Kuichling (1889) and Lloyd–Davies (1906). By the end of the late nineteenth century, engineers possessed various design concepts and methods for wastewater disposal systems, and, for the next 100 years, these would become the standard tools used in urban drainage design throughout the world. Since the 1960s, rapid developments have occurred in urban drainage practice and this can be directly linked to the introduction of the computer and associated electronics such that it is now possible to calculate flows in sewer and drain networks with high precision and resolution to support cost-effective design, analysis, and operation.

Whilst these advances have helped in coping with urban flooding and have substantially improved the health of urban citizens, progress in water quality considerations and particularly those addressing the impact of increasing urban populations and their activities upon both surface and groundwaters have been much slower. Unfortunately, the processes that control water quality in drainage systems are much more complex and less deterministic than those which control flow rates (see McCutcheon, **Chapter 100, Water Quality Modeling, Volume 3**). More recently, major changes in drainage design and operation philosophy have been introduced as a result of

- the introduction and adoption of the concepts of ecological integrity and sustainable approaches to environmental and water resources management set within a watershed-wide framework;
- acceptance of the need to consider urban drainage, wastewater systems, and receiving waters in a holistic, integrated manner;
- the continuing development of computing power and an associated range of new analytical and real-time control techniques.

However, almost two-thirds of the world's population has no inherited sewered infrastructure, and many struggle with recurrent flooding and the daily need to find a place to carry out the most fundamental personal ablutions. Many would argue that the techniques and paradigms on which the wastewater disposal systems were developed in the developed nations are totally irrelevant to the needs and circumstances of developing urban populations and that new urban drainage paradigms are required. The major objectives for urban drainage remain public hygiene, flood protection, and environmental enhancement, although the emphasis in developed countries has been firmly placed to date on flood and pollution control. However, at the beginning of the twenty-first century, urban drainage has evolved to become much more than the simple transport and treatment of urban runoff and the time is ripe for the introduction of new paradigms based on more long-term sustainable strategies.

CHARACTERIZATION OF URBAN DRAINAGE

Introduction

Urban runoff includes dry weather sewage baseflow, stormwater, combined sewer overflows (CSO), as well as industrial effluent discharges and has been identified as a source of receiving water pollution for nearly 50 years. However, it is only in the last 20 years that national efforts within North America, Europe, Japan, and Australia have been made to identify and quantify the various pollutants and urban land uses responsible for such contamination. Surface Water Outfalls (SWOs) are essentially generated by storm rainfall conveying stormwater over impermeable urban surfaces to the separate surface water sewer system although non-wet weather flow can occur due to blockages, line breaks, vandalism, or misconnections. CSOs represent the combined volume of wastewater and stormwater runoff entering combined sewer systems, which exceeds the conveyance and treatment capacity of the drainage network, and is diverted to the receiving water by overflow regulators.

Urban Surface Water Pollutants

The range of pollutant concentrations and loadings associated with stormwater runoff from impermeable urban surfaces indicates that such surface water discharges can be highly variable in quality (Table 1), with standard deviations frequently being 75% (equivalent to a coefficient of variation, C_v, of 0.75) of the average event mean concentration (EMC) value. Table 1 would suggest that EMCs are frequently close to, if not exceeding the minimum NOEL (no observable effects limit) value and thus potentially present a problem to receiving water ecology. However, land use pollutant loading relationships do have a high degree of site specificity, and regional extrapolations on a continental scale cannot be readily applied. Further, the impact is varied with organism; some are victim to chronic low level exposure, others to acute higher concentration flushes. Nevertheless, the land use mean EMC value $\times$ Runoff Volume approach provides a convenient and appropriate screening-level methodology for estimation

Table 1 Pollutant concentrations and loadings for urban stormwater runoff

Pollutant parameter	Event mean concentration and range (mg L^{-1})		Load per unit area (kg imp.ha^{-1} $year^{-1}$)		Minimum concentration causing observable biological effects
	Residential & commercial	Motorways & trunk roads	Residential/ commercial	Motorways & trunk roads	
Total suspended solids	190	261	487		
	(1–4582)	(110–5700)	(347–2340)	(815–6289)	25 mg L^{-1}
BOD	11	24	59		N/A
	(0.7–220)	(12.2–32.0)	(35–172)	(90–172)	
COD	85		358		N/A
	(20–365)	(128–171	(22–703)	(181–3865)	
NH_4-N	1.45		1.76		1.7 μg L^{-1}
	(0.2–4.6)	(0.02–2.1)	(1.2–25.1)	(0.8–6.1)	
Total nitrogen	3.2		9.9		N/A
	(0.4–20.0)		(0.9–24.2)		
Total phosphorus	0.34		1.8		N/A
	(0.02–14.3)		(0.5–4.9)		
Total lead	0.21	0.96	0.83		12.26 μg L^{-1}
	(0.01–3.1)	(2.41–34.0)	(0.01–1.91)	(1.1–13.0)	
Total zinc	0.3	0.41	1.15		30 μg L^{-1}
	(0.01–3.68)	(0.17–3.55)	(0.21–2.67)		
Total hydrocarbons	1.9	28	1.8		
	(0.04–25.9)	(2.5–400)	(0.01–43.3)		
PAH	0.01	(0.03–6.0)	0.002	140	
Fecal coliforms (*Escherichia Coli*)	6430		2.1		
	(40–500 000)	$10–10^3$	(0.9–3.8)		N/A
	MPN per 100 m L^{-1}	MPN per 100 m L^{-1}	$\times 10^9$ counts ha^{-1}		

(Table compiled from: USEPA (1983) *Final Report of the Nationwide Urban Runoff Program*, Vol. 1, US EPA: Washington; USEPA (2004) *Impacts and Control of CSOs and SSOs*, Report 833-R-04-001, US EPA, Washington; Deutsch J. C. and Hemain J. C. (1984) Main results of the French National Programme of urban runoff quality measurement. *Proceedings of the 3rd International Conference Urban Storm Drainage*, Chalmers University: Gothenburg, pp. 939–946; Marsalek J. (1991) Pollutant loads in urban stormwater. *Water Pollution Research Journal Canada*, **23**(3), 360–378; House M. A., Ellis J. B., Herricks E. E., Hvitved–Jacobsen T., Seager J., Lijklema L., Aalderink, H. and Clifforde, I. T. (1993) Urban drainage: impacts on receiving water quality. *Water Science Technology*, **27**(12), 117–158; D'Arcy J. B., Ellis J. B., Ferrier R. C., Jenkins A. and Dils R. (Eds.) (2000) *Diffuse Pollution Impacts: The Environmental and Economic Impacts of Diffuse Pollution in the UK*, Terence Dalton Publication (CIWEM), Lavenham.)

of annual loads and their confidence limits (Marsalek, 1991; Ellis and Mitchell, 2005). Mean EMCs can be calculated from observations or transported from existing databases such as the US NURP database (USEPA, 1983); runoff volume is produced by hydrological modeling. The volume is then multiplied by the mean EMC to obtain the loading, and estimate bounds derived from multiplication by the upper and lower confidence limits.

Properties of EMCs are of further interest in load and impact estimations. Firstly, the US NURP data indicate that geographic location and land use were of little utility in explaining site-to-site variability, or predicting data for unmonitored sites (US EPA, 1983). Under such circumstances, best EMCs are obtained by pooling data for all sites. Secondly, EMCs were found to be statistically independent of runoff event volumes, which implies that loads can be derived by multiplication of the mean EMC by runoff volume, and, furthermore, when sampling runoff events, randomly selected events of any magnitude are acceptable. Thirdly, analysis of EMC data in the NURP program (US EPA, 1983) and in other studies indicated that stormwater constituent concentrations as well as their EMCs are log-normally distributed. This fact can be used in estimating means of censored concentration data, and in estimates of loads and quantiles (Van Buren *et al.*, 1997). Many urban water quality planning and design tasks require spatially and temporally distributed data on stormwater, municipal sewage, and combined sewage flows in urban areas, for the analysis of existing sewerage systems or planning and design of new ones. Such tasks require the use of urban simulation models, which have greatly evolved during the past 30 years. There are many such models currently in use, but several broadly used software packages stand out in the modeling practice (listed alphabetically):

InfoWorks CS (collection system) of Wallingford Software (**http://www.wallingfordsoftware.com**), MOUSE of the Danish Hydraulic Institute (**http://www.dhi.com**), and the Storm Water Management Model of the US Environmental Protection Agency (**http://www.epa.gov/ceampubl/**). These models are modularly structured, and, in general, they calculate runoff and wastewater flows, their quality, route the flows and water quality constituents through transport, storage, management, and treatment facilities, and simulate the fate of effluents in receiving waters. Modules for simulation of real-time control of sewer systems are also available. In general, the available modeling tools serve well the needs of urban modelers and facilitate easy input data import from GIS or other databases (Zoppou, 2001).

Combined Sewer Overflow (CSO) Pollutants

Table 2 shows the range of pollutant concentrations associated with global CSO discharges, which are not that dissimilar between the various countries quoted, given the variation in geographic and climatic situation for the differing data locations. The pollutants in CSOs come from domestic sources (especially BOD_5, TSS and nutrients), trade effluents (especially fats, grease, metals, and synthetic organic compounds), as well as from atmospheric washout and impermeable surface water runoff during wet weather events. Concentrations can vary substantially on a diurnal basis, both within and between stormflow events, as well as from community to community. These pollutants impact the aquatic environment in various ways, as indicated later in the article, with solids, for example, aggravating fish gill tissue, increasing turbidity, and entombing embryos in the bed gravel. Oil can severely affect wildlife through ingestion following preening as well as from loss of external waterproofing, whilst both organics and metals can induce toxic effects upon the biota.

Aesthetic pollutants such as sanitary products, toilet tissue, and faeces also characterize CSO discharges with loads, depending on the magnitude and frequency of overflow events, watershed characteristics, as well as population size and character. Floatables, including sanitary products, litter, and detritus are also characteristic of CSOs and can have an adverse impact on wildlife, primarily through entanglement or ingestion as well as having adverse aesthetic impacts. A recent concern has arisen over the incidence of sewage contaminants associated with pharmaceuticals and personal care products (PPCPs) such as chelating agents (e.g EDTA), antibiotics, antiinflammatories, steroids, and endocrine disrupters that are being found at levels well above the widely accepted $1\,\mu g\,L^{-1}$ limit (Marsalek *et al.*, 2002).

Combined sewer networks should not be regarded simply as conveyance systems as they also serve as physical and chemical reactors having the potential to alter and modify the quality of received urban runoff. The sudden flow influx into a CSO brought on by a rainfall (or snowmelt) event can create a first-flush effect, which occurs when pollutants washed from impermeable urban surfaces combine with pollutants resuspended from in-pipe sediment. Studies of sewer entry–exit mass loads have shown that exchanges with in-pipe pollutant stocks make up a principal source of wet weather flow pollutants for solids, BOD/COD, hydrocarbons, and soluble metals such that they present a prime source of acute oxygen depletion in the receiving water (Ashley *et al.*, 2004).

Construction and Urban Runoff

The increases in sediment load associated with urbanization have been well documented, and it has been suggested that construction causes 50% of the urban sediment yield with as much as 10 tonnes per capita per annum (and averaging between 116 to 157 tonnes $ha^{-1}\,year^{-1}$) being transported in receiving waters during the initial construction phases (D'Arcy *et al.*, 2000). This sediment yield declines substantially as the urban area matures such that TSS concentrations can decline (especially for small urban watersheds) below levels observed in rural catchments.

Table 2 CSO pollutant concentrations

	TSS ($mg\,L^{-1}$)	BOD ($mg\,L^{-1}$)	COD ($mg\,L^{-1}$)	Cd ($\mu g\,L^{-1}$)	Cu ($\mu g\,L^{-1}$)	Pb ($\mu g\,L^{-1}$)	Zn ($\mu g\,L^{-1}$)	P_{total} ($mg\,L^{-1}$)	N_{total} ($mg\,L^{-1}$)	*E. Coli* (100 mL)
US	237–635	43 – 95	120–560			150–290			2.9–4.8	
Canada	190							1.4	8.3	
UK	***425***	***90***				***250***	***870***	***10***	***8.3***	
	176–647	43–225	260–507			80–450	100–1070	6.5–14.0	2.1–28.5	$10^6 - 10^8$
Europe	105–721	39.9 – 200	148–530	1.1–9.6	37–170	42–450	357–1070	2.4–4.0	2.1–14.4	$10^7 - 10^8$

(Table compiled from: Ellis J. B. (1986) Pollutional aspects of urban runoff. *Urban Runoff Pollution, NATO Technical Series,* Torno H. Marsalek J. and Desbordes M. (Eds.), Springer–Verlag: Berlin, pp. 1–34; WRC (1991) *Sewer Quality Archive Data,* Report FR0203, Foundation for Water Research, Medmenham; Arnberg–Neilsen K., Hvitved–Jacobsen T., Johansen B. N., Mikkelsen P. S., Poulsen B. K., Rauch W. and Schlutter F. (2000) *Stormwater Concentrations in Foul Sewers,* Milloproject 532, Danish EPA; EPA (2004) *Impacts and Control of CSOs and SSOs,* Report to Congress, EPA 833-R-04-001, Office of Water, Washington.)

Highway construction can result in equally large increases in sediment yield. For example, in addition to recorded illegal off-site discharges, TSS concentrations of between 3235 and 20 340 mg L^{-1} and 8635 and 46 800 mg L^{-1} have been recorded respectively for the Annan and Kirtle Water over the period April 1993 to May 1995 during the construction of the M74 motorway in Scotland, UK (D'Arcy *et al.*, 2000). The cumulative effect of these nonpoint sediment discharges resulted in 3.9 km of the River Annan and 16.2 km of the Kirtle Water being downgraded from the highest water quality classification standard.

Urbanization and Surface Runoff Flows

The transformation of a watershed from a rural to an urban condition produces three major changes in the hydrological characteristics of receiving streams:

- An increase in flow volume primarily due to the reduction in infiltration following increases in impermeable surfacing with increases in runoff volume being greatest for frequently occurring small, intense storms.
- A decrease in lag time due to a combination of impervious surface runoff and the expansion of the urban drainage net, which reduces hydraulic roughness and increases the velocity of overland flow.
- An increase in peak discharge; this is the combined result of increases in runoff volume and decreases in lag time. A full urbanized watershed with some 50% impervious cover will increase the peak discharge of a 2-year storm by approximately four times.

The relative increase in 0.01 through 0.2 probability storms for runoff peaks and volumes, that is, 100 through 5-year return period storms are 1.8 to 3.0 times the runoff from undeveloped land (for example, see `http://udfcd.org/techpapers.htm`). As the recurrence probability increases to 0.5, that is, a 2-year storm, the runoff after urbanization is about 40 to 60 times or more the undeveloped rate. However, it is the annual (or 1.5-year recurrence interval flow) and more frequently occurring smaller storms that are the dominant channel forming events and which generally shape the watercourse along with being responsible for delivering the majority pollutant load to the receiving water. Further detail on flow effects and related urban drainage design arising from urbanization is also provided by Endreny in **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**.

Given the need to capture and convey large floods up to the 100 RI storm event that result from increased surface runoff, traditional engineering of urban waterways has been to employ straightened concrete-lined channels with safety walls or gabion buttresses. Flows along these watercourses are flashy and potentially hazardous to local residents, thus longitudinal slopes need to be reduced using grade control drop structures, low/trickle flow channels, on-line riffle/pool sequences, sinuous low-flow paths, and so on.

Flow and Quality Pathways

Figure 1 illustrates the various source inputs (heavy-bordered boxes), outputs (or sinks shown by dashed boxes) and pathways of water and pollutants for both natural and anthropogenic sources that are encountered within an urban catchment. There may be unintentional pathways whereby flows leave the sewer pipes via exfiltration or where elevated groundwater levels act as a source and add water into the sewer system via pipe infiltration. The former loss is generally of a small scale and less than 2 to 3% of total flow volume due to joint sealing by sediment and biofilm growth in the sewer pipe (Ellis *et al.*, 2004). The latter gains by infiltration can be much more substantial, particularly following prolonged rainfall periods that elevate the catchment water table above the level of the sewer pipe invert.

URBAN RUNOFF IMPACTS ON RECEIVING WATERS

Introduction

The various impacts of urbanization on the water cycle are independent, and they have a synergy that reinforce each other and lead to a general deterioration and loss in water use. This yields the paradox that it is urban areas, which have the greatest requirements in terms of water and aquatic amenity use, water quality, and flood protection, but are characterized by the highest flood and pollution risks and have the most highly degraded aquatic environments. Some detail on the relationship between infrastructure and resulting water quality is also given by Baker in **Chapter 188, Land Use and Water Quality, Volume 5**.

The most significant urban receiving water impacts are caused by discharges from both separately sewered SWOs and CSOs with the nature and magnitude of the impact being dependent on the characteristics of the generating watershed and the interactions with the receiving waterbody. Such impacts need to be evaluated in terms of specific characteristics at each site, including physical habitat changes, water quality changes, sediment and toxic pollutant impacts, impacts on biological communities, and groundwater impacts. Discharges of fecal bacteria also pose health risks, particularly during and immediately after wet weather events. Such physical, chemical, and biological effects will operate at varying temporal and spatial scales. Temporal scales correspond to the nature of acute (short-term) and chronic (long-term accumulative) impacts with intermittent water quality criteria normally related to exposure duration and return period of the impact-causing event

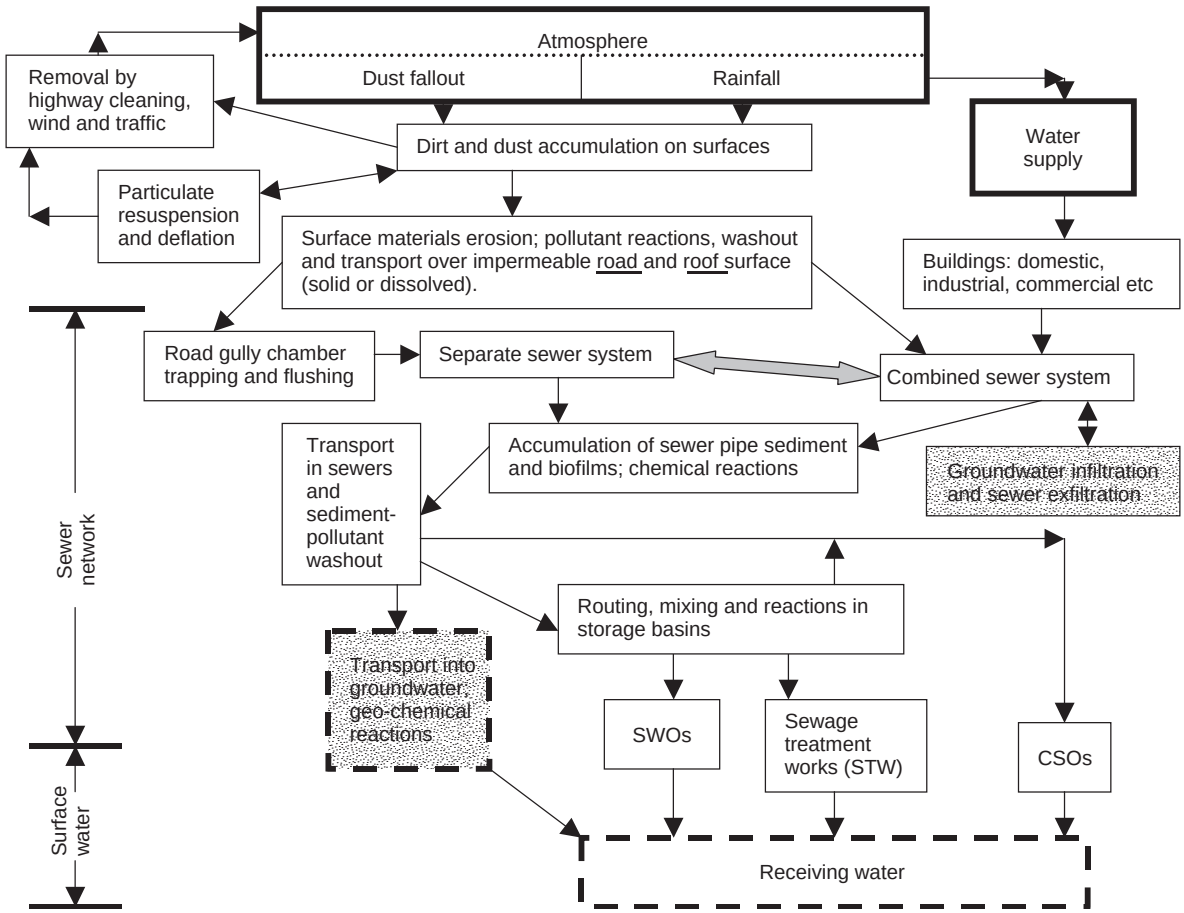


Figure 1 Urban runoff pathways

as well as the pollutant concentration. The typical recovery time after a CSO event is on the order of 5 to 7 days. However, it cannot be concluded that compliance with such criteria will provide guaranteed long-term protection as continued episodic exposure and perturbation can lead to a permanent weakening of the aquatic ecosystem and prevent ultimate recovery.

Receiving Water Impacts

Physical habitat changes Urbanization can permanently modify the nature, form, habitats, and behavior of receiving water bodies that are frequently "canalized" or heavily modified to contain the flood channel and improve storm flow conveyance. Such regulated channels will have altered fluvial dynamics typified by increased sedimentation and high erosion potential, which, in turn, influences stream morphology and channel characteristics as well as in-stream habitat and substrate conditions. Bed sedimentation also limits exchange between surface and underground waters across the hyporheic zone.

Water quality changes Dissolved oxygen (DO) depletion from intermittent urban discharges is a well recognized phenomenon with the soluble organics transported in the water phase exerting an immediate DO depletion with the scouring effect of increased flow on basal sediments adding to this effect, which can be further exacerbated by the presence of ammonia. Settleable solids accumulate on the bed and can result in delayed DO depletion due to an increase in the sediment oxygen demand (SOD) as well as facilitating anoxic conditions under ice cover during winter months. Pollutants discharged from both CSOs and SWOs contribute a range of adsorbable and settleable pollutants derived from sewer deposits, wastewater effluents, and urban surfaces. Owing to the nature and amount of biodegradable organics, anaerobic conditions may prevail in receiving water sediments and accumulated metals, hydrocarbons, and bacteria can then impose long-term, chronic impacts on the sediment community. Approaches to multidimensional water quality modeling of waterbodies receiving polluting discharges are outlined by Lin and Falconer in **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**.

Ecological changes The generic characteristics of urban receiving water ecology are habitat instability and ecotoxicity. The urban stream is dominated by taxa, which can tolerate successive erosional–depositional sequences and transient, low-quality food sources with limited leaf decomposition and short retention times for organic matter. Numerous studies have demonstrated the adverse biotic effects resulting from episodic urban discharges with suppression of ecological diversity occurring downstream of outfalls. The analysis of ecological diversity and associated community structures, together with benthic toxicity testing, provide powerful tools for assessing urban runoff impacts (Rochfort *et al.*, 2000; Ellis, 2000). However, many urban drainage studies have failed to demonstrate with any statistical certainty that water concentrations downstream of SWOs are any more toxic than upstream of the discharge. Acute toxicity tests undertaken by the Canadian National Water Research Institute (Marsalek *et al.*, 1999) on 58 stormwater and 65 CSO samples showed the majority to be nontoxic or only potentially toxic at 67% and 93% for stormwater and CSOs respectively. Similar results have been obtained in studies of highway runoff in the United Kingdom (Moy *et al.*, 2003).

The inhibition of acute toxicity observed by many stormwater studies may simply reflect the pollutant-complexing and binding effects, which occur in the presence of organic rich effluent. Genotoxicity and longer-term chronic toxicity may present more severe problems for urban waterbodies receiving stormwater runoff and CSO discharges. Bioavailability can be locally enhanced by sediment organic carbon content and pH as well as particle size, with interstitial waters being the principal route of uptake for sediment-associated contaminants. It may be the cumulative and interactive effects of water and toxic sediment quality as well as fluctuating flow and physical habitat constraints, which collectively lead to patterns of reduced biotic community status and diversity in urban receiving waters.

Public health risks The design of CSOs and SWOs means that untreated sanitary waste and contaminated effluents discharge to urban receiving waters, and it is widely recognized that urban runoff contains a wide variety and frequently high numbers of pathogenic bacteria and viruses which raise potential public health risks. During 2002, 21% of US beach closures were due to bacterial discharges associated with stormwater runoff in comparison to CSOs, which were responsible for only 1% of closures (US EPA, 2004).

Groundwater impacts There is little clear evidence of any substantial or widespread impact of urban runoff or sewer exfiltration on groundwaters (Ellis *et al.*, 2004), although both winter salting and herbicide applications in urban areas can generate levels above the drinking water standard in adjacent groundwaters.

Aesthetic deterioration Research into the public's perception of urban receiving water quality and the potential for the sustainable management of water uses for amenity, recreation, and nature conservation has shown that they generally perceive most urban rivers as being polluted, even when the chemical and ecological quality may be acceptable (House, 1996).

INTEGRATED URBAN DRAINAGE MANAGEMENT

Concepts and Main Issues

Urban discharges may cause numerous adverse effects on receiving waters with the impacts being exacerbated by traditional drainage systems and end-of-pipe solutions, which often appear to be expensive and inefficient. Increased concerns about such impacts have led to the development of new solutions based on the general concept of integrated urban water management (Figure 2), which provides a holistic integration of flood protection, water supply management and protection, groundwater quality, wastewater management, and receiving water quality. Such total urban water cycle management must be firmly linked to the question of urban sustainability (Lawrence *et al.*, 1999; Marsalek and Chocat, 2002). The emerging issue is probably to minimize the impacts of construction through better planning and design of the urban development itself. This idea is central in the concepts of low impact development (LID), sustainable development design (SDD), and "smart growth" planning (SGP). A characteristic of SDD/SGP smart growth development is a reduced footprint, leaving intervening and adjoining land available for open space, habitat development, and off-site drainage controls such as wetlands or detention basins.

Different Kinds of Actions

Sustainability implies an equilibrium among the three sets of demands; the needs of environmental protection, economic needs, and the needs of the society. So far, most attention has focused on environmental needs (e.g. attenuation of increased flows, sediment exports, chemical and bacteria fluxes). Relatively little is known about the economic and social aspects of new water and wastewater management systems needed to facilitate a full development of urban water resources to meet the needs of society. However, the introduction of sustainable principles for future urban developments being required by many regional and national planning administrations is driving new agendas and approaches for strategic urban infrastructure including the implementation of alternative drainage designs and integrated water resource management approaches, as indicated in Figure 2.

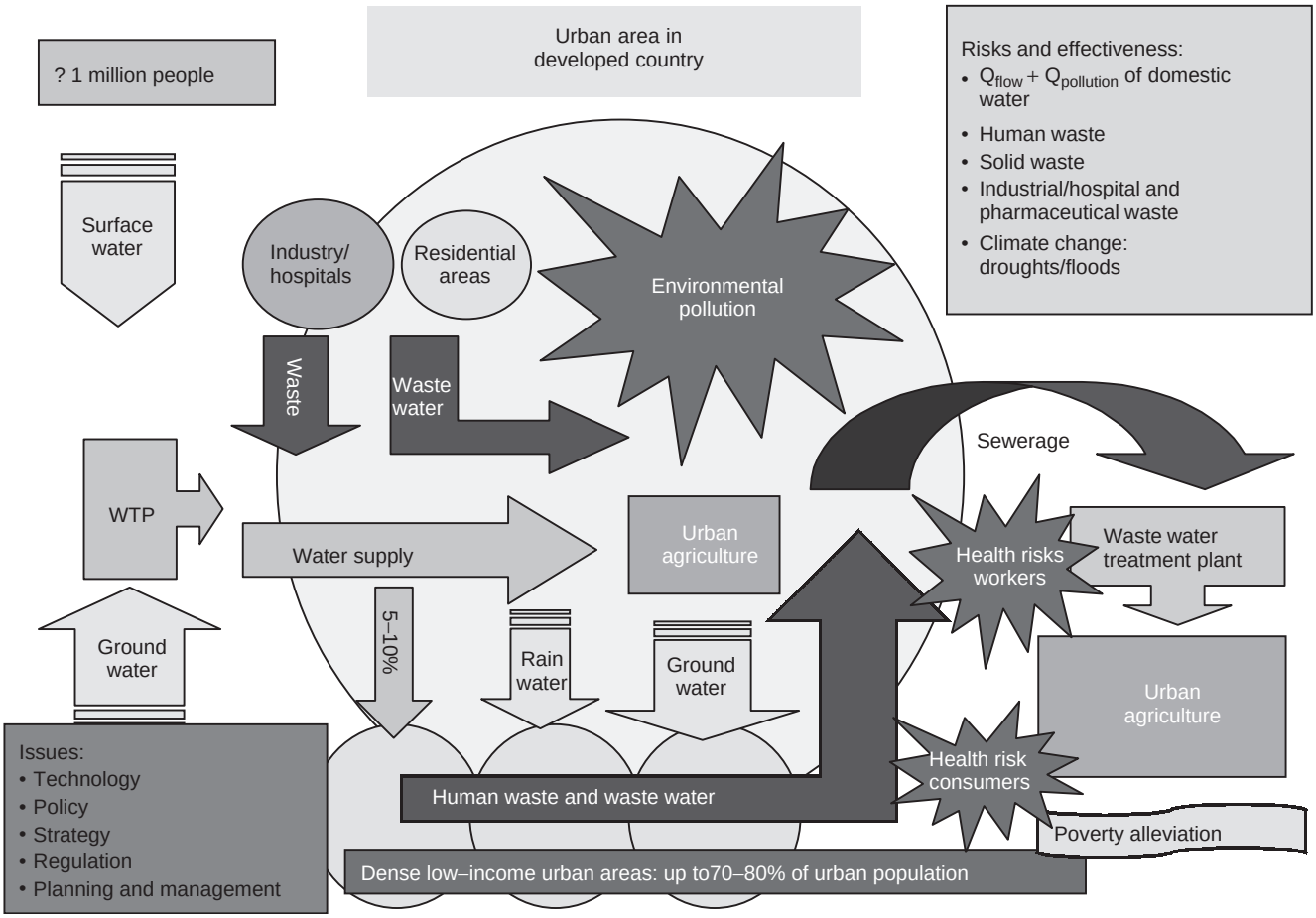


Figure 2 Sustainable urban drainage management. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Integrated Urban Water management implies actions at four levels, which can be referred to as follows:

1. *Policies and nonstructural controls*: These proactive measures are generally highly cost-effective, and for that reason are considered in all stormwater management plans and include public awareness/education/participation; urban development planning; management of material use, exposure, and disposal controls; spill prevention and cleanup; prevention/elimination of illegal dumping and illicit connections; and street and stormwater facilities maintenance (ASCE, 1998). These approaches require a close cooperation between planners, drainage designers, and community stakeholders from the early stages of land development.
2. *Best Management Practices* (BMPs) *for stormwater control and treatment*: These include a variety of reactive structural source and site controls offering cost-effective flow and quality control performance (**www.bmpdatabase.org**; **www.wsud.org.au**) as well as potential ecological and amenity benefits (**www.ciria.org.uk/suds**):

- Lot-level source controls; such measures include enhanced rooftop detention, flow restrictions at catchbasins to enhance local storage/detention, measures to slowdown runoff flow and enhance infiltration along with implementing stormwater harvesting and reuse.
- Local stormwater storage either on roofs or in small cisterns or reservoirs.
- Biofiltration by grass filters, swales, and pocket wetlands; these measures reduce runoff volume by infiltration and enhance runoff quality by such processes as settling, filtration, adsorption, and biouptake resulting in TSS reductions of at least 50%.
- Infiltration facilities; these BMPs serve to reduce the volume and rate of runoff, reduce pollutant transport and recharge groundwater.
- Permeable and porous pavements; introduced within urban areas in order to reduce runoff from impermeable surfaces. Total outflow TSS

concentrations for such structures are typically 10 to 20 mg L^{-1} with both solid metals and organics being reduced by 60 to 80%.

- Water quality inlets, which provide some stormwater treatment by sedimentation and skimming of floatables and oil. French and UK experience with these systems indicates very low effectiveness, except for interception of accidental oil spills with low pollutant removal rates and release of captured pollutants being reported (Bardin *et al.*, 2001).
- Filters; stormwater sand filters have been introduced in the United States with considerable success, although reports from Australia and New Zealand have been less encouraging. They are effective in removing pollutants (Urbonas, 1994), but, to maintain their effectiveness, they may have to be back-washed regularly and the risk of clogging should be reduced by stormwater pretreatment.

3. *Community-level BMPs:*

- Community infiltration facilities; these facilities comprise infiltration trenches and basins of somewhat larger scales than those provided at the site level.
- Stormwater management ponds; stormwater ponds (or wet retention basins) are used widely in Australia, Canada, Western Europe, and the United States to provide various types of controls, including flow control (reduction of flow peaks), sedimentation, and removal of dissolved pollutants by marginal aquatic plants. Outflow pollutant concentrations from these facilities are normally a function of the influent concentrations, but reductions of an order of magnitude are feasible for solids and solid-associated pollutants. Ponds may accumulate large quantities of contaminated sediments, which may be polluted with heavy metals and persistent organic pollutants, including polycyclic aromatic hydrocarbons (PAHs). Both metals and PAHs in deposited sediments may be released into the water column in response to changes in the water quality and flow-through rates.
- Constructed wetlands; wetland BMPs provide stormwater detention and treatment by various processes, including filtration, infiltration, and biosorption, and serve as cost-effective treatment systems for both particulate and dissolved pollutants.
- Extended detention (dry) basins; such basins are widely applicable and can provide stormwater settling in those areas where it is difficult to maintain wet facilities.

These BMPs are often used within multiple (treatment-train) systems. In these hybrid systems, two or more BMPs may be stacked vertically or in a series, to increase the system performance or reliability, or to reduce the maintenance. Such multiple treatment train systems are rapidly becoming the norm in new greenfield and brownfield urban developments as they provide an effective approach to full effluent treatment.

4. *Watershed-level Measures*: The watershed or catchment is a logical unit for water and wastewater management planning and forms the basis, for example, for most European drainage regulation and management. Urban drainage and stormwater management strategy should be included in watershed plans, developed in a hierarchical manner, using an ecosystem approach and providing a basis for the development of more detailed drainage management plans. Yet, in many countries, such an approach is difficult to implement efficiently since organizations involved in urban management are frequently different from those involved in watershed management. The development of new strategies or technologies is strongly impeded by economic problems (costs, financing), sociological problems (acceptance by the public, fragmentation of duties and responsibilities), urban planning challenges (integration into the landscape), organizational cooperation, problems with policies and regulations, and so on. Nevertheless, such integrated, source-control strategies represent a sustainable approach for both developed and developing countries, enabling the adverse effects of urbanization to be addressed at an affordable cost.

WATER AND WASTEWATER REUSE

Background

The concept of total water cycle management in urban areas provides a logical context for water reuse and recycling (Lawrence *et al.*, 1999). The extent to which such measures are practiced depends on water availability, economic incentives, regulatory feasibility, and public acceptance. Reuse can be either direct (reclaimed water is transported to the points of reuse), or indirect, whereby reclaimed water is first discharged into receiving waters or aquifers and then reused (see Figure 2).

Wastewater Effluent Reuse

Stormwater Reuse: Rainwater/stormwater reuse is currently practiced in many countries as a result of the widespread use of stormwater management, which often involves various forms of rainwater/stormwater reuse and thus provides double benefits – mitigation of runoff and its pollution, and provision of subpotable water supply. Typical examples of stormwater reuse include collection and reuse of residential and commercial roof runoff for irrigation or toilet flushing, collection of roof water from dome stadiums for toilet

flushing and landscape irrigation, and treated stormwater reuse for industrial processes, boiler feed, or cooling waters. In stormwater reuse, the most feasible source appears to be roof runoff, which represents the source with the best water quality; other sources of stormwater, particularly runoff from streets and highways, may be too polluted and expensive to treat for reuse. Even in the case of roof runoff, there are concerns (Eriksson *et al.*, 2002) about its quality, mostly due to heavy metals (from roofing materials, depending on the rainwater pH), chemicals in dry atmospheric deposition (depending on local or remote sources and air transport), and fecal bacteria (bird droppings). Collected roof water is usually treated, using such processes as filtration, screening, settling, and UV disinfection.

Greywater Reuse: In the management of domestic wastewater, one of the options receiving much attention in recent years is the at-the-source separation into two separate flows; blackwater, or toilet waste, and greywater representing all remaining household wastewater. Greywater has been studied as an alternative source of water for nonpotable applications, including irrigation, and toilet flushing. Examples of greywater reclamation and reuse include subsurface irrigation, greywater treatment, and disinfection for toilet flushing, greywater reuse in experimental housing, and a major in-building water recycling scheme at the Millennium Dome (UK), where the system provided 55% of the water demand at the site in the form of greywater from washroom sinks, rainwater from the Dome roof, and groundwater (Hills *et al.*, 2002).

Unseparated Wastewater Reuse: General wastewater reclamation and reuse has been called "the greatest challenge of the twenty-first century" (Asano, 2002). It has the potential to bring about two great benefits – (i) provide a reliable source of water and (ii) keep wastewater pollutants out of receiving waters. Basic principles of wastewater reuse include three underpinning principles (i) providing reliable treatment corresponding to the intended reuse, (ii) protecting public health, and (iii) winning public acceptance (Asano, 2002). The approach taken to wastewater reuse depends on the intended category of reuse, with seven types of reuse commonly practiced: (i) agricultural irrigation, (ii) landscape irrigation, (iii) groundwater recharge, (iv) industrial process water, (v) environmental and recreational uses, (vi) subpotable urban uses, and (vii) indirect or direct potable reuse.

Recharge and Direct Reuse

Groundwater recharge is another large-scale application practiced by spreading/infiltration basins or direct injection to groundwater aquifers. The ideal soils for soil-aquifer treatment (SAT) balance rapid recharge (i.e. a coarse-textured soil) with efficient contaminant adsorption and removal. Recreational and environmental (ecological) uses involve nonpotable uses related to land-based water features such as the development of recreational lakes, wetlands, and stream augmentation (Asano, 2002). Nonpotable urban uses include fire protection, air-conditioning, toilet flushing, construction water, flushing of sanitary sewers, heat source or sink (in heating, air-conditioning or snowmelting), snow making, and landscape irrigation.

The most challenging category of wastewater reuse is potable reuse, practiced either by replenishment of water supply storage, or by direct input of highly treated reclaimed water into the water distribution system. Although direct reuse has been demonstrated in the City of Windhoek, Namibia (Harhoff and van der Merwe, 1996), similar applications in industrial countries are highly unlikely, mostly because of the lack of public acceptance and concerns about a safe and complete removal of new chemicals of concern (e.g., endocrine disruptors, pharmaceuticals, personal care, and therapeutic products) and pathogens from the reclaimed water (Marsalek *et al.*, 2002).

FUTURE PERSPECTIVES

Today it is generally accepted that urban surface water, groundwater, and wastewater should be considered in relation to each other and to their interactive impacts on water flows, receiving water pollution and aquatic ecology. This is recognized in the very high investment effort being made in many countries to reduce urban flooding and pollution risks and to provide sustainable urban drainage systems. Since the passage of the US Clean Water Act in 1972, the EPA, states and local water pollution control agencies have undertaken numerous actions and initiatives to reduce both CSO and SWO impacts. Some $11 billion has been invested since 1998 on injunctive relief schemes and $75 million on urban drainage improvement schemes (US EPA, 2004). The Urban Wastewater Treatment Directive and forthcoming Water Framework Directive within Europe is likewise driving stepwise improvements for urban drainage infrastructure with capital investment within the United Kingdom alone reaching £2.8 billion over the 2000–2005 period.

However, in any analysis of future trends and drivers affecting urban water systems, sustainable development is only one, albeit important, influencing factor. Population and demographic trends are of equal significance. The number of megacities with greater than 10 million inhabitants is expected to increase to over 20 by 2050 (with 80% being located in developing countries), with some 70% of the world population living in urban areas. Inevitably such large-scale urbanization has severe implications for urban water conditions and requirements, particularly given that on a global scale, only about 15% of wastewater is treated (**www.thewaterpage.com**). In western cities,

although major demographic shifts including ageing populations and falling average household occupancy are likely to keep population levels relatively stable, such trends are also likely to lead to higher per capita water and wastewater infrastructure demands. The rise of consumerism, individualism, and increased disposable incomes in western societies will also generate new drivers for future water and wastewater resource supply and management as exemplified by the substantial rise in bottled water supplies, water-using domestic appliances, separated waste streams and waste recycling schemes.

A major influencing factor on future urban water resource management will be climatic change. Such change will have major implications for water resources, urban flooding, and receiving water pollution as predicted in the recent UK government Foresight Future Flooding report (OST, 2004). This analysis of urban drainage risk was conducted within the context of scenarios of differing socioeconomic futures, as depicted in Figure 3. The vertical dimension shows the system of governance, ranging from autonomy where power remains at the regional/national level to globalization or interdependence where power increasingly moves to other institutions such as the European Union. The horizontal dimension in the figure shows social values, ranging from individualistic values to community-oriented values. Central to the scenario identification is the recognition that different mechanisms of state and market regulation will influence decision-making. The driver with the greatest influence on future urban drainage was precipitation and its spatial and temporal change. Other important physical drivers were identified as urban creep, increases in groundwater infiltration, and receiving water pollution in addition to regulation, public attitudes to flooding and the ability/willingness to pay for future infrastructure improvements.

Chocat *et al.* (2004) have proposed the evolution of four possible future scenarios for urban drainage. The "green" scenario, dominated by decentralized source-control approaches with minimum sewer connections and extensive wastewater recycling and water conservation, is equivalent to the community-based, local stewardship quadrant of Figure 3. The conservative "technocratic" scenario adopting centralized advanced technology, monopolistically retained, and managed within the public sector, reflects the regionalized, national enterprise socioeconomy of Figure 3. The "privatization" scenario is clearly consumer and world market oriented and represents the predominant strategy operating within many developing countries at present. The final scenario suggested by Chocat *et al.* (2004) is that of "business-as-usual" which stumbles between the technocratic tradition and green ideas as well as picking up varying degrees of privatization.

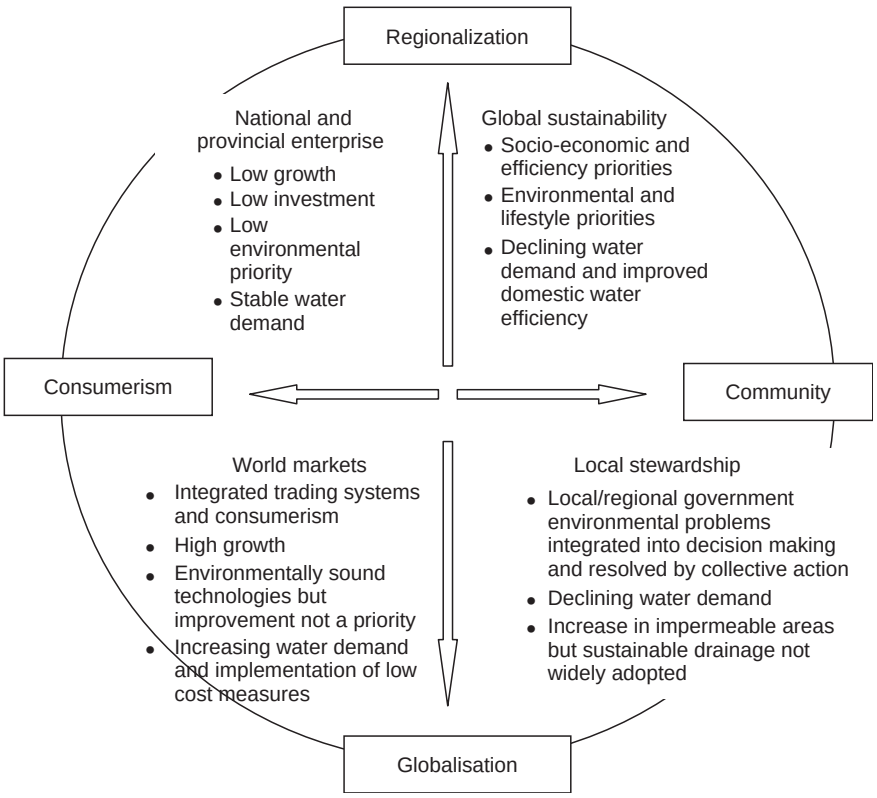


Figure 3 Socioeconomic futures

The scenario analysis approach provides no firm indication of which future might be more probable than another, although it is important to note that technical solutions do not map clearly to particular socioeconomic futures and that the high existing sunk asset value of urban drainage means that there is always likely to be considerable inertia and conservatism in the water industry. Thus, it is feasible to visualize a "no-change" (or little change) future scenario with urban water resource managers having a lack of control with respect to land use planning and chemical usage as well as being underfunded.

Alternative perspectives have been widely canvassed, with some indicating a complete revolution in future urban water systems in terms of new paradigms, new contexts, and new methodologies (Maksimovic and Tejada-Guibert, 2001). This thinking is based on the introduction of multidisciplinary, integrated approaches to urban drainage incorporating sustainable principles, the adoption of network, risk, and vulnerability analysis, complex modeling, incorporation of educational and social values as well as anticipatory and contingency scenarios for addressing new impacts such as climate change.

FURTHER READING

Asano T. (Ed.) (1998) *Wastewater Reclamation and Reuse*, Water Quality Management Library, Vol. 10, Technomic Publishing Company: Lancaster.

Butler D. and Davies J.W. (2000) *Urban Drainage*, E & F N Spon: London.

Herricks E.E. (Ed.) (1995) *Stormwater Runoff and Receiving Systems*, CRC Lewis Publishers: Boca Raton.

Makepeace D.K., Smith D.W. and Stanley S.J. (1995) Urban stormwater quality: summary of contaminant data. *Critical Reviews in Environmental Science and Technology*, **25**, 93–139.

Marsalek J. (1998) Challenges in urban drainage. In *Hydroinformatics Tools for Planning, Design, Operation and Rehabilitation of Sewer Systems*, *NATO ASI Series, 2. Environment – Vol. 44*, Marsalek J., Maksimovic C., Zeman E. and Price R. (Eds.), Kluwer Academic Publishers: Dordrecht/Boston/London, pp. 1–23.

Moffa P.E. (Ed.) (1990) *Control and Treatment of Combined Sewer Overflows*, Van Nostrand Reinhold: New York.

Rowney A.C., Stahre P. and Roesner L.A. (Eds.) (1997) *Sustaining Urban Water Resources in the 21st Century*, American Society of Civil Engineers: Reston.

Torno H., Marsalek J. and Desbordes M. (Eds.) (1986) *Urban Runoff Pollution*, *NATO Technical Series*, Springer-Verlag: Berlin.

Urbonas B. (Ed.) (2001) Linking stormwater BMP designs and performance to receiving water impacts mitigation. *Proceedings of Engineering Foundation Conference*, American Society of Civil Engineers: Snowmass.

REFERENCES

Asano T. (2002) Water from (waste) water – the dependable water resource. *Water Science and Technology*, **45**(8), 23–33.

ASCE (1998) Urban runoff quality management. *WEF Manual of Practice No. 23, ASCE Manual and Report on Engineering Practice No. 87*, American Society of Civil Engineers: Reston.

Ashley R.M., Bertrand-Krajewski J.L., Hvitved-Jacobsen T. and Verbanck M. (2004) *Solids in Sewers*, Scientific and Technical Report No. 14, IWA Publishing, London.

Bardin J.P., Gautier A., Barraud S. and Chocat B. (2001) The purification performance of infiltration basins fitted with pretreatment facilities: a case study. *Water Science and Technology*, **43**(5), 119–128.

Burian S.T., Nix S.J., Pitt R. and Durrans S.R. (2000) Urban wastewater management in the United States: past, present and future. *Journal of Urban Technology*, **7**(3), 33–62.

Chocat B., Ashley R., Marsalek J., Matos M.R., Rauch W., Schilling W. and Urbonas B. (2004) Urban drainage; out-of-sight-out-of mind. *Proceedings NOVATECH 5th International Conference, Sustainable Techniques and Strategies in Urban Water Management*, GRAIE: Lyon, pp. 1659–1690.

D'Arcy B.J., Ellis J.B., Ferrier R.C., Jenkins A. and Dils R. (2000) *Diffuse Pollution Impacts*, Terence Dalton Publishers: Lavenham.

Ellis J.B. (2000) Risk assessment approaches for ecosystem response to transient pollution events in urban receiving waters. *Chemosphere*, **41**, 85–91.

Ellis J.B. and Mitchell G. (2005) Urban diffuse pollution: key management issues for the water framework directive. *Journal Chartered Institution Water and Environmental Management*, (In press).

Ellis J.B., Revitt D.M., Blackwood D.J. and Gilmour D.J. (2004) Leaky sewers: assessing the hydrology and impact of exfiltration in sewers. In *Hydrology in the 21st Century*, Maksimovic C. (Ed.), *Proceedings BHS National Conference*, Imperial College, London, IAHS Press: Wallingford, pp. 266–271.

Eriksson E., Auffarth K., Henze M. and Ledin A. (2002) Characteristics of grey wastewater. *Urban Water*, **4**(1), 85–104.

Harhoff J. and van der Merwe B. (1996) Twenty-five years of wastewater reclamation in Windhoek, Namibia. *Water Science and Technology*, **33**(10–11), 25–35.

Hills S., Birks R. and McKenzie B. (2002) The millenium dome "Watercycle" experiment to evaluate water efficiency and customer perception at a recycling scheme for 6 million visitors. *Water Science and Technology*, **46**(6–7), 233–240.

House M.A. (1996) Public perception and water quality management. *Water Science and Technology*, **34**, 25–32.

Lawrence A.I., Ellis J.B., Marsalek J., Urbonas B. and Phillips B.C. (1999) Total urban water cycle based management. In *Proceedings 8th International Conference on Urban Storm Drainage*, Joliffe I.B. and Ball J.E. (Eds.), Australia Society of Engineers: Sydney, pp. 1142–1149.

Maksimovic C. and Tejada-Guibert J.A. (Eds.) (2001) *Frontiers in Urban Water Management: Deadlock or Hope?* IWA Publishing: London.

Marsalek J. (1991) Pollutant loads in urban stormwater; review of methods for planning level estimates. *Water Resources Bulletin*, **27**(2), 283–291.

Marsalek J. and Chocat B. (2002) International report: stormwater management. *Water Science and Technology*, **46**(6/7), 1–17.

Marsalek J., Rochfort Q., Mayer T., Servos M., Dutka B.J. and Brownlee B. (1999) Toxicity testing for controlling urban wet-weather pollution; advantages and limitations. *Urban Water*, **1**, 91–103.

Marsalek J., Schaefer K., Exall K., Brannen L. and Aidun B. (2002) *Water Re-use and Recycling*, Canadian Council of Ministers of the Environment, Winnipeg, Manitoba, CCME Linking Water Science to Policy Workshop Series, Report No. 3, p. 39.

Moy F., Crabtree R.W. and Simms T. (2003) *The Long Term Monitoring of Pollution from Highway Runoff*. R&D Technical Report P2-038/TR1, Environment Agency, Bristol.

Rochfort Q.J., Grapentine L., Marsalek J., Brownlee B., Reynoldson T., Thompson S., Milani D. and Logan C. (2000) Using benthic assessment techniques to determine combined sewer overflow and stormwater impacts in the aquatic ecosystem. *Water Quality Research Journal of Canada*, **35**(3), 365–397.

OST (2004) *Foresight Flood and Coastal Defence Project*, Office of Science and Technology, Office Deputy Prime Minister: London.

Urbonas B. (1994) Assessment of stormwater BMPs and their technology. *Water Science and Technology*, **29**(1–2), 347–353.

US EPA (1983) *Results of the Nationwide Urban Runoff Program*, Volume I: Final Report PB84-185552, Water Planning Division, US Environmental Protection Agency, Washington.

US EPA (2004) *Impacts and Control of CSOs and SSOs*, Report to Congress, EPA 833-R-04-001, Office of Water, Environmental Protection Agency: Washington.

Van Buren M.A., Watt W.E. and Marsalek J. (1997) Application of the log-normal and normal distributions to stormwater quality parameters. *Water Research*, **31**(1), 95–104.

Zoppou C. (2001) Review of urban stormwater models. *Environmental Modeling and Software*, **16**, 195–231.

98: Pathogens

CHARLES P GERBA

Department of Soil, Water and Environmental Science, University of Arizona, Tucson, AZ, US

Water plays a major role in human disease by serving as the vehicle for the transmission of disease-causing microorganisms (pathogens), and as a habit for some of their insect hosts. Worldwide, it is estimated that a child dies every 8 s from water-related diseases. New pathogens, which are transmitted by water, are recognized on a continuing basis providing challenges to both watershed management and water treatment technologies. Microorganisms transmitted by the fecal-oral route or waterborne pathogens are the most widespread since many may occur in the feces of both infected humans and animals. All surface waters can be expected to contain waterborne pathogens at one time or another. Feces from animals and sewage discharges contaminate surface waters. Groundwater may be contaminated from infiltration from land disposed human and animal wastes or water contaminated with these wastes, septic tanks, and unlined landfills. Contamination of surface and groundwater is always greatest after rainfall events due to the flushing of accumulated animal feces and infiltration of soils. Waterborne pathogens may survive for days to months in the environment. Survival of pathogens in the water environment is largely dependent on temperature and sunlight exposure.

INTRODUCTION

Disease-causing microorganisms are referred to as *pathogens*. A very small number of microorganisms are capable of causing disease in man and animals, but they have a major impact on human health. Pathogenic microorganisms are still the leading cause of death worldwide, and pathogens in which water is related to their spread are major contributors to mortality and morbidity. The World Health Organization estimated in 1996 that every 8 s a child died from water-related disease and that each year more than five million people died from illness linked to unsafe drinking water or inadequate sanitation (WHO, 1996). Waterborne outbreaks of disease continue to occur as well in the developed world. New waterborne pathogens have been discovered almost yearly over the last 20 years, while others have been introduced into the developed world from increased globalization of travel and food resources. This has created an increased awareness of the importance of watershed and water treatment management for their control. Control of diseases spread by water will continue to be a major challenge in the foreseeable future.

Waterborne diseases are those transmitted through the ingestion of contaminated water (Table 1). They are usually excreted in the feces of infected persons in large numbers and can survive in the water environment from a few days to many weeks. Thus, waterborne diseases are transmitted through the fecal-oral route, from human to human or animal to human. Waterborne diseases include the agents of cholera, typhoid fever, and infectious hepatitis.

Water-based diseases are caused by pathogens that either spend all (or essential parts) of their lives in water or depend upon aquatic organisms for the completion of their life cycles. Examples of such organisms are the parasitic helminth, Schistosoma and the bacterium *Legionella pneumophila*, which cause schistosomiasis and Legionnaires diseases respectively.

Water-related diseases, such as West Nile, yellow fever, St. Louis encephalitis, dengue, and malaria, are transmitted by insects that breed in water (e.g. mosquitoes that carry malaria) or live near water (e.g. the flies that transmit the filarial infection onchocerciasis). The major waterborne and water-based pathogens and diseases they cause are shown in Tables 2–5. Viruses consist solely of nucleic acid (which contains the genetic information) surrounded by a protective protein coat called a *capsid*. The nucleic acid may be either ribonucleic acid (RNA) or deoxyribonucleic acid (DNA). They cannot grow outside of living cells

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Type of water-related illnesses associated with pathogens

Type	Cause	Example
Waterborne	Pathogens that originate in fecal material and are transmitted by ingestion	Cholera, typhoid fever, *Giardia*
Water-based	Organisms that originate in water or spend part of their life cycle in aquatic animals and come in direct contact with humans in water or by inhalation	Schistosomiasis, *Legionella*
Water-related	Organisms with life cycles associated with insects that live or breed in water	Yellow fever, West Nile virus

Table 2 Major waterborne enteric viruses

Virus group	Serotypes or types	Some disease caused
Enteroviruses Poliovirus	3	Paralysis, fever, aseptic meningitis
Coxsackievirus	29	Herpangia, paralysis, aseptic meningitis, heart disease, fever
Echovirus	34	Respiratory illness, aseptic meningitis, diarrhea, heart disease, rash
Enteroviruses (68–71)	4	Meningitis, eye infection, respiratory illness
Astroviruses	5	Diarrhea
Hepatitis E	1	Hepatitis
TT virus hepatitis	1	Hepatitis
Reoviruses	3	Respiratory illness
Rotaviruses	5	Diarrhea
Adenoviruses	49	Respiratory disease, eye infections, diarrhea

(animals, plants, bacteria) and do not need food to survive. Viruses transmitted by the fecal-oral route are capable of surviving for long periods of time in water. Viruses that infect bacteria are called *bacteriophages*, and those bacteria that infect coliform bacteria are called *coliphages*. Bacteria are single-celled organisms surrounded by a membrane and cell wall. Bacteria that grow in the human intestinal or gastrointestinal (GI) tract are referred to as *enteric bacteria*. Enteric bacterial pathogens usually cannot survive for prolonged periods of time in the environment. Protozoa are single-celled animals. Protozoan parasites that infect humans are capable of producing environmentally resistant cysts or oocysts. These cysts or oocysts have very thick walls, which make them very resistant to disinfection. Helminths (literally worms) are multicellular animals that parasitize humans. They include roundworms, hookworms, tapeworms, and flukes. Helminths lay eggs that are excreted in the feces of infected persons. These eggs are capable of prolonged survival in water and soil.

Table 3 Major waterborne bacteria

Bacterium	Major disease
Salmonella typhi	Typhoid fever
Salmonella paratyphi	Paratyphoid fever
Salmonella spp	Diarrhea
Shigella	Bacillary dysentery
Vibrio Cholerae	Cholera
Pathogenic *E. coli*	Diarrhea, Hemopytic uremic syndrome
Yersinia enterocolita	Diarrhea
Campylobacter jejuni	Diarrhea
Helicobacter pylori	Common ulcer
Leptospira	Leptospirosis (Weil's disease)

Table 4 Major water-based pathogens

Organism	Major disease
Legionella pneumophila	Acute respiratory illness (Legionnaires' disease) Pontiac fever
Mycobacterium avium	Respiratory illness
Pseudomonas aeruginosa	''Swimmer's ear'' skin infections
Naegleria fowleri	PAM[a]

[a]PAM – Primary amoebic meningoencephalitis.

TYPES OF PATHOGENS

Viruses

More than 140 different types of viruses are known to infect the human intestinal tract and are excreted in the feces. Viruses that multiply in the intestines are referred to as *enteric viruses*. Some enteric viruses are capable of replicating in other organs such as the liver, heart, eye, skin, respiratory tract, and nerve tissue. For example, hepatitis

Table 5 Some characteristics of waterborne/water-based transmitted parasites

Organism	Infective form	Mechanism of transmission	Reservoirs
Giardia lamblia	Cyst	Person – person Waterborne Foodborne	Humans, beavers, muskrats, voles
Cryptosporidium parvum	Sporulated oocyst	Person–person Waterborne Foodborne	Many vertebrates, especially cattle
Entamoeba histolytica	Cyst	Person–person Waterborne Foodborne	Usually humans (potentially pigs, primates, and dogs)
Naegleria fowleri	Trophozoite	Trophozoite swims up nasal cavity	None: free living in aquatic or soil environment
Cyclospora cayetanesis	Sporulated oocyst	Waterborne Foodborne	None known
Enterocytozoon bieneusi	Spore	Fecal-oral	None known
Encephalitozoon hellem	Spore	Urine-oral	None known
Encephalitozoon cuniculi	Spore	Fecal/urine-oral	Laboratory rabbits, rodents, dogs
Encephalitozoon intestinalis	Spore	Fecal/urine-oral	None known
Toxoplasma gondii	Sporulated oocyst	Oral ingestion from soil or litterbox (oocyst) Undercooked meats (tissue cysts)	Cats: definitive host humans, sheep, goats, pigs, cattle, birds, intermediate hosts
Ascaris lumbricoides	Embroynated egg	Oral from soil contact	Humans
Tricuris Trichiura	Embroynated egg	Waterborne Foodborne	Humans
Necator americanus	Filariform	Skin penetration	Humans
Schistosoma mansori	Cercariae	Penetrate skin	Humans Intermediate host: snail

A virus (HAV) infects the liver, causing hepatitis. Enteric viruses are generally very host-specific; therefore, human enteric viruses cause disease only in humans and sometimes in other primates.

Enteroviruses have been the most studied waterborne viruses because of the ease at which they can be grown in the laboratory. The more common enteroviruses include poliovirus (3 types), Coxsackieviruses (30 types), and the echoviruses (34 types). Although these pathogens are capable of causing a wide range of serious illness, most infections are mild. Infectious viral hepatitis is caused by HAV and hepatitis E (HEV). These viruses are very common causes of infections in the developing world. HEV has been associated with large waterborne outbreaks in Asia and Africa, and cases of high mortality in pregnant women (20 to 30% of those ill). Noroviruses and Saporoviruses are members of the calicivirus family. These viruses are now believed to be the major cause of adult viral diarrhea in the world. They are believed to be the major cause of food-borne diarrhea and have been responsible for numerous outbreaks of waterborne diseases worldwide. Human caliciviruses are only known to infect humans and have never been grown in the laboratory. They cause a diarrhea lasting from one to three days with a low mortality. Rotaviruses are the major cause of acute diarrhea in children under two years of age and a major cause of childhood deaths in the developing world and have been responsible for water- and food-borne illness in adults. Adenoviruses have been responsible for recreational water outbreaks of eye and throat infections associated with swimming pools. Drinking water outbreaks have recently been documented.

The ingestion of just a few viruses are capable of causing infection and for this reason low numbers of viruses in water are considered a potential public health hazard. For this reason, large volumes of water are usually processed for their detection. For example, from 10 to 1000 L of water must be processed. Viruses are concentrated from the water by adsorption onto pleated cartridge filters (Hurst *et al.*, 2002). The viruses are then desorbed from the filters in 1 to 2 L of an eluent, which is concentrated to a few milliliters. Next, the concentrate is assayed by using cell culture or molecular methods, such as the polymerase chain reaction (PCR). The viruses are detected by the destruction of the cell culture (called *cytopathogenic effects* or *CPE*), but this may require days to weeks. PCR can detect and identify viruses in less than a day, but cannot be used to assess their viability (ability to cause infection).

Bacteria

Unlike most enteric viruses, enteric bacteria usually infect man and a wide variety of animals. The exception is bacteria of the genus *Shigella*, which infect only humans. Major waterborne bacteria and the illnesses they cause are shown in Table 3. Food is the most common source of infection in humans, although untreated drinking water (or break downs in water treatment) and recreational waters are also sources of infection.

Campylobacter jejuni is one of the most common identified causes of bacterial diarrhea in the United States. It is a common infection of birds, which is believed to be the major source of human exposure. It has been isolated from 22% of coastal and estuary waters ranging from 10 to 230 campylobacters per 100 mL (Bolton *et al.*, 1982). It has been found in 10 to 44% of pond water samples (Carter *et al.*, 1987). It is thought that virtually all surface waters contain *Campylobacter*. Recovery rates from surface waters are highest in the fall and winter months and lowest during the spring and summer months. *C. jejuni* dies off rapidly at warm temperatures (37 °C), but may survive for over 120 days at 4 °C.

Salmonella is a very large group of bacteria comprising over 2400 known types. All these types are pathogenic to humans and can cause a wide range of symptoms from mild diarrhea to severe illness and death. They are capable of infecting a large variety of both cold- and warm-blooded animals. Typhoid fever, caused by *S. typhi*, and paratyphoid fever caused by *S. paratyphoid*, are both enteric fevers that occur only in humans and primates. *Salmonella* can be detected in almost any nondisinfected domestic sewage or animal wastes.

Escherichia coli is found in the GI tract of all warm-blooded animals and is usually considered a harmless organism. However, several strains are capable of causing diarrhea. One group referred to as enterohemorrhagic *E. coli* (EHEC) is capable of causing serious life-threatening illness in young children and the elderly. EHEC is now recognized as a serious problem in developed countries. EHEC almost always belongs to the single serological type 0157 : H7. The illness usually includes severe cramping and diarrhea, which is initially watery, but becomes grossly bloody. The illness usually lasts only eight days, but, in some individuals, it develops into hemolytic-uremic syndrome or HUS, which results in permanent kidney failure. Cattle feces are the major source of human exposure.

Shigella is closely related to *E. coli. Shigella sonnei* causes a mild diarrhea and is the type most often causing illness in the United States. *Shigella* is a common cause of recreational waterborne outbreaks associated with swimming in lakes and rivers. They do not survive long in natural waters and are very sensitive to inactivation by disinfectants.

The genus *Vibrio* comprises a large number of species, but only a few of these species cause illness in humans. The most famous member of the group is *Vibrio cholerae*, the bacterium that causes cholera. Cholera remains prevalent in many parts of South America, Asia, and Africa. *V. cholerae* serotype 01 produces a toxin, which causes severe diarrhea and, in some cases, can cause death in a few hours because of rapid dehydration. Humans are the only known natural host, although environmental reservoirs may exist, apparently in association with copepods or phytoplankton.

Helicobacter pylori is the major cause of stomach ulcers in humans. The bacterium colonizes the stomach mucosa and produces a layer of mucus to protect itself from the low pH and digestive enzymes present in the stomach. Long-term infection may lead to stomach cancer. While person-to-person transmission is possible, several epidemiological studies have suggested that swimming in and drinking nondisinfected water are associated with infections of this organism (Klein *et al.*, 1991). Studies have shown that drinking untreated groundwater was related to infection with this organism (Baker and Hegarty, 2001).

Legionella pneumophila is the causative agent of Legionnaires disease and Pontiac fever.

Legionnaires disease is a severe respiratory infection, which has a 15% fatality rate in hospitalized cases. Pontiac fever is a milder infection characterized by fever and headache. *Legionella* occurs naturally in water, but appears capable of causing the disease only after growing in water above 20 °C. *Legionella* survives well above 50 °C. Hot water systems, air-conditioning cooling towers, hot tubs, and decorative fountains have been traced to outbreaks of *Legionella*. The bacterium survives for months in tapwater. The primary route of transmission is thought to be through the inhalation of aerosols.

Detection of pathogenic bacteria in water is difficult and time consuming and is not done routinely. The recent developments of molecular methods such as the PCR have made identification more practical. Traditional methods involve the use of enrichment media to select for the growth of the pathogen followed by growth on selective media followed by biochemical tests and the use of specific antibodies. Usually 100 to 1000 mL volumes are tested.

Protozoa

Although first described in 1907, *Cryptosporidium* was not recognized as a human pathogen until 1980. Since that time, it has been responsible for numerous outbreaks of waterborne diseases in the United States and the United Kingdom. In 1993, it was responsible for the largest waterborne outbreak ever documented in the United States. The outbreak, which occurred in Milwaukee, WI, resulted in 400 000 cases of diarrhea and almost 100 deaths (Mackenzie *et al.*, 1994). The source of the outbreak was traced to the occurrence of the organism in treated drinking water.

The protozoan parasite produces an environmentally stable oocyst that is resistant to inactivation at chlorine levels normally used in drinking water treatment. The oocysts range in size from 3 to 6 μm in diameter. After they are ingested, they release four sporozoites, which invade cells in the intestines and cause diarrhea. Cattle and other animals are major sources of oocysts in the environment and commonly occur in surface waters. *Cryptosporidium* can be expected to occur in all types of surface freshwater environments. In the United Kingdom, all finished drinking water treatment plants are continuously monitored for the presence of *Cryptosporidium* to ensure treatment performance. In the United States, limited monitoring of water sources for drinking water treatment plants is conducted every five years to ensure that the treatment process can reduce the oocysts to levels considered safe. Because of the resistance of the oocysts to chlorine, adequate filtration is needed to control the occurrence of the oocysts in treated drinking water. Ultraviolet light has been found to be an effective disinfectant for the control of this organism in water.

Giardia lamblia is the most frequently identified intestinal parasite in the United States (Adam, 1991). Humans become infected with *G. lamblia* by ingesting the environmentally resistant stage, the cyst. Once ingested, it passes through the stomach and into the upper intestines, releasing two trophozoites, causing diarrhea. When *Giardia* cysts enter the environment, they can survive for prolonged periods of time. They have been documented to survive for up to 77 days at 8 °C in distilled water (Bingham *et al.*, 1979). Beavers and muskrats are believed to be sources of *Giardia* in surface waters, but humans are also likely to be a major source. The cysts, measuring 8 to 16 μm, are significantly more resistant to chlorine and ozone disinfection than enteric bacteria and viruses, and outbreaks have been associated with unfiltered drinking water and recreational waters. *Giardia* is also very sensitive to inactivation by ultraviolet light. Because of its large size, it is easily removed by filtration.

Entamoeba histolytica is the cause of amebic dysentery. Humans are the main host, although pigs, monkeys, and dogs have also been found to serve as reservoirs. There are two cysts, small (5–9 μm) and large (10–20 μm) with each cyst producing eight trophozoites. This organism is generally only a problem in developing countries where sanitation is substandard. No waterborne outbreaks have occurred in the United States in more than 30 years. It is easily inactivated by common disinfectants.

Waterborne outbreaks of *Cyclospora cayetanesis*, Microsporidia, and *Toxoplasma gondii* have been documented, but how important water is in the transmission of these protozoa is currently uncertain.

Naegleria fowleri is a protozoa changing between a cyst, amoeba, and flagellate with the amoeba being dominant. The free-living protozoa are ubiquitous and found in fresh waters throughout the world (John, 1982). Cysts are usually present in low numbers, but when the water temperatures exceeds 35 °C (hot springs and warm stagnant waters), the amoeba transforms to the flagellated form rapidly, which enables the organism to swim. Infections are usually associated with children swimming in warm nondisinfected natural springs, lakes, and swimming pools. The flagellate swims into the nose of humans and makes its way via the nerves to the brain, producing a toxin that liquefies the brain, causing death. Infection usually occurs in children. Fortunately, it is a fairly rare infection.

Protozoan parasite detection in water involves collecting the organisms from 10 to 100 L of water by passage through cartridge filters, which collect them by size exclusion (Hurst *et al.*, 2002). The organisms are then eluted from the filters and further concentrated through centrifugation and capture on magnetic beads coated with antibodies for the specific organism to be detected. The organisms are then stained with fluorescently labeled antibodies specific for the organism to be detected. The cyst or oocyts are identified under an ultraviolet light microscope for the fluorescing organism and further characterized by shape, size, and internal structures.

Helminths

Helminths are multicellular animals that are transmitted by water and are parasitizing to humans. They include the Nematoda (roundworms) and the Platyhelminths, which are divided into two subgroups: the Cestoda (tapeworms) and the Trematoda (flukes). In these parasitic helminths, ova or eggs constitute the infectious stage. Excreted in the feces of infected persons and spread by water, soil, or food, these ova are very resistant to environmental stresses and to disinfection. The most important parasitic helminths are listed in Table 5. The ones causing the most common infections are largely transmitted by contact with fecally contaminated soil or food, although transmission by contaminated water may occur.

Ascaris is a large intestinal roundworm and a major cause of infections in humans worldwide. Worldwide estimates indicate that between 800 million and one billion people are infected, with most infections being in the tropics and subtropics. In the United States, most infections tend to occur in the Gulf Coast. Although most infections are mild, death can result from intestinal blockage. The life cycle of this parasite includes a phase in which the larvae migrate through the lungs and the throat before being swallowed. Although the eggs are dense, and hence readily removed by sedimentation in wastewater treatment plants, they are very resistant to the action of chlorine.

Schistosoma spp. are worms (trematodes or flukes) that have two deep suckers, which are used for attachment. The life cycles are complex, requiring an intermediate host (snail) to complete their life cycle; humans are the definitive host, excreting eggs in the feces. *S. mansoni* is the most important worldwide and is responsible for more than 200 million infections worldwide and causes up to 200 000 deaths annually (Hopkins, 1992). Cercariae are released from infected snails into the water environment and penetrate the human skin and migrate from the lungs to the liver. The adult worms then migrate to the circulatory system to other organs. The eggs are excreted in the feces or urine.

Helminths are detected in water by centrifuging a 100-mL to 1-L sample and resuspension in a smaller volume, which is examined under the microscope for specific ova or eggs characteristic of the individual species (Ayres and Mara, 1996).

INDICATOR MICROORGANISMS

Because of the difficulty and time necessary for the detection of waterborne pathogens, more easily detected enteric bacteria or bacteriophages (viruses which infect enteric bacteria) are used as indicators of fecal contamination and potential presence of enteric pathogens. Developed at the beginning of the twentieth century, the indicator concept depends on the fact that certain nonpathogenic bacteria occur in the feces of all warm-blooded animals. These bacteria can easily be isolated and quantified by simple laboratory techniques. The coliform group has been used as the standard for assessing fecal contamination of drinking water for almost a hundred years. This group includes all aerobic and facultatively anaerobic, gram-negative, nonspore-forming, rod-shaped bacteria that produce gas upon lactose fermentation in prescribed culture media within 48 h at 35 °C. Through experience it has been learned that absence of these organisms in 100 mL of drinking water ensures the prevention of enteric bacterial disease outbreaks. However, it has been learned that threats of disease may still exist from enteric viruses and protozoa, which have a greater resistance to disinfectants and survive longer in the environment. In addition, coliforms may originate from nonfecal sources and regrow in water under some conditions. Fecal coliforms include the genera *Escherichia* and *Klebsiella*, and are differentiated in the laboratory by their ability to ferment lactose with the production of acid and gas at 44.5 °C within 24 h. Fecal coliforms have been used as standards to indicate the quality of recreational waters. *E. coli* and enterococci (another group of enteric bacteria found in human feces) have been shown to be more specific indicators of fecal contamination, and concentrations of them in recreational water have been shown to be related to risk of gastroenteritis from swimming in surface waters (Cabelli *et al.*, 1982).

Because of their constant presence in sewage and polluted waters, the use of bacteriophages as appropriate indicators of fecal pollution has been proposed (Havelaar *et al.*, 1993). Bacteriophages have also been suggested as indicators of human pathogenic enteric viruses and have been used as models to assess human enteric virus behavior during water treatment and in the environment. Bacteriophages of *E. coli*, referred to as *coliphages*, have been the ones most commonly studied. The use of coliphages as indicators of fecal pollution is based on the assumption that their presence in water denotes the presence of bacteria capable of supporting the replication of the phage. Two groups of phage have been studied: the somatic coliphage, which infects *E. coli* strains through cell wall receptors, and the F-specific RNA coliphage, which infects strains of *E. coli* through the F+ or sex pili. Most F+ coliphages are similar in shape and size, as most of the waterborne enteric viruses and tests have been developed to specifically detect this type of coliphage.

FATE AND TRANSPORT IN THE ENVIRONMENT

Once released from the infected host, waterborne pathogens must survive long enough to infect another host. Enteric pathogens may survive from a few days to many months in the environment, depending on a number of important factors. Generally, helminths eggs survive the longest followed by viruses, protozoa, and bacteria. Enteric bacterial survival is the most limited because they need nutrients and have difficulty in competing with indigenous microflora outside the host. Since water is their natural habit, water-based pathogens increase in number under the proper environmental conditions.

Temperature is the most important factor controlling waterborne pathogens. Generally, the lower the temperature, the longer is the survival time of pathogens. Freezing is detrimental to the survival of helminths eggs, bacteria, and protozoa. However, at freezing conditions, viruses may survive for many years.

Surface Waters

Sources

Enteric pathogens in surface waters primarily originate from runoff during rain events and sewage discharges. Livestock is the primary source of pathogens on agricultural land (Khaleel *et al.*, 1980). Rainfall results in the movement of accumulated animal fecal material on land to nearby water bodies. In some urbanized areas, runoff is collected by the sewage collection systems during major rainfall events and the ability of sewage treatment plants to treat these combined wastes may be exceeded, resulting in the release of untreated sewage.

Table 6 Percentage of animals shedding selected waterborne zoonotic pathogens

Pathogens	Cattle	Calves	Sheep	Pig	Rodents	Waterfowl
Cryptosporidium sp.		20–90	8–40	5 – 20	30	13–100
Giardia sp.	57–97					
Campylobacter spp.					10–95	6–50
Salmonella spp.	13		4–15	7 – 22	1–10	1–10
E. coli O157 : H7	2–3.5		2	1.5 – 9		

Modified from Medema *et al.* (2003).

Livestock are a well-known source of waterborne pathogens (Table 6). There is a direct relationship between the presence of cattle on agricultural land and fecal coliforms in surface waters and direct animal access may have more of an impact on streams than stocking densities (Tiedemann *et al.*, 1987). Wild animals, largely mammals and birds, are another source of human pathogens. In well-protected surface water catchments, upland reservoirs, and mountain streams, wildlife may be the most important source of fecal pollution. In the Netherlands, it has been shown that reservoirs are at greatest risk during late winter/early spring, when bird numbers on the (partly frozen) reservoirs are high. When thaw sets in, the bird feces that have collected on the ice enter the water, leading to a peak contamination with *Campylobacter, Cryptosporidium*, and *Giardia* (Medema *et al.*, 2003).

The concentration of pathogens in raw sewage is dependent upon the incidence of infection within the population, time of year, per capita water usage, and the socioeconomic status of the population. The concentration of pathogens found in feces of infected persons is shown in Table 7. The peak incidence of many enteric infections is seasonal in temperate climates. Thus, the highest incidence of enterovirus infection is during the late summer and early fall; rotavirus infections tend to peak in the early winter, and *Cryptosporidium* infections peak in the early spring and fall. A greater incidence of enteric infections occurs in lower socioeconomic groups, particularly where lower standards of sanitary conditions prevail. The concentration of pathogens in raw sewage in the developed world and polluted surface waters is shown in Table 8. Concentrations of enteric pathogens are much greater in sewage in the developing world than the industrialized world. Sewage treatment by activated sludge or stabilization ponds may reduce the concentration of pathogens by 90 to 99%. Disinfection with chlorine can significantly further reduce the concentration of bacterial pathogens, but significant concentrations of viruses and protozoan parasites may remain. Filtration to remove turbidity, cysts, and oocysts and reduction of organic matter by chemical flocculation and extended disinfection are necessary to reduce pathogens below detection levels.

Table 7 Concentration of enteric pathogens indicator microorganisms in feces

Organism	Per gram of feces
Protozoan parasites	
Helminths *Ascaris*	10^4–10^7
Enteric viruses	
Enteroviruses	10^4–10^5
Rotavirus	10^{10}
Enteric bacteria	
Salmonella spp.	10^4–10^{10}
Shigella	10^5–10^9
Indicator bacteria	
Coliform	10^7–10^9
Fecal coliform	10^6–10^9

Table 8 Estimated levels of enteric organisms in sewage and polluted surface water in the United States

Organism	Raw sewage	Polluted stream water
Coliforms	10^7–10^9	10^5
Fecal coliforms	10^6–10^7	–
Fecal streptococcus	10^5–10^7	–
Entercocci	10^4–10^5	–
Coliphages	10^2–10^3	–
Enteric viruses	10^2	1–10
Giardia	10–10^2	0.1–1
Cryptosporidium	1–10	0.1–10^2

Modified from U.S. Gerba (2000).

Transport

In temperate lakes, thermal stratification may occur during summer and winter. This reduces the exchange of water between the upper and lower layers of lake water. In summer, the quality of the water at the bottom slowly deteriorates due to settling. When de-stratification occurs in autumn, water from the upper and lower layers mixes. This process causes settled particles with coliforms to reenter the water. In one lake, for example, it was reported that autumn de-stratification led to a 10-fold increase in coliform densities for several weeks; from a level consistently below 10/100 mL in summer to more than 100/100 mL (Geldreich *et al.*, 1989).

The concentration of pathogens and indicator bacteria is closely linked to rainfall events. Rainfall and rainfall intensity are important factors in the release of pathogens from

fecal matter and the subsequent mobilization into surface waters (Ferguson *et al.*, 2003). Higher land-surface slopes generally have more overland transport (Tate *et al.*, 2000).

Global climate change has resulted in an increase in heavy rainfall events. This increase has a direct impact of pathogen loading in watersheds. A significant statistical relationship between the increase incidence of rainfall and waterborne disease has been demonstrated in the United States (Curriero *et al.*, 2001).

Pathogens tend to accumulate in sediments due to direct settling or attachment to suspended particulates, which form sediment deposits. Survival of pathogens in sediments can be significantly greater than in surface waters. As long as pathogens remain adsorbed to sediments, there is little threat to public health. However, resuspension of the upper levels of sediments may increase the concentration of pathogens in the water. Sediment resuspension may occur through boat activity, swimming, currents, dredging, storm events, or animal activity. Studies in coastal areas indicate that 10- to 10 000-fold greater concentrations of enteroviruses may occur in sediments than in the overlying seawater on a volume basis (Gerba *et al.*, 1977). Fecal coliforms have been found in freshwater sediments at concentrations 100 to 1000 fold higher than the overlying water (Geldreich, 1970). In duckweed ponds and multispecies wetlands used to treat activated sludge effluents, the concentration of *Giardia* cysts was found to be 150- to 550-fold greater in the sediments than the overlying water (Karim and Gerba, 2002). *Cryptosporidium* oocyst concentrations were 50 to 70 times greater in the sediments than the overlying water. Settling is probably the major mechanism of cyst and oocyst accumulation in sediments. This explains the greater concentration of the larger *Giardia* cysts in the sediments compared to the oocysts.

Survival

Temperature and sunlight intensity are major factors controlling the survival of pathogens in surface waters (Table 9). In general, survival is prolonged when water temperatures are low. In a study in Alaska, little enterovirus inactivation over a distance of 150 miles under the ice in an Alaskan River in the winter was observed (Dahling and Safferman, 1979). Enteric bacteria are sensitive to inactivation by ultraviolet light in sunlight and predictive models for fecal coliform survival in surface waters have been developed (Kapuscinski and Mitchell, 1982). Adsorption of enteric bacteria and viruses to particles prolongs their survival probably through protection from antagonistic factors in water and, in the case of viruses, enhanced thermostabilization (Liew and Gerba, 1980). Survival of poliovirus in marine sediments was found to be 4- to 96-fold greater than in seawater (LaBelle and Gerba, 1980). In contrast, *Giardia* cyst survival was found to be less in duckweed pond sediments than the overlying water (Karim and Gerba, 2002). Indigenous bacteria and protozoa have been found in both marine and freshwater antagonistic to enteric bacterial and viral survival. In general, pathogen survival in marine waters is less than that in freshwater and greater in the presence of sewage. Typical die-off rates of waterborne pathogens in surface waters are shown in Table 10.

Table 9 Factors affecting the persistence of pathogens in soil

Factor	Comments
Temperature	Longer survival at low temperatures; longer survival in winter than in summer.
Water-holding capacity	Survival is lower in sandy soils with lower water-holding capacity.
Light	Lower survival at soil surface because of UV lights
Soil texture	Clays and humic materials increase water retention by soils and thus enhance microbial survival.
pH	May indirectly control survival by controlling adsorption to soils, particularly for viruses.
Cations	Some (e.g. Mg^{2+}) may thermally stabilize viruses
Organic matter	May influence bacterial survival and regrowth. Virus survival is longer in the presence of organic matter.
Biological factors	
Antagonism from soil microflora	Increased survival in sterile soils. No clear trend as regards the effect of soil microflora on viruses.

Adapted from Gerba and Bitton (1984).

Table 10 Die-off rates and reduction time for selected microorganisms in surface water

Microorganism	Die-off (Log_{10} day^{-1})
Parasites	
Cryptosporidium sp	0.0057–0.046
Giardia sp.	0.023–0.23
Viruses	
Enteroviruses	0.01–0.2
Hepatitis A	0.05–0.2
Rotavirus	0.24–0.48
Bacteria	
Salmonella spp	1–7
E. coli	0.23–0.46
Coliforms	0.77
Enterococci	0.17–0.77

Modified from Medema *et al.* (2003).

Subsurface Environments

Sources

Almost half of all the waterborne disease outbreaks documented every year in the United States are due to contaminated groundwater (Lee *et al.*, 2002). Pathogens enter

the groundwater from septic tank leachfields, land application of wastewater for irrigation and disposal, landfills, riverbank infiltration, leaking sewer lines, waste oxidation ponds, and direct injection into the groundwater. Septic tanks are believed to be the major cause of outbreaks and a recent epidemiological study found a direct relationship between diarrhea and septic tank density in communities served by septic tanks (Borchardt *et al.*, 2003). In a nationwide study of utility drinking water wells in the United States, viruses, coliphages, and indicator bacteria were found in 31, 21, and 15% of the samples respectively (Abbaszadegan *et al.*, 2003).

Transport

Microorganisms are colloids (biocolloids) that behave as particulates in their transport in the subsurface. Because of this, their movement may be much different than solutes, depending on the nature of the subsurface matrix. In uniform material, they may experience greater retardation than solutes; however, where fractures are present, they may travel at much greater rates (Hinsby *et al.*, 1996).

Owing to their size, helminths ova, protozoa, and bacteria are easily filtered out during their transport through soil. Bacterial removal by soils is inversely proportional to the particle size of soils. Bacteria may also adsorb to soil. The degree of adsorption is influenced by a number of water quality and soil properties, which are listed in Table 11. Rainfall significantly influences the transport of bacteria through soil and greater numbers are seen in wells after rainfall events. Under field conditions, pathogenic bacteria are efficiently retained by most soils and rarely detected in groundwater. Factors controlling transport of microorganisms in the subsurface are shown in Table 11.

Because of their small size, viruses may travel further than any other pathogen in the subsurface. The major factor controlling virus transport through soil is adsorption. Both electrostatic and hydrophobic interactions are involved in virus adsorption. The charged carboxyl and amine groups on the surface of the viral protein coat are involved in electrostatic interactions and aliphatic groups in the hydrophobic interactions (Gerba, 1984). Clay soils generally have a greater capacity than sandy soils. Greater virus removal occurs near the soil surface because of virus retention onto iron oxides and other metal oxides that form in the presence of oxygen (Chu *et al.*, 2001). Adsorption varies with the type and strain of virus (Goyal and Gerba, 1979). Some viruses (poliovirus type 1) adsorbs very well to most soils, while others (echovirus 1 and coliphages MS-2) adsorb poorly. Viruses adsorb poorly to soils in low ionic strength conditions and rainfall causes viral desorption promoting transport. Transport is also promoted by the presence of organic matter. Thus, virus adsorption is less when suspended in sewage. Virus transport is greater above pH 7.0. Changes in pH can also result in viral de-adsorption (Bales *et al.*, 1995).

Both laboratory and field studies have shown that sewage sludge (biosolid) application does not result in virus transport to groundwater. Viruses have never been detected in groundwater beneath sludge application sites (Straub *et al.*, 1993). Viruses appear to be tightly bound to the sewage sludge because of the presence of proteins produced by bacteria, which have a high affinity for binding viruses (Sano, 2003).

Several models have been proposed for predicting viral and bacterial transport through soils and aquifers (Yates and Ouyang, 1992). These models vary in complexity and require environmental (temperature), soil/hyrogeologic (texture, adsorption coefficient for the organism), and microbiological transport (e.g., microbial type, inactivation rate). Most of these models simulate microbial transport under saturated flow conditions.

Survival

The main factors controlling the survival of pathogenic bacteria in soil are temperature, moisture content, sunlight, pH, organic matter, bacterial type, and antagonistic microflora, which include indigenous soil bacteria and predatory protozoa. Soil desiccation is probably the major

Table 11 Summary of the main factors governing the transport of pathogens through soil

Factor	Comments
Soil Type	Fine-textured soils retain microorganisms more effectively than light-textured soils. Iron oxides increase the adsorptive capacity of soils.
Filtration	Straining of bacteria and protozoan parasites at soil surface limits their movement.
pH	Generally, adsorption increases when pH decreases.
Cations	Adsorption increases in the presence of cations. Rainwater may desorb viruses from soil owing to its low conductivity.
Soluble organics	Generally compete with microorganisms for adsorption sites. Humic and fulvic acid reduce virus adsorption to soils.
Microbial type	Adsorption to soils varies with microbial type and strain.
Flow rate	The higher the flow rate, the lower the microbial retention.
Saturated versus unsaturated flow	Virus movement is less under unsaturated flow conditions.

Modified from Gerba and Bitton (1984).

Table 12 Die-off rates of enteric microorganisms in groundwater

Microorganism	Die-off rate (log_{10} day^{-1})
Viruses	
Hepatitis A	0.10–0.33
Poliovirus 1	0.013–0.77
Coxsackievirus	0.19
Rotavirus SA11	0.36
Coliphage MS-2	0.063–0.75
Bacteria	
E. coli	0.063–0.36
Faecal streptococci	0.03–0.24
Salmonella typhimurium	0.13–0.22

Modified from Medema *et al.* (2003).

factor controlling bacterial survival in soils. Indicator bacteria (coliforms, fecal coliforms, *E. coli*) often increase in soil after rainfall events (Pepper *et al.*, 1993).

Temperature and soil moisture are the major factors controlling virus survival in soil. Viruses in land applied sewage sludge were found to survive for 23 weeks during the winter in Denmark, but only for 2 to 4 weeks in the summer and fall in Florida (Bitton, 1999). Die-off rates for waterborne enteric pathogens in groundwater are shown in Table 12.

CONCLUDING REMARKS

The rapid emergence of new waterborne pathogens in recent decades has created a greater need to better understand their fate and transport in aquatic environments. Pathogens are opportunistic and will always find means to coexist with humans. Applications of molecular methods to identify and track the source of water-related pathogens provide better insight into the exposure via surface and groundwaters. This in combination with microbial risk assessment should provide better means to control their occurrence in our water supplies.

REFERENCES

Abbaszadegan M., Lechevallier M. and Gerba C. (2003) Occurrence of viruses in US groundwaters. *Journal of the American Water Works Association*, **95**, 107–120.

Adam R. (1991) The biology of *Giardia* spp. *Microbiological Reviews*, **55**, 706–732.

Ayres R.M. and Mara D.D. (1996) *Analysis of Wastewater for use in Agriculture*, World Health Organization: Geneva.

Baker K.H. and Hegarty J.P. (2001) Presence of *Helicobacter pylori* in drinking water is associated with clinical infection. *Scandinavian Journal of Infectious Diseases*, **33**, 744–746.

Bales R.C., Li S., Maguire M.T., Yahya M.T., Gerba C.P. and Harvey R.W. (1995) Virus and bacteria transport in sandy aquifer, Cape Cod, MA. *Groundwater*, **33**, 653–881.

Bingham A.K., Jaroll E. and Meyer E. (1979) *Giardia* spp: physical factors of excystation in vitro and excystation vs. eosin exclusion as determinants of viability. *Experimental Parasitology*, **47**, 284–291.

Bitton G. (1999) *Wastewater Microbiology*, John Wiley: New York.

Bolton F.J., Coats P.M.D. and Robertson L. (1982) A most probable number method for estimating small numbers of campylobacters in water. *Journal of Hygiene*, **89**, 185–190.

Borchardt M.A., Chyou P.H., DeVries E.O. and Belongia E.A. (2003) Septic tank density and infectious diarrhea in a defined population of children. *Environmental Health Perspectives*, **111**, 742–748.

Cabelli V.J., Dufour A.P., McCabe L.J. and Levin M.A. (1982) Swimming associated gastroenteritis and water quality. *American Journal of Epidemiology*, **115**, 606–616.

Carter A.M., Pacha R.E., Clark G.W. and Williams E.A. (1987) Seasonal occurrence of *Campylobacter* spp. in surface waters and their correlation with standard indicator bacteria. *Applied and Environmental Microbiology*, **53**, 523–526.

Chu Y., Jin Y., Flury M. and Yates M.V. (2001) Mechanisms of virus removal during transport in unsaturated porous media. *Water Resources Research*, **37**, 253–263.

Curriero F.C., Patz J.A., Roe J.L. and Lele S. (2001) The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994. *American Journal of Public Health*, **91**, 1194–1194.

Dahling D.R. and Safferman R.S. (1979) Survival of enteric viruses under natural conditions. *Applied and Environmental Microbiology*, **38**, 1103–1110.

Ferguson C., de Roda Husma A.M., Altavilla N., Deere D. and Ashbolt N. (2003) Fate and transport of surface water pathogens in watersheds. *Critical Reviews in Environmental Science and Technology*, **33**, 299–361.

Geldreich E.E. (1970) Applying bacteriological parameters to recreational water quality. *Journal of the American Water Works Association*, **62**, 113–120.

Geldreich E.E., Nash H.D., Spino D.F. and Reasoner D.J. (1989) Bacterial dynamics in a water supply reservoir: a case study. *Journal of the American Water Works Association*, **62**, 13–120.

Gerba C.P. (1984) Applied and theoretical aspects of virus adsorption to surfaces. *Advances in Applied Microbiology*, **30**, 133–168.

Gerba C.P. (2000) Domestic wastes and waste treatment. In *Environmental Microbiology*, Maier R.M., Pepper I.L. and Gerba C.P. (Eds.), Academic Press: San Diego, pp. 505–534.

Gerba C.P. and Bitton G. (1984) Microbial pollutants: their survival and transport pattern to groundwater. In *Groundwater Pollution Microbiology*, Bitton G. and Gerba C.P. (Eds.), John Wiley and Sons: New York, pp. 65–88.

Gerba C.P., Smith E.M. and Melnick J.L. (1977) Development of a quantitative method for detecting enteroviruses in estuarine sediments. *Applied and Environmental Microbiology*, **34**, 158–163.

Goyal S.M. and Gerba C.P. (1979) Comparative adsorption of human enteroviruses, simian rotavirus, and selected bacteriophages to soils. *Applied and Environmental Microbiology*, **38**, 241–247.

Havelaar A.H., van Plphe M. and Dorst Y.C. (1993) F-specific RNA bacteriophages are adequate model organisms for enteric viruses in fresh water. *Applied and Environmental Microbiology*, **59**, 2956–2962.

Hinsby K., McKay L.D., Jorgensen P., Lenczewski M. and Gerba C.P. (1996) Fracture aperture measurements and migration of solutes, viruses, and immiscible creosote in a column of clay rich till. *Groundwater*, **34**, 1065–1075.

Hurst H.J., Crawford R.L., Knudsen G.R., McInerney M.J. and Stetzenbach L.D. (2002) *Manual of Environmental Microbiology*, *Second Edition*, ASM Press: Washington.

Hopkins D.R. (1992) Homing in on helminths. *American Journal of Tropical Medical Hygiene*, **46**, 626–634.

John D.T. (1982) Primary amebic meningoencephalitis and the biology of *Naegleria fowleri*. *Annual Reviews of Microbiology*, **36**, 101–103.

Kapuscinski R.B. and Mitchell R. (1982) Sunlight-induced mortality of viruses and *Escherichia coli* in coastal seawater. *Environmental Science and Technology*, **41**, 1–6.

Karim M.R. and Gerba C.P. (2002) Fate of viruses and protozoan parasites in aquatic sediments. In *Encyclopedia of Environmental Microbiology*, Bitton G. (Ed.), John Wiley: New York, pp. 1252–1256. In: Ed.).

Khaleel R., Reddy K.R. and Overcash O.R. (1980) Transport of potential pollutants in runoff water from land receiving animal wastes: a review. *Water Research*, **14**, 421–436.

Klein P.D., Graham D.Y., Gailour A., Opekun A.R. and Smith E.O. (1991) Water source as risk factor for *Helicobacter pylori* infection in Peruvian children. *Lancet*, **337**, 1503–1506.

LaBelle R.L. and Gerba C.P. (1980) Relationships between environmental factors, bacterial indicators, and the occurrence of enteric viruses in estuarine sediments. *Applied and Environmental Microbiology*, **39**, 588–596.

Lee S.H., Levy D.A., Craun G.F., Beach M.J. and Calderon R.L. (2002) Surveillance of waterborne-disease outbreaks-United States, 1999–2002. Surveillance Summaries. *Morbidity and Mortality* **51**,(SS-8) 1–47.

Liew P.F. and Gerba C.P. (1980) Thermostabilization of enteroviruses by estuarine sediment. *Applied and Environmental Microbiology*, **40**, 305–308.

Mackenzie W., Hoxie N., Procter M., Gradua M., Blair K., Peterson D., Kazmierczak J., Addis D., Fox K. and Rose J. (1994) A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply. *New England Journal of Medicine*, **331**, 161–167.

Medema G.J., Shaw S., Waite M., Snozzzi M., Morreau A. and Grabow W. (2003) Catchment characterization and source water quality. *Assessing Microbial Safety of Drinking Water. Improving Approaches and Methods*, Organization for Economic Co-operation and Development: Paris, pp. 111–158.

Pepper I.L., Josephson K.L., Bailey R.L., Burr M.D. and Gerba C.P. (1993) Survival of indicator organisms in Sonoran Desert soil amended with sewage sludge. *Journal of Environmental Science and Health*, **A28**, 1287–1302.

Sano D. (2003) *Discovering, Characterizing and Cloning of Virus-binding Proteins in Activated Sludge Culture for the Innovative Development of Virus Removal Technology*. Ph.D. Dissertation, Tohoku University, Japan.

Straub T.M., Pepper I.L. and Gerba C.P. (1993) Hazards from pathogenic microorganisms in land-disposed sewage sludge. *Reviews of Environmental Contamination and Toxicology*, **132**, 55–91.

Tate K.W., Atwill E.R., George M.R., McDougald M.K. and Larsen R.E. (2000) *Cryptosporidium* transport from cattle fecal deposits on California rangelands. *Journal of Range Management*, **53**, 295–299.

Tiedemann D.A., Higgins D.A., Quiegley T.M., Sanderson H.R. and Marx D.B. (1987) Responses of fecal coliforms in streamwater to four grazing strategies. *Journal of Range Management*, **40**, 322–329.

WHO (1996) *Water and Sanitation: WHO Fact Sheet no. 112*, World Health Organization: Geneva.

Yates M.V. and Ouyang Y. (1992) VIRTUS, a model of virus transport in unsaturated soils. *Applied and Environmental Microbiology*, **58**, 1609–1616.

99: Salinization

JOHN RUPRECHT AND SHAWAN DOGRAMACI

Department of Environment, Salinity and Land Use Impacts Branch, Resource Science Division, East Perth, Australia

Salinization of land and water resources is a major environmental and economic problem facing many parts of the world. Soil salinization and the consequent degradation of agricultural land are proceeding so fast that many countries may not be able to achieve sustainable agriculture to feed their population in the foreseeable future. Furthermore, many important ecosystems and internationally recognized wetlands are under serious threat from the increased salinity.

Salinization is the process by which the concentration of total dissolved solids in water or soil is increased because of natural or human-induced processes.

Many aspects of hydrological science underpin the understanding of salinization. The knowledge and understanding of the movement of water and salt from the atmosphere, through landscapes and rivers are essential to understanding salinization.

The main approaches to assessing the impacts of land and water salinization and to evaluating potential solutions include process studies, catchment studies, modeling, and statistical analysis. Each of these is discussed in more detail in the article. A summary of the models available for salinity studies is also outlined.

INTRODUCTION

Salinization of land and water resources is a major environmental and economic problem facing many parts of the world. One of the essential elements to sustain human life on earth is the capacity of the soil to produce sufficient food for the population. Soil salinization and the consequent degradation of agricultural land are proceeding so fast that many countries may not be able to achieve sustainable agriculture to feed their population in the foreseeable future. Furthermore, many important ecosystems and internationally recognized wetlands are under serious threat from the increased salinity.

Many aspects of hydrological science underpin the understanding of salinization. The knowledge and understanding of the movement of water and salt from the atmosphere, through landscapes and rivers, are essential. The aims of this section are to describe the salinization of soil and water, to present a synopsis of the chemical and physical processes causing the increased salinity levels, and to describe the tools and approaches that are available to evaluate and manage these degraded resources.

SALT ON THE LAND

Physical and chemical processes such as weathering, erosion, flooding, and deposition together with climatic aridity cycles have occurred throughout the history of the Earth. These processes are responsible for the concentrations of salt in various parts of the landscape. Extensive salt accumulation over time has occurred in the lower parts of landscapes resulting in saline soils and salt lakes. Salinization has also been a reversible process, with natural ecosystems flushing out the salt during periods of wetter climatic conditions.

In recorded history, several major occurrences of human-induced salinity (secondary salinity) have affected societies. The earliest and most serious occasion occurred in the latter part of the fourth millennium BC in ancient Mesopotamia, now southern Iraq (Jacobsen and Adams, 1958). Development of agricultural land in semiarid Mesopotamia required massive investments in major engineering schemes such as diversion canals. Large volumes of water were diverted from the Euphrates and Tigris rivers through canals to fields leading to increased seepage, flooding, and a consequent

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

rise in the watertable. Declining wheat production and cultivation of more salt-tolerant species, such as barley, were associated with an increase of soil salinity. Yield reduction, and the resulting inability to meet the needs of the population, devastated these old centers of civilization. Other historical examples of the impacts of irrigation salinization on the decline and depopulation of agricultural areas have occurred in China, the Indus River Basin, and South America (Casey, 1972).

More recent examples of the impacts of large-scale irrigation schemes and land clearing on the salinity development include the San Joaquin Valley in California, Australia, Pakistan, and the Amu Darya and Syr Darya rivers in Turkmenistan and Uzbekistan (Ghassemi *et al.*, 1995). In Australia alone, approximately 5.6 million ha of arable land are currently affected by salinity, and approximately 12 million more ha are threatened by increased salinity in the next 100 years (Australian Agriculture Assessment, 2001).

SALT IN WATER

Approximately 97% of the water on the Earth occurs as saline water in the seas and oceans. Neglecting the amount of the freshwater held in polar ice caps and glaciers, groundwater accounts for approximately half of the remaining freshwater resources on a global scale. The water in rivers, streams, and lakes that contributes frequently and actively to rainfall, evaporation, and stream flow accounts for only a small percentage (0.3%) of the global total freshwater resources (Todd, 1970; Freeze and Cherry, 1979). Most groundwater on the continents is meteoric in origin, that is, it has been derived from the atmosphere as rain or snow.

Naturally occurring saline groundwater may occur in sediments deposited in the ocean where seawater is trapped between the grains; or during weathering where highly saline water is formed and chemically combined in the weathered products; or during volcanism and metamorphism. An example of the latter case is the carbon-dioxide-rich springs in California, which contain significant a component of metamorphic water (Barnes, 1970).

Most of the groundwater resources globally are affected by extensive and intensive anthropogenic activities such as land clearing and irrigation, resulting in secondary salinization. This is particularly true in dry regions of the world, in hyper-arid and subhumid regions that are mostly warm to hot, and the potential evaporation far exceeds rainfall (Ghassemi *et al.*, 1995). For example, Datta and Jong (2002) describe the impacts of irrigation by groundwater pumping on the accumulation of salt in the root zone in the northwest region of Haryana in India. Funakawa *et al.* (2000) discuss the salinization caused by irrigation schemes and diversion of river water in Central Asia. In Australia, clearing and irrigation have caused substantial increases in groundwater and stream salinity (Australian Agricultural Assessment, 2001). Salinization of groundwater has been part of the water resource problem over the last 2000 years in southern Iraq (Jacobsen and Adams, 1958). Overpumping of aquifers and changing groundwater budgets in the coastal areas have resulted in salinization of groundwater by seawater intrusion (Ergil, 2000), and mixing of saline aquifers with fresher aquifers as a result of the groundwater head change (Hopkins and Richardson, 1999; Marie and Vengosh, 2001).

Significant changes to the nature of many inland waters have occurred over the last 200 years. Many inland waters, particularly in semiarid and arid regions, are becoming more saline from human activity.

The need for better recognition of the costs of salinization and its social and environmental effects and more effective managements is stressed by the many publications over the last decade (Ghassemi *et al.*, 1995; Dogramaci and Waterhouse, 2004).

DEFINITION OF SALINIZATION

The term salinity refers to the presence in soil and water of various electrolytic mineral solutes. Most common among these solutes are the dissociated cations Na^+, K^+, Ca^{2+}, Mg^{2+}, and the anions Cl^-, SO_4^{2-}, NO_3^-, HCO_3^-, and CO_3^{2-}. In addition, hypersaline waters may contain trace concentrations of the elements B, Se, Sr, Li, Rb, F, Mo, Mn, Ba, and Al, some of which may be toxic to plants and animals (Tanji, 1990; Hillel, 2000).

Salinization is the process by which the concentration of total soluble salts (TSS) in water or soil is increased because of natural or human-induced processes. Natural and human-induced salinization are also called *primary and secondary salinization* respectively. Salt-affected soils of primary origin are formed as a result of the long-term influence of natural processes leading to an accumulation of salts within a landscape. The presence of salts can be the result of gradual weathering, submergence of soils under seawater, and accession of salts bought inland by rainfall. All continents have extensive areas affected by primary salinization (Ghassemi *et al.*, 1995).

Secondary salinization of land is also the result of an increased concentration of salts (TSS) in the soil profile. This occurs primarily because of the evapotranspiration of recharge, then mobilization by groundwater. The salt stored at the root zone in the soil profile is being mobilized because of excess recharge from human activities such as irrigation or land clearing. The additional water raises watertables or creates an upward pressure on semiconfined or confined aquifers. When the watertable is within 1–2 m of the ground surface, groundwater evaporates, or capillary action brings water to the surface where it evaporates leaving salts behind. The salt stored in a soil profile can be dissolved

and moved laterally or vertically by groundwater and surface runoff towards watercourses, and increase salinity in streams, rivers, or terminal lakes.

Units of Salinity Measurement

Stream and groundwater salinity refers to the concentration of TSS in water. The term *salinity* is also used interchangeably with *salinization* to mean the effects of saline water. Peck (1983) provided a stricter definition of salinity as the concentration of the major ions (Na^+, K^+, Ca^{+2}, Mg^{+2}, Cl^-, HCO_3^-, CO_3^{-2} and SO_4^{-2}) in a solution.

Salinity may be expressed in a number of ways depending on the method of measurement. The most accurate method of determining the salinity of a particular water sample is to sum the measured quantities of the individual ions, which is expressed as TSS. Another method of expressing salinity is to weigh the evaporated residue of a given sample after filtering it to remove sediments. Salinity can also be expressed as Total Dissolved Solids (TDS). Both methods are expressed in units of milligrams per liter ($mg\,L^{-1}$).

For reasons of analytical convenience, a practical index of salinity is electrical conductivity (EC), usually expressed in units of milliSiemens per meter ($mS\,m^{-1}$). Electrical conductivity values are always expressed at a standard temperature of 25 °C to enable comparisons of readings taken under varying climatic conditions. A descriptive classification of water salinity is shown in Table 1.

Stream and groundwater salinity can be measured by:

- Intermittent sampling: samples taken on an ad-hoc basis to measure baseline salinity levels.
- Regular sampling: daily, weekly, or monthly samples. Regular sampling can also encompass automatic pump samplers that sample water at predetermined times, or with rate of rise in water level.
- Continuous monitoring: the continuous measurement of electrical conductivity with values stored on a data logger at regular time intervals, usually ranging from less than 5 min to more than 24 h. The conductivity meters most commonly used are either toroidal cells or four electrode.

Soil salinity can be extremely heterogeneous within short distances horizontally and vertically (Moore, 1998). Soil salinity can be measured in a laboratory by soil testing or with geophysical methods (e.g. electromagnetic induction meters). Soil salinity is measured by extracting liquid from a saturated soil paste and TSS measured as for water. However, because this method is laborious and expensive, a cheaper alternative is to measure electrical conductivity of a 1:5 soil-water mixture (ECe). A calculated ECe can be derived by multiplying the EC (1:5) by a factor related to soil texture (Moore, 1998). Soil salinity can also be defined by the effects on plants (Table 2).

Soil salinity on a large scale can be mapped with an electromagnetic (EM) conductivity meter that estimates the bulk electrical conductivity of the soil, which depends on the salinity of the soil solution, its water content, and the type and amount of clay in the soil.

Table 2 Assessment of saline land with soil salt content (Adapted from USSL, 1954)

Soil salinity ($mg\,L^{-1}$)	Rating	Effect on plants
0–1000	Non – saline	Salinity effects mostly negligible
1000–2100	Slightly saline	Yield of sensitive crops reduced
2100–4500	Moderately saline	Yield of many crops reduced
4500–9200	Highly saline	Only salt-tolerant crops yield satisfactory
>9200	Extremely saline	Only very salt-tolerant plants yield satisfactory

Table 1 Descriptive classification of water salinity (Adapted from Hillel (2000))

Water quality	TDS ($mg\,L^{-1}$)	Comment
Fresh	<500	Regarded as good quality drinking water based on taste Acceptable for most irrigation
Slightly brackish	500 to 1000	Acceptable as drinking water based on taste Acceptable for most irrigation – The acceptable level for irrigation will range with type of crop, type of soils, and level of drainage.
Brackish	1000 to 2000	Acceptable for most stock and some irrigation. The acceptable level for irrigation will range with type of crop, type of soils, and level of drainage.
Moderately saline	2000 to 5000	Primary drainage
Saline	5000 to 10 000	Limited farm use – unacceptable for most stock unless as an emergency supply
Highly saline	10 000 to 35 000	Very saline groundwater
Hypersaline	>35 000	Seawater, may have use in industry or mining.

In saline areas, the depth to the watertable can be used as an alternative method of assessing potential increase in soil salinity levels. The critical depth is the depth of the watertable from which water can reach the soil surface by capillary rise in sufficient quantities to cause a salt problem. The critical depth, which ranges from 1–6 m under dryland conditions (Nulsen, 1981), will vary with the concentration and composition of the solutes, the frequency and amount of rainfall, soil physical properties, and type of crop (Table 3).

CAUSES OF SALINIZATION

Sources of Salt

The major natural sources of salt are considered to be:

- *Marine sedimentary rocks or seawater intrusion*

Marine salt can be derived from seawater intrusion or ancient marine sedimentation. The chemistry and the ionic composition of seawater are distinctive; therefore, the relative contributions of salt sources can be derived from proportional relationships. If marine salts from recent intrusions or connate salt are the primary sources of salt in sediments or groundwater, then the concentration of specific anions and heavy isotopes in the mixture will vary according to the degree of mixing (Herczeg *et al.*, 2001).

- *Weathering of rocks*

Host rocks experience both physical and chemical weathering. Physical weathering may produce larger soil particles than chemical weathering, which may produce soluble ions, colloidal gels, microcrystalline substances, and clay minerals. The soluble ions may then be leached from the weathering profile, or bound to clay minerals. The chemical composition of soil water and groundwater can be related to the weathering products if the hydrochemistry and the geology is well known (Szabolcs, 1989). The correlation between concentrations of anions and cations produced by chemical weathering should be strong if chemical weathering is the main source of dissolved salts in groundwater (Sami, 1992). The lack of chloride ions can be used as an indication of salinity source more likely to be from weathering, as chloride rarely originates from the weathering processes (Herczeg *et al.*, 1993).

Table 3 Critical depth to saline water table (from Nulsen, 1981)

Depth to saline groundwater (m)	Effect of salinity on ground cover
>2	Negligible effect
<1.8	Wheat yield decreased
<1.5	Barley yield decreased
<1.2	Ryegrass growth affected, replaced with salt – tolerant species.

- *Deposition of sea salt carried in rain and wind*

Marine salt may be transported by the wind as an aerosol and deposited by rainfall or dry fallout. This salt is often referred to as *meteoric salt*. In semiarid to arid areas and near-coastal areas, this can be the dominant source of salt. A portion of this salt is dissolved again in subsequent rainfall, which may recharge aquifer systems. Such contribution of salt is limited by distance from the sea (see Figure 1). In Australia, the rate of meteoric salt fallout ranges from 130–175 kg ha^{-1} y^{-1} at or near-coastal areas to 30–70 kg ha^{-1} y^{-1} 300 to 500 km inland.

Meteoric salt is characterized by an ion composition similar to that of the ocean (dominated by Na and Cl ions) from which it originated (Mazor and George, 1992; Dogramaci and Yesertener, 2001). Hingston and Gailitis (1976) and Blackburn and McLeod (1983) provided detailed mineral ion ratios for rainfall and dry fallout for Western Australia and the Murray–Darling Basin (Australia), respectively.

Salt Accumulation and Leaching

Anthropogenic activities, such as irrigation and clearing native vegetation for dryland agriculture, may result in mobilizing salt in the landscape and groundwater, which can lead to salinization through:

- dissolution of naturally occurring salts in the soil
- salts applied to the soil in artificial and natural fertilizers
- the accumulation of salt in the soil profile as applied water is lost by evaporation and evapotranspiration
- capillary action where groundwater rises and evaporates leaving salts in the upper soil profile.

Salt in the soil solution moves by molecular or ionic diffusion in response to the concentration gradients within the solution and by advection due to the mass flow of the soil solution. The processes of diffusion and advection can occur simultaneously, either in the same direction or in the opposite direction. The transport of salt is further complicated by hydrodynamic dispersion, which is the mixing process during flow due to different flow velocities through the pore spaces. The differences in flow velocities and diffusion in the direction of decreasing concentration result in the dispersion of solutes. Transport of salts may also be affected by the exchange or adsorption onto the surfaces of soil particles. For soil water that is dominated by chloride and sodium ions, chloride is generally regarded as nonreactive while sodium undergoes exchange in some soils.

Salt Balance in Soil

Salt from the surface is transported into the soil by infiltrating water. Salt output from the soil is via throughflow and groundwater discharge or through very minor absorption by

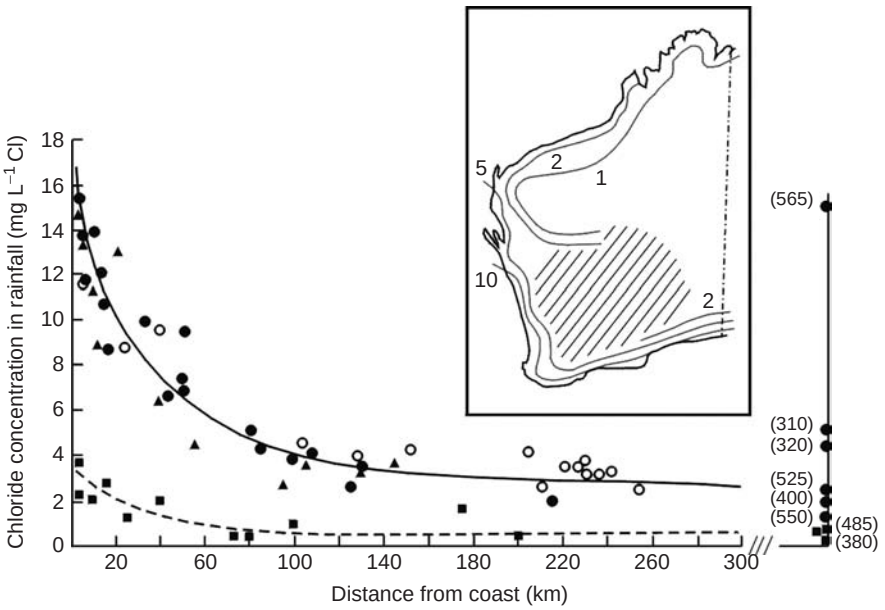


Figure 1 Decrease in concentration of chloride (mg L^{-1} Cl) in rainfall with shortest distance from the coast. The insert map is the chloride concentration in rainwater in western Australia. Hatched area indicates the region of variable values (2–10 mg L^{-1} Cl) (After Hingston and Gailitis, 1976 reproduced by permission of CISRO publishing) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

plant roots. Transport of salt through uptake by vegetation and its return of salt to the soil are largely cyclic. The salinization of soil occurs when salt input exceeds salt output and salt accumulates in the soil. The distribution of salt in the soil profile is influenced by factors that include magnitude of salt input, average annual rainfall, depth of soil, presence or absence of saturated zones, slope, and soil properties such as texture, structure, and permeability. Because the uptake of salt by vegetation is very small in saline areas, it can generally be ignored in the soil salt balance.

The salt balance is the summation of all salt inputs and outputs for a defined volume of soil during a specified period. If salts are conserved, in that they are neither generated nor decomposed chemically in the soil, then the difference between total input and output must equal the change in salt content of the soil zone assessed.

The following simple equation applies to the amount of salt in the liquid phase of the root zone per unit area of land:

$$\Delta M_s = \Delta_w(V_r c_r + V_i c_i + V_g c_g - V_d c_d) \qquad (1)$$

where ΔM_s is the change in mass of salt in the root zone; V_r, V_i, V_g, V_d are the volumes of rain, irrigation, groundwater rise, and drainage respectively, with corresponding salt concentrations of c_r, c_i, c_g, c_d.

As an example, assume:

- Rainfall with a salinity of 20 mg L^{-1} occurred in winter of 350 mm.
- Capillary rise from saline groundwater in spring and autumn totaled 100 mm at a concentration of 2000 mg L^{-1}.
- Irrigation was applied during the summer season and amounted to 900 mm with a salinity of 400 mg L^{-1}.
- Drainage during the irrigation season amounted to 200 mm with a salinity of 800 mg L^{-1}.
- Assume the salt applied in the fertilizer and possible dissolution of soil minerals and salts within the soil are negligible.

Using the above data, the root zone is accumulating salt at a rate of 0.52 kg m^{-2} per year.

Several recent models have been proposed to relate the concentration of drainage water to some clearly definable value of salinity that affects the crop directly. Hoffman and van Genuchten (1983) determined the linearly averaged mean root zone salinity by solving the appropriate equation for one-dimensional vertical flow of water through soil, assuming an exponential soil water uptake function.

The ratio of the linearly averaged concentration of the root zone (c) to the concentration of the irrigation water (c_i) is:

$$\frac{c}{c_i} = \frac{1}{L} + \frac{d}{ZL} + \ln(L + (1\text{-}L))e^{-Z/d} \qquad (2)$$

where L is the leaching fraction, Z is the depth of the root zone, and d is the empirical constant set to 0.2Z.

When more water is applied than can be held by the soil in the crop root zone, the excess water drains below the root system, leaching salts with it. The more water applied in excess of the crop water requirements, the less salt is left in the root zone despite the fact that more salt has actually been added to the field. The term leaching fraction is used to relate the fraction or percent of water applied to the field that actually drains below the root zone. Leaching and drainage are the main mechanisms for removing of salts from the root zone as the uptake of salts by crops contributes little to the removal of salts.

The effectiveness of leaching salt from the root zone in irrigation areas is measured by % efficiency. Surface irrigation systems have leaching efficiencies ranging from 50–80%, and most irrigation systems have leaching efficiencies of 60% or less. The higher the salinity of the irrigation water, the larger the volumes and more frequent applications of water are required. The normal inefficiencies of irrigation systems are often more than sufficient for adequate leaching of salts out of the root zone. However, salt buildup in the root zone resulting from the presence of shallow water tables can lead to salinization despite high theoretical leaching rates.

The leaching of salts and other chemicals out of the root zone needs to be managed in an environmentally responsible way to avoid undue rises of groundwater levels and contamination of underlying groundwater, and to protect surface water into which drainage water is eventually discharged after it leaves the aquifer through natural drainage or constructed drainage or pumping systems (Bouwer, 2002).

In the context of a whole catchment, leaching refers to the time for soil salt to be leached to a stream or river after a change in the net input of water. Peck and Hurle (1973) and, more recently, Hatton *et al.* (2002) estimated leaching times from some of the salinity-affected catchments in Western Australia to range from 30 to more than 500 years.

EFFECTS OF SALINIZATION

Effects on Soil Physical and Chemical Properties

The accumulation of salts in soils can lead to irreversible damage to soil structure. The effects are most extreme in clay soils, where the presence of sodium can result in sodic soils. This occurs by adsorption of sodium ions to clay particles, which, in severely affected soils, can bring about collapse of the soil structure. Sodic soils make growing conditions very poor, soils very difficult to work, and prevent reclamation by leaching using standard techniques. Gypsum in the irrigation water or mixed in the soil can be used to reduce the sodium content of sodic soils.

Quirk and Schofield (1955) proposed the concept of a threshold concentration of irrigation water, which is defined as the concentration of solutes required to maintain soil permeability at an acceptable level relative to that measured with a strong salt solution for a particular value of Sodium Adsorption Ratio (SAR).

$$\mathrm{SAR} = \frac{\mathrm{Na}}{(\sqrt{(\mathrm{Ca} + \mathrm{Mg})/2})} \tag{3}$$

where the ion concentrations are expressed in millimoles per liter.

Problems in soil structure occur when there is excess sodium relative to calcium and magnesium ions in the water applied or in rainfall. Calcium and magnesium ions are divalent cations and are more tightly associated with clays than monovalent sodium cations, and are preferentially adsorbed onto most clays, which generally have negative surface charges.

The excess sodium ions will displace cations from the clay surface because of the law of mass action. This will cause the dispersion of clays and result in the collapse in pore structure. The resultant decreased pore size will decrease the soil permeability and increase the potential for waterlogging. Consequently, the combination of a high EC and a high SAR can create poor soil conditions for agriculture.

Effects on Plants

The effects of salts on plants have been reviewed by Maas (1990) and Pessarakli (1993). Salt-affected plants are usually stunted and may have darker green leaves. As the salt concentration increases above a threshold level, both the growth rate and ultimate size of the plant decrease (Maas and Hoffmann, 1977).

A plant exposed to a high concentration of salts in the root zone will respond almost immediately by reducing the rate of leaf expansion (Moore, 1998). This rapid response is due to the increased osmotic potential hindering water uptake by the roots. Over the long term, excess salt in soils results in a buildup of salt in the leaves, especially the old leaves, resulting in death (necrosis). Salt also causes nutrient deficiencies by hindering nutrient uptake in the root zone because of reduced root growth, water uptake, and competitive displacement by sodium ions.

The salt tolerance of agricultural plants and crops varies widely (Tables 2 and 3). In general, crops tolerate salinity to a threshold level, beyond which yields decrease approximately linearly with increasing salt concentrations. In addition, the combination with other factors such as waterlogging can exacerbate the impacts of salinity.

Effects on Ecosystems

The remaining terrestrial and aquatic ecosystems are threatened by secondary salinization. The overall biological effects are considered to be:

- changes to the natural character (processes and structure)
- loss of biodiversity
- taxonomic replacement (less salt-tolerant species are replaced by more tolerant species).

All of these effects are undesirable and largely irreversible: they cause more or less permanent degradation. The loss of biodiversity is probably greater than generally realized since recent work on both the naturally fresh and saline waters in semiarid and arid regions indicates they have much greater biodiversity than was thought (Williams, 1999). Recent investigations into the impacts of increased salinity on the biodiversity of southwestern Australia suggest that the richness of most biological groups show an inverse relationship with salinity (Hammer, 1986). With respect to shallow lakes, salinization has resulted in the loss of freshwater species of submerged macrophytes and the dominance of a small number of more salt-tolerant species (Davis, 2004).

Many rivers of southwestern Australia have suffered a 20-fold increase in salinity (Schofield *et al.*, 1988), which has led to changes to the biological communities of aquatic ecosystems (Halse *et al.*, 2003). Recent data confirms these trends and further substantial changes are likely to occur, despite the flora and fauna of the southwest of Australia being comparatively well adapted to the existence of salinity in the landscape. Possibly, one-third of wetland and river invertebrates, large numbers of plants, and a substantial proportion of the waterbird fauna will disappear because of the impacts of salinization (Halse *et al.*, 2003).

Other examples of increasing salinity include the Colorado River (southwestern United States) and the Syr and Amu Darya, Central Asia, which discharge into the Aral Sea (Williams, 1999). The salinity of the lower reaches of the Syr Darya has increased by sevenfold in the past 90 years. Ecological effects are largely similar to those for freshwater lakes and wetlands. It is often difficult to separate ecological effects because of secondary salinization from those resulting from the other two major anthropogenic disturbances to rivers and streams in drylands, flow regulation, and diversion. Economic effects are more easily distinguished, when salinities reach 1000 mg L^{-1}. At this level, the water becomes unsuitable as drinking water. Thus, the economic costs may be very large. Managers become alarmed when even small rises occur, as they are, for example, with regard to the River Murray (southeastern Australia), a major source of irrigation water and public water supplies for the city of Adelaide (Ghassemi *et al.*, 1995).

CURRENT EXTENT OF SALINIZATION

Land Salinization

The estimation of the extent of salinization is very difficult due to the dynamic nature and lack of data and inconsistencies between the data provided by various organizations. FAO (2000) provided a global estimate of approximately 400 million ha of land considered to be salt-affected (Table 4). Nearly 50% of the salt-affected land is within Asia, Pacific, and Australia region and includes China, India, and Pakistan, which have the highest portion of irrigation salinity. Australia on the other hand is one of the worst affected by dryland salinity with approximately 5.7 Mha already saline, and it is predicted that a further 17 Mha will be salt-affected in the next 100 years.

River Salinization

The salinity of surface waters is naturally highly variable and there is no global reference value to assess the level of salinity. The average salinity of some of the world's major rivers in the global dataset for rivers (Walling and Webb, 1983) is ~250 mg L^{-1} TDS. Some rivers, such as the Blackwood and Avon rivers in Australia, have high natural salinities or have experienced great increases due to human-induced change (Table 5). In temperate regions, rivers originate in high rainfall zones and the salinity increases from up-gradient to the discharge areas. However, salinity variations along rivers in southwest Australia highlight an atypical characteristic of rivers rising in low rainfall regions – they become fresher as they reach the coast due to fresh tributaries diluting the more saline water coming from inland.

Eighty percent of the rivers within the global dataset have a chemical composition made up of four ions (HCO_3, SO_4, Ca, and SiO_2), which indicate mineral weathering as the major source of salts. Calcium is the most common cation found in the surface waters of many European rivers, and its dominance is mainly a function of geology, especially when carbonate or gypsum deposits are present in the watershed.

Table 4 Extent of salt-affected land in million hectares Mha (After FAO, 2000)

Continent	Land area (Mha)[a]	Salinity-affected(Mha)
Africa	1900	39
Asia, the Pacific, and Australia	3107	195
Europe	2011	7
Latin America	2039	61
Near east	1802	92
North America	1924	5
TOTAL	12 783	399

[a]Mha denotes million hectares.

Table 5 Salinity and constituent concentrations for major rivers worldwide (From GEMS/WATER and Hatton and Ruprecht, 2001)

River	Discharge	TDS	Constituent concentrations ($mg\,L^{-1}$)							
	($10^9\ m^3$)	($mg\,L^{-1}$)	Na	K	Ca	Mg	Cl	SO_4	HCO_3	SiO_2
South America										
Amazon	6450	43.5	1.9	0.8	5.4	0.9	2.2	4.5	21.0	6.9
Orinoco	1135	24.8	1.5	0.7	2.6	0.7	0.9	2.3	10.0	6.3
North America										
Mississippi	580	287	21.5	3.1	40.7	11.3	25.1	54.1	124	7.6
Colorado	0.5	703	95.0	5.0	83.0	24.0	82.0	270	135	9.3
Asia										
Brahmaputra	630	99	2.1	1.9	14.0	3.8	1.1	10.0	58	7.8
Irrawaddy	428	201	30.0	2.0	10.0	6.0	18.0	5.0	120	10.0
Africa										
Congo	1350	42	2.2	1.5	2.7	2.1	3.3	1.5	17.1	11.2
Nile	30	204	8.1	3.2	22.0	5.3	5.7	12.0	135	12.8
Europe										
Danube	214	307	9.0	1.0	49.0	9.0	19.5	24.0	190	5
Volga	214	299	17.9	1.6	50.2	9.9	18.9	62.1	134	4.0
Rhine	74.5	600	91.9	6.4	80.5	11.4	173	74	158	5.2
Australia										
Murray-Darling	7.9	453	101	6.0	21.0	17.0	171	38	94.0	5.0
Blackwood	0.6	1900	505	11.4	38	85.5	958	91.2	127	60.8
Avon	0.4	5200	1383	31.2	120	250	2746	234	302	99

Rivers in Australia, on the other hand, are notable for their marked difference from this global average. In particular, the rivers of southwest Australia are very saline, with the dominant components being Na and Cl. Generally, the concentrations of sodium and chloride are highly correlated in the surface waters in coastal areas, particularly in arid and semiarid regions, indicating marine origin (Herczeg *et al.*, 2001) (Table 5, Figure 2).

Stream salinity in the coastal areas in Australia ranges from 100 in fully forested areas to over $20\,000\,mg\,L^{-1}$ TDS depending on the extent of native vegetation clearing, mean annual rainfall, and geomorphic characteristics (Schofield *et al.*, 1988; Schofield and Ruprecht, 1989). Globally, the mean annual salt load in river discharge to the ocean is $39.5\,t\,km^{-2}$ (Figure 2), though for specific areas there are significant variations. For example, the mean annual salt load can vary from less than $10\,t\,km^{-2}$ to more than $250\,t\,km^{-2}$ for cleared Western Australian catchments in low-rainfall areas (<350 mm). The relatively higher salt load discharging in Australian rivers is mainly due to disturbance to the natural water balance by clearing of native vegetation for agriculture development (Figure 3).

The secondary salinization of rivers and streams is closely related to land salinization. This is because rivers and streams either arise in salinized catchments and discharge salt into lakes and wetlands or flow from them with additional salt loads. Both small streams and large rivers may be subject to secondary salinization. The Blackwood River in southwestern Australia provides a well-documented example of a river that has been salinized during the past 80 years by catchment changes; its catchment has been extensively cleared for agricultural purposes. Less than a century ago, its salinity was $<500\,mg\,L^{-1}$, it is now mostly $>3000\,mg\,L^{-1}$.

Lake and Wetland Salinization

Many freshwater lakes and wetlands occur in semiarid and arid regions and may also suffer secondary salinization after extensive changes to land use in their catchments (Williams, 2001a). Irrigation and clearance of the natural vegetation are the most important reasons. Both processes commonly mobilize subsurface salt that salinizes freshwater lakes and wetlands. Rising saline groundwaters may threaten floodplain wetlands in certain regions. Examples of secondarily salinized freshwater lakes and wetlands include Lake Qarun, a formerly freshwater lake in southwest of Cairo, Egypt, which has become saline because of salt inflows in drainage water from nearby agricultural land (Williams, 2001b). Increases in the salinity of some freshwater lakes in the Rift Valley of Ethiopia have been attributed to irrigation, river diversions, and deforestation (Williams, 2001b).

Many large and permanent salt lakes have also been modified and affected by secondary salinization. The human-induced increase in salinity of these salt lakes is also caused by diversion from inflowing rivers: with decreased freshwater inflows, lake volumes decrease and so solutes may become more concentrated (Kefford, 2000).

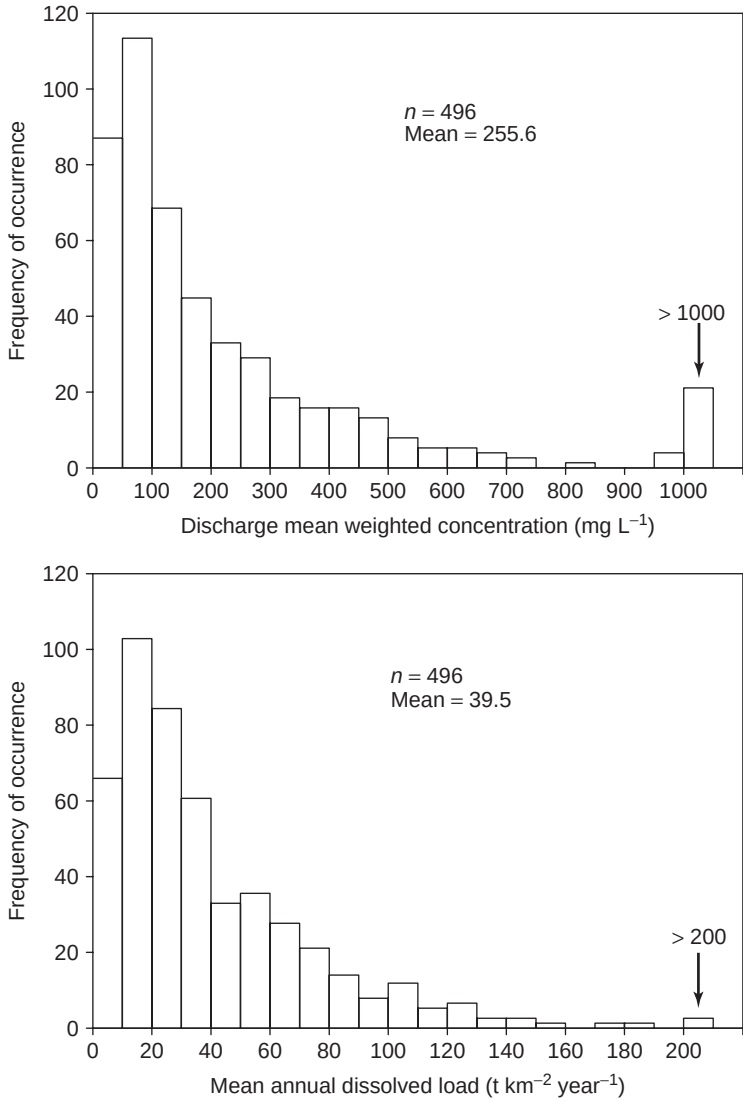


Figure 2 Frequency distribution of discharge-weighted mean total dissolved solids concentration and mean total dissolved load for a sample of 496 world rivers (Adapted from Walling and Webb, 1983)

A few examples of human-induced change on ecosystems are also known from temperate regions. Some of the artificial lakes created by open-cast coal-mining in Germany are now being salinized by saline groundwater, and salt mines have salinized several freshwater lakes in Cheshire, UK. Reservoirs too can be threatened by secondary salinization. Waters impounded by the Imperial Dam on the lower reaches of the River Colorado in the United States, will soon have salinities $>1100\,\mathrm{mg\,L^{-1}}$ (Ghassemi *et al.*, 1995). Salinity is a major environmental problem for this river and its many impoundments. Even small salinity rises can be significant for freshwater lakes and wetlands since their biota generally has a limited salt tolerance. The disappearance of macrophytes and riparian trees is one of the first signs of secondary salinization.

Apart from ecological damage (changes in natural character, biodiversity loss, taxonomic replacement), secondary salinization causes many other problems. Economic losses include the loss or reduced value of the lake or wetland for water supplies. Decreases in conservation, amenity, aesthetic, and general environmental values also occur. For floodplain wetlands, losses include degradation of the riverine system as a whole because of the close hydrological and ecological relationship between floodplain wetlands and rivers (e.g. Davies *et al.*, 1994).

MANAGEMENT OF SALINITY

Throughout many salinity-affected areas (irrigation and dryland salinity), the change in water balance has raised

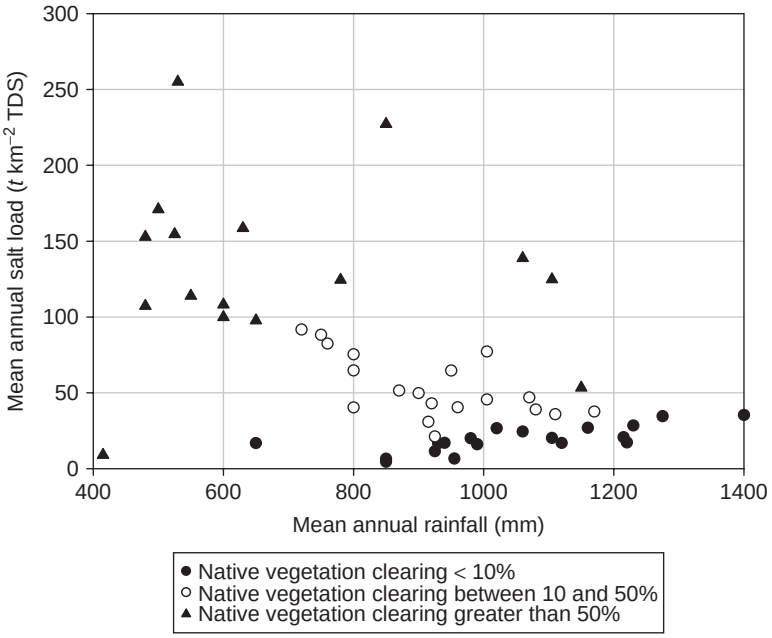


Figure 3 Relationship between mean annual rainfall and mean annual salt load for rivers in southwest Australia

saline groundwater levels, resulting in groundwater discharging to the surface and into natural waterways. These discharge areas become saline, often waterlogged, support only salt-tolerant vegetation, and suffer from soil erosion. Shallow watertables are as much an environmental problem as an issue to agriculture and infrastructure. Management of saline watertable in these areas aims to:

- stabilize watertables at harmless levels in irrigation areas with good quality groundwater,
- significantly retard rising groundwater by control of groundwater recharge, and, where necessary, adapt to higher saline conditions in irrigation areas with saline groundwater,
- significantly reduce accessions to the groundwater systems and to substantially improve salinity problems associated with localized groundwater systems
- protect, and where appropriate, rehabilitate high-value wetlands and other significant environmental features.

The main purpose of salinity management options is to restore the water balance by either reducing the infiltration of excess water (recharge management), increase discharge by engineering options, or combining both approaches. Because irrigation salinity is intrinsically associated with increased recharge, the primary management option is to enhance discharge by engineering work such as drains and groundwater pumping. For more discussion on irrigation management for salinity control, see the reviews of Kruse *et al.* (1990) and Hillel (2000).

Engineering Options

Drainage

Drainage has been essential for managing rising watertables in most irrigation schemes around the world. It is considered necessary for irrigation to maintain plant growth, and prevent waterlogging and soil salinization. The major types of surface drainage are seepage interceptor drains, horizontal subsurface drains, and vertical or tube-well drains. One of the major problems associated with drainage is the loss of water by leakage resulting in inefficient management of groundwater. Up to 40% of drainage water is lost via evaporation and leakage in some drainage schemes (Ghassemi *et al.*, 1995). Improving irrigation water efficiency by means such as irrigation channel lining or piping, land forming, and grading, changing the method of irrigation, or by irrigation scheduling can all reduce land and water salinization substantially.

The ranges of depth and spacing generally used for the placement of drains in field practice are shown in Table 6.

Groundwater Pumping

There is a growing awareness of the salinity management benefits of groundwater pumping and the need for integrated groundwater and surface water solutions. Thousands of groundwater pumping bores have been installed to lower the watertable in most of the irrigation schemes of Pakistan and India (Ghassemi *et al.*, 1995). In areas where a river drains saline aquifer, interception schemes can be used to reduce the discharge of the saline groundwater to the river.

Table 6 Prevalent depths and spacing of drains in different soil types (After Hillel, 2000)

Soil type	Hydraulic conductivity (m day^{-1})	Spacing of drains (m)	Depth of drains (m)
Clay	1.5	10–20	1–1.5
Clay loam	1.5–5	15–25	1–1.5
Loam	5–20	20–35	1–1.5
Fine, sandy loam	20–65	30–40	1–1.5
Sandy loam	65–125	30–70	1–2
Peat	125–250	30–100	1–2

The basic design principle assumes that the extraction of groundwater from an aquifer hydraulically connected to the river, using a line of pumping wells positioned close and roughly parallel to the river, will decrease the hydraulic gradient of the aquifer toward the river.

The release of freshwater from reservoirs in a river basin during periods of low river flow that coincide with high river salinity can also reduce river salinity (Schofield *et al.*, 1988).

Disposal of Saline Water

- *Reuse option*

Drainage water can be reused repeatedly to irrigate more salt-tolerant crops. Drainage water can also be reused for repeated irrigation to the extent that it still has value for use by a crop of higher salt tolerance (Rhoades, 1984). When the drainage quality is such that its potential for reuse is exhausted, then the drainage water could be disposed of to evaporation ponds.

- *Disposal to rivers and lakes*

This option is widely practiced by either controlled or uncontrolled means. In controlled disposal, the volume, salinity, and time of disposal are managed. Effluent from the drainage facilities is collected into holding basins or evaporation basins and released to the river during periods of high river flow.

- *Disposal to evaporation basins*

Drainage water can also be disposed directly to naturally occurring evaporation basins or ponds, (formed from natural depressions, saline lakes) or constructed ones. They can vary in size from a few hectares serving individual farms to very large basins. The design, maintenance, and cost-benefit analysis of evaporation basins have been the focus of a detailed investigation in the Murray Darling Basin in Australia (Leaney *et al.*, 2000; Jolly *et al.*, 2000; Dowling *et al.*, 2000). Leakage, laterally or vertically, is a major problem for evaporation basins. Lateral leakage is often the principal undesired consequence because of its effect on surrounding land within a relatively short period of time. Vertical leakage beneath a basin creates a groundwater mound that may raise groundwater levels next to the basin. It is important to note that the evaporative capacity of the saline disposal basins is reduced with increasing salinity due to evaporation (Table 7).

Agronomic and Vegetation Options

Agronomic Practices

The water balance of a catchment can be manipulated to some extent by altering the agronomic practices. One of the best documented examples is the replacement of shallow-rooted grasses with deep-rooted alfalfa on recharge areas of the Northern Great Plains of the United States (Halvorson and Reule, 1980). Saline seeps that affected a large area of productive dryland agriculture in the region have been controlled with an intensive cropping system and deep-rooted crops. Alfalfa, when grown on about 80% of the recharge area, reduced the deep percolation of soil water and provided hydrologic control within one year of establishment (Halvorson and Reule, 1980).

Productive Use of Salt-Affected Soils and Saline Water

There has been considerable research into the productive use of saline land by revegetation and other means. Salt-tolerant crops and pastures (grasses and legumes) have considerable potential for the adaptation of saline land. Saline aquaculture is also a productive option for the safe storage of saline water.

Revegetation Options

Tree-based land management strategies can contribute significantly to minimizing the spread of salinity and even to recovering land from salinization. The hydrological response of catchments to tree planting varies enormously, making it difficult to provide specific advice on strategies to achieve salinity benefits. In some catchments previously cleared for agriculture, local strategic plantings can have significant and almost immediate effects on waterlogging and groundwater discharge. In others, afforestation of the majority of the catchment is required to significantly limit salinization, and the timescale to do so may be very long: up to hundreds of years (Hatton *et al.*, 2002).

Revegetation can control salinization by either reducing recharge or by increased water use at discharge sites.

Table 7 Reduction of evaporation potential of water with different salinity

Salinity (mg L^{-1} TDS)	Percentage of freshwater evaporation (%)
Freshwater	100
30 000 (seawater)	99
140 000	90
300 000 (saturation)	59

Revegetation strategies can include discharge plantings, dense plantations covering a high proportion of area, wide-space plantations, and strips or small blocks strategically placed but covering a small proportion of area.

Hatton *et al.* (2002) described the key diagnostic properties of a catchment with respect to its hydrologic response to tree planting to be:

- the discharge capacity of the aquifer
- the nature of the recharge process: diffuse compared to localized recharge, its spatial distribution, and the rainfall environment
- the size of the groundwater system
- the salinity of the soil and the groundwater.

Policy and Management Options

Policy options such as water pricing, transferable water entitlements, and integrated natural resource or catchment management are also important and should be considered in any strategy of salinity mitigation for both dryland and irrigation salinity.

Possible economic instruments for addressing salinity include tradable emissions permits, auction-based systems for allocating rights, charges, and subsidies, regulations, or voluntary programs (Pannell, 2000).

In Australia, a recent initiative has been the National Action Plan for Salinity and Water Quality, which is a partnership between Commonwealth and State governments, and regional management bodies (Australian Agricultural Assessment, 2001). The purpose of the action plan is to identify high priority, immediate actions to deal with salinity, particularly dryland salinity, and deteriorating water quality in key catchments and regions across Australia. The plan includes:

- targets and standards for natural resource management
- integrated catchment/regional management plans
- capacity building for communities
- an improved governance framework
- clearly articulated roles for the Commonwealth, State/Territory, local government, and the community
- a public communication program.

ASSESSING LAND AND WATER SALINIZATION

The main approaches to assessing the impacts of land and water salinization and to evaluating potential solutions include process studies, catchment studies, modeling, and statistical analysis.

Process Studies

Process studies identify the hydrologic processes causing the increased land and water salinity. This approach is used to evaluate the impacts of salinity management on the various components of the hydrological balance at small paddock and large catchment scale. For studies of water quantity and quality and the required measurements and methods, see Calder (1992). Process studies may include studies of the effectiveness of vegetation or engineering options to manage salinity such as:

- Afforestation
- Agronomic options
- Deep drainage
- Groundwater pumping
- Siphon wells

Process studies are particularly valuable when associated with properly designed catchment experiments so that the reasons for the observed changes to the hydrology and salinity can be determined.

An example of a process study is one conducted near Kyabram in Victoria (Australia), which monitored a tree plantation with groundwater conductivity in the range 3000 to 5000 $\mu S\,cm^{-1}$ for more than 20 years (Silberstein *et al.*, 1999). Over this time, the trees lowered the watertable from 1 to 5 m below the natural surface. The local depression in the watertable continued to deepen as long as groundwater use exceeded the rate of lateral inflow from the surrounding irrigated pasture. Tree water use estimates over the initial 10 years indicated that the trees used approximately 730 mm per year, of which 430 mm came from the watertable (Silberstein *et al.*, 1999). However, sap flow readings five years later suggested that groundwater use was falling as salt accumulated in the root zone. If this trend continued, the watertable would start to rise again. These field studies, when combined with process modeling, indicate that the tree plantation at the Kyabram site is unsustainable due to salt accumulation. In addition, Silberstein *et al.* (1999) suggest the general rule that, in situations where the accumulated salt cannot be exported, tree plantations grown in areas with a groundwater salinity of 3000 $\mu S\,cm^{-1}$ or greater should be considered as unsustainable.

Catchment Experiments

Catchment studies are designed to investigate the effects or define major processes of salinization at a catchment scale. Paired catchment studies that use a control to evaluate the impacts of salinization are preferable. In a single-catchment experiment, the effects of a land use change are measured by comparing measurements made before and after the change occurred. In a paired catchment experiment, land use is held constant on the control catchment and changed on the treatment catchment. The differences in the hydrologic response and water quality can then be compared with catchments experiencing the same weather patterns. In single-catchment experiments, it

may be difficult to separate the effect of land use change from the effects of different weather patterns before and after the land use change. Paired catchment studies are more costly, but provide more accurate results because the effects of change are being measured under the same weather conditions on the two catchments.

It is critical in the design of catchment experiments, particularly those concerned with the water and salt balance, that the experimental error attached to the effect being studied is not larger than the effects themselves (Calder, 1992).

Catchment Experiment – Case Study

Long-term coupled experimental catchment studies are fundamental in understanding the long-term impact of clearing on the development of land and stream salinity. In 1973, the Western Australian Government and CSIRO (Commonwealth Scientific and Industrial Research Organisations) established five experimental catchments (Table 8) to quantify the salinity impacts of agricultural development. Three of the catchments were cleared in 1976 using a range of clearing strategies and the other two catchments were left untreated and considered as control catchments.

The major research findings, summarized in Peck and Williamson (1987), Ruprecht and Schofield (1989) and Ruprecht and Schofield (1991a, b) are considered to include:

- a rapid rise in groundwater level following clearing, even where the preclearing level was more than 10 m from the ground surface. Although the response time-lag varied, groundwater levels at the range of depths and landscape positions responded rapidly to land clearing in both the high and medium rainfall zones.
- preferred flow paths were the dominant routes for water recharging aquifers postclearing
- the characteristic shape of salt profiles identifies the nature of infiltration in the unsaturated zone and groundwater recharge
- groundwater discharge areas are source areas for runoff and streamflow. The increase in streamflow following clearing was strongly correlated with the development of areas of saturated soil associated with groundwater seepage (Ruprecht and Schofield, 1989).
- streamflow and salt load increased immediately after clearing at the high rainfall site (Wights), with a fourfold increase in streamflow and a 15-fold increase in salt load compared to preclearing levels
- streamflow and salt load from the Lemon Catchment increased dramatically when saline groundwater reached the soil surface in the valley of the catchment (Ruprecht and Schofield, 1991a).

Table 8 Summary of Collie research catchments

Catchment	Treatment	Rainfall zone / salt risk
Salmon	Control	High – low salt risk
Wights	100% cleared	High – low salt risk
Ernies	Control	Medium – high salt risk
Lemon	50% cleared	Medium – high salt risk
Dons	38% cleared in parkland and strips	Medium – high salt risk

Geophysical Approaches to Salinity Mapping

Airborne geophysical techniques, primarily magnetic and radiometric, have been used as the primary methods of mineral exploration since the 1950s. The development of new geophysical technologies, namely, *airborne electromagnetic systems* (AEM) and significant improvements in AEM data processing software and data visualization technique provided the possibility of rapidly obtaining salinity related data over large areas that could assist in dryland salinity investigations. Electromagnetic surveys provide data on the nature of the regolith and the distribution of salt in the landscape. Application of these datasets for land management began in the mid-1980s, although significant analysis and interpretation for farm and catchment planning has been undertaken since 1993 (Rhoades *et al.*, 1997; George, 1998).

Airborne geophysical datasets provide a cost-effective method for mapping, monitoring, and providing better understanding of the surficial geology, particularly when dealing with high-value assets. Research has primarily been concentrated in identifying high conductivity areas, which may suggest the presence of high salt concentrations and the detection of geological structures that impede or facilitate enhanced groundwater flow. These areas can be targeted as priority areas in managing dryland salinity on catchment scale.

Modeling Approaches

Modeling approaches that describe and quantify basic hydrological processes and phenomena under a range of conditions are very useful in salinity investigations, particularly in the assessment of management options. Jakeman *et al.* (1987) delineated the types of models required, and discussed some problems in their construction and calibration.

Groundwater models are useful if a management strategy is required to consider the effects of rainfall, irrigation, cropping activity, groundwater pumping, and land use change on groundwater levels, and on land and stream salinity. Hydrologic routing and surface water quality models can be used to predict downstream salinity concentrations, for management strategies such as salinity dilution by release of additional discharge from upstream storages; to provide advance warning of salinity levels that are too high for irrigation usage; and for quantifying saline inflow within a river reach.

The complexity of model development for specific cases and the lack of required field data are major reasons why some models are used more frequently than others. For example, groundwater, stream routing, and surface water models are used more frequently than complex groundwater solute transport models and unsaturated flow and transport models.

The response of salinity or salt concentration to changes in stream and river flow has been studied in several ways, but investigation of rating relationships between salinity concentration and streamflow represents the traditional approach (Walling and Webb, 1986). Such relationships typically take the form:

$$c = aQ^b \tag{4}$$

where c is solute concentration, Q is streamflow, a is the regression constant, and b is the regression exponent. Examples of more complex relationships of salinity to streamflow are given by Walling and Webb (1986).

Several factors may influence the form and timing of stormy-period chemographs and, in turn, generate hysteretic behavior. In particular, a "flushing effect" has been noted in many rivers (e.g. Walling and Foster, 1975) by which soluble material accumulated during the prestorm period is mobilized and transported to the stream where it influences solute concentrations during the early stages of a storm event. In some cases, dilution in solute concentrations associated with a storm event may be preceded by increasing concentrations. Solute accumulation before a flood event will be affected by evaporation of soil moisture, capillary rise, and the buildup of dry fallout deposits, leaf residues, and dead plant material (Walling and Foster, 1975). Flushing effects are particularly pronounced in autumn storms following solute accumulation over the summer period, and Klein (1981) has demonstrated that solute concentrations in overland subsurface flows are systematically related to the length of the dry period preceding a storm.

In contrast to the flushing effect, solute stores within a catchment may exhaust during a sequence of closely spaced flood peaks (Walling and Webb, 1986). Progressive exhaustion of solute stores may cause a systematic shift in the position of the hysteretic loop as reported by Cornish (1982) from an investigation of German's Creek catchment in New South Wales.

The form and timing of solute responses may be significantly affected by contrasts in the chemical concentration of surface and subsurface flow components (Anderson and Burt, 1982). Water flowing by different routes over and through the soil at varying rates in separate phases of a storm event will have differential access to soluble material distributed unevenly within the soil profile (Walling and Webb, 1986).

There are several examples of models developed in South Africa to simulate the salinity of river inputs to reservoirs (Hall and Gorgens, 1979) in which solute pickup subroutines have been added to existing runoff models.

The exponential form of leaching was used by Hall and Gorgens (1979) for a large catchment in South Africa. The basic equation of leachate concentration was:

$$C = C_o e^{(-R.t/W)} \tag{5}$$

where R is recharge
W is the soil water storage
C is instantaneous leachate concentration
C_o is initial leachate concentration

Soil Water Modeling

At a more detailed level, Bresler (1981) and Raats (1981) reviewed the process of transport and residence times of salt and water through soil systems. Raats (1981) considered the input–output relationships for a two-dimensional saturated soil system and found an exponential distribution of arrival times. It has also been used in theoretical discussions of input/output relationships (Eriksson, 1971), leaching of solutes from a laboratory model (Peck, 1973), and the chloride balance of some farmed and forested catchments in southwestern Australia (Peck and Hurle, 1973).

Process models of water and solute movement include steady-state soil salinity models to predict salinity in the root zone from irrigation application and associated water salinity (Oster and Rhoades, 1990); models that simulate the dynamics of salt movement (accumulation and leaching) in the soil profile and plant response to salinity (Simunek and Suarez, 1994; Huston *et al.*, 1990; Simunek *et al.*, 1994); models that simulate irrigation return flow and its salinity (Aragues *et al.*, 1990); and models that simulate solute transport in the saturated zone (Table 9).

Catchment Modeling

Although there has been considerable progress in understanding the complexity of water and salt flows, most models are site specific and do not have wider applicability. In particular, more work is still required in the validation of models and on increasing their applicability for management decisions.

Small-scale, physically based catchment models that include salinity processes include TOPOG (Vertessy *et al.*, 1996) and WEC-C (Croton and Barry, 2001). In Australia, TOPOG and WEC-C have been applied to small experimental catchments (Silberstein *et al.*, 1999; Bari and Croton, 2002). LASCAM (Sivapalan *et al.*, 1996b) is a conceptual hydrology model that can represent the dryland salinity processes following land use changes of large catchments (applied on catchments up to 100 000 km^2 in area). This model has been applied to catchments affected by dryland salinity in Australia (Sivapalan *et al.*, 1996a, b).

More recently, the groundwater flow model MODFLOW (Michael *et al.*, 1988) has been used to predict the extent

Table 9 Summary of some models applicable to salinization

Model group and name	Description	Reference
Process models		
AgET	1-D soil water model for various vegetation and soil profiles	Argent and George (1997)
UNSATCHEM	2-D transport model for variably saturated porous media with major ion chemistry	Simunek and Suarez, 1994
LEACHM	Leaching estimation and chemistry model – A process-based model of water and solute movement, plant uptake, and chemical reactions in the unsaturated zone	Wagenet and Hutson, 1992
DRAINSAL	Two-dimensional, finite element model that provides long-term predictions of the desalinization of a tile-drained soil and the associated changes in the quality of groundwater and drain effluent	Kamra and Rao (1994)
Hydrus 2D	HYDRUS-2D is a Windows-based modeling environment for analysis of water flow and solute transport in variably saturated porous media.	Simunek and van Genuchten, 1999
SALTMOD	Predicts soil salinity and salt content of groundwater and drainage effluent in irrigated agricultural land.	Rao *et al.* (1992)
DRAINMOD	Simulate the hydrology of poorly drained, high-watertable soils on an hourly or daily basis.	Skaggs (1980)
Catchment models		
CATSALT	Water balance and salt transport modeling at property and catchment scales	Beale *et al.* (2000)
LASCAM	Large-scale catchment model that predicts daily water volume and salt loads	Sivapalan *et al.* (1996 a, b)
WEC-C	Physically based small catchment model that predicts movement of water and salinity	Croton and Barry (2001)
TOPOG - DYNAMIC	Daily time-step model that describes how water moves through landscape. Includes solute movement	Vertessy *et al.* (1996)
IHACRES	Catchment scale rainfall-runoff model based on unit hydrograph approach	Jakeman *et al.* (1990)
MODFLOW	Groundwater flow model	Michael *et al.* (1988)
MIKE SHE	MIKE SHE is a modeling system describing the flow of water and solutes in a catchment in a distributed physically based way.	Refsgaard and Storm (1995)
GIS other related models		
MAGIC	GIS-based steady-state water and salt catchment model	Mauger (1996)
WSIBal	Lake model that models water and salt balance	Barr *et al.* (2000)
IQQM	River basin model of water and salt flows and stores for regulated and nonregulated systems	DLWC (1995)
SWAGMAN	Models impacts of management options and climate on watertable, salinization, and yield.	Meyer *et al.* (1996)
Flowtube	Simple groundwater model to allow evaluation of management options	Clarke *et al.* (1999)

of saline land and evaluate the various engineering options such as drains and groundwater pumping to control land salinization (Salama *et al.*, 1993; Dogramaci *et al.*, 2002). The results of these studies indicate that MODFLOW is a powerful tool to delineate and map areas underlain by a shallow watertable (the salinity risk area).

To cover all processes that contribute to salinity development in large-scale catchments with one model, the MIKE SHE (Refsgaard and Storm, 1995) modeling system can be utilized. The model is an integrated and distributed, physically based, finite difference model. MIKE SHE comprises a number of flow modules, which may be combined to describe integrated surface water–groundwater dynamics and flow within the entire land-based part of the hydrological cycle, or tailored to studies focusing on areas of particular interest.

Statistical Approaches

Helsel and Hirsch (1992) provide a summary of applied statistical methods appropriate for river salinity analysis. The approaches include comparison of datasets, describing uncertainty, several approaches to developing relationship between variables such as flow and salinity (including nonlinear regression and LOWESS), and trend analysis.

Richter *et al.* (1993) provide a comprehensive discussion of both graphical and statistical techniques of data analysis that can be utilized to analyze the water quality data and aid in identification of sources of groundwater salinization.

CONCLUDING COMMENTS

Salinization of land and water resources is a major environmental and economic problem facing many parts of the world. Soil salinization and the consequent degradation of agricultural land are proceeding so fast that, unless policies and approaches change, many countries will not be able to achieve sustainable agriculture to feed their populations in the foreseeable future. Many important ecosystems and internationally recognized wetlands are also under serious threat from the increased salinity.

Research on the fundamentals of managing land and water resources threatened by salinity through anthropogenic activity needs to be advanced to guarantee a sustainable environment. This research needs to be solution focused, leading to the development of sustainable land and water management practices that meet not only production goals, but also good social and environmental outcomes.

FURTHER READING

Genuchten M.T.h, F.J. Leij and L. Wu (Eds.) (1999) *Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, University of California: Riverside, 1523–1536.

George R.J. and Dogramaci S. (2000) Toolibin – A life and death struggle for the last freshwater Wheatbelt lake: Hydro 2000. *3rd International Hydrology and Water Resources Symposium of the Institution of Engineers*, Vol. 2, Australia, 733–738.

Hart B.T., Bailey P., Edwards R., Hortle K., James K., McMahon A., Meredith C. and Swadling K. (1990) A review of the salt sensitivity of the Australian freshwater biota. *Hydrobiologia*, **210**, 105–144.

Huston J.L., Dudley L.M. and Wagenet R.J. (1990) Modelling transient root zone salinity. In: *Agricultural Salinity Assessment and Management*, Tanji K.K. (Ed.), American Society of Civil Engineers: New York, pp. 482–503.

REFERENCES

Anderson M.G. and Burt T.P. (1982) The contribution of throughflow to storm runoff: An evaluation of a chemical mixing model. *Earth Surface Processes and Landforms*, **7**, 565–574.

Aragues R., Tanji K.K., Quilez D. and Faci J. (1990) Conceptual irrigation return flow hydrosalinity model. In: Tanji K.K. (Ed.), *Agricultural Salinity Assessment and Management*, American Society of Civil Engineers: New York, pp. 504–529.

Argent R.M. and George R.J., (1997) *Wattle – A Water Balance Calculator for Dryland Salinity Management. MODSIM 97*, International Congress on Modelling and Simulation: Hobart.

Australian Agricultural Assessment (2001) *National Land & Water Resources Audit, A Program of National Heritage Fund*, National Land and Water Resources Audit, C/-Land & Water, Australia, Vol. 1, ISBN: 0642371296.

Bari M. and Croton J. (2002) Assessing the effects of valley reforestation on streamflow and salinity using the WEC-C model. *27th Hydrology and Water Resources Symposium*, Melbourne, May 20–23, 2002.

Barnes I. (1970) Metamorphic waters from the Pacific Tectonic Belt of the west coast of the United States. *Science*, **168**, 973–975.

Barr A.D., Turner J.V. and Townley L. (2000) WSIBal: A coupled water, conservation solute and environmental isotope mass balance model for lakes and other surface water bodies. *Tracers and Modelling in Hydrogeology*, *Proceedings of the TraM 2000 Conference*, IAHS Publication No. 262: Liege, pp. 539–544, held in May 2000.

Beale G.T.H., Beecham R., Harris K., O'Neill D., Schroo H., Tuteja N.K. and Williams R.M., (2000) *Salinity Predictions for NSW Rivers within the Murray-Darling Basin*, Centre for Natural Resources (CNR), NSW Department of Land and Water Conservation (DLWC): Australia.

Blackburn G. and McLeod S. (1983) Salinity of atmospheric precipitation in the Murray-Darling Drainage Division, Australia. *Australian Journal of Soil Resources*, **21**, 411–434.

Bouwer H. (2002) Integrated water management for the 21st century: Problems and solutions. *Journal of Irrigation Drainage Engineering ASCE*, **128**, 193–202.

Bresler E. (1981) Transport of salts in soils and subsoils. *Agricultural Water Management*, **4**, 35–62.

Calder I.R. (1992) Hydrologic effects of land-use change. In *Handbook of Hydrology*, Maidment D. (Ed.), McGraw-Hill, pp. 13.1–13.50.

Casey H.E. (1972) *Salinity Problems in Arid Lands Irrigation: A Literature Review and Selected Bibliography*, University of Arizona, Office of Arid Lands Studies: Tuscon Arizona, p. 300.

Clarke C., George R., Reggiani P. and Hatton T. (1999) *The Effect of Recharge Management on the Extent and Timing of Dryland Salinity in the Wheatbelt of Western Australia*, Preliminary computer modelling. Report to WA Salinity Council, June, 1999.

Cornish P.M. (1982) The variations of dissolved ion concentration with discharge in some New South Wales streams. In: *The First National Symposium on Forest Hydrology*, O'Loughlin E.M. and Bren L.J., 1982 Melbourne 11–13 May, 67–71.

Croton J.T. and Barry D.A. (2001) WEC-C: a distributed deterministic catchment model-theory formulation and testing, *Environmental Modelling and Software*, **16**, 583–599.

Datta K.K. and Jong C. (2002) Adverse affect of waterlogging and soil salinity on crop and land productivity in northwest region of Haryana, India. *Agricultural Water Management*, **57**(3), 223–238.

Davies B.R., Thomas M.C., Walker K.F., O'Keefe J.H. and Gore J.A. (1994) Dryland rivers: Their ecology, conservation and management. In *The Rivers Handbook*, Vol. 2, Calow P. and Petts G.E. (Eds.), Blackwell Science: Oxford.

Davis J. (2004) Valleys of salt, channels of water, pools of life. In *Engineering Salinity Solutions*, 1st *National Salinity Engineering Conference*, Dogramaci S. and Waterhouse A. (Eds.), Perth, Western Australia, pp. 367–373, ISBN, 085825834X.

DLWC (1995) *Integrated Quantity-Quality Model (IQQM)*, Reference Manual. DLWC Report No. TS94.048, Parramatta.

Dogramaci S. and Waterhouse A. (Eds) (2004) Engineering salinity solutions. *1st National Salinity Engineering Conference*, Engineers Australia, p. 554, ISBN, 085825834X, 9–12 November 2004.

Dogramaci S.S. and Yesertener C. (2001) *The Origin of Dissolved Solutes in the Blackwood River*, XXXI International Association of Hydrogeologists Congress: Munich, 10–14 September.

Dogramaci S., Mauger G.W. and George R.J. (2002) *Engineering Options as Tools For Salinity Management in the Spencer Gully Catchment*, *Salinity Land Use Impacts, Series, SLUI 5*, *Water and Rivers Commission*, p. 35.

Dowling T., Walker G. Jolly I., Christen E. and Murray E. (2000) *On-Farm and Community-Scale Salt Disposal Basins on the Riverine Plain, Testing a GIS-Based Suitability Approach for Regional Planning*, CSIRO Land and Water Technical Report 3/00. p. 121.

Ergil M.E. (2000) The salinization problem of the Guzelyurt Aquifer, Cyprus. *Water Research*, **34**(4), 1201–1214.

Eriksson E. (1971) Compartment models and reservoir theory. *Annual Review of Ecology and Systematics*, **2**, 67–84.

FAO (2000) *Extent and Causes of Salt-affected Soils in Participating Countries*, `http://www.fao.org/ag/agl/agll/spush/topic2.htm`.

Freeze R.A. and Cherry J.A. (1979) *Groundwater*, Prentice-Hall: New Jersey, p. 604.

Funakawa S., Suzuki R., Karbozova E., Kosaki T. and Ishida N. (2000) Salt-affected soils under rice-based irrigation agriculture in southern Kazakhstan. *Geoderma*, **97**(1–2), 61–85.

GEMS/WATER (2004), `http://www.cciw.ca/atlas-gwq/forward-e.html.`.

George R.J. (1998) *Evaluation of Airborne Geophysics for Catchment Management*, National Airborne Geophysics Project, Published electronically, Toolibin: Western Australia, p. 86.

Ghassemi F., Jakeman A.J. and Nix H.A. (1995) *Salinization of Land and Water Resources. Human Causes, Extent, Management and Case Studies*, University of New South Wales Press: Sydney.

Hall G.C. and Gorgens A.H.M. (1979) Modelling runoff and salinity in the Sundays River, Republic of South Africa. In: *The Hydrology of Areas of Low Precipitation*, Vol. 120, International Association of Hydrological Sciences, pp. 323–330.

Halse S.A., Ruprecht J.K. and Pinder A.M. (2003) Salinization and prospects for biodiversity in rivers and wetlands of south-west Western Australia. *Australian Journal of Botany*, **51**, 673–688.

Halvorson A.D. and Reule C.A. (1980) Alfalfa for hydrologic control of saline seeps. *Soil Science Society of America Journal*, **44**, 370–374.

Hammer U.T. (1986) *Saline Lakes Ecosystems of the World*, Dr W. Junk Publishers: Dordrecht, p. 616.

Hatton T., Reggiani P. and Hodgson G. (2002) The role of trees in the water and salt balances of catchments. In *Trees, Water and Salt: An Australian Guide to Using Trees for Healthy Catchments and Productive Farms*, Stirzaker R., Vertessy R. and Sarre A. (Eds.), Joint Venture Agroforestry Program, Publication No. 01/086, Rural Industries Research and Development Corporation, p. 159.

Hatton T.J. and Ruprecht J.K. (2001) Watching the rivers flow. *Wheatbelt Valleys Conference*, Merredin July 31-August 2, 2001.

Helsel D.R. and Hirsch R.M. (1992) *Statistical Methods in Water Resources*, Elselvier Science Publishers.

Herczeg A.L., Dogramaci S.S. and Leaney F.W. (2001) Origin of dissolved salts in a large semi-arid groundwater system: Murray Basin, Australia. *Marine Freshwater Resources*, **52**, 41–52.

Herczeg A.L., Simpson H.J. and Mazor E. (1993) Transport of soluble salts in a large semiarid basin: River Murray, Australia. *Journal of Hydrology*, **144**, 59–84.

Hillel D. (2000) *Salinity Management for Sustainable Irrigation: Integrating Science, Environment and Economics*, The International Bank for Reconstruction and Development, The World Bank: Washington, p. 92.

Hingston F.J. and Gailitis V. (1976) The geographic variation of salt precipitated over western Australia. *Australian Journal of Soil Resources*, **14**, 319–350.

Hoffman G.J. and van Genuchten M.T.h (1983) Soil properties and efficient water use: Water management for salinity control. In *Limitations to Efficient Water Use in Crop Production*, Taylor H.M., Jordan W.R. and Sinclair T.R. (Eds.), American Society of Agronomy: Madison.

Hopkins D.G. and Richardson J.L. (1999) Detecting a salinity plume in an unconfined sandy aquifer and assessing secondary soil salinization using electromagnetic induction techniques, North Dakota, USA. *Hydrogeology Journal*, **7**(4), 380–392.

Jacobsen T. and Adams R.M. (1958) Salt and silt in ancient Mesopotamian agriculture. *Science*, **128**(3334), 1251–1258.

Jakeman A.J., Littlewood I.G. and Whitehead P.G. (1990) Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology*, **117**, 275–300.

Jakeman A.J., Thomas G.A., Ghassemi F. and Dietrich C.R. (1987) Salinity in the river Murray basin: Management and modelling approaches. *Search*, **18**(4), 183–188.

Jolly I., Christen E., Gilfedder M., Leaney F., Trewhella B. and Walker G. (2000) *On-Farm and Community-Scale Salt Disposal Basins on the Riverine Plain, Guidelines for Basin Use*, CSIRO Land and water Technical Report 12/00, p. 121.

Kamra S.K. and Rao K.V.G.K., (1994) *Modelling Long-term Impacts of Sub-surface Drainage in India.–GRID – IPTRID Network Magazine Issue 4*, HR Wallingford, Wallingford, pp. 8–9.

Kefford B.J. (2000) The effect of saline water disposal: Implication for monitoring programs and management, *Environmental Monitoring and Assessment* **63**(2), 313–327.

Klein N. (1981) Dissolved material transport – the flushing effect in surface and subsurface flow. *Earth Surface Processes and Landforms*, **6**, 173–178.

Kruse E.G., Willardson L. and Ayars J. (1990) On-farm irrigation and drainage practices. In *Agricultural Salinity Assessment and Management Manual*, Tanji K.K. (Ed.), ASCE: New York, pp. 349–371.

Leaney F., Christen E., Jolly I. and Walker G. (2000) *On-Farm and Community-Scale Salt Disposal Basins on the Riverine*

Plain, Guidelines Summary. CSIRO Land and Water Technical Report 24/00, p. 33.

Maas E.V. (1990) Crop salt tolerance. In *Agricultural Salinity Assessment and Management*, ACSE Manuals and reports on engineering practice No. 71, Tanji K.K. (Ed.), ASCE: New York, pp. 262–304.

Maas E.V. and Hoffmann G.J. (1977) Crop salt tolerance – current assessment. *ASCE Journal of Irrigation and Drainage Division*, **103**(IR2), 115–134.

Marie A. and Vengosh A. (2001) Sources of Salinity in Groundwater from Jericho Area, Jordan Valley. *Groundwater*, **39**(2), 240–248.

Mauger G.W. (1996) *Modelling Dryland Salinity with MAGIC System*, Report No. WRT7, *Water and Rivers Commission*, p. 17.

Mazor E. and George R.J. (1992) Marine airborne salts applied to trace evapotranspiration, local recharge and lateral groundwater flow in Western Australia *Journal of Hydrology*, **139**, 63–77.

Meyer W.S., Godwin D.C. and White R.J.G. (1996) SWAGMAN Destiny. A tool to project productivity change due to salinity, waterlogging and irrigation management. *Proceedings of 8th Australian Agronomy Conference*, *Toowoomba*, 425–428.

Michael G., McDonald G., Harbaugh A., (1988), *A Modular Three-Dimensional Finite-Difference Ground-water Flow Model: Techniques of Water-Resources Investigations of the United States Geological Survey*, Book 6, Modeling Techniques.

Moore G. (1998) *Soilguide. A handbook for Understanding and Managing Agricultural Soils*. Agriculture Western Australia Bulletin No. 4343.

Nulsen R.A. (1981) Critical depth to saline groundwater in non-irrigated situations. *Australian Journal of Soil Resources*, **19**, 83–86.

Oster J.D. and Rhoades J.D. (1990) Steady state root zone salt balance. In *Agricultural Salinity Assessment and Management Manual*, Tanji K.K. (Ed.), ASCE: New York, pp. 469–481.

Pannell D.J. (2000) *Market-Based Mechanisms, Financial Incentives and Other Institutional Innovations: Assessing their Potential for Addressing Dryland Salinity*, SEA Working Paper 2000/9, Agricultural and Resource Economics, University of Western Australia.

Peck A.J. (1973) Analysis of multi-dimensional leaching. *Soil Science Society of America Proceedings*, **37**, 320.

Peck A.J. (1983) Response of groundwaters to clearing in Western Australia. *Papers of the International Conference on Groundwater and Man*, Vol. 1, Australian Water Resource Council Conference Series, Sydney, 327–336.

Peck A.J. and Hurle D.H. (1973) Chloride balance of some farmed and forested catchments in south-western Australia. *Water Resources Research*, **9**, 648–657.

Peck A.J. and Williamson D.R. (1987) Hydrology and salinity in the Collie River Basin, Western Australia, (Editors). *Journal of Hydrology*, **94**, 1–181.

Pessarakli M. (1993) *Handbook of Plant and Crop Stress*, Marcel Dekker Ltd: New York.

Quirk J.P. and Schofield R.K. (1955) The effect of electrolyte concentration on soil permeability. *Journal of Soil Science*, **6**, 163–178.

Raats P.A.C. (1981) Residence times of water and solutes within and below the root zone. *Agricultural Water Management*, **4**, 63–82.

Rao K.V.G.K., Oosterbaan R.J., Boonstra J., (1992) *Regional Agro-Hydro-Salinity Model: Description of Principles – Joint Unpublished Research Report*, Central Soil Salinity Research Institute, Karnal, India and International Institute for Land Reclamation and Improvement (ILRI): Wageningen.

Refsgaard J.C. and Storm B. (1995) Computer Models of Watershed Hydrology. In *Water Resources Publication*, Singh V.P. (Ed.), pp. 809–846.

Rhoades J.D. (1984) Reusing saline drainage waters for irrigation: A strategy to reduce salt loading of rivers. In *Salinity in Watercourses and Reservoirs*, French R.H. (Ed.), *Proceedings of International Symposium State-of-the-Art Control of Salinity*, Salt Lake City, pp. 455–464, 12–15 July 1983.

Rhoades J.D., Lesch S.M., Lemert R.D. and Alves W.J. (1997) Assessing irrigation / drainage / salinity management using spatially referenced salinity measurements. *Agricultural Water Management*, **35**, 147–165.

Richter B.C., Kreitler C.W. and Bledsoe B.E. (1993) *Geochemical Techniques for Identifying Sources of Ground-Water Salinization*, CRC Press, Inc.: Boca Raton, p. 258.

Ruprecht J.K. and Schofield N.J. (1989) Analysis of stream flow generation following deforestation in southwest Western Australia. *Journal of Hydrology*, **105**, 1–17.

Ruprecht J.K. and Schofield N.J. (1991a) Effects of partial deforestation on hydrology and salinity in high salt storage landscapes. I. Extensive block clearing. *Journal of Hydrology*, **129**, 19–38.

Ruprecht J.K. and Schofield N.J. (1991b) Effects of partial deforestation on hydrology and salinity in high salt storage landscapes. II. Strip, soils and parkland clearing. *Journal of Hydrology*, **129**, 39–55.

Salama R.B., Laslett D. and Farrington P. (1993) Predictive Modelling of Management Options for Control of Dryland Salinity in a first Order Catchment in The Wheatbelt of Western Australia. *Journal of Hydrology*, **145**, 19–40.

Sami K. (1992) Recharge mechanisms and geochemical processes in semi-arid sedimentary basin, Eastern Cape, South Africa. *Journal of Hydrology*, **139**, 27–48.

Schofield N.J. and Ruprecht J.K. (1989) Regional analysis of stream salinization in south-west Western Australia. *Journal of Hydrology*, **112**, 19–39.

Schofield N.J., Ruprecht J.K. and Loh I.C. (1988) *The Impact of Agricultural Development on the Salinity of Surface Water Resources of South-West Western Australia*, For the Steering Committee for Research on Land Use and Water Supply, Report No. WS27, Water Authority of Western Australia, p. 83.

Silberstein R., Vertessy R., Morris J. and Feikema P. (1999) Modelling the effects of soil moisture and solute conditions on long-term tree growth and water use: A case study from the Shepparton irrigation area, Australia. *Agricultural Water Management*, **39**, 283–315.

Simunek J. and Suarez D.L. (1994) Two dimensional transport model for variably saturated porous media with major ion chemistry. *Water Resources Research*, **30**, 1115–1133.

Simunek J. and van Genuchten M.Th. (1999) *Using the HYDRUS-1D and HYDRUS-2D Codes for Estimating Unsaturated Soil Hydraulic and solute Transport Parameters*, in van Genuchten M.Th., Leij F. J. and Wu L. (Eds.) Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media, University of California, Riverside, CA, pp. 1523–1536.

Simunek J., Vogel T. and van Genuchten M.Th. (1994) *The SWMS_2D Code for Simulating Water Flow and Solute Transport in Two-Dimensional Variably Saturated Media*, Version 1.21. Research Report No. 132, U.S. Salinity Laboratory, USDA, ARS: Riverside, p. 197.

Sivapalan M., Viney N.R. and Ruprecht J.K. (1996a) Water and salt balance modelling to predict the effects of land use changes in forested catchments, 2, Coupled model of water and salt balances. *Hydrological Processes*, **10**, 413–428.

Sivapalan M., Viney N.R. and Jeevaraj C.G. (1996b) Water and salt balance modelling to predict the effects of land use changes in forested catchments, 3, The large catchment model. *Hydrological Processes*, **10**, 429–446.

Skaggs R.W. (1980) *DRAINMOD Reference Report – Methods for Design and Evaluation of Drainage-Water Management Systems for Soils With High Water Tables*, USDA SCS, South National Technical Center: Texas.

Szabolcs I. (1989) *Salt – Affected Soils*, Florida CRC Press: Boca Raton, p. 274.

Tanji K.K. (1990) *Agricultural Salinity Assessment and Management*, American society of Civil Engineers: New York.

Todd D.K. (1970) *The Water Encyclopedia*, Water Information Centre: Port Washington, New York.

USSL (1954) *Diagnosis and Improvement of Saline and Alkali soils*. U.S.D.A. Handbook No. 60. p. 160.

Vertessy R.A. Dawes W.R., Zhang L., Hatton T.J. and Walker J. (1996) *Catchment Scale Hydrologic Modelling to Assess the Water and Salt Balance Behaviour of Eucalypt Plantations*, CSIRO Division of Water Resources: Technical Memorandum 96/1, p. 23.

Wagenet R.J. and Hutson J.L. (1992) LEACHM: Leaching Estimation and Chemistry Model. *A Process-based Model of Solute Movement, Transformations, Plant Uptake, and Chemical Reactions in the Unsaturated Zone*, Version 3, Research. Series No. 92-3, Department of Soil, Crop, and Atmospheric Sciences, Cornell University: Ithaca.

Walling D.E. and Webb B.W. (1983) The dissolved loads of rivers: A global overview. In: *Dissolved Loads of Rivers and Surface Water Quantity and Quality Relationships*, Publication No. 141, International Association of Hydrological Sciences, pp. 3–20.

Walling D.E. and Webb B.W. (1986) Solutes in rivers systems. In *Solute Processes*, Trudgill S.T. (Ed.), John Wiley and Sons: Great Britain, pp. 252–327.

Walling P.E. and Foster I.D.L. (1975) Variations in the natural chemical concentration of river water during flood flows and the long effect: Some further comments. *Journal of Hydrology*, **26**, 237–244.

Williams W.D. (1999) Salinization: A major threat to water resources in the arid and semi-arid regions of the world. *Lakes & Reservoirs: Research and Management*, **4**, 85–91.

Williams W.D. (2001a) Anthropogenic Salinization of Inland Waters. *Hydrobiologia*, **466**(1–3), 329–337.

Williams W.D. (2001b) Salinization: Unplumbed salt in a parched landscape. *Water Science and Technology*, **43**(4), 85–91.

100: Water Quality Modeling

STEVEN C MCCUTCHEON

Faculty of Engineering and Warnell School of Forest Resources, University of Georgia on assignment from the US EPA National Exposure Research Laboratory, Athens, GA, US

Water quality forecasts are vital to determining limitations for point source waste loads and diffuse or nonpoint source loads required to meet water quality standards in streams, lakes, and estuaries. Only mechanistic water quality modeling is widely applicable for forecasting. Although surface water quality modeling has a sound interdisciplinary scientific basis, the use of these analysis tools are limited in regulatory and resource decision making by a lack of a professional consensus defining model application protocols. To analyze water quality problems, (i) the uses for a water body must be evaluated, (ii) models must be calibrated and tested to relate impaired water quality to the waste sources, (iii) a formal uncertainty analysis must define a margin of safety to protect human and ecological health, and (iv) when necessary, an economic analysis must be performed to determine if the designated uses are reasonably achievable.

SCIENCE AND ART

The engineering art of surface water quality modeling has a sound interdisciplinary scientific basis, but the use of these analysis tools for forecasts are limited in decision making by a lack of a professional consensus defining what application protocols should be followed. Model forecasts are important elements of determining point source waste loads and nonpoint source loads required to meet water quality standards in streams, lakes, and estuaries. *See* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1** for a broader discussion of these modeling arts for all of hydrology.

At least three fundamental principles are used to formulate mechanistic water quality models for forecasting. These include Newton's Second Law from which the conservation of momentum equation is derived and three other conservation principles – water continuity, constituent mass conservation, and conservation of energy in temperature simulations and occasionally, turbulence closure. Fundamental to all mechanistic water quality models is the conservation of water and the conservation of constituent mass, whether dissolved or particulate. The most primitive models do a simple accounting for water and constituent mass present in the water body. Less primitive models use the continuity equation for water, and the hydraulic slope and water body morphometry to derive flows or velocities at a point and to account for changes in water volume due to flows entering and leaving the water body. Most state-of-the-art models also solve the conservation of water momentum and continuity equations to actually simulate flows from a specification of inflows, outflows, boundary roughness, and wind drag. The most advanced hydraulics and hydrodynamics models use conservation of energy and other semiempirical turbulence closure techniques to simulate the dynamic distribution of water volumes and flows throughout the water body. The allied water quality algorithms or separate models are based on the conservation of mass equations and Fick's law to derive a separate advective-diffusion equation for each dissolved or particulate constituent. The advective-diffusion equation uses simulations or specifications of water velocity and semi-empirical or empirical expressions of each important kinetic process that creates or destroys constituent mass. Lin and Falconer (**Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**) write each one of these governing equations. On the horizon of applied science and expected to become a part of practice in the coming decade is the use of the Second Law of Thermodynamics to define unit stream power and simulate changes in channel geomorphology. These stream and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

watershed geomorphology models are expected to add a fifth fundamental principle to practical water quality modeling.

Although the empirical and semiempirical relationships are well known in water quality modeling, the practical procedures for defining the overall limitations of simulations at a particular site are not defined as well as necessary to support sound decision making. A high degree of experience in the art of water quality modeling is necessary to collect, compile, and derive water quality, flow, morphometric and meteorological data and parameters. The art of collecting, compiling, and selecting this information is driven by the specific resource management decisions to be made for the water body. Because resource management and environmental decision making are dynamic arts as well, the highest degree of practice requires iterative, interactive communication with resource managers, as project and program needs and objectives mature and change. *See* (**Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**) for a broader view of the experience necessary in hydrologic modeling.

The empirical and semiempirical parameters and coefficients used in water quality modeling must by definition be selected by calibration. Although, some data collection protocols and laboratory analysis of samples can be used to specifically determine some parameters, almost all of these numerical values still need to be confirmed during model calibration. Given the large number of empirical parameters involved, all calibrated water quality models require testing to define, at least in a limited sense, the range of applicability for a particular water body. Terms such as "verification" and "validation" have been used to describe the testing procedure. Occasionally, only calibration may be required for some modeling projects. When less precision and accuracy is needed, hydrodynamic and hydraulic models of water movement and conservative tracers may only require calibration checks of roughness coefficients and dispersion parameters, especially in large water bodies in which boundary friction and small-scale mixing is not vitally important.

Despite the general agreement on the need for calibration and testing (Thomann and Mueller, 1987), few well-defined calibration, testing, and hind casting or forecasting protocols have been agreed upon. This fact alone requires a high degree of art and experience in modeling. Thus, the most important element in water quality analysis is not the computer codes applied, but the expertise of the modeler to interpret the calculations and translate the numerical results into information which decision makers and mangers can use. Perhaps the most significant lag in the development of practice is the failure to agree upon and use statistical tests to define the precision and accuracy or uncertainty of all model applications. Guessing and then labeling water quality model calibrations as "good," "good enough," or in similar qualitative terms is not sufficiently informative for decision makers. Formally defining scientific uncertainty for decision makers is not an easy task, but is one that is vital to overcoming past misuses of water quality models in resource management. Plus, there are unknowable events in water quality analysis as in any analysis that is beyond the current imagination of humankind. However, this reality must not prevent practitioners from consistently and reliably describing the known limitations of models and simulation uncertainty for each application. Furthermore, because very few water quality models used in practice have been formally peer reviewed, site-specific definition of simulation uncertainty and definition of model limitations will remain necessary and vital.

Probabilistic modeling can also be linked to a formal uncertainty analysis. Like the statistical testing of calibrations and other simulations, the practical use of probabilistic approaches (Reckhow and Chapra, 1983) is significantly overdue in water quality forecasting (NRC, 2001).

The following two sections support this introduction to the science and art of water quality modeling. The next section reviews some of the practical models in use today to illustrate that uncertainty analyses are not adequately applied in practice. The final section briefly clarifies an application protocol for these models when used to determine a waste load allocation for point sources or nonpoint sources, or both, to illustrate procedurally when and how statistical testing and uncertainty analysis should be applied in water quality modeling.

PRACTICAL COMPUTER CODES

Table 1 lists representative practical models used in water quality analysis. These represent a consistent set of models for simulating the steady state and dynamic water quality of streams and rivers, lakes and reservoirs, estuaries, and watershed loads. This selection was limited to those US codes in the public domain or available to most government agencies, for which misapplication may be most probable. As practice in Europe and elsewhere avoids some misuse by mostly relying upon custom model formulations for each site investigated (*see* **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**), Table 1 was limited to public domain codes to illustrate the weakness in US water quality modeling practice. Shanahan *et al.*, (1998) notes the lack of standard model development protocols that currently limit European practice in water quality modeling (**`http://harmoniqua.wau.nl`**).

Most striking from Table 1 is that only one practical model includes a formal first order uncertainty analysis,

Table 1 Some practical water quality models

Water body and dynamics	Model	Creation date	Uncertainty analysis	Comments
Steady state one-dimensional streams	QUAL2EU	1973	1st order error analysis	Developed by the US Environmental Protection Agency and is in the public domain at (**http://www.epa.gov/ceampubl/swater/qual2eu/**). The original code QUAL-II (Roesner *et al.*, 1973) has undergone (i) line by line peer review of the code and manual (NCASI, 1980) and (ii) peer review of the manual and testing using diverse data from three rivers (McCutcheon, 1983).
Dynamic one-dimensional streams and rivers (RIV1)	CE-QUAL-RIV1	1985	None	Developed by the US Army Corps of Engineers and available upon request (**http://el.erdc.usace.army.mil/products.cfm?Topic=model&Type=watqual)**. No known peer review of code and manual but the governing equations, solution schemes, and some simulation results has been published in the peer-reviewed literature.
Dynamic one-dimensional, two-dimensional, and three-dimensional lakes and estuaries	WASP7	1981	None	Developed by the US Environmental Protection Agency (Di Toro *et al.*, 1981; Connolly and Winfield, 1984; Ambrose *et al.*, 1988) and in the public domain (**http://www.epa.gov/ceampubl/swater/wasp**) and (**http://www.epa.gov/athens/wwqtsc/html/wasp.html**). Earlier version peer reviewed by NCASI (1987).
Dynamic one-dimensional lakes and reservoirs (R1)	CE-QUAL-R1	1985	None	Developed by the US Army Corps of Engineers from the Water Quality for River-Reservoir Systems computer code derived from the original slab model by Orlob (HEC, 1978) and available upon request but no longer maintained by the Army (**http://el.erdc.usace.army.mil/products.cfm?Topic=model&Type=watqual**). No known peer review of code and manual.
Dynamic two-dimensional water bodies (lakes and estuaries, W2)	CE-QUAL-W2	1975	None	Developed by the US Army Corps of Engineers from the Laterally Averaged Reservoir Model (Buchak and Edinger, 1981) and available upon request (**http://el.erdc.usace.army.mil/products.cfm?Topic=model&Type=watqual**). No known peer review of code and manual but (**http://www.ce.pdx.edu/w2/index.html?workshop.html**) indicates that over 516 applications of the code are known, many of which are independent of the Corps of Engineers.
Dynamic three-dimensional estuaries	CE-QUAL-ICM (Integrated Compartment Model)	1982 and 1989	None	Developed by the US Army Corps of Engineers and some versions are in the public domain (**http://el.erdc.usace.army.mil/products.cfm?Topic=model&Type=watqual**). CE-QUAL-ICM derived from the same box-model approach as WASP using a different numerical solution. No known peer review of code and manual but the Chesapeake Bay simulations have been reviewed by at least two independent panels and also published in the peer-reviewed literature.
Dynamic three-dimensional estuaries	EFDC (Environmental Fluid Dynamics Code)	1992	None	Developed by the US Environmental Protection Agency and in the public domain (**http://www.epa.gov/ATHENS/wwqtsc/html/efdc.html**). No known peer review of code and manual but Beach *et al.* (2002) documents a peer review of the conceptual basis of EFDC to simulate polychlorinated biphenyl fate and remediation in the Housatonic River.

(*continued overleaf*)

Table 1 (*continued*)

Water body and dynamics	Model	Creation date	Uncertainty analysis	Comments
Dynamic watersheds	HSPF11 (Hydrologic Simulation Package-Fortran)	1966	Has been used with Monte Carlo analysis (Golder Associates, 2003)	Developed by the US Environmental Protection Agency and US Geological Survey (Donigian *et al.*, 1984; Bicknell *et al.*, 1993) and is in the public domain at (`http://www.epa.gov/ceampubl/swater/hspf/`). No known peer review of code and manual. Independent experts tested the model with both field data and model experiments (Sullivan and Schueler, 1982; Schueler, 1983; Song *et al.*, 1983; Hicks *et al.*, 1985; Nichols and Timpe, 1985; Weatherbe and Novak, 1985; Udhiri *et al.*, 1985; Motta and Cheng, 1987; Moore *et al.*, 1988; Chew *et al.*, 1991; Chen *et al.*, 1998a,b).

Notes: Much of the information is from Martin and McCutcheon (1999). This table focuses on practical US water quality models. Rosbjerg and Madsen (**Chapter 10, Concepts of Hydrologic Modeling, Volume 1**) discuss select European models that are equally practical.

and only two codes have been formally peer reviewed. Yet, despite being the first and only model that includes the necessary uncertainty analysis, the QUAL2EU code has not been adapted to a Windows™ environment, and the US Environmental Protection Agency has developed a newer version that does not include uncertainty analysis (`http://www.epa.gov/athens/wwqtsc/html/qual2k.html`). While all other practical models should be used in an uncertainty analysis framework, only the HSPF seems to have actually been used in this manner (Table 1) in North America. A Monte Carlo analysis was used to generate probabilistic distributions of simulated and forecast water quality conditions (Golder Associates, 2003). This lack of formal uncertainty analysis tools incorporated into all practical software used for water quality modeling makes it vital in the following section to clarify that sound protocols must include uncertainty analysis and development of objective margins of safety in modeling practice. Rosbjerg and Madsen (**Chapter 10, Concepts of Hydrologic Modeling, Volume 1**) review the methods of uncertainty analysis for modeling.

The very limited number of peer reviews of water quality computer codes and manuals is disturbing and inefficient. Furthermore, the two codes, QUAL-II and Water Quality Analysis Simulation Program (WASP), which were peer reviewed, were examined in 1980–1983 and 1987. Since that time, both codes have been updated several times and these updates have not been peer reviewed and the revised codes independently retested. Thus, the scientific credibility of water quality modeling derives almost exclusively from independent case studies by experts in the field. These case studies, more extensive for some models than others, form only an irregular patchwork of *ad hoc* credibility for model use over the full simulation domains. Thus, almost all water quality modeling must only be undertaken by experts who understand the extent of the *ad hoc* credibility and are scientifically qualified to extend the patchwork credibility on a case-by-case basis. This requirement leaves few applications that can be undertaken by neophytes who have not mastered the history and practice. Much standardization and rigorous review are necessary to accommodate widespread use of water quality models in significant resource management and environmental protection decisions. The following protocol for water quality modeling stresses how these shortcomings must be overcome until the scientific bases can be updated, some standardization put in place, and resulting computer programs and manuals independently peer reviewed and tested over the full domain of intended use.

CLARIFICATION OF APPLICATION PROTOCOLS

The art of water quality modeling predominantly comes into play in applications of models to specific sites to address specific management issues. There are two dominant reasons why the relatively complete science of water quality modeling requires considerable art and experience. First, there is a lack of consensus on procedures to calibrate, test, and apply water quality models for a given quantity and quality of data. Practitioners typically ignore some procedures such as model testing and rarely describe the limitations of calibrated and applied models in terms that resource managers understand. Second, statistical criteria

to judge model calibrations and tests are diverse and are rarely used in practice. Sensitivity and uncertainty analysis are rarely used. When applied, analyses are not applied holistically with statistical testing to define well understood margins of safety that are now required for total maximum daily load analysis to be considered scientifically valid in the US (NRC, 2001) and should be required for all water quality design work. Instead, limited knowledge and investigation of conservative assumptions in model conceptual frameworks are typically used to assume that the required margin of safety is qualitatively covered. This implicit assumption is rarely stated, much less discussed in terms that resource managers seem to comprehend for use in decision making.

To overcome these severe limitations, the following general protocol should be followed for water quality analyses, such as a waste load allocation or total maximum daily load calculation for nonpoint source pollution:

1. Evaluate the designated uses for a water body or segment. The water quality criteria that define these uses typically define the water quality issues to be investigated. Designated uses may include water supply, provision of habitat, irrigation water use, recreation, industrial use, and other uses. The water quality criteria to protect the designated uses have typically been translated into standards by the responsible State, Tribe, or Environmental Protection Agency Region in the US. If the water segment has been classified as impaired and listed on the US 303(b) list [40 CFR (Code of US Federal Regulations) Part 130], then the evaluation should look at the reasons for listing, including the quantity and quality of data used for this purpose. Delisting should be an option when current assessment data no longer indicate impairment (Thomann and Mueller, 1987; Shabman and Smith, 2003).
2. Relate potential or existing water quality impairments of a water segment to the point and nonpoint source loads (Chapra, 2003), including legacy sources due to past waste and land management practices that might include contaminated sediments. Development of rational cause-and-effect models for forecasting requires data collection, model calibration, and testing to define model limitations and uncertainty using well-defined sensitivity and statistical testing (*see* **Chapter 10, Concepts of Hydrologic Modeling, Volume 1**; **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**; **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**). The data used to test a model must be independent of the data used to calibrate the program. Sensitivity analysis is used to define an operational model for forecasting within the defined limits, or other analyses so as to hindcast causes of problems. The limits of an operational or calibrated model are defined by the conditions used in the calibration and testing of the model. However, no protocols seem to have been agreed upon to define the limits of a calibrated and tested model. Statistical evaluation using consistent statistics (McCutcheon *et al.*, 1990) and preselected levels of allowable error in decision making is required to determine when a model is adequately calibrated and to test a model with independent data sets. Alternatively, the level of significance for the calibration and independent testing can be determined statistically and used to convey the level of uncertainty to decision makers. If they deem the uncertainty too high, then adaptive management can be employed to make some load reductions, collect more calibration and testing data, and repeat the calibration, testing, and formulation of an operation model.
3. Application of the calibrated and tested model involves iterative simulations of scenarios that represent decreases in point and nonpoint sources to ensure that water quality standards and criteria are met. A formal uncertainty analysis must be used to define the vital margin of safety or factor of safety (Thomann and Mueller, 1987; NRC, 2001; Reckhow, 2003; Walker 2003), along with knowledge of the resource management and environmental protection objectives. This objective margin of safety is generally defined in terms of the following:
   (a) Reserve waste assimilation of the water body
   (b) Potential for acute and chronic human health effects associated with certain designated uses (the greater the risks, the larger the margin of safety should be)
   (c) Life history of all important aquatic species effects for certain designated uses
   (d) Habitat effects for certain designated uses
   (e) Known uncertainties in water quality simulations
   (f) Risk of upsets in point source treatment plants leading to occasional discharge of untreated waste waters
   (g) Reliability of best management practices that may be selected for nonpoint source control.
4. Where required, an economic analysis should be performed to determine the impact on waste load allocations and total maximum daily load allocations. Basic data should be provided for water quality managers to weight costs of achieving or maintaining water quality standards compared to the benefits of the designated uses, consistent with the antidegradation or nondegradation policies associated with the US Clean Water Act [see 40 CFR §131.12 and (**`http://www.epa.gov/OCEPAterms/nterms.html`**)].

Acknowledgments

The reviews of Tim Wool of the US Environmental Protection Agency Watershed and Water Quality Modeling Technical Support Center, the editor, and anonymous reviewers are appreciated. Mark S. Dortch, Tom Cole, and Dotty Tillman of the Water Quality and Contaminant Modeling Branch of the US Army Waterways Experiment Station provided information on the availability and peer review of the Army models in Table 1. Dr. Earl Hayter of the US Environmental Protection Agency National Exposure Research Laboratory provided information on the EFDC. The electronic software templates for this article were prepared from information on the web sites of the Water Quality and Contaminant Modeling Branch of the US Army Waterways Experiment Station and Portland State University, and from the sites of the US Environmental Protection Agency Center for Exposure Research Modeling and the Watershed and Water Quality Modeling Technical Support Center. This paper has been reviewed in accordance with the US Environmental Protection Agency peer and administrative review policies and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFERENCES

Ambrose R.B., Wool T.A., Connolly J.P. and Schanz R.W. (1988) *WASP4, A Hydrodynamic and Water Quality Model – Model Theory, User's Manual, and Programmer's Guide*, U.S. Environmental Protection Agency: Athens, EPA/600/3-87-039.

Beach R.B., Clough J., Craig P.M., DiNitto R., Donigian A.S., Fischenich C., Lawrence G., McGrath R.A., Stoddard A., Svirsky S.C., and Wallen C.M. (2002) *Responsiveness Summary to Peer Review of the Modeling Framework Design and Quality Assurance Project Plan: Modeling Study of PCB Contamination in the Housatonic River DCN: GE-053102-ABAQ*, Roy F. Weston, Inc., US Army Corps of Engineers, Contract No. DACW33-00-D-0006, Task Order 0003, Concord.

Bicknell B.R., Imhoff J.C., Kittle J.L., Donigian A.S. and Johanson R.C. (1993) *Hydrological Simulation Program - FORTRAN (HSPF): Users Manual for Release 10*, U.S. Environmental Protection Agency: Athens, 30605. EPA-600/R-93/174.

Buchak E.M. and Edinger J.E. (1981) *User Guide and Development Document for LAEM2, a Longitudinal-Vertical, Time-Varying Hydrodynamic Estuary Model*, J.E. Edinger Associates: Wayne, Contract No. DACW39-81-M-2788, U.S. Army Corps of Engineers Waterways Experiment Station: Vicksburg.

Chapra S.C. (2003) Engineering water quality models and TMDLs. *Journal of Water Resources Planning and Management*, **29**(4), 247–256.

Chen Y.D., Carsel R.F., McCutcheon S.C. and Nutter W.L. (1998a) Stream temperature simulation of forested riparian areas: I. Watershed-scale model development. *Journal of Environmental Engineering*, **124**(4), 304–315.

Chen Y.D., McCutcheon S.C., Norton D.J. and Nutter W.L. (1998b) Stream temperature simulation of forested riparian areas: II. Model application. *Journal of Environmental Engineering*, **124**(4), 316–328.

Chew Y.C., Moore L.W. and Smith R.H. (1991) Hydrologic simulation of Tennessee's North Reelfoot Creek watershed. *Journal of Water Pollution Control Federation*, **63**, 10–16.

Connolly J.P. and Winfield R. (1984) *A User's Guide for WASTOX, a Framework for Modeling the Fate of Toxic Chemicals in Aquatic Environments. Part 1: Exposure Concentration*, U.S. Environmental Protection Agency: Gulf Breeze, EPA-600/3-84-077.

Di Toro D.M., Fitzpatrick J.J. and Thomann R.V. (1981) rev. 1983 *Water Quality Analysis Simulation Program (WASP) and Model Verification Program (MVP)-Documentation*, Hydroscience: Westwood, for U.S. Environmental Protection Agency: Duluth, Contract No. 68-01-3872.

Donigian A.S., Imhoff J.C., Bicknell B.R. and Kittle J.L. (1984) *Application Guide for the Hydrologic Simulation Program - FORTRAN*, U.S. Environmental Protection Agency: Athens, EPA 600/3-84-066.

Golder Associates. (2003) *Calibration of the HSPF Water Quality Model for the Oil Sand Region in Northeastern Alberta*, Consulting Report IP2-2501 (1000), Calgary.

Hicks C.N., Huber W.C., and Heaney J.P. (1985) Simulation of possible effects of deep pumping on surface hydrology using HSPF. In *Proceedings of Stormwater and Water Quality Model User Group Meeting*, Barnwell T.O. Jr (Ed.), U.S. Environmental Protection Agency: Athens, EPA-600/9-85/016.

Hydrologic Engineering Center (HEC) (1978) *Water Quality for River-Reservoir Systems*, U.S. Army Corps of Engineers: Davis, Preliminary Report.

Martin J.L. and McCutcheon S.C. (1999) *Hydrodynamics and Transport for Water Quality Modeling*, Lewis – CRC Press: Boca Raton.

McCutcheon S.C. (1983) *The Evaluation of Selected One-Dimensional Stream Water-Quality Models with Field Data*, U.S. Waterways Experiment Station Report E-83-11, Army Corps of Engineers: Vicksburg.

McCutcheon S.C., Zhu D.W. and Bird S.L. (1990) Model calibration, validation and use. In *Chapter II, Technical Guidance Manual for Performing Waste Load Allocations, Book III, Estuaries, Part I: Estuaries and Waste Load Allocation Models*, Section 5 Ambrose R., Martin J.L. and McCutcheon S.C. (Eds.), U.S. Environmental Protection Agency Office of Water: Washington.

Moore L.W., Matheny H., Tyree T., Sabatini D. and Klaine S.J. (1988) Agricultural runoff modeling in a small west Tennessee watershed. *Journal - Water Pollution Control Federation*, **60**, 242–249.

Motta D.J. and Cheng M.S. (1987) The Henson Creek watershed study. In *Proceedings of Stormwater and Water Quality Users Group Meeting*, Torno H.C. (Ed.), Charles Howard and Assoc: Victoria.

National Council for Air and Stream Improvement, Inc. (NCASI). (1980) *A Review of the Mathematical Model QUAL-II and Guidance for its Use*, Stream Improvement Bulletin No. 338, New York.

National Council for Air and Stream Improvement, Inc. (NCASI). (1987) *A Review of the Mathematical Model WASP and Guidance for its Use*, Stream Improvement Bulletin, New York.

National Research Council (NRC) (2001) *Assessing the TMDL Approach to Water Quality Modeling*, National Academy Press: Washington.

Nichols J.C. and Timpe M.P. (1985) Use of HSPF to simulate dynamics of phosphorus in floodplain wetlands over a wide range of hydrologic regimes. In Barnwell T.O. Jr (Ed.), *Proceedings of Stormwater and Water Quality Model Users Group Meeting*, U.S. Environmental Protection Agency: Athens, EPA-600/9-85/016.

Reckhow K.H. (2003) On the Need for Uncertainty Assessment in TMDL Modeling and Implementation. *Journal of Water Resources Planning and Management*, **29**(4), 245–246.

Reckhow K.H. and Chapra S.C. (1983) *Engineering Approaches for Lake Management – Volume 1: Data Analysis and Empirical Modeling*, Ann Arbor Science, Butterworth Publishers: Boston.

Roesner L.A., Monser J.R. and Evenson D.E. (1973) *User's Manual for Stream Water Quality Model (QUAL-II)*, U.S. Environmental Protection Agency: Washington.

Schueler T.R. (1983) *Seneca Creek Watershed Management Study, Final Report, Volumes I and II*, Metropolitan Washington Council of Governments: Washington.

Shabman L. and Smith E. (2003) Implications of applying statistically based procedures for water quality assessment. *Journal of Water Resources Planning and Management*, **29**(4), 330–336.

Shanahan P., Henze M., Koncsos L., Rauch W., Reichert P., Somlyódy L. and Vanrolleghem P. (1998) River water quality modelling: II. Problems of the art. *Water Science and Technology*, **38**(11), 245–252.

Song J.A., Rawl G.F. and Howard W.R. (1983) Lake Manatee watershed water resources evaluation using hydrologic simulation program FORTRAN (HSPF). In Beron P. and Barnwell T. (Eds.), *Colloque sur la Modelisation des Eaux Pluviales*, GREMU 83/03 Ecole Polytechnique de Montreal: Quebec.

Sullivan M.P. and Schueler T.R. (1982) The Piscataway Creek watershed model: A stormwater and nonpoint source management tool. In *Proceedings Stormwater and Water Quality Management Modeling and SWMM Users Group Meeting*, Wisner P.E. (Ed.), Univ. of Ottawa, Dept. Civil Engr.: Ottawa.

Thomann R.V. and Mueller J.A. (1987) *Principles of Surface Water Quality Modeling and Control*, Harper Collins: New York.

Udhiri S., Cheng M.-S. and Powell R.L. (1985) The impact of snow addition on watershed analysis using HSPF. In Barnwell T.O. Jr (Ed.), *Proceedings of Stormwater and Water Quality Model Users Group Meeting*, U.S. Environmental Protection Agency: Athens, EPA-600/9-85/016.

Walker W.W. (2003) Consideration of variability and uncertainty in phosphorus total maximum daily loads for lakes. *Journal of Water Resources Planning and Management*, **29**(4), 337–344.

Weatherbe D.G. and Novak Z. (1985) Development of water management strategy for the Humber River. In James E.M. and James W. (Eds.), *Proceedings Conference on Stormwater and Water Quality Management Modeling*, Computational Hydraulics Group, McMaster University: Hamilton.

PART 9

Ecological and Hydrological Interactions

101: Ecosystem Processes

ALAN P COVICH

Institute of Ecology, University of Georgia, Athens, GA, US

Biological production in inland waters is limited by flows of energy and nutrients across landscapes that undergo complex spatial and temporal dynamics driven by hydrological processes. Primary production (by photosynthesis or chemosynthesis), nutrient uptake and cycling, and decomposition of organic matter are all interconnected by the movement of water and organisms among interconnected ecosystems. Hydrology influences rates of transport, deposition, and recycling of inorganic materials, which, in turn, influence the various species and their associated ecosystem processes. Ratios of nutrients limit productivity and influence foraging behavior of consumer species. The ratio of carbon to nitrogen to phosphorus (C:N:P) available to plants often limits rates of photosynthesis. In many inland aquatic ecosystems, the scarcity of phosphorus limits plant productivity, although nitrogen or light are often limiting factors in lakes and shaded streams. Suspended sediment transported by runoff carries adsorbed nutrients and influences light penetration that, in different combinations, can limit rates of photosynthesis.

Dispersal of individuals to high-quality habitats and assembly of natural communities sustain functional groups of species and ecosystem processes following disturbances (such as severe floods and droughts or accidental spills of toxins). This dispersal is often limited by hydrologic connections among different habitats. Some "fugitive species" are especially well adapted for dispersal, and may include other nonaquatic modes of movement. These highly mobile species rapidly colonize new habitats or disturbed habitats and are widespread, generalist species in terms of their roles in ecosystem functioning. Other species have more specialized ecosystem roles in particular habitats and are less adapted for dispersal. These species have restricted distributions and different adaptations for persistence, in relatively isolated ecosystems with limited habitat connectivity.

If population densities increase beyond carrying capacity in optimal habitats, the movement of species from their high-quality habitats to low-quality habitats creates sink populations where individuals survive but do not reproduce and cannot sustain ecosystem services. Source populations are those that continue to reproduce in high-quality habitats and to produce individuals who disperse to locations with lower or variable habitat quality. The occasional flow of genes among these usually isolated subpopulations (metapopulations) influences the long-term evolution of species and aggregations of partially isolated communities (metacommunities). These biotic processes and their associated ecological services depend on sustainable flows of water throughout habitats in entire drainage networks. However, high degrees of connectivity also allow introduced, nonnative species to spread and possibly displace native species. Over time, the loss of native species which are well adapted for long-term variability in environmental conditions can, in some catchments, lead to instability and unreliable provisioning of ecosystem services by nonnative species.

INTRODUCTION

A close look at any drainage basin reveals significant spatial heterogeneity that can influence both hydrological and ecological processes. From micro- to macroscales, there are different distributions of horizontal and vertical elements of the landscape that are important to the structure and function of aquatic ecosystems. These elements influence how water flows across the terrain's surfaces and subsurfaces as well as the distributions of species that perform essential ecosystem processes. Discharge, infiltration, and residence times of water are all influenced by

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

these surfaces. Their combined flowpaths affect weathering, sediment transport, nutrient cycling, as well as biological diversity and productivity. Thus, geological, hydrological, and biological interactions all determine the distributions and dynamics of plants and animals over time and space. The succession of biological communities has predictable patterns of nutrient retention and storage that can be best understood once the underlying physical template is defined and examined in terms of eco-hydrological connectivity.

This review focuses primarily on biological structure and productivity within and among aquatic habitats that are interconnected within drainage networks. The relevance of aquatic connectivity for the dispersal of species and the assembly of plant and animal communities is widely recognized and well documented. The determination and understanding of these general relationships of ecosystem structure and function are the result of more than a century of field studies and analyses of complex natural systems. However, conceptualization of numerous hierarchical linkages in both horizontal and vertical dimensions among aquatic ecosystems has developed most rapidly in the last 10 years.

The varied degrees of isolation and connectivity among populations of aquatic species are strongly influenced by hydrology and topography. Following disturbances, numerous factors determine how species reassemble into the biological communities that influence ecosystem-level processes. The patchy distributions of many aquatic ecosystems (ponds, floodplain lakes, wetlands) result in some degree of spatial separation for varied periods of time. Frequently, however, isolated components of aquatic and terrestrial ecosystems are interconnected by floods or by the ways water resources are managed. For example, flood control can eliminate connections among rivers and floodplain habitats and alter groundwater connections.

The loss of native species which are well adapted for long-term environmental conditions can lead to instability and unreliable provisioning of ecosystem services (Giller *et al.*, 2004a). Water management decisions that alter hydrologic regimes can increase the loss of native species and replacement by nonnative species. Sometimes, these species can functionally substitute for the native species for periods of time. However, these nonnative species may not be well adapted in the long run as conditions continue to change. The intrinsic variability of natural ecosystems often requires a complex mix of species that are capable of sustaining ecological processes in a wide range of conditions. In some cases, there are species interactions that facilitate ecological process through serial processes that favor mutual interactions among species. The loss of one of these species can result in a cascading chain of losses or declines in rates of ecological processes (Covich *et al.*, 2004a). Thus, the distributions of specific producer and consumer species in biotic communities influence rates and types of biological productivity of inland waters (Table 1). This spatial complexity of drainage basins (Table 2) is essential for sustaining the natural processes and ecological services provided by aquatic ecosystems.

Ecosystem Characteristics

Inland aquatic ecosystems are open, self-organized, potentially resilient systems composed of interacting abiotic and biotic components (Wetzel, 2001). Internal ecosystem dynamics of nutrient cycling and flows of energy are regulated by a complex series of interactions among species within food webs and by physical and chemical controls. The biological productivity and biodiversity of any lake, river, or wetland reflect their locations within a drainage network and their histories of disturbances. The rates and timing of abiotic inputs of energy, water, and nutrients to aquatic ecosystems are determined by regional climate, geology, and topography. Erosion, transport, and deposition of sediments in drainage basins regulate complexity of substrata used by the aquatic biota (Amoros and Bornette, 2002; Malard *et al.*, 2002). Assembly of biotic components is determined by evolutionary adaptations, biogeographical relationships, and modifications of habitats over time. Colonization by species from within the region and by introductions of nonnative species influence species composition and their cumulative effects on productivity. Major species losses can occur from overharvesting of certain species and by diseases spread by introduced species (Malmqvist and Rundle, 2002; Holeck *et al.*, 2004; Giller *et al.*, 2004b; Padilla and Williams, 2004).

Interactions among distinct inland aquatic ecosystems emerge as characteristics associated with catchment-scale relationships often driven by climate, geology, and hydrology across a topographic landscape (Montgomery, 1999; Prosser *et al.*, 2001; Church, 2002; Richards *et al.*, 2002; Benda *et al.*, 2004). The physical template that determines how species move through habitats is dynamic and responsive to natural climatic variability and managed flows. Water diversions, channelization, dam construction, agricultural and urban land development greatly modify flows of water and materials while potentially reducing species diversity (Malmqvist and Rundle, 2002; Pringle, 2003a; Meyerbeck, 2004).

Conceptual Developments

Although early studies (Forbes, 1887; Forel, 1892) usually focused on single ecosystems and provided detailed analysis of a particular lake, river or wetland, this initial perspective clearly recognized that connections between terrestrial and aquatic components were important. For example, Forbes (1887) emphasized that lakes represented a natural "microcosm" with a great deal of internal, self

Table 1 Ecosystem terminology

Biogeochemical cycles: Flows of nutrients such as carbon, nitrogen, and phosphorus through distinct pathways within ecosystems.

Chemoautotrophy: Primary production by some microbes is based on energy-yielding chemical reactions as an energy source (rather than solar energy used by photoautotrophs) when synthesizing organic matter from inorganic materials.

Decomposition: The breakdown of dead organic matter (detritus) by bacteria and fungi that recycles nutrients.

Ecosystem: The set of interacting abiotic and biotic components in a spatially defined location where operational boundaries determine measurements of terrestrial, aquatic, and atmospheric inputs and outputs. A concept often used for examining specific questions regarding large-scale changes in the physical environment or shifts in assembly of species within a biotic community.

Ecosystem processes: Flows of energy, cycling of materials, and production of plants and animals are related to determine how ecosystem functioning reflects structure.

Food webs: Species are connected by flow of energy and cycling of materials from producers to consumers in ''flow maps'' composed on multiple food chains (single pathways through the entire food web).

Functional groups: Those species that perform similar roles in an ecosystem process as ''primary producers'' or ''decomposers''.

Heterotrophy: Production based on consumption of externally synthesized organic matter produced by other organisms (such as by herbivores, carnivores, and detritivores).

Photoautotrophy: The most common form of primary production (by green plants such as algae, macrophytes, and by certain bacteria) is based on solar energy when synthesizing organic matter from inorganic materials (carbon dioxide or bicarbonate ions).

Productivity: Organic synthesis by plants (primary productivity) and animals (secondary productivity) per unit area per unit time (annual productivity is usually expressed in units of Kcal m^{-2} $year^{-1}$) provides a basis for general comparisons across different ecosystems and over time.

Succession: Patterns of change in biotic assemblages within ecosystems from one season to another or over longer time periods (years, decades).

Trophic cascades: Changes in the presence or absence (or shifts in abundance) of predators results in altered production at lower trophic levels: primary and secondary productivity is alternately constrained or unconstrained by feeding activities of consumers at upper trophic levels (top-down control).

Trophic levels: Groups of organisms that share similar roles in food chains and food webs defined by the input and output of a shared-source energy. Energy transfers among adjacent groups provide measures of efficiency that are viewed as, sometimes, limiting the length of food chains (primary producers, primary consumers, secondary consumers, etc). A single species may be represented on multiple trophic levels if sources of energy change as a result of changes in their life history and modes of foraging.

Table 2 Drainage basin structure and consequences

Attributes	Measurements	Effects
Basin size	Large versus small area	Amount of runoff increases with basin size
Basin shape	Elongate versus circular	Runoff is faster in narrow basins
Network Pattern	Reticulate versus linear	Runoff is faster in linear drainage networks
Drainage Density	High versus low	Runoff may lag in high-density networks
Channel	Wide versus narrow width	Runoff begins in narrow tributaries (low-order) channels, flows into wider channels (higher-order) channels. Shading limits light in forested narrow channels, and primary productivity is relatively lower than in wide channels.

regulation of their biotic communities. Yet, he also noted that some lakes in Illinois (located within the Mississippi, Illinois, and Ohio River drainages in the central United States) were often isolated in small drainage basins while others were closely interconnected with rivers: "The fuviatile lakes. . .are appendages of the river systems. . . being situated in the river bottoms and connected with the adjacent streams by periodic overflows. Their fauna is, therefore, substantially that of rivers themselves, and the two should, of course, be studied together". Regional studies of lake districts made other early contributions to a landscape perspective. For example, Juday and Birge (1933) compared numerous lakes in northern Wisconsin that shared a glacial origin but differed in morphometry (depth, volume, shape, surface area), species composition and productivity. Thus, regional studies recognizing spatial and temporal dynamics at multiple scales have a long history in aquatic ecology.

Field studies of phosphorus cycling in inland waters by Hutchinson and his colleagues (Hutchinson and Bowen, 1947, 1950) in Connecticut, northeastern USA, used

isotopic tracers to quantify rates of primary productivity and cycling of the phosphorus available for plant uptake (dissolved orthophosphate) in the context of lake biogeochemistry. Related field studies of inland waters contributed additional fundamental concepts regarding biogeochemical relationships between ecosystem structure and function relative to energy flow. Lindeman (1942) provided an innovative framework for comparing how energy flows through food webs based on linked trophic levels; this work continues to stimulate studies that focus on how components within and among ecosystems are connected by energy flows. Odum (1957) quantified energy flows, mineral cycling, and productivity with methods that continue to integrate whole ecosystem dynamics by measuring dissolved oxygen production (photosynthesis) and oxygen uptake (respiration). Many studies began to use isotopes of carbon to measure rates of primary production (photosynthesis) and cycling of carbon.

Light availability and ratios of available nutrients can limit growth of primary producers, composed of suspended algae (phytoplankton) and rooted plants (littoral vegetation). Secondary producers, the herbivores and carnivores, compose the open-water (zooplankton), bottom-dwelling (zoobenthos), and fish communities, that are, in turn, limited by the availability of primary production. Redfield (1958) and his colleagues demonstrated relatively consistent ratios of carbon, nitrogen, and phosphorus that characterized open-water marine organisms (biomass). These ratios are based on the unifying concept of stoichiometry in which all biomass is composed of biochemically well-defined proteins, lipids, carbohydrates and nucleic acids that require basic elemental building materials. Such relationships provide reliable means for predicting where and when consumers might likely forage in different locations to obtain essential nutrition (Sterner and Elser, 2002). In general, these fundamental ecological and biogeochemical processes are often limited by nutrient transport, suspended sediments, and dispersal of organisms that are all interconnected by hydrological processes. For example, suspended sediments transport adsorbed nutrients and limit light penetration; concentrations of dissolved nutrients are altered by the quantity and source areas of runoff.

The importance of viewing lakes and rivers as integrated ecosystems composed of interacting groups of producers and consumers is now widely recognized as being based on coupled energy flows and cycling of nutrients. These ideas of a common energy currency coupled to nutrient cycling emphasize that lakes, river drainages, and wetlands are clearly linked to terrestrial and aquatic ecosystems with flows of energy and nutrients moving in different directions at different times. Odum (1969) determined the ways in which ecosystems change over time. He emphasized that early phases of biological community development are characterized by high rates of primary productivity with relatively few species and linear food webs. These early ecosystems generally have few mechanisms for retaining nutrients, and biomass storage is low. As the community increases in species diversity, the food web becomes further reticulated and rates of primary productivity decline. The more mature ecosystems have additional mechanisms for recycling and retaining nutrients as biomass storage increases.

On the basis of detailed local studies, conceptual advances have examined patterns of productivity by integrating analyses of nutrient budgets, trophic levels, and mass balances of energy (following the first law of thermodynamics). These energy and nutrient budgets also required detailed hydrologic budgets for comprehensive interpretations. Ecologists have had considerable success defining and integrating these budgets in relatively small catchments where sufficient data were available to provide reliable inputs. For example, Mirror Lake, in the Hubbard Brook Valley of the White Mountains, New Hampshire, USA, is one of the best-studied lake basins. Long-term hydrological and limnological studies provide a solid basis for placing biotic relationships in a well-documented ecosystem context that makes this glacial lake basin a testing area for ecological concepts (Likens, 1985, 2004). Information on tributary and atmospheric inputs of nutrients as well as surface and groundwater inputs are integrated over decadal timescales and related to biological productivity and biodiversity. Values of inputs and outputs provide nutrient and water balances that clearly document the important role of Mirror Lake as a key integrator of inflow relative to its outflow to Hubbard Brook.

Large-scale studies in limnology provided opportunities for ecologists, hydrologists, geologists, and atmospheric scientists to work together because interdisciplinary approaches were recognized as essential for a comprehensive analysis. For example, inputs of nitrogen and carbon from precipitation and wind-blown dust often provide sufficient amounts of these essential elements to entire regions within a lake district or large watershed. In contrast, phosphorus is usually scarcely distributed in catchments because of the localization of its geological deposits. Consequently, phosphorus is often the limiting nutrient in many inland waters because it has no atmospheric gas phase like carbon and nitrogen and can enter only as particulates or as runoff (Wetzel, 2001). Phosphorus limitation also results when this nutrient is stored in plant biomass and taken out of circulation until herbivores consume the plant biomass and recycle the organic phosphorus back into solution as orthophosphate. Also, there is considerable inorganic storage in lake sediments where phosphorus remains as an insoluble form (often bound by colloids and by chelates such as ferric hydroxide, or in various mineralized forms such as calcium phosphate in oxidized sediments). Frequently, phosphorus, nitrogen, and carbon are also available from locations of

human-derived, point sources (e.g. sewage-treatment plant effluent) and nonpoint sources (diffuse fertilizer-enriched runoff) in many drainage areas that require even more interdisciplinary perspectives on land use and water management.

The emerging conceptual framework of landscape-level ecosystem dynamics connected to hydrologic processes provided important insights for understanding and managing entire drainage basins (Cummins, 1974; Hynes, 1975). Terrestrial inputs of water and nutrients alter productivity of aquatic ecosystems in complex direct and indirect relationships. Food-web connections modify how energy flows affect nutrient cycling as well as nutrient storage in populations of the top-trophic level consumers. Fish biomass (standing stocks) can be managed to modify nutrient cycling and thus provide biological control of excessive plant productivity. For example, the presence of certain fish predators are know to alter their prey which in turn can alter the rates of consumption by their prey (herbivorous zooplankton feeding on phytoplankton) that alter plant productivity, nutrient cycling, and storage (Carpenter *et al.*, 2001). These indirect linkages among trophic levels are termed *trophic cascades* and provide a means for controlling nutrient levels and rates of eutrophication by managing fish populations. Regulating the harvesting and stocking of certain types of predatory fish results in different types and quantities of herbivores and plants predominating in biological communities. This biomanipulation of food webs is an important tool for maintaining water quality for recreation and drinking water supplies (Shapiro, 1990).

In summary, managing land use in riparian areas and in entire land-water food webs has become an important ecosystem-level, concept-based approach for controlling nutrient availability in aquatic habitats that are highly interconnected with terrestrial habitats through hydrologic processes (Malmqvist and Rundle, 2002; Power and Dietrich, 2002; Thoms and Parsons, 2003). The effects of wetland, floodplain, and groundwater connections emerge as critical links in understanding a landscape-level framework of analysis (Ward and Stanford, 1995; Smith, 1998; Benke *et al.*, 2000; Junk and Wantzen, 2004). Ecologists and hydrologists continue to explore in-depth connections among ecosystems (Hershey *et al.*, 1999; Benke, 2001; Ward *et al.*, 2002) and emphasize that landscape-level network connections (Rodriguez-Iturbe, 2000; Gomi *et al.*, 2002; Fisher *et al.*, 2004) are critical to understanding how changes in land use or climate alter whole drainage basins (Winter, 2001; Likens, 2004). Although many land-water connections were recognized in earlier studies, important linkages within drainage networks were rarely documented relative to rapidly changing land use or at the scale of nested catchments. The increased use of spatially explicit modeling has provided a means to incorporate patchiness and degrees of isolation into more comprehensive analysis of catchment connections (Gergel *et al.*, 2002; Poole, 2002).

THE PHYSICAL TEMPLATE AS CONTROLS ON ECOSYSTEMS

The geologic origins, landscape topography, and connections among inland aquatic ecosystems all form the physical template for processes that alter residence times of water as well as light and nutrient availability. These physical processes influence biological productivity within any drainage basin. The origin of aquatic ecosystems by geologic or other means (Hutchinson, 1957) often determines topographic locations and connectivity to biogeochemical processes within natural landscapes (Swanson *et al.*, 1988; D'Angelo *et al.*, 1997; Grimm *et al.*, 2003; McClain *et al.*, 2003). Subsequent events further connect different ecosystem components (e.g. rivers, lakes, groundwaters, and wetlands) so that coupled physical, chemical, and biological interactions develop over time (Bishop, 1995; Winter, 2001; Richards *et al.*, 2002; Likens, 2004). Atmospheric inputs, runoff, and residence times of water within drainage basins are bounded by the topographic template that constrains species distributions and biological productivity (Hynes, 1975). Thus, the directions, rates, and quantities of water movements influence patterns of species diversity, food-web structure, and function, which, in turn, regulate ecosystem dynamics such as biological productivity, decomposition, and cycling of nutrients and other materials.

Spatial relationships within landscapes strongly control biogeochemical nutrient cycling and biotic productivity (Sorrano *et al.*, 1999; Hershey *et al.*, 1999; Kling *et al.*, 2000). Recent studies on these coupled physical and biological relationships have stimulated new research perspectives (Rice *et al.*, 2001a,b; Poole, 2002; Fisher *et al.*, 2004). As a result, the International Hydrological Program recognized the importance of eco-hydrological connectivity as a major focus for research and management (Janauer, 2000). The interdisciplinary conceptual integration needed to understand how drainage networks can regulate ecosystem processes and biological productivity is emerging as an important area of synthesis. This review first considers how individual streams and rivers are hydrologically linked and then includes the additional complexity of connected lakes and floodplains. Finally, isolated standing waters are considered as additional components that have indirect hydrologic connections.

River Drainage Networks

Running waters are organized in a hierarchy of flow paths from small streams to large rivers so that their combined erosive energy creates a drainage network regulated by the geology, climate, and disturbance history of a catchment

basin (Nakamura *et al.*, 2000; Osterkamp, 2002; Kinner and Stallard, 2004). Drainage density (the ratio of the total stream channel lengths of a drainage network relative to the drainage basin area) is one indicator (Table 2) of how water moves across a landscape, altering runoff and infiltration over geological timescales (Osterkamp, 2002). Along with climate, the topology of the network controls rates of sediment, nutrient transport, and routes of dispersal by many species, which, in turn, influences species distributions.

Drainage networks are dynamic in space and time with surface and subsurface waters interconnected vertically and horizontally. Channel locations at the confluences of tributaries (nodes) have physical attributes that are different from sections of river channels (links) within the drainage network. These confluences result in the combined transport and deposition of sediments that are abrupt and nonlinear in terms of the sizes and amounts of substrata (Benda *et al.*, 2004). The size distributions of substrata and their residence times (stability of organic and inorganic deposits) influence biotic habitats and abundances of bottom-dwelling (benthic) species.

A multidimensional framework is needed to integrate the flows within the network because of the complex horizontal linkages among pools and riffles within the channel and many types of lateral connections with floodplains (discussed in the following sections). Vertical connections within many sediment-filled river channels also link subsurface flows through groundwater and the porous sediments within channels (the hyporheic zone) to create a highly complex set of flow paths through interstitial deposits (Boulton *et al.*, 1998; Poole, 2002; Fisher *et al.*, 2004; Poole *et al.*, 2004). This subsurface zone is often an active site for exchange of water, dissolved nutrients, and organic matter that contains a high diversity of benthic invertebrates and microbes. Up-welling waters typically contain dissolved nutrients that increase algal production at the channel surface. Down-welling zones often transport dissolved oxygen and organic matter into depths where microbial and invertebrate communities process the materials (Stead *et al.*, 2004).

Drainage network characteristics provide numerous other conceptual bases for linking physical and biotic characteristics. One of the first examples was the River Continuum Concept (RCC). The RCC considered tributary sizes and linkages (stream order) together with channel widths as predictors of biotic community composition and function (Vannote *et al.*, 1980). Although this initial emphasis on a linear profile in the RCC was useful as a basis for forecasting the distributions of functional feeding groupings of invertebrate species (primarily aquatic insects) along some rivers, it was insufficient for understanding how other riverine food webs functioned (Perry and Schaeffer, 1987; Thorp and Delong, 1994; Fisher, 1997; Fisher *et al.*, 1998). One of the main features of the RCC was a focus on connectivity between riparian forest inputs of leaf litter and headwater stream energy sources for detritivores. Another feature was the longitudinal coupling of species based on stream orders. The pattern of organic-matter processing emphasizes that species (shredders) in headwaters (upper tributaries that comprise first- and second-order streams) break down coarse leaf material into fine suspended particulates that are consumed by downstream, filter-feeding species. These processing chains in forested headwaters depend on adequate flow and turbulent transport. These biotic linkages provide an example of flow-mediated ecosystem processing that is drought sensitive, in temperate-zone streams (Whiles and Dodds, 2002) and even in tropical rain forests (Crowl *et al.*, 2001; Covich *et al.*, 2003).

Rice *et al.* (2001a,b) developed the Link Discontinuity Concept (LDC) on the basis of the distribution of distinct habitats along elevational profiles of river drainages that are associated with sediment grain sizes resulting from various types and sizes of tributary confluences. These nodes in the drainage network are characterized by distinct substrata that are preferred by certain species. The differences in substrata identified in the LDC were not previously examined in detail within the RCC. The LDC conceptualization does not compress various stream orders into a linear elevational gradient (as in the RCC). Rather, the LDC emphasizes disjunct distributions of stream power as a framework for predicting species distributions in the drainage network as was also recognized by earlier studies (Ward and Stanford, 1995; Montgomery,1999).

There are numerous case studies describing how drainage networks are structured and how local geomorphic features influence ecological productivity (Hershey *et al.*, 1999; Gomi *et al.*, 2002; Parsons *et al.*, 2003). A general pattern occurs among drainages in high-relief, montane catchments where the particular locations of waterfalls create barriers for upstream migrations of fish, crustacean, and molluscan species (Covich and McDowell, 1996; Covich *et al.*, 1996; Smith *et al.*, 2003). For example, in the Luquillo Long Term Ecological Research Project in the Luquillo Mountains of Puerto Rico, the locations of the first waterfalls as a function of distance from the coast are the best predictors of biotic communities and food-web structure (Figure 1). Numerous species of predatory fishes live in the coastal estuaries and many can move various distances upstream into freshwater rivers (Covich and McDowell, 1996). Other predators such as wading birds and large decapod crustaceans (freshwater crabs and shrimps) are important consumers that are most diverse at low elevations, and some species are widely distributed in upstream portions of the drainage network. The abundance and size distributions of invertebrates in many headwater pools results from a large number of juveniles that migrate upstream from estuaries. As these invertebrates move upstream and grow, they are subject to fish predation

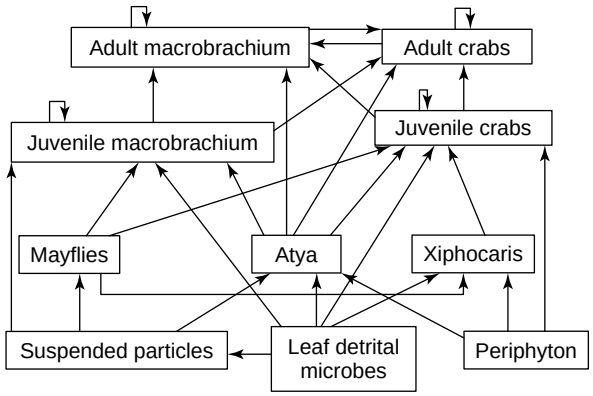


Figure 1 Food-web connections in headwater tropical streams of the Luquillo Mountains, Puerto Rico (Modified from Covich and McDowell, 1996) Arrows indicate how energy flows from in-stream primary production by algae (periphyton) and from riparian inputs of leaf litter (processed by bacteria and fungi) to consumers (insects and decapods). Loops indicated cannibalism by different ages of freshwater shrimp (*Macrobrachium carcinus*) and crabs (*Epilobocera sinuatifrons*)

by mountain mullets (*Agonostomus monticola)* and eels (*Anguilla rostrata*) at lower elevations. However, the upstream distribution of these two dominant fish predators are limited by the location of steep waterfalls ($>8-10$ m in height) that create barriers. Typically, the lowest-elevation waterfalls are relatively small and many are submerged during floods. These periods of inundation allow for upstream movements by migrating fishes. However, a threshold exists in the height of waterfalls that forms persistent barriers. Those waterfalls that are too high to be submerged are the primary geomorphic features that spatially limit assembly of food webs and cause upper tributaries to differ from lower portions of the drainage network. Thus, the local, spatial differences of topographic barriers in various catchments can modify upstream dispersal and formation of food webs. Generally, geomorphic features can help forecast ecosystem structure and processing rates by the ways terrain controls distributions of various key species.

River Drainage Networks with Standing-Water Connections

Direct, visible connections among flowing-water and standing-water habitats form complex patterns that have critical linkages with terrestrial habitats. The complex digitations of shorelines formed by reservoirs and many natural lakes increase the terrestrial transfers across these zones of runoff and organic detritus. These land-water connections are altered by changes in water levels that occur frequently in lakes, rivers, and wetlands during periods of seasonal floods and droughts. Longer-term, climate-induced drying out of springs, tributaries, and other habitats further change patterns of connectivity as a result of changing annual water balances. River inflows from drainage areas, inflows from groundwaters, and direct precipitation (on their lake surface areas) are the total annual inputs while evaporation and outflows are the main outputs that determine the annual lake water balance. Water balances are also modified by vegetative cover and land use that influence sediment transport and infilling of lakes and reservoirs, especially during peak periods of runoff and flood pulses.

The number and distribution of in-flowing and out-flowing rivers and the management of dams, diversions, and reservoirs are major variables that determine lake levels, water quality, flood control, and connectivity with shoreline habitats. For example, the impact of metropolitan Atlanta on the downstream river channel of the Chattahoochee River (Figure 2) provides a case study of one of the most impacted rivers in North America. Models of reservoir productivity and complex food-web dynamics provide useful predictions needed for management and decision making (Osidele and Beck, 2004). The reservoirs upstream and downstream of Atlanta have extremely different levels of nutrient cycling and productivity. These complex hydrological interactions among multiple reservoirs and linked rivers within a drainage network create challenges for management of alternative uses of the reservoirs for recreation, flood control, hydropower production and municipal water supplies (Couch *et al.*, 1996). These challenges are especially acute during periods of floods and drought.

The capacity for water storage and long-term continuity of outflows in rivers and lakes is decreased as suspended materials are transported by rivers and deposited in channels and lake basins. These rates of sedimentary infilling are especially high in steep terrain with sparse vegetative cover, intensive land use, and extremes in precipitation. Reservoirs are often prone to rapid sedimentation in drainage areas with poorly managed land and, thus have a limited lifetime storage capacity. A few large stormflows can transport large inputs of sediments and require expensive dredging to restore the storage capacity. Changes in climate and alterations of vegetative cover and land use can result in major shifts in erosion (from weathering and landslides) that alter rates of sediment and nutrient accumulation, water storage capacities, and turbidity (Waters, 1995; Turner *et al.*, 2003). Particulates with absorbed phosphorus, nitrogen, and carbon as well as trace elements are transported downstream into lakes and reservoirs (Waters, 1995; Smith, 1998). The delivery of sediments to lakes or from one lake to another along river connections is hydrologically determined by variations in precipitation and runoff.

Large proportions of the total nitrogen and phosphorus inputs to a drainage basin are retained in small headwater streams by sediments and microbes through a process of nutrient spiraling in which nitrogen and phosphorus move in and out of a series of abiotic and biotic components within

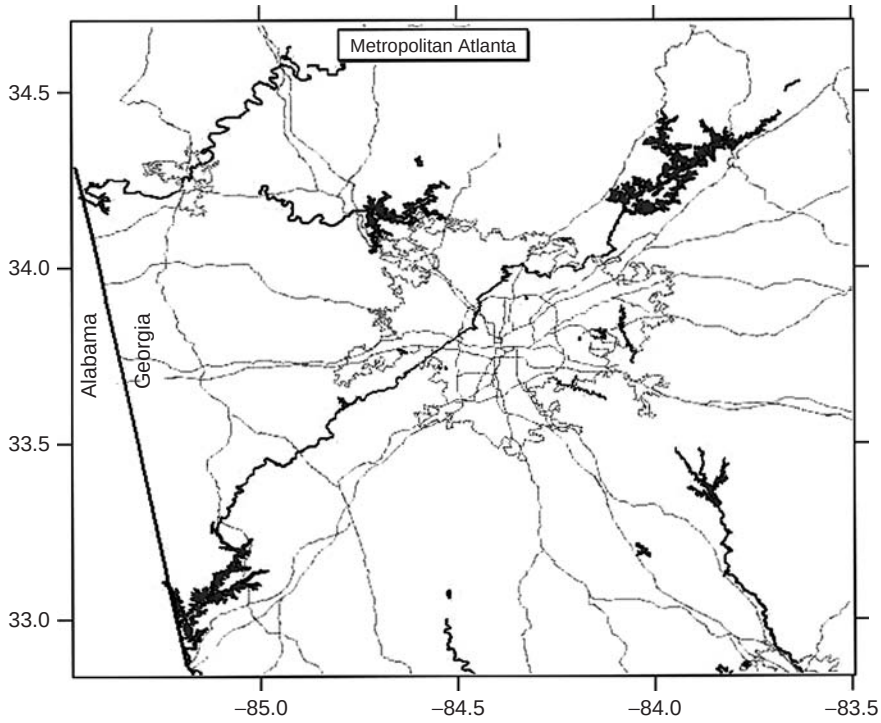


Figure 2 Effects of multiple reservoirs on downstream ecosystems. The Chattahoochee River in northeast Georgia flows from Lake Sidney Lanier through Atlanta to West Point Lake on the border with Alabama. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the channel (Newbold *et al.*, 1983; House, 2003). Nutrients are also transported by animals as they move from one habitat to another; some species of fish and other consumers travel considerable distances within drainages (Hobbs, 1999; Fry, 2002; Vanni, 2002; Vanni and Headworth, 2004). Significant percentages of nitrogen are transformed through the process of denitrification into ammonia gas and released into the atmosphere (Alexander *et al.*, 2000; Thomas *et al.*, 2001). Locations (termed "hot spots") with relatively rapid nutrient transformations by microbial communities are associated with hydrologic zones characterized by relatively low dissolved oxygen and warm water (Grimm *et al.*, 2003; McClain *et al.*, 2003).

River-connected Oxbow Lakes and Floodplain Lakes

Differences in locations among rivers, lakes, and wetlands greatly influence how nutrients and organisms move within drainage basins. River-connected lakes associated with floodplains (parafluvial habitats) have distinct associations with seasonal flooding that inundate their otherwise isolated basins (Benke *et al.*, 2000; Junk and Wantzen, 2004). Floodplain-associated wetlands are spatially and temporally interconnected in complex ways that reflect geology, terrain, and management of surface and groundwaters. Floodplains are among the most intensively studied wetlands because the frequency and duration of inundation clearly alters transport and deposition of nutrient-enriched sediments and inflows of river waters, often reflecting predictable annual patterns of seasonal inundation (Lewis *et al.*, 2001; Knowlton and Jones, 2003; Hamilton *et al.*, 2004). These predictable seasonal processes increase access to both standing- and flowing-water habitats by fishes and other organisms that are well adapted to seasonal migrations and dispersal (Bornette *et al.*, 1998; Amoros and Bornette, 2002). The spatiotemporal complexity of flooding events also creates increased habitat heterogeneity that is critically important in maintaining high biodiversity and productivity (Benke *et al.*, 2000; Junk and Wantzen, 2004). Floodplain lakes and cutoffs of river-meander channels (ox-bowl lakes) are usually shallow and highly productive ecosystems, especially for fish that can readily move among these inundated habitats at different times of years to forage and reproduce. The development of lake stratification and subsurface flows among floodplain lakes and wetlands adds to habitat complexity and results in variable nutrient cycling and plant productivity.

The cessation or alteration of the timing and extent of natural flood events by flow management (such as channelization, diversions of river water and construction of locks and dams) greatly alters the natural flow regime and downstream biotic productivity (Poff *et al.*, 1997; Pringle, 2003b; Tockner *et al.*, 2000; Tockner and Stanford, 2002;

Thoms, 2003; Thoms and Parsons, 2003; Lytle and Poff, 2004). Cumulative effects of wetland drainage and water diversions result in losses of ecosystem services naturally provided by ecological processes that sustain wildlife populations, filter nutrients, and reduce risks of flooding to society (Kiker *et al.*, 2001; Covich *et al.*, 2004b).

River-connected Lake Chains and Lake Outflows

Some drainage basin networks are characterized by chains of interconnected lakes and rivers (Kling *et al.*, 2000; Webster *et al.*, 2000) and chains of reservoirs (Elser and Kimmel, 1985). In montane regions, these chains have hydrologic connections based on major differences in elevation that create complex spatial and temporal relationships in terms of water temperatures, transparency, nutrient concentrations, conductivity, salinity, and biotic productivity. Lakes connected by large, low-gradient rivers often have significant lags in transport of materials and residence times of nutrients. These lags are especially important in those lakes and reservoirs that are modified for flood control, navigation, and other purposes. For example, the upper Mississippi River (from Minneapolis, Minnesota to St. Louis, Missouri, USA) has been converted into a series of 29 reservoirs through construction of locks and dams for regulating water levels that enhance barge navigation. This section is quite different from the free-flowing lower Mississippi River (from the confluence of the Ohio River to the Gulf of Mexico) where the channel is constrained by flood-control levees. High discharge in naturally unconstrained rivers that connect a series of lake basins can create backflows, over-bank flooding of wetlands, and complex groundwater flowpaths in watersheds with low-gradients. Low-gradient connections that depend primarily on groundwater or small tributaries have complex but predictable down-slope patterns of water chemistry and productivity. Large-river ecosystem models are available to compare continuous flowing reaches (Thorp and Delong, 1994) with those that have distinct interruptions of managed flows (Ward and Stanford, 1995; Thoms and Parsons, 2003).

The Northern Lakes Long Term Ecological Research project in Wisconsin, USA, provides some of the first studies to document these surface and subsurface connections in low-relief terrain (Kratz *et al.*, 1991; Kratz *et al.*, 1997; Sorrano *et al.*, 1999; Baines *et al.*, 2000; Riera *et al.*, 2000; Webster *et al.*, 2000). Another well-studied example of a natural chain of lakes is located in the Arctic Long Term Ecological Research project in the northern foothills of the Brooks Mountain Range, Alaska, USA. The Toolik Lake catchment drains 66.9 km^2 in several subbasins. The largest subbasin covers 46.6 km^2 and its longest river drainage network contains 10 lakes and these flows are linked prior to entering Toolik Lake. The mean depths of these 10 lakes range for 4.9 to 25 m and have distinct differences in residence times and water balances during the spring thaw. Each inlet and outlet alters the downstream connections among these linked ecosystems through changes in alkalinity, pH, conductivity, nutrients, and dissolved organic carbon. As the drainage area increases along the downstream locations, there are predictable coherent relationships among abiotic variables and material processing (Kling *et al.*, 2000).

The Kissimmee River is another example of a complex natural river ecosystem that historically connected Lake Kissimmee with Lake Okeechobee in Florida. The flows among these connected ecosystems, the Kissimmee Chain of Lakes, were altered between 1962 and 1971 as part of the Central and Southern Florida Food Control Project. The meandering river was channelized so that 167 km of its natural channel was transformed into a 90-km-long drainage canal (9-m deep and 100-m wide). This project drained more than 21 000 ha of floodplain wetlands. Flow between lakes was regulated by canals, drainage ditches, levees, and pumping stations to provide water for irrigation as well as flood control, creation of pastures, and diversions of out-flowing nutrients to minimize eutrophication of downstream wetlands in the Everglades. Meeting the demands for urban and rural water uses resulted in greatly modified discharge to coastal zones, estuaries, and associated wetlands. Although the natural flow regime was altered, it was not irreversible. It is expensive to attempt restoration of both ecosystem structure and function and to reestablish a natural community again. A 70-km section of the river that influences 11 000 ha of wetlands is being restored to attempt to recreate a more natural flow regime (Dahm *et al.*, 1995; Steinman and Rosen, 2000; Warne *et al.*, 2000).

The Colorado River is another example of a large, complex river system in North America that now contains two of the largest reservoirs in the United States (Lake Mead, Lake Powell). These reservoirs in the lower Glen Canyon and Grand Canyon were designed to regulate complex flow regimes for generating hydropower and water storage to supply agricultural and municipal demands for water. A controlled flood in 1996 was used to restore substrata downstream of Glen Canyon Dam and Lake Powell by moving sediments and recreating beaches and backwater habitats (Patten *et al.*, 2001; Schmidt *et al.*, 2001). This experiment was well designed and documented, but the results indicated that relatively brief, infrequent, small floods are insufficient as a management solution for restoring the coarse-grained habitats needed by endemic species of fishes for spawning or for restoring the riparian vegetation dynamics needed by endangered wildlife (e.g. the southwestern willow flycatcher). This event was also insufficient to create persistent beaches for recreational uses such as river rafting. The study also helped in determining how the loss of volume in the reservoir altered water quality (Hueftle and Stevens, 2001) as well as loss of ecosystem services such as hydropower generation. Additional studies

were conducted to learn more about sediment transport and how dam operations can be used to manipulate downstream water quality. In general, restoration of river drainage connections within chains of lakes and reservoirs requires a perspective that recognizes that multiple, habitat-forming hydrologic events are needed at large spatial and temporal scales (Clarke *et al.*, 2003; Robinson *et al.*, 2004; Ewel, 2001).

Precipitation and weathering of rocks and soils within a drainage basin contribute inorganic elements to flowpaths that converge as surface or subsurface network flows into channels and basins. Webster *et al.* (2000) defined 3 functional classes among spatially connected lake networks: (i) spatially uniform lakes have a high degree of synchrony; (ii) spatially structured lakes have local synchrony among those within similar elevations; (iii) unstructured lakes lack synchrony. Thus, downstream flows among interconnected lakes create a wide range of variability in synchronous responses to disturbances based on differences in lake-basin morphometry and residence times (Kratz *et al.*, 1997; Sorrano *et al.*, 1999). One of the major characteristics of lake chains is that the inflows to rivers from lake outlets differ from inflows derived from tributaries within the drainage basin. Lake outlets are characterized by dynamic changes in water temperatures, nutrient concentrations, algal production, and suspended organic matter from density-stratified lakes and reservoirs (Richardson and MacKay, 1991; Cattaneo, 1996; Robinson and Burgher, 1999).

Closed Lake Basins and Wetlands

Depending on locations and shapes of closed lake basins (those with no outflows), these standing waters differ significantly with regard to their annual water balances. These basins lose water volume primarily through evaporation and sometimes through groundwater seepage. There is a great deal of research focused on vernal pools, wetlands, and closed lakes because they are important components in the hydrologic cycle (Bullock and Acreman, 2003). However, their water quality is changing and they are vanishing from many landscapes as a result of draining natural basins, diversions of water to urban uses, and infilling of these habitats during dry periods (Leibowitz, 2003; Whigham and Jordan, 2003). The hydrologic dynamics of these ecosystems also take on legal considerations as legislation is passed that is intended to protect them under "no net loss" guidelines or to devalue their ecological importance because they do not fit in certain classifications (Winter and LaBaugh, 2003). Smaller closed basins are particularly sensitive to drought and other changes in precipitation (Fritz *et al.*, 2000; Winter, 2001). However, many larger lacustrine and wetland habitats are also seasonal with varied periods of persistence ranging from weeks to months each year. Playa lakes and wetlands in the semiarid western USA are examples of isolated, closed aquatic habitats that vary greatly in size and permanence with complex dispersal mechanisms controlling species diversity (Hall *et al.*, 2004).

The duration of wetland habitats (hydroperiod) influences how many species can consistently survive in these temporary waters (Schneider and Frost, 1996; Spencer *et al.*, 1999; Batzer *et al.*, 2004). These species typically have adaptations such as resting stages for long diapause or rapid dispersal and short life spans that allow them to reestablish in these isolated sites once water levels rise (Jeffries, 1994). Although drying out of these wetlands limits the dispersal of some species, this factor alone is not sufficient to account for the distributions of species in these communities. Community structure, particularly the presence or absence of predators and the diversity of predators, also often determines total species diversity and how these ecosystems function (Wissinger, 1999; Batzer *et al.*, 2004).

Large closed basins reflect long-term changes in climate as the ratios of evaporation and precipitation shift and result in major changes in lake levels, such as are recorded in the geologic record of closed basins such as the Great Salt Lake in Utah and Mono Lake in California, USA. Because many features influence water balances within lake districts composed of closed drainage basins and variable precipitation, some lake levels may decrease while others increase even if they experience identical annual precipitation. For example, the seasonal distribution of rainfall and snowfall alters rates of infiltration and runoff as do the types of vegetation and land use. Disturbances such as fire and landslides further interact with the geomorphically modified terrain to control inflows to closed basins. The shapes and sizes of the lake basins also control what proportion of the lake's volume is open to evaporation from its surface so that changes in lake levels are complex but predictable (Hutchinson, 1957).

Seeps and groundwater flows transport nutrients in various directions within rivers, lakes, and wetlands in response to floods and droughts as well as human impacts such as water diversions, groundwater pumping, and water storage in surface and subsurface reservoirs. Water quality in closed basins reflects the increased salinities and concentrations of toxins that accumulate from long periods of evaporation. For example, draining wetlands has concentrated wildlife in remaining habitats. However, diversion of irrigation waters to create wetlands, such as the Kesterson Refuge in the San Joaquin Valley of California, resulted in evaporative accumulation of selenium, a toxic metal that was lethal to many species of waterfowl. The geologic deposits of selenium are distributed widely in the arid western United States and these sites have similar risks for waterfowl. These natural processes can have detrimental consequences for managed wetlands (Pringle, 2003c).

HYDROLOGY AND ECOSYSTEM STRUCTURE

Ecosystem processes in aquatic ecosystems are often regulated by flow rates, water temperatures, and water balances that reflect regional patterns of precipitation and runoff regimes. Thus, hydrology influences how biotic components form functional groups of species that perform specific ecosystem processes such as primary and secondary productivity. Water-transported nutrients and sediments control production of plant matter, which then controls food availability for grazers and omnivores. Food webs provide maps of energy flows within the ecosystem and indicate how particular species interact to play key roles in the context that results from different degrees of hydrologic connectivity.

Primary and Secondary Productivity

Total or gross productivity is the amount of dry biomass (or energy equivalents) accumulated per unit area per unit time. Units for annual productivity are usually expressed in terms of biomass ($kg\,m^{-2}\,year^{-1}$) or energy ($Kcal\,m^{-2}\,year^{-1}$). The energy stored as biomass and which is not lost when some maintenance energy is allocated to respiration is termed *standing stock biomass* or net productivity. This harvestable material can be used to calculate the total energy fixed by also measuring respiration rates in plants and respiration and assimilation rates in animals, then adding those values to obtain the gross or total productivity. This plant and animal biomass and its associated storage of nutrients can be retained in the same location where primary and secondary production occurred. A portion is often transported downstream by river discharge to lakes or estuaries. Moreover, some of the productivity moves upstream as fish biomass. Many types of invertebrates migrate upstream to disperse and reproduce in headwater reaches. Some amount of the productivity flies or crawls off as insects and other invertebrates disperse throughout the catchment, thereby linking the aquatic and terrestrial ecosystems.

The large diversity of riparian and aquatic plants and animals play major roles in primary and secondary productivity within aquatic ecosystems. Dominance by a few species is typical of many aquatic communities, especially those that are physiologically stressful such as those with low dissolved oxygen, high salinity, high temperature, and high nutrients. Thus, the relationships between species diversity and ecosystem processes are important because changes in connectivity influence population and community dynamics. These changes, in turn, can alter rates of productivity (Wallace and Hutchens, 2000; Naeem and Wright, 2003; Covich *et al.*, 2004b; Giller *et al.*, 2004a; Morin and McGrady-Steed, 2004). Inland aquatic ecosystems provide potentially good testing areas for conceptual frameworks relating dispersal, biodiversity, and ecosystem functioning because of their different degrees of connectivity and varied community structure.

Aquatic Productivity, Nutrient Cycling, and Ecosystem Connectivity

The total amounts and proportions of available macro- and micronutrients that limit plant and animal productivity are derived from catchments and delivered by distinct flow paths (Smith, 1998; Sterner and Elser, 2002; Dodds *et al.*, 2004). The resulting changes in plant and animal productivity are measurable, visible signs of how ecosystems respond to hydrologic inputs and associated nutrient transport and cycling. Rates of plant productivity are also influenced by nutrient cycling (external and internal to specific ecosystems) that is regulated by microbial processes, sediment reworking, and grazing by herbivores at different locations within the drainage network (McClain *et al.*, 2003; Poole *et al.*, 2004).

Periods of high nutrient availability (eutrophication) are often accompanied by low flows that prevent dilution and flushing of excessive nutrients. This low-flow condition can cause algal blooms and massive growths of aquatic vegetation that undergo rapid population growth by a few dominant plant species (Smith, 1998). Plant material accumulates if not consumed by herbivores or washed out of the channel or basin. Plant matter also increases because often the dominant, fast-growing plant species are distasteful to consumers and some are toxic so that rates of removal by grazing often decrease. Pulses of excessive plant productivity can lead to oxygen depletion as plants die off and decompose. Microbial respiration during decomposition can require oxygen, yet dissolved oxygen in density-stratified waters is not mixed from the uppermost surface to these lower layers. The lack of dissolved oxygen, in turn, can result in death (and decomposition) of fish and other consumer species, depleting oxygen even further. This process of excessive growth and decay continues until the lake or pond stratification is broken down by seasonal changes in temperatures that mix the entire volume and more dissolved oxygen enters the water from the atmosphere or until some major inflow of water with different densities (determined by relative temperatures and salinities of the river inflows and the standing waters) destabilizes the density stratification and washes out accumulated nutrients in the basin or channel. As a result of these spatial and temporal relationships, fisheries' yields and the quality of fish harvests can differ markedly from one site to another on the basis of geologic origins, geographic locations, sizes of drainage areas, nutrient inputs, and specific composition of food webs (Cabana and Rassmussen, 1996; Power and Dietrich, 2002; Rassmussen and Vander Zanden, 2004).

Dispersal and Colonization

The assembly of species in a biotic community initially reflects colonization from a regional pool of species that is further modified over time as the assembly is subject to hydrological and geomorphic changes (Lamouroux *et al.*, 2002; Merigoux and Doledec, 2004). Changes in the number of species and their relative abundance within a biotic community reflect spatial relationships among suitable habitats (Osborne and Wiley, 1992; Poff, 1997; Matthews and Robinson, 1998; Olden *et al.*, 2001). Thus, biotic communities are subsets of species that occur in local habitats and move from one location to another in response to disturbances and in search of high-quality habitats (Holt, 2002, 2004). As some species become locally extinct, other species are likely to replace them. The history of how different species disperse among alternative high-quality habitats provides important insights regarding assembly and persistence of natural communities. For example, dispersal of species through drainage networks provides for rapid recovery following disturbances (such as floods and droughts as discussed in the following sections). Following disturbances, especially those that can lead to local extinctions of subpopulations, mobility is essential for reassembly and for sustaining these functional units together with their associated ecosystem processes. Subpopulations in productive habitats have the potential to be source populations. Source populations have access to high-quality resources and can reproduce a surplus of individuals who provide dispersing individuals to locations with variable habitat quality. The movement of species to low-quality habitats creates sink populations where individuals survive but do not reproduce and cannot sustain ecosystem services throughout the entire drainage.

The flow of genes among subpopulations (metapopulations) influences the long-term evolution of species. Populations that are isolated for periods of time but disperse and reproduce among subpopulations from time to time form interconnected metapopulations in complex drainage networks (Gotelli and Taylor, 1999). Similarly, aggregations of partially isolated communities (metacommunities) that perform key ecosystem processes may be altered by degrees of connectivity and flow of genetic information and capacity for adaptation to changing environmental conditions (Pulliam, 1988; Holt, 2002). These interacting local communities comprise a larger-scale metacommunity that includes multiple species capable of dispersing and recolonizing available habitats (Cottenie *et al.*, 2003; Mouquet and Loreau, 2003).

On the basis of means of dispersal, the biotic components of inland aquatic ecosystems consist of two major groups: (i) Species that require interconnecting streams and rivers for active or passive aquatic dispersal and colonization. These include mostly vertebrates and invertebrates such as certain mussels that rely on fish for dispersal; and (ii) Species that are adapted to overland dispersal during some phase of their life history. These include amphibious vertebrates and highly mobile groups of invertebrates, such as some crustacean species, as well as those that have aerial dispersal as adults, such as many aquatic insects (Bilton *et al.*, 2001). Distances and rates of passive dispersal are not well documented among most aquatic species but some long-distance transport is known (Bohonak and Jenkins, 2003).

Migratory movements of consumer species (such as many types of fishes and decapod crustaceans) generally result in a long-term biotic integration of spatially defined subsystems (Schindler and Scheuerell, 2002; Vadeboncoeur *et al.*, 2002). Thus, the species composition of any ecosystem is initially determined by dispersal abilities of species within a regional pool of colonizers and then their persistence depends on successful competition or coexistence with resident species and avoidance of predators. Sources of replacement species are influenced by connectivity, dispersal adaptations, frequency and intensity of disturbances, and the resulting heterogeneity characteristic of aquatic habitats.

Surface and subsurface connections serve as dispersal routes for many aquatic species and specialized propagules that are adapted for buoyancy and resistant to desiccation. These adaptations affect the dispersal of species that form new biotic communities as additional species are added or lost over time. For example, small habitats (such as isolated ponds or headwater streams above high waterfalls or low-water dams) typically have fewer species of fish because the absence of horizontal riverine connections or high vertical barriers prevent dispersal to these sites (Cumming, 2004). Because movements by fish dispersal also affect dispersal of some invertebrate species (such as some riverine mussels that rely on fish for transporting their larvae), these isolated communities often have fewer total species and are distinct from other similar, but spatially interconnected ecosystems. However, many of these isolated ecosystems are characterized by species capable of migrating from nearby locations without direct aquatic connections (Bilton *et al.*, 2001; Bohonak and Jenkins, 2003). For example, many small ponds and intermittent streams have diverse assemblages of invertebrates, in part because of the ability of these species to use other means (such as aerial dispersal) to colonize the habitats that do not require continuous aquatic connectivity (MacKay, 1992).

Studies of zooplankton distributions and genetics in small ponds indicate a combination of mechanisms and result in rapid colonization of new habitats, as well as recolonization of persistent ponds from other local ponds (Figuerola and Green, 2002; Havel *et al.*, 2002; Cohen and Shurin, 2003; Louette and De Meester, 2004). Geographic location and the length of time since the origin of a habitat greatly influence species richness, especially in larger lake and river

ecosystems that generally are more persistent over geologic and evolutionary timescales (Malmqvist, 2002).

The oldest lakes and rivers are centers for speciation as well as refugia from regional extinctions (Wilson *et al.*, 2004). These communities differ markedly from younger assemblages in having higher numbers of species in certain groups that have radiated over millennia and are specialized in their functions and with uniquely restricted distributions (Cohen, 2000; Lundberg *et al.*, 2000). These specialized, endemic species are thought to be at risk from drainage basins changes, especially from eutrophication and deposition of eroded sediments at the mouths of rivers flowing into ancient lakes, such as Lake Tanganyika in the African Rift Valley (Donohue and Irvine, 2004).

ECOSYSTEM RESPONSES TO HYDROLOGIC CONDITIONS

Changing water quality such as depletion of dissolved oxygen, warmer temperatures, or increased turbidity can alter habitats for species that require certain conditions. Unlike many terrestrial or marine habitats, most continental waters are often transitory, impermanent habitats that may dry out during drought (Stanley *et al.*, 1997; Fritz and Dodds, 2004; Stanley *et al.*, 2004), greatly altering their ecosystem processes (Dahm *et al.*, 2003; Lake, 2003). Shallow basins may fill in with sediments following decades of erosion and deposition by variable flow regimes (Fritz *et al.*, 2000). Rates of erosion are especially sensitive to management practices and natural events such as climate change associated with intense rainfall during wet years or prolonged drought during extremely dry years. The effects of insect damage or fire on terrestrial vegetation can lead to severe erosion and increased nutrient loading that affect riparian, river, and lake ecosystems, thereby altering fish distributions (Dunham *et al.*, 2003). Organisms adapted to persist in inland waters, thus, usually have some mechanism for dispersing among continental and insular habitats. These different adaptations are relevant for understanding the rates of recolonization following chronic disturbances and catastrophic loss of species that occur in response to extreme natural events as well as to human-caused disturbances.

Functional Resilience and Species Dispersal

The ways in which species have evolved life histories in response to natural flow regimes and drought over millennia are known to alter their patterns of dispersal and their abilities to respond and persist following disturbances (Lake, 2003; Lytle and Poff, 2004). To resolve local and regional issues of water management as well as protection of rare species throughout their distributions, it is widely recognized that a comprehensive, landscape-level approach to management is needed (Poiani *et al.*, 2000; Gergel *et al.*, 2002; Pringle, 2003b).

The biotic changes associated with chronic pollution such as eutrophication or climate change can lead to major shifts in dominance of resident species or to gradual evolutionary changes among species in habitats with relatively restricted access by dispersing species. Recovery rates of biotic communities and their associated ecosystem functions following a catastrophic event (e.g. extreme floods, droughts, or accidental spills of toxins) are especially influenced by the degree of connectivity among similar habitats and their associated hydrological processes and residence times. The downstream drift or upstream migrations of various species generally result in rapid recovery (Gjerlov *et al.*, 2003). Stream communities are typically resilient to infrequent, local disturbances but may be less capable of recovery following frequent and intense disturbances (Resh *et al.*, 1988). Location and distance from sources, along with other factors such as age of the ecosystem are known to be very important for recolonization. For example, small lakes are often more isolated than streams, and natural rates of extinctions within fish communities are known to be greater than rates of colonization or recolonization (Tonn, 1990; Magnuson *et al.*, 1998; Cumming, 2004).

The physical locations and number of connections among habitats at the landscape level also greatly influence how flow-based disturbances modify food-web composition and function (Schlosser, 1991; Fausch *et al.*, 2002; Power and Dietrich, 2002; Woodward and Hildrew, 2002). For example, effects of extreme flood events are usually of short duration (days to months) because many species either: (i) resist washout from floods by occupying microhabitats such as structurally complex woody or rocky substrata (Townsend *et al.*, 1997); or (ii) rapidly recolonize interconnected habitats within the drainage network (MacKay, 1992; Covich *et al.*, 1996). However, effects of extreme drought and other persistent low-flow events (such as water diversions) on stream food webs can persist for several years (Stanley *et al.*, 1997; Covich *et al.*, 2003; Lake, 2003). Lowered discharge can dry out waterfalls, disrupt upstream migrations, and restrict flow-based chemical communication (Crowl and Covich, 1994). Similarly, low discharge greatly decreases habitat quality and availability by decreasing pool volumes and hydrologic connectivity (Bencala, 1993; Fausch *et al.*, 2002). Reduced flows lead to accumulations of organic and inorganic materials that further decrease pool volumes and quality by infilling with deposits.

As water moves across a landscape and through a drainage network, chemical characteristics are modified with varied concentrations of dissolved and suspended materials being transported downstream and utilized by the biota (Wallace and Hutchens, 2000; Whiles and Dodds,

2002). These transported materials reflect various combinations of geologic history, current and past land use as well as direct and indirect effects of pollution. Deposits of these materials accumulate in lake basins and wetlands that record changes over time in how the drainage network functions in response to atmospheric acidification, climate change and land uses (Fritz *et al.*, 2000).

Connectivity and Vulnerability to Invasive Nonnative Species

Invasive species can spread quickly across continents if the drainage network of freshwater habitats is highly interconnected (Malmqvist and Rundle, 2002; Gido *et al.*, 2004; Holeck *et al.*, 2004; MacIsaac *et al.*, 2004). The ability of nonnative species to disperse rapidly through a drainage network depends on the structure of the network, the location of the initial introduction, the dispersal ability of the nonnative species, and its competitiveness. Once established, nonnative species can modify the habitat, eliminate native species, and set in motion a series of changes that favors complete competitive dominance by nonnative species through alteration of water quality or substrata. All these features are being widely studied as a result of increased global trade and movement of many species in the ballast water of ships and aquarium trade shipments (Kolar and Lodge, 2002; MacIsaac *et al.*, 2004; Padilla and Williams, 2004). Models predicting spread of species with planktonic larvae such as nonnative zooplankton and mussels (especially from the Caspian Sea) incorporate measures of habitat connectivity and variable habitat quality. These models usually also consider transport by people associated with recreational uses of isolated lakes (dumping of bait buckets or movement of boats) as well as other nonintentional, passive types of dispersal. Intentional introductions of fishes and crayfishes for aquaculture production as well as fish (such as trout) for recreational fishing have occurred repeatedly in many parts of the world with numerous impacts on native species of fishes, amphibians, and aquatic invertebrates (Englund and Polhemus, 2001; Pilliod and Peterson, 2001; Knapp, 2005). These species can disperse into locations where they are not wanted, especially in locations with variable flows (Light, 2003). They often outcompete native species, alter relative abundances of persistent coexisting species, or spread diseases that result in differential mortality among native species.

Introductions of species that migrate upstream (such as trout, carp, and crayfish) or drift downstream (such as zebra mussels, *Dreissena polymorpha*) can rapidly disperse and alter substrata, food webs, and ecosystem processes (Beekey *et al.*, 2004). If the nonnative species are introduced at nodes in complex drainage networks with multiple linkages (such as interconnected lake chains and dendritic patterns of tributaries), the dispersal from these point sources can be relatively rapid and difficult to control. However, in some cases, restoration of native species is possible (Vredenburg, 2004).

Determining the sources of nutrition for individuals and documenting where they are capable of growing rapidly within a drainage area can be improved by the use of stable isotopes (Fry, 2002). For example, Fry and Allen (2003) sampled zebra mussels in the Mississippi River along a north–south transect from Minnesota to Louisiana. They detected "local zones of influence" by measuring stable isotopes of carbon, nitrogen, and sulfur in these filter feeders and by relating the sources of nutrients to land use in different tributaries. These connectivities were especially evident in the upper Illinois River that was influenced by municipal point sources of nutrients while other sections of tributaries were most likely influenced by agricultural nonpoint sources of nutrients.

The connectivity among floodplain lakes (as recognized very early for distributions of native fishes by Forbes, 1887) within river networks takes on considerable importance in ongoing studies of invasive species such as Asian carp that are moving upstream within large rivers of North America (Chick and Pegg, 2001; Pegg *et al.*, 2002). For example, the high degree of connectivity among tributaries, floodplain lakes, and off-channel wetlands with the main channels of the Upper Mississippi River and of coastal rivers in the southeastern United States provides a large-scale illustration of a complex drainage basin where native species are at risk because of their need for interconnected habitats (Fry, 2002). Predicting the probabilities of dispersal among communities of existing native species and developing strategies for protection of isolated rare species require careful consideration of aquatic connectivity (Scheerer, 2002; Koel, 2004; Cooke *et al.*, 2005). Other well-connected habitats include vast wetlands such as the Everglades in southern Florida. Models of fish dispersal and persistence also include the importance of various degrees of spatial complexity across large areas of low relief. Small differences in topography have large effects during periods of floods and droughts because refugia are limited to specific sites that reflect geological history and biotic modifications of the terrain (DeAngelis *et al.*, 1997).

Because most oceanic islands are relatively isolated and geologically young, they generally have fewer native freshwater species than continental drainages; their drainage areas are also smaller and less complex, so that predictions based on these relatively simple systems are more likely to be useful in forecasting ecosystem-level changes (Smith *et al.*, 2003). For example, geologically recent, insular drainages on the Caribbean and South Pacific islands, typically, have fewer than 10 fish species (Covich and McDowell, 1996; March *et al.*, 2003). These insular streams are likely to be vulnerable to multiple additions of nonnative

species because some of the more specialized mainland species may be able to outcompete the natives species if they have a wider niche than the invading species (Sax and Brown, 2000). Invasive nonnative species can rapidly find niches in these relatively small drainage networks and some can potentially displace native resident species, especially where refugia for native species or food resources are limited. Release from predators, parasites, and diseases found in mainland rivers may allow nonnative species to spread rapidly throughout the drainage basin. These types of displacements of native species by introduced species are commonly observed in island drainages (Englund and Polhemus, 2001; Brasher, 2003; Craig, 2003; Smith *et al.*, 2003). In some cases, however, the native species can persist because they are often especially well adapted for highly variable flows that may occur only rarely but have significant impacts on the invasive species. For example, crayfish (*Procambarus clarkii*) was introduced to Hawaii from the United States and occurs in numerous small streams on some islands but this species has not yet dispersed into larger, fast-flowing rivers in the mountains; similarly, the Tahitian prawn (*Macrobrachium lar*) occurs widely in low-gradient coastal rivers but has not yet dispersed into steeper montane channels where high flows appear to limit its distribution and where native decapod species persist (Larned *et al.*, 2003). However, in other much older islands and those that were once connected to continents with high species richness, the number of endemics (native species restricted to a single island) is higher (such as in Madagascar), and some nonnative fish species have also been able to disperse widely and rapidly even though 143 native fishes in 54 genera and 21 families occur there and 65% are endemic (Benstead *et al.*, 2003).

Large, continental river basins contain numerous species of fishes reflecting the size and complexity of inland drainage networks. This heterogeneous, physical template has been critical for the evolution of fishes and other freshwater species. Structural heterogeneity of freshwater habitats and reshuffling of species during flood events have led to the evolution of a disproportionate number of inland fishes (Lundberg *et al.*, 2000; Hoeinghaus *et al.*, 2003). Despite the very low proportion of freshwater (0.01%) relative to the planet's total amount of water, freshwater fishes now number more than 10 000 species. This number is approximately 40% of all fish species (Lundberg *et al.*, 2000). Many of these species are in ancient tropical rivers, such as the Amazon River, the largest river in the world. Here the locations of tributaries and floodplain lakes are important over evolutionary timescales because geological and hydrological controls alter habitat availability (Costa *et al.*, 2001; Fritz *et al.*, 2004). For example, 43 species of electric fishes are known to occur within the Amazon River drainage network and recent analysis indicates that their localized diversity is enhanced by tributaries. The number of species in the main channel increases downstream of tributary confluences, representing a nodal model of community organization within the drainage network (Fernandes *et al.*, 2004). However, the expanding rates of deforestation, agricultural growth, road construction, dams, diversions, and growth of cities will have cumulative effects on dispersal and function of these diverse tropical species of fishes (Bojsen and Barriga, 2002; Wright and Flecker, 2004). Loss of species diversity of fishes and other freshwater biota over time will likely have irreversible consequences for ecosystem processes on very large scales (Giller *et al.*, 2004a,b; Meybeck, 2004).

CONCLUSIONS

Biological production in inland waters is limited by flows of energy and nutrients across landscapes that continually undergo complex dynamics driven by hydrological processes. Primary production (by photosynthesis or chemosynthesis), nutrient uptake and cycling, and decomposition of organic matter are all connected by the movement of water and organisms among coupled ecosystems. Inputs of water from precipitation, runoff, and infiltration to groundwater all combine to influence rates of transport, deposition, and recycling of inorganic materials which, in turn, influence habitat quality. The relative and absolute abundance of various species and their roles in functional groups that regulate ecosystem processes are controlled by a large number of hydrological variables that relate to habitat quality.

Distances among aquatic habitats, as well as the types and numbers of hydrologic linkages connecting them, influence the timing of different amounts of water and nutrients that enter and cycle through aquatic ecosystems. Flowing-water habitats (streams, rivers, estuaries) are linked to standing-water habitats (ponds, lakes, wetlands) by horizontal surface and subsurface connections in lake chains, lateral floodplains, and salt marshes. Rates of flows among these connected habitats are often more complex and variable than among small, isolated standing-water ecosystems (marshes, bogs, ponds) where residence times are often longer.

The dispersal of species through drainage networks provides for rapid recovery following disturbances (such as extreme floods and droughts). Dispersal to high-quality habitats by many species that comprise natural communities is essential to sustain reassembly of these functional units and their associated ecosystem processes, following disturbances. Source populations are those that continue to reproduce and, thus, provide dispersing individuals to locations with variable habitat quality. The movement of species to low-quality habitats creates sink populations where individuals survive but do not reproduce and cannot

sustain ecosystem services throughout the entire drainage. The flow of genes among subpopulations (metapopulations) influences the long-term evolution of species. Similarly, aggregations of partially isolated communities (metacommunities) that perform key ecosystem processes are also altered by degrees of connectivity and flow of genetic information, and capacity for adaptation to changing environmental conditions.

High degrees of connectivity also allow introduced, nonnative species to spread throughout a catchment. If the introduced species displace native species, rates of critical ecosystem processes can change. These processes and their associated ecological services depend on sustainable flows of water as well as on well-managed catchments where flow management, land use, and introductions of nonnative species can rapidly alter ecosystem productivity. Over time, the loss of native species which are well adapted for long-term variability in environmental conditions can lead to instability and unreliable provisioning of ecosystem services.

Management of drainage basins for hydropower, stormflow runoff from urban landscapes, and rapid economic development typically alter these critical natural connections by modifying the ecohydrology of riverine networks, especially headwater tributaries (Walters *et al.*, 2003). Loss of habitat and species are increasing as a result of these changes on a global scale (Meyer and Wallace, 2001; Pringle, 2003b). For example, reductions in stream density (stream length per drainage basin area) occur when natural stream channels are lost by substituting subsurface pipes for natural fluvial geomorphology. These areas are often then developed and paved so that infiltration of groundwater from channels is lost as are the natural processes that break down organic matter and retain organic matter *in situ* (Paul and Meyer, 2001). Increased surface runoff increases in paved-over areas of the catchment as the proportion of impervious surface increases. Continued paving of roads and building of rooftops, driveways, and parking lots all combine to increase peak flows, severity of flooding and bank erosion (Arnold and Gibbons, 1996; Gergel *et al.*, 2002; Brattebo and Booth, 2003). The integration of hydrology and aquatic ecology may help resolve some of these issues through enhanced environmental understanding and more coherent planning of biospheric and hydrospheric connections.

Acknowledgments

The manuscript has benefited from comments by Wyatt Cross, and an anonymous reviewer. Figures were prepared by Tamara Heartsill-Scalley and Gretchen Loeffler. Research on food webs, species functions, and drainage networks in Puerto Rico and Hawaii was supported by the US National Science Foundation.

REFERENCES

Alexander R.B., Smith R.A. and Shwartz G.E. (2000) Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico. *Nature*, **403**, 758–761.

Amoros C. and Bornette G. (2002) Connectivity and biocomplexity in waterbodies of riverine floodplains. *Freshwater Biology*, **47**, 761–776.

Arnold C.L. and Gibbons C.J. (1996) Impervious surface coverage: the emergence of a key environmental indicator. *American Planners Association Journal*, **62**, 243–258.

Baines S.B., Webster K.E., Kratz T.K., Carpenter S.R. and Magnuson J.J. (2000) Synchronous behavior of temperature, calcium and chlorophyll in lakes of northern Wisconsin. *Ecology*, **81**, 815–825.

Batzer D.P., Palik B.J. and Buech R. (2004). Relationships between environmental characteristics and macroinvertebrate communities in seasonal woodlands ponds of Minnesota. *Journal of the North American Benthological Society*, **23**, 50–68.

Beekey M.A., McCabe D.J. and Marsden J.E. (2004) Zebra mussel colonization of soft sediments facilitates invertebrate communities. *Freshwater Biology*, **49**, 535–545.

Bencala K.E. (1993) A perspective on stream-catchment connections. *Journal of the North American Benthological Society*, **12**, 44–47.

Benda L., Poff L., Miller D., Dunne T., Reeves G., Pess G. and Pollock M. (2004) The network dynamics hypothesis: how channel networks structure riverine habitats. *Bioscience*, **54**, 413–427.

Benke A.C. (2001) Importance of flood regime to invertebrate habitat in an unregulated river-floodplain ecosystem. *Journal of the North American Benthological Society*, **20**, 225–240.

Benke A.C., Chaubey I., Ward G.M. and Dunn E.L. (2000) Flood pulse dynamics of an unregulated river floodplain in the southeastern US coastal plain. *Ecology*, **81**, 2730–2741.

Benstead J.P., Rham P.H., Gattolliat J.-L., Gibon F.M., Loiselle P.V., Sartori M., Sparks J.S. and Stiassny M.L.J. (2003) Conserving Madagascar's freshwater biodiversity. *Bioscience*, **53**, 1101–1111.

Bilton D.T., Freeland J.R. and Okamura B. (2001) Dispersal in freshwater invertebrates. *Annual Review of Ecology and Systematics*, **32**, 159–181.

Bishop P. (1995) Drainage rearrangement by river capture, beheading and diversion. *Progress in Physical Geography*, **19**, 449–473.

Bohonak A.J. and Jenkins D.G. (2003) Ecological and evolutionary significance of dispersal by freshwater invertebrates. *Ecology Letters*, **6**, 783–796.

Bojsen B.H. and Barriga R. (2002) Effects of deforestation on fish community structure in Ecuadorian Amazon streams. *Freshwater Biology*, **47**, 2246–2260.

Bornette G., Amoros C. and Lamouroux N. (1998) Aquatic plant diversity in riverine wetlands: the role of connectivity. *Freshwater Biology*, **39**, 267–283.

Boulton A.J., Findlay S., Marmonier P., Stanley E.H. and Valett H.M. (1998) The functional significance of the hyporheic

zone in streams and rivers. *Annual Review of Ecology and Systematics*, **29**, 59–81.

Brasher A.M.D. (2003) Impacts of human disturbances on biotic communities in Hawaiian streams. *Bioscience*, **53**, 1052–1060.

Brattebo B.O. and Booth D.B. (2003) Long-term stormwater quantity and quality performance of permeable pavement systems. *Water Research*, **37**, 4369–4376.

Bullock A. and Acreman M. (2003) The role of wetlands in the hydrological cycle. *Hydrology and Earth System Sciences*, **7**, 358–398.

Cabana G. and Rassmussen J.B. (1996) Comparing the length of aquatic food chains using stable N isotopes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10 844–10 847.

Carpenter S.R., Cole J.J., Hodgson J.R., Kitchell J.F., Pace M.L., Bade D., Cottingham K.L., Essington T.E., Houser J.N. and Schindler D.E. (2001) Trophic cascades, nutrients, and lake productivity: experimental enrichment of lakes with contrasting food webs. *Ecological Monographs*, **71**, 163–186.

Cattaneo A. (1996) Algal seston and periphyton distribution along a stream linking a chain of lakes on the Canadian Shield. *Hydrobiologia*, **325**, 183–192.

Chick J.H. and Pegg M.A. (2001) Invasive carp in the Mississippi River Basin. *Science*, **292**, 2250–2251.

Church M. (2002) Geomorphic thresholds in riverine landscapes. *Freshwater Biology*, **47**, 541–557.

Clarke S.J., Bruce-Burgess L. and Wharton G. (2003) Linking form and function: towards an eco-hydromorphic approach to sustainable river restoration. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **13**, 439–450.

Cohen A.S. (2000) Linking spatial and temporal changes in the diversity of ancient lakes: examples from the ecology and palaeoecology of the Tanganyikan ostracodes. *Advances in Ecological Research*, **31**, 521–537.

Cohen G.M. and Shurin J.B. (2003) Scale-dependence and mechanisms of dispersal in freshwater zooplankton. *Oikos*, **103**, 603–617.

Cooke S.J., Bunt C.M., Hamilton S.J., Jennings C.A., Pearson M.P., Cooperman M.S. and Markle D.F. (2005) Threats, conservation strategies, and prognosis for suckers (Catostoidae) in North America: insights form regional case studies of a diverse family of non-game fishes. *Biological Conservation*, **121**, 317–331.

Costa J.B.S., Bemerguy R.L., Hasui Y. and Borges M.S. (2001) Tectonics and paleography along the Amazon River. *Journal of South American Earth Sciences*, **14**, 335–347.

Cottenie K., Michels E., Nuytten N. and De Meester L. (2003) Zooplankton metacommunity structure: regional versus local processes in highly interconnected ponds. *Ecology*, **84**, 991–1000.

Couch C.A., Hopkins E.H. and Hardy P.S. (1996) *Influences of Environmental Settings on Aquatic Ecosystems in the Apalachicola-Chattahoochee-Flint River Basin*, U.S. Geological Survey Water Resources Investigations Report 95–4278, USGS, Atlanta.

Covich A.P., Austen M.C., Barlocher F., Chauvet E., Cardinale B.J., Biles C.L., Inchausti P., Dangles O., Solan M., Gessner M.O., *et al.* (2004a) The role of biodiversity in the functioning of freshwater and marine benthic ecosystems. *Bioscience*, **54**, 767–775.

Covich A.P., Ewel K.C., Hall R.O., Giller P.G., Merritt D. and Goedkoop W. (2004b) Ecosystem Services Provided by Freshwater Benthos. In *Sustaining Biodiversity and Ecosystem Services in Soils and Sediments*, Wall D. (Ed.), Island Press: Washington, pp. 137–159.

Covich A.P., Crowl T.A., Johnson S.L. and Pyron M. (1996) Distribution and abundance of tropical freshwater shrimp along a stream corridor: response to disturbance. *Biotropica*, **28**, 484–492.

Covich A.P., Crowl T.A. and Scatena F.N. (2003) Effects of extreme low flows on freshwater shrimps in a perennial tropical stream. *Freshwater Biology*, **48**, 1199–1206.

Covich A.P. and McDowell W.H. (1996) The stream community. In *The Food Web of a Tropical Rain Forest*, Reagan D.P. and Waide R.B. (Eds.), University of Chicago Press: Chicago.

Craig D.A. (2003) Geomorphology, development of running water habitats, and evolution of black flies on Polynesian islands. *Bioscience*, **53**, 1079–1093.

Crowl T.A. and Covich A.P. (1994) Responses of a freshwater shrimp to chemical and tactile stimuli from a large decapod predator. *Journal of the North American Benthological Society*, **13**, 291–298.

Crowl T.A., McDowell W.H., Covich A.P. and Johnson S.L. (2001) Species-specific responses in leaf litter processing in a tropical headwater stream (Puerto Rico). *Ecology*, **82**, 775–783.

Cumming G.S. (2004) The impact of low-head dams on fish species in Wisconsin, USA. *Ecological Applications*, **14**, 1495–1506.

Cummins K.W. (1974) Structure and function of stream ecosystems. *Bioscience*, **24**, 631–640.

Dahm C.N., Baker M.A., Moore D.I. and Thibault J.R. (2003) Biogeochemistry of surface waters and alluvial ground waters in streams and rivers during drought. *Freshwater Biology*, **48**, 1219–1231.

Dahm C.N., Cummins K.W., Valett H.M. and Coleman R.L. (1995) An ecosystem view of the restoration of the Kissimmee River. *Restoration Ecology*, **3**, 225–238.

D'Angelo D.J., Gregory S.V., Ashkenas L.R. and Meyer J.L. (1997) Physical and biological linkages within a stream geomorphic hierarchy: a modeling approach. *Journal of the North American Benthological Society*, **16**, 480–502.

DeAngelis D.L., Loftus W.F., Trexler J.C. and Ulanowicz R.E. (1997) Modeling fish dynamics and effects of stress in a hydrologically pulsed ecosystem. *Journal of Aquatic Ecosystem Stress and Recovery*, **6**, 1–13.

Dodds W.K., Marti E., Tank J.L., Pontius J., Hamilton S.K., Grimm N.B., Bowden W.B., McDowell W.H., Peterson B.J., Valett H.M., *et al.* (2004) Carbon and nitrogen Stoichiometry and nitrogen cycling rates in streams. *Oecologia*, **140**, 458–467.

Donohue I. and Irvine K. (2004) Seasonal patterns of sediment loading and benthic invertebrate community dynamics in Lake Tanganyika, Africa. *Freshwater Biology*, **49**, 320–331.

Dunham J.B., Young M.K., Gresswell R.E. and Rieman B.E. (2003) Effects of fire on fish populations: landscape perspective on persistence of native and non-native fish invasions. *Forest Ecology and Management*, **178**, 183–196.

Elser J.J. and Kimmel B.A. (1985) Nutrient availability for phytoplankton production in a multiple-impoundment series. *Canadian Journal of Fisheries and Aquatic Sciences*, **42**, 1259–1370.

Ewel K.C. (2001) Managing critical transition zones. *Ecosystems*, **4**, 452–460.

Englund R.A. and Polhemus D.A. (2001) Evaluating the effects of introduced rainbow trout (*Oncorhynchus mykiss*) on native stream insects on Kauai island, Hawaii. *Journal of Insect Conservation*, **5**, 265–281.

Fausch K.D., Torgensen C.E., Baxter C.V. and Li J.W. (2002) Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *Bioscience*, **52**, 473–482.

Fernandes C.C., Podos J. and Lundberg J.G. (2004) Amazonian ecology: tributaries enhance the diversity of electric fishes. *Science*, **305**, 1960–1962.

Figuerola J. and Green A.J. (2002) Dispersal of aquatic organisms by waterbirds: a review of past research and priorities for future studies. *Freshwater Biology*, **47**, 483–494.

Fisher S.G. (1997) Creativity, idea generation, and the functional morphology of streams. *Journal of the North American Benthological Society*, **16**, 305–318.

Fisher S.G., Grimm N.B., Marti E. and Gomez R. (1998) Hierarchy, spatial configuration, and nutrient cycling in streams. *Australian Journal of Ecology*, **23**, 1–52.

Fisher S.G., Sponseller R.A. and Heffernan J.B. (2004) Horizons in stream biogeochemistry: flowpaths to progress. *Ecology*, **85**, 2369–2379.

Forbes S.A. (1887) The lake as a microcosm. *Bulletin Science Association of Peoria, IL*, **188**, 77–87.

Forel F.A. (1892) *Lac Leman: Monographie Limnologique*, Lausanne: Rouge.

Fritz S.C., Baker P.A., Lowenstein T.K., Seltzer G.O., Rigsby C.A., Dwyer G.S., Tapia P.M., Arnold K.K., Ku T.-L. and Luo S. (2004) Hydrologic variation during the last 170 000 years in the southern hemisphere tropics of the South America. *Quaternary Research*, **61**, 95–104.

Fritz K.M. and Dodds W.K. (2004) Resistance and resilience of macroinvertebrate assemblages to drying and flood in a tallgrass prairie stream system. *Hydrobiologia*, **527**, 99–112.

Fritz S.C., Ito E., Yu Z., Laird K.R. and Engstrom D.R. (2000) Hydrologic variation in the northern Great Plains during the last two millennia. *Quaternary Research*, **53**, 175–184.

Fry B. (2002) Stable isotopic indicator of habitat use by Mississippi River fish. *Journal of the North American Benthological Society*, **21**, 676–685.

Fry B. and Allen Y.C. (2003) Stable isotopes in zebra mussels as bioindicators of river-watershed linkages. *River Research and Applications*, **19**, 683–696.

Gjerlov C., Hildrew A.G. and Jones J.I. (2003) Mobility of stream invertebrates in relation to disturbance and refugia: a test of habitat templet theory. *Journal of the North American Benthological Society*, **22**, 207–223.

Gergel S.E., Turner M.G., Miller J.R., Melack J.M. and Stanley E.H. (2002) Landscape indicators of human impacts to riverine systems. *Aquatic Sciences*, **64**, 118–128.

Gido K.B., Schaefer J.F. and Pigg J. (2004) Patterns of fish invasions in the Great Plains of North America. *Biological Conservation*, **118**, 121–131.

Giller P.S., Covich A.P., Ewel K.C., Hall R.O. Jr and Merritt D.M. (2004b) Vulnerability and management of ecological services in freshwater systems. In *Sustaining Biodiversity and Ecosystem Services in Soils and Sediments*, Wall D. (Ed.), Island Press: Washington, pp. 137–159.

Giller P.S., Hillebrand H., Berninger U.G., Gessner M.O., Hawkins S., Inchausti P., Inglis C., Leslie H., Malmqvist B., Monaghan M.T., *et al.* (2004a) Biodiversity effects on ecosystem functioning: emerging issues and their experimental test in aquatic environments. *Oikos*, **104**, 423–436.

Gomi T., Sidle R.C. and Richardson J.S. (2002) Understanding processes and downstream linkages in headwater streams. *Bioscience*, **52**, 905–916.

Gotelli N.J. and Taylor C.M. (1999) Testing metapopulation models with stream-fish assemblages. *Evolutionary Ecology Research*, **1**, 835–845.

Grimm N.B., Gergel S.E., McDowell W.H., Boyer E.W., Dent C.L., Groffman P., Hart S.C., Harvey J., Johnston C., Mayorga E., *et al.* (2003) Merging aquatic and terrestrial perspectives of nutrient biogeochemistry. *Oecologia*, **137**, 485–501.

Hall D.L., Willig M.R., Moorhead D.L., Sites R.W., Fish E.B. and Mollhagen T.R. (2004) Aquatic macroinvertebrate diversity of playa wetlands: The role of landscape and island biogeographic characteristics. *Wetlands*, **24**, 77–91.

Hamilton S.K., Sippel S.J. and Melack J.M. (2004) Seasonal inundation patterns in two large savanna floodplains of South America: The Llanos de Moxos (Bolivia) and the Llanos del Orinoco (Venezuela and Colombia). *Hydrological Processes*, **18**, 2103–2116.

Havel J.E., Shurin J.B. and Jones J.R. (2002) Estimating dispersal from patterns of spread: spatial and local control of lake invasions. *Ecology*, **83**, 3306–3318.

Hershey A.E., Gettel G.M., McDonald M.E., Miller M.C., Mooers H., O'Brien warningmW.J., Pastor J., Richards C. and Schuldt J.A. (1999) A geomorphic-trophic model for landscape control of arctic lake food webs. *Bioscience*, **49**, 887–897.

Hobbs K.A. (1999) Tracing origins and migrations of wildlife using stable isotopes. *Oecologia*, **120**, 314–326.

Hoeinghaus D.J., Layman C.A., Arrington D.A. and Winemiller K.O. (2003) Spatiotemporal variation in fish assemblage structure in tropical floodplain creeks. *Environmental Biology of Fishes*, **67**, 379–387.

Holeck K.T., Mills E.L., MacIsaac H.J., Dochoda M.R., Colautti R.I. and Ricciardi A. (2004) Bridging troubled waters: biological invasions, transoceanic shipping, and the Laurentian Great Lakes. *Bioscience*, **54**, 919–929.

Holt R.D. (2002) Food webs in space: on the interplay of dynamic instability and spatial processes. *Ecological Research*, **17**, 261–273.

Holt R.D. (2004) Implications of system openness for local community structure and ecosystem function. In *Food Webs at the Landscape Level*, Polis G.A., Power M.E. and Huxel G.R. (Eds.), University of Chicago Press: Chicago, pp. 96–114.

House W.A. (2003) Geochemical cycling of phosphorus in rivers. *Applied Geochemistry*, **18**, 739–748.

Hueftle S.J. and Stevens L.E. (2001) Experimental flood effects on the limnology of Lake Powell reservoir, southwestern USA. *Ecological Applications*, **11**, 644–656.

Hutchinson G.E. (1957) *A Treatise on Limnology, Vol. 1. Geography, Physics, and Chemistry*, Wiley and Sons: New York.

Hutchinson G.E. and Bowen V.T. (1947) A direct demonstration of the phosphorus cycle in a small lake. *Proceeding of the National Academy of Sciences, Washington*, **33**, 148–153.

Hutchinson G.E. and Bowen V.T. (1950) Limnological studies in Connecticut. IX. A quantitative radiochemical study of the phosphorus cycle in Linsley Pond. *Ecology*, **31**, 194–203.

Hynes H.B.N. (1975) The stream and its valley. *Verhandlungen Internationale Vereinigung für Theoretische und Angerwandte Limnologie*, **19**, 1–15.

Janauer G.A. (2000) Ecohydrology: fusing concepts and scales. *Ecological Engineering*, **16**, 9–16.

Jeffries M. (1994) Invertebrate communities and turnover in wetland ponds affected by drought. *Freshwater Biology*, **32**, 603–612.

Juday C. and Birge E.A. (1933) The transparency, the color and the specific conductance of the lake waters of northeastern Wisconsin. *Transactions of the Wisconsin Academy of Sciences, Arts and Letters*, **23**, 233–248.

Junk W.J. and Wantzen K.M. (2004) The Flood Pulse Concept: new aspects, approaches, and applications – an update. In *Proceedings of the Second International Symposium on the Management of Large Rivers for Fisheries: Volume 2*, Welcomme R.L. and Petr T. (Eds.), Food and Agriculture Organization and Mekong River Commission. FAO Regional Office for Asia and the Pacific: Bangkok, pp. 117–149, RAP Publication 2004/16.

Kiker C.F., Milon J.W. and Hodges A.W. (2001) Adaptive learning for science-based policy: the Everglades restoration. *Ecological Economics*, **37**, 403–416.

Kinner D.A. and Stallard R.F. (2004) Identifying storm flow pathways in a rainforest catchment using hydrological and geochemical modeling. *Hydrological Processes*, **18**, 2851–2875.

Kling G.W., Kipphut G.W., Miller M.M. and O'Brien W.J. (2000) Integration of lakes and streams in a landscape perspective: the importance of material processing on spatial patterns and temporal coherence. *Freshwater Biology*, **43**, 477–497.

Knapp R.A. (2005) Effects of nonnative fish and habitat characteristics on lentic herptofauna in Yosemite National Park, USA. *Biological Conservation*, **121**, 265–279.

Knowlton M.F. and Jones J.R. (2003) Connectivity influences temporal variables of limnological conditions in Missouri River scour lakes. *Lake and Reservoir Management*, **19**, 160–170.

Koel T.M. (2004) Spatial variation in fish species richness of the upper Mississippi River system. *Transactions of the American Fisheries Society*, **133**, 984–1003.

Kolar C.S. and Lodge D.M. (2002) Ecological predictions and risk assessment for alien fishes in North America. *Science*, **298**, 33–1236.

Kratz T.K., Benson B.J., Blood E.R., Cunningham G.L. and Dalhlgren R.A. (1991) The influence of landscape position on temporal variability in four North American ecosystems. *American Naturalist*, **138**, 355–378.

Kratz T.K., Webster K.E., Bowser C.J., Magnuson J.J. and Benson B.J. (1997) The influence of landscape position on lakes in northern Wisconsin. *Freshwater Biology*, **37**, 209–217.

Lake P.S. (2003) Ecological effects of perturbations by drought in flowing waters. *Freshwater Biology*, **48**, 1161–1172.

Lamouroux N., Poff N.L. and Angermeier P.L. (2002) Intercontinental convergence of stream fish community traits along geomorphic and hydraulic gradients. *Ecology*, **83**, 1792–1807.

Larned S.T., Kinzie R.A. III, Covich A.P. and Chong C.T. (2003) Detritus processing by endemic and non-native Hawaiian stream invertebrates: a microcosm study of species-specific effects. *Archiv fur Hydrobiologie*, **156**, 241–254.

Leibowitz S.G. (2003) Isolated wetlands and their function: an ecological perspective. *Wetlands*, **23**, 517–531.

Lewis W.M. Jr, Hamilton S.K., Rodriquez M.A., Sanders J.D. III and Lasi D.H. (2001) Ecological determinism on the Orinoco floodplain. *Bioscience*, **50**, 681–692.

Light T. (2003) Success and failure in a lotic crayfish invasion: the roles of hydrologic variability and habitat alteration. *Freshwater Biology*, **48**, 1886–1897.

Likens G.E. (Ed.) (1985) *An Ecosystem Approach to Aquatic Ecology: Mirror Lake and its Environment*, Springer-Verlag: New York.

Likens G.E. (2004) Some perspectives on long-term biogeochemical research from the Hubbard Brook ecosystem. *Ecology*, **85**, 2355–2362.

Lindeman R.L. (1942) The trophic dynamic aspect of ecology. *Ecology*, **23**, 399–418.

Louette G. and De Meester L. (2004) Rapid colonization of a newly created habitat by cladocerans and the initial build-up of a *Daphnia*-dominanted community. *Hydrobiologia*, **513**, 245–249.

Lundberg J.G., Kottelat M., Smith G.R., Stiassny M.L.J. and Gill A.C. (2000) So many fishes, so little time: an overview of recent icthyological discovery in continental waters. *Annals of the Missouri Botanical Garden*, **87**, 26–62.

Lytle D.A. and Poff N.L. (2004) Adaptation to natural flow regimes. *Trends in Ecology and Evolution*, **19**, 94–100.

MacKay R.J. (1992) Colonization by lotic macroinvertebrates – a review of processes and patterns. *Canadian Journal of Fisheries and Aquatic Sciences*, **49**, 617–628.

MacIsaac H.J., Borbely J.V.M., Muirhead J.R. and Graniero P.A. (2004) Backcasting and forecasting biological invasions of inland lakes. *Ecological Applications*, **14**, 773–783.

Magnuson J.J., Tonn W.M., Banerjee A., Tolvonen J., Sanchez O. and Rask M. (1998) Isolation vs. extinction in the assembly of fishes in small northern lakes. *Ecology*, **79**, 2941–2956.

Malard F., Tockner K., Dole-Oliver M.-J. and Ward J.V. (2002) A landscape perspective of surface-subsurface hydrological exchanges in river corridors. *Freshwater Biology*, **47**, 621–640.

Malmqvist B. (2002) Aquatic invertebrates in riverine landscapes. *Freshwater Biology*, **47**, 679–694.

Malmqvist B. and Rundle S. (2002) Threats to the running water ecosystems of the world. *Environmental Conservation*, **29**, 134–153.

March J.G., Benstead J.P., Pringle C.M. and Scatena F.N. (2003) Damming tropical island streams: problems, solutions, and alternatives. *Bioscience*, **53**, 1060–1078.

Matthews W.J. and Robinson H.W. (1998) Influence of drainage connectivity, drainage area and regional species richness of

fishes of the interior highlands in Arkansas. *American Midland Naturalist*, **139**, 1–19.

McClain M.E., Boyer E.W., Dent C.L., Gergel S.E., Grimm N., Goffman P.M., Hart S.C., Harvey J.W., Johnston C.A., Mayorga E., *et al.* (2003) Biogeochemical hot spots and hot moments at the interface of terrestrial and aquatic ecosystems. *Ecosystem*, **6**, 301–312.

Merigoux S. and Doledec S. (2004) Hydraulic requirements of stream communities: a case study of invertebrates. *Freshwater Biology*, **49**, 600–613.

Meyer J.L. and Wallace J.B. (2001) Lost linkages and lotic ecology: rediscovering small streams. In *Ecology: Achievement and Challenge*, Chap. 14, Press M., Huntly N. and Levin S. (Eds.), Blackwell Science: Oxford, pp. 295–317.

Meybeck M. (2004) The global change of continental aquatic systems: dominant impacts of human activities. *Water Science and Technology*, **49**, 73–83.

Montgomery D.R. (1999) Process domains and the river continuum. *Journal of the American Water Resources Association*, **35**, 397–410.

Morin P.J. and McGrady-Steed J. (2004) Biodiversity and ecosystem functioning in aquatic microbial systems: a new analysis of temporal variation and species richness, predictability relations. *Oikos*, **104**, 458–466.

Mouquet N. and Loreau M. (2003) Community patterns in source-sink metacommunities. *American Naturalist*, **162**, 544–557.

Naeem S. and Wright J.P. (2003) Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. *Ecology Letters*, **6**, 567–579.

Nakamura F., Swanson F.J. and Wondzell S.M. (2000) Disturbance regimes of stream and riparian systems–a disturbance-cascade perspective. *Hydrological Processes*, **14**, 2849–2860.

Newbold J.D., Elwood J.W., O'Neill R.V. and Sheldon A.L. (1983) Phosphorous dynamics in a woodland stream ecosystem – a study of nutrient spiraling. *Ecology*, **64**, 1249–1265.

Odum H.T. (1957) Trophic structure and productivity of Silver Springs, Florida. *Ecological Monographs*, **27**, 55–112.

Odum E.P. (1969) The strategy of ecosystem development. *Science*, **164**, 262–270.

Olden J.D., Jackson D.A. and Peres-Neto P.R. (2001) Spatial isolation and fish communities in drainage lakes. *Oecologia*, **127**, 572–585.

Osborne L.L. and Wiley M.J. (1992) Influence of tributary spatial position on the structure of warm water fish communities. *Canadian Journal of Fisheries and Aquatic Sciences*, **49**, 671–681.

Osidele O.O. and Beck M.B. (2004) Food web modeling for investigating ecosystem behaviour in large reservoirs of the south-eastern United States: lessons from Lake Lanier, Georgia. *Ecological Modelling*, **173**, 129–158.

Osterkamp W.R. (2002) Geoindicators for river and river-valley monitoring in the humid tropics. *Environmental Geology*, **42**, 725–735.

Padilla D.K. and Williams S.L. (2004) Beyond ballast water: aquarium and ornamental trades as sources of invasive species in aquatic ecosystems. *Frontiers in Ecology and the Environment*, **2**, 131–138.

Parsons M., Thoms M.C. and Norris R.H. (2003) Scales of macroinvertebrate distribution in relation to the hierarchical organizational of rivers systems. *Journal of the North American Benthological Society*, **22**, 105–122.

Patten D.T., Harpman D.A., Volta M.I. and Randle T.J. (2001) A managed flood on the Colorado River: background, objectives, design, and implementation. *Ecological Applications*, **11**, 635–643.

Paul M.J. and Meyer J.L. (2001) Streams in the urban landscape. *Annual Review of Ecology and Systematics*, **32**, 333–365.

Pegg M.A., Lemke A.M. and Stoeckel J.A. (2002) Establishment of bighead carp in an Illinois River floodplain lake: a potential source population for the Illinois River. *Journal of Freshwater Ecology*, **17**, 161–163.

Perry J.A. and Schaeffer J.A. (1987) The longitudinal distribution of riverine benthos: a river discontinuum? *Hydrobiologia*, **148**, 257–268.

Pilliod D.S. and Peterson C.R. (2001) Local and landscape effects of introduced trout on amphibians in historically fishes watersheds. *Ecosystems*, **4**, 322–333.

Poff N.L. (1997) Landscape filter and species traits: towards mechanistic understanding and prediction in stream ecology. *Journal of the North American Benthological Society*, **16**, 391–409.

Poff N.L., Allan J.D., Bain M.B., Karr J.R., Prestegaard K.L., Richter B.D., Sparks R.E. and Stromberg J.C. (1997) The natural flow regime: a paradigm for river conservation and restoration. *Bioscience*, **47**, 769–784.

Poiani K.A., Richter B.D., Anderson M.G. and Richter H.E. (2000) Biodiversity conservation at multiple scales. Functional sites, landscapes, and networks. *Bioscience*, **50**, 133–146.

Poole G.C. (2002) Fluvial landscape ecology: addressing uniqueness within the river discontinuum. *Freshwater Biology*, **47**, 641–660.

Poole G.C., Stanford J.A., Running S.W., Frissell C.A., Woessner W.W. and Ellis B.K. (2004) A patch hierarchy approach to modeling surface and subsurface hydrology in complex floodplain environments. *Earth Surface Processes and Landforms*, **29**, 1259–1274.

Power M.E. and Dietrich W.E. (2002) Food webs in river networks. *Ecological Research*, **17**, 451–471.

Pringle C. (2003a) The need for a more predictive understanding of hydrologic connectivity. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **13**, 467–471.

Pringle C.M. (2003b) Interacting effects of altered hydrology and contaminant transport: emerging ecological patterns of global concern. In *Achieving Sustainable Freshwater Systems: A Web of Connections*, Holland M., Blood E. and Shaffer L. (Eds.), Island Press: Washington, pp. 85–107.

Pringle C.M. (2003c) What is hydrologic connectivity and why is it ecologically important? *Hydrological Processes*, **17**, 2685–2689.

Pulliam R.H. (1988) Sources, sinks, and population regulation. *American Naturalist*, **132**, 652–661.

Prosser I.P., Rutherfurd I.D., Olley J.M., Young W.J., Wallbrink P.J. and Moran C.J. (2001) Large-scale patterns of erosion

and sediment transport in river networks, with examples from Australia. *Marine and Freshwater Research*, **52**, 81–99.

Rassmussen J.B. and Vander Zanden M.J. (2004) The variation of lake food webs across the landscape and its effect on contaminant dynamics. In *Food Webs at the Landscape Level*, Polis G.A., Power M.E. and Huxel G.R. (Eds.), University of Chicago Press: Chicago, pp. 169–182.

Redfield A.C. (1958) The biological control of chemical factors in the environment. *American Scientist*, **46**, 205–221.

Resh V.H., Brown A.V., Covich A.P., Gurtz M.E., Li H.W., Minshall G.W., Reice S.R., Sheldon A.L., Wallace J.B. and Wissmar R.C. (1988) The role of disturbance in stream ecology. *Journal of the North American Benthological Society*, **7**, 433–455.

Rice S.P., Greenwood M.T. and Joyce C.B. (2001a) Tributaries, sediment sources, and the longitudinal organization of macroinvertebrate fauna along river systems. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 824–840.

Rice S.P., Greenwood M.T. and Joyce C.B. (2001b) Macroinvertebrate community changes at coarse sediments recruitment points along two gravel bed rivers. *Water Resources Research*, **37**, 2793–2803.

Richards K., Brasington J. and Hughes F. (2002) Geomorphic threshold in riverine landscapes. *Freshwater Biology*, **47**, 541–557.

Richardson J.S. and Mackay R.J. (1991) Lake outlets and the distribution of filter feeders- an assessment of hypotheses. *Oikos*, **62**, 370–380.

Riera J.L., Magnuson J.J., Kratz T.K. and Webster K.E. (2000) A geomorphic template for the analysis of lake districts applied to the Northern Highland Lake District, Wisconsin *Freshwater Biology*, **43**, 301–318.

Robinson C.T. and Burgher P. (1999) Seasonal disturbance of a lake outlet benthic community. *Archiv fur Hydrobiologie*, **145**, 297–315.

Robinson C.T., Uehlinger U. and Monaghan M.T. (2004) Stream ecosystem response to multiple experimental floods from a reservoir. *River Research and Applications*, **20**, 359–377.

Rodriguez-Iturbe I. (2000) Ecohydrology: a hydrologic perspective of climate-soil-vegetation dynamics. *Water Resources Research*, **23**, 349–357.

Sax D.F. and Brown J.H. (2000) The paradox of invasion. *Global Ecology and Biogeography*, **9**, 363–371.

Scheerer P.D. (2002) Implications of floodplain isolation and connectivity on the conservation of an endangered minnow, Oregon chub, in the Williamette River, Oregon. *Transactions of the American Fisheries Society*, **131**, 1070–1080.

Schindler D.E. and Scheuerell M.D. (2002) Habitat coupling in lake ecosystems. *Oikos*, **98**, 177–189.

Schlosser I.J. (1991) Stream fish ecology: a landscape perspective. *Bioscience*, **41**, 704–712.

Schmidt J.C., Parnell R.A., Grams P.E., Hazel J.E., Kaplinski M.A., Stevens L.E. and Hoffnagle T.L. (2001) The 1996 controlled flood in Grand Canyon: Flow, sediment transport, and geomorphic change. *Ecological Applications*, **11**, 657–671.

Schneider D.W. and Frost T.M. (1996) Habitat duration and community structure in temporary ponds. *Journal of the North American Benthological Society*, **15**, 64–86.

Shapiro J. (1990) Biomanipulation: the next phase- making it stable. In *Biomanipulation-Tool for Water Management*, Gulati R.D., Lammens E.H.R.R., Meijer M.L. and Donk E. (Eds.), Kluwer Academic Publishers: pp. 13–27.

Smith V.H. (1998) Cultural eutrophication of inland, estuarine, and coastal waters. In *Succession, Limitations, and Frontiers in Ecosystem Science*, Pace M.L. and Groffman P.M. (Eds.), Springer-Verlag: New York, pp. 7–49.

Smith G.C., Covich A.P. and Brasher A.M.D. (2003) An ecological perspective on the biodiversity of tropical island streams. *Bioscience*, **53**, 1048–1051.

Spencer M., Blaustein L., Schwartz S.S. and Cohen J.E. (1999) Species richness and the proportion of predatory animal species in temporary freshwater ponds: relationships with habitat size and permanence. *Ecology Letters*, **2**, 157–166.

Sorrano P.A., Webster K.E., Riera J.L., Kratz T.K., Bason J.S., Bukaveckas P.A., Kling G.W., White D.S., Caine N., Lathrop R.C., *et al.* (1999) Spatial variation among lakes within landscapes: ecological organization along lake chains. *Ecosystems*, **2**, 395–410.

Stanley E.H., Fisher S.G. and Grimm N.B. (1997) Ecosystem expansion and contraction in streams. *Bioscience*, **47**, 427–435.

Stanley E.H., Fisher S.G. and Jones J.B. Jr (2004) Effects of water loss on primary production: a landscape-scale model. *Aquatic Sciences*, **66**, 130–138.

Stead T.K., Schmid-Araya J.M. and Hildrew A.G. (2004) The contribution of subsurface invertebrates to benthic density and biomass in a gravel stream. *Archiv fur Hydrobiologie*, **160**, 171–191.

Steinman A.D. and Rosen B.H. (2000) Lotic-lentic linkages associated with Lake Okeechobee, Florida. *Journal of the North American Benthological Society*, **19**, 733–741.

Sterner R.S. and Elser J.J. (2002) *Ecological Stoichiometry*, Princeton University Press: Princeton.

Swanson F., Kratz T.K., Caine N. and Woodmansee R.G. (1988) Landform effects on ecosystem patterns and processes. *Bioscience*, **38**, 92–98.

Thorp J.H. and Delong M.D. (1994) The riverine productivity model: an heuristic view of carbon sources and organic processing in large river ecosystems. *Oikos*, **70**, 305–308.

Thomas S.A., Valett H.M., Mulholland P.J., Fellows C.S., Webstrer J.R., Dahm C.N. and Peterson C.G. (2001) Nitrogen retention in headwater streams: the influence of groundwater-surface water exchange. *The Scientific World*, **1**(S2), 623–631.

Thoms M.C. (2003) Floodplain-river ecosystems: lateral connections and the implications of human interference. *Geomorphology*, **56**, 335–349.

Thoms M.C. and Parsons M. (2003) Eco-geomorphology: an interdisciplinary approach to river science. *International Association of Hydrological Sciences*, **276**, 113–120.

Tockner K., Malard F. and Ward J.V. (2000) An extension of the flood pulse concept. *Hydrological Processes*, **14**, 2861–2883.

Tockner K. and Stanford J.A. (2002) Riverine flood plains: present state and future trends. *Environmental Conservation*, **29**, 308–330.

Tonn W.M. (1990) Climate change and fish communities: a conceptual framework. *Transactions of the American Fisheries Society*, **119**, 337–352.

Townsend C.R., Doleddec S. and Scarsbrook M.R. (1997) Species traits in relationship to temporal and spatial heterogeneity in streams: a test of habitat templet theory. *Freshwater Biology*, **37**, 367–387.

Turner B.J., Stallard R.F. and Brantley S.L. (2003) Investigation of in situ weathering of quartz diorite bedrock in the Rio Icacos basin, Luquillo Experimental Forest, Puerto Rico. *Chemical Geology*, **202**, 313–341.

Vadeboncoeur Y., Vander Zanden M.J. and Lodge D.M. (2002) Putting the lake back together: reintegrating benthic pathways into lake food web models. *Bioscience*, **52**, 44–54.

Vanni M.J. (2002) Nutrient cycling by animals in freshwater ecosystems. *Annual Review of Ecology and Systematics*, **33**, 341–370.

Vanni M.J. and Headworth J.L. (2004) Cross-habitat transport of nutrients by omnivorous fish along a productivity gradient: integrating watersheds and reservoir food webs. In *Food Webs at the Landscape Level*, Polis G.A., Power M.E. and Huxel G.R. (Eds.), University of Chicago Press: Chicago, pp. 43–61.

Vannote R., Minshall G.W., Cummins K.W., Sedell J.R. and Cushing C.E. (1980) The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 130–137.

Vredenburg V.T. (2004) Reversing introduced species effects: experimental removal of introduced fish leads to rapid recovery of a declining frog. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7646–7650.

Wallace J.B. and Hutchens J.J. (2000) Effects of invertebrates in lotic ecosystem process. In *Invertebrates as Webmasters in Ecosystems*, Coleman D.C. and Hendrix P.F. (Eds.), CABI Publishing: Oxon, pp. 73–96.

Walters D.M., Liegh D.S. and Bearden A.B. (2003) Urbanization, sedimentation, and the homogenization of fish assemblages in the Etowah River Basin, USA. *Hydrobiologia*, **494**, 5–10.

Ward J.V. and Stanford J.A. (1995) Ecological connectivity in alluvial river ecosystems and its disruption by flow regulation. *Regulated Rivers*, **11**, 105–119.

Ward J.V., Tockner K., Arscot D.B. and Claret C. (2002) Riverine landscape diversity. *Freshwater Biology*, **47**, 517–539.

Warne A.G., Toth L.A. and White W.A. (2000) Drainage-basin-scale geomorphic analysis to determine reference conditions for ecological restoration- Kissimmee River, Florida. *Geological Society of America Bulletin*, **112**, 884–899.

Waters T.F. (1995) *Sediment in Streams: Sources, Biological Effects, and Control*, Monograph 7, American Fisheries Society: Bethesda.

Webster K.E., Sorrano P.A., Baines S.B., Kratz T.R., Bower C.J., Dillon P.J., Cambell P., Fee E.J. and Hecky R.E. (2000) Structuring features of lake districts: geomorphic and landscape controls on lake chemical responses to drought. *Freshwater Biology*, **43**, 499–515.

Wetzel R.G. (2001) *Limnology: Lake and River Ecosystems*, Academic Press: San Diego.

Whigham D.F. and Jordan T.E. (2003) Isolated wetlands and water quality. *Wetlands*, **23**, 541–549.

Whiles M.R. and Dodds W.K. (2002) Relationships between stream size, suspended particles, and filter-feeding macroinvertebrates in a Great Plains drainage network. *Journal of Environmental Quality*, **31**, 1589–1600.

Wilson A.B., Glaubrecht M. and Meyer A. (2004) Ancient lakes as evolutionary reservoirs: evidence from the thalassoid gastropods of Lake Tanganyika. *Proceedings of the Royal Society of London series. B-Biological Sciences*, **271**, 529–536.

Winter T.C. (2001) The concept of hydrologic landscapes. *Journal of American Water Resources Association*, **37**, 335–349.

Winter T.C. and LaBaugh J.W. (2003) Hydrologic considerations in defining isolated wetlands. *Wetlands*, **23**, 532–540.

Wissinger S.A. (1999) Ecology of wetland invertebrates: synthesis and applications for conservation and management. In *Invertebrates in Freshwater Wetlands of North America*, Batzer D.P., Rader R.D. and Wissinger S.A. (Eds.), John Wiley and Sons: New York, pp. 1043–1086.

Woodward G. and Hildrew A.G. (2002) Food web structure in riverine landscapes. *Freshwater Biology*, **47**, 777–798.

Wright J.P. and Flecker A.S. (2004) Deforesting the riverscape: the effects of wood on fish diversity in a Venezuelan piedmont stream. *Biological Conservation*, **120**, 439–447.

102: Trophic Dynamics

CORY C CLEVELAND, ALAN R TOWNSEND AND DIANE M MCKNIGHT

Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, US

The carbon (C) and water cycles are intimately linked in terrestrial ecosystems. Thus, an understanding of the processes regulating transfers of water in terrestrial ecosystems requires an understanding of the carbon cycle, and in particular, the factors constraining carbon movement in the soil-plant-atmosphere continuum and through trophic levels in ecosystems. The linkages between the C and the water cycles are mediated primarily through biological processes, and are bidirectional in nature. For example, precipitation (and hence ecosystem water availability) strongly regulates plant growth and biogeochemical cycling in soils. Subsequently, plant growth and soil biogeochemistry strongly influence evaporation and atmospheric water vapor (and hence precipitation). Plant growth and soil processes are also cyclically linked. Thus, plant and soil interactions can have important implications for water cycling. However, while major climatic variables (including precipitation) may drive biological patterns and processes at large scales, other ecological interactions also regulate both plant and soil processes. An appreciation of these ecological factors is important to understanding the relationship between C and water, and to predicting how global environmental change is likely to affect the interactions between the C and water cycles.

INTRODUCTION

From an ecological perspective, plant photosynthesis regulates the carbon balance and productivity of the biosphere, controls fluxes of C between the biosphere and the atmosphere, and is responsible for virtually all of the biochemical production of organic matter. From an anthropomorphic perspective, biomass produced by autotrophic organisms via photosynthesis is the source of all food, fiber, and fuel in the biosphere, and thus sustains all other life on earth, including humans. Of all the carbon fixed by plants (gross primary production; GPP), some is stored and some is respired. The fraction of C that is fixed but not respired represents net primary production (NPP). Much of the biomass produced by plants is consumed by individuals of higher trophic levels, or moves into soil where it provides substrate for heterotrophic organisms (net ecosystem production; NEP). Inputs of organic matter from plants to soil influence nutrient cycling and soil water retention, and therefore affect further plant production. Understanding the ecological controls on organic matter cycling in terrestrial ecosystems is fundamental to our understanding of ecosystem water balance, both at present and in the future.

PRIMARY PRODUCTION

Gross Primary Production

In 1942, Lindeman outlined the fundamental ecological concepts of energy flow in ecosystems. At any trophic level, from producer to consumer, energy flow is mediated through the individual organism. Energy is consumed, some is lost as feces, urine, or gas, and part is assimilated and respired or used for the production and growth of new biomass. During each transfer from a lower to a higher trophic level, ~10% of the consumed biomass is directly converted to new biomass; the balance is respired (Smith, 1996). Primary plant production is conceptually similar. Photosynthetically active radiation (PAR; radiation in the 400–700 nm wave band in the visible light spectrum that is utilizable by plants during photosynthesis) may be intercepted by a plant and used to drive photosynthesis. Ultimately, only a tiny fraction of total available light energy is used to convert inorganic carbon to organic molecules

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

within plants; the majority is lost as long-wave radiation and through sensible and latent heat fluxes. However, this small fraction is sufficient to drive plant production that sustains all organisms occupying higher trophic levels. The total organic carbon produced during plant photosynthesis is known as *gross primary production* (*GPP*).

Net Primary Production

The concept of energy flow through ecological trophic levels suggests that not all C fixed by plants during photosynthesis is allocated to growth and biomass production (Figure 1). Plant respiration is also necessary to maintain the energy demands of biomass autotrophic organisms.

Plant respiration, which involves the mitochondrial oxidation of carbohydrates to form ATP, the most important energetic molecule in cells, is the necessary cost of maintaining nonphotosynthetic plant biomass (at all times) and photosynthetic leaves (at night; Chapin *et al.*, 2002); roughly half of the C fixed by any given individual plant during GPP is respired directly by plants as carbon dioxide (CO_2) back into the atmosphere (Figure 1). The difference between total plant C fixed (GPP) and the C lost through plant respiration represents plant net primary production (NPP). Odum (1971) defined NPP as "the rate of storage of organic matter in plant tissues in excess of the respiratory utilization by plants". By this definition, NPP includes increases in plant biomass (e.g. growth of roots, shoots, and leaves), incidental losses of fixed C through roots (i.e. root exudation), transfers of fixed C to symbiotic or associated microorganisms (e.g. mycorrhizal fungi and rhizosphere bacteria), and plant production and losses of biogenic hydrocarbons. In Lindeman's energy transfer model, NPP also represents the amount of photosynthetically fixed C that is available to the first heterotrophic level in an ecosystem (Lindeman, 1942).

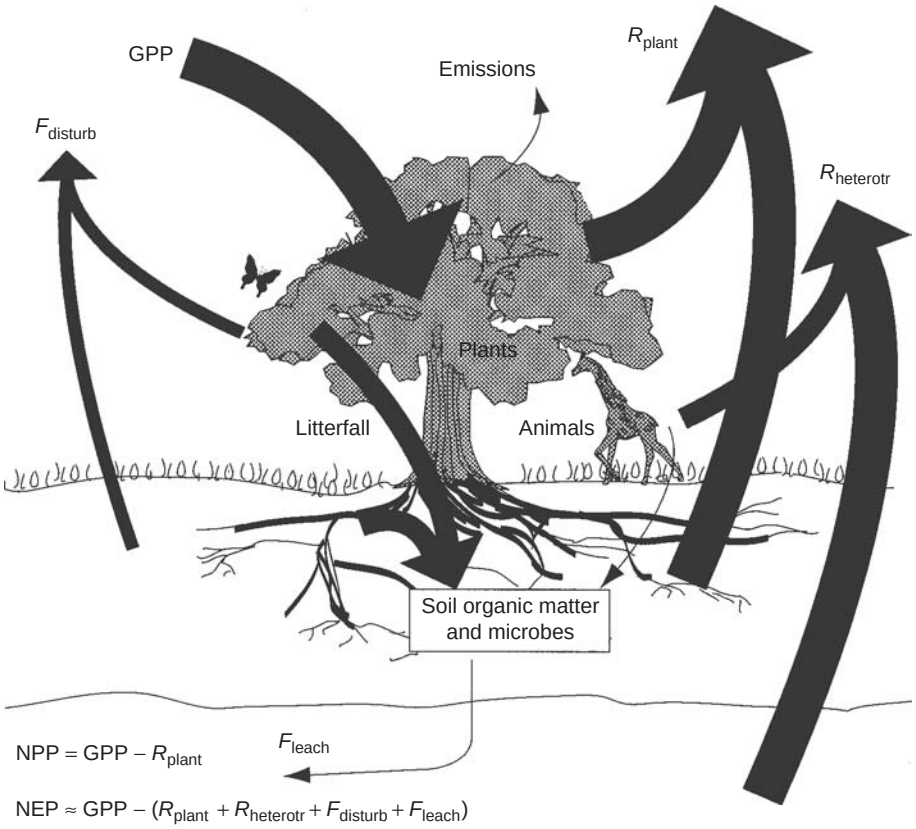


Figure 1 Overview of the major carbon fluxes in an ecosystem. Carbon enters an ecosystem as gross primary production (GPP) through plant photosynthesis. Roots and aboveground plant parts return roughly half of this carbon back to the atmosphere as plant respiration (R_{plant}). Net primary production (NPP) is the difference between GPP and R_{plant}. Most of NPP is transferred to soil organic matter as litterfall, root death, root exudation, and root transfers to symbionts; some NPP is lost to higher trophic levels (e.g. animals). Most carbon entering the soil is lost through microbial respiration, which together with animal respiration, is called *heterotrophic respiration* (R_{hetero}). Additional carbon is lost from soils through leaching and disturbance. Net ecosystem production (NEP) is the net carbon accumulated by an ecosystem; it equals the carbon inputs from GPP minus the various avenues of carbon loss (Reproduced from Chapin *et al.*, 2002 by permission of Springer)

Table 1 Productivity of different ecosystems per day and per unit leaf area

Biome	Season length (d)	Daily NPP per ground area ($g\,m^{-2}\,d^{-1}$)	Total LAI ($m^2\,m^{-2}$)	Daily NPP per leaf area
Tropical forests	365	6.8	6.0	1.14
Temperate forests	250	6.2	6.0	1.03
Boreal forests	150	2.5	3.5	0.72
Mediterranean shrublands	200	5.0	2.0	2.50
Tropical savannas and grasslands	200	5.4	5.0	1.08
Temperate grasslands	150	5.0	3.5	1.43
Deserts	100	2.5	1.0	2.50
Arctic tundra	100	1.8	1.0	1.80
Crops	200	3.1	4.0	0.76

Data from Gower, 2002.

Global Distribution of NPP

NPP is usually measured at the scale of the ecosystem and is expressed in grams of biomass or C per square meter per year (Sala and Austin, 2000). An accurate estimate of NPP represents one of the most fundamentally important characteristics of an ecosystem, and as such, estimates of NPP both within and between ecosystems are plentiful (but see Clark *et al.*, 2001a for discussion of problems in estimating NPP). On the global scale, NPP is spatially heterogeneous, but growing season length is the most important factor explaining differences in biome-specific NPP (Chapin *et al.*, 2002). Like most enzymatic biological processes, photosynthesis is sensitive to extremes in temperature and moisture, and many ecosystems experience times that are too dry, too wet, too cold, or too hot to allow photosynthesis and plant growth to occur. However, when biome-level estimates of NPP are adjusted for growing season length, NPP in different forested ecosystems are nearly identical ($\sim 5\,g\,m^{-2}$ $year^{-1}$), and NPP in the most productive systems (tropical rain forest ecosystems) is only three times NPP in the least productive systems (desert ecosystems; Table 1). These calculations suggest the importance of growing season length in controlling rates of NPP on a global scale (Chapin *et al.*, 2002).

The Fate of NPP: Plant Allocation

What is the fate of NPP? Plant allocation of the products of NPP varies in both space and time, and from species to species. For example, climatic variability over the range of a single plant species influences the timing of C allocation for leaf bud production and foliar growth, leaf senescence, and flowering and fruiting times, among others. Similarly, allocation patterns in deciduous tree species that annually grow new leaves differ from allocation patterns in evergreen species in an area experiencing identical climatic conditions (e.g. aspens in the spruce-fir ecotone). Thus, while generalizations about plant C allocation are difficult, in general, plants allocate carbon to minimize limitation by any single resource (Chapin *et al.*, 2002). The relative demand for different resources regulates the direction and flow of carbon through plants (Aber and Melillo, 2001). Plants allocate resources to roots, shoots, or leaves depending on the relative availability of aboveground and belowground resources. When belowground resources (e.g. water or nutrients) are most limiting to plant production, plants may allocate a greater percentage of C to the production and maintenance of belowground biomass. For example, in deserts, water is often the most limiting resource, and light availability is high. Thus, many desert plants produce extensive root systems that include roots at the surface to capture episodic precipitation, and deep roots that effectively capture more consistent water supplies at depth (Chapin *et al.*, 2002). In contrast, plants growing in closed canopy, light limited forests may partition more C to aboveground tissues to maximize their ability to reach the canopy and capture available light (Table 2).

Abiotic Controls on NPP

As discussed, growing season length explains much of the variability in NPP between different ecosystems. What other factors influence rates of NPP? As the driver of the photosynthetic process, light availability strongly regulates ecosystem NPP. Early work by Montieth (1977) demonstrated that the productivity of well-watered and well-fertilized crop plants was linearly related to the amount of light they absorbed. This basic concept has much utility; it combines meteorological constraints of light impinging on a surface with the physiological constraints of light absorption and use by a leaf (Sala and Austin, 2000). Energy absorbed by a leaf (absorbed PAR; APAR) integrates seasonal and diurnal variation in sunlight and climate, and also implicitly includes the quantity of vegetation that is absorbing radiation, or the leaf area index (LAI; Sala and Austin, 2000). A conversion efficiency factor can then be used to convert APAR to growth, or biomass produced.

Table 2 Components of NPP in 12 old-growth tropical rain forest sites

Site	Aboveground biomass (Mg C/ha)	Component of aboveground NPP (Mg C/ha/y)			
		Fine litterfall	Losses to consumers	Volatile organic C	Estimated aboveground NPP
Ivory Coast: L'Anguédédou Forest	151.5	9.3	1.1	0.2	14.3
Thailand: Khaochong	167.0	5.9	0.7	0.2	9.9
USA: Hawaii (Puu Kolekole)	68.5	4.4	0.5	0.2	7.6
Columbia: Magdelena slope	162.9	4.4	0.5	0.3	7.5
Brazil: Egler Reserve	203.0	3.7	0.4	0.2	6.4
Puerto Rico: Colorado Forest	84.8	3.4	0.1	0.2	5.6
Venezuela: San Carlos	118.7	2.8	0.3	0.3	5.2
Jamaica: Blue Mountain Mull Ridge	156.0	2.8	0.3	0.2	5.0
India: Kagneri	230.0	2.0	0.2	0.3	3.9
Mexico: Chamela	40.0	1.7	0.4	0.2	3.2
Puerto Rico: Pico del Este	23.8	1.6	0.1	0.2	2.9
USA: Hawaii (Site 5)	61.5	0.9	0.1	0.2	1.4

Data from Clark *et al.*, 2001b.

Table 3 Primary production and biomass estimates in global terrestrial ecosystems

Ecosystem type	Area (10^6 km)	Mean NPP (g C/m^2/year)	Mean biomass (Kg C/m^2)
Tropical rain forest	17.0	900	20
Tropical seasonal forest	7.5	675	16
Temperate evergreen forest	5.0	585	16
Temperate deciduous forest	7.0	540	13.5
Boreal forest	12.0	360	9.0
Woodland, shrubland	8.0	270	2.7
Savanna	15.0	315	1.8
Temperate grassland	9.0	225	0.7
Tundra	8.0	65	0.3
Desert scrub	18.0	32	0.3
Rock, ice, sand	24.0	1.5	0.01
Cultivated land	14.0	290	0.5
Swamp, marsh	2.0	1125	6.8
Lake, stream	2.5	225	0.01

Data from Whittaker and Likens, 1973.

Because APAR can be used to generate estimates of NPP on large scales, combined with the ability to obtain values of APAR using remote sensing, estimates of APAR measured from space are routinely used to generate estimates of plant production.

Large-scale (e.g. biome- to global-scale) patterns in terrestrial primary productivity can also be well explained by climatic variables, most notably average annual temperature and precipitation. In general, primary production is high in ecosystems with warm, moist climates, and low in ecosystems characterized by cold, dry climates (Table 3). Analyses of the empirical relationships between NPP and climate (Schuur, 2003; Lieth, 1975) further validate this observation, and suggest the importance of temperature and precipitation in regulating NPP (Figure 2). Across a temperature gradient ranging from $<10\,^{\circ}$C to $\sim30\,^{\circ}$C, NPP increases linearly with increases in temperature (Schuur, 2003). However, precipitation is also highly correlated with ecosystem NPP (Figure 2). This strong positive relationship between ecosystem annual precipitation inputs and NPP represents the first link between C and water cycles depicted in Figure 3.

The relationship between NPP and precipitation has also been verified experimentally. Sala *et al.* (1988) found that mean annual precipitation explained 90% of the variability in mean aboveground NPP in 100 major land resource areas across the central grassland region of North America. These observations illustrate the overwhelming importance of water availability as a control on NPP, at least in some ecosystems. However, while increases in precipitation in

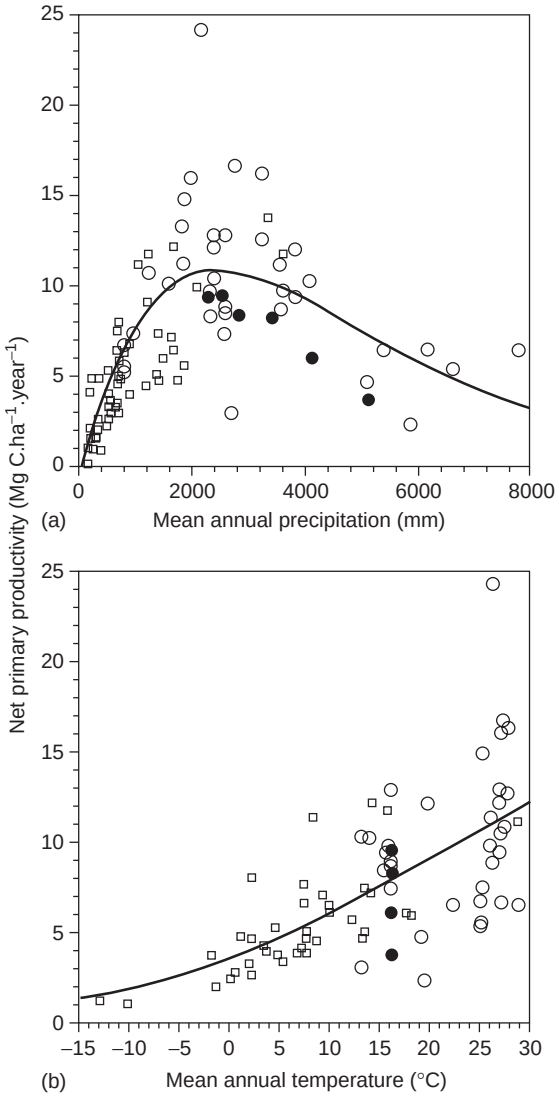


Figure 2 Relationships between net primary production and (a) mean annual precipitation and (b) mean annual temperature in temperate ecosystems (Reproduced from Schuur, 2003 by permission of Ecological Society of America)

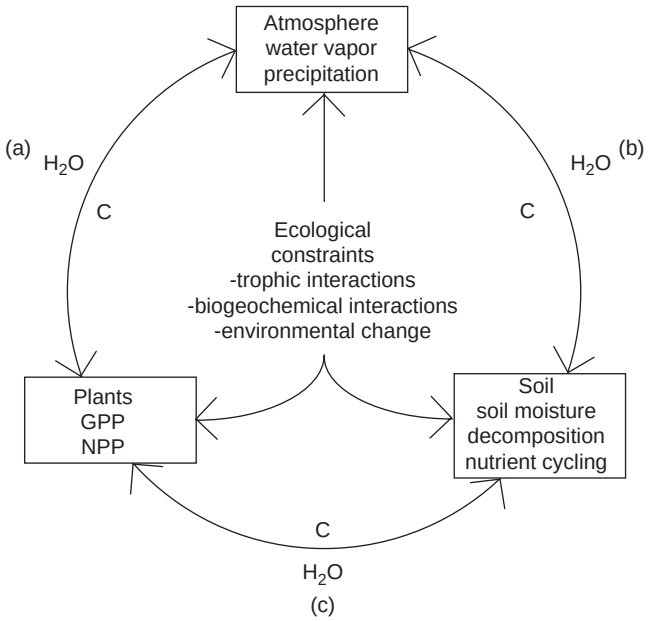


Figure 3 Schematic illustration of the coupling of the carbon and water cycles in terrestrial ecosystems. (a) NPP and evapotranspiration regulate the flow of water and C between plants and the atmosphere; (b) decomposition and evaporation control the flow of water and C between the soil and the atmosphere; and (c) biogeochemistry and soil processes regulate fluxes of C and water in the plant–soil continuum. However, ecological constraints also strongly regulate the fluxes of C and water in terrestrial ecosystems

many arid and semiarid ecosystems correlate with dramatic increases in NPP, in extremely mesic ecosystems, excessive precipitation inputs may actually correlate with decreases in NPP (Schuur, 2003). Together, these data suggest interesting interactions between precipitation and NPP. Namely, NPP increases with temperature and precipitation, but this effect diminishes in wet, warm ecosystems, where further increases in precipitation or temperature may not be balanced by increases in NPP (Figure 2). The fact that NPP decreases in ecosystems characterized by extremely high precipitation inputs suggests that excessive water availability either directly or indirectly affects plant growth through other feedback mechanisms.

Evapotranspiration: The Return of Water to the Atmosphere

The emerging linkages between C and water cycling, which include the strong correlation between water availability and NPP, suggest that the relationship may also operate in reverse. In other words, high NPP may also affect atmospheric water content. Empirical evidence suggests a strong relationship between NPP and ecosystem water losses, and these losses, in turn, complete the cycle connecting plant C cycling with atmospheric water content (e.g. Schimel *et al.*, 1997). Respiratory losses of CO_2 are not the only resource costs of maintaining an autotrophic lifestyle; losses of water via evapotranspiration link primary production (C cycle) with the hydrological cycle (*see* **Chapter 42, Transpiration, Volume 1**). Losses of water from terrestrial ecosystems to the atmosphere occur via two pathways: evaporation, or the direct return of water to the atmosphere from open water bodies and from the land surface (mainly from rock and soil and plant surfaces); and transpiration, the incidental loss of water through plant leaf stomatal openings (*see* **Chapter 70, Transpiration and Root Water Uptake, Volume 2**). During photosynthesis, plants obtain atmospheric CO_2 for fixation in photosynthesis through their stomata, small pores in the leaf surface.

The size of the stomatal opening regulates stomatal conductance, and high stomatal conductance equates with high fluxes of CO_2 available for photosynthesis. However, as a consequence of high stomatal conductance, a decreasing moisture gradient from the inside of the leaf to the outside creates a flux of water out of the leaf, thus resulting in plant water loss to the atmosphere. Thus, in maximizing CO_2 uptake for photosynthesis, plants necessarily subject themselves to high water losses via transpiration.

Together, transpiration and evaporation are known as *evapotranspiration*, and rates of actual evapotranspiration (AET) are very well correlated with NPP (Figure 4). The high correlation between NPP and AET represents an important feedback loop connecting water and carbon cycling in ecosystems. High rates of precipitation and inputs of solar radiation (energy) stimulate high NPP. Subsequently, high rates of NPP in mesic ecosystems lead to high rates of water loss to the atmosphere via evapotranspiration, which have been shown to complete the feedback system by fueling precipitation into ecosystems. The importance of this feedback between NPP, soils, and precipitation is illustrated by an example from tropical rain forests. In Amazonia, the climate and vegetation are tightly coupled. High rates of evapotranspiration from Amazon forests provide water vapor to the atmosphere. Through convective processes, evaporated and transpired water vapor becomes cumulus clouds and precipitation (Nobre *et al.*, 1991). Much of the daily rainfall in the Amazon Basin is quickly reevaporated and transpired into the atmosphere, maintaining high atmospheric humidity and generating clouds that provide the following day's rain. Without forest vegetation, most rainfall would enter rivers, resulting in progressive drying of the air, and affecting NPP.

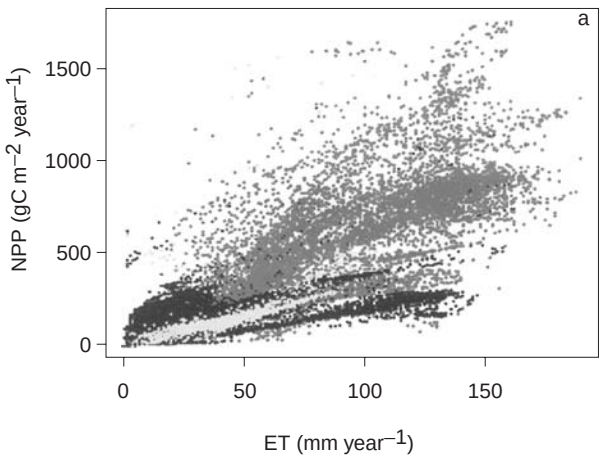


Figure 4 The relationship between NPP and ET ($r^2 = 0.71$) from an integration of the Century ecosystem model for the Northern Hemisphere. Points in dark grey are forest ecosystems, light grey indicates grasslands, and black indicates ''mixed'' ecosystems that include grasses and trees or shrubs (e.g. savannas) (From Schimel, D.S., Braswell, B.H. and Parton, W.J. (1997). Equilibration of the terrestrial water, nitrogen and carbon cycles. Proceedings of the National Academy of Sciences, 94, 8280–8283. ©1997 National Academy of Sciences USA)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Net Ecosystem Production

With respect to organic C balance, ecosystems can be aggrading (i.e. C accumulating in soils and vegetation), degrading (i.e. net loss of C), or in equilibrium (i.e. carbon inputs matched by carbon losses). The balance between carbon entering and leaving an ecosystem is net ecosystem productivity (NEP; also known as *net ecosystem exchange* [NEE]), and represents the difference between NPP and heterotrophic respiration by animals and microorganisms (Figure 1). Most carbon enters an ecosystem though NPP, and the majority of C that enters an ecosystem is eventually lost through respiration. However, in some ecosystems, leaching of dissolved organic carbon (DOC) or dissolved inorganic carbon (DIC) through soils and into aquatic ecosystems may be important loss vectors for C (Neff and Asner, 2001). This is particularly true in arctic terrestrial ecosystems, where roughly 20% of the CO_2 produced in soils is lost as CO_2 from lakes, streams, or groundwater (Chapin *et al.*, 2002). Losses of fixed C via methane or plant produced hydrocarbon emissions are also significant loss pathways for C in some ecosystems. For example, in wetland systems, or systems where anoxic conditions prevent the aerobic decomposition of organic matter, ecosystem losses of C *via* methanogenesis can be significant. Similarly, many plants have been shown to produce C-rich secondary compounds that are lost as volatile hydrocarbons from ecosystems. Volatile organic carbon (VOC) emissions (e.g. terpenoid and isoprenoid compounds from plants) may account for C losses up to 5% of NPP (Figure 1; Chapin *et al.*, 2002).

DECOMPOSITION

Soil Organic Matter Decomposition

In most ecosystems, the majority of the C fixed during NPP leaves the ecosystem through the activity of heterotrophic microorganisms (Figure 1). The most important conversions of ecosystem organic C to CO_2 (i.e. real losses) occur during heterotrophic respiration; microorganisms utilize dead plant biomass and detritus to respire and build biomass. Decomposition is a fundamental ecological and biogeochemical process; it returns fixed C to the atmosphere, and also returns critical nutrients stored in organic matter back to the soil, thus providing the main source of

annual plant nutrient availability (Paul and Clark, 1996) (*see* **Chapter 96, Nutrient Cycling, Volume 3**). Organic matter decomposition is also tightly linked to water inputs, and represents another important link between the C and the water cycles (Figure 3).

Following litterfall and litter accumulation on the soil surface, the first step in organic matter decomposition is leaching of soluble organic material through the litter (and other plant biomass) layer and into the soil. Soluble fluxes of organic C from throughfall (leaching from live plant leaves and other aboveground biomass) and through the fallen litter layer may be 1–19% of the total litterfall C flux, and may represent 1–5% of NPP (References in Neff and Asner, 2001). The rate at which C solubilizes, leaches, and enters the soil is linked to precipitation driving fluxes; the direct movement of water through the vegetation and litter layers drives C fluxes. Experimental evidence suggests that DOC concentrations decrease through the soil horizon, and that the majority of C that enters the soil is biologically available and is utilized and respired by microorganisms (Cleveland *et al.*, 2004).

Abiotic Controls on Decomposition

The decomposition of nonleached plant material accumulated as soil organic matter is strongly linked to climate. For example, temperature affects decomposition directly through its effects on soil microbial activity. Over a broad range, increasing temperature elicits exponential increases in soil respiration (Figure 5). Temperature affects both the physicochemical characteristics of the soil environment as well as the physiological reaction rates of cells. Specifically, in a basic way, microbial decomposition of organic matter must follow simple reaction kinetics, with reaction rates increasing with temperature to a maximum, and decreasing at excessive temperatures that degrade enzymes. Freeze–thaw cycles can also contribute to decomposition, through the physical effects of freezing on the microorganisms decomposing dead plant material and soil organic matter.

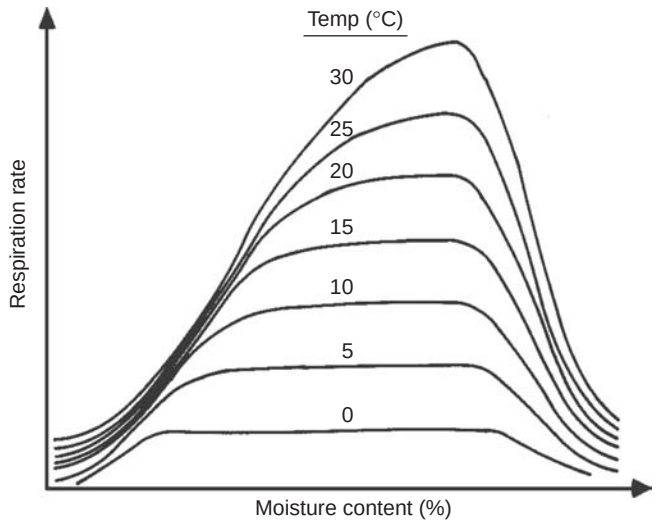


Figure 5 Semiperspective plot of computed soil respiration rate at different soil temperature and soil moisture levels (Reprinted from Paul and Clark, 1996, ©1996 with permission from Elsevier)

Soil moisture also strongly regulates decomposition. Like plants, decomposer organisms are most productive when water availability is high, provided sufficient oxygen is available in the soil (Paul and Clark, 1996). Microbial decomposition of organic matter is dependent upon diffusion processes in the soil environment. Excessive drying of soil prevents both enzymatic processes and physicochemical processes from allowing decomposition and nutrient mineralization to occur. Evidence suggests that decomposition declines at soil moisture values <30–50%, owing to the reduction in moisture necessary to allow rapid diffusion of necessary resources (Paul and Clark, 1996). Similarly, excess water in soils may also affect soil organic matter decomposition rates, but in the opposite direction. Excessive soil moisture can lead to soil anoxia, and decreased oxygen availability has been linked to decreased rates of decomposition (Schuur, 2003; Schuur *et al.*, 2001). The relationship between water availability and decomposition rate is another important link between the C and water cycles (Figure 3).

The decomposition rate of soil organic material is influenced by many factors. For example, soil temperature, soil moisture, organic matter quality, soil oxygen availability, and the activity of specific microorganisms all influence the rate of conversion of organic C to CO_2. However, as with NPP, an analysis of decomposition at the global scale is useful for determining the role of climate in regulating large-scale variations in decomposition. Meentemeyer (1978) demonstrated that while litter quality strongly correlates with decomposition rates, the slope of the relationship between litter quality and decomposition rate decreases coincidentally with ecosystem actual evapotranspiration (AET; Meentemeyer, 1978). While AET most appropriately represents water losses from a system, it is also a useful proxy for the amount of water (and energy) entering an ecosystem (Schimel *et al.*, 1997). Thus, the observed relationship between AET and decomposition integrates the effect of soil moisture and precipitation inputs on decomposition, and further demonstrates the linkage between precipitation and soil processes, including organic matter decomposition. *In situ* soil respiration, which includes plant root and soil microbial respiration, is also a useful proxy for heterotrophic decomposition. Like with NPP, at large scales, soil respiration is positively related to soil moisture availability (Raich and Schlesinger, 1992; Figure 5).

Precipitation exerts control on soil processes (including C decomposition) through AET and soil moisture, demonstrating another important linkage between the C cycle and water cycle. However, abundant precipitation not only affects soil processes, but also soil moisture availability, and this constitutes another feedback mechanism linking soil water availability, soil C cycling, and water losses to the atmosphere. Specifically, evaporative losses of water directly from soil are necessarily higher when soil moisture is higher (i.e. the difference between AET and potential evapotranspiration (PET) results from differences in soil moisture) (*see* **Chapter 42, Transpiration, Volume 1**). High precipitation in mesic ecosystems contributes to high soil moisture, and high soil moisture in ecosystems may perpetuate high precipitation through high rates of AET. Soil is a huge reservoir for water in terrestrial ecosystems, and evaporative processes from soil represent a significant portion of AET. Moreover, variable decomposition rates can affect the amount of organic matter accumulation in ecosystems. Because organic matter in soils often increases water-holding capacity, the linkage between water and C cycles in the atmosphere and soil is completed.

PLANT–SOIL INTERACTIONS

There are important linkages between water availability and both plant and soil processes, and factors affecting these interactions may lead to important feedbacks that further influence the fluxes of water and carbon into and out of ecosystems (Figure 3). However, there are also important interactions that occur between plants and soil, and these interactions affect and are affected by ecosystem C cycling and water status, and hence fluxes of water and C between terrestrial ecosystems and the atmosphere. For example, most inputs of organic matter into ecosystems are the result of primary production. While the fate of soil organic matter is strongly dependent on climatic (e.g. temperature and moisture) and edaphic (e.g. soil texture, soil chemistry) factors, ecosystems with high rates of NPP will have correspondingly high organic matter inputs to soil. Soil organic C decomposition, which is linked to the water cycle through its strong relationship to soil moisture, is also linked to NPP and inputs of organic carbon as the main source of carbon and energy for heterotrophic microorganisms. CO_2 flux from soil is highly correlated with NPP, which supplies organic molecules to decomposers. Across the world's major biomes, Raich and Schlesinger (1992) showed a direct relationship between soil respiration and NPP. Experimental manipulations of C also consistently reveal increases in soil respiration following C inputs (Schlesinger and Andrews, 2000). Thus, precipitation regulates NPP, NPP regulates decomposition and soil C, soil C regulates soil moisture, which in turn, may regulate precipitation *via* AET (Figure 3).

The interactions between NPP and soil processes also operate in reverse. Precipitation exerts control on soil organic matter decomposition, which in turn represents the major mechanism for nutrient mineralization and nutrient cycling in terrestrial ecosystems. On an annual basis, most of the nutrient demands by plants are met through the activity of soil microorganisms, which liberate important nutrients during soil organic matter (SOM) decomposition. Thus, soil moisture can indirectly affect NPP through decomposition and nutrient mineralization. High levels of NPP also provide increased C substrate and fuel higher levels of soil microbial biomass. In systems where nutrients are extremely rare, high levels of microbial biomass can positively affect nutrient cycling and plant growth (Paul and Clark, 1996).

Some interesting data suggest complexities of the interactions between C and water cycles, and provide evidence for the feedbacks depicted in Figure 3. For example, data from grasslands suggest fascinating interactions between NPP, soil processes, and water availability. First, when grassland sites spanning a precipitation gradient are compared, average NPP increases with precipitation (References in Chapin *et al.*, 2002). Moreover, in a single grassland site, NPP increases during wet years, and responds to experimental water addition, indicating that NPP in grasslands is water limited. However, the apparent water limitation is really nutrient limitation; increased soil moisture increases decomposition rates and nutrient mineralization rates, and hence nutrient supply (Chapin *et al.*, 2002). Arid grasslands are never as productive in wet years as grasslands that regularly receive high moisture inputs, suggesting that arid grasslands lack biomass, species, or soil fertility to exploit high moisture years (Laurenroth and Sala, 1992). Thus water controls NPP, but soil moisture determines NPP in three ways: direct stimulation of NPP, its effect on nutrient supply, and its constraints on species composition and the productive capacity of an ecosystem. This example illustrates the complexity of the interactions between C and water cycles in terrestrial ecosystems.

DEVIATIONS FROM ''THE MODEL'': BIOTIC REGULATION OF C AND WATER CYCLES

Abiotic factors, including climate and growing season length strongly regulate fluxes of water and carbon into and out of ecosystems. At least on the global scale, there are robust relationships between NPP, soil processes, and precipitation, and these processes link the C and water cycle in ecosystems. Water is a basic requirement for biological organisms. Thus, it is not surprising that water availability regulates biological processes such as NPP and decomposition. These relationships between climatic

variables and ecosystem processes are extremely useful, as they allow estimation of NPP without direct measurement (e.g. Schuur, 2003; Lieth, 1975). However, while climate drives much of the large-scale variability in C cycling, the variability of the data compiled to generate these relationships suggests that other factors influence processes at the plot- to ecosystem-levels (Figure 2). For example, precipitation inputs alone explain 56% of the variability in NPP at the global scale, while temperature alone explains only 47% (Schuur, 2003). Thus, while the relationships are fairly robust, the variability of the data highlights the inability of climate to accurately predict NPP values at small scales. The inability to capture variability at smaller scales suggests that other factors influence plant and soil processes and their effects on C and water cycling.

ECOLOGICAL CONSTRAINTS ON ECOSYSTEM PROCESSES

We refer to the factors other than temperature, moisture, and radiation that may affect plant and soil processes as ecological constraints. An understanding of the effects of ecological constraints on ecosystem processes is critical for predicting interactions between C and water on small scales. The ecological constraints on plant and soil processes can take many forms, and their effects are often complex. Ecological controls on ecosystem processes can vary both spatially and temporally, can have positive or negative consequences, and can have rapid and dramatic effects on ecosystem C cycling. As a result, ecological constraints on plant and soil processes can have profound implications for water cycling at many levels of organization, from the community to the ecosystem scale.

Most ecological constraints on soil and plant processes fall into three general categories: trophic interactions, or the activity of other organisms in an ecosystem that affect plants and soils directly or indirectly (top-down controls); biogeochemical interactions, or the effects of soil nutrient status and soil biogeochemistry on C cycling (and hence water cycling) of an ecosystem (bottom-up controls); and anthropogenic/environmental interactions, or the influence of the many facets of global environmental change (including those driven by human activity) on ecosystem processes. Although many of the consequences of global change on C and water cycling are largely speculative (due to the rate at which the changes are occurring and ecosystems are responding), it is useful to consider some of the major changes and to discuss the state of the science predicting how global change may affect the interactions between C and water cycling.

Trophic Interactions

Herbivory is the consumption of living plant tissue (primary producer) by an organism occupying another tropic level (consumer), and may influence ecosystem NPP. By the broadest definition, consumers include: parasitic and phytophagous microorganisms (e.g. fungi and algae); phytophagous invertebrates (e.g. stem and foliage feeding insects, root-feeding insects and nematodes and seed predators); and browsing and grazing vertebrates (Barbour *et al.*, 1999). Estimates suggest that on a global scale, ~10–20% of terrestrial NPP is consumed by herbivores, and that the percentage varies by ecosystem. Values of NPP reduction from herbivory in natural ecosystems range from 2–3% in deserts and alpine tundra, 4–7% for forests, 10–15% in lightly grazed grasslands, and 30–60% in heavily grazed grasslands (Barbour *et al.*, 1999). However, periodic outbreaks of herbivore pests such as pathogens or insects may consume up to 100% of NPP.

Grazing

Herbivory in managed ecosystems results in significantly greater reductions in NPP, and thus the associated feedbacks between NPP and the C and water cycles may be impacted to a greater degree than in natural ecosystems (*see* **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**). In particular, grazing may decrease productivity directly, through a decrease in the potential photosynthetic plant material capable of fixing atmospheric CO_2 into new biomass. An analysis of 276 cases of grazed–ungrazed comparisons over 236 sites, Milchunas and Lauenroth (1993) demonstrated the significance of grazing in heavily managed ecosystems. In general, grazing had predominantly negative impacts on NPP. Grazing decreased aboveground NPP (ANPP) by 44%, 55%, 51%, and 60% in grassland, shrubland, mountain (e.g. alpine), and forest ecosystems, respectively (Milchunas and Lauenroth, 1993).

However, predicting the net effects of decreasing NPP from grazing on C and water cycles are complex. For example, the removal of vegetation by grazing can lead to profound changes in soil properties that can affect water balance. Diminished NPP due to leaf herbivory may contribute to decreased losses of water *via* transpiration. However, leaf herbivory that removes canopy vegetation may alter the balance between light absorbed and reflected (i.e. changes in albedo) and increase the importance of evaporation from soils. Additionally, the effects of grazing and the associated feedbacks on the C and water cycles are not limited to the effects on NPP. Soil trampling and vegetation removal can decrease soil organic matter and soil water-holding capacity. Furthermore, soil compaction can increase soil bulk density and break up soil aggregates, reduce water infiltration, soil stability, soil food-web structure, and nutrient cycling rates (Warren *et al.*, 1986; Ingham *et al.*, 1989). Finally, vegetation removal can increase the amount of bare soil, and leads to soil erosion through both hydrological

and aeolian processes (*see* **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**; **Chapter 119, Land Use and Land-cover Effects on Runoff Processes: Forest Harvesting and Road Construction, Volume 3**). All of these potential interactions highlight the complexity of the relationship underpinning the linkage between C and water cycling, and the potential importance of trophic interactions in regulating the relationship.

Insect Herbivory

There are multiple lines of evidence suggesting that grazing of arid lands can have profound effects on ecosystem C cycling which, in turn, could dramatically affect water cycle in terrestrial ecosystems. However, grazing is not the only top-down regulator of ecosystem C cycling. Herbivorous insects and pathogens also play a key role in regulating vegetation growth and dynamics, and represent one of the most important and pervasive agents of disturbance in ecosystems worldwide (Logan *et al.*, 2003). Insects are important components of most ecosystems, and can have both positive and negative effects on ecosystem processes. However, insects have the capacity for extremely high population growth. Thus, when outbreaks of herbivorous insects occur, the effects can be profound. In North American forests, insects affect an area more than 50 times larger than fire on an annual basis, and with substantially greater economic repercussions (Dale *et al.*, 2001).

Insect outbreaks can alter the interactions between C and water cycling in several important ways. First, significant foliage removal or tree mortality resulting from insect herbivory may lead to declines in NPP. In extreme cases, insect defoliation can remove 100% of NPP. At the scale of the individual, decreases in leaf area may lead to declines in plant water uptake and transpiration. At larger scales, this effect can lead to changes in the ways water cycles through the affected ecosystem. For example, landscape-level insect denudation of plant biomass may favor increased water losses as runoff and, thus, lead to decreases in soil moisture across the landscape. Such declines in soil moisture may, in turn, lead to declines in evaporation, plant stand relative humidity and nutrient cycling. Insect outbreaks resulting in large-scale stand mortality may further influence C and water cycling *via* negative effects on nesting birds, mycorrhizal fungi (References in Dale *et al.*, 2001), and on climate (which may affect competitors and natural enemies regulating the abundance of pests and pathogens), creating another potential feedback on insect distribution and abundance (*see* **Chapter 103, Terrestrial Ecosystems, Volume 3**; **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**). Large-scale stand mortality may also accelerate stand fire regimes, further affecting short-term C and water cycling processes.

Interactions between insect outbreaks and water availability are also complex, and are not unidirectional. Specifically, evidence suggests that the potential ecological effects of insect herbivory not only influence ecosystem water balance, but there is also mounting evidence that insect outbreaks in terrestrial ecosystems are related to climate fluctuations, and in particular, water availability (Speer *et al.*, 2001). For example, reductions in soil water potential from low precipitation and soil moisture (drought) have been linked to declines in tree production and health, and such environmental stresses often predispose plant species to attack by insects and other parasites (Hanks *et al.*, 1999). Decreases in tree health resulting from drought often reduce tree resistance to colonization by insects. Thus, in many cases, drought weakens trees, making them more vulnerable to insect infestations. In contrast, using a tree-ring reconstruction, Ryerson *et al.* found evidence that outbreaks of spruce budworm, an insect affecting mixed conifer forests, correspond to increases in moisture. Grassland ecosystem insect susceptibility has also been demonstrated to correlate with changes in precipitation inputs.

A growing body of evidence also suggests that insect herbivory can directly alter nutrient cycling in terrestrial ecosystems, with profound consequences for NPP. Tree damage caused by insect herbivory leads to obvious C losses from trees. These C debits can, in turn, have negative effects on mycorrhizal fungi (Chapin *et al.*, 2002). Mycorrhizal fungi form symbioses with nearly all plants, and their function in the symbiosis is to absorb and translocate important nutrients to their plant host. Mycorrhizal fungi are capable of absorbing mineral nutrients, like phosphorus (P) at much lower concentrations than nonmycorrhizal plant roots. However, the advantages of mycorrhizal infection are reciprocated through a flow of photosynthetically produced C from the host plant to the fungus, providing the C substrates necessary for the fungus to meet much of its energy requirement to grow and sustain biomass. Declines in C available to sustain such symbioses as a result of insect defoliation have clear implications for nutrient acquisition, and thus may also affect ecosystem water fluxes.

Plant Pathogens

Insect infestations also can affect the interactions between C and water cycling *via* interactions with other organisms. For example, plant disease can have a major impact on terrestrial ecosystem dynamics. Chestnut blight in North America and jarrah dieback in Western Australia have severely impacted plant species populations, leading to a cascade of changes to many affected forest ecosystems (References in Rizzo and Garbelotto, 2003). *Phytophthora cinnamomi*, a plant pathogen affecting jarrah, has eliminated most tree species over hundreds and thousands of hectares of the eucalyptus forests of Western Australia, converting them to grassland or shrubland (Rizzo and Garbelotto, 2003).

The severity of such diseases is often related to insect outbreaks. Dutch elm disease, which decimated the population in North America, was spread through the population by elm beetles that carried fungal spores from individual to individual. The wholesale conversion of forest to grassland has profound implications for the water cycle in these affected ecosystems (*see* **Chapter 103, Terrestrial Ecosystems, Volume 3**). The effects of such feedbacks are likely to increase as introduced exotic diseases spread and impact native plant populations (Rizzo and Garbelotto, 2003).

Biogeochemical Constraints on Ecosystem Processes: NPP

The observed decreases in NPP in ecosystems with high precipitation (Figure 2) suggest that in mesic systems, either excess water inhibits NPP, or that in these ecosystems, other biotic or abiotic processes limit NPP. In many terrestrial ecosystems, nutrient availability strongly regulates NPP and decomposition, and thus the nutrient status of an ecosystem is necessary to understand ecosystem C and water balance. However, nutrient limitation to ecosystem processes is often complex and thus generalizations about nature of nutrient limitation in ecosystems are difficult to make. A substantial amount of data from temperate and high latitude ecosystems suggests that nitrogen (N) commonly limits plant growth (Aber *et al.*, 1991; Vitousek and Howarth, 1991). While nitrogen gas (N_2) comprises ~80% of the atmosphere, N_2 gas must be fixed into reactive forms to be utilized by plants. In the absence of human influence, two major processes convert N_2 into biologically available forms: lightning and biological nitrogen fixation (BNF) by microorganisms (both free-living and in symbiotic associations with plants). However, N_2 fixation by terrestrial BNF is roughly an order of magnitude greater than that by lightning (Galloway *et al.*, 1995) and is therefore the dominant source of newly fixed N to the landscape.

The N status of ecosystems is largely dependent on the presence and activity of the organisms that can convert N_2 into usable forms. In temperate ecosystems, relatively frequent glaciations effectively remove accumulated N by removing soils and vegetation. The succession of plant and microbial species (and hence soil N accumulation) following deglaciation is relatively slow (10^3–10^4 years), and this phenomenon results in relatively N-poor ecosystems where NPP is N limited. Evidence for N limitation of NPP comes from a wide variety of data, including manipulations of nutrient availability (Figure 6). The pervasiveness of N limitation, and the positive effect of N fertilization on primary production are also clearly illustrated by the magnitude of N applied as fertilizer to the world's agricultural systems. Thus, in temperate ecosystems, an understanding of the factors regulating NPP and the associated feedbacks on water cycling in terrestrial ecosystems requires an understanding of the effects of nutrient status, particularly N status, on primary productivity.

In contrast to temperate and high latitude ecosystems, many tropical forests lie on extremely weathered soils (Vitousek and Sanford, 1986). Many tropical ecosystems have not been influenced by large-scale natural phenomena (e.g. glaciations), and as a result, many tropical ecosystems are rich with free-living N fixers and symbiotic N-fixing species. In these ecosystems, NPP is commonly limited by rock-derived elements (Vitousek and Sanford, 1986). Over millions of years, high rates of precipitation result in rapid chemical and physical weathering of soil, and lead to losses of important, relatively "non-renewable" elements like calcium (Ca), phosphorus (P), magnesium (Mg), or potassium (K). Low levels of these

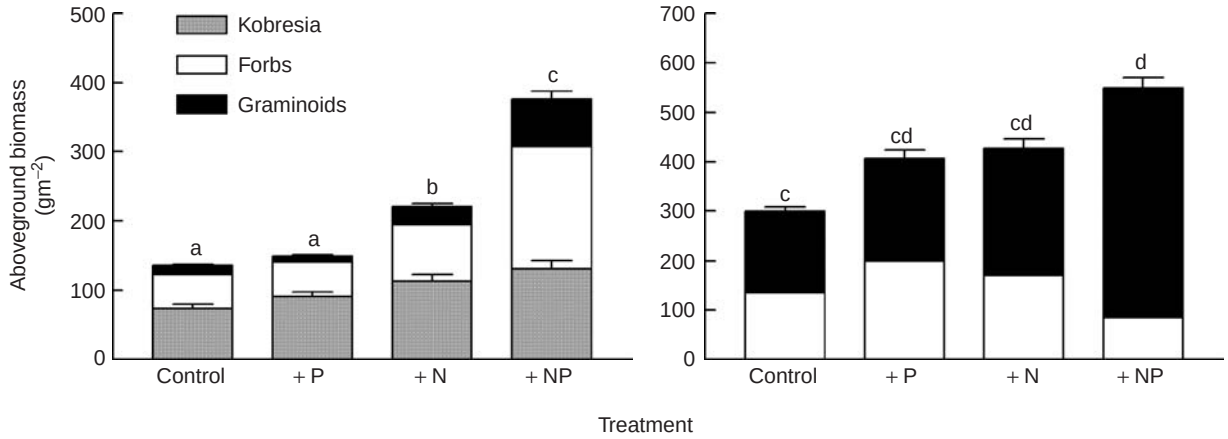


Figure 6 Aboveground biomass production for two alpine tundra communities, dry and wet meadows, under four treatments: control (C), phosphorus (P), nitrogen (N), and N + P treatments. Letters above bars indicate homogeneous means as determined by a Tukey's multiple range test (Reproduced from Bowman *et al.* (1993), by permission of Ecological Society of America)

nutrients may contribute to declines in NPP in tropical ecosystems with high precipitation. Phosphorus is generally believed to be the most limiting element in the majority of tropical forests on older soils (Vitousek, 1984). Indirect attempts to assess nutrient limitation, such as foliar element ratios, strongly suggest P constraints on NPP in many tropical forests on such soils (*see* **Chapter 96, Nutrient Cycling, Volume 3**). Long-term fertilizations in Hawaiian rainforests have also shown that NPP on older soils is clearly limited by phosphorus (Vitousek and Farrington, 1997), and root in-growth studies under different fertilizer applications suggested phosphorus (and possibly calcium) limitation on one oxisol in the Venezuelan portion of Amazonia (Cuevas and Medina, 1988). In contrast, fertilizer additions to a younger soil in the mountains of western Venezuela showed N to be the primary limiting nutrient (Tanner *et al.*, 1998). Similar results were found on younger soils in Hawaii (Vitousek and Farrington, 1997). The latter data suggest the complexity of nutrient controls on NPP, but clearly illustrate the potential for nutrient constraints of NPP.

Biogeochemical Constraints on Ecosystem Processes: Decomposition

The strong relationship between organic matter decomposition and AET (Meentemeyer, 1978) suggests the importance of ecosystem water availability on nutrient cycling and hence NPP. However, in some cases, microbial decomposition of plant material may also be strongly nutrient limited, and in such cases, may exacerbate plant nutrient limitation. For example, in many nutrient poor ecosystems, the pool of actively cycling nutrients is insufficient to allow rapid microbial decomposition. Hobbie and Vitousek (2000) showed that in a phosphorus limited tropical forest in Hawaii, decomposition was also limited by phosphorus (*see* **Chapter 96, Nutrient Cycling, Volume 3**). While microbial nutrient immobilization may prevent important nutrient losses in the long-term, in the short-term, the superior ability of soil microorganisms to compete for available soil nutrients may exacerbate plant nutrient limitation, thus impacting NPP.

C AND WATER CYCLES: THE ROLE OF ENVIRONMENTAL CHANGE

The carbon budget of terrestrial ecosystems is tightly coupled to the water cycle, and changes in C cycling in ecosystems can have potentially profound impacts on the water cycle. Moreover, while climate exerts strong control on C cycling at coarse scales, finer scale variations in C cycling (and hence water cycling) are greatly influenced by ecological processes and interactions. Thus, an understanding of these interactions is critical in assessing the future water balance of terrestrial ecosystems.

Water is a fundamentally important natural resource vital for ecosystem functioning and human well-being. Human use of fresh water, which is expected to triple in the next two decades, and contamination are stressing this important resource, and perturbations to the hydrologic cycle may have profound consequences for people and the environment (Graedel *et al.*, 2001). Consequently, the Natural Research Council of the National Academies of the United States recently suggested that accurate prediction of "changes in fresh water resources and the environment caused by floods, droughts, sedimentation, and contamination in a context of growing demand on water resources" was one of the "Grand Challenges in Environmental Science" (Graedel *et al.*, 2001). Success in this endeavor depends on an appreciation for the tight coupling of the C and water cycles, and an understanding of the complexity of these interactions in a world that is experiencing unprecedented, rapid environmental change.

The global environment is undergoing rapid modification. As the human population and the magnitude of global change continue to grow, an understanding of the structure and functioning of ecosystems will not be possible without an understanding of the strong influence of human activity (Vitousek, 1994). Human activities, including agriculture, industry, fishing, and international commerce have changed the land surface, perturbed global biogeochemical cycles, and altered species dynamics in most of earth's ecosystems. While some of the direct effects are well documented, many of the indirect effects of global change on ecosystem structure and functioning are much more tenuous. However, some of these interactions will undoubtedly result in alterations to global C and water cycles. Specifically, three major categories of global change will play important roles regulating C and water cycling in ecosystems: land transformations, alterations of global biogeochemical cycles, and biotic changes (Vitousek, 1994).

Land Transformations

Humans have caused dramatic changes to the earth's land surface, and this trend is likely to increase well into the future. At present, it is estimated that ~35% of the earth's continental surface (55 million km^2) has been cleared and converted to cropland, pasture, and urban settlements (Figure 7; Foley *et al.*, 2003). To date, the majority of human-induced land conversion has occurred in temperate terrestrial ecosystems; the only three significant remaining areas to be exploited are tropical rain forests (in South America, Asia, and Africa), boreal forests (in Canada and Russia), and deserts (Foley *et al.*, 2003). These ecosystems face increasing development pressure as the rising human population demands more forest and agricultural

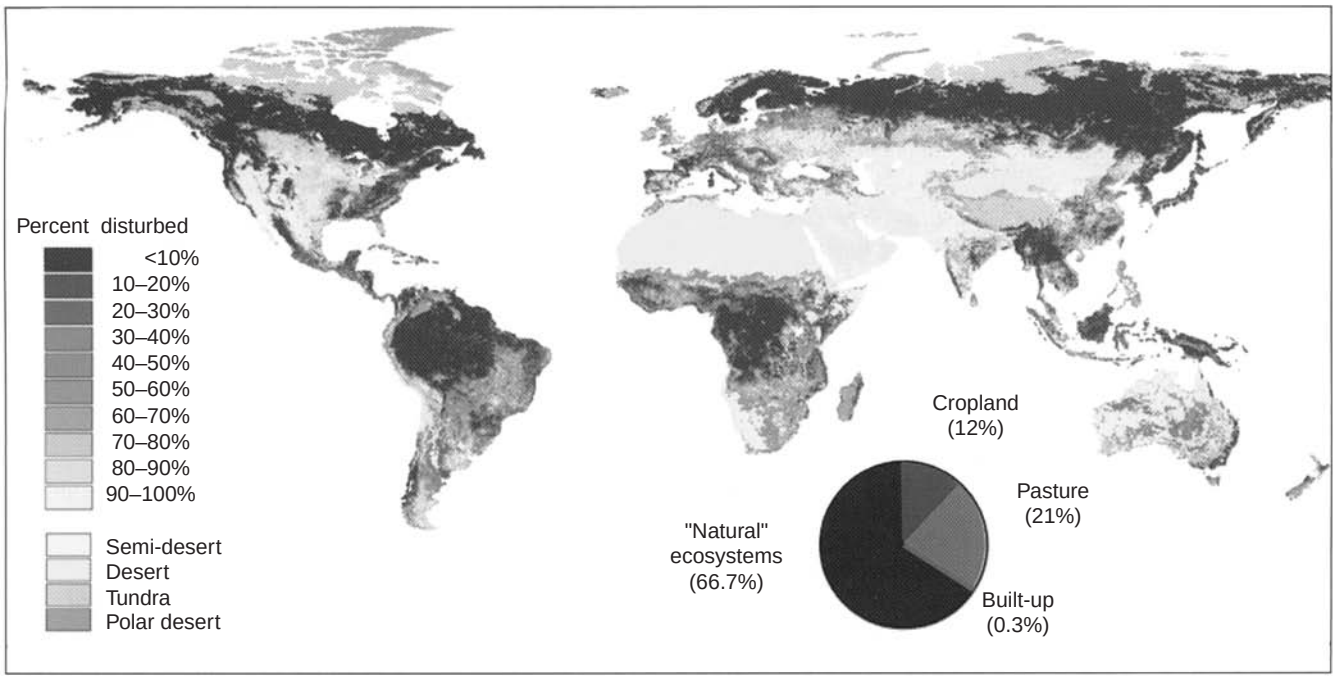


Figure 7 The global extent of human land use, including croplands, pastures, and urban areas across the world (Reproduced from Foley *et al.,* 2003 by permission of Ecological Society of America). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

products. Activities such as wood extraction, forest conversion to pasture, burning, irrigation, and livestock and poultry management will all affect the biogeochemical cycling of elements, including C and water (*see* **Chapter 103, Terrestrial Ecosystems, Volume 3**).

Rates of human-induced land conversion are reaching rates unprecedented in history, and the ecological consequences of land conversion and land use change are numerous (*see* **Chapter 57, Land-cover Classification and Change Detection, Volume 2**). Land conversion leads to losses of indigenous species, introductions of exotic species, rerouting of hydrological flows, and industrial contamination of water, air, and land (Graedel *et al.*, 2001). In addition, land conversion often leads to wholesale changes in the biogeochemical cycling of important elements, including C and water (Vitousek, 1994). However, the specific effects of land use change on C and water cycling are often ecosystem-dependent. For example, in North America, changes in C fluxes following land conversion are related to the fate of converted lands, or the proportion of irrigated to nonirrigated croplands (*see* **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**) or grazed lands (Ojima *et al.*, 1994). These differences, in turn, affect regional climatic patterns, including precipitation inputs to the landscape (Pielke *et al.*, 1997).

Recent evidence suggests that land conversion and the accompanying ecological changes in tropical rain forests may lead to profound feedback that alter water cycling in these ecosystems (*see* **Chapter 103, Terrestrial Ecosystems, Volume 3**). Moist tropical forests comprise one of the world's largest and most diverse biomes, and exchange more carbon, water, and energy with the atmosphere than any other ecosystem. In recent decades, tropical forests have also become one of the world's most threatened biomes, subjected to exceptionally high rates of deforestation and land degradation (e.g. Nepstad, 1999). The effects of land conversion and land use change in tropical rain forests have been thoroughly investigated in recent years (References in Foley *et al.*, 2003). The climatic impacts of such changes are typically evaluated using linked general circulation and biophysical land surface models (Shukla *et al.*, 1990). Most of these analyses suggest that large-scale deforestation and conversion of tropical rain forest will result in a significant temperature increase and decreases in annual evapotranspiration and rainfall (Shukla *et al.*, 1990). Climatic changes are driven largely by shifts in surface energy, water, and momentum balance that accompany deforestation (Figure 8). Lower surface roughness, leaf area, and root depth in pastures relative to forests reduce evapotranspiration, resulting in a decrease in evaporative cooling and surface temperature increases (*see* **Chapter 45, Actual Evaporation, Volume 1**).

Modeled reductions in precipitation following large-scale tropical deforestation also result from changes in ecosystem water and energy balance. For example, reduced APAR

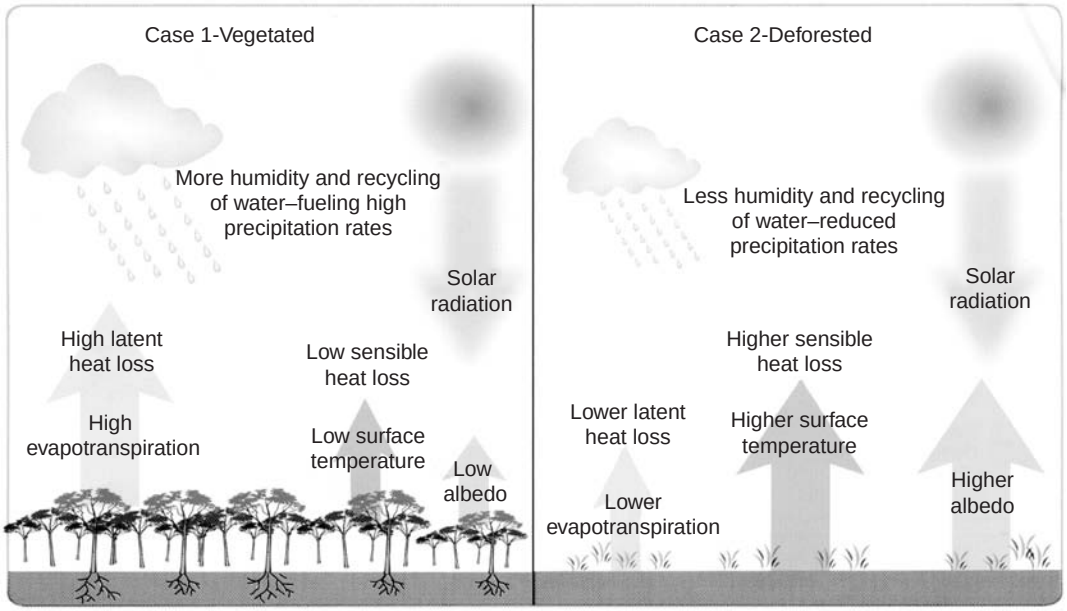


Figure 8 Climatic effects of tropical deforestation on water balance, boundary-layer fluxes, and climate. In vegetation-covered areas (a), the low albedo of the forest canopy provides ample energy for the plants to photosynthesize and transpire, leading to a high latent heat loss that cools the surface. In deforested areas (b), bare soil's higher albedo reduces the amount of energy absorbed at the surface. Latent heat loss is severely reduced and the surface warms, as it has no means of removing the excess energy through transpiration (Reproduced from Foley *et al.*, 2003 by permission of Ecological Society of America). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and elevated surface temperatures contribute to a decrease in the net radiative heating of the land, or a decrease in the energy available to fuel atmospheric circulation and convective precipitation (e.g. Wang and Eltahir, 2000a). By reducing AET, land conversion of forests to pastures results in lower atmospheric water availability, contributing to a decrease in rainfall, and may set into motion the feedbacks (Figure 8). For example, water deficits in the Brazilian Amazon are also linked to increased fire susceptibility, another disturbance mechanism that can affect C and water cycling in these areas (Figure 8).

The potentially profound effects of land conversion and land use change on C and water cycling are also observed in analyses of desert ecosystems. For example, the pronounced and persistent drought conditions present in the Sahel region of West Africa for the past three decades have drawn considerable attention. While there are two theories on the physical mechanism behind this drought, the most widely accepted is that intense human activity in the region have significantly altered land cover at the regional scale and caused the prolonged drought conditions (Wang and Eltahir, 2000b). In particular, it is suggested that ecosystem dynamics play an important role in regulating the climate of the region.

Using a coupled biosphere-atmosphere model that incorporated ecosystem dynamics and feedback between the biosphere and atmosphere, Wang and Eltahir (2000b) explored the relationship between land use change and drought in the Sahel. As discussed previously, vegetation plays a prominent role in the exchange of carbon, water, and energy between the land surface and the atmosphere. Vegetation removal can modify the local carbon, water, and energy balance. Specifically, in the Sahel region, overgrazing, overcultivation of marginal land, slash-and-burn agricultural practices, logging, and poor irrigation techniques have led to widespread land degradation and desertification. Following desertification, ET (and hence surface latent heat flux) significantly decreases, as does precipitation. In simulations, drought conditions are the result of feedbacks involving the response of the natural ecosystem to imposed changes in land cover (Wang and Eltahir, 2000b). Further, when damage to an ecosystem reaches a threshold, climate changes significantly and leads to deterioration of other "healthy" natural ecosystems (Wang and Eltahir, 2000b; Figure 9). Research also suggests that sea surface temperature (SST) is also critical in regulating Sahel climate, but that ecosystem dynamics may dampen the response of the climate system to recover following changes in SST that would promote a return to wetter conditions (Wang and Eltahir, 2000b). Thus, similar to tropical rain forests, land conversion in arid and semiarid ecosystems may lead to changes in regional climate that continue to have feedbacks on C and water cycling.

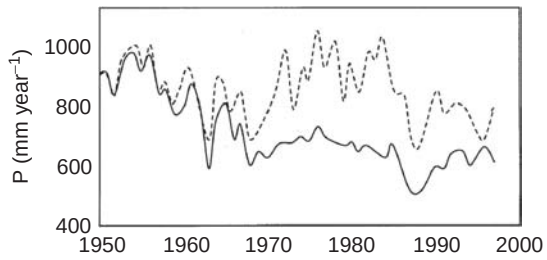


Figure 9 Drought initiation by desertification. Rainfall average over the Sahel region, in the control (dashed line) and desertification (solid line) experiments (Reproduced from Wang and Eltahir, 2000b, by permission of American Geophysical Union)

Alterations to Global Biogeochemical Cycles

Human activities are leading to profound alterations to the earth's major biogeochemical cycles. For example, fertilizer use, legume crop production and fossil fuel combustion have doubled the amount of N entering ecosystems *via* all natural pathways combined (Galloway *et al.*, 1995). Next, it is estimated that humans use 10 to 55% of annual terrestrial NPP for food and fiber or to support grazing animals (Rojstaczer *et al.*, 2001). Perhaps, the most well-documented aspect of global environmental change is the recent increase in atmospheric CO_2 (Schlesinger, 1997), and rising CO_2 could have profound implications for ecosystem water balance. Rising CO_2 has been postulated to affect plant growth, ecosystem structure, and ecosystem function in many ways. In both laboratory and field experiments, plant photosynthesis often increases under elevated CO_2, and this can lead to changes in C cycling at the ecosystem scale (References in Hungate *et al.*, 1997). Increased CO_2 can lead to decreases in plant stomatal conductance, and corresponding decreases in plant transpiration and evapotranspiration. Changes in stomatal conductance can, in turn, have profound consequences for ecosystem water cycling. Studies conducted in both herbaceous and woody ecosystems have demonstrated reductions in evapotranspiration leading to increases in soil water content and increased plant water-use efficiency (Figure 10). Increased CO_2 effects on plant water use and transpiration could have feedbacks ranging from changes in soil processes, increased water yield in watersheds, or to altered precipitation regimes. In systems like the Amazon Basin where evapotranspiration provides moisture for convective storms, increased water use efficiency with increased CO_2 may feed back to lower regional precipitation (Shukla *et al.*, 1990; Nobre *et al.*, 1991).

Increasing CO_2 concentrations in the atmosphere has two potential effects on climate. First, the radiative effect of increased CO_2 leads to tropospheric warming (the greenhouse effect). Next, the radiative effect of increased CO_2 warming the troposphere may also increase atmospheric water content (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**) (through more evaporation) and intensify the global water cycle. Such changes would be beneficial to vegetation in regions where water availability or low temperatures limit plant NPP. However, there is also direct impact of changes in atmospheric CO_2 concentration on vegetation through the physiological effects, or through decreases in evapotranspiration. In arid regions, where climate is often driven by vegetation dynamics, relatively small changes in plant physiology resulting from increased atmospheric CO_2 could have profound implications on climate and particularly on precipitation regime in these systems. Thus, changes to the overall water balance of an ecosystem represent the net effects of the biotic and abiotic responses to elevated CO_2 concentrations. Using a coupled biosphere-atmosphere model to simulate the effects of increased atmospheric CO_2 on climate in the Sahel region of Africa, Wang and Eltahir (2002) found that elevated CO_2 did result in significant declines in plant transpiration due to increased plant water-use efficiency. However, the radiative effects of CO_2, which enhance precipitation, overshadowed the physiological effects, and led to more rainfall to the Sahel (Wang and Eltahir, 2002). Under this simulation, precipitation increased and led to increased NPP. In this specific ecosystem, increased water

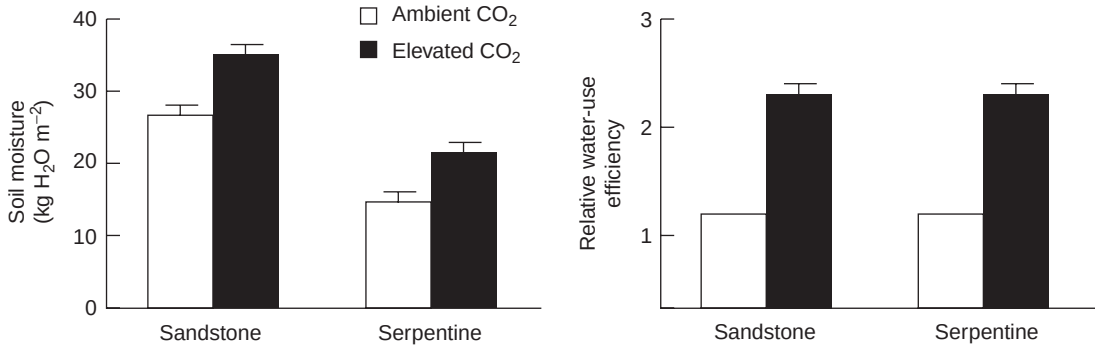


Figure 10 Soil moisture (a) and relative water-use efficiency (b) in untreated and elevated CO_2-treated plots. Values are means ± S.E. (Reproduced from Hungate *et al.*, 1997 by permission of Springer)

availability led to denser vegetation growth at higher CO_2 concentrations, causing subsequent increases in evapotranspiration (i.e. as a result of an increase in transpiring leaf surface area) and reinforced higher levels of precipitation. This simulated feedback again illustrates the close interaction between C and water cycles, and provides clear evidence of the potential effects of a biogeochemical perturbation (increased CO_2) on regional hydrologic cycles.

Biotic Changes

Human activity has greatly modified the earth's biological resources, and changes to its species and genetically distinct populations are on the rise (Vitousek, 1994). While extinction is a natural phenomenon, current rates of species loss are much higher than "background" levels, and this accelerated extinction rate results in part from the activity of humans. However, human enterprise is not only affecting species loss rates, but causing substantial species redistribution by transporting "exotic" species into areas where they do not occur naturally. In some terrestrial ecosystems, plant invasions have led to complete shifts in species composition, and caused dramatic changes in the structure and functioning of ecosystems. In freshwater ecosystems, accidental species introductions have led to widespread invasions, and often lead to wholesale changes in the hydrologic cycle. In many arid regions in North America and Africa, exotic species have established extensively along watercourses. While harsh growing conditions in arid lands generally limit the extent of exotic plant invasions, riparian areas provide suitable habitat for some exotic species, and their establishment can have profound ecological and hydrological consequences.

For example, saltcedar (*Tamarix*) was first introduced into the western United States, and now dominates many waterways throughout the southwest (Sala *et al.*, 1996). Saltcedar invasion has profound consequences for water cycling in arid ecosystems. Saltcedar traps and stabilizes alluvial sediments, and can greatly decrease river channel width (Graf, 1978). An increase in stabilized deposits along stream channels can decrease the ability of the channel to adjust during high flow events, leading to increased flooding. Saltcedar invasion can also directly affect water cycling in invaded areas. Recent research suggests that saltceder uses water much more inefficiently than native vegetation, and the increased water losses as a result of saltcedar invasion are profound. In riparian areas where water availability is high, dense saltcedar stands are characterized by extremely high rates of evapotranspiration, and AET in these areas can exceed potential evapotranspiration by a factor of 2 in nonriparian sites within the ecosystem (Sala *et al.*, 1996). Increased riverine water losses due to saltcedar invasion could be extremely important in arid regions that depend on annual reservoir replenishment *via* riverine inputs to meet the water demand of their growing populations. In this case, saltcedar invasion decreases water that would otherwise be available to meet the growing agricultural and urban water demands in arid regions.

Acknowledgments

We thank G. Wang, W. Bowman, and E. Schuur for providing copies of original figures. We also thank R. Rawlinson for providing valuable comments on early versions of the work. Financial support to C.C. was provided by NSF Grant #0089447.

FURTHER READING

Aber J. and Melillo J. (2001) *Terrestrial Ecosystems*, *First Edition*, Harcourt/Academic Press: Burlington.

Cuevas E. and Medina E. (1985) Nutrient dynamics with Amazonian forest ecosystems I: Nutrient flux in fine litter fall and efficiency of nutrient utilization. *Oecologia*, **76**, 222–235.

Hungate B.A., Chapin F.S. III, Zhong H., Holland E.A. and Field C.B. (1996) Stimulation of grassland nitrogen cycling under carbon dioxide enrichment. *Oecologia*, **109**, 149–153.

Lieth H. and Whittaker R.H. (1975) *Primary Productivity of the Biosphere*, Springer-Verlag: New York.

Sala O.E., Jackson R.B., Mooney H.A. and Howarth R.W. (2000) *Methods in Ecosystem Science*, Springer: New York.

REFERENCES

Aber J.D., Melillo J.M., Nadelhoffer K.J., Pastor J. and Boone R.D. (1991) Factors controlling nitrogen cycling and nitrogen saturation in northern temperate forest ecosystems. *Ecological Applications*, **1**, 303–315.

Aber J.D. and Melillo J.M. (2001) *Terrestrial Ecosystems*, *Second Edition*, Harcourt/Academic Press: Burlington.

Barbour M.G., Burk J.H., Pitts W.D., Gilliam F.S. and Schwartz M.W. (1999) *Terrestrial Plant Ecology*, *Third Edition*, Benjamin/Cummings: Menlo Park.

Bowman W.D., Theodose T.A., Schardt J.C. and Conant R.T. (1993) Constraints of nutrient availability on primary production in two alpine tundra communities. *Ecology*, **74**, 2085–2097.

Chapin F.S. III, Matson P.A. and Mooney H.A. (2002) *Principles of Terrestrial Ecosystem Ecology*, Springer: New York.

Clark D.A., Brown S., Kicklighter D.W., Chambers J.Q., Thomlinson J.R. and Ni J. (2001a) Measuring net primary production in forests: Concepts and field methods. *Ecological Applications*, **11**, 356–370.

Clark D.A., Brown S., Kicklighter D.W., Chamber J.Q., Thomlinson J.R., Ni J. and Holland E.A. (2001b) Net primary production in tropical forests: An evaluation and synthesis of existing field data. *Ecological Applications*, **11**, 371–384.

Cleveland C.C., Neff J.C., Townsend A.R. and Hood E. (2004) Composition, dynamics and fate of leached dissolved organic

matter in terrestrial ecosystems: results from a decomposition experiment. *Ecosystems*, **7**, 275–285.

Cuevas E. and Medina E. (1988) Nutrient dynamics with amazonian forests II. Fine root growth, nutrient availability, and leaf litter decomposition. *Oecologia* **76**, 222–235.

Dale V.H., Joyce L.A., McNulty S., Neilson R.P., Ayres M.P., Flannigan M.D., Hanson P.J., Irland L.C., Lugo A.E., Peterson C.J., *et al.* (2001) Climate change and forest disturbances. *Bioscience*, **51**, 723–734.

Foley J.A., Costa M.H., Delire C., Ramankutty N. and Snyder P. (2003) Green surprise: How terrestrial ecosystems could affect earth's climate. *Frontiers in Ecology and the Environment*, **1**, 38–44.

Galloway J.N., Schlesinger W.H., Levy H., Michaels A. and Schnoor J.L. (1995) Nitrogen fixation: Anthropogenic enhancement-environmental response. *Global Biogeochemical Cycles*, **9**, 235–252.

Gower S.T. (2002) Productivity of terrestrial ecosystems. In *Encyclopedia of Global Change*, Mooney H.A. and Canadell J. (Eds.), Blackwell Scientific: Oxford.

Graedel T.E., Alldredge A., Barron E., Davis M., Field C., Fischhoff B., Frosch R., Gorelick S., Holland E.A., Krewski D., *et al.* (2001) *Grand Challenges in Environmental Sciences*. National Academy of Sciences: Washington, DC, p. 96.

Graf W.L. (1978) Fluvial adjustments to the spread of tamarisk in the Colorado Plateau region. *Geological Society of America Bulletin*, **89**, 1491–1501.

Hanks L.M., Paine T.D., Millar J.G., Campbell C.D. and Schuch U.K. (1999) Water relations of host trees and resistance to the phloem-boring beetle *Phoracantha semipunctata* F-(Coloeptera: Cerambycidae). *Oecologia*, **119**, 400–407.

Hobbie S.E. and Vitousek P.M. (2000) Nutrient limitation of decomposition in Hawaiian forests. *Ecology*, **81**, 1867–1877.

Hungate B.A., Holland E.A., Jackson R.B., Chapin F.S., Mooney H.A. and Field C.B. (1997) The fate of carbon in grasslands under carbon dioxide enrichment. *Nature*, **388**, 576–579.

Ingham E.R., Coleman D.C. and Moore J.C. (1989) An analysis of food-web structure and function in a shortgrass prairie, a mountain meadow, and a lodgepole pine forest. *Biology and Fertility of Soils*, **8**, 29–37.

Laurenroth W.K. and Sala O.E. (1992) Long-term forage production on North American shortgrass steppe. *Ecological Applications*, **2**, 397–403.

Lieth H. (1975) Modeling the primary productivity of the world. In *Primary Productivity of the Biosphere*, Lieth H. and Whittaker R.H. (Eds.), Springer-Verlag: New York, pp. 237–265.

Lindeman R.L. (1942) The trophic-dynamic aspects of ecology. *Ecology*, **23**, 399–418.

Logan J., Regniere A.J. and Powell J.A. (2003) Assessing the impacts of global warming on forest pest dynamics. *Frontiers in Ecology and Evolution*, **1**, 130–137.

Meentemeyer V. (1978) Macroclimate and lignin control of litter decomposition rates. *Ecology*, **59**, 465–472.

Milchunas D.G. and Lauenroth W.K. (1993) Quantitative effects of grazing on vegetation and soils over a global range of environments. *Ecological Monographs*, **63**, 327–366.

Montieth J.L. (1977) Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London*, **281**, 277–294.

Neff J.C. and Asner G.P. (2001) Dissolved organic carbon in terrestrial ecosystems: Synthesis and a model. *Ecosystems*, **4**, 29–48.

Nepstad D.C. (1999) Large-scale impoverishment of Amazonian forests by logging and fire. *Nature*, **398**, 505–508.

Nobre C.A., Sellers P.J. and Shukla J. (1991) Amazonian deforestation and regional climate change. *Journal of Climate*, **4**, 957–988.

Odum E. (1971) *Fundamentals of Ecology*, Saunders: Philadelphia.

Ojima D.S., Galvin K.A. and Turner B.L. (1994) The global impact of land use change. *Bioscience*, **44**, 300–304.

Paul E.A. and Clark F.E. (1996) *Soil Microbiology and Biochemistry*, *Second Edition*, Academic Press: San Diego.

Pielke R.A., Lee T.J., Copeland J.H., Eastman J.L., Ziegler C.L. and Finley C.A. (1997) Use of USGS-provided data to improve weather and climate simulations. *Ecological Applications*, **7**, 3–21.

Raich J.W. and Schlesinger W.H. (1992) The global carbon dioxide flux in soil respiration and its relationships to vegetation and climate. *Tellus*, **44B**, 81–99.

Rizzo D.M. and Garbelotto M. (2003) Sudden oak death: Endangering California and Oregon forest ecosystems. *Frontiers in Ecology and the Environment*, **1**, 197–204.

Rojstaczer S., Sterling S. and Moore N.J. (2001) Human appropriation of photosynthesis products. *Science*, **294**, 2549–2552.

Sala A., Smith S.D. and Devitt D.A. (1996) Water use by *Tamarix ramosissima* and associated phreatophytes in a Mojave Desert floodplain. *Ecological Applications*, **6**, 888–898.

Sala O.E. and Austin A.T. (2000) Methods of estimating aboveground net primary productivity. In *Methods in Ecosystem Science*, Sala O.E., Jackson R.E., Mooney H.A. and Howarth R.W. (Eds.), Springer: New York, pp. 31–43.

Sala O.E., Parton W.J., Joyce L.A. and Lauenroth W.K. (1988) Primary production of the central grassland region of the United States. *Ecology*, **69**, 40–45.

Schimel D.S., Braswell B.H. and Parton W.J. (1997) Equilibration of the terrestrial water, nitrogen, and carbon cycles. *Proceedings of the National Academy of Science*, **94**, 8280–8283.

Schlesinger W.H. (1997) *Biogeochemistry: An Analysis of Global Change*, *Second Edition*, Academic Press: San Diego.

Schlesinger W.H. and Andrews J.A. (2000) Soil respiration and the global carbon cycle. *Biogeochemistry*, **48**, 7–20.

Schuur E.A. (2003) Productivity and global climate revisited: The sensitivity of tropical forest growth to precipitation. *Ecology*, **84**, 1165–1170.

Schuur E.A., Chadwick O.E. and Matson P.A. (2001) Carbon cycling and soil carbon storage in mesic to wet Hawaiian montane forests. *Ecology*, **82**, 3182–3196.

Shukla J., Nobre C. and Sellers P. (1990) Amazon deforestation and climate change. *Science*, **247**, 1322–1325.

Smith R.L. (1996) *Ecology and Field Biology*, *Fifth Edition*, HarperCollins Publishers: New York.

Speer J.H., Swetnam T.W., Wickman B.E. and Youngblood A. (2001) Changes in pandora moth outbreak dynamics during the past 622 years. *Ecology*, **82**, 697–697.

Tanner E.V.J., Vitousek P.M. and Cuevas E. (1998) Experimental investigation of nutrient limitation of forest growth on wet tropical mountains. *Ecology*, **79**, 10–22.

Vitousek P.M. (1984) Litterfall, nutrient cycling, and nutrient limitation in tropical forests. *Ecology*, **65**, 285–298.

Vitousek P.M. (1994) Beyond global warming: Ecology and global change. *Ecology*, **75**, 1861–1876.

Vitousek P.M. and Farrington H. (1997) Nutrient limitation and soil development: Experimental test of a biogeochemical theory. *Biogeochemistry*, **37**, 63–75.

Vitousek P.M. and Howarth R.W. (1991) Nitrogen limitation on land and sea: How can it occur. *Biogeochemistry*, **13**, 87–115.

Vitousek P.M. and Sanford R.L. (1986) Nutrient cycling in moist tropical forest. *Annual Review of Ecology and Systematics*, **17**, 137–167.

Wang G. and Eltahir E.A.B. (2000a) Biosphere-atmosphere interactions over West Africa. I: Development and validation of a coupled dynamic model. *Quarterly Journal of the Royal Meteorological Society*, **126**, 1239–1260.

Wang G. and Eltahir E.A.B. (2000b) Ecosystem dynamics and the Sahel drought. *Geophysical Research Letters*, **27**, 795–798.

Wang G. and Eltahir E.A.B. (2002) Impact of CO_2 concentration changes on the biosphere-atmosphere system of West Africa. *Global Change Biology*, **8**, 1169–1182.

Warren S.D., Thurow T.L., Blackburn W.H. and Garza N.E. (1986) The influence of livestock trampling under intensive rotation grazing on soil hydrologic characteristics. *Journal of Range Management*, **39**, 491–495.

Whittaker R.H. and Likens G.E. (1973) Carbon in the biota. In *Carbon and the Biosphere*, CONF 720510, Woodwell G.M. and Pecan E.V. (Eds.), National Technical Information Service: Washington, DC, pp. 281–302.

103: Terrestrial Ecosystems

CHRISTINA L TAGUE[1], LAWRENCE E BAND[2] AND JANET FRANKLIN[1]

[1] *San Diego State University, San Diego, CA, US*
[2] *University of North Carolina, Chapel Hill, NC, US*

The science of ecohydrology centers on understanding the linkages and feedbacks between terrestrial ecosystems and the hydrologic cycle. Within the disciplines of hydrology, ecosystem ecology, and community ecology, a wide range of conceptual, mathematical, and field-based approaches have evolved to address specific ecohydrologic research questions. In this article, we discuss key elements of these ecohydrologic approaches. We begin by focusing on conceptual frameworks that are organized at the patch level, which define spatially homogeneous units. We compare a set of models in terms of the specific hydrologic and ecologic processes considered and the degree of coupling between them. We then address issues of landscape heterogeneity and compare approaches to the resolution of spatial patterns of soil, topography, vegetation, and atmospheric conditions and the representation of the interactions and feedbacks between neighboring patches. The hierarchical structure and function of terrestrial ecohydrologic systems involve processes varying over very different time and space scales. Finally, we discuss challenges and next steps in developing an integrated framework that can address the range of spatial and temporal scales implied by hydrologic feedbacks on both shorter-term ecosystem material and energy cycling and longer-term ecosystem structure and evolution.

INTRODUCTION

Ecohydrology is an interdisciplinary area of research concerned with the interactions between the hydrologic cycle and ecological processes. Water as a resource, media, reactant, and geomorphic agent is central to many ecosystem functions. These include the influence of water availability on ecosystem material cycling, solute and sediment transport, productivity, trophic systems, ecological community patterns and dynamics, as well as direct and indirect controls of disturbance (e.g. floods, mass wasting, drought, fire). In turn, ecological systems influence water cycling through their control of processes including canopy interception, evapotranspiration, infiltration, and flow resistance. This area has evolved over a period of decades from a set of research themes including forest hydrology and ecology, agricultural water management, acid rain impacts on watersheds, ecosystem budgets of small catchments, biogeography, and aquatic ecology (Bonell, 2002). A set of recent reports has cited the importance of understanding how the mutually regulating cycles of water, carbon, and nutrients (WCN) evolve under different climates and geomorphic conditions as well as disturbance regimes (e.g. NRC, 1998, 1999a; Gupta *et al.*, 2000; WCSG, 2001). Coupling ecological to hydrologic systems through their integration with biogeochemical cycling is a key interdisciplinary area required to address a series of important scientific and societal issues including nonpoint source pollution, freshwater availability, impacts of increased nitrogen deposition and the response of terrestrial and aquatic ecosystems to land use and climate change (WCSG, 2001). Differences in disciplinary conceptual representation of hydrologic and ecological states and processes have been a barrier to effective integration (Benda *et al.*, 2002), which needs to be addressed as part of a new research paradigm.

In this article, we address multiple approaches to understand the geophysical and biological processes of interest to predominantly physical science disciplines (hydrology, geomorphology), community-landscape ecology, and ecosystem science. Our emphasis is on the interactions between hydrologic processes with terrestrial ecological patterns of vegetation and soil disturbance. In particular, we compare how conceptual and operational models in each discipline

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

typically represent interaction between physical and biological processes, and spatial relationships within networks of patches in the landscape. Patches are generally defined as the fundamental spatial unit of study. Patches may be grid cells or irregular shapes. The scale and rules used to define patches vary across disciplines and for specific applications in each discipline. We discuss current integrated approaches that consider feedbacks between WCN cycling and ecosystem structure and function at the patch to landscape and watershed scales, and then explore potential "next steps". While this area necessarily explores coupling and feedback with carbon and nutrient cycling, we do not discuss the details of biogeochemical transformations, transport, or export, which are covered in **Chapter 96, Nutrient Cycling, Volume 3**. Additional, detailed treatment of various aspects of ecohydrology can be found in the recent texts by Baird and Wilby (1999), Tenhunen and Kabat (1999), and Eagleson (2002).

CONCEPTUAL OVERVIEW OF ECOHYDROLOGIC SYSTEMS

Conceptual frameworks in the disciplines of hydrologic and ecosystem science use a systems approach to define stores, fluxes, and transformations of water in the case of hydrology, and WCN in the case of ecosystem cycling, within specific control volumes. The control volumes are defined as a soil body, a canopy, a patch (combining soil and canopy), hillslope, water body, or watershed. Material stores are the amount of a given substance within a control volume (ML^{-3}, where M are units of mass, and L are units of length), while fluxes are the rate of transfer of materials across an interface into, out of, and between control volumes ($ML^{-2}T^{-1}$). Transformations are considered as *in situ* processes occurring within a control volume ($ML^{-3}T^{-1}$). Each of these terms is linked and constrained by mass balance.

In community and disturbance ecology, the concept of stores and fluxes translates into representation of patterns of organisms (e.g. population counts, density, diversity, composition) and the processes that alter those patterns. Ecological patch dynamics (Pickett and White, 1985) incorporates the spatial and temporal evolution of patch patterns by the slow processes of recruitment, growth, aggradation, and succession, as well as disturbance processes involving a rapid change of patch conditions or creation of new patch patterns. The patch dynamics processes of recruitment, mortality, and growth of individual organisms and populations are linked as components in carbon and nutrient budgets. In human dominated ecosystems, these processes are augmented by import, export, and management of carbon and nutrient stocks.

Table 1 shows a set of key state and process variables frequently of interest in each of the three disciplines of hydrology, ecosystem ecology and community/landscape ecology. The control volume is implicitly taken to be at the scale of an ecological patch with a vertical extent from the top of the canopy to some datum below the surface, often the maximum root depth. The hydrologic flux processes in Table 1 include the major input/output components contributing to the canopy and soil water balance of a patch. The three sets of state and process variables either have direct overlap or indirect overlap by mutual regulation and mass conservation. Biogeochemical cycling or ecosystem ecology focuses on short to long-term fluxes and transformation of water and nutrients. Each of the carbon and nitrogen cycling processes interacts, either directly or indirectly, with the storage and flux of water within the patch. In this sense, ecohydrologic coupling emphasizes soil moisture as a major control of ecosystem processes including the role of water as

1. a limiting resource influencing moisture stress
2. a transport mechanism controlling the supply of other limiting resources such as dissolved nutrients
3. an influence on rates of biogeochemical transformations through regulation of soil chemical conditions and microbial populations

Short-term (e.g. daily or subdaily) flux and transformation processes redistribute mass between atmosphere, canopy, litter, and soil. The carbon stores fixed by canopy photosynthesis at the patch scale form the base of the trophic hierarchy, providing feedback through primary productivity to ecological state variables and processes in the right-hand column such as herbivory, predation, decomposition, and mineralization by terrestrial and soil organisms. Long-term aggradation and change of canopy structure and composition, litter decomposition, and soil development in turn feed back to short-term WCN cycling. Water storage and flux at the patch level also influence disturbance processes with effects on fire ignition and spread, erosion, and mass wasting. Therefore, interactions between hydrologic processes, biogeochemical cycling (ecosystem ecology), and community structure and function involve feedbacks across a range of timescales.

Beyond the scale of individual patches, networks of patches exchange water, nutrients, carbon, genetic material (e.g. seeds, pollen), microbial communities, and are in turn grazed by mobile fauna. Some of the exchange is concentrated along hydrologic flowpaths (Figure 1) including engineered flowpaths for water supply, drainage, and transportation infrastructure. Lateral transport along these flowpaths has a significant impact on space/time soil moisture, nutrient concentrations, and seed dispersal, providing a significant influence on ecological patch dynamics. However, a substantial amount of patch dynamics exchange is independent of hydrologic flowpaths, and includes atmospheric transport, and transport by fauna

Table 1 State and process variables commonly used within the three disciplines of hydrology, ecosystem ecology, and community/disturbance ecology

	Hydrology	Ecosystem ecology	Community and landscape ecology
Stores/Patterns	***Above Ground*** Surface water storage in canopy, litter, detention stores (e.g. rooftop, soil surface, impervious surface)	***Above Ground*** Physiologic WCN in foliar and stem biomass, litter water, carbon, and nitrogen content	***Above Ground*** Patch community characteristics – species composition, abundance, diversity, age and size class, fine fuel load and flammability
	Subsurface Unsaturated zone soil moisture, rooting zone soil moisture, groundwater depth, and storage	***Subsurface*** Fine and coarse root carbon and nitrogen, soil organic matter (SOM), microbial C and N pools	***Subsurface*** Soil organism abundance, diversity, taxonomic composition, functional types, soil shear strength
Fluxes/Dynamics	***Vertical***	***Vertical***	***Vertical***
	Precipitation, interception, evaporation, transpiration, infiltration, exfiltration, throughfall, stemflow, runoff, unsat-sat drainage, capillary rise	Photosynthesis, respiration, allocation (above ground and below ground), exudate production, solute leaching, litterfall, mortality, root turnover, decomposition, mineralization, other *in situ* BGC transformations	Reproduction, seed production, recruitment, growth[a], competition, mortality, disturbance (e.g. fire, windstorm)
	Lateral Hortonian, and saturation excess overland flow, through flow, groundwater flow	***Lateral*** Transport and export of particulate and dissolved C and N constituents	***Lateral*** Seed dispersal, organism movement/migration, disturbance propagation (e.g. fire, pathogen, and insect spread), flood, mass wasting

[a]maturation, change in age structure.

and human populations, with each providing detectable feedbacks to both hydrologic and ecosystem processes. This provides multiple, overlapping patterns and space and timescales of patch dynamics, leading to a hierarchical structuring of ecosystem form and change (Wu and Loucks, 1995).

ECOHYDROLOGIC INTERACTIONS AT THE PATCH SCALE

The concept of the ecological patch, delimited by the top of canopy, a defined depth into the soil (often the rooting zone) and bounded by the extent of a vegetation stand (or other land cover), has been used as a basic unit and control volume (or linked set of control volumes) to study the vertical exchange of mass, energy, and momentum and forest species dynamics. Most often, the patch as control volume is accepted as a conceptual framework at a range of spatial scales, without consideration of internal heterogeneity of soils, canopy or atmospheric conditions, or the lateral exchange of water, carbon, and nutrients. The patch paradigm has been applied to a wide range of features and scales ranging from research plots and small catchments, to grid cells used to represent land-atmosphere interaction in regional or global atmospheric circulation models. At all these scales, soil moisture availability is an important regulator of canopy water and carbon exchange with the atmosphere, with water shortage resulting in stomatal closure and in more significant stress events by a drop in leaf display. The actual mechanisms by which soil water conditions influence leaf stomata may not be fully understood (Roberts, 2000), but explicit or implicit representation of this control is one of the most fundamental mechanistic links between hydrologic and ecosystem processes. A set of soil vegetation atmosphere transfer (SVAT) models (discussed below) has been developed in the hydrologic, atmospheric, and ecological disciplines that incorporate some level of soil moisture control on canopy carbon and water cycling, commonly simulating vertical exchange of energy and mass, and neglecting lateral flux. Many field investigations have been patterned on this conceptual model. Trade-offs in detail of soil and canopy representation in terms of number and arrangement of stores, flux, and transformation processes and the simplicity of use, analysis, and measurement distinguish many of these investigations.

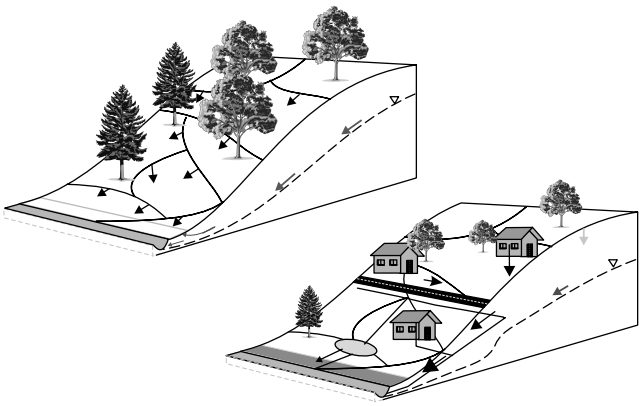


Figure 1 Hillslope ecohydrologic system showing a network of ecosystem patches connected through natural and engineered flowpaths. Note the riparian patches (base of hill) that are maintained by recharge from above. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

A simple method of illustrating first-order links between hydrologic and ecological systems can be presented by inspecting the Penman–Monteith combination equation (Monteith, 1965) for transpiration:

$$LE = \frac{\Delta R_{\mathrm{n}} + \gamma C_{\mathrm{p}} \rho_{\mathrm{a}} g_{\mathrm{a}} VPD}{\rho_{\mathrm{a}} \lambda (\Delta + \gamma (1 + g_{\mathrm{a}}/g_{\mathrm{c}}))} \quad (1)$$

where LE is the latent heat-flux of transpiration, Δ is the slope of the temperature-saturation humidity curve, R_{n} is canopy net radiation, γ is the psychrometric constant, C_{p} is the heat capacity of the air, g_{a} is the canopy aerodynamic conductance, ρ_{a} is the air density, VPD is the atmospheric vapor pressure deficit, λ is the latent heat of vaporization and g_{c} is the canopy conductance. Canopy radiation transmission and absorption are influenced by leaf form, albedo, canopy leaf area index (LAI), and canopy clumping. The canopy conductance term is the leaf conductance integrated over the LAI, and is influenced directly or indirectly by a set of environmental factors of VPD, soil water availability, absorbed photosynthetically active radiation (PAR), and temperature. Both of the conductance terms and the net radiation in the Penman–Monteith equation are dependent on patch level canopy structure, species composition, stand age, and soil conditions as discussed below. The stomatal conductance is regulated by photosynthetic assimilation rates, which are subject to leaf temperature, absorbed PAR, the atmosphere to leaf CO_2 gradient, and enzymatic limitations due to nitrogen availability (Farquhar *et al.*, 1989). Consequently, transpiration rates and photosynthetic assimilation are tightly linked, with both a function of similar biotic and abiotic patch conditions. Therefore, a set of simplified approaches to ecohydrologic linkages have used reductions in canopy and species specific evapotranspiration rates compared to potential evapotranspiration as a surrogate for limitations on productivity. Further discussion of hydrologic controls on transpiration can be found in **Chapter 42, Transpiration, Volume 1**.

Stochastic Water Budget Approaches

Eagleson (1978a,b,c; 1982a,b) introduced an approach based on the coupling of transpiration and productivity assuming soil moisture stress as a rate-limiting factor. In a classic set of papers, Eagleson developed stochastic models of a relatively simple, one-dimensional soil water balance operating at the daily time step with a two-zone soil moisture store characterized by the unsaturated zone and a deep groundwater table. This approach was designed to derive the temporal distribution function of root zone soil moisture conditional on prevailing climate, soil, and plant characteristics. Climate conditions were characterized by probability density functions of interstorm periods and daily precipitation, with prescribed interstorm potential bare soil evaporation and canopy evapotranspiration rates that are modified by available soil moisture. These models allowed the optimization of vegetation type, density, and canopy properties (including root depth) to minimize water stress, and thus maximize productivity, for different vegetation life forms under prevailing climate and soil conditions.

The approach of deriving stochastic models of soil water balance to investigate impacts on water-limited ecosystems has been further developed by Rodriguez-Iturbe and others (e.g. Rodriguez-Iturbe, 2000; Porporato *et al.*, 2002). With an emphasis on water-limited ecosystems, a series of articles extended the approach of using probabilistic one-dimensional water balance conditional on climate, soil, and vegetation characteristics to investigate a range of ecosystem dynamics. The probabilistic approach allows the derivation of both seasonal mean root zone soil moisture conditions, as well as the frequency of water-stress events. Multispecies systems are incorporated by specifying different rooting depths, maximum rates of evapotranspiration, and different thresholds for the onset of water-stress limitations to evapotranspiration. By necessity, these approaches emphasize simplified process representation of water balance in order to develop analytical solutions to the water balance stochastic differential equations incorporating observed frequency distributions of storm and interstorm characteristics as driving functions of ecosystem water stress. Following Eagleson's original work, in most applications ecological response is assumed to be based on an optimization of species type and cover to minimize soil moisture stress, and do not directly address carbon budgets or nutrient availability. However, recently D'Odorico *et al.* (2003) and Porporato *et al.* (2003) have extended the stochastic framework with carbon and nitrogen cycling expressions conditional on soil moisture levels to incorporate the development of soil organic matter (SOM). These models emphasize optimal states of the canopy (assuming

the water-stress optimization principle) to soil conditions and stationary growing season precipitation statistics. As such, the approach does not address transient response to climate or disturbance perturbations, although interesting research addressing the effects of interannual variability in precipitation regime has been used to investigate multimodal soil moisture regimes and resulting shifts in optimal vegetation conditions (D'Odorico *et al.*, 2000).

These stochastic methods offer a data and computationally efficient approach, based on physical principles, for analysis of the influence of daily to interannual variability in climate conditions on ecosystem optimal structure and behavior. By assuming an optimal ecosystem state, as canopy cover or carbon and nitrogen cycling stores, however, these approaches do not extend to investigate long-term growth and aggradation that treat system transience due to disturbance. The stochastic optimization approaches do not incorporate patch dynamical evolution by growth, succession, or mortality. To address long-term feedback to local moisture conditions, as a patch or population of patches develop, representation of transient response rather than optimal state would need to be included.

Canopy Vegetation Atmosphere Transfer Models

In contrast to this approach that emphasizes simplicity in canopy and soil representation, and time-averaged optimal response to precipitation regime, Deardroff (1978) and others developed more complex one-dimensional models of SVAT of energy and water including plant canopy impacts on radiative transfer and multilayer canopy and soil representation. These models are forced by meteorological time series and are operated as land surface process (LSP) models coupled to an atmospheric model (typically at the resolution of the atmospheric model grid), or off line using prescribed meteorological data. The level of process detail incorporated into these models precludes analytical solutions, but allows the representation of more complete coupling of water and energy carbon cycling. In addition to representation of energy, water, and momentum exchange, ecosystem photosynthetic and respiration processes have been incorporated into the biological components and controls of surface exchange (e.g. Dickinson, 1986; Sellars *et al.*, 1986). Resolution of stomatal physiology and stand structural effects on radiation and turbulent exchange (e.g. height, leaf and plant area, leaf form, foliar and soil albedo) produce a greater dependency on both physiological and aerodynamic characteristics of the vegetation. These models typically emphasize short-term (sub-diurnal to seasonal) coupling of atmosphere with the surface, typically with a prescribed vegetation layer. While incorporating assimilation and respiration fluxes of carbon between atmosphere, canopy, and soil, these models do not generally operate over a long enough period to simulate ecosystem aggradation and structural or compositional change. Exceptions include those LSP models that incorporate crop growth models with explicit phenology (e.g. Tsvetsinskaya *et al.*, 2001). For additional discussion, *see* **Chapter 104, Satellite-Based Analysis of Ecological Controls for Land-Surface Evaporation Resistance, Volume 3**.

Ecosystem Carbon Budget and Ecological Community (Gap) Models

An additional set of one-dimensional models of patch level processes were developed in the ecosystem community that incorporate integrated water, carbon, and nutrient cycling at the plot to watershed scale (e.g. Aber and Federer, 1992; Parton *et al.*, 1987; Running and Coughlan, 1988; Running and Hunt, 1993). These models were designed to operate with comprehensive climate data from meteorological stations or climate models, and to resolve diurnal to seasonal patterns, but with the ability and emphasis on long-term ecosystem growth ranging to decadal or centennial timescales. Consequently, time steps typically range from daily to monthly and capture soil moisture dynamic effects on a more complete cycling of carbon and nutrients subject to soil, vegetation type, and climate conditions. In this case, the atmosphere has typically been prescribed, so that there is no feedback from the surface physiological and hydrologic conditions to the atmosphere. These models follow an ecosystem, rather than population or community ecology approach, with an emphasis on water, carbon, and nutrient pools within a stand, instead of representation of individual organisms. While they do track ecosystem aggradation of biomass, as well as allocation of carbon to above ground and below ground pools, there is typically a simple, prescribed mortality loss rate, and only rudimentary treatment of competition and compositional change. However, by solving a full carbon budget with explicit treatment of transient evolution of organic matter pools, some aspects of dynamic changes in one or more patches are incorporated. This includes aggradation of litter over long time periods, as well as daily scale adjustments of litter moisture, with potential use for computing fire ignition and spread probabilities, discussed below.

Patch models arising from the ecological, rather than hydrologic community, include models based on populations of individual trees (stems) growing under a set of water, temperature, nutrient, and light constraints and simulate size class and species dynamics of the stand, but without the detailed representation of WCN flux and transformation processes (e.g. Botkin *et al.*, 1972; Shugart, 1984; Pastor and Post, 1986). These gap models concentrate on stem recruitment, germination, mortality, and growth as influenced by simple water (and other resource) availability reduction factors that are typically set as patch characteristics, although more detailed treatment of patch soil

moisture conditions and other constraints have improved the ability of these models to replicate biogeographic and temporal variability in stand dynamics (e.g. Pacala *et al.*, 1996; Bonan, 1989).

Hybridization of the different model approaches discussed above has been carried out in the last decade. Several researchers have sought to directly link the strengths of the ecophysiological and gap succession models (e.g. Friend *et al.*, 1993; Huston, 1991; Keane *et al.*, 1996), directly linking short-term water and carbon cycling processes with the long-term growth, aggradation, and succession of vegetation (typically forest) patches. Mesoscale and global atmospheric circulation models have been coupled with both ecophysiological and community ecology models (e.g. Foley *et al.*, 1998; Pielke *et al.* 1998) to address land use and climate change effects on land–atmosphere coupling, including shifts in water availability.

All of the above patch models emphasize vertical exchange of water, carbon, and nitrogen. Eagleson (2002) and others recognized that the one-dimensional approach fails to adequately represent hydrologic behavior for a set of conditions in which significant lateral flux of water is a critical element of system behavior, such as in riparian areas. A number of researchers have criticized the assumptions of the one-dimensional approach, stating that both modeling and measurement of lateral divergence of water, nutrients, carbon, and energy are critical components of many ecosystems (e.g. Band *et al.*, 1993; Wood, 1999). While the magnitude of vertical flux of moisture may substantially exceed lateral flux at and below the surface, persistent gradients in moisture conditions developed by lateral flux produces repeated patterns and variance in ecohydrologic behavior over the landscape. In scaling from patch models to watersheds and landscapes, techniques for including this spatial heterogeneity become an important part of ecohydrology studies.

SMALL CATCHMENT RESEARCH

Much of our knowledge base on ecosystem and hydrologic process interactions has been developed from monitoring and experimentation at the scale of a small research plot or below. A major challenge to the community is to learn how system behavior at this level can be linked to landscape or regional level systems. Small catchment research presents an empirical approach to scaling up to regions by investigating the behavior of these units as basic elements. As first-order catchments typically comprise more than half the area of a watershed and a substantial amount of the total stream length, an important question is whether understanding the dynamics of the population of headwater catchments provides sufficient information to understand the behavior of large watersheds. Certainly, downstream processes in higher order streams as well as alluvial floodplain sites that have unique ecohydrologic interactions are not incorporated in this research approach.

Small watershed monitoring and experiments originated with the assumptions of the plot paradigm implicit in their design, simply substituting the full catchment as a metapatch. This approach parallels the lumped hydrology paradigm common in many hydrologic models and allows a conceptualization of the catchment with uniform or dominant soil and vegetation attributes, and the assumption of uniform (measured) precipitation and atmospheric deposition inputs, as well as measured streamflow outputs. This allows for estimation of catchment evapotranspiration rates and nutrient sinks and sources through mass balance. Most small catchment experiments were designed assuming that the net groundwater exchange was minimal, although in many cases it has been found that groundwater may be a significant and unmeasured output. In the United States, the Department of Agriculture and the National Science Foundation have set up a series of experimental watersheds over the last several decades to investigate the interactions between forest management and catchment hydrology. These include a set of the current Long-Term Ecological Research (LTER) sites, including the Coweeta Hydrologic Laboratory, HJ Andrews Experimental Forest, Hubbard Brook as well as the Agricultural Research Service (ARS), US Geological Survey (USGS) and other agency catchments. Similar experimental watersheds have been operated in other countries and used to investigate effects of vegetation management on water yield and quality.

A set of planned and inadvertent experiments in vegetation manipulation and disturbance effects on water yield have been carried out focusing on water use of different species and life forms, as well as ecophysiological conditions of the canopy. Emphasis has been placed on the effects of forest management (timber rotation, species conversion, age class distributions) and disturbance (e.g. wildfire) on water yield and quality. Langford (1976) documented long-term reductions in catchment water yield when old growth Mountain Ash Eucalypt stands were replaced by young stands following wildfire in the Melbourne water supply catchments. Reduction in stand transpiration and photosynthetic rates in old growth Douglas Fir has also been documented in the Pacific Northwest (Yoder *et al.*, 1994), with important implications for regional water supply as old growth is harvested. These reductions have been attributed to a set of factors, including reduced hydraulic conductance in stem tissue, reduced leaf area, and changes in stand structure. Forest conversion to grass cover in the Coweeta Hydrologic Laboratory was shown to increase water yield over the long term, although transpiration rates and water yield were sensitive to nutrient availability, with fertilized grass cover water yields returning close to those of the original forest cover (Hibbert, 1969), linking photosynthetic to transpiration rates. Conversion of mature broadleaf cover

to white pine in a paired catchment experiment at Coweeta has shown persistent declines in annual water yield over a period of decades (Swank *et al.*, 2001). Experiments in the San Dimas Experimental Forest in Southern California converted chaparral to grass cover, to increase water yield. While water yield did increase substantially with the conversion to the more shallow rooted grass cover, reduction in soil strength by decay of the chaparral roots coupled with increased soil moisture led to massive increases in mass wasting and sedimentation from these catchments (Rice *et al.*, 1969). Similar increases in both water yield and mass wasting have been documented following timber harvest, with reduction in soil strength due to loss of root cohesion cited as a key factor (Ziemer, 1981; Roering *et al.*, 2003).

There is also a substantive history of small catchment experiments linking hillslope hydrologic processes and ecosystem biogeochemical cycling (reviewed in **Chapter 96, Nutrient Cycling, Volume 3**)

Valuable compendiums of the long-term research linking hydrologic and ecosystem processes in small catchments in the Coweeta and Hubbard Brook LTER sites can be found in Swank and Crossley (1988) and Likens and Bormann (1995).

APPROACHES TO SCALING ECOHYDROLOGIC INTERACTIONS TO THE LANDSCAPE

Scaling from homogeneous patches to heterogeneous landscapes raises two issues. First, the distribution of patch characteristics within the landscape as a control on aggregate response must be considered, as a set of carbon and nutrient cycling processes have a nonlinear response to soil moisture and canopy conditions. Second, the potential for interactions resulting from gradient driven fluxes of water, nutrients, and organic matter between heterogeneous patches must be explored. The former source of heterogeneity includes local edaphic factors such as soil properties, topography, and microclimate, while the latter source is dependent on the drainage sequence of patches arrayed over the landscape. More general discussion of techniques for representing heterogeneity in hydrology can be found in **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**.

Grid-based Mosaic Approaches

A first-order approach to incorporate landscape heterogeneity in many terrestrial ecosystem process models has been the implementation of ecosystem process models for a cellular grid of patches. Running *et al.* (1989) developed a grid-based implementation of FOREST-BGC (Running and Coughlan, 1988) to simulate carbon and water cycling over a 1200 km^2 area of western Montana over which there are strong gradients in topographically modulated temperature, precipitation and insolation. Similar applications have been implemented for other ecosystem models (e.g. PNET Aber *et al.*, 1995). Similarly, in atmospheric LSP models, characterizing heterogeneity in surface processes by distributed parameterization of spatial grid cells has been termed the *mosaic approach* (Avissar and Pielke, 1989). In these implementations, there is no transfer of water and nutrients from one location to another such that persistent recharge and discharge zones, and patterns of relative soil moisture, are not maintained. Therefore, these approaches capture spatial variation in ecosystem function due to local patch conditions, but do not necessarily capture variation in ecohydrologic dynamics due to position within hydrologic flowpaths. In this respect, the distribution of patches within the landscape could be reshuffled with no impact on aggregate behavior.

Incorporating Hillslope Hydrology to Ecosystem Models

Hillslope scale catenary patterns between soil, water, and vegetation form and process are evident in many ecosystems and critical to the behavior of the full catchment. The partial or variable source area concept in catchment research (Hewlett and Hibbert, 1967; Dunne and Black, 1970) was developed from small catchment research by recognition of distinct and repeatable patterns of soil moisture, maintained by the lateral (and vertical) redistribution of water within hillslopes. Landscape scale spatial patterns of soil moisture produce complex interactions between soil water content, canopy physiology, and ecologically limiting processes of water stress, root aeration, and nutrient availability. As soil water patterns are forced by combinations of climate, soil, and substrate hydraulic properties and topography, a dynamic zonation of ecological limiting conditions develops, which is strongly tied to site geomorphology. Recognition of this zonation is often implicit in ecosystem process studies that focus on particular features such as riparian areas in terrestrial systems.

To extend patterns of terrestrial ecosystem patterns and process to the landscape level necessitates coupling with hydrologic models (conceptual and quantitative) of spatial patterns of moisture conditions. These approaches have ranged from the use of simple soil water indices such as the TOPMODEL topographic/soil wetness index (Beven and Kirkby, 1979) and TOPOG (O'Loughlin, 1981, 1986), to progressively more detailed representation of vertical and lateral unsaturated and saturated zone moisture exchange along hillslope flowpaths including the Distributed Hydrology Soil Vegetation Model – DHSVM (Wigmosta *et al.*, 1994), the Regional HydroEcological Simulation System – RHESSys (Band *et al.*, 1993; Nemani *et al.*, 1993), Dynamic TOPMODEL (Beven *et al.*, 2001),

and MIKE-SHE (Refsgaard and Storm, 1995). For more extended treatment, *see* **Chapter 14, Hydroinformatics and its Contributions to Hydrology: From Computation to Communication, Volume 1**.

Additional coupling of carbon and nutrient balances, as well as long-term growth and aggradation of canopy and soil conditions have been implemented in RHESSys (Mackay and Band, 1997; Tague and Band, 2004; Band *et al.*, 1993, 2001) and TOPOG (Vertessy *et al.*, 1996), as well as Macaque (Watson *et al.*, 1999) and Stieglitz *et al.* (2003). These systems were designed to operate over years to decades to reflect the longer timescales involved in organic matter cycling and ecosystem aggradation. Therefore, these approaches allow for the development of persistent spatial patterns in carbon and nutrient dynamics following hydrologic flowpaths, as well as recovery from disturbance events such as logging, fire, and road construction (e.g. Watson *et al.*, 1999; Tague and Band, 2001).

Linking Hillslope Hydrology with Community Ecology

Community ecology studies seek to understand the dynamics of ecosystem structure and function from a species perspective rather than that of biogeochemical cycling. Studies of the distribution of species assemblages and community characteristics along resource gradients, including soil moisture, have been pursued for decades (Whittaker, 1973). Biogeographic and community studies commonly use hydroclimatologic indices as an environmental variable in multivariate analysis of community species patterns or analysis of species-environment relationships (Franklin, 1995; Guisan and Zimmerman, 2000; Wilson and Gallant, 2000). The models assume equilibrium between biotic distributions and environment, and the hydrologic variables are typically static, and are used to summarize geomorphic (topographic wetness) and climate controls (radiation and temperature indices) on soil moisture variation. As with coupled hydrologic ecosystem process approaches, the use of hillslope hydrologic models can be used to extend these models by adding a dynamic hydrologic component. Spatially distributed hydrologic model predictions of average soil moisture conditions can then be related to species and community patterns, for example, TOPMODEL (Ostendorf and Reynolds, 1998) and DYNWET (Wilson and Gallant, 2000) for soil moisture, as can models of topographically distributed solar insolation related to evaporative demand (Dubayah and Rich, 1995).

Meentenmeyer and Moody (2002) used dynamic simulations of water balance around complex watersheds in Southern California to compute both average and extreme stress conditions as controls on chaparral species distribution. Using RHESSys with a prescribed crop (uniform, low height, and LAI) to remove local effects of variable canopy conditions, they mapped the frequency of extreme stress events defined as predawn leaf water potentials below 5 MPa. (Figure 2). Sampled chaparral community composition showed that species were well arrayed along this stress gradient according to physiological thresholds for embolism (Figure 3). In addition, disturbance-dependent species frequency increases along this gradient, likely reflecting increased fire frequency or intensity.

In landscape ecology, a more dynamic approach has been taken to modeling communities. Spatially explicit models focus on disturbance regimes and community dynamics. In many cases, time steps of these models preclude temporally dynamic hydrology (Mladenoff and Baker, 1999). Landscape models of community dynamics that are not primarily driven by hydrology nonetheless usually characterize the landscape in terms of gradients of moisture availability and nutrients, for example, Mladenoff and He (1999), and other references in Mladenoff and Baker (1999). This characterization of site conditions, whether it is spatially continuous or discrete, is used to determine potential species response (probability of establishment, growth rate, competitiveness), and acts as a template for community dynamics.

Incorporating Disturbance and Management

A major goal of ecohydrologic research is to link community-based ecological concepts such as trophic level, biodiversity, succession, and disturbance with a storage-flux framework that represents the distribution of significant sources, sinks, and transformations of water, carbon, and nutrients within hydrologic flowpaths. This goal has implications for

1. the selection of specific ecohydrologic processes and feedbacks to consider in the development of models and field experiments,
2. resolution of mixed time and space scales of interest,
3. landscape representation, from both a conceptual and operational model standpoint. The models discussed above have typically emerged from the individual disciplines of hydrology, ecosystem process ecology, or community ecology. While interactions between ecology and hydrology are included, none of these models have been designed to integrate a storage-flux WCN framework with community-based ecologic concepts nor do they fully explore the integration of hydrologic flowpaths as both a driver of ecological gradients and as a response to long-term ecosystem evolution.

Representation of disturbance is limited in most ecohydrologic approaches. While community-based ecologic models (e.g. Lake, 2000; Poff *et al.*, 1997) rely on disturbance frequencies, the use of hydrologic models to predict flood and to a lesser extent fire frequencies has primarily been done off-line (where each model is run

Figure 2 Topographic distribution of extreme water-stress frequency in the mission canyon watershed. Disturbance-free species dominate mesic shaded slopes (a), ephemeral drainages (b), and moderately sheltered, high elevation slopes (c). Disturbance-dependent species dominate exposed slopes at low to midelevations (d) and ridges (e). Mixed stands form on exposed high elevation slopes (f) (Reproduced from Meentenmeyer and Moody (2002), by permission of Opulus Press AB)

separately) and do not include feedbacks between vegetation structure and dynamics and hydrologic control of disturbance extent and frequency (*see also* **Chapter 119, Land Use and Landcover Effects on Runoff Processes: Forest Harvesting and Road Construction, Volume 3** and **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**).

For example, Keane *et al.* (1996) used the ecosystem carbon and water cycling of FOREST-BGC to develop fuel loads, and linked carbon budget information to a growth and succession model. However, given the timescales of the succession model, fuel load component moisture conditions were prescribed. Additional linkage of the litter layer moisture balance to the fire ignition and spread model would provide spatial dynamics based on forest canopy, soil, and topographic modulation of the litter. Similarly, applications of models of fire–vegetation species interactions such as LANDIS (Landscape Disturbance Simulator) have incorporated static predictions of soil moisture spatial characteristics (Franklin, 2003). These approaches could be

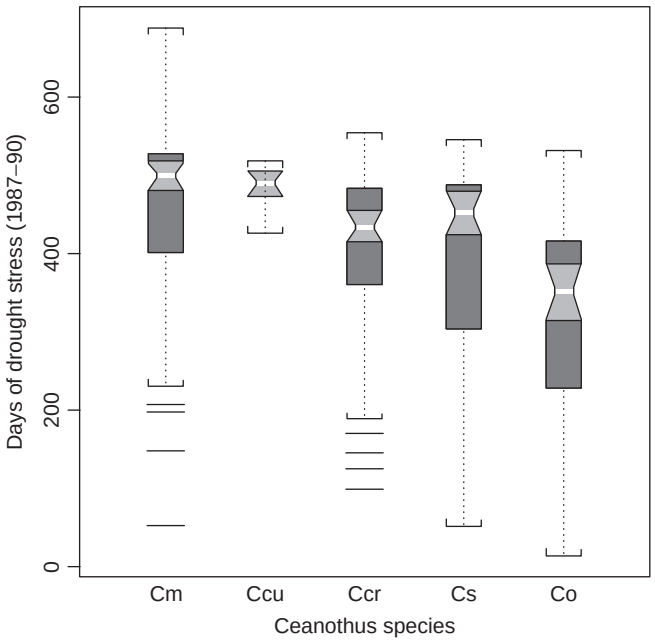


Figure 3 Plot of ceonothus species arranged by their lab-measured susceptibility to soil water tension against ''days of drought stress'' gradient. *Ceanothus megacarpus* (Cm) has 100% xylem embolism at a soil water tension of −13.8 MPa; *C. cuneatus* (Ccu) at −12.9 MPa; *C. crassifolius* (Ccr) at −11.6 MPa; *C. spinosus* (Csp) at −9.2 MPa; and *C. oliganthus* (Co) at 08.2 MPa (Reproduced from Meentenmeyer and Moody (2002), by permission of Opulus Press AB)

extended by coupling with dynamic models of soil and litter moisture patterns that include feedbacks from vegetation, disturbance, and hydrology.

NEXT STEPS: INTEGRATION WITH SOCIETAL PROCESSES AND APPLICATIONS TO ECOSYSTEM MANAGEMENT

Our ability to understand and effectively address critical societal issues ranging from nonpoint source pollution, the maintenance of freshwater supplies, storm water, soil erosion, and sedimentation, and resulting impacts on habitat and biodiversity is dependent on developing a framework to integrate hydrologic science with the structure and function of terrestrial ecosystems at the watershed scale. The development of solutions to these problems also requires that we understand their generation as societal issues, and the interaction of human society as part of the watershed ecosystem. In extending to more fully integrated ecohydrologic frameworks at the landscape level, the selection of appropriate scales of process and coupling representation needs to address the range of scales over which different controlling variables and human regulatory instruments and management techniques are organized. For example, recent literature on terrestrial ecosystem nutrient cycling emphasize fine scale (m) nexus of WCN cycling or "hot spots" and "hot moments" within riparian zones linked to microbiologic response to wetting and drying cycles (McClain *et al.*, 2003). The dynamics of these patches cannot be explained in isolation. At the landscape level, these "hotspots" are a response to space/time variation of hillslope flowpath and stream flow dynamics, which in turn respond to a combination of hydrogeomorphic changes that impact biogeochemical cycling and community structure/function in these zones. Further incorporation of community dynamics to reflect species composition within the riparian zone may necessitate consideration of disturbance regimes such as flooding, fire, insect infestation, or policy and regulation (e.g. riparian buffers) that occur at scales well beyond that of the patch or hillslope. Groffman *et al.* (2002, 2003) have investigated current riparian zone nitrogen retention in urban areas as a function of larger scale development processes, including historical catchment scale land cover and channel change, and the evolution of drainage infrastructure (*see also* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**).

The example of regulatory and infrastructure effects on riparian zones highlights the significance of building in explicit consideration of human individual and institutional activity as an integral part of the ecohydrologic system. Just as ecological communities respond and provide feedback to hydrologic systems, so do human communities by water supply, drainage, fertilization, vegetation management, and the development of various types of best management practices and ecological restoration efforts.

A set of models has begun to integrate patch dynamic processes at overlapping patterns and scales of hydrologic ecosystem and community ecology, with human community and regulatory regimes. The RHESSys model represents landscape elements and processes by progressively nesting watersheds, climate zones, hillslopes, patches, and vegetation strata as a containment hierarchy, with different hydrologic, ecological, climatic, and social/regulatory processes and features defined for specific levels (Band *et al.*, 2000; Tague and Band, 2001). This framework incorporates human management activity as well as disturbance events that can reset ecohydrologic system state differentially within different levels and elements of the landscape hierarchy. DeAngelis *et al.* (1998) linked a spatially explicit ecosystem process model with a multilayer trophic model including herbivores and predators for an area of the Florida Everglades with water management options. Costanza and Voinov (2004) have presented a set of spatially explicit grid-based ecosystem models linked to land use change models for human dominated ecosystems. These models attempt to integrate land value derived from spatial econometric models as a driver of land use change. A general unit ecosystem

model is then parameterized for each grid cell by local land use, and embedded within a grid-based hydrologic routing scheme to estimate watershed scale integration of water, nutrient, and carbon dynamics.

While these issues of heterogeneity and process complexity present challenges for the ecohydrologic community, they also present opportunities for advancement in understanding how ecohydrologic systems evolve and respond to change. Within the developing framework of "hydrological patch dynamics", treating the patch mosaic as a structured set of ecosystem units arranged along flowpaths, a set of questions with both important theoretical and practical implications can be posed:

1. How do ecohydrologic systems coevolve in response to specific disturbance, regulatory and management regimes?
2. How do the spatial arrangements of patch type and conditions impact on ecosystem behavior at the watershed scale, and are there specific arrangements that are more resilient to expected disturbances?
3. Do specific spatial patterns of patch conditions optimize ecohydrologic functions, such as landscape level water yield or retention, ecosystem productivity, or nutrient retention?

Resolution of these issues would address landscape scales at which environmental management is focused (e.g. subdivision, planning district) and promote ecohydrologically based restoration design and implementation.

Acknowledgments

Development of concepts presented in this chapter was partially supported by funding from the National Science Foundation, DEB-97-14835, and BCS-0095796. We are grateful to Aaron Moody for discussion of his research and contribution of figures, and to Peter Groffman for a careful reading and critique, which significantly improved the manuscript.

FURTHER READING

Benda L., Poff N.L., Miller D., Dunne T., Reeves G., Pess G. and Pollock M. (2004) The network dynamics hypothesis: how channel networks structure riverine habitats. *Bioscience*, **54**, 413–427.

Dubayah R. (1994) Modeling a solar radiation topoclimatology for the rio grande river basin. *Journal of Vegetation Science*, **5**, 627–640.

Rodriguez-Iturbe I., D'Odorico P., Porporato A. and Ridolfi L. (1999) On the spatial and temporal links between vegetation, climate and soil. *Water Resources Research*, **35**, 3709–3722.

Waring R. and Running S.W. (1998) *Forest Ecosystems: Analysis at Multiple Scales*, Academic Press: San Diego.

Yeakley J.A., Swank W.T., Swift L.W., Hornberger G.M. and Shugart H.H. (1998) Soil moisture gradients and controls on a southern Appalachian hillslope from drought through recharge. *Hydrology and Earth System Sciences*, **2**, 41–49.

REFERENCES

Aber J.D. and Federer C.A. (1992) A generalized, lumped parameter model of photosynthesis, evapotranspiration and net primary production in temperate and boreal forest ecosystems. *Oecologia*, **92**, 463–474.

Aber J.D., Ollinger S.V., Federer C.A., Reich P.B., Goulden M.L., Kicklighter D.W., Melillo J.M. and Lathrop R.G. Jr (1995) Predicting the effects of climate change on water yield and forest production in the Northeastern U.S. *Climate Research*, **5**, 207–222.

Avissar R. and Pielke R.A. (1989) A parameterization of heterogeneous land-surface for atmospheric numerical models and its impact on regional meteorology. *Monthly Weather Review*, **117**, 2113–2136.

Baird A.J. and Wilby R.L. (Eds.) (1999) *Eco–Hydrology: Plants and Water in Terrestrial and Aquatic Environments*, Routledge: London, p. 402.

Band L.E., Patterson P., Nemani R. and Running S.W. (1993) Ecosystem processes at the watershed scale: incorporating hillslope hydrology. *Agricultural and Forest Meteorology*, **63**, 93–126.

Band L.E., Tague C.E., Brun S.E., Tenenbaum D.E. and Fernandes R.A. (2000) Modeling watersheds as spatial object hierarchies: structure and dynamics. *Transactions in GIS*, **4**, 181–196.

Band L.E., Tague C.E., Groffman P.S. and Belt K. (2001) Ecosystem processes at the watershed scale: Hydrological and ecological controls of nitrogen export. *Hydrological Processes*, **15**, 2013–2028.

Benda L., Poff N.L., Tague C., Palmer M.A., Pizzuto J., Cooper S., Stanley E. and Moglen G. (2002) How to avoid train wrecks when using science in environmental problem solving. *Bioscience*, **52**, 1127–1136.

Beven K. and Kirkby M. (1979) A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Beven K.J. and Freer J. (2001) A dynamic TOPMODEL. *Hydrological Processes*, **15**, 1993–2012.

Bonan G.B. (1989) Environmental factors and ecological processes controlling vegetation patterns in boreal forests. *Landscape Ecology*, **3**, 111–130.

Bonell M. (2002) Ecohydrology – a completely new idea? *Hydrological Sciences Journal*, **47**, 809–810.

Botkin D.B., Janak J.F. and Wallis J.R. (1972) Some ecological consequences of a computer model of forest regrowth. *Journal of Ecology*, **60**, 849–872.

Costanza R. and A. Voinov (Eds.) (2004) *Landscape Simulation Modeling: A Spatially Explicit, Dynamic Approach*, Springer: New York.

DeAngelis D.L., Gross L.J., Huston M.A., Wolff W.F., Fleming D.M., Comiskey E.J. and Sylvester S.M. (1998) Landscape modeling for Everglades ecosystem restoration. *Ecosystems*, **1**, 64–75.

Deardroff J.W. (1978) Efficient prediction of ground temperature and moisture with inclusion of a layer of vegetation. *Journal of Geophysical Research*, **83**, 1889–1903.

Dickinson R.E., Henderson-Sellars A., Kennedy P.J. and Wilson M.F. (1986) *Biosphere-Atmosphere Transfer Scheme for the NCAR Community Climate Model*, Technical Report NCAR/TN-275+STR, NCAR, Boulder, p. 69.

D'Odorico P., Laio F., Porporato A. and Rodriguez-Iturbe I. (2003) Hydrologic controls on soil carbon and nitrogen cycles. II a case study. *Advances in Water Resources*, **26**, 59–70.

D'Odorico P., Ridolfi L., Porporato A. and Rodriguez-Iturbe I. (2000) Preferential states of seasonal soil moisture: the impact of climate fluctuations. *Water Resources Research*, **36**, 2209–2220.

Dubayah R. and Rich P.M. (1995) Topographic solar radiation for GIS. International. *Journal of Geographic Information Systems*, **9**, 405–419.

Dunne T. and Black R.D. (1970) Partial area contributions to storm runoff in a small New England watershed. *Water Resources Research*, **6**, 1296–1311.

Eagleson P. (2002) *Ecohydrology*, Cambridge University Press, p. 443.

Eagleson P.S. (1978a) Climate, soil, and vegetation, 1. Introduction to water balance dynamics. *Water Resources Research*, **14**, 705–712.

Eagleson P.S. (1978b) Climate, soil, and vegetation, 4. The expected value of annual evapotranspiration. *Water Resources Research*, **14**, 731–740.

Eagleson P.S. (1978c) Climate, soil, and vegetation, 6. Dynamics of the annual water balance. *Water Resources Research*, **14**, 749–764.

Eagleson P.S. (1982a) Ecological optimality in water limited natural soil-vegetation systems, 1. Theory and hypothesis. *Water Resources Research*, **18**, 325–340.

Eagleson P.S. (1982b) Ecological optimality in natural soil-vegetation systems, 2. Tests and applications. *Water Resources Research*, **18**, 341–354.

Farquhar G.D., von Caemmerer S. and Berry J.A. (1989) A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species. *Planta*, **149**, 79–90.

Foley J.A., Levis S., Costa M.H., Cramer W. and Thompson S.L. (1998) Coupling dynamic models of climate and vegetation. *Global Change Biology*, **4**, 561–579.

Franklin J. (1995) Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, **19**, 474–499.

Franklin J. (2003) Clustering versus regression trees for determining ecological land units in the southern California mountains and foothills. *Forest Science*, **49**, 354–368.

Friend A.D., Shugart H.H. and Running S.W. (1993) A physiology based gap model of forest dynamics. *Ecology*, **74**, 792–797.

Groffman P.M., Bain D.J., Band L.E., Belt K.T., Brush G.S., Grove J.M., Pouyat R.V., Yesilonis I.C. and Zipperer W.C. (2003) Down by the riverside: urban riparian ecology. *Frontiers in Ecology*, **1**, 315–321.

Groffman P.M., Bouware N.J., Zipperer W.C., Pouyat R.V., Band L.E. and Colosimo M.F. (2002) Soil nitrogen cycle processes in urban riparian zones. *Environmental Science and Technology*, **36**, 4547–4552.

Guisan A. and Zimmerman N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Gupta V.K., Duffy C., Grossman R., Krajewski W., Lall U., McCaffrey M., Milne B., Pielke R. Sr, Reckhow K., Duke University, Swanson F. (2000) *WEB: A Framework for Reassessment of Basic Research and Educational Priorities in Hydrologic Science*. WEB: **http://cires.colorado.edu/hydrology/**.

Hewlett J.D. and Hibbert A.R. (1967) Factors affecting the response of small watersheds to precipitation in humid regions. In *Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 275–290.

Hibbert A.R. (1969) Water yield changes after converting a forested catchment to grass. *Water Resources Research*, **5**, 634–640.

Huston M.A. (1991) Use of individual based forest succession models to link physiological whole tree models to landscape scale ecosystem models. *Tree Physiology*, **9**, 293–306.

Keane, R.E., Morgan P. and Running S.W. (1996) *FIRE-BGC – A Mechanistic Ecological Process Model For Simulating Fire Succession on Coniferous Forest Landscapes of the Northern Rocky Mountains*, USDA Forest Service, Intermountain Research Station, Research Paper INT-RP-484.

Lake P.S. (2000) Disturbance, patchiness and diversity in streams, Journal of North. *American Benthological Society*, **19**, 573–592.

Langford K.J. (1976) Changes in yield of water following a bushfire in a forest of Eucalyptus regnans. *Journal of Hydrology*, **29**, 87–114.

Likens G.E. and Bormann F.H. (1995) *Biogeochemistry of a Forested Ecosystem*, *Second Edition*, Springer-Verlag: New York, p. 159.

Mackay D.S. and Band L.E. (1997) Ecosystem processes at the watershed scale: Dynamic coupling of distributed hydrology and canopy growth. *Hydrological Processes*, **11**, 1197–1217.

McClain, M.E., Boyer E.W., Dent C.L., Gergel S.E., Grimm N.B., Groffman P.M., Hart S.C., Harvey J.W., Johnston C.A., Mayorga E., *et al.* (2003) Biogeochemical hot spots and hot moments at the interface of terrestrial and aquatic ecosystems. *Ecosystems* **6**, 301–312.

Meentenmeyer R.K. and Moody A. (2002) Distribution of plant life-history types in California chaparral: the role of topographically determined drought severity. *Journal of Vegetation Science*, **13**, 67–78.

Mladenoff D.J. and Baker W.L. (1999) *Spatial Modeling of Forest Landscape Change*, Cambridge University Press: Cambridge.

Mladenoff D.J. and He H.S. (1999) Design, behavior and application of LANDIS, as object-oriented model of forest landscape disturbance and succession. In *Spatial Modeling of Forest Landscape Change: Approaches and Applications*, Mladenoff D.J. and Baker W.L. (Eds.), Cambridge University Press: Cambridge, pp. 125–162.

Monteith J. (1965) Evaporation and environment. *Symposia of the Society for Experimental Biology*, **29**, 205–234.

National Research Council (1998) *Hydrologic Sciences: Taking Stock and Looking Ahead*, National Academy Press: Washington, p. 138.

National Research Council (1999a) *Hydrologic Science Priorities for the US Global Change Research Program*, National Academy Press: Washington, p. 34.

Nemani R., Running S.W., Band L.E. and Peterson D.L. (1993) Regional hydroecological simulation system: An illustration of the integration of ecosystem modeling with a GIS. In *Environmental Modeling with GIS*, Gloodchild M.F., Parks B.O. and Steyart L.T. (Eds.), Oxford University Press, pp. 296–304.

O'Loughlin E.M. (1981) Saturation regions in catchments and their relations to soil and topographic properties. *Journal of Hydrology*, **53**, 229–246.

O'Loughlin E.M. (1986) Prediction of surface saturation zones in natural catchments by topographic analysis. *Water Resources Research*, **22**, 794–804.

Ostendorf B. and Reynolds J.F. (1998) A model of arctic tundra vegetation derived from a topographic gradient. *Landscape Ecology*, **13**, 187–201.

Pacala S.W., Canham C.D., Sapanora J., Silander J.A., Kobe R.K. and Ribbens E. (1996) Forest models defined by field measurements: Estimates, error analysis and dynamics. *Ecological Monographs*, **66**, 1–43.

Parton W.E., Schimel D.S., Cole C.V. and Ojima D.S. (1987) Analysis of factors controlling soil organic matter levels in the Great Plains grasslands. *Soil Science Society of America Journal*, **51**, 1173–1179.

Pastor J. and Post W.M. (1986) Influence of climate, soil moisture and succession on forest carbon and nitrogen cycles. *Biogeochemistry*, **2**, 3–27.

Pickett S.T.A. and White P.A. (1985) *The Ecology of Natural Disturbance and Patch Dynamics*, Academic Press: New York.

Pielke R.A., Avissar R., Raupacj M.R., Dolman H., Zeng X. and Denning S. (1998) Interactions between the atmosphere and terrestrial ecosystems: influence of weather and climate. *Global Change Biology*, **4**, 461–475.

Poff N.L., Allan J.D., Bain M.B., Karr J.R., Prestegaard K.L., Richter B.D., Sparks R.E. and Stromberg J.C. (1997) The natural flow regime: a paradigm for river conservation and restoration. *Bioscience*, **47**, 769–784.

Porporato A., D'Odorico P., Laio F., Ridolfi L. and Rodriguez-Iturbe I. (2002) Ecohydrology of water-controlled ecosystems. *Advances in Water Resources*, **25**, 1335–1348.

Porporato A., D'Odorico P., Laio F. and Rodriguez-Iturbe I. (2003) Hydrologic controls on soil carbon and nitrogen cycles. I Modeling scheme. *Advances in Water Resources*, **26**, 45–58.

Refsgaard J.C. and Storm B. (1995) MIKE SHE. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Ranch.

Rice R.M., Corbett E.S. and Bailey R. (1969) Soil slips related to vegetation, topography and soil in southern California. *Water Resources Research*, **5**, 647–659.

Roberts J. (2000) The influence of physical and physiological characteristics of vegetation on their hydrological response. *Hydrological Processes*, **14**, 2885–2901.

Rodriguez-Iturbe I. (2000) Ecohydrology: a hydrologic perspective of climate-soil-vegetation dynamics. *Water Resources Research*, **36**, 3–10.

Roering J.J., Schmidt K.M., Stock J.D., Dietrich W.E. and Montgomery D.R. (2003) Shallow landsliding, root reinforcement, and the spatial distribution of trees in the Oregon Cost Range. *Canadian Geotechnical Journal*, **40**, 237–253.

Running S. and Coughlan J. (1988) A general model of forest ecosystem processes for regional applications I. Hydrologic balance, canopy gas exchange and primary production processes. *Ecological Modeling*, **42**, 125–154.

Running S.W. and Hunt E.R. (1993) Generalization of a forest ecosystem process model for other biomes, BIOME-BGC and an application for global scale models. *Scaling Physiological Processes: Leaf to Globe*, Academic Press.

Running S.W., Nemani R.R., Peterson D.L., Band L.E., Potts D.F., Pierce L.L. and Spanner M.A. (1989) Mapping regional forest evapotranspiration and photosynthesis by coupling satellite data with ecosystem simulation. *Ecology*, **70**, 1090–1101.

Sellars P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A simple biosphere model (SiB) for use within general circulation models. *Journal of Atmospheric Science*, **43**(6), 505–531.

Shugart H. (1984) *A Theory of Forest Dynamics: The Ecological Implications of Forest Succession Models*, Springer-Verlag: New York.

Stieglitz M., Shaman J., McNamara J., Engel V., Shanley J. and Kling G.W. (2003) An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport. *Global Biogeochemical Cycles*, **17**, 1105.

Swank W.T. and Crossley D.A. Jr (1988) Forest hydrology and ecology at Coweeta. *Ecological Studies*, Springer-Verlag: New York.

Swank W.T., Vose J.M. and Elliott K.J. (2001) Long-term hydrologic and water quality responses following commercial clearcutting of mixed hardwoods on a southern Appalachian catchment. *Forest Ecology and Management*, **143**, 163–178.

Tague C. and Band L. (2004) RHESSys: Regional Hydro-ecologic simulation system: An object-oriented approach to spatially distributed modeling of carbon, water and nutrient cycling. *Earth Interactions*, **8**(19), 1–42.

Tague C.L. and Band L.E. (2001) Simulating the impacts of road construction and forest harvesting on hydrologic response. *Earth Surface Processes and Landforms*, **26**, 135–151.

Tenhunen J.D. and Kabat P. (1999) *Integrating Hydrology, Ecosystem Dynamics, and Biogeochemistry in Complex Landscapes*, John Wiley & Sons: New York, p. 367.

Tsvetsinskaya E., Mearns L.O. and Easterling W. (2001) Investigating the effect of seasonal plant growth and development in 3-dimensional atmospheric simulations. Part II: atmospheric response to crop growth and development. *Journal of Climate*, **14**, 711–729.

Vertessy R.A., Hatton T.J., Benyon R.J. and Dawes W.R. (1996) Long term growth and water balance predictions for a mountain ash (E. regnans) forest subject to clearfelling and regeneration. *Tree Physiology*, **16**, 221–232.

Watson F.G.R., Vertessy R. and Grayson R.B. (1999) Large scale modelling of forest hydrological processes and their long term effect on water yield. *Hydrological Processes*, **13**, 689–700.

WCSG (2001) *Plan for a New Science Initiative on the Global Water Cycle*, Chair: G. Hornberger, U.S. Global Change Research Program: Washington.

Whittaker R.H. (1973) Direct gradient analysis. In *Handbook of Vegetation Science 5: Ordination and Classification of Communities*, Whittaker R.H. (Ed.), Junk: The Hague, pp. 9–50.

Wigmosta M., Vail L. and Lettenmaier D. (1994) Distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, **30**, 1665–1679.

Wilson J. and Gallant J. (2000) *Terrain Analysis: Principles and Applications*, John Wiley & Sons: New York.

Wood E.F. (1999) The role of lateral flow: Over or underrated? In *Integrating Hydrology, Ecosystem Dynamics, and Biogeochemistry in Complex Landscapes*, Tenhunen J.D. and Kabat P. (Eds.), John Wiley & Sons, pp. 197–215.

Wu J. and Loucks O.L. (1995) From balance-of-nature to hierarchical patch dynamics: A paradigm shift in ecology, *Quarterly Review of Biology*, **70**, 439–466.

Yoder B., Ryan M.G., Waring R.H., Schoettle A.W. and Kaufmann M.R. (1994) Evidence of reduced photosynthetic rates in old trees. *Forest Science*, **40**, 513–527.

Ziemer R.R. (1981) Roots and the stability of forested slopes. In *Erosion and Sediment Transport in Pacific Rim Steeplands*, Davies T.R.H. and Pearce A.J. (Eds.), International Association of Hydrological Sciences Publication No. 132, IAHS Press: Wallingford, Oxon, pp. 343–361.

104: Satellite-Based Analysis of Ecological Controls for Land-Surface Evaporation Resistance

STEVEN W RUNNING AND JOHN S KIMBALL

Department of Ecosystem and Conservation Sciences, NTSG University of Montana, Missoula, MT, US

Lack of available water constrains ecological processes for two-thirds of the earth's biosphere. These water limitations are manifested as either physical water deficiency or as a chemical unavailability of water as a result of being frozen. This article summarizes global principles of water limitations on the biosphere, and physiological limitations on plants. It presents hydrometeorological principles of evapotranspiration and organizing logic of the soil–vegetation–atmosphere transfer models commonly used to compute evapotranspiration. We then introduce remote sensing from both optical/thermal and active/passive microwave sensors for calculating landscape scale evapotranspiration. Finally, we offer a multisensor-based integrated surface resistance to define landscape water availability under all conditions.

GLOBAL EXTENT OF WATER LIMITATIONS ON THE BIOSPHERE

Water in an available physical–chemical state is a fundamental property of the land surface in Earth systems science. Presence of liquid water is a requirement for life, so the activity of the biosphere is intimately related to the ever-changing conditions of water on the landscape. Water availability is the primary limiting factor for vegetation growth over roughly two-thirds of the Earth (Nemani *et al.*, 2003). At the most general, global scale, the role of water in the biosphere has two dimensions. First, is the abundance of water, or most commonly the limitations of seasonally suboptimal water supply on vegetation physiology. A second consideration is the physical state, or seasonal duration of liquid water on the landscape. Vast areas of the high latitudes have an abundance of water, but that water is in a frozen state for much of the year and largely unavailable for most biological activity.

Two other localized water limitations must be acknowledged before proceeding. Water inundation from flooding or poor topographic drainage exerts a different control on ecosystems, producing anaerobic soil conditions that cause many plants literally to drown from lack of root aeration. Also, water with high salt concentrations occasionally found in desert systems can have osmotic potentials too low for plant tolerance, producing a different kind of chemical water limitation. This article will not deal with these conditions that can be locally severe, but are limited in global scope.

The purpose of this article is to develop a more integrated analysis of primary water limitations on biospheric activity. We can define vegetation by its relative abundance (a forest) or absence (a desert), and by its biological activity, the growing season when the ecosystem biogeochemical cycles are active, as contrasted with the dormant season, when water, carbon, and nutrient fluxes are minimal to zero. In like manner, we can define water also by its activity. The severe absence of water, a drought condition, limits ecosystem activity to a near dormant state until rainfall returns. Water in a frozen state is rather equivalent to vegetation in its dormant state, and in fact freezing of water is a primary cause for vegetation to enter dormancy.

So this more comprehensive measure of water availability to the biosphere requires measuring both the presence of water and its physical frozen or liquid state. The history of remote sensing for these land surface properties

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

is concentrated on optical/thermal and passive microwave sensors for defining water abundance, and active–passive microwave measures for surface freeze–thaw condition. This article will attempt an analysis combining aspects of all these sensors toward a single measure of water mobility on the landscape. We will recommend a multisensor satellite-based algorithm that can provide an integrated global, year-round evaluation of water availability for the biosphere.

Global Biospheric Patterns and Hydrologic Limits

Biogeographers have for decades related global vegetation distribution to broad indices of temperature and water (Walter, 1979). At global scales, the most obvious explanation for the difference in vegetation between deserts and forests was persistence of available water for plant growth. Water availability was most simply described by supply of annual precipitation, but it was quickly apparent that a more refined approach had to include evaporative demand, or a water balance, precipitation minus potential evapotranspiration (PET). Although the logic for use of these water balances as an environmental index is sound, there are many different formulations and temporal domains. PET can be computed in many different ways using radiation, temperature, humidity, wind speed, and other variables for formulations of varying complexity, and computed for time spans from subhourly to annually (Vorosmarty *et al.*, 1998).

Grier and Running (1977) quantified a more explicit functional relationship between climate and vegetation, observing that leaf area indices (LAI) of forests in Northwest America were directly correlated to an annual water balance. LAI is a more useful definition of vegetation than biomass or height because the functional leaf surface area for evapotranspiration, and canopy interception of radiation and precipitation is quantified directly. Nemani and Running (1989b) further developed the soil–vegetation–atmosphere systems logic relating soil water holding capacity and vegetation LAI to meteorological water balances. Similar ideas have also been pursued by Stephenson (1990) and Eagleson (2002).

Woodward (1987) expanded this analysis to global scales, computing a more sophisticated index of vegetation water balance, and using global climate data to predict both biome distributions and LAI. Prentice *et al.* (1992) and Nielson (1995) use similar water balance computations to predict global vegetation distribution patterns, in what are now known as *Dynamic Global Vegetation Models* (DGVMs). Kergoat (1998) introduced remote sensing to calculating global hydrologic equilibriums for predicting LAI, using biweekly NDVI data to follow the vegetation growing season, or phenology. Most recently, Nemani *et al.* (2003) evaluated climatic controls on global vegetation productivity, and determined that over 40% of the Earth's vegetated surface is limited by low water availability, while approximately 33% is limited by cold temperatures and water in it's frozen state, also limiting water availability for plant growth (Figure 1).

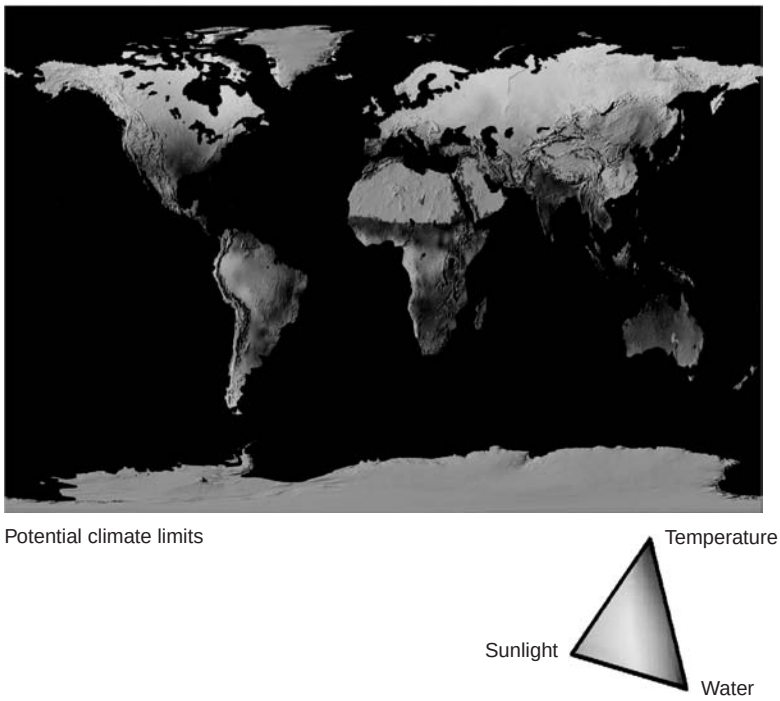


Figure 1 Global analysis of the relative balance of climatic controls on evapotranspiration of vegetation. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

All of these studies relate water limitations to vegetation structural development over relatively long time scales (i.e. decades). We next evaluate how water controls vegetation biophysics and physiology over more immediate daily to seasonal time domains.

Ecological Principles of Water Limits on Plants

Vegetation responds to water deficits in many ways (Waring and Running, 1998). Even mild soil water deficits begin to inhibit cellular expansion, xylem water flow from roots to leaves and phloem sugar transfer in stems of growing plants. Plant water deficits induce progressive leaf stomatal closure, which reduces plant water loss while bringing photosynthesis, transpiration, and canopy-atmosphere gas exchange nearly to a halt. Sustained drought will produce early leaf senescence and shedding, and may impact ecosystem leaf area for a number of years. Additionally, dry soils reduce soil litter decomposition rates, and related nutrient mobilization. Subzero temperatures and frozen water induce the same ecosystem responses, although the biological mechanics are somewhat different; plant cellular growth is inhibited; stomata are closed, limiting photosynthesis, gas exchange, and transpiration; water movement across root membranes, xylem sap flow, soil decomposition, and plant nutrient uptake are severely restricted. When plants are in a fully frost tolerant state, water molecules are hydroscopically bound to cell walls, minimizing ice crystal formation and cellular rupture. It is the lack of these cold tolerant mechanisms that restricts tropical plants to climates that never experience freezing temperatures. Consequently, lack of mobile water, either from the absence of liquid water or from freezing, has similar impacts on reducing plant leaf area, transpiration, growth, and related ecosystem activity.

Hydrometeorological Principles of Vegetation Limits on Evapotranspiration

For meteorological processes, land surface water status and frozen or thawed state determine whether the incident radiant energy is dissipated by heating the air and soil, or by evaporating or sublimating water. From energy balance theory it is clear that:

$$R_n = H + LE + G \tag{1}$$

Net radiation, R_n, being absorbed by the land surface is predominantly partitioned into H, sensible heat, or LE, latent energy evaporating water ($W\,m^{-2}$), while energy loss through soil conductance and photosynthesis, G, are relatively minor components of the energy budget (<5%) over vegetation so are often ignored. Thus, the summary Bowen ratio:

$$B = \frac{H}{LE} \tag{2}$$

can quantify the wide range of energy partitions found from completely dry surfaces where B approaches infinity, that is, all net radiation is translated into heat, to $B = 0$, an open water surface where nearly all net radiation is used to evaporate water. Quantifying the available water at the land surface and the resistance to surface evaporation from soil and vegetation is key to computing LE.

For hydrologic processes, evaporation is quantified as a component of the land surface water budget,

$$PPT = ET + Q + \Delta S \tag{3}$$

where PPT, ET, Q, and ΔS represent precipitation, evapotranspiration, outflow, and a change in water storage ($kg\,m^{-2}$) per unit time, respectively. The storage term, ΔS, includes both surface and subsurface water storage, including water in the subsurface soil profile that ultimately drains to streamflow or groundwater recharge.

For ecological processes, soil moisture is important for determining litter decomposition rates and soil CO_2 evolution, and for providing available liquid water to plants for transpiration demands. Ultimately, therefore, the availability of water for land surface evaporation is a fundamental constraint and integrator of meteorological, hydrological, and ecological processes.

The most widely accepted formula for computing land surface evaporation is the Penman–Monteith (P–M) equation, which combines the key meteorological drivers for evaporation, net radiation, humidity, and windspeed, with the key surface attributes that control evaporation rates. The surface attributes are defined as resistances, one biologically mediated by leaf area and leaf stomatal dynamics, the other physically related quantifying canopy roughness and aerodynamic exchanges. The P–M equation computes latent energy, or λE

$$\lambda E = \frac{\Delta \cdot r_e \cdot (R_n - G_0) + \rho C_p \cdot (e_{sat} - e)}{r_e \cdot (\gamma + \Delta) + \gamma \cdot r_i} \tag{4}$$

where e and e_{sat} are actual and saturation vapor pressure respectively; γ is the psychrometric constant, and Δ is the rate of change of saturation vapor pressure with temperature (i.e. $\partial e_{sat}(T)/\partial T$); r_i is the bulk surface internal resistance and r_e is the external or aerodynamic resistance. In the above equation, it is assumed that the turbulent exchange characteristics for heat and vapor transfer are the same. The Penman–Monteith equation is strictly valid only for a closed vegetation canopy; however, the formula can be adapted for sparse vegetation or even a bare soil surface with properly defined bulk internal diffusion and turbulent transfer resistances.

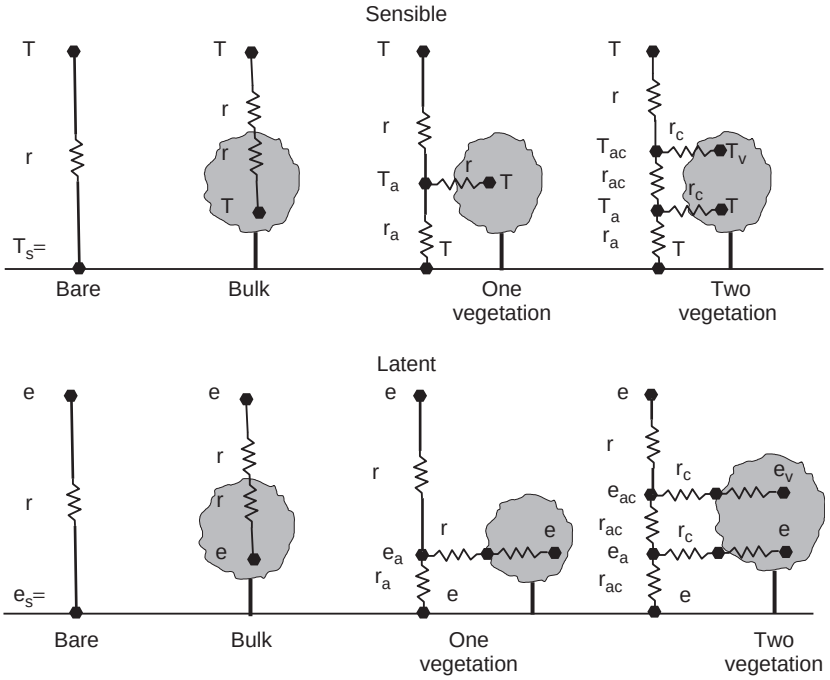


Figure 2 Electrical analog theory used to depict components of meteorology and resistance in transfer of sensible and latent heat from vegetation to the atmosphere. ($T_{a,s,v,g}$ = temperature of air, soil, vegetation, ground; $e_{a,v,g}$ = vapor pressure at atmosphere, vegetation, ground; $r_{a,c,s}$ = resistances at aerodynamic, canopy, surface) (Reproduced from Bonan, 2002 by permission of Cambridge University Press)

Systems Analysis of Soil–Vegetation–Atmosphere Transfer

Critical to computing accurate surface-atmosphere energy transfer and evaporation rates is the representation of various constraints in the soil–vegetation–atmosphere system. These constraints have classically been defined as resistances using electrical analogue theory (Figure 2). Surface resistance includes a variety of surface properties that impede the evaporation of water from the surface, and can incorporate both physical and biological impediments. For example, vegetation height exerts an aerodynamic roughness that decreases resistance, soil litter exerts a bulk diffusion resistance and vegetation physiological stomatal control increases surface resistance as stomata close in reaction to water deficits. Although the P–M equations were derived from electrical analogue theory incorporating resistances, we can simply transform these resistances to their reciprocal, surface conductance, with the advantage that surface conductances are then directly, rather than inversely proportional to the final evaporation rates.

Beginning in the 1970s these concepts were organized into a variety of SVAT (Soil–Vegetation–Atmosphere Transfer) models (Waring and Running, 1998). Some of these SVAT models were incorporated within atmospheric General Circulation Models, GCMs, like the BATS Biosphere Atmosphere Transfer Scheme of Dickinson (1996) and SiB Simple Biosphere model (Sellers *et al.*, 1986). The Project for Intercomparison of Land Surface Models analyzed the various logics and structures for these models (Pitman, 2003). Other SVAT models were designed to work in ecological biogeochemistry modeling such as FOREST-BGC (Running and Coughlan, 1988) or in hydroecological modeling (Band *et al.*, 1993).

A simple SVAT model flowchart is shown in Figure 3 (Bonan, 2002). Critical hydrologic components include precipitation, snowpack, and soil water storage, surface, and subsurface runoff. These topics are covered in detail elsewhere and in other articles. For this article, we now focus on evaporation and transpiration processes, and the control of these rates by the surface resistance. The total, or bulk surface resistance, integrates aerodynamic, soil, and canopy resistances to evaporation, and the physiological control of stomatal dynamics to vegetation transpiration. These resistance components are often dealt with separately, but landscape evapotranspiration rates are constrained by this total surface resistance.

REMOTE SENSING PRINCIPLES FOR EVAPOTRANSPIRATION

Remote sensing provides a particularly valuable methodology for evaluating total surface resistance, because the satellite sensor inherently views the entire land surface, not separating vegetation from soil. The reflectance, emission,

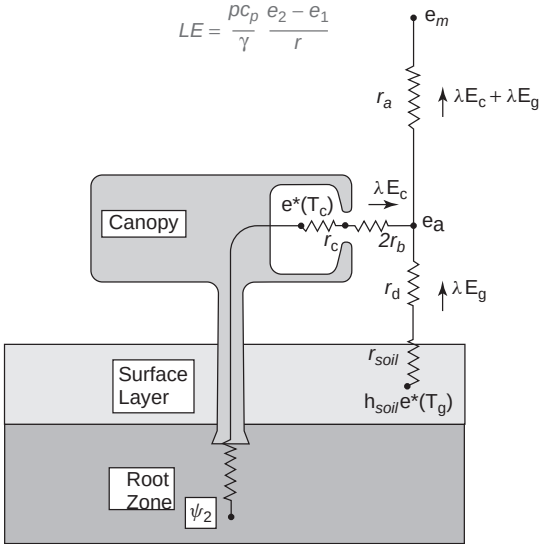


Figure 3 Diagram of a simple SVAT or Soil–Vegetation–Atmosphere Transfer model. ($\lambda E_{c,g}$ = latent energy from canopy, ground; $r_{a,c,b}$ = resistance from aerodynamic, canopy, boundary layer. Other symbols in text) (Reproduced from Bonan, 2002 by permission of Cambridge University Press)

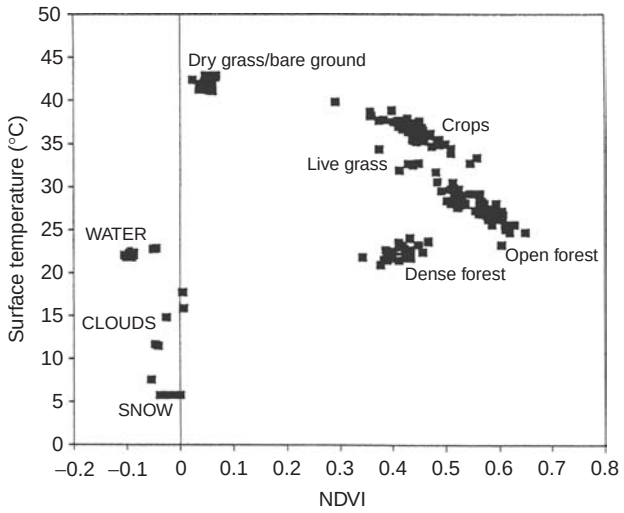


Figure 4 Ranges of radiometric surface temperature LST, and NDVI observed for a 106 000 km^2 landscape of western Montana on 14 July 1987 from AVHRR. Water clouds and snow all have low NDVI and very cool radiometric temperatures compared to the vegetated landscape, with dry bare ground the warmest (Reproduced from Nemani and Running, 1989a by permission of American Meteorological Society)

or backscatter of electromagnetic energy as observed from a satellite sensor provides an integrated view of the landscape dependant on the spatial resolution of the dataset. With any sensor of spatial resolution larger than approximately 5 m, individual plants are not resolved, and a combined view of multiple plants and the surrounding soil or litter surfaces are always included. Rather than attempting to laboriously separate vegetation from soil, or limiting analysis only to completely vegetated or bare soil surfaces, satellite remote sensing lends itself well to evaluating the total landscape resistance to surface-atmosphere energy and mass exchange.

The evaporation theory summarized above is very well developed, and measurement of land surface evaporation and conductance constraints is rather routine for plot scales. The challenge now is to represent regional and continental scales with this theory. The meteorological drivers in the P–M equation are regularly available from mesoscale to global-gridded meteorological models. Representing the extreme spatial and temporal heterogeneity of surface resistances is the biggest limitation for accurate measures of evaporation at landscape to continental scales. At these scales, remote sensing data are readily available, and consequently a variety of sensors have been explored for both direct and indirect assessment of evaporation and associated surface resistances to water and energy movement.

Optical/Thermal Sensors

Optical and thermal sensors provide two key attributes for quantifying land surface evaporation. First, these sensors are sensitive to photosynthetic biomass and surface temperature, and are useful for quantifying vegetation cover and canopy leaf area index. Structural, phenological, and other morphological differences among major vegetation types (e.g. forests, grasslands, croplands) are defined by land cover classes distinguished using optical/thermal sensors like MODIS (Figure 4). Evaluating seasonal trajectories of vegetation indices such as NDVI and land surface temperature (LST) provides an additional dimension of biome discrimination, as more open and arid vegetation types show dramatically higher LST (Figure 5). Global 1 km resolution datasets of general land cover classes that include simple vegetation biome discrimination are now regularly available and are updated annually to reflect land cover changes (Friedl *et al.*, 2002). Higher resolution land cover data are also available based on 30-m resolution Landsat data, allowing more precise vegetation cover discrimination (Vogelmann *et al.*, 2001). Newer generation land cover mapping methods avoid logical classification errors by computing continuous fields of fractional vegetation cover that are then directly interpretable for defining surface roughness, albedo, and LAI for evaporation calculations (Hansen *et al.*, 2002).

NIR/red spectral band ratios such as the NDVI are widely used to estimate vegetation LAI over large areas (Myneni *et al.*, 2002). Increasing LAI provides greater cumulative surface area for canopy water interception and evaporation, and for transpiration by vegetation canopies. Higher LAI also infers more robust vegetation with deeper

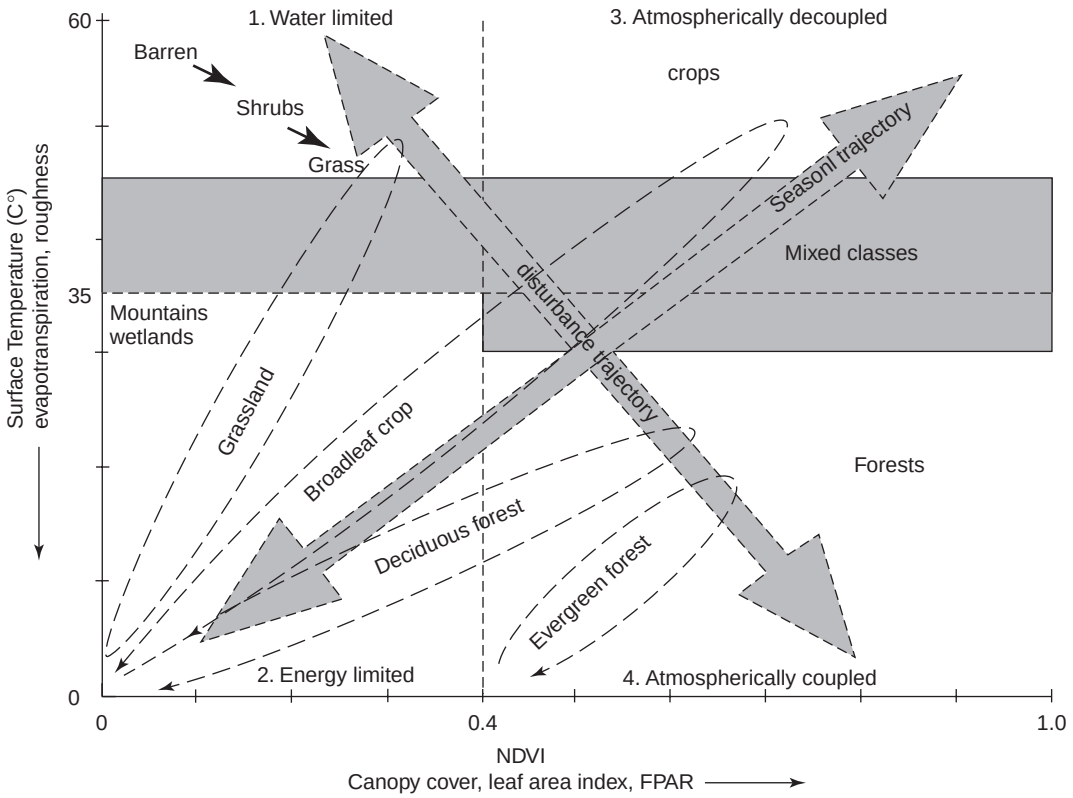


Figure 5 Separation of biome types possible by following the seasonal time series of radiometric surface temperature LST and NDVI. As vegetation grows in the spring NDVI and LST rise, but at different rates depending on the surface energy exchange characteristics of the vegetation. As vegetation desiccates in late summer, or with disturbances, NDVI falls and LST rises (Reproduced from Nemani and Running, 1997 by permission of Ecological Society of America)

rooting zones that are able to tap water deeper in the soil profile, thus allowing an indirect measure of deeper soil water availability than surface observations alone can directly provide.

Thermal remote sensing of radiometric land surface temperature, LST, most directly measures the sensible heat component of the energy balance, and is thus inversely proportional to latent energy and evaporation rates (Wan *et al.*, 2002). As previously discussed, the Bowen ratio (H/LE) is a relatively simple parameter summarizing the relationship between sensible and latent heat flux from a surface. Land surfaces can vary from a very high Bowen ratio when completely dry, often more than 10, to as low as 0.1 when wet. These changes in land surface energy partitioning can occur slowly as a vegetated landscape dries out over many weeks during summer, and then abruptly as the Bowen ratio drops following a substantial rainfall event. Thermal remote sensing can provide an integrated look at land surface evaporation, although overpass timing is critical, as midafternoon radiant heating of the land surface provides the most useful signal. Historically, the NOAA AVHRR sensor with 1430 local overpass time has been most valuable, and now the MODIS sensor on the 1330 Aqua platform is providing high-quality LST data. For some purposes, geostationary satellite data from GOES also can be used to derive LST and surface ET every hour under cloud-free conditions (Norman *et al.*, 2003; Diak *et al.*, 2004).

Many studies have combined both spectral vegetation index and LST information from satellite optical/infrared sensors to infer land surface evaporation rates, typically as a ratio of LST/NDVI (Gillies *et al.*, 1997; Sandholt *et al.*, 2002; Moran *et al.*, 1994; Goward *et al.*, 2002). The denominator, NDVI or equivalent, quantifies potential *increases* in evaporation due to more evaporating surface of LAI, while the numerator, LST infers *decreases* in evaporative conductance from increases in sensible heat. If we examine a single scene of a complex natural landscape in midsummer, a logical pattern of NDVI and LST is discernible (Figure 4, Nemani and Running, 1989a). In this Montana mountainous landscape, clouds and snow are shown with low NDVI and low LST, mesic forests have a high NDVI but rather cool LST, and dry grasses or bare land show low NDVI but high LST. This surface discrimination is strongest during midsummer conditions under relatively high surface solar energy loading with

spatially heterogeneous energy partitioning and associated LSTs. As a surface dries, the LST pattern becomes progressively higher and can easily exceed 60 °C, even with screen-height air temperatures below 40 °C (Figure 6a). By quantifying temporal changes in the slope of the maximum LST/NDVI curve, an inference of surface resistance can be made (Figure 6b; Nemani *et al.*, 1993; Nishida *et al.*, 2003a, 2003b).

However, when we attempt to monitor a region regularly through the vegetation growing season, the most fundamental constraint is the regular interruption by cloudiness and atmospheric aerosols that limit clear imaging of the land surface. In humid tropical areas, cloud cover limits optical/infrared remote sensing more than 80% of the year, and latitudes above 50 degrees also have extended periods of summer cloudiness (Figure 7). For these situations of extended cloudiness and/or inadequate seasonal solar illumination, microwave sensors provide the potential

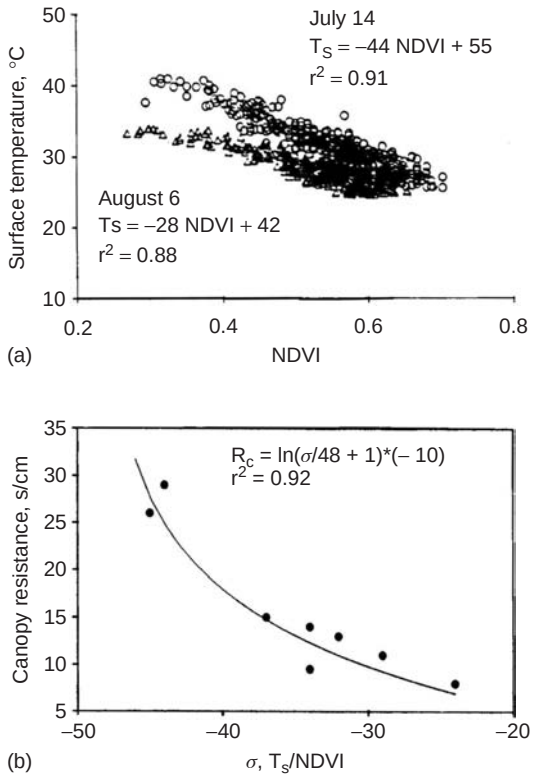


Figure 6 (a) Change in the slope of the LST/NDVI relationship from 14 July, after extended drought, and 6 August, after 3.2 cm of rain, for a natural landscape in western Montana, as observed by AVHRR. A reduction in Bowen ratio is evident in the reduced LST after substantial rainfall. (b) A canopy resistance, simulated by the SVAT model, FOREST-BGC, was highly correlated with seasonal changes in slope of the LST/NDVI relationship, suggesting a way to infer surface resistances by satellite over large regions. (Reproduced from Nemani and Running, 1989a; Nemani *et al.*, 1993, © American Meteorological Society)

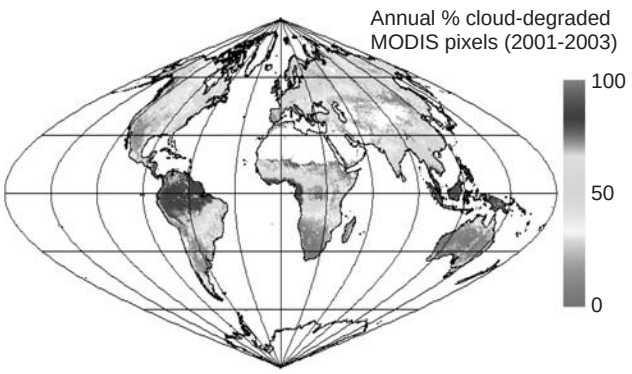


Figure 7 The fraction of annual MODIS land surface optical reflectance data degraded by cloud cover for 2003. Data from Zhao *et al.* (2005). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

for all-weather imaging of the land surface. Remote sensing at microwave wavelengths is largely independent of solar illumination, cloud cover, and other atmospheric attenuation impacts that can significantly degrade remote sensing capabilities at optical and infrared wavelengths. However, microwave wavelengths also provide different information than optical/infrared wavelengths and must be evaluated in a different way.

Active/Passive Microwave Sensors

Microwave remote sensing occurs over much longer spectral wavelengths (0.1 to 30 cm) than optical/NIR remote sensing (0.4 to 10 μm). The signal detected by an active or passive microwave sensor is a function of the emission, absorption, transmission, scattering, and reflectivity of electromagnetic (EM) energy by the landscape and intervening atmosphere at a given spectral wavelength (Ulaby *et al.*, 1986). These properties, in turn, are sensitive to changes in the physical characteristics of the landscape including, surface structure, roughness, orientation, and dielectric properties. At the microwave wavelengths most commonly used for Earth remote sensing, the atmosphere is largely transparent with minimal attenuation of EM energy.

The dielectric constant characterizes the electrical properties and propagation of EM energy in the landscape (El-Rayes and Ulaby, 1987). The interaction of EM energy with a dielectric material has its origin in the response of charged particles to the applied field. The displacement of these particles from their equilibrium positions gives rise to induced dipoles that respond to the applied field. In addition, polar materials contain permanent dipoles caused by the asymmetric charge distribution within the molecules themselves. Water is relatively unique in that it has strong molecular (positive, H^{+2}, and negative, O^{-2}) polarity, exhibiting a dielectric constant that dominates the microwave response of natural landscapes (Kraszewski, 1996). Water molecules

in a liquid state align themselves with an applied EM field, while water molecules in ice are bound in a crystalline lattice and cannot freely rotate, resulting in a substantially lower dielectric constant.

Most natural materials have a dielectric constant ranging from 3 to 8 when dry, while liquid water has a dielectric constant of approximately 80 (Ulaby *et al.*, 1986). Thus, short-term variability in landscape radar backscatter and microwave emissions is to a large extent, a function of changes in the moisture status of vegetation, surface, and soil media. Variations in the predominant frozen or thawed state of the landscape also have a major impact on microwave emissions and reflectivity because frozen water has a very low dielectric constant of approximately 2, which is similar to very dry soil.

In seasonally frozen environments, the most dynamic temporal changes in microwave emissions and reflectivity occur in response to seasonal freeze–thaw transitions

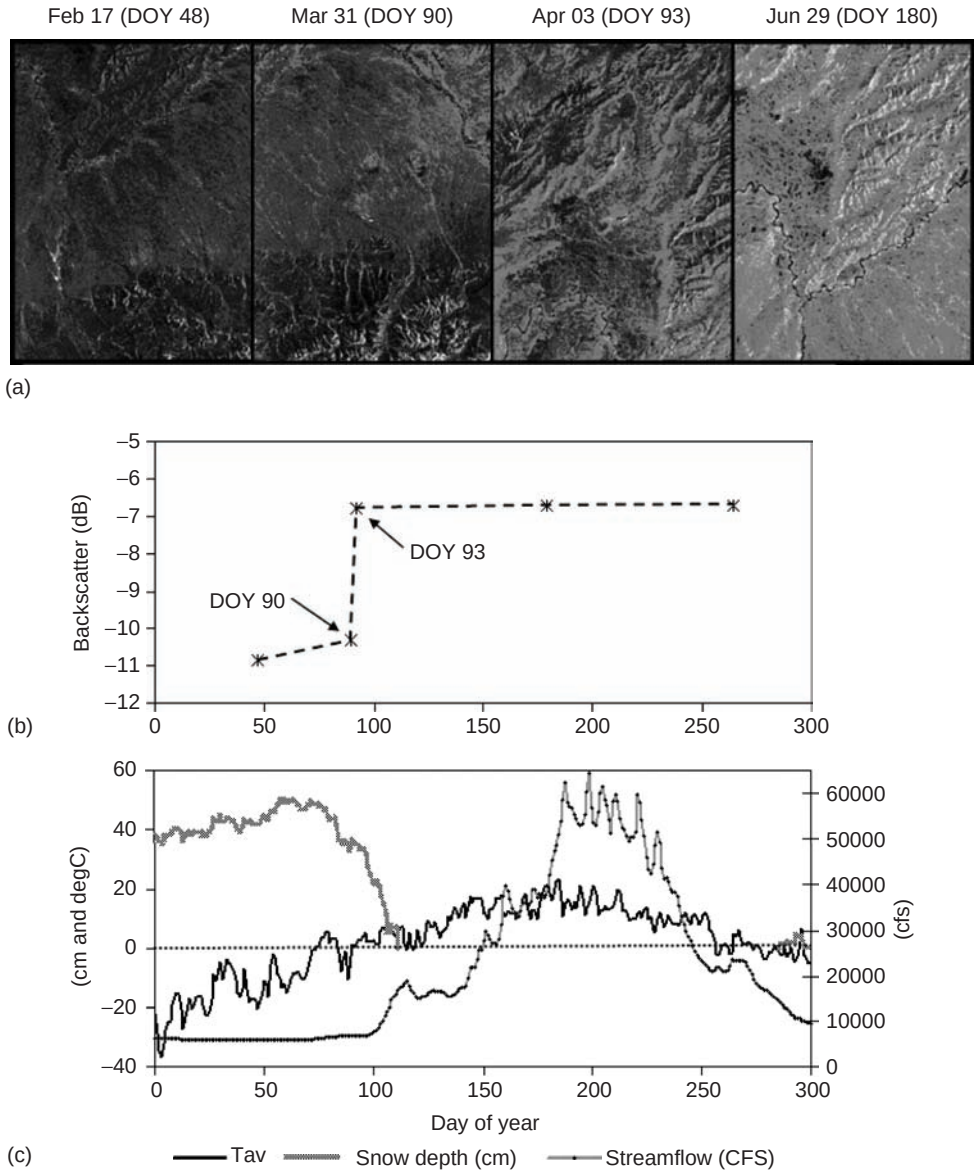


Figure 8 (a)Temporal series of JERS-1 SAR images acquired over interior Alaska boreal forest. The image series shows the landscape radar backscatter transition from predominantly frozen to thawed conditions during spring, 1998. The middle plot (b) shows the radar backscatter temporal response averaged over four sites within the domain. (c) Freeze–thaw transitions dominate the 3–5 dB seasonal variation in radar backscatter. A pronounced backscatter increase between days (DOY) 90 and 93 coincides with thawing air temperatures (T_{av}), seasonal snowmelt and the new release of water on the landscape indicated by local weather station data and USGS streamflow records (redrawn from Entekhabi *et al.*, 2004). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

(Way *et al.*, 1997; Zhang *et al.*, 2003). Spring thaw and the disappearance of seasonal snow cover in these environments is often rapid due to moderate air temperatures and increasing solar radiation, resulting in large decreases in surface albedo of up to 80% between snow covered and snow free conditions (Zhang *et al.*, 2003). The images in Figure 8 show a temporal sequence of JERS-1 satellite L-Band Synthetic Aperture Radar (SAR) images that capture a seasonal freeze–thaw transition event in the boreal forest region of Alaska (Entekhabi *et al.*, 2004). The pronounced increase in radar backscatter between days 90 and 93 occurs in response to seasonal thawing, snowmelt, and the new release of water on the landscape as indicated by seasonal increases in regional stream flow. This event and other freeze–thaw transitions dominate the 5–6 dB seasonal variation in radar backscatter. The surface resistance to evaporation under frozen conditions is similar in magnitude to that of arid, desert environments, while resistance is minimal following snowmelt, with evaporation approaching potential rates. These changes are most dramatic at high latitudes where seasonal increases in surface energy and liquid water available for evaporation have important consequences for regional weather patterns, hydrological, and biospheric processes (Bonan *et al.*, 1995; Chapin *et al.*, 2000; Kimball *et al.*, 2001, 2004).

Active and passive microwave remote sensing techniques have proven sensitive to vegetation phenology, water stress, and other changes in vegetation structure and water content, snow cover, and soil moisture (Waring *et al.*, 1995; Kane *et al.*, 1996; Kimball *et al.*, 2004). The ability of microwave remote sensing to detect changes in the moisture status of these various landscape components is strongly dependent on sensor wavelength, polarization, and spatial resolution. Landscape topography, vegetation structure, soil type, the presence/absence and structure of snow cover also influence retrieval accuracies. Longer wavelengths (e.g. L-band) are generally sensitive to a greater volume of surface vegetation and soil media, relative to shorter (Ku-, C-bands) wavelengths under similar conditions. Longer microwave wavelengths are also directly related to soil moisture as long as the overlying vegetation water content is low (i.e. $<5–6\,kg\,m^{-2}$). Soil moisture sensitivity is also generally limited to the top 5–10 cm of the surface soil layer even in bare soil environments. At biomass levels roughly above that of a fully developed corn crop, however, the ability of microwave remote sensing to detect soil moisture decreases, while sensitivity to vegetation increases.

Vegetation cover is a major impediment limiting direct microwave remote sensing detection of soil moisture over much of the globe. The relatively coarse spatial scales, band widths, and orbital geometries of all current and planned satellite microwave remote sensing platforms also limit capabilities for resolving subgrid scale differences in the moisture content of individual landscape components. However, these scales are generally optimal for regional assessment and monitoring of bulk, landscape surface resistances to evaporation and a fundamental limitation to latent energy and water exchange with the atmosphere.

INTEGRATED REMOTE SENSING OF LANDSCAPE WATER MOBILITY

Regional to Global Scaling

The regional patterns and temporal dynamics of water mobility and surface resistances of the landscape have important consequences for troposphere boundary layer development, weather and global climate. For example, the seasonal transition between frozen (high surface resistance) and nonfrozen (low surface resistance) conditions in the spring is relatively abrupt and coincident with seasonal snowmelt and runoff, large decreases in surface albedo and the initiation of the growing season at high latitudes and upper elevations. The timing and spatial extent of this seasonal shift is of critical importance to the establishment and location of the polar front and regional to global scale weather patterns (Bonan *et al.*, 1995; Chapin *et al.*, 2000). The timing of this frozen/thawed transition is also a major control on vegetation productivity and regional source-sink strength for atmospheric CO_2 (Kimball *et al.*, 2004; McDonald *et al.*, 2004).

At continental scales, satellite-based approaches have been developed to monitor changes in plant available moisture on the landscape (Nemani *et al.*, 1993; Nishida *et al.*,

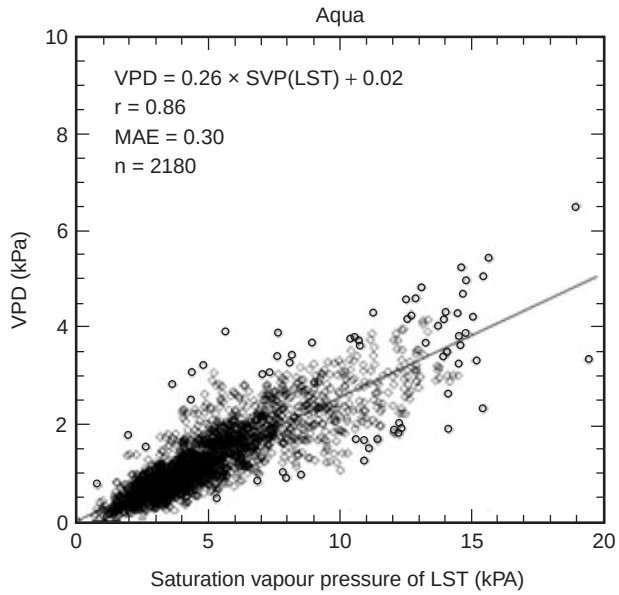


Figure 9 Relationship between a composited maximum 8-day LST from MODIS Aqua at 1330 overpass time and global land surface VPD measured at 2180 surface weather stations for August 2003. (From H. Hashimoto unpublished data) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2003a, 2003b). These advancements offer the potential for finer temporal sampling and mitigation of atmospheric interference. For example, improved MODIS radiometric calibration and cloud-screening algorithms allow the assessment of daily LST temporal dynamics for individual sensor footprints. The scatterplot in Figure 9 shows the relationship between Aqua MODIS LST and the atmospheric vapor pressure deficit (VPD), which is the primary driver of land-atmosphere latent energy and water fluxes. An 8-day time composite of the MODIS Aqua maximum LST for the period provides a metric of changing land surface evaporation resistance that can be mapped at 1 km resolution over regional, continental, and global extents (Figure 10). These technological advances offer potentially simpler techniques for monitoring surface resistances to water mobility compared to earlier methods developed from NOAA AVHRR data using the LST/NDVI ratio technique (e.g. Nemani *et al.*, 1993).

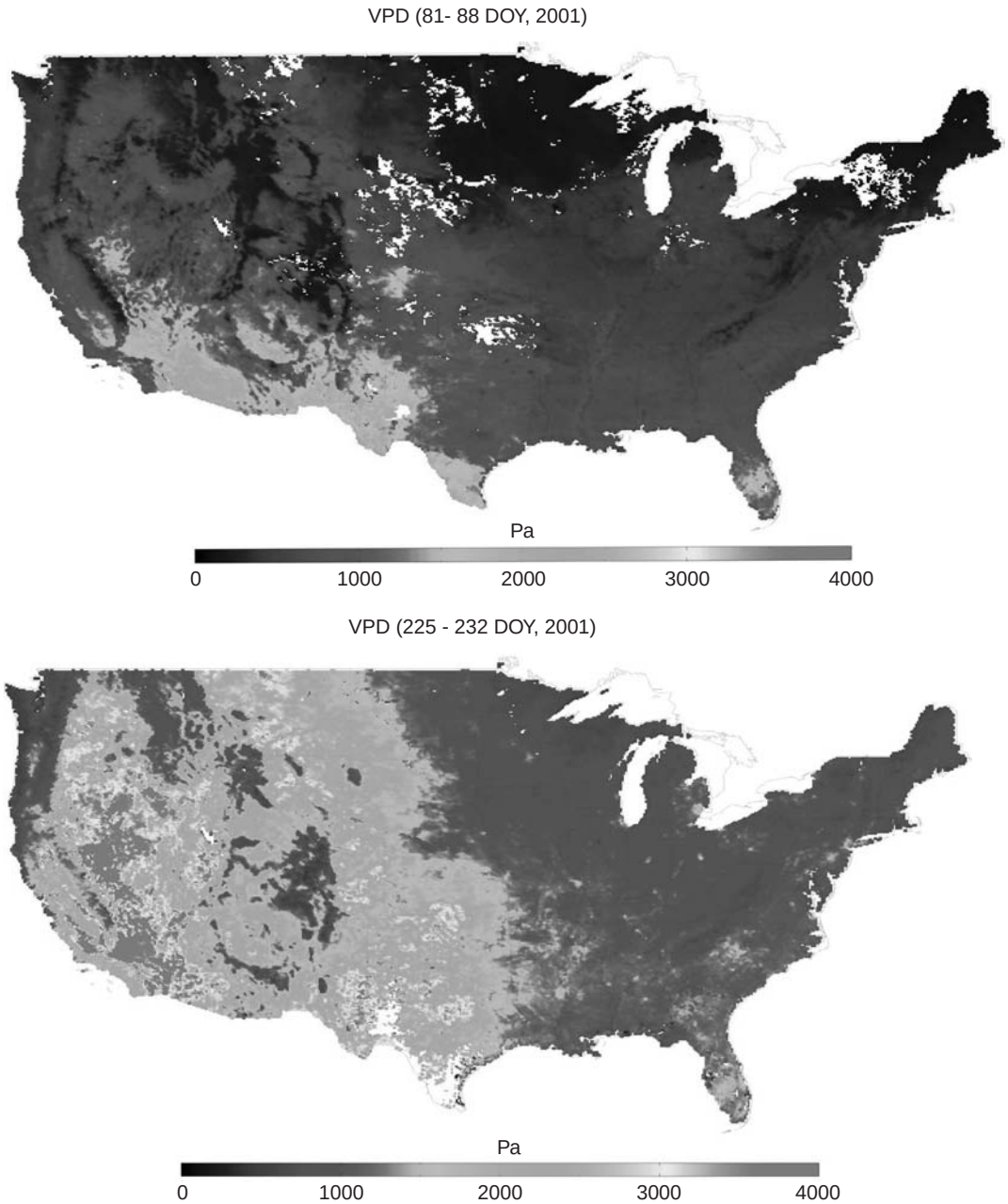


Figure 10 Mapping of maximum weekly VPD during spring and midsummer at 1 km across the continental United States using MODIS Aqua and the relationship in Figure 9. (Hashimoto, unpublished data). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Global satellite remote sensing of landscape water mobility at continental scales requires relatively high (e.g. daily) temporal repeat and moderate spatial scales consistent with the development and dynamic nature of regional weather patterns (Schmugge *et al.*, 2002). Evolving weather systems are influenced by land surface characteristics through surface coupling of energy and water fluxes to the atmospheric boundary layer (ABL). The ABL integrates and responds to surface fluxes on the order of 10 km or less (Albertson and Parlange, 2000). The major limitation on the use of satellite optical/infrared methods to monitor landscape water mobility lie primarily with the coarse temporal compositing required to mitigate the effects of clouds, smoke, and other atmospheric aerosols.

The maps in Figure 11 show seasonal extremes of global plant soil water availability for representative early spring and midsummer conditions estimated using an ecosystem process model (Biome-BGC) driven by daily meteorological inputs from an Atmospheric GCM. Areas of low and high soil water availability represent respective surrogates for high and low resistances to surface latent energy, water, and gas exchange with the atmosphere. In winter and early spring, high latitude and upper elevation landscapes are predominantly frozen and have both low water mobility and high resistance to surface-atmosphere exchanges that are of the same order of magnitude as arid desert regions. Areas of low water mobility (high surface resistances) such as arid deserts and frozen regions are dominated by sensible

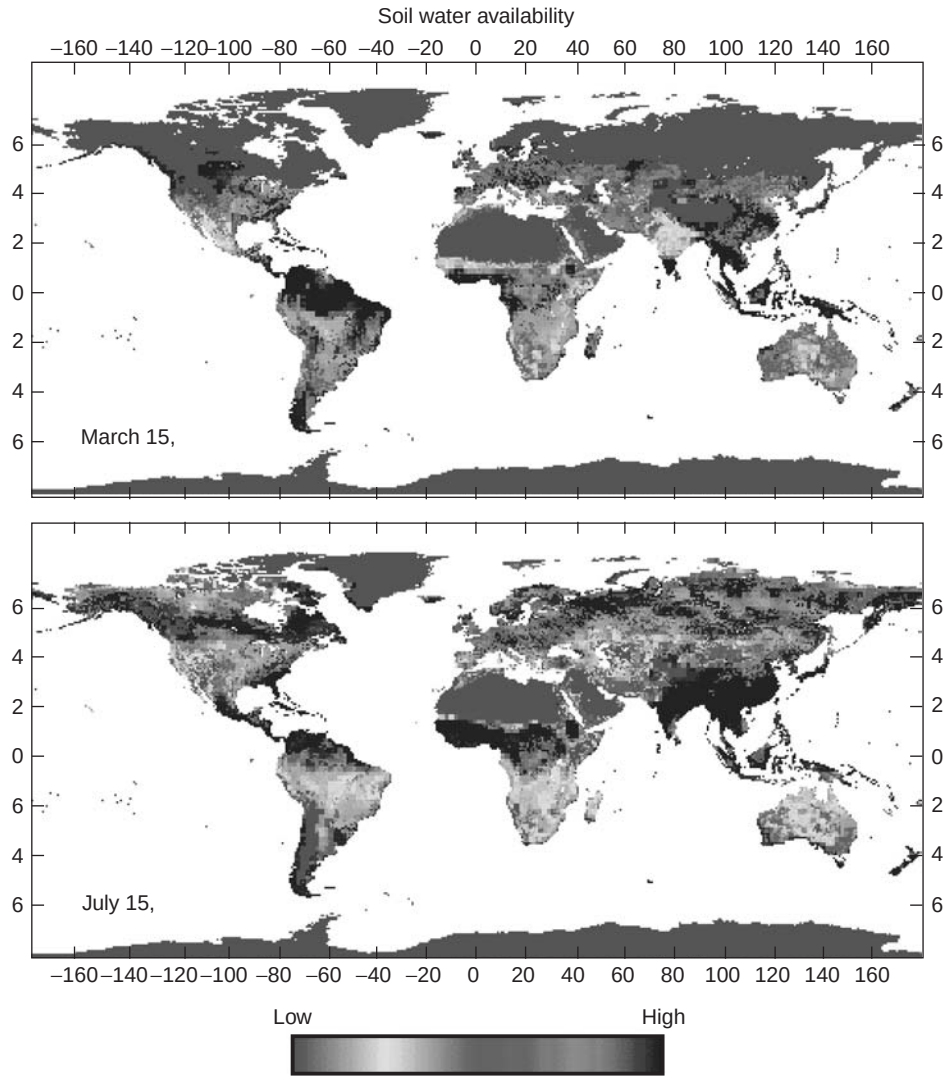


Figure 11 Simulation of a relative integrated soil water availability using NCEP reanalysis daily surface meteorology and the SVAT model Biome-BGC model for 1999. In late winter, 15 March, surface resistance for ET is very high in high latitudes because of freezing temperatures, and in dry deserts. In midsummer, surface resistance is not temperature limited, so high surface resistances are primarily a result of soil water deficits. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

energy exchange, surface temperature extremes and relatively stable atmospheric conditions. In contrast, areas of high water mobility (low surface resistance) are dominated by latent energy exchange and moderate surface air temperatures. Boreal and arctic regions have relatively high seasonal variability in surface resistances to water mobility, while arid deserts and regions of permanent ice and snow maintain relatively persistent low water mobility, high surface resistance conditions.

Future Sensors

A number of recent and planned satellite sensors potentially satisfy the spatial and temporal demands for global monitoring of terrestrial water mobility. The NASA EOS MODIS sensors onboard Terra and Aqua satellites provide both LST and spectral vegetation index information twice daily on a global basis useful for derivation of surface resistances (Nishida *et al.*, 2003a, 2003b). These data are strongly degraded by the presence of clouds and other atmospheric aerosols as well as low solar illumination, shadowing, and the presence of snow cover that often occur at high latitudes and upper elevations. However, the two MODIS sensors operating in tandem onboard Terra and Aqua offer potential for twice daily global observations at relatively high (1 km) spatial resolution (Justice *et al.*, 1998). MODIS also has improved spectral and radiometric resolution compared to older NOAA AVHRR series data. Operational continuity of MODIS capabilities will be provided by the NPOESS (National Polar Orbiting Environmental Satellite System) designed for global daily coverage in the 1330 h overpass time, and scheduled for first launch in 2010 (Townshend and Justice, 2002). NPOESS will have a sensor, VIIRS (Visible Infrared Imager/Radiometer Suite) that will have equivalent spectral capabilities of MODIS in visible, infrared, and thermal infrared wavelengths.

Satellite microwave remote sensing also offers the potential for continuous monitoring of the land surface, with direct sensitivity to landscape moisture, freeze–thaw state and aerodynamic roughness for monitoring landscape water mobility. The NOAA SSM/I sensor series has been providing global daily observations since 1988 at coarse (~25 km) spatial scales (Armstrong and Brodzik, 1995). The AMSR-E sensor recently launched onboard the NASA Aqua satellite also provides global daily observations at similar spatial scales (Kawanishi *et al.*, 2003). Active microwave sensors such as SeaWinds, ERS, and Radarsat are also available and offer the potential for global mapping of freeze–thaw state and surface aerodynamic roughness at spatial resolutions on the order of 25 km or less (McDonald *et al.*, 2004; Way *et al.*, 1997; Le Toan *et al.*, 1992).

The National Aeronautics and Space Administration (NASA) is scheduled to launch the Hydrosphere States Mission (HYDROS) in 2010. This combined active–passive microwave remote sensing satellite is specifically designed to study the state of the global terrestrial hydrosphere including soil moisture and freeze–thaw state (Entekhabi *et al.*, 2004). The HYDROS L-band sensors will provide global mapping of landscape freeze–thaw state at spatial and temporal scales of 3 km and 1 to 3 days. Surface (<5 cm) soil moisture will also be distinguished to within 4% volumetric accuracy, allowing discrimination between dry and saturated conditions over the sparsely vegetated areas of the global land surface at spatial and temporal scales of 10 km and 3 days or less. More importantly, the HYDROS sensor will provide an integrated surface resistance measure that will give full global coverage under all conditions of cloud cover, solar illumination, and vegetation cover through integration of optical/thermal and microwave-based algorithms as discussed here.

The sensitivity of these data to different and complementary landscape biophysical variables such as surface temperature, soil moisture, photosynthetic leaf area and vegetation stress provide a potentially effective means for monitoring ecological controls to surface resistance and water mobility for regional hydro-meteorological research and future water management.

Acknowledgments

We acknowledge financial support from NASA Earth Science Enterprise, and contributions to the text from Drs. Ramakrishna Nemani, Hirofumi Hashimoto, and Maosheng Zhao.

FURTHER READING

Carlson T.N., Gillies R.R. and Schmugge T.J. (1995) An interpretation of methodologies for indirect measurement of soil water content. *Agricultural and Forest Meteorology*, **77**, 191–205.

Running S.W., Nemani R.R., Heinsch F.A., Zhao M., Reeves M. and Hashimoto H. (2004) A continuous satellite-derived measure of global terrestrial primary production. *Bioscience*, **54**, 547–560.

Way J.B., Rignot E., Oren R., Kwok R., McDonald K., Dobson M.C., Bonan G., Viereck L. and Roth J.E. (1994) Evaluating the type and state of Alaskan taiga forests with imaging radar to use in ecosystem flux models. *IEEE Transactions in Geoscience and Remote Sensing*, **32**, 353–370.

REFERENCES

Albertson J.D. and Parlange M.B. (2000) Natural integration of scalar fluxes from complex terrain. *Advances in Water Resources*, **23**(3), 239–252.

Armstrong R.L. and Brodzik M.J. (1995) An earth-gridded SSM/I data set for cryospheric studies and global change monitoring. *Advances in Space Research*, **16**(10), 155–163.

Band L.E., Patterson P., Nemani R.R. and Running S.W. (1993) Forest ecosystem processes at the watershed scale: incorporating hillslope hydrology. *Agriculture Forest Meteorology*, **63**, 93–126.

Bonan G. (2002) *Ecological Climatology*, Cambridge University Press: p. 678.

Bonan G.B., Chapin F.S. and Thompson S.L. (1995) Boreal forest and tundra ecosystems as components of the climate system. *Climatic Change*, **29**, 145–167.

Chapin F.S. III, McGuire A.D., Randerson J., Pielke R. Sr., Baldocchi D., Hobbie S.E., Roulet N., Eugster W., Kasischke E., Rastetter E.B. and Running S.W.(2000) Arctic and boreal ecosystems of western North America as components of the climate system. *Global Change Biology*, **6**,: 211–223.

Diak G.R., Mecikalski J.R., Anderson M.C., Norman J.M., Kustas W.P., Torn R.D. and DeWolf R.L. (2004) Estimating land surface energy budgets from space: Review and current efforts at the University of Wisconsin-Madison and USDA-ARS. *Bulletin of the American Meteorological Society*, **85**, 65–78.

Dickinson R.E. (1996) Land surface processes and climate modeling. *Bulletin of the American Meteorological Society*, **76**, 1445–1448.

Eagleson P. (2002) *Ecohydrology*, Cambridge University Press: p. 443.

El-Rayes M.A. and Ulaby F.T. (1987) Microwave dielectric spectrum of vegetation, Part I: Experimental observations. *IEEE Transactions in Geoscience and Remote Sensing*, **GE-25**(5), 541–549.

Entekhabi D., Njoku E., Houser P., Spencer M., Doiron T., Smith J., Girard R., Belair S., Crow W., Jackson T., *et al.* (2004) The Hydrosphere State (HYDROS) Mission Concept: An earth system pathfinder for global mapping of soil moisture and land freeze/thaw. *IEEE Transactions in Geoscience and Remote Sensing*, **42**, 2181–2195. (in press).

Friedl M.A., McIver D.K., Hodges J.C.F., Zhang X.Y., Muchoney D., Strahler A.H., Woodcock C.E., Gopal S., Schneider A., Cooper A., Baccini A., Gao F., Schaaf C. (2002) Global landcover mapping from MODIS: algorithms and some early results. *Remote Sensing of Environment*, **83**, 287–302.

Gillies R.R., Carlson T.N., Cui J., Kustas W.P. and Humes K.S. (1997) A verification of the 'triangle' method for obtaining surface soil water content and energy fluxes from remote measurements of the Normalized Difference Vegetation Index (NDVI) and surface radiant temperature. *International Journal of Remote Sensing*, **18**(15), 3145–3166.

Goward S.N., Xue Y. and Czajkowski K.P. (2002) Evaluating landsurface moisture conditions from the remotely sensed temperature/vegetation index measurements. *Remote Sensing of Environment*, **79**, 225–242.

Grier C. and Running S.W. (1977) Leaf area of mature northwestern coniferous forests: relation to site water balance. *Ecology*, **58**(4), 893–899.

Hansen M.C., DeFries R.S., Townshend J.R.G., Sohlberg R., Dimiceli C. and Carroll M. (2002) Towards an operational MODIS continuous field of tree cover algorithm: examples using AVHRR and MODIS data. *Remote Sensing of Environment*, **83**, 303–319.

Justice C.O., Townshend J.R.G., Vermote E.F., Masuoka E., Wolfe R.E., Saleous N., Roy D.P. and Morisette J.T. (2002) An overview of MODIS Land data processing and product status, *Remote Sensing of Environment*, **83**, 3–15.

Kane D.L., Hinzman L.D., Haofang Y. and Goering D.J. (1996) The use of SAR satellite imagery to measure active layer moisture contents in arctic Alaska. *Nordic Hydrology*, **27**(1–2), 25–38.

Kawanishi T., Imaoka K., Sezai T., Ito Y.l, Shibata A., Miura M., Inahata H. and Spencer R. (2003) The advanced microwave scanning radiometer for the earth observing system (AMSR-E), NASDA's contribution to the EOS for global energy and water cycle studies. *IEEE Transactions in Geoscience and Remote Sensing*, **41**, 184–194.

Kergoat L. (1998) A model for hydrologic equilibrium of leaf area index on a global scale. *Journal of Hydrology*, **212–213**, 268–286.

Kimball J.S., McDonald K., Keyser A.R., Frolking S. and Running S.W. (2001) Application of the NASA Scatterometer (NSCAT) for determining the daily frozen and nonfrozen landscape of Alaska. *Remote Sensing of Environment*, **75**, 113–126.

Kimball J.S., McDonald K.C., Running S.W. and Frolking S.E. (2004) Satellite radar remote sensing of seasonal growing seasons for boreal and subalpine evergreen forests. *Remote Sensing of Environment*, **90**, 243–258.

Kraszewski A. (1996) *Microwave Aquametry: Electromagnetic Wave Interaction With Water-Containing Materials*, IEEE Press: Piscataway.

Le Toan T., Beaudoin A., Riom J. and Guyon D. (1992) Relating forest biomass to SAR data. *IEEE Transactions in Geoscience and Remote Sensing*, **30**(2), 403–411.

McDonald K.C., Kimball J.S., Njoku E., Zimmermann R. and Zhao M. (2004) Variability in springtime thaw in the terrestrial high latitudes: Monitoring a major control on the biospheric assimilation of atmospheric CO_2 with spaceborne microwave remote sensing. Earth Interactions, **8**(20) 1–23.

Moran M.S., Clarke T.R., Inoue Y. and Vidal A. (1994) Estimating crop water deficit using the relation between surface air temperature and spectral vegetation index. *Remote Sensing of Environment*, **49**, 246–263.

Myneni R.B., Hoffman S., Knyazikhin Y., Privette J., Glassy J., Tian Y., Wang Y., Song X., Zhang Y., Smith G.R., etal (2002) Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*, **83**, 214–231.

Norman J.M., Anderson M.C., Kustas W.P., French A.N., Mecikalski J.R., Torn R., Diak G.R., Schmugge T. and Tanner B.C.W. (2003) Remote sensing of surface energy fluxes at 10-m pixel resolutions. *Water Resources Research*, **39**, 1221, doi:10.1029/2002WR001775.

Nemani R., Keeling C., Hashimoto H., Jolly W., Piper S., Tucker C., Myneni R. and Running S. (2003) Climate-Driven Increases in Global Terrestrial Net Primary Production from 1982 to 1999. *Science*, **300**, 1560–1563.

Nemani R.R., Pierce L.L. and Running S.W. (1993) Developing satellite-derived estimates of surface moisture status. *Journal of Applied Meteorology*, **32**, 548–557.

Nemani R.R. and Running S.W. (1997) Land cover characterization using multi-temporal red, near-IR and thermal-IR data from NOAA/AVHRR. *Ecological Applications*, **7**, 79–90, 1560–1563.

Nemani R.R. and Running S.W. (1989a) Estimation of regional surface resistance to evapotranspiration from NDVI and thermal infrared AVHRR data. *Journal of Applied Meteorology*, **28**, 276–284.

Nemani R.R. and Running S.W. (1989b) Testing a theoretical climate-soil-leaf area hydrologic equilibrium of forests using satellite data and ecosystem simulation. *Agriculture and Forest Meteorology*, **44**, 245–260.

Nielson R.P. (1995) A model for predicting continental-scale vegetation distribution and water balance. *Ecological Applications*, **5**, 362–385.

Nishida K., Nemani R.R., Glassy J.M. and Running S.W. (2003a) Development of an evapotranspiration index from Aqua/MODIS for monitoring surface moisture status. *IEEE Transactions on Geoscience and Remote Sensing*, **41**(2), 493–501.

Nishida K., Nemani R., Running S. and Glassy J. (2003b) An Operational Remote Sensing Algorithm of Land Surface Evaporation. *Journal of Geophysical Research*, **108**(D9), 4270, 10.1029/2002JD002062.

Pitman A.J. (2003) The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology*, **23**, 479–510.

Prentice I.C., Cramer W., Harrison S.P., Leemans R., Monserud R.A. and Solomon A.M. (1992) A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*, **19**, 117–134.

Running S.W. and Coughlan J.C. (1988) A general model of forest ecosystem processes for regional applications. *Ecological Modeling*, **42**, 125–154.

Sandholt I., Rasmussen K. and Anderson J. (2002) A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote Sensing of Environment*, **79**, 213–224.

Schmugge T.J., Kustas W.P., Ritchie J.C., Jackson T.J. and Rango A. (2002) Remote sensing in hydrology. *Advances in Water Resources*, **25**, 1367–1385.

Sellers P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A simple biosphere model for use with general circulation models. *Journal of the Atmospheric Sciences*, **43**, 505–531.

Stephenson N.L. (1990) Climatic control of vegetation distribution: The role of the water balance. *The American Naturalist*, **135**(5), 649–670.

Townshend J.R.G. and Justice C.O. (2002) Towards operational monitoring of terrestrial systems by moderate resolution remote sensing. *Remote Sensing of Environment*, **83**, 351–359.

Ulaby F.T., Moore R.K. and Fung A.K. (1986) *Microwave Remote Sensing-Active and Passive*, Vol. III, Artec House: Dedham.

Vogelmann J.E., Howard S.M., Yang L., Larson C.R., Wylie B.K. and Van Driel N. (2001) Completion of the 1990s National Land Cover Data Set for the Conterminous United States. *Photogrammetric Engineering and Remote Sensing*, **67**, 650–652.

Vorosmarty C.J., Federer C.A. and Schloss A.L. (1998) Potential evaporation functions compared on U.S. watersheds. Possible implications for global-scale water balance and terrestrial ecosystem modeling. *Journal of Hydrology*, **207**, 147–169.

Walter H. (1979) *Vegetation of the Earth and Ecological Systems of the Geo-Biosphere*, *Second Edition*, Springer-Verlag: New York.

Wan Z.M., Zhang Y., Zhang Q. and Li Z. (2002) Validation of the land surface temperature retrieved from Terra MODIS data. *Remote Sensing of Environment*, **83**, 163–180.

Waring R. and Running S.W. (1998) *Forest Ecosystems: Analysis at Multiple Scales*, Academic Press: San Diego.

Waring R.H., way J.B., Hunt E.R., Morrissey L., Ranson K.J., Weishampel J.F., Oren R. and Franklin S.E. (1995) Imaging radar for ecosystem studies. *Bioscience*, **45**(10), 715–723.

Way J.B., Zimmermann R., Rignot E., McDonald K. and Oren R. (1997) Winter and spring thaw as observed with imaging radar at BOREAS. *Journal of Geophysical Research*, **102**, 29,673–29,684.

Woodward F.I. (1987) *Climate and Plant Distribution*, Cambridge Univ Press: p. 174.

Zhang T., Scambos T., Haran T., Hinzman L.D., Barry R.G. and Kane D.L. (2003) Ground-based and satellite derived measurements of surface albedo on the north slope of Alaska. *Journal of Hydrometeorology*, **4**, 77–91.

Zhao M., Heinsch F.A., Nemani R.R., Running S.W. (2005) Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment*. (in press).

105: Microbial Transport in the Subsurface

TIMOTHY R GINN[1], TERRI CAMESANO[2], TIMOTHY D SCHEIBE[3], KIRK E NELSON[1], T PRABHAKAR CLEMENT[4] AND BRIAN D WOOD[5]

[1]Department of Civil and Environmental Engineering, University of California, Davis, CA, US
[2]Department of Chemical Engineering, Worcester Polytechnic Institute, Worcester, MA, US
[3]Environmental Technology Directorate, Pacific Northwest National Laboratory, Richland, WA, US
[4]Department of Civil Engineering, Auburn University, Auburn, AL, US
[5]Environmental Engineering, Oregon State University, Corvallis, OR, US

We review the processes involved in the transport of bacteria in the saturated subsurface. Our goal is to communicate priority topics in the basic science of bacterial transport processes, and this leads to a focus on the nanometer to micron scales of observation. The scope of this review is the set of physical, chemical, and biological processes involved in the transport of individual cells, and the constitutive theory underlying conceptual depiction of these processes.

INTRODUCTION

Concern about pathogen contamination of groundwater and the use of bacterial agents in the cleanup of groundwater has highlighted the need for an improved understanding of the fate and transport of microbes in the subsurface. In particular, *in situ* bioremediation of contaminated groundwater may involve microbial transport promoted by intrinsic bioremediation (as part of natural attenuation, e.g. Chapelle and Lovley, 1990; Semprini *et al.*, 1992; Wiedemeier *et al.*, 1995; Clement *et al.*, 2002)), biostimulation (by the addition of substrates or electron donors, e.g. Semprini and McCarty, 1992; Huesemann and Truex, 1996), or bioaugmentation (by the introduction of microbial cells with specific function to the subsurface; e.g. Mayotte *et al.*, 1996; Duba *et al.*, 1996a,b). Bioaugmentation in this context includes both injection of bacterial suspensions in the saturated zone near a contaminant plume and emplacement of solid media with attached bacteria as a "biobarrier" through which mobile contaminant is expected to pass and be degraded (Kao and Yang, 2000; Dybas *et al.*, 2002). Both biostimulation and bioaugmentation approaches have been widely applied to *in situ* remediation of chlorinated solvents, and the science of underlying processes is well within the topical public news (e.g. Bowman, 2002). The relative effectiveness of biostimulation versus bioaugmentation approaches was recently debated in Nyer *et al.* (2003) and in Major *et al.* (2003).

In this article, we focus on the physical, chemical, and biological processes involved in the transport of bacteria in the saturated subsurface. Since our main focus is on the basic scientific understanding of bacterial transport processes, we do not treat in great detail either conventional mathematical models of bacterial attachment/detachment kinetics (recently reviewed in Clement *et al.*, 1999, and in Murphy and Ginn, 2000; here treated briefly in the Section on "Conventional models of bacterial attachment/detachment kinetics"), or continuum-scale (e.g. laboratory column) experiments elucidating apparent effects of aqueous or mineralogical physicochemical conditions (such as grain size distribution, presence of mineral oxides or organics, aqueous ionic strength, pH, and velocity; recently reviewed in Ginn *et al.*, 2002, and in Harvey and Harms, 2001). Thus, the scope of this review is limited to physical, chemical, and biological processes involved in bacterial transport, and the constitutive theory underlying their conceptual depiction. Transport of other microorganisms (e.g. viruses) is often treated by considering approaches similar to those used for bacteria, but will

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

not be considered in detail in this work (the interested reader is referred to Schijven and Hassanizadeh, 2000). Also in our attempts at concision we focus on the references most generally applicable to subsurface fate and transport sometimes at the cost of identifying the first references.

In Figure 1, we have presented a schematic illustration of a sequence of length scales that might apply to microbial transport in the subsurface. At the most fundamental scale (Level I of Figure 1; hereafter referred to as the "cell scale" or "cytoscale"), transport is influenced by the interactions between the complex heterogeneous cell surface and the solid phase (which is itself a complex, heterogeneous material). These interactions, although manifest at the scale of nanometers, is nevertheless one of the controlling features of microbial transport at the largest scale of interest (the "field scale", Level IV in Figure 1), which might represent the scale associated with an aquifer remediation zone or the depth of a bioreactor such as a tricking filter. In between these two length scales, one may further identify several other scales of interest: the "pore scale" (Level II) where cell transport is governed by the conservation of mass and momentum within the porous matrix, and where groups of cells may become evident as biofilms within the porous media; and the "Darcy scale" where the details of individual cells are not apparent, and the cell mass is treated as a homogeneous chemical phase.

We will first review conceptual models of bacterial phases in the subsurface, and then the processes controlling fate and transport on short (e.g. bioremediation) timescales. Finally, we briefly review field-scale bacterial transport experiments and discuss a number of issues that impact the application of current process descriptions and models at the field scale.

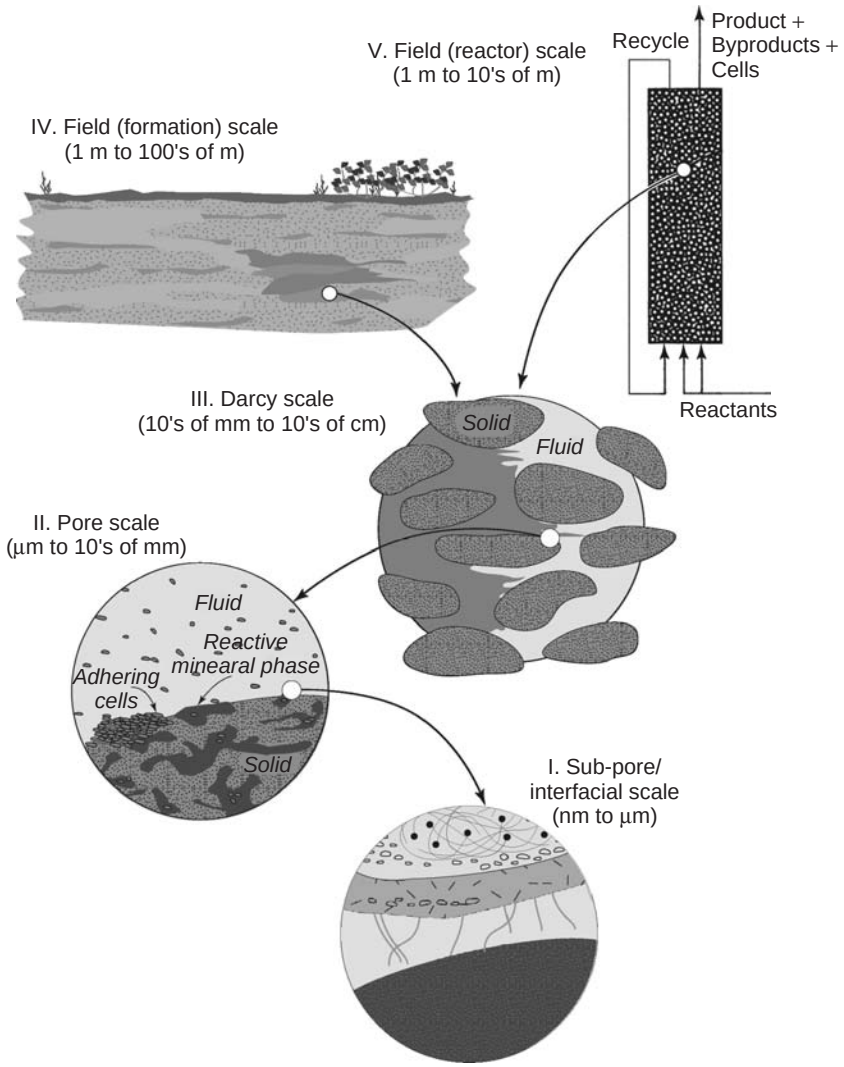


Figure 1 A hierarchy of scales associated with microbial transport in the subsurface. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CONCEPTUAL AND MATHEMATICAL REPRESENTATION OF SUBSURFACE BIOMASS

The transport and fate of microbes in the subsurface is naturally dependent on the distribution of biomass in the aqueous and solid phases. Thus, the mathematical description of the kinetics of biomass transformation resulting from processes such as growth and cellular attachment to surfaces first requires a conceptual model of the biotic phases. Aqueous-phase biomass is commonly treated as a dilute suspension of "free-living" (i.e. aqueous or planktonic) cells, and is typically mathematically represented as a dilute reactive solution (e.g. Murphy *et al.*, 1997). The assumption of a dilute solution neglects interactions between aqueous cells, and this assumption will not be valid when there are significant cell interactions such as those encountered in cell clumping and quorum sensing (Ward *et al.*, 2001). Approaches that invoke the interactions among cells have not yet been developed; however, such effects have been rigorously incorporated in the description of the transport of simple colloids (e.g. Peters, 1990; Peters and Ying, 1991). A similar analysis for representing the interactions among microorganisms represents a significant challenge, and is an open area for continuing research.

Biomass associated with the solid phase can be modeled using three conceptual approaches representing major schools of thought, as discussed originally by Molz (1987). Later, Baveye and Valocchi (1989) formalized these concepts into (1) unstructured macroscopic and (2) structured microscopic models, where the biofilm geometry was accounted for, at least conceptually. The structured models specify the geometrical aspects of the attached biophase in particular, for example, as "pointwise" colonies, or as a biofilm described with a thickness and porosity, and so on. These models allow quantification of momentum and mass transfer at the cytoscale.

Momentum transfer between the solution and structured biophase is important for evaluating forces involved in biomass growth that lowers permeability, that is, bioclogging. Unstructured models treat biomass as another mixture component (like a solute and/or sorbate) without specification of physical structure. Clement *et al.* (1996b) showed that permeability reduction can also be modeled using unstructured macroscopic approaches and pointed out some of the similarities and differences between macroscopic- and biofilm-based bioclogging models. MacDonald *et al.* (1999a,b) examine the role of growth in a structured subsurface biofilm in the clogging of porespaces during bioremediation. Dupin *et al.* (2001a,b) combine pore-scale biodegradation and bacterial growth, using a structured approach, with quantitative treatment of the mechanics of biofilm deformation. Biomass is modeled as a continuous uniform isotropic hyperelastic material, using both structure colony ("aggregate") and biofilm conceptual models, whose expansion and deformation are governed by material mechanics stress-strain relations. Results in the context of a 2-D lattice porous medium suggest that aggregates have a much greater potential impact on clogging than biofilms, so the geometry of the attached biophase (and its representation in models) matters to the simulation and prediction of clogging.

In the unstructured biophase approach, no structural presumption is imposed on the biophase (MacQuarrie *et al.*, 1990; Sudicky *et al.*, 1990; Semprini and McCarthy, 1992; Widdowson *et al.*, 1988; Clement *et al.*, 1999) and the biomass is treated as a suspended, but kinetically sorbing/desorbing species (Zysset *et al.*, 1994; Murphy *et al.*, 1997). In this case, the biomass is viewed as a mixture component that is not a separate phase, occupies no volume, and is fully mixed with other aqueous species. It is also assumed that a linear relation between mass of substrate consumed and biomass produced, and that no diffusion limitations affect the transfer of substrate mass from solution into the biomass. This approach has been taken in model construction (Corapcioglu and Haridas, 1984; Borden and Bedient, 1987; MacQuarrie *et al.*, 1990), in column studies that focus on bacterial transport (Lindqvist *et al.*, 1994; Tan *et al.*, 1994) and in intermediate-scale flow cell studies of growth and coupled transport (Wood *et al.*, 1994; Murphy *et al.*, 1997). Challenges in validating the conceptual model are associated with the difficulty in gathering data at the proper (pore) scale, with nonlinearities in models associated with biofilm growth, and with the tremendous spatial and temporal variability of biofilm properties. Peyton and Characklis (1995) report that published biofilm for density values range from 10 to 130 $\mathrm{kg\,m^{-3}}$ of bulk porous media. These values typically vary even more under conditions of active biodegradation (Murphy and Ginn, 2000). However, a quantitative foundation for understanding microbial processes in porous media may best be achieved through structured models explored at the pore scale and below (e.g., Wood and Whitaker, 1998; 2000a; 2000b).

Within the context of mass-transfer processes, the definition of how to best model biological phases in porous media has been the topic of much discussion and debate (see, e.g. the original research articles by Moltz *et al.* (1986), Baveye and Valocchi (1989), Taylor and Jaffé (1990), and subsequent comments on these articles by Widdowson (1991), Baveye and Valocchi (1991), Baveye *et al.* (1992), Jaffé and Taylor (1992)). Although initially much of the focus of these debates was related to the geometry and distribution of microorganisms associated with the solid phase (and, subsequently, the kinds of conceptual models that were relevant to these various geometries), it was pointed out by Rittmann (1993) that in practice the actual geometry may make little difference in the determination of the effective reaction rate for mass transfer associated with biodegradation. What

does matter is whether the biotic phase itself limits diffusive mass transfer of degradable compounds and reactants; when there is no diffusion limitation between extra-biofilm fluid and intra-biofilm fluid, the biofilm is said to be in local mass equilibrium (LME) with the fluid phase. This assumption implies the absence of significant concentration gradients within the biofilm itself, and when valid leads to a single conservation equation for reactive solutes in the aqueous phase. When mass transfer between the extra-biofilm fluid and the biofilm fluid is diffusion-limited, a two-region nonequilibrium model (leading to two coupled conservation equations for reactive solutes, one for the fluid phase and one for the biological phase) is indicated.

The first model is consistent with what has also been called the *strictly macroscopic model* (Baveye and Valocchi, 1989). This model has been used extensively in the literature (e.g. MacQuarrie *et al.*, 1990; Sudicky *et al.*, 1990; Zysset *et al.*, 1994; Wood *et al.*, 1994, 1995; Clement *et al.*, 1996a,b; Zysset *et al.*, 1994; Murphy *et al.*, 1997), and has recently been derived by upscaling the physical and biological processes at the microscale via volume averaging (Wood *et al.*, 2002, Wood, 2004). For the second model, the situation is slightly more complex. If it is assumed that the characteristic time for changes in the biofilm phase is much smaller than those for the fluid phase, then the biofilm can be treated as being quasi-steady. This approach has been adopted widely, with minor differences in the conceptual picture of the biomass distribution, by a large number of researchers (e.g. Molz *et al.*, 1986; Widdowson *et al.*, 1988; Baveye and Valocchi, 1989; Dykaar and Kitanidis, 1996; Chen *et al.*, 1992). If it is further assumed that the concentration profile in the biofilm is quasi-steady but not necessarily fully penetrating, then the reaction rate can be developed in terms of an "effectiveness factor" that corrects the fully penetrating model to account for a nonuniform substrate distribution in the biofilm. Such approaches have been presented by fewer researchers (e.g. Taylor and Jaffé, 1990; Kim *et al.*, 2004). Finally, for the case of relatively thick biofilms where the mass-transfer effects within the microbial phase itself may be important (e.g. near an injection well in which a pulsing scheme is used for adding substrates; within a bioreactor), it may be necessary to use a fully transient two-region nonequilibrium model. This model would be similar to the "mobile–immobile" modes that are often used for describing solute transport and reactions in highly heterogeneous media (e.g. van Genuchten *et al.*, 1976). However, other than the very general models just mentioned, no two-region models particular to biodegradation in the subsurface have yet been proposed.

Although the structure and geometry of biofilms at the pore scale may not be of paramount importance for predicting reaction rates under the fully penetrated biofilm assumption, it may be important for systems that are not fully penetrated or where the two-equation model applies. Additionally, even for the fully penetrated model, the biofilm geometry and structure within the pore space does affect the effective dispersion tensor that is observed (Wood, 2004). Experimental work has been conducted to expose the structure of biofilms within experimental systems. Observations have ranged from distributions of cell aggregates or "microcolonies" (Vandevivere and Baveye, 1992; Dupin and McCarty, 2000) to relatively thick and continuous films (e.g. Paulsen *et al.*, 1997; Sharp *et al.*, 1999; Dupin and McCarty, 2000; Vayenas, 2002). These studies have focused primarily on two-dimensional structure observed either in etched micromodels or on select surfaces of porous materials. More recently, the three-dimensional structure of thick biofilms has been observed using nuclear magnetic resonance, as reported by Seymour *et al.* (2004) and Wood (2004). Studies of this type should help to elucidate ultimately the conditions under which the LME and nonequilibrium models must be adopted, and representations that allow one to predict the change in the effective parameters of these models (e.g. reaction rates, effective dispersion tensors, effectiveness factor) for these two conditions.

PHYSICAL PROCESSES

Microbial transport in the subsurface involves a host of complex and interacting processes. It is thus not surprising that the literature contains many inconsistencies regarding the effects of bacterial variables such as size, shape, hydrophobicity, and electrostatic charge (Lawrence and Hendry, 1996). As such, it is not currently possible to state definitive correlations between bacterial properties and transport. Because microbes are living organisms, their transport in the subsurface is more complex than is the case for abiotic colloids. Not only are they subject to the same physicochemical phenomena as are colloids, but there are also a number of strictly biological processes that affect their transport (e.g. temporal changes in surface properties due to changes in metabolic state; predation by other subsurface organisms). Because of the complexity of the combined physicochemical and biological processes, these processes are described in separate sections in the material following.

In comparison to the biological processes, physical processes affecting microbial transport such as advection, dispersion, straining, and physical filtration have been the focus of numerous experimental and numerical modeling studies. These important processes provide the framework of bacterial transport and reaction in porous media. In addition to the following, readers are referred to reviews by Harvey and Garabedian (1991) and McDowell-Boyer *et al.* (1986) for more thorough discussions of these processes.

Transport

Microbes undergo convective transport as a particulate suspension moving with the pore water whose velocity is governed by the hydraulic pressure gradient, porosity, and permeability distribution. The occurrence of nutrient and/or electron acceptor as a solute undergoing transport may be coupled to the transport process through the effects of these constituents on the fluid properties of density and viscosity. Convective transport in porous media is also associated with hydrodynamic dispersion, the mixing process arising from the variations in velocity along the tortuous convective paths compounded by diffusional displacements of cells across streamlines of convective transport.

Straining and Filtration

Straining and physical filtration represent the removal of microbes from solution by physical (geometric and intermolecular/surface) forces. Straining is the trapping of microbes in pore throats that are too small to allow passage and is exclusively a result of pore geometry (Corapcioglu and Haridas, 1984). Estimates based on purely geometric relations between the effective diameter of biocolloids and the diameter and packing (coordination number) of grains suggest that mass removal by straining is not significant where the colloid diameter is less than 5% of the porous media grain diameter (Corapcioglu and Haridas, 1984; Harvey and Garabedian, 1991; McDowell-Boyer *et al.*, 1986). It should be noted that in natural heterogeneous porous media, a fraction of the pore diameters might be small enough to cause straining of colloidal particles even though the average grain diameter passes this rule of thumb for nonsignificance of straining.

Physical filtration is the removal of particle mass from solution via transport to and deposition on the porous media surface (note that in this context, "transport" means the conveyance of a microbe from the aqueous phase to the solid phase of the porous medium); here the term includes both attachment and sedimentation. Attachment of bacteria in the natural subsurface via filtration is often treated using colloid filtration theory (CFT; e.g. Yao *et al.*, 1971; Rajagopalan and Tien, 1976), which posits the kinetic rate of attachment as

$$\text{kinetic attachment rate} = \frac{3}{2}\,\frac{1-\varepsilon}{d}\alpha\eta\|v\|C_{A\gamma} \qquad (1)$$

Here, $C_{A\gamma}$ is the bulk aqueous concentration of mobile microbes, ε is porosity, d is average diameter of the porous media grains, η is the collection efficiency (defined as the fraction of microbes approaching an idealized porous media grain that collides with the grain) and α is the collision (or "sticking") efficiency (defined as the fraction of microbes colliding with the idealized porous media grain that attaches to its surface). Generally, η is calculated *a priori* based on bacterial and porous media properties, and then α is calibrated with the use of data from column experiments. The literature contains a helpful clarification on the use of the analytical expressions for η (Logan *et al.*, 1995), as well as a summary of CFT's major assumptions and the implications for modeling bacterial transport in natural porous media (Nelson and Ginn, 2001).

The most commonly used equation for η (the RT equation) was developed by Rajagopalan and Tien (1976). The RT equation describes transport to the porous media surface due to interception, sedimentation, diffusion, and Derjaguin Landau Verwey Overbeek (DLVO) surface interaction forces. Interception is filtration due to convection and the finite size of the colloidal particle. In other words, if a particle is riding exactly on a fluid streamline, then it will make contact with the grain if the streamline comes within one particle radius of the grain.

Sedimentation is enhanced filtration due to gravity (Corapcioglu and Haridas, 1984; McDowell-Boyer *et al.*, 1986) and depends on particle buoyancy (Wan *et al.*, 1995). Many natural bacteria and viruses are neutrally buoyant, in which case sedimentation is negligible. However, cultured microorganisms are typically larger and sometimes more dense than their native counterparts (Harvey *et al.*, 1995) and may involve sizable buoyancy-driven filtration. Sedimentation has been approximately quantified in Harvey *et al.* (1995) by augmenting the advective pore-water velocity with an additional downward component whose magnitude is given by the classical Stokesian velocity of a dense sphere falling through a fluid; $v_s = (\rho_s - \rho)gd_s/18\,\mu$, where v_s is the sedimentation velocity (acting vertically downward), ρ_s is the cell density, ρ is the solution density, g is the gravitational acceleration, μ is the dynamic viscosity, and d_s is the cell diameter (treated as a sphere). It should be pointed out that if CFT is employed and the RT equation for η is used with a microbial cell density greater than that of water, then the CFT kinetic rate of attachment (equation 1) will contribute an *additional* velocity magnitude; however, this CFT sedimentation velocity necessarily acts in the same direction as the principal flow direction, which is inconsistent with general application of the advective-dispersion equation. This is due to the orientation of the Happel sphere-in-cell model (Happel, 1958) that was used in the RT equation's derivation.

Diffusion is filtration due to the Brownian motion of the suspended particles. Since smaller particles will be moved farther by the random molecular bombardments of surrounding water molecules, the effect of diffusion becomes more important as particle size decreases. Indeed, diffusion is the dominant filtration mechanism for submicron passive particles. The RT equation assumes that the effect of diffusion (as described by the Smoluchowski-Levich assumptions that neglect interception and external forces)

can be considered separately from the effects of the other mechanisms. Recent studies (Tufenkji and Elimelech, 2004; Nelson and Ginn, 2005) have indicated that this assumption can result in significant errors for Brownian particles. For its representation of DLVO forces, the RT equation neglects the effect of the electrostatic double layer force, resulting in a surface interaction potential that consists of the attractive London-van der Waals force.

Another important aspect of the CFT is that it is based on an analysis of particle transport within an idealized pore space, and therefore this analysis must be upscaled for application in a macroscopic transport equation such as the advective-dispersion equation. Equation (1) represents the formula that has typically been employed for upscaling the CFT analysis. Recent work (Nelson and Ginn, 2005) has shown that this conventional methodology for upscaling the microscopic CFT analysis contains two inconsistencies. These issues and a proposed alternative methodology are presented in the Section on "Upscaling the Smoluchowski equation to the Darcy scale".

Size Exclusion

Size exclusion is a phenomenon where transported particles move faster than the mean pore-water velocity due to the size or charge of the material conveyed. When the phenomenon is observed in one-dimensional macroscopic scale, it appears as an earlier breakthrough of excluded particles. As a change in speed without change in macroscopic direction this is also termed "speedup" (Ginn, 2002). Size exclusion effects have been observed in laboratory columns (Engfield and Bengtsson, 1988; Pyle, 1979; Shonnard *et al.*, 1994; Hornberger *et al.*, 1992; Mayotte *et al.*, 1996) and in field experiments (Harvey *et al.*, 1989; Pyle and Thorpe, 1981; Wood and Ehrlich, 1978). Identification of exclusion from observations of tracer and particle breakthrough is not straightforward and has led to some confusion in the literature. The most reliable indicators of exclusion are (i) a significant difference in fitted advective velocities in a model parameter estimation exercise, or (ii) significantly higher normalized concentrations (C/C0) of suspended microbes during the rising limb of the breakthrough curve. These issues are discussed further by Zhang *et al.* (2001a) and DeBorde *et al.* (1999). With regard to processes causing exclusion, one may distinguish anionic and size effects, and further divide size effects into classical chromatographic and "pore exclusion" processes. Anion exclusion involves velocity enhancement by channeling of anionic molecular-scale solute particles in finer-grained porous media away from pore walls due to electrochemically repulsive forces that act on nanometer scales, and as such are not generally significant for bacterial transport. In size exclusion, microbes and large colloids, by virtue of their size characteristics alone, preferentially experience the higher velocities near pore centerlines, yielding an average convective velocity that is higher than that of a dissolved tracer. The occurrence of exclusion typically requires the effective bacterial diameter to be less than 1% of the characteristic pore throat length scale (e.g. mean grain diameter) that is common for transport in sandy aquifers (de Marsily, 1986; Dodds, 1982). When the colloidal particle is of the same scale as a significant fraction of pore channels, not all pores are accessible. The presumed rerouting of particles to alternate pore throats (or alternate porous materials at the macroscopic scale) in this case occurs on a relatively larger detour scale than does chromatographic or ionic exclusion. Termed *pore exclusion* (Woessner *et al.*, 2001), for virus transport), this has been suggested as a velocity enhancement of the excluded material in natural aquifers. Recent research based on visualization of microsphere transport within pore-scale micromodels (Sirivithayapakorn and Keller, 2003) indicates that size exclusion processes (which focus colloids near pore throat centers) also cause colloids to move along a different set of flowpaths than solute molecules, which provides a potential mechanism for preferential transport of colloids along selected flowpaths (i.e. pore exclusion) such that the two processes are in fact closely related. Note that pore exclusion has also been called *size exclusion* by some authors (Rehmann *et al.*, 2000), as well as various other terms such as *volume exclusion* (Bales *et al.*, 1989), *pore-size exclusion* (Sinton *et al.*, 2000), or *size exclusion chromatography* (Harvey *et al.*, 1989). The term *differential advection* has also been introduced (Zhang *et al.*, 2001a,b) to generally describe the phenomenon of earlier breakthrough of colloids relative to a solute tracer without specific regard to the mechanism.

The mechanics and modeling of pore exclusion have been recently debated (Ginn, 2000b; Rehmann *et al.*, 2000). One approach, also used in anion exclusion, involves reducing the kinematic porosity (that associated with aqueous phase moving in response to hydraulic gradient) by an "excluded" fraction that is unavailable to the anion. The remaining porosity is then divided into mobile and immobile fractions (Gvirtzman and Gorelick, 1991). However, the simple porosity reduction approach is unsatisfactory because unrealistically large porosity reductions would be required in cases where a high degree of speedup is observed (e.g. Dong *et al.*, 2002). Ginn (2002) developed a mathematical approach to incorporating exclusion effects in a lagrangian context, relating the distribution of excluded particle travel times to that of nonexcluded, or "ideal" particles, such as reflected by a dye tracer test. The analysis involves a constitutive speedup function that tells how travel times of nonexcluded particles map to those of excluded particles. This approach was used to analyze *Cryptosporidium* breakthrough data in saturated columns (Harter *et al.*, 2000). Scheibe and Wood (2003) modeled exclusion phenomena observed in sand column experiments using a modified

particle-tracking approach. In the latter approach, the distribution of local dispersive displacements (corresponding to the value of dispersivity estimated from conservative tracer breakthrough observations) was truncated at the lower end to represent the exclusion of bacteria from regions of the pore space with very small local velocity (i.e. very near pore walls). This approach was demonstrated to be effective at simulating observed large exclusion effects (bacterial velocities nearly double that of conservative tracers) with minimal truncation (on the order of 5%) of the dispersive displacement distribution. They also found a decrease in apparent dispersion of the bacteria relative to conservative tracers, consistent with theoretical considerations and observations (Sinton *et al.*, 2000). James and Chrysikopoulos (2003) derived similar relationships using a particle-tracking model of colloidal transport in a uniform fracture, and also determined that exclusion leads to decreased effective dispersion.

SURFACE INTERACTIONS INFLUENCING MICROBIAL TRANSPORT

Although the physical processes described above are well understood by most hydrologists, the influence of electrochemical interactions between microorganisms and solid surfaces are not as familiar. These forces may act over characteristic lengths that are only fractions of nanometers to microns, but ultimately they determine how microorganisms adsorb and desorb from the solid surface, and thus can dramatically affect microbial transport at even the largest scales. In this section, we review some of the current research and perspectives on these processes.

Understanding the Interaction Potential: is the DLVO Appropriate for Microbes?

CFT, introduced above, assumes that attachment is a two-step process. First, the bacterium must be transported to the porous media grain (the "collector"). Second, the physicochemical and microbiological interactions that occur upon contact of the two surfaces determine if the bacterium attaches to the surface of the collector. Traditionally, the collection efficiency parameter, η, which represents the transport step, is calculated *a priori* from characteristics of the bacterium, the porous media, and the flow field. The sticking efficiency (α, representing the attachment step in filtration theory) is then treated as a fitting parameter in conjunction with experimental data. This conceptual model may be erroneous in that the transport and attachment steps may in fact be coupled (McClaine, 2001; Nelson and Ginn, 2001).

Once a bacterium is transported to within a separation distance on the order of fractions of its own radius away from a collector, a complex set of interactions occurs. The DLVO theory of colloid stability has been widely employed as a model for describing the interaction forces between a microbe and a solid surface. The DLVO force potential is the sum of the London-van der Waals force and the electrostatic force (i.e. the "double layer" force which may be repulsive or attractive depending on the charges of the two interacting surfaces).

Although the DLVO theory is a useful conceptual model of the interactions between microorganisms and solid surfaces, Ninham (1999) and Boström *et al.* (2001) have pointed out that it is not generally applicable for microbial cells whose surfaces are not molecularly smooth, solid, and inert. Atomic force microscopy (AFM) measurements of the forces between bacteria and solid surfaces often show substantial disagreement with the predictions of DLVO theory both in magnitude and decay distance (Camesano and Logan, 2000a). Similar conclusions have been reached by the observation of *Escherichia coli* taxis along a glass surface; the bacteria-surface interactions could not be explained by DLVO theory (Viegeant and Ford, 1997). Further, bacterial adhesion data often deviates from the predictions of the extended DLVO model, which allow for electron donor–acceptor interactions to be accounted (van Oss, 1994; Jucker *et al.*, 1998a).

In addition, recent work has suggested that DLVO theory is not appropriate for microbes because their soft surfaces allow for the exchange of ions through the membrane. This has led to the development of "soft-particle" DLVO theory (Ohshima, 1989), in which the colloid is assumed to consist of an ion-penetrable, charged polyelectrolyte layer around a rigid core (Hayashi *et al.*, 2001). Energy barriers calculated on the basis of classical and soft-particle DLVO theory can be significantly different for some microbes. For example, the critical height of the energy barrier below which spontaneous adhesion of *E. coli* JM-109 could adhere to sand was $49\,k_B T$ (where k_B is the Boltzmann constant and T is absolute temperature) based on soft-particle theory, or $72\,k_B T$ based on conventional, or rigid-particle DLVO analysis (Abu-Lail and Camesano, 2003a). Despite the improved applicability in the application of soft-particle DLVO theory to bacteria, there are still large discrepancies between DLVO predictions and measurements of bacterial adhesion. There are also some microorganisms that cannot be well described by the soft-particle DLVO theory, such as the fungus *Candida parapsilosis* (Emerson and Camesano, 2004). The latter probably has a more rigid cell wall than many bacterial cells, and therefore it may not be as ion-penetrable.

It has been suggested that the failures of DLVO, soft-particle DLVO, and extended DLVO theory are most likely caused by the presence of polymers, other macromolecules, and structures such as pili and flagella on the bacterial surface (Jucker *et al.*, 1998a; Otto *et al.*, 1999; Camesano and Logan, 2000b; Ginn *et al.*, 2002). Thus, an understanding

of polymer-mediated interactions is a priority for the successful prediction of bacterial attachment rates in porous media. Observation of polymer interactions with surfaces and related attachment of bacteria have been under way for over 30 years (e.g. Fletcher and Floodgate, 1973). Bacterial polymer layers can cause a steric repulsive force (van Loosdrecht *et al.*, 1990; Rijnaarts *et al.*, 1999; Simoni *et al.*, 1998; Jucker *et al.*, 1998a) as well as an attractive bridging phenomena (Rijnaarts *et al.*, 1995a; Williams and Fletcher, 1996; Jucker *et al.*, 1998a,b). The occurrence of attraction or repulsion depends on the coverage degree, polymer characteristics, and the type of solvent (Rijnaarts *et al.*, 1995a). Of critical importance to the interplay of attractive and repulsive tendencies is the relative affinity of the polymers for the solid surface and for water. Both types of steric interaction may operate for a given polymer, and since the attraction and repulsion potentials will depend differently on the separation distance, the net result is likely to change signs as separation varies (Rijnaarts *et al.*, 1995a). Heterogeneity in the physical and chemical properties of the polymers on a bacterium would yield a force potential that is distributed over the cell surface. For example, the presence of a small number of long polymer chains, extending far beyond the reach of the bulk polymer layer would not encounter the same energy barrier as the rest of the polymer layer. If only a few of these long chains form hydrogen bonds, the binding force could be sufficient for irreversible attachment. This scenario has been proposed as an explanation for the observed adhesion of *Stenotrophomonas maltophilia* (Jucker *et al.*, 1998b), and a similar argument was made for *Pseudomonas fluorescens* (Williams and Fletcher, 1996; Smets *et al.*, 1999). Larger cell-surface structures such as flagella may also behave in a similar manner (e.g. Otto *et al.*, 1999).

In the last decade, atomic force microscopy (AFM) has surfaced as a viable technology for making the interaction force measurements required for the development of a predictive theory of polymer-mediated interactions (see overview in Ginn *et al.*, 2002). In addition to force measurements, AFM can be used to characterize cell-surface properties such as polymer conformations (e.g. Abu-Lail *et al.*, 2002; Camesano *et al.*, 2002). For a review of AFM cell-surface characterization in general, see Dufrêne, 2001. The basic setup involves a probe, attached to a cantilever; this unit is moved toward the microbial sample (or vice versa) until contact is made, and then retracted. The force is calculated from the deflection of the cantilever via Hooke's Law (or more complicated analyses that incorporate cantilever geometry). The conceptual model for biopolymer interactions with surfaces and associated AFM approach/retraction data is illustrated in Figure 2.

In general, data from the approach can provide information on the transport step, and data from the retraction can provide information on the attachment step or on the physicochemical properties of the biopolymers (Abu-Lail and Camesano, 2002; Camesano and Abu-Lail, 2002). In data collected to date, the approach curve represents almost exclusively repulsive forces for a wide variety of microbes that have been studied, such as *Burkholderia cepacia* (Camesano and Logan, 2000b; Ginn *et al.*, 2002), *Pseudomonas spp*(Camesano and Abu-Lail, 2002), *E. coli* (Ong *et al.*, 1999; Abu-Lail and Camesano, 2003b; Burks *et al.*, 2003), and *Streptococcus mitis* (Vadillo-Rodriguez *et al.*, 2003). The force curve is also commonly hysteretic; that is, the retraction curve is substantially different from the approach curve (in fact, the retraction curve often shows the presence of *attractive* forces). Such curves have been observed for purified biopolymers (Frank and Belfort,

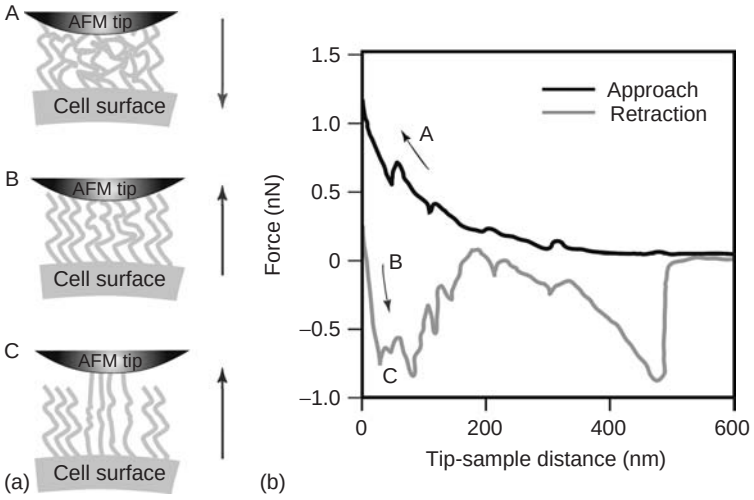


Figure 2 (a) Schematic of biopolymer/surface interactions as measured by atomic force microscopy After (Ginn *et al.*, 2002). (b) An example of the associated force curve showing hysteresis (Adapted from Abu-Lail and Camesano, 2002)

1997), and it is likely that macromolecules on the cell surface are responsible for this attraction between the probe and the microbe. The presence of such macromolecules has been confirmed in some cases by SEM and TEM micrographs. For example, micrographs of *B. cepacia* show a 30–50 nm thick polymer layer in its dehydrated state (e.g. Ginn *et al.*, 2002). Several researchers (e.g. Camesano and Abu-Lail, 2002; Ginn *et al.*, 2002, Section 2.2) have hypothesized that the repulsive force on approach is due to rearrangement of the macromolecules as the tip approaches the surface. Although attractive forces may exist between the tips of individual macromolecules and the probe tip surface, they are overwhelmed by the net repulsive force caused by such steric interactions. On retraction, the force gradually decreases (although not necessarily along the same trajectory as the approach curve). At some point, the net force will reach zero and then begin to become attractive, due to adhesion of the macromolecules to the probe tip. This description is supported by the shape of the retraction curve, which shows some distinctive peaks in the interaction force as individual or clusters of macromolecules break away from the probe surface (Butt *et al.*, 1999).

The complex nature of bacterial surface biomacromolecules has led to much of the difficulty in using simple potential functions to describe the tendency of microorganisms to adsorb to surfaces. In many instances, the length of the surface macromolecules (or surface structures) is greater than the characteristic distance of interaction forces predicted by conventional theories such as the DLVO model. In these instances, the microbial adsorption process is dominated by the interactions between the macromolecules and the solid surface; the bulk of the cell itself may not play a significant role in terms of the interaction forces. Similarly, conventional theories do not predict potential functions that are hysteretic, and this may be a crucial step in understanding microbial adsorption and adhesion. Hysteresis in microbial interactions with solid surfaces is probably due to the interaction of a variety of complex phenomena, including (i) formation of polymer bridges, (ii) biopolymer surface heterogeneity (both in the physical and chemical characteristics of the polymers), and (iii) the dynamic elastic properties of the cell. Hence, the development of a hysteretic interaction potential function will require an understanding of the biomechanics of bacterial cells from a variety of perspectives.

Recent Progress on Modeling and Understanding Microbial Adsorption and Adhesion

Some initial progress in representing the role of exopolymeric substances (EPS) on adsorption and adhesion has been made. Ortiz and Hadziioannou (1999) directly measured the entropic elasticity of individual polymer chains *via* AFM. These data were fitted to two different models of polymer elasticity, the "freely jointed chain" (FJC) model and the "wormlike chain" (WLC) model. Camesano and Abu-Lail (2002) and Abu-Lail and Camesano (2002) performed similar experiments on individual surface polymers of *Pseudomonas putida* KT2442. Their retraction curve data was fitted to the FJC model, and the results indicated that biopolymer heterogeneity on a single cell cannot be explained by differences in molecular weights alone but may be influenced by chemical differences as well. They hypothesized that the presence of multiple polymers with different properties on a cell surface may be the chief cause for the difficulty in predicting bacterial adhesion.

While, some bacterial polymers are well described by the freely jointed chain model (Abu-Lail and Camesano, 2003a), which accounts for the entropic stretching of the polymer chain, other bacterial polymers are better described by the extensible freely jointed chain model, in which entropic and enthalpic effects are considered (Van der Aa *et al.*, 2001). Electrostatic interactions appear to have an indirect effect on bacterial interaction forces, by altering the conformation of the biopolymers (Abu-Lail and Camesano, 2003a). Many bacterial polysaccharides have a negative charge and tend to coil in high salt solutions and extend in lower salt solutions. This type of electrostatic interaction cannot be accounted for with any of the existing DLVO-type models. The ability to measure and quantify conformational changes in bacterial polymers has brought forth a new perspective on cytoscale interaction forces: the standing challenge is to upscale these cytoscale forces, that is, quantify their net macroscopic-scale force, to understand larger scale behavior.

Some inroads on this problem have been made. A number of researchers have investigated the relationship between bacterial polymer properties and attachment likelihood via adhesion experiments on smooth surfaces. The isoelectric point (IEP) of a bacterium was found to be correlated with adhesion to Teflon and glass (Rijnaarts *et al.*, 1995b). An IEP $\leq$ 2.8 indicated the significant presence of cell-surface polysaccharides containing negatively charged phosphate and/or carboxyl groups, which may inhibit adhesion. An IEP $\geq$ 3.2 indicated the absence of polymers that inhibit adhesion. In a detailed study of bacterial surface polysaccharides, lipopolysaccharides (LPS) were extracted from five gram-negative bacterial strains and the adhesion of the isolated LPS to SiO_2, TiO_2, and Al_2O_3 was studied (Jucker *et al.*, 1998b). However, even with detailed information on the physicochemical structure of the LPS molecules, their adhesion could not be predicted.

In summary, the experimental results and theoretical observations of several researchers suggest that the DLVO theory of colloid stability does not adequately describe such interactions when the colloids in question are microbes, and that macromolecules on the cell surface may play an important role in the interactions governing the adhesion of microbes to porous media surfaces. Elucidating the ways in

which cell-surface macromolecules influence the adhesion process is a topic of current research.

Combining Physical and Electrochemical Phenomena Through the Smoluchowski Equation

The transport of microorganisms within a fluid can be described by a convection-dispersion type equation when the suspension is dilute and the particles are far from phase interfaces. However, in porous media the transport microorganisms may be significantly influenced by the presence of solid–fluid interfaces as described above. Under these conditions, the microbial transport phenomena at the subpore scale must be explicitly accounted. Here we introduce a basic approach to combining some of the forces described in the previous section under some simplifying assumptions, the first of which is the absence of hysteresis in the cell-surface force model.

For systems with physically realistic interaction force functions and that do not exhibit hysteresis, it is possible to develop a continuum transport equation that accounts for the particle-surface interactions. From a statistical-mechanical analysis (e.g. Peters, 1990; Peters and Ying, 1991), one can show that for incompressible flows the relevant transport equation for particles takes the form:

Smoluchowski Equation

$$\frac{\partial C_A}{\partial t} + \nabla\cdot(C_A\mathbf{v_o}) = \nabla\cdot(\mathbf{D_o}\cdot\nabla C_A) + \nabla\cdot\left[\left(\frac{1}{kT}\mathbf{D_o}\cdot\nabla\Phi_A\right)C_A\right] \quad \text{in the fluid phase} \quad (2)$$

with the boundary condition

$$\mathbf{n}_{\gamma\kappa}\cdot\left[\mathbf{D}_o\cdot\nabla C_A + \left(\frac{1}{kT}\mathbf{D}_o\cdot\nabla\Phi_A\right)C_A\right] = 0, \quad \text{at the solid–fluid interface} \quad (3)$$

where C_A is the volumetric aqueous particle concentration, $\mathbf{v}_o$ is the velocity field for the particles due only to fluid motion, $\mathbf{D}_o$ is the (position dependent) diffusion tensor for the particles, k is the Boltzmann constant, T is the temperature (in K), and Φ_A is the (position dependent) potential function for the particle-solid surface interactions. In principle, Φ_A can also be a function of the local aqueous and surface concentrations of particles and of time. Such time dependence (if it were known) would allow one to account for changes in the microbe-surface interactions as cell physiology changes in time. The terms of equation (2) correspond to change in local mass (or number) cell concentration, advection, dispersion, and surface interaction, respectively. Note that this equation applies at the interfacial length scale associated with Level I in Figure 1, as well as at the pore scale associated with Level II in Figure 1. As such, the application of equations (2) and (3) in porous media systems is an inherently multiscale exercise. At the subpore scale, equation (2) provides the complete description of the evolution of a dilute suspension of particles. There are two main difficulties in applying equation (2) to applications of microbial transport in porous media: (i) it is a microscale equation, and therefore applies at the pore scale rather than at the Darcy scale; (ii) the length scale associated with the potential function may be much smaller than the length scales associated with convection in the pore space, and this makes the problem difficult to solve.

Upscaling the Smoluchowski Equation to the Darcy Scale

Ideally, the complex physicochemical and biological interactions that are manifest at the molecular, cell, and pore scale would be formally included in a description of cell transport at the Darcy Representative Elementary Volume (REV) scale and above. This connection can in principle be made using various upscaling methods that link subpore-scale (and below) processes to their effective representation at the Darcy scale (and above). However, because of the complexity of the processes there is still substantial research to be done for describing microbial transport in the subsurface. Recent work has been initiated (e.g. Wood *et al.*, 2001, Wood, 2004), and there has been some success in upscaling the Smoluchowski equation (equation 2) to provide a conventional form of a convection-dispersion equation with an effective reaction at the fluid–solid boundary. The essential motivation behind the upscaling is to replace the conservation equations specified by equations (2) and (3), by a simple conservation equation with a reaction boundary condition. In other words, one seeks to replace equations (2) and (3) by the effective representation

$$\frac{\partial C_{A\gamma}}{\partial t} + \nabla\cdot(C_{A\gamma}\mathbf{v}_\gamma) = \nabla\cdot(\mathscr{D}_\gamma\nabla C_{A\gamma}) \quad \text{in the fluid phase} \quad (4)$$

$$-\mathbf{n}_{\gamma\kappa}\cdot(\mathscr{D}_\gamma\cdot\nabla C_{A\gamma}) = k_f C_{A\gamma} - k_r C_{As} \quad \text{at the fluid–solid boundary} \quad (5)$$

$$\frac{\partial C_{As}}{\partial t} = k_f C_{A\gamma} - k_r C_{As} \quad \text{at the fluid–solid boundary} \quad (6)$$

Here, $C_{A\gamma}$ is the concentration of particles in the fluid phase, C_{As} is the surface concentration of particles, $\mathbf{v}_\gamma$ is the average particle velocity in the pore space (note that this is in general different from the fluid velocity that would be observed in the absence of the particles), $\mathscr{D}_{A\gamma}$ is the particle diffusion coefficient, $\mathbf{n}_{\gamma\kappa}$ is the unit normal vector pointing outward from the fluid phase toward the solid phase, k_f is the forward (adhesion) kinetic constant, and k_r is the reverse (detachment) kinetic coefficient.

The transition from equations (2) and (3) to equations (4) to (6) is accomplished by upscaling. Various approaches have been used to determine the forward effective reaction rate parameter (k_f) from the Smoluchowski equation and a variety of boundary conditions; some of the classical works on this topic include studies by Kramers (1940), Brinkman (1956), Ruckenstein and Prieve (1973), Spielman and Friedlander (1974), Rajagopalan and Tien (1976), Shapiro *et al.* (1990), and Song and Elimelech (1993). It should be noted that the classical studies have focused on nonliving colloids, but one recent study (Wood, 2004) has presented an approach for the determination of k_f for cells.

Of special relevance is the upscaling by idealized Happel sphere-in-cell conceptual model (Yao *et al.*, 1971; Rajagopalan and Tien, 1976; Tufenkji and Elimelech, 2004; Nelson and Ginn, 2005), because it underlies CFT that is often used as a conceptual (at least) starting point for bacterial transport in porous media. The latter two recent studies provide a return to the Smoluchowski basis of the Rajagopalan and Tien (1976) approximated analysis, and resolve the equation in the context of the Happel model, in eulerian (Tufenkji and Elimelech, 2004) and lagrangian (Nelson and Ginn, 2005) coordinates respectively. Both find significant but moderate departures in the predicted attachment rate factor (the collision efficiency) in different ways. Thus, the current numerical solutions of the Smoluchowski equation in the simplified and idealized Happel sphere conceptual model exhibit some disparities, although the general trends in the dependence of the attachment rate on pore-water velocity and cell-to-grain size ratio are consistently captured in all three approaches.

As noted in the Section on "Straining and filtration", recent work (Nelson and Ginn, 2005) has shown that the conventional methodology (i.e. equation 1) for upscaling the CFT microscale results contains inconsistencies. The primary difficulty is that it has been common to calculate the rate of particle deposition based on the Happel sphere-in-cell model, while the total rate of particle flux through the idealized pore space is computed based on the isolated sphere model. The different geometries of these models result in a definition of k_f that may yield physically impossible rates of deposition such that more microbes are being deposited on the surface than the number flowing through the pore space. Therefore, it is proposed that the CFT upscaling equation be always expressed so that both the deposition rate and the total flux rate are defined based on the Happel model (which provides a better representation of porous media than the isolated sphere model). A secondary flaw is that the conventional CFT upscaling equation is an approximation of the exact solution. In virtually all cases of practical application, the approximation will be indistinguishable from the exact solution. However, it is conceivable that the approximation may incur significant errors for some experimental systems. Therefore, it is recommended that the exact solution be used. Incorporating both the consistent Happel model definition of fluxes and the exact solution yields the following CFT upscaling equation for the forward effective reaction rate parameter (Nelson and Ginn, 2005),

$$k_f = -\left[\frac{3}{2}\,\frac{(1-\varepsilon)^{1/3}\|v\|}{d}\ln(1-\alpha\eta)\right] \tag{7}$$

Equation (7) multiplied by $C_{A\gamma}$ should replace the right-hand side of equation (1). It is important to note that this definition of k_f requires that the formula for η is also based on the consistent Happel definition of fluxes. This corresponds to the definition given by equation (14) of Rajagopalan and Tien (1976) and equation (16) of Logan *et al.* (1995).

Now we turn to upscaling in the more realistic setting of arbitrarily complex solid surfaces. Although equations (4) to (6) eliminate the interfacial length scale associated with the interaction potential function, it is still a pore-scale representation of microbial transport. In general, a Darcy scale description of microbial transport (at the scale of the large circle in Level III of Figure 1) is the form that would be useful in applications. If one assumes that equations (4) to (6) provide a reasonable representation of the subpore-scale processes affecting microbial transport at the pore scale, it can be shown that in general the Darcy scale representation takes a form that is classically used to describe microbial transport in porous media. Again, neglecting the details we present only the result that has been reported elsewhere (Wood *et al.*, 1999; Ginn *et al.*, 2002).

Darcy Scale Transport Equation

$$\frac{\partial\langle C_{A\gamma}\rangle^{\gamma}}{\partial t} + \langle\mathbf{v}_{\gamma}\rangle^{\gamma}\cdot\nabla\left(\langle C_{A\gamma}\rangle^{\gamma}\right) = \nabla\cdot\left(\mathbf{D}_{eff}\cdot\nabla\langle C_{A\gamma}\rangle^{\gamma}\right) - \frac{a_{\gamma}}{\varepsilon_{\gamma}}\left(k_f\langle C_{A\gamma}\rangle^{\gamma} - k_r\langle C_{A\gamma}\rangle_{\gamma\kappa}\right) \tag{8}$$

Here $\langle C_{A\gamma}\rangle^{\gamma}$ represents the volume averaged intrinsic (pore-water) concentration, $\langle C_{A\gamma}\rangle_{\gamma\kappa}$ is the surface-averaged surface concentration, $\langle\mathbf{v}_{\gamma}\rangle^{\gamma}$ represents the volume averaged intrinsic (pore-water) velocity of the microorganisms, $\mathbf{D}_{eff}$ is the effective dispersion tensor, a_{γ} is the area per unit volume of the fluid–solid interface, ε_{γ} is the porosity, k_f is the adsorption kinetic parameter, and k_r is the desorption kinetic parameter (in the context of linear reversible kinetics for attachment/detachment). Although the averaging that leads to equation (6) as it applies to the Darcy scale has been established, the determination of the effective dispersion tensor for this problem has not yet been studied in detail.

BIOLOGICAL PROCESSES

Biological processes affecting microbial transport are expressed through the growth/decay process and include active adhesion/detachment, chemotaxis, and predation. The biological nature of these processes presents a challenge for transport modeling in that one biological mechanism is often dependent on or influenced by another biological mechanism. Thus, it may be necessary to consider the interdependency among the various biological processes.

Active Adhesion/Detachment

Active adhesion/detachment is treated here as a biologically driven process. Several studies have reported that microorganisms exhibit active adhesion/detachment processes that may be responses to local nutrient availability (Dawson *et al.*, 1981; Kjelleberg and Hermansson, 1984; van Loosdrecht *et al.*, 1989, 1990), survival mechanisms (Dawson *et al.*, 1981; Kjelleberg and Hermansson, 1984; Gilbert and Brown, 1995; Wrangstadh *et al.*, 1990), and/or growth (Jenneman *et al.*, 1985; Reynolds *et al.*, 1989; Sharma *et al.*, 1993). No generally accepted quantitative treatment of dynamic biologically mediated adhesion/detachment processes exists. It is possible that the evidently dynamic attachment/detachment processes are in fact ramifications of EPS mechanics or other cell-driven factors under transient metabolic states. Smets *et al.* (1999) reported experimental results indicating the adhesion of a pseudomonad to glass was significantly more favorable in the exponential growth phase than in the stationary or decay phase, which they hypothesized was due to differences in the cell-surface structure, the cell physicochemistry, or the hydrodynamic behavior of the cells. Thus, the distinction between a microorganism's response to nutrient availability, survival stress, and growth are not necessarily separable or independent processes.

Chemotaxis

Microorganisms that have the capability to move in response to a chemical gradient are termed *chemotactic*. Both random motility (taxis in the absence of a chemical gradient) and chemotaxis have been cited as potential means of transport for subsurface organisms (Barton and Ford, 1995; Corapcioglu and Haridas, 1984; Jenneman *et al.*, 1985; Mercer *et al.*, 1993; Reynolds *et al.*, 1989). Quantitatively, random motility is an effective diffusive flux for microorganisms that depend on the local spatial gradient in aqueous microorganism concentration, whereas chemotaxis is a flux of microorganisms associated with the gradient in nutrient supply. Chemotaxis requires energy and therefore is closely linked to growth processes in porous media. In oligotrophic environments nutrient gradients will be quite small and will likely be associated with either preferential flow paths (if the nutrients arise from recharge) or solid-phase chemical heterogeneity. Chemotaxis may be a very important transport mechanism in these low-nutrient environments. Mercer *et al.* (1993) found that bacteria subjected to oligotrophic conditions displayed enhanced chemotactic response. A contaminant plume will result in large chemical gradients that may also contribute to microbial transport via chemotaxis. Like virtually all microbial characteristics, tactic capability varies widely among organisms. In addition, the chemotactic capability of a given organism can vary depending on the nutrient to which it is being attracted. However, only a small fraction of the many organism-nutrient systems found in natural subsurface conditions and bioremediation schemes has been studied (see Lewis and Ford, 2001 for a compilation of random motility and chemotactic sensitivity coefficients that have been reported for a few different organisms and nutrients). Consequently, these organism-specific and nutrient-specific transport characteristics have not been incorporated into predictive models of microbial transport applicable for field-scale hydrogeological applications.

Much work has been done on developing basic models of chemotactic transport of cell populations in response to gradients in aqueous-phase nutrients. The interested reader is referred to Ford and Cummings (1992), which delineates the relationships between the various models and derives a reduced form of the rigorous three-dimensional cell balance equations of Alt (1980), as well as the review of Ford and Cummings (1998), which includes a comparison of model predictions with experimental results. How to apply our understanding of chemotaxis in a bulk aqueous medium to the modeling of chemotaxis within a porous medium is an open question. However, Barton and Ford (1995, 1997) derived expressions that relate the standard random motility and chemotactic sensitivity coefficients for bulk water to effective values that reflect the impact of the porous medium on an individual cell's swimming behavior. The application of their model to experimental data of bacterial migration through sand cores was consistent with the observed reduction in random motility and provided an explanation (the chemoattractant gradient was too small) for the observed lack of a chemotactic response. The Barton and Ford model modifies the advection-dispersion transport operator of equation (6) to represent random motility and chemotaxis as microbial transport fluxes driven by the random motility coefficient (d_μ) and the chemotactic velocity (v_χ), respectively. This modeling approach is an adaptation for porous media of the classical chemotaxis models reviewed in Ford and Cummings (1992) and Ford and Cummings (1998).

Predation

Tracer studies at field sites have both demonstrated that predation by native protozoan populations on injected

bacteria can have a significant impact on transport distances and concentrations. At South Oyster (described below), a dramatic increase in protozoan populations was observed in response to the injection of bacteria (Choi *et al.*, 2000). Although the total effect is difficult to quantify because of limited observations of attached protozoan populations, it was shown that an increase in apparent attachment (or removal) rate coincided with the timing of the protozoan bloom at a time when a decrease in apparent attachment (or removal) rate would have been expected based on other factors (Zhang *et al.*, 2001b).

Metabolic Effects on Microbial Transport

Some macroscopic-scale evidence suggests that bacterial attachment/detachment kinetics are metabolically mediated, in ways distinct from EPS-associated mechanisms described above. In an experiment conducted in an intermediate-scale flow cell ($100 \times 20 \times 10$ cm dimensions), a substrate pulse resulted in an increase in aqueous-phase bacteria (Murphy *et al.*, 1997, Figure 1), similar to observations in field bioremediation efforts (DOE, 1993; Balkwill *et al.*, 2001). Column experiments have suggested that the response observed in Murphy *et al.* (1997) may be cell-division-mediated transport, a mechanism long recognized in the microbiology literature (cf. references cited in the Section on "Active adhesion/detachment"). Cell-division-mediated transport has also been referred to as *mother–daughter* or *shedding cells* and occurs when the "mother" cell, attached perpendicular to the mineral surface, grows, divides, and the "daughter" cell is released into the aqueous phase (Marshall, 1996). The mother cell remains attached.

Many investigators have noted that starvation or nutrient availability can stimulate a change in the partitioning of a microbial community between the solid and aqueous phases (e.g. Kjelleberg and Hermansson, 1984; Kjelleberg, 1993). A contaminant plume creates a dynamic nutrient environment, but it is not clear whether the corresponding response in partitioning of the microbial community will have any effect at all on the actual contaminant degradation. Therefore, Ginn *et al.* (1998) investigated the relative importance of dynamic partitioning of the bacterial phase on contaminant degradation by numerically modeling the response of a consortium of anaerobic bacteria involved in the degradation of chlorinated hydrocarbons. This example concerns the stimulation of a natural subsurface microbial community that would be initially associated with the solid phase. When substrate is present, as in a contaminant plume, the propionate degrader displays dynamic partitioning (e.g. the forward attachment rate, k_f, changes with the level of metabolic activity). The simulation results showed enhanced degradation under dynamic conditions due to the aqueous partitioning of one member of the consortium that resulted in an increasing population moving with the plume, and hence increasing degradation. This simulation illustrates the importance of the potential partitioning of bacteria under dynamic growth conditions during *in situ* biodegradation.

OVERVIEW OF MICROBIAL TRANSPORT FIELD EXPERIMENTS

Microbial transport processes as described above are conceptualized and defined at the pore to local continuum scale. Most of the experimental work on which these definitions are based has been performed in the laboratory on small porous columns (commonly homogeneously repacked) under one-dimensional flow conditions. Translation of these representations to field-scale transport of bacteria in natural aquifers involves additional factors for consideration, such as the effects of three-dimensional flow fields, multiple scales of heterogeneity in aquifer and groundwater properties, interactions with native microbial communities, and temporal variations due to natural transient fluctuations. These factors confound the straightforward extension of laboratory results to field-scale predictions.

Microorganisms were first used as environmental tracers to delineate subsurface flow pathways, particularly in fractured or karst environments. A number of field studies have also been performed to determine the potential for transport of pathogenic organisms from sources such as septic systems or artificial recharge areas to drinking water supply wells. More recently, local-scale controlled injection and recovery experiments have been conducted at a number of sites to directly study the mechanisms of microbial transport at the field scale in porous aquifers. Harvey and Harms (2001) review the historical use of microbial agents as tracers, and describe several methods used for field-scale bacterial transport injection and recovery tests. Schijven and Hassanizadeh (2000) present a detailed review of viral transport models, processes, and parameters derived from field experiments.

Detailed, local field-scale bacterial transport experiments were pioneered at the United States Geological Survey's Cape Cod Toxic Substances Hydrology Research Site. Several forced- and natural-gradient tests were conducted in saturated glacial outwash sediments using bacteria, viruses, protozoa, inorganic colloids, and solute tracers. Harvey *et al.* (1989) provide an initial overview of transport results using bacteria-sized fluorescent microspheres, native bacteria, and a conservative tracer. A forced-gradient experiment was conducted using a divergent flow design, with multilevel samplers (MLS) 1.7 and 3.2 m from the injection well. A natural-gradient experiment was conducted in a contaminated zone of the aquifer using microspheres only; samples were collected in a row of MLS 6.9 m down-gradient of the injection well. Transported bacteria and microspheres of varying diameter were observed at both sample collection

points in the forced-gradient test, although relative breakthrough was less than 1% at the closest sampling point. Breakthrough of bacteria prior to a conservative bromide tracer was observed, particularly at the distant MLS. Peak abundance of microspheres increased with increasing diameter, qualitatively consistent with filtration theory. Behavior of bacteria and similar-sized microspheres was observed to be significantly different. Harvey and Garabedian (1991) applied filtration theory to quantitatively model transport of bacteria in a subsequent natural-gradient experiment at the same site, and estimated a collision efficiency of 5×10^{-3} to 1×10^{-2} at a travel distance of 6.9 m. Harvey *et al.* (1993) demonstrated the importance of physical heterogeneity in controlling field-scale bacterial transport. Observations of microsphere, bromide, and bacterial transport 6 m down-gradient of the injection point varied significantly in three sampling ports separated by a total vertical distance of 0.7 m. Bales *et al.* (1995) coinjected bromide tracer, viruses (bacteriophage), bacteria, and latex microspheres under natural gradient and observed breakthrough at several multilevel sampling points. The three colloidal materials behaved in a significantly different fashion, with bacteriophage most strongly attenuated by attachment (five orders of magnitude reduction in 6 m travel distance) under ambient pH conditions (5.7). They found that the attached phage could be remobilized by subsequently injecting a high-pH (8.3) pulse of water, consistent with similar laboratory results. Harvey *et al.* (1995) studied the transport of groundwater protozoa at the Cape Cod site, and found that the protozoa were attenuated more rapidly than bacteria although the protozoa were of the optimal size for transport based on experiments with microspheres of various diameters. Pieper *et al.* (1997) performed injections of bacteriophage at the site, and found that attenuation was significantly lower in zones of the aquifer containing high concentration of sewage-derived organic matter. They attributed this observation to the blocking of attachment sites in the organic-rich zone. Similar findings are reported in Harvey *et al.* (1989) and Powelson *et al.* (1991) (for virus tracers), and explored experimentally in Johnson and Logan (1996).

Several subsequent experiments at other sites have further explored field-scale phenomena associated with microbial transport in a variety of geologic settings. DeFlaun *et al.* (1997) report preliminary results of forced-gradient bacterial injection and recovery experiments conducted in a sandy aquifer (marine shoreface sediments) near Oyster, Virginia. Although bacterial breakthrough was observed at the furthest sampling point (4 m from the injection), the large majority of bacteria (greater than 99%) were retained between the injection well and the first sampler only 50 cm away. A dual subpopulation model was proposed as a possible explanation for the observed behavior.

Three forced-gradient bacterial injection and recovery experiments were conducted from 1999 to 2001 at two locations near the town of Oyster, referred to generally as the South Oyster site. These locations are near that of the preliminary experiments reported by DeFlaun *et al.* (1997). Two separate flow cells with different groundwater chemistry were utilized to evaluate the transport of aerobic and facultative iron-reducing bacteria under both oxic and suboxic (reducing) groundwater conditions. Balkwill *et al.* (2001) give an overview of the site characteristics, experiments conducted, and results. Under oxic conditions, bacteria were observed to closely follow high-permeability flow paths, as identified using high-resolution geophysical methods, in a similar manner as a conservative tracer but attenuated by attachment (Hubbard *et al.*, 2001). However, under suboxic conditions, a simple relationship between physical heterogeneity and bacterial transport pathways was not observed. Predictions of the extent of bacterial transport based on laboratory column experiments using intact sediment cores from the site were overly conservative; observed transport at the field scale was much greater than predicted (Scheibe *et al.*, 2001). Estimates of bacterial collision efficiencies based on field-scale modeling decreased as a function of distance traveled (Mailloux *et al.*, 2003).

Schijven *et al.* (2000) injected viruses and bacteria into a deep sandy aquifer in the Netherlands and observed breakthrough at four monitoring wells ranging from 8 to 38 m distant from the injection. Although one bacterium (*E. coli*) was observed only at the closest well, the second (spores of *Clostridium bifermentans*) was observed at all four wells. The attenuation of two bacteriophages was intermediate between that of *E. coli* and *Clostridium* spores. A systematic decrease in attachment rate coefficients with distance was inferred using one-dimensional models of the experiment, with much higher rates of attachment occurring in the first 8 m of transport (attributed to different geochemical conditions near the injection well). Sinton *et al.* (1997) introduced bacteria, bacteriophage, and rhodamine WT dye into an alluvial gravel aquifer near Christchurch, New Zealand and observed transport to wells as far as 400 m down-gradient. Groundwater flow rates in this aquifer are rapid, as high as 86 m day^{-1}. Transport from the ground surface through the vadose zone to the groundwater table was also observed in an irrigation experiment. They represented bacterial attenuation by an apparent decay constant (incorporating attachment losses as well as true bacterial decay), and estimated values on the order of 1.0/day for coliform bacteria in groundwater. Pang *et al.* (1998) describe results of a second experiment in which rhodamine WT, Cl^-, and a bacterium were injected below the water table. They observed faster velocities of the bacteria relative to inert tracers, and estimated total removal rates of 2.4–9.4/day. Sinton *et al.* (2000) further examined the apparent velocities of microbial (bacteria and virus) and inert tracers, and

found that apparent velocity increased, and apparent dispersion decreased, with increasing particle size.

McKay *et al.* (2000) conducted a natural-gradient transport experiment with microspheres, viruses, and bacteria in a fractured shale saprolite aquifer. All of the colloids were detected 13.5 m from the injection well, and the bacteria and microspheres were detected as far as 35 m from the injection well. In this fractured system, microbes exhibited early arrival (at velocities as much as 500 times greater than solute tracers) and variable rates of attenuation corresponding to highly heterogeneous aquifer properties. Becker *et al.* (2003) conducted forced-gradient transport experiments in saturated fractured crystalline bedrock using several bacterial strains and microspheres. They found that relatively small differences in physical characteristics of bacteria (e.g. cell size, Gram type, and motility) corresponded to large differences in field-scale transport behavior.

Conventional Models of Bacterial Attachment/Detachment Kinetics

Modeling of attachment/detachment kinetics is extensively reviewed in Clement *et al.* (1999), Murphy and Ginn (2000), and Harvey and Harms (2001). A brief summary is given here. The linear attachment/detachment kinetic model appearing (as the last two terms on the right-hand side) in equation (8) is perhaps the most basic approach, and reflects the essentially universal treatment of bacterial attachment/detachment as a kinetically controlled process. This model may be viewed as a modification of the CFT approach noted above, wherein a linear detachment process is included (no such detachment occurs in original CFT), and wherein dependence of the rate coefficient for attachment on pore-water velocity is ignored. Murphy *et al.* (1997) extend the linear reversible model to include nonlinear dependence of the rate coefficients on ionic strength of solution, manifest in intermediate-scale experiments. Tan *et al.* (1994), Lindqvist *et al.* (1994), and Saiers and Hornberger (1994) all introduce the classical site-saturation limiting factor on the attachment rate coefficient, in order to account for potential depletion of available surface sites as attached microbe densities increase, which may occur when aqueous microbes cannot attach to attached microbes. Clement *et al.* (1997) performed soil column experiments to evaluate the classical first-order detachment rates of a denitrifying consortium under various growth conditions. The data presented in this work supported the hypothesis that the bacterial detachment rates in porous media systems will increase with increase in the value of specific growth rates. Johnson *et al.* (1995) provide an accounting of the effects of residence time on detachment by zeroing the detachment rate for microbes whose residence time exceeds a particular threshold in a finite difference model of the fate and transport. Ginn (2000a) modeled such non-Markovian (e.g. residence-time-dependent) attachment/detachment kinetics apparent in experiments of McCaulou *et al.* (1994) using the exposure-time approach of Ginn (1999).

Impact of Physical and/or Chemical Heterogeneity

Physical heterogeneity (spatial variations in grain size, permeability, and porosity of the porous medium) clearly plays a significant role in controlling patterns of field-scale transport. Harvey *et al.* (1993) reported a dramatically different character of bacterial breakthrough at three adjacent elevations in a single multilevel sampler, separated by only tens of centimeters. In controlled experiments in two dimensions, Murphy *et al.* (1997) demonstrated how bimodal heterogeneity induced spatial variability in microbial growth through mixing of limiting nutrients and consequently induced significant variations in microbial transport. At the field scale at the South Oyster site, significant effort was invested in characterization of the spatial structure of physical aquifer properties using geological, hydrologic, and geophysical techniques (Chen *et al.*, 2001; Hubbard *et al.*, 1999, 2001). It was demonstrated that transport patterns of both bacteria and bromide were closely related to patterns of permeability and that apparent dispersivities were consistent with correlation lengths inferred from tomographic observations (Hubbard *et al.*, 2001).

Many sandy aquifers contain variable amounts of iron and other metal oxide minerals, commonly in the form of amorphous coatings on quartz grain surfaces. Under pH conditions commonly observed, quartz surfaces are negatively charged while iron oxides are positively charged. This leads to an enhancement of bacterial attachment to surfaces with iron oxide coatings, which has been demonstrated in laboratory experiments (Knapp *et al.*, 1998; Mills *et al.*, 1994; Scholl and Harvey, 1992). However, these effects can be modified at the field scale by variable flow paths caused by physical heterogeneity or by other factors such as blocking of iron oxide sites by attached organic carbon (Johnson and Logan, 1996). Identification of the relative importance of physical and chemical heterogeneity is confounded by cross correlation between the two; iron-rich minerals and coatings tend to be concentrated in finer-grained zones that also have lower permeability and higher collector efficiency.

Apparent Decrease in Attachment with Distance

The apparent rate of attachment (as usually quantified by fitting a collision efficiency parameter or attachment rate coefficient to breakthrough observations) has been observed to decrease with distance traveled in almost all field studies of microbial transport (e.g. Harvey and Garabedian, 1991; Harvey *et al.*, 1995; DeFlaun *et al.*, 1997; Schijven *et al.*, 2000; Woessner *et al.*, 2001). This appears to be the key factor limiting the direct application of laboratory core experiments to predict the extent of field-scale

transport. The phenomenon has occasionally been interpreted as reflecting *in situ* growth (Harvey and Garabedian, 1991) or heterogeneity in aquifer properties (Schijven *et al.*, 2000; Schijven and Hassanizadeh, 2000). Most often, however, it is interpreted as reflecting variability in the cell-surface properties within a monoclonal bacterial population (Albinger *et al.*, 1994; DeFlaun *et al.*, 1997; Bolster *et al.*, 2000; Schijven and Hassanizadeh, 2000), which would cause the more "sticky" bacteria to attach rapidly at short travel distances while a smaller subpopulation with less propensity for attachment transports much farther. Baygents *et al.* (1998) and Glynn *et al.* (1998) observed a bimodal distribution of surface-charge density in selected bacterial strains, providing experimental support for this hypothesis. Redman *et al.* (2001a,b) proposed a multiscale (fractal) distribution of filtration length scales arising from heterogeneity in surface properties of microbes and collectors, and provided experimental support for the model based on column experiments of virus transport. Mailloux *et al.* (2003) estimated a field-scale distribution of apparent collision efficiencies for the oxic flow cell at the South Oyster site. The estimated values clearly decreased with distance, and the distribution inferred from small-scale core experiments was shown to be a subset of the field-scale distribution with the highest values. However, subsequent experiments at the suboxic flow cell conducted by (Dong *et al.*, 2005) (submitted) indicated that intrapopulation variability in collision efficiency is not responsible for the observed decrease in apparent collision efficiency with transport distance at the field scale. Models of bacterial transport in heterogeneous aquifer systems suggest that local-scale heterogeneity in physical and geochemical aquifer properties can lead to apparent enhancement of microbial transport over larger scales, particularly if there exists an inverse correlation between hydraulic conductivity and attachment parameters (Rehmann *et al.*, 1999; Bhattacharjee *et al.*, 2002; Scheibe, 2002). If cell-to-cell surface interactions are more favorable than cell–grain surface interactions, initial attachment of cells near the injection point could increase subsequent attachment of more cells, thereby causing an apparent "ripening" effect that would lead to increased attachment rates nearer the point of injection. However, because attachment parameters for viable cells are not yet in general predictable from basic property measurements, it is also possible that the apparent systematic trend reflects some fundamental shortcoming with the assumed attachment model (generally treated as first-order kinetic). This is especially true when the bacterial population changes from oligotrophic to eutrophic conditions (Murphy *et al.*, 1997). At any rate, it is clear that field-scale transport of microbes will usually be underestimated if predictions are based directly on attenuation rates estimated from laboratory column experiments (Schijven and Hassanizadeh, 2000).

Extended Tailing/Detachment

Typically, rates of detachment of attached cells are low and do not contribute significantly to observed bacterial concentrations during the main breakthrough pulse. However, slow release of attached bacteria at low concentrations over long periods of time may contribute significantly to transport over longer distances. At the South Oyster site, inferred detachment rate coefficients were similar in magnitude to attachment rate coefficients (Zhang *et al.*, 2001b). This contrasts with laboratory experimental results reported in the literature, in which estimated detachment rate coefficients are one to three orders of magnitude smaller than attachment rate coefficients (Zhang *et al.*, 2001b, Table 2). Nevertheless, the inferred detachment rate coefficients were similar to those reported in the literature, so the investigators attribute the observed ratio to unusually low attachment rates of the organism, which had been selected for adhesion deficiency. Numerical simulations using the derived rate coefficients demonstrated that bacterial detachment could have a significant impact on long-term distribution of attached bacteria in the subsurface. Model fits to experimental data also support the hypothesis that there exist both reversible and irreversible attachment events (corresponding to heterogeneity in grain or bacterial cell-surface properties). Zhang *et al.* (2001b) estimated that 70% of the attachment events were irreversible under forced-gradient conditions, increasing to nearly full irreversibility under low-velocity natural-gradient conditions. Other factors that may impact microbial detachment include changes in aquifer geochemistry (Bales *et al.*, 1995), growth of attached bacteria (Murphy *et al.*, 1997), and cell division associated with the edge of an injected groundwater plume (Johnson and McIntosh, 2003).

SUMMARY AND CONCLUSIONS

This paper has reviewed the progress that has been made toward understanding and predicting microbial transport in natural porous media at a variety of characteristic length scales. This review has also revealed the many gaps that still exist in our understanding of the many interrelated processes affecting transport and how these processes fit together. Some of the findings to date and needs for future research are as follows:

1. Aqueous-phase biomass is generally treated as a dilute suspension of cells with no cell-to-cell interactions, an assumption that may sometimes be violated. Attached phase biomass has been treated by different models, and the importance of biofilm spatial structure in porous media is also an area in need of additional research.

2. The physicochemical processes affecting transport have been studied extensively and are reasonably well understood. Although conventional CFT does incorporate many important physical and chemical processes, it also neglects some important ones such as the complex electrochemical phenomena associated with bacterial cells and particle detachment. Attempts to remedy this deficiency have not yielded a definitive approach. Thus, the study of microbial processes as they relate to filtration theory merits further attention.
3. The use of volume averaging to upscale mass transport and reaction processes for both biofilm formation and for microbial transport may be useful in refining our understanding of the macroscopic-scale manifestations of microbial transport and the role of biophase structure in the subsurface. Such formulations will have to be coupled to an improved understanding of the electrochemical processes that affect microbial adhesion at the nanoscale.
4. The effects of multiscale physicochemical aquifer heterogeneity on microbial transport can lead to interesting and complex behavior. The interplay between subsurface heterogeneity and macroscopic-scale transport of microorganisms is an area that is in need of additional research.

The biological processes affecting transport are less well understood. Needs for future work include the following:

1. Development of a transport theory that can account for the space- and time-nonlocal dependence a population of microorganism, particularly as it relates to phenotypical variations caused environmental changes (e.g. nutrient status, quorum sensing).
2. Incorporation of the methods of molecular biology into research pertaining to microbial transport and adhesion. Such methods could help elucidate exactly which mechanisms are responsible for changes in cell-surface properties as environment conditions change.
3. Studies on the possibly coupled nature of the effects of nutrient availability, survival stress, and growth on active adhesion/detachment.
4. Continued cell-scale, pore-scale, and column-scale studies on chemotaxis for relevant bacteria-nutrient systems.
5. Studies on the interactions between bacterial surface polymers and solid surfaces, and methods to reliably quantify cell-surface properties for bacterial species.
6. Increased research on the partitioning of bacteria under dynamic growth conditions and the transient movement of bacteria under changing chemical conditions. Ideally such work would focus on more realistic conditions involving subsurface materials, groundwater analog solutions, and natural consortia of bacterial populations.

Acknowledgments

TRG acknowledges the support of the National Science Foundation under Project No. EAR-0416194, Collaborations in Mathematical Geosciences: Toward Understanding the Transfer of Genetic Information in Subsurface Hydrology. BDW acknowledges the support of the National Science Foundation under Projects No. 0072427, and 0327705. The authors acknowledge and appreciate the review comments of W. P. Johnson by which the manuscript was much improved. TDS acknowledges the support of the Department of Energy under the Natural and Accelerated Bioremediation Research (NABIR) Program. TRG acknowledges support of the National Science Foundation under project numbers EAS/Mathematics (CMG) -0417555 and EAR-0420374.

REFERENCES

Abu-Lail N.I. and Camesano T.A. (2002) Elasticity of *Pseudomonas putida* KT2442 biopolymers probed with single-molecule force microscopy. *Langmuir*, **18**, 4071–4081.

Abu-Lail N.I. and Camesano T.A. (2003a) Role of ionic strength on the relationship of biopolymer conformation, DLVO contributions, and steric interactions to bioadhesion of *Pseudomonas putida* KT2442. *Biomacromolecules*, **4**, 1000–1012.

Abu-Lail N.I. and Camesano T.A. (2003b) The role of lipopolysaccharides in the adhesion, retention, and transport of *Escherichia coli* JM109. *Environmental Science and Technology*, **37**, 2173–2183.

Albinger O., Biesemeyer B.K., Arnold R.G. and Logan B.E. (1994) Effect of bacterial heterogeneity on adhesion to uniform collectors by monoclonal populations. *FEMS Microbiology Letters*, **124**(3), 321–326.

Alt W. (1980) Biased random walk models for chemotaxis and related diffusion approximations. *Journal of Mathematical Biology*, **9**, 147–177.

Bales R.C., Gerba C.P., Grondin G.H. and Jensen S.L. (1989) Bacteriophage transport in sandy soil and fractured tuff. *Applied and Environmental Microbiology*, **55**(8), 2061–2067.

Bales R.C., Li S., Maguire K.M., Yahya M.T., Gerba C.P. and Harvey R.W. (1995) Virus and bacteria transport in a sandy aquifer, Cape Cod, MA. *Ground Water*, **33**(4), 653–661.

Balkwill D., Chen J., DeFlaun M., Dobbs F., Dong H., Fredrickson J., Fuller M., Green M., Ginn T., Griffin T., *et al.* (2001) Breakthroughs in field-scale bacterial transport. *EOS Transactions AGU*, **82**(38), 417 423–425.

Barton J.W. and Ford R.M. (1995) Determination of effective transport coefficients for bacterial migration in sand columns. *Applied and Environmental Microbiology*, **61**(9), 3329–3335.

Barton J.W. and Ford R.M. (1997) Mathematical model for characterization of bacterial migration through sand cores. *Biotechnology and Bioengineering*, **53**(5), 487–496.

Baveye P. and Valocchi A. (1989) An evaluation of mathematical models of the transport of biologically reactive solutes in

saturated soils and aquifers. *Water Resources Research*, **25**(6), 1413–1421.

Baygents J.D., Glynn J.R., Albinger O., Biesemeyer B.K., Ogden K.L. and Arnold R.G. (1998) Variation of surface charge density in monoclonal bacterial populations: Implications for transport through porous media. *Environmental Science and Technology*, **32**(11), 1596–1603.

Becker M.W., Metge D.W., Collins S.A., Shapiro A.M. and Harvey R.W. (2003) Bacterial transport experiments in fractured crystalline bedrock. *Ground Water*, **41**(5), 682–689.

Bhattacharjee S., Ryan J.N. and Elimelech M. (2002) Virus transport in physically and geochemically heterogeneous subsurface porous media. *Journal of Contaminant Hydrology*, **57**(3–4), 161–187.

Bolster C.H., Mills A.L., Hornberger G. and Herman J. (2000) Effect of intra-population variability on the long-distance transport of bacteria. *Ground Water*, **38**(3), 370–375.

Borden R.C. and Bedient P.B. (1987) In situ measurement of adsorption and biotransformation at a hazardous waste site. *Water Resources Research*, **23**(4), 629–636.

Boström M., Williams D.R.M. and Ninham B.W. (2001) Specific ion effects: Why DLVO theory fails for biology and colloid systems. *Physical Review Letters*, **8716**, 168103.

Bowman C. (2002) Nature's way. In *Sacramento Bee*, Sacramento, p. B1.

Brinkman H.C. (1956) Browninan motion in a field of force and the diffusion theory of chemical reactions. *Physica*, **22**.

Burks G.A., Velegol S.B., Paramanova E., Lindenmuth B.E., Feick J.D. and Logan B.E. (2003) Macroscopic and nanoscale measurements of the adhesion of bacteria with varying outer layer surface composition. *Langmuir*, **19**, 2366–2371.

Butt H.-J., Kappl M., Mueller H. and Raiteri R. (1999) Steric forces measured with the atomic force microscope at various temperatures. *Langmuir*, **15**, 2559–2565.

Camesano T.A. and Logan B.E. (2000a) Probing bacterial electrosteric interactions using atomic force microscopy. *Environmental Science and Technology*, **34**(16), 3354–3362.

Camesano T.A. and Logan B.E. (2000b) Imaging modified bacterial cells using atomic force microscopy. *Langmuir*, **16**, 4563–4572.

Camesano T.A. and Abu-Lail N.I. (2002) Heterogeneity in bacterial surface polysaccharides, probed on a single-molecule basis. *Biomacromolecules*, **3**, 661–667.

Chapelle F.H. and Lovley D.R. (1990) Rates of microbial-metabolism in deep coastal-plain aquifers. *Applied and Environmental Microbiology*, **56**, 1865–1874.

Chen J., Hubbard S. and Rubin Y. (2001) Estimating the hydraulic conductivity at the South Oyster Site from geophysical tomographic data using Bayesian techniques based on the normal linear regression model. *Water Resources Research*, **37**(6), 1603–1613.

Chen Y.M., Abriola L.M., Alvarez P.J.J., Anid P.J. and Vogel T.M. (1992) Modeling transport and biodegradation of benzene and toluene in sandy aquifer material – comparisons with experimental measurements. *Water Resources Research*, **28**, 1833–1847.

Choi K.-H., Smith D.W. and Dobbs F.C. (2000) Top-down controls on bacterial transport in oxic and suboxic subsurface environments. *EOS Transaction AGU Supplement*, **81**(48), F184.

Clement T.P., Hooker B.S. and Skeen R.S. (1996a) Numerical modeling of biologically reactive transport from a nutrient injection well. *ASCE Journal of Environmental Engineering*, **122**(9), 833–839.

Clement T.P., Hooker B.S. and Skeen R.S. (1996b) Macroscopic models for predicting changes in saturated porous media properties caused by microbial growth. *Ground Water*, **34**, 934–942.

Clement T.P., Peyton P.M., Skeen R.S., Hooker B.S., Petersen J.M. and Jennings D. (1997) Microbial growth and transport in porous media under denitrification conditions: Experiment and simulations results. *Journal of Contaminant Hydrology*, **24**, 269–285.

Clement T.P., Peyton B.M., Ginn T.R. and Skeen R.S. (1999) Modeling bacterial transport and accumulation processes in saturated porous media: a review. In *Advances in Nuclear Science and Technology*, Lewins J. and Becker M. (Eds.), Vol. 26, Kluwer Academic/ Plenum Publishers: New York, pp. 59–78.

Clement T.P., Truex M.J. and Lee P. (2002) A case study for demonstrating the application of U.S. EPA's monitored natural attenuation screening protocol at a hazardous waste site. *Journal of Contaminant Hydrology*, **59**(1–2), 133–162.

Corapcioglu M.Y. and Haridas A. (1984) Transport and fate of microorganisms in porous media: a theoretical investigation. *Journal of Hydrology*, **72**, 149–169.

Dawson M.P., Humphrey A. and Marshall K.C. (1981) Adhesion: a tactic in the survival strategy of a marine vibrio during starvation. *Current Microbiology*, **6**, 195–199.

de Marsily, G., 1986 *Quantitative Hydrogeology*, translated by G. de Marsily, Academic: San Diego, p. 440.

DeBorde D.C., Woessner W.W., Kiley Q.T. and Ball P. (1999) Rapid transport of viruses in a floodplain aquifer. *Water Research*, **33**(10), 2229–2238.

DeFlaun M.F., Murray C.J., Holben W., Scheibe T., Mills A., Ginn T., Griffin T., Majer E. and Wilson J.L. (1997) Preliminary observations on bacterial transport in a coastal plain aquifer. *FEMS Microbiology Reviews*, **20**, 473–487.

Dodds J. (1982) La chromatographie hydrodynamique. *Analysis*, **10**, 109–119.

Dong H., Onstott T.C., DeFlaun M.F., Fuller M.E., Scheibe T.D., Streger S.H., Rothmel R.K. and Mailloux B.J. (2002) Relative dominance of physical versus chemical effects on the transport of adhesion-deficient bacteria in intact cores from South Oyster, Virginia. *Environmental Science and Technology*, **36**(5), 891–900.

Dong H., Scheibe T.D., Johnson W.P., Monkman C.M. and Fuller M.E. (2005) Change of collision efficiency with transport distance in bacterial transport experiments. *Ground Water*, revised and resubmitted April 2005.

DOE (1993) *Cleanup of VOC's in Non-aride Soils-The Savannah River Integrated Demonstration*., U.S. Department of Energy, Environmental Restoration and Waste Management Office of Technology Development.

Duba A.G., Jackson K.J., Jovanovich M.C., Knapp R.B. and Taylor R.T. (1996a) TCE remediation using in situ, resting-state

bioaugmentation. *Environmental Science and Technology*, **30**, 1982–1989.

Duba A.G., Jackson K.J., Jovanovich M.C., Knapp R.B. and Taylor R.T. (1996b) TCE remediation using resting-state bioaugmentation. *Environmental Science and Technology*, **30**, 1982–1989.

Dufrêne Y.F. (2001) Application of atomic force microscopy to microbial surfaces: from reconstituted cell surface layers to living cells. *Micron*, **32**, 153–165.

Dupin H.J. and McCarty P.L. (2000) Impact of colony morphologies and disinfection on biological clogging in porous media. *Environmental Science and Technology*, **34**, 1513–1520.

Dupin H., Kitanidis P.K. and McCarty P. (2001a) Simulations of two-dimensional modeling of biomass aggregate growth in network models. *Water Resources Research*, **37**(12), 2981–2994.

Dupin H., McCarty P. and Kitanidis P.K. (2001b) Pore-scale modeling of biological clogging due to aggregate expansion: a material mechanics approach. *Water Resources Research*, **37**, 2965–2979.

Dybas M.J., Hyndman D.W., Heine R., Tiedje J., Linning K., Wiggert D., Voice T., Zhao X., Dybas L. and Criddle C.S. (2002) Development, operation, and long-term performance of a full-scale biocurtain utilizing bioaugmentation. *Environmental Science and Technology*, **36**(16), 3635–3644.

Dykaar B.B. and Kitanidis P.K. (1996) Macrotransport of a biologically reacting solute through porous media. *Water Resources Research*, **32**, 307–320.

Emerson R.J. and Camesano T.A. IV (2004) A nanoscale investigation of pathogenic microbial adhesion in biomaterial systems. *Applied and Environmental Microbiology*, **70**, 6012–6022.

Engfield C.G. and Bengtsson G. (1988) Macromolecular transport of hydrophobic contaminants in aqueous environments. *Ground Water*, **26**, 64–70.

Fletcher M. and Floodgate G.D. (1973) An electron-microscopic demonstration of an acid polysaccharide involved in the adhesion of a marine bacterium to solid surface. *Journal of General Microbiology*, **74**, 325–334.

Ford R.M. and Cummings P.T. (1992) On the relationship between cell balance equations for chemotactic cell populations. *SIAM Journal of Applied Mathematics*, **52**(5), 1426–1441.

Ford R.M. and Cummings P.T. (1998) Mathematical models of bacterial chemotaxis. In *Mathematical Modeling in Microbial Ecology*, Koch A.L., Robinson J.A. and Milliken G.A. (Eds.), Chapman & Hall: New York, pp. 228–268.

Frank B.P. and Belfort G. (1997) Intermolecular forces between extracellular polysaccharides measured using the atomic force microscope. *Langmuir*, **13**, 6243–6240.

Gilbert P. and Brown M.R.W. (1995) Some perspectives on preservation and disinfection in the present day. *International Biodeterioration and Biodegradation*, **36**, 219–226.

Ginn T.R. (1999) On the distribution of multi-component mixtures over generalized exposure time in subsurface flow and reactive transport: foundations, and formulations for groundwater age, chemical heterogeneity, and biodegradation. *Water Resources Research*, **35**(5), 1395–1407.

Ginn T.R. (2000a) On the distribution of multi-component mixtures over generalized exposure-time in subsurface flow and reactive transport: batch and column applications involving residence-time distributions and non-Markovian reaction kinetics. *Water Resources Research*, **36**, 2895–2904.

Ginn T.R. (2000b) Comment on "Stochastic analysis of virus transport in aquifers," by L. L. C. Rehmann, Claire Welty, and Ronald W. Harvey. *Water Resources Research*, **36**(7), 1981–1982.

Ginn T.R. (2002) A travel-time approach to exclusion on transport in porous media. *Water Resources Research*, **38**(4), 1041.

Ginn T.R., Scheibe T.D., Murphy E.M., DeFlaun M.F. and Onstott T.C. (1998) Effects of chemical heterogeneity on subsurface fate and transport involving biotic reaction systems: two Examples. *EOS, Abstracts for the 1998 American Geophysical Union Fall Meeting*, **79**, F294.

Ginn T.R., Wood B.D., Nelson K., Scheibe T.D., Murphy E.M. and Clement T.P. (2002) Processes in microbial transport in the natural subsurface. *Advances in Water Resources*, **25**(8–12), 1017–1042.

Glynn, J.R., Belongia B.M., Arnold R.G., Ogden K.L. and Baygents J.C. (1998) Capillary electrophoresis measurements of electrophoretic mobility for colloidal particles of biological interest, *Applied and Environmental Microbiology*, **64**(7), 2572–2577.

Gvirtzman H. and Gorelick S.M. (1991) Dispersion and advection in unsaturated porous media enhanced by anion exclusion. *Nature*, **352**, 793–795.

Happel J. (1958) Viscous flow in multiparticle systems: slow motion of fluids relative to beds of spherical particles. *Journal of American Institute of Chemical Engineers*, **4**(2), 197–201.

Harter T., Wagner S. and Atwill E.R. (2000) Colloid transport and filtration of *Cryptosporidium parvum* in sandy soils and aquifer sediments. *Environmental Science and Technology*, **34**, 62–70.

Harvey R.W. and Garabedian S.P. (1991) Use of colloid filtration theory in modeling movement of bacteria through a contaminated sandy aquifer. *Environmental Science and Technology*, **25**(1), 178–185.

Harvey R.W., George L.H., Smith R.L. and LeBlanc D.R. (1989) Transport of microspheres and indigenous bacteria through a sandy aquifer: Results of natural- and forced-gradient tracer experiments. *Environmental Science and Technology*, **23**(1), 51–56.

Harvey R.W., Kinner N.E., MacDonald D., Metge D.W. and Bunn A. (1993) Role of physical heterogeneity in the interpretation of small-scale laboratory and field observations of bacteria, microbial-sized microsphere, and bromide transport through aquifer sediments. *Water Resources Research*, **29**(8), 2713–2721.

Harvey R.W., Kinner N.E., Bunn A., MacDonald D. and Metge D. (1995) Transport behavior of groundwater protozoa and protozoan-sized microspheres in sandy aquifer sediments. *Applied and Environmental Microbiology*, **61**(1), 209–217.

Harvey, R.W. and Harms H. (2001) Transport of microorganisms in the terrestrial subsurface: in situ and laboratory methods. In *Manual of Environmental Microbiology*, *Second Edition*, Hurst C.J., Crawford R.L., Knudsen G.R., McInerney M.J. and Stetzenbach L.D. (Eds.), pp. 753–776.

Hayashi H., Tsuneda S., Hirata A. and Sasaki H. (2001) Soft particle analysis of bacterial cells and its interpretation of cell

adhesion behaviors in terms of DLVO theory. *Colloids and Surfaces B: Biointerfaces*, **22**, 149.

Hornberger G.M., Mills A.L. and Herman J.S. (1992) Bacterial transport in porous media: evaluation of a model using laboratory observations. *Water Resources Research*, **28**, 915–938.

Hubbard S., Rubin Y. and Majer E. (1999) Spatial correlation structure estimation using geophysical and hydrogeological data. *Water Resources Research*, **35**, 1809–1826.

Hubbard S.S., Chen J., Peterson J., Majer E.L., Williams K.H., Swift D.J., Mailloux B. and Rubin Y. (2001) Hydrogeological characterization of the South Oyster bacterial transport site using geophysical data. *Water Resources Research*, **37**(10), 2431–2456.

Huesemann M.H. and Truex M.J. (1996) The role of oxygen diffusion in passive bioremediation of petroleum contaminated soils. *Journal of Hazardous Materials*, **51**(1–3), 93–113.

Jaffé P.R. and Taylor S.W. (1992) Biofilm growth and the related changes in the physical-properties of a porous-medium 1. *Experimental Investigation – Reply, Water Resources Research*, **28**, 1483–1484.

James S.C. and Chrysikopoulos C.V. (2003) Effective velocity and effective dispersion coefficient for finite-sized particles flowing in a uniform fracture. *Journal of Colloid and Interface Science*, **263**, 288–295.

Jenneman G.E., McInerney M.J. and Knapp R.M. (1985) Microbial penetration through nutrient-saturated Berea sandstone. *Applied and Environmental Microbiology*, **50**(2), 383–391.

Johnson W.P., Blue K.A., Logan B.E. and Arnold R.G. (1995) Modeling bacterial detachment during transport through porous media as a residence-time-dependent process. *Water Resources Research*, **31**(11), 2649–2658.

Johnson W.P. and Logan B.E. (1996) Enhanced transport of bacteria in porous media by sediment-phase and aqueous-phase natural organic matter. *Water Research*, **30**, 923–931.

Johnson W.P. and McIntosh W.O. (2003) Tracking of injected and resident (previously injected) bacterial cells in groundwater using ferrographic capture. *Journal of Microbiological Methods*, **54**(2), 153–164.

Jucker B.A., Harms H. and Zehnder A.J.B. (1998b) Polymer interactions between five gram-negative bacteria and glass investigated using LPS micelles and vesicles as model systems. *Colloids and Surfaces B: Biointerfaces*, **11**(1–2), 33–45.

Jucker B.A., Zehnder A.J.B. and Harms H. (1998a) Quantification of polymer interactions in bacterial adhesion. *Environmental Science and Technology*, **32**(19), 2909–2915.

Kao C.M. and Yang L. (2000) Enhanced bioremediation of trichloroethene contaminated by a biobarrier system. *Water Science and Technology*, **42**, 429–434.

Kim H.S., Jaffe P.R. and Young L.Y. (2004) Simulating biodegradation of toluene in sand column experiments at the macroscopic and pore-level scale for aerobic and denitrifying conditions. *Advances in Water Resources*, **27**, 335–348.

Kjelleberg S. (1993) *Starvation in Bacteria*, Plenum Press: London.

Kjelleberg S. and Hermansson M. (1984) Starvation-induced effects on bacterial surface characteristics. *Applied and Environmental Microbiology*, **48**, 497–503.

Knapp E.P., Herman J.S., Hornberger G.M. and Mills A.L. (1998) The effect of distribution of iron-oxyhydroxide grain coatings on the transport of bacterial cells in porous media. *Environmental Geology*, **33**(4), 243–248.

Kramers H.A. (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, **7**, 284–304.

Lawrence J.R. and Hendry M.J. (1996) Transport of bacteria through geologic media. *Canadian Journal of Microbiology*, **42**, 410–422.

Lewis P. and Ford R.M. (2001) Quantification of random motility and chemotaxis bacterial transport coefficients using individual-cell and population-scale assays. *Biotechnology and Bioengineering*, **75**(3), 292–304.

Lindqvist R., Cho J.S. and Enfield C.G. (1994) A kinetic model for cell density dependent bacterial transport in porous media. *Water Resources Research*, **30**, 3291–3299.

Logan B.E., Jewett D.G., Arnold R.G., Bouwer E.J. and O'Melia C.R. (1995) Clarification of clean-ben filtration models. *Journal of Environmental Engineering*, **121**, 869–873.

MacDonald T.R., Kitanidis P.K., McCarty P.L. and Roberts P.V. (1999a) Effects of shear detachment on biomass growth and in situ bioremediation. *Ground Water*, **37**, 555–563.

MacDonald T.R., Kitanidis P.K., McCarty P.L. and Roberts P.V. (1999b) Mass-transfer limitations for macroscale bioremediation modeling and implications on aquifer clogging. *Ground Water*, **37**, 523–531.

MacQuarrie K.T.B., Sudicky E.A. and Frind E.O. (1990) Simulation of biodegradable organic contaminants in groundwater 1. Numerical formulation in principal directions. *Water Resources Research*, **26**(2), 207–222.

Mailloux B.J., Fuller M.E., Onstott T.C., Hall J., Dong H.L., DeFlaun M.F., Streger S.H., Rothmel R.K., Green M., Swift D.J.P., *et al.* (2003) The role of physical, chemical, and microbial heterogeneity on the field-scale transport and attachment of bacteria. *Water Resources Research*, **39**(6), 1142–1157.

Major D., Edwards E., McCarty P., Gossett J., Hendrickson E., Loeffler F., Zinder S., Ellis D., Vidumsky J., Harkness M., *et al.* (2003) Discussion of environment vs. bacteria or let's play name that bacteria. *Groundwater Monitoring and Remediation*, **23**(2), 33–48.

Marshall, K.C. (1996) Chapter 3, Adhesion as a strategy for access to nutrients. In *Bacterial Adhesion, Molecular and Ecological Diversity*, Fletcher M. (Ed.), John Wiley & Sons, New York, pp. 59–87 361.

Mayotte T.J., Dybas M.J. and Criddle C.S. (1996) Bench-scale evaluation of bioaugmentation to remediate carbon tetrachloride contaminated aquifer materials. *Ground Water*, **34**, 358–367.

McCaulou D.R., Bales R.C. and McCarthy J.F. (1994) Use of short-pulse experiments to study bacteria transport through porous media. *Journal of Contaminant Hydrology*, **15**(1–2), 1–14.

McClaine J.W. (2001) *Influence of Flagellar Rotation on Cell Transport and Adsorption Kinetics for Various Fluid Conditions*. Ph.D. thesis, University of Virginia, Department of Chemical Engineering.

McDowell-Boyer L.M., Hunt J.R. and Sitar N. (1986) Particle transport through porous media. *Water Resources Research*, **22**, 1901–1921.

McKay L.D., Sanford W.E. and Strong J.M. (2000) Field-scale migration of colloidal tracers in a fractured shale saprolite. *Ground Water*, **38**(1), 139–147.

Mercer J.R., Ford R.M., Stitz J.L. and Bradbeer C. (1993) Growth rate effects on fundamental transport properties of bacterial populations. *Biotechnology and Bioengineering*, **42**(11), 1277–1286.

Mills A.L., Herman J.S., Hornberger G.M. and DeJesus T.H. (1994) Effect of solution ionic-strength and iron coatings on mineral grains on the sorption of bacterial cells to quartz sand. *Applied and Environmental Microbiology*, **60**(9), 3300–3306.

Molz F.J., Widdowson M.A. and Benefield L.D. (1986) Simulation of microbial-growth dynamics coupled to nutrient and oxygen-transport in porous-media. *Water Resources Research*, **22**, 1207–1216.

Molz F.J. (1987) Microbial processes and subsurface contamination (meeting report). *EOS Transactions AGU*, **68**, 203.

Murphy E.M. and Ginn T.R. (2000) Modeling microbial processes in porous media. *Hydrogeology Journal*, **8**, 142–158.

Murphy E.M., Ginn T.R., Chilakapati A., Resch C.T., Phillips J.L., Wietsma T.W. and Spadoni C.M. (1997) The influence of physical heterogeneity on microbial degradation and distribution in porous media. *Water Resources Research*, **33**(5), 1087–1103.

Nelson K.E. and Ginn T.R. (2001) Theoretical investigation of bacterial chemotaxis in porous media. *Langmuir*, **17**, 5636–5645.

Nelson K.E. and Ginn T.R. (2005) Colloid filtration theory and the Happel sphere-in-cell model revisited with direct numerical simulation of colloids. *Langmuir*, **21**(6), 2173–2184.

Ninham B.W. (1999) On progress in forces since the DLVO theory. *Advances in Colloid and Interface Science*, **83**, 1–17.

Nyer E.K., Payne F. and Suthersan S. (2003) Environment vs. bacteria or let's play "Name that Bacteria". *Groundwater Monitoring and Remediation*, **23**(1), 36–45.

Ohshima H.K. (1989) Approximate analytical expression for the electrophoretic mobility of colloidal particles with surface-charge layers. *Journal of Colloid and Interface Science*, **130**, 281–282.

Ong Y.-L. and Razatos A. *et al.* (1999) Adhesion forces between *E. coli* bacteria and biomaterial surfaces. *Langmuir*, **15**, 2719–2725.

Ortiz C. and Hadziioannou G. (1999) Entropic elasticity of single polymer chains of poly(methacrylic acid) measured by atomic force microscopy. *Macromolecules*, **32**(3), 780–787.

Otto K., Elwing H. and Hermansson M. (1999) The role of type 1 fimbriae in adhesion of *Escherichia coli* to hydrophilic and hydrophobic surfaces. *Colloids and Surfaces B: Biointerfaces*, **15**, 99–111.

Pang L.P., Close M. and Noonan M. (1998) Rhodamine WT and Bacillus subtilis transport through an alluvial gravel aquifer. *Ground Water*, **36**(1), 112–122.

Paulsen J.E., Oppen E. and Bakke R. (1997) Biofilm morphology in porous media, a study with microscopic and image techniques. *Water Science and Technology*, **36**, 1–9.

Peters M.H. (1990) Adsorption of interacting Brownian particles onto surfaces: II. Results for hydrodynamic interactions. *Journal of Colloid and Interface Science*, **138**, 451–464.

Peters M.H. and Ying R. (1991) Phase-space diffusion equation for single Brownian particle motion near surfaces. *Chemical Engineering Communications*, **108**, 165–184.

Peyton B.M. and Characklis W.G. (1995) Microbial biofilms and biofilm reactors. In *Cell Adhesion: Fundamentals and Biotechnological Applications*, Hjortso M.A. and Roos J.W. (Eds.), Marcel Dekker: New York, pp. 187–231.

Pieper A.P., Ryan J.N., Harvey R.W., Amy G.L., Illangasekare T.H. and Metge D.W. (1997) Transport and recovery of bacteriophage PRD1 in a sand and gravel aquifer: Effect of sewage-derived organic matter. *Environmental Science and Technology*, **31**(4), 1163–1170.

Powelson D.K., Simpson J.R. and Gerba C.P. (1991) Effects of organic matter on virus transport in unsaturated flow. *Applied Environmental Microbiology*, **57**, 2192–2196.

Pyle B.H. (1979) Bacterial movement-experience at Heretaunga. In *Groundwater*, *The Quality and Movement of Groundwater in Alluvial Aquifers of New Zealand, Department of Agriculture Microbiology Technical Publication No. 2*, Noonan M.J. (Ed.), Lincoln College: Canterbury, New Zealand, pp. 105–115.

Pyle B.H. and Thorpe H.R. (1981) Evaluation of the potential for microbiological contamination of an aquifer using a bacterial tracer. *Proceedings Conference On Ground-Water Pollution*, Australian Water Resources Council Conference Series, 1, Sydney, pp. 213–224.

Rajagopalan R. and Tien C. (1976) Trajectory analysis of deep-bed filtration with the sphere-in-cell porous media model. *Journal of American Institute of Chemical Engineers*, **22**(3), 523–533.

Redman J.A., Grant S.B., Olson T.M. and Estes M.K. (2001a) Pathogen filtration, heterogeneity, and the potable reuse of wastewater. *Environmental Science and Technology*, **35**(9), 1798–1805.

Redman J.A., Estes M.K. and Grant S.B. (2001b) Resolving macroscale and microscale heterogeneity in virus filtration. *Colloids and Surfaces A – Physicochemical and Engineering Aspects*, **191**(1–2), 57–70.

Rehmann L.L.C., Welty C. and Harvey R.W. (1999) Stochastic analysis of virus transport in aquifers. *Water Resources Research*, **35**(7), 1987–2006.

Rehmann L.L.C., Welty C. and Harvey R.W. (2000) Reply to comment on "Stochastic analysis of virus transport in aquifers". *Water Resources Research*, **36**(7), 1983–1984.

Reynolds P.J., Sharma P., Jenneman G.E. and McInerney M.J. (1989) Mechanisms of microbial movement in subsurface materials. *Applied and Environmental Microbiology*, **55**(9), 2280–2286.

Rijnaarts H.H.M., Norde W., Lyklema, J. and Zehnder A.J.B. (1995a) The isoelectric point of bacteria as an indicator for the presence of cell surface polymers that inhibit adhesion. *Colloids and Surfaces Benefits: Biointerfaces*, **4**(4), 191–197.

Rijnaarts H.H., Norde W., Lyklema J. and Zehnder A.J. (1999) DLVO and steric contributions to bacterial deposition in media of different ionic strengths. *Colloids and Surfaces B: Biointerfaces*, **14**, 179–195.

Rijnaarts H.H.M., Norde W., Bouwer E.J., Lyklema J. and Zehnder A.J.B. (1995b) Reversibility and mechanism of bacterial adhesion. *Colloids and Surfaces Benefits: Biointerfaces*, **4**(1), 5–22.

Rittmann B. (1993) The significance of biofilms in porous media. *Water Resources Research*, **29**, 2195–2202.

Ruckenstein E. and Prieve D.C. (1973) Rate of deposition of brownian particles under the action of london and double-layer forces. *Journal of the Chemical Society–Faraday II*, **69**, 1522–1536.

Saiers J.E. and Hornberger G.M. (1994) First- and Second-order kinetics approaches for modeling the transport of colloidal particles in porous media. *Water Resources Research*, **30**, 2499–2506.

Scheibe T.D., Chien Y.-J. and Radtke J.S. (2001) Use of quantitative models to design microbial transport experiments in a sandy aquifer. *Ground Water*, **39**(2), 210–222.

Scheibe T.D. (2002) Correlation between heterogeneous bacterial attachment rate coefficients and hydraulic conductivity and impacts on field-scale bacterial transport. *Proceedings of the International Groundwater Symposium*, March 25–28, Berkeley, pp. 440–441.

Scheibe T.D. and Wood B.D. (2003) A particle-based model of size or anion exclusion with application to microbial transport in porous media. *Water Resources Research*, **39**(4), 1080, doi: 10.1029/2001WR001223.

Schijven J.F., Medema G., Vogelaar A.J. and Hassanizadeh S.M. (2000) Removal of microorganisms by deep well injection. *Journal of Contaminant Hydrology*, **44**(3–4), 301–327.

Schijven J.F. and Hassanizadeh S.M. (2000) Removal of viruses by soil passage: overview of modeling, processes, and parameters. *Critical Reviews in Environmental Science and Technology*, **30**(1), 49–127.

Scholl M.A. and Harvey R.W. (1992) Laboratory investigations on the role of sediment surface and groundwater chemistry in transport of bacteria through a contaminated sandy aquifer. *Environmental Science and Technology*, **26**, 1410–1417.

Semprini L. and McCarty P.L. (1992) Comparison between model simulations and field results for in-situ biorestoration of chlorinated aliphatics: Part 2. Cometabolic transformations. *Ground Water*, **30**, 37–44.

Semprini L., Hopkins G.D., McCarty P.L. and Roberts P.V. (1992) In-situ biotransformation of carbon tetrachloride and other halogenated compounds resulting from biostimulation under anoxic conditions. *Environmental Science and Technology*, **26**, 2454–2460.

Seymour J.D., Gage J.P., Codd S.L. and Gerlach R. (2004) Anomalous fluid transport in porous media induced by biofilm growth. *Physical Review Letters*, **93**, 198103.

Shapiro M., Brenner H. and Guell D.C. (1990) Accumulation and transport of Brownian particles at solid surfaces: Aerosol and hydrosol deposition processes. *Journal of Colloid and Interface Science*, **136**, 552–573.

Sharma P.K., McInerney M.J. and Knapp R.M. (1993) In situ growth and activity and modes of penetration of Escherichia coli in unconsolidated porous materials. *Applied and Environmental Microbiology*, **59**, 3686–3694.

Sharp R.R., Cunningham A.B., Komlos J. and Billmayer J. (1999) Observation of thick biofilm accumulation and structure in porous media and corresponding hydrodynamic and mass transfer effects. *Water Science and Technology*, **39**, 195–201.

Shonnard D.F., Taylor R.T., Hanna M.L., Boro C.O. and Duba A.G. (1994) Injection-attachment of methylosinus trichosporium OB3b in a two-dimensional miniature sand-filled aquifer simulator. *Water Resources Research*, **30**(1), 25–36.

Simoni S.F., Harms H., Bosma T.N.P. and Zehnder A.J.B. (1998) Population heterogeneity affects transport of bacteria through sand columns at low flow rates. *Environmental Science and Technology*, **32**(14), 2100–2105.

Sinton L.W., Finlay R.K., Pang L. and Scott D.M. (1997) Transport of bacteria and bacteriophages in irrigated effluent into and through an alluvial gravel aquifer. *Water Air and Soil Pollution*, **98**(1–2), 17–42.

Sinton L.W., Noonan M.J., Finlay R.K., Pang L. and Close M.E. (2000) Transport and attenuation of bacteria and bacteriophages in an alluvial gravel aquifer. *New Zealand Journal of Marine and Freshwater Research*, **34**(1), 175–186.

Sirivithayapakorn S. and Keller A. (2003) Transport of colloids in saturated porous media: a pore-scale observation of the size exclusion effect and colloid acceleration. *Water Resources Research*, **39**(4), 1109.

Smets B.F., Grasso D., Engwall M.A. and Machinist B.J. (1999) Surface physicochemical properties of *Pseudomonas fluorescens* and impact on adhesion and transport through porous media. *Colloids and Surfaces B: Biointerfaces*, **14**, 121–139.

Song L. and Elimelech M. (1993) Dynamics of colloid deposition in porous media: modeling the role. *Colloids and Surfaces A: Physiochemical and Engineering Aspects*, **73**, 49–63.

Spielman L.A. and Friedlander S.K. (1974) Role of the electrical double layer in particle deposition by convective diffusion. *Journal of Colloid and Interface Science*, **46**, 23–31.

Sudicky E.A., Schellenberg S.L. and MacQuarrie K.T.B. (1990) Assessment of the behaviour of conservative and biodegradable solutes in heterogeneous porous media. In *Dynamics of Fluids in Hierarchical Porous Media*, Cushman J.H. (Ed.), Academic Press: New York, pp. 429–462.

Tan Y., Gannon J.T., Baveye P. and Alexander M. (1994) Transport of bacteria in an aquifer sand: experiment and model simulations. *Water Resources Research*, **30**, 3243–3252.

Tufenkji N. and Elimelech M. (2004) Correlation equation for predicting single-collector efficiency in physicochemical filtration in saturated porous media. *Environmental Science and Technology*, **38**(2), 529–536.

Taylor S.W. and Jaffé P.R. (1990) Substrate and biomass transport in a porous-medium. *Water Resources Research*, **26**, 2181–2194.

Vadillo-Rodriguez V. and Busscher H.J. *et al.* (2003) On relations between microscopic and macroscopic physicochemical properties of bacterial cell surfaces: an AFM study on *streptococcus mitis* strains. *Langmuir*, **19**, 2372–2377.

Van der Aa B.C., Michel R.M., *et al.* (2001) Stretching cell surface macromolecules by atomic force microscopy. *Langmuir*, **17**, 3116–3119.

van Genuchten M.T. and Wierenga P.J. (1976) Mass transfer studies in sorbing porous media: I analytical solutions. *Soil Science Society of America Journal*, **40**, 473–480.

van Loosdrecht M.C.M., Lyklema J., Norde W. and Zehnder A.J.B. (1989) Bacterial adhesion: physicochemical approach. *Microbial Ecology*, **17**, 1–15.

van Loosdrecht M.C.M., Lyklema J., Norde W. and Zehnder A.J.B. (1990) Influence of interfaces on microbial activity. *Microbiological Reviews*, **54**(1), 75–87.

Vandevivere P. and Baveye P. (1992) Saturated hydraulic conductivity reduction caused by aerobic bacteria in sand columns. *Soil Science Society of America Journal*, **56**, 1–13.

van Oss C.J. (1994) *Interfacial forces in aqueous media*, Marcel Dekker: New York.

Vayenas D.V., Michalopoulou E., Constantinides G.N., Pavlou S. and Payatakes A.C. (2002) Visualization experiments of biodegradation in porous media and calculation of the biodegradation rate. *Advances in Water Resources*, **25**, 203–219.

Viegeant M.A.S. and Ford R.M. (1997) Interactions between motile *Escherichia coli* and glass in media with various ionic strengths, as observed with a three-dimensional-tracking microscope. *Applied and Environmental Microbiology*, **63**(9), 3474–3479.

Wan J., Tokunaga T.K. and Tsang C.F. (1995) Bacterial sedimentation through a porous medium. *Water Resources Research*, **31**, 1627–1636.

Ward J.P., King J.R., Koerber A.J., Williams P., Croft J.M. and Sockett R.E. (2001) Mathematical modelling of quorum sensing in bacteria. *IMA Journal of Mathematics Applied in Medicine and Biology*, **18**, 263–292.

Widdowson M.A., Molz F. and Benefield L. (1988) A numerical transport model for oxygen- and nitrate-based respiration linked to substrate and nutrient availability in porous media. *Water Resources Research*, **24**, 1553–1565.

Widdowson M.A. (1991) An evaluation of mathematical-models of the transport of biologically reacting solutes in saturated soils and aquifers – comment. *Water Resources Research*, **27**, 1375–1378.

Wiedemeier T.H., Swanson M.A., Wilson J.T., Kampbell D.H., Miller R.N. and Hansen J.E. (1995) Patterns of intrinsic bioremediation at two U.S. Air Force Bases. In *Third International In Situ and On-Site Bioreclamation Symposium*, Hinchee R.E., *et al.* (Eds.), Battelle Press: Columbus, pp. 31–51.

Williams V. and Fletcher M. (1996) Adhesion and transport through porous media of *Pseudomonas fluorescens* is affected by lipopolysaccharide composition. *Applied and Environmental Microbiology*, **62**, 100–104.

Woessner W.W., Ball P.N., DeBorde D.C. and Troy T.L. (2001) Viral transport in a sand and gravel aquifer under field pumping conditions. *Ground Water*, **39**(6), 886–894.

Wood B.D. (2004) Calculation of the darcy-scale effective diffusion and dispersion tensors for porous media: volume averaging with 3-dimensional closure. *EOS Transactions AGU, Fall Meeting Supplement, 85*, Abstract H32A-05.

Wood B.D., Dawson C.N., Szecsody J.E. and Streile G.P. (1994) Modeling contaminant transport and biodegradation in a layered porous media system. *Water Resources Research*, **30**, 1833–1845.

Wood W.W. and Ehrlich G.G. (1978) Use of baker's yeast to trace microbial movement in ground water. *Ground Water*, **16**, 340–398.

Wood B.D., Ginn T.R. and Dawson C.N. (1995) Effects of microobial lag in contaminant transport and biodegradation modeling. *Water Resources Research*, **31**, 553–563.

Wood B.D., Quintard M., Whitaker S. and Minard K. (2001) Bioremediation in porous media: upscaling from the pore to the continuum scales via volume averaging. *EOS (Abstracts For The AGU Fall Union Meeting)*, **82**, 573.

Wood B.D., Quintard M., Golfier F. and Whitaker S. (2002) Biofilms in porous media: development of macroscopic transport equations via volume averaging with closure. In *Computational Methods in Water Resources*, Vol. 2, Hassanizadeh S.M., *et al.* (Eds.), Elsevier: Amsterdam, pp. 1195–1202.

Wood B.D. and Whitaker S. (1998) Diffusion and reaction in biofilms. *Chemical Engineering Science*, **53**, 397–425.

Wood B.D. and Whitaker S. (1999) Cellular growth in biofilms. *Biotechnology Bioengng.*, **64**, 656–670.

Wood B.D. and Whitaker S. (2000a) Erratum to "Diffusion and reaction in biofilms", chemical engineering science 53 (1998) 397–425. *Chemical Engineering Science*, **55**, 2349.

Wood B.D. and Whitaker S. (2000b) Multi-species diffusion and reaction in biofilms. *Chemical Engineering Science*, **55**, 3397–3418.

Wrangstadh M., Szewzyk U., Oestling J. and Kjelleberg S. (1990) Starvation-specific formation of a peripheral exopolysaccharide by a marine Pseudomonas sp., strain S9. *Applied and Environmental Microbiology*, **56,**, 2065–2072.

Yao K.-M., Habibian M.T. and O'Melia C.R. (1971) Water and wastewater filtration: concepts and applications. *Environmental Science and Technology*, **57**(11), 1105–1112.

Zhang P., Johnson W.P., Piana M.J., Fuller C.C. and Naftz D.L. (2001a) Potential artifacts in interpretation of differential breakthrough of colloids and dissolved tracers in the context of transport in a zero-valent iron permeable reactive barrier. *Ground Water*, **39**(6), 831–840.

Zhang P., Johnson W.P., Scheibe T.D., Choi K.-H., Dobbs F.C. and Mailloux B.J. (2001b) Extended tailing of bacteria following breakthrough at the narrow channel focus area, Oyster, Virginia. *Water Resources Research*, **37**(11), 2687–2698.

Zysset A., Stauffer F. and Dracos T. (1994) Modeling of reactive groundwater transport governed by biodegradation. *Water Resources Research*, **30**(8), 2423–2434.

NOMENCLATURE

Roman Letters

a_γ	is the area per unit volume of the fluid–solid interface (m^{-1})
$C_{A\gamma}$	= the concentration of particles in the fluid phase (kg m^{-3})
C_{As}	= the surface concentration of particles (kg m^{-2})
$\langle C_{A\gamma}\rangle^{\gamma}$	= volume averaged intrinsic (pore-water) concentration (kg m^{-3})

$\langle C_{A\gamma}\rangle_{\gamma\kappa}$ = surface-averaged surface concentration ($\mathrm{kg\,m^{-2}}$)
$\mathscr{D}_{A\gamma}$ = microbial diffusion coefficient in the fluid phase ($\mathrm{m^2\,s^{-1}}$)
$\mathbf{D_o}$ = (position dependent) diffusion tensor for particles (including microorganisms) ($\mathrm{m^2\,s^{-1}}$)
d_μ = random motility coefficient ($\mathrm{m^2\,s^{-1}}$)
d_s = diameter of microorganism (m)
d = diameter of particle (collector, m)
g = acceleration due to gravity ($\mathrm{m\,s^{-2}}$)
k_B = Boltzmann constant ($\mathrm{JK^{-1}}$)
k_f = forward (adhesion) kinetic constant ($\mathrm{s^{-1}}$)
k_r = reverse (detachment) kinetic coefficient ($\mathrm{s^{-1}\,m^{-1}}$)
$\mathbf{v}$ = pore-water velocity ($\mathrm{m\,s^{-1}}$)
$\mathbf{v}_\mathrm{o}$ = the velocity field for the particles due only to fluid motion ($\mathrm{m\,s^{-1}}$)
$\langle \mathbf{v}_\gamma\rangle^\gamma$ = volume averaged intrinsic (pore-water) velocity of the microorganisms ($\mathrm{m\,s^{-1}}$)
v_s = sedimentation velocity ($\mathrm{m\,s^{-1}}$)

Greek Letters

α = sticking coefficient in filtration theory
ε_γ = porosity (volume fraction of the fluid phase)
η = collision factor in filtration theory
ρ = fluid density ($\mathrm{kg\,m^{-3}}$)
ρ_s = biocolloid particle density ($\mathrm{kg\,m^{-3}}$)
ω = generalized exposure time (here residence time; s)
Φ_A = the (position dependent) potential function for the microbe-solid surface interactions ($\mathrm{kg\,m^{-2}\,s^{-1}}$)

106: Groundwater Microbial Communities

JAMES P MCKINLEY[1], JAMES K FREDRICKSON[2] AND FREDERICK S COLWELL[3]

[1] *Chemical Sciences Division, Pacific Northwest National Laboratory, Richland, WA, US*
[2] *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, US*
[3] *Biological Sciences Division, Idaho National Laboratory, Idaho Falls, ID, US*

Microbial communities in aquifers consist of diverse interactive individuals which break down complex organic matter for metabolic energy. Microbes are adapted to function over environmental conditions ranging from freezing to boiling, acidic to alkaline. They can use oxygen as a reducible metabolite during organic carbon oxidation, but, since oxygen is rapidly depleted in subsurface environments, different groups of organisms may also rely upon other compounds such as reducible metals or upon fermentation. Community members are interdependent, but compete for resources, and communities often have predominant groups that rely on recognizable chemical pathways such as sulfate reduction. The predominant group varies spatially and temporally, as the available nutrients change or are depleted.

INTRODUCTION

We will address the classification of microbial communities in groundwater, their relation to aquifer geochemistry, and some approaches to investigating microbial effects on groundwater chemistry. Other sections discuss microbial activity with respect to water quality (*see* **Chapter 91, Water Quality, Volume 3**), nutrient cycling (*see* **Chapter 96, Nutrient Cycling, Volume 3**), water quality remediation, and the degradation of contaminants. Microbes catalyze oxidation-reduction reactions to gain energy for metabolism. The different forms of microbial life are widely distributed, and the local abundance of the individual forms is determined by their environment (Stevens, 2002). In aquifers, the substrates for metabolism may be organic or inorganic compounds, but most communities rely ultimately on the oxidation of complex organic compounds deposited along with the solid components of the aquifer. Individual microbes occupy relatively narrow niches in the trophic web that transforms complex organic matter to carbon dioxide, and in natural groundwater systems the community of individuals will be in some sort of equilibrium with its surroundings. Characterization of the environment according to relevant chemical and physical parameters should therefore provide an understanding of the microbial community. Because the microbial community is diverse and adaptive and the environment is heterogeneous and complex, studies of microbial communities usually include microbial and geochemical components.

In an undisturbed environment, the equilibrium community will contain a stable, disparate population that can efficiently exploit and utilize the nutrients available to it. When a disturbance occurs, the population will change to exploit changes in environmental conditions and will evolve to regain an equilibrium condition. Also, the interplay between the microbial community and its environment affects the environment itself, so that each is driven eventually to a condition that is relatively stable. This can occur after abrupt disturbances such as contamination events, or during longer-term changes such as the exhaustion of a particular natural nutrient.

The microbial community is composed of individuals living within the limitations of their metabolic capability, competing with other individuals to most efficiently utilize the available nutritional resources in an aquifer, and also cooperating with other microbes to the mutual benefit of various community elements. Because of the qualitative interdependence of microbes and environment, microbes can be classified according to the environments that harbor them and the chemical transformations they effect.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

MICROBIAL CLASSIFICATION

Microbial taxonomy may be based on properties such as cell morphology, staining characteristics, biochemical transformations, or molecular biology (Chapelle, 1993). Molecular methods have recently become powerful and accessible tools for characterizing microbes at structural and functional levels. Using these methods, changes in nucleotide sequences in ribosomal RNA (16S rRNA) have been used to measure the evolutionary relationships among cells, defining three distinct cellular lineages in nature, the *Bacteria, Archaea*, and *Eukarya* (Woese, 2002). Organisms within these groups may be further associated into subgroups. Bacteria, for example, have been resolved into 12 major lineages by ribosomal RNA sequencing (Madigan *et al.*, 1997). In applications to microbial ecology, the molecular methods liberate the investigator from cultivation-based characterization and provide the means to determine the microbial composition of any environmental niche. The use of these genetic associations is limited, however, by the limited ability to predict physiology based on phylogeny; identifying which microorganisms are present does not say what they are doing (Embley and Stackebrandt, 1996).

Among the three evolutionary branches, *Bacteria* comprise most of the commonly encountered microbes and can have profound effects on the aquifer environment. They inhabit oxic and anoxic aquifer sediments and rely on organic chemicals or inorganic chemicals for energy. Most *Archaea* are anaerobes, and may thrive under unusual conditions such as hot springs. They also can transform the subsurface environment and introduce chemical components through metabolism. The *Bacteria* and *Archaea* usually coexist in aquifer systems and may be indistinguishable in their environmental effects. Microbes within the *Eukarya* are the least significant with respect to aquifers and aquifer chemistry. They include algae, fungi, and protozoa. Algae are photosynthetic and do not function in the subsurface, although they may be present there (Sinclair and Ghiorse, 1989). Fungi are not abundant in subsurface ecosystems (Madsen and Ghiorse, 1993). Protozoa are motile and feed on other organisms or organic particles and have a limited effect on the aquifer environment (Kinner *et al.*, 2002).

Descriptive, practical classification systems are based on microbial attributes that can be observed in the field or laboratory. These include metabolic processes, favored physical conditions, tolerance to oxygen, and the inorganic chemical transformations they make in the environment.

METABOLISM

Metabolically, microbes acquire energy through an electron donor, such as organic carbon, that can provide electrons and become oxidized. They also require an electron acceptor, a substance that can accept electrons and become reduced. The net energy gained in an oxidation-reduction reaction is used for metabolic function and growth. Microbes are classified metabolically according to their limitations in obtaining organic carbon for cell mass during this process. For example, microbes that require organic compounds as a carbon source are termed *heterotrophs. Autotrophs* can utilize CO_2 as the sole carbon source, while catalyzing inorganic reactions for energy. *Mixotrophs* utilize inorganic substances as energy sources and organic carbon as a source of cellular carbon. *Chemolithotrophs* obtain energy by oxidizing reduced inorganic chemicals, and *photolithotrophs* obtain energy from light. *Fermenters* use organic compounds as donors and acceptors.

PHYSICAL ENVIRONMENT

The range of conditions in the subsurface places limits on the existence of microbial life, and in general, the subsurface seems unsuited to microbes. Photosynthesis is impossible, organic carbon and other nutrients are usually scarce, the temperature may be extreme, and interpore space may be cramped. Nonetheless, microbial life has been observed in environments where pore spaces exceed 0.5 μm in diameter (Fredrickson *et al.*, 1997). Microbes have adapted to live not only where conditions are moderate and nutrients abundant, but also where temperatures, pressure, acidity, and energy sources/sinks are extreme, in all lithologic environments (Colwell, 2001; Stevens, 2002). The preferred environment may thus be used to describe microbes. Examples include microbes observed at temperatures as high as 113 °C (Stetter *et al.*, 1993): they are *thermophilic*; to grow at pH below 2 (Reysenbach *et al.*, 2002; Tyson *et al.*, 2004): they are *acidophilic*; above pH 10 (Grant and Tindall, 1986): they are *alkaliphilic*; and are found in deep-sea environments at pressures up to >100 MPa (Yayanos, 1995): they are *barophilic*. The deep-sea environment at mid-ocean ridges harbors bacteria in an environment that is at high pressure and ranges in temperature from below freezing (i.e. the bacteria are *psychrophilic*) to above boiling (Horikoshi and Tsujii, 1999).

OXYGEN TOLERANCE

Dissolved oxygen, coupled to an electron donor as an oxidant in an oxidation-reduction reaction, provides the highest yield of metabolic energy. Because oxygen is only sparingly soluble in water (saturation occurs at approximately 9 mg of O_2 per liter), it is rapidly depleted in most aquifer environments. Many microbes have evolved to function in environments where oxygen has been removed and may not tolerate the presence of dissolved oxygen. Microbes may therefore be described according to their oxygen tolerance. Those that use oxygen are *aerobic* and those that use

alternative compounds as oxidants are *anaerobic*. In addition, aerobes may be *obligate*, requiring oxygen; *facultative*, preferring but not requiring oxygen; or *microaerophilic*, requiring oxygen at levels lower than saturated. Anaerobes also may be *aerotolerant*, not requiring oxygen and growing no better when oxygen is present; or *obligate*, requiring an absence of oxygen to avoid damage or death (Madigan *et al.*, 1997).

CHEMICAL FUNCTION

A fourth useful classification system is based on the chemical transformations that microbes carry out. The study of microbial ecology began with investigations of chemical transformation by unidentified natural organisms. As early as 1839, the ability of microbes entrained in soil to oxidize hydrogen gas was established, although hydrogen-oxidizing bacteria were isolated much later. Toward the end of the nineteenth century, individual organisms were isolated from natural materials and identified by their ability to accomplish inorganic chemical transformations. The subsurface microbes were thus usually described not by their shape or genetics or other taxonomic features, but by their relative chemical function. In addition, since numerous species can effect the same inorganic transformation, the microbes were organized for descriptive purposes according to the observable chemical transformations with which they are associated in natural environments, regardless of how closely they were related genetically (which was unknown). These groupings are called *functional groups* or *guilds* such as "sulfate reducers" or "methanogens". Functional groups provide a convenient and useful means of describing and investigating microbial communities in groundwater because they are directly dependent on the chemical environment and make many chemical transformations that can be observed in groundwater.

MICROBIAL COMMUNITIES

Within the community ecosystem, there are numerous chemical pathways and chemical resources that can be exploited for energy, and multiple individual species may use the same chemical resources for metabolic energy. The population grouped into guilds comprise the community's guild structure (Atlas and Bartha, 1993), where, for example, a guild may consist of all species that utilize sulfate reduction as a metabolic pathway. Thus, while diversity is necessary for community stability, the loss of a single species will not disrupt the overall community structure because multiple species may serve the same function.

The microorganisms within the community compete for resources, and they have a number of survival strategies that allow them to compete and maintain themselves. They may, for example, rely upon high reproduction rates (Andrews and Hall, 1986). An organism with a rapid reproduction rate would dominate when resources were temporarily abundant, where the benefits of reproduction outweigh other competitive adaptations. This could be important in ecosystems with a seasonal or other temporal fluctuation in nutrient availability, and growth-adapted organisms might be expected to flourish in remediation scenarios where the disturbance of the local environment causes an influx of nutrients. In subsurface environments where the supply of resources is relatively constant, organisms compete by having adaptations that allow them to efficiently utilize specific nutrients. In resource-limited environments especially, these organisms would be expected to be successful.

Microbial communities are structured and function according to their composition, dominant species, and population dynamics (Atlas and Bartha, 1993). Different microbial communities occupy different environments, members of a community compete and are interdependent, and the community members respond to environmental changes. In the application of microbial ecology to hydrology, we are perhaps most interested in the function of subsurface microbial communities; that is, the effect the microbial community has on the hydrologic system. Chemical properties and variations are readily obtainable and are often understandable directly in terms of the microbial community's effect on the aquifer environment.

TERMINAL ELECTRON-ACCEPTING PROCESSES

We can compare organisms and their relative dominance in subsurface systems according to the chemical reactions they mediate and the amount of energy they derive from these reactions. To do this, we must establish a uniform basis for comparison.

In groundwater environments, microbes derive metabolic energy from remnant clastic organic carbon or from electron-donating inorganic compounds. The vastly predominant energy source is photosynthesized remnant organic carbon deposited during sedimentation. Photosynthesis proceeds by the disproportionation of energy into localized centers of highly negative $p\varepsilon$ and a reservoir of O_2 (Stumm and Morgan, 1996), where $p\varepsilon$ is the relative tendency of a solution to accept or transfer electrons. Photosynthesis thus imposes a state of disequilibrium on the chemical components that comprise the microbial ecosystem. Non-photosynthetic organisms exploit this disequilibrium, and act to restore it, by using enzymes to decompose the unstable organic products of photosynthesis. They may do this by respiration, using an external acceptor (e.g. oxygen or sulfate) for electrons liberated from electron donors, or by fermentation, in which an organic compound is oxidized

(electron removed) inside a cell and the liberated electron is transferred to another, partially oxidized compound to reduce it (Ehrlich, 1990); that is, a complex organic molecule is split into relatively oxidized and reduced parts. In either case, the microbes facilitate thermodynamically feasible chemical reactions, realizing a fraction of the available free energy for metabolic purposes.

The rapid depletion of oxygen forces anaerobic metabolism to prevail within most groundwater systems, and the geochemically significant reactions are those that involve inorganic terminal electron acceptors such as sulfate. Measurements with respect to microbial community function are therefore based on chemical species that incorporate the products or reactants within microbially catalyzed chemical reactions. For example, in the examination of a subsurface environment where sulfate-reducing bacteria are active, the overall reaction can be represented by

$$2CH_2O + SO_4^{2-} + H^+ + 2H_2O = HS^- + 2CO_2 + 4H_2O \quad (1)$$

Of the reactants and products, the obvious target for the determination that sulfate is reduced is dissolved sulfide, since sulfide is not a common abiotic component of groundwaters. Sulfate reduction will not occur in the absence of sulfate, of course, and the abundance of the electron acceptor could be an indicator of the potential for sulfate reduction (Stevens and McKinley, 1995; Stevens *et al.*, 1993).

The composition of subsurface microbial communities is strongly linked to the chemical environment, which controls the relative abundance of potential electron acceptors and donors (Berner, 1981; Chapelle, 1993; Chapelle *et al.*, 1995). The carbon flow in subsurface environments is regulated by microbial food chains. No individual microbe has the capability to degrade all organic compounds present in groundwater environment, so individual microbes are specialized to perform particular metabolic processes (Chapelle, 1993). In most environments, organic matter degradation of such materials as lignin is begun by fermentation, yielding simpler compounds that can be broken down by other organisms, and the community of microorganisms accomplishes complete degradation to CO_2, with the final degradation step represented by terminal electron-accepting processes (TEAPs). Degradation is accomplished by microbial consortia (Chapelle *et al.*, 1987). The most common TEAPs in groundwater systems, constructed from constituent half-reactions (Champ *et al.*, 1979; Stumm and Morgan, 1996), are listed in Table 1 in order of net energy yield. (The derivation of free energies for comparison is summarized in the Appendix.) In environments where the microbial community has reached an equilibrium or steady state with respect to the supply of electron donors, competition for the available donors controls the predominance of community members through the relative free-energy relationships. For example, in a natural environment where ferric iron and sulfate were degrading organic matter, the iron-reducing bacteria would gain more net energy ($-116\,kJ\,mol^{-1}$) than would sulfate-reducing bacteria ($-105\,kJ\,mol^{-1}$), and the predominant geochemical process would be iron reduction (Chapelle and Lovley, 1992). If ferric iron were not present, sulfate reduction would predominate. The diverse microbial community may quickly transition to predominance by members using alternative TEAPs as geochemical conditions change.

In most groundwaters, the supply of organic compounds and other nutrients is severely limited, and these environments are termed *oligotrophic*. Spatial zonation in community composition may develop as organisms characterized by a particular TEAP locally outcompete other organisms. The study of zonation in natural aquifers may be difficult and require extensive analysis of sediments and groundwaters over extensive geographical areas (Chapelle *et al.*, 1987; Chapelle and Knobel, 1983), although a systematic approach has been developed and demonstrated (Chapelle *et al.*, 1995). Zonation is more commonly investigated in disturbed (contaminated) environments, where the introduction of nutrients causes variation over relatively short distances (Bjerg *et al.*, 1995; Chapelle *et al.*, 1996). Even in contaminant plumes, there is often general agreement with the expected TEAP succession (Baedeker *et al.*, 1993; Chapelle *et al.*, 1996; Vroblesky and Chapelle, 1994).

Table 1 Redox processes in closed systems (modified after Champ *et al.*, 1979)

Reaction	Equation	$\Delta G^\circ(W)$, kJ mol^{-1} e^{-1}
Aerobic respiration	$CH_2O + O_2 = CO_2 + H_2O$	−502
Denitrification	$5CH_2O + 4NO_3^- + 4H^+ = 2N_2 + 5CO_2 + 7H_2O$	−476
Nitrate reduction	$2CH_2O + NO_3^- + 2H^+ = NH_4^+ + 2CO_2 + 3H_2O$	−343
Manganese reduction	$CH_2O + 2MnO_2 + 4H^+ = 2Mn^{2+} + CO_2 + 3H_2O$	−340
Iron reduction	$CH_2O + 8H^+ + 4Fe(OH)_3 = 4Fe^{2+} + 11H_2O + CO_2$	−116
Fermentation	$3CH_2O + H_2O = 2CH_3OH + CO_2$	−113
Sulfate reduction	$2CH_2O + SO_4^{2-} + H^+ + 2H_2O = HS^- + 2CO_2 + 4H_2O$	−105
CO_2 reduction	$4H_2 + CO_2 = CH_4 + 2H_2O$	−104
Methane fermentation	$2CH_2O = CH_4 + CO_2$	−93

An interesting application of groundwater investigations of microbial communities uses the relationship of fermenting bacteria (which use no external electron acceptor) at the base of the microbial food chain, to the predominant TEAP, controlled by free energy and the availability of the relevant electron acceptor. Because the supply of electron donors is controlled by fermentation, the predominant organism will be the one degrading the electron donor most efficiently. Along with simple organic substrates, fermentation produces H_2, which can be used by many organisms also as an electron donor. Those organisms that can gain the most energy from H_2 oxidation – in the presence of their favored electron acceptor – will predominate by driving the equilibrium H_2 concentration to the lowest utilizable level, and to a level that is not easily utilized by less efficient competitors. The dissolved concentration of H_2 may therefore be used as an indicator of the dominant environmental TEAP (Chapelle *et al.*, 1996; Lovley and Goodwin, 1988), or may be used in conjunction with measurements of other substrates (Chapelle *et al.*, 1996).

Although the succession of TEAPs may be hypothetically and conceptually straightforward, exceptions to the free-energy ordering of TEAPs have been observed. In contaminated environments, in particular, the free-energy succession may not be followed, or two TEAPs may significantly occur simultaneously. The variations in observed TEAPs and in the TEAP succession may be explained by partial thermodynamic equilibrium (Postma and Jakobsen, 1996). According to this approach to evaluating subsurface processes, the fermentation of sedimentary organic matter is considered to be rate limiting, and the TEAPs are all considered to be close to thermodynamic equilibrium. In evaluating which TEAPs are operating in a given environment, all geochemical variables are accounted for in calculating permissible organic matter oxidation pathways, and multiple processes are simultaneously allowed if they are thermodynamically permissible (have a negative free energy). By explicitly including thermodynamic evaluations, the partial equilibrium model thus explains systems in which sulfate reduction and iron reduction occur simultaneously, and where the theoretical succession of TEAPS is not observed. This approach was used to explain simultaneous sulfate and iron reduction (i.e. the simultaneous appearance in the sediment column of Fe(II) and S^-) in lake-bottom pore waters (Postma and Jakobsen, 1996).

ENVIRONMENTAL INVESTIGATIONS OF GROUNDWATER COMMUNITIES

The investigation of microbial communities has often focused on community function, revealed by chemical transformations in groundwater or its enclosing sediment, along with microbial investigations through culturing or molecular methods. As noted previously, the direct investigation of microbiological communities is complicated by the biases inherent in the investigation methods: traditional culturing may not describe all members of the community, and molecular methods do not determine the relative significance of the community's members. Following are examples of some biogeochemical investigations.

REGIONAL AQUIFERS

A number of research efforts have focused on the evolution of aqueous geochemistry along a groundwater flow path (Chapelle *et al.*, 1987; Chapelle *et al.*, 1988; McMahon and Chapelle, 1991; Murphy *et al.*, 1992). Those studies confirmed the generality that the microbial community was dependent on geochemical variables and that the geochemical environment was simultaneously modified by bacteria contained within it. Redox zonation, the segregation of different TEAPs during the degradation of sedimentary organic matter, was shown along groundwater flowpaths or between distinct sedimentary environments.

An analysis of the Middendorf aquifer in South Carolina (Murphy and Schramke, 1998; Murphy *et al.*, 1992) using chemical analysis of groundwater, sediments, and microbial culturing methods showed the regional effects of microbial metabolism. The combination of regional and detailed local chemistry and microbiology was required to understand the microbial succession and its effects on aquifer geochemistry. The geochemistry was shown to be controlled by a complex interplay of chemical and microbial factors, as illustrated and summarized in Figure 1 (Murphy and Schramke, 1998). High-lignite zones were noted, along with organic carbon-stained sands, and inorganic carbon increased along the flowpath because of the oxidation of organic carbon. From the point of recharge, aerobes were most numerous, overall, in zones without lignite, and *vice versa*, suggesting that lignite oxidation was coupled to oxygen depletion. Along the flowpath, oxygen was eventually depleted and organic acids were produced by fermenters, and the sulfate-reducing and iron-reducing bacteria were in turn supported by these fermentation-produced electron donors. A modified system of redox zonation was observed: the aquifer underwent transition from aerobic waters to anaerobic waters containing Fe^{2+} and Mn^{2+}, and then to waters with increasing SO_4^{2-} and decreasing Mn^{2+} and Fe^{2+}. Zonation was not completely predicted by the principle of TEAP succession or by the nature of iron-reducing microbes, however. Ferrous iron (Fe^{2+}) increased along the flowpath in an oxygenated portion of the aquifer (Figure 1); this should not occur because iron reduction is an obligately anaerobic process, where oxygen should reach undetectable levels before Fe^{2+} is observed. Sediment analysis showed that the occurrence of Fe^{2+} was

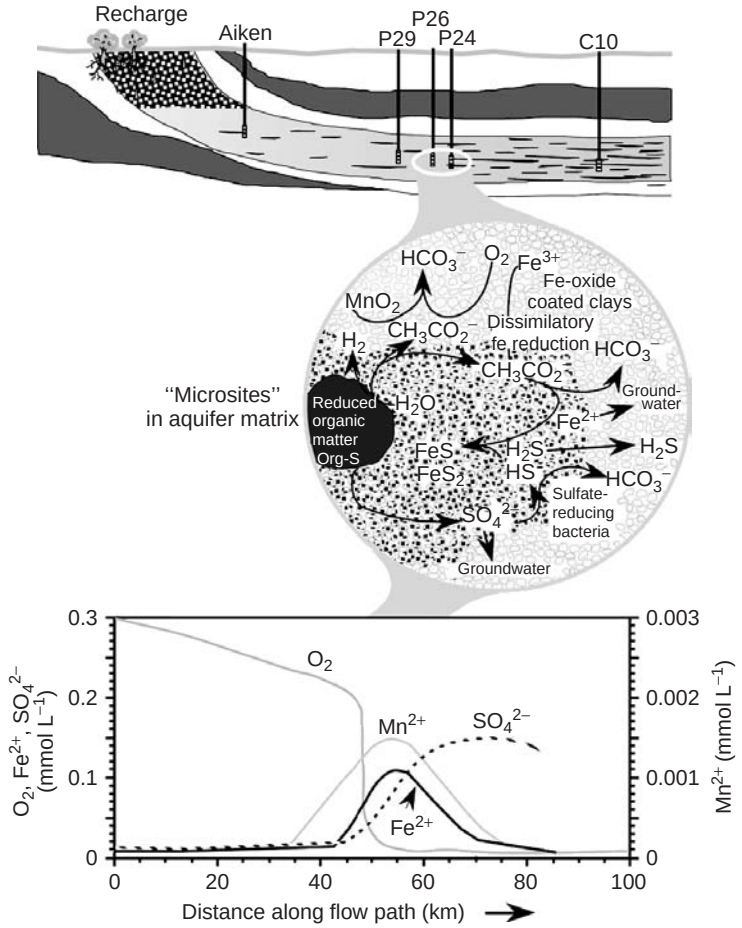


Figure 1 Groundwater evolution and the significance of microbial microsites in the Middendorf aquifer, South Carolina (Murphy *et al.*, 1992). Redox zonation occurred along the aquifer flowpath, and microbial microsites were inferred by the occurrence together of dissolved oxygen and reduced metal species (By courtesy of E. Murphy). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

coupled to the occurrence of high-lignite zones in the aerobic portion of the aquifer. Careful observation, modeling, and deduction indicated the existence, as illustrated in the figure, of a complex fine-scale relationship between anaerobic microsites, where dissimilatory iron-reducing bacteria (DIRB) could use fermentation-produced organic acids and contribute ferrous iron as a solute observable in oxygenated groundwaters. The macroscopic transitions between TEAPs were thus obscured by processes acting at fine scale over a broad area to produce an overlap of the products of microbial metabolism.

Microbial microsites, physically separate, small-scale, chemically distinct sedimentary domains harboring distinct microbial communities and producing anomalous chemical signatures in groundwater, may be important components of an overall microbial community. For example, in a natural system containing the fermentation products H_2, propionate, and butyrate, the concentrations of H_2 were calculated, on thermodynamic grounds, to be too high to permit the syntrophic degradation of propionate and butyrate (Conrad *et al.*, 1986). These compounds were known to be degraded. The conclusion was drawn that an environmental niche – resulting from fine-scale heterogeneity – must have existed where the processes were favored. Specifically, the anaerobic oxidation of butyrate and propionate must have taken place where zones of active methanogenesis (drawing down H_2 concentration) drove the pertinent oxidation reactions through mass action on the overall system. The hyporheic zone of the Columbia River, Washington, a microbially active subsurface interface between groundwater and surface water, was similarly deduced to harbor anaerobic microsites (Moser *et al.*, 2003). The sediment at the river bottom, overlying an aerobic aquifer, was sampled by freezing and intact removal. Relatively high numbers of heterotrophic aerobes were present, along with significant numbers of anaerobic nitrate, sulfate, and iron-reducing microbes. The presence of acid-volatile sulfide indicated that the sulfate-reducing microbes, at least, were

metabolically active. The presence of anaerobic microsites in the otherwise aerobic sedimentary environment was required for the production of sulfide.

In each of the cases cited, the use of the term *microsite* was used without reference to an absolute scale, and it is often used to describe heterogeneity below the scale of observation. In the studies of the Middendorf aquifer (Murphy and Schramke, 1998; Murphy *et al.*, 1992), for example, inferences were made from the information that was available, including compositions derived from groundwater sampling wells that tapped significant vertical intervals in the aquifer and from smaller scale core studies. Under favorable circumstances at other locations, the scales of microbial heterogeneity have been related to aquifer properties measured at the same scale; for example, at the centimeter to decimeter level (Bekins *et al.*, 2001; Smith *et al.*, 1991). As sampling technology progresses, aquifer measurements at or below the scale of community heterogeneity may become more acheivable. With access to information at the scale of heterogeneity distinctions could be made between mixed redox conditions due to permissible free energies for multiple reaction, redox conditions varying over centimeter to decimeter scale controlled by aquifer properties such as sediment texture or compositions, and μm-scale regions or sites controlled by chemical diffusion.

SEDIMENTARY TRANSITIONS

Sedimentary interfaces represent an abrupt transition in depositional environments. Because natural communities are in equilibrium with their environments, it follows logically that they are limited by the supply of some nutrient or by some physical condition. Where the texture and composition of adjacent lithologies contrasts sharply, particularly where a carbon-rich lithology is juxtaposed with a carbon-poor lithology, the microbial environments may favor communities with markedly different structures, and relatively fine-scale investigations of these systems can reveal the interdependence of geochemistry and microbial function. A microbiological study of aquifer and aquitard sediments in the Atlantic Coastal Plain (McMahon and Chapelle, 1991), for example, indicated that fermentation reactions in aquitard sediments produced simple organic compounds and H_2. In the aquitard, the production of organic acids outpaced respiration, and in the aquifer, respiration outpaced fermentation. The diffusive transport of organic acids from the aquitard was calculated to be sufficient to support the observed rates of microbial respiration within the aquifer. Similar studies of a deeply buried, anaerobic, organic-rich paleosol in an otherwise aerobic aquifer indicated that the environment within the paleosol was electron-acceptor limited, and that the diffusion of electron acceptors into the anaerobic stratum drove the metabolism of fermentation-derived organic acids (Fredrickson *et al.*, 1995; Kieft *et al.*, 1995; McKinley *et al.*, 1997).

Deeper, more indurated, and older aquifer systems show the same interformational interaction. A detailed study of a deep aquifer consisting of a transgressive-regressive sequence of sandstones and shales in New Mexico showed that the aquifer was anaerobic, and was fed by regional infiltration of meteoric water from areas at higher elevation (Walvoord *et al.*, 1999). The sediments were sampled by argon-rotary core drilling, and found to include shales that were relatively rich in organic carbon, interbedded with organic carbon-poor sandstones (Figure 2). The microbial environment across the lithologic boundary changed dramatically at the centimeter scale. The shale pore sizes in particular limited the viability of microbes within them (Fredrickson *et al.*, 1997). Chemical measurements of the two lithologies included aqueous compositions and dissolved gases, using a passive multilevel sampler (McKinley, 2001; Ronen *et al.*, 1987), and of sulfate reduction directly within the cores using a silver foil technique (Krumholz *et al.*, 1997). Sulfate reducers, detected by the production of sulfide directly from core samples (Figure 2), were most active at the shale-sandstone interfaces, where fermentation-produced organic carbon from the organic-rich shales diffused into the more transmissive, sulfate-rich sandstone aquifers. The concentration of methanogen methane was lowest in the transmissive sandstones, and highest in the less-transmissive shales where sulfate reduction was apparently limited by the diffusive transport of sulfate. A later study of the same sequence using sediments incubated within a passive multilevel sampler showed that the abundances major archael guilds shifted dramatically over centimeter distances (Takai *et al.*, 2003).

DISTURBED ENVIRONMENTS

Much interest in microbial communities is related to the potential for those communities to remediate contamination, either through natural processes or after stimulation by the addition of the necessary nutrients. *Bioremediation* may degrade organic contaminants and lower the valence of reducible metals. The remediation of contaminated sites by *in situ* organisms without amendment has been termed *intrinsic bioremediation* and has been of particular interest in sites contaminated with organic compounds (Madsen, 2001). There are a number of well-studied sites contaminated with organic compounds, including the landfill at Vejen, Denmark, and a site at Cape Cod, Massachusetts.

At the Vejen site, the distribution of inorganic and organic compounds was determined and a series of investigations tracked the movement and degradation of the contaminant plume associated with the landfill (Baun *et al.*, 2003). An analysis of microbial community structure before and after

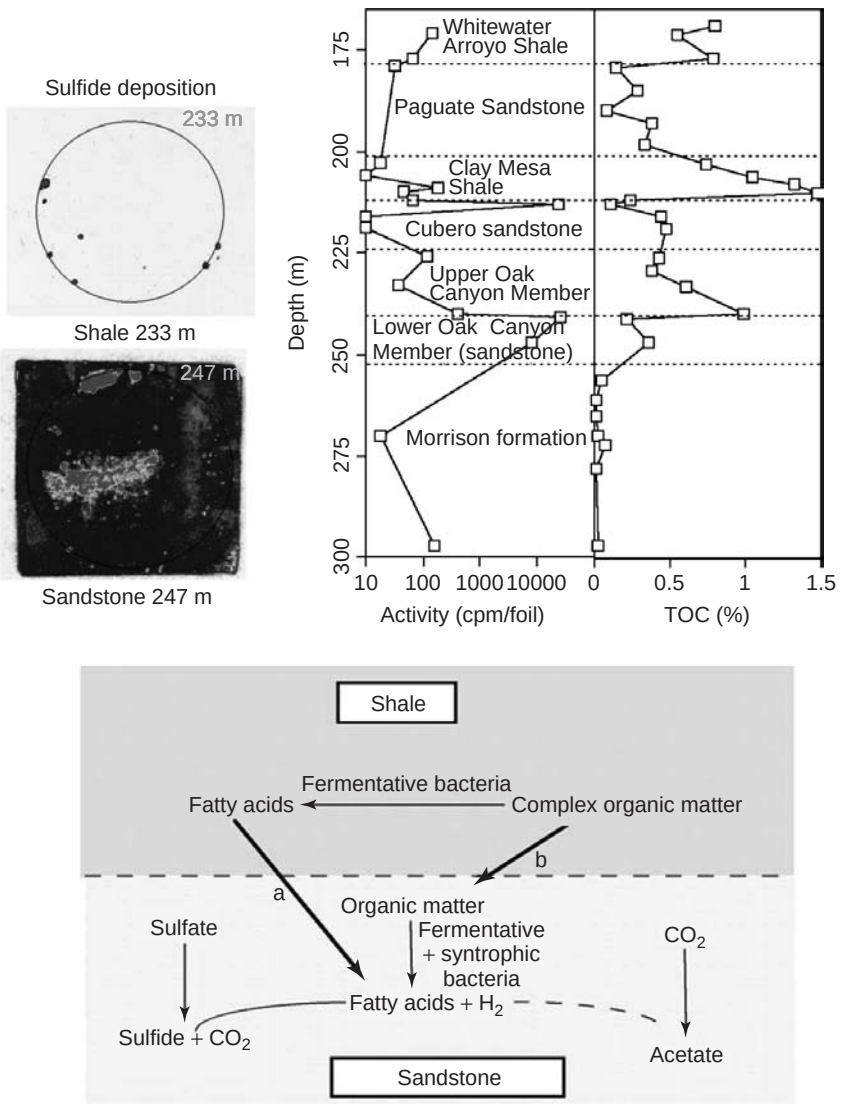


Figure 2 Sulfate-reducing activity as part of a complex interaction across lithologic boundaries in a deep sandstone-shale aquifer near Seboyeta in New Mexico (Krumholz *et al.*, 1997). Silver foils (left), inserted into broken sediment cores, were used to map sulfate reduction. The transgressive-regressive sequences of sandstones and shales harbored interdependent communities (By courtesy of L. Krumholz). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

prolonged exposure to contaminant herbicides indicated that the overall structure of the community was altered, but the community diversity was not (de Lipthay *et al.*, 2004). Related review articles of the biogeochemistry of landfills (Christensen *et al.*, 2001) and of the characterization of redox conditions within landfill plumes (Christensen *et al.*, 2000) contain useful additional information. The long-term and detailed investigation of the Cape Cod site (USGS, 2005) evaluated the fate of organic and inorganic contaminants. The site included a large plume of contaminated groundwater due to secondary treatment and disposal of domestic sewage to a shallow unconfined aquifer. Biodegradation of labile organic compounds was observed, but many trace organics such as trichloroethene persisted for more than 30 years (Barber *et al.*, 1988).

The bioreduction of metals and other inorganic compounds may also be important as a bioremediation approach. The metal-reducing microbes have been extensively studied (Lovley, 1991; Lovley, 1993), and have been shown to reductively immobilize contaminant species such as pertechnetate and uranium. Because uranium is a widespread contaminant at the former weapons production facilities in the United States (Riley and Zachara, 1992), its bioreduction to insoluble and immobile U(IV) has been extensively investigated. The microbial reduction of U(VI) to U(IV) has been demonstrated at mill tailings sites and in

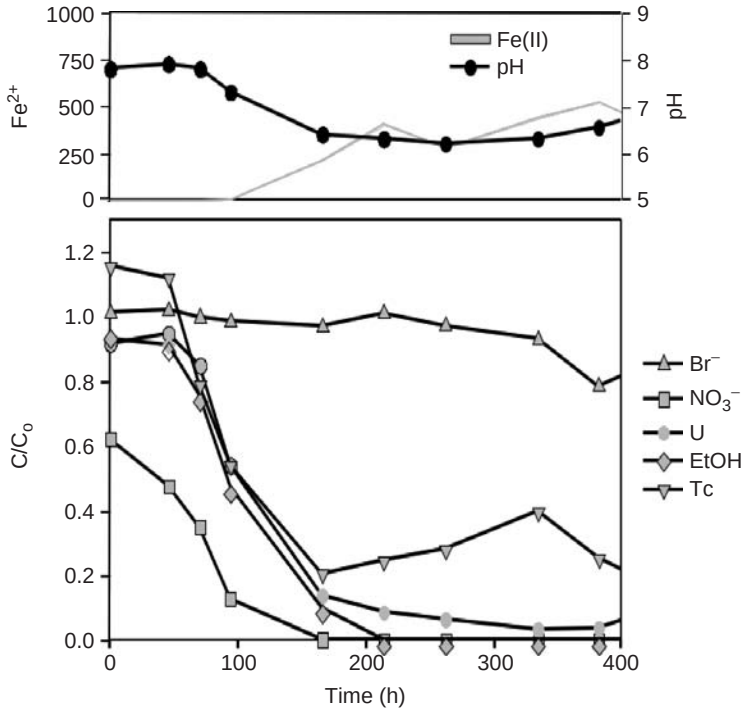


Figure 3 Results of a push-pull test at Oak Ridge, Tennessee, demonstrating the ethanol-stimulated reduction of nitrate, technetium, uranium, and ferric iron. The injected solution was withdrawn over a long period; the results normalized to initial concentrations indicate the stimulation of *in situ* microbial activity (Reprinted from Istok 2005. © 2005 with permission of American Chemical Society)

related laboratory studies (Anderson *et al.*, 2003; Finneran *et al.*, 2002; Senko *et al.*, 2002), particularly where the zone of contamination was stimulated by the addition of an electron donor (Anderson *et al.*, 2003). Uranyl (UO_2^{2+}), the predominant soluble U(VI) ion, was found to be reduced by several metal-reducing microbes, but the presence of nitrate (NO_3^-) and the production of nitrite (NO_2^-) during nitrate reduction was observed to cause U(IV) reoxidation (Senko *et al.*, 2002).

Uranyl reduction can be rapidly stimulated *in situ*, as demonstrated in numerous field studies, including the following example using a hydrologic method. A novel method that can stimulate and determine the direct bioalteration of contaminants has been extensively demonstrated by researchers at Oregon State University (Istok, 2005). The method, a *push-pull test*, uses a single monitoring well, into which is injected an aqueous tracer along with an electron donor and other amendments as necessary. After injection of a significant test volume, ground water is withdrawn and sampled over time and the tracer concentration is used to correct for mixing and to monitor the relative abundance of other dissolved components, including the electron donor. An application of the method is illustrated in Figure 3, for an aquifer contaminated with uranium and technetium at the US Oak Ridge site, Tennessee. Details of the series of injections and their interpretation are presented elsewhere (Istok *et al.*, 2004). The concentration of the electron donor (ethanol) and of nitrate, technetium, uranium, ferrous iron, nitrite, and pH were monitored over the time of water withdrawal (up to 800 h), along with that of the bromide tracer (Figure 3). Nitrate was rapidly reduced as ethanol was oxidized (note their positions below the mixing-normalized tracer concentration), and technetium and uranium were rapidly reduced and removed from solution, while ferrous iron was introduced to solution through sedimentary iron reduction. Nitrate thermodynamically inhibited uranium reduction, indicated by the substantial removal of nitrate prior to the decrease in uranium. The experiments were bounded by injections without donor, and, along with companion injections (not shown), demonstrate directly the biostimulation of an *in situ* microbial community to reductively immobilize technetium and uranium.

Acknowledgments

This work was supported by the Natural and Accelerated Bioremediation Research Program (NABIR), Office of Biological and Environmental Research (OBER), of the US Department of Energy (DOE). Pacific Northwest National Laboratory is supported by the Battelle Memorial Institute for the DOE. The authors are grateful to an anonymous reviewer who provided valuable and helpful comments and suggestions.

Appendix

Microbial organisms act as redox catalysts, mediating the electron transfer between an electron-donating (oxidation) half-reaction and a coupled electron-accepting (reduction) half-reaction. It is convenient to combine the ideas of $p\varepsilon$ and half-reactions into a simple thermodynamic method for evaluating the feasibility of hypothetical reactions for driving microbial metabolism (Stumm and Morgan, 1996).

The quantity $p\varepsilon$ can be defined in a way that is analogous to pH,

$$p\varepsilon = -\log\{e^-\} \tag{2}$$

giving the hypothetical electron activity, and indicating the tendency of a solution to accept or transfer electrons. In a highly reducing solution the electron activity is relatively high, and in an oxidizing solution the electron activity is relatively low. The $p\varepsilon$ can be related to the standard electrode potential and free energy of reaction by the relations

$$E_H^\circ = -\Delta G^\circ/nF = (RT/nF)\ln\ K = (2.3\ RT/nF)$$
$$\log\ K = (2.3\ RT/F)\,p\varepsilon^\circ \tag{3}$$

where ΔG° is the standard Gibbs free-energy change in the reduction reaction, K is the thermodynamic equilibrium constant, n is the number of electrons in the reaction as written, R is the gas constant, and F is the Faraday constant. The energetics of half-reactions may be characterized by the value of K for the reaction, by the standard potential (E_H°), or by the $p\varepsilon^\circ$. By summing appropriate half-reactions, the net potential for any given oxidation-reduction reaction can be readily derived.

A convenient modification of $p\varepsilon$ as a tool for understanding energy-yielding reactions in natural waters is to define a subsidiary constant, $p\varepsilon^\circ(W)$. This quantity is analogous to $p\varepsilon^\circ$, except that $\{H^+\}$ is assigned its activity in neutral water to unambiguously negate the effects of variable pH on comparisons based on $p\varepsilon^\circ$. Standard conditions for $p\varepsilon^\circ(W)$ are thus defined at 25 °C, unit activities of oxidant and reductant, and pH = 7.00. $p\varepsilon^\circ(W)$ is defined by

$$p\varepsilon^\circ(W) = p\varepsilon^\circ + \log[H^+]^n \tag{4}$$

where n is the number of moles of protons exchanged per mole of electrons, and $[H^+]$ equals 1×10^{-7} mol L^{-1}. Using this convention, half-reactions can be evaluated according to $p\varepsilon^\circ(W)$, so that a ranked list will contain reactions ordered in such a way that any higher (more positive) system will oxidize any lower (more negative) system. For example, given a table of half-reactions with values for $p\varepsilon^o(W)$, one can readily establish that NO_3^- can oxidize HS^- to SO_4^{2-}. Details of this approach along with tables and examples may be found elsewhere (Champ *et al.*, 1979; Stumm and Morgan, 1996). The construction of ranked lists of half-reactions has become widespread, and has proved useful in understanding redox reactions that support life, the competition and interplay between microbial metabolic processes, and in exploring the possibilities of life elsewhere (e.g. Gaidos *et al.*, 1999; Klass, 1984; Lovley and Goodwin, 1988; Madigan *et al.*, 1997; Stumm and Morgan, 1996).

REFERENCES

Anderson R.T., Vrionis H.A., Ortiz-Bernad I., Resch C.T., Long P.E., Dayvault R., Karp K., Marutzky S., Metzler D.R., Peacock A.D., *et al.* (2003) Stimulating the in situ activity of Geobacter species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Applied and Environmental Microbiology*, **69**, 5884–5891.

Andrews J.H. and Hall R.F. (1986) *r*- and *K* selection and microbial ecology. *Advances in Microbial Ecology*, **9**, 99–147.

Atlas R.M. and Bartha R. (1993) *Microbial Ecology Fundamentals and Applications*, Benjamin/Cummings Publishing.

Baedeker M.J., Cozzarelli I.M., Eganhouse R.P., Siegel D.I. and Bennet P.C. (1993) Crude oil in a shallow sand and gravel aquifer – III. Biogeochemical reactions and mass balance modeling in anoxic groundwater. *Applied Geochemistry*, **8**, 569–586.

Barber L.B.I., Thurman E.M. and Schroeder M.P. (1988) Long-term fate of organic micropollutants in sewage-contaminated groundwater. *Environmental Science and Technology*, **22**, 205–211.

Baun A., Reitzel L.A., Ledin A., Christensen T.H. and Bjerg P. (2003) Natural attenuation of xenobiotic organic compounds in a landfill leachate plume (Vejen, Denmark). *Journal of Contaminant Hydrology*, **65**, 269–291.

Bekins B.A., Cozzarelli I.M., Godsy E.M., Warren E., Essaid H.I. and Tuccillo M.E. (2001) Progression of natural attenuation processes at a crude oil spill site: II. Controls on spatial distribution of microbial populations. *Journal of Contaminant Hydrology*, **53**, 387–406.

Berner R.A. (1981) A new geochemical classification of sedimentary environments. *Journal of Sedimentary Petrology*, **51**, 359–365.

Bjerg P.L., Rugge K., Pedersen J.K. and Christensen T.H. (1995) Distribution of redox-sensitive groundwater quality parameters downgradient of a landfill (Grindsted, Denmark). *Environmental Science and Technology*, **29**, 1387–1394.

Champ D.R., Gulens J. and Jackson R.E. (1979) Oxidation-reduction sequences in ground water flow systems. *Canadian Journal of Earth Sciences*, **16**, 12–23.

Chapelle F.H. (1993) *Ground-Water Microbiology and Geochemistry*, John Wiley & Sons.

Chapelle F.H., Haack S.K., Adriaens P., Henry M.A. and Bradley P.M. (1996) Comparison of Eh and H_2 measurements for delineating redox processes in a contaminated aquifer. *Environmental Science and Technology*, **30**, 3565–3569.

Chapelle F.H., Joseph L., Zelibor J., Grimes D.J. and Knobel L.L. (1987) Bacteria in deep coastal plain sediments of Maryland: a possible source of CO_2 to groundwater. *Water Resources Research*, **23**(8), 1625–1632.

Chapelle F.H. and Knobel L.L. (1983) Aqueous geochemistry and the exchangeable cation composition of glauconite in the Aquia Aquifer, Maryland. *Ground Water*, **21**(3), 343–352.

Chapelle F.H. and Lovley D.R. (1992) Competitive exclusion of sulfate reduction by Fe(III)-reducing bacteria: a mechanism for producing discrete zones of high-iron ground water. *Ground Water*, **30**(1), 29–36.

Chapelle F.H., McMahon P.B., Dubrovsky N.M., Fujii R.F., Oaksford E.T. and Vroblesky D.A. (1995) Deducing the distribution of terminal electron-accepting processes in hydrologically diverse groundwater systems. *Water Resources Research*, **31**(2), 359–371.

Chapelle F.H., Morris J.T., McMahon P.B. and Zelibor J.L. (1988) Bacterial metabolism and the $\delta^{13}C$ composition of ground water, Floridan aquifer system, South Carolina. *Geology*, **16**, 117–121.

Christensen T.H., Bjerg P., Banwart S.A., Jakobsen R., Heron G. and Albrechtsen H.-J. (2000) Characterization of redox conditions in groundwater contaminant plumes. *Journal of Contaminant Hydrology*, **45**, 165–241.

Christensen T.H., Kjeldsen P., Bjerg P., Jensen D.L., Christensen J.B., Baun A., Albrechtsen H.-J. and Heron G. (2001) Biogeochemistry of landfill leachate plumes. *Applied Geochemistry*, **16**, 659–718.

Colwell F.S. (2001) Constraints on the distribution of microorganisms in subsurface environments. In *Subsurface Microbiology and Biogeochemistry*, Fredrickson J.K. and Fletcher M. (Eds.), Wiley-Liss: pp. 71–96.

Conrad R., Schink B. and Phelps T.J. (1986) Thermodynamics of H_2-consuming and H_2-producing metabolic reactions in divers methanogenic environments under *in situ* conditions. *FEMS Microbiology Ecology*, **38**, 353–360.

de Lipthay J.R., Johnsen K., Albrechtsen H.-J., Rosenberg P. and Aamand J. (2004) Bacterial diversity and community structure of a sub-surface aquifer exposed to realistic low herbicide concentrations. *FEMS Microbiology Ecology*, **49**, 59–69.

Ehrlich H.L. (1990) *Geomicrobiology*, Marcel Dekker.

Embley T.M. and Stackebrandt E. (1996) The use of 16S ribosomal RNA sequences in microbial ecology. In *Molecular Approaches to Environmental Microbiology*, Pickup R.W. and Saunders J.R. (Eds.), Ellis Harwood: pp. 39–62.

Finneran K.T., Anderson R.T., Nevin K.P. and Lovley D.R. (2002) Potential for bioremediation of uranium-contaminated aquifers with microbial U(VI) reduction. *Soil and Sediment Contamination*, **11**, 339–357.

Fredrickson J.K., McKinley J.P., Bjornstad B.N., Long P.E., Ringelberg D.B., White D.C., Krumholz L.R., Suflita J.M., Colwell F.S., Lehman R.M., *et al.* (1997) Pore-size constraints on the activity and survival of subsurface bacteria in a Late Cretaceous shale-sandstone sequence, northwestern New Mexico. *Geomicrobiology Journal*, **14**, 183–202.

Fredrickson J.K., McKinley J.P., Nierzwicki-Bauer S.A., White D.C., Ringelberg D.B., Rawson S.A., Li S.-M., Brockman F.J. and Bjornstad B.N. (1995) Microbial community structure and biogeochemistry of Miocene subsurface sediments: implications for long-term microbial survival. *Molecular Ecology*, **4**, 619–626.

Gaidos E.J., Nealson K.H. and Kirschvink J.L. (1999) Life in ice-covered oceans. *Science*, **284**(5420), 1631–1633.

Grant W.D. and Tindall B.J. (1986) The alkaline saline environment. In *Microbes in Extreme Environments*, Herbert R.A. and Codd G.A. (Eds.), Harcourt Brace Jovanovich.

Horikoshi K. and Tsujii K. (1999) *Extremophiles in Deep-Sea Environments*, Springer: p. 316.

Istok J.D. (2005) Groundwater Research Laboratory. **`https://web.engr.oregonstate.edu/~istokj/grl-main.htm`**.

Istok J.D., Senko J.M., Krumholz L.R., Watson D., Bogle M.A., Peacock A.D., Chang Y.-J. and White D.C. (2004) In situr bioremediation of technetium and uranium in a nitrate-contaminated aquifer. *Environmental Science and Technology*, **38**, 468–475.

Kieft T.L., Fredrickson J.K., McKinley J.P., Bjornstad B.N., Rawson S.A., Phelps T.J., Brockman F.J. and Pfiffner S.M. (1995) Microbiological comparisons within and across contiguous lacustrine, paleosol, and fluvial subsurface sediments. *Applied and Environmental Microbiology*, **61**(2), 749–757.

Kinner N.E., Harvey R.W., Shay D.M., Metge D.W. and Warren A. (2002) Field evidence for a protistan role in an organically-contaminated aquifer. *Environmental Science and Technology*, **36**, 4312–4318.

Klass D.L. (1984) Methane from anaerobic fermentation. *Science*, **223**, 1021–1028.

Krumholz L.R., McKinley J.P., Ulrich G.A. and Suflita J.M. (1997) Confined subsurface microbial communities in Cretaceous rock. *Nature*, **386**, 64–66.

Lovley D.R. (1991) Dissimilatory Fe(III) and Mn(IV) reduction. *Microbiological Reviews*, **55**(2), 259–287.

Lovley D.R. (1993) Dissimilatory metal reduction. *Annual Reviews in Microbiology*, **47**, 263–290.

Lovley D.R. and Goodwin S. (1988) Hydrogen concentrations as an indicator of the predominant terminal electron-accepting reactions in aquatic sediments. *Geochimica et Cosmochimica Acta*, **52**, 2993–3003.

Madigan M.T., Martinko J.M. and Parker J. (1997) *Brock Biology of Microorganisms*, Prentice-Hall.

Madsen E.L. (2001) Intrinsic bioremediation of organic subsurface contaminants. In *Subsurface Microbiology and Biogeochemistry*, Fredrickson J.K. and Fletcher M.K. (Eds.), Wiley-Liss: pp. 249–278.

Madsen E.L. and Ghiorse W.C. (1993) Groundwater microbiology: subsurface ecosystem processes. In *Aquatic Microbiology: An Ecological Approach*, Ford T.E. (Ed.), Blackwell: pp. 167–213.

McKinley J.P. (2001) The use of geochemistry and the importance of sample scale in investigations of lithologically heterogeneous microbial ecosystems. In *Subsurface Microbiology and Biogeochemistry*, Fredrickson J.K. and Fletcher M.K. (Eds.), Wiley-Liss: pp. 173–192.

McKinley J.P., Stevens T.O., Fredrickson J.K., Zachara J.M., Colwell F.S., Wagnon K.B., Smith S.C., Rawson S.A. and Bjornstad B.N. (1997) Biogeochemistry of anaerobic lacustrine

and paleosol sediments within an aerobic unconfined aquifer. *Geomicrobiology Journal*, **14**, 23–29.

McMahon P.B. and Chapelle F.H. (1991) Microbial production of organic acids in aquitard sediments and its role in aquifer geochemistry. *Nature*, **349**, 233–235.

Moser D.P., Fredrickson J.K., Geist D., Arntzen E.V., Peacock A.D., Li S.M., Spadoni C.M. and McKinley J.P. (2003) Biogeochemical processes and microbial characteristics across groundwater-surface water boundaries of the Hanford Reach of the Columbia River. *Environmental Science and Technology*, **37**, 5127–5134.

Murphy E.M. and Schramke J.A. (1998) Estimation of microbial respiration rates in groundwater by geochemical modeling constrained with stable isotopes. *Geochimica et Cosmochimica Acta*, **62**(21/22), 3395–3406.

Murphy E.M., Schramke J.A., Fredrickson J.K., Bledsoe H.W., Francis A.J., Sklarew D.S. and Linehan J.C. (1992) The influence of microbial activity and sedimentary organic carbon on the isotopic geochemistry of the Middendorf aquifer. *Water Resources Research*, **28**, 723–740.

Postma D. and Jakobsen R. (1996) Redox zonation: equilibrium constraints on the Fe(III)/SO_4-reduction interface. *Geochimica et Cosmochimica Acta*, **60**, 3169–3175.

Reysenbach A.L., Gotz D. and Yernool D. (2002) Microbial diversity of marine and terrestrial thermal springs. In *Biodiversity of Microbial Life*, Staley J.T. and Reysenbach A.-L. (Eds.), Wiley-Liss.

Riley R.G. and Zachara J.M. (1992) *Chemical Contaminants on DOE Lands and Selection of Contaminant Mixtures for Subsurface Science Research*, DOE/ER-0547T, US Department of Energy: p. 77.

Ronen D., Magaritz M. and Levy I. (1987) An in situ multilevel sampler for preventive monitoring and study of hydrochemical profiles. *Ground Water Monitoring Review*, **7**, 69–74.

Senko J.M., Istok J.D., Suflita J. and Krumholz L.R. (2002) In-situ evidence for uranium immobilization and remobilization. *Environmental Science and Technology*, **36**, 1491–1496.

Sinclair J.L. and Ghiorse W.C. (1989) Distribution of aerobic bacteria, protozoa, algae, and fungi in deep subsurface sediments. *Geomicrobiol*, **7**, 15–31.

Smith R.L., Harvey R.W. and LeBlanc D.R. (1991) Importance of closely spaced vertical sampling in delineating chemical and microbiological gradients in groundwater studies. *Journal of Contaminant Hydrology*, **7**, 285–300.

Stetter K.O., Huber R., Blochl E., Kurr M., Eden R.D., Fielder M., Vance H. and Vance I. (1993) Hyperthermophilic Archaea are thriving in deep North Sea and Alaskan oil reservoirs. *Nature*, **365**, 743–745.

Stevens T.O. (2002) The Deep Subsurface Biosphere. In *Biodiversity of Microbial Life*, Staley J.T. and Reysenbach A.-L. (Eds.), Wiley-Liss: pp. 439–474.

Stevens T.O. and McKinley J.P. (1995) Lithoautotrophic microbial ecosystems in deep basalt aquifers. *Science*, **270**, 450–454.

Stevens T.O., McKinley J.P. and Fredrickson J.K. (1993) Bacteria associated with deep, alkaline, anaerobic groundwaters in southeast Washington. *Microbial Ecology*, **25**, 35–50.

Stumm W. and Morgan J.W. (1996) *Aquatic Chemistry*, John Wiley & Sons.

Takai K., Mormile M.R., McKinley J.P., Brockman F.J., Holben W.E., Kovacik W.P. Jr and Fredrickson J.K. (2003) Shifts in archaeal communities associated with lithological and geochemical variations in subsurface Cretaceous rock. *Environmental Microbiology*, **5**, 309–320.

Tyson G.W., Chapman J., Hugenholtz P., Allen E.E., Ram R.J., Richardson P.M., Solovyev V.V., Rubin E.M., Rokhsar D.S., and Banfield J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

USGS (2005) United States Geological Survey. *Cape Cod Bibliography*, **`http//toxics.usgs.gov/sites/cape cod page.html`**.

Vroblesky D.A. and Chapelle F.H. (1994) Temporal and spatial changes of terminal electron-accepting processes in a petroleum hydrocarbon-contaminate aquifer and the significance for contaminant bioremediation. *Water Resources Research*, **30**(5), 1561–1570.

Walvoord M.A., Pegram P., Phillips F.M., Person M., Kieft T.L., Fredrickson J.K., McKinley J.P. and Swenson J. (1999) Groundwater flow and geochemistry in the southeastern San Juan Basin: Implications for microbial transport and activity. *Water Resources Research*, **35**, 1409–1424.

Woese C.R. (2002) Perspective: microbiology in transition. In *Biodiversity of Microbial Life*, Staley J.T. and Reysenbach A.-L. (Eds.), Wiley-Liss.

Yayanos A.A. (1995) Microbiology to 10 500 meters in the deep sea. *Annual Review of Microbiology*, **49**, 777–805.

107: Natural and Constructed Wetlands

CHRISTOPHER B CRAFT

School of Public and Environmental Affairs, Indiana University, Bloomington, IN, US

Wetlands are characterized by periodic to continuous inundation or saturation with water (wetland hydrology), soils that are periodically deficient in oxygen (hydric soils), and vegetation that is adapted to periods of anaerobic or anoxic soil conditions (hydrophytic vegetation). Wetlands include herbaceous marshes and fens, forested swamps, and Sphagnum-dominated peat bogs. Wetland plants possess morphological adaptations (aerenchyma, buttressed trunks, pneumatophores, hypertrophied lenticels) to facilitate transport of oxygen to the roots and metabolic pathways to respire anaerobically without toxic effects. Duration, depth, and frequency of inundation, water source (precipitation, surface flow, groundwater), and water quality (fresh, saline, oligotrophic, eutrophic) determine the distribution, productivity, and nutrient cycling characteristics of wetland vegetation. Floodplain swamp forests and estuarine marshes and mangroves receive large quantities of water and nutrients during overbank flooding and tidal inundation that enhances plant productivity and sediment, nutrient, and pollutant retention as compared to wetlands that receive most of their water from precipitation. The ability of natural wetlands to filter pollutants has led to construction of artificial wetlands to treat wastewater, stormwater, acid mine drainage, nonpoint runoff and other sources of pollutants. Wetlands also are constructed to replace important ecological functions (water storage, biological productivity, biodiversity) that are lost when natural wetlands are damaged or destroyed by human activities.

INTRODUCTION

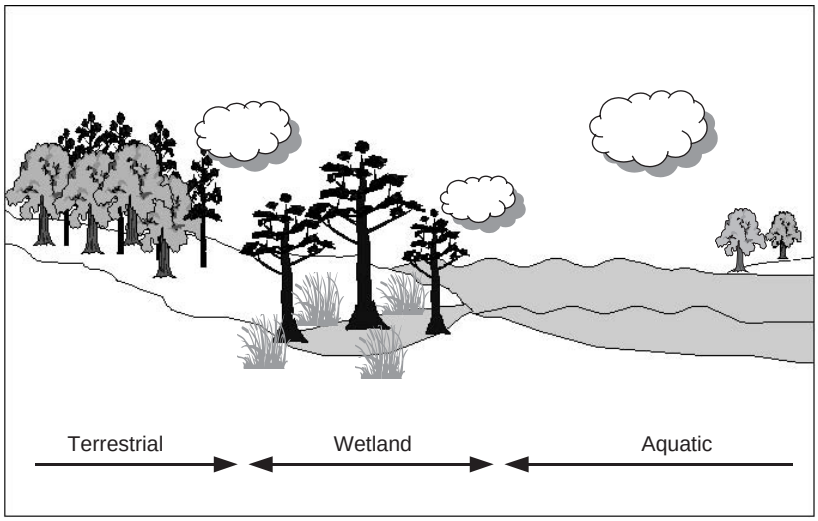
Wetlands are transitional between terrestrial and aquatic ecosystems and are periodically to continuously inundated or saturated with water, producing anaerobic soil conditions and favoring the growth of vegetation adapted to periods of low or no soil oxygen. Wetlands require inundation or soil saturation long enough during the growing season, usually 7 to 21 consecutive days, to produce anoxic (anaerobic) soil conditions. When soils become reduced, heterotrophic microorganisms use nitrate (NO_3^-), ferric iron (Fe^{3+}), sulfate (SO_4^{2-}) and other terminal electron acceptors to support growth and metabolism. In wetlands, extended periods of soil anoxia stresses and kills terrestrial plant species, favoring species adapted to low oxygen. By definition, a wetland must have the three following characteristics: periodic to continuous inundation or saturation with fresh or saline water (*wetland hydrology*), soils that are periodically anoxic (*hydric soils*), and rooted emergent vegetation adapted to periods of low or no oxygen (*hydrophytic vegetation*) (USACE, 1987). Thus, wetlands differ from terrestrial ecosystems by the presence of wetland hydrology, hydric soils, and hydrophytic vegetation. They differ from aquatic ecosystems by the presence of emergent vegetation that is rooted in soil (Figure 1).

Wetlands receive water of varying proportion from precipitation, groundwater, and surface water. The source of water as well as hydroperiod – the duration, depth, and frequency of inundation – interact to determine the plant species that will grow there. Hydrophytic vegetation ranges from herbaceous emergent plants such as the ubiquitous cattail (*Typha*) to woody trees such as bald cypress (*Taxodium distichum*). Hydrophytic species possess morphological and metabolic adaptations that enable them to survive and thrive in anaerobic soils.

CLASSIFICATION OF WETLANDS BASED ON VEGETATION

At the broadest level, wetlands are classified based on inundation with freshwater (freshwater wetlands) or partially diluted seawater (estuarine wetlands). Below this level,

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.



Hydrology:		
Dry	Intermittently to permanently inundated or saturated	Permanently inundated
Redox status:		
Oxidized	Intermittently to permanently reduced	Permanently reduced
Biogeochemistry:		
Source of materials	Source, sink, or transformer	Sink
Plant productivity:		
Low to medium	Generally high but sometimes low	Generally low

Figure 1 Hydrologic, biogeochemical, and biological functions of wetlands relative to terrestrial and aquatic ecosystems. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

wetlands often are separated on the basis of the dominant vegetation because it is a conspicuous feature of the landscape that reflects underlying environmental gradients. Freshwater wetlands include bogs, fens, marshes, and swamps. They are found in a variety of locations, often on regional topographic low spots near rivers, streams, and lakes and on higher parts of the landscape, on broad flats and depressions that do not readily drain. Estuarine wetlands occur in coastal regions and include salt, brackish, and tidal freshwater marshes and, tropical and subtropical regions, mangrove forests.

Freshwater Wetlands

Bogs and fens are the most common wetlands across large areas of the northern hemisphere, above 45° north (Matthews and Fung, 1987; Aselmann and Crutzen, 1989). Bogs are dominated by peat forming vegetation, notably peat mosses of the genus *Sphagnum* (Clymo, 1978; Crum, 1995). Over time, peat accumulation of several meters or more elevates the bog surface above that of the surrounding land such that precipitation is the major source of water and nutrients to these wetlands (Figure 2). Bogs are acidic, nutrient poor (Clymo, 1978) and dominated by acidophilic vegetation such as *Sphagnum*, ericaceous shrubs, and conifers like black spruce (*Picea mariana*) (Table 1, Crum, 1995). In Canada, black spruce-*Sphagnum*-dominated bogs are known as *muskegs*. Fens are often found in association with bogs but at lower elevations on the landscape where groundwater is a more important source of water (Figure 2). Because groundwater often contains greater concentrations of dissolved minerals such as calcium, the pH of fens is higher than in bogs, usually near circumneutral (Bridgham *et al.*, 1996). Fens are dominated by graminoid (grass-like) vegetation, grasses, sedges, and rushes (Table 1).

Like fens, marshes also are dominated by herbaceous emergent vegetation (Mitsch and Gosselink, 2000). However, marsh vegetation is found in almost any geomorphic

Table 1 Common freshwater and estuarine wetlands based on vegetation type. Primary water source, hydroperiod, and dominant vegetation also are presented

Wetland Type	Primary Water Source	Hydroperiod	Dominant Vegetation
Freshwater:			
Bog	precipitation	near-continuous saturation	peat mosses (*Sphagnum* spp.)
Fen	groundwater	near continuous inundation or saturation	grasses (Poaceae), sedges (Cyperaceae)
Swamp	surface water	periodic inundation	bald cypress (*Taxodium distichum*) and other trees
Marsh	precipitation, surface water, groundwater	periodic to continuous inundation or saturation	grasses (Poaceae), sedges (Cyperaceae), rushes (Juncaceae), reeds (*Phragmites, australis, Phalaris* sp.), cattail (*Typha* spp.)
Estuarine:			
Salt marsh	surface water	regular (2x daily) tidal inundation	cordgrass (*Spartina alterniflora*)
Brackish marsh	surface water	regular to irregular tidal inundation	needlerush (*Juncus roemerianus*), *Spartina* spp., salt grass (*Distichlis spicata*)
Tidal freshwater Marsh	surface water	regular tidal inundation	cattail (*Typha* spp.), *Panicum* spp., pickerelweed (*Pontedaria cordata*), *Sagittaria* spp.
Mangrove	tidal inundation	regular to irregular tidal inundation	trees (*Rhizophora* sp., *Avicennia* sp.)

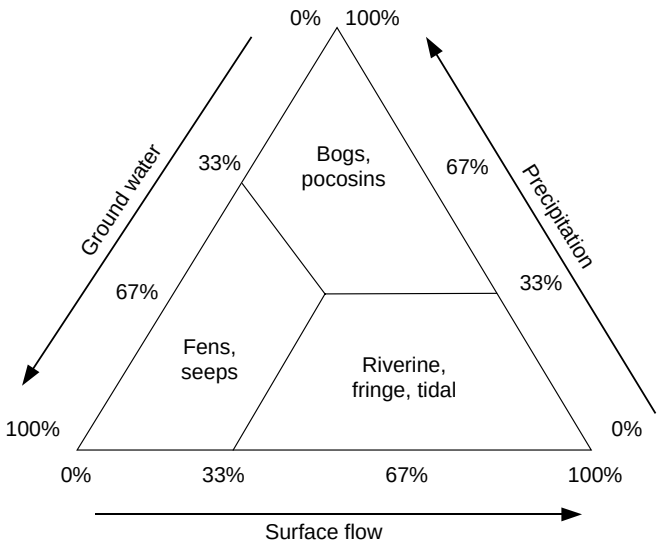


Figure 2 Relationship between water source and wetland vegetation. Modified from Brinson (1993)

setting, including shorelines of ponds, lakes, and rivers as well as seepage slopes and depressions. Common marsh species include the sedges, *Carex* and *Cyperus* spp., rushes, *Juncus* spp., and other species like cattail (*Typha* spp.).

Swamps are forested wetlands that exist along floodplains of varying size and on low-lying or flat geomorphic settings. Along river corridors, swamp forests are classified as bottomland, alluvial, or riparian forest on the basis of duration and frequency of river flooding. Bottomland forests are located in old river channels and oxbows of large river floodplains where inundation is of long duration (Figure 3). Bald cypress (*Taxodium distichum*) and Tupelo gum (*Nyssa aquatica*) are common tree species of bottomland floodplain swamps of the southern US (Wharton *et al.*, 1982). Alluvial forests are found on higher elevations of the floodplain, older terraces of large rivers, and floodplains of smaller rivers where the flood pulse is of shorter duration (Figure 3), several weeks to several months annually. Plant species diversity is greater in alluvial versus bottomland hardwood swamp forests because of the shorter duration of flooding and soil anoxia (Wharton *et al.*, 1982). Riparian forests form narrow corridors along perennial and intermittent streams (Brinson *et al.*, 1981b). These wetlands are inundated for short duration, hours to several days at a time. Riparian forests are important because they maintain stream water quality by shading and by intercepting sediment and pollutants that are transported from the surrounding terrestrial landscape (Naiman and Decamps, 1997).

Estuarine Wetlands

Estuarine wetlands are found in coastal regions where freshwater discharged from rivers interacts with and is diluted by seawater. Estuarine wetlands exist along a gradient of salinity and are characterized by frequent, cyclic flooding from astronomical tides (Figure 4). In estuarine wetlands, salinity and sulfides impose additional stresses on the wetland vegetation in addition to stress imposed by soil anoxia. Most estuarine wetlands are dominated by herbaceous emergent vegetation although forest vegetation replaces herbaceous emergent species in tropical and subtropical estuaries where mangroves dominate.

Salt marshes are found near the estuarine-ocean interface where seawater salinity is diluted from 35 ppt down to 20–30 ppt (Wiegert and Freeman, 1990). Salt marshes are flooded twice daily by the astronomical tides. Predictable tidal pulsing of water, sediment, and nutrients

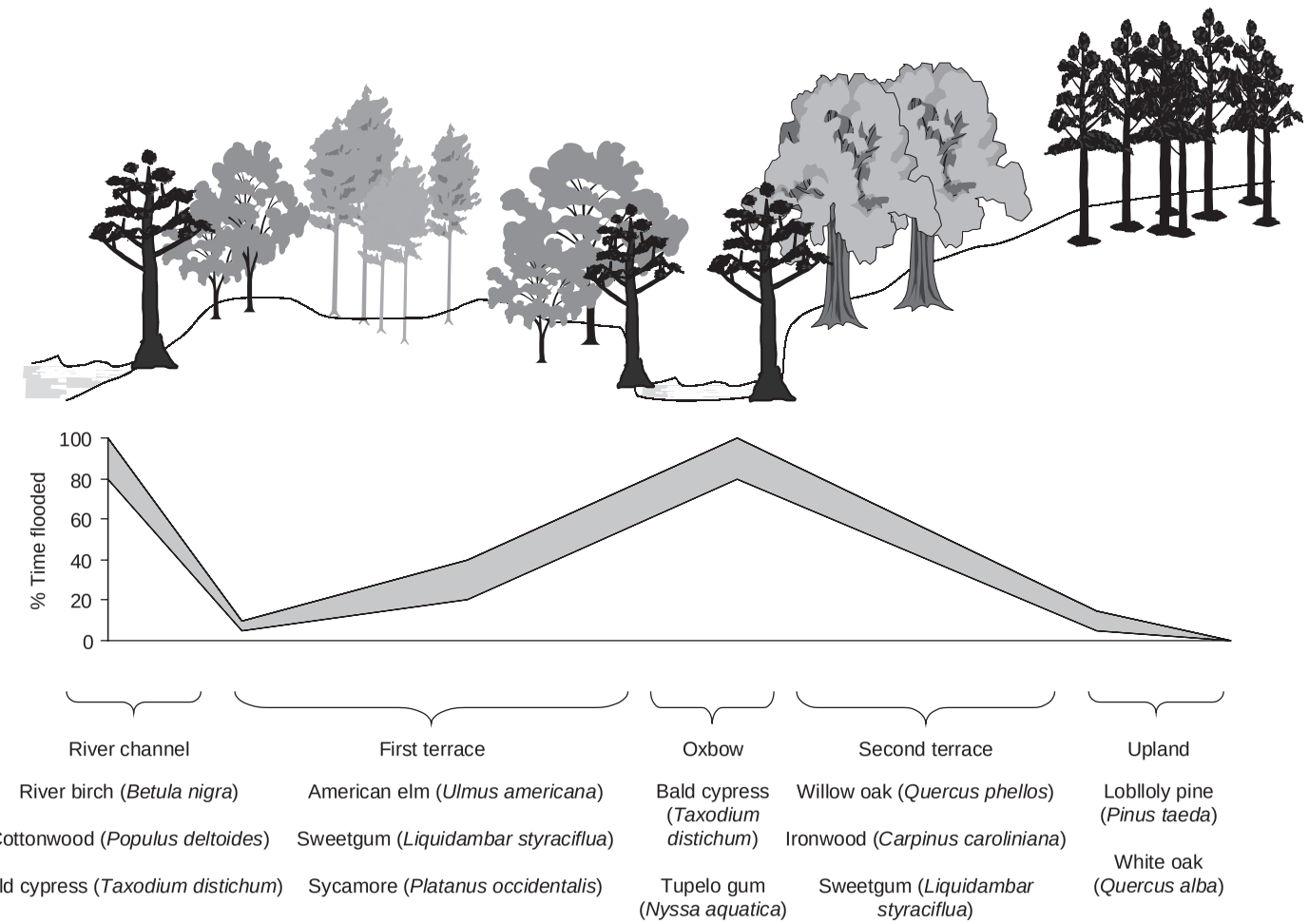
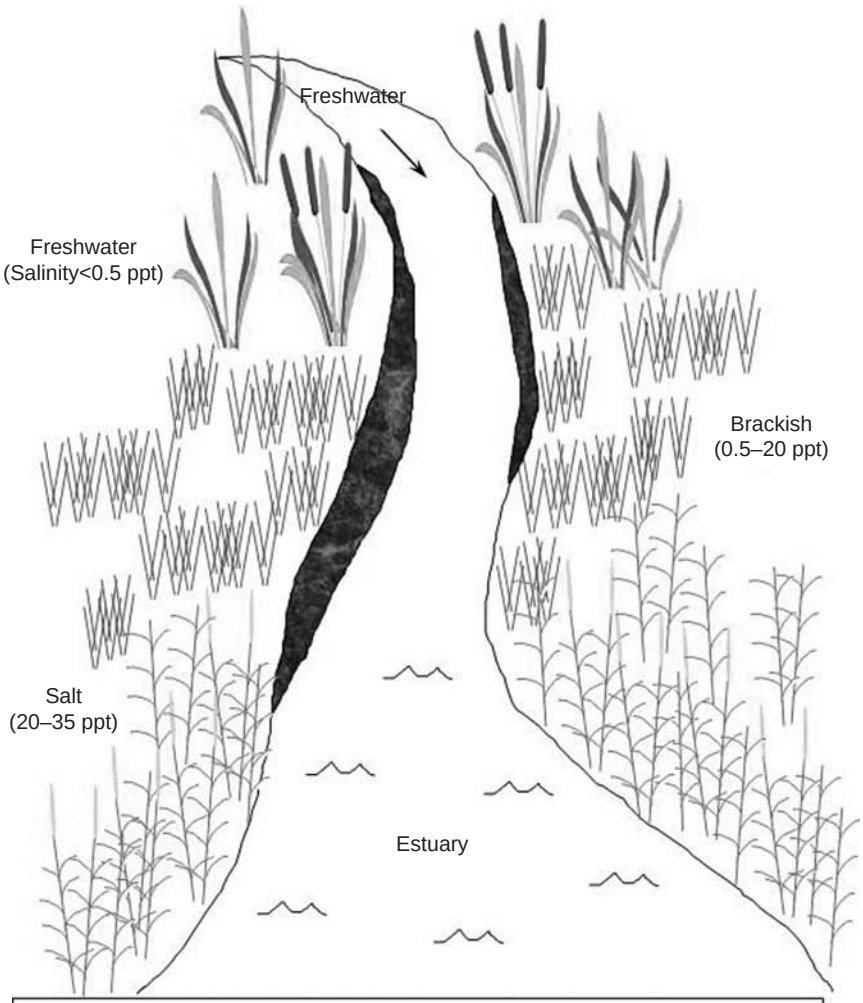


Figure 3 Relationship between forested wetland species and hydroperiod in a southeastern US swamp forest. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs



Plant species	
Freshwater marsh	Cattail (*Typha* spp.), maidencane (*Panicum hemitomon*), wild rice (*Zizania aquatica*), pickerelweed *(Pontedaria cordata*), arrow arum (*Peltandra virginica*), duck potato (*Sagittaria* spp.)
Brackish marsh	Needlerush (*Juncus* spp.), saltmeadow cordgrass (*S. patens*), giant cordgrass (*S. cynosuroides*), spikegrass (*Distichlis spicata*) marsh elder (*Iva frutescens*)
Salt marsh	Smooth cordgrass (*Spartina alterniflora*), glasswort (*Salicornia* spp.)
Mangrove (not shown)	Red (*Rhizophora mangle*), black (*Avicennia germinans*) and white (*Laguncularia racemosa*) mangrove

Figure 4 Relationship between wetland plant communities and salinity along the estuarine continuum. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

makes these wetlands among the most productive in the world. Salt marshes are dominated by nearly monotypic stands of cordgrass *Spartina alterniflora* in the eastern US, *S. foliosa* on the pacific coast, and *S. anglica* in Europe. Other salt-tolerant species such as saltwort (*Salicornia* spp.) are also present.

Brackish marshes exist in areas upstream and landward of salt marshes where salinity is lower, 0.5 ppt to 20 ppt (Figure 4) (Eluterius, 1984). Brackish marshes often are the dominant wetlands in areas of reduced astronomical tides such as the Gulf Coast region and the Albemarle–Pamlico estuaries of North Carolina. Plant species

diversity of brackish-water marshes is greater than in salt marshes because salinity stress is less. Vegetation of brackish marshes includes needlerush (*Juncus roemerianus, J. gerardi*) spp., saltmeadow cordgrass (*Spartina patens*), giant cordgrass (*Spartina cynosuroides*), spikegrass (*Distichlis spicata*), and salt-tolerant woody shrubs (*Iva frutescens*) (Stout, 1984, 1988).

Tidal freshwater marshes occur at the landward (most upstream) end of the estuary where seawater rarely penetrates (salinity <0.5 ppt) but where tidal inundation still occurs (Figure 4). Tidal freshwater marshes are common in large river estuaries such as the Hudson River, Delaware River, Chesapeake Bay, and Mississippi delta of Louisiana (Odum, 1988). Tidal freshwater marshes often have much higher-plant species diversity than either brackish marshes or salt marshes (Simpson *et al.*, 1983; Odum, 1988). Common species include emergents like cattail (*Typha* spp.), giant cutgrass (*Zizaniopsis milaceae*), wild rice (*Zizania aquatica*), maidencane (*Panicum hemitomon*), saw grass (*Cladium jamaicense*), bulrushes (*Schoenoplectus* spp.), and others as well as soft-stemmed species like pickerelweed (*Pontedaria cordata*), arrow arum (*Peltandra virginica*), and duck potato (*Sagittaria* spp.) (Odum *et al.*, 1984; Odum, 1988).

Mangrove forests are the dominant wetlands of estuaries in tropical and subtropical regions (Lugo and Snedaker, 1974; Odum *et al.*, 1982). In the intertidal zone, mangrove forests are dominated by red mangrove, *Rhizophora mangle*. Inland from the red mangrove zone, where tidal inundation is less frequent and less tidal flushing occurs, black mangrove (*Avicennia germinans*) and white mangrove (*Laguncularia racemosa*) are found (Odum *et al.*, 1982). Because mangroves are intolerant of below freezing temperatures, they are restricted to latitudes between about 25° north and 25° south (Mitsch and Gosselink, 2000).

PLANT ADAPTATIONS TO WATERLOGGING AND ANOXIA

In order to grow and reproduce in anaerobic soil environments, wetland plants must possess adaptations that maintain aerobic respiration to support cell growth, maintenance, and reproduction. In the absence of oxygen, vegetation must respire via anaerobic biochemical pathways that produce less energy to support cell maintenance and growth and that also produce metabolites (e.g. ethanol) that are toxic unless eliminated from the system. Because most respiration occurs in the roots of plants (Nobel, 1999), many adaptations to anoxic conditions involve morphological changes in stem or leaf structure that are designed to facilitate oxygen transport to the roots. Other adaptations to anoxia involve alteration of metabolic processes and pathways that reduce or eliminate toxic metabolites produced during anaerobic respiration.

Morphological Adaptations

Morphological adaptations involve physical changes in plant structure to facilitate oxygen transport to the roots. These adaptations include aerenchyma, hypertrophy leading to buttressed (swollen) trunks, pneumatophores, adventitious roots, shallow root systems, root aeration, pressurized ventilation, radial oxygen loss from roots and storage of carbohydrate reserves.

Aerenchyma is spongy tissue containing large hollow conduits in roots, stems, and leaves that transport oxygen-rich air to the roots (Armstrong, 1978). Most herbaceous wetland species, including cattail (*Typha* spp.) and many woody species, possess aerenchyma tissue (Crawford, 1993). ***Hypertrophy*** is swelling of the stem base in woody and herbaceous plants that occurs when aerenchyma is formed. In woody plants such as bald cypress and Tupelo gum (*Nyssa aquatica*), it leads to the formation of buttressed or swollen trunks (Figure 5). ***Pneumatophores*** and prop roots are erect roots that extend aboveground in mangrove species such as black mangrove (*Avicennia* spp.) and red mangrove (*Rhizophora mangle*) (Figure 6). The freshwater tree, bald cypress, possesses knees that vaguely resemble pneumatophores but whose function is less well understood. Buttressed trunks and pneumatophores increase the surface area and number of lenticels (pores in the stems of woody plants through which gases are exchanged) that facilitate oxygen to diffuse into the plant. They also stabilize wetland trees in waterlogged and soft soils.

Wetland plants also produce ***shallow root systems*** and ***adventitious roots*** to enhance oxygen uptake in anaerobic soils. In the case of hydrophytic vegetation, roots are concentrated at or near the soil surface in order to maximize oxygen diffusion from the air or overlying water column.

Figure 5 Cypress-gum swamp forest in southwestern Georgia USA. Note the buttressed or swollen trunks of bald cypress (*Taxodium distichum*) and Tupelo gum (*Nyssa aquatica*) Photo by Christopher Craft

Figure 6 Mangrove forest in Jervis Bay, South New Wales, Australia. Note the pneumatophores of black mangrove, *Avicennia* sp. Photo by Jim Lynch

When wetlands are flooded, some species produce adventitious roots that grow above ground into the water column or along the soil surface to assimilate oxygen directly from the water column or air (Crawford, 1993).

Morphological adaptations that enhance ***root aeration*** are important for sustaining growth in deep waters. The roots of cattail (*Typha*), spikerush (*Eleocharis* spp.) and other species contain porous cortices and thin steles that favor increased gas diffusion, deeper rooting and fine lateral root development relative to other species, adaptations that enable it to dominate in areas of deep water and long hydroperiod (Sorrell *et al.*, 2000). In the Florida Everglades, enhanced root aeration is thought to contribute to invasion by cattail into areas previously dominated by saw grass (*Cladium jamaicense*), whose roots contain less porous cortices, thicker steles and, thus, reduced gas transport (Sorrell *et al.*, 2000).

Some wetland plants augment diffusion of O_2 to the roots by means of ***pressurized ventilation*** (Dacey, 1981) or convective diffusion (Brix, 1993a). Pressurized ventilation occurs when gradients of temperature and pressure between the atmosphere and roots drive oxygen-rich air through the stomata of young leaves, down the shoots to the roots. Air containing respired compounds such as CO_2 is forced up more porous older stems and leaves where it is expelled back into the atmosphere. Pressurized ventilation requires aerenchyma tissue and it occurs in many wetland species, including yellow waterlily (*Nuphar luteum*) (Dacey, 1981), common reed (*Phragmites australis*), *Typha* spp., *Juncus* sp. *Eleocharis*, and other emergents with cylindrical stems and linear leaves (Armstrong and Armstrong, 1991; Brix, 1992).

Radial oxygen loss (ROL) is another adaptation to anoxic conditions driven by diffusion. Radial oxygen loss involves leakage of oxygen from the roots, oxidizing the area of soil around the roots and increasing the soil's oxidation-reduction potential (Armstrong, 1978; Ernst, 1990). Radial oxygen loss benefits the plant by oxidizing reduced metals (Fe^{2+}, Mn^{2+}) and sulfides that, otherwise, may be toxic. Evidence for ROL is observed in the rust colored (Fe^{3+}-rich) zones in the immediate vicinity of live roots. Radial oxygen loss has been observed in a variety of wetland plants, including woody, herbaceous emergent, and floating leaved species (Michaud and Richardson, 1989; Cronk and Fennessy, 2001).

One mechanism to combat short-term anoxia involves production of ***large carbohydrate reserves*** that are stored in roots. These reserves can be used to support anaerobic metabolism when soils are flooded or saturated for short periods of time. Plants with large rhizomes and carbohydrate reserves such as *Typha* and *Phragmites* can survive longer in flooded soils than species with small reserves (e.g. *Juncus, Carex*) (Barclay and Crawford, 1982).

Metabolic Adaptations

When soils are flooded and become anoxic, plants respond by shifting from aerobic metabolism, where CO_2 is the primary end product, to anaerobic metabolism, where ethanol, lactic acid, and other compounds are produced (Crawford, 1993; Summers *et al.*, 2000). Anaerobic respiration produces only 2 mol of ATP for every mole of glucose oxidized whereas aerobic respiration produces 38 mol of ATP (Atlas and Bartha, 1981). Thus, anaerobic respiration does not provide the high levels of energy (ATP) needed to maintain growth and metabolism that aerobic respiration does. One anaerobic pathway, ethanol fermentation, enables wetland plants to maintain high energy levels at least temporarily (Mendelssohn and Burdick, 1988). But, because ethanol is toxic to plants, wetland vegetation must be able to remove excess ethanol or convert it to a nontoxic compound such as malate. While some studies suggest malate production as an alternative metabolic end product (McManmon and Crawford, 1971), other researchers have observed that malate levels do not always increase during anaerobic metabolism (Saglio *et al.*, 1980; Menegus *et al.*, 1989). Also, ethanol may not be as toxic to plants as previously believed (Cronk and Fennessy, 2001) because when flooded, ethanol diffuses out of the roots and into the rooting medium where its toxicity is diluted (Summers *et al.*, 2000).

Davies (1980) suggested that short-term tolerance to anaerobic condition involves regulation of cellular pH to prevent cytoplasmic acidosis rather than accumulation of nontoxic compounds such as malate produced during anaerobic respiration. In plants, cytoplasmic acidosis occurs within minutes of the onset of anoxia as anaerobic metabolism kicks in and pyruvate is converted to lactic acid, leading to a decrease in cellular pH. In flood tolerant and intolerant plants, lactic acid is shunted to the vacuole, where it is isolated from the cytoplasm. In flood intolerant

plants such as maize, after a short period of time (10 h), acid leaks from the vacuole and acidifies the cytoplasm (Roberts, 1988). In flood tolerant species, lactic acid and other anaerobic respiration products are stored in vacuoles. Some data suggests that isolation of lactic acid in vacuoles may not be that important because, for some species, compounds such as succinate accumulate rather than lactic acid (Menegus *et al.*, 1989; Summers *et al.*, 2000).

HYDROLOGIC CONTROL OF PLANT DISTRIBUTION, PRODUCTIVITY, AND NUTRIENT CYCLING

Hydrology not only determines which plant species will survive and thrive in wetlands, it also determines their distribution within the wetland as well as their productivity and nutrient uptake characteristics. Plants vary in their tolerance to flooding and anoxic soil conditions, leading to their classification as obligate (OBL), facultative wet (FACW), facultative (FAC), facultative upland (FACU), or upland (UPL) species based on their degree of tolerance (Reed, 1988). Obligate species are highly tolerant of anoxic soils and, thus, are found in areas of long to continuous inundation within the wetland (Table 2). Upland species, in contrast, are intolerant of flooding and are, seldom if ever, found growing in wetlands (Table 2).

Species Distribution and Zonation

Within wetlands, vegetation frequently exhibits strong patterns of zonation that correspond to depth, duration, and/or frequency of inundation. For example, freshwater marsh vegetation exhibits striking patterns of zonation related to depth and duration of inundation (Figure 7) (Zedler, 1987; Kantrud *et al.*, 1989; Schalles and Shure, 1989). In deeper water where inundation is continuous, floating aquatic plants dominate. In shallower water where inundation is frequent and of long duration, herbaceous emergents (*Typha*) dominate. Hydrophytes tolerant of short hydroperiod and shallow flooding dominate along the wetland:upland boundary. Terrestrial vegetation dominates in unsaturated, aerobic soils above the upper limit of flooding.

Similar patterns of zonation are observed in forested wetlands. In floodplain forests of the southern US, areas of permanent inundation such as oxbows and abandoned river channels are dominated by bald cypress (*Taxodium distichum*) and tupelo gum (*Nyssa aquatica*) (*see* Figure 3 for example, Wharton *et al.*, 1982). Areas inundated by periodic overbank flooding are dominated by hardwoods such as black gum (*Nyssa sylvatica*), green ash (*Fraxinus pennsylvanicus*), and red maple (*Acer rubrum*). The highest, driest areas in the floodplain are dominated by species such as sweet gum (*Liquidambar styraciflua*) and sycamore (*Platanus occidentalis*) (Wharton *et al.*, 1982).

Plant Productivity

Hydroperiod, along with other factors such as salinity, pH, and nutrients, also determines plant productivity – the quantity of new leaves, wood, and roots produced per unit area and time. Odum *et al.* (1979) and others (Brinson *et al.*, 1981a; Odum *et al.*, 1995) suggested that periodic flooding subsidizes plant productivity by supplying nutrients to the wetland and removing toxins from the soil. Continuous flooding, however, acts as a stressor on vegetation by producing anoxic soil conditions that inhibit plant growth. The effects of inundation on plant productivity are best described by the subsidy-stress model (Odum *et al.*, 1979), which states that periodic inundation enhances

Table 2 Classification of plant species based on their frequency of occurrence in wetlands. Developed by Reed (1988) for US Fish and Wildlife Service

Indicator Status	Definition	Examples
Obligate (OBL)	Found in wetlands 99 times out of 100. (In uplands under natural conditions, is found less than once in 100 times.)	*Typha latifolia* (cattail), *Taxodium distichum* (bald cypress), many sedges (*Carex* spp. *Cyperus* spp.), and rushes (*Juncus* spp.)
Facultative wet (FACW)	Found in wetlands 67–99% of the time	*Fraxinus pennsylvanica* (green ash), *Quercus palustris* (pin oak), *Impatiens capensis* (jewel weed)
Facultative (FAC)	Found in wetlands 33–67% of the time. (Equally likely to be found in uplands.)	*Acer rubrum* (red maple), *Festuca rubra* (red fescue), *Ambrosia trifida* (ragweed)
Facultative upland (FACU)	Found in wetlands 1–33% of the time	*Liriodendron tulipifera* (tulip tree), *Cornus florida* (dogwood), *Prunus serotina* (black cherry)
Upland (UPL)	Found in wetlands <1% of the time	*Quercus prinus* (chestnut oak), *Pinus ponderosa* (ponderosa pine), *Zea mays* (corn)

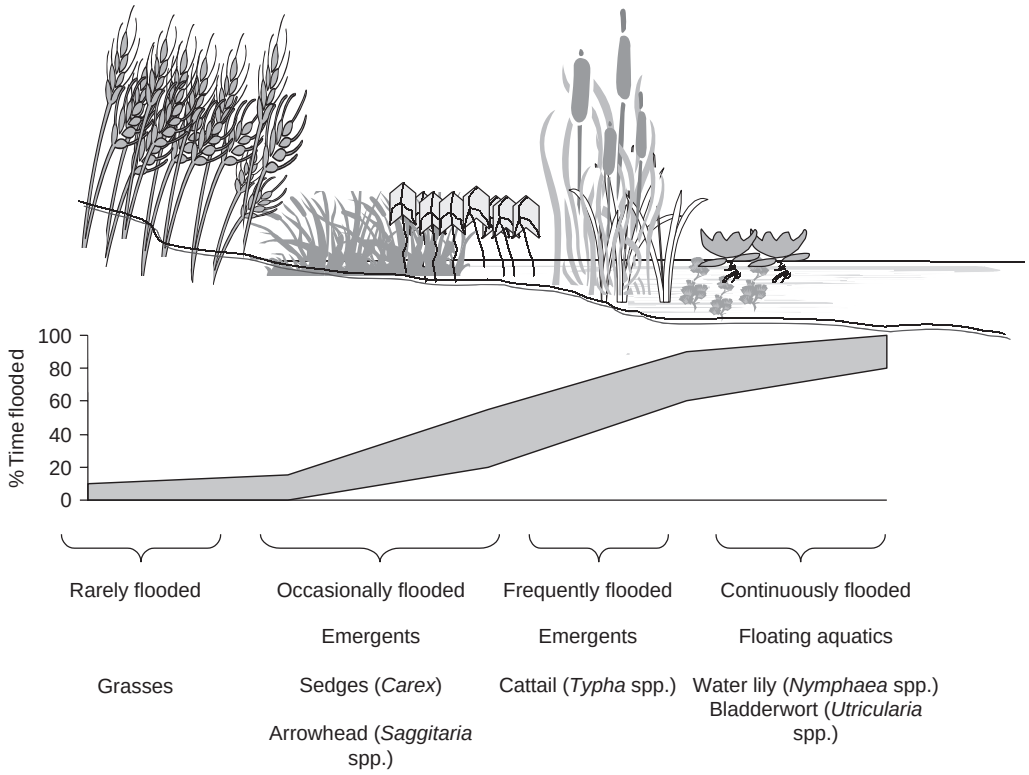


Figure 7 Relationship between zonation of plant species and hydroperiod in a freshwater marsh. (Reprinted with permissions from Brix, H (1993a). Macrophyte mediated oxygen transfer in wetlands: transport mechanisms and rates. In Constructed Wetlands for Water Quality Improvement, Morshiri, GA (Ed) pp 393–398. © CRC Press, Boca Raton, Florida, USA). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

plant growth, whereas continuous inundation or absence of inundation lead to reduced productivity (*see* Figure 8, for example). This hypothesis was developed to describe the high productivity frequently observed in tidal marshes and floodplain forests that are exposed to rhythmic inundation, ranging from twice daily inundation in salt mashes (Steever *et al.*, 1976) to annual inundation in floodplain forests (Broadfoot, 1967; Brinson *et al.*, 1981a). Mitsch and Rust (1984) refined the subsidy-stress model for alluvial floodplain forests by suggesting winter flooding is more beneficial to plant growth than summer flooding. According to this hypothesis, winter flooding provides a nutrient subsidy, but because the plants are dormant, anaerobic conditions caused by prolonged inundation do not stress the plant. Periodic flooding also may reduce stressors such as acidity by increasing soil pH through reduction of ferric (Fe^{3+}) iron to ferrous (Fe^{2+}) iron that consumes H^+ ions.

The subsidy-stress hypothesis of wetland plant production is supported by comparison of bald cypress (*Taxodium distichum*) productivity along a hydroperiod gradient from well drained to stagnant (standing water) hydrology. Cypress growth is greatest at intermediate (periodically inundated) hydroperiod as compared to well-drained or continuously flooded soils (Figure 8, Conner and Day, 1976, 1982; Mitsch and Ewel, 1979; Taylor *et al.*, 1990). A

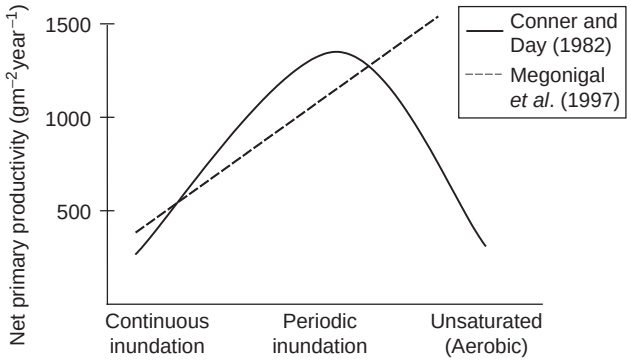


Figure 8 Relationship between aboveground net primary production and hydroperiod of floodplain swamp forest vegetation

recent study by Megonigal *et al.* (1997), however, found that growth was greatest in well-drained floodplain soils and decreased with increasing hydroperiod and anoxic conditions (Figure 8). Thus, the results of this study suggest that the stress caused by soil anoxia may be greater than the "subsidy" of nutrients provided by flooding.

Primary production also depends on hydrodynamics, especially water, and nutrient loads. Loading rate, the

quantity of nutrients that pass through the wetland, depends on the quantity of water and concentration of nutrients in water that flows across the wetland. Catchment size, land use in the watershed and connectivity to flowing aquatic ecosystems such as rivers and estuaries determine loading rates. Floodplain forests and tidal marshes have strong connectivity to aquatic ecosystems. The high productivity ascribed to tidal marshes, mangrove forests, and floodplain forests is attributed, in part, to greater water and nutrient inputs; that is, the subsidy part of the subsidy-stress hypothesis (Lugo and Snedaker, 1974; Gosselink and Turner, 1978; Mitsch *et al.*, 1979, 1991; Odum, 1980; Brinson *et al.*, 1981a,b; Hopkinson, 1992). Brown (1981), for example, observed a positive relationship between bald cypress production and phosphorus load in cypress swamps that varied from low precipitation-driven, low nutrient cypress domes to highly enriched cypress swamps receiving wastewater. Wetlands that lack connectivity include bogs and depressional wetlands, where precipitation is the primary water source. Plant productivity of these wetlands often is low relative to wetlands with strong connections (Moore and Bellamy, 1974; Craft, 2001).

Nutrient Cycling

Increased nutrient loadings associated with greater connectivity also accelerate nutrient cycling by vegetation. As nutrient loading increases, wetland vegetation assimilates more nutrients. However, the efficiency of assimilation decreases and the wetland becomes "leaky" with respect to nutrient retention. Comparison of phosphorus cycling between low connectivity cypress dome and high connectivity floodplain bald cypress swamps illustrates this point (Table 3). In the floodplain cypress swamp, greater connectivity from river flooding resulted in water and P loadings that were 32 and 140 times greater, respectively, than in than the low connectivity, precipitation-driven cypress dome.

Net primary production of the floodplain cypress swamp was 2.5 times greater than in the cypress dome and P uptake by vegetation was 3 times greater in the floodplain swamp ($1.22\,g\,m^{-2}\,yr^{-1}$) than in the cypress dome ($0.41\,g\,m^{-2}\,yr^{-1}$). The low connectivity cypress dome was highly efficient with respect to P cycling. Nearly half (44%) of the P taken up annually by vegetation was translocated belowground at the end of the growing season to be used the following growing season. In contrast, in the high connectivity floodplain swamp, only 25% of the P taken up each year was translocated belowground. The remaining 75% of P was lost to senescence and leaching, indicative of the "leakiness" of this highly connected, high nutrient (P) loaded system.

Hydrology-based Classification System for Wetlands

Recognition of the importance of hydrology to wetland nutrient and carbon cycles led to the development of a hydrology-based classification system for wetlands. The hydrogeomorphic (HGM) approach was developed by Brinson (1993) and others (Hauer and Smith, 1998) to assess hydrologic, biogeochemical, and habitat functions of wetlands. Using the HGM approach, wetlands are classified based on their geomorphic setting, dominant water source, and dominant hydrodynamics (Table 4). For example, an important function of wetlands is their ability to improve

Table 3 Carbon and phosphorus cycling characteristics of a low connectivity (Okefenokee GA) and a high connectivity (Cache River IL) cypress (*Taxodium distichum*) swamp forest (Data from Mitsch *et al.*, 1979 and Hopkinson *et al.*, 1992)

	Okefenokee GA (low connectivity)	Cache River IL (high connectivity)
Hydrology:		
Water source	Precipitation	River flooding
Hydrologic input	$170\,cm\,yr^{-1}$	$5500\,cm\,yr^{-1}$[a]
Nutrient input	Low ($0.57\,g\,P\,m^{-2}\,yr^{-1}$)	High ($80\,g\,P\,m^{-2}\,yr^{-1}$)[b]
Carbon Cycling:		
Aboveground biomass ($kg\,C\,m^{-2}$)	13.8	16.7
Net primary production ($g\,C\,m^{-2}\,yr^{-1}$)	311	801
Phosphorus Cycling:		
Uptake ($g\,P\,m^{-2}\,yr^{-1}$)	0.41	1.22
Aboveground storage ($g\,P\,m^{-2}$)	4.6	6.3
Translocation belowground (%)[c]	0.18 (44%)	0.31 (25%)
Senescence & Death (%)[c]	0.16 (39%)	0.77 (63%)
Loss to leaching (%)[c]	0.07 (17%)	0.14 (12%)
Retention in vegetation ($\%\,yr^{-1}$)[d]	72%	1.5%[e]

[a]Surface flooding accounts for $5300\,cm\,yr^{-1}$.
[b]$3.6\,g\,P\,m^{-2}\,yr^{-1}$ were retained in the wetland.
[c]Percent of annual uptake.
[d]Percent of nutrient input to wetland.
[e]34% of the P retained in the wetland ($3.6\,g\,m^{-2}\,yr^{-1}$) was retained in vegetation.

Table 4 Hydrogeomorphic (HGM) classification system for wetlands to assess ecological functions such as water storage, pollutant removal, and habitat (From Brinson, 1993). With respect to the ecological function *pollutant removal*, wetlands characterized by lateral surface flow have greater capacity to trap sediment than precipitation or groundwater fed wetlands

Geomorphic setting	Water Source	Hydrodynamics	Wetland type	Ecological function (sediment removal)
Depressional	precipitation	vertical	prairie potholes, vernal pools	low
Mineral soil flats	precipitation	vertical	wet pine flats	low
Organic soil flats	precipitation	vertical	peat bogs	low
Riverine	surface water	unidirectional, lateral	floodplain forests	high-very high
Estuarine fringe	surface water	bidirectional, lateral	salt marshes, mangroves	medium-high
Lacustrine fringe	surface water	bidirectional, lateral	Great lakes marshes	medium-high
Slope	groundwater	unidirectional lateral	fens, seepage wetlands	low

or maintain water quality by trapping sediment and pollutants. Highly connected wetlands such as floodplain forests have great potential for trapping sediment because of the high loadings of water and sediment during overbank flooding. Based on the HGM classification system, riverine, and fringe wetlands have greater potential to remove sediments than organic soil flats (peat bogs) or depressional wetlands that receive water mostly from precipitation (Table 4). Conversely, peat bogs have greater potential to sequester carbon through peat accretion, an important ecological function that absorbs atmospheric carbon dioxide, as compared to wetlands in other hydrogeomorphic settings.

CONSTRUCTED WETLANDS

Constructed wetlands are designed and built for water quality remediation and to replace wetland dependent functions lost when natural wetlands are destroyed. The ability of natural wetlands to filter or purify water has led to the construction of wetlands to treat wastewater, stormwater, mine drainage, animal waste, nonpoint runoff from agricultural lands and other sources (Hammer, 1989; Moshiri, 1993; Dubowy and Reaves, 1994; Kadlec and Knight, 1996; Pries, 2002). "Treatment" wetlands have been constructed throughout Europe, the United States and elsewhere to treat wastewater by removing suspended solids, coliform bacteria, BOD, and nutrients (N, P) (Vymazal *et al.*, 1998; Vymazal, 2001; Pries, 2002). In some instances, constructed wetlands are used to "polish" secondarily treated wastewater by removing nutrients (Kadlec and Knight, 1996; Vymazal *et al.*, 1998).

In the past 20 years, wetlands have been created and restored to mitigate for ecological functions (i.e. hydrologic, biological productivity, biogeochemical cycling, habitat) lost when natural wetlands are degraded or destroyed by human activities such as drainage and placement of fill in wetlands (Zelazny and Feierabend, 1988; NRC, 1992). Although laws have been implemented to protect wetlands from these activities, wetlands continue to be lost as a result of numerous small-scale actions that frequently are exempt from regulations designed to protect them (NRC (2001)).

Treatment Wetlands

Treatment wetlands consist of cells containing various planting substrates (pea gravel, sand, muck) planted with fast-growing hydrophytic vegetation. The type of planting substrate depends on whether the system is constructed for surface flow or subsurface flow (Figure 9). Hydrophytic vegetation is planted to provide oxygen to the rooting zone to enhance decomposition (Wolverton *et al.*, 1983; Brix, 1993a) and nitrification (Reddy *et al.*, 1989) and to stabilize the hydraulic conductivity of the soil (Brix, 1993b). Vegetation with high growth rates and that oxygenate the rooting zone are most often employed in treatment wetlands. Construction of treatment wetlands requires precise knowledge of water and nutrient loadings in order to size them appropriately. Water depths need to be shallow enough so that emergent vegetation is not submerged and drowned, but hydrologic retention times need to be long enough to filter and remove incoming pollutants.

Treatment wetlands utilize a variety of hydrophytic vegetation, including free floating, floating leaved, submersed, and emergent plants. Free-floating vegetation systems employ water hyacinth (*Eichhornia crassipes*), duckweed (*Lemna* spp), water lettuce (*Pistia stratiotes*), or pennywort (*Hydrocotyle umbellata*) (Figure 9a). These plants have very high rates of growth and nutrient uptake (Reddy,

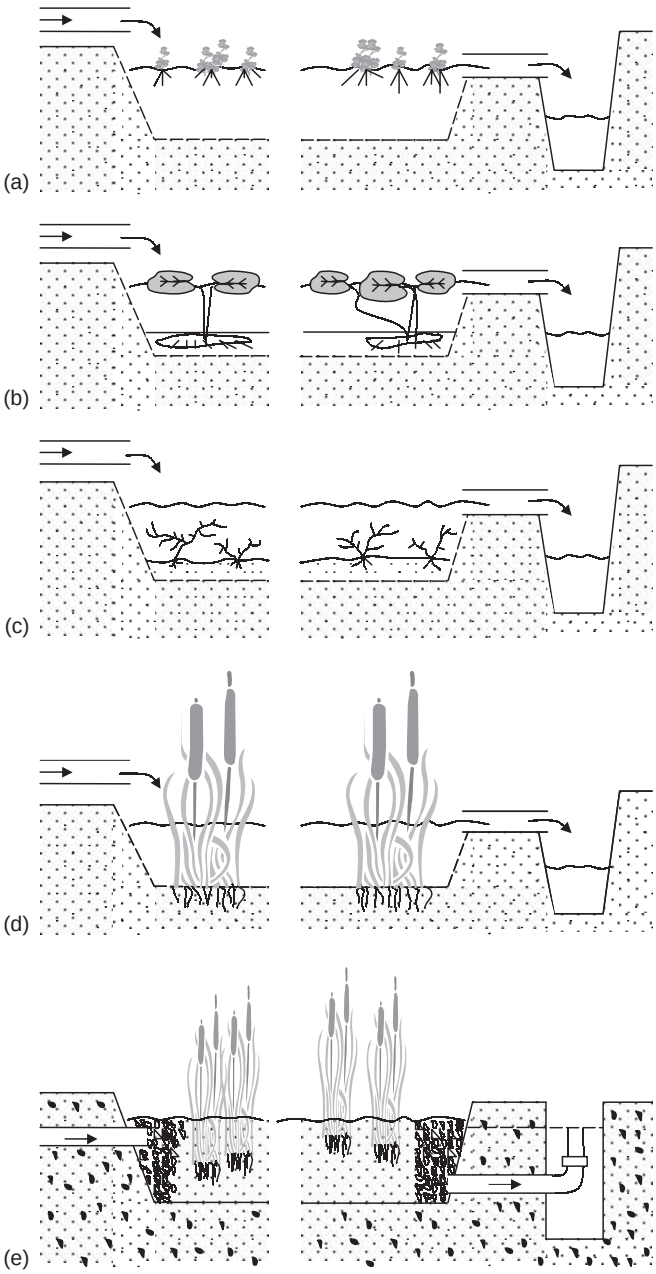


Figure 9 Design of (a) free floating, (b) floating leaved, (c) submersed, (d) surface flow and (e) subsurface flow treatment wetlands. Modified from Brix (1993b) (Reprinted with permissions from Brix, H (1993b). Wastewater treatment in constructed wetlands: system design, removal processes and treatment performance. In *Constructed Wetlands for Water Quality Improvement*, Moshiri G.A. (Ed.), Lewis Publishers: Boca Raton, pp. 9–22. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

1983) and, with periodic harvesting of plant biomass, N and P removal rates are high. Free-floating treatment systems vegetated with water hyacinth are limited to tropical and subtropical regions because of frost damage and reduced plant growth at temperatures below 10 °C (Vymazal, 1998). Systems vegetated with duckweed, which is more tolerant of cold temperatures, have been constructed in temperate regions (Brix, 1993b).

Floating leaved treatment wetlands rely on plants such as water lily (*Nymphaea* spp.), cow lily (*Nuphar lutea*), and lotus (*Nelumbo nucifera*) that are rooted in the substrate and whose leaves float on the surface (Figure 9b). These systems are not widely used because plant growth is less vigorous as compared to other hydrophytes. Treatment wetlands that rely on submersed vegetation also have been constructed (Figure 9c) but, like floating leaved wetlands, are not widely employed. These wetland systems are planted with waterweed (*Elodea* spp.), coontail (*Ceratophyllum* spp.), pondweed (*Potamogeton* spp.), hydrilla (*Hydrilla verticillata*), milfoil (*Myriophyllum heterophyllum*). Submersed vegetation needs high clarity, well-oxygenated water, so these systems are used to "polish" wastewater following primary (physical settling) and secondary treatment (microbial breakdown of organic matter) (Brix, 1993b).

Wetlands planted with emergent vegetation are the most common treatment wetlands. Fast-growing cosmopolitan species such as *Phragmites communis, Typha* spp., *Scirpus* spp., and *Juncus* spp. are usually planted in them. These species are widely distributed, possess high growth rates and are proficient at oxygenating the root zone (Gersberg *et al.*, 1986). Two types of emergent-based treatment wetlands are employed; surface flow or "free water" systems and subsurface flow systems (Figures 9d and e). In surface flow wetlands, water flows across the vegetated surface. Suspended solids settle out the under the low flow, quiescent conditions. Biochemical oxygen demand (BOD) is reduced by aerobic and anaerobic microbial decomposition. Nitrogen is removed by nitrification in aerobic surface waters followed by denitrification in the anaerobic soils (Wolverton *et al.*, 1983; Brix, 1993b; Hunt *et al.*, 2003). Removal of phosphorus in treatment wetlands is low relative to N and is controlled by sorption and precipitation with Al in acid soils and Ca in circumneutral soils (Reddy, 1983; Richardson and Craft, 1993). Uptake and storage of nutrients in plants represents a temporary sink for N and P because when vegetation senesces and dies, N and P are released back into the water column. Permanent storage of N and P occurs by accumulation of organic matter in soil, sorption, and precipitation with minerals (P) and removal from the system by denitrification (N).

In subsurface flow wetlands, wastewater moves either laterally (Figure 9e) or vertically through the soil or planting substrate. The substrate of subsurface flow wetlands is composed of gravel or crushed stone to maintain high hydraulic conductivity and reduce clogging with suspended solids. Subsurface flow treatment wetlands are widely used throughout Europe. These treatment wetlands are

Table 5 Mean removal efficiencies of surface flow and subsurface flow constructed data are from the North America database of constructed wetlands in Table 26–4 of Kadlec and Knight (1996)

	Concentration ($mg\,L^{-1}$)			Mass ($kg\,ha^{-1}\,day^{-1}$)		
	Inflow	Outflow		Load	Removed	
Surface flow wetlands (n = 69):						
BOD[a]	30.3	8.0	(74%)	7.2	5.1	(71%)
Suspended solids	45.6	13.5	(70%)	10.4	7.0	(70%)
Total N	9.03	4.27	(53%)	1.94	1.06	(55%)
Organic N	3.45	1.85	(46%)	0.90	0.51	(56%)
NH_4–N	4.88	2.23	(54%)	0.93	0.35	(38%)
NO_x^-–N[2]	5.56	2.15	(61%)	0.80	0.40	(50%)
Total P	3.78	1.62	(57%)	0.50	0.17	(34%)
PO_4–P	1.75	1.11	(37%)	0.29	0.12	(41%)
Subsurface flow wetlands (n = 25):						
BOD[a]	27.5	8.6	(69%)	29.2	18.4	(63%)
Suspended solids	48.2	10.3	(79%)	48.1	35.3	(74%)
Total N	18.92	8.41	(56%)	13.19	5.85	(44%)
Organic N	10.11	4.03	(60%)	7.28	4.05	(56%)
NH_4–N	5.98	4.51	(25%)	7.02	0.62	(9%)
NO_x^-–N[b]	4.40	1.35	(69%)	3.10	1.89	(61%)
Total P	4.41	2.97	(32%)	5.14	1.14	(22%)
PO_4–P[c]	–	–	–	–		

[a]Biochemical oxygen demand.
[b]NO_3–N + NO_2–N.
[c]No data.

also referred to as "reed bed treatment systems" because common reed (*Phragmites communis*) frequently is planted. Advantages of subsurface flow systems over surface flow wetlands include even distribution and rate of flow through the system, reduced breeding opportunities for mosquitoes and less nuisance odor problems (Kadlec and Knight, 1996). However, subsurface flow systems are more susceptible to clogging than surface flow wetlands (Kadlec and Knight, 1996) and also are less effective at N removal because of insufficient oxygen supplied by the roots to support nitrification (Brix, 1993b).

Comparison of nutrient removal efficiencies for North American surface flow and subsurface flow treatment wetlands are shown in Table 5. Generally, removal efficiencies for most wastewater constituents are similar for both types of wetlands. However, NH_4–N removal is greater for surface flow wetlands owing to greater nitrification in surface waters. Phosphorus removal is low (<50%) in both surface flow and subsurface flow wetlands primarily because of limited sorption in the sand or gravel planting substrate.

Hybrid systems employing subsurface horizontal and vertical flow wetlands in a series of cells have been developed to achieve higher treatment efficiency (Brix, 1993b). In these hybrid systems, wastewater first enters a subsurface vertical flow wetland where suspended solids and BOD are removed. It then flows into a subsurface vertical flow wetland cell where ammonium is nitrified to nitrate. The terminal cell consists of a subsurface horizontal flow wetland where nitrate is denitrified. Some systems are designed such that if nitrogen removal is incomplete, wastewater is recycled back to the first cell and treated again.

Mitigation Wetlands

In recent years, wetlands have been created and restored to replace ecological functions such as hydrologic storage and recharge, biological productivity, biodiversity, carbon sequestration, and water quality remediation that are lost when natural wetlands are degraded or destroyed by human activities. Mitigation for wetland loss is required under US Federal law as set forth in Section 404 of the Clean Water Act (NRC, 1992, 2001) and mitigation wetlands must be of the same type as the natural wetlands that they replace. For example, if a bottomland hardwood forest is drained and converted to upland habitat, then the law requires that the mitigation wetland is of similar habitat, for example, bottomland hardwood forest. Because mitigation wetlands frequently provide a lower level of function than natural wetlands, the amount of wetland acreage created or restored often is greater than the acreage lost (NRC, 2001). A ratio of up to 5:1 of mitigation to natural wetland is used for impacts to endangered species habitat (NRC, 2001).

Studies of mitigation wetlands revealed that development of ecological functions does not occur instantaneously. Wetlands with predictable hydroperiod or emergent vegetation develop faster than those with less predictable hydroperiod or woody vegetation. For example, development of wetland dependent functions proceeds faster on constructed salt marshes where tidal inundation is predictable and frequent

Table 6 Plant biomass, soil properties, microbial processes, and benthic infauna community composition of constructed *Spartina alterniflora* salt marshes versus natural salt marshes. Values separated by the same letter (**in bold**) are not significantly different ($p = 0.05$) based on pair-wise *t*-tests Data are from Craft *et al.* (2003) and unpublished data (methane production, denitrification)

	Constructed Marshes		
	1–3-year-old	26–28-year-old	Natural Marshes
Aboveground biomass ($g\,m^{-2}$)	340–670 **a**	850–1130 **b**	750–1240 **b**
Belowground biomass ($g\,m^{-2}$)[a]	130–450 **a**	2290–4250 **b**	2090–6210 **b**
Decomposition rate[b]	0.7–0.8 **a**	2.1–2.3 **b**	1.0–2.9 **b**
Methane production[c]	0.3–0.4 **a**	3.1–4.3 **b**	1.2–4.3 **b**
Denitrification[d]	7–14 **a**	70–105 **b**	18–42 **b**
Infauna density (no. $m^{-2} \times 1000$)[e]	19–52 **a**	118–138 **b**	95–157 **b**
Infauna taxa richness[f]	2.6–4.5 **a**	7.2–9.2 **b**	7.8–8.6 **b**
Soil organic C($g\,m^{-2}$, 0–30 cm)	440–570 **a**	2610–2620 **b**	3580–10 020 **c**
Soil total N ($g\,m^{-2}$, 0–30 cm)	50–60 **a**	180–210 **b**	250–520 **c**
Soil total P ($g\,m^{-2}$, 0–30 cm)	61–80	30–60	28–45

[a]0–30 cm depth.
[b](umol $CO_2\,g^{-1}$ soil sec^{-1}).
[c](femtomoles g^{-1} soil sec^{-1}).
[d](ng g^{-1} soil day^{-1}).
[e]0–5 cm depth.
[f](no. $7.07^{-1}\,cm^{-2}$ core, 0–5 cm depth).

than on isolated wetlands where hydroperiod is less predictable (Campbell *et al.*, 2002; Craft *et al.*, 2002; NRC, 2001). Similarly, wetlands dominated by emergent vegetation develop faster than forested wetlands because it takes a long time to grow a mature forest (Niswander and Mitsch, 1995; NRC, 2001). Even in mitigation salt marshes, some components of the ecosystem, such as soil organic matter and heterotrophic food webs, may take 25 years or more to develop (Craft and Sacco, 2003; Craft *et al.*, 2003).

Biomass of emergent vegetation in young (1–3-year-old) constructed salt marshes is less than in older (24–28-year-old) constructed marshes and natural marshes (Table 6). Microbial processes are reduced in young constructed marshes and, density and diversity of benthic infauna (invertebrates) is lower than in older constructed marshes and natural marshes. Even after 28 years, standing stocks of soil organic C and N are significantly less in constructed marshes as compared to natural marshes (Table 6). In constructed salt marshes, development of heterotrophic activity (microbial processes, benthic infauna) depends on accumulation of soil organic matter that is driven, by inputs of organic matter from the developing plant community and by anaerobic soil conditions brought on by frequent tidal inundation that slows or inhibits decomposition of organic matter (Craft *et al.*, 2002)

In some cases, mitigation wetlands never achieve functional equivalence to natural wetlands, because of inadequate hydrologic restoration (Galatowitsch and van der Valk, 1996; Magee *et al.*, 1999; Cole and Brooks, 2000; Brown and Veneman, 2001), initial soil characteristics such as low organic matter or nutrients (N) (Langis *et al.*, 1991; Shaffer and Ernst, 1999) or land use changes in the watershed that alter water, sediment, and contaminant loadings and facilitate invasion by exotic species (NRC, 1992; Zedler and Callaway, 1999).

Successful replacement of wetland dependent functions using mitigation wetlands requires that, in addition to recreating hydrologic conditions, soil characteristics, and plant communities of the natural wetland, one also must reestablish energy flows and nutrient cycles that link wetlands with aquatic and terrestrial ecosystems, a goal that is increasingly difficult to achieve in a world increasingly altered by human activities.

REFERENCES

Armstrong W. (1978) Root aeration in the wetland condition. In *Plant Life in Anaerobic Environments*, Hook D.D. and Crawford R.M.M. (Eds.), Ann Arbor Science Publishers: Ann Arbor, pp. 269–297.

Armstrong J. and Armstrong W. (1991) A convective throughflow of gases in *Phragmites australis* (Cav.) Tin ex steud. *Aquatic Botany*, **39**, 75–88.

Aselmann J. and Crutzen P.J. (1989) Global distribution of natural freshwater wetlands and rice paddies, their net primary productivity, seasonality and possible methane emissions. *Journal of Atmospheric Chemistry*, **8**, 307–358.

Atlas R.M. and Bartha R. (1981) *Microbial Ecology: Fundamentals and Applications*, Addison-Wesley Publishing: Reading.

Barclay A.M. and Crawford R.M.M. (1982) Plant growth and survival under strict anaerobiosis. *Journal of Experimental Botany*, **33**, 541–549.

Bridgham S.D., Pastor J., Jannsens J.A., Chapin C. and Malterer T.J. (1996) Multiple limiting gradients in peatlands: a call for a new paradigm. *Wetlands*, **16**, 45–65.

Brinson M.M., Lugo A.E. and Brown S. (1981a) Primary productivity, decomposition and consumer activity in freshwater wetlands. *Annual Review of Ecology and Systematics*, **12**, 123–161.

Brinson M.M., Swift B.L., Plantico R.C. and Barclay J.S. (1981b) *Riparian Ecosystems: Their Ecology and Status*, FWS/OBS-81/17, U.S. Fish and Wildlife Service, Biological Services Program: Washington.

Brinson M.M. (1993) *A Hydrogeomorphic Classification for Wetlands*, Wetlands Research Program technical report WRP-DE-4, US Army, Corps of Engineers, Waterways Research Station. National Technical Information Service, Department of Commerce, Springfield, VA.

Brix H. (1992) Internal pressurization and convective gas flow in some emergent freshwater macrophytes. *Limnology and Oceanography*, **37**, 1420–1433.

Brix H. (1993a) Macrophyte mediated oxygen transfer in wetlands: transport mechanisms and rates. In *Constructed Wetlands for Water Quality Improvement*, Moshiri G.A. (Ed.), Lewis Publishers: Boca Raton, pp. 393–398.

Brix H. (1993b) Wastewater treatment in constructed wetlands: system design, removal processes and treatment performance. In *Constructed Wetlands for Water Quality Improvement*, Moshiri G.A. (Ed.), Lewis Publishers: Boca Raton, pp. 9–22.

Broadfoot W.M. (1967) Shallow-water impoundment increases soil moisture and growth of hardwoods. *Soil Science Society of America Journal*, **31**, 562–565.

Brown S.L. (1981) A comparison of the structure, primary productivity and transpiration of cypress ecosystems in Florida. *Ecological Monographs*, **51**, 403–427.

Brown S.C. and Veneman P.L.M. (2001) Effectiveness of compensatory wetland mitigation in Massachusetts. *Wetlands*, **21**, 508–518.

Campbell D.A., Cole C.A. and Brooks R.P. (2002) A comparison of created and natural wetlands in Pennsylvania. *Wetlands Ecology and Management*, **10**, 41–49.

Clymo R.S. (1978) Peat: mires, swamp, bog, fen and moor. In *Ecosystems of the World 4A*, Gore A.J.P. (Ed.), Elsevier Scientific Publishing: New York, pp. 159–224.

Cole C.A. and Brooks R.P. (2000) A comparison of the hydrologic characteristics of natural and created mainstem floodplain wetlands in Pennsylvania. *Ecological Engineering*, **14**, 221–231.

Conner W.H. and Day J.W. Jr (1976) Productivity and composition of a bald cypress-water tupelo site and a bottomland hardwood site in a Louisiana swamp. *American Journal of Botany*, **63**, 1354–1364.

Conner W.H. and Day J.W. Jr (1982) The ecology of forested wetlands in the southeastern United States, In *Wetlands: Ecology and Management*, Gopal B., Turner R.E., Wetzel R.G. and Whigham D.F. (Eds.), National Institute of Ecology and International Scientific Publications: Jaipur, pp. 69–87.

Craft C.B. (2001) Biology of wetland soils. In *Wetland Soils: Their Genesis, Hydrology, Landscape and Separation into Hydric and Nonhydric Soils*, Richardson J.L. and Vepraskas M.J. (Eds.), CRC Press: Boca Raton, pp. 107–135.

Craft C.B., Broome S.W. and Campbell C.L. (2002) Fifteen years of vegetation and soil development following brackish-water marsh creation. *Restoration Ecology*, **10**, 248–258.

Craft C., Megonigal P., Broome S., Stevenson J., Freese R., Cornell J., Zheng L. and Sacco J. (2003) The pace of ecosystem development of constructed *Spartina alterniflora* marshes. *Ecological Applications*, **13**, 1417–1432.

Craft C.B. and Sacco J.N. (2003) Long-term succession of benthic infauna communities on constructed *Spartina alterniflora* marshes. *Marine Ecology–Progress Series*, **257**, 45–58.

Crawford R.M.M. (1993) Root survival in flooded soils. In *Mires, Swamp, Bog, Fen and Moor*, *Ecosystems of the World*, *Vol. 4A*, Gore A. (Ed.), Elsevier Science: Amsterdam, pp. 257–283.

Cronk J.K. and Fennessy M.S. (2001) *Wetland Plants: Biology and Ecology*, Lewis Publishers: Boca Raton.

Crum H. (1995) *A Focus on Peatlands and Peatmosses*, The University of Michigan Press: Ann Arbor.

Dacey J.W.H. (1981) Pressurized ventilation in the yellow waterlily. *Ecology*, **62**, 1137–1147.

Davies D.D. (1980) Anaerobic metabolism and production of organic acids. In *The Biochemistry of Plants: A Comprehensive Treatise*, Vol. 2, Stumpf P.K. and Conn E.E. (Eds.), Academic Press: New York, pp. 581–611.

Dubowy P.J. and Reaves R.P. (1994) *Constructed Wetlands for Animal Waste Management*, Purdue Research Foundation: West Lafayette.

Eluterius L.N. (1984) Autecology of the black needlerush, *Juncus roemerianus*. *Gulf Research Reports*, **7**, 339–350.

Ernst W.H.O. (1990) Ecophysiology of plants in waterlogged and flooded environments. *Aquatic Botany*, **38**, 73–90.

Galatowitsch S.M. and van der Valk A.G. (1996) Characteristics of recently restored wetlands in the prairie pothole region. *Wetlands*, **16**, 75–83.

Gersberg R.M., Elkins B.V., Lyon R.S. and Goldman C.R. (1986) Role of aquatic plants in wastewater treatment by artificial wetlands. *Water Resources*, **3**, 363–368.

Gosselink J.G. and Turner R.E. (1978) The role of hydrology in freshwater wetland ecosystems. In *Freshwater Wetlands: Ecological Processes and Management Potential*, Good R.E., Whigham D.F. and Simpson R.L. (Eds.), Academic Press: New York, pp. 63–78.

Hammer D.A. (Ed.) (1989) *Constructed Wetlands for Wastewater Treatment*, Lewis Publishers: Boca Raton.

Hauer F.R. and Smith R.D. (1998) The hydrogeomorphic approach to functional assessment of riparian wetlands: evaluating impacts and mitigation on river floodplains in the U.S.A. *Freshwater Biology*, **40**, 517–530.

Hopkinson C.S. Jr (1992) A comparison of ecosystem dynamics in freshwater wetlands. *Estuaries*, **15**, 549–562.

Hunt P.G., Matheny T.A. and Szogi A.A. (2003) Denitrification in constructed wetlands used for treatment of swine wastewater. *Journal of Environmental Quality*, **32**, 727–735.

Kadlec R.H. and Knight R.L. (1996) *Treatment Wetlands*, Lewis Publishers: Boca Raton.

Kantrud H.A., Millar J.B. and van der Valk A.G. (1989) Vegetation of wetlands of the prairie pothole region. In *Northern Prairie Wetlands*, van der Valk A.G. (Ed.), Iowa State University Press: Ames, pp. 132–187.

Langis R., Zalejko M. and Zedler J.B. (1991) Nitrogen assessments in a constructed and a natural salt marsh of San Diego Bay. *Ecological Applications*, **1**, 41–50.

Lugo A.E. and Snedaker S.C. (1974) The ecology of mangroves. *Annual Review of Ecology and Systematics*, **5**, 39–64.

Magee T.K., Ernst T.L., Kentula M.E. and Dwire K.A. (1999) Floristic comparison of freshwater wetlands in an urbanizing environment. *Wetlands*, **19**, 517–534.

Matthews E. and Fung I. (1987) Methane emission from natural wetlands: Global distribution, area and environmental characteristics of sources. *Global Biogeochemical Cycles*, **1**, 61–86.

McManmon M. and Crawford R.M.M. (1971) A metabolic theory of flooding tolerance: the significance of enzyme distribution and behavior. *New Phytologist*, **70**, 299–306.

Megonigal J.P., Conner W.H., Kroeger W. and Shartitz R.R. (1997) Aboveground production in southeastern floodplain forests: a test of the subsidy-stress hypothesis. *Ecology*, **78**, 370–384.

Mendelssohn I.A. and Burdick D. (1988) The relationship of soil parameters and root metabolism to primary production in periodically inundated soils. In *The Ecology and Management of Wetlands*, *Ecology of Wetlands*, *Vol. 1*, Hook D.D., McKee W.F. Jr, Smith H.K., Gregory K., Burrell V.G. Jr, Devoe M.R., Sojka R.E., Gilbert S., Banks R., Stolzy L.H., Brooks C., Matthews T.D. and Shear T.A. (Eds.), Timber Press: Portland, Oregon, pp. 398–428.

Menegus F., Cattaruzza L., Chersi A. and Fronza G. (1989) Differences in the anaerobic lactate-succinate production and in the changes of cell sap pH for plants with high and low resistance to anoxia. *Plant Physiology*, **90**, 29–32.

Michaud S.M. and Richardson C.J. (1989) Relative radial oxygen loss in five plant plants. In *Constructed Wetlands for Wastewater Treatment*, Hammer D.A. (Ed.), Lewis Publishers: Chelsea, pp. 501–507.

Mitsch W.J., Dorage C.L. and Wiemhoff J.R. (1979) Ecosystem dynamics and a phosphorus budget of an alluvial cypress swamp in southern Illinois. *Ecology*, **60**, 1116–1124.

Mitsch W.J. and Ewel K.C. (1979) Comparative biomass and growth of cypress in Florida wetlands. *American Midland Naturalist*, **101**, 417–426.

Mitsch W.J. and Gosselink J.G. (2000) *Wetlands*, Van Nostrand Reinhold: New York.

Mitsch W.J. and Rust W.G. (1984) Tree growth responses to flooding in a bottomland forest in northeastern Illinois. *Forest Science*, **30**, 499–510.

Mitsch W.J., Taylor J.R. and Benson K.B. (1991) Estimating primary productivity of forested wetland communities in different hydrologic landscapes. *Landscape Ecology*, **5**, 75–92.

Moore P.D. and Bellamy B.J. (1974) *Peatlands*, Springer-Verlag: New York.

Moshiri G.A. (Ed.) (1993) *Constructed Wetlands for Water Quality Improvement*, Lewis Publishers: Boca Raton.

Naiman R.J. and Decamps H. (1997) The ecology of interfaces: riparian zones. *Annual Review of Ecology and Systematics*, **28**, 621–658.

Niswander S.F. and Mitsch W.J. (1995) Functional analysis of a two-year-old in-stream wetland: hydrology, phosphorus retention, and vegetation survival and growth. *Wetlands*, **15**, 212–225.

Nobel P.S. (1999) *Plant Physiology*, Academic Press: New York.

(NRC) National Research Council (1992) *Restoration of Aquatic Ecosystems*, National Academy Press: Washington.

(NRC) National Research Council (2001) *Compensating for Wetland Losses under the Clean Water Act*, National Academy Press: Washington.

Odum E.P. (1980) The status of three ecosystem-level hypotheses regarding salt marsh estuaries: tidal subsidy, outwelling and detritus-based food chains. In *Estuarine Perspectives*, Kennedy V.S. (Ed.), Academic Press: New York, pp. 485–495.

Odum E.P., Finn J.T. and Franz E. (1979) Perturbation theory and the subsidy-stress gradient. *BioScience*, **29**, 349–352.

Odum W.E. (1988) Comparative ecology of tidal freshwater and salt marshes. *Annual Review of Ecology and Systematics*, **19**, 147–176.

Odum W.E., McIvor C.C. and Smith T.J. III (1982) *The Ecology of the Mangroves of South Florida: A Community Profile*, Technical Report FWS/OBS/81-24, U.S. Fish and Wildlife Service, Division of Biological Services, Washington.

Odum W.E., Odum E.P. and Odum H.T. (1995) Nature's pulsing paradigm. *Estuaries*, **18**, 547–555.

Odum W.E., Smith T.J. III, Hoover J.K. and McIvor C.C. (1984) *The Ecology of Tidal Freshwater Marshes of the United States East Coast: A Community Profile*, FWS/OBS-84/17, U.S. Fish and Wildlife Service: Washington.

Pries J. (Ed.) (2002) *Treatment Wetlands for Water Quality Improvement: Quebec 2000 Conference Proceedings*, Pandora Press: Waterloo, Ontario.

Reddy K.R. (1983) Fate of nitrogen and phosphorus in a waste-water retention reservoir containing aquatic macrophytes. *Journal of Environmental Quality*, **12**, 137–141.

Reddy K.R., Patrick W.H. Jr and Lindau C.W. (1989) Nitrification-denitrification at the plant root - sediment interface in wetlands. *Limnology and Oceanography*, **34**, 1004–1013.

Reed P.B. (1988) *National List of Plant Species that Occur in Wetlands: National Summary*, Biological report 88(24), U.S. Fish and Wildlife Service, Washington.

Richardson C.J. and Craft C.B. (1993) Effective phosphorus retention in wetlands: Fact or fiction? In *Constructed Wetlands for Water Quality Improvement*, Moshiri G.A. (Ed.), Lewis Publishers: Boca Raton, pp. 271–282.

Roberts J.K.M. (1988) Cytoplasmic acidosis and flooding tolerance in crop plants. In *The Ecology and Management of Wetlands*, *Ecology of Wetlands*, *Vol. 1*, Hook D.D., *et al.* others (Eds.), Timber Press: Portland, Oregon, pp. 393–397.

Saglio P.H., Raymond P. and Pradet A. (1980) Metabolic activity and energy charge of excised maize root tips under anoxia. *Plant Physiology*, **66**, 1053–1057.

Schalles J.F. and Shure D.J. (1989) Hydrology, community structure and productivity patterns of a dystrophic Carolina bay wetland. *Ecological Monographs*, **59**, 365–385.

Shaffer P.W. and Ernst T. (1999) Distribution of soil organic matter in freshwater emergent open water wetlands in the Portland, Oregon metropolitan area. *Wetlands*, **19**, 505–516.

Simpson R.L., Good R.E., Leck M.A. and Whigham D.F. (1983) The ecology of freshwater tidal wetlands. *BioScience*, **33**, 255–259.

Sorrell B.K., Mendelssohn I.A., McKee K.L. and Woods R.A. (2000) Ecophysiology of wetland plant roots: a modeling comparison of aeration in relation to species distribution. *Annals of Botany*, **86**, 675–685.

Steever E.Z., Warren R.S. and Niering W.A. (1976) Tidal energy subsidy and standing crop production of *Spartina alterniflora*. *Estuarine and Coastal Marine Science*, **4**, 473–478.

Stout J.P. (1984) *The Ecology of Irregularly Flooded Salt Marshes of the Northeastern Gulf of Mexico: A Community Profile*, Biological Report 85(7.1), U.S. Fish and Wildlife Service, Washington.

Stout J.P. (1988) Irregularly flooded salt marshes of the gulf and Atlantic Coasts of the United States. In *The Ecology and Management of Wetlands*, *Ecology of Wetlands*, *Vol. 1*, Hook D.D., *et al.* others (Eds.), Timber Press: Portland, Oregon, pp. 511–525.

Summers J.E., Ratcliffe R.G. and Jackson M.B. (2000) Anoxia tolerance in the aquatic monocot *Potamogeton pectinatus*: absence of oxygen stimulates elongation in association with an usually large Pasteur effect. *Journal of Experimental Botany*, **51**, 1413–1422.

Taylor J.R., Cardamone M.A. and Mitsch W.J. (1990) Bottomland hardwood forests: their functions and values. In *Ecological Processes and Cumulative Impacts: Illustrated by Bottomland Hardwood Wetland Ecosystems*, Gosselink J.G., Lee L.C. and Muir T.A. (Eds.), Lewis Publishers: Chelsea, pp. 13–86.

(USACE) United States Army Corps of Engineers (1987) *USACE Wetland Delineation Manual*, Technical Report Y-87-1, U.S. Army Corps of Engineers Waterways Experiment Station, Washington.

Vymazal J. (1998) Introduction. In *Constructed Wetlands for Wastewater Treatment in Europe*, Vymazal J., Brix H., Cooper P.F., Green M.B. and Haberl R. (Eds.), Backhuys Publishers: Leiden, pp. 1–15.

Vymazal J. (Ed.) (2001) *Transformations of Nutrients in Natural and Constructed Wetlands*, Backhuys Publishers: Leiden.

Vymazal J., Brix H., Cooper P.F., Green M.B. and Haberl R. (Eds.) (1998) *Constructed Wetlands for Wastewater Treatment in Europe*, Backhuys Publishers: Leiden.

Wharton C.H., Kitchens W.M., Pendleton E.C. and Sipe T.W. (1982) *The Ecology of Bottomland Hardwood Swamps of the Southeast: A Community Profile*, FWS/OBS-81/37, U.S. Fish and Wildlife Service, Biological Services Program, Washington.

Wiegert R.G. and Freeman B.J. (1990) *Tidal Salt Marshes of the Southeast Atlantic Coast: A Community Profile*, Biological Report 85(7.29), U.S. Fish and Wildlife Service, Washington.

Wolverton B.C., McDonald R.C. and Duffer W.R. (1983) Microorganisms and higher plants for wastewater treatment. *Journal of Environmental Quality*, **12**, 236–242.

Zedler P. (1987) *The Ecology of Southern California Vernal Pools: A Community Profile*, Biological Report 85 (7.11), U.S. Fish and Wildlife Service, Biological Services Program, Washington.

Zedler J.B. and Callaway J.C. (1999) Tracking wetland restoration: do mitigation sites follow desired trajectories? *Restoration Ecology*, **7**, 69–73.

Zelazny J. and Feierabend J.S. (Eds.) (1988) *Increasing our Wetlands Resources*, National Wildlife Federation: Washington.

108: Lake Ecosystems (Stratification and Seasonal Mixing Processes, Pelagic and Benthic Coupling)

DARREN L BADE

Center for Limnology, University of Wisconsin, Madison, WI, US

Seasonal stratification in lakes is caused by the differential heating of surface waters, causing them to become less dense and buoyant above colder, denser water. Other forces oppose stratification, including wind energy and loss of heat from the surface. Wind imparts many types of motions to lake waters, each having distinct patterns, energy, and scales of mixing. Theoretical considerations are presented in detail on the expected response of lakes to physical forcing. In addition, empirical equations are given for diverse regions of the world to predict epilimnetic and thermocline depth and timing of stratification. Lake morphometry along with climatic conditions determines the depth of epilimnetic mixing. The seasonal patterns of mixing and stratification are observed in a northern temperate lake (Sparkling Lake, Wisconsin, USA). The influence of stratification and mixing on the coupling of benthic and pelagic zones through processes of sedimentation, resuspension, eddy diffusivity, and hypolimnetic entrainment are discussed. Through these mechanisms, the hydrodynamics of lakes directly and indirectly influence the chemical and biological environment of lake ecosystems.

INTRODUCTION

Lake ecosystems are highly controlled by external physical forcing. Forces such as wind can cause mixing of water within lakes. Opposing the forces of mixing are the buoyant forces created by vertical density differences in the water column. Temperature and salinity both influence the density of water. If the buoyant force of the water exceeds the antagonistic forces causing mixing, a lake becomes stratified. Both mixing and stratification affect the distribution of nutrients, particulates, and biota within the water column of lakes.

Under conditions of stratification, the bottom waters can be in effect "sealed off" from surface water as vertical transport of dissolved substances is greatly reduced. Thus, oxygen diffuses only slowly into bottom waters, and only a small portion of nutrients released in the bottom waters diffuse into surface waters. These are just two examples where stratification limits the coupling between the benthic and pelagic areas of lakes. Mixing processes tend to increase this coupling by resupplying materials that may have been lost during stratification. Lake mixing patterns determine the strength of coupling between benthic and pelagic materials and processes. This strength of coupling may lead to differences in the levels of metabolism and secondary productivity. Most of these functional differences are manifest in the cycling of nutrients, but they also lead to variations in aquatic biota (Carpenter, 2003) (*see* **Chapter 96, Nutrient Cycling, Volume 3**; **Chapter 102, Trophic Dynamics, Volume 3**). In an extreme case, lakes that do not experience seasonal patterns of stratification cycle phosphorus in ways that are fundamentally different from those of lakes that do experience stratification.

Because physical forces are usually exerted at the lake surface, lake morphology (e.g. fetch, depth, etc.) constrains many of the mixing and stratification patterns of a lake. Another factor, regional climate, determines wind and heat fluxes, and is perhaps more interesting because of temporal dynamics. Therefore, understanding physical processes and their impact on chemical and biotic factors has timely relevance to the concern over the impact of global climate

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

change on lakes (McKnight *et al.*, 1996; Magnuson *et al.*, 2000). Changes in physical forcing of climate may impact lake ecosystem productivity (O'Reilly *et al.*, 2003) and viability of aquatic species (Shuter and Lester, 2004).

This chapter will review the conditions and properties of (i) **stratification** and (ii) **mixing**, and then consider (iii) **mixing patterns** and (iv) their influence on **benthic and pelagic coupling.** The goal is to provide a basic understanding of how physical processes within lakes affect chemical and biological processes. Understanding the physics of stratification and mixing and its relationship to pelagic-benthic coupling is needed for the conceptualization of how lake ecosystems function and respond to change.

STRATIFICATION

Vertical stratification of a lake develops because of the density differences between surface and bottom waters. These density differences can be caused by physical, chemical, or biological processes. The warming of surface waters that become buoyant above colder, deeper water represents the most common form of stratification. The introduction of water with salinities different from those of the lake can also impart stratification. Finally, the production of salts by the biological processes of decomposition can also create density differences, thereby stratifying lakes (Wetzel, 1973).

The density of water is related to temperature in a nonlinear fashion (Figure 1). The density of water is maximum at approximately 3.98 °C. Between the freezing point and 3.98 °C, density increases slightly, and above 3.98 °C density decreases. Therefore, it is possible for stratification to exist with warm water overlying cold water (for water >3.98 °C), or cold water above warmer water (for water <3.98 °C).

The density of water ρ^P(g cm^{-3}) for an applied pressure P(bar) is a function of temperature T (°C) and salinity S (grams of salt per kg of water), and has the form (Chen and Millero, 1986):

$$\rho^P = \rho^0 \left(\frac{1-P}{K}\right)^{-1} \tag{1}$$

where ρ^0 is the density at sea level

$$\begin{aligned}\rho^0 = {} & 0.9998395 + 6.7914 \times 10^{-5} T \\ & - 9.0894 \times 10^{-6} T^2 + 1.0171 \times 10^{-7} T^3 \\ & - 1.2846 \times 10^{-9} T^4 + 1.1592 \times 10^{-11} T^5 \\ & - 5.0125 \times 10^{-14} T^6 \\ & + (8.181 \times 10^{-4} - 3.85 \times 10^{-6} T + 4.96 \times 10^{-8} T^2) S\end{aligned} \tag{2}$$

and K (bar) is

$$\begin{aligned}K = {} & 19652.17 + 148.113T \\ & - 2.293T^2 + 1.256 \times 10^{-2} T^3 \\ & - 4.18 \times 10^{-5} T^4 \\ & + (3.2726 - 2.147 \times 10^{-4} T + 1.128 \times 10^{-4} T^2) P \\ & + (53.238 - 0.313T + 5.728 \times 10^{-3} P) S\end{aligned} \tag{3}$$

Because of the nonlinear relationship, a 1 °C difference in water temperature creates a larger difference in density in warmer water than in colder water, for $T > 4$ °C. This has important implications for the development of stratification. Though the buoyant forces could potentially cause stratification in either case, stratification would be more stable in the case of the warmer water. In addition, the amount of kinetic energy needed to disrupt stratification increases as the temperature gradient increases between water layers.

The largest flux of heat in lakes is through the surface, except for lakes with large volumes of inflowing or outflowing water. The heat flux is the balance of net radiation, latent heat of evaporation, and sensible heat exchange (Ragotzkie, 1978). The flux of heat H^* (W m^{-2}) creates a buoyancy flux B (m^2 s^{-3}),

$$B = \frac{\alpha g H^*}{C_p \rho_w} \tag{4}$$

where α is the coefficient of thermal expansion (°C^{-1}), g is the gravitational acceleration (m s^{-2}), C_p is the specific heat of water (J kg^{-1} °C^{-1}), and ρ_w is the density of water (kg m^{-3}). When heat is gained at the surface of a lake, potential energy is gained. Loss of heat from the surface causes potential energy to be converted to kinetic

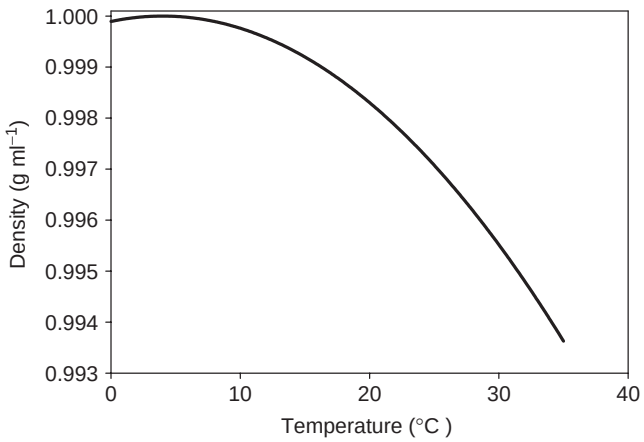


Figure 1 An example of the density change of water as a function of temperature

energy, creating convection (discussed further in section on Mixing).

The amount of work required to mix the entire water column of a stratified lake without the addition or removal of heat is the thermal stability (S).

$$S_t = \frac{g}{A_o} \int_{z_o}^{z_{max}} (\rho_z - \rho_m)(z - z_g) A_z \, dz \qquad (5)$$

where S_t = thermal stability ($J\,m^{-2}$), g = gravitational acceleration ($m\,s^{-2}$), A_o = surface area of the lake (m^{-2}), z = depth (m), z_o = surface of the lake, z_{max} = maximum depth, A_z = area at depth z, z_g = depth of the center of gravity if the water column was fully mixed, ρ_z = density ($kg\,m^{-3}$) at depth z, and ρ_m = mean density or the value of ρ at z_g (Idso, 1973). From equation (2), the importance of lake morphometry for thermal stratification becomes obvious. In order for a lake to thermally stratify, the environmental conditions must add heat to create stability. If the lake is shallow, only a small amount of energy is needed to mix the entire water column and stratification will be transient or nonexistent. The fetch of a lake (length of the lake in the direction of the wind) determines the amount of wind energy that is available for turbulent mixing. Therefore seasonal stratification in large lakes generally requires the lake to be relatively deep.

The interdependence of lake size and depth on thermal stratification was described empirically by Lathrop and Lillie (1980) and further explained by Gorham and Boyce (1989). The latter study determined theoretically a threshold depth which the lake's maximum depth z_{max} (m) must exceed for seasonal stratification to exist at the time of maximum heat content. This threshold,

$$z_{max} > 3.4 \left(\frac{\tau}{g \Delta\rho} \right)^{0.5} F^{0.5} \qquad (6)$$

shows a similar dependence with fetch F (m) as was found empirically ($z_{max} > 0.34 F^{0.5}$) for northern temperate lakes of Minnesota, Poland, and Ontario. In the above equation, the value in parentheses is related to the depth scale (m) of turbulent forces into buoyant fluid, where τ represents wind shear stress (described below), g, the acceleration due to gravity and $\Delta\rho$ the density jump between layers.

The timing of the onset of stratification in lakes is not only a function of lake morphometric characteristics, but is also dependent on local temperature. For 70 temperate and subtropical lakes of the Northern Hemisphere, Demers and Kalff (1993) empirically derived two equations to determine the onset of stratification using mean annual air temperature and the surface area to mean depth ratio (R). The first, for lakes that experience a period of ice cover, is

$$\text{Day of year} = 160 - 5.14 \times T + 5.74 \times \log(R) \qquad (7)$$

and the second, for lakes without ice cover, is

$$\text{Day of year} = 123 - 3.429 \times T + 20.636 \times \log(R) \qquad (8)$$

Some outliners from these models include lakes located in wind-protected valleys (e.g. Lake Zurich and Lake Ikeda) that attain stratification much earlier than predicted by equations (7) and (8) (Figure 2). Extremely large lakes (e.g. Lake Michigan and Lake Baikal) that experience extreme cooling before ice cover require larger inputs of heat to reach the point of thermal stratification. Finally, coastal lakes (e.g. Lake Washington) may experience greater wind stresses and less variation about the mean temperature and therefore stratify later than expected from these regressions.

The classic view of lake stratification often depicts three distinct vertical zones (Figure 3). The upper mixed layer, known as the *epilimnion*, is marked by temperatures that vary little over depth ($\delta T/\delta z \cong 0$). Below the epilimnion is a zone of transition known as the *metalimnion*. Here temperatures rapidly decrease from the warm epilimnion to the cooler bottom waters of the hypolimnion. This zone can be defined as the area where $\delta^2 T/\delta z^2$ varies from maximum to minimum. The thermocline is marked as a plane at the depth of maximum temperature change ($\delta T/\delta z \cong \min$; $\delta^2 T/\delta z^2 = 0$) (Hutchinson, 1957). Below the metalimnion lies the hypolimnion, which also shows little change in temperature with depth ($\delta T/\delta z \cong 0$). More comprehensive views of lake stratification illustrate additional complexity. First, the epilimnion is not always a well-mixed layer, as secondary stratification may develop

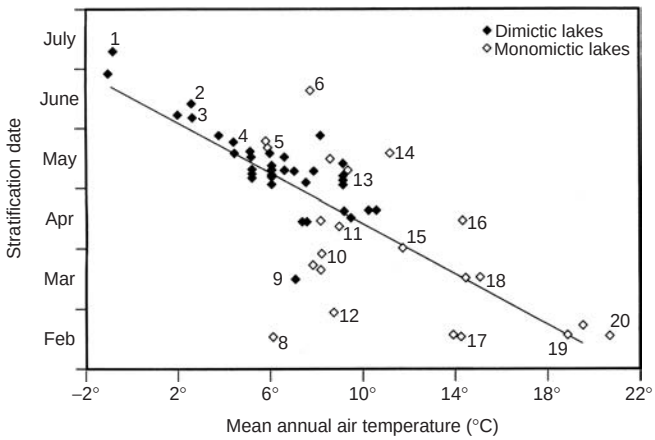


Figure 2 Relationship between mean annual air temperature and date of onset of stratification for 70 north temperate zone lakes. Identified lakes referenced in the text: 1)Baikal, Russia; 6) Michigan, USA; 8)Ikeda, Japan; 12)Zurich, Switzerland; 14) Washington, USA. (Kalff, J. Limnology, 1st Edition, ©2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ)

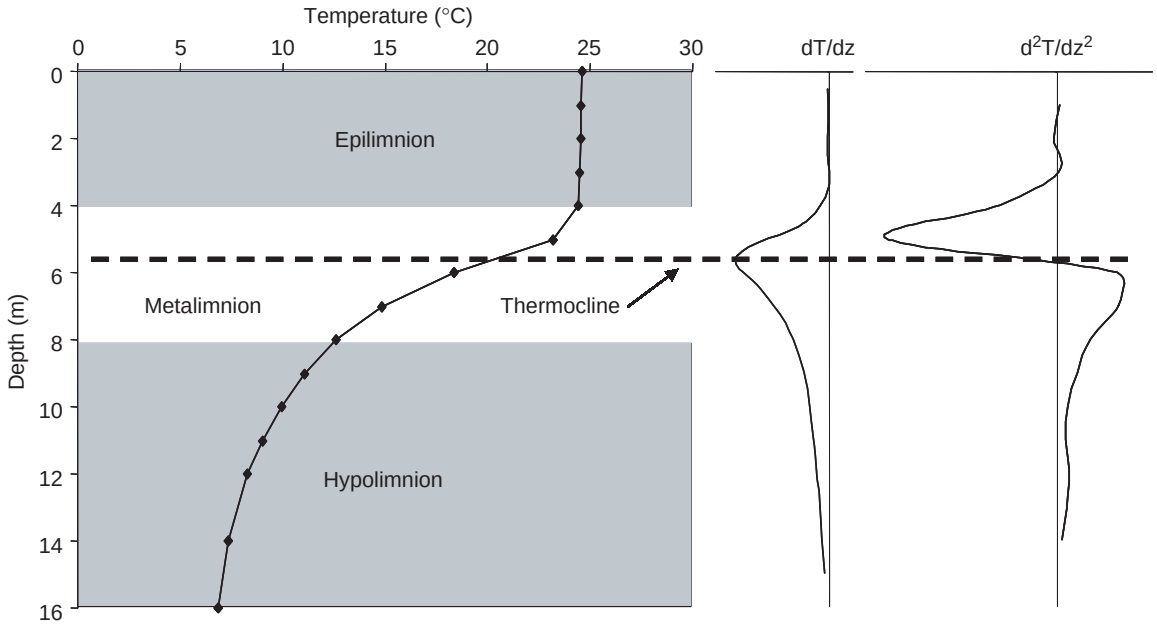


Figure 3 Typical midsummer temperature profile of Sparkling Lake, Wisconsin. Profiles of dT/dz and d^2T/dz^2 are used to determine the thermocline and extent of the metalimnion

in this zone. Diurnal thermoclines may develop with daytime heating and be broken down during evening cooling. Several thermoclines may exist above the deeper seasonal thermocline (Imberger, 1985). The hypolimnion, often thought of as quiescent, contains areas of mixing near the lake bottom. The turbulence in the benthic boundary layer creates important sites for flux of materials (Eckert *et al.*, 2002).

Not all lakes develop these three distinct zonations. Some lakes, where the ratio of epilimnion depth to maximum depth is between 0.5 and 1, may only have an epilimnion and metalimnion. Few lakes with stable stratification have been observed with thermocline depths greater than 70% of the maximum depth (Patalas, 1984; Gorham and Boyce, 1989). Turbulence caused in the benthic boundary layers inhibits the formation of a stable hypolimnion when the mixed layers become this deep.

Because the physical separation of water masses, which occurs when lakes stratify, is important to lake ecosystem function, numerous studies have been conducted to determine the depth of the epilimnion or the thermocline. Hanna (1990) presents an evaluation of many empirical models created for lakes from various temperate parts of the world. Most of these models follow some form of power law relating fetch or area to epilimnion or thermocline depth. In general, they explain a reasonable amount of variation in the response variable. Although these models yield a good first approximation to the mixing depth of many lakes, they should be used with caution, preferably within the limits of the geographical and morphological constraints of the source data. One notable exception to the general relationships is the relatively shallow mixing depths of the Laurentian Great Lakes, given their large surface area. As the Coriolis effect has greater relative influence on water movements in large lakes (see next section on Mixing), mixing by internal seiches may be dampened (Gorham and Boyce, 1989). The large lakes of Africa follow more closely a power law relationship between fetch and mixing depth (Kling, 1988). These lakes are less affected by Coriolis because of their proximity to the equator.

Much attention in limnological research is focused on temperate lakes of the Northern Hemisphere. Lakes in subtropical and tropical regions, however, do not fit the general patterns of temperate lakes. In addition, temperate lakes of the Southern Hemisphere are more influenced by coastal patterns because large landmasses that contribute to the continental climate in the Northern Hemisphere are far larger than in the Southern Hemisphere. Table 1 presents models predicting thermocline depth from diverse regions of the world.

Since thermal stratification is dependent upon heat absorption, differences in the vertical attenuation coefficient of solar radiant energy K_E (Kirk, 1983) should influence thermocline depths (Hocking and Straškraba, 1999). Turbidity, biological particulates, and dissolved colored material (e.g. colored dissolved organic carbon) increase K_E and therefore significantly influence the mixed layer depth (Mazumder and Taylor, 1994; Fee *et al.*, 1996; Xenopoulos and Schindler, 2001). Increased water clarity results in increased thermocline depth, after the effect of area

Table 1 Models predicting thermocline depth or mixed layer/epilimnion depth (meters) for lakes in different regions. Thermocline depth (Z_t), epilimnion depth (E_d), fetch (F), area (A), maximum effective length (MEL), maximum effective width (MEW), effective length (EL), maximum depth (Z_{max}), volume (V), maximum length (ML), maximum width (MW). Expanded from Hanna (1990)

Region	Model	R^2	n	units	Author (year)
Alaska	$Z_t = 1.14(F) + 5.05$	0.5	26	km	Edmundson and Mazumder (2002)
Argentina	$Z_t = 5.95(F)^{0.5} + 12.93$	0.82	26	km	Baigún and Marinone (1995)
	$Z_t = 1.60(A)^{0.5} + 23.68$	0.81	26	km^2	
Cameroon	$Z_t = 9.94(F)^{0.300}$	0.83	52	km	Kling (1988)
	$\log(Z_t) = 0.168\log(A) + 0.996$	0.72	52	km^2	
East Germany/ Baltic Lowland	$E_d = 4.72(EL)^{0.39}$		30	Km	Ventz (1973)
	$EL = (MEL + MEW)/2$				
ELA	$E_d = 0.12(F)^{0.5}$	0.71	20	M	Cruikshank (1984)
	$F = MEL$				
	$E_d = 1.54(A)^{0.28}$	0.64	20	km^2	
	$E_d = 0.69(Z_{max})^{0.56}$	0.48	20	m	
	$E_d = 1.96(V)^{0.22}$	0.72	20	10^5 m^3	
Japan	$E_d = 3.56(ML)^{0.46}$	0.48	19	km	Straškraba (1980)
Japan	$E_d = 6(F)^{0.33}$		32	km	Arai (1964)
	$F = (A)^{0.5}$				
Japan	$E_d = 6.22(F)^{0.304}$	0.53	36	km	Arai (1981)
	$F = (A)^{0.5}$				
New Zealand	$E_d = 7.00(MEL)^{0.42}$	0.79	33	km	Green *et al.* (1987)
New Zealand	$E_d = 7.69(F)^{0.463}$	0.94	22	km	Davies-Colley (1988)
	$F = (A)^{0.5}$				
	$E_d = 6.85(F)^{0.446}$	0.92	22	km	
	$F = (ML + MW)/2$				
	$Z_t = 9.52(F)^{0.425}$	0.95	22	km	
	$F = (A)^{0.5}$				
	$Z_t = 8.58(F)^{0.408}$	0.93	22	km	
	$F = (ML + MW)/2$				
North America	$Z_t = 0.298\ln(F) + 1.82$	0.66	73	km	Shuter *et al.*(1983)
	$F = MEL$				
Poland/Baltic Lowland	$E_d = 4.55(EL)^{0.455}$	0.88	53	km	Straškraba (1980)
	$EL = (MEL + MEW)/2$				Patalas (1960, 1961)
Northern Poland	$E_d = 4.4(F)^{0.5}$	0.88	53	km	Patalas (1961)
	$F = (MEL + MEW)/2$				
Poland, Central Canada	$E_d = 4.6(F)^{0.41}$	0.85	88	km	Patalas (1984)
Yukon–Yellowknife	$E_d = (F)^{0.41}$		23	km	Patalas (1984)
TCMA, Minnesota	$E_d = 3.02(A)^{0.5} + 1.108$	0.98	8	km^2	Osgood (1988)
Temperate lakes	$\log(Z_t) = 0.336\log(MEL) - 0.245$		123	km	Hanna (1990)
Wisconsin, Central Canada	$E_d = 4(F)^{0.5}$		18	km	Ragotzkie (1978)
Worldwide	$Z_t = \ln((A)^{0.5}/43)^{2.35}$	0.66	150	m^2	Gorham and Boyce (1980)

has been accounted for by regression (Figure 4). Although variable, the pattern of decreasing mixed layer depth with increasing K_E is most prevalent in smaller lakes where wind mixing is less important (Fee *et al.*, 1996). Similar responses in thermocline depth have been noted in individual lakes that experienced changes in water clarity due to experimental or management manipulations (Edmundson and Lehman, 1981; Yan, 1983; Fee *et al.*, 1996).

Variables that represent geography, morphometry, and water clarity can predict thermocline depths in lakes. Generally, however, 20–30% of variance is unexplained in these models, likely attributable to local climatic conditions. The variation due to climate does not affect all lakes equally. For example, the thermal structure of clear lakes during abnormally warm years changed, while colored lakes tended to be more stable (Snucins and Gunn, 2000).

Apart from empirically based models, process-based models are often used to simulate the lake thermal structure. These models require the input of extensive meteorological data, but in turn can simulate vertical zonations of temperatures quite accurately with little or no parameterization. DYRESM (Dynamic Reservoir Simulation Model; **`http://www2.cwr.uwa.edu.au/~ttfadmin/model/dyresm1d/`**) is one example of this type of model (Imberger *et al.*, 1978; Spigel and Imberger, 1980). DYRESM is a one-dimensional numerical model that accounts for the budget of turbulent kinetic energy in the surface mixed layer to determine the depth of mixing that is countered

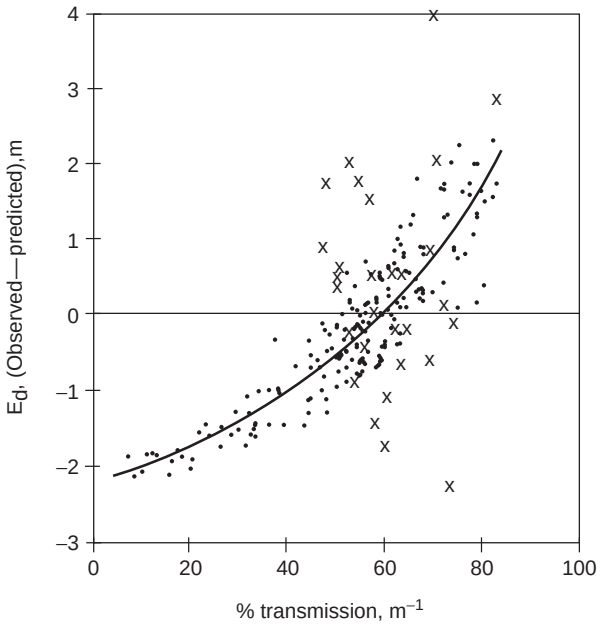


Figure 4 The residuals of a linear regression of lake size on mixing depth versus percent transmission of light within the lake. Lakes <500 ha – •; larger lakes – x. The curve is a least squares fit for lakes <500 ha [$-3 + 0.80$ * $\exp(0.022 T_{\%})$, $r^2 = 0.87$, $n = 186$ (Reproduced from Fee *et al.*. 1996; © 1996 by the American Society of Limnology and Oceanography, Inc.)

by buoyant forces. A recent update has created a pseudo two-dimensional model that accounts for internal and benthic boundary-layer mixing and has been tested on lakes ranging in size from 3.6 to 4800 ha (Yeates and Imberger, 2003). DYRESM often forms the basis for other water quality models (Hamilton and Schladow, 1997). The model is applicable to small to medium size lakes and reservoirs because these systems are more amenable to the assumptions necessary for one-dimensional modeling.

MIXING

The two main forces that contribute to the physical mixing of lakes are wind and the loss of heat from the surface. These forces impart kinetic energy and create several types of organized water movements. Over time, length scales of these water movements cascade into smaller and smaller eddies until they become turbulent kinetic energy (TKE) and dissipate into heat due to viscous forces. Whereas advective forces may transport materials over great distances in lakes, TKE is responsible for the mixing that creates spatial homogeneity. A large body of literature exists on the theory of movements and mixing of stratified lake waters. Some key reviews on this topic are given by Hutchinson (1957), Mortimer (1974), Csanady (1975), Hutter (1983), Imberger and Patterson (1990), Imberger (1994), Imboden and Wüest (1995), Imberger (1998), and Imboden (2004).

The flow of water, for example, in a pipe or in the mixed layer of a lake, ranges from laminar to turbulent. When flow is laminar, parcels of water move past one another in an orderly, smooth fashion, held together by the viscous property of the fluid. When flow becomes turbulent, the viscous forces are overcome, vortices form, and flow becomes chaotic or random. The nondimensional Reynolds number, *Re*, defines the interaction of the inertial forces of flowing water and the viscous property of water as

$$Re = \frac{UL}{\eta} \tag{9}$$

where U is the mean current velocity ($m\,s^{-1}$), L is the depth or thickness of the layer (m), and η is the kinematic viscosity of water ($m^2\ s^{-1}$). Flow is laminar at $Re < 500$ and turbulent at $Re > 2000$. Although flow is rarely laminar in aquatic systems, during calm conditions the scale of the turbulent vortices is small.

Two fluids of differing density moving across one another with differing velocities have the potential to create shear stress. The force associated with shear stress is perpendicular to the direction of flow of the fluid and creates turbulent vortices. When a lake is stratified either thermally or chemically, the potential energy of the stratification counters the kinetic energy that creates shear. The ratio of these two forces, the Richardson number (Ri), can be used to indicate when flow becomes unstable between the two layers, causing mixing to occur. For the mixed layer, the Richardson number is defined as

$$Ri = \frac{g(\mathrm{d}\rho/\mathrm{d}z)}{\overline{\rho}(\mathrm{d}u/\mathrm{d}z)^2} \tag{10}$$

In equation (10), buoyancy is represented by the buoyancy frequency, $N(s^{-1})$, the timescale for a displaced particle of differing density to return to its original position due to the force of gravity:

$$N^2 = \frac{g(\mathrm{d}\rho/\mathrm{d}z)}{\overline{\rho}} \tag{11}$$

where g is the acceleration due to gravity ($cm\,s^{-2}$), $\bar{\rho}$ is the mean density ($kg\,m^{-3}$) and ($d\rho/dz$) is the density change across the fluid of depth z (m). The other terms represents the shear forces created by flow, where u is the horizontal current velocity ($m\,s^{-1}$). For $Ri < 0.25$, flow becomes unstable and mixing occurs between the two layers. For $Ri > 0.25$, flow remains stable and the two layers of water flow past each other without the formation of vortices or mixing.

Wind-induced turbulence is generally the dominant force that creates mixing in larger lakes, except perhaps during

periods of low wind speeds, when convectively induced motion may be important. Wind causes two types of motions in lakes, those with periodicity such as waves, and those lacking periodicity such as currents. Wind associated with storm events imparts the most energy to lakes and is responsible for a large amount of mixing that occurs in lakes. Storm related winds cause surface waves and currents, as well as internal waves in stratified lakes.

The movement of wind across the surface of a lake conveys energy to the underlying water due to friction. The wind stress τ_w ($N\,m^{-2}$), which transfers energy from the wind to the water, is defined as:

$$\tau_w = \rho_a C_{10} (U_{10})^2 \tag{12}$$

where ρ_a is the density of air ($kg\,m^{-3}$), C_{10} is the drag coefficient, and U_{10} is the wind speed ($m\,s^{-1}$) at a height 10 m above the surface. The frictional velocity (u_* ($m\,s^{-1}$) of the water associated with this wind stress is

$$u_* = \left(\frac{\tau_w}{\rho_w}\right)^{0.5} \tag{13}$$

where ρ_w is the density of water ($kg\,m^{-3}$). Since the drag coefficient is a measure of the roughness of the water surface, the wind stress is not only a function of wind velocity but also of the wave field development. Therefore, the drag coefficient also varies with wind speed (Wüest and Lorke, 2003). Wind stress not only creates surface waves, but a small portion of the energy (1–4%) also goes into large-scale water motions (Smith, 1979)

Surface waves are the most conspicuous water motion observed on lakes during windy periods. However, surface waves transfer only moderate amounts of TKE, except when they break near the shore. Short surface waves are periodic and are characterized by wave height (H), amplitude (A), length (λ), period (T) and frequency (f) (Figure 5). Wave period is the time (s) for two consecutive wave crests (or troughs) to pass a fixed point. Wave frequency is the reciprocal of wave period. Waves of $\lambda < 1.8$ cm are generally considered ripples, since surface tension has greater influence on the wave than gravity

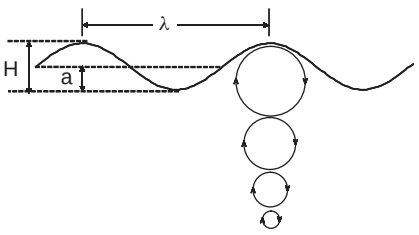


Figure 5 Diagram of short surface gravity waves. Symbols are wavelength (λ), wave height (H) and wave amplitude (a)

(Hutchinson, 1957). Waves of $\lambda > 1.8$ cm are referred to as *gravity* waves. Heights of gravity waves may be greater than 1 m during sustained winds over large lakes. The maximum wave height H_{max}, (cm) obtainable in lakes has been empirically related to the square root of the fetch F (cm) (Hutchinson, 1957):

$$H_{max} = 0.105\sqrt{F} \tag{14}$$

This relationship, however, gives no indication of the influence of bottom effects on wave height. In deep water, wavelengths are approximately 20 times the wave height. Movement caused by surface gravity waves in deep water is cycloidal. These movements decay with depth at approximately half the cycloid diameter for every increase in depth of $\lambda/9$ (Wetzel, 2001). Therefore, the motion induced by surface waves declines rapidly with depth, and the depth to which surface waters are mixed by surface wave movements is usually considered to be about 0.5 λ (Smith and Sinclair, 1972). Small distortion of the orbital particle motion of surface waves causes a small net movement of water in the direction of the wave.

In shallower waters the motion caused by surface waves begins to interact with the lake bottom. This causes the particle motions to become more elliptical as the vertical motion is suppressed and the horizontal motion is increased. In addition, the ratio of wave height to wavelength ($H{:}\lambda$) increases, causing waves to become less symmetrical, and they begin to break (Hutchinson, 1957).

Wind also imparts surface drift or large-scale currents to lake waters. The speed of water current is generally much reduced compared with the wind speed because of the difference in densities between air and water. The direction of water currents usually matches the wind direction until the currents encounter the lake boundaries or opposing topography within the lake. In large lakes, the earth's rotation (Coriolis force) causes currents to deflect to the right of the wind in the Northern Hemisphere and to the left in the Southern Hemisphere. The Coriolis force is strongest at the poles and zero at the equator.

The Eckman spiral (Figure 6) is created by the interaction of current Coriolis deflection and changing current speed over depth. The deflection due to Coriolis forces continues with depth, but currents decrease exponentially until the point of no net flow. Below this there is generally a return flow that is much slower than the surface flow. These surface currents have a large influence on the horizontal advective stirring in lakes.

Langmuir (1938) currents are another surface water movement that has a large impact on distributing materials. Langmuir cells are helical currents, parallel with the wind, which are created from an interaction between surface waves and surface currents (Plueddemann *et al.*, 1996)

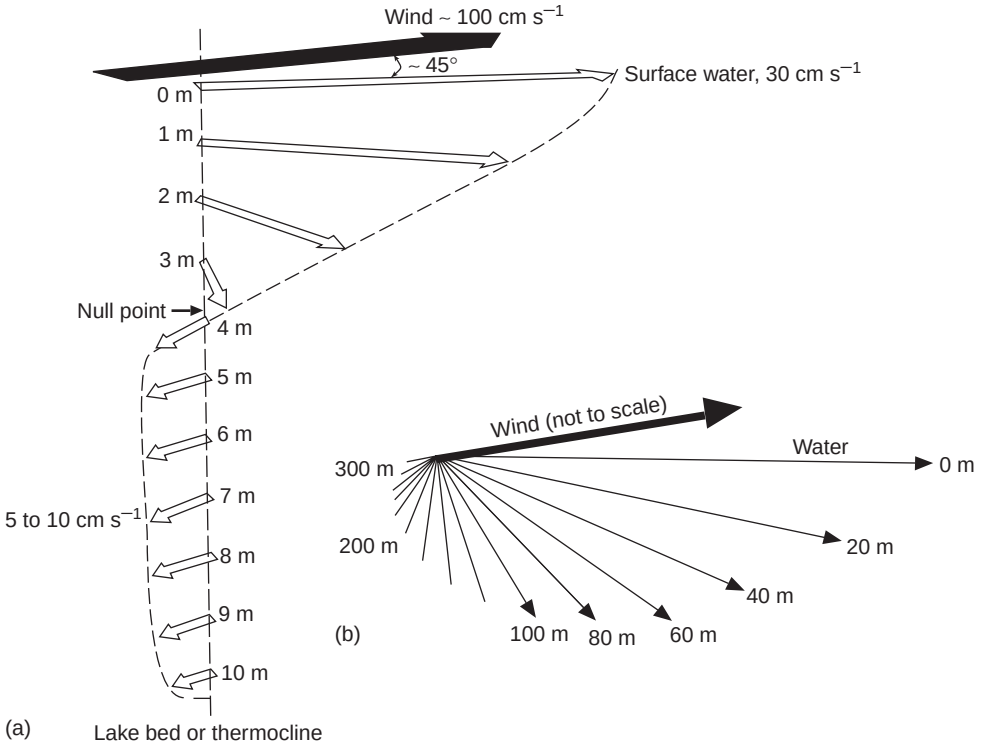


Figure 6 (a) Idealized and (b) measured diagrams of an Eckman spiral. The measured values are from the Pacific Ocean where depths are much deeper than what would be found in lakes (Reproduced from Horne and Goldman, 1994 by permission of McGraw Hill Education)

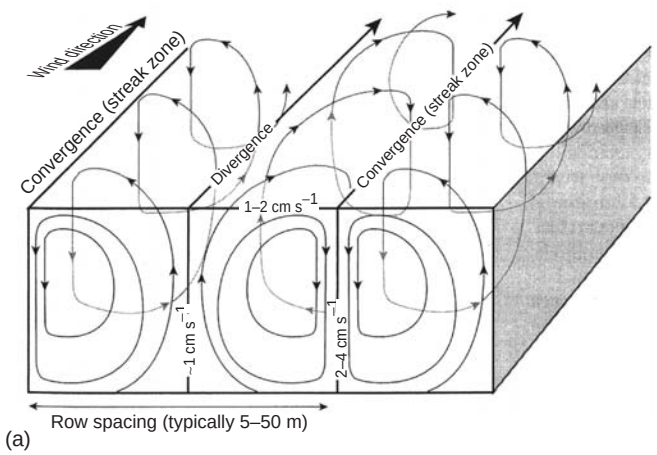
(Figure 7(a)). This pattern of Langmuir cells is observable (at wind speed greater than $3-4\,\mathrm{m\,s^{-1}}$) as bands or windrows of foam and other particulates that accumulate in areas of downwelling (Figure 7(b)). The vertical currents of Langmuir cells are strongest and most concentrated in the downwelling zone and weaker and more diffuse in the upwelling areas (Imboden and Wüest, 1995). Under conditions of strong winds, Langmuir currents may penetrate to the depth of the thermocline. These currents can have considerable influence on the distribution of detritus and planktonic organisms.

Another surface phenomenon is the formation of surface standing waves or surface seiches (barotropic waves). As with surface gravity waves, the turbulent kinetic energy in these waves also is low. These waves are usually set up, as sustained winds blow in a steady direction causing water to accumulate on the leeward side of the lake. At the same time, there is a decrease in water depth on the windward side. As this occurs, the water picks up momentum and overshoots the equilibrium tilt that can be sustained by the force of the wind. To regain equilibrium, the water level now increases on the windward side and decreases on the leeward edge. This rocking motion is usually centered on one or more nodes that move little in the vertical direction, but exhibit the greatest horizontal velocities. The antinodes experience little horizontal water movement but maximum vertical displacement. Length and mean depth influence the period of these waves (in lakes where length greatly exceeds mean depth) such that:

$$T_s = \frac{1}{n}\frac{2L}{\sqrt{g\bar{z}}} \tag{15}$$

where T_s is the period (s), L is the basin length (m), g is gravitational acceleration ($\mathrm{m\,s^{-2}}$) and $\bar{z}$ is the mean depth (m). Equation (15) is used for lake basins approximating a rectangular shape. The period can also be estimated for alternative basin shapes (Hutchinson, 1957).

Surface seiches are responsible for creating internal seiches (baroclinic waves). As surface waters accumulate on the leeward side of the lake, the metalimnetic waters are depressed to compensate for the increase in pressure. Similar to a surface seiche, the internal seiche is set into motion once the wind eases, and the water level oscillates towards its mean level. Internal seiches are much more important for mixing than surfaces seiches because they attain greater amplitudes than surface standing waves. Internal seiches may have amplitudes of several meters. Theoretically, the potential energy stored in the internal seiche exceeds the potential energy of the surface seiche



(a)

(b)

Figure 7 (a) Diagram of Langmuir currents (Kalff, J. Limnology, 1st Edition, ©2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.) that create the noticeable surface streaks or wind rows seen (b) on Lake Mendota. (photo: Center for Limnology, Madison, WI). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

by $(\Delta\rho/\rho_0)^{-1}$, where $\Delta\rho$ is the change in density between the surface mixed layer and the base of the metalimnion (ρ_0) (Imboden, 2004). However, in natural conditions, internal seiches may not reach this theoretical bound, as morphometric characteristics tend to decrease the actual magnitude of the seiche. Like surface seiches, internal seiches are periodic and the period (T_i) can be calculated as

$$T_\mathrm{i} = 2L\left[\frac{g' z_\mathrm{e} z_\mathrm{h}}{(z_\mathrm{e} + z_\mathrm{h})}\right]^{-0.5} \tag{16}$$

where L is the length of the basin, z_e and z_h are the depth of the epilimnion and metalimnion. The reduced gravity, $g' = g\Delta\rho/\rho_0$, is the reduced gravitational acceleration due to the change in density (Spigel and Imberger, 1980).

In large lakes (minimum fetch >2–3 km), internal seiches are influenced by Coriolis forces (Mortimer, 1974). The earth's rotation should significantly influence the motion of the internal seiche if the period of the internal wave is less than the inertial period, ($T_\mathrm{i} < T_\Omega$) (Csanady, 1975), where

$$T_\Omega = \frac{2\pi}{f} \text{ and } f = 2\Omega\sin\phi \tag{17}$$

The angular frequency of earth's rotation Ω is 7.29×10^{-5} s^{-1}, and ϕ is the geographical latitude. The simplest modification created by Coriolis forces imparts a counterclockwise (in the N. Hemisphere) motion to the seiche. The seiche pivots about a single node and is termed a *Kelvin wave*. The amplitude of Kelvin waves decreases exponentially with distance from the shore. This motion creates coastal jets that run parallel to the shoreline and are important for distributing materials that enter near the coastline. In very large lakes where internal waves are influenced by Coriolis forces but not impacted significantly by coastal morphometry, *Poincaré waves* are generated. These waves can be visualized as alternating cells of hills and valleys that undulate and rotate clockwise (Mortimer, 1974). These waves do not decrease in amplitude with distance from the shore and therefore can impart large motions to metalimnetic waters.

All of the above phenomena contain kinetic energy that cascades into TKE. The cascade of energy into TKE may be different for the different motions. One mechanism for the formation of turbulent kinetic energy is shear created by currents of different densities flowing over one another. Shear can create billows that may become unstable and break. This process is referred to as *Kelvin–Helmholtz* instability and these billows are areas of increased turbulent mixing that incorporate denser water into the surface mixed layer (Mortimer, 1974).

Because internal seiches can have large amplitudes (and can contain a large amount of energy) they are extremely important for mixing and the entrainment of deeper waters into the epilimnion. To examine the response of the internal seiche to a wind event, the Richardson number (equation 10) can be modified to characterize the mixed layer as a whole, and include episodic, lake-specific mixing patterns. The modification is termed the *Wedderburn number* (W_e)(Imberger and Hamblin, 1982),

$$W_\mathrm{e} = \frac{g' h^2}{u_*^2 L} \tag{18}$$

where g' is the reduced gravitational acceleration, h is the depth of the mixed layer, u_* is the surface shear velocity (equation 13) and L is the fetch of the metalimnion in the direction of the wind. When W_e is small ($W_\mathrm{e} < 1$), the wind energy is large compared to the strength of stratification, and stratification may break down. When W_e is large,

stratification remains stable (Spigel and Imberger, 1980). A similar term that applies to the entire lake and may have a more mechanistic applicability is the lake number L_N. This dimensionless number incorporates the Schmidt stability index (equation 5) as well as forces other than wind that can disrupt stability (Imberger and Patterson, 1990).

There are a number of mechanisms caused by the movement of internal seiches that can contribute to turbulent mixing. Horn *et al.* (2001) put forth five possible regimes that might exist during the degeneration of internal seiches. These regimes are classified by W_e and the ratio of the mixed layer depth to the maximum depth. Two regimes were observed frequently in lakes. The most common regime observed involves the nonlinear steepening of the internal seiche, forming groups of solitary waves (solitons). The other observed regime was the dampening of the internal seiche before the formation of solitons could occur. Conditions of supercritical flow, Kelvin–Helmholtz billows, or the formation of bores and billows were not observed. The solitary waves transfer their energy to turbulence as they propagate and break on the sloping lake bottom (Imberger, 1998). Therefore, turbulent kinetic energy generated from internal seiches is most prevalent where the thermocline intersects the side of the lake. Combining the pathways by which TKE is produced shows that TKE is most prevalent in the surface mixed layer and at the sides and bottom of the lake, whereas turbulence is weak in the interior of stratified lakes (Wüest and Lorke, 2003).

Cooling, or in the case of saline waters, evaporation, causes surface water to become denser than underlying water (*see* **Chapter 44, Evaporation from Lakes, Volume 1**). This produces convection currents, as the denser water sinks. In small lakes, or large lakes that experience periods of low wind, penetrative convection may be the dominant source of kinetic energy and mixing. The buoyancy flux (B in equation 4) determines the penetrative convection velocity ($m\,s^{-1}$), $w_* = (Bh)^{1/3}$, where h is the depth of the mixed layer. The convection velocity can be compared to wind-generated shear velocity, by examining the ratio $(u_*/w_*)^3$. The depth at which convection and surface wind shear contribute equally is the *Monin–Obukhov length* L_M (m):

$$L_M = \frac{u_*^3}{B} \tag{19}$$

For a negative heat flux, the ratio of $L_M : h$ determines the dominant source of shear at the mixed layer. If $L_M : h > 1$, wind shear is dominant. If $L_M : h < 1$, the ratio is the relative depth of the mixed layer, influenced by u_*, and below L_M, convection dominates. For a positive heat flux, L_M describes the depth at which wind mixing is balanced by the generation of potential energy.

Whereas the onset of thermal stratification requires periods of calm and positive heat flux, the destruction of thermal stratification takes place during periods of cooling and episodes of high wind. As surface waters cool and turbulent forces cause the surface waters to mix into the metalimnion, a deeper thermal stratification will develop. When the density difference between the epilimnion and hypolimnion become small enough that buoyancy is overcome by turbulence, stratification breaks down completely. Empirically, the timing of autumnal mixis in stably stratified lakes depends upon several factors, such as hypolimnetic temperature, mean depth, and latitude adjusted for altitude (Nürnberg, 1988). Hypolimnetic temperature explains the largest portion of the variation in date of turnover for lakes located in similar regions; lakes with colder hypolimnetic temperatures turning over later than lakes with warm hypolimnia. Increasing mean depth also leads to later turnover, and finally, an increase in adjusted latitude causes lakes to mix earlier. These factors are likely related to the water column stability, the mass of water that must be cooled and mixed, and the local climatic conditions, respectively. Interannual variation in climate may be the cause of some of the unexplained variation in these empirical models (Nürnberg, 1988).

PATTERNS OF STRATIFICATION AND MIXING

Not all lakes may experience the same patterns of stratification and mixing due to geographic location or lake morphometry. Schemes for classifying lakes based on stratification patterns have been constructed since the early periods of limnology. A review of these early schemes is provided by Hutchinson (1957). There have been many modifications of these schemes. For example, Lewis (1983) suggests eight patterns of seasonal stratification and mixing that can be found in lakes throughout the world. These patterns are: amictic, cold monomictic, continuous warm and cold polymictic, discontinuous warm and cold polymictic, dimictic, and warm monomictic. Lakes that do not mix completely on a seasonal basis are meromictic lakes.

Amictic

Amictic lakes are perennially ice-covered and therefore are not subject to the direct effects of seasonal temperature variations or wind. Although the name implies that these lakes never mix, inflows from streams or melting ice can add heat or turbulence to the lake. If the ice cover is clear, solar radiation may penetrate to the water or sediments below and cause convective currents. Amictic lakes are found mainly in the Antarctic, although examples are also found in the Northern Hemisphere.

Cold Monomictic

Lakes with one period of circulation, and temperatures that do not reach much greater than 4 °C are termed

cold monomictic. These lakes are generally located at high latitude and are ice-covered for most of the year. Only during a brief period of no ice does complete mixing occur. Because there is little net gain of heat during this period, wind exposure generally does not allow periods of stratification during the ice-free season. However, regional climatic variation can create years when stratification during the ice-free season does occur, and two periods of circulation (dimictic; see below) are present.

Continuous Cold and Warm Polymictic

These lakes have relatively modest depth, or are wind exposed. Cold polymictic lakes rarely reach temperatures greater than 4 °C, but thermal inputs can be sufficient on calm, sunny days to create brief (up to several days) periods of stratification. Warm polymictic lakes reach temperatures greater than 4 °C, but because of excessive winds or large cooling events, stratify only intermittently.

Discontinuous Warm and Cold Polymictic

Lakes that grade from polymictic to dimictic because of increasing duration of stratified intervals are classified as discontinuous.

Dimictic

Many lakes in the temperate regions that have an ice-covered period fall under this category. There are two periods of mixing, one in the spring after the ice has thawed, and another in the fall after the summer thermal stratification has broken down.

Warm Monomictic

Lakes that do not experience ice cover but stratify stably only once annually are termed *warm monomictic*. Compared to their polymictic counterparts, these lakes are generally deeper and have a greater heat storage capacity, preventing intermittent mixing.

Meromictic Lakes

Lakes that experience mixing throughout the entire water column are termed *holomictic*. In certain lakes, the mixing patterns described above are interrupted by a chemocline that cannot be overcome by forces of wind mixing. These perennially stratified lakes are defined as *meromictic*. Meromictic lakes contain an unmixed bottom layer, the monimolimnion, and an upper layer, the mixolimnion, which may experience patterns of thermal stratification. There are generally two situations that result in meromixis. The first, ectogenic meromixis, is brought on by the intrusion of saltwater into a freshwater lake, or *vice versa*. Secondly, biogenic or endogenic meromixis is generally created when decomposition products in the monimolimnion create water of greater density. Many of the deep, highly productive equatorial lakes are meromictic (Wetzel 2001). Lakes may experience periods of intermittent meromixis if conditions prevent full mixing of salts accumulated in the hypolimnion over the stratified period. This allows continued increase in bottom water densities, requiring even greater mixing forces to disrupt the chemocline and mix the entire lake.

An Example of Stratification in a North Temperate Dimictic lake

Because a large number of lakes are in the temperate zone and fall under the category of dimictic, it is worth discussing the annual patterns of stratification and mixing experienced by these lakes. In addition, these lakes experience nearly the entire range of thermal conditions that exist in lakes. Therefore, the patterns of other lakes can quite easily be inferred from the behavior of dimictic lakes. Sparkling Lake is located in the Northern Highland Lake District of Wisconsin, USA. (Lenters *et al.*, 2005) (Figure 8). The lake has a surface area of 64 ha, a maximum depth of 20 m and a mean depth of 10.9 m. Ice covers Sparkling Lake for an average of 141 days each year. Daily average temperatures were recorded with a thermistor chain suspended from a buoy located centrally in the lake. Temperature was recorded at a depth of 0.1 m, then every meter from 1 to 12 m, and finally at 14 and 16 m.

At the time of spring ice melt, lake water is isothermal near the temperature of maximum density (~4 °C; the lake was ice free on 24 April 2002). After ice-off, the surface water is mainly heated by the absorption of solar energy. The thermal resistance to mixing at this time is insufficient to counteract the turbulent mixing created by the force of wind. This creates a period of gradual warming and mixing of the entire water body. Cooling at night creates convection currents that aid in creating an isothermal period of spring mixis. In Sparkling Lake, this period lasted until about 25 May 2002, at which point the bottom of the lake had warmed to a temperature of 6.5 °C.

As the temperature of the entire water body warms, further increases in the surface water temperature cause this water to become increasingly resistant to the forces of mixing with cooler water below. This is due to the nonlinear relationship between water density and temperature increases above 4 °C. In Sparkling Lake, intermittent stratification exists from 25 May until 6 June. After a relatively calm, warm period, the surface warms considerably, buoyancy forces resist further wind mixing of the entire water column, and stable summer stratification sets in. In deep lakes with small surface areas that experience ice cover late

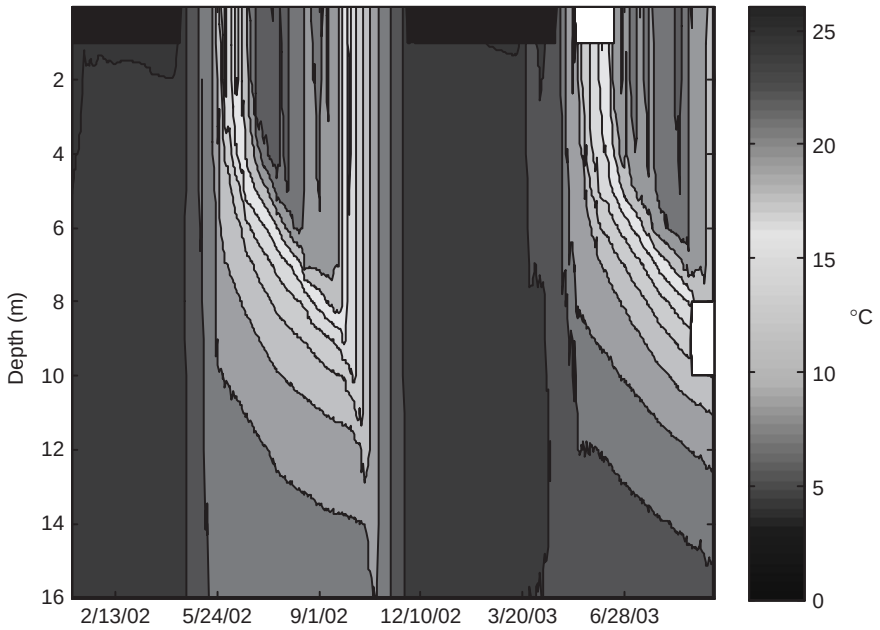


Figure 8 Depth-time diagram of isotherms in Sparkling Lake, Wisconsin. Each isotherm represents a 2 °C change in temperature. Black areas are ice cover (not to scale) and white areas are no data. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

into the spring, heat flux to the lake may be great enough that stratification develops before complete mixing occurs. In Sparkling Lake, the period of mixing is short enough that the deepest waters warm only a few degrees before stratification. In moderately deep lakes (>10 m), stratification is generally stable for the entire summer. Strong wind events and net loss of heat later in the summer season can deepen the thermocline, but stratification generally does not break down. Periods of calm may create secondary thermoclines nearer the surface, but buoyant forces are generally not strong enough for secondary thermoclines to persist.

A change to net negative heat flux in late summer and autumn gradually cools the mixed layer. This aids in eroding the thermocline due to a decrease in the buoyant forces. Generally the entire water body does not mix until the surface water has nearly reached the temperature of the hypolimnion. In Sparking Lake, the hypolimnion gradually warms throughout the entire summer reaching a maximum of about 8 °C (10 October 2002), at which point the epilimnion has cooled to a similar value and the lake is isothermal. The lake then experiences a period of autumnal mixis during which the water temperature continues to fall into the winter months.

Although water reaches its density peak at ~4 °C, cooling beyond this point does not create stable stratification of the less dense colder water above the 4 °C isotherm. This is because the density differences are small at low water temperatures and the resistance to mixing is also low. The temperature at which ice begins to form on water bodies is partly a function of the fetch (F) of the lake (for F (km) <9, $T(^{\circ}\mathrm{C}) = 1.36F^{-0.49}$; Kalff, 2002). Ice can form on small lakes, such as Sparkling, when the temperature of most of the water is between 3 and 4 °C. Large lakes continue to cool below this point and often do not freeze until the water temperature is below 1 °C. After ice formation, the lake is protected from the forces of the wind, and an inverse thermal stratification is set up due to a temperature gradient from ~0 °C very near the ice to higher temperatures of up to 4 °C near the bottom. Inputs of heat from sediments, surface, or groundwater, as well as solar radiation passing through the ice, can create currents and mixing during periods of ice cover (e.g. Likens and Hasler, 1962). As spring approaches again, a large amount of the lake's heat budget goes to melting the ice. This is an important distinction when comparing lakes that do not undergo a period of ice cover (e.g. warm monomictic lakes), because this heat would otherwise be used to establish the next thermal stratification event. In early spring (~14 April 2003), the ice on Sparkling Lake was abnormally clear and allowed for solar heating of the water beneath the ice. Interestingly, the temperature at 1 m reached a daily average high of nearly 6 °C on 15 April, and temperatures of greater than 5 °C extended several meters deep. The ice did not recede until 24 April.

BENTHIC AND PELAGIC COUPLING

Patterns of stratification and mixing have great importance for the dynamics of lake ecosystems. Most obvious is the

physical environment that is created at different times of the year, but this has large ramifications for chemical and biological conditions within lakes. For example, nonmotile algal cells that are negatively buoyant rely on turbulent forces in the mixed layer to maintain their position within the photic zone. If cells sink out of the mixed layer and beyond the photic zone, the cells can no longer grow and have no means to regain their position within the photic zone. On the other hand, the density difference found within a less turbulent metalimnion may allow some species of algae and bacteria to exist in discrete layers that satisfy their chemical and light environment needs. Movement of gases and nutrients must proceed by the slow process of diffusion when turbulent energy is not available. The slow movement of materials through the metalimnion compared to their production or consumption in the hypolimnion has the apparent effect of sealing the hypolimnion of lakes from receiving oxygen from the atmosphere, and traps the decomposition products in the deeper water. Consequently, cold-water fishes living in temperate zones rely on lakes that are stratified to provide cold hypolimnetic waters, but their distribution can be limited because of the oxygen limitation in these environments.

Perhaps the most striking feature of stratification is the chemical separation of the epilimnion from the hypolimnion, or the profundal benthic zone from the pelagic mixed layer. Although the littoral benthic zone is generally in contact with the mixed layer, the profundal benthic zone is physically separated from interactions with surface waters during stratified periods. While surface waters are in contact with the atmosphere and receive solar radiation for primary production, the profundal zone is dominated by heterotrophic activities. Nutrients are lost from surface water as particles sink, but remineralization taking place within the profundal can release nutrients back into the water column. Recycling of nutrients in this manner can be an important term in nutrient budget of some lakes (Poister *et al.*, 1994).

Several processes are of importance for understanding the influence of stratification on lake ecosystems. These processes, described in the following text, are sedimentation, resuspension, vertical and horizontal eddy diffusion, and hypolimnetic entrainment.

Sedimentation is of great importance to algal cells (Smayda, 1970; Reynolds, 1984). Although particles in the mixed layer of a lake continuously experience turbulent forces allowing them to remain suspended in the water, at times turbulent forces will be diminished. If the cell sinks into the metalimnion or hypolimnion, turbulent forces may be minimal and flow may be nearly laminar. Thus the probability of sedimentation is the product of the number of particles, the settling velocity, and the depth of the mixed layer. The settling velocity of particles is described by the Stoke's equation (Reynolds, 1984). Settling is dependent upon the particle density, radius, and viscosity and is defined as

$$\omega_s = \frac{2gr^2(\rho_s - \rho_w)}{9\mu\varphi} \tag{20}$$

where ω_s is the settling velocity ($m\,s^{-1}$), g is gravitational acceleration ($m\,s^{-2}$), r is the radius of a sphere with volume identical to that of the particle (m), ρ_s and ρ_w are the densities of the particle and water ($kg\,m^{-3}$), μ is the dynamic viscosity ($kg\,m^{-1}\ s^{-1}$), and ϕ is the coefficient of form resistance. Originally Stoke's equation described only spherical particles. Therefore, the coefficient of form resistance can be solved as the deviation of ω_s from that of a sphere with identical density and volume to the respective particle.

Once a particle has sedimented out of the mixed layer or photic zone, the nutrients contained in the particle can no longer be recycled for use by other photosynthetic organisms. This is especially important for nutrients that often limit productivity in lakes, such as P and N, and Si (for diatoms). The particle size distribution of algal communities, because of its impact on sedimentation rate, is an important factor in the rate of decline in P concentration in the mixed layer (Guy *et al.*, 1994). In addition, losses of nutrients from the photic zone are subject to seasonal trends associated with annual blooms of algae (Poister *et al.*, 1994).

Resuspension of sediments that have reached the benthic zone have further ecological importance. Resuspended sediments increase turbidity and may be a source of nutrients or contaminants, or they can act to "scrub" certain chemicals from the water column as they are adsorbed to the particles (Gunatilaka, 1982). Resuspended material can constitute a large proportion of particulate matter in lakes (Evans, 1994). Lake morphometry may dictate some of the properties of sediment resuspension (Carpenter, 1983; Håkanson and Jansson, 1983). For example, resuspended sediments are often considered most important in shallow lakes. However, some amount of sediment resuspension takes place in all lake settings. For sediments to become resuspended, current velocities must create enough shear stress to overcome gravity and the cohesiveness of the sediments (Bloesch, 1995). The shear stress τ ($N\,m^{-2}$) can be quantified directly as

$$\tau_s = \rho_w C_1 u_1^2 \tag{21}$$

where ρ_w is water density ($kg\,m^{-3}$), C_1 is the drag coefficient 1 m above the bottom and u_1 is the current velocity ($m\,s^{-1}$) 1 m above the bottom. The shear stress is often presented, however, as a shear velocity $u_* = (\tau_s/\rho_w)^{0.5}$. Average current velocities in the bottom of stratified lakes are generally not great enough resuspend most particles. However, Gloor *et al.* (1994) found that

occasional bursts of higher current velocities associated with internal seiches could cause resuspension of organic particles up to 100 μm and inorganic particles up to 1 μm, thus maintaining a well-mixed benthic boundary layer. In lakes that stably stratify, resuspension is especially prevalent during periods of overturn. Other factors may reduce the ability of currents to resuspend sediments. Chemical precipitates can create more cohesive sediments that are less likely to be resuspended. Biological activity may physically disturb sediments, or it may also consolidate materials. Examples of consolidation include bacterial mats that bind particles and fecal pellet production by benthic invertebrates.

Within lakes, the greatest amount of resuspension and redistribution of sediments takes place primarily along the littoral benthic zone (Shteinman *et al.*, 1997). Here, surface waves that are deep enough to interact with benthic sediments resuspend material into the mixed layer where it can be redistributed by currents. Deeper areas undergo more episodic events of sediment resuspension. These are related to the shoaling of solitary internal waves at the sloping lake edge (Horn *et al.*, 2001; Boegman *et al.*, 2003). Distinct zones of sediment resuspension and accrual occur because surface waves and currents routinely resuspend particles near-shore shallow areas, and sediments are less likely to be resuspended in deep waters (Håkanson and Jansson, 1983). This causes sediments to move laterally away from the shoreline with time until they are permanently buried, a process known as *sediment focusing*. On the other hand, littoral zones in sheltered bays can be areas of increased sedimentation and decreased resuspension. In addition, macrophytes may impart control on water movements and therefore create areas of increased sedimentation (Madsen *et al.*, 2001).

The release of soluble forms of nutrients from the sediments to the overlying water column may dictate the level of productivity in a lake. For phosphorus, this has classically been viewed as a chemically driven process depending on redox potentials and availability of Fe and S (Caraco *et al.*, 1991 and references therein). However, there is also the indirect mediation of redox potentials by bacteria and the direct role of bacterial utilization and release of P (Kalff, 2002). Movement of nutrients (or contaminants) from sediment to water is generally controlled by molecular diffusion driven by steep concentration gradients at the sediment–water interface. Bioturbation, convective porewater exchange, sediment resuspension, and bubble ebullition may increase the flux of substances across the sediment–water interface. In contrast, the formation of chemical and biological aggregates, such as iron oxides and oxyhydroxides (FeOOH), create barriers to molecular diffusion.

The movement of heat and soluble materials in the absence of advective forces is governed by vertical and horizontal eddy (turbulent) diffusion. Vertical movements during stratified periods are opposed by the stability of the water column imparted by the thermal density gradients. Therefore horizontal eddy diffusion coefficients (K_H) are generally much larger than the vertical coefficients (K_V). Since horizontal movements of water and dissolved materials are not usually impeded by density gradients and often are aided by large advective forces, horizontal movement of materials is generally not considered limiting. However, the magnitude of K_V is often a limiting factor in the supply of dissolved substances from the hypolimnion to the epilimnion.

Eddy diffusion of materials is a result of TKE. At steady state in a stratified system, the production of turbulent kinetic energy is balanced by losses to buoyant fluxes (B) created by density differences and dissipation of turbulent kinetic energy (ε) due to viscous forces (Imboden, 2004). The buoyancy flux, defined from Fick's First Law of diffusive transport, is, $B = -K_V N^2$. The ratio of the buoyant flux to the turbulent kinetic energy dissipation (ε) is the mixing efficiency (γ_{mix}). Rearranging this ratio reveals how the eddy diffusion coefficient is related to energy dissipation and the stability of stratification:

$$K_V = \gamma_{mix} \frac{\varepsilon}{N^2} \tag{22}$$

Thus K_V decreases as the intensity of stratification increases. The mixing efficiency, determined experimentally, ranges form 0.05 to 0.25, and ε can be determined from direct measurement to provide point estimates of K_V (Imboden, 2004). Empirical relationships of the form $K_V = a(N^2)^b$ have shown that b can range from -0.3 to -0.9 (Imberger and Patterson, 1990). Since K_V can vary in space and time, these point measurements may not accurately describe the basin scale vertical flux of materials.

Other methods for estimating K_V involve examining the rate of change of temperature or a tracer concentration, and providing an estimate that integrates over a longer period of time. Quay *et al.* (1980) found, in two lakes, that K_V in the metalimnion (5×10^{-5} and 8×10^{-4} cm^2 s^{-1}) was much less than in the hypolimnion (1.7×10^{-3} and 1.8×10^{-2} cm^2 s^{-1}). These differences were related to difference in the stability of the hypolimnion and metalimnion. Because the two lakes examined were small and well protected, the production of turbulence was small and the eddy diffusion through the metalimnion was very near that of molecular diffusion. In the larger Greifensee, Imboden and Emerson (1978) found a higher K_V (2.5×10^{-2} cm^2 s^{-1}) in the metalimnion. Using this K_V and the P concentration gradient between the hypolimnion and the epilimnion, they determined that the internal loading of P was of the same magnitude as external sources during periods of highest productivity. Within the metalimnion K_V can vary spatially, generally increasing in areas where the metalimnion

intersects the lake bottom, because higher amounts of turbulent kinetic energy are found there (MacIntyre *et al.*, 1999). Therefore, the importance of high eddy diffusion rates for benthic and pelagic coupling are dependent upon their interception with gradients of materials such as nutrients. In Mono Lake, the flux of ammonium from boundary mixing was enough to sustain the observed deep chlorophyll maxima (MacIntyre *et al.*, 1999).

A final process important in benthic and pelagic coupling is the entrainment of hypolimnetic water during periods of thermocline deepening. Factors controlling nutrient transport to the epilimnion during entrainment are: concentrations in the metalimnion, thermal stability of the water column, and timing and pattern of storm events. If concentrations are low in the entrained water, then entrainment has little influence in the overall budget. In Lake Mendota, entrainment was a dominant source of P during a year with average external loading from runoff and supplied nearly as much P to the epilimnion as external loading during a year of above-average runoff (Soranno *et al.*, 1997). Over a 20-year period in Lake Mendota, lake stability was an important predictor for midsummer water clarity (Lathrop *et al.*, 1999). In Lake Tanganyika, increased surface temperatures and decreased winds have increased stability and reduced mixing and input of hypolimnetic nutrients, thereby reducing productivity (O'Reilly *et al.*, 2003).

CONCLUSION

The physical processes of stratification and mixing are important for the chemical and biological functioning of lake ecosystems. Through these physical processes, geomorphology and climatic conditions can have direct and indirect impacts on water quality and biotic interactions. Geomorphology exerts a strong influence on the presence of seasonal stratification, as well as its timing. For example, the size and depth of a lake determines whether stratification will occur. Nutrient cycles are generally changed significantly if lakes stratify (Carpenter, 2003). Climatic conditions generate temporal variation in physical forcing and thereby create variation in the benthic and pelagic coupling within and among lakes. The impact of global climate change on the physical environment of lakes is an area of recent interest (McKnight *et al.*, 1996; Magnuson *et al.*, 2000). Mechanistic models that account for changes in the climate-driven physical forcing will be a useful tool to help understand the potential ramifications of climate change on the biology of lakes (De Stasio *et al.*, 1996). Therefore, further advances in the understanding of stratification and mixing processes and their relationship to benthic and pelagic coupling will increase our ability to predict a lake's response or resilience to natural and human-induced disturbances.

REFERENCES

Arai T. (1964) Some relations between the thermal property of lake and its fetch size. *Geographical Review of Japan*, **37**, 131–137.

Arai T. (1981) Climatic and geomorphological influences on lake temperature. *Verhandlung Internationale Vereinigung Limnologie*, **21**, 130–134.

Baigún C. and Marinone M.C. (1995) Cold-temperate lakes of South America: do they fit northern hemisphere models? *Archiv für Hydrobiologie*, **135**, 23–51.

Bloesch J. (1995) Mechanisms, measurement and importance of sediment resuspension in lakes. *Marine and Freshwater Research*, **46**, 295–304.

Boegman L., Imberger J., Ivey G.N. and Antenucci J.P. (2003) High-frequency internal waves in large stratified lakes. *Limnology and Oceanography*, **48**, 895–919.

Caraco N.F., Cole J.J. and Likens G.E. (1991) A cross-system study of phosphorus release from lake sediments. In *Comparative Analyses of Ecosystems: Patterns, Mechanisms and Theory*, Cole J.J., Lovett G.M. and Findlay S.E.G. (Eds.), Springer-Verlag: pp. 239–258.

Carpenter S.R. (1983) Lake geometry: implications for production and sediment accretion rates. *Journal of Theoretical Biology*, **105**, 273–286.

Carpenter S.R. (2003) *Regime Shifts in Lake Ecosystems: Pattern and Variation*, International Ecology Institute.

Chen C.T.A. and Millero F.J. (1986) Precise thermodynamic properties for natural waters covering only the limnological range. *Limnology and Oceanography*, **31**, 657–662.

Cruikshank D.R. (1984) The relationship of summer thermocline depth to several physical characteristics of lakes. *Canadian Technical Report of Fisheries and Aquatic Sciences*, **1248**, iv–33.

Csanady G.T. (1975) Hydrodynamics of large lakes. *Annual Review of Fluid Mechanics*, **7**, 357–385.

Davies-Colley R.J. (1988) Mixing depths in New Zealand lakes. *New Zealand Journal of Marine and Freshwater Research*, **22**, 517–527.

De Stasio B.T., Hill D.K., Kleinhans J.M., Nibbelink N.P. and Magnuson J.J. (1996) Potential effects of global climate change on small north-temperate lakes: physics, fish, and plankton. *Limnology and Oceanography*, **41**, 1136–1149.

Demers E. and Kalff J. (1993) A simple model for predicting the date of spring stratification in temperate and subtropical lakes. *Limnology and Oceanography*, **38**, 1077–1081.

Eckert W., Imberger J. and Saggio A. (2002) Biogeochemical response to physical forcing in the water column of a warm monomictic lake. *Biogeochemistry*, **61**, 291–307.

Edmundson W.T. and Lehman J.T. (1981) The effects of changes in nutrient income on the condition of Lake Washington. *Limnology and Oceanography*, **26**, 1–29.

Edmundson J.A. and Mazumder A. (2002) Regional and hierarchical perspectives of thermal regimes in subarctic, Alaskan lakes. *Freshwater Biology*, **47**, 1–17.

Evans R.D. (1994) Empirical evidence of the importance of sediment resuspension in lakes. *Hydrobiologia*, **284**, 5–12.

Fee E.J., Hecky R.E., Kasian S.E.M. and Cruikshank D.R. (1996) Effects of lake size, water clarity, and climatic variability

on mixing depths in Canadian Shield lakes. *Limnology and Oceanography*, **41**, 912–920.

Gloor M., Wüest A. and Münnich M. (1994) Benthic boundary mixing and resuspension induced by internal seiches. *Hydrobiologia*, **284**, 59–68.

Gorham E. and Boyce F.M. (1980) *The Influence of Lake Surface Area and Depth upon Thermal Stratification and the Depth of the Summer Thermocline*, Limnological Research Center, University of Minnesota, p. 49, Contribution 177.

Gorham E. and Boyce F.M. (1989) Influence of lake surface area and depth upon thermal stratification and the depth of the summer thermocline. *Journal of Great Lakes Research*, **15**, 233–245.

Green J.D., Viner A.B. and Lowe D.J. (1987) The effect of climate on lake mixing patterns and temperatures. In *Inland Waters of New Zealand*, Viner A.B. (Ed.), DSIR Bulletin 241, Department of Scientific and Industrial Research: New Zealand.

Gunatilaka A. (1982) Phosphate adsorbtion kinetics of resuspended sediments in a shallow lake, Neusiedlersee, Austria. *Hydrobiologia*, **91-2**, 293–298.

Guy M., Taylor W.D. and Carter J.C.H. (1994) Decline in total phosphorus in the surface waters of lakes during summer stratification, and its relationship to size distribution of particles and sedimentation. *Canadian Journal of Fisheries and Aquatic Sciences*, **51**, 1330–1337.

Håkanson L. and Jansson M. (1983) *Principles of Lake Sedimentology*, Springer-Verlag.

Hamilton D.P. and Schladow S.G. (1997) Prediction of water quality in lakes and reservoirs. 1. model description. *Ecological Modelling*, **96**, 91–110.

Hanna M. (1990) Evaluation of models predicting mixing depth. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 940–947.

Hocking G.C. and Straškraba M. (1999) The effect of light extinction on thermal stratification in reservoirs and lakes. *International Review of Hydrobiology*, **84**, 535–556.

Horn D.A., Imberger J. and Ivey G.N. (2001) The degeneration of large-scale interfacial gravity waves in lakes. *Journal of Fluid Mechanics*, **434**, 181–207.

Horne, A.J. and Goldman C.R. (1994) *Limnology*, *Second Edition*, McGraw-Hill.

Hutchinson G.E. (1957) *A Treatise on Limnology*, Vol. I, Wiley.

Hutter K. (1983) *Hydrodynamics of Lakes*, Springer-Verlag.

Idso S.B. (1973) On the concept of lake stability. *Limnology and Oceanography*, **18**, 681–683.

Imberger J. (1985) The diurnal mixed layer. *Limnology and Oceanography*, **30**, 737–770.

Imberger J. (1994) Transport processes in lakes: a review. In *Limnology Now: A Paradigm of Planetary Problems*, Margalef R. (Ed.), Elsevier: pp. 99–193.

Imberger J. (1998) Flux paths in a stratified lake: a review. In *Physical Processes in Lakes and Oceans*, Imberger J. (Ed.), American Geophysical Union: pp. 1–18.

Imberger J. and Hamblin P.F. (1982) Dynamics of lakes, reservoirs and cooling ponds. *Annual Review of Fluid Mechanics*, **14**, 153–187.

Imberger J. and Patterson J.C. (1990) Physical limnology. *Advances in Applied Mechanics*, **27**, 303–475.

Imberger J., Patterson J.C., Hebbert B. and Loh I. (1978) Dynamics of reservoirs of medium size. *Journal of Fluid Mechanics*, **78**, 489–512.

Imboden, D.M. (2004) The motion of lake waters. In *The Lakes Handbook: Limnology and Limnetic Ecology*. O'Sullivan P.E. and Reynolds C.S. (Eds.), Blackwell: pp. 115–152.

Imboden D.M. and Emerson S. (1978) Natural radon and phosphorus as limnologic tracers: Horizontal and vertical eddy diffusion in Greifensee. *Limnology and Oceanography*, **23**, 77–90.

Imboden, D.M. and Wüest A. (1995) Mixing mechanisms in lakes. In *Physics and Chemistry of Lakes*, Lerman A., Imboden D.M. and Gat J.R. (Eds.), Springer-Verlag, pp. 83–138.

Kalff J. (2002) *Limnology*, Prentice Hall.

Kirk J.T.O. (1983) *Light and Photosynthesis in Aquatic Ecosystems*, Cambridge University Press.

Kling G.W. (1988) Comparative transparency, depth of mixing, and stability of stratification in lakes of Cameroon, West Africa. *Limnology and Oceanography*, **33**, 27–40.

Langmuir I. (1938) Surface motion of water induced by wind. *Science*, **87**, 119–123.

Lathrop R.C., Carpenter S.R. and Robertson D.M. (1999) Summer water clarity responses to phosphorus, Daphnia grazing, and internal mixing in Lake Mendota. *Limnology and Oceanography*, **44**, 137–146.

Lathrop R.C. and Lillie R.A. (1980) Thermal stratification of Wisconsin lakes. *Transactions Wisconsin Academy of Science, Arts and Letters*, **68**, 90–96.

Lenters J.D., Kratz T.K. and Bowser C.J. (2005) Effects of climate variability on lake evaporation: Results from a long-term energy budget study of Sparkling Lake, northern Wisconsin (USA). *Journal of Hydrology*, doi:10.1016/j.jhydrol.2004.10.028.

Lewis W.M. (1983) A revised classification of lakes based on mixing. *Canadian Journal of Fisheries and Aquatic Sciences*, **40**, 1779–1787.

Likens G.E. and Hasler A.D. (1962) Movements of radiosodium (Na^{24}) within an ice-covered lake. *Limnology and Oceanography*, **7**, 48–56.

MacIntyre S., Flynn K.M., Jellison R. and Romero J.R. (1999) Boundary mixing and nutrient fluxes in Mono Lake, California. *Limnology and Oceanography*, **44**, 512–529.

Madsen J.D., Chambers P.A., James W.F., Koch E.W. and Westlake D.F. (2001) The interaction between water movement, sediment dynamics and submersed macrophytes. *Hydrobiologia*, **444**, 71–84.

Magnuson J.J., Robertson D.M., Benson B.J., Wynne R.H., Livingstone D.M., Arai T., Assel R.A., Barry R.G., Card V., Kuusisto E., *et al.* (2000) Historical trends in lake and river ice cover in the Northern Hemisphere. *Science*, **289**, 1743–1746.

Mazumder A. and Taylor W.D. (1994) Thermal structure of lakes varying in size and water clarity. *Limnology and Oceanography*, **39**, 968–976.

McKnight D., Brakke D.F. and Mulholland P.J. (1996) Freshwater ecosystems and climate change in North America. *Limnology and Oceanography*, **41**,(special issue), 815–1149.

Mortimer C.H. (1974) Lake hydrodynamics. Mitteilungen. *Internationalen Vereinigung für Theoretische und Angewandte Limnologie*, **20**, 124–197.

Nürnberg G.K. (1988) A simple model for predicting the date of fall turnover in thermally stratified lakes. *Limnology and Oceanography*, **33**, 1190–1195.

O'Reilly C.M., Alin S.R., Plisnier P.D., Cohen A.S. and McKee B.A. (2003) Climate change decreases aquatic ecosystem productivity of Lake Tanganyika, Africa. *Nature*, **424**, 766–768.

Osgood R.A. (1988) Lake mixis and internal phosphorus dynamics. *Archiv für Hydrobiologie*, **113**, 629–638.

Patalas K. (1960) Mieszanie wody jako czynnik okreslajacy intensywnose krazenia materiiw roznych morfologicanie jeziorach okolic Wegorzewa. *Roczniki Nauk Rolniczych*, **B77**, 223–242.

Patalas K. (1961) Wind und morphologiebedingte Wasserwegungstypen als bestimmender Faktor für die Intensität des Stoffkreislaufes in nordpolnischen seen. *Verhandlungen der Internationalen Vereinigung für Theoretische und Angewandte Limnologie*, **14**, 59–64.

Patalas K. (1984) Mid-summer mixing depths of lakes of different latitudes. *Verhandlungen der Internationalen Vereinigung für Theoretische und Angewandte Limnologie*, **22**, 201–218.

Plueddemann A.J., Smith J.A., Farmer D.M., Weller R.A., Crawford W.R., Pinkel R., Vagle S. and Gnanadesikan A. (1996) Structure and variability of Langmuir circulation during the surface wave processes program. *Journal of Geophysical Research*, **101**, 3525–3543.

Poister D., Armstrong D.E. and Hurly J.P. (1994) A 6-year record of nutrient element sedimentation and recycling in three north temperate lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **51**, 2457–2466.

Quay P.D., Broecker W.S., Hesslein R.H. and Schindler D.W. (1980) Vertical diffusion rates determined by tritium tracer experiment in the thermocline and hypolimnion of two lakes. *Limnology and Oceanography*, **25**, 201–218.

Ragotzkie, R.A. (1978) Heat budgets of lakes. In *Lakes: Chemistry, Geology, Physics*, Lerman A. (Ed.), Springer-Verlag, pp. 1–19.

Reynolds C.S. (1984) *The Ecology of Freshwater Phytoplankton*, Cambridge University Press.

Shteinman B., Eckert W., Kaganowsky S. and Zohary T. (1997) Seiche-induced resuspension in Lake Kinneret: a fluorescent tracer experiment. *Water Air and Soil Pollution*, **99**, 123–131.

Shuter B.J. and Lester N.P. (2004) Climate change and sustainable lake trout exploitation: predictions from a regional life history model. In *Boreal Shield Watersheds: Lake Trout Ecosystems in a Changing Environment*, Gunn J.M., Steedman R.J. and Ryder R.A. (Eds.), CRC Press: p. 281–291.

Shuter B.J., Schlesinger D.A. and Zimmerman A.P. (1983) Empirical predictors of annual surface water temperature cycles in North American lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **40**, 1838–1845.

Smayda T.J. (1970) The suspension and sinking of phytoplankton in the sea. *Annual Review of Oceanography and Marine Biology*, **8**, 353–414.

Smith I.R. (1979) Hydraulic conditions in isothermal lakes. *Freshwater Biology*, **9**, 119–145.

Smith I.R. and Sinclair I.J. (1972) Deep water waves in lakes. *Freshwater Biology*, **2**, 387–399.

Snucins E. and Gunn J. (2000) Interannual variation in the thermal structure of clear and colored lakes. *Limnology and Oceanography*, **45**, 1639–1646.

Soranno P.A., Carpenter S.R. and Lathrop R.C. (1997) Internal phosphorus loading in Lake Mendota: response to external loads and weather. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 1883–1893.

Spigel R.H. and Imberger J. (1980) The classification of mixed-layer dynamics in lakes of small to medium size. *Journal of Physical Oceanography*, **10**, 1104–1121.

Straškraba M. (1980) The effects of physical variables on freshwater production: analyses based on models. In *The Functioning of Freshwater Ecosystems*, Le Cren E.D. and Lowe-McConnel R.H. (Eds.), Cambridge University Press: p. 13–84.

Ventz D. (1973) Die Einzugsgebietsgrösse, ein Geofaktor für den Trophiczustand stehender Gerwässer. *Fortschnitte der Wasserchemie und ihrer Grenzgebiete*, **14**, 105–118.

Wetzel R.G. (1973) Productivity investigations of interconnected lakes. I. The eight lakes of the Oliver and Walters chains, northeastern Indiana. *Hydrobiologie Studies*, **3**, 91–143.

Wetzel R.G. (2001) *Limnology: Lake and River Ecosystems*, *Third Edition*, Academic Press.

Wüest A. and Lorke A. (2003) Small-scale hydrodynamics in lakes. *Annual Review of Fluid Mechanics*, **35**, 373–412.

Xenopoulos M.A. and Schindler D.W. (2001) The environmental control of near-surface thermoclines in boreal lakes. *Ecosystems*, **4**, 699–707.

Yan N.D. (1983) Effects of changes in pH on transparency and thermal regimes of Lohi Lake, near Sudbury, Ontario. *Canadian Journal of Fisheries and Aquatic Sciences*, **40**, 621–626.

Yeates P.S. and Imberger J. (2003) Pseudo two-dimensional simulations of internal and boundary fluxes in stratified lakes and reservoirs. *International Journal of River Basin Management*, **1**, 297–319.

109: Reservoirs

G RICHARD MARZOLF[1] AND DALE M ROBERTSON[2]

[1] *United States Geological Survey, Office of Ground Water, Reston, VA, US*
[2] *United States Geological Survey, Water Science Center, Middleton, WI, US*

Reservoirs are surface water impoundments that result from the construction of dams on streams and rivers. Limnological principles are applicable to understanding reservoir functions and the processes that control them because they exhibit lake-like characters (see Lake Ecosystems – ***Chapter 108, Lake Ecosystems (Stratification and Seasonal Mixing Processes, Pelagic and Benthic Coupling), Volume 3****). The basic water qualities of reservoirs are dependent, however, on the river that was impounded in their formation. The river's inflow dominates the upstream reaches of reservoirs. Lacustrine processes become more controlling as advective forces diminish and wind-mixing and solar heating become influential forcing factors and as the cross section of the basin widens and deepens downstream. This transition from riverine to lacustrine imposes longitudinal gradients in reservoirs that emerge as their central feature. Designs of dams and their operating schedules have considerable influence on these in-reservoir patterns and on the effects of the dam in the downstream river downstream. Various models have been developed to describe the physical, chemical, and biological processes in reservoirs. Additional studies will lead to further information to improve the models, which in turn will lead to a better understanding and management of reservoirs.*

HISTORICAL BACKGROUND

Reservoirs by themselves are largely a secondary outcome of the development of dams. The earliest uses involved simple structures that provided for the diversion of water into channels so that the water could be released to irrigate crops. Later dams were constructed to control and store water for irrigation in dry seasons and for conversion of kinetic to mechanical energy used to mill grain and to drive machinery for textile production and machine-tool operation.

Most of the large rivers in the United States figured prominently in our history. They were critical avenues that supported the country's expansion and subsequent development. Examples include the Potomac, Delaware, and Hudson Rivers in the seventeenth and eighteenth centuries. The Mississippi, Ohio, and Missouri Rivers were crucial to the postrevolutionary expansion in the nineteenth century, and the full development of the West is clearly tied to the Colorado, Rio Grande, and Columbia Rivers. As the United States expanded, rivers were important transportation routes. The difficulty in moving boats through shoal reaches and up or down steep reaches stimulated the construction of dams and associated locks. There was a clear federal interest in such river engineering and the Army Corps of Engineers has been central to the enterprise since the late nineteenth century. The construction of dams in the headwaters of rivers to store water in wet seasons helped to control flooding and provided for release of stored water during drought and low-flow seasons, thus extending navigational uses. Federal interest in agricultural development in the arid West through irrigation led to the development of the Bureau of Reclamation and additional dam construction. Recently, dams have offered the benefits of electrical power generation and financing for water projects from the sale of this power. The history of water-resource development surrounding American rivers is colorful and the results are a source of national pride.

Early water-resource development activities were dominated by problems associated with the quantity of water available for various uses. Many reservoirs are important as sources of domestic supply, and many have been developed as habitat for commercial and recreational fishery

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

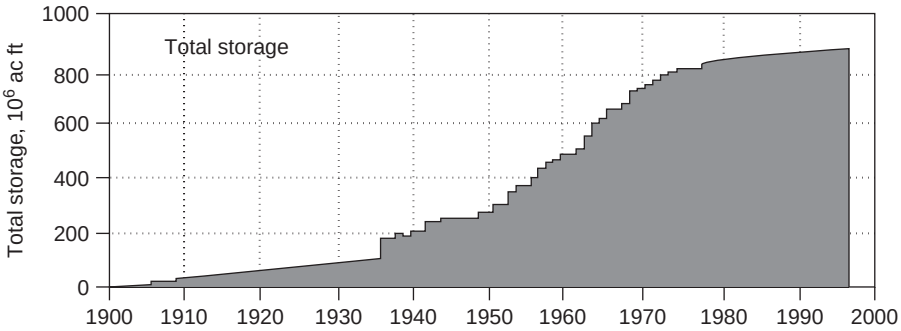


Figure 1 History of increasing total reservoir storage for the continental United States. Data from U.S. Army Corps of Engineers (1996) (Reproduced from Graf, 1999 by permission of American Geophysical Union)

development. Many reservoirs are sites for public recreation because of their lake-like characteristics.

With the emergence of these uses, the view of water-resource managers has shifted to include problems associated with the qualities of water: its temperature; the amount of oxygen dissolved in it; contaminants (e.g., industrial wastes, metals, agricultural chemicals, and pathogens in domestic sewage), and nutrients dissolved in it.

These characteristics of reservoirs have become more important to society as the number of reservoirs has increased, as people use them more for more purposes, and as reservoirs age. Furthermore, many essential characteristics of free-flowing rivers are changed with impoundment, and environmental concerns about these changes have become political, regulatory, and legal. There is general agreement that the era of dam-building in the United States is coming to an end (Figure 1; from Graf, 1999). Construction of new dams is slowing down because of rising costs, a diminishing number of sites where dam construction is feasible, and increasing pressure from environmental groups to limit construction.

There are broader contexts that incorporate dams in world history and use by aboriginal people in North America but the scope of this treatment is limited to development of water resources on the North American continent since European settlement (Worster, 1985).

RESERVOIRS COMPARED TO LAKES

Reservoirs have several basic features in common with natural lakes which means that many principles of traditional lake study (limnology) can be transferred to the study of reservoirs. Basic physical processes occurring in lakes also occur in reservoirs, such as surface warming from solar energy and daily and seasonal cooling, that, when coupled with wind mixing, yield seasonal thermal stratification. Low velocities restrict sediment suspension to the finest particles, allow for a more developed planktonic community, and selects for the survival of a fish community that is unlike that of the parent river. Light penetration defines an upper layer (photic zone) where photosynthesis is carried out by planktonic algae. Organic matter settling into unmixed deeper water may lead to development of anoxic conditions because of the microbial consumption of dissolved oxygen.

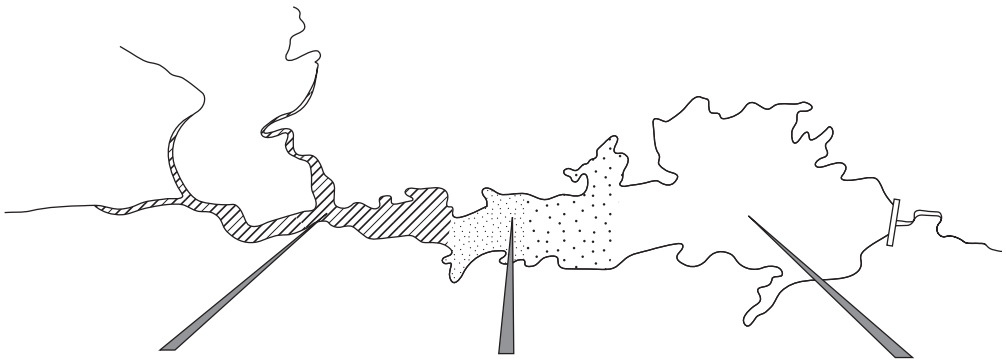
There are, however, some significant differences between reservoirs and natural lakes (Table 1). Reservoirs typically drain larger areas than natural lakes, and typically have shorter retention times than natural lakes. These physical differences often result in a more rapid decline in the water quality of reservoirs.

Natural lakes either have no outlets (seepage lakes) or have one outlet from which water from the surface of the lake exits the system. Reservoirs can also have one outlet; however, water is either released over a dam (spillway), from some intermediate depth (typically through penstocks), or near the bottom (typically through lift gates or release tubes), in addition to water often being transferred for water supply. The quality of the water released from a reservoir is somewhat controllable and can be much different from that of a natural lake because water quality varies so much with depth.

Thermal stratification in water bodies is controlled by solar heating, the temperature and momentum of influent

Table 1 Typical relative differences between lakes and reservoirs

Characteristic	Lakes	Reservoirs
Ratio of surface area to watershed area	Smaller ~10	Larger >500
Retention time	Longer	Shorter
Inflows	None to several	One
Outflows	None or one	One
Transparency	Uniform and clearer	Longitudinal gradients, variable in embayments



Riverine zone	Transitional zone	Lacustrine zone
• Narrow, channelized basin	• Broader, deeper basin	• Broad, deep, lake, like basin
• Relatively high flow	• Reduced flow	• Little flow
• High susp, solids, turbid, low light avail, *Z*p < *Z*m	• Reduced susp, solids, less turbid, light avail, increased	• Rel, clear, light more avail, at depth, *Z*p > *Z*m
• Nutrient supply by advection, rel, high nutrients	• Advective nutrient supply reduced	• Nutrient supply by internal recycling, rel, low nutrients
• Light-limited PPR	• PPR/m^3 rel, high	• Nutrient-limited PPR
• Cell losses primarily by sedimentation	• Cell losses by sedimentation and grazing	• Cell losses primarily by grazing
• Organic matter supply pri-allochthonous, *P* < *R*	• Intermediate	• Organic matter supply primarily autochthonous, *P* > *R*
• More “euthrophic”	• Intermediate	• More "oligotrophic"

Figure 2 Typical water-quality zones in reservoirs with characteristics listed for each zone. [Zp – Photic depth; Zm – Mean depth; P – Production; R – Respiration; PPR – Primary production] (Reproduced from Kimmel, 1990 by permission of John Wiley, Inc)

waters, and wind mixing. Wind mixing in natural lakes commonly establishes a uniform and warmer mixed layer (epilimnion) overlying a colder and denser relatively unmixed layer (hypolimnion). Shallow depths and relatively high flow velocities generally prevent the establishment of stratification in the upstream (riverine zone) reaches of a reservoir (Figure 2). Furthermore, the river flow may remain active and turbulent into the reservoir, further disrupting thermal stability. Frequently, however, if the river inflow is more dense than the water at the surface of the reservoir, the inflowing river water will submerge and enter the reservoir as an underflow (or density current) bounded by the basin and flowing along the bottom. If this river inflow is heavier than surface water but lighter than unmixed bottom water, then it will find its density-equivalent depth and enter the reservoir as an “interflow”, flowing down the long axis of the reservoir until it reaches the dam. Differences in density between river water and reservoir water are typically due to differences in water temperature, but density currents also can occur when inflowing river water contains a high concentration of dissolved materials or suspended sediment.

AGING OF RESERVOIRS

The sediment transported by rivers is often deposited at the upstream ends of reservoirs, with the coarsest particles settling first, thus forming foreset beds on a delta front. At high flow rates during flood stage, these deposits may be overtopped and the delta may be aggraded with topset beds and natural levees. Varying water levels through time create a complicated stratigraphy in the riverine and transitional zones (Figure 2). Through time these deposits provide a setting for invasion of submergent and emergent aquatic vegetation and the establishment of wetlands. Deposits exposed above the controlled water level may be invaded by riparian vegetation. Silt- and clay-sized particles may be carried into the downstream portions of the reservoir, often kept in suspension by wind-driven currents where the basin is shallow. Fine particles deposited in the downstream portions of the reservoir include sediments transported from the reservoir’s watershed and sediments eroded locally due to shoreline processes. Water leaving the reservoir through the dam carries a reduced sediment load relative to the influent river and tributary streams. The effects of sediment reduction on tailwaters and downstream reaches are varied but include: armoring and deepening of the channel bed, erosion of channel-margin deposits, and long-term alteration of habitat characteristics for the river biota. Eventually the reservoirs will fill with sediment and dredging or other management actions will be required.

Eutrophication is the natural process of physical, chemical, and biological “aging” of water bodies associated with the addition of sediment, nutrient, and organic matter. This

natural process often leads to common problems, such as excessive sedimentation, algal blooms, oxygen depletion in the hypolimnetic areas (deeper and cooler layer of water bodies), and excessive aquatic plant growth.

The subject of chemical and biological eutrophication gained substantial attention in the twentieth century, beginning in western Europe with the works of Thienemann (1928) and Naumann (1932) who understood the relations among nutrient contents of lakes, the lake's biota, and the sources of nutrients in the lake's watershed. By mid-century, lakes were frequently classified according to their trophic state, or nutrient content, which is expressed as productivity and biotic composition. Hasler (1947) connected the accelerated eutrophication process in some lakes to nutrient sources from human activity and proximity to urban centers. This accelerated aging caused by human influences is termed "cultural" eutrophication. The nutrients most closely implicated with eutrophication are phosphorus and nitrogen, but other nutrient elements were claimed to be causative as well, for example, carbon and silica.

It is significant and ironic that rivers typically are not identified as eutrophic despite the fact that they often transport high loads of nutrients. The high current velocities (low retention times) and high suspended-sediment loads often do not provide conditions for the expression of high productivities so characteristic of eutrophic lakes. When the rivers are impounded and sediments are deposited, the water becomes more transparent as sediments are deposited, and the retention times are more conducive to planktonic algal growth; as a result, the reservoirs can express the eutrophic state almost immediately.

Because reservoirs often drain much larger land areas than lakes, the eutrophication process often occurs much more rapidly than in lakes and often results in longitudinal gradients in water quality because of the slow downstream movement of water. Figure 2 demonstrates the spatial variability so common in reservoirs with an upstream riverine zone and a more lake-like downstream lacustrine zone. Downstream in the lacustrine zone, the reservoir water often becomes nutrient-depleted, but upstream the nutrient concentrations continue to be driven by advective loading from the impounded river. The position of the transitional zone between these two areas is variable, moving upstream during periods of drought or low inflow and moving downstream during flooding.

Retention time of reservoir segments increases from upstream to downstream along the longitudinal axis of the basin as the cross-sectional area of the reservoir increases. The expression of planktonic production rates is a result of balance between the downstream export of organisms thus depleting populations and the increased rates of production associated with increased light penetration that is adequate for phytoplankton growth.

TRIBUTARY EMBAYMENTS

Reservoirs are formed when river valleys are drowned upon impoundment; therefore, they reflect the dendritic morphology of the parent river's drainage, with isolated embayments extending from the axis of the mainstem reservoir. These embayments often experience limited mixing with the mainstem reservoir that can result in quite different characteristics within the embayments. Embayments are usually protected from winds and thus experience less intense mixing, which often results in greater transparency than in the mainstem reservoir even if the embayments receive additional sources of sediment and nutrients. These simple physical differences often result in higher primary productivity by phytoplankton and/or the invasion of submerged aquatic vegetation. If the embayments have additional sources of nutrients from tributary streams or shoreline human activities, their productivity can be greatly increased.

Embayments are often the focus of reservoir management because of the expression of eutrophication issues, the growth of submerged aquatic plants, and their importance as rearing habitat for the commercial and recreational fisheries. Clearly these same characteristics are valuable to recreational anglers, a feature that focuses additional management attention on embayments. The variability of water qualities among tributary embayments at various places along the main axis of the mainstem reservoir is, or should be, of great interest to reservoir managers. The exchange of water, nutrients, and contaminants between embayments and the mainstem reservoir is, however, not well understood.

MODELING THE PROCESSES IN RESERVOIR

A vast amount of information has been collected describing processes in water bodies and the response of these processes to changes in the external environment. This information can be used to understand these same processes in unmonitored systems, as well as to predict how a reservoir may respond to changes in certain conditions in the future. One way of making these predictions is through the use of models that incorporate results of these specific studies into a predictive tool. Models range in complexity from simple empirical equations that describe changes in a specific parameter to complex three-dimensional dynamic numerical models that describe changes in many parameters over a wide range of spatial and temporal scales.

Empirical models may be simple equations that describe physical characteristics, for example, thermocline depth. More complicated empirical models relate nutrient loading rates and residence time to summer average water-quality conditions, such as phosphorus concentration, chlorophyll *a* concentration, and water clarity. Several

empirical water-quality models are contained in the Wisconsin Lakes Modeling Suite (WiLMS, Panuska and Kreider, 2002, accessible at **http://www.dnr.state.wi.us/org/water/fhp/lakes/laketool.htm**).

Fundamental equations of physics, thermodynamics, and chemistry can be written for water parcels in a reservoir. Moreover, many investigations document specific processes described by these equations. Dynamic models were formed by assembling and linking the fundamental equations with the understanding provided by process studies. The first dynamic models were one-dimensional models that described changes in the thermal structure of lakes and reservoirs; however, these models have evolved to describe changes in the water quality and biology in reservoirs. With advances in the understanding of specific processes associated with improved sampling technology and advances in computer power, two- and three-dimensional models have been created and linked with other watershed and river models.

There are a number of dynamic and semidynamic models that are used to understand the processes in reservoirs. One commonly used semidynamic model is BATHTUB (Walker, 1985; 1986; accessible at **http://www.wes.army.mil/el/elmodels/emiinfo.html**), which performs steady state water and nutrient balance calculations in a spatially segmented hydraulic network to account for advective and diffusive transport, and nutrient sedimentation. Eutrophication-related water-quality conditions are then predicted using empirical relationships derived from assessments of reservoir data. Many commonly used dynamic models are described and made available by the U.S. Geological Survey Surface-Water Quality and Flow-Modeling Interest Group (Rounds, 2004; accessible at **http://smig.usgs.gov/SMIG/**). One commonly used dynamic model is CE-QUAL-W2, which is a water-quality and hydrodynamic model in two dimensions (longitudinal-vertical) for rivers, estuaries, lakes, reservoirs, and river basin systems (Cole and Buchak, 1995; accessible at **http://www.ce.pdx.edu/w2/**). CE-QUAL-W2 simulates eutrophication processes such as relations among temperature, nutrients, algae, dissolved oxygen, organic matter, and sediment.

By describing processes in reservoirs, models can be used to simulate the effects of various management strategies, such as nutrient- or sediment-load reduction, without the costs of implementing each strategy. Models are useful in describing what we know about reservoirs, but they also can be used to help point to what we don't know. Through additional study, we can obtain a better understanding of the processes in reservoirs that will, in turn, lead to better models; an appropriate heuristic cycle.

FURTHER READING

Thornton K.W. (1990) *Reservoir Limnology: Ecological Perspectives*, John Wiley and Sons: New York, p. 246.

REFERENCES

Cole T.M. and Buchak E.M. (1995) *CE-QUAL-W2: A Two-Dimensional, Laterally Averaged, Hydrodynamic and Water Quality Model Version 2.0: Users Manual*, Instruction Report EL-95-1, US Army Engineer Waterways Experiment Station: Vicksburg, Available from National Technical Information Service (ADA298467), tel: 703-605-6000.

Graf W.L. (1999) Dam nation: a geographic census of American dams and their large scale hydrologic impacts. *Water Resources Research*, **35**(4), 1305–1311.

Hasler A.D. (1947) Eutrophication of lakes by domestic drainage. *Ecology*, **28**, 383–395.

Kimmel B.L. (1990) Reservoir primary production. In *Reservoir Limnology: Ecological Perspectives*, Thornton K.W., Kimmel B.L. and Payne F.E. (Eds.), Wiley-Interscience: New York, p. 246.

Naumann E. (1932) Grundzuge der regionalen limnologie. *Die Binnengewasser*, **11**, 1–176.

Panuska J.C. and Kreider J.C. (2002) *Wisconsin Lake Modeling Suite Program Documentation and User's Manual Version 3.3 for Windows*, Wisconsin Department of Natural Resources: PUBL-WR-363–94, p. 32.

Thienemann A. (1928) Der sauerstoff in eutrophen und oligotrophen see. Ein beitrag zur seetypenlehre. *Die Binnengewasser*, **4**, 175.

U.S. Army Corps of Engineers (1996) *Water Control Infrastructure National Inventory of Dams [CD-ROM]*, Federal Emergency Management Agency: Washington.

Walker W.W. (1985) *Empirical Methods for Predicting Eutrophication in Impoundments; Report 3, Phase III: Model Refinements*. Technical Report E-81-9, US Army Engineer Waterways Experiment Station, Vicksburg.

Walker W.W. (1986) *Empirical Methods for Predicting Eutrophication in Impoundments; Report 3, Phase III: Applications Manual*. Technical Report E-81-9, US Army Engineer Waterways Experiment Station, Vicksburg.

Worster D. (1985) *Rivers of Empire*, Pantheon/Random House: New York, p. 402.

110: Paleolimnology and Paleohydrology

SHERILYN C FRITZ[1] AND JASMINE E SAROS[2]

[1]Department of Geosciences & School of Biological Sciences, University of Nebraska, Lincoln, NE, US
[2]Department of Biology, University of Wisconsin, La Crosse, WI, US

The study of the sedimentary record of nonmarine aquatic systems (paleolimnology) provides a tool for reconstructing the history of the basins themselves, as well as the history of the atmospheric and terrestrial systems that influence them. Such records allow us to understand environmental dynamics over periods longer than that of recorded history. Among inland aquatic systems, lakes have been studied more extensively; therefore, this review focuses on insights regarding environmental structure and function gained from studying lake sediments. Studies of the lacustrine paleolimnological record have been used to describe patterns of long-term climatic variation and the response of lakes to changes in climate and catchment vegetation and soils. Although ecological literature from the early decades of the twentieth century suggested that lakes became enriched over time (eutrophic), eventually filling with sediment and becoming wetlands, it is now clear that there are multiple developmental pathways for lakes. Some lakes indeed become enriched, whereas others show a long-term decline in nutrient concentrations and pH. Similarly, succession from lake to bog or wetland is not a universal trend. Paleolimnological studies also have been used to reconstruct the recent history of eutrophication, lake acidification, and atmospheric pollution. In these studies, the history archived in lake sediments provides a tool for measuring the magnitude of human impacts on the environment and for evaluating whether twentieth century patterns fall outside the natural range of environmental variation.

INTRODUCTION

Paleolimnology is the study of the history of nonmarine aquatic systems and how they are influenced by changes in their catchments, by atmospheric inputs, and by climate. Similarly, paleohydrology is the study of past fluctuations in the hydrologic cycle, including precipitation variation, evaporation rates, and surface and groundwater flow. Both are based on the study of lake, wetland, and river deposits, including components within the sediments, as well as geomorphic features of the basins and watershed. Paleohydrology also includes marine systems; however, such studies will not be discussed here. Here, we emphasize studies of sediments deposited in lakes, because these studies comprise the bulk of the paleolimnological literature.

The temporal framework of paleolimnology ranges from studies of contemporary lakes during recent decades or centuries (Smol, 2002) to the geologic record of thousands or millions of years of history of extant or ancient lake basins (Bradbury, 1997; Cohen, 2003; Colman *et al.*, 1995). Because of space limitations, this review will focus on approximately the last 25 000 years, which encompasses the most recent period of widespread continental glaciation in the Northern Hemisphere and the subsequent landscape evolution as the glaciers receded and climate warmed.

The reconstruction of lake history allows us to understand the patterns of variation in lakes over periods much longer than those of the written record and human observation. The longest written time series of lake characteristics span only the last century or so (Magnuson *et al.*, 2000; Nicholson, 1998); the majority are more limited and generally span less than a few decades (Edmondson, 1981; Goldman, 1981; Haworth, 1980; Maberly *et al.*, 1994). During recent decades and centuries, or in some cases millennia, the structure and function of many lakes have been influenced by human activities ranging from local land-use practices in the watershed to the deposition of pollutants transported over great distances via the atmosphere. The impact of these

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

human activities can be disentangled and understood only by reference to archives that predate this influence.

Paleolimnology emerged as a discipline in the early decades of the twentieth century. The pioneering studies explored some of the major ecological questions of the time concerning the extent to which long-term ecosystem development (succession) was self-regulating or driven by external forces. These early paleolimnological studies considered the relationship between changes in catchment vegetation and soils and changes in the morphometry, chemistry, and biology of lakes, primarily in forested regions of the temperate zone. In the late 1960s and 1970s, paleolimnological techniques were used increasingly to address applied questions regarding human impacts on lakes, particularly to determine the magnitude of cultural eutrophication. Later, when acidification of lakes from acid deposition became a major environmental concern, paleolimnological studies became a prominent tool in demonstrating that in some regions pH declines were correlated with increased fossil-fuel combustion and local deposition of atmospheric pollutants. During this environmental debate, the need to quantify reconstructed pH change led to the development of techniques for quantitatively reconstructing water-chemistry change from changes in biotic assemblages (transfer functions – see the following text). These methods are now a routine part of many paleolimnological studies.

Here, we will summarize the major insights regarding lacustrine and environmental structure and function that have emerged from studying the paleolimnological record. These include controls on lake ontogeny, the impact of natural disturbance on lake ecosystems, and the magnitude and nature of human impacts on lakes, including eutrophication, acidification, and alteration of food web structure. We will also consider lakes as important archives of atmospheric processes, particularly of the nature of climate variability, as well as the history of atmospheric pollution.

The Lake as a Sedimentary Archive

Lakes are natural collection basins for material produced within the lake itself and in the adjoining watershed and airshed (Figure 1). Materials enter the lake from outside the basin in surface flow from the catchment, in particulate or wet deposition from the atmosphere, and in groundwater inflow, and also are produced within the lake itself. These materials may sink to the lake bottom, where subsequently they are either recycled into the water column or become a permanent part of the sediment stratigraphy (Figure 2), in either original form or as some degradation product formed by diagenesis. Over time, lake sediments accumulate and preserve an archive of the history of the lake, watershed, and atmosphere. The rate at which sediments accumulate depends on the amount of material that enters the lake from outside, the rate of internal production, and the extent of decomposition of these

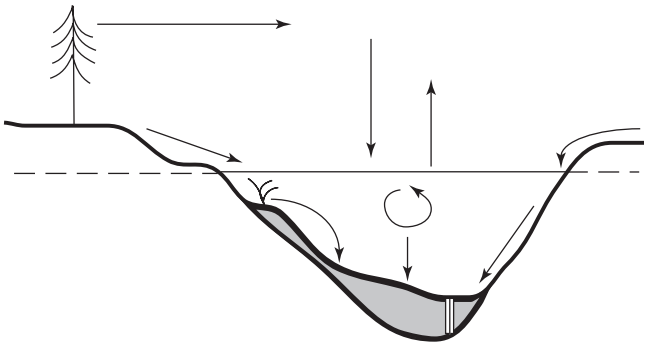


Figure 1 A sketch of sources, as well as losses, of material to lakes and lake sediments. The dotted line is the groundwater table. The shaded area represents accumulated lake sediments, and the striped rectangle is a core sequence through deep-water sediments

components in either the water column or sediments. For example, in relatively undisturbed lakes on hard crystalline bedrock in the Arctic tundra, sediment accumulation rates of $0.1\,\mathrm{mm\,year^{-1}}$ are common, whereas in easily eroded terrains, such as the temperate grasslands of North America, natural accumulation rates often are at least an order of magnitude higher. In disturbed watersheds, such as those dominated by arable agriculture, sediment accumulation rates may be much higher, approaching tens of mm per year.

Commonly, paleolimnological history is reconstructed from a single core from the deepest part of the lake. Studies that compare spatial patterns of deposition in different parts of the lake basin indicate that, in most cases, stratigraphic trends tend to be reproducible from one site to another, given continuous deposition. However, sedimentary concentrations or accumulation rates vary considerably from shallow to deep water because of differences in production and transportation rates. Therefore, basinwide accumulation rates cannot be reconstructed from a single core but require multiple cores from different water depths and depositional environments (Anderson, 1989; Engstrom and Swain, 1986). Multiple cores from within a lake basin have been used to calculate loading rates of pollutants, such as mercury or phosphorus, from watersheds into lakes (Balogh *et al.*, 1999) and thus to quantify the magnitude of human impact on aquatic systems (see the following text).

TOOLS FOR RECONSTRUCTION OF ENVIRONMENTAL HISTORY

Lake sediments sometimes are exposed in section above the contemporary lake surface or may be retrieved by coring. Many types of coring devices have been designed to core lakes of varied size, water depth, and sediment characteristics and to recover sequences that range from just

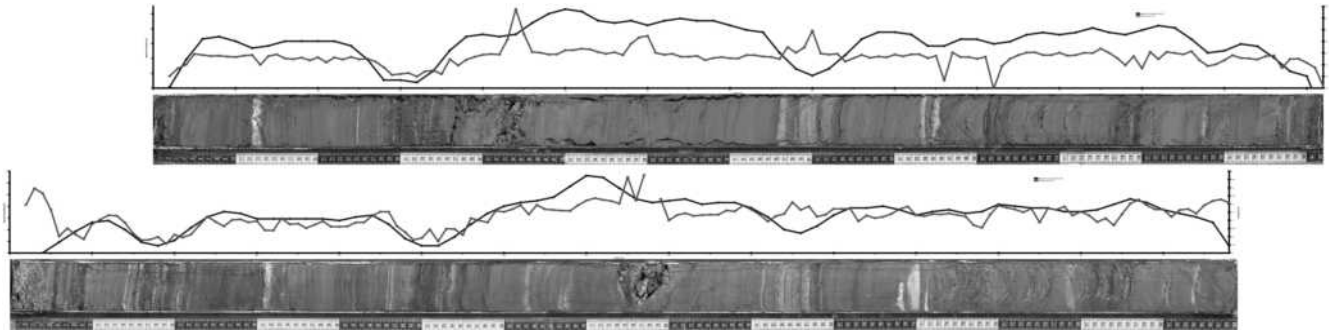


Figure 2 Photograph of sediment cores from Lake Titicaca, Bolivia/Peru. The two cores are from drill holes adjacent to one another and show the reproducibility of the sedimentary sequence. The cores are from an interval of lowered lake level, which resulted in the precipitation of carbonate (light bands) from the water column. The shaded scale bars are each 10 cm, so the upper core is 140 cm in length and the lower is 150 cm. Adjacent to the core photographs are graphs of sediment density (gray line) and magnetic susceptibility (black lines). Magnetic susceptibility is higher in units that have higher minerogenic content. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the uppermost centimeters to many meters of penetration (Glew *et al.*, 2001). In large, deep lakes or in dried basins, drilling techniques may be utilized to recover sedimentary sequences (Leroy and Colman, 2001). Recovered sediments are generally stored at cold temperatures (~4 °C) and subsequently are processed to retrieve the component biological, physical, or chemical information.

Environmental history is reconstructed from visual features of the sediment (laminations or bedding characteristics, burrows, deformation, color) (Figure 2), other physical characteristics (grain size, porosity, density), and analysis of component biological or biogeochemical materials that form the sediment matrix. The analyzed components are often referred to as *proxies*, because each is assumed to represent some dominant environmental process or condition. For example, one cannot directly see past lake level, but changes in diatom groups that are characteristic of deep or shallow water can provide a proxy of lake-level variation through time. Generally, more than one set of conditions or processes can influence any particular organism, sediment component, or structure, and hence there may not be a unique historical sequence that can be derived from an individual proxy. In the best cases, a form of hypothesis testing is used in which multiple sedimentary components are analyzed, and the resultant data are used as a matrix to reject some potential explanations and hence to piece together a coherent sequence of past events or conditions, consistent with the multiple lines of evidence.

Many tools are available for paleolimnological reconstruction (Figure 3). For reconstructions of the history of the lake itself, these include lake organisms that preserve in whole or in part (e.g. diatoms, chrysophyte cysts and scales, chironomids, ostracodes, gastropods, pollen or seeds of aquatic plants) and organic (plant or bacterial pigments, lipids, fatty acids, humic and fulvic acids, isotopes) or inorganic (algal nutrients, major cations and anions, metals, isotopes, minerals) compounds that are formed or reactive with the water column or sediments. Atmospheric history can be reconstructed from chemical compounds in various forms (hydrogen and oxygen isotopes, S, N, Pb, Hg, pesticides) or particulates (carbonaceous particles, loess) that may be of atmospheric origin or atmospherically transported. A variety of tools exist for reconstruction of catchment history, such as chemical compounds that enter in dissolved or adsorbed form; particulate or mineral material that may be eroded or windblown; pollen, spores, charcoal, and macroscopic plant parts (leaves, twigs); and organic compounds of terrestrial origin (leaf waxes, fatty acids). However, because this review is concerned largely with limnological and hydrologic history, we will not consider many of these terrestrial proxies or dwell on studies focused on catchment history alone. In recent years, a series of books and review articles have been published on the methods, uses, and limitation of all major paleolimnological proxies (e.g. Last and Smol, 2001a,b; Smol *et al.*, 2002a,b). Therefore, rather than summarizing individual proxies here, we will rely on examples of paleolimnological reconstructions to illustrate the capabilities and uses of common proxies and techniques.

Statistical Tools

In recent years, statistical techniques have been used increasingly to derive or infer environmental variables, such as pH, salinity, total phosphorus concentration, or temperature from sedimentary proxies, particularly from complex biological assemblages (Birks *et al.*, 1998). Commonly, multivariate statistics, such as canonical correspondence analysis (ter Braak, 1986), are applied to large data sets from contemporary lakes and used to identify environmental variables that are strongly correlated with species distributions. Transfer functions can be generated for variables that have statistically significant and independent

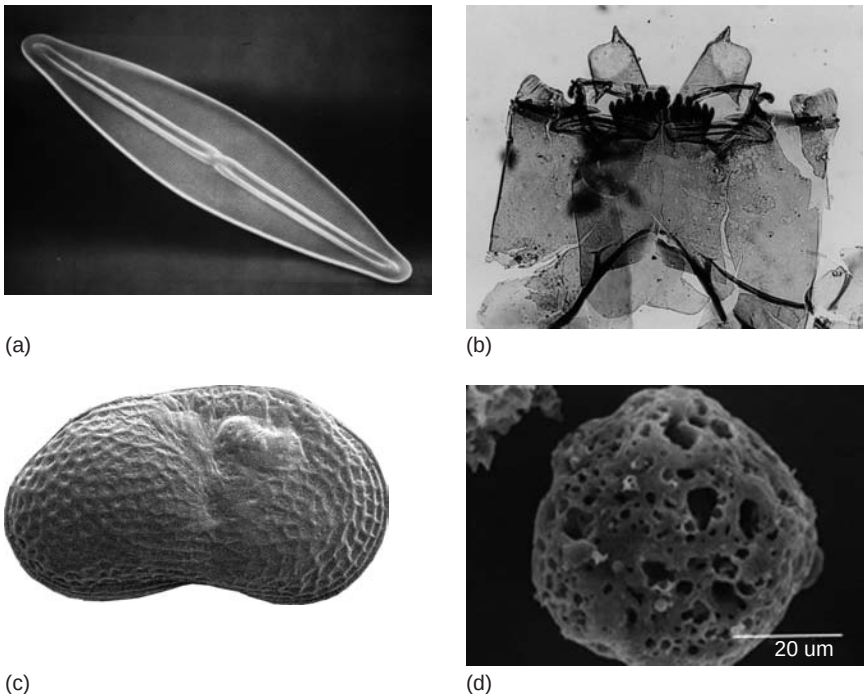


Figure 3 A selection of proxies used in paleolimnological studies: (a) Diatom (Photo: S. Fritz), (b) Chironomid head capsule (Photo: E. Barley), (c) Ostracode (Photo: A. Schwalb), (d) Spheroidal carbonaceous particle from fossil-fuel combustion (Photo: N. Rose). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

influence in these modern data sets and then applied to fossil data to derive a quantitative reconstruction of environmental conditions through time. These multivariate techniques were applied first to infer pH history from the stratigraphy of fossil diatoms in studies of lake acidification (Birks *et al.*, 1990), but are now used widely in paleolimnological studies of nutrient enrichment and climate change (see the following text). Another method used to infer past conditions from biological assemblages is analog matching, in which fossil assemblages are matched to the most similar assemblages in contemporary lakes, and a reconstruction of some environmental variable in the past is then generated from modern measurements in those systems (Smith, 1993). The accuracy of transfer functions can be determined in several ways. Most commonly, the observed environmental values and the values derived from the transfer function are compared for the samples that are used to generate the transfer function, often using bootstrapping or jackknifing techniques to calculate prediction error (Fritz *et al.*, 1999) (Figure 4). Occasionally, sufficient historic data may exist so that reconstructed values from proxies in a sediment core spanning the last century or so may be compared with instrumental measurements of the inferred variable (Dixit *et al.*, 1992; Fritz, 1990; Lotter, 1998; Schelske and Hodell, 1991) (Figure 5) or of climatic parameters that are presumed to drive that variable (Bigler and Hall, 2003; Laird *et al.*, 1998).

Statistical tools also may be used to quantify the pattern, timing, or rate of change in sedimentary variables. For example, multivariate techniques, such as PCA or DCA, can be used to summarize the pattern of change in species composition through time (Figure 6) or to evaluate variation in the rate of change of some proxy variable in the paleolimnological system (Laird *et al.*, 1998). These techniques essentially reduce multiple species or multiple parameters to a single variable or a small number of variables that statistically track the dominant patterns of variation in the larger data set. Similarly, time series analysis can be used to establish the dominant frequencies of temporal variation (e.g. Rodbell *et al.*, 1999).

Chronology

The chronology of environmental change observed in stratigraphic records can be determined via a suite of absolute and relative dating techniques. Some lakes produce annual layers (varves), which can be counted and used to establish a chronology; these are typical only in systems with seasonal inputs and conditions suitable for preservation of sediment structure (Lamoureux, 2001). The most commonly used methods for obtaining absolute ages are radiometric techniques, which include ^{210}Pb for sequences spanning the last ~200 years (Appleby and Oldfield, 1978) and radiocarbon dating of organic material (Björck and Wohlfarth, 2001), which can establish a chronology that

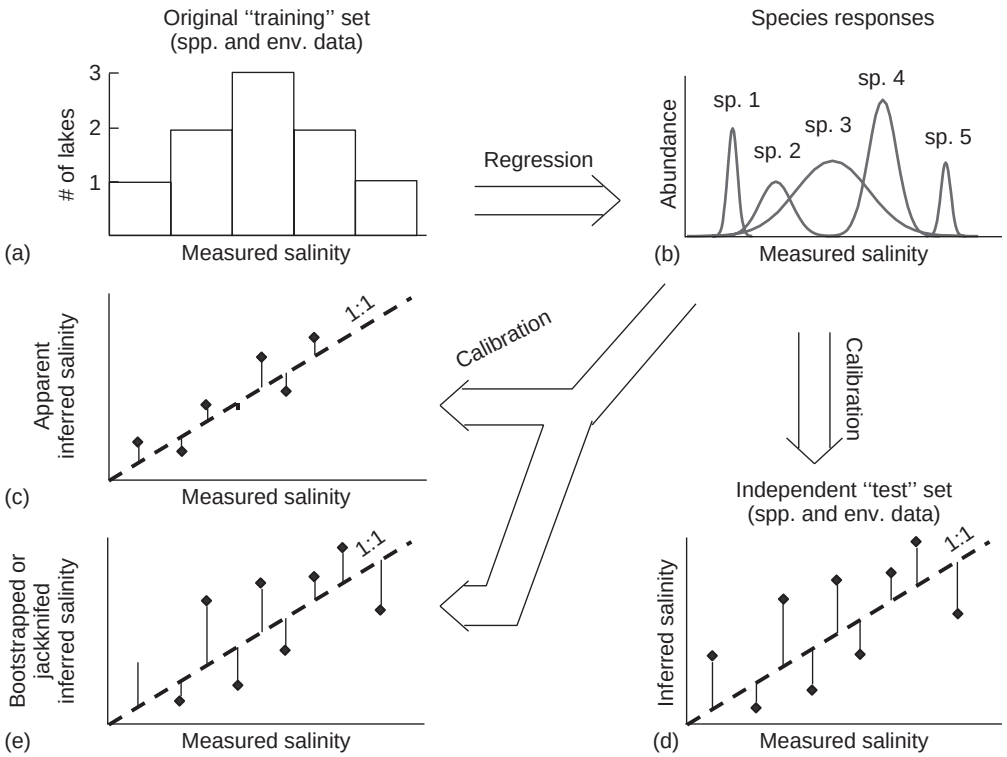


Figure 4 The typical steps involved in developing and assessing the predictive ability of a transfer function, including (a) selection of a modern set of lakes that span an environmental gradient of interest; (b) the regression step, where species responses are estimated based on the species distribution in the modern data set; and the calibration step, where (c) the set of lakes used to estimate species responses are used to generate and evaluate the transfer function; or (d) an independent set of lakes are chosen to evaluate the transfer function model; or (e) computer resampling techniques, such as bootstrapping or jackknifing, are used to evaluate the model (Reproduced from Fritz *et al.*, (1999) by permission of Cambridge University Press)

extends back approximately 50 000 years. Uranium-series dating techniques can be used to establish the age of authigenic carbonates in lacustrine sequences (Israelson *et al.*, 1997), particularly in lakes with evaporite minerals (Fritz *et al.*, 2004a). In suitable circumstances, these tools can be applied to deposits that date back several hundred thousand years. Luminescence dating also has been applied to lake sediments (Berger and Easterbrook, 1993), but this technique relies on high concentrations of quartz or feldspar grains that have been exposed to light (bleached) and consequently is generally only useful where eolian or fluvial input are major sources of sediment.

Some dating techniques produce sediment ages by identification of events or patterns of change in the sediment sequence that are of known age. The expansion of weedy species that spread with land disturbance, such as *Ambrosia* or *Salsola*, has been used to mark the timing of European settlement in North American lake cores that span this period (Bradbury, 1975; Jacobson and Engstrom, 1989). Spikes in the concentration of radionuclides produced by nuclear weapons testing, such as ^{137}Cs, which reached peak atmospheric concentrations in 1963, can be used to identify the age of individual layers of the sediment. In European sequences, profiles of pollutants have been tied to the history of fossil-fuel combustion (Renberg and Wik, 1984) and used to establish a chronology. Volcanic ashes of known mineralogy and age also can be identified in cores and used to establish the age of a single horizon (Sarna-Wojicki *et al.*, 1983; Turney and Lowe, 2001). Less frequently, measurement of paleomagnetic intensity changes are made and correlated to a dated master sequence (King and Peck, 2001).

INSIGHTS FROM THE PALEOLIMNOLOGICAL RECORD

Lake Ontogeny

The evaluation of how lakes age has been a recurring theme in limnology since Pearsall observed lakes of varied nutrient concentration in the English Lake District in the 1920s (Pearsall, 1932) and assumed that they formed a developmental sequence from oligotrophic to eutrophic. Because direct observation of lake development over centuries and millennia is impossible, limnologists have relied

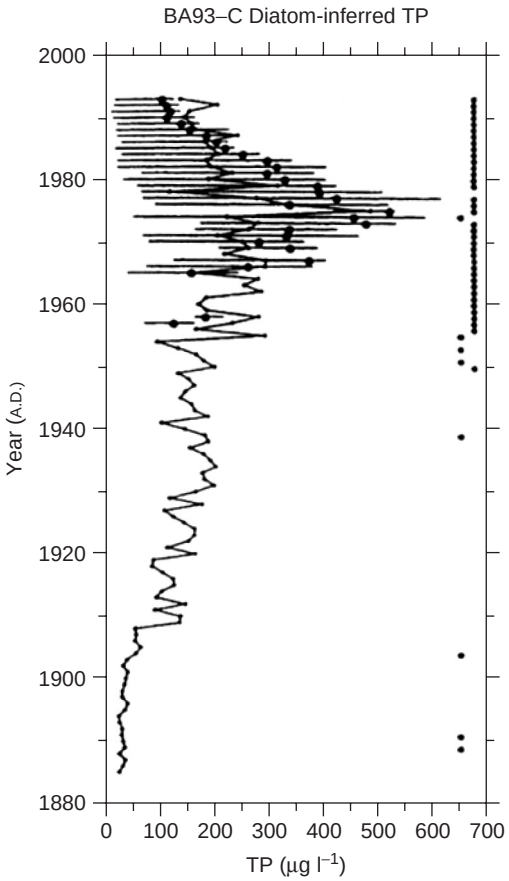


Figure 5 Comparison of diatom-inferred phosphorus (line) with measured concentrations of total phosphorus (TP) in a Swiss lake. The dots represent the measured annual mean TP, and the horizontal lines represent the annual range of TP in the uppermost 15 m of the water column (Reproduced from Lotter (1998) by permission of Arnold)

on sedimentary records to infer patterns of limnological change and to derive hypotheses about the factors controlling the direction and rate of change (Binford and Deevey, 1983). Pioneering paleolimnological studies from lakes in temperate latitudes emphasized changes in lake nutrient concentrations and correlative changes in productivity and hypothesized that lakes become enriched over time, eventually reaching a steady state (Figure 7) or trophic equilibrium (Deevey, 1942). The hypothesis of a trophic equilibrium was a product of one of the major ecological debates of the time regarding whether ecosystems were largely driven by internal dynamics culminating in a steady state (climax) or instead were forced by external, more stochastic, variables, such as climate. As the number of paleolimnological studies has increased in recent decades, it has become clear that there are multiple developmental pathways, rather than a single trajectory of nutrient enrichment with time. Similarly, with the proliferation of data, it is apparent that climate, lakes, and lake catchments are dynamic on all timescales and, therefore, that the concept of steady state is not applicable.

Paleolimnological studies from boreal regions suggest that lakes in areas of base-poor soils with moderate to high precipitation become progressively more acidic as a result of leaching of base cations from catchment soils (Ford, 1990; Pennington *et al.*, 1972; Round, 1961; Whitehead *et al.*, 1989) and/or decreased groundwater inputs caused by podzolization of soils (Engstrom *et al.*, 2000). Associated with the loss of alkalinity is a concomitant increase in water color (dissolved organic carbon, DOC) associated with the buildup of soil organic matter (Engstrom *et al.*, 2000; Pienitz *et al.*, 1999). Some of the early literature from boreal regions also suggested that lakes become progressively more oligotrophic with time, as cations and nutrients are depleted from surface soils (Pennington, 1981). Recent paleolimnological studies of multiple lakes in the boreal rainforest of southeastern Alaska (Figure 6) (Fritz *et al.*, 2004b) and Sweden (Bigler *et al.*, 2003) showed that, although increases in acidity and DOC concentrations over time were widespread in boreal regions of high rainfall, the rate and pattern of water-chemistry change was highly variable because of localized differences in surface hydrology and geomorphology. In southeast Alaska, the pattern of change in nutrient concentrations with time was spatially heterogeneous, and lakes did not show a common developmental sequence during the first few thousand years because of small-scale differences in vegetation change and consequent soil development.

The notion that lakes eventually fill in with sediment and become bogs or wetlands is well entrenched in the scientific literature (Wetzel, 2001). Certainly, infilling is common in some regions, such as in glaciated areas where shallow kettle holes are common and in areas of high precipitation where podzolization of soils and subsequent paludification lead to the growth and expansion of peatlands (Ugolini and Mann, 1979). However, clearly succession from lake to bog or wetland is not a universal trend and occurs only in places where basins are sufficiently shallow to be filled and where erosion rates are sufficiently high to fill the basins before the lakes are destroyed or altered by tectonic, geomorphic, or climatic forces.

Despite over a century of paleolimnological study that now spans many parts of the globe, there has been no recent synthesis of developmental trends in lakes, and most studies of lake ontogeny have focused on lakes in forested zones of north temperate and boreal latitudes. With the increase in sites from grassland, alpine, and arid regions, and from tropical and high-latitude areas, such a synthesis is timely to elucidate the common versus unique features and to further refine our hypotheses about patterns and rates of change and their environmental controls.

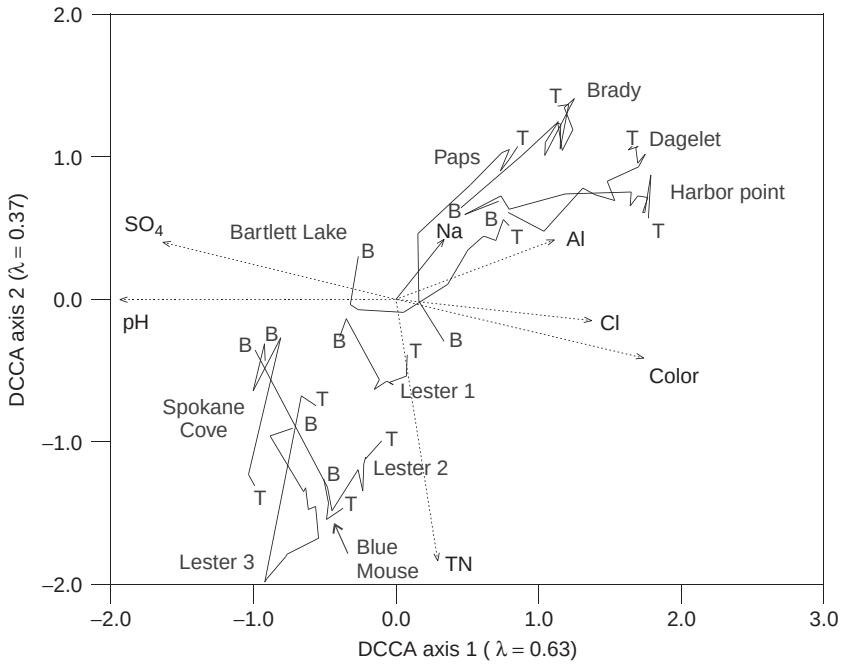


Figure 6 Patterns of change in diatom community structure in lakes from Glacier Bay, Alaska based on multivariate analysis (Detrended Canonical Correspondence Analysis). The arrows show the major water-chemistry variables correlated with the inferred change based on analysis of modern diatom communities. Thus, the first axis of variation is change in pH, whereas variation on the second axis is correlated with lakewater nutrient concentrations (Total Nitrogen). T marks the top of the core or recent times, whereas B marks the base of the cores. The lakes were all formed during the last 2000 years by Neoglacial ice retreat, and thus the pattern of change marks the trajectory of each lake from the time of formation. Note that most lakes show a change along axis 1, suggesting regional pH decline over time, whereas the pattern of change along axis 2 (nutrient gradient) is more variable (Reproduced from Fritz *et al.*, (2004b) by permission of Arnold)

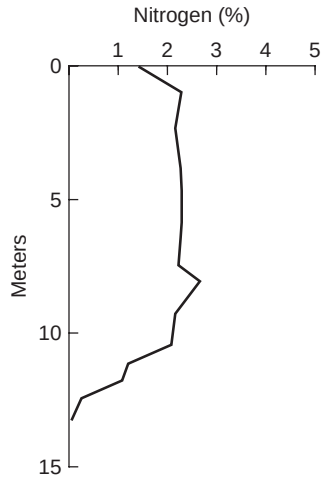


Figure 7 Changes in nutrient concentrations over the last ~10 000 years in Linsley Pond, Connecticut. The stable period is the period of ''trophic equilibrium'' as envisioned by early paleolimnologists (Reprinted from Hutchinson G. E. and Wollack A. (1973) Studies on Connecticut lake sediments. II Chemical analyses of a core from Linsley Pond, North Branford. *American Journal of Science*, **238**, 493–517, by permission of the American Journal of Science)

Response of Lakes to Natural Disturbance

A large portion of the sediment in a lake basin is usually derived from the watershed, and hence lake sediments can be used for reconstructing the history of watershed processes. In some cases, detrital sediments derived from catchment inwash provide a record of the timing of severe storms (Noren *et al.*, 2002) or, in appropriate coastal settings, of hurricane strikes (Liu and Fearn, 2000). These storm events, in turn, can be used to infer synoptic climate variation related to *El Nino* – Southern Oscillation (Rodbell *et al.*, 1999) or the North Atlantic Oscillation, for example.

Paleolimnological data also can be compared with independent evidence of landscape disturbance to evaluate the nature and rate of limnological response and subsequent recovery.

Pollen studies in both North America and Europe have demonstrated abrupt changes in mid-Holocene forest composition associated with pathogens, prehistoric activity, or a combination of the two. These situations provide an opportunity to examine the extent to which forest disturbance affects lakes via changing nutrient inputs, light transparency (erosion), mixing regimes, or hydrologic inputs (surface runoff) and whether or not the lake returns to its former

state and, if so, at what rate. In small lakes in Europe, the widespread decline of elm ~5000 years ago was associated with nutrient enrichment that persisted for several hundred years before recovery to prior conditions (Fritz, 1989). In this case, the role of natural forest disturbance by pathogens relative to Neolithic settlers is debated. In eastern North America, hemlock declined precipitously about 4800 years ago as a result of natural causes (insects and/or climate change), which produced a 2000-year-long change in forest composition before hemlock recovered to its former abundance. Paleolimnological studies from affected sites (Boucherle *et al.*, 1986; Hall and Smol, 1993) show that phosphorus concentrations and productivity, as reconstructed from diatom analysis, increased following the hemlock decline in lakes with large catchments relative to lake size, whereas in small catchments the impacts were limited. Lacustrine productivity subsequently declined as a result of reduced nutrient release during recovery of forest biomass, as forest gaps were colonized by early successional tree taxa. These patterns mirror those seen in experimental studies of the impact of forest clearance on nutrient release to aquatic ecosystems (Bormann *et al.*, 1974). The unique potential of the paleolimnological approach, however, is to show the longer aquatic response over hundreds of years as the early successional taxa were replaced by hemlock. The data suggest that as hemlock abundance increased, lacustrine nutrient concentrations also showed a moderate increase. This suggests that early versus late successional forests differ in their retention of nutrients by vegetation and soils (Hall and Smol, 1993; St. Jacques *et al.*, 2000).

Volcanic eruptions create plumes of airborne tephra that can be transported to great distances from the volcanic center. Tephra deposition in a lake provides a potential source of silica to enhance production by siliceous microorganisms, as well as preservation of siliceous remains in the sedimentary environment. Massive tephra inputs also have the potential for short-term depression of primary production, because of light scattering by particulate material in suspension. Stratigraphic studies reveal that in small lakes, tephra deposition may rapidly change diatom community composition, likely because it alters Si/P ratios and hence nutrient availability; these effects may persist for decades (Abella, 1988; Barker *et al.*, 2000). Studies of lakes affected by the Lacher See tephra, which blanketed large areas of western Europe ~11 000 years ago, suggest catchment lithology affects the impact of ash deposition, although it is unclear whether the lithologic impact is direct or indirect (Birks and Lotter, 1994). In catchments of acidic bedrock, diatom species composition changed, but not in catchments in calcareous terrain.

Climate Change

Lake sediments can be used to reconstruct climate change based on analysis of physical, biological, or chemical variables that preserve in the sediments and are sensitive to climate via either direct or indirect influences (Fritz, 1996). In subhumid to arid regions, paleolimnological studies of variation in lake-level or in ionic concentration (salinity) or composition driven by changes in the balance of precipitation relative to evaporation (P-E) have been instrumental in understanding the natural variability of drought. Drought variability commonly is reconstructed from changes in diatom (Fritz *et al.*, 1999) or ostracode (Smith, 1993) species sensitive to salinity or water depth or from changes in the chemistry (isotopic or trace-element composition) of authigenic or biogenic calcite (Holmes, 2001; Ito, 2001), which reflects the water chemistry at the time of formation. Lake-level change driven by drought also can be reconstructed in transects of cores from shallow into deep water by tracing sediments characteristic of shallow water (with coarse grain size and/or littoral macrophytes) and their migration lakeward at times of lake-level lowering (Digerfeldt *et al.*, 1993).

The magnitude of lake-level change driven by climate in prior millennia is massive by twentieth-century standards. In the tropical Andes, a combination of diatom and geochemical analyses of cores (Baker *et al.*, 2001) and seismic stratigraphy (Seltzer *et al.*, 1998) in Lake Titicaca showed that the lake dropped at least 80 m during a dry phase between 8000 and 4000 years ago. Just to the south in the contemporary salt flat, the Salar de Uyuni, lake levels were as much as 140 m higher during glacial times (Fritz *et al.*, 2004a). Lakes in the northern tropics of Africa underwent lake-level changes of tens of meters (Gasse, 2000; Scholz and Rosendahl, 1988), and paleolake deposits dated between 8000 and 4000 years ago in the now dry Sahara and Sahel (Figure 8) attest to massive hydrological changes caused by intensified monsoon precipitation (Gasse, 2000). An earlier lake-level lowering of tens of meters in East Africa about 11 000 years ago caused desiccation of Lake Victoria, which implies that the diverse assemblage of endemic cichlid fishes must have evolved since that time (Johnson *et al.*, 1996). The paleolimnological record provides an unusual tool for assessing how aquatic biota and ecosystems have responded to climate change in the past, which in turn can contribute to prediction of future responses. However, most paleolimnological studies have used lake records as indicators of climate, rather than analyzing lacustrine response.

During the last few thousand years, the major boundary conditions of the earth's climate system (orbital configuration, atmospheric CO_2 concentrations, extent of continental glaciation) have been roughly similar to those of today. Hence, this period provides a yardstick for reconstructing natural patterns of variation and, in doing so, to evaluate whether twentieth century climate patterns are different from those prior to potential human influences on the climate system. In many regions, paleolimnological

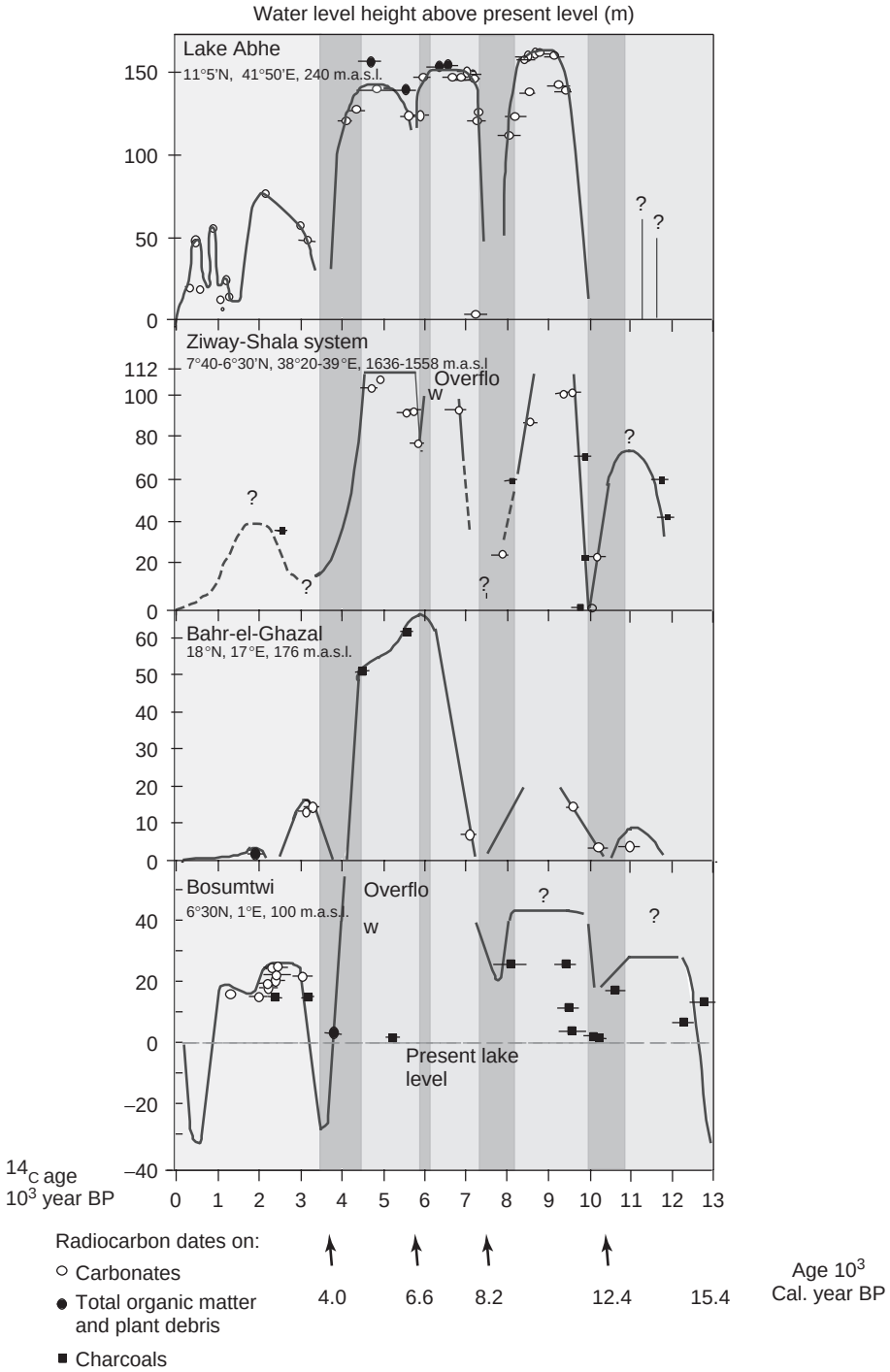


Figure 8 Lake-level changes during the last ~12 000 years in several African lakes in the Sahel and southern Sahara as inferred from paleolimnological data. Note the high lake-level relative to today for much of the period prior to 4000 years ago, as well as the abrupt short-term periods of lake-level lowering indicated by the arrows and shaded zones (Reprinted from Gasse, 2000. ©2000, with permission from Elsevier)

assessments of drought variation at high temporal resolution (decadal or subdecadal) during the last few thousand years indicate that the twentieth century does not encompass the full range of natural variation. In the northern Great Plains of North America, for example, it is clear that intervals of drought as severe and much longer than the Dust Bowl droughts of the 1930s were common in recent millennia (Fritz *et al.*, 2000). In both the Americas and

in Africa, extreme drought was characteristic of Medieval times (Stine, 1994) and coincident with prehistoric cultural collapse or emigration (Figure 9), suggesting that climate change may have contributed to cultural history (Binford *et al.*, 1997; Hodell *et al.*, 2001; Verschuren *et al.*, 2000).

The importance of hydrologic setting in mediating a lake's response to drought has become apparent in a number of paleolimnological studies (Fritz *et al.*, 2000). In a multiproxy study of a lake in central North America, ostracode species composition and shell Mg/Ca concentrations indicated increased lakewater salinity ~5000 years ago, whereas oxygen isotopic values of ostracode calcite became more depleted (Smith *et al.*, 2002). Modeling of basin groundwater flow suggested that during extreme drought, groundwater influx to the lake might have increased because of drying of shallower upstream basins, thereby increasing the input of isotopically depleted water to the lake (Donovan *et al.*, 2002). In contrast, the increased temperatures and evaporation characteristic of midsummer produced increased Mg/Ca in ostracode calcite. Groundwater influences also may affect the magnitude of lake response to decreased effective moisture, such that a lake whose water budget is dominated by groundwater inputs may be relatively stable in water chemistry over time in contrast to a neighboring basin with limited groundwater inflow (Reed *et al.*, 1999). In the Sand Hills of Nebraska, lakes actually formed during extremely dry periods, when migrating sand from destabilized dunes blocked drainage channels, causing dune-dammed lakes to form in response to a rising water table (Loope *et al.*, 1995). These examples of variability in hydrologic response highlight the importance of understanding the hydrogeomorphic setting and of using multiple indicators (Bigler *et al.*, 2002; Rosén *et al.*, 2001) and multiple lakes (Fritz *et al.*, 2000) for well-constrained reconstructions of past climate.

Both temperature and precipitation variation can impact the weathering of catchment soils and hence the input of major cations and anions to lakes. In subhumid or humid regions of base-poor soils, these changes in weathering rates may affect lakewater pH (Sommaruga-Wograth *et al.*, 1997), and hence, pH changes reconstructed from microfossils can be used to infer climate-driven variations in weathering rates (Psenner and Schmidt, 1992) in areas of stable vegetation and soil composition.

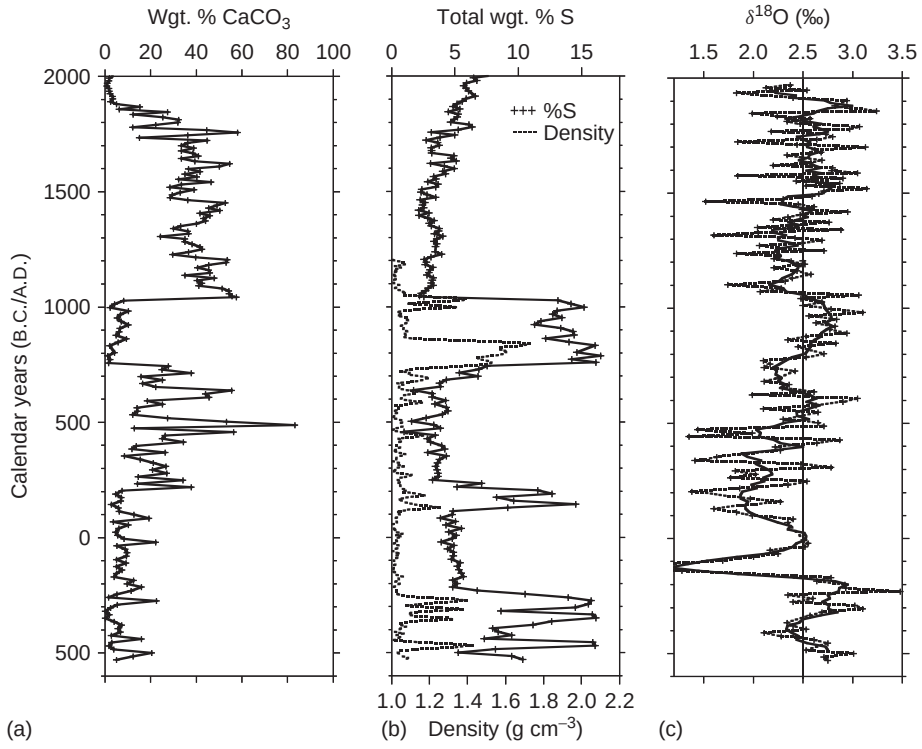


Figure 9 Weight % CaCO3 (a), weight % sulfur (b), sediment bulk density (dashed line), and isotopic concentration of calcite in the gastropod *Pyrophorus coronatus* (c) in a core from Lake Chichancanab, Mexico. Increased values of $\delta^{18}O$ are inferred to represent times of drought (increased evaporation relative to precipitation). The solid line through the isotopic data is a five-point running mean. During times of evaporative concentration of lakewater, gypsum ($CaSO_4$) precipitates from the lake, thereby increasing the % sulfur in the core. The massive increase in % S and enriched isotopic values ~A.D. 800–1000 coincides with the Classic Maya collapse (Reprinted from Hodell, D. A., Brenner, M., Curtis, J. H. and Guilderson, T. (2001). Solar forcing of drought frequency in the Maya lowlands. Science 292, 1367–1370. ©2001 AAAS)

Temperature change affects lakes directly by influencing surface-water temperatures, thermal structure, and the extent and duration of ice cover, and in some cases these variables can be reconstructed from paleolimnological data. In Arctic lakes, the duration of ice cover affects CO_2 exchange with the atmosphere and hence impacts pH in lakes with low buffering capacity. Thus, diatom-inferred pH trends can be used to infer long-term temperature trends (Wolfe, 2002). In Europe, several abrupt cooling episodes that persisted for hundreds of years (Younger Dryas, Preboreal Oscillation, 8.2 event) have been identified between 12 000 and 7000 years ago from records of $\delta^{18}O$ in carbonates, because isotopic fractionation is temperature dependent (Siegenthaler *et al.*, 1984; von Grafenstein *et al.*, 1999). In high northern latitudes, these cooling episodes produced pronounced decreases in lacustrine primary productivity, as evident from the analysis of biogenic silica and organic carbon profiles (Björck *et al.*, 1997). Transfer functions to reconstruct temperature directly from the species composition of chironomids and diatoms (Figure 10) have been derived from modern lakes that span a gradient of temperature and applied to stratigraphic records from lakes, particularly in high latitudes (Joynt and Wolfe, 2001; Levesque *et al.*, 1997; Lotter *et al.*, 1997; Pienitz *et al.*, 1995; Walker *et al.*, 1991). In northern Sweden, however, temperature reconstructions based on diatoms and chironomids were similar during the last 6000 years, but they differed in prior millennia (Figure 10) (Bigler *et al.*, 2002). In this setting, the diatom species composition was strongly affected by pH changes associated with long-term vegetation and soil development following deglaciation (see the preceding text). Thus, pH influences on diatoms overrode the influence of temperature until vegetation and soils stabilized, and therefore, the diatom-based temperature reconstructions during the millennia following lake formation are likely in error.

Temperature change in Arctic regions can be inferred from biological variables (Douglas *et al.*, 1994), as well as from the sedimentology of lakes that receive glacial meltwater and sediments during open-water periods. In Greenland and the Canadian Arctic, these records suggest that twentieth-century temperatures may be higher than at any time during the last 400 years (Lamoureux and Bradley, 1996; Lamoureux *et al.*, 2001; Overpeck *et al.*, 1997), although in some regions they may not be extreme in the context of slightly longer time frames (Hughen *et al.*, 2000).

Human Impacts on Lakes and Their Watersheds

Eutrophication

The first application of paleolimnology to applied limnological questions was the reconstruction of changes in nutrient concentrations over the last few centuries in individual European and North American lakes that were suspected of being enriched from various catchment activities, including sewage influx and land clearance for home building and timbering. Increases in eutrophic diatom and ostracode species and in pigment concentrations, particularly of cyanobacteria, often could be correlated with independent evidence of catchment disturbance (Bradbury, 1975; Engstrom *et al.*, 1985). Such data were pivotal in demonstrating that the nutrient concentrations in some contemporary lakes were considerably higher than their predisturbance state. In the North American Great Lakes, stratigraphic studies of biological indicators (Stoermer *et al.*, 1985) have been coupled with isotopic measurements on organic matter and carbonate (Hodell and Schelske, 1998; Hodell *et al.*, 1998) to evaluate productivity changes associated with enhanced nutrient loading. More recently, transfer functions have been applied to stratigraphic sequences in an attempt to quantify the magnitude of nutrient enrichment of individual lakes (Figure 5) (Bennion *et al.*, 1996; Fritz *et al.*, 1993; Hall and Smol, 1996) or populations of lakes. In the population-based studies, reconstructed nutrient concentrations in presettlement samples (ca. mid-1800s) are compared with modern conditions in multiple sites to assess overall regional trends and to quantify the proportion of regional lakes that have been affected by human activities (Figure 11). In both eastern and midwestern regions of North America, lakes in forested areas generally show little recent change in nutrient concentration, whereas those in regions with significant residential, urban, and agricultural land use show widespread increases in nutrient concentrations, productivity driven pH, and conductivity from road salt input (Dixit *et al.*, 1999; Ramstack *et al.*, 2003; Siver *et al.*, 1999). One of the real values of paleolimnological data is that they provide information on predisturbance conditions and hence can be used in management decisions to set realistic targets for lake restoration (Anderson *et al.*, 1993; Bennion *et al.*, 1996) or to evaluate the impact of efforts to reduce nutrient inputs (Anderson and Rippey, 1994; Hall *et al.*, 1999; Ramstack *et al.*, 2003).

In many stratigraphic sequences from North America, widespread land clearance in the late nineteenth and early twentieth centuries produced massive changes in lake ecosystems (Fritz *et al.*, 1993). However, paleolimnological data have demonstrated that land clearance does not always enrich lakes. A comparison of lakes in both eastern and western Canada in catchments logged for timber production with reference lakes showed that in several regions recent change in lakes with forest clearance in their catchments was no greater than the natural variability in nonimpacted lakes; only very oligotrophic lakes surrounded by steep slopes showed distinct responses to logging and even those responses were small and short-lived (Laird and Cumming, 2001; Laird *et al.*, 2001).

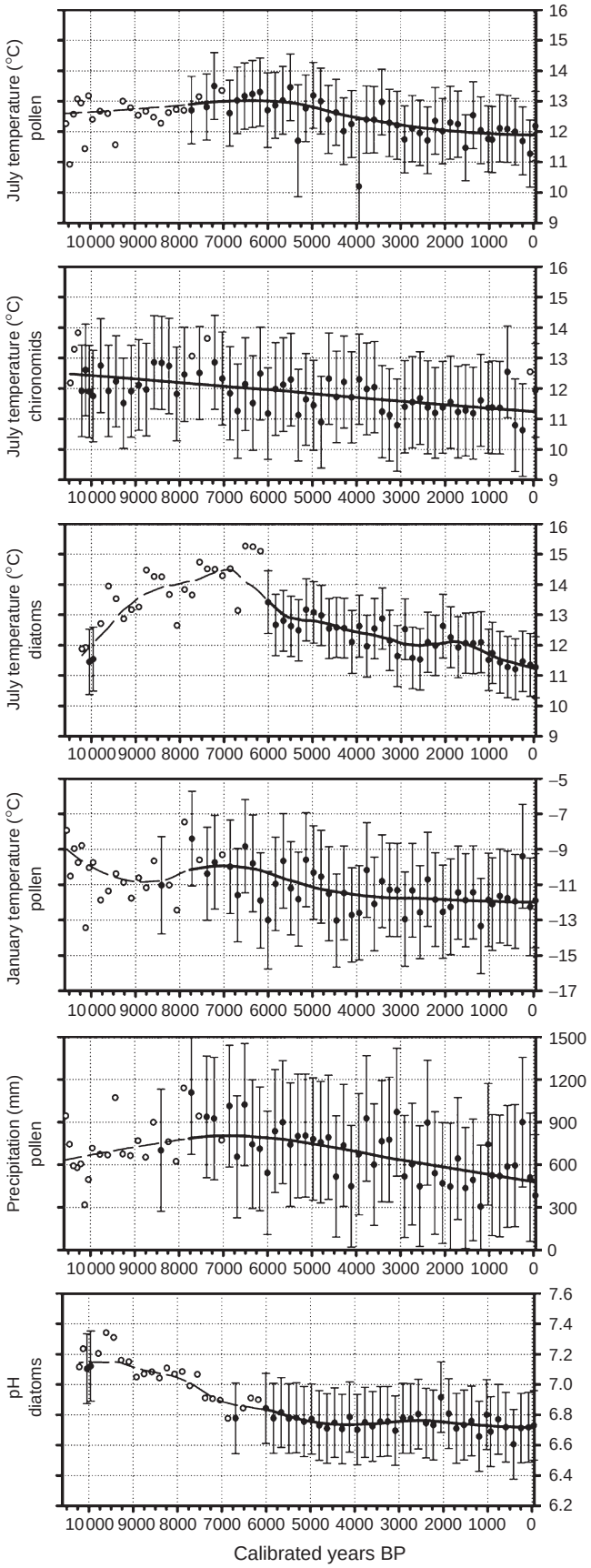


Figure 10 Reconstruction of temperature change in a lake in northern Sweden over the last ~12 000 years based on diatoms, chironomids, and pollen. Note the difference in the diatom temperature reconstruction relative to the other temperature proxies prior to 6000 years ago. Diatom-inferred pH suggests major changes in lake water pH during the early lake history. Thus, it is likely that pH change associated with early soil development exhibited a bigger influence on diatom community structure prior to 6000 years ago than did temperature change (Reproduced from Bigler *et al.*, (2002) by permission of Arnold)

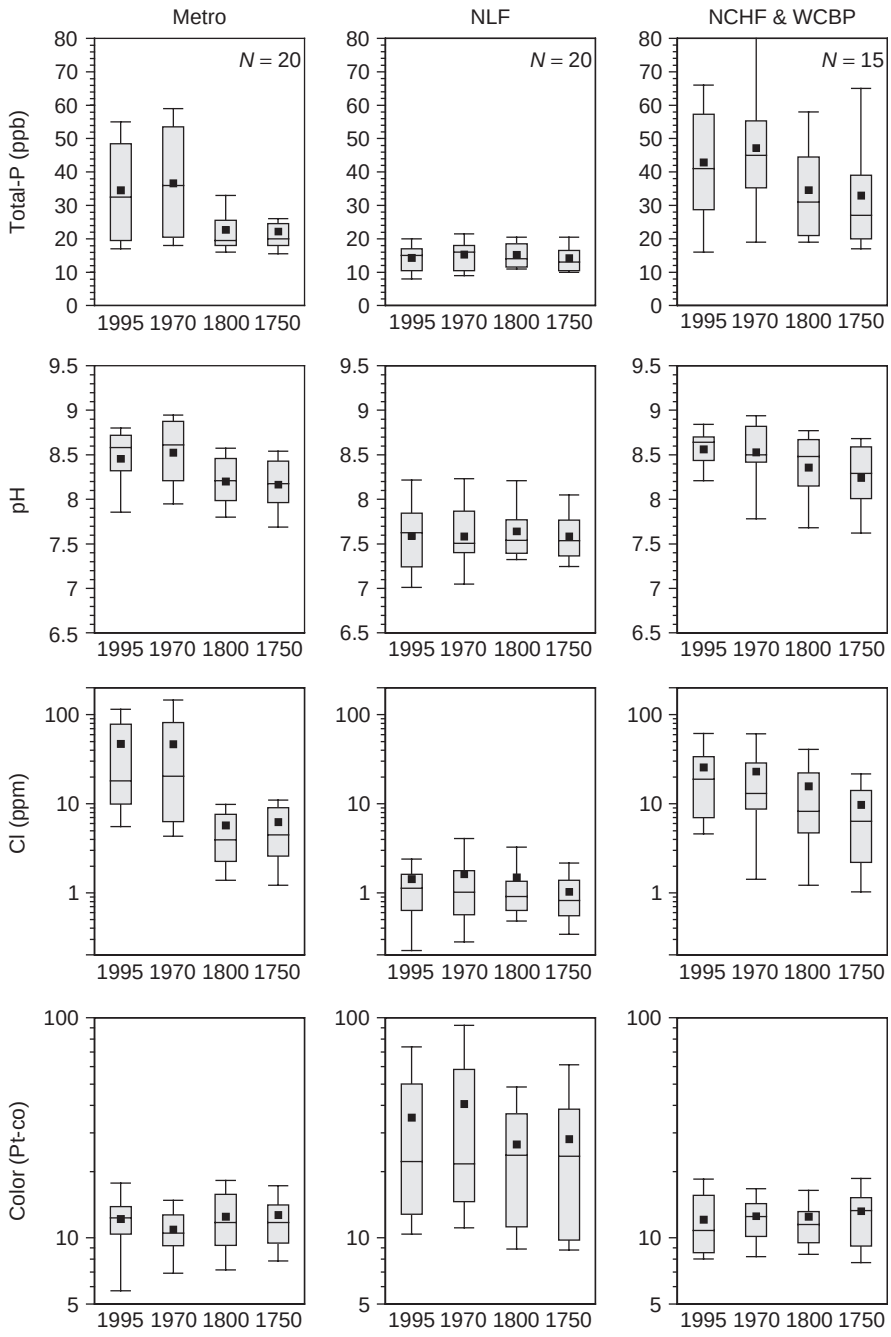


Figure 11 Diatom-based reconstructions of total phosphorus (TP), pH, chloride concentration, and color from core tops (recent times) and core intervals dated at 1970, 1850, and 1700 from a suite of 55 lakes in Minnesota. The lakes are grouped into urban lakes from the Minneapolis–St. Paul metro area, lakes in minimally impacted forested areas, and lakes from agricultural areas of the state. Means are the black squares, gray boxes represent interquartile ranges, the bars are the upper and lower 10%, and the centerline is the median (Reproduced from Ramstack *et al.*, (2004) by permission of NRC Research Press)

Paleolimnological research has demonstrated that in some parts of the world, prehistoric populations also had major impacts on nutrient concentrations (Binford *et al.*, 1987; Ekdahl *et al.*, 2004), particularly in areas of fertile soils. During the Bronze Age (~3000 years ago) in the lowlands of England, for example, the clearance of small patches of land for cereal cultivation within the deciduous forest caused short-lived blooms of cyanobacteria. More extensive land clearance in the Iron Age (~2000 years ago) resulted in massive eutrophication, which was exacerbated

in subsequent centuries with further cultivation and ultimately urban expansion in Anglo Saxon times (Fritz, 1989). One of the classic paleolimnological studies documented the impacts of road building in Roman times on the sedimentation and biotic community of an Italian lake (Hutchinson, 1970).

Clearly, not all nutrient enrichment is human induced. In the grassland regions of North America, for example, a progression from oligotrophic or mesotrophic to eutrophic diatom species parallels the transition from forest to prairie as the climate warmed following retreat of continental glaciers (Radle *et al.*, 1989). Similarly, stratigraphic diatom profiles from a number of European lakes show evidence of modest nutrient enrichment as catchments stabilized following deglaciation and as the climate warmed (Round, 1961).

Food Web Structure

With the implementation of fish stocking programs in western North American lakes in the early twentieth century, questions were raised about how this alteration may have affected food web structure. A comparison of fossil plankton records with contemporaneous limnological data from two manipulated Michigan lakes between 1940 to 1986 revealed that paleolimnological tools could be used to accurately reconstruct shifts in trophic structure (Leavitt *et al.*, 1989). The fish assemblages of Peter and Paul Lakes were manipulated several times during that period; plankton tows revealed that changes in zooplankton and phytoplankton communities accompanied these manipulations. The sediment record accurately reflected these shifts, providing information on changes in total primary production, in community composition of zoo- and phytoplankton, and in size structure of cladoceran populations.

In alpine lakes of the American west, lacustrine sediment records have been used to explore whether shifts in trophic structure have occurred with the implementation of fish stocking programs. In Mount Rainier National Park, sedimentary diatom profiles recorded assemblage changes following the introduction of nonnative trout in the 1920s to previously fishless lakes (Drake and Naiman, 2000). Diatom communities did not return to predisturbance assemblages after fish removal.

Turning the question around, fossil zooplankton assemblages have been used in an alpine lake to determine whether a seemingly native fish population was actually indigenous or became established through an undocumented introduction (Verschuren and Marnell, 1997). The abundance of fossil *Daphnia ephippia* in Avalanche Lake of Glacier National Park suggested heavy zooplanktivory and thus a sustained presence of native cutthroat trout prior to the 1930s. An increase in zooplankton populations in the 1930s and early 1940s indicated a disturbance to food web structure, possibly due to failed attempts at introducing nonnative trout species during this time. Overall, the reconstructions of food web dynamics presented here provide good examples of how paleolimnological tools can be used in hypothesis testing, particularly in remote lakes for which little historical data are available.

Acidification/pH Change

The impact of acid rain on aquatic ecosystems was a major environmental issue in the 1970s and early 1980s. Because historical time series of pH measurements were extremely rare, it was difficult to evaluate from contemporary data whether or not lakes had become more acidic as a result of the deposition of acids derived from fossil-fuel combustion or alternatively whether the acidity was of natural origin (Patrick *et al.*, 1981). Thus, the paleolimnological record became a major tool for historical assessment of long-term water-chemistry trends. The use of paleolimnological data in the political debate demanded rigorous hypothesis testing, high temporal resolution of the sediment record, quantification of the amount of limnological change, and generation of robust error estimates. As a result, paleolimnology rapidly evolved from a largely descriptive science to a more quantitative one that integrated analysis of multiple variables in individual cores.

The reconstruction of pH change from paleolimnological records commonly is based on changes in the species composition of diatoms and chrysophytes. Large data sets of surface-sediment samples (the upper $\sim 0.5-1.0$ cm of mud, which represent modern deposition) and associated water-chemistry measurements were collected from lakes in regions of base-poor soils that were potentially sensitive to acid deposition. Multivariate statistical techniques were used to identify the pH ranges of algal species in these modern samples and to generate transfer functions to reconstruct pH from diatom or chrysophyte assemblages (Birks *et al.*, 1990; Dixit *et al.*, 1993). Subsequently, these transfer functions were applied to fossil assemblages from cores to reconstruct pH conditions in the past. Application of regional transfer functions (Battarbee and Charles, 1987; Charles *et al.*, 1989; Cumming *et al.*, 1994) demonstrated that in acid-sensitive regions of Scandinavia and Great Britain, many lakes had acidified by 0.5–1.5 pH units since the onset of major industrialization in the late 1800s. In North American regions, downwind of major industrial and urban centers, large-scale acidification also had occurred in the decades following increased fossil-fuel combustion, whereas there was little evidence for acidification in acid-sensitive regions with low atmospheric deposition rates, such as in the Sierra Nevada and northern Rocky Mountains. Analysis of sedimentary profiles of trace metals that were components of fossil fuels, such as Pb, Cu, and Zn, as well as carbonaceous particles (Figure 3) resulting from combustion, served to demonstrate that these pH changes were associated with increased local deposition of atmospheric pollutants (Charles and Norton, 1986; Fritz *et al.*, 1989). Paleolimnological techniques also

were used to track the recovery of impacted aquatic systems and to evaluate the factors affecting patterns and rates (Battarbee *et al.*, 1988; Dixit *et al.*, 1992; Smol *et al.*, 1998).

The development of transfer functions for pH inferences also enabled the quantitative reconstruction of natural pH change in the millennia preceding industrialization (Jones *et al.*, 1989; Whitehead *et al.*, 1989). One of the most elegant paleolimnological studies ever conducted analyzed over 700 contiguous diatom samples spanning the last ~12 000 years from a lake in northern Sweden (Figure 12) to evaluate whether the magnitude or rate of recent pH change differed from that associated with natural processes (Renberg, 1990). Long-term pH decline caused by natural soil development had occurred in the millennia following lake formation (see Section on "Lake ontogeny"), but the rapid rate of acidification in the last century was unparalleled in the 12 000-year record. Also apparent in the stratigraphic record is a decrease in acidity about 2300 years ago at the time when agriculture expanded regionally, likely caused by plowing and the erosion of carbonate-bearing subsurface soils (Renberg, 1990; Renberg *et al.*, 2001).

Ultraviolet Radiation

Recent trends in stratospheric ozone depletion have raised concerns about the amount of potentially damaging ultraviolet (UV) radiation, particularly UV-B (280–320 nm), within aquatic environments. This has prompted the development of paleolimnological techniques for reconstructing UV conditions, as once again, direct observations are typically too brief to elucidate the magnitude and effects of changes in UV radiation. Paleolimnological investigations have focused on systems at high latitudes or altitudes, where the atmosphere is thinner, and the position of treeline fluctuates, which alters inputs of DOC to these lakes. The humic and fulvic fractions of DOC strongly absorb UV wavelengths and thus affect UV attenuation in the water column. Transfer functions relating diatom distributions to DOC concentrations and water transparency have been developed for a set of lakes in the Northwest Territories of Canada (Pienitz and Smol, 1993) and applied to sediment cores from boreal treeline lakes in that area to reconstruct DOC concentration (Pienitz *et al.*, 1999). These reconstructions reveal shifts over the Holocene in DOC concentration with changes in watershed vegetation, as confirmed by pollen

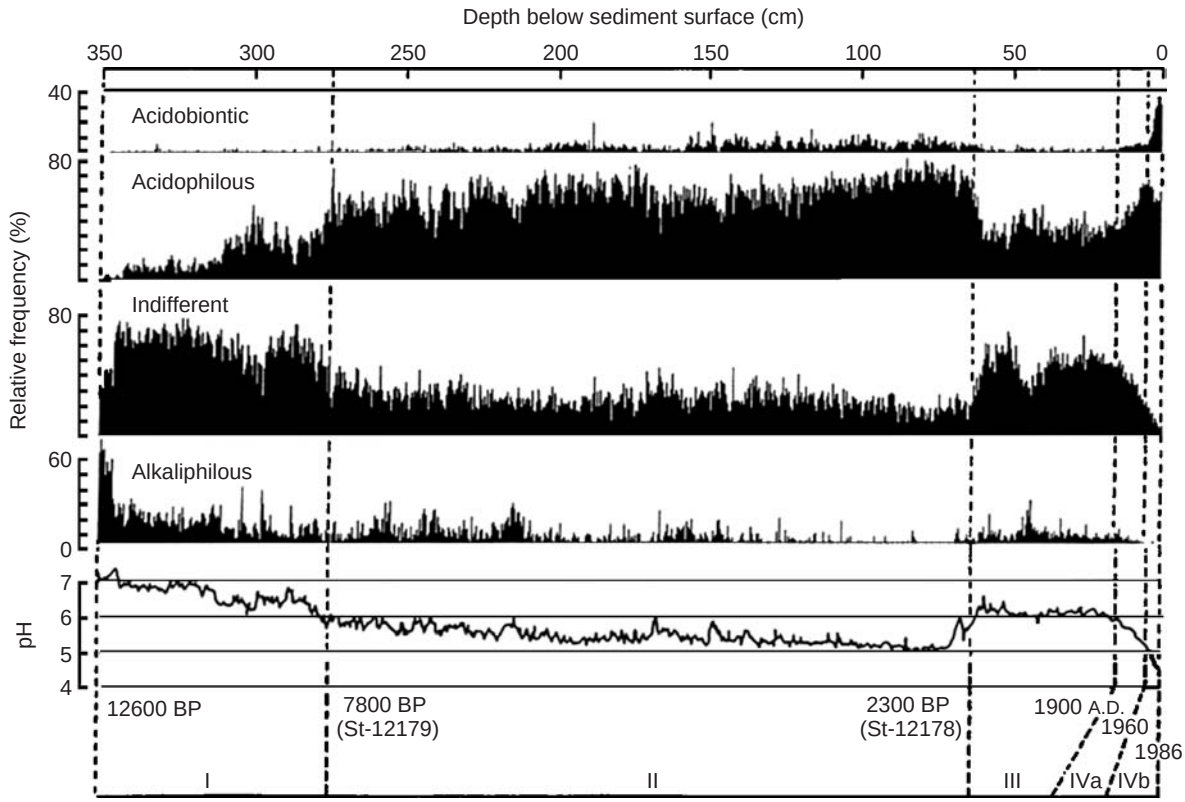


Figure 12 Changes in diatom-inferred pH from Lake Oresjon, Sweden. In the upper part of the figure, the diatoms species are grouped into acid requiring (acidobionitic), acid-preferring (acidophilous), pH indifferent, and alkaline preferring (alkaliphilous) taxa. The lower curve is the pH as reconstructed from these diatom data. Note the increase in pH ~2300 years ago associated with intensified prehistoric agriculture and the unprecedented rapid pH decline after 1900 A.D. (Reproduced from Renberg (1990) by permission of Royal Society of London)

analyses; these DOC changes, in turn, may affect UV penetration and hence aquatic communities.

Fossil pigments also have been used to investigate past UV exposure in aquatic systems, as some benthic algae produce unique pigments when exposed to UV. Two indicator pigments that are produced under high UV conditions have been identified, and the validity of their use in inferring past UV environments was tested in an experimentally acidified boreal lake (Leavitt *et al.*, 1997). Acidification reduced measured concentrations of DOC about fourfold and increased the maximum depth of UV-B penetration eightfold between 1982 and 1990. During this period, the sediment record indicated a 3- to 10-fold increase in these pigments, revealing that these fossil pigment records may provide insight into past UV environments in additional lakes.

Atmospheric Pollutants

Lacustrine sediment records also have provided information on deposition rates and patterns of atmospheric contaminants, particularly of metals, such as lead (Pb) and mercury (Hg). Studies of lead pollution in Swedish lakes have revealed a surprisingly early (about 3000 years ago) increase in Pb deposition, beginning well before industrialization and attributed to metal production (Brännvall *et al.*, 1999; Renberg *et al.*, 1994). Rates of Pb pollution have varied extensively over the past 3000 years, with clear peaks occurring during periods such as Greek–Roman times, the Industrial Revolution, World War II, and the 1970s (Brännvall *et al.*, 2001). Patterns of Pb deposition in lakes across Sweden reveal higher deposition rates in southern Sweden, supporting the idea that the main source of contamination is mainland Europe and Great Britain. Similarly, in the Bolivian Andes stratigraphic studies of metals associated with smelting have been used to document the extent of pre-Incan and pre-Colonial metallurgy (Abbott and Wolfe, 2003).

Mercury deposition patterns in lacustrine sediments of remote and semi-remote lakes of North America and northern Europe reveal that atmospheric Hg has increased by a factor of two to five since the beginning of the industrialized period (Bindler *et al.*, 2001; Lockhart *et al.*, 1998; Lorey and Driscoll, 1999; Lucotte *et al.*, 1995; Swain *et al.*, 1992; Yang *et al.*, 2002). Decreased inputs to midwestern lakes of the US since the 1970s are inferred from sediments in both urban and rural lakes and are attributed to reduced emissions from regional sources (Engstrom and Swain, 1997). In contrast, this decline was not evident in sediments of coastal lakes in southeastern Alaska, implying that global Hg emissions have not abated (Figure 13). Questions have been raised regarding the integrity of sedimentary Hg profiles, as redox conditions affect the form of Hg and the various forms have different mobilities in the sediments. A comparison of lake sediment records of Hg deposition with known histories of Hg contamination in three different Canadian lakes revealed good agreement between the two (Lockhart *et al.*, 2000).

The acidification of aquatic systems via acid deposition was addressed above; however, the nitrogenous component of acid deposition also has the potential to contribute to eutrophication in N-limited aquatic systems. Over the past century, human alteration of the global nitrogen (N) cycle has doubled the amount of fixed nitrogen (predominantly in the forms of ammonium and nitrate) transferred from the atmosphere to land-based ecosystems (Vitousek, 1997). Alpine lakes have served as a focal point for investigating the effects of enhanced N deposition (Psenner and Schmidt, 1992; Saros *et al.*, 2003; Wolfe *et al.*, 2001), because these systems typically have poor buffering capacities, as well

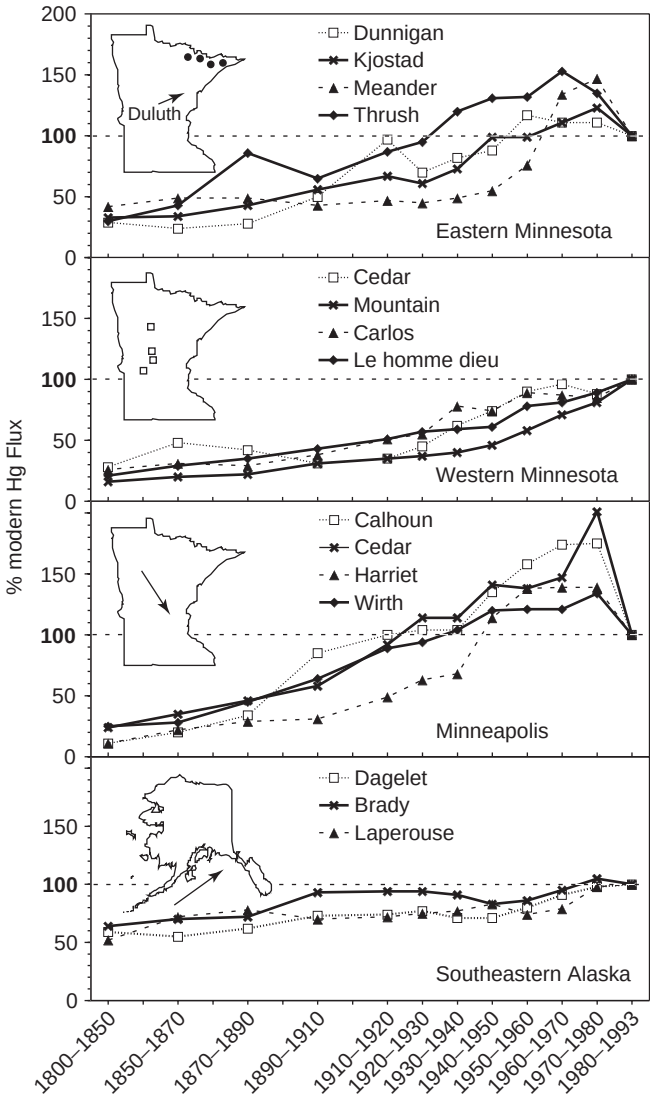


Figure 13 Changes in mercury accumulation rates in the sediments of lakes from 3 regions of Minnesota and from Alaska (Reprinted with permission from Engstrom & Swain 1997. ©1997 American Chemical Society)

as N-limited primary production. In lakes of the Colorado Front Range, shifts in fossil diatom assemblages intensified during the 1950s, with an increase in several diatom taxa typical of mesotrophic lakes (Wolfe *et al.*, 2001). The timing of the shift coincided with the widespread implementation of the Haber–Bosch process for the production of commercial N fertilizers. δ^{15}N analyses of sedimentary organic material in the cores also support the idea that these lakes received increased atmospheric N deposition.

Catchment Processes and Erosion Rates

Reconstruction of catchment erosion rates is possible, based on changes in the accumulation rate of the clastic portion of sediment (Dearing, 1994). Thus, comparison of presettlement versus postsettlement accumulation rates has been used to evaluate the impact of land clearance of varied scale and in different terrains. In the deciduous forests of northwest Europe, catchment sediment yield was low and relatively stable prior to human disturbance but increased about fourfold in the Iron Age and Roman period, when charcoal requirements for smelting of ore led to massive forest clearance and erosion of topsoil. In the Middle Ages, erosion rates returned to predisturbance conditions following reforestation, but accelerated more than 10-fold in progressive deforestation from the eleventh century onward. Interestingly, in this setting, the post-eleventh-century soil erosion rates related to human activities are similar in magnitude to those associated with the cold continental conditions at the end of the last glacial period, in a landscape dominated by unstable soils and tundra or early successional forest. In the Péten region of Guatemala, forest clearance and agriculture during Mayan times eroded topsoils and increased the loading of inorganic material (2–24 times) and phosphorus (1–10 times) into lakes, with increasing sediment accumulation rates, as land use intensified and population numbers increased (Binford *et al.*, 1987). Aquatic productivity apparently decreased, however, possibly because of water column shading associated with a high silt load or, alternatively, reduced availability of phosphorus.

CONCLUSIONS/FUTURE OF THE DISCIPLINE

In the latter decades of the twentieth century, paleolimnology was transformed from a descriptive science with a small number of practitioners to a mature quantitative field with its own journal (Journal of Paleolimnology) and with widespread application to both applied and nonapplied scientific questions. It has played a significant role in developing our understanding of the patterns and magnitude of natural environmental variability of nonmarine systems, their watersheds, and the atmosphere. The paleolimnological record can be used as a tool to describe the nature, magnitude, and rate of response to both natural and human-induced disturbances, and thus it can be used as a yardstick for evaluating the extent to which human-induced changes exceed those produced by natural environmental variation. In this respect, it has been used increasingly in environmental management by federal and state agencies to evaluate human impact, as well as to set suitable targets for restoration efforts. This review has focused on insights about the environment gained from the study of lake sediments and thus uses the term paleolimnology in a restrictive sense. However, many of the generalizations that we make here also apply to other sorts of inland aquatic systems.

The earliest paleolimnological studies focused on describing long-term changes in the lacustrine environment, but in recent decades the paleolimnological record has been used mainly to reconstruct external atmospheric or catchment processes. Thus, in recent times, the paleolimnological record has been underutilized as a tool for understanding how aquatic systems and species respond to perturbations, such as climate change, catchment disturbance, or atmospheric mercury deposition.

The increased emphasis on quantification of trends has led to the proliferation of modern calibration data sets for a variety of organisms and in diverse geographic areas that include both populated and remote regions. These calibration data sets, however, are primarily used to reconstruct past conditions and have not been used to their full potential. For example, although large databases of modern diatom and water-chemistry data are now supported both in the United States and Europe, only cursory attempts have been made to examine diatom distribution across large areas in order to expand our knowledge of the patterns of species distributions and hence their potential ecological and biogeographic controls.

The majority of paleolimnological studies still confine themselves to case studies of an individual lake or small number of lakes. Thus, the extent to which the observed patterns are representative of a broad population of lakes under similar environmental controls is often unclear. Only a relatively small number of studies of twentieth-century human impacts on lakes have utilized stratigraphic data from multiple lakes within a region to quantify the amount and nature of regional variability. Continued generation of statistically robust regional data sets is likely to continue where paleolimnology is applied to lake management at the federal or regional level. Considerably less has been done in terms of describing limnological variation of large regions at longer timescales, and spatial patterns of variation should be considered more explicitly in designing stratigraphic studies. In some cases, the time is ripe for more extensive synthesis of considerable data already generated in individual studies so that we can understand patterns and processes at the landscape scale.

Many paleolimnological records contain a wealth of information, particularly those analyzed at high temporal resolution and that utilize multiple proxies. However,

presently we are unable to interpret many features of stratigraphic records because of an inadequate knowledge of the multiplicity of factors that influence the physical, biological, and chemical structure of lakes. Thus, in many cases we can only speculate about why multiple lakes in a region show different patterns of change through time (Fritz *et al.*, 2000), why reconstructed environments from two different proxies are not identical (Smith *et al.*, 2002), or what the hierarchy of direct or indirect influences are on paleolimnological proxies (Saros and Fritz, 2000) in different settings. Understanding some of the more subtle features of stratigraphic records will require more frequent coupling of paleolimnological and modern process studies to understand the temporal dynamics of lakes. These might include meteorological and hydrologic measurements and modeling to better constrain lacustrine responses to climate (Donovan *et al.*, 2002; Doran *et al.*, 2002; Stone and Fritz, 2004), long-term monitoring of physical, chemical, and biological patterns of change (Interlandi *et al.*, 1999), field experiments of chemical and biological responses to perturbations (Leavitt *et al.*, 1989), and controlled laboratory studies of environmental physiology (Saros and Fritz, 2002).

These inadequately explored issues, such as how lakes respond to disturbance over long timescales, patterns of change at the landscape scale, exploration of species ecology and biogeography, and integration with modern process studies suggest that paleolimnology has incredible untapped potential to yield new insights in the fields of organismal, population, community, and ecosystem ecology, as well as evolution, hydrology, and climatology. The wealth of extant paleolimnological data can be used to describe patterns of variation at various temporal and spatial scales and hence to generate testable hypotheses regarding the behavior of organisms and complex systems. Thus, although paleolimnology has evolved into a mature discipline in its own right, it should strive to explore in greater depth the interactions among the atmosphere, geosphere, biosphere, and hydrosphere and their ultimate drivers, both in the modern world and through geologic history.

REFERENCES

Abbott M.B. and Wolfe A.P. (2003) Intensive pre-Incan metallurgy recorded by lake sediments from the Bolivian Andes. *Science*, **301**, 1893–1894.

Abella S.E.B. (1988) The effect of Mt. Mazama ashfall on the planktonic diatom community of Lake Washington. *Limnology and Oceanography*, **33**, 1376–1385.

Anderson N.J. (1989) A whole-basin diatom accumulation rate for a small eutrophic lake in Northern Ireland and its palaeoecological implications. *Journal of Ecology*, **77**, 926–946.

Anderson N.J. and Rippey B. (1994) Monitoring lake recovery from point-source eutrophication: the use of diatom-inferred epilimnetic total phosphorus and sediment chemistry. *Freshwater Biology*, **32**, 625–639.

Anderson N.J., Rippey B. and Gibson C.E. (1993) A comparison of sedimentary and diatom-inferred phosphorus profiles: implications for defining pre-disturbance nutrient conditions. *Hydrobiologia*, **253**, 357–366.

Appleby P.G. and Oldfield F. (1978) The calculation of lead-210 dates assuming a constant rate of supply of the unsupported lead-210 to the sediment. *Catena*, **5**, 1–8.

Baker P.A., Seltzer G.O., Fritz S.C., Dunbar R.B., Grove M.J., Tapia P.M., Cross S.L., Rowe H.D. and Broda J.P. (2001) The history of South American tropical precipitation for the past 25 000 years. *Science*, **291**, 640–643.

Balogh S.J., Engstrom D.R., Almendinger J.E., Meyer M.L. and Johnson D.K. (1999) History of mercury loading in the upper Mississippi River reconstructed from the sediments of Lake Pepin. *Environmental Science and Technology*, **33**, 3297–3303.

Barker P., Telford R., Merdaci O., Williamson D., Taieb M., Vincens A. and Gibert E. (2000) The sensitivity of a Tanzanian crater lake to catastrophic tephra input and four millennia of climatic change. *The Holocene*, **10**, 303–310.

Battarbee R.W. and Charles D.F. (1987) The use of diatom assemblages in lake sediments as a means of assessing the timing, trends, and causes of lake acidification. *Progress in Physical Geography*, **11**, 552–580.

Battarbee R.W., Flower R.J., Stevenson A.C., Jones V.J., Harriman R. and Appleby P.G. (1988) Diatom and chemical evidence for reversibility of acidification of Scottish lochs. *Nature*, **332**, 530–532.

Bennion H., Juggins S. and Anderson J. (1996) Predicting epilimnetic phosphorus concentrations using an improved diatom – based transfer function and its application to lake eutrophication management. *Environmental Science and Technology*, **30**, 2004–2007.

Berger G.W. and Easterbrook D.J. (1993) Thermoluminescence dating tests for lacustrine, glaciomarine, and floodplain sediments from western Washington and British Columbia. *Canadian Journal of Earth Sciences*, **30**, 1815–1828.

Bigler C., Grahn E., Larocque I., Jeziorski A. and Hall R. (2003) Holocene environmental change at Lake Njulla (999 m a.s.l.), northern Sweden: a comparison with four small nearby lakes along an altitudinal gradient. *Journal of Paleolimnology*, **29**, 13–29.

Bigler C. and Hall R.I. (2003) Diatoms as quantitative indicators of July temperature: a validation attempt at century-scale with meteorological data from northern Sweden. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **189**, 147–160.

Bigler C., Larocque I., Peglar S.M., Birks H.J.B. and Hall R.I. (2002) Quantitative multiproxy assessment of long-term patterns of Holocene environmental change from a small lake near Abisko, northern Sweden. *Holocene*, **12**, 481–496.

Bindler R., Renberg I., Appleby P., Anderson N.J. and Rose N.L. (2001) Mercury accumulation rates and spatial patterns in lake sediments from West Greenland: a coast to ice margin transect. *Environmental Science and Technology*, **35**, 1736–1741.

Binford M., Brenner M., Whitmore T.J., Higuera-Gundy A., Deevey E.S. and Leyden B.S. (1987) Ecosystems, paleoecology

and human disturbance in subtropical and tropical America. *Quaternary Science Reviews*, **6**, 115–128.

Binford M.W. and Deevey E.S. (1983) Paleolimnology: an historical perspective on lacustrine ecosystems. *Annual Review of Ecology and Systematics*, **14**, 255–286.

Binford M., Kolata A.L., Brenner M., Janusek J.W., Seddon M.T., Abbott M. and Curtis J. (1997) Climate variation and the rise and fall of an Andean civilization. *Quaternary Research*, **47**, 235–248.

Birks H.J.B., Frey D.G. and Deevey E.S. (1998) Numerical tools in paleolimnology - progress, potentialities, and problems. *Journal of Paleolimnology*, **20**, 307–332.

Birks H.J.B., Line J.M., Juggins S., Stevenson A.C. and terBraak C.J.F. (1990) Diatoms and pH reconstruction. *Royal Society of London, Philosophical Transactions*, **327**, 263–278.

Birks H.J.B. and Lotter A.F. (1994) The impact of the Laacher See Volcano (11 000 year B.P.) on terrestrial vegetation and diatoms. *Journal of Paleolimnology*, **11**, 313–322.

Björck S., Rundgren M., Ingolfsson O. and Funder S. (1997) The Preboreal oscillation around the Nordic Seas: terrestrial and lacustrine responses. *Journal of Quaternary Science*, **12**, 455–465.

Björck S. and Wohlfarth B. (2001) ^{14}C chronostratigraphic techniques in paleolimnology. In *Tracking Environmental Change Using Lake Sediment*, Smol J.P. (Ed.), Kluwer: Dordrecht, pp. 205–246.

Bormann F.H., Likens G.E., Siccama T.G., Pierce R.S. and Eaton J.S. (1974) The export of nutrients and recovery of stable conditions following deforestation at Hubbard Brook. *Ecological Monographs*, **44**, 255–277.

Boucherle M., Smol J.P., Oliver T.C., Brown S.R. and McNeely R. (1986) Limnologic consequences of the decline in hemlock 4800 years ago in three Southern Ontario lakes. *Hydrobiologia*, **143**, 217–225.

Bradbury J.P. (1975) Diatom stratigraphy and human settlement in Minnesota. *Geological Society of America, Special Paper*, **171**, 1–74.

Bradbury P.J. (1997) A diatom – based paleohydrologic record of climate change for the past 800 k.y. from Owens Lake, California. *Geological Society of America, Special Paper*, **317**, 99–112.

Brännvall M.-L., Bindler R., Emteryd O. and Renberg I. (2001) Four thousand years of atmospheric lead pollution in northern Europe: a summary from Swedish lake sediments. *Journal of Paleolimnology*, **25**, 421–435.

Brännvall M.-L., Bindler R., Renberg I., Emteryd O., Bartnicki J. and Billstrom K. (1999) The Medieval metal industry was the cradle of modern large-scale atmospheric lead pollution in northern Europe. *Environmental Science and Technology*, **33**, 4391–4395.

Charles D.F., Battarbee R.W., Renberg I., van Dam H. and Smol J.P. (1989) Paleoecological analysis of lake acidification trends in North America and Europe using diatoms and chrysophytes. In *Acid Precipitation*, Norton S.A., Lindberg S.E. and Page A.L. (Eds.), Springer-Verlag: New York, pp. 207–276.

Charles D.F. and Norton S.A. (1986) Paleolimnological evidence for trends in atmospheric deposition of acids and metals. *Acid Deposition: Long Term Trends*, National Academy of Sciences Press: Washington, pp. 335–411.

Cohen A.S. (2003) *Paleolimnology: The History and Evolution of Lake Systems*, Oxford University Press: New York.

Colman S.M., Peck J.A., Karabanov E.B., Carter S.J., Bradbury J.P., King J.W. and Williams D.F. (1995) Continental climate response to orbital forcing from biogenic silica records in Lake Baikal. *Nature*, **378**, 769–771.

Cumming B.F., Davey K.A., Smol J.P. and Birks H.J.B. (1994) When did acid – sensitive Adirondack Lakes (New York, USA) begin to acidify and are they still acidifying? *Canadian Journal of Fisheries and Aquatic Sciences*, **51**, 1550–1568.

Dearing J. (1994) Reconstructing the history of soil erosion. In *The Changing Global Environment*, Roberts N. (Ed.), Blackwell: Cambridge, pp. 242–261.

Deevey E.S. (1942) Studies on Connecticut lake sediments. III. The biostratonomy of Linsley Pond. *American Journal of Science*, **240**, 233–264.

Digerfeldt G., Almendinger J.E. and Björck S. (1993) Reconstruction of past lake levels and their relation to groundwater hydrology in the Parkers Prairie sandplain, west-central Minnesota. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **94**, 99–118.

Dixit S.S., Cumming B.F., Birks H.J.B., Smol J.P., Kingston J.C., Uutala A.J., Charles D.F. and Camburn K.E. (1993) Diatom assemblages from Adirondack lakes (New York, USA) and the development of inference models for retrospective assessment. *Journal of Paleolimnology*, **8**, 27–47.

Dixit A.S., Dixit S.S. and Smol J.P. (1992) Algal microfossils provide high temporal resolution of environmental trends. *Water, Air, & Soil Pollution*, **62**, 75–87.

Dixit S.S., Smol J.P., Charles D.F., Hughes R.M., Paulsen S.G. and Collins G.B. (1999) Assessing water quality changes in the lakes of the northeastern United States using sediment diatoms. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 131–152.

Donovan J.J., Smith A.J., Panek V.A., Engstrom D.R. and Ito E. (2002) Climate-driven hydrologic transients in lake sediment records: calibration of groundwater conditions using 20th century drought. *Quaternary Science Reviews*, **21**, 605–624.

Doran P.T., Priscu J.C., Lyons W.B., Walsh J.E., Fountain A.G., McKnight D.M., Moorhead D.L., Virginia R.A., Wall D.H., Clow G.D., *et al.* (2002) Antarctic climate cooling and terrestrial ecosystem response. *Nature*, **415**, 517–520.

Drake D.C. and Naiman R.J. (2000) An evaluation of restoration efforts in fishless lakes stocked with exotic trout. *Conservation Biology*, **14**, 1807–1820.

Douglas M.S.V., Smol J.P. and Blake W. (1994) Marked post-18th century environmental change in high Arctic ecosystems. *Science*, **266**, 416–419.

Edmondson W.T. (1981) The effect of changes in the nutrient income on the condition of Lake Washington. *Limnology and Oceanography*, **26**, 1–29.

Ekdahl E.J., Teranes J.L., Guilderson T.P., Turton C.L., McAndrews J.H., Wittkop C.A. and Stoermer E.F. (2004) Prehistorical record of cultural eutrophication from Crawford Lake, Canada. *Geology*, **32**, 745–748.

Engstrom D.R., Fritz S.C., Almendinger J.E. and Juggins S. (2000) Chemical and biological trends during lake evolution in recently deglaciated terrain. *Nature*, **408**, 161–166.

Engstrom D.R. and Swain E.B. (1986) The chemistry of lake sediments in time and space. *Hydrobiologia*, **143**, 37–44.

Engstrom D.R. and Swain E.B. (1997) Recent declines in atmospheric mercury deposition in the Upper Midwest. *Environmental Science and Technology*, **31**, 960–967.

Engstrom D.R., Swain E.B. and Kingston J.C. (1985) A paleolimnological record of human disturbance from Harvey's Lake, Vermont: geochemistry, pigments, and diatoms. *Freshwater Biology*, **15**, 261–288.

Ford M.S.J. (1990) A 10 000-year. history of natural ecosystem acidification. *Ecological Monographs*, **60**, 57–89.

Fritz S.C. (1989) Lake development and limnological response to prehistoric and historic land-use in Diss, Norfolk, U.K. *Journal of Ecology*, **77**, 182–202.

Fritz S.C. (1990) Twentieth-century salinity and water-level fluctuations in Devils Lake, N. Dakota: a test of a diatom-based transfer function. *Limnology and Oceanography*, **35**, 1771–1781.

Fritz S.C. (1996) Paleolimnological records of climate change in North America. *Limnology and Oceanography*, **41**, 882–889.

Fritz S.C., Baker P.A., Lowenstein T.K., Seltzer G.O., Rigsby C.A., Dwyer G.S., Tapia P.M., Arnold K.K., Ku T. -L. and Luo S. (2004a) Hydrologic variation during the last 170 000 years in the southern hemisphere tropics of South America. *Quaternary Research*, **61**, 95–104.

Fritz S.C., Cumming B.F., Gasse F. and Laird K.R. (1999) Diatoms as indicators of hydrologic and climatic change in saline lakes. In *The Diatoms: Applications for the Environmental and Earth Sciences*, Stoermer E.F. and Smol J.P. (Eds.), Cambridge University Press: pp. 41–72.

Fritz S.C., Engstrom D.R. and Juggins S. (2004b) Patterns of early lake evolution in boreal landscapes: a comparison of stratigraphic inferences with a modern chronosequence in Glacier Bay, Alaska. *The Holocene*, **14**, 828–840.

Fritz S.C., Ito E., Yu Z., Laird K.R. and Engstrom D.R. (2000) Hydrologic variation in the northern Great Plains during the last two millennia. *Quaternary Research*, **53**, 175–184.

Fritz S.C., Kingston J.C. and Engstrom D.R. (1993) Quantitative trophic reconstruction from sedimentary diatom assemblages: a cautionary tale. *Freshwater Biology*, **30**, 1–23.

Fritz S.C., Stevenson A.C., Patrick S.T., Appleby P., Oldfield F., Rippey B., Natkanski J. and Battarbee R.W. (1989) Paleolimnological evidence for the recent acidification of Llyn Hir, Dyfed, Wales. *Journal of Paleolimnology*, **2**, 245–262.

Gasse F. (2000) Hydrological changes in the African tropics since the last Glacial Maximum. *Quaternary Science Reviews*, **19**, 189–211.

Glew J.R., Smol J.P. and Last W.M. (2001) Sediment core collection and extrusion. In *Tracking Environmental Change Using Lake Sediments*, Vol. 1, Last W.M. and Smol J.P. (Eds.), Kluwer: Dordrecht, pp. 73–106.

Goldman C.R. (1981) Lake Tahoe: two decades of change in a nitrogen deficient oligotrophic lake. *International Vereinigung für Theoretische und Angewandte Limnologie, Verhandlungen*, **21**, 45–70.

Hall R.I., Leavitt P.R., Quinlan R., Dixit A.S. and Smol J.P. (1999) Effects of agriculture, urbanization, and climate on water quality in the northern Great Plains. *Limnology and Oceanography*, **44**, 739–756.

Hall R.I. and Smol J.P. (1993) The influence of catchment size on lake trophic status during the hemlock decline and recovery (4800 to 3500 BP) in southern Ontario lakes. *Hydrobiologia*, **269/270**, 371–390.

Hall R.I. and Smol J.P. (1996) Paleolimnological assessment of long – term water – quality in south – central Ontario lakes affected by cottage development and acidification. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 1–17.

Haworth E.Y. (1980) Comparison of continuous phytoplankton records with the diatom stratigraphy in the recent sediments of Blelham Tarn. *Limnology and Oceanography*, **25**, 1093–1103.

Hodell D.A., Brenner M., Curtis J.H. and Guilderson T. (2001) Solar forcing of drought frequency in the Maya lowlands. *Science*, **292**, 1367–1370.

Hodell D.A. and Schelske C.L. (1998) Production, sedimentation, and isotopic composition of organic matter in Lake Ontario. *Limnology and Oceanography*, **43**, 200–214.

Hodell D.A., Schelske C.L., Fahnenstiel G. and Robbins L.L. (1998) Biologically induced calcite and its isotopic composition in Lake Ontario. *Limnology and Oceanography*, **43**, 187–199.

Holmes J.A. (2001) Ostracoda. In *Tracking Environmental Change Using Lake Sediments*, Vol. 3, Smol J.P., Birks H.J.B. and Last W.M. (Eds.), Kluwer: Dordrecht, pp. 125–151.

Hughen K.A., Jiang H.B., Overpeck J.T. and Anderson R.F. (2000) Recent warming in a 500-year palaeotemperature record from varved sediments, upper Soper Lake, Baffin Island, Canada. *The Holocene*, **10**, 9–20.

Hutchinson G.E. (1970) Ianula: an account of the history and development of the Lago di Monterosi, Latium, Italy. *American Philosophical Society, Transactions*, **60**, 1–178.

Interlandi S.J., Kilham S.S. and Theriot E.C. (1999) Responses of phytoplankton to varied resource availability in large lakes of the greater Yellowstone ecosystem. *Limnology and Oceanography*, **44**, 668–682.

Israelson C., Björck S., Hawkesworth C.J. and Possnert G. (1997) Direct U-Th dating of organic- and carbonate-rich lake sediments from southern Scandinavia. *Earth and Planetary Science Letters*, **153**, 251–263.

Ito E. (2001) Application of stable isotope techniques to inorganic and biogenic carbonates. In *Tracking Environmental Change Using Lake Sediments*, Vol. 2, Last W.M. and Smol J.P. (Eds.), Kluwer: Dordrecht, pp. 351–371.

Jacobson H.A. and Engstrom D.R. (1989) Resolving the chronology of recent lake sediments: an example from Devils Lake, North Dakota. *Journal of Paleolimnology*, **2**, 81–98.

Johnson T.C., Scholz C.A., Talbot M.R., Kelts K., Ricketts R.D., Ngobi G., Beuning K., Ssemmanda I. and McGill J.W. (1996) Late Pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes. *Science*, **273**, 1091–1092.

Jones V.J., Stevenson A.C. and Battarbee R.W. (1989) Acidification of lakes in Galloway, south west Scotland: a diatom and pollen study of the post-glacial history of Round Loch of Glenhead. *Journal of Ecology*, **77**, 1–23.

Joynt E.H. and Wolfe A.P. (2001) Paleoenvironmental inference models from sediment diatom assemblages in Baffin Island

lakes (Nunavut, Canada) and reconstruction of summer water temperature. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1222–1243.

King J. and Peck J. (2001) Use of paleomagnetism in studies of lake sediments. In *Tracking Environmental Change Using Lake Sediments*, Vol. 2, Last W.M. and Smol J.P. (Eds.), Kluwer: Dordrecht, pp. 371–390.

Laird K. and Cumming B. (2001) A regional paleolimnological assessment of the impact of clear-cutting on lakes from the central interior of British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 492–505.

Laird K., Cumming B. and Nordin R. (2001) A regional paleolimnological assessment of the impact of clear-cutting on lakes from the west coast of Vancouver Island, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 479–491.

Laird K.R., Fritz S.C. and Cumming B.F. (1998) A diatom-based reconstruction of drought intensity, duration, and frequency from Moon Lake, North Dakota: a sub-decadal record of the last 2300 years. *Journal of Paleolimnology*, **19**, 161–179.

Lamoureux S. (2001) Varve chronology techniques. In *Tracking Environmental Change Using Lake Sediments*, Vol. 1, Last W.M. and Smol J.P. (Eds.), Kluwer: Dordrecht, pp. 247–260.

Lamoureux S.F. and Bradley R.S. (1996) A late Holocene varved sediment record of environmental change from northern Ellesmere Island, Canada. *Journal of Paleolimnology*, **16**, 239–255.

Lamoureux S.F., England J.H., Sharp M.J. and Bush A.B.G. (2001) A varve record of increased 'Little Ice Age' rainfall associated with volcanic activity, Arctic Archipelago, Canada. *Holocene*, **11**, 243–249.

Last W.M. and Smol J.P. (2001a) *Tracking Environmental Change Using Lake Sediments*, *Volume 1: Basin Analysis, Coring, and Chronological Techniques*, Kluwer: Dordrecht.

Last W.M. and Smol J.P. (2001b) *Tracking Environmental Change Using Lake Sediments*, *Volume 2: Physical and Geochemical Methods*, Kluwer: Dordrecht.

Leavitt P.R., Carpenter S.R. and Kitchell J.F. (1989) Whole-lake experiments: the annual record of fossil pigments and zooplankton. *Limnology and Oceanography*, **34**, 700–717.

Leavitt P.R., Vinebrooke R.D., Donald D.B., Smol J.P. and Schindler D.W. (1997) Past ultraviolet radiation environments in lakes derived from fossil pigments. *Nature*, **388**, 457–459.

Leroy S.A.G. and Colman S. (2001) Coring and drilling equipment and procedures for recovery of long lacustrine sequences. In *Tracking Environmental Change Using Lake Sediments*, Vol. 1, Last W.M. and Smol J.P. (Eds.), Kluwer: Dordrecht, pp. 107–136.

Levesque A.J., Cwyner L.C. and Walker I.R. (1997) Exceptionally steep north-south gradients in lake temperatures during the last deglaciation. *Nature*, **385**, 423–426.

Liu K.-b and Fearn M.L. (2000) Reconstruction of prehistoric landfall frequencies of catastrophic hurricanes in northwestern Florida from lake sediment records. *Quaternary Research*, **54**, 238–245.

Lockhart W.L., MacDonald R.W., Outridge P.M., Wilkinson P., DeLaronde J. and Rudd J.W.M. (2000) Tests of the fidelity of lake sediment core records of mercury deposition to known histories of mercury contamination. *The Science of the Total Environment*, **260**, 171–180.

Lockhart W.L., Wilkinson P., Billeck B.N., Danell R.A., Hunt R.V., Brunskill G.J., DeLaronde J. and St. Louis V. (1998) Fluxes of mercury to lake sediments in central and southern Canada inferred from dated sediment cores. *Biogeochemistry*, **40**, 163–173.

Loope D.B., Swinehart J.B. and Mason J.P. (1995) Dune-dammed paleovalleys of the Nebraska Sand Hills: intrinsic versus climatic controls on the accumulation of lake and marsh sediments. *Geological Society of America Bulletin*, **107**, 396–406.

Lorey P. and Driscoll C.T. (1999) Historical trends of mercury deposition in Adirondack lakes. *Environmental Science and Technology*, **33**, 718–722.

Lotter A.F. (1998) The recent eutrophication of Baldeggersee (Switzerland) as assessed by fossil diatom assemblages. *The Holocene*, **8**, 395–406.

Lotter A.F., Birks H.J.B., Hofmann W. and Marchetto A. (1997) Modern diatom, cladocera, chironomid, and chrysophyte cyst assemblages as quantitative indicators for the reconstruction of past environmental conditions in the Alps. I. Climate. *Journal of Paleolimnology*, **18**, 395–420.

Lucotte M.A., Mucci A., Hillaire-Marcel C., Pichet P. and Grondin A.E. (1995) Anthropogenic mercury enrichment in remote lakes of northern Quebec (Canada). *Water, Air, & Soil Pollution*, **80**, 467–476.

Maberly S.C., Hurley M.A., Butterwick C., Corry J.E., Heaney S.I., Irish A.E., Jaworski G.H.M., Lund J.W.G., Reynolds C.S. and Roscoe J.V. (1994) The rise and fall of *Asterionella formosa* in the South Basin of Windermere: analysis of a 45-year series of data. *Freshwater Biology*, **31**, 19–34.

Magnuson J.J., Robertson D.M., Benson B.J., Wynne R.H., Livingstone D.M., Arai T., Assel R.A., Barry R.G., Card V., Kuusisto E. *et al.* (2000) Historical trends in lake and river ice cover in the northern hemisphere. *Science*, **289**, 1743–1746.

Nicholson S.E. (1998) Historical fluctuations of Lake Victoria and other lakes in the northern Rift Valley of East Africa. In *Environmental Change and Response in East African Lakes*, Lehman J.T. (Ed.), Kluwer: Dordrecht, pp. 7–35.

Noren A.J., Bierman P.R., Steig E.J., Lini A. and Southon J. (2002) Millennial-scale storminess variability in the northeastern United States during the Holocene epoch. *Nature*, **419**, 821–824.

Overpeck J.P., Hugen K., Hardy D., Bradley R., Case R., Douglas M., Finney B., Gajewski K., Jacoby G., Jennings A., *et al.* (1997) Arctic environmental change of the last four centuries. *Science*, **278**, 1251–1256.

Patrick R., Binetti V.P. and Halterman S.G. (1981) Acid lakes from natural and anthropogenic causes. *Science*, **211**, 446–448.

Pearsall W.H. (1932) Phytoplankton in English lakes. *Journal of Ecology*, **20**, 242–262.

Pennington W. (1981) Records of a lake's life in time: the sediments. *Hydrobiologia*, **79**, 197–219.

Pennington W., Haworth E.Y., Bonny A.P. and Lishman J.P. (1972) Lake sediments in northern Scotland. *Royal Society of London, Philosophical Transactions*, **264**, 191–294.

Pienitz R. and Smol J.P. (1993) Diatom assemblages and their relationship to environmental variables in lakes from the boreal forest-tundra ecotone near Yellowknife, Northwest Territories, Canada. *Hydrobiologia*, **269/270**, 391–404.

Pienitz R., Smol J.P. and Birks H.J.B. (1995) Assessment of freshwater diatoms as quantitative indicators of past climatic change in the Yukon and Northwest Territories, Canada. *Journal of Paleolimnology*, **13**, 21–49.

Pienitz R., Smol J.P. and MacDonald G.M. (1999) Paleolimnological reconstruction of Holocene climatic trends from two boreal treeline lakes, Northwest Territories, Canada. *Arctic, Antarctic, and Alpine Research*, **31**, 82–93.

Psenner R. and Schmidt R. (1992) Climate-driven pH control of remote alpine lakes and effects of acid deposition. *Nature*, **356**, 781–783.

Radle N.J., Keister C.M. and Battarbee R.W. (1989) Diatom, pollen, and geochemical evidence for the paleosalinity of Medicine Lake, S. Dakota, during the Late Wisconsin and early Holocene. *Journal of Paleolimnology*, **2**, 159–172.

Ramstack J.M., Fritz S.C. and Engstrom D.R. (2004) Twentieth-century water-quality trends in Minnesota lakes compared with pre-settlement variability. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**, 561–575.

Ramstack J.M., Fritz S.C., Engstrom D.R. and Heiskary S.A. (2003) The application of a diatom-based transfer function to evaluate regional water-quality trends in Minnesota since 1970. *Journal of Paleolimnology*, **29**, 79–94.

Reed J.M., Roberts N. and Leng M.J. (1999) An evaluation of the diatom response to Late Quaternary environmental change in two lakes in the Konya Basin, Turkey, by comparison with stable isotope data. *Quaternary Science Reviews*, **18**, 631–646.

Renberg I. (1990) A 12 600 year perspective on the acidification of Lilla Oresjon, southwest Sweden. *Philosophical Transactions, Royal Society of London, Series B*, **327**, 357–361.

Renberg I., Bindler R., Bradshaw E., Emteryd O. and McGowan S. (2001) Sediment evidence of early eutrophication and heavy metal pollution of Lake Malaren, Central Sweden. *Ambio*, **30**, 496–502.

Renberg I., Persson M.W. and Emteryd O. (1994) Pre-industrial atmospheric lead contamination detected in Swedish lake sediments. *Science*, **268**, 323–326.

Renberg I. and Wik M. (1984) Dating recent lake sediments by soot particle counting. *International Vereinigung für Theoretische und Angewandte Limnologie, Verhandlungen*, **22**, 712–718.

Rodbell D., Seltzer G.O., Anderson D.M., Abbott M.B., Enfield D.B. and Newman J.H. (1999) An 15,000 -year record of El Nino-driven alluviation in southwestern Ecuador. *Science*, **283**, 516–520.

Rosén P., Segerstrom U., Eriksson L., Renberg I. and Birks H.J.B. (2001) Holocene climatic change reconstructed from diatoms, chironomids, pollen and near-infrared spectroscopy at an alpine lake (Sjuodjijaure) in Northern Sweden. *The Holocene*, **11**, 551–562.

Round F.E. (1961) The diatoms of a core from Esthwaite Water. *New Phytologist*, **60**, 43–59.

Sarna-Wojicki A.M., Champion D.E. and Davis J.O. (1983) Holocene volcanism in the conterminous United States and the role of silicic volcanic ash layers in correlation of latest - Pleistocene and Holocene Deposits. In *Late Quaternary Environments of the United States: The Holocene*, Wright H.E. (Ed.), University of Minnesota Press: Minneapolis, pp. 52–77.

Saros J.E. and Fritz S.C. (2000) Nutrients as a link between ionic concentration/composition and diatom distributions in saline lakes. *Journal of Paleolimnology*, **23**, 449–453.

Saros J.E. and Fritz S.C. (2002) Resource competition among saline-lake diatoms under varying N/P ratio, salinity and anion composition. *Freshwater Biology*, **47**, 87–95.

Saros J.E., Interlandi S.J., Wolfe A.P. and Engstrom D.R. (2003) Recent changes in the diatom community structure of lakes in the Beartooth Mountain Range (USA). *Arctic, Antarctic, & Alpine Research*, **35**, 18–23.

Schelske C.L. and Hodell D.A. (1991) Recent changes in productivity and climate of Lake Ontario detected by isotopic analysis of sediments. *Limnology and Oceanography*, **36**, 961–975.

Scholz C.A. and Rosendahl B.R. (1988) Low lake stands in lakes Malawi and Tanganyika, East Africa, delineated with multifold seismic data. *Science*, **240**, 1645–1648.

Seltzer G.O., Baker P.A., Cross S., Dunbar R. and Fritz S.C. (1998) High resolution seismic reflection profiles from Lake Titicaca, Peru-Bolivia: evidence for Holocene aridity in the tropical Andes. *Geology*, **26**, 167–170.

Siegenthaler U., Eicher U., Oeschger H. and Dansgaard W. (1984) Comparison of the late glacial oxygen isotope records for sediments in a small Swiss lake with that for Dye 3 Greenland ice core. *Annals of Glaciology*, **5**, 149–152.

Siver P.A., Lott A.M., Cash E., Moss J. and Mariscano L.J. (1999) Century changes in Connecticut, U.S.A., lakes as inferred from siliceous algal remains and their relationships to land-use change. *Limnology and Oceanography*, **44**, 1928–1935.

Smith A.J. (1993) Lacustrine ostracodes as hydrochemical indicators in lakes of the north-central United States. *Journal of Paleolimnology*, **8**, 121–134.

Smith A.J., Donovan J.J., Ito E., Engstrom D.R. and Panek V.A. (2002) Climate-driven hydrologic transients in lake sediment records: multiproxy record of mid-Holocene drought. *Quaternary Science Reviews*, **21**, 625–646.

Smol J.P. (2002) *Pollution of Lakes and Rivers: A Paleoenvironmental Perspective*, Edward Arnold: London.

Smol J.P., Birks H.J.B. and Last W.M. (2002a) *Tracking Environmental Change Using Lake Sediments*, *Volume 3: Terrestrial, Algal, and Siliceous Indicators*, Kluwer: Dordrecht.

Smol J.P., Birks H.J.B. and Last W.M. (2002b) *Tracking Environmental Change Using Lake Sediments, Volume 4: Zoological Indicators*, Kluwer: Dordrecht.

Smol J.P., Cumming B.F., Dixit A.S. and Dixit S.S. (1998) Tracking recovery patterns in acidified lakes: a paleolimnological perspective. *Restoration Ecology*, **6**, 318–326.

Sommaruga-Wograth S., Koining K.A., Schmidt R., Sommaruga R., Tessadri R. and Psenner R. (1997) Temperature effects on the acidity of remote alpine lakes. *Nature*, **387**, 64–66.

St. Jacques J.-M., Douglas M.S.V. and McAndrews J.H. (2000) Mid-Holocene hemlock decline and diatom communities in van

Nostrand Lake, Ontario, Canada. *Journal of Paleolimnology*, **23**, 385–397.

Stine S. (1994) Extreme and persistent drought in California and Patagonia during medieval time. *Nature*, **369**, 546–549.

Stoermer E.F., Wolin J.A., Schelske C.L. and Conley D.J. (1985) An assessment of ecological changes during the recent history of Lake Ontario based on siliceous algal microfossils preserved in the sediments. *Journal of Phycology*, **21**, 257–276.

Stone J.R. and Fritz S.C. (2004) Three-dimensional modeling of lacustrine diatom habitat areas: improving paleolimnological interpretation of planktic:benthic ratios. *Limnology and Oceanography*, **49**, 1540–1548.

Swain E.B., Engstrom D.R., Brigham M.E., Henning T.A. and Brezonik P.A. (1992) Increasing rates of atmospheric mercury deposition in midcontinental North America. *Science*, **257**, 784–787.

ter Braak C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

Turney C.S.M. and Lowe J.J. (2001) Tephrochronology. In *Tracking Environmental Change Using Lake Sediments*, Vol. 1, Last W.M. and Smol J.P. (Eds.), Kluwer: Dordrecht, pp. 451–471.

Ugolini F.C. and Mann D.H. (1979) Biopedological origin of peatlands in southeast Alaska. *Nature*, **281**, 366–368.

Verschuren D. and Marnell L.F. (1997) Fossil zooplankton and the historical status of Westslope Cutthroat Trout in the headwater lake of Glacier National Park, Montana. *Transactions of the American Fisheries Society*, **126**, 21–34.

Verschuren D., Laird K.R. and Cumming B.F. (2000) Rainfall and drought in equatorial east Africa during the past 1 100 years. *Nature*, **403**, 410–414.

Vitousek P.M. (1997) Human alteration of the global nitrogen cycle: sources and consequences. *Ecological Applications*, **7**, 737–750.

von Grafenstein U., Erlenkeuser H., Brauer A., Jouzel J. and Johnsen S.J. (1999) A mid-European decadal isotope climate record from 15500 to 5000 years BP. *Science*, **284**, 1654–1657.

Walker I.R., Smol J.P., Engstrom D.R. and Birks H.J.B. (1991) An assessment of Chironomidae as quantitative indicators of past climatic change. *Canadian Journal of Fisheries and Aquatic Sciences*, **48**, 975–987.

Wetzel R.G. (2001) *Limnology*, Academic Press: San Diego.

Whitehead D.R., Charles D.F., Jackson S.T., Smol J.P. and Engstrom D.R. (1989) The developmental history of Adirondack (N.Y.) lakes. *Journal of Paleolimnology*, **2**, 185–206.

Wolfe A.P. (2002) Climate modulates the acidity of Arctic lakes on millennial time scales. *Geology*, **30**, 215–218.

Wolfe A.P., Baron J.S. and Cornett R.J. (2001) Anthropogenic nitrogen deposition induces rapid ecological changes in alpine lakes of the Colorado Front Range (USA). *Journal of Paleolimnology*, **25**, 1–7.

Yang H., Rose N.L., Battarbee R.W. and Boyle J.F. (2002) Mercury and lead budgets for Lochnagar, a Scottish mountain lake and its catchment. *Environmental Science and Technology*, **36**, 1383–1388.

PART 10

Rainfall-runoff Processes

111: Rainfall Excess Overland Flow

ROGER E SMITH[1] AND DAVID C GOODRICH[2]

[1] *Civil Engineering Department, Colorado State University, Fort Collins, CO, US*
[2] *Southwest Watershed Research Center, ARS USDA, Tucson, AZ, US*

One of the processes that can generate surface runoff is rainfall excess, which is a process controlled at the surface of the soil. This occurs when rainfall reaches the soil at a rate in excess of the soil's ability to absorb, which is called infiltrability. This dynamic property can in uniform soil be described by a rather well-developed infiltration theory. Surface water flow toward a receiving channel may in geometrically simple conditions be described by the kinematic or diffusive wave equations. The surface water is in continuous interaction with the soil's changing infiltrability. Both infiltration theory and surface flow equations are introduced here, and the interactions and complexities arising from spatial variations are discussed. These processes are incorporated in modern hydrologic response models, using numerical solutions as well as analytic solutions. The application of this theory in hydrology, however, must be informed by scale considerations and the appropriate treatment of natural complexities. Some scale limitations and some approximate methods for treating spatial soil variations are illustrated in this article, with reference to relevant literature.

INTRODUCTION

Runoff generated from storm rainfall is largely determined by local interaction of the properties of the rainfall, ground cover, land use, and soil (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3** and **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**). While vegetation, above ground cover, and land use play a critical role in the fate of rainfall, this article focuses on the interactions between rainfall and surface soils. Here we discuss the physical dynamics of those cases where rainfall generates surface runoff at the point where it falls. *Infiltration excess* occurs when the rainfall comes at a rate higher than the rate at which the otherwise unsaturated soil can absorb water. (*Saturation excess*, discussed below, occurs when the soil is saturated or filled with water from a subsoil restriction, such as a shallow bedrock. This mechanism is not rainfall rate dependent.) The runoff or surface water flow resulting from infiltration limitation is called *infiltration excess* runoff, and the (variable) limiting soil intake rate is called the soil *infiltrability*. Infiltration excess runoff is often also referred to as *Hortonian* runoff, after Robert Horton (Horton, 1933).

The infiltration and saturation excess–generating mechanisms are not mutually exclusive on a watershed, nor even mutually exclusive at a point on a watershed. The rainfall rate may exceed the infiltrability for some storms, and for others the rain may come slowly until the surface soil layer is saturated. Climate and geography will determine which mechanism is dominant at a given location and time. Figure 1 shows a world map with climate zones indicated. The predominance of relatively short high intensity storms in most subhumid and semiarid zones means that these areas are more prone to infiltration excess runoff. Conversely, saturation excess runoff is more common in humid areas, usually characterized by greater rainfall volumes but with lower intensities. Increasingly, human activity (e.g. urbanization, compaction, etc.) results in an overall decrease of soil infiltrability resulting in globally increasing areas of infiltration excess runoff generation.

Infiltrability changes with many factors, some of which are described quantitatively here. It can also change because of changes in the soil with freezing, thawing, compaction, and tillage. An intense rain on cultivated soil

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

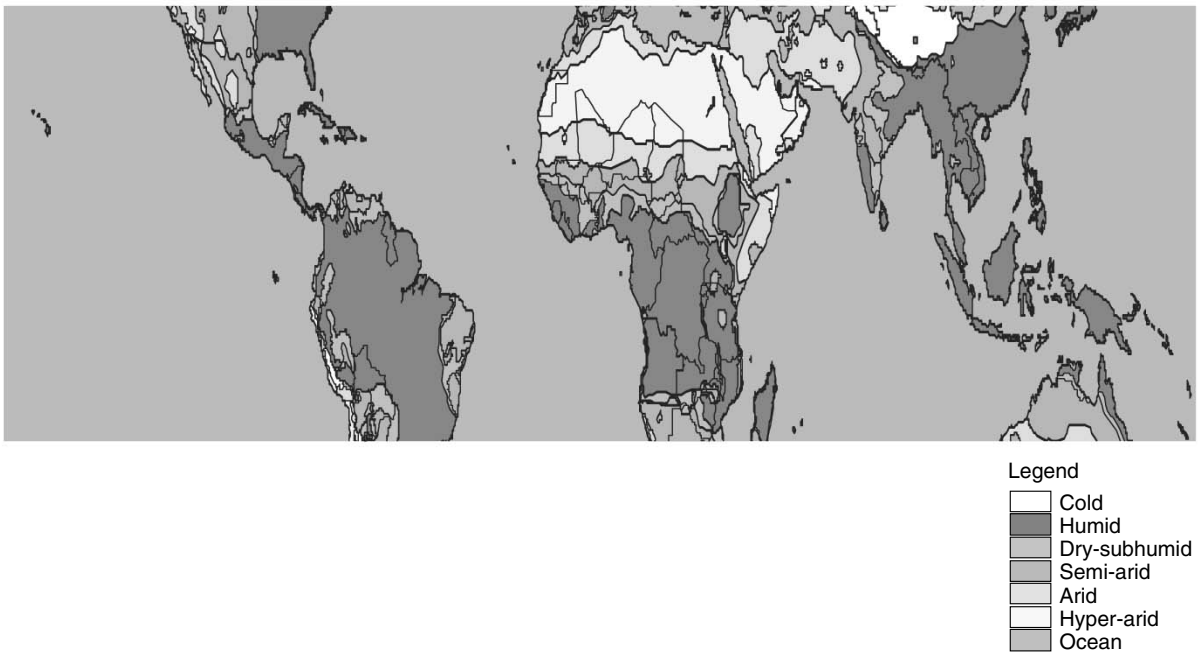


Figure 1 World distribution of land classes, indicating land areas subject to rainfall excess runoff. The largest grey areas correspond to semiarid and arid areas. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

can also create a surface crust or sealed layer. While these processes have been studied and described elsewhere, they are difficult to quantify outside local conditions such as soil cover, and will be omitted from this overview.

The timescale of infiltration and surface flow processes is important, insofar as hydrologic analysis often employs lumping in space and time to approximate the behavior of hydrologic processes. Changes in soil water and changes in runoff rates occur on the order of seconds in some cases, and a timescale of a few minutes is necessary to capture the dynamics of surface runoff. By contrast, the slow movement of water in an aquifer (*see* **Chapter 112, Subsurface Stormflow, Volume 3** and **Chapter 149, Hydrodynamics of Groundwater, Volume 4**) can be characterized by timescales on the order of days or months. Thus, infiltration excess runoff is one of the more dynamic or rapidly changing fluxes in the hydrologic cycle, and as will be seen below, requires knowledge of the temporal pattern of rainfall rates to characterize it accurately.

Historical Notes

Robert Horton was one of the early proponents of the concept that higher intensity rainfall rates on soils of finer texture can exceed the intake rate of the soil and runoff thus reaches upland channels by *overland flow* on the soil surface. He also recognized the mechanism of streamflow generation by flow through the soil mantle. Recent reanalysis of Horton's data from his laboratory watershed in upstate New York indicates, not surprisingly, that the catchment included areas likely to produce a variety of runoff mechanisms, including a small marshy area (Beven, 2004).

Horton termed the limiting soil surface intake rate the "infiltration-capacity" (Horton, 1933). While he recognized that the soil intake rate decayed in value through a storm, he attributed the reduction to surface soil compaction, soil colloidal swelling, or inwashing of fines from the surface (Horton, 1939). Horton made a rather insightful analysis of the hydraulics of overland flow, but his analyses generally used constant infiltration rates. While he was aware of the prescient work of Green and Ampt (1911), he did not believe it related to catchment infiltration:

> "The question naturally arises whether the infiltration-capacity may not vary with the depth of penetration of soil-moisture into a dry soil column. Thus far, the author has not found any definite evidence of an appreciable variation in this respect. The work of Green and Ampt has a bearing on this question although their mathematical analysis of the process appears to be faulty." (Horton, 1936)

Horton further expressed his belief that the counterflow and compression of air negated the ideas of Green and Ampt. Horton's work is a good example of an early appreciation of processes at work in the field, but with a mistaken view of the relative magnitudes involved.

THE DYNAMICS OF INFILTRATION DURING STORM RAINFALL

Unsaturated Soil Hydraulic Properties

Understanding the dynamics of infiltration and runoff requires some understanding of the hydraulics of water in unsaturated soil. Soil is a *porous medium* through which water can flow even when there is air in the spaces between soil grains. Water in unsaturated soil is at a lower pressure than that of the air surrounding it, due to the *capillary pressure*, which is a property of the interface between air and water. This capillary pressure may be expressed as an equivalent (negative) head of water, with the symbol ψ, and is the same force that causes water to rise in an open capillary tube from a free source of water (*see* **Chapter 73, Soil Water Potential Measurement, Volume 2** for a more thorough discussion).

The volumetric water content of a soil, for which the symbol θ [L^3 L^{-3}] is used, varies between zero and the soil porosity, which is the relative volume of the soil not occupied by solid matter. Figure 2(a) illustrates how the soil water pressure decreases dramatically (becomes more negative) as the soil dries. This relation is often referred to as the soil water *retention* relation. Figure 2(b) illustrates by example how the soil water conductivity also falls dramatically as the soil becomes drier. *Hydraulic conductivity, K*, [$L\,T^{-1}$] is defined as the rate of flow of water in soil in response to a unit gradient of hydraulic head. The expression for this is *Darcy's Law*:

$$v = -K\frac{dH}{dz} \tag{1}$$

where v is the rate of flow in units $L\,T^{-1}$, H is hydraulic head in terms of equivalent depth of water (L), and z is the measure of distance (L). The hydraulic head, H, is made up of capillary pressure head and gravitational potential: $H = \psi + z$. Thus, in a homogeneous soil of uniform water content, and thus uniform capillary pressure head (zero capillary head gradient), water may flow downward due to gravity alone. On the other hand, unsaturated flow normal to the gravitational vector moves only in response to capillary head gradients.

In unsaturated soils, as illustrated in Figure 2(b), K is a function of capillary head, $K(\psi)$, expressed as a relative conductivity, *kr,* which is the ratio of $K(\psi)/K(0)$. $K(0)$ is called *saturated hydraulic conductivity*, K_s. K can also be treated as a function of water content, θ, through the retention relation of Figure 2(a). During rainfall infiltration, surface soil water content increases from the initial value, θ_i. For large enough rainfall rates, the surface water can reach *field saturation.* This is the water content at zero capillary head, θ_s. Further discussion of soil capillary properties can be found in article (*see* **Chapter 74, Soil Hydraulic Properties, Volume 2**), in soil physics texts, or in Smith (2002).

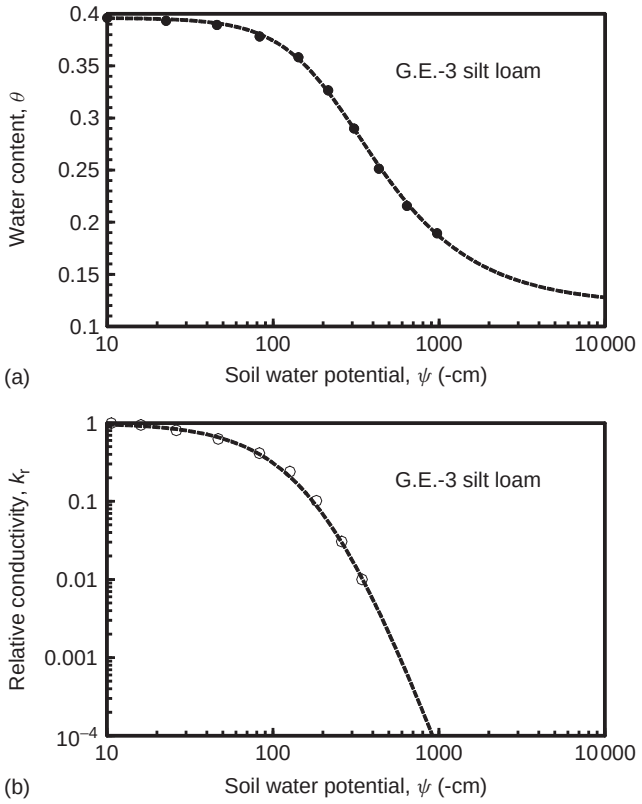


Figure 2 (a) example soil water retention curve, relating water content to soil water potential, and (b) soil relative conductivity relation for the same soil

Physical Basis of Models for Infiltration Under Rainfall

When a hypothetically unlimited supply of water initially reaches the relatively dry soil surface, the wetting at the surface creates extremely large hydraulic head gradients for flow into the soil, and the rate of influx, v, is extremely high, even though overall values of K may be low. As infiltration of rainwater continues, the hydraulic head gradient for absorption decreases, the surface value of K increases, and the rate of infiltration is dramatically reduced. When the capillary gradient at the surface ultimately approaches 0, the only head gradient is gravity, and from equation (1) it can be seen that the ultimate rate of infiltration will equal K_s.

Definitions

Hereafter, we will refer to the vertical infiltration rate at the soil surface with the symbol f. The limiting value of f, given an unlimited supply of surface water, and $\psi = 0$, is called the soil *infiltrability*, with the symbol f_c. Both vary continuously with time during rainfall as the rain intensity

and soil water conditions change. The rainfall rate, varying with time, is *r(t)*. Before a brief presentation of the origin of infiltration relations, we introduce the concept of soil water diffusivity. This concept is analogous to diffusion mathematics, where the flux of a quantity, such as heat, is a function of the gradient of that quantity. In terms of water in the soil, θ, one can use the relation illustrated in Figure 2(a) with equation (1) to obtain

$$v = D(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}z} + K(\theta) \tag{2}$$

by using the definition of diffusivity, D:

$$D(\theta) \equiv K(\theta)\frac{\mathrm{d}\psi}{\mathrm{d}\theta} \tag{3}$$

Approximations for the relation $D(\theta)$ play an important role in the successful derivation of infiltrability relationships given below (Smith, 2002).

Soil Flow Dynamics and Infiltration

At the soil surface, when there is rain, the flow dynamics can be described by combining equation (2) with an expression for continuity of flow, representing the fact that the inflow rate must equal the change in soil water storage at the soil surface:

$$f = \frac{\mathrm{d}}{\mathrm{d}t}\int_0^L (\theta - \theta_\mathrm{i})\,\mathrm{d}z \tag{4}$$

where L is a depth below the advance of the wetting zone, and θ_i is assumed uniform. In this expression, we assume for relative simplicity that the initial water content is small enough for the initial downward (gravitational) flux of water to be negligible. Referring to the infiltrated water depth in the wetted soil adjacent to the surface as I, equation (4) in integral form is

$$I(t) = \int_0^t f\,\mathrm{d}t = \int_0^L (\theta - \theta_\mathrm{i})\,\mathrm{d}z \tag{5}$$

This expression in combination with Darcy's Law in the form of equation (3) yields the infiltration integral, from which at least two basic infiltration models come (see Smith, 2002 for a derivation):

$$I(t) = \int_{\theta_\mathrm{i}}^{\theta_o} \frac{(\theta - \theta_\mathrm{i})D}{v(\theta, t) - K(\theta)}\,\mathrm{d}\theta \tag{6}$$

From equation (6) and realistic assumptions about the form of the highly nonlinear functions $D(\theta)$ and $K(\theta)$ (see Parlange, *et al.*, 1982; Smith, 2002), the expressions for infiltrability are derived in terms of an important integral property of soils, called the *capillary length scale*, or capillary drive, G:

$$G = \frac{1}{K_\mathrm{s}}\int_{-\infty}^{0} K(\psi)\,\mathrm{d}\psi = \int_{-\infty}^{0} k_\mathrm{r}(\psi)\,\mathrm{d}\psi \tag{7}$$

where $\mathrm{K_s}$ is the saturated hydraulic conductivity and k_r is the relative hydraulic conductivity shown in Figure 2(b). This parameter is in effect the k_r-weighted mean value of soil capillary head. Another physically meaningful term that arises in the integration of equation (6) is the *initial water content deficit,* $(\theta_\mathrm{s} - \theta_\mathrm{i})$, hereafter referred to as $\Delta\theta_\mathrm{i}$.

Relations for infiltrability in terms of G, K_s, and the deficit, $\Delta\theta_\mathrm{i}$, include that of Green and Ampt (1911):

$$f_\mathrm{c} = \frac{K_\mathrm{s}(G\Delta\theta_\mathrm{i} + I)}{I} \tag{8}$$

and Smith and Parlange (1978):

$$f_\mathrm{c} = K_\mathrm{s}\left[1 - \exp\left(\frac{-I}{G\Delta\theta_\mathrm{i}}\right)\right]^{-1} \tag{9}$$

The infiltrability relation (8) is derived on the basis of the assumption of a $K(\psi)$ relation that approaches a step function, with K "jumping" from a negligible value to its maximum at some value of ψ (or θ) as the soil wets. This behavior is most like that of a uniform sand or silt. The relation of equation (9) is derived on the basis of the assumption that $K(\psi)$ rises exponentially as ψ increases toward 0. While this does not match the actual measured relation of most real soils, it is in most cases a better approximation than equation (8) (Parlange *et al.*, 1982).

Computational Forms for Infiltration Models

Notably, both equations (8) and (9) express infiltrability in terms of I rather than time, t. While relations between f_c, I, and time, t, can be mathematically derived, the relation of f_c to I is quite important in modeling infiltration during a storm. When a storm is sufficiently intense to create excess, the control on soil f changes from the rainfall to the infiltrability at some point. This is termed the *ponding time*. Owing to the highly nonlinear relation of $D(\theta)$, for most soil hydraulic relations there is a near equality between the infiltration relations $f_\mathrm{c}(I)$ under rainfall and that under unlimited water supply at the surface ("sudden ponding") (Smith, 2002). This allows one to use equations (8) or (9) to predict both the onset of ponding under rainfall, and the decaying function of f_c after that, when the soil exerts control through infiltrability. Because the unifying variable is I, this close approximation has been called the infiltrated depth approximation (IDA) by Smith (2002). It was earlier called the *time condensation* approximation by others (e.g. Sivapalan and Milly, 1989). The principle does not however

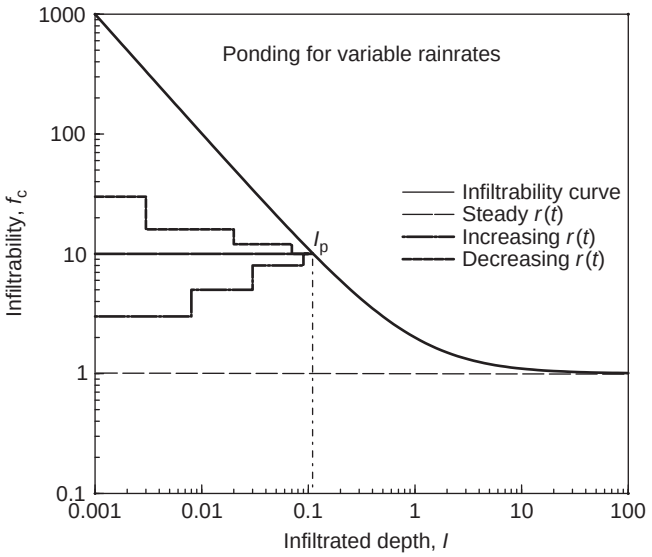


Figure 3 Relation of infiltrability to infiltrated depth, as described by equations (8) or (9), showing how ponding is predicted with this relationship. I_p is the accumulated infiltration depth at ponding

actually relate to the condensation of time. It assumes that the time at which the surface boundary condition changes from rainfall to soil control occurs when the relation between r and I_a matches that for $f_c(I_a)$, where I_a is the accumulated infiltration due to rainfall. This is termed the time of ponding, and subsequent variables with subscript p (such as I_p) refer to the value of the variable at that time. Figure 3 illustrates the fact that in the *f(I)* relation, ponding at I_p may be approached by any number of patterns of rainfall.

Nondimensional Forms

Different soils have a wide range of values of the hydraulic parameters G, K_s, and the initial soil deficit $\Delta\theta_i$ varies between storms. The infiltration equations may be unified and simplified by producing dimensionless forms, scaling I by the value of $G\Delta\theta_i$, and scaling f_c on K_s:

$$f_{c^*} = \frac{f_c}{K_s} \qquad I_* = \frac{I}{G\Delta\theta_i} \tag{10}$$

These dimensionless forms were used in Figure 3. Equation (8), using these variables, is thereby simplified to

$$f_{c^*} = \frac{I_* + 1}{I_*} \tag{11}$$

A dimensionless time may similarly be found:

$$t_* = \frac{t}{G\Delta\theta K_s^{-1}} \tag{12}$$

and rainfall rate may be scaled as for infiltrability: $r_* = r/K_s$. Relations between time and either I_* or f_{c^*} are obtained by substituting the relation $f_* = \mathrm{d}I_*/\mathrm{d}t_*$ into equations (8) or (9). The resulting expressions are implicit in time, but explicit expressions have been presented by Smith (2002), Paige *et al.* (2002), and Li *et al.* (1976), one of which will be presented here.

Infiltrability as a Function of Time

Time-explicit expressions are useful after ponding has been found using the IDA. The simplest approximate expression for post-ponding $f_{c^*}(t_*)$ can be given as (Smith, 2002):

$$f_{c^*} = (1 - \beta) + \sqrt{\beta^2 + \frac{1}{2t'_*}} \tag{13}$$

where β is a weighting parameter. This expression describes a relation intermediate between that of equations (8) and (9) for values of β between 0.5 and 1. Figure 4 illustrates the time-explicit equation (13) and time-f_c relations for both equations (8) and (9). The physics of infiltration requires that at small times the functions must be asymptotic to the line $f_{c^*} = (t_*)^{-1/2}$, and at large times asymptotic to $f_{c^*} = 1$. The variable t' is used to indicate that the time position of this curve must be adjusted to describe f_c after ponding time, t_p, (or ponding depth I_p) so that $f_{c^*} = r_{p^*}$ at $t'_* = t_{p*}$, where r_p is the rainfall rate at which ponding occurs. Using infiltrability expression, equation (9) for example, and solving for $I = I_p$ expressed as an integral of the rainfall rate pattern, the ponding time is found using the

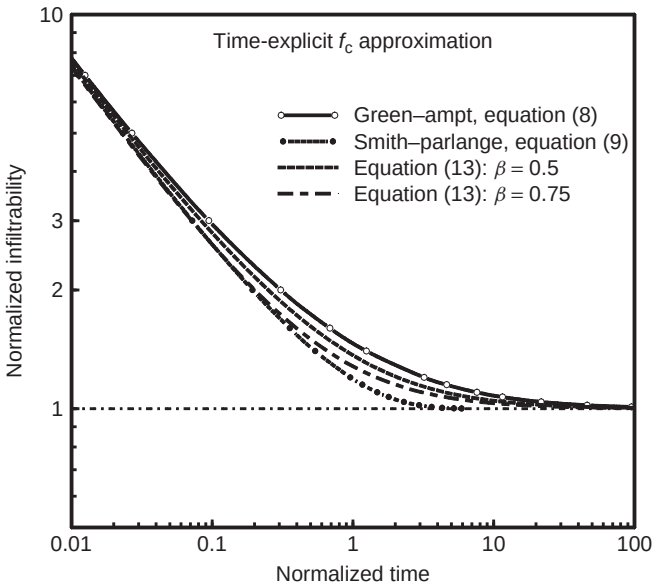


Figure 4 Illustration of the ability of the time-explicit $f_c(t)$ equation (13) to match the behavior of equations (8) and (9) when put in $t(f_c)$ form

IDA method:

$$I_p = \int_0^{t_p} r(t)\,\mathrm{d}t = G\Delta\theta_i \ln\left(\frac{r_*}{r_* - 1}\right) \qquad (14)$$

Clearly, the ponding time $t_p(r)$ can be determined directly for a rainfall of constant rate, for which $I = rt$. After ponding occurs, the infiltrability expression equations (8), (9), or (13) will estimate rainfall excess during the period of time when surface available water, including the rainfall and ponded water, equals or exceeds the infiltrability. This leads to the consideration of the dynamic interaction of infiltration and surface water flow, discussed below.

Irregular Rainfall and Recovery of Infiltrability During Rainfall Hiatus

The above methodology is robust for an unlimited variation in the pattern of rainfall, with the caveat that rainfall must stay greater than K_s, that is, $r_* > 1$. For the more general case, there needs to be an accounting for redistribution of water during breaks in rainfall (periods when $r = 0$), or any time infiltrability exceeds the rainfall. Local conditions can easily be simulated, given the soil hydraulic properties, by solution of the nonlinear convection diffusion equation (Richards' equation; *see* **Chapter 150, Unsaturated Zone Flow Processes, Volume 4** and **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**), but this is not practical for most hydrological catchment models. Some early methods for treating breaks in rainfall patterns simply maintained a value of I through the hiatus (e.g. Mls, 1980). Corradini *et al.* (1997) and Smith *et al.* (1999) have proposed and demonstrated a method of intermediate complexity requiring minimum parameterization of the soil hydraulic properties, which simulates robustly the changes in infiltrability and soil water deficit $\Delta\theta_i$ for periods when rainfall rates fall to any value less than K_s. This method treats the wetting pulse in the soil as a distortable curve with conserved similarity and net volume of water, with Green–Ampt type assumptions on its distortion (changes in depth and surface water content) under low (or negative) surface flux values. The reader is directed to the above references, or Smith (2002), for details.

Surface Water and Infiltration Interdependence

Early methods of treating infiltration excess involved subtracting an infiltration pattern (often assumed constant) from the rainfall pattern and then routing the remainder ("rainfall excess") across the catchment toward a receiving stream (Crawford and Linsley, 1966; HEC-1, 1990). Not only does this ignore the interaction of rainrate and ponding time described above, which was generally unknown, but it also ignores the ability of the soil to continue to infiltrate whatever surface water exists after rainfall has fallen below the infiltrability. The result was an inability to estimate ponding time and an overestimation of the recession flow after runoff production had ceased. Also important in the interactions of surface water and infiltration is the microtopography of the surface and the spatial variability of infiltrability. Microtopography can confine the flowing water to some fraction of the total area, and thus limit the opportunity for infiltration losses during recessions. Significant spatial variability of infiltrability is known to vary on length scales as small as decimeters. These topics will be explored in more detail below.

RUNOFF DYNAMICS

The flow of the rainwater not infiltrated by the soil surface (rainfall excess) is at the small (mm) scale, a complex phenomenon with exact flow directions, unit discharges (discharge per unit flow width), and depths varying widely across the surface (Abrahams and Parsons, 1994; Fiedler and Ramirez, 2000). Runoff is, however, treated at a larger (m) scale as a free-surface hydraulic process. The physical description involves some reasonable assumptions that yield a useful and relatively accurate mathematical description of the runoff from most natural surfaces. Figure 5 illustrates diagrammatically the variables involved in the flow of water along a simplified hillslope.

The Surface Water Flow Equations

The primary assumption is that flow is downslope and locally can be approximated as one-dimensional. At larger scales, it is clearly two-dimensional as the land surface converges and diverges downslope, but even this may be treated as stepwise one-dimensional. The flow is incompressible, and the overall velocity is sufficiently low that the energy is largely in the form of momentum. Thus, the equations used are those that conserve momentum and mass. These equations were first written by de Saint Venant (1871), and are commonly referred to as the Saint Venant equations. The two equations include the conservation of momentum

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + g\frac{\partial h}{\partial x} = g(S_o - S_f) - \frac{u}{A}(r - a_f f) \qquad (15)$$

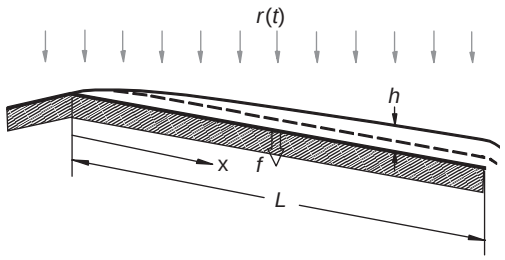


Figure 5 Definition diagram for kinematic approximation for runoff water dynamics

and the conservation of mass

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = (r - a_f f)w \tag{16}$$

where

A is cross-sectional area of flow [L^2]
a_f is fraction of surface covered by infiltrating water
Q is discharge per unit width [$L^2\ T^{-1}$]
x is distance from upstream border [L]
t is time
u is flow velocity [$L\ T^{-1}$]
h is mean depth of flow [L]
g is gravitational acceleration [$L\ T^{-2}$]
S_o is mean bed slope
S_f is friction slope
w is width of flow [L].

The first terms in equation (15) represent changes in momentum and potential energy, and the last term on the right represents the change in momentum due to lateral inflows. The infiltrating area fraction a_f is 1.0 when $r > f_c$, and is reduced on irregular surfaces during flow recession when infiltration comes from surface water. The friction slope represents the shear friction along the flow perimeter, and is exemplified by the Chezy or Manning equation:

$$q = \alpha h^m = \frac{\sqrt{S_f}}{n} h^{5/3} \tag{17}$$

where q is unit discharge [$L^2\ T^{-1}$], α and m are the generalized friction relation parameters, and n is the Manning friction factor. Equation (17) represents experimental results from channels, in which hydraulic radius R replaces h, but the power law relation also reflects considerable experimental data from catchment experiments (e.g. Wu *et al.*, 1978; Abrahams and Parsons, 1994; Emmett, 1970). The form of the friction slope relation is also applicable to the Chezy friction law relationship ($m = 3/2$) (Eagleson, 1970). Note that h represents the mean depth or area per unit width. The relationship described by equation (17) simply states that the unit discharge is a function of a modeled average depth (or storage) in the abstract sheet flow representation of overland flow depicted in Figure 5. It does not imply that overland sheet flow must occur on a hillslope to represent the flow dynamics described in equations (15) through (17). While overland sheet flow has been observed in field settings (e.g. Emmett, 1970), it is not typical as surface water flows usually converge to rivulets in a relatively short distance due to natural microtopography (Dunne, *et al.*, 1991) (also *see* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**). The appropriateness of the discharge-storage relationship implied by equation (17) has been evaluated by several including Wu *et al.* (1978), as well as by microscale numerical simulation with a two-dimensional form of equations (15) and (16) (Fiedler and Ramirez, 2000).

Simplified Forms

The Saint Venant equations are generally applicable to channel and streamflow, but further simplifications are appropriate for overland flow. Because flow is shallow, it can be reasonably assumed that very little energy is contained in the form of inertia, and an order of magnitude analysis demonstrates that the addition of rainfall excess adds negligible net momentum to the flow. With these approximations, equation (15) can be reduced to

$$\frac{\partial h}{\partial x} = S_o - S_f \tag{18}$$

Equations (17) and (18) constitute the *diffusive wave equations* (Morris and Woolhiser, 1980), which are appropriate for certain flow conditions characterized by shallow slopes. A criterion for applicability is given by Morris and Woolhiser (1980).

For larger values of surface slope S_o, the value of dh/dx becomes negligible compared with S_o, and equation (18) reduces to $S_f = S_o$. In this case, the *kinematic wave equations* are appropriate. The flow-depth relationship in this case is a description of the relation of mean effective depth and mean effective discharge, equation (17). This relation plus the continuity of mass equation (16) constitute the kinematic wave equations. Equation (16) with equation (18) may be written in terms of h for unit width of overland flow, rather than area A:

$$\frac{\partial h}{\partial t} + m\alpha h^{m-1} \frac{\partial h}{\partial x} = r - a_f f \tag{19}$$

Woolhiser and Liggett (1967) developed a kinematic wave number, $k = S_o L / h_o F^2$, to measure the ability of equation (19) to represent well the Saint Venant equations. L is plane length, h_o is normal depth for a given flow q, and F is Froude number. Larger values of k indicate that the kinematic wave equations are better approximations. This analysis was further extended by Morris and Woolhiser (1980) to consider a larger range of Froude numbers.

Solving the Kinematic Wave Equations

Equation (19) may be solved under certain assumptions by use of the method of characteristics. This is a mathematical approach to solve sets of partial differential equations by transformation into sets of ordinary differential equations (Lighthill and Whitham, 1955; Wooding, 1965). This article is not an appropriate place to present the details of this method, but it is instructive to look briefly at the solution, as it describes the behavior of a surface water runoff hydrograph.

Characteristics are traces in the solution domain (in this case the x, t plane) along which a partial differential equation is reduced to an ordinary differential equation. The characteristic equations corresponding to equation (19) are as follows:

$$\frac{dx}{dt} = \alpha m \sqrt{S_o} h^{m-1} \tag{20}$$

$$\frac{dh}{dt} = r - a_f f \tag{21}$$

Equation (20) describes the *celerity* or characteristic velocity of the flow. In the characteristic method, equation (21) is valid along the line described in the (x, t) plane by equation (20). In turn, the change in h described by equation (21) changes the slope of the x, t characteristic.

Figure 6 illustrates some of the essential features of the characteristics as runoff begins along a slope. This is a very simplified schematic, but at some time after the start of rainfall, runoff begins. The upstream characteristic starts at $x = 0, t = 0$, and moves at an increasing rate downslope. To the right of this characteristic, the flow is unsteady and uniform. To the left, flow is steady and nonuniform. When the upstream characteristic reaches the lower bound, outflow becomes steady, and this is termed the *time to equilibrium*. In this example, and for many rainfalls and short runoff surfaces, this condition is not reached. After the end of rainfall, the characteristics exhibit decelerating celerities as $h(x)$ is reduced by infiltration, and at some time the characteristic velocity and h fall to zero. The locus of drying is shown on the schematic. Between the end of rainfall and the end of runoff, the hydrograph is in *recession*. For impervious surfaces where $f = 0$, the recession is extended considerably, and the recession characteristics are straight lines of differing slopes that depend on $h(x)$ at the time rainfall excess ends.

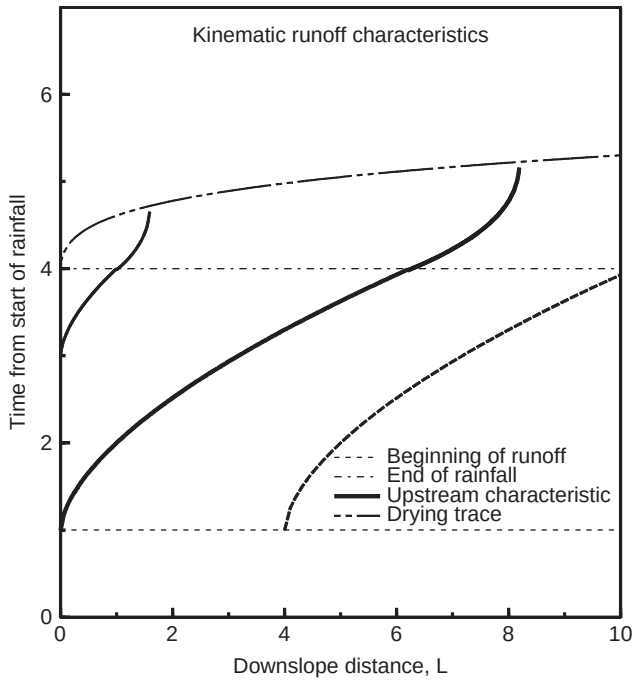


Figure 6 Solution of the kinematic wave equations in the (x,t) plane, for runoff assumed to start at $t = 1$ and end at $t = 4$. The dotted line in Figure 5 illustrates the water depth profile for a time (approximately = 2) when most of the surface is characterized by uniform unsteady flow

Computer models of runoff generally approach the solution of equation (19) with numerical methods, which can treat the time variation of f directly if not analytically. Several early numerical methods have been discussed by Brakensiek *et al.* (1966), and the robust *weighted implicit method* is described by Smith *et al.* (1995b). The characteristic solution is used in this model to estimate the arrival of the upstream characteristic at the first finite difference space node to prevent small solution oscillations. A full discussion of the merits and drawbacks of the various numerical solution techniques is beyond the scope of this article.

SURFACE AND SOIL WATER INTERACTION

It is through the right-hand term in equation (19) that the equations of surface water flow and the equation describing surface water infiltration are linked. From the point of view of the surface water calculations at any point, the input to surface flow is always $r - f$, and becomes negative (a loss) when the rainfall rate, r, falls below f_c, as long as there is free water on the surface. From the point of view of the infiltration calculations after runoff begins, the reduction of rainfall rate below f_c is not noticed, and infiltration rate continues at capacity f_c, as long as this rate can be supplied by surface water plus rainfall rate.

The solution characteristics during recession are also shown schematically in Figure 6. The case of a "flat" infiltrating surface and an impervious surface are two extremes in the behavior of the flow recession. Intermediate to these two is the case of a surface with microtopographic features, either irregular – such as grassed hummocks, or regular, such as furrowed paddocks. In this case, the area available for infiltration during recession flow is a function of the mean depth of surface water. The area is clearly zero at $h = 0$, and at some rather large depth, h_c, the entire area may be covered with water. For $0 < h < h_c$, a fraction of the area will be covered and the loss of water to infiltration during recession will be limited accordingly.

Another case where microtopography plays an important role in surface water interactions is the case of *run-on*. This term refers to situations where water is generated by rainfall excess at a location upslope or upstream of an area with a higher infiltration capacity. Surface microtopography and resulting flow convergence can limit the opportunity for infiltration of the run-on water as the effective wetted

perimeter is reduced. This in turn can affect the extent to which run-on will advance across the irregular surface. Woolhiser *et al.* (1996) studied, for example, the case of run-on caused by a spatial trend in infiltration due to variation in K_s, and showed that significant effects on hydrographs are possible. Small-scale random spatial variations are a somewhat different case.

Spatial Variability and its Effects on Runoff Processes

Like most natural processes, infiltration excess and its associated runoff are affected by the variability of nature. Some natural processes lend themselves to treatment in bulk, with processes characterized by parameters that represent the sum effect of many smaller scale variations. Runoff tends, however, to be a nonlinear process in many locations, and it is often affected by natural spatial variations at scales an order of magnitude less than the scale of interest (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**).

In applying the kinematic flow equations, microtopography can be abstracted to a simple microchannel geometry (Woolhiser, *et al.*, 1996; Smith *et al.*, 1995a). This can represent the ensemble result of extensive measurements of variations in relief across the flow direction, such that a composite geometry may represent the relative surface coverage between mean depths of 0 and some maximum.

Small-scale variations in surface conditions or soil conditions within a catchment can make it difficult to represent the catchment runoff processes by the mathematics described above. The infiltration parameter K_s is a very important one in estimating infiltration and runoff dynamics for Hortonian runoff, and it has been found consistently to exhibit random variation at small scales (cm to m) (Nielsen *et al.*, 1973). Several methods have been reported to deal with the effect of this variation on the runoff for small catchments (Smith and Hebbert, 1979; Sivapalan and Wood, 1986; Sharma *et al.*, 1980; Smith and Goodrich, 2000). In general, the difference between infiltration at a point and the behavior of an ensemble of points is illustrated in Figure 7. This graph uses scaled values as described by equations (10) and (11). The local value of K_s is assumed, on the basis of experimental evidence, to vary randomly with a lognormal distribution. The effect of random variation is to blur the existence of a single time of ponding, as shown, and further to alter the value of the large-time asymptotic value of f_c. The effect of variability on K_s is largely confined to smaller values of rainfall rate; for r_* values of 10 or more, where $r_* \geq$ the highest infiltration capacity, it is negligible. On the basis of the value of the coefficient of variation of K_s [CV(K_s)], Smith and Goodrich (2000) developed an ensemble infiltration model that uses the same parameters as equations (8) or (9), but includes the effect of CV(K_s) and rainfall rate.

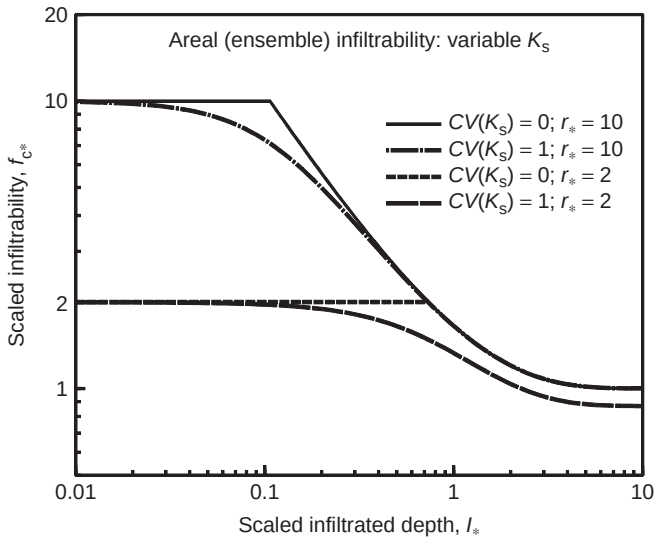


Figure 7 Examples of infiltration patterns for areas characterized by small-scale variability of K_s. Two rainfall rates are shown, and the interaction of variability and rainrate is illustrated

The effect of parameter variations in looking at larger-scale hydrology is an important problem related to *upscaling* in hydrology – simulating or predicting the performance of an ensemble of catchments that make up a larger catchment. Models of hydrology useful at one scale may not be successful at the scale an order or more larger. While the functional representation of small-scale K_s variations improves the prediction of runoff, especially for storms where the runoff is a small fraction of the rainfall, the interactions of variability during runoff are more complicated than the ensemble effect, due to run-on along the flow path. Another method to treat this, at least in a research-level model, distributes finite incremental variations in values of K_s randomly along the flow in a one-dimensional simulation, or divides the catchment into parallel strips of flow, each strip representing a portion of the random distribution expected in the value of K_s (Woolhiser and Goodrich, 1988).

Other characteristics of a catchment may also exhibit significant spatial variability. The effects of variation in slope and surface hydraulic roughness can be modeled at appropriate scales by treating the flow path as a cascade of smaller elements (Smith *et al.*, 1995b; Goodrich *et al.*, 2002; see also `www.tucson.ars.ag.gov/kineros`). This becomes impractical at very small or large scales; however, surface runoff becomes channelized into rills and microchannels before the scale becomes too large to consider such variations.

The scale of variation of runoff intensity is generally larger than that of K_s, but its variation is equally as important for treating infiltration excess on a catchment. Spatial

variability of rainfall and its effects on runoff generation on catchments scales of $1-100\,km^2$ and greater have been relatively widely studied (Balmer *et al.*, 1984; Iwasa and Sueishi, 1990; Niemczynowicz and Sevruk, 1991; Houser *et al.*, 2000; Syed *et al.*, 2002), with those at smaller areal scales typically geared toward urban storm runoff estimation (Berne *et al.*, 2004). Even in an urban setting, the normal scale of rainfall measurements, including those derived from radar, are often considerably larger than rainfall variations that are important for runoff generation. This is particularly true where small intense runoff-producing storm cells dominate runoff production. The USDA-ARS Walnut Gulch Experimental Watershed (WGEW) in southeast Arizona is such an area. In this experimental watershed, Reich and Osborn (1982) concluded that for 5-min maximum storm depths recorded at rain gauges greater then 5 km apart were statistically independent. At a smaller scale within the WGEW (~5 hectare catchment area), Goodrich *et al.* (1995) found rainfall gradients ranging from 0.28 to 2.48 mm/100 m with an average of 1.2 mm/100 m. These gradients represent a 4 to 14% variation of the mean rainfall depth over a 100-m distance but a much larger percentage in terms of total depth of runoff over the catchment due to the relatively low runoff to rainfall ratios that are common in arid and semiarid regions. In a follow-on study, Faurès *et al.* (1995) found that if distributed catchment modeling is to be conducted at the 5-hectare scale, knowledge of the spatial rainfall variability on the same or smaller scale is required. A single raingage with the standard uniform rainfall assumption can lead to large uncertainties in runoff estimation.

Scale Issues and Models of Runoff

Finally, it is important to consider the timescales of the processes of surface runoff. Surface water flows on natural topography do not travel far before converging in some manner into concentrated flow channels. Distributed surface flows prior to some kind of channelization rarely exceed 100 m in length, except in special cases of low-sloped tilled catchments (Dunne *et al.*, 1991; Abrahams and Parsons, 1994). This limitation in spatial scale also imposes a limitation in timescale. The timescale of variations and the rate of flow across the surface means that rainfall rate variations on the order of less than a minute can be important parts of the dynamics of runoff. Surface runoff cannot therefore be treated scientifically with information on rainfall that consists only of the daily depth of rain. For this lumped data approximation, the prediction of runoff will contain enormous uncertainty.

Models of Hortonian runoff generally employ the solution of the kinematic or diffusive wave equations by numerical methods. The wide variety of numerical solution is too numerous to be described here. Readers are referred to Singh (1995) for a good overview of numerical solution techniques. Again, there are practical limits for the numerical subdivision of the flow path, and these impose corresponding limits on the timescales. Moreover, the values of h that are involved in the solution are on the order of a few millimeters, and the solution of equation (19) over values of Δx that are tens or even hundreds of meters or longer are physically realistic treatments of shallow flow dynamics, but are rather simply nonlinear storage models. While some models have employed runoff elements at this scale, a valid numerical model of surface runoff should be confined to the length of flow expected before rills, channels, and other cascading concentrated flow paths begin to dominate the runoff hydraulics. If concentrated flows and rills do not dominate the hydraulics, changes in slope and convergence of flow can both be modeled by the use of cascading one-dimensional surfaces with varying slopes and widths (Woolhiser *et al.*, 1990; Smith *et al.*, 1995b).

SUMMARY

We have briefly shown how the theory of soil and surface water hydraulics has led to a model of surface water generation from rainfall for catchments on the scale of hectares. However, this should not tempt us to believe that a scientific knowledge of rainfall excess can easily be applied to hydrologic problems in general. The models are at best quasi-scientific insofar as nature's ubiquitous heterogeneities do not allow the hydrologist to make direct measurements at any point of many of the important parameters (e.g. surface roughness), or even determine a real effective average of parameters by any remotely sensed data. The theoretical basis of the processes described above does allow some confidence in the robustness of estimated parameters insofar as they are applicable to a wide range of rainfall conditions, but direct measurement of key properties, such as effective catchment values for K_s, remains a difficult undertaking.

Remotely derived topographic data using LIDAR is of sufficient accuracy and resolution that it can be used to aid in estimation of some of the geometric parameters (slopes and slope lengths, etc.) required for excess rainfall-runoff modeling (Carter *et al.*, 2001). Multispectral remotely sensed data has been widely used to estimate land use and ground cover, and if sufficient temporal resolution exists, the variation of ground cover conditions can affect hydraulic roughness and influence infiltration parameters. Land use and land cover data, combined with textural-based estimates of soil hydraulic properties can be used to provide crude initial estimates of needed parameters (Miller *et al.*, 2002; and see **www.tucson.ars.ag.gov/agwa**). However, these estimates are often based on simple look-up table relationships from incomplete field data relating soils and cover to surface and soil hydraulic parameters. This leaves hydrologic science with challenges for field data

collection and realistic characterization of the variability of important parameters for use in the mathematical models of rainfall excess runoff described herein.

REFERENCES

Abrahams A.D. and Parsons A.J. (1994) Hydraulics of interrill overland flow on stone-covered desert surfaces. *Catena*, **23**, 111–140.

Balmer P., Malmqvist P.A. and Sjoberg A. (Eds.) (1984) *Proceedings of the Third International Conference on Urban Storm Drainage*, Vol. 4, Published by Chalmers University of Technology: Goteborg, June 4–8, 1984.

Berne A., Delrieu G., Creutin J.-D. and Obled C. (2004) Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology*, **299**(3–4), 166–179.

Beven K.J. (2004) Surface runoff at the Horton Hydrologic Laboratory. *Journal of Hydrology*, **293**, 219–234.

Brakensiek D.L., Heath A.L. and Comer G.H. (1966) *Numerical Techniques for Small Watershed Flood Routing*, U.S. Department of Agriculture, ARS 41–113.

Carter W., Shrestha R., Tuell G., Bloomquist D. and Sartori M. (2001) Airborne laser swath mapping shines new light on earth's topography. *EOS, Transactions-American Geophysical Union*, **82**(46), 549, 550, 555.

Corradini C., Melone F. and Smith R.E. (1997) A model for infiltration during complex rainfall patterns. *Journal of Hydrology*, **192**, 104–124.

Crawford N.H. and Linsley R.K. (1966) *Digital Simulation in Hydrology*, Stanford Watershed Model IV, Stanford University Technical Report No. 39, Palo Alto.

de Saint Venant B. (1871) Théorie du movement non-permanent des eaux avec application aux crues des revierés *et al.*'introduction des marées dans leur lit. *Academie Des Sciences [Paris] Comptes*, **73**, 148–154, 237–240.

Dunne T., Zhang W. and Aubry B.F. (1991) Effects of rainfall, vegetation, and microtopography on infiltration and runoff. *Water Resources Research*, **27**(3), 2271–2285.

Eagleson P.S. (1970) *Dynamic Hydrology*, McGraw Hill: New York, pp. 325–367.

Emmett W.W. (1970) *The Hydraulics of Overland Flow*, USGS Profession Paper 662-A, US Government Printing Office: Washington.

Faurès J.M., Goodrich D.C., Woolhiser D.A. and Sorooshian S. (1995) Impact of small-scale spatial rainfall variability on runoff simulation. *Journal of Hydrology*, **173**, 309–326.

Fiedler F.R. and Ramirez J.A. (2000) A numerical method for simulating discontinuous shallow flow over an infiltrating surface. *International Journal for Numerical Methods*, **32**, 219–240.

Goodrich D.C., Faurès J.M., Woolhiser D.A., Lane L.J. and Sorooshian S. (1995) Measurement and analysis of small-scale convective storm rainfall variability. *Journal of Hydrology*, **173**, 283–308.

Goodrich D.C., Unkrich C.L., Smith R.E. and Woolhiser D.A. (2002) KINEROS2 – a distributed kinematic runoff and erosion model. *Proceedings of the Second Federal Interagency Hydrologic Modeling Conference*, Las Vegas, p. 12, July 28 – August 1, 2002.

Green W.A. and Ampt G.A. (1911) Studies on soil physics: 1. The flow of air and water through soils. *Journal of Agricultural Science*, **4**, 1–24.

Horton R.A. (1933) The role of infiltration in the hydrologic cycle. *Transactions-American Geophysical Union*, **14**, 446–460.

Horton R.A. (1936) Hydrologic interrelations of water and soils. *Soil Science Society of America Proceedings*, **1**, 401–429.

Horton R.A. (1939) Analysis of runoff-plot experiments with varying infiltration-capacity. *Transactions of the American Geophysical Union*, **20**(Part IV), 693–711.

Houser P.R., Goodrich D.C. and Syed K.H. (2000) Runoff, precipitation, and soil moisture at Walnut Gulch. In *Spatial Patterns in Hydrological Processes: Observations and Modeling*, Chap. 6, Grayson R. and Blosch G. (Eds.), Cambridge University Press: pp. 125–157.

Hydrologic Engineering Center (1990) *HEC-1, Flood Hydrograph Package, Users Manual*, USACE HEC Publication CPD-1A, p. 433.

Iwasa Y. and Sueishi T. (1990) *Proceedings of the Fifth International Conference on Urban Storm Drainage*, Published by University of Osaka: Osaka, July 23–27.

Li R.M., Stevens M.A. and Simons D.B. (1976) Solutions to the Green-Ampt infiltration equation. *Journal of Irrigation and Drainage, ASCE*, **102**(2), 239–248.

Lighthill F.R.S. and Whitham G.B. (1955) On kinematic waves, 1. Flood movement in long rivers. *Proceedings of the Royal Society of London, Series A*, **239**, 281–316.

Miller S.N., Semmens D.J., Miller R.C., Hernandez M., Goodrich D.C., Miller W.P., Kepner W.G. and Ebert D. (2002) GIS-based hydrologic modeling: the automated geospatial watershed assessment tool. *Proceeding of the Second Federal Interagency Hydrologic Modeling Conference*, Las Vegas, p. 12, July 28 – August 1, 2002.

Mls J. (1980) Effective rainfall estimation. *Journal of Hydrology*, **45**, 305–311.

Morris E.M. and Woolhiser D.A. (1980) Unsteady one-dimensional flow over a plane: partial equilibrium and recession hydrographs. *Water Resources Research*, **16**(2), 355–360.

Nielsen D.R., Biggar J.W. and Erh K.T. (1973) Spatial variability of field measured soil water properties. *Hilgardia*, **42**, 215–260.

Niemczynowicz J. and Sevruk B. (1991) Urban rainfall and meteorology. *Atmospheric Research*, **27**(1–3), 215.

Paige G.B., Stone J.J., Guertin D.P. and Lane L.J. (2002) A strip model approach to parameterize a coupled Green-Ampt kinematic wave model. *Journal of American Water Resources Association*, **38**(5), 1363–1377.

Parlange J.-Y., Lisle I., Braddock R.D. and Smith R.E. (1982) The three-parameter infiltration equation. *Soil Science*, **133**(6), 337–341.

Reich B.M. and Osborn H.B. (1982) *Improving Point Rainfall Prediction with Experimental Data, International Statistical Analysis of Rainfall and Runoff. International Symposium on Rainfall/Runoff*, Mississippi State University, Water Resources Publication: Littleton, pp. 41–54.

Sharma M.L., Gander G.A. and Hunt C.G. (1980) Spatial variability of infiltration in a watershed. *Journal of Hydrology*, **45**, 101–122.

Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Water Resources Publications: Highlands Ranch.

Sivapalan M. and Milly P.C.D. (1989) On the relationship between the time condensation approximation and the flux-concentration relation. *Journal of Hydrology*, **105**, 357–367.

Sivapalan M. and Wood E.F. (1986) Spatial heterogeneity and scale in the infiltration response of catchments. In *Scale Problems in Hydrology*, Chap. 5, Gupta V.K., Rodriguez-Iturbe I. and Wood E.F. (Eds.), Reidel: Hingham, pp. 81–106.

Smith R.E. (2002) *Infiltration Theory for Hydrologic Applications*, *Water Resources Monograph 15*, American Geophysical Union: Washington.

Smith R.E., Corradini C. and Melone F. (1999) A conceptual model for infiltration and redistribution in crusted soils. *Water Resources Research*, **35**(5), 1385–1393.

Smith R.E. and Goodrich D.C. (2000) Model for rainfall excess patterns on randomly heterogeneous areas. *Journal of Hydrologic Engineering, ASCE*, **5**(4), 355–362.

Smith R.E., Goodrich D.C. and Quinton J.N. (1995a) Dynamic, distributed simulation of watershed erosion: the KINEROS2 and EUROSEM models. *Journal of Soil and Water Conservation*, **50**(5), 517–520.

Smith R.E., Goodrich D.C., Woolhiser D.A. and Unkrich C.L. (1995b) KINEROS – a kinematic runoff and erosion model. In *Computer Models of Watershed Hydrology*, Chap. 20, Singh V.J. (Ed.), Water Resources Publication: Highlands Ranch, pp. 697–732.

Smith R.E. and Hebbert R.H.B. (1979) A Monte-Carlo analysis of the hydrologic effects of spatial variability of infiltration. *Water Resources Research*, **15**(2), 419–429.

Smith R.E. and Parlange J.-Y. (1978) A parameter-efficient hydrologic infiltration model. *Water Resources Research*, **14**(3), 533–538.

Syed K., Goodrich D.C., Myers D. and Sorooshian S. (2002) Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. *Journal of Hydrology*, **271**(1–4), 1–21.

Wooding R.A. (1965) A hydraulic model for the catchment-stream problem. I. Kinematic wave theory. *Journal of Hydrology*, **3**, 254–267.

Woolhiser D.A. and Goodrich D.C. (1988) Effect of storm rainfall intensity patterns on surface runoff. *Journal of Hydrology*, **102**, 335–354.

Woolhiser D.A. and Liggett J.A. (1967) Unsteady one-dimensional flow over a plane – The rising hydrograph. *Water Resources Research*, **3**, 753–771.

Woolhiser D.A., Smith R.E. and Giraldez J.-V. (1996) Effects of spatial variability of saturated hydraulic conductivity on Hortonian overland flow. *Water Resources Research*, **32**(3), 671–678.

Woolhiser D.A., Smith R.E. and Goodrich D.C. (1990) *KINEROS, A Kinematic Runoff and Erosion Model: Documentation and User Manual*, U.S. Department of Agriculture, Agricultural Research Service: ARS-77, p. 130.

Wu Y.-H., Yevjevich V. and Woolhiser D.A. (1978) *Effects of Surface Roughness and its Spatial Distribution on Runoff Hydrographs*, Colorado State University Hydrology Paper No. 96: Fort Collins, p. 47.

112: Subsurface Stormflow

MARKUS WEILER[1], JEFFREY J MCDONNELL[2], ILJA TROMP-VAN MEERVELD[3] AND TARO UCHIDA[4]

[1]Departments of Forest Resources Management and Geography, University of British Columbia, Vancouver, BC, Canada
[2]Department of Forest Engineering, Oregon State University, Corvallis, OR, US
[3]Ecole Polytechnique Fédérale de Lausanne, School of Architecture, Civil & Environmental Engineering, Lausanne, Switzerland
[4]Research Center for Disaster Risk Management, National Institute for Land & Infrastructure Management, Asahi, Tsukuba, Japan

Subsurface stormflow is a runoff producing mechanism operating in most upland terrains. In a humid environment and steep terrain with conductive soils, subsurface stormflow may be the main mechanism of storm runoff generation. In drier climates and in lowlands with gentler topography, subsurface stormflow may occur only under certain extreme conditions (high antecedent soil moisture), when transient water tables form and induce lateral flow to the channel. While an important contributor to the volume of flow in the stream, subsurface stormflow is also responsible for the transport of labile nutrients into surface water bodies. Since the flow path of water in the subsurface often determines the chemistry of waters discharging into the stream and hence the water quality, characterizing this subsurface flow path and the water's age and origin is important. Subsurface stormflow may also enhance positive pore pressure development in steep terrain and may be responsible for landslide initiation. Thus, subsurface stormflow is of great interest and importance beyond the conventional hydrological literature. This article examines the history of the study of subsurface stormflow processes, reviews theories on the generation of subsurface stormflow, and gives a detailed overview of current research in subsurface flow processes and the implication for future research.

INTRODUCTION

Subsurface stormflow is a runoff producing mechanism operating in most upland terrains. Subsurface stormflow occurs when water moves laterally down a hillslope through soil layers or permeable bedrock to contribute to the storm hydrograph in a river. In humid environments and steep terrain with conductive soils, subsurface stormflow may be the main mechanism of storm runoff generation (Anderson and Burt, 1990b; Gutknecht, 1996). In drier climates and in lowlands with gentler topography, subsurface stormflow may occur only under certain extreme conditions (high rainfall and high antecedent soil moisture), when transient water tables form and induce lateral flow to the channel (Wilcox *et al.*, 1997).

While an important contributor to the volume of flow in the stream, subsurface stormflow is also responsible for the transport of labile nutrients into surface water bodies (e.g. McGlynn and McDonnell, 2003b). Since the flow path of water in the subsurface often determines the chemistry of waters discharging into the stream and hence the water quality, characterizing this subsurface flow path and the water's age and origin is important (Burns *et al.*, 2003). Subsurface stormflow may also enhance positive pore pressure development in steep terrain (Uchida *et al.*, 1999; Wu and Sidle, 1995) and may be responsible for landslide initiation (Montgomery *et al.*, 1997; Sidle and Tsuboyama, 1992). Thus, subsurface stormflow is of great interest and importance beyond the conventional hydrological literature. This article examines the history of the study of subsurface

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

stormflow processes, reviews theories on the generation of subsurface stormflow, and gives a detailed overview of current research on subsurface flow processes and implications for future research.

TERMINOLOGY

Subsurface stormflow is also known in the hydrological literature as interflow, lateral flow, subsurface runoff, transient groundwater, or soil water flow. These multiple terms often confuse the process understanding of subsurface stormflow response to rainfall or snowmelt. While some studies have documented subsurface stormflow as unsaturated flow in the unsaturated zone, most studies have shown that subsurface stormflow is a saturated (or near-saturated) water flow phenomenon – due either to the rise of an existing water table into more transmissive soil above (with ensuing lateral flow) or the transient saturation above an impeding layer, soil-bedrock interface or some zone of reduced permeability at depth (argillic horizon, hard-pan, plough layer, etc.).

The literature on subsurface stormflow includes many references to both soil water and groundwater. Inconsistent definitions of these terms have also led to confusion. Here, we define groundwater or the saturated zone as any area in the soil profile with ≥ 0 kPa matric potential. Soil water, or the unsaturated zone, is the area in the profile with matric potentials of <0 kPa. Conversion from negative to positive potentials may occur very rapidly in the shallow subsurface. Therefore, modifiers such as "transient" groundwater will be used in some instances to indicate parcels of water in the subsurface that change from soil water to groundwater or from being an unsaturated to being a saturated zone, respectively, following our potential-based definition.

HISTORICAL DEVELOPMENT OF IDEAS PERTAINING TO SUBSURFACE STORMFLOW

As part of early studies on the influence of forest management on watershed hydrology in Switzerland, Engler (1919) was among the first hydrologists to recognize the importance of subsurface stormflow on runoff generation in forested environments. He made observations during numerous rainfall events in two neighboring catchments, conducted infiltration experiments, and performed detailed soil physical measurements of porosity, water content, soil texture, and hydraulic conductivity. He concluded from his experiments that overland flow did not occur even during high intensity rainfall. He observed that water infiltrated into the main root zone and flowed laterally in "uncountable veins" in the soil or at the soil-bedrock interface. Figure 1 shows the original conceptualization of these processes by Engler (1919).

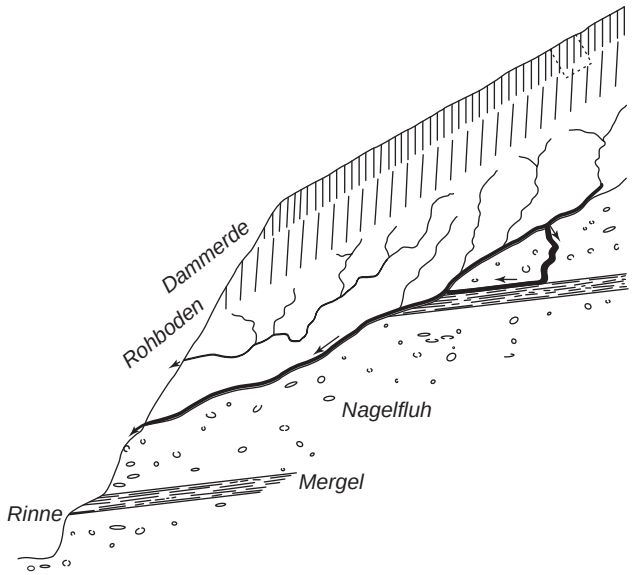


Figure 1 The original perceptual model of subsurface stormflow by Engler (1919). The hatched areas represent the uniform infiltration of water in the humus (*Dammerde*) and the soil (*Rohboden*). Deeper in the profile, water is flowing in ''veins'' laterally. *Nagelfluh* is the bedrock type for a specific geological setting in Switzerland

Hursh and Brater (1941) were the first to quantify the role of subsurface stormflow *senso stricto* in a watershed. This seminal work showed that the stream hydrograph response to storm rainfall at the forested Coweeta experimental watershed consisted of two main components: channel precipitation and subsurface stormflow. Later, Hoover and Hursh (1943) showed that soil depth, topography, and hydrologic characteristics associated with different elevations influenced peak discharge.

The rate of progress in the understanding of subsurface stormflow continued to increase through the International Hydrological Decade (IHD), with key works by Hewlett and Hibbert (1963) on moisture and energy conditions in a sloping concrete-walled hillslope, Whipkey (1965) on lateral preferential flow, Dunne and Black (1970) on subsurface stormflow around the toe of a hillslope and its interaction with near-stream saturated areas, and Weyman (1973) on saturated wedge development. The most important works during the IHD, however, were those that framed subsurface stormflow within the context of the variable source area concept: Hewlett and Hibbert (1967) in the United States, Cappus (1960) in France, and Tsukamoto (1961) in Japan. Later, Anderson and Burt (1978) provided clear field evidence for how topographic hollows were the key hot spots in the landscape for connected subsurface stormflow to the stream channel.

After comments by Freeze (1972) that subsurface stormflow could not be a dominant runoff process based on Darcy-Richards analytical approaches work by Beasley

(1978), Harr (1977) and Mosley (1979) showed clearly that the response time and lateral flux rate of subsurface stormflow on steep forested hillslopes could be fast enough to be a main contributor to channel stormflow in headwater catchments – often via noncapillary pore space or flow in addition to Darcy-Richards like flux in the matrix.

Considerable debate has also surrounded the age and origin of subsurface stormflow. The "Maimai debate" is a classic case study in this regard and, therefore, is used here as an example (see a more extended and detailed review by McGlynn *et al.*, (2002)). Mosley (1979, 1982) conducted the first comprehensive study of subsurface stormflow at the Maimai catchments in New Zealand. He found a close coincidence in the time of the discharge peak in the stream and the time of the subsurface stormflow peaks, suggesting rapid movement of water vertically in the soil profile and in lateral downslope direction in the form of a saturated wedge. The wedge almost intersected the ground surface at the toe of the slope and tapered off in the upslope direction (Mosley, 1979). Dye tracing experiments of Mosley showed that his excavated "pit faces" (i.e. small trenches roughly 1 m wide and dug down to the soil-bedrock interface) displayed points of concentrated seepage during storm events, usually at the base of the B horizon, at which high rates of outflow were observed. At one site, Mosley (1979) observed that water "gushed" out of two pipes discovered at the base of the B horizon. Mosley's perceptual model considered macropore flow (see details in the next section) to be a "short-circuiting" process by which water could move through the soil at rates up to 300 times greater than the measured mineral soil saturated hydraulic conductivity (Figure 2a).

Pearce *et al.* (1986) and Sklash *et al.* (1986) followed with work at Maimai where they collected samples of rainfall, soil water, and streamflow and analyzed for electrical conductivity (EC), chloride ($C1^-$), deuterium (δD), and oxygen-18 (δ^{18}O) composition. They found that: (i) most of the mixing of "old" (pre-event) and "new" (event) water occurred in the hillslope; and (ii) subsurface water discharge to the stream was an isotopically uniform mixture of stored water. In other words, the interpretations offered by Pearce *et al.* (1986) and Sklash *et al.* (1986) directly refuted Mosley's (1979) determination that rapid transmission of new water through macropores formed the majority

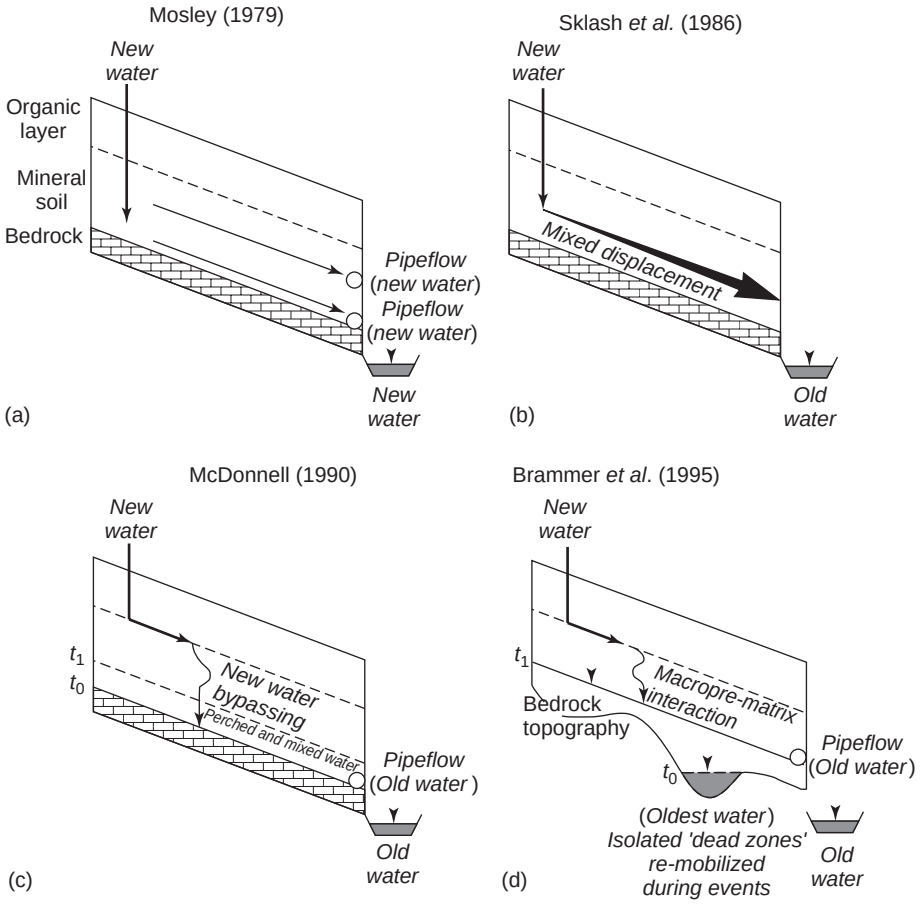


Figure 2 The evolving perceptual model of subsurface stormflow at the Maimai catchment in New Zealand (Reprinted from McGlynn *et al.*, 2002. © 2002, with permission from Elsevier)

of stream runoff. Sklash *et al.* (1986) postulated the conceptual model that saturated wedges on the lower slopes and groundwater ridges in the valley bottoms developed quickly as infiltrating rain converted the tension-saturated zone into groundwater. This perceptual model negated the need to invoke rapid transmission of new water down slope via macropores in order to explain the stream flow response, since stored water was the main component discharged into the stream channel during events (see Figure 2b).

McDonnell (1989, 1990) and McDonnell *et al.* (1991) combined isotope and chemical tracing with detailed tensiometric recording in an effort to explain the discrepancies between the earlier perceptual models. McDonnell found that: (i) water table longevity at the soil-bedrock interface was very short and showed a close correspondence with hillslope throughflow rate; and (ii) the interconnectedness of pipes in those zones was large enough to account for the rapidity of water table decline and pore water pressure dissipation. As a result, McDonnell (1990) proposed a new conceptual model where, as infiltrating new water moved to depth, water perched at the soil-bedrock interface and "backed-up" into the matrix, where it mixed with a much larger volume of stored, old matrix soil water. This water table was dissipated by the moderately well-connected system of pipes at the mineral soil-bedrock interface (see Figure 2c). These studies were followed by Woods and Rowe (1996) and Brammer *et al.* (1995) who showed that the topography of the bedrock surface was a key determinant of where subsurface flow was concentrated spatially across the hillslope (see Figure 2d).

The Maimai experiments, along with other field experiments through the early 1990s, achieved a general consensus that: (i) pre-event water stored in the catchment before the rain event is the dominant contributor to stormflow in the stream – averaging 75% worldwide (Buttle, 1994) (*see* **Chapter 116, Isotope Hydrograph Separation of Runoff Sources, Volume 3**); (ii) vertical (and often also lateral) preferential flow is a ubiquitous phenomenon in natural soils, particularly in steep catchments (Germann, 1990; Tsuboyama *et al.*, 1994) (*see* **Chapter 116, Isotope Hydrograph Separation of Runoff Sources, Volume 3**); and (iii) combining hydrometric, chemical, and isotopic observations in experimental work is necessary to constrain any perceptual model of subsurface stormflow or other runoff producing mechanism (Bonell, 1993; Bonell, 1998).

This scientific debate through the latter half of the twentieth century shows how scientific progress can be made by revisiting and reanalyzing data in one experimental watershed. However, it also highlights the danger in the field of hillslope hydrology of relying upon generalization of findings and processes of only a few selected hillslopes and watersheds. Few studies have compared subsurface flow processes and the dominant flow pathways across many sites (e.g. Beasley, 1978; Scherrer, 1997). The synthesis work of Dunne (1978) still stands as the most complete tabulation and intercomparison of subsurface stormflow data to date.

FLOW REGIMES OF SUBSURFACE STORMFLOW

Subsurface stormflow describes all runoff generation processes in the hillslope close to the soil surface that result in a stream channel hydrograph response during a precipitation or snowmelt event. This response may be coupled directly to flow in preferential pathways like macropores and layers or areas with high permeability. However, rapid subsurface stormflow response may also result from a fast hydraulic response of connected saturated areas in a hillslope in response to infiltrating precipitation (Burt and Butcher, 1985). The main flow regimes at the hillslope scale may be subdivided into homogeneous matrix flow and preferential flow.

Homogeneous Matrix Flow

Lateral matrix flow can be a viable subsurface stormflow process if water is already stored in the soil within connected saturated and close-to-saturated areas. These areas may respond quickly to an increase in hydraulic gradient and cross-sectional area due to infiltrating water. This process may occur in slopes where a high-permeable soil layer with high infiltration capacity is situated (parallel) above a low-permeable soil layer (e.g. bedrock, argillic horizon, etc). Since the water in storage on a typical hillslope is large relative to the rainfall depth, this matrix regime often results in a large contribution of pre-event water to the stream as only a small amount of event water is necessary to increase the hydraulic gradient and cross-sectional area in the slope and create a connected transient groundwater body. This flow process is described in the literature often in terms of translatory flow (Burt, 1989), transmissivity feedback (Rodhe, 1987), or lateral flow at the soil-bedrock interface (Tani, 1997).

Preferential Flow

Preferential lateral flow occurs either in distinctive structures in the soil where water flows only under gravity (macropores) or in areas with a higher permeability than the surrounding soil matrix. Macropores in the soil or fractures in the bedrock that are oriented predominantly slope parallel may transport water efficiently and rapidly from the hillslope to the stream (Beven and Germann, 1982). Laterally oriented macropore flow may dominate in many forest environments where macropores are generated by plant roots and burrowing animals. Macropores that are enlarged by erosion and connected over several meters are often termed *soil pipes* (Anderson and Burt, 1990a; Jones, 1971). If a connected network is developed because of internal erosion

and eluviation and connection of macropores, piping can provide effective drainage augmentation to hillslopes. However, disconnected macropores that connect hydraulically during storms can also result in an effective drainage of the hillslopes (Weiler *et al.*, 2003). If the underlying bedrock is more permeable, water can infiltrate into the bedrock and then percolate vertically into bedrock fissures and cracks – negating the macropore enhancement of subsurface stormflow on the timescale of a rainfall event. The preferential flow regime is described in the literature often in terms of lateral preferential flow (Tsuboyama *et al.*, 1994) and pipeflow (Uchida *et al.*, 1999), or lateral preferential flow at the soil-bedrock interface (McDonnell, 1990).

High permeability layers are areas in the slope with a coarse texture and large pore and void space. These are often found in talus slopes, landslide debris, peri-glacial solifluction deposits, or unconsolidated moraine material. Erosion of fine sediments by turbulent flow in areas with already coarse soil material increases the hydraulic conductivity and makes these areas particularly conductive. This flow regime can be best envisioned by an extension of the surface streams into the hillslope where many "small" subsurface streams connect preferentially the hillslopes with the streams (Sidle *et al.*, 2000). This flow process is described in the literature in terms of flow in "high-permeable layers" (Scherrer, 1997).

CURRENT RESEARCH IN SUBSURFACE FLOW PROCESSES

Subsurface Flow Initiation

Various experimental observations have shown that the water table development near the base of the soil profile can be very rapid following the onset of precipitation. This response is often a key control on the initiation of lateral flow due to the combination of four factors: (i) an increase of the hydraulic gradient; (ii) an increase of the cross-sectional area of flow; (iii) the rise of the water table into more permeable upper soil layers; and (iv) a connection of transient groundwater bodies across the hillslope. There is still debate among scientists about the controls on rapid water table response, but vertical preferential flow during infiltration and a decline of the drainable porosity with soil depth are thought to be the main causes for rapid subsurface flow initiation capable of generating subsurface stormflow.

Vertical Preferential Flow

Vertical preferential flow in natural soils is an almost ubiquitous phenomenon (Flury *et al.*, 1994; Weiler and Naef, 2003). Different types of heterogeneities in soils result in different preferential flow processes. For example, macropore flow is common in soils with animal burrows, roots, and earthworm channels; heterogeneous matrix flow (a type of preferential flow) in soil with spatially heterogeneous soil properties and wettability. How water is transported rapidly in the soil bypassing the unsaturated zone depends on the initiation processes of preferential flow and the interaction of the preferential flow pathways (e.g. macropores) with the surrounding soil matrix. How the two processes determine vertical water delivery into the lower soil profile can be distinguished experimentally. Figure 3 shows results from two sites where a dye (Brilliant Blue FCF) was added to 70 mm of simulated rainfall that stained the flow pathways upon infiltration into the soil. Later, the soil profiles were excavated and pictures were taken of the vertical soil sections. The stained areas can be then classified with image analysis techniques (Weiler and Flühler, 2004). Figure 3(a) shows that infiltration is governed by a high interaction between macropores and the soil matrix and that macropore flow is initiated close to the soil surface. These processes will result in a slow response in the lower profile. In the example in Figure 3(b), macropore flow is also initiated close to the soil profile but interaction between water flow in the macropores and the surrounding soil matrix is very low, resulting in narrowly stained features. Here, water flow in macropores is often turbulent and mostly driven by gravity. Hence, water delivery to the lower soil profile is very rapid. While macropores may comprise only a small part of the total soil porosity (e.g. 0.35–0.77%, e.g. Weiler and Naef (2003)), they account for almost all the water flow at or near saturation within the profile. The resulting water movement in the soils is very heterogeneous and certain areas within the soil may be completely bypassed. These processes often defy the Darcy-Richards formulations that rely exclusively on capillary-driven fluid flow (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2** for full treatment of classical flow in porous media theory).

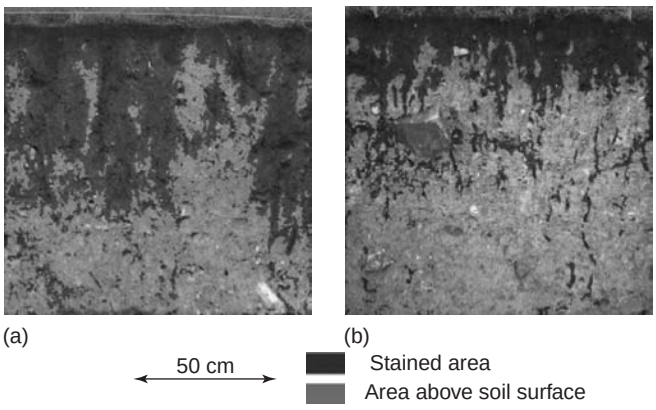


Figure 3 Dye patterns from two different sites: Rietholzbach experimental watershed, Switzerland, (a) and Heitersberg near Zurich, Switzerland (b). Note the spatial heterogeneity of dyed water and preferred nature of water flow vertically within the soil profile. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Another observation sometimes made in soils when water content is close to field capacity is that only a small fraction of water input into the soil results in a strong hydrodynamic response in the lower soil profile. Torres *et al.* (1998) found that this hydrodynamic response can be on an average 15 times greater than the estimated water and wetting front velocities. While these processes remain poorly understood, research to date shows clearly that rapid travel times of fluid pressure head or water content through the unsaturated zone could be mistaken as preferential or macropore flow (Rasmussen *et al.*, 2000; Smith and Hebbert, 1983).

Drainable Porosity

The conversion of unsaturated to saturated conditions or the rise of an existing groundwater table at depth depends on the drainable porosity (or specific yield) of the soil. Drainable porosity is commonly defined as the difference in volumetric water content between 0 kPa and 33 kPa soil water potential (approximately field capacity) (Weiler and McDonnell, 2004). This represents the porosity of the soil draining within a short duration – characteristic of water table rise in the saturated zone. Drainable porosity commonly decreases with depth because of an increase in the bulk density, a decrease in macroporosity, and a change in soil texture and pore structure.

The decrease in drainable porosity with depth can be observed in different soils by comparing measurements of the water content at saturation and field capacity at several depths within the soil profile. Figure 4 illustrates the change of drainable porosity with depth in three different soils. The values show a high-drainable porosity in the topsoil and a slower (site no. D3) or a more rapid (site no. B4) decrease in drainable porosity with depth. Since the drainable porosity cannot be zero within the soil profile, an exponential model

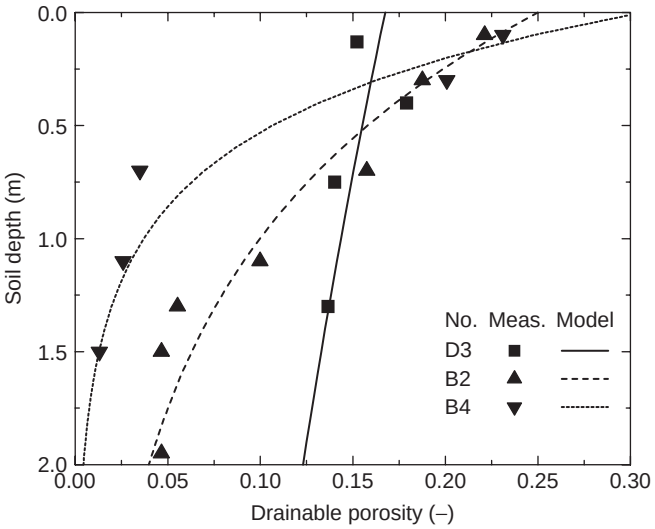


Figure 4 Measured and fitted drainable porosity with depths of three different soils (No. of sites see Table 1)

can be used to describe the observed change with depth:

$$n_d(z) = n_0 \exp\left(-\frac{z}{m}\right) \tag{1}$$

where n_0 is the drainable porosity at the soil surface, m is a decay coefficient, and z is the depth into the soil profile (positive downward). An exponential model can be fitted to the measured values to produce a depth distribution of drainable porosity. Table 1 gives the values of the two fitting parameters n_0 and m for 13 different soils. The fitted values of the parameter m show a pronounced decline of the drainable porosity with depth for almost all sites, except for site no. A4. At this site, the texture changes

Table 1 Values of the two parameters n_0 and m of the exponential depth function for drainable porosity and the goodness-of-fit information (efficiency) for soil data of 13 different sites

No.	Site	n_0	m	Efficiency
Data by Weiler (2001), Grassland soils				
A1	Rietholzbach	0.083	0.294	0.915
A2	Heitersberg	0.037	2.353	0.867
A3	Koblenz	0.058	0.339	0.953
A4	Niederweningen	0.016	−0.457	0.817
Data by Ranken (1974), Forest soils				
B1	Pit 1, upslope	0.268	0.881	0.929
B2	Pit 2, upslope	0.271	1.206	0.946
B3	Pit 5, midslope	0.237	1.717	0.957
B4	Pit 12, downslope	0.356	0.762	0.912
Data by Yee (1975), Forest soils				
C1	Klickitat soil, Pit 2, Granophyric gabbro, 70% slope	0.250	3.152	0.878
C2	Bahannon soil, pit A, Tyee sandstone, 65% slope	0.131	2.756	0.949
Data by Rothacher *et al.* (1967), Forest soils				
D1	Reddish-brown Lateritics over weathered breccia	0.347	1.052	0.984
D2	Yellow–brown Lateritics over rotten rock	0.253	1.146	0.856
D3	Regosol over breccia	0.168	6.496	0.377

to a more sandy soil at depth, resulting in an increase of the drainable porosity with depth. For all other sites, the parameter m ranges from 0.294 (rapid decrease) to 6.496 (slow decrease). The drainable porosity at the soil surface n_0 does not influence the parameter m; however, the grassland soils generally show a lower n_0 than the forest soils. A low drainable porosity in the lower soil profile indicates that only a small amount of water is needed to raise the water table and, thus, to increase the potential for lateral subsurface flow.

For example, a sensitivity study implementing the exponential model for drainable porosity decline in Hillvi (**http://faculty.forestry.ubc.ca/weiler/hillvi.html**), a subsurface hillslope flow model, reveals that for a 10% steep hillslope using m-values of 0.3, 0.5, and 1.5, the response time is 14.4, 3.8, and 1.4 times faster, respectively, than for the same hillslope with a constant drainable porosity. The response time in this example is calculated as the time subsurface flow increases to half of the simulated constant rainfall intensity. This example highlights the importance of the drainable porosity decline in the soil profile for the vertical response of the groundwater level in the hillslope and, hence, the lateral response of subsurface flow.

Topographic Control on Lateral Subsurface Flow

There is now wide consensus that in areas with steep slopes, thin soils, and matrix hydraulic conductivities above the maximum rainfall intensity, water moves vertically to depth (as matrix or preferential flow), perches at the soil-bedrock or an impeding layer at depth, and then moves laterally along the lowest depths of the profile (Freer *et al.*, 2002; McDonnell, 1990; Peters *et al.*, 1995; Sidle *et al.*, 2000; Tani, 1997; Tsukamoto *et al.*, 1982; Uchida *et al.*, 2002). Hence, the bedrock topography may control the direction and accumulation of flow more directly than the topography of the surface. Hillslope experiments using piezometer and tensiometer data as well as flow volumes recorded at a trench face suggest that rapid saturated subsurface flow occurs as narrow "ribbons" of saturated flow along the bedrock topographic surface (Hutchinson and Moore, 2000; Woods and Rowe, 1996). The bedrock surface is the main "pathway" for mobile transient saturated flow during events (Freer *et al.*, 2002; McDonnell *et al.*, 1996).

Isotope composition and major ion concentration of subsurface flow have been used to test the physical mechanisms revealed by the pore water pressure and topographic analysis. ^{18}O analysis of collected trench flow often showed little evidence of "new water" breakthrough during storm events (Burns *et al.*, 1998). The Panola hillslope at Panola, Georgia, USA, is a useful example of hillslope flow behavior. For example, the Panola trench SO_4 chemistry showed no concentration-discharge relationship (Burns *et al.*, 1998). The only variability in concentrations of subsurface flow at the trench face was between neighboring troughs, suggesting waters of slightly different ages were mobilized within the hillslope. Burns *et al.* (1998) found that base cation concentrations in hillslope subsurface stormflow were generally related to flushing frequency, where parts of the trench with the highest bedrock surface drainage area had consistently lower mean base cation concentrations than other trench face positions with lower bedrock surface drainage areas.

The relative contributions of different parts of the Panola hillslope change with total precipitation, antecedent wetness, total flow, and season (Tromp-van Meerveld, 2004). This suggests that the bedrock topography might not be the only dominant control on subsurface flow as was suggested by the analysis of only a few storms by McDonnell (1997) and McDonnell *et al.* (1996) and Freer *et al.* (1997, 2002). Together with bedrock contributing area, soil depth may control subsurface stormflow dynamics during storms, especially for smaller events (Buttle *et al.*, 2004). At the Panola hillslope, subsurface storm flow volume of smaller events (less than the 55 mm rainfall threshold) or events with dry antecedent conditions are controlled largely by soil depth variations in space (Tromp-van Meerveld, 2004), while the bedrock topography is the primary control during medium to large storms. The shift in dominant areas of flow appears to be related to increasing antecedent soil moisture, storm size, and total flow (Freer *et al.*, 2002).

Despite the fact that bedrock topography and soil depth variability are important controls on subsurface stormflow, few models have conceptualized and implemented this into simulation models. Recently, Weiler and McDonnell (2004) incorporated soil depth variability into a subsurface flow model and simulated subsurface stormflow for the Panola hillslope. They could show that soil depth variations not only have a large influence on the spatial variation of subsurface flow but control largely the total subsurface flow volume produced. Figure 5 shows observed and modeled subsurface flow for simulations using Hillvi (**http://faculty.forestry.ubc.ca/weiler/hillvi.html**), a spatially explicit saturated and unsaturated water balance model for the actual measured soil depth variations at the Panola and Maimai experimental hillslopes and for an assumed constant average soil depth at those sites. All other model parameters were kept constant. For the Panola hillslope in particular, the differences in subsurface stormflow response are considerable – changes are also notable for the Maimai hillslope. These differences seem to be related mainly to the total variance of soil depth and the measured spatial correlation length of soil depth.

The preferential nature of flow along the soil-bedrock interface may also partly account for the often observed discrepancy between measurements of saturated hydraulic

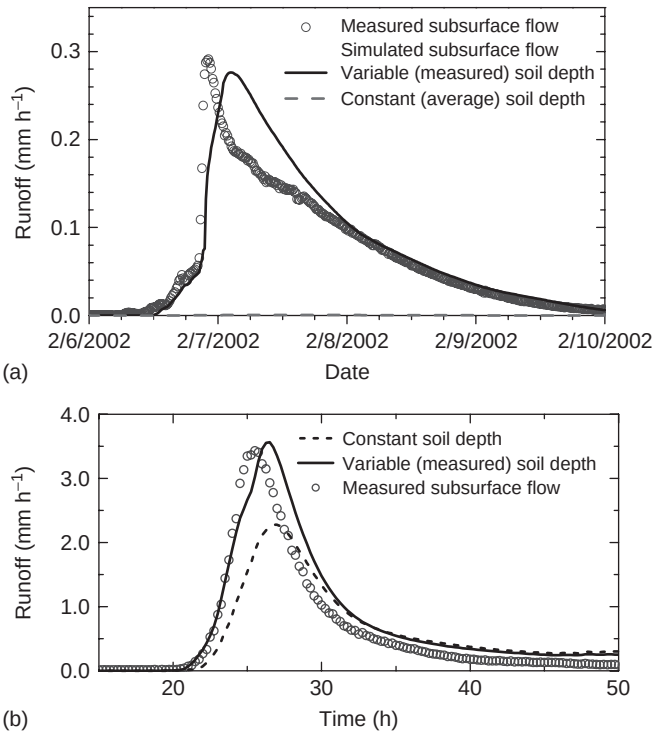


Figure 5 Simulated and measured subsurface flow response for the experimental hillslope at Panola (USA) – (a) and Maimai (NZ) – (b) assuming a constant soil depth and the measured variable soil depth. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

conductivity at the plot scale and the calculated hydraulic conductivity assuming lateral homogeneous matrix flow from measurements of lateral subsurface flow at the hillslope scale (Brooks *et al.*, 2004; Uchida *et al.*, 2002; Weiler *et al.*, 1998). As the results in Figure 5 show, subsurface flow can increase by orders of magnitude because of the variations in soil depth without changing the saturated hydraulic conductivity.

Subsurface Flow in Soil Pipes/Macropores

In many environments, subsurface stormflow is dominated by lateral macropore flow, ranging from subarctic wetland (Woo and DiCenzo, 1989) and boreal forest (Roberge and Plamondon, 1987) to tropical rain forest (Elsenbeer and Lack, 1996) and semiarid land (Newman *et al.*, 1998). These macropores are commonly referred to as *soil pipes* and concentrated subsurface flow in these natural soil pipes is called as *pipeflow* (e.g. Jones, 1971; Jones, 1981). Empirical studies of lateral pipeflow found in the literature show that lateral pipeflow controls hillslope response (e.g. Uchida *et al.*, 1999), nutrient flushing (Buttle *et al.*, 2001), and old water delivery to streams (Freer *et al.*, 2002; McDonnell, 1990) and to riparian zones (McGlynn and McDonnell, 2003a).

Recently, soil pipe morphology has been studied by means of dye tracer, fiberscope, and ground penetrating radar. Results show that: (i) soil pipes in gentle moorland environments are complex and long (>50 m) (e.g. Holden and Burt, 2002; Jones, 1987), while soil pipe connectedness and length mapping in forested steep hillslopes show that soil pipes are highly discontinuous with maximum lengths of usually only a couple of meters (e.g. Michihata *et al.*, 2001; Noguchi *et al.*, 1999); and (ii) the position of soil pipes in gentle moorland environments varies from being shallow to deep within the peat layer. The pipe can be entirely within the peat, at the peat-substrate interface, or entirely within the substrate (e.g. Holden and Burt, 2002). The position of the soil pipes in steep forested hillslopes is often at or near the soil-bedrock interface (e.g. McDonnell, 1990; Terajima, 2002).

Runoff characteristics of pipeflow have been examined at hillslopes in Japan, the United Kingdom, North America, and Peru. These studies suggest that the maximum discharge of pipeflow is determined mainly by soil pipe diameter. Many studies have also shown that a precipitation threshold exists when pipeflow starts to dominate subsurface flow. The observed precipitation threshold depends strongly on the antecedent soil water content (Noguchi *et al.*, 2001; Tromp-van Meerveld, 2004; Uchida *et al.*, 1999). Figure 6 shows the linear relationship between pipeflow and total subsurface stormflow that was observed in several forested hillslopes around the world (Uchida *et al.*, 2005).

Only recently have some models considered the role of pipeflow on runoff generation (Faeh *et al.*, 1997; Jones and Connelly, 2002; Kosugi *et al.*, 2004; Weiler *et al.*, 2003). Nevertheless, these studies have not yet fully incorporated field perceptions into their numerical models. Although they are now viewed as a major subsurface flow control, the effect of soil pipes and other structures on lateral flow and transport at the hillslope scale still awaits good model-process integration.

Thresholds and Nonlinearities

Since the IHD, hydrologists have believed that subsurface stormflow rate changes linearly with changes in rainfall magnitude (Hewlett and Hibbert, 1967). Only recently has the nonlinearity of the subsurface stormflow process been fully realized (Buttle *et al.*, 2004; McDonnell, 2003). Many recent studies have examined the relation between the amount of precipitation and the volume of subsurface storm flow – documenting thresholdlike, nonlinear relationships between the two. In this section, we describe briefly how subsurface storm flow volume changes with rainfall amount and discuss what processes control the change in the hydrological response of subsurface storm flow.

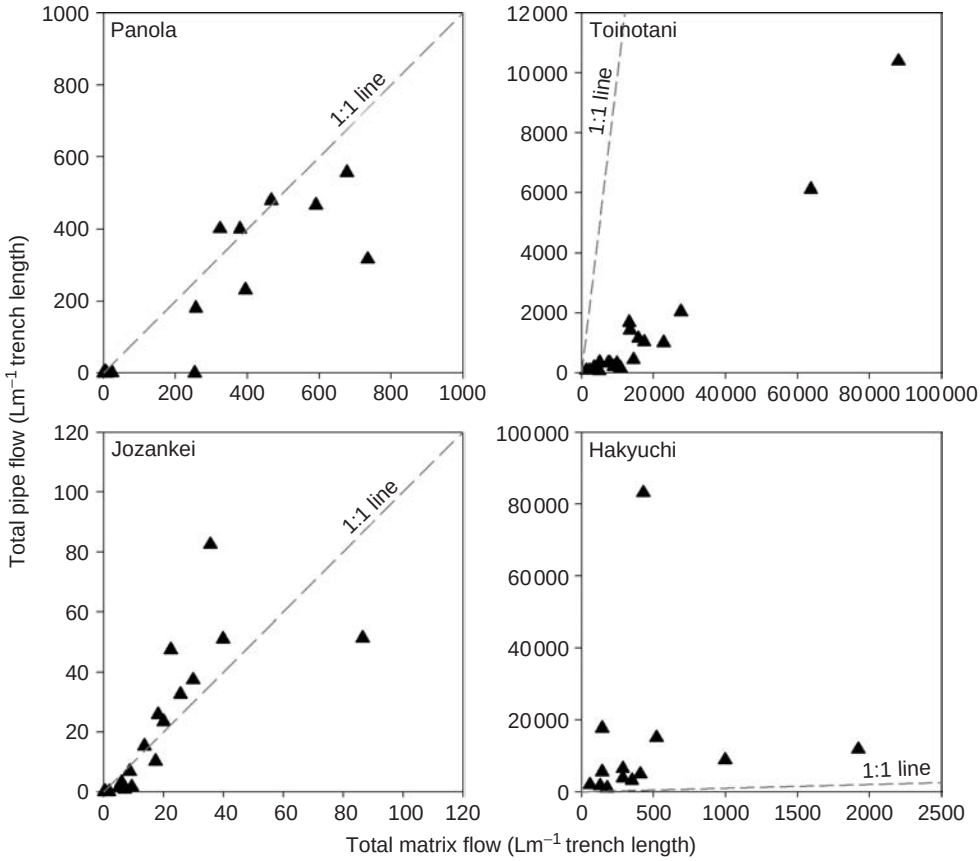


Figure 6 Relationship between total stormflow and total pipe flow of each storm at (a) Panola, Georgia, USA, (b) Toinotani, Kyoto, Japan, (c) Jozankei, Hokkaido, Japan and (d) Hakyuchi, Tokyo, Japan. Data for Jozankei and Hakyuchi compiled from Kitahara *et al.* (1994) and Ohta *et al.* (1981), respectively

Field Evidence of Threshold and Nonlinear Responses

Despite the lack of any formal recognition conceptually, data in the hillslope hydrology literature suggest that precipitation thresholds for subsurface stormflow generation may be a widespread phenomena. Revisiting data from many trenched hillslope studies (e.g. Mosley, 1979; Whipkey, 1965) suggests that the precipitation threshold for subsurface stormflow commencement lies commonly between 15 and 35 mm. More recent studies show that the precipitation threshold depends on the antecedent moisture conditions in the hillslope (Guebert and Gardner, 2001; Noguchi *et al.*, 2001; Peters *et al.*, 1995; Uchida *et al.*, 1999). Tani (1997), for example, has pointed out that after the threshold was reached, there was an almost 1 : 1 relation between precipitation above the threshold and subsurface flow. A simplified summary of the relation between storm total precipitation and the threshold for subsurface flow is given in Figure 7 for four sites. These relations suggest that while hillslopes vary tremendously from place to place, the subsurface stormflow initiation threshold and the slope of the line thereafter may be an emergent property of the hillslope hydrological system.

The recent analysis of 147 storms at the Panola trenched hillslope in the United States is one of the first to explicitly examine and document the relation between rainfall and subsurface stormflow over a large range of storm sizes, seasons, and antecedent wetness conditions (Tromp-van Meerveld and McDonnell, 2004). They found that macropore and matrix flow had very similar thresholds (around 55 mm) for significant (>1 mm) subsurface flow at the trench. The value of the threshold precipitation was related secondarily to antecedent soil moisture. Only 8 of the 147 storms (an average of 4.3 storms per year) exceeded the precipitation threshold to produce significant lateral subsurface stormflow in the Panola watershed. For these storms, the total amount of subsurface flow was between 30% and 80% of the total precipitation after the precipitation threshold.

Searching for First-order Control of Thresholdlike Responses

Notwithstanding the observation of thresholdlike subsurface stormflow, hydrologists still lack a clear understanding of the processes responsible. Recent studies have

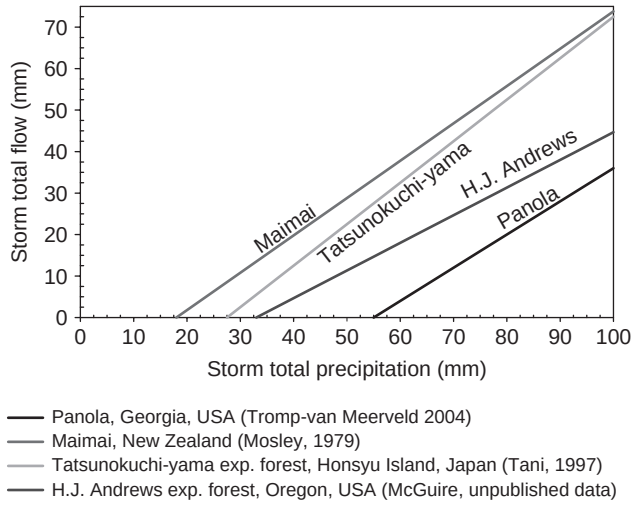


Figure 7 Schematic representation of the thresholdlike relationship between total storm precipitation and storm total flow under wet antecedent conditions. The lines represent the best fit lines through maximum storm total subsurface stormflow data points. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

shown that there are at least two common controls for a threshold response: (i) interconnection of lateral preferential flow paths; and (ii) development and extension of transient saturated areas. Tani (1997) observed a large subsurface runoff response after smaller areas of transient saturated zones in the hillslope became connected. Using data from another catchment in Japan, Sidle *et al.* (2000) postulated that with increasing wetness, subsurface flow expands over greater slope distances, such that during very wet conditions the hillslopes become linked to the channel system to enhance subsurface flow significantly.

Recently, Spence and Woo (2003) and Tromp-van Meerveld and McDonnell (2004) proposed a fill and spill mechanism to account for threshold behavior at the hillslope scale. Tromp-van Meerveld and McDonnell (2004) used a spatially distributed grid of piezometers to obtain spatial and temporal information of transient saturation at the Panola trenched hillslope. Both piezometer data and a two-dimensional finite element model indicated that for storms smaller than the observed precipitation threshold, subsurface saturation occurred first in areas where the soil was shallowest and then expanded to areas with deeper soils, but did not flow further downslope because it was blocked downslope by a bedrock ridge. During larger storms, transient saturation filled up local depressions in the bedrock microtopography before water spilled laterally over the bedrock ridges between the depressions and flowed further downslope through the bedrock lows (Figure 8). Only during storms larger than the precipitation threshold did these transient saturated areas become connected to the trench face, resulting in significant (>1 mm) subsurface flow. Figure 8 illustrates how soil depth and bedrock microtopographic relief (i.e. the bedrock lows) together control the connection of subsurface saturated areas to the trench face causing the observed threshold for subsurface flow (see also Figure 7).

OUTLOOK AND CONCLUSION

Subsurface stormflow is the main mechanism of runoff generation in many catchments around the world and is significant for the initiation of landslides, for the flushing of labile nutrients and hence the water quality changes, and the hydrograph response in streams. Most subsurface stormflow studies published to date have focused on only a few storms at any given site. Up until recently, it has been difficult to derive general hydrologic principles from single research studies within intensively studied small basins (Jones and Swanson, 2001). Comparing different well-instrumented sites allows extraction of the commonalities or major differences between sites and helps define the first-order controls on subsurface stormflow generation and flow pathways.

A question for future research on subsurface stormflow is how we should compare different sites and what tools we should use to define first-order controls. First, we need to continue with our experimental investigations to understand better the internal processes. New experimental technologies like ground-penetrating radar (Huisman *et al.*, 2003) or electromagnetic induction (Sherlock and McDonnell, 2003) will be important. Detailed measurement of flow combined with isotope and chemical tracing in time and space will be essential to constrain our conceptual and numerical models of subsurface stormflow.

Site intercomparison will require an organizational framework to summarize and evaluate the comparative analysis. Scherrer and Naef (2003) proposed a decision tree to define the dominant hydrological flow processes on a variety of grassland sites in Switzerland. They showed, rather convincingly, that the decision tree approach can be a powerful tool to clarify the first-order controls on dominant runoff processes. A decision tree may be a useful organizational framework for subsurface stormflow science to summarize and organize comparative analyses and to provide a structure for defining the hierarchy of process controls that are necessary for model development.

Process controls and new conceptualization for model development will also be a prerequisite for simulating and predicting the links between biogeochemical-hydrological aspects of subsurface stormflow. Recently, attempts have begun to derive the primary controls of transport processes and nutrient flushing in hillslopes dominated by

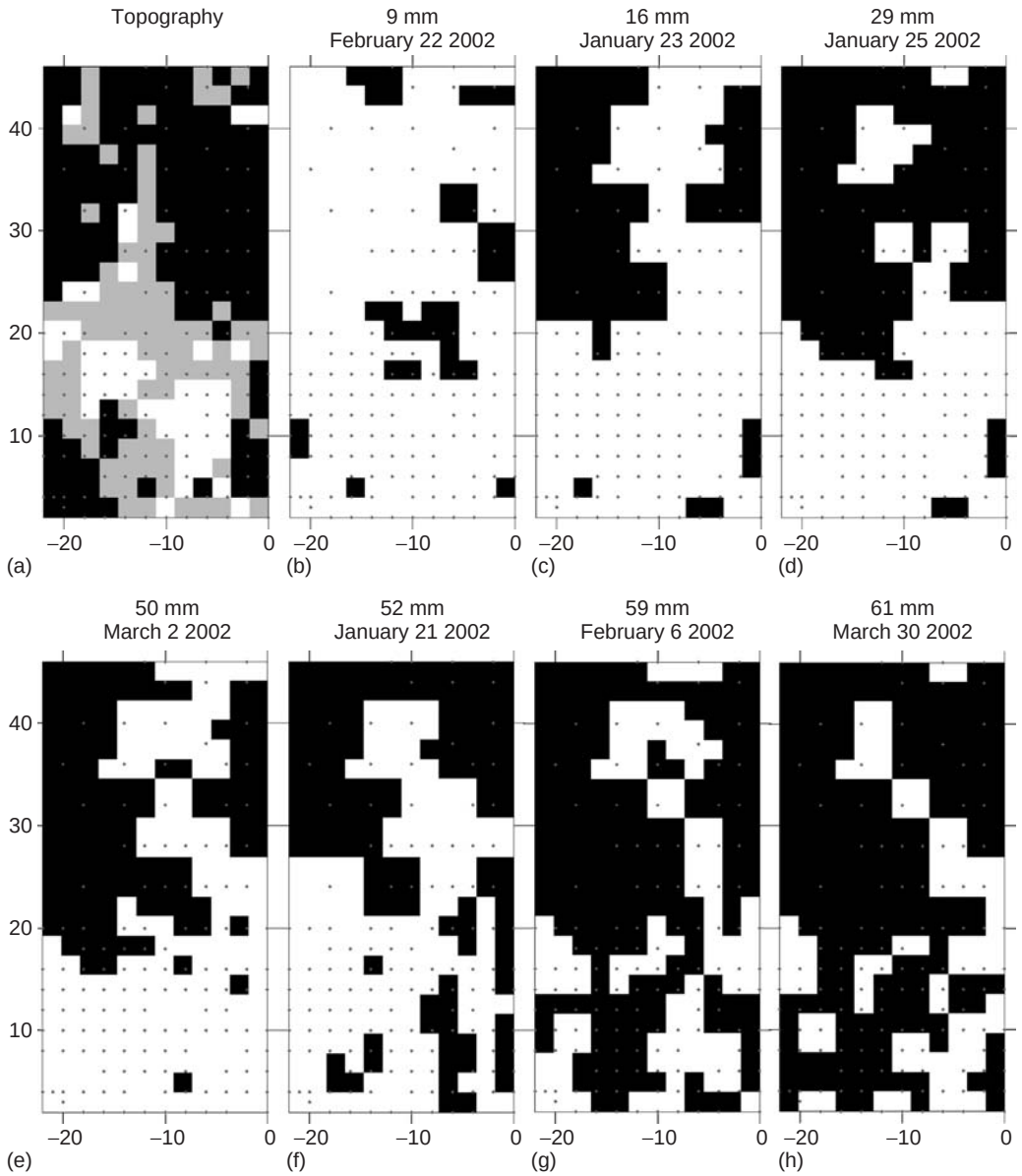


Figure 8 The location of areas of shallow soil depth (<0.75 m) (in black) and the areas with high (>4) bedrock topographic index (grey) (a) and the observed spatial distribution of subsurface saturation at the soil-bedrock interface across the Panola hillslope with increasing precipitation (b–h). The shaded area represents the area where transient subsurface saturation was observed; the unshaded area indicates the area where no subsurface saturation was observed. The diamonds represent the locations of the piezometers. Linear triangulation was used to interpolate between the measurement points. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

subsurface stormflow (e.g. Stieglitz *et al.*, 2003; Weiler and McDonnell, 2005). Nevertheless, the observed variability for soil properties and structural features (e.g. macropores and pipes) still need to be implemented and conceptualized. When introducing variability (as shown, for example, in Figure 5), we may not be able to predict deterministically where and when water is flowing at a specific location, but we may be able to describe the distribution of subsurface flow variability and its effects on water age, origin, and timing.

REFERENCES

Anderson M.G. and Burt T.P. (1978) The role of topography in controlling throughflow generation. *Earth Surfaces Processes and Landforms*, **3**, 331–334.

Anderson M.G. and Burt T.P. (1990a) *Process Studies in Hillslope Hydrology*, John Wiley and Sons: Chichester, p. 539.

Anderson M.G. and Burt T.P. (1990b) Subsurface runoff. In *Process Studies in Hillslope Hydrology*, Anderson M.G. and Burt T.P. (Eds.), John Wiley & Sons: New York, pp. 365–400.

Beasley R.S. (1978) Contribution of subsurface flow from the upper slopes of a forested watershed to channel flow. *Soil Science Society Of America Journal*, **40**, 955–957.

Beven K. and Germann P. (1982) Macropores and water flow in soils. *Water Resources Research*, **18**(5), 1311–1325.

Bonell M. (1993) Progress in the understanding of runoff generation dynamics in forests. *Journal of Hydrology*, **150**, 217–275.

Bonell M. (1998) Selected challenges in runoff generation research in forests from the hillslope to headwater drainage basin scale. *Journal of the American Water Resources Association*, **34**(4), 765–786.

Brammer D.D., McDonnell J.J., Kendall C. and Rowe L.K. (1995) Controls on the downslope evolution of water, solutes and isotopes in a steep forested hillslope. *Transactions of the American Geophysical Union*, **76**(46), 268.

Brooks E.S., Boll J. and McDaniel P.A. (2004) A hillslope-scale experiment to measure lateral saturated hydraulic conductivity. *Water Resources Research*, **40**(W04208), doi:10.1029/2003 WR002858.

Burns D.A., Hooper R.P., McDonnell J.J., Freer J.E., Kendall C. and Beven K. (1998) Base cation concentrations in subsurface flow from a forested hillslope: The role of flushing frequency. *Water Resources Research*, **34**(12), 3535–3544.

Burns D.A., Plummer L.N., McDonnell J.J., Busenberg E., Casile G.C., Kendall C., Hooper R.P., Freer J.E., Peters N.E., Beven K.J. and Schlosser P. (2003) The geochemical evolution of riparian ground water in a forested piedmont catchment. *Ground Water*, **41**(7), 913–925.

Burt T.P. (1989) Storm runoff generation in small catchments in relation to the flood response of large basins. In *Floods, Hydrological, Sedimentological and Geomorphological Implications*, Beven K. and Carling P. (Eds.), John Wiley & Sons: pp. 11–35.

Burt T.P. and Butcher D.P. (1985) Topographic controls of soil moisture distributions. *Journal of Soil Science*, **36**(3), 469–486.

Buttle J.M. (1994) Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins. *Progress in Physical Geography*, **18**(1), 16–41.

Buttle J.M., Dillon P.J. and Eerkes G.R. (2004) Hydrologic coupling of slopes, riparian zones and streams: an example from the Canadian Shield. *Journal of Hydrology*, **287**(1–4), 161–177.

Buttle J.M., Lister S.W. and Hill A.R. (2001) Controls on runoff components on a forested slope and implications for N transport. *Hydrological Processes*, **15**, 1065–1070.

Cappus P. (1960) Bassin expérimental d'Alrance - Étude des lois de l'écoulement - Application au calcul et à la prévision des débits. *La Houille Blanche*, **A**, 493–520.

Dunne T. (1978) Field studies of hillslope flow processes. In *Hillslope Hydrology*, Kirkby M.J. (Ed.), Wiley: Chichester, pp. 227–293.

Dunne T. and Black R.D. (1970) Partial area contributions to storm runoff in a small New England watershed. *Water Resources Research*, **6**(5), 1296–1311.

Elsenbeer H. and Lack A. (1996) Hydrometric and hydrochemical evidence for fast flowpaths at La Cuenca, Western Amazonia. *Journal of Hydrology*, **180**, 237–250.

Engler A. (1919) *Untersuchungen über den Einfluss des Waldes auf den Stand der Gewässer, 12*. Kommissionsverlag von Beer & Cie: Zürich, p. 626.

Faeh A.O., Scherrer S. and Naef F. (1997) A combined field and numerical approach to investigate flow processes in natural macroporous soils under extreme precipitation. *Hydrology and Earth System Sciences*, **1**(4), 787–800.

Flury M., Flühler H., Jury W.A. and Leuenberger J. (1994) Susceptibility of soils to preferential flow of water: a field study. *Water Resources Research*, **30**(7), 1945–1954.

Freer J., McDonnell J., Beven K.J., Brammer D., Burns D., Hooper R.P. and Kendall C. (1997) Topographic controls on subsurface storm flow at the hillslope scale for two hydrologically distinct small catchments. *Hydrological Processes*, **11**(9), 1347–1352.

Freer J., McDonnell J.J., Beven K.J., Peters N.E., Burns D.A., Hooper R.P. and Aulenbach B. (2002) The role of bedrock topography on subsurface storm flow. *Water Resources Research*, **38**(12), 1269, doi:10.1029/2001WR000872.

Freeze A.R. (1972) Role of subsurface flow in generating surface runoff. 2.Upstream source areas. *Water Resources Research*, **8**(5), 1272–1283.

Germann P.F. (1990) Macropores and hydrologic hillslope processes. In *Process Studies in Hillslope Hydrology*, Anderson M.G. and Burt T.P. (Eds.), John Wiley & Sons: New York.

Guebert M.D. and Gardner T.W. (2001) Macropore flow on a reclaimed surface mine: infiltration and hillslope hydrology. *Geomorphology*, **39**, 151–169.

Gutknecht D. (1996) Abflussentstehung an Hängen – Beobachtungen und Konzeptionen. *Österreichische Wasser- und Abfallwirtschaft*, **48**(5/6), 134–144.

Harr R.D. (1977) Water flux in soil and subsoil on a steep forested slope. *Journal of Hydrology*, **33**, 37–58.

Hewlett J.D. and Hibbert A.R. (1963) Moisture and energy conditions within a sloping soil mass during drainage. *Journal of Geophysical Research*, **68**, 1081–1087.

Hewlett J.D. and Hibbert A.R. (1967) Factors affecting the response of small watersheds to precipitation in humid areas. In *Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: New York, pp. 275–291.

Holden J. and Burt T.P. (2002) Piping and pipeflow in a deep peat catchment. *Catena*, **48**, 163–199.

Hoover M.D. and Hursh C.R. (1943) Influence of topography and soil depth on runoff from forest land. *Transactions of the American Geophysical Union*, **2**, 693–698.

Huisman J.A., Hubbard S.S., Redman J.D. and Annanc A.P. (2003) Measuring soil water content with ground penetrating radar: a review. *Vadose Zone Journal*, **2**, 476–491.

Hursh C.R. and Brater E.F. (1941) Separating storm-hydrographs from small drainage-areas into surface- and subsurface-flow. *Transactions of the American Geophysical Union*, **3**, 863–871.

Hutchinson D.G. and Moore R.D. (2000) Throughflow variability on a forested hillslope underlain by compacted glacial till. *Hydrological Processes*, **14**(10), 1751–1766.

Jones J.A.A. (1971) Soil piping and stream channel initiation. *Water Resources Research*, **7**(3), 602–610.

Jones J.A.A. (1981) *The Nature of Soil Piping: a Review of Research*, British Geomorphological Research Group Monograph Series(3).

Jones J.A.A. (1987) The effects of soil piping on contributing area and erosion pattern. *Earth Surface Processes*, **12**, 229–248.

Jones J.A.A. and Connelly L.J. (2002) A semi-distributed simulation model for natural pipeflow. *Journal of Hydrology*, **262**(1–4), 28–49.

Jones J.A. and Swanson F.J. (2001) Hydrologic inferences from comparisons among small basin experiments. *Invited Commentary. Hydrological Processes*, **15**(12), 2363–2366.

Kitahara H., Terajima T., Nakai Y. (1994) Ratio of pipe flow to through flow. *Journal of Japanese Forestry Society*, **76**, 10–17 (in Japanese with English summary).

Kosugi K., Uchida T. and Mizuyama T. (2004) Numerical calculation of soil pipe flow and its effect on water dynamics in a slope. *Hydrological Processes*, **18**(4), 777–789.

McDonnell J.J. (1989) *The Age, Origin and Pathway of Subsurface Stormflow in a Steep Humid Catchment*, University of Canterbury: Christchurch, p. 270.

McDonnell J.J. (1990) A rationale for old water discharge through macropores in a steep, humid catchment. *Water Resources Research*, **26**(11), 2821–2832.

McDonnell J.J. (1997) Comment on "the changing spatial variability of subsurface flow across a hillslide" by Ross Woods and Lindsay Rowe. *Journal of Hydrology, New Zealand*, **36**(1), 97–100.

McDonnell J.J. (2003) Where does water go when it rains? Moving beyond the variable source area concept of rainfall-runoff response. *Hydrological Processes*, **17**(9), 1869–1875.

McDonnell J.J., Freer J., Hooper R., Kendall C., Burns D., Beven K.J. and Peters J. (1996) New method developed for studying flow in hillslopes. *EOS, Transactions of the American Geophysical Union*, **77**(47), 465.

McDonnell J.J., Owens I.F. and Stewart M.K. (1991) A case study of shallow flow paths in a steep zero-order basin. *Water Resources Bulletin*, **27**(4), 679–685.

McGlynn B.L. and McDonnell J.J. (2003a) Quantifying the relative contributions of riparian and hillslope zones to catchment runoff. *Water Resources Research*, **39**(11), doi:10.1029/2003WR002091.

McGlynn B.L. and McDonnell J.J. (2003b) Role of discrete landscape units in controlling catchment dissolved organic carbon dynamics. *Water Resources Research*, **39**(4), doi:10.1029/2002WR001525.

McGlynn B.L., McDonnell J.J. and Brammer D.D. (2002) A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand. *Journal of Hydrology*, **257**, 1–26.

Michihata R., Uchida T., Kosugi K. and Mizuyama T. (2001) An observation of soil pipes morphology at Toinotani hollow in Ashu Experimental Forest. *Forest Research Kyoto*, **73**, 67–70.

Montgomery D.R. *et al.* (1997) Hydrologic response of a steep, unchanneled valley to natural and applied rainfall. *Water Resources Research*, **33**(1), 91–109.

Mosley M.P. (1979) Streamflow generation in a forested watershed. *Water Resources Research*, **15**, 795–806.

Mosley M.P. (1982) Subsurface flow velocities through selected forest soils, South Island, New Zealand. *Journal of Hydrology*, **55**, 65–92.

Newman B.D., Campbell A.R. and Wilcox B.P. (1998) Lateral subsurface flow pathways in a semiarid ponderosa pine hillslope. *Water Resources Research*, **34**(12), 3485–3496.

Noguchi S., Tsuboyama Y., Sidle R.C. and Hosoda I. (1999) Morphological characteristics of macropores and the distribution of preferential flow pathways in a forested slope segment. *Soil Science Society of America Journal*, **63**, 1413–1423.

Noguchi S., Tsuboyama Y., Sidle R.C. and Hosoda I. (2001) Subsurface runoff characteristics from a forest hillslope soil profile including macropores, Hitachi Ohta, Japan. *Hydrological Processes*, **15**, 2131–2149.

Ohta T., Tsukamoto Y. and Noguchi H. (1981) An analysis of pipeflow and landslide. *In Proceedings of Annual Meeting of the Japan Society of Erosion Control Engineering*, 92–93 (in Japanese).

Pearce A.J., Stewart M.K. and Sklash M.G. (1986) Storm runoff generation in humid headwater catchments: 1. Where does the water come from? *Water Resources Research*, **22**, 1263–1272.

Peters D.L., Buttle J.M., Taylor C.H. and LaZerte B.D. (1995) Runoff production in a forested, shallow soil, Canadian Shield basin. *Water Resources Research*, **31**(5), 1291–1304.

Ranken D.W. (1974) *Hydrologic Properties of Soil and Subsoil on a Steep, Forested Slope*, M.S. Thesis, Oregon State University, Corvallis, p. 114.

Rasmussen T.C., Baldwin R.H., Dowd J.F. and Williams A.G. (2000) Tracer vs. pressure wave velocities through unsaturated saprolite. *Soil Science Society of America Journal*, **64**, 75–85.

Roberge J. and Plamondon A.P. (1987) Snowmelt runoff pathways in a boreal forest hillslope, the role of pipe throughflow. *Journal of Hydrology*, **95**, 39–54.

Rodhe A. (1987) *The Origin of Streamwater Traced by Oxygen-18*, PhD Dissertation Thesis, Uppsala University, Uppsala, p. 260.

Rothacher J., Dyrness C.T. and Fredriksen R.L. (1967) *Hydrologic and Related Characteristics of Three Small Watersheds in the Oregon Cascades*, U.S. Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station, Portland.

Scherrer S. (1997) *Abflussbildung bei Starkniederschlägen*, Identifikation von Abflussprozessen mittels künstlicher Niederschläge.

Scherrer S. and Naef F. (2003) A decision scheme to indicate dominant hydrological flow processes on temperate grassland. *Hydrological Processes*, **17**(2), 391–401.

Sherlock M.D. and McDonnell J.J. (2003) A new tool for hillslope hydrologists: spatially distributed groundwater level and soilwater content measured using electromagnetic induction. *Hydrological Processes*, **17**(10), 1965–1977.

Sidle R.C. and Tsuboyama Y. (1992) A comparison of piezometric response in unchanneled hillslope hollows: coastal Alaska and Japan. *Journal of the Japanese Society of Hydrology and Water Resources*, **5**, 3–11.

Sidle R.C., Tsuboyama Y., Noguchi S., Hosoda I., Fujieda H. and Shimizu T. (2000) Stormflow generation in steep forested

headwaters: a linked hydrogeomorphic paradigm. *Hydrological Processes*, **14**(3), 369–385.

Sklash M.G., Stewart M.K. and Pearce A.J. (1986) Storm runoff generation in humid headwater catchments: 2. a case study of hillslope and low-order stream response. *Water Resources Research*, **22**(8), 1273–1282.

Smith R.E. and Hebbert R.H.B. (1983) Mathematical simulation of interdependent surface and subsurface hydrologic processes. *Water Resources Research*, **19**(4), 987–1001.

Spence C. and Woo M.-k. (2003) Hydrology of subarctic Canadian shield: soil-filled valleys. *Journal of Hydrology*, **279**, 151–166.

Stieglitz M. *et al.* (2003) An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport. *Global Biogeochemical Cycles*, **17**(4), 1105, doi:10.1029/2003 GB002041.

Tani M. (1997) Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, **200**, 84–109.

Terajima T. (2002) Subsurface water discharge and sediment yield relevant to pipe flow in a forested 0-order basin, Hokkaido northern Japan. *Transactions, Japanese Geomorphological Union*, **23**, 511–535.

Torres R., Dietrich W.E., Montgomery D.R., Anderson S.P. and Loague K. (1998) Unsaturated zone processes and the hydrologic response of a steep, unchanneled catchment. *Water Resources Research*, **34**(8), 1865–1879.

Tromp-van Meerveld H.J. (2004) Hillslope Hydrology: from Patterns to Processes, Ph.D. dissertation, Oregon State University, Corvallis, 270 p.

Tsuboyama Y., Sidle R.C., Noguchi S. and Hosoda I. (1994) Flow and solute transport through the soil matrix and macropores of a hillslope segment. *Water Resources Research*, **30**(4), 879–890.

Tsukamoto Y. (1961) An experiment on sub-surface flow. *Journal of the Japanese Forestry Society*, **43**, 62–67.

Tsukamoto Y., Ohta T. and Noguchi H. (1982) Hydrogeological and geomorphological studies of debris slides on forested hillslopes in Japan. *IAHS Publication*, **137**, 89–98.

Uchida T., Kosugi K. and Mizuyama T. (1999) Runoff characteristics of pipeflow and effects of pipeflow on rainfall-runoff phenomena in a mountainous watershed. *Journal of Hydrology*, **222**(1–4), 18–36.

Uchida T., Kosugi K.i. and Mizuyama T. (2002) Effects of pipe flow and bedrock groundwater on runoff generation in a steep headwater catchment in Ashiu, central Japan. *Water Resources Research*, **38**(7), doi:10.1029/2001WR000261.

Uchida T., Tromp-van Meerveld H.J. and McDonnell J.J. (2005) The role of lateral pipe flow in hillslope runoff response: An intercomparison of nonlinear hillslope response. *Journal of Hydrology*, (in press).

Weiler M.H. (2001) *Mechanisms Controlling Macropore Flow During Infiltration: Dye Tracer Experiments and Simulations*, Dissertation Thesis, ETH Zürich: Zürich, p. 148.

Weiler M. and Flühler H. (2004) Inferring flow types from dye patterns in macroporous soils. *Geoderma*, **120**(1–2), 137–153.

Weiler M. and McDonnell J. (2004) Virtual experiments: a new approach for improving process conceptualization in hillslope hydrology. *Journal of Hydrology*, **285**(1–4), 3–18.

Weiler M. and McDonnell J.J. (2005) Testing nutrient flushing hypotheses at the hillslope scale: a virtual experiment approach. *Journal of Hydrology*, (in press).

Weiler M. and Naef F. (2003) An experimental tracer study of the role of macropores in infiltration in grassland soils. *Hydrological Processes*, **17**(2), 477–493.

Weiler M., Naef F. and Leibundgut C. (1998) Study of runoff generation on hillslopes using tracer experiments and physically based numerical model. *IAHS Publication*, **248**, 353–360.

Weiler M., Uchida T. and McDonnell J. (2003) Connectivity due to preferential flow controls water flow and solute transport at the hillslope scale, *Proceedings of MODSIM 2003*, Townsville.

Weyman D.R. (1973) Measurements of the downslope flow of water in a soil. *Journal of Hydrology*, **20**, 267–288.

Whipkey R.Z. (1965) Subsurface storm flow from forested slopes. *Bulletin of the International Association of Scientific Hydrology*, **2**, 74–85.

Wilcox B.P., Newman B.D., Bres D., Davenport D.W. and Reid K. (1997) Runoff from a semiarid ponderosa pine hillslope in New Mexico. *Water Resources Research*, **33**, 2301–2314.

Woo M.-k. and DiCenzo P.D. (1989) Hydrology of small tributary streams in a subarctic wetland. *Canadian Journal of Earth Science*, **26**, 1551–1556.

Woods R. and Rowe L. (1996) The changing spatial variability of subsurface flow across a hillside. *Journal of Hydrology (NZ)*, **35**(1), 51–86.

Wu W. and Sidle R.C. (1995) A distributed slope stability model for steep forested basins. *Water Resources Research*, **31**, 2097–2110.

Yee C.S. (1975) Soil and Hydrologic Factors Affecting Stability of Natural Slopes in the Oregon Coast Range, Ph.D. dissertation, Oregon State University, Corvallis, OR, 204 p.

113: Hyporheic Exchange Flows

KENNETH E BENCALA

United States Geological Survey, Menlo Park, CA, US

Water having entered a stream channel from the surrounding catchment may continue to have connections with the catchment. In the stream's hyporheic zone, water "in the channel" exchanges with "groundwater" in the bed of the stream. Hyporheic exchange flows typically occur at scales small relative to the length and volumetric transport characteristics of the stream. Nevertheless, it is well documented that hyporheic exchange flows significantly influence nutrient dynamics. Additionally, there is evidence of hyporheic exchange flows similarly influencing the processes establishing the concentrations of major-ions and metals in stream-catchment systems. It is within the contexts of (i) solute transport and (ii) the continuing connections of streams to their catchments that the hydrologic aspects of hyporheic exchange flows are studied. The Transient Storage Model (TSM), a pseudo-two-dimensional representation of stream and hyporheic zone solute transport, is used to identify characteristics of hyporheic zone, physical size, and solute residence times. The TSM is further extended to simulations of reactive solute transport to understand and interpret the biogeochemical processes of streams influencing the solute dynamics of catchments. Active hydrologic research continues to advance the process basis from which quantitative assessments of the role of hyporheic exchange flows will be made.

INTRODUCTION

Hyporheic exchange flows are one of several mechanisms of the interaction of the groundwater and surface water (Winter *et al.*, 1998, *see* **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4**). In the *hyporheic zones* of streams, water that is flowing in the stream channel flows into the subsurface materials of the streambed and then returns to the stream. Hyporheic zones function as continuing points of connection between the transport of water and solutes in the stream channel and the stream's catchment (Bencala, 1993).

Hyporheic zones are significant in the dynamics of nutrients within the stream-catchment system (*see* **Chapter 96, Nutrient Cycling, Volume 3**) and in the processes establishing the concentrations of major-ions and metals. The influence of hyporheic exchange upon the transport and transformation of solutes occurs in the environments where hydrologic and biogeochemical processes are dynamic and highly heterogeneous. Documenting the biogeochemical function of hyporheic exchange has thus been primarily accomplished in detailed high-sampling-intensity research studies. In spite of the inherent difficulties in isolating the function of hyporheic exchange amidst many other stream and catchment processes, it is clear that hyporheic exchange can be quantified as influencing the establishment of major-ion chemistry and the ongoing transformation of reactive constituents. Numerous investigators have completed extensive field studies documenting the biogeochemical processes active in the hyporheic zones of individual streams. An overview of a small sample of these studies forms the first sections of this article. The understanding of biogeochemical processes within hyporheic zones has further value when utilized in the context of quantitative assessment of solute transport and solute residence within the stream channel and the stream-catchment system. One approach to providing this context, the Transient Storage Model, is discussed below along with current research directions, suggesting that much work remains in interpreting the basic hydrologic mechanisms by which streams and their catchments are connected.

NUTRIENT DYNAMICS

The substantial body of empirical evidence establishing that surface-subsurface exchange significantly influences

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

nutrient dynamics in stream ecosystems is presented in the review article of Mulholland and DeAngelis (2000). As these writers state, "the expectation [of this exchange influence] is based on two factors: (1) high ratios of surface area of sediments to volume of water within sediments should result in large effects of microbial processes on subsurface water, and (2) relatively slow advective flow of water within the subsurface zone retards the downstream movement of soluble materials compared with the surface environment." Using a simple "nutrient spiraling" model, Mulholland and DeAngelis (2000) go on to demonstrate conceptually that the size of the subsurface zone and the rates of water exchange between surface and subsurface zones have substantial effects on nutrient dynamics and uptake lengths in stream ecosystems.

The influences of surface-subsurface exchange upon nutrient dynamics are recognized as being spatially heterogeneous and temporally variable. Butturini *et al.* (2003) present detailed data from a 2-year study of stream-aquifer hydrology and nitrate removal in which they observe a "riparian groundwater system characterized by drastic hydrological changes and by mixing of stream water with hillslope groundwater". In part due to the hydrologic change throughout the year, both nitrate removal and nitrate release are observed within the studied riparian system. The variability in hyporheic zone processes is further evident when considering the range of ecosystems now being studied. Butturini *et al.* (2003) studied an intermittent Mediterranean stream. Edwardson *et al.* (2003) studied Arctic tundra streams. The frank conclusions of Edwardson *et al.* (2003) highlight the difficulty in drawing sweeping inferences about the specific function of hyporheic zone processes. Edwardson *et al.* (2003) conclude: "...we expected that the presence of continuous permafrost in [the] Arctic environment would limit the importance of hyporheic processes, either physically (i.e., through the presence of a restricting thaw bulb in the permafrost) or biogeochemically (i.e., through low temperatures). Instead, we found that biogeochemical processes in the hyporheic zone of [the] Arctic streams are at least as important as it is in similar temperate stream ecosystems".

In addition to the numerous field studies of nutrient dynamics in hyporheic zones, experimental laboratory work has also been reported. Using sediment perfusion cores to study nitrogen transformations, Sheibley *et al.* (2003a,b) have measured the rates of nitrogen transformations in river-bed sediments. These laboratory studies integrate the hydrology of hyporheic exchange with the study of biological transformations to verify that groundwater-surface mixing controls the nitrification–denitrification coupling occurring in the river-bed sediments.

PROCESSES ESTABLISHING THE CONCENTRATIONS OF MAJOR-IONS AND METALS

In addition to influencing the nutrient dynamics of stream-catchment systems, hyporheic exchange flows similarly influence the processes establishing the concentrations of major-ions and metals in streams. The role of hyporheic exchange has been documented in pristine systems as well as in those with a high level of anthropogenic impacts.

Working in a glacial meltwater stream in Antarctica, Gooseff *et al.* (2002) studied the contribution of primary weathering to the in-stream concentrations of silica and potassium. The study of weathering in these streams represents an extreme example of a hydrologic situation in which there were no hillslope processes occurring to influence stream chemistry. Their analysis suggested that the continuous saturation and rapid flushing of the streambed sediment due to hyporheic exchange did facilitate chemical weathering.

Harvey and Fuller (1998), working in a drainage basin in Arizona contaminated through copper mining, found that manganese distributed in the stream–groundwater system affected the transport of trace metals. The cumulative effect of hyporheic exchange in the basin was to remove approximately 20% of the dissolved manganese flowing out of the drainage basin. Their further studies (Fuller and Harvey, 2000) demonstrated that decreased loading of trace metals in the stream was attributable to uptake of the trace metals by manganese oxides in the hyporheic zone that is enhanced by the ongoing manganese oxide formation.

Because the hyporheic zone is an interface between surface and subsurface waters, gradients may exist in oxidation/reduction conditions. For arsenic, the toxicity of this metal can be significantly different for the different oxidation states. Nagorski and Moore (1999) have demonstrated that this gradient has an influence on the form of arsenic mobilized in the hyporheic zone of a contaminated stream in Montana. The continual flux of the reduced form of arsenic to the stream maintains higher concentration of this form than would otherwise be expected in oxygenated surface water.

TRANSPORT MODELING – TRANSIENT STORAGE

The biogeochemical functions of the hyporheic zone occur in the context of solute transport along the stream. Solute transport models are a tool used in conjunction with field data to analyze, and interpret, exchange processes. Models of the transport processes of the interactions of streams and hyporheic zone have been developed over a wide range in physical complexity (Packman and Bencala, 2000). The relatively simple concept of "transient storage", which

has been used by several investigators, is included in the US Geological Survey modeling code OTIS (One-dimensional Transport with Inflow and Storage) (Runkel, 1998). The code and applications information are available at `http://co.water.usgs.gov/otis/`. The Transient Storage Model (TSM) builds upon the standard convection-dispersion model of one-dimensional solute transport down the length of the stream. To the convection-dispersion model, the TSM adds fixed-volume solute storage "boxes" along the stream channel. Solute exchange between the open stream channel and the storage boxes is modeled as controlled through first-order mass-transfer. The storage boxes are effectively the simplest attempt to model the influence that the hyporheic zone has on the transport of solutes; that is, solutes are continually being exchanged between the open stream channel and the storage boxes.

The TSM results in a pseudo-two-dimensional representation of stream and hyporheic zone solute transport; solute transport is one-dimensional (longitudinal) in the stream channel with the storage zones adding a limited second (lateral) dimension. In discussing the TSM, as it is implemented in the OTIS code, Harvey and Wagner (2000) explain both the degree to which the TSM can be used to characterize solute transport in the hyporheic zone and the limitations inherent to this simple representation of complex processes. Harvey *et al.* (1996) present detailed field hydrometric data showing that analysis of solute transport using the TSM leads to a bias in only characterizing the most rapid components of hyporheic exchanges. Wagner and Harvey (1997) then develop guidance for assessing the reliability of TSM characterizations. Scott *et al.* (2003) and Gooseff *et al.* (2005) have recently demonstrated the significance in using objective parameter estimation techniques as part of the application of OTIS to stream tracer data sets.

A further issue in the use of the TSM has been the comparison of results between stream systems. Runkel (2002) discussed several of the comparisons that have been proposed and developed a metric specifically for comparing, stream-to-stream, the significance of transient storage. The TSM can also be implemented to characterize solute transport and storage in terms of temporal moments of concentration–time distributions. Schmid (2003) discusses the development of this approach by several investigators. Wörman (2000) and Schmid (2004) compared the temporal characteristics of TSM solutions to other transport model formulations. Jonsson *et al.* (2004) extended the temporal moments approach to consideration of sorption behavior and long-term retention of reactive solutes in hyporheic zones. In analysis of their field studies, "the method of temporal moments was found to be inadequate for parameter determination, whereas fitting versus the entire tracer breakthrough curves with special emphasis on the tail indicates that [their] proposed model could be used to represent both conservative and reactive transport".

RESIDENCE TIME AND CONNECTION TO THE CATCHMENT

The hyporheic zone influences stream biogeochemistry by increasing solute residence time within the stream-catchment system. The influence of the increased residence time may be enhanced in cases in which the overall biogeochemistry and extent of contact with sediment is distinct within the hyporheic zone compared to the adjacent open stream channel. A variety of approaches are being taken in identifying characteristic solute residence time in hyporheic zones as a significant factor in the connections of streams to their catchments. Harvey *et al.* (2003) used an analysis based on OTIS simulations of five stream tracer experiments to "conclude that relatively simple measurements of channel friction are useful for predicting the response of hydrologic retention in streams to major adjustments in channel morphology as well as changes in streamflow". Identifying characteristic storage zone residence time with a TSM was a step in the work of Thomas *et al.* (2003) to estimating the proportion of reactive nitrogen solute uptake occurring within hyporheic zones.

Although OTIS is used by several investigators and it is the implementation of the TSM that is the context for most of the discussion in this article, the TSM is a highly idealized representation of the function of hyporheic zones. A consequence of an analysis based on OTIS is that a hyporheic zone is characterized by a single residence time distribution. (The single residence time distribution is of exponential form. See Harvey *et al.* (1996) for discussion of this distribution form. The analysis is well beyond the scope of this article to show that a single form is indeed an implicit assumption of the TSM.) Haggerty *et al.* (2002) studied a power-law residence time distribution showing that "the hyporheic zone has a very large range of timescales". Gooseff *et al.* (2003) continued this line of work with analysis of three tracer experiments using a General Residence Time Distribution (RTD) model that is not bound by the implicit assumption of an exponential distribution. The basic result of the analysis is that the General RTD model allowed for a more accurate characterization of longer solute residence times compared to the characterization from OTIS. Gooseff *et al.* (2003) conclude: "Consequently the two models will result in different views of the hyporheic zone and its role in stream ecosystem processes." Haggerty *et al.* (2002) and Gooseff *et al.* (2003) have identified a significant aspect (significant for the understanding of solute transport and transformation in streams) of the complexity of hyporheic zone processes that the TSM's simplicity does not represent.

Residence time in the hyporheic zone is typically considered, as in the discussion above, in the framework of in-stream solute transport. Chanat and Hornberger (2003) use the concept of a "near-stream zone" to develop an analysis showing that mixing in this zone of chemically distinct catchment waters influences the timing of the solute balance flowing into a stream. Although they are using a different terminology for a somewhat different conceptualization of stream-catchment connections, the results of Chanat and Hornberger suggest that residence in the hyporheic zone can also be a factor in catchment-scale transport to a stream.

STREAMS TO RIVERS AND AQUIFERS

Most research papers written on hyporheic zones are undoubtedly about work done in streams, most often low-order, high-gradient streams. A few investigators have started working in the hyporheic zones of higher-order streams or rivers. Paralleling the main discussion above, the listing of a small sample of papers is given here in the sequence: solutes observed in monitoring river hyporheic zones, TSM applications to solute transport in rivers, and investigation of the function of hyporheic zones in connecting streams to aquifers.

Cox *et al.* (2003) completed a tracer experiment in the Upper Santa Clara River (Los Angeles and Ventura Counties, California; 645 square-mile basin) in which tracer added to the surface water was measured in several hyporheic zone sampling locations. In the Willamette River (Oregon; 12 000 square-mile basin) Hinkle *et al.* (2001) demonstrated nitrate reduction occurring along hyporheic flowpaths.

In nine rivers in the Willamette Basin (Oregon), Laenen and Bencala (2001) used OTIS to characterize the transient storage that occurred in 20 tracer additions to surface water. The tracer additions were conducted for other study purposes and there was no direct sampling in likely areas of hyporheic exchange. Thus, the influence of hyporheic exchange could only be inferred from the modeling exercise. However, Fernald *et al.* (2001) completed a study focused upon hyporheic exchange in the Willamette River. Their work included tracer additions, tracer sampling in the subsurface, and modeling analysis with OTIS. They concluded that "significant amounts of water follow flow paths with 0.2–30 hour transient storage zone residence times".

In addition to increasing the scale of studies of hyporheic zones to river systems, the scale can also be increased in the sense of linking hyporheic processes into conjunctive stream-aquifer modeling. Two such model development efforts were presented by Lin and Medina (2003) and Fox and Durnford (2003). Development of these new models will potentially be significant in providing tools for analysis in systems in which the geohydrologic framework is on spatial scales larger than typical for low-order, high-gradient streams.

HYDROLOGIC PROCESSES

The significant environmental influences of hyporheic zones are definitely upon biogeochemical processes. In order to understand these processes, the hydrologic processes that determine the characteristics of hyporheic flow need to also be investigated and understood. Harvey and Bencala (1993) made field measurements of the topography of a streambed. They concluded that solute transport occurred along well-defined subsurface flow paths. They observed hyporheic flow paths as determined by meter-to-meter scale topography, with the flow paths beginning at the transition from pools to steeper channel units. Cardenas and Zlotnik (2003) used ground-penetrating radar to develop a three-dimensional model of channel deposits. The channel deposits were clearly heterogeneous and nonisotropic. Their measurements document that it is an oversimplification to characterize a streambed by a constant width and thickness.

Cardenas *et al.* (2004) took their work further to show that heterogeneity causes significant additional hyporheic zone flux. The presence of heterogeneity can either decrease or increase residence times in the hyporheic zone. Salehin *et al.* (2004) concluded that (when compared to hyporheic exchange in homogeneous streambeds) "the structural heterogeneity of the streambed sediments produces more spatially limited hyporheic exchange that occurs with greater spatial variability and at a higher overall rate".

Solute transport in the hyporheic zone has been predominantly investigated as an adjunct to solute transport in the open stream channel. A groundwater focus has also been taken by a few investigators. Kasahara and Wondzell (2003) used groundwater flow models to investigate the morphologic features that controlled hyporheic exchange flow. Their interpretations are consistent with other studies in finding that surface-visible channel morphologic features controlled the development of the hyporheic zones. Also using groundwater flow models, Storey *et al.* (2003) conclude that the key factors controlling exchange flow within the alluvial (hyporheic) zone were (i) hydraulic conductivity of the alluvium, (ii) the hydraulic gradient between upstream and downstream ends of the riffle, and (iii) the flux of groundwater entering the alluvium from the sides and beneath.

Work to study the environmental fluid mechanics of hyporheic flows has been carried out through experimental tests in recirculating flumes. Bedforms are one control of the rate of solute delivery from the stream to the streambed. Further, bedform-induced pumping can be produced in a

flume to be the dominant exchange mechanism. Marion *et al.* (2002) demonstrate the effects on solute exchange of the longitudinal dimension of bedforms. Zaramella *et al.* (2003) show that the TSM can represent advective exchange with shallow beds that have a defined exchange layer restricted by the presence of an impermeable boundary. However, the TSM does not represent well exchange with a relatively deep sediment bed, where flow along different advective paths in the bed yields a wide distribution of exchange timescale. Clogging of the streambed surface will often isolate deeper regions of the bed from the streamflow. Laboratory flume experiments reported by Packman and MacKay (2003) show kaolinite clay deposition in a sand bed and the resulting alteration of hyporheic exchange flows. Ren and Packman (2004) observed that "reactive colloids can substantially mediate the stream-subsurface exchange of contaminants and the colloid deposition can provide a mechanism of contaminant immobilization that is generally not considered in field studies . . .".

Transient storage of solutes may occur due to features of a stream (e.g. vegetation) that are unrelated to hyporheic flows. In one example of work on this issue, Salehin *et al.* (2003) studied solute transport and hyporheic exchange in vegetated and unvegetated stream reaches. Their work demonstrates the considerable inherent difficulties in trying to differentiate the effects of multiple individual processes from solute breakthrough curves.

CONTINUING WORK

This article focuses upon the hydrologic processes influencing solute transport. The process hydrology of hyporheic exchange flows continues as an area of active investigation. Hyporheic zones have an ecological role in the life cycle of stream organisms. For example, Malcolm *et al.* (2004) have studied the implications for salmon egg survival of hydrological influences on hyporheic water quality. In considering the ecological role of hyporheic waters, it is found that the transport of heat is also influenced. Malcolm *et al.* (2004) observed that "the amplitude of surface water temperature signals was damped considerably with depth into the hyporheic zone". Several authors have provided discussions of the open issues and potential directions for research. Again listed in parallel with the themes of this chapter, examples of these discussions are: (i) Stanley and Jones (2000); potential directions for hydrologic studies needed to support research on the biogeochemistry of hyporheic zones, (ii) Runkel *et al.* (2003); potential directions for development of solute transport modeling in the hyporheic zone, and (iii) Bencala (2000); potential directions for research to understand the hydrological framework of which hyporheic zones are a critical linkage element.

REFERENCES

Bencala K.E. (1993) A perspective on stream-catchment connections. *Journal of the North American Benthological Society*, **12**, 44.

Bencala K.E. (2000) Hyporheic zone hydrological processes. *Hydrological Processes*, **14**, 2797.

Butturini A., Bernal S., Hellin C., Nin E., Rivero L., Sabater S. and Sabater F. (2003) Influences of the stream groundwater hydrology on nitrate concentration in unsaturated riparian area bounded by an intermittent Mediterranean stream. *Water Resources Research*, **39**, 1–13, doi:10.1029/2001WR001260.

Cardenas M.B., Wilson J.L. and Zlotnik V.A. (2004) Impact of heterogeneity, bed forms, and stream curvature on subchannel hyporheic exchange. *Water Resources Research*, **40**, doi:10.1029/2004WR003008.

Cardenas M.B. and Zlotnik V.A. (2003) Three-dimensional model of modern channel bend deposits. *Water Resources Research*, **39**, doi:10.1029/ 2002WR001383.

Chanat J.G. and Hornberger G.M. (2003) Modeling catchment-scale mixing in the near-stream zone – implications for chemical and isotopic hydrograph separation. *Geophysical Research Letters*, **30**, 1091, doi:10.1029/2002GL016265.

Cox M.H., Mendez G.O., Kratzer C.R. and Reichard E.G. (2003) *Evaluation of Tracer Tests Completed in 1999 and 2000 on the Upper Santa Clara River, Los Angeles and Ventura Counties, California*, Investigations Report 03–4277, USGS Water-Resources, **http://water.usgs.gov/pubs/wri/wrir034277/**.

Edwardson K.J., Bowden W.B., Dahm C. and Morrice J. (2003) The hydraulic characteristics and geochemistry of hyporheic and parafluvial zones in Arctic tundra streams, north slope. *Advances in Water Resources*, **26**, 907.

Fernald A.G., Wigington P.J. and Landers D.H. (2001) Transient storage and hyporheic flow along the Willamette River, Oregon: field measurements and model estimates. *Water Resources Research*, **37**, 1681.

Fox G.A. and Durnford D.S. (2003) Unsaturated hyporheic zone flow in stream/aquifer conjunctive systems. *Advances in Water Resources*, **26**, 989.

Fuller C.C. and Harvey J.W. (2000) Reactive uptake of trace metals in the hyporheic zone of a mining-contaminated stream, Pinal Creek, Arizona. *Environmental Science and Technology*, **34**, 1150.

Gooseff M.N., Bencala K.E., Scott D.T., Runkel R.L. and McKnight, D.M. (2005) Sensitivity analysis of conservative and reactive transient storage models applied to field data from multiple-reach experiments. *Advances in Water Resources*, **28**, 497.

Gooseff M.N., McKnight D.M., Lyons W.B. and Blum A.E. (2002) Weathering reactions and hyporheic exchange controls on stream water chemistry in a glacial meltwater stream in the McMurdo Dry Valleys. *Water Resources Research*, **38**, 1279, doi:10.1029/ 2001WR000834.

Gooseff M.N., Wondzell S.M., Haggerty R. and Anderson J. (2003) Comparing transient storage modeling and residence time distribution (RTD) analysis in geomorphically varied reaches in the Lookout Creek basin, Oregon, USA. *Advances in Water Resources*, **26**, 925.

Haggerty R., Wondzell S.M. and Johnson M.A. (2002) Power-law residence time distribution in the hyporheic zone of a 2nd-order mountain stream. *Geophysical Research Letters*, **29**, 1640, doi:10.1029/2002GL014743.

Harvey J.W. and Bencala K.E. (1993) The effect of streambed topography on surface-subsurface water exchange in mountain catchments. *Water Resources Research*, **29**, 89.

Harvey J.W., Conklin M.H. and Koelsch R.S. (2003) Predicting changes in hydrologic retention in an evolving semi-arid alluvial stream. *Advances in Water Resources*, **26**, 939.

Harvey J.W. and Fuller C.C. (1998) Effect of enhanced manganese oxidation in the hyporheic zone on basin-scale geochemical mass balance. *Water Resources Research*, **34**, 623.

Harvey J.W. and Wagner B.J. (2000) Quantifying hydrologic interactions between streams and their subsurface hyporheic zones. In *Streams and Ground Waters*, Jones J.B. and Mulholland P.J. (Eds.), Academic Press, p. 3.

Harvey J.W., Wagner B.J. and Bencala K.E. (1996) Evaluating the reliability of the stream tracer approach to characterize surface-subsurface water exchange. *Water Resources Research*, **32**, 2441.

Hinkle S.R., Duff J.H., Triska F.J., Laenen A., Gates E.B., Bencala K.E., Wentz D.A. and Silva S.R. (2001) Linking hyporheic flow and nitrogen cycling near the Willamette River - a large river in Oregon, USA. *Journal of Hydrology*, **244**, 157.

Jonsson K., Johansson H. and Wörman A. (2004) Sorption behavior and long-term retention of reactive solutes in the hyporheic zone of streams. *Journal of Environmental Engineering*, **130**, 573.

Kasahara T. and Wondzell S.M. (2003) Geomorphic controls on hyporheic exchange flow in mountain streams. *Water Resources Research*, **39**, 1005, doi:10.1029/ 2002WR001386.

Laenen A. and Bencala K.E. (2001) Transient storage assessments of dye-tracer injections in rivers of the Willamette Basin, Oregon. *Journal of the American Water Resources Association*, **37**, 367.

Lin Y. and Medina M.A. Jr (2003) Incorporating transient storage in conjunctive stream–aquifer modeling. *Advances in Water Resources*, **26**, 1001.

Malcolm I.A., Soulsby C., Youngson A.F., Hannah D.M., McLaren I.S. and Thorne A. (2004) Hydrological influences on hyporheic water quality: implications for salmon egg survival. *Hydrological Processes*, **18**, 1543.

Marion A., Bellinello M., Guymer I. and Packman A. (2002) Effect of bed form geometry on the penetration of nonreactive solutes into a streambed. *Water Resources Research*, **38**, 16, doi:10.1029/ 2001WR000264.

Mulholland P.J. and DeAngelis D.L. (2000) Surface-subsurface exchange and nutrient spiralling. In *Streams and Ground Waters*, Jones J.B. and Mulholland P.J. (Eds.), Academic Press, p. 149.

Nagorski S.A. and Moore J.N. (1999) Arsenic mobilization in the hyporheic zone of a contaminated stream. *Water Resources Research*, **35**, 3441.

Packman A.I. and Bencala K.E. (2000) Modeling surface-subsurface hydrological interactions. In *Streams and Ground Waters*, Jones J.B. and Mulholland P.J. (Eds.), Academic Press, p. 45.

Packman A.I. and MacKay J.S. (2003) Interplay of stream-subsurface exchange, clay particle deposition, and streambed evolution. *Water Resources Research*, **39**, 1097, doi:10.1029/ 2002WR001432.

Ren J. and Packman A.I. (2004) Stream-substream exchange of zinc in the presence of silica and kaolinite colloids. *Environmental Science and Technology*, **38**, 6571.

Runkel R.L. (1998) *One-Dimensional Transport with Inflow and Storage (OTIS): A Solute Transport Model for Streams and Rivers*, USGS WRIR 98–4018, **`http://co.water.usgs.gov/otis`**.

Runkel R.L. (2002) A new metric for determining the importance of transient storage. *Journal of the North American Benthological Society*, **21**, 529.

Runkel R.L., McKnight D.M. and Rajaram H. (2003) Modeling hyporheic zone processes. *Advances in Water Resources*, **26**, 901.

Salehin M., Packman A.I. and Paradis M. (2004) Hyporheic exchange with hetergenous streambeds: laboratory experiments and modeling. *Water Resources Research*, **40**, doi:10.1029/2003WR002567.

Salehin M., Packman A.I. and Wörman A. (2003) Comparison of transient storage in vegetated and unvegetated reaches of a small agricultural stream in Sweden: seasonal variation and anthropogenic manipulation. *Advances in Water Resources*, **26**, 951.

Schmid B.H. (2003) Temporal moments routing in streams and rivers with transient storage. *Advances in Water Resources*, **26**, 1021.

Schmid B.H. (2004) Simplification in longitudinal transport modeling: case of instantaneous slug releases. *Journal of Hydrologic Engineering*, **9**, 319.

Scott D.T., Gooseff M.N., Bencala K.E. and Runkel R.L. (2003) Automated calibration of a stream solute transport model: implications for interpretation of biogeochemical parameters. *Journal of the North American Benthological Society*, **22**, 492.

Sheibley R.W., Duff J.H., Jackman A.P. and Triska F.J. (2003a) Inorganic nitrogen transformations in the bed of the shingobee river, minnesota – integrating hydrologic and biological processes using sediment perfusion cores. *Limnology and Oceanography*, **48**, 1129.

Sheibley R.W., Jackman A.P., Duff J.H. and Triska F.J. (2003b) Numerical modeling of coupled nitrification-denitrification in sediment perfusion cores from the hyporheic zone of the shingobee river, MN. *Advances in Water Resources*, **26**, 987.

Stanley E.H. and Jones J.B. (2000) Surface-subsurface interactions: past, present, and future. In *Streams and Ground Waters*, Jones J.B. and Mulholland P.J. (Eds.), Academic Press, p. 405.

Storey R.G., Howard K.W.F. and Williams D.D. (2003) Factors controlling riffle-scale hyporheic exchange flows and their seasonal changes in a gaining stream: a three-dimensional groundwater flow model. *Water Resources Research*, **39**, 1034, doi:10.1029/ 2002WR001367.

Thomas S.A., Valett H.M., Webster J.R. and Mulholland P.J. (2003) A regression approach to estimating reactive solute uptake in advective and transient storage zones of stream ecosystems. *Advances in Water Resources*, **26**, 965.

Wagner B.J. and Harvey J.W. (1997) Experimental design for estimating parameters of rate-limited mass transfer: analysis of stream tracer studies. *Water Resources Research*, **33**, 1731.

Winter T.C., Harvey J.W., Franke O.L. and Alley W.M. (1998) *Ground Water and Surface Water – A Single Resource*, USGS Circular 1139, USGS: **http://water.usgs.gov/pubs/circ/circ1139**.

Wörman A. (2000) Comparison of models for transient storage of solutes in small streams. *Water Resources Research*, **36**, 455.

Zaramella M., Packman A.I. and Marion A. (2003) Application of the transient storage model to analyze advective hyporheic exchange with deep and shallow sediment beds. *Water Resources Research*, **39**, 1198, doi:10.1029/2002WR001344.

114: Snowmelt Runoff Generation

MING-KO WOO

School of Geography and Geology, McMaster University, Hamilton, ON, Canada

Snowmelt runoff has major influence on water supply and hazards for many temperate and polar regions. The magnitude and timing of runoff depend on the processes of snow accumulation and melt, which are controlled by factors that include latitude, topography, and land use. The amount of runoff production is affected by meltwater loss to evaporation and infiltration, the latter is often limited by frozen soil conditions. Runoff is delivered as flow in the snow, overland flow on bare ground, channeled flow in rills and interhummock cracks, or as subsurface flow in organic soils. The flow may be modified by the presence of snow or river ice in the channels, which often causes runoff delay and amplification of peak flows. Streamflow hydrographs for small basins typically show diurnal cycles with high discharge during warm days and declines in the cool and terminating periods of snowmelt. Hydrographs of large basins are an integration of the runoff patterns of their subbasins. In terms of snow management, the control of snowmelt floods is a consideration, yet many arid regions look to snowmelt runoff as a source of water supply.

INTRODUCTION

In winter, snow covers many parts of the temperate latitudes and almost everywhere in the polar regions. For these areas, meltwater runoff is an integral component, if not the principal element, of the hydrological cycle. Here, snowmelt runoff is considered as that part of the snow precipitation, which contributes to stream discharge (see Chow's, 1964 definition on runoff). Snow, as a portion of annual precipitation, increases poleward and hence the incremental importance of snowmelt in runoff generation relative to the contribution of rainfall. In North America, for example, snowfall accounts for under 30% of annual precipitation in temperate zones, rising to over 50% in boreal areas and exceeding three-quarters of total precipitation in the Arctic. Where above-freezing conditions occur frequently in winter, multiple melt runoff events are common, but a protracted snow accumulation season without intervening thaw episodes can eventually lead to late-season melt when the snow collected over months is released rapidly to runoff, often producing the annual flood. The availability of snow, the melt pattern, and the flow delivery mechanisms are major considerations in understanding and predicting snowmelt runoff generation.

SNOW ACCUMULATION AND MELT

The amount and timing of meltwater production depends on the processes of snow accumulation and snowmelt, both of which are controlled by climatic conditions. Of importance is the spatial variability of the snow cover and snowmelt rates; the former is due to uneven snowfall and snow drift and the latter is attributed to variations in energy available to melt the snow (*see* **Chapter 159, Snow Cover, Volume 4** and **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**). Several locational and land use factors affect the regional and microclimate and therefore play an indirect role in influencing melt runoff generation.

Latitude

Many temperate latitude locations experience intermittent snow melt during winter to cause runoff events. At higher latitudes, snowfall generally decreases, but there, winters are long, spanning 6 to 10 months, during which very little melt occurs. The snow can accumulate to considerable amounts before the melt season arrives.

Topography

Snowfall tends to increase with elevation, as does the length of the snow-covered period. In autumn and spring, it is also

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

common to see a snowfall line in mountainous areas, with simultaneous occurrence of snow accumulation above this line and rain-on-snow melt of any snow cover that exists at low elevations. Furthermore, snowmelt in mountainous terrain can span weeks or even months, with runoff generated first at the low altitudes and then proceeding to the high grounds. Snowmelt at high elevations is retarded by an altitudinal decrease in air temperature that reduces sensible heat flux and by frequent presence of clouds that cut down the radiation input. Late runoff from the highlands provides a summer source of water that is often important to crop irrigation on the plains, particularly in the arid and semiarid regions. On a local scale, snow accumulation and melt rate can vary between slopes. Shady slopes, such as those of northern exposure in the northern hemisphere, can have snowmelt delayed by weeks after the melt season has begun. Steep terrain that enhances radiation contrasts between the sunny and the shady aspects further exaggerates the differential melt rates within short distances, leading to asynchronous runoff generation within a basin.

Open Environment

Blowing and drifting of snow is common in the tundra, grasslands, and open fields, resulting in extensive redistribution of snow, with deep snow accumulated in sheltered areas such as gullies and lee slopes, but thin snow on exposed hill tops and windswept slopes (Woo and Marsh, 1978). Unevenness of snow cover quickly leads to its fragmentation when melt commences, and patchy snow distribution favors heat advection from bare ground to accelerate snowmelt (Liston, 1995; Neumann and Marsh, 1998). The area that contributes to snowmelt runoff changes rapidly in the melt season and large accumulation of snow in the valleys not only prolongs melt runoff production but can also temporarily block the flow in stream channels.

Vegetation

Vegetation intercepts precipitation and trees are particularly effective in intercepting snowfall, which can be driven by wind to reach the surface at oblique angles. Some of the snow intercepted by vegetation is sublimated without ever reaching the ground and this reduces the snow precipitation that is available to runoff generation. Interception loss can be substantial. Experiments that weighed entire trees in northern forests during winter demonstrated that, depending on tree type, about one-third of the total snowfall could be sublimated (Pomeroy *et al.*, 1998). Stand closure of a forest also plays a role as it limits the amount of snow that can reach the ground directly as throughfall. However, forest edges may trap the snow that drifts from adjacent open fields to create a narrow belt with deep snow. These considerations increase unevenness of snow distribution, hence the variations in the areas of melt runoff generation.

Snowmelt rates are modified by the type and size of trees, and the density and geometry of the forests, all of which have significant effects on radiation and turbulent heat exchanges that govern the melt energy balance. Although trees, when warmed, emit long-wave radiation to the snow, sometimes creating snow hollows (or tree wells) around the trunks, the trees block off direct beam solar radiation and sun flecks reach the snow only through gaps in the forests (Woo and Giesbrecht, 2000). These conditions render the snow in the forest to become more uneven as melting progresses. Snowmelt rate is often lower in the forest than in open areas because of the reduction in short-wave radiation by tree shadows; but whether the snow stays longer in the forest depends also on the amount of snowfall accumulated under the canopy. Small forest openings usually trap more snow than open fields. In areas where the forests are cut, parts of the logged patches may have residual snow that generates runoff after both the forest and the open fields have become bare.

Built-up Areas

Large variations in snow accumulation pattern are not only caused by local modifications of wind flow because of buildings but also by human removal and compaction of the snow. Urban snow contains many impurities, including salt and sand added by highway crews, and dirt particles from industrial fallout and automobile exhaust. The salts depress the melting point and the dust cover greatly lowers the snow albedo to enhance radiation melt. Snow melt is further complicated by multiple short-wave reflection and long-wave radiation exchanges between the buildings and the snow cover, and by large local variations in turbulent fluxes as well as heat advection from the buildings or bare ground (Semádeni-Davies and Bengtsson, 1998). Snowmelt in built-up zones tends to begin earlier, and the melt is more intense than in rural areas. Urban runoff is generated from a mixture of permeable and impervious areas and flow delivery is affected by artificial drainage that includes sewers, ditches, and pavements that enhance overland flow.

Hydrological models that compute snowmelt runoff generation apply widely different treatments of snow accumulation and melt that may or may not explicitly express the effects of the above factors. Lumped basin models may use melt computation methods that range from degree-days to full energy balance. Melt calculation may be adjusted for elevation and vegetation cover effects (e.g. SSARR model), and distributed models (e.g. SHE model) include topography and cover conditions, but none would have a component to cope with heat advection when the snow becomes patchy. A good summary of the various models with a snowmelt runoff generation component is available in Singh (1995).

FLOW PRODUCTION

From an increase in energy supply in the spring, through snowmelt events, to runoff generation, there are time lags that postpone the flow responses. Water produced by melting may be detained temporarily in the snow, especially if it is dry and has large water retention capacity, or if it is very cold and the meltwater refreezes. This delays meltwater release from the snow cover (*see* **Chapter 161, Water Flow Through Snow and Firn, Volume 4**). The released meltwater may enter the soil, or pond on ground surface before running off. The effects are best illustrated in the Arctic where intense winter coldness delays snowmelt until around the time of summer solstice (time of maximum incoming solar radiation). Figure 1 gives an example of the sequence of events leading to melt and runoff in McMaster River basin (area 33 km^2) near Resolute in arctic Canada (74°45′N; 95°05′W). Solar radiation reached high values around the solstice of 1978, but net radiation received by the snow remained low due to high snow albedo. Snowmelt began at about the time when air temperature rose above 0 °C. Runoff was released to hillslopes after several days of delay when the water-holding capacity of the snow was exceeded. It then took several days before the stream commenced to flow in earnest.

Infiltration is an important process to be considered in snowmelt runoff generation because should meltwater infiltrate the soil, it becomes unavailable to immediate runoff. Highly porous organic soils permit infiltration (Carey and Woo, 2001), while the abundance of soil cracks enhances meltwater entry into the soils. Even crystalline bedrock may be fractured and meltwater can infiltrate through the rock fissures (Spence and Woo, 2002). For most frozen soils, the presence of ice in the pores hinders meltwater infiltration.

Frozen ground conditions prevail during the snowmelt season in most areas other than the warm temperate zone where the ground may not freeze or may undergo alternating freeze–thaw during winter. Frozen soils with low ice content permit rapid infiltration (Kane and Stein, 1983) and this accounts for an absence of meltwater runoff from slopes in silty soils in the subarctic (Carey and Woo, 1999). However, frozen soils with most of their pore space sealed by ice tend to be impermeable to infiltration. Infiltration capacity depends on soil type, moisture content, and the thermal regime, and these factors have been considered in developing equations for frozen soil infiltration. According to Gray *et al.* (1985), infiltration in frozen prairie soils can be distinguished into unlimited (unimpeded infiltration), restricted (very little infiltration), and limited categories. Semiempirical equations that describe limited infiltration would include such independent variables as soil moisture content, soil temperature, and duration when water is available for infiltration (Zhao and Gray, 1999).

In spite of the significance of frozen soil in snowmelt runoff generation, it is rare to find a hydrological model that incorporates a frozen soil routine. In an experiment to demonstrate the role of frozen soil, Gray *et al.* (1985) added a frozen soil infiltration algorithm to the Sacramento model of the United State National Weather Service and obtained an improvement in the calculated runoff for a prairie catchment.

FLOW DELIVERY MECHANISMS

Meltwater released from the snow is subject to evaporative and deep seepage losses, but most of it usually runs off following several modes of delivery.

1. Meltwater can move downslope without first reaching the ground. When the snowpack becomes saturated at its base, it behaves as a porous medium through which meltwater flows laterally down gradient, with the input varying during the course of a day depending on the melt rate (Dunne *et al.*, 1976). Saturation of an entire snowpack will generate flow within and even on the snow surface. One special mechanism of delivery is slush flow, which is a rapid movement of snow–water mixture down slope or down valley (Gude and Scherer, 1998).
2. Unless the soil is impervious, meltwater will infiltrate. The infiltrated water may perch upon the frost that remains in the soil, or it may continue to drain downward to replenish the nonsaturated zone or to enter into groundwater storage. At the surface, water that is not infiltrated will fill the depression storage. In many places, the thinly thawed soil will be saturated and the abundance of water at or near the ground surface provides unlimited moisture supply for evaporation. In some regions, such as the polar deserts where spring melt is the main period with surface water, this can be the time of maximum evaporative loss.
3. After some initial delay to satisfy surface and subsurface storages, runoff is quickly delivered from hillslopes, in the form of overland flow. In impervious areas such as urban locations or frozen soil with concrete frost, overland flow is particularly prevalent. Even in more porous soils, such as of the temperate forests, rapid meltwater production can saturate the ground and generate overland flow on the forest floor (Figure 2a). In low-lying areas and streambank locations, flooding occurs when rapid meltwater production and inflow from high grounds cannot be shed quickly due to flat topography or channel blockage by snow and ice. Overbank flow and spring flooding of deltaic areas or wetlands become a recurrent phenomenon in the cold temperate, subpolar, and polar landscape.

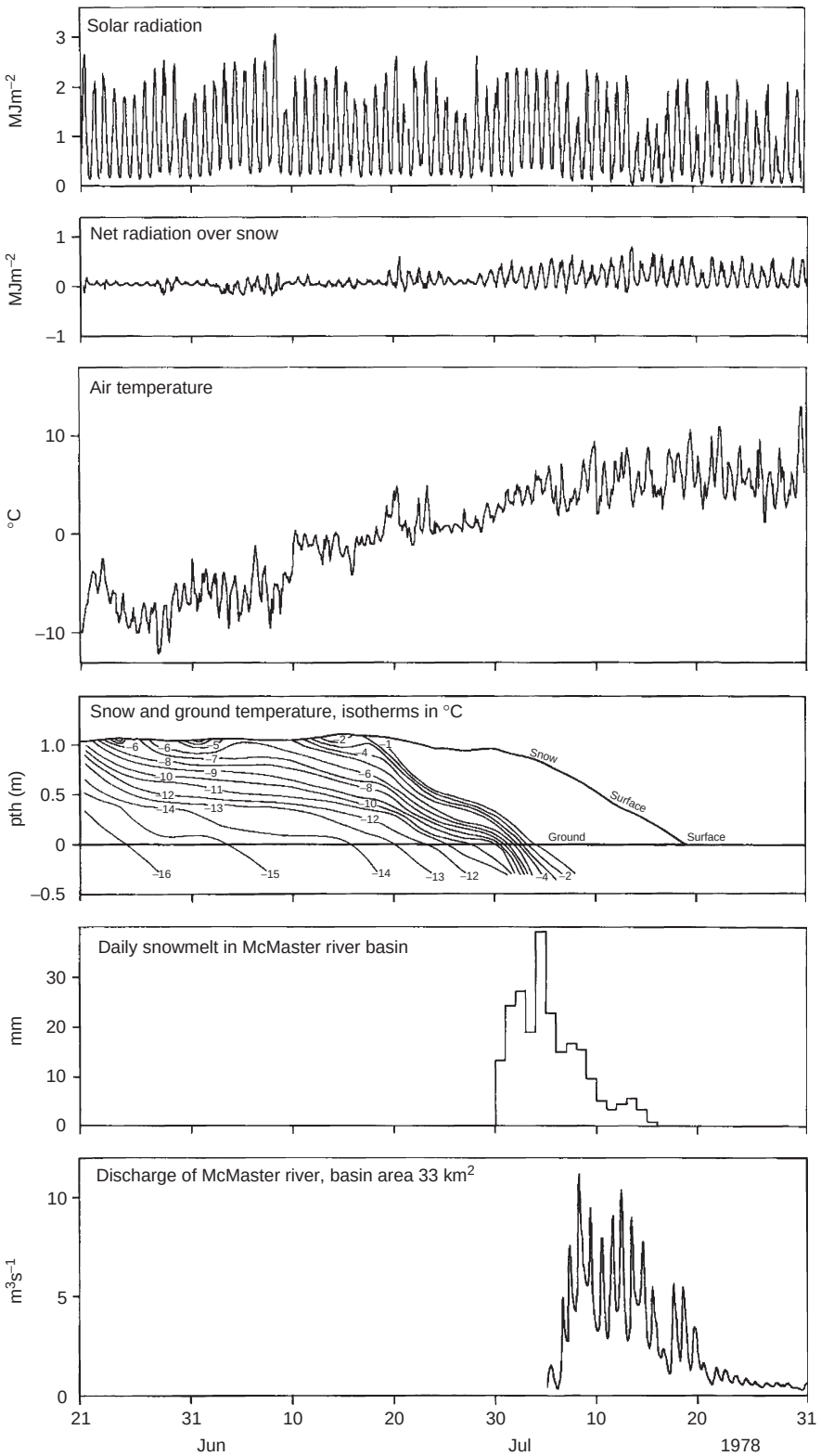


Figure 1 Progressive delay from an increase in melt energy supply to the initiation of streamflow, McMaster River basin, Cornwallis Island, Canada. While solar radiation attains peak values around the summer solstice, net radiation over snow remains low due to high albedo. Only after the air temperature has risen above 0 °C, the snow loses its cold content, followed by delayed runoff. Streamflow shows diurnal fluctuations that reflect the daily melt cycles. Depletion of the snow cover leads to seasonal decline in runoff

(a)

(b)

Figure 2 (a) Meltwater runoff producing overland flow on the floor of a temperate forest. (b) Concentration of meltwater runoff along interconnected shallow depressions between tussocks in an arctic landscape

4. Runoff may follow soil cracks and may concentrate along well-defined or poorly delineated channels on hillslopes, thus activating pathways that are normally dry except during snowmelt or intense rain storms. Preferred paths include cracks between tussocks and earth hummocks. Figure 2(b) shows the concentration of surface flow along shallow, but interconnected, depressions between tussocks. At a larger dimension, runoff may be channelled in rills, gullies, water tracks (which are inconspicuous trough lines on the slope, saturated only in the melt period), and zero-order headwater streams.
5. Water that infiltrates the soil, either in areas without concrete frost or in shallow thawed zone above the frost, moves as subsurface flow, which is often considerably lower in magnitude than overland and channelled flows (Dunne and Black, 1971; Woo and Steer, 1983). Many areas in the cold region have an organic layer composed of lichens and mosses, covering the mineral soils. The organic soil is often highly porous and permits infiltration even when frozen. Upon reaching the frozen mineral substrate, percolation ceases, and a perched saturated zone rises into the organic soil, issuing subsurface flows that include Darcian flow in the soil matrix, and quick flows through macropores that include soil pipes (Carey and Woo, 2001).

SOURCE AREA OF SNOWMELT RUNOFF

Contribution of snowmelt runoff to streamflow depends on (i) location and amount of meltwater production, which change with snow distribution and melt rates, and (ii) connectivity of meltwater to stream channels. Generally, surface runoff conveys water faster and carries more water downslope than subsurface flow, and the area that contributes to quick runoff varies within and between days.

At the beginning of the melt season, snow may cover most or all parts of the hillslopes or small catchments, but, as melt progresses, the snow coverage shrinks unevenly across the land. Not only do the areas that contribute to runoff become discontinuous but also, owing to differential melt rates that arise from topographical and land cover contrasts, the spatial pattern of flow generation becomes increasingly variable. The source area of runoff is dictated by snow distribution and by meltwater infiltration. Some soils with unlimited infiltration may not generate any surface runoff at all.

With frozen soil where most of the pore is sealed by ice (concrete frost), infiltration is limited and overland flow prevails. Where large quantity of meltwater is delivered from upslope, even permeable soils are saturated, and the saturated zone perched on the frozen substrate will rise above ground to generate overland flow. As meltwater production diminishes, such as during cool spells or when the snow cover retreats upslope, the perched water table subsides and only a small amount of subsurface flow is maintained. Where the soil is not frozen, as in some temperate latitude slopes, an abundance of meltwater can infiltrate and then saturate the soil to produce saturation overland flow. Regardless of location, source area varies during and between days and the resulting expansion and contraction of overland flow areas is analogous to the variable source concept that applies to rainfall events (Dunne and Black, 1971). Uneven snow distribution and overland flow patterns are best seen on the barren Arctic slopes, as is illustrated by Figure 3, which shows such a pattern during the melt season for an area in Ellesmere Island, Canada. As melt progresses, snow cover becomes

Figure 3 Uneven snow cover pattern in a barren arctic setting (Hot Weather Creek, Ellesmere Island, Canada) where only topography affects the microclimate to cause differential snow accumulation and melt, and the source area of meltwater runoff changes during the melt season

increasingly patchy and the source area for runoff shrinks. Ultimately, the snow disappears, but, in cold regions, snowbanks may linger well into the summer. Late-lying snow can be a local water source that extends the flow period of streams in high latitudes and high altitudes (Young and Lewkowicz, 1988).

On the scale of small to medium basins, elevation differences, soil and land use factors affect the timing and location of meltwater release. Towards the latter parts of the melt season, the principal source area of melt runoff often shifts from the open, low-lying zones to the forested or higher ground where residual snow remains.

Over a large area, it may not be practical or necessary to examine in detail the snow distribution and differential melt rates during the melt season, though it is useful to consider the changing fraction of snow cover in a basin throughout the melt period. One approach is to devise an empirical snow depletion curve for a basin by plotting the residual fraction of basin snow cover against time or cumulative degree-days or melt, or by plotting the snow-covered fraction against the ratio of basin snow-water equivalent average to the peak snow-water equivalent (Luce *et al.*, 1999). The fraction of snow-covered area in the basin estimated from the depletion curve can be used as a weighting factor to obtain basin snowmelt. One consideration in applying this method is that land use and melt pattern should not vary much from year to year, otherwise a new depletion curve has to be obtained for the basin.

FLOW IN STREAM CHANNELS

Meltwater runoff follows the melt rhythm of snow and therefore responds to energy input that generates melt and to the storage that retards flow delivery. Streamflow may be derived from individual snowfall events, as is common in warm temperate areas where cold spells are quickly followed by melt episodes. Similarly, spring and autumn snowfall in cold temperate regions or summer snowfall in high altitudes and polar latitudes is subject to rapid melting to generate streamflow responses with little delay. These fast responses may be termed snowfall runoff (Yang *et al.*, 1993) in contrast to the snowmelt runoff produced from a snow cover that accumulates over multiple snowfall events over a period of days to months.

With deep snow, the melt season can extend for days or weeks and streamflow usually lags melt events. The initial melt events are subdued. When runoff reaches the valleys, it may encounter several situations in the channels.

1. The stream channels have little snow or ice, so that water can move down the channel network in direct response to the rhythm of snow meltwater input from basin slopes.
2. Snow infills most channels in the Arctic (Woo and Sauriol, 1981) and even in the windswept plains of the temperate region. Snow in the valleys hinders the development of an integrated flow system when the melt season arrives. Streamflow may begin at isolated segments along the valleys, depending on local supply of meltwater and flow blockage by snow jams. A typical example is provided in Figure 4(a) where snow infills channels that are dry in winter (no winter flow). The uneven valley snow blockage temporarily impounds the meltwater that runs off from adjacent slopes, but prevents the linkage of individual ponded segments to form a continuous drainage network. When these snow jams are breached, water retained upstream of the snow dams is released rapidly to generate floods downstream. Such episodes of ponding and snow dam breakage will eventually establish an integrated channel network for flow delivery. For these channels, the rise of streamflow includes an initial low flow stage indicative of the preintegrated phase of flow delivery. Once established, the drainage network along the mostly snow-lined channels will be very effective in carrying meltwater downstream.
3. An ice cover is often formed over the winter in the cold temperate and subarctic rivers, and the ice breaks up in the spring thermally through melting, or mechanically through fracturing. Snowmelt runoff flowing down the channel will cause the river ice to fragment and move downstream, frequently forming ice jams that block the flow. Ice jams cause the upstream water levels to rise and failure of the jams creates high flows to down-stream areas (Beltaos, 2000). Ice breakup lags behind snowmelt, but amplifies peak flows so that flooding due to ice jamming is a common occurrence (Figure 4b). At the time of high flow, ice floes drift downstream until they are jammed, with speeds that often exceed

(a)

(b)

Figure 4 (a) Stream channels in Arctic Canada (Iqaluit, Baffin Island) jammed by snow, causing streamflow to occur in segments until the snow jams are breached to permit the integration of the drainage network. Note the many bare patches on high grounds because melting has proceeded for days to deplete the snow on the ridges and on most slopes. (b) A river-side village in subarctic Canada (Fort Liard) threatened by ice jam flood. Note that many ice floes are much larger than the houses, and these floes move down channel at speeds that often exceed 5 m s^{-1}

5 m/s. The thickness of river ice increases poleward and large ice floes in cold temperate and subarctic regions have tremendous destructive power. Hydraulic models have been developed to forecast peak discharges and their associated flood stages (e.g.Blackburn and Hicks, 2002).

SNOWMELT HYDROGRAPHS

The effect of snowmelt on streamflow increases for basins located in high altitudes and high latitudes. For these areas, the highest flows of the year may be realized in a brief melt period that lasts for about two to three weeks. Where snowmelt exerts a preponderant influence on stream discharge, the term *nival* (or *snow-dominated*) *regime* has been applied to describe the seasonal rhythm of streamflow (Church, 1974). In some warm temperate areas, such as the west coast of British Columbia and the Pacific Northwest of the United States, both snowmelt and winter rainfall are important generators of runoff, and high flows are produced by mixed processes of rainfall and snowfall (Waylen and Woo, 1982). The Coquihalla River (basin area 740 km^2, with an elevation range of about 2500 m), for example, receives much winter precipitation that comes as rainfall at the low elevations and as snowfall at the higher parts of the basin. The rain produces high flows in the fall and the winter and when the snow melts at high elevations, spring melt runoff is generated (Figure 5a).

The hydrograph for small to medium basins that have a nival regime of streamflow typically shows prominent diurnal cycles that reflect runoff contribution from snowmelt (Figure 1). Owing to time lag in runoff delivery to the streams, daily peak discharges are usually attained in the late afternoon or evening and low flows occur in early mornings. A period of intense melt raises the daily peaks, while a cool spell dampens melt and diminishes the amplitudes of the diurnal flow cycles. Depletion of the snow cover is reflected in a gradual reduction in discharge, but a steeper recession for the daily low flow, the latter suggesting a decline in meltwater storage available to feed the baseflow.

A large basin combines the flow regimes of its subbasins. The storage effect of the large basin dampens the short-term fluctuation in discharge, such as the diurnal cycles in the hydrograph. On the other hand, subbasins may have different magnitude and timing of snowmelt due to location, altitude, and land use. The integrated effect is manifested in multiple snowmelt hydrograph rises and falls. Figure 5(b) is an example from the Liard basin in northwestern Canada (area 2 75 000 km^2), showing a downstream increase in snowmelt hydrograph rises and falls as more subbasins with different timing and amounts of flows enter the main trunk valley.

SNOW MANAGEMENT FOR RUNOFF AUGMENTATION

The control of snowmelt flood hazards plays a major role in snow and snowmelt management. This includes the building of dykes to protect structures, increasing the size of culverts and storm drainage system to accommodate large flows, good forecasting of flood arrival, and providing flood warning. It may not be economically feasible to implement flood protection works, and river-side residents may have to adapt to the flood events and clean up their properties

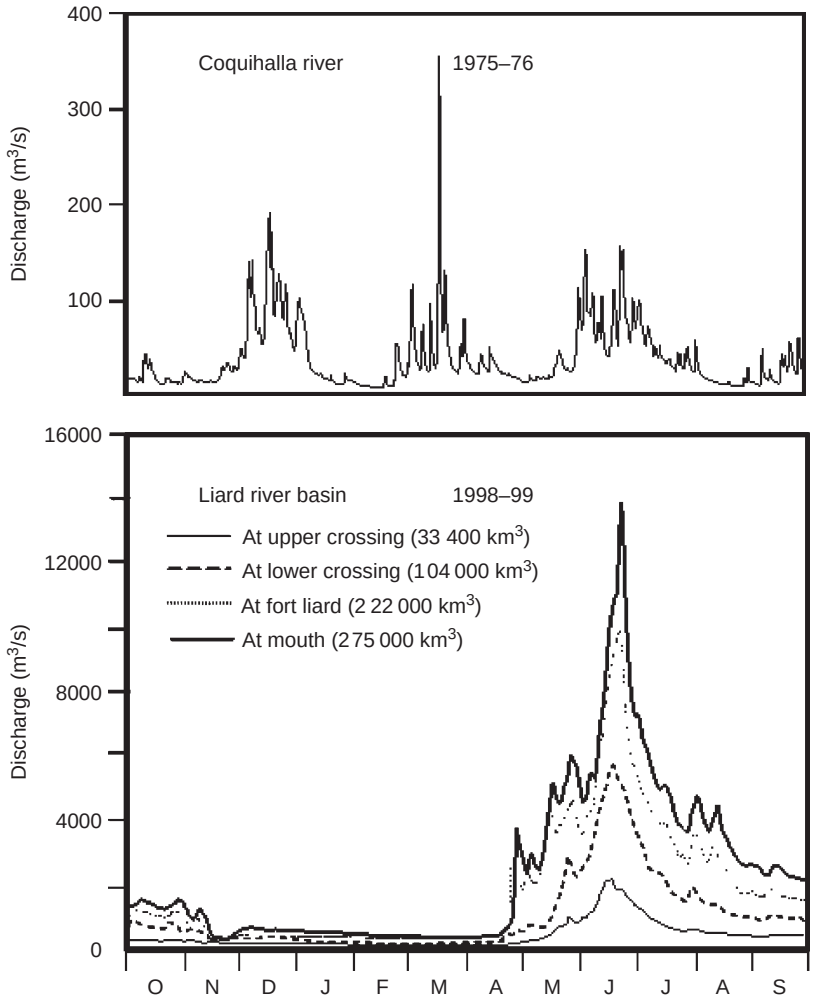


Figure 5 (a) Hydrograph of Coquihalla River in western Canada, a river in the warm temperate region with high flows produced by a mixture of winter rainfall and spring snowmelt events. (b) Hydrographs of the Liard Basin in Canada (a subarctic and mountainous catchment traversed by the 60°N parallel), showing downstream increase in multiple melt-related peaks due to meltwater runoff from its subbasins

after spring melt (e.g. Figure 4b shows the village of Fort Liard in northern Canada encountering ice jam flooding that recurs every spring).

On the other hand, an increase in snowmelt runoff and possibly an extension of the melt period are important to the water supply of many parts of the world. Several examples of snow management practices for runoff augmentation are noted below.

Many parts of Central Asia adopt the ancient practice of using the karez system. The karez or qanat, a water conveyance system developed by the Persians in about sixth century B.C., is a hand-dug subterranean (to limit evaporation loss en route) conduit with a gentle gradient that brings meltwater from the mountains to the oases in arid lowlands, for irrigation and community water supply. Construction of the karez involves the excavation of vertical shafts that are extended laterally to form a shallow tunnel under the desert. The melting of late-lying snow in the mountains provides runoff, which is captured by the karez. This is an example of taking advantage of the natural melt rhythm for water supply.

Agricultural fields in the Canadian prairies can benefit from leaving the crop stubbles over winter to trap snow (Stepphun, 1981). Snow is trapped by the stubbled fields and spring meltwater runoff will increase soil moisture in the semiarid environment. This is a case of encouraging snow accumulation to augment runoff for agriculture. A third example is land-use intervention through logging to modify the snow regime. Experiments have been conducted to examine the logging patterns that alter snow accumulation, melt, and runoff in temperate forests (Anderson and Gleason, 1960; Ffolliott *et al.*, 1989). However, economical and environmental considerations have limited such logging practices and prevented large-scale applications.

CONCLUSIONS

Important to the subject of snowmelt runoff generation are the considerations of timing, quantity, and the location where meltwater is produced. In terms of the magnitude and the source area of runoff, it is necessary to determine the amount of snow accumulation that is subject to sublimation losses, the spatial pattern of snow distribution that is often modified by snow drifting and that changes as melting progresses. The timing and the intensity of meltwater release depend on the melt rate, which, in turn, is energy driven.

Connectivity between meltwater sources and the drainage network dictates whether runoff can be conveyed out of a slope or a basin. Meltwater may be lost to infiltration to become unavailable for immediate runoff, and evaporation reduces the meltwater that would otherwise support flow. The rate of meltwater delivery depends on the flow mechanisms such that, for instance, overland flow is orders or magnitude faster than subsurface flow. Channel blockage may temporarily retard the flow by withholding the water for short periods, but breaching of the blockages would exaggerate the discharge, leading to flood events.

By adapting to or manipulating snow accumulation and melt, snowmelt runoff production can be modified to the advantage of human activities.

REFERENCES

Anderson H.W. and Gleason C.H. (1960) Effects of logging and brush removal on snow water runoff. *International Association of Scientific Hydrology Publication*, **51**, 478–489.

Beltaos S. (2000) Advances in river ice hydrology. *Hydrological Processes*, **14**, 1613–1625.

Blackburn J. and Hicks F. (2002) Combined flood routing and flood level forecasting. *Canadian Journal of Civil Engineering*, **29**, 64–75.

Carey S.K. and Woo M.K. (1999) Hydrology of two slopes in subarctic Yukon, Canada. *Hydrological Processes*, **13**, 2549–2562.

Carey S.K. and Woo M.K. (2001) Slope runoff processes and flow generation in a subarctic, subalpine catchment. *Journal of Hydrology*, **253**, 110–129.

Chow V.T. (1964) *Handbook of Applied Hydrology*, Section 14, McGraw-Hill: New York.

Church M.A. (1974) Hydrology and permafrost with reference to northern North America. *Proceedings, Workshop-Seminar on Permafrost Hydrology*, Canadian National Committee, IHD: Ottawa, pp. 7–20.

Dunne T. and Black R.D. (1971) Runoff processes during snowmelt. *Water Resources Research*, **7**, 1160–1172.

Dunne T., Price A.G. and Colbeck S.C. (1976) The generation of runoff from subarctic snowpacks. *Water Resources Research*, **12**, 677–685.

Ffolliott P.F., Gottfried G.J. and Baker M.B. Jr (1989) Water yield from forest snowpack management: research findings in Arizona and New Mexico. *Water Resources Research*, **25**, 1999–2007.

Gray D.M., Landine P.G. and Granger R.J. (1985) Simulating infiltration into frozen prairie soils in streamflow models. *Canadian Journal of Earth Sciences*, **22**, 464–472.

Gude M. and Scherer D. (1998) Snowmelt and slushflows: hydrological and hazard implications. *Annals of Glaciology*, **26**, 381–384.

Kane D.L. and Stein J. (1983) Water movement into seasonally frozen soils. *Water Resources Research*, **19**, 1547–1557.

Liston G.E. (1995) Local advection of momentum, heat, and moisture during the melt of patchy snow covers. *Journal of Applied Meteorology*, **34**, 1705–1715.

Luce C.H., Tarboton D.G. and Cooley K.R. (1999) Sub-grid parameterization of snow distribution for an energy and mass balance snow cover model. *Hydrological Processes*, **13**, 1921–1933.

Neumann N. and Marsh P. (1998) Local advection of sensible heat in the snowmelt landscape of arctic tundra. *Hydrological Processes*, **12**, 1547–1560.

Pomeroy J.W., Parviainen J., Hedstrom N. and Gray D.M. (1998) Coupled modelling of forest snow interception and sublimation. *Hydrological Processes*, **12**, 2317–2337.

Semádeni-Davies A. and Bengtsson L. (1998) Snowmelt sensitivity to radiation in the urban environment. *Hydrological Sciences Journal*, **43**, 67–89.

Singh V.P. (1995) *Computer Models of Watershed Hydrology*, Water Resources Publications: Highlands Range, pp. 1130.

Spence C. and Woo M.K. (2002) Hydrology of subarctic Canadian shield: bedrock upland. *Journal of Hydrology*, **262**, 111–127.

Stepphun H. (1981) Snow and agriculture. In *Handbook of Snow*, Gray D.M. and Male D.H. (Eds.), Pergamon Press: Toronto, pp. 60–125.

Waylen P. and Woo M.K. (1982) Prediction of annual floods generated by mixed processes. *Water Resources Research*, **18**, 1283–1286.

Woo M.K. and Giesbrecht M. (2000) Simulation of snowmelt in a subarctic spruce woodland: 1. tree model. *Water Resources Research*, **36**, 2275–2285.

Woo M.K. and Marsh P. (1978) Analysis of error in the determination of snow storage for small high arctic basins. *Journal of Applied Meteorology*, **17**, 1537–1541.

Woo M.K. and Sauriol J. (1981) Effects of snow jams on fluvial activities in the high arctic. *Physical Geography*, **2**, 83–98.

Woo M.K. and Steer P. (1983) Slope hydrology as influenced by thawing of the active layer, Resolutem N.W.T. *Canadian Journal of Earth Sciences*, **20**, 978–986.

Yang D.Q., Woo M.K., Liu F.J. and Yang Z.N. (1993) The role of snow in streamflow generation from an alpine permafrost basin in Tianshan, northwestern China. *Proceedings 6th International Conference on Permafrost*, Beijing, pp. 738–743.

Young K.L. and Lewkowicz A.G. (1988) Measurement of outflow from a snowbank with basal ice. *Journal of Glaciology*, **34**, 358–362.

Zhao L. and Gray D.M. (1999) Estimating snowmelt infiltration into frozen soils. *Hydrological Processes*, **13**, 1827–1842.

115: Landscape Element Contributions to Storm Runoff

JAN SEIBERT[1] AND BRIAN MCGLYNN[2]

[1]Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden
[2]Department of Land Resources and Environmental Sciences, Watershed Hydrology Laboratory, Montana State University, Bozeman, MT, US

Hydrological conditions and the corresponding land area contributions to streamflow vary across the landscape. Landscapes are mosaics that can be described in varying levels of detail. This document describes the importance of different landscape elements for storm runoff generation. Since topography is often a first-order control on hydrological processes, topographic indices can be used to describe the spatial variations of hydrological conditions. Another approach is the delineation of landscape elements. We suggest that riparian and hillslope areas are the two most basic landscape elements that need to be considered; however, additional landscape elements including urban areas, floodplains, croplands, wetlands, and so on, can also be important. Landscape elements can be defined through hydrological analysis. This analysis can further define their lateral connections and organization along the stream network, providing valuable interpretation of the dominant controls on hydrological response across the landscape. Additionally, approaches to represent landscape elements in hydrological modeling are reviewed.

INTRODUCTION

Dominant hydrological processes and runoff source areas are spatially variable across the landscape. While it is generally agreed that various landscape elements contribute differently to catchment runoff, quantification of runoff source areas and dominant processes among different landscape elements is poorly understood. We refer to landscape elements as different spatial elements, which have certain hydrologic dynamics. Catchments consist typically of a mosaic of landscape elements. This mosaic can be described in varying levels of detail depending on data and application. One can, for instance, divide a catchment into two landscape elements such as riparian and hillslope zones or into a large number of elements defined by an overlay of spatial information such as topography, soils, land use, and vegetation. Hillslopes might be further delineated according to slope or position relative to the ridge or valley. Other examples of landscape elements might be bedrock outcrops, wetlands, glaciers, or areas largely affected by human activities such as urban areas and agricultural land. The latter might further be separated into pasture, crop, and fallow as well as irrigated (or drained) land. Different dominant hydrological processes follow from the characteristics of various landscape elements. These dominant processes are those needed to define the hydrological functioning of individual landscape elements. Here, we focus on hydrologic function with regard to runoff events, that is, the runoff generation processes for storm runoff.

Capturing the spatial variability of hydrological processes across the landscape is difficult and many modeling approaches have been shown to be only partly successful when tested against spatially distributed field observations. However, while not fully successful, these approaches at least consider the mosaic nature of catchment hydrology. The consideration of different landscape elements, their spatial distribution, and connectivity can provide a more realistic representation of catchment hydrology than a lumped description.

Fully distributed models that typically divide the catchment into small grid cells, have the potential to represent spatial variability. Inadequate input data and computational

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

burden, however, significantly limit this potential in most applications. Lumped models are therefore still widely used and often found to be able to simulate catchment runoff with acceptable accuracy. These models, however, use effective parameters, which are hard to measure and difficult to interpret. Assume a catchment with 10% paved areas (and high infiltration capacities in the remaining catchment). A simple lumped infiltration-capacity based model might be able to simulate runoff in this case, but the required infiltration-capacity parameter value will be an effective value somewhere between the extremes (zero and high infiltration) and the simulations will be some average behavior, but will not reflect reality and the associated variability in runoff processes. The importance of considering spatial variable parameter values instead of a mean (or effective) value has been illustrated by, for instance, Woolhiser *et al.* (1996) and Merz and Plate (1997).

Topography is often a first-order control on the spatial variations of hydrological conditions. It affects the spatial distribution of soil moisture and often groundwater flow. Variables other than topography, such as soils and vegetation, might also largely control hydrological processes. However, topography is often easiest to obtain as spatially distributed data. Therefore, much of the work presented in this article utilizes topographic data although we do acknowledge the importance of other types of data.

Catchment topography is not only important for runoff quantity but also for runoff quality. Beaujouan *et al.* (2002), for instance, applied a model to theoretical catchments and demonstrated that catchment-wide denitrification depended on the topographic shape of the catchment. Despite these findings, the topographic shape of a catchment is not typically represented in lumped models.

In this article, we discuss both the importance of different landscape elements for storm runoff and review approaches to represent landscape elements in hydrological modeling.

SPATIAL AND TEMPORAL SOURCES OF STORM RUNOFF – THE ROLE OF LANDSCAPE ELEMENTS

The relative contributions of different landscape elements to catchment runoff depend to a large degree on prevailing runoff generation processes. In other words, the answer to the question "where does runoff originate?" depends partially on understanding dominant runoff generation processes across the landscape. For example, if **infiltration-excess overland flow** (Horton, 1933, **Chapter 111, Rainfall Excess Overland Flow, Volume 3**) is assumed, areas with the lowest infiltration capacities are the dominant contributors to catchment runoff. Topography in this case is of less importance for the generation of runoff (but still important for the routing of runoff). If **saturation-excess overland flow** (Hewlett and Hibbert, 1967; Dunne and Black, 1970) is assumed to be the dominant runoff process, soil depth and topography are controlling factors on those locations where most catchment runoff is generated. Saturation-excess overland flow is likely on those portions of catchments with water tables near the ground surface – most often, riparian areas and wetlands. **Subsurface flow** (*see* **Chapter 112, Subsurface Stormflow, Volume 3**) is ubiquitous in most catchments, although it is believed that subsurface flow as a storm runoff mechanism is most important in steeply sloped environments with shallow soil. Recent research (McDonnell *et al.*, 1996; Freer *et al.*, 2002) has suggested that both surface and bedrock topography control the magnitudes and spatial variability of subsurface runoff across hillslopes. **Transmissivity feedback** runoff generation (Bishop, 1991; Kendall *et al.*, 1999) refers to increasing magnitude of lateral flow as the water table rises into more transmissive shallow soil. As such, transmissivity feedback depends on water table fluctuation during storm events and therefore is a function of soil depth, hydraulic conductivity profiles with depth, and both surface and subsurface topography.

A realistic representation of the unique landscape elements that incorporates their different contributions to catchment runoff is important for the quantification of spatial and temporal sources of storm runoff. This information can provide insight into the controls on the chemical composition of runoff water since flow pathways and residence times vary for different landscape elements. The identification of different landscape elements might also help researchers design observation networks to best represent the landscape variability (Park and van de Giesen, 2004). The degree of detail with which landscape elements are defined depends on the available information. On the basis of detailed field studies, Naef *et al.* (2002) and Uhlenbrook *et al.* (2004) were both able to distinguish several landscape elements with different runoff generation properties. Such detailed information is usually not available and only more fundamental differences are represented by landscape elements. The most basic catchment discretization separates catchments into two major landscape elements: riparian zones (RZ) and hillslope zones. These zones are often distinct in their dominant runoff processes, water table dynamics, topography, soils, vegetation characteristics, and biogeochemical environments (McGlynn and Seibert, 2003a).

The RZ can be defined as the strip of land between the stream channel and the hillslope. The RZ is sometimes also called *buffer zone* (Lowrance *et al.*, 1985), floodplain (Bates *et al.*, 2000), or near-stream zone (Cirmo and McDonnell, 1997). The RZ differs from upslope zones with regard to hydrology, vegetation, and soils (Swanson *et al.*, 1982; Hill, 1996; Naiman and Décamps, 1997; Butterworth *et al.*, 2000). Distinguishing characteristics are anoxic zones (Pritchett and Fisher, 1987; Megonigal *et al.*, 1993), gleyed

soils (Faulkner and Patrick, 1992; Phillips *et al.*, 2001), color (Blavet *et al.*, 2000), organic content (Mitsch and Gosselink, 1993), breaks in slope (Merot *et al.*, 1995), and near-surface water tables (Brinson, 1993). Because of their location, riparian zones have significant potential to regulate the movement of material in surface and subsurface runoff that flows from upslope areas to the stream (Brinson *et al.*, 1981; Hill, 1996).

McGlynn and McDonnell (2003a,b) suggested that the relative timing of riparian and hillslope source contributions to streams, and the connections and disconnections of dominant runoff contributing areas or landscape elements, were the first-order controls on stream dissolved organic carbon (DOC) fluxes and stream silica from a steeply forested basin at Maimai in New Zealand. The RZ has also been shown to be of high importance for ecological issues such as biodiversity and biogeochemical cycles (Naiman and Décamps, 1997).

The origin of runoff can be described by defining the temporal and spatial sources. A useful temporal distinction is new and old water (also referred to as *event* and *pre-event water*). New water is water entering the system as rainfall or snowmelt in association with the current event, whereas old water is water that resided in the catchment prior to the event. It has been widely demonstrated across the world that old water dominates storm runoff hydrographs in most catchments (Genereux and Hooper, 1998; Buttle, this volume, **Chapter 116, Isotope Hydrograph Separation of Runoff Sources, Volume 3**). Quantifying new and old water fractions in stormflow is important for understanding controls on stormflow geochemistry, interpreting the spatial sources of storm runoff, and understanding dominant runoff processes at the catchment scale. Runoff can also be described by its spatial source. For instance, one can distinguish between hillslope and riparian water. Hillslope water is water originating from upland hillslope zones, whereas riparian water is water originating from riparian zones. Both spatial sources can contribute with both event and pre-event water.

The analysis of landscape element contributions to storm runoff integrates measurements or estimations of spatial and temporal runoff sources, assessment of dominant runoff processes, the spatial extent of landscape elements, and the spatial organization or topology of landscape elements at the catchment scale.

CASE STUDY – CATCHMENT INTERCOMPARISON

The Maimai (New Zealand), Panola Mt. (Georgia, US), and Sleepers River (Vermont, US) research catchments are long-term experimental sites where researchers have sought to elucidate the fundamental controls on runoff sources, flowpaths, and streamwater residence times. The rich history of research at each site provides context for analysis of landscape element contributions to storm runoff. Here we review results of discretization of each catchment into its fundamental landscape elements (hillslopes and riparian zones) and assess the relative contributions of each landscape element to catchment runoff utilizing three-component tracer mass-balance hydrograph separation techniques (McGlynn, in press). Hydrograph separations were based on high-resolution temporal sampling of source water and streamwater ^{18}O, silica, and calcium as tracers (McGlynn, in press; McGlynn and McDonnell, 2003a).

The first step in this analysis was to discretize each catchment into its component landscape elements (McGlynn, in press). In each catchment, riparian zones and hillslopes are the dominant landscape elements. We utilized a combination of detailed field surveys and terrain analysis of high-resolution digital elevation models (DEMs) to delineate riparian zones and hillslope zones (McGlynn, in press; McGlynn and Seibert, 2003a). We found that the riparian zones accounted for 4.5% of the total catchment area (17 ha) at Maimai, 6.4% of total catchment area (40 ha) at Sleepers River, and 10% of total catchment area (40 ha) at Panola Mt. We then separated comparable stormflow hydrographs at each catchment to determine the contributions of each landscape element (riparian zones and hillslope zones) and the proportion of old and new water in each stormflow hydrograph (Figure 1).

We found that old water dominated the storm hydrographs at each site and that the proportion of riparian source water in each storm hydrograph was proportional to the riparian extent in each catchment (Figures 1 and 2). The discretization of each catchment into its component landscape elements, in this case riparian and hillslope zones, and the separation of storm runoff into its spatial sources (landscape elements) and temporal sources (old and new water) provided new insight into the linkages between landscape structure and runoff generation at the catchment scale. We suggest that the combination of landscape discretization and analysis of the contributions of each landscape element to catchment runoff provides a way forward for catchment intercomparison and a potential method for quantifying the relative roles of various landscape elements from the headwaters to the meso-scale (McGlynn *et al.*, 2004).

TOPOGRAPHIC INDICES

Topography has become widely used as a surrogate for the spatial variation of hydrological conditions. Topography has a major impact on the hydrological processes in a catchment and spatially distributed topographic data is readily available (Moore *et al.*, 1991). A major reason for the increased use of topography in hydrological analysis is the wide availability of DEMs, from which various topographic indices can be computed. Usually, gridded elevation data is used for computations, although other methods such as triangulated

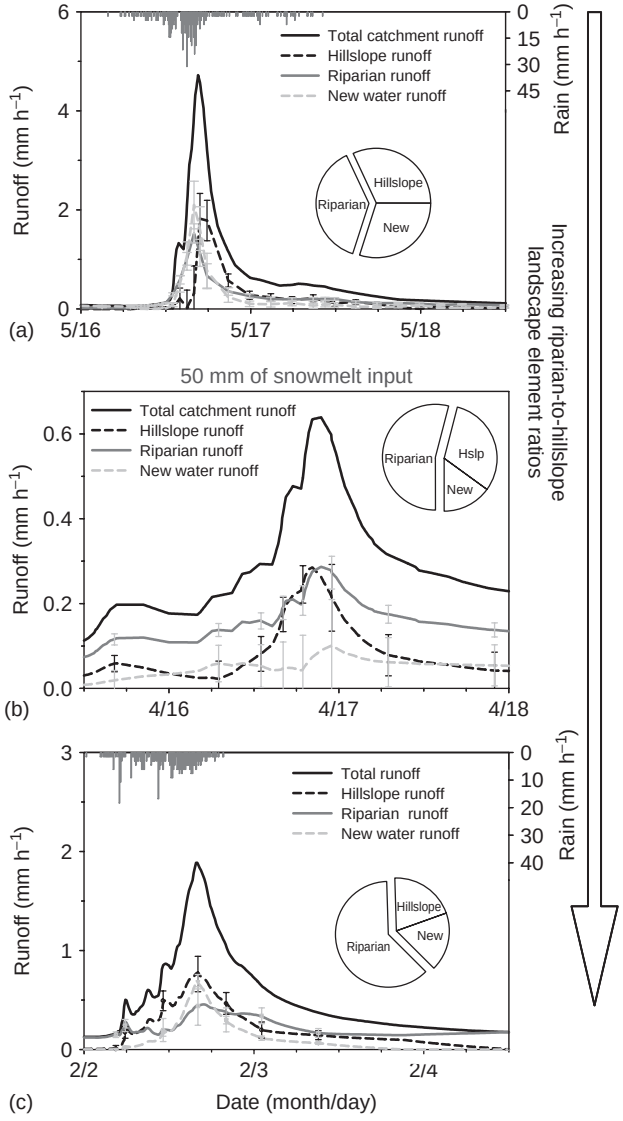


Figure 1 The contributions of riparian and hillslope landscape elements to storm flow in three catchments with increasing riparian-to-hillslope area ratios. (a) Maimai: 65 mm rain event and a runoff ratio of 0.5. (b) Sleepers River: 50 mm snowmelt event and a runoff ratio of 0.5. (c) Panola Mountain: 62 mm rain event and a runoff ratio of 0.54. Adapted from McGlynn (in press)

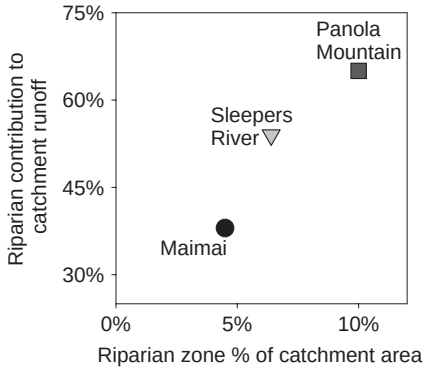


Figure 2 Percent of catchment runoff originating in the riparian landscape element versus riparian landscape element percent of total catchment area. Adapted from McGlynn (in press)

irregular networks (TINs)might be more efficient (Tucker *et al.*, 2001).

Topographic indices allow evaluation and the use of important topographic information. Topographic indices have been used to describe spatial soil moisture patterns (e.g. Burt and Butcher, 1985) and other hydrological variables (*see* Moore *et al.*, 1991 for a review). By topographic index, we refer to spatially distributed values, which have been calculated from a DEM and capture some feature of the topography. These indices can be grouped into locally determined indices such as elevation, slope, aspect, and curvature; indices that also consider the elevations at more distant points, such as upslope accumulated area, distance to stream, or elevation above the stream; and combinations of different indices such as the topographic wetness index (TWI) (as discussed below).

The TWI developed by Beven and Kirkby (1979) within the runoff model TOPMODEL is one of the most widely applied indices. This index is defined as $\ln(a/\tan\beta)$ where a is the local upslope area draining through a certain point per unit contour length and $\tan\beta$ is the local slope. The first is assumed to be a measure of water flowing towards a certain location and the latter a measure of water draining from a certain location. A downslope index has recently been suggested by Hjerdt *et al.* (2004) as an alternative to the local slope. This index is calculated as d/L_d where L_d is the distance to the nearest cell having a height $\geq d$ length units (e.g. 5 m) below the cell. By taking downslope topography into account, the slope of the groundwater table and, thus, the drainage from a certain location might be better estimated by this downslope index than by the local gradient.

The TWI has been used to study spatial scale effects on hydrological processes (Beven *et al.*, 1988; Famiglietti and Wood, 1991; Moore and Grayson, 1991; Sivapalan and Wood, 1987; Sivapalan *et al.*, 1990), prediction of the spatial pattern of saturated areas (Güntner *et al.*, 2004), and identification of hydrological flow paths for geochemical modeling (Robson *et al.*, 1992), as well as biological processes such as annual net primary production (White and Running, 1994), vegetation patterns (Moore *et al.*, 1993), biodiversity (Zinko *et al.*, in press), and forest site quality (Holmgren, 1994).

Topographic indices have also been used to predict the spatial variation of soil characteristics (McKenzie *et al.*, 2000). Pachepsky *et al.* (2001) found soil water retention to be related to topography. Studying a hillslope transect in a boreal landscape, Giesler *et al.* (1998) found

pH and nitrogen to increase with the increasing upslope accumulated area.

Obviously, factors other than topography influence local groundwater levels and soil moisture status. Correlations between TWI and groundwater levels or soil moisture values are therefore usually not very strong (e.g. Seibert *et al.*, 1997; Western *et al.*, 1999). Topographic indices may be improved by including additional information. Pellenq *et al.* (2003), for instance, found soil depth to improve topographically based soil moisture predictions. Güntner *et al.* (2004), however, found only slightly improved predictions of the spatial patterns of saturated areas when they combined different additional information with the topographic wetness index.

HYDROLOGICAL LANDSCAPE ANALYSIS

Catchments can be discretized into their component pieces or dominant landscape elements (McGlynn and Seibert, 2003a). These elements typically consist of, but are not limited to, hillslopes, riparian zones, extended wetlands, and stream reaches. We refer to hydrological landscape analysis (HLA) as the examination of topographic and topologic attributes of DEMs and indices derived from topography and other spatial data. Topology refers to the relative arrangement or configuration of landscape elements or the functional arrangement, that is, hydrologic connections and disconnections.

An important issue is the hydrological functioning of the RZ, with its often distinctive soil properties. The RZ is of special importance for water quality since water from upslope positions passes through the RZ before entering the stream network and, thus, the RZ, to a large degree, controls the chemistry of the runoff contributions along the stream network.

Hydrologic dynamics within landscape elements and the connections between them partially control the sources, flowpaths, amount, and age of water exiting the catchment through each segment of the RZ. Each landscape unit type has characteristic hydrologic and geomorphic attributes that can be assessed through field investigations and topographic analysis of emergent patterns and connections between landscape elements (landscape organization). Landscape analysis coupled with hydrological process investigation provides a context for scaling-up process knowledge from plot- and reach-scale hydrological/hydrochemical investigations to the catchment level. But HLA alone, without being tightly connected to field studies, risks being a theoretical exercise. Thus, there is a need to test landscape descriptors obtained by HLA against field observations.

While the linkage of catchments to water chemistry and aquatic biota is generally accepted, defining this relationship remains a challenge. Large-scale patterns in geology, soils, climate, and land use can define some degree of regional variability in mean water quality parameters (e.g. Lepistö *et al.*, 1995). There remains, however, a significant degree of small-scale variability in water chemistry, especially in the flow-related changes often crucial for the aquatic biota (Levin, 1992). Efforts to understand the controls on flow-related water quality in specific sites have increasingly recognized the riparian (near stream) zone as the key to understanding how water quality is linked to catchments (Hooper *et al.*, 1997). This realization grows out of research from parts of the globe as disparate as Antarctica (Gooseff *et al.*, 2002) and New Zealand (McGlynn and McDonnell, 2003a), on a range of issues from eutrophication to mercury and carbon budgets.

Hillslope hydrology is predominantly controlled by topography in catchments with shallow soil and poorly permeable bedrock, which are typical for boreal regions. McGlynn and Seibert (2003a) recently examined the variability in, and controls on, hillslope inputs to stream networks and the potential for riparian zones to regulate hillslope inputs and thereby both quantitatively and qualitatively buffer, or modify, stream responses to hillslope hydrology. We found that the ratio of riparian zone storage to hillslope inputs was the most important plot-scale measure of the buffering capacity of the riparian zone. One particularly important finding was that the catchment-wide proportion of the riparian area might be misleading. At the 280 ha Maimai research area, the ratio was 0.14. When we calculated this "buffer capacity" for each 20 m stream reach along the stream network, the values were below 0.14 for 75% of the stream length and the median was 0.06 (Figure 3). Using the catchment-wide ratio, thus, would significantly overestimate the "effective" riparian-to-hillslope area ratio.

Clearly, catchments will differ in the degree to which riparian zones buffer the delivery of water from hillslopes to streams, thereby affecting the amount, timing, and quality of hillslope water inputs expressed in streamflow (McGlynn *et al.*, 1999; McGlynn and McDonnell, 2003a). The functioning of the riparian zone depends on the interaction of different variables. Vidon and Hill (2004) studied the nitrate removal in riparian zones and suggested linkages between landscape characteristics and nitrate removal. Important characteristics were topography, riparian soil texture, and the depth of permeable sediments in the riparian as well as the upslope zone.

Field-based mapping of riparian zones is usually not feasible for larger catchments. In this case, a DEM might be used. Calculation of the "elevation-above-stream" along flowpaths to the stream and then specifying a threshold can be used to delineate the riparian zone (Figure 4a). This assessment, when combined with the lateral upslope accumulated area, can be used to determine the reach buffering capacity (Figure 4b).

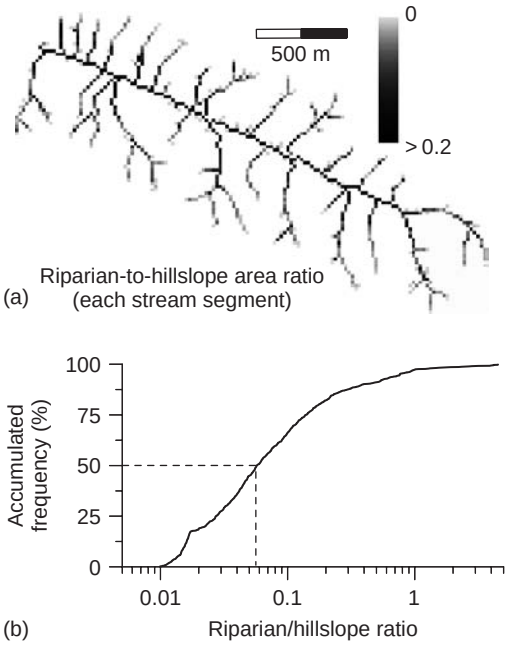


Figure 3 (a) Spatial distribution of riparian buffering potential (riparian/hillslope ratios) along the 280 ha Maimai catchment (Adapted from McGlynn and Seibert, 2003a). (b) Probability distribution of riparian-to-hillslope area ratios corresponding to figure 3(a)

Hydrological landscape analysis can also provide a quantification of the topographic control on water age or residence time distributions. Often, it is expected that mean stream water age for runoff increases with catchment area. However, this has not yet been demonstrated. Both McGlynn *et al.* (2003) and McGuire *et al.* (2005) found that mean streamwater age was not correlated to catchment area. McGlynn *et al.* (2003) found that the median subcatchment area was correlated with mean residence time, whereas McGuire *et al.* (2005) found that the median flowpath length and gradient of all catchment flowpaths were best correlated with residence time (Figure 5, adapted from McGuire *et al.* (2005).

The term *connectivity* is widely used in landscape ecology (Goodwin, 2003), but it is also of importance in hydrology (Western *et al.*, 1998; Stieglitz *et al.*, 2003). Spatial variability of soil thickness might be a major control

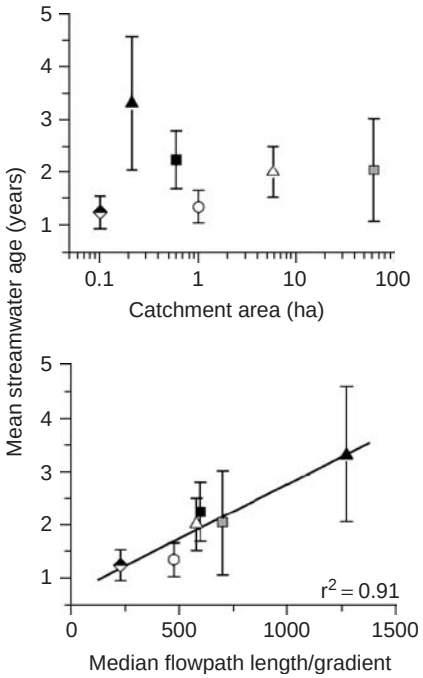


Figure 5 The relationships between mean residence time estimated by modeling $\delta^{18}O$ variations in stream water and catchment area and the ratio of median flowpath length to median flowpath gradient. Median flowpath values were determined from all potential flowpaths defined by a DEM analysis (McGlynn and Seibert, 2003b, adapted from McGuire *et al.*, 2005)

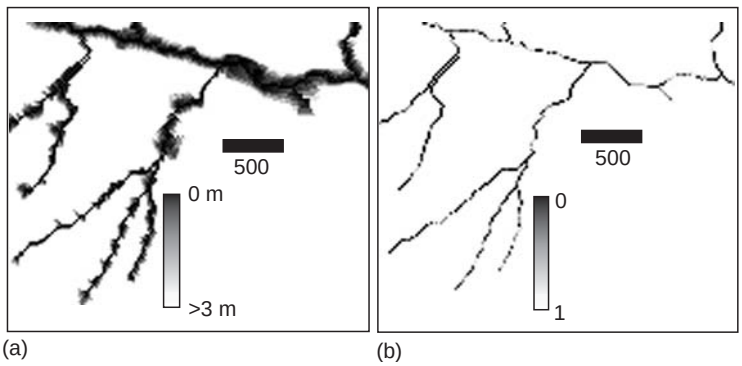


Figure 4 (a) Riparian area delineation on the basis of a threshold elevation (3m) above the stream channel following flowpaths to the stream within the West Fork Catchment, Big Sky Montana. (b) Distributed riparian buffering potential (the ratio of riparian area flowing into each stream cell relative to the hillslope area flowing into each stream cell) along the stream network within the West Fork Catchment, Big Sky, Big Sky Montana (0 = poor potential, 1 = high potential)

on hydrologic connectivity (Buttle *et al.*, 2004). Connections and disconnections of dominant runoff contributing areas are not only important for the runoff volume, but also for its chemical composition. In the latter case, the major process might be a simple mixing of the flux from the hillslopes with the water stored in the riparian zone or a complete biogeochemical resetting of the chemical composition in the riparian zone (Buttle, this issue). For the Maimai catchment, the relative timing of riparian and hillslope source contributions, and the connections and disconnections of dominant runoff contributing areas, were the first-order controls on stream DOC fluxes (McGlynn and McDonnell, 2003b).

LANDSCAPE ELEMENTS IN HYDROLOGICAL MODELING

In hydrologic modeling, various methods are used for dividing catchments into spatial units. The catchment can be treated as single unit (lumped approach) or divided into a large number of small grid cells (fully distributed approach) (**Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**). In between these fall approaches that divide the catchment into functional units on the basis of landscape information. Such units, which can be distributed continuously or discontinuously over the catchment, might include subcatchments, elevation zones, or zones for which similar hydrological behavior is assumed based on landscape characteristics. Terms such as hydrotopes, landscape elements, or hydrological response units (HRUs) are used for these zones, which are delineated on the basis of overlay of different spatial information. The basic concept is to overcome the crude simplification of lumped approaches while avoiding the computational burden of fully distributed approaches. On one hand we know that hydrological conditions vary within a catchment; on the other hand, we also often recognize that there are areas with fairly similar conditions. Recent research using TINs has incorporated the concept of hydrological similarity into TIN derivation (Vivoni *et al.*, 2004). We suggest that the delineation of areas with similar hydrological functioning allows consideration of the spatial variability of the most important controls on hydrological processes.

Dividing the catchment into subcatchments is a common approach to account for some spatial variation. The delineation into different elevation zones is also a common method for the representation of spatial variability. The latter approach is useful for snowmelt modeling as in the HBV model (Bergström, 1976), which is a conceptual model of catchment hydrology that in its simplest application is a purely lumped model. In other applications of the HBV model however, the catchment can be divided into elevation zones, vegetation zones, and subcatchments.

The Precipitation-runoff Modeling System (PRMS) developed at the US Geological Survey (USGS) (Leavesley and Stannard, 1995) divides the catchment into HRUs that are assumed to be homogeneous in their hydrologic functioning. The hydrology of each HRU is simulated using a daily time step and the sum of the runoff of all HRUs, weighted by the respective areas, is computed as daily catchment runoff. The Semidistributed Land-Use-based Runoff Processes (SLURP) model (Kite, 1995) is another example of the approach to divide the catchment both into subcatchments and into hydrologically similar subunits. The latter may be discontinuous within a subcatchment and are aggregated into aggregated simulation areas (ASA). Topography and land use information are used to define these areas.

In many landscape-element-based modeling approaches, there is no lateral flow between the elements, but the assumption is that each element contributes directly to the stream network. Obviously, consideration of lateral flow between elements is important, such as in the connections between upslope and riparian zones. Examples where lateral flow between landscape elements has explicitly been considered are WATBAL (Knudsen *et al.*, 1986), the PRMS application by Flügel (1995), TAC-D (Uhlenbrook *et al.*, 2004) and, for large-scale modeling, Güntner and Bronstert (2004). These models mainly use a semi-distributed approach with the exception of TAC-D, which is a grid-based model in which runoff generation routines and model parameters are assigned to the grid cells on the basis of HRUs.

Topographic indices can be used to delineate landscape elements; however, they can also provide another way to represent spatial variability. The assumption is that the locations with the same index value are hydrologically similar. The topographic index can be computed from gridded digital elevation data at high resolution, but model simulations are not needed for all grid cells. Instead, the simulations are executed for a small number of index value classes and then mapped back into space using the spatial distribution of the topographic index. One example of this approach is the TOPMODEL. The TOPMODEL approach (Beven and Kirkby, 1979; Beven *et al.*, 1995) has become one to be widely used for hydrological catchment modeling, because it allows consideration of topography while avoiding the complexity of fully distributed models. In contrast to the above-mentioned models, the catchment is not divided into homogeneous units, but rather the heterogeneity (i.e. topography) and its effects on hydrological processes are represented by the distribution of the TWI. The local wetness status is computed on the basis of a catchment average and the local TWI value. Therefore, the distribution of wetness states over a catchment can be simulated easily with low computational demands. For runoff simulations, only the distribution function of TWI is necessary, since

it determines the percentage of saturated areas, given a certain catchment average wetness status. Although the spatially distributed predictions of TOPMODEL usually appear reasonable, validation of spatial variations of groundwater levels (or surface wetness) simulated by TOPMODEL have often been unsuccessful (e.g. Burt and Butcher, 1985; Iorgulescu and Jordan, 1994; Seibert *et al.*, 1997).

Topographic indices are widely used for snow modeling (Hock, 2003; **Chapter 114, Snowmelt Runoff Generation, Volume 3**). Most models use elevation zones and a temperature lapse rate to represent the snow dynamics at different elevations. There are, however, other approaches that make more use of topographic data. Winstral *et al.* (2002) used topographic indices to predict the effects of wind-redistribution on spatial snow accumulation. The simple degree-day method for snowmelt modeling can be modified to allow spatially distributed simulations by using topographically based radiation indices (Cazorzi and Dalla Fontana, 1996; Hock, 2003).

While these approaches are promising, it must be recognized that topographic indices have limitations for modeling. An index provides only a part of the information available from a DEM, and important factors other than topography exist. It might be valuable to combine the landscape-element-based modeling approaches with topographic indices, which can be used to describe variability within an element.

As discussed in reference to landscape-unit-based modeling approaches, topographic-index-based modeling approaches usually do not consider lateral connections. In many cases, the spatial pattern of a topographic index map is of no importance other than its distribution function. Hydrological landscape analysis as discussed above (e.g. the computation of riparian-hillslope ratios along the stream network) is a way to consider lateral connections explicitly and, thus, to address the effects of spatial patterns and landscape organization.

LANDSCAPE ELEMENTS AND HETEROGENEITY

The use of landscape elements implies the consideration of heterogeneity and homogeneity. However, it is not simply that heterogeneity is limited to between-element variations and that each unit is homogeneous. Even within elements there is heterogeneity to some degree. The assumption, however, is that this heterogeneity and its effects are small compared to that between elements, or that this heterogeneity can be described statistically. The first applies to the concept of landscape elements as discussed in this article. While there are obviously large variations within, for instance, the riparian and respective hillslope zones, we argue that the major functional difference is between the two landscape elements rather than within each element. If necessary, further division of landscape elements (e.g. concave hillslopes and convex hillslopes) might be necessary to ensure functional homogeneity.

The statistical description of within-unit heterogeneity follows a different and partly a contradictory approach. Instead of seeking (functional) homogeneity, the idea is rather to "average out" the heterogeneity by sampling over a large enough area. The optimal size of such landscape elements for hydrological modeling is an open question. Wood *et al.* (1988) and Beven *et al.* (1988) proposed the representative elementary area (REA) as a fundamental building block of catchment modeling (Blöschl *et al.*, 1995). The REA can be compared to the REV in soil science (Bear, 1972) – an area where variability is minimized. The idea is that, at a certain area the landscape is amalgamated sufficiently to represent a mixture of the important characteristics contributing to variability at smaller spatial scales. While the frequency distribution of those characteristics may still be important, the spatial pattern is less important at this scale (Beven, 1995). In initial studies, the REA was determined on the basis of hydrologic modeling (TOPMODEL), where topography and rainfall were considered to vary in space (Wood *et al.*, 1988). Later, the REA concept was also tested on measurements of runoff and runoff chemistry (Woods *et al.*, 1995; Wolock *et al.*, 1997; Shaman *et al.*, 2004). Fan and Bras (1995) raised doubts about the existence and potential utility of the REA concept. A major problem with the REA approach is that the landscape characteristics vary at a range of different scales and that the size of an REA might depend on the considered variable (e.g. storm flow, baseflow, chemistry, or evaporation). The REA concept represents an attempt to quantify the spatial scale where variability in landscape element topology is subsumed.

CONCLUDING REMARKS

We suggest that landscape analysis might be the basis for new model approaches that fill the gap between fully distributed models using a regular grid and models that are more or less lumped. Hydrological landscape analysis allows for identification of crucial landscape elements and their lateral connections utilizing digital elevation data and the assumption that flow paths follow topography. The wide use of the TWI has promoted the value of topographic data in hydrology. We suggest that there is a suite of additional topographical indices worth testing that can be derived by landscape analysis. On the basis of this, modeling efforts might be distributed more rationally than modeling each grid cell. For instance, it might be reasonable to have a more detailed model for the relatively small riparian zones, whereas a simpler approach might be sufficient for the larger upslope areas. Explicit consideration of lateral connections between landscape elements is critical for

a realistic description of the evolution of water quality along the flow pathways through mixing or biogeochemical processes. Even when landscape elements derived from, for instance, soil or vegetation data are used, the topographical data might still be useful to establish the linkages between these landscape elements. Advances in remote sensing and other measurement techniques as well as improved interpretation and the use of observed spatial patterns (Grayson *et al.*, 2002) will support the further development of hydrological landscape analysis and associated modeling approaches.

REFERENCES

Bates P.D., Stewart M.D., Desitter A., Anderson M.G., Renaud J.P. and Smith J.A. (2000) Numerical simulation of floodplain hydrology. *Water Resources Research*, **36**, 2517–2529.

Bear J. (1972) *Dynamics of Fluids in Porous Media*, American Elsevier Pub. Co.: New York.

Beaujouan V., Durand P., Ruiz L., Aurousseau P. and Cotteret G. (2002) A hydrological model dedicated to topography-based simulation of nitrogen transfer and transformation: rationale and application to the geomorphology-denitrification relationship. *Hydrological Processes*, **16**, 493–507.

Beven K. (1995) Linking parameters across scales – subgrid parameterizations and scale-dependent hydrological models. *Hydrological Processes*, **9**, 507–525.

Beven K.J., Lamb R., Quinn P., Romanowicz R. and Freer J. (1995) TOPMODEL. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Ranch, pp. 627–668.

Beven K.J. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Beven K.J., Wood E.F. and Sivapalan M. (1988) On hydrological heterogeneity – catchment morphological and catchment response. *Journal of Hydrology*, **100**, 353–375.

Bergström S. (1976) *Development and Application of a Conceptual Runoff Model for Scandinavian catchments*, SMHI, RHO No. 7, Norrköping, p. 134.

Bishop K. (1991) *Episodic Increases in Stream Acidity, Catchment Flow Pathways and Hydrograph Separation*, Ph.D., University of Cambridge, Cambridge.

Blavet D., Matheb E. and Lepruna J.C. (2000) Relations between soil colour and waterlogging duration in a representative hillside of the West African granito-gneissic bedrock. *Catena*, **39**, 187–210.

Blöschl G., Grayson R.B. and Sivapalan M. (1995) On the representative elementary area (REA) concept and its utility for distributed rainfall-runoff modelling. *Hydrological Processes*, **9**, 313–330.

Brinson M.M. (1993) *A Hydrogeomorphic Classification for Wetlands*, WRP-DE-4, US Army Corps of Engineers: Springfield, VA.

Brinson M.M., Swift B.L., Plantico R.C. and Barclay J.S. (1981) *Riparian Ecosystems: Their Ecology and Status*, U.S. FWS/OBS-81/17, Fish and Wildlife Service, Biological Services Program: Wasington, D.C.

Burt T.P. and Butcher D.P. (1985) Topographic controls of soil moisture distribution. *Journal of Soil Science*, **36**, 469–486.

Butterworth R., Wilson C.J., Herron N.F., Greene R.S.B. and Cunningham R.B. (2000) Geomorphic controls on the physical and hydrologic properties of soils in a valley floor. *Earth Surfaces Processes and Landforms*, **25**, 1161–1179.

Buttle J.M., Dillon P.J. and Eerkes G.R. (2004) Hydrologic coupling of slopes, riparian zones and streams: an example from the Canadian Shield. *Journal of Hydrology*, **287**, 161–177.

Cazorzi F. and Dalla Fontana G. (1996) Snowmelt modelling by combining air temperature and a distributed radiation index. *Journal of Hydrology*, **181**, 169–187.

Cirmo C.P. and McDonnell J.J. (1997) Linking the hydrologic and biogeochemical controls of nitrogen transport in near-stream zones of temperate-forested catchments: a review. *Journal of Hydrology*, **199**, 88–120.

Dunne T. and Black R.D. (1970) Partial area contributions to storm runoff in a small New England watershed. *Water Resources Research*, **6**, 1296–1311.

Famiglietti J.S. and Wood E.F. (1991) Evapotranspiration and runoff from large land areas – land surface hydrology for atmospheric general-circulation models. *Surveys in Geophysics*, **12**, 179–204.

Fan Y. and Bras R.L. (1995) On the concept of a representative elementary area in catchment runoff. *Hydrological Processes*, **9**, 821–832.

Faulkner S.P. and Patrick W.H. (1992) Redox processes and diagnostic wetland soil indicators in bottomland hardwood forests. *Soil Science Society of America Journal*, **56**, 856–865.

Flügel W.-A. (1995) Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using prms/mms in the drainage basin of the river bröl, Germany. *Hydrological Processes*, **9**, 423–436.

Freer J., McDonnell J.J., Beven K.J., Peters N.E., Burns D.A., Hooper R.P., Aulenbach B. and Kendall C. (2002) The role of bedrock topography on subsurface storm flow. *Water Resources Research*, **38**(12), 1269.

Genereux D.P. and Hooper R.P. (1998) Oxygen and hydrogen isotopes in rainfall-runoff studies. In *Isotope Tracers in Catchment Hydrology*, Chap. 10, Kendall C. and McDonnell J.J. (Eds.), Elsevier Science, pp. 839.

Giesler R., Högberg M. and Högberg P. (1998) Soil chemistry and plants in Fennoscandian boreal forest as exemplified by a local gradient. *Ecology*, **79**, 119–137.

Goodwin B.J. (2003) Is landscape connectivity a dependent or independent variable? *Landscape Ecology*, **18**, 687–699.

Gooseff M.N., McKnight D.M., Lyons W.B. and Blum A.E. (2002) Weathering reactions and hyporheic exchange controls on stream water chemistry in a glacial meltwater stream in the McMurdo Dry Valleys. *Water Resources Research*, **38**(12), 1279, doi:10.1029/2001WR000834.

Grayson R.B., Blöschl G., Western A.W. and McMahon T.A. (2002) *Advances in the Use of Observed Spatial Patterns of Catchment Hydrological Response*, Advances in Water Resources 25th Anniversary Issue (invited paper), pp. 1313–1334.

Güntner A. and Bronstert A. (2004) Representation of landscape variability and lateral redistribution processes for large-scale hydrological modelling in semi-arid areas. *Journal of Hydrology*, **297**, 136–161.

Güntner A., Seibert J. and Uhlenbrook S. (2004) Modeling spatial patterns of saturated areas: An evaluation of different terrain indices. *Water Resources Research*, **40**, W05114, doi:10.1029/2003WR002864.

Hewlett J.D. and Hibbert A.R. (1967) Factors affecting the response of small watersheds to precipitation in humid areas. In *Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: New York, pp. 275–291.

Hill A.R. (1996) Nitrate removal in stream riparian zones. *Journal of Environmental Quality*, **25**, 743–755.

Hjerdt K.N., McDonnell J.J., Seibert J. and Rodhe A. (2004) A new topographic index to quantify downslope controls on local drainage. *Water Resources Research*, **40**, W05602, doi:10.1029/2004WR003130.

Hock R. (2003) Temperature index melt modelling in mountain regions. *Journal of Hydrology*, **282**(1–4), 104–115, doi:10.1016/S0022-1694(03)00257-9.

Holmgren P. (1994) Topographic and geochemical influence on the forest site quality, with respect to pinus sylvestris and picea abies in Sweden. *Scandinavian Journal of Forest Research*, **9**, 75–82.

Hooper R., Aulenbach B., Burns D., McDonnell J.J., Freer J., Kendall C. and Beven K. (1997) Riparian control of stream-water chemistry: implications for hydrochemical basin models. *IAHS Redbook*, **248**, 451–458.

Horton R.E. (1933) The role of infiltration in the hydrologic cycle. *Transactions of the American Geophysical Union*, **14**, 446–460.

Iorgulescu I. and Jordan J.P. (1994) Validation of TOPMODEL on a small Swiss catchment. *Journal of Hydrology*, **159**, 255–273.

Kendall K.A., Shanley J.B. and McDonnell J.J. (1999) A hydrometric and geochemical approach to test the transmissivity feedback hypothesis during snowmelt. *Journal of Hydrology*, **219**, 188–205.

Kite G.W. (1995) The SLURP model. In *Computer Models of Watershed Hydrology*, Chap. 15, Singh V.P. (Ed.), Water Resources Publications: Colorado, pp. 521–562.

Knudsen J., Thomsen A. and Refsgaard J.C. (1986) Watbal – a semi-distributed, physically based hydrological modeling system. *Nordic Hydrology*, **17**, 347–362.

Leavesley G.H. and Stannard L.G. (1995) The precipitation-runoff modeling system – PRMS. In *Computer Models of Watershed Hydrology*, Chap. 9, Singh V.P. (Ed.), Water Resources Publications: Highlands Ranch, pp. 281–310.

Lepistö A., Andersson L., Arheimer B. and Sundblad K. (1995) Effect of geographical factors, forestry activities and deposition on nitrogen load from small forested catchments. *Water Air and Soil Pollution*, **84**, 81–102.

Levin S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Lowrance R.R., Leonard R. and Sheriden J. (1985) Managing riparian ecosystem to control nonpoint source pollution. *Journal of Soil and Water Conservation*, **40**, 87–91.

McDonnell J.J., Freer J., Hooper R., Kendall C., Burns D., Beven K. and Peters J. (1996) New method developed for studying flow in hillslopes. *EOS, Transactions of the American Geophysical Union*, **77**, 465.

McGlynn B.L. The role of riparian zones in steep mountain watersheds. In *Global Change and Mountain Regions*, *"Advances in Global Change Research Book Series"*,, Beniston M. (Ed), Kluwer, In press.

McGlynn B.L. and McDonnell J.J. (2003a) Quantifying the relative contributions of riparian and hillslope zones to catchment runoff and composition. *Water Resources Research*, **39**(11), 1310, Vol. No. 10.1029/2003WR002091.

McGlynn B.L. and McDonnell J.J. (2003b) The role of discrete landscape units in controlling catchment dissolved organic carbon dynamics. *Water Resources Research*, **39**(4), 1090, doi:10.1029/2002WR001525.

McGlynn B.L., McDonnell J.J., Shanley J.B. and Kendall C. (1999) Riparian zone flowpath dynamics during snowmelt in a small headwater catchment. *Journal of Hydrology*, **222**, 75–92.

McGlynn B.L. and Seibert J. (2003a) Distributed assessment of contributing area and riparian buffering along stream networks. *Water Resources Research*, **39**(4), 1082, doi:10.1029/2002WR001521.

McGlynn B.L. and Seibert J. (2003b) *DEM-Based Analysis of Landscape Organization: 1) Riparian to Hillslope Area Ratios*, European Geophysical Union-American Geophysical Union-European Union of Geosciences Joint Assembly: Nice, April 2003.

McGlynn B.L., McDonnell J.J., Seibert J. and Kendall C. (2004) Scale effects on headwater catchment runoff timing, flow sources, and groundwater-streamflow relations. *Water Resources Research*, **40**, W07504, doi:10.1029/2003 WR002494.

McGlynn B.L., McDonnell J.J., Seibert J. and Stewart M.K. (2003) On the relationships between catchment scale and streamwater mean residence time. *Hydrological Processes*, **17**, 175–181.

McGuire K.J., McDonnell J.J., Weiler M., Kendall C., McGlynn B.L., Welker J.M. and Seibert J. (2005) The role of topography on catchment-scale water residence time. *Water Resources Research*, **41**, W05002, doi: 10.1029/2004WR003657.

McKenzie N.J., Gessler P.E., Ryan P.J. and O'Connell D.A. (2000) The role of terrain analysis in soil mapping. In *Terrain Analysis: Principles and Applications*, Chap. 10, Wilson J.P. and Gallant J.C. (Eds.), John Wiley & Sons, p. 479.

Megonigal J.P., Patrick W.H. and Faulkner S.P. (1993) Wetland identification in seasonally flooded forest soils: soil morphology and redox dynamics. *Soil Science Society of America Journal*, **57**, 140–149.

Merot P., Ezzahar B., Walter C. and Aurousseau P. (1995) Mapping waterlogging of soils using digital terrain models. *Hydrological Processes*, **9**, 27–34.

Merz B. and Plate E.J. (1997) An analysis of the effects of spatial variability of soil and soil moisture on runoff. *Water Resources Research*, **33**, 2909–2922.

Mitsch W.J. and Gosselink J.G. (1993) *Wetlands*, *Second Edition*, Van Nostrand Reinhold: New York.

Moore I.D. and Grayson R.B. (1991) Terrain-based catchment partitioning and runoff prediction using vector elevation data. *Water Resources Research*, **27**, 1177–1191.

Moore I.D., Grayson R.B. and Ladson A.R. (1991) Digital terrain modeling – a review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5**, 3–30.

Moore I.D., Norton T.W. and Williams J.E. (1993) Modelling environmental heterogeneity in forested landscapes. *Journal of Hydrology*, **150**, 717–747.

Naef F., Scherrer S. and Weiler M. (2002) A process based assessment of the potential to reduce flood runoff by land use change. *Journal of Hydrology*, **267**, 74–79.

Naiman R.J. and Décamps H. (1997) The ecology of interfaces: Riparian zones. *Annual Review of Ecology and Systematics*, **28**, 621–658.

Pachepsky Y.A., Timlin D.J. and Rawls W.J. (2001) Soil water retention as related to topographic variables. *Soil Science Society of America Journal*, **65**, 1787–1795.

Park S.J. and van de Giesen N. (2004) Soil–landscape delineation to define spatial sampling domains for hillslope hydrology. *Journal of Hydrology*, **295**, 28–46.

Pellenq J., Kalma J., Boulet G., Saulnier G.M., Wooldridge S., Kerr Y. and Chehbouni A. (2003) A disaggregation scheme for soil moisture based on topography and soil depth. *Journal of Hydrology*, **276**, 112–127.

Phillips D.H., Fossa J.E., Stilesa C.A., Trettinb C.C. and Luxmoorec R.J. (2001) Soil-landscape relationships at the lower reaches of a watershed at Bear Creek near Oak Ridge, Tennessee. *Catena*, **44**, 205–222.

Pritchett W.L. and Fisher R.F. (1987) *Properties and Management of Forest Soils*, *Second Edition*, Wiley: New York.

Robson A., Beven K. and Neal C. (1992) Towards identifying sources of subsurface flow: A comparison of components identified by a physically based runoff model and those determined by chemical mixing techniques. *Hydrological Processes*, **6**, 199–214.

Seibert J., Bishop K.H. and Nyberg L. (1997) A test of TOPMODEL's ability to predict spatially distributed groundwater levels. *Hydrological Processes*, **11**, 1131–1144.

Sivapalan M. and Wood E.F. (1987) A multidimensional model of nonstationary space-time rainfall at the catchment scale. *Water Resources Research*, **23**, 1289–1299.

Sivapalan M., Wood E.F. and Beven K.J. (1990) On hydrologic similarity. 3. A dimensionless flood frequency model using a generalized geomorphologic unit hydrograph and partial area runoff generation. *Water Resources Research*, **26**, 43–58.

Shaman J., Stieglitz M. and Burns D. (2004) Are big basins just the sum of small catchments?. *Hydrological Processes*, **18**(16), 3195–3206.

Stieglitz M., Shaman J., McNamara J., Engel V., Shanley J. and Kling G.W. (2003) An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport. *Global Biogeochemical Cycles*, **17**(4), 1105, doi:10.1029/2003 GB002041.

Swanson F.J., Gregory S.V., Sedell J.R. and Campbell A.G. (1982) Land-water interactions: the riparian zone. In *Analysis of Coniferous Forest Ecosystems in the Western United States*, Edmonds R.L. (Ed.), Hutchinson Ross Publishing Company: Stroudsburg, pp. 267–291.

Tucker G.E., Lancaster S.T., Gasparini N.M., Bras R.L. and Rybarczyk S.M. (2001) An object-oriented framework for distributed hydrologic and geomorphic modeling using triangulated irregular networks. *Computers & Geosciences*, **27**, 959–973.

Uhlenbrook S., Roser S. and Tilch N. (2004) Hydrological process representation at the meso-scale: the potential of a distributed, conceptual catchment model. *Journal of Hydrology*, **291**, 278–296.

Vidon P.G.F. and Hill A.R. (2004) Landscape controls on the hydrology of stream riparian zones. *Journal of Hydrology*, **292**, 210–228.

Vivoni E.R., Ivanov V.Y., Bras R.L. and Entekhabi D. (2004) Generation of triangulated irregular networks based on hydrological similarity. *Journal of Hydrologic Engineering*, **9**(4), 288–302.

Western A.W., Blöschl G. and Grayson R.B. (1998) How well do indicator variograms capture the spatial connectivity of soil moisture?. *Hydrological Processes*, **12**, 1851–1868.

Western A.W., Grayson R.B., Blöschl G., Willgoose G.R. and McMahon T.A. (1999) Observed spatial organization of soil moisture and its relation to terrain indices. *Water Resources Research*, **35**, 797–810.

White J.D. and Running S.W. (1994) Testing scale dependent assumptions in regional ecosystem simulations. *Journal of Vegetation Science*, **5**, 687–702.

Winstral A.H., Elder K. and Davis R.E. (2002) Spatial snow modeling of wing redistribution snow using terrain-based parameters. *Journal of Hydrometeorology*, **3**, 524–538.

Wolock D.M., Fan J. and Lawrence G.B. (1997) Effects of basin size on low-flow stream chemistry and subsurface contact time in the Neversink River watershed, New York. *Hydrological Processes*, **11**, 1273–1286.

Wood E.F., Sivapalan M., Beven K.J. and Band L.E. (1988) Effects of spatial variability and scale with implications to hydrologic modelling. *Journal of Hydrology*, **102**, 29–47.

Woods R.A., Sivapalan M. and Duncan M. (1995) Investigating the representative elementary area concept: an approach based on field data. *Hydrological Processes*, **9**, 291–312.

Woolhiser D.A., Smith R.E. and Giraldez J.V. (1996) Effects of spatial variability of saturated hydraulic conductivity on Hortonian overland flow. *Water Resources Research*, **32**, 671–678.

Zinko U., Seibert J., Dynesius M. and Nilsson C. Plant species numbers predicted by a topography based groundwater-flow index. *Ecosystems*, **8**, in press.

116: Isotope Hydrograph Separation of Runoff Sources

JIM M BUTTLE

Department of Geography, Trent University, Peterborough, ON, Canada

Use of isotopic tracers to characterize water ages, origins, and flow pathways in drainage basins is reviewed, along with the assumptions that underlie isotopic hydrograph separations (IHSs). Points of agreement regarding the use of the IHS approach in small drainage basins are examined, as are some of the key challenges associated with application of the approach in future research. Despite a growing awareness of the limitations of IHS, the method offers a powerful tool to examine a number of critical issues in hydrology, particularly when employed in conjunction with hydrometric techniques and the use of other tracers. These issues include the estimation of water residence times, the identification of process thresholds, the assessment of hydrological response to land use change, and the use of isotopic data to calibrate and test hydrological models.

INTRODUCTION

Isotope tracers and isotopic hydrograph separations (IHSs) are important tools for quantifying the age, origin, and pathway of water to streams in drainage basins, and thus for extending our understanding of streamflow generation (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3** and **Chapter 112, Subsurface Stormflow, Volume 3**). They can also provide complementary information on water sources and pathways that are often required in order to draw conclusions about such ecological issues as the degree of closure of nutrient cycles (e.g. Elsenbeer *et al.*, 1995) and the hydrological effects of ecosystem disturbance (e.g. Bruijnzeel, 1990). This article reviews the basic approach and underlying assumptions of IHSs, summarizes the major points of consensus and disagreement regarding results that have been obtained using this approach, examines the challenges facing the use of isotopic tracers, and suggests current research issues in hydrology that would particularly benefit from the use of the IHS technique.

THE ISOTOPE HYDROGRAPH SEPARATION APPROACH

IHSs have generally been conducted using tritium (^{3}H or T), oxygen-18 (^{18}O), and deuterium (^{2}H or D) (Buttle, 1994). Oxygen-18 and D have been used in the majority of IHSs (Sklash, 1990) since they are stable and do not undergo radioactive decay. Ratios of ^{18}O and D to their more common counterparts in the hydrosphere (^{16}O and H) are 1 : 500 and 1 : 6700 (Drever, 1988). Abundance of stable isotopes in a water sample is based on isotopic ratios (e.g. ^{18}O/^{16}O and D/H) and is reported as δ values in parts per thousand (‰ or per mil):

$$\delta^{18}\mathrm{O} \text{ or } \delta\mathrm{D} = \left(\frac{R_{\mathrm{sample}}}{R_{\mathrm{VSMOW}} - 1}\right) \bullet 1000 \qquad (1)$$

where R_{sample} is the ratio of the heavy to light isotope in the sample and R_{VSMOW} is the ratio in the reference standard, which is Vienna Standard Mean Ocean Water (VSMOW) for ^{18}O and D (Kendall and Caldwell, 1998)

Isotope tracers have several unique virtues as water tracers in basin studies:

1. Their natural application over entire basins avoids problems of realistic application extents and rates associated with artificial tracers (Sklash, 1990).
2. Their isotopic ratios are generally unaffected by reactions with soil/regolith at temperatures encountered at or near the earth's surface (Drever, 1988).
3. Fractionation during evaporation results in the relative depletion of heavy isotopes in water vapor while the remaining liquid water becomes progressively enriched in D and ^{18}O; conversely, there is

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

preferential movement of heavy isotopes to the liquid phase during condensation. This leaves the remaining vapor relatively depleted in D and ^{18}O, such that meteoric water has negative δ values that decrease with surface air temperature, increasing latitude, increasing altitude, increasing distance of vapor transport, and increasing amounts of precipitation (Dansgaard, 1964; Ingraham, 1998).

4. Variations in the isotopic signature of precipitation are often dampened during infiltration (Ingraham and Taylor, 1991). Groundwater δ values may approach uniformity in time and space, and are changed only by mixing with waters of different isotopic contents (Sklash, 1990). This often leads to differences between input water δ and that of water stored in the basin before the event.

This difference between the isotopic signature of inputs (event or "new" water) and water stored in the basin before the event (pre-event or "old" water) may allow the stormflow hydrograph to be partitioned into event (new) and pre-event (old) components:

$$Q_t = Q_p + Q_e \tag{2}$$

$$C_t Q_t = C_p Q_p + C_e Q_e \tag{3}$$

$$X = \frac{C_t - C_e}{C_p - C_e} \tag{4}$$

where Q_t is streamflow; Q_p and Q_e are contributions from pre-event and event water; C_t, C_p, and C_e are δ values in streamflow, pre-event, and event waters respectively; and X is the pre-event fraction of streamflow (Figure 1). C_p and C_e must bracket the C_t δ values in order to separate the streamflow hydrograph into event and pre-event components; in addition, IHS is based on several key assumptions.

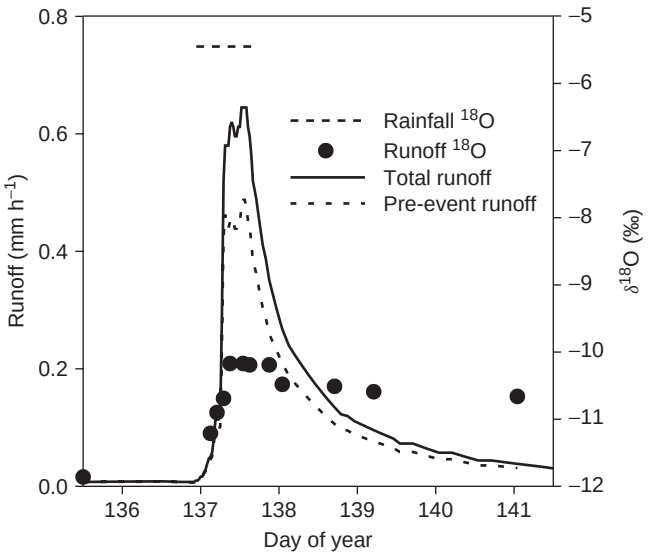


Figure 1 Isotopic hydrograph separation results from the PC1-08 basin during a large spring rainfall, Ontario, Canada (From Peters *et al.*, 1995)

ASSUMPTIONS UNDERLYING THE IHS TECHNIQUE

Use of equations (2)–(4) to solve for the event and pre-event fractions of streamflow assumes:

1. a significant difference between the isotopic content of the event and pre-event components;
2. an event water isotopic signature that is constant in space and time, or that any variations can be accounted for;
3. a pre-event water isotopic signature that is constant in space and time, or that any variations can be accounted for;
4. negligible water contributions from the vadose zone or a soil water isotopic content similar to that of groundwater;
5. negligible contributions to streamflow from surface storage.

Assumption 1 is met in most (if not all) IHS studies, since they have largely been conducted in mid-to-high latitude environments (Buttle, 1994). These regions often experience pronounced annual oscillations in the isotopic signature of precipitation (Figure 2). This increases the possibility of a significant difference between δ for a particular precipitation event and pre-event water, the mean δ of which approximates the mean δ of annual precipitation (Clark and Fritz, 1997; Gremillion and Wanielista, 2000).

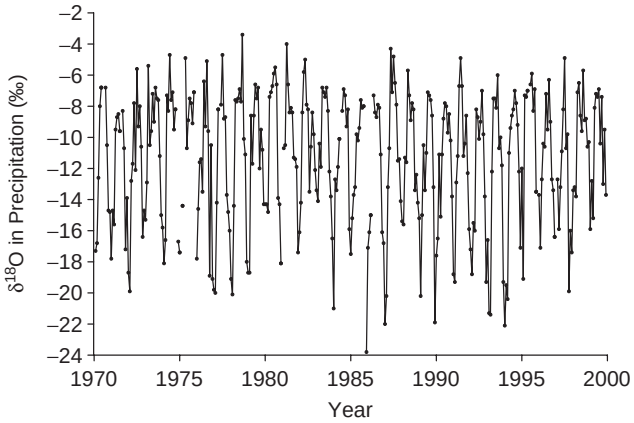


Figure 2 Time series of $\delta^{18}O$ in precipitation sampled at Ottawa, Canada (114 m asl; 45.32 °N latitude). Data from International Atomic Energy Agency/World Meteorological Organization (2001)

Potential violations of the remaining assumptions were often downplayed in many initial IHS studies. Event water δ was often represented by samples from a single rain gauge, snowcore, or snowmelt sample that may have been taken from outside the basin (Unnikrishna *et al.*, 2002), thus assuming spatial uniformity in input δ. The use of bulk rain or snowmelt samples necessitated the frequently untested assumption that temporal variations in event water δ were insignificant. Both assumptions have been shown to be violated in small basins (e.g. McDonnell *et al.*, 1990; DeWalle and Swistock, 1994; Bariac *et al.*, 1995). Precipitation in many IHS studies is measured and sampled at open sites. This presents a problem in forested basins, where interception may enrich the precipitation isotopic signature (Saxena, 1986; DeWalle and Swistock, 1994; Brodersen *et al.*, 2000) and thus reduce the difference between the isotopic content of the event and pre-event components to such a degree as to preclude IHS. Interestingly, many IHS studies have addressed the hypothesis that pre-event water δ was constant in time and space. Sklash (1990) argued that baseflow δ is the best index of C_p on the grounds that baseflow integrates the δ of near-stream groundwater likely to reach the stream during an event. This assumption has been supported by the close correspondence between groundwater and baseflow δ in some studies (e.g. Hooper and Shoemaker, 1986; Hill and Waddington, 1993), but has been called into question by substantial variability in baseflow δ along the stream channel and by significant differences between δ in near-stream groundwater and in baseflow (Buttle *et al.*, 1995; Bonell *et al.*, 1998; Burns and McDonnell, 1998; Gremillion and Wanielista, 2000). These differences have been attributed to spatial variations in groundwater residence times arising from geological complexity (Bonell *et al.*, 1998), and to evaporative enrichment of river water baseflow relative to contributing groundwater (Gremillion and Wanielista, 2000).

Invocation of assumption 4 allowed some early workers (e.g. Sklash and Farvolden, 1979) to assume that X represented both the pre-event and groundwater fraction of total stormflow. This requires that C_p has a constant value, such that C_t equals its pre-storm δ once discharge returns to baseflow values (e.g.Bonell *et al.*, 1990). Several studies have noted a shift in baseflow or groundwater δ in response to inputs during snowmelt or rainfall (e.g. Hooper and Shoemaker, 1986; Buttle *et al.*, 1995; McDonnell *et al.*, 1991a). Under these circumstances, C_p must reflect these temporal variations in baseflow or groundwater δ if X is to be interpreted as the groundwater fraction of discharge at the time of streamflow sampling.

Substantial differences between δ in soil water and in groundwater or baseflow have been shown (e.g. Kennedy *et al.*, 1986; DeWalle *et al.*, 1988; Peters *et al.*, 1995), although the ability of soil water to make significant contributions to stormflow is often unclear. This question has been addressed using two basic approaches. The first infers a soil water component of stormflow on the basis of an inadequate explanation of runoff sources using the standard two-component model (e.g. DeWalle *et al.*, 1988; Ogunkoya and Jenkins, 1993; Hinton *et al.*, 1994). The second uses hydrometric measurements to quantify soil water contributions to basin streamflow (e.g. Buttle and Peters, 1997). In situations where significant contributions to stormflow from one or more additional flow components (e.g. soil water) have been identified, the standard mixing equations have been modified:

$$Q_t = Q_1 + Q_2 + Q_3 + \cdots + Q_n \quad (5)$$

$$C_t Q_t = C_1 Q_1 + C_2 Q_2 + C_3 Q_3 + \cdots + C_n Q_n \quad (6)$$

where

Q_t = streamflow
Q_n = discharge of a particular runoff component
C_t = tracer concentrations in streamflow
C_n = tracer concentration of a particular runoff component.

These equations can be solved using matrix algebra, with the caveat of additional constraints and increased output uncertainty. Thus, solution for three flow components (e.g. event water, soil water, and groundwater) requires either a second tracer or a physical measurement of flow from one component (Genereux and Hooper, 1998). Stable isotope tracers have been used in conjunction with a geochemical hydrograph separation (GHS) to identify contributions from three flow components (e.g. Wels *et al.*, 1991; McDonnell *et al.*, 1991b; DeWalle and Pionke, 1994; Hinton *et al.*, 1994). Unlike isotopic tracers, geochemical tracers provide information about hydrologic flowpaths provided the kinetics of tracer reactivity in the subsurface are known (Burns and McDonnell, 1998; Burns *et al.*, 2001). A good example of attempts to quantify flow for a particular component is that of DeWalle *et al.* (1988), who estimated channel precipitation (direct supply of event water to streamflow) as the product of throughfall and stream surface area. The latter was obtained through regression of measured stream surface areas on the corresponding streamflow rate. In the absence of such hydrometric measurements, n tracers are required to separate stormflow into $n + 1$ components.

Many initial IHS studies purposely avoided basins with appreciable surface storage (e.g. lakes, ponds, wetlands). This appears to have been well advised, since mixing within wetlands can complicate interpretation of results from traditional two-component IHSs (e.g. Buttle and Sami, 1992; Hill and Waddington, 1993; Burns and McDonnell, 1998; Metcalfe and Buttle, 2001).

In addition to greater evaluation of the assumptions underlying the IHS approach, there has been an increased analysis of the errors associated with IHS results. Rodhe (1987) used a first-order uncertainty propagation approach to an explicit error analysis that was later refined by Genereux (1998). Conversely, Bazemore *et al.* (1994) used a Monte Carlo approach to analyze error in hydrograph separation results. Both methods assume that the variability in the flow components signatures follows normal distributions. Joerin *et al.* (2002) combined a Monte Carlo approach with component frequency distributions determined directly from field samples to estimate statistical uncertainty in hydrograph separations. This study is also noteworthy for its distinction between statistical uncertainty and the "model uncertainty" arising from such assumptions as temporal uniformity of flow component signatures.

RESULTS OF IHS STUDIES IN SMALL BASINS

The most important overall finding from the use of IHS in small basins is the significant contribution that pre-event water makes to peak stormflow and the total stormflow hydrograph in many environments (Buttle, 1994; Richey *et al.*, 1998). Although there is a great range in pre-event water fractions (X) of stormflow for various basin sizes, land uses, and event types (Shanley *et al.*, 2002), two salient points have emerged. The first is that X is generally smaller for basins with land uses (e.g. urban – Halldin *et al.*, 1990; Buttle *et al.*, 1995; Gremillion *et al.*, 2000) or surface types (e.g. permafrost – Cooper *et al.*, 1991; Metcalfe and Buttle, 2001) that promote the contribution of surface runoff to stormflow production through infiltration excess overland flow. This process likely generates a greater event water fraction of stormflow from such areas. The second is that X is generally smaller in forested basins in humid temperate climates during spring snowmelt, attributed to greater surface runoff as a result of frozen soils, which may develop during some winters, and maximum extents of saturated near-stream areas.

Beyond these general findings, a broader consensus regarding IHS results is more difficult to discern. For example, there is no agreement on the effects of basin size on new/old water partitioning. Thus, some studies have observed increased new water percentages with increasing basin size (Shanley *et al.*, 2002; McDonnell *et al.*, 1999) while others have found the reverse (Brown *et al.*, 1999). There is also some question as to the degree to which IHS results have been accepted by the broader scientific community. Although some hydrologists have argued that the significant contribution of pre-event to stormflow has promoted a paradigm shift in hydrological thought, there is abundant evidence that IHS results have not been incorporated into many current basin-scale hydrological models. In addition, the pre-event water paradigm largely applies in environments where infiltration excess overland flow is relatively rare, and intense rainfalls in parts of the humid tropics may mean that such overland flow is more prevalent than in the western European and North American studies that have used environmental isotopic tracers (Elsenbeer, 2001).

What other areas of agreement exist with regard to the results of IHS studies? First, we must recognize that the interpretation of IHSs occurs in the context of equifinality (similar outcome generated by a range of alternative mechanisms). Various hydrological processes may be responsible for the rapid delivery of significant quantities of pre-event water to the stream channel, and their degree of importance varies both spatially and temporally within a basin (Buttle, 1994). Equifinality seriously compromises attempts to infer intra-basin processes from the isotopic response at the basin outlet. It also provides a compelling case for integrating isotopic tracers with other tracers as well as with hydrometric techniques to constrain a process interpretation more fully. This leads to the second point of agreement, which is that hydrological science is not advanced appreciably by using IHS alone (Bonell *et al.*, 1998; Rice and Hornberger, 1998; Burns, 2002). We must combine isotopic and geochemical tracers with hydrometric measurements in order to infer the correct hydrological pathways in a basin (Elsenbeer *et al.*, 1995). Third, we must recognize that IHS and GHS studies at the basin scale are black-box approaches that assume homogeneous distributions of flowpaths and other hydrologic properties along with uniform isotope and chemical compositions of input waters (Kendall *et al.*, 2001). The spatial lumping that generally underlies application of the IHS/GHS technique in isolation at the basin scale does not shed light on intrastorm changes in water sources, flowpaths, and the processes acting along these flowpaths (Kendall *et al.*, 2001).

CHALLENGES FACING THE IHS APPROACH

There are a number of conceptual and methodological challenges facing the IHS approach to identifying runoff sources and pathways that need to be addressed in future work:

The Influence of in-channel Processes on Basin Isotopic Response

Many of our conceptual models of streamflow generation envisage that water delivery via various pathways from slope to stream channel occurs normal to the channel margin, and the models often ignore the role of in-channel processes. Nevertheless, complex transfers of water between the stream and its bed and banks during channel flow (hyporheic exchange – Bencala, 2000, *see* **Chapter 113, Hyporheic Exchange Flows, Volume 3**) may alter the isotopic signal of slope runoff from the point of entry to the channel to the sampling point at the basin outlet.

This presents both a challenge and an opportunity. The challenge is determining how divergences between flood wave and water particle travel times are influenced by mixing processes in the hyporheic zone. The opportunity is the potential to use isotopic tracers to distinguish between water residence times on slopes and in the hyporheic zone. Knowledge of the latter would be particularly valuable in aquatic ecological studies given the control water residence time exerts on such key metrics as dissolved organic carbon (DOC) and dissolved oxygen. Research that has addressed the implications of channel processes for IHS results include Bonell *et al.*'s (1998) observation of an initial rise in pre-event water at the start of each hydrograph pulse in the Babinda basins in northeastern Australia. This was attributed to a mechanism first suggested by Nolan and Hill (1990), whereby sudden upstream inputs of new water set up a flood wave composed of pre-event channel water that reaches downstream locations in advance of the translation of the subsequent event water. This process is distinct from the evaporative enrichment of streamflow during passage along the stream channel (Sklash *et al.*, 1976; Gremillion and Wanielista, 2000). Such enrichment would result in the overestimation of X in streamflow by shifting the δ in streamwater toward the groundwater isotopic signature, and should be considered when using IHS in situations where water residence times in the stream channel are appreciable (see below).

Incorporation of Temporal Variations in Event Water δ in IHS Studies

There are a range of approaches to treating temporal variations in event water δ and their influence on IHS results, including volume weighted means, incremental means, and incremental input intensity (McDonnell *et al.*, 1990). Nevertheless, each approach assumes that input water δ early in the event still exerts an influence on the stream water signature by the end of the event. This may be realistic when dealing with short-lived storms. However, these initial water inputs may be exported from the basin before the end of long duration episodes (e.g. entire snowmelt periods), and their δ should not have any influence on the stream water signature at that time. This problem has been addressed through the use of the unit hydrograph concept to estimate event water travel times in the basin (Joerin *et al.*, 2002), and a "runoff corrected event water" approach (Laudon *et al.*, 2002) that estimates input water δ at a given time on the basis of the amount of event water discharged from the basin prior to that time. Both approaches are consistent with Bonell *et al.*'s (1998) call for the use of event water δ that corresponds to the composition of lagged current water inputs. These and other approaches need to be tested across a greater range of basin and water input characteristics.

Incorporation of Temporal Variations in Pre-event Water δ in IHS Studies

Sklash and Farvolden (1979) provided an early example of the use of IHS results to argue for rapid groundwater contributions to streamflow. They assumed a temporally constant C_p signal in their hydrograph separations, while (as noted earlier) numerous studies have observed a shift in the isotopic signature of groundwater during the event. Rodhe (1987) attributed this to the recharge of isotopically different inputs and pre-event soil water and subsequent mixing, and to water table rise into the overlying unsaturated zone and conversion of isotopically different soil water to groundwater. It has also been previously noted that this temporally varying C_p signal has been used in IHS to solve X as the groundwater fraction of discharge. This assumes the instantaneous input of the sampled groundwater to the stream channel, which is physically unlikely. A promising approach to overcoming this problem is that of Weiler *et al.* (2003), who used a transfer function based on the instantaneous unit hydrograph to translate both event and pre-event water into the stream channel (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**). The transfer functions represent the travel time distributions of event and pre-event water within the basin. The method provides a means of assessing the spatial extent of groundwater that could potentially contribute to basin streamflow during the course of an event. This information in turn would provide valuable guidance in deciding what groundwater δ data obtained from spatially distributed samples in a basin should be used to estimate C_p in an IHS, and the associated groundwater travel times. We need to test whether this and other approaches to estimating pre-event water travel times from different points within a basin to the outflow are physically realistic. Such tests could be based on a combination of hydrometric techniques, use of reactive tracers to examine water flowpaths within basins, and the spatial pattern of point-scale estimates of groundwater residence times in a basin (e.g. Buttle *et al.*, 2001a).

The Fate of Event Water Falling on Saturated Areas

It is often assumed that near-stream saturated areas are connected to the stream channel such that event water falling onto these areas can actually reach the channel with little or no change in its isotopic signature (Kendall *et al.*, 2001). However, most overland flows may move only a few meters before infiltrating (Crayosky *et al.*, 1999), such that the reinfiltration and exfiltration of event water inputs to saturated areas may induce differences between the input signature and that of overland flow. Bonell *et al.* (1998) concluded that regions of lateral interflow and exfiltration in basins will inevitably promote the mixing

of event and pre-event water within saturation overland flow, thus modifying the isotopic content of this flow. Observations in the Sleepers River basin in Vermont in the United States suggest that the degree of mixing in the near-stream saturated zone varies down-valley, depending on the local groundwater exfiltration fluxes into this zone (McGlynn *et al.*, 1999). Srinivasan *et al.* (2002) documented the temporal and spatial dynamics of surface saturation areas and surface runoff source areas (generating infiltration excess and saturation excess overland flow) in relation to water table dynamics and slope runoff using innovative surface saturation sensors. The complexity of stormflow generation revealed at the slope scale challenges the simplistic view of stormflow generation based on isotopic, geochemical, and hydrometric observations made at a few locations within a basin and at the basin outfall. There is a need to couple similar sensor arrays with tracer studies to address this question.

The Type of Water Obtained from Standard Soil Water Samplers

DeWalle *et al.* (1988) found insignificant differences in δ from soil water samplers (assumed to favor matrix water) and from pan lysimeters (assumed to favor macropore water). However, Leaney *et al.* (1993) argued that suction lysimetry preferentially removes soil water from larger soil pores, thus biasing samples toward the δ of soil water infiltrating via relatively rapid pathways. This debate can be avoided by using throughflow trenches to sample soil water for IHS and GHS (Burns *et al.*, 2001), thus ensuring sampling of mobile slope water that may participate in stormflow generation. Whatever approach is adopted, the effect of sampling on end member definition and description should be noted in any IHS study.

Rationale for the Inclusion of Additional Runoff Components

Failure of the two-component IHS to describe flow components in a realistic manner may suggest contributions from one or more additional flow components (e.g. soil water). However, inclusion of these components in IHS needs to be supported by independent observations of the hydrological processes; otherwise they simply become mathematical corrections to apparently erroneous IHS results (Bonell *et al.*, 1998). The key issue is our ability to define *a priori* what these additional components might be and to sample them adequately. Uhlenbrook and Leibundgut (2002) developed a conceptual watershed model by defining three components contributing to streamflow in the Brugga basin in Germany. They found that direct runoff (with a mean residence time (MRT) of a few months), shallow groundwater (32 months MRT), and deep groundwater (MRT of 7.1 years) could be blended to give the combined stream signal. They then validated the model output with silica concentration data where each component could be characterized by uniquely different silica end member concentrations. This combination of conceptual watershed model development and runoff component characterization offers a way forward to identifying the minimum set of components that define a given hydrological system.

The Influence of Temporal and Spatial Variations in Hydrologic Linkages Between Landscape Units (Slopes – Riparian Zone – Stream) on a Basin's Isotopic and Chemical Response

We must quantify the processes affecting the spatial distribution of solute concentrations in source water throughout basins if we are to predict the hydrochemical response to such perturbations as forest harvesting and climate change (Welsch *et al.*, 2001). Processes influencing the slope – riparian zone linkage may range from simple mixing where a small volume of slope runoff is diluted by a larger volume of riparian storage during water transit (e.g. Burns *et al.*, 2001) to a complete biogeochemical resetting of solute signatures as slope runoff transits the riparian zone to the stream (e.g. Robson *et al.*, 1992; Hill, 1993). These processes in turn may exert a major control on a basin's hydrochemical response to inputs. Thus, McGlynn and McDonnell (2003) suggested that the relative timing of riparian and slope source contributions, and the connections and disconnections of dominant runoff contributing areas, were the first-order controls on stream DOC fluxes from a forested basin at Maimai in New Zealand. Further study of mixing and geochemical interactions in the riparian zone is required to address this issue.

The Scale Dependence of IHS Results

As noted earlier, there is no agreement on how partitioning between event and pre-event water in streamflow changes with basin scale (Figure 3). Some have noted an increase in X with basin size (Sklash *et al.*, 1976; Brown *et al.* (1999), attributed in the latter case to increased flux of shallow perched pre-event groundwater in larger basins during an intense storm. Conversely, Pearce (1990) suggested that saturated floodplains comprised a greater portion of larger basins at Maimai, thus enhancing the event water contribution to stormflow. McDonnell *et al.*'s (1999) work at Maimai showed that greater pre-event water contributions as one moved from the plot to the small basin scale, but a decrease in X as basin scale increased from $\sim$1 to $\sim$10 km^2. The pre-event fraction of stormflow also decreased with basin scale for forested basins in Vermont in the United States during snowmelt and rainfall, the exception being a small basin largely in pasture (Shanley *et al.*, 2002). Further work is needed to determine whether there is a relationship between basin morphology and the relative

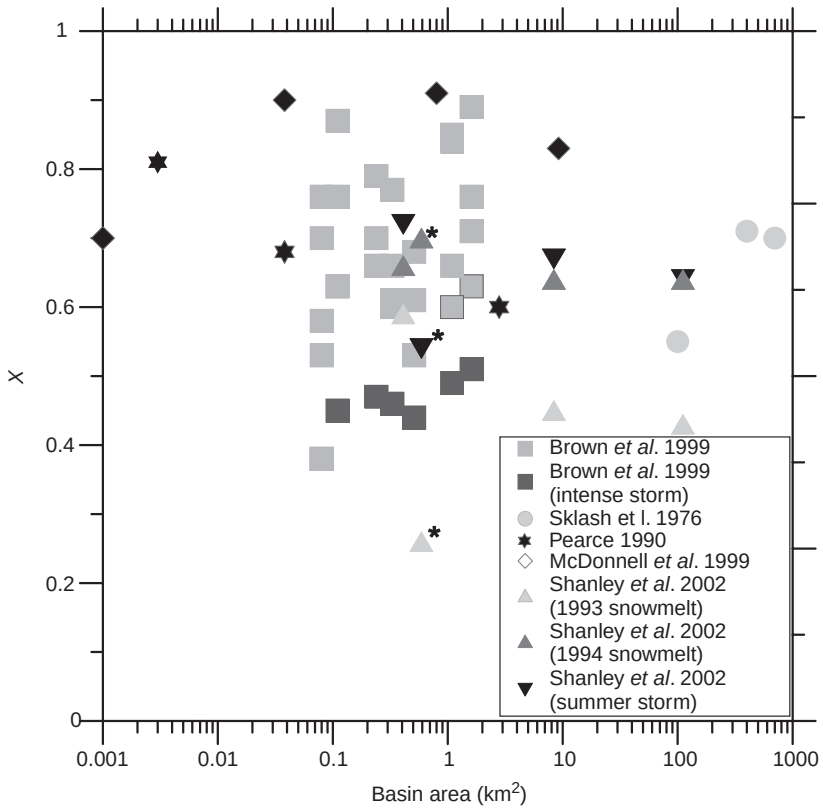


Figure 3 Reported pre-event water fractions in streamflow (X) versus basin area. The * associated with the Shanley *et al.* (2002) results indicates data from a small basin largely in pasture. The other basins from Shanley *et al.* (2002) were forested. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

partitioning of event and pre-event water in stormflow, and the degree to which inter-basin differences in land use, pedologic, and geologic characteristics might influence any scale dependence of IHS results. Another scale-related issue associated with the IHS approach was highlighted by Gremillion *et al.* (2000), who compared the standard IHS approach (that treats equations (2)–(4) as a steady-state model with negligible temporal changes in the volume and isotopic signature of channel storage) with nonsteady-state solutions. They found little difference in predicted pre-event water fractions from a basin in central Florida, but increasing divergence between predicted pre-event water fractions with increasing water residence times in the stream channel. This should be considered when applying IHS to large basins where flow time on slopes is small relative to the residence time of water routed along stream channels (Bras, 1990).

APPLICATION OF IHS TO ADDRESS FUTURE RESEARCH NEEDS

The problems noted above relate to improving our understanding of how hydrological processes govern the isotopic signature of water at the point, slope, and basin scales, as well as more clearly indicating when the IHS method can be appropriately applied to investigate hydrological questions. Despite these challenges, there are a number of outstanding research issues in hydrology that would benefit from the inclusion of environmental isotope tracer techniques, including the following:

Estimation of Water Residence Times at the Point, Slope, and Basin Scales Using Environmental Isotopic Tracers

There is abundant evidence from the literature demonstrating the important control that water residence time exerts on soil water, groundwater, and streamflow chemistry. Isotopic tracers provide a valuable means of estimating MRTs at various scales in basins, and recent examples of estimating MRTs in soil water, groundwater, and streamflow using environmental isotopes include Rodhe *et al.* (1996), McDonnell *et al.* (1999), and Buttle *et al.* (2001a). These MRTs may be used for such purposes as estimating the time required to observe basin response to treatment or disturbance (DeWalle *et al.*, 1997), the proportion of atmospheric NO_3 deposition in streamflow samples (Williard *et al.*, 2001), and the travel time distribution for rainfall

of a particular isotopic composition when conducting IHSs (Bonell *et al.*, 1998 – see above).

Identification of Process Thresholds

Partitioning between event and pre-event water in slope runoff and basin stormflow may have important implications for the transport of reactive substances to receiving water bodies. IHS studies in the same basin have often noted changes in the partitioning of event and pre-event water between events (e.g. Kendall *et al.*, 2001; Buttle *et al.*, 2001b) that have been attributed to shifts in the dominant hydrological processes governing slope runoff in response to changes in input characteristics and soil wetness. Greater understanding of the controls on these process transitions operating at the slope and basin scales is important for our ability to monitor and model basin hydrochemistry, and might be most effectively obtained through the use of controlled field experiments using environmental isotopes. These allow us to explore the role of specific hydrological processes and controlling factors by manipulating input rates and isotopic signatures, thus avoiding the complication of marked temporal variations in the δ of natural precipitation inputs and the resulting overlap between event and pre-event water signatures that precludes IHS (Turton *et al.*, 1995; Collins *et al.*, 2000). Such studies range in the degree of experimental control from the Coos Bay artificial irrigation of a deforested slope in the Coast Range in Oregon in the United States (Anderson *et al.*, 1997; Montgomery *et al.*, 1997; Torres *et al.*, 1998; Anderson and Dietrich, 2001), through the Gårdsjön-covered basin experiments in south-western Sweden (Nyberg, 1995; Rodhe *et al.*, 1996; Lange *et al.*, 1996), to the Hydrohill 490 m^2 artificial basin experiment in China (Kendall *et al.*, 2001).

Identification of Hydrological Response to Land Use Change

Hydrological response to land use change has generally been characterized using such metrics as soil moisture content, evaporation fluxes, and hydrograph properties (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**; **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3** and **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**). However, shifts in the relative proportion of event and pre-event water before and after disturbance or land use alteration also provide a useful detection tool, offering the potential to quantify changes in hydrograph *composition* following disturbance. Vertical profiles of MRTs estimated from soil water $\delta^{18}O$ time series during snowmelt in central Ontario indicated that forest harvesting restricted deep infiltration of inputs and diverted a portion of incoming event water laterally downslope, with important consequences for the quantity and quality of slope runoff reaching receiving waters (Murray and Buttle, in press). A combined IHS–GHS approach to studying water flowpaths in forested and deforested (pasture) basins in French Guiana showed peak flow from the forest basin was largely via subsurface flow, whereas flow through the superficial soil layers dominated peak streamflow in the pasture basin (Bariac *et al.*, 1995). The two-component mass balance approach in equations (2)–(4) also assists studies of the effects of forest road construction on water rerouting at the basin scale. Ditch flow can be separated into direct road runoff (event water) and intercepted subsurface flow from the cut back (pre-event water), enabling these relative inputs to be quantified at specific cross drain and road culvert sections. Ziegler *et al.* (2001) demonstrated the value of this approach in studies of forest road runoff in the humid tropics of Thailand, and Luce (2002) has called for greater use of hydrograph separation in land use change studies involving roads. Urban and suburban development may also result in a shift in stormflow pathways from largely subsurface during the predevelopment phase to overland and channelized flow after development (Buttle, 1990). IHS provides a useful tool for identifying water sources in urban and suburban basins, particularly if the link between event water generation and infiltration excess overland flow from modified surface cover can be made (Halldin *et al.*, 1990; Buttle *et al.*, 1995). For example, Gremillion *et al.* (2000) noted greater event water contributions to stormflow from a suburban sub-basin of the Econlockhatchee River in central Florida, attributed to an increased proportion of surface runoff in the storm hydrograph. This change in water flowpaths to the river may alter groundwater flow through riparian zones, with implications for river water quality and riparian zone ecology (Gremillion *et al.*, 2000). Information on event and pre-event water partitioning of stormflow also assists in interpreting and modeling basin export of surface-applied chemicals, such as radionuclides deposited in fallout from the Chernobyl accident (Halldin *et al.*, 1990) and deicing salts (Buttle *et al.*, 1992).

Integration of Isotopic and Geochemical Tracers and Hydrometric Techniques with Greater Consideration of Topographic Properties

Increased availability of digital topographic information can help estimate where most of the hydrological/hydrochemical action is going to take place in a basin (Kendall *et al.*, 1999; Welsch *et al.*, 2001), and can be used to design field experiments to maximize the amount of information obtained. For example, topographic indices such as the $\ln(\alpha/\tan\beta)$ index of Beven and Kirkby (1979) have been used to estimate the depth to groundwater (Moore

and Thompson, 1996; Seibert *et al.*, 1997) and to interpret spatial variations in the $\delta^{18}O$ of groundwater (Rodhe *et al.*, 1996). This approach should be expanded to include the use of different types of topographic data (e.g. surface topography vs. bedrock topography – Freer *et al.*, 1997) and different topographic indices (e.g. Barling *et al.*, 1994; Chaplot *et al.*, 2000).

Explicit Integration of Models into Our Study Designs

Hydrologists continue to call for greater integration between field and modeling studies (Dunne, 2001). This integration would encourage the adoption of sampling strategies that generate environmental isotope data that could be used to test a range of models (Hooper, 2001). It would also allow us to extend the model calibration process (traditionally based on "hard" data such as the streamflow hydrograph) to incorporate tracer information and hydrograph separation results as "soft" data in a multicriteria model calibration exercise (Seibert and McDonnell, 2002).

CONCLUSIONS

The initial enthusiasm with which some hydrologists greeted the advent of the IHS approach has since been tempered by greater recognition of its assumptions and limitations, and the need to integrate the use of IHS with other tracing and hydrometric methods. There are numerous challenges facing the application of the IHS approach to address hydrological questions, and hydrologists must be alert to these when interpreting the results of tracing studies using environmental isotopes. Nevertheless, isotope hydrograph separation provides a powerful addition to the hydrologist's toolkit of approaches available to examine water age, origin, and pathway during its movement through drainage basins to receiving water bodies.

REFERENCES

Anderson S.P. and Dietrich W.E. (2001) Chemical weathering and runoff chemistry in a steep headwater catchment. *Hydrological Processes*, **15**, 1791–1815.

Anderson S.P., Dietrich W.E., Montgomery D.R., Torres R., Conrad M.E. and Loague K. (1997) Subsurface flow paths in a steep, unchanneled catchment. *Water Resources Research*, **33**, 2637–2653.

Bariac T., Millet A., Ladouche B., Mathieu R., Grimaldi C., Grimaldi M., Hubert P., Molicova H., Bruckler L., Bertuzzi P. *et al.* (1995) Stream hydrograph separation on two small Guianese catchments. *Tracer Technologies for Hydrological Systems*, IAHS Publication 229, International Association of Hydrological Sciences: Wallingford, pp. 193–209.

Barling R.D., Moore I.D. and Grayson R.B. (1994) A quasi-dynamic wetness index for characterizing the spatial distribution of zones of surface saturation and soil water content. *Water Resources Research*, **30**, 1029–1044.

Bazemore D.E., Eshleman K.N. and Hollenbeck K.J. (1994) The role of soil water in stormflow generation in a forested headwater catchment: synthesis of natural tracer and hydrometric evidence. *Journal of Hydrology*, **162**, 47–75.

Bencala K.E. (2000) Hyporheic zone hydrological processes. *Hydrological Processes*, **14**, 2797–2708.

Beven K.J. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Bonell M., Barnes C.J., Grant C.R., Howard A. and Burns J. (1998) High rainfall, response-dominated catchments: a comparative study of experiments in tropical northeast Queensland with temperate New Zealand. In *Isotope Tracers in Catchment Hydrology*, Kendall C. and McDonnell J.J. (Eds.), Elsevier: Amsterdam, pp. 347–390.

Bonell M., Pearce A.J. and Stewart M.K. (1990) The identification of runoff-production mechanisms using environmental isotopes in a tussock grassland catchment, eastern Otago, New Zealand. *Hydrological Processes*, **4**, 15–34.

Bras R.L. (1990) *Hydrology: An Introduction to Hydrologic Science*, Addison Wesley: Reading.

Brodersen C., Pohl S., Lindenlaub M., Leibundgut C. and v. Wilpert K. (2000) Influence of vegetation structure on isotope content of throughfall and soil water. *Hydrological Processes*, **14**, 1439–1448.

Brown V.A., McDonnell J.J., Burns D.A. and Kendall C. (1999) The role of event water, a rapid shallow flow component, and catchment size in summer stormflow. *Journal of Hydrology*, **217**, 171–190.

Bruijnzeel S. (1990) *Hydrology of Moist Tropical Forests and Effects of Conversion: A State of Knowledge Review*, UNESCO International Hydrological Program: Paris.

Burns D.A. (2002) Stormflow-hydrograph separation based on isotopes: the thrill is gone – what's next? *Hydrological Processes*, **16**, 1515–1517.

Burns D.A. and McDonnell J.J. (1998) Effects of a beaver pond on runoff processes: comparison of two headwater catchments. *Journal of Hydrology*, **205**, 248–264.

Burns D.A., McDonnell J.J., Hooper R.P., Peters N.E., Freer J.E., Kendall C. and Beven K. (2001) Quantifying contributions to storm runoff through end-member mixing analysis and hydrologic measurements at the Panola Mountain Research Watershed (Georgia, USA). *Hydrological Processes*, **15**, 1903–1924.

Buttle J.M. (1990) Effects of suburbanization upon snowmelt runoff. *Hydrological Sciences Journal*, **35**, 285–302.

Buttle J.M. (1994) Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins. *Progress in Physical Geography*, **18**, 16–41.

Buttle J.M., Hazlett P.W., Murray C.D., Creed I.F., Jeffries D.S. and Semkin R. (2001a) Prediction of groundwater characteristics in forested and harvested basins during spring snowmelt using a topographic index. *Hydrological Processes*, **15**, 3389–3407.

Buttle J.M., Lister S.W. and Hill A.R. (2001b) Controls on runoff components on a forested slope and implications for N transport. *Hydrological Processes*, **15**, 1065–1070.

Buttle J.M. and Peters D.L. (1997) Inferring hydrological processes in a temperate basin using isotopic and geochemical hydrograph separation: a re-evaluation. *Hydrological Processes*, **11**, 557–573.

Buttle J.M. and Sami K. (1992) Testing the groundwater ridging hypothesis of streamflow generation during snowmelt. *Journal of Hydrology*, **135**, 53–72.

Buttle J.M., Taylor C.H. and Vonk A.M. (1992) Environmental isotope hydrograph separation during snowmelt in a suburban catchment. *Proceedings of the Eastern Snow Conference*, **49**, 1–12.

Buttle J.M., Vonk A.M. and Taylor C.H. (1995) Applicability of isotope hydrograph separation in a suburban basin during snowmelt. *Hydrological Processes*, **9**, 197–211.

Chaplot V., Walter C. and Curmi P. (2000) Improving soil hydromorphy prediction according to DEM resolution and available pedological data. *Geoderma*, **97**, 405–422.

Clark I.D. and Fritz P. (1997) *Environmental Isotopes in Hydrogeology*, CRC Press: Boca Raton.

Collins R., Jenkins A. and Harrow M. (2000) The contribution of old and new water to a storm hydrograph determined by tracer addition to a whole catchment. *Hydrological Processes*, **14**, 701–711.

Cooper L.W., Olsen C.R., Solomon D.K., Larsen I.L., Cook R.B. and Grebmeier J.M. (1991) Stable isotopes of oxygen and natural and fallout radionuclides used for tracing runoff during snowmelt in an Arctic watershed. *Water Resources Research*, **27**, 2171–2179.

Crayosky T.W., DeWalle D.R., Seybert T.A. and Johnson T.E. (1999) Channel precipitation dynamics in a forested Pennsylvania headwater catchment (USA). *Hydrological Processes*, **13**, 1303–1314.

Dansgaard W. (1964) Stable isotopes in precipitation. *Tellus*, **16**, 436–468.

DeWalle D.R., Edwards P.J., Swistock B.R., Aravena R. and Drimmie R.J. (1997) Seasonal isotope hydrology of three Appalachian forest catchments. *Hydrological Processes*, **11**, 1895–1906.

DeWalle D.R. and Pionke H.B. (1994) Streamflow generation on a small agricultural catchment during autumn recharge: II. Stormflow periods. *Journal of Hydrology*, **163**, 23–42.

DeWalle D.R. and Swistock B.R. (1994) Differences in oxygen-18 content of throughfall and rainfall in hardwood and coniferous forests. *Hydrological Processes*, **8**, 75–82.

DeWalle D.R., Swistock B.R. and Sharpe W.E. (1988) Three component tracer model for stormflow on a small Appalachian forested catchment. *Journal of Hydrology*, **104**, 301–310.

Drever J.I. (1988) *The Geochemistry of Natural Waters*, Prentice-Hall: Englewood Cliffs.

Dunne T. (2001) Problems in measuring and modeling the influence of forest management on hydrologic and geomorphic processes. In *Land Use and Watersheds: Human Influence on Hydrology and Geomorphology in Urban and Forest Areas*, Wigmosta M.S. and Burges S.J. (Eds.), American Geophysical Union: Washington, pp. 77–83.

Elsenbeer H. (2001) Hydrologic flowpaths in tropical rainforest soilscapes – a review. *Hydrological Processes*, **15**, 1751–1759.

Elsenbeer H., Lorieri D. and Bonell M. (1995) Mixing model approaches to estimate stormflow in an overland flow dominated rainforest catchment. *Water Resources Research*, **31**, 2267–2278.

Freer J., McDonnell J., Beven K.J., Brammer D., Burns D., Hooper R.P. and Kendall C. (1997) Topographic controls on subsurface storm flow at the hillslope scale for two hydrologically distinct small catchments. *Hydrological Processes*, **11**, 1347–1352.

Genereux D.P. (1998) Quantifying uncertainty in tracer-based hydrograph separations. *Water Resources Research*, **34**, 915–919.

Genereux D.P. and Hooper R.P. (1998) Oxygen and hydrogen isotopes in rainfall-runoff studies. In *Isotope Tracers in Catchment Hydrology*, Kendall C. and McDonnell J.J. (Eds.), Elsevier: Amsterdam, pp. 319–346.

Gremillion P., Gonyeau A. and Wanielista M. (2000) Application of alternative hydrograph separation models to detect changes in flow paths in a watershed undergoing urban development. *Hydrological Processes*, **14**, 1485–1501.

Gremillion P. and Wanielista M. (2000) Effects of evaporative enrichment on the stable isotope hydrology of a central Florida (USA) river. *Hydrological Processes*, **14**, 1465–1484.

Halldin S., Rodhe A. and Bjurman B. (1990) Urban storm water transport and wash-off of caesium-137 after the Chernobyl accident. *Water, Air, and Soil Pollution*, **49**, 139–158.

Hill A.R. (1993) Base cation chemistry of storm runoff in a forested headwater wetland. *Water Resources Research*, **29**, 2663–2673.

Hill A.R. and Waddington J.M. (1993) Analysis of storm run-off sources using oxygen-18 in a headwater swamp. *Hydrological Processes*, **7**, 305–316.

Hinton M.J., Schiff S.L. and English M.C. (1994) Examining the contributions of glacial till water to storm runoff using two- and three-component hydrograph separations. *Water Resources Research*, **30**, 983–993.

Hooper R.P. (2001) Applying the scientific method to small catchment studies: a review of the Panola Mountain experience. *Hydrological Processes*, **15**, 2039–2050.

Hooper R.P. and Shoemaker C.A. (1986) A comparison of chemical and isotopic hydrograph separation. *Water Resources Research*, **22**, 1444–1454.

Ingraham N.L. (1998) Isotopic variations in precipitation. In *Isotope Tracers in Catchment Hydrology*, Kendall C. and McDonnell J.J. (Eds.), Elsevier: Amsterdam, pp. 87–118.

Ingraham N.L. and Taylor B.E. (1991) Light stable isotope systematics of large-scale hydrologic regimes in California and Nevada. *Water Resources Research*, **27**, 77–90.

International Atomic Energy Agency/World Meteorological Organization (2001) *Global Network of Isotopes in Precipitation. The GNIP Database*, Accessible at `http://isohis.iaea.org`.

Joerin C., Beven K.J., Iorgulescu I. and Musy A. (2002) Uncertainty in hydrograph separations based on geochemical mixing models. *Journal of Hydrology*, **255**, 90–106.

Kendall C. and Caldwell E.A. (1998) Fundamentals of isotope geochemistry. In *Isotope Tracers in Catchment Hydrology*, Kendall C. and McDonnell J.J. (Eds.), Elsevier: Amsterdam, pp. 51–86.

Kendall C., McDonnell J.J. and Gu W. (2001) A look inside 'black box' hydrograph separation models: a study at the Hydrohill catchment. *Hydrological Processes*, **15**, 1877–1902.

Kendall K.A., Shanley J.B. and McDonnell J.J. (1999) A hydrometric and geochemical approach to test the transmissivity feedback hypothesis during snowmelt. *Journal of Hydrology*, **219**, 188–215.

Kennedy V.C., Kendall C., Zellweger G.W., Wyerman T.A. and Avanzino R.J. (1986) Determination of the components of stormflow using water chemistry and environmental isotopes, Mattole river basin, California. *Journal of Hydrology*, **84**, 107–140.

Lange H., Lisheid G., Hoch R. and Hauhs M. (1996) Water flow paths and residence times in a small headwater catchment at GÅrdsjön, Sweden, during steady state storm flow conditions. *Water Resources Research*, **32**, 1689–1698.

Laudon H., Hemond H.F., Krouse R. and Bishop K.H. (2002) Oxygen 18 fractionation during snowmelt: Implications for spring flood hydrograph separation. *Water Resources Research*, **38**, 1258, doi: 10.1029/2002/WR0015010.

Leaney F.W., Smettem K.R.J. and Chittleborough D.J. (1993) Estimating the contribution of preferential flow to subsurface runoff from a hillslope using deuterium and chloride. *Journal of Hydrology*, **147**, 83–103.

Luce C.H. (2002) Hydrological processes and pathways affected by forest roads: what do we still need to learn? *Hydrological Processes*, **16**, 2901–2904.

McDonnell J.J., Bonell M., Stewart M.K. and Pearce A.J. (1990) Deuterium variations in storm rainfall: implications for hydrograph separation. *Water Resources Research*, **26**, 455–458.

McDonnell J.J., Rowe L. and Stewart M. (1999) *A Combined Tracer-Hydrometric Approach to Assessing the Effects of Catchment Scale on Water Flowpaths, Source and Age*, International Association of Hydrological Sciences, Publication 258, IAHS, pp. 265–274.

McDonnell J.J., Stewart M.K. and Owens I.F. (1991a) Effect of catchment-scale mixing on stream isotopic response. *Water Resources Research*, **27**, 3065–3073.

McDonnell J.J., Stewart M.K. and Owens I.F. (1991b) A case study of shallow flow paths in a steep zero-order basin, a physical-chemical-isotopic analysis. *Water Resources Bulletin*, **27**, 679–685.

McGlynn B. and McDonnell J.J. (2003) The role of discrete landscape units in controlling catchment dissolved organic carbon dynamics. *Water Resources Research*, **39**, doi:10.1029/2002WR001525.

McGlynn B.L., McDonnell J.J., Shanley J.B. and Kendall C. (1999) Riparian zone flowpath dynamics during snowmelt in a small headwater catchment. *Journal of Hydrology*, **222**, 75–92.

Metcalfe R.A. and Buttle J.M. (2001) Soil partitioning and surface store controls on spring runoff from a boreal forest peatland basin in north-central Manitoba, Canada. *Hydrological Processes*, **15**, 2305–2324.

Montgomery D.R., Dietrich W.E., Torres R., Anderson S.P., Heffner J.T. and Loague K. (1997) Hydrologic response of a steep, unchanneled valley to natural and applied rainfall. *Water Resources Research*, **33**, 91–109.

Moore R.D. and Thompson J.C. (1996) Are water table variations in a shallow forest soil consistent with the TOPMODEL concept? *Water Research Research*, **32**, 663–669.

Murray C.D. and Buttle J.M. (In press) Infiltration and soil water mixing on forested and harvested slopes during spring snowmelt, Turkey Lakes Watershed, central Ontario. *Journal of Hydrology*.

Nolan K.M. and Hill B.R. (1990) Storm-runoff generation in the Permanente Creek drainage basin, west central California – an example of flood-wave effects on runoff composition. *Journal of Hydrology*, **113**, 343–367.

Nyberg L. (1995) Water flow path interactions with soil hydraulic properties in till soil at GÅrdsjön, Sweden. *Journal of Hydrology*, **170**, 255–275.

Ogunkoya O.O. and Jenkins A. (1993) Analysis of storm hydrograph and flow pathways using a three-component hydrograph separation model. *Journal of Hydrology*, **142**, 71–88.

Pearce A.J. (1990) Streamflow generation processes: an Austral view. *Water Resources Research*, **26**, 3037–3047.

Peters D.L., Buttle J.M., Taylor C.H. and LaZerte B.D. (1995) Runoff production in a forested, shallow soil, Canadian Shield basin. *Water Resources Research*, **31**, 1291–1304.

Rice K. and Hornberger G. (1998) Comparison of hydrochemical tracers to estimate source contributions to peak flow in a small, forested headwater catchment. *Water Resources Research*, **34**, 1755–1766.

Richey D.G., McDonnell J.J., Erbe M. and Hurd T. (1998) A critical appraisal of published chemical and isotopic hydrograph separations from New Zealand, North America and Europe. *Journal of Hydrology (New Zealand)*, **37**, 95–111.

Robson A., Beven K. and Neal C. (1992) Towards identifying sources of subsurface flow: a comparison of components identified by a physically based runoff model and those determined by chemical mixing techniques. *Hydrological Processes*, **6**, 199–214.

Rodhe, A. (1987) *The Origin of Streamwater traced by Oxygen-18*, PhD. Thesis, Department of Physical Geography, Division of Hydrology, Report Series A 41, Uppsala University, p. 260, + appendices.

Rodhe A., Nyberg L. and Bishop K. (1996) Transit times for water in a small till catchment from a step shift in the oxygen 18 content of the water input. *Water Resources Research*, **32**, 3497–3511.

Saxena R.K. (1986) Estimation of canopy reservoir capacity and oxygen-18 fractionation in throughfall in a pine forest. *Nordic Hydrology*, **17**, 251–260.

Seibert J., Bishop K.H. and Nyberg L. (1997) A test of TOPMODEL's ability to predict spatially distributed groundwater levels. *Hydrological Processes*, **11**, 1131–1144.

Seibert J. and McDonnell J.J. (2002) On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multi-criteria model calibration. *Water Resources Research*, **38**, 1241, doi:10.1029/2001WR000978.

Shanley J., Kendall C., Smith T., Wolock D. and McDonnell J.J. (2002) Controls on old and new water contributions to stream flow at some nested catchments in Vermont, USA. *Hydrological Processes*, **16**, 589–610.

Sklash M.G. (1990) Environmental isotope studies of storm and snowmelt runoff generation. In *Process Studies in Hillslope*

Hydrology, Anderson M.G. and Burt T.P. (Eds.), Wiley: Chichester, pp. 401–435.

Sklash M.G. and Farvolden R.N. (1979) The role of groundwater in storm runoff. *Journal of Hydrology*, **43**, 45–65.

Sklash M.G., Farvolden R.N. and Fritz P. (1976) A conceptual model of watershed response to rainfall, developed through the use of oxygen-18 as a natural tracer. *Canadian Journal of Earth Sciences*, **13**, 271–283.

Srinivasan M.S., Gburek W.J. and Hamlett J.M. (2002) Dynamics of stormflow generation on a hillslope – a hillslope-scale field study in east-central Pennsylvania, USA. *Hydrological Processes*, **16**, 649–665.

Torres R., Dietrich W.E., Montgomery D.R., Anderson S.P. and Loague K. (1998) Unsaturated zone processes and the hydrologic response of a steep, unchanneled catchment. *Water Resources Research*, **34**, 1865–1879.

Turton D.J., Barnes D.R. and de Jesus Navar J. (1995) Old and new water in subsurface flow from a forest soil block. *Journal of Environmental Quality*, **24**, 139–146.

Uhlenbrook S. and Leibundgut C. (2002) Process-oriented catchment modelling and multiple-response validation. *Hydrological Processes*, **16**, 423–440.

Unnikrishna P.V., McDonnell J.J. and Kendall C. (2002) Isotope variations in a Sierra Nevada snowpack and their relation to meltwater. *Journal of Hydrology*, **260**, 38–57.

Weiler M., McGlynn B., McGuire K. and McDonnell J. (2003) How does rainfall become runoff? A combined tracer and runoff transfer function approach. *Water Resources Research*, **39**, 1315–1327.

Wels C., Cornett R.J. and LaZerte B.D. (1991) Hydrograph separation: a comparison of geochemical and isotopic tracers. *Journal of Hydrology*, **122**, 253–274.

Welsch D.L., Kroll C.N., McDonnell J.J. and Burns D.A. (2001) Topographic controls on the chemistry of subsurface stormflow. *Hydrological Processes*, **15**, 1925–1938.

Williard K.W.J., DeWalle D.R., Edwards P.J. and Sharpe W.E. (2001) ^{18}O isotopic separation of stream nitrate sources in mid-Appalachian forested watersheds. *Journal of Hydrology*, **252**, 174–188.

Ziegler A., Giambelluca T., Sutherland R.A., Vana T. and Nullet M. (2001) Horton overland flow contribution to runoff on unpaved mountain roads, a case study in Northern Thailand. *Hydrological Processes*, **15**, 3203–3208.

117: Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development

THEODORE A ENDRENY

Program in Hydrological Systems Science & Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY, US

Urban and suburban development alters watershed imperviousness and connectivity, resulting in alterations to the hydrologic and hydraulic runoff response. Stormwater conveyance devices have been incorporated into most development to address the goal of floodwater removal for protection of life and property. Urban and suburban development laws generally consider the stormwater drainage system as part of the basic infrastructure and leave right of way and space to locate the component parts. Drainage systems are designed to protect life, enable vehicular access, extend pavement life cycle, control the volume and velocity of stormwater runoff along curbs and gutters, and convey stormwater runoff to natural or human constructed drainage ways and receiving waters. Engineering equations have been developed to design construction for optimal runoff behavior, such as timing, volume, and rate, into, through, and from conveyance devices. Design equations are also utilized for assessing the performance of constructed environments, to ascertain whether stormwater devices need rehabilitation or expansion. New frontiers in stormwater include model spatially distributed development that incorporates high-density elevation maps, radar rainfall, and details of remotely sensed infrastructure as well as ecological restoration simulation, including groundwater and river remediation for wildlife and recreation.

HYDROLOGICAL FEATURES OF DEVELOPED AREAS

INTRODUCTION

Urban areas are distinct from suburban based on the type and density of development, where urban typically exceeds 50% impervious cover, and includes industrial uses in addition to commercial and residential cover. There is a continuum, however, of hydrological impacts due to these development schemes. Estimates of urban runoff timing, depths, and volumes are often initially addressed by civil and water resources engineers as part of the design for development of the urban and suburban area. Design objectives for runoff are typically to control against flooding, standing water, and device scour (ASCE, 1993; Mays, 1999; Chin, 2000) within the constructed environment, focusing primarily on storm flows, and have been expanded to simultaneously address water quality and restoration concerns (WEF/ASCE, 1998; Chin, 2000, *see* **Chapter 97, Urban Water Quality, Volume 3** by Bryan Ellis for more on urban water quality and **Chapter 191, Environmental Flows: Managing Hydrological Environments, Volume 5** for more on urban restoration). According to Urbonas and Roesner (1993), urban and suburban development laws consider the stormwater drainage system as part of the basic infrastructure and leave right of way and space to locate the component parts, such that drainage systems can (i) remove stormwater to enable vehicular access and extend pavement life cycle, (ii) control the volume and velocity of stormwater runoff along curbs and gutters to reduce human hazards and pavement damage, and (iii) convey stormwater runoff to natural or man-made drainage ways and receiving waters. Design equations are also utilized for assessing the performance of constructed environments, to ascertain whether stormwater devices need rehabilitation or expansion. Bedient and Huber (2002) define three separate, but incrementing tasks faced by the urban water resources engineer, as prediction of peak flows, runoff volumes, and complete hydrographs.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Engineering design formula for estimating the timing and magnitude of runoff includes hydrologic and hydraulic approaches. Hydrologic design includes formula from both empirical design, such as Curve Number (CN) based rainfall abstractions, as well as those derived through reduction of the phenomena via the Reynolds Transport Theorem (see Chin, 2000) in a Eulerian control volume analysis of mass, momentum, and energy. Hydraulic design is comprised primarily of the latter approach, but also includes a Lagrangian analysis via differentiation of the changes in mass, momentum, and energy along the entire system. In summary, urban environments were planned to uphold properties of runoff timing and magnitude that generally complied with engineering design equations.

The as-built condition of constructed environments and the actual machinations of the runoff phenomena within the urban and suburban neighborhoods contain elements that deviate from the design. A combination of hydrologic- and hydraulic-based equations is common in practice, weighted more toward the more readily solved and stable hydrologic approach (ASCE, 1993; Maidment, 1993; AASHTO, 1999; Mays, 1999; Chin, 2000). Designed or otherwise, urbanization has been demonstrated to reduce the degree of infiltration and increase the volume of stormwater runoff by changing slope, form, or cover, alter the amount of depression storage by regrading terrain, reduce the amount of evapotranspiration by removing vegetation and reducing infiltration, and reduce the travel time to a receiving waterbody by increasing the efficiency of surface and subsurface drainage networks (ASCE, 1993; WEF/ASCE, 1998; Butler and Davies, 2000).

Runoff processes considered in this article are primarily storm flow across and within constructed conveyance systems, including stormwater management facilities such as gutters, culverts, stormsewers, swales, infiltration trenches, bioretention filters, wet and dry detention basins, and retention ponds. The chapter focuses on the most common engineering design equations used in estimating runoff across the constructed environment, and also addresses the utilization of current research methods that enhance the analysis of runoff processes.

Density of Residents

Urban and suburban areas and their constructed runoff devices are built within, and surrounded by, natural hydrological regimes, making delineation of the developed and natural runoff phenomena often difficult. Delineation is important because it provides a control volume within which the engineering runoff equations are applied, and through which the external forcing of precipitation, evaporation, and water supply and removal are considered. Population density is the traditional means by which government census bureaus define an urban area, and these numbers vary between nations. In 2000, urban areas for the US Census Bureau were defined as populations of at least 390 people/km^2, while for the Canadian Census they were defined as concentrated populations of at least 1000 with no fewer than 400 people/km^2. The US Census defines metropolitan statistical areas (MSA) as concentrated populations of at least 10 000, and when 50% or more of a county resides within an MSA, the entire county is classified as an MSA. All other land is defined as rural, with limited distinction of suburban features as a lower density settlement along the fringe of the urban area.

Table 1 Category of suburban density based on dwellings and effective impervious area (Modified from Dinicola, 1989)

Classification	Dwellings hectare^{-1}	Effective impervious area (%)
Low density	<2.5	<4
Medium density	2.5–10.0	4–10
Suburban density	10.0–20.0	10–24
High density	20.0–40.0	24–48
Urban residential	>40.0	>48%

Constructed development intensity as an urban metric might be useful in stratifying the impact of construction on the runoff processes. One definition of development intensity considers dwellings per unit area (see Table 1), where the area in the denominator might be the designed lot size or the local watershed size (Dinicola, 1989), and is structured discretely from low to high. Ideally, in assessing runoff alterations, the location of the development relative and its impact on recharge, discharge, or conveyance should be considered in addition to its footprint size.

Imperviousness and Constructed Land Cover

Another definition of intensity is to consider impervious area (Schueler, 1994), proposed by the Water Environment Federation (WEF) and American Society of Civil Engineers (ASCE), as a single unifying measure for the effect of urbanization on watershed hydrology (WEF/ASCE, 1998). Impervious area is structured continuously from 0 to 100%. The total or effective impervious area estimate may be based on *in situ* or remote observation (Lunetta and Elvidge, 1998; Doyle, 2001; Steitz and Chandler, 2001), or on empirical functions that assume a road, house, sidewalk, and driveway area for each dwelling or resident.

Imperviousness of the urban area, as surveyed in 11 residential areas in Olympia, Washington, comprises between 63 and 70% transportation-related structures and 30 to 37% occupation-related structures. In regard to water quality, Schueler (1994) has delineated 10% impervious area as the extent where watersheds move from sensitive to impaired, and 25% impervious area as the extent where they become nonsupporting (CWP, 1998). Prior to advanced mapping sources, Brater (1968) attempted to identify hydrologically

significant impermeable area (HSIA) ratio for drainage basins, ranging in size from 23 to 475 km^2, in various stages of urbanization around Detroit, Michigan, by dividing the observed total volume of discharge by the computed effective precipitation volume across the entire watershed. The HSIA ratio is then multiplied by the total basin area, which gives the impermeable area of watershed contributing runoff. In an era when impermeable surfaces were difficult to map, this technique provided a first estimate.

Constructed area stormwater conveyance extent is an important urban attribute not provided by either lot size or impervious area estimates. Leopold (1968) used the extent of stormwater conveyance together with impervious area to estimate the ratio of post- to predevelopment runoff peaks. Rather than just impervious area, more emphasis is now placed on estimates of impervious areas directly connected to the stormwater system and the receiving water, referred to as *directly connected impervious area* (DCIA) or as *the effective impervious area* (CWP, 1998; WEF/ASCE, 1998). DCIA estimates provide the basis for adjusting Curve Number values central in the simple engineering runoff models such as the TR-55 approach (Chin, 2000), which is explained more at the end of this article. Urban stormwater conveyance devices, along with other water resource infrastructure such as water delivery and waste removal, are typically associated with a minimum density of 3500 people/km^2 (McGhee, 1991). Dinicola (1989) has shown that dwellings per unit area can be used to predict impervious cover, but estimating the impervious cover from population numbers is a moving target; the US Census Bureau has reported that the number of residents per dwelling has diminished from 3.1 in 1970 to 2.7 in 2002 as the average floor size per dwelling increased from 140 to 205 m^2. The US Environmental Protection Agency estimates that approximately 6 m of stormwater sewer is needed per person (USEPA, 2001b).

Data products used in estimating development intensity and change primarily originated from airborne and satellite imagery (Lunetta and Elvidge, 1998). USGS and USEPA National Land Cover Database (NLCD) maps, derived from LandSat Thematic Mapper Multi-Resolution Land Cover data (see Table 2), is used to simply denote low-intensity residential as class 21, and high-intensity residential as class 22, with transportation, industrial, and commercial lumped into class 23. These NLCD values are for approximately 30×30 m pixels (picture elements), and per area assessment can be aggregated for an entire watershed. NLCD 2001 uses Landsat 7 ETM+ and high-resolution imagery to assess the subgrid spectral signals within the 30×30 m raw data, providing vegetation cover and impervious cover estimates as a second layer in the variable attribute data files (Huang *et al.*, 2001; Yang *et al.*, 2003). Other projects are not as in-depth. 4-m horizontal resolution IKONOS satellite imagery has been privately used to delineate road networks (Gibson, 2003), and advances in aerial orthoimagery are now used to map entire states at submeter resolution (New York State GIS Clearinghouse, 2003) and provide oblique angle views of city landscapes in places like Los Angeles, CA, and Syracuse, NY (Pictometry, 2003). Advanced remote sensing work in Switzerland by Fankhauser (1999) has also demonstrated detection opportunities with orthoimagery.

Table 2 Category of suburban density for 900 m^2 pixels based on constructed material and vegetation coverage

Classification	Constructed material (%)	Vegetation (%)
Low intensity	30–80	20–70
High intensity	80–100	<20

Drainage Infrastructure of Urban Development

Features of the landscape that distinguish an urban center and suburban area, beyond higher human density and the general class of impervious area, include transportation networks, residential, public, commercial, and industrial buildings, power supply grids, water supply facilities and supply lines from groundwater and/or surface reservoirs, wastewater sewers, facilities for removal of gray and black wastes from the buildings, and receiving water modification to coastal, lake, wetland, and river areas. Receiving water impacts are a significant impact of urban development, with rivers comprising the largest category of receiving water (WEF/ASCE, 1998). Waterside development sites are common given that the services of water supply, waste removal, transportation, and hydropower, and activities of farming and fishing, which established and supported the urban area and its suburban surroundings, either required or benefited from proximity to water resources (Mays, 1999; Butler and Davies, 2000).

Graff (1977) noted that a key hydrological feature of urban development are road networks that add numerous artificial surface channels, separate from stormsewers, designed to efficiently collect and route stormwater such that the lag time and kurtosis of storm hydrographs shift to the left and become flashier. Not only do the gridded networks redirect runoff from predevelopment paths, but the impermeable and relatively smooth paved surfaces prohibit infiltration and reduce friction to increase runoff magnitude and velocity. In short, the urban surface has elements with natural runoff processes and yet has a greater impervious surface cover and hydraulically more efficient drainage networks than undeveloped areas. Urbonas and Roesner (1993) and Mays (1999) provide an overview of the typical hydrologic design for urban drainage and flood control, identifying changes to the channel, floodplain, and watershed surface and subsurface that enhance drainage and flood

control, with more recent alternative design methods also treating degraded urban water quality.

History of Drainage

A brief history of runoff conveyances and urban drainage networks adds context to the importance of this phenomenon. Walesh (1989) and Mays (1999) recount that storm drains were constructed as far back as 1000 B.C., and into the mid- to late 1800s most subsurface masonry or wooden drains were constructed solely to remove stormwater, while wastewater was deposited into privies, streets, and courtyards. Burian *et al.* (2000) go on to describe the evolution of combined sanitary sewers in their historical review of wastewater management in the United States, explaining that the earliest systems of the 1850s in Chicago, Illinois, Brooklyn, and New York were fashioned after the first Combined Sewer Overflow (CSO) built in Germany. According to Burian *et al.* (2000), while separate sewer systems (SSS) were advocated at this same time, they were not built due to cost and a lack of precedent in European cities. Today in the United States, there are approximately 800 000 km of stormwater sewers in the cities (USEPA, 2001b), and the 20-year cost of updating these systems reaches into trillions of US dollars (Congressional Budget Office, 2002).

When vitrified clay replaced brick, wood, and stone sewer material, sewers could be constructed at smaller diameters, SSS costs became more competitive, and sewers were built in Lenox, Massachusetts, in 1875, and in Memphis, Tennessee, in 1880. Despite some failures in the early SSS designs, the technology disseminated, and by 1892, of the 27 cities with wastewater treatment works, 26 had adopted SSS designs to reduce treatment volumes and by the 1930s most cities were building with SSS designs. Combined sewers reduce the volume of stormwater reaching the local stream by routing the mixed waters to the treatment plant, capturing the entire storm for small events and spilling a fraction of the storm during larger events.

External Impacts of Construction

Land-use types altered by urban and suburban development, and associated runoff processes, are spatially extensive and locally significant, yet not considered central to the urban focus in this chapter. Such alterations are listed, however, for identification. Transportation corridor linkages with exterior communities are an alteration associated with urban areas of all sizes to facilitate transport of supplies, equipment, and labor to the development site. Features of larger projects that often accompany the on-site construction are the development of separate sites, locally, regionally, or internationally, as iron ore mines, stone and sand quarries, and timber harvests to provide construction materials and the local clearing of staging areas for storing and assembling development components. Final components for major urban development that extend beyond the urban area include the tapping of new water supplies, either surface or subsurface, with associated conveyance systems; the deposition of stormwater runoff and treated wastewater into receiving water supplies, at times creating interbasin transfers; the utilization of new energy supplies, in many regions using water withdrawals for hydropower or cooling, and removal of trees for transmission lines to the urban area; and the excavation and compaction of soils using new garbage and trash-handling facilities.

Delineation of Developed Catchments

McPherson and Schneider (1974) identified the difficulty of delineating the urban catchment area due to determining drainage divides in stormsewer areas when development grading, cut and fill, and road curb and gutter construction may have changed mapped elevations. Stormsewers may also run at subterranean slopes that are opposite in direction to surface terrain slopes. Djokic and Maidment (1991) recognized similar limitations in delineating urban watersheds when man-made structures divert stormwater from the predevelopment map prediction of steepest descent to alternate conveyance systems. Rather than collect current and dense terrain data of the current urban surface, the authors presented an algorithm based on integrated analysis of triangular irregular network (TIN) terrain datasets (Jones *et al.*, 1990), stormwater intakes, and the drainage network, within the Environmental Systems Research Institute (ESRI) Arc geographical information system software (Djokic and Maidment, 2000). The integrating algorithm is a privately owned expert system shell called *Nexpert Object*, and is demonstrated in the city of Asheville, North Carolina, where the required drainage network and stormwater intake maps are available.

An alternative approach is to find elevation maps with spatial resolution that capture the path of steepest descent, including 2% transverse sloping roads from the crown at approximately 3 m per lane width, 6% sloping gutters, 25 cm curb breaks, and stormwater intake depressions (AASHTO, 2002), and then use the standard terrain analysis algorithms, such as D8, multiple flow (MF), and D-Infinity, discussed in this article. Aerial orthophotos with 2.54 cm to 330 m scale imagery provided 0.6 m contours to test the use of this approach in two $<1\,km^2$ suburban watersheds in New York City's Croton watershed drinking water supply area within Westchester and Putnam Counties. Road longitudinal slopes ranged from flat to 6% grades, and predictions of flow path across 2 m grid cells provided relatively high success in capturing observed surface flow paths (Endreny *et al.*, 2002). Subsequent use of 0.3 m contours, generated by the same photogrammetric technique at a scale of 2.54 cm to 165 m, in a $<1\,km^2$ urban watershed in Syracuse, NY, failed to capture some road crowns

and curb breaks that delineated drainage inlet networks, necessitating site visits and manual delineation.

RUNOFF FROM URBAN AND SUBURBAN CATCHMENTS

Evaporation

Evaporation rates and lengths from urban land cover types are generally smaller than those from forested, grass, and agricultural land cover types. *In situ* and remote sensing studies have shown that urban structures are primarily non-vegetative and create a heat island effect (Lo *et al.*, 1997) due to a higher heat capacity than vegetation and a tendency to reemit heat into the surrounding environment (Stone and Rogers, 2001). Further, the absence of a large area of moist vegetation and soil restricts net radiation fluxes to partition predominantly into sensible heat fluxes, rather than latent fluxes via evaporation. In rural areas, moist vegetation and soil land cover allow for a significant percentage of the heat energy flux to go toward latent heat of vaporization and increase the atmospheric water vapor, cooling the surrounding air. Dow and DeWalle (2000) found that evaporation declined by 12 cm as urbanization increased from 0% up to 60% from 1920 to 1990 for 51 eastern United States watersheds. Because of the different patterns of irrigation supply and utilization of stormwater retention basins, there was significant variability in watershed response. Patterns in urban climate are under greater study as part of the National Science Foundation support of Long-term Ecological Research watersheds in Baltimore, Maryland and Phoenix, Arizona (Brazel *et al.*, 2000).

Precipitation

Urban heat island sources that are advected and diffused through the urban area contribute to the creation of temperature gradients and upstream and downstream thermals, an effect that has been examined in laboratory models under calm and stratified conditions (Noto, 1996). Any intense updraft of this thermally warm core, with greater moisture content, contributes to vertical clouds, moist convection, and storm formation (Baik *et al.*, 2001). This work showed that as the urban heat island intensity increased, cloud water formation times decreased and horizontal proximity to the heating center decreased. Baik and Chun (1997) also noted that when the heating depth is greater, precipitation is greater downstream.

Precipitation alterations have been confirmed in the field with gage, ground-based radar, and satellite radar detection systems. Changnon (1984) analyzed 50 years of gage measured rainfall data, as part of the 1970 Metropolitan Meteorological Experiment, for the Chicago metropolitan area and determined that there were 76% more occurrences of 2.54 cm or greater rains in the city than in the surrounding rural areas. In addition to the favorable thermal regime, the difference was explained by greater accumulations of required condensation nuclei through urban atmospheric particulate matter, which is a trend now identified in other United States' metropolitan regions (Changnon and Easterling, 2000). Satellite analysis by Shepherd *et al.* (2002) of Tropical Rainfall Measuring Mission's (TRMM) precipitation radar measurements for several storms over five southern United States cities identified a 28% average rainfall increase between 30 and 60 km downwind of the urban area, and a 6% average rainfall increase in the urban area. Important radar-based urban rainfall research from Europe has been reported by Thomas and Schmitt (2001).

Infiltration

Soils in suburban and urban systems are observed to have greater incidences of compaction and lower permeability, thereby reducing infiltration rates. The ability to predict the distribution and infiltration behavior of disturbed soils has been investigated with field and laboratory tests by Hamilton and Waddington (1999) in the Philadelphia, Pennsylvania area and by Pitt and Lantrip (2000), and Pitt *et al.* (2003) in the Birmingham, Alabama area. These studies have demonstrated that soils of similar texture and class showed significantly different infiltration rates, where infiltration declined with field- and lab-determined increases in compaction. Significant variability in infiltration was found between adjoining yards (Hamilton and Waddington, 1999), and the scatter for a single soil texture was great enough that decay functions could be replaced with horizontal lines with random scatter. Pitt and Lantrip (2000) conclude that very large errors in soil infiltration rates should be expected if published soil maps are used in conjunction with recommended texture-based infiltration model parameters for the Horton exponential decay model, and recommend that knocking and cone penetrometer studies be used to adjust parameters based on compaction. Dr. David Goodrich provides more details on infiltration dynamics and associated runoff in this series (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3** in this series).

Urban and Suburban Hydrological Storages

Depression storage is the depth or volume of ponded water accumulated at a low point with no possibility of escape as runoff, but is lost to the atmosphere by evaporation or to the soil by infiltration. Detention storage is different, and is defined as the depth or volume of water in motion over the land and contributes to runoff and possibly to infiltration and evaporation (AASHTO, 1999). Depression storage for different urban land cover types is provided in Table 3. A functional relationship for mean depression storage, d (cm), across generic urban land cover has been derived based on

Table 3 Depression storage (mm) for various land cover types

Land cover type	Depression storage range (mm)
Large paved area	2–4
Roofs – flat	3–8
Roofs – sloped	2–3
Lawn grass	5–13
Wooded areas/open fields	5–15

Table 4 Ratio of peak runoff rates after and before development from single family residential sites (Modified from Urbonas and Roesner, 1993)

Return period	Washington DC	Denver, Colorado	Canberra, Australia
2	10.8	57.0	9.0
10		3.1	4.7
15			
100	4.1	1.9	1.9

slopes by Kidd (1978), as an exponential decay function based on the area's percent watershed slope, S, with data from Holland, United Kingdom, and Sweden fitting with a correlation coefficient of 0.85.

$$d = 0.077 \cdot S^{-0.49} \quad (1)$$

Viessman *et al.*, (1977) used data from small impervious areas near Baltimore, Maryland and constructed a linear relationship that is approximated for depression storage d (cm), based on percent watershed slope, S, by

$$d = 0.341 - 0.076 \cdot S \quad (2)$$

STORMWATER GENERATION, COLLECTION AND DRAINAGE

Runoff Timing and Volume

Butler and Davies (2000) explain that the effects of urbanization on runoff are to change the relative proportions of water that infiltrates, evaporates, travels as subsurface flow, and is carried as surface stormwater. In some urban areas, the magnitudes are changed not only due to evaporation, but also due to alterations of the convective cycle and increases in precipitation as well as from interbasin transfers of drinking water that is released within the constructed area through services and irrigation, such as pipe exfiltration, septic discharges, car washes, sprinklers, and so on (Lerner, 2002). Calder (1993) provides a series of interception and evaporation equations to predict the hydrologic effects of land-use change, focusing the review primarily on afforestation and deforestation, but does not address the replacement of these natural systems with suburban or urban infrastructure.

Studies from the 1960s onward have documented that urbanization increases the volume of direct runoff and magnitude of peak discharge, decreases the time to peak discharge, and decreases the volume of subsurface runoff and baseflow (Lazaro, 1979; Bedient and Huber, 2002). Pre- and postdevelopment assessment of stormwater runoff has been reported for a small set of watersheds, some of which are reported in Table 4. In short, flood peaks of shorter recurrence intervals (less than a 50-year event or greater than a 0.02 probability) are increased by urbanization, creating new flood frequency distribution curves that shift to the left. Paired watershed monitoring is another method for identifying urban effects on hydrology, but given the challenge in controlling for covariates, computer modeling has become the most favorable method for urban effects assessment (Walesh, 1989). The difference in runoff efficiency between impervious and pervious surfaces decreases with increasing storm intensity, duration, and return period. As would be expected, as events extremity increases, soils become increasingly saturated and depression storage is more completely filled, causing a direct translation of runoff into the major drainage.

Impervious cover and compacted soils (Hamilton and Waddington, 1999; Pitt *et al.*, 2003) are the principal urban features causing runoff to increase as a result of development. Such a scenario is captured in Figure 1. The values of runoff provided in this figure are estimates based on assumed cover classes and precipitation regimes, and the subtleties of how these change with effective impervious area connectedness and storm intensity and duration are not addressed. The efficiency with which the urban and suburban impervious areas convert rainfall to runoff correlates strongly with the percent impervious cover, but more recent studies have increased this correlation by limiting the analysis to impervious areas directly connected to the drainage system (e.g. DCIA) and not surrounded by pervious surfaces.

Anderson (1970) observed that typical drainage system improvements, such as road networks and stormsewers, reduced lag time to one-eighth that of an undeveloped watershed, creating a "flashiness" in response, where the flood centroid arrives sooner. The flashiness combined with increased impervious cover resulted in flood peaks increasing by a factor of 2 to 8 times that of a natural watershed. Lazaro (1979) reports on studies that corroborated these findings, noting a tripling of peak flows and a reduction in rise time by one-third depending on channel modifications, impervious cover extent, and drainage facilities. Regional, and possibly dated, regression equations that predict increase in flood peak, lag time of flood peak, and magnitude of mean annual flood based on the watershed runoff coefficient, slope, and area were developed by Carter

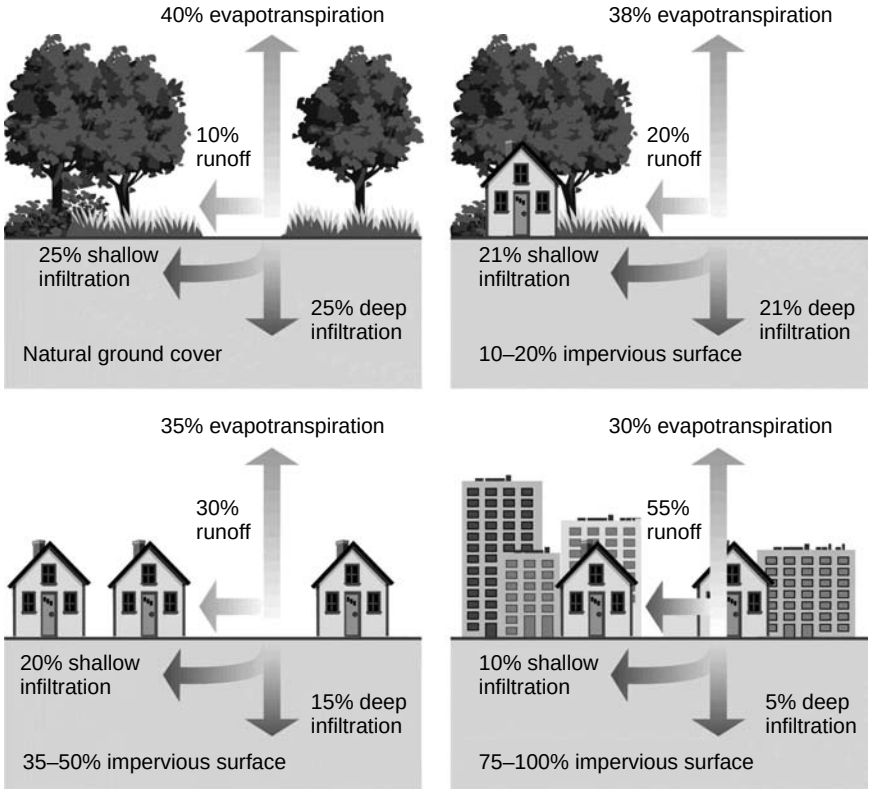


Figure 1 The augmentation of runoff from 10 to 55% that generally corresponds with land-use change toward urban development is schematically shown. Borrowed from the USEPA

(1961) based on watersheds in the Washington, DC area. Urbonas and Roesner (1993) quantitatively demonstrate how the runoff coefficient increases with impervious cover, which was developed for Denver, Colorado area and may have limited application in other areas. Regional limits are common to empirical methods.

Runoff Generation and Routing

Several approaches are used to estimate the runoff from urban areas as discharge, Q ($m^3\,s^{-1}$), and include the unit hydrograph, the Epsey–Altman hydrograph, the Natural Resources Conservation Services (NRCS) dimensionless hydrograph, the kinematic-wave model, the nonlinear reservoir model, and the Santa Barbara urban hydrograph model. Runoff prediction can take the form of field-measured responses, where the ordinates in the unit hydrograph are multiplied by effective rainfall based on different storm durations, or the synthetic urban hydrograph shown below may be employed to estimate instantaneous discharge, Q ($m^3\,s^{-1}$), for each time, t (s), based on peak discharge, Q_p ($m^3\,s^{-1}$), time to peak (s), and an exponential decay parameter, k (s^{-1}) (Butler and Davies, 2000).

$$Q = \frac{t}{t_p} \cdot Q_p \text{ for } 0 < t < t_p \tag{3}$$

$$Q = Q_p \cdot e^{-\frac{t-t_p}{k}} \tag{4}$$

Epsey–Altman 10-min unit hydrographs estimation (Epsey and Altman, 1978) is a technique derived from 41 small watersheds ($<40\,km^2$) with impervious cover extending from 2 to 100%. Runoff hydrograph shape and magnitude were estimated as a function of drainage area, A (km^2), impervious cover, I (%), main channel length, L (m), main channel slope, S, and a dimensionless catchment conveyance factor, Φ. Time to peak, T_p (s), peak discharge, Q_p ($m^3\,s^{-1}$), time of base, T_b (s), and hydrograph width at 50% and 75% peak discharge, W_{50} (s) and W_{75} (s), are given below (Chin, 2000),

$$T_p = 246 \frac{L^{0.23} \Phi^{0.57}}{S^{0.25} I^{-0.18}} \tag{5}$$

$$Q_p = 359 \frac{A^{0.96}}{T_p^{1.07}} \tag{6}$$

$$T_b = 98700 \frac{A}{Q_p^{0.95}} \tag{7}$$

$$W_{50} = 15120 \frac{A^{0.93}}{Q_p^{0.92}} \tag{8}$$

$$W_{75} = 5700\frac{A^{0.79}}{Q_p^{0.78}} \quad (9)$$

Figure 2 provides conveyance values for different main channel Manning's n values and percent impervious cover. Width of flow, as a time interval, at 50% and 75% Q_p is split such that one-third precedes the peak and two-thirds are allocated to the falling limb.

Natural Resources Conservation Services (NRCS, formerly the SCS) dimensionless unit hydrographs, while developed for natural catchments, are frequently applied to urban catchments (Chin, 2000), particularly in development planning. Storm event hydrograph plotting positions are derived from estimates of time to peak, T_p (h), and peak discharge, Q_p ($m^3 s^{-1}$), using estimates of excess rainfall duration, t_r (h), time lag, t_l (h), watershed area, A (km^2), and the dimensionless ratios in Table 5, as given by Chin (2000),

$$T_p = 0.5t_t + t_l \quad (10)$$

$$Q_p = 2.08\frac{A}{T_p} \quad (11)$$

If time lag is not known, an estimate is obtained from the time of concentration t_c (h), given below, and in urban catchments should account for increased efficiencies of stormsewers and drainage ways.

$$t_l = 0.6t_c \quad (12)$$

The NRCS triangular hydrograph is often used to approximate the dimensionless hydrograph, which has the same peak discharge and volume under the rising and falling limb, but the triangular hydrograph uses a total event time of 2.67 T_p, rather than 5 T_p as in the dimensionless estimation. Figure 3 shows the NRCS triangular and dimensionless hydrographs.

The kinematic-wave model is derived from a one-dimensional solution of the continuity and momentum equations, where the momentum terms for local acceleration, convective acceleration, and pressure slopes are neglected, leaving friction as the dominant term. Overland flow depth, y (m), is solved as a function of flow per unit width, q ($m^2 s^{-1}$), based on effective rainfall intensity, i_e ($m s^{-1}$), coefficient α, and a term for laminar or turbulent flow, m. The combined continuity and momentum equation is given as

$$\frac{\partial y}{\partial t} + \alpha m y^{m-1}\frac{\partial y}{\partial x} = i_e \quad (13)$$

The solution for flow depth, based on effective rainfall intensity, is then substituted into the momentum equation

Table 5 Dimensionless unit hydrograph ratios of time to time to peak (T_p) and discharge to peak discharge (Q_p)

t/T_p h/h	Q/Q_p $m^3 s^{-1}/m^3 s^{-1}$
0	0
0.2	0.1
0.4	0.31
0.6	0.66
0.8	0.93
1	1
1.2	0.93
1.4	0.78
1.6	0.56
1.8	0.39
2	0.28
2.2	0.207
2.4	0.147
2.6	0.107
2.8	0.077
3	0.055
3.4	0.029
4.2	0.01
4.6	0.003
5	0

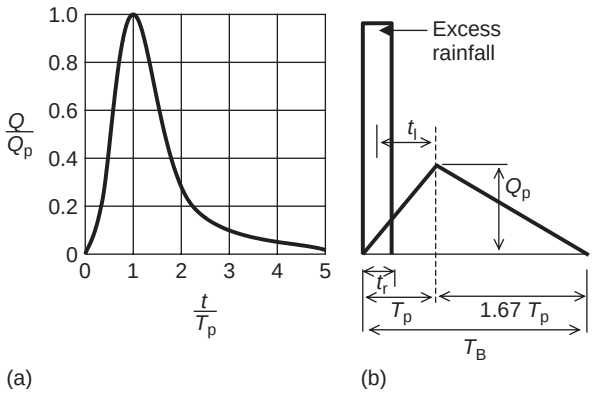


Figure 3 NRCS (a) dimensionless and (b) triangular unit hydrographs

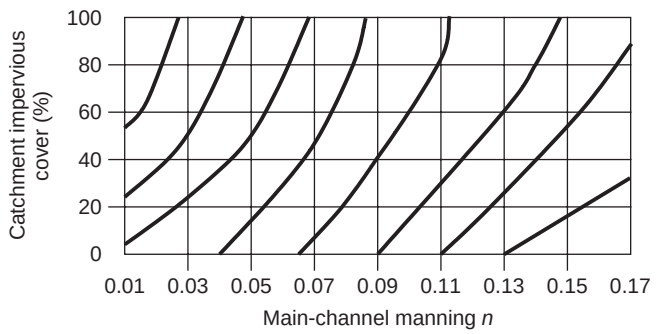


Figure 2 Conveyance factor Φ nomograph based on main channel Manning roughness and impervious extent (%)

to find runoff per unit width,

$$q = \alpha y^m \tag{14}$$

A finite element approximation of the kinematic-wave model is given as

$$y_x^t = q_1 \Delta t + y_x^{t-1} - \alpha m \left(\frac{\Delta t}{\Delta x}\right)\left(\frac{y_x^{t-1} + y_{x-1}^{t-1}}{2}\right)^{m-1} \times \left(y_x^{t-1} - y_{x-1}^{t-1}\right) \tag{15}$$

The kinematic-wave model does not simulate hydrograph attenuation, which is negligible in urban drainage systems, making the equation useful in urban settings. Overland inflow entering a channel per unit length of channel, q_0 ($m^2\,s^{-1}$) and cross-sectional flow area, A (m^2), are used in the kinematic-wave equation for channel flow, given as

$$\frac{\partial A}{\partial t} + \alpha m A^{m-1}\frac{\partial A}{\partial x} = q_0 \tag{16}$$

The nonlinear reservoir model simulates the urban catchment as a shallow reservoir with inflow equal to rainfall excess and outflow a nonlinear function of flow depth y (m), and depression storage, y_d (m). Estimates of runoff depth are given as a function of average excess rainfall over the time step, i_e (m), a representative catchment width, W (m), Manning's roughness, n, an average catchment slope, S_o, and watershed area, A (m^2), given in finite difference form by Chin (2000) as

$$\frac{y_2 - y_1}{\Delta t} = i_e - \frac{W S_o^{0.5}}{An}\left(\frac{y_1 + y_2}{2} - y_d\right)^{5/3} \tag{17}$$

Equation (17) is solved iteratively for y_2, the depth of flow at the end of the time step, which is then substituted into the following discharge equation as y.

$$Q = \frac{W}{n}(y - y_d)^{5/3} S_o^{1/2} \tag{18}$$

The Santa Barbara Urban Hydrograph (SBUH) model assumes that the runoff is derived from the directly connected impervious fraction of the catchment, and may neglect abstractions such as depression storage and rainfall on pervious areas or account for them with a submodel. While the model was developed for Santa Barbara, CA, it has been applied for other urban areas in the United States (Chin, 2000). Instantaneous runoff, q ($m^3\,s^{-1}$), is computed for each time interval, Δt (s), based on watershed area, A (m^2), fraction of watershed that is directly connected impervious, x, rainfall intensity, i ($m\,s^{-1}$), and rainfall excess intensity, i_e ($m\,s^{-1}$), given as

$$q = [ix + i_e(1.0 - x)]A \tag{19}$$

Instantaneous runoff is distributed through time using a routing constant, K_r, based on the time of concentration, t_c (s), given as,

$$K_r = \frac{\Delta t}{2t_c + \Delta t} \tag{20}$$

Ordinates of the routed runoff hydrograph, Q ($m^3\,s^{-1}$), are then computed for each time step, j, as

$$Q_j = Q_{j-1} + K_r(q_{j-1} + q_j - 2Q_{j-1}) \tag{21}$$

Depending on the application, the most common equations available for engineering design are the Manning equation and the Darcy–Weisbach equation (Mays, 1999; Chin, 2000; Haestad, 2002). In stormwater runoff computations, the Manning equation is the predominant form for estimating discharge (ASCE, 1993; Mays, 1999; AASHTO, 2000; Haestad, 2002), and is derived as a function of flow cross-sectional area, A (m^2), Manning's roughness, n, hydraulic radius, R (m), and water surface slope, S.

$$Q = \frac{A}{n}R^{2/3}S^{1/2} \tag{22}$$

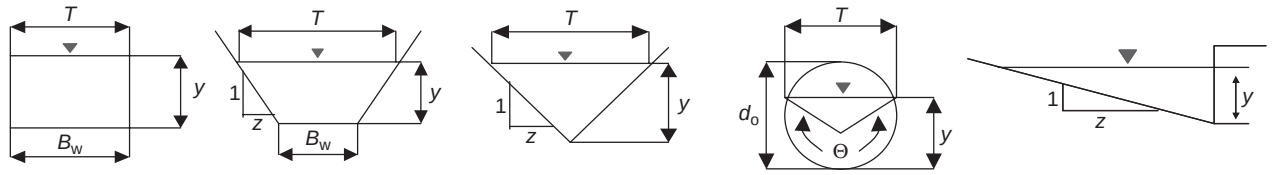
Equation (22) application should be limited to normal flow conditions, where water surface slope, or the energy grade line, parallels channel slope, and when flow is hydraulically rough (Chin, 2000). Hydraulically rough flow is maintained when the following condition is true (Chin, 2000),

$$n^6 > \sqrt{RS} \geq 1.9E - 13 \tag{23}$$

Solution of runoff equations typically requires basic geometric properties for the conveyance devices, which are sketched in Figure 4, along with the equations needed to compute top width, T (m), area, A (m^2), perimeter, P (m), and hydraulic radius, R (m). Values for Manning's roughness factor are provided in Table 6, but a single value should be used with caution. Instead, a range should be used to represent the uncertainty in their values, as noted by Johnson (1996), who documented field estimated values for n have errors between 5 and 35%.

Runoff in Minor and Major Systems

Urban and suburban areas are designed to handle stormwater runoff for frequent (e.g. minor events based on intensity for a given duration) and infrequent (e.g. major events) systems. The minor drainage system is comprised of gutters, drainage inlets, and subsurface stormsewers, and when its capacity is exceeded, the major drainage system of surface runoff along the streets serves as a conveyance route. In both systems, the velocity and discharge are estimated using a form of the conservation of momentum equation. Urban



Dimension	Rectangle	Trapezoid	Triangle	Circle	Gutter
Top width	$T = B_w$	$T = B_w + 2 \cdot z \cdot y$	$T = 2 \cdot z \cdot y$	$T = \left(\sin\left(\frac{\theta}{2}\right)\right) \cdot d$	$T = y \cdot z$
Area	$A = B_w \cdot y$	$A = (B_w + zy)y$	$A = z \cdot y^2$	$A = \frac{1}{8} \cdot (\theta - \sin(\theta)) \cdot d^2$	$A = y^2 \cdot \frac{z}{2}$
Perimeter	$P = B_w + 2y$	$P = B_w + 2 \cdot y \cdot \sqrt{1 + z^2}$	$P = 2y \cdot \sqrt{1 + z^2}$	$P = \frac{1}{2} \cdot \theta \cdot d$	$P = y \cdot \left(1 + \sqrt{1 + z^2}\right)$
Hydraulic radius	$R = \frac{B_w \cdot y}{B_w + 2y}$	$R = \frac{(B_w + zy)y}{B_w + 2 \cdot y \cdot \sqrt{1 + z^2}}$	$R = \frac{z \cdot y^2}{2y \cdot \sqrt{1 + z^2}}$	$R = \frac{1}{4} \cdot \left(1 - \frac{\sin(\theta)}{\theta}\right) \cdot d$	$R = \frac{y \cdot \frac{z}{2}}{1 + \sqrt{1 + z^2}}$

Figure 4 Basic elements in urban stormwater runoff with equations for their top width, area, perimeter, and hydraulic radius

highways, streets, or roads are constructed with fairly standardized engineering plans (NYSDOT, 2001; AASHTO, 2002; Dewberry, 2002), and are classified as principal arterial interstates, principal arterial expressways, minor arterials, collectors, and locals, which have traffic volumes, in vehicles of one-way traffic per hour, that exceed 500 and may be less than 200. Urban local and collector roads are typically not elevated above the 100-year or 500-year floodplain as are principal arterial interstates and expressways, and often are located below the immediate surroundings (Lazaro, 1979) such that sections may flood during high-frequency events.

Drainage is incorporated into road design to maintain structural integrity by removing water (NYSDOT, 2001). A cross section of a typical design is shown in Figure 5, where the road is constructed with a concrete or asphalt top, about 300 mm of well-graded sand and gravel compacted subbase layer, about 100 mm of a coarse aggregate permeable base layer, and a 100 mm diameter edgedrain surrounded by compacted filter material, with the invert extending upward of 100 mm above the top of subbase in limiting slopes but preferably 200 mm below the subbase.

Rooftop Runoff

Drainage systems that begin with roof drainage in a residential house first collect into the eaves gutter, which is a 75-mm half-round channel, and which feed into down pipes that are closed 75–100 mm rectangular channels (Butler and Davies, 2000). In larger urban buildings, the roof drainage is often directly connected to the subsurface stormsewer system, but more innovative, green building design has now incorporated rooftop gardens to capture and evapotranspire the rainfall.

Street and Gutter Runoff

Surface runoff along streets in the United States has a maximum design spread or extension into the traffic lane set by the Federal Highway Authority based on a 5-, 10-, and 50-year return interval storm. Design spreads depend on the design vehicular velocity and traffic volume for that road, and for local streets, generally extend across half a driving lane, whereas for highways, they are typically constrained to at most 1 m into the driving lane while the shoulder is designed to route the majority of the flow (AASHTO, 1999). Lane widths in urban areas vary between 2.7 and 3.7 m depending on the number of vehicles of one-way traffic per hour, and maximum longitudinal slopes range from 5 to 11% depending on the design maximum vehicular speed and the topographical relief (NYSDOT, 2001). Road crown sections are normally sloped at 2%, but the superelevation of the crown may reach a maximum slope of 8% from the shoulder on certain ramps (AASHTO, 2002). In certain designs the crown will be flattened and the entire road will slope between 2 and 8% away from the point of rotation to improve vehicle traction and promote drainage (NYSDOT, 2001). Parking lanes have cross slopes of 4%, and shoulder slopes are normally 6%. Slopes from edge of shoulder along the embankment are normally 2% and limited to a maximum of 3.3% to prevent vehicle turnover.

Minor stormsewer drainage systems typically require curbing, defined as a narrow, raised element placed at the

Table 6 Manning *n* roughness estimates for channel and pipe flow above a variety of materials

Material	Manning *n*	Material	Manning *n*
Natural channels		Excavated earth channels	
Clean and straight	0.03	Clean	0.022
Major rivers	0.035	Gravelly	0.025
Sluggish with deep pools	0.04	Weedy	0.03
		Stony, cobbles	0.035
Metal channels		Floodplains	
Brass	0.011	Pasture, farmland	0.035
Cast iron	0.013	Light brush	0.05
Smooth steel	0.012	Heavy brush	0.075
Corrugated metal	0.022	Trees	0.15
Engineered material channels			
Glass	0.01	Finished concrete	0.012
Clay tile	0.014	Unfinished concrete	0.014
Brickwork	0.015	Gravel	0.029
Masonry	0.025	Planed wood	0.012
		Unplaned wood	0.013
Plastic material			
Corrugated polyethylene (PE) with smooth inner walls			0.009–0.015
Corrugated polyethylene (PE) with corrugated inner walls			0.018–0.025
Polyvinyl chloride (PVC) with smooth inner walls			0.005–0.009
Overland flow (model calibrated)		Overland flow (field measured)	
Smooth asphalt	0.012	Concrete or asphalt	0.011
Asphalt of concrete paving	0.014	Bare sand	0.01
Packed clay	0.03	Graveled surface	0.02
Light turf	0.20	Bare clay loam	0.02
Dense turf	0.35	Range (natural)	0.13
Dense shrubbery and forest litter	0.4	Bluegrass sod	0.45
		Short-grass prairie	0.15
		Bermuda grass	0.41

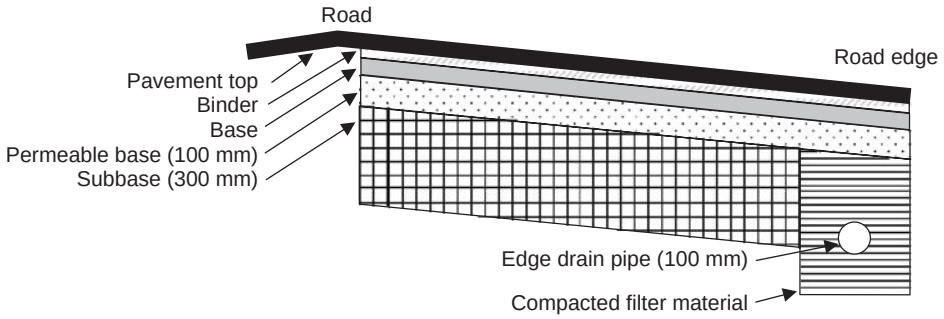


Figure 5 Road cross sections for asphalt pavement showing the crown, edge, top of pavement, binder, base, permeable base, subbase, compacted filter material, and edge drain pipe

edge of pavement, to control drainage and vehicular access, and gutters to channel flow from the road to the storm-sewer. Gutters are defined as a broad shallow ditch placed at the edge of a paved area and designed to primarily collect and carry surface water to a drainage system, and are sometimes partially congruent with shoulder or parking area functions. In the absence of curbs, runoff will continue along the transverse road and should slope into ditches that have a vertical depth set about 1.2 m below the edge of the traveled way to provide subbase drainage more than to provide hydraulic capacity (NYSDOT, 2001). Drainage from beneath the road via the edgedrain is released by connections to ditches, stormsewers, and other drainage devices.

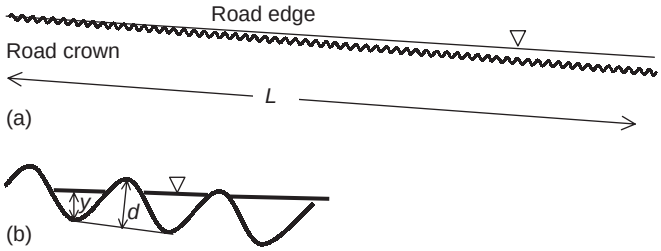


Figure 6 (a) Water film increasing in thickness as it flows from the road crown toward the road edge. (b) Detailed view showing that initially, the water depth does not exceed the road texture depth

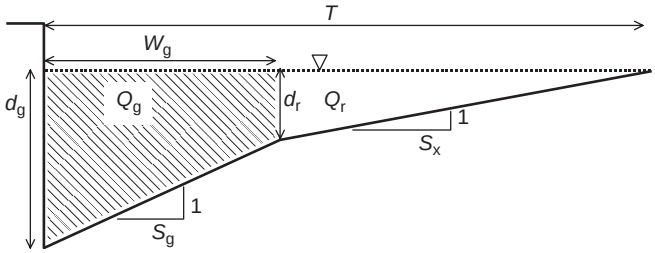


Figure 7 Composite gutter and road section showing distinct flow elements, with the gutter shaded in hash marks, the water surface indicated by a dotted line and inverted triangle, and the pavement and curb boundary indicated by the thick solid line

Rainfall on sloped pavement forms a thin film at the upper slope or road crest, and increases in thickness as it flows toward the edge of the pavement (see Figure 6). At the beginning of the storm, the film depth, y (mm), is beneath the pavement roughness depth, d (mm), and is considered a negative depth and is retained or infiltrated depending on the porosity of the surface. As the runoff length, L_f (m), increases and flow accumulates, it covers the macroasperities, and the film depth increases and behaves as channel flow (AASHTO, 1999). Runoff across a road has a final resultant flow path length and slope, S_r, that is determined on the basis of the road width, L (m), cross slope, S_x, and longitudinal gradient, S_l, as shown in equation (24).

$$S_r = \sqrt{S_x^2 + S_l^2} \tag{24}$$

$$L_f = L(S_r + S_x) = L\sqrt{1 + \left(\frac{S_l}{S_x}\right)^2} \tag{25}$$

Film depth is not appreciably affected by the road longitudinal grade (AASHTO, 1999) because as the grade is steepened, the flow path is lengthened and the flow velocity increases, which offsets increases in depth. As a result, the film water depth does not significantly change at the pavement edge. The film depth, y (mm), is estimated based on the above parameters as well as rainfall intensity, I (mm h^{-1}).

$$y = 0.01485\cdot(d^{0.11}\cdot L_f^{0.43}\cdot I^{0.59}\cdot S_r^{-0.42}) - d \tag{26}$$

Curb flow in composite sections is predicted for each section, illustrated in Figure 7, with two separate equations (Dewberry, 2002; Haestad, 2002). For the gutter section, with a depth of water at the curb, d_g (m), and a depth of water adjacent to the roadway section, d_r (m), discharge, Q_g (m^3 s^{-1}), is computed as a function of the two vertical depths, the road longitudinal slope, S_l, the Manning's roughness, n, and the gutter cross slope, S_w, as

$$Q_g = \frac{0.376\cdot(d_g^{8/3} - d_r^{8/3})\cdot S_l^{1/2}}{n\cdot S_g} \tag{27}$$

For the road section, with a road cross slope of S_x, discharge, Q_r (m^3 s^{-1}), is computed as follows:

$$Q_r = \frac{0.376\cdot d_r^{8/3}\cdot S_l^{1/2}}{n\cdot S_x} \tag{28}$$

The depth of gutter flow along the curb, d_g, can be estimated with the following relation:

$$d_g = T\cdot S_x + W_g\cdot(S_g - S_x) \tag{29}$$

The depth of road flow adjacent to the gutter section, d_r, can be estimated as,

$$d_r = (T - W_g)\cdot S_x \tag{30}$$

Drainage Inlet Runoff

Roadside stormwater is traditionally directed toward storm drain inlets that are designed to capture the street runoff and deliver it to the subsurface stormsewer. Drainage inlets are typically grates in the gutter, openings along the curb, slots within the street, or combinations of grate and gutter (see Figure 8).

Haestad Methods Software Company provides simulation tools (Haestad, 2002) to examine how typical drainage gutter and inlet sections perform for different return interval storms. Drainage inlets are placed either along sloped roads, called *on grade*, or at local minimum elevations where water gathers, called *on sag*. On-grade drainage inlets have a capture efficiency that varies with inlet design type (grate or curb and roughness and sizing of the material), the longitudinal and transverse slope, and the runoff volume, while on-sag inlets pond the water until all is delivered to

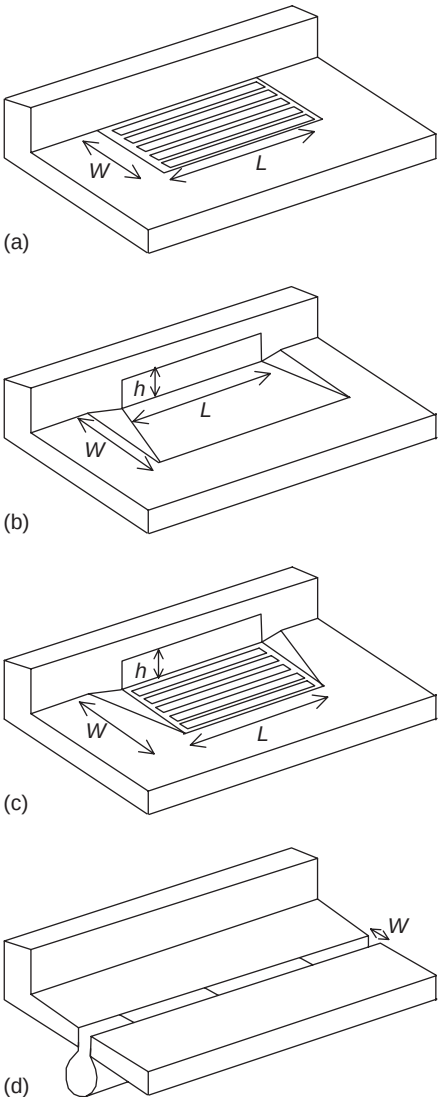


Figure 8 Street runoff drainage inlet styles of (a) grate inlet, (b) curb inlet, (c) combination inlet, and (d) slotted drain inlet

the stormsewer. Such ponding may increase the spread of runoff into the road. Stormsewer manholes used to access stormdrains are typically spaced at a minimum interval of 90 m, and are also placed at system origin, changes in direction, changes in gradient, changes in size, and major junctions (Butler and Davies, 2000).

Stormwater systems are preferably gravity driven rather than pump based to reduce the chance of failure. System slopes and dimensions are selected to maintain subcritical flows, which reduce the potential of scour, but may enhance the chance of backwater effects at junctions and constrictions (Walesh, 1989). As reported by Walesh (1989), safety measures recommended for surface conveyance systems with high flows include fencing or guardrails along steep channel sidewalls, ladders along armored channel walls to allow for escape if swept away, drop structures within channels to reduce flow velocity, and freeboard extending above the design stage.

Curb inlets experience reduced efficiencies when clogging debris blocks the flow, and have increased efficiency when placed to capture flows less than $0.08\,m^3\,s^{-1}$, longitudinal slopes are 2% or less, and local depressions surround the curb opening (ASCE, 1993). The curb operates in three flow regimes – as a weir with depths equal to the opening height, transitionally, and as an orifice with depths greater than 1.4 times the opening height (Chin, 2000). Effective head, d (m), for the curb inlet is computed as a function of the throat type for the curb, depending on the inclination of the curb, θ, the curb throat opening height, h (m), and the depth at the lip of the curb, d_i (m), as illustrated in Figure 9.

$$d_o = d_i - \frac{h}{2}(\sin\theta) \tag{31}$$

For a curb inlet operating as a weir with a depressed gutter width, W(m), an opening length, L(m), a depth of flow, d(m), and a weir coefficient, C_w, discharge, Q ($m^3\,s^{-1}$) are given by (Chin, 2000)

$$Q = C_w(L + 1.8W)d^{1.5} \tag{32}$$

$$C_w = 1.27 \tag{33}$$

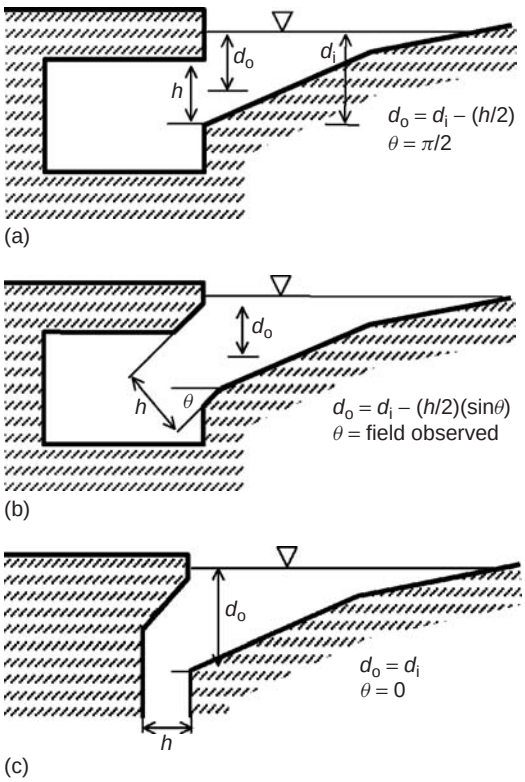


Figure 9 Different types of curb inlets, with (a) a horizontal throat, (b) an inclined throat, and (c) a vertical throat

Without a depressed gutter under weir flow, the discharge is given by (Chin, 2000)

$$Q = C_w L d^{1.5} \tag{34}$$

Curb inlets operate under orifice flow when the depth is $1.4h$, where h (m) is the opening height. Orifice flow estimates require an orifice coefficient, C_o, where discharge is given by (Chin, 2000)

$$Q = C_o A \sqrt{2g\left(d - \frac{h}{2}\right)} \tag{35}$$

$$C_o = 0.67 \tag{36}$$

Grate inlets also operate as weir or orifice flow, and the effective head, d (m), is simply the height of flow along the curb. Weir flow is achieved with a depth of flow, d (m), less than or equal to 0.12 m, and is based on the perimeter of grate opening excluding any length adjacent to the curb, P (m). Discharge is given by (Chin, 2000)

$$Q = C_w P d^{1.5} \tag{37}$$

$$C_w = 1.66 \tag{38}$$

For grates that are within a depressed gutter or other local depression, the perimeter, P (m), in the following equation is estimated as a function of the grate length, L (m), and width, W (m), (Haestad, 2002).

$$P = L + 1.8W \tag{39}$$

When depths of flow exceed 0.43 m, flow is orifice controlled and the area of the grate opening, A (m^2), is then used to determine discharge, given by (Chin, 2000)

$$Q = C_o A \sqrt{2gd} \tag{40}$$

$$C_o = 0.6 \tag{41}$$

For transitional flow at $0.12\,\text{m} < d \leq 0.43\,\text{m}$, the discharge through the grate inlet is predicted by the tangent line connecting the bounds of the weir and orifice curves (AASHTO, 2000; Chin, 2000).

Runoff not captured by drainage inlets on grade is passed back into gutter flow. Grate drainage inlets along the grade have a limited capture efficiency, which is higher for frontal flow than side flow, and is thereby increased by extending the grate toward the road. The ratio of intercepted frontal flow to the total frontal flow is a function of the gutter velocity up gradient of the inlet, V ($m\,s^{-1}$), the inlet splash-over velocity that jumps the grate and is obtained by grate manufacturers, V_o ($m\,s^{-1}$), and the frontal coefficient, k_f, given as (Haestad, 2002),

$$R_f = 1 - k_f(V - V_o) \tag{42}$$

$$k_f = 0.295 \tag{43}$$

For side flow, the length of the grate or depression, L (m), the road cross slope, S_x, and the side coefficient, k_s, are required as well as the velocity in the gutter at the inlet. The ratio of intercepted side flow to the total side flow is given as (Chin, 2000; Haestad, 2002)

$$R_s = \left(1 + \frac{k_s V^{1.8}}{S_x L^{2.3}}\right)^{-1} \tag{44}$$

$$k_s = 0.0828 \tag{45}$$

Culvert Runoff

Culvert flows route water collecting in ditches and along gutters under roads, driveways, or other structures. Culvert hydraulic equations are traditionally approached with conservation of energy equations to solve for headwater depths and flow rates. Nomographs derived from field experiments are typically used to find the design discharge, velocity, and headwater depth for different culvert barrel diameters (AASHTO, 2000), but the Federal Highway Authority has also developed a set of equations, in US customary units, to approximate design conditions (Haestad, 2002). The computational challenges stem primarily from the many flow regimes that might exist in the culvert including gradually varied flow, such as backwater condition, and rapidly varied flow, such as a hydraulic jump. Submerged culvert inlet flow typically provides more discharge capacity than unsubmerged inlets, but unsubmerged inlet flow is often the design target when the crown of the culvert barrel forms the base of a roadway (Chin, 2000). Roadway overtopping does occur in some cases, and is quantifiable (AASHTO, 2000).

Inlet-controlled culverts are characterized by turbulence energy losses near the culvert entrance when the flow passes from subcritical to supercritical depths, rather than to friction energy losses in the culvert barrel. Inlet-controlled flow is characterized by one of three conditions, an unsubmerged inlet, a transitionally submerged inlet, and a submerged inlet above $1.2D$, where D is the diameter of the culvert barrel (Chin, 2000). Several scenarios for inlet-controlled discharge are shown in Figure 10, with each case described in text. In culvert design, the headwater depth is the unknown variable, but for stormwater runoff analysis, the culvert equations are used to solve for discharge.

For cases (a) and (b), the culvert behaves similarly to a weir, with an unsubmerged inlet with a headwater depth, h_i (m), upstream velocity, V_i ($m\,s^{-1}$), critical flow depth, h_c (m), entrance losses, h_e (m), the entrance loss coefficient, k_e (which varies between 0.2 and 0.9 depending on the

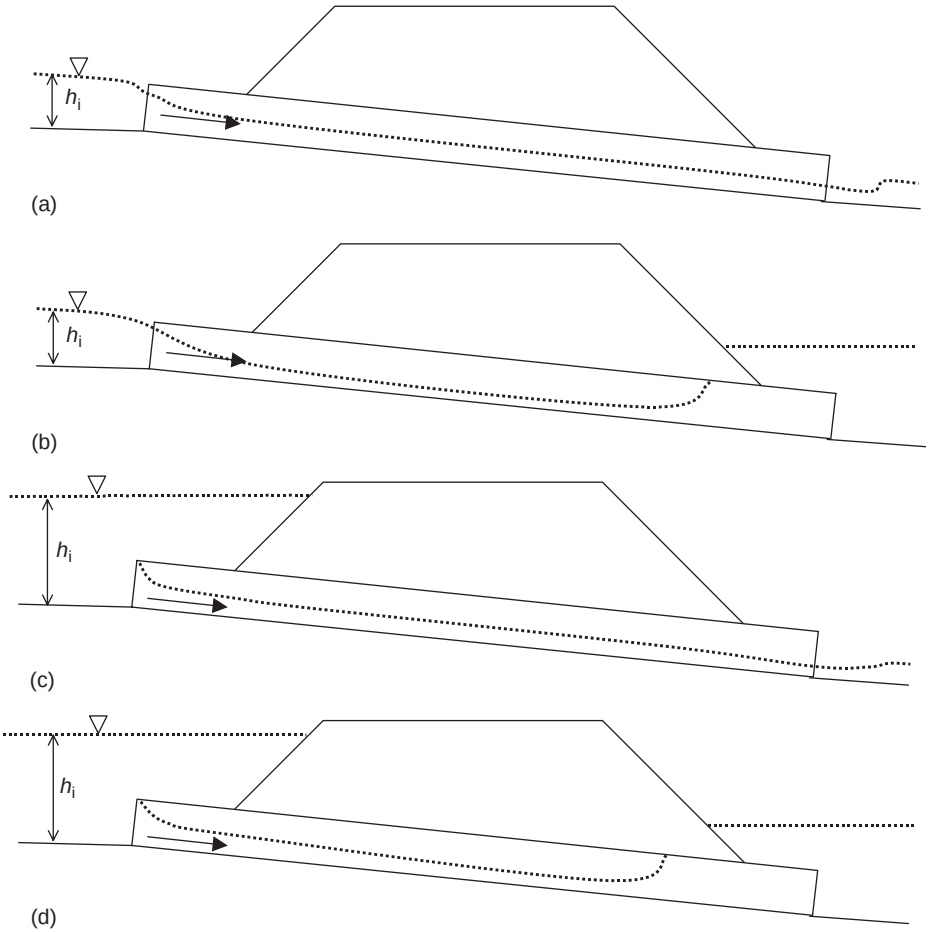


Figure 10 Inlet-controlled culverts passing beneath a road with (a) unsubmerged inlet and free flowing outlet, (b) unsubmerged inlet and submerged outlet, (c) submerged inlet and free flowing outlet, and (d) submerged inlet and submerged outlet. Water surface is shown as a dotted line, with the arrow indicating flow direction

entrance material and shape), a flow area at the critical flow section, A_c (m^2), and gravitational acceleration of g ($m\,s^{-2}$). Discharge is given by (Chin, 2000)

$$Q = A_c \sqrt{2g\left(\Delta h + \frac{V_i^2}{2g} - h_e\right)} \tag{46}$$

$$\Delta h = h_i - h_c \tag{47}$$

$$h_e = k_e \frac{V_1{}^2}{2g} \tag{48}$$

For cases (c) and (d), the culvert behaves similarly to an orifice, with a submerged inlet, with a coefficient of discharge, C_d, that ranges between 0.62 for a square edged entrance and 1 for a well-rounded entrance, a culvert cross-sectional area, A (m^2), and a vertical distance from the center of the culvert inlet to the depth of water at the inlet, Δh. Discharge is given by (Chin, 2000)

$$Q = C_d \cdot A\sqrt{2g\Delta h} \tag{49}$$

$$\Delta h = h_i - h_m \tag{50}$$

Outlet-controlled culverts are characterized by energy losses along the barrel length as well as some entrance losses. Several scenarios for inlet-controlled discharge are shown in Figure 11, with each case described in text. In cases where the flow does pass below critical depth (cases (d) and (e)), this occurs near the culvert outlet rather than near the inlet. For cases (a) and (b) with a submerged outlet and case (c) with an unsubmerged outlet, discharge is controlled by the driving energy head, Δh (m) equal to the difference between the headwater and tailwater elevations, the hydraulic radius, R (m), the length of the culvert, L (m), and the Manning's roughness, n. Discharge is given

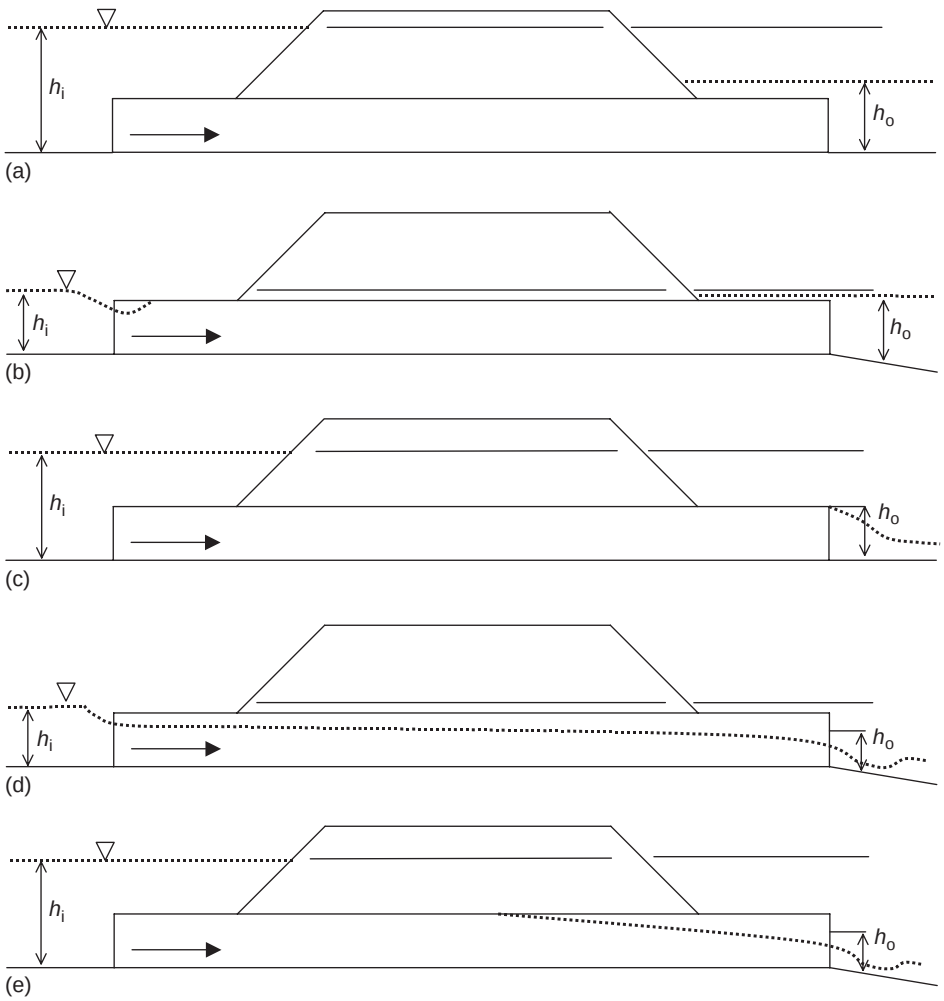


Figure 11 Outlet-controlled culverts passing beneath a road with (a) submerged inlet and outlet under full flow, (b) unsubmerged inlet and submerged outlet with full flowing barrel, (c) submerged inlet and unsubmerged outlet under full flow with normal depth greater than barrel diameter, (d) unsubmerged inlet and free flowing outlet below critical depth, and (e) submerged inlet and free flowing outlet below critical depth. Water surface is shown as a dotted line, with the arrow indicating flow direction

by (Chin, 2000)

$$Q = A\sqrt{\frac{2g\Delta h}{19.62n^2\dfrac{L}{R^{4/3}} + k_e + 1}} \tag{51}$$

$$\Delta h = h_i - h_o \tag{52}$$

For cases (d) and (e), the culvert experiences critical flow near the outlet, and discharge is predicted with an equation similar to that used for inlet-controlled flow, but depends on the friction losses in the barrel. Solutions to the discharge are implicit, as the friction losses are a function of the barrel velocity, V ($\mathrm{m\,s^{-1}}$), as well as many terms defined above. Discharge is given by (Chin, 2000)

$$Q = A_c\sqrt{2g\left(\Delta h + \frac{V_i^2}{2g} - h_e - h_f\right)} \tag{53}$$

$$h_f = \frac{n^2V^2L}{R^{4/3}} \tag{54}$$

$$\Delta h = h_i - h_o \tag{55}$$

Runoff Surcharge

Surcharge, or stormsewer discharge into lower level residences or streets, may occur when the hydraulic grade line (pressure and elevation energy slopes) rises above the crown of the stormsewer system (Mays, 1999). In such cases, drainage inlet discharge is reversed, and additional runoff may be released into gutters and roadways.

Surcharge is more common in older systems, where they may be undersized for the increased development within the drainage area and have lower capacities than in newer designs built to handle growing populations.

Stormsewer Runoff

Bisecting the urban watershed is the hydraulically improved drainage system. Drainage is often based on a combination of artificial and natural conveyance structures that collect and dispose of water. Details of stormsewer runoff exceed the scope of this chapter, and are primarily solved with application of the coupled hydraulic continuity and momentum equations to simulate the unsteady one-dimensional gradually or rapidly varied flow typical of stormwater systems. The Water Wave equations or Saint Venant equations are the coupled set of conservation of mass and momentum equations, in implicit or explicit structure depending on known variables, solved with finite difference, finite element, or the method of characteristics as described by Zoppou (2001).

The momentum equation has four components, known as *local acceleration slope, convective acceleration slope, pressure slope*, and *friction slope*. Complex routing is also achieved with simplifications to the momentum equation. The Diffusion Wave Model ignores the convective acceleration and local acceleration terms in the momentum equation, and the kinematic-wave equation, developed above, neglects the local acceleration, convective acceleration, and pressure slopes in the momentum equation. Models useful in applying these equations include StormCAD developed by Haestad Methods Software Company (Haestad, 2002) and the Storm Water Management Model (SWMM) developed for the US Environmental Protection Agency (Huber and Dickinson, 1992).

Runoff Detention and Storage

Urbonas and Roesner (1993) provide a broad overview on detention and retention facilities, which became components in urban stormwater management to limit the magnitude of hydrograph peaks in the early 1970s. Engineers size the volume of basins according to their type (e.g. detention or retention) (Mays, 1999; AASHTO, 2000; Chin, 2000), local ordinances, and the climatic patterns for rainfall. Detention basins (see Figure 12a) are designed to dry out by drainage through a low head culvert or weir outlet and infiltration through the base soils before the next rainfall event. Retention basins (see Figure 12b) are designed to remain wet between events, but drain excess storm volume through a high head riser or weir. Either basin design is able to control a range of events, from the 1.25 cm rain depths to 1-, 2-, 5-, 10-, and 100-year return interval events. Studies on the physical placement of basins have shown that when located near the mouth of a small watershed, the timing of their release may overlap with headwater flood wave arrival, which can exacerbate the flood peak, suggesting that central and headwater placement is more efficient (Urbonas and Roesner, 1993). Inflow and outflow structures for the basins must control for erosion, deposition, and scour, and need regular maintenance to remove windblown and storm-carried debris and trash (WEF/ASCE, 1998).

Storage basins attenuate the increases in runoff peak and time to peak from upslope urban development inflow

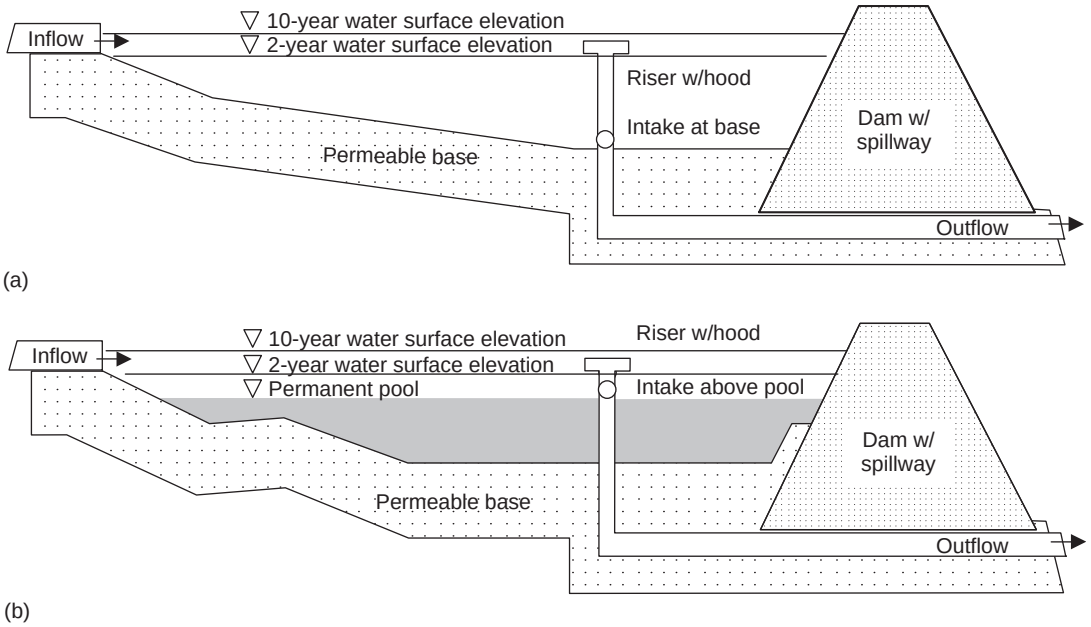


Figure 12 Storage basins for the (a) detention and (b) retention of stormwater, with the inflow, water surface elevation, riser and intake, permeable base, dam, and outflow illustrated

hydrograph by delaying the outflow hydrograph through increases in storage, as illustrated in Figure 13. Storage device outflow hydrographs as a function of time, O_t ($m^3\,s^{-1}$), are predicted given the inflow hydrographs as a function of time, I_t ($m^3\,s^{-1}$), the hydrograph time step, Δt (s), and the storage volume as a function of time, S_t (m^3), using the level pool routing continuity equation (Mays, 1999), otherwise known as *the storage indication method* (Chin, 2000),

$$\left(\frac{2S_{t=2}}{\Delta t} + O_{t=2}\right) = (I_{t=1} + I_{t=2}) + \left(\frac{2S_{t=1}}{\Delta t} - O_{t=1}\right) \tag{56}$$

Solution requires an initial estimate of the outflow hydrograph at time step 1, which is typically set to zero. The above equation is solved using a look-up table relating storage volume, or the storage stage, h (m), to discharge. Such relations often take the form of a weir or an orifice discharge function based on the outlet device. Weir-type discharge requires information about the crest length of the weir, b (m), the elevation of the water surface above the weir crest, h (m), the weir discharge coefficient (frequently set to 0.62), and are often formulated as (Chin, 2000)

$$Q = \frac{2}{3} C_d \sqrt{2g} \cdot b h^{3/2} \tag{57}$$

Orifice type discharge through a fully submerged culvert requires information on the cross-sectional area of the culvert, A (m^2), the elevation of the water surface above the center of the culvert entrance, and the orifice discharge coefficient, C_d, and is given as (Chin, 2000)

$$Q = C_d A \sqrt{2gh} \tag{58}$$

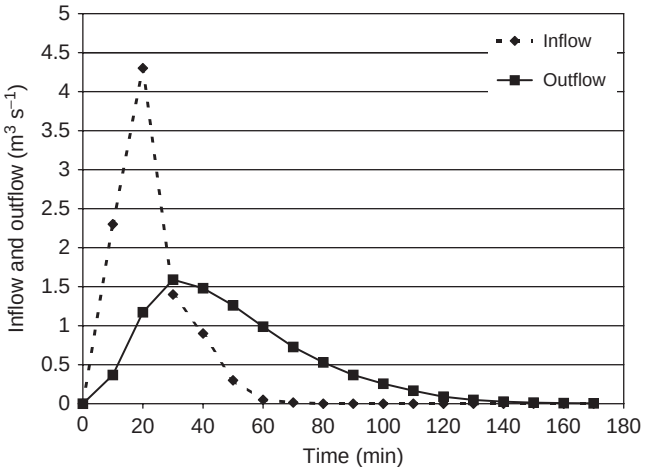


Figure 13 Illustration of how the storage in a basin can attenuate an inflow peak hydrograph by extending the falling limb of the outflow hydrograph

Subsurface Runoff Processes

Urban soils are generally more compact than rural soils, are often construction-based fill materials, receive less infiltration, have artificial conveyance devices bisecting the subsurface environment, and in places, have on-site treatment and disposal systems (OSTDS) or septic systems that provide a replenished subsurface reservoir of drainage. In areas with relatively undisturbed soil matrices, the processes of subsurface flow behave according to that described by Dr. Jeffrey McDonnell (*see* **Chapter 112, Subsurface Stormflow, Volume 3** in this series). In disturbed areas, the subsurface moisture dynamics drain toward decreasing piezometric pressure and increasing suction, but the details have not been fully characterized. Lerner (2002) points out that the dynamics of urban recharge are poorly characterized, and that much more research is needed to quantify the impacts on subsurface water stores.

An important feature for designing artificial subsurface conveyance capacity is the diffuse addition of new water sources. Infiltration and inflow (I and I) that enter storm and sanitary sewers through poor joints, cracked pipes, and the walls of manholes are a cause for some system failures and subsequent surcharge (McGhee, 1991). Inflow enters perforated manhole covers, roof drains connected to sewers, and drains from flooded cellars. Inflow is associated with runoff events, while infiltration is drawn from the soil and may occur in dry weather. Inflow may be detected by smoke tests and video to find sources (Butler and Davies, 2000). Specifications for new sewer projects limit infiltration to $45\,L\,km^{-1}\,day^{-1}\,mm^{-1}$ diameter pipe, but older systems have been as high as $35\,000–115\,000\,L\,km^{-1}\,day^{-1}\,mm^{-1}$ diameter pipe (McGhee, 1991).

Collected Urban Runoff, Channel Flow, and Receiving Waters

Runoff from urban areas is engineered to collect within river channels that leave the urban area to abate local flooding (Mays, 1999) rather than within infiltration devices to increase groundwater recharge and delay the hydrograph peak (WEF/ASCE, 1998). Infiltration devices include detention and retention basins (AASHTO, 2000), swales, infiltration trenches, bioretention devices, and percolation or exfiltration trenches (USEPA, 1999a; USEPA, 1999b; USEPA, 1999c). Porous pavements are also increasingly used to encourage infiltration on-site (Urbonas and Roesner, 1993; AASHTO, 2000).

Channel Runoff Volume

Rose and Peters (2001) analyzed stream flow and precipitation data between 1958 and 1996 on how urbanization affected river runoff between stream flow and precipitation recorded for larger (hundreds of square kilometers) urban and nonurban watersheds in the greater Atlanta, Georgia

area. The authors found that annual runoff coefficients for the urban watershed, with more than 50% of its land cover classified as urban, were not significantly different from the less urbanized watersheds, but that for the 25 largest storm flows, peak flows were 30–100% greater, storm recession was 40–100% greater, and baseflow was 25–35% smaller in the urban watershed than in the nonurban watersheds, trends explained by increased impervious cover, shorter lag times, and less infiltration. Brun and Band (2000) found similar trends in baseflow decreasing by 20%, while runoff ratios stayed relatively constant during an 18-year longitudinal study of suburban growth in a 170 km^2 watershed northwest of Baltimore, Maryland. The authors did not test for the magnitude of increases in peak flow, which likely occurred to balance the water budget.

Channel Degradation

Urbonas and Benik (1995) review stream degradation and erosion patterns resulting from urbanization in semiarid regions and provide recommendations on management strategies to check erosion and return the stream to stable conditions. They illustrate how urbanized runoff to ambient runoff ratios increase from close to 1.5 for 100-year events to 3 for 5-year, and 50 for 2-year storms in Denver, Colorado. Damages to the urban waterway, both natural and man-made, from significant increases in event peaks and volumes include gullied channels, collapsing infrastructure, eroded riparian habitat, and scoured aquatic habitat.

Several researchers have recorded a variety of stream adjustments to cross-sectional area, width, and depth due to altered runoff processes within urban and suburban watersheds. In general these studies examined how unarmored streams responded at different rates and in different directions to the increased frequency of bank full flow, watershed erosion, and alternating patterns of in-channel aggradation and degradation (Lazaro, 1979). Walesh (1989) reports on how urbanization has encouraged the shortening of rivers, thereby increasing the channel slope and increasing discharge efficiency, as predicted by discharge equations such as Manning's formulas. Such alterations are typically designed to increase the efficiency of the floodway. In the United States, the National Flood Insurance Act of 1968 generally uses the 100-year floodplain to define the floodway zone.

Hammer (1972) studied streams in 50 suburban watersheds and 28 rural watersheds of 1 km^2 in the Philadelphia region and found an enlargement ratio of stream cross-sectional area that ranged from 1 to 7, increasing with valley slopes and the extent of stormsewers. Hammer (1972) noted that the changes in stream adjustment were detectable after 4 years of development, most pronounced for development between 4 and 30 years and less severe for development greater than 30 years, possibly due to losses in the deteriorating stormwater infrastructure and healthier pervious areas with larger trees and greater infiltration. Leopold (Leopold, 1973) noted during 20 years of urbanization for Maryland's 9.5 km^2 Watts Branch watershed, the average stream cross-sectional area decreased by a ratio of 0.8.

Receiving Waters

Urban receiving waters are predominantly streams and rivers, followed by estuaries and oceans, and then lakes, small ponds, and backwater areas (WEF/ASCE, 1998). The effects of stormwater runoff on stream ecosystems is to increase the magnitude and/or frequency of runoff events, increase surface runoff and reduce subsurface runoff, and increase storm flow velocity (WEF/ASCE, 1998). The morphology of the stream changes to have an increased cross-sectional area appropriate for the higher storm flows, accelerated down-cutting of the channel base if floodplain infrastructure prevents widening, increased sediment loads due to bank erosion and possibly watershed construction, modification of the streambed material to finer material, and increased slopes by straitening. The water quality impacts of stormwater discharge on receiving waters are not discussed in this section.

MODELS FOR URBAN AND SUBURBAN CATCHMENTS

Spatially Lumped Flow Paths

Lumped subcatchment areas delineated by stormwater drainage inlets rather than DEM pixels, are often used to collect and then deliver to the stormsewer. The use of this approach is rather common, and employed by the USEPA SWMM (Metcalf and Eddy Incorporated, 1971; Huber and Dickinson, 1992), and is illustrated in Figure 14. In this approach, the effective width of the contributing area determines both the flow path length and the time of concentration. Given this direct influence on hydrograph timing and the fact that the actual width and length arrangement of irregularly shaped subcatchment areas delivering runoff to the drainage inlet are not traditionally measured, the width parameter, W, is often adjusted to calibrate the model. Figure 15 shows several possible arrangements for a subcatchment area and how these arrangements influence the time of concentration for the stormwater runoff event.

Spatially Detailed Flow Paths

An alternative to this approach is to use explicit routing across each DEM pixel. The D8 (O'Callaghan and Mark, 1984), MF (Quinn *et al.*, 1991), DEMON (Costa-Cabral and Burges, 1994), and D-Infinity (Tarboton, 1997) algorithms process raster DEM data to identify local slopes and create flow direction maps charting a path between each pixel and one or more of its adjacent eight neighbors. Pixel-specific flow direction data are then passed to a standard routing algorithm that maps flow from its source to sink.

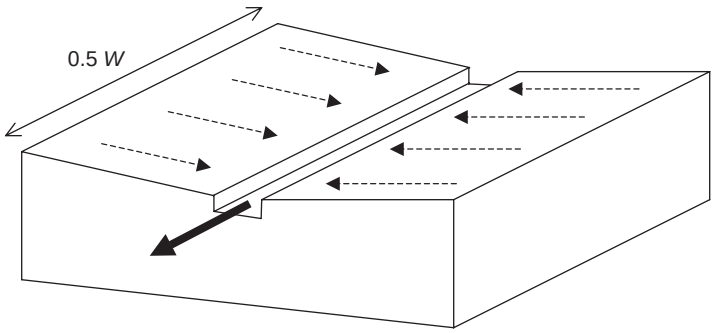
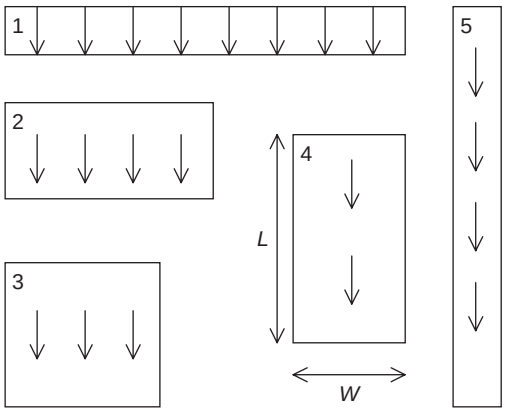


Figure 14 Aggregated subcatchment bisected by the street gutter drainage, where the total width of the contributing pixel is twice the street length. Runoff from the subcatchment is accumulated, but not simulated, as shown by the dashed arrows, and computed as the solid black arrow that goes to the drainage inlet



Shape	Area (ha)	W (m)	L (m)	t_c (s)
1	0.372	244.00	15. 25	3.7
2	0.372	122.00	30. 50	5.7
3	0.372	61.00	61. 00	8.6
4	0.372	30.50	122. 00	13.0
5	0.372	15.25	244.00	19.7

Figure 15 Subcatchment 1, 2, 3, 4, and 5 have the same areas of 0.372 ha, but different widths, *W* (greatest for 1 and least for 5), and lengths, *L*. Each area has a 1% watershed slope, 100% imperviousness, 0 cm of depression storage, a Manning's roughness of 0.02, a rainfall intensity of 25.4 mm h^{-1}, an equilibrium outflow of 0.026 $m^3 s^{-1}$, and a simulation time step of 300 s. The table illustrates how time of concentration, t_c, increases as *W* decreases

The D8 algorithm uses a "nearest neighbor" approach to compute elevation differences between the central pixel and its eight surrounding neighbors, and uses a steepest descent method to point to the neighbor joined by the steepest downward slope. Endreny and Wood provide a description of each algorithm and demonstrate its stability given vertical uncertainties in elevation estimates (Endreny and Wood, 2001).

Simulation of overland flow with D8, or single-sided flow in the path of steepest descent, requires an ordering scheme to route the water toward the outlet. Smith and Brilly (1992) developed an automated DEM grid-ordering algorithm for GIS-based routing to get connectivity network. The algorithm requires the filling of all sinks within the DEM, uses the D8 methodology of steepest descent to identify each grid's overland flow direction, delineates the watershed based on flow directions, identifies the number of neighbor grids that inflow into each DEM grid, flags all grids with no inflow as start grids, and flags all grids with two or more inflows as junction grids. The algorithm then makes many passes across DEM, tracing overland paths from start grids to the watershed outlet. For each start grid path, start grid and junction grid path segments are recorded. Start grid path segments begin with the start grid and end with the first detected junction grid. Junction grid path segments begin with a junction grid and end at the first down-gradient junction grid. A countervariable called *inflow* is assigned to each junction grid and increased each time the algorithm has traced an overland path through it. Listing of watershed paths proceeds by assigning the top left start grid a path number of one. When that path reaches the first junction grid, path number one terminates and path number two begins.

In the Smith and Brilly (1992) algorithm, all grids in each path are numbered based on the path number and sequence in the path, where the sequence begins at 1 and ends when a junction or outlet grid is reached. When the outlet is reached, the ordering continues at the next start grid. For this second and all subsequent passes through the DEM, if the junction grid has an inflow counter value of 1 or greater, a junction grid path segment is not created, but the inflow counter is increased. Once all start grids have been assigned a path number, they are ranked from 1 to N, where N is the number of start grids, for the connectivity network. Junction grids are then assigned a rank from $N + 1$ to $N + M$, where M is the number of junction grids, with the first rank assigned to the lowest inflow counter, and the last rank assigned to the highest inflow counter. For junction grids with identical inflow counters,

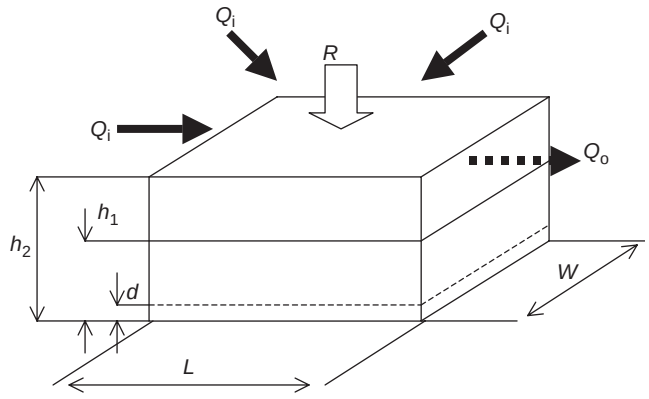


Figure 16 Cell dimensions, where the solid black arrows signify inflow from adjoining cells, the dashed black arrow signifies outflow, and the box white arrow signifies precipitation. Depth of depression storage is given as d, depth of water at the beginning and end of the time step is given as h_1 and h_2, respectively. Cell length is given as L, and width as W

rank is increased. Smith and Brilly (1992) demonstrated the utility of this connectivity network in an urban nonlinear reservoir stormwater simulation of the 9.4 hectare Gray Haven watershed near Baltimore, Maryland, where they closely matched the observed hydrograph.

Runoff Continuity

Smith (1993) developed an urban stormwater model that couples the continuity equation and Manning's turbulent flow equation for wide shallow channels to route overland flow to stormsewers, where a time shift approach rather than explicit dynamics is used. On the basis of Figure 16, a grid of width W (m), length L (m), is defined to have a depression storage depth, d (m), a water depth at the beginning of the time step, h_1 (m), with net rainfall R (m). The element has a known slope, S, Manning's surface roughness, n, and inflows, Q_i ($m^3\,s^{-1}$). Unknowns for each time step, t (s), are water depth at the end of the time step, h_2 (m), and outflow, Q_o ($m^3\,s^{-1}$). The continuity equation for the grid is written as

$$h_2 = h_1 + R + \left(\frac{\sum Q_i}{W \cdot L} - \frac{Q_o}{W \cdot L}\right) \cdot \Delta t \qquad (59)$$

$$Q_o = (h_2 - d)^{5/3} \cdot S^{1/2} \cdot \frac{1}{n} \cdot W \qquad (60)$$

Combining and rearranging the above equations gives a function of continuity,

$$f = -h_2 + h_1 + R + \frac{\sum Q_i}{W \cdot L} \cdot \Delta t - \frac{\left(\frac{h_2 + h_1}{2} - d\right)^{5/3} \cdot S^{1/2}}{n \cdot L} \cdot \Delta t \qquad (61)$$

The Newton–Raphson procedure is used to solve for h_2, which requires differentiation of the above equation.

$$\frac{df}{dh_2} = -1 - \frac{5}{6} \cdot \left(\frac{h_2 - h_1}{2} - d\right)^{2/3} \cdot S^{1/2} \cdot \frac{1}{n \cdot L} \cdot \Delta t \qquad (62)$$

$$h_{2est} = h_1 + 0.5 \Delta R \qquad (63)$$

$$h_{2est}^{new} = h_{2est}^{old} - \frac{f(h_{2est}^{old})}{\frac{df}{dh_2}(h_{2est}^{old})} \qquad (64)$$

$$\left|h_{2est}^{old} - h_{2est}^{new}\right| < \varepsilon \qquad (65)$$

The improvement achieved with each iteration of the Newton–Raphson procedure typically diminishes, and when an improvement increment is below the minimum set increment, ε, then the iteration stops.

Simulation Model Synthesis

Engineering models and methods for estimating urban runoff were presented throughout the previous sections; however, several key techniques that remain unidentified are introduced in this section. Beven provides more details on the range of modeling options for rainfall runoff simulation (see **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3** in this series). McPherson and Schneider (1974) articulate several difficulties in modeling urban watersheds, including supply and inflow volumes associated with groundwater pumping and drinking supply reservoirs, supply associated with rainfall at adequate temporal and spatial resolution to match time of concentration, nonconsumptive use patterns by residential, commercial, industrial, and municipal facilities, loss from abstractions such as infiltration, depression storage, interception in the urban environment, and routing through stormsewer networks and surface conveyance systems. It could be contended that the urban hydrological watershed is more complex than the rural watershed due to the layering of these additional issues over the traditional permeable hydrological landscape (Overton and Meadows, 1976).

Nix (1994) categorizes urban stormwater models as simple, simple routing, and complex routing based on varying data input demands and temporal and spatial output options. Simple models without routing of runoff are typically based on empirical relations, and include the Rational Method and TR-55 Curve Number Method. Simple routing, often referred to as hydrologic routing, utilizes continuity principles to move water across space and time, but at larger temporal and spatial scales than hydraulic models. The Linear Reservoir Method or Level

Pool Method is used for simulating the outflow hydrograph based on an inflow hydrograph through a flat-water surface such as a reservoir, while the Muskingum Method is used for simulating the outflow hydrograph based on an inflow hydrograph through a wedge of water such as a river flood wave.

TR-55 is a widely adopted engineering approach based on the Curve Number method from the NRCS (formerly the SCS) used to assess the runoff impacts of watershed change (McCuen, 1998; Chin, 2000). The TR-55 approach provides estimates of effective precipitation, P_e (cm), time to peak discharge, T_p (h), and peak discharge, q_p ($m^3\,s^{-1}$) based on watershed area, A (km^2), measured precipitation, P (cm), the duration of the precipitation event, D (h), time of concentration, t_c (h), and estimated land cover Curve Number values. The CN-based approach is empirically based, and an uncalibrated model intercomparison study in Denver, Colorado and Seattle, Washington (Zarriello, 1998) demonstrated its inability to capture observed runoff as well as mechanistic models. The TR-55 method is derived from empirical studies of 24-h rainfall events in primarily range, forested, and agricultural areas, but has been adapted for storms as short as 6 h and regularly used in simple engineering assessments of urban and suburban development change for estimating hydrological impacts. Curve Number values, presented in Table 7 for various cover types for typical antecedent moisture conditions, are critical for obtaining estimates of long storm duration storage, S, that represents initial abstractions, I_a, of interception and depression storage, and storm event infiltration. When antecedent soil moisture conditions are exceptionally dry or wet, the Curve Number values should be adjusted.

$$S = \frac{2540}{CN} - 25.4 \tag{66}$$

$$I_a = 0.2S \tag{67}$$

$$P_e = \frac{(P - I_a)^2}{(P - I_a) + S} \text{ for } P > 0.2S \tag{68}$$

$$T_p = 0.5 \cdot D + 0.6 \cdot t_c \tag{69}$$

$$q_p = \frac{2.11 \cdot A \cdot P_e}{T_p} \tag{70}$$

Pilgrim and Cordery (1993) provide a review of studies that cast doubt on the application of the Curve Number method, and note that the model predictions are highly sensitive to estimates of the time of concentration and the antecedent soil moisture condition, which are not entered directly but determine the values for model input parameters. Initial abstraction in the TR-55 method is set at 20% of long-term storage, which is likely an overestimate in urban runoff analysis, but has been partially adjusted for with the selected urban and suburban Curve Number values.

The Rational Method was developed for estimating urban peak runoff, q_p ($m^3\,s^{-1}$), based on an assumed uniform intensity precipitation, i ($mm\,h^{-1}$), event across the entire watershed area, A (km^2) and an estimated runoff coefficient, C. Runoff coefficients for urban and suburban areas are provided in Table 8. When the Rational Method is used in design for sizing stormwater conveyance components, estimates of the rainfall intensity are taken from Intensity-duration-frequency (IDF) curves for a given duration and frequency event, which show rainfall intensity decrease with increasing frequency and exponentially decay with increasing duration. The frequency for urban infrastructure design is traditionally specified in local ordinances between 2 and 25 years, most commonly at 10 years, while the duration is often taken as the watershed time of concentration, t_c, to achieve the equilibrium condition of the entire watershed contribution and the maximum peak discharge.

$$q_p = 0.278C \cdot i \cdot A \tag{71}$$

Time of concentration t_c estimates for an urban watershed are obtained from numerous methods, described in detail by Chin (2000), and are most frequently derived by separately analyzing and then summing overland flow travel times, t_i (h), and conveyance elements, such as gutters or stormsewers, and travel times, t_t (h). The kinematic-wave equation is shown below and provides an estimate for the time of concentration for overland flow based on the overland flow path length, L (m), the Manning's roughness factor, n, the design rainfall intensity, i ($mm\,h^{-1}$), and the slope of the overland flow path, S. Manning's equation provides an estimate for conveyance element velocities, which are then divided into flow path length to get the time of concentration for the conveyance element.

$$t_c = \frac{6.92 \cdot L^{0.6} \cdot n^{0.6}}{i^{0.4} \cdot S^{0.3}} \tag{72}$$

The SWMM (Huber and Dickinson, 1992) is a comprehensive open-source urban runoff model developed for the USEPA. SWMM has a wide consortium of users and several private software retailers who offer enhanced graphical user interface options, such as XP Software Inc. In addition to water quality routines, sanitary flows, storage, and treatment options in SWMM, the model offers stormwater flow routing characteristics that are a function of the simulation block selected. The three blocks – the Runoff, Transport, and Extended Transport – provide a range of hydrologic and hydraulic features, which are listed in Table 9. Backwater effects, characterized by the upstream element adjusting momentum terms based on downstream element conditions, are only available in the Extended Transport Block, while the Runoff and Transport Blocks upstream elements simply

Table 7 Runoff curve numbers based on hydrologic soil type (A is well drained and D is poorly drained) for suburban and urban land use for antecedent moisture condition II and initial abstraction equal to 0.2 storage

Land-use type		A	B	C	D
Lawns, parks, golf courses, cemeteries					
Good condition, grass cover >74%		39	61	74	80
Fair condition, grass cover 50–74%		49	69	79	84
Commercial or business area, 85% impervious		89	92	94	95
Industrial districts, 73% impervious		81	88	91	93
Residential					
Lot size	% Impervious				
<0.05 ha	65	77	85	90	92
0.10 ha	38	61	75	83	87
0.13 ha	30	57	72	81	86
0.20 ha	25	54	70	80	85
0.40 ha	20	51	68	79	84
Paved parking lots, roofs, driveways		98	98	98	98
Streets and roads					
Paved with curbs and stormsewers		98	98	98	98
Gravel		76	85	89	91
Dirt		72	82	87	89

Table 8 Typical rational method runoff coefficients for a 2-year to 10-year frequency design

Surface type	Runoff coefficients
Business	
Downtown area	0.70–0.95
Neighborhood area	0.50–0.70
Residential	
Single family Areas	0.30–0.50
Multiunits, detached	0.40–0.60
Multiunits, attached	0.60–0.75
Residential (suburban)	0.25–0.40
Apartment dwelling areas	0.50–0.70
Parks, cemeteries	0.10–0.25
Playgrounds	0.20–0.35
Streets	
Asphalt	0.70–0.95
Concrete	0.80–0.95
Brick	0.70–0.85
Drives and walks	0.75–0.85
Roofs	0.75–0.95
Lawns, sandy soil	
Flat, <2%	0.05–0.10
Average, 2–7%	0.10–0.15
Steep, >7%	0.15–0.20
Lawns, heavy (clay and silt) soil	
Flat, <2%	0.13–0.17
Average, 2–7%	0.18–0.22
Steep, >7%	0.25–0.35

provide inflow for downstream elements. SWMM guidelines recommend that the Transport or Extended Transport Blocks are applied to simulate larger diameter (>762 mm) pipes and trunk sewers, and that the Runoff Block can be used for smaller diameter (<762 mm) pipes and roadside curb and gutter flow in extremely small catchments (Huber and Dickinson, 1992).

Several other comprehensive urban runoff models have been developed, but are less widely used than SWMM. Chow and Yen (1976) evaluated the Illinois Urban Storm Runoff (IUSR) method as well as the Rational Method, Unit Hydrograph Method, Chicago Hydrograph Method, Transport and Road Research Laboratory Method (TRRL), Cincinnati Urban Runoff Model (UCUR), Dorsch Hydrograph Volume Method, and SWMM in their ability to capture volumes and flowrates for several storm events, where SWMM was found to be one of the most capable. Chen and Shubinski (1971) developed an urban stormwater runoff model that routes rainfall excess via the surface runoff system given a preconstructed connectivity matrix. At each rainfall time step, the model accounts for infiltration with Horton's exponential decay function, updates detention storage based on constant values for pervious and impervious areas, and then iteratively solves for continuity between depth and Manning's overland flow from each watershed element, which is then routed to gutter flow. Gutter input is then routed as Manning's concentrated flow by iteratively solving for gutter depth and accumulates as a storm hydrograph, where all gutters join the receiving water. Watt and Kidd (1975) developed and tested an urban stormwater model in Kingston, Ontario, with surface and sewer routing that subdivided the watershed

Table 9 Flow routing characteristics of runoff, transport, and extended transport blocks (Modified from Huber and Dickinson, 1992)

Simulation feature	Runoff block	Transport block	Extended transport block
Flow Routing method	Nonlinear reservoir, cascade of circuits	Kinematic wave, cascade of conduits	Complete continuity and momentum equations, interactive conduit network
Computational expense	Low	Moderate	High
Attenuation of hydrograph peaks	Yes	Yes	Yes
Time displacement of hydrograph peaks	Weak	Yes	Yes
In-conduit storage	Yes	Yes	Yes
Backwater or downstream control effects	No	No, unless as horizontal water surface behind a storage element	Yes
Flow reversal	No	No	Yes
Surcharge	Weak	Weak	Yes
Pressure flow	No	No	Yes
Branching tree network	Yes	Yes	Yes
Looped network	No	No	Yes
Number of preprogrammed conduit shapes	3	16	8
Alternative hydraulic elements (weirs, regulators, pumps)	No	Yes	Yes
Dry weather flow and infiltration generation (base flow)	No	Yes	Yes

by drainage inlets and for each subcatchment land cover was categorized as directly connected impervious area, street-side pervious areas, disconnected impervious areas or backyards, and noncontributing areas. From detailed measurement of subcatchment areas the model first estimates Horton infiltration and allocates depression storage, and then generates surface flow based on hydrograph convolution integrals for 1-min rain interval data. The hydrographs from the multiple subcatchments are time-offset, based on pipe lengths, and routed through sewers with a Manning's velocity equation to generate the watershed cumulative hydrograph.

FUTURE DIRECTIONS IN URBAN RUNOFF

Advances in Model Simulation

Stormwater simulation models, such as the federally supported SWMM (Huber and Dickinson, 1992) and Federal Highway Administration road and gutter computations (AASHTO, 2000) or the private industry Haestad Methods Software suite (e.g. StormCAD, CulvertMaster, PondPack) (Haestad, 2002), have been central to the analysis and design of gutters and curbs, stormwater basins, culverts, drainage inlets and catch-basins, stormsewer networks, and estimation of pollutant loads. A review of urban stormwater models (USEPA, 1997; Wang *et al.*, 2000; Zoppou, 2001) reveals that they range from simple peak discharge rate and total discharge volume methods (e.g. Rational Method and TR-55) to water quality models with complex unsteady wave propagation (e.g. SWMM) or continuous, mixed land use, semidistributed, hydrologic routing (e.g. Hydrologic Simulation Program – Fortran). In general, models that utilized standard engineering design equations are capable of predicting the runoff changes to the system, such as moving from a pasture to an urban lot as illustrated in Table 10.

A new generation of engineering ecology and ecohydrology stormwater models are needed, and might borrow in part from the forest ecosystem models, such as the

Table 10 Runoff differences between a 0.4 ha developed and undeveloped site (Data from Schueler, 1994)

Runoff quantity and quality	Parking lot	Meadow
Curve number	98	58
Runoff coefficient	0.95	0.06
Time of concentration (s)	288	864
Peak discharge 2 year/24 h ($m^3 s^{-1}$)	0.12	0.01
Peak discharge 100 year/toc ($m^3 s^{-1}$)	0.36	0.09
Runoff volume 1in storm (m^3)	97.77	6.18
Runoff velocity 2 year storm ($m s^{-1}$)	2.44	0.55

Regional Hydro-ecologic Simulation System (Band *et al.*, 1991). Several advances in urban stormwater simulation are already established, and provide a basis for the modeling advances in this research. Integration of stormwater models with geographical information system (GIS) data, such as the USEPA BASINS tool kit (Endreny, 2001; USEPA, 2001a), has allowed for several improvements that enable more detailed simulation of infrastructural controls and basic ecosystem processes. Such tools enable rapid evaluation of how variation in the spatial (30-m to 0.1-m pixel side) and radiometric resolution (panchromatic to color-infrared) of the increasingly available remotely sensed land cover GIS products affect simulation of urban watershed soil moisture dynamics and stormwater predictions (Endreny *et al.*, 2003). More detailed and accurate data sets of urban infrastructure, terrain, and land cover (vegetation and paved surfaces) are coming along to advance the stormwater model development. Rauch and Bertrand-Krajewski (2002) have developed REBEKA to provide deterministic and stochastic simulation of complex urban areas, thereby integrating the latest GIS-derived information with engineering algorithms.

Another development in urban modeling that has advanced ecosystem modeling was the creation of a variable source area (VSA) model, traditionally applied to forested areas. VSA-based runoff prediction, which captures the relation between subsurface and surface moisture dynamics, was applied to urban areas by Valeo and Moin (2000) as TOPURBAN, a modification of the TOPMODEL method from Beven and Kirkby (Beven and Kirkby, 1979). TOPURBAN adapted the topographic index to remove impervious surface contributing area (e.g. DCIA) and thereby disconnect it from subsurface discharge, routing it directly to the outlet. Field studies in suburbanizing watersheds with OSTDS have shown that drainage tile leaching predicts saturation more accurately than the topographic index estimates, and further research into urban ecosystem effects are needed.

Multimedia integrated modeling systems (MIMS) provide a framework for simulating the interconnections between stormwater dynamics with vadose zone, groundwater, and receiving water dynamics. MIMS theory can provide a basis for establishing linkages at the common boundary between the different media (Gilliland, 2000), and allow for long-term simulation of storm flow and low flow conditions. Initial developments in this area are being tested in the Neuse River Basin of North Carolina, linking the atmosphere, soils, rivers, and estuary (Peters-Liddard, 2000), but an urban application is not under way.

Ecological Stresses Associated with Urban Runoff

Ecohydrology is loosely defined as the mutual interaction between the hydrological cycle and ecosystems, but has been implemented primarily as a coupled set of climate–soil–vegetation dynamic equations that attempt to replicate soil moisture and plant patterns in space and time (Rodriguez-Iturbe, 2000; Eagleson, 2002). The set of linked equations used to simulate micrometeorological forcing and plant evolution adroitly synthesizes what Harte (2002) identified as disparate Newtonian and Darwinian worldviews of the previously uncoupled engineer and ecologist. To date, the ecohydrological simulations intentionally simplify some areas, by neglecting the activities of bacteria, fungi, and animals (including human management) and the constraints of soil chemistry, to focus on complexities of vegetation form and function. In its next step ecohydrology should address features of the urban environment, by simulating human activities such as stormwater management with the associated chemical constraints, such as road salt toxicity (Wegner and Yaggi, 2001) and the limits on sorption of urban metals (Davis *et al.*, 2003).

Rodriguez-Iturbe (2000; 2003) has identified a major application area for the new field as representing hydrologic control on ecological processes through simulation, where water may be a limiting factor due to scarcity or intermittent and unpredictable appearance. Water often has such a signature in urban environments (Collins *et al.*, 2000; Zhou *et al.*, 2002), making urban applications a natural extension for ecohydrological simulation. In the modeling applications, simulations capture years to decades of soil moisture dynamics, typically at a daily time step, which take on probabilistic patterns defined by the distributions of precipitation, evaporation, soil texture, and root growth observed in nature (Guswa *et al.*, 2002; Porporato *et al.*, 2002). Future work should address how the ecohydrological simulations might represent the shorter hydraulic and hydrologic time step of urban stormwater that determines allocation between runoff and infiltration.

Hart and Finelli (1999) provide a review of several decades of work that establishes how accelerated fluctuations in flow regimes in a stream channel impact dispersal, habitat use, resource acquisition, competition, and predator–prey interactions, arguing that flow is a master variable in stream ecosystems. Booker (2003) used a 3-D computational fluid dynamics (CFD) model, SSIIM, to simulate urban flows and their associated hydraulic forces in two reaches of the upper Tame River, UK to examine whether fish species' maximum sustained swimming speed was exceeded by an overpowering river with excessive shear. Paul and Meyer (2001) identified the amount of impervious surface as a predictor of stream ecological impacts, with replacement of vegetation along riparian corridors with urban structures and impervious cover having greater impacts than development farther removed from the stream.

Finkenbine *et al.* (2000) demonstrated in streams around Vancouver, BC, that even at 5% impervious cover fish survival was impacted, but with in-stream large woody debris and riparian habitat, stream fish communities began to re-equilibrate. The study did not explain the mechanisms for this trend, but noted the establishment of coarser material in urban streams as providing some fish habitat counterbalanced by an absence of large woody debris and fine sediments common in rural reaches. McMahon and Cuffney (2000) were able to use the National Water Quality Assessment (NAWQA) to identify a relationship between intensity of urbanization based on land cover, but not explicitly impervious surface, and biological, chemical, and physical water quality parameters, in a variety of geographic areas. Ecologically based urban hydrology research in Europe, such as that by Krebs and Larsen (1997) and findings from the German-hosted Urban Drainage Modelling conference series, is pioneering the use of sustainability metrics for urban drainage design, and signals new frontiers in stormwater investigations.

REFERENCES

AASHTO (1999) *Highway Drainage Guidelines (Metric Edition)*, American Association of State Highway and Transportation Officials (AASHTO): Washington.

AASHTO (2000) *Model Drainage Manual (Metric Edition)*, American Association of State Highway and Transportation Officials (AASHTO): Washington.

AASHTO (2002) *A Policy on Geometric Design of Highways and Streets*, American Association of State Highway and Transportation Officials: Washington.

Anderson D.G. (1970) *Effect of Urban Development on Floods in Northern Virginia*, US Geological Survey: Reston, Water Supply Paper No. 2001C. p. 22.

ASCE (1993) *Design and Construction of Urban Stormwater Management Systems: ASCE Manuals and Reports on Engineering Practice No. 77*, American Society of Civil Engineers: Reston, p. 724.

Baik J.J. and Chun H.Y. (1997) A dynamical model for urban heat Islands. *Boundary-Layer Meteorology*, **83**(3), 463–477.

Baik J.J., Kim Y.H. and Chun H.Y. (2001) Dry and moist convection forced by an urban heat Island. *Journal of Applied Meteorology*, **40**(8), 1462–1475.

Band L., Peterson D., Running S., Coughlan J., Lammers R., Dungan J. and Nemani R. (1991) Forest ecosystem processes at the watershed scale: basis for distributed simulation. *Ecological Modeling*, **56**, 171–196.

Bedient P.B. and Huber W.C. (2002) *Hydrology and Floodplain Analysis, ThirdEdition*, Prentice Hall: Upper Sadle River, p. 763.

Beven K. and Kirkby J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**(1), 43–69.

Booker D.J. (2003) Hydraulic Modeling of Fish Habitat in Urban Rivers During High Flows. *Hydrological Processes* **17**, 577–599.

Brater E.F. (1968) Steps toward a better understanding of urban runoff processes. *Water Resources Research*, **4**(2), 335–347.

Brazel A., Selover N., Vose R. and Heisler G. (2000) The tale of two Climates – Baltimore and phoenix urban LTER sites. *Climate Research*, **15**, 123–135.

Brun S.E. and Band L.E. (2000) Simulating runoff behavior in an urbanizing watershed. *Computers, Environment and Urban Systems*, **24**, 5–22.

Burian S.J., Nix S.J., Pitt R. and Durrans S.R. (2000) Urban wastewater management in the united states: past, present, and future. *Journal of Urban Technology*, **7**(3), 33–62.

Butler D. and Davies J.W. (2000) *Urban Drainage*, E and FN Spon: New York, p. 489.

Calder I.R. (1993) Hydrologic effects of land use change. In *Handbook of Hydrology*, Maidment D.R. (Ed.), McGraw Hill: New York, p. 50.

Carter R.W. (1961) *Magnitude and Frequency of Floods in Suburban Areas*, US Geological Survey: Reston, Professional Paper No. 424-B. p. 11.

Changnon S.A. (1984) Flooding on the increase in Illinois: a call for awareness and action. *Illinois Municipal Review*, **January**(7–8), `http://www.sws.uiuc.edu/docs/journals/3MS_2221AtmosSciPol.pdf`.

Changnon S.A. and Easterling D.R. (2000) U.S. water policies pertaining to weather and climate extremes. *Science*, **289**, 2053–2055.

Chen W. and Shubinski R.P. (1971) Computer simulation of urban stormwater runoff. *Journal of the Hydraulic Division, ASCE*, **97**(HY2), 289–302.

Chin D.A. (2000) *Water Resources Engineering*, Prentice Hall: Upper Saddle River, p. 750.

Chow T.V. and Yen B.C. (1976) *Urban Stormwater Runoff: Determination of Volumes and Flowrates*, US Environmental Protection Agency, Office of Research and Development, Municipal Environmental Research Laboratory: Cincinnati.

Collins J.P., Kinzig A., Grimm N.B., Fagan W.F., Hope D., Wu J. and Borer E.T. (2000) A new urban ecology. *American Scientist*, **88**(5), 416–425.

Congressional Budget Office (2002) *Future Investment in Drinking Water and Wastewater Infrastructure*, The Congress of the United States: Congressional Budget Office: Washington, p. 58.

Costa-Cabral M.C. and Burges S.J. (1994) Digital Elevation Model Networks (DEMON): a model of flow over hillslopes for computation of contributing and dispersal areas. *Water Resources Research*, **30**, 1681–1692.

CWP (1998) *Rapid Watershed Planning Manual*, Center for Watershed Protection: Ellicott, p. 312.

Davis A.P., Shokouhian M., Sharma H., Minani C. and Winogradoff D.A. (2003) Water quality improvement through bioretention: lead, copper, and zinc removal. *Water Environment Research*, **75**(1), 73–82.

Dewberry (2002) *Land Development Handbook* McGraw Hill: New York, p. 1124.

Dinicola R.S. (1989) *Characterization and Simulation of Rainfall Runoff Relations for Headwater Basins in Western King and Snohomish Counties*, U.S. Geological Service: Washington, WRIR 89–4052. p. 52.

Djokic D. and Maidment D.R. (1991) Terrain analysis for urban stormwater modelling. *Hydrological Processes*, **5**, 115–124.

Djokic D. and Maidment D.R. (2000) *Hydrologic and Hydraulic Modeling Support with Geographic Information Systems*, ESRI: Redlands.

Dow C.L. and DeWalle D.R. (2000) Trends in evaporation and bowen ratio on urbanizing watersheds in the Eastern United States. *Water Resources Research*, **36**(7), 1835–1843.

Doyle R. (2001) Sprawling into the third millennium. *Scientific American*, **284**(3), 25.

Eagleson P.S. (2002) *Ecohydrology: Darwinian Expression of Vegetation Form and Function*, Cambridge: New York, p. 443.

Endreny T.A. (2001) BASINS toolkit for hydrological monitoring, modelling, and assessment. *Hydrological Processes: Hydrology Today*, **16**(6), 1331–1335.

Endreny T., Hassett J.M., Hassett J.P., Mitchell D., Siegel M., Burns D. and Heisig P.M. (2002) Runoff timing and quality as a function of suburbanization. *AWRA, American Water Resources Association 2002 Annual Water Resources Conference*, Philadelphia, TPS-02-4, p. 328.

Endreny T., Somerlot C. and Hassett J.M. (2003) Hydrograph sensitivity to estimates of map impervious cover: a Win-HSPF BASINS case study. *Hydrological Processes*, **17**(5), 1019–1034.

Endreny T. and Wood E.F. (2001) Representing elevation uncertainty in runoff modelling and flow path mapping. *Hydrological Processes*, **15**, 2223–2236.

Epsey W.H. and Altman D.G. (1978) *Nomographs for Ten-minute Unit Hydrographs for Small Urban Watersheds: Addendum 3 of Urban Runoff Control Planning*, U.S. Environmental Protection Agency: Washington, EPA-600/9-78-035.

Fankhauser R. (1999) Automatic determination of imperviousness in urban areas from digital orthophotos. *Water Science and Technology*, **39**(9), 81–86.

Finkenbine J.K., Atwater J.W. and Mavinic D.S. (2000) Stream health after urbanization. *Journal of the American Water Resources Association*, **36**(5), 1149–1160.

Gibson L. (2003) Finding road networks in IKONOS satellite imagery. *Proceedings of ASPRS 2003 Conference Anchorage*, Alaska, p. 13.

Gilliland A. (2000) Introduction: MIMS ecosystem modeling component integration efforts. *Cross-Discipline Ecosystem Modeling and Analysis Workshop Research*, Triangle Park, p. 7.

Graff W.L. (1977) Network characteristics in suburbanizing streams. *Water Resources Research*, **13**(2), 459–463.

Guswa A.J., Celia M.A. and Rodriguez-Iturbe I. (2002) Models of soil moisture dynamics in ecohydrology: a comparative study. *Water Resources Research*, **38**(9), 510–515.

Haestad (2002) *Computer Applications in Hydraulic Engineering, FifthEdition,* Haestad Methods: Waterbury, p. 374.

Hamilton G.W. and Waddington D.V. (1999) Infiltration rates on residential lawns in central Pennsylvania. *Journal of Soil and Water Conservation*, **54**(3rd Quarter), 564–568.

Hammer T.R. (1972) Stream channel enlargement due to urbanization. *Water Resources Research*, **8**(6), 1530–1540.

Hart D.D. and Finelli C.M. (1999) Physical-Biological coupling in streams: the pervasive effects of flow on benthic organisms. *Annual Review of Ecology and Systematics*, **30**, 363–395.

Harte J. (2002) Toward a synthesis of the Newtonian and Darwinian worldviews. *Physics Today*, **55**(10), 29–34.

Huang C., Yang L., Wylie B.K. and Homer C.G. (2001) A strategy for estimating tree canopy density using landsat 7 ETM+ and high resolution images over large areas. *Third International Conference on Geospatial Information in Agriculture and Forestry*, Denver, p. 10.

Huber W.C. and Dickinson W.T. (1992) *Storm Water Management Model, Version 4, User's Manual*, U.S. Environmental Protection Agency: Athens, EPA-600-3-88-001a. p. 569.

Johnson P.A. (1996) Uncertainty in hydraulic parameters. *Journal of Hydraulic Engineering*, **22**(2), 112–114.

Jones N.L., Wright S.G. and Maidment D.R. (1990) Watershed delineation with triangle-based terrain models. *Journal of Hydraulic Engineering*, **116**(10), 1232–1251.

Kidd C.H.R. (1978) *Rainfall-Runoff Processes Over Urban Surfaces*, Institute of Hydrology, Proceedings International Workshop: Wallingford.

Krebs P. and Larsen T.A. (1997) Guiding the development of urban drainage systems by sustainability criteria. *Water Science and Technology*, **35**(9), 89–98.

Lazaro T.R. (1979) *Urban Hydrology: A Multidisciplinary Perspective*, Ann Arbor Science Publishers: Ann Arbor, p. 249.

Leopold D.J. (1968) *Hydrology for Urban Land Planning: A Guidebook on the Hydrologic Effects of Urban Land Use*, U.S. Geological Survey Circular 554, U.S. Geological Survey, Reston, p. 18.

Leopold L.B. (1973) River channel change with time: an example. *Geological Society of America Bulletin*, **84**, 1845–1860.

Lerner D.N. (2002) Identifying and quantifying urban recharge: a review. *Hydrogeology Journal*, **10**(1), 143–152.

Lo C.P. Quattrochi D.A. and Luvall J.C. (1997) Application of High-Resolution thermal infrared remote sensing and GIS to assess the urban heat Island effect. *International Journal of Remote Sensing*, **18**(2), 287–304.

Lunetta R.S. and Elvidge C.D. (1998) *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications*, Ann Arbor Press: Michigan, p. 318.

Maidment D.R. (1993) *Handbook of Hydrology*, McGraw-Hill: New York.

Mays L.W. (1999) *Hydraulic Design Handbook*, McGraw Hill: New York.

McCuen R.H. (1998) *Hydrologic Analysis and Design*, *Second Edition,* Prentice-Hall: Englewood Cliffs, p. 814.

McGhee T.J. (1991) *Water Supply and Sewerage*, McGraw Hill, p. 602.

McMahon G. and Cuffney T.F. (2000) Quantifying urban intensity in drainage basins for assessing stream ecological conditions. *Journal of the American Water Resources Association*, **36**, 1247–1262.

McPherson M.B. and Schneider W.J. (1974) Problems in modeling urban watersheds. *Water Resources Research*, **10**(3), 434–440.

Metcalf and Eddy Incorporated, University of Florida, and Water Resources Engineers (1971) *Storm Water Management Model*,

U.S. Environmental Protection Agency: Washington, EPA-11024 DOC 07/71.

New York State GIS Clearinghouse (2003) *New York State Statewide Digital Orthoimagery Program*, Office for Technology: NYS Cyber Security and Critical Infrastructure Coordination, `http://www.nysgis.state.ny.us/orthoprogram.htm`. Last Update.

Nix S.J. (1994) *Urban Stormwater Modelling and Simulation*, Lewis Publishers: Boca Raton.

Noto K. (1996) Dependence of heat Island phenomena on stable stratification and heat quantity in a calm environment. *Atmospheric Environment*, **30**(3), 475–485.

NYSDOT (2001) *Highway Design Manual: Typical Sections*, New York State Department of Transportation: Albany.

O'Callaghan J.F. and Mark D.M. (1984) The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics, and Image Processing*, **28**, 323–344.

Overton D.E. and Meadows M.E. (1976) *Stormwater Modeling*, Academic Press: New York.

Paul M.J. and Meyer J.L. (2001) Streams in the Urban Landscape. *Annual Review of Ecology and Systematics*, **32**, 333–365.

Peters-Liddard C.D. (2000) Applications of the TOPLATS land surface hydrology model, including strategies for modeling large river basins and coupling to saturated groundwater models. *Cross-Discipline Ecosystem Modeling and Analysis Workshop Research*, Triangle Park, p. 8.

Pictometry (2003) *See Everywhere, Measure Anything, Plan Everything*, `http://www.pictometry.com/news.asp`. Last Update.

Pilgrim D.H. and Cordery I. (1993) Flood runoff. In *Handbook of Hydrology*, Maidment D.R. (Ed.), McGraw-Hill: New York.

Pitt R., Chen S.E., Clark S., Lantrip J., Ong C.K. and Voorhees J. (2003) Infiltration through compacted urban soils and effects on biofiltration design. In *Practical Modeling of Urban Water Systems*, James W. (Ed.), CHI: Guelph, pp. 217–252.

Pitt R. and Lantrip J. (2000) Infiltration through disturbed urban soils. In *Applied Modeling of Urban Water Systems*, James W. (Ed.), CHI: Guelph, pp. 1–22.

Porporato A., Ridolfi L., Rodriguez-Iturbe I. and D'Odorico P. (2002) Ecohydrology of Water-Controlled ecosystems. *Advances in Water Resources*, **25**(8), 8–12.

Quinn P., Beven K., Chevallier P. and Planchon O. (1991) The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological Processes*, **5**, 59–79.

Rauch W. and Bertrand-Krajewski J.L. (2002) Deterministic modelling of integrated urban drainage systems. *Water Science and Technology*, **45**(3), 81–94.

Rodriguez-Iturbe I. (2000) Ecohydrology: a hydrologic perspective of climate-soil-vegetation dynamics. *Water Resources Research*, **36**(1), 3–9.

Rodriguez-Iturbe I. (2003) Hydrologic dynamics and ecosystem structure. *Water Science and Technology*, **47**(6), 18–24.

Rose S. and Peters N.E. (2001) Effects of urbanization on streamflow in the Atlanta area (Georgia, USA): a comparative hydrological approach. *Hydrological Processes*, **15**, 1441–1457.

Schueler T. (1994) The importance of imperviousness. *Watershed Protection Techniques*, **2**(4), 100–111.

Shepherd J.M., Pierce H. and Negri A.J. (2002) Rainfall modification by major urban areas: observations from space borne rain radar on the TRMM satellite. *Journal of Applied Meteorology*, **41**, 689–701.

Smith M.B. (1993) GIS-based distributed parameter hydrologic model for urban areas. *Hydrological Processes*, **7**(1), 45–61.

Smith M.B. and Brilly M. (1992) Automated grid element ordering for GIS-Based overland flow modeling. *Photogrammetric Engineering and Remote Sensing*, **58**(5), 579–585.

Steitz D.E. and Chandler L. (2001) *New Satellite Maps Provide Planners Improved Urban Sprawl Insight*, National aeronautics and space administration: Washington, p. 2.

Stone B.J. and Rogers M.O. (2001) Urban form and thermal efficiency: how the design of cities influences the urban heat Island effect. *Journal of the American Planning Association*, **67**(2), 186–198.

Tarboton D.G. (1997) A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research*, **33**, 309–319.

Thomas M. and Schmitt T.G. (2001) Usage of radar measured rainfall intensity distributions in urban drainage modeling. *World Water Environmental Resources Congress, Urban Drainage Symposium*, Orlando.

Urbonas B. and Benik B. (1995) Stream stability under a changing environment. In *Stormwater Runoff and Receiving Systems: Impact, Monitoring, and Assessment*, Herricks E.E. (Ed.), CRC Lewis Publishers: New York, pp. 77–101.

Urbonas B.R. and Roesner L.A. (1993) Hydrologic design for urban drainage and flood control. In *Handbook of Hydrology*, Maidment, D.R. (Ed.), McGraw Hill, New York, pp. 5.1–5.52.

USEPA (1997) *Compendium of Tools for Watershed Assessment and TMDL Development*, US Environmental Protection Agency: Washington, EPA 841-B-97-006.

USEPA (1999a) *Storm Water Technology Fact Sheet: Bioretention*, US Environmental Protection Agency: Office of Water, Washington, EPA-832-F-99-012. p. 8.

USEPA (1999b) *Storm Water Technology Fact Sheet: Infiltration Trench*, US Environmental Protection Agency: Washington, EPA 832-F-99-019. p. 7.

USEPA (1999c) *Storm Water Technology Fact Sheet: Vegetated Swales*, US Environmental Protection Agency: Washington, EPA 832-F-99-006. p. 7.

USEPA (2001a) *Better Assessment Science Integrating Point and Nonpoint Sources: BASINS Version 3.0 User's Manual*, U.S. Environmental Protection Agency: Office of Water (4305), Washington, EPA-823-B-01-001. p. 337.

USEPA (2001b) *Drinking Water Infrastructure Needs Survey: Second Report to Congress*, US Environmental Protection Agency: Office of Water, Washington, EPA- 832-R-97-003. p. 87.

Valeo C. and Moin S.M.A. (2000) Variable source area modelling in urbanizing watersheds. *Journal of Hydrology*, **228**(1–2), 68–81.

Viessman J.W., Knapp J.W., Lewis G.L. and Harbaugh T.E. (1977) *Introduction to Hydrology*, Harper and Row: New York.

Walesh S.G. (1989) *Urban Surface Water Management*, John Wiley & Sons: New York, p. 518.

Wang J., Hassett J.M., Endreny T.A. and McDonnell J.J. (2000) *Criteria for Selection of Models for Water Quality Management*

in Urbanizing Areas, College of Environmental Science and Forestry: Syracuse, p. 141.

Watt W.E. and Kidd C.H.R. (1975) QUURM - A realistic urban runoff model. *Journal of Hydrology*, **27**, 225–235.

WEF/ASCE (1998) *Urban Runoff Quality Management*, Water Environment Federation and American Society of Civil Engineers: Alexandria, p. 259.

Wegner W. and Yaggi M. (2001) Environmental impacts of road salt and alternatives in the New York city watershed. *Stormwater*, **2**(5), 24–31.

Yang L., Huang C., Homer C.G., Wylie B.K. and Coan M.J. (2003) An approach for mapping Large-Area impervious surfaces: synergistic use of landsat 7 ETM+ and high resolution imagery. *Canadian Journal of Remote Sensing*, **29**(2), 230–240.

Zarriello P.J. (1998) Comparison of nine uncalibrated runoff models to observed flows in two small urban watersheds. *Subcommittee on Hydrology of the Interagency Advisory Committee, First Federal Interagency Hydrologic Modeling Conference*, Las Vegas, pp. 163–170.

Zhou S.L., McMahon T.A., Walton A. and Lewis J. (2002) Forecasting operational demand for an urban water supply zone. *Journal of Hydrology*, **259**(1–4), 189–202.

Zoppou C. (2001) Review of urban storm water models. *Environmental Modelling and Software*, **16**(3), 195–231.

118: Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects

TIMOTHY P BURT[1] AND MICHAEL C SLATTERY[2]

[1]*Department of Geography, University of Durham, Durham, UK*
[2]*Institute of Environmental Studies, Texas Christian University, Fort Worth, TX, US*

The impact of agriculture on runoff processes cannot be underestimated. The greatest impact is usually on soil surface conditions, specifically infiltration. Undisturbed soils have much higher infiltration capacity than soils in agricultural land. Once cultivated, soils become easily compacted as they are laid bare to the elements until the next crop develops a protective cover. Infiltration can fall to levels where infiltration-excess overland flow can occur even in rainfall of modest intensity. Successive rainfall events compact the bare soil further and impermeable crusts often form. Heavy machinery used in agriculture exacerbates soil compaction, particularly at depth, via the development of plough pans. These compacted, dense layers have significantly higher bulk density and lower total soil porosity than the soil directly above or below it and frequently become preferential zones of runoff generation. Temporary saturation of shallow topsoil layers leads to topsoil saturation overland flow that, if connected to the channel network, may become a significant contributor to event response. Linear zones of compaction and reduced hydraulic conductivity that occur as a result of vehicular traffic have similar impacts.

CONTROLS ON STORM RUNOFF GENERATION

Whenever land is cleared for agriculture, the impact on local hydrology is likely to be immediate and obvious. The greatest impact is usually on soil surface conditions: deforestation exposes the soil surface to rainfall, and even where revegetation is rapid, the hydrology of the newly vegetated area, whether grass or shrubs, is very different from that of a mature forest canopy. Even if left to recover naturally, it can be several years before the system recovers to its previous state. Where present, grazing animals will compact the soil surface, which is then in a condition very different from the erstwhile open structure of the undisturbed soil. If the soil is being farmed for arable crops, the soil surface is continually laid bare to the elements, vulnerable until the next crop develops a protective leaf cover. Later improvement to farmland may further alter hydrological condition, for example, through the introduction of ditches or underground drainage. Compared to an area of, say, natural forest, agriculture affects the entire hydrological cycle: the water balance changes because the balance between evaporation and runoff is altered; usually, the land becomes more responsive to rainfall and storm runoff increases. If cultivation ceases, changes in broadly the opposite direction will occur.

The dominant controls of storm runoff generation are climate and soil, with topography as an important secondary variable at the subcatchment scale (Dunne, 1978; Anderson and Burt, 1990). Kirkby (1978) identified the crucial role of soil hydraulic conductivity in relation to rainfall intensity. He recognized domains dominated either by infiltration-excess overland flow, in places where infiltration capacity is low compared to rainfall intensity, or by a combination of subsurface stormflow and saturation-excess overland flow in places where rainfall intensities are low and the soil is relatively permeable. Of course, in some situations, certain soils will straddle the boundary between these two domains. As will be seen later, this will very often occur on farmland when deterioration of soil surface condition results in a much lower infiltration capacity than ought to be the case.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Runoff theories based on the occurrence of infiltration-excess overland flow were championed by Horton (1933, 1945), a hydraulic engineer who worked mainly in the heavily farmed regions of the American Midwest. Horton argued that the soil surface divides rainfall, one part flowing quickly to the stream as overland flow, while the other part infiltrates the soil, where it is stored, eventually draining to the stream or evaporating. Horton (1933) showed that for all soils, infiltration capacity fell during the course of a storm, eventually reaching a constant minimum level. High-intensity rain might produce overland flow immediately, while lower intensity rain would only produce overland flow once the initially high infiltration capacity had declined; low-intensity rainfall might produce no surface runoff at all. The Partial Area model of Betson (1964) remains the best guide to the location of runoff source areas for infiltration-excess overland flow. Horton argued that surface runoff would be widespread across an area, but this can only be the case if both rainfall intensity and infiltration capacity are uniform though space and time. In fact, overland flow is both unsteady and spatially varied since it is supplied by rainfall and depleted by infiltration, neither of which is necessarily constant with respect to time and location (Emmett, 1978). Betson stressed that overland flow is localized, generated from only certain parts of a drainage basin. In contrast to the Variable Source Areas described in the subsequent sections, partial source areas of overland flow appear relatively fixed in location within a given storm event.

Subsurface stormflow may be generated by two mechanisms: by non-Darcian flow through macropores or pipes; and by Darcian flow through the soil matrix. Horton assumed a semi-infinite soil where storage effects were unimportant, but in most soils, the character of the different horizons dictates that hydraulic conductivity declines with depth, often dramatically at a horizon boundary or where there is impermeable bedrock immediately beneath the soil (Note that this decline with depth is not always the case, for example, in areas of blanket peat or glacial till). Once percolating water begins to build within the soil, unless the ground is flat, soil water will begin to drain downslope. In appropriate circumstances, the subsurface response can be rapid enough to generate stormflow (Anderson and Burt, 1990). If soil water storage is limited so that the upper soil horizon becomes saturated, any excess rainfall must pond on the surface and will then produce saturation-excess overland flow; this can occur at rainfall intensities that are much lower than infiltration capacity. Soil saturation will tend to occur in relatively localized zones: in hillslope hollows; at the foot of slopes, especially where the slope profile is concave; and where the soil is thin. However, unlike the source areas for infiltration-excess overland flow, which are relatively fixed during a given storm event, source areas for saturation-excess overland flow can expand during a storm as the areas of soil saturation expand. On this basis, John Hewlett defined his Variable Source Area model, to describe surface runoff generation in areas where soil storage rather than infiltration capacity controls the occurrence of surface runoff. Whilst Hewlett's model was developed in the forested basins of the southern Appalachians, it is equally applicable in nonforested basins too. Note that the source areas for saturation-excess overland flow and subsurface stormflow are essentially identical since the same processes control the occurrence of both types of runoff.

PLOT SCALE: LAND USE AND OVERLAND FLOW ON AGRICULTURAL LAND

To elaborate on Horton's model for partitioning rainfall, when rainfall that has not been intercepted by vegetation reaches the soil surface, it will all infiltrate the soil unless the rainfall intensity exceeds the infiltration rate. In the latter case, only part of the rain infiltrates, part fills small surface depressions, and the remainder, if any, moves across the surface as overland flow. Infiltration rate is defined as the maximum rate at which water can enter into the soil. It is highly variable, depending on rainfall intensity, soil, vegetation, and slope gradient. At any given point, it will vary among storms depending on the soil-moisture content and the changing character of the soil surface, and will decrease during the course of a storm event. Infiltration rate is measured using a ring infiltrometer or a rainfall simulator (Rawls *et al.*, 1993); in neither case is the area of measurement usually very large compared to that of even a small drainage basin, so most reported values are in essence point values.

Given the importance of overland flow for erosion, there have been many studies of infiltration in agricultural soils; the effect of land use on infiltration and surface runoff is immediately apparent. For example, Table 1 shows results from an area of silty-clay loam soils. These soils have a naturally high infiltration capacity, but are easily compacted so that infiltration can fall to a level where infiltration-excess overland flow can occur even in rainfall of modest intensity. Several points are worth making in relation to the results presented in Table 1. Perhaps the most important is that undisturbed soils in woodland have much higher infiltration capacity than soils in agricultural land; intensity of land use correlates broadly and inversely with infiltration capacity. Measurements show that in this area infiltration rate can fall to a very low level if (wet) soils become compacted by mismanagement – by grazing or by the use of heavy machinery at inappropriate times. Bare soil is not necessarily the worst case, although, as will be seen later, the lack of vegetation cover does render the soil very vulnerable to overland flow and erosion. High infiltration rates may occur where tillage or shrinkage cracks have opened up the soil surface.

Table 1 The effect of land use on surface runoff at Slapton, Devon, England

Land use	Rainfall intensity ($mm\,h^{-1}$)	Infiltration rate ($mm\,h^{-1}$)	Surface bulk density ($g\,cm^{-3}$)
Temporary grass	12.5	12.3	0.96
Barley	12.5	11.0	1.08
Rolled, bare soil	12.5	4.0	0.93
Lightly grazed permanent pasture	12.5	5.9	1.12
Heavily grazed permanent pasture	12.5	0.1	1.18
Permanent pasture[a]		9 (range: 3–36)	
Freshly ploughed soil[a]		50	
Woodland soil[a]		180	

Data from Heathwaite *et al.* (1990), Burt *et al.* (1983), Burt and Butcher (1985).
[a]Results obtained using a rainfall simulator except where indicated.

Table 2 Final minimum infiltration rate in relation to soil texture. Based on information contained in Rawls *et al.* (1993)

SCS hydrologic soil group	USDA soil textures included in group	Final infiltration rate ($mm\,h^{-1}$)
A	Sand, loamy sand. Sandy loam	7.6
B	Silt loam, loam	3.8–7.6
C	Sandy clay loam	1.3–3.8
D	Clay loam, silty-clay loam, sandy clay, silty clay, clay	0–1.3

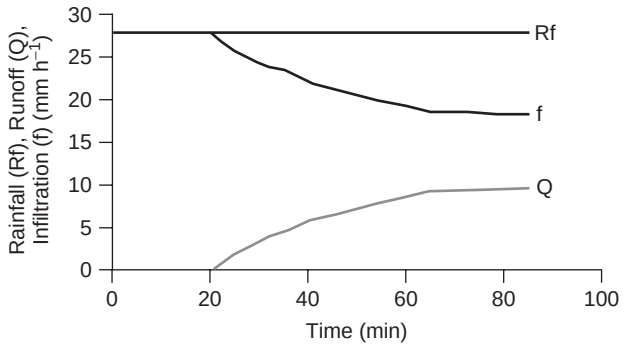


Figure 1 Results from a rainfall simulator experiment for a loess soil near Heidelberg, Germany, showing rainfall (Rf), runoff (Q) and infiltration (f) rates

Table 2 indicates the broad relationship between soil texture and final minimum infiltration rate. The hydrologic soil groups are those defined by the US Soil Conservation Service and the USDA (US Department of Agriculture) soil texture classes normally assigned to each group are indicated. It is clear that fine-grained soils have generally lower infiltration rates, unsurprising given the general link between soil texture and pore size. Whilst the values given can only be broadly correct, they do nevertheless indicate the order of magnitude for soil infiltration rates; the values suggest that, even in temperate regions, rainfall intensities can regularly exceed infiltration rate. It used to be argued that infiltration-excess overland flow was only commonplace in tropical regions, where very high rainfall intensities occur often, or in places where the soil had been mismanaged. However, many field observations over recent decades have shown that infiltration-excess overland flow can be commonplace in very many different circumstances and is by no means the rarity it was once claimed to be.

Figure 1 shows a typical infiltration curve for a loess soil; the results were obtained from a drip-type rainfall simulator (Bowyer-Bower and Burt, 1989). Rainfall intensity was constant throughout the experiment at 27.8 mm per h. No runoff occurs until 20 min after the start of the rainfall; before this point in time the infiltration rate would have been too high for ponding to occur. Once infiltration rate is exceeded, water ponds on the surface; soon after there is overland flow. After approximately 80 min there is equilibrium between rainfall and runoff, infiltration rate having fallen to a final, constant minimum. In this case, the decline in infiltration rate may just not be an effect of the soil becoming wetter, but also because the bare soil surface is compacted and sealed by the impact of raindrops; loess soils are particularly vulnerable in this respect. Imeson and Kwaad (1990, Figure 2) showed how, over the course of a growing season, infiltration rate could fall from an initially very high figure immediately after tillage to almost zero; successive rainfall events compact the bare soil and in between each event an impermeable crust forms. There is a "window of opportunity" for a crust to form while the soil is unprotected by vegetation cover: if the new crop grows quickly and there are few heavy falls of rain, a crust may not form before the soil surface becomes protected by leaf cover (30% cover is usually reckoned to be "safe", although it should be noted that this yardstick is without precise scientific foundation). However, once a crust is present, it may persist until the soil is ploughed again; overland flow can therefore be generated regularly over a large fraction of the crop cycle if a combination of heavy rainfall and poor crop growth prevails early in the growing season. This

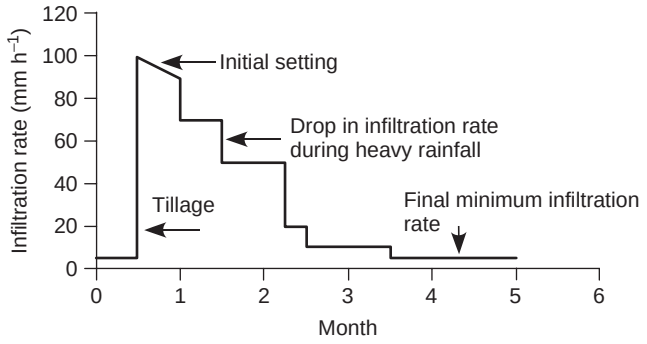


Figure 2 Change in infiltration capacity over a crop cycle for a soil with a tendency to crust (Reproduced from Imeson and Kwaad, 1990 by permission of John Wiley & Sons, Ltd)

shows the need for a new crop to become well established before the wet season and for periods of bare soil to be minimized where possible. Slattery and Burt (1996) showed that the crust can breakdown later on, disturbed by root growth or frost action.

Note that in some countries, frozen soils in agricultural areas can be a factor encouraging overland flow during several months of the year (Harris, 1972). This may be important, not only in relation to the generation of storm runoff, but also in controlling soil-moisture levels during and after the melt season (Granger *et al.*, 1984; Willis *et al.*, 1961). On the other hand, freezing conditions can also break up crusts, so that the infiltration capacity is much higher afterwards (Burt and Slattery, 1996). In some climates, duricrusts or iron pans may form at depth, in some cases limiting infiltration and in others percolation at depth (Goudie, 1985). The effect of compacted layers at depth is further explored in the next section.

TILLAGE IMPACTS AND THE GENERATION OF SATURATION-EXCESS OVERLAND FLOW

In modern agriculture, heavy machines used to pull implements, apply fertilizers, or harvest crops can create yield-limiting soil compaction. Certain tillage implements such as the moldboard plough and the disk harrow compact the soil below their working depth even as they lift and loosen the soil above. Repeated use of these implements at the same depth can form plough pans (often called *tillage pans*), dense zones immediately below the ploughed layer (Figure 3). Plough pans are generally relatively thin (3–6 cm), compacted soil layers having a significantly higher bulk density and a lower total porosity than the soil directly above or below it. Following 10 years of continuous corn tillage on a clay loam soil at Waseca, Minnesota (Bauder *et al.*, 1981), a dense compacted layer was detected

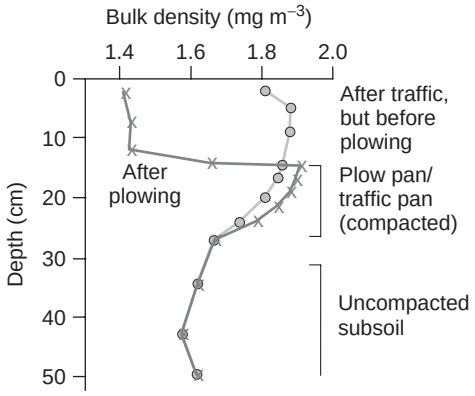


Figure 3 Change in bulk density with soil depth before ploughing (circles) and after ploughing (x's) showing the formation of a combined plough pan and traffic pan at 15–25 cm. The traffic pan is compaction at depth from tires of tractors and other heavy equipment. Bulk densities in excess of 1.8 Mg m^{-3} prevented the penetration of cotton roots in this case (after Voorhees, 1984). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

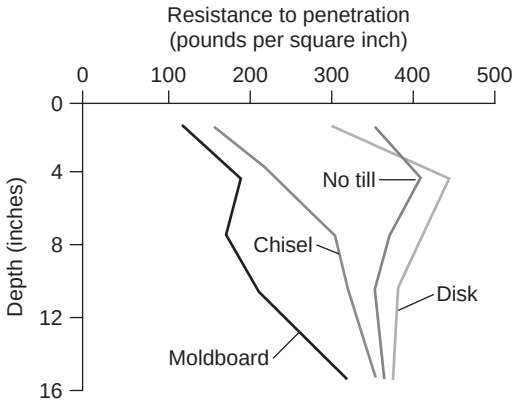


Figure 4 Soil resistance to penetration after 10 years of continuous tillage for four tillage systems (after Bauder *et al.*, 1981). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

just below the depth of tillage (4 inches) on the disk treatment (Figure 4). A cone penetrometer was used to register soil strength, and the greatest force required to penetrate the top 12 inches was measured on the plot that was spring-disked and had no other tillage. Other implements such as chisel ploughs do not press down upon the soil beneath them and thus are useful in breaking up plough pans and stirring the soil with a minimum of compaction.

While tillage pans may not have a significant effect on crop production, and can be alleviated by varying depths of tillage over time or by special tillage operations, they do become preferential zones of runoff generation, specifically as it relates to saturation-excess overland flow. This is

because plough pans are characterized by significantly decreased hydraulic conductivity. On the Texas High Plains, for example, infiltration studies reveal that some soils with a high bulk density plough pan will absorb only 6 cm of moisture in 24 h; similar soils without a plough pan will absorb more than three times the amount in the same period. In these situations, soil saturation begins to buildup above the plough pan; percolating water accumulates above the low-conductivity layer to form a temporarily saturated zone that is not connected to a regional groundwater flow (Figure 5). This situation is referred to as a *perched zone of saturation* (see Burt (1986) for a detailed description of the process mechanisms involved). Bergsma *et al.* (1984) called this *topsoil saturation overland flow*, as shown in Figure 5, where temporary saturation of a shallow topsoil layer occurs above the relatively impermeable pan. In this case, subsurface stormflow can be a significant contributor to event response, especially if it occurs in areas that are well connected to the stream. Saturation-excess overland flow may occur towards the foot of the slope if the subsurface discharge exceeds the transmissivity of the soil above the plough pan.

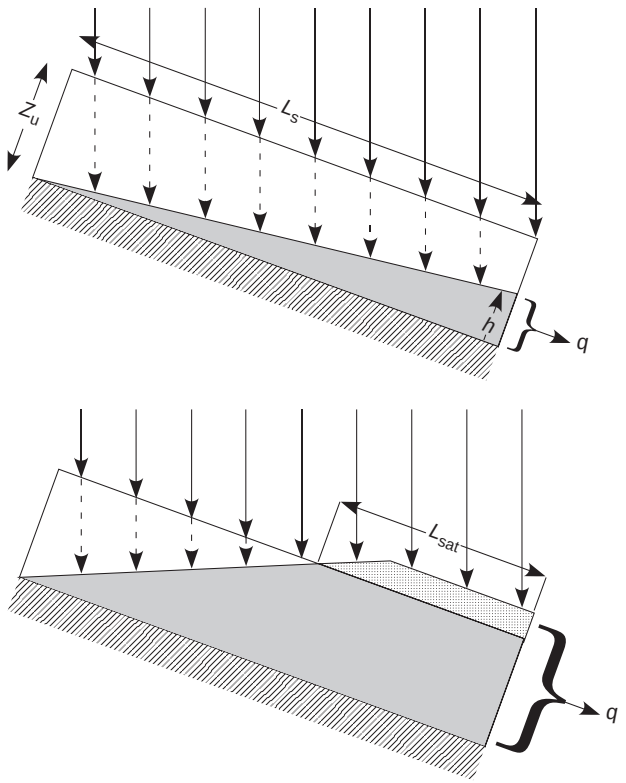


Figure 5 Cross section showing the formation of a perched saturated zone above low-conductivity layer, such as a plough pan, with constant slope and soil thickness. Z_u is soil depth, L_s is slope length, L_{sat} is the length of the zone of soil surface saturation, h is the depth of the saturated wedge, and q is the slope discharge

SCALING UP FROM HILL SLOPES TO CATCHMENTS

While plot studies have significantly enhanced our understanding of runoff generation, it is difficult to extrapolate the results from bounded plots to the scale of the field or catchment. Plots are hydraulically simple compared to the hydrological behavior of watersheds, and they fail to reproduce the complex topography that commonly characterizes both individual fields and whole catchments. Of greater significance, however, is the lack of *connectivity* inherent in bounded plots – that is, the study of runoff generation from a small plot tells us nothing about the delivery of runoff to the stream channel. A basic question central to our conceptualization of how catchments work remains simply this: how are runoff source areas linked to the stream?

In the context of agricultural catchments, features that help link slope and stream include: (i) linear zones of compaction and reduced hydraulic conductivity that occur as a result of vehicular traffic; (ii) convergent topography; and (iii) plough pans. However, it is also important to consider features that interrupt flow paths: these include hedgerows, ponds, and riparian buffer zones. Further discussion on landscape sensitivity in relation to runoff production can be found in Burt (2001).

While rills are generally located on valley-side slopes as a result of the hydraulic shear stress exerted by runoff, they often form along the base of depressions and dry valley bottoms where surface runoff concentrates because of ground surface convergence. These rills have been referred to as *ephemeral gullies, thalweg (or talweg) gullies, and valley-bottom rills* (Boardman, 1988; Poesen and Govers, 1990; Auzet *et al.*, 1993), and are features of many erodible agricultural landscapes, particularly in northwestern Europe, where they have been widely reported. Of course, it is not the agricultural activity *per se* that causes flow convergence in valley bottoms. Rather, it is the tractor wheelings (the compacted lines of soil formed by the passage of vehicles) that generates disproportionate volumes of surface runoff, which then converges in these valley bottoms and scours the soil surface down to the plough pan. Slattery *et al.* (1994), for example, describe an extensive thalweg rill system in a field in north Oxfordshire, UK, supplied with runoff and sediment by a series of "feeder rills" developed along the steeper valley-side slopes that had incised into the soil surface along vehicle wheelings. The importance of wheelings in contributing to runoff has been noted in many areas of the United Kingdom, such as the sandy soils of the West Midlands (Reed, 1983) and the silty soils of the South Downs (Boardman and Favis-Mortlock, 1992). In the North Oxfordshire study, the vehicle wheelings were characterized by very low infiltration levels, $5\,mm\,h^{-1}$ compared to $64\,mm\,h^{-1}$ on the adjacent cultivated surface (Slattery *et al.*, 1994). Wheelings were also shown to be significant zones of runoff generation (and sediment transport) on a

Figure 6 Concentrated runoff and erosion along vehicle wheelings on tobacco, North Carolina. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 7 Overland flow along a rill discharging directly into a drainage ditch on North Carolina coastal plain cropland. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

4.5-ha field planted to tobacco on the North Carolina coastal plain (Slattery *et al.*, 1997; Figure 6).

The thalweg rill system described by Slattery *et al.* (1994) developed on cultivated slopes that had been left bare throughout the summer months and drilled in late August with winter wheat. This produced smooth, highly crusted slopes that generated widespread surface runoff often under relatively low-intensity winter rainfall. In this case, runoff from the thalweg rill flowed into adjacent downslope fields but then infiltrated. In other instances, however, runoff from source areas apparently isolated from the stream network can become connected to the main river channel. This raises the complex issue of source-flow path connectivity and challenges current thinking on rainfall-runoff relations, particularly the rather simplistic, mechanistic notions underlying the variable source area and partial area concepts.

Slattery *et al.* (in press) demonstrated direct slope-channel connectivity, and the importance of compaction zones along vehicle tracks, in a study of runoff generation in a 19.4 ha watershed on North Carolina coastal plain cropland. These authors monitored a rill system draining cultivated cotton fields on low-energy slopes <0.003. The main rill, shown in Figure 7, incised along a vehicle track and discharged directly into the main channel, some 100 m from the basin outlet. The study was unusual because the slope runoff was measured during a tropical storm and two hurricanes. The slope and basin hydrographs from the Tropical Storm Arthur are shown in Figure 8. The storm generated 22.9 mm of rain over a period of 14 h, although 9.9 mm fell during a 10-min downpour between 16:48 and 16:57. Runoff was generated within the rill almost instantaneously, reaching a discharge of $28.9\,l\,s^{-1}$ at 16:53, with 4.8 mm of rain having fallen in the three minutes prior to

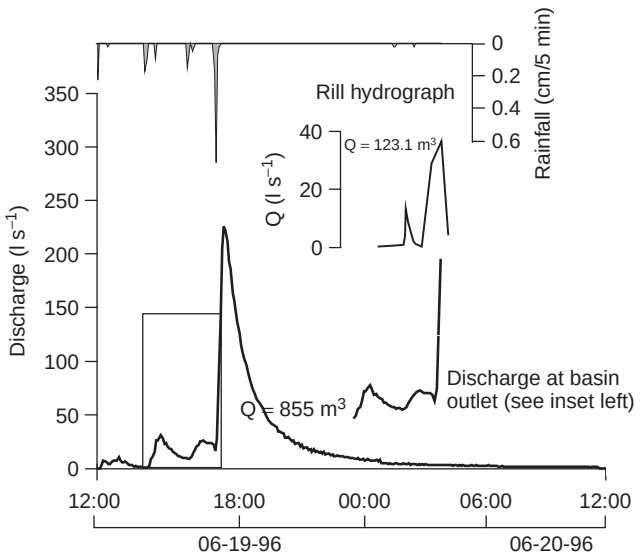


Figure 8 Comparison of hillslope and basin hydrographs during Tropical Storm Arthur in the Clayroot basin, North Carolina (after Slattery *et al.*, in press)

this measurement at a maximum intensity of $96.5\,mm\,h^{-1}$. Flow then quickly broke the banks of the rill channel and no further measurements were possible until the runoff had become reconfined within the rill walls. An estimated peak rill discharge of $44\,l\,s^{-1}$ at 17:08 corresponds to a discharge in the main channel at this time of $179.1\,l\,s^{-1}$ (peaking 7 min later at $224.9\,l\,s^{-1}$). In total, the rill on this field generated $123.1\,m^3$ of runoff, or 14.4% of the storm total for the entire basin. It is noteworthy that this single rill, with a total contributing area of <3% that of the basin area, contributed such a disproportionately large volume of runoff to the total event flow. As noted earlier, Betson (1964) first

formalized this view as the partial area concept, and in rural basins these areas are mainly roads and tracks plus the occasional field where a compacted soil surface causes surface flow. In either case, the total area involved is very often no more than 5% of the basin (Burt and Slattery, 1996).

The work referred to in the preceding section indicates that connectivity in catchments is clearly critical for determining the travel distances of overland flow and hence the amount of flow that exits the slope as channel discharge. If linear features such as roads and vehicle wheelings are fragmented, because of the presence of barriers within the landscape (Burt, 2001), the runoff will infiltrate a short distance downslope, as was the case in the North Oxfordshire thalweg rill example described earlier. Greater connectivity, on the other hand, allows the runoff to travel further and hence become stream discharge. The task then becomes one of elucidating hydrologic connections and disconnections across various scales within complex landscapes (McDonnell, 2003; Burt and Pinay, in press).

REFERENCES

Anderson M.G. and Burt T.P. (1990) Subsurface runoff. In *Process Studies in Hillslope Hydrology*, Anderson M.G. and Burt T.P. (Eds.), Wiley: Chichester, pp. 365–400.

Auzet A.V., Boiffin J., Papy F., Ludwig B. and Maucorps J. (1993) Rill erosion as a function of the characteristics of cultivated catchments in the North of France. *Catena*, **20**, 41–62.

Bauder J.W., Randall G.W. and Swan J.B. (1981) Effect of four continuous tillage systems on mechanical impedance of a clay loam soil. *Soil Science Society of America Journal*, **45**, 802–806.

Bergsma E., Charman P., Gibbons F., Hurni H., Moldenhauer W.C. and Panichapong S. (1984) *Terminology for Soil Erosion and Conservation*, International Society of Soil Science.

Betson R.P. (1964) What is watershed runoff? *Journal of Geophysical Research*, **69**, 1541–1542.

Boardman J. (1988) Severe erosion on agricultural land in East Sussex, UK, October 1987. *Soil Technology*, **1**, 333–348.

Boardman J. and Favis-Mortlock D. (1992) Soil erosion and sediment loading of watercourses. *SEESOIL*, **7**, 5–29.

Bowyer-Bower T.A.S. and Burt T.P. (1989) Rainfall simulators for investigating soil response to rainfall. *Soil Technology*, **2**, 1–16.

Burt T.P. (1986) Runoff processes and solutional denudation rates on humid temperate hillslopes. In *Solute Processes*, Trudgill S.T. (Ed.), Wiley: pp. 193–249.

Burt T.P. (2001) Integrated management of sensitive catchment systems. *Catena*, **42**, 275–290.

Burt T.P. and Butcher D.P. (1985) The role of topography in controlling soil moisture distributions. *Journal of Soil Science*, **36**, 469–486.

Burt T.P., Butcher D.P., Coles N. and Thomas A.D. (1983) Hydrological processes in the Slapton Wood catchment. *Field Studies*, **5**, 731–752.

Burt T.P. and Pinay G. Linking hydrology and biogeochemistry in complex landscapes. *Progress in Physical Geography*, in press.

Burt T.P. and Slattery M.C. (1996) Time-dependent changes in soil properties and surface runoff generation. In *Advances in Hillslope Processes*, Anderson M.G. and Brooks S.M. (Eds.), Wiley: pp. 79–95.

Dunne T. (1978) Field studies of hillslope flow processes. In *Hillslope Hydrology*, Kirkby M.J. (Ed.), Wiley: Chichester, pp. 227–293.

Emmett W.W. (1978) Overland flow. In *Hillslope Hydrology*, Kirkby M.J. (Ed.), Wiley: Chichester, pp. 145–176.

Goudie A.S. (1985) Duricrusts and landforms. In *Geomorphology and Soils*, Richards K.S., Arnett R.R. and Ellis S. (Eds.), GeorgeAllen & Unwin: London, pp. 37–57.

Granger R.J., Gray D.M. and Dyck G.E. (1984) Snowmelt infiltration to frozen prairie soils. *Canadian Journal of Earth Sciences*, **21**, 669–677.

Harris A.R. (1972) Infiltration rate as affected by soil freezing under three cover types. *Soil Science Society of America Proceedings*, **36**, 489–492.

Heathwaite A.L., Burt T.P. and Trudgill S.T. (1990) Land-use controls on sediment production in a lowland catchment, south west England. In: Boardman J., Foster I.D.L. and Dearing J. (Eds.), *Soil Erosion on Agricultural Land*, Wiley, pp. 69–86.

Horton R.E. (1933) The role of infiltration in the hydrological cycle. *Transactions, American Geophysical Union*, **14**, 446–460.

Horton R.E. (1945) Erosional development of streams and their drainage basins∼: hydrophysical approach to quantitative morphology. *Bulletin of the Geological Society of America*, **56**, 275–370.

Imeson A.C. and Kwaad F.J.P.M. (1990) The response of tilled soils to wetting by rainfall and the dynamic character of soil erodibility. In *Soil Erosion on Agricultural Land*, Boardman J., Foster I.D.L. and Dearing J. (Eds.), Wiley: Chichester, pp. 3–14.

Kirkby M.J. (1978) Implications for sediment transport. In: Kirkby M.J., (Ed.), *Hillslope Hydrology*, Wiley: Chichester, 325–363.

McDonnell J.J. (2003) Where does water go when it rains? Moving beyond the variable source area concept of rainfall-response. *Hydrological Processes*, **17**, 1869–1875.

Poesen J. and Govers G. (1990) Gully erosion in the Loam Belt of Belgium: typology and control measures. In *Soil Erosion on Agricultural Land*, Boardman J., Foster I.D.L. and Dearing J. (Eds.), Wiley: Chichester, pp. 513–530.

Rawls W.J., Ahuja L.R., Brakensiek D.L. and Shirmohammadi A. (1993) Infiltration and soil water movement. In *Handbook of Hydrology*, Maidment D.R. (Ed.), Mcgraw-Hill: pp. 5.1–5.51.

Reed A.H. (1983) The erosion risk of compaction. *Soil and Water*, **11**, 29–33.

Slattery M.C. and Burt T.P. (1996) On the complexity of sediment delivery in fluvial systems: results from a small agricultural catchment, North Oxfordshire, UK. In *Advances in Hillslope*

Processes, Anderson M.G. and Brooks S.M. (Eds.), Wiley: pp. 635–656.

Slattery M.C., Burt T.P. and Boardman J. (1994) Rill erosion along the thalweg of a hillslope hollow: a case study from the Cotswold Hills, central England. *Earth Surface Processes and Landforms*, **19**, 377–385.

Slattery M.C., Burt T.P. and Gares P.A. (1997) Dramatic erosion of a tobacco field at Vanceboro, North Carolina. *Southeastern Geographer*, **XXXVII**(1), 85–90.

Slattery M.C., Gares P.A. and Phillips J.D. Multiple modes of storm runoff generation in a North Carolina coastal plain watershed. *Hydrological Processes*, in press.

Voorhees W.B. (1984) Soil compaction, a curse or a cure? *Solutions*, **28**, 42–47.

Willis W.O., Carlson C.W., Alessi J. and Haas H.J. (1961) Depth of freezing and spring runoff as related to fall soil-moisture level. *Canadian Journal of Soil Science*, **41**, 115–124.

119: Land Use and Landcover Effects on Runoff Processes: Forest Harvesting and Road Construction

LA SAMPURNO BRUIJNZEEL

Department of Hydrology and Geo-Environmental Sciences, Vrije Universiteit, Amsterdam, The Netherlands

The hydrologic impacts of various forestry operations are reviewed, notably the effects of thinning, selective harvesting, clear-cutting with and without roads, and removal of understorey or riparian vegetation. Thinning and selective logging have a relatively modest effect on amounts of rainfall reaching the forest floor unless at least 70% of the standing biomass is removed. The effect on soil water and streamflow is smaller still, mostly because of the increased vigor of the remaining vegetation. Removing understorey vegetation produces a similarly small effect. The annual water yield is invariably increased in proportion to the amount of biomass harvested, with the exception of forests subject to high fog incidence or old-growth forest of low vigor. Cutting riparian vegetation produces a larger increase in streamflow (compared to cuts away from the streams) under subhumid conditions and shallow water tables only. The largest seasonal increases in flow after clear-cutting are normally observed at times when water deficits in the undisturbed forest are highest. The gains in streamflow after clear-cutting gradually disappear when regrowth is allowed, the effect lasting longer where soils are deeper and forest regeneration slower. Mechanized harvesting practices and road building both increase stormflow volumes and peak discharges. The relative effect diminishes with storm size in the case of clear-cutting only, but increases in the case of roads only. Similarly, effects of forest cutting per se on stormflows are "diluted" at larger scales, but the effect of the associated road network is not. Various measures are discussed to minimize the adverse hydrologic impacts of timber harvesting operations.

INTRODUCTION

Trees and forests are widely valued for timber and other forest products, for biodiversity, as well as for the general well being most people derive from being in their proximity. It has long been appreciated by both scientists and the general public that a forest cover influences micro climatic and soil conditions and therefore the amount of water flowing from forested areas. In the words of Zon (1927): "Of all the direct influences of the forest, the influence upon the supply of water in streams and upon the regularity of their flow is the most important in human economy". However, with populations rising explosively in some parts of the world, and per capita demands for water, timber, and other forest products increasing in others as living standards continue to improve, the pressure on the world's remaining forests is increasing steadily. In many areas, headwater forests serve as the traditional suppliers of high quality water and there is a widespread belief, particularly in the tropics, that mechanized logging operations or forest removal *automatically* results in increased flooding and massive soil degradation. Although this traditional view of the hydrologic functioning of forest came under (judicious) scrutiny in the early eighties of the last century (Hewlett, 1982; Hamilton and King, 1983; Bruijnzeel, 1986), there is a tendency among the current, more policy-, and socioeconomically oriented generation of forest hydrologic "myth busters" (e.g. Calder, 1999; Calder and Aylward, 2004; Kaimowitz, 2004; *see* **Chapter 187, Land Use Impacts**

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

on Water Resources – Science, Social and Political Factors, Volume 5) to overemphasize the greater water use of forests and to undervalue some of the more positive hydrologic aspects of a well-developed forest cover, such as improved infiltration and moderation of stormflow volumes in all but the most extreme cases, as well as protection against surface erosion and shallow landslides (Bruijnzeel *et al.*, 2004). When discussing the environmental impacts of logging or forest clearing, it is essential to take into account the specific climatic, topographic, and soil conditions in the area under consideration rather than apply blanket generalizations (Bruijnzeel, 2004). Given the current pressure to exploit the world's remaining forests, a sound understanding of the hydrologic functioning of forests (both pristine and "managed" ones) is arguably more important than ever.

This article aims to review the current state-of-knowledge with respect to the ways in which the main forest hydrologic processes are affected by various timber harvesting operations and related activities, such as thinning, selective logging, clear-felling, and road building. In doing so, the principal focus will be on the more humid parts of the world (both temperate and tropical), while paying particular attention to the changes in amounts and timing of streamflow. Hydrologic processes (including runoff generation) under conditions of snow and frost are dealt with in articles **Chapter 162, Hydrology of Snowcovered Basins, Volume 4**, **Chapter 114, Snowmelt Runoff Generation, Volume 3**, and **Chapter 71, Freezing and Thawing Phenomena in Soils, Volume 2**, whereas issues of water quality and biogeochemical cycles are discussed, *interalia*, in article **Chapter 93, Effects of Human Activities on Water Quality, Volume 3**. Before presenting the evidence, the article starts with a definition of terms (forest hydrologic processes, methods of evaluating changes in streamflows, timber harvesting techniques). It concludes with a summary of "best management practices" (BMPs) that are recommended if timber harvesting operations are to be compatible with the maintenance of hydrologic values and high quality water supplies (see Dykstra and Heinrich, 1996; Cassells and Bruijnzeel, 2004).

The Forest Hydrologic Cycle

The principal components of the forest hydrologic cycle are shown in Figure 1. Rainfall (P) is the most important input of water to most forests, with snow added at higher elevations and latitudes (*see* **Chapter 162, Hydrology of Snowcovered Basins, Volume 4**), and so-called "occult" precipitation (fog) in coastal or montane fog belts (*see* **Chapter 38, Fog as a Hydrologic Input, Volume 1**). A small proportion of the precipitation reaches the ground without touching the canopy ("direct" throughfall). Another (usually small) part flows along branches and stems as stemflow (Sf). A considerable portion of the precipitation hitting the canopy evaporates back into the atmosphere during and shortly after the storm (interception loss, Ei), while the remainder reaches the forest floor as crown drip. Because it is impractical to separate direct throughfall and crown drip, the two are usually lumped and simply referred to as *throughfall* (Tf). The sum of throughfall and stemflow is often termed *net precipitation* and normally substantially smaller than P, unless there are significant (and usually unmeasured) contributions by occult precipitation (CW; *see* Bruijnzeel, 2000; **Chapter 38, Fog as a Hydrologic Input, Volume 1**).

If the intensity of net precipitation falling on sloping ground exceeds the capacity of the soil to absorb the water, the excess runs off as "Hortonian" or "infiltration-excess" overland flow (HOF). Given the generally high absorption capacity of the organic-rich topsoil in most forests, this type of flow is rarely observed in undisturbed forest unless there is an unusually dense clayey substrate or an excessive concentration of stemflow. Not all of the water infiltrating into the soil emerges again as streamflow. A considerable part is taken up by the roots of the vegetation and returned to the atmosphere via transpiration (Et). The term evapotranspiration (ET) is often used to indicate the sum of transpiration (evaporation from a dry canopy), interception loss (evaporation from a wet canopy), and evaporation from the litter and soil surface (Es) (*see* **Chapter 42, Transpiration, Volume 1** and **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). The latter term is usually small, especially in dense forests where little radiation penetrates to the forest floor; humidity is maintained at high levels and ventilation is limited (*see* **Chapter 39, Surface Radiation Balance, Volume 1** and **Chapter 45, Actual Evaporation, Volume 1**).

If not obstructed by an impermeable layer in the soil, the water not taken up by the vegetation will move to the groundwater table via vertical and lateral downslope percolation and then flow laterally to the nearest stream as groundwater (Figure 1). Alternatively, it may be deflected upon meeting impermeable subsoil or rock. Such laterally draining "throughflow" helps to maintain the typically wet conditions found around headwater streams, thereby accounting for the stream's "baseflow". During rainfall, infiltrated water may take one of several routes to the nearest stream channel, depending on the soil's hydraulic conductivity profile with depth, slope form, and steepness, and the spatial distribution of soil moisture already present (Figure 2). "Saturation overland flow" (SOF) is caused by rain falling onto an already saturated soil, as typically occurring in hillside depressions or on concave footslopes near the stream (Figure 1). More widespread hillside SOF (i.e. outside the riparian zone and depressions) has been observed sometimes during intense rainfall in the

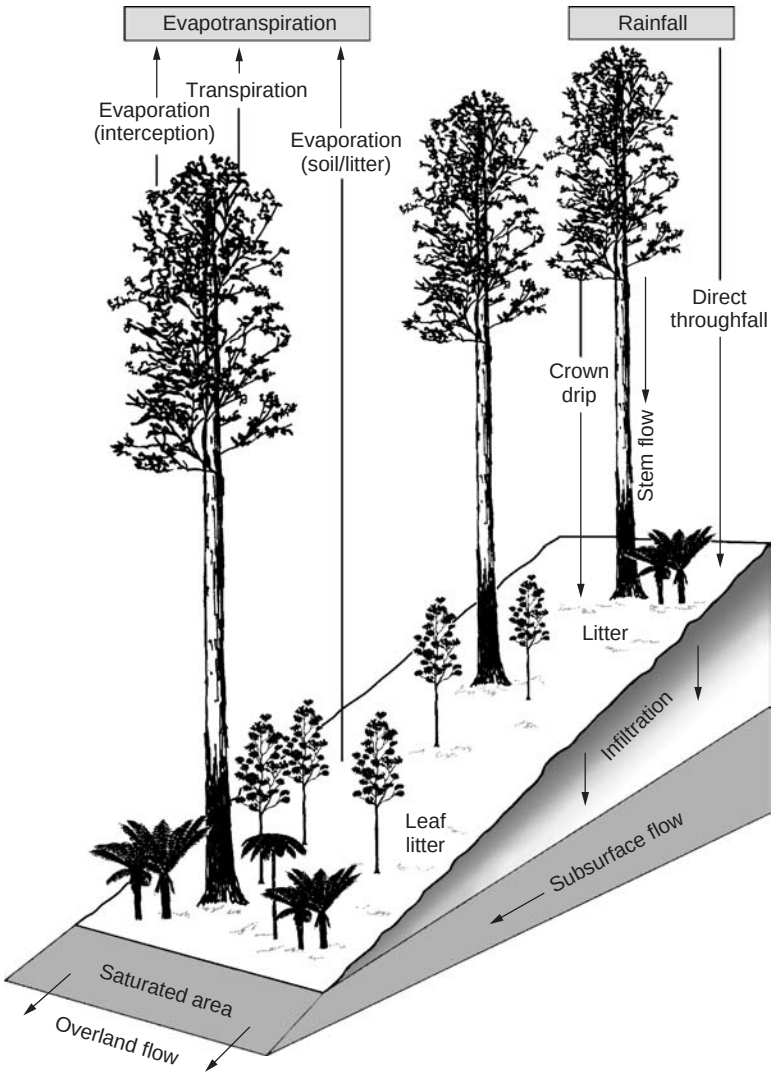


Figure 1 Key hydrological processes on a forested hillslope (after Vertessy *et al.,* 1998)

tropics in places with an impeding layer at shallow depth. Rapid throughflow during storms ("subsurface stormflow", SSF) usually represents a mixture of "old" (i.e. already present in the soil before the rain) and "new" water travelling through "macropores" and "pipes" (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**; **Chapter 112, Subsurface Stormflow, Volume 3**). Because of the contributions by SOF, SSF, and in extreme or disturbed cases also HOF, streamflow usually increases rapidly during rainfall. The increase above baseflow levels is called "stormflow" or "quickflow", whereas the highest discharge is referred to as "peak flow". Peak discharges may be reached during the rainfall event itself or as late as a few days afterwards, depending on catchment size and steepness, soil depth, and water content, and the duration, intensity, and quantity of the rainfall (*see* **Chapter 112, Subsurface Stormflow, Volume 3**). The total amount of streamflow discharged from a catchment area over a certain period of time (usually a month, season, or year) is called *water yield* (expressed in millimeters).

Evaluating Catchment Hydrologic Effects of Forest Disturbance or Removal

Within the limitations imposed by flow measurement techniques, the practical overall influence exerted by forests (or their manipulation) on hydrologic processes is arguably most clearly borne out by a comparison of streamflow amounts from catchment areas with contrasting proportions or types of forest. Forestry activities and natural disturbances both have the potential to more or less seriously alter forest evaporation and infiltration processes, and thus change the amount and timing of streamflow. Because of complications caused by local climatic contrasts (rainfall,

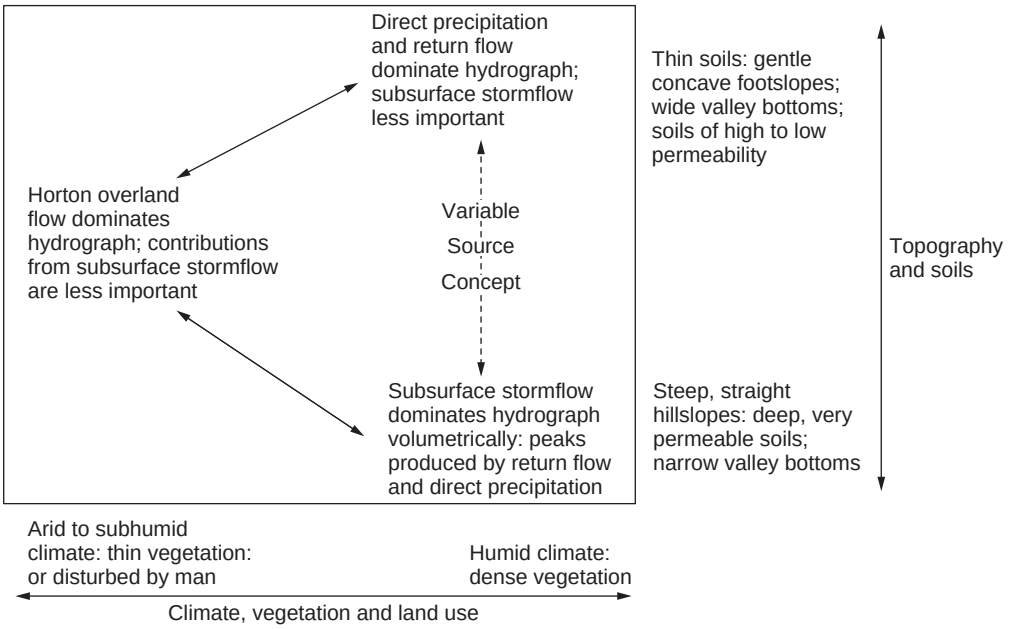


Figure 2 Schematic representation of the occurrence of various streamflow generating processes in relation to their major controls (Reproduced from Dunne, 1978, by permission of John Wiley & Sons Ltd). Note: ''direct precipitation and return flow'' are equivalent to saturation overland flow, *SOF*

exposure to radiation and air streams) and ungauged subterranean transfers of water from one catchment to another, a "direct" comparison of streamflows from catchments with contrasting covers can be problematic. The same applies to a comparison of flows from a single catchment before and after a change in cover. The classic response to such problems has been the "paired catchment experiment" in which the streamflow from two (preferably adjacent) catchments of comparable geology, topography, exposure, and vegetation are expressed in terms of each other (using regression analysis) during a "calibration phase". Once a robust baseline calibration relationship has been obtained, one of the catchments is subjected to manipulation of its cover (e.g. strip cutting or clear-felling) while the other catchment remains undisturbed as the "control" (Figure 3). Throughout this "treatment" phase, streamflow from both catchments continues to be monitored. Any effects of the treatment are evaluated by comparing the actually measured flow totals from the manipulated catchment with the flows that would have occurred if the catchment had remained undisturbed. This is usually achieved by inserting streamflow totals determined for the control catchment into the calibration relationship (Figure 3; Hewlett and Fortson, 1983). Although a more rigorous comparison between catchments is obtained in this way than in the case of a "direct" comparison of flows, the tacit underlying assumption is that differences in leakage between the two catchments remain unchanged with time, regardless of catchment cover status. Also, to avoid unjustified extrapolation of the calibration relationship to accommodate extremes in streamflow during the treatment phase (e.g. because of excessive rainfall or drought), it is imperative for the calibration period to include both wet and dry years. This makes the paired catchment method a time-consuming (typically >5 years) and thus an expensive affair. Moreover, the method is essentially a "black-box" requiring additional process research to reveal the relative importance of different causative factors to explain the observed changes in streamflow. All this, plus the limited resolution afforded by the paired catchment approach (usually more than 20% cover change is required for effects on streamflow to be detectable for small headwater catchments), has led to a general decline in the number of such studies in the last few decades and a gradually increasing emphasis on computer simulations (*see* **Chapter 132, Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change, Volume 3** and **Chapter 121, Intersite Comparisons of Rainfall-runoff Processes, Volume 3**).

Forest Harvesting Techniques

Harvesting of forest products may range from the manual collection of a particular component of the forest (such as fruits, fuelwood, or rattan), usually by forest dwellers or farmers living near the forest, through the removal of litter and topsoil for use as fertilizer or animal bedding, to the mechanized extraction of timber whose impacts are the main focus of this article. The hydrologic effects of the

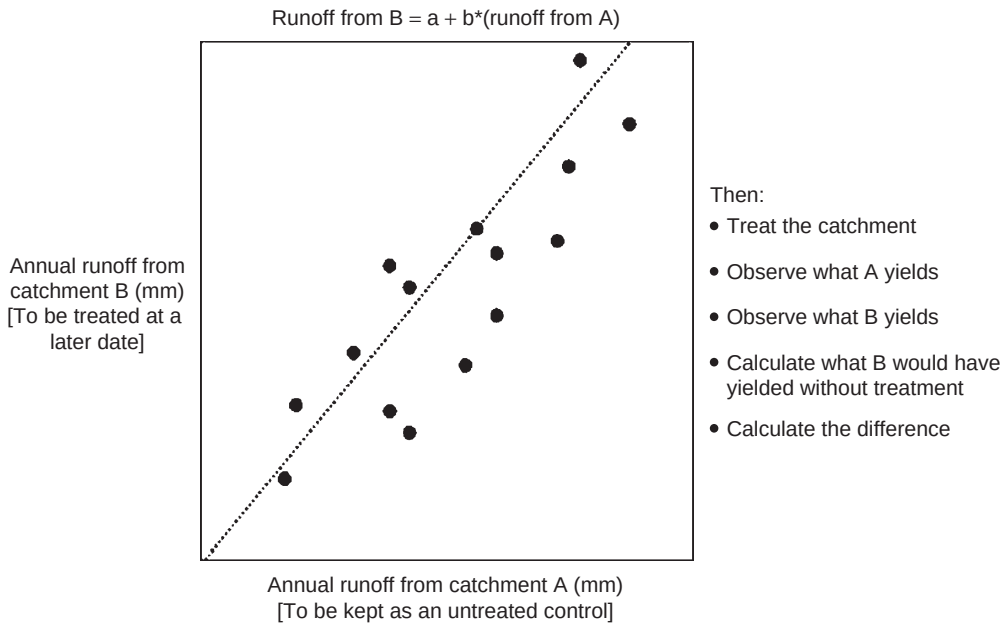


Figure 3 The paired catchment technique to evaluate the effect of land cover change on streamflow. Data shown represent flows as measured during the calibration period; the derived calibration relationship links the flows from the two catchments

various disturbances can be expected to be commensurate with the degree of damage inflicted upon the soil and standing vegetation. Whilst the manual collection of fuelwood is relatively harmless, the repeated removal of litter and topsoil (as widely practiced in South China and neighboring countries) may cause dramatic increases in HOF and surface erosion. Contrary to popular belief, it is the combined protection afforded by understorey vegetation and a well-developed litter layer rather than the main canopy that prevents the soil from being eroded. The erosive power of crown drip often exceeds that of rainfall in the open because of the associated increase in drop sizes, with the largest increases observed for large-leaved trees such as teak (Wiersum, 1985; Hall and Calder, 1993). Several years of uninterrupted litterfall and decomposition are required for the recovery of such a degraded forest floor (Zhou *et al.*, 2001).

Mechanized timber extraction not only opens up the canopy but also causes a number of disturbances to the soil surface. Disturbance occurs particularly during the construction and subsequent use of haulage roads, tractor tracks, and log landings, but also in the form of scars by uprooted and falling trees, especially in steep terrain. It is important to make a distinction between selective logging and clear-felling operations. In the former, damage to the residual vegetation needs to be kept to a minimum to ensure future timber yields, and hence much more of the vegetation is left standing. By contrast, upon clear-cutting, most (if not all) of the forest is felled and removed, with only very steep or wet parts of the terrain being left alone. In addition, logging debris in clear-felling operations may be chopped or "windrowed" by machinery and subsequently burned to gain access for the planting of the new rotation of trees. Up to 80% of the surface in small headwater catchments may be affected by such activities although at larger scales a value of 25 to 33% is more typical (Grip *et al.*, 2004). The (adverse) hydrologic effects of fire are discussed in **Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**.

In the context of selection logging, it has been estimated that for every harvested tree, another tree is killed and a third one damaged beyond recovery. Furthermore, the higher the intensity of the logging, the larger the proportion of the surface temporarily laid bare and greater the damage to the residual vegetation (although the relationship is not linear). In the richly stocked rain forests of Southeast Asia, where timber harvesting intensities of $100-150\,m^3\ ha^{-1}$ are not uncommon, up to 30% of the soil surface may be more or less seriously disturbed. In less intensively exploited forests, soil disturbance is closer to 5–10% (Bruijnzeel, 1992; Bruenig, 1996). Soil compaction by rubber-tyred skidders is much more pronounced than that by tracked vehicles and typically extends to a depth of 15–20 cm. Recovery of saturated hydraulic conductivity values (K*) upon regeneration of the vegetation on former tractor tracks is often slow and may, depending on the intensity of the previous use, require more than 50 years. HOF occurs much more frequently and in larger volumes on such compacted tracks, some of which may effectively

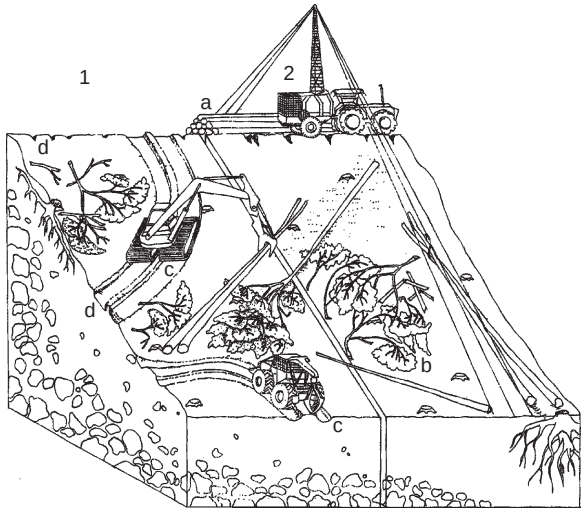


Figure 4 Timber harvesting generally uses heavy machinery such as caterpillars (1) or other tracked or wheeled vehicles; cable crane systems (2) generally have less impact on the soil. (a) Log landing on ridge top, (b) residual slash, (c) soil surface compaction and (d) sheet and rill erosion on tractor tracks (Reproduced from Mackensen *et al.*, 2003 Ambio No. 2, Vol.32, p. 107. by permission of Royal Swedish Academy of Sciences)

act as an extension of the regular drainage network (Grip *et al.*, 2004). Soil impacts during logging may be reduced by the manual skidding of logs on wooden rails in rare cases (e.g. in flat terrain in the tropics where labor costs are low) or by the application of aerial cable yarding techniques (Figure 4). However, because of the high level of damage to the residual vegetation associated with cable yarding, the method is more suitable for use in clear-cut operations. Helicopter-based logging is only rarely economically profitable (Bruenig, 1996; Dykstra and Heinrich, 1996).

HYDROLOGIC EFFECTS OF THINNING AND SELECTIVE LOGGING

In the following paragraphs, the hydrologic effects of (i) various forms of forest management (thinning and selective logging, removal of understorey or riparian vegetation, clear-cutting followed by regrowth or replanting) and (ii) road construction will be reviewed. The effects on the amounts of net precipitation are discussed prior to the effects on soil water, tree water use, and the amounts and timing of streamflow, both on an annual (or seasonal) basis and on an event basis. With a few exceptions, the discussion below is based primarily on experimental rather than simulated results.

Effects of Forest Thinning on Rainfall Interception and Net Precipitation

It has long been recognized that amounts of net precipitation tend to be inversely related to the stocking of a forest, both in the context of thinning or logging operations (Rogerson, 1967; Teklehaimanot *et al.*, 1991; Asdak *et al.*, 1998) and in terms of forest age (Delfs, 1955; Helvey, 1967). In other words, the denser the canopy the smaller the amount of rainfall reaching the forest floor and the greater the amount of rainfall interception. This not only reflects the greater leaf surface area of older stands but also their enhanced aerodynamic roughness (leading to increased atmospheric turbulence and wet canopy evaporation; *see* **Chapter 39, Surface Radiation Balance, Volume 1**). The effect can be substantial. For example, net precipitation in 60-year-old stands of white pine (*Pinus strobus*) in southeastern United States was ca. 220 mm year^{-1} less than in 10-year-old forest, despite the fact that stemflow decreased somewhat with forest age (Helvey, 1967). By contrast, no such decline with age was found for deciduous forests in the same area. Helvey and Patric (1965) concluded from the very similar interception totals derived for 11-year-old oak coppice and 30 and 50-year-old mixed yellow poplar-hickory forests that the younger stand had already acquired comparable leaf biomass and roughness characteristics to the older forests. A rather different picture has been obtained in evergreen mountain ash forests (*Eucalyptus regnans*) in Australia (Figure 5). Here, interception increased rapidly to a value of about 25% of the rainfall during the first 30 years, followed by a gradual decline to about 15% in old-growth forests (marking a difference of about 190 mm year^{-1}). Such changes in interception over time reflect concurrent changes in forest structure: during the first few decades, the regenerating forest consisted of numerous closely spaced trees with little understorey vegetation. In later years, the trees were much more widely spaced, whereas a well-developed understorey had been established as well (Haydon *et al.*, 1996).

Interception by deciduous forests during the dormant season is lower than during the growing season, but

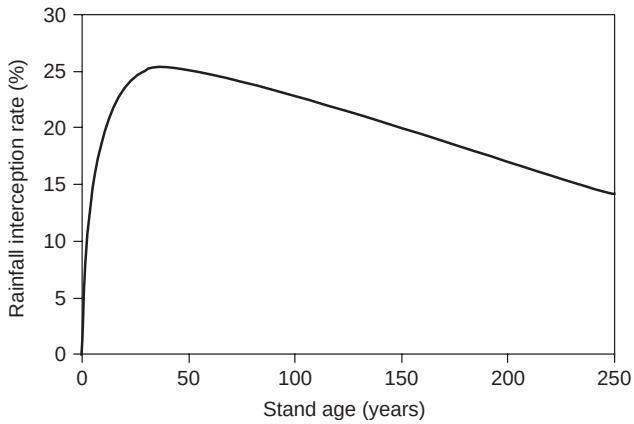


Figure 5 Relationship between mountain ash rainfall interception rate and stand age (Reproduced from Haydon *et al.*, 1996, by permission of Elsevier)

the typically observed increase in throughflow of 5–10% during winter is by no means proportional to the reduction in leaf area that can be up to sixfold (Helvey and Patric, 1965). Likewise, decreases in rainfall interception following forest thinning are often three to four times less than would be expected based on the degree of canopy opening alone. For example, a 50% reduction in basal area of a Douglas fir forest in France resulted in only a 13% reduction in interception (Aussenac *et al.*, 1982), whereas a 13-fold reduction in basal area in a dense Sitka spruce (*Picea sitchensis*) plantation (corresponding to a change in planting interval from 2×2 m to 8×8 m) in the United Kingdom was followed by a less than fourfold reduction in interception (Teklehaimanot *et al.*, 1991). Such findings cannot be explained entirely in terms of the actual decreases in intercepting canopy surface area but must also be ascribed to the gradual decrease in turbulent exchange between the trees and the surrounding air after opening up of the canopy. A strong negative correlation has been reported to exist between the aerodynamic conductance (g_a) of the boundary layer above the forest and the spacing of trees (Teklehaimanot *et al.*, 1991; *see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**).

Results from the humid tropics by and large confirm the above findings. Both Veracion and Lopez (1976) and Florido and Saplaco (1981) found increases in throughfall after thinning 30–50% of (natural) pine forests in the Philippines to be negligible. Significant increases were obtained only after as much as 70% of the trees had been removed. Asdak *et al.* (1998) found rainfall interception in undisturbed and heavily logged lowland rain forest in Indonesia to be 11% and 6% of incident rainfall, respectively. However, because the comparison did not involve the same forest before and after logging, it is not clear whether the reduction in interception primarily reflects the change in canopy cover or the preexisting structural differences between the two stands. Throughfall in tropical rain forests is notoriously variable in space and typically requires large numbers of (roving) gauges to narrow down the standard error of the mean throughfall estimate (Lloyd and Marques-Filho, 1988; *see* **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**). Elsewhere in Borneo, Chappell *et al.* (2001) reported no significant difference in throughfall beneath undisturbed rain forests and in logging gaps dominated by ca. 10-year-old pioneer vegetation. However, throughfall was significantly reduced in heavily logged-over forests that had been invaded by vines (Chappell *et al.*, 2001).

To summarize, the degree of forest thinning needs to be quite substantial (up to 70% of stand basal area) before a significant increase in net precipitation is achieved.

Effects of Thinning and Selective Logging on Soil Water, Transpiration, and Water Yield

Whilst the effect of thinning on amounts of precipitation reaching the forest floor is thus rather limited, the effect on soil water (and ultimately streamflow) can be expected to be even smaller. If the felled trees are left to decompose *in situ* and protect the soil against the impact of rainfall or crown drip, the slash is bound to intercept part of the gain in throughfall. More importantly, when opening up a stand the penetration of radiation to, and turbulent exchange with, the understorey vegetation and the forest floor are increased, thereby enhancing evaporation. In addition, the roots of the surrounding vegetation will start to compete for the extra moisture supplied by the increased throughfall in the newly created gaps. The magnitude and the duration of such effects will differ between locations, depending on the vigor of over- and understorey vegetation, climatic conditions (including site exposure and rainfall), and the configuration of the cutting (Stogsdill *et al.*, 1992; Parker, 1985; see Figure 6 below).

No change was detected in the streamflow from a deciduous hardwood forest catchment of southeasterly exposure at Coweeta (southeastern US) after a 27% reduction in basal area, and only a 4.3% rise in flows was observed after a 53% selective cut. Also, the removal of the entire understorey (representing 22% of the total basal area) from a 28 ha catchment of northwesterly exposure in the same area produced an equally modest change (Johnson and Kovner, 1956). Apparently, the moisture gained by removing one component of the forest is rapidly taken up by another. Further evidence for this contention comes from

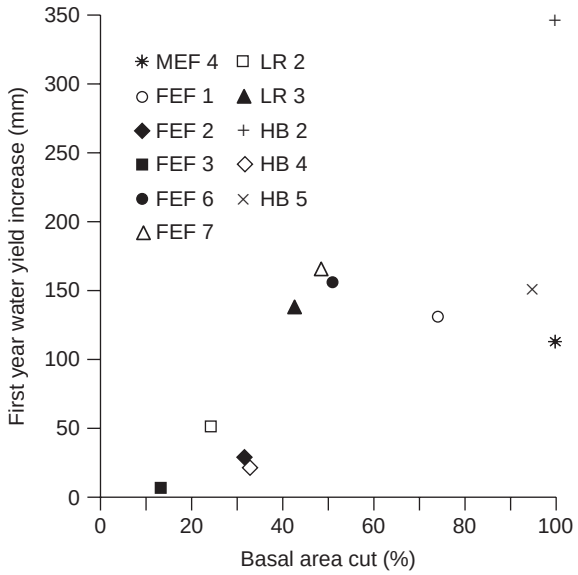


Figure 6 First-year increases in water yield in response to forest cutting in northeastern United States (after Hornbeck *et al.*, 1993, © Elsevier)

the Pacific Northwest of the United States where substantial soil water deficits tend to develop below the Douglas fir forest over the summer. Here, Black *et al.* (1980) observed "remarkably similar" transpiration rates for unthinned stands with little or no undergrowth and thinned stands with a well-developed understorey. In a related study, Kelliher *et al.* (1986) reported that, after removing the undergrowth, transpiration by the Douglas fir trees increased by 30–50%, with the greatest increases being found where undergrowth biomass had been highest. The net effect of the removal of the understorey vegetation on soil water content was insignificant. Whitehead *et al.* (1984) measured tree water uptake rates in two 40-year-old Scots pine (*P. sylvestris*) plantations of similar height, but with a more than fivefold difference in stocking in the northern United Kingdom. Transpiration in the more widely spaced plantation was on an average 67% of that in the denser stand. However, relative transpiration rates *per tree* were 3.3 times higher in the thinned plot, and intermediate in magnitude between the relative increases in average basal sapwood area per tree (2.9 times) and leaf area per tree (4.2 times), compared to the unthinned stand. Therefore, although water uptake by the thinned plantation had not attained prethinning levels yet, the large increases in both leaf and sapwood areas of the remaining trees may be viewed as to represent a tendency toward complete reequilibration following a set of tree physiologic relationships aimed at maximum site utilization and production (Whitehead *et al.*, 1984; *see* **Chapter 42, Transpiration, Volume 1**). Another striking example comes from South Africa where any potentially positive effects on streamflow of three rounds of heavy thinning (46%, 34%, and 50% at age three, five, and eight years) in plantations of *Eucalyptus grandis* were masked entirely by the steady reduction in flows resulting from the vigorous growth of the trees (Lesch and Scott, 1997).

Under wet tropical conditions, Gilmour (1977b) was unable to detect any changes in streamflow after "lightly" logging ($20\,m^3\ ha^{-1}$) a rain forest in northern Queensland, Australia. Although the removal of 33 and 40% of the commercial stocking in a rain forest in Peninsular Malaysia caused increases of 40 and 70% in annual water yield ($100–150\,mm\ year^{-1}$), there was no sign of the expected decline in streamflow gain during subsequent forest recovery (Abdul Rahim and Zulkifli, 1994). This result contradicts the more general experience that evaporation rates associated with young secondary growth in the humid tropics, although initially less than that of the surrounding forest, tend to increase very rapidly and may even exceed water use by old-growth forests within ca. 5 years (Figure 7; see Parker, 1985). Such findings illustrate the limitations of paired catchment studies in the absence of detailed process studies.

The influence of the *position or configuration* of the cutting on the magnitude and duration of any increases

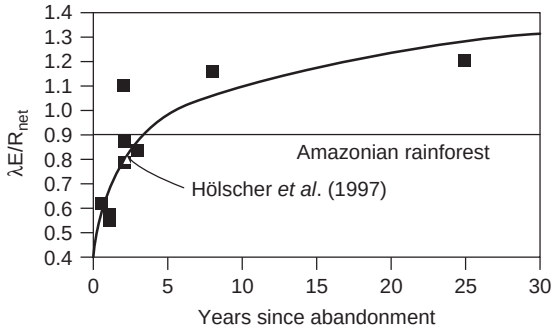


Figure 7 Ratio of energy used for evapotranspiration and net radiation in secondary vegetation as a function of years since abandonment of clearing for sites in eastern Amazonia and northern Thailand (Reproduced from Giambelluca, 2002, by permission of John Wiley & Sons Ltd)

in flow has been investigated in some detail. In eastern United States, the extraction of 24% of the basal area from catchment LR 2 at Leading Ridge (Pennsylvania) caused a nearly twofold larger increase in streamflow than cutting 33% on catchment HB 4 at Hubbard Brook (New Hampshire) or catchment FEF 2 at Fernow Experimental Forest (West Virginia) (Figure 6). The cutting at Leading Ridge consisted of a single block in the lowermost part of the catchment; at Hubbard Book it was effected through a series of strips situated half way up the catchment, while that at Fernow involved harvesting trees from all over the catchment (Hornbeck *et al.*, 1993). Therefore, increases in streamflow associated with strip cutting are smaller than for single blocks. This agrees with the suggestion of increased transpiration by surrounding trees after opening up of the canopy and the limited effect on net precipitation by thinning discussed earlier.

No significant differences in streamflow increases were observed between the cutting *en bloc* of the upper half of a catchment (such as in catchment FEF7 at Fernow, Figure 6) or the lower half (catchment FEF 6 in Figure 6) (Hornbeck *et al.*, 1993). Likewise, elimination of the riparian vegetation at Coweeta (southeastern US) did not produce an increase in water yield above that associated with the removal of an equal area of forest elsewhere in the catchment (Dunford and Fletcher, 1947). A comparable result was obtained for a catchment with similarly deep soils in the summer-rainfall zone of South Africa (Scott and Lesch, 1996), but not for an area in the winter-rainfall zone where streamflow gains after cutting pines and wattle trees in the riparian zone exceeded those associated with the harvesting of pines away from the stream by 30% (Dye and Poulter, 1995). Such contrasting results can be explained in terms of average soil water surplus or deficit, depth to the groundwater table, and slope morphology. Where rainfall is abundant, slopes steep and convex, and the water table is rather deep ($>3\,m$); no major spatial effect is

expected. Conversely, where soil water is scarcer, slopes gentle and concave, and depth to the water table is shallow; a more pronounced effect is possible. This is because trees located closer to the stream will have a better possibility to supplement soil water reserves during dry periods by their more ready access to the water table. Such trees can therefore be expected to maintain higher rates of water uptake than their counterparts away from the stream. These ideas were tested in a modeling experiment conducted for subhumid (MAP 700 mm) conditions, gentle slopes, and high water tables in southeastern Australia by Vertessy *et al.* (2003). Although the simulation aimed to assess the effect of forest planting at different positions in the landscape rather than forest removal, the results may be interpreted in an analogous manner. The predicted effect of tree planting differed markedly, depending on whether planting started at the top of the slope and moved progressively downward, or *vice versa*. The curves shown in Figure 8 suggest that under the prevailing subhumid conditions planting or removing trees in the lower 30% of the catchment would have a much more pronounced effect on flows than the same activity in the uppermost 30%.

Effects of Forest Clear-felling and Regrowth on Annual and Seasonal Water Yield

Well over 100 paired catchment studies have been carried out to date (mostly in the humid temperate zone) to examine the nature and extent of changes in streamflow associated with various forestry operations (mostly clear-felling). The literature on catchment treatment experiments (summarized by Bosch and Hewlett, 1982; Stednick, 1996; Best *et al.*, 2003) allows several generalizations to be made about the direction (increase or decrease) and approximate magnitude of the changes in flow. These are discussed briefly below.

Firstly, increases in total water yield during the first few years after treatment prove roughly proportional to percentage reductions in stand basal area the world over, provided the latter exceed a threshold of 20–25% (Figure 9) and rainfall remains above ca. 800 mm year^{-1} (see Figure 8). The reason for this is that tall forest vegetation evaporates significantly more water than shorter vegetation (such as pioneers in gaps) or a bare surface, partly because of differences in aerodynamic roughness and evaporating surface area and partly because of contrasts in rooting depth (*see* **Chapter 186, Water and Forests, Volume 5**). Because rainfall interception totals are higher in wetter years, the impact of forest clearance (i.e. near-zero interception) on flows also tends to increase with mean annual rainfall (Figure 10). Generally speaking, most of the increase in streamflow will show up in the form of increased baseflows unless the soil surface becomes seriously disturbed over a significant proportion of the catchment.

The database allows a tentative distinction between coniferous and deciduous hardwood forests, suggesting that increases in flow associated with the cutting of conifers are higher than for deciduous hardwoods (ca. 40 vs. 25 mm per 10% change in cover, respectively; Figure 9). The contrast is generally thought to reflect the dense evergreen habit and higher interception exhibited by coniferous forests, compared with deciduous forests that lose their leaves

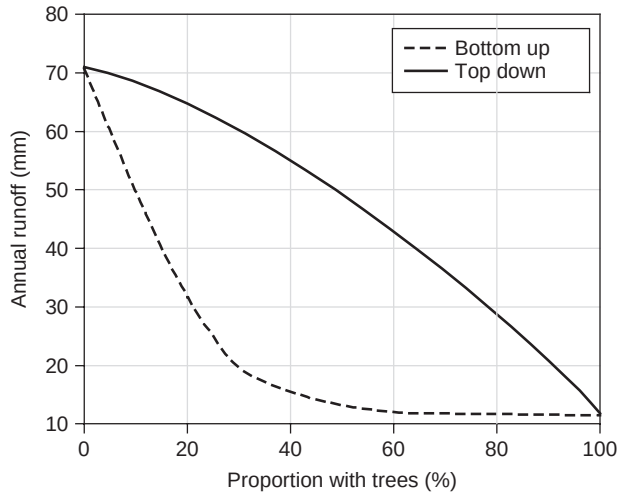


Figure 8 Results from a numerical modeling experiment showing two sets of predictions of annual streamflow after planting trees on a catchment under pasture in central New South Wales, Australia (mean annual rainfall 700 mm). The upper curve (solid line) shows changes in annual flow with forestation starting at the top of the catchment and progressing downslope. The lower curve (dashed line) shows the comparative response when forestation starts at the bottom of the catchment and progresses upslope (Reproduced from Vertessy *et al.*, 2003, by permission of Institute of Foresters of Australia)

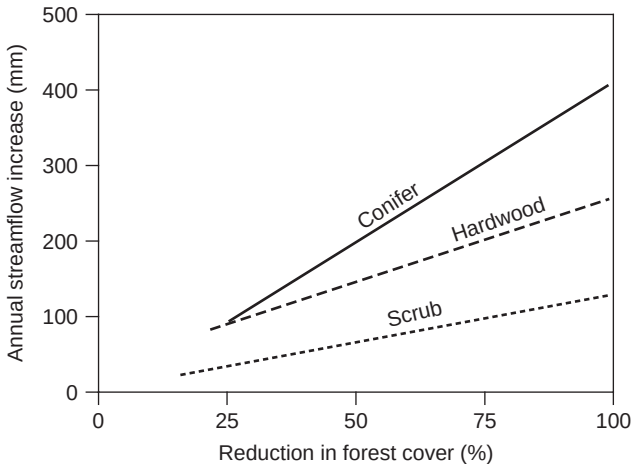


Figure 9 Relationships between reduction in forest cover and increase in annual catchment water yield. The general trend lines show the respective relationships for three types of woody vegetation (adapted from Bosch and Hewlett, 1982, © Elsevier)

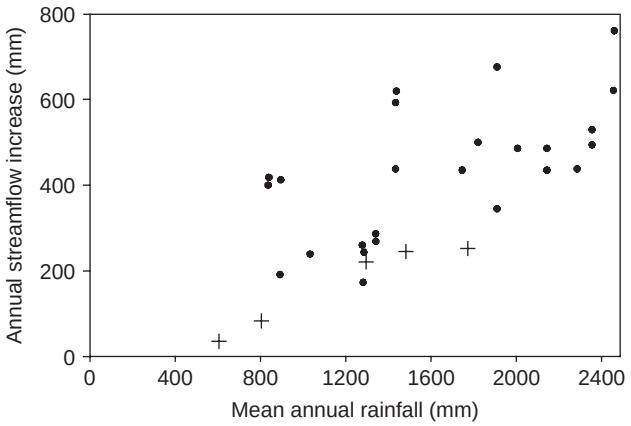


Figure 10 Effect of mean annual rainfall on increases in total water yield caused by total clearance of conifer and scrub vegetation (adapted from Bosch and Hewlett, 1982)

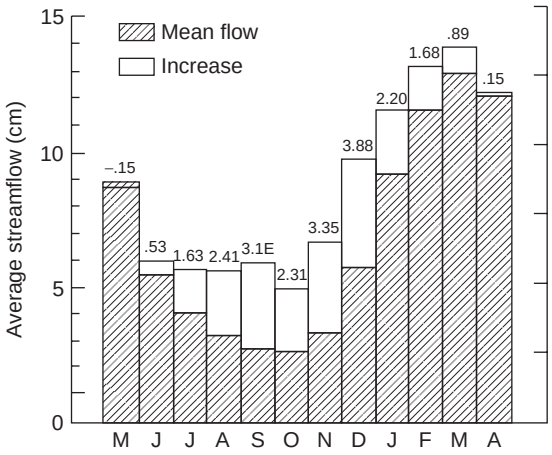


Figure 11 Mean monthly streamflow from Watershed 17 at Coweeta, southeastern United States before clear-cutting and increases during a 7-year period of annual recutting (Reproduced from Swank *et al.*, 1988, by permission of Springer-Verlag GmbH)

in winter or during an extended dry season. However, the scatter around the two tentative regression lines is large. The effect of the clearing of natural scrub (found in areas with low rainfall) is smaller still (ca. 10 mm per 10% cover change; Figure 9) and again highly dependent on actual rainfall conditions (Figure 10). Stednick (1996) approached the problem of the heterogeneity in results for 95 paired catchment studies from the United States by deriving separate regression equations for geographically homogeneous regions. In doing so, a clearer picture was obtained of the regional variation in yield increases that ranged from ca. 10 mm per 10% cover change in the Rocky Mountains to over 60 mm in the Central Plains. An even larger variation in initial flow increases after clear-felling (ca. 10–80 mm per 10% cover change) has been derived for humid tropical conditions. Partly this must reflect the larger differences in rainfall among different parts of the tropics although the combination of high intensity rainfall and extensive surface disturbance (leading to a larger stormflow component, see below) may also play a role (Bruijnzeel, 2004). One of the few exceptions to the general rule of increased streamflows after forest harvesting is the finding of Ingwersen (1985) that summer flows in a fog-ridden part of the Pacific Northwest of the United States declined instead of increasing after a patch clear-cut operation. The effect disappeared after 5 to 6 years when the regenerating trees had become tall enough again to capture sufficiently large amounts of water from the passing clouds. Another exception relates to the situation in which old-growth forests of low vigor and water use are replaced by young, vigorously growing forest (see Figure 7; Vertessy *et al.*, 2001; Giambelluca, 2002).

Apart from contrasts in rainfall, tree species, and age, part of the observed variability in the change in streamflow after clear-cutting relates to differences in catchment exposure and soil depth. The importance of exposure is illustrated by the difference in first-year streamflow gains after cutting differently exposed deciduous hardwood forest catchments at Coweeta. Flows from southerly exposed catchments increased by about 130 mm year^{-1} but increases from northerly exposed catchments sometimes exceeded 400 mm year^{-1} (Swank *et al.*, 1988). As such, the effect of clearing is inversely proportional to solar radiation inputs. As for the effect of soil depth, Trimble *et al.* (1963), discussing differences in duration of the changes in streamflow following forest clearing in West Virginia, suggested that: "the deeper the soil the longer it takes roots of new growth or the expanding root systems of residual plants to occupy the soil mantle. When the soil mantle is reoccupied, transpiration reaches maximum levels again, and increases in streamflow disappear". However, changes in forest structure (main canopy vs. undergrowth biomass and leaf area) and tree physiology with forest age (such as reduced vigor) may also play a role (see Figure 7; Giambelluca, 2002; Vertessy *et al.*, 2001). The role of soil depth is illustrated further by an examination of the timing of the maximum and minimum increases in flow after forest clearance during the year at Coweeta (Figure 11). Here, the ample rainfall is distributed evenly over the year, and the loamy soils are very deep (up to 6 m) and thus capable of holding large amounts of moisture. Under such conditions, the soil is not fully recharged until early spring (April–May) after which the presence or absence of a forest has little effect on the amount of flow (the soil being at or above field capacity anyway). However, during the summer, the contrast in streamflow from forested and cleared areas increases as the growing season advances. The effect continues well into the winter months (November–January), even though the trees have lost their leaves by then and evaporation is much reduced.

This reflects the development of the much larger soil water deficit during summer below forests that take an additional two months (February–March) to be recharged again by the winter rains (Figure 11). By contrast, in areas with shallow or sandy soils (having less water storage and lower water retention capacity), soil water recharge and depletion patterns alternate much more rapidly and increases in flow after forest clearance follow the pattern of forest water use much more closely (Hornbeck *et al.*, 1970).

The effect of the *timing of the cut* and the importance of controlling regrowth are demonstrated by a comparison of the results obtained for catchments HB2 and HB5 at Hubbard Brook (Figure 6). On catchment HB 2, the forest was clear-cut during the dormant season and regrowth was eliminated by herbicide application the following spring. Together, this caused a first-year increase in water yield of nearly 350 mm. On catchment HB 5, the removal of the trees took nearly a full year, while regrowth was not suppressed. The resulting first-year increase in flow was only ca. 150 mm, or 44% of that observed for catchment HB 2 (Figure 6). In addition, the gain in streamflow had largely disappeared after three years (Hornbeck *et al.*, 1970, 1993). Such contrasting findings again illustrate the effect of vigorously regenerating vegetation on water use and streamflow (see Figure 7; Vertessy *et al.*, 2001). It should be noted, however, that regeneration was particularly vigorous in the example from Hubbard Brook as it took place mainly via sprouting. As demonstrated by comparative work elsewhere in eastern United States, elevated streamflow levels last much longer (two to three times) when regeneration originates from seeds rather than via sprouting (Hornbeck *et al.*, 1993). Also, when a regenerating hardwood forest at Coweeta was cut a second time after 23 years (when streamflow was still ca. 80 mm year^{-1} above that observed for mature forest), the initial increase in water yield was nearly identical to that of the first cut (375 vs. 362 mm). However, the subsequent decline in (extra) water yield was distinctly faster. The difference was attributed to the much more rapid recovery of vegetation biomass during the second regeneration period owing to the greater sprouting potential of even-aged forest. The projected total duration of the second period of increased flows was estimated to be 18 years less than the 35 years associated with the first cut (Swank *et al.*, 1988). Such long durations reflect the time required by the roots to reoccupy the very deep soils at Coweeta compared with the much shallower soil at Hubbard Brook where effects rather last 3 to 9 years (Hornbeck *et al.*, 1993).

Effects of Forest Harvesting and Road Construction on Catchment Response to Rainfall

Part of the increases in annual streamflow totals after forest logging or clear-cutting (Figures 6, 8–11) reflects an increase in catchment response to rainfall (i.e. stormflows). In the absence of compacted surfaces such as roads, tractor tracks, and log landings, storm runoff is normally generated in the wettest parts within the landscape, such as depressions and riparian areas (*see* **Chapter 115, Landscape Element Contributions to Storm Runoff, Volume 3**). Topography permitting (see Figure 2), these runoff-producing areas will be enlarged after forest removal because of the increased amounts of precipitation reaching the ground, and the reduced uptake of water from the soil. Again, the strongest contrasts are expected during the time of the year when the soils are normally driest (see Figure 11; Hewlett, 1982). Also, the wetter soil conditions associated with cleared areas will be more conducive to flow and thus generally more responsive to rainfall. Relative increases in stormflow volumes after forestry operations *with minimum soil disturbance* (e.g. using skyline logging techniques) are largest for small rainfall events (up to three times the predisturbance volumes), intermediate for medium-sized events (+25–50%), but not too different (and lesser than 10–20% increase) for large events (Pearce *et al.*, 1980; Hewlett, 1982). As such, the effect of the presence or absence of a forest cover is inversely related to the size of the rainfall even generating the stormflow (Figure 12). This can be explained in terms of variations in the soil's capacity to store additional rain. Where previous uptake of soil water by the trees has depleted reserves, the storage capacity will be relatively high, but once the soil has become thoroughly wetted by frequent rains, opportunities to absorb large additional amounts of rain will be limited (see Figure 11). Likewise, as rainfall events increase in size, so does the relatively fixed maximum storage capacity of the soil become less influential (Figure 12). In other words, under conditions of extreme rainfall and soil wetness, the

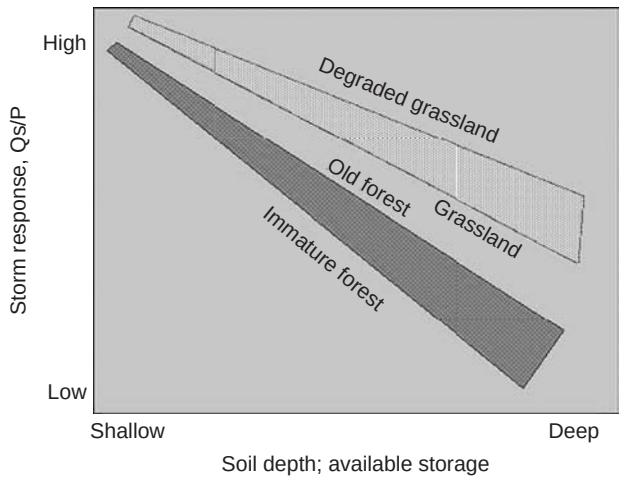


Figure 12 Postulated generalized relationship between catchment storage capacity and stormflow response to rainfall, as affected by vegetation cover (Reproduced from Scott *et al.*, 2004, by permission of Cambridge University Press). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

presence or absence of a forest cover is no longer decisive. Catchment runoff response is then governed primarily by the soil's physical capacity to store and transmit water (Hewlett, 1982; Grip *et al.*, 2004).

Similarly, *peak discharges* immediately downstream of forestry operations are increased temporarily (on an average by 5–50% in the absence of major road effects) as the wetter soil conditions after clearing cause an extension of the drainage network during rainfall. However, such effects on both stormflow volume and peak flows diminish subsequently with the establishment of the new vegetative cover (Hewlett, 1982; Grip *et al.*, 2004).

However, "typical" clear-cut operations involve road construction, mechanized harvesting using wheeled or tracked vehicles, and, in some cases, site preparation (chopping or windrowing and burning of slash) and machine planting (Hewlett and Doss, 1984; Swindel *et al.*, 1983). Under such conditions, soil disturbance can be expected to be much more severe and widespread, and potentially impair the catchment's capacity to absorb rainfall in a more structural manner. The actual effect on stormflow response will depend on the degree of topsoil compaction by machinery, the proportion of the catchment affected, and the degree of connectivity between roads and tracks and the drainage network (Van der Plas and Bruijnzeel, 1993; Jones and Grant, 1996; Grip *et al.*, 2004).

Equally important is the hillslope hydrological setting prior to disturbance, as illustrated by the following example from French Guyana (Figure 13; Grip *et al.*, 2004). Here, stormflow response for a series of small rain forested catchments varied roughly fivefold (7–34% of rainfall), depending on the proportion of the catchment underlain by free-draining or poorly drained soils, and the presence or absence of a marshy valley bottom. After mechanized clear-felling, the strongest *relative* impacts were observed on those catchments having the lowest storm runoff under forested conditions (150–200% increase), and *vice versa*. As such, catchments with free-draining soils proved much more sensitive to disturbance than catchments with impeded drainage and marshy valley bottoms (65–85% increase). This most probably reflects the fact that changes in the surface hydraulic properties of the free-draining soils (notably compaction by machinery) caused a shift in dominant runoff pathways during rainfall, that is, from predominantly vertical percolation under forests to more rapid surface routes (notably HOF) after clearing. By contrast, the soils with impeded drainage already exhibited ample rapid shallow lateral subsurface flow under forested conditions which, although partly displaced upward toward the surface after clearing, had less overall impact on the size of the storm hydrograph (see Gilmour, 1977b). On the other hand, *absolute* annual increases in stormflow totals were greater on the catchments with impeded drainage (310–360 mm) than on free-draining catchments (230–270 mm). However, despite these pronounced initial effects, storm runoff decreased to original values within four to five years for all of the forestry treatments imposed in this experiment (natural regrowth after logging and/or clearing, tree plantation establishment following clearing and burning; Figure 13) although it is pertinent to note that no permanent roads were constructed within the catchments (Grip *et al.*, 2004).

Needless to say, *roads* are usually highly compacted and have very low rainfall intake rates (typically $<15\,\mathrm{mm\,h^{-1}}$) and therefore very high annual runoff coefficients ($>65\%$) (Grip *et al.*, 2004). In addition, road-cut slopes can interrupt downslope-moving subsurface flow which is then converted to more rapid surface flow (Jones and Grant, 1996). As such, the layout of the road system and the proper drainage of road and track surfaces assume great importance if large off-site increases in stormflow are to be avoided (Adams and Andrus, 1990; Dykstra and Heinrich, 1996; see Figure 14 below). Indeed, such increases after forest harvesting in the Pacific Northwest of the United States were shown to be significantly enhanced if more than 10% of the catchment was occupied by roads compared with catchments having only 3–5% of road surface (Harr *et al.*, 1975). Similarly, increases in peak discharges after 25% clear-cutting with roads had the same effect as 100% clear-cuts without roads in the same area, although the underlying hydrologic mechanisms were different (Jones and Grant, 1996). At the same time, the downstream effects of forestry operations on catchment runoff response to rainfall at a larger scale are often thought to be modest to undetectable. This is because any adverse local effects of forest removal on all but the largest stormflow response (see Figure 12) tend to be "diluted" by more modest flows from other areas receiving less rain or being less disturbed (Hewlett, 1982; Bruijnzeel and Bremmer, 1989). However, in a thorough analysis of long-term peak flow records for $60–600\,\mathrm{km^2}$ catchments with logged-over forest in various stages of regeneration (again in the Pacific Northwest), Jones and Grant (1996) concluded that peak discharges had increased by 100% in large basins over the past 50 years versus a 50% increase in small (60–100 ha) basins. These increases were attributed mainly to changes in flow routing because of roads rather than to changes in soil water storage because of vegetation changes (as in some of the experiments discussed above).

Naturally, it is difficult to separate the effect of forest roads on flooding from those of logging in the *case of larger basins* (mostly because the two interventions usually happen more or less simultaneously). La Marche and Lettenmaier (2001), therefore, used a physically based, spatially distributed hydrologic model to explore the relative effects of the two types of disturbance in an extensively logged $150\,\mathrm{km^2}$ catchment (Deschute River, Washington), and in nine subcatchments therein ($2–21\,\mathrm{km^2}$). The model predicted increases in mean annual peak discharge of ca.

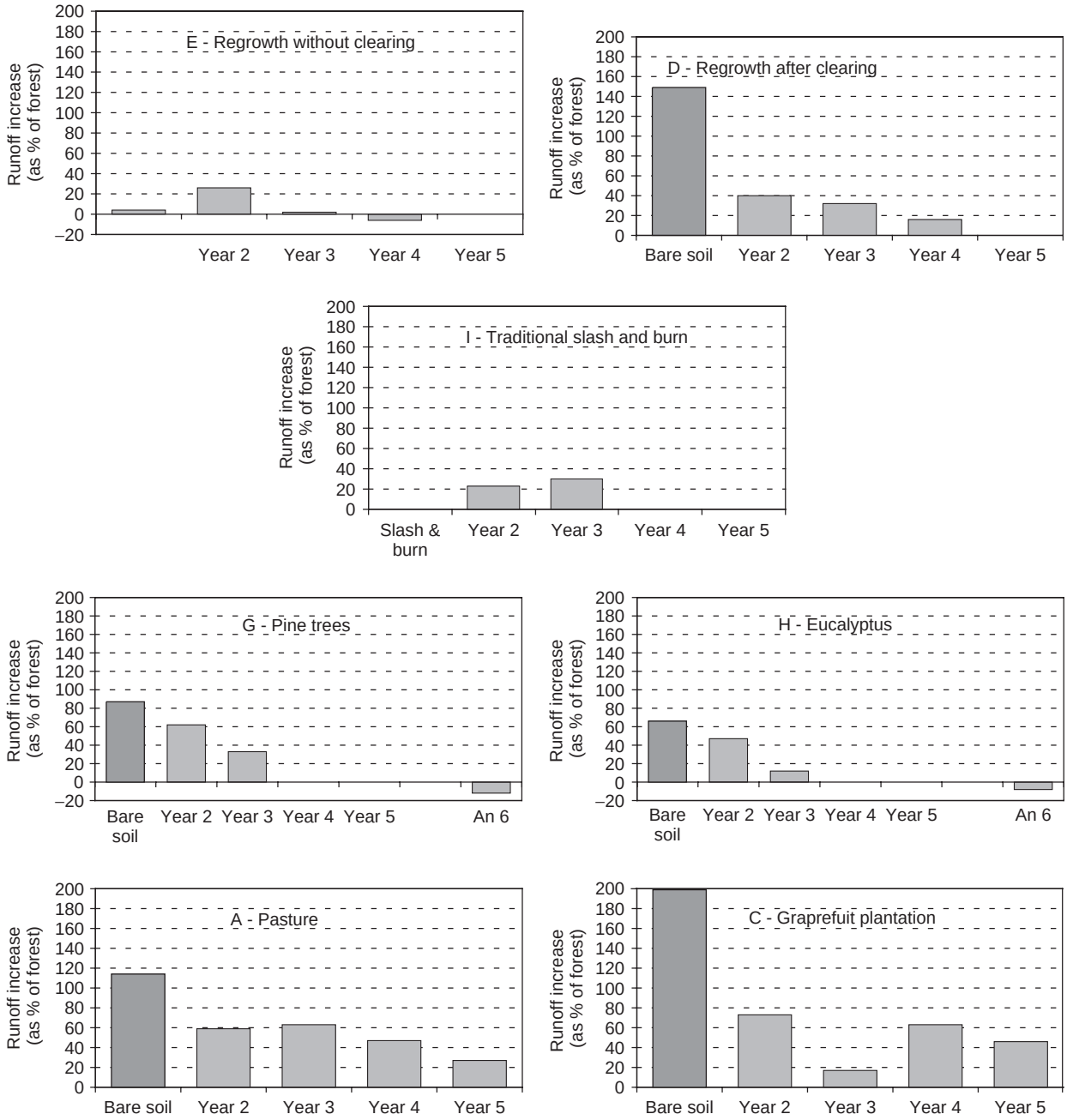


Figure 13 Changes in stormflow with time for the respective treatments within the ECEREX Experiment, French Guyana. Results are expressed as a percentage of corresponding runoff totals under forested conditions. (a) Regrowth following logging, (b) regrowth after logging and land clearing, (c) traditional slash and burn agriculture, (d) conversion to pine plantation, (e) conversion to eucalypt plantation, (f) conversion to pasture, and (g) conversion to grapefruit plantation (adapted from Fritsch, 1993). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2–10% because of roads alone. These increases were similar to the ones incurred by clear-cutting only. Interestingly, modeled road-only effects were slightly greater (ca. 3–12% increase) for the 10-year-event, whereas those for logging-only decreased with flood return period (i.e. conform the conceptual model of Figure 12). In other words, the effect of roads tends to be exacerbated in larger storms.

In view of these recent findings from the Pacific Northwest, the frequently heard but rarely demonstrated claim that logging tropical forest increases "flood" flows may contain more truth than has been acknowledged by the scientific community (see Hamilton and King, 1983; Bruijnzeel and Bremmer, 1989; Calder and Aylward, 2004), particularly when taking into account that logging operations in the

tropics rarely follow guideline prescriptions (Bruenig, 1996; Cassells and Bruijnzeel, 2004) and rainfall intensities there tend to be much higher than in the Pacific Northwest. It remains to be seen, however, whether the related claim of reductions in dry season flows after logging (due to the seriously hampered water intake capacity of the soil over a large portion of the catchment) in the tropics is as likely. Although such reductions in low flows have been demonstrated in the case of long-term, post-forest land degradation or after widespread urbanization (Bruijnzeel, 2004), regrowth after timber harvesting operations in the tropics is usually prolific and tends to reduce stormflow production from non-road surfaces (see Giambelluca, 2002). Preliminary modeling work has suggested that for dry season baseflows to become seriously reduced, infiltration-excess overland flow would need to be increased from marginal values under undisturbed forest to as much as 30–40% of the rainfall. Further work is needed to settle the debate (see Bruijnzeel *et al.*, 2004).

MINIMIZING ADVERSE HYDROLOGIC IMPACTS OF FORESTRY OPERATIONS

Generally, the more intensive a forestry operation, the more the soil and residual vegetation tend to be damaged. The relationship is not linear, however, and a poorly planned and executed harvest of a relatively small volume of timber may do as much damage as a more carefully executed operation removing twice as much timber. There is ample evidence that the key to minimizing damage to remaining trees and soil is the careful planning, preparation, and execution of any operation. The same holds for limiting the adverse off-site impacts (enhanced peak discharges, stream sedimentation) of, especially, road networks. Without going into detail here, the following measures are generally considered essential for a forestry operation to be successful in terms of minimizing both costs and environmental damage (Adams and Andrus, 1990; Pearce and Hamilton, 1986; Bruijnzeel, 1992; Dykstra and Heinrich, 1996; Bren, 2000):

- Prelogging assessment of potentially unstable (wet depressions, very steep slopes) areas or particularly erodible soil types; delineation of catchment boundaries and drainage lines (including channels that carry water during storms only); evaluation of seasonal distribution of rainfall to identify wettest periods.
- Preplanning of the extraction road and track network in relation to terrain characteristics, the natural drainage network, and type of logging system to be used (tractor/winch lorry/cable yarding/animal traction); keeping to ridges (roads, log landings) and staying away from streams (roads, log landings, and tracks) and steep sections as much as possible.
- Timing of road construction to conform to the period of least rainfall; allowing sufficient time for earthworks to stabilize before intensive use; provide adequate drainage of roads and tracks to prevent accumulation of large volumes of surface runoff that may initiate gully erosion upon being discharged.
- Whenever possible, yard logs uphill rather than downhill (Figure 14); use winch ropes rather than have tractors/skidders clear an approach to every log; directional felling into existing gaps; minimize the number of passes; raise the leading end of logs to keep them from ploughing into the soil; suspend tractor logging during wet periods to avoid excessive soil compaction.

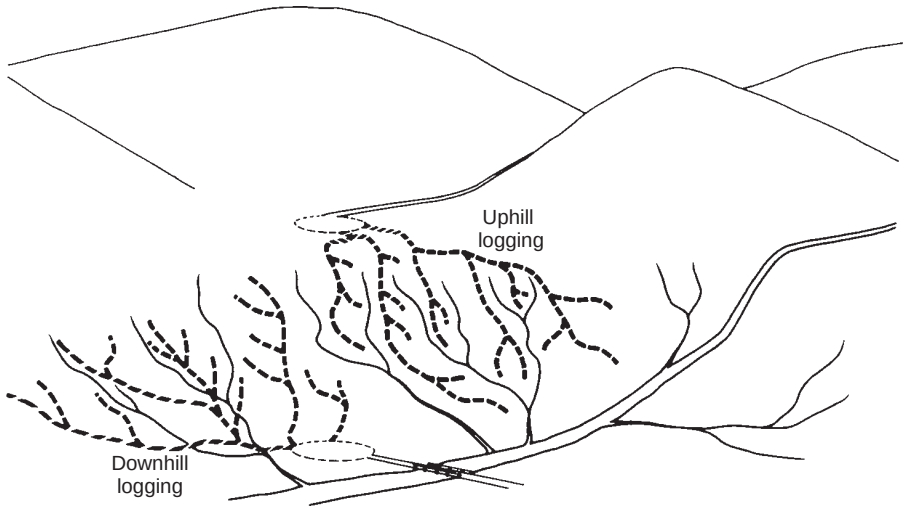


Figure 14 Hypothetical skidder track patterns (dashed lines) for uphill and downhill logging (Reproduced from Gilmour, 1977a, by permission of the Food and Agriculture Organization of the United Nations). Note how uphill logging tends to divert runoff and sediment away from streams, in contrast with the flow convergence promoted by downhill extraction

- Maintain streamside buffer strips to trap runoff and sediment generated upslope and to provide shade to the aquatic ecosystem; recommended widths will vary between terrains, but should be at least 10–20 m on either side and not only cover major streams but also ephemeral channels and other runoff generating areas where possible.
- Postlogging treatment of skidder tracks (removal of temporary stream crossings, construction of cross drains on critical tracts) and log landings (breaking up/reseeding); road maintenance; reseeding of roadside earth works.

Comprehensive sets of guidelines that include most or all of the above aspects have been developed in a number of countries. Their application in northern Australia in an area experiencing some of the world's most erosive rainfall intensities, has been shown to be highly effective in meeting the objectives of minimum extra costs and minimum environmental impacts, both on-site and downstream (Gilmour, 1977a; Cassells *et al.*, 1984). Naturally, to be truly effective, the various measures listed above should be applied with a rigor that matches the intensity of local rainfall and steepness of terrain. In many cases, this is easier said than done, as the forestry industry tends to resist change, being guided by short-term economic gains rather than long-term considerations of environmental sustainability. However, with the introduction of educational programmes, combined with the application of a realistic penalty structure and, more recently, the introduction of timber certification schemes, major improvements in harvesting practices have been achieved in a number of areas (Cassells *et al.*, 1984; Bruenig, 1996; Cassells and Bruijnzeel, 2004).

REFERENCES

Abdul Rahim N. and Zulkifli Y. (1994) Hydrological response to selective logging in Peninsular Malaysia and its implications for watershed management. In *Proceedings of the International Symposium on Forest Hydrology 1994*, Ohta T., Fukushima Y. and Suzuki M. (Eds.), IUFRO: Tokyo, pp. 263–274.

Adams P.W. and Andrus C.W. (1990) Planning secondary roads to reduce erosion and sedimentation in humid tropic steeplands. *International Association of Hydrological Sciences Publication*, **192**, 318–327.

Asdak C., Jarvis P.G., Van Gardingen P. and Fraser A. (1998) Rainfall interception loss in unlogged and logged forest areas of Central Kalimantan, Indonesia. *Journal of Hydrology*, **206**, 237–244.

Aussenac G., Granier A. and Naud R. (1982) Influence of thinning on growth and water balance. *Canadian Journal of Forest Research*, **12**, 222–231.

Best A., Zhang L., McMahon T., Western A. and Vertessy R.A. (2003) *A Critical Review of Paired Catchment Studies with Reference to Seasonal Flows and Climate Variability*, CSIRO Land and Water Technical Report, 25/03, CSIRO Land and Water, Canberra, p. 44.

Black T.A., Tan C.S. and Nnyamah J.U. (1980) Transpiration rate of Douglas-fir trees in thinned and unthinned stands. *Canadian Journal of Soil Science*, **60**, 625–631.

Bosch J. and Hewlett J.D. (1982) A review of catchment experiments to determine the effects of vegetation changes on water yield and evapotranspiration. *Journal of Hydrology*, **55**, 3–23.

Bren L.J. (2000) A case study in the use of threshold measures of hydrologic loading in the design of stream buffer strips. *Forest Ecology and Management*, **132**, 243–257.

Bruenig E.F. (1996) *Conservation and Management of Tropical Rainforests. An Integrated Approach to Sustainability*, CAB International: Wallingford.

Bruijnzeel L.A. (1986) Environmental impacts of (de)forestation in the humid tropics: a watershed perspective. *Wallaceana*, **46**, 3–13.

Bruijnzeel L.A. (1992) Managing tropical forest watersheds for production: where contradictory theory and practice co-exist. In *Wise Management of Tropical Forests*, Miller F.R. and Adam K.L. (Eds.), Oxford Forestry Institute: Oxford, pp. 37–75.

Bruijnzeel L.A. (2000) Forest hydrology. In *The Forests Handbook*, Evans J.C. (Ed.), Blackwell Scientific: Oxford, pp. 301–343.

Bruijnzeel L.A. (2004) Hydrological functions of tropical forests: not seeing the soil for the trees? *Agriculture, Ecosystems & Environment*, **104**, 185–227.

Bruijnzeel L.A., Bonell M., Gilmour D.A. and Lamb D. (2004) Tropical forests, water and people – an emerging view. In *Forests, Water and People in the Humid Tropics*, Bonell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge.

Bruijnzeel L.A. and Bremmer C.N. (1989) *Highland-Lowland Interactions in the Ganges Brahmaputra River Basin: A Review of Published Literature*, ICIMOD Occasional Paper No. 11, International Centre for Integrated Mountain Development: Kathmandu, p. 136.

Calder I.R. (1999) *The Blue Revolution*, Earthscan Publications: London.

Calder I.R. and Aylward B. (2004) *Forests and Floods: Perspectives on Watershed Management and Integrated Flood Management*, Food and Agriculture Organization of the United Nations: Rome.

Cassells D.S. and Bruijnzeel L.A. (2004) Guidelines for controlling the soil and water Impacts of timber harvesting in the humid tropics: a critical appraisal. In *Forests, Water and People in the Humid Tropics*, Bonell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge.

Cassells D.S., Gilmour D.A. and Bonell M. (1984) Watershed forest management practices in the tropical rainforests of north-eastern Australia. In *Effects of Forest Land Use on Erosion and Slope Stability*, O'Loughlin C.L. and Pearce A.J. (Eds.), IUFRO: Vienna, pp. 289–298.

Chappell N.A., Bidin K. and Tych W. (2001) Modelling rainfall and canopy controls on net precipitation beneath selectively-logged tropical forest. *Plant Ecology*, **153**, 215–229.

Delfs J. (1955) Die Niederschlagszurückhaltung im Walde. *Mitteilungen des Arbeitskreises "Wald und Wasser"*, **2**, 1–54.

Dunford E.G. and Fletcher P.W. (1947) The effect of removal of streambank vegetation upon water yield. *Transactions of the American Geophysical Union*, **28**, 105–110.

Dunne T. (1978) Field studies of hillslope flow processes. In *Hillslope Hydrology*, Kirkby M.J. (Ed.), John Wiley & Sons: Chichester, pp. 227–293.

Dye P.J. and Poulter A.G. (1995) A field demonstration of the effect on streamflow of clearing invasive pine and wattle trees from a riparian zone. *South African Forestry Journal*, **173**, 27–30.

Dykstra D. and Heinrich R. (1996) *FAO Model Code of Forest Harvesting Practice*, Food and Agriculture Organization of the United Nations: Rome.

Florido L.V. and Saplaco S.R. (1981) Rainfall interception in a thinned 10-15-year-old natural Benguet pine (*Pinus kesiya* Roy. ex Gordon) stand. *Sylvatrop Philippines Forest Research Journal*, **6**, 195–201.

Fritsch J.M. (1993) The hydrological effects of clearing tropical rainforest and of implementation of alternative land uses. *International Association of Hydrological Sciences Publication*, **216**, 53–66.

Giambelluca T.W. (2002) The hydrology of altered tropical forest. Invited Commentary. *Hydrological Processes*, **16**, 1665–1669.

Gilmour D.A. (1977a) Logging and the environment, with particular reference to soil and stream protection in tropical rainforest situations. In *FAO Conservation Guide*, Vol. 1, Kunkle S.H. (Ed.), Food and Agriculture Organization of the United Nations: Rome, pp. 223–235.

Gilmour D.A. (1977b) Effects of rainforest logging and clearing on water yield and quality in a high rainfall zone of north-east Queensland. *Proceedings of the Brisbane Hydrology Symposium 1977*, Institution of Engineers Australia: Canberra, pp. 156–160.

Grip H., Fritsch J.M. and Bruijnzeel L.A. (2004) Soil and water impacts during forest conversion and stabilisation to new land use. In *Forests, Water and People in the Humid Tropics*, Bonell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge.

Hall R.L. and Calder I.R. (1993) Drop size modification by forest canopies: measurements using a disdrometer. *Journal of Geophysical Research*, **98**, 18,465–18,470.

Hamilton L.S. and King P.N. (1983) *Tropical Forested Watersheds: Hydrologic and Soils Response to Major Uses or Conversions*, Westview Press: Boulder.

Harr R.D., Harper W.C. and Krygier J.T. (1975) Changes in storm hydrographs after road building and clear-cutting in the Oregon coast range. *Water Resources Research*, **11**, 436–444.

Haydon S.R., Benyon R.G. and Lewis R. (1996) Variation in sapwood area and throughfall with forest age in mountain ash (*Eucalyptus regnans* F. Muell.). *Journal of Hydrology*, **187**, 351–366.

Helvey J.D. (1967) Interception by eastern white pine. *Water Resources Research*, **3**, 723–729.

Helvey J.D. and Patric J.H. (1965) Canopy and litter interception of rainfall by hardwoods of eastern United States. *Water Resources Research*, **1**, 193–206.

Hewlett J.D. (1982) Forests and floods in the light of recent investigations. *Hydrological Processes of Forested Areas*, National Research Council of Canada Publication No. 20548, NRCC: Ottawa, pp. 543–560.

Hewlett J.D. and Doss R. (1984) Forests, floods, erosion: a watershed experiment in the southeastern Piedmont. *Forest Science*, **30**, 424–434.

Hewlett J.D. and Fortson J.C. (1983) The paired catchment experiment. In *Forest Water Quality*, Hewlett J.D. (Ed.), School of Forest Resources, University of Georgia: Athens, pp. 11–14.

Hornbeck J.W., Adams M.B., Corbett E.S., Verry E.S. and Lynch J.A. (1993) Long-term impacts of forest treatments on water yield: a summary for northeastern USA. *Journal of Hydrology*, **150**, 323–344.

Hornbeck J.W., Pierce R.S. and Federer C.A. (1970) Streamflow changes after forest clearing in New England. *Water Resources Research*, **6**, 1124–1132.

Ingwersen J.B. (1985) Fog drip, water yield, and timber harvesting in the Bull Run municipal watershed, Oregon. *Water Resources Bulletin*, **21**, 469–473.

Johnson E.A. and Kovner J.L. (1956) Effect on streamflow of cutting a forest understory. *Forest Science*, **2**, 82–91.

Jones J.A. and Grant G.E. (1996) Peak flow responses to clear-cutting and roads in small and large basins, western Cascades, Oregon. *Water Resources Research*, **32**, 959–974.

Kaimowitz D. (2004) Useful myths and intractable truths: the politics of the link between forests and water in Central America. In *Forests, Water and People in the Humid Tropics*, Bonell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge.

Kelliher F.M., Black T.A. and Price D.T. (1986) Estimating the effects of understory removal from a Douglas fir forest using a two-layer canopy evapotranspiration model. *Water Resources Research*, **22**, 1891–1899.

La Marche J.L. and Lettenmaier D.P. (2001) Effects of forest roads on flood flows in the Deschutes River, Washington. *Earth Surface Processes and Landforms*, **26**, 115–134.

Lesch W. and Scott D.F. (1997) The responses in water yield to the thinning of *Pinus radiata, Pinus patula* and *Eucalyptus grandis* plantations. *Forest Ecology and Management*, **99**, 295–307.

Lloyd C.R. and Marques-Filho A.deO. (1988) Spatial variability of throughfall and stemflow measurements in Amazonian rain forest. *Agricultural and Forest Meteorology,* **42**, 63–73.

Mackensen J., Klinge R., Ruhiyat D. and Fölster H. (2003) Assessment of management-dependent nutrient fluxes in tropical industrial tree plantations. *Ambio*, **32**, 106–112.

Parker, G.G. (1985) *The Effect of Disturbance on Water and Solute Budgets of Hillslope Tropical Rainforest in Northeastern Costa Rica*, PhD Thesis, Department of Ecology, University of Georgia, Athens, p. 161.

Pearce A.J. and Hamilton L.S. (1986) Water and soil conservation guidelines for land-use planning. *Summary of Seminar-Workshop on Watershed Land-Use Planning*, Gympie. Environmental & Policy Institute: Honolulu, p. 43, May 1985.

Pearce A.J., Rowe L.K. and O'Loughlin C.L. (1980) Effects of clearfelling and slashburning on water yields and storm hydrographs in evergreen mixed forests, western New Zealand.

International Association of Hydrological Sciences Publication, **130**, 119–127.

Rogerson T.L. (1967) Throughfall in pole-sized loblolly pine as affected by stand density. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon Press: Oxford, pp. 187–190.

Scott D.F., Bruijnzeel L.A. and Mackensen J. (2004) The hydrological and soil impacts of forestation in the tropics. In *Forests, Water and People in the Humid Tropics*, Bonell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge.

Scott D.F. and Lesch W. (1996) The effects of riparian clearing and clearfelling of an indigenous forest on streamflow, stormflow and water quality. *South African Forestry Journal*, **175**, 1–14.

Stednick J.D. (1996) Monitoring the effects of timber harvest on annual water yield. *Journal of Hydrology*, **176**, 79–95.

Stogsdill W.R., Wittwer R.F., Hennessey T.C. and Dougherty P.M. (1992) Water use in thinned loblolly pine plantations. *Forest Ecology and Management*, **50**, 233–245.

Swank W.T., Swift L.W. Jr and Douglas J.E. (1988) Streamflow changes associated with forest cutting, species conversions, and natural disturbances. In *Forest Hydrology at Coweeta*, *Ecological Studies 66*, Swank W.T. and Crossley D.A. (Eds.), Springer Verlag, pp. 297–312.

Swindel B.F., Lassiter C.J. and Riekerk H. (1983) Effects of different harvesting and site preparation operations on the peak flows of streams in *Pinus elliottii* flatwood forests. *Forest Ecology and Management*, **5**, 77–86.

Teklehaimanot Z., Jarvis P.G. and Ledger D.C. (1991) Rainfall interception and boundary layer conductance in relation to tree spacing. *Journal of Hydrology*, **123**, 261–278.

Trimble G.R., Reinhart K.G. and Webster H.H. (1963) Cutting the forest to increase water yields. *Journal of Forestry*, **61**, 635–640.

Van der Plas M.C. and Bruijnzeel L.A. (1993) Impact of mechanized selective logging of rainforest on topsoil infiltrability in the Upper Segama area, Sabah, Malaysia. *International Association of Hydrological Sciences Publication*, **216**, 203–211.

Veracion V.P. and Lopez A.C.B. (1976) Rainfall interception in a thinned Benguet pine (*Pinus kesiya*) forest stand. *Sylvatrop Philippines Forest Research Journal*, **1**, 128–134.

Vertessy R.A., Watson F.G.R. and O'Sullivan S.K. (2001) Factors determining relations between stand age and catchment water yield in mountain ash forests. *Forest Ecology and Management*, **143**, 13–26.

Vertessy R.A., Watson F.G.R., O'Sullivan S.K., Davis S., Campbell R., Benyon R.G. and Haydon S.R. (1998) *Predicting Water Yield From Mountain Ash Forest Catchments*, CRCCH Industry Report 98/4, Cooperative Research Centre for Catchment Hydrology, Clayton.

Vertessy R.A., Zhang L. and Dawes W.R. (2003) Plantations, river flows and river salinity. *Australian Forestry*, **66**, 901–907.

Whitehead D., Jarvis P.G. and Waring R.H. (1984) Stomatal conductance, transpiration, and resistance to water uptake in a *Pinus sylvestris* spacing experiment. *Canadian Journal of Forest Research*, **14**, 692–700.

Wiersum K.F. (1985) Effects of various vegetation layers in an *Acacia auriculiformis* forest plantation on surface erosion in Java, Indonesia. In *Soil Erosion and Conservation*, El-Swaify S.A., Moldenhauer W.C. and Lo A. (Eds.), Soil Conservation Society of America: Ankeny, pp. 79–89.

Zhou G.Y., Morris J.D., Yan J.H., Yu Z.Y. and Peng S.L. (2001) Hydrological impacts of reafforestation with eucalyptus and indigenous species: a case study in southern China. *Forest Ecology and Management*, **167**, 209–222.

Zon R. (1927) *Forests and Water in the Light of Scientific Investigation*, US Government Printing Office.

120: Land Use and Land Cover Effects on Runoff Processes: Fire

CHARLES H LUCE

Rocky Mountain Research Station, Boise, ID, US

Fire dramatically alters hydrologic processes in many regions of the world. Individual fires reduce vegetation and change soil characteristics, sometimes producing dramatic runoff events in the years shortly after a fire. The greatest determinant of the effect of fire on runoff generation is the severity of the fire, which relates to the frequency of fires and other climatic and vegetation characteristics. Severe fires can produce hydrophobic soils or increase risk of soil surface sealing, reducing infiltration rates. Measurements of the spatial pattern of water repellent soils are useful for estimating potential runoff from postfire storms. The most severe events occur during convective storms, so the spatial extent of individual postfire floods is generally limited in extent. Recovery of water repellent soils is relatively rapid, with significant reductions occurring within a few years. Longer-term changes to hydrology are related to the reduced evapotranspiration caused by loss of vegetation biomass. In forests, changes to annual water balances may last decades.

INTRODUCTION

In some parts of the world, fire is an important natural disturbance to landscapes. Its very nature can cause substantial changes in hydrological processes in an area, as it consumes vegetation biomass and sometimes affects the soil characteristics directly. One of the most spectacular hydrologic results of fire is the combination of water repellent soils followed by thunderstorms, which can create locally severe flooding and erosion (e.g. Klock and Helvey, 1976; DeBano, 1981; Swanson 1981; Moody and Martin, 2001; Istanbulluoglu *et al.*, 2002; Miller *et al.*, 2003). Effects of fire on vegetation, soil, and hydrologic processes can be extraordinarily variable, ranging from nearly no noticeable effect to extreme flood events with results such as those shown in Figure 1. The degree of effect depends on the severity of the fire, or how hot and long it burned, and the spatial extent and patchiness of high severity fire. While there are clear deterministic influences on fire behavior, such as fuel amount and condition, air temperature, humidity, and wind, it can be treated stochastically, and in this sense, fire can be thought of as a weather phenomenon itself. Like other hydrologically relevant weather parameters, fire can be considered both from the perspective of regime, with return frequency and severity metrics, and as an event. Fire regime is essentially the climatic context of fire, and just as it would be somewhat nonsensical to discuss flooding processes without mentioning the aridity of the landscape, so it is that fire regime is an important concept for fire effects. This article first discusses fire regimes and their relationship to hydrology and expectation for fire events, followed by a discussion of the hydrologic responses that might be expected from a given fire event.

FIRE REGIMES AND HYDROLOGY

Metrics for fire regimes relate to the frequency with which fire visits an area and what it tends to do to dominant vegetation in the area. The effects of fire can range from minor damage to the dominant vegetation to stand replacement, where all vegetation in the area is consumed (Hessburg and Agee, 2003). Where there is a strong tendency for one type of outcome versus the other, the severity can be classified as "nonlethal" or as "stand replacing." Where the nature of effects tends to change from fire to fire, the severity is classified as "mixed," and a spatially patchy vegetation structure can result. Within

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

Figure 1 Mouth of Wren Creek in the Boise National Forest, Idaho. The watershed burned in 1994, and a severe thunderstorm passed over the basin in the summer of 1995, initiating a hyperconcentrated flow event in this and neighboring streams. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

a given vegetation community type (e.g. shrub steppe, or forest), severity tends to go hand in hand with frequency. Very frequent regimes, with mean fire intervals less than 25 years, tend to have little fuel to consume with each event, and frequent kills of small trees do not allow for creation of complex vegetation canopies that can act as ladders from ground fuels to the dominant vegetation species. Conversely, very infrequent regimes allow time for buildup of significant fuels and complex canopies, and crown fires are more common in areas where fires occur less frequently.

Fire regimes have a profound influence on vegetation characteristics, so knowledge of the relationship between plant species and the kinds of fire regimes that they tend to occur with can provide some information about the nature of fires one might expect. Frequency and severity of disturbance can be important in the determination of species (Franklin and Dyrness, 1973), which in turn can affect the probability, severity, and continuity of successive fires. Fire-adapted plant species have strategies to either avoid impacts from flames or to quickly capitalize on freshly burned sites. Trees with thick bark (e.g. Ponderosa Pine, *Pinus ponderosa*) are common in locations with frequent low severity burns, where the bark protects the cambium from the effects of a quickly moving ground fire, and the height of the tree keeps the crown from catching fire. Trees with serrotinous cones (e.g. lodgpole pine, *Pinus contorta*, or Jack pine, *Pinus banksiana*) are more common in locations with rare but severe fires that cover large areas, as the serrotinous cones release seeds into a nutrient rich environment with little competition from more distant seed sources. Invasive fire-adapted grass species (e.g. Cheatgrass, *Bromus tectorum*) can set up a frequent fire regime that prevents regeneration of deeper-rooted native shrub species (Young and Evan, 1985; Billings, 1994).

One of the effects of fire on runoff generation processes may occur where shifts in fire regime force changes in plant communities that affect soil properties and the hydrologic cycle. Fire regime is a function not only of vegetation assemblages but also of the climate. Fire both drives and is driven by vegetation changes in response to climate change. A variety of stratigraphic evidence has shown that substantial variations in precipitation and temperature can occur on long timescales, producing periods of shifting fire regimes and vegetation within an area (Meyer and Pierce, 2003; Whitlock *et al.*, 2003).

Within arid forests and rangelands, physiological adaptations to seasonal aridity are obviously important in determining relative success and spatial distribution of species within the landscape. The ecohydrological optimality principles (*see* **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**; **Chapter 101, Ecosystem Processes, Volume 3**; and **Chapter 103, Terrestrial Ecosystems, Volume 3**) that apply well to more humid landscapes (Eagleson, 2002) could potentially represent arid and semiarid landscapes better if fire disturbance, essentially a hydroclimatology phenomenon itself, was included in the conceptualization.

HYDROLOGY AFTER FIRE EVENTS

The effects of an individual fire event on hydrologic processes are tied primarily to the loss of vegetation, loss of organic matter at the soil surface, and the chemico-physical changes to shallow soil horizons that lead to water repellency. The degree of effect is greatly affected by the characteristics of the fire and the fuels it is burning through. Rate and duration of energy releases are key characteristics, and are affected by fuel size and moisture distributions, the amount of fuel available to the fire, volatility of the fuels, and weather (temperature, humidity, wind speed) at the time of burning (Albini *et al.*, 1996). The patchiness of the resulting burn is also important to runoff generation and is tied to continuity and availability of fuels. Effects on soil organic matter and water repellency are less common and depend on soil and vegetation characteristics. One of the primary differences between purposefully set "prescribed" fires and wildfire is that the decision about when to set the fire allows for some degree of control of many of the important factors, including weather, fuel moisture, and soil moisture, which allows for some control on the degree of disturbance to the soil.

Water Repellency

One of the most commonly cited effects of fire on runoff generation processes is the formation of a water repellent layer in the soil, sometimes termed *hydrophobicity* (e.g. DeBano, 1981, 2000). Water repellency is a condition where soil not only loses its usual capillary draw on water, but actually resists entry of water into the soil (*see* **Chapter 68, Water Movement in Hydrophobic Soils, Volume 2**). This condition increases the amount of infiltration excess overland flow (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**). Water repellency occurs naturally in some soils, but seems to be increased in severity, strength, degree, persistence, and continuity by fire (DeBano, 2000). Typically, water repellency occurs in locations with severe heating of the soil surface.

Postfire water repellency is hypothesized to occur by translocation of waxes and other organic compounds with hydrophobic properties from upper layers of the soil and organic matter into lower layers by vaporization where temperatures are high at the surface and condensation on soil particles lower in the soil profile where temperatures are cooler (DeBano *et al.*, 1976). As one might expect, sufficient temperature and duration of heating are necessary for the formation of water repellent layers, and there is some indication that vegetation type affects the formation of water repellent soils (Doerr *et al.*, 2000). The effectiveness of the hypothesized coating process is dependent on the amount of material to be coated, and there is some effect of soil specific surface, generally as measured by grain size. Coarse soils tend to be more susceptible to water repellency than fine grained (DeBano, 1981).

Water repellency is sensitive to soil moisture, and the soil does not impede water movement once wetted (Imeson *et al.*, 1992; Doerr *et al.*, 2000). The requirement for low soil moistures tends to make water repellency a dry season phenomenon, with little effect on runoff generation during snowmelt. Typically, the concern is intense precipitation events during the summer months leading to brief, severe flash flood events. Once the wettable surface layers are wetted, water repellent soils can yield substantial runoff as an infiltration excess process. Over the course of a storm, the infiltration capacity of the soil tends to increase, at least during early times; this is in direct contrast to normally wettable soils that see a decrease in infiltration capacity during a storm (Imeson *et al.*, 1992). After being wetted, soils can become water repellent again if dried.

There are three timescales fundamental to the degree of water repellency, the time for wetting during a storm event (minutes to hours), the variations due to annual wetting and drying, and a longer-term decay of water repellency. While there are a number of examinations of the shortest timescale (e.g. Imeson *et al.*, 1992), and a fairly well known wetting and drying relationship, the longer-term persistence of fire-induced water repellency is not well understood. A water repellent layer may break down due to microbial activity, dissolution during wetting and drying, or physical disturbances like freeze thaw, bioturbation, and soil creep. Severe erosion events (often in the form of rilling) induced by water repellency are a key process for removal of water repellent layers and result in a spatial organization of water repellent and nonrepellent soils, where the upslope interrill patches are repellent but the rills are nonrepellent. There are few published observations of the persistence of water repellency. Dyrness (1976) observed water repellency in a burned area six years after the fire. Personal observations have shown extensive water repellency still existing seven years after severe fire under a subalpine fir stand, and spotty repellency 25 years after a prescribed fire in a coastal Douglas-fir stand. I have also seen extensive water repellency under a subalpine fir stand that had no fire in the last 200 years, but there is no certainty that the repellency originated with a fire in this stand. The erosion mechanism, mentioned above, seems to be the fastest mechanism for removing large areas of water repellent soils. A clearer picture of processes and timescales for recovery from fire-induced water repellency is needed to better understand long-term risks of flooding posed by fires (Doerr and Moody, 2004).

While substantial study has gone into research on water repellency at point and plot scales, an understanding of how it contributes to runoff generation even in small catchments is largely unexplored (Shakesby *et al.*, 2000; Doerr

and Moody, 2004). Most of the research on water repellency has focused on methods for measuring the "strength" of the water repellency (Letey *et al.*, 2000). Such measures include water contact angles, head needed to penetrate, or water drop penetration times. At scales of one to a few meters, the tendency of water repellent soils to form preferential flow paths or fingers of wetting has been noted (Imeson *et al.*, 1992; Ritsema and Dekker, 2000). Conceptually, this idealization can apply at larger scales as well, where topology and runoff–runon relationships must be considered (Shakesby *et al.*, 2000; Doerr and Moody, 2004). This conceptualization would argue that if we were interested in the potential for runoff production from a watershed during intense storm events, we would want to measure the fractional area that is water repellent. This approach has seen some success in estimating location of gully initiation sites (Istanbulluoglu *et al.*, 2002).

Soil Surface Sealing

Surface sealing is another frequently suggested mechanism for reductions in infiltration capacity and increases in overland flow (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**) following fire (e.g. Rowe, 1948; Swanson, 1981; Benavides-Solorio and MacDonald, 2001; Meyer and Pierce, 2003). Surface sealing has not received as thorough a treatment for postfire periods as it has in the literature addressing agricultural and severely disturbed soils (e.g. Mohammed and Kohl, 1987; Bosch and Onstad, 1988; Luce, 1997). Surface sealing occurs when raindrop impact and rapid wetting break up soil aggregates, effectively reducing the surface grain size and hydraulic conductivity, and potentially forming a crust. Erosion initiated with the loss of the protective surface organics can also cause relocation of surface fines into macropores, reducing their capacity to move water into deeper layers quickly. Luce (1997) noted reductions in hydraulic conductivity in excess of 70%. The degree of reduction depends on clay content and type, the kinetic energy of the precipitation, and the duration of exposure. Soils with high clay content (nondispersive clays) and high organic matter content tend to have stronger aggregates (Kemper and Koch, 1966). Reduced surface hydraulic conductivity can lead to the initiation of infiltration excess (Horton) overland flow during rainfall events with intensities greater than the hydraulic conductivity. Organic matter reductions are patchy, with organic matter consumption generally related to local burn conditions. If soils in a watershed are susceptible to surface sealing, the hillslope scale runoff generation will depend on the degree of surface sealing, and the proportion of the hillslope and downslope continuity of patches where the organic matter is completely consumed.

Vegetation Loss

The effects of vegetation canopy loss are similar to other land use effects such as forest harvest or rangeland chaining (*see* **Chapter 119, Land Use and Landcover Effects on Runoff Processes: Forest Harvesting and Road Construction, Volume 3**). These effects include reductions in evapotranspiration, reduced interception of liquid and solid precipitation, and increased snowmelt rates during periods of solar dominated melt (e.g. Reifsnyder and Lull, 1965; Harr, 1976; Waring and Schlesinger, 1985 and *see* **Chapter 42, Transpiration, Volume 1**, **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**, **Chapter 39, Surface Radiation Balance, Volume 1**). Reductions in evapotranspiration and interception generally lead to higher soil moistures (Johnston, 1970; Klock and Helvey, 1976) and greater annual water yields (Megahan, 1983; Troendle and King, 1985; Kuczera, 1987; Watson *et al.*, 1999). The result is greater low flow generation during summer, with springs active higher in watersheds, and more opportunity for production of peak flows (Harr, 1976; Campbell and Morris, 1988). Reduced shading by canopy can substantially increase snowmelt rates leading to increased peak flows in snowmelt-dominated systems (*see* **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**). Although we understand that standing dead trees can inhibit turbulent exchange between the snowpack and the atmosphere, the strength of the effect of wildfire on turbulent heat transfers is less well researched. Reductions in organic matter on the soil surface accompanying other vegetation loss primarily yield a reduction in water interception, which can be important during brief precipitation events. While some effect of vegetation loss on runoff generation is expected for almost every fire, the degree of the effects listed is greatly determined by the degree of vegetation loss. A crown fire may remove branches and needles from a tree, where a ground fire may result in patchy mortality and gradual dying of leaves and needles.

Scale of Effects

Although fire creates soil and vegetation conditions that are more conducive to severe hydrologic behavior and rapid runoff forming processes, the more catastrophic postfire runoff events also require substantial precipitation or snowmelt events. There are generally some limitations on the spatial and temporal scales of extreme events (*see* **Chapter 1, On the Fundamentals of Hydrological Sciences, Volume 1**). Although we have seen large fires (e.g. greater than 20 000 ha) in some parts of the world in recent years, it is not uncommon for severe hydrologic events to be confined to a small portion of the fire, even to a small portion of the severely burned areas, suggesting that the characteristic patch scale of intense

precipitation and rapid snowmelt events is generally smaller than that of large wildfires (Miller *et al.*, 2003). The fact that these events tend to be limited in their scale is of great consequence to fish, which have evolved migratory life histories and metapopulation strategies to cope with fire related disturbances (Dunham *et al.*, 2003; Rieman *et al.*, 2003).

Flash flood events and related hyperconcentrated flows are partially constrained in scale by the size of the thunderstorm causing the event. Infiltration excess runoff generation processes require that a threshold precipitation intensity be exceeded in order for runoff to occur, and intense precipitation from thunderstorms cover a limited extent. Consequently, it is not uncommon to have only a few small patches within a burned area, typically less than 20 square kilometers each, affected by severe runoff processes (Moody and Martin, 2001; Miller *et al.*, 2003). This dependence on area is not a new concept, as evidenced by a long and prolific literature on the subject of spatial scaling for design precipitation events (e.g. Rodriguez-Iturbe and Mejia, 1974; Rodriguez-Iturbe, 1986; Sivapalan and Blöschl, 1998; Seed *et al.*, 1999 as a small sample). In most engineering applications, the areal reduction factor (ARF) approach has been used. The ARF relates a decrease in storm intensity to the size of the basin. To examine the change in probability of an event of particular intensity and duration as a function of the area under consideration, storm-centered ARFs are needed. Statistics for both storm-centered and basin-centered ARFs can be developed from radar precipitation images of a series of storms. The size of individual basins affected by hyperconcentrated flows is effectively governed by tributary junctions with larger stream channels, with major deposition occurring when what constitutes a major event to a tributary is insignificant to the receiving channel.

Debris flows related to rapid snowmelt characteristically have larger patch dimensions (Miller *et al.*, 2003). Rapid snowmelt generally occurs in concert with large synoptic scale precipitation events covering patches a few hundred kilometers in extent, and in practicality, rapid snowmelt is constrained in scale by elevation. Rapid snowmelt is driven by the high winds during these events and occurs only where air temperatures over a snowpack are above freezing, in one major storm, the elevation range was less than 500 m (Miller *et al.*, 2003). The severe landslide events related to it are further constrained by the land slope, and the final extent of related debris flows is constrained by channel slope, which generally restricts them to headwater basins.

Less dramatic changes to subsurface flow runoff generation processes caused by loss of vegetation or changes in vegetation density are probably effectively constrained by elevation as well. Most measurements of changes in annual water yield only show differences in humid climates. In the interior west of the United States, changes have been seen in mixed conifer and subalpine vegetation types, with little to no change occurring in montane systems at lower elevations (e.g. compare Troendle and King, 1985 to Megahan, 1983).

SUMMARY

Fire is fundamentally intertwined with hydrology. Its occurrence is controlled by seasonal and longer timescale hydroclimatology, and it greatly affects hydrologic processes through its controls on vegetation and soil conditions.

Fire regime presents some expectation for the nature of fire events that might occur in an area. Specifically, some idea of the degree to which vegetation will be removed, the degree to which the soil might be altered, and the patchiness of those effects are tied to the fire regime. Fire regimes also control the type and density of vegetation present in an area.

After a fire, the potentially most destructive runoff generation process is infiltration excess runoff generation, which is influenced by the fractional area with water repellent soils and by the degree to which surface sealing occurs. Observations of these quantities in burned areas are largely missing. In addition, we can expect a greater amount of subsurface flow contribution to streams because of reduced interception and evapotranspiration. The magnitude of these effects is largely controlled by the severity and heterogeneity of the burn, with large homogeneous severe burns having the greatest potential for severe runoff events.

Although postfire runoff generation processes can be spectacular in their magnitude and results, they seem to be generally limited in the areal extent of their effect. This characteristic is critical to the evolved ecology of aquatic ecosystems in response to natural disturbances.

REFERENCES

Albini F., Amin M.R., Hungerford R.D., Frandsen W.H. and Ryan K.C. (1996) *Models for Fire-Driven Heat and Moisture Transport in Soils*, USDA Forest Service General Technical Report INT-GTR-335, U.S. Department of Agriculture, Forest Service, Intermountain Research Station, Ogden.

Benavides-Solorio J. and MacDonald L.H. (2001) Post-fire runoff and erosion from simulated rainfall on small plots, Colorado front range. *Hydrological Processes*, **15**, 2931–2952.

Billings W.D. (1994) Ecological impacts of Cheatgrass and resultant fire on ecosystems in the Western Great Basin. In *Proceedings – Ecology and Management of Annual Rangelands*, Monsen S.B. and Kitchen S.G. (Eds.), USDA Forest Service Intermountain Research Station: Ogden, pp. 22–30.

Bosch D.D. and Onstad C.A. (1988) Surface seal hydraulic conductivity as affected by rainfall. *Transactions of the American Society of Agricultural Engineers*, **31**, 1120–1127.

Campbell W.G. and Morris S.E. (1988) Hydrologic response of the Pack river, Idaho, to the sundance fire. *Northwest Science*, **62**, 165–170.

DeBano L.F. (1981) *Water Repellent Soils: A State-of-the-Art*, Pacific Southwest Research Station: USDA Forest Service General Technical Report PSW-46.

DeBano L.F. (2000) The role of fire and soil heating on water repellency in wildland environments: a review. *Journal of Hydrology*, **231–232**, 195–206.

DeBano L.F., Savage S.M. and Hamilton A.D. (1976) The transfer of heat and hydrophobic substances during burning. *Soil Science Society of America Journal*, **40**, 779–782.

Doerr S.H. and Moody J.A. (2004) Hydrological effects of soil water repellency: on spatial and temporal uncertainties. *Hydrological Processes*, **18**, 829–832.

Doerr S.H., Shakesby R.A. and Walsh R.P.D. (2000) Soil water repellency, its characteristics, causes and hydro-geomorphological consequences. *Earth-Science Reviews*, **51**, 33–65.

Dunham J.B., Young M.K., Gresswell R.E. and Rieman B.E. (2003) Effects of fire on fish populations: landscape perspectives on persistence of native fishes and nonnative fish invasions. *Forest Ecology and Management*, **178**, 183–196.

Dyrness C.T. (1976) *Effect of Wildfire on Soil Wettability in the High Cascades of Oregon*, USDA Forest Service Research Paper PNW-202, USDA Forest Service Pacific Northwest Forest and Range Experiment Station: Portland, Oregon.

Eagleson P.S. (2002) *Ecohydrology: Darwinian Expression of Vegetation Form and Function*, Cambridge University Press: Cambridge.

Franklin J.F. and Dyrness C.T. (1973) *Natural Vegetation of Oregon and Washington*, USDA Forest Service General Technical Report PNW-8, U.S. Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station, Portland, Oregon.

Harr R.D. (1976) *Hydrology of Small Forest Streams in Western Oregon*, USDA Forest Service General Technical Report PNW-55, USDA Forest Service, Pacific Northwest Forest and Range Experiment Station, Portland.

Hessburg P.F. and Agee J.K. (2003) An environmental narrative of Inland Northwest United States forests, 1800–2000. *Forest Ecology and Management*, **178**, 23–59.

Imeson A.C., Verstraten J.M., Van Mullingen E.J. and Sevink J. (1992) The effects of fire and water repellency on infiltration and runoff under Mediterranean type forests. *Catena*, **19**, 345–361.

Istanbulluoglu E., Tarboton D.G., Pack R.T. and Luce C.H. (2002) A probabilistic approach for channel initiation. *Water Resources Research*, **38**, 1325, doi:1310.1029/2001WR000782.

Johnston R.S. (1970) Evapotranspiration from bare, herbaceous, and aspen plots: a check on a former study. *Water Resources Research*, **6**, 324–327.

Kemper W.D. and Koch E.J. (1966) *Aggregate Stability of Soils from the Western United States and Canada*, USDA-ARS Technical Bulletin No. 1355, USDA Agricultural Research Service: Washington.

Klock G.O. and Helvey J.D. (1976) Debris flows following wildfire in north central Washington. *Proceedings of the Third Federal Interagency Sedimentation Conference. U.S. Subcommittee on Sedimentation*, Denver.

Kuczera G.A. (1987) Prediction of water yield reductions following a bushfire in ash-mixed species eucalypt forest. *Journal of Hydrology*, **94**, 215–236.

Letey J., Carrillo M.L.K. and Pang X.P. (2000) Approaches to characterize the degree of water repellency. *Journal of Hydrology*, **231–232**, 61–65.

Luce C.H. (1997) Effectiveness of road ripping in restoring infiltration capacity of forest roads. *Restoration Ecology*, **5**, 265–270.

Megahan W.F. (1983) Hydrologic effects of clearcutting and wildfire on steep granitic slopes in Idaho. *Water Resources Research*, **19**, 811–819.

Meyer G.A. and Pierce J.L. (2003) Climatic controls on fire-induced sediment pulses in yellowstone National Park and Central Idaho: a long-term perspective. *Forest Ecology and Management*, **178**, 89–104.

Miller D.J., Luce C.H. and Benda L. (2003) Time, space, and episodicity of physical disturbance in streams. *Forest Ecology and Management*, **178**, 121–140.

Mohammed D. and Kohl R.A. (1987) Infiltration response to kinetic energy. *Transactions of the American Society of Agricultural Engineers*, **30**, 108–111.

Moody J.A. and Martin D.A. (2001) Post-fire, rainfall intensity-peak discharge relations for three mountainous watersheds in the western USA. *Hydrological Processes*, **15**, 2981–2993.

Reifsnyder W.E. and Lull H.W. (1965) *Radiant Energy in Relation to Forests*, Technical Bulletin 1344, USDA.

Rieman B.E., Lee D.C., Burns D., Gresswell R., Young M., Stowell R., Rinne J. and Howell P. (2003) Status of native fishes in the Western United States and issues for fire and fuels management. *Forest Ecology and Management*, **178**, 197–211.

Ritsema C.J. and Dekker L.W. (2000) Preferential flow in water repellent sandy soils: principles and modeling implications. *Journal of Hydrology*, **231–232**, 308–319.

Rodriguez-Iturbe I. (1986) Scale of fluctuation of rainfall models. *Water Resources Research* **22**, 15S–37S.

Rodriguez-Iturbe I. and Mejia J.M. (1974) On the transformation from point rainfall to areal rainfall. *Water Resources Research*, **10**, 729–735.

Rowe P.B. (1948) *Influence of Woodland Chaparral on Water and Soil in Central California*, California Department of Natural Resources, Division of Forestry, and USDA Forest Service: Sacramento.

Seed A.W., Srikanthan R. and Menabde M. (1999) A space and time model for design storm rainfall. *Journal of Geophysical Research*, **104**, 31 623–31 630.

Shakesby R.A., Doerr S.H. and Walsh R.P.D. (2000) The erosional impact of soil hydrophobicity: current problems and future research directions. *Journal of Hydrology*, **231–232**, 178–191.

Sivapalan M. and Blöschl G. (1998) Transformation of point rainfall to areal rainfall: Intensity-duration-frequency curves. *Journal of Hydrology*, **204**, 150–167.

Swanson F.J. (1981) Fire and geomorphic processes. In *Fire Regimes and Ecosystem Properties, Proceedings of the Conference*, Mooney H.A., Bonnicksen T.M., Christensen N.L., Lotan J.E. and Reiners W.A. (Eds.), USDA Forest Service General Technical Report WO-26, USDA: Honolulu, pp. 401–420.

Troendle C.A. and King R.M. (1985) The Fool Creek watershed – thirty years later. *Water Resources Research*, **21**, 1915–1922.

Waring R.H. and Schlesinger W.H. (1985) *Forest Ecosystems, Concepts and Management*, Academic Press: Orlando.

Watson F.G.R., Vertessy R.A. and Grayson R.B. (1999) Large-scale modelling of forest hydrological processes and their long-term effect on water yield. *Hydrological Processes*, **13**, 689–700.

Whitlock C., Shafer S.L. and Marlon J. (2003) The role of climate and vegetation change in shaping past and future fire regimes in the northwestern US and the implications for ecosystem management. *Forest Ecology and Management*, **178**, 5–21.

Young J.A. and Evan R.A. (1985) Demography of *Bromus tectorum* in Artemisia communities. In *The Population Structure of Vegetation*, White J. (Ed.), Dr W. Junk Publishers: Dordrecht, pp. 489–502.

121: Intersite Comparisons of Rainfall-runoff Processes

JULIA JONES

Department of Geosciences, Oregon State University, Corvallis, OR, US

This article argues that despite the limitations of rainfall-runoff data, there are compelling reasons for hydrologists to conduct many more intersite comparisons of rainfall-runoff data. Inferences about hydrologic processes are drawn from an unnecessarily narrow subset of temporal scales, spatial scales, and geographic conditions given the range of data available. In intersite comparison of rainfall-runoff data, we face the same challenge as hydrologic modelers, namely: how can we discriminate among alternative mechanistic explanations for any given rainfall-runoff (or runoff-runoff) dataset? This article (1) provides a justification for additional intersite comparisons given the history and lessons learned from rainfall-runoff and related studies, (2) demonstrates how intersite comparison helps discriminate among alternative hydrologic mechanisms, and (3) outlines steps involved in conducting intersite comparative analysis with a particular focus on analysis of primary data. The article argues that the best approach to intersite hydrology is (re)analysis of original data, which allows us to (1) expand the sample size to include data not yet analyzed and replicate hypothesis tests; (2) ask new questions, even of old data, posed by the current context for hydrology; (3) make new comparisons among records not formerly juxtaposed; (4) use novel statistical approaches to reveal hitherto obscure features of the data, and (5) use ancillary data to refine hypothesis tests about hydrologic mechanisms. The article provides suggestions for the steps involved in intersite comparison including (!) identifying a question, (2) developing a study design, (3) selecting, accessing, and merging datasets, and (4) choosing statistics for comparison of rainfall and runoff data. The endeavor of intersite comparison of rainfall and runoff data is analogous and complementary to the parallel quest for methods of parameter identification and model structure selection in hydrologic modeling.

OBJECTIVE AND SCOPE OF THIS ARTICLE

Of the tools available to hydrologic science – process studies, modeling, and analysis of rainfall-runoff data – data analysis has taken a far back seat over the past few decades. Reasons for this neglect are varied, but many hydrologists are deeply disparaging of the potential for new insights from rainfall-runoff data. A common view is that it is impossible to gain insights about hydrologic mechanisms from rainfall-runoff data because they are inevitably limited to a black-box form of analysis.

This study takes the opposite view, namely, that a great deal can be learned from intersite comparison of rainfall-runoff data. Our inferences about hydrologic processes are drawn from an unnecessarily narrow subset of temporal scales, spatial scales, and geographic conditions, given the range of data available. A distressingly large number of publications simply repeat or challenge the results of a few primary analyses, without exploring unexamined datasets.

In analysis of rainfall-runoff data, we face the same problem as that faced by hydrologic modelers. Hydrologic modelers have great difficulty in identifying the correct parameters from among the very large set of parameters that can optimize any given rainfall and/or runoff dataset applied to a given model structure (Beven, 2000). Moreover, hydrologic modelers must find some means of discriminating among alternative model structures, other than optimizing models to fit rainfall and/or runoff datasets. In intersite comparison of rainfall-runoff data, we face the same challenge, namely: **how can we discriminate among alternative**

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

mechanistic explanations for any given rainfall-runoff (or runoff–runoff) dataset?

The creative analysis of multiple rainfall-runoff records can help test hypotheses about hydrological processes. This optimism is based on the following observations:

1. *A great many rainfall-runoff datasets have never been examined*. Even in datasets that have been analyzed, there exist time periods and temporal resolutions that have never been examined. Because many bitter controversies in interpretation of hydrologic processes hinge upon hypotheses tested using very small samples, replicating these hypothesis tests with enlarged samples of basins and time periods is an obvious first step toward constructing hydrologic theory from analysis of observed data.
2. *New questions can be asked of rainfall-runoff records*, even those that have been analyzed to answer old questions. Old questions that have dominated the analysis of rainfall-runoff data include (i) (forest harvest) treatment effects, (ii) tests of unit hydrograph, and (iii) flood and water yield forecasting. New questions include tests for (i) trends related to climate and/or vegetation and (ii) scaling of rainfall and runoff in space and time. Prediction of runoff in ungaged basins is an ongoing question.
3. *New comparisons may be made among records not previously compared*. The most common intersite comparison has been of runoff in small, treated-control basin pairs within a site. However, many opportunities exist for other kinds of comparisons. For example, (i) two or more control basins can be compared to reveal the influence of some factor that differs between them, (ii) a small basin may be compared to a larger basin that contains it to reveal the effect of spatial scale and routing, or (iii) two or more large basins may be compared to reveal the influences of factors that differ between them. It may even be possible to compare (iv) two basins in very different, distantly spaced sites, to infer the progression of some external forcing, such as climate change.
4. *New forms of statistical analyses may reveal new aspects of rainfall-runoff data*. Intersite comparisons of rainfall-runoff data have been limited to a few statistical methods, which provide limited insights into the patterns of rainfall and runoff. Linear regression, such as of runoff at a treated versus a control basin, has dominated analyses in the past. However, many other statistical techniques such as analysis of variance, autocorrelation methods, cross-correlations, and even spectral analysis or wavelets, can reveal hitherto unknown aspects of rainfall-runoff relationships among sites.
5. *Ancillary datasets are now available to reveal processes operating inside instrumented basins.* With improved computer technology and Internet access, rainfall and runoff data can be readily accessed and merged with rainfall-runoff data. Examples of ancillary data include historical climate data, digital elevation data, maps of geology, soils, and vegetation, and process studies and experiments. In combination with greater computer power and the ready availability of multiple statistical analysis tools, this information can be used to stratify and subdivide rainfall records to more precisely test hypotheses about the hydrologic mechanisms operating in instrumented basins of all sizes and types.

Intersite comparison of rainfall-runoff data involves **analysis of primary datasets, using new data, new questions, novel comparisons, novel statistical approaches, and ancillary datasets to discriminate among alternative mechanistic hypotheses about hydrologic processes**. Records from all gaged basins are relevant, irrespective of basin size, climate type, ecosystem or vegetation type, continent, time period, or length of record. Until recently, primary analyses of rainfall-runoff data have attempted to draw inferences on the basis of the behavior of streamflow, precipitation, or runoff ratios using a single basin at one site, or a single paired basin. *Intersite hydrology comparison studies enlarge this perspective to include multiple basins within a site or across sites, or multiple paired-basin experiments within a site, or across sites.*

The objective of this article is to describe intersite analyses of rainfall-runoff processes, and evaluate their potential and limitations for contributing to basic advances in hydrology. The article consists of

- justification for additional intersite comparisons given the history and lessons learned from rainfall-runoff and related studies
- how intersite comparison helps discriminate among alternative hydrologic mechanisms, and
- an outline of steps involved in conducting intersite comparative analysis with a particular focus on analysis of primary data.

JUSTIFICATION FOR INTERSITE COMPARISON STUDIES

Analysis of primary rainfall and runoff data is by far the most practical and promising approach to intersite comparisons aimed at elucidating fundamental hydrologic processes. Analysis of primary data is necessary because to date, analyses of rainfall-runoff data have not produced accepted generalizations or theory. On the contrary, rainfall-runoff studies, as well as reviews and metaanalyzes of these studies, have left a great many questions and controversies about mechanisms that control runoff.

Some authors have approached intersite comparison by conducting literature reviews or summarizing findings

(e.g. Robinson *et al.*, 2003). A somewhat more formal, but still qualitative, approach involves tabulating streamflow studies. These approaches to intersite comparison are weak because they are qualitative, and rely upon the original authors' assessments of whether findings were significant, independent of sample size or the type of test used. Such approaches also may lead to incorrect synthesis, particularly when they involve many studies with small sample sizes (e.g. short hydrologic records).

Metanalysis overcomes some of the limitations of reviews or summary tables, but it is still limited to answering the questions posed by the original authors. Metanalysis is a quantitative approach to synthesizing results from multiple studies (Gurevich and Hedges, 2001). In metanalysis, studies that share common methodologies – usually an experimental treatment – are compared, so that collective results are not confounded by differences in approaches. Metanalysis techniques estimate the average effect of some treatment across a range of studies, determine whether that average effect is significantly different from zero, and examine whether differences in effects can be attributed to certain characteristics of studies. The "average effect of the treatment" is a measure of the difference between treated and control sample units (e.g. instrumented watersheds) and may be a simple difference or a log-transformed difference (Gurevich and Hedges, 2001). These differences may be weighted to account for differences in sample sizes.

Reviews, summaries, and metanalysis of hydrologic data have focused on predicting the effects of forestry treatments on water yield, flooding, and nutrient/sediment fluxes, although many other questions could potentially be examined. For example, Hibbert (1967) examined the relationship between annual water yield responses and forestry treatments, and concluded that water yield was inversely related to forest cover, although responses to forestry treatments were highly variable and unpredictable. Bosch and Hewlett (1982) expanded Hibbert (1967) study, collecting data on the maximum increase in water yield in the first five years following reduction of forest cover from 94 paired watershed experiments around the world. They concluded that streamflow increases were weakly positively related to the proportion of forest cover removed, and the strength of the relationship between percent forest removed and annual streamflow increase was greatest for conifers, followed by hardwood and scrub vegetation. Bosch and Hewlett (1982) also noted that annual streamflow response was positively related to mean annual precipitation, at least for the sample of watersheds with conifers. The inverse relationship between forest cover and water yield from these studies gave rise to a large set of studies examining the potential for water yield augmentation from forestry treatments. These studies (e.g. Harr, 1983) revealed that short-term water yield increases were not preserved as forests regenerated, and, furthermore, streamflow surpluses often occurred in wet seasons when downstream users did not need additional water. More recent reviews, such as Stednick (1996), also have evaluated the effects of forestry treatments on water yield. More recently, metaanalyzes have examined water yield and flooding consequences of reforestation. For example, based on a metanalysis of 28 small basin studies across Europe, Robinson *et al.* (2003) concluded that reforestation was unlikely to have much effect on floods or lowflows, except where conifer plantations were established on poorly drained soils.

Many intersite comparisons in the form of reviews, summaries, or metaanalyzes have addressed the relationships of rainfall-runoff processes to nutrient and sediment. Binkley and Brown (1993) extended the metanalysis approach to examine the effect of forestry treatments on stream temperature, dissolved oxygen, nitrate, and suspended sediment based on studies in the United States. They concluded that "best management practices" were able to mitigate undesirable changes in most of these properties, except for occasional large storm event-related sediment pulses. Martin *et al.* (1984) and Hornbeck *et al.* (1997) summarized and evaluated the implications of forestry treatments for water quality with a focus on the eastern United States. Binkley *et al.* (2004) summarized the nitrogen and phosphorus concentrations of streams draining forests in the United States, but found no factors that explained the variation among sites.

Reviews, summaries, and metanalysis are an inadequate approach to intersite hydrology comparison in many instances. Reviews are limited to repeating the conclusions drawn by the authors, which may be limited or biased, depending on the experimental design, sample size, and analysis. Frequently, experiments combined in a review or summary have small sample sizes, which lack the power to detect treatment effects. Thus, a summary or evaluation of a set of experiments may provide misleading conclusions biased toward no detected effect (see discussions in Hedges and Olkin, 1985). Although metanalysis is quantitative, and does not rely upon the conclusions of the original authors (Gurevich and Hedges 2001), it is still a weak form of synthesis because it is limited to estimates of effects determined by the original authors of the studies.

HOW INTERSITE COMPARISON HELPS DISCRIMINATE AMONG ALTERNATIVE HYDROLOGIC MECHANISMS

The essence of an intersite comparison is that it examines primary rainfall and/or runoff data to determine the difference, or change, between two or more datasets, at two or more time periods, in order to elucidate hydrologic processes. The best-understood example of this approach is paired-basin forest harvest treatment experiments, in which the ratio of the runoff at the treated and control

basin is compared between the pretreatment and the treated period to infer how the treatment (removal or replacement of vegetation) has affected runoff. *In this article, intersite comparisons encompass all analyses in which rainfall and/or runoff is compared among any two or more sites or two or more time periods, with an aim to inferring hydrologic mechanisms.*

Intersite comparison, when it involves analysis of primary data, can be a useful tool for discriminating among alternative process explanations for observed rainfall-runoff relationships. Analysis of primary rainfall-runoff data can overcome many of the deficiencies of secondary analyses such as literature review and metaanalyzes. Analysis of original data allows us to

- expand the sample size to include data not yet analyzed and replicate hypothesis tests;
- ask new questions, even of old data, posed by the current context for hydrology;
- make new comparisons among records not formerly juxtaposed;
- use novel statistical approaches to reveal hitherto obscure features of the data; and
- use ancillary data to refine hypothesis tests about hydrologic mechanisms.

Replicating Hypothesis Tests and Expanding Sample Size

Intersite comparison studies can expand the sample size of records tested for certain hydrologic responses, and thereby replicate hypotheses tested in earlier studies. Our inferences about hydrologic processes are drawn from an unnecessarily narrow subset of temporal scales, spatial scales, and geographic conditions, given the range of data available (Figure 1).

A vast amount of rainfall and runoff data have been collected, but never analyzed (Figure 1). Rainfall and runoff data have been collected for centuries to aid in predictions of water yield, flooding, and other hydrological properties. Over the past century, and particularly the past 50 years, streamflow gaging technology has permitted the acquisition of continuous records with very fine temporal resolution (5 or 15 minutes). Sites with streamflow gaging have proliferated and number in the millions, with more than 1.5 million sites gaged historically by the United States Geological Survey alone. Continued developments in computer database technology and informatics have permitted data to be digitized and stored for ready access. Many rainfall and runoff records, especially those managed by government agencies, are now publicly available on the Internet.

The fact that rainfall and runoff records are held by a variety of agencies may partly explain why some records have never been analyzed. Streamflow and precipitation have been monitored systematically in basins in the United States and abroad since the early to mid twentieth century. In the United States, such monitoring has occurred under the auspices of individual states, the US Forest Service, the US Geological Survey, the USDA Agricultural Research Service, and state agencies. In the United Kingdom, monitoring is conducted under the auspices of the Center for Ecology and Hydrology. In Australia, monitoring is conducted by the water resources branch of the CSIRO, Forestry Tasmania, and state and municipal agencies. In Japan, streamflow monitoring is conducted by the Forestry and Forest Products Research Institute and Japanese universities. All countries have some form of streamflow monitoring, although record length, quality, and availability vary.

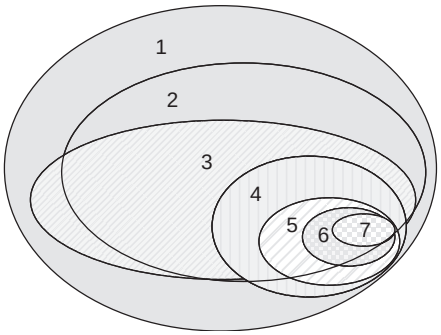


Figure 1 Most inferences from data about hydrologic processes are drawn from an unnecessarily narrow subset of temporal scales, spatial scales, and geographic conditions, given the range of data available. A distressingly large number of publications simply repeat (or challenge) the results of a few primary analyses without exploring unexamined datasets. Only a small fraction of the published studies in hydrology (1) utilize any measured streamflow data (2) (boxes not to scale). The majority of these studies are based on data from small basins (3) despite the enormous numbers of gaged large basins. Of the small, instrumented basins whose records are noted in any publication (3), only a fraction have been analyzed (4), so that many basin records remain unexamined. Few of these studies involve analysis of data at finer than annual time steps (5), and only a fraction of studies utilize records longer than a decade (6), so very few studies examine daily data over long periods comparable to most hydrologic modeling efforts (7). Apart from analyses of instantaneous peak flows, virtually no studies have compared streamflow data among sites at finer than daily time steps. Thus, there are at least four opportunities for intersite comparisons involving hypothesis testing and expanded data use: (a) using data to verify modeling and process studies, which have not involved measured data; (b) analyzing large basin records that have never been looked at; (c) analyzing small basin records that have never been looked at; and (d) extending the time periods, or increasing the temporal resolution (e.g. from annual to daily, daily to hourly) of analyses for basins whose records have been examined. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Also, intersite comparisons may be made more difficult by the fact that agency jurisdiction and consequent objectives and study designs in monitored basins vary with basin size. Small basins are more likely to be instrumented for research and forest or agricultural land use management, whereas large basins are more likely to be instrumented for water management and supply assessment objectives. In the United States, most basins monitored for research purposes by federal agencies (USFS of the USDA, NCRS of USDA, USGS research branch) are small, usually $<1\,km^2$, usually with a single land owner and one dominant vegetation type. In contrast, basins monitored for water supply and flood control purposes may drain areas up to millions of square kilometers and typically encompass multiple land uses and vegetation types.

Intersite comparisons could replicate a number of hypotheses that have been advanced about hydrologic responses, mostly based on experiments designed to test the effect of vegetation removal on runoff. These hypotheses are based on analyses conducted from records using a small number of basins, limited record length, basins of only one – usually small – size, and only one temporal resolution – usually annual runoff (Figure 1). Several key hypotheses merit replicated testing, including statements (from Hibbert, 1967; Bosch and Hewlett, 1982; Jones and Grant, 1996; Beschta *et al.*, 2000) that the hydrologic response to vegetation treatment is positively related to

- the amount of vegetation removed;
- the amount of precipitation;
- the evergreen-ness of the vegetation;

and negatively related to

- the amount of runoff;
- the size of the basin.

Analyses of primary rainfall and runoff data using more sites, longer records, and finer temporal resolutions of data could replicate tests of these hypotheses, and resolve contradictions among them. Reviews (e.g. Hibbert, 1967, Bosch and Hewlett, 1982) have included many small paired-basin records in the initial posttreatment period. However, the primary data from many of these small, paired basins have not been reanalyzed, although much of these data are now available through the Internet *(see* e.g. the Clim-DB, Hydro-DB website **`http://www.fsl.orst.edu/climhy/`**). Moreover, the primary records from other small basin sites that were not part of paired-basin experiments, recently established paired-basin experiments, and a great many large basins monitored for flood control and water yield have not been analyzed (Figure 1). In cases where streamflow and climate records are continually expanding, new trends may emerge from examination of longer records. The treated-control relationships in small paired-basin experiments summarized by Bosch and Hewlett (1982) could be reanalyzed using longer records to ascertain how vegetation succession alters water yield over time. Examination of finer resolution data may help discriminate multiple scales of pattern indicative of several interacting hydrologic mechanisms. For example, Jones and Post (2004) showed that basins with similar hydrologic response to forest harvest at the annual streamflow resolution had very different day-to-day patterns of hydrologic response. Continuous records of streamflow or precipitation that may have been analyzed at one temporal resolution (such as annual or monthly) can be resampled at finer resolutions (such as storm, daily, or hourly) to seek evidence for hydrologic processes that are not detectable at monthly or annual time scales.

Posing and Answering New Questions

Often, intersite hydrology involves asking new questions about old data. Typically, streamflow monitoring studies were established for reasons that are no longer principal motivating factors for research or management. For example, an original objective for many USGS instrumented basins included establishing flood frequency curves for designing downstream engineering structures such as dams or bridges. However, in many cases, the construction of engineering works has been completed, while upstream land use may have changed the flood frequency distributions. Also, many USFS sites established paired watershed experiments to determine how water quantity would be affected by forest harvest treatments. However, forest harvest and road building practices have changed and no longer resemble the treatments imposed in the experiments.

Changing contexts have created important new questions about rainfall-runoff processes. New areas of concern include considerations of climate change, scaling, and the ongoing challenge of prediction in ungaged basins. For example, we can ask how the rainfall-runoff relationship has changed over time, and interpret those findings to infer changes in climate and vegetation. Also, we can examine how the rainfall-runoff relationship varies with spatial scale to understand how perturbations propagate in streamflow networks. We can combine these two general questions to determine how the rainfall-runoff relationship varies as a function of basin scale and the temporal resolution of the data to help scale up our understanding of hydrologic processes to that of global climate models. We can also ask how the geography of basins influences the rainfall-runoff relationship based on intersite comparisons, and perhaps derive principles on which to base predictions of runoff in ungaged basins.

Making Novel Intersite Comparisons

A key component of intersite comparison of rainfall and runoff involves the choices of basins to compare. Intersite hydrology involves comparisons of rainfall and runoff

Table 1 Novel comparisons among basins represent new experimental designs for intersite comparison to test a variety of hydrologic mechanisms influencing rainfall-runoff relationships. Examples are based on runoff data, but could be extended to rainfall/runoff ratios in cases where both rainfall and runoff data exist for all the sites

Compare	To	Study question	Hydrologic interpretation
Runoff at treated basin (vegetation removal, roads)	Runoff at control basin	Was there a detectable effect of an experimental treatment?	Water balance altered by vegetation removal and/or roads routing water
Runoff at treated basin, by time period since treatment	Runoff at control basin, by time period since treatment	How did the treatment effect change over time?	Water balance affected by vegetation recovery or other covarying factors
Runoff at treated basin, by season	Runoff at control basin, by season	How did the treatment effect vary by season?	Water balance affected by seasonal variation in wetness, temperature, and/or other factors
Runoff at treated, by event size	Runoff at control, by event size	How did the treatment effect vary by the size of the runoff event?	Water balance affected by antecedent wetness, precipitation, snow, and/or other factors
Runoff at multiple treated basins, by time since most recent pretreatment disturbance	Runoff at multiple control basins, by time since most recent pretreatment disturbance	How did the treatment effect differ among basin pairs according to the time since most recent pretreatment disturbance?	Water balance affected by changes in vegetation structure and species composition over succession
Runoff at small basin	Runoff at large basin containing the small basin	How does runoff vary with basin scale?	Hydrograph affected by routing of water in network and/or differences in vegetation type and cover and/or roads small to large
Runoff at basin in climate type a	Runoff at basin in climate type b	How does runoff vary with climate?	Water balance affected by differences in timing of moisture, freezing
Runoff at basin in vegetation type a (or land use type a)	Runoff at basin in vegetation type b (or land use type b)	How does runoff vary with vegetation type (or land use)?	Water balance affected by differences in timing, amount of transpiration, evaporation, interception

among basins according to the differences in those basins, whether differences arise from "controlled" or inadvertent experiments, or geographical differences. It involves examining flow regimes and responses to treatment across divergent vegetation, hydrologic systems. Intersite comparison studies could make important contributions to inferring the importance of various hydrologic processes by examining novel pairings of basins that have not been compared before (Table 1). Obvious comparisons involve (i) multiple neighboring or nearby basins, and (ii) nested basins.

Multiple small, neighboring, or nearby basins are a logical choice for intersite comparisons. Small basins are instrumented at many United States Forest Service-sponsored experimental forests (e.g. Andrews, Caspar Creek, Coweeta, Fernow, and Hubbard Brook Experimental Forests) as well as United States Department of Agriculture Agricultural Research Service sites (e.g. Reynolds Creek, Walnut Gulch). When multiple basins are instrumented in a single site, they allow examination of environmental variation and treatment effects on streamflow. Some sites have implemented one or more "paired watershed experiments" in which a small control and one or more small treated (usually adjacent) basins are monitored over some pretreatment period, and then a treatment, such as vegetation removal, is imposed on one basin, and the monitoring of both the treated and control basin is continued for some posttreatment period. Bosch and Hewlett (1982) identified nearly 100 such treated/control basin pairs in several dozen locations with experiments involving modification of forest, shrub, and grassland vegetation.

Multiple large basins also could be compared using intersite analysis. Large basins generally are instrumented for flood and water yield forecasting. Large paired-basin experiments do not exist because of the difficulty of implementing a treatment over a large area and because of the lack of large basins in a "control" state. However, neighboring or nearby large basins with contrasting vegetation histories can be considered as parts of inadvertent experiments (*see* e.g. Jones and Grant, 1996; Thomas and Megahan, 1998). Nested basin comparisons are possible when small, instrumented basins occur inside large, instrumented basins (see e.g. the Andrews Forest, Hubbard Brook, Coweeta). Nested basins allow the testing of scaling relationships (Gupta and Waymire, 1998).

Table 2 Novel statistical approaches, multiple sites, and longer records permit asking novel questions in intersite comparisons of rainfall and runoff data

Questions	Number of Variables	Sites	Periods	Analysis tool	Illustrative citations
1. What is the mean value of rainfall, runoff?	One	One	One	Mean	Post *et al.* (1998)
2. What is the variation in rainfall, runoff?	One	One	One	Variance, standard deviation, coefficient of variation	Post and Jones (2001)
3. What is the shape of the distribution of rainfall, runoff?	One	One	One	Mean, variance, and higher moments; quantiles	Gupta and Waymire (1989)
4. How has the mean value of rainfall, runoff varied over time?	One	One	Two or more	Means	Andreassian *et al.* (2003)
5. How is rainfall, runoff related to itself over time?	One	One	Two or more	Time-series methods: autocorrelation, spectral analysis, wavelets	Kirchner *et al.* (2000) Lafrenière and Sharp (2003) Tague and Grant (2004)
6. How is runoff (rainfall) related across sites?	One	Two or more	One	Linear regression	Hibbert (1967), Bosch and Hewlett (1982), Andreassian (2004)
7. How does the shape of the rainfall or runoff distribution change across sites or time?	One	Two or more	One or more	Statistical self-similarity; quantiles	Gupta and Waymire (1993, 1998), Andreassian *et al.* (2003)
8. How does the rainfall or runoff relationship between two sites change over time?	One	Two	Two or more	Linear regression for each time period; anova by time period	Jones and Grant (1996), Thomas and Megahan (1998), Beschta *et al.* (2000), Jones (2000), Jones and Post (2004)
9. How is rainfall related to runoff?	Two	One	One	Rainfall-runoff ratio	Post and Jones (2001), Kokkonen *et al.* (2004)
10. How is rainfall related to runoff at a site over time?	Two	One	Two or more	Cross-correlation; cross-spectral, cross-wavelet	Post and Jones (2001), Lafrenière and Sharp (2003)

Novel Statistical Approaches Reveal Hitherto Unquantified Aspects of Data

Most of the published analyses of rainfall and runoff data are based on a narrow set of statistical tools. Linear regression has been widely used. However, other techniques, including time-series analysis, autocorrelation, and cross-correlation approaches have the potential to reveal aspects of rainfall and runoff data that could be compared among sites, and lend themselves to interpretations of physical or biological processes (Table 2).

New Ancillary Data to Refine Hypotheses

Counter to the assertion that rainfall-runoff data are a "dead end" for determining hydrologic processes, this article contends that intersite comparisons of rainfall and runoff data can shed light on the hydrologic processes operating within the basin. While such interpretations are not definitive, rainfall, runoff, and ancillary data can be used to narrow the range of possible interpretations of hydrologic mechanisms operating at a given site for a given time period. In so doing, researchers are engaged in a process analogous to that conducted by hydrologic modelers in selecting the appropriate model structure and identifying correct parameters.

Intersite comparisons involve drawing inferences about the water balance – revealing or inferring the processes inside the black box. The water balance is the basic theoretical tool for the interpretation of analyses of rainfall-runoff data (Figure 2). Any analysis of rainfall and runoff confronts the difficulty of discriminating which of the large number of terms in the water balance explain the hydrologic processes at a site under a given set of conditions (Figure 2a).

Intersite comparisons can make progress by utilizing ancillary data to narrow down the number of possible terms in the water balance that could be operating under certain conditions, thereby reducing the number of possible mechanistic interpretations (Jones and Post, 2004). If certain moisture, temperature, or eco-physiological conditions are

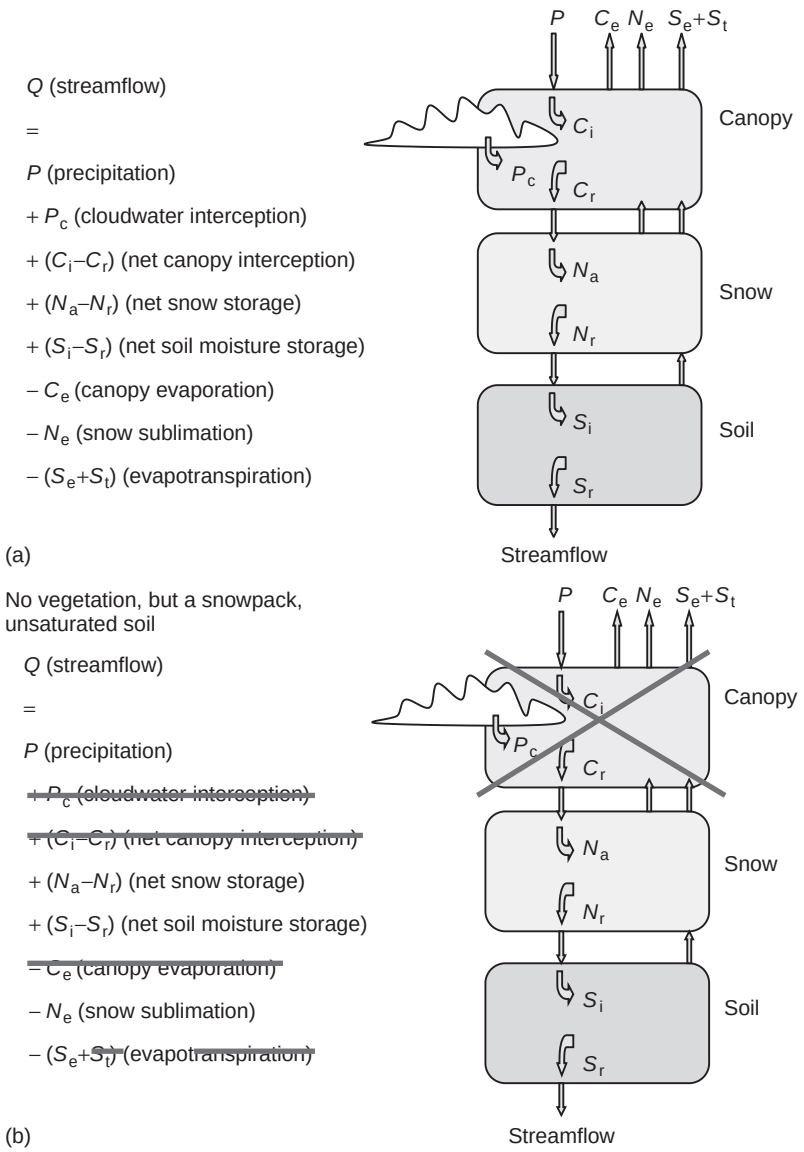


Figure 2 An approach to simplifying the number of possible hydrologic processes that can explain rainfall-runoff data involves partitioning data into subsets, each of which can be argued to involve a subset of terms in the water balance. Thus, the complete water balance for a given site (a) can be simplified into a set of water balances for conditions when (b) vegetation is absent, or (c) both vegetation and snowpack are absent, or (d) snowpack is absent and soils are near-saturated, or (e) vegetation is not actively transpiring, snowpack is absent, and soils are near-saturated. Analyses of rainfall-runoff relationships for each of these subsets of data can help bracket the range of values for the terms in the water balance. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

met, key terms can be effectively removed from consideration in the water balance for a given place or time period (Figure 2b–2e). For example, if vegetation is absent, soils are unsaturated, and a snowpack is present, any observed change in streamflow relative to precipitation must involve snow and soil moisture storage only (Figure 2b). If, in addition, no snowpack is present, the water balance can be simplified to consider only the effects of the soil moisture storage (Figure 2c). Alternatively, when vegetation is present, but snow is absent, and if the soil is saturated, then the water balance is regulated by interception, evaporation, and transpiration processes in the vegetation canopy (Figure 2d). To further simplify the water balance, it may be possible to identify periods when vegetation is not actively transpiring, and the water balance can be reduced to the interception and evaporation terms (Figure 2e).

This kind of approach can guide intersite comparison studies, because data are collected on related climate variables in addition to precipitation and streamflow at most

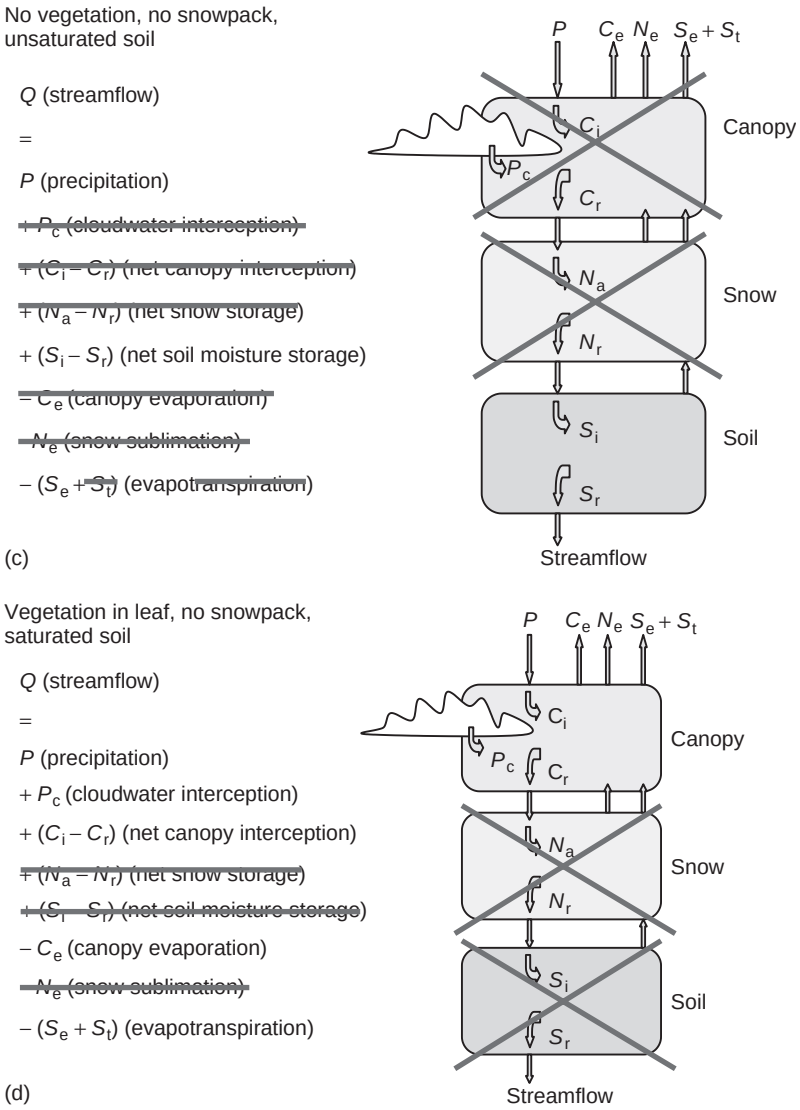


Figure 2 (*continued*)

instrumented sites (see the Reynolds Creek dataset publications for an excellent example, described in Slaughter *et al.*, 2001; Pierson *et al.*, 2001). Instrumented variables differ among sites, and, over time, within a site, depending upon the objectives for the site and whether the data have been or are being utilized for research and/or management.

AN OUTLINE OF STEPS INVOLVED IN CONDUCTING INTERSITE COMPARATIVE ANALYSIS

The steps involved in conducting an intersite comparison are

- identifying a question;
- developing a study design;
- selecting, accessing, and merging datasets; and
- choosing statistics for comparison of rainfall and runoff data.

Identifying a Question

Many questions are possible for intersite comparisons (Tables 1 and 2). The list below (and Tables 1 and 2) contains questions that hydrologists can answer using intersite comparison. Answers to these questions would represent important contributions to the current state of knowledge in hydrology.

1. *Are basins unique, or are there consistent types of behavior among basins?*

Intersite comparisons among multiple instrumented basins may reveal consistent groupings of basin behavior. If

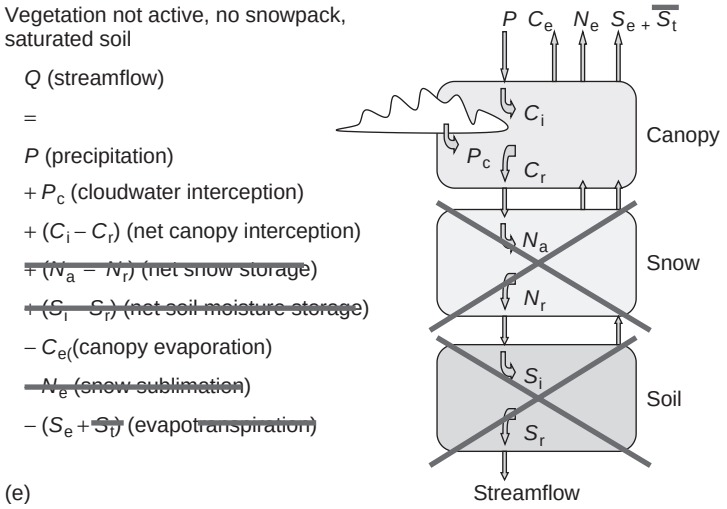


Figure 2 (*continued*)

consistent types of rainfall-runoff behavior can be identified for groups of instrumented basins, a classification system may be developed, which could facilitate the prediction of runoff from ungaged basins (see the IAHS PUB initiative).

2. *What factors (climate, vegetation, subsurface properties, land use history) explain similarities and differences in streamflow among basins?*

Intersite comparisons can involve *post hoc* superexperiments, in which rainfall-runoff responses associated with one factor (such as vegetation) can be examined while controlling for other factors (such as climate, land use history). Understanding, and separating, the roles of climate, vegetation, and other factors will help elucidate hydrologic mechanisms that account for streamflow in ungaged basins.

3. *Can streamflow be predicted in ungaged basins based on a priori knowledge?*

Records from multiple instrumented basins could be examined in such a way as to test models for predicting streamflow in ungaged basins.

4. *Are precipitation, streamflow, or rainfall-runoff relationships changing over time?*

Rainfall-runoff relationships are the basis for predictions of water yield and flooding and other practically relevant aspects of hydrology. Intersite comparison of long-term rainfall-runoff records can reveal changes in these relationships, and, potentially, the factors that cause the changes.

5. *How does the variability of streamflow and rainfall change with time scale – diurnal, storm, seasonal, or the time scales of vegetation succession and climate change?*

Existing records of streamflow and precipitation have high temporal resolution, potentially revealing processes occurring at multiple time scales. However, most published analyses of these records have dealt with annual or mean annual data. Intersite analyses of these records may reveal hydrologic processes that produce streamflow responses at diurnal or storm time scales, while other analyses may reveal processes operating at the time scales of vegetation succession and climate change.

6. *Are changes over time in streamflow occurring in response to vegetation change, climate change, or other drivers?*

Intersite comparisons can involve examination of changes in basins whose vegetation and climate history is well documented, potentially revealing responses to vegetation, climate change, natural and anthropogenic disturbances, and their influences on streamflow.

7. *Do basins respond differently to similar perturbations or treatments? If so, what accounts for differential responses?*

Intersite comparisons can investigate the causes of divergent streamflow responses to similar treatments (such as vegetation removal). For example, the presence of snow, soil depth and texture, and the seasonality of vegetation water use; all may produce different responses to the same treatment in different basins.

8. *Do basins respond predictably to perturbations or treatments?*

Very few of the existing long-term rainfall/runoff records are regularly used for testing hydrological models. Yet, the lengths of these records, the detailed associated knowledge about these basins, and the differences among them provide many opportunities for testing hydrologic model predictions.

9. *Are streamflow responses to perturbations or treatments linear or nonlinear? Are there thresholds of response?*

Intersite analysis of rainfall-runoff records offers the opportunity to seek nonlinear responses, or thresholds, in hydrologic system behavior.

Developing a Study Design

The second step in an intersite hydrology comparison is to establish an experimental design. The experimental design involves identifying a comparison that appropriately tests the study question (Table 1). It also involves selecting the number of basins, variables, and time periods for the analysis (Table 2). In an intersite comparative analysis of primary rainfall and runoff, data are included from a sample of basins selected because the hydrologic processes are expected to differ among these basins in a predictable fashion. Study basins may be selected to replicate the effects of a given treatment (deliberate or inadvertent), or to examine rainfall or runoff relationships across a gradient of conditions.

A number of decisions are made in designing any intersite comparison. These include the following:

1. Choosing the study basins.
2. Choosing a temporal resolution.
3. Stratifying data by time period, season/climatic conditions, event size, or other factors.

The study basins should be chosen so as to allow testing of the question of interest, usually by selecting basins that differ from one another according to the factor of interest (whether or not this difference was the result of a human-imposed treatment) (Table 1). The temporal resolution of the data (ranging from 15 minutes to annual or decadal) also should be chosen so as to permit testing of the question of interest. For example, questions involving the timing and routing of runoff (e.g. flooding) focus on the time scale of individual storms (peak discharges, storm hydrographs), whereas questions involving the amounts of water in various components of the water balance may focus on hourly, daily, or annual data. Because the influence of hydrologic mechanisms varies according to the conditions, it can be quite important to stratify data by time period (e.g. age of vegetation, or time since treatment), by season or climatic conditions (e.g. subzero vs warm temperatures, wet vs dry conditions), by event size (e.g. amount of precipitation or runoff), or other factors.

Selecting, Accessing, and Merging Datasets

Selecting, accessing, and merging datasets is the third step in an intersite comparison. Data availability poses some restrictions on the types of intersite hydrology comparisons that can be made (Figure 3). Streamflow, precipitation, temperature, snow, vegetation, geology, topography, and other records are all pertinent to intersite comparisons. Each form of data has its own advantages and weaknesses.

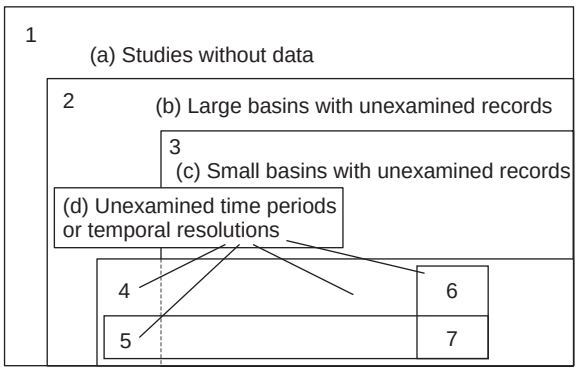


Figure 3 Data availability poses some restrictions on the types of intersite hydrology comparisons that can be made (ellipses not to scale). Many basins have streamflow records (1), but only some have been digitized (2). Ancillary data on precipitation and temperature, not all of which is digitized, are available only for some basins (3). Of the total number of gaged basins, only a few involve treated/control pairs, and not all of these records are digitized (4). Ancillary data on vegetation, snow, and other ecological factors is available for only a subset of basins, whether they are treated-control pairs or not (5). Online datasets that involve streamflow, precipitation, temperature, and ancillary data on vegetation, snow, and other ecological factors are only a small fraction of the total data, and these are limited to daily time resolutions (6). A very few sites have online data available at finer than daily time steps (7)

Streamflow measurements vary in accuracy and precision, and these differences among sites determine the minimum change that can be detected in an intersite comparison of runoff data. In small basins, streamflow is measured using flumes or weirs with controlled cross sections, so stage height observations are readily converted to discharge. The greatest uncertainty, however, is associated with the largest flows for which measurements are most difficult to obtain. In large basins, discharge is calculated on the basis of the measured stage height on a known cross section; changes in the cross section associated with flooding add additional uncertainty to discharge estimates from large basins. Streamflow records typically are continuous, so, if the original charts are saved, or the record is digitized, it is possible to capture streamflow patterns at very fine temporal resolutions, including small diurnal fluctuations (*see* e.g. **`www.fsl.orst.edu.lter/data/hydrology`**). The weir, flume, or cross section determines the ability of the streamflow record to capture small variations in discharge. For example, trapezoidal flumes typically cannot capture very small fluctuations as well as V-notch weirs. Thus, the minimum detectable change in a given streamflow record is smallest for records collected in small basins using V-notch

weirs, and largest for records collected in large basins where cross sections change frequently.

Precipitation and air temperature records are also typically available from sites adjacent or near to streamflow gages, but it is challenging to match point measurements of precipitation with area measurements of streamflow. Capturing the spatial variability of precipitation (or air temperature) is problematic, particularly for large basins, and especially for particular time periods in the past. Most small basin streamflow studies involve one or more precipitation and air temperature gages. When multiple precipitation (air temperature) gages are present, it is possible to spatially interpolate precipitation or temperature over the basin (Daly *et al.*, 2002; Smith, 2002). Maps of interpolated precipitation can be found at **http://www.ocs.orst.edu/prism/**, and illustrative maps of temperature interpolations can be found at **http://www.ocs.orst.edu/pub/smithjw/hja/**).

Accurate representation of precipitation is more difficult in large compared to small basins, and this difficulty is compounded by the fact that precipitation gages in large basins are maintained by different agencies than those that manage streamflow data collection. Thus, for example, in the United States, the US Geological Survey manages streamflow data collection from large basins, but state climate services or the National Oceanic and Atmospheric Administration manage precipitation records.

In an intersite comparison study, it is generally possible to match fine-resolution temporal streamflow data with precipitation data at the small basin scale, because these records usually were collected as part of the same monitoring effort. However, at the large basin scale (above 1 km^2), it may not be possible to match historical streamflow with precipitation at temporal resolutions finer than monthly, particularly if spatial heterogeneity is important. Obtaining historical precipitation data for periods of record that match gaged periods in large basins can be a significant challenge. Nevertheless, historical climate records have been compiled at the national scale in some cases (*see* e.g. for the United States, **http://cdiac.esd.ornl.gov/epubs/ndp019/ndp019.html**). Temporal resolution of precipitation and air temperature data at the small basin scale is usually high, and may even match that of streamflow (e.g. 15-minute data). However, obtaining spatially interpolated historical – or even current – precipitation and air temperature data at temporal resolutions finer than monthly may be difficult for many large basin sites (*see* **http://www.ncgc.nrcs.usda.gov/branch/gdb/products/climate/data/index.html**).

Other data, such as vegetation cover and type, soils, geology, and land use history may be critical for interpreting intersite hydrology comparisons. However, these data are much less likely than streamflow or climate data to be available or tabulated in a format that is easily interpreted. Some good examples of vegetation, soils, and other spatial datasets linked to streamflow and precipitation can be seen at long-term ecological research (LTER) sites (**www.lternet.edu**), especially those that were originally experimental forests, in the United States. Even if they are available, vegetation, soils, and land use history data are very unlikely to have been collected in a consistent framework across multiple sites, increasing the difficulty of making use of such data in an intersite hydrology comparison. When conducting an intersite hydrology comparison, it is usually necessary to contact researchers or managers at individual sites for help in obtaining and interpreting this sort of data.

Intersite hydrology comparisons are limited by the availability of data (Figure 3). Efforts are underway to systematically collect and provide hydrologic data (see hydrodb website, **http://www.fsl.orst.edu/hydrodb/**). However, hydrologic data may still be in original analog formats, or not yet publicly available, at many sites of interest for intersite hydrology studies. Researchers will need to seek out data of interest from these individual sites, and they will also need to support and participate in efforts to make such data publicly available.

Intersite comparisons require that hydrologic data be collected using comparable measurement methods, units, and data records. Relative to other kinds of comparative studies, such as vegetation or stream chemistry, intersite hydrology and climate studies are made easier by the fact that streamflow data collection methods are largely standardized and common units (e.g. cfs, cms) are used. Both precipitation and streamflow data are typically collected at 15-minute or finer resolution, and are tabulated at daily, monthly, and annual resolution.

An intersite hydrology comparison also may require overlapping records, but certain intersite hydrology comparisons may not. Runoff ratios require overlapping precipitation and streamflow periods of record, and intersite comparisons of runoff ratios probably should be made controlling for climate by using a common period of record. Paired-basin experiments require the collection of simultaneous records at treated and control basins. Intersite comparisons looking for trends over time due to climate change also require simultaneous records. On the other hand, intersite hydrology comparisons of paired-basin experiments need not have records from the same time periods, since the questions of interest relate to the periods before and after imposition of a treatment.

The lengths of record available limit the scope of what can be determined from an intersite hydrology comparison (Figure 3). Short records (i.e. a few years) lend themselves to calculations of metrics, such as runoff ratios, which may be compared to other sites, or examined to compare

hydrologic processes at finer than annual temporal resolutions such as seasonal, storm, or daily time scales. In contrast, long records allow examination of changes over time in metrics, such as runoff ratios, that may be related to longer-term processes, including vegetation succession, disturbance, or climate change.

The numbers of sites with records of hydrology, climate, and environmental variables also limit the scope of what can be learned from intersite hydrology comparisons (Figure 3). There is no complete listing of small monitored basins, or of small paired-basin experiments. Only a couple of dozen small basins ($<1\,km^2$) exist in the United States where streamflow and precipitation data have been consistently monitored for periods of several decades and where good records exist describing vegetation cover, soils, and landscape history (see `http://www.fsl.orst.edu/hydrodb/`). Bosch and Hewlett (1982) noted a total of 26 sites in 16 states in the United States where small paired-basin experiments had been conducted, but many of these records are not available in digital form. On the other hand, hundreds of large basins (10^1 to $10^6\,km^2$) have long-term streamflow records collected by the USGS. However, these sites may not have matching precipitation records, and they do not have compiled records of vegetation cover, soils, or landscape history.

Choosing Statistics for Comparison of Rainfall and Runoff Data

Lots of ink and breath have been expended in statistical controversies in hydrology. Some argue that discussion of statistical methods distracts from the essential questions. However, statistical approaches force researchers to clearly articulate the questions and hypotheses they are testing. Many statistical techniques are now available (Table 2). As in any statistical analysis, the analysts must keep in mind that if data are not independent, or if data are not normally distributed, certain statistical findings will be biased. Thus, tests of autocorrelation, and data transformations, are an important component of intersite comparisons.

The statistics used may be measures of single variables, or they may be measures involving two or more variables. Statistics that are useful for intersite comparisons include means, variances, general linear models (regression and analysis of variance), autocorrelation methods (including spectral analysis and wavelets), and cross-correlation methods (Table 2). Single-variable measures include means, variability (e.g. standard deviation), distributions or measures of distribution shape, or measures of autocorrelation. Bivariate measures include linear regression, analysis of variance, rainfall-runoff ratios, and cross-correlation.

Single-variable Analysis – Comparison of Means

Mean daily (or monthly, or annual) streamflow may be compared across a range of sites in an intersite hydrology study (Table 2). For example, Post *et al.* (1998) compared mean annual precipitation and mean annual runoff across sites. This kind of comparison of long-term averages can reveal differences in climate, or which types of vegetation use larger or smaller proportions of water input as precipitation. Post *et al.* (1998) showed that rainfall-runoff ratios were similar across four forested sites with strongly contrasting climates.

Single-variable Analysis – Comparison of Variance or Variation

A second single-variable approach to comparison involves measures of variability (Table 2). Standard deviations, standard errors, or coefficients of variation (standard deviation divided by the mean) of streamflow can be evaluated to compare variability among sites or time periods. For example, Post and Jones (2001) conducted an intersite comparison based on the coefficients of variation of daily, monthly, and annual streamflow at small, forested basins in the United States. They replicated the common finding of decreased variability with increasing temporal aggregation, noting that streamflow variability was much higher at daily timescales than at seasonal or annual time scales, at four sites with temperate deciduous, temperate conifer, and tropical forest. More interestingly, from the point of view of intersite comparison, Post and Jones (2001) showed that the greatest variability in rainfall and runoff varied among sites for any give time scale: variability at the daily time scale was greatest at eastern United States temperate forest sites in mesic climates, variability at the seasonal time scale was greatest at western United States conifer forest sites in xeric (Mediterranean) climates, and annual variability was greatest at tropical forest sites in hurricane-affected Puerto Rico.

Single-variable Analysis – Comparison of Higher Moments of Distributions or Quantiles

A third single-variable approach involves creating and comparing measures of the shapes of distributions of precipitation, streamflow or rainfall-runoff ratios (Table 2). For example, Andreassian *et al.* (2003) created measures of the proportions of the daily streamflow distributions that fell above or below certain benchmarks (e.g. flood flows, low flows, base flows) and compared how these proportions changed over time for a number of small, forested basins. This kind of comparison can be used to detect changes over time, or cross-site differences, in low flows, high flows, or other parts of the flow frequency distribution. Andreassian *et al.* (2003) showed that flow frequency distributions had remained relatively constant at a recently reforested basin in Ohio, USA (Coshocton) and a burnt shrubland basin in southern France (Real Collobrier), but low flows and baseflow had shifted downwards over time at an old-growth basin in Oregon, USA (Andrews WS 2). Gupta and

Waymire (1998) used the moments of the distributions of rainfall and runoff to test for self-similarity across a set of basins arranged in increasing size.

Single-variable Analysis – Comparison of Autocorrelation

A fourth single-variable approach involves measures of how streamflow or rainfall are related to themselves over time (Table 2). Measures of autocorrelation can be used to characterize and compare streamflow, precipitation, or rainfall/runoff ratios among sites at various time scales, which can then be compared across sites or time periods. For example, Post and Jones (2001) conducted an intersite comparison based on the autocorrelation coefficients for a range of lags using both daily and monthly streamflow at four forested sites. They were able to detect differences among sites in the seasonal and daily timescales rainfall-runoff relationships that were linked to differences in climate and vegetation. Tague and Grant (2004) applied autocorrelation analysis to streamflow records from two large basins in Oregon, showing how the seasonality of flows differed between two basins with contrasting geology.

Single-variable Analysis – Comparison of Spectra or Wavelets

More complex single-variable approaches to intersite comparison of rainfall and runoff data involve the use of spectral analysis or wavelets (Table 2). Kirchner *et al.* (2000) constructed log–log plots of the spectral power of rainfall, streamflow, and stream chloride concentrations in catchments in Plynlimon (Wales), and showed that the plot of spectral power of chloride concentration was fractal (had a positive slope). Lafrenière and Sharp (2003) used wavelet analysis of streamflow from glacially controlled lakes to infer and separate the influences of snowmelt from rainfall on discharge.

Bivariate Analysis – Comparison Using Linear Regression

Linear regression has been widely used to compare two sites over two or more time periods (Table 2). The slope of a regression line comparing e.g. runoff at site 1 versus site 2 (y/x) is a ratio that can be compared across sites or time periods. The earliest studies of paired-basin experiments used a regression of streamflow at the treated (y-axis) versus the control (x-axis) to characterize the treatment effect. Beschta *et al.* (2000) used linear regression to reexamine data from the small basins in Jones and Grant (1996). Jones and Grant (1996) used regression analysis to examine the ratio of change in peak discharge versus change in cumulative area harvested in large (60 to 600 km^2) basins, as a way to detect whether forest harvest was associated with changes in peak discharges. Thomas and Megahan (1998) used analysis of covariance to reexamine data from large basins in Jones and Grant (1996).

Bivariate Analysis – Comparison Using Analysis of Variance

Analysis of variance has been less widely used than linear regression, but it also can be used to compare two sites over two or more time periods (Table 2). The ratio of e.g. runoff at site 1 versus site 2 can be tested for significant changes by time period using ANOVA. A more useful ratio is the log-transformed ratio, i.e. $\ln(y/x)$, which is equivalent to the difference of the log-transformed values, i.e. $\ln(y) - \ln(x)$. Eberhardt and Thomas (1991) recommend that the log-transformed ratio be used in analysis of streamflow from paired-basin experiments. For example, Jones and Grant (1996) and Jones (2000) used the log-transformed ratio of streamflows at the treated versus the control basins as a dependent variable in an analysis of paired-basin treatment effects on peak discharges. Using this ratio, they were able to show how peak discharges changed in the treated versus the control in small paired-basin experiments over 5-year periods after forest harvest (Jones and Grant, 1996; Jones, 2000). Jones and Post (2004) used this ratio to calculate how daily flows changed in the treated versus the control basins in small paired-basin experiments over 5-year periods after forest harvest.

Bivariate Analysis – Comparison Using Rainfall-runoff Ratios

Rainfall-runoff ratios are a second form of ratio that has been widely used in intersite comparisons, but only at certain time scales (Table 2). This kind of comparison allows examination of the fraction of water released from the basin relative to precipitation inputs, over time or among sites. For example, Post and Jones (2001) used mean annual runoff ratios (streamflow divided by precipitation) as a basis for comparison across sites. Annual runoff ratios also can be used as dependent variables in more complex analyses involving ratios. For example, Post and Jones (2001) conducted an intersite comparison based on the slopes of annual rainfall-runoff regressions across sites for a group of small, forested basin sites. Using annual runoff ratios in long (up to 50 year) records, they were able to show that forested basins respond to interannual variation in precipitation by taking up more or less water, and water use was similarly responsive to interannual variations in precipitation for all four forest types (Post and Jones, 2001).

As these examples illustrate that ratios of streamflow in treated versus control basins are the most commonly used measures for assessing treatment effects in paired-basin experiments. However, ratios also have the potential to be used for comparisons among basins that have not been subjected to formal treatments, but whose behavior may be diverging as the result of inadvertent treatments (e.g. Table 1). Inadvertent treatments in such "out of control" basins may include forest mortality from disease or disturbances like windthrow or fire, climate change, or

changes in species composition and leaf area associated with forest growth and succession. The potential for learning about rainfall-runoff relationships by applying standard treated/control comparison methods to pairs of control basins (which may actually be "out of control") deserves further exploration.

Runoff ratios have potential uses that have not been explored. For example, an intersite comparison could examine the change over time in the runoff ratio (e.g. Kokkonen *et al.*, 2004), and compare this measure across basins. Or, runoff ratios could be calculated for finer than annual time periods, such as seasons or even days. This latter approach resembles the calculations involved in water balance estimates.

Bivariate Analysis – Comparison Using Cross-correlation

The cross-correlation is a form of the rainfall-runoff ratio that takes into account the temporal lag between precipitation input and streamflow output in a basin, or the lag in rainfall or runoff between one basin and another (Table 2). Cross-correlations of rainfall versus runoff at a basin estimate the lags involved between input and output in a given basin, while cross-correlations of runoff (or precipitation) across sites can be used to determine the delays in transmission of a pulse of rainfall or runoff through a stream network. For example, Post and Jones (2001) used cross-correlation of daily and monthly precipitation versus streamflow as a measure for comparing across small, forested basins. Streamflow responses to precipitation lasted for only a day or two at deciduous forest sites with mesic climates and shallow soils, but they persisted for almost two weeks at conifer sites with a Mediterranean climate and deep soils. Bond *et al.* (2002) correlated hourly sapflow with streamflow at various lags to compare across two adjacent small basins with old-growth and young conifer forest (Andrews, Oregon, US). They were able to show that diurnal variations in streamflow were most pronounced and occurred within a few hours of maximum sapflow in the early summer, but, by late summer, the relationship had diminished and the lag between maximum sapflow and streamflow had lengthened to eight hours. Moore (2003) correlated hourly precipitation, soil moisture, sapflow, and streamflow at various lags to extend this comparison of two, small, adjacent forested basins. She showed that diurnal streamflow variations were related to soil moisture, sapflow, and vapor pressure. However, strong lagged relationships were apparent during only one of several periods in three successive summers. Tague and Grant (2004) used cross-correlation of streamflow in two, large, adjacent basins to demonstrate how contrasting geology produced consistent lags in runoff at the basin whose geology permitted longer storage and slower transmission of water.

SUMMARY

Despite the limitations of rainfall-runoff data, there are compelling reasons for hydrologists to conduct many more intersite comparisons of rainfall-runoff data. Studies of rainfall-runoff processes in hydrology have generated many important predictions. Our inferences about hydrologic processes are drawn from an unnecessarily narrow subset of temporal scales, spatial scales, and geographic conditions, given the range of data available. Too many publications simply repeat (or challenge) the hypotheses inferred from a few analyses of primary data, without exploring unexamined datasets. However, verification of these hypotheses depends on replicating them using rainfall and streamflow records from locations and time periods not included in the original studies. Many thousands of basin-years of data from a wide variety of ecosystems, climate types, and basin sizes have been collected, and much of these data are now available on the Internet. However, only a fraction of these data have been analyzed. Existing published analyses of these records do not address all of the questions of current interest to hydrologists and researchers from related disciplines, including climate science, geomorphology, and ecology. New questions, novel combinations of basins, new statistical approaches, and ancillary data can all be used to more carefully test important hypotheses regarding hydrologic mechanisms operating inside basins. The endeavor of intersite comparison is analogous and complementary to the parallel quest for methods of parameter identification and model structure selection in hydrologic modeling.

Acknowledgment

This work has been supported by grants to the H.J. Andrews Long-term Ecological Research program, and for Intersite Hydrology comparisons, from the Long-term Studies program of the National Science Foundation, and by a Bullard Fellowship at Harvard Forest. Younes Alila provided helpful comments on a draft of this manuscript.

REFERENCES

Andreassian V. (2004) Waters and forests: from historical controversy to scientific debate. *Journal of Hydrology*, **291**, 1–27.

Andreassian V., Parent E. and Michel C. (2003) A distribution-free test to detect gradual changes in watershed behavior. *Water Resources Research*, **39**(9), 1252–1262.

Beschta R.L., Pyles M.R., Skaugset A.E. and Surfleet C.G. (2000) Peakflow responses to forest practices in the western cascades of Oregon, USA. *Journal of Hydrology* **233**, 102–120.

Beven K.J. (2000) Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, **4**(2), 203–213.

Binkley D. and Brown T.C. (1993) Forest practices as nonpoint sources of pollution in North America. *Water Resources Bulletin*, **29**(5), 729–741.

Binkley D., Ice G.G., Kaye J. and Williams C.A. (2004) Nitrogen and phosphorus concentrations in forest streams of the United States. *Journal of the American Water Resources Association*, **40**(5), 1277–1291.

Bond B.J., Jones J.A., Moore G., Phillips N., Post D. and McDonnell J.J. (2002) The zone of vegetation influence on baseflow revealed by diel patterns of streamflow and vegetation water use in a headwater basin. *Hydrological Processes*, **16**, 1671–1677.

Bosch J.M. and Hewlett J.D. (1982) A review of catchment experiments to determine the effect of vegetation changes on water yield and transpiration. *Journal of Hydrology*, **55**, 3–23.

Daly C., Gibson W.P., Taylor G.H., Johnson G.L. and Pasteris P. (2002) A knowledge-based approach to the statistical mapping of climate. *Climate Research*, **22**, 99–113.

Eberhardt L.L. and Thomas J.T. (1991) Designing environmental field studies. *Ecological Monographs*, **61**(1), 53–73.

Gupta V.K. and Waymire E. (1989) Statistical self-similarity in river networks parameterized by elevation. *Water Resources Research*, **25**(3), 463–476, doi:10.1029/88WR04115.

Gupta V.K. and Waymire E. (1993) A statistical analysis of mesoscale rainfall as a random cascade. *Journal of Applied Meteorology*, **32**, 251–267.

Gupta V.K. and Waymire E. (1998) Scale invariance and regionalization in floods. In *Scale Dependence and Scale Invariance in Hydrology*, Sposito G. (Ed.), Cambridge University Press: Cambridge, pp. 88–135.

Gurevich J. and Hedges L.V. (2001) Meta-analysis: combining the results of independent experiments. In *Design and Analysis of Ecological Experiments*, Scheiner S.M. and Gurevich J. (Eds.), Oxford University Press: New York, pp. 347–369.

Harr R.D. (1983) Potential for augmenting water yield through forest practices in western Washington and western Oregon. *Water Resources Bulletin*, **19**(3), 383–393.

Hedges L.V. and Olkin I. (1985) *Statistical Methods for Meta-Analysis*, Academic Press, New York.

Hibbert A.R. (1967) Forest treatment effects on water yield. In *International Symposium on Forest Hydrology*, Sopper W.E. and Lull H.W. (Eds.), Pergamon, Oxford.

Hornbeck J.W., Bailey S.W., Buso D.C. and Shanley J.B. (1997) Streamwater chemistry and nutrient budgets for forested watersheds in New England: variability and management implications. *Forest Ecology and Management*, **93**, 73–89.

Jones J.A. (2000) Hydrologic processes and peak discharge response to forest removal, regrowth, and roads in ten small experimental basins, western Cascades, Oregon. J.A. Jones. *Water Resources Research*, **36**(9), 2621–2642.

Jones J.A. and Grant G.E. (1996) Peak flow response to clearcutting and roads in small and large basins, western Cascades, Oregon. *Water Resources Research*, **32**, 959–974.

Jones J.A. and Post D.A. (2004) Seasonal and successional streamflow response to forest cutting and regrowth in the northwest and eastern United States. *Water Resources Research*, **40**, W05203, doi:10.1029/2003WR002952.

Kirchner J.W., Feng X. and Neal C. (2000) Fractal stream chemistry and its implications for contaminant transport in catchments. *Nature*, **403**, 524–527.

Kokkonen T., Koivusalo H., Karvonen T., Croke B. and Jakeman A. (2004) Exploring streamflow response to effective rainfall across event magnitude scale. *Hydrological Processes*, **18**(8), 1467–1486.

Lafrenière M. and Sharp M. (2003) Wavelet analysis of inter-annual variability in the runoff regimes of glacial and nival stream catchments, Bow Lake, Alberta. *Hydrological Processes*, **17**, 1093–1118.

Martin C.W., Noel D.S. and Federer C.A. (1984) Effects of forest clearcutting in New England on stream chemistry. *Journal of Environmental Quality* **13**(2), 204–210.

Moore G.W. (2003) *Drivers of Variability in Transpiration and Implications for Stream Flow in Forests of Western Oregon*, PhD. dissertation, Forest Science, Oregon State University.

Pierson F.B., Slaughter C.W. and Cram Z.K. (2001) Long-term stream discharge and suspended-sediment database, Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resources Research*, **37**(11), 2857–2861, doi:10.1029/2001WR000420.

Post D.A., Grant G.E. and Jones J.A. (1998) Ecological hydrology: expanding opportunities in hydrologic sciences. *EOS*, **79**(43), 517, 526.

Post D.A. and Jones J.A. (2001) Hydrologic regimes at four long-term ecological research sites in New Hampshire, North Carolina, Oregon, and Puerto Rico. *Advances in Water Resources*, **24**(9–10), 1195–1210.

Robinson M., Cognard-Plancq A.-L., Cosandey C., David J., Durand P., Fuhrer H.-W., Hall R., Hendriques M.O., Marc V., McCarthy R., *et al.* (2003) Studies of the impacts of forests on peak flows and baseflows: a European perspective. *Forest Ecology and Management*, **186**, 85–97.

Slaughter C.W., Marks D., Flerchinger G.N., Van Vactor S.S. and Burgess M. (2001) Thirty-five years of research data collection at the Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resources Research*, **37**(11), 2819–2823.

Smith J.W. (2002) *Mapping the Thermal Climate of the H.J. Andrews Experimental Forest, Oregon*, MS thesis, Department of Geosciences, Oregon State University.

Stednick J.D. (1996) Monitoring the effects of timber harvest on annual water yield. *Journal of Hydrology*, **176**(1–4), 79–95.

Tague C. and Grant G.E. (2004) A geological framework for interpreting the low-flow regimes of cascade streams, Willamette River Basin, Oregon. *Water Resources Research*, **40**, W04303, doi:10.1029/2003WR002629.

Thomas R.B. and Megahan W.F. (1998) Peak flow responses to clear-cutting and roads in small and large basins, western Cascades, Oregon: A second opinion. *Water Resources Research* **34**(12), 3393–3403.

PART 11

Rainfall-runoff Modeling

122: Rainfall-runoff Modeling: Introduction

KEITH J BEVEN

Department of Environmental Science, and Lancaster Environment Center, Lancaster University, Lancaster, UK

This section provides an introduction to the theme of rainfall-runoff modeling in hydrology. It provides a summary of the purposes of rainfall-runoff modeling; a classification of rainfall-runoff models; a brief account of process descriptions in rainfall-runoff models; and a short history of rainfall-runoff modeling. It also discusses the problem of model choice from the wide range of possibilities available, and the important question of model calibration using different types of observational data before considering the prediction of the effects of future change (when no observations are possible). Finally, a guide is given to the future of rainfall-runoff modeling. Reference to the other sections in this article are given throughout, as well as to relevant available texts.

INTRODUCTION

The Many Purposes of Rainfall-runoff Modeling

No introduction to the range of rainfall-runoff modeling techniques can be presented without considering the purposes for which the modeling is carried out. All modeling involves some form of extrapolation in either time or space, but different modeling techniques are appropriate for different hydrological problems. At the end of this article, there are reviews of modeling for flood forecasting (both in real time and for predicting flood frequencies), for flood inundation modeling, for integrated basin management, and for the prediction of the effects of change. There are also sections that present different types of modeling methodologies in more technical detail.

It has to be recognized straight away that, despite the very real demand for good quantitative rainfall-runoff predictions for these very practical problems, the degree to which current models can satisfy that demand is limited. Much of hydrology measurement technique is constrained, since so much of what is interesting in the hydrological response of a catchment to rainfall takes place beneath the ground surface. The development and application of models is, therefore, also limited by the available measurement techniques. The result is that it remains very difficult to predict the hydrological response of any arbitrary catchment area based on mapped information about the local topography, soils, and land use alone. This "prediction of ungauged catchments" problem is still essentially unsolved (*see* **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**). Thus, there is still a considerable body of modeling work that depends on local *calibration* of models using measured rainfall inputs, and comparing observed and predicted discharges with a view to improving the predictions. In the past it had been considered sufficient to find some "optimal" model to make predictions. Today, the uncertainties inherent in the measured inputs to the model, model definition, and model calibration are more widely recognized, and techniques of assessing uncertainties in the predictions are being developed (*see* **Chapter 130, Fuzzy Sets in Rainfall/Runoff Modeling, Volume 3** and **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**).

Part of this uncertainty arises from the model definition itself. There are now very many different models in hydrology, representing variants on several different lines of development or families of models of different degrees of complexity (see Section "A classification of rainfall-runoff models"). They all remain, however, simplifications of what we perceive as the real complexity of hydrological processes. One of the major questions that arises in rainfall-runoff modeling is how best to approximate the perceived complexity of the real catchments. Ironically, perhaps, making models more complex to take account of some of the perceived complexities does not necessarily lead to

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

more precise and less uncertain predictions. This is because adding complexity in general, introduces more parameter values that are not easily measured directly or calibrated on the basis of indirect information. More complex models, therefore, in general have more degrees of freedom in representing the catchment, and this might result in greater uncertainty in prediction. Thus, there is a problem of trying to achieve the right balance between model complexity and prediction uncertainty in situations that might have different types of data available for model calibration or evaluation.

It must be stressed that this is an ongoing research area in hydrology. This is reflected in the sections on recent modeling approaches, including fuzzy modeling techniques, "top–down" approaches to modeling and the assessment of uncertainties in model calibration and model predictions. These new approaches have been developed out of a recognition of some of the difficulties and limitations of modeling all the hydrological processes in any specific catchment area; difficulties that become important when trying to predict hydrological responses under more extreme (wet or dry) conditions than measured before, or trying to predict the hydrological impacts of future change in land management or climate inputs to inform policy and management decisions. A view on the future of rainfall-runoff modeling is expressed below in "Predicting the effects of change" and a guide to further reading in "The future of rainfall-runoff modeling".

A Classification of Rainfall-runoff Models

There are a variety of different ways of classifying rainfall-runoff models (see e.g. Clarke, 1973; Wheater *et al.*, 1993), but the most fundamental distinction that is usually made is between *lumped* models and *distributed* models. Lumped models deal with a catchment as a single unit, attempting to relate precipitation inputs to discharge outputs without any consideration of the spatial patterns of the processes and characteristics within the catchment. Distributed models attempt to take account of the spatial patterns of hydrological response within a catchment area.

Rainfall-runoff modeling started with lumped models back in the nineteenth century with the "*rational method*" that related discharge (usually peak discharge) directly to a measure of rainfall inputs, catchment area, and a runoff coefficient (see Beven, 2001b). The runoff coefficient served the purpose of what would now be called a model *parameter* that could be varied to reflect local conditions in particular applications. The difficulty in applying the rational method was therefore a difficulty of deciding on a value for the runoff coefficient. This was a particular difficulty in this prototype model because this parameter would vary not only from one catchment to another but also with the magnitude of an event and the state of the catchment prior to an event (the *antecedent conditions*).

In the twentieth century, therefore, more complex lumped models were developed to reflect the perceptions of different hydrologists about the processes of hydrological responses to rainfall. One of the most important developments was that of the *unitgraph* (now called the *unit hydrograph*) that attempted to predict the time distribution of discharge as well as the peak. The unitgraph represented the time distribution of one unit of *effective* or *excess* rainfall. The effective rainfall was that part of the total rainfall inputs that contributed to the storm hydrograph for an event (though it is known that in many catchments much of the hydrograph is made up of water that was stored in the catchment prior to an event, and which is displaced out of storage during the event). There was still a runoff coefficient problem in predicting how much of the rainfall would become effective rainfall for a given magnitude event and antecedent conditions, on an event-by-event basis. The unit hydrograph is a *linear* approximation, which means that once a unit hydrograph has been determined for a catchment area, different effective rainfalls will be predicted as having the same time distribution, but with a simple linear scaling of the magnitudes. The linear approximation then produces the result that two units of *effective* rainfall (rather than the total input) will produce twice as much discharge.

The linear assumption of the unit hydrograph has been much criticized in the past, but has served rainfall-runoff modelers quite well. Similar principles are still used in *transfer function* models today (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**). The greater problem is that of estimating how much of the total input to a catchment should form the effective rainfall. The effect of antecedent conditions on the effective rainfall is very much a nonlinear problem. Twice as much rainfall with the same antecedent conditions will generally produce more than twice the discharge, while the same rainfall under dry antecedent conditions will produce much less runoff than under wet antecedent conditions. This nonlinearity problem was partly solved by the introduction of continuous simulation models, particularly as digital computers started to become more widely available in the 1960s. One of the earliest lumped continuous simulation models was the Stanford Watershed Model, developed by Norman Crawford and Ray Linsley at Stanford University in 1962 (for a fuller history see Section "A short history of rainfall-runoff modeling" and Beven, 2001b). This structure still survives, with many added components, in the HSPF (Hydrologic Simulation Package Fortran) package now distributed by the US Environmental Protection Agency. It is an example of a lumped explicit soil moisture accounting (ESMA) model in that it has several storage elements to represent different hydrological processes and storages (interception, upper soil moisture, lower soil moisture, groundwater etc.). The fluxes between these stores are

controlled by a variety of different parameters. Application of the model, therefore, requires that these parameters be estimated or calibrated for each application area. The temporal changes in the antecedent conditions are taken into account by continuously calculating the changes in storage in the model, but at the expense of introducing a considerable number of parameter values (more than 30 in some versions of the Stanford Watershed Model).

This type of model can be made more distributed by applying it to different subcatchments and then routing the subcatchment outputs to the point of interest using a river routing component. This allows the spatial patterns of inputs to be taken into account, at least at the subcatchment level. This might be important where there are strong spatial patterns in inputs resulting from strong changes in rainfall with elevation, or the seasonal patterns of snowmelt as a function of elevation, aspect, snow accumulation, and so on. The only problem with such a strategy is that each subcatchment then requires its own set of parameter values to be estimated or calibrated.

A similar problem applies to fully distributed models which attempt to make predictions of the response of all the elements in a spatial discretization of a catchment area. The elements might be based on a square grid discretization (an approach that has become common with the widespread availability of raster digital elevation data) or more irregular elements or *control volumes*. Each element can have its own inputs and parameter values. The more complex models also involve multiple layers in the vertical to predict fluxes in three dimensions. Thus, these models can have literally hundreds or thousands of parameter values that must be defined to run the model. Even so, these models cannot reflect all the perceived complexity of the hydrological processes. In particular, they still require some parameterization of the variability in soil water, runoff, and evapotranspiration fluxes at the sub-element scale. Beven (1989) has therefore, suggested that even the most distributed hydrological models are still effectively lumped representations at the element scale. The *representative elementary watershed* approach to distributed modeling, described in (**Chapter 13, Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, Volume 1** and **Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**), is an attempt to deal directly with this issue in a physically rigorous way.

A final classification of rainfall-runoff models is into models that are *deterministic* and models that are *stochastic* in their predictions. A deterministic model will, with a given input sequence, produce a single prediction of all its output variables. The majority of model applications in hydrology are still run deterministically, even though the uncertainties associated with such predictions are now widely appreciated. A stochastic model allows for uncertainty in the inputs or parameterizations or outputs, such that a given input sequence will produce an uncertain prediction of the output variables. Unfortunately it can be very difficult to separate out the effects of the various sources of uncertainty that contribute to the prediction process, particularly for this type of nonlinear system. There is, therefore, no agreement about what approach should be used in assessing predictive uncertainty. Those methods that have a good theoretical base often have restrictive assumptions that are difficult to justify in real applications. Other methods are not so restrictive in their assumptions, but do not have such a strong basis in theory. These issues are discussed further in (*see* **Chapter 130, Fuzzy Sets in Rainfall/Runoff Modeling, Volume 3** and **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**).

Process Descriptions in Rainfall-runoff Models

It is important to recognize the difference between the perceived complexity of water flow pathways in a catchment and the relatively simplistic models that are used to predict water fluxes and stream discharges. The hydrologist can appreciate the complexity of the real processes in a qualitative way (the *perceptual* model); it is much more difficult to represent that complexity in a quantitative mathematical description (the *formal* or *conceptual* model). Thus, any rainfall-runoff model must be necessarily only an approximate representation of the real processes. This is perhaps easiest to appreciate for lumped catchment models, though these can still make useful predictions if the only variable of real interest is the discharge hydrograph. The same consideration also applies, however, to even the most distributed rainfall-runoff model, since they can only approximate the responses at the sub-control volume scale. Thus, all process descriptions in rainfall-runoff models rely on some parameterization of the sub-control volume complexities and heterogeneities.

In all models, it is necessary to maintain mass continuity. Water, as represented by the model, should not be gained or lost. Theoretically, this should also be the case for continuity of energy and momentum as well, so that hydrological models could claim to preserve the second law of thermodynamics. Unfortunately, this is much more difficult to ensure, with energy and momentum losses being accounted for implicitly by model parameters such as resistance coefficients (but *see* **Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**). What is often not realized, however, is that even for mass continuity, the water balance kept by the model may not be the same as that kept by a real catchment or control volume. This is because there are components of the real mass balance that cannot easily be measured accurately, especially as integral fluxes over an area. This is particularly true of precipitation inputs over a catchment when rain falls from locally intensive convective cells or

snow accumulation is affected by wind redistribution. It is also true of evapotranspiration fluxes from heterogeneous vegetation covers on hillslopes. This is not to suggest that models should not maintain mass continuity (although some models include multipliers to adjust the input rainfalls or potential evapotranspiration rates as parameters), only that these physical laws must be viewed as theoretical capacities of the system that might be difficult to verify by measurement (see discussion in Beven, 2002a).

In many conceptual models, the representation of processes at the control volume scale is expressed in terms of some function relating flux to storage in the control volume. Most lumped catchment models are of this type, including the ESMA models. Many routing models, such as the unit hydrograph, can also be interpreted as a sequence of storage elements in series or in parallel. These storage elements can have storage-flux relationships that are either linear or nonlinear. With one or two exceptions, these relationships are not based on secure theoretical reasoning, but on subjective choices that seem to give the right type of behavior. A linear store has often been used, for example, to represent the recession curve from a catchment area. The equation of a linear storage gives a recession curve for which discharge declines exponentially with time. There is no physical reason to suggest that the recession of the integral drainage fluxes of all sources of subsurface flow in the catchment or control volume should have the form of a linear store (except under some the very limiting assumptions: *see* **Chapter 126, Modeling Recession Curves and Low Streamflows, Volume 3**) but the form appears to work for many cases.

But not for all cases: other nonlinear storage elements seem to be more appropriate for some catchments. Some models have used first order or second order hyperbolic functions in time (see e.g. Tallaksen, 1995; Lamb and Beven, 1997, for a recent discussion of the different forms that result from different assumptions about the subsurface flow domain). The important point to take from these different parameterizations is that we do not have the investigative measurement techniques necessary to be secure about what form these relationships should take, given that we know that the sub-control volume processes are subject to considerable complexity, except by seeing which functions might be appropriate in reproducing the discharges at the catchment outlet (where we can take a measurement). It then follows that different types of function might be appropriate for different catchments or control volumes within a catchment. This does not only apply to subsurface drainage but also to any other hydrological flux within a control volume. Any subjective choice of function to represent the fluxes at the control volume scale can usually only be calibrated and corroborated by comparison with measured discharges in this way.

The subjective nature of model design was recognized early in the history of computer-based rainfall-runoff modeling. It did not seem to be really scientific. There was then a conscious attempt to make model definition more objective by representing the processes by the physical "laws" used to describe those processes. The seminal paper describing such an approach was the "blueprint for a physically based digitally simulated model" of Freeze and Harlan (1969). Very briefly, they showed how a model could be developed using the Richards' equation (based on Darcy's law) for partially saturated subsurface fluxes and the St. Venant equations for depth-integrated surface fluxes. These equations could be coupled though appropriate boundary conditions. The equations are both nonlinear partial differential with boundary conditions that vary in space and time as hydraulic potentials and therefore infiltration rates, seepage faces, and other features of the system change dynamically. There are no general analytical solutions to these equations; therefore, the model must be based on approximate numerical solution algorithms (originally finite difference methods but now more usually finite element or finite volume techniques). This requires that the system be discretized into a large number of control volumes to represent the dynamic distributed responses and the way in which the fluxes will vary in response to patterns of hydraulic potentials on the hillslopes and in the stream network.

Following the Freeze and Harlan blueprint, the model description is objectively based on physical principles. It does not, however, entirely eliminate the subjective element in model definition. This is because the characteristics of all the control volumes representing the catchment must be defined in terms of appropriate parameter values before the model can be run. In principle, these parameters have physical meaning (they include, for example, hydraulic conductivity and porosity of the soil, and the roughness coefficient for channel flow). But we have no techniques that will allow the values of those parameters to be measured at the control volume scale. Small-scale measures of hydraulic conductivity, for example, reveal that there may be small-scale variability of orders of magnitude within a control volume, especially for soils containing macropore structures. This then implies that there will be sub-control volume variability in the hydraulic gradients and velocities of flow in the soil. Because of the highly nonlinear relationships between unsaturated hydraulic conductivity and flow rate, this then implies that the small-scale physics (as represented here by Darcy's law so that flux per unit area is assumed to be directly proportional hydraulic gradient and local unsaturated hydraulic conductivity) should not be used at the control volume scale. In a heterogeneous domain, the same form of relationship cannot represent the nonlinear variability of local velocities at the control volume scale except in the special

(and unrealistic) case of a soil with spatially homogeneous characteristics.

Darcy's law is still being used, however, in nearly all rainfall-runoff models based on the Freeze and Harlan blueprint, because without detailed knowledge of the small-scale heterogeneity in soil structure and characteristics, it is difficult to develop an adequate alternative description. The continued use of Darcy's law to represent control volume fluxes, therefore, depends on an implicit (and subjective) assumption that it is an adequate approximation to the integral fluxes at the control volume scale. Physically, this is much easier to justify for flows in the saturated zone than in the unsaturated zone. Similar considerations apply to the representation of surface flows because of the way in which velocities and depths can vary rapidly across a hillslope (especially if covered with short vegetation) or in the variable cross-sections of the channel network. In all these cases, effective values of the model parameters will be required at the control volume scale, but these effective values cannot easily be measured and their physical basis has been undermined by the effects of the sub-control volume nonlinearities. These laws are now just one possible choice for a sub-control volume parameterization, awaiting the development of something more realistic. For a modern interpretation of the Freeze and Harlan blueprint, however, (*see* **Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**).

In addition, the Freeze and Harlan blueprint is not complete (Beven, 2002b). There are processes that were not included in the original description, including preferential flows in the soil and on the surface. Other processes might also be important, such as surface crusting, the concentration of rainfalls reaching the soil surface as a result of the structure of the vegetation canopy, and the separation of different parts of the saturated flow domain due to fracturing of the bedrock or small-scale undulations in the bedrock surface. There are many field studies that attest to the perception of such processes as important in different catchments, but the development of generally acceptable descriptions of such processes has proven to be very difficult. It is, for example, extremely difficult to obtain adequate descriptions of the bedrock surface or soil structure even in research catchments. Some models have added preferential flow components or depth variations in surface flow within their sub-control volume parameterizations, but the effective parameter values will be difficult to estimate.

The difficulty of applying this type of model, even with intensive field measurements to try and determine model parameter values, is attested by the continuing story of modeling the R5 catchment at Chickasha, Ohio, by Keith League and his coworkers. Their original model was based on the perception that the stream hydrograph was the result of a Hortonian infiltration excess runoff process. They made several attempts to predict the stream hydrographs using a model based on this concept and with more and more detailed information about the spatial patterns of local infiltration characteristics in the catchment. The measurement scale was, however, still smaller than the model control volume scale. None of these attempts was considered entirely successful (see League and Kyriakidis, 1997) and the story is now continuing with a new approach based on coupled surface and subsurface modeling (van der Kwaak and League, 2001). Similar problems have been encountered in trying to represent distributed field measurements of soil water contents or water table levels in other studies (e.g. Lamb *et al.*, 1998; Anderton *et al.*, 2002). It should be noted, however, that such comparisons are also subject to scale problems since local water content measurements and even local water table measurements in shallow soils might not be directly equivalent to the variable of the same name predicted by the model at the control volume scale.

There is one additional important consideration in the formulation of process representations of rainfall-runoff models. In many field natural and artificial tracer experiments, based, for example on the variations in concentrations of the environmental isotopes of oxygen and hydrogen in the water molecule, it has been shown that much of the storm hydrograph is made up of water that was stored in the catchment prior to an event, and not the rainfall that fell in that event. The stored water is being displaced into the stream by the incoming water. A model that is predicting only discharge does not necessarily need to take this into account, but it might be an indication that the process descriptions being used in a model are not appropriate (for example, if a model is based on an infiltration excess runoff description but the tracer data indicate a high proportion of pre-event water in the hydrograph). It does become important if predictions of water quality are required, since the pre-event water and the event water might have quite different quality characteristics (though it has to be remembered that the event water might interact geochemically with the vegetation and soil before reaching the stream even if it does contribute to the storm hydrograph). In research studies, it has proven very difficult to achieve adequate simulations of both runoff and some quality characteristics, even for conservative tracers such as the oxygen and hydrogen isotopes. This is, at least in part, because the effective storage volumes for the prediction of fluxes might be quite different to the effective storage volumes for the prediction of concentrations, and neither might relate easily to the storages that can actually be measured (locally) in the soil profile. Thus, additional model parameterizations and additional effective parameter values, will be required for the prediction of concentrations in addition to fluxes.

A Short History of Rainfall-runoff Modeling

The nineteenth century origins of rainfall-runoff model in the rational method were noted in the Section "A classification of rainfall-runoff models". The rational method, in its very simple equation,

$$Q_p = CAP \tag{1}$$

where Q_p is peak discharge, P is the volume of input precipitation over the catchment area A in a defined time period, and C is a runoff coefficient parameter, already has all the elements of a rainfall-runoff model and, indeed, the method is still in use today. The problem in its application, and the reason why we need more complex methods, lies in the parameter C. C cannot be treated as a constant for a catchment area, but will vary from event to event. It must incorporate the effects of the space and time variability in antecedent conditions, precipitation inputs and surface and subsurface runoff generation and the effects of routing of that runoff to the catchment outlet or where predictions are required.

There was obvious scope, therefore, in later developments of rainfall-runoff models to separate out the processes of runoff generation from runoff routing. This was first done in the 1920s by C. N. Ross in Australia when he used the idea of zones of different travel time to the outlet in a catchment to allow for different runoff coefficients in different zones. The runoff generation in each zone could then be delayed by the appropriate travel time and summed to produce a total discharge. A similar approach was used by Zoch and Clark in the United States, and by Richards (1944) in the United Kingdom in one of the very first books on rainfall-runoff modeling. The approach is dependent on the assumption that the routing times do not change for different events. It is, therefore, equivalent to a linear *transfer function* for the predicted runoff generation. Sherman (1932) took this idea and proposed that the transfer function could be represented as a time distribution for routing runoff from a catchment area without any direct link to the areas involved. He called the transfer function the *unitgraph*, now more commonly referred to as the *unit hydrograph*. The unit hydrograph approach to runoff routing is still widely used, and can be treated within a general linear transfer function methodology (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**).

There remained the (nonlinear) question of how much of the input rainfalls would be available as effective rainfall to create the output hydrograph from a catchment. At the same time as Sherman proposed the unitgraph, Robert Horton (1933) came up with a model of runoff generation that has since been very widely applied. He proposed that the storm hydrograph, in excess of a baseflow component, was predominantly made up of surface overland flow in excess of the local infiltration capacity of the soil. He provided a function to describe the infiltration capacity of the soil, with two parameters to be identified for each soil type, though he viewed this function as very much a representation of the surface controls on infiltration, rather than a control resulting from flow through the bulk soil as it is represented using the Richards' equation in most process based models today. The combination of Horton's and Sherman's ideas then provided one of the first models to predict the full storm hydrograph and has been very widely used. It has also been widely modified with different types of runoff generation function, such as the US Soil Conservation Service method that has its origins in the analysis of runoff data from plots and small catchment areas and does not explicitly assume an infiltration excess mechanism (see the summary of this and other methods in Beven, 2001b).

We now know, of course, that runoff generation is more complicated than that (see the Section below on "Process descriptions in rainfall-runoff models"), that surface runoff can be generated by saturation excess as well as infiltration excess mechanisms and that in many environments much of the storm hydrograph is made up of water that was stored in the catchment prior to an event and which is then displaced out of subsurface storage by the input rainfalls. Thus, particularly as digital computers allowed more complex calculations to be made, there were new models developed that tried to take more explicit account of the various storages in a catchment and their responses to rainfalls. These ESMA models, referred to in the Section "A classification of rainfall-runoff models" have been very widely used and many different variants can be found. In the 1960s and 1970s there were probably nearly as many ESMA models as there were hydrological modelers. All of them had parameters of the functions controlling the fluxes between the storage elements that had to be calibrated for a particular catchment, although they varied in the number of parameters to be calibrated. They all worked, more or less, because given a historical period of rainfall and discharge data the parameters could be varied until a good or atleast acceptable fit was obtained between observed and predicted discharges.

This calibration was either carried out manually, with a visual evaluation of how well the model was fitting the data (this was how the Stanford Watershed Model was calibrated) or by an automatic optimization method, using a quantitative measure of model performance, such as the sum of the squared errors between observed and predicted discharges during the calibration period. The manual method had the advantage of using "hydrological reasoning" and the experience of the modeler to improve the model performance (though this is difficult in practice once more than 4 or 5 parameters are being varied); the automatic optimization method has the advantage of objectivity (at least once the quantitative criterion of performance has been chosen). But, using either technique it was often difficult

to decide whether one model or parameter set was really better than another.

At the end of the 1960s, an alternative approach was proposed with a view to avoiding some of these problems by making the representation of the processes as "physically based" as possible as in the Freeze and Harlan (1969) blueprint for a physically based distributed description of surface and subsurface flow processes in time and space. Although, with even the best computers available in the early 1970s, only very approximate (and indeed inaccurate) solutions of these equations could be achieved there were many perceived advantages of this family of models and they have continued to be developed and applied to the present day. The best known current example of such a model is the Système Hydrologique Européen (SHE) which is now available in a number of different versions. The advantages of such models included the potential to apply the models using only measured (rather than calibrated) parameters, the ability to take account of spatial variability of inputs and catchment characteristics in their correct spatial context; the ability to make spatially distributed predictions of water fluxes that could then be used as the basis for other predictions of sediment and contaminant transport. All of these advantages appeared to make the model applications more realistic and more objective.

Unfortunately, for a number of reasons, it has proven difficult to take full advantage of these attractive features of the Freeze and Harlan distributed model blueprint. It is now more widely appreciated that the equations of these models, when applied at the model grid scale, are still only very approximate representations of the actual processes. Thus, it is still generally necessary to resort to some form of calibration to apply these models in real catchments. There are very few applications reported in the literature where results based only on measured and directly estimated parameter values are reported and these results have generally not been that accurate for predictions of either discharge or internal state variables such as water table levels or soil moisture storage. For internal state variables, there is, in fact, a problem in making such a comparison, since the local measurements are themselves generally at a much smaller scale to the element scale used in the distributed model.

These difficulties, however, do not take away the demand for predictions that are distributed in space, so there has also been a family of models that have attempted to take account of the spatial heterogeneity in catchment areas, but in a parsimonious way (that is, with only a small number of parameters). This type of model represents the variability in responses by some statistical distribution function of characteristics within the area of the catchment rather than making spatially distributed predictions directly. In some cases, these models have been constructed so that the predictions can still be mapped back into their correct spatial context and compared with spatial observations. Because of the simplifications made, however, the spatial predictions are expected to be much more approximate than for the fully distributed models. The best known examples of this type of model are the Probability Distributed Model (PDM) in which a distribution function of storage capacities is based on a purely statistical distribution with parameters to be determined by calibration; and TOPMODEL in which a distribution function of an index of hydrological similarity is based on an analysis of catchment topography. In the latter case, the similarity assumptions mean that predictions only need to be made for representative values of the topographic index, but knowing the pattern of the topographic index allows the predictions to be mapped back into space (see Beven, 2001b, for a full description of this and other distribution function models).

The Problem of Model Choice

Given this wide variety of different rainfall-runoff models available, there is a real problem of model choice for the practitioner in any practical application. Beven (2001b) summarizes a number of criteria on which to base model choice.

- Is a particular model readily available, or could it be made available if the investment of time and money appeared to be worthwhile?
- Does the model predict all the variables required for a particular aims of a project?
- Are the assumptions of the model likely to be limiting given what is known about the nature of the hydrological responses at a site?
- Can all the inputs required by the model (flow domain, boundary and initial conditions, model parameters) be provided within the time and cost constraints of the project?

More correctly these are criteria for model rejection, and it is all too common that all the available models could be rejected on the basis of one or more of these criteria. This is clearly not very helpful in practical applications, so some compromise may have to be reached. It is important, however, that the user be aware of what compromises are being made so that the associated uncertainties can, as far as possible, be evaluated.

The Problem of Model Calibration

Nearly, all model applications require some form of model calibration to identify the model parameter values appropriate to a site. This is because we do not have the techniques that would allow the direct measurement or independent estimation of parameter values at the scale required by the model (i.e. the control volume scale for a distributed model

or the catchment scale for a lumped model). There is no general agreement, however, about how best to approach the model calibration problem. This is because the calibration problem is associated with many different sources of error that cannot be easily separated. There is the error associated with the model structure, in that even the best model structures available are only an approximate representation of the actual processes. There are the errors associated with the definition of the subsurface flow domain and its boundary conditions. There are the errors associated with the specification of the initial storages within the flow domain. There are the errors associated with the measurements of the input precipitation (and other forcing variables) to the model, and their spatial and temporal variability. There are the errors associated with measurements of the discharges (and other internal variables) with which the predictions of the model will be compared. Normally, there is no way of estimating any of these different sources of error independently. We can evaluate how well a particular model predicts the observed discharges, given the specified inputs, but we do not know what is the real source of the modeling error.

In many rainfall-runoff modeling situations, this then makes it quite difficult to use the theories of parameter inference that have been well developed in statistics. These theories require that the structure of the different sources of error be defined, so that the probability or likelihood function of predicting a particular observation, given the model, can be assessed. The form of the likelihood function depends on the assumptions that are made about the nature of the errors. It is possible to lump all sources of error into a single "modeling" error, assume a particular structure for that total error, and use the statistical approach. This is an objective approach in that the validity of the assumptions about the modeling error can be tested in comparison with the actual modeling results to see if those assumptions are justified, at least approximately.

Some progress has been made in using this type of approach, particularly in real time forecasting of river flows when predictions are only required for a few time steps into the future (e.g. Young, 2002; Krzysztofowicz and Kelly, 2000). However, for the case of longer time rainfall-runoff simulations, using nonlinear distributed models, it is far more difficult to justify simple structures for the errors in space and time, especially in the face of potential errors in both the inputs and available model structures. Thus, some alternative approaches have been developed, generally based on Monte Carlo simulation in which very large numbers of model runs are made using random choices of inputs and/or parameter sets and compared with the available observations. Those models that perform well in this evaluation are then kept for use in prediction. However, this approach also cannot separate out the different sources of error and does not have the same firm theoretical basis for model evaluation as is the case where strong assumptions can be made about the nature of the errors in statistical inference.

What is important in modern model calibration, however, is that some attempt should be made to evaluate the uncertainty in the model predictions. It is no longer sufficient to find an optimal model and make a single deterministic prediction of the rainfall-runoff response. There is then too much chance of being wrong in prediction. The problem of model calibration and uncertainty estimation is discussed further in (**Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**).

Predicting the Effects of Change

One of the most interesting problems in rainfall-runoff modeling is predicting the impacts of future change (*see* **Chapter 132, Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change, Volume 3**). This might be change in land management strategies or change in climate as inferred from the global predictions reported by the International Panel on Climate Change (IPCC, 2001; *see* `www.ipcc.ch`). In both cases, it may not only be changes in runoff and subsurface storage that are of interest, but also other variables that depend on runoff such as sediment production, nutrient export, plant water stress, and so on. In addition, it may be changes in the extreme hydrological responses (floods and droughts) that may be of most interest to water managers. The interesting thing about such prediction problems is that there are no data available from the future for recalibration of the hydrological model.

Prediction of the effects of change for a particular catchment is, therefore, essentially an extrapolation problem; extrapolation from what can be learned about the catchment responses today into some future scenario. This can be difficult because there will certainly be additional uncertainties involved, but it may be very difficult to estimate the magnitude of those uncertainties. For example, the various global model outputs reported by the IPCC usually provide estimates of the change in monthly rainfalls to be expected later in the twenty-first century as a result of changes in atmospheric composition. Since these models take so long to run even on the best of today's supercomputers, such predictions are not associated with any direct estimates of uncertainty (though different global models from different modeling groups currently produce quite widely differing predictions). Thus, it is impossible to associate any probability estimate with these scenarios. In addition, to model the impact of change on the hydrological extremes, it will be necessary to interpret these predicted mean monthly changes into changes in the distributions of rainfalls (and other climate variables) when it is not clear if the relationships seen between atmospheric and ocean circulation patterns and rainfall extremes in a region today

will be the same under a changed climate. Many different scenarios could be envisaged for how these relationships might change, but again it will be impossible to associate an uncertainty measure with a potential scenario. Similar considerations hold for potential changes in land management that might have more to do, for example, with future agricultural subsidy policies than with the impacts of climate change.

Thus, we know that predictions of future change will be uncertain but cannot easily quantify that uncertainty. About the only strategy that can be taken in this situation is to assess the uncertainty in the predictions of a model under current conditions and then run different future scenarios with that model taking account of similar input and parameter uncertainty assumptions. Any relative possibility estimate for different scenarios will be necessarily subjective, but such a procedure will give the most realistic assessment of the potential future hydrological responses. The implication, however, is that it will be very important to continue monitoring catchments into the future so that as the responses change they can be used to revise and update the scenario predictions.

The Future of Rainfall-runoff Modeling

New computer technologies seem likely to change the way that rainfall-runoff models are constructed and used as components of large-scale integrated modeling frameworks for environmental management. In particular, the possibility of using large-scale (the GRID; *see* **www.gridcomputing.com**; **www.gridforum.org**; **www.globus.org**) computer networking to link together distributed database and computational engines means that it will be possible to couple together models of many more different environmental systems across disciplinary boundaries and across national administrative boundaries. This is, in fact, already possible and is already happening on a limited basis as demonstrated, for example, in the regional water resources models under construction in Denmark or the national models for environmental management being used in the Netherlands. In Japan, the very large-scale "Earth Simulator" is aiming to implement global scale models (see Earth Simulator Center, 2003; see **www.es.jamstec.go.jp/esc/eng/ES/index.html**).

There is then, however, a real question raised about how this type of interdisciplinary model might be best implemented. In the past, comprehensive modeling systems have been constructed as large complex computer programmes. These programmes were intended to be general, but have proven to be expensive to develop, difficult to maintain and difficult to apply because of their data demands and needs for parameter identification. With GRID computing technology, it will be possible to continue in the same vein, but with more coupled processes and finer spatial and temporal resolutions for the predictions. It is not clear, however, whether this will result in a real improvement in model accuracy and use because the problems inherent in the current generation of distributed environmental models do not necessarily easily go away with improvements in space and time resolutions of the component models.

We can distinguish two different (albeit overlapping) types of models here in terms of constraints. In the first, the accuracy of solutions is still constrained by computational resources; in the second model accuracy is primarily a result of lack of knowledge of appropriate process representations and boundary conditions. In the first type of models, real advances may still be possible as computational constraints are relaxed. In atmospheric modeling, for example, there is still scope for improvements of the representation of local convection and rainfall forecasts, in the representation of sub-grid spatial variability of energy fluxes, and in the representation of topography by finer grid scales. Ultimately, however, this type of model will be constrained by the need to know finer and finer detail of boundary conditions and parameter values, as is already the case in the second type of model. An example of this second type is classical distributed hydrological models.

New developments in environmental modeling will allow a new approach to be taken to this problem based on scale-dependent model objects, databases, and spatial objects in practical applications to specific places. One of the most exciting benefits of the possibilities provided by the GRID in environmental modeling is the potential to implement models available from different institutions as a process of *learning about specific places*. It will be possible, in fact, to have models of all places of interest. However, as argued by Beven (2000, 2001a, 2002b, 2004), as a result of scale, nonlinearity, and incommensurability issues, the representation of place will be inherently uncertain so that this learning process should be implemented *within an uncertainty estimation framework*.

Sites of interest for a particular prediction can be implemented as *active objects*, seeking information across the GRID to achieve a specified purpose, and using the power of parallel computing resources to estimate the uncertainty associated with the predictions as constrained by *site-specific observations*, including those accessed over the GRID in real time. Initially model results, based perhaps on only GIS databases and limited local information, may be relatively uncertain but experience in monitoring and auditing of predictions will gradually improve the representation of sites and boundary conditions. It is this learning process that will be critical in the development of a new generation of environmental models that are geared toward the management of specific places, rather than general process representations.

That is not to say that models of places will not require process representations, but there is a real research question about how detailed a process representation is necessary to

be useful in predicting the dominant modes of response of a system, given the uncertainties inherent in representing the processes in places that are all unique. This appropriate complexity issue has become obscured by the desire to build more and more scientific understanding into models, including physical, chemical, and biological components. This desire is perfectly understandable, it is a way of demonstrating that we do understand the science of the environment, but it results in models that have lots of parameter values that cannot be easily measured or estimated in applications to real places. There is always a certain underlying principle in science, that as we add more understanding and eliminate empiricisms, then the application of scientific principles should become simpler and more robust. This does not seem to have been the experience in the practical application of environmental models.

Events such as the river flooding in the UK in 2000, 2002 and 2004 and the consequences of the 2001 fires in the US, have demonstrated the need for a new generation of systems for environmental forecasting. The subtle (and sometimes not so subtle) coupling between atmospheric forcing, catchment response, river runoff, and transport processes requires the dynamical coupling of many components to capture these subtleties. Components would be a representation of the regional atmosphere and the terrestrial surface and subsurface hydrology that would interact through different boundary conditions. Built on the fluxes within those models air and water pollutant transport models and biogeochemical models could be implemented locally within the regional-scale domain. Each component would be able to assimilate data transmitted from field sites and assess the uncertainty in the predictions. The components would share 4-D/5-D visualization tools with appropriate interactive user interfaces. Users will be able to access the current data, visualize predictions for particular locations and play what-if scenario games over different timescales. The structure of the system would be such as to facilitate and even stimulate improvements to the representation of different components and the constraint of predictive uncertainty by field data collection and data assimilation. The potential capabilities of the GRID underlie all these components, though much could already be achieved using the Web technology of today. Examples of steps toward this type of integrated system (albeit essentially raster based) include the US Inter-Agency Object Modeling System (OMS), (*see* **Chapter 129, Rainfall-runoff Modeling for Integrated Basin Management, Volume 3**).

Such an integrated system should operate both in real time, assimilating data and boundary conditions from larger scale models, and displaying the "current state of the environment", as well as providing the potential to update model predictions into the future under different scenarios. Initiatives such as the European Union Water Framework Directive (see **www.europa.eu.int/comm/environment/water-framework/index-en.html**; **www.defra.gov.uk/environment/water/wfd**) are increasing demands for predictions of this type about the *responses of specific locations to change* in a way that integrates hydrological and ecological considerations in management. The system would need to be powerful enough to be used for assessing *uncertainties in model predictions and the consequent risks of potential outcomes*. It should also be able to be used off-line for "what-if" management purposes or decision support including developing strategies for risk-based sustainable management in the context of climate and other changes. This will include the evaluation of management of subsystems including for licensing of airborne emissions and effluents to water courses; strategies for remediation of contaminated land, rivers and estuaries, and so on.

An essential element of this strategy will be the need, as far as possible, to "future proof" the model and database systems used; avoiding, for example, a strict raster based approach or a commitment to one particular modeling framework. The key will need to be flexibility. Raster databases will continue to be driven by remote sensing imaging inputs to the modeling process, and, in some cases, by convenient numerical solution schemes for partial differential equations. However, it is often inappropriate to force an environmental problem into a raster straightjacket. Treating places as flexible active objects might be one way around this future proofing problem. Defining the spatial domain of a prediction problem would allow that place, as an active object, to search on the GRID for appropriate methods and data for resolving that problem, and also for appropriate methods and data for providing the boundary conditions for the problem (which might then involve other modeling or data extrapolation techniques).

There are some interesting implications of such an approach. One is that the variety of modeling methods available across the GRID to solve a prediction problem might be able to be compared more readily, leading to better understanding of issues of appropriate model complexity for different modeling problems. This will especially be the case if, as part of the learning process, simulations are saved to be compared with later observations of the real outcome. This use of "post-audit" analysis has been rarely used in environmental modeling, but has been instructive in the field of groundwater modeling (Konikow and Bredehoeft, 1992; Anderson and Woessner, 1992) and is routine in atmospheric modeling in the evaluation of forecast skill (although the evaluation of global climate model predictions still requires an element of compromise at the regional level).

To be useful, however, the process of model application will require the definition of a self-coding system attached to places to record and retrieve the methods that have been applied to (or by) that place in the past so that they can be easily reviewed and evaluated by the user. There is then

a further interesting question that arises as to how far the place, once defined for a problem, can learn about itself from the data and model predictions available; using methods such as fuzzy classification or genetic programming as tools to extrapolate data from that and other sites to develop predictive methods of appropriate complexity to the problem at hand within the limitations of the uncertainties implied by the data available. This approach has been advocated, for example, by the proponents of "hydroinformatics" (e.g. Abbott, 1991).

The learning framework that underlies this framework is best suited to systems that are not changing. In that way new data should allow a refinement of the feasible model representations and reduction in the predictive uncertainty. Many of the predictions required of environmental systems, however, involve questions of current or future change under different scenarios. Such predictions will be even more uncertain than the simulation of current conditions, but there has been very limited work on estimating the uncertainties of potential outcomes in future scenario simulations, and less on the conditioning of those predictions as monitoring of changing conditions changes. Continued monitoring and recursive data assimilation, in this framework then becomes a tool for following drift in system response (within the limitations of data uncertainties.

Perhaps a theme can be identified in running through this discussion of environmental models of everywhere: the focus on data. Data will be required to characterize places, to drive model predictions, to evaluate the results of model predictions, to reject some models previously considered feasible, and to monitor changes in system response. Once all places are represented within the flexible GRID-based system envisaged, the data may assume a greater importance than model structures as a means to refine the representation of each place within a learning framework. The result may be a new way of looking at environmental modeling that transcends the traditional goal of incorporating all our understanding of the complexity of coupled environmental systems into a single mathematical framework with a multitude of parameters that cannot easily be identified for any particular place (Beven, 2003, 2004).

Further Information

There have been a number of reviews of rainfall-runoff models published in the last decade. For more details, the reader is referred to Singh (1995); Abbott and Refsgaard (1996); and Beven (2001b). Concise descriptions of particular models can be found in Singh (1995) and Singh and Frevert (2002a,b). Extensive discussions of the model calibration problem can be found in Anderson and Bates (2001).

REFERENCES

Abbott M.B. (1991) *Hydroinformatics, Information Technology and the Aquatic Environment*, Avebury Technical: Aldershot and Brookfield.

Abbott M.B. and Refsgaard J.-C. (Eds.) (1996) *Distributed Hydrological Modelling*, Kluwer: Dordrecht.

Anderson M.G. and Bates P.D. (Eds.) (2001) *Model Validation: Perspectives in Hydrological Science*, Wiley: Chichester.

Anderson M.P. and Woessner W.W. (1992) The role of the post-audit in model validation. *Advances in Water Resources*, **15**, 167–174.

Anderton S., Latron J. and Gallart F. (2002) Sensitivity analysis and multi-response, multi-criteria evaluation of a physically based distributed model. *Hydrological Processes*, **16**, 333–353.

Beven K.J. (1989) Changing ideas in hydrology: the case of physically based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (2000) Uniqueness of place and the representation of hydrological processes. *Hydrology and Earth System Sciences*, **4**(2), 203–213.

Beven K.J. (2001a) How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, **5**(1), 1–12.

Beven K.J. (2001b) *Rainfall-runoff Modelling – The Primer*, Wiley: Chichester.

Beven K.J. (2002a) Towards a coherent philosophy of environmental modelling. *Proceedings of the Royal Society of London*, **458**, 1464–1485.

Beven K.J. (2002b) Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system. *Hydrological Processes*, **16**(2), 189–206.

Beven K.J. (2003) On environmental models of everywhere using the GRID. *Hydrological Processes (HPToday)*, **17**, 171–174.

Beven K.J. (2004) Towards Environmental Models of Everywhere: Advances in modelling and data assimilation. In *Hydrology: Science and Practice for the 21st Century*, Webb B., Arnell N., Onof C., McIntyre N., Gurney R. and Kirby C. (Eds.), Vol. 1, 244–250.

Clarke R.T. (1973) A review of mathematical models used in hydrology, with some observations on their calibration and use. *Journal of Hydrology*, **19**, 1–20.

Earth Simulator Center (2003) *Annual Report of the Earth Simulator Centre, April 2002-March 2003*, The Earth Simulator Center: Yokohama.

Freeze R.A. and Harlan R.L. (1969) Blueprint for a physically-based digitially simulated hydrologic response model. *Journal of Hydrology*, **9**, 237–258.

Horton R.E. (1933) The role of infiltration in the hydrological cycle. *Transactions-American Geophysical Union*, **14**, 446–460.

Intergovernmental Panel on Climate Change, (IPCC) (2001) *Climate Change, 2001*, Vol. 4 Cambridge University Press: Cambridge.

Konikow L.F. and Bredehoeft J.S.D. (1992) Groundwater models cannot be validated. *Advances in Water Resources*, **15**, 75–83.

Krzysztofowicz R. and Kelly K.S. (2000) Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water Resources Research*, **36**(11), 3265–3277.

Lamb R. and Beven K.J. (1997) Using interactive recession curve analysis to specify a general catchment storage model. *Hydrology and Earth System Sciences*, **1**, 101–113.

Lamb R., Beven K.J. and Myrabø S. (1998) Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. *Advances in Water Resources*, **22**(4), 305–317.

Loague K.M. and Kyriakidis P.C. (1997) Spatial and temporal variability in the R-5 infiltration data set: déjà vu and rainfall-runoff simulations. *Water Resources Research*, **33**, 2883–2896.

Richards B.D. (1944) *Flood Estimation and Control*, Chapman & Hall: London.

Sherman L.K. (1932) Streamflow from rainfall by unit-graph method. *Engineering News Record*, **108**, 501–505.

Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Water Resource Publications: Littleton.

Singh V.P. and Frevert D.K. (Eds.) (2002a) *Mathematical Models of Large Watershed Hydrology*, Water Resources Publications: Highlands Ranch.

Singh V.P. and Frevert D.K. (Eds.) (2002b) *Mathematical Models of Small Watershed Hydrology and Applications*, Water Resources Publications: Highlands Ranch.

Tallaksen L.M. (1995) A review of baseflow recession analysis. *Journal of Hydrology*, **165**, 349–370.

Van der Kwaak J.E. and Loague K. (2001) Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model. *Water Resources Research*, **37**, 999–1013.

Wheater H.S., Jakeman A.J. and Beven K.J. (1993) Progress and directions in rainfall-runoff modelling. In *Modelling Change in Environmental Systems*, Jakeman A.J., Beck M.B. and McAteer M.J. (Eds.), Wiley: Chichester, pp. 101–132.

Young P.C. (2002) Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society of London*, **B360**, 1433–1450.

123: Rainfall-runoff Models for Real-time Forecasting

EZIO TODINI

Department of Earth and Geo-Environmental Sciences, University of Bologna, Bologna, Italy

This article describes the use of rainfall-runoff models within the frame of real-time flood forecasting and warning. After assessing the requirements of real-time flood forecasting systems, some of the more well known ones available for operational purposes are briefly mentioned. The different types of rainfall-runoff models used in practice are then presented, within the framework set on the one hand by the operational system requirements, and on the other hand by the historical evolution of research and development in the fields of hydrological and meteorological forecasting. After discussing the problems of real-time updating and of the assessment of the predictive uncertainty, two example applications are illustrated. The article concludes with a description of current and future research themes for extending the forecasting lead time, assessing forecasting uncertainty, and appropriately communicating it to the end users.

INTRODUCTION

Definition of Real-time Flood Forecasting Systems

Real-time flood forecasting systems are tools aimed at reducing the uncertainty on the evolution of future events thus allowing decision makers to take the most effective decisions under uncertainty (Raiffa and Schlaifer, 1961; Todini, 1999). This means that, to be operational, flood forecasting systems are not only required to be timely and sufficiently accurate within predetermined time horizons but should also provide a usable quantification of the forecasting uncertainty.

The following example (Figure 1), relevant to the decision of issuing a flood warning, can demonstrate why a flood forecast without a measure of its uncertainty can be of no value and sometimes even dangerous. In the example, if the water level in the channel rises above the dykes, damages will occur with costs increasing with the dot–dashed line. The forecast says that the expected value will stay below the top of the dyke, which seems to imply that the expected damage costs will be zero and no action should then be taken. Unfortunately, because of the forecasting uncertainty always embedded in a forecast (depicted in Figure 1 as a probability density function with mean on the forecasted value), one can immediately notice that the expected damage (the integral of the product of the density times the cost) is not null and the most appropriate decision may then be totally different.

Uncertainty increases with the forecasting horizon, which is a function of the time required to implement the flood protection measures. The forecasting horizon is a characteristic of each individual problem and may vary from a few hours needed to safeguard a small village to more than four days, as it is required to protect the city of Canton, given that the dykes of the Bei Jiang river must be blown with dynamite after evacuating 2 million inhabitants from their villages in the water detention areas to the surrounding hills.

Whenever the forecasting horizon required for the implementation of the flood management interventions is sufficiently short, when compared to the characteristic response time of the river, one should use hydraulic routing models to compute downstream discharge and/or level forecasts from the upstream measured ones, given that the hydraulic models tend to be more accurate than the hydrologic rainfall-runoff models. Unfortunately, in most situations the lead time provided by the hydraulic models is insufficient to allow for the implementation of the protection measures. In these cases, the use of rainfall-runoff models is inevitable. If the forecasting chain starts from

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

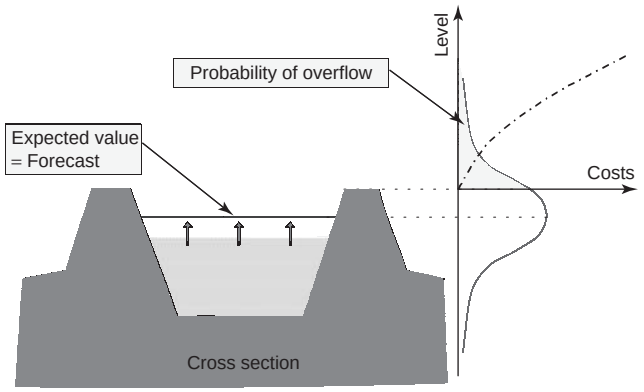


Figure 1 The problem of issuing an alert under flood forecasting uncertainty. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the measured rainfall that is to be used as input to a rainfall-runoff model, one gains forecasting lead time as a function of delays, mostly because of soil storage filling and surface waters' travel time, but inevitably introduces additional errors and uncertainty because of the simplifications of the processes' representation and the parameterizations typical of the hydrological rainfall-runoff models. Finally, when the time horizon gained using the measured rainfall is still insufficient, as in small steep or urban catchments or more generally when the required lead time is very large compared to the characteristic space–time dimensions of the phenomena, one must resort to use RADAR-based nowcasting (0–6 h) (MUSIC, 2001, **http://www.geomin.unibo.it/orgv/hydro/music/**) or quantitative precipitation forecasts (QPF) provided by numerical weather prediction (NWP) models (1–4 days) (EFFS, 2003, **http://effs.wldelft.nl**). Inevitably, the level of uncertainty increases greatly, and overcoming this problem currently represents the frontier of research efforts (see Section on "The HEPEX Project").

The assessment of the problem relevant to the estimation of a measure of uncertainty to be provided to the decision makers, which would incorporate all the different sources of uncertainty, has been analyzed in detail (Krzysztofowicz, 1999), and a Bayesian combination approach has been proposed. Although very rewarding from a speculative point of view, its practical implementation (Krzysztofowicz and Kelly, 2000; Krzysztofowicz and Herr, 2001; Krzysztofowicz, 2001) is still under evaluation.

Operational Systems

One must acknowledge that a wide gap still exists today between operational flood forecasting and research in the related field. Most people working in operational flood forecasting still perceive the forecast as a "deterministic" value to be compared to a threshold in order to select the most appropriate decision. They tend to look for simple packages with extremely simple solutions that would fit into their way of thinking. On the other hand, researchers tend to overparameterize the models, to overfit the data, and to produce extremely complex outputs or pieces of information instead of trying to integrate the complexity of the uncertainty estimation into some simple decision rule to be easily understood by the decision makers.

In this context, it is easy to find operational systems that, in reality, never worked properly because of their complexity while some very simple approaches, although unsatisfactory from the scientific point of view, have been fully understood and properly used by decision makers.

Nonetheless, the definition of a real-time flood forecasting system as a tool for reducing the uncertainty in the future events implies that all the components included in the system must be (i) robust, since a forecast must be provided, (ii) sufficiently accurate, (iii) adaptive, and (iv) capable of providing, at each step in time, an overall measure of uncertainty. Moreover, the operational system must run round the clock, accept inputs arising from several sources generally stored in different databases, detect incoming data errors, and rebuild the missing data. As one can imagine, from the schematic representation of one of the commercially available real-time flood forecasting systems (Figure 2), this can be quite a sophisticated and complex piece of software that must guarantee its continuous operation to the user.

As can be seen from Figure 2, the European Flood Forecasting Operational Real Time System (EFFORTS), is a complex real-time flood forecasting system that incorporates automatic data validation, sampling, and reconstruction; stochastic extrapolation of rainfall and upstream inflows; rainfall-runoff modeling and 1-D flood routing; stochastic assessment of uncertainty using the Mutually Interactive State and Parameter (MISP) estimation algorithm of Todini (1978).

As a matter of fact, the number of existing operational real-time flood forecasting systems is not too large and, moreover, not many of these systems can be taken as representative of a specific approach, this is why this article will only briefly mention the most widely known commercial packages. Several operational systems are now available starting from the historical ALERT system based upon the Sacramento model, widely used in the United States, since the end of the 1970s, to the Flood Works (FW) developed by Wallingford Software (**http://www.wallingfordsoftware.com/products/rivers/**) and based on the Real Time Flood Forecasting System (RFFS) (**http://www.nwl.ac.uk/ih/www/research/mfloodfore.html**) developed at CEH, to the SMHI Real-time Flood Forecasting system based on the HBV model (Bergström, 1976, 1992, 1995), to the more

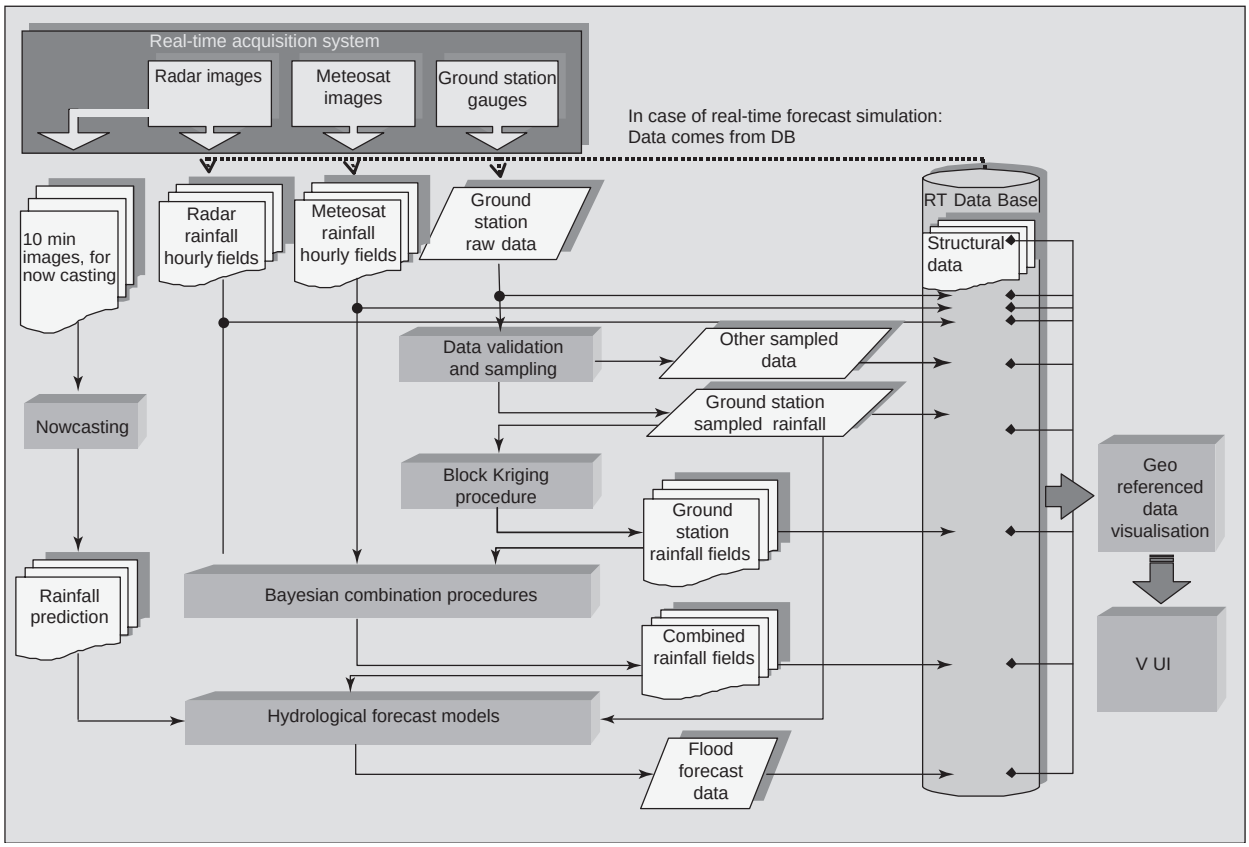


Figure 2 Schematic representation of EFFORTS an operational real-time flood forecasting system (by permission of PROGEA (Protezione e Gestione Ambientale) Srl). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

recent MIKE FLOODWATCH (DHI, 2000a), which couples the NAM model (DHI, 2000c) with MIKE 11 FF (DHI, 2000b) into an operational GIS-based flood forecasting system, and to EFFORTS in its two GIS-based versions, the first one based on the ARNO model (Todini, 1996; 2002a) and the second one based on the TOPKAPI (Todini and Ciarapica, 2002; Liu and Todini, 2002). Last, but not least, a shell for incorporating alternative real-time flood forecasting models and in particular, LISFLOOD (De Roo *et al.*, 1998; 2000), although not yet fully operational, is currently under development at Delft Hydraulics within the framework of project EFFS.

RAINFALL-RUNOFF MODELS FOR REAL-TIME FLOOD FORECASTING

It is not easy to generalize a typology for the rainfall-runoff models used for real-time flood forecasting because the choice of the model heavily depends on the scale of the catchment, the rapidity of the phenomena, and the required forecasting lead time. As can be easily understood, one thing is to deal with real-time forecasting in some small flashing catchment or alternatively in very large catchments such as those of the Po, the Rhine, the Danube, the Mississippi, or the Yangtze rivers.

In addition, there are two lines of thought in the development of rainfall-runoff models for flood forecasting. The first one, mostly originated by the continuous flow of available data generally collected round the clock, places more emphasis on measurements and develops extremely simple statistical or parametric models, the parameters of which may be estimated and continuously updated by means of observations. The second line of thought, still recognizing the importance of the available data, believes that the highly nonlinear behavior of the rainfall-runoff processes may not be fully learned from the limited set of data used for calibration (the learning set) and must be introduced as *a priori* knowledge into the models in order to reduce uncertainty and improve the reproduction of the phenomena beyond the range of observations (Kitanidis and Brass, 1980a,b; Brath and Rosso, 1993).

The consequence to this dichotomy is that one can find all sorts of operational systems that range from extremely simplified models, possibly recalibrated in real-time, to systems based on extremely complex rainfall-runoff models. In addition, real-time flood forecasting models developers

still discuss the use of event type models as opposed to continuous type explicit soil moisture accounting (ESMA) models.

During recent decades, most of the operational flood forecasting systems based on rainfall-runoff models were developed using continuous type ESMA models (WMO, 1975; Todini, 1988; Franchini and Pacciani, 1991). The main reason for this is because ESMA-type models are capable of continuously updating the soil moisture conditions on the basis of the water balance equation. The major limit to the use of simple linear models lies in the fact that they do not correctly represent the nonlinear threshold-type soil behavior and the split of rainfall into infiltration and surface runoff. Similarly, the event type models, not performing a continuous mass balance computation, must be initialized with actual soil moisture conditions, which heavily condition the forecasts in real time and can hardly be measured at a catchment scale. From the original ESMA models developed in the 1960s and early 1970s (WMO, 1975) needing an exaggerated number of parameters (generally more than 20), new types of models were developed in the 1980s needing only three to six parameters but fully representing the dynamics of saturated areas, which was recognized as a main cause of runoff generation in large floods. This result originates from the finding that, for the purposes of flood forecasting, the most important effect that models must represent is the overall soil filling and depletion mechanisms and the resulting variation in the extension of the basin saturated areas; in fact, when large portions of the basin reach saturation the rainfall that falls over these areas has a direct effect on the flood magnitude without being attenuated by the soil's absorption capacity. These concepts have spawned the new ESMA-type models, such as for instance the PDM model (Moore and Clarke,1981; Moore, 1985) or the Arno model (Todini, 1996; 2002a) that have been widely used and extensively described in the literature. In general, these models have been developed as semidistributed models, although the problem of their adaptation to ungauged subcatchments has not yet been properly resolved.

More recently, deterministic spatially distributed rainfall-runoff models have been used more widely. Apart from the recent availability of GIS tools and more powerful computers, there are several reasons for this tendency. The main reason lies in the fact that the physical meaning of processes and relevant parameters can only be preserved at small scales, while this physicality is lost when the processes are aggregated at subcatchment or catchment scales. On the basis of equations representing the storage and the movement of water in the soil and on the surface, these models are potentially easier to calibrate by relating the values for their physically meaningful parameters to additional information provided in the form of Digital Elevation Maps, Soil Maps, and Land Use Maps. These models, which attempt to represent more realistically the different processes at scales finer than the subcatchment, account for the spatial distribution of rainfall (now available at pixels of between 1×1 km and 10×10 km from RADAR images or in terms of QPF from NWP models) and are better equipped to address the problem of extrapolating the parameter values to the ungauged subcatchments, a problem for which a proper solution has not yet been found.

All these potential properties, combined to the increased efficiency of computers, has given rise to a number of fully distributed rainfall-runoff models, some of which have been (while others are in the process of being) incorporated into real-time flood forecasting systems.

Black-box Models

Black-box models have a long history in hydrology. In 1932, Sherman introduced the concept of the unit hydrograph based on the principle of superposition of effects. Even though it had not yet been formalized at that stage, the superposition principle involved many assumptions: a basin, under the effects of rainfall-runoff transformation, should behave like a linear, dynamic, time invariant causative system. This unit hydrograph definition nevertheless excited the interest of hydrologists who by then were able to provide not just an estimate of the peak discharge as with the rational method but also the behavior of flood flows under the action of rainfall with a more or less complex pattern.

Yet, the unit hydrograph method presented a certain number of problems that were solved using more or less subjective methods like

- the separation of surface runoff from base flow;
- the determination of "effective" rainfall, namely, that portion of the rainfall that is not lost via evaporation and transpiration and does not filter down through the soil; and
- the actual derivation of the unit hydrograph from the available data.

The leap in quality only came about in the 1950s when hydrologists mastered systems theory techniques. They realized that the unit hydrograph was merely the response of a linear, dynamic, time invariant, causative system and that the use of techniques such as Z transforms and Laplace or Fourier transforms enabled them to derive this response from an analysis of the system's input and output data.

This was the period in which "conceptual" models came into being. The derivation of the unit hydrograph based on sampled input and output data, known as the *inverse problem*, was still a complex problem both because of the less than perfectly linear behavior of the basin system and also because of gross errors in the measurement of input and output quantities. To overcome these problems, hydrologists found that the forms of the unit hydrograph

could be obtained as the solutions of differential equations describing the temporal behavior of simple systems, such as a linear reservoir or a cascade of linear reservoirs, for example, the Nash model (Nash, 1958, 1960). The unit hydrograph thus came to be expressed as a function of a few parameters that could be estimated using statistical techniques: moments, cumulants, regressions, maximum likelihood, and so on, thereby permitting regionalization and extension to basins without measurements by regressing the parameters obtained as a function of the physiographic characteristics of the basin. A plethora of these models gave rise to an incredible variety of solutions: a cascade of linear reservoirs, linear channels, linear channels and reservoirs, nonlinear reservoirs (Prasad, 1967), and so on.

On the other hand, the derivation of the form of the unit hydrograph by means of transforms based on observed data still presented enormous difficulties as far as its application to real cases was concerned: in a classic article, Rao and Delleur (1971) highlighted the extent of the effects induced, on a unit hydrograph estimate made using Fourier transforms, by errors of measurement on input and output data. Only after the works of Wiener (1949) and Tikhonov (1963a,b) and the introduction at the estimation stage, of elementary physical concepts, such as mass balance or as regularity and nonnegativity of unit hydrograph ordinates (Eagleson *et al.*, 1965; Natale and Todini, 1976a,b; Bree, 1978; Bruen and Dooge, 1984) were realistic and reliable unit hydrograph estimates obtained.

At the same time, research into nonlinear systems led to representations based either on Volterra integrals of an order greater than the first, or by means of projections based on orthogonal polynomials (Amorocho and Orlob, 1961) or based on piecewise linearizations (Todini and Wallis, 1977; Todini, 2002b) aimed at reproducing the sharp soil saturation effect.

Box and Jenkins (1970) provided hydrologists with new instruments for expressing unit hydrographs in a discretized form by means of autoregressive exogenous variables models (ARX) or autoregressive moving average models (ARMA) (Clarke, 1973; Ubertini, 1982), and analogies with existing conceptual models were pointed out by Spolia and Chander (1974). In subsequent developments of the applications by means of the techniques pioneered by Box and Jenkins there was, however, a loss of "physicality" in the models (particularly when using ARIMA and Intervention Analysis models) in favor of mere mathematical and system identification techniques. The loss of physicality is even more accentuated with the new Artificial Neural Network (ANN) approaches, which can be viewed as nonlinear analogues of the transfer function models. This can turn into poor forecasts when the events are larger than the one experienced in the training set (Cameron *et al.*, 2002).

Although research is very active in this domain as, for instance, Shamseldin (1997) and Dawson and Wilby (2001), who provide an interesting review on application of ANN to rainfall-runoff modeling, very few examples of existing operational systems are presently based on ANN (Garcìa-Bartual, 2002). More recently, a Data Based Mechanistic (DBM) modeling approach was introduced by Young (2002), which, although deriving the model structure and the parameter values from the input and output data by means of system engineering identification and parameter estimation techniques, attempts to go beyond the black-box concept by selecting those (not necessarily linear) model structures that are considered physically meaningful (Young, 2001, 2002). Adaptive DBM models were used as the basis for an operational flood forecasting system for the river Nith in Scotland (see Lees *et al.*, 1994). Although, the DBM modeling approach recognizes the importance of the physical coherence of the identified model structure, it derives it from the observations, thus disregarding *de facto* the results of at least 50 years of research efforts aimed at specifying the hydrological physical mechanisms that generate floods. This might cause problems in forecasting beyond the range of observations that are available in calibrating a DBM model, or where those data are highly uncertain in times of flood, but the ease of implementing these models in an adaptive form may mitigate against this. A full application of the Bayes principle would suggest that all possible *a priori* knowledge (in this case our knowledge on the hydrological processes and possibly on the parameter values as well as past observations) should be combined with the current observations to obtain less uncertain *a posteriori* forecasts, but it remains a research question as to how best to do so.

The ESMA-type Models

In the 1960s, new types of hydrological cycle models were introduced. In an attempt to achieve a better physical interpretation of the phenomena, experts tried to represent the behavior, at basin level, of each individual component in the hydrological cycle by using a group of interconnected conceptual elements, each of which represent the response of a particular subsystem. The need for this kind of approach was due to newly emerging requirements implying the extension of hydrological models to continuous simulation on larger catchments by overcoming the complexity and subjectivity of the separation of surface and base flows and of the derivation of the effective rainfall.

A large number of these models were developed: Dawdy and O'Donnell (1965), the Stanford Watershed Model IV (Crawford and Linsley, 1966), the Sacramento (Burnash *et al.*, 1973), the SSARR (Rockwood, 1982), the Tank (Sugawara, 1967, 1995), and so on, which represented in different ways the response mechanism of the various phenomena and the interconnections between the various subsystems, and which at that time were regarded as the very best that could be achieved with models. Figure 3

shows a block diagram of the different processes generally represented in the ESMA models.

In theory, the parameters of these models, such as the storage coefficients, roughness coefficients, or the different thresholds, could be somehow set in relation to the physiographic quantities of the basins. In reality, when the parameter estimates are made on the basis of objective functions to be minimized (e.g. the sum of squared deviations), groups of unrealistic parameters are obtained which incorporate both data measurement errors and the errors present in the structure of the model itself; in addition the observability conditions are not always guaranteed, particularly for multiple input–output hydrological models, as pointed out by Sorooshian and Gupta (1983) and by Singh and Woolhiser (2002). ESMA-type models have given rise to several real-time flood forecasting systems. In particular, among them four specific models have to be recalled, since they have given rise to operational real-time systems. The first one is the National Weather Service-River Forecast System (NWS-RFS) also known as the Sacramento model (Burnash *et al.*, 1973). The NWS-RFS was used as the basis of the ALERT flood forecasting

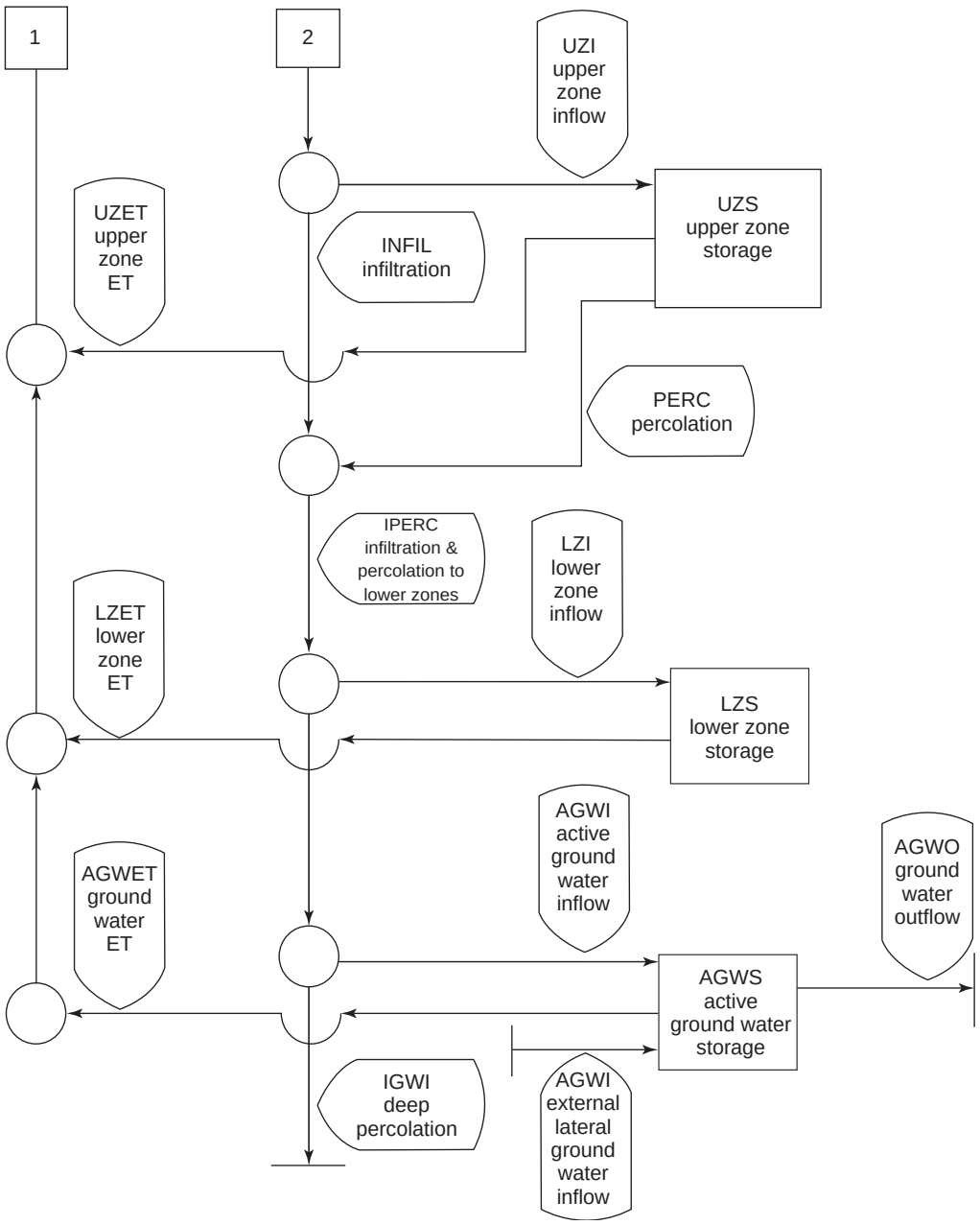


Figure 3 A schematic representation of storages and processes in an ESMA-type model

system, which is widely used in the United States, and for the development of several flood forecasting systems around the world, such as for example, the forecasting systems of the Hindus (WMO/UNDP, 1980) and the Nile in Sudan (Elzeim and Adam, 1996). A second model, the SSAR (Rockwood, 1982) is operationally used for flood forecasting in the Mekong (Tanaka, 1999). Another widely used model, the HBV (Bergström, 1976, 1992, 1995), which is part of the Swedish SMHI flood forecasting system, has found several applications in many countries and in particular, in the North European countries. Last but not least is the NAM model (Nielsen and Hansen, 1973; DHI, 2000c), which is included as the rainfall-runoff model in the MIKE FLOODWATCH software package (DHI, 2000a), which is becoming more and more widely used throughout the world.

The Variable Contributing Area Models

A new generation of ESMA-type models, the variable contributing area models, appeared around the end of the 1970s with the work of Zhao (1977) which originated the Xinanjiang model used in China for forecasting floods on the Xinanjiang river. More or less at the same time, and quite independently, Moore and Clarke (1981) (see also Moore, 1985), proposed the Probability Distribution Moisture (PDM) model, while Juemou *et al.* (1987) developed the Synthesized Constrained Linear Systems model (SCLS) by combining the Xinanjiang soil water production function with the Constrained Linear Systems (CLS) model (Natale and Todini, 1976a,b; Todini and Wallis, 1977; Todini, 2002b). Few years later, by modifying the Xinanjiang soil water production function, Todini (1996, 2002a) developed the ARNO model, from which Wood *et al.* (1992) derived the VIC by increasing the number of soil layers (Liang *et al.*, 1996a,b). All these models were developed as lumped semidistributed models on the assumption of a spatially variable storage capacity for infiltration at the soil surface: namely, all the precipitation enters the soil and surface runoff is originated by saturation of the first soil layer.

The core of these models is a two-parameter curve representing the relation between the total volume of water stored in the soil and the extension of the saturated areas. Unfortunately, the parameterization of this curve, as well as that of the other processes represented (drainage, percolation, groundwater flow, etc.) is based on empirical parameters to be estimated from the data. Beven and Kirkby, (1979) originated a more physically meaningful lumped model by following an alternative approach to derive a similar relation. They originated the TOPMODEL on the assumption that the accumulation of soil moisture can be approximated by successive steady states of the water table originating in a first layer of soil. On this basis, they derived a new relation between the volume of water stored in the soil and the extent of areas at saturation on the basis of physically meaningful parameters.

All these models, generally combined with simplified hydrological overland flow and channel routing components, have been transformed into semidistributed models and are used for the setting up of operational real-time flood forecasting systems. For instance, after the applications of the original Xinanjiang system, most real-time flood forecasting systems in China have been modeled using the SCLS, which recently came out first in a Chinese nationwide competition.

In the United Kingdom, the Center for Ecology and Hydrology combined the use of the PDM model and RADAR rainfall data producing the RFFS, 2005, (**`http://www.nwl.ac.uk/ih/www/research/mfloodfore.html`**), which constitutes the core of the FW system, a generic software package oriented to real-time flood forecasting developed by Wallingford Software, 2005 (**`http://www.wallingfordsoftware.com/products/rivers/`**). In Italy, the ARNO model, which incorporates linear parabolic routing on the catchment slopes and in the channels, has been used for many years as the core of the EFFORTS. The system, originally developed for the Fuchun river in China (ET&P, 1992), was successively applied to the Danube in Germany as well as to several Italian rivers.

The Distributed Models

The lack of any real "physical" connection between lumped models and reality, which emerged in the work of Wooding (1965a,b; 1966) and Woolhiser and Liggett (1967) where kinematic models were used for the study of small urban basins, prompted Freeze and Harlan (1969) to propose, albeit only as a future project, the creation of a mathematical model based on distributed physical knowledge of the phenomena by means of the numerical integration of the various systems of differential equations describing surface flow, flow in unsaturated soils, and flow in water tables, matching the solutions of a subsystem with the boundary conditions of another. Unfortunately, the applications, not the least because of the limited computational resources available at that time, were directed only at small, almost impermeable mountain basins and were designed to evaluate flow in the catchment-closing section without highlighting the characteristics peculiar to this type of approach.

The need to evaluate the effects of modifications to the hydrological cycle engendered by changing agricultural practices and more generally by anthropogenic intervention prompted several research institutes such as the Danish Hydraulic Institute (DK), the Institute of Hydrology at Wallingford (UK), and SOGREAH (F) to return to Freeze and Harlan's proposal, and to produce a model that expressed the various phenomena by integrating the differential equations with partial derivatives that express the continuity of mass and momentum

in reciprocal compliance with the exchanged boundary conditions (Abbott *et al.*, 1986a,b).

The result of these efforts, the SHE (Système Hydrologique Européen), now available in two versions, the MIKE-SHE of DHI (Refsgaard and Storm, 1995) and the SHETRAN of the University of Newcastle upon Tyne (Ewen *et al.*, 2000), is a powerful and interesting tool that has opened the interest to distributed modeling at the catchment scale.

With the recent advances in remote sensing, geographic information system, and computer technology, together with the natural evolution of research in hydrologic modeling, physically based distributed hydrologic models with simple and parsimonious parameterizations, such as WATFLOOD (Kouwen, 2000), DHSVM (Wigmosta *et al.*, 1994), TOPKAPI model (Todini, 1995b; Todini and Ciarapica, 2002; Liu and Todini, 2002), FEWS NET Stream flow Model (Verdin and Klaver, 2002), LISFLOOD (De Roo *et al.*, 1998, 2000), tRIBS (Vivoni, 2003), and InHM (van der Kwaak and Loague, 2001) were developed in recent years and represent interesting alternatives to MIKE-SHE and SHETRAN, which cannot yet be incorporated into real-time flood forecasting systems because of their computational time requirements.

Although interesting from a scientific point of view, not all the above-mentioned models are used in practice for real-time flood forecasting. Presently, only the TOPKAPI model can be considered fully operational, as it has been incorporated into the EFFORTS within the frame of project MUSIC and is operationally active on the Arno, the Reno, and nine additional smaller rivers in Italy. Another model, LISFLOOD FF is under development and test within the frame of project EFFS, but cannot yet be considered as an operational tool, while MIKE-SHE will be probably incorporated into a real-time flood forecasting system within the frame of a recently EC-funded project FLOODRELIEF.

REAL-TIME ADJUSTMENT OF RAINFALL-RUNOFF MODELS

The Different Types of Real-time Adjustment

The need for on-line adjustments in real-time flood forecasting models was underlined at the WMO Simulated Real-Time Intercomparison of Hydrological Models workshop (WMO, 1992), which was held at the University of British Columbia.

When real-time flood forecasting is performed using a nonstructural model, for instance, a "data-driven model", namely, a model for which the structure is not imposed *a priori* but identified from the observations, it is common practice to perform at the same time both the hydrological model and noise model identification, as well as the estimation of their relevant parameter values. As a matter of fact, real-time updating of state variables and parameter values is an essential feature that allows the models to more closely reproduce the observations. Recursive estimation algorithms are used that are generally based upon Kalman filter-type algorithms or other similar on-line estimation techniques. For instance, Lees *et al.* (1994) developed an adaptive flood warning scheme for the river Nith at Dumfries where they used a stochastic time-variable parameter (TVP) estimation algorithm (Young, 1984).

The problem is definitely more complex when using "structural models" (Kendall and Stuart, 1967), namely those models, such as for instance, the ESMA or the physically meaningful distributed models, for which a (generally complex) model structure has been hypothesized. In this case, apart from bluntly matching the latest model output to the most recent observations, there are basically three known procedures for adjusting forecasts in real time.

The first procedure is to correct the observed input to the model, given its large uncertainty (for instance, areal precipitation) and rerun the model until there is a reasonable match between model output and observed runoff (Bergström, 1976). Although several objections may be raised to this method, its rationale is to modify the state variables (the different water storages present in the model such as soil moisture storage content, groundwater content, etc.) in order to allow the hydrological model to be consistent with the observed output. Unfortunately, the method has a limited validity, which is restricted to reservoir type models.

The second procedure is to modify the parameters of the model by means of a real-time recalibration approach. This can be performed either limiting it to the transfer function relevant to the generally linear routing component of the model, or by recalibrating in real time all the model parameters, as proposed by Georgakakos for the SACRAMENTO model, by means of an Extended Kalman Filter (Georgakakos, 1986a,b). Following the results of the WMO (1992) Simulated Real-Time Intercomparison of Hydrological Models, this approach should not be recommended, not only because of the computational requirements of implementing such complex nonlinear Kalman Filters but chiefly because it often leads to highly unstable parameter estimates with a consequent increase in flood forecasting uncertainty.

The third procedure, based upon the development of a noise model, can be approached either by identifying the structural model parameter values and the noise model at the same time or by identifying the noise model conditionally to a given structural model and its parameter values. Owing to the complex nonlinear and/or threshold-type structures of the hydrological models and the parameter estimation difficulties expressed above, the conditional noise model approach is presently more widely used.

The noise model, conditional to the chosen model structure and parameter values, which implies the derivation

of the hydrological model residuals for a long historical record, can be achieved in several ways, ranging from ARMA and ANN models to Nearest Neighbour techniques (Brath *et al.*, 2002). More recently, Bayesian approaches have been increasingly used. They are based either upon the application of Kalman Filters, containing both the observed and the modeled flows in the state vector (EFFORTS, 2005, **http://www.progea.net**) or, as in the case of the hydrological processor introduced by Krzysztofowicz (1999), on a regression between observed and modeled flows after transforming them into a Standard Normal space using the Normal Quantile Transform (Van der Waerden, 1952, 1953; Kelly and Krzysztofowicz, 1997).

As an example of the effect of the noise model, Figure 4 shows graphically the reduction of uncertainty, obtainable by means of a two-dimensional Kalman Filter on the observed and modeled flows using the MISP algorithm (Todini, 1978). In Figure 4, the white histogram represents the forecasting standard error when the QPF from a NWP model is used for extrapolating the discharge forecast up to 12 h in advance at the Casalecchio gauging station on the Reno river (as further discussed in the Section on "Input uncertainty"), for which it is difficult to reliably provide more than 4–5 h in advance forecasts because of the relatively short catchment concentration time. One can see in Figure 4 (white histogram) that the standard error of the hydrological model remains unchanged around $50\,m^3\,s^{-1}$ up to 4 h of advance, while it reaches $100\,m^3\,s^{-1}$ when the forecast is issued 12 h in advance. The black histogram shows the reduction of the forecasting standard error when MISP is used. It is easy to verify, as one could expect, that the advantage provided by the noise model is much larger in the first few hours and tends to vanish with the increasing forecasting lead time.

The Kalman Filter

Given its importance in real-time updating of rainfall-runoff models, it was felt essential to provide in this section the basic elements of the Kalman Filter, while a more extensive derivation can be found in Jazwinski (1970) or Gelb (1974).

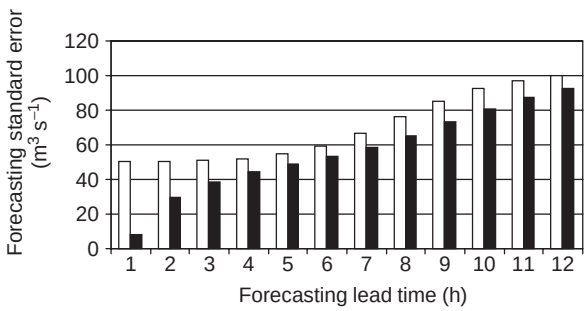


Figure 4 Reduction of forecasting uncertainty at Casalecchio on the river Reno using MISP up to 12 h in advance. Without MISP (white histogram), with MISP (black histogram)

The Kalman Filter (KF) (Kalman, 1960; Kalman and Bucy, 1961) is the recursive extension of the Wiener Filter (Wiener, 1949) to linear (or locally linearized in time) stationary as well as nonstationary processes. Its original derivation descends from the classical state-space formulation of dynamic systems in its time discretized form:

$$\mathbf{x}_t = \mathbf{\Phi}_{t-1,t}\mathbf{x}_{t-1} + \mathbf{\Gamma}_t\boldsymbol{\eta}_t$$

known as the "model" or "system" equation (1)

and

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \boldsymbol{\varepsilon}_t \quad \text{known as the "measurement" equation} \tag{2}$$

where $\mathbf{x}_t[n, 1]$ is the state vector, namely, the vector containing all the n state variables used to represent the dynamic system; $\mathbf{\Phi}_{t-1,t}[n, n]$ is the state transition matrix that may vary at each step in time; $\boldsymbol{\eta}_t[p \le n, 1]$ is an unknown random Gaussian time independent process, with mean $\overline{\boldsymbol{\eta}}_t$ and covariance matrix $\mathbf{Q}_t$, used to represent the model error, while matrix $\mathbf{\Gamma}_t[n, p]$ is an appropriate matrix to relate the dimensions; $\mathbf{z}_t[m \le n, 1]$ is the measurement vector, namely, the vector containing the m observations and matrix $\mathbf{H}_t[m, n]$ is an appropriate matrix to relate the dimensions; $\boldsymbol{\varepsilon}_t[m, 1]$ is the measurement error, represented as a random Gaussian time independent process with mean $\overline{\boldsymbol{\varepsilon}}_t$ and covariance matrix $\mathbf{R}_t$, which is taken as being also independent from the model error $\boldsymbol{\eta}_t$.

For the sake of simplicity, following the original derivation by Kalman (1960) the "control" term has been omitted from the model equation, while a measurement error term has been added. The Kalman Filter aims at finding $\hat{\mathbf{x}}_{t|t}$, the unbiased minimum variance estimate of the unknown state $\mathbf{x}_t$, together with its error covariance matrix $\mathbf{P}_{t|t}$, conditional to the knowledge of an unbiased *a priori* state estimate $\hat{\mathbf{x}}_{t|t-1}$, together with its covariance matrix $\mathbf{P}_{t|t-1}$ (which fully represent the stochastic process because of the hypothesis of Gaussian errors), and the latest noise-corrupted measurements $\mathbf{z}_t$ together with its measurement error statistics (again mean and covariance are sufficient due to the Gaussian hypothesis). This estimate is obtained using the following equations, a neat derivation of which can be found in Gelb (1974).

At each step in time, the estimates of the state and of the error covariance are extrapolated from the previous step:

$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{\Phi}_{t|t-1}\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{\Gamma}_t\overline{\boldsymbol{\eta}}_t$$

state extrapolation (3)

$$\mathbf{P}_{t|t-1} = \mathbf{\Phi}_{t|t-1}\mathbf{P}_{t|t-1}\mathbf{\Phi}_{t|t-1}^T + \mathbf{\Gamma}_t\mathbf{Q}_t\mathbf{\Gamma}_t^T$$

error covariance extrapolation (4)

Then the following quantities are estimated:

$$\boldsymbol{v}_t = \mathbf{z}_t - \overline{\boldsymbol{\varepsilon}}_t - \mathbf{H}_t \hat{\mathbf{x}}_{t|t-1}$$

known as the *Innovation* (5)

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^T (\mathbf{H}_t \mathbf{P}_{t|t-1}\mathbf{H}_t^T + \mathbf{R}_t)^{-1}$$

known as the *Kalman Gain* (6)

Finally the *a priori* estimates can be updated to include the latest measurements:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t \boldsymbol{v}_t$$

state update (7)

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t)\mathbf{P}_{t|t-1}$$

error covariance update (8)

With respect to the Wiener Filter, the advantage of the Kalman Filter lies in the fact that optimality (in terms of unbiasedness and minimum variance) is recursively imposed at each step in time as opposed to the batch form required by the Wiener Filter. This also allows the nonlinear process dynamics to be locally linearized in time to produce what is known as the *Extended Kalman Filter* (*EKF*) (Jazwinski, 1970).

The basic problem in applying the KF is that the optimality conditions only hold when the state transition matrix $\boldsymbol{\Phi}_{t-1,t}$, together with the model and measurement error statistics ($\overline{\boldsymbol{\eta}}_t$, $\mathbf{Q}_t$, $\overline{\boldsymbol{\varepsilon}}_t$, $\mathbf{R}_t$) are fully known. Mehra (1970) provided a solution to the problem of estimating the unknown error statistics by imposing the time independence of the innovation, a condition associated with the KF optimality. The estimation of the state transition matrix parameters (generally known as the *hyper-parameters*) proved to be a far more complex problem. Several approaches can be found in the literature for solving the nonlinear estimation problem originated by the simultaneous estimation of both state and parameter values, ranging from developing the KF in the parameter space (Mayne, 1965; Sage and Husa, 1969) to the use of the Extended Kalman Filter on a state vector enlarged with the parameters (Jazwinski, 1970) or from the use of Maximum Likelihood (ML) (Gupta and Mehra, 1974; Cooper and Wood, 1982a,b; Wood and O'Connell, 1985) to the method of moments (Wojcik, 1993) and to full Bayesian approaches (Mantovan *et al.*, 1999). Following the Instrumental Variables (IV) approach (Kendall and Stuart, 1967; Young, 1974; Young and Whitehead, 1977), Todini realized that the posterior state estimate $\hat{\mathbf{x}}_{t|t}$, is the best possible IV, being totally independent from measurement noise and a minimum variance estimator of the true unknown state due to optimality of the KF. Accordingly, he developed the Mutually Interactive State Parameter (MISP) estimation technique by using two mutually conditional KFs: one in the space of the state conditional to the previous step parameter estimates and one in the space of the parameters conditional to the previous and the last updated state estimates (Todini, 1978). MISP was recently found by Mantovan *et al.* (1999) to be superior to the method of moments and very close to ML but much less demanding in computer time, while a full Bayesian approach, requiring the use of the Gibbs sampler (Gelfand and Smith, 1990; Carter and Khon, 1994) to produce the posterior distributions, had to be abandoned because of its exaggerated computer time requirements.

Several examples of application of Kalman Filters in hydraulic and hydrological applications can be found in Chao-Lin (1978), while specific applications to rainfall-runoff models can be found in Hino (1973), Szollosy-Nagy (1976) for on-line estimation of linear transfer function models, in Todini and Wallis (1978) for on-line state and parameter estimation of a threshold-type multiple input single output ARX model, and in Georgakakos (1986a,b) for the on-line update of the Sacramento model parameters using an EKF.

Recent developments in Kalman Filtering aim at overcoming the EKF limitations because of error Gaussianity requirements and model equation linearization. These new developments include the Particle Filter (PF) (Salmond *et al.*, 1993), the Ensemble Kalman Filter (EnKF) (Evensen, 1994, 2003) presently used for assimilating data in large nonlinear ocean and atmospheric models, and its simplification known as the *Unscented Kalman Filter* (*UKF*) (Julier and Uhlmann, 1997; Wan and van der Merwe, 2000).

Before concluding this section, it is worthwhile pointing out that the KF being a "filter" is not a "predictor". Therefore, it is an excellent tool for assimilating observations and calibrating state and parameter values, but rapidly becomes very poor when used for "forecasting" at future times, when observations are no longer available. In addition, it cannot be directly used for assessing what will be defined as "forecasting uncertainty", unless a "pseudo measurement" is available at future times, as described in the Section on "The definition of forecasting uncertainty". Unfortunately, these problems have been overlooked for many years in real-time flood forecasting.

ASSESSMENT OF FORECASTING UNCERTAINTY

The Definition of Forecasting Uncertainty

As pointed out in the introduction, one of the most relevant aspects of real-time flood forecasting is to allow better decision making under uncertainty. Krzysztofowicz, (1999) points out that "*Rational decision making (for flood warning, navigation, or reservoir systems) requires that the total uncertainty about a hydrologic predictand (such as river*

stage, discharge, or runoff volume) be quantified in terms of a probability distribution, conditional on all available information and knowledge" and that *"Hydrologic knowledge is typically embodied in a deterministic catchment model*."

These statements very clearly underline two aspects usually not clearly understood.

The first one is that the objective of forecasting is to provide the description of *the uncertainty of the future values of water stage, discharge, runoff volume, and so on, and not the uncertainty of predictions generated by the hydrological forecasting models.*

The second one is that this uncertainty, generally expressed in terms of *probability density* (or probability distribution) *function, must be "conditional" upon the forecasting model prediction, which is now seen as the available, although uncertain, knowledge of the future*. In other words, the forecasting model prediction is now taken as an *error affected pseudo measurement* of future values of water stage, discharge, runoff volume, and so on.

If, at a given river cross section of interest, variable y_t represents the observations of our "hydrologic predictand" (water stage, discharge, runoff volume, etc.) while $\hat{y}_t$ is the corresponding hydrological model output, the scope of uncertainty assessment can be defined as the estimation of the following conditional probability density:

$$f\{y_{t^*+i\Delta t} \,|\hat{\mathbf{y}}_{t^*+i\Delta t}, \mathbf{y}_{t^*}\} \tag{9}$$

where $y_{t^*+i\Delta t}$ represents the unknown future value of the predictand some $i\Delta t$ time steps after the present time, t^*; $\mathbf{y}_{t^*}$ represents the vector of all available observations of the predictand up to the present time, while $\hat{\mathbf{y}}_{t^*+i\Delta t}$ represents the vector of all the hydrological model forecasts available up to the required forecasting horizon.

Unfortunately, instead of the conditional probability density given in equation (9), it is common practice to provide the following probability density:

$$f\left\{\hat{y}_{t^*+i\Delta t} \,\middle|\hat{\mathbf{y}}_{t^*+(i-1)\Delta t}, \mathbf{y}_{t^*}\right\} \tag{10}$$

namely, the uncertainty of our forecast, which is not the proper uncertainty measure to be used in the decision making process.

Going back to the preference given by many hydrologists to the use of structural rainfall-runoff models (ESMA or physically meaningful distributed models), it must be noted that when the observations end, and the models are run in predictive mode, the nonstructural models will forecast future values by projecting (extrapolating) in the future the latest state estimate. Given that there are no more measurements to be used to "filter" these extrapolations (namely to reduce their variance) at each step in time, they will tend more or less rapidly to degrade as a function of the state transition probabilities. On the contrary, if one uses the structural modeling approach combined with a Bayesian processor, one can initially extend the structural model predictions beyond the measured record because of the concentration time of the catchment and/or to the availability of QPF. The structural model predictions can thus be taken as uncertain measures of the flows that are expected to occur and used to derive their probability density, or at least their expected value and variance, conditional upon the model predictions, by means of Bayesian or Kalman filtering techniques, which inevitably leads to a reduction of forecasting uncertainty.

Unfortunately, most of the existing operational real-time flood forecasting systems do not incorporate such a measure of uncertainty and the few operational real-time flood forecasting systems that provide, together with the forecast, a measure of uncertainty are essentially based upon a stationary measure of uncertainty derived from the analysis of performances on past historical data. This stationary measure is sometimes updated in real-time by means of a Kalman filter (see the Section on "The Kalman filter"), or similar recursive tools, in order to incorporate some information on the most recent performances but the final result is not always satisfactory.

With respect to the sources of uncertainty, as described by (Krzysztofowicz, 1999), they can be categorized as operational, hydrologic, and input. Excluding operational uncertainty, such as the one caused by erroneous or missing data, human processing errors, unpredictable interventions and so on, total uncertainty in real-time flood forecasting, which deals with forecasting horizons ranging from a few hours to a few days in advance, can be schematized as the combination of (i) hydrological uncertainty and (ii) input uncertainty.

Hydrological Uncertainty

Uncertainties definitely arise from imperfections of the hydrological model: its structure and relations, incorrect values of model parameters, incorrect estimates of other inputs taken as deterministic, errors in measurements of physical quantities, and so on. All these uncertainties are incorporated into what is referred to as the hydrological uncertainty.

Extensive research has been carried out on parameter estimation and on the assessment of parameter uncertainty. Examples are the Generalised Likelihood Uncertainty Estimation (GLUE) due to Beven and Binley (1992), NLFIT by Kuczera (1994), the more recent Bayesian Recursive Estimation (BaRE) algorithm introduced by Thiemann *et al.* (2001), or the Shuffled Complex Evolution Metropolis (SCEM-UA) by Vrugt *et al.* (2003).

Unfortunately, these approaches, although useful tools for estimating parameters and describing parameter uncertainty given a specific model, do not entirely represent the forecasting uncertainty since they neglect or treat implicitly

many other sources of uncertainty, such as the measurement and the areal precipitation estimate uncertainty or the uncertainty due to the choice of the model. Moreover, marginalizing with respect to all the possible sources seems presently an overwhelming task that makes the approach hardly usable to assess the *predictive uncertainty* of the predictand of interest (described by the probability density of equation 9). When observations are no more available, the only practical possibility to describe the predictive uncertainty is to use the presently selected model with a fixed set of parameter values (a given value or the last updated one) as well as all the other uncertainty sources, except the rainfall forecast, taken as deterministic. In other words, the forecast will necessarily be *conditional upon the chosen model structure, the parameter values as well as to all other error sources except the precipitation forecasts*. This conditional uncertainty, the *hydrological uncertainty*, can thus be more simply ascertained by using the chosen model with fixed parameter values and by comparing the observed and predicted values in the historical record using the observed rainfall as input. Krzysztofowicz (1999), Krzysztofowicz and Herr (2001) provide a technique based upon a Bayesian combination of a Markov model of water levels with a regression model relating the future levels with the predicted as well as the last observed ones. The approach, which is applied to data transformed in a Gaussian space via Normal Quantile Transform (Van der Waerden, 1952, 1953; Kelly and Krzysztofowicz, 1997), allows in the end to generate the conditional density of equation (9).

After assessing the *input uncertainty*, it is then possible to obtain the predictive uncertainty by combining in Bayesian sense, the input, and the hydrological uncertainty (Krzysztofowicz, 2001).

Input Uncertainty

Input uncertainty encompasses all exogenous variables and internal states (initial conditions) whose values vary from one forecast time to the next; they exclude parameters whose values remain fixed for a given river basin. Input uncertainty includes the uncertainty of the inputs causing the most significant effects on the forecasts (within the relatively short forecasting horizon), such as the future rainfall forecasts, while disregarding inputs generating minor effects over the forecasting horizon, such as evapotranspiration, the effect of which is appreciable in terms of its integral over longer periods of time.

Kelly and Krzysztofowicz (2000) provide a simple probabilistic approach based upon a two-component Weibull distribution in order to account for the probability of no future rainfall as an alternative to Monte Carlo approaches (Lardet and Obled, 1994). Nonetheless, a large amount of research work is still needed, taking into account the quality of the presently available quantitative precipitation forecasts (QPF) arising from Nowcasting or from numerical weather prediction (NWP) models.

In their work, Kelly and Krzysztofowicz (2000) and Krzysztofowicz and Herr (2001) assume that a probabilistic quantitative precipitation forecast (PQPF) is available and that it is actually representative of the precipitation uncertainty.

Unfortunately, because of several reasons among which the fact that the amount of precipitated water is not a state variable in NWP models (with the consequence that they cannot be updated in the models by means of precipitation observations) the QPFs have not yet reached the accuracy needed for issuing reliable flood forecasts when directly introduced as inputs to rainfall-runoff models. Moreover, as it was clearly shown in project EFFS the PQPFs originating from the NWP ensemble forecasts tend to be highly biased, therefore providing a very poor and distorted representation of future precipitation input uncertainty to be used within the proposed Bayesian processor.

Research is presently concentrated at finding more realistic alternatives to the direct use of NWP ensembles as inputs to flood forecasting models (see Section on "The HEPEX Project").

As an alternative to the direct coupling proposed by Georgakakos (1989), within the frame of the EU Funded Flood Forecasting Research Project AFORISM project (A Comprehensive Forecasting System for Flood Risk Mitigation and Control) (Todini, 1995a), an approach for linking a Limited Area Meteorological models (LAMs) to real-time flood forecasting models was introduced, which was then set up and tested under the EU funded TELFLOOD Project (Todini and Cerlini, 1999). The technique uses a Bayesian combination of an ensemble of past measured precipitation traces, conditional to the latest ground measurements of rainfall, derived with a phase-space approach known as the *Nearest Neighbour technique* (Yakowitz, 1987; Yakowitz and Karlsson, 1987; Todini, 1999), with the LAM forecasts that are taken as biased and noise-corrupted measures of future precipitation. The procedure was tested on nine flood events observed on the Reno river at Casalecchio (1051 km^2), which were chosen as to represent various rain and flood typologies.

The QPF forecasts from 1 to 24 h in advance were used as input to a rainfall-runoff model of the Reno river and the computed bias and standard deviation of errors of forecast are shown in Figure 5.

Using the Nearest Neighbours technique, 50 rainfall series of 24 h were drawn from the past observed values. Again, the bias and standard deviation of errors of forecast were computed and are shown in Figure 6.

From Figure 6, it is possible to notice that after a few hours with virtually no bias and a more or less constant standard deviation of errors of about 50 $m^3 s^{-1}$, which measures the uncertainty due to the rainfall-runoff model, the standard deviation rapidly reaches the value of 160 $m^3 s^{-1}$. Although of poor quality, the synthetically

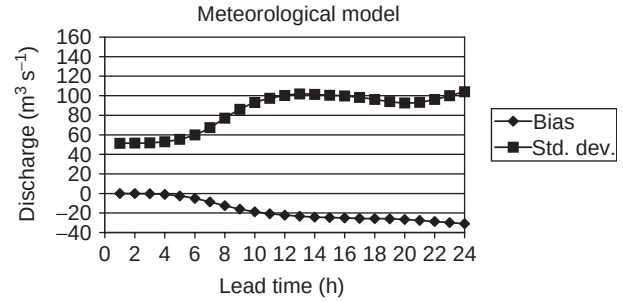


Figure 5 Bias and standard deviation of discharge forecasting errors for the Reno river at Casalecchio using LAM QPFs

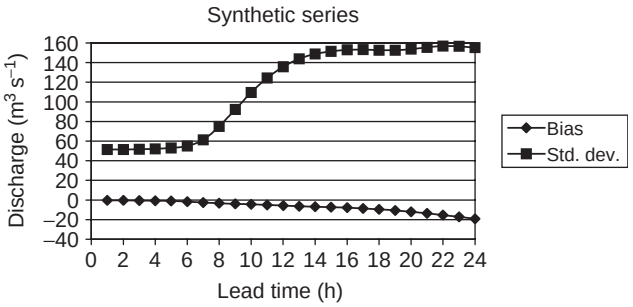


Figure 6 Bias and standard deviation of discharge forecasting errors for the Reno river at Casalecchio using the Nearest Neighbour rainfall forecasted scenario

generated scenarios, which can be considered as our "*a priori* best guess", can now be corrected using the QPF provided by the NWP model, taken as an "imperfect measure" of future rainfall.

Following the Bayesian approach described in Berger (1980) under the assumption of Gaussian errors, a correction scheme, typical of a Kalman filter (see Section "The Kalman filter") can be set up.

Given z_1 and z_2, two different measures of the unknown quantity x both affected by errors, namely:

$$z_1 = x + \varepsilon_1 \qquad z_2 = x + \varepsilon_2 \tag{11}$$

where: $\varepsilon_1 \cong N\{\mu_1, \sigma_1^2\}$, $\varepsilon_2 \cong N\{\mu_2, \sigma_2^2\}$ are mutually independent, namely:

$$E\{(\varepsilon_1 - \mu_1)(\varepsilon_2 - \mu_2)\} = 0 \tag{12}$$

A Bayesian combination of the two measures starts by defining a posterior estimate $\hat{x}$, as a linear combination of the two measures (minus their respective measurement bias):

$$\hat{x} = k_1(z_1 - \mu_1) + k_2(z_2 - \mu_2) \tag{13}$$

The two weights, k_1 and k_2 can be found by imposing that $\hat{x}$ must be unbiased and minimum variance, which implies that $\tilde{x} = \hat{x} - x$, the estimation error, must fulfill the following requirements: $E\{\tilde{x}\} = 0$ and $E\{\tilde{x}^2\} = Min$.

The first condition gives: $k = k_2 = 1 - k_1$, while the minimization of the error variance, provides:

$$k = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{14}$$

This coefficient, also known as the *Kalman gain*, is then used in equation (1) to give:

$$\begin{aligned}\hat{x} &= \left(1 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)(z_1 - \mu_1) + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(z_2 - \mu_2) \\ &= \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}(z_1 - \mu_1) + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(z_2 - \mu_2)\end{aligned} \tag{15}$$

As can be easily verified, the posterior estimate $\hat{x}$ is unbiased and its variance:

$$\begin{aligned}E\{\tilde{x}^2\} &= \left(1 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2 \sigma_1^2 + \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2 \sigma_2^2 \\ &= \left(\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^2 \sigma_1^2 + \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2 \sigma_2^2 \\ &= \sigma_1^2\sigma_2^2 \frac{\sigma_1^2 + \sigma_2^2}{(\sigma_1^2 + \sigma_2^2)^2} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}\end{aligned} \tag{16}$$

is definitely smaller than the single variances σ_1^2 and σ_2^2 of the two original measures.

By taking each member of the discharge ensemble generated by the Nearest Neighbor approach as measurement z_1 and the discharge obtained by using the QPF provided by the NWP model as measurement z_2, the Bayesian approach can be used after estimating μ_2 and σ_2 by means of historical data.

The results of the experiment on the Reno river were quite encouraging. Figure 7 shows that the bias, present

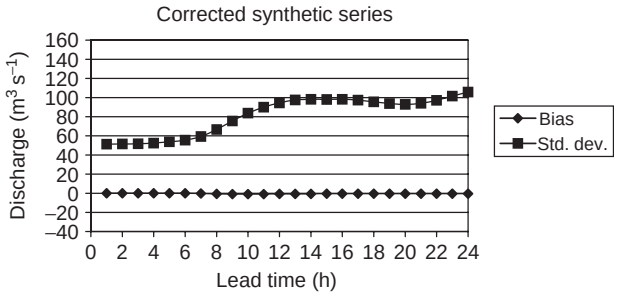


Figure 7 Bias and standard deviation of discharge forecasting errors for the Reno river at Casalecchio after the Bayesian correction

in the generated precipitation traces, has been totally eliminated, while the standard deviation of errors is being reduced at lags larger than three hours, while no improvement is obtained at lags up to three hours because of the relatively better quality of the *a priori* Nearest Neighbor forecasts on a short-term basis.

These encouraging results, together with the results obtained by Obled *et al.* (2002) and Obled and Djerboua (2001) using meteorological pressure fields for sorting the analogues, can be viewed as the starting point for estimating an improved input uncertainty, to be used within Bayesian integration techniques, such as the one proposed by Krzysztofowicz (1999, 2001), as described in the Section on "Examples of real-time flood forecasting systems". This approach seems realistically possible given the large amount of available weather reanalyses produced by most of the Meteorological Forecasting Centers, such as he NCEP-NCAR (Kalnay *et al.*, 1996) or the ECMWF.

Communicating Forecasting Uncertainty to the Users

A major problem associated with the use of predictive probability densities, instead of deterministic forecasts, is how to communicate understandable information to the stakeholders (the authorities responsible for taking decisions in real-time) as well as to the general public.

At present, there is ample evidence that the performance of several flood warning systems is rather poor. The information that they are designed to disseminate fails to reach a large part of the target audience, and the people who make up this audience are not satisfied with the service that they receive. This poor performance is a function of a weak link in the chain that connects the flood forecast with the person who is designated to obtain this information: the flood manager and particularly the public at large risk from the flooding that is being forecasted. Without comparable strength in this dissemination link in the chain, the performance of the whole system is degraded. The advancement of the accuracy of flood forecasting, and better targeting of the dissemination of ensuing message, is not for the benefit of the forecasters but for the public (Penning-Rowsell and Tunstall, 1997). As pointed out by several authors (Penning-Rowsell and Fordham, 1994; Handmer and Penning-Rowsell, 1990; Parker and Fordham, 1996), a flood warning system can be conceptualized as the combination of a flood forecasting subsystem (the technological part) and a flood warning dissemination subsystem (the social part), interlinked by a feedback arrangement. The messages that are generated by the technological part of the system, also and hopefully in terms of predictive probabilities, must be converted into simple and understandable messages, not necessarily the same, to the flood managers and to the public. While extensive research is still needed in this context in collaboration with sociologists and flood managers, a number of potential approaches are delineated in the sequel.

The Bayesian Approach

The more natural approach to incorporate flood forecasting uncertainty into a simple decision scheme, as evaluated in the MUSIC project, is the classical Bayesian utility function approach.

Traditionally, flood alert thresholds are computed in a preliminary planning phase and compared to water stage or discharge forecasts taken as "deterministic", implicitly assuming that they are not affected by forecasting errors. This is a reasonably good approach when the forecast is extremely reliable. Nevertheless, when the forecast is highly uncertain, particularly when using QPF to extend the forecasting horizon, the approach may give rise to too many false alarms as well as missed alarms. It is therefore reasonable to incorporate the forecasting uncertainty into the decision scheme.

With reference to the flood warning problem of Figure 1, the following utility functions shown in Figure 8 can be defined to represent (i) the perception of losses that could occur if a warning is not appropriately issued $g_{na}(y)$; (ii) the cost of implementation of the measures taken as a consequence of an alert, plus the losses, in the case of dykes being overtopped, assumed smaller than in case of no alert $g_a(y)$. By means of the probability density of future water stages (or discharges) conditional on the hydrological model forecasts given by equation (9), it is possible to estimate at each step in time the expected value of both costs and choose to issue an alert if, within an established horizon, the inequality of equation (17) holds.

$$\int_0^{+\infty} g_a(y_{t^*+i\Delta t})\, f\{y_{t^*+i\Delta t}|\hat{\mathbf{y}}_{t^*+i\Delta t}, \mathbf{y}_{t^*}\}\, \mathrm{d}y_{t^*+i\Delta t} \leq \int_0^{+\infty} g_{na}(y_{t^*+i\Delta t})\, f\{y_{t^*+i\Delta t}|\hat{\mathbf{y}}_{t^*+i\Delta t}, \mathbf{y}_{t^*}\}\, \mathrm{d}y_{t^*+i\Delta t} \tag{17}$$

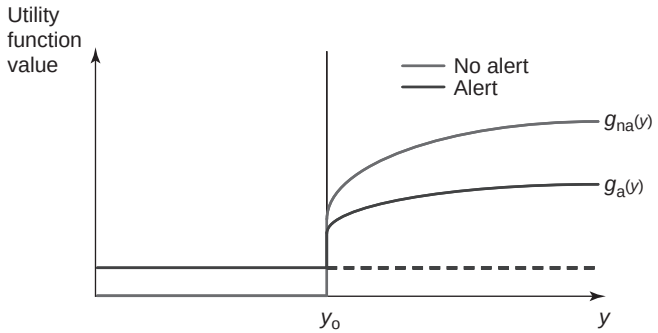


Figure 8 Utility functions for the dyke overtopping decision scheme. y_0 water stage (or discharge) at which overtopping occurs; $g_{na}(y)$ losses in case of no alert; $g_a(y)$ losses in case of alert. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

In this approach, it is not necessary to develop extremely correct estimates of the different cost utility functions. The technique is more sensitive to the actual overtopping thresholds (that can be established reasonably well on the basis of hydraulic analyses) than to the shape of the cost functions, that can be adjusted in order to maximizing the desired success rate by simulating the decision mechanism over the historical record.

The Alert Decision Support Matrix

The major aims in communicating flood forecasting to the stakeholders are clarity and simplicity. People in charge of decisions during flood events are not necessarily technical people; therefore efforts have to be made to convert probability densities into meaningful decision tables or support matrices. Within the frame of project MUSIC, the University of Newcastle upon Tyne developed a framework for converting flood forecasting uncertainty into alert decision support matrices similar to the one shown in the following Table 1.

The proposed methodology requires the preliminary definition, during a planning phase, of a number of alert levels (blue, yellow and red) corresponding to bankfull (blue), inundation of property (yellow), and potential loss of life (red). The decision to take (or not to take) action is derived as a function of (i) the alert level, (ii) the likelihood of the event, and (iii) the reliability of the forecast. The likelihood of the event is defined as a function of the probability of exceedence (Very low, Low, Medium, High, Very high) of each specific alert level, while the reliability (Very low, Low, Medium, High, Very high) of the forecast is defined in terms of its coefficient of variation (CV).

The rationale for this approach is that on the one hand it becomes more essential to take the right decision when the Red level is reached and that on the other hand one has to be more careful when the uncertainty of the forecast is low. The results are synthesized into three simple decision support matrices (one for each alert level) similar to the ones shown below, where A stands for "action" and NA for "no action".

The Development of Conditional Precipitation Thresholds for Early Flood Warning

Another aspect, currently under investigation, is the possibility of deriving precipitation thresholds, conditional on soil moisture, to be used as a simple basis for early warning. Today, in several countries early flood warning is based on unconditional precipitation thresholds such as: if 50 mm of rainfall is forecasted for the next 24 h, a warning must be issued. It is obvious that the effects of 50 mm of rain in 24 h are quite different if they fall over a dry or a saturated catchment. Bearing in mind that the weak link in real-time flood forecasting systems is not the technical component but rather the communication to the stakeholders, it would certainly be useful to provide them with extremely simple but sound and robust procedures based on rainfall thresholds, conditional to the initial soil moisture status; these thresholds can in fact be easily understood and used by nontechnical people.

In order to incorporate this information into a simple decision support scheme for flood warning authorities, extensive Monte Carlo simulations (10 000 years) are needed using synthetic rainfall generators and rainfall-runoff models.

Table 1 An example of flood decision support matrices

	Likelihood					
Reliability	Blue	Very low	Low	Medium	High	Very high
	Very low	NA	NA	NA	NA	NA
	Low	NA	NA	NA	NA	NA
	Medium	NA	NA	NA	NA	A
	High	NA	NA	A	A	A
	Very high	NA	NA	A	A	A
	Likelihood					
Reliability	Yellow	Very low	Low	Medium	High	Very high
	Very low	NA	NA	NA	NA	NA
	Low	NA	NA	NA	NA	NA
	Medium	NA	NA	A	A	A
	H	NA	A	A	A	A
	Very high	NA	A	A	A	A
	Likelihood					
Reliability	Red	Very low	Low	Medium	High	Very high
	Very low	NA	NA	NA	NA	NA
	Low	NA	NA	NA	A	A
	Medium	NA	A	A	A	A
	High	A	A	A	A	A
	Very high	A	A	A	A	A

The basic idea is in fact to define a critical level threshold at a specific cross section of a river: for instance, the level can be of 0.50 m (or 1.00 m to be conservative) below the top of the levee. From that level, one can estimate the corresponding maximum flood discharge and, via stochastic simulation, it is possible to derive, for a number of alternative soil humidity classes, cumulated rainfall threshold values on increasing length time horizons. Only if the forecasted rainfall cumulated in time overtops one of the corresponding threshold values a warning is then issued. A first attempt was produced by DIIAR (2002) for the case of the Arno river in Italy, by using a lumped rainfall-runoff model based on the USDA-CN concept.

Libralon (2002) developed a decision theory oriented approach by combining a Neyman-Scott stochastic model to the lumped version of TOPKAPI. From the joint probability densities of discharge and rainfall conditional to the initial status of the soil moisture, two probability densities were derived. The first one is the density of rainfall that does not produce a discharge (or a water level) overtopping the threshold, while the second one is the density of rainfall that does produce a discharge (or a water level) overtopping the threshold (Figure 9).

It is then possible to determine a rainfall threshold by maximizing the probability of issuing correct forecasts, or, better, by minimizing a Bayesian utility function as described in the Section on "The Bayesian approach". This becomes clear when a threshold value is introduced in Figure 10: the two densities give rise to the four probabilities, namely the probability of issuing a true alert (TT), of issuing a false alert (TF), of not issuing a true alert (FT), and of not issuing a false alert (FF). The rainfall threshold can be defined as the one that maximizes the two probabilities of success, namely TT and FF. In the work

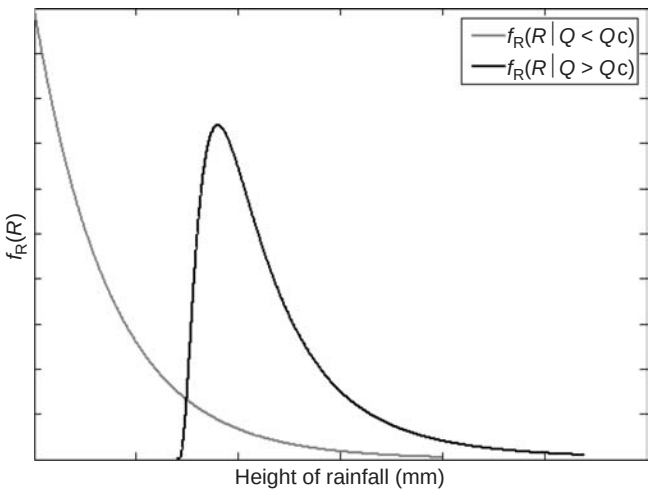


Figure 9 The probability densities of cumulated rainfall conditional to the overtopping of a discharge (or water stage) threshold

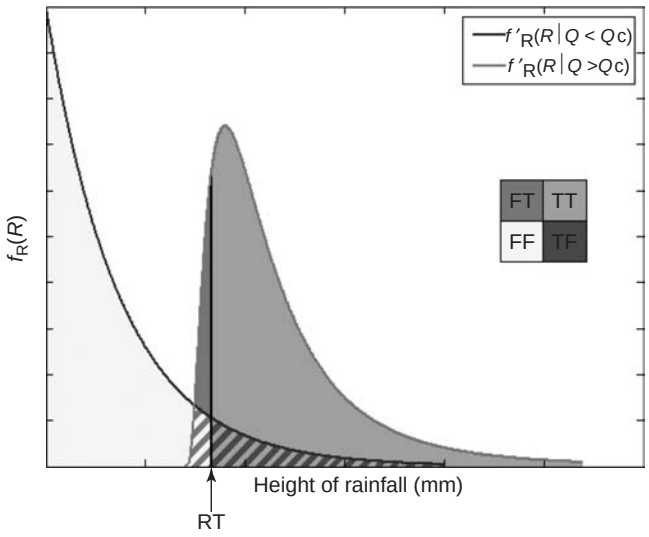


Figure 10 The four probabilities of alert: TT correctly issued alert; FF correctly nonissued alert; FT missed alert; TF wrongly issued alert. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

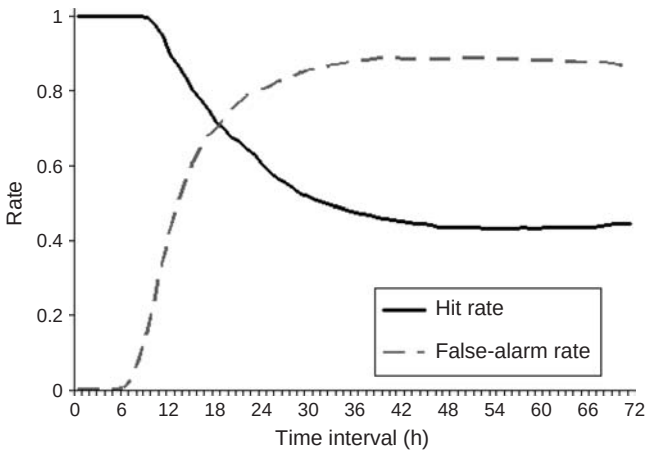


Figure 11 The Hit Rate (percent of successfully forecasted threshold overtopping) and the False Alarm Rate (percent of false alerts) for the Sieve river, a tributary of the Arno river, at Fornacina

of Libralon, the threshold was more generally defined as the one that minimizes the expected value of a Bayesian utility function that will trade off between the reduction of damages because of successful warnings and the actual cost of actions descending from activating the warning and a psychological cost expressing the loss of credibility generated by a large number of false alarms. The results, so far obtained on the assumption of error free rainfall forecasts, show a very high rate of success. Figure 11 shows the results obtained for the Sieve river, a steep and rather small tributary of the Arno river (ca. 800 km^2 of catchment area) on which it is very difficult to issue a reliable forecast for

more than 3–4 h in advance. As can be seen, the hit rate remains practically equal to 100% up to 9 h in advance, with only 10% of false alarms.

The present limitation in the use of the rainfall thresholds approach lies in the implicit assumption that the forecasted rainfall totals are not affected by errors. Current research is thus investigating the possibility of incorporating the rainfall forecasting uncertainty in the estimation of the expected damage curves to be minimized.

EXAMPLES OF REAL-TIME FLOOD FORECASTING SYSTEMS

The Upper Po River Flood Forecasting System

The first example relates to the Piemonte Region operational flood forecasting system developed for the upper Po catchment closed at the exit of the Piedmont Region (Figure 12).

The system was developed on the basis of MIKE FLOODWATCH (DHI, 2000a), which couples the NAM hydrological model (Nielsen and Hansen, 1973; DHI, 2000c) with MIKE 11 FF (DHI, 2000b) into an operational GIS-based flood forecasting system (Figure 13).

On the basis of approximately 300 rainfall real-time reporting raingauges, in order to catch the spatial variability of the meteorological forcing, the upper Po catchment was divided into 187 subcatchments with an average area of 200 km^2, for each of which a NAM model was calibrated.

Forecasts at several river sections are issued in a qualitative mode up to 48 h ahead using the hydrological model coupled with QPF provided by a limited area meteorological model and, in a quantitative management mode up to 6–12 h in advance using the observed rainfall and water levels.

Figure 14 shows the results of such operational forecasts at the cross section of Torino (catchment area of 5362 km^2 and concentration time of ca. 24 h) during the flood event of October 13–16, 2000. Five days of heavy rainfall hit the northern part of the Piemonte region with an average

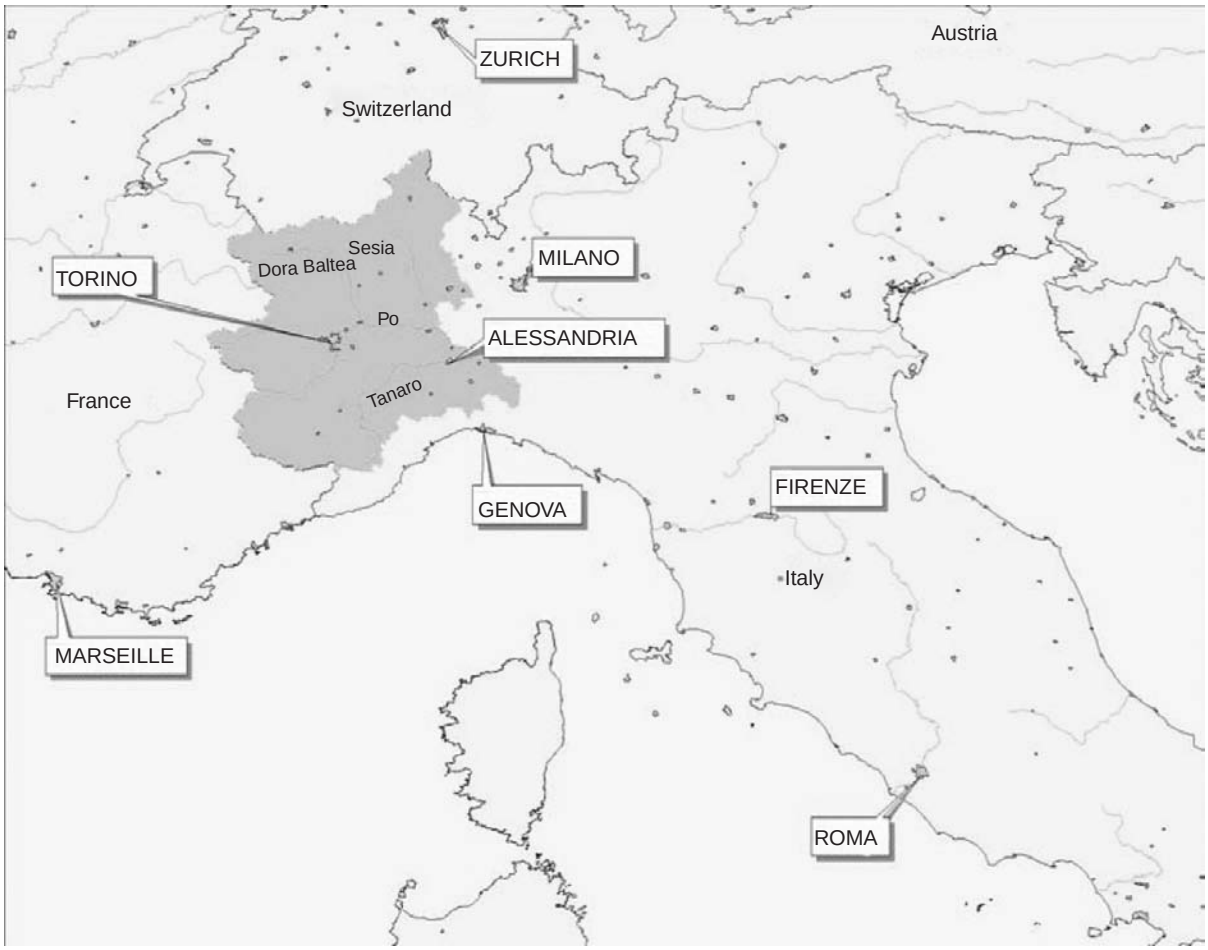


Figure 12 The Piemonte region in Italy (Courtesy of ARPA Piemonte). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

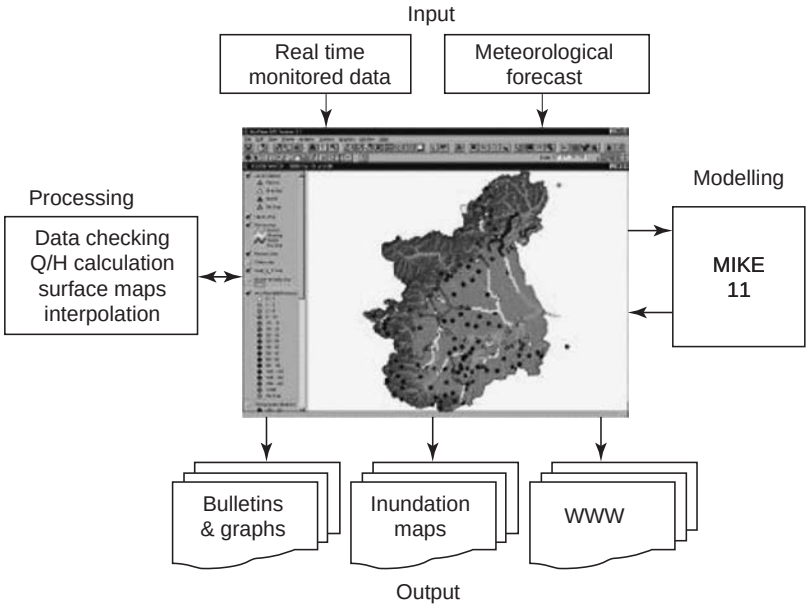


Figure 13 Schematic representation of MIKE FLOODWATCH applied to the Po river in the Piemonte Region (Courtesy of DHI Water & Environment). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

total rainfall larger than 300 mm, with peaks of 700 mm, producing large flood waves in most of the river network (Rabuffetti and Barbero, 2005).

It can be observed that forecasts up to 12 h in advance are quite reliable, particularly in the rising limb, and can be successfully used to trigger the warning level (grey) and the alarm level (black).

The Reno River Flood Forecasting System

The second example relates to the Reno river flood forecasting system operational at the Emilia-Romagna Hydro-Meteorological Service. The system, developed using the EFFORTS package within the frame of the EU funded project MUSIC, is relevant to the Reno river whose drainage basin has a total surface area of 4930 km^2, more than half of which pertains to the mountain basin.

Figure 15 shows the catchment and its subcatchments, while the available observation network is summarized in the following Table 2.

As shown in Table 2, in addition to the availability of a meteorological RADAR, METEOSAT-based rainfall estimates are available through the Institute of Atmospheric Sciences and Climate (ISAC-CNR). As described in MUSIC Deliverable 9.1 (**`http://www.geomin.unibo.it/orgv/hydro/music/`**), the distributed version of TOPKAPI (Todini, 1995b; Todini and Ciarapica, 2002; Liu and Todini, 2002) was the hydrological model used for the Reno catchment.

The model was developed at 1 × 1 km scale on the basis of three available maps: a digital elevation map (Figure 16), a map of soil types (Figure 17 and Table 3), and a land use map (Figure 18 and Table 4).

Table 2 The different measurement systems available on the Reno river

Measurement station type	Numerosity
Telemetering thermometers	15
Telemetering raingauges	47
Telemetering stage gauges	28
Rating curves	12
Radar	1
Satellite	1

As shown in Figure 19 for the upper Reno closed at the gauging station of Casalecchio di Reno, the operational forecasting system, which runs around the clock, is continuously fed not only by all the telemetering gauges (point measurements) but also by the RADAR measurements (fine mesh) as well as by the METEOSAT-based rainfall estimates (coarser mesh). An innovative Block–Kriging/Bayesian Combination of the three measurement systems aimed at producing an unbiased minimum variance estimator for rainfall (Todini, 2001; Mazzetti and Todini, 2002) was developed within the frame of project MUSIC and integrated into the EFFORTS system. At each time step, the system checks the availability of the different measurements and integrates the ones that are available to produce an estimate of rainfall over each model pixel.

In order to extrapolate flood forecasts in the future, the system uses the QPF generated by NWP whenever available, while a combination of the meteorological forecast

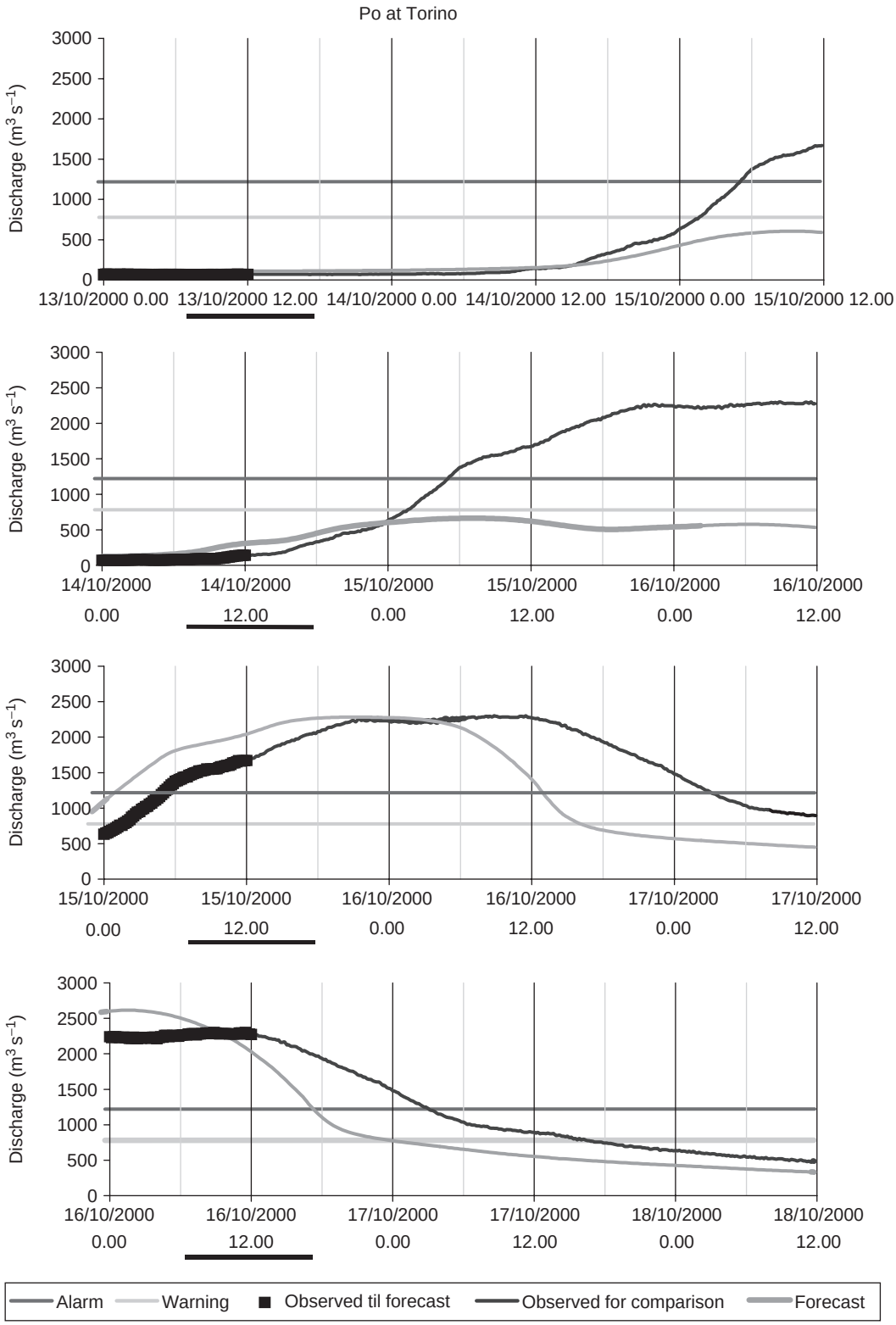


Figure 14 Successive flood forecasts of Po at Torino, Italy (Courtesy of ARPA Piemonte). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

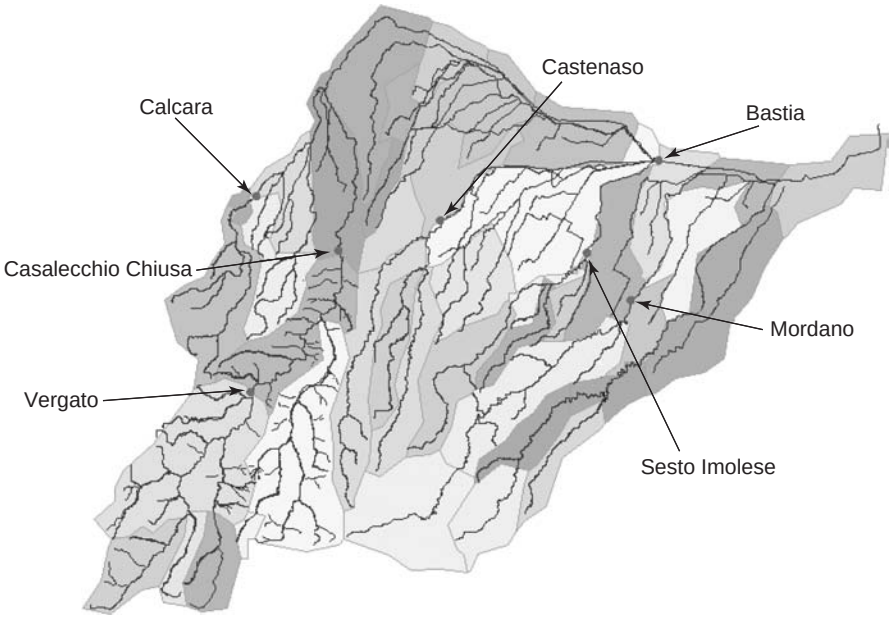


Figure 15 The Reno river catchment area and its subcatchments. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

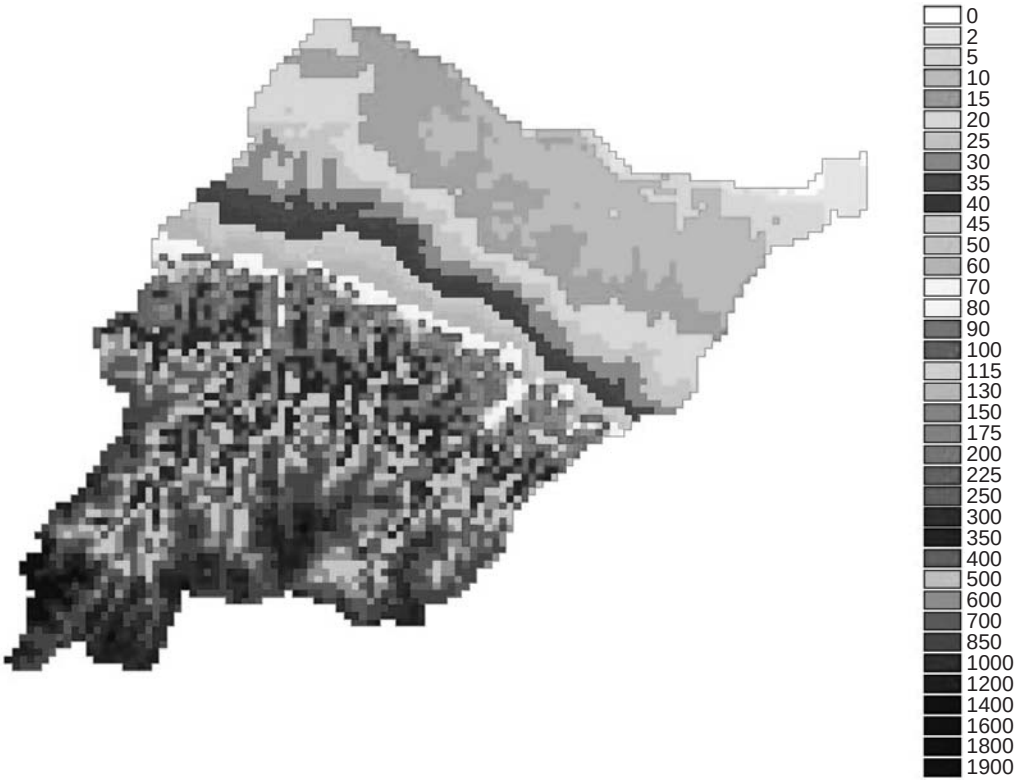


Figure 16 The digital elevation model of the Reno river catchment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

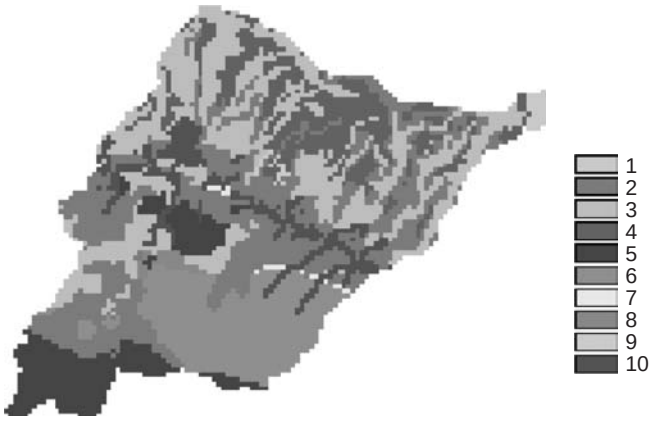


Figure 17 The soil types map of the Reno river catchment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Table 3 The 10 classes of soil types in the Reno river catchment

n	Soil type	Percentage
1	Sand	0.59
2	Clay loam	28.23
3	Silt loam	25.86
4	Silty clay	12.12
5	Sandy loam	10.12
6	Loam	14.42
7	Clay loam	0.38
8	Loam sand	0.75
9	Sand	0.36
10	Gravel	7.17

Figure 18 The land use map of the Reno river catchment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

with an analogue-based technique is presently under test. Future precipitation is obtained by combining the QPF generated by a limited area model with 50 series of precipitation derived from past observations with a technique similar to the one proposed by Obled *et al.* (2002) and Obled and Djerboua (2001). The technique aims at finding the closest analogues in terms of geopotential fields at 500 hPa (which was shown to provide the best results) over a wide spatial domain conditioning the weather conditions over the catchment in the following hours (Figure 20).

The treatment of uncertainty follows the Krzysztofowicz (1999) approach. The deterministic model forecasted flows are converted into a Gaussian space via the Normal Quantile Transform and are used to produce the probability density of the future values of discharge conditional on the hydrological forecasts. Figure 21 shows an example of successive 6 h in advance forecasts compared to the

Table 4 The nine classes of land use in the Reno river catchment

n	Land use	Percentage
1	Continuous urban, industrial and business areas, roads, and railways, airports, rock, basins	1.38
2	Discontinuous urban, dumps, green urban areas, sports/recreation areas	3.31
3	Extraction area, apamoors and shrubs, evolving vegetation, sparse vegetation	5.53
4	Crops: farm, heterogeneous, annual/permanent, nonirrigated seed	63.17
5	Vineyards	0.08
6	Fruit orchards	0.46
7	Meadows, natural pasture	1.84
8	Woods: broadleaf, conifer, mixed, beaches	23.64
9	Watercourses, inland swamps, lagoons	0.61

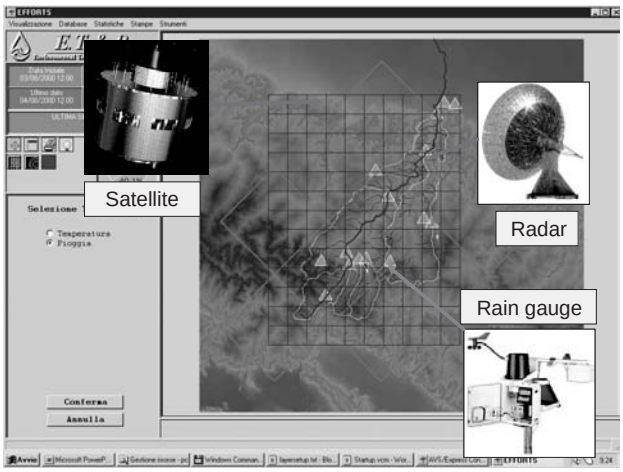


Figure 19 The different rainfall measurement systems (raingauges, RADAR, METEOSAT) integrated in EFFORTS and the relevant measurement points and meshes (by permission of PROGEA (Protezione e Gestione Ambientale) Srl). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

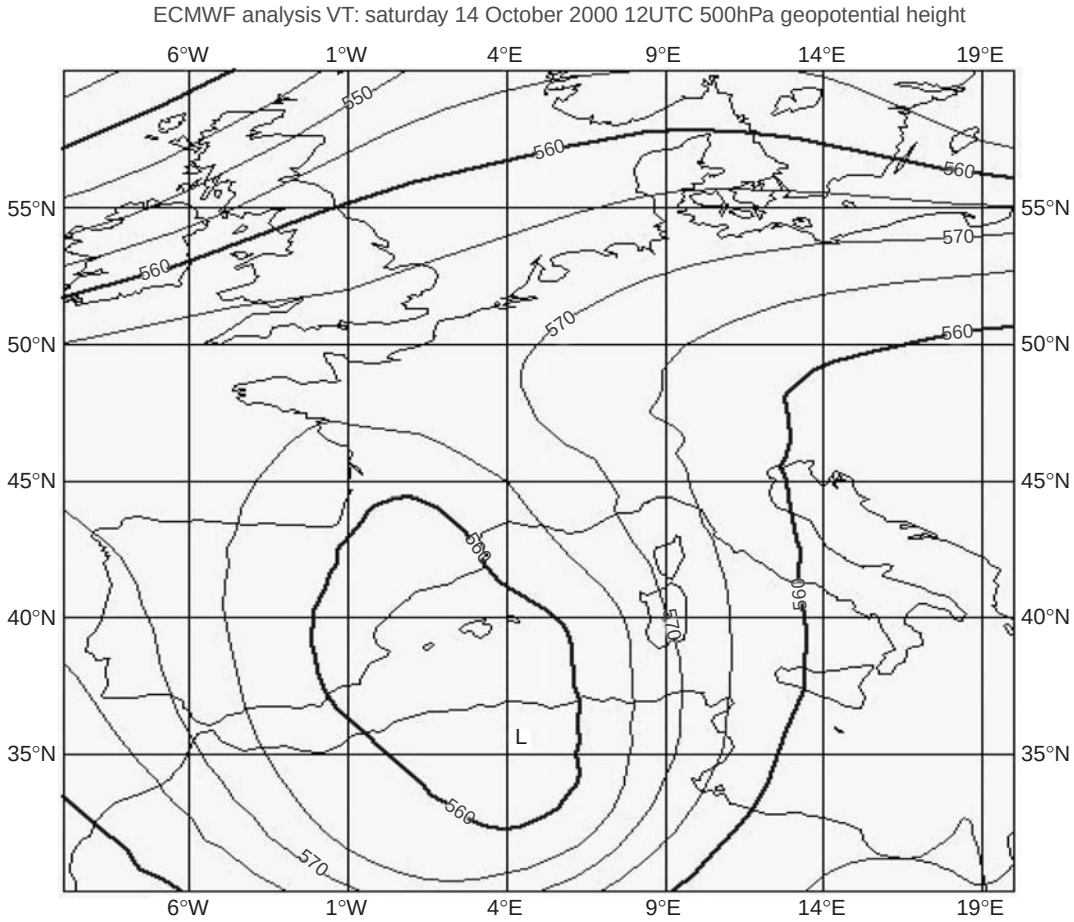


Figure 20 An example of the geopotential field at 500 hPa used for deriving the meteorological analogues and the corresponding precipitation series. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

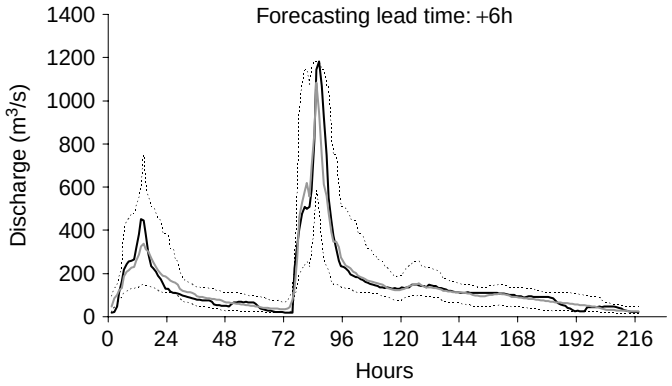


Figure 21 Successive six hours in advance forecasts on the Reno river at Casalecchio (grey) compared to the observed discharge values (black) and conditional uncertainty expressed as lower and upper quantiles (dashed) corresponding to $\pm 2\sigma$ in the Standard Normal space. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

observed discharge values (unknown at the time of forecast) together with quantiles corresponding to plus or minus two standard deviations of errors in the Normal space. Although the example refers to one of the largest events recorded in the Reno river in the latest years (November 6–7, 2000), it can be noted how the observed value falls always well inside the uncertainty band and that the band results as being asymmetric and nonstationary in time.

FUTURE DEVELOPMENTS

The HEPEX Project

A new international project, the Hydrological Ensemble Prediction Experiment (HEPEX, 2004) (**http://www.ecmwf.int/newsevents/meetings/workshops/2004/HEPEX/index.html**) was recently initiated, which aims at bringing the international hydrological and meteorological communities together to demonstrate how to produce reliable hydrological ensemble forecasts that can be used with confidence by the emergency management and water resources sectors to make decisions that have important consequences for the economy, public health, and safety.

The main scientific theme of HEPEX will be how hydrologic forecast uncertainty can reliably be quantified at each step of the forecast process and then communicated to and applied by the end users. Reliable quantification of forecast uncertainty is the key science issue that gives the program a unique and important role that at the present is not fulfilled by any existing hydrological prediction related program.

Similar to meteorological ensemble predictions, HEPEX will study the possibility of issuing hydrologic ensemble prediction by integrating many sources of uncertain information and accounting for how hydrologic processes would behave in response to it. Because hydrological models are imperfect and because of limitations in representing important sources of uncertainty, it is expected that hydrological ensemble forecasts will contain complex biases that must be removed to meet user requirements for reliable ensemble forecast information.

Data assimilation is required in hydrological ensemble prediction to process available observations to produce the best possible probabilistic estimates of initial hydrological conditions. These estimates must include ensemble members as well as probabilistic distributions of individual state variables. The ensemble members must represent the appropriate joint variable structure, both among state variables at a given location and spatially of equally likely possible initial states.

The relevance of HEPEX lies in its aim of giving an answer to most hydrologic modeling issues discussed in this article, such as:

- How can meteorological forecasting uncertainty be accounted for in hydrological forecasts?
- What are the sources of uncertainty in hydrological models?
- What are the implications of hydrological models being imperfect representations of real hydrological systems?
- How can uncertainties in hydrological models, model parameters, and hydrological initial conditions be represented in hydrological ensemble prediction?
- How can forecasting uncertainty be communicated to the users and how it can be used for improving decisions?

CONCLUSIONS

In this chapter, the description of rainfall-runoff models used in real-time flood forecasting has been placed into the perspective of the more complex problem of the effectiveness of the forecasts. The problem of real-time flood forecasting is, in fact, only partly dependent on the correct choice of the most appropriate rainfall-runoff model, although these models constitute the essential core of most of the existing operational systems. The effectiveness of the flood warnings and the success of the flood management decisions depend mostly upon the capability of a forecasting system to provide simple messages (e.g. “issue the warning” or “do not issue the warning”), which must be derived on the basis of a correct use of the predictive probabilities, to be assessed and estimated by combining all the possible uncertainties involved. Starting from an *a priory* subjective guess of these probabilities, it is then possible to revise them as a function of all the uncertain information gathered (such as input precipitation data) or generated via a model (such as the rainfall-runoff model) using a Bayesian integration technique, in order to obtain a posterior predictive probability density. This density, which is the best technical information one can provide, must then be used to produce simple messages by minimizing the expected value of a utility function (or several for the different purposes) associated to the scope of the required decision to be taken.

Given the importance of the objective at stake, extensive research is still needed in this domain, aiming at reducing flood damages and related casualties by improving the overall efficiency of the flood forecasting systems.

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J.A., O’Connell P.E. and Rasmussen J. (1986a) An introduction to the European hydrological system – Système Hydrologique Européen, “SHE”, 1: history and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 45–59.

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986b) An introduction to the European Hydrological System – Système Hydrologique Européen, "SHE", 2: structure of physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

ALERT, (2005) **www.alertsystems.org**.

Amorocho J. and Orlob G.T. (1961) *Nonlinear Analysis of Hydrologic Systems*, Contributors 40, Water Resources Centre, University of California: Berkeley.

Berger J.O. (1980) *Statistical Decision Theory Foundations, Concepts, and Methods*, Springer-Verlag: New York.

Bergström S. (1976) *Development and Application of Conceptual Runoff Model for Scandinavian Catchments*, Report No. 7, SMHI, Norrköping.

Bergström S. (1992) *The HBV Model – its Structure and Applications*, Report No. 4, SMHI, Norrkoping.

Bergström S. (1995) Chapter 13: The HBV model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Littleton.

Beven K.J. and Binley A.M. (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K.J. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Science Bull*, **24**, 43–69.

Box G.E.P. and Jenkins G.M. (1970) *Time Series Analysis Forecasting and Control*, Holden Day: San Francisco.

Brath A., Montanari A. and Toth E. (2002) Neural networks and non-parametric methods for improving realtime flood forecasting through conceptual hydrological models. *Hessische*, **6**(4), 627–640.

Brath A. and Rosso R. (1993) Adaptive calibration of a conceptual model for flash flood forecasting. *Water Resources Research*, **29**(8), 2561–2572.

Bree T. (1978) The stability of parameter estimation in the general linear model. *Journal of Hydrology*, **37**, 47–66.

Bruen M. and Dooge J.C.I. (1984) An efficient and robust method for estimating unit hydrograph ordinates. *Journal of Hydrology*, **70**, 1–24.

Burnash R.J.C., Ferral R.L. and Mc Guire R.A. (1973) *A General Streamflow Simulation System – Conceptual Modelling for Digital Computers*, Report by the, Joint Federal State River Forecasts Center, Sacramento.

Cameron D., Kneale P. and See L. (2002) An evaluation of a traditional and a neural net modeling approach to flood forecasting for an upland catchment. *Hydrological Processes*, **16**(5), 1033–1046.

Carter C.K. and Khon R. (1994) On Gibbs sampling for state-space models. *Biometrika*, **81**, 641–653.

Chao-Lin C. (Ed.) (1978) *Application of Kalman Filter to Hydrology, Hydraulics and Water Resources*, University of Pittsburgh: Penn.

Clarke R.T. (1973) *Mathematical Models in Hydrology*, Irrigation and drainage Paper No. 19, FAO: Rome.

Cooper D.M. and Wood E.F. (1982a) Identification of multivariate time series and multivariate input-output models. *Water Resources Research*, **18**(4), 937–946.

Cooper D.M. and Wood E.F. (1982b) Identification of multivariate time series and multivariate input-output models: application to rainfall-runoff processes. *Water Resources Research*, **18**(5), 1352–1364.

Crawford N.H. and Linsley R.K. (1966) *Digital Simulation in Hydrology*, Stanford Watershed model IV, Technical Report No. 39, Department Civil Engineering Stanford University.

Dawdy D.R. and O'Donnell T. (1965) Mathematical models of catchment behavior. *Journal of the Hydraulic Division Proceedings ASCE.* **91**(HY4), 123–131.

Dawson C.W. and Wilby R.L. (2001) Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, **25**(1), 80–108.

De Roo A.P.J., Wesseling C.G. and Van Deursen W.P.A. (1998) Physically based river modelling within a GIS. The LISFLOOD model. In *Geo-Computation CD-ROM*, Produced by Abrahart R.J., *Proceedings of 3rd International Conference on Geo-Computation*, ISBN 0-9533477-0-2, **http:/www.geocomputation.org/1998/06/gc$_{06}$.htm**.

De Roo A.P.J., Wesseling C.G. and Van Deursen W.P.A. (2000) Physically-based river basin modelling within a GIS: the LISFLOOD model. *Hydrological Processes*, **14**, 1981–1992.

DHI (2000a) *FLOODWATCH, User Guide and Reference Manual*, DHI Water and Environment: Horsholm.

DHI (2000b) *MIKE 11, User Guide and Reference Manual*, DHI Water and Environment: Horsholm.

DHI (2000c) *NAM Technical Reference and Model Documentation*, DHI Water and Environment: Horsholm.

DIIAR Politecnico di Milano (2002) *Valutazione Delle Soglie idrometeorologiche di allarme delle piene fluviali del Bacino dell'Arno*, Report of Politecnico di Milano (in Italian).

Obled Ch. and Djerboua A. (2001) Quantitative Precipitation Forecasts: a real-time exercise during the MAP experiment. In Ranzi R. and Bacchi B. (Eds.), *Hydrological aspects in the Mesoscale Alpine Programme-SOP experiment*, Technical Report N. 10. Dept. of Civil Engineering, University of Brescia.

Eagleson P.S., Mejia R. and March F. (1965) *The Computation of Optimum Realizable Unit Hydrographs from Rainfall and Runoff Data*, Report No. 84 – MIT, Hydrodynamics Laboratory.

EFFORTS (2005) *European Flood Forecasting Operational Real Time System*, Progetti di Gestione Aziendale Srl, **www.progea.net**.

EFFS (2003) *An European Flood Forecasting System*, Delft Hydraulics. **http://effs.wldelft.nl**.

Elzeim, A.R.S. and Adam I.S. (1996) Sudan Flood Early Warning System (FEWS). *El Mouhandis*, **2**(5), 23–24. Also on **www.geocities.com/TheTropics/5379/fews.html**.

ET&P (1992) *The Fuchun River Project – A Computer Based Real-time System*, EC-China Cooperation. Final Report Research Contract CI13-0004-I.

Evensen G. (1994) Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**(C5), 10 143–10 162.

Evensen G. (2003) The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, **53**, 343–367.

Ewen J., Parkin G. and O'Connell P.E. (2000) SHETRAN: distributed river basin flow and transport modeling system. *Journal of Hydrologic Engineering*, **5**(3), 250–258.

FLOODRELIEF (2003) *A Real-time Decision Support System Integrating Hydrological, Meteorological and Radar Technologies*, DHI, **http://projects.dhi.dk/floodrelief/index2.htm**.

Franchini M. and Pacciani M. (1991) Comparative analysis of several conceptual rainfall runoff models. *Journal of Hydrology*, **122**, 161–219.

Freeze R.A. and Harlan R.L. (1969) Blueprint for a physically-based Digitally-simulated hydrologic response model. *Journal of Hydrology*, **9**, 237–258.

Garcìa-Bartual R. (2002) Short term river forecasting with neural networks. *Integrated Assessment and Decision Support Proceedings of the 1st Biennial Meeting of the International Environmental Modelling and Software Society*, **2**, 160–165. ISBN: 88-900787-0-7).

Gelb A. (Ed.) (1974) *Applied Optimal Estimation*, The M.I.T. Press: Cambridge.

Gelfand A.E. and Smith A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 389–409.

Georgakakos K.P. (1986a) A generalized stochastic hydrometeorological model for flood and flash-flood forecasting. 1. Formulation. *Water Resources Research*, **22**(13), 2083–2095.

Georgakakos K.P. (1986b) A generalized stochastic hydrometeorological model for flood and flash-flood forecasting. 2. Case studies. *Water Resources Research*, **22**(13), 2096–2106.

Georgakakos K.P. (1989) Real time coupling of hydrological and meteorological models for flood forecasting. In *Recent Advances in the Modelling of Hydrological Systems*, Bowles D. and O'Connell P.E. (Eds.), Reidel Publishing Company.

Gupta N.K. and Mehra R.K. (1974) Computational aspects of maximum likelihood estimation and reduction of sensitivity function calculations. *IEEE Transactions on Automatic Control*, **19**, 774–783.

Handmer J.W. and Penning-Rowsell E.C. (Eds.) (1990) *Hazard and Communication of Risk*, Gower Technical Press: Aldershot.

HEPEX (2004) *Hydrological Ensemble Prediction Experiment*, **http://www.ecmwf.int/newsevents/meetings/workshops/2004/HEPEX/index.html**

Hino M. (1973) On-line prediction of hydrologic systems, *Proceedings of XV Conference IAHR*, Istanbul, pp. 121–129.

Jazwinski A. (1970) *Stochastic Processes and Filtering Theory*, Academic Press: New York.

Juemou W., Ruifang Z. and Guanwu X. (1987) Synthesised Constrained Linear System (SCLS), *Journal of Hydraulic Engineering*, **7**.

Julier S.J and Uhlmann J.K. (1997) A new extension of the Kalman filter to nonlinear systems, *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Controls*.

Kalman R.E. (1960) A new Approach to linear filtering and prediction problems. *Journal of Basic Engineering Transaction ASME*, **82**(D), 35–45.

Kalman R.E. and Bucy R.S. (1961) New results in linear filtering and prediction theory. *Journal of Basic Engineering Transaction ASME*, **83**(D), 95–108.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin M., Iredell M., Saha S., White G., Woolen J., *et al.* (1996) The NCEP/∼ NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77**(3), 437–471.

Kelly K.S. and Krzysztofowicz R. (1997) A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrology and Hydraulics*, **11**, 17–31.

Kelly K.S. and Krzysztofowicz R. (2000) Precipitation uncertainty processor for probabilistic river stage forecasting. *Water Resources Research*, **36**(9), 2643–2653.

Kendall M.G. and Stuart A. (1967) *The Advanced Theory of Statistics. Vol II*, Charles Griffin & Company: London.

Kitanidis P.K. and Brass R.L. (1980a) Real-Time forecasting with a conceptual hydrological model, 1. Analysis of uncertainty. *Water Resources Research*, **16**(6), 1025–1033.

Kitanidis P.K. and Brass R.L. (1980b) Real-Time forecasting with a conceptual hydrological model, 2. Application and results. *Water Resources Research*, **16**(6), 1034–1044.

Kouwen N. (2000) *WATFLOOD/SPL: Hydrological Model and Flood Forecasting System*, University of Waterloo: Waterloo. Department of Civil Engineering, Also see **http://www.civil.uwaterloo.ca/watflood/**

Krzysztofowicz R. (1999) Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, **35**(9), 2739–2750.

Krzysztofowicz R. (2001) Integrator of uncertainties for probabilistic river stage forecasting: precipitation-dependent model. *Journal of Hydrology*, **249**, 69–85.

Krzysztofowicz R. and Herr H.D. (2001) Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *Journal of Hydrology*, **249**, 46–68.

Krzysztofowicz R. and Kelly K.S. (2000) Hydrologic uncertainty processor for probabilistic river stage Forecasting. *Water Resources Research*, **36**(11), 3265–3277.

Kuczera G. (1994) *NLFIT: A Bayesian Nonlinear Regression Program Suite*, University of Newcastle: Callaghan. Department. of Civil Engineering and Survey.

Lardet P. and Obled C. (1994) Real-time flood forecasting using a stochastic rainfall generator. *Journal of Hydrology*, **162**, 391–408.

Lees M., Young P., Ferguson S., Beven K. and Burns J. (1994) An adaptive flood warning scheme for the river Nith at Dumfries. In *Proceedings of 2nd International Conference on River Flood Hydraulics*, White W.R. and Watts J. (Eds.), John Wiley & Sons.

Liang X., Lettenmaier D.P. and Wood E.F. (1996a) One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the Two-Layer variable infiltration capacity model. *Journal of Geophysical Research*, **101**(D16), 21 403–21 422.

Liang X., Wood E.F. and Lettenmaier D.P. (1996b) Surface soil moisture parameterization of the VIC-2L model: evaluation and modifications. *Global and Planetary Change*, **13**, 195–206.

Libralon A. (2002) *Studio Delle Possibilità D'uso Del Modello TOPKAPI Per La Valutazione Delle Soglie Di Rischio Pluviometrico*, Unpublished Thesis of the University of Bologna, (In Italian).

Liu Z. and Todini E. (2002) Towards a comprehensive physically-based Rainfall-runoff model. *Hessische*, **6**(5), 859–881.

Mantovan P., Pastore A. and Tonellato S. (1999) Recursive estimation of system parameter in environmental time series. In *Classification and Data Analysis*, Vichi M. and Optiz O. (Eds.), Springer: pp. 311–318.

Mayne D.Q. (1965) Optimal Non-Stationary estimation of parameters of a linear system with Gaussian inputs. *Journal of Electronics Control*, **14**, 101–112.

Mazzetti C. and Todini E. (2002) Combining different sources of precipitation measurements using a Bayesian approach, *Proceeding of GEOenv 2002*, 4th *European Conference on Geostatistics for Environmental Application*, Barcelona.

Mehra R.K. (1970) On the Identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, **15**, 175–184.

Moore R.J. (1985) The probability-distributed principle and runoff production at point and basin scales. *Hydrological Science Journal*, **30**(2), 273–297.

Moore R.J. and Clarke R.T. (1981) A distribution function approach to Rainfall-runoff modelling. *Water Resources Research*, **17**(5), 1367–1382.

MUSIC. (2001) *Multi-Sensor Precipitation Measurements Integration, Calibration and Flood Forecasting*, **http://www.geomin.unibo.it/orgv/hydro/music/**.

Nash J.E. (1958) *The form of the Instantaneous Unit Hydrograph*, *IUGG General Assembly of Toronto, Vol. III – IAHS Publications, 45*, pp. 114–121.

Nash J.E. (1960) *A Unit Hydrograph Study, with Particular Reference to British Catchments*, Proceedings the Institution of Civil Engineers.

Natale L. and Todini E. (1976a) A stable estimation for linear models – 1. Theoretical development and Monte-Carlo experiments. *Water Resources Research*, **12**(4), 667–671.

Natale L. and Todini E. (1976b) A stable estimator for linear models – 2. Real world hydrologic applications. *Water Resources Research*, **12**(4), 672–675.

Nielsen S.A. and Hansen E. (1973) Numerical simulation of the Rainfall-runoff process on a daily basis. *Nordic Hydrology*, **4**, 171–190.

Obled C.h, Bontron G. and Garçon R. (2002) Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analog sorting approach. *Journal of Atmospheric Research*, **63**(3–4), 303–324.

Parker D.J. and Fordham M. (1996) An evaluation of flood forecasting, warning and response systems in the European Union. *Water Resources Management*, **10**, 279–302.

Penning-Rowsell E.C. and Fordham M. (Eds.) (1994) *Floods Across Europe: Bhazard Assessment, Modelling and Management*, Middlesex University Press: London.

Penning-Rowsell E.C. and Tunstall S.M. (1997) The weak link in the chain: flood warning dissemination, *Proceeding RIBAMOD Workshop*, Padua.

Prasad R. (1967) A nonlinear hydrologic system response model. *Proc. ASCE*, **93**(HY-4), 202–221.

Rabuffetti D. and Barbero S. (2005) Operational meteo-hydological warning and real-time flood forecasting. *The Piemonte Region Case Study*. Hessische.

Raiffa H. and Schlaifer R. (1961) *Applied Statistical Decision Theory*, The MIT Press: Cambridge.

Rao R.A. and Delleur J.W. (1971) *The Instantaneous Unit Hydrograph: Its Calculation by the Transform Method and Noise Control by Digital Filtering*, Technical Report No. 20, Purdue University Water Resources Research Centre.

Refsgaard J.C. and Storm B. (1995) Chapter 23: MIKE SHE. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Re-sources Publications: Littleton.

RFFS, (2005) *Real Time Flood Forecasting System*, Center for Ecology and Hydrology: Wallingford, **http://www.nwl.ac.uk/ih/www/research/mfloodfore.html**.

Rockwood D.M. (1982) Theory and practice of the SSARR model as related to analyzing and forecasting the response of hydrologic systems. In Singh V.P. (Ed.), *Applied modeling in catchment hydrology*, Water Resources Publications, Littleton, Colo., pp. 87–106.

Sage A.P. and Husa G.W. (1969) Adaptive filtering with unknown prior statistics, *Proceeding 1969 Joint Automation Control Conference*, Boulder (Co), 760–796.

Salmond D.J., Gordon N.J. and Smith A.F.M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings – F*, **140**(2), 107–113.

Shamseldin A.Y. (1997) Application of neural network technique to Rainfall-Runoff modelling. *Journal of Hydrology*, **199**, 272–294.

Sherman L.K. (1932) Streamflow from rainfall by the unit graph method. *Engineering News Record*, **108**, 501–505.

Singh V.P. and Woolhiser D.A. (2002) Mathematical modeling of watershed hydrology. *Journal of Hydrologic Engineering*, **7**(4), 270–292.

Sorooshian S. and Gupta V.K. (1983) Automatic calibration of conceptual rainfall-runoff models: the question of parameter observability and uniqueness. *Water Resources Research*, **19**, 260–268.

Spolia S.K. and Chander S. (1974) Modeling of surface runoff systems by an ARMA model. *Journal of Hydrology*, **22**, 317–332.

Sugawara M. (1967) The flood forecasting by a series storage type model. *International Symposium Floods and Their Computation*, IAHS: pp. 1–6.

Sugawara M. (1995) Chapter 6: Tank model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Littleton.

Szollosy-Nagy A. (1976) An adaptive identification and prediction algorithm for real-time forecasting of hydrological time series. *Hydrological Sciences Bulletin*, **3**, 163–176, XXI.

Tanaka H. (1999) Flood forecasting of the Mekong river in 1997. *FAO/RAP Publication 1999/14*, FAO: Bangkok, pp. 109–117.

Thiemann M., Trosset M., Gupta H. and Sorooshian S. (2001) Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research*, **37**(10), 2521–2535.

Tikhonov A.N. (1963a) Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, **4**, 1035–1038.

Tikhonov A.N. (1963b) Regularization of incorrectly posed problems. *Soviet Mathematics*, **4**, 1624–1627.

Todini E. (1995a) *AFORISM – A Comprehensive Forecasting System for Flood Risk Mitigation and Control*, Final Report of Contract EPOC-CT90-0023.

Todini E. (1995b) New trends in modelling soil processes from hillslope to GCM scales. In *The Role of Water and the Hydrological Cycle in Global Change*, *NATO ASI Series, Series I: Global Environmental Change, Vol. 31*, Oliver H.R. and Oliver S.A. (Eds.), Springer-Verlag pp. 317–347.

Todini E. (1978) Mutually interactive State/Parameter Estimation (MISP). In *Application of Kalman Filter to Hydrology, Hydraulics and Water Resources*, > Chao-Lin C. (Ed.), University of Pittsburgh: Penn, pp. 135–151.

Todini E. (1988) Rainfall runoff modeling: past present and future. *Journal of Hydrology*, **100**, 341–352.

Todini E. (1996) The ARNO Rainfall-Runoff model. *Journal of Hydrology*, **175**, 339–382.

Todini E. (1999) Using phase-space modelling for inferring forecasting uncertainty in non-linear stochastic decision schemes. *Journal of Hydroinformatics*, **1**, 75–82.

Todini E. (2001) A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hessische*, **5**(2), 187–199.

Todini E. (2002a) The ARNO model. In *Mathematical Models of Large Watershed Hydrology*, Chap. 16, Singh V.P., Frevert D.K. and Meyer S.P. (Eds.), Water Resources Publications: Littleton, pp. 687–716.

Todini E. (2002b) The CLS model. In *Mathematical Models of Large Watershed Hydrology*, Chap. 20, Singh V.P., Frevert D.K. and Meyer S.P. (Eds.), Water Resources Publications: Littleton, pp. 861–886.

Todini E. and Bongioannini Cerlini P. (1999) *TELFLOOD: Technical Report*, DISTGA University of Bologna.

Todini E. and Ciarapica L. (2002) The TOPKAPI model. In *Mathematical Models of Large Watershed Hydrology*, Chap. 12, Singh V.P., Frevert D.K. and Meyer S.P. (Eds.), Water Resources Publications: Littleton, pp. 471–506.

Todini E. and Wallis J.R. (1977) Using CLS for daily or longer period Rainfall-Runoff modelling. In *Mathematical Models for Surface Water Hydrology*, Ciriani T.A., Maione U. and Wallis J.R. (Eds.), John Wiley & Sons: London, pp. 149–168.

Todini E. and Wallis J.R. (1978) A real time rainfall-runoff model for an on-line flood warning system. In *Application of Kalman Filter to Hydrology, Hydraulics and Water Resources*, Chao-Lin C. (Ed.), University of Pittsburgh: Penn, pp. 355–368.

Ubertini L. (1982) Modelli stocastici di trasformazione Afflussi-Deflussi. *Valutazione dellle Piene*, Publication No. 165, Progetto Finalizzato CNR Conservazione del suolo – Dinamica Fluviale, pp. 159–195, (in Italian).

van der Kwaak J.E. and Loague K. (2001) Hydrologic-response simulations for the r-5 catchment with a comprehensive physics-based model. *Water Resources Research*, **37**, 999–1013.

Van der Waerden B.L. (1952) Order tests for two-sample problem and their power. *Indagationes Mathematicae*, **14**, 253–458.

Van der Waerden B.L. (1953) Order tests for two-sample problem and their power. *Indagationes Mathematicae*, **15**, 303–316.

Verdin J. and Klaver R. (2002) Grid-cell-based crop water accounting for the famine early warning system. *Hydrological Processes*, **16**(8), 1617–1630.

Vivoni E.R. (2003) *Hydrologic Modeling using Triangulated Irregular Networks: Terrain Representation, Flood Forecasting and Catchment Response*, Ph.D. Thesis MIT, Cambridge.

Vrugt J.A., Gupta H.V., Bouten W. and Sorooshian S. (2003) A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrological model parameters. *Water Resources Research*, **39**, 1201, doi:10.1029/2002WR001642.

Wallingford Software, (2005) (**http://www.wallingfordsoftware.com/products/rivers/**).

Wan E.A. and van der Merwe R. (2000) The unscented Kalman filter for nonlinear estimation. *Proceedings of Symposium 2000 on Adaptive Systems for Signal Processing*, Communication and Control (AS-SPCC).

Wiener N. (1949) *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, John Wiley and Sons: New York.

Wigmosta M.S., Vail L.W. and Lettenmaier D.P. (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, **30**(6), 1665–1679.

Wojcik P.J. (1993) On-line estimation of signal and noise parameters and the adaptive Kalman filtering. In *Approximate Kalman Filtering*, Chen G. (Ed.), World Scientific Publishing Company: Singapore.

WMO (1975) *Intercomparison of Conceptual Models Used in Operational Hydrological Forecasting*, Operational Hydrology Report 7; WMO 429, WMO: Geneva.

WMO (1992) *Simulated Real-Time Intercomparison of Hydrological Models*. OHR-38. WMO No. 779.

WMO/UNDP (1980) *Improvement of River Forecasting and Flood Warning System for the Indus River Basin in Pakistan, Project Findings and Recommendations*, PAK/74/027, PAK/75/WMO TF, Geneva.

Wood E.F. and O'Connell P.E. (1985) Chapter 15 – Real Time Forecasting. In *Hydrological Forecasting*, Anderson M.G. and Burt T.P. (Eds.), John Wiley & Sons.

Wood E.P., Lettenmaier D.P. and Zartarian V.G. (1992) A Land-surface hydrology parameterization with subgrid variability for general circulation models. *Journal of Geophysical Research*, **97**(D3), 2717–2728.

Wooding R.A. (1965a) A hydraulic model for the catchment-stream problem: I. Kinematic wave theory. *Journal of Hydrology*, **3**, 254–267.

Wooding R.A. (1965b) A hydraulic model for the catchment-stream problem: II. Numerical solutions. *Journal of Hydrology*, **3**, 268–282.

Wooding R.A. (1966) A hydraulic model for the catchment-stream problem: III. Comparison with runoff observation. *Journal of Hydrology*, **4**, 21–37.

Woolhiser D.A. and Liggett J.A. (1967) Unsteady, one-dimensional flow over a plane – the rising hydrograph. *Water Resources Research*, **3**(3), 753–771.

Yakowitz S. (1987) Nearest neighbour methods for time series analysis. *Journal of Time Series Analysis*, **8**(2), 235–247.

Yakowitz S. and Karlsson M. (1987) Nearest neighbour methods for time series with application to rainfall-runoff prediction. In *Stochastic Hydrology*, Mac Neil J.B. and Humphries G.H. (Eds.), Reidel: Hingham, pp. 149–160.

Young P.C. (1974) Recursive approach to time series analysis. *Bulletin of the Institute of Mathematics and its Applications*, **12**(5/6), 209–224.

Young P.C. (1984) *Recursive Estimation and Time Series Analysis*, Springer Verlag: Berlin.

Young P.C. (2001) Data-based mechanistic modelling and validation of rainfall-flow processes. In *Model Validation: Perspectives in Hydrological Science*, Anderson M.G. and Bates P.D. (Eds.), Wiley: Chichester, pp. 117–161.

Young P.C. (2002) Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences* **360** 1433–1450.

Young P.C. and Whitehead P. (1977) A recursive approach to time-series analysis for multivariable systems. *International Journal of Control*, **25**(3), 457–482.

Zhao R.J. (1977) *Flood Forecasting Method for Humid Regions of China*, East China College of Hydraulic Engineering: Nanjing.

124: Flood Routing and Inundation Prediction

PAUL D BATES

School of Geographical Sciences, University of Bristol, Bristol, UK

This contribution considers the methods available to hydrologists to simulate flow routing and inundation in natural channels. Whilst the fluid dynamics of free surface flow in rivers and over floodplains may be complex, the methods available to treat such problems range widely in complexity from simple empirical models to full solutions of three-dimensional Navier–Stokes equations with sophisticated representations of flow turbulence. These methods are reviewed in detail, followed by a consideration of their requirements for topography, friction, initial condition, boundary condition, and validation data. In particular, the contribution discusses how these needs are being met increasingly using remote sensing techniques. Lastly, the contribution considers techniques to evaluate uncertainties in flow modeling.

THE NEED FOR FLOW ROUTING AND INUNDATION MODELS

Stream channel hydraulic process may need to be represented in rainfall-runoff models for two main reasons. First, flow routing may be required to translate runoff that has arrived in a stream channel to some point of interest further down the stream network. Second, models may be required to translate a predicted point hydrograph, as produced by a rainfall-runoff model, into predictions of practical interest to catchment managers and civil protection authorities. These may include predictions of flood depth and extent, flow velocity, and possible defence failure. In both cases, the hydrologist draws on, and contributes to, the body of research into open channel flows, discussed in detail elsewhere in this volume (*see* **Chapter 135, Open Channel Flow – Introduction, Volume 4**; **Chapter 136, Hydrodynamic Considerations, Volume 4**; **Chapter 137, Uniform Flow, Volume 4**; **Chapter 138, Unsteady Flow, Volume 4**; and **Chapter 141, Computer Modeling of Overbank Flows, Volume 4**). For basin scale hydrologic forecasting and most flood prediction problems the hydrologist is generally concerned with flow routing and inundation prediction over relatively long river reaches or whole river basins of the order of $10^1 - 10^3$ km or more, and this strongly influences the types of hydraulic model adopted. Moreover, at large basin scales the travel time of flood waves within the channel system may become the dominant process in hydrological forecasting.

For real-time forecasting problems (see e.g. Young, 2002; Brath *et al.*, 2002) computational speed is a key requirement and this leads hydrologists to favor models based on simple wave equations or approximations to these. The same constraint also holds for assessments of predictive uncertainty (see e.g. Aronica *et al.*, 1998) where Monte Carlo analysis of model response to parameter variations may be required, thus resulting in a need to compute many different realizations with the same model. However, for design and planning studies where there is no requirement to minimize computational costs, two and three-dimensional hydraulic models may be applied, even when large domains are under consideration. For example, Sauvaget *et al.* (2000) report the application of a finite element solution of the two-dimensional Shallow Water equations to a 15-km reach of the Dordogne-Isle confluence plain to produce an optimum design for flood relief culverts. Sauvaget *et al.* use a very detailed unstructured grid consisting of 60 000 elements ranging in size from 2 m in the culverts to 150 m on flat floodplain areas. This leads to a single model simulation taking 10 days on a UNIX workstation.

This section will first consider the hydraulics of flood waves in both simple and compound channels in order to identify the processes operating at particular space and time scales and which define our current perceptual model of open channel flow. It will then examine the range of models adopted for both flow routing and inundation prediction in

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

hydrology and the ways in which these models make simplifying assumptions according to our best understanding of process dominance. The article concludes by examining data sources for inundation prediction models, and in particular the ways in which remotely sensed data are beginning to affect modeling opportunities, before finally considering uncertainties and uncertainty handling in such schemes.

FLOOD WAVE HYDRAULICS

Floods can be thought of as a long, low amplitude wave passing through a compound channel of complex geometry. In the very largest basins, such waves may be up to $\sim 10^3$ km in length or greater but with an amplitude of only $\sim 10^1$ m, and may take several months to traverse the whole system. Flood waves are translated downstream with speed or celerity, c (LT^{-1}), and attenuated by frictional losses and floodplain storage such that in downstream sections the hydrograph is flattened out (see Figure 1). Wave speeds can be shown (see NERC, 1975) to vary with discharge such that maximum wave speed occurs at approximately two-thirds bankfull capacity (Knight and Shiono, 1996). Typical observed values for c reported by NERC (1975) and Bates *et al.* (1998) for UK rivers are in the ranges 0.5 to 1.8 ms^{-1} and 0.3 to 0.67 ms^{-1} respectively.

Below the scale of the flood wave itself, other significant in-channel process can be identified, each with a characteristic length scale. These include shear layers forming at the junction between the main flow and slower moving "dead zones" at the scale of the channel platform (Hankin *et al.*, 2002), secondary circulations at the scale of the channel cross section (Bridge and Gabel, 1992; Nezu *et al.*, 1993), and turbulent eddies ranging from heterogeneous structures at the scale of roughness elements and obstructions on the bed (Ashworth *et al.*, 1996; McLelland *et al.*, 1999; Shvidchenko and Pender, 2001) down through the turbulent energy cascade (Hervouet and Van Haren, 1996) to the

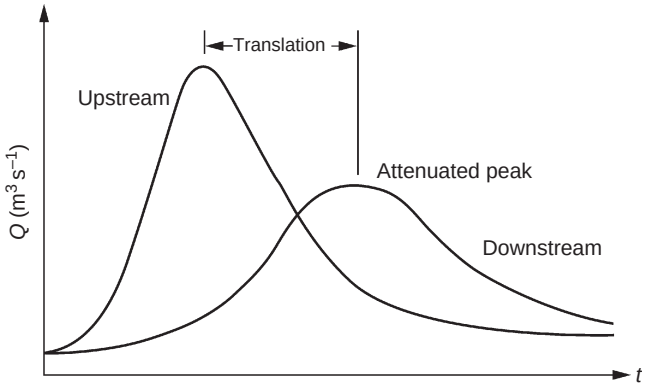


Figure 1 Translation and attenuation of a flood wave between two gauging stations

Kolmogorov length scale, η (L), where turbulent kinetic energy is finally dissipated. In typical channel flows, these very smallest eddies may be only $\sim 10^{-2}$ mm across (Hervouet and Van Haren, 1996) and are highly transient. All these scales may be fully described with the Navier–Stokes equation (Schlichting, 1979):

$$\rho \frac{D\mathbf{u}}{Dt} = -\nabla p + \mu \nabla^2 \mathbf{u} + F \tag{1}$$

where ρ is the fluid density (with dimensions ML^{-3}); $\mathbf{u}$ is the velocity vector (LT^{-1}); t is the time (T), p is the pressure ($ML^{-1}T^{-2}$); μ is the viscosity ($ML^{-1}T^{-2}$) and F is the set of terms (e.g. gravity, coriolis, and friction) to be included in the specification of a particular problem.

When equation (1) is combined with the equation of continuity:

$$\nabla \cdot \mathbf{u} = 0 \tag{2}$$

one obtains a system of equations that can be solved to yield the three-dimensional velocity vector $\mathbf{u} = (u\ v\ w)$; where u, v, and w are the three components of $\mathbf{u}$ in the x, y, and z directions respectively, and pressure, p, for a given point in time and space.

Solution of equations (1) and (2) over a suitably refined grid (i.e. $\Delta x \ll \eta$) using a suitably small time step can, in theory, simultaneously simulate all flow features described above. However, for typical flood wave flows (i.e. unsteady, nonuniform flows of high Reynolds number in a complex geometry), the direct numerical simulation of the Navier–Stokes equations is computationally prohibitive. Hydrologists have therefore sought to isolate from this complex assemblage those processes that are central to the problem of flood routing and inundation prediction in order to build appropriate models.

A starting point for such approximations is to consider the forces acting on the flow. For fluvial flood-routing problems the dominant driving and resisting forces are clearly gravity and friction. This realization led in the eighteenth and nineteenth centuries to the development of resistance laws for steady open channel flow by Chezy, Manning, and Darcy–Weisbach amongst others (see Chanson, 1999, pp. 72–91 for a discussion). However, to describe unsteady flows requires development of equations of motion based on principles of mass and momentum conservation for which the critical decision is the number of dimensions in which significant flow field variation occurs. Clearly, for the hydrologist concerned with calculating the translation and attenuation of a flood wave over long river reaches, significant variations in the flow field occur dominantly in the streamwise direction. Hence, despite known three-dimensional processes, channel flows are often considered as one-dimensional in the streamwise direction and variations in velocity in the vertical and cross-stream directions

may be neglected. For many channel flood-routing applications, this is a reasonable approximation, however, Knight and Shiono (1996) suggest that when flow exceeds bankfull capacity and begins to spread over adjacent low-lying floodplains, a series of two- and three-dimensional processes are initiated which may be important in some circumstances.

One such process is momentum exchange between fast moving water in the channel and slower moving water on the floodplain (Knight and Shiono, 1996). This momentum exchange occurs across the shear layer created at the interface between channel and floodplain which is manifest (see Figure 2) as a series of vortices with vertically aligned axes (Sellin, 1964; Fukuoka and Fujita, 1989; Shiono and Knight, 1991). Knight and Shiono (1996) draw attention to the three-dimensional structure of this interface zone, whilst Ervine and Baird (1982) conclude that failure to account for the resulting momentum exchange can lead to errors of up to ±25% in the discharge calculated using uniform flow formulae. Further vigorous momentum exchange occurs during out-of-bank flow in meandering compound channels (see Sellin and Willetts, 1996 for a discussion). Here, water spills from the downstream apex of channel bends and flows over meander loops before interacting with channel flow in the next meander (see Figure 3). These three-dimensional interactions modify secondary circulations within the channel and represent an additional energy loss in the near channel area. Floodplain flows beyond the meander belt will not be subject to such energy losses and this region may provide a route for rapid flow conveyance. The impact of these additional energy losses and storage of water on the floodplain leads to a reduction in wave speed for out-of-bank flows (see Figure 4). Minimum wave speeds occur at a particular shallow overbank stage when the interaction between main channel and floodplain is at its greatest (Knight and Shiono, 1996), before slowly increasing again as the whole floodplain and valley floor begins to behave as a single channel unit.

Flow in the floodplain itself is clearly a two-dimensional process, given that flows may be several kilometers in horizontal extent but only a few meters deep, and many inundation prediction studies make the assumption that the variation in vertical velocity is negligible compared to variation in the horizontal. Such shallow water flows over low-lying topography are characterized by rapid extension and retreat of the inundation front over considerable distances, potentially with distinct processes occurring during the wetting and drying phases (see Nicholas and Mitchell, 2003). Correct treatment of this moving boundary problem is, therefore, important both to capture adequately the shallow water energy losses (which may be high due to large relative roughness) and because flood extent is a common prediction requirement from hydraulic codes. Flow interactions with microtopography (see Walling *et al.*,

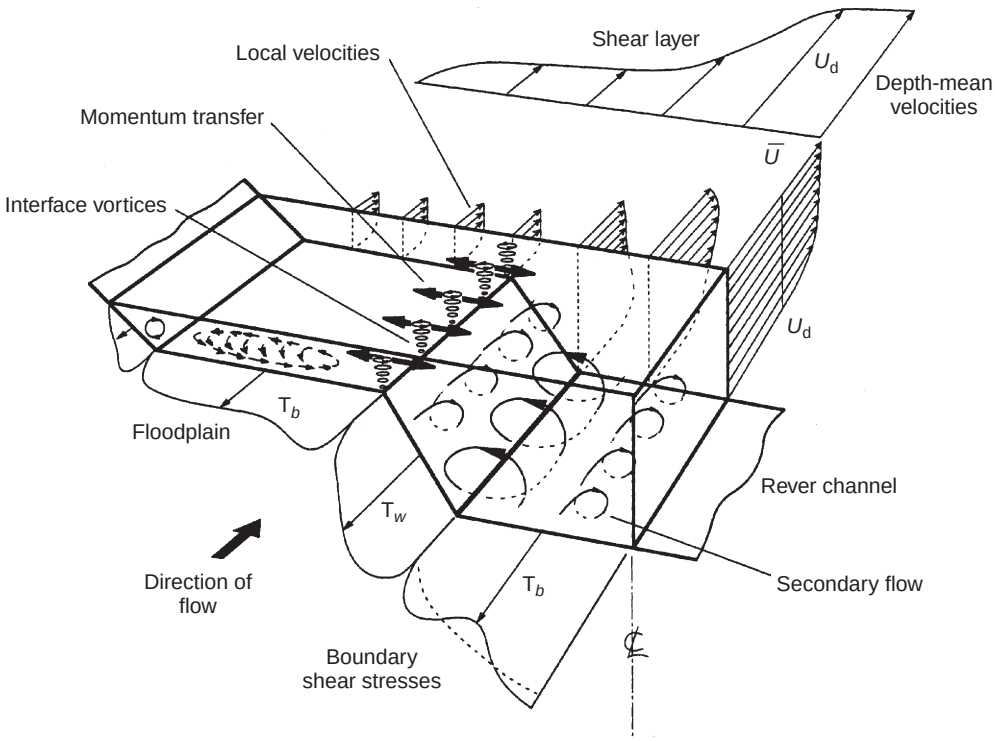


Figure 2 Hydraulic processes associated with overbank flow in a straight compound channel (after Shiono and Knight, 1991, © Cambridge University Press)

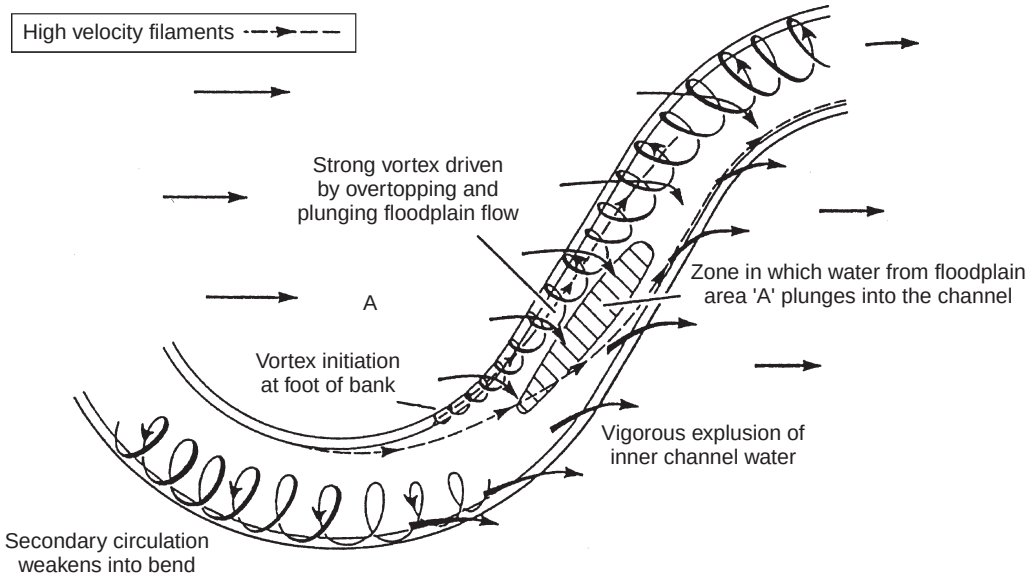


Figure 3 Representation of principal flow structures occurring during overbank flow in a meandering compound channel (after Sellin and Willetts, 1996, © John Wiley & Sons Ltd)

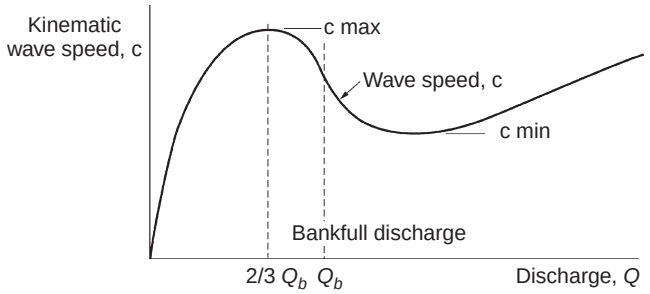


Figure 4 Typical kinematic wave speed versus discharge curve for a compound channel (after Knight and Shiono, 1996, © John Wiley & Sons Ltd)

1986), vegetation (Lopez and Garcia, 2001) and structures (Meselhe *et al.*, 2000) may all be important thereby giving a complex modeling problem. In particular, where the floodplain acts as a route for flow conveyance rather than just as storage, energy losses are dominated by vegetative resistance. In particular, hedges and walls may have a large impact on the effective floodplain roughness. Yet, despite a small number of pioneering studies (see e.g. Kouwen, 1988; Nepf, 1999; Ghisalberti and Nepf, 2002; Wilson and Horritt, 2002), the interaction between plant form, plant biomechanics, energy loss, and turbulence generation is at present relatively poorly understood. Moreover, most flood-routing models assume that the channel bed is fixed over the course of the event, and for very large floods this may not be the case as embankment failure or geomorphic change may considerably affect the propagation of the flood wave.

Lastly, whilst typical hydraulic models do not consider water exchanges with the surrounding catchment, for whole catchment modeling or flood inundation simulation over long river reaches such exchanges may, at particular times, become important (e.g. Stewart *et al.*, 1999; Woessner, 2000). Such processes include direct precipitation or runoff to the floodplain surface (e.g. Mertes, 1997), evapo-transpiration losses, so-called bank-storage effects (Pinder and Sauer, 1971; Squillace, 1996) resulting from interactions between the river water and alluvial groundwater contained within the hyporheic zone (Stanford and Ward, 1988; Castro and Hornberger, 1991; Wroblicky *et al.*, 1998), subsurface contributions to the floodplain groundwater from adjacent hillslopes (e.g Bates *et al.*, 2000; Burt *et al.*, 2002) and flows along preferential flow paths, such as relict channel gravels, within the floodplain alluvium (e.g. Haycock and Burt, 1993; Poole *et al.*, 2002). Over particular reaches and in particular environments, integration of some or all of these processes with flood-routing models may be required and necessitate complex modeling structures (e.g. Stewart *et al.*, 1999; Kohane and Welz, 1994).

In summary, it is common to assume that in-channel routing may be treated as a one-dimensional flow in the streamwise direction, and this is the approach taken in most industry standard hydraulic models (see Havnø *et al.*, 1994), for example, HEC-RAS (**http://www.scisoftware.com/environmental_software/product_info.php?products_id=182**); MIKE11 (**http://www.dhisoftware.com/mike11/**); ISIS Flow (**http://www.wallingfordsoftware.com/products/isis/**); SOBEK (**http://www.sobek.nl/**). For out-of-bank flows, such consensus does not yet exist, and the choice

of model is best considered as being dependent on scale and user needs. Clearly, floodplain flow is at least a two-dimensional processes, yet it is also known that in particular zones, such as the channel-floodplain interface, flow is strongly three-dimensional. A working assumption has evolved (see e.g. Feldhaus *et al.*, 1992; Bates *et al.*, 1992; Bates *et al.*, 1998) that reach scale flows are best treated with two-dimensional methods, whilst for shorter sections or applications such as sediment transfer to overbank sections, where the details of the channel-floodplain momentum transfer are clearly critical, a three-dimensional approach may be more appropriate. Such working assumptions are, however, provisional and waiting to be refined on the basis of better data or understanding.

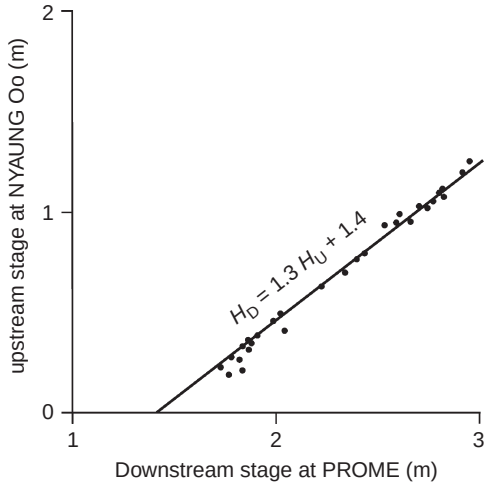


Figure 5 Correlation between peak flood stages recorded for 35 events at the Prome and Nyaung Oo gauging station on the River Irrawaddy, Burma (after Shaw, 1983)

FLOW ROUTING IN RAINFALL-RUNOFF MODELS

In rainfall-runoff modeling applied to relatively large river basins it is often necessary to model explicitly the translation and attenuation of a flood wave downstream through the channel network. As discussed above, for in-channel flows a consensus exists that one-dimensional models are an appropriate tool for such problems and to achieve this a number of empirical and physically based methods are available. These can be broadly classified as: (i) empirical or nonstorage models; (ii) full and simplified solutions of the one-dimensional St. Venant equations, and (iii) hydrological storage models. For flood-routing problems where flow is in an out-of-bank condition a number of methods exist which attempt to predict the additional energy losses and mass transfers in a way that can be incorporated in one-dimensional codes (Knight and Shiono, 1996, pp. 155–159; De Roo *et al.*, 2001). In the future, it may also be possible to integrate recently developed simplified two-dimensional codes (see e.g. Bates and De Roo, 2000) with hydrologic models. Such schemes have already been used to represent river-floodplain reaches up to 60 km in length and may allow simultaneous simulation of flood routing and inundation prediction at the whole catchment scale.

Empirical or Nonstorage Flood-routing Models

At the most basic level, flood-routing predictions may be achieved by building regression models to predict stage or discharge at a particular point on the river network based on observed stage or discharge at some upstream point (see Fread, 1984, p. 439). For example, Shaw (1983, p. 379) includes an example (see Figure 5) showing good correlation between 35 flood peaks recorded over a five-year period at two gauging stations 345 km apart on the River Irrawaddy, Burma. Such simple linear relationships may work well under certain circumstances, but like all regression models they are not easy to extrapolate to conditions beyond those for which training data are available. For example, Shaw (1983) points out that floods of the same peak flow but different volume may show very different routing behavior as they move downstream through a river network. Recently, this type of approach has been revisited in a more sophisticated way using Transfer Function models (e.g. Young, 1986; Lees *et al.*, 1994; Young, 2002; Young, **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**) to relate upstream and downstream levels. These consist of linear and nonlinear functions that are the discrete-time equivalents of differential equations and describe the temporal behavior only at selected spatial nodes within the catchment system. For example, Lees *et al.* (1994) describe a model to forecast flood levels in sections of the River Nith, Dumfries, UK, based only on upstream levels measured up to 5 h ahead and a general noise term as part of an adaptive real-time flood warning scheme. The overall system achieved mean prediction accuracy in terms of peak stage of 0.08 m rmse over nine large flood events. Such models can also be combined with distributed models to yield distributed predictions, as demonstrated by Romanowicz and Beven (1998) for the case of flood inundation forecasting. However, where the system will be used to undertake environmental design or forecast the effects of change then distributed, physically based models have an advantage. These are described in the following section.

Full and Simplified Solutions of the One-dimensional St. Venant Equation

As outlined previously, in-channel flood routing is often assumed to be a one-dimensional shallow water flow in the streamwise direction (e.g. Fread, 1984; Samuels, 1990;

Fread, 1993; Singh, 1996). Equations for one-dimensional channel flow can be derived by considering mass and momentum conservation between two cross sections Δx apart. This yields the well-known one-dimensional St. Venant or shallow water equations:

Conservation of momentum

$$\frac{\partial Q}{\partial t} + \frac{\partial (Q^2/A)}{\partial x} + gA\left(\frac{\partial h}{\partial x} + S_f - S_o\right) = 0 \qquad (3)$$

where Q is the flow discharge ($L^3\,T^{-1}$); A is the flow cross-section area (L^2), g is the gravitational acceleration (LT^{-2}), S_f is the friction slope (LL^{-1}) and S_o is the channel bed slope (LL^{-1}).

Conservation of mass

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = q \qquad (4)$$

where q is the lateral inflow or outflow per unit length ($L^2\,T^{-1}$).

Equations (3) and (4) have no exact analytical solution, but with appropriate boundary and initial conditions they can be solved using numerical techniques (e.g. Preissmann, 1961; Abbott and Ionescu, 1967) to yield estimates of Q and h in both space and time. The river reach in one-dimensional St. Venant models is discretized as a series of irregularly spaced cross sections (see Figure 6). Boundary conditions consist of specifying the scalar velocity, u (LT^{-1}) and h at each inflow or outflow boundary. In practice, however, measurements of u are rarely available from river gauging stations and we generally have only h. This can be used with an approximate rating curve and cross-sectional survey to give values of Q and A from which u can be derived. This rating curve can be measured, or derived from some uniform flow formulae such as the Manning equation, if the appropriate roughness coefficient is given. In this case uniform flow represents an additional approximation that implies the water surface is parallel to the bed slope. This is not the case for a full dynamic wave, where the water-surface slope should be steeper than the bed slope on the rising limb of the hydrograph, and less steep on the falling limb. Thus, it is important to be very clear what assumptions and approximations are being made in any particular boundary condition specification. For critical flow where the Froude number, $F_r = u/\sqrt{gh}$, exceeds 1 where u is a scalar velocity (LT^{-1}), waves cannot propagate upstream within the domain and no downstream boundary condition need therefore be prescribed. These equations form the basis of most standard commercial hydraulic modeling softwares such as HEC-RAS (**http://www.scisoftware.com/environmental_software/product_info.php?products_id=182**); MIKE11 (**http://www.dhisoftware.com/mike11/**); ISIS Flow (**http://www.wallingfordsoftware.com/products/isis/**); and SOBEK (**http://www.sobek.nl/**) and assume (NERC, 1975; Beven, 2001) that:

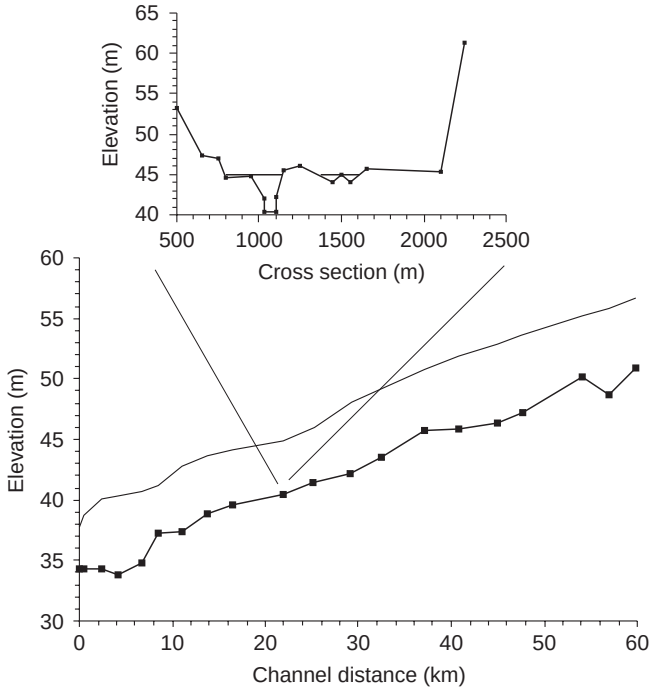


Figure 6 A typical discretization from a one-dimensional hydraulic model, in this case the HEC-RAS scheme applied to a 60-km reach of the River Severn in the United Kingdom by Horritt and Bates (2002) © Elsevier. Shown are the bed elevation profile with model cross-section locations marked by squares and a typical simulated water free surface profile. A sample cross section is also shown

- the flow can be adequately represented by the average flow velocity and average flow depth at any cross section;
- the water-surface slope varies gradually downstream so that pressure distribution in the vertical can be assumed hydrostatic;
- the water is incompressible and of constant temperature and density;
- the friction slope may be estimated using one of the standard uniform flow formulae.

Despite recent advances in computer power that allow the full dynamic flood wave contained in equations (3) and (4) to be solved with relative ease, there are many situations where this complexity is not required. Simplification of the full St. Venant equations may therefore be possible and can be achieved merely by neglecting particular terms in equation (3) to yield an approximate solution to the full equation. The two main classes of approximate solution are the kinematic and diffusion forms of equation (3).

The kinematic model was first described by Lighthill and Whitham (1955) and its use in hydrology has recently

been comprehensively reviewed by Singh (2001). The model assumes that the first three right-hand-side terms in equation (3) are small compared to the fourth term and the momentum equation can be simplified to:

$$S_f - S_o = 0 \tag{5}$$

This implies that the momentum of an unsteady flow is equivalent to that of a steady uniform flow as described by the Chezy or Manning equations where discharge is a single valued function of depth and the water surface is always parallel to the bed (Fread, 1984). If velocity, **u**, is evaluated using the Manning equation, the kinematic wave speed, c_k (LT^{-1}) in a wide, rectangular channel will be (Graf and Altinakar, 1998, pp. 274–275):

$$c_k = \frac{5}{3}\mathbf{u} \tag{6}$$

The kinematic approximation to the full one-dimensional St. Venant equations has three important limitations which arise from the assumption that the water slope is always parallel to the bed:

1. In channels of uniform geometry the kinematic model can simulate wave translation only and not the attenuation effect, such that although the shape of a kinematic wave may change as it propagates through a reach, its peak discharge will not. In variable geometry, however, both the peak and wave shape will change.
2. Hysteresis in the discharge-stage relationship cannot be represented.
3. Similarly, the effect of disturbances can only propagate in a downslope direction (Beven, 2001) and the kinematic model cannot, therefore, predict backwater effects in subcritical flows.

All the above three effects require the water and bed slope be nonparallel, and limitation 3 also means that for a kinematic flood-routing model the bed slope must be everywhere negative and a downstream boundary condition is not required. However, despite these constraints, kinematic equations form the basis of many flood-routing codes (see Singh, 1996; Bates and De Roo, 2000; Cameron *et al.*, 2002; Cooper *et al.*, 2002). This is not only because of their computational efficiency but also because Lighthill and Whitham (1955) have shown for the subcritical flow with $F_r < 1$, as is typical of most floods, that dynamic waves decay exponentially and kinematic waves come to dominate.

The diffusion form of the St. Venant equations is based on the assumption that only the inertia terms, $\partial Q/\partial t + \partial(Q^2/A)/\partial x$, in equation (3) are insignificant and thus the momentum equation can be approximated as:

$$\frac{\partial h}{\partial x} + S_f - S_o = 0 \tag{7}$$

Inclusion of the water-surface slope term, $\partial h/\partial x$, allows the model, unlike the kinematic approximation, to describe the flood wave attenuation effect as described in the Section on "Flood wave hydraulics". Attenuation can be thought of as a diffusion process, hence giving the model its name (see Figure 7). It also follows that the diffusion model can deal with stage-discharge hysteresis, nonnegative slopes, and backwater effects and also that numerical solution of the equations (4) and (7) require specification of a downstream boundary condition. Diffusion wave models are also widely utilized in hydrology (Moussa, 1997; Horritt and Bates, 2001a) and continue to be the subject of development research (see Cappelaere, 1997; Moussa and Bocquillon, 2001). However, Fread (1984) notes that inclusion of the additional inertia terms to derive a full dynamic wave model actually adds relatively little to the computational burden over solution of the diffusion wave equation. Moreover, comparison of different wave routing approximations for channel flow by Horritt and Bates (2001a) showed that the computational cost of a kinematic solution was an order of magnitude less than an alternative diffusion wave model. There may therefore be relatively little advantage, in terms of either flow physics or computational cost, in using the diffusion wave approximation over a full dynamic wave.

Fread (1984), Ponce *et al.* (1978), Vieira (1983), Moussa and Bocquillon (1996), Graf and Altinakar (1998, pp. 280–281), and Singh (2001) all discuss the range of validity for kinematic, diffusion, and full dynamic wave equations for flood routing. Singh (2001) argues that in a channel flow, all wave types exist at any given position, with their relative significance changing with the changing nature of the flow as characterized by the Froude number, F_r.

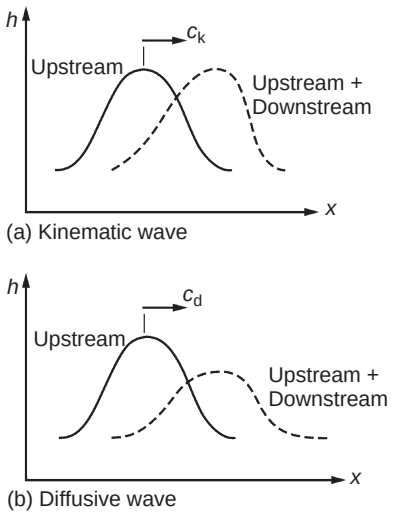


Figure 7 Translation and attenuation of a flood wave under different hydraulic approximations (after Graf and Altinakar, 1998) where c_k is the speed of a kinematic wave and c_d is the speed of a diffusive wave

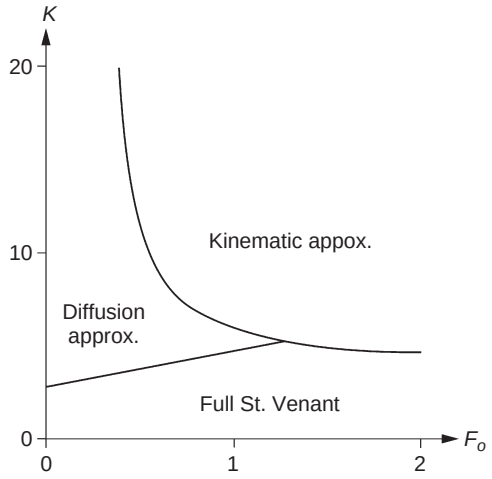


Figure 8 Ranges of validity of approximations to the one-dimensional full St. Venant equations in terms of the dimensionless Froude (F_o) and kinematic wave (K) numbers (after Vieira, 1983, © Elsevier)

Vieira (1983) presents a similar analysis in terms of the dimensionless Froude and kinematic wave numbers (see Figure 8), whilst Ponce *et al.* (1978) conclude that bottom slope and wave shape determine the range of applicability. The consensus of these studies (Graf and Altinakar, 1998) is that for typical flood flows where $F_r < 1$ the kinematic approximation holds, but on very low angle slopes where $F_r \ll 1$ a flood must be approximated as a diffusion wave as backwater effects become significant. Moreover, Beven (2001, p. 116) argues that uncertainty over effective parameter values and boundary conditions may mean that simplified models may even be useful under a wider range of conditions than this. It may thus not be possible to conclusively discriminate that a more complex treatment of the wave physics gives better performance, particularly if the model is calibrated. For a full discussion of these studies the reader is referred to Singh (2001, pp. 682–685).

Hydrological Storage Models

Even greater simplification of equations (3) and (4) can be achieved by replacing the momentum equation with an arbitrary functional relationship that describes storage in a reach as a function of its inflow and outflow. This is often referred to as hydrological storage flood routing, and a complete background to such methods is given by Shaw (1983, pp. 386–391). Hydrological storage models treat the river reach in question as a "black box" (NERC, 1975), and assume that the geometry and frictional resistance are too complex to be modeled in detail. The first step in implementing such a procedure is to replace the mass conservation equation (4) with:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = I - O \tag{8}$$

where S is the storage within a river reach (L^3), I is the inflow to the reach ($L^3\,T^{-1}$) and O is the outflow ($L^3\,T^{-1}$).

The momentum equation (3) is then replaced with a relationship which expresses storage as a linear function of the inflow I, outflow O, and one or more parameters which are specific to the river reach being studied. A general form of this relationship is given (Fread, 1984) as:

$$S = K[\lambda I + (1-\lambda)O] \tag{9}$$

where K is a travel time constant (T) and λ is a dimensionless weighting parameter indicating the relative importance of I and O in determining the storage S within the reach. λ has lower and upper limits of 0 and 0.5, and typical values in the range 0.2 to 0.4 (Shaw, 1983, p. 388).

Similar to the kinematic wave equation, all hydrological routing models are limited (Fread, 1984) to applications where there is unique relationship between depth and discharge. Thus, they cannot consider backwater effects or "looped" depth-discharge rating curves. The earliest hydrological storage routing method was proposed by McCarthy (1938) and is termed *the Muskingum method*. This considers flow routing through a reach between time $t = 1$ and $t = 2$. Using this notation, equation (8) is first recast in centered finite difference form:

$$\frac{S_2 - S_1}{\Delta t} = \frac{I_1 + I_2}{2} - \frac{O_1 + O_2}{2} \tag{10}$$

where the subscripts refer to the time. Equation (9) is then substituted into equation (10) and rearranged to give an equation in terms of the desired unknown, namely outflow from the reach at $t = 2$.

$$O_2 = C_a I_2 + C_b I_1 + C_c O_1 \tag{11}$$

with

$$C_a = \frac{\Delta t + 2K\lambda}{\Delta t + 2K - 2K\lambda} \tag{12}$$

$$C_b = \frac{\Delta t - 2K\lambda}{\Delta t + 2K - 2K\lambda} \tag{13}$$

$$C_c = \frac{-\Delta t + 2K - 2K\lambda}{\Delta t + 2K - 2K\lambda} \tag{14}$$

As $\sum C = 1$, the outflow from the reach at the end of a time step is a weighted sum of the starting inflow, the starting outflow, and the ending inflow. Young (1986) points out that this is now a Transfer Function model (see the Section on "Empirical or nonstorage flood-routing models" and Young, **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**), but not a good one as it does not include a time delay. The Muskingum method can therefore be viewed as just one structure in a general

class of autoregressive moving average (ARMA) models (Beven and Wood, 1993, p. 103).

To implement the Muskingum method the coefficients K and λ must be calibrated based on observed values of I and O. The methods can thus only be applied between gauging stations and only the average flood wave shape is reflected in the fitted routing parameters (Fread, 1984). Moreover, best-fit Muskingum parameters will often have an impulse response with nonphysical negative ordinates unless the reach is split down into the right number of subreaches to get a time delay compatible with the time step used (Young, 1986).

An improvement on the Muskingum method which removes the need for calibration of K and λ was proposed by Cunge (1969) who showed that K was approximately equal to the travel time of the flood wave through the reach:

$$K = \frac{\Delta x}{c_k} \tag{15}$$

Where x is the reach length and c_k is the kinematic wave speed. Given equation (15), Cunge then derived an expression for λ based on channel properties:

$$\lambda = 0.5\left[\frac{1 - q_o}{c_k S_o \Delta x}\right] \tag{16}$$

where q_o is a unit width reference discharge ($L^2\,T^{-1}$) obtained by dividing the mean peak discharge by the mean channel width.

While the Muskingum–Cunge method does not require observed inflow and outflow hydrographs to establish the routing coefficients, Fread (1984) notes that best results are obtained if the wave speed is determined from actual flow data and the criticism of Young (1986) still applies. Despite these limitations, the Muskingum–Cunge scheme is still frequently used for both flood routing (Barry and Bajracharya, 1995; Tang *et al.*, 1999a,b; Franchini *et al.*, 1999; Litrico, 2001) and catchment modeling (Koussis *et al.*, 1998; Takeuchi *et al.*, 1999; Lange *et al.*, 1999; Coe *et al.*, 2002) problems.

Flood Routing for Overbank Flows

Whilst it has been demonstrated that for overbank flows the one-dimensional flow field assumption breaks down, for many long reach flood-routing problems use of higher dimensional hydraulic codes may be both computationally prohibitive and unnecessary. This is particularly so if the additional distributed predictions generated by the higher order code are not required for the application in hand. However given, we are often interested in routing relatively large flood events, overbank flow is likely to be encountered. Here, hydrologists have three main options. First, one can assume that the additional approximations involved in continuing to treat out-of-bank flow as if it were one-dimensional are small compared to other uncertainties (for a discussion see Ali and Goodwin, 2002). Second, one can attempt to correct one-dimensional flow routing methods to account for the additional energy losses and/or mass transfers. Lastly, one can use a simplified two-dimensional code that incorporates some, but not all, of the additional processes known to occur during overbank flow.

Horritt and Bates (2002) demonstrate for simulation of a flood event over a 60-km reach of the River Severn, UK, that one-dimensional, simplified two-dimensional, and full two-dimensional models perform equally well in simulating flow routing and inundation extent given uncertainties over inflow, topography, and validation data. This suggests that although gross assumptions are made regarding the flow physics incorporated in a one-dimensional model applied to out-of-bank flows, the additional energy losses can be compensated for using a calibrated effective friction coefficient.

The second class of approaches are reviewed by Knight and Shiono (1996, pp. 155–159) and are based on subdividing the channel in the streamwise direction and then calculating the conveyance in each section using uniform flow formulae. The subarea conveyances are then summed to give the total conveyance. Knight and Shiono (1996) identify three main variations on the channel division method which aim to simulate the channel-floodplain interaction more exactly. These are: (i) modification of the subarea wetted perimeters (Wormleaton *et al.*, 1982); (ii) calculation of discharge adjustment factors for each subarea based on a "coherence" concept (Ackers, 1993), and (iii) quantification of the apparent shear stresses on the subarea division lines (Knight and Hamed, 1984). However, these methods have been developed to estimate the depth-discharge rating curve at particular cross sections and are yet to be incorporated in standard flood-routing models. One exception here is the LISFLOOD-FF model of De Roo *et al.* (2001). Here, a correction of the Manning roughness value is applied to simulate the momentum exchange, which occurs across the shear layer between main channel and floodplain flows.

Lastly, recent advances in higher order modeling have led to the development of simplified two-dimensional codes (see e.g. Bladé *et al.*, 1994; Estrela and Quintas, 1994; Bechteler *et al.*, 1994; Romanowicz *et al.*, 1996; Bates and De Roo, 2000; Venere and Clausse, 2002; Dhondia and Stelling, 2002) that, because of their computational efficiency, may potentially be applied to long reach flood-routing problems. Such models typically treat in-channel flow with some form of the one-dimensional St. Venant equations, but treat floodplain flows as two-dimensional using a storage cell concept first described by Cunge *et al.* (1980). Here the floodplain is discretized as a series of regions, with flows between regions calculated using analytical uniform flow formulae such as the Manning equation.

Initially, this approach was implemented in standard one-dimensional river routing packages such as ISIS (Wicks *et al.*, in press) by defining the storage cells as large polygonal areas (surface areas of $\sim 10^0 - 10^1\,\mathrm{km}^2$) representing discrete flooding compartments (e.g. polders, storage basins etc.) that are subjectively identified by the user. Recent developments in topographic data capture have, however, allowed high (cell size $\sim 10^0$ m) resolution Digital Elevation Models (DEMs) of floodplain areas to be produced (see the Section on "Data sources for inundation models and model data assimilation"). This has allowed storage cells to be discretized as a high-resolution grid, for example, the LISFLOOD-FP raster flood-routing model of Bates and De Roo (2000). LISFLOOD-FP treats in-channel flow using either kinematic (equations 4 and 5) or diffusion (equations 4 and 7) wave routing. Floodplain flows are similarly described in terms of continuity and mass flux equations, discretized over a grid of square cells, which allows the model to represent two-dimensional dynamic flow fields on the floodplain. Flow between two cells is assumed simply to be a function of the free surface height difference between those cells (Estrela and Quintas, 1994):

$$Q_x^{i,j} = \frac{h_{\mathrm{flow}}^{5/3}}{n}\left(\frac{h^{i-1,j} - h^{i,j}}{\Delta x}\right)^{1/2} \Delta y \qquad (17)$$

Water depths in each cell are then updated at each time step based on the sum of the fluxes over the four faces of the cell.

$$\frac{\mathrm{d}h^{i,j}}{\mathrm{d}t} = \frac{Q_x^{i-1,j} - Q_x^{i,j} + Q_y^{i,j-1} - Q_y^{i,j}}{\Delta x \Delta y} \qquad (18)$$

where $h^{i,j}$ is the water free surface height at the node (i, j), Δx and Δy are the cell dimensions, n is the effective grid scale Manning's friction coefficient ($\mathrm{L}^{1/3}\mathrm{T}^{-1}$) for the floodplain, and Q_x and Q_y describe the volumetric flow rates between floodplain cells. Q_y is defined analogously to equation (6). The flow depth, h_{flow}, represents the depth through which water can flow between two cells, and is defined as the difference between the highest water free surface in the two cells and the highest bed elevation (this definition has been found to give sensible results for both wetting cells and for flows linking floodplain and channel cells.) This approach is similar to diffusive wave propagation, but differs marginally due to the decoupling of the x- and y-components of the flow. Whilst this approach does not accurately represent diffusive wave propagation, it is computationally simple and has been shown to give very similar results to a more accurate finite difference discretization of the diffusive wave equation (Horritt and Bates, 2001a).

Equation (17) is also used to calculate flows between floodplain and channel cells, allowing floodplain cell depths to be updated using equation (18) in response to flow from the channel. These flows are also used as the source term, q, in the channel flow submodel (see equation 4), effecting the linkage of channel and floodplain flows. Thus, only mass transfer between channel and floodplain is represented, and this is assumed to be dependent only on relative water-surface elevations. Whilst this neglects effects such as channel-floodplain momentum transfer and the effects of advection and secondary circulation on mass transfer, it is the simplest approach to the coupling problem and should reproduce the dominant behavior of the real system.

The channel in the LISFLOOD-FP model occupies no floodplain pixels, but instead represents an extra flow path between pixels lying over the channel. Thus floodplain pixels lying over the channel have two water depths associated with them: one for the channel and one for the floodplain itself. This new scheme (referred to as the *Near Channel Floodplain Storage,* or NCFS model by Horritt and Bates, 2001b) has proved more suitable for situations where large floodplain grid spacings are used in conjunction with a narrow (width $< \Delta x$) channel since channel width and raster grid size are decoupled. Since the NCFS scheme will also calculate floodplain flows between cells occupied by the channel, extra flow routing in near channel regions will also be represented.

Such cellular approaches are capable of high-resolution application to relatively long river reaches up $\sim 10^2$ km in length and may provide an alternative to one-dimensional codes for channel routing in rainfall-runoff models. One such application is described by Coe *et al.* (2002) who model the whole of the Amazon basin using a regular grid of resolution 9 km with Muskingham–Cunge routing in-channel and floodplain flow treated using a storage cell approach. Similarly, the LISFLOOD-FP storage cell code has been successfully used to model flow routing along a 60-km reach of the River Severn in the United Kingdom (see Horritt and Bates, 2002) and a 35-km reach of the River Meuse in The Netherlands (see De Roo *et al.*, 2003 and Figure 9).

INUNDATION PREDICTION MODELS

In addition to modeling the downstream translation and attenuation of a flood wave, hydrologists may also be required to predict distributed hydraulic quantities, such as inundation extent, for use by catchment managers and civil protection authorities. Potentially this may be a difficult problem, as it requires prediction of the extension and retreat of a flow over large and relatively flat areas. Small errors in predicted water-surface elevation may result in large errors in the position of the predicted inundation front, and comparison of hydraulic models to inundation extent data may therefore be a sensitive test of their abilities. This task requires not only accurate floodplain topographic data but also models capable of treating dynamic wetting and

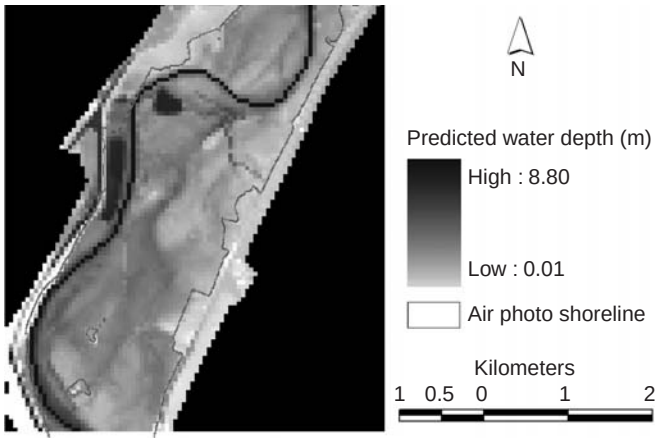


Figure 9 Predicted water depths and inundation extent simulated using the LISFLOOD-FP two-dimensional storage cell model applied at 50 m resolution to a 35 km river-floodplain reach of the River Meuse, The Netherlands over 20 days of a 1 in 63-year recurrence interval flood event that occurred in January 1995. Observed flow at the gauging station from the 19 January onwards is used as a model boundary condition and predicted inundation is compared to that observed by aerial photography taken on 27 January. The model correctly classifies as wet or dry 85% of pixels in the above image. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

drying in a numerically stable fashion (see Lynch and Gray, 1980 for a discussion).

Whilst floodplain flow is clearly a two-dimensional phenomena (as the flow paths cannot be predicted *a priori*, see Nicholas and Mitchell, 2003), overbank flow can still be treated using one-dimensional models as noted in the Section on "Flood routing for overbank flows". Such codes predict water depth at each model cross section at each time step, and it is then a relatively simple matter to calculate the predicted inundation extent by reprojecting the predicted water-surface elevations onto a DEM (see e.g. Puech and Raclot, 2002). The success of this procedure therefore depends strongly on the quality of the DEM, but with the advent of high-resolution, high-accuracy topographic data capture techniques (see the Section on "Data sources for inundation models and model data assimilation") this is becoming less of a problem. For example, Werner (2001) demonstrates a simple GIS method to generate inundation extent and depth maps from the predictions of a standard one-dimensional code. Similarly, Horritt and Bates (2002) show that with such an interpolation step in place, the performance of a one-dimensional code in predicting inundation extent for a 60-km reach of the River Severn, UK, was indistinguishable from the performance of two-dimensional storage cell and full two-dimensional models.

The success of such interpolation methods indicates that in certain situations one may not even need a model at all to predict inundation extent. Given gauged water-surface elevations along a reach, or water-surface elevations predicted on the basis of flood frequency analysis, one can perform a similar interpolation to that used by Werner (2001). This approximates the flood wave as a plane (or series of planes) which are intersected with the DEM to give extent and depth predictions. Bates and De Roo (2000) compared this method to two-dimensional storage cell and full two-dimensional hydraulic models for a 35-km reach of the River Meuse, The Netherlands and found that in certain situations the planar approximation performed almost as well as hydraulic modeling. In this application, the methods were used to simulate the January 1995 event ($Q_{\text{peak}} = \sim 2700\,\text{m}^3\,\text{s}^{-1}$, 1 in 63-year recurrence interval) and validated by comparison to air photo imagery of flood extent taken at around the time of maximum flooding. For regions close to gauging stations where the recorded water level was used as a height control, the planar approximation achieved accuracy of 81% pixels correctly predicted as wet or dry, compared to 85% for a two-dimensional storage cell model. Away from the gauging station, the planar approximation performed less well, and using a hydraulic model produced a better result in all circumstances examined, even if at times this difference was marginal. Clearly, the planar approximation will work well for reaches that are short compared to the wavelength of the flood and where there are good gauged data to constrain the position of the plane. Even in these circumstances, however, lack of mass conservation will mean that the areas that are not hydraulically connected to the channel are predicted as flooded. Nevertheless, this may be a useful method under some circumstances, and provides a benchmark level of performance that all hydraulic models should exceed to be considered skillful.

Despite the potential success of planar and one-dimensional models, a general solution to the problem of predicting flood inundation extent requires a model with a dimensionality to match our perceptual understanding of the flow physics. As discussed previously, this requires at least a two-dimensional code (see Figure 10). Numerous classes of two-dimensional approaches have been developed, and these include storage cell codes, full solutions of the two-dimensional St. Venant or shallow water equations (Gee *et al.*, 1990; Feldhaus *et al.*, 1992; Bates *et al.*, 1998; Nicholas and Mitchell, 2003) and simplified shallow water models (see e.g. Molinaro *et al.*, 1994) where certain terms, such as inertia, are omitted from the controlling equations. The two-dimensional St. Venant equations are derived from the full three-dimensional Navier–Stokes equations (1 and 2) by first averaging each dependent variable in time to yield the Reynolds Averaged Navier–Stokes or (RANS) equations and then integrating over the flow depth (see Bates and Anderson, 2001, pp. 328–332 for a discussion). Reynolds averaging leads to the introduction of

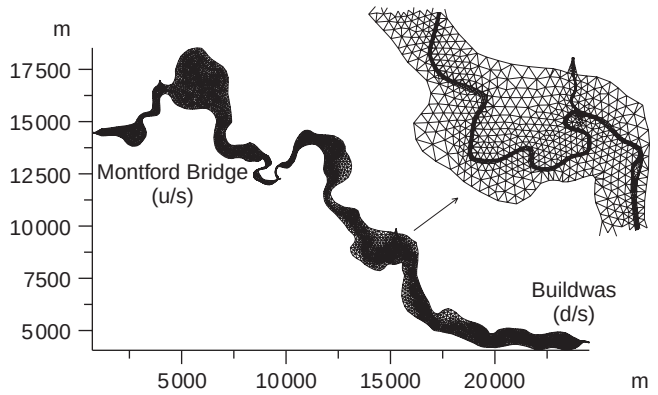


Figure 10 Two-dimensional unstructured mesh discretization constructed for a typical reach scale river-floodplain application. In this case the mesh was constructed by Horritt and Bates (2002) for use with the TELEMAC-2D finite element solution of the two-dimensional shallow water equations for the 60-km reach of the River Severn, UK, shown in Figure 6

new terms representing the shear stress on the mean flow due to turbulence. These so-called Reynolds stresses may be given, for example, as:

$$\bar{\tau}_{xz} = -\rho(\bar{u'}\bar{v'}) \tag{19}$$

where $\bar{\tau}_{xz}$ is the Reynolds stress in the $x - z$ plane.

As these depend on the instantaneous velocity fluctuations, their existence is a problem for a time-averaged model as these values are unknown. To close the RANS equations we need to provide values for the Reynolds stresses by introducing some model of turbulence. A variety of schemes for these are available but the most simple are those based on the Boussinesq approximation and this approach is therefore favored in flood inundation codes given the typical scale of the computational domain. This method treats turbulent momentum transfer in a similar manner to the approach adopted for viscous forces, as both can be thought of as shear stresses. Here a turbulent viscosity coefficient μ_t, is introduced with dimensions similar to the molecular viscosity μ. The overall fluid shear stress, τ ($ML^{-1}T^{-2}$) including molecular and turbulent components then becomes:

$$\tau = (\mu + \mu_t)\frac{d\bar{u}}{dy} \tag{20}$$

As with molecular viscosity, for convenience μ_t is often expressed in kinematic form, v_t, with dimensions (L^2T^{-1}). Simple models of turbulence, such as the zero equation, mixing length, one equation and k-ε (two equation) schemes, merely attempt to provide a value for v_t. The two-dimensional St. Venant may then be given in nonconservative form as:

Continuity equation

$$\frac{\partial h}{\partial t} + \vec{u_d} \cdot \overrightarrow{grad}\,(h) + h\,div(\vec{u_d}) = 0 \tag{21}$$

Momentum equations

$$\frac{\partial u_d}{\partial t} + \vec{u_d} \cdot \overrightarrow{grad}\,(u_d) + g\frac{\partial h}{\partial x} - div(v_t \cdot \overrightarrow{grad}\,(u_d)) = S_x - g\frac{\partial Z_f}{\partial x} \tag{22}$$

$$\frac{\partial v_d}{\partial t} + \vec{u_d} \cdot \overrightarrow{grad}\,(v_d) + g\frac{\partial h}{\partial y} - div(v_t \cdot \overrightarrow{grad}\,(v_d)) = S_y - g\frac{\partial Z_f}{\partial y} \tag{23}$$

where u_d, v_d are the depth-averaged and time-averaged velocity components (with dimensions LT^{-1}) in the x and y Cartesian directions (L); Z_f is the bed elevation (L); v_t is the kinematic turbulent viscosity (L^2T^{-1}) and S_x, S_y are the source terms (friction, coriolis force, and wind stress).

Given appropriate boundary and initial conditions, equations (21) to (23) can then be solved numerically to obtain predictions of the water depth, h, and the two components of the depth-averaged velocity, u_d and v_d.

Both storage cell and numerical models can be discretized using structured (Bates and De Roo, 2000; Nicholas and Walling, 1997; Nicholas and Mitchell, 2003) or unstructured grids (Sen, 2002; Hervouet and Van Haren, 1996), and more complex schemes even allow the grid to deform to follow the moving inundation front through time (Lynch and Gray, 1980; Kawahara and Umetsu, 1986; Benkhaldoun and Monthe, 1994). Fixed grid approaches require additional algorithms to treat spurious mass and momentum conservation effects that occur in partially wet cells at the flow field boundary (King and Roig, 1988; Leclerc *et al.*, 1990; Defina *et al.*, 1994; Ip *et al.*, 1998; Tchamen and Kawahita, 1998; Bates and Hervouet, 1999; Defina, 2000). Given the computational cost of remeshing and potential problems in maintaining numerical stability, to date fixed grid approaches have been preferred. This also has the advantage that it allows use of a more resolved spatial discretization and thus a greater degree of topographic complexity to be included in the model. This is clearly critical, as topography is a major control on the inundation process.

Lastly, for relatively short river-floodplain reaches (<5 km) three-dimensional approaches may be used. Here either the three-dimensional RANS equations or the three-dimensional shallow water equations (Hervouet and Van Haren, 1996, p. 191) are solved for a structured or unstructured grid using either finite difference, finite element or finite volume techniques. Correct grid spacing

in the vertical is typically obtained through the σ-transform (see Hervouet and Van Haren, 1996, p. 206) to give a normalized thickness of vertical layers irrespective of bottom topographic variations. One such application is reported by Stoesser *et al.* (2003) who report the application of a three-dimensional RANS model to a 3.5 km channel-floodplain reach of the River Rhine in Southwest Germany. The governing equations were solved for a general three-dimensional geometry, discretized by the finite volume method on curvilinear coordinates. The study used a relatively high-resolution mesh of 1 98 144 cells comprising $258 \times 64 \times 12$ cells in the streamwise, cross-streamwise, and vertical directions respectively (Figure 11), giving an approximate cell size of 13 m by 3 m by 0.5 m. The model successfully reproduced reach scale floodplain velocity observations determined from dye tracing experiments taken during a 1 in 100-year recurrence interval flow with discharge of $\sim 3600\,m^3\,s^{-1}$ and water levels for an independent verification event when discharge equaled $\sim 2400\,m^3\,s^{-1}$. The study provided evidence of the computational feasibility of a three-dimensional approach for river reaches of this scale for steady-state flows and demonstrated the range of hydraulic quantities that can be calculated (see Figure 12). Three-dimensional approaches may suffer from stability problems during dynamic flows as cell height-length ratios become highly distorted as water depth h approaches zero. Solution of this problem requires deformation of the three-dimensional mesh that may be computationally prohibitive.

Higher order turbulence modeling approaches for three-dimensional flow in compound channels have also been attempted and include algebraic stress (Shao *et al.*, 2003) and Large Eddy Simulation (Thomas and Williams, 1995; Bradbrook *et al.*, 2000) schemes. However these approaches incur large computational costs and have yet to be applied to anything other than experimental channels of regular geometry. Full field-scale models of complex natural geometry will no doubt become possible in the future, but may be over-specified for the problem in hand given the quality of the available validation data.

DATA SOURCES FOR INUNDATION MODELS AND MODEL DATA ASSIMILATION

The above discussion has made clear the variety of models available for flow routing and inundation prediction. Choice of model has been shown to depend on the scale of the problem, the computational resources available and the needs of the user. However, any model is only as good as the data used to parameterize, calibrate, and validate it, and in this section the data sources available for flow routing and inundation models are discussed along with the methods available to assimilate these data into hydraulic models. The data required by any hydraulic model

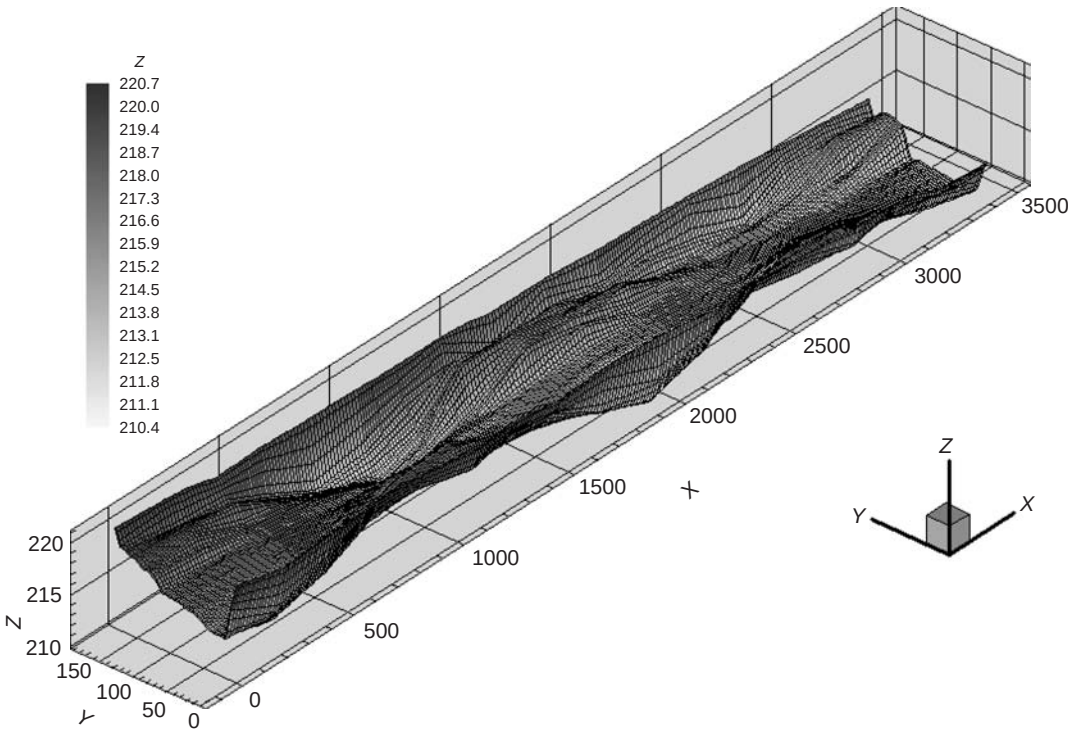


Figure 11 Three-dimensional computational grid used to simulate a 3.5-km reach of the River Rhine, Southwest Germany by Stoesser *et al.* (2003) © IWA Publishing

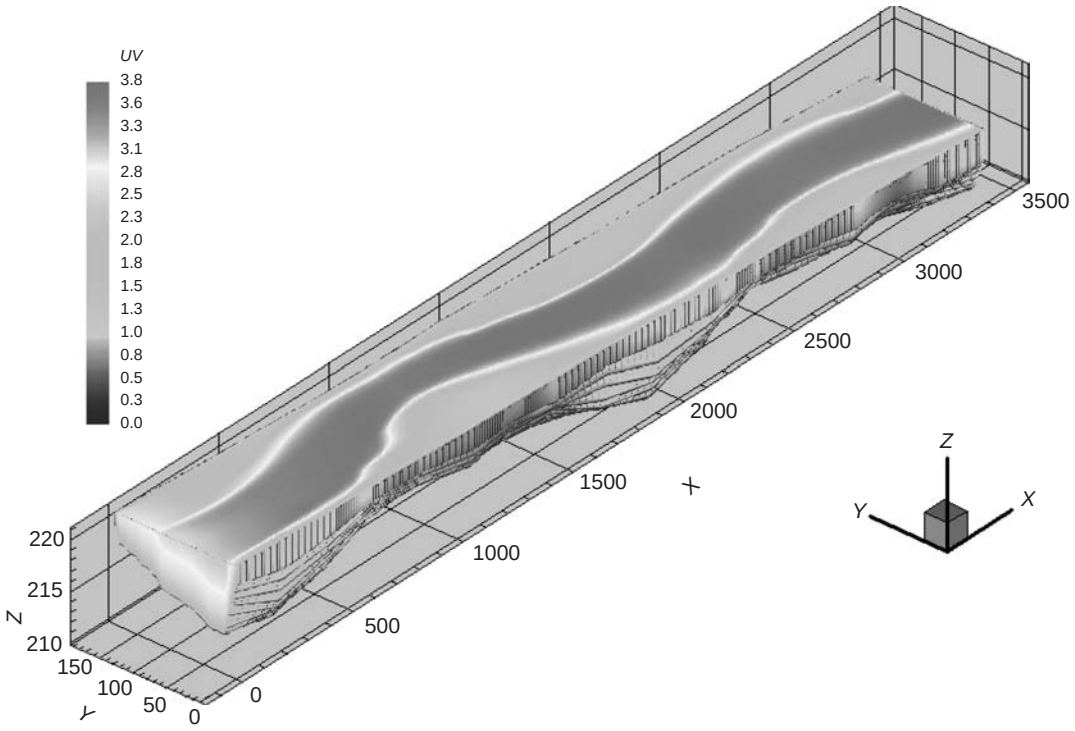


Figure 12 Scalar product of u and v velocities in ms^{-1} predicted by Stoesser *et al.* (2003) for the 3.5-km reach of the River Rhine shown in Figure 11 using a three-dimensional solution of the full Navier–Stokes equations with k-ε turbulence closure. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

are principally: (i) boundary condition data; (ii) initial condition data; (iii) topography data; (iv) friction data; and (v) hydraulic data for use in model validation.

Boundary Condition Data

Boundary condition specification in hydraulic models consists of assigning values for each dependent variable at each boundary node at each time step. Hence, for a two-dimensional model this would consist of values for the water depth, h, and the two components of the depth-averaged velocity vector, u_d and v_d. Typically, values for u_d and v_d are unavailable, and a rating curve is used to convert gauged stage values into discharge, Q, which in conjunction with a surveyed cross section can be used to calculate the velocity. If the rating curve was derived using the assumption of locally uniform flow at the boundary (i.e. with a uniform flow relationship such as the Manning equation) this introduces a further approximation into the model. For kinematic treatments and super-critical flows where $F_r > 1$ the absence of backwater effects means that the downstream boundary condition is unnecessary. Boundary condition data can also be taken from design floods calculated on the basis of flood frequency data or from the output of a rainfall-runoff model. Along the lateral boundary, exchanges with the adjacent catchment or floodplain (variable q in equation 4) are either set to zero, parameterized using data on average lateral inflow rates from previous floods for the given reach, or assumed proportional to the upstream inflow rate (e.g. O'Donnell, 1985).

Initial Condition Data

Initial conditions for a hydraulic model consist of values for each model dependent variable at each computational node at time $t = 0$. In practice, these will be incompletely known, if at all, and some additional assumptions will therefore be necessary. For steady-state simulations any reasonable guess at the initial conditions is usually sufficient, as the simulation can be run until the solution is in equilibrium with the boundary conditions and the initial conditions have ceased to have an influence. However, for dynamic simulations this will not be the case and whilst care can be taken to make the initial conditions as realistic as possible, a "spin up" period during which model performance is impaired will always exist. For example, initial conditions for a flood simulation in a compound channel are often taken as the water depths and flow velocities predicted by a steady-state simulation with inflow and outflow boundary conditions at the same value as those used to commence the dynamic run. Whilst most natural system are rarely in steady state, careful selection of simulation periods to

coincide their start with near steady-state conditions can minimize the impact of this assumption.

Topography Data

A conclusion that can be drawn from the studies discussed above is that the key data set for flow routing and inundation modeling is topography. High-resolution, high-accuracy topographic data are essential to shallow water flooding simulations over low slope floodplains with complex microtopography, and such data sets are increasingly available from a variety of remotely mounted sensors. Traditionally, hydraulic models have been parameterized using ground survey of cross sections perpendicular to the channel at spacings of between 100 and 1000 m. Such data are accurate to within a few millimeters in both the horizontal and vertical, and integrate well with one-dimensional hydraulic models. However, ground survey data are expensive and time consuming to collect and of relatively low spatial resolution. Moreover, topographic data available on national survey maps tends to be of low accuracy with poor spatial resolution in floodplain areas. For example, in the United Kingdom, nationally available contour data are only recorded at 5-m spacing to height accuracy of ± 1.25 m, and for a hydraulic modeling study of a typical river reach Marks and Bates (2000) report finding only three contours and 40 unmaintained spot heights within a $\sim 6\,km^2$ floodplain area. When converted into a DEM such data lead to relatively low levels of inundation prediction accuracy in hydraulic models (see Wilson and Atkinson, in press).

Considerable potential therefore exists for more automated, broad-area mapping of topography from satellite, and more importantly airborne platforms. Three techniques which currently show reasonable potential for flood modeling are aerial stereo-photogrammetry (Baltsavias, 1999; Lane, 2000; Westaway *et al.*, 2003), airborne laser altimetry or LiDAR (Krabill *et al.*, 1984; Gomes Pereira and Wicherson, 1999) and airborne Synthetic Aperture Radar (SAR) interferometry (Hodgson *et al.*, 2003). Radar interferometry from sensors mounted on space-borne platforms, and in particular the Shuttle Radar Topography Mission (SRTM) data (Rabus *et al.*, 2003), may also in the future provide a viable topographic data source for hydraulic modeling in large, remote river basin where the flood amplitude is large compared to the topographic data error. The characteristics of these instruments relevant to flood inundation modeling are compared in Table 1.

LiDAR in particular has attracted much recent attention in the hydraulic modeling literature (Marks and Bates, 2000; Bates *et al.*, 2003; French, 2003; Charlton *et al.*, 2003). Major LiDAR data collection programmes are underway in a number of countries, including The Netherlands and the United Kingdom, where so far approximately 20% of the land surface area in England and Wales has been surveyed. In the United Kingdom, helicopter-based

Table 1 Characteristics of remotely sensed topographic data relevant to flood inundation modeling

Type	Typical sensor altitude	Spatial resolution	Vertical accuracy	Survey rate	Approximate relative cost per km^2
Airborne stereo-photogrammetry	~1000 m	User defined, only partially automated	5–20-cm rmse depending on photo scale	Depends on the point spacing	1
Airborne laser altimetry (LiDAR)	~1000 m	~2 m	10–30-cm rmse	~50 km^2 per day	~0.1
Airborne synthetic aperture radar interferometry	~6000 m	~5 m	0.3–3.0-m rmse	~500 km^2 per day	~0.01
Satellite synthetic aperture radar interferometry (ERS, RADARSAT, ENVISAT, etc.)	~800 km	12–25 m	2–10-m rmse	Global repeat coverage every 24–35 days depending on the satellite	–
SRTM	~230 km	30 m	~5-m rmse ~2-m rmse pixel-to-pixel variations	Data collected over an 11-day Space Shuttle mission in February 2000 targeting 80% of earth's land mass (between 60°N and 56°S). Acquired at a rate of ~10.6 M km^2 per day	~0.0003

LiDAR survey is also beginning to be used to monitor in detail (~0.3-m spatial resolution) along critical topographic features such as flood defences, levees, and embankments. LiDAR systems operate by emitting pulses of laser energy at very high frequency (~5–33 KHz) and measuring the time taken for these to be returned from the surface to the sensor. Global Positioning System data and an onboard Inertial Navigation System are used to determine the location of the aircraft in space and hence the surface elevation. As the laser pulse travels to the surface, it spreads out to give a footprint of $\sim 0.1\,m^2$ for typical operating altitude of ~800 m. On striking a vegetated surface, part of the laser energy will be returned from the top of the canopy and part will penetrate to the ground. Hence, an energy source emitted as a pulse will be returned as a waveform, with the first point on the waveform representing the top of the canopy and the last point (hopefully) representing the ground surface. The last returns can then be used to generate a high resolution "bare earth" DEM (see Figure 13).

In general, all the above sensors only monitor topography above the water surface. Subsurface bathymetry is also important and can be monitored by either traditional ground survey at low water, boat surveys or for large water bodies, sidescan sonar (e.g. Howe *et al.*, 2001). Bathymetric and topographic data then require careful "stitching" together to create a single terrain model.

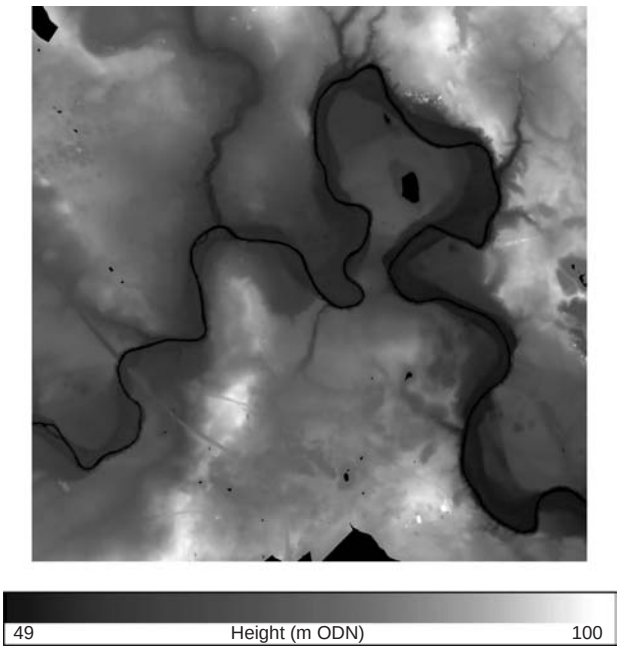


Figure 13 Digital Elevation Model at 10-m resolution derived from LiDAR data for a 6×6 km region of the River Severn floodplain in the United Kingdom using the processing algorithm of Cobby *et al.* (2003)

Friction Data

Friction is usually the only unconstrained parameter in a hydraulic model. Two- and three-dimensional codes, which use a zero-equation turbulence closure may additionally require specification of an "eddy viscosity" parameter which describes the transport of momentum within the flow by turbulent dispersion; however, this prerequisite disappears for most higher order turbulence models of practical interest. Hydraulic resistance is a lumped term that represents the sum of a number of effects: skin friction, form drag, and the impact of acceleration and deceleration of the flow. These combine to give a lumped energy loss coefficient, which in hydraulics is usually expressed in terms of resistance coefficients such as Manning's n and Chezy's C, and are derived from uniform flow theory. This assumes that the rate of energy dissipation for nonuniform flows is the same as it would be for uniform flow at the same water- surface (friction) slope. The precise effects represented by the friction coefficient for a particular model depend on the model's dimensionality and resolution. Thus, the extent to which form drag is represented depends on how the channel cross-sectional shape, meanders, pool-riffle sequences (see Beven and Carling, 1992) and long profile are incorporated into the model discretization. For example, a one-dimensional code will not include frictional losses due to channel meandering in the same way as a two-dimensional code. In the one-dimensional code, these frictional losses need to be incorporated into the hydraulic resistance term. Similarly, a high-resolution discretization will represent explicitly a greater proportion of the form drag component than a low-resolution discretization using the same model. Complex questions of scaling and dimensionality hence arise which may be somewhat difficult to disentangle.

Certain components of the hydraulic resistance are, however, more tractable. In particular, skin friction for in-channel flows is a strong function of bed material grain size, and a number of relationships exist which express the resistance coefficient in terms of the bed material median grain size, D_{50} (e.g. Hey, 1979). Equally, on floodplain sections where conveyance rather than storage processes dominate the drag due to vegetation is likely to form the bulk of the resistance term (Kouwen, 2000). Determining the drag coefficient of vegetation is, however, rather complex, as the frictional losses result from an interaction between plant biophysical properties and the flow. For example, at high flows the vegetation momentum absorbing area will reduce due to plant bending and flattening. Moreover, vegetation is highly spatially variable at scales typically smaller than the model grid, and correct treatment of features such as hedges will be important in assigning an effective roughness coefficient. To account for such effects, Kouwen and Li (1980) and Kouwen (1988) calculated the Darcy–Weisbach friction factor f for short

vegetation, such as floodplain grasses and crops, by treating such vegetation as flexible, and assuming that it may be submerged or nonsubmerged. f is dependent on water depth and velocity, vegetation height, and a product *MEI*, where M is the number of stems per unit area, E is the stem modulus of elasticity and I is the stem area's second moment of inertia. Whilst *MEI* often cannot be measured directly, it has been shown to correlate well with vegetation height (Temple, 1987).

Similar to topography, ground survey of grain size and vegetation parameters is extremely time consuming and, recently, research has begun to consider the use of remotely sensed techniques to determine certain of the above data. For example, photogrammetric techniques to extract grain size information from ground-based or airborne photography are currently under development (Butler *et al.*, 2001). Similarly, recent research developments allow extraction of specific plant biophysical parameters from LiDAR data. For vegetation >4 m in height, this latter technique uses the timing difference between the first and last points on the returned LiDAR waveform to determine the height of the canopy. For vegetation <4 m the method uses the local standard deviation of the last return heights (see Figure 14). Cobby *et al.* (2000, 2001) demonstrate that the height of short vegetation up to 1.4 m high can be estimated with such a technique to 0.14 m rmse. Taller vegetation (>10 m) is subject to greater height estimation error ($\sim 2-3$ m) as canopies are typically denser and it is less likely that the laser pulse will penetrate the full depth of the canopy, however, for the purpose of determining hydraulic resistance this is less of a problem. Given that other plant biophysical properties, for example, *MEI* correlate with plant height, Mason *et al.* (2003) have presented a methodology to calculate time and space distributed friction coefficients for flood inundation models directly from LiDAR data. Much further work is required in this area; however, such studies are beginning to provide methods to explicitly calculate important elements of frictional resistance for particular flow routing problems. This leads to the prospect of a reduced need for calibration of hydraulic models and therefore a reduction in predictive uncertainty (see Section "Uncertainties in flood-routing and inundation prediction"). Given the complexity of the flow being simulated, calibration of local effective friction values may still improve model results, but the methodology outlined above may be useful in constraining uncertainty in prior estimates of parameter values.

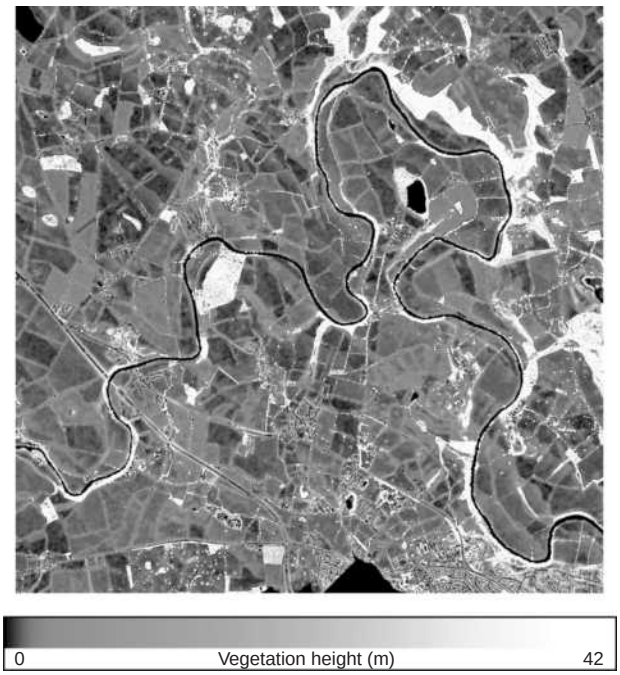


Figure 14 Vegetation height map at 10-m resolution derived from LiDAR data for a 6×6 km region of the River Severn floodplain in the United Kingdom using the processing algorithm of Cobby *et al.* (2003)

Model Validation Data

Traditionally, hydraulic models of reach scale flood inundation are validated using either water levels or discharge from maintained national gauging stations (Gee *et al.*, 1990; Bates *et al.*, 1992; Feldhaus *et al.*, 1992; Bates *et al.*, 1998). More rarely, individual flood levels are recorded during major events, although the reliability of such records can be questioned they have been relatively less used in the validation process. The spacing of national gauging stations is defined by their flood warning role and these are typically between 10 and 40 km apart. Such networks were never designed with hydraulic model validation in mind and the low density means that very few data points are available internal to a reach scale model domain. Whilst such data, particularly stage, are accurate, they are only 1D in time and 0D in space and do not directly test a model's ability to predict distributed hydraulics. Moreover, prior to the further development of the techniques for resistance parameter determination, calibration of hydraulic models by manipulating the friction coefficient will remain a necessity. Calibration and validation against bulk routing data can be simply achieved with spatially lumped friction parameters (Bates *et al.*, 1998) that represent the "average" frictional loss through the reach. In reality, friction parameters will vary in time and space, and many spatial patterns of frictional resistance will match the observed bulk flow data equally well yet give different distributed predictions. Such equifinality can only be reduced through the use of distributed validation data such as inundation extent, water depths, and flow velocities. This needs to cover a broad area with many points, and it again gives an advantage to remotely sensed data capture.

Inundation data provides a potential solution to the problem of hydraulic model validation as whilst these data

are 0D in time their 2D spatial format could provide "whole reach" data for both distributed calibration and validation of distributed predictions. Inundation extent can also be seen as a sensitive test of a hydraulic model, as small errors in predicted water-surface elevation would lead to large errors in shoreline position over flat floodplain topography. As well as data from post-flood ground surveys, such data can also be obtained using remote sensing. The sensors available for this task are reviewed by Bates *et al.* (1997), Pearson *et al.* (2001), and Bates *et al.* (2004) and are:

- Optical imagers on airborne and satellite platforms.
- SAR imagers on airborne and satellite platforms.
- Digital video cameras mounted on surveillance aircraft.

Processing such data, particularly SAR, to yield accurate shoreline information is not straightforward (see e.g. Horritt *et al.*, 2001 and Figure 15), and tends to be prone to misclassification. Yet, such data are at least wide area and distributed, and are therefore increasingly used in hydraulic model validation (see Horritt, 2000).

Despite the utility of inundation extent information, Mason *et al.* (2003) identify a situation where discrimination between competing models is impossible with satellite-derived SAR data but would be relatively easy given information on distributed velocities across the floodplain. Present methods of determining water velocities rely on making in-situ point measurements, which are necessarily limited in spatial extent, can be dangerous to undertake and difficult to interpret in terms of the time and space averaged velocities predicted by a given model. A resolution to this problem is potentially provided by Microwave Doppler radar. When the transmitted radar signal is scattered from a rough water surface, a Doppler frequency shift is produced by the centimeter-length surface waves that backscatter. The magnitude of the shift is a function of the stream current at the surface. The principle has been used to map surface velocities of coastal currents for a number of years (e.g. Hwang *et al.*, 2000), and has recently been applied to the measurement of the spatial distribution of river currents using a radar mounted on a van parked next to a river (Costa *et al.*, 2000). By aiming the radar across the river in two directions about 30 degrees apart, one looking upstream and the other downstream, Costa *et al.* found that the difference in the Doppler shifts of the two return signals gave the downstream surface velocity component. However, similar to velocity meter data, surface velocities must still be related to the time and space averaged velocity components predicted by a model.

Experiments are also now starting to be conducted on the use of along-track airborne radar interferometry for measuring water-surface velocities (Srokosz, 1997). These employ aircraft having microwave coherent real-aperture radar with two antennae aimed a few degrees apart in the along-track direction. If the two radars are arranged to look across- rather than along-track, it is possible to perform across-track interferometry instead, allowing a map of water-surface elevations to be constructed remotely. Such studies are still at the research stage and a useable technique is still some way off. Nevertheless, the model validation potential of such surface current measurements is considerable and deserving of further exploration.

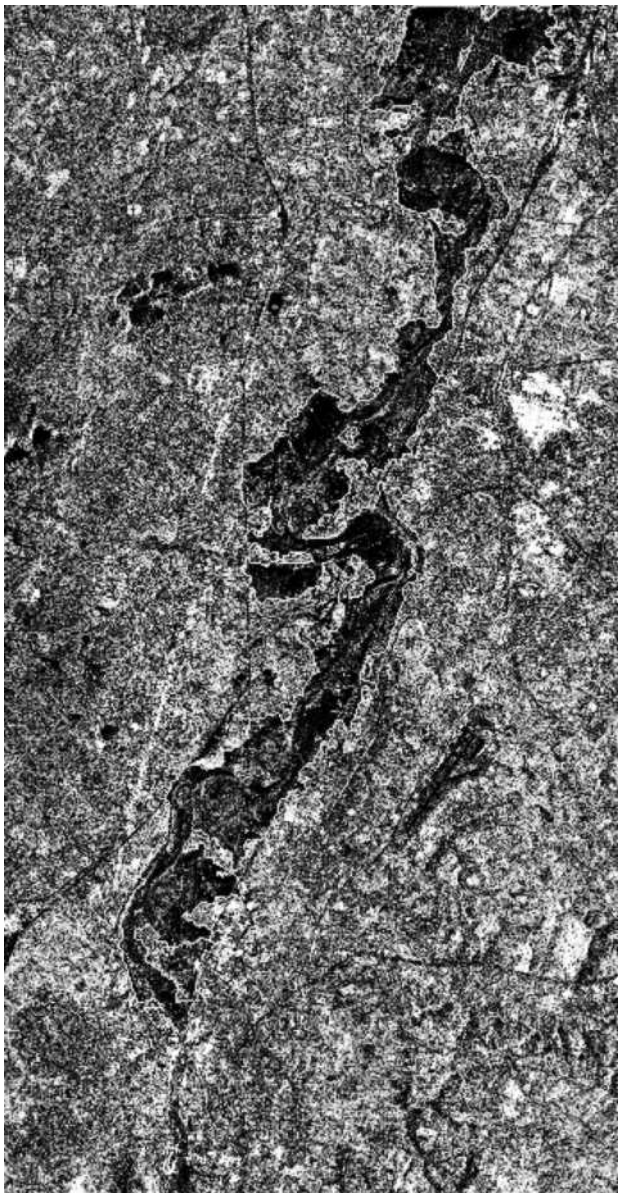


Figure 15 Satellite Synthetic Aperture Radar image of the January 1995 flooding on the River Meuse, The Netherlands. The image size is $\sim 70 \times 30$ km and the waterline extracted from the data using the statistical active contour processing algorithm of Horritt *et al.* (2001) is shown. The high speckle and low signal to noise ratio of this data leads to difficulties in accurate waterline determination

Model Data Assimilation

Increasing use of the above remote sensing techniques has caused a rapid shift in hydraulic modeling from a data-poor to a data-rich and spatially complex modeling environment with attendant possibilities for model testing and development. Despite an increase in computational power, the relative resolution of model and topography data has now reversed for most codes typically used to simulate flood inundation at the reach scale (see Bates and De Roo, 2000 for a review). A newly emergent research area is therefore how to integrate such massive data sets with lower resolution numerical inundation models, in an optimum manner that makes maximum use of the information content available. This is the direct opposite to the problem that most environmental modelers have traditionally faced (see Grayson and Blöschl, 2001 for a general discussion).

For example, Marks and Bates (2000) describe the integration of a LiDAR data set with a two-dimensional hydraulic model where the unstructured mesh discretization was derived independent of the topography. The topography was then assimilated into the model in an *a posteriori* step using weighted nearest neighbor interpolation to assign an elevation value to each computational node. This is typical of finite element mesh construction in many fields, but may not produce a mesh that captures those attributes of the original surface that are critical to the modeling problem in hand and may also lead to high data redundancy. To overcome these problems, Bates *et al.* (2003) describe a processing chain for high-resolution data assimilation into a lower resolution unstructured model grid. This consists of: (i) variogram analysis to determine significant topographic length scales in the data; (ii) identification of topographically significant points in this data set; (iii) incorporation of these points into an unstructured model grid that provides a quality solution for the relevant numerical solver; and (iv) uses the data left over from the mesh generation process to parameterize the Bates and Hervouet (1999) subgrid scale algorithm for dynamic wetting and drying. This method is demonstrated for the case of LiDAR topographic data but is general to any data type or model discretization. Cobby *et al.* (2003) take this process further and develop an automatic mesh generator that produces an unstructured grid, refined according to vegetation features (hedges, stands of trees etc.) on the floodplain (see Figure 16) identified automatically from LiDAR.

Similarly for friction parameters, Mason *et al.* (2003) use an area-weighing method to calculate area effective frictional resistance from the high-resolution height information contained in LiDAR data (see Figure 17). This method aims to yield model parameters that are appropriate to the particular discretization used, rather than being scale and discretization independent although this has yet to be fully tested.

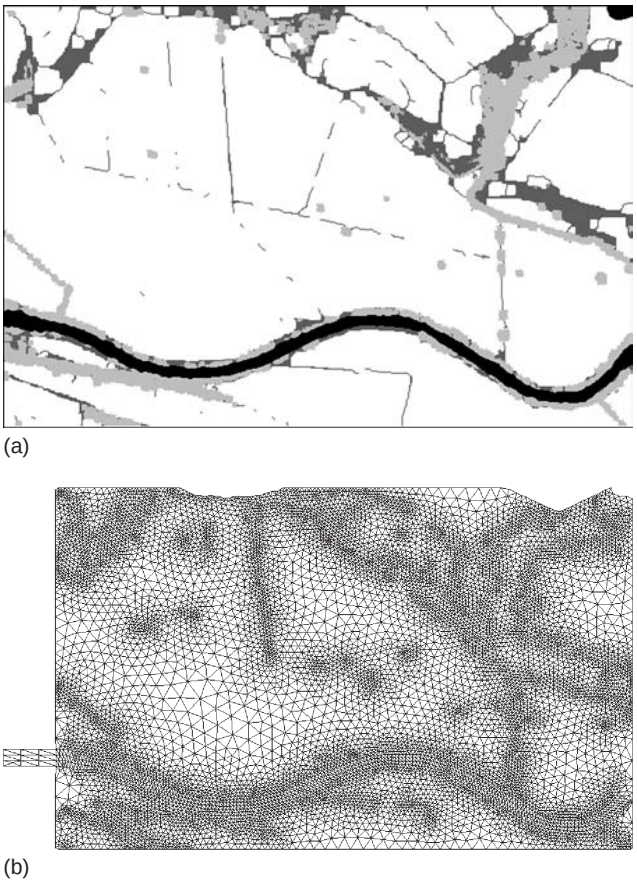


Figure 16 Results of the automatic unstructured two-dimensional mesh generator developed by Cobby *et al.* (2003). The software takes as input a LiDAR derived vegetation height map and produces an unstructured grid decomposed according to vegetation features (hedges, stands of trees, etc.) on the floodplain

UNCERTAINTIES IN FLOOD ROUTING AND INUNDATION PREDICTION

Flood inundation predictions used to aid design and planning decisions are typically derived from single realizations of numerical hydraulic models applied in a forward modeling framework (e.g. Anastasiadou-Partheniou and Samuels, 1998; Bates *et al.*, 1998). Rigorous calibration studies are undertaken to define a single parameter set that optimizes the model fit to some observed data. Typically, these data consist of flows or levels at a small set of discrete points, and only rarely are distributed flood extent data available that can be used to test the model over a wide area. If sufficient data (level or extent) are available, best practice is to undertake some form of split sample analysis, as advocated by Klemeš (1986) in the context of catchment hydrologic models, to test the validity of the optimized parameters for an independent set of observed data. The

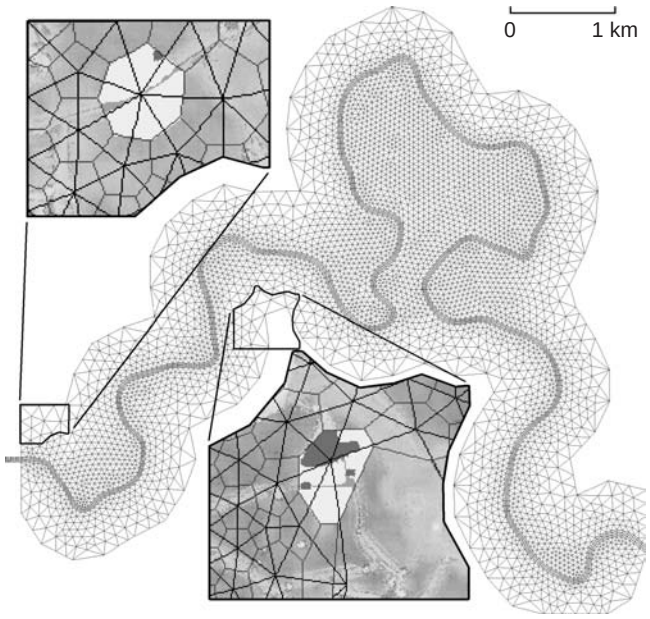


Figure 17 Assignment of an ''effective'' friction value for each element of a finite element mesh using the vegetation height data shown in Figure 14 and the friction mapping algorithm of Mason *et al.* (2003). For each mesh node, an instantaneous friction factor is calculated at each time step, given the frictional material in the neighborhood of the node and the current water depth and flow velocity there. A node's neighborhood is defined as a polygon whose vertices are the centroids of the elements surrounding the node. Subregions of connected regions in the vegetation height map are formed by intersecting the node polygon map with the vegetation height map. Each subregion in a polygon may contain either sediment or one of three vegetation height classes. If a subregion contains vegetation its region's average vegetation height is attributed to the subregion (after Mason *et al.*, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

success of this process requires that there is a single, well-defined optimum that can be identified by the calibration procedure and which is stationary between events. If this is not the case, a further period of calibration and testing is required, which at present is based on subjective intervention by a skilled operator (see the discussion in Werner, 2002). The identified single optimum is then used to predict flood extent and level for a design flood for which calibration and validation data are unlikely to be available. The result is a single deterministic prediction of flood extent for the design event, which is impossible to validate directly.

However, recent work using multiple simulations of different parameter sets (Romanowicz *et al.*, 1996; Romanowicz and Beven, 1998; Aronica *et al.*, 1998; Aronica *et al.*, 2002; and Romanowicz and Beven, 2003) has shown for a variety of flood prediction models that many combinations of model structure, input data, and parameters may fit sparse validation data equally well. This has been shown by modeling within a Monte Carlo framework, where an extensive sampling of the model parameter space reveals multiple acceptable (or "behavioral") solutions and even multiple discrete optima in different parts of the parameter space. By contrast, traditional calibration methodologies seek only a single optimum parameter set. Thus, when the parameter space is comprehensively searched, optima are unlikely to be well defined (e.g. Aronica *et al.*, 2002) and may occur at multiple points on the response surface (e.g. Horritt and Bates, 2001a).

Despite these difficulties, for any parameter identification problem, an optimum parameter set will exist. It may, however, be sensitive to errors in the input data and model definition and to the choice of performance measure used in the optimization. Monte Carlo simulation, whilst generally more computationally demanding, does allow the identification of multiple behavioral models with the advantage that using such multiple models in prediction allows a more meaningful estimation of the risk of flooding at any point on the floodplain during a future event. This is the basis for the Generalized Likelihood Uncertainty Estimation (GLUE) methodology originally developed by Beven and Binley (1992) and applied to the flood inundation problem by Romanowicz *et al.* (1996), Romanowicz and Beven (1998), Aronica *et al.* (2002), and Romanowicz and Beven (2003). In reality, the risk of flooding for a particular design event is better conceived as a "fuzzy" map in which there will be significant spatial structure. Hence, the uncertainty in flood risk at a given point on the floodplain can be thought of as the sum of input data uncertainty, model structural error, and parameterization uncertainty.

An example of this is provided by Aronica *et al.* (2002) who use the GLUE procedure to estimate the spatially distributed uncertainty in flood extent prediction, as a result of uncertainty over friction parameterization in the LISFLOOD-FP two-dimensional storage cell model. The binary pattern of wet/dry areas predicted for a 1 in 5-year flood event on a 3-km reach of the River Thames in the United Kingdom was compared to an observed inundation map obtained from a SAR image of the flood to yield a zero-dimensional global performance measure for each of 500 simulations in a Monte Carlo ensemble. To unearth the spatial uncertainty in model predictions for the particular flood being modeled, Aronica *et al.* (2002) took the flood state as predicted by the model for each pixel in each realization, and weighted it according to the measure of fit, $F^{\langle 2\rangle}$, between observed (A_{obs}) and modeled (A_{mod}) flooded areas:

$$F^{\langle 2\rangle} = \frac{A_{\text{obs}} \cap A_{\text{mod}}}{A_{\text{obs}} \cup A_{\text{mod}}} \tag{24}$$

for that simulation.

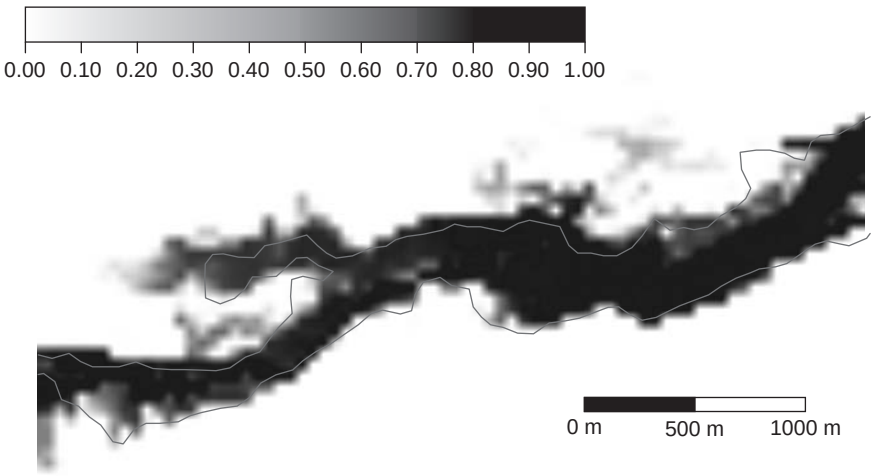


Figure 18 Probability map of predicted inundation P_i^{flood}, calculated using 500 realizations of the LISFLOOD-FP two-dimensional storage cell model for a 1 in 5-year recurrence interval event on a 3-km reach of the upper River Thames, UK. The observed shoreline derived from satellite SAR data is shown. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

This information could then be used to give a flood probability for each pixel i, P_i^{flood}.

$$P_i^{\text{flood}} = \frac{\sum_j f_{ij} F_j^{(2)}}{\sum_j F_j^{(2)}} \tag{25}$$

where f_{ij} takes a value of 1 for a flooded pixel and is zero otherwise and $F_j^{\langle 2\rangle}$ is the global performance measure for simulation j. P_i^{flood} will assume a value of 1 for pixels that are predicted as flooded in all simulations and 0 for pixels always predicted as dry. Model uncertainty (here defined by the interaction of the global performance measure and spatially distributed probabilities of the event being modeled) will manifest itself as a region of pixels with intermediate values, maximum uncertainty being indicated by pixels with $P_i^{\text{flood}} \approx 0.5$. The result is a distributed uncertainty measure P_i^{flood}, mapped over real space (see Figure 18). Such mapping can even be done in real time with updating of the probabilities using Bayes equation (Romanowicz and Beven, 1998). The advantage of the approach is that it both captures distributed uncertainty and contains information on global likelihood that can be used to condition predictions of further events for which observed data are not available. The method reveals the spatial structure in simulation uncertainty and simultaneously enables mapping of flood probability predicted by the model.

CONCLUSIONS

Standard methods of flood-routing and inundation extent modeling are currently undergoing critical re-evaluation and development as a result of the new opportunities afforded by remotely sensed data and the application of uncertainty analysis techniques in hydraulics. Future research in these areas is capable of yielding major improvements in our ability to parameterize and validate hydraulic models, and may lead to greater understanding of the dominant physical processes relevant to flood wave hydraulics at a variety of scales. However, uncertainty is unlikely to be eliminated and dealing with this challenge in a way that can be readily communicated to environmental managers and decision makers needs to be an integrated component of any hydraulic modeling study.

REFERENCES

Abbott M. and Ionescu F. (1967) On the numerical computation of nearly horizontal flows. *Journal of Hydraulic Research*, **5**(2), 97–117.

Ackers P. (1993) Stage-discharge functions for 2-stage channels – the impact of new research. *Journal of the Institution of Water and Environmental Management*, **7**(1), 52–61.

Ali S. and Goodwin P. (2002) The predictive ability of 1D models to simulate floodplain processes. In *Hydroinformatics 2002: Proceedings of the Fifth International Conference on Hydroinformatics, Model Development and Data Management, Vol. 1*, Falconer R.A., Lin B., Harris E.L. and Wilson C.A.M.E. (Eds.), IWA Publishing: London, pp. 247–252.

Anastasiadou-Partheniou L. and Samuels P.G. (1998) Automatic calibration of computational river models. *Proceedings of the Institution of Civil Engineers, Water Maritime and Energy*, **130**, 154–162.

Aronica G., Bates P.D. and Horritt M.S. (2002) Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrological Processes*, **16**, 2001–2016.

Aronica G., Hankin B.G. and Beven K. (1998) Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data. *Advances in Water Resources*, **22**, 349–365.

Ashworth P.J., Bennett S.J., Best J.L. and McLelland S.J. (1996) *Coherent Flow Structures in Open Channels*, John Wiley & Sons: Chichester, p. 733.

Baltsavias E.P. (1999) A comparison between photogrammetry and laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing*, **54**(2–3), 83–94.

Barry D.A. and Bajracharya K. (1995) On the Muskingum-Cunge flood routing method. *Environment International*, **21**(5), 485–490.

Bates P.D. and Anderson M.G. (2001) Validation of hydraulic models. In *Model Validation: Perspectives in Hydrological Science*, Anderson M.G. and Bates P.D. (Eds.), John Wiley & Sons: Chichester, pp. 325–356.

Bates P.D., Anderson M.G., Baird L., Walling D.E. and Simm D. (1992) Modelling floodplain flow with a two-dimensional finite element scheme. *Earth Surface Processes and Landforms*, **17**, 575–588.

Bates P.D. and De Roo A.P.J. (2000) A simple raster-based model for floodplain inundation. *Journal of Hydrology*, **236**, 54–77.

Bates P.D. and Hervouet J.-M. (1999) A new method for moving boundary hydrodynamic problems in shallow water. *Proceedings of the Royal Society of London, Series A*, **455**, 3107–3128.

Bates P.D., Horritt M.S., Cobby D. and Mason D. (2004) Flood inundation modelling using LiDAR and SAR data. In *Spatial Modelling of the Terrestrial Environment*, Kelly R., Drake N. and Barr S. (Eds.), John Wiley & Sons: Chichester, 79–106.

Bates P.D., Horritt M., Smith C. and Mason D. (1997) Integrating remote sensing observations of flood hydrology and hydraulic modelling. *Hydrological Processes*, **11**, 1777–1795.

Bates P.D., Marks K.J. and Horritt M.S. (2003) Optimal use of high-resolution topographic data in flood inundation models. *Hydrological Processes*, **17**, 5237–5557.

Bates P.D., Stewart M.D., Desitter A., Anderson M.G., Renaud J.-P. and Smith J.A. (2000) Numerical simulation of floodplain hydrology. *Water Resources Research*, **36**, 2517–2530.

Bates P.D., Stewart M.D., Siggers G.B., Smith C.N., Hervouet J.-M. and Sellin R.H.J. (1998) Internal and external validation of a two-dimensional finite element model for river flood simulation. *Proceedings of the Institution of Civil Engineers, Water Maritime and Energy*, **130**, 127–141.

Bechteler W., Hartmaan S. and Otto A.J. (1994) Coupling of 2D and 1D models and integration into Geographic Information Systems (GIS). In *Proceedings of the 2nd International Conference on River Flood Hydraulics*, White W.R. and Watts J. (Eds.), John Wiley & Sons: Chichester, pp. 155–165.

Benkhaldoun F. and Monthe L. (1994) An adaptive nine-point finite volume Roe scheme for two-dimensional Saint Venant equations. In *Modelling Flood Propagation over Initially Dry Areas*, Molinaro P. and Natale L. (Eds.), American Society of Civil Engineers: New York, pp. 30–44.

Beven K.J. (2001) *Rainfall-Runoff Modelling: The Primer*, John Wiley & Sons: Chichester, p. 360.

Beven K. and Binley A. (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K.J. and Carling P. (1992) Velocities, roughness and dispersion in the Lowland River Severn. In *Lowland Rivers: Geomorphological Perspectives*, Petts G.F. and Carling P. (Eds.), John Wiley & Sons: Chichester, pp. 71–93.

Beven K. and Wood E.F. (1993) Flow routing and the hydrological response of channel networks. In *Channel Network Hydrology*, Beven K. and Kirkby M.J. (Eds.), John Wiley & Sons: Chichester, p. 319.

Bladé E., Gómez M. and Dolz J. (1994) Quasi-two dimensional modelling of flood routing in rivers and flood plains by means of storage cells. In *Modelling of Flood Propagation over Initially Dry Areas*, Molinaro P. and Natale L. (Eds.), American Society of Civil Engineers: New York, pp. 156–170.

Bradbrook K.F., Lane S.N., Richards K.S., Biron P.M. and Roy A.G. (2000) Large Eddy simulation of periodic flow characteristics at river channel confluences. *Journal of Hydraulics Research*, **38**(3), 207–215.

Brath A., Montanari A. and Toth E. (2002) Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. *Hydrology and Earth System Science*, **6**(4), 627–639.

Bridge J.S. and Gabel S.L. (1992) Flow and sediment dynamics in a low sinuosity, braided river – Calamus river, Nebraska sandhills. *Sedimentology*, **39**(1), 125–142.

Burt T.P., Bates P.D., Stewart M.D., Claxton A.J., Anderson M.G. and Price D.A. (2002) Water table fluctuations within the floodplain of the River Severn, England. *Journal of Hydrology*, **262**, 1–20.

Butler J.B., Lane S.N. and Chandler J.H. (2001) Automated extraction of grain-size data from gravel surfaces using digital image processing. *Journal of Hydraulic Research*, **39**(5), 519–529.

Cameron D., Kneale P. and See L. (2002) An evaluation of a traditional and a neural net modelling approach to flood forecasting for an upland catchment. *Hydrological Processes*, **16**(5), 1033–1046.

Cappelaere B. (1997) Accurate diffusive wave routing. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **123**(3), 174–181.

Castro N.M. and Hornberger G.M. (1991) Surface-subsurface water interactions in an alluviated mountain stream channel. *Water Resources Research*, **27**, 1613–1621.

Chanson H. (1999) *The Hydraulics of Open Channel Flow: An Introduction*, Arnold: London, p. 495.

Charlton M.E., Large A.R.G. and Fuller I.C. (2003) Application of airborne LiDAR in river environments: The River Coquet,

Northumberland, UK. *Earth Surface Processes and Landforms*, **28**(3), 299–306.

Cobby D.M., Mason D.C. and Davenport I.J. (2001) Image processing of airborne scanning laser altimetry for improved river flood modelling. *ISPRS Journal of Photogrammetry and Remote Sensing*, **56**(2), 121–138.

Cobby D.M., Mason D.C., Davenport I.J. and Horritt M.S. (2000) Obtaining accurate maps of topography and vegetation to improve 2D hydraulic flood models. *Proceedings of the EOS/SPIE Symposium on Remote Sensing for Agriculture, Ecosystems, and Hydrology II*, Barcelona, 25–9 September, pp. 125–136.

Cobby D.M., Mason D., Horritt M.S. and Bates P.D. (2003) Two-dimensional hydraulic flood modelling using a finite element mesh decomposed according to vegetation and topographic features derived from airborne scanning laser altimetry. *Hydrological Processes*, **17**, 1979–2000.

Coe M.T., Costa M.H., Botta A. and Birkett C. (2002) Long-term simulations of discharge and floods in the Amazon Basin. *Journal of Geophysical Research – Atmospheres*, **107**(D20), Art. No. 8044.

Cooper D.M., House W.A., Reynolds B., Hughes S., May L. and Gannon B. (2002) The phosphorus budget of the Thame catchment, Oxfordshire: 2. Modelling. *Science of the Total Environment*, **282**, 435–457.

Costa J.E., Spicer K.R., Cheng R.T., Haeni F.P., Melcher N.B., Thurman E.M., Plant W.J. and Keller W.C. (2000) Measuring stream discharge by non-contact methods: a proof-of concept experiment. *Geographical Research Letters*, **27**(4), 553–556.

Cunge J.A. (1969) On the solution of the Muskingum method of flood routing equation. *Journal of Hydraulic Research*, **7**(2), 205–230.

Cunge J.A., Holly F.M. Jr and Verwey A. (1980) *Practical Aspects of Computational River Hydraulics*, Pitman: London, p. 420.

Defina A. (2000) Two-dimensional shallow flow equations for partially dry areas. *Water Resources Research*, **36**(11), 3251–3264.

Defina A., D'Alpaos L. and Matticchio B. (1994) A new set of equations for very shallow water and partially dry areas suitable to 2D numerical models. In *Modelling Flood Propagation over Initially Dry Areas*, Molinaro P. and Natale L. (Eds.), American Society of Civil Engineers: New York, pp. 72–81.

De Roo A.P.J., Bartholmes J., Bates P.D., Beven K., Bongioannini-Cerlini B., Gouweleeuw B., Heise E., Hils M., Hollingsworth M. Holst B. *et al.* (2003) Development of a European flood forecasting system. *International Journal of River Basin Management*, **1**(1), 49–59.

De Roo A., Odijk M., Schmuck G., Koster E. and Lucieer A. (2001) Assessing the effects of land use changes on floods in the Meuse and Oder catchment. *Physics and Chemistry of the Earth Part B – Hydrology, Oceans and Atmospheres*, **26**(7–8), 593–599.

Dhondia J.F. and Stelling G.S. (2002) Application of one-dimensional-two-dimensional integrated hydraulic model for flood simulation and damage assessment. In *Hydroinformatics 2002: Proceedings of the Fifth International Conference on Hydroinformatics*, *Model Development and Data Management, Vol. 1*, Falconer R.A., Lin B., Harris E.L. and Wilson C.A.M.E. (Eds.), IWA Publishing: London, pp. 265–276.

Ervine D.A. and Baird J.I. (1982) Rating curves for rivers with overbank flow. *Proceedings of the Institution of Civil Engineers Part 2 – Research and Theory*, **73**, 465–472.

Estrela T. and Quintas L. (1994) Use of GIS in the modelling of flows on floodplains. In *Proceedings of the 2nd International Conference on River Flood Hydraulics*, White H.R. and Watts J. (Eds.), John Wiley & Sons: Chichester, pp. 177–189.

Feldhaus R., Höttges J., Brockhaus T. and Rouvé G. (1992) Finite element simulation of flow and pollution transport applied to a part of the River Rhine. In *Hydraulic and Environmental Modelling; Estuarine and River Waters*, Falconer R.A., Shiono K. and Matthews R.G.S. (Eds.), Ashgate Publishing: Aldershot, pp. 323–344.

Franchini M., Lamberti P. and Di Giammarco P. (1999) Rating curve estimation using local stages, upstream discharge data and a simplified hydraulic model. *Hydrology and Earth System Science*, **3**(4), 541–548.

Fread D.L. (1984) Flood routing. In *Hydrological Forecasting*, Anderson M.G. and Burt T.P. (Eds.), John Wiley & Sons: Chichester, pp. 437–503.

Fread D.L. (1993) Flood routing. In *Handbook of Applied Hydrology*, Maidment D.R. (Ed.), Mc-Graw Hill: New York, pp. 10.1–10.36.

French J.R. (2003) Airborne LiDAR in support of geomorphological and hydraulic modelling. *Earth Surface Processes and Landforms*, **28**(3), 321–335.

Fukuoka S. and Fujita K. (1989) Prediction of flow resistance in compound channels and its application to the design of river courses. *Proceedings of the Japan Society of Civil Engineers*, **411**, 63–72.

Gee D.M., Anderson M.G. and Baird L. (1990) Large scale floodplain modelling. *Earth Surface Processes and Landforms*, **15**, 512–523.

Ghisalberti M. and Nepf H.M. (2002) Mixing layers and coherent structures in vegetated aquatic flows. *Journal of Geophysical Research – Oceans*, **107**(C2), Art. No. 3011.

Gomes Pereira L.M. and Wicherson R.J. (1999) Suitability of laser data for deriving geographical data: a case study in the context of management of fluvial zones. *Photogrammetry and Remote Sensing*, **54**, 105–114.

Graf W.H. and Altinakar M. (1998) *Fluvial Hydraulics: Flow and Transport Processes in Channels of Simple Geometry*, John Wiley & Sons: Chichester, p. 692.

Grayson R. and Blöschl G. (2001) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Cambridge University Press: Cambridge, p. 416.

Hankin B.G., Holland M.J., Beven K.J. and Carling P. (2002) Computational fluid dynamics modelling of flow and energy fluxes for a natural fluvial dead zone. *Journal of Hydraulic Research*, **40**(4), 389–401.

Havnø K., Madsen N. and Dørge J. (1994) MIKE11 – a generalised river modelling package. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Colorado, pp. 733–781.

Haycock N.E. and Burt T.P. (1993) Role of floodplain sediments in reducing the nitrate concentration of subsurface run-off: A case study in the Cotswolds, UK. *Hydrological Processes*, **7**, 287–295.

Hervouet J.-M. and Van Haren L. (1996) Recent advances in numerical methods for fluid flows. In *Floodplain Processes*, Anderson M.G., Walling D.E. and Bates P.D. (Eds.), John Wiley & Sons: Chichester, pp. 183–214.

Hey R.D. (1979) Flow resistance in Gravel Bed rivers. *American Society of Civil Engineers, Journal of Hydraulics Division*, **105**(4), 365–379.

Hodgson M.E., Jensen J.R., Schmidt L., Schill S. and Davis B. (2003) An evaluation of LIDAR- and IFSAR-derived digital elevation models in leaf-on conditions with USGS Level 1 and Level 2 DEMs. *Remote Sensing of the Environment*, **84**(2), 295–308.

Horritt M.S. (2000) Calibration of a two-dimensional finite element flood flow model using satellite radar imagery. *Water Resources Research*, **36**, 3279–3291.

Horritt M.S. and Bates P.D. (2001a) Predicting floodplain inundation: raster-based modelling versus the finite element approach. *Hydrological Processes*, **15**, 825–842.

Horritt M.S. and Bates P.D. (2001b) Effects of spatial resolution on a raster based model of flood flow. *Journal of Hydrology*, **253**, 239–249.

Horritt M.S. and Bates P.D. (2002) Evaluation of 1-D and 2-D numerical models for predicting river flood inundation. *Journal of Hydrology*, **268**, 87–99.

Horritt M.S., Mason D.C. and Luckman A.J. (2001) Flood boundary delineation from synthetic aperture radar imagery using a statistical active contour model. *International Journal of Remote Sensing*, **22**, 2489–2507.

Howe J.A., Overnell J., Inall M.E. and Wilby A.D. (2001) A side-scan sonar image of a glacially-overdeepened sea loch, upper Loch Etive, Argyll. *Scottish Journal of Geology*, **37**, 3–10.

Hwang P.A., Krabill W.B., Wright W., Swift R.N. and Walsh E.J. (2000) Airborne scanning lidar measurement of ocean waves. *Remote Sensing of the Environment*, **73**, 236–246.

Ip J.T.C., Lynch D.R. and Friedrichs C.T. (1998) Simulation of estuarine flooding and dewatering with application to Great Bay, New Hampshire. *Estuarine Coastal and Shelf Science*, **47**(2), 119–141.

Kawahara M. and Umetsu T. (1986) Finite element method for moving boundary problems in river flows. *International Journal of Numerical Methods in Fluids*, **6**, 365–386.

King I.P. and Roig L. (1988) Two-dimensional finite element models for floodplains and tidal flats. In *Proceedings of an International Conference on Computational Methods in Flow Analysis*, Niki K. and Kawahara M. (Eds.), Okayama, pp. 711–718.

Klemeš V. (1986) Operational testing of hydrologic simulation models. *Hydrological Sciences Journal*, **31**, 13–24.

Knight D.W. and Hamed M.E. (1984) Boundary shear in symmetrical compound channels. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **110**(10), 1412–1430.

Knight D.W. and Shiono K. (1996) River channel and floodplain hydraulics. In *Floodplain Processes*, Anderson M.G., Walling D.E. and Bates P.D. (Eds.), John Wiley & Sons: Chichester, pp. 139–182.

Kohane R. and Welz R. (1994) Combined use of FE models for prevention of ecological deterioration of areas next to a river hydropower complex. In *Computational Methods in Water Resources X*, Peter A., Wittum G., Meissner U., Brebbia C.A., Gray W.G. and Pinder G.F. (Eds.), Vol. 1, Kluwer: The Netherlands, pp. 59–66.

Koussis A.D., Smith M.E., Akylas E. and Tombrou M. (1998) Groundwater drainage flow in a soil layer resting on an inclined leaky bed. *Water Resources Research*, **34**(11), 2879–2887.

Kouwen N. (1988) Field estimation of the biomechanical properties of grass. *Journal of Hydraulic Research*, **26**(5), 559–568.

Kouwen N. (2000) Closure of 'effect of riparian vegetation on flow resistance and flood potential'. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **126**(12), 954.

Kouwen N. and Li R.M. (1980) Biomechanics of vegetative channel linings. *American Society of Civil Engineers, Journal of Hydraulics Division*, **106**(6), 713–728.

Krabill W.B., Collins J.G., Link L.E., Swift R.N. and Butler M.L. (1984) Airborne laser topographic mapping results. *Photogrammetric Engineering and Remote Sensing*, **50**, 685–694.

Lane S.N. (2000) The measurement of river channel morphology using digital photogrammetry. *Photogrammetric Record*, **16**(96), 937–957.

Lange J., Leibundgut C., Greenbaum N. and Schick A.P. (1999) A noncalibrated rainfall-runoff model for large, arid catchments. *Water Resources Research*, **35**(7), 2161–2172.

Leclerc M., Bellemare J.-F., Dumas G. and Dhatt G. (1990) A finite element model of estuarine and river flows with moving boundaries. *Advances in Water Resources*, **13**, 158–168.

Lees M., Young P., Ferguson S., Beven K. and Burns J. (1994) An adaptive flood warning scheme for the River Nith at Dumfries. In *2nd International Conference on River Flood Hydraulics*, White W.R. and Watts J. (Eds.), John Wiley & Sons: Chichester, pp. 65–75.

Lighthill M.J. and Whitham G.B. (1955) On kinematic waves: 1. Flood movement in long rivers. *Proceedings of the Royal Society of London Series A*, **229**, 281–316.

Litrico X. (2001) Nonlinear diffusive wave modeling and identification of open channels. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **127**(4), 313–320.

Lopez F. and Garcia M.H. (2001) Mean flow and turbulence structure of open-channel flow through non-emergent vegetation. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **127**(5), 392–402.

Lynch D.R. and Gray W.G. (1980) Finite element simulation of flow deforming regions. *Journal of Computational Physics*, **36**, 135–153.

Marks K. and Bates P.D. (2000) Integration of high resolution topographic data with floodplain flow models. *Hydrological Processes*, **14**, 2109–2122.

Mason D., Cobby D.M., Horritt M.S. and Bates P.D. (2003) Floodplain friction parameterization in two-dimensional river flood models using vegetation heights derived from airborne scanning laser altimetry. *Hydrological Processes*, **17**, 1711–1732.

McCarthy G.T. (1938) The unit hydrograph and flood routing. Paper presented at the *Conference of the North Atlantic Division of the U.S. Army Corps of Engineers*, New London.

McLelland S.J., Ashworth P.J., Best J.L. and Livesey J.R. (1999) Turbulence and secondary flow over sediment stripes in weakly

bimodal bed material. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **125**(5), 463–473.

Mertes L.A.K. (1997) Documentation and significance of the perirheic zone on inundated floodplains. *Water Resources Research*, **33**, 1749–1762.

Meselhe E.A., Weber L.J., Odgaard A.J. and Johnson T. (2000) Numerical modeling for fish diversion studies. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **126**(5), 365–374.

Molinaro P., Di Filippo A. and Ferrari F. (1994) Modelling of flood wave propagation over flat dry areas of complex topography in presence of different infrastructures. In *Modelling of Flood Propagation over Initially Dry Areas*, Molinaro P. and Natale L. (Eds.), American Society of Civil Engineers: New York, pp. 209–225.

Moussa R. (1997) Geomorphological transfer function calculated from digital elevation models for distributed hydrological modelling. *Hydrological Processes*, **11**(5), 429–449.

Moussa R. and Bocquillon C. (1996) Criteria for the choice of flood-routing methods in natural channels. *Journal of Hydrology*, **186**(1–4), 1–30.

Moussa R. and Bocquillon C. (2001) Fractional-step method solution of diffusive wave equation. *American Society of Civil Engineers, Journal of Hydrologic Engineering*, **6**(1), 11–19.

Nepf H.M. (1999) Drag, turbulence, and diffusion in flow through emergent vegetation. *Water Resources Research*, **35**(2), 479–489.

NERC (1975) *Flood Studies Report*, Vol. 5, Natural Environment Research Council: London.

Nezu I., Tominaga A. and Nakagawa H. (1993) Field-measurements of secondary currents in straight rivers. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **119**(5), 598–614.

Nicholas A.P. and Mitchell C.A. (2003) Numerical simulation of overbank processes in topographically complex floodplain environments. *Hydrological Processes*, **17**(4), 727–746.

Nicholas A.P. and Walling D.E. (1997) Modelling flood hydraulics and overbank deposition on river floodplains. *Earth Surface Processes and Landforms*, **22**, 59–77.

O'Donnell T. (1985) A direct three-parameter Muskingum procedure incorporating lateral inflow. *Hydrological Sciences Journal*, **30**(4), 479–496.

Pearson D., Horritt M.S., Gurney R.J. and Mason D.C. (2001) The use of remote sensing to validate hydrological models. In *Model Validation: Perspectives in Hydrological Science*, Anderson M.G. and Bates P.D. (Eds.), John Wiley & Sons: Chichester, pp. 163–194.

Pinder G.E. and Sauer S.P. (1971) Numerical simulation of flood wave modification due to bank storage effects. *Water Resources Research*, **7**, 63–70.

Ponce V.M., Li R.M. and Simons D.B. (1978) Applicability of kinematic and diffusion models. *American Society of Civil Engineers, Journal of the Hydraulics Division*, **123**(HY3), 353–360.

Poole G.C., Stanford J.A., Frissell C.A. and Running S.W. (2002) Three-dimensional mapping of geomorphic controls on flood-plain hydrology and connectivity from aerial photos. *Geomorphology*, **48**(4), 329–347.

Preissmann A. (1961) Propagation of translatory waves in channels and rivers. *Proceedings of the 1st Congress de l'Association Francaise de Calcul*, Grenoble, pp. 433–442.

Puech C. and Raclot D. (2002) Using geographical information systems and aerial photographs to determine water levels during floods. *Hydrological Processes*, **16**, 1593–1602.

Rabus B., Eineder M., Roth A. and Bamler R. (2003) The shuttle radar topography mission – a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, **57**(4), 241–262.

Romanowicz R. and Beven K.J. (1998) Dynamic real-time prediction of flood inundation probabilities. *Hydrological Sciences Journal*, **43**, 181–196.

Romanowicz R. and Beven K. (2003) Estimation of flood inundation probabilities as conditioned on event inundation maps. *Water Resources Research*, **39**(3), Art. No. 1073.

Romanowicz R., Beven K.J. and Tawn J. (1996) Bayesian calibration of flood inundation models. In *Floodplain Processes*, Anderson M.G., Walling D.E. and Bates P.D. (Eds.), John Wiley & Sons: Chichester, pp. 333–360.

Samuels P.G. (1990) Cross section location in one-dimensional models. In *International Conference on River Flood Hydraulics*, White W.R. (Ed.), John Wiley & Sons: Chichester, pp. 339–350.

Sauvaget P., David E., Demmerle D. and Lefort P. (2000) Optimum design of large flood relief culverts under the motorway in the Dordogne-Isle confluence plain. *Hydrological Processes*, **14**, 2311–2329.

Schlichting H. (1979) *Boundary-Layer Theory*, *Seventh Edition*, McGraw-Hill: New York, p. 817.

Sellin R.H.J. (1964) A laboratory investigation into the interaction between the flow in the channel of a river and that over its floodplain. *La Houille Blanche*, **7**, 793–801.

Sellin R.H.J. and Willetts B.B. (1996) Three-dimensional structures, memory and energy dissipation in meandering compound channel flow. In *Floodplain Processes*, Anderson M.G., Walling D.E. and Bates P.D. (Eds.), John Wiley & Sons: Chichester, pp. 255–298.

Sen D. (2002) An algorithm for coupling 1D river flow and quasi-2D flood inundation flow. In *Hydroinformatics 2002: Proceedings of the Fifth International Conference on Hydroinformatics*, *Model Development and Data Management, Vol. 1*, Falconer R.A., Lin B., Harris E.L. and Wilson C.A.M.E. (Eds.), IWA Publishing: London, pp. 103–108.

Shao X.J., Wang H. and Chen Z. (2003) Numerical modelling of turbulent flow in curved channels of compound cross-section. *Advances in Water Resources*, **26**(5), 525–539.

Shaw E.M. (1983) *Hydrology in Practice*, *Second Edition*, Van Nostrand and Reinhold: London, p. 539.

Shiono K. and Knight D.W. (1991) Turbulent open-channel flows with variable depth across the channel. *Journal of Fluid Mechanics*, **222**, 617–646.

Shvidchenko A.B. and Pender G. (2001) Macroturbulent structure of open-channel flow over gravel beds. *Water Resources Research*, **37**(3), 709–719.

Singh V.P. (1996) *Kinematic Wave Modelling in Water Resources: Surface Water Hydrology*, John Wiley & Sons: New York, p. 1399.

Singh V.P. (2001) Is hydrology kinematic? *Hydrological Processes*, **16**, 667–716.

Squillace P.J. (1996) Observed and simulated movement of bank storage water. *Groundwater*, **34**, 121–134.

Srokosz M. (1997) Ocean surface currents and waves and along-track interferometric SAR. *Proceedings of a Workshop on Single-Pass Satellite Interferometry*, Imperial College: London, 22 July 1997.

Stanford J.A. and Ward J.V. (1988) The hyporheic habitat of river ecosystems. *Nature*, **335**, 64–66.

Stewart M.D., Bates P.D., Anderson M.G., Price D.A. and Burt T.P. (1999) Modelling floods in hydrologically complex lowland river reaches. *Journal of Hydrology*, **223**, 85–106.

Stoesser T., Wilson C.A.M.E., Bates P.D. and Dittrich A. (2003) Application of a 3D numerical model to a river with vegetated floodplains. *Journal of Hydroinformatics*, **5**, 99–112.

Takeuchi K., Ao T.Q. and Ishidaira H. (1999) Introduction of block-wise use of TOPMODEL and Muskingum-Cunge method for the hydro-environmental simulation of a large ungauged basin. *Hydrological Sciences Journal*, **44**(4), 633–646.

Tang X.N., Knight D.W. and Samuels P.G. (1999a) Variable parameter Muskingum-Cunge method for flood routing in a compound channel. *Journal of Hydraulic Research*, **37**(5), 591–614.

Tang X.N., Knight D.W. and Samuels P.G. (1999b) Volume conservation in variable parameter Muskingum-Cunge method. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **126**(5), 610–620.

Tchamen G.W. and Kawahita R.A. (1998) Modelling wetting and drying effects over complex topography. *Hydrological Processes*, **12**, 1151–1183.

Temple D.M. (1987) Closure of 'velocity distribution coefficients for grass-lined channels'. *American Society of Civil Engineers, Journal of Hydraulic Engineering*, **113**(9), 1224–1226.

Thomas T.G. and Williams J.J.R. (1995) Large-Eddy simulation of turbulent flow in an asymmetric compound open channel. *Journal of Hydraulic Research*, **33**(1), 27–41.

Venere M. and Clausse A. (2002) A computational environment for water flow along floodplains. *International Journal of Computational Fluid Dynamics*, **16**(4), 327–330.

Vieira J.D. (1983) Conditions governing the use of approximations for the Saint-Venant equations for shallow surface water flow. *Journal of Hydrology*, **60**, 43–58.

Walling D.E., Bradley S.B. and Lambert C.P. (1986) *Conveyance Loss of Suspended Sediment within a Floodplain System*, Vol. 159, IAHS Publication: pp. 119–132.

Werner M.G.F. (2001) Impact of grid size in GIS based flood extent mapping using a 1D flow model. *Physics and Chemistry of the Earth Part B – Hydrology, Oceans and Atmospheres*, **26**(7–8), 517–522.

Werner M.G.F. (2002) Uncertainties in floodplain inundation modelling. In *Hydroinformatics 2002: Proceedings of the Fifth International Conference on Hydroinformatics*, *Software Tools and Management Systems, Vol. 2*, Cluckie I.D., Han D., Davis J.P. and Heslop S. (Eds.), IWA Publishing: London, pp. 1347–1354.

Westaway R.M., Lane S.N. and Hicks D.M. (2003) Remote survey of large-scale braided, gravel-bed rivers using digital photogrammetry and image analysis. *International Journal of Remote Sensing*, **24**(4), 795–815.

Wicks J., Mocke R., Bates P.D., Ramsbottom D., Evans E. and Green C. (in press) Selection of appropriate models for flood modelling. *Proceedings of the 38th Department for Environment, Food and Rural Affairs Annual Flood and Coastal Management Conference*, Department of Food and Rural Affairs: London.

Wilson M. and Atkinson P. (in press) The use of elevation data in flood inundation models: a comparison of landmap and ordnance survey land-form PROFILE™. *Computers and Geosciences*.

Wilson C.A.M.E. and Horritt M.S. (2002) Measuring the flow resistance of submerged grass. *Hydrological Processes*, **16**(13), 2589–2598.

Woessner W.W. (2000) Stream and fluvial plain ground water interactions: re-scaling hydrogeologic thought. *Groundwater*, **38**, 423–429.

Wormleaton P.R., Allen J. and Hadjipanos P. (1982) Discharge assessment in compound channel flow. *American Society of Civil Engineers, Journal of Hydraulics Division*, **108**(HY9), 975–993.

Wroblicky G.J., Campana M.E., Valett H.M. and Dahm C.N. (1998) Seasonal variation in surface-subsurface water exchange and lateral hyporheic area of two stream-aquifer systems. *Water Resources Research*, **34**, 317–328.

Young P.C. (1986) Time-series methods and recursive estimation in hydrological systems analysis. In *River Flow Modelling and Forecasting*, Karijenhof D.A. and Moll J.R. (Eds.), Reidel: Dordrecht, pp. 129–180.

Young P.C. (2002) Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society, Series A: Mathematical, Physical and Engineering Sciences*, **360**(1796), 1433–1450.

125: Rainfall-runoff Modeling for Flood Frequency Estimation

ROB LAMB

JBA Consulting – Engineers & Scientists, Skipton, North Yorkshire, UK

Rainfall-runoff models are amongst the most important tools for the practical solution of flood estimation problems, as well as for theoretical investigations of controls on the flood frequency curve or for analysis of catchment and climate change. This article aims to provide a discussion of the use of rainfall-runoff modeling for flood estimation, with an emphasis on overall methodology rather than on specific hydrological models. After a brief introduction to some basic concepts of flood analysis, the article describes the two main approaches for applying rainfall-runoff models in flood estimation, namely, event-based modeling and continuous simulation. The theoretical basis for each approach is described along with typical model structures. Topics covered include rainfall models, hydrological model calibration and the estimation of parameters for modeling at ungauged or data-poor sites. The article concentrates on the development of the continuous simulation method, and includes a number of illustrations of its application, both from research studies and applied hydrology.

INTRODUCTION

Rainfall-runoff models are used for flood frequency estimation to simulate storm events or continuous time series of runoff for which the probability of exceeding a given flow can be determined. Event-based procedures simulate flows for discrete "design" storms and are an established complement to statistical methods for flood frequency estimation. Continuous simulation has developed from process-based ideas about modeling, in which the runoff that produces flood flows is modeled, at least in principle, as part of a wider range of catchment responses. In this way, the approach may be distinguished from the statistical and design-event flood estimation methods where specific features of the catchment response (flood peaks and individual event hydrographs, respectively) are separated out as the objects to be modeled. Methods that seek to integrate rainfall-runoff modeling and statistical frequency analysis can also be referred to as "derived distribution" approaches.

This article considers flood estimation in non-urban settings, although many of the techniques reviewed have parallels in urban drainage modeling. It gives a brief overview of event-based hydrograph models, but does not seek to duplicate the many detailed descriptions available in standard textbooks and technical manuals (*see*, for example, Natural Environment Research Council, 1975; Institution of Engineers, 1987; Chow *et al.*, 1988; Beven, 2001). The article instead concentrates on the rapidly progressing continuous simulation approach. There are now a great number of (often very similar) hydrological models that could be described in this context. Rather than reproducing a long list of models, the aim here is to explore how rainfall-runoff modeling fits within a whole framework for flood estimation, which may also need to include rainfall and flow routing models. Before examining the development of rainfall-runoff models for flood estimation, it will be useful to survey very briefly some basic concepts and features of statistical methods used for flood frequency analysis.

CONCEPTS OF FLOOD FREQUENCY

Flood frequency estimation is concerned with the quantification of rarity. This can be expressed in terms of the probability of observing a flood of a given size at any given time. Consider a gauging site where there is a long, continuous record of flow data, Q. There will be some probability

$$p = \Pr(Q > Q') \tag{1}$$

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

of observing flows greater than a given value Q'. This "exceedance probability" is commonly expressed as a return period,

$$T = \frac{1}{p} \tag{2}$$

Generally the return period is the long-term average interval between flows greater than Q'. Often it is convenient to define the size of floods in terms of the largest flow recorded in any year, known as the *annual maximum or AMAX series*. In this case, the return period has the more specific interpretation of the long-term average interval between years containing a flood peak larger than Q' and p is the annual exceedance probability. An alternative is to define the series of flood peaks above a given threshold, where the threshold is set so as to include an average of, say, three peaks per year. This "Peaks-Over-Threshold" (POT) or "partial duration" series generally provides a larger sample size for a given record length.

The relationship between Q and T (or p) is referred to as the flood frequency curve. The current probability of exceeding a specified flow Q' could easily be determined with confidence if a river had been gauged over a very long period of stationary physical and climatic conditions, during which time the flow Q' had been observed many times. These conditions are not usually true in practice. For this reason, the exceedance probabilities associated with peak flows have to be estimated, especially for flows close to or greater than the maximum on record. All determinations of flood rarity are therefore estimates, and "flood frequency estimation" will be used here to refer both to estimates of the size of floods of a given rarity and to the estimation of rarity for specified flood flows.

STATISTICAL APPROACHES

Flood Frequency Distributions

River flows have long been measured and used by engineers in the design of hydraulic structures and for planning development in, or for management of, the floodplain. Flood frequency analysis has been the basis for engineering design and economic analysis, and methods have been developing since at least the early twentieth century (Benson, 1968). The usual approach has been to regard the size of floods in a catchment, when arranged as a series of independent peak flows, as a sample drawn from some underlying parent distribution. Many statistical distributions have been used at different times (see, for example, Bobée and Rasmussen, 1995, Clarke, 1994; Institute of Hydrology, 1999; Reimann, 1989).

In flood frequency analysis, measured peak flows can be used to estimate the form and parameters of the parent distribution. Once the parent distribution is established in this way, it can be used as a basis for extrapolation to estimate flood quantiles for very rare conditions, especially peak flows whose return period is much longer than the period of recorded flows. Typically, this is done under an assumption that the flood-producing regime of the catchment does not change greatly over time (the flood "regime" includes here both runoff production mechanisms and the driving climatic patterns). However, the stationarity condition can be relaxed. For example, Clarke (2003) has incorporated trend into estimates of the probability of future extreme events on both annual maximum and POT scales, while Cunderlik and Burn (2003) show that even weakly significant trends may "seriously bias" flood estimates. This could be a problem in practice even over a relatively short design life of 10 or 20 years. Whilst the statistical tools exist to fit probability distributions in the presence of trend, there has often been considerable uncertainty over the exact parameterization, future persistence and even the existence of such trends (whether they are associated with changes in climate or catchment land use). Useful surveys of the issues have been provided by Robson and Kundzewicz (2000) and Kundzewicz (2004, and papers cited therein).

Sophisticated methods are available for fitting the distribution for a flood series. The conventional methods are maximum likelihood and equating moments of the theoretical distribution with sample moments calculated from flow data. More recently, Probability Weighted Moments or PWMs (Davison and Smith, 1990; Wang, 1991) and robust linear combinations of PWMs called *L moments* (Hosking, 1990) have been widely used. The factors that limit the presumed accuracy of a flood estimate are fundamentally the quality and amount of measured data, the suitability of the chosen distribution and the degree to which the statistical assumptions are met, including the robustness of the distribution over different flow ranges and over time. The simplest case is where a "single-site" distribution is fitted to flow data from a gauged location. Consistent, high-quality flow series rarely span more than 50 years or so in most parts of the world. Typical flow records therefore do not provide large samples of peak data and the uncertainty in a single-site distribution can be large, particularly when the distribution is extrapolated to estimate floods typically adopted for "design", say, of $T \sim 100$ years.

One response to this problem of record length is to add information from other locations to increase the effective sample size. This technique relies on assuming that data from sites within a geographical region, or from catchments of similar physical and climatic characteristics, can be combined to fit a distribution curve typical of those catchments. It is usually necessary to separate the scale and the shape of the frequency curve, where the scale is based on a single-site estimate of

an index flood of short return period, say the mean or median annual maximum flow, whilst the shape of the frequency curve is based on a dimensionless regional growth curve. A T-year flood estimate is then obtained as the product of the local index flood and the regional growth curve.

Combining data within fixed geographical regions is referred to as "regionalization". In some cases, the regionalization has been based on convenient administrative units, such as the boundaries of hydrological authorities (Natural Environment Research Council, 1975). A more recent development is to move away from using fixed geographical regions and instead, for any subject site, to pool peak flow data from a group of catchments that are similar in terms of hydrologically significant characteristics rather than geographical proximity (Acreman and Wiltshire, 1989; Burn, 1990). As with the fixed regionalization approach, the "pooling group" must contain catchments in which the flood regime is homogeneous. Pooling has become the recommended technique for flood estimation in the United Kingdom (Institute of Hydrology, 1999). Heterogeneity caused by particular features (such as engineered storage or differences in seasonality, for example) would reduce confidence in the pooled flood analysis, but the catchments forming a pooling group can be scrutinized to exclude these effects.

Ungauged Sites

One of the most important and challenging problems in flood frequency analysis is estimation at ungauged sites, where there are no data available for fitting a local model directly. This statement applies equally to statistical and other flood estimation methods (and, indeed, to models used for other hydrological purposes, *see* **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**). In statistical flood estimation, a regionalized or pooled growth curve can be applied for an ungauged or data-poor subject site in much the same way as for a gauged site, based on physical similarity. However, the index flood obviously cannot be determined from flow data at an ungauged site, and so has to be estimated in a different way. This may be done using a regression model that relates values of the index flood at gauged sites to catchment properties, or by a nearest-neighbor approach, in which the index flood is computed as a weighted average from gauged sites most similar to the subject site.

Statistical methods for flood frequency analysis are very powerful tools, and are widely used and accepted in practice. Advantages include the direct empirical support through the gauged data, the ability to incorporate data from many locations, the relative ease of calculation and the appealing theoretical basis for representing floods using extreme value distributions. There are some limitations, however. Uncertainty arising from limited sample size remains, albeit moderated by the use of regionalized or pooling methods. Perhaps more limiting is the treatment of "floods" as discrete, independent observations of a stochastic process. Although it is possible to relax these assumptions, for example, by building distributional models that include autocorrelation and trend, the result is more difficult to work with than the conventional statistical methods.

Despite their empirical basis, it would be far from true to assert that statistical methods take no account of hydrological realities. Many studies using flood frequency distributions have been informed by knowledge of the catchment situation; for example, the analysis of multicomponent distributions that reflect different types of flood (e.g. Connell and Pearson, 2001; Rossi *et al.*, 1984; Waylen and Woo, 1982). Indeed, at a practical level, when deriving a flood frequency curve, the experienced analyst will make a careful study of the quality of any available local data. Often, this will include scrutiny of rating curves, which may reflect hydraulic characteristics such as transitions between channel and floodplain flows that can, in turn, affect the shape of the flood frequency curve (Naden, 1992a). However, statistical approaches to flood estimation are not structured in a way that corresponds in any detail to hydrologists' understanding of the processes that generate flood flows.

Statistical approaches become more difficult to apply where flood generation includes storage or threshold effects, or where the concept of flood frequency is really being used as a way of determining the hydrological conditions associated with a specified level of risk that results from a combination of factors. In many practical situations, it is necessary to model the response of a complex system to the prescribed risk, for example, in designing flood management schemes or for the purposes of planning where there are pressures to build or otherwise use the floodplain for development. Often, the flood response will be influenced by natural or engineered storage, such as floodplain inundation, storage behind raised embankments or in artificial retention ponds. In these cases, the volume of water and its timing may be as important as the peak flow.

EVENT-BASED RAINFALL-RUNOFF METHODS

Unit Hydrograph

To provide flood estimates in cases where dynamic processes are important, a point estimate of peak flows is unlikely to provide a realistic basis for analysis and a hydrograph is needed. A typical approach to this problem has

been to simulate "design" flood events using a rainfall-runoff model. Event models have evolved from the unit hydrograph theory of Sherman, 1932 (*see* Dooge, 2003, for a detailed discussion). The unit hydrograph (UH) is an empirical model that specifies the delays and attenuation of runoff in response to a given rainfall input. The model is linear so that the response to a rainfall input of N units is held to be N times larger than the response to a single unit input. The form of the unit response is also generally assumed to remain fixed over time. Consequently, the responses to successive unit inputs can be added together to provide a total response.

It has not always been considered necessary to specify a "hydrograph-shaped" UH; standard engineering applications often make use of a triangular approximation to simplify the calculations (e.g. Institute of Hydrology, 1999). The three principles of linearity, time invariance and superposition are illustrated in Figure 1. The linearity assumption is known to be physically only an approximation, both for flow routing (where the flood wave velocity changes with discharge for several reasons, *see* **Chapter 124, Flood Routing and Inundation Prediction, Volume 3**) and, in particular, for runoff production, which will be discussed further below. However, it has proven to be a useful approximation, given other sources of error in estimating flood magnitudes, especially where calibrated, using data from reasonably large flood events (return periods of ∼10 years or more), on catchments large enough for some averaging of runoff responses to take place (typically ∼10 to ∼100 km^2) and not too large for the distribution of rainfall to be significant. The UH model has therefore been successful enough to continue in routine engineering use around the world up to the present day (*see also* the general discussion of transfer function modeling techniques in **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**).

Calculating Losses in Event Models

The accumulated response of runoff to rainfall is highly nonlinear because of the influence of soil moisture storage (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2**– **Chapter 79, Assessing Uncertainty Propagation Through Physically based Models of Soil Water Flow and Solute Transport, Volume 2**), which may cause a transient "loss" of rainfall within the event timescale, and losses to groundwater (*see* **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4** and **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**) and evaporation (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1** and **Chapter 42, Transpiration, Volume 1**). The input to the UH response function should not be the total rain falling on the catchment, but instead some effective

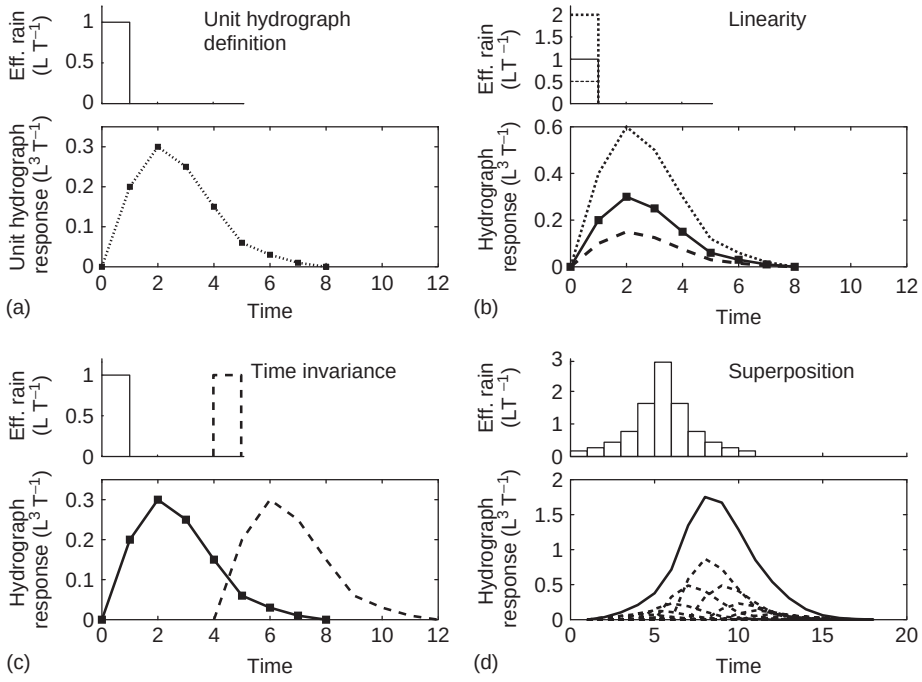


Figure 1 Principles of unit hydrograph theory: (a) definition of a triangular unit response as used in UK flood studies; (b) principle of linearity; (c) time invariance of the unit response; and (d) superposition, allowing the convolution of a storm rainfall profile with the unit hydrograph (Reproduced from Institute of Hydrology, 1999 by permission of Centre for Ecology and Hydrology (CEH))

rainfall that accounts for these losses. The UH is therefore generally coupled with a losses model to provide some variation in runoff rates as a function of an assumed antecedent catchment soil moisture status and of evaporation losses.

There are many possible forms for the losses model. Some seek to represent process understanding of the physics of interception, infiltration, and evaporation processes. Others are conceptual in nature, mixing parameters that have a physical basis (such as a hydraulic conductivity) with empirical ones (such as storage coefficients). Simple empirical methods are usually applied to represent losses in event-based models (which are anyway most often applied in situations where detailed physical modeling would be impractical because of lack of data). One such empirical method is the model of the USDA Soil Conservation Service (1985). This method calculates direct storm runoff

$$Q = \frac{(P - S_1)^2}{(P - S_1 + S_2)} \approx \frac{(P - 0.2S_2)^2}{(P + 0.8S_2)} \tag{3}$$

as a function of rainfall P, an initial rainfall loss S_1 occurring before the onset of runoff, and a maximum potential loss S_2. When the method was developed, losses were assessed by a graphical comparison of measured rainfall and runoff volumes, and interpreted as comprising interception, depression storage, an initial volume of infiltration and the volume of infiltration after runoff commenced. The great utility of the method is that S can be estimated from soil and land cover data, on the basis of relationships that were established from analysis of hydrographs for agricultural plots and small catchments in the Midwestern USA. The parameter S, originally defined in inches, is more commonly transformed into the curve number, $CN = 10\,000/(S + 10)$, to provide a more convenient scale. Whilst this general model derives from a consideration of basin water balance, it is entirely empirical in its application and not necessarily appropriate beyond the region in which it was calibrated.

In the United Kingdom, an alternative approach was developed as part of the Flood Studies Report (FSR, Natural Environment Research Council, 1975), again based on empirical analysis of recorded events. The FSR uses the concept of "percentage runoff" (*PR*) to represent losses as a combination of a standard percentage runoff parameter, *SPR* (which is treated as a fixed property of the catchment) and a "dynamic" modification to this that depends on an index of catchment wetness and on storm rainfall depth. In fact, the modification is only "dynamic" in the sense that it attempts to account for differences in *PR* between different storms; the *PR* is then applied as a constant proportional loss rate throughout the duration of the rainfall event.

The FSR percentage runoff model is

$$PR = SPR + DPR_{\mathrm{CWI}} + DPR_{\mathrm{RAIN}} \tag{4}$$

where DPR_{CWI} represents the contribution of antecedent catchment wetness, expressed as a catchment wetness index (CWI) and DPR_{RAIN} is a function of storm depth (and is zero for storms of less than a threshold depth). This approach was formulated with the explicit intention of providing a means of simulating individual runoff events, and DPR_{CWI} can be estimated from antecedent rainfall and soil moisture deficits. In the more common case where a hydrograph has to be produced for a statistically derived design rainfall, often at an ungauged location, then CWI and its associated *PR* contribution are estimated from an empirical function of long-term average rainfall. Like the curve number method, the parameters of the losses model are based on empirical relationships with catchment characteristics.

Design Inputs

Event-based procedures conform to the basic generic structure of any conceptual rainfall-runoff model, combining loss and attenuation modules. Typically, variables such as the shape and time-to-peak of the hydrograph and the runoff coefficient tend to be parameterized empirically (i.e. they are calibrated from analysis of gauged events). The event-based method makes good use of rainfall data, which itself brings potential advantages in regions where the rain gauge network is more extensive than the flow gauge network, or where longer records exist for precipitation.

For flood frequency estimation, a statistical rainfall model is used to estimate the storm that would cause a flood hydrograph of a given probability to occur. However, by making the flood-producing rainfall a probabilistic element, a link has to be made to the probability of the resulting flood. From a process point of view, this can become a complicated issue because the relationship between rainfall and flood probability must account for the dynamics of catchment storage (both within and between events). These are controlled by precipitation history, soils and topography, evaporation and other processes. Even in a relatively straightforward catchment not influenced by snowmelt, groundwater, or tidal boundary conditions, a complete event model would ideally account for the joint probability distributions of rainfall intensity, storm duration, storm profile, antecedent soil moisture conditions and evaporation losses (both being functions of past rainfall). Additionally, the model would have to represent the transformation of the joint inputs to produce a flood hydrograph, including rapid storm runoff and more slowly responding "baseflow" components.

The joint controls on the flood probability have generally been resolved by the specification of "design packages" of inputs that combine to generate flood events of specified probability, rather than attempting to resolve the joint probability issues explicitly. In the FSR, for example, a Monte Carlo analysis was undertaken to simulate the

outputs of many different combinations of design inputs to the UH model. The Monte Carlo simulation (MCS) allowed a link to be established between the probability of the rainfall event used as input to the UH model and the desired flood probability.

The four design variables that form the inputs to the FSR event model are rainfall duration, rainfall depth (or return period), design storm profile and antecedent CWI. MCS on 98 catchments demonstrated that the simple event model structure could reproduce the probability distributions derived from gauged annual maximum series. There followed a set of experiments in which three of the four input variables were fixed and the remaining variable optimized to agree with the flood magnitudes simulated by the Monte Carlo method. It was found that the flood magnitude was most sensitive to rainfall depth and antecedent wetness. For ungauged sites, the parameters that define the simplified triangular UH were related empirically to catchment characteristics and to the duration of the design rainstorm.

When antecedent wetness was also fixed, a relationship between the return periods of the storm depth and the resultant flood was established, and found to be "similar between catchments" (Natural Environment Research Council, 1975, p. 455). The relationship is shown in Figure 2. This analysis was a considerable achievement in that it provided users of the method with workable, scientifically based rules to apply the event rainfall-runoff model in practice. It is also, in effect, a form of calibration used to circumvent the joint probability issues in modeling. Because the method relies on this calibration, updating its component modules can affect its performance, as discussed by Ashfaq and Webster (2002). A more explicit solution of the joint probability issues has been demonstrated (Rahman *et al.*, 2002) for the Design Event Approach used in Australia (Institution of Engineers, 1987).

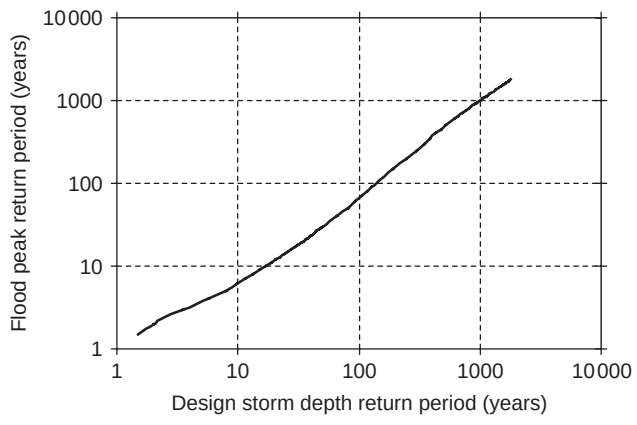


Figure 2 Calibrated relationship between return period for design storm depth and required flood peak return period for the UK event-based rainfall-runoff method (Reproduced from Institute of Hydrology, 1999 by permission of Centre for Ecology and Hydrology (CEH))

DERIVED DISTRIBUTIONS AS A GENERALIZATION OF RAINFALL-RUNOFF MODELING

Event-based methods provided a practical engineering solution for flood estimation based on statistical and theoretical approaches established, for the most part, in the early part of the twentieth century. Meanwhile, opportunities to advance the underlying theoretical approaches were also being investigated. In an important paper, Eagleson (1972) introduced the concept of derived flood distributions. The derived distribution approach was a step towards a flood frequency methodology that would combine a treatment of the dependence between causative factors with a conceptual model more closely linked to physical process. The three main steps in building a more "physically based" model of flood frequency were:

- a model of the rainfall process (or gauged rainfall data),
- a model for runoff production and routing,
- a method to combine the component models to derive the flood frequency distribution.

The basis of the methodology was to combine statistical distributions for effective rainfall with a kinematic wave model of runoff propagation. An important feature of the methodology was that it integrated analytically the probabilities governing rainfall with the runoff production functions in an attempt to derive a theory of flood frequency related to physical catchment parameters.

The intensity of rainfall excess after losses, i_e, and the duration of the rainfall excess, t_e, were assumed to be independent random variables described by the exponential joint probability density function

$$f(i_e, t_e) = \frac{b\lambda}{\kappa} e^{-(bi_e/\kappa + \lambda t_e)} \tag{5}$$

In developing equation (5), Eagleson made the simplifying assumption that storm rainfall is effective over a fixed proportion of the catchment area A_r that is assumed to generate direct runoff. The partial area A_r was assumed to be determined solely by physical characteristics of the catchment and the long-term climate (although, in reality, it would be expected to vary as a function of soil moisture changes between and during rainfall events). The treatment of rainfall excess also assumed a constant (in time) and uniform (in space) loss rate.

Both overland and channel flow were represented using the kinematic wave model on an idealized catchment consisting of two symmetrical overland flow planes draining into a linear stream channel. From a hydrological

and hydraulic point of view, the kinematic wave model is a highly simplified representation of the movement of water. However, it provides an acceptably accurate description over a range of typical physical conditions for overland and channel flow (Singh, 2002). The kinematic wave can be derived from the full 1-D Saint-Venant equations for open channel flow (*see* **Chapter 135, Open Channel Flow – Introduction, Volume 4**, **Chapter 136, Hydrodynamic Considerations, Volume 4**) and its parameters therefore have a physical interpretation.

If the discharge anywhere on a one-dimensional channel or overland flow "plane" of unit width is q and the depth of flow is y then the continuity equation can be written

$$\frac{\partial y}{\partial t} + \frac{\partial q}{\partial x} = i \tag{6}$$

where x is distance in the downslope or downstream direction, t is time and i is a rate of inflow. In the derived flood frequency analysis, the inflow for the overland flow (q_o) on each slope was rainfall excess i_e over duration t_e. For flow in the stream channel, the inflow was taken to be $2q_o$. Equation (6) can be rewritten as

$$\frac{\partial q}{\partial t} + c\frac{\partial q}{\partial x} = ic \tag{7}$$

where the wave celerity

$$c = \frac{\partial q}{\partial y} \tag{8}$$

is the speed with which a disturbance will propagate in the downslope or downstream direction. The total differential $\mathrm{d}x = c \cdot \mathrm{d}t$ describes the passage of a disturbance or wave at any location. Eagleson (1972) used the kinematic wave model formed by combining equation (7) with a flow law of the generic form

$$q = \alpha y^m \tag{9}$$

where the resistance parameter α can be related to the surface slope angle for overland flow or the channel slope for stream flow. For overland flow, the chosen parameterization was $m = 2$. For flow in the stream channel, $m = 3/2$ was chosen, which allows equation (9) to be interpreted as the Chezy equation for flow resistance. Equations (7) and (9) can be solved for constant and uniform inputs by integrating along the characteristic curves in the (x, t) plane defined by

$$\frac{\mathrm{d}y}{\mathrm{d}t} = i \tag{10}$$

for constant x and

$$\frac{\mathrm{d}x}{\mathrm{d}t} = c = m\alpha\ y^{m-1} \tag{11}$$

for characteristics of constant y. In carrying out the integration, three dominant event scenarios can be defined in terms of the duration of rainfall excess relative to the times of concentration of the catchment or hillslope, as follows:

- the duration of rainfall excess is greater than the catchment time of concentration,
- the duration of rainfall excess is less than the catchment time of concentration but greater than the time of concentration of overland flow, and
- the duration of rainfall excess is less than the time of concentration of overland flow.

Each of these scenarios leads to an expression for the maximum peak flow rate $Q_{\max}$ of the general form

$$Q_{\max} = g(i_e, t_e, \boldsymbol{\theta}) \tag{12}$$

where $Q_{\max}$ is a function of given values of the random variables for rainfall excess (intensity, i_e, and duration, t_e) and the set of fixed catchment parameters $\boldsymbol{\theta}$ (which includes the parameters for flow resistance and the proportion of the catchment assumed to generate direct storm runoff).

Equation (12) describes a curve on a graph of i_e against t_e beneath which the peak of a hydrograph will be less than the maximum possible value. The principle of the analytical derived distribution approach was then to integrate the joint probability distribution for the rainfall excess (equation (5)) over the region within which $Q < Q_{\max}$, that is, the region bounded above by the function g. This integration results in the derived flood frequency distribution

$$\begin{aligned} F'(Q_{\max}) &= \Pr(Q \le Q_{\max}) \\ &= \iint_{Q<Q_{\max}} f(i_e, t_e)\,\mathrm{d}i_e\,\mathrm{d}t_e \\ &= 1 - \Pr(Q > Q_{\max}) \end{aligned} \tag{13}$$

which may be compared with the general form in equation (1). Eagleson (1972) also added an annual average baseflow contribution (which does not alter the probability integration described above) to derive the flood frequency distribution in terms of total peak flows.

The derived distribution theory was shown to be reasonably successful in reproducing the frequency distributions of three river basins in Connecticut. The theory also attempted to relate flood frequency to the underlying mechanics of runoff production. However, the assumptions needed to derive the flood frequency distribution analytically may be considered to exclude some physical processes known to occur in reality. Two of the most conspicuous examples of this are the assumption of independence between the duration and mean intensity of rainfall excess and the fixed partial area for runoff production.

A number of authors have also developed derived flood frequency distributions using geomorphological unit hydrograph (GUH) theory (Rodriguez-Iturbe and Valdes, 1979; Rodriguez-Iturbe *et al.*, 1982; Rodriguez-Iturbe, 1993) as a flood routing mechanism with a link to physical basin parameters. Whilst the GUH is a concise way of building basin characteristics into a conceptual routing model, it also requires a method to estimate the actual depth or rates of runoff to be routed. Hebson and Wood (1982) applied the GUH with a fixed partial area mechanism for runoff production. Cordova and Rodriguez-Iturbe (1983) combined the GUH with a calculation of runoff production based on Soil Conservation Service Curve Numbers. Working with a triangular UH shape and initially assuming constant intensity of rainfall within a storm of duration t_r, Cordova and Rodriguez-Iturbe derived an expression for the peak of the response hydrograph Q' as

$$Q' = 2.42\frac{R_P A}{\pi}\left(1 - 0.218\frac{t_r}{\pi}\right) \qquad (14)$$

where A is the catchment area, R_p is the direct runoff depth generated by a rainfall depth P, as computed using the appropriate SCS curve for the catchment and π is a geomorphic parameter incorporating information derived from the geometry of catchment drainage network, but also dependent nonlinearly on R_p and t_r. Annual maximum rainfall depths were extracted from gauged records for a range of durations. By applying equation (14), a corresponding series of flood peaks were obtained for the different durations, and the largest value for each year was chosen as the annual maximum flood peak for that year. An extreme value distribution was then fitted to this synthetic series of annual maximum flood peaks. Wood and Hebson (1986) adopted similar methods, but specified the probability distribution for the rainfall excess used to generate runoff, rather than using gauged records.

The derived distribution theories cited above made an explicit link between flood frequency and physical basin parameters. This link was based not on empirical considerations, but on conceptual representations of the catchment processes. However, the representations had to be simple ones (at least compared to contemporary knowledge of physical processes) to allow exact solutions to be found for the derived flood frequency distributions. To obtain solutions in a convenient form necessarily places limits on the processes that can be represented within the model, in terms of both inputs and runoff production/routing. In particular, it is difficult to account analytically for different sequences of inputs occurring over different timescales. As noted by Beven (1986), the derived distribution models were, out of necessity, not "satisfactory descriptions of the runoff production process at the basin scale".

A solution to this problem would be to model the catchment inputs and outputs explicitly, requiring a stochastic approach to the modeling of precipitation (effectively generating a synthetic record of inputs) and a runoff model capable of transforming the input series into flood hydrographs. It would then be possible to generate relatively long sequences of inputs and use the catchment runoff model to keep track of the evolution of soil moisture, groundwater, or channel storage. Hence, the approach would be one of continuous simulation of streamflow responses.

CONTINUOUS SIMULATION METHODS

Although computer simulation models of catchment hydrology had been available since at least 1960, by the end of the 1970s, models of catchment runoff, along with processing and data storage capabilities, had advanced to the stage where it would become feasible to link them with stochastically generated input data. Free of the need to derive analytical solutions, the dynamics of catchment storage and threshold effects could be represented with greater flexibility, as could spatial variation in runoff responses. The three elements of the continuous simulation method are:

- a series of climatological inputs, derived either from observed data or using a stochastic or process model,
- a model of the catchment processes to be simulated,
- the extraction from the simulated data of important features of the flood response.

It will be seen that the three elements are similar to those of a "physically based" derived distribution approach to flood frequency. The main difference is that the elements need not be combined analytically to derive a flood frequency distribution. The end product is a time series of flow data, from which it is easy to extract different features of interest, including peak flows, volumes above a threshold and measures of flow duration. Potentially, any rainfall-runoff model could be used, ranging from a detailed, distributed physically based model to a black-box transfer function approach. At the center of most developments of continuous simulation are models of catchment hydrology that incorporate some structural knowledge or theory to describe the dynamics of runoff production, but that are simple enough to be computationally efficient.

One of the motivations for continuous simulation, as noted by Bras *et al.* (1985), is that rainfall data can often provide more information than streamflow data, at least in the sense that rain gauge networks may be denser, more extensive and longer established. Data are also required for the calculation of potential evaporation, and would obviously be needed to support snow accumulation and melt models in some regions. The cumulative effect of evaporative forcing may affect the flood regime in some catchments by controlling soil moisture deficits (and, in greater physical detail, aspects of vegetation cover and soil structure).

Evaporation must therefore be an important part of the continuous simulation approach, and compared to current process understanding, a simple uniform loss rate would seem to be too simple a model (*see* **Chapter 40, Evaporation Measurement, Volume 1**– **Chapter 45, Actual Evaporation, Volume 1**). However, it will generally be the case on an event-by-event basis that rainfall inputs dominate the hydrological response. Indeed, it has been found (Calder *et al.*, 1983; Fowler, 2002) that relatively simple representations of evaporation are often good enough for stream flow modeling.

Rainfall inputs are clearly of paramount importance. In some cases, there may be an interest in simulating flows directly from gauged rainfall records where there are limited flow records (but good rainfall data) to fit a frequency distribution; an early example using the Stanford Watershed Model (Crawford and Linsley, 1966) was reported by James (1965).

Often, however, rainfall records will be too short to drive the length of simulation that would be needed to sample floods of "design" magnitude. Whilst continuous simulation would still provide a means of estimating the whole flood hydrograph without needing to invoke baseflow separation or design package assumptions, there would be some justification for arguing that a better estimate of flood magnitudes could be obtained by making use of pooled statistical analysis. To overcome the limitations of gauged records, modeled rainfall can be used as the driving input for continuous simulation. In this way, a long artificial streamflow series can be generated, including floods of a magnitude that may not have been observed in the shorter, gauged record. As with event-based flood analysis, the use of a rainfall model (albeit of a different kind this time) provides a way of exploiting the information on extremes contained in the rainfall records.

RAINFALL MODELS FOR FLOOD FREQUENCY ESTIMATION BY CONTINUOUS SIMULATION

It is not the purpose of this article to offer an in-depth survey on rainfall modeling, but it is worth summarizing a few approaches that have been found suitable for use in continuous simulation for flood frequency. Cox and Isham (1994) identified three main approaches to rainfall modeling: purely empirical statistical models, dynamic meteorological models and intermediate stochastic models that have a structure based conceptually on physical phenomena. There is also a class of models based on fractal representations of rainfall (Lovejoy and Schertzer, 1985; Puente, 1996). The suitability of a model will depend on the space and timescales that are important for the particular application. For example, Acreman (1990) noted that Markov chain models that predict rainfall depths conditional on past rainfall can present a large number of parameters to be estimated when applied at relatively fine timescales (for instance, the hourly or subhourly scales often needed to resolve flood flows), where dependence can extend over many time steps. For catchments of area $\sim$10 to $\sim$1000 km^2, the chosen rainfall model may need to reproduce short bursts of high rainfall intensity embedded within medium-term and seasonal accumulations that set the antecedent conditions for runoff generation.

Point Rainfall Models

For a catchment that is small enough (relative to the dominant spatial structures in rainfall), rainfall can be considered as an areally averaged quantity. A one-dimensional, "point" model can then be used. Typically, such models simulate storms and dry periods separately to build up a continuous sequence of rainfall depths. In the simplest cases, storm intensities, durations, and interarrival times are simulated from prescribed distributions, and a typical profile is adopted for each storm event.

The approach requires suitable profiles to be generated for each storm, and also that storms can be generated that reflect the mixture of rainfall types encountered over a catchment. In this respect, it is useful to specify separate distributions for storms of "low" and "high" intensity, conceptually representing a distinction between steady, frontal rainfall, and intense convective storms (e.g. Cameron *et al.*, 2001). Further generalizations can be made, such as accounting for dependence between storm mean intensity and duration using a suite of duration classes, each with its own distributions of intensity. Seasonality can be incorporated by separate parameterization for different parts of the year, and different profile mechanisms can be used for storms of different intensity.

An alternative structure for stochastic rainfall models is based on simulation of pulses of rainfall. In this structure, rainstorm exteriors are modeled in terms of durations and arrival times, but each exterior contains individual rain cells, comprising pulses of individual intensity and duration. The total rainfall intensity at any time is the sum of the intensities of the active pulses. Two popular stochastic pulse mechanisms are the Neyman-Scott and Bartlett-Lewis models. These differ in the temporal pattern of rain cell arrival and survival, but in practice generate similar results.

The contrasting model types have been investigated by Cameron *et al.* (2000a) who compared a modified version of the exponential point model originally proposed by Eagleson (1972), a semiempirical profile-based model developed by Cameron *et al.* (1999) and the random-pulse Bartlett-Lewis gamma model of Onof and Wheater (1994). The three models were fitted to point data from three contrasting raingauge sites in the United Kingdom.

The exponential storm model was found to be too simple to represent features of the gauged records at a range of timescales. The model developed by Cameron *et al.*

(1999) performed well, benefiting from using empirical cumulative distributions for storm arrival times, durations, and intensities. Storm profiles were based on sampling a "library" of observed profiles. For data-rich situations, such as the test sites adopted, the empirical distributions would be expected to perform well, as was confirmed. A potential problem with all such empirical distribution models is extrapolation beyond data contained in the gauged record. This problem was addressed by Cameron *et al.* (1999) by use of a theoretical extension for the upper tails. However, the model was limited where storms were generated of durations or intensities for which there was only a limited sample of profile shapes available. Nonetheless, the rainfall model was thought to be suitable as an input for continuous runoff simulation.

The rectangular pulses Bartlett-Lewis model was found to work well, apart from its simulation of short-duration extreme rainfalls. Cameron *et al.* (2000a) attributed this to limitations of the gamma distribution, fitted on a seasonal or annual basis, to model pulse intensities. A suggested solution was to separate rain cells into "frontal" and "convective" intensity types. Hashemi *et al.* (2000) have demonstrated good performance for a Neyman-Scott pulse model structured in this way. In effect, the comparisons made in the United Kingdom, for rainfall stations featuring combinations of stratiform and convective rainfall, seem to indicate that either the storm-profile or rectangular-pulse algorithms may be suitable point models for generating continuous simulation inputs. Cowperthwaite *et al.* (1996) have reported a generalized parameterization of a point model, allowing regional application.

Spatial and Cluster Models

Structures in rainfall are also reflected in spatial patterns. In a recent review of progress in fluvial flood modeling, Wheater (2002) concluded that "the spatial and temporal variability of rainfall can be extremely important in influencing flood hydrograph shape and volume, but that the importance will vary greatly as a function of catchment and rainfall properties". As contrasting examples, this review cited Naden (1992b), who found that rainfall variability was smoothed out and had little influence on the flood responses in the largely permeable Thames catchment in southern England, whereas Michaud and Sorooshian (1994) had shown that a high spatial resolution (2 km) was needed to simulate flood peaks caused by convective rainfall in an arid environment in the southwest United States. Modeling studies, such as those of Ngirane-Katashaya and Wheater (1985) and Watts and Calver (1991) have shown the potential importance of storm velocities in determining flood magnitudes.

Wheater *et al.* (2000) have summarized key aspects of spatially structured stochastic rainfall models. One such model, by Cox and Isham (1988), features a hierarchical structure in which the highest-level objects are rainfall fields that are modeled as a temporal stochastic process, within which storms are modeled as a spatial process. The lowest level objects are rain cells, which are generated within each storm and may be clustered in space. Objects at all three levels may be assigned velocities. Clusters can be generated by creating cells as a temporal Poisson process from an origin, with the process ending after some time according to a prescribed distribution. Northrop (1997) has further generalized this approach to add more realistic spatial structure in which the origin of each storm cell has an additional spatial displacement.

Spatial-temporal models of this type are complex and require some ingenuity in order to fit the parameters. The ideal situation would be to make use of radar rainfall fields for model fitting; notwithstanding issues of radar calibration, such data offer a reasonable approximation to the continuous space-time fields that are represented in the model and thus allow derivation of the required statistical parameters. The data processing implications are substantial. If spatial-temporal rainfall models are to be applied successfully in driving rainfall-runoff models for flood simulation, there may be situations where rain gauge data have to be relied upon for parameterization. (Arguably, rain gauge data should in any case be incorporated in model fitting and testing, as these are a direct measurement of precipitation reaching the ground, unlike radar fields.)

A different approach designed to represent the movement of storm fronts over a catchment is the Modified Turning Band (MTB) method of Mellor (1996). This approach models the rainfall process with a layered structure. The first layer generates sets of intersecting parabolic prisms that move in the spatial plane and define a basic structure underlying a potential field for raincell generation. The second layer adds a modulating function that travels in the direction of the rainstorm and is used to represent the storm extent, as well as generating rainfall banding within the moving storm. Individual raincells are then generated as a Poisson process with occurrences according to the potential field derived from the larger scale layers. Mellor *et al.* (2000a) report that the method represents rainfall fields that display features observed in radar data, including raincells, cluster regions, and rain bands.

In an important application, Mellor *et al.* (2000b) have also shown how the MTB spatial rainfall model, calibrated on radar rainfall data, can be used to drive ensemble simulations of flood hydrographs. The authors made use of two rainfall-runoff models, the first being a detailed distributed process-based model (SHETRAN) and the second being the ARNO conceptual model, which handles runoff production in a way that is similar to the "semidistributed" approach taken in the Probability Distributed Model (PDM) (*see* the text below). The ARNO model was supplied with rainfall aggregated up to the catchment scale. It was concluded that

simulated floods were sensitive to the spatial distribution of rainfall, particularly after dry conditions. Whilst the Mellor *et al.* (2000b) study was set in the context of flood forecasting, rather than flood frequency, it would be a natural extension of the approach to simulate long rainfall sequences to provide design flood estimates. A significant finding was also that the simpler, semidistributed model of runoff production was adequate for flood generation.

RAINFALL – RUNOFF MODELS FOR CONTINUOUS SIMULATION

The choice of a continuous simulation rainfall-runoff model depends on the proposed application and available data. Models that have been most useful in continuous simulation studies tend to have been in the category of "semidistributed" conceptual models (*see* Beven (2001) and **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3– Chapter 134, Downward Approach to Hydrological Model Development, Volume 3** for a taxonomy and further discussions of rainfall-runoff models). Such models have structures that are based on process knowledge, some facility to describe spatial and temporal variations in runoff production, and, perhaps above all, are designed to be parametrically efficient and hence capable of robust calibration. These models are also, of course, far from complete representations of catchment hydrology and their parameters are therefore not generally measurable directly in the field. Indeed, there have been a number of independent studies (Beven, 1993, 2002; Grayson *et al.*, 1992), arguing that even in explicitly physically structured models that build on the grid-based modeling "blueprint" of Freeze and Harlan (1969), physical parameters such as hydraulic conductivity and porosity become effective parameters.

The following models have been selected as representative of those applied to date to simulate flood frequency curves, although it should be understood that this is by no means a definitive selection. The models are notable for having been used for continuous simulation in different contexts, ranging from very detailed, catchment-specific engineering studies, through explorations of process controls to generalized, national applications. There are many other models, often based on similar concepts, that could also be described, but the chosen examples are typical. Boughton and Droop (2003) have recently reviewed various hydrological models applied in software systems used for continuous simulation in Australia, the United States, Europe, South Africa and elsewhere. The emphasis here is on the conceptual basis of the modeling approaches.

Distribution Function Models

TOPMODEL

As a compromise between a conceptual structure based on process knowledge and a parsimonious parameterization, the catchment model TOPMODEL (Beven, 2001; Beven and Kirkby, 1979; Beven *et al.*, 1995) is well suited to continuous simulation in environments where its assumptions are reasonable. The main assumptions are that the distribution of subsurface storage is driven by gravity drainage that follows topographic control, and that the transient dynamics of the saturated zone rapidly lead to a spatial equilibrium in which local storage has a unique relationship with the catchment average storage, and a few soils and topographic parameters. This concept works well in environments where runoff production mechanisms involve a shallow, near-surface water table that may intersect the surface at the base of slopes or in areas of flow convergence, causing saturated areas of rapid storm runoff as overland or shallow subsurface flow. The basic structure can be adapted to include other runoff production mechanisms, such as infiltration-excess overland flow, and also to relax some of the usual assumptions to cater for deep soils, spatially variable soil properties and other factors.

The distribution of storage in the simplest form of TOPMODEL is determined by topography via an index of upslope drained area, a, divided by slope, $\tan\beta$, which is used to approximate the hydraulic gradient of the saturated zone. This index can be calculated conveniently from a digital elevation model (DEM) using automated procedures. Figure 3 shows in schematic form how the area/slope index is defined and one, typical, arrangement of stores to account for subsurface moisture. Each "block" shown in cross section corresponds to a discrete class of values of the topographic index.

In most implementations of the TOPMODEL concepts, the local storage deficit due to gravity drainage (D) is expressed as a function of the form

$$D = \overline{D} + m\left[\gamma - \ln\left(\frac{a}{T_0 \tan\beta}\right)\right] \tag{15}$$

where $\bar{D}$ is the areal average storage deficit, γ is the areal average of the index $\ln(a/T_0\tan\beta)$ and T_0 is the lateral transmissivity of the soil profile when $D = 0$. The parameter m controls the rate of an assumed exponential decrease in transmissivity with depth, which corresponds to a local downslope saturated flow rate

$$q = (T_0 \tan\beta)e^{-D/m} \tag{16}$$

for a slope element of unit width. It is often assumed that the lateral transmissivity of the soil when just saturated has the same value throughout the catchment, although versions of TOPMODEL used to model flood frequency have instead allowed saturated transmissivity to be described by a statistical distribution. The saturated zone is often modeled as a lumped nonlinear store and the unsaturated zone can

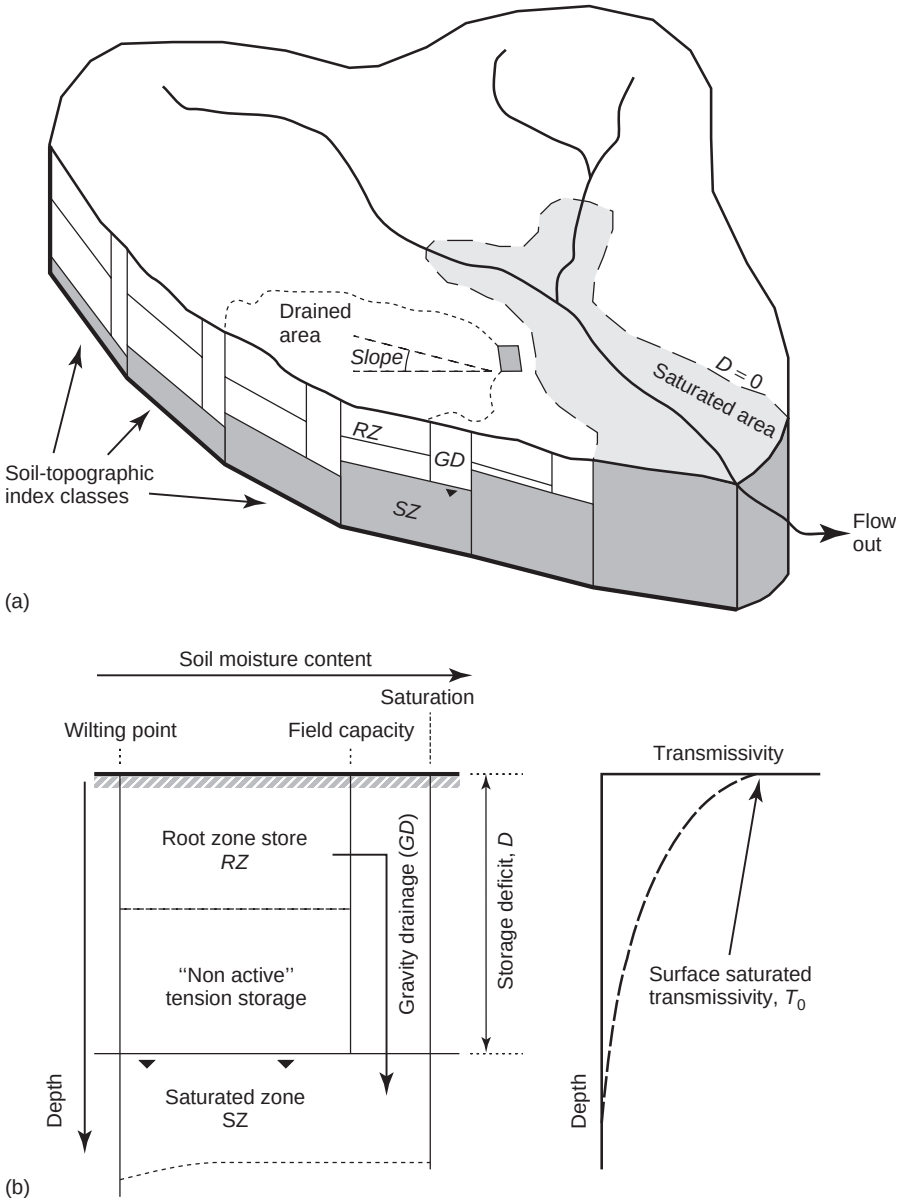


Figure 3 Schematic diagram of TOPMODEL

be represented in simple form by a root zone store, which has a uniform maximum depth. When filled by rainfall, any excess is routed to the saturated zone via a simple linear time delay. Evaporation is lost from the root zone as a function of the potential rate (supplied to the model as an input) and storage in the root zone.

The TOPMODEL concepts are simple representations of runoff production that can be applied with the estimation of as few as four parameters. Research applications for flood frequency estimation include those of Beven (1986, 1987), Blazkova and Beven (1995, 1997, 2002, 2004), Cameron *et al.* (1999, 2000b,c), Robinson and Sivapalan (1997) and Sivapalan *et al.* (1990). Most descriptions of TOPMODEL rely on the conceptualization of the catchment in terms of storage elements, although an alternative derivation (Kirkby, 1997) begins with the kinematic wave theory, with runoff being produced from linear strips on the hillslope draining into the stream network. The storage-based accounting generally used to compute rainfall losses, and the additional complexity of the spatial arrangement of the hillslope elements, preclude complete analytical solution of the theoretical model.

The PDM

The Probability Distributed Model (PDM) was developed by Moore and Clarke (1981). The distribution function

used to model runoff production in the PDM is based on statistical arguments. The PDM has been used in the development of a prototype national application of continuous simulation for flood frequency estimation in Britain (Calver *et al.*, 1999). The distinguishing feature of the model is that soil moisture storage is represented by a continuous distribution of storage capacity, c. This is shown schematically in Figure 4, where the stores are arranged on the left in rank order of capacity from zero to a maximum depth $c_{\max}$ and the probability distribution $f(c)$ for the storage capacity is shown to the right. The soil moisture store permits runoff to respond nonlinearly to rainfall inputs by varying the proportion of the catchment generating fast, "direct" runoff according to the amount of water stored in the soil. In the example shown in Figure 4, rainfall of depth P has fallen over an initially dry population of stores. The storage capacity will be filled for all parts of the catchment for which c is less than a critical value, c^* (which equals P for the dry initial condition). At subsequent times, c^* can be updated according to the time sequence of rainfall inputs and losses from the soil moisture store. The proportion of the catchment for which $c \leq c^*$ can be calculated from the cumulative distribution function

$$F(c) = \Pr(c \leq c^*) = \int_{\min(c)}^{c^*} f(c)\,\mathrm{d}c \tag{17}$$

This partial area is taken to be the area producing direct storm runoff, which is then routed, typically via one or more linear storage functions, before contributing as "quick flow" to the total catchment runoff. The remaining rainfall enters the soil moisture store, which is depleted by evaporation. Drainage from the soil moisture store (representing recharge) then takes place, typically via a nonlinear storage function, and generates baseflow after being routed by a further "slow flow" store.

The form of the storage functions used in the PDM can be varied as described by Institute of Hydrology (1996). A particular choice of functions may be best suited to any given catchment. The soil moisture storage capacity is often assumed to follow a Pareto distribution, for which the distribution function is

$$F(c) = 1 - \left(1 - \frac{c}{c_{\max}}\right)^b \tag{18}$$

where the parameter b controls the variability and the minimum value of storage capacity found in the catchment has been assumed to equal zero. Moore (1985) describes how the state of the soil moisture store is integrated over the distribution at each time step, as c^* varies, to calculate runoff and the volume of water retained in storage.

Rainfall inputs to the soil moisture store can be multiplied at every time step by a volume adjustment parameter. Although introduced notionally to compensate for errors arising in the estimation of spatially averaged rainfall from point gauge data, this parameter can also effectively compensate for components of the overall water balance not fully represented in the model, including losses to regional groundwater or gauge bypassing. Losses due to evaporation can be calculated as a function of potential evaporation and the status of the soil moisture store.

For the development of a generalized continuous simulation method, the choice of model may be determined by the need for a standardized formulation with a small number of parameters. The generic PDM structure has been

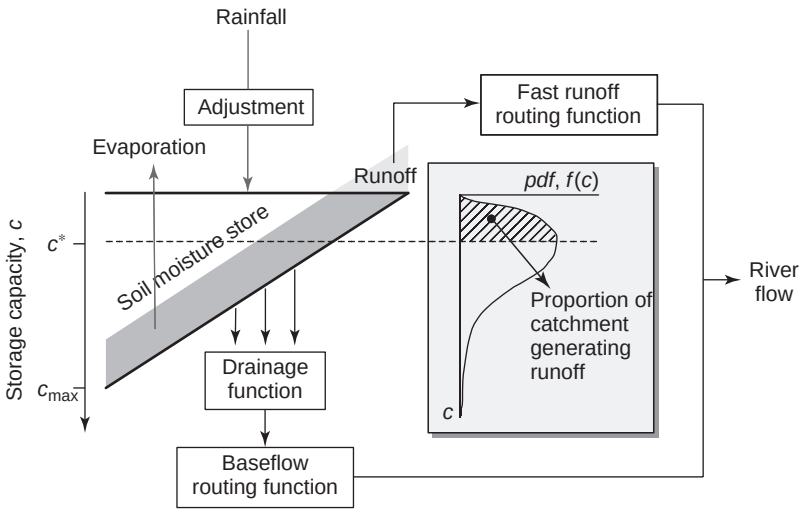


Figure 4 The Probability Distributed Model soil moisture storage concept of Moore and Clarke (1981)

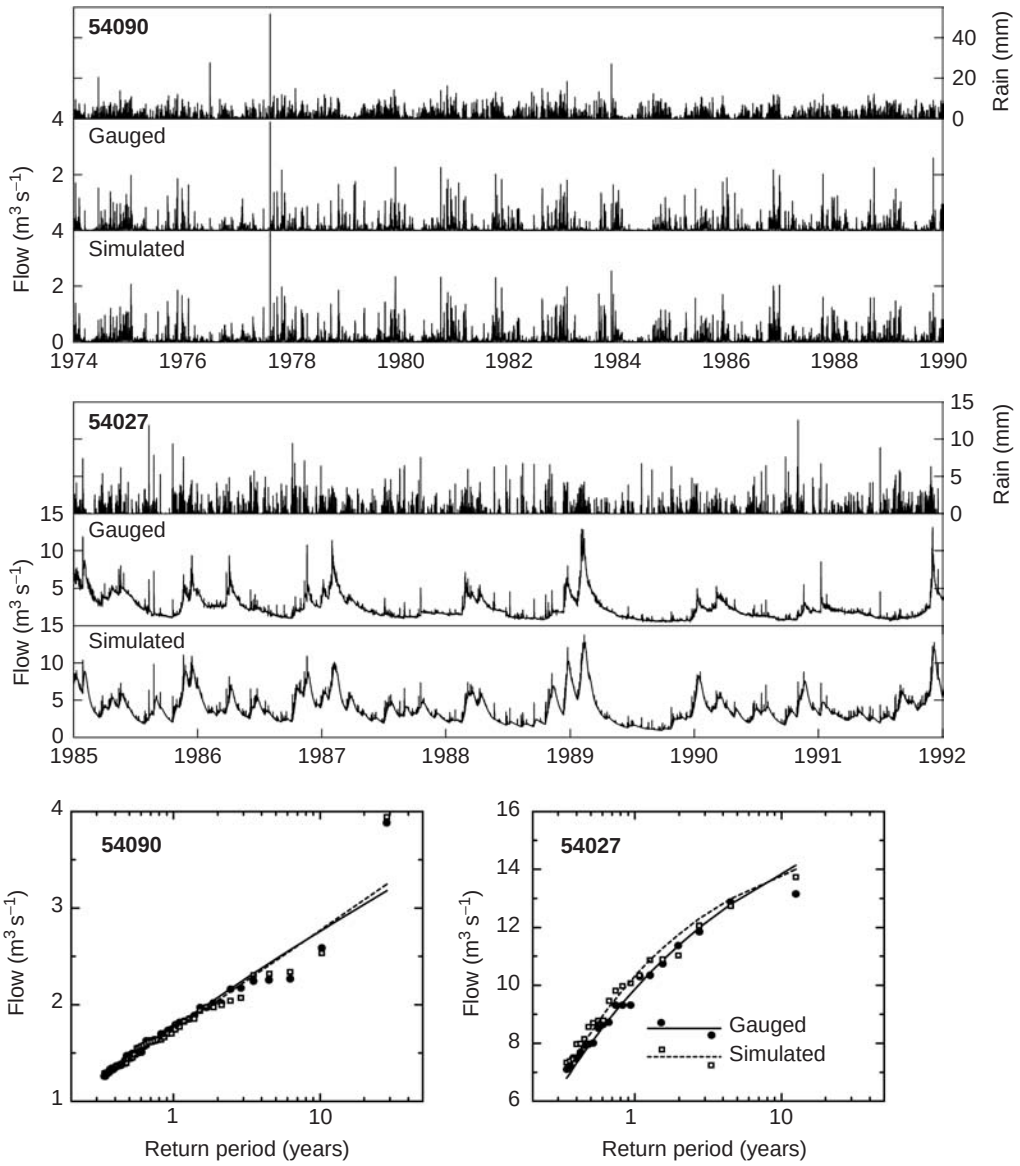


Figure 5 Long time series simulations and derived flood frequency distributions using the PDM rainfall-runoff model (Reproduced from Lamb, R & Calver, A 'Continuous simulation as a basis for national flood frequency estimation' in Continuous River Flow Simulation: Methods, Applications & Uncertainties edited by Littlewood, I. British Hydrological Society Occasional Paper 13 (2002) 67–75 by permission of British Hydrological Society)

found flexible enough to derive flood frequency distributions that agree well with distributions fitted to gauged flow records (up to the limit imposed by the record length) for a range of catchment types. Two contrasting examples are shown in Figure 5. The top panel shows rainfall, observed flows and flows simulated by the PDM after calibration for the Tanllwyth stream, a very small ($\sim$1 km^2) catchment with average annual rainfall of 2425 mm, shallow soils and a largely impermeable solid geology. The bottom panel shows calibrated model results for the River Frome, a 198 km^2 catchment with annual average rainfall of 827 mm and a highly permeable geology in its headwaters.

Although not strictly an application to flood estimation, the concept of the probability-distributed soil moisture store can also be used directly to derive expressions for the cumulative distribution of discharge (but not necessarily of peak flows). Hosking and Clarke (1990) derived such a result by assuming a Markov chain model for the occurrence of rainfall, probability distributions for rainfall and evaporation depths and a linear routing function for runoff to reach the catchment outfall. There are a number of other models in widespread use that apply distribution function concepts similar to those of the PDM, including the VIC (Variable Infiltration Capacity) model (Liang *et al.*,

1994), the Xinanjiang model (Zhao, 1992) and the ARNO model of Todini (1996).

The Time-Area Topographic Extension (TATE) Model

The "Time-Area Topographic Extension" (TATE) model is a parsimonious transfer function model developed by Calver (1993, 1996). It has been used alongside the PDM in the national continuous simulation pilot study of Calver *et al.* (1999). The principles of the TATE model can be traced back to the time-Area approach of Clark (1945). However the TATE model includes further recognition of hydrological processes by allowing runoff responses to vary over the catchment. The motivation for this was the observation that wetter conditions tend to prevail close to the channel leading to an increase in the wave speed of runoff responses, and a consequent reduction in response times. The reference to topography in the model's name follows from the use of the slope and stream channel configuration to define the relationship between drainage area and distance from the channel or catchment outlet, given a distribution of response velocities.

In principle, the relationships between distance from the channel/outlet, response velocity and drained area could be derived empirically from the analysis of topographic slopes and flow paths. However, Calver (1996) adopted a simple functional form in which the relationship between the response velocity, c (effectively a wave speed), and distance from the channel or outlet, s, was expressed as

$$c = k\left(1 - \frac{s}{s_{\max}}\right) = kx \tag{19}$$

where s was scaled by the maximum flow path or hillslope length $s_{\max}$ and $x = 1 - (s/s_{\max})$ is the dimensionless distance from the top of the hillslope. This relationship applies on a hypothetical flow plane or slope element. The relationship between a position on the slope and the corresponding cumulative fractional response area of the catchment, $A_s/A_{\max}$, was defined as

$$x = \left(1 - \frac{A_s}{A_{\max}}\right)^{1/\chi} \tag{20}$$

where χ is a parameter. Calver (1996) showed that Equation (20) with $\chi \approx 3$ would be a good approximation for a large number of catchments in Britain.

Equation (19) describes a physical response on a hillslope, or more generally along a flow pathway (which may include parts of the channel network), while equation (20) relates this response to the spatial layout of the catchment. Combining the two functions gives the expression

$$\frac{A_s}{A_{\max}} = 1 - e^{-\chi k\tau} \tag{21}$$

which defines the runoff response time, τ, of the partial catchment area $A_s/A_{\max}$. Equation (21) allows the catchment to be split into drainage area increments of equal response time, which can be convolved with a time series of rainfall excess to compute runoff from the catchment. In this respect, the TATE model is a form of UH, expressed as a linear transfer function. In its derivation, it shares some concepts with the GUH, which involves the idea of a distribution of time delays for runoff on hillslopes, but is also predicated more explicitly on the topology of the stream channel network.

The basic physical response described by equation (19) can also be related to the physical interpretation for runoff responses on a hillslope provided by the kinematic model discussed earlier. The wave speed is given by equation (11), which applies along characteristics of constant flow depth. The limiting characteristic defined by $y(x = 0, t = 0) = 0$ originates at the top of the slope at the start of rainfall excess and the plane reaches steady state when this characteristic reaches the base of the slope. After this time, $q = xi$ for constant and uniform rainfall excess. By substituting for q in the flow law specified by equation (9), the response velocity of a disturbance under conditions of unit rainfall excess can therefore be written

$$c = \alpha m\left[\frac{x}{\alpha}\right]^{(m-1/m)} \tag{22}$$

Equation (22) shows in general that the response velocity increases downslope, which is a premise of the TATE model. However, equation (19) specifies that the increase in response velocity should be a linear function of distance. The only flow law that can satisfy this assumption is a logarithmic form

$$q = \alpha e^{my} \tag{23}$$

in which case $c = mx = m(1 - s/s_{\max})$ for unit rainfall excess at steady state. This type of flow law (which is also assumed in TOPMODEL, *see* equation (16)) has been shown by Kirkby (1997) to be a suitable form for simplifying the dynamics of runoff responses as a succession of steady states because it tends to cause transients caused by changes in rainfall rate to disappear rapidly. In the TATE model, the simplified physical response is then used to define the form of a transfer function for the runoff response at the catchment scale.

The TATE model provides a compact and parsimonious structure that can be related to some aspects of catchment form, adding to its potential for generalization to ungauged sites. Its conceptual stores can be adapted as a basis for further functions, including calculation of groundwater recharge and levels (Calver, 1997), where the additional complexity is warranted.

Gridded Modeling

The three models described above have been presented as applied on a catchment basis. Adopting catchment units as the spatial basis for estimating flood frequency is convenient and consistent with existing approaches applied for event-based modeling. However, there are potential advantages in adopting a grid-based model that can be linked directly to raster data for land use, rainfall, and the river network topology. Indeed, gridded applications exist for the basic runoff production concepts used in the models already described. Another approach in rainfall-runoff modeling has been to develop models that attempt to solve the full equations for flow (or appropriate simplifications) in two or three dimensions. This "physics-based" approach to modeling is often traced back to the early work of Freeze and Harlan (1969) and is discussed in detail in **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3** and **Chapter 127, Rainfall-runoff Modeling: Distributed Models, Volume 3**. Models of this type have not been as widely used as conceptual models for simulation of flood frequency (although a physically based model, not requiring calibration, could clearly be a very useful tool for flood estimation). One exception in flood frequency analysis has been applications of the SHETRAN model (Kilsby *et al.*, 1998).

Climate and Land Use Scenario Simulation in Catchments (CLASSIC)

A grid-based model that has been used for flood frequency estimation is called "CLASSIC" (Climate and Land use Scenario Simulation In Catchments), which was developed by Crooks *et al.* (1996). This model is mentioned here because it has been used in the United Kingdom for studies of climate change impacts (Reynard *et al.*, 2001) that have informed government planning policy guidance on flood risk (Office of The Deputy Prime Minister, 2001). Modeling the impacts of change often requires a continuous simulation approach to account for changes in hydrological response caused by the combination of changes in boundary conditions at different spatial and temporal scales (*see* **Chapter 132, Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change, Volume 3** and further discussion below).

CLASSIC works on grid cells of comparable scale to an elemental subcatchment unit; in the United Kingdom, a 40 km^2 grid cell size is a reasonable choice, and corresponds to the grid scale adopted for calculation of potential evaporation within the Met Office Rainfall and Evaporation Calculation System (MORECS). Other grid scales have also been used. The CLASSIC model accounts for losses using a simple, conceptual storage-based soil water balance module to calculate an effective rainfall sequence. Runoff production and attenuation is then computed using a linear transfer function model within each grid cell; the transfer function formulation employed is that of the "IHACRES" model (Jakeman *et al.*, 1990; Littlewood and Jakeman, 1994).

The conceptual soil moisture model has minimal parameters that have been related to broad land use categories within any grid square. A notable feature of the model is that the parameters of the runoff component, initially calibrated at 22 test catchments in the United Kingdom, have also been related to catchment properties, in this case to digital soils and land surface data, available for Britain on a 1×1 km grid. These relationships open the way for generalized applications to ungauged catchments.

The hydrological model performs soil moisture and runoff calculations separately within each component grid square. Runoff is then routed to the basin outlet using the network width function approach of Kirkby (1976). This approach makes use of the density of channels in the basin at each spatial step away from the outlet, which is convolved with a "standard" response function to produce a network response function (i.e. a form of UH). In CLASSIC, each grid square has its own network width function. The "standard" routing function, which applies over all grid cells, is based on a solution of the diffusion equation used by Naden (1992b) as a suitable routing model in moderately shallow sloping rivers. The result is that each grid square has its own UH for routing to the main basin outlet, with more distant grid squares naturally exhibiting greater attenuation and lag, as illustrated schematically in Figure 6.

CLASSIC has performed well in modeling daily streamflow and flood frequency (based on daily mean flows) within three contrasting basins of up to nearly 10 000 km^2, covering most of the main types of soils, geology, and land use in the British Isles. The grid-based structure makes this type of model particularly suitable for integration with GIS tools, and its direct link to soils data and simple representation of losses has been exploited to model effects of land use change. In general, calibration has been carried out for the main basin outlet and some key internal gauging stations. However, the model has been found to give good matches to gauged data at subcatchments where no calibration was carried out. Figure 7, for example, shows gauged and modeled daily flows for the Thames at Kingston (9948 km^2), a location that was used for calibration. Results are also shown for the Mole at Esher (470 km^2) and the Cherwell at Oxford (907 km^2) where the results depend on the parameterization derived from land use and soils data (after global calibration at Kingston); Figure 6 shows the locations cited.

INVESTIGATING CONTROLS ON THE FLOOD REGIME

As a research tool, early developments of continuous simulation stemmed from the potential to improve upon the

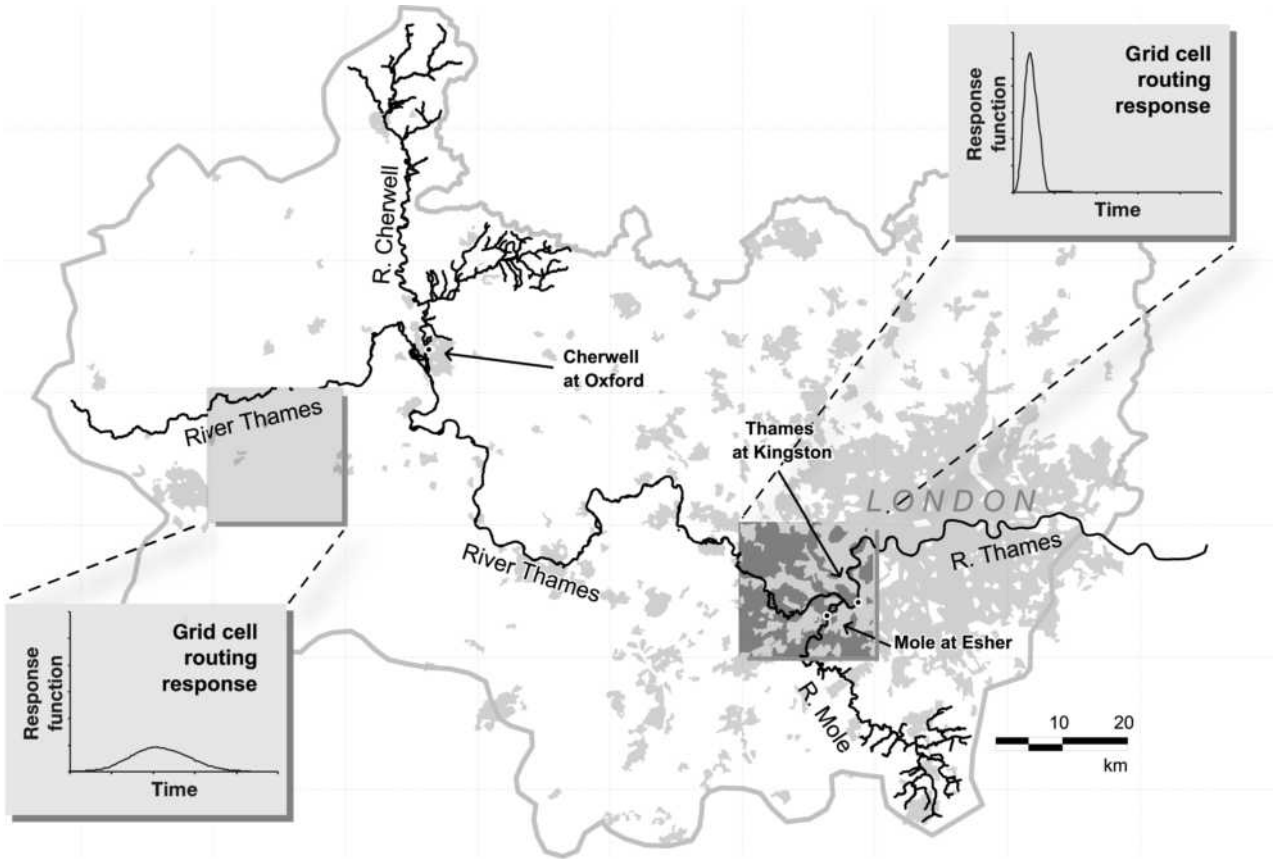


Figure 6 Application of the ''CLASSIC'' model of Crooks *et al.* (1996) to the Thames catchment, England. Triangle symbols show the locations of model simulations plotted in Figure 7. The expanded schematic plots show typical network response functions at a downstream point for two grid squares, with a more attenuated response being produced for runoff from distant parts of the catchment

catchment process representations in derived distribution models. If flood frequency can be modeled using physically based concepts, then there is the possibility of using such models to improve understanding of controls on flood regimes. This was the route explored by Beven (1986), who combined a stochastic rainfall model, based on the distributional assumptions used by Eagleson (1972), with the conceptual rainfall-runoff model TOPMODEL. In this study, the aim was to investigate the controls on catchment flood regime. The version of TOPMODEL applied in this case included both infiltration-excess and saturation-excess overland flow mechanisms. The study areas were three small ($\sim$1 to $\sim$10 km^2) catchments in United Kingdom in an area experiencing typically $\sim$1500 mm annual rainfall. Modeled runoff responses were dominated by saturation-excess flow, as would be expected from field knowledge for the catchments chosen in the study. Saturation excess produced rapid responses with high runoff coefficients. The dominant control on the flood frequency curve was, therefore, the shape of the rainfall growth curve, rather than catchment parameters.

A more complex response was noted by Sivapalan *et al.* (1990), again using TOPMODEL distribution function concepts. Experimentation with different modeled rainfall regimes suggested that the flood frequency curve was controlled at lower magnitudes by saturation-excess runoff production, which, having a "memory" of antecedent conditions, modified the rainfall growth curve. However, for larger, lower-probability, events, it appeared that an infiltration-excess mechanism would dominate runoff production, passing the slope of the rainfall growth curve more directly into the flood frequency curve.

Robinson and Sivapalan (1997) used an exponential stochastic rainfall model with a linear storage model of catchment runoff production to investigate effects of different temporal rainfall patterns and scaling. The study also examined scaling of catchment size (by an empirical scaling law for time of concentration). Catchments were categorized into distinct regimes based on the ratio of storm durations and inter-storm periods to catchment response time. The analysis showed that within-storm variation has greatest impact on flood frequency in fast-responding

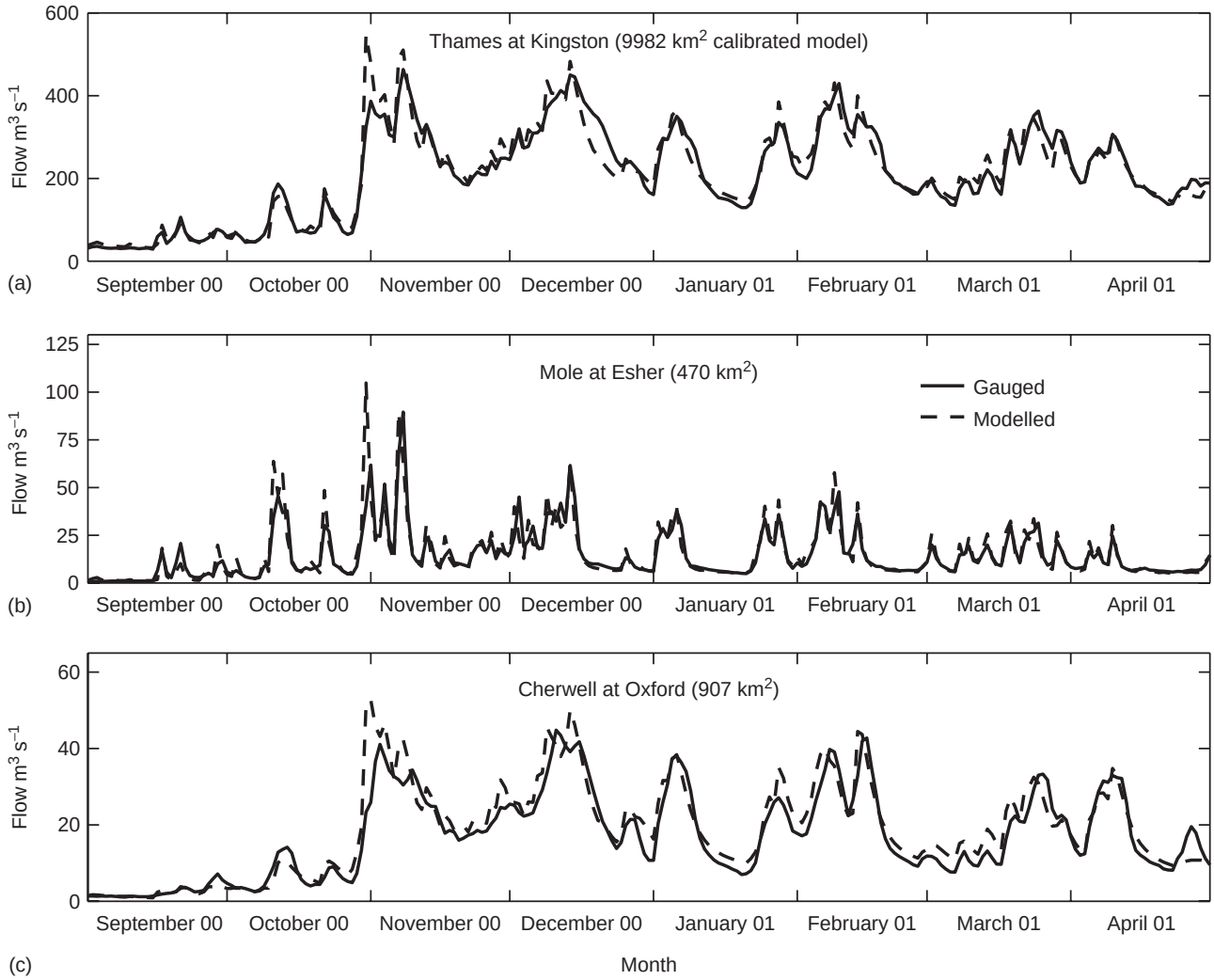


Figure 7 Daily mean flows simulated using CLASSIC

catchments, whereas multiple storms and seasonality have greater influence in slower catchments. The scaling of the mean and coefficient of variation of flood magnitude with area varied according to which "response time regime" a catchment occupied. In general, it would appear that nonlinearity in the runoff production and routing processes contributes to steeper flood growth curves (Bloeschl and Sivapalan, 1997). Such nonlinearity is likely to be associated with variable antecedent wetness, at least in temperate climatic zones.

The discussion here has centered on using modeling to investigate how the interaction of rainfall patterns and runoff production mechanisms controls the flood frequency curve. Continuous rainfall-runoff modeling is useful in this context because it is a technique that takes explicit account of patterns of precipitation and loss mechanisms in time (and, to some degree, in space). The effects of sequences of rainfall events or wet and dry periods can therefore be modeled, and associated with probabilities if the driving variables are modeled in a stochastic framework. The same features of continuous simulation can also be exploited for modeling the effects of climate or land use change on the flood regime. Rainfall-runoff modeling for assessing the impacts of climate and land use change is discussed in detail in **Chapter 132, Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change, Volume 3**.

CALIBRATION

The calibration of hydrological models has been a subject of extensive research (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**). A few key points relevant to continuous simulation for flood frequency are noted here. Calibration strategies in the context of continuous simulation vary according to the type of model selected and the quality and quantity of rainfall and flow

data available at gauged sites. Parameter estimation for ungauged or data-poor sites is considered separately below.

An advantage of using models that have parameters based on some conceptual representation of catchment processes is that it is possible to consider the plausibility of the fitted parameter values. However, the parameters are in general not truly physical properties because of the scaling, averaging, and (often considerable) approximation of physical processes within hydrological models. Calibration is therefore required to seek model parameter values that maximize agreement between simulated and observed flows. The following calibration strategies are well known:

- visual assessment of fit;
- manual optimization of an objective function;
- automatic optimization of an objective function;
- use of a "split sample" to calibrate and test on separate periods of record;
- "multiobjective" methods in which a set of Pareto optimal solutions are identified;
- Bayesian approaches, which seek a posterior distribution for model parameters, conditioned on sample data.

The development of Bayesian and multiobjective approaches have been motivated by the observation that there is often not a single, global optimum solution (i.e. set of parameter values) when calibrating a hydrological model, but rather a set of possible solutions, leading to uncertainty about the model predictions. As discussed in **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**, uncertainty in calibration can be reflected in the problems often encountered in trying to identify a single "best-fit" parameterization for a model. For instance, there might be long meteorological and flow records available for calibration at a gauged catchment, and hence no lack of data; however, even a conceptually well-founded catchment model is unlikely to be able to simulate every time step of the flow record successfully (especially at the short time steps often required for flood analysis) because of errors in measured data, errors in model specification and also errors arising from the differences in scales over which processes occur and are modeled.

Measures of fit based on the sums of squared errors at each time step have commonly been used to try to optimize rainfall-runoff models, or to report on the goodness of fit of simulations. Although this approach may work well for individual events or short records, errors about longer river flow time series tend to exhibit complicated structures including unstable variances and serial autocorrelation. These features reduce the efficiency of objective functions based on summing the squares of the residuals (for theoretical discussions of the reasons why, and possible alternatives, *see* Sorooshian and Dracup, 1980; Kuczera, 1983; Romanowicz *et al.*, 1994).

Working with the PDM rainfall-runoff model and time series of hourly data of up to 20 years, Lamb (1999) showed how attempting to optimize the well-known Nash and Sutcliffe (1970) efficiency measure was not very effective in calibrating for flood frequency estimation. A variety of alternative objective functions that give greater or exclusive weight to flood peaks were more useful, but nevertheless still gave rise to uncertainty reflected in a lack of any unique, global "optimum" parameter set. One partial solution to this problem was to consider a combination of objectives, such as fitting to the magnitudes of ranked flood peaks and flow duration curve quantiles, and even visual inspection of the plausibility of the overall hydrograph. There are still uncertainties associated with trade-off solutions in such multiobjective problems (*see* Gupta *et al.*, 2002; Vrugt *et al.*, 2003 and **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**).

A different perspective on model calibration is to reject the idea that any single calibration solution exists. Instead, many different combinations of parameter values may be found that lead to equally or similarly plausible predictions. This "equifinality" concept suggests that model parameters can only be identified as distributions and that predictions should be expressed as intervals rather than a "best-fit" point or line. Applications of this approach to continuous simulation of flood frequency characteristics are provided by Cameron *et al.* (2000b,c) and Blazkova and Beven (2002, 2004).

In some cases, it may be possible to arrive at a satisfactory model calibration by a combination of tactics if a single solution is demanded for a particular application. However, with the possible exception of using continuous simulation to provide a more complete picture of flood volumes over the period of record at a gauged location, it is likely that real applications would require some extrapolation in time (using modeled rainfall), or in space (to ungauged locations). In both cases, the potential impacts of uncertainty in calibration are more serious.

GENERALIZED PARAMETERIZATION

A key requirement for a general method of flood estimation is to be able to carry out modeling across a range of catchments of different scale and type. As noted already, practical applications often include ungauged sites, where measurements of river flows are not available, and model calibration is therefore not possible. Overcoming this problem can be thought of as "spatial generalization". Some of the derived distribution studies attempted to do this by associating model parameters directly with physical properties. Generalization of rainfall-runoff models used for continuous simulation has recently been approached on

a more empirical basis, in common with the regionalization methods often used for statistical and event-based flood estimation.

Attempts have been made to generalize conceptual hydrological models in a number of studies, not necessarily motivated by flood estimation (e.g. Abdulla and Lettenmaier, 1997; Post *et al.*, 1998; Sefton and Howarth, 1998). Spatial generalization of model parameters remains a difficult problem for which there is no standard methodology. However, a typical approach is as follows:

- Select a set of gauged catchments covering a range of characteristics and with good quality rainfall and flow records.
- Calibrate hydrological model parameters at each gauged catchment.
- Fit predictive relationships between the calibrated parameters and physical catchment properties.
- Use the fitted relationships to estimate model parameters at ungauged locations.

Applying this approach in Britain, Calver *et al.* (1999) fitted univariate regression functions between calibrated "best" estimates of model parameters and catchment properties. This approach assumed that model parameters were independent, which is known to be an approximation. However, results at test catchments indicated for the first time that the continuous simulation approach could be developed successfully for flood frequency estimation on a national basis. Modeled peak flow series were generally consistent with statistical distributions of gauged flows and the results were considered promising for simulation of sites where calibration had not been carried out directly, and, by implication, for ungauged locations.

The same study also noted the difficulty of identifying unique, independent best parameter estimates for continuous simulation (even though the models applied, PDM and TATE, were only moderately complex). A response to this problem is to simplify the hydrological model structure with the hope of creating a more robust model, albeit at the expense of the description of catchment processes. Following their pilot study, Calver *et al.* (2001) therefore applied a simplified version of the TATE rainfall-runoff model, in which a number of the original parameters were fixed on the basis of average values (informed by earlier calibration) or as ratios, reducing the number of parameters for calibration from seven to three.

Table 1 shows the results obtained in terms of the unsigned differences between values on frequency curves fitted to gauged and modeled flood peaks, expressed as a proportion of the gauged value at specified return periods. The mean and standard deviation of the errors in Table 1 are aggregated over 40 test sites. The values shown are for return periods up to 20 years, which was chosen as a maximum value in view of the length of gauged records

Table 1 Summary statistics of generalized flood frequency estimation for Britain using the TATE rainfall-runoff model (Reproduced from Lamb, R & Calver, A 'Continuous simulation as a basis for national flood frequency estimation' in Continuous River Flow Simulation: Methods, Applications & Uncertainties edited by Littlewood, I. British Hydrological Society Occasional Paper 13 (2002) 67–75 by permission of British Hydrological Society)

Return period (years)		1	2	5	10	20
Full model[a]	Mean % error[b]	31	31	33	35	38
	sd % error	29	31	33	35	36
Simplified model	Mean % error	25	24	25	28	31
	sd % error	23	22	23	26	33

[a]Results after Calver *et al.* (1999).
[b]The quoted percentage errors are the difference between ordinates of distributions fitted to gauged and modeled flood peaks, expressed as a proportion of the gauged value. Quoted figures are mean and standard deviation (sd) over a sample of catchments, used both for calibration and subsequent testing of generalized model.

at the majority of test sites. It is clear that the simplified version of the TATE model gives better results than the full model when applied without site calibration (i.e. as if for an ungauged site). This is expressed both in terms of a reduced average percentage error at all of the return periods tested and in the somewhat reduced spread of errors over the group of test sites.

A second response to the problems posed by dependence between parameters in fitting a model for ungauged locations is to adopt a more explicitly multivariate approach. This can be done by concurrently seeking to fit model parameters at each site whilst relating the parameters to regional basin characteristics. Fernandez *et al.* (2000) attempted such an approach using 33 basins in the United States and found that it did not result in improvements in the ability to model streamflow at ungauged sites, although it did produce improved regional relationships for the model parameters.

An alternative approach to account for the interaction between parameters is a "sequential generalization" algorithm, in which model parameter predictor equations are derived one after another, with calibration of parameters later in the sequence made conditional on the generalized estimates for the earlier parameters. The approach is outlined in Figure 8 and described in more detail by Lamb *et al.* (2000). Clearly, the results depend on the order in which parameters are related to catchment properties. The sequence can be decided by visual inspection of profile plots showing the marginal distribution of goodness-of-fit functions against each model parameter, selecting the parameter that is most sensitive in terms of the goodness-of-fit function at each step.

Table 2 summarizes the trials of this method. Results are given first for the Calver *et al.* (1999) pilot study. The second set of results was obtained using the sequential method

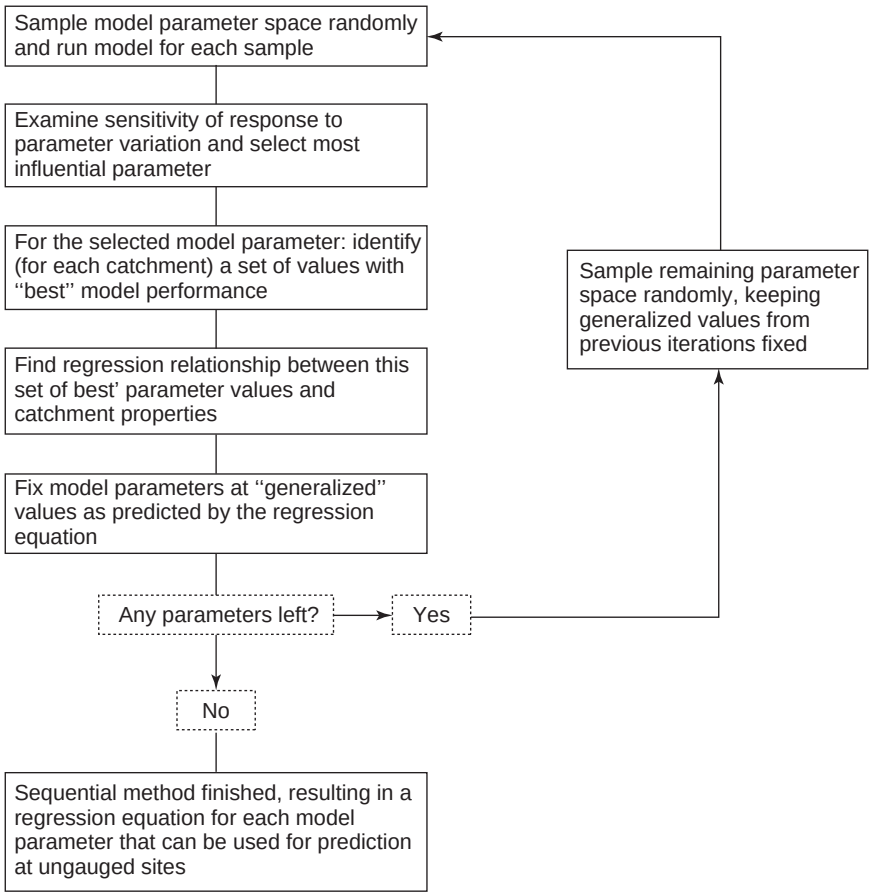


Figure 8 Outline flowchart for sequential calibration and generalization (Reproduced from Lamb, R & Calver, A 'Continuous simulation as a basis for national flood frequency estimation' in Continuous River Flow Simulation: Methods, Applications & Uncertainties edited by Littlewood, I. British Hydrological Society Occasional Paper 13 (2002) 67–75 by permission of British Hydrological Society)

outlined above. It can be seen that the revised generalization procedure, integrating "calibration" and "regionalization", has improved performance to a similar degree as the simplification of model structure discussed above. Figure 9 shows comparisons of flood frequency distributions fitted to gauged data and to flood peaks modeled using the generalized continuous simulation method (i.e. effectively treated as ungauged sites).

The overall conclusion from these experiments to explore the generalized application of continuous simulation has been summarized by Wheater (2002), who noted that the results provided "...the basis of a national methodology for continuous simulation modelling...". The same author commented that such modeling also provides the "...capability to analyse regional datasets and hence gain insight into the variability of hydrological response". In future, the link between model parameterization and process controls may be further tightened by regional analysis of the residuals in models to identify model deficiencies in a way that may not be apparent at an individual catchment.

Table 2 Summary statistics of generalized flood frequency estimation for Britain using the PDM rainfall-runoff model (Reproduced from Lamb, R & Calver, A 'Continuous simulation as a basis for national flood frequency estimation' in Continuous River Flow Simulation: Methods, Applications & Uncertainties edited by Littlewood, I. British Hydrological Society Occasional Paper 13 (2002) 67–75 by permission of British Hydrological Society)

Return period (years)		1	2	5	10	20
Univariate[b]	Mean % error[a]	42	40	40	42	45
	sd % error	35	33	34	37	44
Multivariate sequential	Mean % error	22	23	24	26	27
	sd % error	18	18	20	21	23

[a]The quoted percentage errors are the difference between ordinates of distributions fitted to gauged and modeled flood peaks, expressed as a proportion of the gauged value. Quoted figures are mean and standard deviation (sd) over a sample of catchments, used both for calibration and subsequent testing of generalized model.
[b]Results after Calver *et al.* (1999).

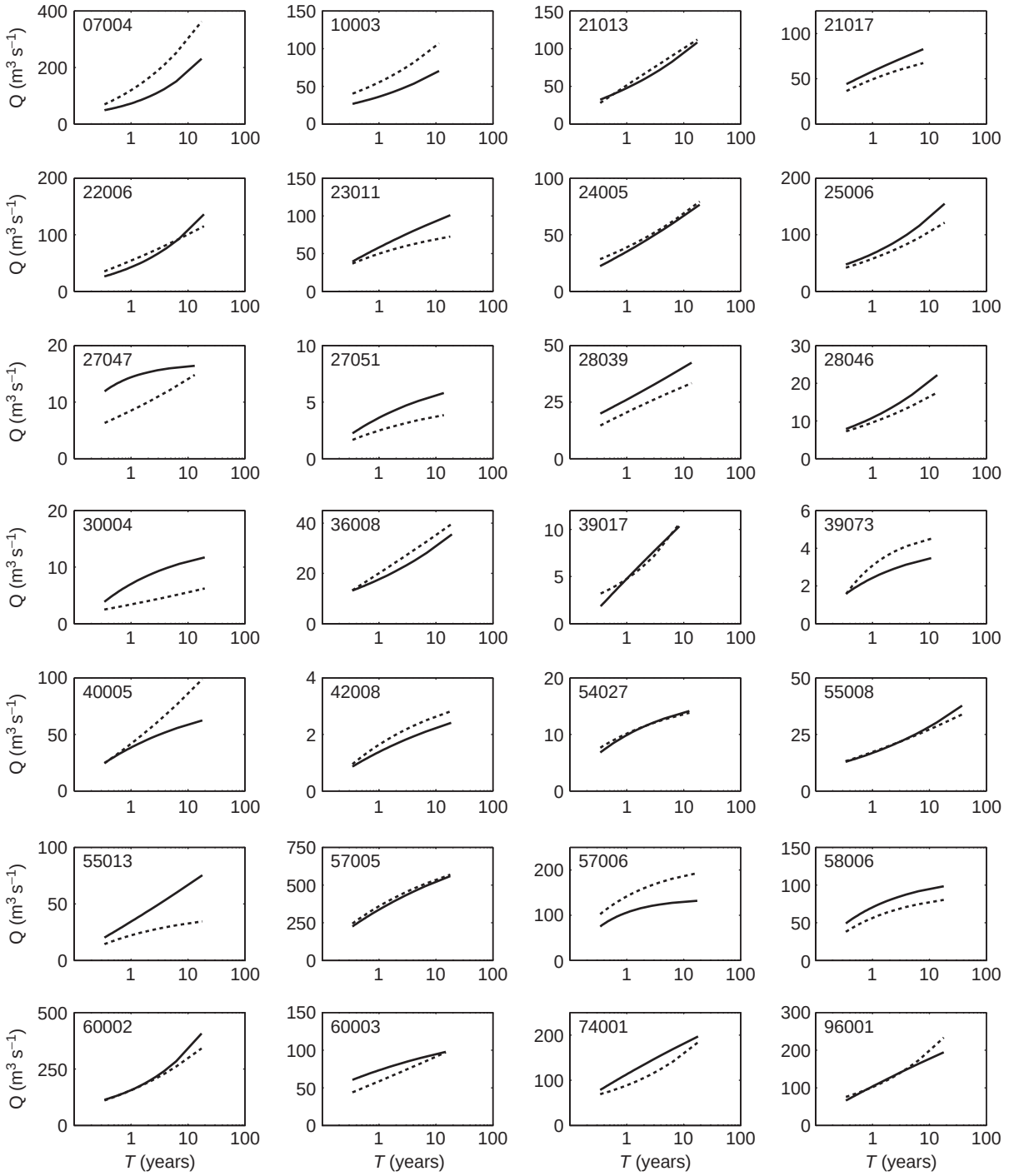


Figure 9 Flood frequency curves modeled using spatially generalized continuous simulation for sites in Great Britain. Solid lines are Generalized Pareto distribution (GPD) fitted to gauged flow data. Dotted lines are continuous simulation modeled results, with model parameters estimated for each site as if ungauged

FLOOD ROUTING AND HYDRAULIC CALCULATIONS

Flood routing is discussed in detail in **Chapter 124, Flood Routing and Inundation Prediction, Volume 3**; a brief summary will be given here to relate routing approaches directly to rainfall-runoff modeling for flood frequency.

Routing will be required for larger catchments and as a means of linking distributed inputs from discrete subcatchments. The type of approach adopted depends on

the time and space resolution of available data, the types of predictions required, and the physical characteristics of the river basin. The derived distribution studies discussed above include simplified flow routing, both for overland flows and water in the channel, based on the kinematic approximation and the GUH theory. Channel routing may be needed where flood storage, diversion channels, extensive systems of embankments, or changes in channel conveyance are important features, and an event or continuous simulation model is needed to provide input hydrographs.

Flow Routing

Simple hydrological routing can be based on conceptual stores that act to lag and attenuate an input hydrograph. The UH, although often derived from transfer function principles, is also equivalent to a linear storage-based routing approach. In many continuous simulation rainfall-runoff models, combinations of stores are arranged in series or in parallel. The approach can be likened to level-pool routing in that the conceptual stores neglect hydraulic slope, friction, and momentum, instead relating the rate of flow uniquely to the depth of storage. The approach only modifies flows, rather than water levels, but can provide suitable hydrograph attenuation, especially when calibrated.

The kinematic wave is also essentially equivalent to storage-based routing in that the dynamic equation neglects local accelerations and is based only on a functional relationship between depth (or storage) and flow rate. The kinematic wave routes a flood without changing the peak discharge unless there are lateral inflows or changes in channel characteristics, and becomes a weaker approximation for small Froude numbers, shallow slopes, and high resistance (Daluz Viera, 1983), but does have the advantage of providing a direct parametric link to hydraulic theory. A more complete physical approximation is the diffusion wave, in which the energy gradient and channel slope can differ. To generate a catchment-wide network response, these simplified flow routing functions can be solved for an explicit network of channel segments, as would commonly be done in engineering river models, or used within the network width function approach, described above as part of the CLASSIC model.

Conversion of Flows to Water Levels

In engineering applications, it will often be necessary to calculate the stage or absolute water level within the channel for design or planning purposes. The simplified routing methods discussed above need to be combined with separate hydraulic calculations to do this, unless an empirical relationship between flow and stage (a rating curve) is available, which may be the case at a flow gauging station or other hydraulic control. Where no such relationship exists, an alternative is to use a hydraulic model to construct a rating curve, for which local information will be needed on channel cross sections and surface roughness coefficients. Where detailed local data are not available, it may be sufficient to rely on simplified calculations assuming steady, uniform flow and applying the Manning or Chezy flow law.

The most detailed approach for channel flows generally used in practice is numerical solution of the hydrodynamic Saint-Venant equations (*see* **Chapter 124, Flood Routing and Inundation Prediction, Volume 3**), which account for friction, local accelerations, and momentum. The hydrodynamic approach can be used to compute flows and water levels throughout a reach. It represents the varying relationship between depth and flow rate as a flood wave passes through the reach, which may be particularly important where there are relatively shallow bed slopes and high resistance, for example, when water flows out of bank onto a floodplain.

Raster Two-dimensional Routing

An alternative routing approach is to make use of digital elevation data that can be produced in grid form to a high spatial resolution and vertical accuracy using LIDAR remote sensing. In this raster approach, each grid cell can be treated as an individual water store, with the flows between cells being calculated using a suitable representation, such as Manning's equation. Inflows can be added to any cell, and flows are propagated between cells by explicit time-stepping to model the evolution of flood extents. This is primarily a floodplain modeling approach, and where the channel itself needs to be modeled in greater detail (e.g. where it contains hydraulically significant structures), a one-dimensional channel model can be combined with the raster floodplain method to good effect. A full discussion of this approach to flood modeling is given in **Chapter 124, Flood Routing and Inundation Prediction, Volume 3**. However, the technique is robust and capable of being applied over large areas, with boundary conditions being provided by a rainfall-runoff model.

CONFIDENCE AND UNCERTAINTIES

Uncertainty at Gauged Catchments

One of the main sources of uncertainty in statistical flood estimation is typically the limited sample size, especially for single-site analysis. The use of pooled data helps overcome this to bring greater confidence to flood estimates. For an ungauged site, however, there is a further issue related to the confidence with which the index flood can be estimated, as well as the uncertainty about the pooling group or regional growth curve. Analytical expressions are available to derive confidence intervals for many such models; for example,

the shape of the likelihood function in the vicinity of the maximum likelihood parameter estimate can be used for a fitted single-site distribution. For more involved models, such as pooled or multisegment growth curves, resampling methods may be needed to provide a general means of establishing confidence intervals (*see* Faulkner and Jones (1999) for an application to rainfall frequency analysis).

Where rainfall-runoff models are used for flood estimation, the situation becomes rather more complicated. One of the main sources of uncertainty in predictions is the lack of uniqueness in calibrated parameters discussed in the context of calibration. Rainfall-runoff models applied in continuous simulation tend to be highly nonlinear (including thresholds in their responses) and the probability distributions of their parameters can be complex. Conceptual models are also capable of simulating multiple output variables (e.g. flood peaks, low flows, soil moisture) against which observational data may be compared and different measures of performance evaluated.

As a consequence of these complications, measures of uncertainty have tended to be based on importance sampling using MCS methods, or concepts of Pareto optimality (Gupta *et al.*, 2002). For example, Cameron *et al.* (1999, 2000b) used the GLUE (Generalized Likelihood Uncertainty Estimation) method of Beven and Binley (1992), a Bayes Monte Carlo approach, to compute uncertainty bounds for flood frequency curves. A number of approaches were used to "measure" model error, including the likelihood function for statistical distributions fitted to observed and simulated flood peak data. The rainfall-runoff model used by Cameron was TOPMODEL. Even though it has few parameters, it was shown that there could still be interactions within the model parameter space leading to uncertainty in the modeled flood frequency curves.

Generalized Model Uncertainty

A further issue is the uncertainty associated with generalization of continuous simulation modeling. In a recent study, Lamb and Kay (2004) have used MCS to generate confidence intervals for flood frequency curves at ungauged catchments. The sequential procedure discussed earlier was used to develop regression relationships between hydrological model parameters and catchment properties. Random samples were then generated from the distribution of the residuals surrounding each regression equation and used to calculate confidence intervals at a set of test catchments.

The confidence intervals calculated for the generalized continuous simulation were compared with intervals derived for a statistical distribution fitted to gauged data. The degree of uncertainty in the spatially generalized continuous simulation varied between sites but was often no greater than for the site-specific frequency curve fitted directly to the gauged flows. This procedure only took account of calibration uncertainty at the gauged sites in that the "calibrated" parameter estimates were regarded as if they were samples, subject to random error. However, it has provided a first assessment of the confidence that could be expected in continuous simulation flood estimates at ungauged sites.

BROADSCALE MODELING

The Problem

Flood frequency analysis is routinely used for engineering studies carried out at a very local scale. Point estimates of flood flows can often be used in such circumstances. At the wider catchment scale, however, upstream and downstream flood processes and management measures interact to determine the flood risk at any given location. For example, economic development may affect the production of runoff whilst the effects of engineered or natural storage may be propagated some distance along the river network.

In many places, there are formalized policies for the sustainable management of whole river basins and the environmental implications of flood management at the catchment scale have to be considered within this context. Hence, there is a need to model the joint effects of runoff production, runoff routing, and interactions with engineered and natural environments on a "whole catchment" basis. Modeling of this type can rarely consider every location in as much detail as a local study would allow. Instead, the methods employed need to be sufficiently generalized to cope with both data-rich and data-poor situations, and can be defined as "broad-scale modeling" (although this general framework might include, for example, very detailed hydrodynamic modeling of specific river reaches within a broad conceptualization of catchment hydrological response).

Rainfall-runoff Modeling

Currently, typical engineering practice would be to use event-based rainfall-runoff models to represent runoff production within a broadscale model. Modeling event hydrographs in this way provides a suitable input to routing or full hydrodynamic river models. Event models offer the facility to investigate, albeit crudely, possible effects of land management scenarios by adjustment of parameters such as the runoff coefficients and time-to-peak of the UH.

The event-based approach also has drawbacks in this context. Simulation on a catchment scale may require inputs to a river model in many places. Runoff production can be modeled independently at each node for a specified local rainfall (and hence flow) exceedance probability. Parameters of the design input package (e.g. storm duration and antecedent soil conditions) can be calibrated to provide a worst-case total flow when combined at a downstream point. However, the link between a specified level of risk

and the design inputs needed to generate that risk becomes tenuous unless the joint probabilities of the inputs are modeled. This is a considerable challenge. The problem is essentially that the concept of a catchment-wide "T-year storm" is not always a realistic concept, but is difficult to avoid with event-based methods.

Continuous Simulation

The continuous simulation method appears an ideal approach for broadscale modeling to represent the generation and propagation of floods, including the broad spatial variation of runoff production. Each of the main elements of the modeling may be as simple or sophisticated as the situation demands (or as the data allow). One of the first practical applications of at least a quasi-continuous simulation approach was motivated by the need to model a complex catchment system, including the operation of engineered river management schemes, in what could be interpreted as a form of broadscale model. This study, reported by Bras *et al.* (1985), involved a 46 397 km^2 basin in Kentucky, which was divided into 17 subbasins. A stochastic rainfall model was developed in which storm exteriors were modeled using a choice of distributions (dependent upon season) and with spatial correlation structures represented within the subbasins. Storm interiors were generated conditionally on a subset of the exterior properties by a stochastic process that preserved the observed within-storm temporal autocorrelation. Rainstorms were generated in chronological order, but, to cut down on the amount of ineffective data, were screened using a very simple "surrogate" rainfall-runoff model to retain only the most hydrologically important.

Bras *et al.* simulated runoff responses using an empirical model to generate daily subbasin runoff in consecutive 6-h blocks. The runoff data were transformed by a UH procedure and then routed through a hydraulic model that included the operation of reservoirs and simulated conditions at specified damage locations. The approach allowed different operational scenarios to be compared to examine management of the reservoir system. Bras *et al.* noted that "the advantage of the methodology is that it permits otherwise impossible analysis. It not only exploits available rainfall records which are many times better than their streamflow counterparts, but it also allows study of the system under varying operational scenarios".

The same comments could be applied to a study in which a network of rainfall-runoff models were driven by long series of spatially structured rainfall data, and the resulting flows taken as a proxy record from which the system state under differing levels of risk could easily be computed. This complete broadscale modeling framework would bring together the following elements:

- generalized, spatially consistent rainfall models
- generalized, conceptually realistic runoff production models
- integrated data storage and analysis tools.

All three of the elements listed above have been developed at least to a proof-of-concept stage, and in some cases as prototype applications.

At a practical level, continuous simulation has already proven to be useful in circumstances where flood risk

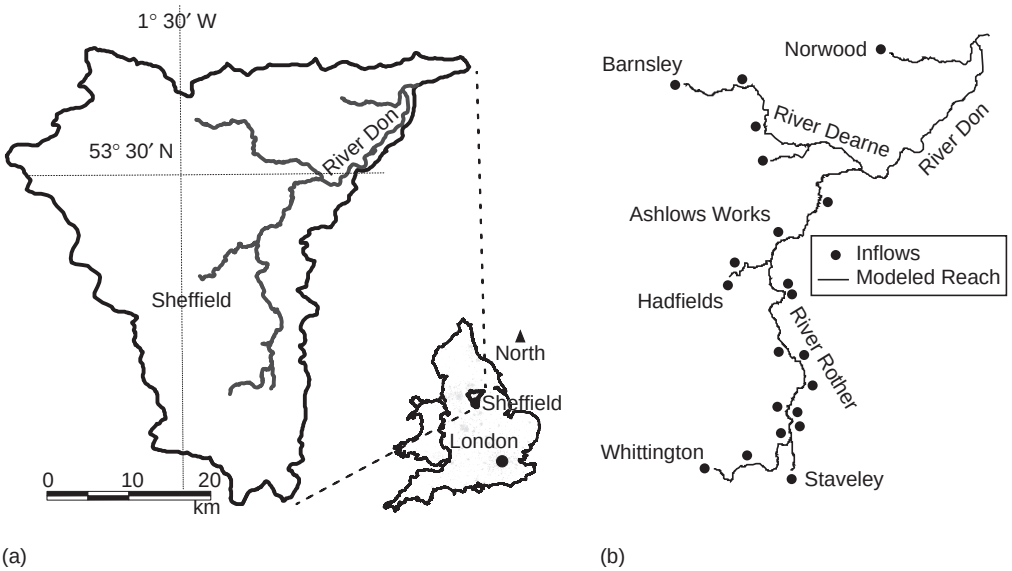


Figure 10 (a) Location of the River Don catchment. (b) Main river network and location of continuous simulation inputs to hydraulic model. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

within the catchment depends on the operation of engineered systems, where the flood risk at key locations is created by combinations of hydrological inputs, and where events can combine in different sequences. These issues alone may be sufficient to justify considering a continuous simulation approach for analysis of flood risk. If the flows at key locations within the catchment are sensitive to the combined inputs, and those locations are also physically sensitive to the flows (e.g. because defences could be overtopped), then continuous simulation is likely to be the best modeling approach. Modeling will be enhanced by any available gauged flow data that can be used for calibration, even relatively short records.

The following case study, reported in detail by Faulkner and Wass (2005), describes a situation where the above conditions all existed. The aim of the study was to generate maps indicating the extent of flooding at given probabilities for the catchment of the River Don, located in the central-eastern part of the United Kingdom (see Figure 10). This 1256 km^2 basin has various subcatchments of different elevation, slope, and hydrological response. It has a complex system of regulators and engineered storage (Figure 11), developed since 1958, which need to

(a)

(b)

Figure 11 Typical flood management structures within the River Don catchment. (a) Flow regulators to divert flow into washland storage. (b) Raised defences. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

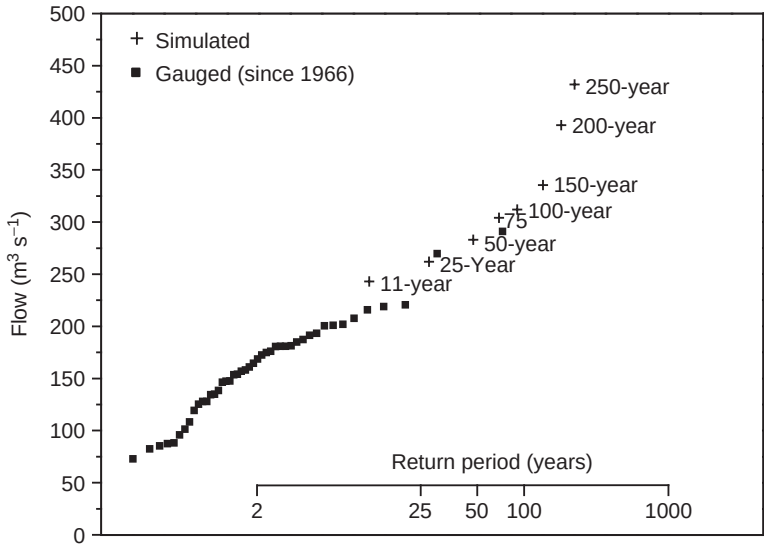


Figure 12 Gauged and simulated annual maximum (AMAX) flows as a function of return period using Gringorten plotting positions

be accounted for in any modeling, and they also make statistical frequency estimation difficult.

For the Don study, a point stochastic rainfall model was calibrated and applied to 20 subcatchment runoff models, with spatial variation in the rainfall to allow for differences in altitude. Each runoff model was a version of the PDM. Calibration was carried out at gauged subcatchments. For ungauged subcatchments, most parameters were transferred from calibrated catchments on the basis of catchment similarity. (The store parameters controlling the hydrological routing of storm hydrographs were estimated using empirical relationships derived with respect to existing UH models, for which general formulae are available.) A calibrated hydrodynamic model was used to connect the inputs in this study and to represent the operation of flood defences (although in other applications, simpler routing models could also be used for much of a catchment river network).

This catchment scale model was run for long (1000 year) sequences of synthetic rainfall data. The largest floods were extracted from the simulations and ranked to estimate the levels associated with specified return periods. Comparison of peak flows at the catchment outlet (Figure 12) with data gauged since 1966 (after the onset of regulation) would appear to support the results of the modeling.

What is significant about this study is the adoption of a complete continuous simulation approach, including rainfall, runoff, and river hydraulics, to solve a difficult catchment scale modeling problem. Whilst the initial outputs of the study were flood risk maps, it has been found that the large synthetic data set has other applications, such as the optimization of defences and the setting of flood warnings.

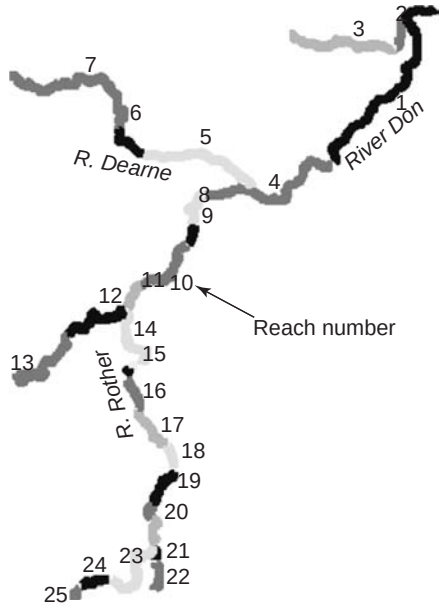


Figure 13 Don catchment main river network showing numbered reaches in which the 100-year flow arose from the same simulated event within each reach

The results illustrated the considerable spatial variation in the rank of the modeled event that constituted the local 100-year flood (Figure 13). There is clearly no assumption of a single, catchment-wide T-year event using this method, and the individual "events" making up the 100-year flood at each point vary in terms of antecedent conditions, hydrograph shape, volume, duration, and peak flow.

Concluding Remarks

Event-based rainfall-runoff models continue to be applied on a routine basis and are easily implemented in software packages. Despite some theoretical difficulties, the approach provides a practical solution for broadscale flood modeling when volumes and timing are important. Further development is likely to focus on resolving joint probabilities for design inputs and incorporating new data into event procedures to improve flood frequency estimates.

Continuous simulation models of runoff production are now becoming more widely used and understood, as are the techniques to apply them in combination with rainfall and river models. The most important challenge for continuous simulation is parameterization for ungauged catchments to provide generalized modeling of runoff production. Allied to this is the need for generalized rainfall models to allow flood estimation for long return periods. Continuous simulation is not yet widely accessible in "off-the-shelf" software form, but as the studies cited here indicate, the approach is now well established, both as a research method and in practice.

Acknowledgments

The author is grateful for help given by the following individuals and organizations during the writing of this review: Keith Beven (Lancaster University), Ann Calver, Sue Crooks, Alison Kay and Lisa Stewart (CEH Wallingford), Paul Wass and Duncan Faulkner (JBA Consulting – Engineers & Scientists). The two anonymous reviewers are thanked for their comments.

FURTHER READING

Franchini M., Hashemi A.M. and O'Connell P.E. (2000) Climatic and basin factors affecting the flood frequency curve: PART II – a full sensitivity analysis based on the continuous simulation approach combined with a factorial experimental design. *Hydrology and Earth System Sciences*, **4**, 483–498.

REFERENCES

Abdulla F.A. and Lettenmaier D.P. (1997) Development of regional parameter estimation equations for a macroscale hydrologic model. *Journal of Hydrology*, **197**, 230–257.

Acreman M.C. and Wiltshire S.E. (1989) The regions are dead, long live the regions: methods of identifying and dispensing with regions for flood frequency analysis. In *FRIENDS in Hydrology Bolkesjø (Norway)*, Roald L., Nordseth K. and Anker Hassel K. (Eds), IAHS Press: Wallingford, pp. 175–188.

Acreman M.C. (1990) A simple stochastic model of hourly rainfall for Farnborough, England. *Hydrological Sciences Journal*, **35**, 119–148.

Ashfaq A. and Webster P. (2002) Evaluation of the FEH rainfall-runoff method for catchments in the UK. *Journal of the Chartered Institute of Water and Environmental Management*, **16**, 223–228.

Benson M.A. (1968) Uniform flood-frequency estimating methods for federal agencies. *Water Resources Research*, **4**, 891–908.

Beven K. (1986) Runoff production and flood frequency in catchments of order n: an alternative approach. In *Scale Problems in Hydrology*, Gupta V.K. (Ed.), D. Reidel: Dortdrecht, pp. 107–131.

Beven K.J. (1987) Towards the use of catchment geomorphology in flood frequency predictions. *Earth Surface Processes and Landforms*, **12**, 69–82.

Beven K. (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, **16**, 41–51.

Beven K. (2001) *Rainfall-runoff Modelling – The Primer*, Wiley: Chichester.

Beven K.J. (2002) Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modeling system. *Hydrology Processes*, **16**, 189–206.

Beven K. and Binley A. (1992) The future of distributed models – model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Beven K., Lamb R., Quinn P., Romanowicz R. and Freer J. (1995) Topmodel. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Colorado, pp. 627–668.

Blazkova S. and Beven K.J. (1995) Frequency version of TOPMODEL as a tool for assessing the impact of climate variability on flow sources and flood peaks. *Journal of Hydrology and Hydromech*, **43**, 392–411.

Blazkova S. and Beven K. (1997) Flood frequency prediction for data limited catchments in the Czech Republic using a stochastic rainfall model and topmodel. *Journal of Hydrology*, **195**, 256–278.

Blazkova S. and Beven K.J. (2002) Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty). *Water Resources Research*, **38**(8), 1139, doi:10.1029/2001WR000500.

Blazkova S. and Beven K.J. (2004) Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic. *Journal of Hydrology*, **292**, 153–172.

Bloeschl G. and Sivapalan M. (1997) Process controls on regional flood frequency: Coefficient of variation and basin scale. *Water Resources Research*, **33**, 2967–2980.

Bobée B. and Rasmussen P.F. (1995) Recent advances in flood frequency analysis. *Reviews of Geophysics*, **33**(S1), 1111–1116.

Boughton W. and Droop O. (2003) Continuous simulation for design flood estimation – a review. *Environmental Modelling and Software*, **18**, 309–318.

Bras R.L., Gaboury D.R., Grossman D.S. and Vincens G.J. (1985) Spatially varying rainfall and flood risk analysis. *Journal of Hydraulic Engineering*, **111**, 754–773.

Burn D.H. (1990) Evaluation of regional food frequency analysis with a region of influence approach. *Water Resources Research*, **26**, 2257–2265.

Calver A. (1993) The time-area formulation revisited. *Proceedings of the Institution of Civil Engineers: Water, Maritime and Energy*, **101**, 31–36.

Calver A. (1996) Development and experience of the 'TATE' rainfall-runoff model. *Proceedings of the Institution of Civil Engineer: Water, Maritime and Energy*, **118**, 168–176.

Calver A. (1997) Recharge Response Functions. *Hydrology and Earth System Sciences*, **1**, 47–53.

Calver A., Lamb R. and Morris S.E. (1999) River flood frequency estimation using continuous runoff modelling. *Proceedings of The Institution of Civil Engineers. Water, Maritime and Energy*, **136**, 225–234.

Calver A., Lamb R., Kay A.L. and Crewett J. (2001) *The Continuous Simulation Method for River Flood Frequency Estimation*, *Defra Project FD0404 Final Report*, Centre for Ecology and Hydrology: Wallingford, November 2001.

Cameron D., Beven K. and Tawn J. (2000a) An evaluation of three stochastic rainfall models. *Journal of Hydrology*, **228**, 130–149.

Cameron D., Beven K. and Naden P. (2000b) Flood frequency estimation by continuous simulation under climate change (with uncertainty). *Hydrology and Earth System Sciences*, **4**, 393–405.

Cameron D., Beven K., Tawn J. and Naden P. (2000c) Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation). *Hydrology and Earth System Sciences*, **4**, 23–34.

Cameron D., Beven K. and Tawn J. (2001) Modelling extreme rainfalls using a modified random pulse Bartlertt-Lewis stochastic rainfall model (with uncertainty). *Advances in Water Resources*, **24**, 203–211.

Cameron D., Beven K., Tawn J., Blazkova S. and Naden P. (1999) Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). *Journal of Hydrology*, **219**, 169–187.

Calder I.R., Harding R.J. and Rosier P.T.W. (1983) An objective assessment of soil-moisture deficit models. *Journal of Hydrology*, **60**, 329–355.

Chow V.T., Maidment D.R. and Mays L.W. (1988) *Applied Hydrology*, McGraw Hill: New York, p. 572.

Clark C.O. (1945) Storage and unit hydrograph. *Transactions of the American Society of Civil Engineers*, **110**, 1416–1446.

Clarke R.T. (1994) *Statistical Modelling in Hydrology*, Wiley: New Jersey, p. 412.

Clarke R.T. (2003) Frequencies of future extreme events under conditions of changing hydrologic regime. *Geophysical Research Letters*, **30**(3), 1124, doi:10.1029/2002GL016214.

Connell R.J. and Pearson C.P. (2001) Two-component extreme value distribution applied to Canterbury annual maximum flood peaks. *Journal of Hydrology (New Zealand)*, **40**, 105–127.

Cordova J.R. and Rodriguez-Iturbe I. (1983) Geomorphoclimatic estimation of extreme flow probabilities. *Journal of Hydrology*, **65**, 159–173.

Cowperthwaite P.S.P., O'Connell P.E., Metcalfe A.V. and Mawdsely J.A. (1996) Stochastic point process modelling of rainfall (II). Regionalisation and disaggregation. *Journal of Hydrology*, **175**, 47–65.

Cox D.R. and Isham V. (1988) A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society London*, **A415**, 317–328.

Cox D.R. and Isham V. (1994) Stochastic models of precipitation. In *Statistics for the Environment 2: Water Related Issues*, Barnett V. and Turkman K. (Eds.), Wiley: Chichester, pp. 3–18.

Crawford N.H. and Linsley R.K. (1966) *Digital Simulation in Hydrology: Stanford Watershed Model IV* , *Technical Report 39* , Stanford University, California.

Crooks S.M., Naden P.S., Broadhurst P. and Gannon B. (1996) *Modelling the Flood Response of Large Catchments: Initial Estimates of the Impacts of Climate and Land Use Change*, Institute of Hydrology Report: Wallingford.

Cunderlik J.M. and Burn D.H. (2003) Non-stationary pooled flood frequency analysis. *Journal of Hydrology*, **276**, 210–223.

Daluz Viera J. (1983) Conditions governing the use of approximations of the St. Venant equations for shallow subsurface water flow. *Journal of Hydrology*, **60**, 43–58.

Davison A.C. and Smith R.L. (1990) Model for exceedances over high thresholds. *Journal the Royal Statistical Society Series B*, **52**, 393–442.

Dooge J.C.I. (2003) *Linear Theory of Hydrologic Systems*, EGU Reprint Series No. 1, European Geosciences Union: Katlenburg-Lindau, p. 323.

Eagleson P.S. (1972) Dynamics of Flood Frequency. *Water Resources Research*, **8**, 878–898.

Faulkner D.S. and Jones D.A. (1999) The FORGEX method of rainfall growth estimation. III: Examples and confidence intervals. *Hydrology and Earth System Sciences*, **3**, 205–212.

Faulkner D. and Wass P. (2005) Flood estimation by continuous simulation in the Don catchment, South Yorkshire, UK. *Journal of the Chartered Institution of Water and Environmental Management*, **19**(2), 78–84.

Fernandez W., Vogel R.M. and Sankarasubramanian A. (2000) Regional calibration of a watershed model. *Hydrological Sciences Journal*, **45**, 689–708.

Fowler A. (2002) Assessment of the validity of using mean potential evaporation in computations of the long term soil water balance. *Journal of Hydrology*, **256**, 248–263.

Freeze R.A. and Harlan R.L. (1969) Blue-print for a physically-based digitally simulated hydrologic response model. *Journal of Hydrology*, **9**, 237–258.

Grayson R.B., Moore I.D. and McMahon T.A. (1992) Physically-based hydrologic modelling. 2. Is the concept realistic. *Water Resources Research*, **28**, 2659.

Gupta H.V., Bastidas L.A., Vrugt J. and Sorooshian S. (2002) Multiple criteria global optimization for watershed model calibrations. In *Advances in the Calibration of Watershed Models*, Duan Q., Gupta H.V., Sorooshian S., Rousseau A. and Turcotte R. (Eds.), Monograph Series on Water Resources, American Geophysical Union.

Hashemi A.M., Franchini M. and O'Connell P.E. (2000) Climatic and basin factors affecting the flood frequency curve: part I – a simple sensitivity analysis based on the continuous

simulation approach. *Hydrology and Earth System Sciences*, **4**, 463–482.

Hebson C. and Wood E.F. (1982) A derived flood frequency distribution. *Water Resources Research*, **18**, 1509–1518.

Hosking J.R.M. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal Royal Statistical Society Series B*, **52**, 105–124.

Hosking J.R.M. and Clarke R.T. (1990) Rainfall-runoff relations derived from the probability theory of storage. *Water Resources Research*, **26**, 1455–1463.

Institution of Engineers (1987) *Australian Rainfall and Runoff: A Guide to Flood Estimation*, Editor-in-chief Pilgrim D.H. (Ed.), Revised Edition 1987 (Reprinted edition 1998), ACT: Barton.

Institute of Hydrology (1996) *A Guide to The PDM (Version 1.0)*, Institute of Hydrology: Wallingford.

Institute of Hydrology (1999) *Flood Estimation Handbook (In Five Volumes)*, Institute of Hydrology: Wallingford.

James L.D. (1965) Using a digital computer to estimate effects of urban development on flood peaks. *Water Resources Research*, **1**, 223–234.

Jakeman A.J., Littlewood I.G. and Whitehead P.G. (1990) Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology*, **117**, 275–300.

Kirkby M.J. (1976) Tests of the random network model, and its implications for basin hydrology. *Earth Surface Processes*, **1**, 197–212.

Kirkby M.J. (1997) Topmodel: a personal view. *Hydrological Processes*, **11**, 1087–1097.

Kilsby C.G., O'Connell P.E. and Fallows C.S. (1998) Producing rainfall scenarios for hydrological impact modelling. In *Hydrology In A Changing Environment*, Wheater H.S. and Kirby C. (Eds.), Wiley: Chichester, pp. 33–42.

Kuczera G. (1983) Improved parameter inference in catchment models 1: evaluating parameter uncertainty. *Water Resources Research*, **19**, 1151–1162.

Kundzewicz Z.W. (2004) Editorial – searching for change in hydrological data. *Hydrological Sciences Journal*, **49**, 3–6.

Lamb R. (1999) Calibration of a conceptual rainfall-runoff model for flood frequency estimation by continuous simulation. *Water Resources Research*, **35**, 3103–3114.

Lamb R. and Kay A.L. (2004) Confidence intervals for a spatially generalised, continuous simulation flood frequency model for Great Britain. *Water Resources Research*, **40**, W07501, doi:10.1029/2003WR002428.

Lamb R., Crewett J. and Calver A. (2000) Relating hydrological model parameters and catchment properties to estimate flood frequencies from simulated river flows. *Proceedings of the BHS 7th National Hydrology Symposium, Newcastle, September 2000*, Institute of Hydrology: Wallingford.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A Simple hydrologically based model of land surface water and energy fluxes for GSMs. *Journal of Geophysical Research*, **99**(D7), 14,415–14,428.

Littlewood I.G. and Jakeman A.J. (1994) A new method of rainfall-runoff modelling and its applications in catchment hydrology. In *Environmental Modelling*, Zannetti P. (Ed.), Computational Mechanics Publications: Southampton, pp. 143–171.

Lovejoy S. and Schertzer D. (1985) Generalized scale invariance in the atmosphere and fractal models of rain. *Water Resources Research*, **21**, 1233–1250.

Mellor D. (1996) The modified turning bands (MTB) model for space-time rainfall: I Model definition and properties. *Journal of Hydrology*, **175**, 113–117.

Mellor D., Sheffield J., O'Connell P.E. and Metcalfe A.V. (2000a) A stochastic space-time rainfall forecasting system for real time flow forecasting I: development of MTB conditional rainfall scenario generator. *Hydrology and Earth System Sciences*, **4**, 603–616.

Mellor D., Sheffield J., O'Connell P.E. and Metcalfe A.V. (2000b) A stochastic space-time rainfall forecasting system for real time flow forecasting II: application of SHETRAN and ARNO rainfall runoff models to the Brue catchment. *Hydrology and Earth System Sciences*, **4**, 617–626.

Michaud J. and Sorooshian S. (1994) Effects of rainfall sampling errors on simulation of desert flash floods. *Water Resources Research*, **30**, 2765–2775.

Moore R.J. (1985) The probability-distributed principle and runoff production at point and basin scales. *Hydrological Sciences Journal*, **30**, 273–297.

Moore R.J. and Clarke R.T. (1981) A distribution-function approach to rainfall-runoff modelling. *Water Resources Research*, **17**, 1367–1382.

Naden P.S. (1992a) Analysis and use of peaks-over-threshold data in flood estimation. In *Floods and Flood Management*, Saul A.J. (Ed.), Kluwer Academic: Dordrecht, pp. 131–143.

Naden P.S. (1992b) Spatial variability in flood estimation for large catchments: the exploitation of channel network structure. *Hydrological Sciences Journal*, **37**, 53–71.

Natural Environment Research Council (1975) *Flood Studies Report (In Five Volumes)*, Natural Environment Research Council: London.

Nash J.E. and Sutcliffe J.V. (1970) River flow forecasting through conceptual models, 1, a discussion of principles. *Journal of Hydrology*, **10**, 282–290.

Ngirane-Katashaya G. and Wheater H.S. (1985) Hydrograph sensitivity analysis to storm kinematics. *Water Resources Research*, **21**, 337–345.

Northrop P. (1997) *A clustered spatial-temporal model of rainfall*, Research Report No. 177, Department of Statistical Science, University College London.

Office of The Deputy Prime Minister (2001) *Planning Policy Guidance 25: Development and flood risk*, London. `http://www.odpm.gov.uk/`.

Onof C. and Wheater H.S. (1994) Improvement to the modelling of British rainfall using a modified random parameter Bartlett-Lewis rectangular pulses model. *Journal of Hydrology*, **157**, 177–195.

Post D.A., Jones J.A. and Grant G.E. (1998) An improved methodology for predicting the daily hydrologic response of ungauged catchments. *Environmental Modelling and Software*, **13**, 395–403.

Puente C.E. (1996) A new approach to hydrologic modeling: derived distributions revisited. *Journal of Hydrology*, **187**, 65–80.

Rahman A., Weinmann P.E., Hoang T.M.T. and Laurenson E.M. (2002) Monte Carlo simulation of flood frequency curves from rainfall. *Journal of Hydrology*, **256**, 196–210.

Reynard N.S., Prudhomme C. and Crooks S.M. (2001) The flood characteristics of large UK rivers: potential effects of changing climate and land use. *Climatic Change*, **48**, 343–359.

Reimann J. (1989) *Mathematical Statistics with Application in Flood Hydrology*, Akade'miai Kiado': Budapest, p. 330.

Robinson J.S. and Sivapalan M. (1997) Temporal scales and hydrological regimes: implications for flood frequency scaling. *Water Resources Research*, **33**, 2981–2999.

Robson A. and Kundzewicz Z.W. (Ed.) (2000) *Detecting Trend and Other Changes in Hydrological Data*, World Meteorological Organization, Geneva.

Rodriguez-Iturbe I. (1993) The geomorphological unit hydrograph. In *Channel Network Hydrology*, Beven K. and Kirby M.J. (Eds.), John Wiley & Sons: Chichester.

Rodriguez-Iturbe I., Gonzalez-Sanabria M. and Bras R.L. (1982) A geomorphoclimatic theory of the instantaneous unit hydrograph. *Water Resources Research*, **18**, 877–886.

Rodriguez-Iturbe I. and Valdes J. (1979) The geomorphologic structure of the hydrologic response. *Water Resources Research*, **15**, 1409–1420.

Romanowicz R., Beven K.J. and Tawn J.A. (1994) Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach. In *Statistics for the Environment 2: Water Related Issues*, Barnett V. and Feridum-Turkman K. (Eds.), Wiley: New York, pp. 297–317.

Rossi F., Fiorentino M. and Versace P. (1984) Two component extreme value distribution for flood frequency analysis. *Water Resources Research*, **20**, 847–856.

Sefton C.M. and Howarth S.M. (1998) Relationships between dynamic response characteristics and physical catchment descriptors of catchments in England and Wales. *Journal of Hydrology*, **211**, 1–16.

Sherman L.K. (1932) Streamflow from rainfall by unit-graph method. *Engineering News Record*, **108**, 501–505.

Singh V.P. (2002) Is hydrology kinematic? *Hydrological Processes*, **16**, 667–716.

Sivapalan M., Wood E.F. and Beven K.J. (1990) On hydrologic similarity 3: a dimensionless flood frequency model using a generalized geomorphological unit hydrograph and partial area runoff generation. *Water Resources Research*, **26**, 43–58.

Sorooshian S. and Dracup J.A. (1980) Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: correlated and heteoscedastic error cases. *Water Resources Research*, **16**, 430–442.

Todini E. (1996) The ARNO rainfall-runoff model. *Journal of Hydrology*, **175**, 339–382.

USDA Soil Conservation Service (1985) *Estimation of Direct Runoff from Storm Rainfall*, Section 4, National Engineering Handbook: Hydrology.

Vrugt J.A., Gupta H.V., Bastidas L.A., Bouten W. and Sorooshian S. (2003) Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resource Research*, **39**, 1214.

Wang Q.J. (1991) The POT model described by the generalized Pareto distribution with Poisson arrival rate. *Journal of Hydrology*, **129**, 263–280.

Watts L.G. and Calver A. (1991) Effects of spatially-distributed rainfall on runoff for a conceptual catchment. *Nordic Hydrology*, **22**, 2001–2014.

Waylen P.R. and Woo M. (1982) Prediction of annual floods generated by mixed processes. *Water Resources Research*, **18**, 1283–1286.

Wheater H.S. (2002) Progress in and prospects for fluvial flood modelling. *Philosophical Transactions of the Royal Society, Series A.*, **360**, 1409–1431.

Wheater H.S., Isham V.S., Cox D.R., Chandler R.E., Kakou A., Northrop P.J., Oh L., Onof C. and Rodriguez-Iturbe I. (2000) Spatial-temporal rainfall fields: modelling and statistical aspects. *Hydrology and Earth System Sciences*, **4**, 581–601.

Wood E.F. and Hebson C. (1986) On hydrologic similarity 1: derivation of the dimensionless flood frequency curve. *Water Resources Research*, **22**, 1549–1554.

Zhao R.-J. (1992) The Xinanjiang model applied in China. *Journal of Hydrology*, **135**, 371–381.

126: Modeling Recession Curves and Low Streamflows

CHARLES KROLL

Environmental Resources and Forest Engineering, SUNY College of Environmental Science and Forestry (ESF), Syracuse, NY, US

Low streamflow events are critical periods for water quantity and water quality management purposes. During most low streamflow events, the stream is comprised solely of groundwater sources. This section focuses on both theoretical and empirical methods for modeling groundwater discharge to a stream and the resulting streamflow recession. A review is provided of solutions to the Boussinesq equation for hillslope discharge, and methods for determining when specific solutions may be most appropriate. In addition, how to estimate baseflow recession constants that aid in the modeling of groundwater discharge to streams, has been discussed. This involves the presentation of numerous methods to parameterize baseflow recession characteristics, how best to represent error in these models, and experiments that have been performed to determine the most appropriate methodology.

INTRODUCTION

Importance of Low Streamflows

Understanding the frequency and duration of low streamflow events is critical to the efficient management of water resources throughout the world. This is especially important for water quality management, where low streamflows provide the necessary dilution of nonpoint source and point source pollution discharges, and water quantity management, where low streamflows greatly influence water use policy. Low streamflow predictions and forecasts are used to plan water supply, hydropower, and irrigation systems, design cooling-plant facilities, site treatment plants and sanitary landfills, determine waste-load allocations, and make decisions regarding interbasin transfers of water and allowable basin withdrawals. In addition, low streamflow events are often critical periods for aquatic habitats owing to potentially low dissolved oxygen concentrations and/or high concentration of pollutants. Low streamflows are also used in health risk assessments that quantify the long-term risk associated with exposure to carcinogenic substances in rivers.

Low streamflow and associated droughts can have a devastating impact on a society. Often a major impact of droughts is a reduction in agricultural production, which can cause large monetary impacts as well as widespread malnutrition, displacement of peoples, and loss of life. For some statistics regarding the worldwide impacts droughts (and other natural disasters), the reader is encouraged to access the Emergency Events Database (EM-DAT), which was developed by the World Health Organization (WHO) and the Belgium Government (EM-DAT, 2003). It is important to note that many drought "events" are not reported or quantified since they often occur over many years and the societal impacts are more difficult to assess than other natural disasters.

Difficulties with Lowflow Analyses

While flood frequency analyses have received considerable attention in the research literature, the estimation of low streamflow statistics has received relatively little. In some regions, there is evidence that droughts may be more influenced by climatic fluctuation and anthropogenic effects than floods. For instance, by examining historical records for the last 50 years, Douglas *et al.* (2000) found significant trends in lowflow series in some regions of the United States, while they found no such trends for flood flow series in any regions of the United States. Lowflow analyses are often complicated owing to precision problems regarding

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

recorded data. Analyses of historic streamflow records can involve the presence of intermittent streamflows, resulting in a truncation or censoring of the lower tail of the frequency distribution (which complicates many log-based analyses). Often low streamflow estimates are based on extrapolating rating curves (stage-discharge relationships), thus introducing additional error.

The physical processes responsible for lowflow events are also complex. Lowflows are primarily due to groundwater discharge to a stream. Despite advancements in measurement technology, "our knowledge of what goes on underground is still very limited" (Beven, 2001). The nonlinear partial differential equations describing these processes have only been solved analytically for idealized systems with specific boundary conditions. These solutions do not address the common heterogeneity of hydrogeologic properties observed within most watersheds, nor the complex physical structure of watersheds and aquifers. In addition, the spatial and temporal variability of infiltration and percolation processes complicates our understanding of groundwater recharge characteristics. Despite these complexities, we are often forced to model groundwater discharge processes to aid in the management of low streamflow events.

Rainfall-runoff Models: Baseflow Recession

There are many analyses related to lowflows and streamflow recessions (Smakhtin, 2001). Other sections of this encyclopedia cover many of these topics, such as frequency analyses and the estimation of low streamflow statistics at gauged and ungauged river sites. In this section, several topics related to understanding and parameterizing streamflow recession characteristics for rainfall-runoff analyses are discussed. There are a variety of ways to represent groundwater discharge, which is commonly referred to as baseflow, to streamflow. Baseflow can be defined in a number of ways, including just groundwater contributions, or more commonly groundwater and other delayed sources (Hall, 1968).

This section of is broken into two subsections: theoretical models of groundwater discharge, and parameterizing baseflow recession characteristics. The first subsection ("Theoretical Models of Groundwater Discharge") will discuss solutions to the Boussinesq equation, and variations of the Brutsaert and Nieber (1977) technique for analyzing the appropriateness of the Boussinesq solutions. This subsection presents theoretical relationships between groundwater discharge parameters and watershed characteristics, and ways in which the parameters of these models may be estimated and/or regionalized. The second subsection (Parameterizing Baseflow Recession Characteristics) will focus on estimating the parameters of some common groundwater discharge models employed in rainfall-runoff modeling. This will include discussions of the impact of error on recession constant estimators, as well as attempts to compare these estimators. Different rainfall-runoff models model baseflow in varying manners. A number of these techniques are discussed below. Tallaksen (1995) provides a review of additional baseflow recession analyses.

Some event-based rainfall-runoff models ignore baseflow contributions or consider them constant. In these models, it is the flood peak that is typically of interest (which can be orders of magnitude larger than the baseflow contributions), or stormflow contributions (e.g. to design detention ponds that delay and attenuate excess storm runoff). A common approach is to consider groundwater contributions to streamflow as a single linear reservoir, such as in HEC-1 (Feldman, 1995) (note that HEC-1 is now called HEC-HMS) and GWLF (Haith *et al.*, 1992). A linear reservoir means that groundwater discharge is modeled as a linear function of storage, resulting in an exponential decay of discharge with time. Another common technique is to assume two or more linear reservoirs to account for slower and quicker groundwater contributions. SAC-SMA (Burnash and Ferral, 2002), UBC (Quick, 1995), Tank (Sugawara, 1995), and HBV (Bergstrom, 1995) models employ this type of groundwater discharge model. Groundwater discharge in PRMS (Leavesley and Stannard, 1995) is modeled in a similar manner, but shallow subsurface flow is simulated as a 2nd order quadratic function.

Using the TOPMODEL assumptions that downslope transmissivity varies with depth of saturation as an exponential function of storage deficit, and that the water table remains parallel to the surface, a baseflow recession follows a first-order hyperbolic relationship (Beven *et al.*, 1995), though this was later generalized to other transmissivity profiles (Iorgulescu and Musy, 1997; Lamb *et al.*, 1997). In many spatially distributed models, groundwater is routed cell-by-cell and a distributed groundwater response is modeled. This distributed response can be based on Darcy's law and surface gradients, such as in DHVSM, (Wigmosta *et al.*, 1994); a nonlinear Boussinesq response, such as in SHE (Bathurst *et al.*, 1995); or approximated using a kinematic assumption, such as in THALES (Grayson *et al.*, 1995). Franchini and Pacciani (1991) and Singh (1995) review the components (including groundwater discharge) of numerous rainfall-runoff models.

THEORETICAL MODELS OF GROUNDWATER DISCHARGE

One approach to modeling groundwater discharge is by using a groundwater flow model. Distributed models, such as the finite difference model MODFLOW or the finite element model FEMWATER, are typically employed to examine small-scale groundwater transport problems. These

three-dimensional models require a large number of parameters and are usually not employed across an entire watershed, especially in the application of rainfall-runoff models. As mentioned previously, in most rainfall-runoff models groundwater discharge is modeled as a more simplified, lumped process. In this section, hillslope models based on the Boussinesq equation are discussed.

Solutions to the Boussinesq Equation

Partially saturated hillslope subsurface drainage can be described by the Richards equation (Brutsaert and El-Kadi, 1984). Since the resulting solutions from this method cannot easily be parameterized in practice, a hydraulic approach is typically taken (Brutsaert, 1994). In 1877 Boussinesq derived an expression for one-dimensional flow from an unconfined sloping aquifer (Childs, 1971):

$$q = -kh\left[\left(\frac{\partial h}{\mathrm{d}x}\right)\cos(\theta) + \sin(\theta)\right] \tag{1}$$

where q is the flow in the x-direction per unit width of aquifer, k is the hydraulic conductivity, and h is the thickness of the water layer perpendicular to the underlying impermeable layer with slope θ. This equation is based on the assumption that capillary effects and evapotranspiration are negligible. Combining this expression with the continuity equation, and neglecting any recharge from the unsaturated zone, one obtains:

$$\frac{\partial h}{\partial t} = \frac{k}{f}\left[\cos(\theta)\frac{\partial}{\partial x}\left(h\frac{\partial h}{\partial x}\right) + \sin(\theta)\frac{\partial h}{\partial x}\right] \tag{2}$$

where f is the drainable porosity. Equation (2) is usually referred to as the Boussinesq equation, and is based on assuming k, f, and θ are constant. This equation satisfies the Dupuit–Forchheimer conditions (horizontal flow lines and hydraulic gradient equal to the slope of the water table). Brutsaert (1994) used a linear approximation of equation (2) to derive an expression for drainage of a sloped, fully saturated aquifer. Brutsaert referred to this as the "unit response" (Green's function) of the aquifer.

Mizumura (2002) recently used the method of variations to approximate the solution of equation (2). Verhoest and Troch (2000) applied the approach of Brutsaert (1994) to time-varying recharge rates. This solution employs a quasi steady-state approach, where varying water table profiles between successive steady-state conditions can be computed. Troch *et al.* (2003) expanded the original Boussinesq equation (equation (2)) by reformulating the problem in terms of soil water storage as opposed to water table height. This leads to a new one-dimensional equation, the so-called hillslope-storage Boussinesq (hsB) equation, which accounts for the plan shape (convergent/divergent) and profile curvature (concave/convex) to describe groundwater flow and saturation in complex hillslopes. Paniconi *et al.* (2003) examined how solutions of the one-dimensional hsB equation compare with solutions to the three-dimensional Richard's equation. A more common solution is based on assuming the impermeable layer is horizontal ($\theta = 0$), resulting in:

$$\frac{\partial h}{\partial t} = \frac{k}{f}\left[\frac{\partial}{\partial x}\left(h\frac{\partial h}{\partial x}\right)\right] = \frac{k}{f}\left[h\frac{\partial^2 h}{\partial x^2} + \left(\frac{\partial h}{\partial x}\right)^2\right] \tag{3}$$

Solutions to this equation are based on the idealized aquifer presented in Figure 1, with a fully penetrating stream channel (stream channel bottom coincident with impermeable layer. There are three analytical solutions to this equation that have been applied (Brutsaert and Nieber, 1977). The first is based on the assumption that either the slope of the water table is nearly horizontal ($\partial \mathrm{h}/\partial \mathrm{x} = 0$) or that $\partial^2 \mathrm{h}/\partial \mathrm{x}^2 \gg (\partial \mathrm{h}/\partial \mathrm{x})^2$. Under this situation, equation (3) is linearized, and the resulting groundwater outflow from an initially saturated aquifer per unit length of channel can be written:

$$q = \frac{2kpD(D - D_c)}{B}\sum_{n=1,3,\ldots}^{\infty}\exp\left(-\frac{n^2\pi^2 kpDt}{4fB^2}\right) \tag{4}$$

where D is the depth of the aquifer, pD is the mean depth of the water table, D_c is the depth in the stream, and B is the aquifer breadth. After a long time this equation reduces to

$$q = \frac{2kpD(D - D_c)}{B}\exp\left(\frac{\pi^2 kpDt}{4fB^2}\right) \tag{5}$$

Equation (5) is often referred to as the linear solution of Boussinesq, who solved this equation in 1903 (Brutsaert and Nieber, 1977). Boussinesq also obtained an exact solution to equation (3) by assuming the water level in the channel was zero ($D_c = 0$), and the water table is initially described by an inverse incomplete beta function. The resulting groundwater discharge per unit length of channel

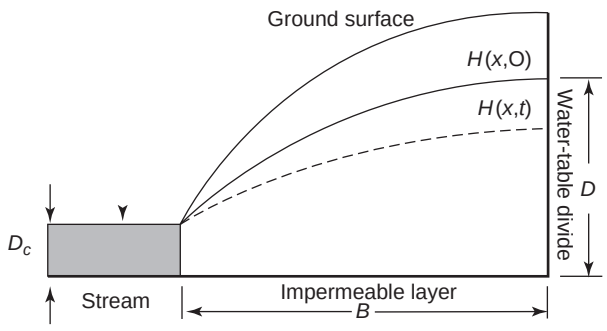


Figure 1 Idealized physical model of a hillslope

for this case is:

$$q = \frac{0.862kD^2}{B\left(1 + \frac{1.115kDt}{fB^2}\right)^2} \quad (6)$$

Equation (6) is not applicable when the aquifer is fully saturated, and thus is only applicable after a long time t. If one assumes the aquifer is infinitely wide ($B = \infty$) and again $D_c = 0$, then a solution for a small t can be found as:

$$q = \frac{1}{2}D_o t^{-1/2} = \frac{0.3321(kf)^{1/2}D^{3/2}}{t^{1/2}} \quad (7)$$

where D_o is the desorptivity (Brutsaert and Nieber, 1977):

$$D_o = 0.6642(kf)^{1/2}D^{3/2} \quad (8)$$

There are therefore three commonly cited solutions to the Boussinesq equation: a linear solution, and two nonlinear solutions (one for long t and one for small t). Of interest is which of these solutions is most appropriate for a given watershed or region of watersheds.

Hogarth *et al.* (1999) developed a new approximate solution to the one-dimensional Boussinesq equation for a semi-infinite aquifer with a time-varying hydraulic head as a boundary condition. A further approximation to the Boussinesq equation arises if it is assumed that the downslope hydraulic gradient stays constant at the slope angle. This then gives rise to a kinematic wave equation, which was used by Beven (1982) to model subsurface stormflow analytically for simple initial conditions and rainstorms and different conductivity profiles.

Power Relationship and the Boussinesq Solutions

Brutsaert and Nieber (1977) suggested a power equation relating outflow from the basin, Q, to the volume of water stored in the aquifer, S:

$$Q = mS^n \quad (9)$$

where m and n are constants. If $n = 1$ in equation (9), then the discharge-storage relationship is linear; such a system is described as a "linear reservoir". Combining this equation with a lumped continuity equation one obtains

$$Q = Q_0 K_b^t \quad (10)$$

for the case when $n = 1$. Q_0 is the outflow at time $t = 0$, and K_b is a constant, commonly referred to as the baseflow recession constant. This solution is consistent with the linearized solution to the Boussinesq equation, where

$$K_b = \exp\left(\frac{-\pi^2 kpD}{4fB^2}\right) \quad (11)$$

Thus the linear solution to the Boussinesq equation is consistent with the modeling baseflow as a linear reservoir, as is done in many circumstances. For the nonlinear solutions ($n \neq 1$), a general solution is:

$$Q = Q_0(1 + ct)^{n/(1-n)} \quad (12)$$

where c is a constant. With $n = 2$, this equation is consistent with equation (6), the nonlinear solution to the Boussinesq equation for a long time t. Wittenberg and Sivapalan (1999) used (12) to identify and quantify the main components of the groundwater balance using observed time series.

To eliminate the time frame (and thus avoid the estimation of Q_0), Brutsaert and Nieber (1977) suggested not examining the hydrograph itself, but instead its slope, $\mathrm{d}Q/\mathrm{d}t$, using a power function relating slope to discharge:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -aQ^b \quad (13)$$

where a and b are constants. These constants are related to equation (12) by:

$$b = \frac{2n-1}{n} \quad (14)$$

and

$$a = \frac{n}{n-1}cQ_0^{(1-n)/n} \quad (15)$$

By assuming some geometric similarity of the catchment drainage patterns, such that the groundwater outflow per unit length of channel is

$$q = \frac{Q}{2L} \quad (16)$$

where L is the total stream length in the catchment, and the aquifer breadth is

$$\mathrm{B} = \frac{\alpha \mathrm{A}}{2L} \quad (17)$$

where α is the fraction of catchment underlain by aquifers and A is the drainage area, the three solutions to the Boussinesq equation can be put in the form of equation (13) (Brutsaert and Nieber, 1977). For the linear solution (equation (5)),

$$b = 1 \quad \text{and} \quad a = \frac{\pi^2 kpDL^2}{f(\alpha A)^2} \quad (18)$$

In this situation, the aquifer is being modeled as a single linear reservoir, and the parameter $a = \ln(K_b)$, where K_b is the baseflow recession constant. For the nonlinear solution for a long period (equation (6)), one obtains

$$b = \frac{3}{2} \quad \text{and} \quad a = \frac{4.804k^{1/2}L}{f(\alpha A)^{3/2}} \quad (19)$$

while for the nonlinear solution for a short time period (equation (7)), one obtains

$$b = 3 \quad \text{and} \quad a = \frac{1.133}{kfD^3L^2} \tag{20}$$

Brutsaert and Lopez (1998) noted α equals 1 in most watersheds, and thus usually can be removed from these expressions. Szilagyi and Parlange (1998) developed a method of baseflow separation using the two nonlinear solutions to the Boussinesq equation.

By examining b, the slope of a log–log plot of $\mathrm{d}Q/\mathrm{d}t$ versus Q, one can potentially assess the appropriateness of various solutions of the Boussinesq equation to describing groundwater discharge to a stream. This approach is sometimes referred to as the Brutsaert–Nieber procedure (Tallaksen, 1995), and is presented in the next section. In addition, the y-intercept, a, is a function of watershed characteristics, thus potentially allowing the parameters of a groundwater discharge model to be regionalized or estimated from watershed characteristics. In addition, the converse may be true, where watershed characteristics (such as regional hydrogeologic properties) may be estimated from the parameters of calibrated groundwater discharge models.

Appropriateness of Boussinesq Solutions

There are a number of difficulties associated with examining the appropriateness of solutions to the Boussinesq equation. One problem is determining when streamflow sources other than baseflow cease and when recharge from the unsaturated zone becomes negligible (i.e. when the streamflow is entirely baseflow and may be described by a Boussinesq solution). Another problem is determining how to fit a line to a log–log plot of $\mathrm{d}Q/\mathrm{d}t$ versus Q to obtain an estimate of the slope, b. This issue is further complicated by the precision with which the streamflow data are recorded. A few of the studies addressing this problem are discussed below.

Brutsaert and Nieber (1977) examined six catchments in the Finger Lakes region of New York. They considered only streamflow data 5 days after rainfall as baseflow, and fit only the "lower envelope" of data in a $\mathrm{d}Q/\mathrm{d}t$ versus Q plot (i.e. the lowest $|\mathrm{d}Q/\mathrm{d}t|$ at a given Q). The rationale for fitting the lower envelope is to remove data points that might be influenced by sources other than groundwater. A value of $b = 3/2$ appeared to have a good "visual" fit at these sites. Using these fits, a value of a in equation (13) was estimated at each site. These six values were regressed then against $\mathrm{LA}^{-3/2}$ to obtain a regional estimate of the quantity $k^{1/2}/(f\alpha^{3/2})$ using equation (18). Brutsaert and Nieber note that one would expect k, f and α to vary from site to site, yet felt a satisfactory relationship existed between a and $\mathrm{LA}^{-3/2}$ for 5 of the 6 sites. One obvious shortcoming of this study was the limited number of sites employed.

Troch *et al.* (1993) applied the Brutsaert and Nieber technique to 10 years of streamflow data at a single, well studied catchment in Belgium. Recessions began two days after the cessation of rainfall, which was approximately 2 times larger than the estimated time of concentration for the watershed. They examined fitting an ordinary least squares (OLS) regression line to all recession points, finding it to correspond well to $b = 3/2$. They also fit the lower envelope of the data with lines of $b = 3/2$ and $b = 3$, and performed a sensitivity analysis of the lower envelop fit by examining solutions excluding 5% and 10% of the data. Troch *et al.* noted that where the lines of $b = 3/2$ and $b = 3$ cross represents an estimator of the discharge threshold under which the long time solution to the Boussinesq equation is appropriate. Assuming a range of values for drainable porosity, they developed a catchment-scale estimator of effective depth to the water table (representing the initial storage in the aquifer) and hydraulic conductivity.

Brutsaert and Lopez (1998) examined recessions at 22 USDA subbasins in Oklahoma. They proposed fitting the lower envelop with a line they referred to as the line of organic correlation. Others have called this line the unique solution and the reduced major axis (Hirsch, 1982). Brutsaert and Lopez proposed this fitting technique owing to the error in both the independent and dependent variables in the regression. They used streamflows 2 days after the cessation of rainfall. In contrast to the other two studies mentioned above, they found in this region the long time response (corresponding to the smallest values of Q), was described by the linear Boussinesq equation ($b = 1$), while for shorter time periods (larger values of Q) $b = 3$. They thus had 2 estimates of a for each catchment (from the two lines of fit). Using these equations Brutsaert and Lopez obtained effective basin-scale estimators of various groundwater parameters, which allowed estimation of watershed hydraulic conductivity and drainable porosity. A watershed-scale parameter represents the integration of groundwater processes across the entire watershed, characterizing an equivalent homogeneous response. Eng and Brutsaert (1999) found similar results at 11 catchments in Kansas that were part of the same geomorphic region as the sites used by Brutsaert and Lopez; this result indicates the potential for regionalizing groundwater discharge characteristics.

One interesting observation in Eng and Brutsaert was a "stair-step" pattern in a log–log plot of $\mathrm{d}Q/\mathrm{d}t$ versus Q, which was also apparent in the studies by Vogel and Kroll (1992) and Kroll (1989). Below certain values of Q, there was a threshold below which no values of $\mathrm{d}Q/\mathrm{d}t$ occurred. This was because of the precision with which the US Geological Survey data employed is reported. For instance,

discharge values between 10 and 10 000 cfs are reported as integers, while values between 1 and 10 cfs are reported with 1 decimal. Thus it was impossible to have values of dQ/dt calculated from the data less than 1 if the discharges were greater than 10 cfs, but it is possible to obtain such a value at lower discharges. This phenomenon creates data voids in the dQ/dt versus Q plots, further complicating the analyses. There is also potential for uncertainty and nonstationarity in the rating curve at a gauging station at low flows. One should be keenly aware of the precision of the reported data when performing such an analysis.

Vogel and Kroll (1992) performed a similar analysis at 23 catchments in Massachusetts. They did not examine rainfall data to determine baseflow recessions, but instead examined removing a certain percentage of the recession to remove impacts of direct runoff. Vogel and Kroll found that when 30% of the hydrograph was removed, a log–log plot of dQ/dt versus Q produced a slope of $b = 1$ across the region. Vogel and Kroll fit all of the data and not just the lower envelope, since fitting only the lower envelope of the data neglects the presence of measurement error in the data. They developed a catchment lowflow response using the linear solution to the Boussinesq equation, and showed the parameters of such an equation are significant in a regional regression model for low streamflow statistics.

One common theme across these studies was the goal of developing relationships between the parameters of theoretical groundwater discharge equations and watershed characteristics. Unfortunately, these studies represent the beginning of such efforts. Assumptions regarding the homogeneity of hydrogeologic characteristics (such as the assumption of a watershed hydraulic conductivity required by the Boussinesq solutions) and the simplified physical structure of an aquifer used in these derivations are known to be incorrect in most watersheds. Also, these investigations have been performed in only a few regions, typically with only a small number of watersheds. It is difficult to obtain far-reaching conclusions based on these limited studies, but they do provide a framework by which one may continue further investigations of regional and at-site groundwater discharge response.

PARAMETERIZING BASEFLOW RECESSION CHARACTERISTICS

Lowflows in Model Calibration

One concern is how to estimate the baseflow recession parameters of rainfall-runoff models. A potential way to address this is to estimate these parameters when calibrating the model. Often estimating these baseflow parameters is not a priority, even for continuous simulation models. Most loss functions that are minimized in the calibration of rainfall-runoff models (such as minimizing bias or root mean square error of streamflow estimators) are developed to favor flood events, which are often of primary concern. In these situations, the model improves the fit of larger streamflow events at the expense of fitting small flows poorly. In addition, measured low streamflows, which are considered as "truth" in the loss function, often have large amounts of error (as do large streamflow events). To improve the fit of low streamflow events, one can employ either a log-spaced or relative loss function (such as relative bias, which is a percent); in these situations, though, the model fit to larger streamflow events is typically reduced. One should always consider the goals of the model when developing a calibration methodology.

Linear Reservoir Models

Model Form and Error Terms

Another common procedure is to estimate the baseflow recession parameters using available streamflow data. Barnes (1939) assumed that hydrograph recessions follow the basic relations

$$B_t = K_b B_{t-1} \tag{21}$$

$$O_t = K_o O_{t-1} \tag{22}$$

where B_t and O_t are the baseflow and other component of streamflow, and K_b and K_o are the baseflow and other component recession constants, respectively. Thus the streamflow during these recession periods is a combination of these two components

$$Q_t = B_t + O_t \tag{23}$$

Equation (21) corresponds to the linear solution to the Boussinesq Equation presented in equation (5), and is the "linear reservoir" (exponential decay) model employed in many rainfall-runoff models. Barnes (1939) was interested in hydrograph separation procedures, and estimated the recession constants graphically, though his methods contained many subjective judgments (Kulandaiswamy and Seetharaman, 1969; Anderson and Burt, 1980). Hall (1968) provides a review of many early recession analyses that were based on improved graphical procedures. Tallaksen (1995) presents many alternative forms of equation (21).

Recession constants can be estimated from an individual hydrograph or ensembles of hydrographs. Interpreting individual streamflow hydrographs is difficult owing to the poor quality of the data, the small number of data points, and the high variability in recession behavior both within and between catchments (Hall, 1968). This may lead one to favor ensembles of streamflow recessions. Perzyna (1990, as cited in Tallaksen, 1995) suggested a minimum of 10 years of streamflow data was necessary to provide reliable estimates of recession parameters. Regardless, one

should consider these estimates as random variables, and thus attempt to address their statistical properties (Vogel and Kroll, 1996).

If one was to wait a sufficiently long period of time, then the streamflow contribution of other sources, O_t, will approach zero in equation (23). Thus streamflow is comprised of only baseflow and might then be modeled as a single reservoir, the drainage of which might be described by a simple parametric function. In addition, the assumptions made regarding the form of the error in equations (21) and (22) impact the derived estimators. Some methods to estimate recession constants for single or double linear reservoir models are discussed below.

Consider equation (23) as a single linear model ($O_t = 0$) with an additive error term:

$$Q_t = K_b B_{t-1} + \varepsilon_t = K_b Q_{t-1} + \varepsilon_t \tag{24}$$

where ε_t are assumed independent error terms with mean zero and constant variance, and represent both time sampling and model errors. Equation (24) describes a first-order autoregressive process (AR(1)) (Box and Jenkins, 1976), where K_b is the autoregressive parameter. James and Thompson (1970) employed this model. If instead one assumes an additive error in log-space, equation (23) becomes:

$$Q_t = K_b B_{t-1} \exp(\varepsilon_t) = K_b Q_{t-1} \exp(\varepsilon_t) \tag{25}$$

Taking the logarithms of both sides of equation (25) results in an integrated moving average model (ARIMA(0,1,0)) with a constant drift parameter of $\ln(K_b)$ (Box and Jenkins, 1976).

Now consider streamflow recession to comprise both baseflow and other components (as in equation (23)), with an additive error term in real-space:

$$Q_t = K_b B_{t-1} + K_o O_{t-1} + \varepsilon_t \tag{26}$$

Vogel and Kroll (1996) show how this model can be rewritten as an AR(2) process (Box and Jenkins, 1976):

$$Q_t = \phi_1 Q_{t-1} + \phi_2 Q_{t-2} + \varepsilon_t \tag{27}$$

where:

$$K_b = \frac{1}{2}\left[\phi_1 + (\phi_1^2 + 4\phi_2)^{1/2}\right] \tag{28}$$

$$K_o = \frac{1}{2}\left[\phi_1 - (\phi_1^2 + 4\phi_2)^{1/2}\right] \tag{29}$$

Vogel and Kroll discuss the parameter conditions that must be satisfied for this process to be stationary. Spolia and Chander (1974, 1979) showed that if a watershed is represented as a cascade of n equal or unequal linear reservoirs, the streamflow process is equivalent to an ARMA(n, $n-2$) model; here $n = 2$. Another potential AR(2) model is to assume an additive error in log-space:

$$Q_t = (\phi_1 Q_{t-1} + \phi_2 Q_{t-2}) \exp(\varepsilon_t) \tag{30}$$

James and Thompson (1970) modeled recessions as an ARMA(1,1) and ARMA(2,2) process, with the autoregressive and moving average parameters equal; these models both contain parameter redundancy and the derived process reduces to white noise, that is, an uncorrelated random process with zero mean and finite variance (Box and Jenkins, 1976).

Least Squares Estimators

One possible estimator of the recession constants is to obtain an estimate for each individual recession, and then take the average (or a weighted average) of the ensemble to obtain a single estimator for the site. Vogel and Kroll (1996) note how the time series estimators of the parameters are downwardly biased when sample size is small, as is the case for an individual recession. They suggested instead an ensemble of data from all recessions be employed to obtain a single estimator for a site using least squares regression procedures, as was employed by James and Thompson (1970). For the AR(1) model of equation (24), the least squares estimator of K_b is:

$$K_b = \frac{\sum_{t=1}^{n} Q_t Q_{t-1}}{\sum_{t=1}^{n} Q_t^2} \tag{31}$$

where n is the number of pairs of consecutive streamflow recessions. For the ARIMA(0,1,0) model, the least squares estimator of K_b is:

$$K_b = \exp\left[\frac{1}{n}\sum_{t=1}^{n}(\ln(Q_t) - \ln(Q_{t-1}))\right] \tag{32}$$

For the AR(2) model with an additive error in, one could estimate ϕ_1 and ϕ_2 as ordinary least squares (OLS) estimators. Vogel and Kroll (1996) found the error terms to be heteroscedastic (nonconstant variance), and thus the OLS estimators would not be the best linear unbiased estimators (BLUEs). By transforming the model to:

$$\frac{Q_t}{Q_{t-1}} = \phi_1 + \phi_2 \frac{Q_{t-2}}{Q_{t-1}} + \eta_t \tag{33}$$

where the transformed error terms η_t are approximately homoscedastic. ϕ_1 and ϕ_2 can now be estimated from a simple linear regression model, and K_b and K_o then

estimated by equations (28) and (29), respectively. For the AR(2) model with additive error terms in log-space, any standard two-dimensional search algorithm may be employed to estimate the parameters. Kroll (1989) used an iterative least squares approach to estimate this model's parameters.

Recall that then when $b = 1$ in equation (13), the parameter $a = \ln(K_b)$. Using this relationship with an error term additive in log-space Vogel and Kroll (1996) derived another least squares estimator of K_b:

$$K_b = \exp\left\{-\exp\left[\frac{1}{n}\sum_{t=1}^{n}\left\{\ln(Q_{t-1} - Q_t) - \ln\left(\frac{1}{2}(Q_{t-1} + Q_t)\right)\right\}\right]\right\} \quad (34)$$

which they referred to as the linear reservoir estimator.

Comparison of Least Squares Estimators

It is often difficult to determine conclusively the best estimator of a statistic among several competing estimators. One method to do this is a Monte Carlo simulation, which requires definition of an underlying model. Since groundwater processes are complex, such a model is difficult to construct, and most models would resemble one of the models presented above, thus most likely favoring that estimator. The many studies that have investigated baseflow recession modeling (see examples in Tallaksen, 1995), have typically examined only a limited numbers of recessions and/or sites, and thus have been unable to develop conclusive recommendations for baseflow recession estimation methodology.

Vogel and Kroll (1996) attempted to assess a baseflow recession constant estimator's performance based on its ability to describe low streamflow statistics in regional regression models. The watershed model employed was developed by Vogel and Kroll (1992) based on a linear solution to the Boussinesq equation. In this model, lowflow statistics are related to three watershed characteristics: drainage area, basin slope, and the baseflow recession constant. One would expect a better baseflow recession constant estimator to produce a better low streamflow regional regression model.

Vogel and Kroll (1996) performed this experiment with 23 watershed from Massachusetts for the 7-day, 10-year ($Q_{7,10}$) and 7-day, 2-year ($Q_{7,2}$) low streamflows. All five of the baseflow recession constant estimators proposed in the section "Least Squares Estimators" above were analyzed as well as a sixth estimator, which was the summation of baseflow recession constants estimated from individual recessions with an additive error term in log-space. One major conclusion from this analysis was that for the baseflow recession constant estimators based on the AR(1) and AR(2) models with error terms additive in real-space, model performance was worse than using the other four estimators, all of which assumed additive error terms in log-space. The AR(1) model was shown to have heteroscedastic error terms, a violation of the model assumptions. While the performance of the four log-space error term estimators was difficult to distinguish, based on the standard error of estimators across both models, the estimator from an AR(2) with an additive error in log-space (equation (30)) was best. The linear reservoir model estimator (equation (34)) was second best, and slightly worse than the AR(2) model with log-space error. The estimator based on average estimates from individual recessions performed nearly as well, and had the benefit of allowing one to easily assess the variance of the estimator. The AR(2) model is more complicated to implement in practice, and produced only slight parameter improvement; in practice using one of the other three log-space error term estimators is recommended, with the linear reservoir model (equation (34)) preferred. Recently Kroll *et al.* (2004) examined the use of baseflow recession constants based on equation (34) for estimating low streamflow statistics across the United States. They found that by using 3-day moving averages instead of daily averages in this model, lowflow regional regression models were improved in all regions. The use of 3-day moving averages reduces some of the precision problems of recorded low streamflows discussed previously, thus producing a better baseflow recession constant estimator.

Alternative Models and Indices

The linear reservoir models presented in section "Linear Reservoir Models" are not applicable in all circumstances. Often, as has been mentioned previously, the groundwater discharge from a watershed does not exhibit tendencies consistent with the linear solution to the Boussinesq equation. A number of alternative theoretically motivated models of groundwater discharge exist. In addition, researchers have suggested a variety of empirical relationships to model groundwater discharge. A number of these models are discussed below.

Based on the storage-discharge relationship given in equation (9), if the $n \neq 1$ then the general baseflow discharge equation given in equation (12) is:

$$Q = Q_0(1 + ct)^{n/(1-n)} \quad (35)$$

Brutsaert and Nieber (1977) note that Maillet first employed this equation to model low flows on the Vanne River in 1905. Wittenberg and Sivapalan (1999) used an iterative least squares procedure to determine c and n from observed recession data. Hornberger *et al.* (1970) used this

equation with $n = 2$ to model outflow from an unconfined aquifer. A value of $n = 2$ is consistent with the nonlinear solution to the Boussinesq equation for a long time t.

Beven *et al.* (1995) noted that if a linear relationship between storage deficit and depth to the water table holds with a nonlinear (exponential) relationship between transmissivity and storage deficit, as is assumed using TOPMODEL concepts, then groundwater discharge has a first-order hyperbolic relationship to time:

$$\frac{1}{Q_t} = \frac{1}{Q_o} + \frac{t}{m} \tag{36}$$

Using a plot of $1/Q_t$ against time, a linear relationship should exist. The estimated slope of a simple linear fit would be an estimator of 1/m (Beven *et al.*, 1995). Similar results for other transmissivity functions are given by Iorgulescu and Musy (1997) and Duan and Miller (1997).

Werner and Sundquist (1951) modeled the outflow from a confined aquifer as:

$$Q_t = Q_o \sum_{i=1}^{n} b_i \exp(-a_i t) \tag{37}$$

where a_i and b_i are constants. They suggested only the first term was needed, resulting in the linear reservoir model discussed previously. A model similar to equation (37) was employed by Nutbrown (1975), who applied a normal-mode analysis to groundwater discharge to a partially penetrating stream. Barnes (1939) suggested modeling the entire streamflow recession with $n = 3$.

There have also been a number of empirical relationships proposed to model groundwater discharge. Tallaksen (1995) provides a review of many of these models. Tallaksen notes that one-parameter models are preferable when comparing catchments and their responses; as more parameters are added, the fit to observed data improves, but the uncertainty of predictions from the model may increase and parameter sets may not be unique (Beven, 1989). One common empirical model is:

$$Q_t = (Q_o - b) \exp\left(\frac{-t}{C}\right) + b \tag{38}$$

where b and C are constants. With this model, Q_t asymptotically approaches b for large t. Toebes and Strang (1964) suggested this model for areas with snow and ice present. Otnes (1978, as cited in Tallaksen, 1995), found that Norwegian catchments, which have a high lake percentage, are modeled well by the empirical expressions

$$Q_t = at^{-r} \tag{39}$$

or

$$Q_t = at^{-r} + b \tag{40}$$

where a, b and r are constants.

It is also possible to develop an empirical nonparametric recession curve relationship with relative storage under the assumption that recharge from the unsaturated zone is negligible and that only baseflow contributes to the stream discharges. Traditionally, multiple recession curves for a catchment were overlain manually. A set of computer routines for this purpose was reported by Lamb and Beven (1997).

The baseflow recession constant, K_b, is often thought of as a hydrogeologic index, since it attempts to represent concisely the complex hydrogeologic processes in a catchment. In the linear reservoir model with a daily time step and no groundwater recharge, K_b represents the fraction of yesterday's baseflow, that is expected today. Another common hydrogeologic index is the baseflow index, which is an estimate of the fraction of annual streamflow that is comprised of baseflow (Institute of Hydrology, 1980). To estimate the baseflow index, streamflow is divided into 5 day blocks, and the minimum discharge of each of these blocks is flagged. The flagged discharges are put in a time series, and grouped in overlapping groups of three. If 0.9 times the central value of the group is less than both of the outer values, then the central value is identified as a turning point. By connecting all the turning points with straight lines, and estimating the area under the lines, an estimate of baseflow discharge, V_B, is obtained. After calculating the total volume of streamflow discharge using the original daily streamflows, V_A, the baseflow index can be estimated as:

$$\text{BFI} = \frac{V_B}{V_A} \tag{41}$$

The annual values of BFI have been found to be relatively stable, with a typical standard deviation of 0.04 (Institute of Hydrology, 1980).

Estimating Hydrogeologic Indices from Watershed Characteristics

Of great interest is how to estimate hydrogeologic indices, such as K_b and BFI, from available watershed information. This would allow parameterization of these indices at ungauged river sites. Various researchers have attempted to examine the geologic controls on groundwater storage and baseflow discharge. These studies have met with very limited success, and have often been qualitative in nature (see Tallaksen, 1995, for a list of citations). Some of these studies are discussed below.

Baseflow is a function of many factors, including topography, geology, soils, vegetation, and climate (Lacey and Grayson, 1998). These factors vary both spatially and/or

temporally in most watersheds, thus making it difficult to estimate an integrated watershed baseflow response from these factors. Zecharias and Brutsaert (1988) examined geomorphic impacts on groundwater discharge in 19 Appalachian watersheds. They found a function of baseflow recession, which they called the reaction factor, to be significantly correlated with basin ground surface slope. Aquifer breadth and the effect of a soil parameter, hydraulic conductivity divided by drainable porosity, were not significant at a 5% level.

Nathan *et al.* (1996, as cited in Lacey and Grayson, 1998) performed a regression analysis of baseflow index (BFI) versus watershed characteristics at 164 catchments in Victoria, Australia. They found drainage area, stream length, and elevation of the catchment centroid to be significant variables. Lacey and Grayson (1998) attempted to examine likely physical controls on baseflow response. They related BFI to a number of dimensionless topographic and climatic parameters, and a classification scheme representing geology, vegetation, and soils. Unfortunately, they found no trends in plots of BFI versus any of their explanatory variables.

While our instincts tell us that hydrogeologic indices should be a function of watershed characteristics, our ability to develop such relationships using actual data has been difficult. This may be due to a number of reasons, including difficulties in measuring watershed properties (such as hydrogeology), the lack of homogeneity and complex physical structure of most watersheds, and uncertainty in the derived hydrogeologic indices. In spite of this, this topic is ripe for further research activity, and will aid in our understanding and ability to predict hydrologic response at ungauged watersheds.

REFERENCES

Anderson M.G. and Burt T.P. (1980) Interpretation of recession flow. *Journal of Hydrology*, **46**, 89–101.

Barnes B.S. (1939) The structure of discharge-recession curves. *Transactions of American Geophysical Union*, **20**, 721–725.

Bathurst J.C., Wicks J.M. and O'Connell P.E. (1995) The SHE/SHESED basin scale water flow and sediment transport modelling system. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 563–594.

Bergstrom S. (1995) The HBV model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 443–476.

Beven K.J. (1982) On subsurface stormflow: predictions with simple kinematic theory for saturated and unsaturated flows. *Water Resources Research*, **18**(6), 1627–1633.

Beven K.J. (1989) Changing ideas in hydrology - the case of physically-based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (2001) *Rainfall-Runoff Modelling: The Primer*, John Wiley & Sons: Chichester.

Beven K.J., Lamb R., Quinn P., Romanowicz R. and Freer J. (1995) Topmodel. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resource Publications: pp. 627–668.

Box G.E.P. and Jenkins G.M. (1976) *Time Series Analysis: Forecasting and Control*, Holden-Day: Oakland.

Brutsaert W. (1994) The unit response of groundwater outflow from a hillslope. *Water Resources Research*, **30**(10), 2759–2763.

Brutsaert W. and El-Kadi A. (1984) The relative importance of compressibility and partial saturation in unconfined groundwater flow. *Water Resources Research*, **20**, 4000–4408.

Brutsaert W. and Lopez J.P. (1998) Basin-scale geohydrologic drought flow features of riparian aquifers in the southern great plains. *Water Resources Research*, **34**(2), 233–240.

Brutsaert W. and Nieber J.L. (1977) Regionalized drought flow hydrographs from a mature glaciated plateau. *Water Resources Research*, **13**, 637–643.

Burnash, R., and Ferral, L. (2002) *Conceptualization of the Sacramento Soil Moisture Accounting Model*. Available on World Wide Web: **`http://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/html/2sac-sma.htm`**

Childs E.C. (1971) Drainage of groundwater resting on a sloping bed. *Water Resources Research*, **7**, 1256–1263.

Douglas E.M., Vogel R.M. and Kroll C.N. (2000) Trends in floods and low flows in the United States: impact of spatial correlation. *Journal of Hydrology*, **240**, 90–105.

Duan J. and Miller N.L. (1997) A generalised power function for the subsurface transmissivity profile in TOPMODEL. *Water Resources Research*, **33**, 2559–2562.

EM-DAT. (2003) *International Disaster Database*, Brussels. Available on World Wide Web: **`http://www.cred.be/emdat`**

Eng K. and Brutsaert W. (1999) The generality of drought flow characteristics within the Arkansas River basin. *Journal of Geophysical Research*, **19**, 435–441.

Feldman A.D. (1995) HEC-1 flood hydrograph package. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 119–150.

Franchini M. and Pacciani M. (1991) Comparative analysis of several conceptual rainfall-runoff models. *Journal of Hydrology*, **122**, 161–219.

Grayson R.B., Bloschl G. and Moore I.D. (1995) Distributed parameter hydrologic modelling using vector elevation data: THALES and TAPES-C. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 669–696.

Haith D.A., Mandel R. and Wu R.S. (1992) *GWLF: Generalized Watershed Loading Functions, Version 2.0, User's Manual*. Cornell University: Ithaca.

Hall F.R. (1968) Base flow recessions-a review. *Water Resources Research*, **4**, 973–983.

Hirsch R.M. (1982) A comparison of four streamflow record extension techniques. *Water Resources Research*, **18**(4), 1091–1088.

Hogarth W.L., Parlange J.Y., Parlange M.B. and Lockington D. (1999) Approximate analytical solution of the Boussinesq equation with numerical validation. *Water Resources Research*, **35**(10), 3193–3197.

Hornberger G.M., Ebert J. and Remson I. (1970) Numerical solution of the Boussinesq equation for aquifer-stream interaction. *Water Resources Research*, **6**(2), 601–608.

Institute of Hydrology (1980) *Low-Flow Studies*, Report No. 1, Institute of Hydrology, Wallingfor.

Iorgulescu I. and Musy A. (1997) Generalisation of TOPMODEL for a power law transmissivity profile. *Hydrological Processes*, **11**, 1353–1355.

James L.D. and Thompson W.O. (1970) Least squares estimation of constants in a linear recession model. *Water Resources Research*, **6**(4), 1062–1069.

Kroll C.N. (1989) *The estimation and usage of baseflow recession constants*. Master of Science Thesis, Tufts University, Medford, p. 131.

Kroll C.N., Luz J.G., Allen T.B., and Vogel R.M. (2004) Developing a watershed characteristics database to improve low streamflow prediction. *Journal of Hydrologic Engineering*, **9**(2), 116–125.

Kulandaiswamy V.C. and Seetharaman S. (1969) A note on Barnes' method of hydrograph separation. *Journal of Hydrology*, **9**, 222–229.

Lacey G.C. and Grayson R.B. (1998) Relating baseflow to catchment properties in south-eastern Australia. *Journal of Hydrology*, **204**, 231–250.

Lamb R. and Beven K.J. (1997) Using interactive recession curve analysis to specify a general catchment storage model. *Hydrology and Earth System Sciences*, **1**, 101–113.

Lamb R., Beven K.J. and Myrabø S. (1997) Discharge and water table predictions using a generalised TOPMODEL formulation. *Hydrological Processes*, **11**(9), 1145–1168.

Leavesley G.H. and Stannard L.G. (1995) The precipitation-runoff modeling system-PRMS. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 281–310.

Mizumura K. (2002) Drought flow from hillslope. *Journal of Hydrologic Engineering*, **7**, 109–114.

Nathan R.J., Austin K., Crawford D. and Jayasuriya N. (1996) The estimation of monthly yield in ungauged catchments using a lumped conceptual model. *Australia Journal of Water Resources*, **1**(2), 65–75.

Nutbrown D.A. (1975) Normal mode analysis of the linear equation of groundwater flow. *Water Resources Research*, **11**(6), 979–987.

Otnes J. (1978) T ⇑ rrv ≥ rskurven. In *Hydrologi i Praksis*, Otnes J. and R ≥ stad E. (Eds.), Ingenio ⇑ rforlaget: Oslo, pp. 228–233.

Paniconi C., Troch P.A., van Loon E.E. and Hilberts A.G.J. (2003) The hillslope-storage Boussinesq model for subsurface flow and variable source areas along complex hillslopes: 2, intercomparison with a 3D Richards equation model. *Water Resources Research*, **39**(11), 1317.

Perzyna G. (1990) *Derived Frequency Distribution of Low Flows*, Doctor of Science Thesis, Institute of Geophysics, University of Olso, Oslo.

Quick M.C. (1995) The UBC watershed model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 233–280.

Singh V.P. (1995) *Computer Models of Watershed Hydrology*, Water Resource Publications: Highlands Ranch.

Smakhtin V.U. (2001) Low flow hydrology: a review. *Journal of Hydrology*, **240**, 147–186.

Spolia S.K. and Chander S. (1974) Modelling of surface runoff systems by an ARMA model. *Journal of Hydrology*, **22**, 317–332.

Spolia S.K. and Chander S. (1979) Modelling of surface runoff systems by an ARMA model – a correction. *Journal of Hydrology*, **41**, 187.

Sugawara M. (1995) Tank model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publications: Highlands Ranch, pp. 165–214.

Szilagyi J. and Parlange M.B. (1998) Baseflow separation based on analytical solutions of the Boussinesq equation. *Journal of Hydrology*, **204**, 251–260.

Tallaksen L.M. (1995) A review of baseflow recession analysis. *Journal of Hydrology*, **165**, 349–370.

Toebes C. and Strang D.D. (1964) On recession curves, 1-recession equations. *Journal of Hydrology, New Zealand*, **3**(2), 2–15.

Troch P.A., De Troch F.P. and Brutsaert W. (1993) Effective water table depth to describe initial conditions prior to storm rainfall in humid regions. *Water Resources Research*, **29**(2), 427–434.

Troch P.A., Paniconi C. and van Loon E.E. (2003) The hillslope-storage Boussinesq model for subsurface flow and variable source areas along complex hillslopes: 1. formulation and characteristic response. *Water Resources Research*, **39**(11), 1316.

Verhoest N.E. and Troch P.A. (2000) Some analytical solutions of the linearized Boussinesq equation with recharge for a sloping aquifer. *Water Resources Research*, **36**(3), 793–800.

Vogel R.M. and Kroll C.N. (1992) Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics. *Water Resources Research*, **28**(9), 2451–2458.

Vogel R.M. and Kroll C.N. (1996) Estimation of baseflow recession constants. *Water Resources Management*, **10**, 303–320.

Werner P.W. and Sundquist K.J. (1951) On the groundwater recession curve for large watersheds. *IAHS Publication*, **33**, 202–212.

Wigmosta M.S., Vail L.W. and Lettenmaier D.P. (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, **30**(6), 1665–1679.

Wittenberg H. and Sivapalan M. (1999) Watershed groundwater balance estimation using streamflow recession analysis and baseflow separation. *Journal of Hydrology*, **219**, 20–33.

Zecharias Y.B. and Brutsaert W. (1988) Ground surface slope as a basin scale parameter. *Water Resources Research*, **21**(12), 1895–1902.

127: Rainfall-runoff Modeling: Distributed Models

PAOLO REGGIANI[1] AND JAAP SCHELLEKENS[1,2]

[1] *Inland Water Systems, Foundation Delft Hydraulics, Delft, The Netherlands*
[2] *Earth and Life Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*

The present article presents a discussion of distributed modeling used in watershed hydrology from a modern perspective. The need for distributed modeling is explained on the basis of situations encountered in hydrological practice. The basic equations governing water and energy transfer in hydrology are introduced at the point- or microscale. It is explained how these equations are used in the different approaches through change in spatial scale from the microscale to an intermediate scale, the macroscale, or the megascale, which corresponds to the spatial dimension of some smaller or larger parts of the catchments. The strengths and limitations that have been recognized for the various approaches will be pointed out. As a possible alternative a flux-based approach, formulated entirely at the megascale, is highlighted. Its potential scope for practical applications will be discussed in the context of the closure problem for mass momentum and energy fluxes in hydrology.

INTRODUCTION

The demands by policymakers, implementers, and stakeholders with respect to the prognostic capabilities of hydrological models become increasingly sophisticated. Experts are asked to predict hydrological variables, such as water table depth, y, soil saturation, s, saturated catchment surface fraction ω, pore pressure p, flow velocity $\vec{v}$, and hydrological fluxes like surface runoff R, infiltration I, bare soil evaporation E_s, plant transpiration E_v, rainfall interception E_i, recharge of the water table, and capillary rise, q, in a spatially distributed fashion and at a level of accuracy not demanded previously. These demands are primarily driven by two reasons:

Firstly, it is widely recognized that decisions on land-use practices, deployment of infrastructure, exploitation of natural resources, or change of climate can irreversibly perturb the local and the large-scale equilibrium of a hydrological system and, subsequently, alter runoff behavior, yield and water quality, with potentially adverse socioeconomic implications for stakeholders, economic actors, and current, as well as future generations living within or downstream of the watershed boundaries.

Secondly, there is a common belief, which is, however, not shared by several experts (see e.g. the essay by Grayson *et al.*, 1992 or the debate between Beven, 1996a, b; Refsgaard and Storm, 1996), that the recently experienced leap in availability and affordability of radar-, space-, and airborne high-resolution datasets will lead to significant progress in the representation of hydrological systems through digital simulation models, thus bringing a "sufficiently accurate" description within reach.

By the same token, the prognosis of the spatially distributed hydrological variables y, s, ω, p, and $\vec{v}$ cannot be achieved at all using *lumped models*, or at least not at the level of spatial resolution required by the end-user. Among those models, we list the pioneering Stanford Watershed model by Crawford and Linsley (1966), and its spin-off, the Sacramento model, the largely operationally used HBV model described by Bergström (1995), the LASCAM model by Viney and Sivapalan (1999, 2001) or the Xinanjiang model by Zhao (1977) and Zhao and Liu (1995). This argument also applies to their possible grid-based, rasterized applications.

The zero-dimensional structure of lumped models is based on the conservation of mass and represents entire portions of the landscape as interconnected reservoirs, for

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

which the mass fluxes exchanged between entities are parameterized in terms of power-law-type relationships. Conservation of momentum or energy is not considered explicitly, nor is the effect of topography and its relation with gravity. Nevertheless, the role of such models as robust forecasting tools for operational watershed management is invaluable if merely discharges or water quality at a catchment outlet need to be predicted over a restricted period into the future. For those periods, changes within the system remain negligible and the statically calibrated lumped model structure continues to give acceptable simulations.

The listed hydrological variables, together with the hydrodynamic properties of the shallow subsurface, may control the hydrological fluxes R, I, E_s, E_v, E_i, and q. These act as source or sink terms for moisture and energy stored in the subsurface, the canopy, and the planetary boundary layer. Therefore, they have a direct impact on the mass and energy budget of the soil-vegetation-atmosphere continuum, and consequently on weather formation and its long-term pattern, the climate.

In this context, it is important to note that predicting the behavior of the hydrological variables or the fluxes can be required in response to forcing by individual events ("forcing" can be either wetting during storms or drying during inter-storm periods). A typical example of this kind of application is the use of a hydrological model for flood or drought forecasting purposes.

Alternatively, the long-term behavior of the variables and fluxes may need to be studied. In this case, the modeler is not interested in their fluctuations induced by single events, but in their mean behavior over a critical period, typically a year or longer. The aim of such an analysis is to understand catchment-average response to particular forcing patterns classified collectively as "climate", which is expressed in terms of time-averaged fluxes and long-term water yield. Such an analysis, based on particular restrictive hypotheses, has been carried out by Eagleson (1978a, b, 1982) and Eagleson and Tellers (1982). Salvucci and Entekhabi (1995) extended the theory of Eagleson to an entire hillslope in the presence of a shallow water table. They carried out case studies for a variety of different hillslope shapes, soil parameters, and climatic forcing. By applying restrictive hypotheses yielding mathematically tractable expressions for the various time-averaged fluxes, it is deduced that the climate, together with local factors such as the geology and topography, exerts control on the hydrological fluxes and ultimately on the yield, expressed in terms of a single hydrological state variable, namely, the space and time average saturation of the soil column.

At a practical level, such hydrological equilibrium theories and underlying hypotheses would require verification via distributed modeling tools that are able to explicitly account for spatial variability of land use, geology, and topography. Their actual implementation, however, will be hampered by the difficulty of determining the necessary parameterizations for the fluxes (see e.g. Beven, 1989, and also Grayson *et al.*, 1992 and Beven, 1993, 2001).

As well as circulating water and energy, the hydrological cycle is also the vehicle for other substances, such as nutrients, pesticides, and agricultural fertilizers that influence water quality in rivers, lakes, coastal zones, and the shallow subsurface. The impact of policies on fertilizer and pesticide application at a national or international level (e.g. the EU Water Framework Directive) is strongly linked to the hydrological cycle. The effect of agricultural management practices on water quality can be studied efficiently only with hydrological modeling tools permitting the prognosis of y, s, p, ω, and $\vec{v}$, and of the resulting fluxes R, I, E_s, E_v, E_i, q at a sufficiently high spatial resolution, such as, for example, the characteristic size of land-use patterns of forested and cultivated areas.

Similar arguments can be put forward for the study of other problems related to the hydrological cycle. Amongst those we list the quantification of the impacts of agricultural practices on soil erosion in the tropics, shown by Bruijnzeel (1990, 2004), or the problem of dryland salinity in Australia, studied by Peck and Hurle (1973).

Finally, the continuous development of global circulation models (GCM) used to forecast future climate scenarios (for relevant references see e.g. the IPCC, 2001 report), requires an increasingly detailed description of the soil-vegetation-atmosphere feedback for an accurate representation of the mass, momentum, and energy fluxes exchanged with the atmospheric boundary layer (ABL). These constitute the lower boundary conditions for the equations governing atmospheric mass, momentum, and energy circulation. This should be achieved by representing the hydrological fluxes in a spatially distributed manner by taking the shape of drainage basins and the landscape explicitly into account (see e.g. Koster *et al.*, 2000), rather than through a lumped representation of landscape entities as conceptual buckets matching the GCM grid cell size.

The issues addressed here are complicated further because the phenomena evolve over vastly different spatial scales that can range from a few centimeters for a wetting front infiltrating into the unsaturated zone, up to spatial scales of several kilometers for mass, momentum, and energy fluxes for land–atmosphere interaction processes. Similar arguments apply to temporal scales that can extend from a few seconds, typical for eddy fluctuation of an air current flowing over a forest canopy, to the several years or decades that it can take a pesticide to reach the water table through filtration across the unsaturated zone.

The present article looks at commonly used distributed modeling approaches, suited for the prediction of a subset of y, s, p, ω, and $\vec{v}$, whereby an attempt is made to explain the approaches starting from first principles: the balance laws for a continuum formulated at the *microscale*. By averaging

the equations, these are mapped to a particular, usually larger, scale of interest. From this perspective, it will be shown how various authors of distributed hydrological models have cast the balance laws into a form, which would suit the description of a hydrological system in mathematical and physical terms. They did so with the aid of a series of hypotheses, and by choosing the spatial domain through a sensible partitioning of the landscape into averaging regions and domains of integration.

BALANCE LAWS

Distributed hydrological models are commonly based on the conservation equations for mass, momentum, and energy, whereby the mass balance is employed in virtually all model formulations. The momentum equation is difficult to apply to hydrological systems because it is vectorial and it is generally seen as difficult to close the various force terms via suitable parameterizations. The most common use of the momentum balance in hydrology is made through Darcy's law, a parameterization of the up-scaled conservation of momentum for water contained within the voids of a porous medium at the scale of a conceptual *Representative Elementary Volume* (REV; Bear, 1979; Hassanizadeh and Gray, 1980).

People describing the hydrodynamics of open water systems such as lakes, rivers, and estuaries feel more comfortable with the momentum equation, especially as they have the clear advantage of working with one and the same medium. In addition, observation permits a far better knowledge of the boundary conditions and geometry in terms of, for example, the bathymetry and wind shear on the open water surface, than for hydrologists, who often do not know where the water is (for more discussion see e.g. Beven, 2001).

The energy equation, especially the conservation of thermal energy, is necessary for describing land–atmosphere interaction across the soil-vegetation-atmosphere continuum. This is particularly relevant in view of the fact that in dry areas of the world up to 95% of the mean annual precipitation leaves the system through evapo-transpiration and not via surface runoff. In this superficial part of the system, which is exposed to the incoming radiation energy, it is essential to capture the terrestrial energy balance, as it greatly influences the bare soil evaporation E_s, rainfall interception E_i, and plant transpiration E_v. The thermal energy balance is rarely explicitly included in hydrological model formulations, mainly because it plays no significant role when merely runoff processes are investigated.

Balance laws formulated at the microscale describe the conservation of a physical property at a point in a phase, where with "*phase*" we understand water, air, or the porous medium soil matrix. The mass, momentum, and energy conservation laws are written as follows:

Balance of mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = R \tag{1}$$

where ρ is the mass density for the phase and $\vec{v}$ is the phase velocity, a vector.

Balance of momentum:

$$\frac{\partial (\rho \vec{v})}{\partial t} + \nabla \cdot (\rho \vec{v}\vec{v} - \vec{T}) - \rho \vec{g} = 0 \tag{2}$$

where $\vec{T}$ is the stress tensor for the phase and $\vec{g}$ is the gravity vector.

Balance of total energy:

$$\rho \frac{\partial}{\partial t} \left(E + \frac{1}{2} v^2 \right) + \nabla \cdot (\vec{q} - \vec{T} \cdot \vec{v}) + h - \rho \vec{g} \cdot \vec{v} = 0 \tag{3a}$$

where E is the internal energy, $\vec{q}$ is the heat-flux vector, and h is the external energy supply (e.g. radiation). Scalar multiplication of equation (2) with $\vec{v}$ and subtraction from (3a) yields the thermal energy equation:

Balance of thermal energy:

$$\rho \frac{\partial}{\partial t} E + \nabla \cdot \vec{q} + h = 0 \tag{3b}$$

For the water in overland or channel flow, only a single phase is present. In the subsurface, different phases coexist and the balance laws hold at every point in each phase. The laws form the theoretical basis for a quantitative description of a physical system.

CHANGING SPATIAL SCALE

In the representation of environmental systems and their subset, hydrological systems, we commonly map the balance laws from a representation at the *microscale* to larger scales of interest, referred to as the *macroscale* and the *megascale* by Gray *et al.* (1993). The *macroscale* D is defined as an intermediate scale of interest that is larger than the *microscale* l, but smaller than the size of the entire modeling domain, denoted as the *megascale,* L. It is not uniquely defined, but its choice depends on the particular needs of the modeler. In the context of the present article, with *megascale,* we understand the spatial scale corresponding to a small part of a catchment. However, in the context of a channel or overland flow sheet, for example, the dimension of the cross section or the depth of the flow sheet is also considered *megascopic* because it represents one or two dimension(s) of the domain.

The mutual relationship between *micro, macro*, and *megascale* obeys to the relation:

$$l < D < L \tag{4}$$

While in a *macroscale* representation of a system, internal variations of properties in space are considered, at the *megascale* spatial variations are ignored and properties are expressed as their averages over the domain.

These concepts can be explained by looking at a few approaches used typically in the description of environmental systems. For example, the integration of the microscale balance equations (1), (2), and (3) over different phases, such as water, solid, and air contained within a spherical REV, yields a macroscale representation of multiphase flow (Bear, 1979; Hassanizadeh and Gray, 1979a, b). The upscaled equations represent the conservation laws at a new *macroscale* point with a diameter equal to that of the REV. The REV is larger than the *microscale* but smaller than the size of the domain. Application of commonly accepted hypotheses for porous media flow yields a generalized form of Darcy's law for multiphase flow.

The microscopic balance equations can be scaled to the macroscale over two dimensions out of three, and to the *megascale* over the third dimension. For example, integration over an upright cylindrical REV across a confined aquifer leads to horizontal flow equations that are *macroscopic* along the two horizontal directions (*macroscopic* dimension D) and *megascopic* over the vertical dimension of the aquifer (*megascopic* dimension L).

Alternatively, we can map conservation equations directly from the *microscale* to the *macroscale* in two dimensions, while preserving a *microscale* representation along the third dimension. For example, if (1) and (2) are integrated over a slab of infinitesimal thickness representing the cross section of a river channel, we obtain a description of the system that is *megascopic* over the cross section and *microscopic* along the channel axis. Application of the classic hypotheses assumed for open channel flow reduces the scaled balance laws to the Barré de Saint-Venant (1871) equations used to model one-dimensional flow in a channel.

This perspective allows mapping physical conservation laws consistently between scales of representation without requiring *a-priori* assumptions. Such assumptions may lead to the neglect of important terms early in the scaling process, thereby restricting the application to the particular situations for which the hypotheses hold.

This framework also casts a particular view on model formulations underlying the Freeze and Harlan (1969) blueprint and model (Freeze 1971, 1972a, b), SHE Abbott *et al.* (1986a, b), InHM (VanderKwaak and Loague, 2001), TOPOG (O'Loughlin, 1986), Vertessy *et al.* (1993), TOPMODEL (Beven and Kirkby, 1979), and other distributed modeling tools. A formulation developed entirely at the *megascale* and proposed as the *Representative Elementary Watershed* (REW) approach by Reggiani *et al.* (1998, 1999, 2000) is based on up-scaling the balance equations from the micro to the megascale and given here as a possible alternative distributed model concept.

TRADITIONAL DISTRIBUTED FORMULATIONS

In the following subsections on "The Freeze–Harlan (FH) blueprint and derived models" and "Other distributed models", we introduce the most commonly used distributed modeling tools that are based on the solution of the upscaled balance equations by means of partial differential equations (PDE) (e.g. SHE; InHM) or via analytical expressions (e.g. TOPMODEL or the TOPKAPI model of Liu and Todini, 2002). Using PDE's in hydrology is commonly seen as a significant effort, mainly because of the difficulty associated with determining the necessary *effective* parameter values in a spatially distributed fashion for a heterogeneous domain like a watershed. Also, the simultaneous solution of coupled PDEs is numerically and computationally demanding. The hydrological variables are calculated at the nodal points of a model mesh, at a spatial resolution that is not required for most practical applications. The full validation of the results is data-intensive, as pointed out by Refsgaard (2000) or Beven (2002a, b), and is rarely performed for many distributed model applications. Thus, many distributed models remain insufficiently validated. Limited verifications are presented only in terms of outflow hydrograph comparisons instead of a satisfactory prediction of spatially distributed variables, such as, for example, piezometric heads, soil moisture, or the spatial pattern of overland flow generation. Such a validation is also often precluded by the lack of sufficient high-quality spatial data.

Another difficulty in applying the equations lies in the representation of processes occurring at scales smaller than the scale of the discretization grid (e.g. Finite Element Method (FEM) or Finite Difference Method (FDM) mesh size) or other spatial model entities (e.g. flow tubes, TOPOG elements, or the pixel size). Hydrologists refer to these processes collectively as *subgrid* variability. Subscale processes are usually less relevant for simulating a lumped variable such as the outflow hydrograph. It becomes relevant for the representation of other spatially distributed hydrological processes such as runoff generation due to seepage or infiltration excess or localized groundwater recharge and the formation of perched systems, to mention just a few. These subscale processes can hardly be represented in terms of the independent variables calculated at the model grid nodes. The main reason is that the high heterogeneity of the soils and subsurface pathways is difficult to capture via parameters such as the saturated hydraulic conductivity, lumped at the scale of a *REV* or model mesh

size, despite the fact that the assumption of the REV forms the basic hypothesis for the description of porous media flow and is generally accepted for the description of a variety of hydrological processes. A common example for the just mentioned difficulties is the procedure of expressing a water retention curve at the scale of a model mesh element in terms of soil structural and textural parameters identified at the laboratory scale and inserting the resulting expression into a Richard's equation solver, which is implemented at the grid size of a kilometer.

It is, however, understandably argued by Refsgaard and Storm (1996) that to date these tools constitute the *best there is* and that improvement of future air or space-borne measurement techniques will help to bring data resolution and scale of process representation closer together. The following paragraphs will briefly review the most commonly used distributed modeling approaches.

The Freeze–Harlan (FH) Blueprint and Derived Models

The first and most radical attempt to propose a complete distributed modeling framework is known as *Freeze–Harlan Blueprint* (1969) (see also Freeze, 1978). The authors of the blueprint proposed using the conservation equations (1) and (2) to describe flow within a watershed. Most of these equations originated from various disciplines that are linked for historical reasons to engineering problems. It is only later that the fields of their original application have been extended to more general hydrological systems. Typically, the Saint-Venant equations (Barré de Saint-Venant, 1871) come from the domain of river engineering and Richards' equation from soil physics (Richards, 1931) as an extension of the empirical Darcy's Law for saturated flow through a porous medium (Darcy, 1856). Lighthill and Whitham (1955) were the first to show through a mathematical analysis that the kinematic wave equation could be applied to predict flow in rivers under circumstances for which the spatial derivatives of stage and velocity could be considered negligible (in contrast to dynamic waves for which this is not the case). Henderson and Wooding (1964) extended the use of the kinematic wave to predict flow on hillslopes and subsurface flow.

Freeze and Harlan aggregated these equations into a modeling framework, which would cover most flow types observed in a watershed through an appropriate continuum description in terms of PDE and suitable parameterizations of the fluxes. The equations constitute up-scaled versions of equations (1) and (2), averaged to the macroscale over appropriate REVs (Darcy's law for saturated or unsaturated porous media flow) or integrated to the *megascale* over one or two spatial dimensions (to represent averaged variables over the flow depth or flow cross section respectively). In those days, putting the approach into practice seemed to be limited solely by computer power and insufficient availability of spatial data.

Different interpretations and numerical implementations of the Freeze–Harlan approach led to various catchment models based on the outlined blueprint. The most widely known is the *Système Hydrologique Europeén* (SHE) model developed by different European laboratories. The SHE model has been described first by Abbott *et al.* (1986a, b) and is currently distributed by the Danish Hydraulic Institute under the name of MIKE-SHE (*see* e.g. Refsgaard and Storm, 1995) and by Newcastle University under the name SHETRAN (e.g. Birkinshaw and Ewen, 2000; Bathurst *et al.*, 2004). In SHE, the governing PDE equations are solved through finite difference schemes. Through recent work by Christiansen *et al.* (2004), the model has been adapted in order to be able to take macropore flow into account.

The *Institute of Hydrology Distributed Model* (IHDM), described in the report by Beven *et al.* (1987) and summarized by Calver and Wood (1995), is based on similar design principles as SHE, but relies on the FEM in variable-width hillslope elements for the solution of PDEs.

Similar models are still being proposed and implemented. Among the most recent contributions, we list the *Quasi-physically Based Rainfall-runoff Model* (QPBRRM) based on simplified governing equations and applied by Loague and Kyriakidis (1997) to reproduce the response of the 0.1 km^2 R5 Washita River Experimental Watershed catchment, Oklahoma. The even more numerically sophisticated 3-D PDE-based *Integrated Hydrology Model* (InHM) (**`http://inhm.org`**) accounts for variably saturated flow and transport in macropores and 2-D flow and transport over the land surface and in open channels. First-order coupling effects can be accounted for explicitly. The tool has been used for research work (VanderKwaak and Loague, 2001; Loague and VanderKwaak, 2002) on the R5 experimental watershed and is currently being applied also to large-scale applications including solute transport (Loague and VanderKwaak, 2004). Figure 1 shows a typical example of a 3-D model element mesh, in this particular case, the *InHM* finite element grid, used to model the R5 catchment, and Figure 2 shows some results on flow velocities and water depth computed for the overland flow, given a particular precipitation event.

Similarly, the SHE model has been employed on the Rimbaud catchment in France by Parkin *et al.* (1996) or on the 440 km^2 Karup catchment in Denmark (Refsgaard, 1997). Paniconi and Wood (1993) proposed another 3-D unsaturated–saturated simulation code, for which an application in combination with data assimilation techniques (Newtonian nudging) for distributed physically based hydrological modeling is shown in Paniconi *et al.* (2003).

All these tools have in common that they solve essentially coupled PDE-based governing equations through

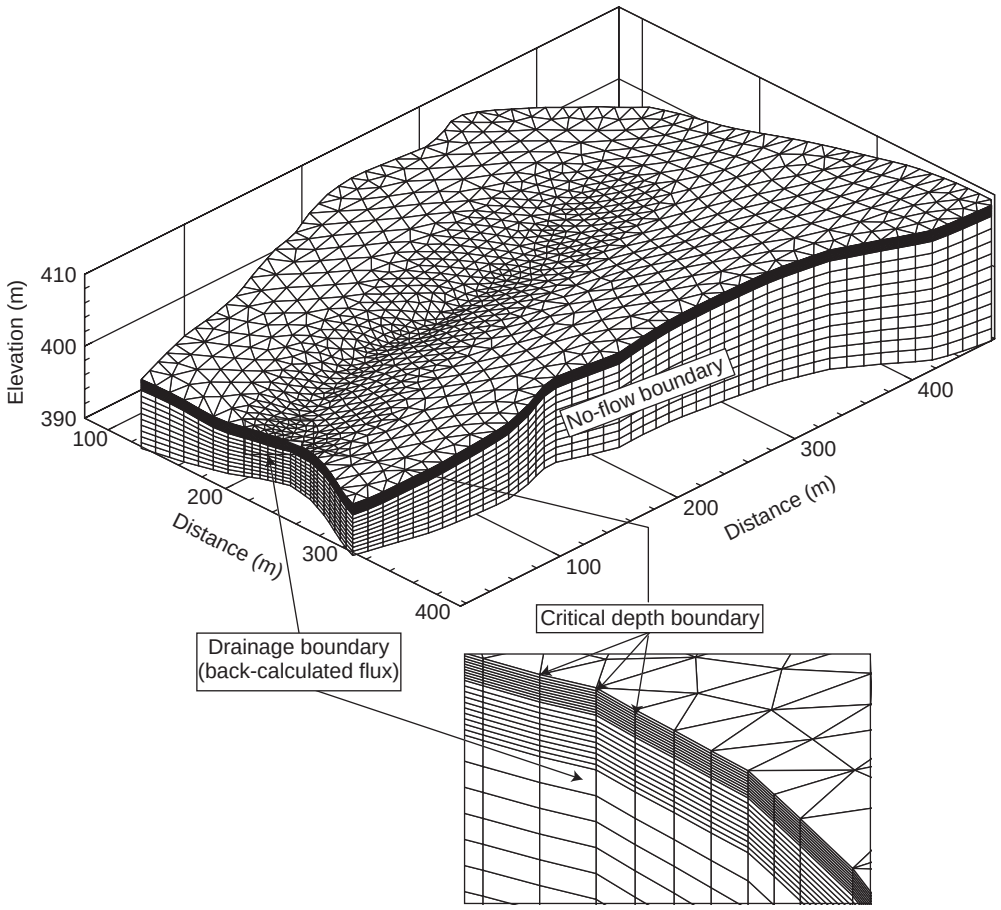


Figure 1 The Integrated Hydrology Model (InHM) Finite Element Mesh used to model the R5 Watershed (JE VanderKwaak and K Loague, Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model, Water Resources Research, **37**(4), pp. 999–1013 (2001) American Geophysical Union. Reproduced by permission of American Geophysical Union)

finite element or finite difference schemes and focus on describing flow processes at the *macroscale* or various combinations of *micro* and *macroscale.*

It is interesting to observe that the choice of titles for the series of publications by Loague and coworkers (e.g. Loague and Freeze 1985; Loague and Kyriakidis 1997 or Loague *et al.*, 2000) and others (Beven, 1989, 1993, 2001; Woolhiser, 1996) is very much an indication for the challenges faced in reproducing hydrologic behavior even for a small basin like the R5 at Chickasha, Oklahoma, for which data are available in quantities and at a spatial resolution that are limited to only a few exceptionally monitored experimental watersheds.

However, the principal criticisms expressed with respect to physically based distributed modeling approaches focus on the Freeze–Harlan (1969) blueprint itself, rather than on any particular derived model. Several concerns have been raised during the last two decades by Beven (1989, 1993, 1996a, b), or Grayson *et al.* (1992). A flux-based approach with a representation at the *megascale (see* Section "The REW approach"), along the lines of the REW concept, is envisaged as a possible alternative to the Freeze–Harlan (FH) blueprint.

TOPMODEL

TOPMODEL (Beven and Kirkby, 1979) is a distributed simulation tool, which calculates runoff production because of saturation excess along 1-D flow tubes representing the catchment hillslopes. The concept of flow tubes was later replaced by a terrain analysis based on raster grid data, for which a topographic index is calculated on a per-pixel-basis using the digital terrain analysis program DTM Analysis (see `http://www.es.lancs.ac.uk/hfdg/topmodel.html`). In TOPMODEL, the transmissivity, or vertically integrated saturated hydraulic conductivity at a particular location, is estimated by assuming an exponential decline of the conductivity over depth and by vertically integrating the resulting mathematical expression over the depth of the saturated soil layer. The authors also assume that the shallow subsurface flow is driven by a hydraulic gradient equal

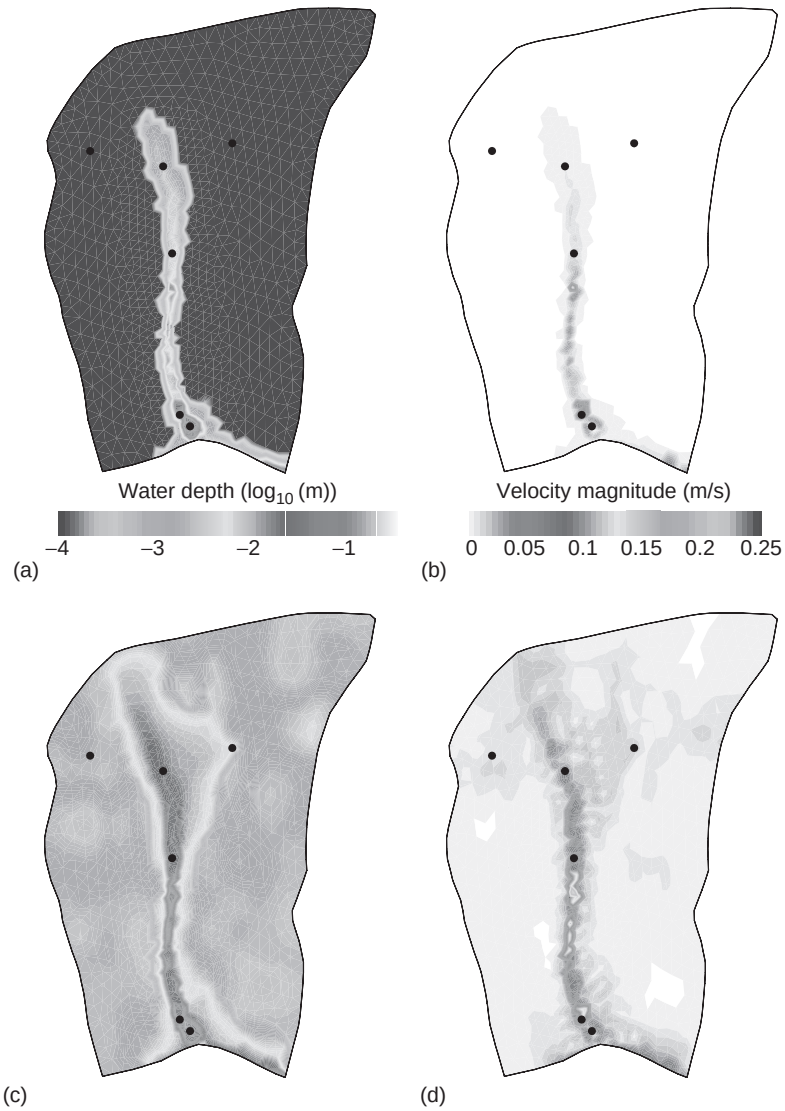


Figure 2 Plan view plots across the R-5 catchment of the I*n*HM-simulated surface water depths. (a) and (c) and velocities (b) and (d) for a particular precipitation event; (a) and (b) are the initial conditions, (c) and (d) are at the time of peak stream flow (JE VanderKwaak and K Loague, Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model, Water Resources Research, **37**(4), pp 999–1013 (2001) American Geophysical Union Reproduced by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

to the local topographical slope. Thanks to the exponential form of the conductivity, an analytical expression for the transmissivity is found. This expression is subsequently used to define a water table shape for a given catchment storage as if local discharge was equal to a constant recharge to the water table integrated over the catchment area accumulated uphill of a particular contour segment.

The authors manipulate this equality, recasting it in terms of (i) a topographic index, (ii) an average water table depth over the catchment, and (iii) the local moisture deficit (or water table depth) at the particular point. They establish a criterion that allows regions that are unable to discharge the subsurface runoff from uphill to be identified, because either the local water table is too high or the local hydraulic gradient is too small. The excess flow leads to local saturation of the subsurface and is evacuated as surface runoff. From a deterministic perspective, the authors utilize a *macroscopic*, vertically averaged momentum balance equation in steady state to estimate local subsurface runoff, while the average water table is controlled by a *megascopic* (catchment-scale) water balance equation for the entire catchment. The surface excess water budget and the channel routing component are also controlled by respective mass balance equations at the *megascale*.

This fundamental concept has served as platform for many other spin-off projects. Just to cite a few, an extension to include a dimensionless version (Sivapalan *et al.*, 1987), to handle spatially variable soil depths (e.g. Saulnier *et al.*, 1998), water quality issues (e.g. Robson *et al.*, 1992; Hornberger *et al.*, 1994), or the study of particular features such as spatial variability of subsurface runoff (e.g. Woods *et al.*, 1997). A further important application is the use of the TOPMODEL concept for representation of lower boundary conditions in GCMs by Koster *et al.* (2000).

TOPOG and THALES

TOPOG is founded on a contour-based terrain analysis method pioneered by O'Loughlin (1986) and Moore *et al.* (1988) at CSIRO. The approach was initially developed for the study of Australian catchments. The concept underlying the original idea of TOPOG is quite similar to the one of TOPMODEL and based on the local transmissivity to identify areas of the catchment that are subject to saturation and give raise to saturation excess runoff. The main difference between the initial concept of TOPOG and TOPMODEL is the way in which the landscape is dissected. While TOPMODEL is based on geometrical elements in the form of flow tubes and, more recently, pixels for the evaluation of the saturated areas, in TOPOG irregular elements are identified by breaking flow tubes in smaller interconnected and irregularly shaped elements by intersecting the lines of steepest descent with the contour lines. Figure 3 shows a typical example of the TOPOG model mesh. The elements are shadowed based on spatially distributed soil properties and surface slope classes. These terrain analysis algorithms are distributed through the stand-alone software package TAPES-C. The original TOPOG design has undergone substantial changes and expansion since its initial conception. The most recent versions distributed by CSIRO Land and Water have, with exception of the terrain analysis component, little in common with the original concepts. For example, a finite element-based groundwater module has been added next to a Richard's equation solver to study moisture dynamics in the unsaturated zone, yielding an entirely

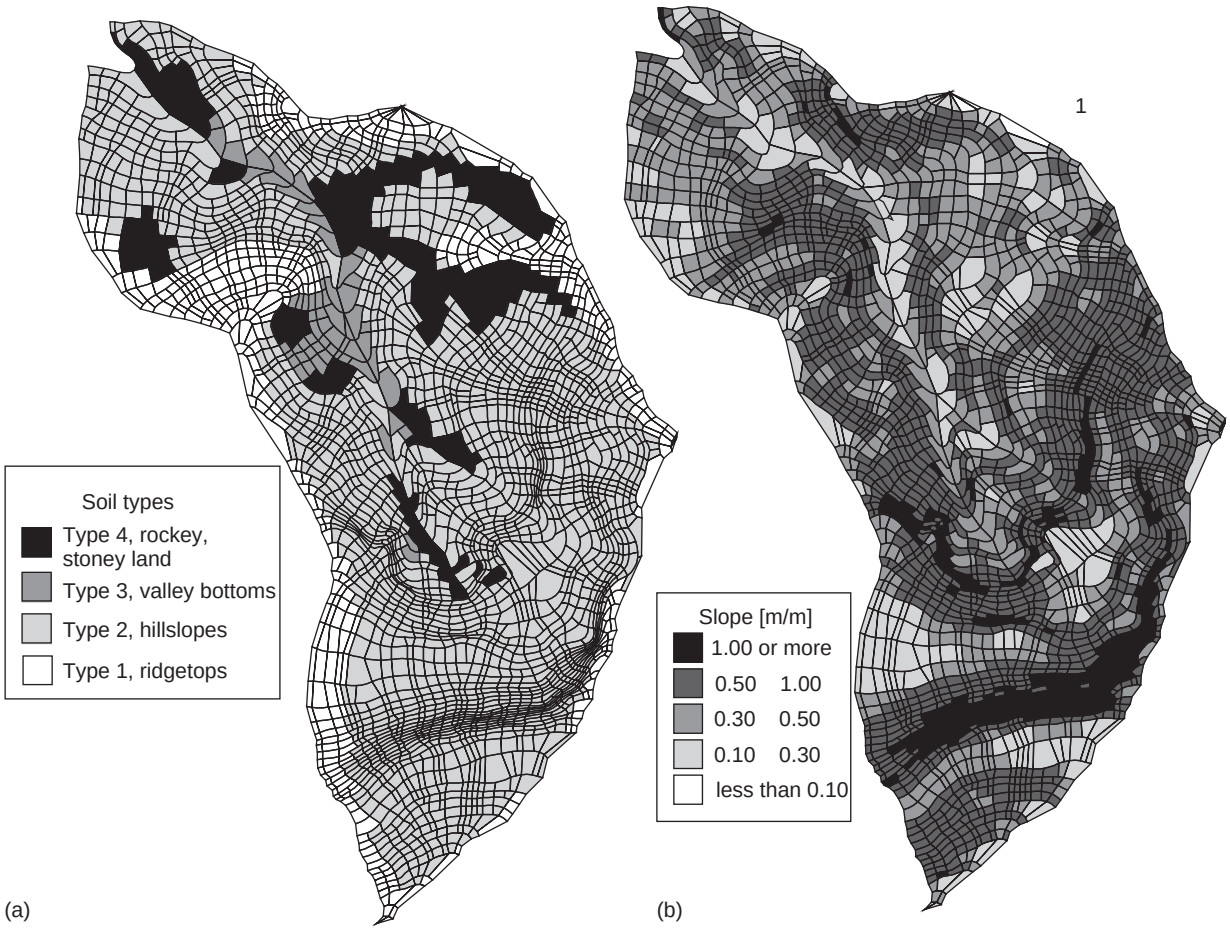


Figure 3 View of the irregularly shaped mesh of the TOPOG model obtained by dissecting the landscape on the basis of contour and flow lines. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

micro- or *macroscale*-based physical description. Explicit representation of transpiration for various plant species, including the carbon and the nutrient cycle (see e.g. Vertessy *et al.*, 1993) is also incorporated.

TOPOG has now been applied to many catchments outside of Australia, including a small Amazonian catchment (Vertessy and Elsenbeer, 1999) and, in an adapted version, to a small semiarid watershed in China by Zhu *et al.* (1999). Schellekens and Bruijnzeel (2004) used TOPOG in a small six-ha catchment in Puerto Rico to model runoff response in combination with extensive hydrometric and tracer validation of the runoff sources in the catchment (see also Schellekens *et al.*, 2004).

The TAPES-C routine is also used in another Australian-distributed modeling tool akin to TOPOG, called *THALES*. The developers of THALES use predominantly analytical steady-state solutions for the balance laws at the *micro* or the *macroscale* for estimation of infiltration, saturation excess, and subsurface runoff within model elements. References on THALES can be found in Grayson *et al.* (1995).

Other Distributed Models

Beyond the instruments mentioned here, other distributed modeling tools such as the TOPKAPI model (Liu and Todini, 2002) can be found in the literature. TOPKAPI integrates the point-scale conservation equations for mass and momentum over individual pixels of digital elevation maps (DEM). Restrictive assumptions reduce the governing equations to kinematic waves for subsurface, overland, and channel flow, for which analytical solutions are proposed.

Giving more detailed descriptions or listing even more models would go beyond the scope of this contribution. The reader is referred further to the books by Singh (1995) and the more recent publications by Singh and Frevert (2002a, b) for a more complete overview. Finally, it is relevant to observe that recent (Palacios-Vélez, 1998) and ongoing (Vivoni *et al.*, 2004) research is focussing on the use of triangular irregular networks (TIN) for discretizing a landscape into smaller hydrologic units. The discretization method is based on Delaunay triangulation and was proposed first in a hydrologic context by Palacios-Vélez and Cuevas-Renaud (1986). The TIN-based distributed modeling tool TRIBS currently developed at MIT seems to be promising in terms of the efficiency attributable to the method of spatial subdivision of the landscape. However, the relevant research is still at an early stage in order to allow drawing further conclusions.

THE REW APPROACH

As an alternative to the distributed approaches listed above, Reggiani *et al.* (1998, 1999, 2000) have introduced recently a distributed hydrological modeling concept formulated directly at the megascale. The motivation to undertake this research was the lack of a consistent formulation at a spatial scale larger than the macroscale, which would differ from the Freeze–Harlan (1969) blueprint by not requiring the solution of aggregated sets of PDE equations.

The authors of the REW approach went back to the conservation equations for mass (1), momentum (2), and energy (3) at the *microscale* and mapped them to the *megascale* by integration over appropriately chosen averaging regions, called *REWs*. It is important to note that in doing so, one of the main objectives of the REW approach is to provide a general framework, which would also be able to accommodate the wide range of temporal scales encountered in hydrological systems. In this way, the approach becomes suitable for studying instantaneous hydrological problems as well as long-term average behavior and yield (see Section "Introduction"), for which fluctuations at smaller timescales are less relevant. Consequently, the equations have also been averaged in time over a characteristic interval that is problem-specific and dictated by the particular application. A year could be such a typical averaging period.

The averaging regions or REWs are defined in such a way as to allow the definition to be generally applicable and scale-independent. Three-dimensional spatial regions were chosen that are determined by subcatchments of a particular size and that are delimited laterally by a prismatic mantle surface. They can be seen as extracted with a cookie cutter from the landscape.

A concrete example of a REW is depicted in Figure 4. In practice, a REW can be extracted from a digital terrain map with the threshold area method (Band, 1986) or by means of similar criteria (Tarboton, 1997) by imposing the first or higher Horton–Strahlerstream order as threshold for identifying a given number of subcatchments.

A REW communicates with neighboring REWs across the prismatic mantle surface via lateral exchange of groundwater. On the surface, the REWs are interconnected through the stream network that drains surface water towards a common catchment outlet. This mutual interconnection of REWs for a generic watershed structure is visualized in Figure 5. The definition of the REW allows a scale-independent subdivision of a watershed into irregular modeling elements, which are not tied to a square grid, but allow the natural footprint of the landscape to be preserved in the discretization.

The REW could be seen as a kind of Hydrologic Response Unit (HRU), used frequently as a concept in the discretization of hydrological systems. HRUs are based on superimposition of particular characteristics in the landscape such as topography, aspect, soil properties, or land

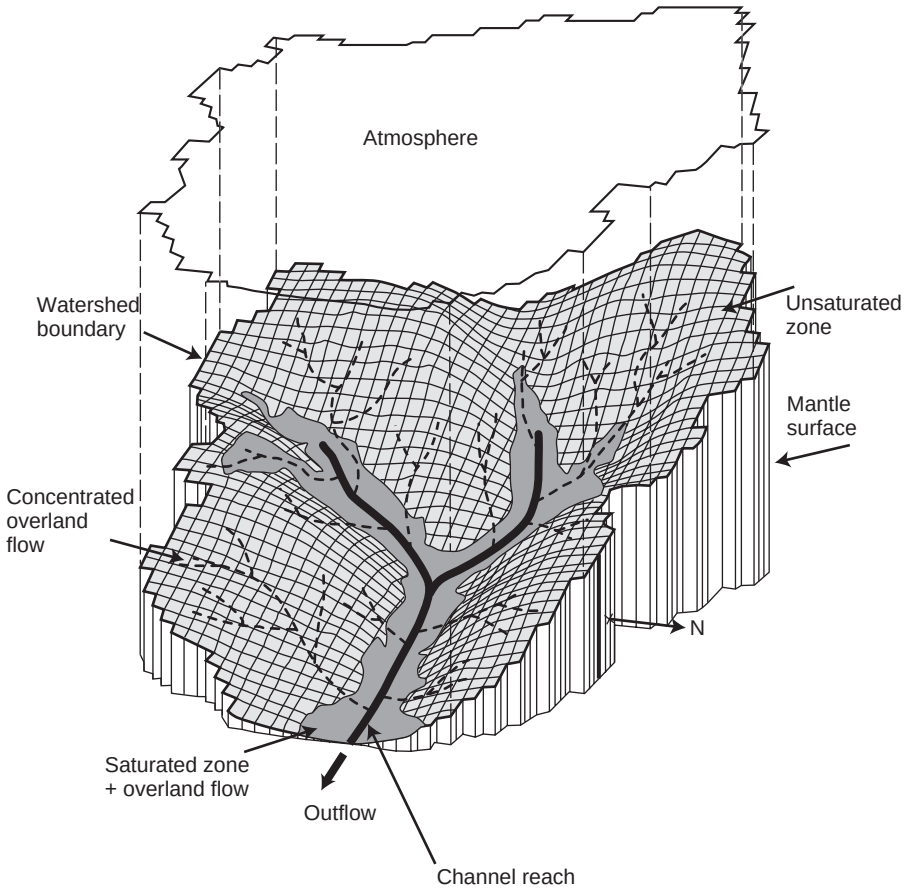


Figure 4 View of a Representative Elementary Watershed (REW) as a 3-D spatial entity

use with the aim of identifying spatial regions that can be considered compact hydrological entities for which the mass balance can be applied. The definition of REW is based on scalable spatial entities, whose reoccurring structure can be recognized at scales that range form a watershed of a few hectares to much larger entities, which may measure several tens of square kilometers. These entities represent hydrologically meaningful volumes over which the conservation equations for mass, momentum, and energy are integrated. Given a chosen spatial scale, suitable closure expressions for the respective REW-scale fluxes need to be parameterized. The parameterized expressions depend on the size of the entity and the degree of heterogeneity of the hydrological entity to be modeled.

The point-scale equations (1), (2), and (3) are integrated over particular zones of the REW that are identified on the basis of their hydrological role (i.e. porous medium or just water) and their hydrodynamic behavior (i.e. unsaturated porous media flow, channel flow). In the case of the saturated zone, the entire saturated soil within a REW is considered as integration volume; in the same way as the unsaturated zone, the overland flow volume or the channel volume constitutes separate integration volumes.

The resulting *megascale* governing equations are obtained without making any *a-priori* assumptions and represent scale-independent ordinary differential equations (ODEs), which conserve physical properties for a REW in terms of spatially and temporally averaged variables.

Equations (1), (2), and (3) can be restated more generically in terms of a variable ψ, a non-convective interaction $\vec{i}$, and a rate of external supply for ψ, f, which are placeholders for the properties summarized in Table 1. Integration of the microscopic conservation equations over a phase volume V and a characteristic time interval $(t - \Delta t, t + \Delta t)$ yields the following generic *megascale* (or REW-scale) conservation law, from which specific conservation equations for mass, momentum, and energy can be deduced:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{t-\Delta t}^{t+\Delta t}\int_{V}(\rho\psi)\,\mathrm{d}V\,\mathrm{d}\tau + \sum_{j}\int_{t-\Delta t}^{t+\Delta t}\int_{A_j} [\rho\psi(\vec{v}-\vec{w}) - \vec{i}]\cdot\vec{n}\,\mathrm{d}A\,\mathrm{d}\tau - \int_{t-\Delta t}^{t+\Delta t}\int_{V}\rho f\,\mathrm{d}V\,\mathrm{d}\tau = 0 \tag{5}$$

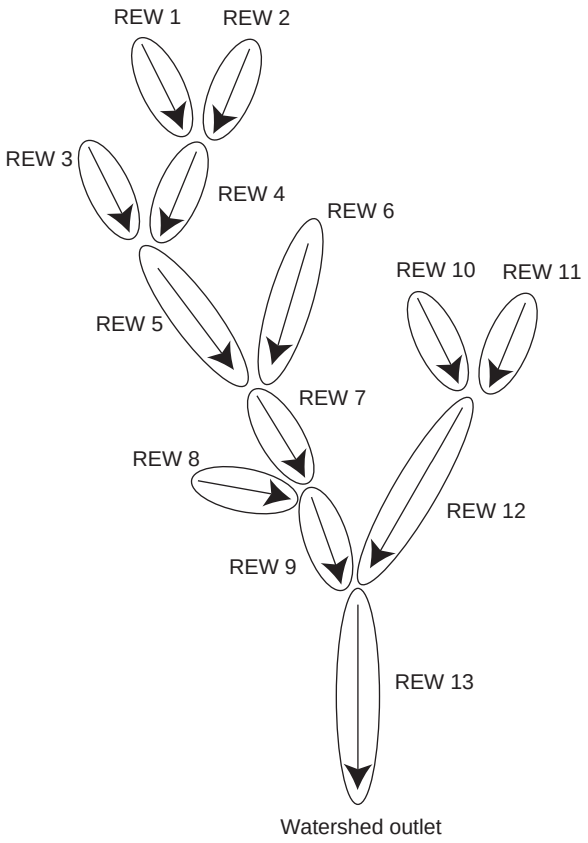


Figure 5 Aggregation of 13 REWs into a binary tree structure for representation of a catchment

Table 1 Microscopic properties in balance laws

	ψ	$\vec{i}$	f
Mass	*1*	*0*	*0*
Momentum	$\vec{v}$	$\vec{T}$	$\vec{g}$
Energy	$E + v^2/2$	$\vec{q}$	h

In (5), $\sum_j A_j$ is the sum of boundary surface segments delimiting the volume V and $\vec{w}$ the surface velocity. The second integral on the left-hand side represents the sum of fluxes of the various properties across the boundary surface segments delimiting the control volume. These flux terms constitute unknown exchange terms for mass, force or energy between phases and system compartments that result from the integration procedure. Such terms are frequently encountered in the mathematical description of environmental systems. Their expressions can be evaluated in closed form only in rare cases, for which the geometry and the velocity of the fluid and the interfaces are uniform and known. The field of fluid mechanics is well acquainted with handling such unknown flux terms (e.g. see Stull, 1988 for applications in ABL meteorology).

THE CLOSURE PROBLEM AND HYDROLOGY

The parameterization of unknown flux terms in space and time is challenging in practice. This is a well recognized problem in fluid mechanics, where perturbation and subsequent averaging in time of the governing equations, under the assumption of statistically stationary and homogeneous turbulence, gives rise to correlations of small-scale fluctuations of the independent variables known as *eddy correlations*. These terms are responsible for transfer of significant quantities of mass, momentum, and energy within the system and are measurable with relative ease at a point (e.g. eddy fluctuation sampling from a measurement tower for ABL experiments).

For representation in a numerical model, the correlation terms, which constitute excess unknowns with respect to the available number of prognostic equations, need to be expressed via constitutive relationships as functions of the independent variables. The independent variables are usually the (computation cell) mean velocity, the temperature or saturation, calculated from the solution of the prognostic equations at the nodal points of the model grid. The constitutive relationships are collectively referred to as *closure* schemes. A typical example for such a closure is the $K - \varepsilon$ scheme (for reference, see e.g. Stull, 1988), relating the eddy momentum diffusivity to the turbulent kinetic energy, for which there is a prognostic equation.

Hydrologists encounter similar problems, but do not approach them in the same systematic fashion as people in fluid mechanics, the reason being that hydrologists are confronted with a much higher degree of heterogeneity in their systems and with processes characterized by a wide range of temporal scales. *Closure* in hydrology means mainly parameterization of mass and energy fluxes exchanged between different system compartments and between land surface and canopy and the ABL. The closure is not achieved systematically, through closure schemes, but rather heuristically via *ad hoc* parameterizations, for which suitable parameter values are tuned by calibration. A rare example, where closure of mass fluxes between the saturated and the unsaturated store in a hillslope is exposed as a problem *per se* and approached directly, can be found in an article by Duffy (1996). The infinite choice of possible parameter sets in reproducing the response of a hydrological system (and thus its internal mass fluxes) for a given model architecture opens a window on the difficulties and arbitrariness hydrologists are facing already when attempting to close just the mass balance (e.g. Beven and Binley, 1992; Beven, 2002a).

In fluid or continuum dynamics, the geometry, the fluxes or stresses, and the material properties are generally well known, or can be estimated with sufficient accuracy thanks to a reasonable level of homogeneity in the system. This

is rarely the case in hydrology where the flow processes occurring in highly heterogeneous soils are difficult to predict and generally do not allow to translate parameter values, measured and validated at the laboratory scale or *in situ* to larger spatial scales used in the models.

Flux profiling for the determination of eddy diffusivity coefficients in a well-mixed or uniformly stratified water body or a homogeneous clay layer, for which a Brooks–Corey water retention curve can be fitted, are examples of privileged situations, for which analogues are rarely encountered in a hydrological context. Hydraulic conductivity values measured through an infiltration experiment in the field or in the laboratory cannot be extrapolated simply to larger spatial scales, as the processes at those scales have good chances not to obey to the laws governing their behavior at smaller scales (see e.g. Beven, 1993). This uncertainty on process representation underlying hydrological models is amply discussed throughout works by Beven (e.g. 1993, 2002a, 2002b) and needs to be born in mind when pursuing distributed hydrological modeling, whatever the approach.

Finally, it needs to be emphasized that the closure of fluxes in a hydrological context precludes, due to the inherent heterogeneity of the landscape and soils, the arbitrary scaling of parameterized closure expressions form a measurement point to the model grid scale, as it is common practice in fluid mechanics. These need to be adapted to unique hydrologic situations and to the chosen spatial discretization. The sensitivity of model behavior with respect to given parameterizations will remain the principal criterion for accepting or rejecting a model based on a specific set of flux parameterizations and a topic for further research.

DISCUSSION

What can be thus said about traditional distributed hydrological models based on the FH69 blueprint *vis à vis* a megascopic flux-based formulation such as the REW approach? In general, both formulations, one based on a mix of micro, macro, and megascale representation and the other one exclusively on megascale physics, are consistent with established physical principles in the form of stated assumptions and balance laws (Beven, 2002a). However, to be truly physically based, a model based on either theory needs to satisfy the additional condition of consistency with available hydrological observation at the spatial scale at which the governing equations are formulated. It has been shown exhaustively during the last 30 years in various simulation experiments for data-rich experimental watersheds (e.g. the R5 watershed amongst others, Woolhiser, 1996) that this is not the case for FH69-based models.

It has become clear that, notwithstanding high-quality data from dense logging networks (including information on soil texture and structure, piezometer levels, weirs, surface runoff rates, infiltration rates, and precipitation), achieving improvements of the results beyond a certain point is incommensurate with the efforts to be invested. In other words, it is like trying to squeeze a couple of horse powers more out of an engine that is already operating at its design limits. Achieving added model skill is an expectedly asymptotic effort in terms of the increased spatial resolution of input data and the computational resources required for a successively improved model run, with the hypothetical end-of-the-line being the solution of the full microscale Navier–Stokes equations. This fact could eventually lead to the FH69 blueprint being abandoned like the search of the Spanish conquistadores for "El Dorado" (Woolhiser, 1996) under pressure of having to go back to the drawing board.

In contrast, the *megascale* REW or other control volume–based approaches (Duffy, 1996) are based on a formulation, where a change in spatial scale from the *micro*- to the *megascale* transforms the gradients in the PDE equations into fluxes across the boundary of the control volumes (second left-hand side terms in equation (5)). A successful application lies in representing these fluxes and the hysteretic behavior of the control volume in response to changes in boundary fluxes in the best way possible, whereby the already emphasized difficulties in the necessary parameterizations remain to be addressed. But at least it may point into a viable direction for a way-out from the modeling dilemma experienced in the context of FH69.

The hypothetical measurement at the REW-scale and the correct parameterization of mass or energy fluxes across large surfaces identified between zones in a REW, such as the water table and the land surface or the REW mantle segments, pose the principal challenges in applying this approach and giving it a status of a blueprint for a physically based solution at the *megascale* in its own right (see Beven, 2002a for a discussion of the issues in formulating an alternative blueprint).

However, the flux-based philosophy of the formulation rests on scale-independent and generally valid conservation equations and should draw the attention of the hydrologist towards improving constitutive relationships for fluxes, as currently done in other disciplines, rather than towards reducing mesh sizes, improving numerical algorithms, or attempting to employ data at denser spatial resolution in line with the FH69 philosophy. It needs to be said, however, that improving the megascale representation of hydrological fluxes will remain a significant scientific challenge that needs to be tackled systematically from both the theoretical and the practical side. Some fluxes, such as evaporation, may eventually become measurable thanks to improved technology. Other fluxes, particularly in the subsurface, perhaps never will become measurable, and, thus, need to be estimated indirectly via inverse techniques based on known megascale fluxes or internal state variables. Finally,

it is important to note that the uncertainty intrinsic to the estimation of these quantities will need to be taken into account and dealt with systematically.

GIS, DATABASES, AND DEM DATA IN DISTRIBUTED MODELING

A discussion on distributed hydrological modeling must be accompanied by a note on Geographical Information Systems (GIS) and Databases (DB), tools that play an increasingly important role in the implementation and operation of distributed models. All distributed models require input data of different kind that extend from DEM to soil properties, digital land-use charts, infrastructure typology maps, and a variety of hydraulic properties. In addition, models need to be initialized by assigning initial conditions, need to be driven by spatially distributed environmental forcing, and require the imposition and update of spatially distributed boundary conditions. During run-time, data assimilation procedures can be applied (see e.g. Paniconi *et al.*, 2003). After termination of the model runs, large amounts of data distributed in space and time need to be processed. The handling of this substantial amount of information is facilitated through GIS that allow storing, converting and analyzing spatial data sets arranged in vector or raster formats.

At the moment, world-wide coverage of hydrologically sound digital terrain data is available at a 1×1 km resolution (the HYDRO1k Elevation Derivative Database is accessible at **`http://edcdaac.usgs.gov/gtopo30/hydro/`**). At the same time, more detailed data is freely available at 30×30 m resolution from the ASTER sensor on board of the EO-1 satellite. To date, nobody has compiled a hydrologically sound world-wide coverage from this data source, which remains, however, a significant undertaking. In addition, the IFSAR programme will use the Shuttle Radar Topography Mission (SRTM) data to construct a 30×30 m resolution DEM of high quality (accessible at: **`http://www.fas.org/irp/program/collect/ifsar.htm`**). It is foreseen that this data set will be the primary focus for building DEM data for hydrological modeling in the coming years. This fact will stimulate modelers to reduce the resolution of distributed models and to drive them with an increasingly large amount of information. In this context, GIS are becoming progressively an essential component of hydrological models.

Similar arguments are valid for DB tools. The access and storage of long time-series of historical or simulated discharge, temperature, water level or water quality data are facilitated significantly with the use of third-hand party information tools that are optimized in the handling of large data sets (at the date of publication in the order of Gigabytes). For this reason, distributed hydrological models are required to interface routinely with such instruments.

Alternatively, entire distributed models can be run from a GIS environment, whereby the GIS is not just used as a simple user interface. This is the case for programmable GIS, in which "traditional" programming statements can be applied to entire maps (a drainage networks). One such system is PCraster (Van Deursen, 1995) that has been used to build a variety of hydrological models for large catchments, such as RHINEFLOW (Van Deursen and Kwadijk, 1993) and LISFLOOD (De Roo *et al.*, 2000). Although running a distributed hydrological model through a GIS can have important advantages such as easy programming, and raster data integration from various sources, it remains essentially based on conventional process formulations and respective parameterization. It provides, however, a user-friendly platform for building, running, and presenting a variety of models relatively quickly.

Finally, it must be said that the rapid growth in quantity and the reduction in cost of data and relevant information will not provide an answer to the basic scientific questions underlying the use of data in a hydrological context discussed above. Data are recorded by sensors, which capture electromagnetic signals and convert them into digital ones. To digital signals, numerical values are assigned. These signals can be distorted at the source through a series of factors that decrease the quality of a scan. A good example is the presence of snow or topography interfering with an active radar signal for precipitation-intensity measurements or distortion induced by the atmosphere for altitude measurements of a topographic relief.

Measured quantities employed in hydrological models, such as soil moisture content, water levels, or rainfall intensity, are attributed to the recorded digital signal with the aid of *interpretative* models. This procedure leaves a large space for erroneous interpretation. Once more, the use of radars to predict the intensity of precipitation fields is a typical example for a situation, where significant effort over the last decades has still not led to a robust and reliable calibration procedure. The same can be said for remote sensing of land use and airborne soil moisture measurements, although considerable progress has been made in recent years in the quantitative use of passive microwave in space-borne soil moisture determination (see e.g. Owe *et al.*, 2001; De Jeu and Owe, 2003). However, it is to be expected that these measurements will only deliver surface soil moisture affected by uncertainty that provides limited information on change of moisture storage over the soil depth profile at the model grid scale.

In the near future, the availability of remotely sensed data from which hydrologically relevant information (e.g. soil moisture, land-cover features, leaf-area index) can be derived via interpretation models is bound to grow. The risk that the information will be misused for modeling purposes due to too large a trust in what is promised by suppliers and too much emphasis on the aspects related to

data processing will grow proportionally. As a general rule, common sense and a critical attitude from the side of the hydrologist will never go astray, however sophisticated the data logging instruments and however affordable the data might become.

CONCLUSIONS

In the late 1960s and during the 1970s, the rapid development of computer power led hydrologists and environmental modelers to believe that it was only a matter of time, until a complete distributed catchment model, based on the solution of PDEs, would become available. It was envisaged that, in line with growing CPU speed, improved numerical algorithms, falling hardware prices, and more sophisticated measurement techniques, these models would improve and allow detailed simulations of hydrological systems. This belief was justified at the time, in particular in view of the rapid progress made in other areas, such as in the fluid mechanics of oceanography and meteorology, and in continuum mechanics, where these developments have, in fact, led to impressive predictive skills of the models.

In hydrology, however, more than a quarter of a century later, this initial optimism has been visibly dampened. Repeated applications on heavily monitored catchments at higher spatial resolution and with ever more powerful computers than the one available earlier, have shown that, despite being able to solve the coupled equations numerically, other problems started to appear. Hydrologists using models based on the Freeze–Harlan blueprint could not share the enthusiasm of their colleagues in the fields of fluid dynamics or continuum mechanics.

The reasons are manifold. The processes represented at the micro or macroscale require parameterization of subgrid variability (the variability of the system at scales smaller than the scale of the numerical discretization grid), an extremely challenging if not impossible task for highly heterogeneous systems such as natural soils or a vegetated land surface. Lumping of highly nonlinear processes into parameter representations at larger scales is conceivable in theory in analogy to more homogeneous systems, but has turned out to be unsatisfactory for applications in hydrological practice.

The difficulty underlying the parameterization, in addition to the rather large number of individual processes to be represented, result in a problem of *equifinality* of models and parameter sets that are consistent with the observed data available (e.g. Beven, 1993, 2005; Beven and Freer, 2001). Equifinality leads to various sets of parameter values, which might give a good calibration result, but are often far apart from the values that are observed in the field, raising the question about how *physical* the description actually is. This fact will continue to characterize distributed hydrological modeling and will be exacerbated even further by increasing the spatial resolution of the models.

The *REW* approach offers an alternative by proposing a flux-based method that makes use of appropriately chosen control volumes. The *megascopic* nature of the approach is oriented towards specifying the fluxes across the volume boundaries, rather than towards trying to increase the detail of the process description within the system. In addition, fluxes, such as evaporation, might also have a greater chance of being measurable over large areas, thus aiding the closure procedure. A higher level of detail in the REW approach can be potentially obtained through a breakdown of the areas, for which the fluxes are estimated. In the interest of handling model uncertainty systematically in hydrological modeling *sensu* Beven (2002a), the REW approach could, through an appropriate choice of alternative closure schemes and respective parameterizations, serve as platform for the construction of a multidimensional model and parameters space for the simulation of hydrological systems. However, the aspect of the uncertainty inherent to any attempt of physical description will persist and remain a fact to be catered for systematically in hydrological studies based on megascopic modeling concepts.

REFERENCES

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986a) An introduction to the European hydrologic system – système hydrologique Européen, "SHE" 1: history and philosophy of a physically based, distributed modelling system. *Journal of Hydrology*, **87**, 45–59.

Abbott M.B., Bathurst J.C., Cunge J.A. and O'Connell P.E. (1986b) An introduction to the European hydrologic system – système hydrologique Européen, "SHE" 2: structure of a physically based, distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

Band L.E. (1986) Topographic partition of watersheds with digital elevation models. *Water Resources Research*, **22**(1), 15–24.

Barré de Saint-Venant A.J.C. (1871) Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lits. *Comptes rendus des séances de l'Académie des Sciences*, **73**, 147–154 237–240.

Bathurst J.C., Ewen J., Parkin G., O'Connell P.E. and Cooper J.D. (2004) Validation of catchment models for predicting land-use and climate change impacts. 3. Blind validation for internal and outlet responses. *Journal of Hydrology*, **287**, 74–94.

Bear J. (1979) *Hydraulics of Groundwater*, McGraw-Hill: New York.

Bergström S. (1995) The HBV model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Park, pp. 443–476.

Beven K.J. (1989) Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (1993) Prophecy, reality and uncertainty in distributed hydrological modeling. *Advances in Water Resources*, **16**, 41–51.

Beven K.J. (1996a) A discussion on distributed modelling. In *Distributed Hydrological Modelling*, Chap. 13A, Refsgaard C.J. and Abbott M.B. (Eds.), Kluwer: Dodrecht, pp. 255–278.

Beven K.J. (1996b) Response to comment on "A discussion on distributed modelling" by C.J. Refsgaard *et al.*. In *Distributed Hydrological Modelling*, Chap. 13C, Refsgaard C.J. and Abbott M.B. (Eds.), Kluwer: Dodrecht, pp. 255–278.

Beven K.J. (2001) How far can we go in distributed hydrological modelling? (Dalton Lecture). *Hydrology and Earth System Sciences*, **5**(1), 1–12.

Beven K.J. (2002a) Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological Processes*, **16**, 189–206.

Beven K.J. (2002b) Towards a coherent philosophy for modelling the environment. *Proceedings of the Royal Society of London. Series A*, **485**, 2465–2484.

Beven K.J. (2005) A manifesto for the equifinality thesis. *Journal of Hydrology*, (in press).

Beven K.J. and Binley A.M. (1992) The future of hydrological models: model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

Beven K.J., Calver A. and Morris E. (1987) *The Institute of Hydrology Distributed Model, U.K.*, Report No. 98, Institute of Hydrology.

Beven K.J. and Freer J. (2001) Equifinality, data assimilation and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, **249**, 11–29.

Beven K.J. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**(1), 43–69.

Birkinshaw S.J. and Ewen J. (2000) Modelling nitrate transport in the slapton wood catchment using SHETRAN. *Journal of Hydrology*, **230**, 18–33.

Bruijnzeel L.A. (1990) *Hydrology of Moist Tropical Forests and Effects of Conversion: A State-of Knowledge Review*, UNESCO International Hydrological Programme: Paris.

Bruijnzeel L.A. (2004) Hydrological functions of tropical forests, not seeing the soil for the trees? *Agriculture, Ecosystems and Environment*, **104**, 185–228.

Calver A. and Wood W.L. (1995) The institute of hydrology distributed watershed model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Park, pp. 595–626.

Christiansen J.S., Thorsen M., Clausen T., Hansen S. and Refsgaard J.C. (2004) Modelling of macropore flow and transport processes at catchment scale. *Journal of Hydrology*, **299**, 136–158.

Crawford N.H. and Linsley R.S. (1966) *Digital Simulation in Hydrology: The Stanford Watershed Model IV*, Technical Report No. 39, Department of Civil Engineering, Stanford University Press, Palo Alto.

Darcy H. (1856) *Les Fontaines Publiques de la Ville de Dijon*, Dalmont.

De Jeu R.A.M. and Owe M. (2003) Further validation of a new methodology for surface moisture and vegetation optical depth retrieval. *International Journal of Remote Sensing*, **24**, 4559–4578.

De Roo A.P.J., Wesseling C.G. and Van Deursen W.P.A. (2000) Physically-based river basin modelling within a GIS: the LISFLOOD model. *Hydrological Processes*, **14**, 1981–1992.

Duffy C.J. (1996) A two-state integral-balance model for soil moisture and groundwater dynamics in complex terrain. *Water Resources Research*, **32**, 2421–2434.

Eagleson P.S. (1978a) Climate soil and vegetation, 1, introduction to water balance dynamics. *Water Resources Research*, **14**(5), 705–712.

Eagleson P.S. (1978b) Climate soil and vegetation, 6, dynamics of the annual water balance. *Water Resources Research*, **14**(5), 749–764.

Eagleson P.S. (1982) Ecological optimality in water-limited natural soil-vegetation systems, 1, theory and hypothesis. *Water Resources Research*, **18**(2), 325–340.

Eagleson P.S. and Tellers T.E. (1982) Ecological optimality in water-limited natural soil-vegetation systems, 2, tests and applications. *Water Resources Research*, **18**(2), 341–354.

Freeze R.A. and Harlan R.L. (1969) Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of Hydrology*, **9**, 237–258.

Freeze R.A. (1971) Three-dimensional, transient, saturated-unsaturated flow in a groundwater basin. *Water Resources Research*, **7**, 347–366.

Freeze R.A. (1972a) Role of subsurface flow in generating surface runoff: 1. Base flow contributions to channel flow. *Water Resources Research*, **8**, 609–623.

Freeze R.A. (1972b) Role of subsurface flow in generating surface runoff: 2. Upstream source areas. *Water Resources Research*, **8**, 1272–1283.

Freeze R.A. (1978) Mathematical models of hillslope hydrology. In *Hillslope Hydrology*, Kirkby M.J. (Ed.), John Wiley & Sons: Chichester, pp. 177–225.

Gray W.G., Leijnse A., Kolar R.L. and Blain C.A. (1993) *Mathematical Tools for Changing Spatial Scales in the Analysis of Physical Systems*, CRC Press: Boca Raton.

Grayson R.B., Moore I.D. and McMahon T.A. (1992) Physically-based hydrological modelling 2. Is the concept realistic? *Water Resources Research*, **26**(10), 2659–2666.

Grayson R.B., Blöschl G. and Moore I.D. (1995) Distributed parameter hydrologic modelling using vector elevation data: THALES and TAPES-C. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Park, pp. 595–626.

Henderson F.M. and Wooding R.A. (1964) Overland flow and groundwater flow from a steady rainfall of finite duration. *Journal of Geophysical Research*, **69**(8), 1531–1540.

Hassanizadeh S.M. and Gray W.G. (1979a) General conservation equations for multiphase systems: 1. Averaging procedure. *Advances in Water Resources*, **2**, 131–144.

Hassanizadeh S.M. and Gray W.G. (1979b) General conservation equations for multiphase systems: mass, momenta, energy and entropy equations. *Advances in Water Resources*, **2**, 191–203.

Hassanizadeh S.M. and Gray W.G. (1980) General conservation equations for multiphase systems: 3. constitutive relationships. *Advances in Water Resources*, **2**, 25–40.

Hornberger G.M., Bencala K.E. and McKnight D.M. (1994) Hydrological controls on dissolved organic carbon during

snowmelt in the Snake river near Montezuma, Colorado. *Biogeochemistry*, **25**, 147–165.

IPCC Third Assessment Report (2001) *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J. and Xiaosu D. (Eds.) Cambridge University Press, UK. pp. 944.

Koster R.D., Suarez M.J., Ducarne A., Stieglitz M. and Kumar P. (2000) A catchment-based approach to modeling land surface processes in a general circulation model: 1. model structure. *Journal of Geophysical Research*, **105, 20**, 809–822.

Lighthill M.J. and Whitham G.B. (1955) On kinematic waves I: flood movements in long rivers. *Proceedings of the Royal Society of London. Series A*, **229**, 281–316.

Liu Z. and Todini E. (2002) Towards a comprehensive physically-based rainfall-runoff model. *Hydrology and Earth System Sciences*, **6**(5), 859–881.

Loague K. and Freeze R.A. (1985) A comparison of rainfall-runoff modeling techniques on small upland catchments. *Water Resources Research*, **21**, 229–248.

Loague K., Gander G.A., VanderKwaak J.E., Abrams R.H. and Kyriakidis P.C. (2000) Simulating hydrologic response for the R-5 catchment: a never ending story. *Floodplain Management*, **1**, 57–83.

Loague K. and Kyriakidis P.C. (1997) Spatial and temporal variability in the R-5 infiltration data set: Déjà vu and rainfall-runoff simulations. *Water Resource Research*, **33**, 2883–2895, Special issue on Scale Problems in Hydrology.

Loague K. and VanderKwaak J.E. (2002) Simulating hydrologic response for the R-5 catchment: comparison of two models and the impact of the roads. *Hydrological Processes*, **16**, 1015–1032.

Loague K. and VanderKwaak J.E. (2004) Physics-based hydrologic response simulation: platinum bridge, 1958 Edsel, or useful tool. *Hydrological Processes*, **18**, 2949–2956.

Moore I.D., O' Loughlin E.M. and Burch G.J. (1988) A contour-based topographic model for hydrological, geomorphological and ecological applications. *Earth Surface Processes and Landforms*, **13**, 305–320.

O'Loughlin E.M. (1986) Prediction of surface saturation zones in natural catchments by topographic analysis. *Water Resources Research*, **22**(5), 794–804.

Owe M., De Jeu R.A.M. and Walker J.P. (2001) A Methodology for surface soil moisture and vegetation optical depth retrieval using the microwave polarization difference index. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1643–1654.

Palacios-Vélez O. and Cuevas-Renaud B. (1986) Automated river course, ridge and basin delineation from digital elevation data. *Journal of Hydrology*, **86**, 299–314.

Palacios-Vélez O., Gandoy-Bernasconi W. and Cuevas-Renaud B. (1998) Geometric analysis of surface runoff and the computation order of unit elements in distributed hydrological models. *Journal of Hydrology*, **211**, 266–274.

Paniconi C. and Wood E. (1993) A detailed model for simulation of catchment scale subsurface hydrologic processes. *Water Resources Research*, **29**(6), 1601–1620.

Paniconi C., Marrocu M., Putti M. and Verbunt M. (2003) Newtonian nudging for a Richards equation-based distributed hydrological model. *Advances in Water Resources*, **26**(2), 161–178.

Parkin G., O'Donnell G., Ewen J., Bathurst J.C., O'Connell P.E. and Lavabre J. (1996) Validation of catchment models for predicting land-use and climate change impacts: 2. case study for a Mediterranean catchment. *Journal of Hydrology*, **175**, 595–613.

Peck A.J. and Hurle D.A. (1973) Chloride balance of some farmed and forested catchments in Southwestern Australia. *Water Resources Research*, **9**(3), 648–657.

Refsgaard J.C. (1997) Parameterization, calibration and validation of distributed hydrological models. *Journal of Hydrology*, **198**, 69–97.

Refsgaard J.C. (2000) Towards a formal approach to calibration and validation of models using spatial data. In *Spatial Pattern in Catchment Hydrology: Observations and Modelling*, Grayson R. and Blöschl G. (Eds.), Cambridge University Press: pp. 329–354.

Refsgaard J.C. and Storm B. (1995) MIKE SHE. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Park, pp. 809–846.

Refsgaard J.C. and Storm B. (1996) Comment on "A discussion on distributed modelling" by K.J. Beven. In *Distributed Hydrological Modelling*, Chap. 13B, Refsgaard C.J. and Abbott M.B. (Eds.), Kluwer: Dodrecht, pp. 279–287.

Reggiani P., Sivapalan M. and Hassanizadeh S.M. (1998) A unifying framework of watershed thermodynamics: 1 balance equations for mass, momentum, energy and entropy and the second law of thermodynamics. *Advances in Water Resources*, **22**(4), 367–398.

Reggiani P., Sivapalan M. and Hassanizadeh S.M. (2000) Conservation equations governing hillslope responses. *Water Resources Research*, **38**(7), 1845–1863.

Reggiani P., Sivapalan M., Hassanizadeh S.M. and Gray W.G. (1999) A unifying framework of watershed thermodynamics: 2 constitutive relationships. *Advances in Water Resources*, **23**(1), 15–39.

Richards L.A. (1931) Capillary conduction of liquids through porous mediums. *Physics*, **1**, 318–333.

Robson A., Beven K.J. and Neal C. (1992) Towards identifying sources of subsurface flow: a comparison of components identified by a physically-based runoff model and those determined by chemical mixing techniques. *Hydrological Processes*, **6**, 199–214.

Salvucci G.D. and Entekhabi D. (1995) Hillslope and climatic controls on hydrologic fluxes. *Water Resources Research*, **31**(7), 1725–1739.

Saulnier G.M., Obled C.h and Beven K. (1998) Including spatially variable effective soil depths in TOPMODEL". *Journal of Hydrology*, **202**, 158–172.

Schellekens J. and Bruijnzeel L.A. (2004) Modelling water yield and runoff response of a small tropical rain forest catchment using a physically-based distributed model. *Hydrological Processes*, (in press).

Schellekens J., Scatena F.N., Bruijnzeel L.A., Van Dijk A.I.J.M., Groen M.M.A. and van Hoogezand R.J.P. (2004) Stormflow generation in a small rainforest catchment in the luquillo

experimental forest, Puerto Rico. *Hydrological Processes*, **18**, 505–530.

Singh V.P. (1995) *Computer Models of Watershed Hydrology*, Water Resources Publications: Highlands Park.

Singh V.P. and Frevert D.K. (2002a) *Mathematical Models of Large Watershed Hydrology*, Water Resources Publications: Highlands Park.

Singh V.P. and Frevert D.K. (2002b) *Mathematical Models of Small Watershed Hydrology*, Water Resources Publications: Highlands Park.

Sivapalan M., Beven K.J. and Wood E.F. (1987) On hydrologic similarity: 2. A scaled model of storm runoff production. *Water Resource Research*, **23**(12), 2266–2278.

Stull R.B. (1988) *An Introduction to Boundary Layer Meteorology*, Kluwer Academic Publishers: Dordrecht.

Tarboton D.G. (1997) A new method for the determination of flow directions and contributing areas in grid digital elevation models. *Water Resources Research*, **33**(2), 309–319.

VanderKwaak J.E. and Loague K. (2001) Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model. *Water Resources Research*, **37**(5), 999–1013.

Van Deursen W.P.A. (1995) *Geographical Information Systems and Dynamic Models: Development and Application of a Prototype Spatial Modelling Language*, Ph.D. Thesis, Utrecht University, NGS 190.

Van Deursen W.P.A. and Kwadijk J.C.J. (1993) RHINEFLOW: an integrated GIS water balance model for the river Rhine. Applications of geographic information systems. In *Hydrology and Water Resources Management, HydroGIS 1993*, Kovar K. and Nachtnebel H.P. (Eds.), IAHS Publication No. 211, IAHS, pp. 507–518.

Vertessy R.A. and Elsenbeer H. (1999) Distributed modelling of storm flow generation in an Amazonian rainforest catchment: effects of model parameterization. *Water Resources Research*, **35**(7), 2173–2187.

Vertessy R.A., Hatton T.J., O'Shaughnessy P.J. and Layasuria M.D.A. (1993) Predicting water yield from a mountain ash forest using a terrain analysis based catchment model. *Journal of Hydrology*, **150**, 665–700.

Viney N.R. and Sivapalan M. (1999) A conceptual model of sediment transport: application to the Avon river basin in Western Australia. *Hydrological Processes*, **13**, 727–743.

Viney N.R. and Sivapalan M. (2001) Modelling catchment processes in the Swan-Avon river basin. *Hydrological Processes*, **15**, 2671–2685.

Vivoni E.R., Ivanov V.Y., Bras R.L. and Entekhabi D. (2004) Generation of triangulated irregular networks based on hydrological similarity. *Journal of Hydrologic Engineering*, **9**(4), 288–302.

Woods R., Sivapalan M. and Robinson J. (1997) Modelling the spatial variability of subsurface runoff using a topographic index. *Water Resources Research*, **33**(5), 1061–1073.

Woolhiser D.A. (1996) Search for physically-based runoff model – a hydrologic El Dorado. *Journal of Hydraulic Engineering ASCE*, **122**, 122–129.

Zhao R.J. (1977) *Flood Forecasting Method for Humid Regions of China*, East China College of Hydraulic Engineering: Nanjing.

Zhao R.J. and Liu X.R. (1995) The Xinanjiang model. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications: Highlands Park, pp. 809–846.

Zhu T.X., Band L.E. and Vertessy R.A. (1999) Continuous modeling of intermittent stormflows on a semi-arid agricultural catchment. *Journal of Hydrology*, **226**, 11–29.

128: Rainfall-runoff modeling: Transfer Function Models

PETER C YOUNG

Centre for Research on Environmental Systems & Statistics (CRES), Lancaster University, Lancaster, UK (Also, Centre for Resource & Environmental Studies, Institute of Advanced Studies, Australian National University, Canberra ACT 2020, Australia)

This article discusses continuous and discrete-time transfer function (TF) models within the context of the Data-Based Mechanistic (DBM) modeling of hydrological systems. Although, at a superficial level, TF models simply provide a convenient and concise way of presenting differential or difference equations in an input-output form, they are much more than this. In particular, they allow for the consideration of linear dynamic systems in simple algebraic terms and for their analysis within a dynamic systems and control context. This is useful in various ways, particularly where a higher order TF model is identified and estimated directly from hydrological data. Often, this data-based model can then be decomposed into serial, parallel, and feedback connections of first-order systems that can be interpreted in hydrologically meaningful terms, thereby facilitating the model's use in hydrological systems analysis. The article also shows that TF models can be considered in time-variable and state-dependent parameter (SDP) form, so allowing for them to describe nonstationary and nonlinear systems. And, since it is straightforward to consider all TF models in stochastic terms, they provide a powerful vehicle for uncertainty, forecasting, and risk analysis. The practical utility and power of DBM TF models is illustrated by two real hydrological examples.

INTRODUCTION

From a conceptual standpoint, most mathematical models of hydrological systems are formulated on the basis of natural laws, such as dynamic conservation equations, often expressed in terms of continuous-time (CT) linear or nonlinear differential equations. The objective of *Data-Based Mechanistic* (DBM) modeling (e.g. Young and Lees, 1993; Young, 1998 and the prior references therein) is to infer the nature and structure of such models directly from hydrological data, using powerful methods of statistical identification and estimation, and to then interpret the model equations in physically meaningful terms. One important generic model class that facilitates such DBM modeling studies is the Transfer Function (TF) family of models, which is the subject of this article. The article starts by reviewing briefly some of the background to the use of TF models in hydrology, concentrating on prior publications concerned with the modeling of rainfall-flow (RF) processes. This is followed by a tutorial section that discusses the formulation of linear, constant parameter TF models in the derivative operator $s = \mathrm{d}/\mathrm{d}t$, which are simply the CT, TF form of ordinary differential equations. It then proceeds to develop and discuss the discrete-time (DT) equivalents of these CT models, namely, TFs in the backward shift operator z^{-1} (see later text), before outlining methods for the statistical identification and estimation of both CT and DT models from noisy time series data. In this manner, the models take on a natural stochastic form that is appropriate to their use in applications such as hydrological forecasting, uncertainty, and risk analysis. Two practical examples help demonstrate the practical utility of the TF approach in hydrology, one concerned with RF modeling and the other with the dynamics of solute transport and dispersion in rivers. Finally, the article also shows that TF models can be extended to incorporate time-variable and "state-dependent" parameters, so allowing them to describe nonstationary and nonlinear stochastic systems.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

TF MODELS IN HYDROLOGY

The most common application of TF models in hydrology is connected with the modeling of RF processes, where the objective is to characterize the dynamic relationship between rainfall and *total* measured flow in the river. Note that the term "rainfall-flow" is used here, with an emphasis on the total flow in the river rather than the ubiquitous term Rainfall-Runoff (RR). As we shall see later, this is because the most recent work on the application of data-based TF models in this context has shown how they are able to quantify other components of the flow; most importantly, those arising from groundwater processes that dominate the base flow in the river.

The characterization of the nonlinear dynamic relationship between rainfall and river flow is one of the most interesting and challenging modeling problems in hydrology. It has received considerable attention over the past 40 years, with mathematical and computer-based models ranging from simple black-box representations to complex, physically based catchment models. Other articles in this Encyclopedia will deal more comprehensively with this topic and there are many books that consider it, either in whole, or in part: for example, Anderson and Burt (1985: see particularly, the chapter by Wood and O'Connell); Shaw (1994), Singh (1995) and Beven (2001). But, in attempting to consider the historical context of this topic, we should recall Beven's comments in the latter book:

> "It is now virtually impossible for any one person to be aware of all the models that are reported in the literature, let alone know something of the historical framework of the different initiatives."

On this basis, it is clear that any treatment of such extensive and quite often controversial literature will tend to be somewhat subjective.

Wheater *et al.* (1993) have categorized RF models into four, broad types: conceptual models, physics-based models, metric models and Hybrid Metric-Conceptual (HMC) models. Conceptual and physics-based models tend to be driven by deterministic, reductionist thinking. As a result, they are often very large and suffer from problems of overparameterization that make rigorous statistical identification and estimation from the available data difficult or even impossible. At the other extreme, metric models, such as the neural network (e.g. Hsu *et al.*, 1993) and *neuro-fuzzy* types (e.g. Jang *et al.*, 1997), are the epitome of "black-box" modeling, revealing very little of their internal structure that has any physical meaning.

The HMC models are an attempt to combine the ability of metric models to efficiently characterize the observational data in statistical terms (the "principle of parsimony" (Box and Jenkins, 1970); or the "Occam's Razor" of antiquity), with the advantages of *simple* conceptual models that have a prescribed physical interpretation within the current scientific paradigm. Transfer functions have become a very useful tool in HCM modeling because their outward "black-box" nature obscures an underlying ability to explain the more physically meaningful mechanisms that underlie this input-output description; mechanisms that can be revealed using the "system theoretic" procedures of TF decomposition. In one important sense, TF models provide a rather natural model form for hydrologists because the impulse response of a continuous-time TF is equivalent to the Instantaneous Unit Hydrograph (IUH); while the pulse response of the discrete-time TF is equivalent to the discrete-time Unit Hydrograph (UH) (or TUH): see later Section "The unit hydrograph and finite impulse response TF models".

System theoretic modeling ideas have had quite a strong influence in hydrology over many years: to name but a few contributions, for example, see Amorocho and Hart (1964); Salas *et al.* (1980); Wood (1980); O'Connell and Clarke (1981); Sorooshian (1983); Gupta (1984); Bras and Rodriguez-Iturbe (1985) and Singh (1988). However, such models do not often attempt to represent the nonlinear dynamics inherent in the transformation of rainfall to runoff and, therefore, may not always perform well over long periods of time when the catchment storage and soil wetness is changing over a wide range. Sometimes, piecewise linearity has been assumed to handle some aspects of the nonlinear behavior (e.g. Natale and Todini, 1976a,b); while, in other cases, the model parameters have been allowed to vary with time (Young, 1974; Whitehead and Young, 1975; Whitehead *et al.*, 1976; Young and Wallis, 1985; Young, 1986).

The latter Time-variable Parameter (TVP) approaches to TF estimation have led on to a more overt treatment of nonlinearity, sometimes by the use of nonlinear black-box models (e.g. the early example of Hsu *et al.*, 1993), and sometimes by the extension of TF models to allow for nonlinearity, as considered in this article. As we shall see, this latter approach has the advantage of allowing for the easier inclusion of conceptual ideas and the retention of physical meaning within the estimated model structure. This desire to associate a physical meaning with TF models has a rather natural appeal to hydrologists and was probably first adumbrated by Young (1974) and Whitehead and Young (1975). The idea of a Physically Realizable Transfer Function (PRTF: see Han, 1991) tackled this problem by ensuring that the estimated TF model maintained mass conservation and did not have any nonphysical, oscillatory modes of dynamic behavior. It was also the basis for the work of Jakeman *et al.* (1990) and provided the philosophical and methodological underpinning of the DBM approach to modeling RF processes (Young, 1993; Young and Beven, 1994) that is discussed further in the following sections of this article.

Finally, it must be emphasized that TF models are not restricted to RF processes: since they provide a generic tool for modeling linear and nonlinear, stochastic dynamic

systems, they can be applied to data-based modeling problems in many areas of science, engineering, and the social sciences (see the wide variety of examples in Young, 1998). In the hydrological context, they have already been used for flow routing down rivers (see Section "Continuous-time TF models") so that, together with the rainfall-flow TF models, they allow for characterization of a whole river catchment. As we shall see in a later example, TF models are also fundamental to the Aggregated Dead Zone (ADZ) modeling of pollution transport and dispersion in rivers and soils. Given this flexibility, it seems likely that they will be used even more widely in the future as a basic tool for modeling in hydrological and environmental sciences.

CONTINUOUS-TIME TF MODELS

In order to introduce TF models, let us consider first a conceptual catchment storage equation in the form of a continuous-time, linear storage (tank or reservoir) model: see, for example, the review papers by O'Donnell, Dooge and Young in Kraijenhoff and Moll (1986); or more comprehensive treatments, such as the recent books by Beven (2001) and Dooge and O'Kane (2003). Here, the rate of change of storage in the channel is defined in terms of water volume entering the linear storage element (e.g. river reach) in unit time, minus the volume leaving in the same time interval, namely,

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = GQ_i(t-\tau) - Q_o(t) \tag{1}$$

where $Q_i(t-\tau)$ represents the input flow rate delayed by a pure time or "transport" delay of τ time units to allow for pure advection; and G is a gain parameter inserted to represent gain (or loss) in the system. Making the reasonable and fairly common assumption that the outflow is proportional to the storage at any time, that is,

$$Q_o(t) = TS(t)$$

and substituting into (1), we obtain,

$$T\frac{\mathrm{d}Q_o(t)}{\mathrm{d}t} = GQ_i(t-\tau) - Q_o(t) \tag{2}$$

This equation is a first-order, linear differential equation model whose response, from an initial steady flow condition, to a unit impulsive change Q_i^{imp} of the input flow at time $t = t_0$, is given by

$$Q_o(t) = Q_e + Q_i^{\mathrm{imp}}\mathrm{e}^{-(t-t_0)/T}$$

where Q_e is the initial steady flow level. This has a typical hydrograph recession shape, with a decay *Time Constant*, T, that defines the *Residence Time* of the model. As we shall see later, combinations of two or more such first-order models exhibit a typical UH form (e.g. Dooge, 1959; Beven, 2001).

By introducing the derivative operator s, that is, $s = \mathrm{d}/\mathrm{d}t$, and collecting like-terms together, it is easy to see that equation (2) can be written as,

$$(1+Ts)Q_o(t) = GQ_i(t-\tau)$$

so that, dividing throughout by $1+Ts$, we obtain the following continuous-time TF form of equation (2),

$$Q_o(t) = \frac{G}{1+Ts}Q_i(t-\tau) \tag{3}$$

where,

$$H(s) = \frac{G}{1+Ts}$$

represents the TF in terms of the derivative operator s. Note that the same letter s (or sometimes p) is used to represent the related Laplace transform operator. Considered in these Laplace transform terms, it is possible to utilize Laplace transform methods to handle initial conditions on the variables in the model and analytically compute its response (here $Q_o(t)$) to variations in input variable (here $Q_i(t)$). However, this is not essential in the present context, although the interested reader should find the article a good primer for the study of Laplace transform methods (e.g. Schwarzenbach and Gill, 1979).

Physically Interpretable Parameters

The TF model 3 is characterized by three parameters: G, T, and τ. However, there are five, physically interpretable model parameters associated with the model that are worth discussing. The Steady State Gain (SSG), denoted by G, is obtained by setting the s operator in the TF to zero (i.e. $\mathrm{d}/\mathrm{d}t = 0$ in a steady state). It shows the relationship between the equilibrium output and input values when a steady input is applied. For this reason, *if the input and output have similar units*, G is ideal for indicating the physical losses or gains occurring in the system. In the case of a flow-routing model, for example, it indicates whether water has been added ($G > 1$) or lost ($G < 1$) between the upstream and downstream boundaries; and the percentage of water lost or gained can be defined by *Loss Efficiency* $LE = 100(1-G)$, which will be negative if $G > 1.0$. As pointed out above, the *Residence Time* or *Time Constant* T is the time required for the storage element output to decay to e^{-1} or 0.3679 of its maximum value in response to an impulsive input. Finally, the pure *Advective Time Delay* τ indicates the time it takes for a flow increase upstream to be first detected downstream: and $T_t = T + \tau$ defines the *Travel Time* of the system. These five parameters typify the

equilibrium and dynamic characteristics of the TF model and provide a physical interpretation of the TF model in terms of its mass transfer and dispersive characteristics.

TF Manipulation and Block Diagrams

The first-order TF model (3) is often written in the form,

$$Q_o(t) = \frac{g_0}{s + f_1} Q_i(t - \tau) \quad i.e. \quad H(s) = \frac{g_0}{s + f_1} \qquad (4)$$

where,

$$g_0 = \frac{G}{T}; \qquad f_1 = \frac{1}{T}$$

because this is the form in which the model is normally estimated (see later). Typically, a Channel or Flow-routing model for a river catchment will contain a number of elemental models, such as 3 or 4, connected in a manner that relates to the structure of the catchment. For instance, a serial connection of n such elements constitutes the lag-and-route model of a single river channel (Meijer, 1941; Dooge, 1986) and, with $\tau = 0$ and all elements identical, it becomes the well known "Nash Cascade" model (Nash, 1959). More complex river systems can be represented by a main channel of this type, with tributaries modeled in a similar manner. Also, a typical TF model between effective rainfall and flow often contains a parallel connection of two or more such storage elements (see Young, 1992; 2001b,c, 2002a, 2003). Related serial and parallel arrangements characterize the mass conservation equation of the ADZ model for the transport and dispersion of a solute in a river channel (e.g. Wallis *et al.*, 1989 and the prior references therein). And ADZ-type models can lead to more complex interactive water quality models, including chemical and biological interaction, such as dissolved oxygen-biological oxygen demand (DO-BOD) models (e.g. Beck and Young, 1975; Whitehead and Young, 1975; Young, 1999b). Practical examples of both RF and ADZ modeling are described later in Section "STATISTICAL IDENTIFICATION AND ESTIMATION OF TF MODELS".

The TF formulation allows for the visual representation of a total system model in the form of a *Systems Block Diagram*. Figure 1 is a typical example of such diagram that represents a catchment model consisting of a effective RF submodel involving two *different* first-order TFs of the form (3) with $\tau = 0$,

$$H_1(s) = \frac{G_1}{1 + T_1 s}; \quad H_2(s) = \frac{G_2}{1 + T_2 s}$$

that are connected in parallel. The flow output of this submodel forms the input to a flow-routing submodel composed of two *identical*, first order TFs,

$$H_3(s) = \frac{G_3}{1 + T_3 s}$$

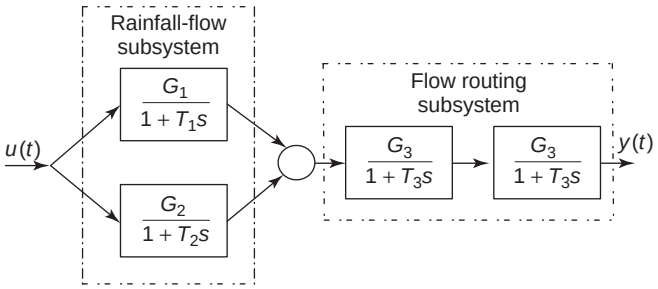


Figure 1 Block diagram of a hypothetical catchment model consisting of a parallel pathway, rainfall-flow subsystem in series with a two reach, flow-routing subsystem

again of the form (3) with $\tau = 0$, but this time connected in series as a "Nash Cascade". The upstream input to this complete system is an effective rainfall measure, represented by $u(t)$, that is generated from the gauged rainfall $r(t)$ (not shown) by a nonlinear relationship $N(s)$ involving a measure of the catchment wetness (see later example). The downstream output of the complete system is denoted by $y(t)$.

One advantage of the TF formulation of the model shown in Figure 1 is that it allows for the computation, using *Block Diagram Algebra*, of a single, multiorder TF that represents the total system. Here, TFs connected in parallel are additive; while those connected in series are multiplicative. Consequently, in this case, the two submodels can be represented by the following composite TFs:

$$H_{rf}(s) = H_1(s) + H_2(s); \quad H_{ff}(s) = H_3(s).H_3(s)$$

So that, with the above definitions of $H_1(s)$, $H_2(s)$ and $H_3(s)$,

$$H_{rf}(s) = \frac{G_1}{1 + T_1 s} + \frac{G_2}{1 + T_2 s} = \frac{(G_1 + G_2) + (G_1 T_2 + G_2 T_1)s}{(1 + T_1 s)(1 + T_2 s)}$$

and

$$H_{ff}(s) = \frac{(G_3)^2}{(1 + T_3 s)(1 + T_3 s)}$$

Now, since $H_{rf}(s)$ and $H_{ff}(s)$ are connected in series, the TF of the total system $H(s)$ is obtained as the multiplication of these two composite TFs: that is,

$$H(s) = H_{rf}(s).H_{ff}(s) = \frac{(G_1 + G_2) + (G_1 T_2 + G_2 T_1)s}{(1 + T_1 s)(1 + T_2 s)} \times \frac{(G_3)^2}{(1 + T_3 s)(1 + T_3 s)}$$

Multiplying out these expressions, we see that the complete $H(s)$ is a 4th order TF that can be manipulated to the form:

$$H(s) = \frac{g_0 + g_1 s}{s^4 + f_1 s^3 + f_2 s^2 + f_3 s + f_4} \tag{5}$$

where we will leave the definition of the parameters $f_i, i = 1, 2, \ldots, 4$ and $g_j, j = 0, 1$ in 5, as an exercise for the reader (*hint*: carry the above analysis with the TF representation 4 rather than 3: then it will be clear why the former representation is better for analysis, although it lacks the direct physical interpretation of the latter, which provides a better form for the block diagram representation).

The General, Multi-Order TF Model

It is clear from the above example that, in general, serial, parallel (or even feedback) connections of elemental first order TF models, such as 3 or 4, lead to a multi-order TF model. The block diagram algebra for a feedback connection is a little more complicated but, since it is not particularly relevant in the current context, the interested reader should consult a standard text on the subject (e.g. Schwarzenbach and Gill, 1979) to find the details. In the present context, the general multi-order TF form is as follows:

$$x(t) = \frac{G(s)}{F(s)} u(t - \tau) \quad y(t) = x(t) + \xi(t) \tag{6}$$

where $F(s)$ and $G(s)$ are polynomials in s of the following form:

$$F(s) = s^p + f_1 s^{p-1} + f_2 s^{p-2} + \cdots + f_p$$
$$G(s) = g_0 s^q + g_1 s^{q-1} + g_2 s^{q-2} + \cdots + g_q$$

in which p and q can take on any positive integer values. Here $u(t)$ and $x(t)$ denote the deterministic input and output signals of the system at its upstream and downstream boundaries, respectively; τ is a pure advective time (transport) delay affecting the input signal $u(t)$; and $y(t)$ is the observed output, which is assumed to be contaminated by a noise or stochastic disturbance signal $\xi(t)$. This noise is assumed to be independent of the input signal and it represents the aggregate effect, at the downstream boundary, of all the stochastic inputs to the system, including distributed unmeasured inputs, measurement errors, and modeling error. Multiplying throughout equation (6) by $F(s)$ and converting the resultant equation to alternative ordinary differential equation form, we obtain:

$$\frac{d^p y(t)}{dt^p} + f_1 \frac{d^{p-1} y(t)}{dt^{p-1}} + \cdots + f_p y(t) = g_0 \frac{d^q u(t - \tau)}{dt^q} + \cdots + g_q u(t - \tau) + \eta(t) \tag{7}$$

where $\eta(t) = F(s)\xi(t)$ is a modified noise signal generated by the manipulation of the equation. The structure of this model, in either form (6) or (7), is defined by the triad $[p\ q\ \tau]$.

If the noise $\xi(t)$ can be assumed to have specific statistical characteristics, then this can help in obtaining low or minimum variance estimates of the TF model parameters. But, as we see later when discussing the identification and estimation of TF models, such assumptions are not an essential requirement of statistical estimation and good, low variance parameter estimates can be obtained without invoking them. Also, depending on the objectives of the modeling study, it may be necessary, in a complete system consisting of many sub-elements such as (3) or (4), to consider noise inputs *within* the system, associated with collections of subelements that have distinct physical meaning: for example, stochastic lateral inflows.

Discrete-time, Sampled Data TF Models

To date, the most popular form of TF modeling has been carried using the DT equivalents of the models (4) and (6). In the case of equation (4), this discrete-time TF model takes the form:

$$Q_{o,k} = \frac{b_0}{1 + a_1 z^{-1}} Q_{i,k-\delta} \tag{8}$$

Here, $Q_{o,k}$ is the downstream flow measured at the kth sampling instant, that is at time $k\Delta t$, where Δt is the sampling interval in time units. $Q_{i,k-\delta}$ is the input flow at time $(k - \delta)\Delta t$ time units previously, where δ is the advective time delay, normally defined as the nearest integer value of $\tau/\Delta t$ (thus incurring a possible approximation error); and z^{-1} is the backward shift operator, that is, $z^{-r} Q_{o,k} = Q_{o,k-r}$. Of course, this model can be written in its "difference equation" form, namely:

$$Q_{o,k} = -a_1 Q_{o,k-1} + b_0 Q_{i,k-\delta} \tag{9}$$

which is obtained by simple cross multiplication and application of the z^{-1} operator. This reveals that the flow $Q_{o,k}$ at the kth sampling instant is a proportion $-a_1$ (note that, in the present context, a_1 will be a negative number less than unity, so that this is a positive proportion) of its value $Q_{o,k-1}$ at the previous $(k-1)^{th}$ sampling instant, plus a proportion b_0 of the delayed upstream flow input $Q_{i,k-\delta}$ measured δ sampling instants previously.

The values of the parameters a_1 and b_0 in equations (8) and (9) can be related to the parameters of the model (4) in various ways depending upon how the input flow $Q_i(t)$ is assumed to change over the sampling interval between the measurement of $Q_{i,k-1}$ and $Q_{i,k}$ (since it is not measured over this interval). The simplest and most common assumption is that it remains constant over this

interval (the so-called zero-order hold, ZOH, assumption), in which case the relationships are as follows:

$$a_1 = -\exp(-f_1 \Delta t) \quad b_1 = \frac{g_0}{f_1}\{1 - \exp(-f_1 \Delta t)\} \quad (10)$$

Note that, because these relationships are functions of the sampling interval Δt, for every unique CT model such as (4), there are infinitely many DT equivalents (8), depending on the choice of Δt, all with different parameter values defined in equation (10). Following from the definition of this first-order DT model at the chosen Δt, the general multiorder equivalent of the general CT model (6) is defined as follows:

$$x_k = \frac{B(z^{-1})}{A(z^{-1})} u_{k-\delta} \quad y_k = x_k + \xi_k \quad (11)$$

where,

$$A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}$$
$$B(z^{-1}) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_m z^{-m}$$

Normally $n = p$ but m may be equal or greater than q. The structure of this DT model is defined by the triad $[n\ m\ \delta]$. Note finally that the TFs in both (11) and (3) are composed of ratios of polynomials (in z^{-1} and s respectively), so they are sometimes referred to as *Rational Transfer Functions*.

The Unit Hydrograph and Finite Impulse Response TF Models

The reader can verify that the division of the numerator polynomial $B(z^{-1})$ by the denominator polynomial $A(z^{-1})$ in the TF model (2) normally results in a infinite dimensional polynomial $G(z^{-1}) = g_0 + g_1 z^{-1} + g_2 z^{-2} + \cdots + g_\infty z^{-\infty}$, so that the equation can also be written in the alternative form:

$$y_k = \sum_{i=\delta}^{\infty} g_i u_{k-i} + \xi_k; \quad g_j,\ j = 0, 1, \ldots, \delta - 1 = 0 \quad (12)$$

This will be recognized as the discrete-time form of the convolution integral equation associated with the solution of differential equations (here with a pure time delay of δ sampling intervals or $\delta \Delta t$ time units) and, once again, we see that the TF model is simply a discrete-time equivalent of a continuous-time differential equation model.

Equation (12) has important connotations in hydrology because, if the noise $\xi_k = 0$ for all k, then its unit impulse response (that is the response of the equation to a unit impulse input: $u_k = 1.0$ for $k = 1; u_k = 0$ for all $k > 0$), is equivalent to the hydrological UH. Sometimes, indeed, the equation (12) is referred to as a TF model, although its infinite dimensional nature is an obvious restriction. Despite this disadvantage, some hydrologists have used the equation directly in RF modeling by considering it in a Finite Impulse Response (FIR) form, in which the upper limit of the summation is set to some value $r < \infty$ where it is considered that the ordinates of the impulse response have become small enough to be ignored.

Unfortunately, it can be shown that the number of FIR model parameters (or "weights"), $g_i, i = 1, 2, \ldots, r$, is nearly always much larger than the number of parameters in the rational TF form (11). As a result, the statistical estimates of these parameters will normally have unacceptably high variance and have to be constrained in some manner. For instance, Natale and Todini (1976) use quadratic programming to compute the constrained least squares estimates of the FIR parameters in order to ensure that they are all positive and, if necessary for conservation purposes, they all sum to unity (i.e. the model has unity steady state gain). Even with this approach, however, it is difficult to recommend the FIR model because the rational TF model (11), normally with far fewer parameters, is not only entirely equivalent to the infinite dimensional IR model, without any approximation, it is also easier to estimate from RF data using the methodology discussed in the next Section "STATISTICAL IDENTIFICATION AND ESTIMATION OF TF MODELS".

STATISTICAL IDENTIFICATION AND ESTIMATION OF TF MODELS

In statistical terms, the modeling of a system using a TF model is a problem of time-series analysis (e.g. Box and Jenkins, 1970; Young, 1984) that involves identification and estimation of the model from time-series data. Here, Identification means the definition of the most appropriate model structure: in particular, the values of the elements in the triads $[p\ q\ \tau]$ and $[n\ m\ \delta]$. *Estimation* is then the estimation of the parameters that characterize this structure ($f_i, i = 1, 2, \ldots, p$; $g_j, j = 0, 1, \ldots, q$; and τ in the CT case; $a_i, i = 1, 2, \ldots, n$; $b_j, j = 0, 1, \ldots, m$; and δ in the DT case). Much has been written about the identification and estimation of the DT model (11) but much less on the CT model 6. This is almost certainly for two reasons: first, it is easier to analyze the DT model in statistical terms; second, in these days of digital instrumentation and computers, the DT model provides a more convenient model form. For these reasons, attention will be concentrated initially on the DT model (11) and we will return to the more physically transparent CT modeling later.

Discrete-time TF Model Identification and Estimation

Much has been written on the identification and estimation of discrete-time TF models and the reader is advised to

consult one or more of the many texts on the subject (e.g. Ljung and Söderstrom, 1983; Young, 1984; Norton, 1986). Here, we will not attempt to review this literature but simply refer briefly to a number of topics that are particularly pertinent to the use of TF models in hydrology.

The Box–Jenkins Model

The best known work on the identification and estimation of the DT model (11) is by Box and Jenkins (1970: see Chapter 10), to which the reader is directed for a comprehensive description of how the model can be estimated using the Maximum Likelihood (ML) approach, under the assumption that the noise ξ_k has rational spectral density. This assumption leads to the formulation of the following *Box–Jenkins* model:

$$y_k = \frac{B(z^{-1})}{A(z^{-1})}u_{k-\delta} + \xi_k \quad \xi_k = \frac{D(z^{-1})}{C(z^{-1})}e_k \quad e_k = N(0, \sigma^2)$$

or,

$$y_k = \frac{B(z^{-1})}{A(z^{-1})}u_{k-\delta} + \frac{D(z^{-1})}{C(z^{-1})}e_k \quad e_k = N(0, \sigma^2) \qquad (13)$$

where,

$$C(z^{-1}) = 1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_s z^{-s}$$
$$D(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} + \cdots + d_r z^{-r}$$

and e_k is defined as normally distributed "white noise", defined in statistical terms as $N(0, \sigma^2)$: that is a zero mean, serially uncorrelated sequence of normally distributed random variables with variance (an additional unknown parameter) σ^2. We see, therefore, that the "colored" noise ξ_k is assumed to be generated from the white noise e_k by passing it through a "filter" that has a similar TF form to the system model, but with both polynomials in z^{-1} having leading unity elements (monic polynomials). This particular type of filter is called a AutoRegressive Moving Average (ARMA) model (see Box and Jenkins, 1970). With the noise process defined in this manner, it is clearly necessary to identify the order of the noise model polynomials $C(z^{-1})$ and $D(z^{-1})$ and estimate the associated parameters $c_i, i = 1, 2, \ldots, s$ and $d_i, i = 1, 2, \ldots, r$ and σ^2.

The ARMAX Model

In the systems and control literature, the most widely used approach to DT model estimation is probably the Prediction Error Minimization (PEM) method (e.g. Ljung and Söderstrom, 1983), which is the preferred method in the Matlab™ Identification Toolbox. PEM estimation is most often applied to the alternative AutoRegressive Moving Average eXogenous (ARMAX) TF model form: that is,

$$A(z^{-1})y_k = B(z^{-1})u_{k-\delta} + C(z^{-1})e_k; \quad e_k = N(0, \sigma^2) \qquad (14)$$

Despite the popularity of this ARMAX model in the systems literature, there are good theoretical and practical reasons for preferring the model (13), as discussed below.

Refined Instrumental Variable (RIV) Estimation

One approach to TF model estimation that is simpler and more robust, because it does not rely on the assumption that the noise has rational spectral density, is the Refined Instrumental Variable (RIV) method, which has proven particularly useful and robust in hydrological applications. The RIV approach (see Young and Jakeman, 1979; Young, 1984) is based on the TF models (11) and (13) and was initially developed with environmental applications in mind, where robustness to the kinds on nonstandard conditions met in the environment, including hydrology, is very important. The RIV method has three primary advantages that are pertinent to the discussion here.

1. In the case of the full Box–Jenkins model (13) and following from an important theorem by Pierce (1972), the estimates of the "system" parameters in $A(z^{-1})$ and $B(z^{-1})$ are asymptotically independent of those for the "noise" parameters in $C(z^{-1})$ and $D(z^{-1})$. This has important consequences on parameter estimation (Jakeman and Young, 1981, 1983) that justify the RIV estimation of the TF model (13) rather than the ARMAX model (14). They also expose deficiencies in the ARMAX model estimates if the data are generated by the model (13).
2. Like all instrumental variable methods (see e.g. Young, 1984), the RIV approach does not necessarily require the concurrent estimation of a noise model and, indeed, remains statistically efficient (i.e. the parameter estimates are optimal in the sense that amongst all estimates, they have the minimum variance) if the noise ξ_k is itself white (i.e. $C(z^{-1}) = D(z^{-1}) = 1.0$). Used in this manner, the RIV algorithm is, for obvious reasons, called the Simplified RIV (SRIV) algorithm (Young, 1985).
3. As pointed out by Wellstead (1978) and considered further by the present author and his colleagues (Young *et al.*, 1980; Young, 1989), model structure identification (definition of the triad $[n\ m\ \delta]$) is greatly facilitated by the special properties of the Instrumental Product Matrix (IPM) that is a fundamental component of IV estimation and has optimal statistical properties in the case of RIV/SRIV estimation. This has been exploited in the development of the Young Information Criterion (YIC) order identification criterion mentioned later.
4. The RIV/SRIV algorithms for DT models utilize a special form of adaptive prefiltering. This is further exploited in the RIV method for continuous-time systems (the RIVC algorithm: Young and Jakeman, 1980; Young, 2002b) to bypass the need for direct

differentiation of the input and output signals. This algorithm is discussed further in the following section.

Continuous-time TF Model Identification and Estimation

Following on the point 4. in the previous subsection, the statistical estimation of the CT model (6) is straightforward if completely continuous-time data are available. However, the hydrologist is normally confronted by discrete-time, sampled data and the problem of modeling continuous-time models such as this, based on discrete-time, sampled data at sampling interval Δt, is not so obvious. This problem can be approached in two main ways.

- *The Indirect Approach:* Here, the identification and estimation steps are first applied to the DT model (11) or (13). This estimated model is then converted to the CT model (6), again using some assumption about the nature of the input signal $u(t)$ over the sampling interval Δt. In the first-order, ZOH case, this conversion is given by the relationships in equation (10) but, in more general terms, the multiorder conversion must be carried out in a computer, using a conversion routine such as the *d2cm* algorithm in the Matlab Control Toolbox.

- *The Direct Approach:* Here, the RIVC algorithm, together with an appropriate order identification criterion (see the following text), is used to identify the most appropriate, identifiable CT model structure defined by the triad $[p \;\; q \;\; \tau]$; and then estimate the TF parameters $f_i, i = 1, 2, \ldots n$, $g_j, j = 0, 1, \ldots m$ and τ that characterize this structure. Of course, some approximation will be incurred in this estimation procedure because the inter-sample behavior of input signal $u(t)$ is not known and it must be interpolated over this interval in some manner (see the preceding text).

Model Structure Identification

Finally, as mentioned previously, an important aspect of TF modeling, in both continuous and discrete-time, is *Model Structure Identification* (the definition of the model structure triads). This can be approached in various ways. For example, Box and Jenkins (1970) discuss the topic at length and Akaike (1974) suggested an alternative approach that has since spawned several other, related methods. Within the context of the instrumental variable methods discussed previously, however, model structure identification is based conveniently on two statistical criteria. First, the *Coefficient of Determination*, R_T^2, a normalized measure based on the variance of the error between the sampled output data y_k and the simulated (CT or DT) model output at the same sampling instants (this is normally equivalent to the well known *Nash-Sutcliffe Efficiency* measure (Nash and Sutcliffe, 1970); and $100 \times R_T^2$ is the percentage of the variance of the output data explained by the model). Note that R_T^2 should not be confused with its well known relative R^2, as used in classical regression and time-series analysis, which is defined in terms of the one-step-ahead prediction errors. In general, R_T^2 is a more discerning measure of model adequacy in a dynamic systems context than R^2 since it quantifies the ability of the model to explain the whole of the output data from the input data, without any reference to the output data. In contrast, R^2 only measures the ability of the model to predict one-step-ahead on the basis of the latest input and output data. The second statistic is the YIC, which is a heuristic measure of model identifiability and is based on the properties of the IPM matrix (see the previous discussion). These statistics are discussed in more detail in Young (1989) and Appendix 3 of Young (2001b).

TWO PRACTICAL EXAMPLES

TF models can be applied in many areas of hydrology and provide a good basis for DBM modeling. Two typical examples are outlined in this section. The first is a TF model of the relationship between effective rainfall and flow in an Australian river catchment. This DBM model could provide a direct basis for flow forecasting at the flow measurement station or it could be extended to incorporate flow-routing TF elements down the river in order to allow for forecasting over the whole catchment area (that may involve other RF modules). The second example is an ADZ model for the transport of a conservative dye tracer in a Florida wetland area. This DBM model describes the physical aspects of solute advection and dispersion in the wetland and so could provide the physical basis for more complex, nonconservative water quality modeling involving biological and chemical mechanisms.

Rainfall-flow Modeling

Figure 2 shows a portion of effective rainfall $u(t)$ and flow $y(t)$ data from the River Canning, an ephemeral River in Western Australia. A longer set of these data has been analyzed comprehensively in Young *et al.* (1997), using DT modeling. The effective rainfall series plotted in part (b) of Figure 2 is generated from the effective rainfall nonlinearity identified in this earlier study, with the effective rainfall generated by the equation

$$u(t) = r(t)y(t)^{\gamma} \tag{15}$$

where $r(t)$ is the measured rainfall and γ is a power law parameter. Here, the flow $y(t)$ is acting as a surrogate measure of the catchment storage, rather than attempting to model this storage separately in some conceptual manner (see e.g. Jakeman *et al.*, 1990; Wheater *et al.*, 1993; Young 2001b, 2003). In the following, we obtain direct and indirectly estimated, linear CT models between the effective

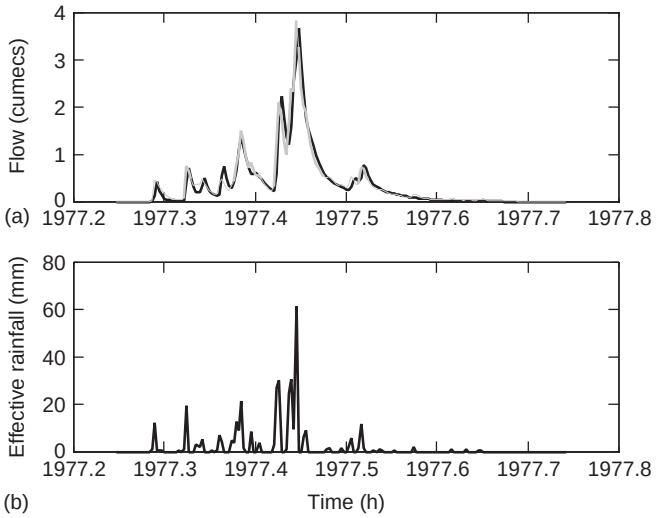


Figure 2 Daily effective rainfall (b) and flow (a) data for the River Canning in Western Australia. The dotted line appears to me to have come out as a grey line (a) is the simulated output of the DBM model (see text)

rainfall $u(t)$, as defined in the above manner, and the flow $y(t)$. Of course, the DT model is useful in its own right and provides an excellent basis for applications such as flow forecasting (see e.g. Lees *et al.*, 1994; Lees, 2000a,b; Young, 2001c, 2002a). However, the transparency of the CT model in revealing, more obviously, the physical interpretation of the model makes it more useful as a vehicle in DBM modeling and the use of the model in wider catchment studies.

Direct CT Identification and Estimation

Let us consider first identification and estimation based on the measured daily data shown in Figure 2. The RIVC algorithm, in combination with the R_T^2 and YIC statistics, identifies a [2 3 0], second order model with the parameter estimates:

$$\begin{array}{ll} \hat{f}_1 = 0.457 \ (0.032) & \hat{f}_2 = 0.0248 \ (0.0045) \\ \hat{g}_0 = 0.0138 \ (0.001) & \hat{g}_1 = 0.0505 \ (0.002) \\ \hat{g}_2 = 0.0046 \ (0.0008) & \end{array}$$

where, here and elsewhere in the paper, the figures in parentheses are the estimated standard errors obtained from the RIVC algorithm. This model output (shown as the dotted line in (a) of Figure 2) explains the data very well with $R_T^2 = 0.980$ (98% of the output flow variance explained by the simulated CT model output) and a residual variance of 0.00723 (standard deviation 0.085 cumecs, where the maximum flow is 3.86 cumecs). The model can also be decomposed by standard TF decomposition (see earlier text) into a parallel pathway form. The three pathways are: (i) an instantaneous effect, accounting for 7.4% of the flow, which is a measure of rapid flow that occurs in the very short term (within one sampling interval; here one day); (ii) a quick pathway, accounting for 54.1% of the flow, that represents a linear store with a residence time of about 2.53 days, probably the result of surface and shallow, subsurface processes in the catchment; and (iii) a slow or base-flow pathway, accounting for 38.5% of the flow, that passes through a much longer 15.9 day residence time, linear store, probably representing the effects of deeper groundwater processes and the displacement of old water.

Indirect CT Identification and Estimation

The first step of indirect CT modeling involves DT identification using the RIV algorithm (see earlier). This identifies a similar [2 3 0] structure TF with the following parameter estimates:

$$\begin{array}{ll} \hat{a}_1 = -1.6034 \ (0.008) & \hat{a}_2 = 0.6244 \ (0.007) \\ \hat{b}_0 = 0.0140 \ (0.001) & \hat{b}_1 = 0.0151 \ (0.002) \\ \hat{b}_2 = 0.0252 \ (0.0013) & \end{array}$$

However, since this is a discrete-time model, it has to be converted to continuous-time form. The *d2cm* algorithm in Matlab (see Section "CONTINUOUS-TIME TF MODELS".) accomplishes this conversion, using a ZOH approximation (i.e. the input effective rainfall is assumed to be constant over the sampling interval). The resulting CT model parameter estimates are:

$$\begin{array}{lll} \hat{f}_1 = 0.4711 & \hat{f}_2 = 0.0264 & \\ \hat{g}_0 = 0.0140 & \hat{g}_1 = 0.0514 & \hat{g}_2 = 0.0049 \end{array}$$

In this case, the model is quite similar to the directly estimated CT model in the previous subsection (i). But note that there are no standard errors quoted here: this is because these are not available after the conversion and must be computed separately in some manner, usually by the less convenient and lengthy Monte Carlo simulation analysis.

This lack of directly available estimation statistics is not the only disadvantage of the indirect approach. Unlike the CT model parameters, which are independent of the sampling interval Δt, the estimated DT model parameters are a function of Δt, so that different numerical estimates of the parameters are obtained for different values of Δt. Moreover, the *quality* of these estimates is also a function of Δt: in particular, if the sampling interval is too small, then the estimates are poorly defined statistically; the resulting CT model parameter estimates will be impaired; and the whole estimation process is less robust (see Young, 2004a, for discussion on the reasons for this). Conveniently, it is under these same, rapid sampling (small Δt) conditions that the direct CT approach works best. So, in practical terms, indirect CT model estimation works very well when the data are not sampled too rapidly but direct CT estimation is preferable at rapid sampling intervals.

Aggregated Dead Zone (ADZ) Modeling of Tracer Data

Tracer experiments can provide important information on how solutes are transported and dispersed in hydrological systems. This example is concerned with TF modeling of the input-output data shown in Figure 3 (TF modeling is also applied to a different set of tracer data in the article **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3**). These tracer data were obtained from a bromide tracer experiment carried out in a Florida wetland area receiving treated domestic wastewater for further nutrient removal. The experiment was part of a study carried out by Chris Martinez and Dr. William R. Wise of the Environmental Engineering Sciences Department, University of Florida for the City of Orlando. The bromide tracer was injected 765 meters upstream of a weir, at which samples were taken with a sampling interval Δt of 2 h. As we shall see, TF models are not only able to explain these tracer data well, but following the requirements of DBM modeling, they can also be interpreted in ADZ model terms that have physical meaning (see Wallis *et al.*, 1989). Here, we will consider the discrete-time TF model and utilize the SRIV algorithm (see earlier text) to identify the model order and estimate the parameters. However, continuous-time TF estimation using the continuous-time SRIV algorithm yields similar (albeit not identical) results and the discrete-time analysis is used here simply for illustration.

Although the impulsive ("gulp") input does not perturb the system continually (in the systems literature, it is said to be "not persistently exciting"), the SRIV algorithm has no difficulty identifying and estimating a low order model. The best identified TF, based on the R_T^2 and YIC criteria, is either 3rd or 4th order, but subsequent analysis, described later, suggests that the latter is superior from a physical standpoint. The estimated [4 2 22] (4th order denominator, 2nd order numerator and a 22 sampling interval pure time delay) takes the form:

$$y_k = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u_{k-22} + \xi_k \tag{16}$$

where,

$$\hat{A}(z^{-1}) = 1 - 3.67z^{-1} + 5.06z^{-2} - 3.11z^{-3} + 0.72z^{-4}$$
$$\hat{B}(z^{-1}) = 0.00103 - 0.00101z^{-1}$$

Here the "hat" denotes the estimated value; y_k is the observed tracer concentration at the weir and u_k is the impulsive input of tracer (186.33 mg l^{-1}), both measured at the kth sampling instant. Note that, although the standard errors on the parameters are available they are not cited specifically. However, the uncertainties on the physically meaningful, derived ADZ model parameters are discussed later. Note also that the large advective time delay of 22 sampling intervals (44 h) is the time taken for the solute to first reach the weir. The noise ξ_k is small and so the model explains the data very well with $R_T^2 = 0.997$ (i.e. 99.7% of the output variance is explained by the model).

Unfortunately, despite its ability to describe the data very well, the model (16) is not immediately acceptable from a DBM standpoint, primarily because the eigenvalues are $\{0.988, 0.964, 0.860 \pm 0.132j\}$, where $j = \sqrt{(-1)}$ is the complex number. Here, the pair of complex roots are associated with oscillatory response and this is difficult to justify in ADZ modeling terms. In particular, the elemental, single reach ADZ model is a first-order differential equation based on mass conservation (Wallis *et al.*, 1989) and so, other than in exceptional circumstances, multiple reach ADZ models must be characterized by real eigenvalues when considered in TF terms.

In the present circumstances, the most obvious approach is to re-estimate the model in a form where the eigenvalues are constrained to be real. This was carried out by means of constrained, nonlinear least squares optimization using the *leastsq* optimization procedure in Matlab. To ensure that the most parametrically efficient model was obtained, both [3 2 22] and [4 2 22] models were considered in this analysis but the latter yielded much the best constrained model, which has the following form:

$$y_k = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u_{k-22} + \xi_k \tag{17}$$

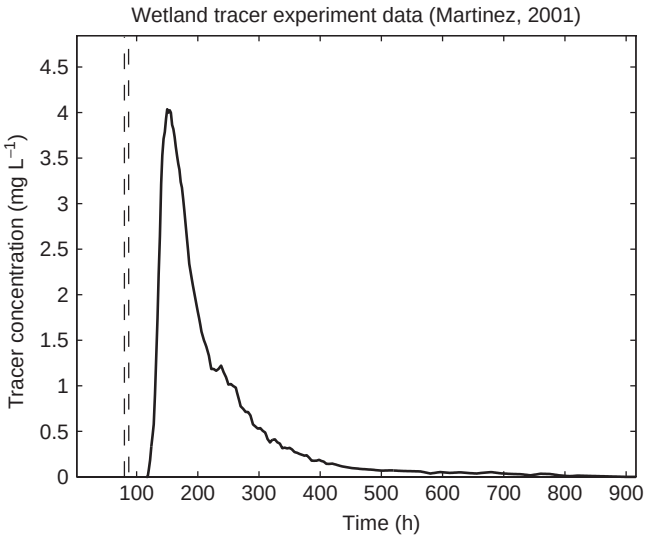


Figure 3 Wetland tracer experiment data: the input u_k is an impulsive or "gulp" application of bromide tracer (dashed line) and the output y_k is the concentration of bromide measured every 2 h at a downstream weir (full line)

where,

$$\hat{A}(z^{-1}) = (1 - 0.980z^{-1})(1 - 0.855z^{-1})^3$$
$$\hat{B}(z^{-1}) = 0.00127 - 0.00121z^{-1}$$

This model is well defined statistically and it explains 99.7% of the experimental data ($R_T^2 = 0.997$), the same as the unconstrained model (1). Figure 4 compares the model output (full line) with the measured output y_k (circular points).

Unlike the TF model (16), the model (17) not only has four real eigenvalues, as required, but three of these are repeated, so defining three identical ADZ reaches. These eigenvalues define ADZ residence times (time constants) of 99 h and 12.8 h (×3), giving a total estimated residence time for the wetland cell of 137.4 h ($99 + 3 \times 12.8$). One particular, physically meaningful, decomposition and interpretation of the model defined in this manner is again obtained by partial fraction expansion of the TF in equation (17). This decomposition is shown in the block diagram of Figure 5 which, it will be noted, has exactly the same form as Figure 1, except for the addition of the pure advective time delay, τ, demonstrating the widely applicable, generic nature of TF models such as these.

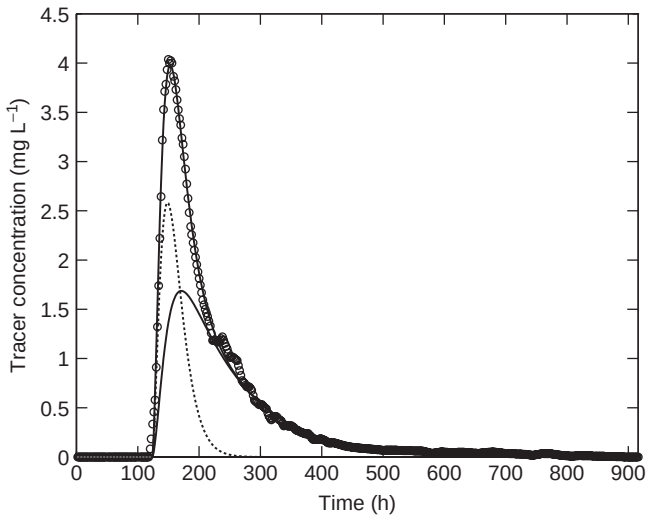


Figure 4 Comparison of the DBM model output (full line) and tracer experiment data (circular points). Also shown are the inferred slow-flow component (dashed line) and quick-flow component (dotted line)

The block diagram in Figure 5 shows that the solute is partitioned initially into a parallel pathway form with a "quick-flow" pathway, consisting of a single ADZ reach with a residence time of 12.8 h, and a "slow-flow" pathway with a much longer residence time of 99 h. The two pathways then add together and the solute passes further downstream, through two more ADZ reaches, each with a residence time of 12.8 h. The total *travel time* for this complete system is 181.5 h (the sum of the 44 h advective time delay and the cumulative overall time constant of 137.4 h). This means that the *dispersive fraction* (see Wallis *et al.*, 1989; Young and Wallis, 1994; Young, 1999a) is 0.76 (i.e. 137.4 ÷ 181.5): in other words, 76% of the water appears to be effective in dispersing the solute. This is a very high proportion, reflecting the nature of the system in this case, with a much higher potential for dispersion of tracer than in normal, faster moving streams, where the dispersive fraction is normally in the range 0.3 to 0.4. The inferred responses of the two pathways are plotted in Figure 4: the dotted line shows the estimated concentration changes in the quick-flow pathway (effectively three 12.8 h residence time ADZs in series), which accounts mainly for the initial response measured at the weir; while the dashed line shows the estimated changes in the slow-flow pathway (effectively a 99 h residence time ADZ in series with two 12.8 h ADZs), and these are mainly responsible for the raised tail of the measured response.

It is possible to compute estimates of other physical attributes associated with the model. First, the steady state gains associated with the two parallel pathways define the partitioning of the flow, with 33% of flow associated with the quick pathway and 67% with the slow pathway. And since the flow rate is known in this example, the Active Mixing Volumes (AMVs: Young and Lees, 1993), based on the estimated partitioned flow, are 361 m^3 in the quick

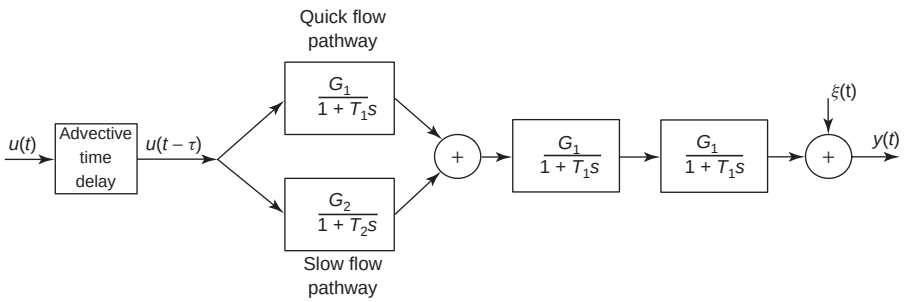


Figure 5 Block diagram of a transfer function decomposition for the ADZ model that can be interpreted in physical terms as a parallel-serial decomposition of ADZ reaches. $T_1 = 12.8$ h; $T_2 = 99$ h; $\tau = 22$ h

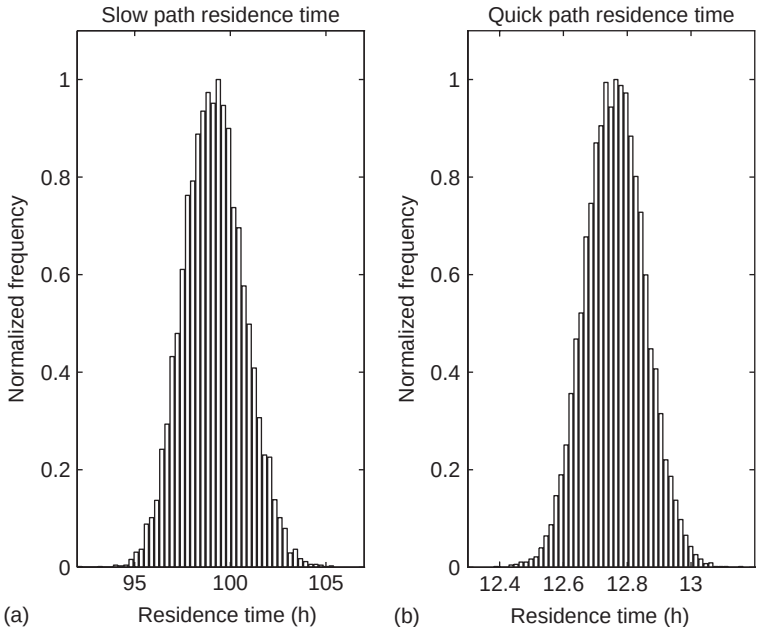


Figure 6 MCS analysis results: normalized histograms of the slow-flow (a) and quick-flow (b) pathway residence times

pathway and $5656\,m^3$ in the slow pathway. As a result, the total estimated AMV is $5656 + 3 \times 361 = 6739\,m^3$, which seems reasonable when compared with the $9749\,m^3$ for the total volume of the wetland, estimated by physical measurement. This suggests that about 70% of the wetland is important in dispersing the tracer (and, therefore, the waste water) and compares reasonably with the dispersive fraction derived percentage of 76%.

Of course, all of the results above are statistical estimates and so they are inherently uncertain. The advantage of the DBM/TF approach is that we can quantify and consider the consequences of this uncertainty (see Young, 1999a). For instance, based on the covariance matrix of the parameters produced by the SRIV estimation analysis, empirical probability distributions, in the form of histograms, can be computed for the "derived" physical parameters, such as the residence times, partition percentages, AMVs, total AMV, and steady state gain, using Monte Carlo Simulation (MCS) analysis. Figure 6 is a typical example of such analysis: it shows the normalized empirical distributions for the two residence times obtained by MCS using 10 000 random realizations (the procedure used here is discussed in Young, 1999a).

Of course, it should be noted that the parallel decomposition of the estimated TF used here is not unique: there are other decompositions that are just as valid and give precisely the same input–output response. For example, two other examples are: (i) a serial connection of the three ADZs with residence times of 12.8 h, in parallel with another serial connection, this time of an ADZ with residence time of 99 h in series with two 12.8 h residence time ADZs the same as those in the first parallel pathway; (ii) various decompositions including feedback processes. However, the latter seem less supportable in physical terms and are rejected according to the DBM ethos.

Finally, how can decompositions of ADZ reaches, such as those shown in Figure 5 be interpreted in terms of the wetland system? The most plausible mechanism is that the pure time delay allows for the flow-induced pure advection of the solute (chemical engineers would call this "plug-flow"). The quick-flow parallel pathway then represents the "main stream-flow" that is relatively unhindered by the vegetation; while the slow-flow pathway represents the solute that is captured by the heavy vegetation and so dispersed more widely and slowly before rejoining the main flow and eventually reaching the weir. It is this latter pathway, which we have shown above accounts for some 67% of the flow, together with the large associated dispersive fraction of 76%, that is most useful in terms of nutrient removal, since it allows more time for the biological activity to take place.

NONSTATIONARY AND NONLINEAR TF MODELS

Although the RF model discussed in Section "DISCRETE-TIME TF MODEL IDENTIFICATION AND ESTIMATION" is characterized by an input nonlinearity, the example has emphasized the constant parameter, linear TF part of the model (which defines the shape of the catchment UH). However, the nonlinearity is very important because

it defines the way in which the model responds under different catchment wetness conditions. So, in general, it is necessary to consider extensions to the constant parameter, linear TF model that allow for the consideration of more complex hydrological processes such as this. Fortunately, it is possible to extend TF models to handle "nonstationary" situations described by TVP models or "nonlinear" situations characterized by State-Dependent Parameter (SDP) models. The latter model class encompasses a wide variety of nonlinear, stochastic, dynamic phenomena, including chaotic systems.

Although such TVP and SDP models are beyond the scope of this article, they are introduced briefly in the following subsections. Both types of models can be considered in either continuous or discrete-time. However, since statistical estimation is more straightforward in the DT case, it often provides the more practical approach. Consequently, the brief descriptions that follow are limited to this DT situation.

Nonstationary Time-Variable Parameter (TVP) Models

In the DT case, the TVP form of the TF model (11) can be written as follows:

$$x_k = \frac{B(k, z^{-1})}{A(k, z^{-1})} u_{k-\delta} \quad y_k = x_k + \xi_k \tag{18}$$

where,

$$A(k, z^{-1}) = 1 + a_{1,k} z^{-1} + a_{2,k} z^{-2} + \cdots + a_{n,k} z^{-n}$$
$$B(k, z^{-1}) = b_{0,k} + b_{1,k} z^{-1} + b_{2,k} z^{-2} + \cdots + b_{m,k} z^{-m}$$

Here, all the parameters are assumed to be functions of the time index k, that is it is assumed that they may vary over time in an unknown manner that needs to be estimated from the data.

Some of the latest research on TVP estimation is reported in Young (1999b, 2000, 2004b) where the reader will find a complete description of the recursive estimation algorithms that allow for the estimation of the time-variable parameters. A simple example would be a model such as (8) in which the "gain" parameter $b_0 = b_{0,k}$ is assumed to vary over time in order to reflect any changes in the dynamic behavior of the channel that may occur over a wide range of flow conditions. An approach similar to this is suggested by Cluckie (1993), who refers to the gain as a "model scaling factor" and points out that it is analogous to the "variable proportional loss method" of defining effective rainfall. An example of adaptive gain adjustment that has been implemented in practice is the adaptive flood warning system for the River Nith at Dumfries, in south west Scotland (Lees *et al.*, 1993, 1994). Here, a recursive estimation algorithm has been used for over a decade to continually update a gain parameter, based on telemetered data from the catchment, in order to provide an adaptive capability. In this case, however, the adaptive gain parameter is only required to "fine-tune" an *already computed* effective rainfall input generated from the measured rainfall by a "state-dependent parameter" nonlinearity, as discussed in the next section.

Nonlinear State-Dependent Parameter (SDP) Models

Again, in the discrete-time case, the SDP form of the TF model (11) can be written as follows:

$$x_k = \frac{B(\mathbf{w}_k, z^{-1})}{A(\mathbf{v}_k, z^{-1})} u_{k-\delta} \quad y_k = x_k + \xi_k \tag{19}$$

where,

$$A(\mathbf{v}_k, z^{-1}) = 1 + a_1(v_{1,k}) z^{-1} + a_2(v_{2,k}) z^{-2} + \cdots + a_n(v_{n,k}) z^{-n}$$
$$B(\mathbf{w}_k, z^{-1}) = b_0(w_{0,k}) + b_1(w_{1,k}) z^{-1} + b_2(w_{2,k}) z^{-2} + \cdots + b_m(w_{m,k}) z^{-m}$$

Here, the possibility that the parameters may be functions of other "state" variables is investigated. In other words, it is assumed that any parameter in the model may vary over time as a function of one or more other variables (e.g. the input u_k or output y_k and their past values), so introducing nonlinear behavior into the model. In the SDP model (19), these variables are denoted by $v_{i,k}, i = 1, 2, \ldots, n$ and $w_{j,k}, j = 0, 1, \ldots, m$ and they are, respectively, the elements of the vectors $\mathbf{v}_k$ and $\mathbf{w}_k$.

The first applications of SDP estimation in hydrology are described in Young (1993); Lees *et al.* (1993, 1994) and Young and Beven (1994); while the latest research on this topic is reported in Young (2000, 2001a) and Young (2002). A simple example of SDP estimation is provided by the estimation of the effective rainfall nonlinearity (15) of the RF model discussed in Section "DISCRETE-TIME TF MODEL IDENTIFICATION AND ESTIMATION". Here, SDP estimation identifies that, in order to explain the RF data well, the numerator parameters b_0, b_1 and b_2 in a *linear* TF model of the RF data need to vary significantly as functions of the flow: that is, they are identified as SDPs $b_0(y_k)$, $b_1(y_k)$ and $b_2(y_k)$. However, the inclusion of a power law nonlinearity such as equation (15), operating on the input rainfall, removes the need for these variations, leaving the model as a serial connection of the nonlinearity (15) and the linear, constant parameter TF, as described in Section "Discrete-time TF model identification and estimation". Other related examples have been discussed recently in Young (2001b,c, 2002a, 2003).

CONCLUSIONS

At one level, TF models are simply elegant and convenient representations of continuous-time, stochastic differential equation models, or their discrete-time equivalents. This article demonstrates, however, that they are much more than this: they also provide a powerful, generic method of modeling hydrological systems. This method provides a useful block diagram interpretation of the input–output dynamics of the system under study that can be decomposed straightforwardly into physically meaningful sub-systems that are in an ideal form for DBM modeling studies. The utility of the method is further enhanced by the powerful methods of statistical identification and parameter estimation that have been developed for such TF models over many years. These are not only robust in practical application, but they also provide valuable information about the stochastic nature of the system and allow for the quantification of the uncertainty associated with the model parameter estimates and its output predictions. Moreover, these estimation methods are normally available in a recursive form that allows for their use in on-line, real-time applications, where they provide a natural basis for adaptive modeling and data assimilation. Recent developments in recursive time-variable and state-dependent parameter estimation have enhanced the utility of TF models still further, so that they are now able to characterize both nonstationary and nonlinear stochastic systems.

Finally, it should be emphasized that DBM models, in general, and data-based TF models, in particular, naturally require real time-series data for their identification and estimation. Also, their primary use is for improving our understanding of the system dynamics and/or for implementing model-based prediction, as in flood warning and forecasting applications, where the main changes that are likely to occur in future relate to the input variables (measured rainfall and upstream flows). They are not particularly appropriate, at least in their basic form, for assessing the impacts of changes to the nature of the system ("what-if" studies) unless the derived, physically meaningful parameters associated with the DBM model (as, for example, in the case of the Florida wetland example) can be associated with physical environmental characteristics, such as the topography or other physical features of a catchment ("regionalization studies": see e.g. Kokkonen *et al.*, 2003 and the **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3** in this Encyclopedia). In this regard, it is likely that the derived parameters will be "bulk" parameters that relate to the behavior of the system at the catchment scale, rather than the kind of "process" parameters used in hydrological simulation models.

It is clear, however, that when suitable, measured time-series data are available, DBM/TF models should always be used as an adjunct to simulation model synthesis because they define well the "dominant mode" behavior of the system that is identifiable from the historical data (*see* **Chapter 134, Downward Approach to Hydrological Model Development, Volume 3** in this Encyclopedia). In this way, they can alert the model builder to those aspects of the simulation model that are not well defined in relation to the available data and so are of a more speculative nature.

Acknowledgments

The author is grateful to Professor Tony Jakeman of the Centre for Resource and Environmental Studies, Australian National University for providing the Canning River data; and to Chris Martinez and Dr. William R. Wise of the Environmental Engineering Sciences Department, University of Florida, for providing the tracer experiment data.

FURTHER READING

O'Donnell T (1986) Deterministic catchment modelling. In *River Flow Modelling and Forecasting*, Kraijenhoff D.A and Moll J.R (Eds.), D. Reidel: Dordrecht, pp. 11–37.

Young P.C, McKenna P and Bruun J (2001) Identification of nonlinear stochastic systems by state dependent parameter estimation. *International Journal of Control*, **74**, 1837–1857.

Young P.C, Parkinson S.D and Lees M (1996) Simplicity out of complexity in environmental systems: Occam's Razor revisited. *Journal of Applied Statistics*, **23**, 165–210.

Young P.C and Pedregal D (1999) Recursive and en-bloc approaches to signal extraction. *Journal of Applied Statistics*, **26**, 103–128.

Young P.C and Tomlin C.M (2000) Data-based mechanistic modelling and adaptive flow forecasting. In *Flood Forecasting: What Does Current Research Offer the Practitioner?* BHS Occasional paper No 12, Lees M.J and Walsh P (Eds.), Centre for Ecology and Hydrology on behalf of the British Hydrological Society: pp. 26–40.

REFERENCES

Akaike H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **ACl9**, 716–722.

Anderson M.G. and Burt T.P. (Eds.) (1985) *Hydrological Forecasting*, John Wiley: Chichester.

Amorocho J. and Hart W.E. (1964) A critique of current methods of hydrologic systems investigation. *Transactions of the American Geophysical Union*, **45**, 307–321.

Beck M.B. and Young P.C. (1975) A dynamic model for DO-BOD relationships in a non-tidal stream. *Water Research*, **9**, 769–776.

Beven K.J. (2001) *Rainfall-Runoff Modelling: the Primer*, John Wiley and Sons: Chichester.

Box G.E.P. and Jenkins G.M. (1970) *Time-series Analysis, Forecasting and Control*, Holden-Day: San Francisco.

Bras R.L. and Rodriguez-Iturbe I. (1985) *Random Functions and Hydrology*, Addison-Wesley: Reading.

Cluckie I.D. (1993) Real-time flood forecasting using weather radar. In *Concise Encyclopedia of Environmental Systems*, Young P.C. (Ed.), Pergamon Press: Oxford, pp. 291–298.

Dooge J.C.I. (1959) A general theory of the unit hydrograph. *Journal of Geophysical Research*, **64**, 241–256.

Dooge J.C.I. (1986) Theory of flood routing. In *River Flow Modelling and Forecasting*, Kraijenhoff D.A. and Moll J.R. (Eds.), D. Reidel: Dordrecht, pp. 39–65.

Dooge J.C.I. and O'Kane J.P. (2003) *Deterministic Methods in Systems Hydrology*, Balkema: Lisse.

Gupta V.K. (1984) The Identification of Conceptual Watershed Models, Ph.D. Dissertation, Case Western Reserve University, Cleveland.

Han D. (1991) Weather Radar Information Processing and Real-Time Flood Forecasting, Ph. D. thesis, Department of Civil Engineering, University of Salford.

Hsu K., Gupta H.V. and Sorooshian S. (1993) Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, **29**, 1185–1194.

Jang J.S.R., Sun C.T. and Mizutani E. (1997) *Neuro-Fuzzy and Soft Computing*, Prentice Hall: NJ.

Jakeman A.J. and Young P.C. (1981) On the decoupling of system and noise model parameter estimation in time-series analysis. *International Journal of Control*, **34**, 423–431.

Jakeman A.J. and Young P.C. (1983) Advanced methods of recursive time-series analysis. *International Journal of Control*, **37**, 1291–1310.

Jakeman A.J., Littlewood I.G. and Whitehead P.G. (1990) Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology*, **117**, 275–300.

Kokkonen T.S., Jakeman A.J., Young P.C. and Koivusalo H.J. (2003) Predicting daily flows in ungauged catchments model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. *Hydrological Processes*, **17**, 2219–2238.

Kraijenhoff D.A. and Moll J.R. (1986) *River Flow Modelling and Forecasting*, D. Reidel: Dordrecht.

Lees M.J. (2000a) Advances in transfer function based flood forecasting. In *Flood Forecasting: What Does Current Research Offer the Practitioner?* Lees M.J. and Walsh P. (Eds.), BHS Occasional paper No 12, Centre for Ecology and Hydrology on behalf of the British Hydrological Society: pp. 41–55.

Lees M.J. (2000b) Data-based mechanistic modelling and forecasting of hydrological systems. *Journal of Hydroinformatics*, **2**, 15–34.

Lees M., Young P.C. and Ferguson S. (1993) Adaptive flood warning. In *Concise Encyclopedia of Environmental Systems*, Young P.C. (Ed.), Pergamon Press: Oxford, pp. 234–236.

Lees M., Young P.C., Beven K.J., Ferguson S. and Burns J. (1994) An adaptive flood warning system for the River Nith at Dumfries. In *River Flood Hydraulics*, White W.R. and Watts J. (Eds.), Institute of Hydrology: Wallingford, pp. 65–75.

Ljung L. and Söderstrom T. (1983) *Theory and Practice of Recursive Estimation*, MIT Press: Cambridge.

Meijer H. (1941) Simplified flood routing. *Civil Engineering*, **11**, 306–307.

Nash J.E. (1959) Systematic determination of unit hydrograph parameters. *Journal of Geophysical Research*, **64**, 111–115.

Nash J.E. and Sutcliffe J.V. (1970) River flow forecasting through conceptual models: discussion of principles. *Journal of Hydrology*, **10**, 282–290.

Natale L. and Todini E. (1976a) A stable estimator for linear models: 1. Theoretical development and Monte-Carlo experiments. *Water Resources Research*, **12**(4), 667–671.

Natale L. and Todini E. (1976b) A stable estimator for linear models: 2. real world hydrologic applications. *Water Resources Research*, **12**(4), 672–675.

Natale L. and Todini E. (1976) A stable estimator for linear models: 2. real world hydrologic applications. *Water Resources Research*, **12**, 672–676.

Norton J.P. (1986) *An Introduction to Identification*, Academic Press: London.

O'Connell P.E. and Clarke R.T. (1981) Adaptive hydrological forecasting–a review. *Hydrological Sciences Bulletin*, **26**, 179–205.

Pierce D.A. (1972) Least squares estimation in dynamic disturbance time-series models. *Biometrika*, **59**, 73–78.

Salas J.D., Delleur J.W., Yevjevich V. and Lane W.L. (1980) *Applied Modeling of Hydrologic Time Series*, Water Resources Publications: Littleton.

Schwarzenbach J. and Gill K.F. (1979) *Systems Modelling and Control*, Edward Arnold: London.

Shaw E.M. (1994) *Hydrology in Practice*, *Third Edition*, Chapman & Hall: London.

Singh V.P. (1988) *Hydrologic Systems: Rainfall-Runoff Modeling*, Vol. 1, Prentice-Hall: Englewood Cliffs.

Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Colorado: Water Resources Publications.

Sorooshian S. (1983) Surface water hydrology: on-line estimation. *Review of Geophysics and Space Physics*, **21**, 706–721.

Wallis S.G., Young P.C. and Beven K.J. (1989) Experimental investigation of the aggregated dead zone model for longitudinal solute transport in stream channels. *Proceedings of the Institution of Civil Engineers, Part 2*, **87**, 1–22.

Wellstead P.E. (1978) An instrumental product moment test for model order estimation. *Automatica*, **14**, 89–91.

Wheater H.S., Jakeman A.J. and Beven K.J. (1993) Progress and directions in rainfall-run-off modelling. In *Modelling Change in Environmental Systems*, Jakeman A.J., Beck M.B. and McAleer M.J. (Eds.), Wiley: Chichester, pp. 101–132.

Whitehead P.G. and Young P.C. (1975) A dynamic-stochastic model for water quality in part of the Bedford-Ouse River system. In *Computer Simulation of Water Resources Systems*, Vansteenkiste G.C. (Ed.), North-Holland: Amsterdam, pp. 417–438.

Whitehead P.G., Young P.C. and Hornberger G.H. (1976) A systems model of stream flow and water quality in the Bedford Ouse River – Part l: stream flow modeling *Water Research*, **13**, 1155–1169.

Wood E.F. (Ed.) (1980) *Workshop on Real Time Forecasting/Control of Water Resource Systems*, Pergamon Press: New York.

Young P.C. (1974) Recursive approaches to time-Series analysis. *Bulletin of the Maths and its Applications*, **10**, 209–224.

Young P.C. (1984) *Recursive Estimation and Time-series Analysis*, Springer-Verlag: Berlin.

Young P.C. (1985) The instrumental variable method: a practical approach to identification and system parameter estimation. In *Identification and System Parameter Estimation*, Barker H.A. and Young P.C. (Eds.), Pergamon Press: Oxford, pp. 1–16.

Young P.C. (1986) Time-series methods and recursive estimation in hydrological systems analysis. In *River Flow Modelling and Forecasting*, Kraijenhoff D.A. and Moll J.R. (Eds.), D. Reidel: Dordrecht, Chap. 6, pp. 129–180.

Young P.C. (1989) Recursive estimation, forecasting and adaptive control. In *Control and Dynamic Systems: Advances in Theory and Applications*, Vol. 30, Leondes C.T. (Ed.), Academic Press: San Diego, Chap. 6, pp. 119–166.

Young P.C. (1992) Parallel processes in hydrology and water quality: a unified time series approach. *Journal of the Institute of Water and Environmental Management*, **6**, 598–612.

Young P.C. (1993) Time variable and state dependent modelling of nonstationary and nonlinear time series. In *Developments in Time Series Analysis*, Subba Rao T. (Ed.), Chapman & Hall: London, pp. 374–413.

Young P.C. (1998) Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. *Environmental Modelling and Software*, **13**, 105–122.

Young P.C. (1999a) Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis. *Computer Physics Communications*, **117**, 113–129.

Young P.C. (1999b) Nonstationary time series analysis and forecasting. *Progress in Environmental Science*, **1**, 3–48.

Young P.C. (2000) Stochastic, dynamic modelling and signal processing: time variable and state dependent parameter estimation. In *Nonstationary and Nonlinear Signal Processing*, Fitzgerald W.J., Walden A., Smith R. and Young P.C. (Eds.), Cambridge University Press: Cambridge, pp. 74–114.

Young P.C. (2001a) The identification and estimation of nonlinear stochastic systems. In *Nonlinear Dynamics and Statistics*, Mees A.I. (Ed.), Birkhauser: Boston, pp. 127–166.

Young P.C. (2001b) Data-based mechanistic modelling and validation of rainfall-flow processes. In *Model Validation: Perspectives in Hydrological Science*, Anderson M.G. and Bates P.D. (Eds.), John Wiley: Chichester, pp. 117–161.

Young P.C. (2001c) Advances in Real-time Forecasting. Centre for Research on Environmental Systems and Statistics, Report No. TR/176, Lancaster University, p. 38.

Young P.C. (2002a) Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*, **360**, 1433–1450.

Young P.C. (2002b) Comments on the estimation of continuous-time transfer functions. *International Journal of Control*, **75**, 693–697.

Young P.C. (2003) Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale. *Hydrological Processes*, **17**, 2195–2217.

Young P.C. (2004a) Identification and estimation of continuous-time hydrological models from discrete-time data, *Proceedings, British Hydrological Society BHS2004 Conference*, Imperial College, London.

Young P.C. (2004b) Identification of time varying systems, *Encyclopedia of Life Support Systems (EOLSS)*, UNESCO-EOLSS: (**www.eolss.net**).

Young P.C. and Beven K.J. (1994) Data-based mechanistic modelling and the rainfall-flow nonlinearity. *Environmetrics*, **5**, 335–363.

Young P.C. and Jakeman A.J. (1979) Refined instrumental variable methods of recursive time-series analysis: Parts I, single-input, single-output systems. *International Journal of Control*, **29**, 1–30.

Young P.C. and Jakeman A.J. (1980) Refined instrumental variable methods of recursive time-series analysis: Part III, extensions. *International Journal of Control*, **31**, 741–764.

Young P.C. and Lees M. (1993) The Active Mixing Volume (AMV): a new concept in modelling environmental systems. In *Statistics for the Environment*, Barnett V. and Turkman K.F. (Eds.), John Wiley: Chichester, Chap. 1, pp. 3–44.

Young P.C. and Wallis S.G. (1985) Recursive estimation: a unified approach to identification, estimation and forecasting of hydrological systems. *Applied Mathematics and Computations*, **17**, 299–334.

Young P.C. and Wallis S.G. (1994) Solute transport and dispersion in channels. In *Channel Networks*, Beven K.J. and Kirby M.J. (Eds.), John Wiley: Chichester, Chap. 6, pp. 129–173.

Young P.C., Jakeman A.J. and McMurtrie R. (1980) An instrumental variable method for model order identification. *Automatica*, **16**, 281–294.

Young P.C., Jakeman A.J. and Post D.A. (1997) Recent advances in data-based modelling and analysis of hydrological systems. *Water Science and Technology*, **36**, 99–116.

129: Rainfall-runoff Modeling for Integrated Basin Management

GEORGE LEAVESLEY

United States Geological Survey, Denver, CO, US

Integrated basin management is concerned with the interactions of physical, ecological, economic, and social systems as they affect the operation, planning, and policy making processes inherent in the management of land and water resources. Systems of integrated hydrological, chemical, biological, ecological, and socioeconomic models are typically used to assess the effects of proposed management alternatives on basin resources, or to manage basin resources in real time. Water is a common thread linking many of the components among these models. The ability to adequately simulate rainfall-runoff processes and their interactions with processes related to other system components significantly affects the integrated system results. Model complexity and compatibility, data availability, and a number of error sources that include data, model parameters, and model structure, are major concerns in the use of integrated modeling systems. The effects of these error sources on the uncertainty in simulation results become even more complex in integrated system application in which output from one model is typically used as input to other models. The number and complexity of available integrated modeling systems represent a wide variety of approaches with different models, operational features, user interfaces, data formats, and analytical tools. For this review, integrated modeling systems have been characterized as using either a (i) dedicated-system approach with a predefined set of models or (ii) modeling-framework approach that typically enables the application of a user-selected set of models. A number of available integrated systems within these two classifications are described and an example of application of a modular-framework approach is presented.

Integrated basin management can be described as a coordinated form of management that considers the interactions of physical, ecological, economic, and social systems as they affect the operation, planning, and policy making processes inherent in the management of land and water resources (*see* **Chapter 185, Integrated Land and Water Resources Management, Volume 5**). Integrated basin management issues are the concern of local, national, and international organizations. In the United States, integrated basin management has been advanced in something termed the *watershed approach* to environmental management by the federal government's Environmental Protection Agency (EPA, 2005). In Europe, the European Union's Water Framework Directive directs member states to begin implementing the principles of sustained development and integrated river basin management (European Commission, 2004). Comparable efforts have been, or are being initiated in a large number of river basins around the world.

Water is a common thread linking many of the issues in integrated basin management. The ability to predict the results of existing or proposed actions on all components of a basin system is dependent to a large degree on the ability to adequately simulate rainfall-runoff processes and their interactions with processes related to other system components. This entails linking hydrological models with a variety of chemical (*see* **Chapter 100, Water Quality Modeling, Volume 3**), biological, ecological (*see* **Chapter 103, Terrestrial Ecosystems, Volume 3**; **Chapter 107, Natural and Constructed Wetlands, Volume 3**; **Chapter 108, Lake Ecosystems (Stratification and Seasonal Mixing Processes, Pelagic and Benthic Coupling), Volume 3**), and socioeconomic models. The development and application of these types of models for integrated basin

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

management has been the topic of a number of conferences and symposia (e.g. Simonovic *et al.*, 1995; Schumann *et al.*, 2001; Hays and McKee, 2001; Marino and Simonovic, 2001).

Systems of integrated models are generally used to address "what if" questions, such as the effects of proposed management alternatives on basin resources, or to manage basin resources in real time. Most integrated system applications to date have been for the assessment of resource-management alternatives or the potential effects of natural disasters such as fire, floods, or droughts. However, where real-time meteorological, streamflow, and other natural resource data are available, integrated systems can be applied for both management-alternative assessment and real-time resource management. Real-time applications have the advantage of continuous measures of model performance. However, this also necessitates identifying the sources of simulation errors and making the appropriate corrections or adjustments to the input data or/and model state variables to keep the model results consistent with observations.

Error sources include data, model parameters, and model structure. Integrated modeling applications are often on large basins where many types of data may have limited availability. Such data limitations can constrain or adversely affect the ability to adequately calibrate parameters for some models (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**; **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**). In some basins, an absence of data for some processes may prevent calibration and limit parameter determination to estimation from related physical measures or expert opinion (*see* **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3**). The effects of these error sources on the uncertainty in simulation results are always a concern in the application of individual models to the decision making process (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**). The issue of uncertainty becomes even more complex, however, in the integrated system application where output from one model is typically used as input to other models. The development of methods and tools to provide measures of the uncertainty resulting from the cascade of errors through multiple models is an area of current research.

The linking of models and/or model processes can be accomplished using either a loosely or tightly coupled approach. Loosely coupled models provide for information flow in only one direction. Models are run sequentially with the output of one model providing the input to the next model. Loosely coupled models are most appropriate for applications where there is no need to consider feedback among the models and processes being coupled. Tightly coupled models provide for information flow in two directions and support the simulation of feedback between or among processes. The number and complexity of coupled processes range among models, but can include the integration of surface water, groundwater, sediment, nutrient, pesticide, and biological processes.

The coupling of models to build an integrated modeling system is not a trivial task. Many individual models have been developed for problem specific and/or site-specific applications. They have been constructed using a wide variety of operational features, user interfaces, data formats, analytical tools, and documentation standards. To overcome some of these complexities of integrating such a diverse set of models and tools, a variety of modular strategies have been developed. These modular approaches support the linking of loosely and tightly coupled models and model processes into integrated modeling systems for basin management applications. Most integrated modeling systems can be characterized as using either a (i) dedicated-system approach with a predefined set of models or (ii) modeling-framework approach that typically enables the application of a user-selected set of models.

DEDICATED SYSTEMS

Dedicated systems typically have monolithic-type architecture where process algorithms, individual models, data management components, and analytical tools are hard-wired together into a comprehensive modeling system. These systems are usually developed for specific types of application and are limited in scope to the process models and tools included in the system.

Stand-alone watershed models that couple rainfall-runoff processes with chemical and ecological processes represent the most basic form of the dedicated systems. Some of the more comprehensive models include the Hydrologic Simulation Program – Fortran (HSPF) (Bicknell *et al.*, 1997), the Soil and Water Assessment Tool (SWAT) (Arnold *et al.*, 1998), the Soil and Water Integrated Model (SWIM) (Krysanova *et al.*, 1998), and the Integrated Catchments (INCA) model (Whitehead *et al.*, 1998). All are river basin scale models whose major components include hydrology, weather, sedimentation, nutrients, and agricultural management. HSPF and SWAT also include a pesticide component.

Differences among stand-alone models are primarily a function of the number and degree of complexity of process conceptualizations included in the model. The more complex models are typically those with fully coupled surface/subsurface modeling capabilities. SHETRAN (Parkin *et al.*, 2000) for example, is a 3D, coupled surface/subsurface, physically based, spatially distributed,

finite-difference model for coupled water flow, multifraction sediment transport and multiple, reactive solute transport in river basins.

The functionality and complexity of the existing models can be increased by adding or replacing selected components. MODFLOW (McDonald and Harbaugh, 1988; Harbaugh and McDonald, 1996) has been used to replace simpler groundwater components in SWAT (Sophocleous and Perkins, 2000) and in the Precipitation–Runoff Modeling System (PRMS) (Markstrom *et al.*, 2002) to provide a more physically based, fully distributed, groundwater component in each of these models.

Some dedicated modeling systems allow the user to select the degree of complexity by selecting process conceptualizations from a module library to construct a model of appropriate complexity. Selection is limited, however, to the modules provided. The MIKE SHE modeling system (Danish Hydraulics Institute, 2005) offers the choice of several different approaches ranging from simple, lumped, and conceptual to distributed and physically based. In its most complex configuration, it provides a coupled surface/groundwater system using a 2D or 3D finite-difference groundwater model that is similar to MODFLOW. In a future release, it will include MODFLOW as an option.

In contrast to the single model approach, some dedicated systems use a multiple, stand-alone model approach. These systems provide a common graphical user interface, database structure, and a set of analysis tools in support of a set of selected models. Links between or among models are accomplished using the common database and models may be loosely coupled in most systems. The US Environmental Protection Agency's (EPA's) Better Assessment Science Integrating Point and Nonpoint Sources (BASINS) (U.S. Environmental Protection Agency, 2001a) software system integrates a GIS package, meteorological and digital data coverages of the United States, and the models HSPF, SWAT, QUAL2E (Barnwell and Brown, 1987; U.S. Environmental Protection Agency, 1995) and PLOAD (U.S. Environmental Protection Agency, 2001b) into a single package. BASINS was developed to support the assessment of total maximum daily loads (TMDLs), a required analysis for all river systems in the US. The Watershed Modeling System (WMS) (Environmental Modeling Systems, 2005) is a dedicated modeling environment that provides tools to support HEC-1 (U.S. Army, 1998), HEC-HMS (U.S. Army, 2001), TR-20 (U.S. Natural Resource Conservation Service, 1992), TR-55 (U.S. Natural Resource Conservation Service, 1986), and HSPF.

FRAMEWORK MODELING SYSTEMS

The second approach to integrated modeling system development is the modeling framework. The framework provides a flexible platform in which to integrate a variety of models and tools of the user's choice. A library of modules, models, and data management and analytical tools is normally provided. When selected library components are not available, options are typically provided for adding new components to the framework library. In theory, the applicability of the modeling-framework approach is limited only by the models and process modules that can be incorporated into the framework library.

In some frameworks, existing models are considered to be individual modules and are incorporated into the framework with little or no modification. In other frameworks, existing models are disaggregated into their process components, which then become process modules in the system library. Considerations in the selection and linking of process modules to build a model include problem objectives, types of data available, and spatial and temporal scales of application. These issues must also be considered in the selection of individual models to be coupled or linked within the framework.

Another consideration in module and model selection is the level of complexity in process simulation needed to address the desired management objectives. For example, the simulation of runoff volume and timing may be accomplished using a relatively simple modeling approach. However, if the simulation of water-quality constituents in the runoff is also an issue, a more process-based watershed model may be needed. This might be a model that considers flow paths, residence times, and biogeochemical processes, in order to provide the level of detail in process simulation and output needed by other models selected within the framework.

One of the earlier modular frameworks was the U.S. Geological Survey (USGS) Modular Modeling System (MMS) (Leavesley *et al.*, 1996a). MMS supports the development of models using modules from a system library and the coupling of multiple models for application to complex, multidisciplinary problems. A major limitation of MMS is that, given its age, it is object-based and not fully object oriented. To address this limitation, the concepts of MMS have been used to create the Object Modeling System (OMS) (U.S. Department of Agriculture, 2003) which is a fully object-oriented framework. OMS is a Java-based system that uses the NetBeans IDE as its framework platform. A Java wrapper is used to integrate model and module code in other programming languages.

Other framework development efforts in the United States include the Framework for Risk Analysis in Multimedia Environmental Systems (FRAMES) (Whelan *et al.*, 1997) and the Dynamic Information Architecture System (DIAS) (Sydelko *et al.*, 2001). FRAMES supports the linking of a variety of existing models for use in assessing the impacts of hazardous and radioactive releases on the environment. DIAS is an object-oriented, Java-based framework

that supports the coupling of complete models and/or process modules to build integrated models for a wide range of environmental management problems. DIAS also uses a Java-wrapper approach to integrate models and modules in other programming languages.

Comparable framework developments are also being conducted in Australia. The Interactive Component Modeling System (ICMS) is an object-oriented framework that supports model integration and the development of custom interfaces for each application (Reed *et al.*, 1999). Tarsier (Watson *et al.*, 2001) is an object-oriented framework that supports a broader range of models and data handling and analysis tools than ICMS but is limited in that it only supports model code developed in C++. A collaborative effort is being conducted to integrate ICMS and Tarsier to build a more comprehensive framework called the *Catchment Modeling Toolkit* (Argent and Vertessy, 2002).

In Europe, the Generic Framework (GF) (Blind *et al.*, 2001) is an open framework for linking models, which is being developed by a large consortium of groups in the Netherlands. The basic design principles of the GF focus on a modular approach to separate model, tool, and data components. The Open Modeling System (**`http://www.dutch-oms.org/intro.html`**) is another framework being developed in the Netherlands to support 2D and 3D simulation of flow and transport processes. The Open Modeling System architecture is designed to support data exchange and synchronization between legacy codes. Current development is focused on the integration of the Delft3D system (Delft Hydraulics, 2003) and the SIMONA system (National Institute for Coastal and Marine Management, 2000).

With the proliferation of modeling frameworks, some of the original difficulties in the linking of individual models, as noted above, have also been recognized in attempts to share modules, tools, and other resources among frameworks. In an attempt to resolve these problems, coordinated efforts by collaborative groups of research and operational agencies have been initiated to develop more generic frameworks and standards. In the United States, a Memorandum of Understanding (MOU) has been signed by 10 federal agencies to establish a framework for facilitating cooperation and coordination in research and development of multimedia environmental models and related tools and databases, and to facilitate their application to human and environmental health risk assessment (**`http://www.iscmem.org`**).

In Europe, support of the integrated basin management requirements of the EU Water Framework Directive has spawned a number of group efforts. The HarmonIT project is a collaborative effort among 14 government, university, and private organizations to develop and implement a European Open Modeling Interface and Environment (OpenMI), which will simplify the linking of models and their application to integrated basin management (Gijsbers *et al.*, 2002). As part of this effort, a state-of-the-art review was conducted to evaluate existing capabilities in linking models and resolving issues within the project (Hutchings, 2002). In a related effort, the Benchmark Models for the Water Framework Directive project is by a group of eight organizations that are developing a scheme for selecting and benchmarking existing model systems for integrated basin management applications (Finnish Environmental Administration, 2002).

APPLICATIONS

Integrated systems are applied at basin scales ranging from a few tens of square kilometers to several thousands of square kilometers. One can enter the name of almost any major river basin in the world into a web search tool, along with key words such as "integrated basin modeling", and find that one or several local, national, and/or international organizations are supporting the development and application of some type of integrated modeling and decision support system. The small number of these that are currently operational, as compared with the total number of systems being developed, indicates the number of challenges and limitations in the development and application of such systems. These include limited data, process knowledge, financial resources, consensus on approach, and social and political agreement on resource allocation. However, the large number of systems in development also indicates the recognition of the need for objective tools to address water and environmental resource allocation problems and a willingness to use integrated basin management systems to address these problems.

A sample application of an integrated modeling system that is operational is the Watershed and River System Management Program (WARSMP) (Leavesley *et al.*, 1996b, **`http://wwwbrr.cr.usgs.gov/warsmp`**; **`http://www.usbr.gov/pmts/rivers/warsmp`**). The WARSMP is a cooperative effort between the USGS and the Bureau of Reclamation (BOR) to develop an operational, database-centered, decision support system for application to complex water and environmental resource-management issues. The USGS MMS has been coupled with the BOR RiverWare software (Zagona *et al.*, 2001; **`http://cadswes.colorado.edu/riverware`**) using a shared relational database. RiverWare is an object-oriented reservoir and river-systems modeling framework developed to provide tools for evaluating and applying optimal water allocation and management strategies.

The WARSMP is operational on the Gunnison River Basin in Colorado, where MMS provides forecasts of daily inflows from 15 headwater basins to a series of three reservoirs (Leavesley *et al.*, 2002). RiverWare uses these inflow

forecasts to manage the reservoirs for power generation, flood control, irrigation, fisheries, and recreation. Streamflow forecasts are made using the Ensemble Streamflow Prediction methodology (ESP) (Day, 1985). Snowmelt is the dominant source of streamflow in the Gunnison Basin and so decisions on reservoir storage levels and releases rely on the forecasts of the volume and timing of snow accumulation and melt. Optimal operation of the Gunnison Basin maximizes power generation, irrigation water volume, and environmental water uses, while minimizing flooding below the reservoirs. Use of the WARSMP decision support tools, as compared with the historic management approach, was evaluated by Ryan (1996). His evaluation demonstrated an increase in power generation revenues and decrease in flooding as a result of the use of these tools.

The WARSMP tools are also being applied to the Yakima River Basin in Washington (Maston and Vaccaro, 2002; U.S. Geological Survey, 1998; **http://wa.water.usgs.gov/projects/yakimawarsmp**); the Rio Grande Basin in Colorado and New Mexico; and the Truckee River Basin in California and Nevada. Integrated water-management issues in these basins include efficiency of water-resources management, environmental concerns such as meeting flow needs for endangered species, groundwater/surfacewater interactions, water-quality issues related to irrigated agriculture, and optimizing operations within the constraints of multiple objectives such as power generation, irrigation, and water conservation.

REFERENCES

Argent R. and Vertessy R. (2002) *Catchment Modeling Toolkit*, Bulletin No. 2. **http://www.toolkit.net.au/**.

Arnold J.G., Srinivasin R., Muttiah R.S. and Williams J.R. (1998) Large area hydrologic modeling and assessment: part I. Model development. *Journal of the American Water Resources Association*, **34**(1), 73–89.

Barnwell, T.O. and Brown, L.C., (1987) *The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS: Documentation and User Manual*, EPA/600/3-87/007.

Bicknell B.R., Imhoff J.C., Kittle J.L. Jr, Donigian A.S. Jr and Johanson R.C. (1997) *Hydrological Simulation Program–Fortran: User's manual for version 11*, U.S. Environmental Protection Agency, National Exposure Research Laboratory, Athens, EPA/600/R-97/080.

Blind, M.W., Wentholt, L., van Adrichem, B. and Groenendijk, P. (2001) *The Generic Framework – An open framework for model linkage and rapid decision support system development*, presented at the ModSim 2001 conference, Canberra, Australia. Also available on the World Wide Web: **http://www.genericframework.org/uk/techdocu.htm**.

Danish Hydraulics Institute (2005) **http://www.dhisoftware.com/mikeshe**.

Day G.N. (1985) Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management, ASCE*, **111**, 157–170.

Delft Hydraulics (2003) **http://www.wldelft.nl/soft/d3d/**.

Environmental Modeling Systems (2005) WMS 7.1 Overview, **http://www.ems-i.com/WMS/WMS_Overview/wms_overview.html**.

European Commission (2004) **http://europa.eu.int/comm/environment/water/water-framework/index_en.html**.

Finnish Environmental Administration (2002) **http://www.vyh.fi/eng/research/euproj/bmw/homepage.htm**.

Gijsbers, P.J.A., Moore, R.V. and Tindall, C.I. (2002) *HarmonIT: Towards OMI, an open modeling interface and environment to harmonise European developments in water related simulation software*. Hydroinformatics 2002 Conference., **http://www.harmonit.org/docs/hic2002_harmonitpaper.pdf**.

Harbaugh A.W. and McDonald M.G. (1996) User's documentation for MODFLOW-96, an update to the U.S. Geological Survey modular finite-difference ground-water flow model, U.S. Geological Survey: *U.S. Geological Survey Open-File Report 96-485*.

Hays, D.F. and McKee, M. (2001) *Decision Support Systems for Water Resources Management*, Proceedings American Water Resources Association Specialty Conference, Snowbird, Utah, June 27–30, 2001.

Hutchings, C. (2002) State of the art review. HR Wallingford Ltd., Report SR 598, **http://www.harmonit.org/docs/repu_01_09_soa_review_approved.pdf**.

Krysanova V., Becker A. and Mueller-Wohlfeil D.I. (1998) Development and testing of a spatially distributed hydrological / water quality model for mesoscale watersheds. *Ecological Modeling*, **106**, 261–289.

Leavesley G.H., Markstrom S.L., Brewer M.S. and Viger R.J. (1996b) The modular modeling system (MMS) – the physical process modeling component of a database-centered decision support system for water and power management. *Water Air and Soil Pollution*, **90**, 303–311.

Leavesley G.H., Markstrom S.L., Restrepo P.J. and Viger R.J. (2002) A modular approach to addressing model design, scale and parameter estimation issues in distributed hydrological modeling. *Hydrological Processes*, **16**, 173–187.

Leavesley G.H., Restrepo P.J., Markstrom S.L., Dixon M. and Stannard L.G. (1996a) *The Modular Modeling System – MMS: User's Manual*, U.S. Geological Survey Open File Report 96–151, Also available on the World Wide Web: **http://wwwbrr.cr.usgs.gov/mms**.

Marino M.A. and Simonovic S.P. (2001) *Integrated Water Resources Management*, IAHS Pub. No. 272, IAHS.

Markstrom S.L., Boyle D.P., Pohll G.M., Viger R.J., Fluegel W., Leavesley G.H. and McConnell J.R. (2002) A modular approach to coupling surface water and ground water models. *Eos Transactions AGU*, **83**(47), Abstract H12A-0912 F548.

Maston M.C. and Vaccaro J.J. (2002) *Documentation of Precipitation-Runoff Modeling System modules for the Modular Modeling System for the Watershed and River System Management Program*, U.S. Geological Survey Open-file Report 02–362, Also available on the World Wide Web: **http://water.usgs.gov/pubs/of/ofr02362**.

McDonald M.G. and Harbaugh A.W. (1988) *A Modular Three-Dimensional Finite-Difference Ground-Water Flow Model*. U.S. Geological Survey Techniques of Water-Resources Investigations, Book 6, Chapter A1.

National Institute for Coastal and Marine Management (2000) **http://www.netcoast.nl/projects/netcoast/tools/rikz/simona.htm**.

Parkin G., Ewen J. and O'Connell P.E. (2000) SHETRAN: a coupled surface/subsurface modelling system for 3D water flow and sediment and solute transport in river basins. *American Society of Civil Engineers Journal of Hydrologic Engineering*, **5**, 250–258.

Reed M., Cuddy S.M. and Rizzoli A.E. (1999) A framework for modeling multiple resource management issues – an open modeling approach. *Environmental Modeling and Software*, **14**, 503–509.

Ryan T. (1996) *Development and Application of a Physically based Distributed Parameter Rainfall Runoff Model in the Gunnison River Basin*, U.S. Bureau of Reclamation, National Technical Information Service: Springfield, p. 64.

Schumann A.H., Acreman M.C., Davis R., Marino M.A., Rosbjerg D. and Jun X. (2001) *Regional Management of Water Resources*, IAHS Pub. No. 268, IAHS.

Simonovic S.P., Kundzewicz Z., Rosbjerg D. and Takeuchi K. (1995) *Modeling and Management of Sustainable Basin-Scale Water Resource Systems*, IAHS Publication No. 231, IAHS.

Sophocleous M.A. and Perkins S.R. (2000) Methodology and application of combined watershed and ground-water models in Kansas. *Journal of Hydrology*, **236**(3–4), 185–201.

Sydelko P.J., Hlohowskyj I., Majerus K., Christiansen J.H. and Dolph J. (2001) An object-oriented framework for dynamic ecosystem modeling: applications for integrated risk assessment. *The Science of the Total Environment*, **274**, 271–281.

U.S. Army (1998) *HEC-1 Flood Hydrograph Package User's Manual*, Corps of Engineers, Hydrologic Engineering Center, Davis, CPD-1A. Also available on the World Wide Web: **http://www.hec.usace.army.mil/software/legacysoftware/hec1/documentation/hec1user.pdf**.

U.S. Army (2001) *Hydrologic Modeling System HEC-HMS User's Manual*, Corps of Engineers, Hydrologic Engineering Center: Davis, CPD-74A, Also available on the World Wide Web: **http://www.hec.usace.army.mil/software/hec-hms/documentation/hms_user.pdf**.

U.S. Department of Agriculture (2003) *The Object Modeling System (OMS)*, Agricultural Research Service, Fort Collins, **http://oms.ars.usda.gov/index.html**.

U.S. Environmental Protection Agency (1995) *QUAL2E Windows Interface Users Guide*, EPA/823/B/95/003. Also available on the World Wide Web: **http://www.epa.gov/waterscience/basins/bsnsdocs.html**.

U.S. Environmental Protection Agency (2005) **http://www.epa.gov/owow/watershed**.

U.S. Environmental Protection Agency (2001a) *Better Assessment Science Integrating Point and Nonpoint Sources: BASINS Version 3.0 User's Manual*, EPA/823/B/01/001, Also available on the World Wide Web: **http://www.epa.gov/waterscience/basins/bsnsdocs.html**.

U.S. Environmental Protection Agency (2001b) *PLOAD version 3.0 An ArcView GIS Tool to Calculate Nonpoint Sources of Pollution in Watershed and Stormwater Projects: User's Manual*, **http://www.epa.gov/waterscience/basins/b3docs/PLOAD_v3.pdf**.

U.S. Geological Survey (1998) *Watershed and River System Management Program: Application to the Yakima River Basin, Washington*, USGS Fact Sheet 037–98. Also available on the World Wide Web: **http://water.usgs.gov/pubs/FS/FS-037-98**.

U.S. Natural Resource Conservation Service (1986) *Urban Hydrology for Small Watersheds TR-55*, **http://www.wcc.nrcs.usda.gov/hydro/hydro-tools-models-tr55.html**.

U.S. Natural Resource Conservation Service (1992) *TR-20 Computer Program for Project Formulation Hydrology*, **http://www.wcc.nrcs.usda.gov/hydro/hydro-tools-models-tr20.html**.

Watson F., Rahman J. and Seaton S. (2001) Deploying environmental software using the Tarsier modelling framework. *Proceedings 3rd Australian Stream Management Conference*, Rutherfurd I., Sheldon F., Brierly G. and Kenyon C. (Eds.), Brisbane, 27–29 August, 2001, 631–637. Also available on the World Wide Web: **http://science.csumb.edu/~tarsier/**.

Whelan G., Castleton K.J., Buck J.W., Hoopes B.L., Pelton M.A. Strenge D.L., Gelston G.M. and Kickert R.N. (1997) *Concepts of a Framework for Risk Analysis in Multimedia Environmental Systems*, Pacific Northwest National Laboratory: PNNL-11748, Also available on the World Wide Web: **http://mepas.pnl.gov:2080/FRAMESV1/frames_doc.pdf**.

Whitehead P.G., Wilson E.J. and Butterfield D. (1998) A semi-distributed integrated Nitrogen model for multiple source assessment in catchments (INCA): Part I – model structure and process equations. *Science of the Total Environment*, **210/211**, 547–558.

Zagona EdithA., Fulp TerranceJ, Shane Richard, Magee Timothy and Morgan Goranflo H. (2001) RiverWare: a generalized tool for complex reservoir systems modeling. *Journal of the American Water Resources Association*, **37**(4), 913–929.

130: Fuzzy Sets in Rainfall/Runoff Modeling

ANDRÁS BÁRDOSSY

Universität Stuttgart, Institut für Wasserbau, Stuttgart (Vaihingen), Germany

Fuzzy sets offer an alternative to statistical methods to quantify uncertainties in modeling natural systems. Three types of applications for hydrological modelling can be considered:

- *the quantification of the uncertainties related to inexact knowledge of the model parameters that describe the underlying hydrological processes using fuzzy numbers and the extension principle*
- *alternative description of individual hydrological processes using fuzzy rules*
- *introduction of model performance measures which reflect soft information coded with fuzzy membership functions.*

One of the major problems in the application of fuzzy sets is the assessment of the membership functions. Model parameters and performance measures are usually estimated subjectively. Fuzzy rules can be assessed using automated learning algorithms. Examples show that fuzzy sets, if they are applied in the combination with hydrological expertise, can be well used for rainfall runoff modelling.

INTRODUCTION

The rainfall runoff process is a complicated nonlinear process. Even though one is dealing with a physical system, its modeling contains considerable uncertainties. Uncertainty is usually treated with the help of statistical methods. In the last few years, the theory and application of fuzzy sets has advanced considerably. Fuzzy sets offer an alternative approach to the treatment of uncertainties that is often complementary to statistical methods.

The application of fuzzy sets is not widely spread in hydrology. Besides "hard" hydrological data, such as rainfall or discharge measurements, "soft" information such as expert knowledge on soils and/or governing processes is useful for modeling. Very often only the hard information is considered – leading to an unrealistic approach. The coding of expert knowledge in the form of fuzzy sets offers an interesting alternative. On the other hand, fuzzy sets should not be misused to introduce arbitrariness into hydrological models.

In this article, the application of fuzzy methods for three types of uncertainties in hydrological modeling will be discussed.

1. A part of these uncertainties is related to an inexact knowledge of the model parameters that describe the underlying physical process. Soil hydraulic properties, for example, are highly variable in space (and to some extent even variable in time). Measurements and laboratory experiments are carried out on small scales, while the processes they are part of take place on large scales. This leads to difficulties in the estimation of model parameters. This uncertainty can be considered as partly due to spatial variability. However, often a small amount of direct measurement can be combined with "soft" information, such as soil type, in an informative way. In this case, expert knowledge can be transferred to model parameters. One possible method that can be applied in this context is the idea of fuzzy sets.
2. Besides the inherent uncertainty in the model parameters, there is a considerable uncertainty in the mathematical process description of hydrological models. The physically based concepts are usually not applicable on larger scales. Expert knowledge is here again an important source for enriching the modeling process.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

It can be used not only in finding appropriate parameters or forms of equations, but also in the form of case-dependent rules. Fuzzy rules provide appropriate tools to code expert knowledge and thus are also well applicable for rainfall/runoff modeling.

3. Rainfall/runoff models are (as are most environmental models) simplified descriptions of complicated natural processes. Therefore, they cannot perfectly describe all aspects of the processes. One has thus to select appropriate measures to identify model parameters (calibration) and to judge the performance of models (validation). Model performance measures are usually based on statistical principles. One of the most frequent measures applied is the Nash–Sutcliffe efficiency – which is a purely statistical measure based on a least squares principle. In the case of model calibration and validation, expert knowledge might also play a very important role. Qualitative measures of model performance exploiting this expert knowledge can be introduced in the form of fuzzy measures.

The purpose of this contribution is to provide examples and guidelines for the application of fuzzy sets for the above mentioned three types of problems. The article is divided into six parts. After the introduction of fuzzy sets, fuzzy arithmetic and fuzzy rules are presented. In Section "Fuzzy Hydrological Data and Model Parameters", problems of applying fuzzy parameters in rainfall/runoff models are discussed. Section "Fuzzy Rules for Hydrological Modeling", presents the application of fuzzy rules for rainfall/runoff related processes. In Section "Fuzzy Calibration Measures", qualitative and quantitative performance measures are discussed in a fuzzy context. Finally, in Section "Conclusions", further research needs are discussed.

FUZZY SETS, FUZZY NUMBERS, AND FUZZY RULES

Fuzzy sets were first introduced by Zadeh (1965). The impact of this paper was enormous. Among others, environmental scientists found the idea very appealing and tried to apply it to different problems. A short introduction to the basic principles is given in the next sections.

Membership Functions

A fuzzy set is a set of objects without clear boundaries or without perfectly (precisely) defined characteristics. In contrast with ordinary sets, where for each object it can be decided whether it belongs to the set or not, a partial membership in a fuzzy set is possible. An example of a fuzzy set could be "the set of days with heavy rain". There are days with rainfall amounts that clearly belong to the above set, and others (for example, dry days) that clearly cannot be considered as days with heavy rain. But, if the concept of heavy rain is not exactly defined (for example, daily total >30 mm), then there is a certain "gray" zone where the judgment outcome is not obvious. An exact definition of heavy rain is possible but does not necessarily conform with the human way of thinking. Why should a day with 30.1 mm rain be a day with heavy rain and another with 29.9 mm not be? The normal human observer cannot perceive the difference – only precise measurements could decide on the membership and then a nearby gauge would read differently anyway. However, in real life one can use the concept even without taking any measurements. Weather forecasts use the expression *heavy rain* without quantitative definition.

Formally, a fuzzy set (or often called a fuzzy subset) is defined as follows:

Definition: Let X be a set (universe). A is called a fuzzy subset of X if A is a set of ordered pairs:

$$A = \{(x, \mu_A(x)); x \in X, \mu_A(x) \in [0, 1]\}$$

where $0 \leq \mu_A(x) \leq 1$ is the grade of membership of x in A. The function $\mu_A(x)$ is called the membership function of A.

The closer $\mu_A(x)$ is to 1 the more x is considered to belong to A – the closer it is to 0 the less it is taken as belonging to A. If [0,1] is replaced by the two-element set $\{0, 1\}$, then A can be regarded as an ordinary subset of X because then x is definitely either in A or not in A.

Fuzzy Numbers

Special cases of fuzzy sets are fuzzy numbers, which are generalizations of our usual concept of numbers. While Boolean operations such as union, intersection, or complement can be performed on general fuzzy sets, arithmetic operations such as addition or multiplication can only be performed on fuzzy sets defined on the set of real numbers. The simplest generalizations of real numbers are fuzzy numbers:

Definition: A fuzzy subset A of the set of real numbers is called a fuzzy number if there is at least one z such that $\mu_A(z) = 1$ (normality assumption) and for every real number a, b, c with $a < c < b$

$$\mu_A(c) \geq \min(\mu_A(a), \mu_A(b)) \tag{1}$$

Any real number can be regarded as a fuzzy number with a single point support, and is called a *'crisp number'* in fuzzy mathematics. The simplest fuzzy numbers are

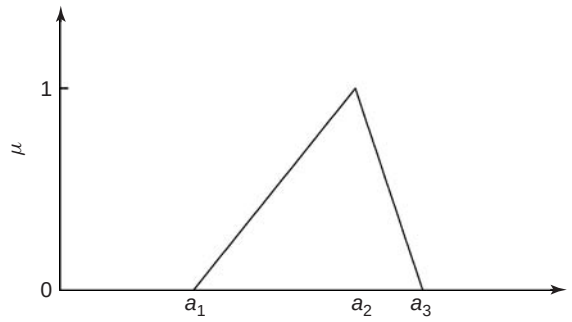


Figure 1 Membership function of the triangular fuzzy number $A = (a_1, a_2, a_3)_T$

those known as *triangular fuzzy numbers*. The membership function of the triangular fuzzy number consists of an increasing and a decreasing linear function – forming a triangle. The formal definition is:

Definition: The fuzzy number $A = (a_1, a_2, a_3)_T$ with $a_1 \leq a_2 \leq a_3$ is a triangular fuzzy number if its membership function can be written in the form:

$$\mu_A(x) = \begin{cases} 0 & \text{if } x \leq a_1 \\ \dfrac{x - a_1}{a_2 - a_1} & \text{if } a_1 < x \leq a_2 \\ \dfrac{a_3 - x}{a_3 - a_2} & \text{if } a_2 < x \leq a_3 \\ 0 & \text{if } a_3 < x \end{cases}$$

Figure 1 shows the membership function of a triangular fuzzy number.

The *extension principle* (Zadeh, 1965) is a method of extending point to point operations to fuzzy sets. It is the basic tool for the development of fuzzy calculations and fuzzy arithmetic.

Definition:

If X and Y are two sets, and f is a point to point mapping from X to Y

$$f: X \longrightarrow Y \quad \text{for every} \quad x \in X \quad f(x) = y \in Y$$

then f can be extended to operate on fuzzy subsets of X in the following way:

Let A be a fuzzy subset of X with membership function μ_A, then the image of A in Y is the fuzzy subset B with the membership function

$$\mu_B(y) = \begin{cases} \sup\{\mu_A(x); y = f(x), x \in X\} \\ 0 \quad \text{if there is no } x \in X \quad \text{such that } f(x) = y \end{cases} \tag{2}$$

Figure 2 illustrates the extension principle. The highest membership of the three elements x_1, x_2 and x_3 all with $f(x_i) = y$ is assigned as membership to y.

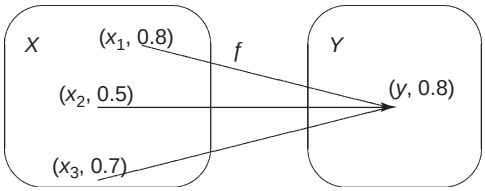


Figure 2 The extension principle

Using the extension principle for the bivariate functions

$$f_a(x_1, x_2) = x_1 + x_2$$
$$f_m(x_1, x_2) = x_1 \cdot x_2$$
$$f_d(x_1, x_2) = \frac{x_1}{x_2}$$

one can derive the rules of fuzzy arithmetic (Kaufmann and Gupta, 1985). For example, for fuzzy addition one can derive from the extension principle that:

$$(a_1, a_2, a_3)_T + (b_1, b_2, b_3)_T = (a_1 + b_1, a_2 + b_2, a_3 + b_3)_T \tag{3}$$

h-level sets of fuzzy sets are useful when one applies the extension principle:

Definition: The h-level set of a fuzzy set A is

$$A(h) = \{x; \mu_A(x) \geq h\}$$

The h level sets are defined for $0 < h \leq 1$.

One can prove that the extension principle leads to the equation:

$$f(A(h)) = B(h) \tag{4}$$

This means that the h level set of the image of a set can be calculated directly from the h level set of the original set. This fact makes the application of the extension principle in general much simpler.

The extension principle is a kind of *pessimistic* way of treating uncertainties. In modeling contexts, it can be regarded as a systematic sensitivity analysis.

The extension principle can be applied to complicated functions either directly or in a step-by-step manner. Consider for example, the case of $g(x) = f_1(x) + f_2(x)$. There are two possible ways of calculating the membership function of $g(A)$ for a given fuzzy number A:

1. The first one calculates the fuzzy sets $B_1 = f_1(A)$ and $B_2 = f_2(A)$ using the extension principle. Then the extension principle is applied to the bivariate function $y = x_1 + x_2$ with the fuzzy arguments B_1 and B_2 to obtain the fuzzy set $g(A) = B$.

2. The extension principle is directly applied to the univariate function g.

The first approach is computationally much simpler, but leads to an increased uncertainty due to the fact that a univariate function is treated as a bivariate one when the two fuzzy numbers B_1 and B_2 are added. This example shows the problem of applying fuzzy arithmetic mainly based on bivariate operations – without considering whether some of the arguments are related (or even identical) or not. The application of fuzzy arithmetic in hydrological modeling is therefore not recommended because in most of the models the same parameters are used for different time steps. Fuzzy arithmetic would lead to a very fast growth in uncertainty as demonstrated in Dou *et al.* (1995), leading to useless results after a few time steps.

Fuzzy Rules and their Assessment

In contrast with, or in complement to, a probabilistic approach, we may consider the treatment of imprecision (or vagueness) in hydrology, corresponding, for example, to the statements: "runoff increases with higher antecedent moisture".

The most popular fuzzy direct reasoning technique is Mamdani's method (Mamdani, 1974). It is composed of direct IF-THEN statements with fuzzy premises and fuzzy conclusions. It is widely applied in fuzzy control (Sugeno and Yasukawa, 1993), however, its use in fuzzy modeling is still restricted to a few specific case studies (Bardossy and Disse, 1993). Mamdani's method is very well suited to describe the complicated, case-dependent nature of hydrological/environmental systems.

A Mamdani type fuzzy rule consists of a set of arguments $A_{i,k}$ in the form of fuzzy sets with membership functions $\mu_{A_{i,k}}$ and a consequence B_i also in the form of a fuzzy set.

$$\text{If } a_1 \text{ is } A_{i,1} \odot a_2 \text{ is } A_{i,2} \odot \ldots \odot a_K \text{ is } A_{i,K} \text{ then } B_i \tag{5}$$

Note that in general, the order in which a statement is evaluated plays a central role. The operator $\odot$ stands for a logical combination of the arguments such as AND or OR or XOR. The statement "a_k is $A_{i,k}$ " is, for simplicity, replaced by $A_{i,k}$ in the following text. The rule (equation 5) will thus be written as

$$\text{If } A_{i,1} \odot A_{i,2} \odot \ldots \odot A_{i,K} \text{ then } B_i \tag{6}$$

The truth value corresponding to the fulfillment of the conditions of a rule for given premises $(a_1, \ldots, a_K)$ is called the *degree of fulfillment* (DOF) of that rule.

Since the DOF depends on the premise vector $(a_1, \ldots, a_K)$ the DOF of rule i is denoted $D_i(a_1, \ldots, a_K)$.

Knowledge of the DOFs does not provide the information about the results – it only indicates what kind of responses could occur. In order to obtain a "unique" result the fuzzy responses have to be combined, and, for the sake of a single response, defuzzified.

Verbal rules are often translated into fuzzy rules using linguistic variables.

It has been shown by Wang and Mendel (1992) that all continuous real functions can be approximated by fuzzy rules. This means that in theory each hydrological model can be well approximated by fuzzy rules.

Another class of fuzzy rules are the Takagi–Sugeno rule systems. In this case, the response of a fuzzy rule is not a fuzzy set but a function. This means that the rule has the form:

$$\text{If } a_1 \text{ is } A_{i,1} \odot a_2 \text{ is } A_{i,2} \odot \ldots \odot a_k \text{ is } A_{i,K} \text{ then } f_i(a_1, \ldots, a_K) \tag{7}$$

This type of formulation is advantageous as it allows a partly functional relationship between input and output variables. Further, the use of more than one rule means that the functional relationships are not necessarily universal, their validity is limited to the cases where the rule is applicable.

Fuzzy rules can be assessed directly from expert knowledge or by derivation from observed data. A combination with artificial neural nets in some cases enables an automatic derivation of fuzzy rules.

Fuzzy rule-based modeling shows much potential in cases when: a causal relationship is well established, but difficult to calculate under real-life conditions; data are scarce and imprecise; and/or when a given input vector may have several contradictory responses, which may be true to varying degrees. These features are often present in hydrology and water resources problems.

FUZZY HYDROLOGICAL DATA AND MODEL PARAMETERS

The uncertainty of a hydrological model parameter or an input variable can be described using fuzzy numbers. For example, if there are no measurements of soil hydraulic conductivity available at a given site, one can use a soil map. The knowledge of the soil type does not provide an exact value of the hydraulic conductivity. Instead, we can have a "best" estimate, and we can also provide a range in between which the soil hydraulic conductivity might be. This information can be translated to a triangular fuzzy number by assigning a membership value of one to the most likely value, zero to the lowest and the highest possible values, and assuming a linear membership function between these values as in Figure 1. One could also assign a fuzzy

number to the rainfall amounts. In this case, the systematic undercatch of the rain gauges would lead to a nonsymmetric triangular fuzzy number.

The uncertainty of the fuzzy soil hydraulic conductivity and the fuzzy rainfall amounts however have a different character. One can assume that the soil hydraulic properties do not change appreciably in time (at least below the surface and during one precipitation event). Therefore, even though their value is uncertain, we can assume that it remains much the same for all-time steps. In contrast, the rainfall uncertainty is time dependent. The systematic errors corresponding to different time steps might be very different, and should be treated accordingly. This leads to a different way of treating these uncertainties in models. If this is not correctly taken into account in modeling, then the uncertainty grows rapidly in time as presented by Dou *et al.* (1995) and Kunstmann and Kinzelbach (1997).

The difference between the application of fuzzy arithmetic (Kaufmann and Gupta, 1985) and the extension principle of Zadeh (1965) is demonstrated in the simple example presented in the next section.

Example – The Fuzzy Unit Hydrograph

We assume that the unit hydrograph is not known exactly, instead only imprecisely. In the case when one applies statistical methods, the statistics of the deviations are usually used to characterize the inprecision. In the fuzzy case, an *envelope* type approach is selected. One would like to obtain a set of functions that approximate the hydrograph both from below and above. For each time t, $u(t)$ is assumed to be a fuzzy number. This means that in the discrete form (2) the fuzzy discharge is calculated as:

$$\hat{Q}_u(t_n) = \sum_{i=0}^{n} p(t_i) \cdot \hat{u}(t_n - t_i) \tag{8}$$

$\hat{u}(t)$ being a fuzzy number, and $p(t_i)$ is the sequence of (crisp) depths of effective rainfall. The fuzzy direct discharge $\hat{Q}_u(t_n)$ can be calculated using two different methods:

1. Using fuzzy arithmetic (Kaufmann and Gupta, 1985) or
2. Using the extension principle (Zadeh, 1965).

The two methods are illustrated with the following example:

Assume that $u(t_0) = (1, 2, 3)_T$ and $u(t_1) = (0, 1, 2)_T$ and that there are effective rainfalls of $p(t_0) = 1$ and $p(t_1) = 2$

Then by the application of fuzzy arithmetic we obtain:

$$\begin{aligned} Q(t_0) &= 1 \cdot (1, 2, 3)_T = (1, 2, 3)_T \text{ m}^3\,\text{s}^{-1} \\ Q(t_1) &= 2 \cdot (1, 2, 3)_T + 1 \cdot (0, 1, 2)_T \\ &= (2, 4, 6)_T + (0, 1, 2)_T = (2, 5, 8)_T \\ Q(t_2) &= 2 \cdot (0, 1, 2)_T = (0, 2, 4)_T \end{aligned}$$

The highest possible discharge is $8\,\text{m}^3\text{s}^{-1}$ at time t_1. However, if we investigate this discharge we can see that it is the result of taking $u(t_0) = 3$ and $u(t_1) = 2$. This choice, however, would violate the mass conservation constraint being

$$p(t_0) + p(t_1) = u(t_0) + u(t_1) = 3.$$

Consequently, the range of the calculated discharge increases leading to a false estimation of uncertainty.

This problem can be avoided if one considers the extension principle. In this case, a membership value is assigned to each fuzzy number $u(t)$ under the constraint that the mass conservation is not violated. This leads to the membership function of the discharge at time t:

$$\mu_{Q_u(t)}(Q) = \max \left\{ \begin{array}{l} \min(\mu(u(t_0))..., \mu(u(t_n)) \\ \text{such that } \; Q = \sum p(t_i) u(t - t_i) \\ \text{and} \sum_{i=0}^{n} u(t_i) \; = \gamma A_E \end{array} \right\}$$

As one can see, the difference between the use of fuzzy arithmetic and the extension principle lies in the additional constraint on the integral of the unit hydrograph. This means that by introducing $u(t_0) = x$, mass conservation leads to $u(t_1) = 3 - x$

$$Q(t_1) = x \cdot 2 + (3 - x) \cdot 1 = x + 3$$

if $1 < x \leq 2$ then for this x we have $\mu_{u_1}(x) = x - 1$ and $1 \leq 3 - x < 2$ leading to $\mu_{u_2}(3 - x) = 2 - (3 - x) = x - 1$ if $2 < x < 3$ then $\mu_{u_1}(x) = 3 - x$ and $0 < 3 - x < 1$ $\mu_{u_2}(3 - x) = 3 - x\; x + 3$ has the membership value (in the result) of $x - 1$ if $1 < x < 2\; x + 3$ membership of $3 - x$ if $2 < x < 3$ leading to the result:

$$Q(t_1) = (4, 5, 6)_T$$

This result is much crisper than the previous one $(2, 5, 8)_T$ and reflects the physical constraint of mass conservation. The application of fuzzy arithmetic in models generally leads to an unrealistic increase of uncertainty. Therefore, it is suggested that the extension principle should be used in all fuzzy calculations.

The application of the extension principle requires additional effort. The calculations can to some extent be simplified if one uses h-level sets.

Assessment of the Fuzzy Model Parameters

The fuzzification of hydrological models using fuzzy parameters is in most cases not a very complicated exercise. The identification of the parameters from observations is in this case a much more complicated task.

Rainfall/runoff models cannot perfectly reproduce measured discharges. The deviations have different causes, for example, the simplification of the process descriptions, the uncertainties associated with the assessment of the model parameters (for example, spatial variability or temporal variations of parameters assumed to be time independent).

The fuzzification of the parameters means that a part of the uncertainty should be treated as fuzzy, another part as probabilistic. Ozelkan and Duckstein (2001) presented different fuzzy conceptual rainfall/runoff models. They suggested the use of a fuzzy linear regression type approach (Tanaka *et al.*, 1982; Bardossy *et al.*, 1990) for the estimation of the model parameters. Using the fuzzy parameters for each time step, a fuzzy discharge is calculated. The observed discharge should have a reasonably high membership level in the calculated fuzzy discharge value set. This condition could be fulfilled by infinitely many parameter sets – one tries to find the crispest among them. For this purpose, an acceptable level and a measure of crispness have to be defined. Ozelkan and Duckstein (2001) suggest a multiobjective methodology. The parameter estimation procedure in this context, as in the statistical type of approach, leads to an optimization problem. Owing to the fact that the uncertainty is in this case explicitly considered in the fuzzy parameters, the number of unknowns to be identified increases. For example, instead of a simple linear reservoir storage parameter K, in this case a triangular fuzzy number $(K_1, K_2, K_3)_T$ would have to be identified.

A combination of statistical and fuzzy uncertainty for the identification of the parameters offers an interesting possibility. Some of these aspects are discussed in the next section.

Fuzzy Versus Stochastic Uncertainty

As stated in the introduction of this section, different types of uncertainties should be treated differently. Unknown systematic errors and natural variability have a very different influence on modeling results. The major difference between the two approaches lies in the treatment of the interdependence between variables. While if considered as fuzzy the dependence of the variables is considered in a pessimistic way leading to large uncertainties, the probabilistic approach allows a whole range of different dependence structures.

For example, for a given time interval, the precipitation amount can be treated both as a random variable or as a fuzzy quantity. The systematic undercatch of the raingauges can be well modeled as a fuzzy number. On the other hand, the uncertainty due to spatial and/or temporal variability of can be better described using random variables. In many cases, a combination of fuzzy and stochastic methods could yield an interesting quantification of uncertainties. For example, spatial variability of soil parameters can be well described using geostatistical methods. The expert knowledge in the form of fuzzy parameters can be combined using fuzzy Kriging (Piotrowski *et al.*, 1995) or fuzzy conditional geostatistical simulation. A fuzzy number can be used to describe the unknown mean of a field while stochastic methods are used for the spatial variability. A joint treatment of the methods requires the combination of Monte Carlo methods with the appropriate application of the extension principle. This effort is however rewarded by the insight gained though the separation of the different kinds of uncertainty in the model results. Expected values, exceedence probabilities, or other statistical quantities become fuzzy numbers, their memberships depending on the fuzzy uncertainty, while the memberships of certain physical properties (for example, travel times, peak values, etc.) are characterized using their probabilistic type distribution. The author of this article does not know any application of the above sketched methodology in hydrology. It is subject of future research to explore the capacity of these approaches.

FUZZY RULES FOR HYDROLOGICAL MODELING

Fuzzy rules can describe relationships between variables in a linguistic form. Owing to their case-specific validity, rules can be usefully employed to describe complicated nonlinear relationships. Thus, they have a good potential in rainfall/runoff modeling. As rules are universal approximators (Wang and Mendel, 1992), all hydrological models could be formulated in terms of fuzzy rules. Hundecha *et al.* (2001) used this principle and developed a fuzzy rule-based version of the HBV model (Bergstrom and Forsman, 1973). The fuzzification of existing rainfall/runoff models is an interesting intellectual exercise, but does not necessarily advance hydrological modeling. The replacement of complicated models by fuzzy rules might however be advantageous for processes, where the traditional formulation of the model requires extensive calculations.

For example, the modeling of water dynamics in the unsaturated zone, a common approach is to solve the Richards equations using suitable algorithms. The main idea of using fuzzy rules is that the water movement at a certain time and at a selected point depends to a very high degree only on the conditions (moisture content) in the immediate neighborhood of the point. The fuzzy rules to describe the vertical water movement can be formulated with the help of relative soil moisture contents. The parameter set can be reduced by extracting two *linear coefficients*, the saturated hydraulic conductivity K_s and the moisture content θ_i, with the idea of including all the nonlinearities into the rule system. A rule be formulated verbally as:

If the relative soil moisture in element i is $A_{1,i}$ and the relative moisture content of the adjacent element below is $A_{2,i}$ then the normalized flux between the elements is B_i.

Normalized flux means in this case that a specific value K_s is selected for which the rules are specified. As the flux is proportional to the K_s value, a simple multiplication makes the rule applicable to the case of a value of K_s different from the selected reference value. In Bardossy and Disse (1993), the above unsaturated flow model is described in detail. Schulz and Huwe (1999) presented a fuzzy model for the transport in the unsaturated zone.

The application of fuzzy rules in hydrological modeling requires an assessment of the rules. The rules can be assessed from experiments or even from numerical models. The first approach would fully disregard the physical knowledge. Furthermore, measurement errors and inaccuracies would influence the rule system. Using numerical models has the advantage, that rules for arbitrary conditions can be assessed. A combination of experimental and model results is also possible.

As the processes are complicated and depend on a large number of parameters the derivation of rules from datasets is not a trivial task. The membership functions of the arguments can be defined as triangular or trapezoidal fuzzy numbers over the range of the possible parameter values. For each parameter, a certain number of membership functions should be allowed. If one considers all possible combinations of the parameter membership functions, this might lead to an explosion of the rule system. The available information might not be enough for a reasonable assessment of the rules; in the case of rule systems one should try to be parsimonious. In hydrology the number of influencing variables is usually high, but the dominant processes depend on only a few of them. If rule systems are built by considering only the important variables for modeling, the number of rules might be reduced to a maximum of 10 to 40. Bardossy *et al.* (2003), presented an automated methodology for the assessment of a small number of rules based on simulated annealing. Neuro-fuzzy approaches offer another possibility for the automatic assessment of fuzzy rules and the corresponding membership functions. An application of this methodology for flood forecasting is presented by Stuber and Gemmar (1997).

Fuzzy rules can also be used to combine the results of different models. Each hydrological model has its own strengths and weaknesses. If one can recognize *a priori* whether in a given situation, the model can be trusted or not, then a case-dependent combination of the model results can lead to substantial improvements. Xiong *et al.* (2001) developed a Takagi-Sugeno rule system to combine the results of flood forecasting models. Their results showed that the combination lead to substantial improvements. The nonlinearity of the combination played a central role in their approach.

FUZZY CALIBRATION MEASURES

Expert knowledge can be very valuable for the identification of hydrological model parameters and for the judgement as to what extent a model can be regarded as a reasonable representation of the natural system under study.

Hydrological models calculate the different components of the water cycle in a spatial resolution, which is usually finer than the resolution of the observations. Further, there are often processes that are only modeled and not quantitatively measured. Despite there being no hard data available, these variables might be used for model calibration. Expert knowledge can give plausible ranges for these variables and maximum and minimum possible values for individual states or for time integrals. These expert opinions can be described with the help of fuzzy sets, for example, as trapezoidal fuzzy numbers. For each calibration variable, a soft measure in the form of a fuzzy set A_k with the membership function $\mu_{A_k}(S_k)$ is defined. The membership value is defined on the state variable S_k. The value of S_k depends on the parameter vector $\mathbf{p}$ of the hydrological model. The calibration procedure can now be transformed to an optimization problem with constraints or to a bivariate optimization problem. The first is:

$$\mu_{A_k}(S_k(\mathbf{p})) \geq h_0 \quad k = 1, \ldots, K$$
$$O(\mathbf{p}) \rightarrow \max \tag{9}$$

where $O(\mathbf{p})$ represents the "crisp" performance measure of the hydrological model. Alternatively, a bivariate optimization problem can be formulated:

$$\min(\mu_{A_k}(S_k(\mathbf{p}))\ ;\ k = 1, \ldots, K) \rightarrow \max$$
$$O(\mathbf{p}) \rightarrow \max \tag{10}$$

Fuzzy sets are used for multiobjective calibration as a kind of value function. The performance measures are transformed to the unit interval and are interpreted as fuzzy membership values. This approach was used by Cheng *et al.* (2002) for the calibration of the Xinanjiang model. Yu and Yang (2000) presented a multiobjective fuzzy methodology for model calibration. Fuzzy sets are not essential for such purposes, but they provide a kind of unified framework for treating the different measures of model performance in a reasonable balance.

Samanta and Scott Mackay (2003) present an interesting methodology for the calibration of hydrological models, which considers the set of acceptable model parameters as a fuzzy set.

CONCLUSIONS

Fuzzy sets provide a methodology to handle uncertainty in mathematical models describing natural systems. Until now, their application in hydrology has been limited to a relatively small number of cases. The possibility of including information and knowledge into the model-building and calibration procedure, which due to its uncertainty was not used before, offers a great potential for hydrological modeling and makes fuzzy sets attractive. The correct and appropriate use of the methods requires further research. The author believes that the number of successful applications of fuzzy sets in hydrology will increase substantially in the future.

REFERENCES

Bardossy A., Bogardi I. and Duckstein L. (1990) Fuzzy regression in hydrology. *Water Resources Research*, **26**(7), 1497–1508.

Bardossy A. and Disse M. (1993) Fuzzy rule-based models for infiltration. *Water Resources Research*, **29**, 373–382.

Bardossy A., Haberlandt U. and Krysanova V. (2003) Automatic fuzzy-rule assessment and its application to the modelling of nitrogen leaching for large regions. *Soft Computing*, **7**, 370–385.

Bergstrom S. and Forsman A. (1973) Development of a conceptual deterministic rainfall-runoff model. *Nordic Hydrology*, **4**, 174–170.

Cheng C.T., Ou C.P. and Chau K.W. (2002) Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall-runoff model calibration. *Journal of Hydrology*, **268**, 72–86.

Dou C., Woldt W., Bogardi I. and Dahab M. (1995) Steady state groundwater flow simulation with imprecise parameters. *Water Resources Research*, **31**(11), 2709–2719.

Hundecha Y., Bardossy A. and Theisen H.W. (2001) Development of a fuzzy logic-based rainfall-runoff model. *Hydrological Sciences Journal*, **46**(3), 363–376.

Kaufmann A. and Gupta M.M. (1985) *Introduction to Fuzzy Arithmetic: Theory and Applications*, Van Nostrand Reinhold: New York.

Kunstmann H. and Kinzelbach W. (1997) Methoden zur quantifizierung von unsicherheit in grundwassermodellen. In *Modellierung in der Hydrologie*, Schmitz G. (Ed.), Dresden, pp. 104–118.

Mamdani E.H. (1974) Applications of fuzzy algorithms for control of a simple dynamic plant. *Proceedings of IEEE*, **121**(12), 1585–1588.

Ozelkan E.C. and Duckstein L. (2001) Fuzzy conceptual rainfall-runoff models. *Journal of Hydrology*, **253**, 41–68.

Piotrowski J.A., Bartels F., Salski A. and Schmidt G. (1995) Geostatistische Regionalisierung Hydrogeologischer Parameter Mit FUZZY-kriging, Konferenzmat. 62. Tagung Arbeitsgemeinschaft Nordwestdeutscher Geologen, 12–19.

Samanta S. and Scott Mackay D. (2003) Flexible automated parametrization of hydrologic models using fuzzy logic. *Water Resources Research*, **39**, SWC 1–13.

Schulz K. and Huwe B. (1999) Uncertainty and sensitivity analysis of water transport modelling in a layered soil profile using fuzzy set theory. *Journal of Hydroinformatics*, **1**, 127–138.

Stuber M. and Gemmar P. (1997) An approach for data analysis and forecasting with neuro fuzzy systems – demonstrated on flood events at river Mosel, *International Conference on Computational Intelligence, 5th Fuzzy Days in Dortmund*, Dortmund, April 28–30.

Sugeno M. and Yasukawa T. (1993) A fuzzy logic based approach to qualitative modelling. *IEEE Transactions on Fuzzy Systems*, **1**(1), 7–31.

Tanaka H., Uejima S. and Asai K. (1982) Linear regression analysis with fuzzy model. *IEEE Transactions on Systems Man and Cybernetics*, **SMC-12**, 903–907.

Wang L. and Mendel J.M. (1992) Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man and Cybernetics*, **22**(6), 1414–1427.

Xiong L., Shamseldin A.Y. and O'Connor K.M. (2001) A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. *Journal of Hydrology*, **245**, 196–217.

Yu P.-S. and Yang T.-C. (2000) Fuzzy multi-objective function for rainfall-runoff model calibration. *Journal of Hydrology*, **238**, 1–14.

Zadeh L. (1965) Fuzzy sets. *Information and Control*, **8**, 338–353.

131: Model Calibration and Uncertainty Estimation

HOSHIN V GUPTA[1], KEITH J BEVEN[2] AND THORSTEN WAGENER[1,3]

[1]Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, US
[2]Department of Environmental Science, and Lancaster Environment Centre, Lancaster University, Lancaster, UK
[3]Now at Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, US

All rainfall-runoff models are, by definition, simplifications of the real-world system under investigation. The model components are aggregated descriptions of real-world hydrologic processes. One consequence of this is that the model parameters often do not represent directly measurable entities, but must be estimated using measurements of the system response through a process known as model calibration. The objective of this calibration process is to obtain a model with the following characteristics: (i) the input-state-output behavior of the model is consistent with the measurements of catchment behavior, (ii) the model predictions are accurate (i.e. they have negligible bias) and precise (i.e. the prediction uncertainty is relatively small), and (iii) the model structure and behavior are consistent with current hydrologic understanding of reality. This article describes the historic development leading to current views on model calibration, and the algorithms and techniques that have been developed for estimating parameters, thereby enabling the model to mimic the behavior of the hydrologic system. Manual techniques as well as automatic algorithms are addressed. The automatic approaches range from purely random techniques, to local and global search algorithms. An overview of multiobjective and recursive algorithms is also presented. Although it would be desirable to reduce the total output prediction error to zero (i.e. the difference between observed and simulated system behavior) this is generally impossible owing to the unavoidable uncertainties inherent in any rainfall-runoff modeling procedure. These uncertainties stem mainly from the inability of calibration procedures to uniquely identify a single optimal parameter set, from measurement errors associated with the system input and output, and from model structural errors arising from the aggregation of real-world processes into a mathematical model. Some commonly used approaches to estimate these uncertainties and their impacts on the model predictions are discussed. The article ends with a brief discussion about the current status of calibration and how well we are able to represent the effects of uncertainty in the modeling process, and some potential directions.

THE NATURE OF RAINFALL-RUNOFF MODELS

The hydrology of any catchment involves complex interactions driven by a number of spatially distributed and highly interrelated water, energy, and vegetation processes. Any computer-based model intended to represent the behavior of a catchment must, therefore, conceptualize this reality using relatively simple mathematical equations that involve parameters to be specified for any particular application. Two characteristics of the modeling process are relevant to our discussion. First, all rainfall-runoff (RR) models, regardless of how spatially explicit, are to some degree lumped, so that their equations and parameters describe the processes as aggregated in space and time. As a consequence, the model parameters are typically not directly measurable, and have to be specified through an indirect process of parameter estimation. This process of parameter estimation is often called model *calibration* if values of parameters are adjusted to fit some observations made

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

on the system of interest. Rainfall-runoff models usually contain several such "conceptual" parameters. While many of these parameters cannot be assumed to have direct physical (measurable) interpretations, they are often assumed to have physical relevance, insofar as they are related to inherent and invariant properties of the hydrologic system. Second, the structure of the RR model is generally specified prior to any attempt to model the catchment being undertaken (Wheater *et al.*, 1993). While this specification is usually based on observed characteristics of the catchment, other factors that play a major role include the availability of data, modeling goal, and a variety of subjective considerations including personal preference and experience. This article will not address the issue of model structure specification, but will focus on the difficulties of model calibration, assuming that a suitably representative and acceptably accurate model structure has already been selected. The article reviews the historical development leading to current views of model calibration, and discusses the estimation and propagation of uncertainties in RR modeling.

MODEL CALIBRATION

Calibration is a process in which parameter adjustments are made so as to match (as closely as possible) the dynamic behavior of the RR model to the observed behavior of the catchment (Figure 1). The process therefore requires measurements of catchment behavior, usually in terms of the inputs (rainfall) and the outputs (e.g. streamflow at the catchment outlet). Because, the outputs of RR models are usually related to the parameters in a nonlinear way, explicit linear-regression-type solutions are generally not possible, and some degree of directed iterative guesswork is required to arrive at a suitable solution (estimates for the parameters). Necessary conditions for an RR model to be "well-calibrated" are that it has (at least) the following three characteristics: (i) the input-state-output behavior of the model is consistent with the measurements of catchment behavior, (ii) the model predictions are accurate (i.e. they have negligible bias) and precise (i.e. the prediction uncertainty is relatively small), and (iii) the model structure and behavior are consistent with a current hydrologic understanding of reality. Note that, for the second requirement to be met, some method for estimating and reporting model prediction uncertainty must be included (Figure 2). Further, the third requirement is critical if the model is to be used to estimate the effects of perturbations to the structure of the real system (e.g. land-use changes).

It is important to stress that the process of model identification should not be understood as that of simply finding a model structure and parameter set that "fits the model to the data". It is actually a process of progressive model identification in which the initial (large) uncertainty in our knowledge of what constitutes a good model structure and good parameter estimates is sequentially reduced while constraining the model to be structurally and functionally (behaviorally) consistent with the available qualitative (descriptive) and quantitative (numerical) information about the catchment. Because, as mentioned before, any selected model will be (at best) a structural and functional approximation of the true (unknown) watershed structure and function, the calibrated estimates of the parameters and the resulting predictions will always contain some remaining uncertainty. These and other uncertainties will also lead to the result that our model will generally not be able to fit the data perfectly, that is, we will not be able to perfectly track the observed system behavior with our model.

Early methods for the calibration of RR models were based on manual, so-called "trial-and-error" procedures. The manual calibration process can be considered to have three levels (Boyle *et al.*, 2000). In Level Zero, the initial uncertainty of the estimates is defined by selecting feasible

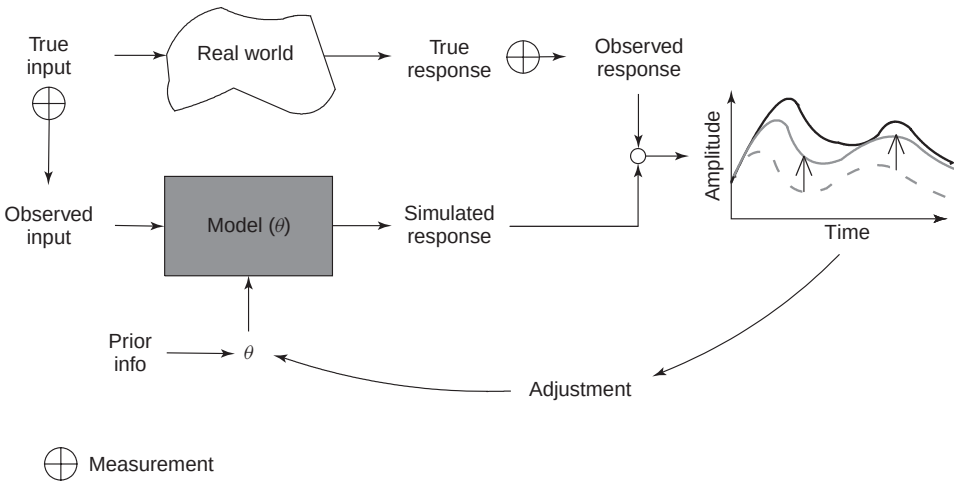


Figure 1 Strategy for model calibration. The model parameter set is represented by θ

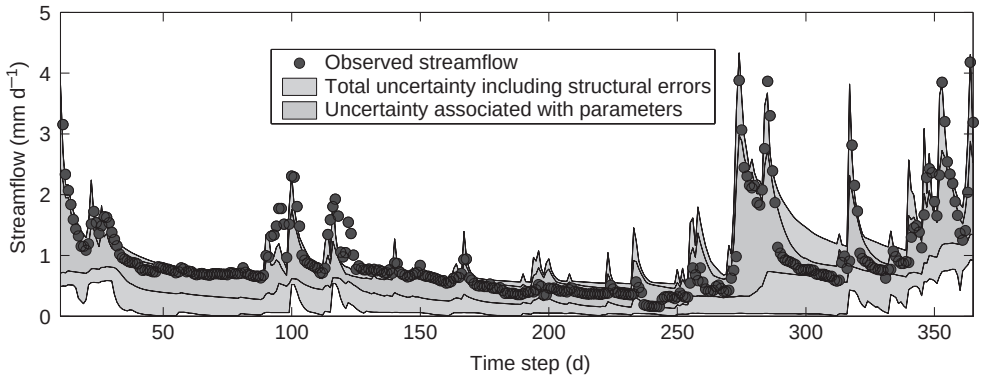


Figure 2 Probabilistic streamflow prediction. Flow is shown in transformed space. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

ranges for each parameter, using estimates from similar catchments, look-up tables, maps, and databases. In Level One, the hydrologist attempts to reduce the size of the parameter uncertainty by adjusting one parameter at a time to try and match the particular segments of the catchment input-output (IO) response to which those parameters are most sensitive. Parameter interaction is usually ignored at this stage. Finally, in Level Two, the behavior of the entire hydrograph is examined, and the parameters further adjusted to allow for parameter interactions, while further reducing the distance between simulated response and the observed catchment behavior. This last stage is the most difficult, owing to the complex nature of the parameter interactions and the nonlinear nature of the model. A systematic approach to Level Two parameter estimation requires (i) a strategy to define (measure) the closeness between the observations and the model response, and (ii) a strategy to reduce the size of the feasible parameter space.

While the manual approach to model calibration (as described above) is based on subjective judgment and expertise, a trained and experienced hydrologist can often obtain excellent results, so that the model response generates a realistic simulation of the response of the catchment. However, the process can be very time consuming, and because it involves subjective decisions by the modeler, requires considerable training and practice. Further, the knowledge and skills so obtained are not easily transferred from one person to another. These limitations have led to interest in methods for model calibration that can be carried out automatically using the speed and power of a digital computer.

The goal of the automatic calibration approach is to use a computer to perform the difficult Level Two stage of the three-level strategy outlined above. Level Zero is still performed manually to provide a crude description of the feasible parameter space, and the Level One stage is generally ignored (see Wagener *et al.*, 2003a, for a discussion of this problem). The potential advantages of a computer-based approach are not difficult to enumerate – properly designed computer algorithms can be fast and objective, while handling more complex problems (e.g. stronger nonlinearities and larger numbers of parameters). The closeness between the simulated and observed responses is typically measured by one (sometimes two or more) mathematical measures (called *objective functions*; OFs) and the parameters are adjusted by an iterative search procedure (called an *optimization algorithm*) towards the optimal value(s) of the OF(s).

An objective function (OF) is a numerical summary of the magnitude of the residuals, that is, the difference between the observed (measured) and the simulated system response (typically the streamflow hydrograph). The goal of calibration is usually to minimize (or maximize depending on definition) the value of this OF. The residuals are calculated as follows,

$$e(\theta) = y_t{}^{\text{obs}} - y_t(\theta) \tag{1}$$

where $y_t{}^{\text{obs}}$ is the observed response at time t, $y_t(\theta)$ is the simulated response, and θ is a vector of parameter values. The residuals are usually aggregated using a prespecified function,

$$F(\theta) = F\{e_t(\theta), \quad t = 1, 2, \ldots, n\} \tag{2}$$

The most commonly applied OFs are of the Weighted Least Squares (WLS) type (e.g. Sorooshian and Gupta, 1995), derived from regression theory,

$$F(\theta) = \sum_{t=1}^{n} w_t \cdot [y_t{}^{\text{obs}} - y_t(\theta)]^2 \tag{3}$$

where w_t is the weight at time t, and n is the number of data points analyzed. In absence of additional information, the weights are commonly set to the value 1.0 for all-time steps. This leads to the Simple Least Squares (SLS)

OF that yields unbiased parameter estimates when the following assumptions regarding the residual distribution are valid: (i) the residuals are independent and identically distributed (i.i.d.), (ii) the distribution shows a homogeneous variance, and (iii)] the residual distribution follows a normal distribution with zero mean (e.g. Gershenfeld, 1999). Because these assumptions are often violated in RR modeling, several researchers have tested the use of alternative error models, and therefore alternative OFs (e.g. Sorooshian and Dracup, 1980; Sorooshian *et al.*, 1983; Kavetski *et al.*, 2003). Examples include the Heteroscedastic Maximum Likelihood Estimator (HMLE, Sorooshian and Dracup, 1980; Sorooshian *et al.*, 1983), which considers a heteroscedastic variance in the system response measurements, and the Bayesian Total Error Analysis (BATEA, Kavetski *et al.*, 2003), which also considers errors in the input data. A more general form of OF, which can be adjusted to be consistent with different error models, was introduced to hydrologic modeling by Thiemann *et al.* (2001), and is based on an exponential power density (Box and Tiao, 1973).

Research into automatic methods (by the hydrologic community) began in the 1960s and 1970s. At that time, it was assumed that the problem of model calibration is similar to that of estimating the parameters in a nonlinear statistical regression. As mentioned above, the OF was typically selected to be some kind of l-norm, such as the weighted sum of squared errors ($l = 2$) shown in equation (3). The minimization of the OF was typically carried out using a "Local Search" algorithm, beginning with an initial parameter guess and employing a preprogrammed iterative strategy to move the parameter search in a direction of local improvement.

In general, Local Search strategies belong to one of two classes – derivative-free (direct) methods and derivative-based (gradient) methods. Examples of popular direct methods include the Downhill Simplex (Nelder and Mead, 1965), the Pattern Search (Hooke and Jeeves, 1961), and the Rotating Directions (Rosenbrock, 1960) algorithms. Gradient methods are potentially more powerful than direct methods, because they use additional information – estimates of the local downhill direction based on the first and/or second derivative of the response surface with respect to the model parameters (Bard, 1974). Although Gupta and Sorooshian (1985) and Hendrickson *et al.* (1988) showed that analytical or numerical derivatives could be computed, even for complex conceptual RR models, Hendrickson *et al.* (1988) found that, in practice, gradient methods do not perform better than direct methods. The use of local search algorithms for model calibration has been tested extensively (Ibbitt, 1970; Johnston and Pilgrim, 1976; Pickup, 1977; among many others), with the general conclusion that such methods are unable to provide a reliable estimate of the globally optimal solution to the RR model minimization problem. Instead, the solution to a Local Search is typically strongly dependent on the accuracy of the initial guess.

Initial responses, during the 1980s, to the failure of automatic calibration methods based on Local Search, were to try and put the optimization problem onto a more rigorous statistical footing. Two (related) directions can be found in the literature, one based on the use of maximum likelihood theory (e.g. Sorooshian and Dracup, 1980) and the other based on the use of Bayesian theory (e.g. Kuczera, 1983a,b). However, neither of these directly addressed the causes of the inability to find the optimum for a selected OF.

Towards the end of the 1980s, with the advent of easier access to powerful digital computers, attention shifted to the testing of "Global Search" algorithms (e.g. Brazil and Krajewski, 1987). A characteristic of many Global Search algorithms is to begin with a number of initial guesses distributed throughout the feasible parameter space, and to evolve this population of guesses iteratively towards promising regions of the OF response surface. Global search algorithms that have been tested include Adaptive Random Sampling (Masri *et al.*, 1980; Brazil, 1988), Simulated Annealing (Kirkpatrick *et al.*, 1983; Thyer *et al.*, 1999), Controlled Random Search (Price, 1987; Klepper *et al.*, 1991) and the Genetic Algorithm (Holland, 1975; Goldberg, 1989; Wang, 1991). Duan *et al.* (1992) conducted a detailed analysis of the properties of the OF response surface associated with a typical RR model and found that:

- It contains more than one main region of attraction.
- It has many local optima within each region of attraction (Figure 3).
- It is rough with discontinuous derivatives.
- It is flat near the optimum with significantly different parameter sensitivities.
- Its shape includes long and curved ridges.

These insights were incorporated into the design of a novel Global Search procedure called the *Shuffled Complex Evolution* (SCE-UA) algorithm, which combines elements of a number of different strategies, including the Downhill Simplex, Controlled Random Search, and Competitive evolution with the newly proposed idea of "Complex Shuffling" (Duan *et al.*, 1992, 1993, 1994; Sorooshian *et al.*, 1993). Extensive testing of the SCE-UA method by numerous researchers has proven its effectiveness, ability to consistently find the global optimum), and efficiency (low probability of failure of any trial) in reliably finding the global solution, when a unique solution exists.

However, these studies have also demonstrated that numerous parameter sets usually exist, widely distributed throughout the feasible parameter space, which have very similar values for the selected OF. This poses difficulties for local or global optimization methods, referred

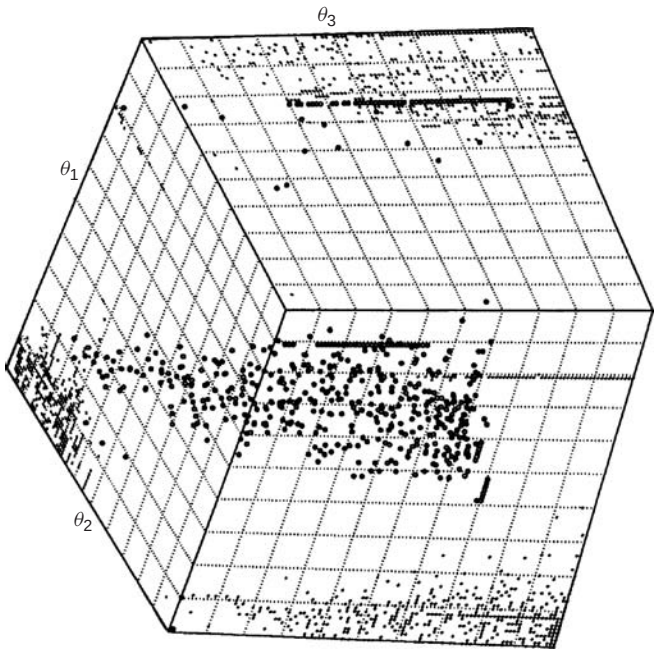


Figure 3 Three-parameter (θ_i) subspace of a simple conceptual catchment model (SIXPAR, Duan *et al.*, 1992), showing locations of multiple local optima

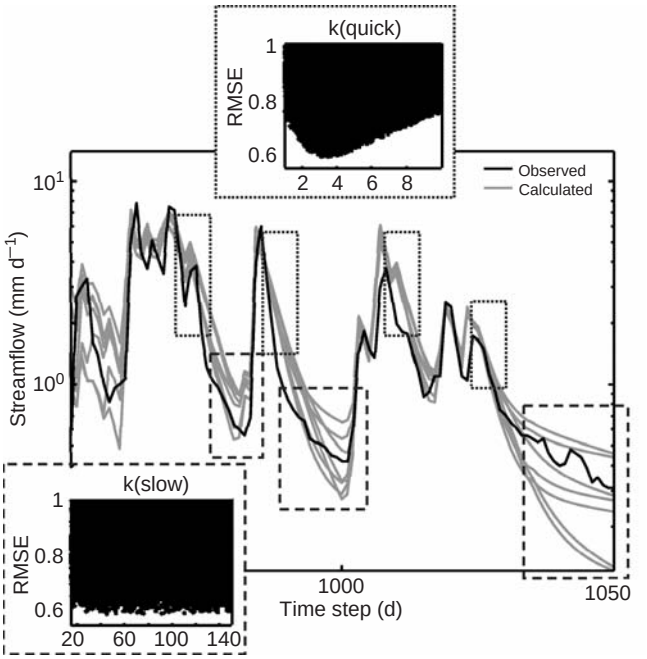


Figure 4 Hundred days extract of six years of daily streamflow data. Observed flow in black, seven different model realizations in grey. Inlets show dotty plots for the linear reservoir residence times k(quick) and k(slow) versus their corresponding Root Mean Squared Error (RMSE) values. The model structure used consists of a Penman soil moisture accounting and a parallel routing component of linear reservoirs with fixed flow distribution

to in the optimization literature as problems of non-uniqueness, indeterminacy, or nonidentifiability. In reviewing this problem, Beven (1993) applied the term "equifinality" to describe the generic nature of finding multiple feasible models that make predictions consistent with any observations available for model calibration in any reasonably complex modeling problem.

The consequences of this progression of events have been some shifts in the underlying philosophy and perceived objectives of model calibration. Model calibration procedures were traditionally based on an attempt to find a "best" (most likely) estimate of the parameter values conditioned on the data, and a subsequent best (most likely) estimate/prediction of the catchment response. Efforts were concentrated on finding the most efficient techniques for doing so with a view to saving computer time in model calibration. The findings reported above have helped to make it clear that the inherent uncertainty (indeterminacy) in the estimated parameter values must be explicitly considered during both calibration and prediction. Approaches to do so are discussed in the section "Considering parameter uncertainty".

Another response has been to look for causes of the parameter indeterminacy and to design ways to address them. In particular, the traditional way to pose the model calibration problem relies on the specification of a single OF, which provides an aggregate measure of the mean distance between observed and calibrated hydrograph over the whole length of the data time-series, as the measure of performance. This action is now understood to result in considerable loss of important information that can be used to distinguish between competing parameter sets. For example, Figure 4 shows a number of model-simulated hydrographs (from different parameter sets) that produce identical OF values, but are clearly visually and behaviorally different (Wagener *et al.*, 2003c). Based on this, Gupta *et al.* (1998) suggest that a calibration approach with higher discriminative power is required and proposed use of a multiobjective methodology. This approach is described further in the section "Considering structural uncertainty".

A special case of parameter estimation arises in predicting the response of ungauged catchments. These are catchments for which observed records of the variable of interest, usually streamflow, are either too short or nonexistent. The main approaches to estimate the parameters of lumped rainfall-runoff models for ungauged catchments are through *physical reasoning* (e.g. Koren *et al.*, 2003), *statistical analysis* (e.g. Jakeman *et al.*, 1992; Wagener *et al.*, 2003b), or a mixture of both. Physical reasoning in this context means that parameters are derived from catchment properties, either directly or through empirical equations. Koren *et al.* (2003) suggest that reasonable initial estimates

for the parameters of the Sacramento soil moisture accounting (SAC-SMA) model can be derived from data such as average field capacity or wilting point of the soil in a catchment. The main approach to ungauged modeling however is the derivation of statistical relationships between model parameters and catchment characteristics for a large number of gauged catchments. Typically, a single model structure is selected that is assumed suitable to represent all the available catchments. The parameters of this structure are derived through a calibration process for all the gauged catchments, and an attempt is then made to develop regression relationships between model parameters and catchment characteristics. It is hoped that these statistical relationships can then be used to predict the parameters in ungauged catchments. See Wagener *et al.* (2003b) and **Chapter 133, Rainfall-runoff Modeling of Ungauged Catchments, Volume 3** for a discussion on this type of approach.

It should be noted that, regardless of the chosen approach, it will generally be impossible to reduce the total output prediction error, that is, the difference between observed and simulated system behavior to zero. Even if we could be sure that we had the correct model equations, errors in the input data and observations used in calibration will result in a residual prediction error. Indeed, experience suggests that reducing the error during one system response mode, often leads to an increase in the error during another mode (Gupta *et al.* 1998). In the following sections we will discuss the components and characteristics of the error.

It should also be noted that streamflow is not the only catchment response that can be used in model calibration. There is a wide range of studies where alternative hydrological variables are used for this purpose. Examples are Lamb *et al.* (1998) and Blazkova *et al.* (2002b) using distributed groundwater levels; Kuczera and Mroczkowski (1998) using groundwater level observations and stream salinity data; Franks *et al.* (1998) and Blazkova *et al.* (2002a) using information on saturated areas; and Seibert and McDonnell (2003) using "soft" data on the nature of catchment responses.

ON THE SOURCES AND THE NATURE OF THE TOTAL ERROR IN RAINFALL-RUNOFF MODELING

There is a problem, in any modeling application, of trying to understand the origins of the error between model predictions of a variable and any observational data of the same variable. The difficulty arises because there are a variety of sources for the error but (at any given time) only one measure of the deviation or residual between prediction and observation (i.e. the "total error"). Thus, disaggregation of the error into its source components is difficult, particularly in cases common to hydrology where the model is nonlinear and different sources of error may interact to produce the measured deviation.

Obvious sources of error in the modeling process include the errors associated with the model inputs and boundary conditions, the errors associated with the model approximation of the real processes, and the errors associated with the observed (measured) variables. A less obvious source of error is when the variable predicted by a model is not the same quantity as that measured (even though they might be referred to by the same name) because of scale effects, nonlinearities or measurement technique problems. A soil moisture variable, for example, might be predicted as an average over a model grid element several 100 m in spatial extent and over a certain time step; the same variable might be measured at a point in space and time by a small gravimetric sample, or by time domain reflectrometry integrating over a few tens of cm, or by a cross-borehole radar or resistivity technique, integrating over several tens or hundreds of meters. Only the latter might be considered to approach the same variable as described by the model (but will also be associated with its own interpretation errors of the geophysical signal in obtaining a soil water estimate). Fortunately, in rainfall-runoff modeling, the predictions are most usually compared with the measured discharges at the outlet from a catchment area. The measured discharges may be considered to be essentially the same variable as that predicted by the model, although subject to measurement errors.

In general, no satisfactory approach to separate the sources of error that contribute to the total error has yet been proposed. In line with traditional statistical estimation theory, the total error is often treated as a single lumped additive variable of the form:

$$Q(t) = M(\Theta, t) + \varepsilon(t) \tag{4}$$

where $Q(t)$ is a measured variable, such as discharge, at time t; $M(\Theta,t)$ is the model prediction of that variable using parameter set Θ; and $\varepsilon(t)$ is the total remaining error. Note that a multiplicative error form could also be used if appropriate by taking logs of the observation and the predicted variables. The additive form allows application of the full range of statistical estimation techniques to model calibration and uncertainty analysis, subject to the limitations of dealing with the nonlinearity of the model.

Implicit in this approach, however, is the assumption that the model structure is at least adequate, if not correct. In catchment RR modeling we cannot generally be sure of this. Further, it is generally necessary to make assumptions about the structure of the total errors $\varepsilon(t)$ – typical assumptions include normality of the underlying error distribution, constancy of variance and simplicity of the correlation structure. While such assumptions are convenient to the application of statistical theory, they have usually not been supported by the actual series of model residuals, which

may show variations in bias (nonstationarity), variance (heteroscedasticity), and correlation structures under different hydrologic conditions.

In the following sections we analyze the main contributors to model error and to the uncertainty in predictions of our current models in general. Approaches to deal with these are outlined and open research questions are discussed.

CONSIDERING PARAMETER UNCERTAINTY

It was mentioned above that a large number of widely different parameter sets could, in many cases, yield practically identical results with respect to a particular OF. Three main responses to this problem of perceived equifinality can be found in the literature.

First, the finding of parameter indeterminacy can be interpreted as an indication that the chosen model structure is overly complex given the information about hydrologic behavior actually observable in the data. Following this interpretation, various researchers have tested and successfully applied simpler model structures such that the number of associated model parameters is only so large as to allow confidence in the results of the calibration (Wheater *et al.*, 1993; Jakeman and Hornberger, 1993; Young *et al.*, 1996; Young, 2001; Wagener *et al.*, 2002, 2003c).

Second, the finding can be interpreted as supporting the need for set theoretic approaches, which assume that all plausible models should be retained unless and until evidence to the contrary becomes apparent. Many of these set theoretic approaches are related to the Regional Sensitivity Analysis (RSA; also sometimes called the *Hornberger-Spear-Young approach*) concept advanced by Spear and Hornberger (1980) that evaluates the sensitivity of the model output to changes in parameters without referring to a specific point in the parameter space (such as a most likely value for a parameter). These techniques commonly apply random sampling procedures to explore the feasible parameter space in search for plausible (behavioral) models. Examples of the set theoretic approach applied to RR modeling include the Generalized Likelihood Uncertainty Estimation (GLUE) technique of Beven and Binley (1992), the Dynamic Identifiability Analysis (DYNIA) approach of Wagener *et al.* (2003a), the PIMLI approach of Vrugt *et al.* (2002), the Monte Carlo set membership (MCSM) approach of van Straten and Keesman (1991), the explicit Bayesian approach of Kuczera and Mroczkowski (1998), the Bayesian Recursive Estimation (BARE) technique of Thiemann *et al.* (2001), and the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm of Vrugt *et al.* (2003b).

Third, the finding can be attributed to a failure to properly specify the (automatic) calibration problem in such a way as to properly exploit the information contained in the data. There are two (related) responses to this. On the one hand, a multicriteria approach can result in better exploitation of the information in the data, while the resulting optimal population solution (optimal in a multicriteria sense) defines a kind of parameter uncertainty attributable to model structural errors. This is discussed further in the section "Considering structural uncertainty". In addition to the improved use of information, recursive processing of the data can provide better extraction of the information in the data, because the temporal aggregation associated with batch processing of data is reduced. Examples of recursive algorithms that can be applied to RR models for the estimation of parameter uncertainty include the Kalman Filter and its extensions (e.g. Kitanidis and Bras, 1980a,b; Beck, 1987), the PIMLI approach (Vrugt *et al.*, 2002), the DYNIA approach (Wagener *et al.*, 2003a), the BaRE approach (Thiemann *et al.*, 2001) and the application of Transfer Functions (TF) with time-varying parameters identified using Instrumental Variable techniques (Young, 2001). The GLUE methodology can also be applied recursively with appropriate choice of the Likelihood criterion. Recursive approaches can also provide a method for checking violations of the underlying assumption that the parameters are constant – for example, both Wagener *et al.* (2003a) and Misirli (2003, BaRE2) reported that certain parameters of the models they tested displayed significant temporal variations, and suggested that this may be an indication of model structural error.

We now take a closer look at four approaches representative of the second and third responses to dealing with parameter uncertainty, that is, the use of set theoretic methods, GLUE and SCEM-UA, and the application of recursive approaches, DYNIA and BaRE. Common to these approaches is the selection or identification of a *set* (population) of models (different combinations of model structures and parameter values), and assignment of some relative degree of believability to each member of the set. That degree of believability is translated into interval estimates of the uncertainty (confidence) in model simulations/predictions. The approaches differ in the suite of assumptions underlying each technique, based on which the methods used to compute the relative degree of believability are derived. The idea is that, in principle, the sensitivity of the predictions and associated uncertainty to the underlying assumptions of the methods are testable and can be evaluated.

The GLUE methodology evolved out of early Monte Carlo studies of different realizations of parameter sets in rainfall-runoff modeling as a way of estimating the sensitivity of model predictions to different parameter values (see, e.g., Hornberger and Spear, 1981). The Hornberger-Spear-Young approach to sensitivity analysis involves the classification of many different parameter sets into those that are behavioral and those that are nonbehavioral in some predefined way. The parameter distributions for each

of these sets are then examined to check for significant differences between them.

The additional step in the GLUE methodology of Beven and Binley (1992) is to associate each of the behavioral simulations with a likelihood measure (as a way of quantifying model believability), to estimate the uncertainty associated with the model predictions as conditioned on some calibration data used in the likelihood value calculation. Models that perform well in the calibration period will be associated with a high likelihood weight in prediction, those that perform less well will carry a low likelihood weight. Those that are considered nonbehavioral will not be used in prediction. Such an approach allows for the equifinality of different parameter sets in fitting the available calibration data, it allows that different types of likelihood measure might be used in model evaluation, and it allows for the likelihood weights to be updated as new calibration data become available (see the review of applications by Beven and Freer, 2001). The steps in the GLUE methodology are as follows:

- Decide on a model structure or structures to be used.
- Sample multiple sets of values from prior ranges or distributions of the unknown parameters by Monte Carlo sampling, ensuring independence of each sample in the model space.
- Evaluate each model run by comparing observed and predicted variables.
- Calculate a likelihood measure or measures for those models considered behavioral.
- Rescale the cumulative likelihood weights over all behavioral models to unity.
- Use all the behavioral models in prediction, with the output variables being weighted by the associated rescaled likelihood weight to form a cumulative distribution of predictions, from which any desired prediction quantiles can be extracted.

The methodology depends on obtaining an adequate sample of behavioral models. In some model applications this may require from many thousands to even billions of simulations to adequately define the behavioral regions in the multidimensional model space (Iorgulescu *et al.*, 2005). In most problems it remains quite difficult to define prior distributions for effective values of the different parameters, even in terms of simple means and covariances. Thus, most applications of GLUE define only a range for each parameter and sample uniformly within that range. For cases where the region of behavioral models is relatively small this will result in inefficient sampling and alternatively sampling strategies may be required. However, in many applications it has been found that behavioral models are scattered widely in the model space so that efficiency in sampling behavioral models is difficult to achieve. It is, of course, possible that no behavioral models will be found, particularly where models must satisfy multiple criteria to remain behavioral (see for example Freer *et al.*, 2003) giving a good indication that there are problems either with the data set or with the model structure. A statistical estimation strategy might still find an optimal model in such a case, but would generally assign the deficiencies to a large "total error" component (unless other statistical inadequacy components had been added, as for example in Kennedy and O'Hagan, 2001).

In fact, the GLUE methodology is general, in that statistical error assumptions and likelihood functions can be used where the assumptions are satisfied, but a much wider range of likelihood measures including fuzzy measures can also be used. The only requirements are that the relative likelihood value should increase monotonically with improving model performance and that nonbehavioral models should be given a likelihood of zero. These are much less stringent requirements than those of statistical theory and the GLUE approach does avoid the assumption that the model is correct (implicit where a statistical error model with zero bias is assumed in model calibration). As a result, unlike a statistical approach, GLUE does not attempt to estimate the probability of predicting an observation given the (optimal) model. Instead, it predicts the probability of a model prediction, conditioned on the ranges of model structures and parameter values considered, the period of calibration (or evaluation) data used and the likelihood measures used in model evaluation. The method therefore assumes that the error structures associated with a particular behavioral model parameter set will remain "similar" during any prediction period, so that the likelihood weights determined in calibration can be used to weight the predictions of any variable of interest. In this way, distributional assumptions are avoided, and nonlinear changes in predicted distributions of variables are allowed (as demonstrated in Freer *et al.*, 1996, where such changes between high and low discharges are shown to be consistent with expectations of system response). The emphasis is on the parameter set in obtaining a behavioral model, rather than on individual parameters and their covariation. The approach can be extended to including multiple model structures, provided that different models can be evaluated with respect to the same likelihood measures. Different applications of the GLUE methodology are described in Beven *et al.* (2000), Beven and Freer (2001), and Beven (2001).

Vrugt *et al.* (2003b) extended the SCE-UA algorithm (described above) to allow for both the estimation of the most likely parameter set, and also for its underlying posterior distribution. The authors replaced the Downhill Simplex method used for population evolution by Duan *et al.* (1992) with the Metropolis Hastings (MH) algorithm. By merging the strengths of the MH algorithm, controlled random search, competitive evolution, and complex shuffling, the SCEM-UA is designed to evolve to a stationary

posterior target distribution of the parameters. The stochastic nature of the MH annealing scheme avoids the tendency of the SCE-UA algorithm to collapse into a single region of attraction (i.e. the global minimum), while the information exchange (shuffling) between parallel sequences allows the search to be biased in favor of better regions of the solution space. Examination of the posterior parameter distribution allows the user to detect whether multiple and/or large regions of the parameter space continue to remain consistent with the current data, and this parameter uncertainty can be projected into the output space as uncertainties on the model predictions (in a manner similar to both BaRE and GLUE). A detailed description of this algorithm appears in the article by Vrugt and Dane (**Chapter 77, Inverse Modeling of Soil Hydraulic Properties, Volume 2**).

The dynamic identifiability analysis (DYNIA) approach developed by Wagener *et al.* (2003a) is a recursive parameter identification approach, based on elements of the GLUE methodology, the popular Regional Sensitivity Analysis (RSA; Spear and Hornberger, 1980), aspects of wavelet analysis (e.g. Gershenfeld, 1999), and the use of Kalman filtering for hypothesis testing as applied by Beck (1987). Monte Carlo sampling based on a uniform prior distribution is used to examine the feasible parameter space. In contrast to the GLUE procedure, DYNIA uses only a window period rather than the full data period to calculate a measure of performance. This window is moved through the available time-series in a step-wise fashion, that is, at time step t one considers the residuals between $t - n$ and $t + n$. The size of n is selected depending on the length of time over which the parameter is influential, and on the quality of the data. The performance measure is then used to condition the marginal parameter distribution at that particular time step. A threshold is applied to consider only the best performing parameter sets (e.g. best 10%), which represent the peak of the parameter distributions. The shape of the resulting distribution is projected into the time-parameter space and variation of its shape in time can be visualized. This methodology can be applied to track the variation of parameter optima in time, to separate periods of information and noise, or to test whether model components (and therefore parameter values) represent those processes they are intended to represent (Wagener, 2003).

The Bayesian Recursive Estimation (BaRE) algorithm, developed by Thiemann *et al.* (2001) based on assumptions similar to those used in the SCEM-UA batch calibration algorithm, employs a recursive scheme for tracking the conditional probabilities associated with several competing parameter sets (models) in an on-line mode instead of searching for a single best solution in an off-line mode. The parameter probabilities are used to compute probabilistic predictions of the desired output variables (Figure 2). Probability updating, via Bayes theorem, facilitates the assimilation of new data as they become available. The BaRE algorithm belongs to a broad class of ensemble methods, which include the Ensemble Kalman Filter (EnKF, see e.g. Evensen, 1994; Madsen and Canizares, 1999; Reichle *et al.*, 2002) and which use multiple possible model realizations (possibly involving multiple parameter sets, model structures, error sequences, etc.). The main difference is that BaRE employs a full nonlinear updating procedure, while the EnKF uses a linear correlation updating rule. The initial BaRE algorithm suffered from several shortcomings (Beven and Young, 2003; Gupta *et al.*, 2003; Misirli, 2003), the most important being that the parameter distribution collapsed onto a single point owing to an insufficient sampling density. Misirli (2003) addressed this and other problems by developing BaRE2, an improved version of the original BaRE algorithm, introducing (among other things) a resampling procedure to ensure an appropriate sampling density in the high probability region. The development of BaRE has helped to stimulate some discussion about appropriate methods for handling various sources of uncertainty, including model structural uncertainty. Please see the comment and reply on this topic published recently in Water Resources Research (Beven and Young, 2003; Gupta *et al.*, 2003).

An alternative to Monte Carlo based approaches to the estimation and propagation of uncertainty are the "point" methods based on first-order analysis. Such techniques can be used to calculate the mean and variance of the predicted variable based on the mean and variance of uncertain inputs and parameters only. They do not require computer intensive Monte Carlo schemes to estimate the shape of the response surface that is then mapped into the output space. These techniques are thus particularly attractive for practical applications. An overview of such methods can be found in Melching (1995). They commonly apply a Taylor series expansion of the OF or the model output around a specific point in the parameter space, usually truncated after the first-order term, hence the term first-order analysis. In the Mean-value First-Order Second-Moment (MFOSM) method, the selected point is the mean value. Numerical and sometimes even analytical derivatives can be used (Melching, 1995) to calculate the expected value and the variance of the predicted variable (e.g. streamflow). An advantage of this approach is its relative simplicity and computational efficiency. The main weakness is *the* assumption that *a single linearization of the system performance function at the central values of the basic variables is representative of the statistical properties of system performance over the complete range of basic variables* (Melching, 1995). This is a difficult assumption to make in RR modeling where the system under investigation usually exhibits a strongly nonlinear behavior. The Advanced First-Order Second-Moment (AFOSM) improves on the MFOSM approach by using a "likely" point in the

parameter space, instead of the mean. Rosenblueth's point-estimation method (Rosenblueth, 1981) uses the mean and the covariance of the variables in a Taylor series expansion and does not require the calculation of derivatives, as do MFOSM and AFOSM. For applications of Rosenblueth's approach to hydrological models, see, for example, Rogers *et al.* (1985), Binley *et al.* (1991), Melching (1992), or McIntyre *et al.* (2002). Harr's point-estimation method (Harr, 1989) reduces the number of simulations required for Rosenblueth's method from 2^p to 2p, where p is the number of model parameters. See McIntyre *et al.* (2002) for details on this method and for an application in hydrology.

All of the point-estimation methods described above, are, however, limited by some assumptions made during their application, most importantly approximate linearity of the model. The first two moments of the predicted variable can sometimes be calculated accurately using these approaches if the nonlinearity of the model (and model structural error) and the uncertainty in the model parameters is not too large (Garen and Burges, 1981; Kuczera, 1988; Høybye, 1998).

CONSIDERING STRUCTURAL UNCERTAINTY

RR models are, by definition, simplifications of reality that aggregate catchment processes into simpler representations. The process of defining a perceptual model of the catchment and translating this model into mathematical code depends on the imagination and hydrologic understanding of the modeler. To corroborate or reject a model as suitable for the anticipated purpose, a procedure of model testing and evaluation must be applied. The imperfect model representation arising from the aggregation process introduces a degree of uncertainty into the model predictions, which is difficult to quantify. However, some of the consequences of this uncertainty can be detected and even used for improvements in the model structure.

A major consequence of model structural imperfection is that the model is incapable of reproducing all aspects and portions of the hydrograph equally well with a single parameter set. Because the classical manual modeling/calibration approach seeks a single "best" parameter set, the hydrologist is forced to select a trade-off between the errors in fit to different parts of the hydrograph, thereby arriving at some suitable compromise parameter set that meets the needs and objectives of the modeling exercise. The manual approach typically depends on visual examination of the "local" fit between various segments of the simulated and observed hydrographs, while also checking to see that some selected "global" OFs take on values that are within acceptable distance of their "optimal" values. The goal is to find a parameter set that produces a realistic hydrograph shape while giving an acceptable level of overall (statistical) performance. The classical single OF automatic calibration approaches result, in essence, in an implicit (difficult to specify or control) aggregate weighting of different aspects of hydrograph fit, which, in practice, tends to produce simulations that are biased towards specific aspects of the observed hydrograph (e.g. high or low flows). To date, it has not become clear if the complex thought processes that lead to successful manual calibration could be encapsulated into a single OF. This has fueled the recent research on multicriteria approaches.

Gupta *et al.* (1998) argued that the calibration of RR models is inherently a multiobjective problem. Their multiobjective approach offers a way forward by emulating the ability of Manual-Expert calibration to employ a number of complementary ways of evaluating model performance, thereby compensating for various kinds of model and data errors, and extracting greater amounts of information from the data. The outcome is a set of models that are constrained (by the data) to be structurally and functionally consistent with available qualitative and quantitative information and which simulate, in an uncertain way, the observed behavior of the watershed. By maintaining the independence of the various performance criteria, and by performing a full multicriteria optimization, the entire set of Pareto optimal solutions is identified. Chankong and Haimes (1993) define the concept of a Pareto optimum as follows: "A solution is said to be Pareto optimal (also synonymously known in the literature as efficient, noninferior, and nondominated) if the value of any OF cannot be improved without degrading at least one of the other Ofs." In simple language, it is possible to divide the parameter space into 'good' and 'bad' solutions, but one cannot objectively favor one of the good solutions, since there is always another one that is better in a certain aspect, that is, with respect to another OF.

Yapo *et al.* (1998), and later Vrugt *et al.* (2003a), presented algorithms capable of solving, in a single optimization run, the multiobjective problem posed by this approach. Yapo *et al.* (1998) developed the multiobjective complex evolution (MOCOM-UA) algorithm that uses a population evolution strategy (similar to the one employed by the SCE-UA algorithm) to converge to the Pareto set. In brief, the MOCOM-UA method involves the initial selection of a "population" of p points distributed randomly throughout the n-dimensional feasible parameter space Θ. In the absence of prior information about the location of the Pareto optimum, a uniform sampling distribution is used. For each point, the multiobjective vector $E(\theta)$ is computed, and the population is ranked and sorted using a Pareto-ranking procedure suggested by Goldberg (1989). Simplexes of $n + 1$ points are then selected from the population according to a robust rank-based selection method. The MOSIM procedure, a multiobjective extension of the Downhill Simplex method (Nelder and Mead, 1965), is used to evolve each simplex in a multiobjective improvement direction. Iterative application of the ranking and evolution procedures causes the entire population to converge towards the Pareto

optimum. The procedure terminates automatically when all points in the population become nondominated.

Vrugt *et al.* (2003a) replaced the Downhill Simplex approach in the MOCOM-UA algorithm with an efficient Markov Chain Monte Carlo sampler (similar to the one employed in the SCEM-UA algorithm), which additionally allows for the estimation of the underlying probability distributions with respect to the different objective functions. They also improved on some of the weaknesses of the original approach such as the tendency of the MOCOM-UA algorithm to cluster the Pareto solutions in the most compromise region among the objectives, and premature convergence of the algorithm for cases involving a multitude of parameters and highly correlated performance measures. The multiobjective implementation of the SCEM-UA algorithm is termed *MultiObjective SCEM-UA* (MOSCEM-UA).

Several researchers used the fact that different parameter sets are required to represent different response modes of the hydrologic system in a more structured manner (e.g. Beck, 1985; Young, 2001; Wagener *et al.*, 2003a). Beck (1987) applied the extended Kalman filter (EKF) to recursively estimate optimum parameter values, and used the variation of these optima in time to detect structural inadequacies. He reported that the EKF was inadequate for this purpose owing to a lack of robustness and owing to restrictions imposed by the filter assumptions. Wagener *et al.* (2003a) used the DYNIA methodology described earlier for the same purpose and showed structural problems in a RR model having typical structural elements by tracking parameter variations in time.

Kennedy and O'Hagan (2001) on the other hand, focused on deriving a more complex error model. They include a model inadequacy function in their Bayesian calibration framework. The goal, similar to the approach by Sorooshian and Dracup (1980; see next section), is to produce an error series having the desirable properties of constant variance and independence in time and space, so that unbiased estimates of the various parameters and correction terms can be more easily estimated. If the model structure is at least approximately correct, the statistical approaches lead to a concentration on finding a probabilistic description for the optimal values of the model and error parameters. Where there are model structural errors on the other hand, the statistical error model will, to some extent, be required to compensate for those errors. It is therefore important to carry out postcalibration diagnostic tests to ensure that the model and error assumptions are sound.

Another approach, arising from the equifinality concept (where more than one model or model structure appear to provide acceptable representations of the available observations; Beven, 1993), suggests that complications arising owing to the presence of nonlinear model structural error make it difficult to properly apply a rigorous statistical estimation procedure. This approach rejects the idea that an "optimal" model exists and concentrates instead on the task of finding a *set* of models that are *behavioral* in the sense of being acceptably consistent with the observations (however acceptable might be defined), or, more importantly, rejecting all those models that can be shown to be nonbehavioral. As discussed earlier, this is the basis for the Generalized Likelihood Uncertainty Estimation methodology and other set theoretic approaches to model calibration, which are easily extended to consider multiple objectives and multiple model structures, at the expense of significant additional computer run time. Within this framework, as noted earlier, it is possible that all the models tried will be rejected where such consistency with the observations cannot be demonstrated (Freer *et al.*, 2003; Beven, 2005), leading to the serious reconsideration of model structure, input data, or calibration data that would be justified in such a case.

Other researchers have also explored the same premise, that is, the need to consider multiple model structures. Neuman (2002), for example, suggests that not allowing for different system conceptualizations can lead to statistical bias and an underestimation of uncertainty. He introduces an approach based on Bayesian model averaging (Hoeting *et al.*, 1999) to account for this problem, where each model is treated as if it were the correct structure in trying to maximize its contribution to the averaging process.

CONSIDERING DATA UNCERTAINTY

Data used for RR modeling are measurements of the input and output fluxes of the hydrologic system, and sometimes of its states. The input is precipitation, usually as rainfall or snow, while output data are streamflow and potential evapotranspiration. The latter is sometimes replaced by measurements of temperature. State variables that are of potential use in rainfall-runoff modeling are, for example, measurements of groundwater levels or soil moisture content.

Measurement errors with respect to streamflow occur owing to underflow and bypassing of gauging structures, and rating curve inaccuracies, especially at very high and very low flows. Sorooshian and Dracup (1980) addressed the problem of heteroscedastic errors in streamflow measurements by deriving a likelihood estimator, which uses a power transformation to stabilize the variance. However, research with respect to data uncertainty and its effect on the predictions of RR models has focused mainly on errors in the precipitation. This focus is based on the assumption that the dominant source of error stems from poor knowledge of the rainfall input. One of the earliest examples in this respect is the work of Crawford and Linsley (1966), who used a rain gauge scaling factor as a calibration parameter in the Stanford Watershed Model, with values up to 1.1 to

account for wind or orographic influences. The use of such a factor can still be found in some of today's modeling exercises. Kavetski *et al.* (2003), for example, included an adjustment coefficient for each rainstorm within a statistical estimation framework.

Much of the error in precipitation measurements is related to the inability of available gauging networks to properly capture the amount and variability of precipitation in space and time. Not surprisingly, Beven and Hornberger (1982) and Obled *et al.* (1994) found that a correct assessment of the global volume of rainfall input in a variable pattern is more important than a rainfall pattern (by itself) for simulating streamflow hydrographs. However, the variability of the rainfall pattern can exert a strong influence on the timing of the hydrograph peak at a downstream gauging station. Initial studies on precipitation error focused on how well networks of rain gages were capable of estimating the actual total rainfall, first using synthetic data (Wei and Larson, 1971; Troutman, 1983; Watts and Calver, 1991) and then real data (Krajewski *et al.*, 1991; Ogden and Julien, 1993, 1994; Shah *et al.*, 1996a). Such studies commonly assume that the highest resolution data available are approximately representative of the real pattern of precipitation behavior. The use of radar rainfall estimates has now become increasingly common and has been tested either individually or in combination with gauged data (e.g. Smith *et al.*, 1996; Moore and Hall, 2000; Morin *et al.*, 2001; Carpenter *et al.*, 2001). Also under investigation is the use of satellite-based remotely sensed information for deriving precipitation estimates (Hsu *et al.*, 1997; Sorooshian *et al.*, 1993, 2000), which has the potential of providing global estimates of precipitation. Satellite-based precipitation estimates enable some knowledge of rainfall over the relatively extensive ungaged portions of the world (including locations that are difficult to access or are blocked from radar coverage owing to topography), and are particularly useful for large (regional and continental) scale hydrologic studies. Both radar and satellite estimates of rainfall, however, are dependent on interpretative models of the recorded signals that also have parameters subject to calibration. It is also important to consider that errors in the precipitation data will also introduce a bias on the parameter estimates that in turn impacts the model predictions (e.g. Troutman, 1983; Andréassian *et al.*, 2001; Kavetski *et al.*, 2003).

The results of the above mentioned studies suggest that the importance of capturing the spatial variability of rainfall depends significantly on whether the catchments are infiltration- or saturation-excess dominated (e.g. Ogden and Julien, 1994; Koren *et al.*, 1999). Spatial variability of rainfall seems to be of particular importance for infiltration-excess dominated catchments. In such catchments, the location of runoff production typically shows a stronger correlation with the location of high rainfall intensity (Michaud and Sorooshian, 1994; Winchell *et al.*, 1998). On the other hand, other factors such as the topography of the catchment can have a stronger influence on the location of runoff production in saturation-excess dominated catchments. The spatial distribution of rainfall can also be of higher importance in cases where the catchment is dry (Shah *et al.*, 1996b). Smith *et al.* (1993) suggest that results related to the importance of estimating precipitation variability should be treated with some degree of caution at this early stage of research, and that it needs to be demonstrated that the sensitivity of the models used in the aforementioned studies is actually representative of the sensitivity of the real catchments.

DISCUSSION AND CONCLUSIONS

This article began with the premise that all rainfall-runoff models are (at some level) lumped and conceptual representations of a real-world system. The main consequences of this premise are that the model structure is defined, prior to any modeling being undertaken, by the modeler's understanding of the natural system, and that estimates of the model parameters must be provided. Such parameter estimates are typically derived through a process of model calibration using observed system behavior. Three necessary conditions for a rainfall-runoff model to be considered as being properly calibrated are that: (i) the input-state-output behavior of the model is consistent with the measurements of catchment behavior, (ii) the model predictions are accurate (i.e. they have negligible bias) and precise (i.e. the prediction uncertainty is relatively small), and (iii) the model structure and behavior are consistent with the hydrologists understanding of reality. With respect to all three aspects, problems have been encountered that are still not satisfactorily solved to this day. Regarding the first point, it has been found that different parameter sets are required to simulate different behaviors of the natural system. This is usually taken to be an indication of model structural problems. Multiobjective approaches can be used to consider this problem, and recursive methods can be applied to more objectively track parameter variation in time. At this time, however, structured approaches to improve model representations are only available for certain simple types of models (e.g. linear). With respect to the second aspect, it is often found that the estimate of prediction uncertainty is relatively large and depends on the type of approach chosen to analyze it. To be blunt, there is currently no unifying framework that properly addresses uncertainty in hydrological modeling. Statistical approaches (such as SCEM and BaRE) require assumptions that are often difficult to justify; alternative approaches (such as GLUE and DYNIA) require subjective decisions about model evaluations that can also be difficult to justify. And finally, the issue of the realism of hydrologic models in current practice is receiving increasing attention. Many modeling approaches

have been based on the powerful mathematics of regression and systems theory with insufficient consideration for the conceptual nature of the model structure and parameters. Recent developments suggest that proper consideration of this issue is required if successful prediction of ungauged catchments or those undergoing land-use changes is to be achieved (Beven, 2002, 2005).

We remain confident, of course, that other strategies to understanding and dealing with the various sources of error, including those arising from model structural deficiencies, will emerge as increasing numbers of intelligent and energetic minds are brought to bear on the problem. If nothing else, history teaches us that the progress of science is inexorable, and that today's "truths" are all too often tomorrow's "mistakes"! However, the nature of hydrological systems is such that even if new measurement techniques become available in the future, uncertainty in hydrological prediction will not be eliminated. Thus, hydrology as science must learn to be realistic about the uncertainties that arise in the modeling process.

Acknowledgments

Partial support for the first and third authors was provided by SAHRA under NSF-STC grant EAR-9876800, and the National Weather Service Office of Hydrology under grant number NOAA/NA06WH0144. The third author also received support from the German Academic Exchange Service (DAAD) under the auspices of its postdoctoral fellowship program. We thank Mary Black for her editing work. We also greatly appreciate the constructive criticism provided by George Kuczera, which resulted in considerable improvements to the manuscript.

SOFTWARE LINKS TEXTBOX

Hydrologic software, including many of the above-described algorithms (including SCE-UA, SCEM-UA, BaRE, etc.), is available for noncommercial use at **`http://www.sahra.arizona.edu/software.html`** (Wagener *et al.*, 2004). Demonstration GLUE software can be downloaded from **`http://www.es.lancs.ac.uk/glue.html`**. The Monte Carlo Analysis Toolbox (MCAT) is a compilation of several techniques to analyze the parameter and output space including the corresponding uncertainties. Any dynamic mathematical model can be analyzed for which a Monte Carlo sampling or a population evolution procedure can be run. The Toolbox includes the DYNIA methodology, elements of the GLUE procedure and multiobjective plots as explained later. Copies of the MCAT can be obtained from **`http://ewre-www.cv.ic.ac.uk/software.`** Other popular optimization packages, not explicitly mentioned in the text, are NLFIT (developed by George Kuczera, **`http://www.eng.newcastle.edu.au/~cegak/`**) and PEST (developed by John Doherty, **`http://www.sspa.com/pest/`**).

FURTHER READING

Beven K.J. (1989) Changing ideas in hydrology - the case of physically-based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (2001) *Rainfall-Runoff Modelling – The Primer*, John Wiley & Sons: Chichester. (This book provides both a primer for the novice and a detailed and practical description of techniques and difficulties demanded by more advanced users and developers.).

Boyle D.P., Gupta H.V., Sorooshian S., Koren V., Zhang Z. and Smith M. (2001) Towards improved streamflow forecasts: the value of semi-distributed modelling. *Water Resources Research*, **37**, 2739–2759.

Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (2003) *Calibration of Watershed Models*, Water Science and Application Series, 6, American Geophysical Union: Washington. (This edited monograph is a collection of papers representing a state-of-the-art analysis of mathematical methods used in the identification of models for hydrologic forecasting, design, and water resources management. This book is suitable for scientists, researchers and students of watershed hydrology, practicing hydrologists, civil and environmental engineers, and water resource managers.).

Ehrgott M. (2000) *Multicriteria Optimization*, Springer-Verlag: Berlin.

Finnerty B.D., Smith M.B., Seo D.-J., Koren V. and Moglen G.E. (1997) Space-time sensitivity of the Sacramento model to radar-gage precipitation inputs. *Journal of Hydrology*, **203**, 21–38.

Gupta H.V., Sorooshian S. and Boyle D.P. (2001) Assimilation of data for the construction and calibration of watershed models. *Keynote Paper Presented at the International Workshop on Catchment Scale Hydrologic Modelling and Data Assimilation*, Wageningen.

Harlin J. (1991) Development of a process oriented calibration scheme for the HBV hydrologic model. *Nordic Hydrology*, **22**, 15–36.

Hornberger G.M., Beven K.J., Cosby B.J. and Sappington D.E. (1985) Shenandoah watershed study: calibration of the topography-based, variable contributing area hydrologic model to a small forested catchment. *Water Resources Research*, **21**, 1841–1850.

Johnston D., Smith M., Koren V. and Finnerty B. (1999) Comparing mean areal precipitation estimates from NEXRAD and rain gauge networks. *ASCE Journal of Hydrologic Engineering*, **4**(2), 117–124.

Kuczera G. and Williams B.J. (1992) Effect of rainfall errors on accuracy of design flood estimates. *Water Resources Research*, **28**(4), 1145–1153.

Legates D.R. and McCabe G.J. Jr (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, **35**, 233–241.

Madsen H. (2000) Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of Hydrology*, **235**(3–4), 276–288.

Madsen H., Wilson G. and Ammentorp H.C. (2002) Comparison of different automated strategies for calibration of rainfall-runoff models. *Journal of Hydrology*, **261**, 48–59.

Nash J.E. and Sutcliffe J.V. (1970) River flow forecasting through conceptual models, I, A discussion of principles. *Journal of Hydrology*, **10**, 282–290.

Pilgrim D.H. (1983) Some problems in transferring hydrologic relationships between small and large drainage basins and between regions. *Journal of Hydrology*, **65**, 49–72.

Piñol J., Beven K.J. and Freer J. (1997) Modelling the hydrologic response of Mediterranean catchments, Prades, Catalonia. The use of distributed models as aid to hypothesis formulation. *Hydrological Processes*, **11**, 1287–1306.

Popper K. (2000) *The Logic of Scientific Discovery*. First published 1959 by Hutchinson Education, Routledge.

Seibert J. (2000) Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, **4**, 215–224.

Singh V.P. (1997) Effect of spatial and temporal variability in rainfall and watershed characteristics on stream flow hydrograph. *Hydrological Processes*, **1**, 1649–1669.

Smith M.B., Koren V.I., Zhang Z., Reed S.M., Pan J.-J. and Moreda F. (2004) Runoff Response to Spatial Variability In Precipitation: An Analysis Of Observed Data. *Journal of Hydrology*, **298**(1–4), 267–286.

Sorooshian S. and Gupta V.K. (1985) The analysis of structural identifiability: theory and application to conceptual rainfall-runoff model's. *Water Resources Research*, **21**, 487–495.

Sorooshian S., Gao X., Hsu K., Maddox R.A., Hong Y., Imam B. and Gupta H.V. (2002) A diurnal variability of tropical rainfall retrieved from combined GOES and TRMM satellite information. *Journal of Climate*, **15**, 983–1001.

Uhlenbrock S., Seibert J., Leibundgut C. and Rohde A. (1999) Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure. *Hydrological Sciences Journal*, **44**, 779–797.

Wagener T., Wheater H.S. and Gupta H.V. (2003) *Rainfall-Runoff Modelling in Gauged and Ungauged Catchments*, Imperial College Press: London. (This monograph is an extensive treatment of the theory, development and application of conceptual lumped rainfall-runoff models to gauged and ungauged catchments. This book will be valuable for hydrologic researchers, graduate students, and everybody who applies lumped catchment-scale models in operational or research settings.).

Watts L.G. and Calver A. (1991) Effects of spatially-distributed rainfall on runoff for a conceptual catchment. *Nordic Hydrology*, **22**, 1–14.

Wheater H.S., Bishop K.H. and Beck M.B. (1986) The identification of conceptual hydrologic models for surface water acidification. *Hydrological Processes*, **1**, 89–109.

Yan J. and Haan C.T. (1991) Multiobjective parameter estimation for hydrologic models – weighting of errors. *Transaction of the American Society of Agricultural Engineers*, **34**(1), 135–141.

Young C.B., Bradley A.A., Krajewski W.F. and Kruger A. (2000) Evaluating NEXRAD multisensor precipitation estimates for operational hydrologic forecasting. *Journal of Hydrometeorology*, **1**, 241–254.

REFERENCES

Andréassian V., Perrin C., Michel C., Usart-Sanchez I. and Lavabre J. (2001) Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. *Journal of Hydrology*, **250**(1–4), 206–223.

Bard Y. (1974) *Nonlinear Parameter Estimation*, Academic Press.

Beck M.B. (1985) Structures, failure, inference and prediction. In Barker M.A. and Young P.C. (Eds.), *Identification and System Parameter Estimation: Proceedings 7th Symposium Volume 2, July 1985*, IFAC/IFORS, York, UK, pp. 1443–1448.

Beck M.B. (1987) Water quality modelling: a review of the analysis of uncertainty. *Water Resources Research*, **23**, 1393–1442.

Beven K.J. (1993) Prophecy, reality and uncertainty in distributed hydrologic modelling. *Advances in Water Resources*, **16**, 41–51.

Beven K.J. (2001) *Rainfall-Runoff Modelling – The Primer*, John Wiley & Sons: Chichester.

Beven K.J. (2002) Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system. *Hydrological Processes*, **16**(2), 189–206.

Beven K.J. (2005) A manifesto for the equifinality thesis. *Journal of Hydrology*, (in press).

Beven K.J. and Binley A.M. (1992) The future of distributed models: model calibration and uncertainty in prediction. *Hydrological Processes*, **6**, 279–298.

Beven K.J. and Freer J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems. *Journal of Hydrology*, **249**, 11–29.

Beven K.J., Freer J., Hankin B. and Schulz K. (2000) The use of generalised likelihood measures for uncertainty estimation in high order models of environmental systems. In *Nonlinear and Nonstationary Signal Processing*, Fitzgerald W.J., Smith R.L., Walden A.T. and Young P.C. (Eds.), CUP: pp. 115–151.

Beven K.J. and Hornberger G.M. (1982) Assessing the effect of spatial pattern of precipitation in modelling stream flow hydrographs. *Water Resources Bulletin*, **18**, 823–829.

Beven K.J. and Young P.C. (2003) Comment on "Bayesian recursive parameter estimation for hydrologic models". By M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian. *Water Resources Research*, **39**(5), 1116, doi:10.1029/2001WR001183, COM 1–1 to 1–4.

Binley A.M., Beven K.J., Calver A. and Watts L.G. (1991) Changing responses in hydrology: assessing the uncertainty in physically based model predictions. *Water Resources Research*, **27**(6), 1253–1261.

Blazkova S., Beven K.J. and Kulasova A. (2002a) On constraining TOPMODEL hydrograph and simulations using partial saturated area information. *Hydrological Processes*, **16**, 441–458.

Blazkova S., Beven K.J. and Kulasova A. (2002b) On constraining TOPMODEL hydrograph and simulations using partial saturated area information. *Hydrological Processes*, **16**, 441–458.

Box G.E.P. and Tiao G.C. (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley-Longman: Reading.

Boyle D.P., Gupta H.V. and Sorooshian S. (2000) Towards improved calibration of hydrologic models: combining the strengths of manual and automatic methods. *Water Resources Research*, **36**, 3663–3674.

Brazil L.E. (1988) *Multilevel Calibration Strategy for Complex Hydrologic Simulation Models*, Unpublished Ph.D. Thesis, Colorado State University, Fort Collins.

Brazil L.E. and Krajewski W.F. (1987) Optimisation of complex hydrologic models using random search methods. *Engineering Hydrology Proceedings, Hydraulics Division/ASCE/Williamsburg*, Virginia, pp. 726–731.

Carpenter T.M., Georgakakos K.P. and Spersflage J.A. (2001) On the parametric and NEXRAD-radar sensitivities of a distributed hydrologic model suitable for operational use. *Journal of Hydrology*, **253**, 169–193.

Chankong V. and Haimes Y.Y. (1993) Multi-objective optimization: pareto optimality. In *Concise Encyclopaedia of Environmental Systems*, Young P.C. (Ed.) Pergamon Press: pp. 387–396.

Crawford N.H. and Linsley R.K. (1966) *Digital Simulation in Hydrology: Stanford Watershed Model IV*. Department of Civil Engineering, Stanford University, Technical Report No. 39, California, US.

Duan Q., Gupta V.K. and Sorooshian S. (1992) Effective and efficient global optimisation for conceptual rainfall-runoff models. *Water Resources Research*, **28**, 1015–1031.

Duan Q., Sorooshian S. and Gupta V.K. (1993) Shuffled complex evolution approach for effective and efficient global minimization. *Journal of Optimization Theory and Applications*, **76**(3), 501–521.

Duan Q., Sorooshian S. and Gupta V.K. (1994) Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, **158**, 265–284.

Evensen G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**, 10143–10162.

Franks S., Gineste P., Beven K.J. and Merot P. (1998) On constraining the predictions of a distributed model: the incorporation of fuzzy estimates of saturated areas into the calibration process. *Water Resources Research*, **34**, 787–797.

Freer J., Beven K.J. and Ambroise B. (1996) Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Water Resources Research*, **32**, 2161–2173.

Freer J.E., Beven K.J. and Peters N.E. (2003) Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure. In *Advances in Calibration of Watershed Models, AGU Monograph Series, US*, Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (Eds.), pp. 69–87.

Garen D.C. and Burges S.J. (1981) Approximate error bounds for simulated hydrographs. *Journal of the Hydraulics Division, American Society of Civil Engineers*, **107**(HY11), 1519–1534.

Gershenfeld N. (1999) *The Nature of Mathematical Modelling*, Cambridge University Press: Cambridge.

Goldberg D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Reading, MA, p. 412.

Gupta V.K. and Sorooshian S. (1985) The relationship between data and the precision of parameter estimates of hydrologic models. *Journal of Hydrology*, **81**, 57–77.

Gupta H.V., Sorooshian S., Hogue T.S. and Boyle D.P. (2003) Advances in the automatic calibration of watershed models. In *Calibration of Watershed Models*, Water Science and Application Series, 6, Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (Eds.), American Geophysical Union: Washington, pp. 9–28.

Gupta H.V., Sorooshian S. and Yapo P.O. (1998) Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research*, **34**, 751–763.

Gupta H.V., Thiemann M., Trosset M. and Sorooshian S. (2003) Reply to comment by K. Beven and P. Young on "Bayesian recursive parameter estimation for hydrologic models". *Water Resources Research*, **39**(5), 1117, doi:10.1029/2002WR001405, 2003, COM 2-1 to 2-5.

Harr M.E. (1989) Probability estimates for multivariate analyses. *Applied Mathematical Modelling*, **13**, 313–318.

Hendrickson J.D., Sorooshian S. and Brazil L. (1988) Comparison of Newton-type and direct search algorithms for calibration of conceptual rainfall-runoff models. *Water Resources Research*, **24**(5), 691–700.

Hoeting J.A., Madigan D., Raftery A.E. and Volinskiy C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, **14**(4), 382–417.

Holland J.H. (1975) *Adaption in Natural and Artificial Systems*, University of Michigan Press: Ann Arbor.

Hooke R. and Jeeves T.A. (1961) Direct search solutions of numerical and statistical problems. *Journal of the Association for Computing Machinery*, **8**(2), 212–229.

Hornberger G.M. and Spear R.C. (1981) An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management*, **12**, 7–18.

Høybye J.A. (1998) *Uncertainty Modelling in Hydrological Applications*, Unpublished Ph.D. Dissertation, Department of Water Resources Engineering, University of Lund, Sweden.

Hsu K., Gao X., Sorooshian S. and Gupta H.V. (1997) Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, **36**(9), 1176–1190.

Ibbitt R.P. (1970) *Systematic Parameter Fitting for Conceptual Models of Catchment Hydrology*, Unpublished Ph.D. Thesis, Imperial College of Science, Technology and Medicine, London.

Iorgulescu I., Beven K.J. and Musy A. (2005) Data-based modelling of runoff and chemical tracer concentrations in the Haute-Menthue (Switzerland) research catchment. *Hydrological Processes*, in press.

Jakeman A.J. and Hornberger G.M. (1993) How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, **29**, 2637–2649.

Jakeman A.J., Hornberger G.M., Littlewood I.G., Whitehead P.G., Harvey J.W. and Bencala K.E. (1992) A systematic approach to modelling the dynamic linkage of climate, physical catchment descriptors and hydrologic response components. *Mathematics and Computers in Simulation*, **33**, 359–366.

Johnston P.R. and Pilgrim D.H. (1976) Parameter optimisation for watershed models. *Water Resources Research*, **12**(3), 477–486.

Kavetski D., Franks S.W. and Kuczera G. (2003) Confronting input uncertainty in environmental modelling. In *Calibration of Watershed Models*, Water Science and Application Series, 6, Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (Eds.), American Geophysical Union: Washington, pp. 49–68.

Kennedy M.C. and O'Hagan A. (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, **63**(3), 425–464.

Kirkpatrick S., Gelatt C.D. Jr and Vecchi M.P. (1983) Optimisation by simulated annealing. *Science*, **220**, 671–680.

Kitanidis P.K. and Bras R.L. (1980a) Real-time forecasting with a conceptual hydrologic model - part I: analysis of uncertainty. *Water Resources Research*, **16**(6), 1025–1033.

Kitanidis P.K. and Bras R.L. (1980b) Real-time forecasting with a conceptual hydrologic model - part II: applications and results. *Water Resources Research*, **16**(6), 1034–1044.

Klepper O., Scholten H. and van de Kamer J.P.G. (1991) Prediction uncertainty in an ecological model of the Oosterschelde Estuary. *Journal of Forecasting*, **10**, 191–209.

Koren V.I., Finnerty B.D., Schaake J.C., Smith M.B., Seo D.-J. and Duan Q.Y. (1999) Scale dependencies of hydrologic models to spatial variability of precipitation. *Journal of Hydrology*, **217**, 285–302.

Koren V.I., Smith M. and Duan Q. (2003) Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall-runoff models. In *Calibration of Watershed Models*, Water Science and Application Series, 6, Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (Eds.), American Geophysical Union: Washington, pp. 239–254.

Krajewski W.F., Lakshim V., Georgakakos K.P. and Jain S.C. (1991) A Monte Carlo study of rainfall-sampling effect on a distributed catchment model. *Water Resources Research*, **27**(1), 119–128.

Kuczera G. (1983a) Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. *Water Resources Research*, **19**(5), 1151–1162.

Kuczera G. (1983b) Improved parameter inference in catchment models: 2. Combining different kinds of hydrologic data and testing their compatibility. *Water Resources Research*, **19**(5), 1163–1172.

Kuczera G. (1988) On the validity of first-order prediction limits for conceptual hydrologic models. *Journal of Hydrology*, **103**, 229–247.

Kuczera G. and Mroczkowski M. (1998) Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resources Research*, **34**, 1481–1489.

Lamb R., Beven K.J. and Myrabo S. (1998) Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. *Advances in Water Resources*, **22**, 305–317.

Madsen H. and Canizares R. (1999) Comparison of extended and ensemble Kalman filter for data assimilation in coastal area modelling. *International Journal of Numerical Methods Fluids*, **31**, 961–981.

Masri S.F., Bekey G.A. and Safford F.B. (1980) A global optimization algorithm using adaptive random search. *Applied Mathematics and Computation*, **7**, 353–375.

McIntyre N., Wheater H.S. and Lees M.J. (2002) Estimation and propagation of parametric uncertainty in environmental models. *Journal of Hydroinformatics*, **4**(3), 177–198.

Melching C.S. (1992) An improved first-order reliability approach for assessing uncertainties in hydrologic modelling. *Journal of Hydrology*, **132**, 157–177.

Melching C.S. (1995) Reliability estimation. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publishers, pp. 69–118.

Michaud J.D. and Sorooshian S. (1994) Effect of rainfall-sampling errors on simulation of desert flash floods. *Water Resources Research*, **30**(10), 2765–2775.

Misirli F. (2003) *Improving Efficiency and Effectiveness of Bayesian Recursive Parameter Estimation for Hydrologic Models*, Ph.D. Dissertation, Department of Hydrology and Water Resources, The University of Arizona, Tucson.

Moore R.J. and Hall M.J. (Eds.) (2000) HYREX: the hydrology radar experiment. Special Issue of *Hydrology and Earth System Sciences*, **4**(4), 681.

Morin E., Enzel Y., Shamir U. and Garti R. (2001) The characteristic time scale for basin hydrologic response using radar data. *Journal of Hydrology*, **252**, 85–99.

Nelder J.A. and Mead R. (1965) A simplex method for function minimisation. *Computer Journal*, **7**, 308–313.

Neuman S.P. (2002) Accounting for conceptual model uncertainty via maximum likelihood Bayesian model averaging. *Acta Universitatis Carolinae – Geologica*, **46**(2/3), 529–534.

Obled C.H., Wendling J. and Beven K. (1994) The sensitivity of hydrologic models to spatial rainfall patterns: an evaluation using observed data. *Journal of Hydrology*, **159**, 305–333.

Ogden F.L. and Julien P.Y. (1993) Runoff sensitivity to temporal and spatial rainfall variability at runoff plane and small basin scales. *Water Resources Research*, **29**(8), 2589–2597.

Ogden F.L. and Julien P.Y. (1994) Runoff model sensitivity to radar rainfall resolution. *Journal of Hydrology*, **158**, 1–18.

Pickup G. (1977) Testing the efficiency of algorithms and strategies for automatic calibration of rainfall-runoff models. *Hydrological Sciences Bulletin*, **12**(2), 257–274.

Price W.L. (1987) Global optimisation algorithms for a CAD workstation. *Journal of Optimization Theory and Applications*, **55**(1), 133–146.

Reichle R.H., McLaughlin D.B. and Entekhabi D. (2002) Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review*, **130**, 103–114.

Rogers C.C.M., Beven K.J., Morris E.M. and Anderson M.G. (1985) Sensitivity analysis, calibration and predictive uncertainty of the Institute of Hydrology Distributed Model. *Journal of Hydrology*, **81**, 179–191.

Rosenblueth E. (1981) Two-point estimates in probabilities. *Applied Mathematical Modelling*, **5**, 329–335.

Rosenbrock H.H. (1960) An automatic method for finding the greatest or least value of a function. *Computer Journal*, **3**, 175–184.

Seibert J. and McDonnell J.J. (2003) Multi-criteria calibration of conceptual rainfall-runoff models – the quest for an improved dialog between modeler and experimentalist. In *Advances in*

Calibration of Watershed Models, AGU Monograph Series, USA, Duan Q., Gupta H.V., Sorooshian S., Rousseau A.N. and Turcotte R. (Eds.), American Geophysical Union, pp. 301–315.

Shah S.M.S., O'Connell P.E. and Hosking J.R.M. (1996a) Modelling the effects of spatial variability in rainfall on catchment response. 1. Formulation and calibration of a stochastic rainfall field model. *Journal of Hydrology*, **175**, 66–88.

Shah S.M.S., O'Connell P.E. and Hosking J.R.M. (1996b) Modelling the effects of spatial variability in rainfall on catchment response. 2. Experiments with distributed and lumped models. *Journal of Hydrology*, **175**, 89–111.

Smith J.A., Seo D.-J., Baeck M.L. and Hudlow M.D. (1996) An intercomparison study of NEXRAD precipitation estimates. *Water Resources Research*, **19**(3), 791–810.

Sorooshian S. and Dracup J.A. (1980) Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: correlated and heteroscedastic error cases. *Water Resources Research*, **16**(2), 430–442.

Sorooshian S., Duan Q. and Gupta V.K. (1993) Calibration of rainfall-runoff models: application of global optimisation to the Sacramento soil moisture accounting model. *Water Resources Research*, **29**, 1185–1194.

Sorooshian S. and Gupta H.V. (1995) Model calibration. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.) Water Resources Publishers: pp. 23–68.

Sorooshian S., Gupta V.K. and Fulton J.L. (1983) Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models – influence of calibration data variability and length on model credibility. *Water Resources Research*, **19**(1), 251–259.

Sorooshian S., Hsu K.L., Gao X., Gupta H.V., Imam B. and Braithwaite D. (2000) Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bulletin of the American Meteorological Society*, **81**(9), 2035–2046.

Spear R.C. and Hornberger G.M. (1980) Eutrophication in peel inlet, II, identification of critical uncertainties via generalized sensitivity analysis. *Water Research*, **14**, 43–49.

Thiemann M., Trosset M.W., Gupta H.V. and Sorooshian S. (2001) Bayesian Recursive Parameter Estimation for Hydrologic Models, *Water Resources Research*, **37**(10), 2521–2535.

Thyer M., Kuczera G. and Bates B.C. (1999) Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms. *Water Resources Research*, **35**(3), 767–773.

Troutman B.M. (1983) Runoff prediction errors and bias in parameter estimation induced by spatial variability of precipitation. *Water Resources Research*, **19**(3), 947–964.

Van Straten G. and Keesman K.J. (1991) Uncertainty propagation and speculation in projective forecasts of environmental change: a lake-eutrophication example. *Journal of Forecasting*, **10**, 163–190.

Vrugt J.A., Bouten W., Gupta H.V. and Sorooshian S. (2002) Toward improved identifiability of hydrologic model parameters: the information content of experimental data. *Water Resources Research*, **38**(12), 48-1–48-13 Art. No. 1312.

Vrugt J.A., Gupta H.V., Bastidas L.A., Bouten W. and Sorooshian S. (2003a) Effective and efficient algorithm for multi-objective optimization of hydrologic models. *Water Resources Research*, **39**(8), 1214, 10.1029/2002WR001746.

Vrugt J.A., Gupta H.V., Bouten W. and Sorooshian S. (2003b) A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, **39**(8), 1201, 10.1029/2002WR001642.

Wagener T. (2003) Evaluation of catchment models. *Hydrological Processes*, **17**(16), 3375–3378.

Wagener T., Gupta H.V., Carpenter C., James B., Vázquez R., Sorooshian S. and Shuttleworth J. (2004) A hydroarchive for the free exchange of hydrological software. *Hydrological Processes*, **18**(2), 389–391.

Wagener T., Lees M.J. and Wheater H.S. (2002) A toolkit for the development and application of parsimonious hydrologic models. In *Mathematical Models of Large Watershed Hydrology*, Vol. 1 Singh V.P. and Frevert D. (Eds.), Water Resources Publishers: pp. 87–136.

Wagener T., McIntyre N., Lees M.J., Wheater H.S. and Gupta H.V. (2003a) Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. *Hydrological Processes*, **17**(2), 455–476.

Wagener T., Wheater H.S. and Gupta H.V. (2003b) *Rainfall-Runoff Modelling in Gauged and Ungauged Catchments*, Imperial College Press: London.

Wagener T., Wheater H.S. and Gupta H.V. (2003c) Identification and evaluation of watershed models. In *Advances in Calibration of Watershed Models*, Duan Q., Sorooshian S., Gupta H.V., Rouseau A. and Turcotte R. (Eds.), AGU: Monograph, pp. 29–47.

Wang Q.J. (1991) The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resources Research*, **27**(9), 2467–2471.

Watts L.G. and Calver A. (1991) Effects of spatially-distributed rainfall on runoff for a conceptual catchment, *Nordic Hydrology*, **22**, 1–14.

Wei, Tseng C. and Larson C.L. (1971) *Effects of Areal and Time Distribution of Rainfall on Small Watershed Runoff Hydrographs*, Minnesota University, Minneapolis, Water Resources Research Center, Bulletin No. 30, p. 130.

Wheater H.S., Jakeman A.J. and Beven K.J. (1993) Progress and directions in rainfall-runoff modelling. In *Modelling Change in Environmental Systems*, Jakeman A.J., Beck M.B. and McAleer M.J. (Eds.), John Wiley & Sons: Chichester, pp. 101–132.

Winchell M., Gupta H.V. and Sorooshian S. (1998) On the simulation of infiltration- and saturation-excess runoff using radar-based rainfall estimates: effects of algorithm uncertainty and pixel aggregation. *Water Resources Research*, **34**(10), 2655–2670.

Yapo P.O., Gupta H.V. and Sorooshian S. (1998) Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, **204**, 83–97.

Young P.C. (2001) Data-based mechanistic modelling and validation of rainfall-flow processes. In *Model Validation*, Anderson M.G. and Bates P.G. (Eds.), John Wiley & Sons: Chichester, pp. 117–161.

Young P.C., Parkinson S. and Lees M.J. (1996) Simplicity out of complexity in environmental modelling: occam's razor revisited. *Journal of Applied Statistics*, **23**, 165–210.

132: Rainfall-runoff Modeling for Assessing Impacts of Climate and Land Use Change

AXEL BRONSTERT

Institute for Geoecology, University of Potsdam, Potsdam, Germany

Rainfall-runoff models are frequently used as a tool to assess the impacts of climate and land-use changes on the hydrological cycle. This requires that the applied rainfall-runoff model is able to represent the dynamics of the hydrological cycle of the catchment under study, in particular, if land-use change is of concern, that is, the effects of land-use or land-cover change on the runoff generating processes. This may be accomplished by applying either a process-based model or a conceptual model if the latter's parameters can be regionalized with respect to land-use change. Furthermore, the chosen model must be able to "represent" the influence of different climatic/ meteorological boundary conditions. This implies, of course, that we know the main dynamics of the catchment under consideration.

INTRODUCTION

From the viewpoint of systems-theory, the meteorological conditions and the physiographic characteristics of a landscape are the boundary conditions of the hydrological dynamics of a catchment. For example, rainfall or snowmelt rates serve as the meteorological "Input" into this system and the landscape properties (such as vegetation cover, land-use practice, soil properties, basin and river morphology, and the geological setting) significantly control the relative importance of the different processes of the hydrological cycle. The impacts of any changes in these boundary conditions on the hydrological cycle have been of concern ever since the relationships between external forces (boundary conditions) and the internal dynamics (subprocesses and their interactions) of the hydrological cycle have been known. In this Section, the term "environmental change" is used to summarize changes in climatological and land-use conditions. There are other changes to the boundary conditions which could be included in the term "environmental change", for example, river training measures (see Part 12: Open-channel Flow), water retention in reservoirs, or groundwater use (see Part 13: Groundwater), but those are not discussed in this section. There are specific chapters on land use and climate change in this encyclopedia, for example, Part 16 deals with the topic of land use and water management and Part 17 with the topic of climate change. The reader is therefore referred to these articles for a wide and general discussion of land use and climate change impacts on the water cycle.

The focus of this section is on the possible use of rainfall-runoff models for assessing the impacts of the mentioned environmental changes. Therefore, first a summary of the main issues of climate and land-use change relevant for rainfall-runoff modeling is given. Then, the basic model requirements for modeling the environmental changes, related uncertainties, and possible feedback effects between climate and land use are discussed. This article ends with presentation of a few case studies. There is no attempt to give a comprehensive overview of case studies on this subject nor to list all rainfall-runoff models which have been applied until now for such purposes. Lists of this kind will inevitably be incomplete and pretty soon outdated. It is more important to understand the main model requirements as well as the limitations of such model applications.

THE MAIN ISSUES OF CLIMATE CHANGE FOR RAINFALL-RUNOFF MODELING

What is anthropogenic climate change?

The climate system of our planet is composed of different subsystems, such as the atmosphere, biosphere, cryosphere,

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

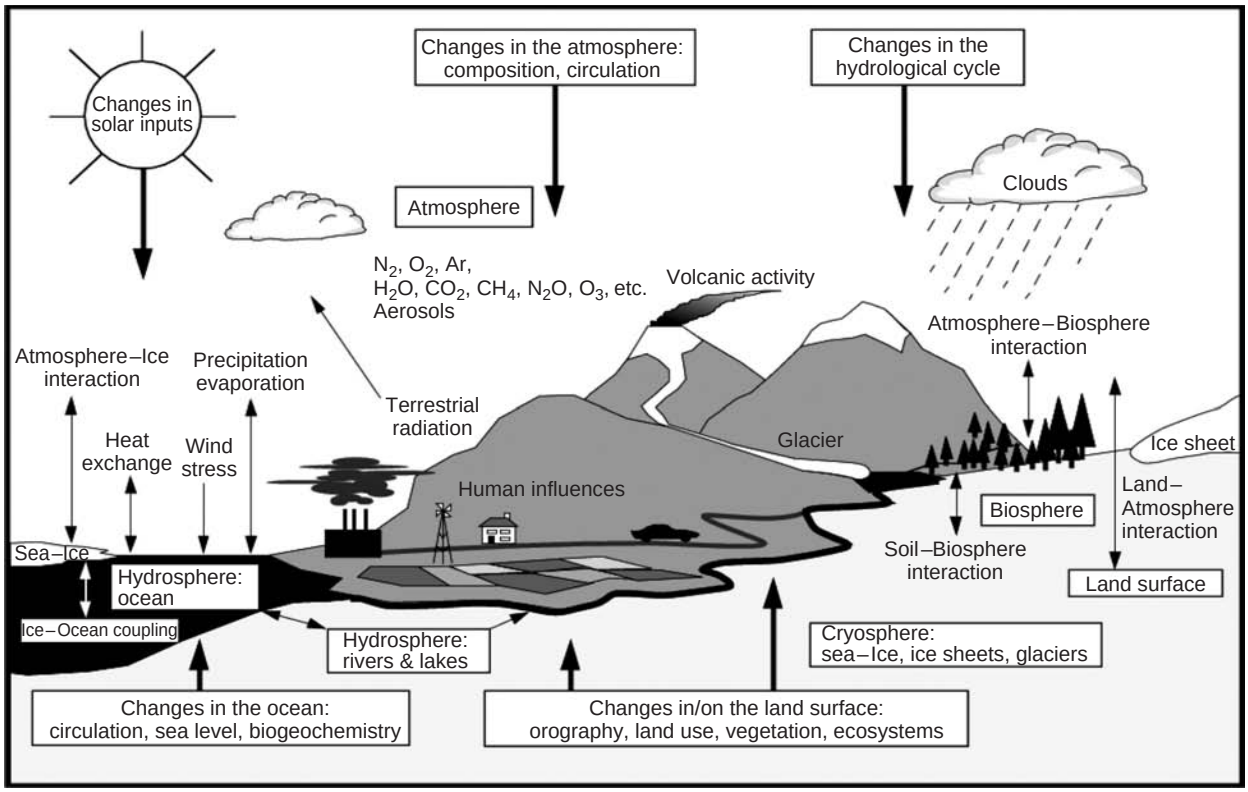


Figure 1 The components of climate system and global change influences (Reproduced by permission of Intergovernmental Panel on Climate Change (IPCC), 2001)

and the terrestrial and maritime hydrosphere. The hydrosphere is one of the most important subsystems, because the hydrological cycle links the energy and water fluxes between the atmosphere, biosphere, and pedosphere (see Figure 1). The climate system was always dynamic, that is, there have always been changes in some of its subsystems or in levels of solar forcing, due to long-term changes of the orbital parameters or the varying activity of the sun. These alterations have caused significant changes in the past in the radiation budget of the atmosphere and the Earth's surface. The changes within the climate system are associated with different scales in space and, more importantly, in time. More details on the climatological timescales can be found in Part 17: Climate Change.

Since the start of industrialization of our Earth (i.e. for about the last 150 years), the use of fossil energy has created a new (additional) source of carbon dioxide (CO_2) emission from the Earth into the atmosphere. Though these emissions are small compared to the natural emissions (mainly from the biosphere), they disturbed the carbon balance of the atmosphere, that is, the CO_2 concentration of the lower atmosphere has increased from a rather stable preindustrial level of 280 ppm to a present level of 360 ppm. Besides CO_2, the concentration of other "greenhouse gases" has increased through human activities. The most important of these increases are for methane (CH_4) which rose from a preindustrial value of 0.7 ppm to a present value of 1.74 ppm, and nitrogen oxide (N_2O), which rose from a preindustrial value of 275 ppb to a present value of 313 ppb. Carbon dioxide and the other "greenhouse gases" affect the atmospheric absorption properties of long-wave radiation, thereby changing the radiation balance. See Part 17: Climate Change for more details on emissions, concentrations, and physical effects off greenhouse gases.

One very important feature of this anthropogenic climate change, compared to natural alterations of the climate system, is the speed of these changes. Never before has the Earth's atmosphere experienced such a rate of change in its chemical composition.

What effects have been observed in recent decades?

Over approximately the last 50 years, a global temperature rise of about 0.6 °C has been observed. Different measurement techniques and data have been used as a basis for this analysis (see IPCC, 2001). The second technical report of the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 1996) concluded, *inter alia*, that the mean summer temperatures in the Northern Hemisphere in the last decade have been the highest experienced since the beginning of

the fifteenth century. On the basis of a wide array of proxy sources (such as ice cores, tree rings, corals, and lake sediments) it has been concluded that the twentieth century was at least as warm as or even warmer than any century of the last millennium. Accordingly, for some places on the earth (especially for high mountain regions), the twentieth century seems to have been the warmest century in several thousand years. Such a trend of temperature increase, as well as an accumulation of warmer years, was not only observed on a global basis, but also in many individual regions of the earth. Higher temperatures are a very important change of the meteorological "input" into rainfall-runoff models. The main point is that higher temperatures will increase calculated evapotranspiration rates, which might cause reduced soil water contents and stronger or earlier water deficit in the root zone. Statistically significant changes of air humidity due to higher temperature (caused by anthropogenic climate change) have not (yet) been monitored. This is the reason why specific air humidity is usually considered unchanged in such modeling studies. Temperature increase can also influence the phenology and composition succession of ecosystems. Such issues are briefly discussed in Section "Interactions of climate change with land-cover conditions".

Regional and long-term observations accompanied by spatial-temporal statistical analysis of precipitation have been carried out in various parts of the globe. Some of them have shown significant trends in precipitation. However, it is very important to realize that differing precipitation trends affect different regions, that is, in some regions there is a notable increase, in some regions there is a decrease, and in other regions no trend has been observed. An example is given in the following text for the Rhine catchment. Measurements over the last 110 years demonstrate an increase in mean precipitation, on a regional scale, in parts of Germany and neighboring areas. According to Engel (1997) the annual precipitation over the Rhine area extending to Cologne has, since 1890, shown a rising overall tendency (increase of ca. 80 mm between 1890 and 1996) with distinctive periodic fluctuations (Figure 2). Apart from the increased amounts of annual precipitation, Engel (1997) also shows a tendency towards a seasonal shift from summer to winter. Combining these trends over the last 100 years, there has been a fairly constant summer precipitation (June–October) and a significant increase in winter precipitation by almost 20% (November–May). The statistical analysis of precipitation trends in Europe between 1891 and 1990 by Rapp and Schönwiese (1996) shows a significant increase in winter precipitation in central and northern Europe and a decrease in southern and southeastern Europe. These results are consistent with the rainfall trends observed in the Rhine basin, which is located in central Europe.

A rather elegant synoptic approach to detect possible regional climatic changes is to analyze the statistics of regional climatic conditions, the so-called *atmospheric circulation patterns* (ACPs), rather than to analyze the different climatic variables independently (Conway and Jones, 1998). For example, Bárdossy and Caspary (1990) and Gerstengarbe *et al.* (2000) have examined the frequency of different types of climate conditions in Europe since 1891 by means of time series derived from daily weather maps. Both studies found a significant increase in the frequency of westerly weather conditions over western Europe, the statistical break point of the time series occurring in the mid-1970s. Because of the lack of earlier weather maps and records, however, it is hardly possible to state to what extent the observed trend (over the past ~120 years) might be explained by longer-term, natural fluctuations.

As an example, the time series of westerly weather conditions and complementary types of weather conditions over western Europe for the months of December and January are shown in Figure 3. The increase of westerly zonal conditions is evident. Since these weather conditions can be considered typical of long-term, large-scale precipitation patterns in western-central Europe, for some catchment areas a correlation has been established between the increase of occurrence of these ACPs and an altered rainfall-runoff behavior of mesoscale catchments. Investigations at many other weather stations in central Europe have confirmed the relationship between westerly/southwesterly weather conditions and heavy precipitation over parts of Europe such as eastern France, Belgium, southwestern Germany, and the western Alps (Bárdossy, 1993).

A similar methodology was applied in a recent study for central Europe by Fricke and Kaminski (2002), who analyzed the ACP frequency of occurrence relevant for heavy summer rainfall over the Danube basin since 1882. They found, for example, for the summer months of June–August, a significant occurrence increase over the past five decades of the ACP "TrM", characterizing atmospheric lows from the Mediterranean to central Europe. This was shown to be associated with an approximate doubling of the probability for heavy and high intensity summer rainfall over the Danube basin.

What development of climatic variables has been projected for the future?

The most relevant climatic variables for rainfall-runoff calculations are precipitation, temperature, air humidity, radiation, and wind velocity, where precipitation and – when snow is melting – temperature are directly relevant for runoff generation at short timescales ("event scale"), and temperature, air humidity, wind velocity, and radiation are relevant for evapotranspiration processes at longer timescales, thus, for example, influencing subsurface runoff and catchment water content preceding a rainfall event.

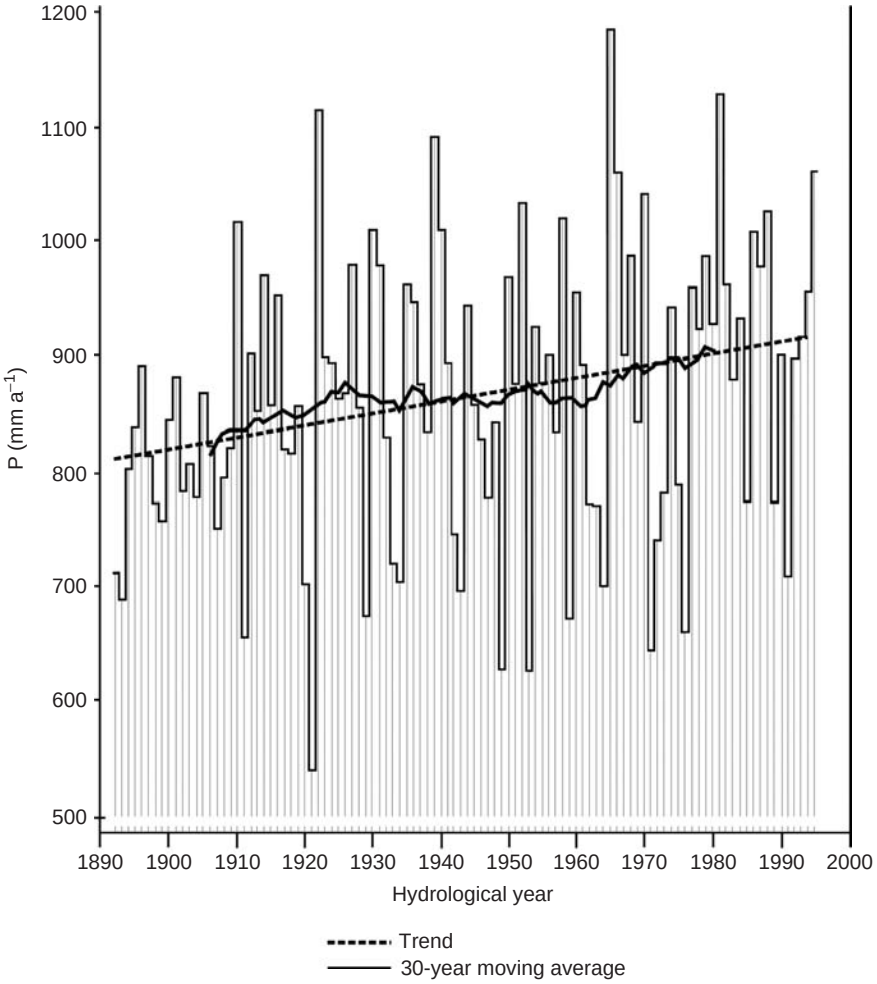


Figure 2 Annual precipitation over the Rhine basin upstream Köln, 1880–1996 (Reproduced from Engel, 1997, Verlag C.F. Müller, Heidelberg)

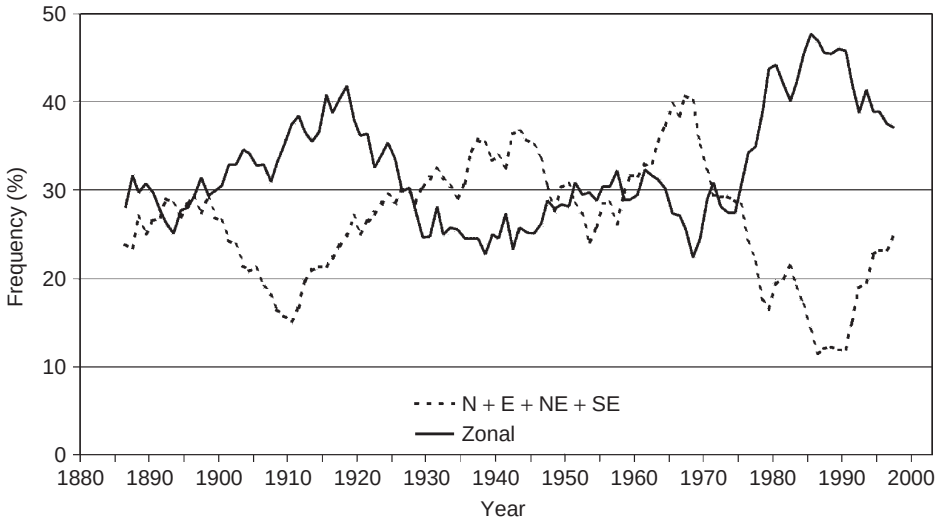


Figure 3 Frequencies of westerly (zonal) weather type and types, north, east, northeast, and southeast (N + E + NE + SE) for December and January from 1881 to 2002 (Bárdossy, personal communication)

The present status of climate change research allows some rather clear statements to be made on the projected global development of radiation and temperature and only a rather vague statement on precipitation, but there is no reliable information on the possible alterations of wind speed and air humidity.

The starting point for an analysis of the development of the global radiation balance is based on scenarios for the emission of greenhouse gases (see Part 17: Climate Change). The third assessment report of the IPCC (IPCC, 2001) lists a range of such scenarios, where the range of emissions varies significantly, depending on the underlying assumptions concerning economic growth, industrial development, and emissions control. On the basis of such emission scenarios it is possible to derive scenarios for the future greenhouse concentration in the atmosphere.

The concentration of greenhouse gases has a direct effect on the atmospheric absorption properties of long-wave (earth) radiation R_{lw}. Shortwave (solar) radiation R_{sw} is almost uninfluenced by these concentrations. Thus, atmospheric absorption of long-wave radiation emitted from the earth increases, resulting in a warming of the lower atmosphere. This yields a reduced long-wave radiation balance at the earth's surface $R_{lw,net}$. For instance, a doubling of the CO_2 concentration is assumed to result in an average reduction of the global long-wave radiation balance ΔR_{lw} of about $3\,W\,m^{-2}$.

According to the projected changes in radiative forcing, a global temperature increase in the range of 1° to 5 °C is to be expected, depending mainly on the emission scenarios; see Figure 4 from IPCC (2001), showing also the range of uncertainty of these projections. IPCC, 2001 also points out that higher latitudes and land surfaces will have to cope with a much higher temperature increase than lower latitudes and sea surfaces. The conclusions are, to some extent, subject to uncertainties, for example, in the prediction of cloud cover and its effects on the global albedo and long-wave radiation.

Predictions of the development of global precipitation are more uncertain than those for temperature. From the viewpoint of global atmospheric energy, one can conclude that an increase of global atmospheric temperature of, say, 3 °C would yield an intensification of the hydrological cycle by about 10%, that is, 10% higher global evaporation and precipitation rates. This is a result of the reduction of net long-wave radiation and an expected decrease of the globally averaged Bowen–Ratio value (see Bronstert *et al.*, 2002 for more details).

The projections of the development of regional precipitation (derived directly from GCM results) are even more uncertain. This is demonstrated by the fact that the results from the different global atmosphere-ocean circulation models (GCMs) show great differences in projected regional precipitation values, and GCMs can rarely reflect the regional precipitation pattern (either spatially or seasonally) for the current, observed climate. Hence, the level of confidence in regional precipitation projections derived directly from GCMs is still rather low. In any case, an assessment of possible changes of rainfall-runoff characteristics resulting from climate change requires reliable

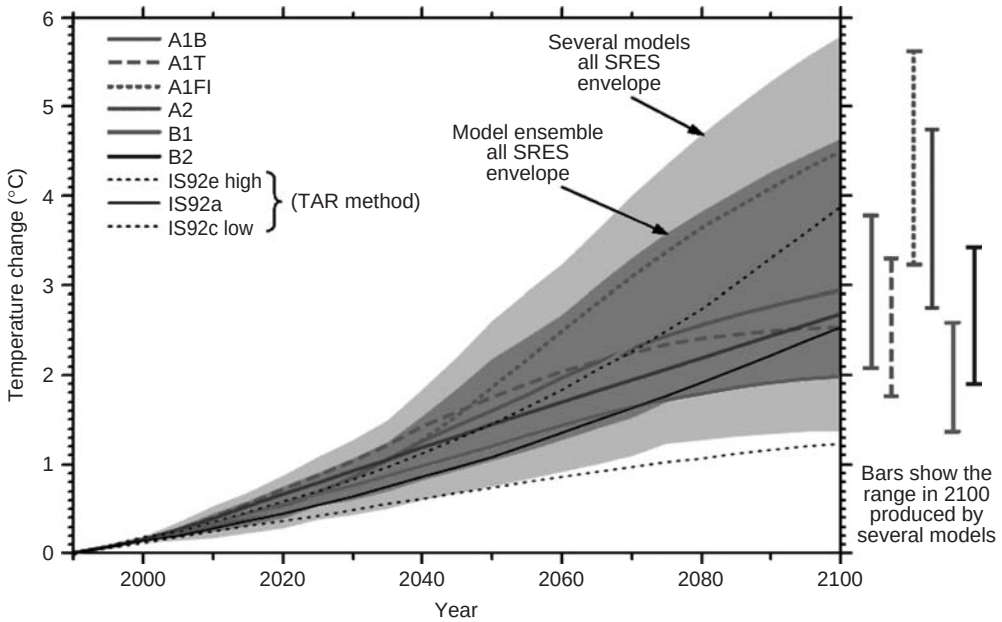


Figure 4 Projections of future development of the global temperature (Reproduced by permission of Intergovernmental Panel on Climate Change (IPCC), 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Table 1 Overview of present knowledge on the projected development of climate variables relevant for rainfall-runoff events. The uncertainty is classified as *very certain, relatively certain, rather uncertain,* and *very uncertain*

Climate variable	Relevance for rainfall runoff	Projected changes due to increase of atmospheric greenhouse gases	Associated uncertainty
Shortwave (solar) radiation balance	Small to moderate	None	Relatively certain
Long-wave (earth) radiation balance	Small to moderate	$-3\,W\,m^{-2}$ (for $2\times CO_2$)	Relatively certain
Air temperature	Moderate	1°–5°	Relatively certain
Precipitation	High	+10% (globally, for $2\times CO_2$)	Globally rather uncertain; regional values very uncertain
Air humidity	Moderate		Very uncertain
Wind speed	Moderate		Very uncertain

information on much smaller scales than the global or continental scale. This is why climate downscaling techniques are an essential tool for climate impact assessment (see Section "Requirements for modeling environmental change impacts on rainfall-runoff events"). Table 1 summarizes present knowledge of the projected development of climate variables relevant for rainfall-runoff events.

Table 1 gives some estimates on the projected changes of climate variables. However, it is important to mention that these numbers are associated with considerable uncertainty and must not be routinely applied in rainfall-runoff modeling for climate impact assessment. Why?

- These numbers show globally averaged values. This scale is not the appropriate one for rainfall-runoff simulations. The regional differences of projected changes of these climate variables are very great and cannot be generalized.
- Furthermore, these numbers show only average values in time. They do not give any information on the extremes (or "anomalies"). The extreme values, however, are of high relevance for rainfall-runoff simulations as the average values.
- The most important variable for rainfall-runoff simulation is precipitation, in particular, precipitation extremes for flood calculations. The associated uncertainty of the statements, however, is especially high for precipitation and highest for extreme values. Even for the simulation of drought events, changes of average (and extreme) precipitation numbers are highly relevant.

THE MAIN ISSUES OF LAND-USE CHANGE FOR RAINFALL-RUNOFF MODELING

Definitions: Land Use, Land Cover, Land Management, River Training

- Land use: According to Mücher *et al.*, (1993), land use can be defined as the human activities that are directly related to land, making use of its resources or having impact on it through interference in ecological processes that determine the functioning of land cover. A very similar definition is given by FAO (1997) and UNEP (1999), who explain land use as "the total of arrangements, activities, and inputs that people undertake in a certain land-cover type".
- Land cover: This term corresponds to a (bio)physical description of the Earth's surface. It is that which overlays or currently covers the ground. This description enables various fundamental biophysical categories to be distinguished, basically areas of vegetation (trees, bushes, fields, lawns), bare soil, hard surfaces (rocks, buildings), and wet areas and bodies of water, for example, watercourses or wetlands (EEA, 2002).
- Land management: This term is directly related to human activities to manage the land's resources. It is influenced by economic, social, environmental, ethical, technological, and other perspectives. This definition sounds similar to the land-use definition given above; however, the various mentioned perspectives may yield categories of land management which have a nonphysical meaning (e.g. "commercial" or "subsistence" land management).
- River training: Human interference with river stretch by means of various river construction measures are summarized by the expression river training. This term is related to the state (including engineering measures) of a river and the corresponding floodplains, that is, it refers only to the river course system and does not include information about the surrounding catchment area. However, river training might be of high importance for the features of rainfall-runoff events.

These definitions are not easily transferable to rainfall-runoff studies. Hence it is proposed to use the following definitions:

- Land use refers to different classes of human activities that are directly related to land (e.g. forest, urban areas,

arable land, meadows). More detailed classifications are possible, for example, distinguishing arable land with/without field drainage or intensively/extensively grazed meadows and so on.
- Land cover refers to the biological and biophysical features of the vegetation covering the land. In this respect, the seasonal dynamics of vegetation or cycles of crop cultivation are of major importance.
- Land management refers to different perspectives, such as economic, social, or technological. There is not always an unambiguous relationship between land management and land use or land cover.
- River training refers to the state of the river and the corresponding floodplains.

With these definitions in mind, it is rather clear that "land use" and "land cover" are of major relevance for rainfall-runoff processes (mainly for runoff-generation processes), river training measures are relevant for the runoff behavior of the river, and land management itself is a less relevant issue in this context.

What is the magnitude of observed land-use change?

The regions of the Earth that are characterized by intense human activity on land resources are almost unavoidably confronted with significant changes of land use. This effect has been proven throughout the entire history of mankind. However, since the era of industrialization and rapid growth of population, land-use change phenomena have accelerated in many regions. Some examples are:

- During the past 50 years, Austria – like almost every other region in Europe – has undergone profound changes in its agriculture, resulting in a reduction of arable land by 14% and an increase of meadows by 24% (Krausmann *et al.*, 2003).
- Between 1938 and 1988, the southwestern German State of Baden-Württemberg underwent a very similar transition: in 1938 58.3% of the land surface was used for agriculture, while 50 years later this figure was only 47.5%. During the same period, urban land use increased from 5.3% to 11.8%.
- Many regions have seen a transition from an agriculturally oriented society to a society dominated by industry. This resulted in an immense increase in urban area. The Ruhr area in Germany, for example, is nowadays one of the most industrialized and most densely populated areas of Europe. The urban area spread over the last century from 6.9% in 1883, to 14% in 1927, to 40% in 1997 (Dosch and Beckmann, 1999).
- Another heavily debated land-use change phenomenon is that of deforestation of tropical forest. For example, DeFries and Achard (2002) reports for the last decade of the twentieth century a deforestation rate of 1.2 million km^2 $year^{-1}$ (0.52×10^6 km^2 in Africa, 0.44×10^6 km^2 in Latin America, and 0.24×10^6 km^2 in Southeast Asia).

Land-use change relevance at what scale and for what climatic conditions?

Part 10: Rainfall-runoff Processes have comprehensively described the different runoff-generation mechanisms and the influence of land-use and land-cover conditions on these. One very important first question for modeling the impact of land-use change on rainfall-runoff processes is the relevance of surface versus subsurface runoff-generation processes in the area under investigation. If the surface runoff processes play an important or even dominant role for rainfall-runoff events, changes on the ground surface may significantly alter those runoff processes. If, however, rainfall-runoff events are characterized mainly by subsurface processes, land-use changes may be of minor importance. This means that for modeling land-use change impacts, it is important that the model represents the main relevant runoff-generation processes, in particular, surface runoff-generation processes.

A second important question is the relevance of surface conditions for the mechanisms by which water produced or triggered by rainfall or snowmelt can travel to a stream ("runoff concentration processes"). These mechanisms (see Part 10: Rainfall-runoff Processes for their explanation) are influenced by the topographic, pedologic–geologic, and vegetational conditions in which they occur. For different climate conditions, different mechanisms of runoff generation and runoff concentration may dominate the rainfall-runoff process. The higher the infiltration-excess overland flow, the greater the importance of ground surface conditions, and therefore the greater the importance of changes to the ground conditions. Infiltration excess is relevant for high intensity rainfall events and/or low conductive soil surfaces. Such conditions are of particular importance (i) in the case of convective rainstorms and (ii) for sealed, crusted, and nonmacroporous soil surfaces. The latter conditions might be exacerbated by removal of vegetation cover. Note that the combination of both conditions is present in many tropical and subtropical regions, while removal of vegetation and the creation of roads and other low permeability surface (e.g. associated with forest logging) is a side effect of land-use change phenomena in the same regions.

In any given catchment, the relative importance of some or all mechanisms of runoff generation and runoff concentration may vary seasonally or even during an individual rainfall-runoff event. Thus the relative importance of land-use changes may vary, too.

As mentioned before, changes in land use and land cover, land management, and river training measures can influence mechanisms of runoff generation, runoff concentration, and

Table 2 Potential impact of land-use changes on surface and near-surface hydrological processes (fluxes or storages) and relevance for runoff components of the hydrological cycle

Process	Potential impact of land-use changes and relevance for components of the hydrological cycle
Interception storage	Greatly affected by vegetation changes (e.g. crop harvest, forest cutting): relevant for evapotranspiration/energy balance
Litter storage	Affected by vegetation changes, in particular, forest cutting: relevant for evapotranspiration/energy balance
Root zone storage	Affected by management practices like tilling method and so on: relevant for evapotranspiration and storm-runoff generation
Infiltration-excess overland flow	Affected by crop cultivation and management practices: relevant for storm-runoff generation in case of high rainfall intensities and low soil conductivity; may be enhanced by soil siltation and crusting and reduced by field drainage and good tillage management
Saturation-excess overland flow	Only slightly affected by land-use changes (process is controlled by topography and subsurface conditions)
Subsurface stormflow	Can be increased by artificial drainage systems, for example, mole or tile drains. Otherwise, subsurface stormflow is only slightly affected by land-use changes (process is controlled by topography and subsurface conditions)
Runoff from urbanized areas	Highly affected by sewer system and sewage retention measures: relevant for storm runoff *from urban areas*
Decentralized retention in the landscape	Affected by landscape structuring and agricultural rationalization of arable land: relevant for storm-runoff concentration *from arable land*

river discharge conditions at different space and timescales. It is evident that a rainfall-runoff event at the catchment scale is a composite of all such mechanisms. Hence, modeling land-use change impacts on such events requires models where the relevant mechanisms are represented by the model. In the following text, the focus is on modeling the impacts on runoff generation and runoff concentration, whereas the impact of river training on discharge conditions is dealt with in Part 12: Open-channel Flow.

As explained before, the influence of land-use practices on runoff generation is most important to the surface or near-surface zones of the soil at least under "normal" conditions; that is, subsurface land uses like mining or underground constructions are not part of these considerations. This means that it is the surface or near-surface fluxes and storages which are mainly affected by land-use changes. This leads to a list of the potential impacts of land-use changes on hydrological fluxes and storages (Table 2).

From Table 2, it is clear that only certain flux and storage processes are *both* affected by land-use changes *and* are primarily relevant for rainfall-runoff events, namely, root zone storage, infiltration-excess overland flow, runoff from urbanized areas, and decentralized retention in the landscape. Artificial drainage of farm land increases subsurface flow velocity but may also increase the storage capacity of the upper soil, which results in a delay or decrease of subsurface storm flow that is, an evaluation of drainage impacts needs to account for both aspects. In any case, an evaluation of land-use change impacts requires identification of the relevant storm-runoff generation mechanisms for the specific *catchment characteristics* and *precipitation conditions*. For different categories of rainstorms (e.g. convective or advective rainstorms), different runoff-generation processes can be relevant and contribute in varying proportions to total runoff. This requires that the modeling of storm-runoff generation has to be both catchment-specific and event-specific. Furthermore, the interactions between precipitation conditions and soil-surface conditions (e.g. soil sealing due to high rainfall intensity) might have to be taken into account.

REQUIREMENTS FOR MODELING ENVIRONMENTAL CHANGE IMPACTS ON RAINFALL-RUNOFF EVENTS

Modeling Climate Change Impacts

Modeling climate change impacts on rainfall-runoff events directly implies that an altered climatic forcing is applied to the rainfall-runoff model. In most climate change impact studies, the mechanisms of the hydrological model remain unchanged, that is, the parameterization of land cover and the algorithms of the hydrological processes remain unchanged. This means that most climate change studies are still based on the assumption that an altered climate system will encounter unmodified land cover and ground surface conditions. Section "Interactions of climate change with land-cover conditions" discusses some examples of effects of climate change on land-cover conditions, that is, for such studies a changed climate forcing *and* changed land-cover parameters have to be applied jointly.

The most important constraint when applying climate change scenarios for hydrological climate change impact analysis is the difference in the spatial scales typical for

climate models and rainfall-runoff models. This means that the best currently available information of future climatic developments, given at a global scale, has to be transformed for use in catchment-scale rainfall-runoff models. This procedure is termed *climate downscaling*.

The most powerful climate models today are coupled GCMs, which carry out three-dimensional calculations of the equations for mass and energy transport, impulse, humidity of the atmosphere, and salt content (of the ocean) for the entire globe. An assessment of possible changes of rainfall-runoff characteristics resulting from climate change requires much smaller scales than the global or continental scale. The following information is required:

- A realistic description of changes in precipitation. This includes both changes of the average value and of the anomalies in space and time. Scenarios that only give changes in the average value are hardly sufficient for an impact analysis, in particular, for flooding conditions;
- A realistic description of changes in temperature. This is particularly important for catchments where rainfall runoff can be composed of *both* rainfall and snowmelt events, which is the case, for example, in many mountainous or high latitude catchments. Besides, temperature may significantly influence evapotranspiration, thereby altering soil water content and long-term characteristics of base flow and low flow periods;
- Information about the uncertainty associated with the climate scenarios. This may form the basis for a thorough uncertainty analysis of the coupled simulations of climate and flood hydrology.

A variety of techniques have been developed to derive climate change scenarios for the catchment scale (or "regional scale" in meteorological terms) required for modeling the hydrological, basin-wide impacts of climate change. The most important of these are regional climate models (or "dynamic downscaling") and "statistical downscaling" methods. In the following paragraph, these methods are summarized, while a more in-depth discussion is given in Part 17: Climate Change.

Regional climate models (RCMs) have been applied for some time. In contrast to general circulation models, they cover only a section of the globe, which can be modeled at a finer spatial resolution. At present, the grid widths used for regional climate models are approximately 50 km or less. In comparison to a GCM resolution, the spatial resolution of an RCM is definitely more adequate for the estimation of runoff-relevant precipitation, particularly with regard to weather conditions connected with large-scale precipitation fields. However, to obtain accurate information on the location, quantity, and intensity of precipitation and on changes in precipitation characteristics caused by climate change, the models are still not sufficiently spatially detailed or accurate for the following reasons (see, e.g. Xu, 1999):

- The boundary conditions of the regional model are obtained from the GCM, and therefore frequently contain a systematic error of atmospheric dynamics, which is transferred to the respective region. Errors in the GCM thus directly limit the capacity of the regional climate model.
- The parameterization of important processes in the RCM, such as the formation of clouds, soil water dynamics, or land-surface interactions, has not yet been resolved in a way that allows for a definition of the natural variability under any weather condition or for the recognition of a possible signal of climate change.
- The resolution of the RCMs is sufficiently detailed to represent large-scale precipitation patterns. However, these resolutions are not sufficient to cover small-scale precipitation, such as convective thunderstorms of local orographic rainfall. Though processes taking place at a smaller scale than represented by the grid box ("subgrid-scale processes") can be parameterized by the subdivision of the grid boxes into a clouded and a cloud-free section, several convective systems cannot be localized within a grid square.
- Regional climate models have not yet been sufficiently tested with regard to how realistically they represent rainfall-runoff events caused by storm rainfall.

Statistical downscaling bridges the two different scales by establishing empirical (statistical) relationships between large-scale features simulated reliably by the GCMs (such as geopotential height fields) and regional or local climate variables (such as temperature and precipitation at a certain location). While dynamic downscaling utilizes physical principles only, the empirical approach is anchored in the observed fact that weather phenomena are often caused by the conditions of the prevailing large-scale atmospheric circulation. The method of GCM downscaling, in its various forms, has by now been well established as an appropriate and necessary tool for impact assessment studies, and the reader is referred for more information to the review article of Wilby & Wigley (1997). Statistical downscaling techniques have been applied in a series of studies (for an overview, see, e.g. Wilby *et al.*, 1999).

There are two main categories of empirical downscaling: deterministic, regression-type methods and weather-type techniques that include stochastic weather generators. Because of the limited correlation between daily circulation and local precipitation, the simulated variability of regression models (including neural networks) is too low, and extreme events cannot be modeled at all (Weichert and Bürger, 1998). Weather-pattern models utilize a finite set of specific circulation patterns, which tend to persist for a certain period of time. Within such a regime, daily precipitation is modeled stochastically using some form of weather generator with regime-dependent parameters. This technique has been successfully applied in a number of studies

(see, e.g., Bárdossy and Plate, 1992; Bárdossy, 1997; Conway and Jones, 1998). Note, however, that the problem of pattern classification introduces subjective elements and the method implicitly assumes that climatic change will not introduce any new weather patterns. *Expanded downscaling* (EDS, Bürger, 1996) is a method midway between the deterministic regression-type models and the stochastic models conditioned on weather types. EDS results from relaxing the unconditional error minimization of regression to allow for the preservation of local variability as a side condition.

As mentioned before, most studies of climate change impacts on hydrology are performed on the basis of the same algorithms for hydrological processes and unmodified land-cover parameterization (see Figure 5). This means that such studies prescribe the altered climate conditions to the same hydrological model and the same land-cover parameterization. This assumption is reasonable as long as the altered climate does not cause significant changes in land-surface conditions (such as a different vegetation cover) or important changes in the dominance of hydrological processes (for example, significant changes of rainfall intensity may change the relevance of infiltration-excess surface runoff). The question of whether the feedback effects between climate, the hydrological conditions of the catchment, and the land cover have to be taken into account is very much related to the timescale of the study (Bronstert *et al.*, 2005). For example, if the study covers short to medium timescales (rainfall-runoff events or event series of a duration, e.g. not longer than a few decades) these feedbacks might be omitted. However, for timescales from a few decades up to centuries, the feedback effects of climate and hydrology may result in significant changes of the land cover. In that case, the alterations of climate, hydrological processes, and land-use conditions have to be analyzed in a coupled manner in order to reflect altered systems conditions of this kind.

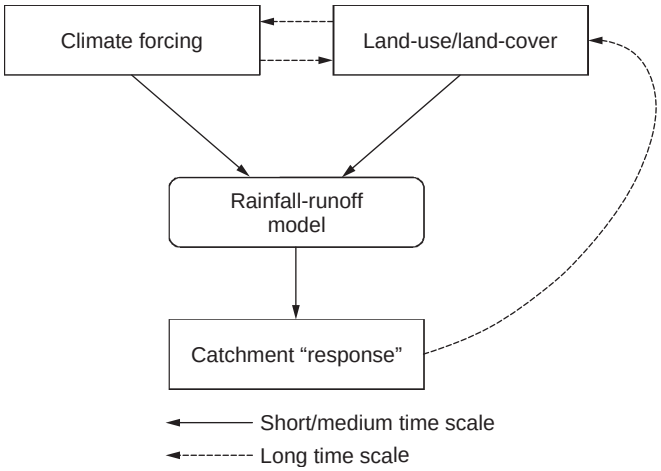


Figure 5 Standard procedure of rainfall-runoff model application for climate and land-use change assessment: the altered hydrological response of a catchment is assessed by application of a hydrological model with altered climatic or boundary conditions. In most cases, the alterations of climate conditions or land-use parameters are performed independently, and the process descriptions of the hydrological model remain unchanged. This approach is appropriate if the modeling study covers short to medium timescales (e.g. not longer then few decades). However, the feedback effects of climate and of the hydrological catchment conditions can result in significant changes in land-cover conditions (e.g. for timescales equal to or longer than a few decades), so that these feedback processes should jointly be taken into account, that is, the alterations of climate, hydrological processes, and land-use conditions have to be analyzed in a coupled manner (indicated by the dotted arrows)

Both climate forcing and hydrological process description may introduce significant uncertainty into the rainfall-runoff modeling. It would therefore be valuable to quantify this uncertainty, or at least, to take this uncertainty into account, (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**). Some remarks on consideration of uncertainty in impact analysis of environmental change are also given in the following text.

Modeling land-use change impacts

As stated before, land-use and land-cover conditions mainly influence surface or near-surface processes; that is, surface or near-surface runoff fluxes are more likely to be affected by land-use changes than the runoff from groundwater. Besides hydrological fluxes (such as infiltration or surface runoff) the capacity of the ground surface to retain water is of importance. In addition to the consideration of fluxes and storage capacity, modeling the effects of land-use change requires that the initial conditions, such as antecedent soil moisture, snow pack, vegetation phenological status, or land cultivation status are appropriately taken into account. Finally, rainfall intensity during the event is of special importance, because it controls the emergence of infiltration excess.

Modeling the impacts of land-use change on rainfall-runoff events requires that the model is able to reflect these surface fluxes, storages, initial and boundary conditions. "Being able" means that the model has to contain algorithms that calculate the relevant processes and storages with model parameters representing the land-use/land-cover conditions and their possible changes.

Hydrological models where fluxes and storage processes are described by equations derived from hydrological process knowledge have been termed *physically based* (e.g. Abbott *et al.* 1986) or *process-based* (e.g. Janssen and Heuberger, 1995) models. Of course, these types of models generally have implicit, empirically derived relations built into them (see e.g. Beven, 2001a). However, compared to empirical or lumped models, these model types bear the

advantage that the different processes can be distinguished and parameterized, which is an essential prerequisite for a model to be able to extract and calculate the effects of land use on hydrological processes.

Thus, assessing land-use change impacts by applying a hydrological model requires a process-based model in which, at least, the hydrological processes on and near the ground surface are included explicitly. However, this does not mean that simplified approaches to describe such processes prohibit the use of this specific model for land-use change impact studies. The necessary and sufficient condition for model selection is the requirement that the relevant rainfall-runoff processes are included and that the specific land-use conditions can be reflected by the model parameters. An overview of some popular hydrological models was put together by, for example, Singh (1995). From the principal requirements stated above, it is possible to specify the features of rainfall-runoff models which are necessary for this kind of impact analysis.

(a) Representation of the soil zone

The condition of the soil surface and the unsaturated zone is crucial for rapid rainfall-runoff processes, for example, infiltration excess, subsurface stormflow, and surface runoff (see Part 10: Rainfall-runoff Processes). Models that do not distinguish different runoff-generation processes (at least, surface, near-surface, and deep-surface processes) and whose parameters do not reflect the state of the soil zone are not advisable, in particular, if effects are to be investigated which may influence the development of a particular process (see Table 2).

(b) Spatial distribution and scale

It is necessary to operate in a spatial resolution of such kind, so that the relevant hydrological processes responsible for the influence of land use on runoff can be represented. This means, that depending on the relevant processes, different spatial resolutions are possible. As an example, if a large-scale change in interception storage is the only relevant process, the spatial resolution can be much coarser compared to a case where small-scale varying Hortonian runoff generation is influenced (amplified or reduced) by the effects of land-use change on the local distribution of soil properties.

A distributed approach is essential if the runoff-generation processes are highly variable in space, especially if this variability can be attributed to soil and vegetation characteristics. A distributed approach is also required if land-use change impacts are to be analyzed in their actual spatial appearance.

(c) Land-use scenarios

While many hydrologists apply rainfall-runoff models with great care and expertise, they tend to be much less aware of the significance of an adequate representation of the boundary conditions controlling the dynamics of the hydrological system. This "representation of the boundary conditions of the systems dynamics" is termed here as *scenario*. Land-use change scenarios need to account for both geophysical and anthropogenic aspects. Designing such scenarios concerning land-use development might consist of two steps: the determination of land-use trends, given as spatially averaged values, and the spatial transformation of these trends into spatially distributed land-use patterns. Many impact studies of environmental changes have neglected neighborhood relationships and the geographic position of land-use features, by adjusting only the proportion of different land-use types. However, if the location of land-use features and changes to these are of importance for the rainfall-runoff model, the land-use change scenario should account for the spatially explicit allocation of land-use change in the landscape, as summarized by for example, Niehoff *et al.* (2002) and described in detail by Fritsch (2002).

(d) Temporal resolution

If rainfall intensity is relevant for runoff generation, this should be reflected in the temporal resolution of both the meteorological data and the modeling time step. This is typically the case in small catchments and/or if Hortonian overland flow contributes significantly to runoff.

(e) Coupling of climate and land cover

If the climate conditions have an *indirect impact on the runoff-generation conditions*, for example, by causing crusting of the soil, altering the characteristics of the vegetation cover and so on, this needs to be taken into account, especially for long-term projections of the development of rainfall-runoff characteristics, see Section "Interactions of climate change with land-cover conditions".

UNCERTAINTY ISSUES

It has to be emphasized that the capability of rainfall-runoff models to forecast the hydrological response of a catchment after changes in land use or climate conditions is severely constrained by the uncertainties involved in the whole rainfall-runoff modeling procedure. In (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**), a methodology to estimate and to take into account uncertainty is presented, where the focus is laid on the uncertainty resulting from difficulties in identifying "optimal" parameter sets. Other sources of uncertainty can be attributed to the model structure and model approaches (due to insufficient knowledge on the physical appearance and the stochastic features of the processes involved) and to data problems (e.g. constrained data availability, insufficient spatial or temporal resolution, measurement or data interpretation errors).

In the case of rainfall-runoff modeling for environmental impact assessment, a very important additional source of uncertainty is introduced by the prescription of the altered land use or climatic boundary conditions, the so-called *land-use change scenarios* and *climate change scenarios*. This will be termed *scenario uncertainty*. In most past rainfall-runoff studies concerned with climate and land-use change, such scenarios have been established in a rather deterministic "if-then" mode, for example, "if" the scenario conditions are prescribed (given, assumed), "then" the altered hydrological response is simulated by the rainfall-runoff model. In most cases the probability of the scenario ("if" case) has not been considered; actually, the probability that the prescribed scenario may become true is usually completely unknown. In rare cases the probability might be estimated from current/observed trends of greenhouse gas concentrations (basis for climate change scenarios) or from socioeconomic development (basis for land-use change), and these trends might be extrapolated in or predicted for the future on the basis of different assumptions with varying probability of the general economic development.

Though knowledge on the scenario uncertainty usually is poor, it is desirable to include it in an integrated uncertainty analysis and estimation. If the purpose of rainfall-runoff model application is to derive standards for design of, for example, water supply schemes, flood protection, or ecological management of floodplains under conditions of environmental change, it is indeed essential to quantify the uncertainty involved in the model application.

In order to assess the uncertainty resulting from model structure and parameterization, a rigorous procedure such as described by Freer *et al.* (1996) can be carried out, *see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3** for details. Cameron *et al.* (2000) applied this approach for a parameter uncertainty estimation for flood frequency simulation under a changing climate and included different emission scenarios from the UK Hadley Centre (Hulme & Jenkins, 1998) into their analysis. Using the General Likelihood Uncertainty (GLUE) procedure, they compared the results obtained by applying different emission scenarios as boundary conditions for TOPMODEL (e.g. Beven, 2001b) while taking account of the parameter uncertainty. This comparison is an important step towards an assessment of the scenario uncertainty, though the shortcoming remains that no probability can be attributed to the different scenarios. From Figure 6, one can extract different kinds of important information:

1. The extrapolation of flood frequency distribution for low probabilities (long return periods), based on observed values (dark solid line), is very uncertain, due to the very limited (or absent) data for such low probabilities and due to the dependence of the chosen distribution function.

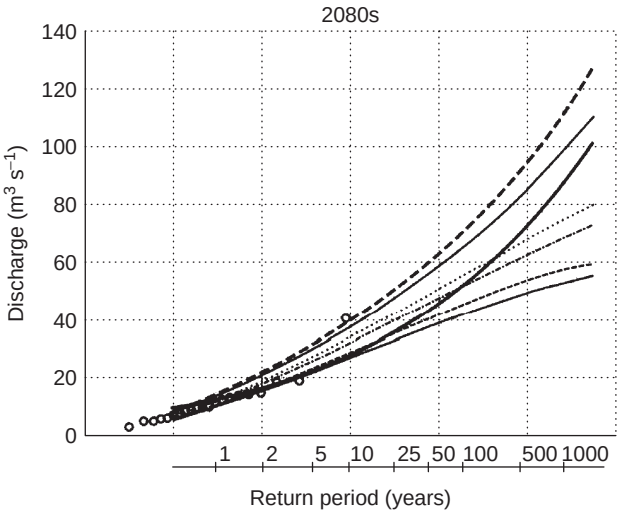


Figure 6 Flood frequency distribution for the Wye catchment (UK) under a changing climate with 90% uncertainty bounds derived from annual maximum peaks obtained by 1000 TOPMODEL-simulations with varying parameter sets. Circles – observed hourly peaks; dark solid line – flood frequency distribution fitted to the observations; dashed lines – 90% uncertainty bounds obtained for climate change scenario (HadCM2 ''medium high'' scenario for 2080s); dotted line – median simulation obtained from climate change scenarios; light solid lines – 90% uncertainty bounds obtained for current conditions; light dash-dot line – median simulation obtained for current conditions, derived from Cameron *et al.* (2000)

2. The median frequencies predicted under current climate conditions (light dash-dot line) yield a distribution which is within the uncertainty bounds for a changed climate (dashed lines) and, of course, within the uncertainty bounds for the current climate conditions.
3. Comparing simulated current climate and changed climate conditions, one can see a slight but notable increase in flood frequency for that particular catchment: for example, from the median results one can derive a return period of 300 years for a flood of 60 m^3 s^{-1} for current climate, and a return period of 200 years under changed climate conditions for the same discharge value.

One has to be aware that, like most other modeling studies of climate change impacts on the hydrological cycle, this study assumed that the functional description of runoff generating processes remains unchanged. That assumption means that an altered climate does not cause a change of the importance of certain specific hydrological processes: if a catchment runoff is fed mainly by saturation-excess overland flow and return flow (a basic TOPMODEL assumption), the same catchment is expected to behave the same under changed climate. Of course, this assumption can be questioned, but for a first step this is acceptable,

because the variations within the existing climate are probably stronger (e.g. between dry and wet years, or between flood and lowflow periods), than the differences between the current and the future climate. In a second step, however, one should check this assumption: does the climate change scenario include altered rainfall intensities, and does this possibly increase the contribution of infiltration-excess overland flow, or may increased evaporation and subsequent dryer soil conditions possibly trigger soil crusting? If so, one can try to introduce such changes into the parameters of the hydrological model, but in doing this, one will definitely introduce a new source of uncertainty.

In general, scenario uncertainty might be dealt with according to the following structure:

- Assessing the overall trend of land use and climate change. In case of climate change, this trend has to be derived from GCM results; in case of land-use change, it might be derived by estimations about the long-term development of the economy, land management, or infrastructure.
- Assessing the uncertainty of the overall trend by evaluating results obtained from different GCMs, or results by one single GCM obtained by using varying initial conditions or varying climate forcing assumptions (multiple scenario ensembles as proposed by, for example, Mitchell & Hulme, 1999), or evaluating the long-term predictions for economic development and land management, respectively.
- Applying a state-of-the-art downscaling method, to express results from GCMs or land-use trends, respectively, on scales more appropriate for rainfall-runoff modeling. Some methods for climate downscaling and for the derivation of spatially distributed land-use scenarios have been mentioned earlier in Section "Requirements for modeling environmental change impacts on rainfall-runoff events".
- Using such "regional scenarios" and the associated uncertainty as climate or land-use boundary conditions in rainfall-runoff modeling.

It should also be noted that uncertainty involved in the assessment of environmental changes is even more pronounced regarding extreme hydrological events (rare events). This is mainly a reflection of the immense natural variability of the climate system, which requires long estimation periods for the statistical derivation of long return periods (= very rare events) and of our restricted hydrological process knowledge for extreme conditions. With the current setting of climate downscaling approaches, using the information of relatively short, observed and simulated circulation fields, one is bound to return to periods in the order of decades. Therefore, Frei *et al.* (2000) suggest that climate impact studies focusing on rare events could profit from analysis of moderately intense events rather than damage-causing extreme events, where the associated uncertainty is very high.

But even for relatively short return periods one must expect that, for a given global emission scenario, different climate models coupled to different subsequent downscaling methods generate a whole spread of local responses. For a thorough uncertainty assessment it is suggested that varying overall trends (of climate or land use) need to be taken into account and that the regional patterns obtained by different downscaling methods should be investigated, which should finally provide a rough guess of the scenario uncertainty.

Additional difficulty arises in modeling land-use change scenarios because, in that case, the parameter set calibrated for observed land-use conditions might be invalid for altered conditions. So, there is the question of assessing if and how the parameter sets will change. The simplest way is to do so entirely deterministically and change the relevant parameters to new estimates. As soon as the parameters are allowed to become uncertain however, this is a lot more difficult because of all the (nonlinear) interactions within the model structure. **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3** presents uncertainty estimation procedures, such as the GLUE, which is a quite generally accepted method to include the uncertainty associated with model- and scale-specific parameter sets. A way to cope with the uncertainty of the model structure (e.g. of the processes and their interactions represented by the model) is to check if the model structure for a specific land use still holds for a different catchment with different land-use conditions, applying an altered parameter set. If the model does not give plausible results in such a test, one has to check which process approaches have to be adjusted or additionally included to obtain a satisfactory result.

When we consider the many possible sources of uncertainty (scenario uncertainty and model, parameter, and data uncertainty), we might conclude that rainfall-runoff modeling is of little value for environmental impact assessment. However, there are ways to achieve a compromise, such as:

- *Comparison of uncertainty* introduced by data, model, and parameter problems with uncertainty due to the unreliability of scenarios.
- Use of *average and statistical changes* obtained by the GCMs and subsequent downscaling techniques (see Bronstert *et al.*, 1999 for a study example), rather than *absolute values* of climate.
- Testing of model structure by applying the same model structure with an adjusted parameter set to a second catchment with different land-use characteristics and/or different meteorological characteristics and assessing the validity/uncertainty of the model structure (representation of processes and feedbacks).

- Application of rainfall-runoff models *not in a forecast sense or as a predictive* tool but rather as a tool to evaluate the *changes in process dynamics* because of an altered land use or climate.
- Comparing *uncertainty bounds* or confidence intervals obtained by rainfall-runoff models with and without altered climate or land-use conditions, such as shown in the work by Cameron *et al.*, 2000 (see as an example Figure 6 or in a recent paper by Thieken and Merz 2003), rather than comparing single hydrographs only.

INTERACTIONS OF CLIMATE CHANGE WITH LAND-COVER CONDITIONS

Though climate and land-use change are summarized under the notion "environmental change", in most model-based environmental change assessments, climate and/or land-use change impacts are modeled independently, which is an implicit assumption that climate and land-use conditions do not influence each other. However, a changing climate results also in an alteration of the land cover, and, moreover, a changed land cover may influence the local or in part the regional climate. Thus, for assessments with a prospect of long timescales, say several decades, such interactions might become rather important and coupled climate and land-use change scenarios can be of interest.

What are the interactions of climate change and land cover and how can they be parameterized for the purpose of rainfall-runoff modeling?

- The effect of a raised concentration of CO_2 is of particular interest. The results of investigations on the effects of increased CO_2 (mostly from growth-chamber studies or from simulations with vegetation models) are rather complex. Typically, the biomass production increases. Furthermore, the stomatal conductance is reduced as a result of raised CO_2 concentration. One consequence is a greater water use efficiency at leaf level. But this effect is, at least to some extent, counterbalanced at the stand level by the greater biomass and leaf area (Kimball *et al.*, 1993; Norby *et al.*, 1996). So the most interesting hydrological effect seems to be the increased biomass production and leaf area, which leads to a greater interception storage capacity, and the decreased stomata conductance resulting in less transpiration per leaf area. Parameters to account for both leaf area and stomata conductance can be found in most rainfall-runoff models, which include evapotranspiration processes. The adequate quantification of the parameters, however, remains a major research challenge in the coupling of climate and hydrological modeling.
- As well as the greater biomass due to raised levels of $\mathbf{CO_2}$, global warming is expected to lead to a longer growing season in high and midlatitudes, an effect which has been observed during recent decades (e.g. Schaber, 2002). In an analysis over several decades, Sharrat (1992) has observed such a trend over the last 65 years in several regions throughout Alaska. The length of the growing season of course is important for various rainfall-runoff processes and rainfall-runoff modeling. It affects mainly evapotranspiration processes, in which the timing of vegetation growth can be reflected by the seasonal variation of the leaf area index.
- Atmospheric warming directly influences the timing and duration of snowfall, frost, and snowmelt. As long as the change of the temperature conditions are given in the climate scenario, such effects can be accounted for rather easily by most rainfall-runoff models, resulting in, for example, an altered timing and duration of spring floods.
- Furthermore, there is a considerable long-term feedback of the climate on the succession and composition of vegetation and its hydrologic characteristics. However, such effects are at present very difficult to quantify or to transfer into parameters of rainfall-runoff models, particularly on the catchment scale. The long-term effects of vegetation change on the soil-surface conditions (e.g. soil crusting, soil stability) are also still very poorly understood. This points out the need for a stronger integration of hydrological research with climate and soil science.
- Finally, it is worth mentioning that large-scale (e.g. regional or even continental) alteration of land cover (such as deforestation) has an effect on the local or regional climate, due to the changes of the energy exchange between atmosphere and land surface. This may also influence the air humidity, wind profiles, and conditions for rainfall generation over that area. Those mechanisms, however, have not yet been investigated in such depth that this knowledge could be used in coupled regional climate-land-use-change scenarios.

CASE STUDIES

Modeling Climate Change Effects on the Runoff Conditions in Different European Catchments

The project European River Flood Occurrence and Total Risk Assessments System (EUROTAS) was aimed at assessing the risk of current and future flooding for Europe under environmental change, including the effects of climate change. With respect to the latter, climate scenarios were derived for five large European catchments with rather different typical climate conditions, from semi-arid/Mediterranean to maritime: the Pinios, the Jizera, the Mulde, the Saar, and the Thames (see Bürger, 2002). The catchments cover a range of European climates: the Pinios

(area 9450 km^2) runs through a mountainous area in central–northern Greece, with heavy precipitation in winter and persistent droughts in summer. The Jizera (2193 km^2) is an upper Labe/Elbe tributary running through a middle mountain area in the northern Czech Republic, having a climate that is dominated by more continental characteristics. Similar is the situation of the Mulde (6170 km^2) which drains into the middle Elbe first through a mountainous and then through a flat area of eastern Germany. The Saar (7363 km^2) in France and Germany is part of the Rhine basin draining into the river Mosel. Its rainfall regime is dominated by westerly cyclones from the Atlantic. The Thames in south England (about 10 000 km^2 upstream of Kingston), finally, is characterized by maritime conditions without strong seasonality.

Subsequently, these climate scenarios were used as the climate forcing for hydrological models of these catchments to assess the possible impacts of climate change on their runoff regime, with a particular emphasis on the extreme events. In the following, first the climate change information conveyed by the scenarios and an assessment of its statistical significance is presented. Second, as an example of a climate change impact assessment, the runoff response of the Mulde catchment is summarized.

Regional Climate Scenarios

The climate scenarios are derived by a statistical-empirical method developed by Bürger (1996), the so-called *Expanded Downscaling*. EDS has merged both regressional and stochastic approaches. It modifies the classical regression principles (i.e. local climate variable vs. large-scale atmospheric field) by relaxing the condition of absolute error minimization to allow for the preservation of local variability in time. As a consequence, when driven by daily atmospheric fields, EDS simulates, by definition, daily local climate variables (e.g. temperature or precipitation) that have almost complete realistic variability and at the same time are as close as possible to the large-scale atmospheric circulation process. For this work, an improved version of the EDS technique was used which included atmospheric humidity in the set of large-scale variable fields, thus reducing the uncertainty of the statistical relationship between the large-scale fields and local conditions. One advantage of the EDS method is that it offers the possibility to generate local climate time series depending on the type of large-scale atmospheric fields. For example, different large-scale atmospheric fields can be used, such as observed atmospheric fields (supplied by NCEP or ECMWF), or results from a GCM (supplied by, e.g. the UK Hadley Centre or the Max Planck Institute for Meteorology, Germany, MPI). GCM results can be based on a projected increase of greenhouse gases (yielding the large-scale climate change scenario) or on a stationary greenhouse gas concentration (yielding the control simulation). Thus it is possible to compare the results of this downscaling technique for past observed atmospheric, control, and climate change conditions, respectively.

Three different large-scale data sets formed the basis of this study: (i) The MPI Hamburg conducted a climate simulation with the current version of their GCM, ECHAM4/OPYC3, (see Roeckner *et al.*, 1996) for the period 1860–2100, where the greenhouse gas concentrations were taken from observations for the period 1860–1990 and from the IS95a scenario (usually termed *business as usual*) for the remaining period. (ii) With the same GCM, an additional 300-year control simulation without changes of greenhouse gas concentration was conducted. (iii) Finally, a data set of "observed" large-scale atmospheric fields, covering the same selected North Atlantic/European section of the globe was available (reanalysis data of the US National Center of Environmental Prediction, NCEP).

The EDS-downscaling method then was applied for all three large-scale data sets, resulting in three different regional climate scenarios:

1. Regional scenario based on the GCM result given a stationary greenhouse gas concentration (acronym *CTL*).
2. Regional scenario based on GCM results given the above-mentioned IS95a emission trend (acronym *SCA*),
3. Regional scenario based on the reanalysis data of the NCEP (acronym *ANA*).

Table 3 summarizes the results concerning precipitation for four out of the five catchments, showing average values: mean daily precipitation in mm (m), the probability of a daily rainfall larger then 0.1 mm (p), and the mean daily intensity in mm day^{-1} (i). The results are given separately for winter (October–March) and summer (April–September). Here the OBS denotes observed precipitation values. The scenario (SCA) results are based on statistical analysis for the period 2061 to 2090.

One can see that for scenario conditions the intensity (i) increases in all catchments, for both seasons. This can have positive effects on the mean rainfall (Saar), or negative effects on the rainfall probability (Jizera), or both (Thames). The mean rainfall increases for all catchments, besides the Pinios. However, this increase is significant only for the maritime-influenced climates of the Saar and the Thames. It is most interesting that the figures shown for the Pinios indicate a dramatic decrease in rainfall probability, both in winter and summer, meaning less frequent but stronger rainfall events, which would be a dramatic and unfavorable change of the local climate conditions. This is consistent with the tendency of the ECHAM GCM to project a general drying for the eastern Mediterranean area. The design and the analysis of the significance test (mainly the combination

Table 3 Statistics of the precipitation downscaling for four European catchments in winter and summer. Arrows indicate significant departure of the climate change scenarios SCA from the joint statistics of OBS + ANA + CTL, after Kolmogorov–Smirnov (↓: decrease, ↑: increase), from Bürger, 2002

		Winter				Summer			
		OBS	ANA	CTL	SCA	OBS	ANA	CTL	SCA
Pinios	m	2.1	2.1	2.4	2.1	1.0	1.1	1.3	1.2
	p	33.4	32.5	32.7	↓25.8	18.4	17.0	18.3	↓13.9
	i	6.4	6.5	7.4	↑8.3	5.4	6.3	6.9	↑8.6
Jizera	m	2.6	2.6	2.9	3.8	2.9	2.9	3.3	3.7
	p	61.6	61.2	61.7	↓57.4	53.7	55.9	54.2	↓47.0
	i	4.2	4.2	4.7	↑6.7	5.3	5.2	6.2	↑7.9
Saar	m	2.3	2.4	2.8	↑3.5	2.2	2.3	2.5	↑2.7
	p	50.4	47.0	49.0	50.5	43.2	47.1	46.2	43.4
	i	4.6	5.2	5.7	↑6.8	5.0	4.8	5.4	6.2
Thames	m	2.3	2.2	2.6	↑3.5	1.9	1.8	2.2	↑2.4
	p	56.2	56.0	54.9	↓50.7	43.6	44.7	44.1	↓38.3
	i	4.0	4.0	4.7	↑6.9	4.4	4.1	4.9	↑6.2

of ANA, OBS, and CTL) are rather simple and of course the subject of debate. However, it can serve as a rule of thumb for the climate change (given the scenario conditions) impacts on precipitation regimes in different European catchments. More information and in-depth explanations of the results are given by Bürger (2002).

In a subsequent step, the analysis of precipitation time series was conducted looking at the extreme values of daily precipitation. On the basis of the obtained precipitation time series for observed data, downscaled respectively from reanalysis fields from GCM results with constant greenhouse gas concentration, and from GCM results with increasing greenhouse gas concentration (OBS, ANA, CTL, SCA), a classical extreme values analysis was performed, yielding precipitation frequency curves. Return periods larger than 30 years were generally not included into the statistical analysis, because the observed data (OBS) covered a period of only that duration. To utilize the information of the 300 years control run, this series was broken up into ten 30-year subperiods, and their cdfs were calculated separately. Accordingly Figure 7 illustrates a comparison of the probability of annual maximum precipitation values (expressed in return intervals) derived from OBC, ANA, 10 CTL and SCA precipitation series, where Figure 7(a) shows the situation for the winter months (October–March), Figure 7(b) for the summer months (April–September).

One can see from Figure 7 that a great uncertainty stems from the widespread of CTL runs, which is both due to natural variation and to GCM-induced uncertainty. Even with a relatively small return interval of 30 years, the variation within the CTL-cdfs is remarkable. Nevertheless, the precipitation statistics of the scenario (SCA) are significantly different, in particular, with respect to the whole range of return periods, for example, they fall outside the variability of the current climate. They show an intensification of about 10 to 20% for most catchments (in particular, for the Jizera and the Thames) and for both seasons, which indicates an increased flood risk and is in accordance with the results given in Table 3.

Rainfall-runoff Modeling for Climate Change Impact Assessment in the Mulde Basin

Within the framework of large-scale hydrological modeling of the river Elbe and its tributaries (approximately 150 000 km^2 both in Germany and the Czech Republic) the river Mulde has been selected for a detailed analysis of climate change impact on flooding (Menzel *et al.*, 2002; Menzel and Bürger, 2002). The hydrological model applied is the distributed version of the conceptual rainfall-runoff model HBV (Bergström, 1995), called HBV-D (Krysanova *et al.*, 1999). HBV-D consists of three main components: (i) snow accumulation and snowmelt, (ii) the simulation of soil moisture and runoff generation, and (iii) a catchment response and river runoff routine. Continuous simulations of river discharge by use of HBV-D were carried out for the Mulde on a daily time step over the eight successive years 1981–1988. HBV-D was repeatedly applied to simulate discharges with different climate input as follows: first, observed daily data from 75 precipitation and 4 climate stations (OBS), second, regional climatology obtained by applying the EDS method (see the preceding text) on the basis of reanalysis atmospheric fields (ANA), and third regional climatology obtained by applying the EDS method (see the preceding text) on the basis of GCM results with emission scenario IS95a (SCA). This procedure is very similar to the one conducted for the other four catchments. However, the GCM results for a constant greenhouse gas concentration (CTL) were not used here.

Using these three different climate forcings for a historical period enables a direct comparison of simulated runoff and associated uncertainties due to rainfall and temperature for measured data, and for results derived from reanalysis data and from GCM results. Thus, it is possible to test

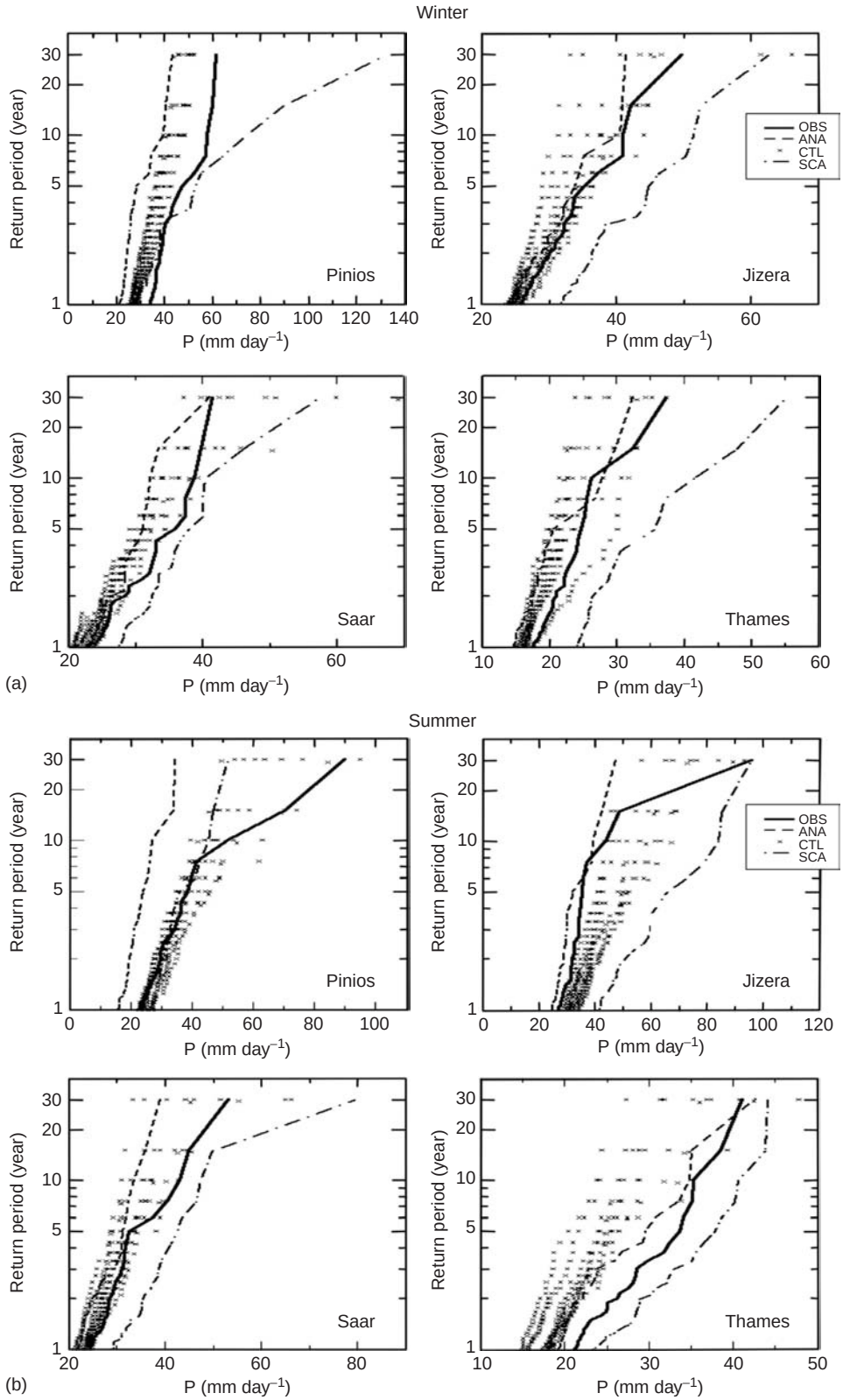


Figure 7 Return periods and associated annual maximum precipitation values for four European catchments (derived from OBS, ANA, CTL, and SCA time series, see text): (a) winter months (October–March); (b) summer months (April–September), from Bürger, 2002. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the performance of the selected climate forcing for regional rainfall-runoff modeling. Finally, the SCA data were used for a climate change impact assessment for the period up to 2100.

Figure 8 shows a comparison between observed runoff, simulated runoff with measured climatology, and simulated runoff with downscaled climatology (based on reanalysis data) for the period 11/1984 up to 10/1988 at the catchment outlet (Bad Düben gauge). One can see that the applied rainfall–runoff model using observed meteorological conditions is able to capture the runoff dynamics of the Mulde catchment quite well (grey-straight line in Figure 8, yielding a Nash–Sutcliff coefficient 0.8). Use of the climate forcing obtained by EDS based on reanalysis data (ANA) yields somewhat less satisfactory results (black-dashed line), but still gives a surprisingly good performance of runoff dynamics. This shows that a large part of precipitation variability can be attributed to the large-scale circulation. However, some flooding events are not covered by the ANA simulation, in particular, during summer periods (e.g. July 1983), implying that convective rainstorms cannot be derived from large-scale atmospheric data. Finally, Figure 8 shows the rainfall-runoff performance using the climate forcing obtained by EDS based on GCM results (SCA) for the same period (grey-dashed line). It can be seen that the simulated overall runoff dynamics are still similar to the measured and to the simulated ANA results, but, of course, the specific runoff events are not matched, because the SCA conditions cannot reflect the timing of single weather events. Nevertheless, the simulated SCA hydrograph seems to show a lower tendency for high runoff (both in frequency and in peak volume), especially during winter and springtime. This is probably due to the lack of capability to recognize areally restricted storms from large-scale data combined with the simulated underestimation of snow accumulation and related shorter melting periods. Summarizing these findings, one can see the influence of large-scale atmospheric fields on catchment climate and hydrology and the increasing uncertainty if local features of the climate forcing are neglected (ANA) or if no observations at all – neither local measurements nor large-scale fields – are used to derive the climate forcing (SCA, period 1985–1988).

The last step of this case study was the application of the same rainfall-runoff model with projected future climate forcing (SCA, period up to 2100). Menzel and Bürger (2002) reported that this simulation reveals a high uncertainty about the future development of flood risk

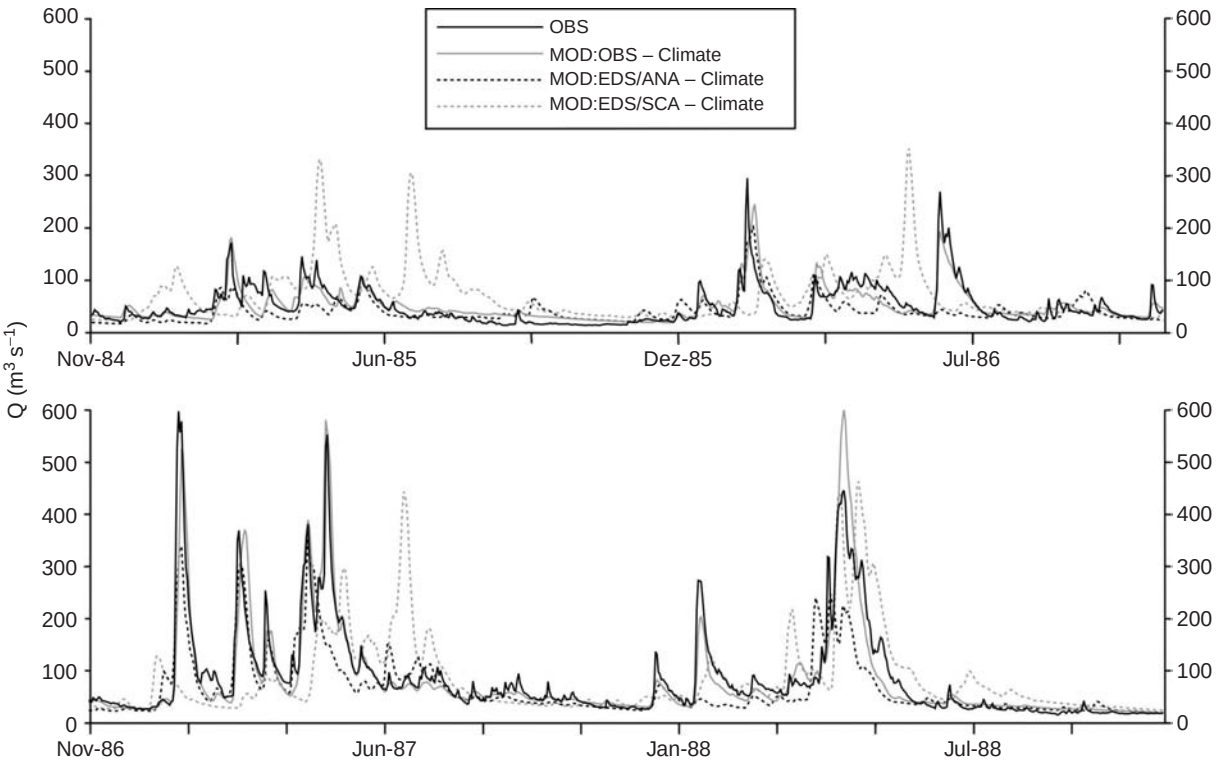


Figure 8 Observed runoff, simulated runoff with measured climatology, and simulated runoff with downscaled climatology (based on reanalysis data and based on GCM results) for the river Mulde at gauge Bad Düben, 11/1984–10/1988, from Menzel *et al.* (2002). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

for catchments. The simulations predict a weak tendency towards reduced peak discharges. This seems to be valid for a major part of the whole Elbe catchment. On the basis of the applied climate change scenario, the simulations indicate that a future problem might arise from reduced water availability in the Elbe catchment. However, it has to be emphasized that such analyses are associated with high uncertainty, in particular, with respect to the probability of very rare events (see also the discussion by Lamb (1999) about difficulties in modeling flood frequency distributions and the detailed elaboration of this issue in Chapter 11.9 of this encyclopedia).

A Comprehensive Modeling Study of Land-use Change Effects on the Runoff Conditions in the Rhine Basin

From 1997 until 2001, an EU-funded project was conducted aiming to quantify flood risk for the Rhine basin under altered environmental conditions. The main focus of the project was on two important questions:

1. To what degree do changes of land cover and river training influence the flood situation in the Rhine basin? and
2. To what degree can flooding be mitigated by water retention measures both in the landscape and along the river courses?

Therefore, the project on the one hand aimed at quantifying the impact of landscape hydrology and river network hydraulics on flooding conditions in the Rhine basin. In this context, the retention capacity of the landscape with its geophysical properties and the river training conditions were of importance. On the other hand, the mitigation potential of specific measures like infiltration ponds, altered agricultural management practices, restoration measures in small tributaries and polders or flooding areas in larger rivers were investigated. The aim was to provide results which could serve as sound scientific figures about a possible increase in flood risk caused by land-use changes or a possible decrease in flood risk induced by specific measures like decentralized water retention or the flooding of polders.

The project required an interdisciplinary and multiscale approach. This was achieved by combining models for different purposes at different spatial scales, allowing the comparison of the impacts of land-cover conditions and the effects of river training activities (including the retention capacity in rivers and floodplains) on the discharge conditions in the river Rhine. The models had either to be adjusted or expanded in order to fulfill the project requirements.

Special attention had to be paid to the coupling of catchment hydrology and flood wave propagation as well as to the linkage between the land-use scenarios and the structure and parameterization of hydrological modeling. The simulations were carried out by process-oriented rainfall-runoff modeling at three different spatial scales covering the Rhine basin from Maxau (southwest Germany) to Lobith (Dutch/German border), which is a total area of more than 110 000 km^2. The mentioned modifications were necessary at the following spatial scales and will be explained hereafter:

1. To include process knowledge at the lower mesoscale (three small catchments covering an area of about 100–500 km^2), which was considered to be relevant for land-use change modeling but was not represented in the original version of the chosen model. This refers in particular to the representation of macropore effects on infiltration conditions and therefore on the possible generation of infiltration-excess overland flow and to the representation of urban storm water processes, such as runoff on impervious areas, retention in sewage systems, and possible transfer to sewage treatment facilities.
2. To transfer process knowledge ("upscaling") simulated at the lower mesoscale to the upper mesoscale (101 larger subcatchments of the Rhine, ranging between 400 and 2100 km^2, covering a total area of 110 000 km^2) where, as a consequence of data scarcity and data management constraints, a rather conceptual modeling approach had to be chosen.
3. To route the runoff from all the different subcatchments of the Rhine catchment in the main river system. This includes possible retention effects within the river corridor and/or in flood polders along the channel system of the Rhine and its main tributaries. This is primarily a matter of hydrodynamic modeling.

The following paragraphs present the models applied and the results achieved at the different scales. The modeling at the lower mesoscale is described more in detail, because at that scale the modeling of land-cover/land-use change effects on runoff processes can be demonstrated best. The other two levels are summarized, followed by an overall discussion of this case study.

Modeling land-cover impact on runoff processes during periods of heavy precipitation at the lower mesoscale

The quantification of the influence of land cover on the runoff process during periods of heavy precipitation at the lower mesoscale was done by applying the process-oriented hydrological model *WASIM-ETH* (Schulla, 1997) on catchments of an area ranging between 100 km^2 to 500 km^2. These were selected because they represent different characteristic land-use patterns with either dominantly urban, agricultural, or forested structure. For these three exemplary catchments, spatially distributed scenarios of future

land use and land cover were generated on the basis of regional analyses of land-use trends.

WASIM-ETH was considered to be an adequate tool for this purpose and this scale, because it includes most of the processes relevant for runoff generation. It considers the spatial distribution of catchment characteristics, and is based on fine-gridded spatial and temporal dynamics of climate variables, topography, soil properties, land use, and vegetation. As mentioned earlier, infiltration and surface runoff-generation processes are crucial processes in modeling land-cover or land-use effects. Particularly under extreme precipitation conditions these processes contribute strongly to the water being transferred to the river system, either directly by generating infiltration-excess overland flow or indirectly by influencing the extent of saturated surface areas, thereby causing saturation-excess overland flow. Therefore, the developed extensions of WASIM-ETH focus on soil and land-cover characteristics, which either increase or reduce the infiltration capacity. Increasing effects are mainly due to the occurrence of macropores (Bronstert, 1999), and reducing effects due to possible siltation at the soil surface. In addition, to account for urban hydrological effects, a model extension was developed, which accounts for the connection of sealed surfaces to the sewer system and the possible diffuse water retention in urban and agricultural areas.

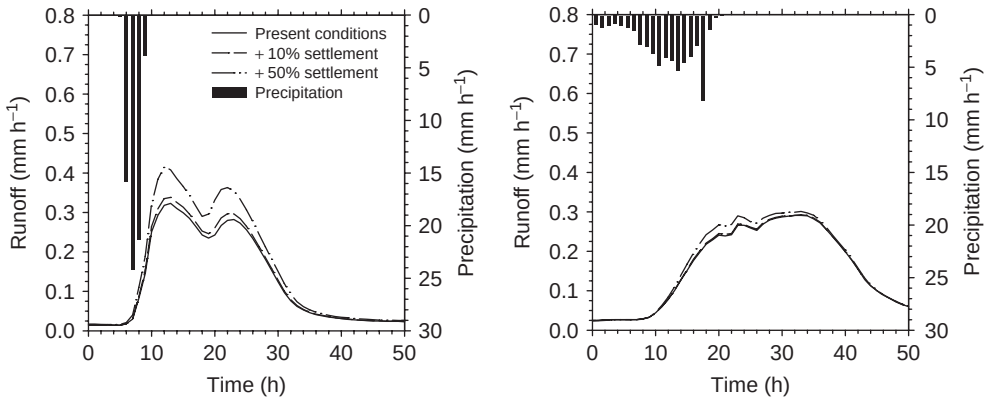
The final model version requires the following parameters to represent the relevant (land-use affected) processes of flood generation:

- soil parameters: soil depth; hydraulic conductivity; saturation content; macroporosity; mean macropore depth; possible reduction of surface conductivity due to siltation
- urban parameters: settlement area; fraction of sealed surface within a settlement; capacity of urban retention reservoirs; controlled discharge from urban retention reservoirs; local catchment area and retention capacity of diffuse storm water control management devices in urban or agricultural areas; maximum seepage rate from these storm water management devices
- land-cover parameters: maximum interception storage; soil cover index; root depth, leaf area index

Niehoff *et al.* (2002) gives an overview of the model and its extensions and Niehoff (2002) provides comprehensive documentation and several application and sensitivity examples for the different extensions.

As an example for the modeling results at the lower mesoscale, the simulated response to a relative increase in urban area of 10% and 50% respectively is described in the following. In the Lein catchment, such an increase corresponds to an absolute growth of these land-use types from 7.4% of the catchment area to 8.1% and 11.1% respectively.

Figure 9 is a comparison of two typical flood events in the Lein catchment, one induced by advective circulation conditions (long lasting but less intense rainfall) and one by convective conditions (i.e. a summer thunderstorm: short rainfall of very high intensity). The figure presents simulation results for present conditions as well as for the two urbanization scenarios. The comparison demonstrates that the increase in flood volume and peak runoff due to urbanization is much more distinct for the *convective* storm event than for the *advective* one, although the precipitation volume as well as the peak flow is in the same order of magnitude for both events and represents a return period of approximately 2 to 3 years in both cases. The markedly less pronounced effect on the advective event is the result of (i) *higher antecedent soil moisture* which levels differences in soil characteristics as well as (ii) *lower precipitation intensities* which prevent an overflow of the sewer system.



(a) June/28/1994, Frankenbach/Lein (b) February/16/1990, Frankenbach/Lein

Figure 9 Simulation of two flood events in the Lein catchment (115 km^2) as a response to (a) a convective storm event (28. 6. 1994) and (b) an advective storm event (15.2.1990) for present conditions and two urbanization scenarios (Reproduced from Niehoff, 2002 by permission of D. Niehoff)

This argument is also supported by a comparison of various advective events with different return periods, as shown in Table 4. The comparison reveals a strong correlation between the impact of urbanization on runoff and the *baseflow contribution* to the flood event, which serves as indicator for high groundwater levels and high soil moisture.

Another comprehensive source of information are the runoff components that are simulated for the different flood events (Figure 10). The two pie charts reveal pronounced differences in the *dominating runoff-generation mechanisms* depending on the event characteristics (rainfall intensities and prestorm moisture conditions). The response to convective storm events is dominated by *sewer overflow* from sealed surfaces as well as a considerable amount of *infiltration excess* mainly from agricultural areas. In contrast to that, for advective events *subsurface flow* processes and *saturation-excess* prevail.

Niehoff and Bronstert (2001) have presented other examples of modeling land-use change impacts at the lower mesoscale, such as urban storm water infiltration, agricultural storm water management, or surface crusting of arable land. The results were similar in the sense that the impacts are generally much more pronounced in case of convective rainstorms than in advective rainstorms.

Upscaling from the lower to the upper mesoscale

In order to simulate land-use change effects at larger scales, the detailed information about the runoff generating processes obtained by applying WASIM-ETH at the lower mesoscale (see the preceding text) has to be transferred to larger subcatchments of the Rhine tributaries (up to 2100 km^2 each) covering the whole Rhine basin between Maxau and Lobith (110 000 km^2). This was approached by applying a rather simple, conceptual hydrological model (extended HBV-model, see the following text) with a generalized parameter set, tailored to represent land-use features. This generalization is achieved by establishing a statistical relationship between the simulated land-use change impacts obtained from the WASIM-ETH model and the HBV parameters.

The HBV-model (Bergström, 1995), is a semidistributed conceptual model. It uses subcatchments as primary hydrological units, and a simple Muskingum flood routing is used

Table 4 Increase in runoff volume and peak due to a 50% increase of settlement and industrial areas in the Lein catchment; the events are sorted by the urbanization impact on runoff volume

Year, month	Increase in runoff compared to present conditions		Simulated baseflow contribution to volume (%)	Duration (h)	Return period approximate (a)
	Peak (%)	Volume (%)			
1990, February	3.4	3.7	19	150	2
1993, December	5.9	2.7	17	250	8
1997, February	3.9	2.7	19	150	7
1982, December	1.7	1.5	27	225	3
1983, May	0.6	0.9	39	300	4
1988, March	0.0	0.0	52	650	3
Mean	2.6	1.8	29	290	4.5

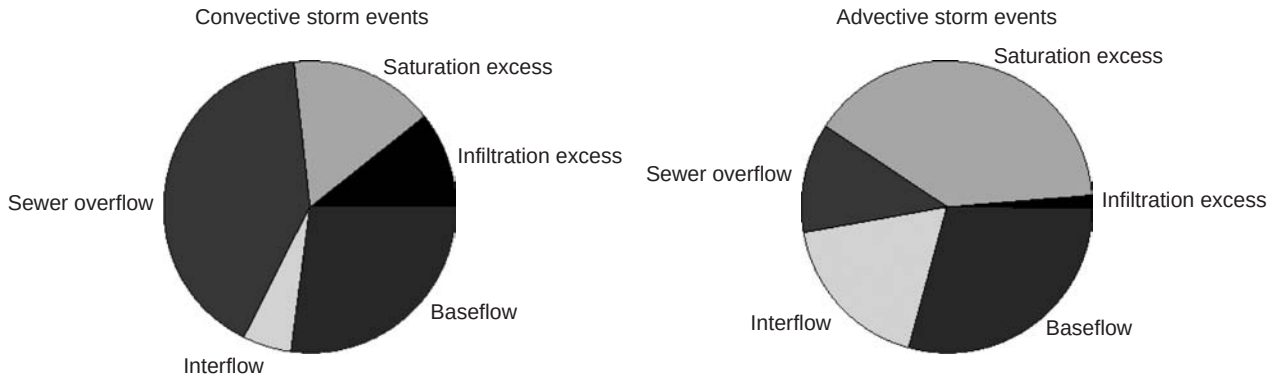


Figure 10 Runoff components simulated with WaSIM-ETH for the Lein catchment for five convective and six advective storm events with return periods between two and eight years (Reproduced from Niehoff and Bronstert, 2001 by permission of Springer)

between the subcatchments to model the runoff from the whole catchment. The subcatchments can further be subdivided into homogeneous zones on the basis of elevation, soil type, vegetation, and so on. The runoff is computed for each zone within the subcatchments using simplified conceptual routines of snow accumulation and snowmelt, soil moisture accounting, and evapotranspiration. The generated runoff is then routed to the outlet of the subcatchments using a lumped runoff concentration model schematized by a series of two reservoirs; a nonlinear upper reservoir that is fed by the runoff generated from the zones and a lower linear reservoir, which is fed by a constant percolation rate from the upper reservoir (Lindström, *et al.*, 1997). The outflow from the upper reservoir simulates the relatively fast runoff from the soil zone near the surface, while the outflow from the lower reservoir simulates the rather slower base flow component. The total outflow from both reservoirs is finally smoothed by a triangular transformation function.

Some extension of the model structure was made to account for the effects of urbanization and possible Hortonian runoff, which is not part of the original model version. A component for impervious area was introduced and runoff generated from such areas is directly routed to the outlet of the subcatchments using the transformation function without entering any one of the reservoirs of the runoff concentration module. Furthermore, at high precipitation intensity, infiltration into (unsealed) soil may be limited by the infiltration capacity of the soil, which might be reduced severely by siltation of the soil surface due to rainfall energy. In order to model this phenomenon, two more parameters were introduced: a threshold precipitation intensity and the fraction of Hortonian runoff contributing area during heavy precipitation intensity. When the intensity of precipitation exceeds the threshold value, part of the precipitation in excess of the threshold produces overland flow. Both parameters are estimated through model calibration. The parameters *impervious area* and *fraction of Hortonian runoff contributing area* are both directly related to the land-use practice. The rainfall intensity is a meteorological parameter, determining the occurrence of Hortonian runoff.

As the HBV-model is a conceptual model, most of its parameters do not have a physical meaning and therefore they have to be estimated through model calibration against observed catchment response. The effect of land-use changes on these parameters can however be estimated only if a relationship between the model parameters and the land-use characteristics of the catchment can be established, which is the main scientific challenge when using a conceptual hydrological model for large-scale (e.g. upper mesoscale) regionalization of process knowledge. In this case, parameters pertaining processes of runoff generation (in each elevation zone of each subcatchment) were estimated as functions of the land-use type or soil type within the zone in question. Parameters of the runoff concentration module were estimated for each subcatchment using a transfer function of the catchment characteristics (see Hundecha and Bárdossy, 2004 for details).

The usual approach of estimating the parameters of the transfer function follows a two-step procedure. The model is first calibrated for many subcatchments independently to obtain "best" model parameter sets, and then the parameters of the transfer function are obtained using optimization. However, this kind of model calibration results in only a single realization among many parameter sets that lead to similar model performance. Therefore, the relationship established between the model parameters and the catchment descriptors is likely to be weak. A different approach was implemented in this study. The model parameters were first expressed in terms of the parameters of the transfer function and thereafter the model was calibrated for many subcatchments with contrasting catchment properties simultaneously. Thus, the model calibration directly yields the parameters of the transfer function, which remain the same for all subcatchments. The lower mesoscale subcatchments modeled by the detailed process-oriented modeling approach (see the preceding text) were also incorporated in the simultaneous model calibration. The calibration for these subcatchments was made against the simulated runoff obtained by the detailed process-oriented model (WASIM-ETH). In order to reproduce the effect of land-use changes obtained by WASIM-ETH, the results obtained for the different land-use scenarios (at the lower mesoscale) were used together with the corresponding land-use distribution. Hundecha and Bárdossy (2004) present details of this regionalization approach and demonstrate simulation results referring land-use change effects on runoff at the upper mesoscale. Some examples of those results are included in the following paragraphs.

Flood routing of the river Rhine and its main tributaries

The macroscale conflation of the tributary subcatchment simulation results was conducted by a flood routing model of the river Rhine and its main tributaries, applying the Muskingum routing model SYNHP (Homagk, 1995) to the main tributaries and the hydrodynamic hydraulic model SOBEK (WL Delft Hydraulics and the Ministry of Transport, Public Works and Water Management, 1997) to the main channel. This allows the river discharge dynamics to be simulated, including impacts of alterations of river channels and retention effects from detention basins and floodplains.

The combined application of these river routing models enables the explicit simulation of changes of cross-section

geometry, river roughness, storage options in floodplains and polders and considers further influencing factors like groundwater exchange. Within the scope of this project, SOBEK was applied to the simulation of the Rhine stretch between Maxau and the Dutch border and the lower parts of the rivers Neckar, Main, and Lahn. SYNHP was applied for the calculation of different retention scenarios in the remaining and upper stretches of the tributaries and furthermore for the upper courses of the relevant Rhine tributaries.

Summary of the modeling results at the large scale

This study is an example of a multiscale process-oriented coupling of different models in order to assess the impacts of land use and river training measures on the runoff of a large catchment. For the first time, it was possible to give quantitative estimates for the impacts of land-use change and river training measures on the flood conditions for the river Rhine. Examples for the process-oriented modeling of land-use change impacts at the lower mesoscale have been demonstrated above; the study as a whole is documented in the Commission for the Hydrology of the River Rhine (CHR) report (Bronstert *et al.*, 2003).

Three different land-use scenarios in the catchment, focussing on urbanization and/or urban storm water treatment, have been taken into consideration:

1. Scenario $D1$ is based on a rather realistic scenario of about a further 10% expansion of urban areas in the Rhine catchment as projected by Dosch and Beckmann (1999),
2. Scenario $D2$ includes the increase of urban area of scenario D1 and, additionally, a planned project for controlled infiltration of urban storm runoff in 2500 km^2 urban areas, as recommended in the flood action plan of the International Rhine Commission (IKSR, 1998), and
3. Scenario $D3$ representing an "extreme scenario" of a 50% increase of urban areas.

All scenarios D1, D2, and D3 also consider the effects of the planned or already constructed flood defense works along the Rhine between Maxau and Lobith. These modeled river training measures include flood polders along the Upper Rhine below Maxau (total volume 79.2×10^6 m^3) and along the Lower Rhine (total volume 65.4×10^6 m^3). However, the planned flood polders along the Upper Rhine upstream of Maxau (207.6×10^6 m^3) have not been assessed in this study.

These three scenarios of land use and/or river training (river retention) measures were simulated driven by *three different scenarios of meteorological forcing*, one observed extreme and large-scale precipitation situation and two designed "extreme meteorological scenarios" in order to test the model system for even more severe meteorological conditions.

1. Scenario $M95$: Meteorological forcing (in its observed spatial and temporal distribution) of January/February 1995, which caused a flood in the Rhine with a return period >100 years,
2. Scenario $M95+$: Meteorological forcing of January/February 1995 *plus* a linear increase of precipitation of 20%,
3. Scenario $M95++$: Meteorological forcing of January/February 1995 *plus* a linear increase of precipitation of 20% *plus* an additional pre-event snow water equivalent of 20 mm over the whole catchment.

In Table 5, a summary of the model results is presented by listing the combined effects of land-use scenarios and meteorological scenarios. The modeled differences in water levels (in cm) at five main gauging stations of the Rhine are given. The values without parentheses are due to the combined effects of land-use change (increase of urban areas) and river training (increase of flood-discharge retention in river polders). The values in parenthesis are due to land-use change only (first value) and due to river training only (second value). Positive numbers imply a decrease in water level and negative numbers imply an increase.

From the differences in water levels listed in Table 5, one can draw several conclusions:

- The increase of flood peak level due to a further moderate (realistic) increase of urbanized areas (D1) is very small (water level increase 2 cm of less) and therefore almost negligible.
- The influence of the proposed management of urban storm water results in a very limited mitigation of flood peaks (water level decrease 2 cm or less) and therefore is almost negligible, too.
- The effects of water retention in flood polders (between Maxau and Lobith) have a stronger but still small effect (water level decrease of 3 cm of less for the M95 scenario, up to 10 cm for the M95+, and up to 17 cm for M95++). It is important to understand that consideration of the possible retention polders upstream of Maxau would yield an additional reduction in the range of 10 cm, in particular, for the Upper Rhine area.
- The unrealistic, extreme land-use scenario (50% increase of urban areas) would result in a water level increase not much more than 10 cm.
- The M95+ and even more the M95++ scenario results in higher reduction of flood levels in the case where the flood polders are active. This is because of the fact that according to the operation rules, the flood polders are to be filled only if flood discharge exceeds the 200-year value. In January/February 1995 (M95) the flood discharge along most stretches of the Rhine was above the 100-year value, but below

Table 5 Modeled changes in water level (cm) at five main gauging stations of the Rhine, due to scenarios of land-use change and river training (river retention) measures for three different meteorological scenarios. Explanations are given in the text (from Bronstert *et al.*, 2003)

Rhine gauging station (km downstream Lake Constance)	Meteorological scenario		
	M95	M95+	M95++
Worms (km 444)			
D1	0 (0/0)	10 (0/10)	16 (0/16)
D2	0 (1/0)	9 (0/10)	16 (−1/17)
D3	0 (−1/1)	−10 (−1/−9)	15 (−1/16)
Kaub (km 546)			
D1	1 (−1/2)	8 (−1/9)	9 (−2/11)
D2	1 (−1/2)	8 (−1/9)	9 (−1/11)
D3	−5 (−7/3)	3 (−6/8)	3 (−9/11)
Andernach (km 614)			
D1	0 (−1/1)	5 (−1/6)	6 (−1/8)
D2	1 (0/1)	6 (−1/6)	7 (−1/8)
D3	−5 (−7/2)	1 (−5/6)	2 (−6/8)
Köln (km 688)			
D1	0 (−2/1)	5 (−1/6)	4 (−2/6)
D2	1 (0/1)	5 (−1/6)	5 (−1/6)
D3	−8 (−9/2)	−1 (−7/6)	−3 (−9/7)
Lobith (km 857)			
D1	2 (−1/3)	2 (−1/3)	2 (−1/3)
D2	2 (−1/3)	3 (−1/3)	2 (−1/3)
D3	−1 (−5/3)	−2 (−6/3)	−5 (−8/3)

D1: Current land-use conditions and a 10% increase of urban area.
D2: Current land-use conditions and increase of urban area (D1) *plus* controlled infiltration of urban storm runoff.
D3: Current land-use conditions and a 50% increase of urban area (''extreme urbanization scenario'').
M95: Meteorological forcing (in its observed spatial and temporal distribution) of January/February 1995.
M95+: Meteorological forcing of January/February 1995 *plus* a linear increase of precipitation of 20%.
M95++: Meteorological forcing of January/February 1995 *plus* increase of precipitation *plus* an increase of snowwater equivalent.

200 years. That is why the polders were used only at a few stretches resulting in small water level reduction only.

Some more general results from the whole study are summarized in the following:

1. At the lower mesoscale level the influence of land use on storm-runoff generation is stronger for convective storm events with high precipitation intensities than for long advective storm events with low precipitation intensities because only storm events associated with high rainfall intensities are at least partially controlled by the conditions of the land cover and/or the soil surface. Convective storm events, however, are of minor relevance for the formation of floods in the large river basins of central Europe because the extent of convective rainstorms is usually restricted to local catchment areas rather than widespread occurrence.
2. An estimated, rather dramatic, further increase of urban areas of about 50% may result in an increase of medium-size flood peak discharge (e.g. return intervals between 2 and 8 years) in catchments of the lower mesoscale (up to ca. 1000 km^2):
   - Between 0% and 4% for advective rainfall events, and
   - Up to 30% for convective rainstorms,
3. The flood impacts due to a more realistic representation of urbanization increase are in the order of 1 cm to 5 cm in the main channel, while the effects are even less for extreme rainfall amounts.
4. The decentralized storage, detention, and infiltration of urban storm water yields reduction of flood peaks in catchments of the lower mesoscale which are about

the same size as the increase due to urbanization (see the preceding text).

5. The superposition of flood waves originating in different subbasins shows that the maximum effect of water retention in the landscape generally occurs in the rising limb of the flood wave in the main channel. The flood mitigation effect at the peak is considerably smaller.
6. Water retention measures in polders along the Upper and Lower Rhine, under the given boundary conditions, yield flood peak attenuation along the Rhine all the way down to Lobith of between 1 cm and 15 cm (see the preceding text). The optimized and coordinated control of the polders can result in a considerably stronger decrease of the peaks.

The results of this study examined the responses to the different scenarios in a purely deterministic way, as a tool to evaluate the changes in process dynamics due to the postulated scenarios of climate and land-use change. A more comprehensive study, as recommended earlier, should also attempt to evaluate the uncertainties in the predictions of change.

CONCLUSIONS

Rainfall-runoff models can be adequate tools to assess the impacts of climate and land-use changes on the hydrological cycle. This implies that the applied rainfall-runoff model is able to "represent" the dynamics of the system to be modeled, that is, the hydrological cycle of the catchment under study, in particular – if land-use change is of concern – the effects of land-use (or land cover) change on the runoff generating processes. This may be accomplished by applying a process-based model (with a rather direct alteration on the model parameter) or by applying a conceptual model if its parameters can be regionalized in connection to land-use change. Furthermore, the chosen model must be able to "represent" the influence of different climatic/meteorological boundary conditions. Like any model, the chosen rainfall-runoff model cannot give a full picture of reality, however, it is important that the model represents the main part of the system dynamics. This implies, of course, that we know (or at least have a well-educated guess) of the main dynamics of the catchment under consideration. The more we know about the catchment to be modeled and the better the model is able to "represent" this knowledge and the more reliable the model results will be.

Different modeling purposes result in different requirements for the rainfall-runoff model to be applied. For instance, before starting the assessment, one has to clearly define the appropriate time and space scale if the focus is on the average water balance or, for example, on hydrological anomalies (floods and droughts). As mentioned above, the boundary conditions are of similar importance to the rainfall-runoff model itself. This means that the whole modeling procedure must contain an adequate, scientifically sound handling of the boundary conditions. A changed climate or an altered land-use condition should not be arbitrarily chosen, instead be based on *reasonable assumptions about the trend of driving forces* influencing climate (e.g. green house gas concentrations) or land use (e.g. economic development), and the *expression of the changed climate or land-use conditions for the space and timescale which is relevant* for the study.

Climate change and land-use change assessment studies using rainfall-runoff models can be treated independently as long as no feedback effects of climate on land use or vice versa are relevant. In most assessment studies, such feedback effects are not on top of the priority list. However, for long-term (decades or centuries) assessment studies the relevance of such feedbacks should be discussed *a priori* and, if necessary, coupled boundary conditions should be defined, or coupled (integrated) models be developed (Bronstert *et al.*, 2005).

Uncertainty assessment is of particular interest for this kind of rainfall-runoff modeling, because both the uncertainty within the rainfall-runoff modeling procedure (data, process, parameter uncertainty), and the uncertainty due to the definition of the scenarios (boundary condition) are relevant. It is an ongoing challenge and an area of present research to reflect the different sources of uncertainty and to quantify it as much as possible.

The ideas presented above are sorted into a few statements yielding – a rather ambitious and often hard to follow – scheme for quality control while modeling the impacts of climate and land-use change:

1. Define the modeling purpose clearly, including the definition of the important hydrological processes, the relevant time and space scale, and the role of hydrological anomalies.
2. Check if the chosen rainfall-runoff model contains/represents the relevant processes of the system to be modeled, under the given purposes.
3. Check model performance, for example, by comparing the model results obtained on the basis of today's-boundary conditions with observed data, and by evaluating model results with several catchments with differing land-use and/or different climate conditions.
4. Define the overall trend of the land use and/or climate change and develop the land use and/or climate scenario in the time and space scale which is relevant for the specific study.
5. Check if feedback effects of climate and land use are important for the specific study, and – if so – develop integrated scenarios.
6. Assess the uncertainty of the whole system modeled, including both the uncertainty within the rainfall-runoff modeling procedure (data, process, parameter

uncertainty), and the uncertainty due to the definition of the scenarios (boundary condition).

Acknowledgments

The author would like to thank Dr. G. Bürger for the downscaling of climate and precipitation scenarios, Dr. Uta Fritsch for her derivation of the land-use scenarios in the Lein catchment and the Rhine basin, Dr. L. Menzel for the climate and land-use change assessment of the Elbe and Mulde catchments, and – last but not least – Dr. D. Niehoff for the rainfall-runoff modeling on the Lein catchment. Thanks are also due to the Commission for the Hydrology of the River Rhine for data provision of the Rhine catchment and the whole LAHOR team for the steady and successful conductance of this project. The work reported here was partly funded by the European Commission (DG XII for the EUROTAS project; DG XVI – IRMA-programme for the LAHOR project) and by the German Federal Environment Agency (UBA).

REFERENCES

Abbott M.B., Bathurst J.C., Cunge A., O'Connell P.E. and Rassmussen J. (1986) An introduction to the SHE, 2: structure of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 61–77.

Bárdossy A. (1993) *Stochastische Modelle zur Beschreibung der raum-zeitlichen Variabilität des Niederschlags*, Mitt. des Inst. für Hydrologie und Wasserwirtschaft, 44, Univ. Karlsruhe: p. 153.

Bárdossy A. (1997) Downscaling from GCMs to local climate through stochastic linkages. *Journal of Environmental Management*, **49**, 7–17.

Bárdossy A. and Caspary H.J. (1990) Detection of climate change in Europe by analyzing European atmospheric circulation patterns from 1881–1989. *Theoretical and Applied Climatology*, **42**, 155–167.

Bárdossy A. and Plate E. (1992) Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, **28**, 1247–1259.

Bergström S. (1995) The HBV model. In *Computer Models of Watershed Hydrology*, Singh, V.P. (Ed.), Water Resources Publications; Highlands Ranch, Colorado, US, pp. 443–476.

Beven K. (2001a) How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, **5**(1), 1–12.

Beven K. (2001b) *Rainfall-Runoff Modelling–The Primer*, Wiley: Chichester, p. 360.

Bronstert A. (1999) Capabilities and limitations of detailed hillslope hydrological modelling. *Hydrological Processes*, **13**, 21–48.

Bronstert A., Bárdossy A., Bismuth C., Buiteveld H., Busch N., Disse M., Engel H., Fritsch U., Hundecha Y., Lammersen R., Niehoff D. and Ritter N. (2003) *Quantifizierung des Einflusses der Landoberfläche und der Ausbaumaßnahmen am Gewässer auf die Hochwasserbedingungen im Rheingebiet*, Reports of the Commission for Hydrology of the River Rhine (CHR), Series II, No. 18, p. 78.

Bronstert A., Bürger G., Heidenreich M., Katzenmaier D. and Köhler B. (1999) Effects of climate change influencing storm runoff generation: basic considerations and a pilot study in Germany. In *The Impact of Climate Change on Flooding and Sustainable River Management*, *Proceedings of the final RIBAMOD Workshop, Wallingford, 26–27 February 1998*, Casale R., Samuels P. and Bronstert A. (Eds.), Office for Official Publications of the European Communities: Luxembourg, pp. 325–340.

Bronstert A., Carrera J., Kabat P. and Lütkemeier S. (Eds.) (2005) *Coupled Models for the Hydrological Cycle–Integrating Atmosphere, Biosphere and Pedosphere*, Springer: Heidelberg, p. 346.

Bronstert A., Niehoff D. and Bürger G. (2002) Effects of climate and land-use change on storm runoff generation: present knowledge and modelling capabilities. *Hydrological Processes*, **16**(2), 509–529.

Bürger G. (1996) Expanded downscaling for generating local weather scenarios. *Climate Research*, **7**, 111–128.

Bürger G. (2002) Selected precipitation scenarios across Europe. *Journal of Hydrology*, **262**(1–4), 99–110.

Cameron D., Beven K. and Naden P. (2000) Flood frequency estimation by continuous simulation under climate change (with uncertainty). *Hydrology and Earth System Sciences*, **4**(3), 393–405.

Conway D. and Jones P.D. (1998) The use of weather types and air flow indices for GCM downscaling. *Journal of Hydrology*, **213**(1–4), 348–361.

DeFries R. and Achard F. (2002) New estimates of tropical deforestation and terrestrial carbon fluxes: result of two complementary studies. *LUCC Newsletter* No. 8, 7–9.

Dosch F. and Beckmann G. (1999) Trends und Szenarien der Siedlungsflächenentwicklung bis 2010. In *Steuerung der Flächennutzung*, Raumordnung B.F.B.U. (Ed.), Bundesamt für Bauwesen u. Raumordnung, Vol. 8, pp. 493–510.

EEA (European Environmental Agency) (2002) EEA multilingual environmental glossary: **http://glossary.eea.eu.int/EEAGlossary/L/land_use**

Engel H. (1997) Die Ursachen der Hochwasser am Rhein – natürlich oder selbstgemacht? In *Hochwasser. Natur im Überfluss?* Immendorf R. (Ed.), Verlag C.F. Müller: Heidelberg, pp. 9–30.

FAO (1997) *State of the World's Forests*, Food and Agriculture Organization: Rome, p. 200.

Freer J., Beven K.J. and Ambroise B. (1996) Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Water Resources Research*, **32**, 2161–2173.

Frei C., Davies H.C., Gurt J. and Schär C. (2000) Climate dynamics and extreme precipitation and flood events in Central Europe. *Integrated Assessment*, **1**, 281–299.

Fricke W. and Kaminski U. (2002) Ist die Zunahme von Starkniederschlägen auf veränderte Wetterlagen zurückzuführen? *GAW Brief des Deutschen Wetterdienstes*, Nr.

12, Meteorologisches Observatorium Hohenpeißenberg: Sept. 2002.

Fritsch U. (2002) *Entwicklung von Landnutzungsszenarien für landschaftsökologische Fragestellungen*, *Brandenburgische Umweltberichte*, 12, University of Potsdam: p. 132.

Gerstengarbe F.-W., Werner P.C., Frädrich K. and Österle H. (2000) Recent climate change in the North Atlantic/European Sector. *International Journal of Climatology*, **20**(5), 463–471.

Homagk P. (1995) *Simulation des Hochwassergeschehens am Oberrhein*, Wasserbau-Mitteilungen der TH Darmstadt, Nr. 40.

Hulme M. and Jenkins G.J. (1998) *Climate Change Scenarios for the UK: Scientific Report*, UKCIP Technical report No. 1, Climate Research Unit, Norwich.

Hundecha Y. and Bárdossy A. (2004) Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology*, **292**(1–4), 281–295.

IKSR (International Commission for the Protection of the Rhine), (1998) Aktionsplan Hochwasser (Flood action plan) [in German] *International Commission for the Protection of the Rhine*, Koblenz, pp. 30.

IPCC (1996) Contribution of working group I to the second assessment of the intergovernmental panel on climate change. In *The Science of Climate Change*, Houghton J.T., Meira Filho L.G., Callender B.A., Harris N., Kattenberg A. and Maskell K. (Eds.), Cambridge University Press: p. 572.

IPCC (2001) Climate change 2001. the scientific basis. *Contribution of Working Group I to the Third Assessment of the Intergovernmental Panel on Climate Change*, Cambridge University Press: p. 881.

Janssen P.H.M. and Heuberger P.S.C. (1995) Calibration of process-oriented models. *Ecological Modelling*, **83**(1–2), 55–66.

Kimball A., Mauney J.R., Nakayama F.S. and Idso S.B. (1993) Effects of increasing atmospheric CO2 on vegetation. In *CO2 and Biosphere*, Rozema J., Lambers H., Van de Geijn S.C. and Cambridge M.L. (Eds.), Kluwer Academic Publisher: Belgium.

Krausmann F., Haberl H., Schulz N.B., Erb K.-H., Darge E., Gaube V. (2003) Land-Use Change and Socio-Economic Metabolism in Austria–Part I: Driving Forces of Land-Use Change: 1950–1995, *Land Use Policy*, **20**(1), 1–20.

Krysanova V., Bronstert A. and Muller-Wohlfeil D.I. (1999) Modelling river discharge for large drainage basins: from lumped to distributed approach. *Hydrological Sciences Journal*, **44**(2), 313–331.

Lamb R. (1999) Calibration of a conceptual rainfall-runoff model fro flood frequency estimation by continuous simulation. *Water Resources Research*, **35**(19), 3103–3114.

Lindström G., Johansson B., Persson M., Gardelin M. and Bergström S. (1997) Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, **201**, 272–288.

Menzel L. and Bürger G. (2002) Climate change scenarios and runoff response in the Mulde catchment (Southern Elbe, Germany). *Journal of Hydrology*, **267**(1–2), 53–64.

Mitchell T.D. and Hulme M. (1999) Predicting regional climate change: living with uncertainty. *Progress in Physical Geography*, **23**(1), 57–78.

Menzel L., Niehoff D., Bürger G. and Bronstert A. (2002) Climate change impacts on river flooding: a modeling study of three meso-scale catchments. In *Climatic Change, Implications for the Hydrological Cycle and for Water Management. Advances in Global Change Research*, Beniston M. (Ed.), Kluwer Academic Publishers: Dordrecht, pp. 249–269.

Mücher C.A., Stomph T.J. and Fresco L.O. (1993) *Proposal for a Global Land Use Classification*, Food and Agricultural Organisation & Wageningen Agricultural University: Rome & Wageningen.

Niehoff D. (2002) *Modellierung des Einflusses der Landnutzung auf die Hochwasserentstehung in der Mesoskala*, Brandenburgische Umweltberichte, 11, University of Potsdam.

Niehoff D. and Bronstert A. (2001) Influences of land-use and land-surface conditions on flood generation: a simulation study. In *Advances in Urban Stormwater and Agricultural Source Controls*, NATO Science Series IV. Earth and Environmental Sciences, Marsalek J., Watt E., Zeman E. and Sieker H. (Eds.), Kluwer Academic Publishers: pp. 267–278.

Niehoff D., Fritsch U. and Bronstert A. (2002) Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany. *Journal of Hydrology*, **267**(1–2), 80–93.

Norby R.J., Wullschleger S.D. and Gunderson C.A. (1996) Tree responses to elevated CO2 and implications for forests. In *Carbon Dioxide and Terrestrial Ecosystems*, Koch G.W. and Mooney H.A. (Eds.), Academic Press: San Diego.

Rapp J. and Schönwiese C.-D. (1996) *Atlas der Niederschlags- und Temperaturtrends in Deutschland 1891-1990*, Frankfurter Geowiss. Arb. Ser. B, Bd. 5.

Roeckner E., Oberhuber J.M., Bacher A., Christoph M. and Kirchner I. (1996) ENSO variability and atmospheric response in a global coupled atmosphere-ocean GCM. *Climate Dynamics*, **12**, 737–754.

Schaber J. (2002) Phenology in Germany in the 20th century: methods, analyses and models. Potsdam-Institute for Climate Impact Research (PIK). *PIK-Report*, **78**, 145.

Schulla J. (1997) *Hydrologische modellierung von flussgebieten zur abschätzung der folgen von klimaänderungen*, Zürcher Geographische Schriften, 69, Geographisches Institut ETH: Zürich.

Sharrat B.S. (1992) Growing season trends in the Alaskan climate record. *Arctic*, **45**, 124–127.

Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Water Resources Publications: Highlands Ranch, p. 1144.

Thieken A. and Merz B. (2003) Klimaänderungsszenarien und hochwasserentwicklung im rheingebiet: modellierung und unsicherheiten. In *'Klima-Hydrologie-Flussgebietsmanagement im Lichte der Flut 2002*, Leibundgut C. (Ed.), *Proceedings of the Hydrology Day 2003*, 20./21.3.2003 in Freiburg; *Forum für Hydrologie und Wasserbewirtschaftung*, **3**(4), 51–58.

UNEP (1999) *Terminology for Integrated Resources Planning and Management*, Food and Agriculture Organization/United Nations Environmental Programme: Rome/Kenya.

Weichert A. and Bürger G. (1998) Linear vs. non-linear techniques in downscaling. *Climate Research*, **10**, 83–93.

Wilby R.L., Hay L.E. and Leavesley G.H. (1999) A comparison of downscaled and raw GCM output: implications for climate change scenarios in the San Juan River basin, Colorado. *Journal of Hydrology*, **225**, 67–91.

Wilby R.L. and Wigley T.M.L. (1997) Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, **21**, 530–548.

WL Delft Hydraulics and the Ministry of Transport, Public Works and Water Management (1997) SOBEK, Technical Reference Manual.

Xu C.Y. (1999) Climate change and hydrologic models: a review of existing gaps and recent research developments. *Water Resources Management*, **13**(5), 369–382.

133: Rainfall-runoff Modeling of Ungauged Catchments

GÜNTER BLÖSCHL

Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Vienna, Austria

Catchments where no runoff data are available are termed ungauged catchments. For these catchments, the parameters of rainfall-runoff models cannot be obtained by the calibration on runoff data and hence need to be obtained by other methods. Model parameters that require calibration are usually transposed from similar gauged catchments. This article reviews concepts for identifying hydrologic similarity as well as methods for transposing the parameters of both event models and explicit soil moisture accounting (ESMA) models. Model parameters that are physically based are usually measured or inferred from other data within the ungauged catchment of interest. This article summarizes the most important methods and discusses the issues of using point scale field data in rainfall-runoff models. Alternatives to runoff data for model calibration are suggested. The value of soft data and qualitative field observations is emphasized.

INTRODUCTION

Most catchments of the world are ungauged. Even though more than 60 000 stream gauges are installed worldwide (WMO, 1995), there are a couple of orders of magnitude more catchments where no runoff data are available. These are termed *ungauged catchments*. In these catchments, rainfall-runoff models must be used to calculate runoff and other variables, given rainfall (and other climate variables) as an input. There are both operational and academic drivers for pursuing rainfall-runoff modeling of ungauged catchments. The former include design applications (of spillways, culverts, and embankments), forecasting applications (flood warning and hydropower operation), and catchment management applications (water allocation, climate impact studies), the latter are geared toward understanding the catchment functioning and how the individual processes combine to produce catchment response.

The main challenge with rainfall-runoff modeling in ungauged catchments is the lack of local runoff data that could be used for calibrating model parameters. Parameter calibration is important for a number of reasons. First, with the exception of the water balance equation, there is no unique hydrological equation that can be derived from first principles, so most of the model equations are empirical in nature and tend to depend on the hydrological setting (Chapter 1 On the fundamentals of hydrological sciences. Volume 1). Calibration can account for the effects of the hydrological setting in a particular catchment. Second, hydrological models are very much dependent on their boundary conditions, and these are often poorly defined. Calibration can adjust for biases in the inputs, for example, as a result of orographic effects and instrument biases. Third, and probably most important, the media properties (both soil and vegetation) are highly heterogeneous and essentially always unknown or at least poorly known. Soil properties can change dramatically in space but change very little with time, so parameter calibration can significantly enhance the performance of rainfall-runoff models. This led Beven (2000) to remark that: "However such models are constructed, they will have some parameters that will need to be defined for each application site". While the calibration on runoff data has served hydrology well in the past, this is not an option in ungauged catchments. Alternatives are needed which are the subject of this article.

The most efficient method of addressing the issue of ungauged catchments is to glean the model parameters (and perhaps the model structure) from analogue catchments in

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the region, that is, from one or more catchments that one can expect to behave similarly to the catchment of interest. This article therefore first discusses hydrological similarity in the context of rainfall-runoff modeling (Section "Hydrological similarity in rainfall-runoff modeling"). Rainfall-runoff models use two broad categories of model parameters: physically based parameters that in principle can be observed or estimated directly from measurements in a catchment; and calibration parameters that appear in empirical relationships and need to be back-calculated from rainfall and runoff data. The calibration parameters need to be transposed from similar, gauged catchments and these methods are reviewed in Section "Transposing calibration parameters from similar, gauged catchments". Physically based parameters, on the other hand, can be observed in the catchment of interest, although with some restrictions. This is discussed in the Section "Measuring or inferring physically based model parameters in an ungauged catchment". Even though in an ungauged catchment no runoff data are available, other hydrologic response data may be available and could be used as an alternative for model calibration and testing. This is dealt with in the Section "Alternatives to runoff data for model calibration".

HYDROLOGICAL SIMILARITY IN RAINFALL-RUNOFF MODELING

If the rainfall-runoff processes in two catchments are similar, their hydrologic responses will be similar too. This is the notion of hydrologic similarity. Similarity of hydrological processes can be defined in various ways. Dunne (1978) suggested that runoff processes are mainly controlled by physioclimatic controls and identified three main types: Infiltration excess runoff which is generated from partial areas where surface hydraulic conductivities are low; saturation excess runoff which is generated in areas with shallow water tables or near-channel wetlands; and subsurface storm flow which is likely to be active and dominant on steep, humid forested hillslopes with very permeable surface soils. In two similar catchments, the relative role of each of these processes would be similar. The characteristics of these processes in natural catchments are never known in full detail so a number of similarity concepts have been proposed in the literature that attempt to represent these processes to various degrees.

1. *Spatial proximity:* In the first concept, catchments that are close to each other are assumed to behave in hydrologically similar manner. The rationale of this concept is that the controls on the rainfall-runoff relationship are likely to vary smoothly in space, so one can expect spatial proximity to be a good indicator of the similarity of catchment response. This has been one of the traditional methods of regionalizing rainfall-runoff model parameters and is usually based on delineating on a map, spatially contiguous regions with approximately homogeneous model parameters. The regions are found from an analysis of a small number of gauged catchments in each of the regions (e.g. Nathan and McMahon, 1990) by exercising expert judgment supported by whatever hydrologic information that is available. Hydrologic information may consist of hydrogeologic maps, climate maps, soil and vegetation maps, and the seasonality of hydrologic response as an indicator of hydrologic processes (e.g. Merz *et al.*, 1999). The regions found in this way will be plausible but it is not straightforward to formalize the procedure. Also, the uncertainty associated with identifying the regions is not usually known. As an alternative to contiguous regions, interpolation methods such as "kriging" have been used more recently as one way of exploiting spatial proximity as a similarity measure (e.g. Merz and Blöschl, 2004, 2005).

2. *Similar catchment attributes:* The second similarity concept consists of using measurable catchment attributes as indicators of hydrological similarity. The assumption is that if catchment attributes such as soil type, vegetation type, and topographic characteristics of two catchments are similar, one would expect the hydrologic response to be similar. Again, this is a plausible assumption, although few studies have conclusively demonstrated the predictive power of catchment attributes. Catchment attributes can be used in different ways. The first method is to use them to define a group of hydrologically similar catchments. The similarity between catchments is usually quantified by a distance measure as a function of the differences of the catchment attributes in the two catchments. The distance measure is zero if the catchment attributes in the two catchments are identical and increases as the attributes get more dissimilar. From the matrix of distance measures, the grouping can then be obtained by a range of statistical methods such as cluster analysis, principal component analysis, and classification trees (Nathan and McMahon, 1990; Bates, 1994; Breiman *et al.*, 1984). One particular variant of this method is the region of influence approach where for each catchment of interest a separate pooling group is formed (Burn and Boorman, 1993). Once the groups are identified, the model parameters (and perhaps the model structure) can be transferred from an analogue gauged catchment to an ungauged catchment within the same group.

The second method of using catchment attributes consists of defining relationships between model parameters and catchment attributes. There is a wide spectrum of methods ranging from multiple regression methods for calibration parameters (see Section "Explicit soil moisture accounting models") to empirical formulae in engineering design (see Section "Event models"). The structure of these relationships is often dictated by parsimony and convenience although some degree of process reasoning can come in. The parameters of the relationships are usually based on the analysis of a large number of gauged catchments. In

multiple regressions, one may encounter problems with multicollinearity if one or more of the catchment attributes is highly correlated with another catchment attribute or with some linear combination of them. Multicollinearity will reduce the reliability of the regression coefficients (Hirsch *et al.*, 1992). One therefore limits the number of catchment attributes used in the regression, sometimes combining a number of attributes into an index that is deemed representative of one particular aspect of the rainfall-runoff model (such as a baseflow index, IH, 1999). Sequential regression may assist in identifying robust parameter estimates (Calver *et al.*, 2004; Lamb and Kay, 2004). The extension of this general approach of using catchment attributes are methods that derive physically based model parameters from landscape or catchment attributes (see Sections "Local scale measurements and relation to landscape attributes" and "Upscaling local measurements to catchment/model element scale"). Sometimes, the first and second methods are combined in using relationships between model parameters and catchment attributes only within regions that are homogeneous with respect to a certain set of catchment attributes (e.g. Burn and Boorman, 1993). A formal way of combining the two methods results in the CART model (Breiman *et al.*, 1984; Laaha and Blöschl, 2005a). In CART models, the independent variables are the catchment attributes and the dependent variables are the model parameters. Regression trees then divide a heterogeneous domain into a number of more homogeneous regions by maximizing the homogeneity of model parameters and catchment attributes within each group simultaneously. Regression trees have a number of advantages over other models. Their structure is nonparametric, small trees are readily interpretable, there is no global sensitivity to outliers, and they are able to handle nonlinear relationships well. However, big trees are difficult to interpret, there is a lack of smoothness, and there are potential problems with overfitting the data, so the trees need to be pruned (see, e.g. Breiman *et al.*, 1984).

The third method of using catchment attributes in defining hydrological similarity has originated from the use of geographical information systems in hydrological modeling. In this method, patches of the landscape are classified into groups of similar topographic slope, aspect, elevation, vegetation type, soil type, and precipitation distribution, for example, by overlaying maps of the different types of information. Each group is termed a *hydrological response unit* (HRU) and all the elements in a group are assumed to exhibit the same hydrological response characteristics (Leavesley and Stannard, 1995). This again is based on the assumption that similar catchment attributes would be associated with similar hydrologic processes. The way the layers of information are combined can have various degrees of process representation. Flügel (1995), for example, combined the layers by reasoning such as "rangeland on gley soil at the valley floor with shallow groundwater over impervious bedrock". HRUs are often the method of choice when inferring physically based model parameters from landscape attributes. The main difficulties with the use of HRUs is that the estimation of parameters for multiple HRUs in a catchment is not simple, either in calibration or for ungauged catchments (*see* **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**), and that measurable landscape attributes and model parameters are often poorly correlated (see Sections "Limitations and challenges" and "Local scale measurements and relation to landscape attributes").

3. *Similarity indices:* Similarity indices are based on some understanding of the structure of runoff generation and runoff routing and are usually defined as a dimensionless number. Two catchments or two parts of the landscape would then behave hydrologically similar if they are associated with the same value of the similarity index (Blöschl and Sivapalan, 1995; Aryal *et al.*, 2002; *see also* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**). Similarity indices differ in terms of the processes they aim to represent. Similarity in climate can be quantified by the aridity index of Budyko (1974), for example, which is the ratio of long-term potential evaporation to precipitation. Atkinson *et al.* (2002) and Farmer *et al.* (2003) used this type of climatic index for analyzing suitable rainfall-runoff model complexities as a function of timescale and climate characteristics. Similarity in channel flow processes can be quantified by the stream ordering and other characteristics of the map view of the channel topology (e.g. Rodríguez-Iturbe and Valdés, 1979; Rinaldo *et al.*, 1995). Similarity in catchment processes can be quantified by topographic indices (Moore *et al.*, 1991). Examples include the topographic wetness index of Beven and Kirkby (1979), which is a function of the area drained per unit contour length at a given point and the local slope gradient. Catchment characteristics such as soil depths may be incorporated in the index. Each patch in the landscape exhibiting the same value of the index is considered to produce the same hydrologic response. Two catchments are then deemed similar if their distributions of the index are identical. On the basis of this similarity index, Sivapalan *et al.* (1987) identified five nondimensional similarity parameters that represented the interrelationships of topography, soil, and rainfall, which lead to similar catchment responses. Larsen *et al.* (1994) and Robinson and Sivapalan (1995) extended this work concentrating on the relative dominance of the infiltration excess and saturation excess mechanisms. The advantage of this type of index is that they are not purely empirical so their structure may be more defendable than that of the empirical relationships between model parameters and catchment attributes discussed previously. The drawback is that the inputs needed for the indices are often difficult to specify in ungauged catchments. Also, tests of some

of these indices against spatial soil moisture data have given mixed results, as their applicability depends on catchment conditions and climate (e.g. Lamb *et al.*, 1998; Western *et al.*, 1999; Blazkova *et al.*, 2002). Similarity indices are therefore currently not widely used in practice although they constitute an active and promising area of research.

The various concepts of hydrological similarity are potentially useful for inferring a suitable model structure in ungauged catchments, but their main application lies in transposing model parameters in space. The following section focuses on the case of calibration parameters.

TRANSPOSING CALIBRATION PARAMETERS FROM SIMILAR, GAUGED CATCHMENTS

Methods

Calibration parameters are those that cannot be measured or inferred from measurements but need to be transposed from gauged catchments in the region. These gauged catchments should be hydrologically similar to the catchment of interest in the sense discussed earlier. Rainfall-runoff models can be either event models or ESMA models (*see* Beven 2001a and **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**). Event models require initial conditions on the soil moisture status of the catchment and are usually run over several hours. ESMA models, in contrast, account for the changes of soil moisture stores in a catchment through taking stock of evapotranspiration, precipitation, and runoff and are usually run over several years. Different model parameters exist in each type of model: runoff coefficient (or loss) parameters and lag time (or time to peak) parameters in the case of event models; and evaporation parameters, runoff generation parameters and storage coefficients in the case of ESMA models. The procedures for transposing these parameters from similar, gauged catchments in the region to an ungauged catchment are, however, similar for the two model types. The transposition typically involves the following steps:

1. Delineation of homogeneous regions and/or identification of one or more gauged catchments, termed *donor catchments*, based on any of the similarity measures discussed in Section "Hydrological similarity in rainfall-runoff modeling".
2. Estimation of model parameters for the donor catchments by manual or automatic calibration on observed runoff data (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**).
3. Selection of catchment attributes that are deemed to affect catchment response to rainfall. This is either based on an *a priori* understanding of what attributes may be relevant to a particular model parameter or on some goodness-of-fit measure.
4. Setting up models relating each rainfall-runoff model parameter to a set of catchment attributes. In this step, commonly, multiple linear regressions are used (see Section "Hydrological similarity in rainfall-runoff modelling"), and some or all of the catchment attributes are (e.g. logarithmically) transformed.
5. Testing the strength of the relationship of (4), for example, by some goodness-of-fit measure such as a correlation coefficient.
6. Estimating each parameter of the rainfall-runoff model for the ungauged catchment from the (regression) model.
7. Simulating runoff for the ungauged catchment of interest by applying the same model as in (2), using the regionally transposed model parameters.
8. Testing the transposition by cross-validation. A gauged catchment is assumed to be ungauged, runoff is simulated as in (7) and then compared with the locally observed runoff.

Some of these steps can be skipped depending on the data availability and the similarity concept chosen. In the simplest case of using model parameters from one or more donor catchments directly in the ungauged catchment, steps (3 – 6) can be skipped. Alternatively, results from the literature can be used for any of the steps. The less the information available in the region of interest, the fewer the steps that can be carried out, but this is likely at the cost of decreasing model performance. In each of these steps it is prudent to not only find the "best" relationships or parameters but also to find some measure of the uncertainty associated with the estimates. In step (1) alternative regions and/or donor catchments can be examined and their suitability can be tested by cross-validation (step 5); in step (2) formal methods of parameter uncertainty analysis can be used or less formal sensitivity analyses and split sample methods (*see* **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**); in step (3) alternative catchment attributes or combinations of catchment attributes can be used and tested in a similar way; and in step (4) alternative transposition models can be tested. The advantage of the cross-validation test in step (5) over other methods of assessing uncertainty, such as sensitivity analyses, is that it examines a combination of a number of sources of uncertainty including the data, the model parameters, and the structure of both the runoff model and the transposition model.

Event Models

As an example for estimating model parameters for event-based rainfall-runoff models, the procedure recommended in the UK Flood Estimation Handbook (FEH) (IH, 1999: `http://www.nwl.ac.uk/ih/feh/`) will be very briefly summarized. The general recommendation in the FEH is to

use runoff data wherever possible including records as short as a year. In ungauged catchments, this is obviously not possible. The preferred method for ungauged catchments is to use the parameters from a similar (donor) catchment in the region. As a last resort, relationships between model parameters and catchment attributes can be used. These involve the largest errors. It is also recommended to combine the last two methods and scale the parameters from the donor catchment by the ratio of the parameter estimates from catchment attributes in the catchment of interest and those in the donor catchment. It is suggested to estimate the time-to-peak parameter, if needed, from the following catchment attributes: mean drainage path slope, mean drainage path length, extent of urban land cover, and a parameter representing the proportion of time when soil moisture deficit was below 6 mm during a standard period. The runoff coefficient can be estimated as the sum of three terms. The first represents the normal capacity of a catchment to generate runoff. It varies between catchments but does not vary between storms and is estimated from the Hydrology of Soil Types (HOST) soil classes (Boorman *et al.*, 1995). The second term represents the variation in runoff depending on the antecedent soil moisture and can be estimated from mean annual precipitation. The third term depends on the storm rainfall depth and represents the nonlinearity in runoff generation. There is also a correction for urbanization.

There are numerous recommendations in the engineering hydrology literature of how to obtain model parameters in ungauged catchments, but they all fall into one of the two categories of transpositions from similar, gauged catchments and empirical formulae (e.g. USACE, 1994; Bates, 1994). In the empirical formulae, lag time parameters are often a function of morphometric parameters such as topographic slope, stream slope, and flow length (e.g. USACE, 1994). Loss parameters are often a function of land cover, soil type, and antecedent soil moisture (such as in the Curve Number method of the US Soil conservation service (SCS, 1973; Mishra and Singh, 2003; Merz *et al.*, 2005). There is also the option of making use of the structure of the empirical equations and adjusting their coefficients by analyzing runoff data in the region which, from a practical perspective, is probably more meaningful than applying any of the (black box) regression equations in use. In most studies, the emphasis is on finding a best-fit curve between the runoff model parameters and their controls. The uncertainty around these best-fit curves is often immense (Cordery and Pilgrim, 1983). Because of this, similar to the suggestion of the FEH, the general recommendation is to analyze runoff data from similar gauged catchments in the region whenever possible and use them by one method or another for estimating the model parameters in the ungauged catchment of interest.

Explicit Soil Moisture Accounting Models

Little is known on the process relationships between the parameters of conceptual ESMA models and the measurable catchment attributes. Because of this, the most widely used method of estimating ESMA model parameters is either by using uniform zones or regions (e.g. lithological zones in Drogue *et al.*, 2002), or by multiple regressions from catchment attributes. However, most of the case studies published in the literature have found rather low correlations. Sefton and Howarth (1998), for example, compared calibrated parameters of the IHACRES model with attributes of 60 catchments in England and Wales. The best correlations they obtained were $R^2 = 0.59$ between a routing parameter and percentage of aquifers, and $R^2 = 0.69$ between an evaporation parameter and mean annual precipitation. For the storage parameters no significant correlations were obtained. Seibert (1999) related the model parameters of the HBV model (Bergström, 1976) to attributes of 11 Swedish catchments within the NOPEX area. The relationships between forest percentage and snow parameters could be interpreted on hydrological grounds but other relationships could not. The rank correlation coefficient between a nonlinearity parameter of runoff generation and catchment area was $R^2 = 0.87$ but most other parameters exhibited hardly any significant correlations with catchment attributes. These typically low correlations are likely to translate into rather low model performances for the ungauged catchment case. In Seibert's (1999) study, the median Nash and Sutcliffe (1970) model efficiency decreased from 0.81 to 0.79 when moving from calibrated parameters to regionalized parameters for the same set of 11 catchments, but the median efficiencies decreased to 0.67 for a separate set of 7 catchments. A model efficiency of 1 indicates a perfect fit while smaller values indicate poorer fits. In a Norwegian study, Beldring *et al.* (2002) used 141 catchments for calibrating a version of the HBV model. They then treated 43 additional catchments as ungauged and regionalized the model parameters as a function of land use classes. For both sets of catchments they found median Nash–Sutcliffe efficiencies of 0.68 and concluded that the regionalization method represented the main features of the landscape well. However, for 20% of the second set of stations the efficiencies were less than 0.3.

In a recent study, Merz and Blöschl (2004) simulated the water balance dynamics of 308 catchments in Austria using a lumped ESMA model involving 11 calibration parameters. The parameters were calibrated separately on two nonoverlapping 11 year periods of daily runoff data to assess the reliability of the calibrated model parameters. Figure 1 shows two of the parameters. Figure 1(a) gives the calibrated *beta* parameter, which is the nonlinearity parameter in the function relating runoff generation to the soil moisture state. There are distinct patterns of low values in the West and high values in the East. The regional

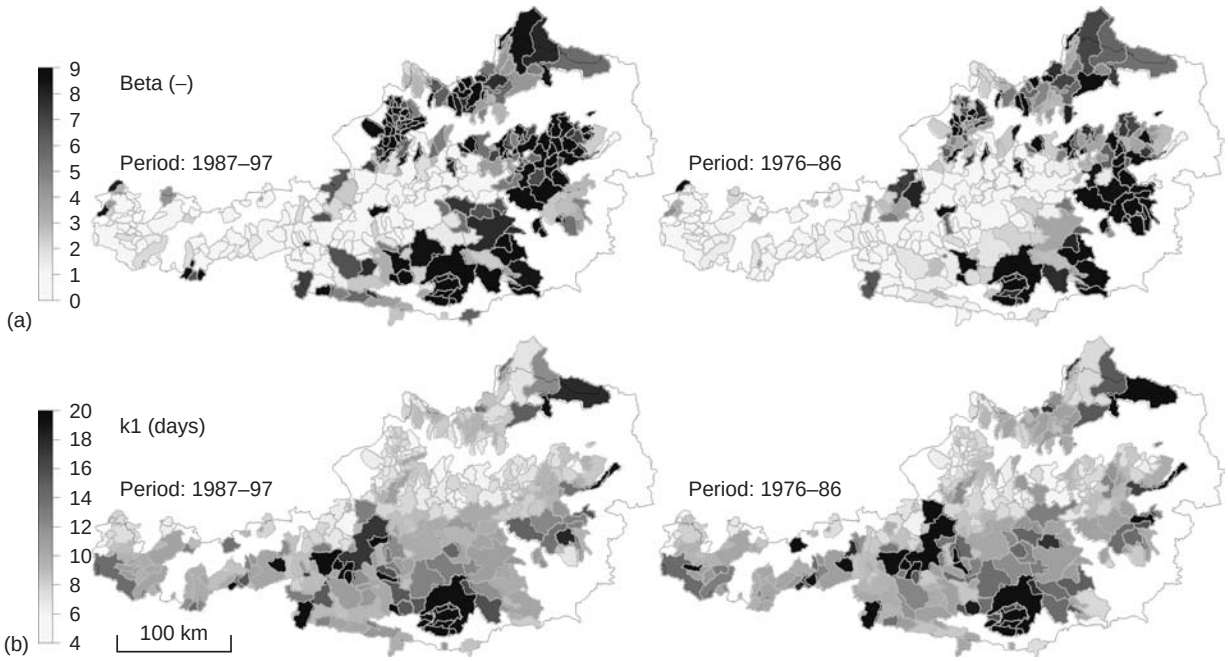


Figure 1 Patterns of calibrated model parameters (left: calibration period 1987–1997, right: calibration period 1976–1986). (a) Nonlinearity parameter, *beta* (–). (b) Fast storage coefficient k1 (days) (From Merz and Blöschl, 2004 by permission of Elsevier). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

differences in *beta* imply a relatively linear rainfall-runoff relationship and large runoff coefficients in the wetter alpine catchment in the West, and a nonlinear rainfall-runoff relationship and small runoff coefficients in the dryer lowland catchments in the East. Figure 1 (b) shows the spatial patterns of the fast storage coefficient. There is a tendency for faster responses in the prealpine catchments of the North than in the alpine catchments of the South. The regional patterns of the parameters for the two periods are similar although large local differences occur. This suggests that the calibrated parameters are able to represent the regional or large-scale differences in the hydrological conditions. One would therefore assume that it is possible to derive regional relationships between the calibrated parameter values and catchment attributes. Merz and Blöschl (2004), however, found that the correlation coefficients were relatively low with maximum values of $R^2 = 0.27$. They then compared various methods for estimating the model parameters in ungauged catchments in terms of their ability of simulating runoff, judged against runoff data (i.e. a cross-validation; Table 1, Figure 2). The regionalization methods significantly increased the predictive performance from the cases of preset parameters (Figure 2a) and global average parameters (Figure 2b). The increase in performance depended on the regionalization method. Average parameters of immediate upstream and downstream neighbors (Figure 2f) and regionalization by kriging (Figure 2 g) performed somewhat better than multiple regressions with catchment attributes (Figure 2d). For the calibration period, the median of the Nash–Sutcliffe model efficiency decreased from 0.67 to 0.57 when moving from gauged to ungauged catchments (Table 1, using kriging for regionalization). For the verification period, the median decreased from 0.63 to 0.56. This means that the uncertainty introduced by moving from gauged to ungauged catchments is about twice the uncertainty of moving from the calibration to the verification period. Of particular interest is the change in the bias of simulated runoff (i.e. volume errors). While the median values are always small, the scatter between catchments (quantified by the difference of the 75% and 25% quantiles) doubles when moving from the calibration to the verification period, but increases by a factor of 5 when moving from gauged to ungauged catchments. This illustrates the value of model calibration in reducing bias in gauged catchments and the uncertainties in ungauged catchments where calibration is not an option Parajka *et al.* 2005a extended the analysis of Merz and Blöschl (2004) by modifying the model structure and testing additional regionalisation methods. Their model efficiencies were higher than those of Merz and Blöschl (2004) but their main findings remained similar.

Limitations and Challenges

There are three potential explanations of the relatively poor correlations between model parameters and catchment attributes, and the relatively poor performance of both event-based and ESMA models when parameters estimated from catchment attributes are used.

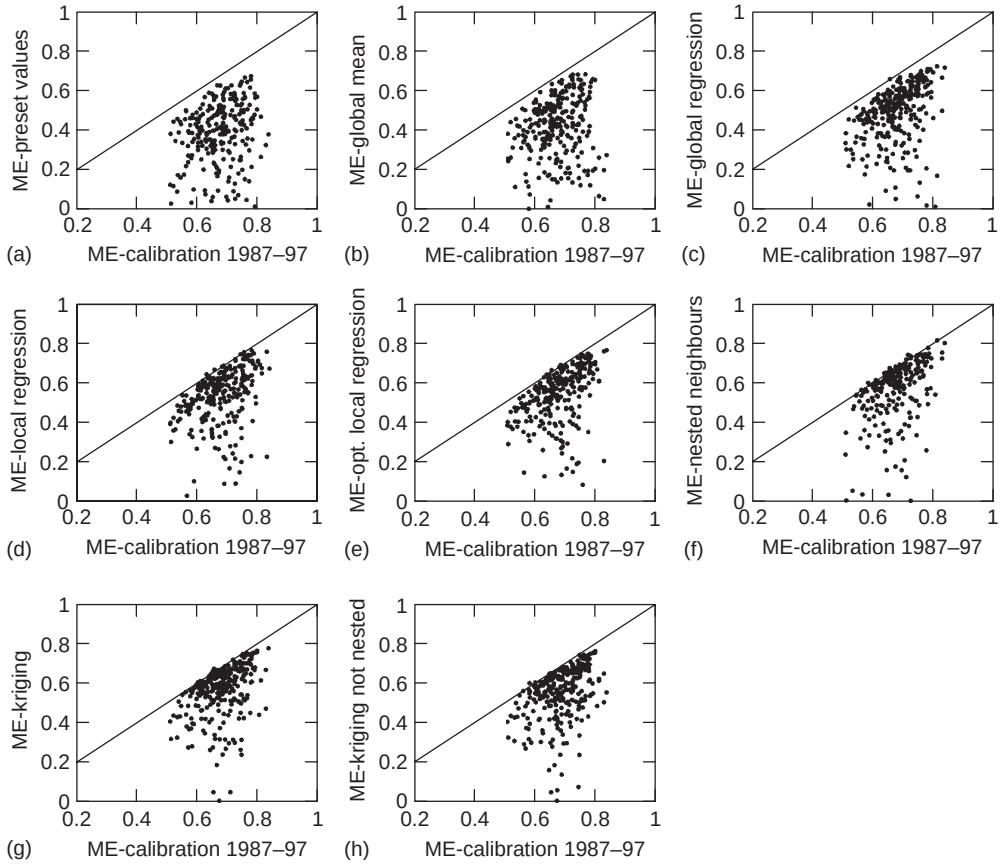


Figure 2 Nash–Sutcliffe model efficiencies of regionalized versus calibrated parameters for the period 1987–1997. (a) preset parameters, (b) global mean of all catchments, (c) global regression with catchment attributes using all catchments, (d) local regression within a 50 km neighborhood, (e) optimized local regression, (f) average parameters of immediate upstream and downstream (nested) neighbours, (g) kriging, (h) kriging without nested neighbours (From Merz and Blöschl, 2004 by permission of Elsevier)

Table 1 Model performance for gauged catchments and ungauged catchments (kriging regionalization) by cross-validation, both for the calibration and the verification periods. ME: median Nash–Sutcliffe efficiency of runoff plus or minus difference of 75% and 25% quantiles of efficiencies, that is, a measure of scatter. Bias: Median relative volume errors of runoff plus or minus difference of 75% and 25% quantiles of relative volume errors. Statistics are for 308 catchments in Austria. (From Merz and Blöschl, 2004 by permission of Elsevier)

	Gauged catchment	Ungauged catchment
ME–Calibration period	0.67 ± 0.10	0.57 ± 0.18
ME–Verification period	0.63 ± 0.11	0.56 ± 0.18
Bias–Calibration period	0.00 ± 0.04	0.04 ± 0.21
Bias–Verification period	-0.01 ± 0.10	0.04 ± 0.21

1. One explanation is that the measurable catchment attributes may not be very relevant for catchment response. This is certainly the case for soil type as reflected by the usually low predictive power of pedotransfer functions (see Section "Local scale measurements and relation to landscape attributes"). Catchment attributes are usually static indicators and may therefore not be very representative of the hydrologic functioning of catchments. Also, much of what is of interest in rainfall-runoff modeling happens beneath the surface while most measurable catchment attributes (such as topographic characteristics and vegetation) may be representative of the surface but are not normally representative of the hydrologically active zone in the subsurface. This will be expanded in Section "Measuring or inferring physically based model parameters in an ungauged catchment" of this article. One avenue to proceed would be to identify catchment attributes that are more representative of the processes relevant to rainfall-runoff modeling, perhaps based on similarity indices, but so far it is not quite clear how to define them at the regional scale.
2. The second explanation is that there may be significant uncertainty in the calibrated parameter values,

which may cloud the underlying relationship between calibrated model parameters and catchment attributes (e.g. Gottschalk, 2002). There are methods of accounting for parameter uncertainty in the regionalization of model parameters (see, e.g. Campbell and Bates (2001), in the case of an event scale rainfall-runoff model illustrated on 39 catchments in southwestern Australia). The examples in Figure 1 suggest that the uncertainty is not very large in this case, but it does seem important to better constrain model parameters in the calibration process. There are two possibilities. The first is to use additional data on state variables in the calibration procedure (Gupta *et al.*, 1998; Mroczkowski *et al.*, 1997; also see **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**). For the study area of Figure 1, Parajka *et al.* (2005b) used snow depth measurements in a multiobjective calibration procedure which lead to better constrained parameters. The second possibility is regional calibration in which the model parameters of a number of catchments are calibrated simultaneously (e.g. Fernandez *et al.*, 2000; Szolgay *et al.*, 2002; Engeland and Gottschalk, 2002). Regional calibration will provide more robust parameters that are likely to translate into a reduction in uncertainty when transposing them to ungauged catchments. Merz *et al.* (2004) also showed that regional calibration may increase the correlations between calibrated model parameters and catchment attributes.

3. The third explanation is that the structure of the model relating catchment attributes and model parameters may not be suitable. Indeed, the choice of a linear regression model is one of convenience rather than one based on know relationships. An additional complication with the linear regression model is that it is an unbiased linear estimator of the model parameters but because the rainfall-runoff models are nonlinear, the regression model results in biased runoff simulations in the ungauged catchment. Also, estimating the model parameters independently from catchment attributes may lead to parameter combinations that are not very realistic. Nonlinear models (such as CART models, Section "Hydrological similarity in rainfall-runoff modelling") exist, and thresholds or bounding conditions may be introduced. For example, in the United Kingdom, the baseflow index (IH, 1999) may take on a full range of values in low rainfall areas of the country but is always low in high rainfall areas of the country. There have been a number of other attempts to include process reasoning into the estimation of calibration parameters of ESMA models. Schumann *et al.* (2000), for example, imposed the structure on some of the equations of estimating the model parameters. Another attempt is that of Koren *et al.* (2003) who derived parameters of the Sacramento model from catchment attributes on the basis of hydrologic reasoning.

There are additional sources of uncertainty including errors in the structure of the rainfall-runoff model and data errors. These are discussed in more detail in Beven (2001a) and also in **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**, **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3**.

MEASURING OR INFERRING PHYSICALLY BASED MODEL PARAMETERS IN AN UNGAUGED CATCHMENT

Methods

Physically based parameters are those that can be observed in the field or estimated directly from measurements of catchment or channel characteristics without the use of runoff data. They possess a physical meaning beyond the particular model used. Examples include roughness parameters such as Manning's n, hydraulic conductivity, soil depth, and surface albedo. There are, however, two types of difficulties with using measured parameters in rainfall-runoff models that are related to scale (Western and Blöschl, 1999; Blöschl, 1999). The first difficulty is that the measurement volume (or *support scale*) is usually much smaller than the model element size. The second is that in most cases there are only a few measurement locations within a catchment and the spacing between the measurements (the *spacing scale*) is hence large. This means that the measured parameter is not defined in exactly the same way as in the model even if it shares the same name (Beven, 1989). In principle, both scale disparities can be addressed by upscaling procedures. In practice, there is no generally accepted upscaling theory (see Section "Upscaling local measurements to catchment / model element scale") and one often neglects the incompatibility related to the support and addresses the incompatibility related to the spacing by some sort of interpolation procedure (see Section "Upscaling local measurements to catchment / model element scale"). In many instances, particularly when remote sensing data are used, model parameters are inferred from surrogates (e.g. roughness from land use). As the model parameters are often not very well correlated with the surrogate, this introduces additional uncertainties.

Because of these difficulties, in gauged catchments physically based parameters are often allowed some degree of calibration within a physically justifiable range to adjust some of the measurement or estimation biases. For ungauged catchments, it may therefore be of value to transfer physically based parameters from similar catchments in the region by one of the methods outlined in Section "Transposing calibration parameters from similar, gauged catchments", in addition to measuring the parameters in the field (or estimating them from remote sensing

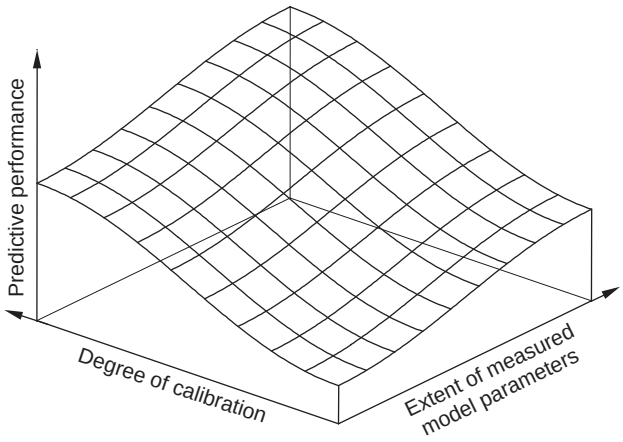


Figure 3 Schematic of the value of calibration model parameters versus measuring model parameters for a typical rainfall-runoff application

data). In fact, estimating the parameters from two or more different sources will always increase their reliability. The distinction between calibration parameters and physically based parameters is hence not a sharp one, there is a gradual transition from what one would call a calibration parameter and a physically based parameter, depending on the type and extent of information available in any particular case. Figure 3 shows a schematic of the relative value of the calibration and the measurement of model parameters for a typical rainfall-runoff model application. Both calibrating and measuring a model parameter tend to enhance the predictive performance of the model. Ideally one should pursue both avenues. The schematic also indicates that, while it is useful to measure physically based parameters, it is even more useful to have access to runoff data in a particular catchment to calibrate or adjust one or more of the model parameters. The following section reviews methods of inferring model parameters from measurements in the catchment of interest.

Local Scale Measurements and Relation to Landscape Attributes

Physically based model parameters can be inferred from field measurements or from remote sensing data. The former are usually local scale measurements and hence best suited for small catchments while the latter are more appropriate for large catchments. Often, parameters are inferred from qualitative information or surrogates. In the following, a brief summary is given of the three main categories of model parameters as used in rainfall-runoff models and how they can be related to surrogates that are more widely available than the model parameters themselves. A practical discussion of some of the parameter estimation methods is given in Duan *et al.* (2001) and Vieux (2001).

1. *Soil hydraulic characteristics*: Soil hydraulic characteristics such as the saturated conductivity, porosity, and soil water release characteristics are usually estimated from infiltration experiments at the plot scale in the field or, alternatively, from laboratory core tests. If no or only a few measurements are available, they are sometimes estimated from relationships to soil type (often defined by % sand, silt, clay, organic matter, and perhaps bulk density). These relationships are termed *pedotransfer functions*. An excellent review of pedotransfer functions and their use in hydrology is given in Wösten *et al.* (2001). The appeal of pedotransfer functions is that soil texture data are now widely available in databases such as the State Soil Geographic Data Base (STATSGO) in North America (USDA, 1991) and the European soil database (Jamagne *et al.*, 2002). The justification of using pedotransfer functions is that the grain-size distribution (defined by the soil type) should also be relevant to the pore-size distribution (which in turn is related to soil hydraulic properties). Unfortunately, this is not often the case because peds and cracks, rather than the grain-size distribution, tend to dominate the hydraulic properties. It is therefore not uncommon for the soil properties to vary as much between soil types as within a soil type (e.g. Warrick *et al.*, 1990) and for other influences such as terrain to be important to soil hydraulic properties (Gessler *et al.*, 1995). This makes Wösten *et al.* (2001) conclude that pedotransfer functions are sufficiently accurate for interpolation purposes between soil hydraulic measurements in the catchment of interest whereas they are not recommended to be used in catchments where no measurements are available. If no measurements are available, a minimum requirement for the use of pedotransfer functions would seem to be that the conditions in the catchment of interest are similar to those in the catchments in which they were derived (see Section "Hydrological similarity in rainfall-runoff modeling"), though it is worth noting that the pedotransfer functions are generally derived from measurements on small soil cores or samples.

2. *Surface roughness*: Surface roughness parameters such as Manning's n are usually determined by *in situ* experiments on irrigation plots (e.g. Hessel *et al.*, 2003). If no measurements are available, tabulated values in the literature are often used, which are a function of land cover and sometimes topographic slope (Engman, 1986). Land-cover type can be obtained either from field surveys or from analyses of satellite data. Roughness can also change with the flow rate (e.g. Sepaskhah and Bondar, 2002). Again, ideally, the conditions of the application should be similar to those where the roughness values were measured.

3. *Vegetation characteristics*: The Leaf Area Index (LAI), the Fraction of Green Vegetation (FGV), and the Fraction of Absorbed Photosynthetic Active Radiation (f_{APAR}) are vegetation characteristics that can be used in rainfall-runoff

models to estimate evapotranspiration. These vegetation characteristics can be related to land-cover classes although the relationships are not always unique (Hall *et al.*, 1995; Gorte, 2000; Kite and Droogers, 2000). Land-cover classes, in turn, can be estimated from a range of satellite data on the basis of indices such as the normalized difference vegetation index (ndvi). There exist numerous satellite-(e.g. AVHRR and Landsat) based land-cover maps such as the European CORINE land-cover data set (Büttner *et al.*, 2002), data sets for North America (see, e.g. Gallo *et al.*, 2001), as well as global data sets (e.g. Hansen *et al.*, 2000; Tucker *et al.*, 2004). Scale incompatibilities may be less stringent with satellite data than with ground data although scale issues do remain (e.g. Brunsell and Gillies, 2003).

Upscaling Local Measurements to Catchment/Model Element Scale

As mentioned previously, the *support* scale incompatibility is usually neglected but some sort of interpolation is often used to bridge the *spacing* scale incompatibility (Blöschl and Sivapalan, 1995; *see also* **Chapter 6, Principles of Hydrological Measurements, Volume 1**). The interpolation procedures usually build on the surrogate variables, discussed in the section "Local scale measurements and relation to landscape attributes", either based on different classes of the surrogates (land-cover class, soil type class, etc.) or on regression relationships or a combination of the two. The classification approach is closely related to the Hydrologic Response Unit concept (Section "Hydrological similarity in rainfall-runoff modeling") where each HRU is a unique combination of attribute classes (land cover, soil type, terrain slope, etc.). Both regressions and classifications based on HRU concepts are widely used in the literature to upscale measurements (see, Blöschl and Grayson, 2000; *see also* **Chapter 11, Upscaling and Downscaling – Dynamic Models, Volume 1**). Examples include Busch *et al.* (1999) and Bormann *et al.* (1999) who identified HRUs based on a cluster analysis of actual evapotranspiration, groundwater recharge, and surface runoff, and then performed one single simulation for each HRU. Upscaling methods based on remotely sensed data are summarized in Stewart *et al.* (1996). Upscaling methods based on geostatistical concepts are reviewed in Bierkens *et al.* (2000) and also in **Chapter 9, Statistical Upscaling and Downscaling in Hydrology, Volume 1**.

There is considerable uncertainty associated with any of these upscaling methods and the magnitude of the uncertainty is not always clear. While the standard procedure is to relate the model parameters to landscape attributes either by classification or regression methods (which are both black box methods), there are also initial attempts at incorporating hydrological understanding. An excellent example is presented in Peschke *et al.* (1999a,b) who, based on many years of field experience, mapped the *type* of runoff generation mechanism that occurred for a given catchment state in the 4.6 km^2 Wernersbach catchment in Germany. They then developed an expert system, known as *FLAB*, that estimates the dominant runoff mechanism for a given point in the landscape and a given event size. The mechanisms included were Hortonian overland flow, saturation area overland flow, interflow, recharge, and storage. The expert system is based on rules. For example, potentially contributing areas are identified on the basis of connectivity to the stream, terrain slope, and soil characteristics including layering. If, for a given event, the saturation deficit is exceeded, these areas become active saturation areas and contribute to saturation area overland flow. The rules are based both on hydrologic reasoning and field mapped runoff mechanisms for certain event types. Figure 4 shows a test of the FLAB model against the field mapped areas for two runoff mechanisms, saturation area runoff (Figure 4 (a)), and interflow (Figure 4 (b)). Each figure is a composite of the observed and FLAB simulated binary pattern of the presence/absence of the given runoff mechanism. Black indicates areas where both model and field observations identify the mechanism, grey indicates areas where presence of the mechanism is modeled but not observed, and light grey indicates areas where the presence of the mechanism is observed but not modeled. Overall, there is a close match. Saturation areas mainly occur close to the stream, interflow occurs at some distance from the stream, and this general pattern is modulated by soils and terrain slope. The strength of the FLAB approach is that the rules are based on *field scale* observations rather than on laboratory-scale measurements. This general method has the potential of being used in ungauged catchments for identifying runoff mechanisms and hence upscaling local measurements. Although the rules are not likely to be universal, identifying these rules for certain catchment regime types seems a promising area of research (Scherrer and Naef, 2003; Woods, 2003).

Soft Data and Qualitative Field Observations

Strategies of upscaling such as those of Peschke *et al.* (1999a,b) are partly based on qualitative information which one may term *"soft data"*. In addition to upscaling local measurements, this type of soft information can be used for estimating or, at least, for making an educated guess on the magnitude of model parameters in ungauged catchments. Clearly, site visits will be instrumental in this type of assessment. The value of qualitative information is illustrated by an example from the Austrian Alps. Wienerbruck and Mitterbach are two adjacent catchments that are similar in size and are similar in terms of their catchment attributes as available at the regional scale (Table 2). Wienerbruck is slightly steeper which would suggest somewhat faster response but the channel lengths are slightly larger, so most empirical equations suitable for ungauged catchments would give similar estimates of time-to-peak model

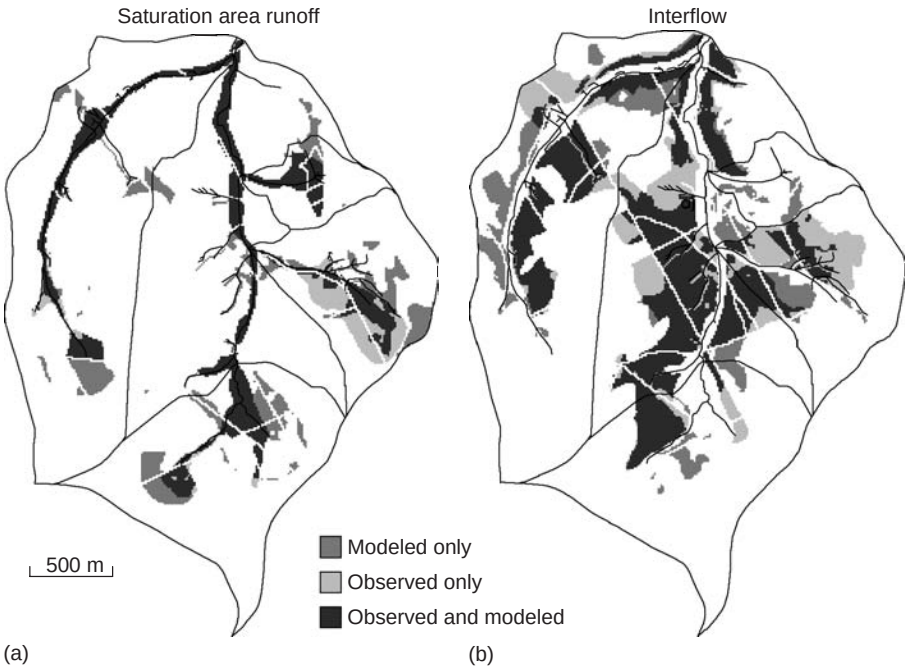


Figure 4 Comparison of the FLAB model estimates against field mapped areas of the presence of two runoff mechanisms (a): saturation area runoff, (b): interflow, for a 5 h, 25 mm event and wet antecedent soil moisture status (Peschke *et al.,* (1999b). Reproduced with permission of IHI Zittau). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

parameters in the two catchments. Similarly, surrogate-based estimates of physically based model parameters would give similar values, as land use and soil types are similar. Since these two catchments are gauged, one can test this assessment. Table 3 shows the event characteristics of the largest event on record (August 2, 1991). While runoff coefficients in the two catchments were similar, the response time (here in terms of the recession constant near peak) in the Mitterbach was almost 10 times that of the Wienerbruck catchment. It would be clearly impossible to predict the differences in catchment response between the two catchments on the basis of the quantitative catchment attributes. In contrast, soft information obtained through a visual examination of the catchments during site visits may help tremendously. Figure 5 shows photographs that are representative of the landforms in each of the catchments. In the slow response catchment (Mitterbach, (a)) the stream channel is mossy. Quite clearly, there is very little hydrologic activity. In contrast, the fast-response catchment (Wienerbruck, (b)) exhibits signs of erosion rills and deeply incised channels, that is, this is a hydrologically highly active catchment. From the visual assessment it would not be impossible to predict the magnitude differences in response timing. It is clear that physical controls do lead to these differences in catchment response but these are not apparent in the quantitative catchment attributes usually available in rainfall-runoff analyses. The geologic database available at the regional scale indicates limestone as the dominant geologic formation in both catchments. A more detailed survey indicates that the difference in response is likely a function of a number of factors, including topography, geology, and the valley bottom infill. In both catchments, dolomites from the Upper Trias prevail although in slightly different variants (Dachsteindolomit in Mitterbach and Wettersteindolomit in Wienerbruck). In Mitterbach, moraines have accumulated on some of the valley bottoms while in Wienerbruck the valley bottoms are narrow without infill. In a small part of the upper Mitterbach catchment, karstic limestone appears to exist giving rise to karstic springs. Soils tend to be deeper in Mitterbach than in Wienerbruck. In a *post hoc* analysis one can speculate that the steeper slopes along with more rainfall have lead to feedback effects between runoff processes, soil development, and vegetation development through some sort of coevolution (see e.g. Hoosbeek and Bryant, 1993; Lucas, 2001; *see also*, **Chapter 12, Co-evolution of Climate, Soil and Vegetation, Volume 1**). In an *a priori* mode, as needed in ungauged catchments, there is clearly a very important role of this type of field assessment.

There have been few attempts at formalizing the soft information from an expert assessment of the catchment functioning of the type in the above example. One notable exception is presented in Seibert and McDonnell (2002). They proposed a method where qualitative knowledge from field surveys is made useful through defining a fuzzy membership function that can be used in calibrating a

Table 2 Catchment attributes of two adjacent catchments in the Austrian Alps

	Ötscherbach at Wienerbruck	Große Erlauf at Mitterbach
Catchment area (km^2)	36.1	29.7
Mean topographic elevation (m)	1013	984
Mean topographic slope (%)	30.0	22.3
Maximum channel length (km)	10.2	8.7
Geology	Limestone	Limestone
Soils	Rendzina	Rendzina
Forest (%)	84.0	81.0
Grass (%)	14.2	16.7
Rock (%)	1.9	0.0
Mean annual precipitation (mm)	1680	1415

Table 3 Characteristics of the largest flood event on record of the two catchments of Table 2 and recession constants of five largest events

	Ötscherbach at Wienerbruck	Große Erlauf at Mitterbach
Event rainfall depth (mm) of event on August 2, 1991	255	167
Runoff coefficient (–) of event on August 2, 1991	0.78	0.73
Recession constant near peak (h) of event on August 2, 1991	3.8	36.1
Recession constant near peak (average of 5 largest events $\pm$ standard deviation) (h)	6.4 ± 1.8	35.9 ± 7.9

rainfall-runoff model. Seibert and McDonnell tested their method in the Maimai research catchment in New Zealand. They achieved very good fits for a three-box model when optimizing the parameter values with only runoff data, but parameter sets obtained in this way showed in general a poor goodness of fit for other criteria such as the simulated new water contributions to peak runoff. Inclusion of soft data criteria in the model calibration process resulted in lower runoff model efficiency values but led to a better overall performance, as interpreted by the experimentalist's view of catchment runoff dynamics. One of the difficulties with this type of analysis is that normally it will be necessary to introduce additional parameters to make use of tracer data (be it artificial tracers or natural tracers including geochemical data) because wave velocities and flow velocities are not necessarily the same as assumed in their study. However, this general approach of including soft information does appear to have considerable potential for application in ungauged catchments provided soft information in the catchment of interest is available.

ALTERNATIVES TO RUNOFF DATA FOR MODEL CALIBRATION

So far this article has focused on methods of parameter estimation that do not use any runoff data for calibration in the catchment of interest as the article is concerned with the case of ungauged catchments. Even if no runoff data are available there may exist other hydrologic response data that can be used to calibrate some or all of the model parameters. In the following, two examples are given to illustrate the potential of hydrologic response data. The first example is set in the 90 km^2 Schneealpe area in the Austrian Alps (Blöschl *et al.*, 2002). In this study, calibration of model parameters on runoff data was not possible because the catchment was highly Karstic and it was unclear what the hydrologic catchment area was. The catchment was therefore treated as ungauged and the model parameters were calibrated on alternative response data. The main interest in the study was on snow processes, so snow cover patterns were deemed to be suitable response data. Snow cover patterns for the years 1998–2000 were derived from Systeme Probatoire pour l'Observation de la Terre SPOT XS images based on an unsupervised isodata technique (Jensen, 1996), separately for different illumination classes. The classification produced three class patterns of snow, no-snow, and partial coverage. The rainfall-runoff model focused on the snow component and simulated the snow pack evolution and snow melt for each grid cell on the basis of an energy balance approach. Wind drift of snow was represented by a wind drift factor, which was a function of terrain elevation, slope, and curvature. For the calibration of the model parameters, four SPOT-derived cover patterns were chosen from each of the years 1998 and 1999. These eight patterns were compared with the simulated patterns for the same dates. The calibration proceeded in two steps. In the first step, snow albedo and the threshold air temperature for separating snowfall and rainfall were calibrated. The objective function was based on differences in the snow-covered area of simulation and observation, both lumped for the entire study area and stratified by terrain slope and aspect. This made it possible to separate the effects of albedo from those of the threshold temperature through differential melting on north and south facing slopes. In the second step, the wind drift factor was refined. To this end, a combined error map of cover for the eight patterns was calculated (Figure 6a). This map represents the average percent error in snow cover estimation and it was assumed that this error stemmed largely from the representation of snow drift. The parameters of the snow-drift model were then individually calibrated for each pixel to minimize the error. Figure 6(b) shows the error pattern for the calibrated model and illustrates the significant improvement over the pattern in Figure 6(a). Remaining error is due to sources other than wind drift, or that cannot be explained by the structure

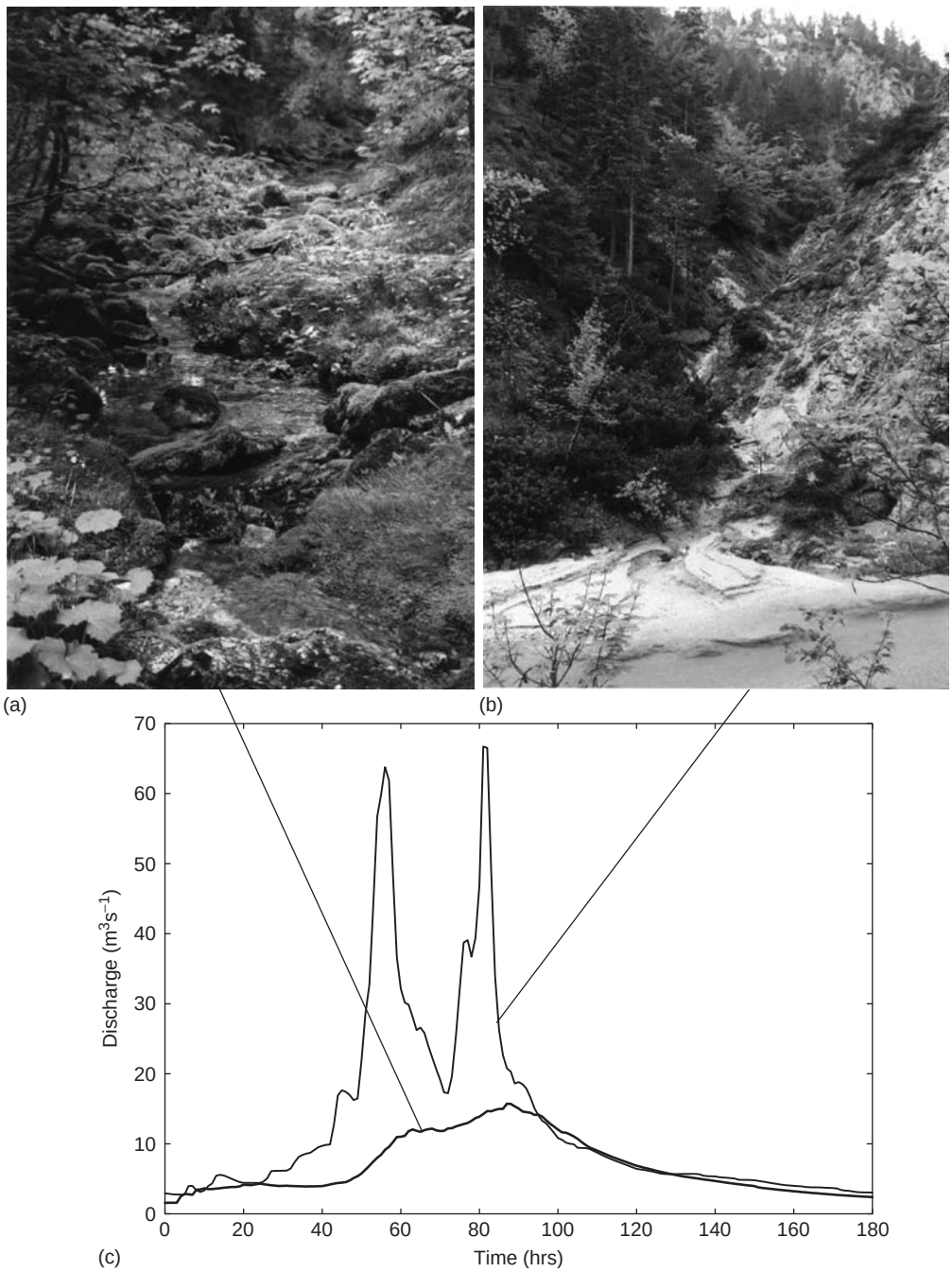


Figure 5 (a) Photographs that are representative of the landforms in the Mitterbach (a) and the Wienerbruck (b) catchments (Tables 2 and 3). (c) Hydrograph of the largest flood on record (thick line: Mitterbach; thin line: Wienerbruck). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of the wind drift representation. The model was then applied to a different snow season (2000) in a classical split sample test. Figure 6(c) shows the error map for this year and illustrates that, while there is some deterioration in simulation compared to the fully calibrated year, the revised snow-drift model is a major improvement over the original form (Figure 6a). While subsurface parameters of the rainfall-runoff model cannot be estimated from a comparison with observed snow patterns, the snow related parameters clearly can.

The second example is taken from Bauer (2004). His study is set in the Okavango Delta which is a large alluvial fan situated in northwestern Botswana. This wetland is fed by the Okavango River, which flows from the tropical

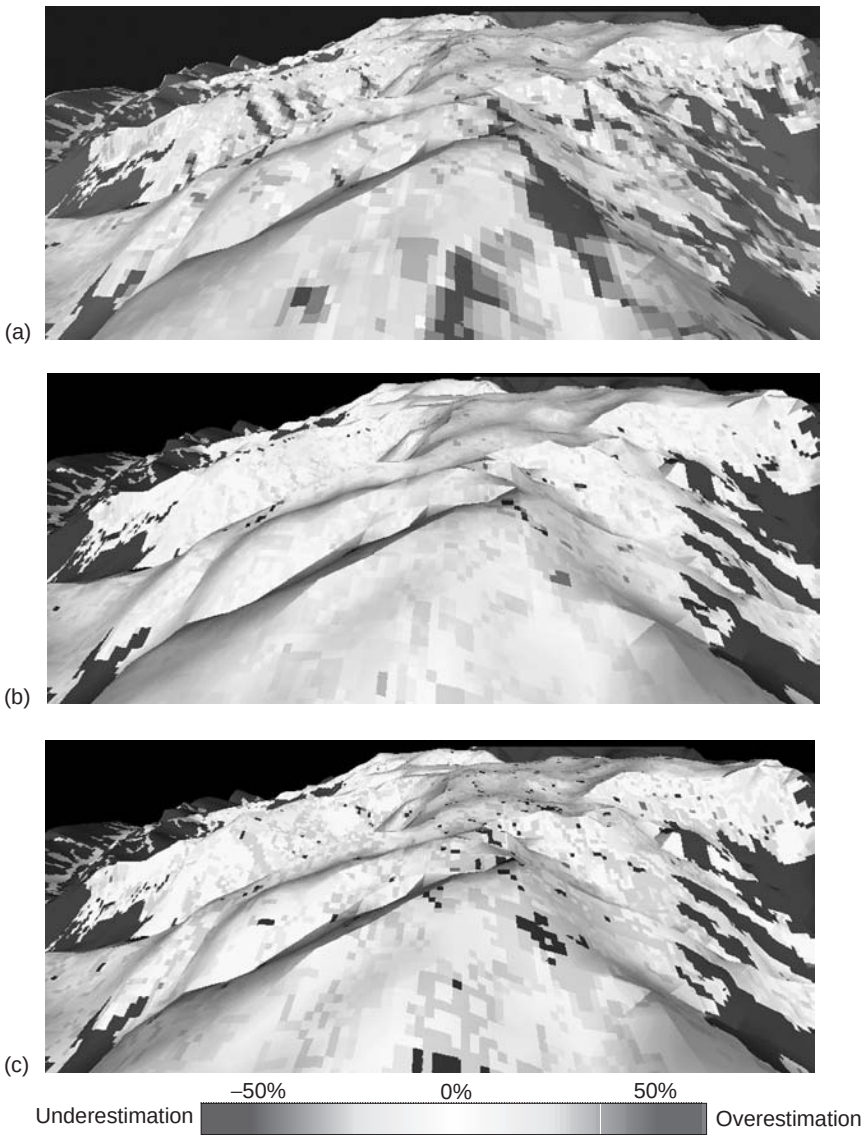


Figure 6 Bias in snow cover patterns based on residuals of comparisons between observed and simulated patterns for the Schneealpe region. (a) precalibration using eight observed patterns in 1998–1999, (b) postcalibration using eight observed patterns in 1998–1999, (c) validation using four observed patterns in 2000 (Blöschl *et al.* (2002). Reproduced with permission of Spinger-Verlag Wien). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

highlands of Angola into the Kalahari basin. The core wetland, which is permanently flooded, is around 6000 km^2 in size; the surrounding seasonal floodplains add another 6000 km^2. Runoff data would be extremely difficult to acquire as the flow occurs in numerous channels and on the marsh land. The catchment was therefore treated as ungauged and the model parameters were calibrated on alternative response data. The main interest in the study was on the runoff processes in the wetland, so inundation patterns were deemed to be suitable response data. The inundation patterns (either land or water) were derived from National Oceanic and Atmospheric Administration (NOAA), Advanced Very High Resolution Radiometer (AVHRR), and Landsat satellite images by unsupervised classification (McCarthy *et al.*, 2003). A total of 150 inundation patterns over the period of 1970–2000 was available for calibration. The hydrological model used was physically based and represented surface routing by Manning's equation, infiltration into the unsaturated and saturated zones by a simple leakage approach, two-dimensional regional groundwater flow, and evaporation from the saturated and unsaturated zones. The main calibration parameters were Manning's roughness of surface flow, the hydraulic conductivity of the swamp cells, the transmissivity of the sand

aquifer, and the vertical leakance between the surface and the subsurface layer (i.e. an infiltration parameter). Three objective functions were used in a manual calibration procedure and they all involved a comparison of simulated and observed inundation patterns for the same date. The first objective function was based on the total size of the flooded area, the second on a pixel-by-pixel comparison, and the third attempted to match the number of fringe points of the observed and simulated inundation patterns. Resulting parameters were all within a physically plausible range.

There are more alternatives to runoff that can be used for calibrating rainfall-runoff models in ungauged catchments. Suitable variables depend on the type of hydrological process of interest (Grayson *et al.*, 2002). Observed soil moisture would be an obvious choice for explicit soil moisture accounting models, but it is difficult to obtain spatially representative soil moisture measurements in catchments (Western *et al.*, 2002, 2004). Ground measurements may be representative of the entire root zone but are usually limited to a few spots in a catchment. Spaceborne estimates of soil moisture can be retrieved for large areas (e.g. Wagner *et al.*, 2003) but their main limitations are the shallow penetration depths, which are much smaller than the root depth represented in many hydrologic models. Although there are methods of dealing with this incompatibility in a simplified way (Houser *et al.*, 2000; Walker *et al.*, 2001; Schuurmans *et al.*, 2003), challenges remain. Various types of field based soft data, as discussed in Section "Soft data and qualitative field observations", are also potentially useful for assisting in model calibration as demonstrated in Seibert and McDonnell (2002). Soft data include saturation areas as mapped in the early work of Dunne and Black (1970) and Dunne *et al.* (1975). There is renewed interest in saturation patterns as illustrated by a number of mapping projects occurring (e.g. Kirnbauer *et al.*, 2005) and their application in constraining parameter estimation in rainfall-runoff modeling (e.g. Franks *et al.*, 1998).

CONCLUDING REMARKS

In ungauged catchments, no runoff data are available for calibrating model parameters, so alternative methods are needed. This article has reviewed numerous methods that depend on the type of parameters, whether they are calibration parameters associated with one particular model, or physically based parameters with some meaning beyond the model used. Calibration parameters are preferably transposed from similar, gauged catchments in the same region. Estimating calibration parameters from catchment attributes is generally not recommended for a number of reasons. Physically based parameters can be inferred from measurements, often based on widely available surrogates such as land use and soil type but, again, there are significant uncertainties and comparisons with adjusted parameters from similar, gauged catchments in the same region are extremely useful. As an alternative, or in addition to these parameter estimation methods, hydrologic response data other than runoff may be available in the catchment of interest and these may assist in calibrating model parameters.

Some guidance for selecting the model structure may also be needed in ungauged catchments. Similar to gauged catchments, model structure choice depends on the problem at hand, data availability, and the runoff processes one is to represent (see **Chapter 122, Rainfall-runoff Modeling: Introduction, Volume 3**). An analysis of similar, gauged catchments in the same region can provide some assistance in the latter step. Acquiring confidence in the model through model tests is not straightforward in the ungauged catchment case but pursuing more than one avenue of parameter estimation can be of assistance. For example, transposing parameters from similar gauged catchments, inferring them from measurements in the ungauged catchment, and testing the model against hydrologic response data that may be available in the ungauged catchment may provide some indications on model reliability through complimentary information. Cross-validation is also useful as it tests a number of uncertainties in a combined way. Additionally, some of the methods of uncertainty assessment reviewed in **Chapter 131, Model Calibration and Uncertainty Estimation, Volume 3** are applicable to the ungauged catchment case. In all these steps there is a very important role of qualitative field observations, if only by visual examination of the landforms. Qualitative field observations may greatly assist in judging the catchment functioning and hence in testing the plausibility of model parameters and the model structure.

It may seem strange to end a review of rainfall-runoff modeling in ungauged catchments with a note on the value of runoff data, but that in my opinion is the state of the science. As highlighted throughout this article, there are methods for dealing with rainfall-runoff modeling in ungauged catchments, but significant uncertainties remain. Short-runoff data series may contain very valuable information (Vogel and Kroll, 1991; Laaha and Blöschl, 2005b) so the single best recommendation on the issue of rainfall-runoff modeling of ungauged catchments may be to install a stream gauge!

Acknowledgment

I am grateful to Ralf Merz for providing valuable comments on an early draft of this manuscript and to Keith Beven and an anonymous reviewer for comments that helped improve the manuscript. I would also like to thank Dieter

Gutknecht, Rodger Grayson, and Siva Sivapalan for the numerous insightful discussions on the subject matter over the years. Financial assistance from the Austrian Academy of Sciences and the Austrian Science Foundation (project no P14478-TEC) is gratefully acknowledged.

FURTHER READING

Beven K.J. (2001b) How far can we go with distributed hydrological modelling? *Hydrology and Earth Systems Sciences*, **5**(1), 1–12.

REFERENCES

Aryal S.K., O'Loughlin E.M. and Mein R.G. (2002) A similarity approach to predict landscape saturation in catchments. *Water Resources Research*, **38**(10), 1208, doi:10.1029/2001WR000864.

Atkinson S.E., Woods R.A. and Sivapalan M. (2002) Climate and landscape controls on water balance model complexity over changing timescales. *Water Resources Research*, **38**(12), 1314, doi:10.1029/2002WR001487.

Bates B.C. (1994) Regionalisation of hydrologic data: a review. *Cooperative Research Centre for Catchment Hydrology*, Monash University: Victoria, p. 61.

Bauer P. (2004) *Flooding and Salt Transport in the Okavango Delta, Botswana: Key Issues for Sustainable Wetland Management*, Diss., Naturwissenschaften, Eidgenössische Technische Hochschule ETH Zürich, Nr. 15436.

Beldring S., Roald L.A. and Voksø A. (2002) *Avrenningskart for Norge (Runoff map for Norway, in Norwegian)*, Report No. 2, Norwegian Water and Energy Directorate, Oslo.

Bergström S. (1976) *Development and Application of a Conceptual Runoff Model for Scandinavian Catchments*, Department of Water Resources Engineering, Lund Institute of Technology, Bulletin Series A-52, Swedish Meteorological and Hydrological Institute: Norrköping, p. 134.

Beven K. (1989) Changing ideas in hydrology – the case of physically based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (2000) Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth Systems Sciences*, **4**(2), 203–213.

Beven K.J. (2001a) *Rainfall-runoff Modelling – The Primer*, John Wiley and Sons, p. 360.

Beven K.J. and Kirkby M.J. (1979) A physically-based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Bierkens M.F.P., Finke P.A. and de Willigen P. (2000) *Upscaling and Downscaling Methods for Environmental Research*, Kluwer Academic Publishers: p. 190.

Blazkova S., Beven K., Tacheci P. and Kulasova A. (2002) Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): the death of TOPMODEL? *Water Resources Research*, **38**(11), 10.1029/2001WR000912.

Blöschl G. (1999) Scaling issues in snow hydrology. *Hydrological Processes*, **13**, 2149–2175.

Blöschl G. and Grayson R. (2000) Spatial observations and interpolation. In *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Chap. 2, Grayson R. and Blöschl G. (Eds.), Cambridge University Press: Cambridge, pp. 17–50.

Blöschl G., Kirnbauer R., Jansa J., Kraus K., Kuschnig G., Gutknecht D. and Reszler Ch. (2002) Einsatz von Fernerkundungsmethoden zur Eichung und Verifikation eines flächendetaillierten Schneemodells (Using remote sensing methods for calibrating and verifying a spatially distributed snow model). *Österreichische Wasser- und Abfallwirtschaft*, **54**, 1–16.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling – a review. *Hydrological Processes*, **9**, 251–290.

Boorman D.B., Hollis J.M. and Lilly A. (1995) *Hydrology of Soil Types: A Hydrologically Based Classification of the Soils of the United Kingdom*, IH Report No. 126, Institute of Hydrology, Wallingford.

Bormann H. Diekkrüger B. and Renschler C. (1999) Regionalisation concept for hydrological modelling on different scales using a physically based model: results and evaluation, *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, **24**(7), 799–804.

Breiman L., Friedman J.H., Olshen R. and Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth International Group: Belmont.

Brunsell N.A. and Gillies R.R. (2003) Scale issues in land-atmosphere interactions: implications for remote sensing of the surface energy balance. *Agricultural and Forest Meteorology*, **117**, 203–221.

Budyko M.I. (1974) *Climate and Life*, Academic Press: Orlando, p. 508.

Burn D.H. and Boorman D.B. (1993) Estimation of hydrological parameters at ungauged catchments. *Journal of Hydrology*, **143**, 429–454.

Busch G., Sutmöller J., Krüger J.-P. and Gerold G. (1999) Regionalization of runoff formation by aggregation of hydrological response units: a regional comparison. In *Regionalization in Hydrology*, Diekkrüger B., Kirkby M.J. and Schröder U. (Ed.), IAHS Publication No. 254, IAHS Press: pp. 45–53.

Büttner G., Feranec J. and Jaffrain G. (2002) *Corine Land Cover Update 2000: Technical Guidelines*, Technical Report No. 89, European Environment agency, Copenhagen.

Calver A., Kay A.L., Jones D.A., Kjeldsen T., Reynard N.S. and Crooks S. (2004) Flood frequency quantification for ungauged sites using continuous simulation: a UK approach. In *Complexity and Integrated Resources Management, Transactions of the 2nd Biennial iEMSs Meeting*, Pahl-Wostl C., Schmidt S., Rizzoli A.E. and Jakeman A.J. (Eds.), International Environmental Modelling and Software Society (iEMSs): Manno, pp. 1214–1218.

Campbell E. and Bates B. (2001) Regionalization of rainfall-runoff model parameters using Markov Chain Monte Carlo samples. *Water Resources Research*, **37**(3), 731–739.

Cordery I. and Pilgrim D.H. (1983) *On the Lack of Dependence of Losses from Flood Runoff on Soil and Cover Characteristics*, *Proceedings of the Hamburg Symposium*, IAHS Publication No. 140, IAHS: Wallingford, pp. 187–195.

Drogue G., El Idrissi A., Pfister L., Leviandier T., Iffly J.-F. and Hoffmann L. (2002) Calibration of a parsimonious rainfall-runoff model: a sensitivity analysis from local to regional scale. *Proceedings of iEMSs 2002 Integrated Assessment and Decision Support 24–27 June 2002 Lugano*, Vol. 1, International Environmental Modelling and Software Society: pp. 464–469.

Duan Q., Schaake J. and Koren V. (2001) A priori estimation of land surface model parameters. In *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling. Water Science and Application*, Vol. 7, American Geophysical Union: pp. 77–94.

Dunne T. (1978) Field studies of hillslope flow processes. In *Hillslope Hydrology*, Kirkby M.J. (Ed.), Wiley: Chichester, pp. 227–293.

Dunne T. and Black R.D. (1970) Partial area contributions to storm runoff in a small New England watershed. *Water Resources Research*, **6**(5), 1296–1311.

Dunne T., Moore T.R. and Taylor C.H. (1975) Recognition and prediction of runoff-producing zones in humid regions. *Hydrological Sciences Bulletin*, **20**, 305–327.

Engeland K. and Gottschalk L. (2002) Bayesian estimation of parameters in a regional hydrological model. *Hydrological Earth System Sciences*, **6**(5), 883–898.

Engman E.T. (1986) Roughness coefficients for routing surface runoff. *Journal of Irrigation and Drainage Engineering*, **112**, 39–53.

Farmer D., Sivapalan M. and Jothityangkoon C. (2003) Climate, soil and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: downward approach to water balance analysis. *Water Resources Research*, **39**(2), 1035, SWC 1–21.

Fernandez W., Vogel R.M. and Sankarasubramanian A. (2000) Regional calibration of a watershed model. *Hydrological Sciences Journal*, **45**(5), 689–708.

Flügel W.A. (1995) Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. *Hydrological Processes*, **9**, 423–436.

Franks S.W., Gineste P., Beven K.J. and Merot P.h (1998) On constraining the predictions of a distributed model: the incorporation of fuzzy estimates of saturated areas into the calibration process. *Water Resources Research*, **34**(4), 787–797.

Gallo K., Tarpley D., Mitchell K., Csiszar I., Owen T. and Reed B. (2001) Monthly fractional green vegetation cover associated with land cover classes of the conterminous USA. *Geophysical Research Letters*, **28**(10), 2089–2092.

Gessler P.E., Moore I.D., McKenzie N.J. and Ryan P.J. (1995) Soil landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information Systems*, **9**(4), 421–432.

Gorte B.G.H. (2000) Land-use and catchment characteristics. In *Remote Sensing in Hydrology and Water Management*, Schultz G.A. and Engman E.T. (Eds.), Springer: Berlin, pp. 133–156.

Gottschalk L. (2002) Advances in observational hydrology – field experiments and modelling. In *Proceedings of Workshop on the Prediction of Ungauged Basins (PUBs) held at 28–29 March 2002 at the Yamanashi University*, Takeuchi K. (Ed.), International Association of Hydrological Sciences: Kofu.

Grayson R., Blöschl G., Western A. and McMahon T. (2002) Advances in the use of observed spatial patterns of catchment hydrological response. *Advances in Water Resources*, **25**, 1313–1334.

Gupta H.V., Sorooshian S. and Yapo P.O. (1998) Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research*, **34**(4), 751–763.

Hall F.G., Townshend J.R. and Engman E.T. (1995) Status of remote sensing algorithms for estimation of land surface state parameters. *Remote Sensing of Environment*, **51**(1), 138–156.

Hansen M., DeFries R., Townshend J.R.G. and Sohlberg R. (2000) Global land cover classification at 1 km resolution using a decision tree classifier. *International Journal of Remote Sensing*, **21**, 1331–1365.

Hessel R., Jetten V. and Guanghui Z. (2003) Estimating Manning's n for steep slopes. *Catena*, **54**(1–2), 77–91.

Hirsch R.M., Helsel D.R., Cohn T.A. and Gilroy E.J. (1992) Statistical analysis of hydrological data. In *Handbook of Hydrology*, Maidment R. (Ed.), McGraw-Hill: New York, pp. 17.1–17.55.

Hoosbeek M.R. and Bryant R.B. (1993) Towards the quantitative modeling of pedogenesis: a review. *Geoderma*, **55**, 183–210.

Houser P.R., Goodrich D. and Syed K. (2000) Runoff, precipitation, and soil moisture at Walnut Gulch. In *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Chap. 6, Grayson R. and Blöschl G. (Eds.), Cambridge University Press: Cambridge, pp. 125–157.

Institute of Hydrology (IH) (1999) *Flood Estimation Handbook*. Institute of Hydrology: Wallingford.

Jamagne M., Deroussin J., Eimberck M., King D., Lambert J.J., Lebas C. and Montanarella L. (2002) Soil geographical data base of Eurasia and Mediterranean countries at 1:1m, Transcations of the 17th World Congress of Soil Science, *The International Union of Soil Sciences, Symposium 44*, Bangkok, paper no 494, 14–20 August.

Jensen J.R. (1996) *Introductory Digital Image Processing. A Remote Sensing Perspective*, *Second Edition*, Prentice-Hall.

Kirnbauer R., Blöschl G., Haas P., Müller G. and Merz B. (2005) Identifying space-time patterns of runoff generation – A case study from the Löhnersbach catchment, Austrian Alps. In *Global Change and Mountain Regions*, Huber U., Reasoner M. and Bugmann H. (Eds.), Springer Wien: New York, pp. 309–320.

Kite G. and Droogers P. (Eds.) (2000) Comparing actual evapotranspiration from satellite data, hydrological models and field data. Special Issue *Journal of Hydrology*, **229**(1–2), 1–100.

Koren V., Smith M. and Duan Q. (2003) Use of a priori parameter estimates in the derivation of spatially consistent parameters sets of rainfall-runoff modes. In *Calibration of Watershed Models. Water Science and Application*, Vol. 6, American Geophysical Union: pp. 239–254.

Laaha G. and Blöschl G. (2005a) A comparison of low flow regionalisation methods – catchment grouping. *Journal of Hydrology*, in press.

Laaha G. and Blöschl G. (2005b) Low flow estimates from short stream flow records – a comparison of methods. *Journal of Hydrology*, **306**(1–4), 264–286.

Lamb R., Beven K.J. and Myrabø S. (1998) Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. *Advances in Water Resources*, **22**(4), 305–317.

Lamb R. and Kay A.L. (2004) Confidence intervals for a spatially generalized, continuous simulation flood frequency model for Great Britain. *Water Resources Research*, **40**, W07501, doi:10.1029/2003WR002428.

Larsen J.E., Sivapalan M., Coles N.A. and Linnet P.E. (1994) Similarity analysis of runoff generation processes in real-world catchments. *Water Resources Research*, **30**(6), 1641–1652.

Leavesley G.H., and Stannard L.G. (1995) The Precipitation-Runoff Modeling System – PRMS. In *Computer Models of Watershed Hydrology*, Singh V.P. (Ed.), Water Resources Publications, Highlands Ranch, pp. 281–310.

Lucas Y. (2001) The role of plants in controlling rates and products of weathering. *Annual Review of Earth and Planetary Sciences*, **29**, 135–163.

McCarthy J.M., Gumbricht Th.R., McCarthy T.S., Frost Ph.E., Wessels K. and Seidel F. (2003) Flooding patterns of the Okavango Wetland in Botswana between 1972 and 2000. *AMBIO: A Journal of the Human Environment*, **32**(7), 453–457.

Merz R. and Blöschl G. (2004) Regionalisation of catchment model parameters. *Journal of Hydrology*, **287**, 95–123.

Merz R. and Blöschl G. (2005) Flood frequency regionalisation – spatial proximity vs. catchment attributes. *Journal of Hydrology*, **302**(1–4), 283–306.

Merz R., Blöschl G. and Parajka J. (2005) Raum-zeitliche Variabilität von Ereignisabflussbeiwerten in Österreich (Spatio-temporal variability of event runoff coefficients in Austria). *Hydrologie und Wasserbewirtschaftung*, in press.

Merz R., Parajka J. and Blöschl G. (2004) Räumliche Muster in der konzeptionellen Wasserbilanzmodellierung – Parameteridentifikation. *Proceedings Tag der Hydrologie in Potsdam*, Universität Potsdam.

Merz R., Piock-Ellena U., Blöschl G. and Gutknecht D. (1999) Seasonality of flood processes in Austria. In *Hydrological Extremes: Understanding, Predicting, Mitigating*, Gottschalk L., Olivry J.C., Reed D. and Rosbjerg D. (Eds.), *Proceedings of the Birmingham Symposium*, IAHS Publication No. 255, IAHS: pp. 273–278, July 1990.

Mishra S.K. and Singh V.P. (2003) *Soil Conservation Service Curve Number (SCS-CN) Methodology*, *Book Series: Water Science And Technology Library, Volume 42* , Kluwer Academic Publishers.

Moore I.D., Grayson R.B. and Ladson A.R. (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5**, 3–30.

Mroczkowski M., Raper G.P. and Kuczera G. (1997) The quest for more powerful validation of conceptual catchment models. *Water Resources Research*, **33**(10), 2325–2335.

Nash I.E. and Sutcliffe I.V. (1970) River flow forecasting through conceptual models, part I. *Journal of Hydrology*, **10**, 282–290.

Nathan R.J. and McMahon T.A. (1990) Identification of homogeneous regions for the purpose of regionalization. *Journal of Hydrology*, **121**, 217–238.

Parajka J., Merz R. and Blöschl G. (2005a) A comparison of regionalisation methods for catchment model parameters. *Hydrology and Earth Systems Sciences Discuss.*, **2**, 509–542. **www.copernicus.org/EGU/hess/hessd/2/509/ SRef-ID: 1812-2116/hessd/2005-2-509**.

Parajka J., Merz R. and Blöschl G. (2005b) Regionale Wasserbilanzkomponenten für Österreich auf Tagesbasis (Regional water balance components in Austria on a daily basis). *Österreichische Wasser- und Abfallwirtschaft*, **57**(3/4), 43–56.

Peschke G., Etzenberg C., Töpfer J., Zimmermann S. and Müller G. (1999a) Runoff generation regionalization: analysis and a possible approach to a solution. In *Regionalization in Hydrology*, Diekkrüger B., Kirkby M.J. and Schröder U. (Eds.), IAHS Publication No. 254, IAHS Press: pp. 147–156.

Peschke G., Etzenberg C., Müller G., Töpfer J. und Zimmermann S. (1999b) *Das wissensbasierte System FLAB – ein Instrument zur rechnergestützten Bestimmung von Landschaftseinheiten mit gleicher Abflußbildung*, IHI-Schriften, H.10, Internat. Hochschulinstitut Zittau.

Rinaldo A., Vogel G.K., Rigon R. and Rodriguez-Iturbe I. (1995) Can one gauge the shape of a basin? *Water Resources Research*, **31**, 1119–1128.

Robinson J.S. and Sivapalan M. (1995) Catchment-scale model of runoff generation by aggregation and similarity analysis. *Hydrological Processes*, **9**(5–6), 555–574.

Rodríguez-Iturbe I. and Valdés J.B. (1979) The geomorphologic structure of hydrologic response. *Water Resources Research*, **15**, 1409–1420.

Scherrer S. and Naef F. (2003) The identification of dominant runoff processes on plot-scale based on field information used to define hydrologic response units on the catchment-scale. *Geophysical Research Abstracts*, **5**, 14 393.

Schumann A.H., Funke R. and Schultz G.A. (2000) Applications of a Geographic Information System for conceptual rainfall-runoff modeling. *Journal of Hydrology*, **240**(1,2), 45–61.

Schuurmans J.M., Troch P.A., Veldhuizen A.A., Bastiaanssen W.G.M. and Bierkens M.F.P. (2003) Assimilation of remotely sensed latent heat flux in a distributed hydrological model. *Advances in Water Resources*, **26**, 151–159.

SCS – Soil Conservation Service (1973) *A Method for Estimating Volume and Rate of Runoff in Small Watersheds*, Technical paper 149, U.S. Department of Agriculture: Washington.

Sefton C.E.M. and Howarth S.M. (1998) Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales. *Journal of Hydrology*, **211**, 1–16.

Seibert J. (1999) Regionalisation of parameters for a conceptual rainfall-runoff model. *Agricultural and Forest Meteorology*, **98–99**, 279–293.

Seibert J. and McDonnell J.J. (2002) On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. *Water Resources Research*, **38**(11), 1241, doi:10.1029/2001WR000978.

Sepaskhah A.R. and Bondar H. (2002) Estimation of manning roughness coefficient for bare and vegetated furrow irrigation. *Biosystems Engineering*, **82**(3), 351–357.

Sivapalan M., Beven K.J. and Wood E.F. (1987) On hydrologic similarity. 2. A scaled model of storm runoff production. *Water Resources Research*, **23**(12), 2266–2278.

Stewart J.B., Engman E.T., Feddes R.A. and Kerr Y. (1996) *Scaling up in Hydrology Using Remote Sensing*, John Wiley: p. 255.

Szolgay J., Hlavková K., Kohnová S. and Kubeš R. (2002) Regional calibration of a water balance model for estimating mean monthly flow in small ungauged catchment, *ERB and Northern European FRIEND Project 5 Conference*, Demänovská dolina.

Tucker C.J., Grant D.M. and Dykstra J.D. (2004) NASA's global orthorectified landsat data set. *Photogrammetric Engineering and Remote Sensing*, **70**(3), 313–322.

United States Department of Agriculture (USDA) (1991) *State Soil Geographic Data Base (STATSGO) – Data User's Guide*, Miscellaneous Publication No. 1492, USDA-Soil Conservation Service: Washington, p. 88.

USACE (1994) *Engineering and Design – Flood-Runoff Analysis*, Publication Number EM 1110–2-1417, U.S. Army Corps of Engineers: Washington.

Vieux B.E. (2001) *Distributed Hydrologic Modeling Using GIS: Water Science And Technology Library*, Vol. 38, Kluwer Academic Publishers.

Vogel R.M. and Kroll C.N. (1991) The value of streamflow record augmentation procedures in low-flow and flood-flow frequency analysis. *Journal of Hydrology*, **125**(3–4), 259–276.

Wagner W., Scipal K., Pathe C., Gerten D., Lucht W. and Rudolf B. (2003) Evaluation of the agreement between the first global remotely sensed soil moisture data with model and precipitation data. *Journal of Geophysical Research-Atmospheres*, **108**(D19), 4611, doi: 10.1029/2003JD003663.

Walker J.P., Willgoose G.R. and Kalma J.D. (2001) One-dimensional soil moisture profile retrieval by assimilation of near-surface observations: a comparison of retrieval algorithms. *Advances in Water Resources*, **24**(6), 631–650.

Warrick A.W., Zhang R., Moody M.M. and Myers D.E. (1990) Kriging versus alternative interpolators: errors and sensitivity to model inputs. In *Field-scale Water and Solute Flux in Soils*, Roth K. etal (Eds.), Birkhäuser Verlag: Basel, pp. 157–164.

Western A.W. and Blöschl G. (1999) On the spatial scaling of soil moisture. *Journal of Hydrology*, **217**, 203–224.

Western A.W., Grayson R.B., Blöschl G., Willgoose G.R. and McMahon T.A. (1999) Observed spatial organisation of soil moisture and its relation to terrain indices. *Water Resources Research*, **35**(3), 797–810.

Western A.W., Grayson R.B. and Blöschl G. (2002) Scaling of soil moisture – a hydrological perspective. *Annual Review of Earth and Planetary Sciences*, **30**, 149–180.

Western A.W., Zhou S.-L., Grayson R.B., McMahon T.A., Blöschl G. and Wilson D.J. (2004) Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes. *Journal of Hydrology*, **286**(1–4), 113–134.

WMO (1995) *INFOHYDRO Manual, Second Edition*, Operational Hydrology Report, No. 28, WMO Report No. 683, World Meteorological Organisation, Geneva.

Woods R.A. (2003) The role of catchment classification in Hydrology science. *Geophysical Research Abstracts*, Vol. 5, 08 192, European Geophysical Society.

Wösten J.H.M., Pachepsky Y.A. and Rawls W.J. (2001) Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology*, **251**(3–4), 123–150.

134: Downward Approach to Hydrological Model Development

MURUGESU SIVAPALAN[1] AND PETER C YOUNG[2,3]

[1]Centre for Water Research, The University of Western Australia, Crawley, Australia
[2]Centre for Research on Environmental Systems and Statistics, Lancaster University, Lancaster, UK
[3]Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia

This article presents the top-down or downward approach to hydrological model development as an alternative to the bottom-up, reductionist approach, which is the current, dominant paradigm in the hydrological sciences. It discusses the philosophical underpinnings as well as the pros and cons of each approach, and illustrates the application of the downward approach through several examples. These examples variously emphasize three key elements in the application of the downward approach. First, the analysis of identifiable "signatures" or "features" in the available data including the definition of a generic model form that can accommodate these features; second, model development including model structure identification and parameter estimation; and last, model refinement or "fingering down" (Klemes, 1983; Jarvis, 1993), in which the model is improved by the use of additional information or by adding more causal factors. Two main approaches to downward modeling are identified and discussed: the classical stochastic systems approach and the deterministic, conceptual modeling approach. The advantages of the downward approach are highlighted, particularly as they relate to the development of models that have sufficient causal basis to make predictions in basins other than those on which the model was developed.

INTRODUCTION

Hydrologists are under considerable pressure to provide answers to questions concerning the effects of future climatic and land-use changes on the quantity and quality of streamflows. Understanding changes of hydrological processes in relation to human-induced climatic and landscape changes and the ability to predict these ahead of time are essential for sustainable catchment management. Hydrologists are, therefore, concerned with the development of hydrological models of various kinds for these purposes. These may include, amongst others: models of long-term water balance needed for integrated water resources management; models of stream water quality (i.e. sediments, nutrients, heavy metals, etc.) to manage the environmental health of receiving water bodies such as lakes, estuaries, and coastal oceans; and models of flood response for the purposes of flood warning and forecasting.

If a model is able to reflect the essence of how a catchment functions hydrologically (for the full range of its possible states), then it may be possible to extrapolate with some confidence beyond the observed conditions and come up with reliable predictions. The issue of obtaining adequate predictive power for future states is, therefore, tantamount to acquiring, and then utilizing, adequate understanding of how the catchment system functions over the full range of states. But just how do we obtain this understanding? Ideally, modeling approaches that are adopted for prediction must be based on a combination of *a priori* theoretical understanding of climate, soil, vegetation, and topography controls on catchment behavior, as well as an insightful interpretation of observed catchment response data. In other words, modelers must rely on accepted knowledge about

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

general catchment behavior, combined with information on actual ground conditions and data collected in real, *specific* catchments.

The current debate on the development of hydrological models that can be used for future predictions centers on the relative merits of two major modeling philosophies (Klemes, 1983; Jarvis, 1993; Sivapalan *et al.*, 2003a). First, the "upward" or "bottom-up" approach, the dominant paradigm at present, leads to models that are invariably based on small-scale theories of hydrologic responses. They often result in large, complex models because of the perceived need to capture, in full, all of the heterogeneities and process complexities known or observed at small space-time scales. In contrast, the "downward" or "top-down" approach concentrates on extracting parametrically efficient or "parsimonious" model forms that successfully reproduce key "signatures" or "features" of observed hydrological variability; or can describe dominant modes of dynamic system behavior that are "identifiable" from available input–output and other (e.g. internal) response data. The objective of this article is to present an overview of the latter "top-down", or "downward approach" to hydrological modeling; to set it apart from the currently more popular "bottom-up" or "upward approach"; to illustrate its application using a number of relevant examples; and, finally, to discuss its relevance to current hydrologic research and practice.

BOTTOM-UP OR UPWARD APPROACH TO MODEL DEVELOPMENT

The defining feature of the bottom-up or upward approach is the attempt to predict the overall catchment response on the basis of understanding and process knowledge acquired on smaller spatial and temporal scales. The consequent "up-scaling" is through an integration or aggregation of the smaller-scale process descriptions up to the catchment scale. In hydrology, the upward approach has been championed by Freeze and Harlan (1969) in their "Blueprint for a physically based, digitally simulated hydrologic response model" and, since then, many other modeling studies have followed this dominant paradigm (see e.g. Abbott and Refsgaard, 1996).

The resulting models are often labeled as "process-based" or "physically based", with the approach to such model development being known variously as *upward, bottom-up, reductionist*, and so on. Klemes (1983) defines the "upward approach" as "the route that attempts to combine, by mathematical synthesis, the empirical facts and theoretical knowledge available at a lower level of scale, into theories capable of predicting events to be expected at a higher, in our case hydrological, level". One can gain further insights into this approach from a description offered by Jarvis (1993), who defines "bottom-up" models as "deterministic, state-of-the-art, mechanistic, process-based models that provide a statement of how a system may function on its spatial and temporal scale, on the basis of knowledge acquired on smaller spatial and shorter temporal scales".

Development of these distributed, process-based or physically based models over the past 25 years has been fueled by the explosion in knowledge about various hydrological processes and the development of theories and models that are able to describe them to a high level of sophistication. It has also been influenced by several other factors: the rapid growth in certain types of data availability (e.g. terrain attributes such as topography, soils, vegetation); the enormous increases in computational power; and the development of advanced statistical methods of data analysis and model parameter/state estimation.

The common argument, indeed abiding hope, of developers of distributed, physically based models has always been that, if a model mirrors in detail how individual parts of a catchment function hydrologically, the resulting predictions at the catchment scale will be reliable, particularly for unobserved states. Unfortunately, this hope has been deceptive, as illustrated by Beven (1989) and Grayson *et al.* (1992), among others, as already anticipated by Freeze and Harlan (1969) themselves.

The main problem seems to be that processes important at one scale (either space or time or both) may not necessarily be important at other, larger or smaller, scales (Blöschl and Sivapalan, 1995). Indeed, because of the change of dominant processes with changing scales, it can be argued that, at large scales, not all the complexity embedded in small-scale models is actually necessary (Sivapalan, 2003). For example, soil heterogeneity may average out as we aggregate from the patch to the catchment scale. Conversely, as we go up in scale, processes not observed at the small scale may become important, such as large-scale, preferential flow paths in the subsurface. Similarly, in the time domain, the way one models an individual event may need to be different from the way one models a population of events occurring over a long period, in view of the multiplicity of timescales and processes, and the changes in the relative dominance of the variety of processes, over that longer period. For example, one frequently comes across such changes in dominant processes when developing models for extreme flood estimation (Jothityangkoon and Sivapalan, 2003). An *a priori* perception of what small-scale processes are important and how they interact, either in the temporal or spatial domain, can be grossly misleading unless underpinned by catchment response data over a broad range of time and space scales, or by some larger-scale process understanding.

Another, equally important, problem is excessive model complexity relative to data availability, and the associated

difficulties in identifying the chosen model structure and estimating its parameters. While in the 1970s and 1980s it was sometimes argued that the more complex models may be appropriate in situations with limited data availability because of their basis in physical representation of the processes (Abbott *et al.*, 1986), the opposite now seems to be true: the more complex a model, the more data are needed to calibrate it and to have confidence in using it in predictive mode (Grayson and Blöschl, 2000). Because of nonexistent or inadequate data for comprehensive model validation exercises, many models of this type tend to be overparameterized, with arbitrary and overly complex model structures.

From a systems perspective, model overparameterization is intimately connected to the question of model "identifiability". A defined model structure is only considered identifiable if its parameters can be estimated uniquely from a given set of data, without ambiguity. In this sense, overparameterized models have more parameters than are required to explain the data, and there is, therefore, a basic ambiguity in the model, leading to multiple model solutions. Because of the possibility of obtaining such multiple model solutions when using large models, Beven (2000) has introduced the term "equifinality" to describe this situation, based on an analogy with the use of this term in geomorphology. The implications of this on model identification and estimation are far reaching. For instance, Young (1978) has argued that, because large and complex models have enormous explanatory power, they can usually be fitted easily to the meager time-series data that often confronts the modeler: many of the "estimated" parameters tend to be ill-defined, and a smaller subset is often sufficient to explain or mimic the observed system behavior (see also, Jakeman and Hornberger, 1993). Not surprisingly, therefore, the converse is also possible: Young *et al.* (1996), for instance, show how a simple, low-order linear model, identified and estimated on the basis of data obtained from experiments on a very large, nonlinear simulation model, is able to mimic almost perfectly the response of the large simulation model.

TOP-DOWN OR DOWNWARD APPROACH TO MODEL BUILDING

As outlined in the previous section, there are considerable difficulties in the application of the bottom-up or reductionist modeling approach to the prediction of catchment behavior. In response to difficulties of this kind, there has been a paradigm shift in scientific thinking regarding descriptions of the natural world, away from the mechanistic or reductionist perspective to a more holistic or systems perspective (Capra, 1996). This shift is supported by experience from numerous examples in nature, including many from hydrology, where one finds apparently complex systems behaving in a simpler manner than might be expected, (e.g. Young, 1978, 1983, 2003; Young *et al.*, 1996; Sivapalan, 2003).

Whereas the bottom-up, reductionist approach emphasizes individual components or processes that make up the whole, the systems approach places more emphasis on whole system behavior, as inferred from appropriate data. It recognizes that the nature of the whole system is different from the sum of its parts, and that properties of individual processes or components are not intrinsic properties of the whole system, and so any study of the individual processes or components is carried out only from the point of view of understanding the whole system. Consequently, the focus is much more on the interactions, feedbacks, and functional relationships amongst various parts, and their relationship to the whole system behavior. In a hydrological context, a holistic approach to spatial processes focuses more on the patterns of connectivity between various parts of the catchment system, such as paddocks, hillslopes, soil catena, stream network structure, downstream hydraulic geometry, and so on, and less on the details of the responses of individual components. Similarly, in the temporal domain, a holistic approach will focus less on individual events, and more on the population of events, and the connectivity between these events over a long period of time through the combined climate-catchment system.

The downward approach, as discussed here, is a holistic approach but with a specific sharper focus. Klemes (1983) defines the downward approach as the "route that starts with trying to find a distinct conceptual node directly at the level of interest (or higher) and then looks for the steps that could have led to it from a lower level". The downward approach is an empirical or data-based approach involving the interpretation of, or learning about, a catchment's functioning from data obtained *at the catchment scale*. This is in contrast to the bottom-up approach, which relies on *a priori* available small-scale theories and/or process understanding, where the role of data is in the validation of model predictions. The defining feature of the downward approach is that any explanation and generalizations are achieved *by fingering down into the (smaller-scale) processes from above* (i.e. at the catchment scale), and this is why models developed in this way are also called *top-down* models (in the hydrological literature, see for example, the papers by Silvert, 1981; Jarvis, 1993).

In other words, starting from available data, the downward approach attempts to identify the processes directly at the scales of interest, and interprets them in terms of properties and processes that may be occurring at the same or finer scales, as required by the objectives of the modeling study. In this sense, the downward approach is different from the classical systems approach to hydrology pioneered by the work of Dooge (1973), for example. In his "Linear theory of hydrologic systems" Dooge (1973, pp. 5–6) states: "In systems analysis, we are concerned only with the way in which

the system converts input to output. If we can describe this system operation, we are not concerned in any way with the nature of the system – with the components of that system, their connection with one another, or with the physical laws which are involved". This is only one "black-box" interpretation of the systems approach – the downward approach to model building is a holistic or systems approach that goes further by emphasizing the need to explain the internal mechanisms of the identified input-output system in a physically meaningful manner.

METHODOLOGIES FOR APPLICATION OF THE DOWNWARD APPROACH

The defining feature of the downward approach to hydrological modeling is the attempt to predict the overall catchment response, as well as the catchment's functioning, based on interpretation of the observed responses directly at the catchment scale. Blöschl and Sivapalan (1995) present the most generally accepted sequence of steps that are involved in the development of hydrological models (see also Mackay and Riley, 1991; O'Connell, 1991): (i) collecting and analyzing data; (ii) developing a conceptual model that describes the important characteristics of a catchment; (iii) translating the conceptual model into a mathematical model; (iv) calibrating (optimizing, estimating) the model to fit a part of the historical data by adjusting the various coefficients in some manual or automatic manner; and (v) evaluating the model by predictive validation, that is, ensuring that it continues to predict the data satisfactorily on data other than that on which it was calibrated. In essence, this is a *hypothetico-deductive* approach to model building that follows from the ideas of the philosopher Popper (1959); see also Young (2002).

In contrast to the hypothetico-deductive approach, a top-down model is normally inferred from the data using an *inductive* process (see Young, 2002). This can be achieved in various ways. For instance, it can involve formal, data-based, statistical *identification* of an appropriate generic model structure (normally differential equations or their discrete-time equivalents, often, in the systems case, considered in stochastic terms), followed by *estimation* of the parameters that characterize this structure, and the *interpretation* of the identified model in physically meaningful terms. Or, less formally, it can involve simulation modeling (usually deterministic) accompanied by judicious selection of the model based on its ability to explain the data in the simplest manner. These procedures help avoid the imposition of too many preconceived ideas about the hydrologic mechanisms involved *prior* to modeling. Depending on the objectives of the model building exercise, however, the downward approach can proceed further – not satisfied with extracting a model structure and parameters from data, it can seek to refine the model by enabling it to use other information available at the same or lower level of scale.

Three main stages can be discerned in previous applications of the downward approach to the development of hydrological models. (Here we have tried to unify the procedures used in stochastic and deterministic top-down modeling into a common strategy that should be acceptable to the proponents of both strategies.)

Signature Analysis and Model Identification

The exploration of signatures in the available information is an attempt to quantify the functional transformation, or "filtering" that takes place during the catchment's response to climate inputs, and helps postulate an appropriate transfer function or conceptual model. Note that filtering is used here in a broad sense. This includes the more specific, systems meaning of the word, namely, the effect of passing a signal though a dynamic system (e.g. a real system or a model) such that it is filtered in frequency domain terms: that is, components at some frequencies are attenuated more than at others. An example is rainfall-flow, where the high frequency components in the rainfall are attenuated as the rainfall leads to changes in the flow of river water down the catchment: as a result, the measured flow becomes progressively smoother (and easier to forecast). Depending on the objectives of the modeling, the key aspects of the input-output transformation that we may be interested in can include, amongst others, time delay, attenuation of flood peaks, the dispersion of the input signal, and loss of mass (due to infiltration losses, evaporation, chemical transformations). In systems and statistical terms, the process of defining an appropriate model is termed "identification" and is normally seen as a method of defining not only a generic model form that is most appropriate but also the best defined dynamic model structure (e.g. dynamic order, time delay).

Model Development

Model Development refers to the *identification* of appropriate models on the basis of signatures extracted from the data, or the *choice* of appropriate conceptual models that can reproduce these signatures. This is followed by the *estimation* of model parameters, either from the data itself (using statistical or less formal methods), or from known catchment characteristics, depending on the type of model chosen. In describing models, we use the two terms, "transfer function models" and "conceptual models" somewhat interchangeably. Essentially, both refer to the mathematical or functional form of the input–output transformation. They differ in their origins in different scientific disciplines, in the way they are identified, and in the way they are parameterized.

The term "transfer function" (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**) represents the dynamic relationship between the input and output of a linear, dynamic system represented, in continuous-time terms, as the ratio of polynomials in the Laplace operator. (Note that in the case where initial conditions on variables are zero, the Laplace operator is equivalent to the derivative operator, so the Laplace transform transfer function is simply a useful shorthand representation of an input–output relationship described by a differential equation.) However, transfer function (TF from hereon) models can also be formulated in discrete-time terms and are then ratios of polynomials in the backward shift operator. TFs, expressed as Laplace transforms or as convolution integrals, have made a considerable impact in hydrology, as can be seen in the seminal work of Dooge (1973). The Nash cascade model (Nash, 1959), consisting of a number of identical, linear reservoirs in series, is a very well-known hydrological model and can be expressed in the form of a continuous or discrete-time TF model (Dooge, 1973; Dooge and O'Kane, 2003). There have also been extensions of such mathematical representations to include nonlinear transformations, through the work of Ven Te Chow and his coworkers in the 1960s, as described in detail in Chow (1964). More recently, Young (2000, 2001a) has extended TF representations to handle a wide class of "state-dependent parameter" nonlinear stochastic systems by allowing the parameters in the transfer function to be dependent on other variables or states in the system (see later example).

The attractive aspect of TF models is that, because of the direct connection between the signatures and model structure, the model structure can be *identified* by careful analysis of the hydrological data. However, there is no pretense that these can be automatically estimated (without further analysis) from observable catchment characteristics. Following the early work of Dooge and others, therefore, subsequent developments in hydrology tended to move away from formal representations of TFs and their estimation from data, in favor of so-called *conceptual models*. Conceptual models tend to be based on how the modeler perceives the nature of the catchment and its behavior, and not on the analysis of data. With the proliferation of conceptual models such as these, however, the moderating step of learning from data has tended to be forgotten and model structures are not often chosen on the basis of these data (i.e. signature or model identification analysis). Recent work on the downward approach by Sivapalan and his coworkers has partially addressed this problem and examples of this work are presented later. In these examples, because of the claimed hydrological basis of the model, the parameters of the chosen model structure may come directly from response data, as well as from observable catchment physical characteristics, which may help toward the "regionalization" (e.g. Kokkonen *et al.*, 2003) of at least some of the parameters to other, ungauged catchments.

Model Refinement (Fingering Down)

An important element of the downward approach, particularly when it is carried out within a deterministic, conceptual framework, is the emphasis placed on the use of model refinement based on additional information. This can relate directly to the input–output transformation and can involve various approaches, such as the establishment and estimation of causal links between the parameters of the identified or chosen model structure with catchment physical characteristics, or through the use of additional response information or surrogate data. This is the hardest part of the methodology and is done in an evolutionary manner. While no formal procedure exists, clearly the work on model refinement must be motivated by the success or otherwise of the model predictions as quantified, for example, by measures of predictive uncertainty.

EXAMPLES OF APPLICATION OF THE DOWNWARD APPROACH IN HYDROLOGY

The terms "top-down" or "downward" represent a general philosophical approach to model building that can be interpreted and used in a variety of different ways, often depending on the background of the hydrologist. On the basis of the three stages in the general methodology outlined previously, three main interpretations can be discerned in current hydrologic practice. First, a *stochastic approach* that exploits advanced statistical and systems methods of modeling to identify model structure and estimate parameters directly by data analysis (e.g. Young, 2001b, and the prior references therein; Schulz and Beven, 2003). Second, a *deterministic conceptual approach* that relies on considerable hydrological expertise in the choice of model structure on the basis of detailed analysis of available data, puts more emphasis on physically based model refinement, but places less emphasis on formal procedures for the extraction of model structure or parameters from the data directly. Because of its deterministic basis, at least some of the parameters in this approach to downward modeling can be estimated from the observable catchment characteristics (e.g. Wittenberg and Sivapalan, 1999; Jothityangkoon *et al.*, 2001). Finally, there is another *deterministic conceptual approach*, that is concerned with models that help in the understanding of interesting phenomena, rather than making predictions in the traditional sense: this is an approach that focuses much more on model refinement or "fingering down" (Milly, 1994; Rodríguez-Iturbe and Valdés, 1979; Jury *et al.*, 1982). Of course, the methodologies utilized in each of these cases are not mutually exclusive and there is hope (Young, 2003) that they can be combined

to form an integrated, top-down approach to hydrological modeling that gains through formalized procedures for data analysis, model development, parameterization, and model refinement. In this section, we discuss a number of different top-down modeling exercises that help exemplify the wide variety and applicability of this approach to model building.

The Aggregate Dead Zone Model: a Data-Based Mechanistic Model of Pollutant Transport in a River

This example and the one that will be discussed in the next subsection "A data-based mechanistic model of the rainfall-runoff process" are illustrations of *Data-Based Mechanistic* (DBM) modeling (see e.g. Young and Lees, 1993; Young, 1998 and the prior references therein). The systematic stochastic approach to the application of the downward approach within this DBM setting is illustrated well by the *Aggregate Dead Zone* (ADZ) model for the transport and dispersion of a dissolved pollutant in a defined reach of a river channel, with the flow rate denoted by Q. For simplicity, let us assume that Q is constant and there is "complete mixing" within an *Active Mixing Volume* (AMV: see Young and Lees, 1993) of defined volume V less than the volume of water in the reach and non-dispersive, purely advective (plug flow) elsewhere. Note that the mixing processes in the reach as a whole are normally imperfect, so that AMV is normally of considerably less volume than the total volume of water Vr in the reach at any time. The ratio V/Vr is called the dispersive fraction (Wallis *et al.*, 1989). The relationship between the measured concentration $C_o(t)$ of a pollutant at the output (or downstream) location in the reach, to the measured concentration $C_i(t)$ at the input (or upstream) location can then be obtained from dynamic mass-conservation considerations. Assuming, again for simplicity, that the loss of pollutant in the reach (i.e. both within the AMV *and* during advection) is proportional to the mass of pollutant in the reach, this mass balance equation takes the following ordinary differential equation form:

$$\underset{\substack{\text{rate of}\\ \text{change of}\\ \text{mass}}}{\frac{\mathrm{d}\{VC_o(t)\}}{\mathrm{d}t}} = \underset{\substack{\text{mass in}\\ \text{per unit}\\ \text{time}}}{Q\mathrm{e}^{-k\tau}C_i(t-\tau)} - \underset{\substack{\text{mass out}\\ \text{per unit}\\ \text{time}}}{QC_o(t)} - \underset{\substack{\text{mass lost}\\ \text{per unit}\\ \text{time}}}{k[VC_o(t)]} \tag{1a}$$

or,

$$\frac{\mathrm{d}C_o(t)}{\mathrm{d}t} = -\left(k + \frac{Q}{V}\right)C_o(t) + \frac{Q}{V}\mathrm{e}^{-k\tau}C_i(t-\tau) \tag{1b}$$

This simple description is the basis of the ADZ model (Beer and Young, 1983; Wallis *et al.*, 1989), and it can be rewritten in the form of the following first-order, stochastic TF model (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**):

$$C_o(t) = \frac{\beta_0}{s+\alpha_1}C_i(t-\tau)\ ;\ C_o^m(t) = C_o(t) + \xi(t) \tag{1c}$$

where $\beta_0 = (Q/V)\mathrm{e}^{-k\tau}$, $\alpha_1 = k + (Q/V)$ and s is the Laplace transform operator, interpreted loosely here as the derivative operator, that is, $s = \mathrm{d}/\mathrm{d}t$. In the above equations, k is the decay rate and τ is the advective time delay introduced to allow for the translational effects associated with the unidirectional river flow. Because we are assuming that the model has been obtained from the analysis of measured data, $C_o^m(t)$ in equation (1c) denotes the *measured* downstream concentration of the solute and $\xi(t)$ represents the combined effect of all unmeasurable stochastic inputs to the system, including measurement noise, unmeasured inputs and errors arising from any limitations in the assumed model structure ("modeling errors").

Often in TF modeling, the TF is represented in its discrete-time, sampled data form, since most data these days are sampled over time at some sampling interval (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**). In this case, the TF equation (1c) becomes:

$$x_k = \frac{b_o}{1+a_1z^{-1}}u_{k-\delta}: \quad x_k = C_o(k\Delta t); \quad y_k = x_k + \xi_k;$$
$$u_{k-\delta} = C_i(k\Delta t - \delta) \tag{1d}$$

where z^{-1} is the backward shift operator, that is, $z^{-r}y_t = y_{t-r}$, and the subscripted variables denote the value of the variable at the kth sampling instant. If the sampling interval is Δt (in appropriate units: hours, days, etc.), then $k = t/\Delta t$ is the sampling index, $k = 1, 2, \ldots, N$ and N is the total number of samples. Also, δ is the pure advective time delay in sampling intervals, defined as the nearest integral value of $\tau/\Delta t$. Note that this is normally an approximation to the time delay, which is the price paid for using the discrete-time rather than the continuous-time TF model. The discrete-time ("difference" or "finite difference" equation) model associated with equation (1d) is obtained by simple cross multiplication and application of the z^{-1} operator, to yield:

$$y_k = -a_1y_{k-1} + b_ou_{k-\delta}: \quad a_1 = \exp(-\alpha_1\Delta t);$$
$$b_o = \frac{\beta_o}{\alpha_1}\{1 - \exp(-\alpha_1\Delta t)\} \tag{1e}$$

The definitions of a_1 and b_0 given here are based on an assumption that the input concentration remains approximately constant over the sampling interval Δt. Equation (1e) shows that in these discrete-time terms, the downstream concentration of the solute at the kth sampling

instant is equal to a fraction ($-a_1$ where a_1 is negative and less that unity) of the concentration measured at the previous $(k-1)$th sampling instant, plus a fraction b_0 of the upstream concentration measured δ samples previously.

In the case of the first-order ADZ models such as (1a–e), there is little that can be deduced about the internal structure of the system other than the quantification of the overall advective and dispersive effects. However, suppose that analysis of the experimental data results in the statistical identification and estimation of an ADZ model in the form of the following second-order TF:

$$C_o(t) = \frac{\beta_0 s + \beta_1}{s^2 + \alpha_1 s + \alpha_2} C_i(t-\tau) \; ; \; C_o^m(t) = C_o(t) + \xi(t) \tag{2a}$$

or in discrete-time TF form:

$$x_k = \frac{b_0 + b_1 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}} u_{k-\delta} : \quad x_k = C_o(k\Delta t);$$
$$y_k = x_k + \xi_k; \quad u_{k-\delta} = C_i(k\Delta t - \delta) \tag{2b}$$

In most practical situations, the TF in equation (2a) can be factorized into the following parallel connection (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**) of two first-order TFs of the form (1b):

$$C_o(t) = \left(\frac{b_q}{s + a_q} + \frac{b_s}{s + a_s} \right) C_i(t-\tau) = \frac{G_q}{1 + T_q s} C_i(t-\tau) + \frac{G_s}{1 + T_s s} C_i(t-\tau) \tag{3}$$

Since, in practice, the residence times of the two parallel, first-order processes almost always are of quite different magnitudes, the subscripts q and s denote "quick" and "slow" processes, respectively. Of course, for any physical interpretation of the parameters in the parallel TFs (e.g. to compute V and k, given a measurement of Q), one must take account of the mass partitioning down the two parallel pathways. The parameters of the TF model (3) are the ADZ residence times or time constants, T_q and T_s, and the steady state gain, G_q and G_s, which account for any gain ($G_q + G_s > 1.0$) or loss ($G_q + G_s < 1.0$) of solute in the ADZ reach. The "travel time" in each parallel pathway is the sum of the associated residence time and the advective time delay τ.

A typical practical example of ADZ modeling is the analysis of data obtained from a dye tracing experiment on the River Conder, near Lancaster in northwest England (Young, 1999). This river is fairly small, with a "cobbled" bed, and the experiment involved the injection of 199 mg of the dye tracer Rhodamine WT, with the measurement locations situated 400 m apart, some way downstream of the injection location to allow for initial mixing. The measurements were taken at regular sampling interval of 0.25 min, and the river flow rate was measured at a constant 1.3 $m^3\,sec^{-1}$. Identification and estimation of an ADZ model from these data does not require any prior assumption about the order of the system and the advective time delay: these properties of the model are identified directly from the data. In this example, a continuous-time, second-order TF model, with a pure advective time delay of 3 min, is identified from the sampled data using the continuous-time *Simplified Refined Instrumental Variable* (SRIV) algorithm in the CAPTAIN Toolbox. This Toolbox is a collection of computational algorithms written for use in the Matlab software environment (see `http://www.es.lancs.ac.uk/cres/captain/`). Such a continuous-time TF estimation involves no approximations (see previous discussion) and seems more appropriate in this example. In any case, discrete-time TF estimation did not yield such good results.

The SRIV-estimated second-order model is of the same form as equation (2a) and it explains the downstream concentration changes very well indeed, as shown in Figure 1, with a *Coefficient of Determination* based on the simulated output (normally equivalent to *Nash-Sutcliffe Efficiency*) of $R_T^2 = 0.998$ (i.e. 99.8% of the measured output variance is explained by the model). Not surprisingly, this model can be decomposed into the parallel pathway form of equation (3) and the resulting ADZ models in the two pathways are defined as follows:

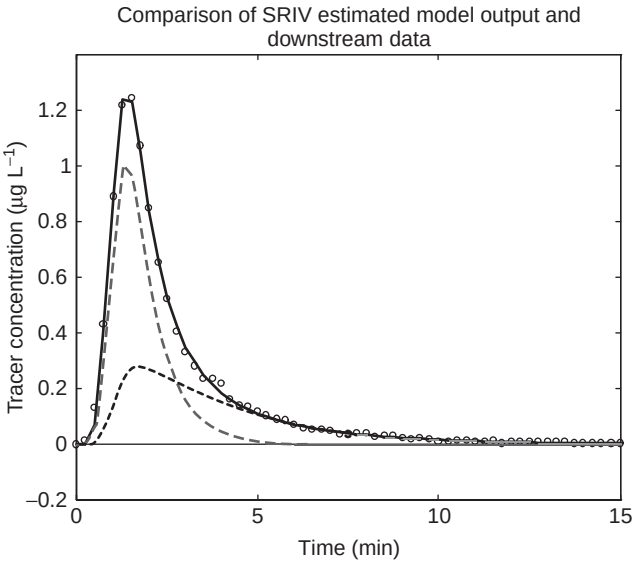


Figure 1 Comparison of the SRIV identified and estimated ADZ model output (full line) and the measured concentration of tracer at the downstream location (circular points). Also shown: inferred quick-flow pathway (dashed) and slow-flow pathway (dash-dot) concentrations

A. Slow Pathway: $x_s(t) = (0.475/1 + 2.818s)u(t-3)$ $G_s = 0.475$; $T_s = 2.818$ mins; $P_s = 42.9\%$
B. Quick Pathway: $x_q(t) = (0.608/1 + 0.590s)u(t-3)$ $G_q = 0.608$; $T_q = 0.590$ mins; $P_q = 57.1\%$

where *u(t)* is the upstream tracer concentration, while $x_s(t)$ and $x_q(t)$ are the inferred tracer concentrations at the output of the slow and quick pathways, respectively.

If it is assumed that the flow is partitioned in the same way as the dye, then the acting mixing volumes associated with the two pathways can be calculated by reference to the flow rate and residence times as $V_q = 26.3\,m^3$ and $V_s = 94.1\,m^3$, respectively. The associated dispersive fractions are then calculated as $DF_q = 0.12$ and $DF_s = 0.56$ (i.e. the acting mixing volumes are 12% and 56% of the total volume of water in each pathway, respectively). In other words, the slow pathway results in a considerably greater dispersion (and longer term detention) of the dye than the quick pathway, as one might expect.

The most obvious physical interpretation of the parallel flow decomposition in this example, as required in DBM modeling, is a form of "two-layer flow", with the slow pathway representing the dye in the water moving in, or adjacent to, the cobbled bed and banks of the river, which is being differentially delayed in relation to the quick pathway, which is associated with the more freely moving surface layers of water. The aggregated effect of each pathway is then an advective transportation delay of 3 min, associated with nondispersive plug flow, and an ADZ, defined by the associated active mixing volume and dispersive fraction in each case, which are the main mechanisms for dispersion of the dye in its passage down the river. Further details of this example, including full discussion of the stochastic aspects of the model, are given in Young (1999). Another ADZ modeling example is described in **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**.

In this ADZ case, model development and refinement are fairly straightforward. Given single reach ADZ models such as (1c), (2a), or the above model, it is possible to synthesize a water quality model for a complete river system from serial and parallel (and even feedback) interconnections of these elements. The nature of the interconnections in this situation will be defined by the physical characteristics of the river catchment. In such applications, the ADZ models obtained from conservative tracer experiments, such as that described previously, effectively provide a "calibration" of the *physical* aspects of solute transport and dispersion in the river. This information can improve the identifiability of other chemical and biological factors that may affect a nonconservative solute as it undergoes physical transport. For example, in the case of the simple decay process in equation (1a), it is clear that estimation of the decay rate k is facilitated by prior and independent estimation of the ADZ model parameters. And the advantages are even more pronounced in more complex situations where chemical and biological interactions are taking place (e.g. *Dissolved Oxygen-Biochemical Oxygen Demand* (DO-BOD) models, as in Beck and Young, 1975).

A Data-Based Mechanistic Model of the Rainfall-Flow Process

Much has been written on the important topic of rainfall-runoff and rainfall-flow modeling (e.g. Beven, 2001 and the prior references on this topic given therein). A typical top-down, DBM model of a rainfall-flow process is shown in the multiple storage element block diagram of Figure 2 (see Young, 2003). Here, the two linear storage elements or "buckets" A and B are first-order, linear TF models of a generic form similar to those used in the water quality model (1d) but, of course, with different physical interpretations. The block marked $f(s_k)$ is a nonlinearity, which is a function of an estimated "catchment soil water storage state" s_k. This storage state is generated from the measured rainfall r_k via the linear soil water storage TF that constitutes the storage element C; and it modulates r_k to yield the input u_k. The model in this form was identified and estimated on the basis of 360 samples of hourly rainfall and flow data from the River Hodder, at Hodder Place in northwest England, using the discrete-time *Refined Instrumental Variable* (RIV) algorithm in the CAPTAIN Toolbox mentioned above. The model was validated over a further, independent set of 120 hourly samples.

Full details of the DBM analysis that leads to the model in Figure 2 are given in Young (2003). In this example, the initial generic modeling stage is carried out in discrete-time terms (although it could equally well be based on a continuous-time model, as in the case of the ADZ model) and it identifies the following second-order, nonlinear TF model that consists of a nonlinearity in series with a linear second-order TF model (a "Hammerstein" model in systems terminology):

$$x_k = \frac{b_0 + b_1 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}} u_{k-\delta} \tag{4a}$$

$$u_k = f(s_k).r_k \tag{4b}$$

where,

$$\hat{a}_1 = -1.821(0.017);\ \hat{a}_2 = 0.822(0.016);$$
$$\hat{b}_0 = 0.13(0.011);\ \hat{b}_1 = -0.135(0.010)$$

In equation (4b), the nonlinearity $f(s_k)$ is estimated by considering the model in *State-dependent Parameter* (SDP) terms (see Young, 2000, 2001a,b). In particular, SDP estimation applied to the measured rainfall-flow time series (x_k, r_k) shows that, if the rainfall r_k is multiplied by a state-dependent parameter, identified from the data

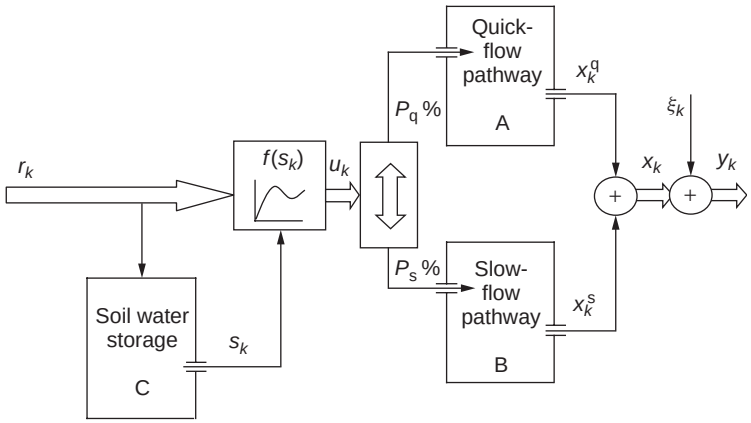


Figure 2 Block diagram of DBM rainfall-runoff model for the River Hodder at Hodder Place

as a power law in s_k (i.e. the nonlinearity $f(s_k) = s_k^\gamma$, where γ is the power law coefficient), then the complete model in equation (4) is able to predict very well the nonlinear flow variations x_k on the basis of the variations in the "effective rainfall" input $u_k = s_k^\gamma . r_k$ (see the following text). And because the introduction of this effective rainfall nonlinearity accounts for most of the nonlinear dynamics in the rainfall-flow data, the relationship between u_k and x_k can be represented quite simply by a linear, constant coefficient, discrete-time TF (the discrete-time equivalent of a linear differential equation).

As mentioned later in Section "Other examples of the downward approach", equation (4a) is the direct equivalent of the discrete-time convolution equation and the pulse response of this discrete-time TF (the response to a unit-volume pulse input of effective rainfall, u_k) can be interpreted in hydrological terms as the discrete-time *Unit Hydrograph* (UH) or, more strictly, the unit-volume pulse response (TUH) of the system. Note that this TUH is invariant in shape because the TF parameters $(\hat{a}_1, \hat{a}_2, \hat{b}_0, \hat{b}_1)$ that define this shape are all constant over time: the variation in the magnitude of the TUH that controls the changes in the nature of the output flow response is a consequence of the gain changes associated with the SDP, as it affects the effective rainfall nonlinearity $f(s_k)$. It is clear, therefore, that by estimating the rainfall-flow model in this manner, one is led naturally to the calibration of an underlying, *invariant* TUH that controls the shape of the storm events occurring over the whole data set.

As in the ADZ modeling example discussed in the previous subsection, the subsequent and very important hydrological interpretation stage of DBM modeling is relatively straightforward in this case. The nonlinearity is clearly associated with the "effective rainfall" dynamics: it accounts for the effect of antecedent rainfall on the soil water storage in the catchment; and this storage is modeled by the TF in block C. The second-order linear TF then defines the unit hydrograph response of the model and it nicely decomposes into a parallel connection of "quick" and "slow" flow pathways, similar to those discussed previously in relation to the ADZ model (3). Each of these is modeled as a first-order TF with estimated gains and residence times that define the quick "storm runoff" and slow "baseflow", which constitute the two major estimated components, or "dominant modes" (Young, 1999), of the *total* gauged flow.

The model in Figure 2 explains the data very well over the validation data set, as we see in Figure 3(a), which compares the model output (full line) with the measured flow in the validation part of the data set (dashed line). The *Coefficient of Determination* (see the preceding text) based on the simulated model output is $R_T^2 = 0.94$ (i.e. 94% of the measured flow variance is explained by the simulated model output); while the more normal coefficient of determination, based on the one-hour-ahead prediction errors is $R = 0.98$. In addition, the physical properties of the two pathways make good practical sense for the Hodder catchment. For instance, the parallel decomposition of the linear transfer function part of the model yields the following description of the two linear storage elements:

C. Slow Flow: $x_k^s = (0.0023/1 - 0.996z^{-1})u_{k-4}$ $G_s = 0.53$; $T_s = 230$h; $P_s = 41\%$
D. Quick Flow: $x_k^q = (0.134/1 - 0.825z^{-1})u_{k-4}$ $G_q = 0.77$; $T_s = 5.2$h; $P_q = 59\%$

where x_k^s and x_k^q are, respectively, the partitioned quick and slow-flow components; G denotes the steady state gain; T the residence time and P the partition percentage (i.e. the inferred percentage flow of water down the pathway in question), with the subscripts differentiating between the slow and quick pathways. The flow components generated by this parallel decomposition are shown in Figure 3(b) and (c). It is clear that the very long term baseflow behavior is being modeled by the slow-flow component (c), which has a very long 230-h residence time. The soil water storage TF also has physically meaningful characteristics, with a

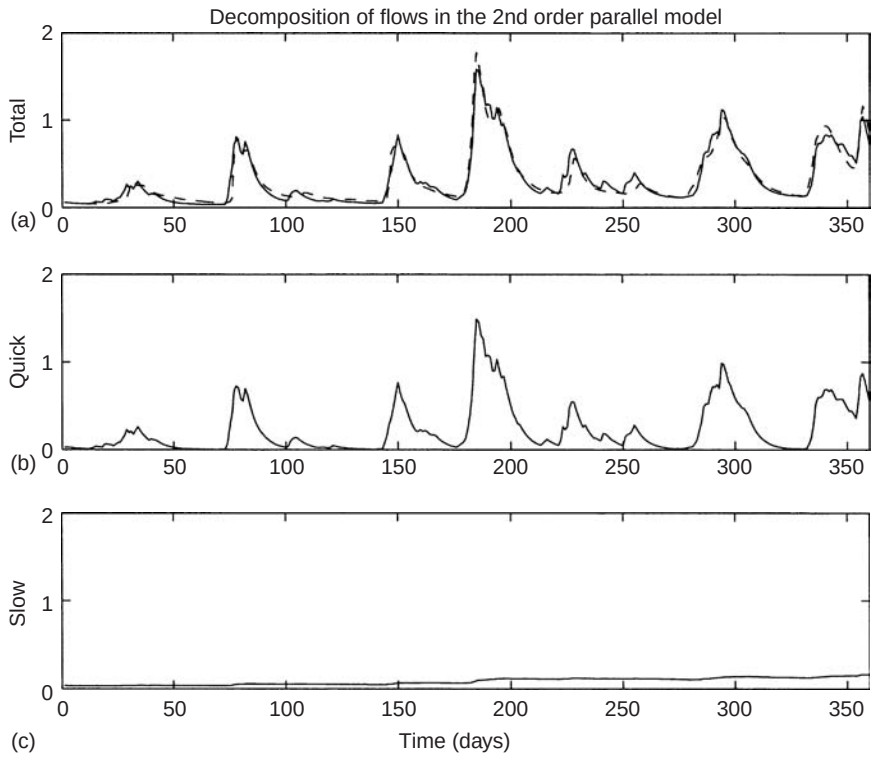


Figure 3 Parallel decomposition of the DBM rainfall-flow model for the River Hodder at Hodder Place, with all components plotted to the same scale for comparison

residence time of 9.1 h that is about twice the residence time of the quick-flow TF. And, finally, the nonlinearity acts as an effective rainfall coefficient that is zero when $s_t = 0$ and rises to a maximum of 0.87 at higher soil water storage level: that is, 87% of the gauged rainfall becomes effective rainfall when the storage is at this higher level (see Figure 8 in Young, 2003). Of course, since the model is stochastic, all of these parameters have associated uncertainties that are quantified in the estimation process (*see* **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**). However, in order to simplify the presentation of the results, they have not been reported here.

As in the previous example, model development and refinement are fairly straightforward because of the modular nature of the rainfall-flow model and the fact that a similar DBM/TF approach can be used for flow routing, that is, modeling the relationship between flow (or level) at adjacent gauge locations on the river (which could include man-made elements such as weirs). Depending on the application, the main river catchment will be modeled by suitably interconnected routing modules, while the rainfall-flow modules will provide inputs to this system. If information is required at a finer scale, for instance flows between gauge locations, then the routing modules can be replaced by a physically appropriate connection of smaller routing modules chosen so that the combined response mimics that of the single module (see the discussion concerning the combined response of serial and parallel processes in **Chapter 128, Rainfall-runoff modeling: Transfer Function Models, Volume 3**).

Water Balance Estimation Through Streamflow Recession Analysis and Baseflow Separation

Wittenberg (1999) and Wittenberg and Sivapalan (1999) present an example of the downward approach to water balance model development that uses an empirical, deterministic methodology, involving analyzing key signatures in the data, proposing simple models capable of reproducing these signatures, and then refining these systematically to capture more aspects of the variability in the observed data. Their work is a forerunner to the approaches adopted in much of the subsequent work in this area (see later) and offers hope that much more information can be extracted from data than can be obtained in the usual model calibration exercises. Wittenberg and Sivapalan present a method of analysis of observed rainfall and streamflow time series with a view to identifying and quantifying the underlying groundwater balance. In a recent effort, Wittenberg (2003) extends this work to include further analyses to decipher the effects of human impacts.

Through the analysis of numerous streamflow recession curves Wittenberg and Sivapalan obtained a nonlinear relationship between groundwater discharge and reservoir

storage in shallow unconfined aquifers. This analysis is similar to methods previously adopted by Brutsaert and Nieber (1977) and Troch *et al.* (1993). Depletion of groundwater by evapotranspiration losses through the uptake of trees was found to bias the recession curves and the estimated reservoir parameters. This biasing was found to be strongly seasonal, due to the reinforcing seasonality of both rainfall and potential evaporation. Analysis of the recession curves, stratified according to time of year, allowed the quantification of evapotranspiration loss as a function of calendar month and stored groundwater storage.

Using the nonlinear storage-discharge relationships, they then separated the subsurface flows from the observed time series of streamflows, and subsequently inferred the corresponding time series of storage and the rates of recharge to groundwater through a numerically stable, inverse routing procedure. Comparison of the computed groundwater recharge hydrographs with measured daily rainfalls revealed an approximately time-invariant travel time (to the water table) distribution of infiltrated water. Using traditional unit hydrograph methods, unit recharge responses to rainfall (i.e. unit hydrographs) were computed by least squares fitting. These estimated unit response functions were found to be comparable to those found in previous work based on theoretical approaches that used assumed soil properties, and others based on lysimeter data.

It can be seen that the work of Wittenberg and Sivapalan (1999) involves systematically analyzing rainfall and streamflow data, and nothing more; and then piecing together the components, such as surface and subsurface runoff, evapotranspiration, groundwater storage, and recharge. The method is very good at identifying the key transformations, and extracting the appropriate transfer functions based on the data, without utilizing anything more than the basic mass balance concept. Clearly, it is an empirical method and yet produces a simple model of the system and involves no subsequent calibration. Of course, the drawback of such a model is that, first, the methodology itself is rather *ad hoc* and can benefit from the use of more standardized methodologies. Second, in the absence of any inferred relationship between the mechanistic interpretation of the model and geographical or other physical aspects of the catchment ("regionalization analysis"), it is difficult to extrapolate to ungauged catchments.

Hierarchical Approach to Development of Conceptual Water Balance Models

Following on from the lead of Wittenberg and Sivapalan (1999), Sivapalan and a number of his coworkers (Jothityangkoon *et al.*, 2001; Atkinson *et al.*, 2002; Farmer *et al.*, 2003; Atkinson *et al.*, 2003a; Atkinson *et al.*, 2003b; Struthers *et al.*, 2003; Eder *et al.*, 2003) extended the deterministic approach further to look at water balance variability at the annual, monthly, daily, and hourly timescales, and through these to develop a more formalized procedure for the development of continuous water balance models of appropriate complexity. The methodology was applied to catchments in a variety of climatic and soil settings, in New Zealand, various parts of Australia, Austria, and Germany.

The similarity with the Wittenberg and Sivapalan (1999) approach is that the adopted methodology remains a hydrologically oriented, deterministic approach to the analysis of the observed data. However, the main difference is that the analysis goes beyond the analysis of individual storms to the comprehensive analysis of a population of storm events and is designed to make inferences about hydrological variability not only at the event scale, but also at longer, intra-annual (seasonal), and interannual variability.

In terms of transformations that one then looks for in the data, the approach is now considerably different since the focus is on a population of storms. Whereas the event response analysis looked for signs of "filtering" behavior at that scale in the traditional sense, the analysis of a population of storms, covering a range of scales of variability, looks for signs of such filtering at longer timescales including, for example, threshold filtering, whereby only those storms that exceed an intensity or depth threshold pass through the system, thus generating an output time series which is a truncated version of the input time series. In order to further the application of the downward approach, Sivapalan and his coworkers invoked a number of *discrete signatures* that characterize water balance variability at these timescales. These are, in sequential order:

- interannual (i.e. year to year) variability of annual runoff;
- mean monthly variation of runoff (also known as the *regime curve*);
- flow duration curve based on daily flows, describing the variability of streamflow within the year;
- streamflow recession curves, limited to deciphering the nature of subsurface flows within the catchment and as a measure of associated time delays.

The downward approach to model development then takes the view that these signatures of the transformation can be extracted from the observed data, and model development can then be pursued with the idea of incorporating the climate, soil, vegetation controls that underpin an explanation of these transformations. One continues with the idea that relatively simple models need to be found that explain and reproduce these signatures of variability.

A basic assumption in the above approach is of an evident hierarchy of controlling influences relating climate-landscape interactions to the prediction of hydrologic responses at various space-timescales of interest. Relatively simple models were often found to be adequate to capture the water balance at large time and space scales, provided they took into account the primary controlling

variables of precipitation, potential evaporation, and soil storage capacity.

When greater levels of prediction are required (i.e. at smaller spatial and/or temporal scales), then consideration must be given to the inclusion of additional variables and processes, in order to address the more subtle causes of variability. For example, when one changes from the annual to the monthly timescale, then the systematic time delay between precipitation and streamflows is an important feature that must be captured, and this is done through the incorporation of subsurface flow pathways. Such features that come into play at a smaller timescale are deemed as *emergent properties*, because they may not have been important to separately capture, if the objective was to reproduce interannual variability alone. Testing of hypotheses with respect to climate and landscape controls on these emergent behaviors are carried out in respect of the key signatures of variability at the monthly, daily, and hourly timescales. Typically, model complexity increases as timescale decreases, since it is increasingly important to introduce additional processes and parameterizations to reproduce these emergent behaviors. The various signatures (interannual variability, mean monthly flows, flow duration curve at the daily timescale, etc.) thus serve as guideposts in the development of models of increasing complexity.

Figure 4 presents an example of a hierarchy of models of increasing complexity explored for Australian catchments by Farmer *et al.* (2003), ranging from simple overflow or Manabe-type buckets, to multiple-bucket models that include both shallow subsurface flow and deep groundwater discharge, as well as explicit treatment of unsaturated zone time delays. An example of such an application of the downward approach to a specific catchment from Australia is presented in Figure 5, also from Farmer *et al.* (2003), while a summary of the levels of model complexity required to reproduce the signatures of water balance variability with changing timescales is presented in Figure 6, confirming an evident hierarchy with respect to timescale. Atkinson *et al.* (2002) found that the required model complexity increased with increasing aridity of the catchment, which is nicely illustrated by the schematic presented in Figure 7. This is also confirmed by the relative ease with which a simple parsimonious model was developed (Eder *et al.*, 2003) for a humid catchment in Austria.

Through the application of the downward approach, Jothityangkoon *et al.* (2001) developed a distributed model of water balance (based on subcatchments) for a large, heterogeneous catchment in semiarid Western Australia. This systematic procedure led to a model with just 6 parameters, most of which could be estimated *a priori* either from landscape properties or through prior analysis of observed streamflow records.

While conceptual models of the type developed by Sivapalan and his coworkers abound in the literature, their modeling approach differs from previous ones in a number of respects: (i) the appropriate model structure is obtained through systematic analysis of the data, (ii) the model is parsimonious for the data that are used in the model development, and (iii) many of the model parameters are estimated from the data. In addition, the systematic analysis of the data also enables the hydrologist/modeler to learn from the data as part of the process of model development.

While the signatures signify the transformations that the modeler attempts to reproduce, there is considerable flexibility in the choice of models that can be adopted to achieve these transformations. In the above examples, "bucket" models of varying degrees of complexity and nonlinearity are used, the parameters of which are estimated from the observed streamflow data, but other model types can likewise be employed. Indeed, Struthers *et al.* (2003) used the downward approach to model deep drainage underneath a large lysimeter in Germany, through the use of a nonlinear bucket model. They found that the bucket model did an excellent job in reproducing long-term (seasonal and multiannual) variability of recharge.

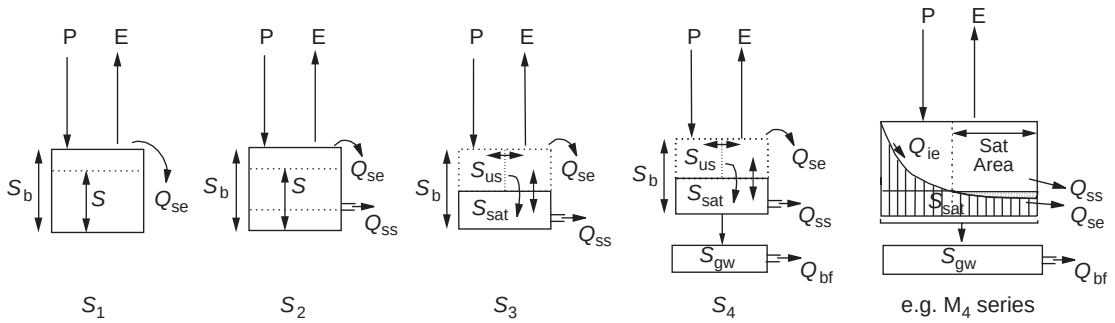


Figure 4 Schematic describing models of increasing complexity, ranging from single, overflow bucket model to multiple-bucket models of the VIC-type with a deep groundwater component (Reproduced from Farmer *et al.* (2003) by permission of American Geophysical Union)

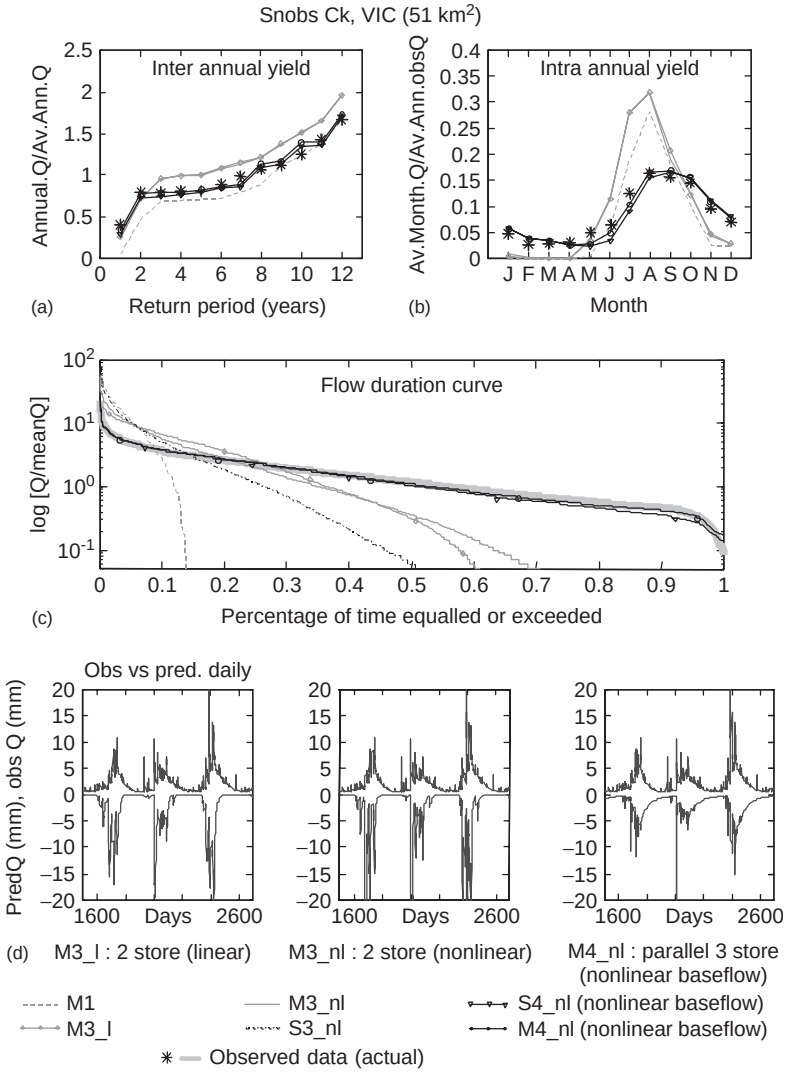


Figure 5 Systematic application of the downward approach to develop conceptual models of increasing complexity through ''fingering down''. (a) interannual variability of annual runoff, (b) mean monthly runoff, (c) flow duration curve, and (d) comparisons of the resulting hydrograph predictions against observed runoff (Reproduced from Farmer *et al.* (2003) by permission of American Geophysical Union) Refer to Figure 4 for explanations of notation of model type. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Other Examples of the Downward Approach

Mean Annual Water Balance and the Budyko Curve

It is well known that annual evapotranspiration approaches annual precipitation in regions where the water equivalent of the energy supply greatly exceeds precipitation (arid regions), whereas in regions where the water equivalent of energy supply is only a small fraction of precipitation (humid regions), annual evapotranspiration is commensurate with the energy supply. On the basis of these considerations, and water balance data obtained from a number of catchments around the world, Budyko (1974) derived a simple relationship describing annual evapotranspiration as a function of annual energy supply (i.e. net radiation) and precipitation. The resulting *Budyko curve*, as the universal relationship is known, represents a simple pattern at the annual timescale and watershed (space) scale, a pattern that is often hidden behind considerable scatter in the observed data but is, nevertheless, extremely critical and not yet fully understood.

Milly (1994) presented a simple mathematical framework to explore the climate, soil, vegetation controls on the shape of the Budyko curve, which is an excellent example of the application of the "downward approach". Milly hypothesized that long-term water balance is determined by interactions of fluctuating water supply and demand, mediated by water storage in the soil. He showed that a simple bucket model based on these considerations was able to

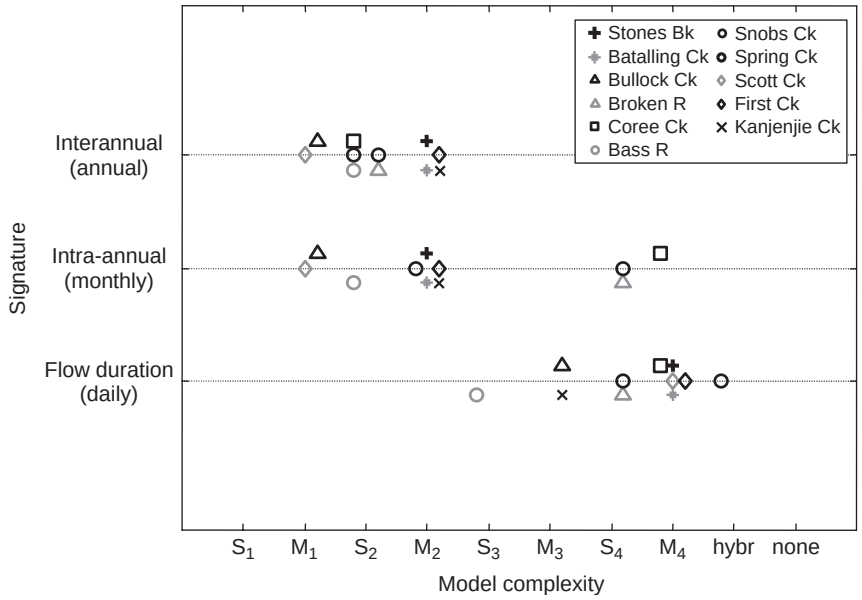


Figure 6 Minimum model complexity required to capture signatures: interannual variability, mean monthly variability, and daily flow duration curve (Reproduced from Farmer *et al.* (2003) by permission of American Geophysical Union)

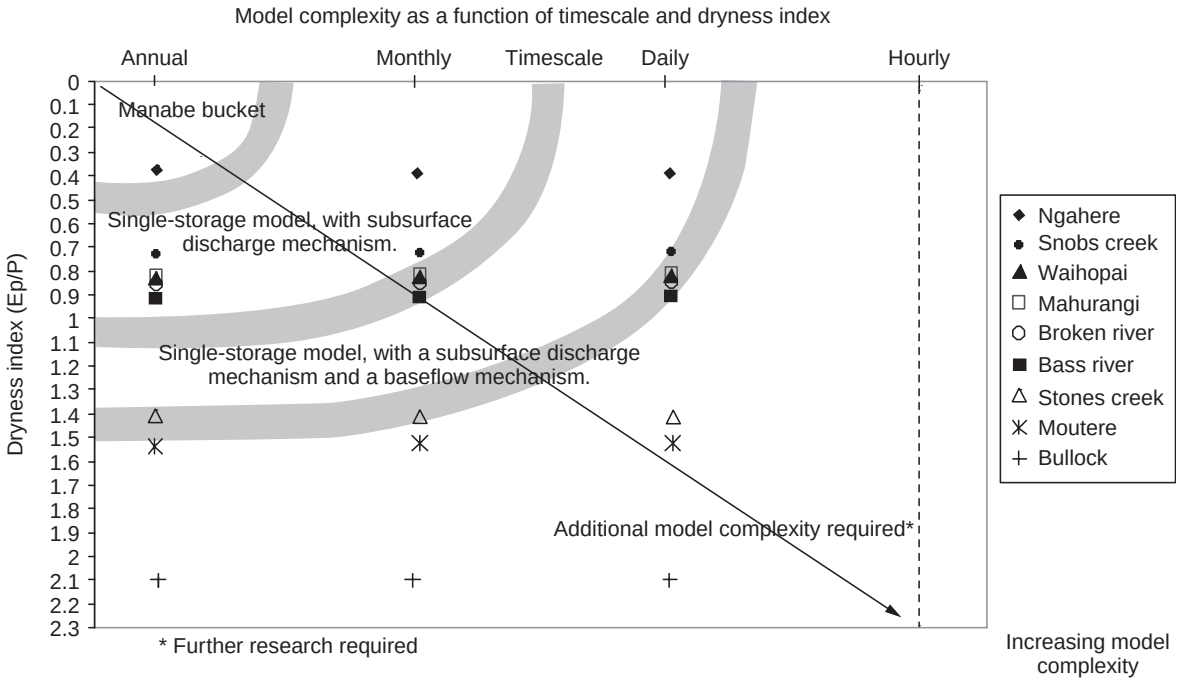


Figure 7 Required model complexity as a function of timescale and climatic dryness index. Model complexity increases with decreasing timescale and increasing aridity of the catchment (Reproduced from Atkinson *et al.* (2002) by permission of American Geophysical Union)

capture over 80% of the spatial variability of annual water balance for that part of the USA east of the Rocky Mountains. Choudhury (1999) proposed an alternative explanation that attributes the Budyko curve to the effects of spatial variability of precipitation and energy supply. In an attempt to quantify the effects of vegetation changes on annual water balance, Zhang *et al.* (2001) developed a simple two-parameter model that related mean annual evapotranspiration to rainfall, potential evapotranspiration, and to plant-available water capacity. This model related

average annual evapotranspiration under the same climatic conditions to vegetation characteristics, and attributed any differences therein to the way different kinds of vegetation use soil water.

Geomorphologic Instantaneous Unit Hydrograph (GIUH)

The unit hydrograph concept is one of the earliest building blocks of hydrologic theory, the one concept that separates hydrology from hydraulics. In place of resorting to the use of established theories of hydraulics (e.g. St. Venant equations) to the entire network of channels (and hillslopes) comprising the catchment, the unit hydrograph concept permits the formulation (through a convolution integral or equivalent linear transfer function relationship) of the catchment's storm response in terms of an (integrated) unit response, called the *instantaneous unit hydrograph* (IUH). Yet, until the work of Rodríguez-Iturbe and Valdés (1979) and the introduction of the geomorphologic instantaneous unit hydrograph (GIUH), the physical basis of the IUH was not explored, nor understood. The GIUH work has been driven by a desire to understand the pattern of hydrologic response at the event scale, and to interpret instantaneous unit hydrographs (IUHs) obtained from observed streamflow data in terms of the elements of catchment structure and functioning "that could have led to it" (Klemes, 1983). The GIUH of Rodríguez-Iturbe and Valdés (1979) made an explicit connection between the IUH and the geometry of the catchment, especially the topology of its channel network as represented by Horton order ratios. The physics of flows in the river network was, however, represented by simple storage routing, through a set of interconnected linear reservoirs, the mean residence times of which were governed by a common velocity parameter.

Subsequent work on interpretations of the GIUH have shed considerable light on the geometric features and component processes operating at lower scales that govern the shape of the IUHs, especially the relative roles of channel network topology, at-site and downstream hydraulic geometry, channel network hydraulics, and the nature of hill slope responses. An alternative scheme to incorporate network geometry was introduced by Lee and Delleur (1976) and Kirkby (1976) on the basis of the catchment's so-called *width* or *area function*. The width or area function is a mechanism to represent the distribution of catchment area along a length dimension. Various models of water transport motivated by hydraulic considerations have been proposed as alternatives to the use of linear reservoirs: (i) pure translation routing based on a constant velocity assumption (Gupta *et al.*, 1980), (ii) linearized diffusion wave routing (Rinaldo *et al.*, 1991), based on the assumption of constant velocity and hydrodynamic diffusivity, and (iii) based on spatially variable velocity and hydrodynamic dispersion (Saco and Kumar, 2002). This body of literature illustrates the use of the downward approach to "finger down" in search of improved physical meaning and realism of a concept established at the catchment scale. Whether the additional process details can be justified against data remains an open question.

Finally, it is interesting to compare this deterministic approach to GIUH analysis with the equivalent stochastic systems approach discussed in Section "A Data-Based Mechanistic Model of the Rainfall-Flow Process". There, the underlying shape of the discrete-time unit hydrograph (TUH) is represented by the pulse response of the estimated linear TF model (4); the changes of the magnitude (but not the shape) of the TUH under different flow and catchment storage conditions is defined by the effective rainfall nonlinearity $f(s_k)$ (a function of the catchment storage variable, s_k); and the interconnected linear reservoirs derive from the TF decomposition into the parallel pathway structure of Figure 2. Similar comments apply to the equivalent, continuous-time TF model, although there the impulse response is precisely the IUH, since it is the response to an instantaneous impulsive input.

Solute Transport in Heterogeneous Media

Transfer function modeling has been pursued within groundwater hydrology area in the 1980s for describing vertical percolation of water and solute transport in heterogeneous soils. Use of models based on deterministic description of the system, such as Richards equation, is the standard approach and has shown encouraging results under homogeneous conditions. However, in heterogeneous media, the success of such models has been rare because of the difficulty in parameterizing both vertical and lateral heterogeneity of soil hydraulic properties and the complicated internal dynamics of such a system. Instead of trying to provide an exact and detailed representation of the system, Jury (1982) developed an alternative modeling approach based on a transfer function model. In the case of solute transport within soils, the transfer function model is able to simulate the average solute concentrations at any depth and time based on the so-called *lifetime density function* (Jury *et al.*, 1982; Jury *et al.*, 1986; White *et al.*, 1986). As in the case of solute transport in river channels discussed above, the form of the TF model is not dictated by the detailed mechanisms of solute transport and the structural properties, and all small-scale processes that lead to an observed solute mass loss rate are implicitly imbedded into this lifetime density function.

Beven and Young (1988) and Young and Beven (1993) describe a rather different stochastic DBM approach to modeling solute transport in heterogeneous soils. Here the TF model is obtained by direct identification and estimation from tracer experiment data (similar to the approach used in

Section "The Aggregate Dead Zone Model: a Data-Based Mechanistic Model of Pollutant Transport in a River") and this is then interpreted in dynamic mass-conservation and dispersion terms.

DISCUSSION AND CONCLUSIONS

In this article, we have given a broad introduction to the downward or top-down approach to model development, distinguishing it from the current dominant paradigm that is the upward or bottom-up approach. We have illustrated its application to hydrological problems through two sets of examples based on the deterministic and stochastic approaches to downward modeling. The deterministic approach focuses more on drawing hydrological inferences from the analysis of the data; whereas the stochastic approach uses much more formal procedures of data-based analysis, drawn from the systems theory and statistics literature. While each of these methodologies may have come from different disciplines and may appear to be superficially different, there are significant similarities between them, as demonstrated by Quimpo (1971), O'Connor (1976), and Young (2003). Therefore, in the future, we believe that the two methodologies should be combined together to yield a common, formalized procedure for the analysis of data and the development of parsimonious, data-based mechanistic models. This will considerably enhance the acceptability of models developed by the downward approach, helping to streamline and make more efficient the required analysis of data and associated statistical inference.

Predictions in Ungauged Basins (PUB) remains one of the most difficult, yet practically important problems in hydrology, and is the focus of a major, new decade-long international initiative (Sivapalan *et al.*, 2003b). Extension of models developed through the systems approach to predict in ungauged basins requires that the model structure and parameterizations have a mechanistic basis, in order that their parameters can be connected to routinely observable climate and landscape properties, thereby enabling their extrapolation (so-called *regionalization*) to other, neighboring catchments. The enhancement of the mechanistic basis naturally requires "fingering down" to seek connections to processes and features operating at a lower level, which is a defining feature of the downward approach.

Klemes (1983) stated 20 years ago: ". . . the most promising route to significant new discoveries in hydrology is to combine the upward and downward search based on the existing facts and knowledge as well as on imagination and intuition, to form testable hypotheses – i.e. to apply the time-honored scientific method". However, there is a limit to how far we can refine or "finger down" the top-down models without changing over completely to the upward approach, and introducing assumptions not justified by, nor identifiable from, the data. This limit is usually reflected in a limit on the model complexity justified by the data. This is in line with the principle of *Occam's Razor*, which is consistent with the concept of parsimonious modeling discussed in this article: this states that one should not make any more assumptions than the minimum needed: that is, when choosing from among models with equal explanatory power, the simplest model is more likely to be correct (Young, 1978; Forster, 2000).

One way of addressing this problem is to move beyond the notion of "trying to model everything" and develop methods that are able to identify the dominant processes that control hydrological response in different environments (landscapes and climates) and at different scales, and then develop models that focus on these dominant processes (Grayson and Blöschl, 2000; Young, 1999; Young *et al.*, 1996, 2004). This may involve the development of some sort of classification or typological system (Woods, 2002) to precede the modeling effort, which is consistent with a downward avenue to hydrologic prediction. It would also be worth adopting the downward approach in a comparative mode in many catchments around the world under different climatic and hydrologic settings. This should allow both the methodology and the insights gained by the top-down modeling to be shared around the world and, hopefully, this will result in top-down models that apply in different climatic or hydrologic settings. Ideally, the order of these settings may be governed by a catchment classification or typology system, such as that proposed by Woods (2002) and McDonnell and Woods (2004).

REFERENCES

Abbott M.B. and Refsgaard J.C. (Eds.) (1996) *Distributed Hydrological Modelling*, Kluwer Academic Publishers: Dordrecht, p. 321.

Abbott M.B., Bathurst J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1986) An introduction to the European hydrologic system – système hydrologique Européen, "SHE", 1, history and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 45–49.

Atkinson S., Sivapalan M., Viney N.R. and Woods R.A. (2003a) Physical controls of space-time variability of hourly streamflows and the role of climate seasonality: Mahurangi catchment, New Zealand. *Hydrological Processes*, **17**, 2171–2193, doi: 10.1002/hyp.1327.

Atkinson S., Sivapalan M., Woods R.A. and Viney N.R. (2003b) Dominant physical controls of hourly streamflow predictions and an examination of the role of spatial variability: Mahurangi catchment, New Zealand. *Advances in Water Resources*, **26**(2), 219–235.

Atkinson S., Woods R.A. and Sivapalan M. (2002) Climate and landscape controls on water balance model complexity over changing time scales. *Water Resources Research*, **38**(12), 1314, doi: 10.1029/2002WR001487, 50.1–50.17.

Beck M.B. and Young P.C. (1975) A dynamic model for DO-BOD relationships in a non-tidal stream. *Water Research*, **9**, 769–776.

Beer T. and Young P.C. (1983) Longitudinal dispersion in natural streams. *American Society of Civil Engineers, Journal of Environmental Engineering*, **109**, 1049–1067.

Beven K.J. (1989) Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology*, **105**, 157–172.

Beven K.J. (2000) On the future of distributed modelling in hydrology. *Hydrological Processes*, **14**, 3183–3184.

Beven K.J. (2001) *Rainfall-Runoff Modelling: The Primer*, John Wiley & Sons: Chichester.

Beven K.J. and Young P.C. (1988) An aggregated mixing zone model of solute transport through porous media. *Journal of Contaminant Hydrology*, **3**, 129–143.

Blöschl G. and Sivapalan M. (1995) Scale issues in hydrological modelling – a review. *Hydrological Processes*, **9**(3/4), 251–290.

Brutsaert W. and Nieber J.L. (1977) Regionalized drought flow hydrographs from a mature glaciated plateau. *Water Resources Research*, **13**, 637–643.

Budyko M.I. (1974) *Climate and Life*, Academic Press.

Capra F. (1996) *The Web of Life: A New Scientific Understanding of Living Systems*, Anchor Books, Doubleday: New York, p. 348.

Choudhury B.J. (1999) Evaluation of an empirical equation for annual evaporation using field observations and results from a biophysical model. *Journal of Hydrology*, **216**, 99–110.

Chow V.T. (1964) Runoff. Chapter 14. In *Handbook of Applied Hydrology*, Chow V.T. (Ed.), McGraw-Hill: New York.

Dooge J.C.I. (1973) *Linear Theory of Hydrologic Systems*, Technical Bulletin No. 1468, US Department of Agriculture, Agricultural Research Service, Washington.

Dooge J.C.I. and O'Kane P. (2003) *Deterministic Methods in Systems Hydrology*, A.A. Balkema: Lisse.

Eder G., Sivapalan M. and Nachtnebel H.P. (2003) Modeling of water balances in Alpine catchment through exploitation of emergent properties over changing time scales. *Hydrological Processes*, **17**, 2125–2149, doi: 10.1002/hyp.1325.

Farmer D., Sivapalan M. and Jothityangkoon C. (2003) Climate, soil and vegetation controls upon the variability of water balance in temperate and semi-arid landscapes: downward approach to hydrological prediction. *Water Resources Research*, **39**(2), 1035, doi: 10.1029/2001WR000328.

Forster M.R. (2000) Key concepts in model selection: performance and generalizability. *Journal of Mathematical Psychology*, **44**, 205–231.

Freeze R.A. and Harlan R.L. (1969) Blueprint for a physically-based, digitally simulated hydrologic response model. *Journal of Hydrology*, **9**, 237–258.

Grayson R.B. and Blöschl G. (2000) Summary of pattern comparison and concluding remarks. In *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Chap. 14, Grayson R. and Blöschl G. (Eds.), Cambridge University Press: Cambridge, pp. 355–367.

Grayson R.B., Moore I.D. and McMahon T.A. (1992) Physically-based hydrologic modelling, 2. Is the concept realistic? *Water Resources Research*, **26**, 2659–2666.

Gupta V.K., Waymire E. and Wang C.T. (1980) A representation of an IUH from geomorphology. *Water Resources Research*, **16**, 885–862.

Jakeman A.J. and Hornberger G.M. (1993) How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, **29**, 2637–2649.

Jarvis P.G. (1993) Prospects for bottom-up models. In *Scaling Physiological Processes: Leaf to Globe.*, Ehleringer J.R. and Field C.B. (Eds.), Academic Press.

Jothityangkoon C. and Sivapalan M. (2003) Towards estimation of extreme floods: examination of the roles of runoff process changes and floodplain flows. *Journal of Hydrology*, **281**, 206–229.

Jothityangkoon C., Sivapalan M. and Farmer D. (2001) Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. *Journal of Hydrology*, **254**(1–4), 174–198.

Jury W.A. (1982) Simulation of solute transport using a transfer function model. *Water Resources Research*, **18**, 363–368.

Jury W.A., Sposito G. and White R.E. (1986) A transfer function model of solute transport through soil. 1. Fundamental concepts. *Water Resources Research*, **22**, 243–247.

Jury W.A., Stolzy L.H. and Shouse P. (1982) A field test of the transfer function model predicting solute transport. *Water Resources Research*, **18**, 369–375.

Kirkby M.J. (1976) Tests of random network model and its application to basin hydrology. *Earth Surface Processes*, **1**, 197–212.

Klemes V. (1983) Conceptualization and scale in hydrology. *Journal of Hydrology*, **65**, 1–23.

Kokkonen T.S., Jakeman A.J., Young P.C. and Koivusalo H.J. (2003) Predicting daily flows in ungauged catchments – model regionalization from catchment descriptors at the Coweeta hydrologic laboratory, North Carolina. *Hydrological Processes*, **17**, 2219–2238.

Lee M.T. and Delleur J.W. (1976) A variable source area model of the rainfall-runoff process based on the watershed stream network. *Water Resources Research*, **12**, 1029–1036.

Mackay, R., and Riley M.S. 1991 The problem of scale in the modelling of groundwater flow and transport processes. In *Chemodynamics of Groundwaters*, *Proceedings of Workshop*, Saint-Odile, November 1991, EAWAG, EERO, PIR "Environment" of CNRS, IMF Universite Louis Pasteur: Strasbourg, pp. 17–51.

McDonnell J.J. and Woods R.A. (2004) Editorial: on the need for catchment classification. *Journal of Hydrology*, **299**, 2–3.

Milly P.C.D. (1994) Climate, soil water storage, and the average annual water balance. *Water Resources Research*, **30**, 2143–2156.

Nash J.E. (1959) Systematic determination of unit hydrograph parameters. *Journal of Geophysical Research*, **64**(1), 111–115.

O'Connell P.E. (1991) A historical perspective. In *Recent Advances in the Modeling of Hydrologic Systems*, Bowles D.S. and O'Connell P.E. (Eds.), Kluwer: Dordrecht, pp. 3–30.

O'Connor K.M. (1976) A discrete linear cascade model for hydrology. *Journal of Hydrology*, **29**, 203–242.

Popper K. (1959) *The Logic of Scientific Discovery*, Hutchinson: London.

Quimpo R.G. (1971) Structural relation between parametric and stochastic hydrology models. *IAHS Publication*, **100**, 151–157.

Rinaldo A., Marani A. and Rigon R. (1991) Geomorphological dispersion. *Water Resources Research*, **27**(4), 513–525.

Rodríguez-Iturbe I. and Valdés J.B. (1979) The geomorphologic structure of hydrologic response. *Water Resources Research*, **15**, 1409–1420.

Saco P.M. and Kumar P. (2002) Kinematic dispersion in stream networks. 1. coupling hydraulic and network geometry. *Water Resources Research*, **38**(11), 1244, doi: 10.1029/2001WR000694.

Schulz K. and Beven K. (2003) Data supported robust parameterisations in land surface – atmosphere flux predictions: towards a top-down approach. *Hydrological Processes*, **17**, 2259–2277, doi: 10.1002/hyp.1331.

Silvert W. (1981) Top-down modelling in marine ecology. In *Progress in Ecological Engineering and Management by Mathematical Modelling*, Dubois D.M. (Ed.), Cebedoc, Liège: Belgium, pp. 259–270.

Sivapalan M. (2003) Process complexity at hillslope scale, process simplicity at the watershed scale: is there a connection? Invited commentary. *Hydrological Processes*, **17**, 1037–1041.

Sivapalan M., Blöschl G., Zhang L. and Vertessy R. (2003a) Downward approach to hydrological prediction. *Hydrological Processes*, **17**, 2101–2111, doi: 10.1002/hyp.1425.

Sivapalan M., Takeuchi K., Franks S.W., Gupta V.K., Karambiri H., Lakshmi V., Liang X., McDonnell J.J., Mendiondo E.M., O'Connell P.E., *et al.* (2003b) IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, **48**(6), 857–880.

Struthers I., Hinz C., Sivapalan M., Deutschman G., Beese F. and Meissner R. (2003) Modelling the water balance of a free-draining lysimeter using the downward approach. *Hydrological Processes*, **17**, 2151–2169, doi: 10.1002/hyp.1326.

Troch P.A., de Troch F.P. and Brutsaert W. (1993) Effective water table depth to describe initial conditions prior to storm rainfall in humid regions. *Water Resources Research*, **29**, 427–434.

Wallis S.G., Young P.C. and Beven K.J. (1989) Experimental investigation of the aggregated dead zone model for longitudinal solute transport in stream channels. *Proceedings of the Institution of Civil Engineers, Part 2*, **87**, 1–22.

White R.E., Dyson J.S., Haigh R.A., Jury W.A. and Sposito G. (1986) A transfer function model of solute transport through soil. 2. Illustrative applications. *Water Resources Research*, **22**, 248–254.

Wittenberg H. (1999) Baseflow recession and recharge as nonlinear storage processes. *Hydrological Processes*, **13**, 715–726.

Wittenberg H. (2003) Effects of season and man-made changes on baseflow and flow recession – case studies. *Hydrological Processes*, **17**, 2113–2123, doi: 10.1002/hyp.1324.

Wittenberg H. and Sivapalan M. (1999) Watershed groundwater balance estimation using streamflow recession analysis and baseflow separation. *Journal of Hydrology*, **219**, 20–33.

Woods R.A. (2002) Seeing catchments with new eyes. *Hydrological Processes*, **16**, 1111–1113.

Young P.C. (1978) A general theory of modeling for badly defined dynamic systems. In *Modeling, Identification and Control in Environmental Systems*, Vansteenkiste G.C. (Ed.), North Holland: Amsterdam, pp. 103–135.

Young P.C. (1983) The validity and credibility of models for badly defined systems. In *Uncertainty and Forecasting of Water Quality*, Beck M.B. and Van Straten G. (Eds.), Springer-Verlag: Berlin, pp. 69–100.

Young P.C. (1998) Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. *Environmental Modelling and Software*, **13**, 105–122.

Young P.C. (1999) Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis. *Computer Physics Communications*, **115**, 1–17.

Young P.C. (2000) Stochastic, dynamic modelling and signal processing: time variable and state dependent parameter estimation. In *Nonlinear and nonstationary signal processing*, Fitzgerald W.J., Walden A., Smith R. and Young P.C. (Eds.), Cambridge University Press: Cambridge, pp. 74–114.

Young P.C. (2001a) The identification and estimation of nonlinear stochastic systems. In *Nonlinear Dynamics and Statistics*, Mees A.I. (Ed.), Boston: Birkhauser, pp. 127–166.

Young P.C. (2001b) Data-based mechanistic modelling and validation of rainfall – flow processes. In *Model Validation: Perspectives in Hydrological Science*, Anderson M.G. and PD Bates (Eds.), John Wiley: Chichester, pp. 117–161.

Young P.C. (2002) Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*, **360**, 1433–1450.

Young P.C. (2003) Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale. *Hydrological Processes*, **17**, 2195–2217, doi: 10.1002/hyp.1328.

Young P.C. and Beven K.J. (1993) Soils: solute transport. In *Concise Encyclopedia of Environmental Systems*, Young P.C. (Ed.), Pergamon Press: Oxford, pp. 551–553.

Young P.C., Chotai A. and Beven K.J. (2004) Data-based mechanistic modelling and the simplification of environmental systems. In *Environmental Modelling: Finding Simplicity in Complexity*, Chap. 22, Wainwright J. and Mulligan M. (Eds.), John Wiley & Sons: Chichester, pp. 371–388.

Young P.C. and Lees M.J. (1993) The Active Mixing Volume (AMV): a new concept in modelling environmental systems. In *Statistics for the Environment*, Chap. 1, Barnett V. and Turkman K.F. (Eds.), John Wiley & Sons: Chichester, pp. 3–44.

Young P.C., Parkinson S. and Lees M.J. (1996) Simplicity out of complexity: Occam's Razor revisited. *Journal of Applied Statistics*, **23**, 165–210.

Zhang L., Dawes W.R. and Walker G.R. (2001) Response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resources Research*, **37**, 701–708.

Encyclopedia of
Hydrological Sciences

Encyclopedia of
Hydrological Sciences

Editor-in-Chief

Malcolm G Anderson
Department of Geography, University of Bristol, Bristol, UK

Senior Advisory Editor

Jeffrey J McDonnell
Department of Forest Engineering, Oregon State University, Corvallis, OR, US

4

EHS

WILEY

PART 12

Open-channel Flow

135: Open Channel Flow – Introduction

WALTER H GRAF AND MUSTAFA S ALTINAKAR

Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

The introduction begins with a presentation of the different types of channels as well as with the corresponding flow regimes. Subsequently, the notions of the distribution of velocity and of pressure are exposed.

INTRODUCTION

Treated will be the flow and flow-related phenomena in artificial and natural channels of simple geometry with a free surface subjected to the atmospheric pressure.

CHANNELS

A channel is a transport system where water flows and where the free surface is subject to atmospheric pressure. Channels are important and integral parts of any hydrologic system.

Two categories of channels are to be distinguished: (i) natural channels and (ii) artificial channels (see Figure 1).

Natural channels are watercourses that exist naturally on (or under) the earth, such as gullies, brooks, torrents, rivers, streams, and estuaries. The geometric and hydraulic properties are generally rather irregular.

Artificial channels are watercourses developed by men on (or under) the earth, such as open channels (navigation channels, power canals, irrigation and drainage channels) or closed channels where flow does not fill the entire section (hydraulic tunnels, aqueducts, drains, sewage canals). The geometric and hydraulic properties of such channels are generally rather regular.

Geometry of Channels

The (transversal) section of a channel is a section in the cross-sectional plane being normal to the direction of flow (see Figure 2). The section, or better the wetted (cross-sectional) surface, A, is the portion of the cross section occupied by the liquid.

A channel whose section does not vary and whose longitudinal slope and roughness remains constant – however, the flow depth may vary – is called a *prismatic* channel; otherwise the channel is a nonprismatic one.

The geometric elements of a section or wetted surface, A, are the following:

The wetted perimeter, P, of the channel, being formed by the length of the line of contact between the wetted surface and the bed and the side walls. The wetted perimeter can be composed of a fixed or immobile bed (concrete, rock) or of a mobile bed (granulates of sediments).

The hydraulic radius, R_h, being the ratio of wetted surface to its wetted perimeter, or

$$R_h = \frac{A}{P} \tag{1}$$

The width, B, of the channel being the width at the free surface.

The hydraulic depth, D_h, of the channel being defined by $D_h = A/B$.

The flow depth, h, or the water height – if not defined otherwise – is considered to be the maximum depth.

Formulas for the geometric elements for different types of channel sections are given in Chow (1959), p. 21. A natural watercourse might have a rather irregular geometric form, but often it can be rather well approximated by a trapezoidal or parabolic section.

Besides the geometric elements, the longitudinal slopes are also to be considered, namely,
the slope of the bed (bottom or floor), S_f,
the slope of the water surface (piezometric), S_w.

The value of the bottom slope depends essentially on the topography of the terrain, being generally weak; it may be expressed by $S_f = tg\alpha \cong \sin\alpha$.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

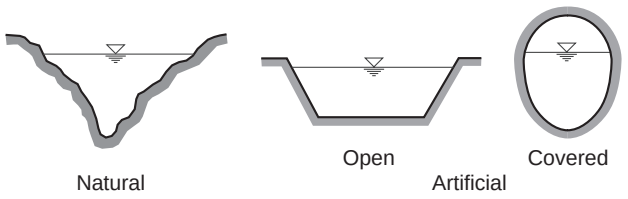


Figure 1 Kinds of channels. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

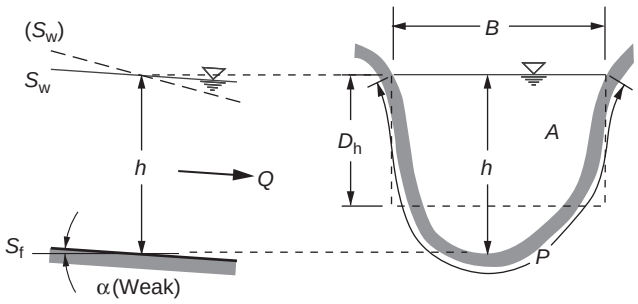


Figure 2 Geometric elements of a channel section. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

FLOW IN CHANNELS

Flow in open channels is essentially due to the inclination (slope) of the bed, while flow in closed conduits (see Graf and Altinakar, 1991, 1995, Chapter PP.1), is due to a difference in the charge between the sections.

Types of Flow

A classification of open channel flow may be done according to the change of the flow depth, h or D_h, with respect to time and space:

$$D_h = f(t, x) \tag{2}$$

Time variation: Flow is steady (stationary or permanent), if the average velocity of flow, U, but also the flow depth, h or D_h, remain invariable with time. Consequently, the discharge remains constant:

$$UA = Q \tag{3}$$

between the different sections of the channel, supposing there is no lateral inflow or outflow. Flow is unsteady if the flow depth, $D_h(t)$, as well as the other parameters vary with time. Consequently, the discharge is no more constant. Strictly speaking, open channel flow is rarely steady. However, the temporal variations are often sufficiently slow and the flow may be assumed to be steady.

Space variation: Flow is uniform if the flow depth, h or D_h, as well as the other parameters, remain unchanged at every section of channel. The line of the bottom slope is thus parallel to the one of the free-water surface, or $S_f \equiv S_w$. Flow is nonuniform or varied if the depth, $D_h(x)$, as well as the other parameters, vary along the length of the channel. The bottom slope is thus different from the slope of the water surface, or $S_f \neq S_w$. Nonuniform flow can be steady or unsteady. Varied flow can be accelerating, $dU/dx > 0$, or decelerating, $dU/dx < 0$, depending on the variation of velocity in the direction of flow. If the flow is a gradually varied one, the depth, $D_h(x) \cong D_h$, varies slowly from one section to another. If the flow is a rapidly varied one, the depth, $D_h(x)$, changes abruptly over a comparatively short distance, sometimes with a discontinuity. This happens generally in the neighborhood of a singularity, such as at a weir or at a change of channel width, and also at a hydraulic jump or a hydraulic drop.

Flow Regimes

The physics of open channel flow is governed basically by the interplay of the inertia forces, the gravity forces, and the friction (viscosity and roughness) forces.

The (reduced) equations of motion (see Graf and Altinakar, 1991, 1995, Section FR.7.2), which bring these forces into play, involve the following dimensionless numbers:

The Froude number, being the ratio of gravity to inertia forces, or

$$\frac{\rho g}{\rho U_c^2 / L_c} = Fr^{-2} \quad \text{and} \quad Fr = \frac{U_c}{\sqrt{g L_c}} \tag{4}$$

The Reynolds number, being the ratio of friction to inertia forces, or

$$\frac{\mu (U_c / L_c^2)}{\rho U_c^2 / L_c} = Re^{-1} \quad \text{and} \quad Re = \frac{U_c L_c}{\nu} \tag{5}$$

Added to these two numbers is still the relative roughness, being the ratio of the roughness height, k_s, to a characteristic length, or

$$\frac{k_s}{L_c} \tag{6}$$

Here, U_c and L_c are the characteristic velocity and length; one takes often $U_c = U$ and $L_c = R_h$ or $L_c = D_h$.

In the hydraulics of open channel flow, one generally defines these dimensionless numbers as

$$Fr = \frac{U}{\sqrt{g D_h}};\ Re = \frac{4 R_h U}{\nu} \quad \text{or} \quad Re' = \frac{R_h U}{\nu};\ \frac{k_s}{D_h} \tag{7}$$

The Reynolds number is used to classify the flow (see Graf and Altinakar, 1991, 1995, Chapter FR.3) as follows:

Laminar flow	$Re' < 500$
Turbulent flow	$Re' > 2000$
Transition flow	$500 < Re' < 2000$

From numerous experiments with different artificial channels (see Chow, 1959, p. 10), it results that flow is turbulent if the Reynolds number, Re', reaches a value of 2000. In general, flow in open channels is a turbulent and often rough flow.

The Froude number is used to classify the flow as follows:

Subcritical (fluvial) flow	$Fr < 1$
Supercritical (torrential) flow	$Fr > 1$
Critical flow	$Fr \equiv Fr_c = 1$

In general, flow in open channels can be thus of the three types.

DISTRIBUTION OF VELOCITY

In flow along a wall (the bottom of a channel), a distribution of velocity (see Graf and Altinakar, 1991, 1995, Chapter FR. 6) is encountered. Being zero at the wall, the point velocity, u, increases towards the free surface; its maximum value is often found slightly below this free surface. The velocity profile is approximately logarithmic.

Steady flow depends in general on the three variables, x, y, and z; this is called three-dimensional flow. If such a channel has a large width, B – large in comparison with the depth, $B > 5\,h$ – flow is considered two-dimensional, with the exception of a small distance close to the vertical side walls. Hydraulic calculations are considerably simplified, if one assumes the flow to be one-dimensional. The average velocity, $U(x)$, in a vertical or in a section, is expressed by

$$U = \frac{1}{h}\int_0^h u(z)\,\mathrm{d}z \quad \text{or} \quad U = \frac{1}{A}\int_0^B \int_0^h u(z)\,\mathrm{d}z\,\mathrm{d}y \tag{8}$$

In open channels of simple geometry, one encounters generally turbulent flow where the point velocity, $u(x, z)$, differs little from the average velocity, $U(x)$. In the steady state, such an hypothesis allows to consider the flow as one-dimensional.

DISTRIBUTION OF PRESSURE

A general expression for the relative (with respect to the atmospheric pressure) pressure on a curved bottom of the channel is given (see Graf and Altinakar, 1998, Chapter 1.4) by

$$p_f = \gamma h' + \rho\frac{U^2}{r}h + \cancel{p_a} = 0 \tag{9}$$

having an hydrostatic and an accelerating component (see Figure 3).

For uniform flow, when the average velocity, U, remains constant and the streamlines are reasonable rectilinear (with $r \to \infty$) , the distribution of pressure is hydrostatic in a section, normal to the bottom (see Figure 4). The expression for the pressure, relative to the bottom, can now be given as

$$p_f = +\gamma h' = \gamma h \cos\alpha \tag{10}$$

For the usually encountered open channels, the inclination, α, is rather weak, that is, $\alpha < 6°$; or $S_f < 0.1$, implying that $\cos\ \alpha \simeq 1$. Consequently, equation (10) reduces to

$$\left(\frac{p}{\gamma}\right)_f = h \tag{11}$$

where h is the flow depth in the channel.

For flow, being (slightly) nonuniform, thus having a curvilinear current of converging or diverging type, there exists an acceleration component caused by the inertia

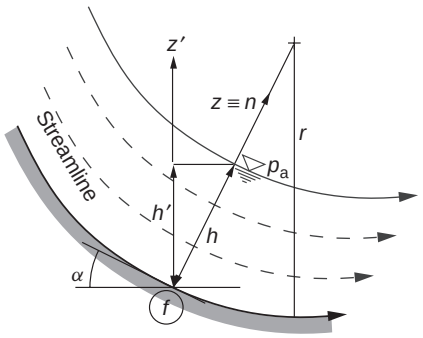


Figure 3 Flow over a concave bottom. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

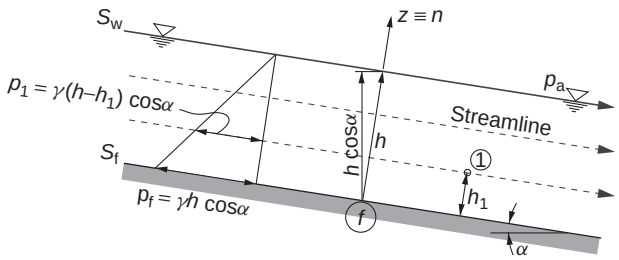


Figure 4 Flow with a uniform current

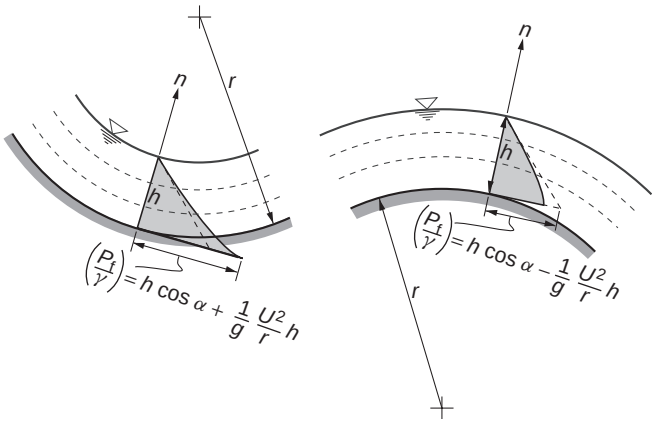


Figure 5 Flow over a concave and a convex bottom. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

forces. The expression for the pressure relative to the bottom is given by

$$p_f = \gamma h' \pm \rho\frac{U^2}{r}h = \gamma h\cos\alpha \pm \rho\frac{U^2}{r}h \qquad (12)$$

being (+) for a concave and (−) for a convex bottom. The distribution of pressure is no more hydrostatic (see Figure 5). For an external concave current, the centrifugal force increases the pressure; while for a convex current, this force decreases the pressure. In the latter case, the pressure could get below the atmospheric pressure, thus causing separation of flow on the channel bed.

REFERENCES

Chow V.T. (1959) *Open Channel Hydraulics*, McGraw-Hill: New York.

Graf W.H. and Altinakar M.S. (1991; 1995) *Hydrodynamique*, Eyrolles: Paris; Presses Polytechniques Romandes: Lausanne.

Graf W.H. and Altinakar M.S. (1998) *Fluvial Hydraulics*, John Wiley & Sons: Chichester.

136: Hydrodynamic Considerations

WALTER H GRAF AND MUSTAFA S ALTINAKAR

Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

The equations of continuity and of energy are developed for the general case. Subsequently, the specific energy, a concept useful for the understanding of different problems, is introduced.

INTRODUCTION

Some fundamental notions of hydrodynamics, being the basis of open-channel hydraulics, shall be presented.

EQUATION OF CONTINUITY

The equation of continuity, one of the basic equations of fluid mechanics, is an expression of the conservation of mass.

A flow of an incompressible fluid, being steady, uniform and almost rectilinear, in an open channel with a free surface and a weak bed slope (see Figure 1) will be studied. Two channel sections are considered. The volume, entering by the first section is $Q\,\mathrm{d}t$; the volume leaving by the second section is $[Q + (\partial Q/\partial x)\mathrm{d}x]\mathrm{d}t$. The variation of the volume between these sections during the time, $\mathrm{d}t$, is consequently $(\partial Q/\partial x)\ \mathrm{d}x\ \mathrm{d}t$. This variation is the result of a modification of the free surface, $\partial h/\partial t$, between the two sections during the time, $\mathrm{d}t$; it is expressed by $(B\mathrm{d}x)\ (\partial h/\partial t)\ \mathrm{d}t$, where $B(h)$ is the width of the channel, $h(x, t)$ is the flow depth and $\mathrm{d}A = B\mathrm{d}h$. Assuming the fluid incompressible, the above two expressions are made equal (see Chow, 1959, p. 525) and one obtains:

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = 0 \tag{1}$$

For a rectangular channel, defining $q = Q/B$ as the unit discharge, equation (1) is:

$$\frac{\partial q}{\partial x} + \frac{\partial h}{\partial t} = h\frac{\partial U}{\partial x} + U\frac{\partial h}{\partial x} + \frac{\partial h}{\partial t} = 0 \tag{2}$$

For steady flow, $\partial A/\partial t = 0$, the equation of continuity, equation (1), reduces to:

$$\frac{\partial Q}{\partial x} = 0 \tag{3}$$

If a supplementary discharge leaves (or enters) the channel between the two sections, equation (1) can be written as:

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} \pm q_\lambda = 0 \tag{4}$$

where q_λ is the supplementary discharge per unit length.

EQUATION OF ENERGY

The equation of energy is an expression of the first principle of thermodynamics. The energy for an element of incompressible fluid, written in homogeneous quantities of length (see Figure 2) – here as the height of a liquid with specific weight $\gamma = \rho g$ – in almost rectilinear and one-dimensional flow, taken with respect to the plane of reference (PdR), is given by:

$$\frac{u^2}{2g} + \frac{p}{\gamma} + z_p = \frac{p_t}{\gamma} = \mathrm{Cte} \tag{5}$$

The different terms represent: the velocity head, $u^2/2g$, the pressure head, p/γ, the elevation (position) of a point P, z_P, and the (mechanical) energy head or the total head, $p_t/\gamma = H$.

Consider the following reasonable assumptions:

1. The piezometric head, $p*/\gamma = p/\gamma + z_P$, is supposed to be constant over a normal to the bed, implying that the distribution of pressure is hydrostatic.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

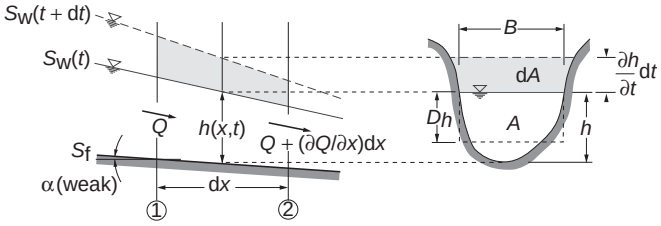


Figure 1 Scheme for the equation of continuity. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

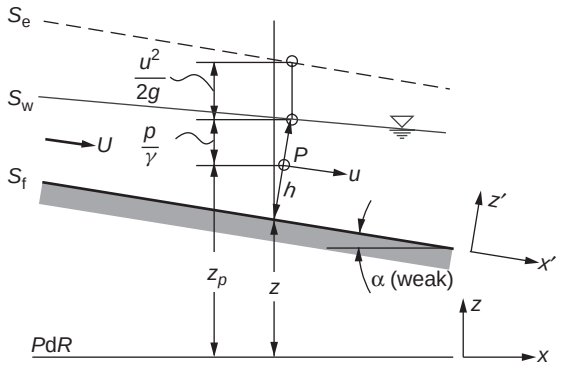


Figure 2 Scheme for the equation of energy in a cross section. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2. The slope (weak) of the channel is given by: $S_f = tg\alpha = -dz/dx \cong \sin\alpha$;
3. If h is the flow depth, the pressure head at the bed of the channel (see equation (10) in **Chapter 135, Open Channel Flow – Introduction, Volume 4**) is: $(p/\gamma)_f = h\cos\alpha$, but for weak slopes, $\alpha < 6°$, where $S_f < 0.1$, one may take $\cos\alpha \cong 1$. The system of the coordinates, xz, is thus almost identical with the one of the coordinates, $x'z'$, (see Figure 2).
4. In a perfect fluid, each fluid element moves with the same velocity, being identical with the average velocity in the section, U.

Making use of these assumptions, the total head in a section is now given by:

$$\frac{U^2}{2g} + h + z = H \tag{6}$$

The equation of energy, equation (6), is a manifestation of the principle of energy if the fluid is *perfect*. From one to another section, each of the three terms in equation (6) can take a different value, but the sum, H, remains constant.

For flow of a *real* fluid with a free surface, being unsteady and nonuniform, the difference of the total head

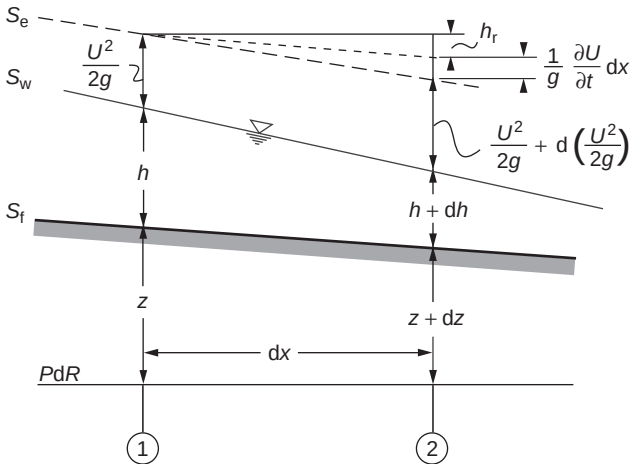


Figure 3 Scheme of the equation of energy, between two sections. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

between two sections, separated by a distance, dx, (see Figure 3) is given as:

$$\alpha_e\frac{U^2}{2g} + h + z = \left[\alpha_e\frac{U^2}{2g} + d\left(\alpha_e\frac{U^2}{2g}\right)\right] + [h + dh] + [z + dz] + \frac{1}{g}\frac{\partial U}{\partial t}dx + \frac{1}{g}\frac{\tau_o}{\rho}\frac{dP}{dA}dx \tag{7}$$

Here $(1/g)\ (\partial U/\partial t)\ dx$ is the term of energy due to unsteady flow, and $(1/g)\ (\tau_o/\rho)\ (dP/dA)\ dx$ is the term of energy or head loss due to friction (see Graf and Altinakar, 1991, 1995, p. 138); where τ_o is the shear stress due to the frictional forces acting on the wetted surface, $dP\ dx$. This last term is usually written as h_r. The kinetic energy correction coefficient, α_e, results from the distribution of the velocity in the section. Its numerical values (see Chow, 1959, p. 28), notably for turbulent flow, are very close to unity, $\alpha_e \cong 1$.

Dividing by the distance, dx, and using partial differentials, the equation of energy, equation (7), can thus be written:

$$\frac{1}{g}\frac{\partial U}{\partial t} + \frac{U}{g}\frac{\partial U}{\partial x} + \frac{\partial h}{\partial x} - S_f = -S_e$$

steady, uniform
steady, non-uniform
unsteady, non-uniform

(8)

where $h_r = S_e\ dx$, with S_e as the energy slope, and $S_f = -(dz/dx)$ as the bottom slope. Equation (8) is the dynamic equation for unsteady and nonuniform flow.

The head loss, h_r, must be evaluated with a formula such as the one of Weisbach–Darcy, equation (7), of Chézy,

equation (8) in **Chapter 137, Uniform Flow, Volume 4**, or of other experimenters. Such relations are only valid for steady, uniform flow; however – for lack of better information – they are also used (see Chow, 1959, p. 217) for unsteady and nonuniform flow.

The equation of continuity, equation (1), and the equation of motion, equation (8), form together the *equations of Saint-Venant* (see Chow, 1959, p. 528). Despite the various simplifications made to obtain these equations, their solutions are often rather complicated. In some cases, which are simple but still realistic, explicit solutions are possible. For flow, which is steady but nonuniform or steady and uniform, equation (8) is used in a reduced form.

The equation of motion, equation (8), can also be obtained by applying the momentum equation. The resulting equation is almost the same (see Chow, 1959, p. 51).

SPECIFIC ENERGY

The total head, H, in a given cross section was defined with respect to an arbitrary horizontal plane. If the plane of reference is placed into the bed slope, S_f, a fraction of the total head, called the *specific energy*, H_s, is defined; one writes now (see Figure 4):

$$\frac{U^2}{2g} + h = H_s \quad \text{or} \quad \frac{Q^2/A^2}{2g} + h = H_s \tag{9}$$

The notion of the specific energy is often very useful; it helps to understand and to solve different problems of free-surface flow.

For a given section in the channel, the area of flow, A, is a function of the flow depth, h, and equation (9) gives a relation of the following form:

$$H_s = f(Q, h) \tag{10}$$

which allows a study of the variation of h with H_s for $Q = \text{Cte}$ or h with Q for $H_s = \text{Cte}$ (see Graf and Altinakar, 1998, Chapter 2.3).

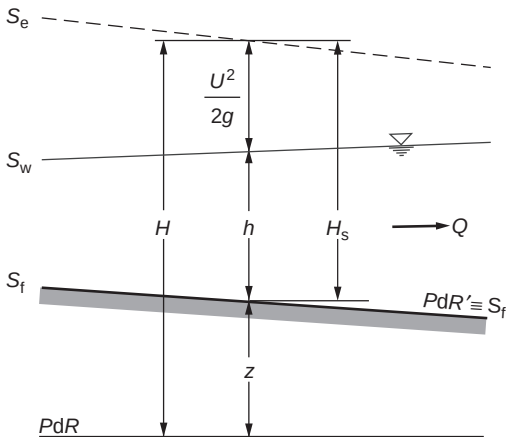


Figure 4 Definition of the total head, H, and the specific energy, H_s. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Specific-energy Curve

Equation (9) gives the evolution of the specific energy, H_s, as a function of the flow depth, h, for a given discharge, $Q = UA$. This curve (see Figure 5) has a horizontal asymptote for $h = 0$ and is asymptotic to the line $h = H_s$ for $h = \infty$. In addition, the curve has a minimum, H_{s_c}, for:

$$\frac{dH_s}{dh} = -\frac{Q^2}{gA^3}\frac{dA}{dh} + 1 = 0 \quad \text{or} \quad \frac{Q^2}{g}\frac{B}{A^3} = \frac{U^2}{gD_h} = 1 \tag{11}$$

since dA/dh is equal to the width of the channel, B, at the free surface, and the hydraulic depth is defined as $D_h = A/B$.

For a channel with a rectangular cross section, one takes $D_h = h$. The flow depth, h, which corresponds to the minimal specific energy, H_{s_c}, is called *critical depth*, h_c.

Following the curve (see Figure 5) one notices that for a given discharge $Q = \text{Cte}$, and for an arbitrary value of

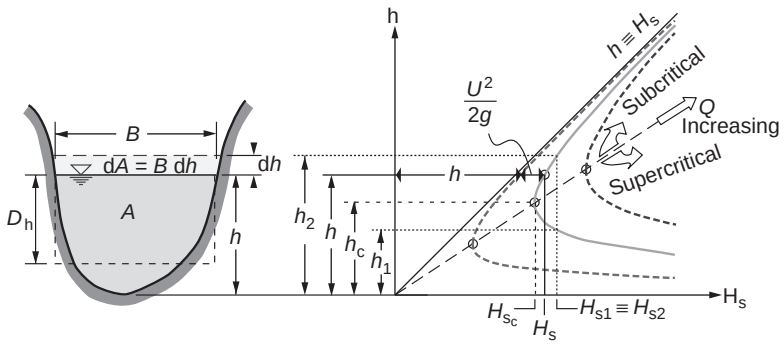


Figure 5 Specific-energy curve, $H_s = f(h)$, for $Q = \text{Cte}$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

specific energy, H_s, – if flow can take place – there are always two solutions for the flow depth, h_1 and h_2. They are called the *corresponding (alternate) depths*; one of which, h_1, is smaller and the other one, h_2, is larger than the critical depth, h_c. Both of these depths are indications of different regimes of flow, thus:

$h < h_c$ supercritical (torrential) regime
$h > h_c$ subcritical (fluvial) regime
$h \equiv h_c$ critical regime

Each curve (see Figure 5) has thus two branches. Consequently, a steady flow in a channel can exist in two different ways, both having the same specific energy, H_s; namely in supercritical regime, where the flow depth is small and the velocity large, and in subcritical regime, where the flow depth is large and the velocity small. For a variation of the discharge, Q, the corresponding curves have the same form; they follow each other for an increase in the discharge, starting at the origin, (see Figure 5).

Discharge Curve

Equation (9) shows also the evolution of the discharge, Q, as a function of the flow depth, h, for a given specific energy, H_s, such as:

$$Q = A\sqrt{2g(H_s - h)} \tag{12}$$

From such a curve (see Figure 6) one obtains a discharge of $Q = 0$ for $h = 0$ and $h = H_s$. The curve has a maximum value, Q_{max}, for:

$$\frac{dQ}{dh} = \frac{2g(H_s - h)(dA/dh) - Ag}{[2g(H_s - h)]^{1/2}} = \frac{gB[2(H_s - h) - D_h]}{[2g(H_s - h)]^{1/2}} = 0 \tag{13}$$

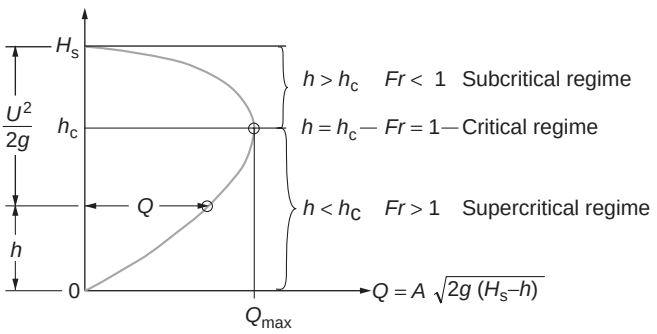


Figure 6 Discharge curve, $Q = f(h)$, for $H_s = \text{Cte}$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

using the definitions $dA/dh = B$ and $D_h = A/B$. Its derivative becomes zero, if:

$$2(H_s - h) - D_h = 0 \tag{14}$$

The values, h and D_h, which do correspond to the maximum discharge, Q_{max}, represent the *critical depth*, h_c and D_{h_c}. For flows smaller than Q_{max}, one finds again the two different flow regimes (see Figure 6 and also Figure 5).

Equation (14) can be used to obtain the critical depth ($h \equiv h_c$ and $H_s \equiv H_{s_c}$). For a channel with a rectangular cross section, $D_h \equiv h$, the critical depth is:

$$h_c = \tfrac{2}{3} H_{s_c} \tag{15}$$

while for channels with a triangular or parabolic cross section, one obtains respectively:

$$h_c = \tfrac{4}{5} H_{s_c} \quad \text{and} \quad h_c = \tfrac{3}{4} H_{s_c} \tag{16}$$

Critical Depth

In summary, the critical depth, h_c, in a channel is the flow depth at which the specific energy is minimal, H_{s_c}, for a given discharge (see Figure 5), and at which the discharge is maximal, Q_{max}, for a given specific energy (see Figure 6).

Between equation (14) and equation (12), the maximum discharge, Q_{max}, is obtained as:

$$Q_{max} = A\sqrt{g D_{h_c}} \tag{17}$$

The average velocity, which corresponds to the critical hydraulic depth, D_{h_c}, is:

$$U_c = \sqrt{g D_{h_c}} \quad \text{or} \quad \frac{U_c^2}{2g} = \frac{D_{h_c}}{2} \tag{18}$$

In critical regime, the velocity head is thus equal to half of the hydraulic depth. Both equation (18) and equation (11) could also be expressed as:

$$\frac{U_c}{\sqrt{g D_{h_c}}} = 1 \tag{19}$$

being precisely the definition of the Froude number (see equation (4) **Chapter 135, Open Channel Flow – Introduction, Volume 4**) in critical regime; here the Froude number is equal to unity, $Fr_c = 1$. Consequently, the Froude number classifies also the different flow regimes, such as:

$Fr > 1$ supercritical regime $U > U_c$
$Fr < 1$ subcritical regime $U < U_c$
$Fr = 1$ critical regime $U \equiv U_c$

The average velocity, U_c, corresponding to the critical depth, h_c, is given by:

$$U_c = \sqrt{g D_{h_c}} = c \tag{20}$$

being equal to the celerity, c, of the propagation of (superficial) infinitesimal gravity waves in a channel of hydraulic depth, D_{h_c} (see Graf and Altinakar, 1998, Chapter 2.4).

Using equation (15) and equation (18), the critical depth for a rectangular channel, $D_h \equiv h$, with a unit discharge, $q = Uh$, is now given by:

$$\frac{h_c}{2} = \frac{q^2}{2gh_c^2} \quad \text{or} \quad h_c = \sqrt[3]{\frac{q^2}{g}} \tag{21}$$

The maximum unit discharge, q, which may produce itself in a channel of rectangular section is equal to:

$$q = \sqrt{g h_c^3} = \sqrt{g \left(\tfrac{2}{3} H_{s_c}\right)^3} \tag{22}$$

According to equation (17) and equation (22), the critical hydraulic depth, D_{h_c}, or the critical flow depth, h_c, depend only on the discharge. Thus it is interesting to use this information for metering flow in open channels; such as is done with the free overfall and the Venturi channel (see Graf and Altinakar, 1998, Chapter 2.3).

REFERENCES

Chow V.T. (1959) *Open Channel Hydraulics*, McGraw-Hill: New York.

Graf W.H. and Altinakar M.S. (1991; 1995) *Hydrodynamique*, Eyrolles: Paris; Presses Polytechniques Romandes: Lausanne.

Graf W.H. and Altinakar M.S. (1998) *Fluvial Hydraulics*, John Wiley & Sons: Chichester.

137: Uniform Flow

WALTER H GRAF AND MUSTAFA S ALTINAKAR

Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

The equation of continuity and of motion will be developed. The different relationships for the determination of the coefficients of friction for fixed and mobile channel beds are subsequently presented. The calculation of discharge is elaborated. Elementary knowledge about flow in curves as well as instabilities at the free-water surface are exposed.

INTRODUCTION

Flow in a channel is considered as uniform and steady, if the flow depth remains invariable in the flow direction as well as in time. In fluvial hydraulics, uniform flow is taken as the base (reference) for all other considerations, and this despite the fact that truly uniform flow is rarely encountered in reality.

HYDRODYNAMIC EQUATIONS

Equation of Continuity

If the flow is uniform and steady, the wetted cross section of the flow, A, remains the same in direction, x, and in time, t. The equation of continuity is given as:

$$\frac{\partial(UA)}{\partial x} = 0 \tag{1}$$

where $Q = UA$ is the discharge and U is the average velocity. Consequently, the discharge remains constant, $Q = Cte$. Between two cross sections (see Figure 1), one has:

$$A_1 U_1 = Q = A_2 U_2 \tag{2}$$

and with $U_1 = U_2$ and $A_1 = A_2$, one writes: $Q = U\,A$.

Equation of Motion

Consider a prismatic channel (see Figure 1). The liquid in motion provokes a friction force at the wetted perimeter, $F_F = \tau_o P\,dx$, by the action of the longitudinal component of the gravity force, $F_G = \gamma A\,dx \sin\alpha = W \sin\alpha$. In uniform flow, there exists an equilibrium between these forces, $\tau_o P\,dx = \gamma A\,dx \sin\alpha$. Consequently, one obtains:

$$\tau_o = \gamma \frac{A}{P} \sin\alpha = \gamma R_h S_f \tag{3}$$

where τ_o is the tension due to the friction forces, called the *shear stress*, which acts on the wetted surface (wall and bed).

The quotient of the wetted cross section, A, and its wetted perimeter, P, defines the hydraulic radius, R_h. The inclination is usually very small; thus one may write $\sin\alpha \cong tg\alpha = S_f$.

In hydrodynamics, one defines, $\tau_o/\rho = u_*^2$, where u_* is the friction velocity. Thus one can also write:

$$u_* = \sqrt{g R_h S_f} \tag{4}$$

Instead of the shear stress, $\tau_o = \rho u_*^2$, one may also use the definition of the *friction coefficient* (see Graf and Altinakar, 1991, p. 433) which is given by:

$$f = \frac{\tau_o}{\rho U^2/8} = 8\left(\frac{u_*}{U}\right)^2 \tag{5}$$

Subsequently one obtains:

$$\left(\frac{f}{8}\right)\rho U^2 = \tau_o = \rho g R_h S_f \quad \text{or} \quad S_f = f\,\frac{1}{4R_h}\frac{U^2}{2g} \tag{6}$$

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

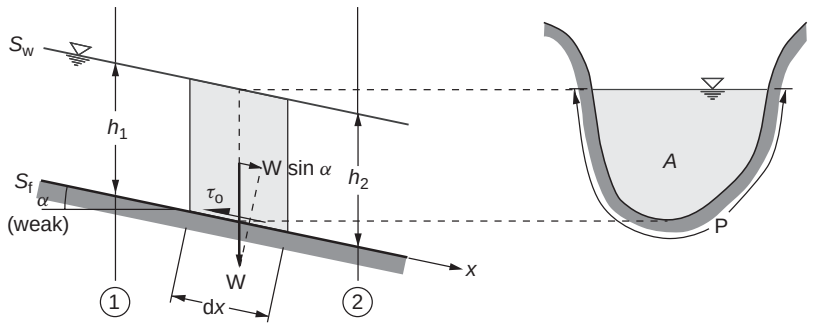


Figure 1 Scheme of uniform flow. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

This relation, which is known as the equation of *Weisbach–Darcy* (see Graf and Altinakar, 1991, sect. FR. 2.1 and sect. PP.2), is also used for calculating flow in pipes. The coefficient, $\boldsymbol{f}$, of friction (head loss) depends on the Reynolds number and the relative roughness, and also on the form of the cross section. The equation of Weisbach–Darcy can also be written as:

$$U = \sqrt{\frac{8g}{\boldsymbol{f}}}\sqrt{R_h S_f} \tag{7}$$

an expression which is frequently given in the form of:

$$U = C\sqrt{R_h S_f} \tag{8}$$

This is called the *relationship of Chézy*, where C is the resistance coefficient of Chézy.

In uniform regime (see equation 7 and equation 8), the flow depth, h, which corresponds to the hydraulic radius, R_h, is defined as being the *normal flow depth*, $h \equiv h_n$.

Different formulae have been elaborated over the years to render expressions for the friction (resistance) coefficients. Herewith some more common formulae will be presented, namely the:

1. coefficient of Weisbach–Darcy,
2. coefficient of Chézy,
3. coefficient of Manning–Strickler,
4. coefficient of friction for mobile bed.

COEFFICIENT OF FRICTION

The two relations, equation (7) and equation (8), give satisfactory results, notably for practical problems in channel flow, if applied correctly and respecting their possible limitations. The ASCE (see Silberman *et al.*, 1963) recommended, however, the use of the equation of Weisbach–Darcy.

Artificial and, particularly, natural channels have all types of the cross-sectional shapes. No parameter exists which would well take care of the variability in form; the use of the hydraulic radius is often not sufficient. An estimation of the friction coefficient for a fixed or immobile bed is already difficult; but still more difficult will be an estimation for a mobile bed.

Coefficient of Weisbach–Darcy

In the above relation, equation (7), the definition of friction coefficient, $\boldsymbol{f}$, is analogous to the one given for circular pipes. For pipes having an industrial roughness, a universal formulation is given by the diagram of Moody–Stanton or the relation of Colebrook–White, for turbulent flow. For cross sections, which are geometrically close to circular sections, one may readily use the experiments performed on pipes. However, some modifications are necessary; the hydraulic radius (see Graf and Altinakar, 1991, p. 439) should be expressed by $4R_h = 4$ (A/P). Thus $4R_h$ becomes the characteristic length to be used in the definition of the Reynolds number, the relative roughness and in the equation of Weisbach–Darcy. For the roughness, k_s, in artificial channels, the equivalent roughness, established for industrial pipes, may be taken (see Graf and Altinakar, 1998, Table 3.1).

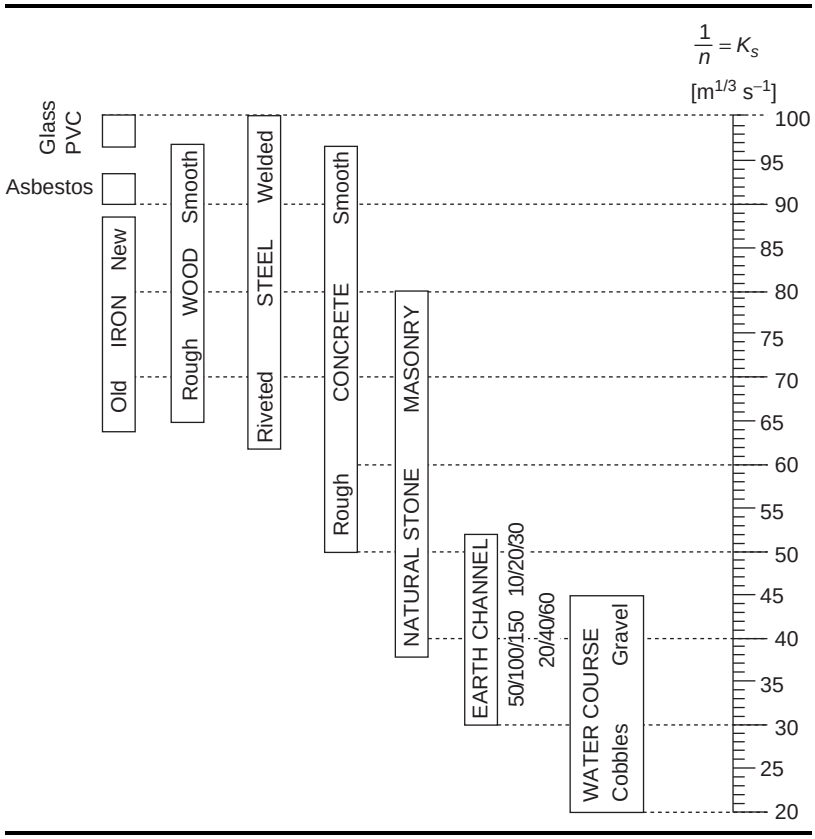
The use of the diagram of Moody–Stanton (see Graf and Altinakar, 1991, Fig. PP.9) with $\mathrm{Re} = 4R_h U/\nu$ and $k_s/4R_h$ gives values for $\boldsymbol{f}$ for laminar and turbulent flow. Subsequently, one obtains the average velocity, U, using equation (7), or the bed slope, $S_f = S_w$, of the channel, using equation (6).

Instead of using the diagram of Moody–Stanton, one may also take the semiempirical *relation of Colebrook–White* (see Graf and Altinakar, 1991, p. 436), valid only for turbulent flow, which is written for channels as follows:

$$\sqrt{\frac{1}{\boldsymbol{f}}} = -2\log\left(\frac{k_s/R_h}{a_f} + \frac{b_f}{Re\sqrt{\boldsymbol{f}}}\right) \tag{9}$$

with $12 < a_f < 15$ and $0 < b_f < 6$, established for different kinds of cross sections, as well as for different types of roughness (see Silberman *et al.*, 1963, p. 104).

Table 1 Coefficients of roughness of Manning and of Strickler



For channels or watercourses whose bed is made up of a granulate, one generally takes $k_s \cong d_{50}$; d_{50} being the diameter equal to 50% of grains in the granulometric curve.

Coefficient of Chézy

For turbulent, rough flow the *formula of Chézy*, equation (8), can be used. However, it cannot be used for laminar or turbulent smooth flow. The coefficient of Chézy, C [m$^{1/2}$/s], is a dimensional expression. Different formulae, being all of empirical nature, have been advanced for the determination of the coefficient of Chézy, C; all of which make use of the hydraulic radius, R_h.

In the practice, one prefers presently exponential relations and one uses commonly the *formula of Manning–Strickler* in the form of:

$$U = K_s R_h^{2/3} S_f^{1/2} \quad \text{with} \quad C = K_s R_h^{1/6} = \frac{1}{n} R_h^{1/6} \quad (10)$$

Here K_s [m$^{1/3}$ s^{-1}] is the coefficient of *Strickler* and n [m$^{-1/3}$ s^1] is the coefficient of *Manning*. The above relation, equation (10), was elaborated using numerous measurements, performed in both natural and artificial channels. The values of n and K_s are given in Table 1. More detailed tables are available in the literature.

Coefficient of Manning

The most popular formula is presently the one of *Manning–Strickler*, often called shortly the *formula of Manning*:

$$U = \frac{1}{n} R_h^{2/3} S_f^{1/2} \quad (11)$$

This is a rather simple relationship, but it must be used only for turbulent rough flow, thus for flow at large Reynolds numbers. In such a case, the coefficient of Manning, n, stays constant for a given roughness, while the coefficient of Chézy, C, depends on the relative roughness, $(R_h^{1/6}/n)$.

Complete tabulations of the coefficient of Manning, n, have been presented by Chow (1959, pp. 110–113) and Graf (1984, pp. 306–309). Furthermore, Chow (1959, pp. 115–123) and Barnes (1967) provide photos of different natural and artificial channels as a visual support, to facilitate the choice of the coefficient of Manning in the range of $0.012 < n < 0.15$. Indicative values are summarized in Table 1. The influence of vegetation on the coefficient of friction is extensively treated by Chow (1959).

It must be pointed out that the values of the coefficient of Manning are the same both in the metric and in the English system. In the latter case, one has to use the following

relation:

$$C = \frac{1.48}{n} R_h^{1/6} \tag{10a}$$

For watercourses, where the bed and walls are made up of a noncohesive granulate, the *formula of Strickler* (see Strickler, 1923, pp. 11–15) may be used:

$$K_s = \frac{21.1}{d_{50}^{1/6}} \quad \text{or} \quad K_s = \frac{26}{d_{90}^{1/6}} \tag{12}$$

where d_{50} or d_{90} [m] are the diameters, being equal to 50 or 90% of the grains in the granulometric curve.

Composite Roughness

The coefficients of friction, $\boldsymbol{f}$, n and C, are valid as long as the entire wetted perimeter has the same roughness; thus the wetted section is homogeneous. In sections where the wetted perimeter is *not homogeneous*, the bed and the sidewalls have different roughness; thus it becomes necessary to compute an equivalent coefficient of friction.

According to *Einstein* (see Chow, 1959, p. 136), one divides – in a reasonable way – the wetted surface, A, in N parts, each one having its wetted perimeter, P_1, $P_2 \ldots\ldots P_N$, and its coefficient of friction, n_1, $n_2 \ldots . n_N$. Furthermore, one assumes that the average velocity of each particular section, A_1, $A_2 \ldots . A_N$, is the same and thus also the same as the average velocity of the entire section, U. Using, for example, the formula of Manning, equation (11), the equivalent coefficient of friction for a composite roughness is given by:

$$n = \left[\frac{\sum_1^N (P_N n_N^{3/2})}{P} \right]^{2/3} \tag{13}$$

Bed Forms

Natural, but also artificial channels may have a *mobile bed*, being a channel bed composed of solid particles (noncohesive granulate), which displace themselves by the action of the flow. The bed may become covered with *bed forms*, commonly called *dunes*.

The geometry of a dune (idealized, since sometimes they are not well apparent) is approximated by a triangular form of length, λ, and of height, ΔH. Indicatives relations (see Graf, 1984, p. 283), made dimensionless by the flow depth, are given as:

$$\frac{\Delta H}{h} < \frac{1}{6}; \quad \frac{\lambda}{h} \approx 5 \tag{14}$$

The presence of bed forms causes an increase in the flow resistance. For the calculation of the total shear stress on the bed, τ_o, one assumes (see Graf, 1984, p. 303) that the contribution of the roughness due to the particles, τ', and the one due to the bed forms, τ'', is additive, namely:

$$\tau_o = \tau' + \tau'' \tag{15}$$

Using the definition of the friction velocity and of the coefficient of friction, one writes:

$$u_*^2 = (u_*')^2 + (u_*'')^2 \quad \text{or} \quad \boldsymbol{f} = \boldsymbol{f}' + \boldsymbol{f}'' \tag{16}$$

but also:

$$n^{3/2} = (n')^{3/2} + (n'')^{3/2} \quad \text{and} \quad \frac{1}{C^2} = \frac{1}{(C')^2} + \frac{1}{(C'')^2} \tag{16a}$$

The total shear stress, τ_o, (see equation 15) varies with the Froude number, Fr.

Coefficient of Friction, Mobile Bed

A quantification of friction coefficients for flow over a mobile bed has, up to now, not been very successful over a large range of flow parameters. Nevertheless, there exist methods, where one calculates the coefficient of friction due to the grain roughness, $\boldsymbol{f}'$ or n', using the formulae presented above; subsequently one determines the coefficient of friction due to the bed forms, $\boldsymbol{f}''$ or n'', using other types of formulae. There exist also methods where one determines directly the *entire* coefficient of friction, $\boldsymbol{f}$ or n.

Different empirical relationships have been elaborated for the calculation of the friction velocity and of the coefficient of friction, u_*'' and $\boldsymbol{f}''$, being due to bed forms, for example, the relation proposed by *Einstein–Barbarossa* which is given usually in graphical form (see Figure 2). Many observations from American rivers, having $0.19 < d_{35}$ [mm] < 4.3 and $1.49 \times 10^{-4} < S_w < 1.72 \times 10^{-3}$ were used. This relationship is expressed (see Graf, 1984, p. 310) by:

$$\frac{U}{u_*''} = f\left(\frac{\rho_s - \rho}{\rho} \frac{d_{35}}{R_h' S_f} \right) = f(\Psi') \tag{17}$$

where R_h' is the hydraulic radius attributed to the grain roughness.

More relations have been presented in Graf (1984, pp. 303–320), Raudkivi (1976, chap. 6) and Graf and Altinakar (1998, p. 84).

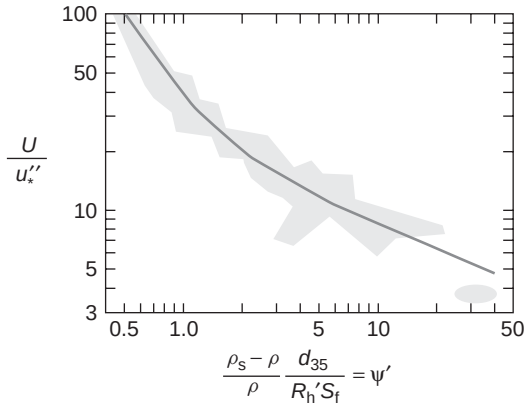


Figure 2 Friction velocity, u_*'', due to bed forms; after Einstein–Barbarossa. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

DISCHARGE CALCULATION, FIXED BED

Determination of the discharge, Q, in a channel with a fixed bed requires the knowledge of the channel geometry, of the roughness coefficient and of the bed slope. Assumed will be that the walls (bed and side walls) of the channel are fixed or immobile, thus not subject to erosion.

Conveyance

The discharge, Q, at uniform flow, given by equation (2), using the corresponding velocity, given by equation (11), can be expressed as:

$$Q = UA = \frac{1}{n} R_h^{2/3} S_f^{1/2} A \tag{18}$$

An expression can be formed such as, $K(h) = (1/n) R_h^{2/3} A$, known as the *conveyance* of the channel (see Bakhmeteff, 1932, p. 13), being only a function of the flow depth, $h \equiv h_n$, known as the *normal depth* for the given discharge, Q. Thus, above expression yields:

$$Q = K(h)\sqrt{S_f} \quad \text{or} \quad \frac{Q}{\sqrt{S_f}} = f(h) \tag{19}$$

For a given form (shape) of the section, this relation can be obtained and plotted point by point (see Figure 3). One can readily calculate the conveyance, K, for geometrically simple sections; for complex ones, a graphical solution is necessary. The normal depth, h_n, increases with the discharge, Q. For identical channels, but having different slopes, S_f, the normal depth increases if the bed slope decreases.

The conveyance, K, characterizes the channel; it represents a measure of the capacity of water transport through the cross section. The curve of the normal depths (see Figure 3) is rather useful in solving different kinds of problems: if two of the three parameters, h_n, Q and S_f, (see equation 19) are known, the third one can be found; *a priori*, the roughness of the walls, n, is taken to be known.

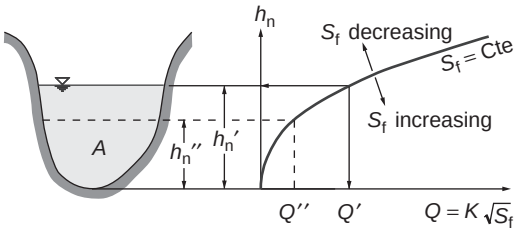


Figure 3 Curve of conveyance or of normal depth. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Normal Depth

The normal depth, h_n, is the flow depth at uniform flow of discharge, Q, at a given bed slope, S_f. (All geometric elements of the cross section, which correspond to the normal depth, h_n, are known as *normal elements*, such as: R_{h_n}, A_n, or P_n)

The normal depth of a channel of a given geometry is calculated using the relation for the discharge, equation (18). This relation shows that uniform flow is only possible in a channel whose bed slope is descending, $S_f > 0$. In a horizontal channel, $S_f = 0$, the normal depth would be infinite.

For a natural watercourse and for rectangular channels whose width, B, is very large ($B \gg h$), one takes $R_h \approx h$ as the hydraulic radius. Subsequently, one obtains for the normal or uniform depth:

$$h_n = \left(\frac{q^2}{C^2 S_f}\right)^{1/3} \quad \text{where } q = \frac{Q}{B} \tag{20}$$

Composite Section

A cross section of a channel can be composed of different subsections (see Figure 4), of which each one can have a different roughness and a different bed slope. This

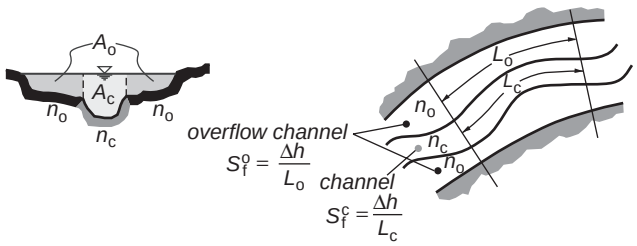


Figure 4 Composite section. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

is frequently the case during flooding, when the flow leaves the channel and enters into the overflow section of the channel.

Such a case can be approximately treated by applying the formula of discharge for each subsection:

$$Q = Q_c + Q_o = \frac{1}{n_c} A_c R_{h_c}^{2/3} \left(\frac{\Delta h}{L_c}\right)^{1/2} + \frac{1}{n_o} A_o R_{h_o}^{2/3} \left(\frac{\Delta h}{L_o}\right)^{1/2} \qquad (21)$$

Note that the wetted perimeters, P_c and P_o, should be calculated for the lines of contact between water and bed.

DISCHARGE CALCULATION, MOBILE BED

Artificial and natural channels, whose flow moves in an alluvium, composed of a granulate, are channels of *mobile bed*. The discharge can be calculated using the coefficient of roughness for a mobile bed.

In such channels, the velocity (in the vicinity of the bed) should:

1. not be superior to a certain critical value, otherwise there is a risk of erosion of the solid particles on the bed: this is the permissible maximum velocity or *velocity of erosion*, also called the *critical velocity*;
2. not be inferior to a certain critical value, otherwise there is a risk of deposition of the solid particles: this is the permissible minimum velocity or *velocity of sedimentation*.

The flow velocity, U, to be selected for a "good functioning" of the channel, must lie between the velocity of erosion, $U_E \equiv U_{cr}$, and the one of sedimentation, U_D:

$U_D < U < U_E$. As is evident in Figure 5, the two velocities, U_D and U_E, will have distinctly different values.

Sedimentation Velocity

The velocity of sedimentation, U_D, is the minimum velocity which is necessary to transport the flow containing solid particles in suspension. A diagram (see Figure 5) which was established by *Hjulstrom* (see Graf, 1984, p. 88) delimits the zone of sedimentation as a function of the diameter of the (monodispersed) granulate.

Critical Velocity

There will be erosion of the bed (and the walls), if one exceeds a certain critical value, expressed with:

1. the average critical velocity, U_{cr}, or the critical velocity, $u_{b_{cr}}$, at or close to the bed, or
2. the critical shear stress, $\tau_{o_{cr}}$.

From a hydraulic view point, it is more reasonable to use the shear stress, τ_o, as a criterion of erosion. Using the definition of shear stress:

$$\tau_o = \gamma R_h S_f \qquad (3)$$

and of the average velocity:

$$U = C\sqrt{R_h S_f} \qquad (8)$$

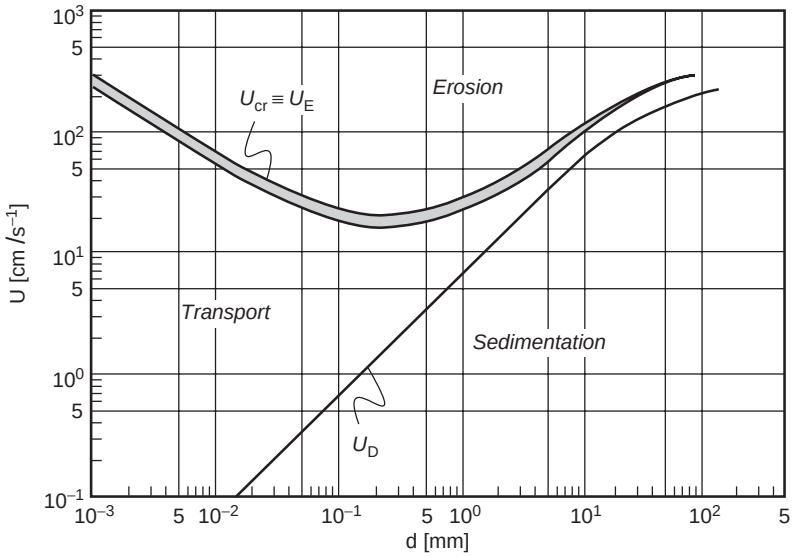


Figure 5 Velocity of sedimentation and of erosion, U_D and U_{cr}, for a uniform granulate, after Hjulstrom

the following ratio is obtained:

$$\frac{U}{\sqrt{\tau_o/\rho}} = \frac{C}{\sqrt{g}}$$

In fluvial hydraulics, one usually uses (see Graf, 1984, p. 91) a dimensionless form of the shear stress, τ_*, or:

$$\frac{\tau_o}{(\gamma_s - \gamma)d} \equiv \tau_* = \frac{\gamma R_h S_f}{(\gamma_s - \gamma)d} \tag{22}$$

where d is the diameter of the granulate (to be specified); γ_s and γ are the specific weight of the granulate and of water, respectively. With this relation, one compares the flow parameters, R_h and S_f, with the granulometric parameters, d and $(\gamma_s - \gamma)$.

Amongst the different formulae for prediction of the critical condition, found in the literature (see Graf, 1984, chap. 6), only the one by *Shields* will be presented. Relying on concepts of the hydrodynamics, *Shields* developed a relation between the dimensionless shear stress, τ_* (see equation 22), and the friction/particle Reynolds number, $Re_* = u_* d/\nu$, such as:

$$\tau_* \equiv \frac{\tau_o}{(\gamma_s - \gamma)d} = f\left(\frac{u_* d}{\nu}\right) \tag{23}$$

where $u_* = \sqrt{\tau_o/\rho}$. The form of this relation using experimental data was determined. An average curve, reasonably well defined, characterizes the begin of erosion, expressed by $\tau_{*_{cr}}$. For the particle diameter, one takes usually $d \equiv d_{50}$. It is to be seen (see Figure 6) that the critical values fall roughly in the range of $0.03 < \tau_{*_{cr}} < 0.06$. The determination of $\tau_{*_{cr}}$ is done using the above relation, equation (23), by successive approximations. The criterion of *Shields* is of great importance for the hydraulic engineer.

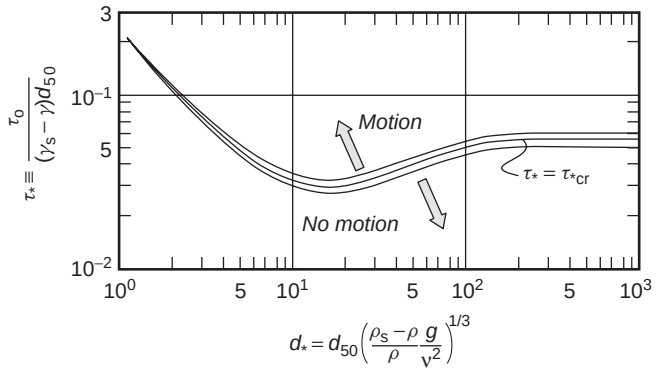


Figure 6 Dimensionless shear stress, τ_*, as a function of the dimensionless diameter, d_*, after Shields–Yalin. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Since a direct use of the relation of *Shields* is laborious, Yalin (1972, p. 82) proposed to use a dimensionless diameter of the granulate, $d_* = d\{[(\rho_s - \rho)/\rho](g/\nu^2)\}^{1/3}$ Consequently, the above relation, equation (23), can be written as:

$$\tau_* = f(d_*) \tag{23a}$$

which is given with Figure 6; one usually takes $d = d_{50}$. If the properties of the fluid, ρ and ν, and of the granulate, ρ_s and d, are known, one can readily determine the corresponding value of $\tau_{*_{cr}}$ and subsequently of τ_{cr}.

For cohesive material, the determination of the critical values, U_{cr} or $\tau_{*_{cr}}$, represents a difficult task; the specialized literature (see Graf, 1984, chap. 12, and Raudkivi, 1976, chap. 9) should be consulted.

Distribution of Shear Stress

The shear stress, τ_o, is given by equation (3). For a channel of large width one may take $R_h \equiv h$.

However, it must be remarked that the shear stress, τ_o, is distributed over the wetted perimeter, P. A typical distribution for a trapezoidal channel (see Chow, 1959, p. 169) is given with Figure 7.

An expression for the critical shear stress on the channel side walls, $(\tau_{o_{cr}})^w$, was proposed by *Lane* (see Graf, 1984, p. 116), being of the following form:

$$(\tau_{o_{cr}})^w = \tau_{o_{cr}}\left[\cos\theta\left(\frac{1 - tg^2\theta}{tg^2\varphi}\right)^{1/2}\right] \tag{24}$$

$\tau_{o_{cr}}$ is the critical shear stress on the bed – given for example, with Figure 6, θ is the inclination of the side wall(s), and φ is the angle of repose. The latter depends on the granulometry and on the cohesion (see Graf, 1984, p. 115); it varies such as $20° < \varphi < 40°$. Evidently: $(\tau_{o_{cr}})^w < \tau_{o_{cr}}$, and for stable side walls: $\theta < \varphi$.

FLOW IN CURVES

A curve or bend, positioned in a rectangular channel, causes a change in the flow direction. If the discharge, Q, remains constant along the curve, the flow velocity, U, as well as the wetted section, A, remain also constant. The sectional

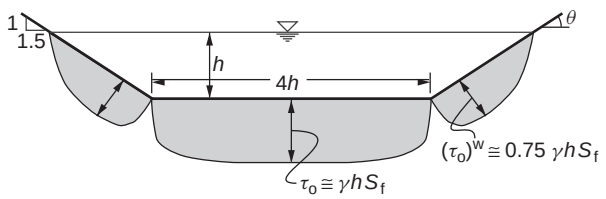


Figure 7 Distribution of the shear stress in a trapezoidal channel. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

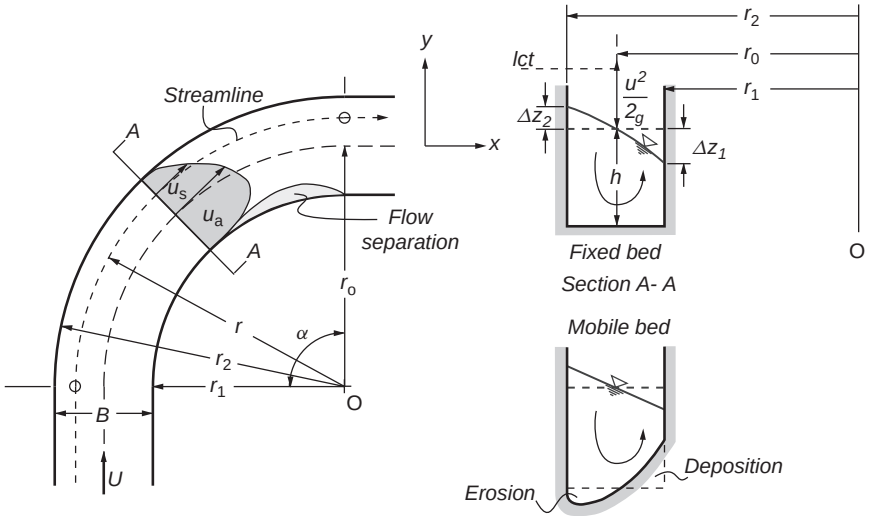


Figure 8 Flow in a curve. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

distribution of the flow depth, $h(y)$, will be responsible for a transversal water slope and a superelevation, Δz, at the outside of the curve.

The distribution of velocity in the curve can be approximated by the one of a free vortex (see Graf and Altinakar, 1991, p. 196). The velocity has a maximum at the inside of the curve (see Figure 8).

Superelevation

In a curve, the streamlines will no longer stay parallel and the flow becomes three-dimensional. This is a complex physical phenomenon, for which an adequate analysis seems difficult. It was proposed (see Kozeny, 1953, p. 223) that the superelevation can be calculated by:

$$\Delta z = \frac{B r_o}{r_1 r_2} \frac{U^2}{2g} \tag{25}$$

If the channel width, $B = r_2 - r_1$, is small compared to the radius of the curve, r_0, one gets the simplified expression of:

$$\Delta z = \frac{B}{r_o} \frac{U^2}{2g} \tag{25a}$$

The transversal water profile is convex; the superelevation, $\Delta z = \Delta z_2 + \Delta z_1$, has its maximal value, $\Delta z = \Delta z_{max}$, usually observed for fluvial flow, $Fr < 1$, at the entrance of the curve and for supercritical flow, $Fr > 1$, at the exit of the curve.

The superelevation at the outside of the curve (see Figure 8) causes a vertical downward current, which comes back to the water surface at the inside of the curve. Such a secondary current will superpose itself on the primary flow and result in helicoidal flow over the entire reach of the curve. If the outside side wall is of mobile material, erosion will take place; on the inside there will be deposition (see Figure 8).

Supercritical Flow

Gravity waves (see Graf and Altinakar, 1998, sect. 3.5.2 and sect. 2.4.3) will establish themselves in a curve, notably if the flow is supercritical, $Fr > 1$. For a rectangular channel, the resulting maximum superelevation, $(\Delta z' + \Delta z)$, due to the gravity waves, can be twice the superelevation, Δz, obtained with equation (25). It is given by:

$$\Delta z' = \frac{B}{r_o} \frac{U^2}{2g} \tag{26}$$

If the superelevations get very large, methods are available to suppress it, like the construction of transition curves or the installation of steps on the channel bed.

Head Loss

In flow over a curve, one encounters not only a head loss due to friction, h_r, but also one due to the curvilinear flow, $h_r{}^c$. This additional head loss is usually expressed by:

$$h_r{}^c = \zeta_c \frac{U^2}{2g} \tag{27}$$

where ζ_c is a coefficient which depends on $\zeta_c = f$ (*Fr, Re,* r_0/B, h/B, α); α being the angle of the curve, *Fr* and *Re* are the number of Froude and of Reynolds, respectively. According to numerous experiments, one takes (see Chow, 1959, p. 443):

$$0.1 \leq \zeta_c \leq 1.1$$

where the larger values of ζ_c are for curves of $r_0/B = 0.5$.

INSTABILITY AT SURFACE

If the channel slope is substantial and/or if the flow is supercritical, the water surface can become unstable. The normal flow depth, h_n, must now be considered as an average value. Such an instability is characterized by a series of gravity waves of small flow depth, called *roll waves*, progressing downstream, and a breaking of these waves, causing an *air entrainment*.

Roll Waves

An instability at the water surface is evidenced by the formation of roll waves. The uniform steady flow becomes locally an unsteady one. The roll waves are superposed on the uniform flow (see Figure 9). They displace themselves towards the downstream – increasing in height and then collapsing – with an absolute celerity, c_w, being larger than the flow velocity, U, or $c_w = U + \sqrt{gh} > U$.

There is no simple criteria available to determine the geometrical dimensions of this type of waves. Their height can, however, attain dimensions of the order of magnitude of the prevailing flow depth (see French, 1986, p. 625). The crests of the roll waves are zones of strong turbulence, while the rest of the waves remains remarkably smooth. Some theoretical considerations for a determination of the (in)stability of uniform flow are presented in Liggett (1975, Chap. 6).

Air Entrainment

For large channel slopes, S_f, – such as exist also on the downstream face of a weir – the flow is usually supercritical and gravity waves appear at the water surface. These waves will break and entrain air into the water. The turbulence will diffuse (mix) the air bubbles across the entire flow depth; and water droplets will escape into the air. In flow of such an air–water mixture, it becomes rather difficult to define the flow depth; the water surface is often covered by *white water*.

A schematic distribution of the concentration, C(z), of air is given with Figure 10. Two regions are to be distinguished: bubbles in the water and droplets in the air.

The equivalent flow depth of water (without the air) is defined by:

$$h = \int_0^\infty (1 - C)\,\mathrm{d}z \cong h_a(1 - \overline{C}) \tag{28}$$

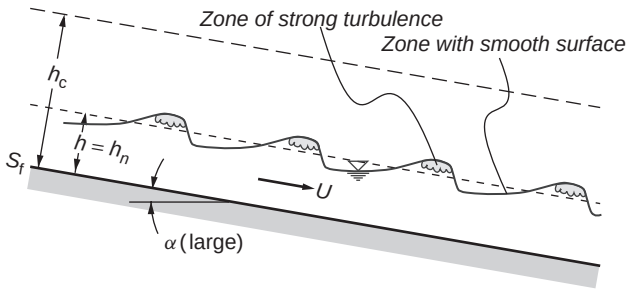


Figure 9 Roll waves. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

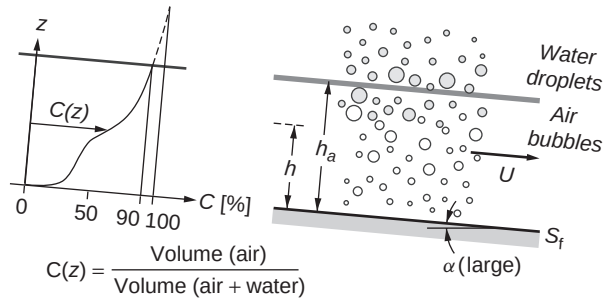


Figure 10 Flow with air entrainment. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and the average velocity of water by $U = q/h$, where q is the unit discharge of the water. The depth of the mixture, h_a, is the depth where the concentration is equal to $C = 90\%$ (see Wood, 1985, p. 21) and $\overline{C}$ is the average concentration in the cross section, given (see Henderson, 1966, p. 185) by the relation:

$$\overline{C} = 0.7 \log\left(\frac{\mathrm{S_f}}{\mathrm{q}^{1/5}}\right) + 0.9 \tag{29}$$

which was obtained from experimental studies performed by *Straub* et *Anderson* for $0.14 < q[\mathrm{m^2/s}] < 0.93$. Usually (see Wood, 1985, p. 21) one takes:

$$\overline{C} = 0.14 \quad \text{for slopes of } \mathrm{S_f} = 7.5°$$
$$\overline{C} = 0.71 \quad \text{for slopes of } \mathrm{S_f} = 75°$$

REFERENCES

Bakhmeteff B.A. (1932) *Hydraulics of Open Channel Flow*, McGraw-Hill: New York.

Barnes H.H. (1967) *Roughness Characteristics of Natural Channels*, Water Supply Paper No. 1849, U.S. Geological Survey.

Chow V.T. (1959) *Open Channel Hydraulics*, McGraw-Hill: New York.

French R. (1986) *Open-Channel Hydraulics*, McGraw-Hill: New York.

Graf W.H. (1971; 1984) *Hydraulics of Sediment Transport*, McGraw-Hill: New York; Water Resources Publications: Littleton.

Graf W.H. and Altinakar M.S. (1991; 1995) *Hydrodynamique*, Eyrolles: Paris; Presses Polytechniques Romandes: Lausanne.

Graf W.H. and Altinakar M.S. (1998) *Fluvial Hydraulics*, John Wiley & Sons: Chichester.

Henderson F.M. (1966) *Open Channel Flow*, Macmillan Company: New York.

Kozeny J. (1953) *Hydraulik*, Springer-Verlag: Wien.

Liggett J. (1975) Stability. In *Unsteady Flow in Open Channels*, Mahmood K. and Yevjevich V. (Eds.), Water Resources Publications: Fort Collins.

Raudkivi A.J. (1976; 1990) *Loose Boundary Hydraulics*, Pergamon Press: Oxford.

Silberman E., Carter R., Einstein H.A., Hinds J. and Powell R. (1963) Friction Factors in Open Channels, Proceedings of the American Society Civil Engineers, Vol. 89, HY2.

Strickler A. (1923) *Beiträge zur Frage der Geschwindigkeitsformeln...*, Mitteilung No. 16, Amtes fur Wasserwirtschaft: Bern.

Wood I. (1985) Air water flows, *Proceedings of the XXI Congress, International Association for Hydraulic Research*, Vol. 6, Melbourne.

Yalin M.S. (1972) *Mechanics of Sediment Transport*, Pergamon Press: Oxford.

138: Unsteady Flow

WALTER H GRAF AND MUSTAFA S ALTINAKAR

Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

Flow is unsteady, if the flow depth as well as other hydraulic parameters vary with time. An unsteady flow is usually also nonuniform. The one-dimensional hydrodynamic equations, known are the equations of Saint-Venant, are developed. The simplified equations of Saint-Venant are used to treat the kinematic wave as well as the diffusive wave. The flood wave is also briefly discussed. Numerical solutions to unsteady flow will be treated in a subsequent chapter (see ***Chapter 139, Numerical Modeling of Unsteady Flows in Rivers, Volume 4****).*

HYDRODYNAMIC EQUATIONS

Equations of Saint-Venant

We (see Figure 1) will consider a channel with flow, having a free water surface, and that is one-dimensional, unsteady, nonuniform (gradually varied) and almost rectilinear. The bed slope, S_f, is fixed and permanent, but also weak; the discharge of the incompressible fluid is given by $Q = UA$, with $U(x, t)$ as the velocity averaged over the cross section, $A(x, t)$. There is no lateral inflow or outflow.

The *equation of continuity* is given (*see* **Chapter 136, Hydrodynamic Considerations, Volume 4**, Figure 1) by

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = A\frac{\partial U}{\partial x} + U\frac{\partial A}{\partial x} + B\frac{\partial h}{\partial t} = 0 \qquad (1)$$

For a rectangular, prismatic channel, the above equation reads

$$h\frac{\partial U}{\partial x} + U\frac{\partial h}{\partial x} + \frac{\partial h}{\partial t} = 0 \qquad B = Cte \qquad (2)$$

The *dynamic equation* for an unsteady and gradually varied flow is given (*see* **Chapter 136, Hydrodynamic Considerations, Volume 4**, Figure 3) by

$$\frac{1}{g}\frac{\partial U}{\partial t} + \frac{U}{g}\frac{\partial U}{\partial x} + \frac{\partial h}{\partial x} + \frac{\partial z}{\partial x} = -S_e \qquad (3)$$

where $S_f = -(\partial z/\partial x)$ is the bed slope, $h_r = S_e\,dx$ is the head loss, and S_e is the energy slope. As an approximation, it is assumed that the energy slope, S_e, can be expressed by a relation established for uniform, steady flow of the Weisbach-Darcy type (*see* **Chapter 137, Uniform Flow, Volume 4**):

$$S_e = \boldsymbol{f}\frac{1}{4R_h}\frac{U^2}{2g} \qquad (4)$$

or of the type of Chézy:

$$S_e = \frac{8g}{C^2}\frac{1}{4R_h}\frac{U^2}{2g} \qquad (5)$$

Where $\boldsymbol{f}$ is the friction coefficient, defined with equation 5 in **Chapter 137, Uniform Flow, Volume 4**. The friction coefficient of Chezy, C, is defined as

$$U = K_s R_h^{2/3} S_f^{1/2} \qquad \text{with} \qquad C = K_s R_h{}^{1/6} = \frac{1}{n} R_h{}^{1/6} \qquad (6)$$

where K_s is the coefficient of Strickler, n the coefficient Manning, and R_h, is the hydraulic radius.

For a river with a weak bed slope, S_f, the two terms due to acceleration can readily be neglected (see Henderson, 1966, p. 364). Consequently, one may write equation 3 as:

$$\frac{\partial h}{\partial x} + \frac{\partial z}{\partial x} = -S_e \qquad \text{where} \qquad -S_w = \frac{\partial h}{\partial x} + \frac{\partial z}{\partial x}$$
$$\text{or} \quad S_e = S_w \qquad (7)$$

If the variation of the flow depth is weak compared to the bed slope, $\partial h/\partial x < \partial z/\partial x$ – this may be the case in

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

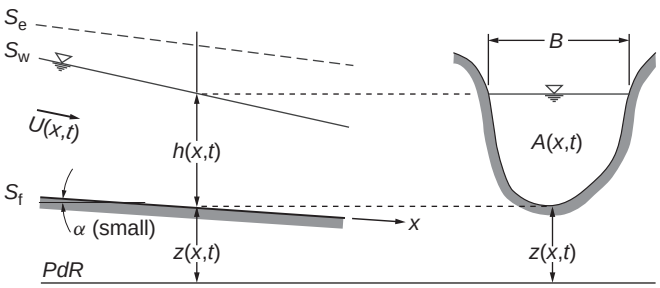


Figure 1 Scheme of an unsteady and nonuniform flow over a slope of a fixed bed, $z(x)$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

waterways of more or less steep bed slopes – the dynamic equation, equation (7), reduces to

$$\frac{\partial z}{\partial x} = -S_e \qquad \text{where} \qquad -S_f = \frac{\partial z}{\partial x} \qquad \text{or} \quad S_e = S_f \tag{8}$$

Using the relation of Chézy, the dynamic equation, equation (3), can also be expressed as

$$U = C\sqrt{R_h S_e} = C\sqrt{R_h\left(S_f - \frac{\partial h}{\partial x} - \frac{U}{g}\frac{\partial U}{\partial x} - \frac{1}{g}\frac{\partial U}{\partial t}\right)} \tag{9}$$

which for a steady and uniform flow becomes

$$U = C\sqrt{R_h S_f} \tag{10}$$

The relation between the velocity, U, or the discharge, Q, and the flow depth, h, or the hydraulic radius, R_h, (see Figure 2a) is given by

- a unique relation for equation (10),
- a nonunique relation with a *loop* for equation (9), where the width of the loop is an indication of the importance of the terms of inertia and of pressure.

The relation, $Q = f(h)$, (Figure 2a), called the *gauging (or rating) curve of the section*, will provide the following information (see Forchheimer, 1930):

- For unsteady flow, the discharge, Q, has two different values for the same flow depth, h, depending on the increase or decrease of the water level.
- If the flow depth, h, increases, the term $\partial h/\partial t$ is positive and the term $\partial U/\partial t$ is negative; consequently the velocity, U, decreases.
- At a given section, one observes (see Figure 2b) the maximum of the average velocity, U_{max}, then the maximum of the discharge, Q_{max}, and finally the maximum of the flow depth, h_{max}.

The equations, equation (1) and equation (3), established first by Saint-Venant (1870), render solutions to problems of unsteady flow, if integration is possible. They constitute a system of two equations of partial derivatives – of the hyperbolic type – in x and t, introducing two unknown functions, h and U. The initial and boundary conditions must be selected to describe adequately the physical problem. An exact integration of the equations of Saint-Venant is, however, very complicated; analytical solutions are rare (see Sobey, 2001, and MacDonald *et al.*, 1997). Nevertheless, there exist different numerical (and graphical) techniques making solutions possible.

It happens that a problem, where flow is unsteady and gradually varied, can be formulated in a way that one or more of the terms in the dynamic equation can be neglected. Considering the different terms in the dynamic equation, equation (3), one usually distinguishes the following types

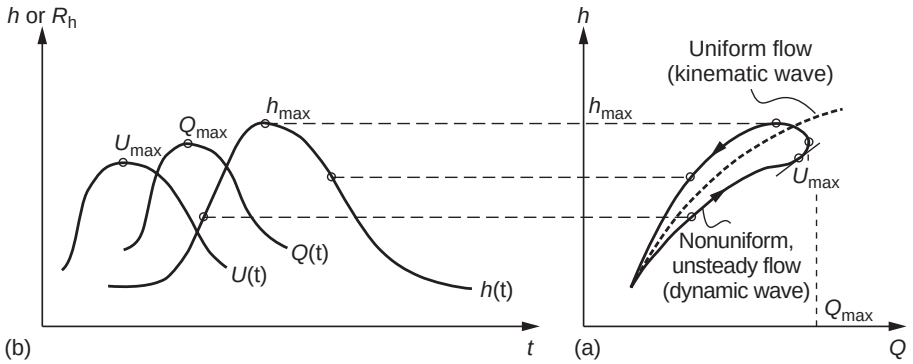


Figure 2 Schematic representation of the relation of $Q = f(h)$ and of $h = f(t)$ for an unsteady and nonuniform flow

of unsteady flow, here called *waves*:

$$\frac{1}{g}\frac{\partial U}{\partial t} + \frac{U}{g}\frac{\partial U}{\partial x} + \frac{\partial h}{\partial x} = S_f - S_e \qquad (3)$$

- kinematic wave (11)
- diffusive wave (12)
- dynamic, quasi-steady wave (13)
- dynamic wave (3)
- simple wave (14)

The equation of continuity, equation (1), remains of course valid for all these waves.

Numerical methods exist to solve the equations of Saint-Venant. The following ones could be mentioned:

- the method of characteristics,
- the explicit method,
- the implicit method.

Usually a *scheme of finite differences* – also schemes of *finite element* and *finite volume* – is employed. The numerical methods can well be dealt with the use of computers. These methods are presented in detail in the books of Liggett and Cunge (1975), Cunge *et al.* (1980) and Abbott and Basco (1989). Practical aspects dealing with fluvial hydraulics are treated by Cunge *et al.* (1980).

KINEMATIC WAVE

Hydrodynamic Equations

The kinematic wave represents a special and simple case of unsteady flow. The equations of Saint-Venant, equation (1) and equation (3), can be simplified, when considering the kinematic wave. The equation of continuity remains valid:

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = A\frac{\partial U}{\partial x} + U\frac{\partial A}{\partial x} + B\frac{\partial h}{\partial t} = 0 \qquad (1)$$

but the equation of motion can be considerably simplified:

$$S_f = S_e \qquad (11)$$

This implies, that in the complete equation of motion, equation (3), the terms due to the variations in velocity, U, and in depth, h, are negligible.

This equation, equation (11), can also be written, using the relation of Chézy, as

$$S_f = \frac{U^2}{C^2}\frac{1}{R_h} \qquad \text{or} \qquad U = C\sqrt{R_h S_f} \qquad (10)$$

The discharge is thus given by: $Q = UA = CA\sqrt{R_h S_f}$, implying that $Q = f(A)$ or $A = f(Q)$. Thus, a single-valued function (Figure 2a) exists between the discharge, Q, and the wetted surface, A, in a cross section at a given abscissa, $x = x_o$. The term, $\partial A/\partial t$, in above equation, equation (1), can be expressed as

$$\frac{\partial A}{\partial t} = \left(\frac{\partial A}{\partial Q}\right)_{x_o} \cdot \frac{\partial Q}{\partial t} \qquad (15)$$

Consequently, the equation of continuity, equation (1), can be written (see Forchheimer, 1930) as follows:

$$\frac{\partial Q}{\partial t} + \left(\frac{\partial A}{\partial Q}\right)_{x_o} \cdot \frac{\partial Q}{\partial x} = \frac{dQ}{dt} = 0 \qquad (16)$$

This is the equation of the *kinematic wave*, where the term

$$\left(\frac{\partial Q}{\partial A}\right)_{x_o} = c_k \qquad \text{or} \qquad c_k = -\frac{\partial Q/\partial t}{\partial Q/\partial x} \qquad (17)$$

is the *celerity of propagation* (or the *wave speed*) of the kinematic wave for a given discharge; it will be different for each section of abscissa, $x = x_o$. According to equation (17), the discharge, Q, is convectively displaced with the celerity, c_k.

The kinematic wave is a type of wave, whose properties follow essentially from the law of conservation of mass, namely, the equation of continuity, equation (1).

Since a variation in flow depth, $\partial h/\partial x$, (see equation 3 and equation 11) is not accounted for, the kinematic wave will not subside (attenuate). The geometric form of a kinematic wave, namely its wave profile – being of limited interest for the engineer – is given by Henderson (1966).

For a numerical simulation of kinematic waves, the methods presented in literature can readily be used

- the explicit method, as given by Jansen *et al.* (1979) and Abbott and Basco (1989);
- the implicit method, as given by Chow *et al.* (1988), Jansen *et al.* (1979) and Abbott and Basco (1989).

For the numerical simulation, one generally assumes (see Dingman, 1984), that the kinematic wave, equation (11), is a good approximation to the complete dynamic wave, equation (3), if

$$\frac{g\ L\ S_f}{U^2} > 10$$

where L is the length of the channel under investigation and, U is the average velocity of the uniform flow.

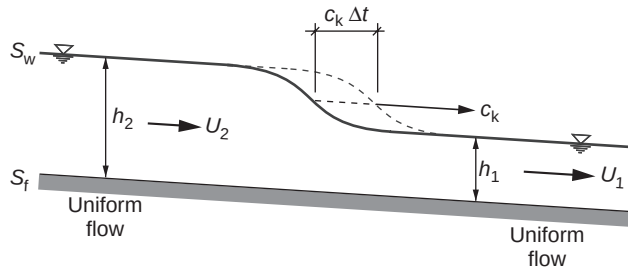


Figure 3 Monoclinal wave (front of a kinematic wave). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Celerity of Propagation

Take a wide rectangular channel (Figure 3) where the flow is uniform, having a flow depth, h_1, and an average velocity, U_1; the discharge, Q_1, is initially constant.

The flow increases suddenly to another discharge, Q_2; it stabilizes itself at another uniform flow, h_2, and U_2. This step-increase in discharge propagates downstream and forms a translatory wave of stable shape, called the *monoclinal* (rising) wave. During an interval of time, Δt, the wave displaces itself over a distance of $\Delta x = c_k \Delta t$. The celerity of this propagation is given as

$$c_k = \frac{Q_2 - Q_1}{A_2 - A_1} = \frac{dQ}{dA} = \frac{1}{B}\frac{dQ}{dh} \tag{18}$$

implying that the celerity of propagation of a kinematic wave is proportional to the rate of change of discharge, Q, with the surface area, A, or the flow depth, h. The above equation can also be expressed (see Lighthill and Whitham, 1955) by

$$c_k = \frac{d(UA)}{dA} = U + A\frac{\partial U}{\partial A} \tag{19}$$

This equation shows that

- in absence of an initial flow, $Q_1 = 0$, implying $U_1 = 0$ and $A_1 = 0$, one has

$$c_k = U_2 \tag{20}$$

- in presence of an initial flow, $Q_1 > 0$, one has

$$c_k > U_1, \quad c_k > U_2$$

The celerity of the kinematic wave, c_k, is always larger than the flow velocities, U_1 and U_2, upstream and downstream of the wave front.

If the flow in a wide rectangular channel is turbulent, the average velocity can be obtained, using the relation of Chézy, equation (10). The above equation (19), can thus been expressed (see Forchheimer, 1930) by

$$c_k = U + h\frac{\partial U}{\partial h} = U + h\frac{C S_f^{1/2}}{2h^{1/2}} = U + \frac{U}{2} = \frac{3}{2}U \tag{21}$$

For a parabolic or triangular channel, one gets, respectively

$$c_k = \frac{4}{3}U \qquad \text{or} \qquad c_k = \frac{5}{4}U \tag{22}$$

Should the same reasoning be done, by using the relation of Manning, equation (6), one obtains for a wide rectangular channel

$$c_k = \frac{5}{3}U \tag{23}$$

The celerity of propagation of a kinematic wave is 3/2 or 5/3 times the flow velocity, U. This conclusion was arrived at using only two relations, namely the equation of continuity, equation (1), and the momentum equation, equation (11), expressed with a relation of friction.

DIFFUSIVE WAVE

Hydrodynamic Equations

The diffusive wave represents another special and simple case of unsteady flow. The equations of Saint-Venant, equation (1) and equation (3), can be simplified, when considering the diffusive wave. While the equation of continuity remains valid,

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = A\frac{\partial U}{\partial x} + U\frac{\partial A}{\partial x} + B\frac{\partial h}{\partial t} = 0 \tag{1}$$

the equation of motion can be simplified:

$$S_f - \frac{\partial h}{\partial x} = S_e \tag{12}$$

This implies, that in the complete equation of motion, equation (3), the terms of inertia are negligible.

The diffusive wave undergoes attenuation (subsidence) due to a (possible) variation of the flow depth, $\partial h/\partial x$; this is not possible for the kinematic wave which cannot subside, but will deform, changing its curvature (see Figure 4).

Above equation, can be written (see equation 9), using the relation of Chézy, equation (10), as:

$$U = C\sqrt{R_h\left(S_f - \frac{\partial h}{\partial x}\right)} \tag{24}$$

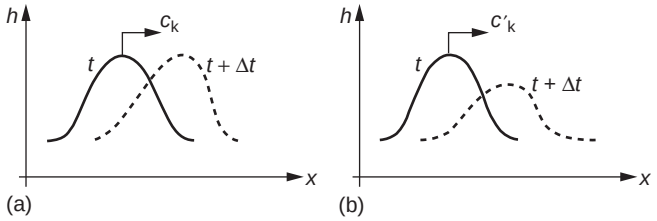


Figure 4 Attenuation of a wave; successive positions

The dynamic equation, equation (12), can also be written as

$$\frac{\partial h}{\partial x} + \frac{\partial z}{\partial x} + \frac{Q|Q|}{K^2} = 0 \tag{25}$$

whereby the conveyance of the channel is taken as

$$K(h) = \frac{1}{n} {R_h}^{2/3} A \tag{26}$$

which in turn gives $Q = K\sqrt{S_f}$ (*see* **Chapter 137, Uniform Flow, Volume 4**, equation 19).

Suppose the channel is rectangular of width $B = Cte$. Upon mathematical manipulation (see Graf and Altinakar, 1998, Section 5.4 and Cunge *et al.*, 1980) the above two equations, equation (1) and equation (25), render a single equation, such as

$$\frac{\partial Q}{\partial t} + \left(\frac{Q}{BK}\frac{\mathrm{d}K}{\mathrm{d}h}\right)\frac{\partial Q}{\partial x} - \left(\frac{K^2}{2B|Q|}\right)\frac{\partial^2 Q}{\partial x^2} = 0 \tag{27}$$

This is the equation of the *diffusive wave*. This partial differential equation is parabolic, having only one dependent variable, $Q(x,t)$. The discharge, Q, is convected with the celerity (see also equation 17) of

$${c_k}' = \left(\frac{Q}{BK}\frac{\mathrm{d}K}{\mathrm{d}h}\right) = \frac{1}{B}\frac{\mathrm{d}Q}{\mathrm{d}h} \tag{28}$$

and is diffused (or attenuated) with a coefficient of

$$C_D = \left(\frac{K^2}{2B|Q|}\right) = \frac{Q}{2BS_f} \tag{29}$$

The equation of the diffusive wave, equation (27), can now be written as

$$\frac{\partial Q}{\partial t} + {c_k}'\frac{\partial Q}{\partial x} - C_D\frac{\partial^2 Q}{\partial x^2} = 0 \tag{30}$$

Assuming that the bottom slope remains constant, Henderson (1966) gives the following relation (see also equation 21) for the celerity, using the relation of Chézy, or

$${c_k}' = \frac{3}{2}U = \frac{3}{2}C\sqrt{R_h\left(S_f - \frac{\partial h}{\partial x}\right)} \tag{31}$$

and for the coefficient of diffusion

$$C_D = \frac{{c_k}' h}{3[S_f - (\mathrm{d}h/\mathrm{d}x)]} \approx \frac{{c_k}' h}{3\ S_f} \tag{32}$$

For a numerical simulation of diffusive waves, the solution methods presented in the literature can readily be used:

- the explicit method, as given by Abbott and Basco (1989);
- the implicit method, as given by Abbott and Basco (1989) and Jansen *et al.* (1979).

The diffusion term in the above equation, equation (27), may be neglected if $\partial^2 Q/\partial x^2$ or/and $Q/(2BS_f)$ (see equation 29) are small. The last condition implies that

$$Q \ll 2BS_f \tag{33}$$

or, when using the Manning relation, equation (6), for a wide rectangular channel:

$$h^{5/3} \ll 2n{S_f}^{1/2} \tag{34}$$

The latter relation shows, that for bottom slopes, S_f, being

- sufficiently *steep*, the diffusion term can be eliminated; the diffusive wave is thus well approximated by a kinematic wave;
- *gentle*, the diffusion term must remain; the diffusive wave is maintained.

In steady state, the discharge, Q, can be calculated according to equation (24), which gives a non-single-valued relation, with a loop (see Figure 2a). The size of this loop is due to the diffusion coefficient, C_D (equation 29), in the equation of the diffusive wave, equation (27).

FLOOD WAVE

Flood waves are phenomena of great importance for the hydraulician. A flood is a type of unsteady flow, whose theoretical study must be deduced from the complete equations of Saint-Venant; methods of solution have been presented in the literature.

The displacement – rising and falling – of such a flood is generally very *slow*; thus certain terms in equation of

motion, equation (3), can be neglected. Consequently the equations of Saint-Venant can be simplified. Usually one makes a distinction (see Henderson, 1966) between

- floods in channels on *steep slopes*, Fr < 1;
- floods in channels on *weak slopes*, Fr $\ll 1$.

A flood wave is thus considered as being a slowly variable flow.

Floods in channels on *steep* slopes, Fr < 1, can be approximated by *kinematic waves* (see Section on "Kinematic Wave"). The *in situ* observations by Seddon in 1900 as well as theoretical considerations by Kleitz in 1858 have this confirmed; it is *known as* the principle of Kleitz-Seddon (see Chow, 1959).

One supposes that a flood rises gradually to a (unique) maximum, and subsequently descends till the initial uniform flow is reached again. Before the rise and after the fall of the flood, the flow is in permanent regime and has the same discharge in all cross sections. This wave, being a kinematic one, is characterized by

- a relation of $Q = f(h)$ being single-valued (Figure 2a);
- the celerity of propagation, given by:

$$c_k = \frac{3}{2}U = \frac{3}{2}C\sqrt{hS_f} \tag{21}$$

  for a wide rectangular section, using the relation of Chézy;
- a geometric form, being invariable and without attenuation.

For such type of a flood (see Figure 5), the major part – the main body of the flood – is well described by a kinematic wave, equation (18). In front and behind the main body of the flood wave, there travel gravity waves (*see* **Chapter 136, Hydrodynamic Considerations, Volume 4**, equation 20). According to Henderson (1966), if Fr $<$ $3/2$ (or Fr < 2), these gravity waves are usually of little importance, since they subside rapidly.

Floods in channels on *weak* slopes, Fr $\ll 1$, can be approximated by *diffusive waves*, being characterized (see Section on "Diffusive Waves") by

- a relation of $Q = f(h)$ that is not single-valued (Figure 2a);
- the celerity of propagation, given by

$$c_k' = \frac{3}{2}C\sqrt{h\left(S_f - \frac{\partial h}{\partial x}\right)} \tag{31}$$

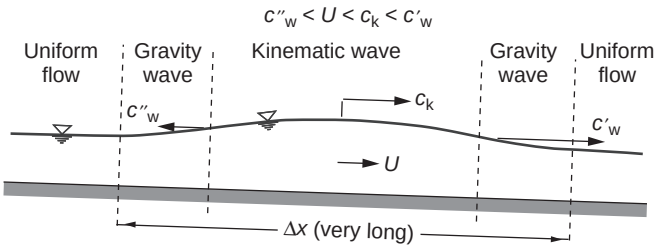


Figure 5 Flood wave (Fr $< 3/2$). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

where C is the Chézy coefficient;
- a geometric form, being variable; thus an attenuation is present.

Various aspects of such waves are presented by Henderson (1966).

REFERENCES

Abbott M.B. and Basco D. (1989) *Computational Fluid Dynamics*, Longman Science and Technology: Harlow.

Chow V.T. (1959) *Open Channel Hydraulics*, McGraw-Hill: New York.

Chow V.T., Maidment D. and Mays L. (1988) *Applied Hydrology*, McGraw-Hill: New York.

Cunge J., Holly F. and Verwey A. (1980) *Practical Aspects of Computational River Hydraulics*, Pitman Publication: London.

Dingman S. (1984) *Fluvial Hydrology*, W. Freeman and Company: New York.

Forchheimer P. (1930) *Hydraulik*, Teubner-Verlag: Leipzig.

Graf W.H. and Altinakar M.S. (1998; 2000; 2002) *Fluvial Hydraulics: Flow and Transport in Channels of Simple Geometry*, John Wiley & Sons: Chichester.

Henderson F.M. (1966) *Open Channel Flow*, Macmillan Company: New York.

Jansen P., Bendegom L., Berg J., de Vrier M. and Zanen A. (1979) *Principles of River Engineering*, Pitman Publications: London.

Liggett J. and Cunge J. (1975) Numerical methods of solution of the unsteady flow equations In *Unsteady Flow in Open Channels*, Mahmood K. and Yevdevich V. (Eds.), Water Resources Publications: Fort Collins.

Lighthill M.J. and Whitham G.B. (1955) On kinematic waves. *Proceedings, of the Royal Society of London. Series A*, **229**, 281–316.

MacDonald I., Baines M., Nicols N.K. and Samuels P. (1997) *Analytic Benchmark Solutions for Open-Channel Flows*, Vol. 123, Proceedings of the American Society of Civil Engineers: JHE11, Reston, VA.

Saint-Venant B. (1870) Demonstration elementaire de la formule de propagation d'un onde, *Compte Rendu des séance de l'Acamidie*, **7**, 186–195.

Sobey R.J. (2001) *Evaluation of Numerical Models of Flood and Tide Propagation in Channels*, Vol. 127, Proceedings of the American Society of Civil Engineers: JHE10, Reston, VA.

139: Numerical Modeling of Unsteady Flows in Rivers

BOMMANNA G KRISHNAPPAN[1] AND MUSTAFA S ALTINAKAR[2]

[1]Aquatic Ecosystem Impacts Research Branch, National Water Research Institute, Environment Canada, Canada Centre for Inland Waters, Burlington, ON, Canada
[2]National Center for Computational Hydroscience and Engineering, The University of Mississippi, MS, US

Solution methods to solve the unsteady flow equations are reviewed in this article. The basis of the method of characteristics is outlined. Some of the commonly used finite difference schemes are reviewed. A detailed description of an unsteady flow model called MOBED is given to highlight the various assumptions and simplifications that are involved in the development of an unsteady flow model. Testing of the model using a laboratory data set measured by Treske is described. Such a data set can serve as a benchmark data for the testing of unsteady flow models. Integral form of governing equations and their properties are introduced. High-resolution conservative finite-volume schemes for modeling flows with discontinuities are briefly discussed. Examples of numerical solutions using a robust one-dimensional upwind finite-volume code are presented.

GOVERNING EQUATIONS

In **Chapter 5, Fundamental Hydrologic Equations, Volume 1**, the governing equations for unsteady flows (i.e. the St. Venant equations) were presented, and their properties were analyzed. In this section, the numerical techniques that have been used to solve these equations are reviewed. The St. Venant equations describe the unsteady flow characteristics in channels, which are prismatic, and which do not have any lateral inflows. If the river is nonprismatic and contains lateral inflows, then the governing equations have to be modified to include additional terms to take into account these extra features. The numerical methods that are reviewed in this section are capable of treating these additional features. The governing equations that are discussed in this section are listed below (Krishnappan, 1981):

$$\frac{\partial Q}{\partial x} + B\frac{\partial h}{\partial t} = q \tag{1}$$

$$\frac{\partial Q}{\partial t} + 2\frac{Q}{A}\left(\frac{\partial Q}{\partial x}\right) - B\frac{Q^2}{A^2}\left(\frac{\partial h}{\partial x}\right) + gA\left(\frac{\partial h}{\partial x}\right) = gA(S_x - S_f) + q\left(U_q - \frac{Q}{A}\right) + \frac{Q^2}{A^2}A_x^y \tag{2}$$

where x is the longitudinal axis placed along the thalweg of the river, t is the time axis, Q is the flow rate, h is the flow depth, B is the top width of the river, A is the flow cross-sectional area, S_x is the slope of the river bed, S_f is the friction slope, q is the lateral inflow rate, U_q is the downstream component of velocity of the lateral inflow, g is the acceleration due to gravity, and A_x^y is the rate of change of A with respect to x when the distance from the bed is held constant. This derivative goes to zero for prismatic channels. A schematic representation of the profile and the cross section of a river that can be modeled using the above equations is shown in Figure 1.

The first equation expresses the conservation of mass while the second equation is based on the conservation of momentum. In the derivation of these equations, a number of simplifying assumptions have been introduced. A good understanding of the underlying assumptions is important to gauge the applicability and the usefulness of the equations to practical problems. For example, the equations were derived on the basis of first – order shallow

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

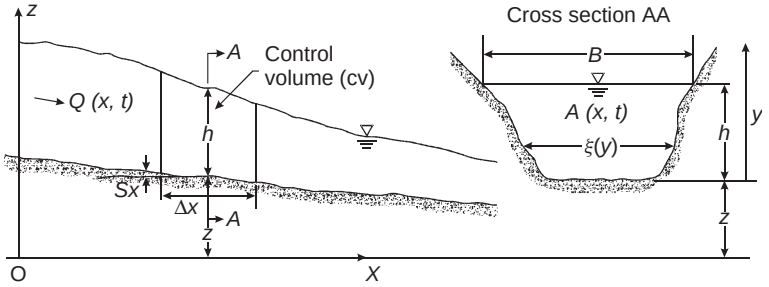


Figure 1 Schematic representation of the longitudinal profile and a flow cross section in a river reach considered for modeling

water theory with the assumption that the wave surface varies gradually. Such an assumption leads to a condition that the pressure distribution in the vertical direction is hydrostatic and the vertical accelerations are small. This implies that the equations are valid, strictly speaking, for long waves only. For the treatment of hydraulic jumps and bores, where the vertical accelerations are substantial, these features are treated as discontinuities of the water surface of infinitesimal length and use the moving hydraulic jump relations to link the upstream and downstream regions where the hydrostatic pressure distribution assumption is not violated (see Cunge *et al.*, 1980). Second assumption used in the derivation of the governing equations deals with the bed friction loss. It is assumed in the derivation that the friction loss in unsteady flows is not significantly different from those in the steady flow and the steady flow friction factor equations such as Chezy and Manning equations are used to evaluate the friction losses in unsteady flows. The other assumptions invoked were: (i) slope of the bed is small so that the sine of the angle that the x axis makes with the horizontal is equal to the tangent of the angle and therefore, the angle itself; (ii) velocity distribution within the cross section does not have appreciable impact on the wave propagation.

The governing equations are partial differential equations of hyperbolic type, and hence can be solved either by the method of characteristics or by approximating the equations using the finite difference equations. A review of these methods is presented in the following subsection.

SOLUTION METHODS

Method of Characteristics

In the method of characteristics, the two partial differential equations are replaced by a system of four ordinary differential equations. To illustrate this method, the simplest case of uniform rectangular channel flow without the friction and lateral inflow, and so on, is considered (see Abbott, 1975). In this case, the mass and momentum conservation equations become

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = 0 \tag{3}$$

and

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + g\frac{\partial h}{\partial x} = 0 \tag{4}$$

where u is the average flow velocity. Multiplying equation (3) by g and substituting $gh = c^2$, we get the following two equations:

$$2\frac{\partial c}{\partial t} + 2u\frac{\partial c}{\partial x} + c\frac{\partial u}{\partial x} = 0 \tag{5}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + 2c\frac{\partial c}{\partial x} = 0 \tag{6}$$

Adding equations (5) and (6) and subtracting equation (5) from equation (6), the following two equations are obtained:

$$\frac{\partial (u+2c)}{\partial t} + (u+c)\frac{\partial (u+2c)}{\partial x} = 0 \tag{7}$$

$$\frac{\partial (u-2c)}{\partial t} + (u-c)\frac{\partial (u-2c)}{\partial x} = 0 \tag{8}$$

Comparing the form of equation (7) and equation (8) with the total variation (total differential) of any quantity A that is a function of x and t, that is,

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{\partial A}{\partial t} + \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\partial A}{\partial x} \tag{9}$$

we get from 5.7,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = u + c \tag{10}$$

and

$$\frac{\mathrm{d}(u+2c)}{\mathrm{d}t} = 0 \tag{11}$$

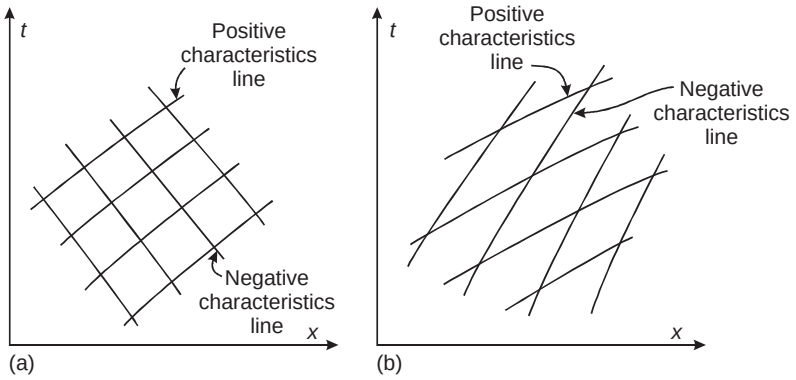


Figure 2 Schematic representation of characteristic lines for subcritical (a) and (b) supercritical flows

From equation (8), we get:

$$\frac{dx}{dt} = u - c \tag{12}$$

and

$$\frac{d(u - 2c)}{dt} = 0 \tag{13}$$

Equation (10) gives the slope (direction) of a line (characteristics) at a point in the $x-t$ plane, while equation (11) expresses the relationship between the dependent variables that is valid along the line. Similarly, equation (12) gives the direction of the second characteristic at the same point in the $x-t$ plane and equation (13) gives the relationship between the variables along this characteristics. Thus, the four ordinary differential equations (i.e. equations 10–13) are equivalent to the two partial differential equations (i.e. equations 3 and 4), and they provide a basis for solving for u and h as a function of x and t at all the intersection points of the characteristics in the $x-t$ plane.

In $x-t$ plane, the line passing through the directions specified by equation (10) is called the *positive characteristics* (see Figure 2), and the line passing through the directions specified by equation (12) is called *the negative characteristics*. For subcritical flows, $|u| < c$, the positive characteristics have positive slopes, and negative characteristics have negative slope (Figure 2a). For supercritical flows, u is greater than c, and therefore, both positive and negative characteristic lines have a positive slope as shown in Figure 2(b). The solution technique based on the method of characteristics involves constructing the positive and negative characteristic lines in the $x-t$ plane and determining the flow parameters u and h at the intersection points of these characteristics. At the boundaries, the boundary conditions needed to calculate the values at the boundary nodes depend on the number of characteristics emanating from the boundaries at a single point. For example, in subcritical flows at the upstream boundary, only one of the characteristics, namely the positive characteristics emanate from the boundary node at a point and the negative characteristics intersecting the boundary at that point emanate from the downstream sections of the flow. Therefore, one boundary condition involving either the flow velocity or the flow depth has to be specified at the upstream boundary. Similarly, at the downstream boundary only one condition has to be specified involving flow velocity or the flow depth. In the case of supercritical flows flowing from left to right, both positive and negative characteristics can emanate from a single point on the upstream boundary. Therefore, two boundary conditions involving both the flow velocity and the flow depth have to be specified at the upstream boundary. There is no need to specify a condition at the downstream boundary for the supercritical flows.

When the governing equations include additional terms to reflect the complexities such as irregular cross sections, nonprismatic channel, friction, lateral inflows, and so on, such as in equations (1 and 2), Liggett (1968) had proposed a generalized method to transform the governing equations into the characteristic form. Liggett (1968) had also considered the off channel storage in his derivation of the characteristic equations. Liggett (1968) solved the characteristic equations using numerical methods because of the complex nature of the equations. The points at which the equations are solved are spaced at uneven time and space intervals. This fact has often been used as a criticism of the method, but in some cases it is an advantage because the net becomes denser in the region of rapid change, where more points need to be calculated. Solution methods based on the method of characteristics had been proposed by other investigators as well. The notable ones are Stoker (1957), Fletcher and Hamilton (1967), and Lai (1965).

Finite Difference Approximations

The numerical methods based on finite difference approximations of the governing equations allow the computation of the flow variables (flow velocity and depth) at fixed grid points in the $x-t$ plane. There are two types of finite difference schemes, namely, Explicit Schemes and Implicit

Schemes. In Explicit Schemes, the finite difference equations are set up in such a way that the solution at a single point in the $x-t$ plane is calculated at a time. In the Implicit Schemes, on the other hand, solutions at all points at a new time level are calculated simultaneously knowing the solutions at the previous time step. Some of the commonly used finite difference schemes are reviewed here. For a comprehensive description of the available schemes, the references Liggett and Cunge (1975) and Cunge *et al.* (1980) can be consulted.

Explicit Schemes:

1. *The leap-frog scheme:*

This scheme was a widely used explicit scheme. It employs the central difference approximations for both time and distance derivative terms as shown below:

$$\frac{\partial f}{\partial t} = \frac{f_i^{j+1} - f_i^{j-1}}{2\Delta t} \tag{14}$$

$$\frac{\partial f}{\partial x} = \frac{f_{i+1}^{j} - f_{i-1}^{j}}{2\Delta x} \tag{15}$$

The function $f(x,t)$ is approximated as follows:

$$f(x,t) \approx f_i^j \tag{16}$$

The subscript i in the above equations denotes the numerical grid points in the x direction and the superscript j denotes the numerical grids along the time axis. Substitution of the equations (14) to (16) in the St. Venant equations results in algebraic equations expressing the values of the variables at a single grid point in the new time level in terms of the variables in the old time level in the vicinity of the grid, and hence can be solved for the variables in the new time level explicitly without having to solve a system of simultaneous equations.

Stability and accuracy of the numerical schemes determine the suitability of the schemes for practical applications. The stability analysis for the leap-frog scheme was carried out by Liggett and Cunge (1975) for the linearized equation without the resistance term using the spectral method. This analysis indicated that the scheme is stable only when the Courant condition,

$$\frac{\Delta t}{\Delta x} \leq \frac{1}{|u \pm c|} \tag{17}$$

is satisfied. If the Courant condition is not satisfied, then the perturbations grow indefinitely making the scheme unsuitable for computational purposes. The analysis further showed that when the LHS of equation (17) is *equal* to the RHS, then the amplification factor of the scheme is unity implying that the perturbations are neither amplified or damped, and the perturbations are propagated with the celerity of the analytical solution, that is, without dispersion. But, when the LHS is less than the RHS, the scheme is stable but dispersive.

2. *Lax scheme:*

This scheme was applied to the homogeneous part of the St. Venant equation expressed in the conservation form. The derivation of the St. Venant equation in the conservation form is given in Cunge *et al.* (1980). The conservation form of the St. Venant equation is specially suited for tackling discontinuities. Details of modeling discontinuities are taken up in Section "Modeling Flows With Discontinuities". The homogeneous part of the conservation form of St. Venant equation is given by

$$\frac{\partial f}{\partial t} + \frac{\partial F(f)}{\partial x} = 0 \tag{18}$$

where $f(x,t)$ is the vector of the variables and F is the flux vector. Lax scheme approximates the derivatives as follows:

$$\frac{\partial f}{\partial t} \approx \frac{f_i^{j+1} - \left[\alpha f_i^j + (1-\alpha)\dfrac{f_{i+1}^j + f_{i-1}^j}{2}\right]}{\Delta t} \tag{19}$$

$$\frac{\partial F(f)}{\partial x} \approx \frac{F_{i+1}^j - F_{i-1}^j}{2\Delta x}; \qquad 0 \leq \alpha \leq 1 \tag{20}$$

Substitution of equations (19) and (20) into equations (18) gives an explicit expression for the unknown vector function *f(x,t)* at the computational points. This scheme yields exact solution of the linearized equation when α is zero and the time and distance steps satisfy the Courant condition. The scheme is first – order accurate for arbitrary values of α, Δx, and Δt.

Implicit Schemes:

1. *Preissmann implicit scheme:*

This has been a very popular scheme. It has been implemented in many operational models such as DWOPER/DAMBRK (US-NWS), CARIMA (SOGREAH), and MOBED (NWRI-EC). In this scheme, the dependent variable and its derivatives are approximated as follows:

$$f(x,t) = \frac{\theta}{2}\left[f_{i+1}^{j+1} - f_i^{j+1}\right] + \frac{1-\theta}{2}\left[f_{i+1}^{j} + f_i^{j}\right] \tag{21}$$

$$\frac{\partial f}{\partial x} = \theta\left[\frac{f_{i+1}^{j+1} - f_i^{j+1}}{\Delta x}\right] + (1-\theta)\left[\frac{f_{i+1}^{j} - f_i^{j}}{\Delta x}\right] \tag{22}$$

$$\frac{\partial f}{\partial t} = \frac{1}{2}\left[\frac{f_{i+1}^{j+1} - f_{i+1}^{j}}{\Delta t}\right] + \frac{1}{2}\left[\frac{f_i^{j+1} - f_i^{j}}{\Delta t}\right] \tag{23}$$

where θ is a weighting coefficient that can assume values from 0 and 1. Substitution of these approximations into

the St. Venant equations gives rise to a system of algebraic equations, which have to be solved simultaneously. When θ is equal to 0, the scheme becomes fully explicit, and when θ is equal to 1, the scheme is fully implicit. The stability and the accuracy of the scheme had been investigated by Cunge (1961), who concluded that the scheme is unconditionally stable for θ values between 0.5 and 1.0 and the accuracy is second order with respect to Δx when θ is equal to 0.5 and first – order for any other value between 0.5 and 1.0. Cunge (1961) further observed that for θ, some oscillations occurred in the solution resembling the phenomenon of numerical instability, and suggested a practical range for θ to be between 0.6 and 1.0. With respect to the computational celerity, it was found to differ from the analytical celerity and the deviation was a function of θ, $\Delta x/L$ and $\Delta t/\Delta x$. When $\Delta x/L$ is small (long waves), the deviation is small as long as $\Delta t/\Delta x$ is closer to 1.

2. *Gunaratnam–Perkins scheme:*

This is a six-point scheme developed by Gunaratnam and Perkins (1970). In this scheme, the derivatives of the variables are approximated as follows:

$$\frac{\partial f}{\partial t} \approx \frac{1}{6}\frac{f_{i-1}^{j+1} - f_{i-1}^{j}}{\Delta t} + \frac{2}{3}\frac{f_{i}^{j+1} - f_{i}^{j}}{\Delta t} + \frac{1}{6}\frac{f_{i+1}^{j+1} - f_{i+1}^{j}}{\Delta t} \tag{24}$$

$$\frac{\partial f}{\partial x} \approx \frac{f_{i+1}^{j+1} - f_{i-1}^{j+1}}{2\Delta x} \tag{25}$$

The scheme, therefore, is fully implicit, and it involves six nodes, namely, three consecutive distance nodes in two consecutive time levels. Substitution of the approximations in the St. Venant equations leads to system of $2N-4$ algebraic equations for $2N$ unknowns, where N is the total number of grid points. With two boundary conditions and two characteristic equations written for the boundary points, the system of equations becomes a closed system and can be solved for the unknowns simultaneously.

AN EXAMPLE OF AN UNSTEADY FLOW MODEL AND ITS TESTING

An unsteady flow model called *MOBED* developed by the first author is described here in detail to highlight the various assumptions and simplifications that have to be introduced in the development of an unsteady flow model. The description also includes a testing of the model using laboratory channel data that can serve as benchmark data for testing of unsteady flow models.

Description of the MOBED Model

The MOBED model is based on the governing equations (1) and (2). These equations were discretized using the implicit finite difference scheme of Preissmann (1961), that is, using equations (21) to (23). The model, therefore, cannot handle discontinuities and is valid only for subcritical flows. The value of θ used in the model was 0.667.

Expressing the value of f at $(j+1)$th time level as a sum of the value at the jth time level and a difference Δf, that is,

$$f^{j+1} = f^{j} + \Delta f \tag{26}$$

the relationships given by equations (21)–(23) can be rewritten as:

$$f(x,t) = \frac{1}{2}[\theta(\Delta f_{i+1} + \Delta f_i) + (f_{i+1} + f_i)] \tag{27}$$

$$\frac{\partial f}{\partial x} = \frac{1}{\Delta x}[\theta(\Delta f_{i+1} - \Delta f_i) + (f_{i+1} - f_i)] \tag{28}$$

$$\frac{\partial f}{\partial t} = \frac{1}{2\Delta t}[\Delta f_{i+1} + \Delta f_i] \tag{29}$$

The superscript for f in the equations (27)–(29) was dropped because all values of f correspond to the time level j. When these equations were substituted in the continuity equation, that is, equation (1), and after linearization we get:

$$a_i \Delta h_{i+1} + b_i \Delta Q_{i+1} = c_i \Delta h_i + d_i \Delta Q_i + e_i \tag{30}$$

where

$$a_i = \left[\frac{B_{i+1} + B_i}{2\Delta t}\right] - \frac{2\theta}{\Delta x}\left[\frac{Q_{i+1} - Q_i}{B_{i+1} + B_i}\right]\frac{\mathrm{d}B}{\mathrm{d}h_{i+1}} + \theta\left[\frac{q_{i+1} + q_i}{B_{i+1} + B_i}\right]\frac{\mathrm{d}B}{\mathrm{d}h_{i+1}} \tag{31}$$

$$b_i = \frac{2\theta}{\Delta x} \tag{32}$$

$$c_i = -\left[\frac{B_{i+1} + B_i}{2\Delta t}\right] + \frac{2\theta}{\Delta x}\left[\frac{Q_{i+1} - Q_i}{B_{i+1} + B_i}\right]\frac{\mathrm{d}B}{\mathrm{d}h_i} - \theta\left[\frac{q_{i+1} + q_i}{B_{i+1} + B_i}\right]\frac{\mathrm{d}B}{\mathrm{d}h_i} \tag{33}$$

$$d_i = \frac{2\theta}{\Delta x} \tag{34}$$

$$e_i = -\frac{2}{\Delta x}[(Q_{i+1} - Q_i) + (q_{i+1} + q_i)] + \theta[(\Delta q_{i+1} + \Delta q_i)] \tag{35}$$

When equations (27)–(29) were substituted in the momentum equation and the resulting equation linearized, we get:

$$a_i' \Delta h_{i+1} + b_i' \Delta Q_{i+1} = c_i' \Delta h_i + d_i' \Delta Q_i + e_i' \tag{36}$$

where

$$a_i' = \frac{\theta}{\Delta x}\left[\frac{Q_{i+1}B_{i+1}}{A_{i+1}^2}(Q_i - Q_{i+1})\right] + \frac{g\theta}{\Delta x}[(h_{i+1} - h_i)B_{i+1} + A_{i+1} + A_i] + \frac{\theta}{2\Delta x}\left[\left(\frac{2B_{i+1}^2Q_{i+1}^2}{A_{i+1}^3} - \frac{Q_{i+1}^2}{A_{i+1}^2}\frac{\mathrm{d}B}{\mathrm{d}h_{i+1}}\right) \times (h_{i+1} - h_i) - \frac{B_{i+1}Q_{i+1}^2}{A_{i+1}^2} - \frac{B_iQ_i^2}{A_i^2}\right] - \frac{g\theta}{2\Delta x} \times [B_{i+1}(z_i - z_{i+1})] + \frac{g\theta}{2}const\left(\frac{R_{i+1}}{D_{65}}\right)^m Fr_{i+1}^n \times \left(B_{i+1}(m - 3n) - \frac{A_{i+1}}{P_{i+1}}\frac{\mathrm{d}P}{\mathrm{d}h_{i+1}}(m - n) + B_{i+1}\right) + \theta\frac{Q_{i+1}^2B_{i+1}}{A_{i+1}^3}A_{x\ i+1}^y \quad (37)$$

$$b_i' = \frac{1}{2\Delta t} + \frac{\theta}{\Delta x}\left[\frac{Q_{i+1}}{A_{i+1}} + \frac{Q_i}{A_i} + \frac{(Q_{i+1} - Q_i)}{A_{i+1}}\right] + \frac{\theta}{2\Delta x}\left[\frac{2Q_{i+1}B_{i+1}}{A_{i+1}^2}(h_i - h_{i+1})\right] + gn\theta const\left(\frac{R_{i+1}}{D_{65}}\right)^m \times Fr_{i+1}^n\frac{A_{i+1}}{Q_{i+1}} - \theta\frac{Q_{i+1}}{A_{i+1}^2}A_{x\ i+1}^y \quad (38)$$

$$c_i' = \frac{\theta}{\Delta x}\left[\frac{Q_iB_i}{A_i^2}(Q_i - Q_{i+1})\right] + \frac{g\theta}{2\Delta x}[(h_{i+1} - h_i)B_i - A_{i+1} - A_i] + \frac{\theta}{2\Delta x}\left[\left(\frac{2B_i^2Q_i^2}{A_i^3} - \frac{Q_i^2}{A_i^2}\frac{\mathrm{d}B}{\mathrm{d}h_i}\right) \times (h_{i+1} - h_i) + \frac{B_{i+1}Q_{i+1}^2}{A_{i+1}^2} + \frac{B_iQ_i^2}{A_i^2}\right] - \frac{g\theta}{2\Delta x} \times [B_i(z_i - z_{i+1})] + \frac{g\theta}{2}const\left(\frac{R_i}{D_{65}}\right)^m Fr_i^n\left(B_i(m - 3n) - \frac{A_i}{P_i}\frac{\mathrm{d}P}{\mathrm{d}h_i}(m - n) + B_i\right) + \theta\frac{Q_i^2B_i}{A_i^3}A_{x\ i}^y \quad (39)$$

$$-d_i' = \frac{1}{2\Delta t} + \frac{\theta}{\Delta x}\left[\frac{(Q_{i+1} - Q_i)}{A_i} - \frac{Q_{i+1}}{A_{i+1}} - \frac{Q_i}{A_i}\right] + \frac{\theta}{2\Delta x} \times \left[\frac{2Q_iB_i}{A_i^2}(h_i - h_{i+1})\right] + gn\theta const\left(\frac{R_i}{D_{65}}\right)^m \times Fr_i^n\frac{A_i}{Q_i} - \theta\frac{Q_i}{A_i^2}A_{x\ i}^y \quad (40)$$

$$-e_i' = \frac{1}{\Delta x}\left[\left(\frac{Q_{i+1}}{A_{i+1}} + \frac{Q_i}{A_i}\right)(Q_{i+1} - Q_i)\right] + \frac{g}{2\Delta x} \times [(A_{i+1} + A_i)(h_{i+1} - h_i)] + \frac{1}{2\Delta x}\left[\left(\frac{B_{i+1}Q_{i+1}^2}{A_{i+1}^2} + \frac{B_iQ_i^2}{A_i^2}\right)(h_i - h_{i+1})\right] + \frac{g}{2\Delta x}[(A_{i+1} + A_i)(z_{i+1} - z_i)] + \frac{g}{2}const\left(\frac{R_{i+1}}{D_{65}}\right)^m Fr_{i+1}^n A_{i+1} + \frac{g}{2}const\left(\frac{R_i}{D_{65}}\right)^m \times Fr_i^n A_i - \frac{1}{2}\frac{Q_{i+1}^2}{A_{i+1}^2}A_{x\ i+1}^y - \frac{1}{2}\frac{Q_i^2}{A_i^2}A_{x\ i}^y \quad (41)$$

In deriving the above relations, the following linearization expressions were utilized:

$$\frac{1}{f_i + \Delta f_i} = \frac{1}{f_i\left(1 + \frac{\Delta f_i}{f_i}\right)} \approx \frac{1}{f_i}\left[1 - \frac{\Delta f_i}{f_i}\right] \quad (42)$$

$$\frac{1}{(f_i + \Delta f_i)^2} = \frac{1}{f_i^2\left(1 + \frac{\Delta f_i}{f_i}\right)^2} \approx \frac{1}{f_i^2}\left(1 - \frac{2\Delta f_i}{f_i}\right) \quad (43)$$

$$(f_i + \Delta f_i)^2 \approx f_i^2 + 2f_i\Delta f_i \quad (44)$$

The friction slope appearing in equations (2) was evaluated using a generalized friction factor relationship proposed by Krishnappan (1985) and it is valid for both rigid and mobile boundary flows. This form of the equation is as follows:

$$S_f = const.\left(\frac{R}{D_{65}}\right)^m\left(\frac{U^2}{gR}\right)^n \quad (45)$$

For rigid boundary flows considered in this review, the value of the first exponent m becomes zero and the value of the second exponent n is equal to one. The value of the const is related to the Chezy's coefficient. The finite difference form of S_f, that is, ΔS_f was evaluated as follows:

$$\Delta S_{f_i} = \frac{2nS_{f_i}\Delta Q_i}{Q_i} + \left[S_{f_i}\frac{B_i}{A_i}(m - 3n) - S_{f_i}\frac{1}{P_i}\frac{\mathrm{d}P}{\mathrm{d}h_i}(m - n)\right]\Delta h_i \quad (46)$$

For any time level, the coefficients appearing in equations (30) and (36) can be evaluated using equations (31)–(35) and (37)–(41) knowing the flow properties Q and h, geometric parameters A, P, R, $\mathrm{d}B/\mathrm{d}h$ and $\mathrm{d}P/\mathrm{d}h$, and the friction parameters const, m and n. Therefore, when equations (30) and (36) are applied to the first $(N-1)$ grid points along the length of the river (N is the total number of grid points) is a system of two $(N-1)$ linear equations involving $2N$ unknowns, namely, ΔQ_1, ΔQ_2, ΔQ_3, …, ΔQ_N and Δh_1, Δh_2, Δh_3, …, Δh_N result. With two boundary conditions (one at the upstream boundary and the other at the downstream boundary in the case of subcritical

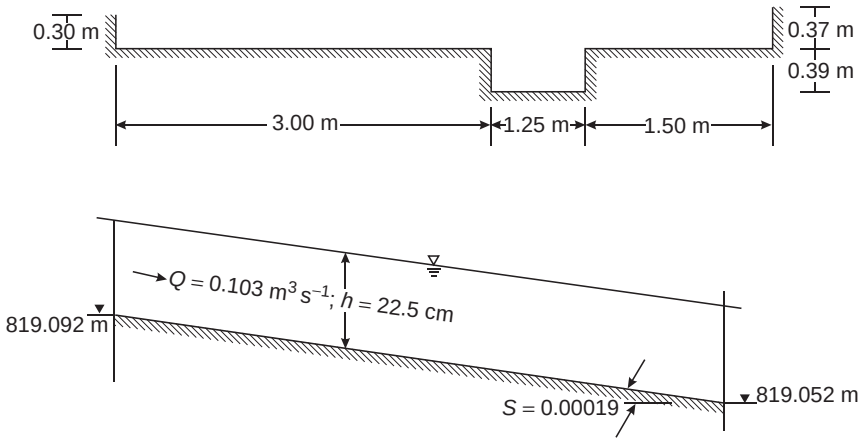


Figure 3 Schematic representation of a cross section and a profile of the compound channel used in Treske's experiments (1980).

flows) the number of equations matches the number of unknowns and hence the system of algebraic equations can be solved to obtain the solution at the next time level. The system of linear algebraic equations was solved using the double sweep method in the MOBED model. The details of the method can be found in Krishnappan (1981).

Testing of MOBED Model

The MOBED model was tested using unsteady flow experiments carried out by Treske (1980) in a laboratory channel. The channel used by Treske is a straight compound channel. The channel is 224 m long. The cross section and the profile of the channel are shown schematically in Figure 3. The main channel section is 39.0 cm deep and 125.0 cm wide. The floodplains on either side of the main channel are of different dimensions. The right side (looking downstream) floodplain is 150.0 cm wide and 37.0 cm deep whereas the left side floodplain is 300.0 cm wide and 30 cm deep. The working length of the channel is 210 m. The slope of the channel bed is 0.019% and the Manning's roughness coefficient was estimated to be 0.012 from uniform flow experiments. Unsteady flow experiments were carried out

Table 1 Measured data for test no. 1

Time in min	Upstream stage in mm	Upstream flow in l s^{-1}	Downstream stage in mm	Downstream flow in l s^{-1}
1	318	103	277	103
2	318	103	277	103
3	318	103	277	103
4	318	103	277	103
5	318	103	277	103
6	324	111	277	103
7	330	124	279	105
8	338	134	285	109
9	346	146	292	114
10	357	157	302	121
11	368	170	312	128
12	380	183	323	137
13	386	186	334	145
14	388	175	346	154
15	388	164	351	158
16	387	153	351	158
17	385	142	348	156
18	379	131	344	153
19	370	119	339	149
20	361	107	331	143
21	357	105	320	134

(*continued overleaf*)

Table 1 *continued*

Time in min	Upstream stage in mm	Upstream flow in l s^{-1}	Downstream stage in mm	Downstream flow in l s^{-1}
22	353	106	312	128
23	348	105	308	125
24	343	106	304	122
25	339	105	301	120
26	335	105	296	117
27	333	104	294	115
28	332	104	292	114
29	330	104	290	112
30	328	104	288	111
31	326	104	287	110
32	325	104	286	110
33	324	104	285	109
34	324	104	284	108
35	323	104	283	107
36	322	104	282	107
37	322	104	281	106
38	322	104	281	106
39	321	104	280	105
40	321	104	280	105
41	320	104	279	105
42	319	104	279	105
43	319	104	279	105
44	319	104	279	105
45	319	104	278	104
46	319	103	278	104
47	319	104	278	104
48	319	103	278	104
49	319	104	278	104
50	319	103	278	104

Table 2 Measured data for test no. 10

Time in min	Upstream stage in mm	Upstream flow in l s^{-1}	Downstream stage in mm	Down stream flow in l s^{-1}
0	309	96	267	96
6	309	95	267	96
12	324	114	277	103
18	349	135	302	121
24	377	156	330	142
30	403	177	357	163
36	429	198	382	183
42	454	220	407	204
48	477	241	431	224
54	493	261	446	237
60	505	282	453	247
66	519	304	464	264
72	533	325	478	286
78	545	346	492	309
84	556	368	507	335
90	567	389	520	358
96	577	411	533	382
102	580	402	540	394
108	579	382	539	393
114	573	361	534	383
120	566	340	525	367
126	558	318	516	357
132	548	297	505	332
138	539	276	494	313
144	528	255	482	293

Table 2 *continued*

Time in min	Upstream stage in mm	Upstream flow in $l\,s^{-1}$	Downstream stage in mm	Down stream flow in $l\,s^{-1}$
150	516	233	470	273
156	504	212	457	253
162	488	193	445	236
168	450	170	412	208
174	410	148	373	176
180	375	126	339	149
186	343	104	307	125
192	324	99	285	109
198	317	98	277	103
204	313	98	272	100
210	312	98	271	99
216	311	98	269	98

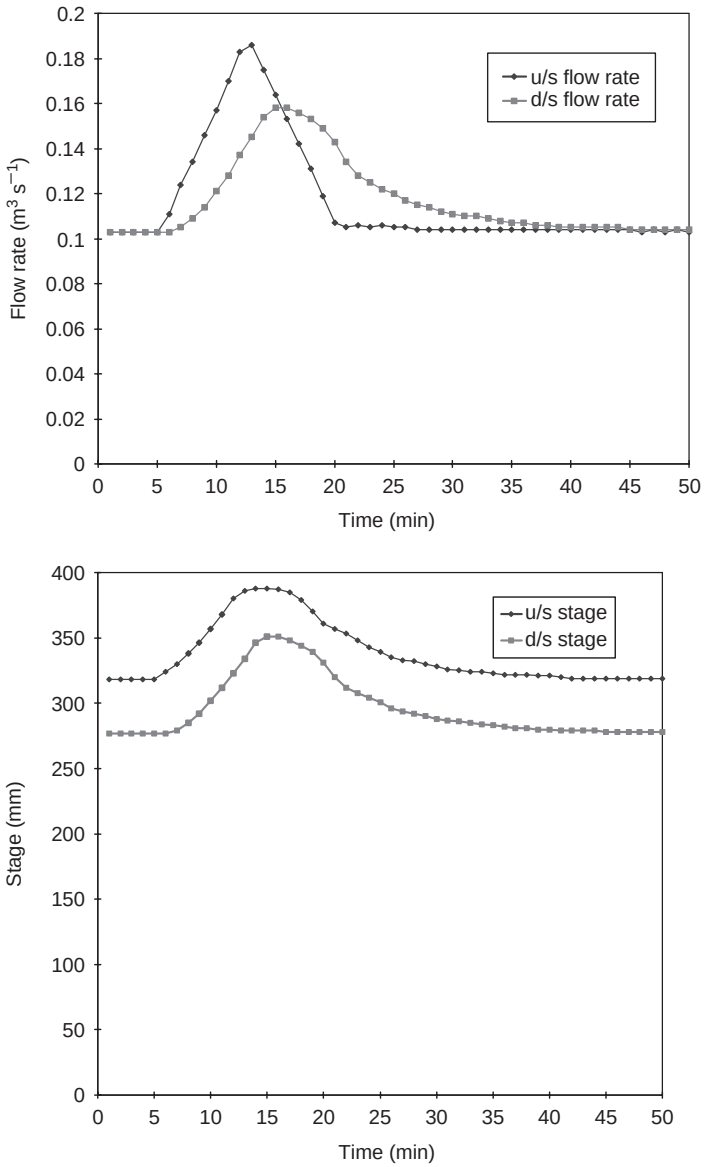


Figure 4 Measured data for test no. 1. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

by first running a steady flow for a period of time and then introducing a predetermined flow hydrograph at the head box. There were two measurement stations. The first one was at an upstream station located at 14.0 m from the head box and the second one was at the downstream end that was located at 224.0 m from the head box. At each measurement station, the flow rate and the flow stage were measured as a function of time. Altogether, 10 different hydrographs were tested in this channel. Out of these, four hydrographs produced flows that were confined to the main channel and the remaining six hydrographs produced flows in the floodplains. For testing the MOBED model two hydrographs were selected. The first one was a main channel flow (test no. 1) and the second one was a compound channel flow (test no. 10). The measured data for these two tests are shown in Figures 4 and 5, and listed in Tables 1 and 2.

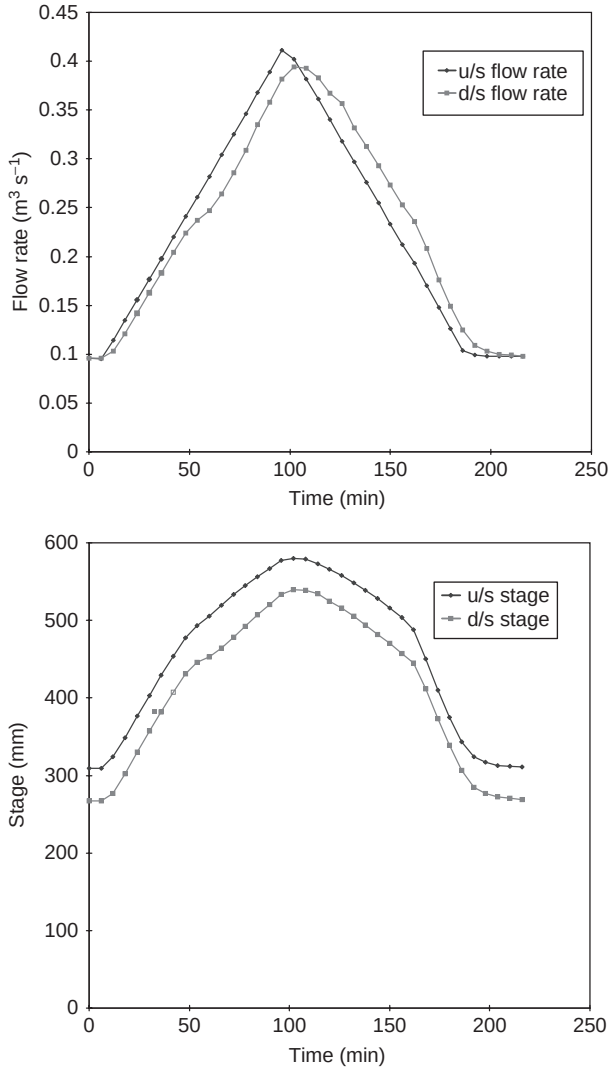


Figure 5 Measured data for test no. 10. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

For simulating the experimental conditions in the model, the reach between the two measurement stations was modeled. The flow rate measured at the upstream station was used as the upstream boundary condition and a stage-discharge relationship established from the measured data at the downstream boundary was used as the downstream boundary condition. This type of boundary condition specification is appropriate for the flows tested as the regime of the flow is subcritical at all times. For test no. 1, a time step of one minute and a distance step of 10 m were used. The stage-discharge relationship was expressed in the following functional form:

$$Q\left(\frac{\mathrm{m}^3}{\mathrm{sec}}\right) = 0.740h(\mathrm{m}) - 0.064 \tag{47}$$

For test no. 10, a time step of three minutes and a distance step of 10 m were used. For this test, a composite stage-discharge relationship established from the measured data was used. The forms of the relationships are shown below:

$$Q\left(\frac{\mathrm{m}^3}{\mathrm{sec}}\right) = 0.833h(\mathrm{m}) - 0.090 \quad \text{for} \quad h < 0.39 \tag{48}$$

$$Q\left(\frac{\mathrm{m}^3}{\mathrm{sec}}\right) = 1.583h(\mathrm{m}) - 0.383 \quad \text{for} \quad h > 0.39 \tag{49}$$

Results

Figure 6 shows a comparison of MOBED's prediction of upstream stage and the measured stage as a function of time for test no. 1. As can be seen from this figure, the model

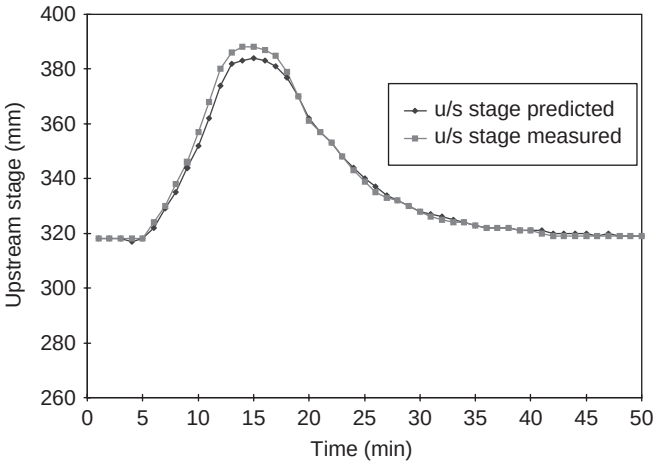


Figure 6 Comparison of predicted and measured upstream stage for test no. 1. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

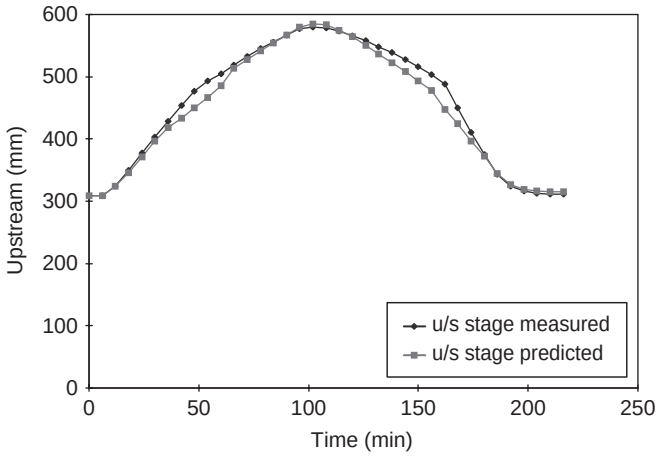


Figure 7 Comparison of predicted and measured upstream stage for test no. 10. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

prediction agrees reasonably well with the measurement, except for the peak stage. The model underestimates the peak stage by about 1%. It should be pointed out that for this test in which the flow is confined to the main channel, no adjustment of the friction factor was made. The measured friction factor under uniform flow conditions was used in the model without any adjustment. Under such condition, the prediction of peak stage within 1% can be considered as a significant result.

Figure 7 shows the agreement between the MOBED's prediction and the measurement for test no. 10. In this test, the flow spilled over into the floodplains when the stage exceeded 482 mm. When simulating the floodplain flow, the friction factor of the flow during the period when the flow was in the floodplain was increased by 25%. Such an increase was necessary to match the predicted peak stage with the measured peak stage, but it failed to match the stages during the rising and falling near the transition. Nevertheless, the agreement between the model prediction and the measurement can be considered satisfactory.

MODELING FLOWS WITH DISCONTINUITIES

Integral (Conservative) Forms of Governing Equations

The governing equations, equations (1) and (2) as well as equations (3) and (4), are written in a differential form. They are only valid in the case of smooth solutions and do not accept discontinuities. These equations, therefore, cannot be used to model flows involving mixed flow regimes of subcritical and supercritical flows, hydraulic jumps, and translatory waves, such as bores and dam-break fronts.

The governing equations for unsteady flows (i.e. the St. Venant equations), however, can be cast in an integral form that accepts discontinuous solutions. To illustrate the ideas for numerical modeling of discontinuities in unsteady flows, let us rewrite the differential form of simplified one-dimensional unsteady open-channel flow equations for rectangular channels (equations 3 and 4) in a slightly different form:

$$\frac{\partial h}{\partial t} + \frac{\partial hU}{\partial x} = q_\lambda \tag{50a}$$

$$\frac{\partial hU}{\partial t} + \frac{\partial}{\partial x}\left(hU^2 + \frac{1}{2}gh^2\right) = -gh(S_o - S_e) \tag{50b}$$

where h is the water depth, U is the average velocity, q_λ is the lateral discharge, and S_o and S_e represent the bed slope and slope of the energy line. These equations can be conveniently cast in a vector form as follows:

$$\mathbf{\Phi}_t + \mathbf{F}(\mathbf{\Phi})_x = \mathbf{S}(\mathbf{\Phi}) \tag{51a}$$

with

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \begin{bmatrix} h \\ hU \end{bmatrix}; \mathbf{F}(\mathbf{\Phi}) = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} hU \\ hU^2 + \frac{1}{2}gh^2 \end{bmatrix};$$
$$\mathbf{S}(\mathbf{\Phi}) = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} q \\ -gh(S_o - S_e) \end{bmatrix} \tag{51b}$$

where $\mathbf{\Phi}$ is the vector of conserved variables, $\mathbf{F}(\mathbf{\Phi})$ the flux vector, $\mathbf{S}(\mathbf{\Phi})$ the vector of source terms, and the subscripts x and t represent derivation with respect to x and t, respectively. Bold characters denote a vector or a matrix.

To present the general ideas behind conservative methods for modeling flows with discontinuities, we will consider only the homogeneous part of the equation (51a).

$$\mathbf{\Phi}_t + \mathbf{F}(\mathbf{\Phi})_x = 0 \tag{52}$$

The inclusion of the source terms will be discussed later. Consider the solution space x–t shown in Figure 8, which is discretized by dividing it into cells of size $\Delta x \times \Delta t$. Integral form of equation (52) can be obtained by integrating it over the control volume $[x_{i-1/2}, x_{i+1/2}] \times [t^n, t^{n+1}]$, represented by the shaded rectangle shown in Figure 8, and making use of Green's theorem:

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{\Phi}(x, t^{n+1})\,\mathrm{d}x = \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{\Phi}(x, t^n)\,\mathrm{d}x - \left[\int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{\Phi}(x_{i+1/2}, t))\,\mathrm{d}t - \int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{\Phi}(x_{i-1/2}, t))\,\mathrm{d}t\right] \tag{53}$$

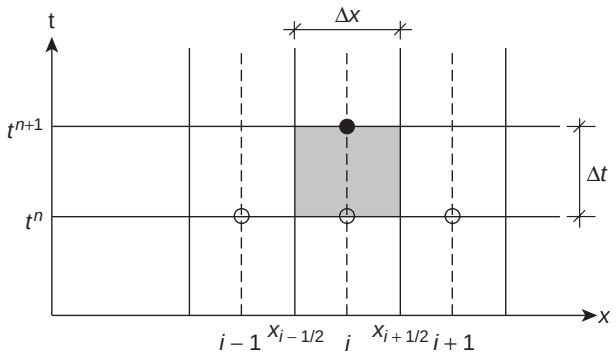


Figure 8 Solution domain x-t, and the control volume used for integration of equation (52)

By making the following definitions:

integral average of $\mathbf{\Phi}(x,t)$ at time $t = t^n$

$$\mathbf{\Phi}_i^n = \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{\Phi}(x, t^n)\,\mathrm{d}x \tag{54a}$$

integral average of $\mathbf{\Phi}(x,t)$ at time $t = t^{n+1}$

$$\mathbf{\Phi}_i^{n+1} = \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{\Phi}(x, t^{n+1})\,\mathrm{d}x \tag{54b}$$

integral average of flux $\mathbf{F}(\mathbf{\Phi})$ at $x = x_{i-1/2}$

$$\mathbf{F}_{i-1/2} = \int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{\Phi}(x_{i-1/2}, t))\,\mathrm{d}t \tag{54c}$$

integral average of flux $\mathbf{F}(\mathbf{\Phi})$at $x = x_{i+1/2}$

$$\mathbf{F}_{i+1/2} = \int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{\Phi}(x_{i+1/2}, t))\,\mathrm{d}t \tag{54d}$$

mesh size and time step, respectively

$$\Delta x_i \equiv x_{i+1/2} - x_{i-1/2}; \Delta t \equiv t^{n+1} - t^n \tag{54e}$$

the integral form of one-dimensional unsteady flow equations is obtained as:

$$\mathbf{\Phi}_i^{n+1} = \mathbf{\Phi}_i^n - \frac{\Delta t}{\Delta x}[\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}] \tag{55}$$

It should be noted that so far the derivation of equation (55) does not involve any approximations. This integral form of governing equations accepts discontinuous solutions.

Properties of Integral Governing Equations

An equivalent but nonconservative form of equation (52) can be written as:

$$\mathbf{\Phi}_t + \frac{\partial \mathbf{F}(\mathbf{\Phi})}{\partial \mathbf{\Phi}}\mathbf{\Phi}_x = \mathbf{\Phi}_t + \mathbf{A}(\mathbf{\Phi})\mathbf{\Phi}_x = 0 \tag{56}$$

where $\mathbf{A}(\mathbf{\Phi})$ is called *the associated Jacobian matrix*, given as:

$$\mathbf{A}(\mathbf{\Phi}) = \frac{\partial \mathbf{F}(\mathbf{\Phi})}{\partial \mathbf{\Phi}} = \begin{bmatrix} \partial f_1/\partial \phi_1 & \partial f_1/\partial \phi_2 \\ \partial f_2/\partial \phi_1 & \partial f_2/\partial \phi_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -U^2 + gh & 2U \end{bmatrix} \tag{57}$$

The system is hyperbolic and has two distinct eigenvalues, which can be determined by setting $|\mathbf{A}(\mathbf{\Phi}) - \lambda \mathbf{I}| = 0$:

$$\lambda_1 = \frac{\mathrm{d}x}{\mathrm{d}t} = U + c;\ \lambda_2 = \frac{\mathrm{d}x}{\mathrm{d}t} = U - c \quad \text{with} \quad c = \sqrt{gh} \tag{58}$$

These eigenvalues are the same as given by equations (10) and (12). The Figure 9 shows the characteristics families for different flow regimes, as has been shown earlier in Figure 2 for the general case. Equations (11) and (13) hold along the characteristic lines λ_1 and λ_2, respectively. The shaded area is the *domain of dependence* for point x_3^{n+1}. The area between two characteristics emanating from a given point is the *domain of influence* for that point.

Equation (52), is written in terms of conserved variables h, and (hU). A quasi-linear form of equations can also be written using primitive variables, h and U:

$$\mathbf{W}_t + \mathbf{E}(\mathbf{W})\mathbf{W}_x = 0 \tag{59}$$

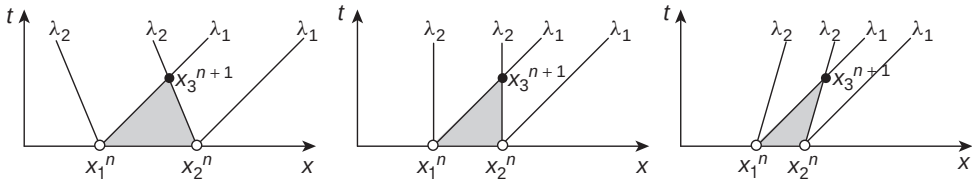


Figure 9 Characteristic lines for various uniform flow regimes

with

$$\mathbf{W} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} h \\ u \end{bmatrix}; \quad \mathbf{B}(\mathbf{W}) = \begin{bmatrix} u & h \\ g & u \end{bmatrix} \tag{60}$$

The $\mathbf{W}$ is the vector of primitive variables, and $\mathbf{E}(\mathbf{W})$ the associated Jacobian matrix. By setting $|\mathbf{E}(\mathbf{W}) - \lambda\mathbf{I}| = 0$ it can be shown that this system has the same two eigenvalues given by equation (58). The right and left eigenvectors can be determined by writing, $\mathbf{E}\mathbf{R}^{(i)} = \lambda\mathbf{R}^{(i)}$ and $\mathbf{L}^{(i)}\mathbf{E} = \lambda\mathbf{L}^{(i)}$, where $i = 1, 2$, respectively:

$$\mathbf{R} = [\mathbf{R}^{(1)}\mathbf{R}^{(2)}] = \begin{bmatrix} \mathrm{r}_1^{(1)} & \mathrm{r}_1^{(2)} \\ \mathrm{r}_2^{(1)} & \mathrm{r}_2^{(2)} \end{bmatrix} = \begin{bmatrix} h & h \\ c & -c \end{bmatrix};$$
$$\mathbf{L} = \begin{bmatrix} \mathbf{L}^{(1)} \\ \mathbf{L}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathrm{l}_1^{(1)} & \mathrm{l}_2^{(1)} \\ \mathrm{l}_1^{(2)} & \mathrm{l}_2^{(2)} \end{bmatrix} = \begin{bmatrix} c & h \\ c & -h \end{bmatrix} \tag{61}$$

Each eigenvalue describes a wave family. The *generalized Riemann invariants* define relationships that hold true across the wave structure. They relate the ratio of the change of dependent variables to the respective component of the eigenvector:

$$\frac{\mathrm{d}w_1}{\mathrm{r}_1^{(i)}} = \frac{\mathrm{d}w_2}{\mathrm{r}_2^{(i)}} \qquad \text{with} \quad i = 1, 2 \tag{62}$$

Using the two right eigenvectors, one can write:

$$\text{For} \quad i = 1 \quad \frac{\mathrm{d}h}{h} = \frac{\mathrm{d}U}{c} \Rightarrow \mathrm{d}U = \frac{c}{h}\,\mathrm{d}h = \frac{\sqrt{g}h^{1/2}}{h} \times \mathrm{d}h = \sqrt{g}h^{-1/2}\,\mathrm{d}h \tag{63a}$$

$$\text{For} \quad i = 2 \quad \frac{\mathrm{d}h}{h} = \frac{\mathrm{d}U}{-c} \Rightarrow \mathrm{d}U = -\frac{c}{h}\,\mathrm{d}h = -\frac{\sqrt{g}h^{1/2}}{h} \times \mathrm{d}h = -\sqrt{g}h^{-1/2}\,\mathrm{d}h \tag{63b}$$

Integrating these equations in the phase plane w_1-w_2 (i.e. u-h), one obtains the following relationships:

$$\text{For } i = 1 \quad \lambda_1 = U + c \Rightarrow U + 2c = const. \tag{64a}$$

$$\text{For } i = 2 \quad \lambda_1 = U - c \Rightarrow U - 2c = const. \tag{64b}$$

These are equivalent to equations (11) and (13), respectively.

Discontinuous Solutions and Riemann Problem

The Riemann problem is an initial-value problem whose formal definition is given as:

$$\mathbf{\Phi}_\mathrm{t} + \mathbf{F}(\mathbf{\Phi})_x = 0$$
$$\mathbf{\Phi}(x, 0) = \begin{cases} \mathbf{\Phi}_\mathrm{L} & \text{if} \quad x < 0 \\ \mathbf{\Phi}_\mathrm{R} & \text{if} \quad x > 0 \end{cases} \tag{65}$$

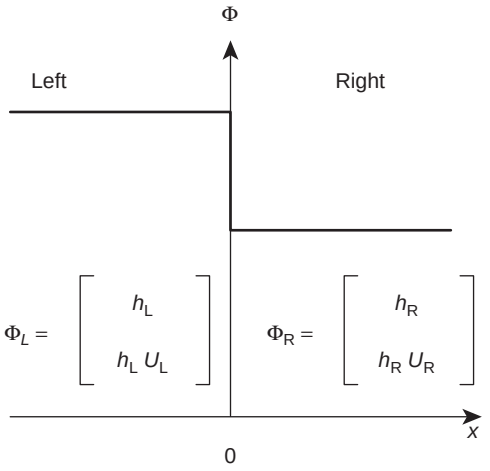


Figure 10 Initial data for the Riemann problem

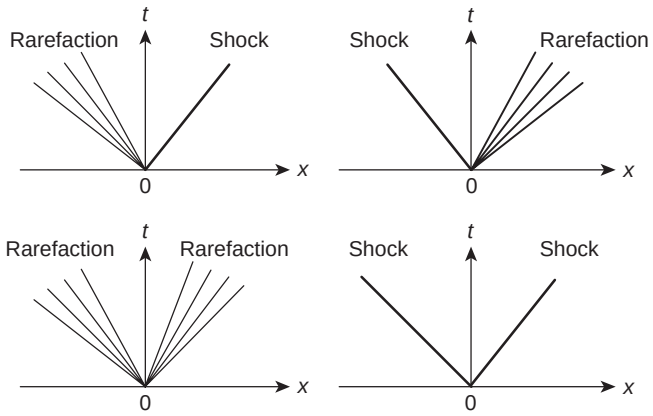


Figure 11 Possible wave patterns for a Riemann problem with both sides containing water

As shown in Figure 10, the initial data represent a discontinuity at $x = 0$, which separates left and right constant states. The Riemann problem is in fact a generalization of the dam-break problem in which the velocities in the left and right constant states are identically zero. The *exact* solution of this problem can be studied in terms of basic wave types which can be either a shock wave or a rarefaction wave. Figure 11 shows four possible wave combinations arising from a Riemann problem.

As shown in Figure 12, the left and right waves, which can be either shock or rarefaction, divide the solution domain into three distinct regions. Left and right states are the undisturbed initial constant data states. The region between the left and right waves is called *star region.* Depending on the type of the wave (shock or rarefaction), functions f_L and f_R, which relate constant values of depth, h_*, and velocity, U_*, in the star region to the values in the undisturbed left and right states, can be derived

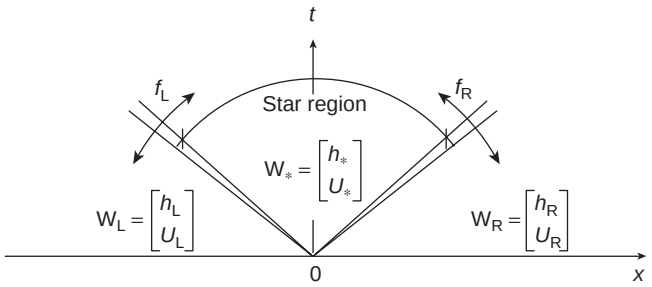


Figure 12 Solution domain used for solving Riemann problem when left and right initial states both contain water. Left and right waves can be either shock or rarefaction. The region between the two waves is called *star region*

using either Riemann invariants, equations (40a,b), or the Rankine–Hugoniot condition for shocks:

$$\mathbf{F}_R - \mathbf{F}_L = a\ (\mathbf{\Phi}_R - \mathbf{\Phi}_L) \tag{66}$$

where "a" is the propagation speed of the shock. Following Toro (2001), the solution for h_* is given by the root of a nonlinear algebraic equation (in terms of f_L, f_R, U_L and U_R) by an iterative numerical method such as Newton–Raphson method. Once h_* is known, the U_* is calculated from a simple algebraic formula. This computational procedure is only valid when the initial left and right constant states have both water ($h_L > 0$ and $h_R > 0$). If one side is dry, then there is no star region and the solution procedure must be modified. The detailed discussion of the solution of Riemann problems for wet–wet and wet–dry type discontinuities can be found in LeVeque (1999), Toro (1999) and Toro (2001). The last reference contains a sample FORTRAN code for the exact solution of Riemann problem for both wet–wet and wet–dry discontinuities.

First – order Numerical Schemes for Modeling Flows with Discontinuities

Nonconservative forms of governing equations are not suitable for dealing with discontinuous solutions. In the presence of discontinuities, nonconservative equations written in primitive variables lead to wrong jump conditions and give erroneous results for shock speed, strength, and position. Moreover, Hou and LeFloch (1994) have shown that nonconservative schemes do not converge to correct solution if the solution involves a shock wave.

Lax and Wendroff (1960) have shown that convergent conservative methods do converge to the *weak solution* of the conservation law. The weak solution is a solution that satisfies the integral form (conservative form) of conservation laws, equation (55), and the Rankine–Hugoniot condition, equation (66). It is, however, important to mention that the uniqueness of a weak solution is not always guaranteed. The weak solution may converge to a physically nonrelevant result. In such cases additional constraints, called *entropy function*, need to be imposed to ensure convergence to a physically relevant solution.

When modeling flows with discontinuities, special attention must be paid not only to the form of equations but also to the numerical scheme to be used. In this respect, the conservative form given by equation (55) constitutes an excellent starting point for developing conservative numerical schemes. Equation (55) admits discontinuous solutions, and also provides a direct link for the development of numerical methods when the flux terms are replaced by their numerical approximations over a solution domain discretized with a step size of $\Delta x = x_{i+1/2} - x_{i-1/2}$. Since information propagates with finite speed, it can be assumed that numerical approximations to flux terms $\mathbf{F}_{i+1/2}$ and $\mathbf{F}_{i-1/2}$ can be devised in terms of the initial data to the left and right of the respective interfaces, provided that the time step is chosen appropriately.

The general idea in selecting a correct time step is that the fastest wave should not travel more than one cell length, Δx, during the time Δt. On the basis of the values known at time $t = t^n$, let the propagation speed of the fastest wave be $a^n_{\max}$. The constraint for the time step can be given as $a^n_{\max} \le \Delta x/\Delta t$, where $\Delta x/\Delta t$ can be interpreted as grid speed. This constraint can be cast in the following convenient form:

$$\Delta t = \frac{C_{CFL}\Delta x}{a^n_{\max}} \quad \text{with} \quad 0 < C_{CFL} \le 1 \tag{67}$$

Table 3 Some centered (symmetric) schemes and their properties

Scheme	Flux term in equation (55)	Properties	
Forward time-centered space (FTCS)	$\mathbf{F}^{FTCS}_{i+1/2} = \frac{1}{2}(\mathbf{F}^n_i + \mathbf{F}^n_{i+1})$	Unconditionally unstable	
Lax–Friedrichs (LF)	$\mathbf{F}^{LF}_{i+1/2} = \frac{1}{2}(\mathbf{F}^n_i + \mathbf{F}^n_{i+1}) + \frac{\Delta x}{2\Delta t}(\mathbf{\Phi}^n_i - \mathbf{\Phi}^n_{i+1})$	Monotone and stable for	$0 < \lvert CFL\rvert \le 1$
Godunov-centered (GODC)	$\mathbf{F}^{GODC}_{i+1/2} = \mathbf{F}(\mathbf{\Phi}^{GODC}_{i+1/2})$, with $\mathbf{\Phi}^{GODC}_{i+1/2} = \frac{1}{2}(\mathbf{\Phi}^n_i + \mathbf{\Phi}^n_{i+1}) + \frac{\Delta t}{\Delta x}(\mathbf{F}^n_i - \mathbf{F}^n_{i+1})$	Stable for monotone for oscillatory for	$\lvert N_{CFL}\rvert \le \sqrt{2}/2$ $1/2 \le \lvert N_{CFL}\rvert \le \sqrt{2}/2$ $0 \le \lvert N_{CFL}\rvert \le 1/2$

where C_{CFL} is the Courant-Friedrichs-Levy (CFL) coefficient, or the Courant number. It is generally taken as $C_{CFL} \cong 0.9$.

Centered numerical schemes provide the simplest approach for expressing fluxes at the cell interfaces. These schemes do not explicitly take into account the wave nature of the solution and the direction of propagation of the information. Table 3 summarizes the flux expressions of some important centered schemes and their properties. First – order schemes are generally highly diffusive and the stability condition is defined by the Courant number.

The centered (symmetric) schemes do not make use of the wave-propagation characteristics of conservation equations. Godunov (1959) devised a first – order upwind method on the basis of the forward-time solution of Riemann problem. Consider the discretized solution domain illustrated in Figure 13(a). The solution $\mathbf{\Phi}^n$ at time $t = t^n$ is known (Figure 13b). A piecewise continuous representation of the initial data is obtained by calculating the integral averages $\mathbf{\Phi}_i^n = \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{\Phi}^n \, dx$. To obtain the Godunov intercell fluxes $\mathbf{F}_{i\pm1/2}^{GOD}$, first the exact Riemann initial-value problem is solved at each cell interface using a local coordinate system with $x = 0$ centered at the interface (Figure 13c). If the time interval is chosen sufficiently small, such that $|C_{CFL}| \leq 1$, the solutions $\mathbf{\Phi}_{i-1/2}(x/t)$ and $\mathbf{\Phi}_{i+1/2}(x/t)$ emanating from left and right sides of the cell do not reach the other side of the cell. The Godunov intercell fluxes can then be simply computed as the physical flux function $\mathbf{F}(\mathbf{\Phi})$ evaluated along the t-axis used for the Riemann problem, which in terms of the local coordinate system corresponds to $x/t = 0$. The results of Riemann problems at the cell interfaces are sampled to obtain:

$$\mathbf{F}_{i+1/2}^{GOD} = \mathbf{F}(\mathbf{\Phi}_{i+1/2}(0)) \quad \text{and} \quad \mathbf{F}_{i-1/2}^{GOD} = \mathbf{F}(\mathbf{\Phi}_{i-1/2}(0)) \tag{68}$$

The conservative form equation (55) is then used for marching the solution to the time $t = t^{n+1}$:

$$\mathbf{\Phi}_i^{n+1} = \mathbf{\Phi}_i^n - \frac{\Delta t}{\Delta x}\left[\mathbf{F}_{i+1/2}^{GOD} - \mathbf{F}_{i-1/2}^{GOD}\right] \tag{69}$$

The exact solution of Riemann problem at each interface involves an iterative solution for h_* and is time consuming. Moreover, the detailed flow field information given by the exact solution of the Riemann problem is not used by the Godunov method. Several methods have, therefore, been proposed for approximate solution of the Riemann initial-value problem. Approximate Riemann solvers can lead to considerable reduction in computational time.

Roe's approximate solver (Roe, 1981) is the earliest and most popular approximate solver. It is based on the idea of replacing the associated Jacobian matrix $\mathbf{A}$ in equation (56), by a linear system with constant coefficients (obtained by freezing $\mathbf{A}$ at some intermediate time step). Roe's approximate solver fails, however, in case of wet-dry discontinuities. This can be avoided by imposing a very small water depth even when there is no water. This solution, however, may sometimes lead to an error in the estimation of the true wave speeds. Roe's approximate Riemann solver also produces an entropy-violating rarefaction shock when the rarefaction wave (negative wave) forms around the local t-axis (sonic or critical conditions). This problem can be cured by imposing an appropriate entropy function. Another important approximate solver is the HLL (Harten *et al.*, 1983) approximate Riemann solver which computes the numerical flux using estimated values of left and right wave speeds. HLL Riemann solver, however, relies on another suitable approximate Riemann solver for computing h_*. A detailed discussion of the approximate solvers is beyond the scope of this article. The interested readers are referred to LeVeque (1999), and Toro (1999) and Toro (2001).

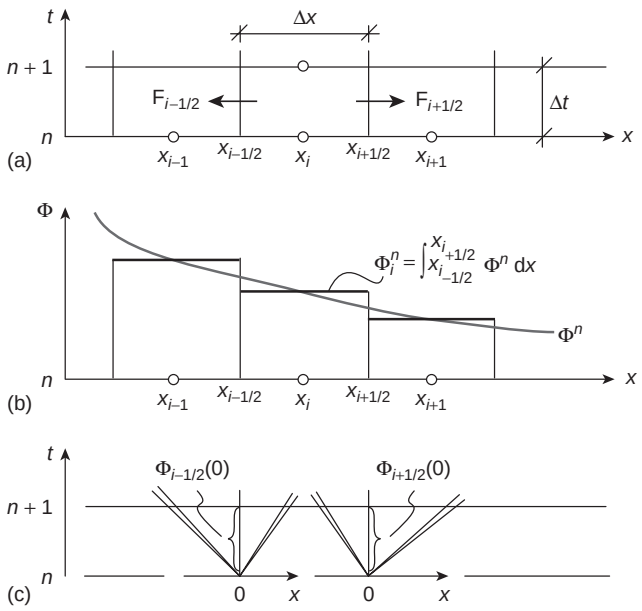


Figure 13 Principles of Godunov upwind scheme for one-dimensional flow: (a) control volume in solution space x–t, and the Godunov fluxes; (b) initial solution at time $t = t^n$ and its piecewise constant representation; (c) Riemann problems are solved in local coordinates with the origin located at the cell interfaces (adapted from Toro, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Higher – order Numerical Schemes for Modeling Flows with Discontinuities

On the basis of the leading term of their local truncation error, it can be seen that the centered schemes and the Godunov upwind scheme presented above are all first – order accurate. These first – order schemes are

very diffusive and the discontinuities smear out as they propagate. One way of fixing this problem is to construct higher – order schemes.

A second – order accurate (in space and time) version of Godunov scheme can be constructed using Weighted Average Flux (WAF) method. The WAF method also relies on the piecewise constant representation of the initial data. Without going into the details, the WAF flux at a cell interface can be regarded as a weighted average of the fluxes at the cell centers on both sides see Toro (1999) and Toro (2001).

$$\mathbf{F}_{i+1/2}^{\mathrm{WAF}} = \frac{1}{2}(\mathbf{F}_i - \mathbf{F}_{i+1}) - \frac{1}{2}\sum_{1}^{N} C_{\mathrm{CFL}}^{(k)} \Delta \mathbf{F}_{i+1/2}^{(\mathrm{k})} \quad \text{with}$$

$$\Delta \mathbf{F}_{i+1/2}^{(\mathrm{k})} = \mathbf{F}_{i+1/2}^{(\mathrm{k}+1)} - \mathbf{F}_{i+1/2}^{(k)} \tag{70}$$

The term $\Delta \mathbf{F}_{i+1/2}^{(k)}$ is the flux jump across the wave k, N the number of conservation equations (two in the present case). The courant number for the wave k is represented by $C_{\mathrm{CFL}}^{(k)} = \Delta t a_{\max}^{(k)}/\Delta x$, where $a_{\max}^{(k)}$ is the maximum propagation speed of the wave k.

Monotone Upstream-centered Scheme for Conservation Laws (MUSCL) (MUSCL or MUSCL-Hancock), applies a different approach for the derivation of a scheme second – order accurate in space and time. MUSCL-Hancock method uses a piecewise linear representation of the initial data as opposed to piecewise constant representation used by the original Godunov method. As a result, the left and right states are not constant but linearly varying. This leads to a so-called *Generalized Riemann Problem* (GRP), which is defined as

$$\mathbf{\Phi}_t + \mathbf{F}(\mathbf{\Phi})_x = 0$$
$$\mathbf{\Phi}(x,0) = \begin{cases} \mathbf{\Phi}_{\mathrm{L}}(x) & \text{if} \quad x < 0 \\ \mathbf{\Phi}_{\mathrm{R}}(x) & \text{if} \quad x > 0 \end{cases} \tag{71}$$

When compared with equation (65) it is seen that the left and right states are now functions of x. The resulting characteristic lines are no longer straight lines but curves in the x–t plane. Without going into the details, one way of solving GRP is to compute left and right states at the half-step ($\Delta t/2$) using a formula involving the values of $\mathbf{\Phi}_{\mathrm{L}}(x)$ and $\mathbf{\Phi}_{\mathrm{R}}(x)$ evaluated at the interface. These intermediate $\mathbf{\Phi}_{\mathrm{L}}(x)$ and $\mathbf{\Phi}_{\mathrm{R}}(x)$ values are then treated as constant left and right states of a piecewise constant data, and the intercell fluxes are computed using an appropriate Riemann solver, the same way it is done in the first–order upstream method of Godunov. In fact, the approach used by MUSCL-Hancock method can be used to construct even higher – order schemes by using other piecewise representations of the data. The piecewise parabolic method of Colella and Woodward (1984), for example, is a scheme third – order accurate in space and time.

The higher – order methods presented above are not monotone and suffer from spurious oscillations near the sharp discontinuities. Total variation diminishing (TVD) methods provide a way to cure this problem. TVD methods are in a way similar to the use of artificial viscosity in some traditional centered schemes. The difference being that in TVD methods artificial viscosity is built into the scheme and is activated in a rather sophisticated way. TVD schemes were first introduced by Harten (1984) for solving the Euler equations in gas dynamics. The TVD schemes are second – order accurate and oscillation free at the discontinuities. A complete discussion of TVD class schemes would be too lengthy to describe here. It will suffice to mention that there are two general approaches to construct TVD schemes: *flux limiter approach* and *slope limiter approach*. The details of these approaches and TVD versions of higher – order upwind schemes can be found in LeVeque (1999), Toro (1999), and Toro (2001).

Treatment of Source Terms

The conservative methods described in the preceding sections dealt exclusively with the homogeneous form of the governing equations given by equation (52). Using fractional step method, these numerical schemes can be easily extended to solve the complete version with source terms as given by equation (51a). The fractional step (or splitting) method requires the solution of two separate equations at each step:

$$\mathbf{\Phi}_t + \mathbf{F}(\mathbf{\Phi})_x = 0 \tag{72a}$$

and

$$\mathbf{\Phi}_t = \mathbf{S}(\mathbf{\Phi}) \tag{72b}$$

First, using the initial data $\mathbf{\Phi}_i^n$ at time t, the equation (72a) is solved to obtain an intermediate solution, $\mathbf{\Phi}_i^{\mathrm{int}}$ at time $t + \Delta t$. This intermediate solution is then used as the initial data for solving equation (72b) to find the final solution $\mathbf{\Phi}_i^{n+1}$ at time $t + \Delta t$. Any appropriate numerical method can be used at this step to solve the ordinary differential equation, equation (72b). The entire solution can be written as:

$$\mathbf{\Phi}_i^{n+1} = \mathbf{\Phi}_i^n - \frac{\Delta t}{\Delta x}(\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}) + \Delta t \mathbf{S}(\mathbf{\Phi}_i^{\mathrm{int}}) \tag{73}$$

The fractional step (or splitting) methods do not give good results for problems involving steady flows or stagnant water. This is because of their inability to correctly model the delicate balance between flux terms and the source terms in each cell.

Recently methods have been developed to incorporate the source terms into the wave-propagation algorithm and thereby to avoid the use of fractional step methods. A

discussion of these methods can be found in LeVeque (1998) and Sanders *et al.* (2003).

AN EXAMPLE OF A CONSERVATIVE UPWIND MODEL AND ITS TESTING

Some of the high – order conservative upwind methods presented above may still present entropy glitches and may not perform well with arbitrary cross sections and bottom variations. For solving practical problems, it is desired to have a computationally efficient robust scheme, which will perform well under a wide range of conditions. The relatively simple upwind scheme proposed by Ying *et al.* (2003) and Ying *et al.* (2004) is very robust and computationally very efficient. The governing equations solved are:

$$\frac{\partial \mathbf{\Phi}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{\Phi})}{\partial x} = \mathbf{S}(\mathbf{\Phi}) \tag{74}$$

with

$$\mathbf{\Phi} = \begin{bmatrix} A \\ Q \end{bmatrix} \quad \mathbf{F}(\mathbf{\Phi}) = \begin{bmatrix} Q \\ \dfrac{Q^2}{A} \end{bmatrix}$$

$$\mathbf{S}(\mathbf{\Phi}) = \begin{bmatrix} 0 \\ -gA\dfrac{\partial Z}{\partial x} - \mathrm{g}\dfrac{Q|Q|}{C^2 R_\mathrm{h} A} \end{bmatrix} \tag{75}$$

where A is the cross-sectional area; Q is the discharge; g the gravitational acceleration; Z the water-surface elevation with respect to a reference datum; C the Chezy coefficient, and R_h is the hydraulic radius. Referring to Figure 8, the conservative form of equation (74) is obtained by integrating it over the control volume $[x_{i-1/2}, x_{i+1/2}] \times [t^n, t^{n+1}]$:

$$\mathbf{\Phi}_i^{n+1} = \mathbf{\Phi}_i^n - \frac{\Delta t}{\Delta x_i}(\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}) + \Delta t \mathbf{S}_i \tag{76}$$

The intercell flux term is given by

$$\mathbf{F}_{i+1/2} = \begin{bmatrix} Q_{i+k}^n \\ \dfrac{(Q_{i+k}^n)^2}{A_{i+k}^n} \end{bmatrix} \text{ where } k = \begin{cases} 0 & \text{if} \quad Q \geq 0 \\ 1 & \text{if} \quad Q \leq 0 \end{cases} \tag{77}$$

and the source term is calculated using

$$\mathbf{S}_i = \begin{bmatrix} 0 \\ -gA_i^{n+1}\left[\omega_1\left(\dfrac{Z_{i+1-k}^{n+1} - Z_{i-k}^{n+1}}{x_{i+1-k} - x_{i-k}}\right) + \omega_2\left(\dfrac{Z_{i+k}^{n+1} - Z_{i-1+k}^{n+1}}{x_{i+k} - x_{i-1+k}}\right)\right] - g\dfrac{Q_i^n|Q_i^n|}{(C_i^n)^2 R_i^n A_i^n} \end{bmatrix} \tag{78}$$

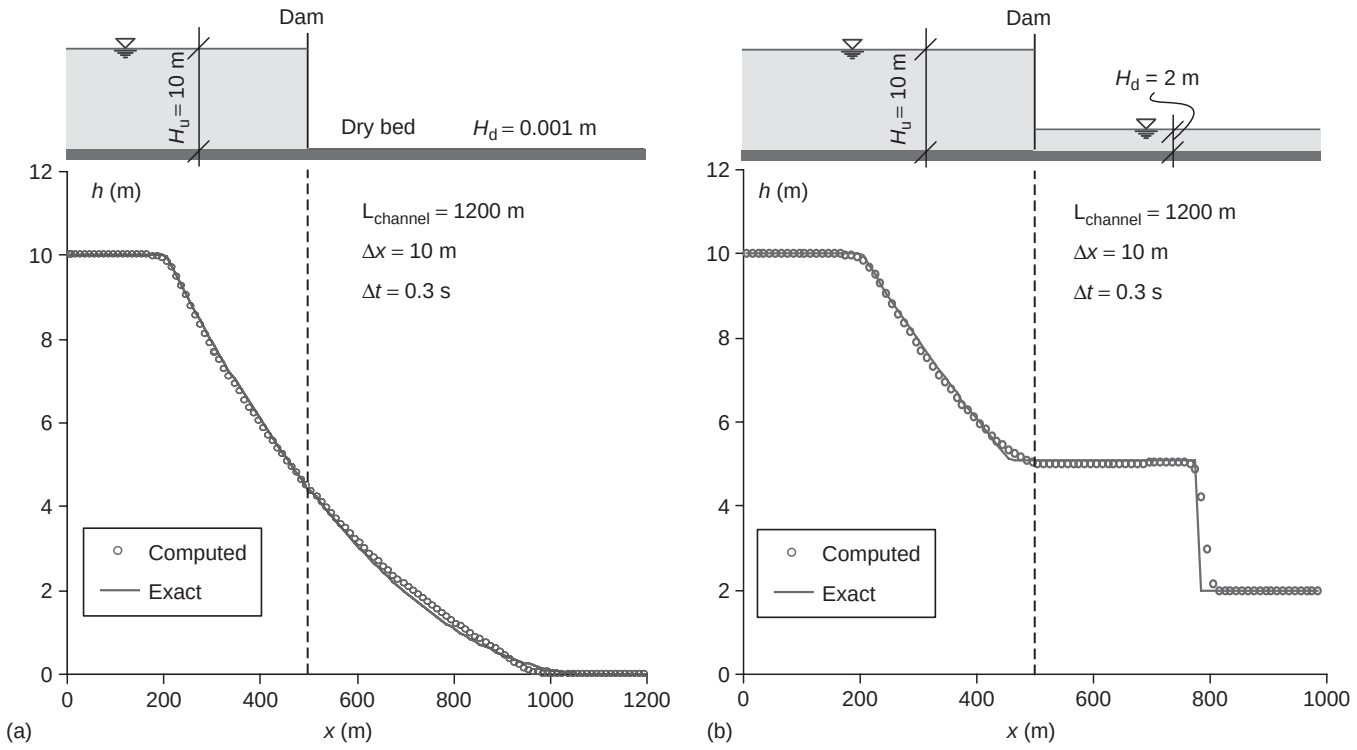


Figure 14 Dam break in a frictionless rectangular channel. (a) downstream is dry; (b) downstream has a water depth of 2 m. Equation (79b) is used to compute weighting factors. Initial conditions and other computational parameters are given in the figure. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

where ω_1 and ω_2, are the weighting factors used for calculating the gradient of the water-surface as a combination of forward and backward differences. Ying *et al.* (2003, 2004) propose two different choices for calculating these weighting factors:

$$\omega_1 = 1 - \frac{\Delta t}{x_{i+1-k} - x_{i-k}} \frac{U_{i+1-k} + U_{i-k}}{2} \quad \text{and}$$
$$\omega_2 = \frac{\Delta t}{x_{i+k} - x_{i-1+k}} \frac{U_{i+k} + U_{i-1+k}}{2} \tag{79}$$

or

$$\omega_1 = 1 - \sqrt{\frac{\Delta t}{x_{i+1-k} - x_{i-k}} \frac{U_{i+1-k} + U_{i-k}}{2}} \quad \text{and}$$
$$\omega_2 = \sqrt{\frac{\Delta t}{x_{i+k} - x_{i-1+k}} \frac{U_{i+k} + U_{i-1+k}}{2}} \tag{80}$$

The analysis shows that the conservative finite-volume scheme of Ying *et al.* (2003, 2004) is stable if the Courant number satisfies the following relationship: $C_{\mathrm{CFL}} = (U + \sqrt{gh})\Delta t/\Delta x$, where $U = |Q|/A$ is the magnitude of the average velocity at a cross section, and h the local water depth. The dry-bed condition is handled by maintaining a very small water depth even when the cross section is dry. Authors report that a general value of $h_{\min} = 0.001$ m can be used for all practical cases without introducing any noticeable error for the wave-propagation speed. The numerical tests show that the model is extremely robust and computationally very efficient. The results of several test cases are illustrated in Figures 14, 15 and 16.

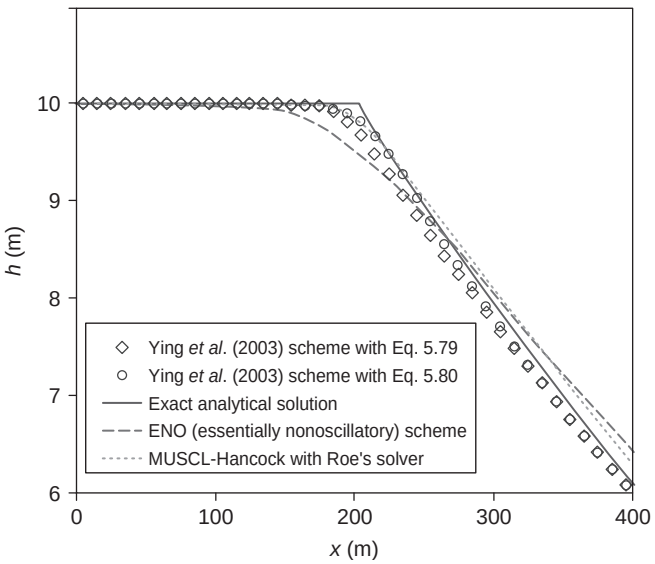


Figure 15 Comparison of Ying *et al.* (2003) scheme with ENO and MUSCLE-Hancock schemes for the case of dam break in a horizontal frictionless channel. The computational parameters are the same as in Figure 14(a). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

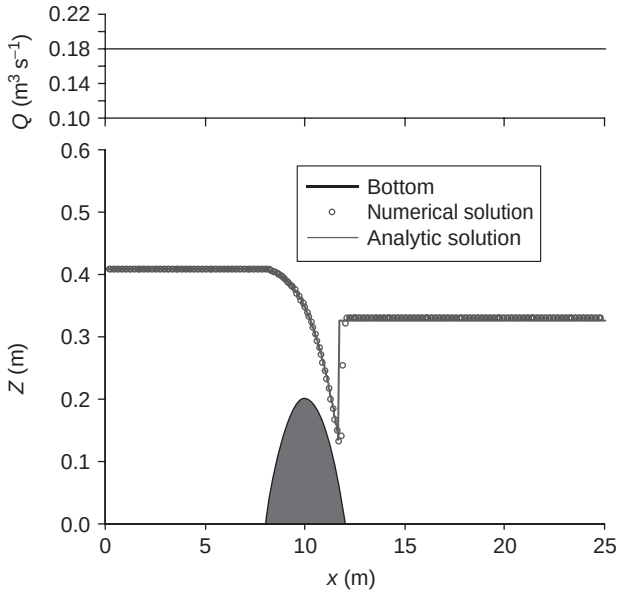


Figure 16 Numerical solution of flow over a bump using Ying *et al.* (2003) scheme. Note that the hydraulic jump downstream of the bump is finely resolved. The plot of discharge as a function of distance yields a constant line indicating an excellent mass conservation. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CONCLUDING REMARKS FOR DISCONTINUOUS SOLUTIONS

The research on shock-capturing conservative finite-volume methods for solving hyperbolic conservation laws has made considerable progress in the recent years. The short introduction given here barely scratches the surface of this rapidly evolving research area. Zoppou and Roberts (2002) gives a short but complete overview of various methods as well as a comparison of the relative computational cost of various schemes. The readers interested on a more complete treatment of the topic are referred to excellent books by LeVeque (1999) and Toro (1999) and Toro (2001). Sample computer codes on exact Riemann solver and Godunov method can be found in the latter.

Conservation LAWs PACKage (CLAWPACK) is a package of Fortran subroutines for solving time-dependent hyperbolic systems of partial differential equations in 1, 2, and 3 space dimensions, including nonlinear systems of conservation laws. The software can also be used to solve nonconservative hyperbolic systems and systems with variable coefficients, as well as systems including source terms. The package includes an MPI version in which the domain

can be distributed among multiple processors, and adaptive mesh refinement versions (AMRCLAW) in two and three space dimensions. This package and the accompanying documentation can be downloaded from the website by LeVeque (2003).

Recently, Sobey (2001) reviewed the physical characteristics of the St. Venant equations expressed in the conservation form, and highlighted the challenges that a numerical model has to address in dealing with a hyperbolic system. He has also developed a series of benchmark test problems that will highlight the significant response patterns of a numerical model, and can be used to evaluate the performance of a numerical model of flood and tide propagation in channels. He stressed the importance of testing the models using such benchmark test problems as a routine and automatic part of any numerical model development exercise in this area.

Lai *et al.* (2002) gives various forms of conservation-form equations of unsteady open-channel flow. An analysis of the Riemann solvers available in the literature was carried out by Delis *et al.* (2000a), and they have compared the predictive ability of these solvers by applying them to the idealized and inhomogeneous dam-break problems. Delis *et al.* (2000b) had examined four implicit high-resolution TVD schemes for solving the St. Venant equations. Delis *et al.* (2000b) presented two new TVD schemes: one based on Harten's *et al.* (1983) modified flux and the second based on Van Leer (1979) MUSCL approach. Delis *et al.* (2000b) tested these two methods and a symmetric TVD proposed by Yee (1987) using the analytical solutions of the steady state benchmark test problems proposed by MacDonald *et al.* (1997). The idealized dam-break problem was also included in the test to check the accuracy of the schemes in an unsteady flow with shocks. From these tests, Delis *et al.* (2000b) concluded that the MUSCL method was the most efficient and robust.

REFERENCES

Abbott M.B. (1975) Method of characteristics. In *Unsteady Flow in Open Channels*, Chap. 3, Mahmood K. and Yevjevich V. (Eds.), Vol. 1. Water Resources Publications: Fort Collins, Colorado, USA.

Colella P. and Woodward P.R. (1984) The piecewise-parabolic method (PPM) for gas dynamical simulations. *Journal of Computational Physics*, **54**, 174–201.

Cunge J.A. (1961) Etude d'un schema de differences finies applique a l'integration numerique d'un certain type d'equation hyperbolique d'ecoulement. Thesis presented to the Faculty of Sciences of Grenoble University.

Cunge J.A., Holly F.M. and Verway A. (1980) *Practical Aspects of Computational River Hydraulics*, Pitman Publishing: London.

Delis A.I., Skeels C.P. and Ryrie S.C. (2000a) Evaluation of some approximate Riemann solvers for transient open channel flows. *Journal of Hydraulic Research, IAHR*, **38**(3), 217–231.

Delis A.I., Skeels C.P. and Ryrie S.C. (2000b) Implicit high-resolution methods for modeling one-dimensional open channel flow. *Journal of Hydraulic Research, IAHR*, **38**(5), 369–382.

Fletcher A.G. and Hamilton W.S. (1967) Flood routing in irregular channel. *Journal of the Engineering Mechanics Division, ASCE*, **94**(EM3), 45–62.

Godunov S.K. (1959) Finite difference methods for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mathematicheskii Sbornik*, **47**(3), 271–306.

Gunaratnam D. and Perkins F.E. (1970) *Numerical Solutions of Unsteady Flow in Open Channels, Hydrodynamic Laboratory T.R. No. 127*, Department of Civil Engineering, MIT: Cambridge.

Harten A. (1984) On a class of high resolution total-variation-stable finite difference schemes. *SIAM Journal of Numerical Analysis*, **21**, 1.

Harten A., Lax P. and van Leer B. (1983) On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Review*, **25**, 35–61.

Hou T.Y. and LeFloch P.G. (1994) Why nonconservative schemes converge to wrong solutions: error analysis. *Mathematics of Computation*, **62**, 497–530.

Krishnappan B.G. (1981) *User Manual: Unsteady, Nonuniform, Mobile Boundary Flow Model-MOBED*, Hydraulics Division, National Water Research Institute, Environment Canada: Burlington, p. 107.

Krishnappan B.G. (1985) Modeling of unsteady flows in alluvial streams. *Journal of Hydraulic Engineering, ASCE*, **111**(2), 257–266.

Lai C. (1965) *Flows of Homogeneous Density in Tidal Reaches: Solution by the Method of Characteristics*, United States Department of the Interior, Geological Survey: Open File Report, Washington.

Lai C., Baltzer R.A. and Schaffranek R.W. (2002) Conservation-form equations of unsteady open-channel flow. *Journal of Hydraulic Research*, **40**(5), 567–578.

Lax P.D. and Wendroff B. (1960) Systems of conservation laws. *Communication On Pure and Applied Mathematics*, **13**, 217–237.

LeVeque R.J. (1998) Balancing source terms and flux gradients in high-resolution Godunov methods: The Quasi-steady Wave-propagation algorithm. *Journal of Computational Physics*, **146**(1), 346–365.

LeVeque R.J. (1999) *Numerical Methods for Conservation Laws*, Birkhäuser Verlag: Basel.

LeVeque R.J. (2003) *A Software Package for Conservation Laws and Hyperbolic Systems*. Downloadable from `http://www.amath.washington.edu/~claw/`.

Liggett J.A. (1968) Mathematical flow determination in open channels. *Journal of Engineering Mechanics Division, ASCE*, **94**(EM4), 947–963.

Liggett J.A. and Cunge J.A. (1975) Numerical methods of solution of the unsteady flow equations. In Mahmood K. and Yevjevich V. (Eds.), *Unsteady Flow in Open Channels*, Vol. 1, Water Resources Publications, Fort Collins.

MacDonald I., Baines M.J., Nichols N.K. and Samuels P.K. (1997) Analytic benchmark solutions for open channel flows. *Journal of Hydraulic Engineering, ASCE*, **123**(11), 1041–1045.

Preissmann A. (1961) *Propagation Des Intumescences Dans Les Cannaux Et Rivieres*, 1er Congres de l'association francaise de calcul: Grenoble.

Roe P.L. (1981) Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, **43**, 357–372.

Sanders B.F., Jaffe D.A. and Chu A.K. (2003) Discretization of integral equations describing flow in nonprismatic channels with uneven beds. *Journal of Hydraulic Engineering*, **129**(3), 235–244.

Sobey R. (2001) Evaluation of numerical models of flood and tide propagation in channels. *Journal of Hydraulic Engineering, ASCE*, **127**(10), 805–824.

Stoker J.J. (1957) *Water Waves*, Interscience Press: New York.

Toro E.F. (1999) *Riemann Solvers and Numerical Methods for Fluid Dynamics*, *Second Edition*, Springer-Verlag: Berlin.

Toro E.F. (2001) *Shock-Capturing Methods for Free-Surface Shallow Flows*, Wiley and Sons: Chichester.

Treske A. (1980) Experimentelle Überprüfung numerischer Berechnungsverfahren von Hochwasserwellen, in blind. *Report of Hydraulics Research Station, TU München*, **44**, 1–133.

Van Leer B. (1979) Towards the ultimate conservative difference scheme. V. A second order sequel to Gudonov's method. *Journal of Computational Physics*, **32**, 101.

Yee H.C. (1987) Construction of explicit and implicit symmetric TVD schemes and their applications. *Journal of Computational Physics*, **68**, 151.

Ying X., Khan A.A. and Wang S.S.Y. (2003) *An Upwind Method for One-Dimensional Dam Break Flows*, Proceedings of XXX Congress of International Hydraulic Research Association, Thessaloniki, pp. 245–252.

Ying X., Khan A.A. and Wang S.S.Y. (2004) An upwind conservative scheme for Saint Venant equations. *Journal of Hydraulic Engineering*, **130**(10), 977–987.

Zoppou C. and Roberts S. (2002) Explicit schemes for dam-break simulations. *Journal of Hydraulic Engineering*, **129**(1), 11–34.

140: Transport of Sediments

WALTER H GRAF AND MUSTAFA S ALTINAKAR

Laboratoire de Recherches Hydrauliques, Ecole Polytechnique Fédérale, Lausanne, Switzerland

A channel bed, made up of a granular material, is usually of a form which varies in space and time. The movement of the sediments, which make up the bed, represents a rather complex phenomenon. The hydrodynamic equations of the flow over a mobile bed are developed. The different modes of the transport of (noncohesive) sediments as bed load and as suspended load are presented. The formulae for the calculations of the transport of the total load are exposed, as well as their domain of application.

GENERALITIES

Notions

The flow of water over a mobile bed has the ability to entrain the sediments; a water–sediment mixture will consequently displace itself in the watercourse. The movement of the sediments – *erosion, transport, deposition* – will modify the flow, but also the channel bed. The interaction between the water and the sediments makes the problem a coupled one. When the bed is a *mobile* one, the fluvial hydraulics must concern itself with both the flow of the liquid phase, namely the mixture, and the movement of the solid phase, namely the sediments in the mixture.

The transport of these sediments plays an important role in all problems of fluvial hydraulics. The formulae, developed for the quantitative determination of the transport of sediments, are based on experimental results, and being often limited, they should be used with much caution, and they must be applied within hydraulic conditions under which they have been established.

Flow of a Mixture

For gravitational flow of a water–sediment mixture, one may distinguish three types of movement:

1. The mixture may be considered Newtonian if the volumic concentration of the particles is very small, $C_s \ll 1\%$. The difference between the density of the mixture and of the water, $\Delta\rho = (\rho_m - \rho) = (\rho_s - \rho)C_s$, remains also small, $\Delta\rho \ll 16\,\mathrm{kg\,m^{-3}}$.
   The *transport of sediments*, as bed load and suspended load, falls into this category. It is this type of transport of solid particles that is most often encountered in watercourses and treated in this article.
2. The mixture behaves like a quasi-Newtonian if the volumic concentration remains small, $C_s < 8\%$; the density difference becomes important, $\Delta\rho < 130\ \mathrm{kg\,m^{-3}}$.
   The transport of sediments as *concentrated suspension* (see Graf, 1971, 1984, pp. 182–186) notably close to the bed, as well as the *turbidity currents* fall into this category.
3. The mixture behaves non-Newtonian, if the volumic concentration becomes of importance, $C_s > 8\%$; the density difference is also very large, $\Delta\rho > 130\,\mathrm{kg\,m^{-3}}$.
   The flow of a non-Newtonian fluid modifies all concepts of Newtonian hydraulics.
   The transport of sediments as *hyperconcentrated suspension* (see Wan and Wang, 1994), the *debris flow* (see Takahashi, 1991), as well as *hyperconcentrated turbidity currents* (see Wan and Wang, 1994) fall into this category.

Modes of Transport (see Figure 1)

The transport of sediments by flow of water is the entire solid transport which passes through a cross section of a watercourse. Traditionally the transport of sediments is

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

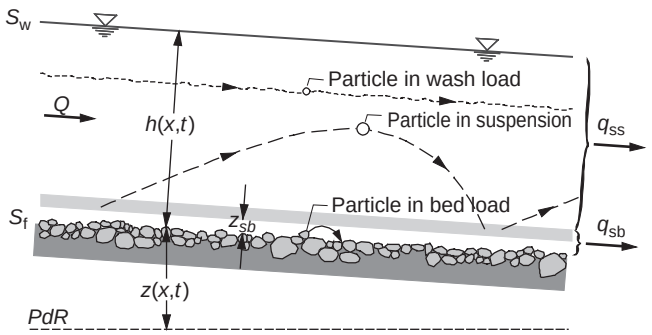


Figure 1 Scheme of the modes of transport. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

classified in different modes of transport that correspond to distinctly different physical mechanisms.

In a watercourse the sediments, namely the solid phase, are transported as:

1. *bed load*, q_{sb} – volumic solid discharge per unit width ($m^3\,s\,m^{-1}$) – when the particles stay in close contact with the bed;
2. *suspended load*, q_{ss}, when the particles stay occasionally in contact with the bed;
3. bed load + suspended load, being the (total) *bed-material load*, $q_s = q_{sb} + q_{ss}$, when the particles stay more or less in continuous contact with the bed.
4. *wash load*, q_{sw}, when the particles are almost never in contact with the bed; the particles are washed through the cross section by the flow.

The transport of sediments, namely the erosion of the bed (*see* **Chapter 137, Uniform Flow, Volume 4**), commences upon attainment of a certain critical value, which can be parameterized by the critical shear stress, τ_{ocr}.

It will be useful to give limiting values for the separation of the different modes of transport, using the ratio of the shear velocity of the flow, u_*, and the settling velocity of the particles, v_{ss} (see Graf, 1971, 1984):

- $u_*/v_{ss} > 0.10$ for beginning of bed-load transport,
- $u_*/v_{ss} > 0.40$ for beginning of suspended load transport.

To determine quantitatively the transport of sediments, there are three possibilities available, namely:

- using existing formulae (see Graf and Altinakar, 1998, Chap. 6),
- obtaining field measurements with adequate instruments (see Graf, 1971, 1984, Chap. 13),
- performing physical models (see Graf, 1971, 1984, Chap. 14).

The different modes of transport of sediments, quantified in form of solid discharge, q_{sb}, q_{ss}, and q_s, should be related to the liquid discharge, q. This will give the relation of the "sedimentological" rating curve for a given cross section of the channel. This curve, together with the "liquid" rating curve, gives a rather complete hydraulic description for a given cross section of a channel having a mobile bed.

The formulae, which are used to calculate the solid discharge, q_s, allow determination of the *capacity* of the transport of sediments for a given flow. Under such conditions, the transport of sediments is said to be in equilibrium.

HYDRODYNAMIC EQUATIONS

The hydrodynamic equations, and some solutions, for flow in an open channel over a mobile bed, when entrainment of sediments take place, will be presented next.

Equations of Saint-Venant–Exner

The equations of Saint-Venant (*see* Figure 2 and **Chapter 138, Unsteady Flow, Volume 4**) for unsteady and nonuniform flow over a *fixed bed* in a prismatic open channel of constant width, $B = Cte$, having a weak bed slope, S_f, were given before; for flow over a *mobile bed*, they can be written as:

$$\frac{\partial h}{\partial t} + h\frac{\partial U}{\partial x} + U\frac{\partial h}{\partial x} = 0 \qquad B = Cte \tag{1}$$

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + g\frac{\partial h}{\partial x} + g\frac{\partial z}{\partial x} = -gS_e \tag{2}$$

The energy slope, S_e, shall be expressed with a relationship established for uniform flow, by using a friction coefficient, $\boldsymbol{f}$, for a mobile bed (*see* **Chapter 137, Uniform Flow, Volume 4**), or:

$$S_e = f(\boldsymbol{f}, U, h) \tag{3}$$

where h is the flow depth, U is the average velocity of the flow, and $z(x,t)$ gives the elevation of the channel bed.

For flow over a *mobile bed* (see Figure 2), the elevation (level) of the channel bed, $z(x,t)$, may vary. According to the relation of *Exner* (see Graf, 1971, 1984, p. 288), such a variation can be expressed in form of a continuity equation for the solid phase, namely:

$$\frac{\partial z}{\partial t} + \left(\frac{1}{1-p}\right)\frac{\partial q_s}{\partial x} = 0 \tag{4}$$

where p is the porosity of the sediments of the bed, being defined as the ratio of the volume of empty space (occupied by water) and of the total volume. $q_s = C_s U h$ is the volumic solid discharge per unit width and C_s is the volumic concentration of the solid phase, being defined

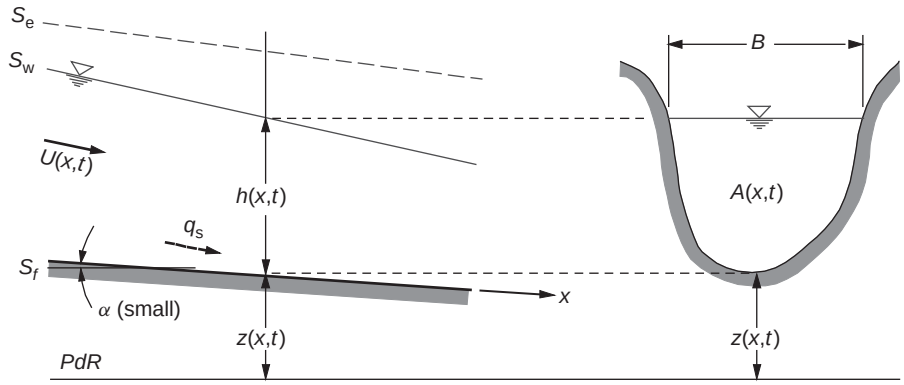


Figure 2 Scheme of unsteady and nonuniform flow over a mobile bed, $z(x, t)$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

by the ratio of the volume of the sediments and of the volume of the mixture. In general one admits that the solid discharge, q_s, is a function – still to be determined – of the liquid discharge, $q = Uh$, or:

$$q_s = f(U, h;\ \text{sediment}) \tag{5}$$

The three equations, equations (1), (2), and (4), contain three unknowns, $U(x, t)$, $h(x, t)$, and $z(x, t)$, with their independent variables, x and t. The two other unknowns, S_e and q_s, have to be expressed with semiempirical relationships. The five relations, equations (1), (2), and (4) together with equations (3) and (5), are the *equations of Saint-Venant–Exner*.

The liquid and the solid phase are *implicitly* coupled by the semiempirical relations, equations (3) and (5). After the solution for the liquid phase, equations (1) and (2), a solution for the solid phase, equation (4), can be obtained, giving the variation of the bed elevation, $z(x, t)$.

To obtain solutions to the equations of Saint-Venant–Exner, analytical methods can be used for simple problems, and numerical methods for complex problems.

Propagation of Perturbations

The propagation of a perturbation, being a wave of small amplitude on the mobile bed, can now be investigated by using the equations of Saint-Venant–Exner. For a rectangular channel, these equations are written as six equations of partial derivatives:

$$\frac{\partial h}{\partial t} + h\frac{\partial U}{\partial x} + U\frac{\partial h}{\partial x} = 0 \tag{1}$$

$$\frac{1}{g}\frac{\partial U}{\partial t} + \frac{U}{g}\frac{\partial U}{\partial x} + \frac{\partial h}{\partial x} + \frac{\partial z}{\partial x} = -S_e \tag{2}$$

$$(1-p)\frac{\partial z}{\partial t} + \frac{\partial q_s}{\partial U}\frac{\partial U}{\partial x} = 0 \tag{4}$$

with:

$$\frac{\partial h}{\partial x}\,\mathrm{d}x + \frac{\partial h}{\partial t}\,\mathrm{d}t = \mathrm{d}h$$

$$\frac{\partial U}{\partial x}\,\mathrm{d}x + \frac{\partial U}{\partial t}\,\mathrm{d}t = \mathrm{d}U$$

$$\frac{\partial z}{\partial x}\,\mathrm{d}x + \frac{\partial z}{\partial t}\,\mathrm{d}t = \mathrm{d}z$$

In writing equation (4), it was assumed that the solid discharge is only a function of the flow velocity, $q_s = f(U)$.

Upon mathematical manipulations of these six equations the following cubic characteristic polynomial (see Vries, 1973, p. 2) is obtained:

$$-c_w^3 + 2Uc_w^2 + \left(gh - U^2 + \frac{g}{1-p}\frac{\partial q_s}{\partial U}\right)c_w - \frac{gU}{1-p}\frac{\partial q_s}{\partial U} = 0 \tag{6}$$

where the absolute celerity (the characteristic) is defined by $c_w = \mathrm{d}x/\mathrm{d}t$.

This equation, equation (6), has evidently three real roots, thus three characteristics. Two roots, c_{w_1} and c_{w_2}, are an expression of the celerity of the perturbation (wave) on the water surface; the third root, c_{w_3}, gives the celerity of a perturbation (undulation) on the mobile bed. The celerity, c_{w_3}, is positive for subcritical flow, $U < \sqrt{gh}$; the form (undulation) of the bed, usually called *dunes*, displaces itself in the same direction as the flow (see Vries, 1973, p. 3). If, c_{w_3}, is negative for supercritical flow, $U > \sqrt{gh}$; the form (undulation) of the bed, usually called *antidunes*, displaces itself in the opposite direction of the flow.

It seems reasonable (see Vries, 1973, p. 4) to assume that for $Fr = U/\sqrt{gh} \neq 1$ the celerities, c_{w_1} and c_{w_2}, of

the waves on the surface are much larger than the celerity, c_{w_3}, of the undulations on the bed.

When studying the perturbations on the bed having such a weak celerity, c_{w_3}, it is now possible to consider the flow of the liquid phase as quasi-steady; thus

$$\frac{\partial U}{\partial t} = 0 \quad \text{and} \quad \frac{\partial h}{\partial t} = 0$$

Consequently, combining equation (1) with equation (2), one can write a single differential equation, which is the equation of the free-surface flow, or

$$\frac{\partial U}{\partial x}\left(U - g\frac{h}{U}\right) + g\frac{\partial z}{\partial x} = -gS_e \tag{7}$$

By eliminating $\partial U/\partial x$ between equation (7) and equation (4), one obtains

$$\frac{\partial z}{\partial t} + c_{w_3}\frac{\partial z}{\partial x} = -c_{w_3}S_e = \mathbf{F}(U) \tag{8}$$

where $\mathbf{F}(U)$ is a friction (roughness) term – being responsible for the decay of the perturbation on the bed – and

$$c_{w_3} = \frac{g}{(1-p)}\frac{(\partial q_s/\partial U)}{(gh/U - U)} = \frac{1}{(1-p)}\frac{U(\partial q_s/\partial U)}{h(1-Fr^2)} \tag{9}$$

where $Fr^2 = U^2/gh$ is the Froude number. For subcritical flow, when $Fr^2 \ll 1$, and expressing the solid discharge by a power law of the form

$$q_s = a_s U^{b_s} \quad \text{and} \quad \frac{dq_s}{dU} = b_s\left(\frac{q_s}{U}\right) \tag{10}$$

one may write

$$c_{w_3} = \frac{1}{(1-p)}b_s\frac{q_s}{h} \tag{9a}$$

It is to be noted (see equation 9) that the celerity of propagation of the undulations, c_{w_3}, on the bed is usually rather small compared to the average velocity, U, of the flow itself.

Analytical Solutions

To obtain analytical solutions to the equations of Saint-Venant–Exner, which are nonlinear and hyperbolic, is a difficult and often impossible task. However, simplifications are possible, if one assumes that for flow at small Froude numbers, $Fr < 0.6$, a *quasi-steadiness* is maintained. This hypothesis of a steadiness of flow can be justified: in general a variation of liquid discharge, $\partial(Uh)/\partial t$, is a short-term phenomenon, while a variation of the bed elevation, $\partial z/\partial t$, is a long-term phenomenon, which produces itself when the variation of the discharge has already disappeared; thus the flow may be considered reasonably constant, $q = Uh = Cte$. Under such conditions, solutions are of great interest, notably if one studies the variation of the bed, $z(x, t)$, as a long-term phenomenon.

Using the hypothesis of quasi-steadiness of the flow, a system of two differential equations can be written as

$$\frac{\partial U}{\partial x}\left(U - g\frac{h}{U}\right) + g\frac{\partial z}{\partial x} = -gS_e \tag{7}$$

$$(1-p)\frac{\partial z}{\partial t} + \frac{\partial q_s}{\partial U}\frac{\partial U}{\partial x} = 0 \tag{4}$$

These two equations are nonlinear ones; only numerical solutions are possible. For certain special cases, analytical solutions (after linearization) can be of help to understand the problem, notably the relative importance of the different terms.

If one further assumes (see Vreugdenhil and de Vries, 1973, p. 8) that the quasi-steady flow is also quasi-uniform, $\partial U/\partial x = 0$, the above equation, equation (7), becomes

$$0 + g\frac{\partial z}{\partial x} = -gS_e = -g\frac{U^2}{C^2 h} = -g\frac{U^3}{C^2 q} \tag{11}$$

where C is the coefficient of Chézy and $q = Uh$ is the unit discharge.

By eliminating $\partial U/\partial x$ after differentiation of equation (11) with respect to x, one obtains for the above equation, equation (4):

$$\frac{\partial z}{\partial t} - K(t)\frac{\partial^2 z}{\partial x^2} = 0 \tag{12}$$

where the coefficient (of diffusion), $K(t)$, being a function of time, is given by

$$K = \frac{1}{3}\frac{\partial q_s}{\partial U}\frac{1}{(1-p)}\frac{C^2 h}{U} \tag{13}$$

This model, equation (12), is a *parabolic* one and is limited to large values of x and of t, namely, for $x > 3h/S_e$ (see Vries, 1973). The expression for the coefficient, equation (13), can also be written – upon linearization, possible for $U \cong U_o$, where the index, o, refers to the uniform (initial) condition and using the power-law expression for the solid discharge, equation (10) – in the following way:

$$K \equiv \frac{1}{3}b_s q_s \frac{1}{(1-p)}\frac{1}{S_{e_o}} \tag{13a}$$

This parabolic model – obtained by using important assumptions – is of considerable interest, since it allows the obtaining of analytical solutions for certain well-defined cases. Depending on the applied mathematical techniques

and on the hypothesis used, some solutions have been communicated in the literature (see Graf and Altinakar, 1998, Chap. 6).

A *hyperbolic* model for quasi-steady flow, but being nonuniform, has been proposed (see Vreugdenhil and de Vries, 1973, p. 5):

$$\frac{\partial z}{\partial t} - K\frac{\partial^2 z}{\partial x^2} - \frac{K}{c_{w_3}}\frac{\partial^2 z}{\partial x \partial t} = 0 \tag{14}$$

where K and c_{w_3} are respectively given by equation (13) and equation (9). Since an analytical solution to equation (14) is rarely possible (see Vries, 1973), this model turns out to be not all that useful.

A model of a *simple wave* (see Vreugdenhil and de Vries, 1973) is obtained by reduction of equation (8), or:

$$\frac{\partial z}{\partial t} + c_{w_3}\frac{\partial z}{\partial x} = 0 \tag{15}$$

where c_{w_3} is given by equation (9). Since the friction term, $\mathbf{F}(U)$, is now neglected, the application of this model remains limited to small values of x and of t, namely, for $x \ll 3h/S_{e_o}$.

Degradation and Aggradation

A *degradation* (or *aggradation*) in a reach of a watercourse is encountered if the entering solid discharge is smaller (or larger) than the capacity of the transport of sediments. The sediments of the bed will be eroded (or deposited) and as a consequence the elevation of the channel bed decreases (or increases). Degradation (erosion) and aggradation (deposition) are long-term processes of the evolution of the channel bed, $z(x, t)$.

The flow, being steady and uniform at the beginning, will also be steady and uniform at the end of the process; in between the flow becomes nonuniform and quasi-steady. If one assumes, that during this transition the flow can be considered as being quasi-uniform, $\partial U/\partial x = 0$, one may make use of the *parabolic model* (equation 12). Note that this model is limited to large values of x and of t, namely, for $x > 3h/S_e$, and for Froude numbers of $Fr < 0.6$. Analytical solutions to the parabolic model (equation 12) can be obtained for cases in which the initial and boundary conditions are well specified. Such solutions will not only clarify a physical problem, but can often be considered as a first tentative for an understanding of the problem. Caution is, however, necessary since this model was established using different assumptions.

The analytical solutions will certainly help to explain the long-term evolution of the bed of the channel, when the variation of the liquid discharge can be readily neglected. A detailed presentation of this model (equation 12), as applied to degradation and aggradation, is found in the literature (see Graf and Altinakar, 1998, Chap. 6).

Numerical Solutions

Analytical solutions of the equations of Saint-Venant–Exner are only possible when the hypothesis of quasi-steadiness of the flow is justified. Furthermore, it is often necessary to assume also quasi-uniformity of the flow. However, these assumptions are *no* more possible, if the temporal variation of the discharge, $\partial(Uh)/\partial t$, and the one of the elevation of the bed, $\partial z/\partial t$, are of the same order of magnitude, namely, relatively rapid.

If the flow is unsteady and nonuniform or steady and nonuniform, no analytical solutions, which are reasonably simple, are available. The system of the equations of Saint-Venant–Exner can be resolved – without making too severe assumptions – by numerical methods; this may be well achieved with the use of computers.

The numerical methods are essentially the same as the ones that are used to solve the equations of Saint-Venant, namely, for flow over a *fixed* bed (*see* **Chapter 137, Uniform Flow, Volume 4** and **Chapter 138, Unsteady Flow, Volume 4**). They become, however, rather complicated if they are applied for the modelization of flow over *mobile* bed.

BED-LOAD TRANSPORT

Notions

Transport as bed load is the mode of transport of sediments (see Figure 1) where the solid particles glide, roll, or (briefly) jump, but stay very close to the bed, $0 < z < z_{sb}$. The displacement of the particles is intermittent.

There exist a number of formulae that can be used for the prediction of the bed-load transport (see Graf, 1971, 1984, Chap. 7; Yalin, 1972 Chap. 5; and Raudkivi, 1976, 1990, Chap. 7). Many of these formulae are of empirical nature, but often have incorporated dimensionless numbers.

Theoretical Considerations

Consider that the bed of a channel (see Figure 1) is plane but mobile, composed of solid particles of uniform size which are being noncohesive. These particles displace themselves under the action of the flow, which is uniform and steady. For such simplified conditions, one tries to obtain functional relations. The form of such functions, being often rather complex, will be established by experiments, which more or less will take care of the reality of the problem.

A dimensionless analysis, using the Π-theorem (see Yalin, 1972, p. 61), shows that the arguments that quantify

the bed-load transport can be expressed by four dimensionless quantities:

$$\Phi = f\left(Re_*, \tau_*, \frac{R_h}{d}, \frac{\rho_s}{\rho}\right) \tag{16}$$

Thus an expression for a dimensionless *intensity of the solid discharge* as the bed load is obtained, or:

$$\Phi \equiv q_{sb*} = \frac{q_{sb}}{\sqrt{(s_s - 1)gd^3}} \tag{17}$$

where $q_{sb}\,\mathrm{m^2\,s^{-1}}$ is the volumic solid discharge per unit width, $Re_* = u_* d/\nu$ is a Reynolds number of the particle, $\tau_* = \tau_o/[(\gamma_s - \gamma)d]$ a dimensionless shear stress, R_h/d a relative depth, and $s_s = \rho_s/\rho$ a relative density. Since some terms, R_h/d and ρ_s/ρ, are included in the term of τ_*, and taking $\tau_* = f(Re_*)$, one can formulate now a rather simple relationship:

$$\Phi = f(\tau_*) \text{ or } \frac{q_{sb}}{\sqrt{(s_s - 1)gd^3}} = f\left(\frac{\tau_o}{(\gamma_s - \gamma)d}\right) \tag{18}$$

which is often written as:

$$\Phi = f(\Psi) \tag{18a}$$

where $\tau_* \equiv \Psi^{-1}$ and Ψ is called the *dimensionless intensity of shear stress*, applied upon the solid particles.

This expression (equation 18) links the solid transport, q_{sb}, to the shear stress, τ_*. Thus an increase in τ_* – passing by $\tau_{*_{cr}}$, where erosion begins – is responsible for an increase in q_{sb}. The form of this function (equation 18) must still be established; it is given by the formulae of bed-load transport, which are established by experiments performed in the laboratory and in the field.

One often assumes that this relation (equation 18) can be expressed in the form of a power law:

$$\Phi = \alpha(\tau_*)^\beta \tag{19}$$

Subsequently an approximate relation (see Vries, 1973) in the form of

$$q_{sb} = a_s U^{b_s} \tag{10}$$

is derived, where a_s, α and $b_s = 2\beta$, β are the coefficients that depend essentially on the granulometry. This simple, but often useful, relation shows that the average velocity, U, of the flow is the predominant parameter for the determination of the solid discharge, q_{sb}.

Bed-load Relations

At present, the formulae for a determination of the solid discharge as bed load give only reasonably satisfying results within a domain of the parameters for which the chosen formula has been established. Consequently, the application and use of such formulae have to be done with great care. A selection of two of the many available formulae (see Graf, 1971, 1984, p. 133) will be given, namely, the ones by Meyer-Peter *et al.* and by Einstein.

From the different empirical formulae which *Meyer-Peter et al.* have developed in 1934 and 1948 (see Graf, 1971, 1984, p. 136), their last one is presented, namely:

$$0.25\rho^{1/3}\frac{(g'_{sb})^{2/3}}{(\gamma_s - \gamma)d} = \frac{\gamma R_{hb}\xi_M S_e}{(\gamma_s - \gamma)d} - 0.047 \tag{20}$$

where $g'_{sb} = g_{sb}\,(\gamma_s - \gamma)/\gamma_s$ is the solid discharge in weight under water and $g_{sb}/\gamma_s = q_{sb}$; R_{hb} is the hydraulic radius of the bed. For a nonuniform granulometry, the mean diameter, $d = d_{50}$, is taken as the equivalent diameter. This relation can also be written in the dimensionless form such as:

$$\Phi = 8(\xi_M \tau_* - \tau_{*_{cr}})^{3/2} \tag{20a}$$

where $\tau_{*_{cr}}$ is the dimensionless critical shear stress. $\xi_M = (K_s/K'_s)^{3/2}$ is the roughness parameter, where K'_s is the roughness of the granulates, evaluated with the formula of Strickler, and K_s is the total roughness of the bed, evaluated with the formula of Manning-Strickler, or:

$$K_s = \frac{U}{R_{hb}^{2/3} S_e^{1/2}} \quad \text{and} \quad K'_s = \frac{26}{d_{90}^{1/6}}$$

In the absence of bed forms, it is recommended to take $\xi_M = 1$; but $1 > \xi_M > 0.35$ if bed forms are present.

This relation (equation 20) is applicable for rather large grain sizes, $d > 2\,\mathrm{mm}$, being uniform as well as nonuniform, and for bed slopes, being moderate to strong (see Table 1).

Using the concepts of hydrodynamics, *Einstein* developed in 1942 and 1950 (see Graf, 1971, 1984, pp. 139–150) a probabilistic model for the transport of sediments as bed load. To express the probability of motion, Einstein (1950, p. 37) postulated the following function:

$$p_e = 1 - \frac{1}{\sqrt{\pi}} \int_{-B_*\Psi_* - 1/\eta_o}^{+B_*\Psi_* - 1/\eta_o} e^{-\xi^2}\,\mathrm{d}\xi \tag{21}$$

where ξ is a variable of integration. B_* and η_o are constants to be determined experimentally. Ψ_* is Einstein's intensity of shear stress, which takes into account the nonuniformity of the granulometry of the bed and its roughness as well as the velocity distribution of flow. For a uniform granulometry, one takes (see Einstein, 1950, p. 36):

$$\Psi_* = \Psi', \text{ with } \Psi' = \frac{(\gamma_s - \gamma)d}{\rho u'^2_*} = \frac{(\gamma_s - \gamma)}{\gamma}\frac{d}{R'_{hb} S_f} = \frac{1}{\tau'_*}$$

Subsequently Einstein (1950) elaborated an equation of bed load by postulating that the rate of erosion is equal to the rate of deposition. This results in the following expression:

$$\left(\frac{p_e}{1-p_e}\right) = A_* \left(\frac{i_{sb}}{i_b}\right) \Phi = A_* \Phi_* \qquad (22)$$

where Φ is the intensity of transport after Einstein (1950) given by equation (17) and A_* is a constant to be determined experimentally. i_{sb} is a fraction (see Figure 4) of the granulometric curve of the unit solid discharge, g_{sb}, in weight and i_b is a fraction (see Figure 4) of the granulometric curve of the bed material. p_e is the probability of erosion of a particle. For a uniform granulometry, one takes (see Einstein, 1950, p. 36) simply:

$$\Phi_* = \Phi$$

The above relations (equations 21 and 22), put together, give now the final form of the *equation of bed load of* Einstein (1950), or:

$$p_e = 1 - \frac{1}{\sqrt{\pi}} \int_{-B_*\Psi_* - 1/\eta_o}^{+B_*\Psi_* - 1/\eta_o} e^{-\xi^2} \, d\xi = \frac{A_*\Phi_*}{1 + A_*\Phi_*} \qquad (23)$$

namely, a functional relation (see equation 18a), such as:

$$\Phi_* = f(\Psi_*) \qquad (23a)$$

The (universal) constants have now to be determined experimentally both for uniform and nonuniform granulometries (see Einstein, 1950, pp. 37 and 43); they are given (see Graf, 1971, 1984, p. 149) as:

$$A_* = 43.6; \quad B_* = 0.143; \quad \eta_o = 0.5 \qquad (23b)$$

The above relation (equation 23) is plotted in Figure 3 together with the data of Meyer-Peter *et al.* and Gilbert. The graphical representation facilitates the use of the relation (equation 23). Since a nonuniform granulometry can be broken down into its fractions, i_{sb}/i_b, this relation is rather flexible. For a quasi-uniform granulometry, an equivalent diameter of $d = d_{35}$ can be taken. The equation of Einstein (equation 23) is well suitable for uniform and nonuniform granulates over a large range of diameters, $d > 0.7$ mm, and of bed slopes (see Table 1). It is used worldwide with great success. It is interesting to remark that the relation of *Einstein* (equation 23) and the one of *Meyer-Peter et al.* (equation 20) give rather similar results (see Graf, 1971, 1984, p. 150), and this notably for $\Phi < 10$.

Granulometry, Armoring

The noncohesive sediments, which make up the bed of a watercourse, are in general of different sizes, being given by the granulometric curve of the bed material (see Figure 4a). This curve can be divided into fractions (percentages), i_{b_i}. For each fraction the average diameter, d_i, is determined

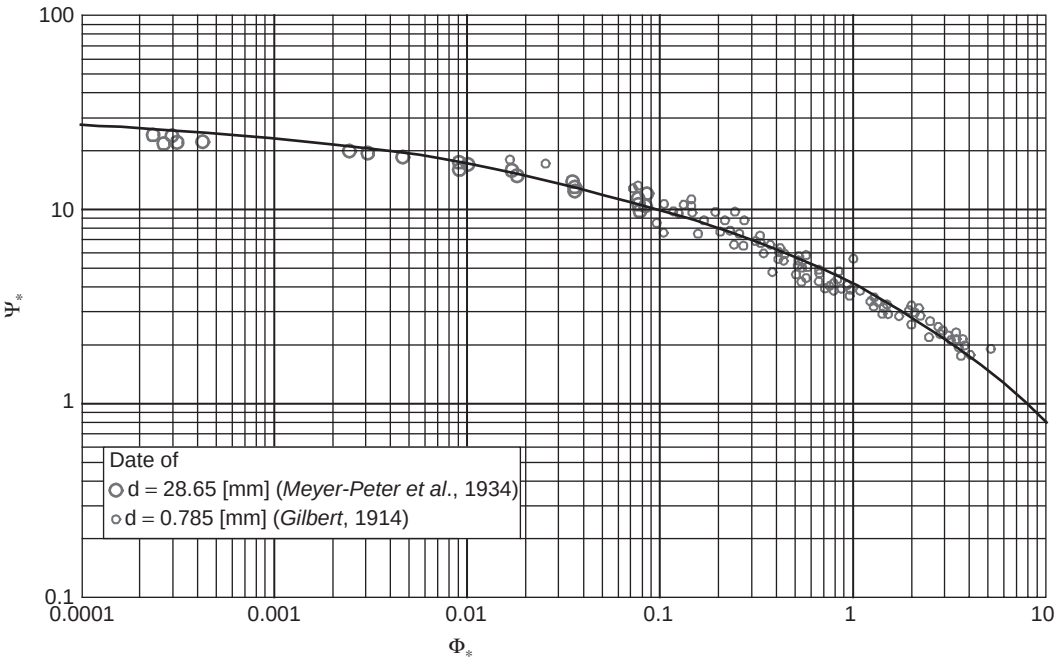


Figure 3 Equation of bed load, $\Phi_* = f(\Psi_*)$, of Einstein (see Graf, 1971, 1984, p. 148). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

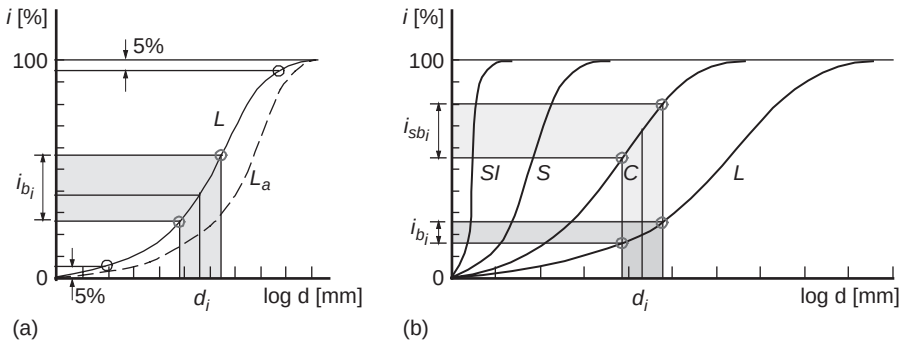


Figure 4 Scheme of granulometric curves for the bed material, L, and the armored bed material, L_a, the bed-load material, C, and the suspended (wash) load material, S (SI). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Table 1 Parameters used for establishing the different formulae

Formula	d [mm]	S_f [−]	d_x [mm], equivalent diameter for a nonuniform granulate
Meyer-Peter *et al.* (equation 20)	3.1–28.6	0.0004–0.020	d_m (d_{50})
Einstein (equation 23)	0.8–28.6	–	d_{35}
Graf and Acaroglu (equation 40)	0.3–1.7 (23.5)	–	d_{50}
Ackers and White (equation 43)	0.04–4.0	$Fr < 0.8$	d_{35}

and the corresponding solid discharge, $i_{sb_i} q_{sb_i}$, is calculated using one of the formulae for bed-load transport. For the entire granulometric mixture, the solid discharge is now obtained as: $q_{sb} = \sum i_{sb_i} q_{sb_i}$.

The granulometric curve of the bed material is in general different from the one of the material moving as bed load or as suspended load (see Figure 4b). Consequently, for an average diameter of the granulate, d_i, the given fraction of the granulometric curve of the bed material, i_{b_i}, will be different from the corresponding fraction of the granulometric curve of the solid discharge, i_{sb_i}. This subtlety was elaborated by Einstein (1950, p. 32) by introducing the ratio, i_{sb}/i_b, into the equation of bed-load transport (equation 23). For a very intensive sediment transport, all sizes (fractions) of particles will readily participate; consequently $i_{b_i} \equiv i_{sb_i}$, since the curves L and C become identical (see Figure 4b).

For cohesive material, the determination of the solid discharge represents a difficult task; the literature specialized on this topic should be consulted (see Graf, 1971, 1984, Chap. 12 and Raudkivi, 1976, 1990, Chap. 9).

On a channel bed of nonuniform granulometry, the smaller particles are more easily eroded than the larger ones and; a grain-size sorting takes place. An accumulation of the remaining larger particles results in an *armoring* of the bed, which subsequently protects the underlying "original" granulate (see Graf, 1971, 1984, p. 102). There exists only limited conclusive information about the ratio of the original granulometry, d, and the granulometry of the armor layer, d_a. For some Swiss rivers, having large bed slopes, $S_f > 0.03$ [−], and large grain sizes, $d_{50} > 6$ mm, an indicative relationship of:

$$\frac{d_{50_a}}{d_{50}} \approx 1.4; \quad \frac{d_{50_a}}{d_{90}} \leq 0.6$$

was developed by Correia and Graf (1988).

SUSPENDED-LOAD TRANSPORT

Notions

Transport of sediments in suspension is the mode of transport where the solid particles displace themselves by making large jumps, but remain in contact with the bed load and also with the bed. The zone of suspension is delimited by $z_{sb} < z < h$ (see Figure 1). The transport as suspended load could be considered as an advanced stage of transport as bed load; however, the analytical methods do not allow a description of these two modes of transport with the same (or single) relationship.

Theoretical Considerations

For steady uniform flow, the vertical distribution of the concentration of the suspended particles, $c_s(z)$, in the fluid can be obtained by using the equation of one-dimensional

diffusion-convection (see Graf, 1971, 1984, p. 166):

$$0 = v_{ss}\frac{\partial c_s}{\partial z} + \frac{\partial}{\partial z}\left(\varepsilon_s \frac{\partial c_s}{\partial z}\right) \tag{24}$$

where $c_s(z)$ is the local volumic concentration, ε_s is the diffusivity of the suspended particles, whose units are $[L^2\ T^{-1}]$, and v_{ss} is the settling velocity of the particles. Integration of the above equation (equation 24) yields:

$$v_{ss}c_s + \varepsilon_s \frac{dc_s}{dz} = Cte = 0 \tag{25}$$

where the constant of integration is taken to be $Cte = 0$, implying that $c_s = 0$ at the water surface for $\varepsilon_s = 0$. Equation (25) is valid only for weak concentrations, namely, for $(1 - c_s) \cong 1$ or $c_s < 0.1\%$. The above equation expresses that, at all levels, $z_{sb} < z < h$, the rate of sedimentation of particles per unit volume is equal to the rate of turbulent diffusion per unit volume.

In the open-channel flow, the turbulence and thus the diffusivity are vertically distributed, $\varepsilon_s(z)$. The distribution of the diffusivity – it was shown (see Graf and Altinakar, 1998, Chapt. 6) that the transport of matter and transport of momentum by turbulence are analogous and close to a solid boundary identical, $\varepsilon_s \approx \nu_t$ – is given (see Graf, 1971, 1984, p. 173) by:

$$\varepsilon_s = \kappa u'_* \frac{z}{h}(h - z) \tag{26}$$

This parabolic relation, established for unidirectional flow, has been experimentally verified (see Raudkivi, 1976, 1990, p. 170). Substitution of equation (26) into equation (25), and separation of the variables, yields:

$$\frac{dc_s}{c_s} = -\frac{v_{ss}}{\kappa u'_*}\left(\frac{h}{h - z}\right)\frac{dz}{z} \tag{27}$$

where one defines the *Rouse* exponent as:

$$\mathcal{Z} = \frac{v_{ss}}{\kappa u'_*} \tag{27a}$$

This expression (equation 27) can now be integrated by parts, within the limits of $a < z < h$ (see Graf, 1971, 1984, p. 173) and renders:

$$\frac{c_s}{c_{sa}} = \left(\frac{h - z}{z}\frac{a}{h - a}\right)^{\mathcal{Z}} \tag{28}$$

where c_{sa} is the concentration at a reference level, a. It gives the distribution of the relative concentration, c_s/c_{sa}, for one single particle size, v_{ss} and $\mathcal{Z}$. Note that in the definition of the Rouse exponent, $\mathcal{Z}$, the friction velocity, u'_*, due to the granulate must be used. In the Rouse exponent one should take the settling velocity, v_{ss}, of the particle in clear and quiescent water (see Graf, 1971, 1984, p. 45), thus being not influenced by turbulence or by concentration.

The equation for the distribution of the relative concentration (equation 28) for different values of the Rouse exponent, $\mathcal{Z}$, is shown in Figure 5. The following is to be observed:

– For small $\mathcal{Z}$-values, the relative concentration is large and tends to become uniform over the entire flow depth, h.

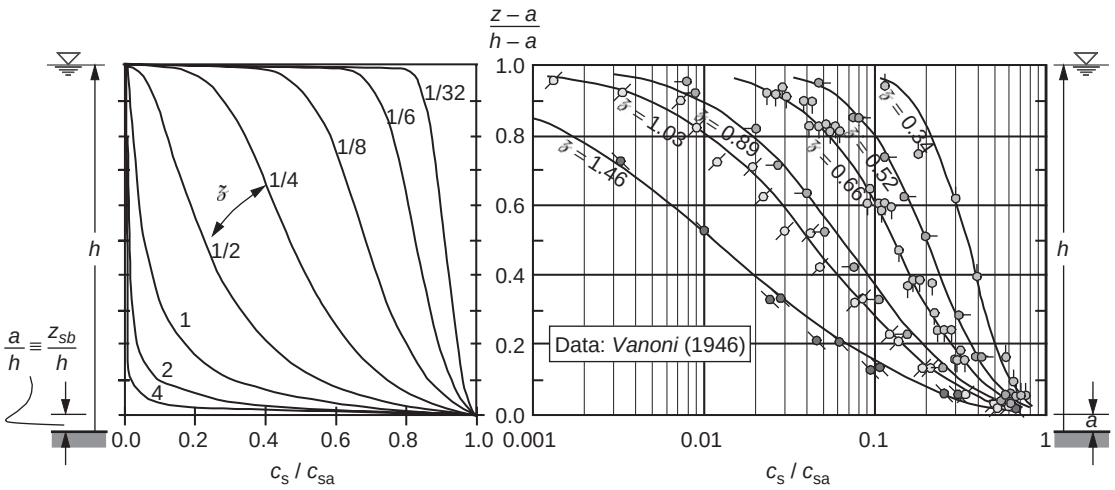


Figure 5 Vertical distribution of the relative concentration, c_s/c_{sa}, in a suspension. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

- For large $\mathcal{Z}$-values, the relative concentration is small at the water surface and is large close to the bed.
- The size of the particles, expressed with the settling velocity, v_{ss}, is directly responsible for these distributions.
- Close to the bed, $z \cong 0$,the concentration goes towards infinity, $c_s = \infty$, thus to an impossible value. Thus one delimits this level usually by $a \equiv z_{sb} \cong 0.05h$ or by $z_{sb} = 2d$, below which there exists the bed load (see Figure 1).
- The reference concentration, c_{sa}, is usually taken at a level of $a \equiv z_{sb}$; it can be calculated (see equation 34) with one of the bed-load formulae, q_{sb}.

Numerous investigations, both in laboratory and *in situ*, give evidence of the validity of the above equation (equation 28) (see Graf, 1971, 1984, p. 175).

Suspended-load Relation

The volumic solid discharge in suspension per unit width, in a region delimited by $z_{sb} < z < h$, is obtained by:

$$q_{ss} = \int_{z_{sb}}^{h} c_s u \, dz \tag{29}$$

where $c_s(z)$ is the local concentration (equation 28), and $u(z)$ is the local velocity. This relation is valid for a single particle size, d or v_{ss}. The distribution of the velocity shall be given by a logarithmic relation (see Einstein, 1950, p. 17), of the form:

$$u(z) = u'_* 5.75 \log\left(30.2 \frac{z}{\Delta}\right) \tag{30}$$

where $\Delta = k_s/\chi$ is a correction term, given in Figure 6, and u'_* is the friction velocity due to the granulate.

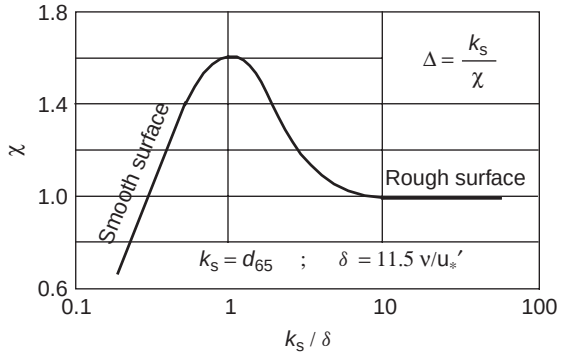


Figure 6 Correction coefficient of velocity distribution

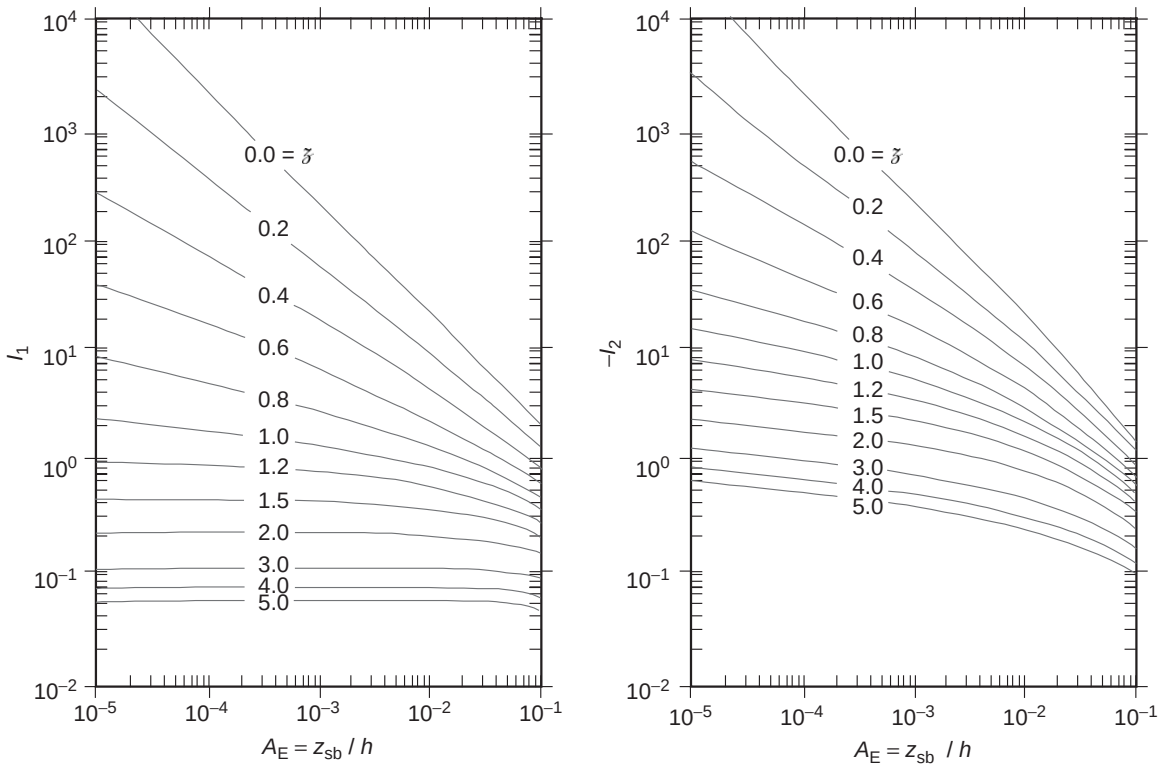


Figure 7 The integrals, $I_1(A_E, \mathcal{Z})$ and $I_2(A_E, \mathcal{Z})$, used in the method of Einstein (1950). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Upon substitution of equations (28 and 30) into the above equation, equation (29), one obtains:

$$q_{ss} = \int_{z_{sb}}^{h} c_{sa} \left(\frac{h-z}{z} \frac{a}{h-a} \right)^{\mathcal{Z}} u'_* 5.75 \log \left(30.2 \frac{z}{\Delta} \right) dz \tag{31}$$

Replacing $(a + z_{sb})$ by a dimensionless expression, $z_{sb}/h = A_E$, using h as the unity of z (see Einstein, 1950, p. 18), and after some mathematical manipulations, one gets:

$$q_{ss} = c_{sa} u'_* 5.75 h \left(\frac{A_E}{1-A_E} \right)^{\mathcal{Z}} \times \left\{ \log \left(30.2 \frac{h}{\Delta} \right) \int_{A_E}^{1} \left(\frac{1-z}{z} \right)^{\mathcal{Z}} dz + 0.434 \int_{A_E}^{1} \left(\frac{1-z}{z} \right)^{\mathcal{Z}} \ln z \, dz \right\} \tag{32}$$

The values of the following integrals

$$I_1 = 0.216 \frac{A_E^{\mathcal{Z}-1}}{(1-A_e)^{\mathcal{Z}}} \int_{A_E}^{1} \left(\frac{1-z}{z} \right)^{\mathcal{Z}} dz;$$

$$I_2 = 0.216 \frac{A_E^{\mathcal{Z}-1}}{(1-A_e)^{\mathcal{Z}}} \int_{A_E}^{1} \left(\frac{1-z}{z} \right)^{\mathcal{Z}} \ln z \, dz$$

are numerically evaluated (see Einstein, 1950, p. 19–24) and this for different values of A_E and $\mathcal{Z}$; they are given in Figure 7. Finally, the above equation, equation (32), can be put into the following form:

$$q_{ss} = 11.6 \, c_{sa} u'_* z_{sb} \left[2.303 \log \left(30.2 \frac{h}{\Delta} \right) I_1 + I_2 \right] \tag{33}$$

where q_{ss} is the volumic solid discharge per unit width of the suspended load.

The *reference concentration*, c_{sa}, shall be taken there, where the concentration distribution (equation 28) lacks any physical sense, particularly very close to the bed. It will thus be positioned within the layer where the bed load moves (see Figure 1). One usually assumes (see Graf, 1971, 1984, p. 191) that the thickness of this layer, called *bed layer*, is twice the grain diameter, $z_{sb} \approx 2d$; for a granulometric mixture, the bed layer takes different values for each granulometric fraction. It is now necessary to establish a relation between the bed-load and suspended-load transport; the reference concentration, c_{sa}, will make this link. Exploiting experiments, Einstein (1950, p. 40) proposed an expression for the reference concentration:

$$c_{sa} = \frac{q_{sb} i_{sb}}{11.6 \, u'_* z_{sb}} \tag{34}$$

Consequently, the solid discharge as suspended load per unit width (equation 33) is given by:

$$q_{ss} i_{ss} = q_{sb} i_{sb} \left[2.303 \log \left(30.2 \frac{h}{\Delta} \right) I_1 + I_2 \right] \tag{35}$$

$q_{ss} i_{ss}$ being the volumic solid discharge per unit width of the suspended load for one single granulometric fraction. This relation (equation 35) establishes the link between bed-load and suspended load transport for all particle sizes, which are found in the granulometric fraction of the bed load.

TOTAL-LOAD TRANSPORT

Notions

Total-load transport of sediments – or better called *total bed-material load transport* – is made up of transport as bed load (see Section "Bed-load transport") and of transport as suspended load (see Section "Suspended-load transport") or:

$$q_s = q_{sb} + q_{ss} (+ q_{sw}) \tag{36}$$

Added should (possibly) be the transport as wash load, q_{sw}.

Different formulae (see Graf, 1971, 1984, Chap. 9 or White *et al.*, 1973) exist, which can be used for the prediction of the bed-material load in a watercourse. The formulae for determination of the total load – just as the ones of the bed load – give only reasonable results in the domain of their established parameters (see Table 1). Thus an application of any formula must be done with great care. Here we will give a selection of some existing formulae.

Total-load Relations

Einstein (1950, p. 40) proposed a formula for bed-load transport (equation 23) and one for suspended-load transport (equation 35); combining these two relations, it is possible to get a formula for the bed-material load transport:

$$q_s i_s = q_{sb} i_{sb} + q_{ss} i_{ss} = q_{sb} i_{sb} \left[1 + 2.303 \log \left(30.2 \frac{h}{\Delta} \right) I_1 + I_2 \right] \tag{37}$$

This relation gives the sediment-transport capacity, but does not, of course, include the wash-load transport. This formula (equation 37) can be used if the hydraulic and sedimentological parameters are known in advance. If in addition a measurement of the suspended load is also available, there exists a modified version (see Graf, 1971, 1984, p. 207) of the above relation.

In many ways, the indirect method of Einstein (equation 37) is hydraulically rather complete, but its application might seem laborious.

Graf and Acaroglu (1968) developed a relation for the direct prediction of the bed-material transport, valid for open-channel flow (but also for flow in pipes).

A *parameter of shear intensity* was elaborated as a criteria of solid transport (see Graf, 1971, 1984, pp. 218 et 443):

$$\Psi_A = \frac{(s_s - 1)d}{S_e R_h} \tag{38}$$

which is the inverse of the dimensionless shear stress. Applying the concept of power (work) of a flow system, a *parameter of transport* was proposed (see Graf, 1971, 1984, pp. 219 and 446):

$$\Phi_A = \frac{C_s U R_h}{\sqrt{(s_s - 1)g d^3}} = \frac{(q_s/q) U R_h}{\sqrt{(s_s - 1)g d^3}} \tag{39}$$

which is similar to the dimensionless intensity of solid discharge, given by equation (17).

It could be shown that a functional relation between these parameters, Ψ_A and Φ_A (see equation 18a), whose form was experimentally determined, is possible. Using close to 800 experiments from the laboratory and close to 80 experiments in the field (see Table 1), all for free-surface flow (and close to 300 experiments for pipe-line flow), the following relationship (see Graf, 1971, 1984, pp. 220 and 448) was established:

$$\Phi_A = 10.39(\Psi_A)^{-2.52} \tag{40}$$

This relationship is found valid for $10^{-2} < \Phi_A < 10^3$ or for $\Psi_A \leq 14.6$; it is also valid when taking an equivalent diameter, $d \cong d_{50}$, if the granulometry is a nonuniform one.

Ackers and White (1973) proposed a direct determination of the total-load transport, q_s, using some sedimentological parameters.

A *parameter of mobility* of sediments was defined as:

$$F_{gr} = \frac{u_*^{n_w}}{\sqrt{(s_s - 1)gd}} \left[\frac{U}{\sqrt{32}\log(10h_m/d)} \right]^{(1-n_w)} \tag{41}$$

which becomes $F_{gr} = \sqrt{\tau_*}$ for very fine particles, where $n_w = 1$. A *parameter of transport* of sediments was postulated as:

$$G_{gr} = C_w \left(\frac{F_{gr}}{A_w} - 1 \right)^{m_w} \tag{42}$$

The total-load transport is now calculated according to:

$$C_s = \frac{q_s}{q} = G_{gr} \frac{d}{h_m} \left(\frac{U}{u_*} \right)^{n_w} \tag{43}$$

where C_s is the volumic average concentration in a section and $h_m = A/B$ is the average flow depth. The coefficients in the above relations were determined by regression analysis, using close to 1000 experiments in the laboratory and close to 250 experiments in the field, with sediments having a uniform and a nonuniform granulometry, $0.04 < d_{50}\,mm < 4.0$ and for flow at $Fr < 0.8$ (see Table 1). The resulting values of these coefficients are the following:

co-efficient	$d_* > 60$ $d > 2.5$[mm]	$1.0 < d_* \leq 60$	$d_* < 1$ $d < 0.04$[mm]
n_w	0.0	$(1 - 0.56 \log d_*)$	1.0
m_w	1.50	$(9.66/d_*) + 1.34$	
A_w	0.17	$0.23/\sqrt{d_*} + 0.14$	
C_w	0.025	$\log C_w = 2.86 \log d_* - (\log d_*)^2 - 3.53$	

Above, the dimensionless particle diameter, $d_* = d[(s_s - 1)(g/\nu^2)]^{1/3}$, is used. For a nonuniform granulometry, one takes $d = d_{35}$ as the equivalent diameter.

Application of Relations

Different formulae for the determination of the solid transport have been presented. However, none of these relations can pretend to translate the intrinsic complexity of the transport of sediments. Most of these formulae should not be used beyond the conditions within which they were established. Table 1 contains a summary of the range of the parameters, d and S_f, investigated for the establishment of each formula. Also listed are the recommendation by the authors for the choice of the equivalent diameter, d_x, if the granulometry is quasi or nonuniform.

The formulae for the transport of sediments are often established using laboratory data and less often using field data. A verification of these formulae in watercourses is a very delicate task, since it is difficult to measure correctly the solid discharge in the field. Furthermore, it is often a rather subjective evaluation, since the zones of the modes of transport cannot easily be separated.

Numerous studies have been reported comparing measurements in watercourses with the different existing formulae. Many (19) of the existing formulae for the calculation of the total transport have been studied by White *et al.* (1973) and compared with experimental results. They evaluated almost 1000 laboratory experiments with uniform and nonuniform sediments of $0.04 < d_{50}\,[mm] < 4.9$, at flow depth of $h < 0.4$ m, and almost 270 experiments in watercourses with sediments of $0.1 < d\,mm < 68.0$ and a width/depth ratio of $9 < B/h < 160$. Each formula was applied to all the data of the solid-discharge measurements. Subsequently a ratio of the values calculated, C_{calc}, and the values observed, C_{obs}, where $C \equiv C_s$ is the total-load transport, expressed in concentration, was established. Some

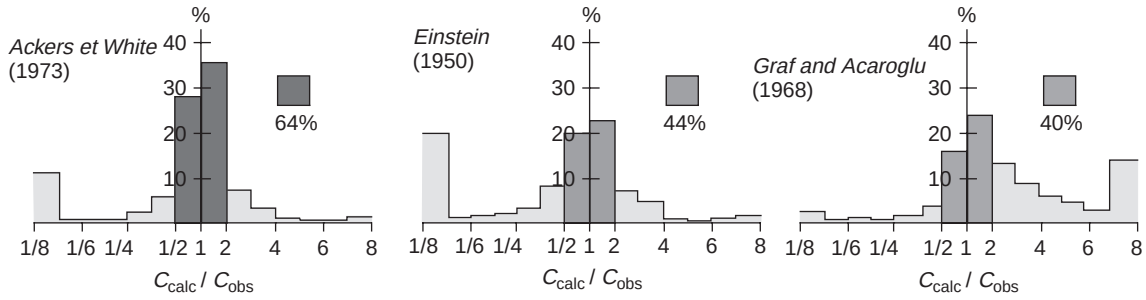


Figure 8 Comparison, with respect to C_{calc}/C_{obs}, of the success of prediction for the presented formulae

results of this investigation are given in Figure 8, where one may see the success of a prediction (in percentage) for different ranges of the ratio, C_{calc}/C_{obs}. For the formulae, which are presented in this article – considering only the range of $1/2 < C_{calc}/C_{obs} < 2$ – it can be seen that the percentage of success for the different formulae are as follows: is for the formula of:

Einstein (1950), equation 37:	44% of success
Graf and Acaroglu (1968), equation 40:	40% of success
Ackers and White (1973), equation 43:	64% of success

This implies that with the formula of Ackers and White, 64% of the experimental data can be predicted in the above-mentioned range. This is usually considered as a good (or a not-so-bad) result; more than half of the studied (19) formulae give results which are not as good, namely <40%.

The comparative study of White *et al.* (1973) is reasonably objective, but certainly not conclusive. Other studies exist (see Raudkivi, 1976, 1990, p. 227) that show clearly that an objective validation is nearly impossible.

Among the different existing formulae for the determination of the total-load transport, but equally for the ones of the bed-load and suspended-load transport, each one will give an answer, but none will be very precise nor very true. Thus the results obtained with these formulae give only valuable guidelines for the engineer. For practical purposes, it is advised to consult more than one formula; the obtained result may, however, render different values (see Graf, 1971, 1984, p. 156).

Wash Load

The wash load, q_{sw}, contains all these particles that are never in contact with the bed and displace themselves by being carried (washed) through the channel by the flow (see Figure 1). This mode of the transport of sediments is limited to the very finest particles, which are rare in the granulometry of the bed material. The distribution of these particles is rather uniform over the entire flow depth. Einstein (1950, p. 7) has proposed that the granulometry of the wash load is the fraction of granulometry of the bed that is smaller than 10%.

Since there exists no physical relationship to the flow, it has been difficult to advance an analytical method for the determination of the wash load. At present, no methods exist for the prediction of the wash load. The wash load depends more on the hydrological, geomorphological, and meteorological conditions within the drainage basin (see Graf, 1971, 1984, p. 232), namely on the overland surface erosion and less on the erosion in the stream bed. In order to obtain quantitative information on the wash load, measurements in the field must be performed. One measures thus the total suspended load, $q_{ss} + q_{sw}$. Subsequently the suspended load, q_{ss} (see Section Suspended-load transport), is calculated, and consequently the suspended wash load, q_{sw}, can be obtained.

REFERENCES

Ackers P. and White W.R. (1973) Sediment transport: new approach and analysis. *Proceedings of the American Society of Civil Engineers*, Vol. 99, HY11.

Correia L. and Graf W.H. (1988) *Grain-size Distribution and Armoring in Gravel-Bed Rivers: A Case Study*, Rapport Annuel, Laboratoire de Recherches Hydrauliques: EPF-Lausanne.

Einstein H.A. (1950) *The Bed-load Function for Sediment Transportation in Open Channel Flows*, US Department of Agriculture, Soil Conservation Service: T.B. No. 1026, Washington.

Graf W.H. (1971; 1984) *Hydraulics of Sediment Transport*, McGraw-Hill: New York; Water Resources Publication: Littleton.

Graf W.H. and Acaroglu E.R. (1968) A physical model for sediment transport in conveyance systems. *Bulletin of the International Association of Scientific Hydrology*, **13**(3B).

Graf W.H. and Altinakar M.S. (1998) *Fluvial Hydraulics*, John Wiley & Sons: Chichester.

Raudkivi A.J. (1976; 1990) *Loose Boundary Hydraulics*, Pergamon Press: Oxford.

Takahashi T. (1991) *Debris Flows*, *IAHR-Monograph*, Balkema Publication, Rotterdam.

Vreugdenhil C.B. and de Vries M. (1973) *Analytical Approaches to Non-Steady Bedload Transport*, Report S-78/N, Delft Hydraulics Laboratory.

Vries M. (1973) *River-Bed Variations – Aggradation and Degradation*, Publication No. 107, Delft Hydraulics Laboratory.

Wan Z. and Wang Z. (1994) *Hyperconcentrated Flow*, *IAHR-Monograph*, Balkema Publication: Rotterdam.

White W.R., Milli H. and Crabbe A. (1973) *Sediment Transport: An Appraisal of Available Methods*, Hydraulics Research Station Report No INT 119, Wallingford.

Yalin M.S. (1972) *Mechanics of Sediment Transport*, Pergamon Press: Oxford.

141: Computer Modeling of Overbank Flows

ALAN ERVINE[1] AND GARETH PENDER[2]

[1]*Department of Civil Engineering, University of Glasgow, Glasgow, UK*
[2]*School of the Built Environment, Heriot–Watt University, Edinburgh, UK*

Computational river models are in widespread use for estimating flood extent and the effectiveness of proposed flood mitigation schemes. This paper investigates the application of commonly used computational models to river floods, highlighting the advantages and shortcomings and current research activity. 1D, quasi-2D, 2D, and 3D computational models are applied to simulate a large-scale physical model tested in the Flood Channel Facility (FCF), Wallingford, as well as field data from the River Severn near Shrewsbury. The results reveal that commonly used 1D models give a reasonable representation of flood levels and flood extent but do not include floodplain/main channel interaction losses associated with compound flows and can be in error in their predictions. This deficit can be overcome in a straightforward manner by using a modified conveyance at each cross section based on the depth-integrated turbulent form of the Navier–Stokes equation. This so-called Lateral Distribution Method (LDM) has been incorporated into a new Conveyance Estimation System (CES) funded by the EA/Defra Joint Research Programme in Flood and Coastal Defence and developed by HR Wallingford Ltd for use with 1D models. Alternatively, 2D models can be more useful when simulation of flood inundation extent is important. The use of 3D models is becoming more common due to their ability to simulate very complex flow patterns over short river reaches. Initial results indicate that the choice of mesh discretization in a 3D model has a significant influence on the ability of the model to reproduce secondary velocity fields correctly. Problems of data requirements, friction and free surface representation still persist in 3D models in general.

INTRODUCTION

The use of computer modeling in flood risk management is continuing to grow. This is related to the rapid increase in desktop computing facilities and the consequent decrease in the cost of computations. Indeed, decisions for many flood risk management strategies are now based exclusively on results from computer model predictions. Inherent in this observation is the assumption that computer simulations are accurate as well as cost-effective; otherwise, basing important strategic decisions on their output would be foolish. The results of computer simulations are only as good as the physical laws incorporated in the governing equations and the boundary conditions used to drive them. From the following, it is clear that these continue to improve.

ONE-DIMENSIONAL (1D) MODELS

These models are used widely for the simulation of flood wave propagation over relatively long lengths of rivers. They are well proven and work well for what they are designed for, that is, calculating flows and water levels during the propagation of floods through the river system. Their governing equations are based on the assumptions that water pressure is hydrostatic, water level is horizontal across the channel, and that the stream-wise component of velocity is dominant. This means that only two equations describing the conservation of mass and momentum along the river's longitudinal axis need be used. Yen (1973) provides a formal derivation of these equations, often referred to as the St. Venant equations. They can be written as,

$$B\frac{\partial \eta}{\partial t} + \frac{\partial Q}{\partial x} = q \tag{1}$$

Conservation of mass

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\beta Q^2}{A}\right) + gA\frac{\partial \eta}{\partial x} + gAS_{\mathrm{f}} = 0 \tag{2}$$

Conservation of momentum

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

where Q is the river discharge, η is the water elevation, B is the flow top breadth, q is the lateral inflow per unit length, A is the flow area, R is the hydraulic radius, S_f is the friction slope, and β the momentum correction factor, which accounts for the fact that the velocity distribution over the cross section is nonuniform. The friction slope can be evaluated using any of the uniform flow friction laws.

$$Q = K(S_f)^{1/2} \tag{3}$$

where K is termed the channel conveyance.

Using Manning's equation

$$Q = \frac{AR^{2/3}}{n}(S_f)^{1/2} \tag{4}$$

$$K = \frac{AR^{2/3}}{n} \tag{5}$$

At certain river sections, such as bridges, weirs, local expansions or contractions, equations (1) and (2) do not apply and empirical laws must be used. These equations, when furnished with suitable initial conditions, may be solved by a numerical method to give flows and stages throughout the modeled river reach. Boundary conditions usually consist of specified flow hydrographs at upstream boundaries and specified water stages or rating curves at downstream boundaries, although alternative boundary conditions are possible. Cunge, *et al.* (1989) give a fuller description of the method and techniques involved.

In recent years, research in 1D modeling has focused on improving the physical representation of energy loss through the development of enhanced conveyance estimation techniques and better prediction of lateral floodplain inundation.

At present, cross-sectional conveyance for 1D river modeling is estimated using the divided channel method and a uniform flow law such as Manning's equation, Figure 1. In this approach, Manning's 'n' is a lumped parameter that attempts to account for all energy losses occurring between each model cross section. The advances in quasi-2D modeling described in the next section have provided the opportunity to improve upon this technique and explicitly include in 1D models the energy loss due to lateral shear stress and other second order effects such as flow expansion, contraction, and recirculation. The method underpins the calculation adopted in the newly developed Conveyance Estimation System (CES), HR Wallingford (2003). Here, equation (6) is solved to provide the depth-averaged velocity distribution across a channel cross section throughout the full range of flood stages. These are then integrated to provide a stage-discharge curve. This can be converted to a stage-conveyance curve using a representative slope (normally taken as the bed slope). Unlike the stage-conveyance curve obtained from the divided-channel method, this technique provides a stage-conveyance curve that includes the influence of secondary energy loss mechanisms. The CES software is currently being evaluated; however, examples of the application of the method can be found in Ervine and MacLeod (1999) and Forbes and Pender (2000).

The assumption that the water surface is horizontal at each channel cross section limits the ability of 1D models to simulate the exchange of water between the main channel and the floodplain. Traditionally, a storage cell approach has been adopted (Pender and Ellis, 1990; Pender, 1992), where the storage ponds are relatively large polygons hydraulically linked to the main channel and each other by weir or uniform flow laws. The increased availability of high-resolution, remotely sensed floodplain topography data has provided opportunities to develop enhanced techniques for the simulation of 2D flood inundation. These fall into two categories – raster flood routing and hydrodynamically linked 1D/2D modules. Considerable effort is currently going into the development and enhancement of raster models. Bates and De Roo (2000) describe the explicitly formulated raster flood routing model LISFLOOD-FP; whereas, Sen (2002) describes an alternative implicit formulation. The popularity of this technique arises from the

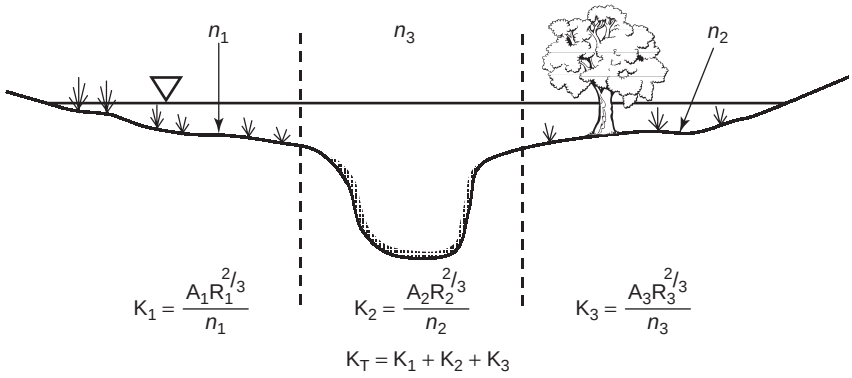


Figure 1 Divided Channel approach in flood modeling

fact that the data format means that results are easily integrated into commercial GIS software to provide a powerful tool for flood risk management decision makers, Wicks *et al.* (2003).

QUASI-2D MODELS

For over-bank flows, the depth-averaged velocity is clearly more complex, as sketched in Figure 2. A one-dimensional model is unable to predict the lateral distribution of velocity or shear stress, and hence the effects of shape and channel/floodplain interaction are not accounted for in 1D models. Improved methods for computing lateral distribution of velocity and shear stress are based on a quasi-2D approach (Wark *et al.*, 1990; Shiono and Knight, 1991; Ervine *et al.*, 2000). These methods are now being incorporated into a CES (HR Wallingford, 2003). Most lateral distribution methods (LDM) are based on the depth-integrated form of the Navier–Stokes equations:

$$\rho g H S_o - \frac{1}{8}\rho f U_d^2 \left(1 + \frac{1}{s^2}\right)^{1/2} + \frac{\partial}{\partial y}\left\{\rho \lambda H^2 \left(\frac{f}{8}\right)^{1/2} U_d \frac{\partial U_d}{\partial y}\right\} = \frac{\partial}{\partial y}\{H(\rho \overline{UV})_d\} \quad (6)$$

The first term in equation (6) represents the weight of fluid component in the longitudinal direction, the second term the bed friction, the third term the lateral shear (eddy viscosity), and the right-hand side term the secondary currents (3D mixing). Equation (6) was solved by Ervine *et al.* (2000) using the assumption that the temporal mean velocity components on the left-hand side of equation (6) are related to the depth-averaged velocity in a simplistic way by:

$$\overline{U} = K_1 U_d \quad (7)$$

$$\overline{V} = K_2 U_d \quad (8)$$

and by extension;

$$\overline{UV} = C_{uv} U_d^2 \quad (9)$$

The term $C_{uv}U_d{}^2$ is used to simulate a complex 3D mixing process that includes a horizontal shear layer, mass exchange in and out of the main channel, as well as expansion and contraction losses. C_{uv} is an empirical coefficient that varies with the geometry and roughness of the boundaries. Clearly, channel geometry has a large part to play in the magnitude of secondary cells. In each case, the transverse component V is closely associated with the depth-averaged velocity U_d, although the exact relationship is unknown. Ervine *et al.* (2000) showed that this simple approximation, when combined with good estimates of transverse turbulent shear stress, could produce much more satisfactory prediction of depth-averaged velocities.

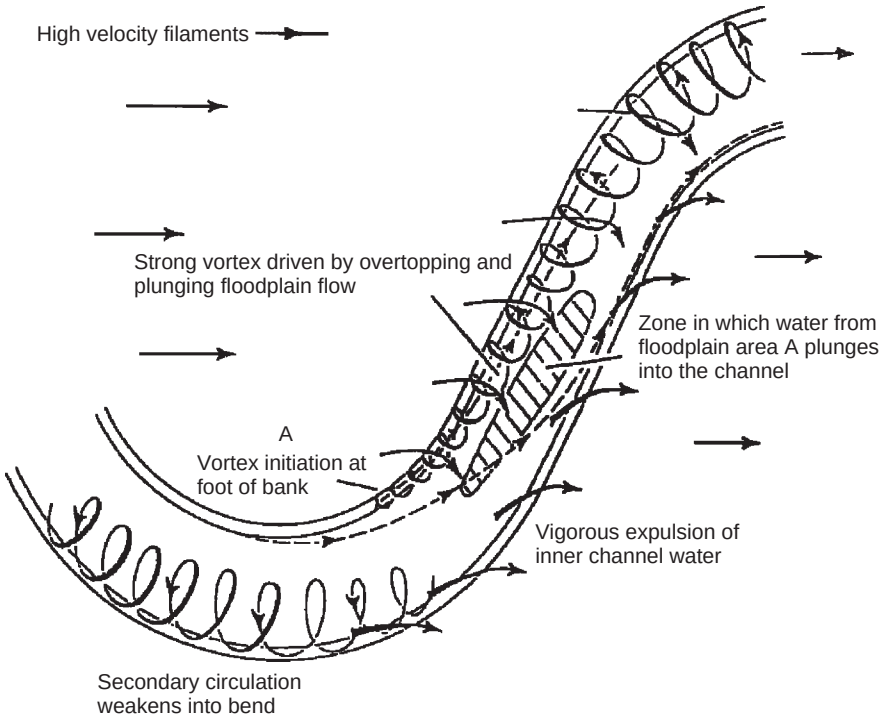


Figure 2 Complex flow patters in meandering overbank flows

To illustrate this, equations (6) and (9) are combined and applied to Flood Channel Facility FCF-Series B data with the comparison shown in Figure 3. This shows the depth-averaged velocity prediction for test B23 (60° cross-over angle meander, with trapezoidal main channel and smooth flood plains, depth = 0.2 m) at a bend apex. Clearly, a C_{uv} value of 5.5% greatly improves predictions at least in the main channel.

A more effective test of the validity of combining equations (6) and (9) is to compare the output with measurements of boundary shear stress. The present analytical solution is used to predict the boundary shear stress distribution

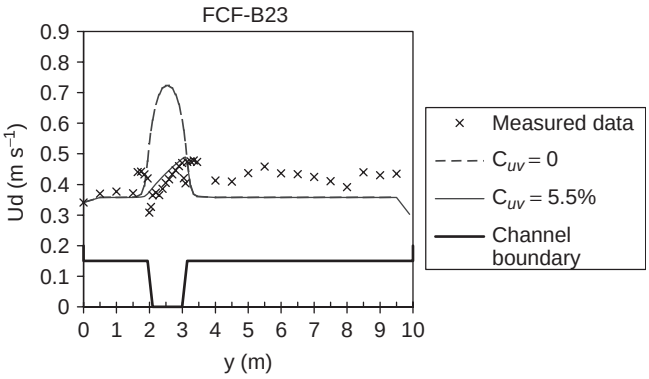


Figure 3 Comparison between analytical and experimental lateral distribution of depth-averaged velocity for meandering overbank flow. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

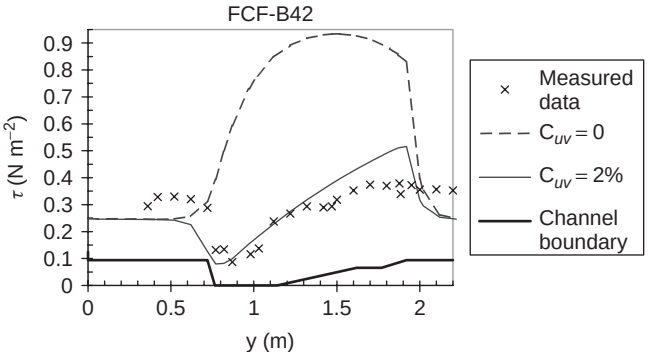


Figure 4 Comparison between analytical and experimental lateral distribution of boundary shear stress distribution for meandering overbank flow. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

for other FCF data. Figure 4 shows the boundary shear stress distribution for experiment FCF-B42, the quasi-natural cross section, focusing in the region of the main channel and both excluding and including the secondary cell term in the main channel. The predictions in Figure 4 focus in the region of the quasi-natural main channel area and only prove satisfactory when the secondary flow term is included.

This method has been applied to predict depth-averaged velocity values in River Severn at Lower Farm near Shrewsbury during a flood of approximately 103 m^3 s^{-1}. The reach modeled is shown in Figure 5. Figure 6 plots the

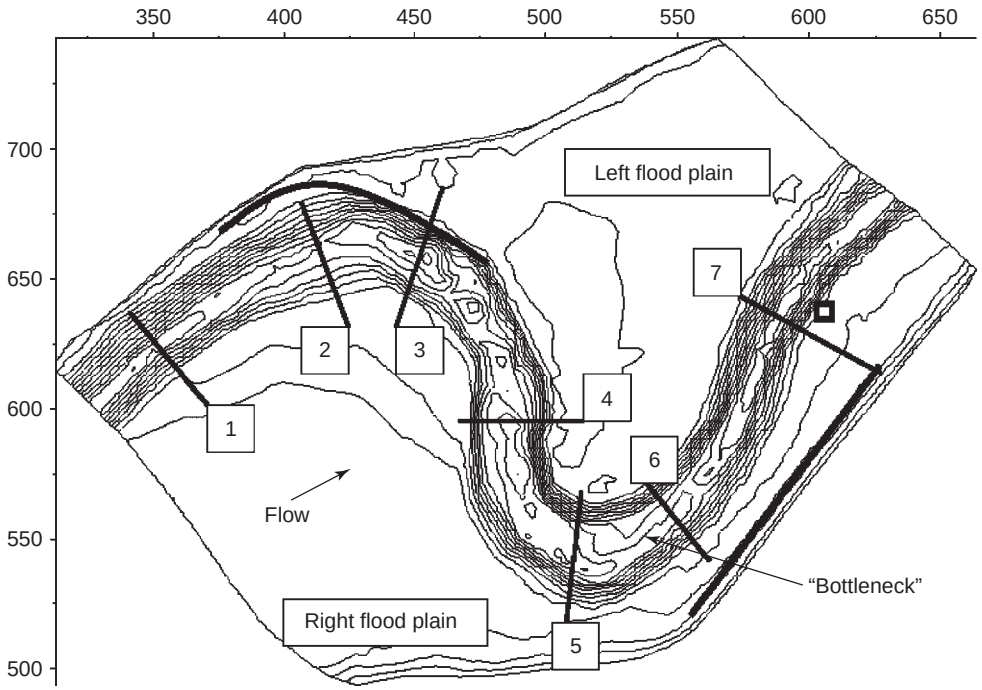


Figure 5 Site of three-dimensional flood measurements on River Severn

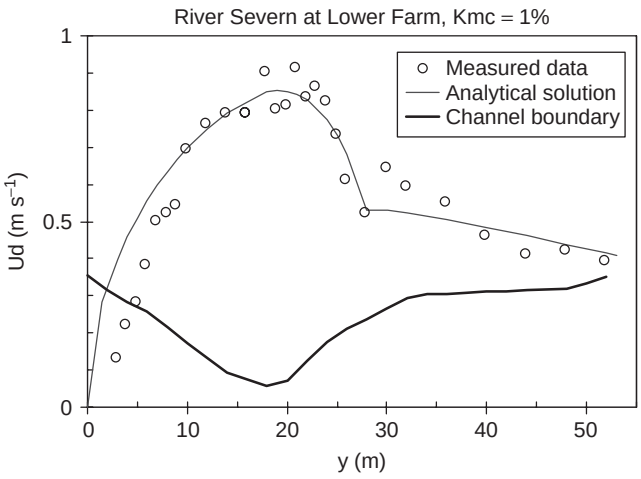


Figure 6 Comparison of measured and computed depth-averaged velocity at the River Severn for C_{uv} of 1%. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

predicted-against-measured depth-averaged velocity using the secondary current term $C_{uv} = 1\%$ for measurements at cross section 7. A reasonable correlation is revealed emphasizing the importance of secondary current term in predicting depth-averaged velocity values. The reduced value of $C_{uv} = 1\%$ reflects the distance downstream of cross section 7 from the previous bend. One would expect that the parameter C_{uv} would reflect the intensity of secondary cells in different channels and, therefore, the C_{uv} value for a meandering compound channel would be much greater than the one for a straight compound channel. The findings obtained so far indicate that $0.2\% < C_{uv} < 0.5\%$ for straight compound flows, and $2\% < C_{uv} < 5\%$ for meandering compound flows at least at a bend apex cross section.

TWO-DIMENSIONAL (2D) MODELS

When the river geometry results in both horizontal velocity components being important and strong vertical mixing promoted by bed roughness, then a depth-averaged form of the Reynolds' Averaged Navier–Stokes equations is appropriate. The depth-averaged equations may be written in a natural coordinate system as

$$\frac{\partial \eta}{\partial t} + \frac{\partial \overline{U}h}{\partial x} + \frac{\partial \overline{V}h}{\partial y} = 0 \tag{10}$$

Conservation of mass

$$\frac{\partial \overline{U}h}{\partial t} + \beta\left[\frac{\partial}{\partial x}(\overline{UU}h) + \frac{\partial}{\partial y}(\overline{UV}h)\right] + gh\frac{\partial \eta}{\partial x} + \frac{\tau_{\mathrm{bx}}}{\rho} + F_x - \frac{1}{\rho}\left[\frac{\partial}{\partial x}\mathrm{T}_{xx} + \frac{\partial}{\partial y}\mathrm{T}_{xy}\right] = 0 \tag{11}$$

Conservation of momentum in x direction

$$\frac{\partial \overline{V}h}{\partial t} + \beta\left[\frac{\partial}{\partial x}(\overline{VU}h) + \frac{\partial}{\partial y}(\overline{VV}h)\right] + gh\frac{\partial \eta}{\partial y} + \frac{\tau_{\mathrm{by}}}{\rho} + F_y - \frac{1}{\rho}\left[\frac{\partial}{\partial x}\mathrm{T}_{yx} + \frac{\partial}{\partial y}\mathrm{T}_{yy}\right] = 0 \tag{12}$$

Conservation of momentum in y direction

where η is the water elevation, $\overline{U}$ and $\overline{V}$ are depth-averaged velocities in the horizontal directions, β is the momentum correction factor, h is the water depth, τ_{bx} and τ_{by} are the bed shear stresses in the horizontal directions, $F_{x,y}$ represents some external body force, and T_{xx} and T_{xy} and so on are the horizontal stresses.

Appropriate boundary conditions for the solution of the above equations are the specification of flow or water elevation or some combination of the two at the open boundaries, and at closed boundaries some relationship describing the flow at water/land interfaces (zero flow normal to the land and a relationship between flow and stress tangential to the land).

The computational effort required to carry out 2D modeling is significantly higher than for 1D modeling, but resulting simulations are better able to represent flow events; such as, the extent of flood inundation following a dam break and flow separation (recirculation and dead zones). The first models used finite difference techniques (Leendertse, 1967; Vreugdenhil and Wijbenga, 1982) and the method characteristics (Townson, 1974). With the development of the finite element method (Brebbia *et al.*, 1978) and associated numerical techniques (Brooks and Hughes, 1982; Hervouet, 1991), enhanced hydraulics codes were produced, such as RMA-2 (King and Norton, 1978) in the early 1970s and TELEMAC-2D in the late 1980s (Hervouet, 1991). The finite element method proved useful in representing complex geometries; however, it is a demanding method to implement numerically and possess limitations in its ability to conserve mass. Recently, the finite volume method has emerged as the preferred numerical technique for 2D flood modeling because of its conservative properties, numerical accuracy, and simplicity in application (Sleigh *et al.*, 1998).

Recent research has focused on the treatment of wetting-and-drying and adaptive meshing to capture flood spreading and inundation (Lynch and Gray, 1980; Akanbi and Katopodes, 1988; Tchamen and Kahawita, 1998). In particular, the work by Bates *et al.* (1993, 1996) has illustrated the capability of the finite element code TELEMAC-2D to reproduce the transit of a flood wave and the corresponding flood map dynamically. Such codes can now support river models of up to 60 km (Bates *et al.*, 1996).

As mentioned earlier, the need to improve predictions of flood inundation extent to support flood risk management decisions has resulted in research into the development of hybrid 1D/2D models, Dhondia and Stelling (2002). These models utilize 1D modeling techniques to simulate flows through the river network with dynamic linking to 2D hydrodynamic models for the simulation of floodplain flows. Dhondia and Stelling (2002) provide a case study of such a technique applied to the Sistan–Baluchistan River Basin in the Islamic Republic of Iran and Wicks *et al.* (2004) discuss the performance of various modeling methods applied to urban inundation in London, UK.

THREE-DIMENSIONAL (3D) MODELS

The complex topography of natural channels and their flood plains results in a velocity field that is highly three-dimensional. This has implications for the mechanics of sediment entrainment and deposition as well as pollutant mixing and dispersion. Models that represent three-dimensional flow are referred to as 3D models. TELEMAC-3D, FLOW 3D, and CFX computational software are typical of this group.

Some three-dimensional flow mechanisms for simple meandering compound channels are shown in Figure 2. The primary flow direction within the main channel (below the bank full level) tends to follow the channel sidewalls (the stream-wise direction), whereas on the floodplain, above bank full level, the flow tends to follow the general valley direction. As a result, a horizontal shear layer develops at around the bank full level in the main channel, especially at the cross-over region. A bulk exchange of fluid between the floodplain and the main channel takes place; this is particularly pronounced in the cross-over region. The fluid near the outside of the main channel emerges onto the adjacent floodplain shortly after each bend apex. On the other hand, the flow from the upstream floodplain is partially entrained by secondary flows in the main channel as it passes over the cross-over region and, consequently, a local plunging zone is found in this locality.

Morvan *et al.* (2000) have used CFX 4 to model the FCF–B23 experiment as discussed in the section on quasi-2D models. The total flow depth is 200 mm, giving a 50-mm depth on the floodplain. The discharge is set at $0.25\,m^3\,s^{-1}$. The roughness height of the mortar surfaces has been calculated from Manning's "*n*" values to be between 0.2 mm and 0.8 mm. A first CFX model was built using a mesh with 49 200 cells to model the downstream half of the FCF flume. During preliminary calculations, it became apparent that the discretization close to the walls was insufficient, and affected the solution. To achieve a better level of mesh independence a second mesh was constructed using 218 120 elements for half the flume length.

At the inlet, a velocity profile was set so that the overall mass flow is equal to $250\,kg\,s^{-1}$. The turbulent kinetic energy was set to 9.0×10^{-4} ($m^2\ s^{-2}$) and the dissipation length scale to 0.1 m. The outlet mass flow matches that of the inlet. The wall boundary is modeled using a simplified rough wall boundary equation as:

$$\frac{u_\tau}{u_*} = \frac{1}{\kappa}\ln(E' y^+) \tag{13}$$

with,

$$E' = \frac{E}{(1 + 0.3k_s^+)} \tag{14}$$

where κ is the von Karman constant, y^+ is the nondimensional distance from the wall, E is the law of the wall constant, and k_s^+ is the nondimensional roughness height. This is in agreement with other correction terms given in the literature for open-channel flows at high Reynolds number, but less satisfactory at the lower end of the turbulent transition region. The free surface is modeled as a rigid lid. This assumption is acceptable at low Froude numbers, provided there are no large oscillations of the free surface. The pressure field is monitored to ensure the correctness of this assumption.

The results for the velocity at the bend apex and cross-over sections obtained using a k-ε turbulence model are shown on Figure 7 (looking downstream). They should be compared with the actual laboratory data in Figure 7. At the bend apex cross section, flow separation from the inner bank of the main channel is reasonably well represented, although the floodplain flow on the left bank seems to have too much impact on the main channel flow in the model. The cross-over region is adequately modeled, and the rotational impact of the floodplain flow is clearly visible. Along the left bank, a very strong vertical velocity field can be seen. The vertical distribution of secondary velocities of the principal secondary cell in the main channel at the cross-over section is shown in Figure 8(a), and can be compared with the measured FCF data in Figure 8(b). The intensity of the rotation at this location is reproduced well.

A similar exercise has been conducted for velocity data at the River Severn, near Shrewsbury. A plan view of the river reach is shown in Figure 5. Figure 9 and 10 show the comparison between the predicted and measured velocity profiles with depth at cross-sections 4 and 5 in the vicinity of the second bend in the reach. These data were recorded during a flood in October 2000 with flood discharge of the order of $100\,m^3\ s^{-1}$. Comparisons exhibit considerable discrepancy between measured and computed, calling into question both measurement accuracy and accuracy of frictional representation and discretization in the model. Clearly, this science is still in its infancy.

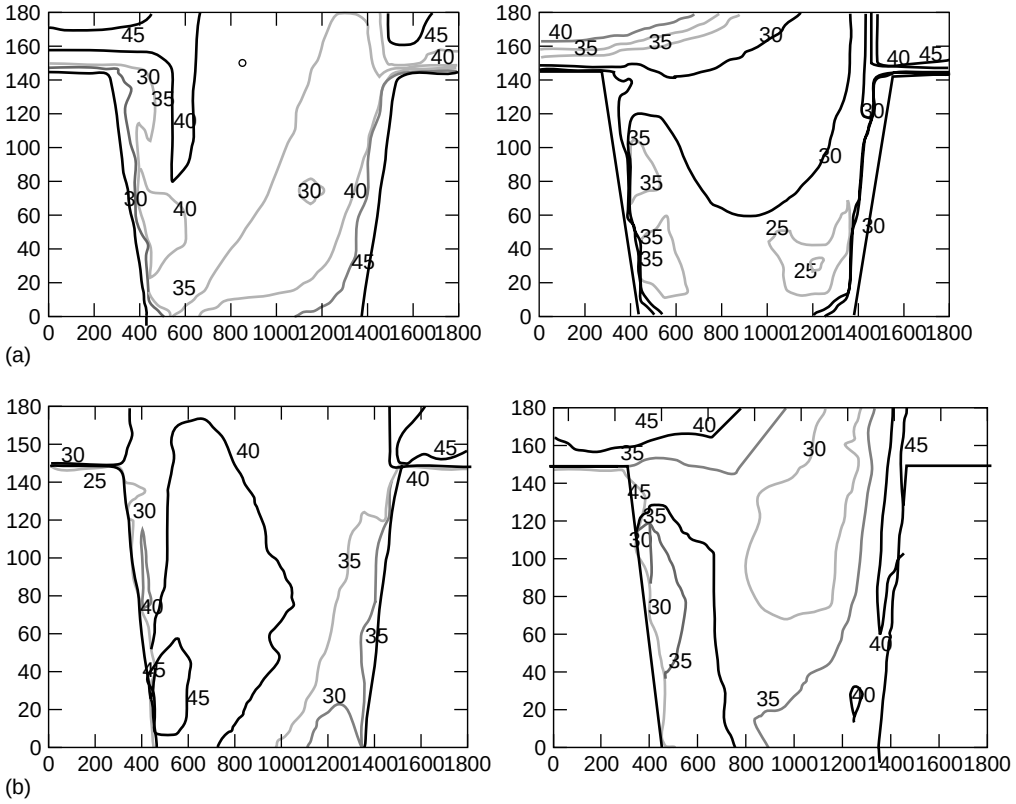


Figure 7 (a) CFX k-ϵ model ($K_s = 0.2$ mm); velocity(cm/s) at cross-sections 5 and 8, (b) FCF B23; experimental velocity data (cm/s) at cross-sections 5 and 8

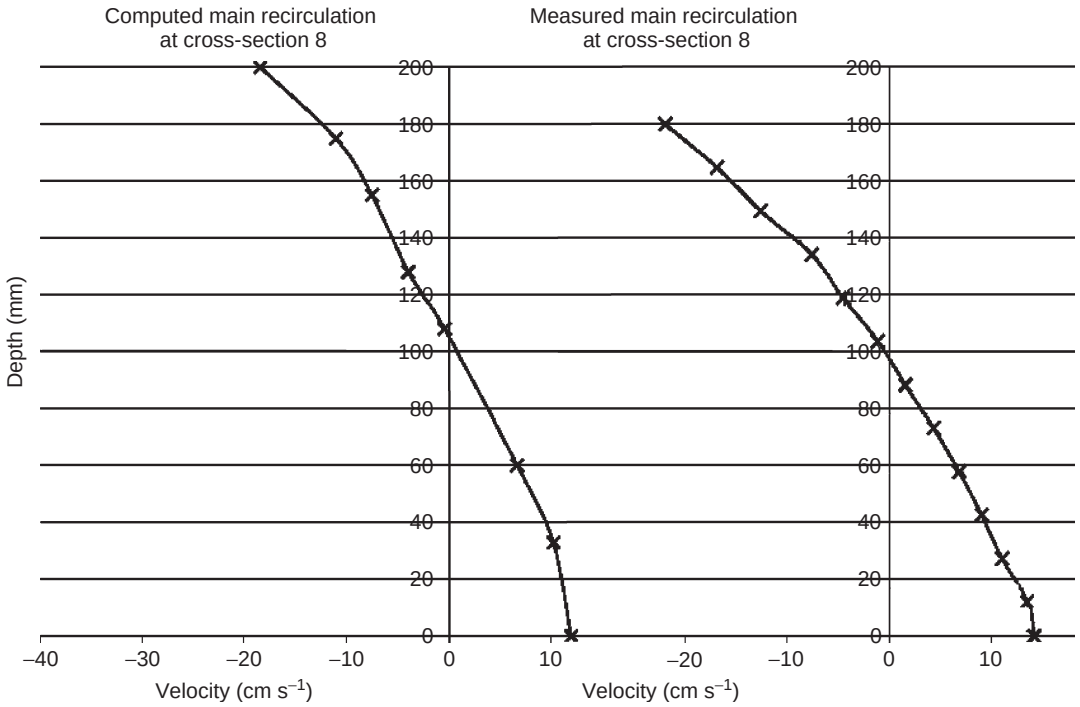


Figure 8 FCF B23-computed and measured model recirculation at cross-over region

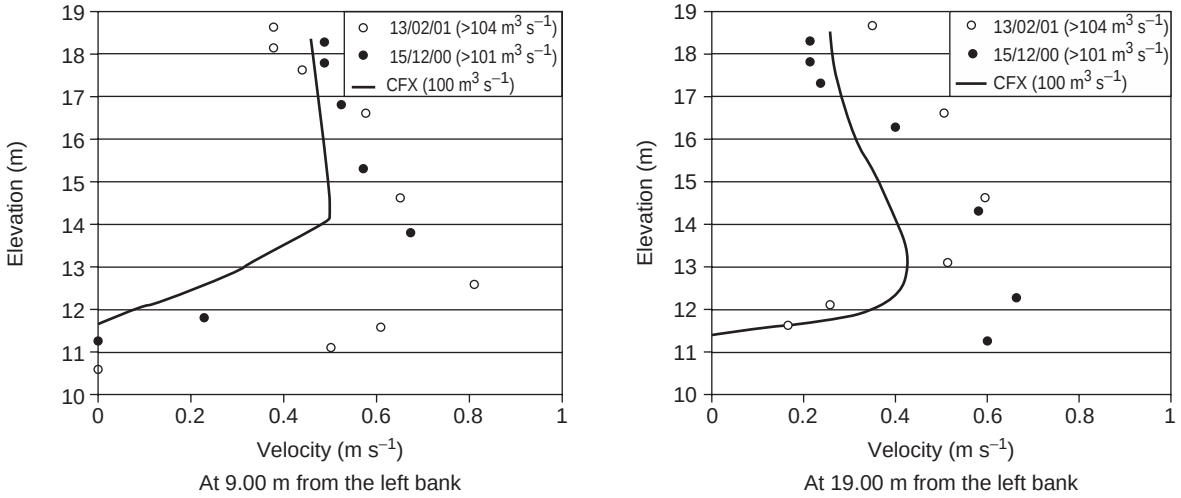


Figure 9 Comparison between Field Data and River Severn CFX Model Predictions at cross section 4

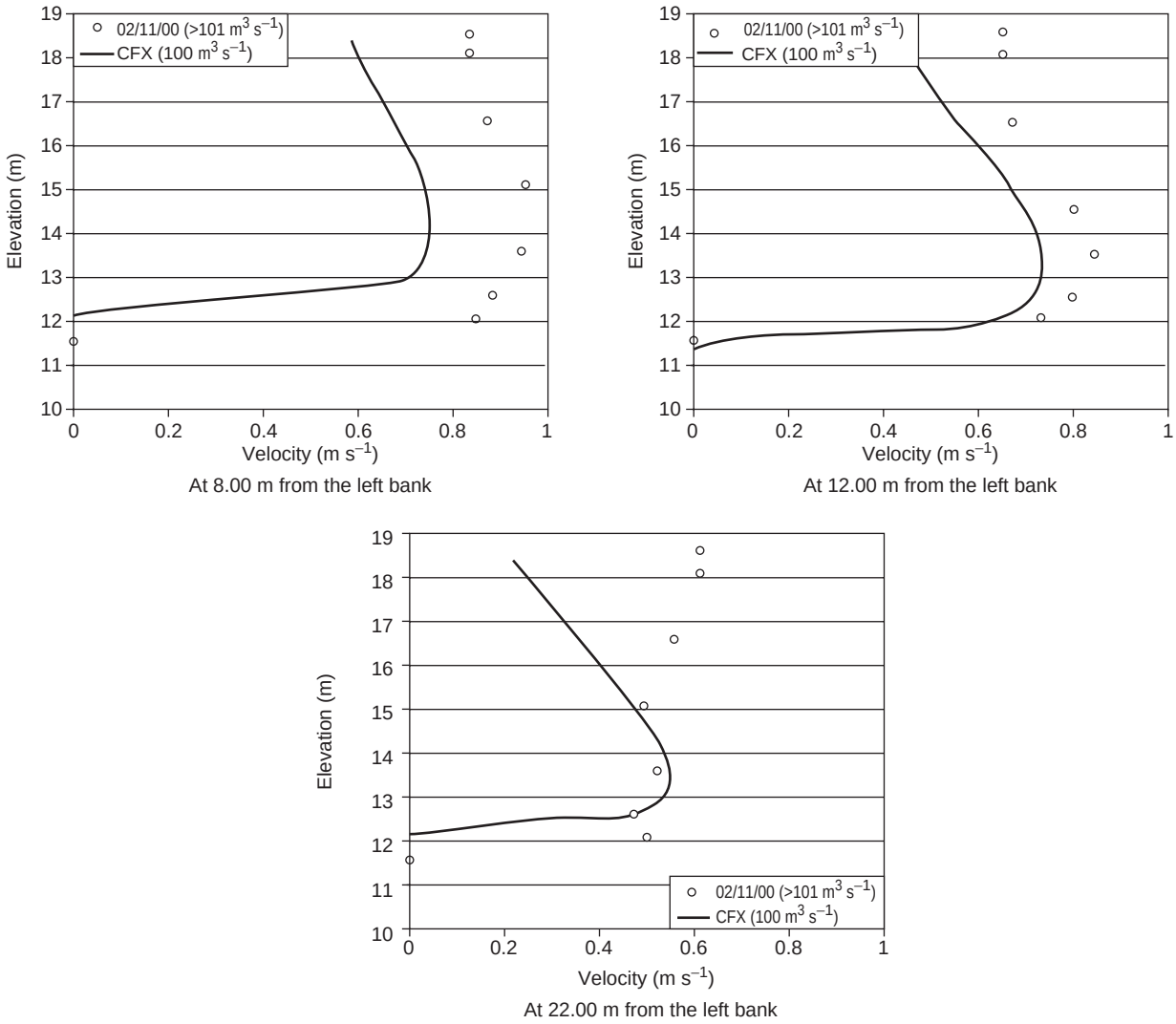


Figure 10 Comparison between Field Data and River Severn CFX Model Predictions at cross section 5

CONCLUSIONS

It is apparent that river flooding is on the increase, and also that there needs to be continued research and development of computational models for predicting flood flows. The choice of model is dictated by the information required to support decision making. 1D models are appropriate for decision support at reach and catchment scales, 2D models are useful where inundation extent and plan velocity fields are important issues, and 3D models are finding greater application in river engineering where detailed 3D velocity fields are required.

Commonly used 1D models are usually adequate to simulate flood levels and, to some degree, lateral extent of flooding. They can provide more useful information when combined with GIS, and future trends indicate an integration of 1D models with 2D flood-spreading algorithms that utilize remotely sensed data. Until now, 1D models have not included floodplain/main channel interaction losses associated with out-of-bank flows. The effect of these interactions has recently been incorporated into ISIS using modified conveyance at each cross section.

The analytical solution to the quasi-2D depth-integrated turbulent form of the Navier–Stokes equation that takes into account the secondary flow term in the formulation, has been shown to predict the distribution of depth-averaged velocity with some accuracy. The value of C_{uv} required to represent secondary cells varies from 0.2% to 0.5% for straight channels and 2% to 5% for meandering channels.

The use of three-dimensional models in natural channels with flood plains is in its infancy. A number of problems arise, including the mode and degree of discretization, handling of roughness and flow resistance, handling of the free surface, and other effects. The FCF flume and the River Severn have been simulated three-dimensionally using the CFX code. The results demonstrate the ability of the CFX model to reproduce complex flow hydrodynamics adequately by using a simple turbulence model, particularly in the FCF. The River Severn comparison is more problematic, emphasizing the need for careful, detailed field data. The results indicate that the choice of mesh discretization has a significant impact on the ability of the numerical scheme to reproduce secondary velocity fields correctly; however, increasing computer power and improved numerical techniques, such as unstructured grids and coupled solvers, suggest that this could be a short-lived concern.

FURTHER READING

Bates P.D., Anderson M.G. and Hervouet J.M. (1995) Initial comparison of two two-dimensional finite element codes for river flood simulation. *Proceedings of the Institution of Civil Engineers-Water and Engineering*, **112**, 238–248.

Berz G. (2000) Flood disasters: lessons from the past-worries for the future. *Proceedings of the Institution of Civil Engineers-Water and Engineering*, **142**, 3–8.

Ramsbottom D., Harpin R. and Spencer P. (1999) The appropriate use of 1D computational river models. *34th MAFF Conference of River and Coastal Engineers*, Keele University.

Morvan H., Pender G., Wright N.G. and Ervine D.A. (2001) Three-dimensional hydrodymanics of meandering compound channels. *Journal of Hydraulic Engineering, ASCE*, **128**(7), 674–682.

Sellin R.H.J., Ervine D.A. and Willetts B.B. (1993) The behaviour of two-stage channels. *Proceedings of the Institution of Civil Engineers-Water Maritime and Energy*, **101**(2), 99–112.

Shiono K. and Knight D.W. (1989) Two dimensional analytical solution for a compound channel. *Proceedings 3rd International Symposium on Refined Flow Modelling and Turbulence Measurements*, Tokyo, Japan, pp. 591–599.

Shiono K. and Muto Y. (1998) Complex flow mechanisms in compound meandering channels with overbank flow. *Journal of Fluid Mechanics*, **376**, 221–261.

REFERENCES

Akanbi A.A. and Katopodes N.D. (1988) Model for flood propagation on initially dry land. *Journal of Hydraulic Engineering-ASCE*, **114**(7), 689–707.

Bates P.D. and Anderson M.G. (1993) A two dimensional finite element model for river flow inundation. *Proceedings of the Royal Society of London Series A*, **440**, 481–491.

Bates P.D., Anderson M.G., Price D.A., Hardy R.J. and Smith C.N. (1996) *Analysis and Development of Hydraulic Models for Floodplain Flows in Floodplain Processes*, Anderson M.G., Walling D.E. and Bates P.D. (Eds.), Wiley.

Bates P.D. and De Roo A.P.J. (2000) A simple raster-based model for floodplain inundation. *Journal of Hydrology*, **236**, 54–77.

Brebbia C.A., Gray W.G. and Pinder G.F. (Eds.) (1978) Finite elements in water resources. *Proceedings of Second International Conference*, WIT Press: London.

Brooks A.N. and Hughes T.J.R. (1982) Streamline upwind/Petrov Galerkin formulations for convection dominated flows with particular emphasis on the incomprehensible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, **32**, 199–259.

Cunge J.A., Holly F.M. and Verwey A. (1989) *Practical Aspects of Computational River Hydraulics*, Pitman Advanced Publishing Program.

Dhondia J.F. and Stelling G.S. (2002) Application of one-dimensional two-dimensional integrated hydraulic model for flood simulation and damage assessment, hydroinformatics, *Proceedings of 5th International Conference on Hydroinformatics*, Vol. 1, Cardiff, 1–5th July 2002.

Ervine D.A., Babaeyan-Koopaei K. and Sellin R.H.J. (2000) A practical two-dimensional analytical solution for straight and meandering overbank flows. *ASCE, Journal of Hydraulic Engineering*, **126**(9), 653–669.

Ervine D.A. and MacLeod A.B. (1999) Modelling a river channel with distant floodbanks. *Proceedings of the Institution of Civil Engineers-Water Maritime and Energy*, **136**, 21–33.

Forbes G. and Pender G. (2000) The application of enhanced conveyance calculations in flood prediction. *Proceedings of International Symposium of Flood Defence*, Kassel.

Hervouet J.-M. (1991) Vectorisation et simplification des algorithmes en elements finis. EDF, Bulletin de la Direction des Etudes et Recherches Serie C, Mathematiques, Informatiques, No. 1, 1991, pp. 1–37, (in French).

HR Wallingford (2003) *Interim Report 2: Review of Methods for Estimating Conveyance*, Project W5A-057, March.

King I.P. and Norton W.R. (1978) Recent application s of RMA's finite element models for two-dimensional hydrodynamics and water quality. In *Finite Elements in Water Resources*, Brebbia C.A., Gray W.G. and Pinder G.F. (Eds.), WIT Press.

Leendertse J.J. (1967) *Aspects of Computational Model for Long Period Water Wave Propagation*, RAND Corporation Report, Report RM 5294-PR, Santa Monica.

Lynch D.R. and Gray W.G. (1980) Finite element simulation of flow in deforming regions. *Journal of Computational Physics*, **36**, 135–153.

Morvan H., Wright N.G., Pender G., Ervine D.A. (2000) Three-dimensional modeling of secondary currents in meandering compound channels. CFX Conference, Reading.

Pender G. (1992) Maintaining numerical stability of flood plain calculations by time increment splitting. *Proceedings of the Institution of Civil Engineers-Water Maritime and Energy*, **96** 35–42.

Pender G. and Ellis J. (1990) Numerical simulation of overbank flooding in rivers. *Proceedings of International Conference on River Flood Hydraulics*, White W.R. (Ed.), Wallingford, England, pp. 403–411, 17–20 September.

Sen D. (2002) An algorithm for coupling 1D river flow and quasi 2D flood inundation flow, hydroinformatics. *Proceedings of 5*th *International Conference on Hydroinformatics*, Vol. 1, Cardiff, 1–5th July 2002.

Shiono K. and Knight D.W. (1991) Turbulent open channel flows with variable depth across the channel. *Journal of Fluid Mechanics*, **222**, 617–646.

Sleigh P.A., Gaskell P.H., Berzins M. and Wright N.G. (1998) An unstructured finite-volume algorithm for predicting flow in rivers and estuaries. *Computer and Fluids*, **27**(4), 479–508.

Tchamen G.W. and Kahawita R.A. (1998) Modelling wetting and drying effects over complex topography. *Hydrological Processes*, **12**, 1151–1182.

Townson J.M. (1974) An application of the method of characteristics to tides in x-y-t space. *Journal of Hydraulic Research*, **12**(4), 499–523.

Vreugdenhil C.B. and Wijbenga J.H.A. (1982) Computation of flow patterns in rivers. *Journal of the Hydraulics Division-ASCE*, **108**(HY11), 1296–1310.

Wark J.B., Samuels P.G. and Ervine D.A. (1990) A practical method of estimating velocity and discharge in a compound channel. In *River Flood Hydraulics*, White W.R. (Ed.), Wiley: pp. 163–172.

Wicks J., Mocke R., Bates P.D., Ramsbottom D., Evans E. and Green C. (2003) Selection of appropriate models for flood modelling. *Proceedings of 38*th *DEFRA Annual Flood and Coastal Management Conference*, Department of Environment, Food and Rural Affairs: London.

Wicks J., Syme B., Hassan M.A.A.M., Lin B. and Tarrant O. (2004) 2D modelling of floodplains – is it worth it? *Proceedings of 39*th *DEFRA Annual Flood and Coastal Management Conference*, Department of Environment, Food and Rural Affairs: London.

Yen B.C. (1973) Open channel flow equations revisited. *Journal of the Engineering Mechanics Division-ASCE*, **99**(EM5), 84–95.

142: Debris Flow

ARONNE ARMANINI, LUIGI FRACCAROLLO AND MICHELE LARCHER

CUDAM and Department of Civil and Environmental Engineering, University of Trento, Trento, Italy

The principal features of debris flow are described. One section is devoted to debris-flow triggering, accounting for the geomechanical criteria by Takahashi (1978) and for most recent hydrological distributed models. In a second section, the rheology of debris flows is described: the theory of dispersive stresses by Bagnold, the kinetic theories of dense gases, and some recent experimental observations about debris flows in equilibrium. A section is devoted to mathematical models and the final section discusses debris flows countermeasures.

INTRODUCTION

Debris flow is a motion of widely sorted debris (from a few millimeters to some meters) inside a watery matrix or also in the presence of clayish mud. One of the most evident facts is the floatation of huge boulders on the surface of the debris flow. Taking up the original definition proposed by Takahashi (1981), debris flows are *massive sediment transport* phenomena that manifest themselves in mountain streams characterized by a steep slope, where the motion of the granular phase is induced directly by gravity. Here, the ratio between the liquid and the solid transport rates is relatively low, and can be zero in case of dry granular mixtures (e.g. dry landslides). On the contrary, in the case of ordinary sediment transport (bed load and suspended load), sediments are not driven directly by gravity, but by the hydrodynamic actions induced by the fluid, and liquid–solid transport ratio is relatively high. There is a wide range of two-phase sediment-laden motions for which this distinction is not so sharp. Other situations outside simple schemes concern cases where, although the solid transport rate is in the debris-flow range, the finest fractions of the grain distribution change the rheology of the interstitial fluid. In fact, the role of fine sediments, because of their size, and in some cases, of their electrochemical properties, is completely different from the coarser ones: they can mix with water, forming a homogeneous interstitial fluid (slurry), typically characterized by high viscosity and in some cases by cohesive behavior. Its composition is strongly connected to the flow dynamics, making it very difficult to define and distinguish the role of the two phases. Further, complications take place when a third phase, constituted by air and other gases, is present inside and affects the flowing mixture, such as in the stout front of advancing debris-flow waves.

The picture presented above is mirrored by manifold classifications found in literature, where massive phenomena involving the liquid and solid phase are reported with different terminology: for example, *debris flows, mud flows, hyperconcentrated flows*, and so on, any of them possibly being preceded by adjectives such as *turbulent, laminar*, and so on. Recently, the knowledge about granular flows has progressed a lot, and nowadays (2003) there is a wide literature regarding both the rheological features of these kinds of flows and the applicative aspects concerning design criteria for defence and protection works.

As a rule, there are at least two fundamental elements that must be present for the triggering of a debris flow: an opportune succession of rain events and the availability of solid material. As better explained afterwards, the weather event has to present absolutely particular characteristics: generally it must be very intense, but in order to be able to move the bulk it must also be preceded by an event long enough to take the sediments to saturation. The concomitance of these happenings gives little predictability to debris-flow events that are relatively rare, but present a remarkable destructive power; they concern above all small basins and in particular alluvial fans that have often undergone a recent urbanization. Basically, they are nonstationary phenomena developing in particularly

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

short times. Because of their unexpected character, their destructive power is often undervalued: small creeks that are affected by very modest solid or liquid events for decades, are sometimes affected by debris flows of huge intensity. The recurrence of such events, in fact, is very difficult to determine; systematic observations show that the debris flows have return times of the order of 50–100 years, a period of time during which the phenomenon is likely not to manifest at all. Unlike floods in water streams, in which some extreme events happen almost regularly every year, even though with a different intensity, for debris flows it is exactly the sudden and often unexpected triggering that makes the phenomenon insidious.

DEBRIS-FLOW TRIGGERING

The assessment of debris-flow risk is very difficult as three main factors participate in the triggering of debris flows: the rainfall intensity, the initial state of the basin soil moisture, and the presence of enough sediment. The analysis of many different events has shown that generally the debris flows manifest themselves after an extreme rainfall event following a rainfall of long duration. In a certain area, in fact, debris flows usually occur when the antecedent rainfall intensity exceeds a certain critical value. Yet, it is necessary that in a suitable previous period (of the order of 7 to 10 days) a congruous volume of rain has fallen, making at least part of the valley saturated.

A more physically based approach consists in using a steady-state shallow subsurface flow model (e.g. *TOPMODEL* by Beven and Kirkby, 1979) to evaluate the saturated areas and their expansion during the storm. However, unlike landslides generated in hillslopes by subsurface flow, debris-flow initiation is mainly due to surface runoff and a complete rainfall-runoff model is required. If design of defence works is implied, a special kind of rainfall representation could be useful in the form of intensity-duration-frequency (IDF) curves that relate the rainfall intensity and duration to a prescribed return time. If just the peak discharge is required, some simplified models are available and are sufficient for this goal (Rigon *et al.*, 2005). According to the theory presented below, once the runoff creates a flow depth greater than a suitable threshold, the slope collapses and the debris-flow run-out starts. This critical value depends obviously on the geometric, geologic, and morphologic characteristics of the debris deposit and of the sublayer on which it lies.

This approach alone is, however, not sufficient to get the total of sediment mass moved. Besides the weather events, there must also be a deep-enough debris storage so that the frequency of debris flows does not coincide with that of intense rainfall events. Hampel (1968) distinguishes between watercourses with a rocky channel bed and those which flow on erodible alluvial deposits. The former require a gradual debris accumulation between an event and the following ones and are temporarily stable after the event; the latter can be destabilized by an event and this can bring about a period of activity with rather high frequencies, followed by the return to a state of rest. The use of stochastic models to assess the contemporary presence of all the conditions required is presently (2003) being studied, but it is still not sufficiently tested to be utilized in field applications (Iida, 1999).

A further possibility of debris-flow generation is obviously due to the dynamic action of landslide in their run-out from hillslopes into channels or the collapse of retention structures existing along the mainstream. The total quantity of debris that rests after the event is usually due both to the size of the source of material (especially if it derives from a landslide) and to the dynamics of the debris flow, which erodes the bed on which it is moving.

Takahashi (1978) first introduced the stability theory regarding loose materials for the study of debris-flow triggering. This theory refers to a noncohesive and uniform granular body and is based on the balance of forces acting in different imbibition conditions. The Takahashi instability condition is written as follows:

$$\tan\alpha \leq \tan\varphi \frac{C_0\Delta}{1 + C_0\Delta + \dfrac{1}{n}\dfrac{h_0}{D}} \tag{1}$$

where α is the inclination angle of the deposit, φ is the friction angle of the material, C_0 is the volume concentration of the particles in a state of rest, h_0 is the flow depth of the water flowing over the deposit (in uniform-flow condition), $\Delta = (\rho_s - \rho)/\rho$ is the relative density of the immersed material, D the mean diameter of the material, and n is a nondimensional coefficient of the order of the unit depending on the shape and distribution of the grains. Note that if $h_0 = 0$, corresponding to a *completely saturated* body with no surface runoff, the stability condition becomes:

$$\tan\alpha \leq \tan\varphi \frac{C_0\Delta}{1 + C_0\Delta} \tag{2}$$

Once in movement, the body is soon transformed into a debris flow. In this situation, the dynamic friction can also move the material lying underneath the layer initially made unstable; this phenomenon is known as *mass entrainment*. Moreover, according to Takahashi (1978) the values of the ratio (h_0/nD) able to generate a real debris-flow range between 0 and 1.33. For values below 0, there is a partially dry deposit which, when it becomes unstable owing to big enough slopes, gives rise to a landslide. For values over 1.33, there is a movement more similar to bed load rather than to debris flow. In short, according to Takahashi, the

slopes for which there is debris-flow range between the two following extremes:

$$\tan\varphi\frac{C_0\Delta}{1+C_0\Delta+1.33}\le\tan\alpha\le\tan\varphi\frac{C_0\Delta}{1+C_0\Delta} \qquad (3)$$

If, for example, for the above-cited parameters we assume the following values that are characteristic of stony, noncohesive material, $C_0 = 0.7$, $\Delta = 1.65$, $\tan\varphi = 0.8$, we obtain:

$$15°9' \le \alpha \le 23°5' \qquad (4)$$

Once the movement is triggered, the debris flows can flow even with slopes smaller than the limit on the right of relation (4). Anyway, with slopes lesser than 3° the debris flows do stop. Consequently, the basins that are more frequently affected by debris flows are those in which the presence of hillslopes or torrents with slopes comprised between the limits listed above are statistically more important.

The analysis of the historical events in the different regions suggests that the debris-flow phenomena regard essentially small catchments, generally from 2 to 10 km^2. Other elements that are important for the debris-flow formation are the geology and the vegetative cover. It is evident that the presence of vegetation inhibits strongly the debris-flow formation; therefore, debris flows are more likely to form above the vegetation limit. After all, it is at higher elevations that there are hillslopes with a slope ranging between the limits cited.

Other mechanisms generating debris flows are given by landslides depositing in the channel bed, or by the collapse of a natural dam which was formed temporarily in the channel bed because of the stop of vegetation transported by the current, or else by the collapse of some retention structures (in particular check dams) built along the mainstream of the river affected by an initial debris flow.

DEBRIS-FLOW RHEOLOGY AND DYNAMICS

Once started, debris flows tend to assume a typical layout, formed by a round-shaped *front* in which the biggest boulders tend to accumulate, followed by a *body* in which the free surface is nearly parallel to the undisturbed debris bed. In this part, the motion can be assumed to be close to the uniform motion. The central part with uniform motion is followed by a *tail*, in which the flow depth becomes thinner and where the zone of erosion is exposed (see Figure 1).

It is possible to study the behavior of the motion in the different parts of the debris flow separately. It is, above all, the part with quasi-uniform motion that becomes significant, since the hypothesis of uniform flow makes it possible to give a detailed determination of kinematic and dynamic

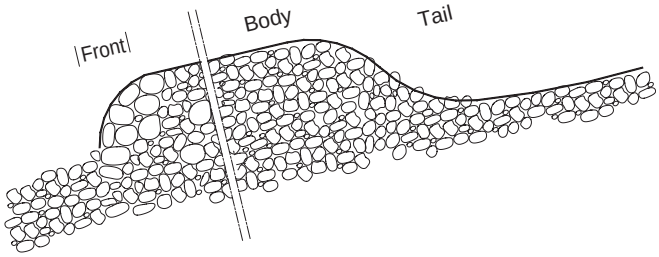


Figure 1 Scheme of the longitudinal section of a debris flow: in the central part (body) the motion is quasi-uniform

conditions. The study of debris-flow dynamics requires the knowledge of the interaction between particles and interstitial fluid, between the contour, the particles, and the interstitial fluid, and among the particles themselves. Therefore, the quantitative description of debris-flow dynamics is very complex. It is often convenient to consider the debris flow as a continuous medium, to which is assigned a suitable rheological law able to simulate the different interactions between particles, fluid, and wall. In fact, a granular material subjected to deformation can determine different types of interactions among grains and, therefore, it can generate stresses by different mechanisms. Individual particles may interact with one another in rigid particle clusters, generating a network of contact forces through sustained rolling or sliding contacts, or by nearly instantaneous collisions, during which momentum is exchanged and energy is dissipated because of inelasticity and friction. The relative importance of these mechanisms may be used as the characteristic defining the various flow regimes (Savage, 1984). Basically, the main interactions can be listed as follows:

- deformation of the mean fluid field by the particles;
- collisions among the particles;
- friction among the particles during long contact periods;
- deformation of the turbulence structure generated by the wall.

Depending on particle concentration and hydrodynamic conditions, some of the mechanisms described above can prevail with respect to the others.

Theory of Dispersive Stresses

Theoretical and experimental findings by Bagnold (1954), regarding the rheology of a gravity-free dispersion of large solid spheres in a Newtonian fluid under shear, probably represent the most extensively cited research over the last 50 years among scientists involved in the study of debris flows, snow avalanches, and of the flow of granular materials in general. Bagnold's theory presents some conceptual limits that have been analyzed in detail in the recent literature, among others by Hunt *et al.* (2002). Yet, it has the merit of being simple and of being the first

one that indicated the physics of the phenomenon clearly; moreover, it has been successfully applied to debris flows, above all by Takahashi (1978).

The theory is based on the observation that when particles collide among themselves or on the containment wall, collisions manifest themselves as an increase in pressure called *dispersive pressure*. According to Bagnold (1954), two extreme situations can occur. The first one occurs in the presence of particles of small size being well dispersed within the interstitial fluid: the debris flow assumes then a *macroviscous* behavior and the viscous stresses prevail with respect to the dispersive ones because of the intergranular collisions.

On the contrary, an opposed situation occurs when the particles have high concentrations and the debris-flow speed is high, and then the collisions are much more frequent. In this regime, called *grain-inertia*, the collisions among grains are the determining factors as to the debris-flow resistance and the effect due to the viscosity of the interstitial fluid is negligible.

Bagnold distinguished *grain-inertia* and *macroviscous* flow regimes on the basis of the nondimensional parameter B_a presented in equation (5), subsequently termed *Bagnold number*, representing the ratio between stresses due to inertia and those due to viscosity.

$$B_a = \rho \frac{D^2(du/dz)}{\mu}\lambda^{1/2} \text{ with } \lambda = \frac{1}{(C_0/C)^{1/3} - 1} \qquad (5)$$

where D is the grain diameter, λ is Bagnold *linear concentration*, C is the volume concentration of the solid fraction, C_0 is the maximum concentration possible (*concentration at rest*), du/dz is the shear rate, μ and ρ are the dynamic viscosity and the mass density of the interstitial fluid. Bagnold called *macroviscous* the regime characterized by small Bagnold numbers ($B_a < 40$), where the shear stresses behave like a Newtonian fluid with a viscosity corrected by the presence of the particles, and *grain-inertia* the other regime, is characterized by $B_a > 450$, where the stresses were independent of the fluid viscosity and proportional to the square of the shear rate and to the square of the linear concentration λ. The intermediate range of Bagnold number ($40 < B_a < 450$) occupies a transitional region.

Bagnold derived simple analyses to explain the rheological behavior in the two limiting regimes. In the viscosity-dominated *macroviscous regime*, shear and normal stresses are linear functions of the shear rate du/dz. Bagnold identified the presence of a normal stress in radial direction, termed *dispersive pressure*, and attributed it to a statistically preferred anisotropy in the spatial particle distributions. He proposed the following relations for the shear stress τ and the normal stress σ:

$$\text{in the macroviscous region}: \tau = 2.25\lambda^{3/2}\mu\frac{du}{dz}$$

$$\sigma = \frac{\tau}{\tan\varphi} \qquad (6)$$

$$\text{and in the grain-inertia region}: \sigma = 0.042\rho\left(\lambda\frac{du}{dz}D\right)^2\cos\varphi$$

$$\tau = \sigma\tan\varphi \qquad (7)$$

It should be noted that both in the macroviscous and in the grain-inertia regime, the normal stress is proportional to the shear stress in the form $\tau = \sigma \tan\varphi$, where φ represents a dynamic friction angle, depending on collision conditions. Such behavior is reminiscent of the Coulomb criterion used to describe the stresses in cohesionless soils under conditions of limited equilibrium. According to Brown and Richards (1970), typical values for φ obtained during quasi-static yielding at low stress levels are close to the angle of repose, that is, about 24° for spherical glass beads and 38° for angular sand grains. Bagnold proposed a dynamic friction angle $\varphi \cong 37°$ for the macroviscous regime and $\varphi \cong 18°$ for the grain-inertia regime.

As a completion of Bagnold's theory, nowadays the macroviscous regime is considered to be practically absent in debris flows, while where the shear rate and the velocity of the fluid are small, a regime called *frictional* or *quasi-static* takes place, in which the contact among particles is quasi-permanent and the ratio between normal and shear stress can be considered roughly constant.

Kinetic Theories

Within a granular flow, the velocity of each particle may be decomposed into the sum of a mean velocity and a random component, taking into account the relative motion of the particle compared to the time-averaged value. Ogawa (1978) introduced first the concept of *granular temperature* T_s, where $3T_s$ is the mean square of particle velocity fluctuations as expressed by equation (8), and Savage and Jeffrey (1981) made the first attempt to make more substantial use of the ideas contained in the previous theoretical work that had dealt with dense gases, for example, Chapman and Cowling (1970).

$$T_s = \frac{1}{3}\langle u'^2 + v'^2 + w'^2\rangle \qquad (8)$$

In analogy with thermodynamic temperature, granular temperature plays similar roles in generating pressures and in governing the internal transport rates of mass, momentum, and energy.

Granular temperature can be generated with two distinct mechanisms (Campbell, 1990). The first, the so-called *collisional temperature generation*, is a by-product of inter-particle collisions, in the sense that two colliding particles

will have resultant velocities depending not only on their initial velocities but also on the type of collision they experienced; therefore, they will contain apparently random velocity components. The second mode of temperature generation, the so-called *streaming temperature generation*, is itself a by-product of the random particle velocities. Following its random path, a particle moving parallel to the local velocity gradient will acquire an apparently random velocity that is proportional to the difference in mean velocity between its present location and the point of its last collision. It should be noted that in both mechanisms of granular temperature generation, the magnitude of the generated random velocities is proportional to the local velocity gradient. However, unlike the collisional temperature generation, the streaming mechanism can generate only the component of random velocity lying in the direction perpendicular to the mean velocity gradient, therefore the generated granular temperature will be anisotropic.

Campbell (1990) describes how the physical similarity between rapid granular flows and kinetic-theory view of gases has led to a great deal of work on creating similar models for granular materials on the basis of the idea of deriving a set of continuum equations (typically mass, momentum, and energy conservation) entirely from microscopic models of individual particle interactions. All the models are based on the assumption that particles interact by instantaneous collisions, implying that only binary or two-particle collisions need to be considered. Particles are usually modeled in a simple way, ignoring surface friction or any other particle interactions tangential to the contact-point, and considering a constant coefficient of restitution to represent the energy dissipated by the impact normal to the point of contact between the particles, even though Lun and Savage (1986) and other researchers showed a strong dependence of the coefficient of restitution on the relative impact velocity. Furthermore, molecular chaos is generally assumed, this implying that the random velocities of particles are independently distributed.

Jenkins and Hanes (1998) apply kinetic theories to a sheet flow in which particles are driven by turbulent fluid and supported by their collisional interactions rather than by the velocity fluctuations of the interstitial fluid. Azanza *et al.* (1999) adapted kinetic theories to a channel flow in which the particles are driven by gravity. Armanini *et al.* (2005) accounted for the interstitial fluid interaction by means of an added mass coefficient.

The constitutive relation for the particle pressures is therefore:

$$\sigma = C\rho_s\left(1+\frac{r\rho}{\rho_s}\right)(1+4Cg_0)T_s \text{ where } r=\frac{1+2C}{2(1-C)} \tag{9}$$

The function $g_0(C)$ describes the variation of the particle collision rate with concentration, and was derived by Carnahan and Starling (1969) from considerations about the nearly geometric form of the virial series for nonattracting rigid spheres.

$$g_0(C)=\frac{(2-C)}{2(1-C)^3} \tag{10}$$

Again, the constitutive relation for the particle shear stress is taken to be the one for a dense molecular gas, in the form:

$$\tau = -\frac{8D(1+r\rho/\rho_s)\rho_s C^2 g_0 {T_s}^{1/2}}{5\pi^{1/2}}\left[1+\frac{\pi}{12}\left(1+\frac{5}{8Cg_0}\right)^2\right]\frac{du}{dz} \tag{11}$$

The balance of particle fluctuation energy is equal to that for the energy of the velocity fluctuations of the molecules of a dense gas. For inelastic particles, the gradient of the vertical component Q of the fluctuation energy flux, expressed by equation (13), is required to balance first the net rate of fluctuation energy production per unit of volume of the mixture, and secondly the rate of collisional dissipation γ_d:

$$\frac{dQ}{dz}=\tau\frac{du}{dz}-\gamma_d \tag{12}$$

where the first term (on the left-hand side) represents the energy diffusion, the second one the net rate of production (the rate of working of the particle shear stress through the mean shear rate) and the last one, γ_d, is the rate of collisional dissipation. Particles are driven into collisions by the mean motion, creating fluctuation energy, while the inelasticity of the collisions dissipates fluctuation energy into real thermal energy.

The constitutive relation for the flux of particle fluctuation energy is taken to be the one for a dense molecular gas, in the form given by Chapman and Cowling (1970), while the rate of collisional dissipation per unit of volume may be calculated using the Maxwellian velocity distribution function, in the form given by Jenkins and Savage (1983):

$$Q=-4\left[1+\frac{9\pi}{32}\left(1+\frac{5\pi}{12Cg_0}\right)^2\right]\frac{C^2 g_0}{\pi^{0.5}}\rho_s\left(1+\frac{r\rho}{\rho_s}\right)D{T_s}^{1/2}\frac{dT_s}{dz} \tag{13}$$

$$\gamma_d = 24(1-e)\frac{C^2 g_0}{\pi^{0.5}}\frac{\rho_s(1+r\rho/\rho_s)T_s^{3/2}}{D} \tag{14}$$

Savage (1998) developed a theory for slow, dense flows of cohesionless granular materials for the case of planar deformations, employing the notion of granular temperature. The conservation equations for mass, momentum, and particle fluctuation energy are employed. At

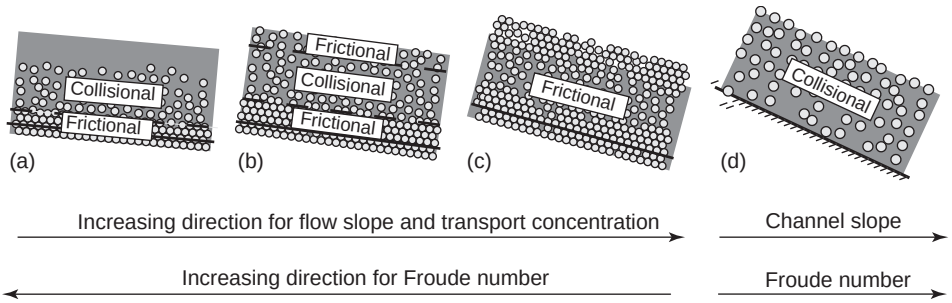


Figure 2 Typology of the flows examined: (a) loose bed, immature (or oversaturated) debris flow; (b) loose bed, mature debris flow; (c) loose bed, plug (or undersaturated) debris flow; (d) rigid bed debris flow

low deformation rates, the apparent form of the constitutive behavior is similar to that of a liquid, in the sense that the actual viscosity *decreases* as the granular temperature augments, contrary to rapid granular flows, in which viscosity *increases* as granular temperature augments.

Stresses are constituted by two parts: a rate-independent, dry friction contribution, and a rate-dependent viscous part, having a quadratic dependence on the shear rate, obtained from the high shear rate granular flow kinetic theories in the form of equations (9) and (11). The magnitude of the rate-independent contribution was chosen so that the sum of the two parts satisfied the overall momentum balance perpendicular to the flow direction.

Savage and Jeffrey (1981) introduced a parameter R, involving the particle diameter D, the square root of the granular temperature and the shear rate, written in the form:

$$R = \frac{D}{T_s^{1/2}} \frac{du}{dz} \tag{15}$$

As discussed in Savage (1998), in both analyses and computer simulations, typically the parameter R is found to be of the order of the unit for granular flows, ranging from purely collisional to slow, predominantly frictional flows.

Experimental Analysis

In contrast with previous experimental studies dealing with solid–liquid flows in annular shear cells, closed ducts, or nonrecirculatory chutes, the experimental analysis carried out in a steady uniform channel flow has shown that the dynamics of a free-surface debris flow constituted of sedimentable material is more articulated than what was predicted by Bagnold for gravity-less granular flows. Contrary to what was presupposed by Takahashi (1978) in adapting Bagnold's theory to debris flows, in the experiments it has been observed that a debris flow can present a series of layers one on top of the other, each governed by a different rheology. The experiments by Armanini *et al.* (2005) permit to distinguish four main regimes of interest (Figure 2): (a) loose bed, immature debris flow; (b) loose bed, mature debris flow; (c) loose bed, plug debris flow; and (d) rigid bed flow.

The first three regimes (a)–(c) exhibit loose-bed equilibrium conditions. The immature debris-flow regime (a) is characterized by the flow of a clear water layer over a fluid-driven sheet of granular material supported by contacts with the stagnant bed. For mature debris flow (b), the entire moving layer is composed of a mixture of liquid and grains. In the "plug" debris flow (c), a partially emerged, quasi-static assembly of grains translates over a liquid-granular shear layer.

The possibility for the flowing mixture to have an underlying bed formed by the same constituents (*movable bed*) leads to the formation of an *equilibrium condition*, since the slope of the bed is a dependent variable dynamically coupled with the flow. For nonequilibrium conditions, the mixture has to flow over a nonerodible bed, and in uniform-flow conditions, the free surface is parallel to the rigid bed of the flume; therefore the presence of the solid phase results somehow in a variable independent of the bed slope. In fact, in this case the solid discharge is smaller than the effective transport capacity of the current. The distinction between *equilibrium* and *nonequilibrium conditions* has of course a strong influence on the rheological processes throughout the cross section of the uniform flow. In fact, considering the profiles of the variables of interest in the stress formation (mean velocity, velocity fluctuations, and solid concentration), one can observe that they change enormously from cases with or without equilibrium. The experimental results made it possible to create a picture of the flow rheology that is rather simple: throughout the flow depth there are layers dominated by collisions among grains, whereas the complementary domains are essentially frictional. When the two rheological mechanisms coexist, the separation between collisional and frictional layers is represented as a rather sharp interface, correspondent to Stokes numbers, defined by equation (16), being equal to $5 \div 10$. Large Stokes numbers characterize collisional regions, while small ones

characterize frictional layers.

$$St = \frac{1}{18}\frac{\rho_s}{\rho}\frac{D^2 \mathrm{d}u/\mathrm{d}z}{\nu} \tag{16}$$

In equation (16), ν is the kinematic viscosity of the interstitial fluid and $D(\mathrm{d}u/\mathrm{d}z)$ represents the relative velocity between adjacent sheared rows of particles.

Moreover, the experimental vertical profiles of energy production, diffusion, and dissipation (Armanini *et al.*, 2003) lead to the conclusion that in the collisional layers energy production balances roughly energy dissipation, so that in equation (12) the diffusive term can be neglected. Exploiting this assumption, normal and shear stress described by equations (9) and (11) depend on the square of the shear rate, exactly like in Bagnold's grain-inertia theory, even if the ratio between shear and normal stresses is perfectly constant only in Bagnold's conjectures.

Global Relations for Debris Flows in Equilibrium

According to the descriptions above, a succession of frictional and collisional layers, one on top of the other, constitutes debris flows. However, considering the debris flow throughout its depth, if the frictional layers, characterized by small mean velocities, give a negligible contribution to the total discharge of the flow, uniform-flow formulas can be obtained referring only to the collisional/grain-inertia layers. For this purpose, it can be assumed (Takahashi, 1978) that the flow is steady and uniform and that the tangential projection of the entire burden of the flow (water + grains) is charged over the solid phase only, therefore assuming that the shear stress in the liquid phase is negligibly small. Moreover, the vertical component of momentum balance for the particle phase requires that the gradient in the particle pressure σ balances the buoyant weight of a unit volume of particles.

$$\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} = C\rho\Delta g\cos\alpha\frac{\mathrm{d}\tau}{\mathrm{d}z} = (C\Delta + 1)\rho g\sin\alpha \tag{17}$$

Under these hypotheses and considering the concentration C to be constant (which corresponds to assume the dynamic friction angle φ to be constant), it is possible to integrate Bagnold's grain-inertia rheological equation (7) combined with equation (17) throughout the flow depth, thus obtaining a relation between the mean velocity and the slope of the current in uniform motion in the same form of Gaukler–Strickler or Chezy formula. In this case, Chezy friction coefficient χ_{DF} is a function of the solid fraction, of the sediment size, of the ratio between fluid and solid density, of the friction angle φ of the material, and of the flow depth h.

$$U = \chi_{DF}\sqrt{h\sin\alpha} \quad \text{with } \chi_{DF} = \frac{2}{5}\frac{h}{\lambda D}\sqrt{g\frac{\rho}{\rho_s}\frac{(C\Delta+1)}{a\sin\varphi}} \tag{18}$$

Bagnold suggests to assume $a = 0.042$, while, on the basis of debris-flow laboratory data, Takahashi (1978) proposed a greater coefficient $a = 0.35$. This value corresponds in case of real debris flows (e.g. for $D = 5\,\mathrm{cm}$, $C = 0.56$, $\varphi = 36°$, $\Delta = 1.65$, $h = 3\,\mathrm{m}$, $C_0 = 0.70$, $\lambda = 13$) to a Chezy coefficient $\chi_{DF} = 11\,\mathrm{m}^{1/2}\,\mathrm{s}^{-1}$. The debris-flow concentration, if assumed to be equal to that of incipient movement in saturated conditions, can be expressed by the following relation as a function of the channel slope and of the relative density of the material:

$$C = \frac{\tan\alpha}{\Delta(\tan\varphi - \tan\alpha)} \tag{19}$$

Takahashi (1991) proposed empirical formulas similar to equation (18) when the regime is clearly not dominated by collisional particle interactions.

In case of viscous debris flow or mudflows, the problem fluid is often treated as a single-phase visco-plastic fluid, characterized by a Herschel–Bulkley rheology (Coussot, 1997). This kind of approach is particularly suitable when the presence of silt and clay is relevant. For low solid concentrations, the flows behave like a Newtonian fluid, while by increasing the concentration the mixture presents a yield stress τ_c and a nonlinear relation between the shear stress and the shear rate, as expressed by equation (20):

$$\tau = \tau_c + k\left(\frac{\mathrm{d}u}{\mathrm{d}z}\right)^n \text{ if } \tau > \tau_c$$
$$\left(\frac{\mathrm{d}u}{\mathrm{d}z}\right)^n = 0 \text{ if } \tau < \tau_c \tag{20}$$

where k and n are parameters depending on the type of fluid and on its concentration, to be determined by means of laboratory tests and by the utilization of a rheometer. Like equation (7), equation (20) expresses a rheological link between shear stress and shear rate, and therefore it can be integrated throughout the flow depth in order to obtain the mean velocity value in uniform-flow conditions.

$$U = \frac{n}{(n+1)h}\left(\frac{\rho_m g\sin\alpha}{k}\right)^{1/n}\left(h - \frac{\tau_c}{\rho_m g\sin\alpha}\right)^{1+(1/n)} \times\left[h - \frac{n}{2n+1}\left(h - \frac{\tau_c}{\rho_m g\sin\alpha}\right)\right] \tag{21}$$

where ρ_m represents the density of the homogeneous mixture. If we assume $n = 1$ and $k = \mu$, relations (20) and (21) are referred to the case of a Bingham fluid.

Debris-flow Peak Discharge

Takahashi (1978) proposed a scheme for the estimation of debris-flow peak discharge based on the hypothesis that, after a short time, the debris-flow front assumes a self-similar profile. In this way, it is possible to make some assessments regarding the ratio between liquid and solid discharge.

The debris-flow front is supposed to move with velocity U_f and also the tail goes downhill with uniform speed U_t, smaller than U_f; in order that debris-flow profile becomes longer without becoming thicker.

With reference to Figure 3, the balances of liquid and solid phase masses give:

$$U_0 h_0 = U_f(1-C)h - U_t(1-C)h - U_t(1-C_0)sa + U_t h_0 \tag{22}$$

$$U_f C h = U_t C h - U_t C_0 a \tag{23}$$

where h is the front depth, a is the depth of the excavation produced by the passing of the tail, s is the degree of saturation of the pores under the debris flow. The liquid feeding from upriver gives rise to a speed U_0 and to a flow depth h_0 and then to a discharge per length unit $q_0 = U_0 h_0$.

From equation (23), we infer that the front speed must be lesser than the tail speed ($U_t < U_f$), and then that the debris flow itself tends to become longer while going downhill.

From equations (22) and (23), which admit that, as the debris flow passes by, the material in the channel bed is completely saturated ($s = 1$), we derive the following relation between the debris-flow discharge and the incoming liquid discharge:

$$U_f h = U_0 h \frac{C_0}{C_0 - C + \dfrac{h_0}{a} C\left(1 - \dfrac{U_0}{U_f}\right)} \tag{24}$$

Equation (24), relative to the propagation of a quasi-stationary debris-flow front, can be utilized to derive the relation between debris-flow discharge and incoming liquid discharge. Supposing that the front speed U_f coincides with the speed of the incoming water, then relation (24) becomes the following:

$$Q_{DF} = Q_0 \frac{C_0}{C_0 - C} \tag{25}$$

where Q_{DF} is the debris-flow peak discharge and Q_0 is the liquid peak discharge. As previously observed, in general, we can assume $C_0 \cong 0.65$ for natural debris.

According to Takahashi (1978), for sufficiently high slopes ($i > 20°$), debris-flow concentration can be assumed to be equal to 90% of the maximum concentration. By doing so, equation (25) becomes $Q_{DF} = 10 Q_0$.

In the case of milder slopes ($i < 20°$), debris-flow concentration is assumed to be equal to that of incipient movement in saturated conditions expressed by equation (19).

In practice, debris-flow concentration ranges between the following limits:

$$0.3 < \frac{C}{C_0} < 0.9 \quad 1.43 Q_0 < Q_{DF} < 10 Q_0 \tag{26}$$

Liquid peak discharge Q_0 can be calculated with common hydrologic methods.

An alternative method used to determine debris-flow peak discharge consists in assimilating the debris-flow front to that which forms after the collapse of a natural dam. The expression obtained is a function of the torrent depth h_m, to be evaluated upstream of the dam, and of the torrent width B.

$$Q_{DF} = \frac{8}{27} B h_m \sqrt{g h_m} \tag{27}$$

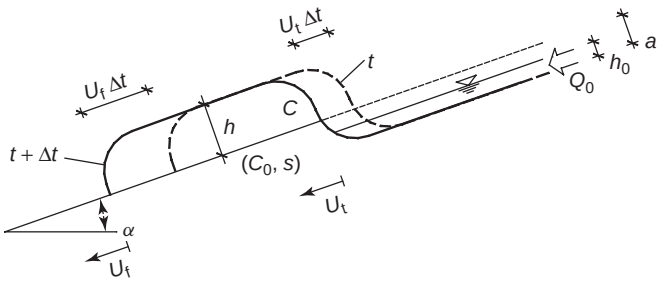


Figure 3 Scheme of quasi-uniform propagation of debris-flow front

MATHEMATICAL MODELING

Mathematical models are tools suitable to describe the general features of initiation, motion, and deposition of a gravity-driven mixture of debris and water and, lastly, the zoning of areas where damage is likely to occur. Owing to the high complexity that characterizes a real event, simplified assumptions in the mathematical modeling have to be introduced. At present in the applications the grain-size composition in the mixture is generally not accounted for, and only the solid concentration, which may change in time and space, remains to characterize the flowing sediments. Therefore, the possibility to represent the variations of the sediment size throughout the debris flow (from its steep front, mainly constituted of boulders, to the tail) is generally very limited.

General three-dimensional models require an extremely high computing time and level of hardware to determine the relevant solutions by means of numerical techniques, and are applicable, at present, to local situations only.

The representations in a reduced two- or one-dimensional frame correspond to depth- or section-integrated models.

This averaging process requires hypotheses about the profiles of the field variables (velocities, pressure, concentration) through the flow depth. These assumptions make it possible to derive models where the flow domain is the basal topography.

The characterization of the fluid mixture takes into account the fact that the fluid is not strictly homogeneous, but made up of different phases. The liquid phase refers to the interstitial slurry that is made up of water and of the finest grain fractions. The solid phase represents the coarser grain fractions. Some models admit that the two phases have different velocity distributions. The coupling between the two phases is present both in the mass conservations and in the momentum equations (due to the solid–liquid stress interaction). When we apply the hypothesis that there is no relative velocity between the two phases, the resulting integrated models present only one momentum equation, referring to a fluid having the bulk mass density.

Since attention is focused on averaged equation models, the rheological properties of the flowing mixture is lumped in the definition of algebraic expressions for the tangential stresses acting on the boundaries of the flow domain, especially on the bottom. Therefore, the boundary stresses are a function of the field variables of the model (such as flow depth, velocity, slope, sediment size) and show up as source terms in the momentum equations.

Mathematical and numerical models are thought of with proper boundary and initial conditions. Debris flows run-out and depositions over fans clearly depend on the incoming hydrograph. Initial conditions may also play a significant role, as in cases where the flow is triggered by an abrupt collapse of barrages (dam break–induced debris flows; Fraccarollo and Capart, 2002).

Motion equations are based on conservation laws. Their number, along with the number of variables, depends on the refinement of the model. The options mainly concern the number of phases and the number of momentum equations, one for each direction and phase-velocity field. In case of a nonhomogeneous fluid, the bed-level changes in time and the morphological evolutions are part of the solution. For sake of simplicity, one-dimensional models are referred to a flow section of unit width and rectangular shape, assuming no frictional influences from the sidewalls.

Modeling of a one-phase system in the one-dimensional framework is herein reported:

$$\begin{cases} \frac{\partial}{\partial t}(h) + \frac{\partial}{\partial x}(Uh) = 0 \\ \frac{\partial}{\partial t}(Uh) + \frac{\partial}{\partial x}\left(\beta U^2 h + g\frac{h^2}{2}\right) + gh\frac{\partial z_b}{\partial x} = -\frac{\tau}{\rho_m} \end{cases} \tag{28}$$

where h is the flow depth, U is the flow-depth-averaged velocity, $\rho_m = \rho(C\Delta + 1)$ is the density of the homogeneous mixture, z_b is bed elevation, τ/ρ_m represents the only source term, providing the shear stress at the bottom, β is the momentum-flux coefficient that takes into account the nonuniform velocity distribution through the flow depth.

In this context, the main point characterizing the fluid is the formulation of the wall friction force, to be deduced from rheological assumptions. Many rheological models have been tested in the past by different authors. Most of them can be classified in the following categories: (i) models for muddy fluid (e.g. Bingham models); (ii) models for mixtures in the inertial regime (e.g. equation 11); (iii) a combination of different dissipation models, including strain-rate independent contributions, to simulate grain-to-grain static prolonged contact, and various strain-rate dependent ones, such as Newtonian, turbulent, or collisional behaviors.

Under the same hypotheses, and moreover assuming $\beta = 1$, the two-dimensional extent of the mass and momentum balances are:

$$\begin{cases} \frac{\partial}{\partial t}(h) + \frac{\partial}{\partial x}(hU) + \frac{\partial}{\partial y}(hV) = 0 \\ \frac{\partial}{\partial t}(hU) + \frac{\partial}{\partial x}\left(\frac{1}{2}gh^2 + hU^2\right) + \frac{\partial}{\partial y}(hUV) + gh\frac{\partial z_b}{\partial x} = -\frac{\tau_x}{\rho_m} \\ \frac{\partial}{\partial t}(hV) + \frac{\partial}{\partial x}(hUV) + \frac{\partial}{\partial y}\left(\frac{1}{2}gh^2 + hV^2\right) + gh\frac{\partial z_b}{\partial y} = -\frac{\tau_y}{\rho_m} \end{cases} \tag{29}$$

where the inner shear stresses have been also neglected, as in most models of this kind, and where τ_x and τ_y are the components of the bottom shear stress and U and V are the components of the velocity.

One of the answers expected from model applications for mapping the risk concerns the description of debris deposition phenomena at the alluvial fans. In case of a homogeneous fluid, the stopping of the flow can be induced by a visco-plastic constitutive law (i.e. Bingham, Herschel–Bulkley models): when the yield stress balances or exceeds the local acting forces, the fluid comes to a local stop throughout the flow depth.

On the contrary, in situations where the debris deposition comes together with a reduction of the water content because of a velocity reduction of the flow over the alluvial fan, it is necessary to exploit models dealing with a nonhomogeneous two-phase fluid. The presence of solid and liquid phases is accounted for by considering the volumetric solid concentration (depth averaged) in both mass and momentum conservations. With a full two-phase model, two mass and two momentum equations should be considered. However, assuming that the velocities of the two phases are correlated, it comes out that a single momentum equation (and a single velocity field), associated to the bulk fluid-density, remains.

Here, it is chosen to represent the formally simpler case, assuming that there is no velocity-phase difference. The

following model is obtained in the one-dimensional case:

$$\begin{cases} \frac{\partial}{\partial t}(h + z_b) + \frac{\partial}{\partial x}(Uh) = 0 \\ \frac{\partial}{\partial t}(Ch + C_b z_b) + \frac{\partial}{\partial x}(CUh) = 0 \\ \frac{\partial}{\partial t}[(C\Delta + 1)Uh] + \frac{\partial}{\partial x}\left[(C\Delta + 1)\left(\beta U^2 h + kg\frac{h^2}{2}\right)\right] \\ \qquad +(C\Delta + 1)gh\frac{\partial z_b}{\partial x} = -\frac{\tau}{\rho} \end{cases} \tag{30}$$

where C is the depth-averaged solid concentration, ρ is the water density, k is the active/passive coefficient employed in soil mechanics, determined by the depth-averaged ratio between the vertical and the longitudinal normal stresses (Savage and Hutter, 1991; Iverson, 1997); k is often assumed equal to one, as for pure fluids.

In two-phase models (equation 30) one more closure-assumption, relevant to the volumetric solid concentration C, is needed with respect to the homogeneous fluid model (equation 28). This relation can be derived from assuming a transport-capacity formulation, under the hypothesis that the processes of debris transport rapidly fit the dynamic changes (hypothesis referred to as *equilibrium*). Regarding the transport-capacity formulation to be used, equation (19) is an example concerning granular debris flows. Many other empirical laws can be considered, spanning from refinements of bed-load formulations often employed in torrential floods (Armanini, 2005), to more specific laws devoted to the massive sediment transport under consideration.

Under the same hypotheses, assuming, moreover, that the β and k coefficients are equal to unity, the two-dimensional extension of the mass and momentum balances can be derived, in which, as in the case of homogeneous fluid, the inner shear stresses can be neglected.

MITIGATION AND RISK REDUCTION MEASURES

The mitigation remedial and defence techniques against debris flows are in their present form relatively recent tools: in the past, in fact, people simply tried not to build in the areas where there had been previous debris-flow events.

Only recently, as a consequence of the frantic development of the tourism settlements, risk areas, frequently located along the alluvial fans, have started to be urbanized. In general, the defence works against the debris flows can be classified in two categories: the active countermeasures and the passive countermeasures.

The former consist basically of interventions aiming at reducing the risk of debris-flow triggering. Then they are meant to give stability to the debris deposits in the torrents beds or in the hillslopes. The latter instead are works built to defend directly the settlements or the zones subject to

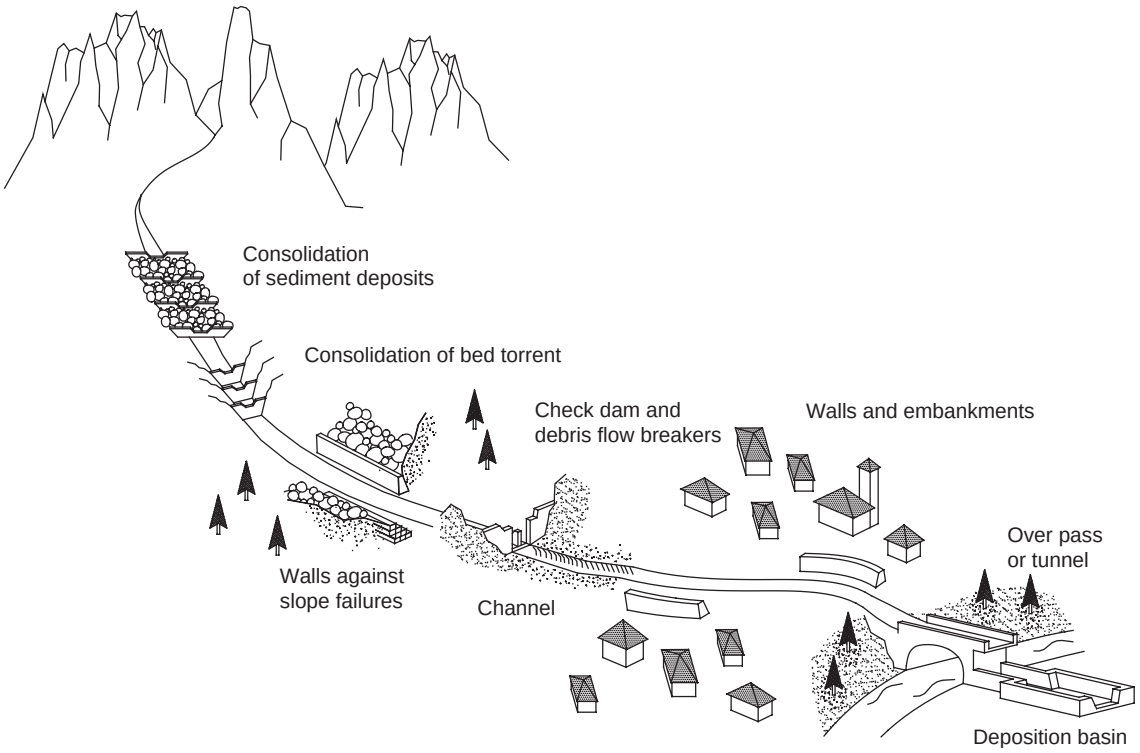


Figure 4 Scheme of the different types of active defence works against debris flows

debris-flow risk or the single structures (Figure 4). There are of course works responding to both criteria.

Consolidation of the Debris Deposits

The stabilization of the debris deposits is obtained either by reducing the slope or by hindering or reducing the possibility of saturation. This objective can be reached by consolidation works transversal to the deposits (endowed, where possible, with effective draining systems), by piling and anchors and, where possible, by vegetation. However, the bulkiest sources of the material mobilized by the debris-flow events are often located in places hardly reachable by the mechanical means; moreover, it is often counterproductive to move the deposits themselves in order to build drainages, as the movement of the material may increase its instability. As a result, the cases of applicability of these devices are often limited.

Consolidation of Riverbed and Lateral Slopes

In case the debris deposit affects directly the channel bed, the more limited the extent of the works is, the easier is the consolidation. The works constructed in these cases are sills or traditional closed check dams designed to reduce and to stabilize the riverbed slope and therefore to reduce the flow velocity. Often, in fact, the channel bed instability initially gives rise to an intense bed load that can degenerate into a debris flow owing to the erosion of the riverbanks or to the accumulations caused by the geometric variations of the sections (natural dam break). A system that augments the channel bed stability is represented by the chains of consolidation dams. The check dams built for this purpose are absolutely similar to the check dams used in case of solid ordinary transport. A phenomenon that characterizes the passing of the debris flow and that must be considered in the dimensioning of the check dam foundations is the amplification of flow depth from bed fluidification that accompanies the passing of the debris flow (Jäggi and Pellandini, 1997).

Debris-flow Breakers

More often, open check dams are built to intercept the solid material. However, debris flows usually occur in streams with a steep slope, where there is often lack of space where the material can be stored. In this case, the objective of the check dam is to diminish the flow velocity in order to reduce its destructive power in case of a dynamic impact. Therefore, these structures must resist above all the dynamic impact and they should also intercept the huge boulders.

Various devices have been proposed for this purpose. The most common structures are slit check dams with quite a large opening often protected by one or more stout buttresses that are dimensioned so that they can resist the dynamic impact (Figure 5).

Figure 5 Slit check dam with a series of debris-flow breaker on Torrent Chieppena – Trentino Italy (Provincia Autonoma di Trento, 2002). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The hydrodynamic working of such structures is not clear and practical criteria suggested by experience are used for their design. The most effective results are obtained by water separation because the reduction of water concentration increases considerably the energy dissipation inside the flow and consequently its velocity.

Yet it is important that the volume upstream of the check dam remains free after each event, therefore the opening must allow the removal of the material accumulated. Accordingly, it can be useful to dimension the opening, so that the excavators meant to remove the material accumulated can pass through it.

The upstream side of the buttresses is inclined with respect to the vertical so as to reduce the dynamic impact of the debris flow that depends on the normal component of velocity. The inclination adopted varies between 45% and 60%.

The overpressure Δp generated in the dynamic impact of the debris-flow front colliding against a vertical wall is obtained by applying the mass balance and the motion quantity balance, thus obtaining (Armanini and Scotton, 1993) the following expression:

$$\Delta p = \rho_{\mathrm{m}} U (U - a_{\mathrm{w}}) \tag{31}$$

where ρ_m is the debris-flow density, U is the mean velocity of the front and a_w is the velocity of propagation of the gravitational wave reflected by the wall.

Very often relation (31) is corrected with an opportune coefficient α_p in order to consider possible secondary effects. The impact coefficient α_p often includes the effect of the reflected wave:

$$\Delta p = \alpha_p \rho_s U^2 \tag{32}$$

The coefficient α_p varies from 2 for quite slow and dense debris flows to 0.7 for faster and more liquid debris flows.

Debris Flows Artificial Channels and Retention Basins

It has been already pointed out that often there is not sufficient space to intercept all the debris-flow volume upstream urbanized areas. Provided that debris-flow breakers and check dams have reduced the flow velocity and intercepted part of the largest boulders, it is necessary to canalize the torrent downstream the retaining works, often by building artificial channels able to deviate the debris flow. These channels usually cross the villages and the alluvial fans where the space available is limited. Yet, it is useful to distinguish between a channel designed for the water discharge only and one affected by remarkable solid discharges or by debris flows. Often, it is the second necessity that is more urgent. Then the channel is supposed to resist to high velocities and strong tangential stresses, and must offer a very stable bed. Therefore, it is often necessary to resort to channels covered and strengthened with concrete. In this case, the covering surface is often smoothed to reduce the wall roughness. This is not the case of the channel for debris flow (Figure 6) where the resistance does not depend on surface roughness, but on particle collisions.

Figure 6 Artificial channel on Torrent Gola – Trentino Italy (Provincia Autonoma di Trento, 1991). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Bends with strong curvatures must be avoided in order to prevent depositions induced by secondary effects due to the supercritical nature of the flow, and in any case free-surface elevation must be properly accounted for, in order to avoid lateral debris-flow flooding.

In other cases, it can be useful to deviate the debris flow on the side to safeguard some sites. This deflection is obtained by the same techniques used for the snow avalanches, that is, by diverting walls or dikes.

The topography of the torrent often does not offer spaces granting a sufficient volume for the deposition of the solid material. In this case, it is necessary to create such storages. The debris retention basins are created through lateral training dikes, protected downstream by a check dam, which, if necessary, is inserted into the body of an artificial banking.

To optimize the volume available, it is necessary to remember that the widening cone of an overcritical current depends on the Froude number of the incoming current.

Acknowledgment

The authors thank Prof. Riccardo Rigon for his help in preparing the section about the triggering of debris flow.

REFERENCES

Armanini A. (2005) Mountain streams. *Encyclopedia of Hydrological Sciences*, John Wiley and Sons.

Armanini A., Capart H., Fraccarollo L. and Larcher M. (2005) Rheological stratification in experimental free-surface flows of granular-liquid mixtures. *Journal of Fluid Mechanics*, **532**, 269–319.

Armanini A., Fraccarollo L. and Larcher M. (2003) Dynamics and energy balances in uniform liquid-granular flows, *Proceedings of FLOWS 2003, International Workshop on Occurrence and Mechanisms of Flows in Natural Slopes and Earthfills*, Sorrento, 14–16 May 2003, L. Picarelli editor. Patron editore: pp. 131–137.

Armanini A. and Scotton P. (1993) On the dynamic impact of a debris flow on structures. *Proceedings of the XXV IAHR Congress*, Tokyo, vol. B, paper no. 1221.

Azanza E., Chevoir F. and Moucheront P. (1999) Experimental study of collisional granular flows down an inclined plane. *Journal of Fluid Mechanics*, **400**, 199–227.

Bagnold R.A. (1954) Experiments on a gravity-free dispersion of large solid spheres in a Newtonian fluid under shear. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, **225**, 49–63.

Beven K. and Kirkby M.J. (1979) A physically-based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Brown R.L. and Richards J.C. (1970) *Principles of Powder Mechanics*, Pergamon Press: London.

Campbell C.S. (1990) Rapid granular flows. *Annual Review of Fluid Mechanics*, **22**, 57–92.

Carnahan N.F. and Starling K. (1969) Equations of state for non-attracting rigid spheres. *Journal of Chemical Physics*, **51**, 635–636.

Chapman S. and Cowling T.G. (1970) *The Mathematical Theory of Non-Uniform Gases*, *Third Edition*, Cambridge University Press.

Coussot P. (1997) *Mudflow Rheology and Dynamics*, A. A. Balkema: Rotterdam.

Fraccarollo L. and Capart H. (2002) Riemann wave description of erosional dam-break flows. *Journal of Fluid Mechanics*, **461**, 183–228.

Hampel R. (1968) *Geschiebeablagerungen in Wildbächen, Dargestellt am Modellversuchen – Teil 1 und 2* . Wildbach- und Lawinenverbau, Nr. 1–2, 32nd Year.

Hunt M.L., Zenit R., Campbell C.S. and Brennen C.E. (2002) Revisiting the 1954 suspension experiments of R. A. Bagnold. *Journal of Fluid Mechanics*, **452**, 1–24.

Iida T.A. (1999) Stochastic hydro-geomorphological model for shallow landsliding due to rainstorm. *Catena*, **34**, 293–313.

Iverson R.M. (1997) The physics of debris flows. *Reviews of Geophysics*, **35**(3), 245–296.

Jäggi M.N.R. and Pellandini S. (1997) *Torrent Check Dams as a Control Measure for Debris Flows. Lecture Notes on Earth Sciences*, *Vol. 64, Recent Developments of Debris Flows*, Armanini A. and Michiue M. (Eds.), Springer-Verlag: pp. 186–205.

Jenkins J.T. and Hanes D.M. (1998) Collisional sheet-flow of sediment driven by a turbulent fluid. *Journal of Fluid Mechanics*, **370**, 29–52.

Jenkins J.T. and Savage S.B. (1983) A theory for the rapid flow of identical, smooth, nearly elastic particles. *Journal of Fluid Mechanics*, **130**, 187–207.

Lun C.K. and Savage S.B. (1986) The effects of an impact dependent coefficient of restitution on stresses developed by sheared granular materials. *Acta Mechanica*, **63**, 15–44.

Ogawa S. (1978) Multitemperature theory of granular materials. *Proceedings of the US-Japan Seminar on Continuum Mechanics and Statistical Approaches to Mechanics of Granular Mater*, Gukujutsu Bunken Fukyukai: Tokyo, pp. 208–217.

Provincia Autonoma di Trento (1991) *Per Una Difesa Del Territorio*, Edizioni Arco Trento.

Provincia Autonoma di Trento (2002) **http://www.sistemazionemontana.provincia.tn.it/**

Rigon R., D'Odorico P. and Bertoldi G. (2005) The peak flows and their geomorphic structure. Submitted to *Water Resources Research*.

Savage S.B. (1984) The Mechanics of Rapid Granular Flows. *Advances in Applied Mechanics*, **24**, 289–366.

Savage S.B. (1998) Analyses of slow high-concentration flows of granular materials. *Journal of Fluid Mechanics*, **377**, 1–26.

Savage S.B. and Hutter K. (1991) The dynamics of avalanches of granular materials from initiation to runout. Part I: analysis. *Acta Mechanica*, **86**, 201–223.

Savage S.B. and Jeffrey D.J. (1981) The stress tensor in a granular flow at high shear rates. *Journal of Fluid Mechanics*, **110**, 255–272.

Takahashi T. (1978) Mechanical characteristics of debris flow. *Journal of Hydraulics Division, ASCE*, **104**(8), 1153–1169.

Takahashi T. (1981) Debris flow. *Annual Review of Fluid Mechanics*, **13**, 57–77.

Takahashi T. (1991) *Debris Flow*, IAHR Monograph, A. A. Balkema: Rotterdam.

143: Mountain Streams

ARONNE ARMANINI

Department of Civil and Environmental Engineering, University of Trento, Trento, Italy

This paper illustrates the main features of the hydraulics of mountain streams. The effects of small submergence and riverbed steepness on sediment motion initiation and transport are discussed together with the effect of big roughness elements and pool-riffle and step-pool sequences on velocity distribution and flow resistance. Finally, the most used torrent defense techniques are described.

INTRODUCTION

A definite distinction between mountain and valley rivers is practically impossible, except for the obvious observation that mountain water courses are those that are located in mountain areas. However, skipping a rigorous definition, there are many features that characterize mountain rivers. First of all, mountain streams are characterized by high slopes: for example, gradients greater than 0.02 along the majority of the length. Other strong differences concern the morphological, geological, and hydrological characteristics. Natural mountain streams present strong boundary variations both in the geometry and in the roughness elements that compose their bed. Generally speaking, these variations are due to geological constraints prevailing over the sedimentological influences that determine the features of valley rivers. Moreover, the hydrological regimes of mountain streams are very unsteady, characterized by intense floods and sediment transport peaks.

According to Wohl (2000) the most relevant characteristics of the mountain rivers are:

- *steep average gradients;*
- *high channel-boundary resistance and high boundary roughness from bedrock and coarse clasts that are more likely to be present along these channels than along low-gradient channels;*
- *highly turbulent flow and stochastic sediment movement resulting from the steep gradient and rough channel boundaries;*
- *strongly seasonal regime with high spatial and temporal discharge variability resulting from the effect of changes in precipitation with elevation and basin orientation;*
- *channel morphology that has high spatial variability because of the external control of geology, but low temporal variability because only infrequent flood or debris flows are able to exceed channel-boundary resistance*.

In particular, the sediments transport is dominated by the bed load and is very unsteady, ranging from long periods of limited sediment supply to intense events able to determine major morphological changes.

However, the concept of mountain river covers a great variety of different situations both from a hydrological point of view, because of different regional conditions, and from a morphological point of view because of different altitudes. Nonetheless, the hydraulic tools used in the analysis of mountain streams are not very different from those employed in valley rivers, apart from the fact that certain parameters play a more important role in these conditions. An evident difference between mountain streams and lower-gradient rivers is due to the fact that the geometry of a valley stream is largely determined by its water and sediment regimes: they are the "*authors of their own geometry*" (Leopold and Langbeim, 1962). In mountain streams, watershed morphology and geology are mainly responsible for the characteristics of their streams, maybe with the exception of their parts flowing into the alluvial fans.

VELOCITY DISTRIBUTION

The analysis of river hydraulics is generally based on the ideal concept of (temporary) steady and (locally) uniform conditions. While the low-gradient rivers for most of the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

time and far away from the control sections are not far from satisfying this condition, at least locally, in mountain rivers it is very rare to find reaches that can be considered uniform, mainly because of the great variability of the geometry of the cross sections. However, the reference to uniform and steady conditions is often necessary in order to apply a very large amount of mathematical and physical findings and technical tools developed in low-gradient hydraulics.

Accordingly, it is possible to affirm that the hydraulics of mountain rivers is based on the same principles as the hydrodynamics of rivers. Generally speaking, one important difference lies in the relative roughness, that in mountain rivers is generally one or more orders of magnitude higher that in low-land rivers. The wall in this situation can be considered hydraulically rough and the flow is determined by the turbulence generated by the roughness wall elements. The velocity distribution in the *inner region* ($y/h <\sim 0.15$) is often expressed by a logarithmic law:

$$\frac{u}{u_*} = \frac{1}{\kappa} \ln \frac{y + y_o}{d_s} + B_r \tag{1}$$

where $u_* = \sqrt{\tau_o/\rho}$ is the friction velocity, y the direction normal to the bed and y_o a reference height, $\kappa \simeq 0.4$ the von Kármán constant and d_s a suitable characteristic grain size. B_r is an integration constant, depending mostly by the wall roughness elements. In the sand-grain bed this constant assumes the value 8.5. In large relative roughness B_r assumes values ranging from 3.25 to 6.26 (Graf, 1989).

In very small relative submergence, the existence of a layer near the top of roughness elements (*roughness layer*) dominated by the wakes generated by the individual grains has been recognized. Within this layer the local velocity distribution along the depth tends to be more uniform, and Reynolds stresses ($\overline{u'_x u'_y}$) tend to be suppressed (Nakagawa *et al.*, 1989). This situation seems to lead to an imbalance of the longitudinal momentum because the longitudinal component of the water weight would not be balanced by a proper shear stress. The roughness layer, however, is characterized by strong space heterogeneity. In order to write the momentum balance correctly, it is useful (Nikora *et al.*, 2001) to perform a double integration (in time and in space along planes parallel to the channel bed). As a consequence of the double integration process in the momentum balance, some extra terms, called *dispersive stresses* (or *form-induced stresses*), appear. Four layers may be identified: (a) interfacial sublayer, (b) form-induced sublayer (c) logarithmic layer and (d) outer layer (Figure 1).

According to the these authors, in the form-induced layer and in the interfacial sublayer the velocity distribution, double-averaged on space and time, $\langle \overline{u}(y) \rangle$, presents a linear

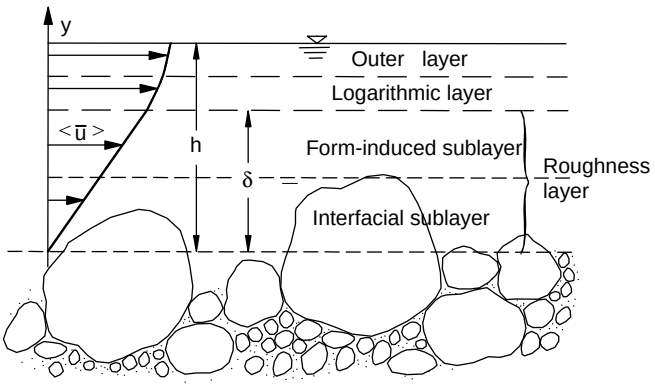


Figure 1 Sketch of layers distribution in small submergence conditions in gravel-bed rivers (Reproduced from Nikora *et al.*, 2001)

distribution

$$\frac{\langle \overline{u} \rangle}{u_*} = C \frac{y}{\delta} \tag{2}$$

while the logarithmic layer is characterized by a velocity profile similar to equation (1).

$$\frac{\langle \overline{u} \rangle}{u_*} = \frac{1}{\kappa} \ln \frac{y}{\delta} + C \tag{3}$$

C is a coefficient depending on roughness that in case of small relative submergence ranges from 3 to 7; δ is the boundary between the linear and the logarithmic regions. The outer layer is similar to the homologous layer for a hydraulically smooth bed.

According to Nikora *et al.*, (2001), in a very small relative submergence flow ($2 \div 5 > h/k_e \geq 1$), where h is the water depth and k_e is the roughness height, it is possible to postulate that the linear relation (2) is valid over the whole depth, and to assume for the height of the linear layer, $\delta = d_{84}$.

HYDRAULIC RESISTANCE

As a rule, the resistance law of a mountain stream can be obtained by cross-section (or depth) integration of velocity distribution (e.g. equation (1) or equations (2) and (3)). In addition to the grain resistance, in very steep channels, a form resistance also should be taken into account.

In the literature, many empirical formulas on hydraulic resistance of mountain rivers (gravel-bed and boulder-bed rivers) can be found. It is advantageous to divide them into two groups: the logarithmic and the power law. The logarithmic resistance formulas have the form of Nikuradse formula for flat channel flow, with different coefficients, and mostly with different expression for the roughness

Table 1 Parameters for the logarithmic formula (4) according to various authors. R_h = hydraulic radius

A	B	Y	Remarks	Authors
2.90	0.70	h/d_{50}	$2.2 \leq h/d_{90} \leq 31$	Bray, (1979)
2.80	1.72	h/d_{65}		
2.49	$2.88 \div 3.02$	h/d_{84}	$1 < R_h/d_{65} < 14$	Hey, (1979)
$2.457(1 - 0.1k_e/R_h)$	0	$12R_h/k_e$	$1.2 < h/d_{84}k_e = 4.5d_{50}$	Thompson and Campbell, (1979)
2.43	2.15	R_h/d_{50}	$1 \leq R_h/d_{50} \leq 200$	Griffiths, (1981)
2.5	2.88	$(R_h/d_{84})(h_m/R_h)^{0.314}$	$0.3 \leq R_h/d_{84} \leq 1 h_m$ = max. flow depth	Bathurst, (1982)
2.48	3.1	R_h/d_{84}	$1 < h/d_{50}$	Bray and Davar, (1987)
3.28	2.43	R_h/d_{84}		Limerinos, (1970)
2.5	3.25	R_h/d_{50}	$1 \leq R_h/d_{50} \leq 50$	Graf *et al.*, (1983)

parameter:

$$\frac{U}{u_*} = A \ln Y + B \tag{4}$$

U is the depth-averaged velocity, Y is a dimensionless parameter representative of relative roughness, A and B are empirical constants. As equivalent roughness, many authors suggest to take $k_e = (3 \div 3.5)d_{84}$. Table 1 reports the parameters of formula (4).

Often, instead of the log formula (4), a power law formula is preferred:

$$\frac{U}{u_*} = A_1 Y^{\beta} \tag{5}$$

The coefficient A_1 and the exponent β depend on the definition of relative submergence adopted and vary from author to author (Table 2).

The first two formulas have been derived for relatively low-gradient rivers, but are widely employed in engineering works in mountain rivers.

Expressions linking directly the discharge to the channel, or basin characteristics are sometimes employed: however, these formulas are calibrated for a specific stream and do not have a universal validity.

THE INITIATION OF SEDIMENT MOVEMENT

The mechanics of the initiation of sediment motion in mountain streams does not differ substantially from that in small-gradient streams, except for the influence of some parameters being usually negligible in the latter: in particular, the effect of slope, small submergence, and grain-size heterogeneity.

The dimensionless groups affecting incipient motion condition may be pointed out by means of dimensional analysis. The motion initiation condition of a particle lying on the stream bed is expressed by a relation between the following nine parameters:

$$\text{funct}(U, h, \tau_o, \rho, \rho_s, g, d_s, \mu, \sigma_s) = 0 \tag{6}$$

Where, ρ and ρ_s are the density of the fluid and of the sediments respectively, g is the gravity acceleration, d_s is

Table 2 Parameters for the power law formula (5) according to various authors

A_1	β	Y	Remarks	Authors
6.74	0.167	R_h/d_{50}	Sand and gravel mixtures	Strickler (1923)
8.30	0.167	R_h/d_{90}	Sand mixtures	Meyer-Peter and Müller (1948)
5.4	0.25	R_h/d_{84}	$1 < h/d_{50}$	Bray and Davar (1987)
3.85	0.281	h/d_{50}	Natural gravel rivers	Bray (1979)
4.19	0.2769	h/d_{65}		
5.03	0.268	h/d_{90}	$2.2 \leq h/d_{90} \leq 31$	
$(B/h)^{7(\lambda_1 - 0.08)}$	2.34	$R_h/(0.365d_{84})$	$R_h/d_{84} < 1.2$	Bathurst (1978)
$\lambda_1 = 0.139 \log 1.91 \frac{d_{84}}{R_h}$			λ_1 = frontal roughness concentration; B = channel width	
1.48	1.80	R_h/d_{84}	$i \sim 0.06$ step-pool streams	Lee and Ferguson (2002)
3.84	0.547	h/d_{84}	$0.002 < i < 0.008$ / $0.008 < i < 0.04$ } $h/d_{84} < 11$	Bathurst (2002)
3.10	0.93			

the size of a diameter representative of bed material and σ_s is a suitable parameter representative of the grain-size distribution (e.g. standard deviation of grain-size distribution). By means of the Π theorem relation, (6) reduces to a relationship between six dimensionless groups:

$$\text{funct}\left(\frac{u_*^2}{g\Delta d_s}, \frac{\rho u_* d_s}{\mu}, \frac{h}{d_s}, \frac{U}{u_*}, \frac{U}{\sqrt{gh}}, \sigma_s\right) = 0 \qquad (7)$$

where $\Delta = (\rho_S - \rho)/\rho$ represents the relative submerged density of the bed particles. The first group is the Shields' mobility parameter, $\tau_{*c} = u_{*c}^2/g\Delta d_s$. The second group is the grain Reynolds number, $R_* = u_{*c}/g\Delta d_s$. The third ratio is the relative submergence, h/d_s. The fourth group is the relative resistance, $\sqrt{8/f} = U/u_*$. The fifth group is the Froude number, $F_r = U/\sqrt{gh}$. The last parameter σ_s is related also to grain size, roughness distribution, and spacing, but, as for flow resistance, the dependence on this parameter can be included in the definition of the representative diameter. The grain-size distribution, however, is very important in all the cases where a selective removal or deposition of material occurs.

The direct dependence on relative resistance U/u_* can be omitted because it depends on relative submergence h/d_s and on Froude number F_r. According to Shields (1936), the motion initiation of the particles laying on a bed of a small-gradient river depends on the mobility parameter and on the grain Reynolds number: $\tau_{*c} = f_{\text{Sh}}(R_*)$ (Figure 2).

Compared with Shields' analysis, then, the critical shear stress (incipient motion) in mountain streams shows a further dependence on relative submergence h/d_s and Froude number F_r (Graf, 1989).

Different combinations of dimensionless groups have been used by different authors. In the mountain restoration technique, the critical slope, the critical discharge (Schoklitsch, 1962; Graf, 1984), and the critical Froude number are often used. It is possible to account for the new parameters by multiplying the original Shields' function f_{Sh} by a special function that becomes a unit function when the influence of these new parameters disappears. The effect of the longitudinal slope is accounted for just by introducing in the force balance the effect of the longitudinal component of the particle submerged weight (Christensen, 1996):

$$\tau_{*c} = f_{\text{Sh}}\left(\cos\alpha - \frac{\rho_s}{\rho_s - \rho}\frac{\sin\alpha}{\tan\varphi}\right) \qquad (8)$$

where φ is the angle of repose of the bed material and α is angle of channel bed.

Experimental evidence shows that the critical mobility decreases when submergence decreases, owing to the effect of small submergence. An empirical relation matching the experimental data is the following (Armanini, 1999):

$$\tau_{*c} = f_{\text{Sh}}\left(1 + 0.67\sqrt{\frac{d_s}{h}}\right) \qquad (9)$$

According to some authors, instead of the mobility parameter, which depends on bed shear stress, it is more advantageous to refer directly to the critical liquid discharge (Schoklitsch, 1962):

$$q_{\text{cr}}^* = \frac{q_{\text{cr}}}{d_s\sqrt{gd_s}} \qquad (10)$$

Bathurst *et al.*, (1987) suggest that, in nearly uniform conditions, the critical discharge depends substantially on the local bed slope:

$$q_{\text{cr}}^* = 0.21 i_f^{-1.12} \qquad (11)$$

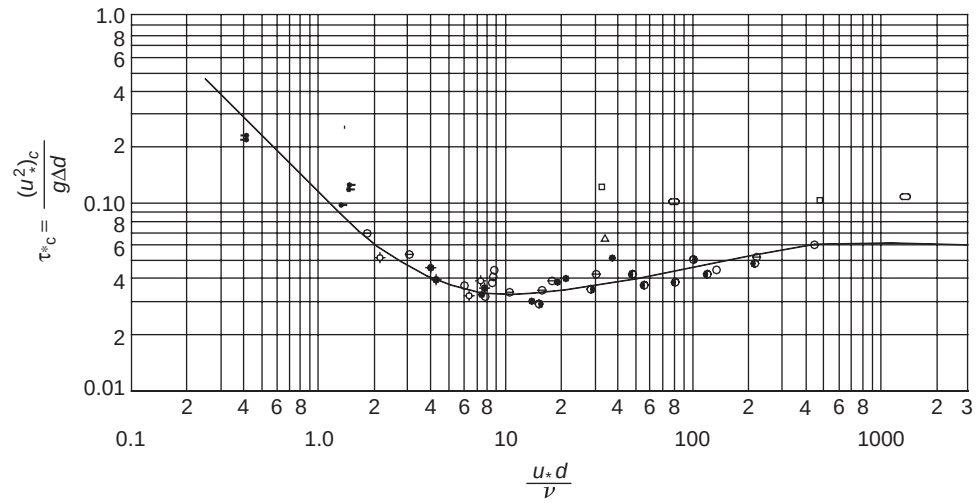


Figure 2 Shields' diagram

where $d_s = d_{50}$ has to be assumed as particle diameter in equation (10); the relation (11) has been calibrated for $0.25\% < i_f < 10\%$. By contrast, according to Whittaker and Jäggi (1986), it is

$$q^*_{\mathrm{cr}} = 0.257 \left(\frac{\rho_s - \rho}{\rho} \right)^{0.5} i_f^{-1.12} \qquad (12)$$

Hiding and Armoring

Mountain rivers are characterized by a wide grain-size distribution. This peculiarity has important consequences both on sediment movement initiation and on sediment transport. In a mixture of large and small particles, the larger ones tend to move more easily than if they were in a bed of uniform size because of their greater exposure, while smaller particles are protected by the larger ones from the movements (Figure 3). This phenomenon is termed *hiding*.

Since according also to Shields' criterion the mobility of a particle is inversely proportional to its size, the exposure effect makes the mixture movement more equalized.

The hiding effect is usually accounted for by means of a suitable *hiding* (or *exposure*) *coefficient* that multiplies the mobility of each individual fraction, calculated as if the bed was uniformly composed of material of this fraction. The hiding coefficient is therefore a function of the size of the individual fraction and of the average grain size (Table 3).

A second effect due to the nonuniformity of the grain size is the bed *armoring*.

It has been frequently observed that in certain conditions the surface of a gravel river is composed of a layer of material coarser than that composing the layers immediately below the bed surface. This effect is due to the fact that when the sediment supply vanishes or nearly vanishes, the finer fractions tend to be removed from the bed surface while the value of the bed shear stress is not high enough to move the coarser fraction. Then the size composition of the surface tends to become coarser. The difference between the surface and the subsurface material in the armored layer is often very strong, and the rupture of the armored layer is an instantaneous event that can move large quantities of material and it can be accompanied by evident erosion and deposition phenomena, which can damage the structures along the water course. The mechanism creating and destroying the armor layer is enhanced during the rising and falling limb of the hydrograph.

The grain-size distribution of the armored layer has been analyzed both theoretically and experimentally giving the

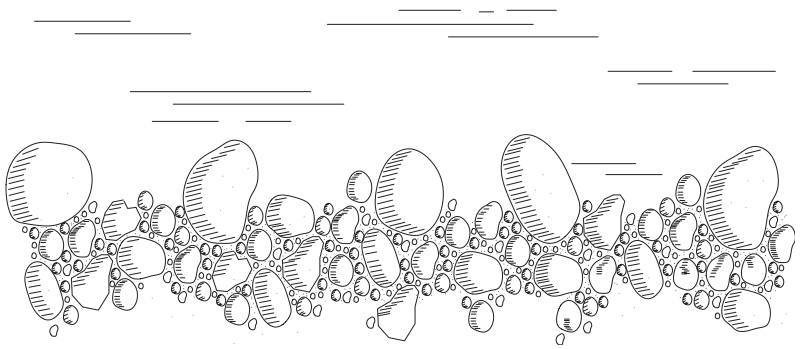


Figure 3 Sketch of hiding

Table 3 Hiding coefficients according to different authors

Formula		Remarks	Authors
$\xi_j = \left(\frac{\log 19}{\log 19 \frac{d_j}{d_{\mathrm{m}}}} \right)^2$	(13)	d_{m} = mean size of the mixture	Egiazaroff (1965)
Equation (13)		for $d_j/d_{\mathrm{m}} \geq 0.4$	
$\xi_j = 0.85 \frac{d_{\mathrm{m}}}{d_j}$	(14)	for $d_j/d_{\mathrm{m}} < 0.4$	Ashida and Michiue (1973)
$\xi_j = \left(\frac{d_{50}}{d_j} \right)^n$	(15)	$n = 0.74$	Ashworth and Ferguson (1989)
		$n = 0.87$	Andrews (1983)
		$n = 0.88$	Ferguson *et al.*, (1989)
		$n = 0.94$	Diplas (1987)
		$n = 0.98$	Parker *et al.*, (1982)

following expression:

$$F_j = f_j \frac{d_j^n}{\sum f_j d_j^n} \tag{16}$$

F_j is the percentage of the jth fraction in the armored layer and f_j is the percentage of the same fraction in the subsurface layer, which is assumed to be identical to the grain-size distribution of transported material. The exponent n in equation (16) is not far from unity ($n = 1.12 \div 1.35$) (Parker and Sutherland, 1990).

Because of the hiding effect in an armored bed, the critical shear stress corresponding to each grain-size fraction $(\tau_{*c})_j$ is considerably affected by the presence of surrounding particles of different size. Monomial expressions are often used to express the fractional mobility

$$(\tau_{*c})_j = (\tau_{*c})_u \left(\frac{d_j}{d_{us}}\right)^{-n} \tag{17}$$

Andrews (1983) proposed n = 0.872 and $d_u = d_{50s}$ (the median diameter of the subsurface bed material), and $(\tau_{*c})_u = 0.0834$ (the relative dimensionless critical shear stress).

From equation (16), it is possible to obtain the critical shear stress of an armored bed. For example, Günter (1979) introduced the following empirical relation between the critical dimensionless shear stress $(\tau_{*c})_{dmo}$ of an armored bed and the dimensionless shear $(\tau_{*c})_e$ for uniform bed material, the mean grain size of bed surface material, d_{ms}, and the mean grain size of bed subsurface material, d_{mo}:

$$(\tau_{*c})_{dmo} = (\tau_{*c})_e \left(\frac{d_{ms}}{d_{mo}}\right)^{0.67} \tag{18}$$

The theoretical analysis of heterogeneous-sized material (Parker and Sutherland, 1990) has highlighted that a surface coarsening occurs also in normal upstream sediment feeding conditions. In this situation, termed *dynamic armoring*, the mean size of the material transported is generally finer than that comprising the channel bed surface, but is slightly coarser than that composing the subsurface layer. The dynamic armoring is due to the nonlinearity of the relation between solid discharge and grain-size distribution of the material.

SEDIMENT TRANSPORT

The small relative submergence and the direct effect of gravity forces on the particles modify also the sediment transport formulas. With respect to the quantities considered in the dimensional analysis of the previous section, one more parameter relevant to sediment transport rate should be introduced. The corresponding dimensionless group is usually given by the Einstein (1950) transport parameter:

$$\Phi = \frac{q_S}{d_s \sqrt{g \Delta d_s}} \tag{19}$$

In the most common formulas used in the small-gradient streams, Φ depends on the mobility parameter τ_*, in some cases on a suitable grain Reynolds number and on the same parameter under critical conditions τ_{*c}. In the case of mountain streams, the transport parameter Φ should depend also on relative submergence and on Froude number. However, it should be noted that the critical mobility parameter τ_{*c} does not appear in the dimensional analysis and that this parameter implicitly accounts for slope and relative submergence. Nonetheless, in some formulas recommended for mountain streams, the bed slope or Froude number explicitly appears (Table 4).

Schoklitsch (Schoklitsch, 1962) type formulas are also recommended for mountain streams with slope $i >\sim 0.01$ (Bathurst *et al.*, 1987) and for steeper channels ($i >\sim$

Table 4 Different sediment transport formulas used in mountain streams

Formula		Remarks	Authors
$\Phi = 8(\tau_*' - \tau_{*c})^{1.5}$	(20)	Proposed for $d_{50} \leq 7$ mm, but widely used also for coarser material	Meyer-Peter and Müller (1948)
$\Phi = 4\left(\frac{d_{90}}{d_{30}}\right)^{0.2} i_f^{0.6} \tau_*^{0.5}(\tau_* - \tau_{*c})\frac{U}{u_*}$	(21)	$0.03 \leq i_f \leq 0.20$	Smart and Jäggi (1983)
$\Phi = \frac{3.1}{\Delta^{0.5}}\left(\frac{d_{90}}{d_{30}}\right)^{0.2} \tau_*^{0.5}(\tau_* - \tau_{*c})F_r^{1.1}$	(22)	$0.004 \leq i_f \leq 0.20$	Rickenmann (1991)
$\Phi = 17\tau_*'^{1.5}\left(1 - \frac{\tau_*}{\tau_{*c}}\right)\left(1 - \sqrt{\frac{\tau_{*c}}{\tau_*}}\right)$	(23)	Used also for debris flows	Ashida and Michiue (1971)
$\Phi = 11.933\tau_*^{1.5}\left(1 - \frac{0.0329}{\tau_*}\right)$	(24)	$\tau_* > 0.06$	Parker (1989)
$\Phi = 10.4\tau_*^{1.5}\left(1 - \frac{0.045}{\tau_*}\right)^{2.5}$	(25)	$d = d_{50}$; $\tau_* > 0.0659$ (the Authors give $\Phi \leq 10^{-2}$)	Graf and Suszka (1987)

$0.05 \div 0.17$) (Rickenmann, 2001). This type of formulas are based on critical liquid discharge of equation (10).

$$q_s = 2.5\frac{\rho}{\rho_s} i^{1.5}(q - q_{cr}) \tag{26}$$

Schoklitsch type equations are preferred in mountain streams because they do not contain explicitly the water depth or the bed shear stress, quantities often difficult to be determined because of the geometric heterogeneity that characterizes mountain water courses.

In some cases, the sediment transport in mountain streams presents a two-phased feature: up to a certain value of the discharge, the transport load is composed of relatively fine material, while, when the discharge exceeds a threshold value, the grain-size composition of the transported material changes rather abruptly and becomes coarser. (Ryan *et al.*, 2002). This phenomenon is associable with the rupture of the armored layer especially in the pool-riffle sequences. The pools feed the finer material that characterizes phase 1; during phase 2, coarser particles are supplied by break-up of armored layer of riffle units.

Many of the anomalies that characterize mountain streams are probably due also to the wide grain-size distribution of the bed material. Often the representative diameter of the bed material is coarser than the mean diameter of the sediment supply causing static and dynamic armoring, two phases sediment transport, apparent non equilibrium between sediment transport and transport capacity, and so on. It is possible in some cases to account separately for the contribution of each single fraction. In order to estimate the global transport rate, one must first calculate the transport rates q^*_{sj} for each single fraction – as if the bed was uniform in size and composed of the material of the individual class – and multiply these values by the percentage f_j of the material of the jth class present on the bed surface, accounting for the hiding factor, and summing this product up to all the classes considered:

$$q_s = \sum_j^M f_j q^*_{s\xi j} \tag{27}$$

where $q^*_{s\xi j}$ is the transport capacity of each single class modified by the hiding factor ξ_j.

MOUNTAIN CHANNEL MORPHOLOGY

The morphology of mountain rivers is extremely variable, ranging from bed rock–dominated systems to braiding or meandering reaches in the alluvial low gradient parts.

In the literature, different channel classification systems for mountain rivers are present. Montgomery and Buffington (1997) indicate step pool, plane bed, and pool riffle as possible categories at reach scale to classify the mountain streams.

In step-pool sequences, steps are usually composed of groups of boulders organized along straight or curved lines. Woody debris steps often occur. Unlike the dunes in low-slope rivers, step and pool geometry is scarcely regular. Typical step length ranges from tens of centimeters to tens of meters.

According to Wohl (2000), step and pool sequences occur at bed gradients ranging from 3 to 10%; Whittaker and Jäggi (1986) suggest a minimum bed slope of 7.5%. Step and pool configurations do not have to be considered as a permanent feature of a mountain reach as they can be suppressed or substantially modified during major flood events, while the most evident sequences form during floods with a return period ranging from 20 to 50 years. Step-pool sequences affect also the local and the average flow resistance: form resistance prevails over grain resistance. The flow regimes change from supercritical just after the steps to subcritical just upstream the chutes, with a hydraulic jump and strong energy dissipation in the transitions.

The morphological conditions in which step-pool sequences form and their formation mechanism are not clear. Two major theories have been suggested (Abrahams *et al.*, 1995). A first hypothesis states that step-pool sequences derive from antidunes disturbed by a large heterogeneity in the grain size and, in particular, by the presence of larger boulders that have fixed some of the antidunes thereby hindering their migration. A second approach is based on the hypothesis that the bed forms must maximize the flow resistance. Flume results and findings on natural step-pool streams suggested the following relations between average step length $\overline{L}$ and height $\overline{H}$ and average channel slope angle α (Figure 4):

$$\frac{\overline{H}}{\overline{L}} = 0.67(\sin\alpha)^{0.682} \tag{28}$$

A similar relation has been suggested also by Wohl and Grodek (1994). These relations are based on the assumption that the step height $\overline{H}$ is independent of the average bed slope. Average pool spacing is about four times the water depth.

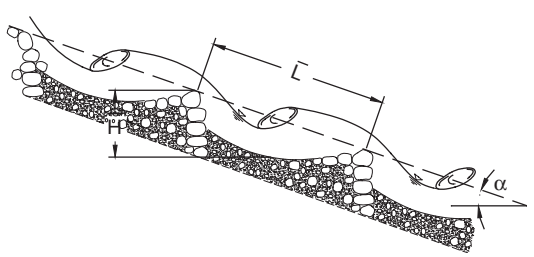


Figure 4 Sketch of step-pool sequence

Plane bed without regular bed forms characterizes mountain reaches with slopes between 0.01 and 0.05. Plane bed reaches have a relatively homogeneous grain-size bed composed of gravel and cobbles and a relatively uniform hydrological regime. The bed, often armored and with the absence of standing bars, probably indicates a limited sediment supply.

Pool-riffle reaches are nearly periodic planimetric structures characterized by sequences of relatively deep depressions (pools) followed by high-gradient shallows (riffles). Pool-riffle sequences occur at moderate slopes (0.05–1%) and are generally unconfined. Typical periodic distance between consecutive pools is about 5–7 channel width (Leopold and Wolman, 1957). Pools, often related to local scouring process due to flow convergence, are composed of relatively finer sediments and characterized by relatively low velocity at low stage discharges. Riffle units are often armored. Different theories on stability of pool-riffle sequences have been proposed in the literature: minimal potential energy loss or minimum power expenditure, energy reversal, and also hydrodynamic instability by deformation of alluvial bars similar to that in low-gradient rivers (Whittaker and Jäggi, 1982).

Montgomery and Buffington (1997) has given an account of forced riffles induced by large wood debris especially in forested channels.

TORRENT CONTROL CRITERIA

Sediment transport, bed and side erosion, and deposition processes in mountain rivers are phenomena that may be several orders of magnitude stronger than in valley rivers. For this reason, the defense or protection criteria in mountain environment concern mostly the control of sediment transport, while in low-gradient rivers, protection devices are designed mostly with respect to liquid discharges and volumes.

The control of sediment transport during floods is essentially a problem of controlling a large quantity of sediments released during short, intense events. In mountain streams, the sediment can be transported as an ordinary sediment transport (*bed load* and *suspended load*), mainly owing to the hydrodynamic forces of the water, or as a sediment gravity flow (*debris-flow* or *mud flow*) depending on the morphology, the concentration and quality of sediments (grain-size distribution), and the modes of mobilization. The gravity sediment transport is typical of very steep torrents, and its treatment requires a particular approach and appropriate techniques, which will be considered in a specific **Chapter 142, Debris Flow, Volume 4** (Armanini *et al.*, 2005).

The following strategies are often proposed in order to reduce the sediment transport movement along the river and to control the erosion/deposition processes:

- Prevention of bed and banks erosion through the so-called *consolidation check dams* as nonerodible sills in a river.
- Prevention of downstream overdeposition by structures capable of capturing the sediments in the upper part of the stream. The usual technique consists in the construction of one or more *retention dams*. The shape of this second kind of dam is often the same as the previous one, but its size and the purpose are different.

In the past, the strategy for preventing the further transport of sediment along rivers, and above all along torrents, was based on the building of high and large dams to trap more sediments and to resist high impact pressure.

The designed discharge of these structures depends on the importance of the work and the settlements to be protected. Often, a discharge corresponding to a return period of 50–100 years is assumed. The central weir of the dam is calculated as a trapezoidal broad-crested weir (Figure 5):

$$Q = \sqrt{\frac{4}{27}}\left(B_g + \frac{4}{5}nh_{\mathrm{m}}\right)\sqrt{2g}h_{\mathrm{m}}^{1.5} \tag{29}$$

It was soon observed that the space available for retaining the sediments is usually filled in a relatively short time, after which the dam practically loses its main purpose. With a traditional check dam, the whole sediment transport is stopped, causing a strong sediment imbalance in the downstream reach of the river. The reappearance of erosion, in sections previously subject to deposition, has been observed many times.

The next step was the development of *open dams* (Figure 6) with different purposes and different shapes. Meanwhile, it was gradually established that the purpose of sediment control was not to stop all the sediment transport but to reduce the sediment peak discharge by means of a temporary retention process. Minor floods should flow undisturbed through the structure.

The first type of open dam was designed with the purpose of retaining the larger stones through a sieving effect. A great variety of shapes and structures have been designed and tested to pursue and optimize this objective.

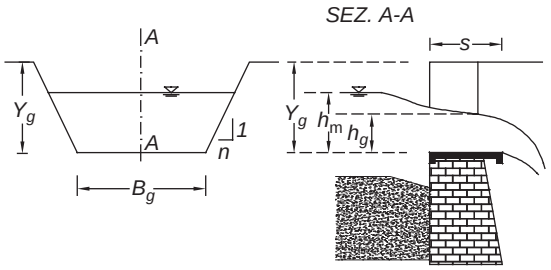


Figure 5 Sketch of the flow over the weir of the check dam

Figure 6 Retention check dam (Provincia Autonoma di Bolzano Alto Adige, 1977) (a) and sequence of consolidation check dams Rio Val de Casa (b) (Provincia Autonoma di Trento, 1991). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Figure 7 Beam dam in Tyrol, Austria. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Basically, this kind of open check dam is composed of wide horizontal openings, which possibly do not narrow the torrent width, guarded by a grid of beams, whose purpose is sieving (*filtering*) sediments and logs: the space among the beams is proportional to the size of the biggest boulders; this check dam was called *grid dam* or *beam dam* (Figure 7).

Although there are no generally accepted rules as to the distance between the beams, as a guiding framework the beam spacing varies between 1.2 and 3 times the maximum grain size being allowed to pass through. In practice, however, the retention effect is strongly reduced by the presence of woody debris: it has been generally observed that beam dams tend to be clogged with logs and other vegetation carried by the current, thus causing complete stoppage even of the finest sediment. For this reason, the distance between the beams in the lower part of the opening is sometimes larger (in order to allow smaller floods to flow undisturbed) than in the upper part (where the filtering effect has to be enhanced).

Once a beam dam is obstructed, it is necessary to remove artificially the logs and large boulders behind the grid by removing the beams or by mechanical excavation of deposited material.

At first, the *slit-dams* were based on similar criteria (Figure 8). This kind of open check dam is provided with only one central opening whose size is proportional to the size of the material to be retained. Also in this case, if the

Figure 8 Slit dam on Rio Inguela, Trentino Italy (Provincia Autonoma di Trento, 2002). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

opening is not large enough, a progressive obstruction tends rapidly to transform the open check dam into a closed one.

It has been recently realized that the retention effect is induced by a hydrodynamic effect rather than through a sieving effect. The cross-section contraction created by the opening inside the dam induced a backwater effect and a hydraulic jump during floods; along this reach the flow velocity is reduced, thus allowing most of the sediments to deposit. The deposited material should be subsequently eroded during minor floods. It is important that the slit allows minor floods to flow undisturbed, with a velocity sufficient to carry the material and possibly to erode the deposit.

A calculation of the optimum opening can be done by imposing the conservation of liquid and solid mass and on the momentum conservation; thus it is possible to find a relation between dam height H_{cd} and opening width b_{cd} (Armanini and Larcher, 2001) (Figure 9).

$$\begin{cases} Q = U_o h_o B_o = U_{cd} h_{cd} b_{cd} \\ Q_s = 8 B_o ((\tau_*)_o - \tau_{*_c})^{1.5} = 8 B_{cd} ((\tau_*)_{cd} - \tau_{*_c})^{1.5} \\ z_o + h_o + \dfrac{Q^2}{2g B_o^2 h_o^2} = h_{cd} + \dfrac{Q^2}{2g B_{cd}^2 h_{cd}^2} \end{cases} \tag{30}$$

Where $(\tau_*)_o$ and $(\tau_*)_{cd}$ are the mobility parameters calculated upstream the dam (section $O - O$ in Figure 9) and inside the slit (Section $Cd - Cd$ in the same figure) respectively. Here, the Meyer-Peter and Müller (1948) equation (20) has been used to calculate the sediment transport rates, but other formulas can be employed. A simplified but working version of this relation is the following:

$$H_{cd} = \frac{Q}{\chi B_o} \left(\tau_{*c} \Delta d_s + \left(\frac{1}{8} \frac{Q_s \Delta}{B_o \sqrt{g}} \right) \right)^{-1/2} \left(\frac{B_o}{B_{cd}} - 1 \right) \tag{31}$$

where χ is Chézy friction coefficient; Q and Q_s are respectively the liquid and the solid design discharges; τ_{*c} is Shields' mobility parameter (see Section "THE INITIATION OF SEDIMENT MOVEMENT"); B_o and b_{cd} are the width of the undisturbed torrent and that of the opening; d_s is the representative size of the bed material.

In comparison with the traditional closed retention dams, the open structures have the undisputed advantages of maintaining the filling space over many years and allowing the flowing through of minor floods. In any case, if necessary, there is the possibility of recreating the deposition space after a filling, if there is access for excavators and trucks. The impact on downstream morphological equilibrium and

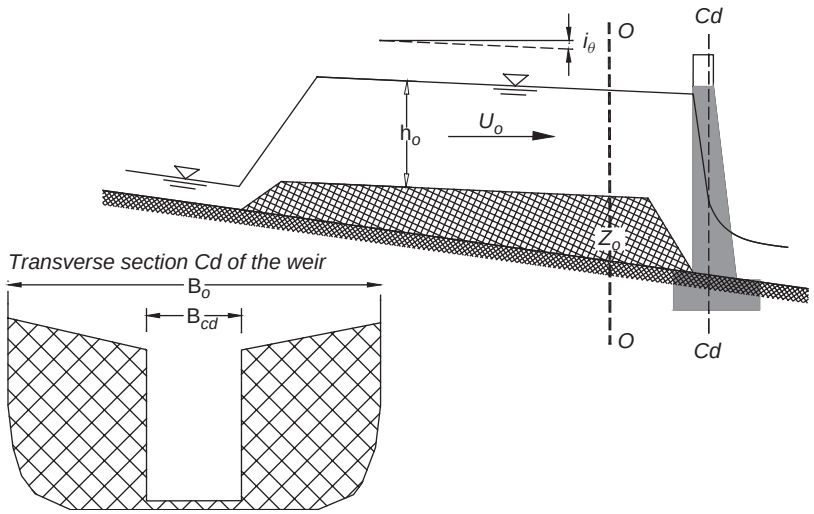


Figure 9 Sketch of the calculation of deposition upstream the silt weir

ecosystems is highly reduced, because a nearly continuous sediment discharge is allowed.

An important problem, that often compromises the efficiency of any device in controlling sediment transport, is the presence of vegetable material such as logs and plants. Special devices, formed by rows of steel ropes taut across the river and inclined toward the bank, have been experimented. In this case the trees are pushed toward the bank, where they can be removed by machines.

REFERENCES

Abrahams A.D., Li G. and Atkinson F.F. (1995) Step-pool: adjustment to maximum flow resistance. *Water Resources Research*, **31**, 2593–2602.

Andrews E.D. (1983) Entrainment of gravel from naturally sorted river-bed material. *Geological Society of America Bulletin*, **94**(10), 1225–1231.

Armanini A. (1999) *Principi Di Idraulica Fluviale*, Bios: Cosenza, p. 152.

Armanini A., Fraccarollo L. and Larcher M. (2005) *Debris Flows, Encyclopedia of Hydrological Sciences*, Jon Whiley & Sons: Chichester.

Armanini A. and Larcher M. (2001) Rational criterion for designing opening of slit-check dam. *Journal of Hydraulic Engineering, ASCE*, **127**(2), 94–104.

Ashida K. and Michiue M. (1971) Sediment Transport Rate and Bed Transportation. *The Disaster Prevention Laboratory of the Kyoto University*, Annual Report, 14-B.

Ashida K. and Michiue M. (1973) Studies on bedload transport rate in open channel flows, *Proceedings of the International Symposium on River Mechanism*, Bangkok, Vol. 1, pp. 407–418.

Ashworth P.J. and Ferguson R.I. (1989) Size-selected entrainment of bed load in gravel streams. *Water Resources Research*, **25**(4), 627–634.

Bathurst J.C. (1978) Flow resistance of large-scale roughness. *Journal of Hydraulics Division, ASCE*, **104**(HY12), 1578–1603.

Bathurst J.C. (1982) Flow resistance in boulders-bed streams. In *Gravel-bed Rivers*, Hei R.D., Bathurst J.C. and Thorne C.R. (Eds.), Wiley & Sons: Chichester, pp. 443–462.

Bathurst J.C. (2002) At-a-site variation and minimum flow resistance for mountain rivers. *Journal of Hydrology*, **269**, 11–26.

Bathurst J.C., Graf W.H. and Cao H.H. (1987) Bed load discharge equations for steep mountain streams. In *Sediment Transport in Gravel Bed Rivers*, Thorne C.R., Bathurst J.C. and Hey R.D. (Eds.), John Wiley and Sons: Chichester, pp. 453–477.

Bray D.I. (1979) Estimating average velocity in gravel-bed rivers. *Journal of Hydraulics Division, ASCE*, **105**(HY9), 1103–1122.

Bray D.I. and Davar K.S. (1987) Resistance to flow in gravel-bed rivers. *Canadian Journal of Civil Engineering*, **14**(2), 77–86.

Christensen B.A. (1996) Discussion of the paper by Chiew, YM and G Parker 1995, Incipient sediment motion on non-horizontal slopes. *Journal of Hydraulic Research, IAHR*, **32, 33**(5), 725–730.

Diplas P. (1987) Bed load transport in gravel bed streams. *Journal of Hydraulic Engineering, ASCE*, **113**(3), 277–292.

Egiazaroff I.V. (1965) Calculation of non-uniform sediment concentrations. *Journal of Hydraulics Division, ASCE*, **91**(HY4), 225–246.

Einstein H.A. (1950) *The Bed-load Function for Sediment Transportation in Open Channel Flow*, Technical Bulletin No. 1026, U.S. Department of Army, Soil Conservation Service, Washington.

Ferguson R.I., Prestegaard K.L. and Ashworth P.J. (1989) Influence of sand on hydraulics and gravel transport in braided gravel bed rivers. *Water Resources Research*, **25**(4), 635–643.

Graf W.H. (1984) *Hydraulics of Sediment Transport*, Water Resources Publications: Littleton.

Graf W.H. (1989) Flow resistance over gravel bed: its consequence on initial sediment movement. In *Fluvial Hydraulics of Mountain Regions*, Armanini A. and Di Silvio G.D. (Eds.), Lecture Notes on Earth Sciences, Vol. 37, Springer-Verlag, pp. 17–32.

Graf W.H., Cao H.H. and Suszka L. (1983) Hydraulics of steep, mobile-bed channels, *Proceedings of the XXI IAHR Congress*, Moscow, Vol. 7, pp. 301–305.

Graf W.H. and Suszka L. (1987) Sediment transport in steep channels. *Journal of Hydroscience and Hydraulic Engineering*, **5**(1), 11–26.

Griffiths G.A. (1981) Flow resistance in coarse gravel-bed rivers. *Journal of Hydraulics Division, ASCE*, **107**(HY7), 899–916.

Günter A. (1979) Die kritische mittlere sohlenschubspannung bei geschiebemischungen unter berücksichtingung der deckschichtbildung und der turbulenzdedingungten sohlschubspannungsschwankungen. *Mitt. Nr. 3 der Versuchsanstalt Für Wasserbau, Hydr. und Glaz.*, ETH: Zürich.

Hey R.D. (1979) Flow Resistance in Gravel-Bed Rivers. *Journal of Hydraulics Division, ASCE*, **105**(HY4), 365–379.

Lee J.A. and Ferguson I.R. (2002) Velocity and flow resistance in step-pool streams. *Geomorphology*, **46**, 59–71.

Leopold L.B. and Langbeim W.B. (1962) *The Concept of Entropy in Landscape Evolution*, U.S. Geological Survey: Professional Papers 500-A.

Leopold L.B. and Wolman M.G. (1957) River channel patterns: Braided, meandering and straight. *U.S.G.S. Prof. Paper 282-B*, U.S. Gov. Print. Off., Washington D.C., pp. 85.

Limerinos J.T. (1970) *Determination of the Manning Coefficient for Measured Bed Roughness in Natural Channels*, Water Supply Paper 1898-B, U.S. Geological Survey: Washington.

Meyer-Peter E. and Müller R. (1948) Formulas for bed-load transport, *Proceedings of the 2nd Meeting on IAHSR*, Stockholm, pp. 1–26.

Montgomery D.R. and Buffington J.M. (1997) Channel-reach morphology in mountain drainage basin. *Geological Society of America Bulletin*, **109**(5), 596–611.

Nakagawa H., Tsujimoto T. and Shimizu Y. (1989) Turbulent flow with small relative submergence. In *Fluvial Hydraulics of Mountain Regions*, Armanini A. and Di Silvio G.D. (Eds.), Lecture Notes on Earth Sciences, Vol. 37, Springer-Verlag, pp. 33–44.

Nikora V., Derek G., McEwan I. and Griffiths G. (2001) Spatially averaged open-channel flow over rough bed. *Journal of Hydraulic Engineering, ASCE*, **127**(2), 123–133.

Parker G. (1989) Surface-based bedload transport relation for gravel rivers. *Journal of Hydraulic Research*, **28**(4), 417–437.

Parker G., Klingeman P.C. and McLean D.G. (1982) Bed-load and size distribution in paved gravel-bed streams. *Journal of Hydraulics Division, ASCE*, **18**(5), 1409–1423.

Parker G. and Sutherland A.J. (1990) Fluvial Armor. *Journal of Hydraulic Research*, **28**(5), 529–544.

Provincia Autonoma di Bolzano Alto Adige (1977) Dai crinali al fondovalle, *Periodico di Informazione del Consiglio e della Giunta Provinciali*, V, II,10.

Provincia Autonoma di Trento (1991) *Per Una Difesa Del Territorio*, Edizioni Arco Trento.

Provincia Autonoma di Trento (2002) **http://www.sistemazionemontana.provincia.tn.it**

Rickenmann D. (1991) Hyperconcentrated flow and sediment transport at steep slopes. *Journal of Hydraulic. Engineering, ASCE*, **117**(11), 1419–1439.

Rickenmann D. (2001) Comparison of bed load transport in torrent and gravel bed streams. *Water Resources Research*, **37**(12), 3295–3305.

Ryan S.E., Porth L.S. and Troendle C.A. (2002) Defining phases of bedload transport using piecewise regression. *Earth Surface Processes and Landforms*, **26**, 971–990.

Schoklitsch A. (1962) *Handbuch des Wasserbaues*, *Third Edition*, Springer-Verlag: Vienna.

Shields A. (1936) *Anwendung der Ähnlichkeitsmechanik und der Turbulenzforschung auf die Geschiebebewegung*, Vol. 26, Mitteil, PVWES: Berlin.

Smart G.M. and Jäggi M.N.R. (1983) Sediment transport in steep slopes. *Mitteilung der Versuchsanstalt für Wasserbau, Hydrologie und Glaziologie*, ETH: Zürich.

Strickler A. (1923) *Beiträge zur Frage der Geschwindigkeitsformel und der Rauhigkszahlen für Ströme, Kanäle und geschlossene Leitungen*, Mitt. des Am. für Wasserwirtschaft, Nr. 16.

Thompson S.M. and Campbell P.L. (1979) Hydraulics of a large channel paved with boulders. *Journal of Hydraulic Research*, **17**(4), 341–354.

Wohl E.E. (2000) *Mountain Rivers*, Water Resources Monograph 14, American Geophysical Union: Washington, p. 320.

Wohl E.E. and Grodek T. (1994) Channel bed-steps along Hahal Yael, Negev desert, Israel. *Geomorphology*, **9**, 117–126.

Whittaker J.G. and Jäggi M.N.R. (1982) Origin pd step-pool systems in mountain streams. *Journal of Hydraulics Division, ASCE*, **108**, 758–773.

Whittaker J.G. and Jäggi M.N.R. (1986) Blockschwellen. *Mitteilung der Versuchsanstalt für Wasserbau, Hydrologie und Glaziologie*, Vol. 91, ETH: Zürich.

144: Regulated Lowland Rivers

HUIB J DE VRIEND

WL Delft Hydraulics, Delft, The Netherlands

The article describes flow, sediment transport, and morphological processes in regulated lowland rivers and the impact of human activities thereon. It gives some elementary examples of the morphological response to deviations from a prismatic straight channel, in an attempt to bridge part of the gap between textbook knowledge and reality. Moreover, it shows that even a regulated river is still to some extent natural, thus variable and unpredictable.

INTRODUCTION

Ever since man has started using rivers for his purposes, he has attempted bringing these rivers under his control. River regulation has now become a common practice all over the world (for example, see Figure 1). A wide variety of enabling structures, ranging from large regulation works to simple groynes and bank protections, has been invented and put into place. Each of these structures has its own specific impact on the river's behavior, an impact we would better know on beforehand, if we prefer to avoid unpleasant and often costly surprises. Therefore, a separate article is dedicated to the phenomena that are typical of regulated lowland rivers.

Figure 1 The Lower Rhine, an example of a regulated lowland river (flow from bottom right to top left). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

UPPER, MIDDLE, AND LOWER REACH

The profile of a river can roughly be divided into: a relatively steep upper reach, often in mountainous terrain and with high flow velocities, a more gently sloping middle reach with gradually finer sediments and inflow from tributaries, and a mildly sloping lower reach with relatively low flow velocities and fine sediments. River regulation in the upper reach concerns either discharge regulation by high dams or sediment retention by check dams. In the middle reach, which often lies in a wide valley, water levels may be regulated by weirs and navigation locks, and the banks protected by riprap or other material. In the lower reach, the river is surrounded by land that is mostly built up from river sediment left behind after migration of the river channel or deposited during inundation events. This part is often aligned by dikes or levees, in order to prevent inundation of riparian zones. As a consequence, there is no more sediment deposition in the riparian zones, whereas the river keeps on building out its delta thus shifting its own downstream boundary. Since the slope remains the same, the bed at a fixed point along the river tends to build up. This may continue until the river bed lies well above the surrounding land. In this situation, the river gets hydrologically disconnected from the surrounding land.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Figure 2 Drainage basin of the Yellow River. Note that the lower reach (about 700 km long) is hydrologically almost disconnected from the surrounding land. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Parts of the Lower Yellow River, for instance, lie more than 15 m above the surrounding land (also see Figure 2). It is not surprising, therefore, that the Chinese call this reach the *suspended river*.

FUNCTIONS AND HUMAN INTERVENTIONS

Lowland rivers in densely populated areas usually have four main functions: (i) safe conveyance of floods, sediment, and ice; (ii) fresh water supply; (iii) navigation; and (iv) discharge of (treated) wastewater. Apart from these four, there can be various other functions such as agriculture, recreation, or cooling water supply. Pressure on space and the demand for safety against flooding have led people to embanking rivers and fixing their course. Navigability requirements led them to reducing the width, so as to have sufficient depth under a wide range of conditions. Width-reduction also helps to increase the flow velocity, thus reducing any probability of ice jamming. Dams are not only used for power generation, or storage for irrigation, but also for discharge regulation (e.g. the Three Gorges Dam in the Yangtze River, China) and sometimes even for sediment-transport regulation further downstream (e.g. the Xiaolangdi Dam on the Yellow River, China; see Figure 3). Barrages and weirs can be used to set up the water level in times of low discharge so as to improve navigability, but also to regulate the discharge distribution between river branches.

Figure 3 Combined release of clean (white) and heavily sediment-laden (dark) water through the Xiaolangdi Dam, Yellow River, China. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

IMPACTS AND RESPONSES

These human interferences with nature have their consequences for the river's behavior. Dams and barrages are barriers to the sediment that tends to pile up in the reservoir upstream, whereas the flow downstream of the structure is undersaturated with sediment and tends to erode the bed. Bend cutoffs, which reduce the river length, can easily be shown to lead to (limited) river bed erosion in the entire upstream reach. According to Lane's balance (Lane, 1955), a reduction of the main channel width produces not only a larger depth, but also a smaller longitudinal slope. Since the normalization of the Rhine branches in the first half of the twentieth century, the bed of the Lower Rhine near the Dutch–German border has come down by more than a meter and the erosion process still continues (Visser *et al.*, 1999). Floodplain encroachments for housing, industry, or other activities that reduce the space for flood water storage, leads to changes in the flood conveyance properties of the river. Urban bottlenecks hamper the passage of flood waves and set up the water level over some distance upstream. Nature development in originally open floodplains (e.g. pasture land) will lead to vegetation growth, which may reduce the flood conveyance capacity. As a consequence, compensating measures are needed, or the vegetation has to be removed from time to time (floodplain rejuvenation; e.g. van der Lee *et al.*, 2001).

River regulation generally tends to evoke further regulation, as it increases safety and facilitates the use of the riparian zones for economic activities, thus attracting more people and more economic activities, which tend to demand more from the river. Although heavily influenced by man, regulated rivers are still to a certain extent natural in that they often experience strong discharge variations and have a mobile bed consisting of natural sediments. At the basin scale, the overall channel pattern is often natural at large, with artifacts here and there (e.g. the Pannerdensch Kanaal in the Rhine branches, or the Atchafalaya diversion in the Mississippi; see Figure 4). Water and sediment inputs into the river are not constant in time, not even in a statistical sense. They depend on the climate and on the land use in the river basin. The downstream boundary is often a shelf sea, or the ocean, with all its inherent water level variations (tides, storm surges, sea level rise). Clearly, real-life rivers seldom resemble the highly schematic situations that are usually considered in simplified models. There are always complications requiring special attention.

Figure 4 About one-quarter of the Mississippi River discharge (at the top) is diverted to the Atchafalaya River via three man-made channels with regulation works. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

PROCESSES AT THE BASIN SCALE

The river basin is the area that is drained by the river. On almost every point on land one can say by which river it is drained. This explains how in the north of France there can be roadmarks indicating the exact location of the drainage divide between the Atlantic and the Mediterranean basins.

The rainfall–runoff relationship is a typical property of a river (sub)basin, indicating how rainfall patterns in the basin translate into discharge hydrographs in the streams and rivers forming the drainage network of that basin. Land use (agriculture, forestry, etc.) and land cover (forests, pasture land, cities) determine the specific properties of this relationship, but are also subject to continuous changes due to changing human activities and climate change. Rainfall-runoff relationships are therefore not necessarily time-invariant (*see* **Part 10: Rainfall-runoff Processes** of this encyclopedia).

On the other hand, it is very difficult, if not impossible, to evaluate such a relationship on the basis of elementary data, such as soil composition, vegetation type, or topographic maps. Empirical methods are therefore inevitable, but they require homogeneous data over a long period of time, that is, data of which the statistical properties do not change through that period. In areas with much human activity, such as densely populated regions, such homogeneous data do not exist, as things keep on changing all the time. In such situations, rainfall–runoff relationships have to be estimated (*see* **Part 11: Rainfall-runoff Modeling** of this encyclopedia).

Another important (sub)basin process is the sediment yield. Sediment is produced by the weathering of rock, by

Figure 5 The Loess Plateau, the vast erosion area where the Yellow River picks up its huge sediment load (flow from top to bottom). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the erosion of loose deposits, or as a residue of organic processes (cf. Figure 5). The production rate depends on geological structure, terrain properties, climate, land cover, land use, and so on. Well-known examples of human influence are landslides, mudflows and erosion due to deforestation, sediment retention by check dams, and sediment blocking by large river dams. Since the sediment yield of the basin determines the sediment input into the river, it also has its impact on river morphology (*see* **Part 7: Erosion and Sedimentation** of this encyclopedia).

FLOOD WAVES

Human interventions in the river bed and the floodplains change the flood conveyance properties of a river. In general, the propagation speed of a flood wave increases if the flow velocity in the main channel increases and if the storage capacity of the floodplain decreases. At the same time, the peak height of the flood wave increases if this capacity decreases. Floodplain encroachment occurs at the expense of storage capacity and is therefore detrimental to a river's flood conveyance capacity. In the Middle Rhine, for instance, many decades of ever further confinement of the river have made flood waves propagate much faster (i.e. less warning time), with much higher peak discharges (i.e. higher maximum water levels). Figure 6 shows the results of a numerical model exercise for the same flood wave in the Middle Rhine with the geometry of 1882/1883 and that of 1995. The changes in geometry reduce the lead time

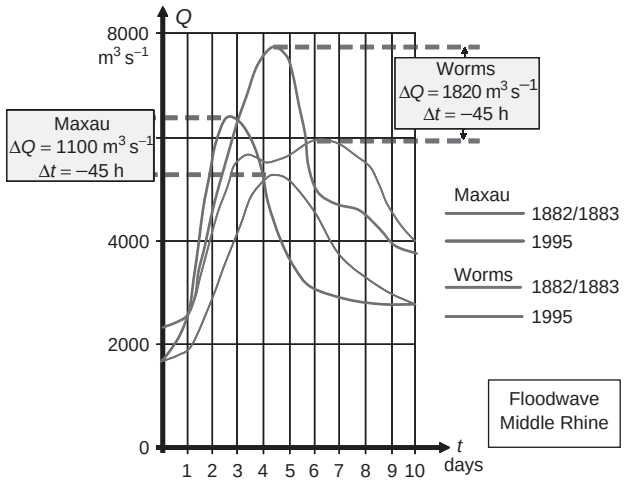


Figure 6 The same floodwave in the Middle Rhine at Maxau and Worms, computed with the geometry of 1882/1883 and that of 1995 (after: KHR, 1993). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

to peak passage by about two days and increase the peak discharge by some 30%.

People have begun to realize that this process may have gone too far. Many present-day river improvement projects, therefore, focus on creating more space for the river in flood situations. Examples of space-creating measures are channel widening, floodplain lowering, digging side channels, building flood bypasses, removing summer obstacles, setting back dikes, and so on (for instance, see Silva *et al.*, 2001).

ROUGHNESS AND FLOW RESISTANCE

An insufficiently known aspect of discharge variations is the so-called *dynamic roughness*, that is the variable bed-form roughness associated with varying discharge. The bed-form dimensions (e.g. Figure 7) exhibit a retarded response to the flow velocity, whence they cannot be related to the instantaneous discharge. Instead, a relaxation-type bed roughness submodel should be included in models predicting flood levels (Wilbers and ten Brinke, 2003). The relationship between bed-form dimensions, bed roughness, and flow conditions is subject to further research at various places around the world.

We do not know to what extent the bed-form changes observed under extreme conditions can be extrapolated to design conditions, which are sometimes far beyond the highest flood on record. In the case of the Dutch Rhine branches, for instance, the design discharge is $16\,000\,\mathrm{m^3\ s^{-1}}$, whereas the highest discharge on record is $12\,600\,\mathrm{m^3\ s^{-1}}$, observed in 1926. Clearly, this introduces

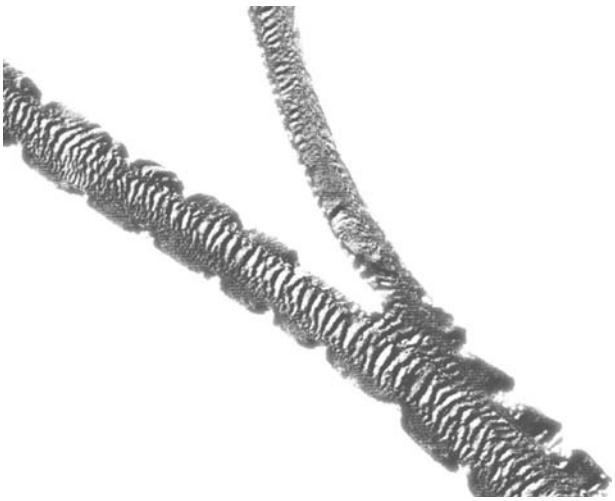


Figure 7 Bed-form pattern around the Pannerdensche Kop bifurcation in the Rhine, as observed under flood conditions with multibeam sonar. Typical wave length: 50 m (Reproduced from Ten Brinke, W., (2004) by permission of Veen Magazines)

significant uncertainties in the prediction of the design water levels.

TRANSPORT OF SUSPENDED AND DISSOLVED MATTER

Rivers transport not only sediment but also many other types of particulate and dissolved matter, often originating from human activities in riparian zones. This includes industrial waste, which used to be discharged untreated into the river until the last few decades since when water quality regulations became stricter. As a consequence, much of the fine sediment deposited in the floodplains is polluted with heavy metals, phosphates, nitrates, and so on. Handling this heavily polluted material sometimes constitutes a major problem to large-scale river restoration schemes (*see* **Part 8: Water Quality and Biogeochemistry** of this encyclopedia).

A particular problem of groyne-regulated rivers is the complexity of the pollutant pathways. This plays a special role in the case of calamitous pollution events, when an amount of pollutant is discharged into the river at a certain location and at a certain point in time. This pollutant is transported down the river and will at a certain moment reach water intakes and other sensitive points. The accurate prediction of that moment and the duration of the dangerous situation can be of major societal importance, especially when health risks are involved. Modeling this type of phenomenon in groyne-regulated rivers is particularly difficult, because part of the material lingers behind in the groyne fields and reenters the main channel at a later stage (van Mazijk, 1996). In 1-D model terms, this means a strong dispersion effect, which does not affect the maximum speed of the pollutant cloud, but has a major effect on the length of its tail, and hence on the duration of the dangerous period at any point along the river (*see* **Chapter 100, Water Quality Modeling, Volume 3**).

SEDIMENT TRANSPORT

The general mechanisms of sediment transport in regulated rivers are not different from those in natural rivers or mobile-bed irrigation channels. The shear stress exerted by the flow on the bed brings the sediment into motion if a certain threshold value is exceeded. Clearly, this threshold depends on the grain size and the sediment density. Once this threshold is exceeded, transport may take place in different regimes: in a thin sheet along the bed, via migrating bedforms (ripples, dunes), rolling downhill while being transported downstream, avalanching down steep slopes, or suspended in the water column. In the latter case, the turbulent water motion delivers the power needed to keep the sediment grains in suspension. Irrespective of the regime, the transport rate is a nonlinear function of the shear stress (for further details see Graf, 1998; *see* **Chapter 140, Transport of Sediments, Volume 4**).

This means that the sediment transport rate increases along with the flow velocity and the bed roughness. In flood conditions, both are large, so large amounts of sediment are set into motion. The probability of branches getting choked with sediment during an extreme flood event is therefore not hypothetical, although we are still not able to quantify it. We do know, however, that in heavily sediment-laden rivers, such as the Lower Yellow River, flood levels cannot be computed without taking due account of the morphological changes during the event (e.g. Kemink *et al.*, 2003).

Another reference is that the total amount of sand deposited in the floodplains of the Rhine branches during the 1995 flood event is estimated at 300 000 m^3 (ten Brinke, 2004), which is of the same order of magnitude as the river's yearly sediment load (*see* **Chapter 84, Floodplain Sedimentation – Methods, Patterns, and Processes: A Review with Examples from the Lower Rhine, the Netherlands, Volume 2**).

If floodplains are lowered as part of a scheme to increase the flood conveyance capacity of a river without dike strengthening, their sediment-trapping capacity is likely to increase. If, in addition, the summer levees bordering the main channel are removed, the inundation frequency of the floodplains will increase, and therewith the amount of sediment deposited there. At some distance from the main channel, this sediment will mainly be fine fractions, such as clay, but closer by, there will also be deposition of sand (cf. Figure 8). In the dry season, this sediment is available for aeolian processes such as dune formation. Given the above quantities of sand deposited in the floodplains during a large

Figure 8 Sand deposited in a floodplain of the Waal River during a flood event. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

flood, this must influence the overall sediment balance of the river.

A specific aspect of groyne-regulated rivers is the sediment balance of the groyne fields. Altogether they store a large amount of sediment and if they all trap or release part of that sediment at the same time, this must influence the main channel. Investigations so far (e.g. Yossef and de Vriend, 2004) indicate that groyne fields tend to import sediment under low-flow conditions via the recirculation zones between the groynes, as well as under high-flow conditions via secondary currents induced by the groynes. This suggests that they would accrete consistently until they have reached some equilibrium state where the net import is zero. In reality, however, it turns out that navigation-induced water motion plays a key role in the sediment balance. Especially, ship-induced water level set-down turns out to be effective in drawing sediment out of the groyne field (ten Brinke, 2004). Consequently, the sediment content of the groyne fields tends towards a nontrivial dynamic equilibrium state, which also depends on the type and the intensity of the navigation (see Figure 9).

GRADED SEDIMENT

Sediment in real-life rivers is hardly ever uniform in grain size. As sediments of different grain sizes respond differently to hydrodynamic and gravitational forcing, a variety of mechanisms tends to segregate coarse and fine sediment.

The dynamic interaction between hydrodynamics, sediment transport, bedforms, and bed roughness is further complicated if the bed sediment is nonuniform in grain size. Bedforms migrate by erosion at the upstream side and

Figure 9 The morphological state of groyne fields is influenced by navigation (flow from top to bottom). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

deposition of the eroded sediment at the downstream side. The latter is usually so steep that deposited grains tend to slide down the slope. This so-called *slip-face* is a very good grain-size selector, as every building contractor knows. The coarsest grains tend to roll further down the slope than the finer ones, so that the sediment composition in the troughs becomes relatively coarse. Thus, the bedforms tend to form a coarse layer at their trough level while migrating down the river (Blom *et al.*, 2003; also see Figure 10). When during an extreme flood the bedforms are high, the troughs are deep, and the coarse layer is formed at a level that is usually not reached under less extreme conditions. The coarse layer formed in the Lower Rhine during the 1995 flood, for instance, is still visible in boxcores taken today.

Once this coarse sublayer has formed, it keeps the troughs from digging deeper into the bed and thus limits the bedform height. This constitutes a self-regulatory feedback system.

Another mechanism is selective transport. Under moderate flow conditions, the coarsest sediment fractions do not move, whereas the finer ones do. Thus, it is possible that in an erosion area the coarser sediment lags behind and forms a so-called *armor layer* on top of finer sediment, thus protecting the latter from being eroded. This phenomenon is often found downstream of sediment-blocking structures, such as dams, barriers, or weirs, or in the scour holes near bridge piers, groyne heads, and so on. The flow there is undersaturated with sediment and therefore tends to pick

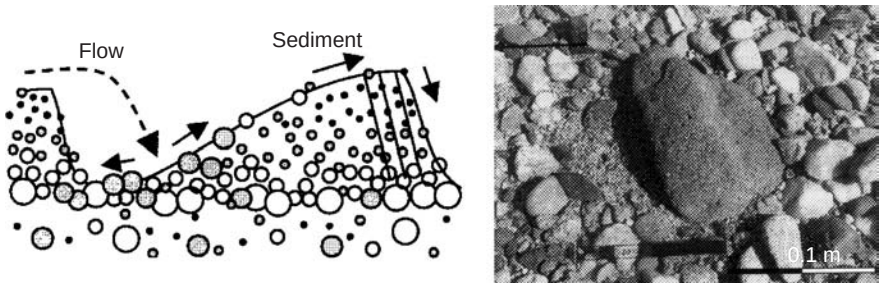


Figure 10 Interaction of water and sediment motion causes sorting of graded sediment. (a) vertical sorting by bedforms (Blom *et al.,* 2003). (b) a coarse armor layer on top of a finer substrate

up sediment from the bed. After some time, the formation of the armor layer keeps the bed from further eroding, thus constituting yet another self-regulatory feedback.

At a larger scale, the bed composition may vary in space and time, under the influence of variations in sediment supply, discharge variations, nonuniformities in the river geometry, bifurcations, confluences, and so on. This means that the bed composition waves migrate more or less continuously through the river (e.g. Ribberink, 1987). This, in its turn, influences the sediment transport, the bed roughness, and the flow. Thus, the bed composition is in continuous interaction with the other elements of the morphodynamic system.

Because of gravitation, sediment is transported more easily downhill than uphill. This downhill gravitational transport component works out differently for grains of different size. This means that in situations with a cross-stream bed slope, like in a bend, or near a bifurcation or a confluence, there is a cross-stream segregation by grain size. Coarser sediment generally stays near the foot of the slope, whereas, the fines are carried uphill. Since sediment transport is quite dispersive, this segregation usually gets undone after a bend or a confluence, but at bifurcations it may strongly influence the grain size of the sediment entering the branches. This so-called *Bulle-effect* (Bulle, 1926; also see Jansen, 1979) depends on the curvature-induced secondary flow, hence, on the shape of the bifurcation.

LARGE-SCALE MORPHOLOGICAL EQUILIBRIUM STATE

A regulated lowland river of which the alignment and the width are fixed has two degrees of freedom left to accommodate the supply of water and sediment from upstream; that is, its water depth, h, and its slope, S. If the discharge, Q, and the sediment input, Q_s, as well as the downstream water level are constant in time, a static equilibrium state would develop. If we assume a rectangular cross section of constant width B, and a power-law relationship between the transport rate per unit channel width and the mean flow velocity, this state can be derived from first physical principles and is given by

$$h_{eq} = \left(\frac{Q_s}{a_s B}\right)^{-1/b_s} \frac{Q}{B}; S_{eq} = \left(\frac{Q_s}{a_s B}\right)^{3/b_s} \frac{Bf}{8gQ} \qquad (1)$$

in which a_s is the coefficient of proportionality in the transport formula, b_s is the exponent of the velocity in that formula, f is the Darcy Weisbach friction factor, and g is the acceleration due to gravity.

Typical values of b_s are between 3 and 5, and a_s can be shown to be inversely proportional to some representative grain size. If the sediment supply is assumed independent of the discharge, the above formulae show that the product of equilibrium, depth, and slope is independent of the discharge. So if the depth increases, the slope decreases. This is exactly what has happened in the Rhine branches after the normalization in the late nineteenth and early twentieth centuries (Visser *et al.*, 1999).

The formulae also show the equilibrium slope to decrease with the discharge and increase with the grain size. This corresponds qualitatively with the general observation that the slope of a river tends to decrease from upstream (low discharge, coarse sediment) to downstream (high discharge, fine sediment). In reality, sediment supply and discharge are coupled to some extent, if only because the inflow from the tributaries also brings in sediment. In general, the transport rate will be an increasing function of the discharge. This explains why the slope in river channels is found to be less than inversely proportional to the bankfull discharge (gravel-bed channels: $S \propto Q^{-0.4}$, sand bed channels: $S \propto Q^{-0.3}$; e.g. Knighton, 1998).

The static equilibrium relationships (1) can be used to construct the equilibrium bed topography in a river. Starting from the given downstream water level, one can construct the water surface, which is a continuous broken line consisting of piecewise straight lines under slope S_{eq}.

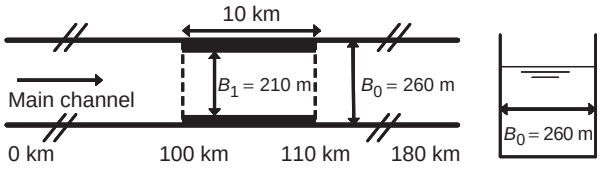


Figure 11 Hypothetical case of a prismatic channel with a long constriction

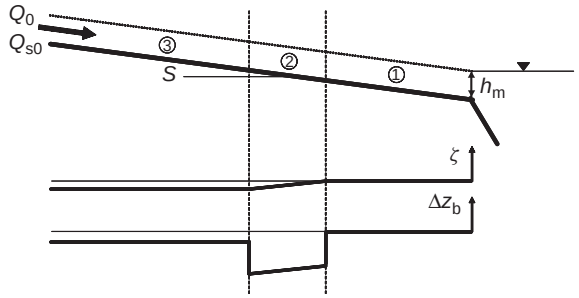


Figure 12 Static equilibrium state for this case. Section 2 is the constricted reach. The top panel shows the initial state, the lower panels show the changes of the water level, $\Delta\zeta$, and the bed level, Δz_b

Subsequently, one can draw the equilibrium bed level at a distance h_{eq} below and parallel to this water surface.

Figures (11) and (12) show how this works out for a large-scale human intervention, in this hypothetical case a 10-km long constriction assumed to be created suddenly in an otherwise prismatic channel with a constant supply of water and sediment and a mobile bed in equilibrium. According to equation (1), the equilibrium water depth will be larger in the constricted reach, and the water surface slope, as well as the bed slope, will be smaller. Consequently, the bed level and the water level in the whole reach upstream of the constriction will come down slightly.

MORPHODYNAMICS

Static equilibrium is a useful theoretical concept that helps understanding river morphology, but it has little to do with reality in most rivers in which such an equilibrium state is never reached. How close the bed topography gets to the equilibrium state depends on the ratio between the rate at which the system tends towards this equilibrium (the morphological time scale) and the characteristic time scale of the natural (e.g. seasonal) variation of conditions. In most rivers, this ratio is such that time-dependent morphological processes are predominant.

Time-dependent morphological processes are rather complicated, because of the dynamic interaction of water motion, sediment transport, and bed level changes. Yet, some of this morphodynamic behavior can be understood by considering two limit cases, one for length scales that are small compared to the length scale involved in the backwater curve (typically $h/3S$) and one for length scales that are large compared to this backwater scale. In the former case, bed perturbations behave like nonlinear kinematic waves in that they propagate downstream at a speed $c_b = b_s Q_s/(Bh)$ and in the meantime undergo a deformation due to the faster propagation of the higher parts. Thus, upstream-facing slopes tend to flatten out and downstream facing ones tend to develop into a shock front (see Jansen, 1979).

The large-scale limit behavior is essentially different and can be shown to have a diffusive character (e.g. Ribberink and van der Sande, 1987). This means that a perturbation at this scale tends to spread out in both directions (upstream and downstream) and to decrease in amplitude at an ever slower rate during the process. This provides a mechanism for morphological effects at large distances upstream of an intervention.

These two effects are clearly visible in the animation of Figure 13. It concerns a numerical model simulation for the above hypothetical case. The animation shows the time evolution of the bed profile starting from the equilibrium situation before the constriction was applied. In the beginning, it shows how the bed erodes very rapidly in the constricted reach, while the erosion products are deposited just downstream of that reach. The hump of sand formed in this way starts expanding in the downstream direction, and when the supply from the eroding constricted

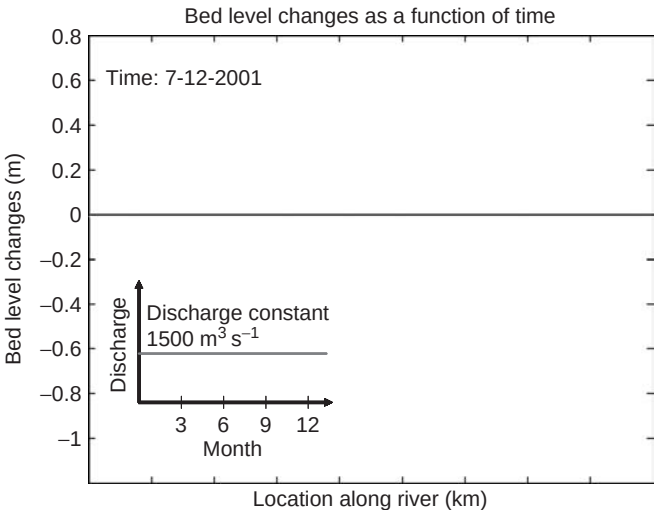


Figure 13 Animation of a morphodynamic simulation for the case shown in Figure 11, with constant discharge and sediment supply (animation goes with this article; courtesy Saskia van Vuren; also see **http://www.sobek.nl** for a free trial copy of the software). A color version of the animation is available at http://www.mrw.interscience.wiley.com/ehs

zone drops off, it migrates further downstream until it has moved out of the model area. Subsequently, there is a much slower large-scale adjustment to the boundary conditions, with the theoretical equilibrium state as a final result.

MORPHOLOGY AND TIME-VARYING INPUTS

Discharge variations inevitably influence the morphological behavior of a river. For example, in Figure 11, the rate of erosion in the constricted reach will be a function of the discharge. Hence, the sediment supply to the downstream reach will depend on the discharge. Discharge variations will therefore cause the formation of bed waves, which will from thereon propagate downstream and ultimately leave the model domain, as shown in the animation of Figure14. The bed waves are of practical importance, as they may hamper navigation, or jeopardize the stability of structures.

Bedwave formation due to discharge variation will generally be induced by geometrical nonuniformities in the river geometry. This may also concern the floodplain geometry, as is shown in the example in Figure 15 for a hypothetical case of floodplain lowering. At every nonuniformity, every discharge variation will yield a morphological response in the form of erosion or accretion. This causes a surplus or a deficit in the sediment supply to the reach downstream of the nonuniformity, forming a hump on, or a hole in, the bed there. These bed disturbances start traveling downstream through the river and may give rise to problems with navigability or structure stability. Recent investigations for the Waal River have shown that their effect on the flood

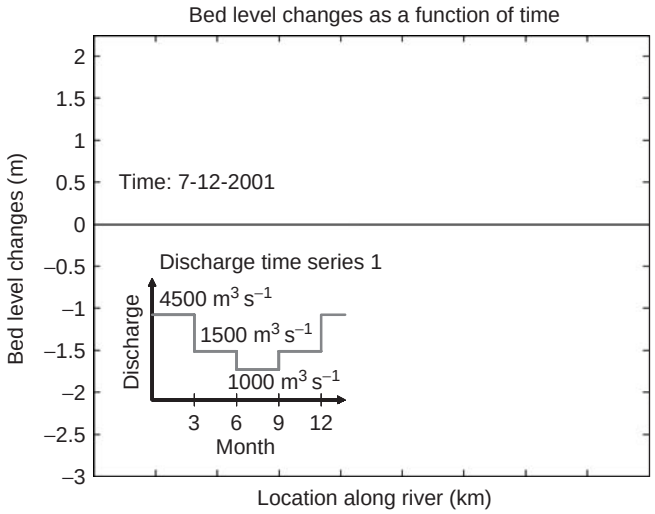


Figure 14 Morphodynamic simulation for the case shown in Figure 11, with a stepwise varying discharge and sediment supply (animation goes with this article; courtesy Saskia van Vuren; also see **http://www.sobek.nl** for a free trail copy of the software). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

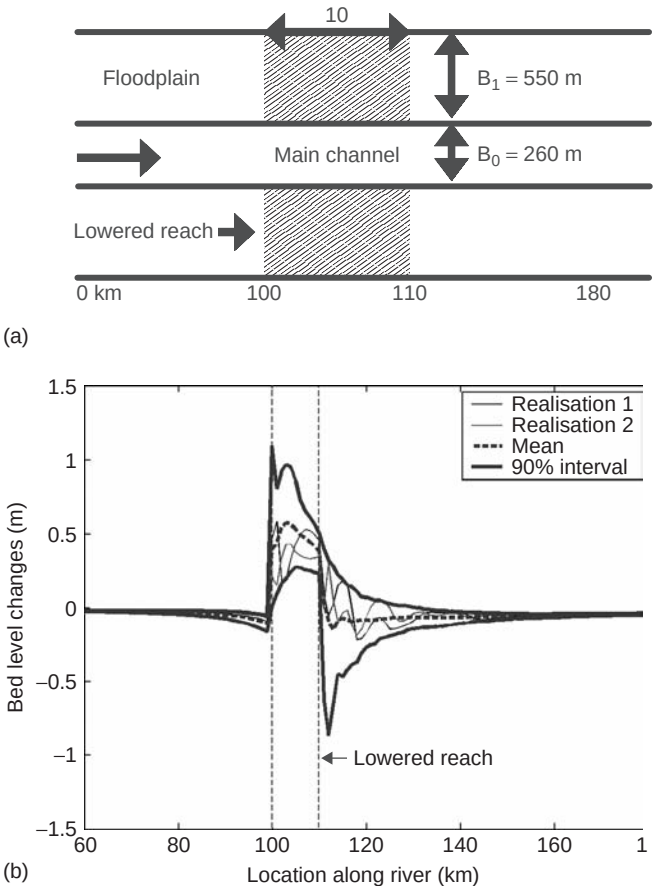


Figure 15 Morphological effect of a geometrical nonuniformity (hypothetical floodplain lowering) in combination with discharge variations. (a) situation. (b) bed-level changes in the main channel for two different hydrographs, and the mean and the 90% confidence interval for a large number of such hydrographs (courtesy Hanneke van der Klis). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

conveyance capacity of this river is minor (van Vuren *et al.*, 2003), except near bifurcations.

Figure 15 also shows that the response is different for every discharge hydrograph. Monte Carlo Simulation with a large number of statistically equivalent hydrographs provides information on the statistical properties of the bed-level variations (see van der Klis, 2003). In the figure, the mean and the 90% bandwidth are indicated. These are useful indicators for the design of structures, or making decisions whether or not to dredge the navigation channel.

Clearly, the bed waves will not be reproduced by any model with a constant discharge. The mean bed-level change due to the floodplain lowering, however, looks rather similar to what would be found in the case of a constant discharge (see Figure 12). Hence one might think that it must be possible to find a constant discharge that yields a good estimate of this mean response. This

will not be the mean discharge, since the water and sediment mechanics underlying the morphological changes are strongly nonlinear. Hence, substituting the average discharge into a transport–discharge relationship will not yield the average transport rate. The discharge that does produce the right average transport rate, called the *dominant discharge*, is generally well above average. Since the sediment balance equation is linear, the right transport field must give the right short-term bed-level changes. Yet, the dominant discharge approach generally does not produce the right equilibrium state, because the nonlinear feedback between the bed-level changes and the water and sediment motion will exert its influence in the longer run. It can be shown from the basic equations for uniform one-dimensional straight channel flow with a mobile bed in equilibrium that the key morphological properties, that is, the bed slope and the water depth in the river mouth, are determined by different statistical properties of the discharge, and that each of these properties is different from the mean discharge and the dominant discharge. In other words, even a simple property as the equilibrium state in a straight channel of uniform width cannot be described with a single representative discharge. For further details, see the discussion on the dominant discharge concept in Jansen (1979).

THREE-DIMENSIONAL MORPHODYNAMICS

Although rivers are sometimes considered as lines in the landscape and their behavior is often described with one-dimensional models, the flow and sediment-transport phenomena are essentially three-dimensional. In fact, the one-dimensional description is cross-sectionally averaged, instead of genuinely 1-D. This cross-sectionally averaged description is useful in many situations, especially when considering large-scale effects. At intermediate and local scales, however, it is not always applicable.

A well-known example of an essentially 3-D phenomenon is bend morphology. Although the alignment of regulated rivers is fixed, it usually contains windings, sometimes referring to the ancient channel pattern, sometimes the result of consecutive realignment measures. The flow pattern in these curved reaches is three-dimensional, due to curvature-induced secondary flows. As a consequence, the flow-induced bed shear stress has a component towards the center of curvature, which tends to force sediment towards the inner bank. There it piles up and forms a so-called *point bar* (cf. Figure 16), until the transverse slope becomes steep enough for the downhill gravitational transport component to balance the secondary-flow induced component.

This model applies only to bends with a constant radius of curvature and an infinite angle of rotation. Such bends do not exist in reality, rather the curvature varies continuously along the river. This means that spatial lag effects come into play: neither the velocity distribution nor the bed topography can adjust immediately to a change in curvature. The adaptation length scale for the flow, λ_w, is determined by the interplay between inertia and bed friction, the one for the transverse bed slope, λ_s, by the channel aspect ratio, the amount of sediment to be moved per unit channel length and the efficiency factor of the downhill gravitational transport component. The ratio of these two scales, called the *interaction parameter IP*, determines the transition behavior of the bed topography (Struiksma *et al.*, 1985).

Figure 16 View of the freely meandering Usumacinta River in Guatemala (flow from top to bottom). Note that the location level of the point bar does not coincide with the maximum curvature. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

$$\lambda_{\mathrm{w}} = \frac{4}{f}h;\ \lambda_{\mathrm{s}} = \frac{\beta}{8}\left(\frac{B}{h}\right)^2 h;\ IP \hat{=} \frac{\lambda_{\mathrm{s}}}{\lambda_{\mathrm{w}}} = \frac{\beta f}{32}\left(\frac{B}{h}\right)^2 \quad (2)$$

in which β is the coefficient of proportionality between the bed slope and the downhill gravitational transport rate.

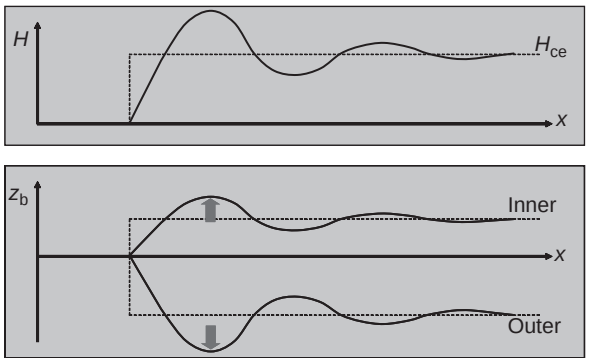


Figure 17 Damped oscillatory transition phenomenon at the entrance to a bend of constant curvature. H indicates the bed-level difference between the inner and the outer bend, H_{ce} the value it would take if the bend would continue infinitely. The lower panel gives the bed along the inner and outer bend. x is the position along the channel axis. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

For a certain range of values of *IP*, the transverse bed slope exhibits a damped oscillatory adaptation to a change in curvature (Figure 17). In system terms, this behavior is called *forced,* referring to the forcing by the channel curvature. It comes with an overshoot of the transverse bed slope in the beginning and just beyond the end of the bend, which may have implications for navigability and bank stability. For curvature variations of a certain wave length, this forced behavior may even be in resonant interaction with the curvature. Some researchers claim this to be the onset of meandering.

The forced behavior described earlier refers to the static equilibrium state, with constant discharge and sediment supply. The interaction of the retarded adjustment of the velocity field and that of the transverse bed slope also gives rise to a form of free (as opposed to forced) behavior, that is, the formation of a downstream migrating system of bars alternately at the left and the right bank of the river (Figure 18). Like the point bar formation in bends described previously, this formation of so-called *alternate bars* is governed by the channel aspect ratio. When this exceeds a certain threshold value, the bars begin to develop spontaneously, insofar as there is no spatio-temporal variation in the system's forcing that imposes this type of behavior. The extent to which the equilibrium amplitude is reached depends on the duration of the state of threshold exceedance, which is usually limited due to stage variations.

Blondeaux and Seminara (1985) came up with a unified theory for free and forced bars based on a weakly nonlinear stability analysis of the coupled water–sediment bed system. Later on, their group at the University of Genoa has done pioneering work on other types of morphodynamic instabilities (e.g. Blondeaux, 2001).

Figure 18 Jaeggi's famous shot of alternate bars in the Rhine in Switzerland (Jaeggi, 1984; flow from bottom to top) © American Society of Civil Engineers (ASCE)

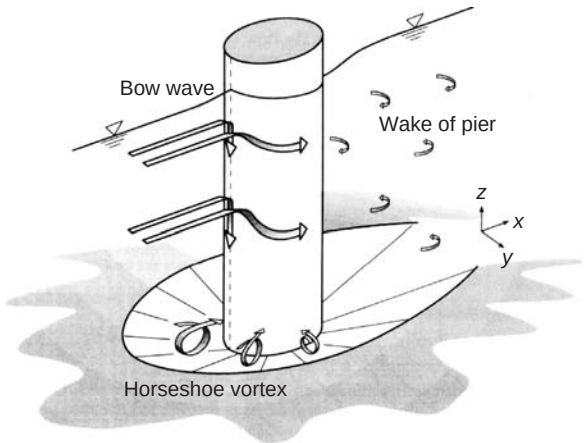


Figure 19 Scour around bridge piers cannot be modeled without taking due account of the 3-D horseshoe vortex (Graf, 1998)

This theoretical work starts from idealized situations that are seldom found in reality. Yet, the insight it gives

means an important support to numerical modeling of more realistic situations. One of the software systems enabling 3-D morphodynamic modeling is Delft3D (see **http://wldelft.nl/soft/d3d** for further reference).

At a local scale, essentially 3-D phenomena occur near structures and obstacles, such as bridge piers (Figure 19) and groyne heads. Here we need to distinguish two modes of scour called *clear water scour* and *live bed scour*, respectively. Clear water scour occurs if the shear stress at the undisturbed bed is below the threshold of motion. Near the structure, this threshold is exceeded and the bed keeps on eroding until the shear stress has been reduced to the threshold level. Strictly speaking, the erosion downstream of a sediment-blocking barrier or weir is also a case of clear water scour, because the critical shear stress is reached via coarsening of the bed sediment, rather than by a decrease of the bed shear stress.

An example of live bed scour is the development of scour holes and deposition areas (so-called *groyne flames*; see Figure 20) in groyne-regulated rivers (e.g. Yossef and Klaassen, 2002). Especially the scour holes near the tips of the groynes cannot be described without considering the 3-D water motion in the eddies shedding from the tip, which are very efficient in picking up sediment from the bed. The practical relevance of these phenomena lies primarily in the potential limitation of navigability due to the groyne flames.

Navigability is an important aspect of many regulated rivers. Bends are critical in this respect, especially because the point bar formation in the inner bend tends to reduce the navigable width exactly in a reach where extra width would be welcome, because the pathway of ship is larger in a bend, and because loaded ships tend to cross over towards the deep outer bend. Various measures have been taken to artificially increase the navigable width in bends, such as maintenance dredging, filling up the outer bend with riprap (Figure 21), increasing the flow resistance in the outer bend by bottom groynes, suppressing the secondary flow by bottom vanes, and so on (also see Klaassen *et al.*, 2002). All these measures, except for the dredging and the vanes, have in common the tendency to shift the maximum transport rate towards the inner bend, and also that they end somewhere near the bend exit. This combination leads to an unnatural transverse distribution of the sediment transport at the end of the structure, and hence to an increased pickup of sediment where the transport coming out of the bend is too low and an increased deposition where the transport is too high. This leads to an anomaly in the transverse bed slope, which decays further downstream (see Figure 21). Depending on the value of the interaction parameter, this can be an exponential or a damping oscillatory process. In any case, it has the effect of shifting the critical point for navigability from the bend to the reach just downstream of the bend.

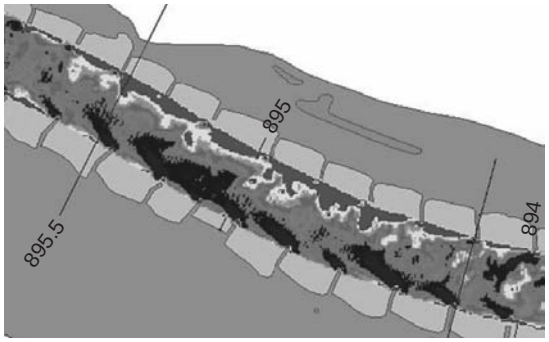


Figure 20 Groyne flames as observed in the Waal west of Nijmegen (flow from right to left). The asymmetry between the left and the right part of the channel is due to point bar formation in this slightly curved reach (courtesy Rijkswaterstaat). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

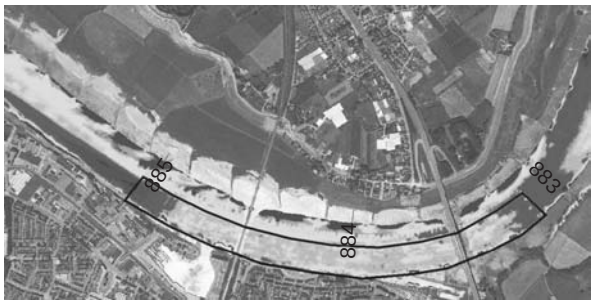


Figure 21 Outer bend fill-up in the River Waal at Nijmegen (dark contour). The flow is from right to left. Note the strong scour at the left bank just downstream of the filled area, and the deposition at the opposite side (courtesy Rijkswaterstaat). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CONCLUSION

Many of the above examples illustrate that effective river regulation is possible only on the basis of a thorough understanding of the dynamics involved in the river's response to the regulation measures. The same goes for the management of regulated rivers. Despite the control measures put in place, the river keeps part of its natural character, including the variability and the limited predictability involved. This has to be borne in mind whenever we are tempted to think that we are in full control of these systems.

REFERENCES

Blom A., Ribberink J.S. and de Vriend H.J. (2003) Vertical sorting in bed forms, Flume experiments with a natural and a tri-model mixture. *Water Resources Research*, **39**(2) 1025.

Blondeaux P. (2001) Mechanics of coastal forms. *Annual Review of Fluid Mechanics*, **33**, 339–370.

Blondeaux P. and Seminara G. (1985) *A unified bar-bend theory of river meanders*. *Journal of Fluid Mechanics*, **157**, 449.

Bulle H. (1926) *Investigations into the Trapping of Bedload in Branching Rivers*, VDI-Verlag, Forschungsarb. Ing. Wesens: Berlin, Heft 283 (in German).

Graf W.H. (1998) *Fluvial Hydraulics*, John Wiley & Sons: Chichester, p. 882, ISBN 0-471-97714-4.

Jaeggi M. (1984) Formation and effect of alternate bars. *Journal of Hydraulic Engineering, ASCE*, **110**(2), 142–156.

Jansen P.P.h (Ed.) (1979) *Principles of River Engineering*, Pitman: London; Facsimile edition: (1994) Delftse Uitgevers Maatschappij: Delft, p. 509, ISBN 90-6562-146-6.

Kemink E., Wang Z.B., de Vriend H.J. and van Beek E. (2003) Modelling of flood defence measures in the Lower Yellow River using SOBEK. In *Proceedings of the International Yellow River Forum*, Hongqi S. (Ed.), The Yellow River Conservancy Publishing House: Zhengzhou, China, pp. 214–223, ISBN 7-80621-676-6.

KHR (1993) *The Rhine under the influence of man – river engineering works, shipping, water management*. Int. Comm. for the Hydrology of the Rhine, Report I-11, 260 pp. ISBN 90-70980-17-7.

Klaassen G.J., Douben K.-J. and van der Wal M. (2002) Novel approaches in river engineering. In *River Flow 2002*, Bousmar D. and Zech Y. (Eds.), Balkema: Lisse, pp. 27–43, ISBN 90-5809-509-6.

Knighton D. (1998) *Fluvial Forms and Processes*, Arnold: London, p. 383, ISBN 0-340-66313-8.

Lane E.W. (1955) The importance of fluvial geomorphology in hydraulic engineering. *Proceedings of the ASCE*, **81**, 1–17.

Ribberink J.S. (1987) *Mathematical Modelling of One-dimensional Morphological Changes in Rivers with Non-Uniform Sediment*, Report No. 87–2, Delft University of Technology, Comm. on Hydraulic and Geotechnical Engineering, p. 200.

Ribberink J.S. and van der Sande J.T.M. (1987) Aggradation and degradation of alluvial channel beds. *Journal of Hydraulic Engineering, ASCE*, **113**(2), 272–276.

Silva W., Klijn F. and Dijkman J. (2001) *Room for the Rhine Branches in The Netherlands: What Research has Taught us*, Report RIZA-2001.031/WL – R 3294, Rijkswaterstaat/WL|Delft Hydraulics, p. 162, ISBN 90-3695-385-5.

Struiksma N., Olesen K.W., Flokstra C. and de Vriend H.J. (1985) Bed deformation in curved alluvial channels. *Journal of Hydraulic Research*, **23**(1), 57–79.

ten Brinke W.B.M. (2004) *The Controlled River*, Veen Magazines: Diemen, p. 228, ISBN 90-76988-65-x (in Dutch).

van der Klis H. (2003) *Uncertainty Analysis Applied to Numerical Models of River Bed Morphology*, PhD thesis, Delft University of Technology, Delft University Press, Delft, p. 147, ISBN 90-407-2447-4.

van der Lee G.E.M., Baptist M.J., Ververs M.J. and Geerling G. (2001) *Application of the Cyclic Floodplain Rejuvenation Strategy to the Waal River*, Technical Report CFR 7, WL|Delft Hydraulics / University of Nijmegen.

van Mazijk A. (1996) *One-dimensional Approach of Transport Phenomena of Dissolved Matter in Rivers*, PhD thesis, Delft University of Technology, p. 310. [also see: **www.chr-khr.org/chr-engels/ram.htm**]

van Vuren S., Kok M. and Ouwerkerk S.J. (2003) Impact of river morphology on extreme flood level prediction: a probabilistic approach, *Proceedings of the European Safety and Reliability Conference*, Maastricht, pp. 1653–1662, ISBN 190-5809-595-9, June 2003.

Visser P.J., Havinga H. and ten Brinke W.B.M. (1999) How do we keep the river navigable? *Land en Water*, **9**, 25–27. (in Dutch).

Wilbers A.W.E. and ten Brinke W.B.M. (2003) The response of subaqueous dunes to floods in sand and gravel bed reaches of the Dutch Rhine. *Sedimentology*, **50**, 1013–1034.

Yossef M. and de Vriend H.J. (2004) Mobile-bed experiments on the exchange of sediment between main channel and groyne fields. In *River Flow 2004*, Greco M., Carravetta A. and Della Morte R. (Eds.), Balkema Publishers: Leiden, pp. 127–133, ISBN 90-5809-658-0.

Yossef M. and Klaassen G.J. (2002) Reproduction of groynes-induced river bed morphology using LES in a 2-D morphological model. In *River Flow 2002*, Bousmar D. and Zech Y. (Eds.), Balkema: Lisse, pp. 1099–1108, ISBN 90-5809-509-6.

PART 13

Groundwater

145: Groundwater as an Element in the Hydrological Cycle

WILLIAM M ALLEY[1], JAMES W LA BAUGH[2] AND THOMAS E REILLY[2]

[1] United States Geological Survey, Office of Ground Water, San Diego, CA, US
[2] United States Geological Survey, Office of Ground Water, Reston, VA, US

Groundwater commonly is portrayed in diagrams of the hydrological cycle as a single pool of water with simple transfers of water to and from the land surface, oceans, and atmosphere. In reality, groundwater is an integral part of a complex hydrological cycle that involves the continuous movement of water on Earth. We begin with a global perspective on the importance of groundwater in the hydrological cycle and then address the importance of understanding the dynamics of groundwater-flow systems. We next turn our attention to groundwater budgets and the effects of humans and climate on these budgets. The final section is devoted to a discussion of interactions between groundwater and surface water, and we highlight the importance of these interactions in the hydrological cycle.

INTRODUCTION

Groundwater occurs almost everywhere beneath the land surface. The widespread occurrence of potable groundwater is a major reason it is used as a source of water supply worldwide. Groundwater also plays a crucial role in sustaining streamflow during dry periods and is vital to many lakes and wetlands. Moreover, many plants and aquatic animals depend greatly upon the groundwater that discharges to streams, lakes, and wetlands.

Groundwater commonly is portrayed in diagrams of the hydrological cycle as a single pool of water with simple transfers of water to and from the land surface, oceans, and atmosphere. In reality, groundwater is an integral part of a complex hydrological cycle that involves the continuous movement of water on Earth. In this article, we begin with a global perspective on the importance of groundwater in the hydrological cycle. The importance of understanding the dynamics of groundwater-flow systems is then addressed. We next turn our attention to groundwater budgets and the effects of humans and climate on these budgets. The final section is devoted to a discussion of interactions between groundwater and surface water, and we highlight the importance of these interactions in the hydrological cycle.

GROUNDWATER IN THE HYDROLOGICAL CYCLE OF THE EARTH

The volume of water contained in groundwater storage often is compared to other major global pools of water within the Earth's hydrological cycle. Published estimates of the volumes of water in major global pools are shown in Table 1. While the estimates of the volume of water in the Earth's oceans and atmosphere shown in Table 1 are relatively similar, the estimates for groundwater storage vary greatly. In part, this variability is due to different considerations of depth and salinity in defining the global groundwater pool, and, in part, the variability reflects less knowledge about groundwater than other global pools of water. The estimates of water contained as soil moisture also illustrate considerable uncertainty.

Note that early estimates by Kalle (1945) of the global groundwater pool greatly underestimated its volume. Kalle's estimates are shown for historical purposes, but are not considered further in the discussion below.

The two major pools of freshwater are groundwater and water contained in glaciers and polar ice. As shown in Table 1, and depending on its definition, the groundwater pool may constitute from less than one-third to more

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

Table 1 Volume of water (in thousand km^3) attributed to different parts of the world water balance (Significant figures largely retained from original sources)

	Source				
Water form	Kalle (1945)[f]	Nace (1967)[g]	Voskresensky (1978)[h]	L'vovich (1979)[i]	Schlesinger (1997)[j]
Ocean	1 372 000	1 320 000	1 338 000	1 370 332	1 350 000
Glaciers, Polar ice	16 500	29 200	24 064	24 000	33 000
Permafrost ice	– [a]	–	300	–	–
Groundwater	250	8350[b]	23 400	60 000	15 300
Fresh	–	–	*10 530*	*4000*[c]	–
Other	–	–	*12 870*	–	–
Lakes	250[d]	229	176.4	280	–
Fresh	–	*125*	*91*	*150*	–
Saline	–	*104*[e]	*85.4*	*125*	–
Reservoirs	–	–	–	*5*	–
Wetlands	–	–	11.47	–	–
Soil moisture	–	67	16.5	83	122
Rivers	[d]	1.25	2.12	1.2	40
Biota	–	–	1.12	–	1
Atmosphere	13	13	12.9	14	13
Total	1 389 013	1 357 860	1 385 984	1 454 710	1 398 476

[a]Presence of hyphen indicates no data provided.
[b]Groundwater to a depth of 4000 m.
[c]Groundwater that is actively exchanged in hydrological cycle.
[d]Value for lakes includes lakes and rivers.
[e]Includes inland seas.
[f]Data from table on page 31 of Kalle (1945) were presented in kg cm^{-2} of the Earth's surface area. These were converted to the values shown here using an estimate of the Earth's surface area of 510.1×10^6 km^2 from page 1 of Kalle and based on the specific gravity of water of 1.0 g per cm^3 at 4 °C.
[g]Data from Table 1 on page 2.
[h]Data from Table 9 on page 43.
[i]Data from Table 1 on page 15 and Table 2 on page 21 and text in Chapter 1.
[j]Data from Table 2.2 and Figure 10.1 on page 346.

than twice the volume of water contained in glaciers and polar ice.

Table 2 focuses on the global freshwater resource and shows the relative amounts of groundwater and other sources of freshwater (ignoring water frozen in glaciers and polar ice). Despite uncertainty in the values, it is clear that groundwater is the major pool of nonfrozen freshwater on Earth, likely composing more than 95% of this resource. Lakes and reservoirs contain much of the remaining nonfrozen (available) freshwater with considerable uncertainty about the volume contained as soil moisture. Rivers contain only a tiny fraction of the available freshwater.

Comparisons also can be made in the annual global fluxes of water as indicated in Table 3. The values for groundwater discharge to oceans shown in Table 3 include only estimates for direct discharge to bays, estuaries, and oceans. Groundwater also reaches coastal waters indirectly by discharge to streams and rivers that drain to coastal waters. The values for runoff to oceans shown in Table 3 include this indirect groundwater discharge to streams and rivers. For example, L'vovich (1979) estimates that 12 of the 41 thousand km^3 per year of runoff to oceans consists of groundwater discharge to streams and rivers.

The residence time of groundwater can vary from a few days to more than 10 000 years. It is difficult, however, to estimate an average global residence time for groundwater, because of the previously noted complexities in defining the global groundwater pool. In comparison, the average residence time for continental runoff in free-flowing rivers likely varies between 16 and 26 days, and regulation by reservoirs has increased the mean age of continental runoff to slightly longer than a month (Vörösmarty and Sahagian, 2000).

Some discussions of the global groundwater pool distinguish between groundwater in "active" exchange with the Earth's surface (commonly referred to as renewable groundwater) and "fossil" water accumulated over tens of thousands of years under different climatic conditions, which, once used, cannot be replenished readily (sometimes referred to as nonrenewable groundwater). Commonly, it is difficult to describe the two categories of groundwater reserves quantitatively and to draw boundaries between them.

Water withdrawals by humans are superimposed upon the natural pools and fluxes of groundwater. Groundwater withdrawals are large and have increased greatly during the past 50 years. Perhaps as many as two billion people depend directly upon groundwater for drinking water, and 40% of the world's food is produced by irrigated agriculture that relies largely on groundwater (United Nations Environment

Table 2 Percent of total nonfrozen freshwater resources attributed to different components of the world water balance. Percent values were derived from volumes presented in Table 1

	Source			
Freshwater form	Nace (1967)	Voskresensky (1978)	L'vovich (1979)	Schlesinger (1997)
Groundwater	97.74	98.86	94.31[a]	98.95
Lakes	1.46	0.85	3.54	–
Reservoirs	–	–	0.12	–
Wetlands	–	0.11	–	–
Soil moisture	0.78	0.15	1.96	0.79
Rivers	0.01	0.02	0.03	0.26
Total	100	100	100	100

[a]Groundwater that is actively exchanged in hydrological cycle.

Table 3 Annual water fluxes (in thousand km^3 per year) between the atmosphere, land, and ocean, and between land and the ocean

	Source					
Water flux form	Hutchinson (1957)[a]	Nace (1967)[b]	Budyko and Sokolov (1978)[c]	L'vovich (1979)[d]	Chahine (1992)[e]	Schlesinger (1997)[f]
Evaporation						
From ocean surfaces	383	350	505	452.6	434	425
From land surfaces	63	70	72	72	71	71
Precipitation						
On ocean surfaces	347	320	458	411.6	398	385
On land surfaces	99	100	119	113.5	107	111
Runoff/Discharge						
Runoff to oceans	36	38[g]	44.7	41	36	40
Direct groundwater discharge to oceans	–	1.6[h]	2.2	2.2	–	–

[a]Data from Table 16 calculated from amounts published in Kalle (1945).
[b]Data from Table 1 on page 2.
[c]Data from Table 189 on page 590.
[d]Data from Table 1 on page 15 and Table 2 on page 21 and text in Chapter 1; groundwater discharge value from page 30.
[e]5 Data from Figure 1.
[f]Data from Figure 10.1 on page 346.
[g]From rivers and icecaps.
[h]Arbitrarily set to about 5% of runoff.

Programme, 2003). Unfortunately, despite the importance of groundwater, estimates of its use in many countries are either unavailable or of poor quality.

Finally, it is worth noting that groundwater depletion may have affected global water balances enough to contribute to sea-level rise during the past century as a result of water pumped from wells that returns to the sea by runoff or evaporation/precipitation (Sahagian *et al.*, 1994). Extensive drainage of wetlands during the past century also may have contributed to sea-level rise.

GROUNDWATER-FLOW SYSTEMS

The three-dimensional body of Earth material saturated with moving groundwater that extends from areas of recharge to areas of discharge is referred to as a groundwater-flow system (Figure 1). Perhaps the most obvious source of water to groundwater-flow systems (or more simply referred to as groundwater systems) is areal recharge from precipitation. Water from areal recharge flows from the water table where the recharge enters the saturated groundwater system through the flow system to the discharge area (which is a stream in Figure 1). Surface water features such as streams, wetlands, and lakes can either be the location of water entering the groundwater system (recharge area) or the location of water leaving the groundwater system (discharge area).

The areal extent of groundwater-flow systems varies from a few square kilometers or less to tens of thousands of square kilometers. The lengths of groundwater-flow paths range from a few meters to tens, and sometimes hundreds, of kilometers. A deep groundwater-flow system with long flow paths between areas of recharge and discharge may be overlain by, and in hydraulic connection with, several shallow, more local, flow systems.

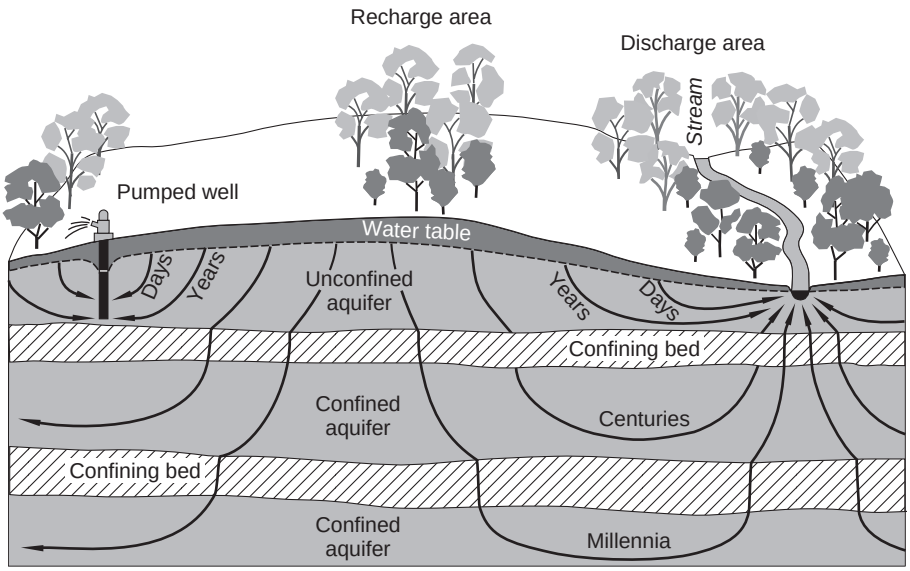


Figure 1 Groundwater system showing generalized flow paths of groundwater movement and the relative age of the water since the time of recharge. (Modified from Heath, 1983 and Winter *et al.*, 1998)

Factors that control Groundwater Movement

Two types of factors control groundwater movement: factors within the porous media and factors along the boundaries of the porous media where the water enters and leaves the groundwater system.

The equation that describes the movement of groundwater within a porous media is known as *Darcy's law*. In simplified terms, the average linear velocity (Freeze and Cherry, 1979) of the water is calculated from Darcy's law as:

$$v = -\left(\frac{K}{n}\right) i \qquad (1)$$

where v is the average linear velocity (referred to as velocity for the remainder of the text) of the groundwater (LT^{-1}), K is the hydraulic conductivity (LT^{-1}), n is the effective porosity (dimensionless), and i is the hydraulic gradient (the change in head per unit of distance; dimensionless), which is a negative number in the direction of flow. As a result, the hydraulic conductivity, effective porosity, and hydraulic gradient are all important in the determination of the movement of groundwater. The hydraulic conductivity, which represents the ability of the geologic framework to transmit water, is a property of the porous media and the fluid contained therein. The hydraulic conductivity of a groundwater system can vary over many orders of magnitude (Figure 2). The larger the hydraulic conductivity of a porous media, the easier it is for water to flow through it. The porosity is the ratio of the volume of the voids divided by the total volume. The effective porosity is the volume of the voids that are interconnected and available for fluid transmission divided by the total volume. The hydraulic gradient is the change in static head per unit of distance in a given direction (Lohman *et al.*, 1972). The static head generally is measured as the water level in a well. The water-level distribution or potentiometric surface describes the hydraulic gradient within an aquifer, which, in turn, determines the direction and rate of groundwater flow. For groundwater systems in cavernous karst terrain or in fractured-rock systems, the validity of Darcy's law, which was developed for porous media, may not be strictly applicable (White, 1993), and other methods for determining the velocity distribution may be required.

Darcy's law describes the rates and directions that water will flow in a groundwater system. The other important factors that are necessary to adequately understand the movement of water in a groundwater system are the sources of water to the system and the stresses (including water use) on the system. The external boundaries of the groundwater system, such as streams and areal recharge, affect the paths of water movement and the hydraulic head distribution (Reilly, 2001). The location of wells and water use (for example, irrigation) also impact where water moves within the groundwater system.

Time of Travel in a Groundwater Flow System

The age (time since recharge) of groundwater varies in different parts of groundwater-flow systems. The age of groundwater increases along a particular flow path through the groundwater-flow system from an area of recharge to

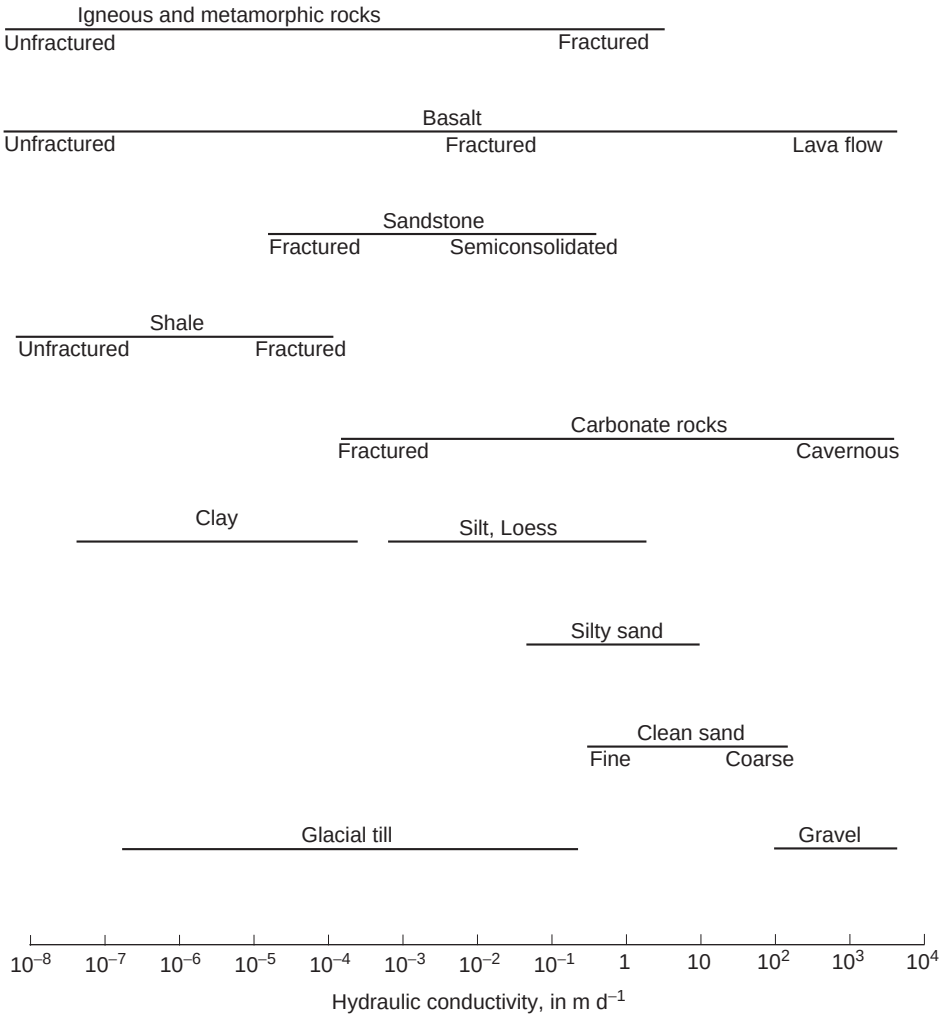


Figure 2 Range of hydraulic conductivity for selected rock types (Modified from Heath, 1983)

an area of discharge. In general, Darcy's law determines the flow velocity, and, therefore, the time of travel. Travel times within groundwater systems can vary considerably (Figure 1). In shallow flow systems, ages of groundwater at areas of discharge can vary from less than a day to a few hundred years. In deep, regional flow systems with long flow paths (tens of kilometers), ages of groundwater may reach thousands or tens of thousands of years or more. The shallower and younger groundwater tends to be more vulnerable to contamination from human activities at the land surface (Focazio *et al.*, 2002).

Fractured-rock systems in bedrock usually have smaller effective porosities than unconsolidated porous media such as sands and gravels, and flow velocities through fractured rock can be relatively fast. For example, travel times of water over distances of several kilometers have been estimated at less than a year for municipal wells completed in fractured dolomite in Wisconsin (Rayne *et al.*, 2001). In such cases, seasonal variations in recharge and pumping affect the variability in travel times. In more sluggish groundwater systems, such as the Bangkok Basin in Thailand (Sanford and Buapeng, 1996), long-term climate and geologic change need to be considered in understanding the movement of groundwater over tens of thousands of years.

Tracer techniques have been applied widely to estimate the residence time of subsurface waters, as well as the amounts and timing of recharge and discharge (Cook and Böhlke, 1999). Most tracer techniques require knowledge (or assumption) of the time history of tracer input at the land surface or the water table. This temporal pattern then is correlated to a concentration-depth pattern in the subsurface at a point in time. Other approaches use information on decay products to determine age. Tracers can occur naturally (chloride, heat, the stable isotopes ^{2}H and ^{18}O), occur in the atmosphere as a result of human activities (tritium, ^{36}Cl, CFCs), or be applied intentionally on the land surface (fertilizers, pesticides). Isotopes of elements,

such as radon, dissolved from host rocks also can be used to estimate residence times and interactions with surface water. Over the past decade, research in age dating and tracking young groundwater (<50 years old) using multiple tracers has resulted in major breakthroughs in understanding the dynamics of groundwater systems.

The time required for water to move through a groundwater system also can be estimated for some systems on the basis of aquifer storage volumes. A three-dimensional groundwater system contains a specific volume of water in the pores and cracks of the media. This volume has been referred to here as the volume in storage or the "pool" of water in storage (V in units of L^3). The rate at which this volume is replaced is related to the rate of recharge entering the system that displaces the water in storage; this is the flux of water entering the system (Q in units of L^3T^{-1}). In an average (or simplified) sense that does not take into account differences in flow paths and the nature of the framework, the time required for water to move through the entire system is simply VQ^{-1}. This approach may yield a poor approximation in complex flow systems, such as karst or fractured rock.

GROUNDWATER BUDGETS

Water budgets are used widely to account for flow and storage changes in various hydrological systems, including rivers, lakes, drainage basins, the land surface, and groundwater systems. In its most basic form, a water budget is expressed simply as:

$$\text{Inflow} - \text{Outflow} = \text{Change in Water Storage} \quad (2)$$

Each of the three main terms in the water-budget equation can include both natural (e.g. precipitation) and human-induced (e.g. imported water) components. Quantities in the budget apply to a predefined volume and can either be flow rates [L^3T^{-1}] or volumes [L^3] for a specified time period.

Understanding water budgets for groundwater systems is both critically important and challenging. Groundwater is hidden from view, and groundwater divides do not necessarily coincide with surface-water divides (Winter *et al.*, 2003). Several aquifers may underlie the land surface at any given location with varying hydraulic properties and locations of recharge.

Under natural (predevelopment) conditions, a groundwater system is in long-term equilibrium. That is, averaged over some period of time (and in the absence of climate change), the amount of water entering or recharging the system is approximately equal to the amount of water leaving or discharging from the system. Inflows (recharge) to a groundwater system under natural conditions include areal recharge from precipitation on the land surface and recharge from losing streams, lakes, and wetlands. Outflows (discharge) from a groundwater system under natural conditions include discharge to surface-water bodies and groundwater evapotranspiration. Because the system is in equilibrium, the quantity of water stored is constant or varies about some average condition in response to annual or longer-term climatic variations. This predevelopment water budget is shown schematically in Figure 3(a).

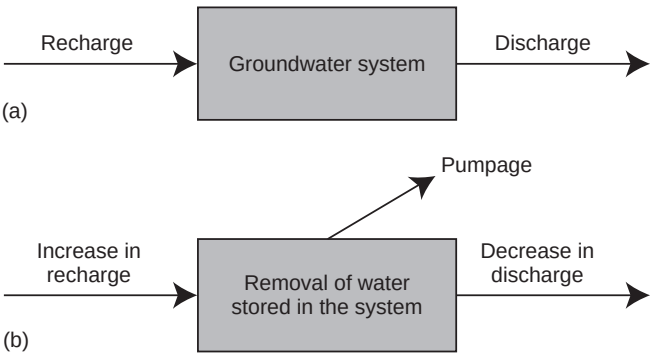


Figure 3 (a) Predevelopment water-budget diagram illustrating that inflow equals outflow. (b) Water-budget diagram showing changes in flow for a groundwater system being pumped. The sources of water for the pumpage are changes in recharge, discharge, and the amount of water stored. (Alley *et al.,* 1999)

Effects of Human Activities on Groundwater Budgets

Withdrawals of groundwater by pumping change the natural (or predevelopment) flow system. Water that is withdrawn must be supplied by some combination of (i) more water entering the groundwater system (increased recharge), (ii) less water leaving the system (decreased discharge), and (iii) removal of water that was stored in the system. This statement, illustrated in Figure 3(b), can be written in terms of rates (or volumes over a specified period of time) as:

$$\text{Pumpage} = \text{Increased recharge} + \text{Water removed from storage} + \text{Decreased discharge} \quad (3)$$

That is, the water pumped from the system must come from some change of flows and from removal of water stored in the predevelopment system (Theis, 1940; Lohman, 1972). Note that the initial predevelopment water-budget values do not enter directly into the budget calculation.

Regardless of the amount of water withdrawn, the system will undergo some drawdown in water levels in pumping wells to induce the flow of water to these wells, which means that some water initially is removed from storage. For most groundwater systems, the change in storage in response to pumping is a transient phenomenon that occurs as the system readjusts to the pumping stress. The relative

contributions of changes in storage, changes in recharge, and changes in discharge evolve with time. If the system can come to a new equilibrium, the changes in storage will diminish to zero and inflows will again balance outflows:

$$\text{Pumpage} = \text{Increased recharge} + \text{Decreased discharge} \quad (4)$$

Thus, the long-term source of water to discharging wells is typically a change in the amount of water entering or leaving the system.

The time that is required to bring a hydrologic system into equilibrium depends on the rate at which the discharge can be captured, which is a function of the characteristics of the aquifer system and the placement of pumping wells (Lohman, 1972). In many circumstances, the dynamics of the groundwater system are such that an approximate equilibrium condition would not be reached for decades or even centuries.

A prevalent view among many hydrologists and nonhydrologists alike is that the development of a groundwater system is considered to be "safe" if the rate of groundwater withdrawal does not exceed the rate of natural recharge. Bredehoeft *et al.* (1982) referred to this concept as the "Water-Budget Myth". It is an oversimplification of the information that is needed to understand the effects of developing a groundwater system. Natural recharge is a critical element for understanding water budgets, is an essential parameter in the modeling of aquifer systems, and its estimation is fundamental to understanding contaminant transport from the land surface. Nonetheless, an estimate of natural recharge, by itself, is of limited value in determining the amount of groundwater that can be withdrawn on a sustained basis. How much groundwater is available for use depends much more upon how changes in inflow and outflow, which result from pumping, affect the surrounding environment and the acceptable trade-off between groundwater use and these changes. Achieving this trade-off in the long term is a central theme in the evolving concept of sustainability (Alley *et al.*, 1999; Sophocleous, 2000; Alley and Leake, 2004).

In determining the effects of pumping and the amount of water available for use, it is critical to recognize that not all the water pumped is necessarily consumed (Kendy, 2003). For example, some of the water pumped for irrigation is consumed by evapotranspiration and some of the water returns to the groundwater system by infiltration, canal leakage, and other means of irrigation return flow. Most other uses of groundwater are similar, in that some of the water pumped is not consumed, but is returned to the system.

Although withdrawal of groundwater is the most direct way in which humans affect groundwater budgets, many other human activities commonly must be considered, particularly those that affect recharge rates. Changes in the patterns and rates of recharge can be caused by irrigation and urban development, removal or changes in the type of vegetation, changes in surface-water flows and storage, and land drainage. Identifying human practices that influence recharge is straightforward, but quantifying the effects of these practices is very difficult. For example, built-up and paved areas promote runoff and inhibit infiltration. The enhanced runoff, however, may be channeled to a retention basin or infiltration gallery, resulting in the relocation of recharge areas and the transition from slow, diffuse recharge to rapid, localized recharge (Lerner, 2002). Wastewater infiltration also is often a major component of overall recharge to aquifers around urban areas, especially in more arid climates (Foster and Chilton, 2004).

Artificial recharge with excess surface water or reclaimed wastewater is increasing in many areas, thus becoming an increasingly important component of the hydrological cycle. In fact, with long-term intensive use, much of the local groundwater may be derived from artificial recharge (Shelton *et al.*, 2001)

Examples of large-scale changes that can occur in the water budgets of regional groundwater systems are illustrated by examining the simulated sources of water (increased recharge, decreased discharge, and changes in storage) that supplied withdrawals from nine major groundwater systems in the United States. These results, shown in Figure 4, are from numerical simulations of major regional aquifer systems for the period of development up through the mid to latter 1980s. The results illustrate large variability in the relative proportions of different sources of water to pumping wells. The Floridan and Edwards–Trinity aquifer systems, which equilibrate rapidly after pumping, were simulated as steady state with no long-term change in storage for the simulation period. In contrast, the Southern High Plains, where most natural discharge occurs far from pumping wells, and the deeply buried Great Plains aquifer system had substantial changes in groundwater storage. In some areas, such as the California Central Valley and Eastern Snake River Plain, return flow of excess irrigation water was the major change in the water budget during the simulation period. Note that the distinction between changes in recharge and changes in discharge commonly is a function of how the system was defined (i.e. a gain to one system may result in a loss from an adjoining system). For example, groundwater withdrawals from confined aquifers (Northern Atlantic Coastal Plain and Gulf Coastal Plain) can cause flow to be diverted (recharged) into the deeper regional flow regime that would otherwise discharge to streams in the outcrop areas.

Effects of Climate on Groundwater Budgets

Climate affects groundwater budgets in several ways. Most directly, climate influences the rates and distribution of

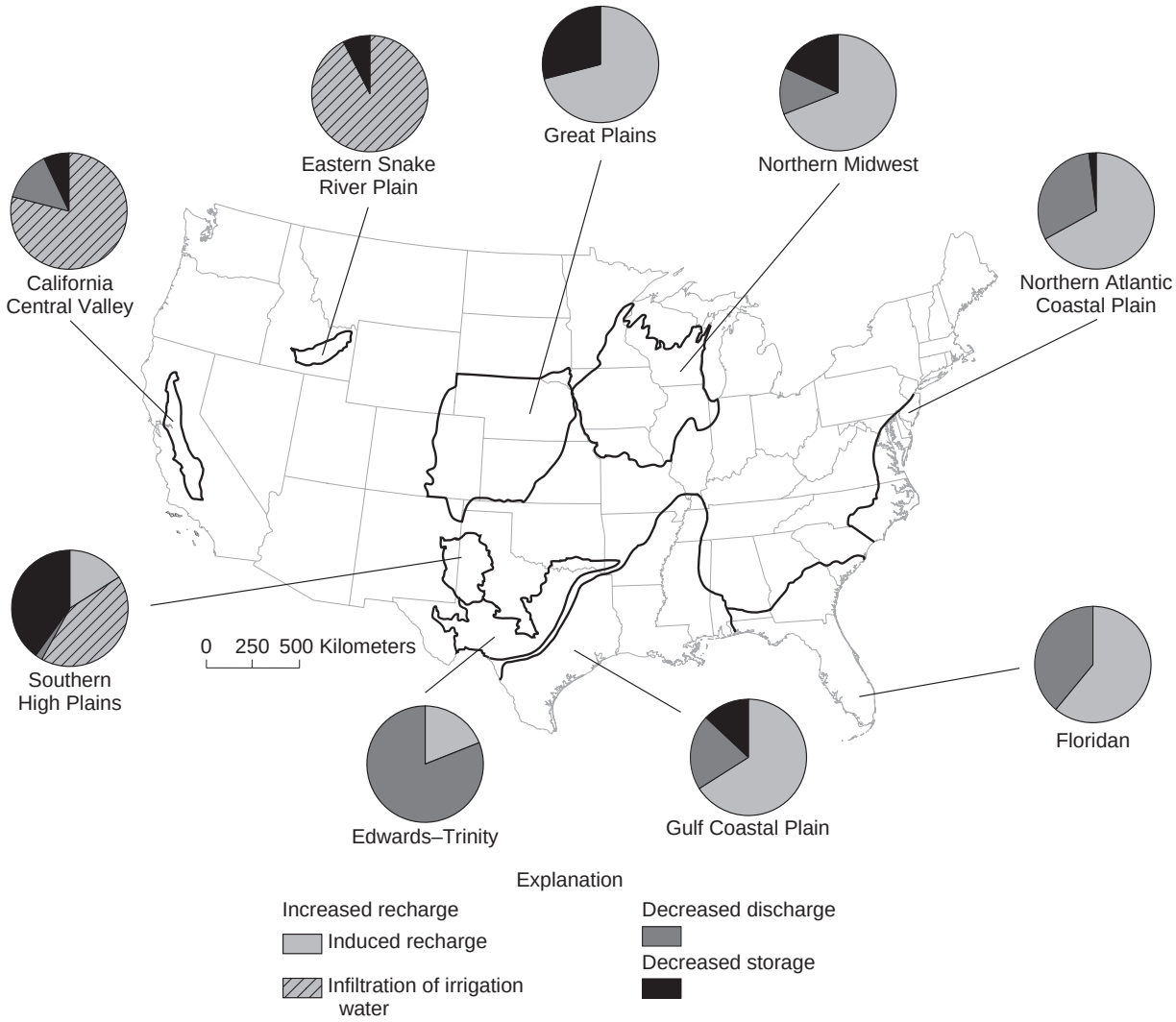


Figure 4 Sources of water that supply withdrawals from major aquifer systems in the United States based on model simulations for various periods through the 1970s and 1980s. The results illustrate the variety of ways in which overall groundwater budgets can change in response to large-scale pumping. The simulations of the aquifer systems were done at different stages in their development, and, in many cases, are not representative of today's conditions given the dynamic nature of groundwater systems. (Data from Johnston, 1999; Modified from Alley *et al.,* 2002)

recharge. Climate also affects human demands for groundwater and affects plant transpiration from shallow groundwater in response to changing inputs of solar energy and changing depths to the water table. Some groundwater responses to climate variability and change may show considerable temporal lag, given the relatively long response times of groundwater systems. For example, long-term trends in the balance between precipitation and evapotranspiration – caused by either long-term variability or by anthropogenic global change – may affect groundwater discharge, but their effects may be attenuated and spread out over time. As an indication of the effects of long-term climate changes on groundwater systems, much of the water pumped today in arid regions comes from aquifers that were recharged at higher rates during wetter or cooler conditions in the past (e.g. Zhu *et al.*, 1998).

Long-term droughts may be viewed as a natural stress on groundwater systems that in many ways have effects similar to groundwater withdrawals; namely, reductions in groundwater storage and accompanying reductions in groundwater discharge to streams and other surface-water bodies. Because climate stress on the hydrologic system is added to the existing or projected human-derived stress, droughts represent extreme hydrologic conditions that may be important in long-term groundwater management.

Human-induced climate change in the coming decades may affect groundwater resources in several ways, including (i) long-term changes in groundwater recharge resulting

from changes in the annual and seasonal distribution of precipitation and temperature, (ii) more severe and longer lasting droughts, (iii) changes in evapotranspiration resulting from changes in vegetation, and (iv) possible increased demands for groundwater as a backup source of water supply. Surficial aquifers that supply much of the water in streams, lakes, and wetlands are likely to be the part of the groundwater system most sensitive to climate change. Climate not only affects groundwater budgets, but to the extent that groundwater systems influence evapotranspiration and runoff, changes in groundwater fluxes also may affect climate (National Research Council, 2004).

Several studies have used atmospheric circulation models and future scenarios of precipitation and temperature to simulate the effects of climate change on groundwater recharge, storage, and discharge. For example, York *et al.* (2002) illustrate the potential importance of links among climate, shallow groundwater levels, and groundwater evapotranspiration using a coupled aquifer-land surface-atmosphere model.

INTERACTIONS BETWEEN GROUNDWATER AND SURFACE WATER

Interactions between groundwater and surface water are an integral part of the hydrological cycle. The interactions need to be viewed at a range of scales and as having large variability in space and time. Many of the early studies of these interactions focused on streams. Once thought to be of little consequence and thus ignored, interactions of groundwater with lakes, wetlands, estuaries, and oceans are now recognized as important processes. The interactions of surface-water bodies with groundwater systems are governed by the positions of the water bodies relative to the groundwater-flow system, the characteristics of surface-water beds and underlying geologic materials, and the climatic setting (Winter *et al.*, 1998).

Streams

Streams interact with groundwater in various ways. Depending upon the relation between groundwater head at the stream–channel interface and the stream stage, a stream may gain water from inflow of groundwater through the streambed (gaining stream; stream stage < groundwater head), lose water to groundwater by outflow through the streambed (losing stream; stream stage > groundwater head), gain water from groundwater on one side of the stream and lose water to groundwater on the other side (flow-through), or possibly have zero exchange (parallel-flow) when the channel stage and groundwater head are equal throughout a cross section (Woessner, 2000). The position of a channel within a floodplain can control the type of exchange. For example, flow-through reaches are most often found when a channel cuts perpendicular to the fluvial plain groundwater-flow field (Wroblicky *et al.,* 1998; Woessner, 2000). Larkin and Sharp (1992) classify stream-aquifer systems based on the predominant regional groundwater-flow component as (i) underflow component–dominated (groundwater flow is largely parallel to and in the same direction as the stream), (ii) baseflow component–dominated (groundwater flow is largely perpendicular to and from the stream), and (iii) mixed. They conclude that the dominant groundwater component (baseflow or underflow) can be inferred from geomorphic data such as channel slope, river sinuosity, degree of river incision through the alluvium, width-to-depth ratio of the bankfull river channel, and the character of the fluvial depositional system.

In some environments, streamflow gain or loss can persist; that is, a stream might always gain water from groundwater, or it might always lose water to groundwater. In many environments, the flow direction can vary along a stream; some reaches receive groundwater and other reaches lose groundwater. Flow directions between groundwater and surface water also can change on daily or shorter timeframes as a result of individual storms causing focused recharge near the streambank, flood peaks moving down the channel, or transpiration of groundwater by streamside vegetation.

Losing streams can be connected to the groundwater system by a continuous saturated zone or can be disconnected from the groundwater system by an unsaturated zone. An important feature of streams that are disconnected from groundwater is that pumping shallow groundwater near the stream does not affect the flow of the stream near the pumped wells.

Even in settings where streams are primarily losing water to groundwater, certain reaches may receive groundwater inflow during some seasons. The amount of water that groundwater contributes to streams can be estimated by analyzing streamflow hydrographs to determine the groundwater (baseflow) component (Rutledge, 1993; Halford and Mayer, 2000). The proportion of stream water derived from groundwater inflow varies considerably among different physiographic and climatic settings, as illustrated in Figure 5 for streams in 10 different regions in the United States.

Water exchange across the interface between groundwater and surface water has been explored in some detail in the past decade (Jones and Mulholland, 2000). Many studies have demonstrated that local geomorphic features such as streambed topography, streambed roughness, meandering, and heterogeneities in sediment hydraulic conductivities can give rise to localized flow systems within streambeds and banks (Harvey and Bencala, 1993; Winter *et al.,* 1998). The near-stream subsurface environment in which there is active

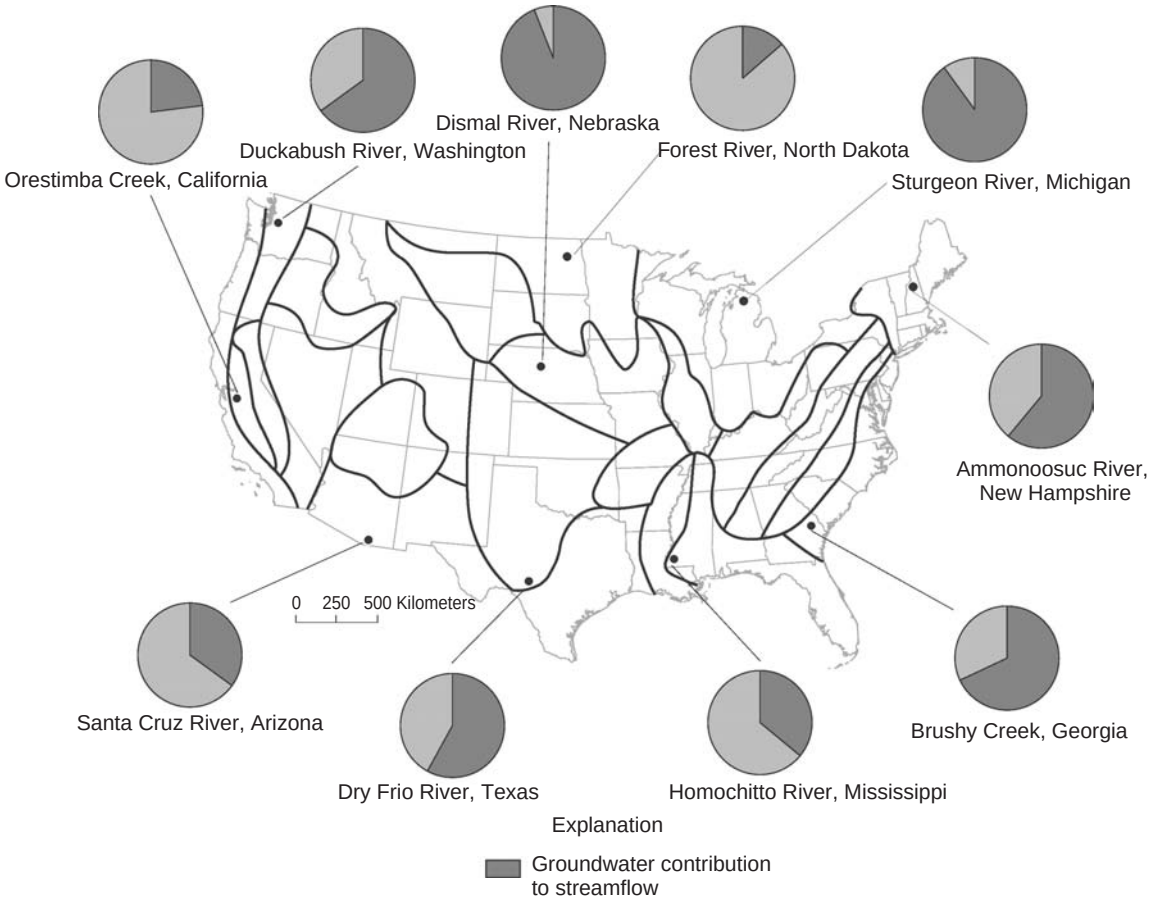


Figure 5 Twenty-four regions were delineated in the conterminous United States where the interactions of groundwater and surface water are considered to have similar characteristics. The estimated groundwater contribution to streamflow is shown for specific streams in 10 of these regions. (Modified from Winter *et al.,*1998)

exchange between surface water and groundwater commonly is referred to as the hyporheic zone. The hyporheic zone can be viewed as a zone in which finer-scale interactions between the stream-channel water and groundwater occur within the context of larger-scale patterns of loss and gain of channel water in drainage basins. Understanding these very localized subsystems within the hydrological cycle is increasingly viewed as important to understanding the chemical composition of surface and subsurface water and stream and riparian ecology (Brunke and Gonser, 1997; Boulton, 2000). For example, an important feature of water exchange between streams and hyporheic zones is that surface water is kept in close contact with chemically reactive mineral coatings and microbial colonies in the subsurface, having resultant effects on biogeochemical processes such as nitrification (Dahm *et al.,* 1998; Hinkle *et al.*, 2001). Localized exchange of surface water and groundwater also can result in thermal effects that influence the distribution of biota and biogeochemical processes. For example, thermal effects of groundwater discharge have been directly related to fish habitat, both in terms of spawning areas and refuge for adults when ice forms in colder environments (Power *et al.,* 1999)

Lakes and Wetlands

The relation of lakes and wetlands to groundwater can be viewed as occurring in three different ways: (i) lakes and wetlands may recharge groundwater without receiving groundwater input, (ii) they may both recharge groundwater and receive discharge water from groundwater, or (iii) they may receive groundwater discharge, but not discharge water to groundwater (Sloan, 1972; Winter, 1976; Boyle, 1994). These relations are influenced by the positions of the lakes or wetlands within regional flow systems (Smith and Townley, 2002). Groundwater-flow patterns in the vicinity of these water bodies can vary in response to shifts in subsurface-water divides that occur during wetter or drier conditions (Holzbecher, 2001).

The amount of groundwater moving into or through lakes and wetlands varies considerably. Negligible to relatively small amounts of groundwater move into or

through lakes and wetlands in poor hydraulically conductive materials, such as some crystalline rocks (Schindler *et al.*, 1976), clayey till (Shaw and Prepas, 1990), lacustrine clays (Zektzer and Kudelin, 1966), or dense organic peats. More considerable amounts of groundwater move into or through lakes and wetlands in more conductive materials such as sandy tills, sands, or sands, and gravels (Shaw and Prepas, 1990), as well as in karst (Lee, 2002).

Groundwater movement into and out of lakes and wetlands can be diffused, focused, or some combination thereof. Focused groundwater discharge into lakes and wetlands is in the form of springs. The direction of flow at the periphery of lakes and wetlands can be affected by transpiration. Flow reversals in response to transpiration have been documented in a variety of geologic materials, such as glacial till (Meyboom, 1967), organic peat (Fraser *et al.*, 2001), and sandy till (Doss, 1993).

Springs

Springs typically are present where the water table intersects the land surface. They may form at topographic depressions; along lithologic contacts where permeable rocks overlie rocks of much lower permeability; along joints, fractures, or faults; or in karst areas (Fetter, 1988). Springs serve as important sources of water to streams and other surface-water features, as well as being important cultural and esthetic features in themselves. The important role of springs in human settlement is evident from the many localities named after the local springs. The constant source of water typical at many springs leads to abundant growth of plants and, in many cases, to unique habitats. Groundwater development can lead to reductions in spring flow, alterations of springs from perennial to ephemeral, or elimination of springs altogether. Springs typically represent points on the landscape where groundwater-flow paths from different sources converge. Groundwater development may affect the amount of flow from these different sources to varying extents, thus affecting the chemical composition of the spring water. Where springs discharge at land surface, they may offer a relatively easy place to measure natural groundwater discharge. However, because springs integrate the signal of hydrologic and geologic processes, sometimes over large spatial areas and long periods of time, they typically provide only an indirect source of information about groundwater systems (Manga, 2001).

Bays, Estuaries, and Oceans

The interface between saltwater and freshwater in groundwater systems is important in determining the amount of freshwater available in the aquifer system. This boundary generally occurs as fresh groundwater discharges into salty bays, estuaries, and oceans; however, the boundary also occurs inland where deeper brines bound an active freshwater flow system. In relatively homogeneous porous media, the denser saltwater tends to remain separated from the overlying freshwater. A transition zone, however, known as the *zone of diffusion* or the *zone of dispersion*, forms between the two fluids. In coastal areas where the porous media is heterogeneous in nature, a system of layered mixing zones can form. Substantial groundwater withdrawals can cause surrounding saltwater to move into areas of groundwater use in coastal (Bear *et al.*, 1999; Barlow, 2003) and inland (Lahm and Bair, 2000) areas, decrease the volume of freshwater available for use, and reduce the amount of freshwater discharged into coastal ecosystems.

Most research to date has been on the landward movement of saltwater in response to pumping in coastal areas. More recently, the seaward flow of fresh groundwater to coastal ecosystems and the role of groundwater in delivering nutrients and other dissolved constituents to these systems have received increased attention (Burnett *et al.*, 2002). As with all the components of the hydrological cycle, this circulation of water from one part of the cycle to another part is important to many issues.

The time required for the saltwater–freshwater interface and freshwater discharge to respond to human and natural changes can range from almost instantaneous to thousands of years. In some coastal areas, such as New Jersey, United States, and Suriname, South America (Kooi and Groen, 2001), relatively fresh groundwater found far off the coast is hypothesized to be water that was recharged during the Pleistocene when sea levels were much lower and has not yet been discharged to the surface.

CONCLUDING REMARKS

During the past 50 years, groundwater depletion has spread from isolated pockets to large areas in many countries throughout the world. Although sometimes viewed as a separate resource, groundwater is in fact an integral part of the hydrological cycle. Groundwater systems are dynamic and continually changing in response to human and climatic stresses. Linkages with surface water and the land surface are numerous and extend over many different timescales.

Future success in understanding the dynamic nature of groundwater systems will rely on continued and expanded data collection at various scales. Despite their importance, groundwater data have received little attention in concerns expressed about the continuity of global water data, primarily because such concerns have focused more on visible atmospheric and surface-water monitoring networks.

Water-level measurements from wells remain the principal source of information on the effects of hydrologic stresses on groundwater systems. Advances in instrumentation now enable the collection of real-time water-level data allowing scientists to observe diurnal and seasonal

trends from well networks across large areas (Cunningham, 2001). To understand the true nature of change in a groundwater system and to differentiate between natural and human-induced changes, records of water-level measurements over substantial periods are required (Taylor and Alley, 2001).

Surface-water depletion is viewed increasingly as the limiting factor to the long-term utilization of groundwater resources, yet the distinctly different temporal and spatial scales at which groundwater and surface-water systems operate present major challenges to their integrated analysis. The locations, quantity, and timing of reductions in surface-water flow resulting from groundwater development are fundamental questions at the scales of years to decades, whereas ecological issues require attention to seasonal and even diurnal changes in groundwater recharge and discharge. There is a continuing need to develop modeling tools that better represent groundwater as an integral part of the hydrological cycle; for example, coupled watershed and groundwater models are needed to better simulate the hydrological continuum from the atmosphere, to the land surface, to the unsaturated and saturated subsurface zones, and then back to surface waters or land surface.

Finally, groundwater systems have value not only as perennial sources of water supply, but also as reservoirs for cyclical injection and withdrawal to modulate the variability inherent in surface-water supplies. Management approaches increasingly involve the use of artificial recharge of excess surface water or recycled water by direct well injection, surface spreading, or induced recharge from streams. Many scientific challenges remain, however, to understand more fully the long-term responses of aquifer systems to alternate management approaches (Galloway *et al.*, 2003).

REFERENCES

Alley W.M., Healy R.W., LaBaugh J.W. and Reilly T.E. (2002) Flow and storage in groundwater systems. *Science*, **296**, 1985–1990.

Alley W.M. and Leake S.A. (2004) The journey from safe yield to sustainability. *Ground Water*, **42**, 12–16.

Alley W.M., Reilly T.E. and Franke O.L. (1999) *Sustainability of Ground-Water Resources*, U.S. Geological Survey Circular 1186. Also available on the World Wide Web at **http://pubs.water.usgs.gov/circ1186**.

Barlow P.M. (2003) *Ground Water in Freshwater-Saltwater Environments of the Atlantic Coast*, U.S. Geological Survey Circular 1262. Also available on the World Wide Web at **http://pubs.water.usgs.gov/circ1262**.

Bear J., Cheng A.H.-D., Soreck S., Ouazar D. and Herrera I. (Eds.) (1999) *Seawater Intrusion in Coastal Aquifers – Concepts, Methods, and Practices*, Kluwer Academic Publishers, Dordrecht.

Boulton A.J. (2000) River ecosystem health down under: assessing ecological conditions in riverine groundwater zones in Australia. *Ecosystem Health*, **6**, 108–118.

Boyle D.R. (1994) Design of a seepage meter for measuring groundwater fluxes in the nonlittoral zones of lakes–Evaluation in a boreal forest lake. *Limnology and Oceanography*, **39**, 670–681.

Bredehoeft, J.D., Papadopulos, S.S. and Cooper, H.H., Jr. (1982) The water-budget myth. In *Scientific Basis of Water Resources Management*, National Academy Press: Washington, pp. 51–57.

Brunke M. and Gonser T. (1997) The ecological significance of exchange processes between rivers and groundwater. *Freshwater Biology*, **37**, 1–33.

Budyko M.I. and Sokolov A.A. (1978) Water balance of the earth. In *World Water Balance and Water Resources of the Earth*, Korzun V.I. (Ed.) UNESCO, Gidrometeoizdat: Leningrad, pp. 586–591.

Burnett B., Chanton J., Christoff J., Kontar E., Krupa S., Lambert M., Moore W., O'Rourke D., Paulsen R. and Smith C. *et al.* (2002) Assessing methodologies for measuring groundwater discharge to the ocean. *Eos Transactions–American Geophysical Union*, **83**, 117–123.

Chahine M.T. (1992) The hydrological cycle and its influence on climate. *Nature*, **359**, 373–380.

Cook P.G. and Böhlke J.-K. (1999) Determining timescales for groundwater flow and solute transport. In *Environmental Tracers in Subsurface Hydrology*, Cook P.G. and Herczeg A.L. (Eds.), Kluwer Academic Publishers: Boston, pp. 1–30.

Cunningham W.L. (2001) *Real-Time Ground-Water Data for the Nation*, U.S. Geological Survey: Fact Sheet 090-01.

Dahm C.N., Grimm N.B., Marmonier P., Valett M.H. and Vervier P. (1998) Nutrient dynamics at the interface between surface waters and groundwaters. *Freshwater Biology*, **40**, 427–451.

Doss P.K. (1993) Nature of a dynamic water table in a system of non-tidal, freshwater coastal wetlands. *Journal of Hydrology*, **141**, 107–126.

Fetter C.W. (1988) *Applied Hydrogeology*, *Second Edition*, Merrill Publishing: Columbus.

Focazio M.J., Reilly T.E., Rupert M.G. and Helsel D.R. (2002) *Assessing Ground-Water Vulnerability to Contamination: Providing Scientifically Defensible Information for Decision Makers*, U.S. Geological Survey: Circular 1224. Also available on the World Wide Web at **http://pubs.water.usgs.gov/circ1224**.

Foster S.S.D. and Chilton P.J. (2004) Downstream of downtown: Urban wastewater as groundwater recharge. *Hydrogeology Journal*, **12**, 115–120.

Fraser C.J.D., Roulet N.T. and Lafleur M. (2001) Groundwater flow patterns in a large peatland. *Journal of Hydrology*, **246**, 142–154.

Freeze R.A. and Cherry J.A. (1979) *Groundwater*, Prentice-Hall: Englewood Cliffs.

Galloway D.L., Alley W.M., Barlow P.M., Reilly T.E. and Tucci P. (2003) *Evolving Issues and Practices in Managing Ground-Water Resources: Case Studies on the Role of Science*, U.S. Geological Survey Circular 1247. Also available on

the World Wide Web at **http://pubs.water.usgs.gov/circ1247**.

Halford K.J. and Mayer G.C. (2000) Problems associated with estimating ground water discharge and recharge from stream-discharge records. *Ground Water*, **38**, 331–342.

Harvey J.W. and Bencala K.E. (1993) The effect of streambed topography on surface-subsurface water exchange in mountain catchments. *Water Resources Research*, **29**, 89–98.

Heath R.C. (1983) *Basic Ground-Water Hydrology*, U.S. Geological Survey: Water-Supply Paper 2220.

Hinkle S.R., Duff J.H., Triska F.J., Laenen A., Gates E.B., Bencala K.E., Wentz D.A. and Silva S.R. (2001) Linking hyporheic flow and nitrogen cycling near the Willamette River – A large river in Oregon, U.S.A. *Journal of Hydrology*, **244**, 157–180.

Holzbecher E. (2001) The dynamics of subsurface water divides–Watersheds of Lake Stechlin and neighboring lakes. *Hydrological Processes*, **15**, 2297–2304.

Hutchinson G.E. (1957) *A Treatise on Limnology: Volume 1, Geography, Physics, and Chemistry*, John Wiley & Sons: New York.

Johnston R.H. (1999) *Hydrologic Budgets of Regional Aquifer Systems of the United States for Predevelopment and Development Conditions*, U.S. Geological Survey: Professional Paper 1425.

Jones J.A. and Mulholland P.J. (Eds.) (2000) *Streams and Ground Water*, Academic Press: San Diego.

Kalle K. (1945) *Der Stoffhaushalt des meeres*, Akademische Verlagsgesellschaft: Leipzig.

Kendy E. (2003) The false promise of sustainable pumping rates. *Ground Water*, **41**, 2–4.

Kooi H. and Groen J. (2001) Offshore continuation of coastal groundwater systems: predictions using sharp-interface approximations and variable density flow modeling. *Journal of Hydrology*, **246**, 19–35.

Lahm T.D. and Bair E.S. (2000) Regional depressurization and its impact on the sustainability of freshwater resources in an extensive mid-continent variable-density aquifer. *Water Resources Research*, **36**, 3167–3177.

Larkin R.G. and Sharp J.M. Jr. (1992) On the relationship between river-basin geomorphology, aquifer hydraulics, and ground-water flow direction in alluvial aquifers. *Geological Society of America Bulletin*, **104**, 1608–1620.

Lee T.M. (2002) *Factors Affecting Ground-Water Exchange and Catchment Size for Florida Lakes in Mantled Karst Terrain*, Water-Resources Investigations Report 02–4033, U.S. Geological Survey.

Lerner D.N. (2002) Identifying and quantifying urban recharge: a review. *Hydrogeology Journal*, **10**, 143–152.

Lohman S.W. (1972) *Ground-Water Hydraulics*, U.S. Geological Survey: Professional Paper 708.

Lohman S.W. *et al.* others (1972) *Definitions of Selected Ground-Water Terms – Revisions and Conceptual Refinements*, U.S. Geological Survey: Water-Supply Paper 1988.

L'vovich M.I. (1979) *World Water Resources and their Future*, English Translation Nace R.L. (Ed.) American Geophysical Union, Lithocrafters Inc., Chelsea.

Manga M. (2001) Using springs to study groundwater flow and active geologic processes. *Annual Review of Earth and Planetary Sciences*, **29**, 201–228.

Meyboom P. (1967) Mass-transfer studies to determine the groundwater regime of permanent lakes in the hummocky moraine of western Canada. *Journal of Hydrology*, **5**, 117–142.

Nace R.L. (1967) *Are We Running Out of Water?*, U.S. Geological Survey. Circular 536.

National Research Council (2004) *Groundwater Fluxes Across Interfaces*, National Academies Press, Washington.

Power G., Brown R.S. and Imhof J.G. (1999) Groundwater and fish–Insights from northern North America. *Hydrological Processes*, **13**, 401–422.

Rayne T.W., Bradbury K.R. and Muldoon M.A. (2001) Delineation of capture zones for municipal wells in fractured dolomite, Sturgeon Bay, Wisconsin, USA. *Hydrogeology Journal*, **9**, 432–450.

Reilly, T.E. (2001) *System and Boundary Conceptualization in Ground-Water Flow Simulation*, Chap. B8, Techniques of Water-Resources Investigations of the United States Geological Survey: Book 3.

Rutledge A.T. (1993) *Computer Programs for Describing the Recession of Ground-Water Discharge and for Estimating Mean Ground-Water Recharge and Discharge from Streamflow Records*, Water-Resources Investigations Report 93–4121, U.S. Geological Survey.

Sahagian D.L., Schwartz F.W. and Jacobs D.K. (1994) Direct anthropogenic contributions to sea level rise in the twentieth century. *Nature*, **367**, 54–57.

Sanford W.E. and Buapeng S. (1996) Assessment of a groundwater flow model of the Bangkok Basin, Thailand, using carbon-14-based ages and paleohydrology. *Hydrogeology Journal*, **4**, 26–40.

Schindler D.W., Newberry R.W., Beaty K.G. and Campbell P. (1976) Natural water and chemical budgets for a small Precambrian lake basin in central Canada. *Journal of the Fisheries Research Board of Canada*, **33**, 2526–2543.

Schlesinger W.H. (1997) *Biogeochemistry: An Analysis of Global Change*, *Second Edition*, Academic Press: San Diego.

Shaw R.D. and Prepas E.E. (1990) Groundwater-lake interactions: nearshore seepage patterns and the contribution of ground water to lakes in Central Alberta. *Journal of Hydrology*, **119**, 121–136.

Shelton J.L., Burow K.R., Belitz K., Dubrovsky N.M., Land M. and Gronberg J.M. (2001) *Low-Level Volatile Organic Compounds in Active Public Supply Wells as Ground-Water Tracers in the Los Angeles Physiographic Basin, California, 2000*, Water-Resources Investigations Report 01–4188, U.S. Geological Survey.

Sloan C.E. (1972) *Ground-Water Hydrology of Prairie Potholes in North Dakota*, U.S. Geological Survey Professional Paper 585-C.

Smith A.J. and Townley L.R. (2002) Influence of regional setting on the interaction between shallow lakes and aquifers. *Water Resources Research*, **38**, 10-1–10-13.

Sophocleous M. (2000) From safe yield to sustainable development of water resources: the Kansas experience. *Journal of Hydrology*, **235**, 27–43.

Taylor C.J. and Alley W.M. (2001) *Ground-Water-Level Monitoring and the Importance of Long-Term Water-Level Data*, U.S.

Geological Survey Circular 1217. Also available on the World Wide Web at **http://pubs.water.usgs.gov/circ1217**.

Theis C.V. (1940) The source of water derived from wells: essential factors controlling the response of an aquifer to development. *Civil Engineering*, **10**, 277–280.

United Nations Environment Programme (2003) *Groundwater and its Susceptibility to Degradation*. Accessed March 18, 2004 at **http://www.unep.org/DEWA/water/groundwater**.

Vörösmarty C.J. and Sahagian D. (2000) Anthropogenic disturbance of the terrestrial water cycle. *BioScience*, **50**, 753–765.

Voskresensky K.P. (1978) Water of the Earth. In *World Water Balance and Water Resources of the Earth*, Korzun V.I. (Ed.), UNESCO, Gidrometeoizdat: Leningrad, pp. 42–56.

White W.B. (1993) Analysis of karst aquifers. In *Regional Ground-Water Quality*, Alley W.M. (Ed.) Van Nostrand Reinhold: New York, pp. 471–489.

Winter T.C. (1976) *Numerical Simulation Analysis of the Interaction of Lakes and Ground Water*, U.S. Geological Survey: Professional Paper 1001.

Winter T.C., Harvey J.W., Franke O.L. and Alley W.M. (1998) *Ground Water and Surface Water – A Single Resource*, U.S. Geological Survey Circular 1139. Also available on the World Wide Web at **http://pubs.water.usgs.gov/circ1139**.

Winter T.C., Rosenberry D.O. and LaBaugh J.W. (2003) Where does the ground water in small watersheds come from? *Ground Water*, **41**, 989–1000.

Woessner W.W. (2000) Stream and fluvial plain ground water interactions: rescaling hydrogeologic thought. *Ground Water*, **38**, 423–429.

Wroblicky G.J., Campana M.E., Valett H.M. and Dahm C.N. (1998) Seasonal variation in surface-subsurface water exchange and lateral hyporheic area of two stream-aquifer systems. *Water Resources Research*, **34**, 317–328.

York J.P., Person M., Gutowski W.J. and Winter T.W. (2002) Putting aquifers into atmospheric simulation models: an example from the Mill Creek Watershed, northeastern Kansas. *Advances in Water Resources*, **25**, 221–238.

Zektzer I.S. and Kudelin B.I. (1966) The methods of determining ground water flow to lakes with special reference to Lake Ladoga. *Proceedings of the International Association of Scientific Hydrology, Symposium of Garda*, *Hydrology of Lakes and Reservoirs Volume I*, Publication Number 70, International Association of Scientific Hydrology: Gentbrugge, pp. 31–38.

Zhu C., Waddell R.K., Star I. and Ostrander M. (1998) Responses of groundwater in the Black Mesa Basin, northeastern Arizona to paleoclimatic changes during the late Pleistocene and Holocene. *Geology*, **26**, 127–130.

146: Aquifer Recharge

JOHN R NIMMO[1], RICHARD W HEALY[2] AND DAVID A STONESTROM[1]

[1] United States Geological Survey, Menlo Park, CA, US
[2] United States Geological Survey, Denver, CO, US

Aquifer recharge is important both for hydrologic understanding and for effective water resource management. Temporal and spatial patterns of unsaturated-zone processes such as infiltration largely determine its magnitude. Many techniques of recharge estimation exist. Water budget methods estimate all terms in the continuity equation except recharge, which is calculated as the residual. Detailed hydrologic models based on water-budget principles can produce recharge estimates at various scales. Empirical methods relate recharge to meteorologic and geographic parameters for a specific location. Surface-water methods include stream-hydrograph analyses to estimate baseflow (groundwater discharge) at lower elevations in a watershed, which is taken to equal the recharge that has occurred at higher elevations. Subsurface methods include analysis of water-table fluctuations following transient recharge events, as well as diverse unsaturated-zone methods. The zero-flux plane method determines the recharge rate from the change in water storage beneath the zero-flux depth, a boundary between water moving upwards due to evapotranspiration and water moving downward due to gravity. Lysimeter methods use buried containers filled with vegetated soil to mimic natural conditions. Water exiting the bottom is considered to be recharge. Darcian methods for estimating flux densities use unsaturated hydraulic conductivities and potential gradients, indicating recharge rates under appropriate conditions. Chemical mass-balance methods use conservative tracers that move with recharging water. Tracer concentrations in deep unsaturated-zone water, together with tracer input rates, indicate recharge rates. Distinct chemical "markers" can indicate travel times, hence, recharge rates. Thermal methods use heat as a tracer. Moving water perturbs temperature profiles, allowing recharge estimation. Geophysical methods estimate recharge based on water-content dependence of gravitational, seismic, and electromagnetic properties of earth materials.

INTRODUCTION

Defined as water that moves from the land surface or unsaturated zone into the saturated zone, aquifer recharge is vitally important for understanding the hydrologic cycle as well as for applications to water-resource management. The definition used here excludes saturated flow between aquifers so it might be more precisely termed "aquifer-system" or "saturated-zone" recharge. This definition avoids double-accounting in large-scale studies. Recharge is commonly expressed as a volume [L^3], typical units being m^3 or acre-ft. Recharge rate expresses either a flux [L^3T^{-1}] into a specified portion of aquifer, or a flux density [LT^{-1}] (volume per unit surface area) into an aquifer at a point. Over the long term, recharge naturally balances the total losses of water from the saturated zone. In some systems that are characterized by low permeabilities and sensitivity to climatic change, this long-term balance can be discerned only on millennial timescales.

Because it represents replenishment of aquifers critical to maintaining water supplies and ecosystems, recharge has obvious practical importance, especially where ground water is extracted for human use. It is a vital component for evaluating sustainability, as streamflow is for surface water. In the hydrologic balance of an aquifer, recharge processes act in opposition to discharge processes, so the relative magnitudes of recharge and discharge rate give a basic indication of the health of the aquifer and related systems.

Quantitative estimation of recharge rate contributes to the understanding of large-scale hydrologic processes. It

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

is important for evaluating the sustainability of groundwater supplies, though it does not equate with a sustainable rate of extraction (Bredehoeft *et al.*, 1982). Where contamination of an aquifer is a concern, the flux of water into the aquifer is essential information; estimating the recharge rate is a first step toward predicting the rate of solute transport to the aquifer. In cases where advection dominates the transport of contaminants (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**), little additional hydraulic information may be needed to estimate travel times and solute fluxes. Moreover, recharging fluxes and their distribution need to be known for assessment of aquifer vulnerability to contamination, prediction of zones of significant contamination, and evaluation of remedial measures.

Our aim in this article is to provide an understanding of recharge in terms of its processes, its role as a component of the hydrologic cycle, and means for its local or regional estimation. We approach it mainly from the perspective of natural science, emphasizing recharge from natural processes, but also with attention to anthropogenic sources such as irrigation, wastewater disposal, and deliberate augmentation of groundwater. The discussion of estimation techniques makes up the largest part of this article, and we have included much description and explanation of basic recharge processes within this discussion.

RECHARGE IN NATURE AND APPLICATIONS

Sources and Basic Processes

The nature of the water source to a large extent determines which subsurface processes are relevant or dominant in the transport of water to the aquifer. In addition to precipitation, which is commonly the dominant source, other possible sources include surface-water bodies, irrigation, and artificial recharge.

In some cases, perennial or ephemeral surface water in the form of rivers, canals, and lakes may enter the aquifer directly (Winter *et al.*, 1998). Surface-water bodies, however, are not always recharge sources; they can instead be associated with aquifer discharge. A reach of a stream is considered "gaining" or "losing" depending on whether aquifer discharge or recharge dominates over that reach. If there is no unsaturated zone, recharge and discharge may proceed by analogous processes, in each case dictated by the net driving force (usually a combination of gravity and the pressure gradient) at the bottom of the surface-water body. Surface water in direct contact with the saturated zone can be treated in terms of surface water-groundwater interaction, for example, with stream-gauging. This is described further below and in **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4**. In some cases, especially in drier climates, water from a body of surface water first travels through the unsaturated zone before reaching the aquifer.

Where recharge occurs from precipitation or other sources through an unsaturated zone, general expectations are (i) that gravity is the dominant driving force for recharge (as matric-pressure gradients would typically act to oppose recharge or would balance out over time) and (ii) that the hydraulic conductivity is markedly less in the unsaturated than in the saturated zone. It is not always possible to make a direct analogy between the recharge and discharge processes as with gaining and losing streams. Capillary forces may draw aquifer water upward into the unsaturated zone, but its future transport and ultimate discharge from the aquifer occur to a large extent through saturated-zone processes. Thus unsaturated-zone processes are critical to the evaluation of recharge in practical circumstances.

Deliberate or artificial-recharge operations frequently employ surface-water ponds or streams to put water into the subsurface, though in some applications the water for recharge is injected at significant depth. Natural and artificial recharge mostly involve the same types of processes, though artificial recharge may easily create situations that would be unusual in nature, for example, a thick unsaturated zone under a pond. Irrigation and other water-use practices can contribute to recharge, typically in ways analogous to precipitation or surface water, depending on the mode of application.

Variation

Recharge rates vary considerably in time and space. Recharge often occurs episodically in response to storms and other short-term, high-intensity inputs. Some, perhaps most, of the water that becomes recharge may be spatially concentrated in narrow channels as it passes through the unsaturated zone. Time- and space-concentration are important because for available water at the land surface, there is competition among alternative possible fates such as recharge, runoff, and evapotranspiration. For a given amount of infiltration, temporal concentration favors recharge because it entails shorter residence times for water in the portions of the soil from which evapotranspiration takes place. Similarly, a larger fraction will become recharge if it is concentrated in narrow channels such as fingers or macropores, not only because this tends to hasten its passage through the unsaturated zone, but also because the water then occupies less of the volume of soil from which evapotranspiration takes place.

Temporal Distribution of Recharge

Temporal variation occurs in general with seasonal or short-term variations in precipitation, runoff, or evapotranspiration. Evapotranspiration, in particular, is important because

it may extract most or all of the water that infiltrates. The time of greatest recharge is normally when much water is present, so that the processes other than recharge, especially evapotranspiration, are overwhelmed. Because flow rates are much greater in soil at high than at low water content (*see* **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**), the downward flow important for recharge is much greater at high water contents. The slower vertical flow occurring with moderate or low water contents allows more water to remain for extraction from the soil without recharge.

Temporal concentration commonly occurs during storms, floods, and snowmelt. Variability is especially evident in thin unsaturated zones, where recharge may occur within a short time, often much less than one year (e.g. Delin *et al.*, 2000), and sometimes much less than one day (e.g. Gburek and Folmar, 1999). In deep unsaturated zones, recharge may be homogenized over many years such that it occurs with constant flux even though fluxes at shallow depths are erratic. Figure 1 illustrates temporal effects with an example from Dreiss and Anderson (1985).

Variation over time can also be significant over a longer term. Climatic change, for example, can systematically alter recharge patterns over hundreds or thousands of years. Sometimes, especially in arid regions with thick unsaturated zones, evidence for such change may be present in the observable distributions of water and other substances within the unsaturated zone (Tyler *et al.*, 1996).

Spatial Distribution of Recharge

Spatial variation occurs with climate, topography, soil, geology, and vegetation. For example, a decrease of slope or increase of soil permeability may lead to greater infiltration and greater recharge.

Topography influences recharge in several ways. It controls the generation of runoff. Steep slopes, especially if they are convex, may decrease infiltration and recharge. Swales and other concavities that tend to collect surface water may increase infiltration and recharge. Spatial concentration and increased recharge typically occur in depressions and channels, where higher water contents promote rapid movement by increasing the hydraulic conductivity (K), the amount of preferential flow, and the downward driving force at a wetting front.

Vegetation influences recharge mainly through its water-distributing activity. Where water is a significant limiting factor of vegetation growth, roots usually extract most of the water that infiltrates. This may be the case in all but the most humid climates. The result is that evapotranspiration may nearly equal infiltration, leaving recharge as a small fraction of infiltration. Artificial control of vegetation may counter natural influences. Agricultural practices, for example, may leave land fallow part of the time, decreasing total annual evapotranspiration from what it would be naturally. The depth range over which roots can extract water is frequently an important issue. Depending on the plant species, soil, and climate, roots may be present and active at depths of 10 m or more.

Another effect of evapotranspiration is that by drying the root zone, it enhances gradients that may be drawing water upward from the water table, in effect causing discharge through the unsaturated zone. The depth of the water table is a major influence in this process. Such discharge is likely to be greater if the water table is within or near the root

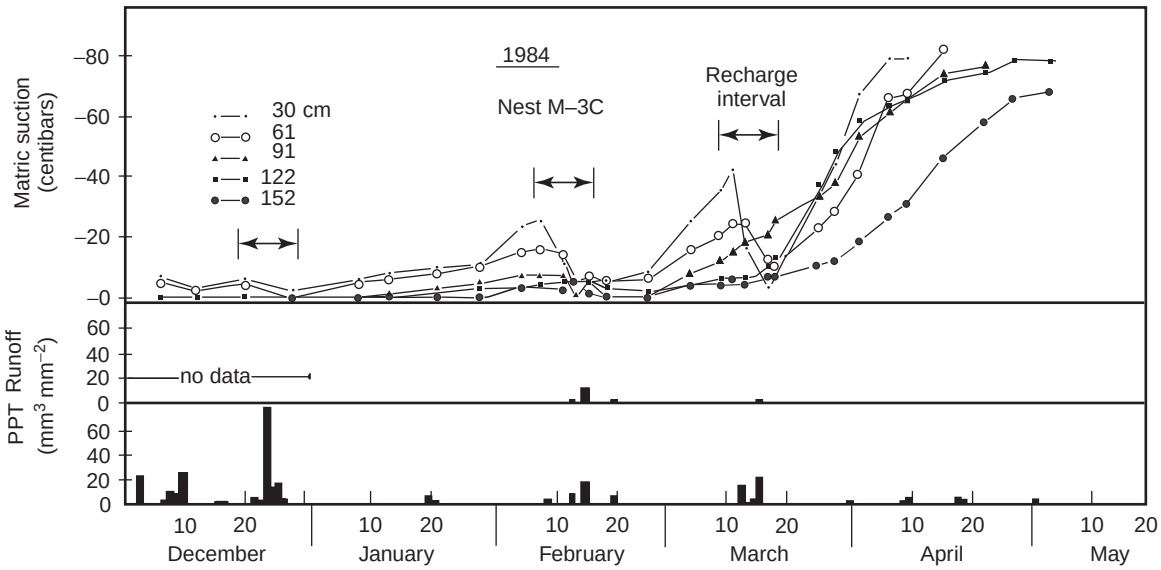


Figure 1 Measured precipitation (PPT), runoff, and matric pressure at several depths for a five-month period on a marine terrace (Reprinted from Ground Water, the National Ground Water Association. Copyright 1985)

zone. The process is thus affected by water-table variation, whether caused by natural processes or artificial recharge or withdrawal.

The character of the subsurface is also a major influence on recharge. Fractures, preferential flow, karst features, and so on, can lead to enhanced recharge by increasing spatial concentration. Features that act to hold water near the surface tend to increase evapotranspiration, and decrease recharge. For example, stratification of the unsaturated zone retards downward flow. Soil water hysteresis also acts during redistribution of water in the unsaturated zone to hold water high in the profile; therefore soils with more pronounced hysteresis of soil water retention may have greater evapotranspiration and less recharge (*see* **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**). The medium also affects the thickness of the zone most affected by evapotranspiration; the presence of bedrock at shallow depth, for example, may confine roots to a thin zone and allow more water to escape their influence by flowing into the bedrock.

Recharge Distribution and Applications

The importance of the variability of recharge in time and space depends on the nature of the application as well as the processes. Water supply issues related to depletion of an aquifer as a whole usually make use of averages over a large area, generally an entire watershed or aquifer. Usually this type of application also requires a time average suitable for predicting long-term sustainability. Thus the period of averaging may include centuries since the last major climatic change. Issues of contaminant transport from a point source usually need a point or highly localized estimate of recharge rate, for a time period about as long as the contaminant would remain present in significant quantities. Aquifer vulnerability assessments and diffuse-source contamination (e.g. from agricultural pesticides) often require knowledge of the distribution of recharge rates over a specified area with high spatial resolution.

Certain time and space scales are inherent in the relevant conceptual models to be chosen or developed. For example, with a deep unsaturated zone in a granular medium, virtually all temporal fluctuations may be averaged out before water enters the saturated zone, making a decadal or longer scale appropriate. If there is a predominance of fracture flow or a shallow unsaturated zone, however, the conceptual model might need to allow for variations on a daily basis.

Methods of estimating recharge produce estimates pertaining to different time and space scales. Many methods fundamentally assess recharge in a way that is highly localized or specific in time and space while others produce recharge estimates that are temporally or areally averaged. For point applications, areal estimates will generally not be suitable unless spatial variability is slight, or unless it is possible to delineate features on the ground that might permit evaluation of likely deviations from the areal average. Additional data relating to the finer scale would be needed to estimate recharge at a given point. For areal applications when the available recharge estimates are point estimates, the area of interest can be divided into subareas within which recharge is more likely to be uniform, and an integrated estimate may be produced using point data in each subarea. Thus, for a given application, a good conceptual model of the hydrologic system and its recharge processes is essential for selecting recharge estimation methods, and for integrating point measurements to represent an area.

ESTIMATION OF RECHARGE

Because of the extreme variability of recharge rates and the fact that many relevant processes cannot be directly observed, recharge is usually difficult to estimate, especially when it involves unsaturated flow. There is no single instrument with the simplicity of a thermometer or pressure gauge that measures unsaturated-zone flux, so estimates are normally based on a combination of various types of data. Particular difficulties in the unsaturated zone are the extreme nonlinearity of basic unsaturated hydraulic properties, the typically extreme heterogeneity of these properties and other relevant features of the unsaturated zone, and the frequent occurrence of preferential flow (*see* **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**). In addition, important processes that compete for water in the unsaturated zone, such as evapotranspiration, often are also difficult to evaluate. A quantitative estimate of recharge may depend on a combination of data that themselves have substantial uncertainty. For these reasons it is wise to apply multiple methods and compare their results.

Much software for simulating surface water, groundwater, and unsaturated-zone transport has been developed that is useful in investigations of aquifer recharge. Public-domain software related to various methods described in this article is available at **`http://water.usgs.gov/software/index.html.`**

Water-budget Methods

Water-budget methods comprise the largest class of techniques for estimating recharge. The water budget is an integral component of any conceptual model of the system under study. Hence, analysis of the water budget is of great value in any recharge study. Simple water budgets can be readily constructed and refined as needed. As noted by Lerner *et al.* [1990], a good method for estimating recharge should explicitly account for all water that does not become recharge.

The water balance for a control volume, such as an aquifer or a watershed, can be stated as (Scanlon *et al.*,

2002)

$$P + Q_{on}^{sw} + Q_{on}^{gw} = ET^{sw} + ET^{uz} + ET^{gw} + Q_{off}^{sw} + Q_{off}^{gw} + Q_{bf} + \Delta S^{snow} + \Delta S^{sw} + \Delta S^{uz} + \Delta S^{gw} \quad (1)$$

where superscripts refer to surface water, groundwater, unsaturated zone, or snow, P is precipitation and irrigation; Q_{on} and Q_{off} are water flow into and out of the basin; Q_{off}^{sw} is surface-water runoff; Q_{bf} is baseflow (groundwater discharge to streams or springs); ET is evapotranspiration; and ΔS is change in water storage. Units may be any of those presented earlier for recharge. Following our definition, as water that enters the saturated zone, recharge can be written as (Schicht and Walton, 1961):

$$R = \Delta Q^{gw} + Q_{bf} + ET^{gw} + \Delta S^{gw} \quad (2)$$

where R is recharge and ΔQ^{gw} is the difference between groundwater flow out of and into the basin. This equation simply states that water arriving at the water table (i) flows out of the basin as groundwater flow, (ii) discharges to the surface, (iii) is lost to evapotranspiration, or (iv) remains in the subsurface, augmenting storage. Substitution into (1) results in another useful form of the water balance:

$$R = P + Q_{on}^{sw} - Q_{off}^{sw} - ET^{sw} - ET^{uz} - \Delta S^{snow} - \Delta S^{sw} - \Delta S^{uz} \quad (3)$$

Water-budget methods include all techniques based, in one form or another, on one of these water balance equations.

Applications of various forms of equations 1 to 3 abound in the literature. An attractive feature of these methods is their flexibility; applications are often simplistic, but can be quite complex. Derivation of equations 1 to 3 requires few assumptions on the mechanisms that control individual components. Hence, through careful adaptation the equations can be applied over a wide range of space and timescales. For example, the water budget equation can be applied to study water movement in a soil lysimeter at scales of centimeters and seconds. At the same time, the water-budget equation is an integral part of Global Climate Models (GCMs) that predict global climate changes over periods of centuries.

The most common way of obtaining an estimate of recharge by these methods is the indirect or "residual" approach whereby all of the variables in the water budget equation, except R, are either measured or estimated and R is set equal to the residual. The accuracy of the recharge estimate is then dependent on the accuracy with which the other water-budget components can be measured or estimated. Consider a simple example where the only significant components of equation 3 are recharge, precipitation, and unsaturated zone ET and ΔS. Using the residual approach, we can write:

$$(R + \varepsilon_R) = (P + \varepsilon_P) - (ET^{uz} + \varepsilon_{ET}) - (\Delta S^{uz} + \varepsilon_S) \quad (3a)$$

Where the ε terms represent measurement errors. The magnitude of the error in the recharge estimate, ε_R, could be as large as $|\varepsilon_P + \varepsilon_{ET} + \varepsilon_S|$. In arid settings where R is generally a small fraction of P or ET, $|\varepsilon_R|$ can exceed the true value of R, even if the other error terms are relatively small. This can be a serious limitation on application of the residual approach. Timescales for application of water-budget methods are important; more frequent tabulations are likely to improve accuracy. Averaging over longer time periods tends to dampen out extreme precipitation events and hence underestimate recharge. For example, in an arid region, potential ET averaged over a week or a month will usually greatly exceed average P over the same period. But on a daily basis, P can exceed potential ET – a necessity for recharge to occur. Narayanpethkar *et al.* (1994) used a water-budget method to derive an estimate of 23 mm yr^{-1} of recharge in a region of India; Steenhuis *et al.* (1985) estimated recharge to be 400 mm at a site in the eastern US.

Direct water-budget approaches are based on measurements of changes in water storage within some compartment (such as a soil column) and partitioning of those changes to recharge and possibly to other components such as evapotranspiration. Included in this class are the zero-flux plane method and the use of lysimeters, both discussed in a subsequent section.

Modeling Methods

Hydrologic models constitute an important class of water-budget methods. The commonly used groundwater and surface-water flow equations are, in fact, variations of equations 1 to 3. Several modeling approaches have been used to estimate recharge.

Watershed models have been applied at a variety of spatial scales. Some of these models provide a single, lumped value of recharge for the watershed. Others allow disaggregation of the watershed into distinct hydrologic zones that may receive varying amounts of recharge. Watershed models generally require daily climatic data and information on land-surface features and land use. Streamflow data, and less frequently groundwater data, are used to calibrate the models. Bauer and Mastin (1997) used a watershed model to estimate recharge in three small watersheds (average drainage area 0.4 km^2) in western Washington. On the other end of the spatial scale, Arnold *et al.* (2000) applied the SWAT model to the entire upper Mississippi River Basin (492 000 km^2). Recharge estimates in different parts of that basin were between 10 and 400 mm yr^{-1}. Figure 2

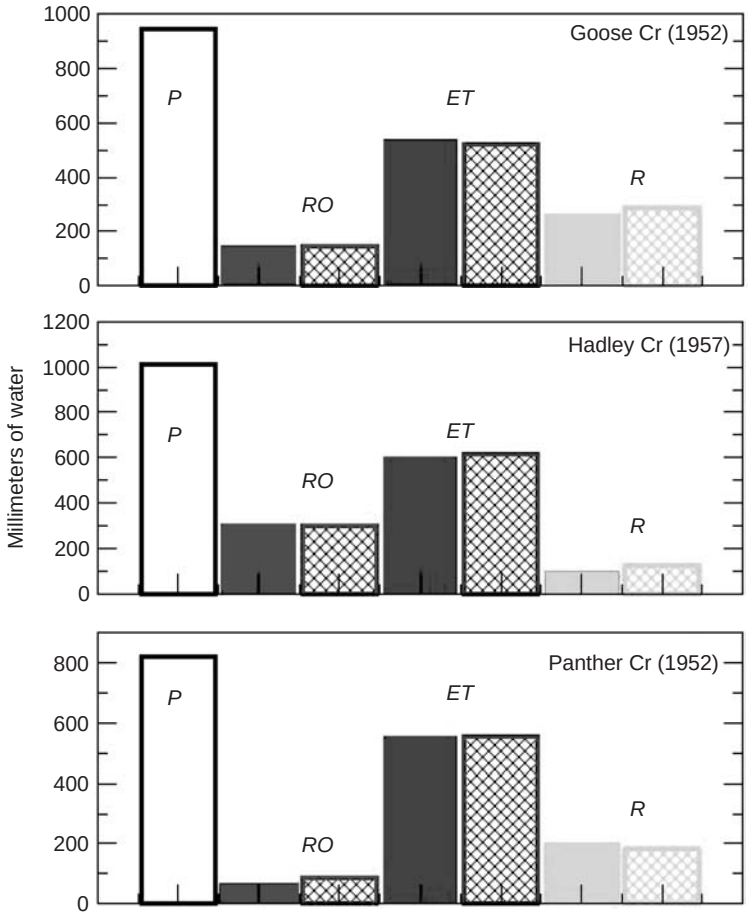


Figure 2 Water-budget components for three watersheds in Illinois: actual (Schicht and Walton, 1961), shown as solid bars; and predicted with the SWAT model (Arnold and Allen, 1996), shown as hatched bars. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

shows one-year water budgets, including recharge, for 3 small streams in Illinois as measured by Schicht and Walton (1961) and as predicted by Arnold and Allen (1996) using the SWAT model.

Recharge is an integral component of groundwater flow models, and estimates of recharge can be generated through model calibration. Recent incorporation of automatic inverse modeling techniques (Hill *et al.*, 2000) to these models has facilitated this process. A concern with these inverse techniques, however, is the possibility that they could generate nonunique solutions (i.e. different parameter values could produce identical results). This situation can be avoided by calibrating with data on both groundwater head and flux, or by independent determination of hydraulic conductivity. Tiedeman *et al.* (1998) developed estimates of recharge for the Albuquerque Basin in New Mexico using this approach.

One-dimensional models of vertical flow through the unsaturated zone have also been used to estimate recharge rates. As with watershed models, daily climate data are usually used. Information on hydraulic properties of sediments within the unsaturated zone is also required. The complexity of these models varies widely, from a simple reservoir (bucket) model to analytical or numerical solutions to Richards' equation. Example applications include those of Scanlon and Milly (1994), Kearns and Hendrickx (1998), and Flint *et al.* (2002).

A powerful feature of all models is their predictive capability. They can be used to gauge the effects of future climate or land-use changes on recharge rates. Sensitivity analyses can also be used to identify model parameters that most affect computed recharge rates (*see* **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4** and **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**).

Empirical equations, such as the Maxey-Eakin equation (Maxey and Eakin, 1949), are often used to obtain quick and inexpensive estimates of recharge. These equations, usually developed for specific regions, relate recharge to more easily measured phenomena such as precipitation,

temperature, elevation, vegetation, and soil type. While the accuracy of these equations may not be sufficient for some studies, the simplicity of the approach facilitates application over large areas using a Geographic Information System (GIS) and remotely sensed data.

Methods Based on Surface-water Data

Several methods are available for estimating recharge from data collected in surface-water bodies. Most of these methods require data on stream discharge at multiple points or times. Stream loss or seepage to the subsurface can be determined indirectly as the difference in discharge between a downstream and an upstream measurement point. This method is referred to as the Channel Water Balance Method and is appropriate for application only on losing streams. Stream or transmission loss does not necessarily equate with recharge. Water lost from a stream could remain as bank storage, be consumed by evapotranspiration, or become recharge. Stream loss falls into the category termed "potential" recharge by Lerner *et al.* (1990). Seepage estimates obtained by this method represent a value that is integrated over the length of stream channel between the two measurement points. Surface/groundwater exchange at specific points in a surface-water body can be directly measured with seepage meters (Kraatz, 1977; Lee and Cherry, 1978).

Stream flow hydrographs for gaining streams can be analyzed to estimate what portion of that discharge can be attributed to baseflow (also known as *groundwater discharge*). The rate of groundwater discharge is closely related to the rate of recharge; however, the two are not necessarily equal (equation 2). It is unlikely that all recharge in a basin would eventually be discharged to streams. Recharge may be lost to evapotranspiration, pumpage, or groundwater flow out of the basin. Care should be taken in understanding runoff processes in any given watershed. Some basins are unsuitable for hydrograph analysis. In particular, basins with upstream regulation, significant groundwater withdrawals, topographically flat terrain, karstification, or losing stream reaches should be avoided. Hydrograph analysis is usually performed for relatively humid regions where streams are perennial, with well-sustained base flow. Analytical techniques include hydrograph separation (Sloto and Crouse, 1996), chemical hydrograph separation (Hooper *et al.*, 1990), recession-curve displacement (Rorabaugh, 1964; Rutledge, 1998), and analysis of flow-duration curves (Kuniansky, 1989). Using an automated hydrograph recession-curve analysis computer program, Rutledge and Mesko (1996) analyzed discharge records from 89 basins in the eastern US and obtained recharge estimates between 152 and 1270 mm per year. Daniel and Harned (1998) analyzed 161 water years of record from 16 surface-water stations with the local minimum method (Pettyjohn and Henning, 1979) and recession-curve displacement and found that the former method provided estimates that were on average 21% less than those of the recession-curve displacement method.

Methods Based on Groundwater Data

Information on groundwater levels and how those levels vary in time and space can be used to estimate recharge. The most widely used of these methods is the water-table fluctuation (WTF) method (Healy and Cook, 2002). Other methods in this group include variations of the Darcy Methods (as discussed below), flow net analysis (Cedergren, 1977), and various semiempirical modeling approaches (e.g. Su, 1994; Wu *et al.*, 1997).

The WTF method uses the following equation to estimate recharge to unconfined aquifers:

$$R = S_y \frac{\mathrm{d}h}{\mathrm{d}t} \approx S_y \frac{\Delta h}{\Delta t} \tag{4}$$

where S_y is specific yield, h is water-table height, $\mathrm{d}h/\mathrm{d}t$ is the derivative of water-table height with respect to time, which is approximated as the difference in h (Δh) over the elapsed time between measurements (Δt). Equation 4 is applied only over periods of water-level rise (i.e. Δh is positive). The method works best over short periods for shallow water tables that display sharp water-level rises and declines. The method inherently assumes that recharge occurs only as a result of transient events; recharge occurring under steady flow conditions cannot be estimated.

Despite its simplicity and an abundance of available data, there are difficulties in application of the WTF method (Healy and Cook, 2002). Phenomena other than recharge can induce fluctuations in the water table. These include evapotranspiration, changes in atmospheric pressure, the presence of entrapped air ahead of a wetting front, extraction or injection of water by pumping, temperature effects, and tidal effects (Todd, 1989). Careful examination of water level records in conjunction with climatic data is necessary in order to correctly identify water level rises that can be attributed to recharge events. Determining values for S_y can be problematic. There are many methods for measuring or estimating values of S_y (e.g. laboratory methods, field methods, water-budget methods, and empirical methods), but all of these have a degree of associated uncertainty. Rasmussen and Andreasen (1959) applied this method to a small basin in the eastern US and determined that over a 2-year period annual recharge averaged 541 mm. Figure 3 shows average weekly water levels and precipitation for that study. The dashed lines in Figure 3 are the extrapolated recession curves (traces that the well hydrograph would have followed in the absence of the rise-producing precipitation). Δh is measured as the difference between a hydrograph peak and the recession curve at the time of that peak.

The zero-flux plane (ZFP) is the plane within the subsurface where the vertical hydraulic gradient is zero;

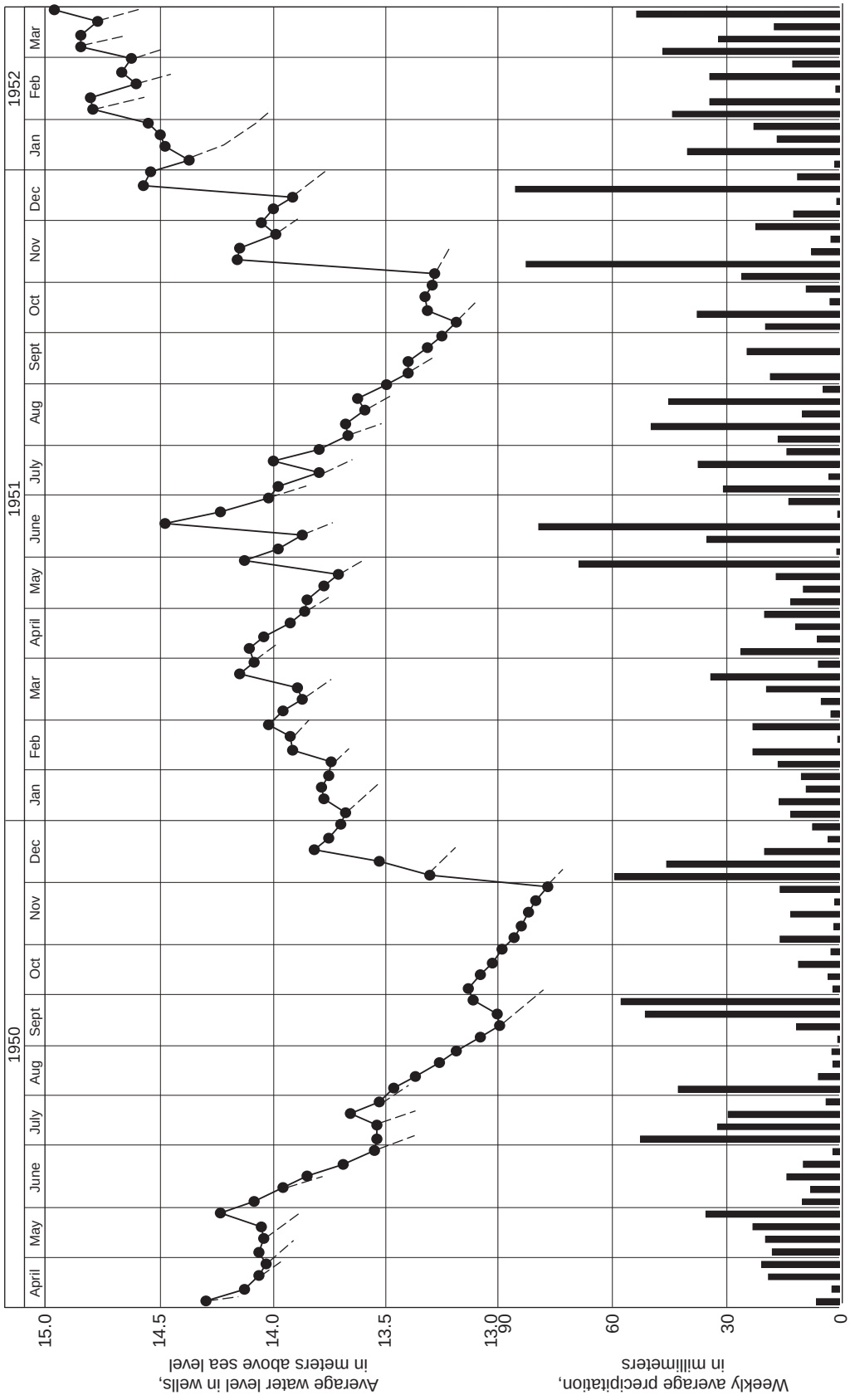


Figure 3 Average water levels in wells and weekly precipitation for Beaverdam Creek basin study of Rasmussen and Andreasen (1959), cited by Healy and Cook (2002)

that is, the plane marks the boundary between water that is moving upwards in response to evapotranspiration and water that is draining downward to become recharge. In applying this method, recharge rate is equated with the rate of change in water storage (drainage) beneath the ZFP (Richards *et al.*, 1956; Roman *et al.*, 1996). Data requirements for this method may be too onerous for some studies. Frequent measurements of soil matric pressure are needed at several depths to accurately locate the ZFP. Soil moisture content must also be measured over the entire thickness of the unsaturated zone. The method cannot be used when water is moving downward throughout the entire profile (i.e. when the ZFP is at land surface). Sharma *et al.* (1991) applied the ZFP method at eight sites in a semiarid region of Western Australia; estimated recharges rates ranged from 34 to 149 mm yr^{-1}.

Lysimeters are containers filled with soil and possibly vegetation and are placed in an environment that mimics natural conditions (Brutsaert, 1982; Young *et al.*, 1996). Fluxes to and from lysimeters are monitored with drains and water-collection vessels or balances that measure changes in weight. The instruments are usually used to measure *ET*. However, if the base of a lysimeter is below the ZFP, drainage from it can be assumed equal to recharge. Kitching *et al.* (1977) used a lysimeter with a surface area of 100 m^2 to measure recharge rates of 342 to 478 mm yr^{-1} for Bunter Sandstone in England.

Darcian Methods

Applied in the unsaturated zone, Darcy's law gives a flux density (q) equal to hydraulic conductivity (K) times the driving force, which equals the recharge rate if certain conditions apply. The basic requirements of the method are to determine K and the driving force, and then to calculate the downward flux density from Darcy's law in a form such as

$$q = -K(\theta)\left(\frac{d\psi}{dz} + \rho g\right) \qquad (5)$$

where θ is the water content, ψ is the matric pressure, z is the vertical coordinate, ρ is the density of water, and g is the gravitational acceleration. K would have SI units of $m^2\,Pa^{-1}\,s^{-1}$. Another common form of equation (5), with ψ in head units, has the gravitational term equal to unity. **Chapter 150, Unsaturated Zone Flow Processes, Volume 4** gives additional information about equation (5) and its variables. To use (5) in estimating recharge, accurate measurements are necessary to know $K(\theta)$ adequately under field conditions at the point of interest. For the total driving force, gravity being known, it is essential to determine the matric-pressure gradient or to demonstrate that it is negligible. For purposes requiring areal rather than point estimates, additional interpretation and calculation are necessary. Even when properly applied, Darcian methods do not necessarily indicate total recharge; some types of preferential flow are inherently non-Darcian and if significant would need to be determined separately.

In the simplest cases, in a region of constant downward flow in a deep unsaturated zone as illustrated in Figure 4(a), gravity alone drives the flow. With a sample from the zone of uniform ψ, K measurements at the original field water content directly indicate the long-term average recharge rate. The depth required for flow to be steady depends on the climate, medium, and vegetation. Fluxes

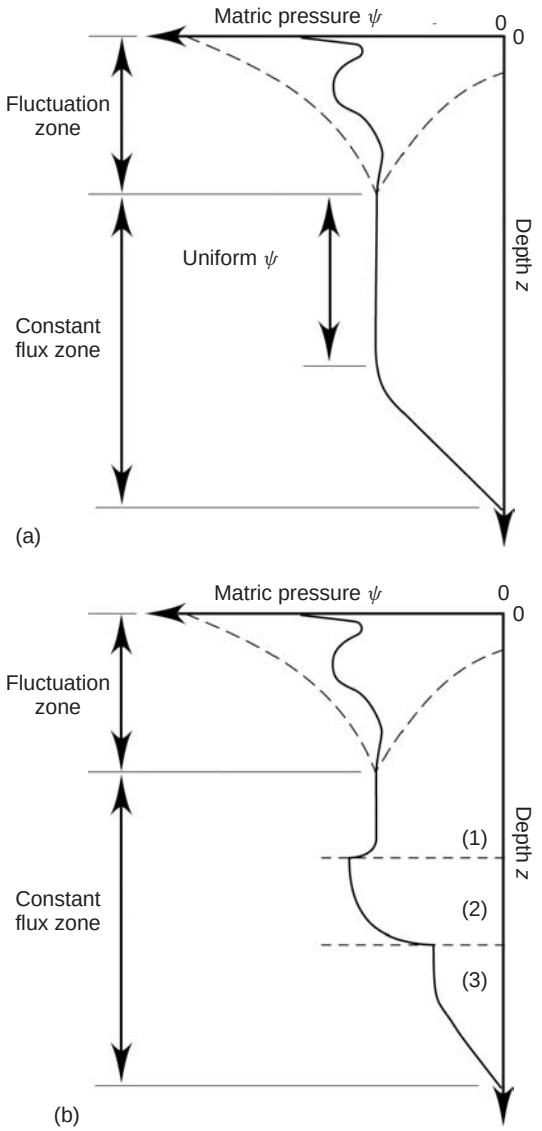


Figure 4 Hypothetical profiles of matric pressure as a function of depth in an unsaturated zone deep enough that its lower portion has a constant downward flux of water in (a) a uniform and (b) a layered profile. Dashed curves indicate possible extremes in the upper portion. The lower horizontal line in each diagram indicates the position of the water table

and water contents fluctuate within the root zone, and possibly to considerably greater depths. Evidence from field experiments (e.g. Nixon and Lawless, 1960) and theoretical assessment of the damping of moisture fluctuations as they move downward (Gardner, 1964) suggest that depths of a few meters are often adequate. In some cases, however, moisture fluctuations have been detected at depths of tens of meters (e.g. Jones, 1978). An obvious way to evaluate steadiness is to measure water content or matric pressure in the zone of interest for a period of a year or more. Where that is not possible, the evaluation of such factors as the homogeneity of the unsaturated zone and the distribution of water within it may help to indicate the degree of steadiness with depth.

In early applications of this technique (Enfield *et al.*, 1973; Stephens and Knowlton, 1986), limited accuracy of K estimates was a significant problem. Accurate laboratory K measurements on core samples, as by the steady-state centrifuge method, are of great value (Nimmo *et al.*, 1994). Figure 5 shows an example of measured hydraulic conductivities with the field water content and its associated K value identified. It happens for sample M30 that two measured $K(\theta)$ points have θ nearly equal to its field value, but even where this is not the case, having measurements at values both higher and lower than the field θ permits the appropriate K to be ascertained by interpolation.

In the examples of Figure 5, as is common, the $K(\theta)$ relation is very steep in the relevant range, meaning that uncertainty in the field θ leads to a greatly magnified uncertainty in the inferred recharge rate. To have reasonable confidence in the recharge estimate, accurate knowledge of the *in situ* water content of the sampled material is essential. In the cases illustrated, it was necessary to work with the water content of the identical soil mass just after sampling, corrected for evaporation between sampling and the first weighing. The oven-dry weight of that same soil must be known, here obtained by drying after K measurements were completed and corrected for losses of dry matter during experimental operations.

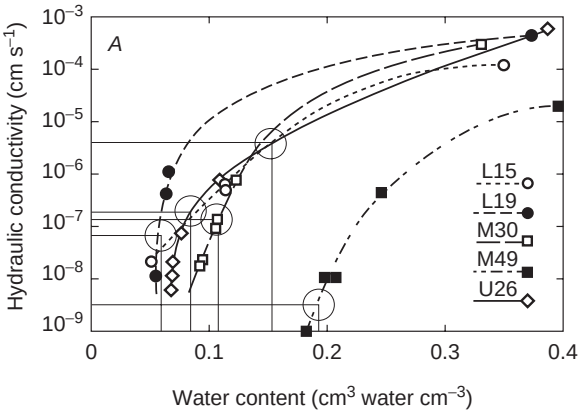


Figure 5 Steady-state centrifuge measurements of hydraulic conductivity versus water content, for five core samples from the unsaturated zone below a desert wash. To indicate recharge flux, the original field water content of each sample is indicated with its corresponding hydraulic conductivity (Reproduced from (Nimmo *et al.*, 2002) by permission of American Geophysical Union)

Variations of the steady-state method are possible for some cases that fall short of the ideal of constant and uniform flow and moisture conditions within a finite portion of the unsaturated zone. One important case is steady flow in a layered medium, in which θ and matric pressure vary spatially as necessary to give K the value needed at each point to maintain a constant Darcian flux density (Figure 4b). Nimmo *et al.* (2002) demonstrated a method for handling this problem using spatially detailed estimates of unsaturated hydraulic properties to compute the uncertainty ascribable to inadequate knowledge of the actual matric-pressure gradient. In such a case one can use geologic data, to the extent possible, to indicate the spatial variability of hydraulic properties.

A more challenging problem is the case where downward fluxes are not steady in time. Some situations may be treated as sometimes-steady, especially if there is evidence that conditions are nearly constant for an extended period such as a growing season. In others, the deviations may be small enough in magnitude that the error from assuming true steadiness is tolerable. It must be acknowledged however, that these approximations compromise not just the magnitude of uncertainty but also the possibility of claiming the inferred recharge rate as a long-term average.

For the general case in which flow is not steady, the appropriate Darcian methods are more complex. Transient water contents and matric pressures must be measured in addition to K (Freeze and Banner, 1970; Healy, 1989). The field site needs to be instrumented with devices for these measurements (*see* **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**). With knowledge of K and the gradient of matric pressure, Darcy's law (equation 5) straightforwardly indicates the flux density at a given time and position, and its downward component is interpretable as recharge if the setting and conditions are such that fates other than recharge can be ruled out. Transient recharge computed with Darcy's law can relate to storms or other short-term events, or provide data for integration into temporal averages. Considerable equipment and labor is required for this, however. It would not usually indicate the long-term average recharge rate accurately, but it can have a valuable payoff in giving a detailed picture of the behavior of water in the unsaturated zone.

Chemical Tracer Methods

Geochemical tracers that have been used for recharge estimation include tritium (^{3}H), oxygen-18 (^{18}O), and deuterium (^{2}H), which are constituents of the water molecule

(H_2O); naturally occurring anions such as chloride (Cl^-) and bromide (Br^-); agriculturally introduced chemicals including nitrate (NO_3^-); applied organic dyes such as fluorescein ($C_{20}H_{12}O_5$); and dissolved gases including chlorofluorocarbons (CFCs), sulfur hexafluoride (SF_6), and noble gases such as helium (He) and argon (Ar). Concentrations of these constituents in pore water are related to recharge by applying chemical mass-balance equations, by taking advantage of distinctive temporal patterns in the composition of infiltrating water, or by determining the "age" of the water (i.e. the time since the water was in contact with the atmosphere). Geochemical tracers provide point and areal estimates of recharge. Of the many geochemical tracers that have been used for research purposes, only a few have found widespread use at scales relevant to aquifer recharge. They are the main subjects of this section.

The chemical mass-balance approach assumes that the tracer is conservative and moves with pore water, without significant retardation or acceleration due to electrochemical interactions with solids. Chemical mass-balance methods can be applied at basin scales using only groundwater and atmospheric-deposition data, (e.g. Dettinger, 1989; Anderholm, 2000). When applied to the unsaturated zone, chemical mass-balance estimates are obtained by equating the time-averaged flux of tracer across the land surface with the tracer flux at a depth sufficient to be unaffected by evapotranspiration and other temporally varying influences:

$$q_{m,z_0} = C_z q_z \tag{6}$$

Here q_{m,z_0} is the average flux of tracer across the land surface (mass of tracer per unit land surface per unit time), C_z is the tracer concentration in pore water at the evaluation depth z (mass of tracer per unit volume of pore water), and q_z is the water-flux density at z (volume of water per unit bulk area normal to the flux direction per unit time). As with the Darcian method (equation 5), q_z equals the recharge rate if certain conditions apply.

Soluble tracers reach the land surface in atmospheric deposition of dust (dry fall) and in precipitation. Infiltration of runoff (streamflow and sheetwash) supplies additional tracer where overland flow occurs. Irrigation water, agricultural chemicals, and engineered structures provide additional tracer to cultivated areas and areas of artificial recharge. The total flux of tracer across the land surface is the sum of the fluxes from constituent sources, that is

$$q_{m,z_0} = C_p P + C_r R + C_i I + C_f F + C_a A \tag{7}$$

where C_p is the effective tracer concentration in precipitation (i.e. including tracer in dry deposition), C_r is the concentration of tracer in infiltrating runoff, C_i is the concentration of tracer in irrigation water, C_f is the concentration in fertilizer and other agricultural applications, C_a is the concentration in artificial-recharge water, and P, R, I, F, and A are the volumes of precipitation, runoff, irrigation, agricultural applications, and artificial-recharge water that infiltrate per unit time per unit land surface. The first term on the right-hand side represents atmospheric sources, the second term overland-flow sources, and the last three terms anthropogenic sources. One or two terms usually dominate in a given environment.

Substituting equation (7) into equation (6) and rearranging yields the water-flux density at depth z (q_z) in terms of measured quantities, that is.

$$q_z = \frac{(C_p P + C_r R + C_i I + C_f F + C_a A)}{C_z} \tag{8}$$

If z is in a zone of steady flux, that is, if climatic and other inputs are approximately stationary over relevant timescales (dictated by travel times between the land surface and the aquifer), then q_z is equal to the recharge rate at the sampled location. Application of equation (8) also assumes negligible preferential flow (or sufficient sampling to provide adequate averaging of preferential pathways), in addition to conservative transport (Wood, 1999; Scanlon, 2000). Collectively, these assumptions imply that tracer fluxes at depth are in approximate steady-state equilibrium with the time-averaged tracer fluxes across the land surface.

Recharge rates can also be determined from apparent travel velocities of a tracer (solute) "marker" through the unsaturated zone, using (Saxena and Dressie, 1984; Gvirtzman and Magaritz, 1986; Williams and Rodoni, 1997)

$$q_z = \bar{\theta}\frac{(z_2 - z_1)}{(t_2 - t_1)} \tag{9}$$

where $\bar{\theta}$ is the average volume fraction of water between z_1 and z_2. Here z_1 and z_2 are the depths of a solute marker at times t_1 and t_2. Like the mass-balance method (equation 8), the tracer-velocity method requires that the marker move conservatively with the water and that water contents at depth remain constant through time.

The most widely used geochemical tracer for recharge estimation is chloride. Chloride is abundant in nature, conservative in hydrologic settings, and readily analyzed. Chloride mass-balance methods have proven especially useful in arid and semiarid regions (Stone, 1984; Edmunds *et al.*, 1988; Allison *et al.*, 1994; Phillips, 1994; Prudic, 1994; Tyler *et al.*, 1996; Roark and Healy, 1998; Maurer and Thodal, 2000; Izbicki *et al.*, 2002; Stonestrom *et al.*, 2003b). Figure 6 (adapted from Stonestrom *et al.*, 2003b) shows an example of a conservative tracer (chloride) applied in three distinct hydrologic settings in an arid environment: beneath native xeric vegetation, beneath an ephemeral stream, and beneath an irrigated crop. Results in Figure 6 are plotted in integral form (as cumulative chloride versus cumulative water, starting at the land surface).

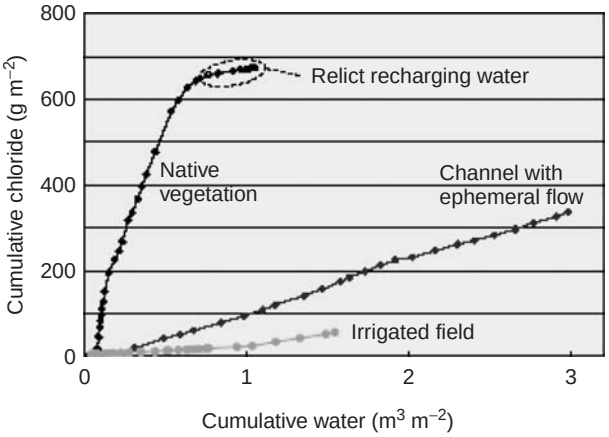


Figure 6 Concentrations of chloride in pore water, given by the slopes of the curves, indicate aquifer recharge for quasi-steady conditions. Recharge beneath irrigated field is higher than beneath ephemeral channel. Profile beneath native vegetation has multiple slopes reflecting climate-induced changes in recharge. Example from the Amargosa Desert, Nevada (Stonestrom *et al.*, 2003b). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

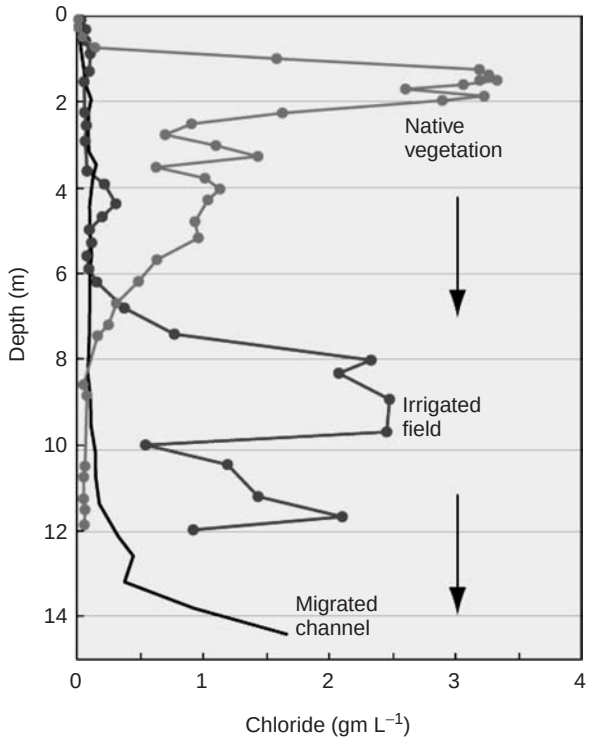


Figure 7 Travel-time of a solute marker indicates aquifer recharge under certain conditions. Here accumulations of chloride, common beneath native vegetation in arid and semiarid environments, served as the marker. Recharge following conversion to agriculture mobilized the marker in one case. Recharge following flood-induced channel migration mobilized the marker in another. Example from the Amargosa Desert, Nevada (Stonestrom *et al.*, 2003b). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Plotted this way, linear relations that extrapolate through the origin are consistent with the assumption that transport of water and solute at depth are in approximate equilibrium with average net inputs across the land surface. Slopes of the relations (equal to pore-water concentrations) are inversely proportional to recharge rates. Profiles beneath the channel and irrigated field meet the criteria for application of the method; the profile beneath native xeric vegetation does not. Water in the deep unsaturated zone here is relict infiltration from a prior climatic period. Figure 7 (adapted from Stonestrom *et al.*, 2004) shows displaced chloride "markers" used in the tracer-velocity method (equation 9).

Distinct isotopic "species" of water also serve as tracers for determining recharge. The isotopic composition of precipitation varies with altitude, season, storm track, and other factors. Recharge estimates using distinct isotopic species of water employ temporal or geographic trends in infiltrating water (Saxena and Dressie, 1984; Gvirtzman and Magaritz, 1986; Williams and Rodoni, 1997). Isotopically distinct water can be used in mass-balance or tracer-velocity methods (equations 8 and 9). Deuterium and oxygen-18 labeled species are conservative tracers. Mass-balance methods can be applied to tritium-labeled species taking radioactive decay into account.

Tritium and other nonconservative tracers can indicate the length of time that water has been isolated from the atmosphere, that is, its "age". In principle equivalent to tracer-velocity methods, age-dating methods are usually applied to the saturated zone. Recharge rates can be inferred from groundwater ages if mixing is small. If ages are known along a flow line, the recharge rate is given by

$$q = \bar{\theta} L (A_2 - A_1)^{-1} \tag{10}$$

where $\bar{\theta}$ is the average volumetric water content along the flow line, A_1 and A_2 are the ages at two points, and L is the separation length. One point is often located at the water table. Preindustrial water can be dated by decay of naturally occurring radioisotopes, including carbon-14 and chlorine-36 (Leaney and Allison, 1986; Phillips *et al.*, 1986). The abundance of tritium, chlorine-36, and other radioisotopes increased greatly during atmospheric weapons testing, labeling precipitation starting in the 1950s (Knott and Olimpio, 1986). Additional anthropogenic tracers for dating postindustrial recharge include chlorofluorocarbons, krypton-85, SF_6 and agricultural chemicals (Ekwurzel *et al.*, 1994; Davisson and Criss, 1996; Busenberg and Plummer, 2000).

Heat-based Methods

Heat is a useful tracer for estimating aquifer recharge. Perturbations of purely conductive propagation of temperature fluctuations from the land surface into the subsurface indicate recharge rates in shallow profiles beneath streams and other sources of water (Rorabaugh, 1954; Lapham, 1989; Stonestrom and Constantz, 2003). Water that moves below the reach of seasonal temperature oscillations perturbs the geothermal temperature distribution at depth. The degree of perturbation thus indicates recharge rates in deep unsaturated zones (Rousseau *et al.*, 1999). Measured temperature profiles are used in analytical or numerical inversions of the equation(s) governing the coupled transport of heat and water, solving for water flux at the water table. Numerical inversions can treat sediment heterogeneities as well as arbitrary initial and boundary conditions.

Daily or annual temperature fluctuations can often be approximated by sinusoidal functions (Figure 8). If recharging fluxes are steady and materials homogeneous, the

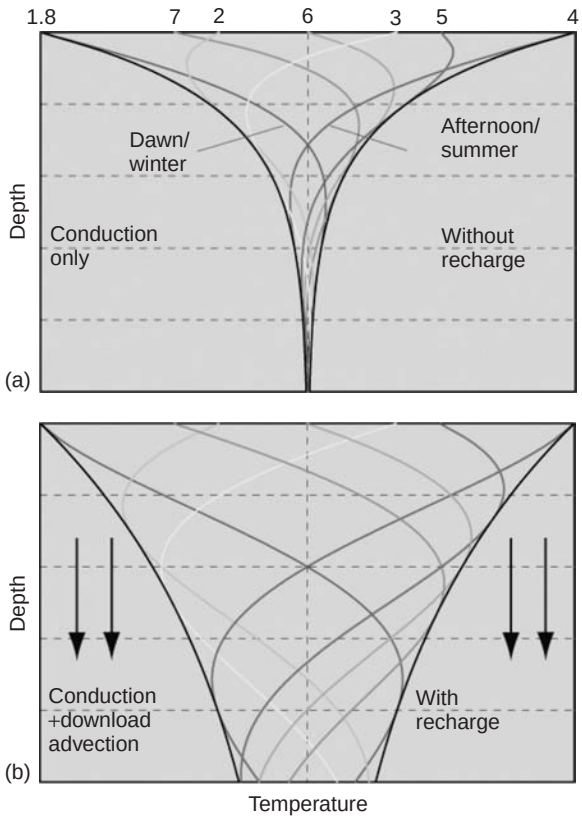


Figure 8 Recharging water advects diurnal and seasonal temperature fluctuations deeper into the profile than conduction alone. Numbered lines show successive temperature profiles over one daily or annual cycle (profiles 1 and 8 are identical, but separated by one cycle). The amount of downward shift in the bounding envelope depends on the rate of aquifer recharge. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

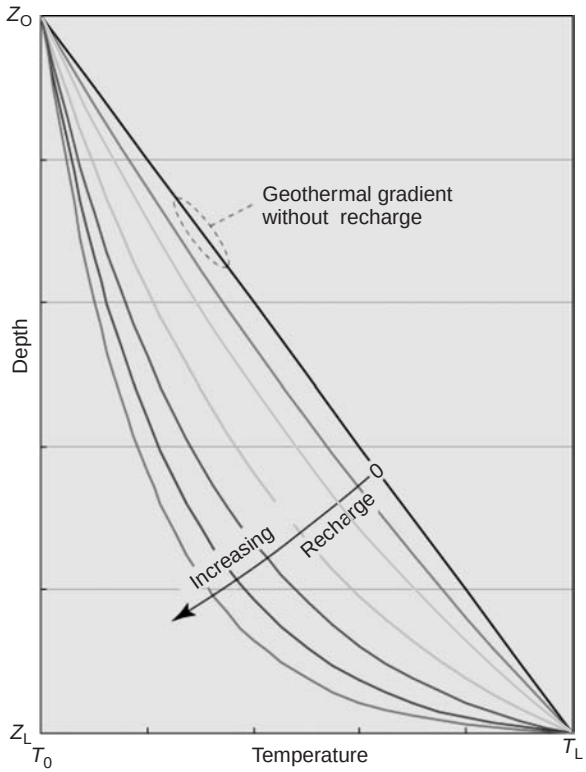


Figure 9 Recharging water in the deep unsaturated zone perturbs the steady geothermal profile produced by conduction alone. The degree of departure from the purely conductive profile indicates the amount of recharge. T_0 and T_L are temperatures at depths, Z_0 and Z_L, beneath the maximum penetration of seasonal fluctuations. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

recharge rate q can be obtained by inversion from the phase shift (b) and attenuation (a) of the temperature waves as they propagate downward from the land surface (Stallman, 1965)

$$T(z) = T_0 + \Delta T \exp(-az) \sin\left(\frac{2\pi}{\tau - bz}\right) \qquad (11)$$

where a and b are related to thermal-conduction and advection constants K' and V' by

$$a = \left[\left(\frac{K' + V'^4}{4}\right)^{\frac{1}{2}} + \frac{V'^2}{2}\right]^{1/2} - V'$$

and

$$b = \left[\left(\frac{K' + V'^4}{4}\right)^{\frac{1}{2}} - \frac{V'^2}{2}\right]^{1/2}$$

Here T_0 and ΔT are the mean and amplitude of the temperature signal at the land surface, V' and K' are defined as

$$K' = \frac{\pi C_b}{\kappa_b \tau}$$

and

$$V' = \frac{qC_w}{2\kappa_b}$$

C_b and C_w are the volumetric heat capacities of the bulk medium and water, κ_b is the thermal conductivity of the bulk medium, and τ is the period of forcing. Any self-consistent system of units (such as SI units) can be used to compute the recharge flux q, which is directly proportional to V'.

Below the depth of seasonal fluctuations, temperature profiles will be bowed away from linear profiles by recharging water, with the degree of curvature related to the magnitude of recharge (Figure 9). In dimensionless form (Bredehoeft and Papadopoulos, 1965)

$$\frac{[T(z) - T_0]}{[T_0 - T_L]} = \frac{[\exp(q/\lambda) - 1)]}{[\exp(Lq/\lambda) - 1)]} \tag{12}$$

where T_L is the temperature at the water table, at depth L, and $\lambda(= \kappa_b/C_b)$ is the thermal diffusivity of the bulk medium. As with equation (11), any self-consistent units can be used to compute the recharge flux q by inversion of the measured temperature profile, in this case the steady profile at depth.

Other Geophysical Methods

Many geophysical techniques provide data relevant to recharge based on the water-content dependence of gravitational, seismic, or electromagnetic properties of earth materials. For quantitative estimates of recharge rate, repeated high-precision gravity (microgravity) surveys indicate changes in mass due to recharge events (Pool and

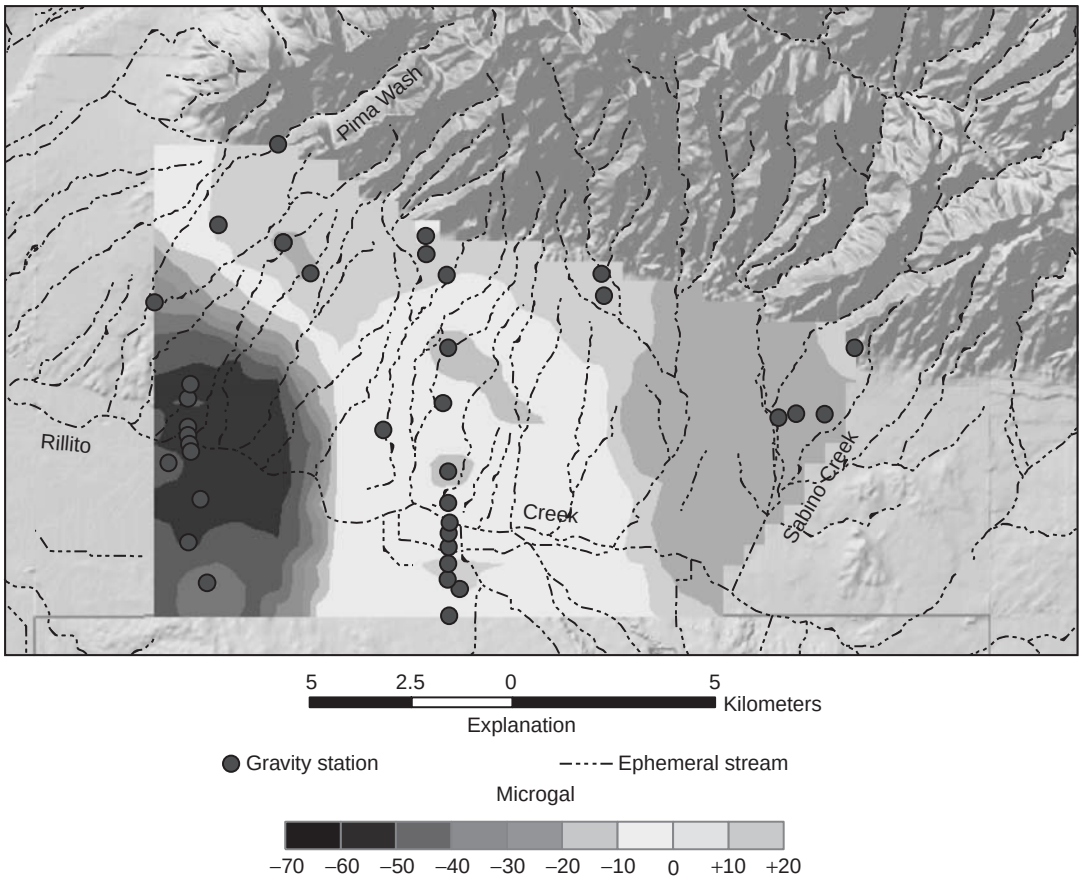


Figure 10 Changes in mass from aquifer recharge can be measured by repeated microgravity surveys. The example shows positive increases in gravity due to rising groundwater levels along the mountain front beneath tributaries to the ephemeral Rillito river, Arizona between June 1999 and March 2002. Pumping-related decreases in groundwater mass are evident downstream, beneath the main stem (unpublished data of Donald R. Pool, USGS, Tucson, AZ). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Schmidt, 1997; Figure 10). Similarly, repeated surveys using seismic or ground-penetrating-radar equipment show recharge-induced changes in water-table elevation (Haeni, 1986; Bohling *et al.*, 1989). Electrical resistance tomography can image zones of high-water content. Resistance tomographs have been used to map recharge areas in three dimensions based on land-surface measurements (Stonestrom *et al.*, 2003a). In addition to surface-based techniques, cross-bore tomographic imaging produces detailed three-dimensional reconstructions of water distribution and movement during periods of transient recharge (Daily *et al.*, 1992).

The next wave of advances in recharge estimation may involve remote sensing (RS) (see Part 05). Although still in the developmental stage, the broad spatial coverage afforded by satellite and aerial-borne instruments makes RS methods particularly attractive. Some of these methods attempt to directly determine changes in the amount of water stored in the subsurface, such as Synthetic Aperture Radar Interferometry (InSAR) (Hoffmann *et al.*, 2001; Lu and Danskin, 2001) and remotely sensed changes in gravity. Other RS methods contribute more indirectly to recharge estimation, for example, where RS data indicate certain components of the water budget equation, such as precipitation and evapotranspiration, or parameter values required in simulation models, such as soil properties, vegetation type and density, and land-use practices.

REFERENCES

Allison G.B., Gee G.W. and Tyler S.W. (1994) Vadose zone techniques for estimating groundwater recharge in arid and semiarid regions. *Soil Science Society of America Journal*, **58**, 6–14.

Anderholm S.K. (2000) *Mountain-front recharge along the eastern side of the Middle Rio-Grande Basin, central New Mexico*, Water-Resources Investigations Report 00-4010, U.S. Geological Survey.

Arnold J.G. and Allen P.M. (1996) Estimating hydrologic budgets for three illinois watersheds. *Journal of Hydrology*, **176**, 57–77.

Arnold J.G., Muttiah R.S., Srinivasan R. and Allen P.M. (2000) Regional estimation of base flow and groundwater recharge in the Upper Mississippi river basin. *Journal of Hydrology*, **227**, 21–40.

Bauer H.H. and Mastin M.C. (1997) *Recharge from Precipitation in Three Small Glacial-Till-Mantled Catchments in the Puget Sound Lowland*, Water-Resources Investigations Report 96–4219, U.S. Geological Survey: Washington.

Bohling G.C., Anderson M.P. and Bentley C.R. (1989) *Use of Ground Penetrating Radar to Define Recharge Areas in the Central Sand Plain*, Wisconsin Water Resources Center Report WIS WRC 89–01, University of Wisconsin-Madison: Madison.

Bredehoeft J.D. and Papadopoulos I.S. (1965) Rates of vertical ground-water movement estimated from the earth's thermal profile. *Water Resources Research*, **1**, 325–328.

Bredehoeft J.D., Papadopoulos S.S. and Cooper H.H. Jr (1982) Groundwater–the water-budget myth. *Scientific Basis of Water-Resource Management*, National Academy Press: Washington, pp. 51–57.

Brutsaert W. (1982) *Evaporation into the Atmosphere*, D. Reidel Publishing Company: Dordrecht.

Busenberg E. and Plummer L.N. (2000) Dating young groundwater with sulfur hexafluoride: natural and anthropogenic sources of sulfur hexafluoride. *Water Resources Research*, **36**, 3011–3030.

Cedergren H.R. (1977) *Seepage, Drainage and Flow Nets*, Wiley: New York.

Daily W., Ramirez A., LaBrecque D. and Nitao J. (1992) Electrical resistivity tomography of vadose water movement. *Water Resources Research*, **28**, 1429–1442.

Daniel C.C. III and Harned D.A. (1998) *Ground-Water Recharge to and Storage in the Regolith-Fractured Crystalline Rock Aquifer System, Guilford County, North Carolina*, Water-Resources Investigations Report 97-4140, U.S. Geological Survey.

Davisson M.L. and Criss R.E. (1996) Stable isotope and groundwater flow dynamics of agricultural irrigation recharge into groundwater resources of the Central Valley, California. In *Isotopes in Water Resources Management, Proceedings of a Symposium on Isotopes in Water Resources Management, 20–24 Mar. 1995*, Vol. 1, International Atomic Energy Agency: Vienna, pp. 405–418.

Delin G.N., Healy R.W., Landon M.K. and Bohlke J.K. (2000) Effects of topography and soil properties on recharge at two sites in an agricultural field. *American Water Resources Association*, **36**, 1401–1416.

Dettinger M.D. (1989) Reconnaissance estimates of natural recharge to desert basins in Nevada, U.S.A., by using chloride-balance calculations. *Journal of Hydrology*, **106**, 55–78.

Dreiss S.J. and Anderson L.D. (1985) Estimating vertical soil moisture flux at a land treatment site. *Ground Water*, **23**, 503–511.

Edmunds W.M., Darling W.G. and Kinniburgh D.G. (1988) Solute profile techniques for recharge estimation in semi-arid and arid terrain. In *Estimation of Natural Groundwater Recharge*, Summers I. (Ed.), D. Reidel: Norwell, pp. 139–157.

Ekwurzel B., Schlosser P., Smethie W.M.J., Plummer L.N., Busenberg E., Michel R.L., Weppernig R. and Stute M. (1994) Dating of shallow groundwater: comparison of the transient tracers $^{3}H/^{3}He$, chlorofluorocarbons, and ^{85}Kr. *Water Resources Research*, **30**, 1693–1708.

Enfield C.G., Hsieh J.J.C. and Warrick A.W. (1973) Evaluation of water flux above a deep water table using thermocouple psychrometers. *Soil Science Society of America Proceedings*, **37**, 968–970.

Flint A.L., Bodvarsson G.S., Flint L.E., Kwicklis E.M. and Fabryka-Martin J.T. (2002) Estimating recharge at Yucca Mountain, Nevada, USA–comparison of methods. *Hydrogeology Journal*, **10**, 180–204.

Freeze A.R. and Banner J. (1970) The mechanism of natural ground-water recharge and discharge 2. Laboratory column experiments and field measurements. *Water Resources Research*, **6**, 138–155.

Gardner W.R. (1964) Water movement below the root zone. *Transactions, 8th International Congress of Soil Science*, International Society of Soil Science: Bucharest, pp. 63–68.

Gburek W.J. and Folmar G.J. (1999) A ground water recharge field study–site characterization and initial results. *Hydrological Processes*, **13**, 2813–2831.

Gvirtzman H. and Magaritz M. (1986) Investigation of water movement in the unsaturated zone under an irrigated area using environmental tritium. *Water Resources Research*, **22**, 635–642.

Haeni F.P. (1986) *Application of seismic-refraction techniques to hydrologic studies*, Open File Report 84-746, U.S. Geological Survey.

Healy R.W. (1989) Seepage through a hazardous-waste trench cover. *Journal of Hydrology*, **108**, 213–234.

Healy R.W. and Cook P.G. (2002) Using groundwater levels to estimate recharge. *Hydrogeology Journal*, **10**, 91–109.

Hill M.C., Banta E.R., Harbaugh A.W. and Anderman E.R. (2000) *MODFLOW-2000, the U.S. Geological Survey Modular Ground-Water Model–User Guide to Observation, Sensitivity, and Parameter-Estimation Processes and Three Post-Processing Programs*, Open-File Report 00–184, U.S. Geological Survey.

Hoffmann J., Zebker H.A., Galloway D.L. and Amelung F. (2001) Seasonal subsidence and rebound in Las Vegas Valley, Nevada, observed by synthetic aperture radar interferometry. *Water Resources Research*, **37**, 1551–1566.

Hooper R.P., Christophersen N. and Peters N.E. (1990) Modeling of streamwater chemistry as a mixture of soilwater endmembers–an application to the Panola Mountain catchment, Georgia, USA. *Journal of Hydrology*, **116**, 321–343.

Izbicki J.A., Radyk J. and Michel R.L. (2002) Movement of water through the thick unsaturated zone underlying Oro Grande and Sheep Creek Washes in the western Mojave Desert, USA. *Hydrogeology Journal*, **10**, 409–427.

Jones T.L. (1978) *Sediment Moisture Relations–Lysimeter Project 1976–1977 Water Year*, Report RHO-ST-15, Rockwell International.

Kearns A.K. and Hendrickx J.M.H. (1998) *Temporal Variability of Diffuse Groundwater Recharge in New Mexico*, Technical Completion Report 43, New Mexico Water Resources Research Institute.

Kitching R., Shearer T.R. and Shedlock S.L. (1977) Recharge to bunter sandstone determined from lysimeters. *Journal of Hydrology*, **33**, 217–232.

Knott J.F. and Olimpio J.C. (1986) *Estimation of Recharge Rates to the Sand and Gravel Aquifer Using Environmental Tritium, Nantucket Island, Massachusetts*, Water Supply Paper 2297, U.S. Geological Survey.

Kraatz D.B. (1977) Irrigation canal lining. *FAO Land and Water Development Series.*, **1**, 99.

Kuniansky E.L. (1989) *Geohydrology and Simulation of Ground-Water Flow in the "400-Foot," "600-Foot," and Adjacent Aquifers, Baton Rouge Area*, Technical Report 49, Louisiana Department of Transportation and Development Water Resources: Louisiana.

Lapham W.W. (1989) *Use of Temperature Profiles Beneath Streams to Determine Rates of Vertical Ground-Water Flow and Vertical Hydraulic Conductivity*, Water-Supply Paper 2337, U.S. Geological Survey.

Leaney F.W. and Allison G.B. (1986) Carbon-14 and stable isotope data for an area in the Murray Basin: its use in estimating recharge. *Journal of Hydrology*, **88**, 129–145.

Lee D.R. and Cherry J.A. (1978) A field exercise on groundwater flow using seepage meters and mini-piezometers. *Journal of Geological Education*, **27**, 6–10.

Lerner D.N., Issar A.S. and Simmers I. (1990) *Groundwater Recharge, a Guide to Understanding and Estimating Natural Recharge*, Report 8, International Association of Hydrogeologists: Kenilworth.

Lu Z. and Danskin W.R. (2001) InSAR analysis of natural recharge to define structure of a groundwater basin, San Bernardino, California. *Geophysical Research Letters*, **28**, 2661–2664.

Maurer D.K. and Thodal C.E. (2000) *Quantity and Chemical Quality of Recharge, and Updated Water Budgets, for the Basin-Fill Aquifer in Eagle Valley, Western Nevada*, Water Resources Investigations Report 99–428, U.S. Geological Survey.

Maxey G.B. and Eakin T.E. (1949) *Ground Water in White River Valley, White Pine, Nye, and Lincoln Counties, Nevada*, Nevada State Engineer: Water Resources Bulletin 8.

Narayanpethkar A.B., Rao V.V.S.G. and Mallick K. (1994) Estimation of groundwater recharge in a basaltic aquifer. *Hydrological Processes*, **8**, 211–220.

Nimmo J.R., Deason J.A., Izbicki J.A. and Martin P. (2002) Evaluation of unsaturated-zone water fluxes in heterogeneous alluvium at a Mojave Basin site. *Water Resources Research*, **38**, 1215, doi:10.1029/2001WR000735, 33-1–33-13.

Nimmo J.R., Stonestrom D.A. and Akstin K.C. (1994) The feasibility of recharge rate determinations using the steady-state centrifuge method. *Soil Science Society of America Journal*, **58**, 49–56.

Nixon P.R. and Lawless G.P. (1960) Detection of deeply penetrating rain water with neutron-scattering moisture meter. *Transactions of the American Society of Agricultural Engineers*, **3**, 5–8.

Pettyjohn W.A. and Henning R. (1979) *Preliminary Estimate of Ground-Water Recharge Rates, Related Streamflow, and Water Quality in Ohio*, Water Resources Center Project Completion Report 552, Ohio State University: Columbus.

Phillips F.M. (1994) Environmental tracers for water movement in desert soils of the American Southwest. *Soil Science Society of America Journal*, **58**, 15–24.

Phillips F.M., Bentley H.W., Davis S.N., Elmore D. and Swanick G. (1986) Chlorine 36 dating of very old groundwater: 2. Milk River aquifer, Alberta, Canada. *Water Resources Research*, **22**, 2003–2016.

Pool D.R. and Schmidt W. (1997) *Measurement of Ground-Water Storage Change and Specific Yield Using the Temporal-Gravity Method Near Rillito Creek, Tucson, Arizona*, Water-Resources Investigations Report 97–4125, U.S. Geological Survey.

Prudic D.E. (1994) *Estimates of Percolation Rates and Ages of Water in Unsaturated Sediments at Two Mojave Desert Sites, California-Nevada*, Water-Resources Investigations Report 94–4160, U.S. Geological Survey.

Rasmussen W.C. and Andreasen G.E. (1959) *Hydrologic Budget of the Beaverdam Creek Basin, Maryland*, Water-Supply Paper 106, U.S. Geological Survey.

Richards L.A., Gardner W.R. and Ogata G. (1956) Physical process determining water loss from soil. *Soil Science Society of America Proceedings*, **20**, 310–314.

Roark D.M. and Healy D.F. (1998) *Quantification of deep percolation from two flood-irrigated alfalfa fields, Roswell Basin, New Mexico*, Water Resources Investigations Report 98-4096, U.S. Geological Survey.

Roman R., Caballero R., Bustos A., Diez J.A., Cartageno M.C., Ballejo A. and Caballero A. (1996) Water and solute movement under conventional corn in central Spain–I. Water balance. *Soil Science Society of America Journal*, **60**, 1530–1536.

Rorabaugh M.I. (1954) *Streambed Percolation in Development of Water Supplies*, U.S. Geological Survey: Ground Water Notes Hydraulics 25.

Rorabaugh M.I. (1964) *Estimating Changes in Bank Storage and Groundwater Contribution to Streamflow*, International Association of Scientific Hydrology Publication 63, International Association of Scientific Hydrology.

Rousseau J.P., Kwicklis E.M. and Gillies D.C. (1999) *Hydrogeology of the Unsaturated Zone, North Ramp Area of the Exploratory Studies Facility, Yucca Mountain, Nevada*, Water Resources Investigations Report 98–4050, U.S. Geological Survey.

Rutledge A.T. (1998) *Computer Programs for Describing the Recession of Ground-Water Discharge and for Estimating Mean Ground-Water Recharge and Discharge from Streamflow Data–Update*, Water-Resources Investigations Report 98–4148, US Geological Survey.

Rutledge A.T. and Mesko T.O. (1996) *Estimated Hydrologic Characteristics of Shallow Aquifer Systems in the Valley and Ridge, the Blue Ridge, and the Piedmont Physiographic Provinces Based on Analysis of Streamflow Recession and Base Flow*, US Geological Survey: Professional Paper 1422-B.

Saxena R.K. and Dressie Z. (1984) Estimation of groundwater recharge and moisture movement in sandy formations by tracing natural oxygen-18 and injected tritium profiles in the unsaturated zone. In *Isotope Hydrology, 1983, Proceedings of an International Symposium on Isotopes in Hydrologic Water Resources Development, 12–16 Sept. 1983*, International Atomic Energy Agency: Vienna, pp. 139–150.

Scanlon B.R. (2000) Uncertainties in estimating water fluxes and residence times using environmental tracers in an unsaturated zone. *Water Resources Research*, **36**, 395–409.

Scanlon B.R., Healy R.W. and Cook P.G. (2002) Choosing appropriate techniques for quantifying ground-water recharge. *Hydrogeology Journal*, **10**, 18–39.

Scanlon B.R. and Milly P.C.D. (1994) Water and heat fluxes in desert soils 2. Numerical simulations. *Water Resources Research*, **30**, 721–733.

Schicht R.J. and Walton W.C. (1961) *Hydrologic Budgets for Three Small Watersheds in Illinois*, Report of Investigation 40, Illinois State Water Survey.

Sharma M.L., Bari M. and Byrne J. (1991) Dynamics of seasonal recharge beneath a semiarid vegetation on the Gnangara Mound, Western Australia. *Hydrological Processes*, **5**, 383–398.

Sloto R.A. and Crouse M.Y. (1996) *HYSEP–A Computer Program for Streamflow Hydrograph Separation and Analysis*, Water-Resources Investigations Report 96–4040, U.S. Geological Survey.

Stallman R.W. (1965) Steady one-dimensional fluid flow in a semi-infinite porous medium with sinusoidal surface temperature. *Journal of Geophysical Research*, **70**, 2821–2827.

Steenhuis T.S., Jackson C.D., Kung S.K. and Brutsaert W. (1985) Measurement of groundwater recharge on eastern Long Island, New York, U.S.A. *Journal of Hydrology*, **79**, 145–169.

Stephens D.B. and Knowlton R. Jr (1986) Soil water movement and recharge through sand at a semiarid site in New Mexico. *Water Resources Research*, **22**, 881–889.

Stone W.J. (1984) *Recharge in the Salt Lake Coal Field Based on Chloride in the Unsaturated Zone*, Open-File Report 214, New Mexico Bureau of Mines and Mineral Resources: Socorro.

Stonestrom D.A., Abraham J.D., Lucius J.E. and Prudic D.E. (2003a) Focused subsurface flow in the Amargosa Desert characterized by direct-current resistivity profiling [abs.]. *Eos, Transactions, American Geophysical Union*, **84**, Fall Meeting Supplement, Abstract F660 H31B-0467.

Stonestrom D.A. and Constantz J. (2003) *Heat as a Tool for Studying the Movement of Ground Water Near Streams*, Circular 1260, U.S. Geological Survey.

Stonestrom D.A., Prudic D.E., Laczniak R.J., Akstin K.C., Boyd R.A. and Henkelman K.K. (2003b) *Estimates of Deep Percolation beneath Irrigated Fields, Native Vegetation, and the Amargosa-River Channel, Amargosa Desert, Nye County, Nevada*, Open-File Report 03–104, U.S. Geological Survey.

Stonestrom D.A., Prudic D.E., Laczniak R.J. and Akstin K.C. (2004) Tectonic, climatic, and land-use controls on ground-water recharge in an arid alluvial basin–Amargosa Desert, U.S.A. In *Groundwater Recharge in a Desert Environment–The Southwestern United States*, Hogan J.F., Phillips F.M. and Scanlon B.R. (Eds.), American Geophysical Union: Washington, pp. 29–47.

Su N. (1994) A formula for computation of time-varying recharge of groundwater. *Journal of Hydrology*, **160**, 123–135.

Tiedeman C.R., Kernodle J.M. and McAda D.P. (1998) *Application of Nonlinear-Regression Methods to a Ground-Water Flow Model of the Albuquerque Basin, New Mexico*, Water-Resources Investigations Report 98–4172, U.S. Geological Survey.

Todd D.K. (1989) *Groundwater Hydrology*, Wiley: New York.

Tyler S.W., Chapman J.B., Conrad S.H., Hammermeister D.P., Blout D.O., Miller J.J., Sully M.J. and Ginanni J.M. (1996) Soil-water flux in the southern Great Basin, United States: temporal and spatial variations over the last 120,000 years. *Water Resources Research*, **32**, 1481–1499.

Williams A.E. and Rodoni D.P. (1997) Regional isotope effects and application to hydrologic investigations in southwestern California. *Water Resources Research*, **33**, 1721–1729.

Winter T.C., Harvey J.W., Franke L.O. and Alley W.M. (1998) *Ground Water and Surface Water–A Single Resource*, Circular 1139, U.S. Geological Survey.

Wood W.W. (1999) Use and misuse of the chloride-mass balance method in estimating groundwater recharge. *Ground Water*, **37**, 2–3.

Wu J., Zhang R. and Yang J. (1997) Estimating infiltration recharge using a response function model. *Journal of Hydrology*, **198**, 124–139.

Young M.H., Wierenga P.J. and Mancino C.F. (1996) Large weighing lysimeters for water use and deep percolation studies. *Soil Science*, **161**(8), 491–501.

147: Characterization of Porous and Fractured Media

PHILIPPE RENARD[1], JAIME GÓMEZ-HERNÁNDEZ[2] AND SOUHEIL EZZEDINE[3]

[1]*Centre for Hydrogeology, University of Neuchâtel, Neuchâtel, Switzerland*
[2]*Department of Hydraulics and Environmental Engineering, Universidad Politécnica de Valencia, Valencia, Spain*
[3]*University of California, Lawrence Livermore National Laboratory, Livermore CA, US*

The characterization of porous or fractured media is a site, scale, and project-specific process aiming at a quantitative description of the geometry and properties of the geological structures controlling groundwater flow and solute transport. The characterization process involves four main steps (i) the definition of the domain and the goals of the characterization; (ii) the collection and analysis of field observations allowing the construction of a geometrical model; (iii) the collection and analysis of field measurements allowing to construct a property model; and (iv) the collection and analysis of field data relative to the state of the system and their integration within the geometrical and property models (inverse problem). When data are sufficient and structures are relatively well known, deterministic techniques of interpolation can be successfully applied to construct the geometric or the parameter models. However, because of the lack of sufficient data, stochastic models are often employed to characterize the heterogeneity that usually exists; such models also facilitate the quantification of the uncertainty in model predictions. Without describing the details of every technique, this article provides an overview of the tools most often used for the characterization of porous or fractured aquifers.

INTRODUCTION

Aquifer characterization can be defined as the process of data acquisition, analysis, and integration leading to a description of aquifer geometry and properties. This process may be relatively straightforward and limited to mapping the extent of an aquifer, its thickness, and estimating average properties such as transmissivity and storativity. It may as well be a much more complex task, integrating data obtained from various field investigations and involving intensive numerical modeling. The level of complexity of the characterization process is related to the goals of the study, the geological conditions, and the level of confidence required by the stakeholders. For example, evaluating the safety of a deep underground nuclear waste repository requires higher characterization efforts than estimating the amount of groundwater exploitable in a small shallow aquifer. Furthermore, the characterization efforts are oriented toward specific aquifer properties and objectives depending on the project. In the nuclear waste repository project, it may be important to have an accurate description of the diffusion properties of the geological materials, while this aspect may be irrelevant for the water resource project. Similarly, the thermal properties of the underground will have to be characterized in the framework of a geothermal project, but are not relevant in the case of groundwater protection.

The above examples demonstrate that characterization is a site, scale, and project-specific activity, the aim of which is not to describe all the properties of the system, but to focus on the major structures and properties relevant to the processes of interest. This requires the collection of specific field observations, their analysis, and their quantitative integration into a synthetic descriptive model of the reality.

Most of the difficulties associated with subsurface characterization stem from the high spatial variability of the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

subsurface environment. This heterogeneity is an intrinsic property of geological formations and results from the complexity of the geological processes (sedimentation, diagenesis, rock deformation, etc.). Additionally, only sparse information is available from outcrops, boreholes, or geophysics. The lack of information combined with the intrinsic heterogeneity is the source of uncertainty that makes the characterization of subsurface hydrogeological systems challenging.

The aim of this article is to present an overview of the usual steps and techniques used for the characterization of porous and fractured media. Subsequent articles within the encyclopedia cover technical aspects of data acquisition and modeling techniques used for data integration. Some specific technical points such as interpolation methods are covered in more detail, as they are not treated elsewhere within the encyclopedia.

The article encompasses six sections. The first section introduces the main steps involved in the characterization and highlights some key features and difficulties. Subsequent sections discuss deterministic, stochastic, and genetic techniques. The final section is devoted to inverse modeling.

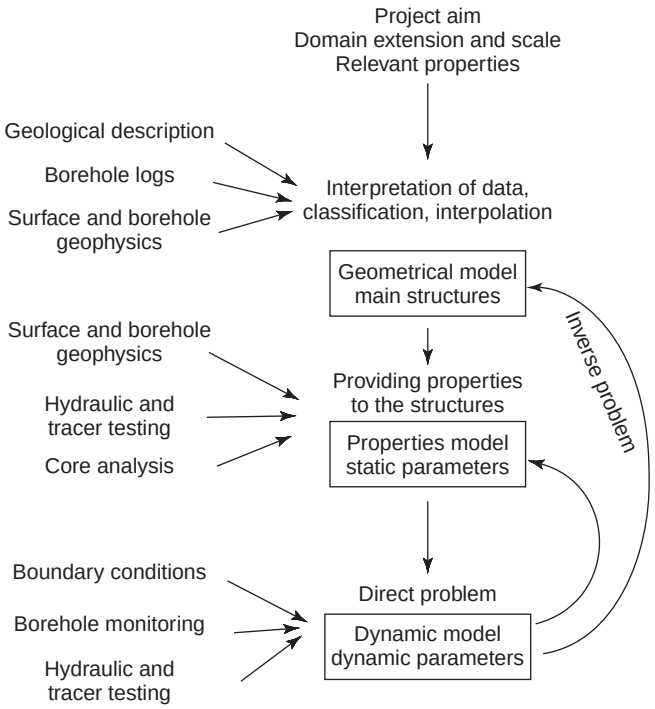


Figure 1 A schematic diagram of the characterization process

CHARACTERIZATION PROCEDURE

The Typical Steps of Characterization

The goal of characterization is to build a model in which the parameters involved in the processes under consideration are specified everywhere within the domain of interest. The four major steps of the characterization procedure are illustrated in Figure 1.

The first step is the definition of the project goals, domain of investigation, and the selection of the relevant processes and variables. Then, characterization consists of collection, interpretation, and analysis of measurements from various data sources. It is useful to distinguish between static (time invariant) and dynamic (time dependent) properties. This definition is not strict: some static properties in a given context (e.g. the aquifer geometry) may be considered as dynamic in another context (e.g. aquifer geometry during land subsidence; *see* **Chapter 158, Anthropogenic Land Subsidence, Volume 4**). Another useful distinction is the separation between measurable quantities (state variables) and physical parameters that cannot be directly measured but that parameterize physical laws (e.g. hydraulic conductivity).

The second step is the definition of the geometry of the structures controlling groundwater flow and transport. It relies on local geological observations, general geological knowledge related to the type of environment encountered, and geophysical investigations (*see* **Chapter 148, Aquifer Characterization by Geophysical Methods, Volume 4**). The end product is a geometric model that encompasses all relevant features: including aquifer, aquitard, channels, faults, lenses, and so on.

In the third step, field experiments such as hydraulic testing (*see* **Chapter 151, Hydraulics of Wells and Well Testing, Volume 4**) laboratory experiments, geochemical sampling, and tracer testing allow determination of the physical properties (static and dynamic) of the main structures. Of course, the separation between geometric modeling and defining properties is not straightforward in practice, as the knowledge of the properties is often required to decide whether a geological object is a relevant structure, and whether it is necessary to define its geometry.

The final step is the integration of the dynamic observations (state variables such as hydraulic head) with the static properties. The ultimate goal is that the geometric model and the property model must be in agreement with the dynamic observations. The main objective is therefore to link the geometry, the physical parameters, and the state variables through a system of partial differential equations that can be solved analytically or numerically (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**; **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**; **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**; and **Chapter 157, Sea Water Intrusion Into Coastal Aquifers, Volume 4**). The model is then used in an

inverse procedure (*see* **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**) in order to improve the property model and/or the geometric model so that the calculated state variables match the observed ones under certain criteria.

Typical Goals of the Characterization

Depending on the type of aquifer, and on the project goals, different properties may be relevant, but in most cases the basic goal is to characterize the water conductive features. This means that for a porous media the most relevant property is the hydraulic conductivity. For fractured media, the relevant properties are the intensity of fractures, their extension, their connectivity, their apertures, and the hydraulic conductivity of the matrix. Table 1 provides a summary of the typical goals for subsurface characterization.

Deterministic versus Stochastic Methods

The geometric model, the property model, and the dynamic model can be defined within a deterministic or stochastic framework. In the deterministic framework, a unique geometry and property map are considered. On the other hand, in the stochastic framework (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**), the unique estimate is replaced by an ensemble of equally probable realizations, generally characterized by a statistical model. The main advantage of the stochastic approach is that it provides a formal means to quantify uncertainty. In practice, however, stochastic and deterministic approaches are often complementary: some parts of the characterization process are described deterministically while others are described statistically.

The Scale Issue

One important difficulty that arises during the characterization process is the integration of observations and measurements that have been collected at different scales. It is important to distinguish four main scales of interest (Haldorsen and Lake, 1982; Dagan, 1989): the microscale, the macro or laboratory scale, the mega or local scale, and the giga or regional scale (Figure 2).

Scale issues arise because a physical law that describes a process at one scale may differ when it is averaged or upscaled over a large volume. Furthermore, when the form of the equation remains identical between different scales, values of the physical parameters have to be averaged in a way that is physically consistent. Most often the relevant parameters are not additive and cannot be simply averaged by a standard mean.

Table 1 Typical goals for the characterization of fractured or porous media

All cases	Porous media	Fractured media	Project-specific
Geological and hydrogeological boundaries	Hydraulic conductivity	Frequency of fracture occurrence / density	Thermal properties
Head/ pressures	Specific storage	Orientation – extension	Diffusion properties
Fluxes	Porosity	Aperture	Geochemistry
	Dispersivity	Transmissivity	Stress/deformation
		Connectivity	Salinity
		Matrix properties	

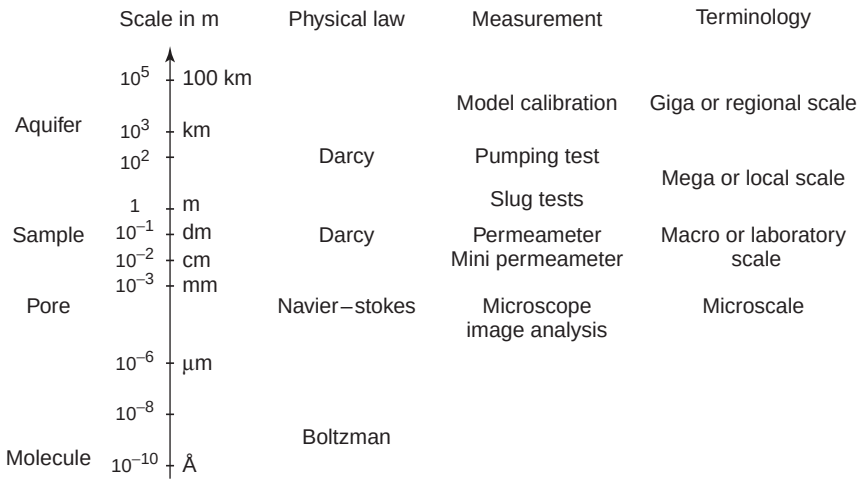


Figure 2 Definition of the characterization scales

To further illustrate these principles, at the microscopic scale the governing equations for groundwater flow are the Navier–Stokes equations (Figure 2). These are parameterized using the fluid viscosity and density, which are the relevant properties at the pore scale. Characterization techniques at this scale include microscopy, X-ray tomography, 3D pore space reconstruction, and so on. Moving to the macroscopic scale, it is possible to demonstrate theoretically that under low Reynolds number, the flux of groundwater through a porous medium obeys Darcy's law (Matheron, 1967; Mei and Auriault, 1989). The relevant physical parameters that remain are the fluid viscosity and density, but the geometry of the microscopic pore network is now characterized by two macroscopic properties, namely, the permeability and the porosity. Again, moving to a larger scale (mega or gigascopic) it is still possible to prove that the new governing equation is identical in form to Darcy's law (Matheron, 1967; Sáez *et al.*, 1989), but the permeability represents now an average of the small-scale permeabilities that account for their spatial distribution in the aquifer, and becomes a tensorial quantity in most cases, even when it is a scalar at the smaller scale. Many upscaling tools exist depending on the type of permeability distribution, but the most accurate techniques require a detailed knowledge of the spatial distribution of the permeability (Renard and de Marsily, 1997).

One of the best examples of a detailed study of scale dependence of the permeability is provided by Tidwell and Wilson (1999). They used an automated mini air permeameter in order to map the permeability of the face of a tuff sample. They measured the permeability with different injection devices (seals) having different radii. In this way, they characterized the permeability field of a sample face at different scales. Figure 3(b) and 3(c) present two permeability maps obtained using different seal sizes. It is apparent from the figures that when the seal size increases, heterogeneities are smoothed out. The statistical description of the permeability field is a function of the scale of observation. The mean slightly decreases with increasing scale (Figure 3e), the variance significantly

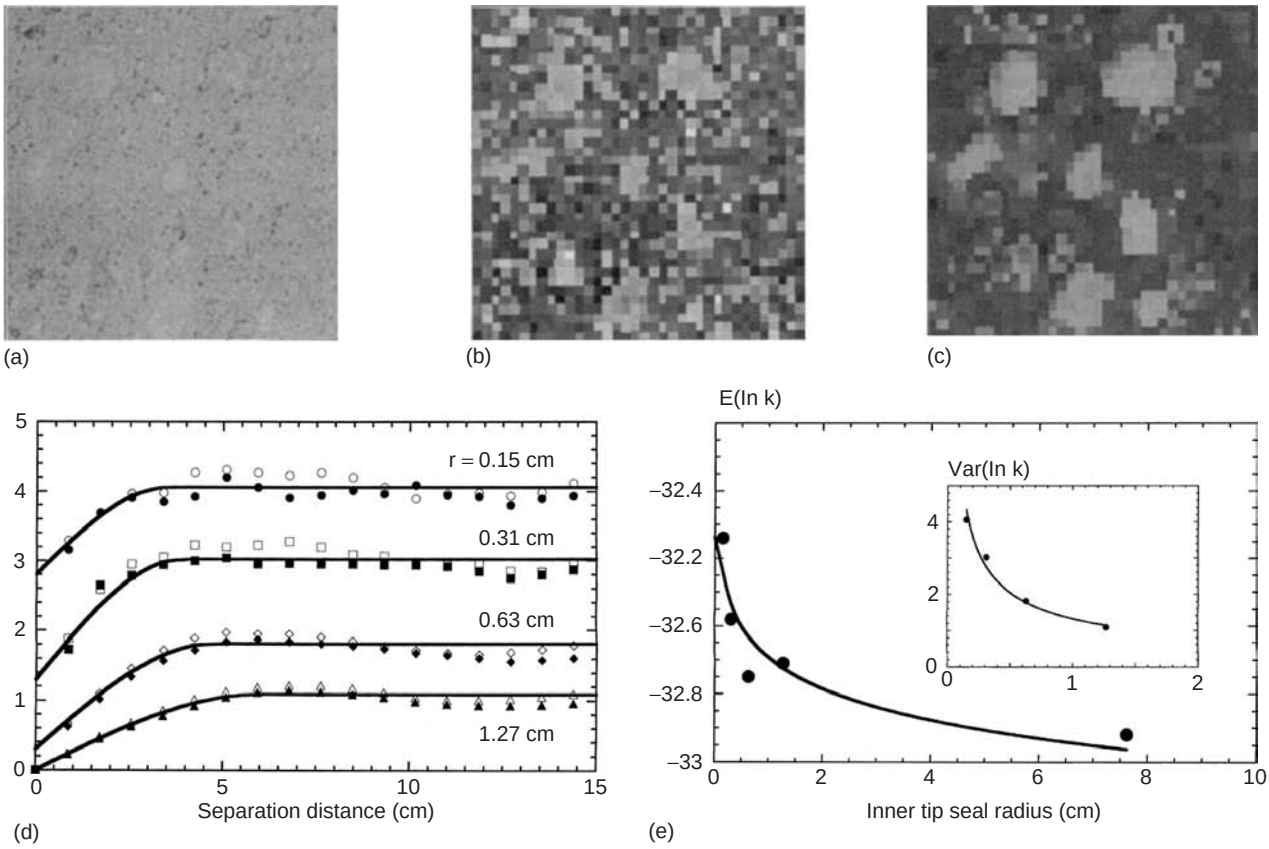


Figure 3 Characterization of the permeability field of a tuff sample at several successive scales. (a) Photograph of the sample face (30 cm times 30 cm); (b) and (c) permeability maps, measured with an injection device having an inner tip seal radius of 0.15 cm and 0.63 cm respectively, showing the smoothing of the permeability contrast with increasing size of measurement device; (d) semi-variograms of the permeability fields as a function of the tip seal radius, r; (f) expectation and variance of the permeability as a function of the tip seal radius (Reproduced from Tidwell and Wilson, 1999 by permission of American Geophysical Union)

decreases (Figure 3e), and the correlation length increases (Figure 3d).

On a much broader range of scales, several authors (Clauser, 1992; Sánchez-Vila *et al.*, 1996; Schulze-Makuch and Cherkauer, 1998) indicate that the average hydraulic conductivity of a formation increases with scale (Figure 4). These observations contradict the results of Tidwell and Wilson (1999). However, the apparent increase must be analyzed with caution because many experimental data at the laboratory scale are biased towards low values (Zlotnik *et al.*, 2000).

The dependence of hydraulic conductivity with scale is not the exception; the characterization of most parameters strongly depends on the scale of observation. Last but not the least, the elements on which the static and dynamic models have been discretized will, in general, be larger than the support on which measurements have been taken. Therefore, the values assigned to the model elements will always represent some type of averaged, or upscaled value of the underlying hydraulic conductivity distribution. This implies that larger the elements, the smoother their spatial distribution. Large elements should be sampled from probability distributions with smaller variance and larger continuity than small elements. When all elements in the model are of the same size and shape, the only decision to make is the choice of a random function model; however, when the model has been discretized with elements of different sizes, care should be taken to ensure the proper spatial variability for each element size. The rigorous method of defining the parameter values is to establish an upscaling rule to allow the transfer of the statistical characterization that can be inferred from the measurements, at the measurement support, up to the simulation support. Unfortunately, in most cases this extrapolation is never made, and the statistical characterization of the measurement values is transferred into the simulation support without any correction, incurring in what some authors have termed "not accounting for the missing scale" (Durlofsky, 1992).

DETERMINISTIC CHARACTERIZATION

In this section, we review a few deterministic tools used to define the geometry of a hydrogeological system and to describe the distribution of properties within the geological formations.

Zonation

The whole domain is splitted volume into subvolumes corresponding to different geological objects that can represent hydrostratigraphic units or subunits.

Recently, automatic algorithms have been developed to construct the zonation in three dimensions using geological observations along outcrops, borehole logs, and interpretative vertical sections (Courrioux *et al.*, 2001). Every point that identifies an interface between two zones is represented in three dimensions by a pair of points located on each side of the interface and labeled with a number corresponding to the identifier of the zone. An initial partition of the three-dimensional space is constructed with the help of a Voronoï diagram and interfaces are subsequently smoothed. An example of the application of this method is provided in Figure 5.

Interpolation

Often data are available at points. The data must be interpolated in order to reconstruct either a geometric surface or the spatial distribution of a parameter. The interpolation problem is widely encountered in many fields of

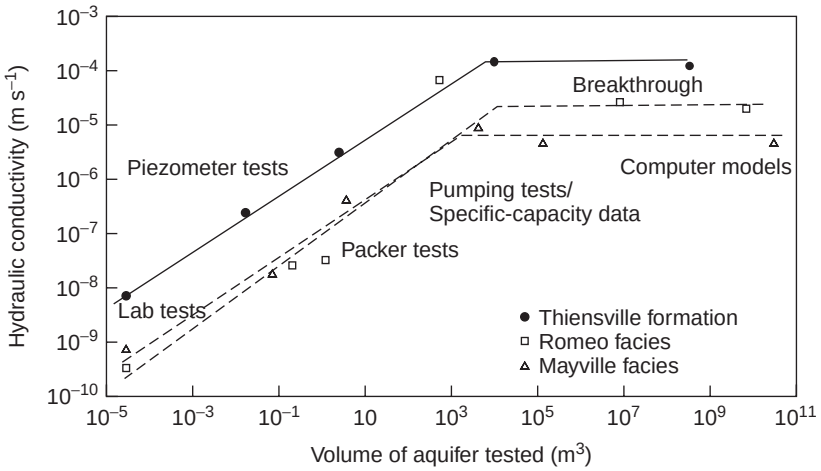


Figure 4 Apparent scale effect in hydraulic conductivity (Reproduced from Schulze–Makuch and Cherkauer, 1998 by permission of Springer-Verlag GmbH)

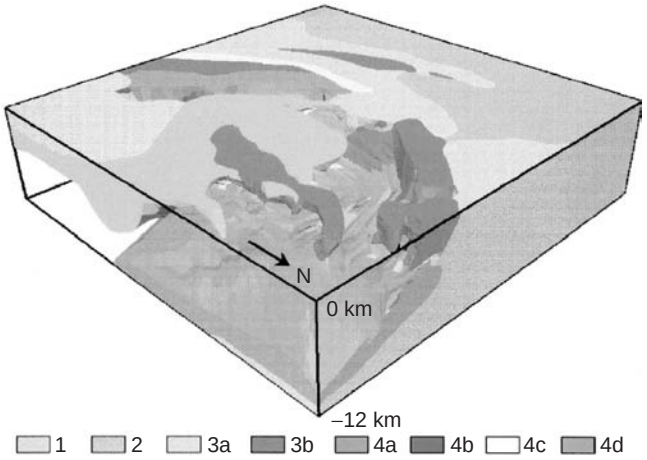


Figure 5 Geometric model of the Cadomian belt constructed automatically using Voronoı diagrams. Every grey level corresponds to a different geological unit ($50 \times 60 \times 12\,km^3$) (Reprinted from Courrioux *et al.*, 2001. © 2001, with permission from Elsevier) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

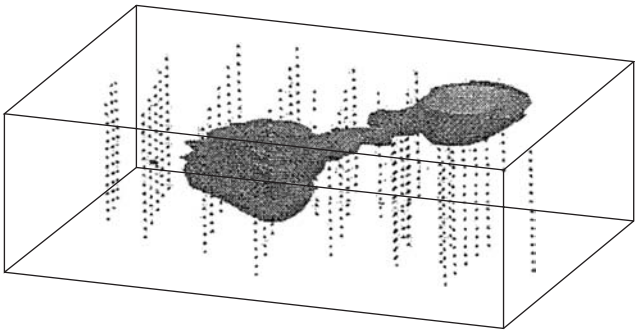


Figure 6 Example of three-dimensional interpolation (kriging) of contaminant concentrations in the cape cod aquifer allowing the location of a contaminant plume to be characterized. The dots represent the data points; the grey zone represents the volume where the interpolated concentrations exceed a threshold (Jones N.L. *et al.*, 2003; reprinted from Ground Water with permission of the National Ground Water Association. © 2003)

sciences. Consequently, a wide variety of techniques are available. Among the techniques most often used are linear piecewise interpolation, inverse distance weighting, polynomial interpolation, splines, natural neighbor, kriging, and radial basis functions. These techniques are implemented in numerous software packages such as Surfer (**http://www.goldensoftware.com**), ArcGis (**http://www.esri.com**), Idrisi (**http://www.clarklabs.org**), GMS (**http://chl.erdc.usace.army.mil**), Earth Vision (**http://www.dgi.com**), and FeFlow (**http://www.wasy.de**)

The relative efficiency of interpolation techniques has been investigated in many articles, one of the most recent is the article by Jones *et al.* (2003) who compared the application of inverse distance weighting, natural neighbor, and kriging for the characterization of four different contaminant plumes in three dimensions (Figure 6). This study illustrated that at three sites the kriging technique gave the lowest error; the inverse distance weighting gave the lowest error at one site and performed well otherwise. The natural neighbor method was the least accurate. Note that kriging is considered here as a deterministic method, since only the interpolated values are used but not the estimated uncertainty. The conclusions of Jones *et al.* are not really surprising since kriging (as will be discussed more in detail in the Section "Geostatistics") is a method whose principle is to minimize estimation errors. What should be assessed is the validity of the variogram analysis of Jones *et al.* when the inverse distance method performed better than kriging. Nonetheless, the main point of interest of their study is that it showed that inverse distance methods, which are very fast and do not require a variogram analysis, perform rather well and can provide an acceptable interpolated map at least in the preliminary stage of a study.

Discrete Smooth Interpolation

A particularly interesting interpolation technique in the framework of geological objects is the so-called Discrete Smooth Interpolation or DSI (Mallet, 2002). The principle of DSI is to construct a discrete representation of an object, a triangulated surface for example, and to impose some constraints on the object. For example, certain nodes are given some fixed positions while others exist, but their location is not known *a priori*. On a node, the orientation of the surface can be defined as a constraint, while its position is unknown. Some constraints such as a minimum distance can be imposed between objects as well. The basis of the DSI algorithm is to minimize the roughness of the discrete object subject to predefined constraints. Generally speaking, the roughness itself is defined as the sum of the squared distance between any point of the graph and the center of gravity of its immediate neighbors.

DSI is an extremely versatile technique. It can be used to interactively model the geometry of complex geological structures including layers and faults. As an example, Figure 7 shows the three-dimensional geometry of the Soûltz horst (Alsace, France). Site characterization is carried out in the framework of a hot dry rock geothermal energy project. In this case, the reservoir geometry was constrained by borehole observations and five seismic profiles. After having identified the different faults and horizons on the seismic profiles, DSI was used to interpolate the triangulated surfaces corresponding to faults and geological boundaries between layers (Renard and Courrioux, 1994).

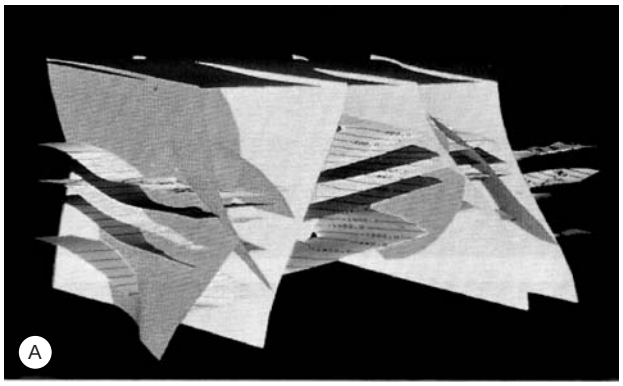


Figure 7 Three-dimensional geometry a network of 18 faults and five stratigraphic layers. The top surface represents the topography, the geometric model extends to a depth of 2 km and has an extension of 6 km by 2 km (Reprinted from Renard and Courrioux 1994. © 1994, with permission from Elsevier) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

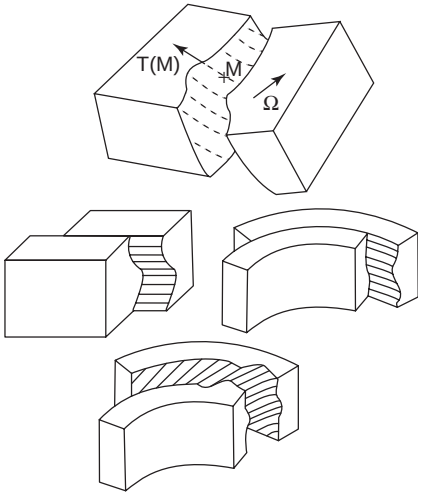


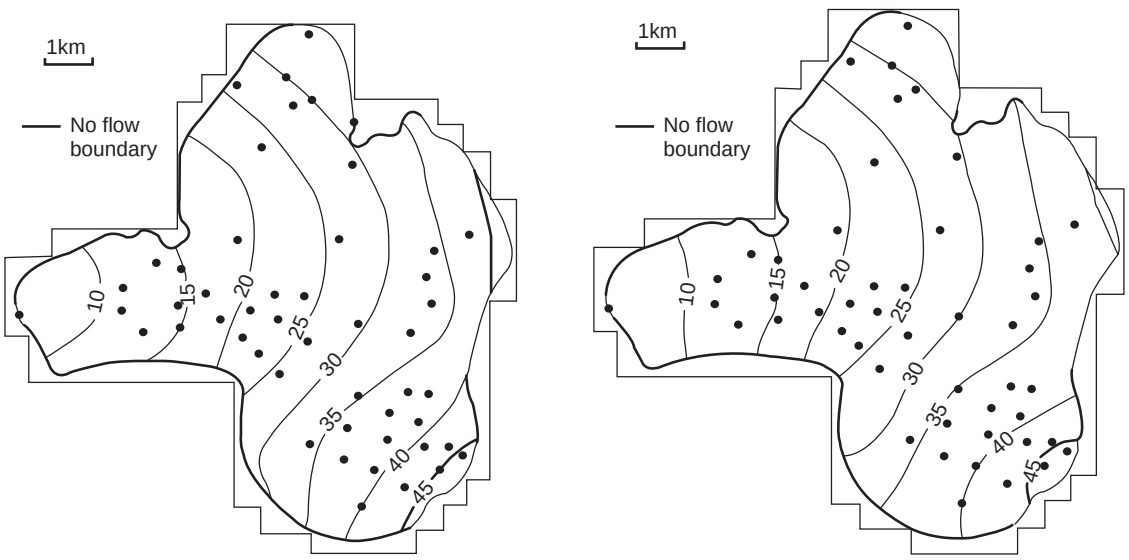
Figure 8 Examples of admissible surfaces for a fault (Reprinted from Thibaut *et al.*, 1996. © 1996, with permission from Elsevier)

Plausibility Constraints

When interpolating, a common problem is to respect not only the raw data but also some knowledge related to the type of variable that is interpolated. For example, the hydraulic conductivity or the concentration cannot be negative. Another example is the interpolation of the geometry of a three-dimensional surface describing a fault from a series of points in space. In order to be acceptable, it must belong to a certain type of surface such as planes, spheres, cylinders, and surfaces of revolution (Figure 8) as the fault surface has been created by the relative movement of two rigid blocks (Thibaut *et al.*, 1996). Mallet (2002) discusses how to implement such constraints within DSI. As a last example, the interpolation of hydraulic head data must honor boundary conditions. Delhomme (1979) demonstrates how the kriging technique can be modified to account for such constraints (Figure 9).

STOCHASTIC CHARACTERIZATION

Stochastic modeling allows assessing uncertainty. In this section, we present an overview of the statistical models



(a) Usual kriging 49 wells

(b) Kriging under boundary conditions 49 wells

Figure 9 Comparison of (a) standard kriging and (b) kriging under boundary conditions to interpolate piezometric head data (By courtesy of JP Delhomme, 1979)

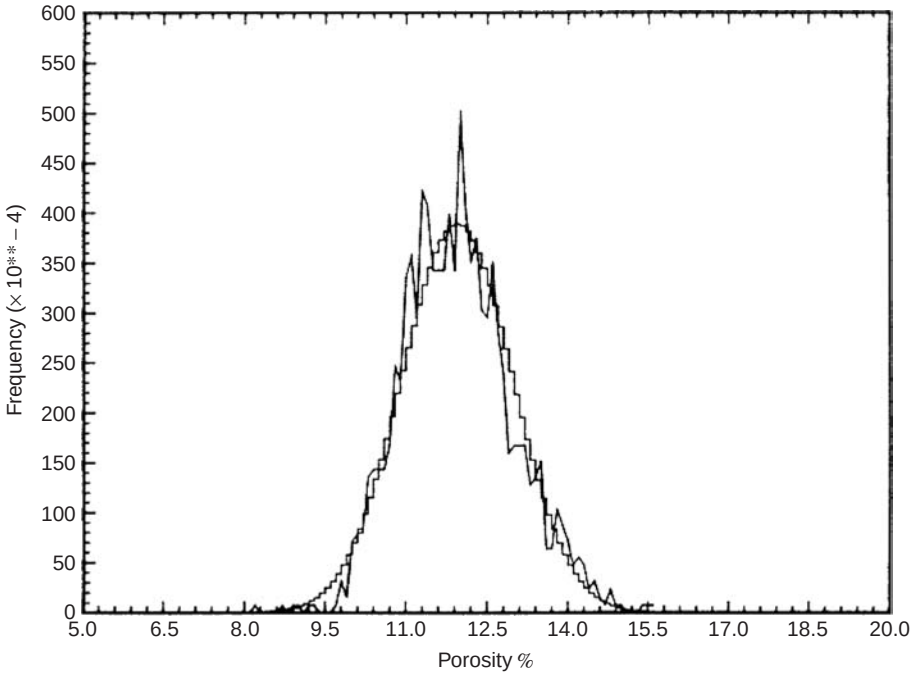


Figure 10 Frequency distribution (experimental and Gaussian model) of porosity of a series of Berea sandstone samples (Reprinted from Bahralolom and Heller 1991. © 1991, with permission from Elsevier)

most currently used for the characterization of porous and fractured media. We start the review with standard statistics, and follow with object-based models, geostatistics, and finally a short overview of the emerging field of multiple points geostatistics.

Statistics

The first kind of stochastic analysis, which is conducted when characterizing a hydrogeological system, is to investigate the univariate and nonspatial statistics. The most simple and complete statistical tool during this first step is to analyze the experimental probability density functions (pdf). The pdf allows the analyst to infer the degree of variability of the property, the type of probability law that would best represent the data, the possible multimodality, and so on. When conducting such analyses, special tools must be used when dealing with data that fall into a finite mathematical space (i.e. compositional data, fracture orientations). In the case of fracture orientation, the field data are a series of orientation angles (strike and dip). The usual way to represent these statistics is a contoured stereographic projection that allows the main families of fractures, their mean orientation, and variability around the mean to be defined (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**).

When the experimental statistics have been analysed, simple statistical models can be used to represent the data. For example, Figure 10 shows an example of porosity pdf taken from a number of sandstone samples. The resulting plot indicates that, in this case, a Gaussian distribution, defined by its mean and variance, can be used to model the porosity distribution.

The statistical analysis of fracture observations along boreholes, tunnel faces, or maps requires some specific tools. For example, the statistics are calculated from data that are usually gathered on 1D or 2D space, but they need to be corrected by stereological techniques in order to estimate the 3D statistics. Chilès and de Marsily (1993) provide an excellent overview of specific statistical (and geostatistical) techniques used to analyze fractures and fracture networks. One of the particular aspects of fracture network statistics is that they often exhibit a very wide range of scales and therefore their statistics can be described by power laws (Figure 11) relating over certain domains (Bonnet *et al.*, 2001; Bour *et al.*, 2002).

A second step of the statistical characterization is to investigate the multivariate statistics to define correlations between them. The variables can be numerical properties and also categorical properties such as lithofacies or hydrofacies indicators. When a large number of numerical variables are available (e.g. geophysical logging), a systematic statistical analysis complemented with the application of a classification algorithm may allow the various relevant litho or hydrofacies identifications in a semiautomatic way. The same statistical techniques are applied to characterize water types within an aquifer (Güler *et al.*, 2002).

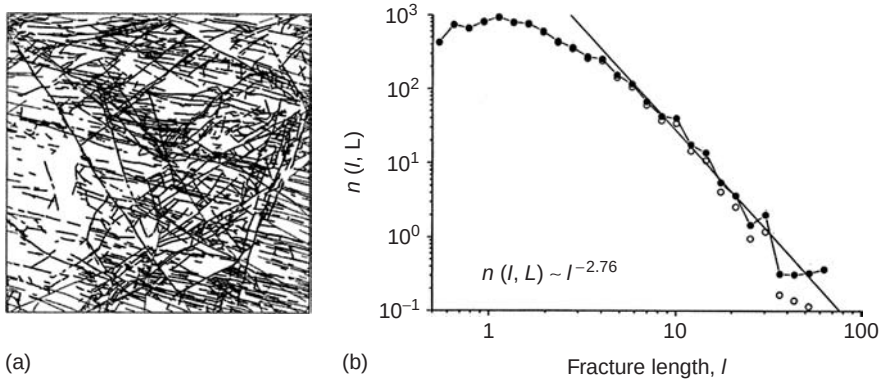


Figure 11 (a) Map of a fracture network in a sandstone outcrop ($90 \times 90\ m^2$) in Norway and (b) density length distribution. The power law begins when the fracture length is greater than 5 m (Reproduced from Bour *et al.*, 2002 by permission of American Geophysical Union)

Object-based Models

Object-based models attempt to reproduce the geological architecture of the aquifer by locating objects with shapes that resemble geological bodies using rules about their position, size, and shape, and also rules of attraction, repulsion, and spatial proportions. These objects are assimilated to specific geologic facies, which are later assigned porosity and hydraulic conductivity values.

The most common algorithms used in object-based models are Boolean models. These models work with deterministic shapes defined by stochastic parameters. For instance, the early models by Haldorsen and Lake (1984) reproduce sand/shale reservoirs in which shales are included as parallelepipeds with random locations and sizes. Generating such a model starts by randomly drawing a point in space and then drawing at random the three sizes of the parallelepiped (representing a shale inclusion) that is located at the drawn point. This procedure is repeated until a predetermined sand/shale proportion is reached. These models have evolved substantially to include (i) more elaborate shapes, which may better resemble the geological bodies (Figure 12) and (ii) complex rules regarding allowed relative positions of the objects, that is, following a stratigraphic sequence, erosion rules, and so on (Jussel *et al.*, 1994; Scheibe and Freyberg, 1995).

Realizations generated with object-based models are appealing to geologists; however, their main drawback is the difficulty to condition these realizations to large amount of data: it is very difficult to randomly draw objects obeying all rules and honoring borehole information. Other drawbacks are that lithofacies do not conform to the simple geometries used, lithofacies are not randomly distributed in space, and that these algorithms are difficult to generalize and must be custom designed for each depositional pattern. Some of these drawbacks have been addressed by Tyler *et al.* (1994) with different degrees of success.

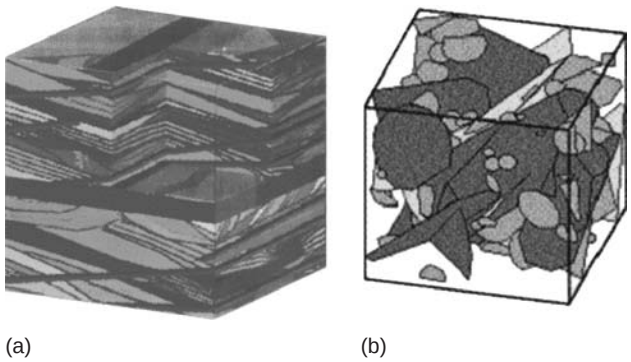


Figure 12 Two examples of object-based models. (a) Simulated structure of a point bar deposits. The dark grey represents low permeability while the light grey represents high permeability region. The cube represents approximately 1 cubic m (Reproduced from Scheibe and Freyberg, 1995 by permission of American Geophysical Union). (b) 3D fracture network. The fractures are assumed to be disks. The network is simulated according to the inferred statistics of fracture orientation, length, and density (Reproduced from Ezzedine, 1994 by permission of American Geophysical Union) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Geostatistics

The word geostatistics, or geographical statistics, was defined by Matheron in 1962 to designate a set of statistical techniques used for ore reserve evaluation. The key concept is to quantify, in statistical terms, how the information provided by a sample located in space and/or time influences the statistics of possible values of the same variable at any distance from this data point. This information is quantified with a variogram or a covariance function. The same tools can also be applied to a set of different variables accounting for spatial cross-correlation between variables. These techniques and formulations are described in various references

(Matheron, 1962; Journel, 1989; Goovaerts, 1997; Kitanidis 1997; Chilès and Delfiner, 1999; *see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**).

In practice, dedicated software is available either as interactive packages such as Isatis (**http://www.geovariances.com**), WinGslib (**http://www.gslib.com**), Gocad (**http://www.gocad.com**), Fsstools (**http://www.fssintl.com**), Earth Vision (**http://www.dgi.com**), or as source libraries such as gslib (**http://www.gslib.com**) and the geostatistical template library GsTL (**http://pangea.stanford.edu/~nremy/GTL/GsTL_home.html**).

Applying geostatistics first involves an exploratory data analysis. If we greatly simplify the procedure, the data exploration consists of analyzing the experimental variogram of the data in order to identify the most appropriate variogram model. The exploratory data analysis can be applied either to a continuous variable or to an indicator variable that represents the presence or absence of a geological object (lithofacies).

When the variogram model is inferred, kriging is used to interpolate at any location conditioned to data within the neighborhood and the variogram. Kriging provides an estimation of the expected value of the variable and its variance that represents the possible error of estimation at the same location. Maps obtained by kriging have already been shown in Figure 6 and Figure 9.

When the characterization process requires an estimation of a property P_1 that does not linearly depend on the property P_2 for which data is available, it is necessary to use nonlinear geostatistical techniques. For example, the statistical expectation E() being a linear operator (i.e. an arithmetic average), and if P_2 is a nonlinear function of P_1, that is,

$$P_2 = f(P_1) \tag{1}$$

then the expectation of P_2 is not equal to the transform of the expectation of P_1:

$$E[P_2] = E[f(P_1)] \neq f(E[P_1]) \tag{2}$$

It means that it is erroneous to apply a nonlinear function to a kriged map in order to estimate the expected value of the transform.

To circumvent this problem, the most general nonlinear geostatistical technique is the use of stochastic simulations. Instead of estimating the expected value, the principle is to generate a series of equiprobable realizations that are constructed in order to honor the data points, the variogram, and the pdf of the data. One can then apply the nonlinear transformation to each of these maps and calculate the statistics of the results. Another important aspect is that the kriged field is smoother than the data (Figure 13). The simulated field, instead, has the same spatial structure (variogram) as observed in the data (Figure 13), as well as the same pdf, but it is only one possible reality, one equiprobable realization. Estimation (kriging) and simulation are therefore not applicable for the same purposes. Kriging is useful to map the expected value and to identify the main trends in a field. Simulations are useful when predictions, such as flow and transport simulations, must be applied on the field.

When applied to generate equiprobable realizations of a continuous variable, some geostatistical models rely heavily on the use of a multi-Gaussian distribution. This multi-Gaussian character has some unwanted side effects that should be carefully considered before use, namely, the lack of connection of the extreme values at the tails of the probability distributions, that is, it is very difficult for multi-Gaussian-based realizations to display flow channels or flow barriers (Gómez-Hernández and Wen, 1998; Wen and Gómez-Hernández 1998).

As an alternative to multi-Gaussian models, indicator-based geostatistical models were developed. In indicator-based geostatistics, the different classes (or categories) in which the range of variability of the parameter under study could be divided, are independently characterized, thus controlling the spatial correlation of all classes, particularly those at the extreme ends of the distribution. Each indicator class is characterized by its own variogram function. One of the earliest applications of indicator-based geostatistics was discussed by Gómez-Hernández and Srivastava (1990) in the context of the simulation of a sand-shale sequence.

The main advantage of all geostatistical methods is their ability to be conditional to parameter measurements. Thus, the realizations not only have the spatial patterns characterized by the variogram function but also honor the parameter data. As a consequence, the larger the number of conditioning data, the more alike are the generated realizations and less the uncertainty on the predictions based on these conditional realizations.

Multiple-point Geostatistics

A recent development of geostatistics that goes beyond the variogram-based geostatistics (whether Gaussian or indicator) deserves a section of its own. The major criticism to geostatistical methods has come from the proponents of object-based simulations criticizing the difficulty in reproducing intricate geological patterns when the only controlling tool is the variogram (a two-point statistics measure). For instance, it is very difficult to generate meandering-like depositional patterns, or realizations respecting certain stratigraphical ordering of the facies generated. A solution to this problem was proposed by Strebelle (2002) that departs from traditional geostatistics. Conditional probability values are computed directly from conditioning data using the exact geometrical pattern of the surrounding data with respect to the point being estimated. (In traditional geostatistics, these conditional probabilities are computed

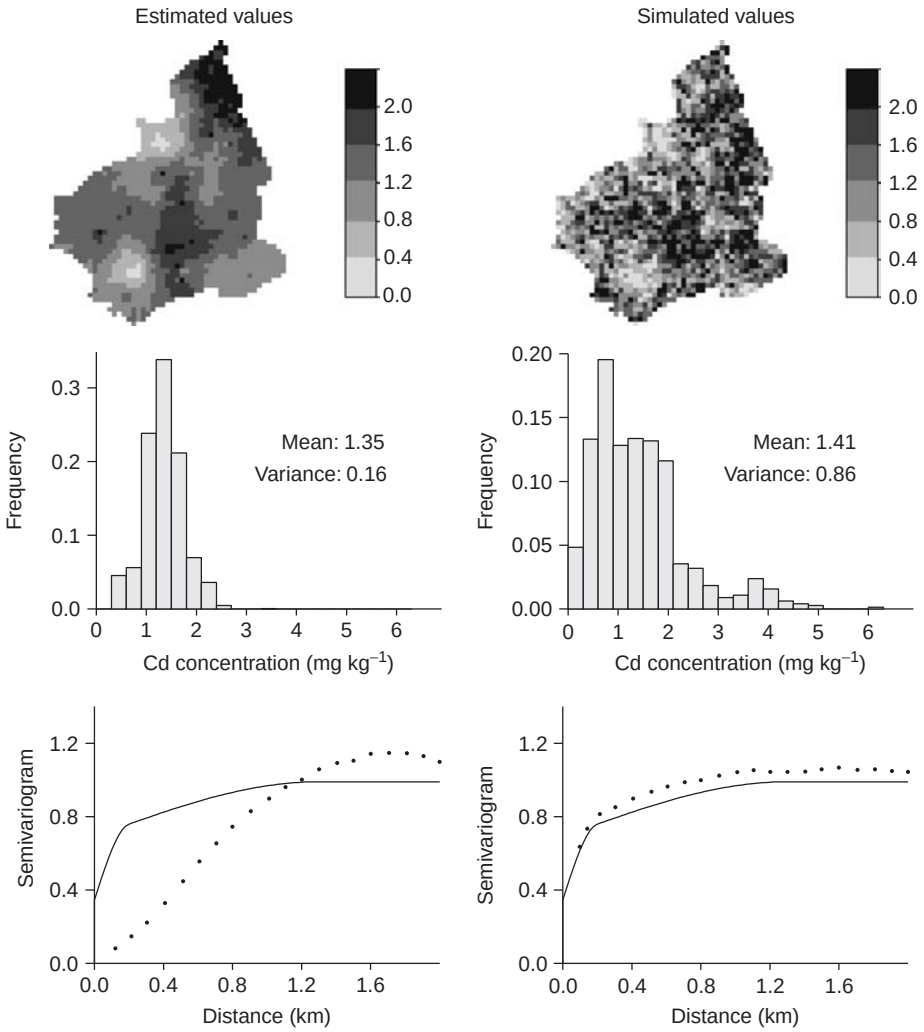


Figure 13 A typical example showing the differences between a map of Cadmium (Cd) interpolated with ordinary kriging (a) and simulation (b). The figure shows the differences in pdfs and variograms calculated *a posteriori* for the two maps. The solid line represents the variogram model and the dotted line represents the experimental variogram calculated *a posteriori* on the interpolated fields (Reproduced from Goovaerts, 1997 by permission of Oxford University Press)

by considering only the separation vectors between each pair of data, and between each datum and the point being estimated.) Evaluating the conditional probability in the way proposed by Strebelle requires establishing the probability distributions for any possible data configuration. Since this is impossible to perform from sample data, Strebelle suggests the use of training images derived from outcrops, expert knowledge, or even a geologist's drawing. He also suggests to only use the nearby data; therefore, reducing the number of conditioning data configurations for which the probability distribution has to be derived.

Multiple-point geostatistics is capable of generating realizations that are very similar to those obtained with object-based algorithms, with the advantage that it can be made conditional by construction, therefore surpassing the main problem of object simulation.

The main criticism to multiple-point geostatistics is the selection of a training image from which to infer the multiple-point statistical model, especially in three dimensions. Outcrops are good for 2D realizations, but there are no three-dimensional outcrops, and it is not trivial to combine multiple 2D training images into a single 3D one. A possible solution to the problem of 3D training images would be to use a genetic model or an object-based model to generate a realization from which to infer the multipoint statistical model. Multiple-point realizations based on such a model will look like the ones obtained from the genetic or object-based models, but will be conditional to well data. Another caveat is that its implementation for practical applications is full of difficulties and computational tricks if CPU-times are to be kept reasonable.

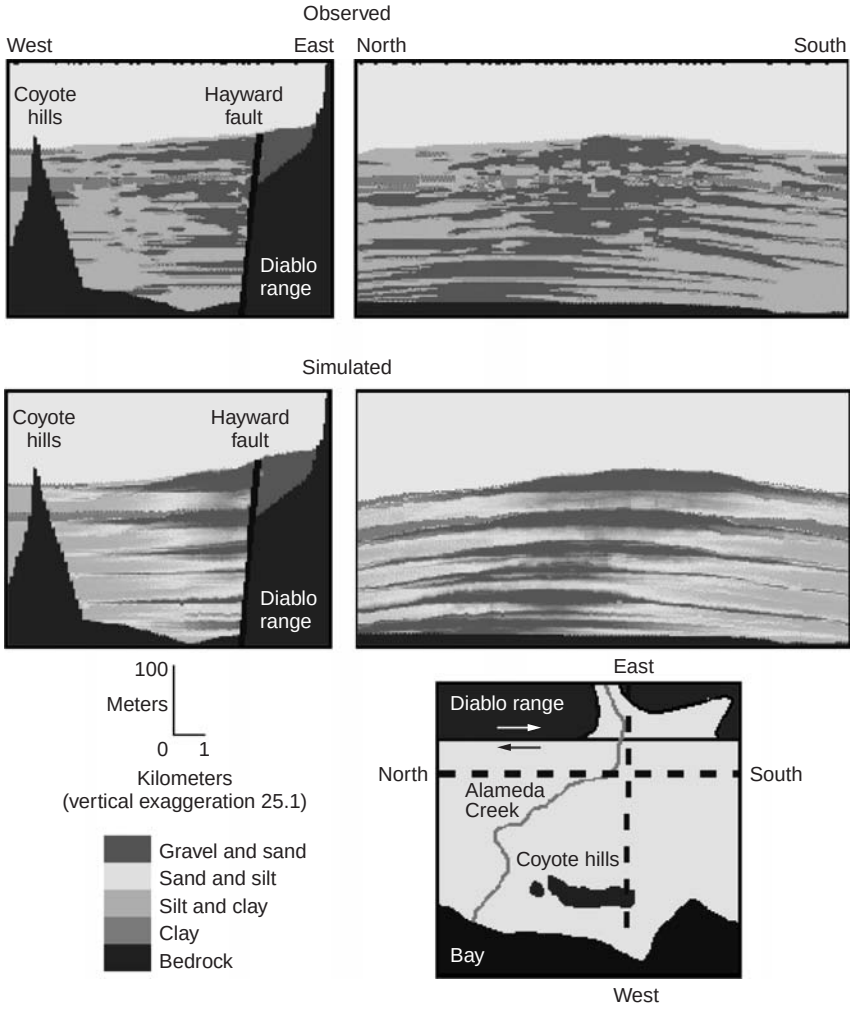


Figure 14 Reconstruction of the internal architecture of the sedimentary complex of the San Francisco Bay, California by a process-imitating model (Reprinted with permission from Koltermann, C. E., and S. M. Gorelick. 1992. Paleoclimatic signature in terrestrial flood deposits. *Science* 256, 1775–1782. © 1992 AAAS) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

GENETIC MODELS

Genetic models, which should not be confused with the genetic algorithms used in global optimization, assign parameter values to the elements of the model by simulating the genesis of the aquifer. One of the first models capable of generating realizations of facies is SEDSIM. This program was developed by Tetzlaff and Harbaugh (1989), and it simulates the genesis of a sedimentary basin by modeling the processes of erosion, transport, and sedimentation. SEDSIM moves fluid particles over a 2-D grid, in which sediments from multiple types continuously mix. The basic principle of SEDSIM is that a fluid element moves down the slope as velocity increases its capacity to erode and pick up sediments, then when it finishes descending the slope, it slows down, its transport capacity decreases, and the sediments are deposited.

These models are mechanicist, in the sense that they reproduce the mechanical processes involved in the genesis of sedimentary basin. They must be run over tens of thousands of years, and require initial and boundary conditions that are difficult to estimate, as well as identifying of the external stresses that drive the processes. For example, information of the initial spatial distribution of the material that will be eroded is needed, as well as information about pluviometry over the entire simulation time. Because most of these inputs are impossible to determine and, at most, they are drawn from predefined probability distributions, these models cannot be called deterministic, even though they use deterministic models to obtain the spatial representation of the parameters.

One of the successful applications of the method (Figure 14) was performed by Koltermann and Gorelick (1992). They simulated the genesis of an alluvial fan-aquifer

system in north-central California (US). For this purpose, they had to collect local and regional geologic and climatic data, and hydrologic history of the study area. They also had to address sea level change, fault motion, sediment loading, compaction, porosity relations, and paleoclimate-driven fluctuations in floods and sediment loads. In addition, they had to simulate flood events using a stochastic streamflow time series. The geometry and geology of the fan-aquifer was simulated for 6 00 000 years. The output grain size distributions from the process model were transformed into porosity and hydraulic conductivity values using petrophysical relations.

Genetic models such as the ones described here are computationally intensive; however, they produce realistic images of large-scale sedimentary structures provided the model inputs are carefully constrained. The realizations are realistic at a large scale; however, locally it is very difficult to condition them to specific porosity or conductivity values at certain locations. This last caveat of genetic methods could be their major drawback. Lately, some researchers have been working on the problem of conditioning, such as Karssenberg *et al.* (2001). Others have developed agent models in order to trigger the behavior of the sediments while reducing the computational load (Teles *et al.*, 2001).

INVERSION METHODS

Inverse theory is concerned with the problem of making inferences about physical systems from data (directly measured or remotely sensed). Since nearly all data are subject to some uncertainty, these inferences are usually statistical. Further, since one can only record finitely many (noisy) data and since physical systems are usually modeled by continuum equations, no inverse problems are really uniquely solvable: if there is a single model that fits the data, there will be an infinite number of them. Our goal then is to characterize the set of models that fit the data and satisfy our prejudices as well as other information. This section describes how to determine model parameter values. Models are assumed to be valid; the only unknowns are parameter values that define the models. For completeness, we introduce some concepts and terminology commonly used in inverse/forward problem community. Detailed coverage of the inverse problem using deterministic tools and stochastic tools is given in **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4** and **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4** respectively.

Well-posed versus Ill-posed Problem

Prediction based on a given set of parameter values is called *forward modeling*. Determination of parameter value from observed data is called *inverse modeling*. Inversion requires minimizing the discrepancy between predictions and observations. Inversion can be achieved in two ways. On one hand, a modeler iteratively modifies parameter values (such as hydraulic conductivity), and runs a forward model (i.e. ModFlow, FeFlow) until attaining best "fit" or "match". This kind of process falls into the trial and error methods. Such forward modeling is sometimes tedious and time consuming. On the other hand, an inverse algorithm can be adopted to automatically or semiautomatically obtain the parameter values from the observed data and an initial set of trial parameters values. The procedure also provides an estimate of parameter uncertainty and resolution.

A well-posed inverse problem requires "existence" of the problem, the "uniqueness", and the "stability" of the solution or algorithm. Obviously, in view of the observed data and our understanding of a real-world physical system, a problem is presumed to exist, for example, detection of contaminant plume in groundwater suggests that contamination must have happened in the past. The question then is how to relate the observation to the migration history of the contaminant. A cause generally has an effect. Can an effect result from different causes? Is it unique in theory or model? Even if it is, have we counted and resolved all parameters that define a model?

Inverse uniqueness has two levels: the model itself and the model-defining parameters. The latter is related to the stability of a solution algorithm. How sensitive are parameters to uncertainty of observed data? Are the errors amplified during inversion? Is the inversion algorithm efficient in terms of ease of usage and cost of running the inversion program (complexity)?

Deterministic versus Stochastic Inversion

An inverse model attempts to obtain a spatial distribution of the parameter values, so that the simulated state of the system, using forward flow and transport models, reproduce the observed state of the system at those locations. Because the relationship between state variables and parameters is not linear, conditioning parameter realizations to state variable data is not trivial, and, in general involves nonlinear optimization algorithms, in which the objective function and its gradient are very expensive to evaluate.

To address the inverse problem, two main frameworks have been developed and they are either deterministic or stochastic. In the deterministic framework, the structure of the spatial variability of the parameters is fixed. For example, the aquifer is divided into a number of zones, and each zone is supposed to have a constant hydraulic conductivity; then, the algorithm seeks the best hydraulic conductivity values for which the solution of the flow equation reproduces the state data (Carrera and Neuman, 1986).

However in a stochastic framework, the spatial variability of the parameters is statistically mapped. For example, the overall average and variance, and the variogram of the final realization are specified; this characterization is not enough to fully determine the parameter values at every cell. Then, a spatial realization is sought meeting the statistical constraints, conditional to the parameter values, and so that the forward model of the state of the system matches the observed values. Many alternative realizations can meet the statistical constraints and reproduce the state data. The self-calibrating method by Gómez-Hernández *et al.* (1997) was developed for this purpose. To make these inferences quantitative in either deterministic or stochastic framework, one must answer three fundamental questions:

1. How accurately is the data known, that is, what does it mean to "fit" the data?
2. How accurately can we model the response of the aquifer system? In other words, have we included all the physics in the model that contribute significantly to the data?
3. Finally, what is known about the system independent of the data? This is called *a priori* information and is essential since for any sufficiently fine parameterization of an aquifer system there will be unreasonable models that fit the data too. Prior information is the means by which we reject or down-weight unreasonable models.

Examples of Stochastic Inversion Methods

It has already been pointed out that the final hydraulic property realization cannot disregard the measurement data; they are the only factual knowledge available about the aquifer. However, aquifers are systems, the state of which is described by the spatial distribution of piezometric heads, and by the concentration of the solutes dissolved in water. In general, there is more information about the state of the system than about the parameters that controls it. Therefore, it appears necessary to generate spatial distributions of the parameters that are not only conditional to parameter values, but also consistent with the (partial) knowledge about the state of the system. We will illustrate this stochastic inversion by constraint through three examples (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4** for more details).

Cokriging methods. Rubin and Dagan (1987a, b) used the analytical approach to solve the perturbed flow equation. They calculate $h' = h - E[h]$ and $Y' = Y - E[Y]$ at the points of h, head, and Y, log of transmissivity, measurements and determined analytically the covariance function of h' and the cross-covariance (h', Y') as a function of the covariance of Y. The covariance of Y is function of a set of parameters q (integral scale of Y and its variance). This is actually sufficient to estimate the transmissivity field by cokriging. The cokriging estimator then gives the optimal estimation of Y at any point as follows:

$$Y(x) = \sum_{i=1}^{n_Y} \lambda_i Y_i + \sum_{j=1}^{n_H} \upsilon_j (h_j - E[h_j]) \qquad (3)$$

where the λ_i and the n_j are optimal weights that depend on the position x. The cokriging equations that provide the value of the optimal weights simply require that the covariance functions of Y, of h, and of $h - Y$ be known. They are developed by Rubin and Dagan (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). Rubin and Dagan then calculate by cokriging all the values of Y at the measurement points where Y is known and where it is therefore possible to compare the known value with the one estimated by cokriging – without using the known value of this point in the cokriging equations. As the cokriging estimator is a function of the q parameters, these parameters can thus be optimized to minimize the errors between the estimated and measured Y values. The Maximum Likelihood method was used for their optimization (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). Once the q parameters are known, the cokriging equations give an estimation of Y at all points and a map of Y is obtained (Figure 15).

Bayesian Inversion. For a statistician, an inverse problem is an inference or estimation problem. The data are finite in number and contain errors, as they do in classical estimation or inference problems; the unknown typically is infinite dimensional, as it is in nonparametric regression. The additional complication in an inverse problem is that the data could be directly and indirectly related to the unknown. Bayesian techniques have become more attractive for the hydrogeological communities through the elegant work of Tarantola (1987). One of the fundamental tenets of Bayesian inference is that uncertainty always can be represented as a probability distribution; in particular, the Bayesian approach treats the model as the outcome of a random experiment. The essential defining property of a Bayesian is to talk about the probability $P(H|E)$ of a hypothesis H, given evidence E. Whether one adheres to a Bayesian view, estimators that arise from the Bayesian approach have an attractive property, that is, the posterior pdf is at least as informative as prior one. In this case, the likelihood function is called *diffusive* or totally noninformative, and the prior estimates are exactly equal to the posterior estimates. It is emphasized that the method does not always guarantee better estimates for a couple of reasons. First, the Bayesian approach

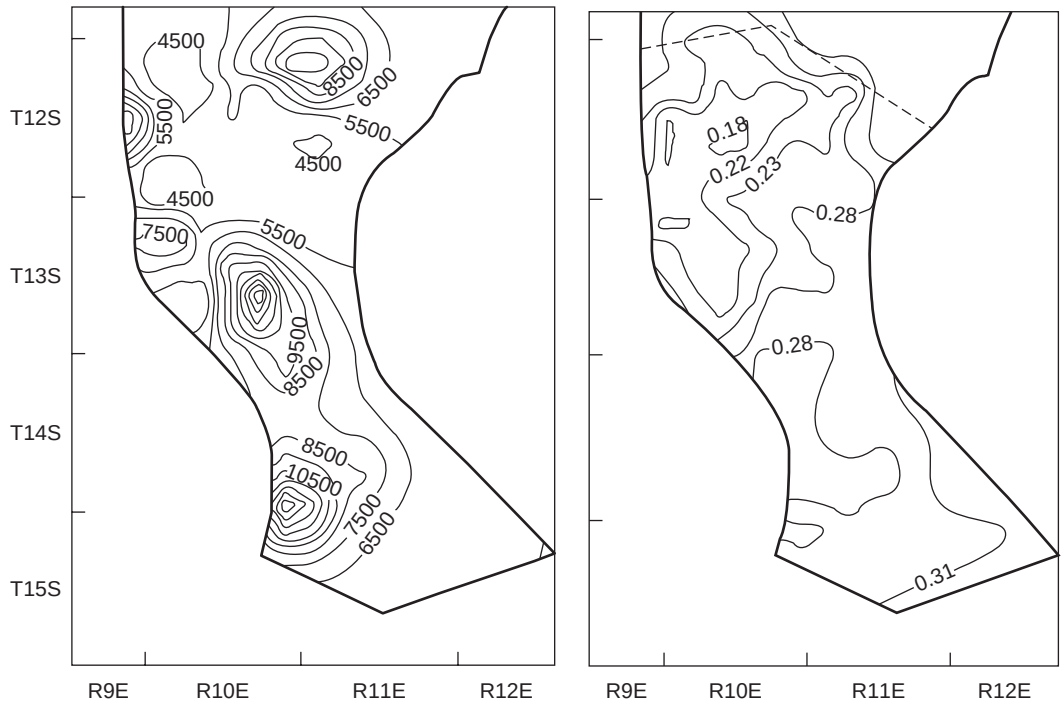


Figure 15 Estimated transmissivity T in ft^2 day^{-1} (a) and the conditional variance of ln[T] (b) based on cokriging and maximum-likelihood estimation (Reproduced from Rubin and Dagan, 1987b by permission of American Geophysical Union)

provides a pdf, not a single-valued estimate. Second, the improvement achieved in the posterior pdf is dictated by the quality of external factors such as the accuracy of the geophysical survey and the petrophysical model in the case of geophysical–hydrogeological stochastic joint inversion.

Bayesian inversion is illustrated in Ezzedine *et al.* (1999). Their hierarchal approach is intended to integrate and transform the well log data to a form in which it can be updated by the geophysical survey, and this tends to be a convoluted process. They started with generating images of the lithology, conditional to well logs. Each lithology image is then used as the basis for generating a series of shaliness images, conditional to well log data. Shaliness images are converted to resistivity images using a site-specific petrophysical model relating between shaliness, resistivity, and lithology, to create the necessary interface with the cross-well resistivity survey. The lithology and resistivity images are then updated using cross-well electromagnetic resistivity surveys. They explored the limits of the approach through synthetic surveys of different resolutions and error levels, employing the relationships between the geophysical and hydrological attributes that are weak, nonlinear, or both. The synthetic surveys closely mimic the conditions at the Lawrence Livermore National Laboratory (LLNL) Superfund site. Ezzedine *et al.* (1999) showed that the proposed stochastic Bayesian approach improves hydrogeological site characterization even when using low-resolution resistivity surveys (Figure 16).

Self-calibrating Stochastic Inversion

The self-calibrated algorithm (Gómez-Hernández *et al.*, 1997) is the first algorithm specifically aimed at the generation of hydraulic conductivity fields conditional to hydraulic conductivity and transmissivity data without resorting to any approximation of the state equation or linearization of the relationship between head and conductivity. It has been later extended to the generation of realizations conditioned to concentration data by Sahuquillo *et al.* (1999) and Hendricks Franssen *et al.* (2003).

In the self-calibrating approach, multiple realizations of the parameters controlling groundwater flow movement and mass transport, that is, hydraulic conductivity, transmissivity, or specific storage, are generated conditioned to values of the parameters and of the state variables. That the realizations are conditioned to the parameter values means that all realizations display the patterns of variability and cross-correlation observed in the field and modeled by a random function, and, at the same time, each realization honors the measured parameter values at their measurement locations. That the realizations are conditioned to the state variables means that the solution of the groundwater flow and mass transport equations with the parameter realizations generated results in the prediction of the state of the system that

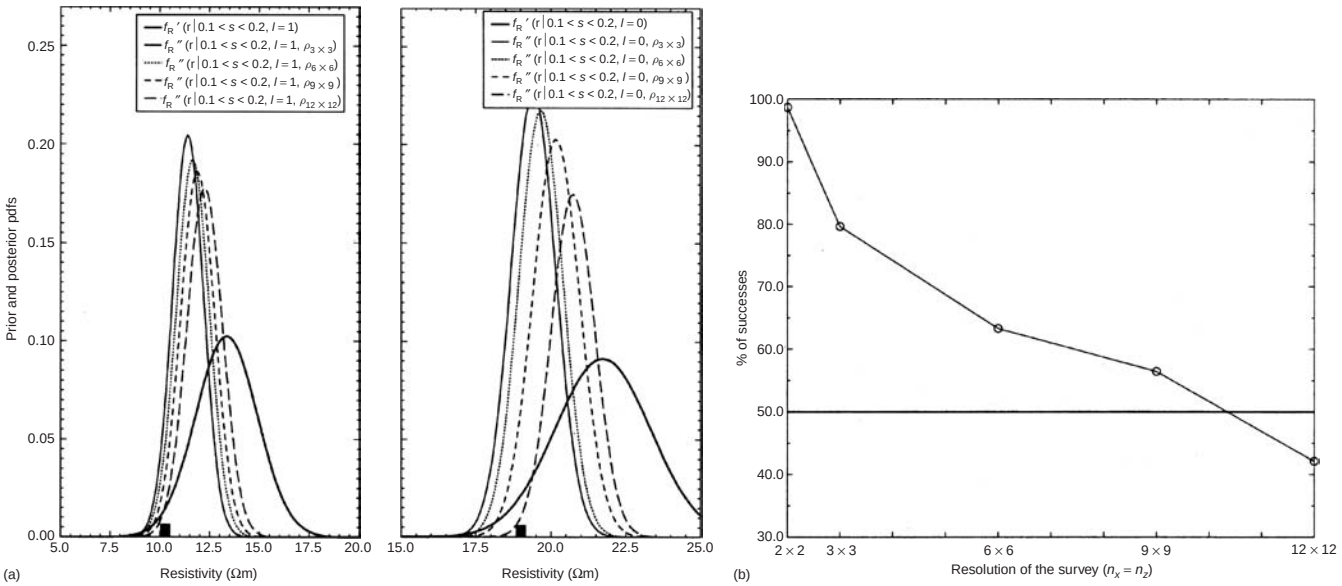


Figure 16 (a) Effect of the block resolution of the $n_x \times n_z$ resistivity survey on the posterior pdf's (prior pdf's are also plotted). The bias in the variance and the mean decrease with the increase of the resolution of the resistivity survey (from 12×12, 9×9, 6×6, to 3×3). The black box denotes ''true'' resistivity values. Prior and posterior pdf's for shaliness between 0.1 and 0.2 in silt (left) and sand (right). (b) Percentage of number of successes of the Bayesian updating approach for different survey resolutions and different errors in the surveys (Reproduced from Ezzedine *et al.*, 1999 by permission of American Geophysical Union)

honors the spatiotemporal measurements of the state variables. Achieving such a dual conditioning amounts to solve a deterministic inverse problem for each realization, something that can only be done after a careful parameterization of the spatial variability of the realizations and efficient computational algorithms.

CONCLUSION

Looking back at the models described here, it is concluded that the best alternative to characterize the spatial variability of a given parameter is through the use of hybrid models. It is important to capture the architecture of the different facies in the aquifer, as it is to capture the variability of the parameters within each facies. Hybrid models start by using a genetic model, an object-based model, or any of the geostatistically-based models capable to generate facies realizations, to generate the spatial distribution of the different facies present in the aquifer; then a geostatistical model (either Gaussian or non-Gaussian) is used to fill in each facies with spatial distributions of the parameters. An example of this approach can be found in Cox *et al.* (1994), who used the cross-sectional geologic images created with a genetic model to estimate the parameters of an indicator-based spatial statistical model. Then, conditioning and generating multiple realizations of hydraulic conductivity were achieved with a geostatistical model.

REFERENCES

Bahralolom I., and Heller J. (1991) Core sample heterogeneity from laboratory flow experiments. In *Reservoir Characterization II*, Lake L.W., Carroll H.B. Jr and Wesson T.C. (Eds.), Academic Press: pp. 77–101.

Bonnet E., Bour O., Odling N.E., Davy P., Main I., Cowie P. and Berkowitz B. (2001) Scaling of fracture systems in geological media. *Reviews of Geophysics*, **39**(3), 347–383.

Bour O., Davy P., Darcel C. and Odling N.E. (2002) A statistical scaling model for fracture network geometry, with validation on a multiscale mapping of a joint network (Hornelen Basin, Norway). *Journal of Geophysical Research*, **107**(B6), 12, 10.1029/2001JB000176.

Carrera J. and Neuman S.P. (1986) Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Research*, **22**(2), 199–210.

Chilès J.-P. and de Marsily G. (1993) Stochastic model of fracture systems and their use in flow and transport modeling. In *Flow and Contaminant Transport in Fractured Rock*, Bear J., Tsang C.-F. and de Marsily G. (Eds.), Academic Press: pp. 169–230.

Chilès J.-P. and Delfiner P. (1999) *Geostatistics: Modeling Spatial Uncertainty*, John Wiley and Sons.

Clauser C. (1992) Permeability of crystalline rocks. *EOS Transactions*, **73**(21), 233–240.

Courrioux G., Nullans S., Guillen A., Boissonnat J.D., Repusseau P., Renaud X. and Thibaut M. (2001) 3D volumetric modelling of Cadomian terranes (Northern Brittany, France):

an automatic method using Voronoï diagrams. *Tectonophysics*, **331**(1–2), 181–196.

Cox D.L., Lindquist C.L., Bargas C.L., Havholm K.G. and Srivastava R.M. (1994) Integrated modeling for optimum management of a giant gas condensate reservoir, Jurassic Eolian Nugget Sandstone, Anschutz Ranch East Field, Utah Overthrust (USA). In *Stochastic Modeling and Geostatistics*, Yarus J.M. and Chambers R.L. (Eds.), American Association of Petroleum Geologists, Vol. 3, pp. 287–321.

Dagan G. (1989) *Flow and Transport in Porous Formations*, Springer-Verlag: New York.

Delhomme J.P. (1979) Kriging under boundary conditions. *American Geophysical Union Fall Meeting*, San Francisco.

Durlofsky L.J. (1992) Representation of grid block permeability in coarse scale models of randomly heterogeneous porous media. *Water Resources Research*, **28**(7), 1791–1800.

Ezzedine S. (1994) *Modélisation des écoulements et du transport dans les milieux fissurés. Approches continues et discontinues*. PhD thesis, Ecole des Mines de Paris.

Ezzedine S., Rubin Y. and Chen J.S. (1999) Bayesian method for hydrogeological site characterization using borehole and geophysical survey data: theory and application to the Lawrence Livermore National Laboratory Superfund site. *Water Resources Research*, **35**(9), 2671–2683.

Gómez-Hernández J.J., Sahuquillo A. and Capilla J.E. (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric head data, 1, Theory. *Journal of Hydrology*, **203**(1–4), 162–174.

Gómez-Hernández J.J. and Srivastava R.M. (1990) ISIM3D: an ANSI-C three-dimensional multiple indicator conditional simulation model. *Computer and Geosciences*, **16**(4), 395–440.

Gómez-Hernández J.J. and Wen X.-H. (1998) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. *Advances in Water Resources*, **21**(1), 47–61.

Goovaerts P. (1997) *Geostatistics for Natural Resource Evaluation*, Oxford University Press.

Güler C., Thyne G.D., McCray J.E. and Turner A.K. (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal*, **10**, 455–474.

Haldorsen H.H. and Lake L.W. (1982) A new approach to shale management in field scale simulation models. *57th Annual Fall Technical Conference and Exhibition of the Society of Petroleum Engineers of AIME*, Society of Petroleum Engineers: New Orleans, p. 10976.

Haldorsen H.H. and Lake L.W. (1984) A new approach to shale management in field scale simulation models. *Society of Petroleum Engineers Journal*, **24**(4), 447–457.

Hendricks Franssen H.-J., Gómez-Hernández J.J. and Sahuquillo A. (2003) Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data. *Journal of Hydrology*, **281**(4), 281–295.

Jones N.L., Davis R.J. and Sabbah W. (2003) A comparison of three-dimensional interpolation techniques for plume characterization. *Ground Water*, **41**(4), 411–419.

Journel A. (1989) *Fundamentals of Geostatistics in Five Lessons*, American Geophysical Union: Vol. 8.

Jussel P., Stauffer F. and Dracos T. (1994) Transport modeling in heterogeneous aquifer: 1. Statistical description and numerical generation of gravel deposits. *Water Resources Research*, **30**(6), 1803–1817.

Karssenberg D., Törnqvist T.E. and Bridge J.E. (2001) Conditioning a process-based model of sedimentary architecture to well data. *Journal of Sedimentary Research*, **71**(6), 868–879.

Kitanidis P.K. (1997) *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press.

Koltermann C.E. and Gorelick S.M. (1992) Paleoclimatic signature in terrestrial flood deposits. *Science*, **256**, 1775–1782.

Mallet J.-L. (2002) *Geomodeling*, Oxford University Press.

Matheron G. (1962) *Traité de Géostatistique Appliquée*, Technip: Paris.

Matheron G. (1967) *Eléments Pour Une Théorie Des Milieux Poreux*, Masson: Paris.

Mei C.C. and Auriault J.-L. (1989) Mechanics of heterogeneous porous media with several spatial scales. *Proceedings of Royal Society of London, A*, **426**, pp. 391–423.

Renard P. and Courrioux G. (1994) Three-dimensional geometric modeling of a faulted domain: the Soultz horst example. *Computers and Geosciences*, **20**(9), 1379–1390.

Renard P. and de Marsily G. (1997) Calculating equivalent permeability: a review. *Advances in Water Resources*, **20**(5–6), 253–278.

Rubin Y. and Dagan G. (1987a) Stochastic identification of transmissivity and effective recharge in steady-groundwater flow. 1. Theory. *Water Resources Research*, **23**(7), 1185–1192.

Rubin Y. and Dagan G. (1987b) Stochastic identification of transmissivity and effective recharge in steady-groundwater flow. 2. Case Study. *Water Resources Research*, **23**(7), 1193–1200.

Sáez A.E., Otero C.J. and Rusinek I. (1989) The effective homogeneous behaviour of heterogeneous porous media. *Transport in Porous Media*, **4**(3), 213–238.

Sahuquillo A., Hendricks Franssen H.-J., Gómez-Hernández J.J. and Capilla J.E. (1999) Computational aspects of the coupled inversion of groundwater flow and mass transport. In *ModelCARE 99*, Stauffer F., Kinzelbach W., Kovar K. and Hoehn E. (Eds.), IAHS Press.

Sánchez-Vila X., Carrera J. and Girardi J.P. (1996) Scale effects in transmissivity. *Journal of Hydrology*, **138**, 1–22.

Scheibe T.D. and Freyberg D.L. (1995) Use of sedimentological information for geometric simulation of natural porous media structure. *Water Resources Research*, **31**(12), 3259–3270.

Schulze-Makuch D. and Cherkauer D.S. (1998) Variation in hydraulic conductivity with scale of measurement during aquifer tests in heterogeneous, porous carbonate rocks. *Hydrogeology Journal*, **6**, 204–215.

Strebelle S. (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, **34**(1), 1–22.

Tarantola A. (1987) *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier.

Teles V., Bravard J.P., de Marsily G. and Perrier E. (2001) Modelling of the construction of the Rhône alluvial plain since 15000 years BP. *Sedimentology*, **48**, 1209–1224.

Tetzlaff D.A. and Harbaugh J.W. (1989) *Simulating Clastic Sedimentation*, Van Nostrand Reinhold: New York.

Thibaut M., Gratier J.P., Léger M. and Morvan J.M. (1996) An inverse method for determining three-dimensional fault geometry with thread criterion: application to strike-slip and thrust faults (Western Alps and California). *Journal of Structural Geology*, **18**(9), 1127–1138.

Tidwell V.C. and Wilson J.L. (1999) Upscaling experiments conducted on a block of volcanic tuff: Results for a bimodal permeability distribution. *Water Resources Research*, **35**(11), 3375–3387.

Tyler K., Henriquez A. and Svanes T. (1994) Modeling heterogeneities in fluvial domains: a review of the influence on production profiles. In *Stochastic Modeling and Geostatistics*, Yarus J.M. and Chambers R.L. (Eds.), American Association of Petroleum Geologists, Vol. 3, pp. 77–89.

Wen X.-H. and Gómez-Hernández J.J. (1998) Numerical modeling of macrodispersion in heterogeneous media: a comparison of multiGaussian and non-multiGaussian models. *Journal of Contaminant Hydrology*, **30**(1–2), 129–156.

Zlotnik V.A., Zurbuchen B.R., Ptak T. and Teutsch G. (2000) Support volume and scale effect in hydraulic conductivity: experimental aspects. In *Theory, Modelling, and Field Investigations In hydrogeology: A Special Volume in Homor of Shlomo P. Neuman's 60th birthday*, Vol. Special Paper 348, Zhang D. and Winter L.L. (Eds.), Geological Society of America: Boulder, pp. 215–231.

148: Aquifer Characterization by Geophysical Methods

HARRY VEREECKEN, ANDREAS KEMNA, HANS-MARTIN MÜNCH, AXEL TILLMANN AND ARRE VERWEERD

Agrosphere Institute (ICG-IV), Forschungszentrum Jülich GmbH, Jülich, Germany

Geophysical methods are increasingly being used to characterize the subsurface. They offer the potential to derive basic characteristics, state variables, and properties of geological formations. In this article, we focus on the application of geophysical methods to derive properties and state variables of aquifer systems. Special attention is given to the assessment of hydraulic conductivity, porosity, and water saturation. Three different groups of methods (seismics, electrical techniques, and electromagnetics) and their combined use are discussed. For each method, relationships between geophysical and hydrogeological quantities are outlined and applications are presented. A separate section is devoted to the development and application of combined approaches in hydrogeophysics.

INTRODUCTION

The shallow subsurface is an extremely important zone that yields much of our water resources (*see* **Chapter 145, Groundwater as an Element in the Hydrological Cycle, Volume 4**). It also serves as the repository for municipal and industrial waste. As sustainable and effective use of the subsurface environment is a major challenge facing modern societies, there is a great need to improve our understanding of the shallow subsurface and, in particular, groundwater systems. In this respect, geophysical methods can contribute in a significant manner.

In the past, application of geophysical methods in groundwater hydrology has focused on mapping geological structures (e.g. clay/sand layers, bedrock valleys), delineation of aquifer boundaries, mapping of fracture zones and chemical pollution plumes, delineation of water-saturated zones, and seepage flow in landslide bodies. For these purposes, standard methods are presently available and well documented in the geophysical literature (e.g. Parasnis, 1997; Reynolds, 1997).

In recent years, increased attention has been given to the use of geophysical methods to derive parameters and state variables characterizing especially near-surface groundwater systems and soils. Research in this direction is mainly driven by the fact that geophysical methods allow continuous mapping of geophysical properties, which can be transferred to parameters or variables characterizing the aquifer system (e.g. water content, porosity, flow velocity, and their changes in time). Classical approaches like drilling, coring, and well testing (*see* **Chapter 151, Hydraulics of Wells and Well Testing, Volume 4**) have shown their limitations in constraining spatial and temporal variabilities of these quantities. Characterizing such variabilities is however of utmost importance for, for example, determining the success of water management strategies or predicting pollution risks for water supply systems.

In this article, we will present the application of geophysical methods to derive parameters (e.g. hydraulic conductivity, porosity, dispersivity) and state variables (e.g. water saturation) of aquifer systems relevant for water flow and solute transport (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**). This area of research is today referred to as hydrogeophysics (e.g. Vereecken *et al.*, 2002; Rubin and Hubbard, 2005). Hydrogeophysics is a rapidly evolving discipline aiming at integrating hydrological, hydrogeological, and geophysical methods and concepts to characterize the subsurface. Rather than covering the whole range of available geophysical methods, we

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

focus on selected methods that are promising in terms of site-specific hydrogeophysical characterization of aquifer systems. Three groups of methods will be discussed: seismics, electrical techniques, and electromagnetics, the latter including ground-penetrating radar. In the last section, we will address the combined use of geophysical and hydrogeological data to characterize aquifer properties.

SEISMIC AQUIFER CHARACTERIZATION

Seismic characterization of the subsurface is based upon the propagation of elastic waves generated through a seismic source at the surface or in boreholes. Owing to the nature of elastic media, there exist different wave types propagating with different velocities. The primary information derived from seismic methods comprises the wave velocities and wave attenuation. From these quantities information can be derived about material properties of the subsurface like porosity, hydraulic conductivity, elastic moduli, and water saturation.

Seismic-Hydrogeological Relationships

Seismic velocities are dependent on aquifer properties (e.g. grain material, texture, and porosity) and state variables (e.g. water saturation). The fast compressional wave (also known as primary wave or P-wave) is related to the bulk and shear moduli of a medium, and its propagation velocity, v_p [m s^{-1}], is

$$v_p = \sqrt{\frac{K + \frac{4}{3}G}{\rho_b}} \tag{1}$$

where K [Pa] and G [Pa] are the bulk and shear moduli, respectively, and ρ_b [kg m^{-3}] is the bulk density. P-waves occur in solids, fluids, and gases. The slower shear wave (secondary or S-wave) does not depend on the bulk modulus; its velocity, v_s [m s^{-1}], is given by

$$v_s = \sqrt{\frac{G}{\rho_b}} \tag{2}$$

Shear waves do not exist in fluids and gases since their shear moduli are zero.

The elastic moduli of sediments depend on the characteristics of the solid matrix, the porosity, and the properties of the pore space. The effect of water saturation on the seismic velocities is well-known and subject of many theoretical (Gassmann, 1951; Biot, 1956a,b; Kuster and Toksöz, 1974a) and experimental (Wyllie *et al.*, 1956; Kuster and Toksöz, 1974b) investigations. Under the constraint that mineral composition and porosity are the primary factors controlling the dry-frame elastic moduli, Nolen–Hoeksema (1993) derived a first-order approximation of the relation between porosity and the dry-frame bulk and shear moduli, K_{dry} [Pa] and G_{dry} [Pa], respectively:

$$K_{dry} = K_{solid}\left(1 - \frac{n}{n_0}\right), \quad G_{dry} = G_{solid}\left(1 - \frac{n}{n_0}\right) \tag{3}$$

Here, K_{solid} [Pa] and G_{solid} [Pa] are the bulk and shear moduli of the solid matrix material, n [−] is the porosity, and n_0 [−] denotes the precompaction porosity, which defines the transition from unconsolidated ($n > n_0$) to consolidated ($n < n_0$) sediments. For media consisting of spheroidal grains, n_0 is in the range of 0.35 to 0.40. A typical rock contains a mixture of grain/pore sizes and shapes, leading to a more complicated relationship between elastic moduli and porosity (see e.g. Toksöz *et al.*, 1976).

For sediments with a homogeneous and isotropic pore space, Gassmann's equation (Gassmann, 1951) describes the influence of fluids on the bulk modulus for low-frequency elastic waves:

$$K = K_{solid}\frac{nK_{dry} + Q}{nK_{solid} + Q} \tag{4}$$

with

$$Q = K_{fluid,eff}\frac{K_{solid} - K_{dry}}{K_{solid} - K_{gas}} \tag{5}$$

where K_{gas} [Pa] denotes the bulk modulus of the gas phase, and $K_{fluid,eff}$ [Pa] is the effective bulk modulus of the fluid–gas mixture. Equations (4) and (5) are valid for consolidated materials, that is, $n < n_0$, and therefore, from equation (3), $K_{dry} > 0$. For a homogeneous fluid–gas mixture in the pore space, $K_{fluid,eff}$ is given by

$$K_{fluid,eff} = \left(\frac{S}{K_{fluid}} + \frac{1 - S}{K_{gas}}\right)^{-1} \tag{6}$$

where S [−] is the fluid (water) saturation and K_{fluid} [Pa] the bulk modulus of the fluid phase. According to equations (4) to (6), an increase of water saturation gives rise to an increase of the bulk modulus of sediments, while water does not contribute to the shear modulus. Therefore, the P-wave velocity (equation (1)) shows a stronger variation caused by saturation changes than the S-wave velocity (equation (2)), which is only influenced through changes in bulk density

$$\rho_b = (1 - n)\rho_{solid} + nS\rho_{fluid} \tag{7}$$

where ρ_{solid} [kg m^{-3}] is the mass density of the solid matrix and ρ_{fluid} [kg m^{-3}] the water density. Accordingly, the v_p/v_s ratio is suitable to distinguish dry from saturated sediments and detect phreatic surfaces, that is, water tables.

For illustration, Figure 1 shows the P-wave and S-wave velocities modeled for a porous medium consisting of

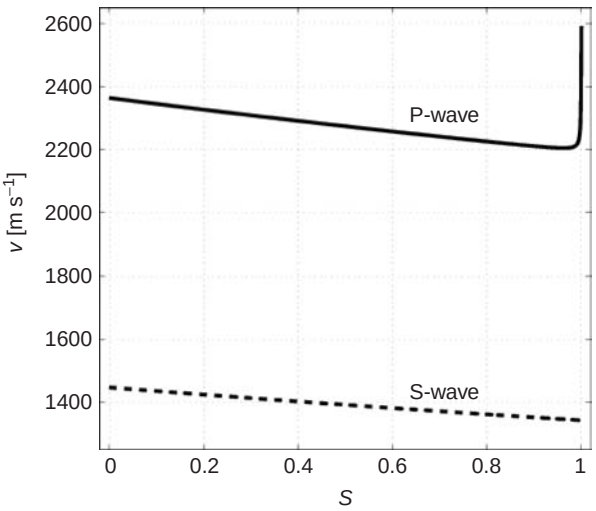


Figure 1 Seismic P- and S-wave velocities as function of water saturation modeled for a porous medium consisting of quartz grains (*see* Tillmann, 2001)

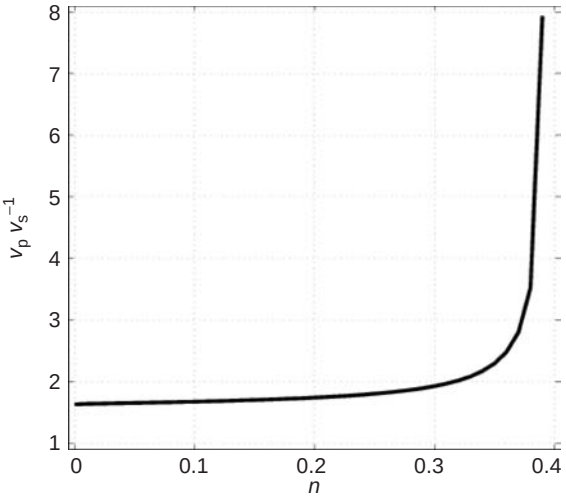


Figure 2 Seismic P-wave to S-wave velocity ratio as function of porosity modeled for a porous medium consisting of quartz grains (*see* Tillmann, 2001)

quartz grains with $n = 0.3$ (for more details, see Tillmann, 2001) as a function of water saturation. The slight decrease of the velocities with increasing saturation at lower to medium saturation degrees is due to the dominant effect of the increase of bulk density with increasing water saturation in the pore space. Near full saturation, the complete filling of pores with water causes a rapid increase of the bulk modulus and the P-wave velocity. Equations (3) to (6) lead to the conclusion that this P-wave velocity increase is larger for high porosities and negligible for low porosities compared to velocity perturbations owing to changes in soil material. The v_p/v_s ratio of the quartz grain model at full saturation as a function of porosity is shown in Figure 2. For n approaching the precompaction porosity, here $n_0 = 0.4$, the ratio becomes infinite since the dry matrix shear modulus becomes zero. Figures 1 and 2 illustrate the potential of seismic methods to detect the transition zones to full saturation and to estimate porosity using a combined interpretation of P-wave and S-wave data.

Empirical relationships between P-wave velocity, porosity, permeability, clay content, bulk density, and water saturation for sandstones and carbonate rocks were derived for example by Koesoemadinata and McMechan (2003) using correlation analyses between the different quantities. Marion *et al.* (1992) obtained velocity-porosity relationships for sediments on the basis of a simple geometric-mechanical model for sand–clay mixtures. From porosity, in turn, permeability can be estimated using the Kozeny–Carman relation (Kozeny, 1927; Carman, 1937).

Seismic Methods

Seismic methods are based on the propagation of elastic waves in the earth. From the experimental setup, different methods can be distinguished (for an overview, *see* e.g. Telford *et al.*, 1990). In the seismic refraction method, the first arrival times of the seismic waves are measured. The derivation of seismic velocities from measured travel times (inversion) is based on Snell's law from ray theory, which implies the assumption of a layered earth. The method is sensitive to velocity changes between two layers as long as critically refracted waves are generated at the layer boundaries. Therefore the method requires an increase in seismic velocity with depth, which is a severe constraint for many shallow applications where low-velocity layers may be encountered.

In contrast to seismic refraction, seismic reflections occur owing to changes in the elastic impedance of the medium, that is, the product of velocity and density in a layer. Therefore, low-velocity layers and density changes can be detected too, as long as the wavelength of the seismic signal is small compared to the layered structure. From this it follows that measurements in shallow sediments have to be performed using signals in the kHz range, which unfortunately are affected by high attenuation.

In seismic crosshole tomography, travel time measurements for a large number of ray paths are collected providing information on the seismic velocity distribution. While the application of this technique in the field involves P-wave travel time measurements between boreholes, in the vertical seismic profiling (VSP) method it is common to use surface-to-borehole shot-receiver configurations. Seismic crosshole tomography gives a relatively detailed velocity model between the boreholes, and does not require the assumption of a layered earth.

Surface wave analysis exploits the fact that a particular wave type is formed owing to the free boundary condition

at the Earth's surface, the amplitude of which drops exponentially with depth. Surface wave interpretation is normally based on the assumption of a horizontally layered earth (Haskell, 1953). Surface waves are dispersive, that is, the propagation velocity as well as the penetration depth depends not only on material parameters but also on frequency. The analysis of this dispersive behavior leads to information about all elastic parameters and the depths of shallow layer boundaries.

Hydrogeological Applications

As pointed out in Section "Seismic-hydrogeological relationships", the typically wide range of elastic properties of sediments normally exceeds any perturbations caused by the presence of water, which makes near-surface water detection by means of seismic methods difficult. Nevertheless, seismic methods have been successfully used for mapping water table depths in aquifers (e.g. Birkelo *et al.*, 1987; Bacharach and Nur, 1998; Baker *et al.*, 2000; Bradford and Sawyer, 2002). However, the water table is not a trivial boundary in seismic terms. Generally, it is defined through the pore fluid properties (such as compressibility, density, and viscosity) and the degree of water saturation (depending on pore pressure and pore-size distribution). Therefore, its location may vary over a relatively short distance in both space and time. Seismic methods can be used to monitor temporal groundwater variations by detecting associated changes in seismic velocity. For instance, Bacharach and Nur (1998) followed this approach for monitoring the water table at a beach under tidal influence. On a longer time scale, Baker *et al.* (2000) monitored seasonal water table changes over one year. Tillmann and Stöcker (2001) presented a joint inversion approach for seismic and electrical data (*see* Section "Electrical aquifer characterization") taking lithological and hydrological parameters into account. In contrast to conventional joint inversion methods (e.g. Hering *et al.*, 1995), their approach can handle the presence of a water table.

As a field demonstration we present the results of a seismic crosshole tomography survey performed to resolve the porosity distribution in a near-surface aquifer. The survey was conducted at the Krauthausen test site, which is located close to Jülich in the western part of Germany. More details about the site can be found in Vereecken *et al.* (2000). In 1997, a series of tomographic measurements was made to characterize the spatial variability of porosity. Figure 3(a) shows the reconstructed P-wave velocity image along a cross section containing several boreholes. Details on the used tomographic reconstruction technique and further results are reported in Dietrich and Fechner (1997). Taking into account additional information from borehole

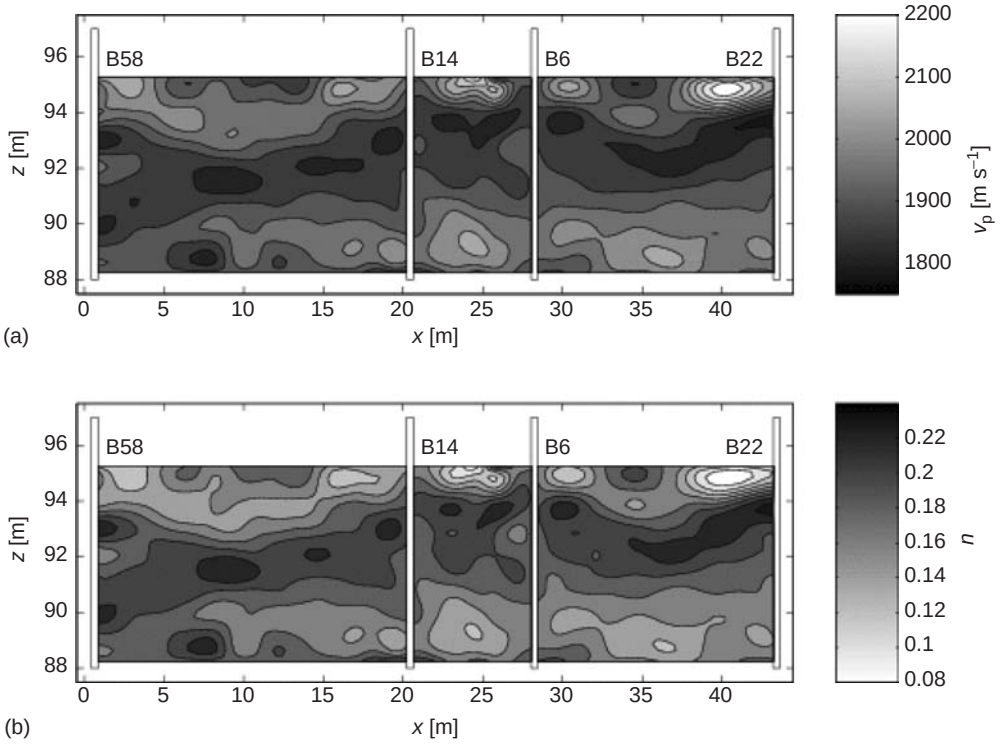


Figure 3 Results of a crosshole seismic tomography survey conducted at the Krauthausen test site along a transect containing four boreholes (B58, B14, B6, B22) to characterize a near-surface aquifer: (a) P-wave velocity distribution, (b) estimated porosity distribution

measurements, the porosity was derived from the P-wave velocity using the approach of Yamamoto *et al.* (1994). The resultant porosity distribution is shown in Figure 3(b). The porosities derived from the seismic data showed good correlation with those determined on cores from the well locations.

ELECTRICAL AQUIFER CHARACTERIZATION

It has been recognized for a long time that measurements of subsurface electrical properties can be used for the investigation of the underground, in particular concerning groundwater resources (e.g. Vacquier *et al.*, 1957). Since solid and aqueous phases of a porous soil/rock system electrically act in a different way, electrical properties of soils and rocks are closely related to the underlying structure. Importantly, in predominantly ion-conductive systems electrical and hydraulic conduction follow similar pathways through the interconnected saturated pores – a circumstance that makes electrical methods particularly attractive for hydrogeological aquifer characterization.

Electrical-Hydrogeological Relationships

In addition to electrolytic conduction in the saturated pores, described by an electrolytic conductivity contribution σ_{el} [S m^{-1}], electrical surface conduction can exist along the interfaces between solid and aqueous phases owing to the presence of electric surface charges and the involved formation of electrical double layers, leading to a surface conductivity contribution σ_s [S m^{-1}]. Following the conceptual model that both conduction mechanisms act in parallel (e.g. Waxman and Smits, 1968; Rink and Schopper, 1974), the bulk electrical conductivity, σ [S m^{-1}], of sediments is given by

$$\sigma = \sigma_{el} + \sigma_s \tag{8}$$

For clay free, saturated formations, the bulk electrical conductivity may be expressed as

$$\sigma_{sat} \cong \sigma_{el,sat} = \frac{\sigma_w}{F} = n_{eff}^m \sigma_w \tag{9}$$

(Archie, 1942; Sen *et al.*, 1981), where σ_w [S m^{-1}] denotes the water electrical conductivity, and F [−] is a formation factor depending on the pore space geometry (porosity, tortuosity, effective pore/grain size), usually written in dependence on effective porosity, n_{eff} [−], and a so-called cementation exponent m [−] (for unconsolidated sands, typically $m \approx 1.3$, *see* e.g. Schön, 1996). The surface conductivity contribution in equation (8) may be expressed as

$$\sigma_{s,sat} = \frac{2\mu_s Q_s}{F\Lambda} \tag{10}$$

(Johnson *et al.*, 1986), where Λ [m] is an intrinsic measure of interconnected pore size, Q_s [C m^{-2}] is the surface ionic charge density, and μ_s [m^2 V^{-1} s^{-1}] is the effective ionic mobility in the electrical double layer around the charged surface. In general, ionic charge density and mobility depend on water chemistry, that is, on σ_w. The pore-size parameter Λ accounts for different surface and bulk tortuosity factors; its inverse is closely related to the surface area to pore volume ratio.

In general, σ may also describe induced electrical polarization (IP) phenomena attributed to ion-selective membrane effects in the pore space, in particular, associated with present clay minerals (e.g. Marshall and Madden, 1959; Vinegar and Waxman, 1984; Lesmes and Frye, 2001). Then, σ becomes a complex-valued quantity, σ^*, comprising real (σ') and imaginary (σ'') parts (or magnitude and phase). In the frequency range relevant for practical applications (typically <100 Hz), however, virtually only σ_s contributes to the imaginary part and, hence, σ'' allows direct access to internal surface characteristics. Importantly, σ_s generally depends on the applied measurement angular frequency ω [s^{-1}]. The relaxation times of transient IP processes directly correspond to the respective spatial scales at which these processes take place in the pore space. This leads to a dispersive behavior $\sigma_s(\omega)$ that contains important additional information for textural characterization (e.g. Kemna *et al.*, 2000; Lesmes and Morgan, 2001; Scott and Barker, 2003).

The dependence of real and imaginary conductivities on water saturation S is normally expressed as a simple power-law relation. For clay free formations, where only the electrolytic conductivity contributes to the real conductivity, it is

$$\sigma' \cong \sigma_{el} = \sigma_{el,sat} S^p \tag{11a}$$

(e.g. Archie, 1942). Here p [−] is the so-called saturation exponent with typically $p \approx 2$ (*see* e.g. Schön, 1996). For the imaginary conductivity, the corresponding relationship is

$$\sigma'' = \sigma''_{sat} S^q \tag{11b}$$

(e.g. Vinegar and Waxman, 1984), but recent studies (e.g. Ulrich and Slater, 2004) indicate that the saturation exponent of the imaginary conductivity, q [−], is significantly smaller than the saturation exponent of the real part (p). However, the saturation dependency of induced polarization is still an issue of current research (e.g. Titov *et al.*, 2004).

The relationships between structural, in particular textural, aquifer properties and the bulk electrical conductivity have motivated the attempt to assess hydraulic conductivity, K_h [m s^{-1}], or permeability, k [m^2], from geoelectrical measurements. Early approaches were based on the linear regression analysis of $\log\sigma - \log K_h$ correlations, which provides calibration functions valid for a particular set of

samples or field site (e.g. Mazáč *et al.*, 1985; Huntley, 1986). For instance, Heigold *et al.* (1979) found the field-scale empirical relation

$$\log K_h \approx -1.28 + 0.93 \log \sigma \tag{12}$$

(K_h in [m s^{-1}], σ in [S m^{-1}]) for the Niantic-Illiopolis aquifer in Illinois, USA. However, generally both a direct and an inverse relation between logσ and logK_h can be observed, depending on lithology type. In relatively clean sands, σ increases with increasing K_h since both properties are similarly controlled by porosity (dominance of σ_{el}). In clayey material, clay content takes over the governing role and, accordingly, an opposite σ-K_h relationship is found: σ increases but K_h decreases with increasing clay content (dominance of σ_s).

Recently, different empirical to semiempirical models were proposed for electrical k estimation (e.g. Börner *et al.*, 1996; de Lima and Niwas, 2000; Purvance and Andricevic, 2000; Slater and Lesmes, 2002). One approach (Börner *et al.*, 1996) is based on an observed direct relation between the surface area to pore volume ratio, A_p [m^{-1}], and σ'' (*see* Börner and Schön, 1991) in conjunction with a modified Kozeny–Carman model,

$$k \propto \frac{1}{F A_p^u} \tag{13}$$

where the formation factor F is estimated from measurements of σ_w. The exponent u [−] can be related to the fractal dimension of the pore surface (Pape *et al.*, 1987), that is, to soil/rock type. In contrast, the controlling influence on k might be attributed to a characteristic grain size, r_{gr} [m], rather than a characteristic pore size (such as A_p or Λ), by adopting a Hazen type model,

$$k \propto r_{gr}^w \tag{14}$$

with generalized exponent w [−], in conjunction with an empirically found inverse relation between r_{gr} and σ'' (Slater and Lesmes, 2002).

Geoelectrical Methods

With geoelectrical methods the subsurface bulk electrical conductivity σ, or resistivity $\rho = \sigma^{-1}$ [Ωm], is measured. Measurements are usually accomplished by injecting a DC, or quasi DC (low-frequency), electric current into the ground using a pair of electrodes and measuring the resultant voltage drop between another pair of electrodes (Figure 4). This defines a transfer resistance (or impedance, if also IP effects are considered) measurement, which contains information about a characteristic region of the underground, primarily depending on the measurement geometry (Figure 5). Accordingly, from a set of measurements with different geometry a model of the subsurface electrical conductivity distribution can be derived.

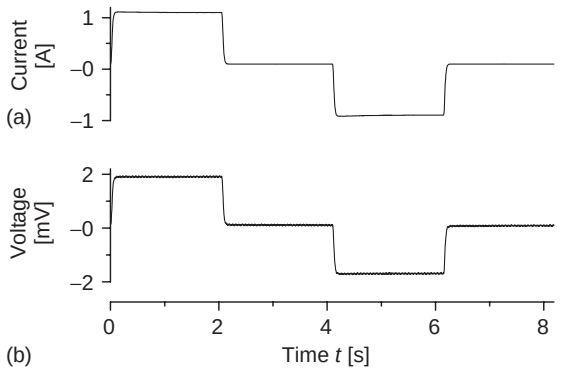


Figure 4 Typical signal waveforms in geoelectrical surveying: (a) injected square-wave current, and (b) observed voltage response after application of a low-pass filter to reduce power line noise

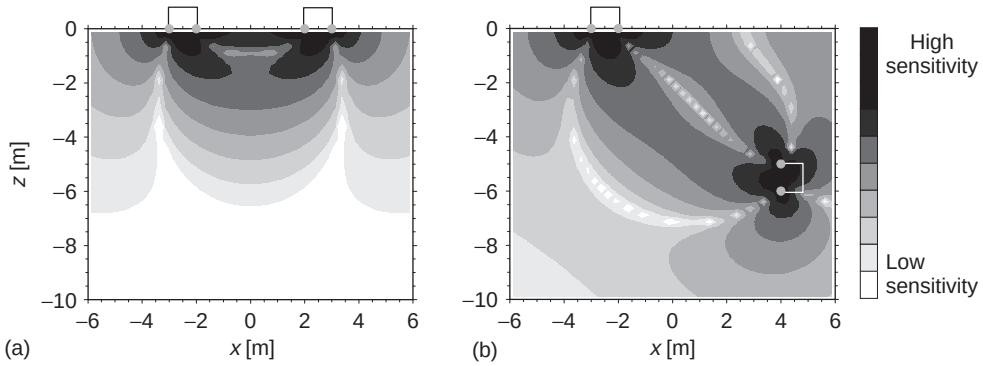


Figure 5 Sensitivity patterns for typical geoelectrical measurements using two separated electrode pairs for current injection and voltage measurement (so-called dipole–dipole configuration) assuming a homogeneous ground: (a) both electrode pairs placed at the surface, (b) electrode pairs placed at the surface and in a borehole, respectively. Grey circles indicate position of electrodes (Reproduced from Kemna, 1996, by permission of Landesamt fur Natur & Umwelt der Lander Schlesing-Holstain)

Early approaches were restricted in terms of the geometric parameterization of the underground as well as electrode placement. For example, in classic geoelectrical soundings (see e.g. Telford *et al.*, 1990) a multilayer earth model is assumed and only surface electrodes are used. However, with the advances in instrumentation and computational capabilities more flexible electrical imaging methods were developed, yielding high-resolution images of the subsurface electrical conductivity distribution based on surface and/or borehole measurements (for an overview, see Binley and Kemna, 2005). Following the terminology in medical imaging, this approach is nowadays usually referred to as electrical resistance (or resistivity) tomography (ERT).

ERT theory is based on the Poisson equation, which relates the electric potential, $\phi(\mathbf{r})$ [V], owing to any current sources to the underlying electrical conductivity distribution, $\sigma(\mathbf{r})$. For a point source (electrode) at the origin, with current strength I [A], it is

$$\nabla \cdot (\sigma \nabla \phi) = -I\delta(\mathbf{r}) \tag{15}$$

where δ denotes the Dirac delta function. For arbitrary 2D and 3D electrical conductivity models, usually finite element or finite difference methods are used to numerically solve the modeling (forward) problem defined by equation (15) subject to given boundary conditions (for an overview, *see* Hohmann, 1988). The ERT imaging (inverse) problem is inherently nonlinear, nonunique, and strongly ill-posed. However, in most imaging algorithms (e.g. Loke and Barker, 1995; LaBrecque *et al.*, 1996) nonuniqueness and ill-posedness are overcome by applying some sort of regularization, that is, imposing some constraints (such as smoothing) on the model. More recently, these approaches have been extended to also include IP effects in the interpretation (Oldenburg and Li, 1994; Kemna *et al.*, 2000).

Hydrogeological Applications

Geoelectrical methods have been extensively used for hydrogeological purposes. Early applications focused on the characterization of aquifers by means of geoelectrical soundings and subsequent hydrogeological interpretation of the derived electrical 1D models (e.g. Kosinski and Kelly, 1981), usually by taking hydraulic (borehole) data into account. Although correlations between, for example, electrical conductivity and hydraulic conductivity provided some calibration capability, these early applications were generally only of qualitative nature. However, with the improved understanding of soil/rock electrical signatures, today electrical-structural models are available that bear the potential for an improved quantitative aquifer characterization, in particular if applied in conjunction with high-resolution electrical imaging methods (Kemna *et al.*, 2004). With further development of these approaches, it may become possible to obtain, for example, images of *in-situ* hydraulic conductivity (for inverse modeling *see* **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**) with relatively high spatial resolution (*see* Figure 6), which obviously is of extreme practical relevance.

In addition to direct structure characterization, geoelectrical methods have been likewise proven to be efficient for the monitoring of subsurface flow and transport processes (*see* **Chapter 78, Models of Water Flow and Solute Transport in the Unsaturated Zone, Volume 2** and **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**) when applied in a time-lapse mode. Earlier studies include the monitoring of groundwater flow by profiling or mapping the electrical response to an injected saline tracer, either from the surface (e.g. White, 1994; Morris *et al.*, 1996) or with the help of borehole electrodes (e.g. Bevc and Morrison, 1991). From the electrical response, for example, the effective groundwater flow direction and velocity can be determined. Using modern ERT methodology, the imaging of subsurface water or solute

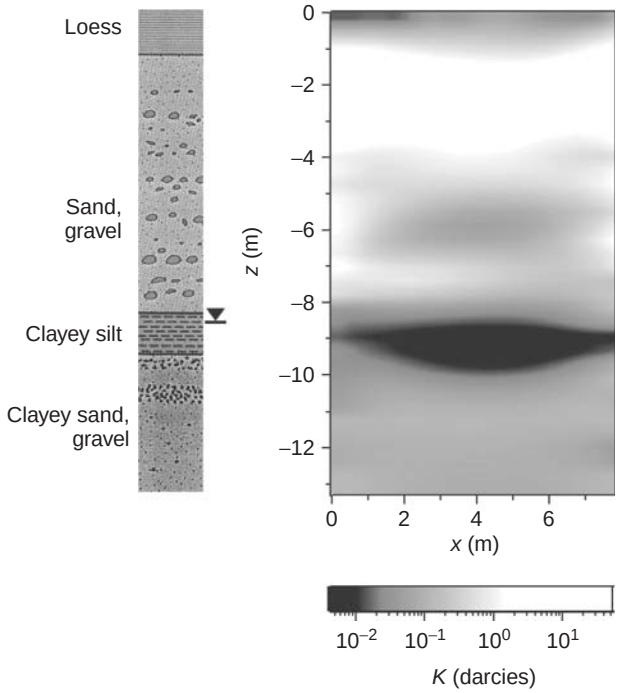


Figure 6 *In-situ* estimate of hydraulic permeability in a typical Quaternary environment (lithological setting shown on the left) as derived from cross-borehole electrical imaging results in conjunction with the application of a semiempirical electrical-hydraulic model according to equation (13) (Reproduced from Kemna *et al.,* 2004, by permission of Society of Exploration Geophysicists)

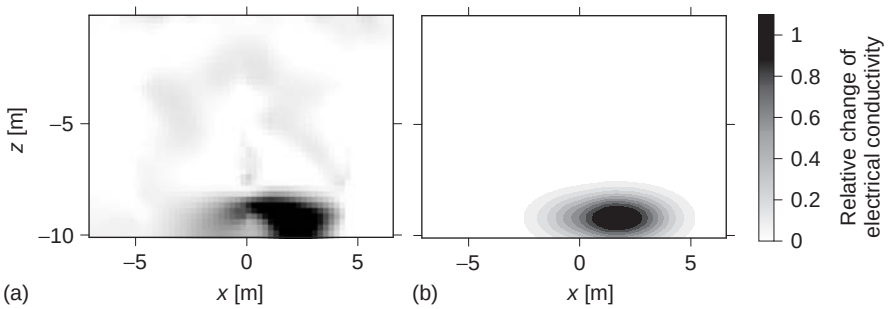


Figure 7 (a) ERT derived map of temporal electrical conductivity changes in the course of a tracer experiment conducted in a heterogeneous aquifer at the Krauthausen test site, Germany. (b) Corresponding result of an equivalent 3D advection-dispersion model fit, assuming a linear relation between relative electrical conductivity change and tracer concentration (Reprinted from Kemna *et al.*, 2002. © Copyright 2002, with permission from Elsevier)

plume movement is possible with high spatial resolution (e.g. Daily *et al.*, 1992; Slater *et al.*, 2000; Yeh *et al.*, 2002). Recently, it was demonstrated how cross-borehole ERT results in the course of a field tracer experiment can be used to quantify the variability of parameters relevant to flow and transport in heterogeneous aquifers (Kemna *et al.*, 2002) (for flow and transport parameters and models *see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4** and **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**). By interpreting images of temporal electrical conductivity changes by means of equivalent transport models (Figure 7), longitudinal and lateral spreading of a tracer plume, as well as the degree of mixing and the heterogeneity of transport within the plume, can be quantified in terms of fitted equivalent dispersivities (Kemna *et al.*, 2002). The spatial variability of equivalent transport parameters contains information about the spatial structure of the flow field, which in turn can be used to infer information about the structure of the hydraulic conductivity field (e.g. Rubin and Ezzedine, 1997; Vanderborght and Vereecken, 2001) (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**).

ELECTROMAGNETIC AQUIFER CHARACTERIZATION

In electromagnetic (EM) methods, electromagnetic fields of various sources (e.g. controlled transmitter coils, radio transmitters, telluric currents) are used to investigate the electromagnetic properties of the subsurface. Numerous different approaches exist depending on the employed source type, the measurement principle (e.g. time domain vs. frequency domain), and the analyzed time/frequency range (*see* e.g. Telford *et al.*, 1990).

In this section, we will distinguish between low-frequency EM methods (typically 1–100 kHz), including frequency-domain and time-domain approaches, and the high-frequency ground-penetrating radar (GPR) method (typically 10–1000 MHz). Both methodological fundamentals as well as hydrogeological applications will be addressed. Since low-frequency EM methods primarily characterize the subsurface in terms of electrical conductivity, it is here again referred to Section "Electrical-hydrogeological relationships" where the basis of its hydrogeological interpretation is provided. In the GPR method, however, primarily the dielectric permittivity is analyzed and therefore in the corresponding section dielectric-hydrogeological relationships are likewise reviewed.

Low-Frequency Electromagnetic Methods

Frequency-domain Methods

A typical frequency-domain (FD) EM system consists of two coils (loops), a transmitter and a receiver coil (Figure 8). In the transmitter coil an alternating current (frequency typically ranging from 1 to 100 kHz) generates a magnetic field, the so-called primary field, H_p^* [A m^{-1}] (where $*$ denotes a complex quantity comprising real and imaginary parts). According to Faraday's induction law, the time-varying primary magnetic field induces a time-varying electric field in the earth. Owing to the conductive nature of the earth, this electric field is associated with time-varying electric currents, the so-called eddy currents, according to Ohm's law. These eddy currents in turn generate a magnetic field, the so-called secondary field, H_s^* [A m^{-1}], which is picked up by the receiver coil together with the primary field. In typical FD-EM methods the mutual coupling of such a two-coil system is measured, which can be described as the ratio of the secondary to primary magnetic fields. From this, information on the electrical conductivity distribution in the subsurface can be derived.

Maxwell's equations form the theoretical basis of the general EM problem. For the considered case of a two-coil system, however, an apparent electrical conductivity (i.e. the conductivity of an equivalent homogeneous half-space),

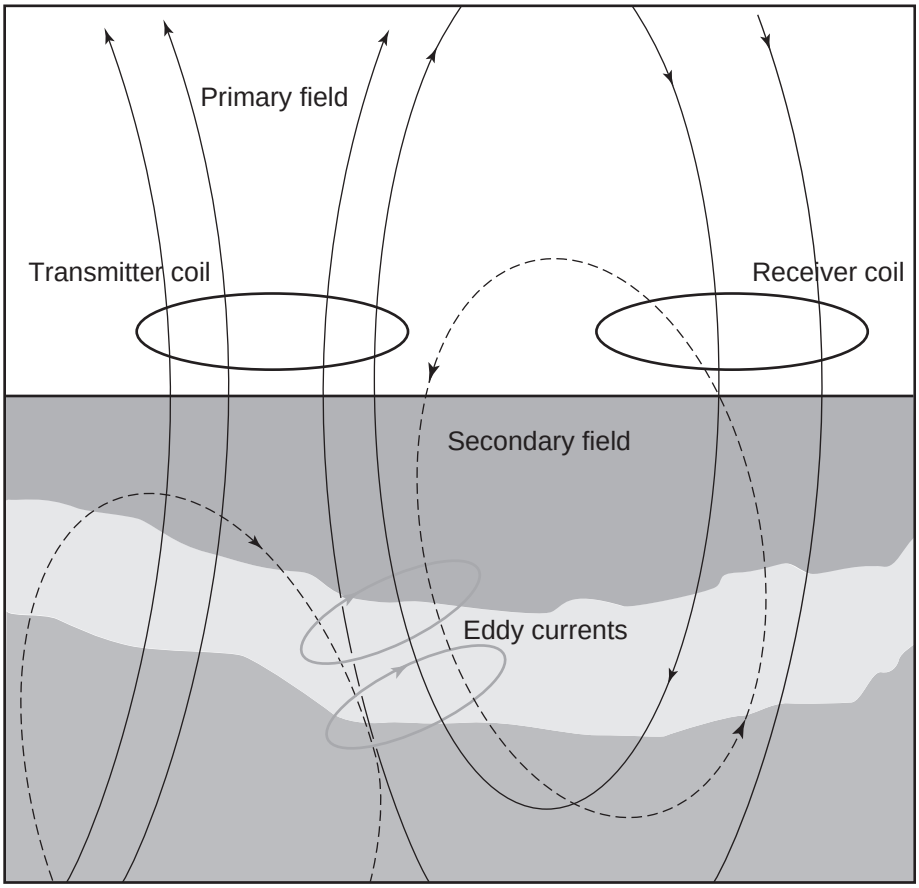


Figure 8 Measurement principle of a typical two-coil (loop–loop) frequency-domain EM system

σ_a [S m^{-1}], can be simply computed from the mutual coupling ratio H_s^*/H_p^* according to (McNeill, 1980a)

$$\sigma_a = \frac{4}{\omega\mu_0 r^2}\left(\frac{H_s^*}{H_p^*}\right)'' \tag{16}$$

if $\mu \approx \mu_0$ is assumed and

$$Q \equiv r\sqrt{\frac{\omega\mu_0\sigma}{2}} \ll 1 \tag{17}$$

where $''$ denotes the imaginary (quadrature) component, ω [s^{-1}] is the measurement angular frequency, $\mu_0 = 4\pi \cdot 10^{-7}$Vs A^{-1}m^{-1} the magnetic permeability of free space, μ [Vs A^{-1}m^{-1}] the subsurface magnetic permeability, r [m] the spacing between transmitter and receiver coils, Q [−] is the so-called induction number, and σ [S m^{-1}] an estimated maximum value for the subsurface electrical conductivity. The assumption of low induction numbers (equation (17)) is valid for most sediments and frequencies of commercially available EM systems operating in the low-frequency range.

While the interpretation of "apparent" electrical conductivities may be sufficient in certain situations, generally knowledge of the "true" electrical conductivity distribution in the subsurface is desired. This can be derived, however, by the application of inversion schemes similar to those used in geoelectrics, as outlined in Section "Geoelectrical methods", in conjunction with appropriate electromagnetic forward modeling routines (for an overview, *see* Oristaglio and Spies, 1999).

The induction number Q is closely related to another important parameter used in FD-EM methods, particularly for survey design, that is, the skin depth (e.g. Telford *et al.*, 1990)

$$\delta = \sqrt{\frac{2}{\omega\mu_0\sigma}} \tag{18}$$

[m]. The skin depth corresponds to the depth where the transmitted EM signal is reduced to e^{-1} of its original value at the surface; it is directly related to the exploration depth of a FD-EM survey. For a homogeneous earth, obviously $\delta = r/Q$. Equation (18) clearly shows the effect of both subsurface electrical conductivity and measurement frequency on the attenuation of EM fields. The more

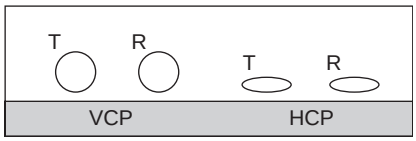


Figure 9 Vertical coplanar (VCP) and horizontal coplanar (HCP) orientations of a loop–loop EM system consisting of transmitter (T) and receiver (R) coils

conductive the earth, the stronger are the EM fields attenuated by Ohmic loss, and vice versa.

Besides the skin depth, also coil spacing and orientation are connected with the exploration depth. With increasing coil spacing, secondary field responses from greater depths are measured owing to the penetration characteristics of EM fields. For rigidly connected transmitter and receiver coils, two different system orientations are commonly used: horizontal coplanar (HCP) and vertical coplanar (VCP) (Figure 9). In the HCP orientation, the vertical component of the secondary field is measured; in the VCP orientation its horizontal component is measured. The horizontal component is more sensitive to the shallow subsurface than the vertical component. Approximate maximum exploration depths are 1.5 and 0.75 times the coil spacing for, respectively, HCP and VCP orientations (McNeill, 1980a).

Since the exploration depth of a FD two-coil system depends on measurement frequency as well as coil spacing and orientation, different surveying modes are possible. Many commercial systems operate in a wide frequency band in order to enable vertical soundings of the subsurface electrical conductivity via systematic frequency variation for a fixed measurement position at the surface. Alternatively, in order to map the electrical conductivity laterally the method is typically applied in a profiling mode by moving the system with fixed coil spacing along a transect while using a single measurement frequency. Via combination of both sounding and profiling modes, and also in conjunction with varying coil spacing, multidimensional imaging can be realized.

Besides the two-coil (loop–loop) configuration outlined above, there exist other FD-EM measurement approaches that utilize external (or far-field), either artificial or natural, sources as the primary field. In magnetotellurics (MT), for instance, natural ionospheric currents are used as external sources. Far-field artificial sources are primarily represented by radio transmitters originally operated for military communication or marine/aerial navigation purposes; they are used in the very-low-frequency (VLF), or radio-magnetotellurics (RMT), method. At the receiver side, in addition to coils picking up magnetic field components, also electrodes may be placed in the ground to measure electric field responses. Depending on the source type and which field components are measured, specific data processing techniques have been developed for the different approaches to infer information on the electrical conductivity distribution in the subsurface (for an overview, *see* Nabighian, 1991).

Importantly, EM methods solely based on magnetic field measurements do not require galvanic contact with the ground and thus bear inherent advantages over geoelectrical methods (*see* Section "Geoelectrical methods"). Accordingly, EM methods are widely used as an airborne reconnaissance method to enable a swift and relatively cheap mapping of the electrical conductivity distribution. Coils can be located either on a plane or helicopter, or on a rigid boom towed by the aircraft. Similarly, marine surveys are possible, where the coils are typically located on a cable suspended in the water or dragged along the seafloor. EM methods may also be superior to electrical methods in highly resistive environments (provided that induction effects are still significant), where high contact resistances at the electrodes often impede a deeper penetration of injected direct currents. In conductive environments, on the other hand, the attenuation of EM fields is relatively large and hence the use of electrical methods may be more advantageous.

Time-domain Methods

Time-domain electromagnetic (TD-EM) methods also use a transmitter and a receiver coil (Figure 10). A static primary magnetic field is generated by a steady current in the transmitter coil. When the current is sharply turned off, the associated abrupt change in the magnetic field induces an electromotive force in the ground, which gives rise to decaying eddy currents ("smoke rings") flowing in subsurface conductors. These currents in turn generate a secondary magnetic field, which is measured at the receiver coil and from which an "apparent" electrical conductivity can be calculated (McNeill, 1980b). By means of application of appropriate modeling and inversion routines, the "true"

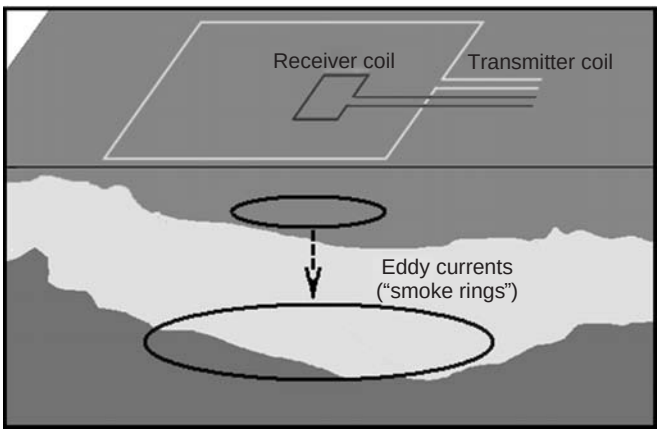


Figure 10 Layout of a typical time-domain EM survey showing induced electric currents in the subsurface

electrical conductivity distribution in the subsurface can also be obtained (Oristaglio and Spies, 1999).

Time-domain EM transmitter coils have relatively large dimensions (coils with an area of several $100\,m^2$ are not uncommon), whereas the receiver coils have smaller dimensions. Especially in urban areas, the signal-to-noise ratio is significantly improved by using small receiver coils. The receiver coil can be located inside the transmitter coil, at a certain distance, or the transmitter coil itself can be used as receiver. After the transmitter current is turned off, measurements may take several seconds. They are then repeated several times to increase the signal-to-noise ratio, each time with a different current polarity. This reduces the influence of polarization effects in the subsurface on the measured signal. TD-EM methods are typically used as a sounding tool by interpreting early and late-time responses, analogous to the spectral analysis in FD approaches.

Hydrogeological Applications

Since all EM methods aim to map the subsurface electrical conductivity distribution, they represent, in principle, an alternative to geoelectrical methods for all identified hydrogeological applications (*see* Section "Electrical aquifer characterization"). EM methods are typically preferred for investigations at a larger (e.g. aquifer) scale (e.g. Edet, 1991; Puranen *et al.*, 1999; Danielsen *et al.*, 2003). Under favorable conditions (little anthropogenic interference, presence of high-conductive dissolved solids in the groundwater, or presence of conductive contaminants/tracers), for instance, an image of the groundwater body or a contaminant plume can be obtained (e.g. Wynn, 2002).

Ground-Penetrating Radar

Ground-penetrating radar (GPR) is an EM pulse reflection method that is based on similar principles as reflection seismics (*see* Section "Seismic methods"). With GPR the travel time, like in conventional sonar and radar, and amplitude of high-frequency (10–1000 MHz) EM waves propagating from a transmitter antenna through the subsurface to a receiver antenna are measured. GPR is sensitive to changes in the dielectric permittivity and electrical conductivity, which cause reflection, refraction, and diffraction, as well as attenuation of the EM waves. Variations in these properties influence the measured travel times and amplitudes. Permittivity and electrical conductivity strongly depend on water content, salinity, porosity, grain size, and clay content of sediments, and therefore many subsurface structural characteristics can be detected by GPR. A benefit of GPR is its relatively high spatial resolution, as well as the possibility of on-line visualization of the measuring results.

Dielectric-Hydrogeological Relationships

Existing dielectric-hydrogeological relationships are based on either empirical, phenomenological, or some sort of mixing models (e.g. Shen *et al.*, 1985; Knoll, 1996; Hagrey and Müller, 2000). Empirical models mostly relate the volumetric water content, θ_v [−], to the relative dielectric permittivity (dielectric constant), ε_r [−], which is defined as $\varepsilon_r = \varepsilon/\varepsilon_0$, where ε [$As\,V^{-1}m^{-1}$] is the dielectric permittivity and $\varepsilon_0 = 8.854 \cdot 10^{-12} As\,V^{-1}m^{-1}$ its value in free space. For instance, the relation of Topp *et al.* (1980) is often applied:

$$\theta_v = -5.3 \cdot 10^{-2} + 2.9 \cdot 10^{-2}\ \varepsilon_r - 5.5 \cdot 10^{-4}\ \varepsilon_r^2 + 4.3 \cdot 10^{-6}\ \varepsilon_r^3 \qquad (19)$$

Volumetric mixing formulas relate the bulk permittivity of a multiphase mixture to the permittivities and volumetric fractions of its constituents. A general form is that of Lichtenecker and Rother (1931) (*see* e.g. Birchak *et al.*, 1974):

$$\varepsilon_b^\chi = \sum_i V_i\ \varepsilon_i^\chi \qquad (20)$$

where ε_b is the bulk relative permittivity, V_i [−] and ε_i are the volumetric fraction and relative permittivity, respectively, of the i-th constituent, and χ is an empirical constant accounting for the geometric shape of the solid matrix (grains). For $\chi = 0.5$, one obtains for a three-component mixture of matrix, water, and air a formula similar to the complex refractive index method (CRIM) (e.g. Shen *et al.*, 1985):

$$\sqrt{\varepsilon_b} = (1 - S)n\sqrt{\varepsilon_{air}} + Sn\sqrt{\varepsilon_w} + (1 - n)\sqrt{\varepsilon_m} \qquad (21)$$

with porosity n, saturation $S = \theta_v/n$, and ε_{air}, ε_w, and ε_m denoting the relative permittivities of air, water, and matrix, respectively. For $\varepsilon_w \approx 80$ and for example, $\varepsilon_m = 4$ (possible value for dry soil) and $S = 1$ (full saturation), equation (21) yields $\varepsilon_b \approx 4 + 28n + 48n^2$.

In general, the dielectric properties can be described by a frequency-dependent, complex dielectric permittivity, $\varepsilon^*(\omega)$, comprising real (ε') and imaginary (ε'') parts. The frequency dependence is taken into account by many phenomenological models such as the Cole–Cole model (Cole and Cole, 1941)

$$\varepsilon^*(\omega) = \varepsilon_\infty + \frac{\varepsilon_0 - \varepsilon_\infty}{1 + (i\omega\tau)^c} \qquad (22)$$

Here, ε_0 and ε_∞ represent the (real-valued) low and high-frequency asymptotes of ε^*, respectively, i denotes the imaginary unit ($i = \sqrt{-1}$), c [−] is the so-called Cole–Cole-exponent and τ [s] a characteristic relaxation time.

Measurement Principles

GPR utilizes short EM pulses of broadband dipole antennas that are emitted into the ground. The propagation velocity,

v [m s^{-1}], of the EM waves is directly related to the real part of the relative dielectric permittivity, ε'_r, via

$$v \approx \frac{c_0}{\sqrt{\varepsilon'_r}} \tag{23}$$

where $c_0 \approx 3 \cdot 10^8 \text{ms}^{-1}$ is the vacuum propagation velocity of EM waves and it is assumed that $\sigma \ll \omega\varepsilon_0\varepsilon'_r$ (with σ [S m^{-1}] being the DC electrical conductivity). Typically, dominant frequencies between 10 and 1000 MHz are used, corresponding to wavelengths, $\lambda = 2\pi v/\omega$ [m], in the range 0.1 to 10 m. The wavelength defines the maximum spatial resolution of GPR; layers smaller than $\lambda/4$ cannot be resolved (Jol, 1995). Obviously, a higher frequency provides a higher resolution.

Solving Maxwell's equations for the electric field strength, $\mathbf{E}^*$ [V m^{-1}], in case of an isotropic, homogeneous medium and sinusoidal time dependence exp (iωt) yields for the transversal plane wave traveling in vertical (z [m]) direction $\mathbf{E}^*(t) = \mathbf{E}_0^* \exp(\text{i}\omega t - \gamma^* z)$, where the complex propagation constant $\gamma^* \equiv \alpha + \text{i}\beta$ [m^{-1}] consists of an attenuation constant

$$\alpha = \frac{\omega}{c_0}\sqrt{\frac{\varepsilon'_r}{2}\left(\sqrt{1+\left(\frac{\sigma}{\omega\varepsilon'}\right)^2} - 1\right)} \tag{24}$$

and a phase constant

$$\beta = \frac{\omega}{c_0}\sqrt{\frac{\varepsilon'_r}{2}\left(\sqrt{1+\left(\frac{\sigma}{\omega\varepsilon'}\right)^2} + 1\right)} \tag{25}$$

(*see* e.g. Hagrey and Müller, 2000).

Equations (23) to (25) exhibit the dependence of the propagation behavior of the EM waves on material properties. Water is critical for the propagation in rocks because its dielectric constant ($\varepsilon'_r \approx 80$) is much higher than that of air ($\varepsilon'_r = 1$) and the solid matrix ($\varepsilon'_r \approx 4 \ldots 10$) (for typical values of different materials *see* e.g. Reynolds, 1997). The electrical conductivity is decisive for the attenuation and thus the exploration depth (skin depth), δ [m], defined as $\delta = 1/\alpha$ (cf. equation (18)). In high-conductive environments, the use of GPR is strongly limited (e.g. van Overmeeren, 1994).

The EM waves emitted by the GPR transmitter antenna into the ground are reflected and refracted at boundaries between regions (layers) with different complex impedance, $Z^* = \sqrt{\mu_0/\varepsilon^*}$ [Ω], leading to multiple waves associated with different subsurface raypaths that are recorded with the receiver antenna. Figure 11 shows the different wavepaths that can be identified for a typical GPR survey geometry with two antennas over a layered underground. The corresponding travel time curves are plotted in Figure 12.

The air wave propagates with the speed of light above the Earth's surface; it can be used as a reference to mark the beginning of the measurement. Directly at the surface, the ground wave propagates with the velocity v of the uppermost layer. The corresponding travel times for air and ground waves are, respectively, $t_{air} = x/c_0$ and $t_g = x/v$, where x [m] denotes the distance in horizontal direction. From simple ray theory, the travel time of the reflected wave is given as

$$t_r = \sqrt{\left(\frac{x}{v}\right)^2 + \left(\frac{2d}{v}\right)^2} \tag{26}$$

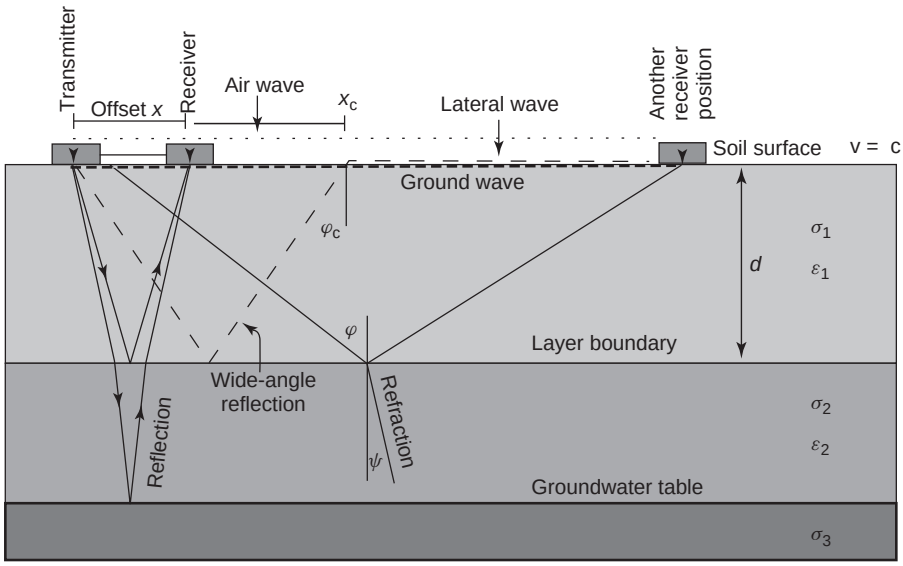


Figure 11 Typical GPR survey geometry showing the different waves that can be observed

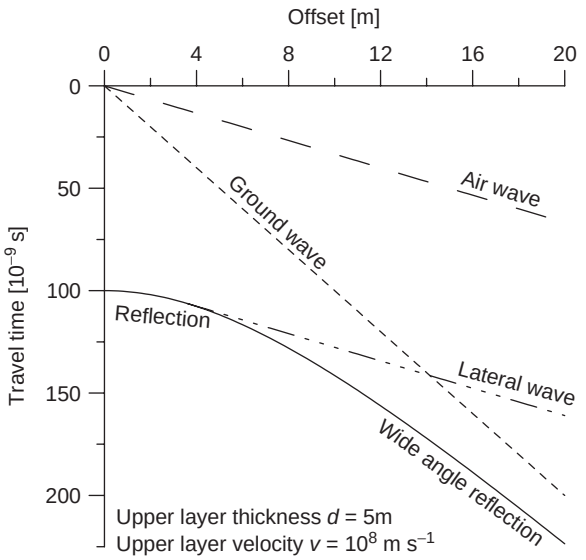


Figure 12 Travel time curves for the different waves shown in Figure 11 for a two-layer model with exemplary parameter values

where d [m] is the thickness of the uppermost layer. When a reflected, up-traveling wave encounters the surface at an angle $\varphi \geq \varphi_{\text{critical}} = \arcsin(v/c_0)$, the so-called lateral wave (*see* e.g. Huisman *et al.*, 2003) with travel time

$$t_l = \frac{x}{c_0} + \frac{2d}{v}\sqrt{1 - \left(\frac{v}{c_0}\right)^2} \tag{27}$$

can be observed, analogous to the head wave in refraction seismics.

The depth of discontinuities can be directly calculated from the corresponding reflection travel times (equation (26)) for given velocity values, which can be obtained from common midpoint (CMP) measurements (i.e. constant midpoint between transmitter and receiver antennas; *see* e.g. Reynolds, 1997). Another measurement mode is the constant-offset gather, where both antennas are moved with fixed separation. Because of the high pulse succession, quasi-continuous profiling is possible with moving antennas. Borehole surveys may include borehole-to-surface and/or borehole-to-borehole measurements (e.g. Binley *et al.*, 2002). Here, data may be collected in a zero-offset profiling mode (transmitter and receiver in different boreholes at equal depths) or in a multiple-offset gather (transmitter and receiver in different boreholes at different depths).

The general conversion of travel times into depths (migration) is equivalent to conversion procedures used in seismics. Prior to time-to-depth migration, however, the recorded signals must be corrected for possible errors caused by incorrect station geometry and zero time, geometric spreading, transmitter radiation pattern, transmitter amplitude, and high-angle raypaths (Peterson, 2001).

Hydrogeological Applications

One of the main applications of GPR is the estimation of water content in the unsaturated zone (e.g. Binley *et al.*, 2001, 2002; Alumbaugh *et al.*, 2002; Huisman *et al.*, 2002, 2003; Schmalz *et al.*, 2002; Schmalholz *et al.*, 2004) using proven petrophysical relationships. Variations in porosity and grain size likewise influence the propagation velocity of EM waves. Therefore, GPR can provide indications of changes in properties connected with textural or hydraulic parameters. Moreover, lithology information can be inferred; typically high-velocity zones correlate to sand and gravel layers, while low-velocity regions correlate to clay and silt layers. Peterson *et al.* (1999) derived 2D high-resolution images of porosity and electrical conductivity from radar tomography data at the Boise hydrogeophysical research site using the CRIM approach (equation (21)). Estimates showed a good agreement with corresponding data from neutron probe measurements.

The large contrast in dielectric permittivity between unsaturated and saturated zones also makes the groundwater table a common target in GPR surveys; it can be mapped quasi-continuously with high spatial resolution. Although structures below the groundwater table can be resolved (Figure 13), information from the saturated zone may be limited owing to the relatively high attenuation of EM waves in this region. However, GPR can be used to investigate various subsurface structures relevant to hydrogeology, such as sedimentary sequences, fractured or karstic zones, faults, and cavities (e.g. Benson, 1995).

The detection of subsurface contamination has also been objective of numerous GPR studies (e.g. Brewster and Annan, 1994; Greenhouse *et al.*, 1993). Contaminants have been found to be indicated by zones where radar reflections are weak or absent (Davis and Annan, 1989) or where additional reflections occur (Francisca and Rinaldi, 2001). Sauck *et al.* (1998) noted that owing to biogeochemical processes the electrical conductivity of hydrocarbon spills changes with time from resistive to conductive; these changes can be mapped with GPR.

Using GPR in a time-lapse manner, dynamic processes such as infiltration of surface water (*see* **Chapter 66, Soil Water Flow at Different Spatial Scales, Volume 2** and **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**) or spreading of a contaminant plume can be studied by mapping the associated changes in dielectric permittivity over time. Recent developments aim at combining GPR with other geophysical methods such as electrical resistivity tomography (Binley *et al.*, 2002).

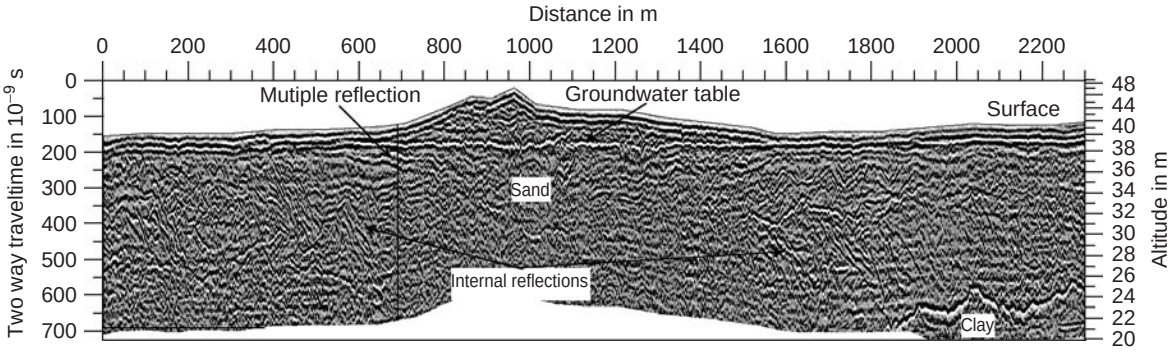


Figure 13 GPR profile measured over aqueoglacial sediments, revealing groundwater table, aquiclude, and reflections from the saturated zone (Reproduced from Bahloul, 2000, by permission of Shaker Verlag)

COMBINED HYDROGEOPHYSICAL APPROACHES

Combined hydrogeophysical approaches aim at the improved characterization of subsurface structures or processes by using geophysical methods in conjunction with hydrogeological and/or hydrological data. These data are usually obtained from well bore measurements (*see* **Chapter 151, Hydraulics of Wells and Well Testing, Volume 4**) and therefore carry only limited spatial information. Most of the combined hydrogeophysical approaches require relationships between hydrogeological and geophysical properties. In this part, we will focus on aquifer systems although considerable work on combining geophysical with hydrogeological data is presently available in the vadose zone literature (e.g. Binley *et al.*, 2001; Alumbaugh *et al.*, 2002; Huisman *et al.*, 2002).

Earlier work in hydrogeophysics has focused on synthetic case studies showing the potential of using geophysical data for aquifer characterization. Rubin *et al.* (1992) for example presented a method to identify the permeability distribution in near-surface aquifers based on a hydrogeophysical inversion technique. They incorporated, in addition to sparsely sampled pressure and hydraulic conductivity data, densely sampled seismic data, derived from a reflection or tomography survey in combination with empirical relationships between seismic and hydraulic properties. The feasibility of their procedure is illustrated using a few synthetic case studies. Copty *et al.* (1993) showed that seismic information improved the hydraulic conductivity estimates even in cases where a random error was added to the seismic velocities. They proposed a Bayesian updating of the hydraulic conductivity field using seismic velocity-hydraulic conductivity-pressure relationships, thereby accounting for estimation uncertainty. Based on these findings, further studies were conducted on the joint use of geophysical and well bore data to characterize hydraulic properties of aquifers. Copty and Rubin (1995) used a stochastic procedure to incorporate seismic surface reflection data and well data into the identification of the log hydraulic conductivity field and the spatial distribution of lithofacies. This procedure was tested on a synthetic case made up of a geological section consisting of a series of sandy aquifers separated by clay layers. Hyndman *et al.* (1994) used a combination of seismic and tracer data to estimate spatial patterns of aquifer properties such as hydraulic conductivity and dispersivity (*see* **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**) for two synthetic aquifers with the same seismic profile but with different hydraulic conductivity profiles in a hypothetical lithological cross section with two different images. Their proposed split inversion method (SIM) does not require knowledge on the relationship between seismic velocity and hydraulic conductivity but it assumes the existence of some relationship for large-scale lithological zones. This method was then used to estimate the lithological zonation of the Kesterson aquifer (San Joaquin Valley, California), the hydraulic conductivity of each zone, and its dispersivity (Hyndman and Gorelick, 1996). Hyndman and Gorelick (1996) showed that combining seismic and tracer data has the potential to provide high-resolution estimates of aquifer zonation and hydraulic properties. Hubbard *et al.* (1999) performed numerical analyses of synthetic case studies where the scale of the geophysical measurements (tomographic radar and seismics) was varied relative to the scale of hydraulic conductivity. Site-specific petrophysical models were used to relate hydraulic conductivity to either seismic or radar velocities. The study suggested that collection of a few tomographic profiles and interpretation of these profiles together with limited well bore data can provide information on the correlation structure of hydraulic conductivity.

In the last years, a few studies appeared that deal with real case studies. McKenna and Poeter (1995) used geological, geophysical, and hydrological data to identify hydrofacies and to define their spatial distribution for a field site located in Golden, Colorado. Hydrofacies are defined as homogeneous possibly anisotropic geological units being

hydrogeologically meaningful. They allow the construction of subsurface models with sharp interfaces in hydraulic conductivities. McKenna and Poeter (1995) used multiple-indicator, stochastic simulation conditioned on hard and soft information to produce realizations of hydrofacies geometries. Hard data are those data obtained in the vertical direction from wells whereas soft data are available between wells (e.g. geophysical data derived from surface or cross-borehole surveys). Each of these geometries then serves as zonation pattern for inverse flow modeling in order to discard implausible realizations and to characterize each zonation by a specific hydraulic conductivity value. Hyndman and Gorelick (1996) applied SIM to estimate the effective hydraulic conductivity of the observed lithologies at the Kesterson aquifer and the dispersivity value for the entire domain, as already outlined above. In their approach, they combined seismic, hydraulic, and tracer data using SIM. For the Kesterson aquifer, SIM adjusts in total six parameters (two seismic velocity parameters identifying three lithological classes with specific velocity, three hydraulic conductivities, and a regional value for the dispersivity) to best fit tracer data. A first step in the approach was to estimate seismic slowness (inverse seismic velocity) from available seismic cross sections obtained from well bore measurements. From these data, vertical and horizontal variograms for the region were derived. Sequential Gaussian simulation was used to develop multiple 3D conditional seismic slowness realizations. The SIM was then used to identify three lithological classes within each realization, each with a specific slowness value. The effective hydraulic conductivity was then estimated for each lithological zone such that observed tracer concentrations were optimally matched, and a value of dispersivity for the whole region was derived. The obtained results demonstrate that combining seismic and tracer data has the potential to derive structural information and hydraulic properties of aquifers.

Another interesting study was presented by Chen *et al.* (2001), who explored the possibility of using ground-penetrating radar and seismic tomography to estimate the hydraulic conductivity at the South Oyster site using a Bayesian approach. Their approach uses a-priori information on the hydraulic conductivity measured at selected wells using flowmeter technique. Geophysical data were only used at those wells where hydraulic conductivity was available. Correlation analysis showed positive correlations between GPR velocity and natural log conductivity of 0.68 and between seismic velocity and natural conductivity of 0.67. GPR attenuation and log conductivity appeared to be uncorrelated. To estimate log conductivity, a stochastic framework in which log conductivity, GPR velocity, GPR attenuation, and seismic velocity are considered as spatial random functions was adopted. Normal linear regression models are used to update conditional probability distribution functions (pdf) of seismic velocity, attenuation, and GPR velocity needed in the Bayesian likelihood approach using collocated hydrological (in this case hydraulic conductivity) and geophysical data. Chen *et al.* (2001) showed that especially velocity data hold potential in improving estimation of hydraulic conductivity even in cases where both hydraulic conductivity and geophysical tomography data vary in narrow ranges. Important for this approach is, however, the presence of correlations between the various properties. Hubbard *et al.* (2001) used a suite of methods such as surface GPR and seismic crosshole tomography, cone penetrometer, and borehole flowmeter to interpret subregional and local stratigraphy, to provide high-resolution hydraulic conductivity estimates and log conductivity spatial correlation functions. They derived horizontal and vertical intregal scales for hydraulic properties using a Bayesian approach (*see* e.g. Chen *et al.*, 2001) including tomographic data (e.g. seismic and radar velocity) and a-priori knowledge of the hydraulic conductivity pdf. Using first-order linear stochastic theory, the longitudinal dispersion coefficient of an inert solute plume in an aquifer with properties derived from the Oyster site was calculated. From this value, the longitudinal plume length was estimated and compared with visual evaluation of the plume length of a bromide plume observed at the Oyster site. Both length scales were found to be in reasonable agreement.

Gloaguen *et al.* (2001) estimated the water content of an unconfined sandy aquifer underlain by a 20 m thick clay layer using GPR. The 2D distribution of the saturated/unsaturated thickness was derived from a combination of GPR reflection, piezometric, and stratigraphic data using co-kriging. Once the piezometric levels and the clay layer depth were determined, travel times were used to compute the velocity field, which was then transferred to water content using the CRIM relation (equation (21)). From observed GPR attenuations and electrical conductivities observed in piezometer wells, porosity was determined using Archie's law (equation (9)).

OUTLOOK

Noninvasive characterization of aquifer properties and hydrological processes using geophysical methods will become more and more important in the next years. Geophysical methods in combination with hydrogeological and hydrological models and concepts may help to overcome some of the unresolved problems in subsurface research such as, for example, upscaling issues and characterization of space-time structures of subsurface properties and processes. This requires a better integration of hydrological, hydrogeological, and geophysical knowledge. In this article, research activities in the area of combined hydrogeophysical approaches were outlined, which represent a first step in tackling these unresolved problems. Further steps involve the development of data fusion techniques to

effectively integrate available information from the various sources, but also the development of novel technologies and methodologies with improved imaging and characterization capabilities. More recent methodological advances include, for instance, the surface nuclear magnetic resonance (SNMR) method – enabling direct access to subsurface water – and the magneto-electrical resistivity imaging technique (MERIT) – utilizing also magnetic fields of impressed electric currents (for an overview, *see* Yaramanci *et al.*, 2005). The establishment of hydrogeological test sites providing detailed information on geophysical, hydrological, and hydrogeological quantities is essential to develop and validate novel approaches.

Acknowledgments

We are grateful to an anonymous reviewer whose comments helped to improve this article.

REFERENCES

Alumbaugh D., Chang P.Y., Paprocki L., Brainard J.R., Glass R.J. and Rautman C.A. (2002) Estimating moisture contents in the vadose zone using cross-borehole ground penetrating radar: a study of accuracy and repeatability. *Water Resources Research*, **38**, 45-1–45-12.

Archie G.E. (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. *Transactions of the American Institute of Mining and Metallurgical Engineers*, **146**, 54–62.

Bacharach R. and Nur A. (1998) High-resolution shallow-seismic experiments in sand, part I: watertable, fluid flow, and saturation. *Geophysics*, **63**, 1225–1233.

Bahloul F. (2000) *Kartierung oberflächennaher Grundwasservorkommen mit dem elektromagnetischen Reflexionsverfahren (EMR): Bestimmung der elektrischen Kenngrößen durch kombinierte Auswertung von Gleichstromgeoelektrik und elektromagnetischen Reflexionsmessungen*, Ph.D. thesis, Münster University (ISBN 3-8265-7427-3).

Baker G.S., Steeples D.W., Schmeissner C. and Spikes K.T. (2000) Ultrashallow seismic reflection monitoring of seasonal fluctuations in the water table. *Environmental & Engineering Geoscience*, **6**, 271–277.

Benson A.K. (1995) Applications of ground penetrating radar in assessing some geological hazards: examples of groundwater contamination, faults, cavities. *Journal of Applied Geophysics*, **33**, 177–193.

Bevc D. and Morrison H.F. (1991) Borehole-to-surface electrical resistivity monitoring of a salt water injection experiment. *Geophysics*, **56**, 769–777.

Binley A., Cassiani G., Middleton R. and Winship P. (2002) Vadose zone flow model parametrisation using cross-borehole radar and resistivity imaging. *Journal of Hydrology*, **267**, 147–159.

Binley A. and Kemna A. (2005) DC resistivity and induced polarisation methods. In *Hydrogeophysics*, Rubin Y. and Hubbard S. (Eds.), Springer.

Binley A., Winship P., Middleton R., Pokar M. and West J. (2001) High-resolution characterization of vadose zone dynamics using cross-borehole radar. *Water Resources Research*, **37**, 2639–2652.

Biot M.A. (1956a) Theory of elastic waves in a fluid-saturated porous solid, I: low frequency range. *Journal of the Acoustical Society of America*, **28**, 168–178.

Biot M.A. (1956b) Theory of elastic waves in a fluid-saturated porous solid, II: high frequency range. *Journal of the Acoustical Society of America*, **28**, 179–191.

Birchak J.R., Gardner C.Z.G., Hipp J.E. and Victor J.M. (1974) High dielectric constant microwave probes for sensing soil moisture. *Proceedings of the IEEE*, **62**, 93–98.

Birkelo B.A., Steeples D.W., Miller R.D. and Sophocleous M.A. (1987) Seismic reflection study of a shallow aquifer during a pumping test. *Ground Water*, **25**, 703–709.

Börner F.D. and Schön J.H. (1991) A relation between the quadrature component of electrical conductivity and the specific surface area of sedimentary rocks. *The Log Analyst*, **32**, 612–613.

Börner F.D., Schopper J.R. and Weller A. (1996) Evaluation of transport and storage properties in the soil and groundwater zone from induced polarization measurements. *Geophysical Prospecting*, **44**, 583–601.

Bradford J.H. and Sawyer D.S. (2002) Depth characterization of shallow aquifers with seismic reflection, part II: prestack depth migration and field examples. *Geophysics*, **67**, 98–109.

Brewster M.L. and Annan A.P. (1994) Ground-penetrating radar monitoring of a controlled DNAPL release: 200 MHz radar. *Geophysics*, **59**, 1211–1221.

Carman P.C. (1937) Fluid flow through a granular bed. *Transactions of the Institution of Chemical Engineers and the Chemical Engineer*, **15**, 150–166.

Chen J., Hubbard S. and Rubin Y. (2001) Estimating the hydraulic conductivity at the South Oyster site from geophysical tomographic data using Bayesian techniques based on the normal linear regression model. *Water Resources Research*, **37**, 1603–1613.

Cole K.S. and Cole R.H. (1941) Dispersion and absorption in dielectrics. *Journal of Chemical Physics*, **9**, 341–351.

Copty N. and Rubin Y. (1995) A stochastic approach to the characterization of lithofacies from surface seismic and well data. *Water Resources Research*, **31**, 1673–1686.

Copty N., Rubin Y. and Mavko G. (1993) Geophysical-hydrological identification of field permeabilities through Bayesian updating. *Water Resources Research*, **29**, 2813–2825.

Daily W.D., Ramirez A.L., LaBrecque D.J. and Nitao J. (1992) Electrical resistivity tomography of vadose water movement. *Water Resources Research*, **28**, 1429–1442.

Danielsen J.E., Auken E., Jørgensen F., Søndergaard V. and Sørensen K.I. (2003) The application of the transient electromagnetic method in hydrogeophysical surveys. *Journal of Applied Geophysics*, **53**, 181–198.

Davis J.L. and Annan A.P. (1989) Ground-penetrating radar for high-resolution mapping of soil and rock stratigraphy. *Geophysical Prospecting*, **37**, 531–551.

de Lima O.A.L. and Niwas S. (2000) Estimation of hydraulic parameters of shaly sandstone aquifers from geoelectrical measurements. *Journal of Hydrology*, **235**, 12–26.

Dietrich P. and Fechner T. (1997) *Erkundung Hydraulisch Relevanter Strukturen im Testfeld Krauthausen*, Internal Report, Forschungszentrum Jülich GmbH, IB-41329576.

Edet A.E. (1991) Application of photogeologic and electromagnetic techniques to groundwater exploration in northwestern Nigeria. *Journal of African Earth Sciences*, **11**, 321–328.

Francisca F.M. and Rinaldi V.A. (2001) The potential application of the GPR to detect organic contaminants in sand. *Proceedings of the 15th International Conference on Soil Mechanics and Geotechnical Engineering*, Vol. 1, Balkema Publishers: Istanbul, 27–31 August 2001.

Gassmann F. (1951) Über die Elastizität Poröser Medien. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, **96**, 1–23.

Gloaguen E., Chouteau M., Marcotte D. and Chapuis R. (2001) Estimation of hydraulic conductivity of an unconfined aquifer using cokriging of GPR and hydrostratigraphic data. *Journal of Applied Geophysics*, **47**, 135–152.

Greenhouse J., Brewster M., Schneider G., Redmann D., Annan P., Olhoeft G., Lucius J., Sander K. and Mazella A. (1993) Geophysics and solvents: the Borden experiment. *The Leading Edge*, **12**, 261–267.

Hagrey S.A. al and Müller C. (2000) GPR study of pore water content and salinity in sand. *Geophysical Prospecting*, **48**, 63–85.

Haskell N.A. (1953) The dispersion of surfaces waves in multilayered media. *Bulletin of the Seismological Society of America*, **43**, 17–34.

Heigold P.C., Gilkeson R.H., Cartwright K. and Reed P.C. (1979) Aquifer transmissivity from surficial electrical methods. *Ground Water*, **17**, 338–345.

Hering A., Misiek R., Gyulai A., Ormos T., Dobroka M. and Dresen L. (1995) A joint inversion algorithm to process geoelectric and surface wave data, part A: basic ideas. *Geophysical Prospecting*, **43**, 135–156.

Hohmann G.W. (1988) Numerical modeling for electromagnetic methods of geophysics. In *Electromagnetic Methods in Applied Geophysics*, *Theory, Vol. 1*, Nabighian M.N. (Ed.), Society of Exploration Geophysicists: pp. 313–363.

Hubbard S., Chen J., Peterson J., Majer E., Williams K., Swift D., Mailloux B. and Rubin Y. (2001) Hydrogeological characterization of the South Oyster bacterial transport site using geophysical data. *Water Resources Research*, **37**, 2431–2456.

Hubbard S., Rubin Y. and Majer E. (1999) Spatial correlation structure estimation using geophysical and hydrogeological data. *Water Resources Research*, **35**, 1809–1825.

Huisman J.A., Hubbard S.S., Redman J.D. and Annan A.P. (2003) Measuring soil water content with ground penetrating radar: a review. *Vadose Zone Journal*, **2**, 476–491.

Huisman J., Snepvangers J., Bounten W. and Heuvelink G. (2002) Mapping spatial variation in surface soil water content: comparison of ground-penetrating radar and time domain reflectometry. *Journal of Hydrology*, **269**, 194–207.

Huntley D. (1986) Relations between permeability and electrical resistivity in granular aquifers. *Ground Water*, **24**, 466–474.

Hyndman D.W. and Gorelick S.G. (1996) Estimating lithologic and transport properties in three dimensions using seismic and tracer data: the Kesterson aquifer. *Water Resources Research*, **32**, 2659–2670.

Hyndman D.W., Harris J.M. and Gorelick S.M. (1994) Coupled seismic and tracer test inversion for aquifer property characterization. *Water Resources Research*, **30**, 1965–1977.

Johnson D.L., Koplik J. and Schwartz L.M. (1986) New pore-size parameter characterizing transport in porous media. *Physical Review Letters*, **57**, 2564–2567.

Jol H.M. (1995) Ground penetrating radar antennae frequencies and transmitter powers compared for penetration depth, resolution and reflection continuity. *Geophysical Prospecting*, **43**, 693–709.

Kemna A. (1996) *Tomographische Inversion des spezifischen Widerstandes in der Geoelektrik*, 3. DGG-Seminar Umweltgeophysik, Neustadt/Weinstraße, 30 August – 2 September 1994, Deutsche Geophysical Gesellschaft, Sonderband III, pp. 1–17. (ISSN 0947–1944).

Kemna A., Binley A., Ramirez A. and Daily W. (2000) Complex resistivity tomography for environmental applications. *Chemical Engineering Journal*, **77**, 11–18.

Kemna A., Binley A. and Slater L. (2004) Cross-borehole IP imaging for engineering and environmental applications. *Geophysics*, **69**, 97–107.

Kemna A., Vanderborght J., Kulessa B. and Vereecken H. (2002) Imaging and characterization of subsurface solute transport using electrical resistivity tomography (ERT) and equivalent transport models. *Journal of Hydrology*, **267**, 125–146.

Knoll M.D. (1996) *A Petrophysical Basis for Ground Penetrating Radar and Very Early Time Electromagnetics: Electrical Properties of Sand-Clay Mixtures*, Ph.D. thesis, University of British Columbia, (ISBN 0-612-14774-6).

Koesoemadinata A. and McMechan G.A. (2003) Correlations between seismic parameters, EM parameters, and petrophysical/petrological properties for sandstone and carbonate at low water saturations. *Geophysics*, **68**, 870–883.

Kosinski W.K. and Kelly W.E. (1981) Geoelectric soundings for predicting aquifer properties. *Ground Water*, **19**, 163–171.

Kozeny J. (1927) Über kapillare Leitung des Wassers im Boden – Aufstieg, Versickerung und Anwendung auf die Bewässerung. *Sitzungsberichte der Akademie der Wissenschaften Wien, Mathematisch Naturwissenschaftliche Abteilung*, **136**, 271–306.

Kuster G.T. and Toksöz M.N. (1974a) Velocity and attenuation of seismic waves in two-phase media, part I: theoretical formulations. *Geophysics*, **39**, 587–606.

Kuster G.T. and Toksöz M.N. (1974b) Velocity and attenuation of seismic waves in two-phase media, part II: experimental results. *Geophysics*, **39**, 607–618.

LaBrecque D.J., Miletto M., Daily W., Ramirez A. and Owen E. (1996) The effects of noise on Occam's inversion of resistivity tomography data. *Geophysics*, **61**, 538–548.

Lesmes D.P. and Frye K.M. (2001) The influence of pore fluid chemistry on the complex conductivity and induced-polarization responses of Berea sandstone. *Journal of Geophysical Research*, **106**, 4079–4090.

Lesmes D.P. and Morgan F.D. (2001) Dielectric spectroscopy of sedimentary rocks. *Journal of Geophysical Research*, **106**, 13329–13346.

Lichtenecker K. and Rother K. (1931) Die Herleitung des logarithmischen Mischungsgesetzes aus allgemeinen Prinzipien der stationären Strömung. *Physikalische Zeitschrift*, **32**, 255–260.

Loke M.H. and Barker R.D. (1995) Least-squares deconvolution of apparent resistivity pseudosections. *Geophysics*, **60**, 1682–1690.

Marion D., Nur A., Yin H. and Han D. (1992) Compressional velocity and porosity in sand-clay mixtures. *Geophysics*, **57**, 554–563.

Marshall D.J. and Madden T.R. (1959) Induced polarization, a study of its causes. *Geophysics*, **24**, 790–816.

Mazáč O., Kelly W.E. and Landa I. (1985) A hydrogeophysical model for relations between electrical and hydraulic properties of aquifers. *Journal of Hydrology*, **79**, 1–19.

McKenna S.A. and Poeter E.P. (1995) Field example of data fusion in site characterization. *Water Resources Research*, **31**, 3229–3240.

McNeill J.D. (1980a) *Electromagnetic Terrain Conductivity Measurement at Low Induction Numbers*, Technical Note, 6, Geonics.

McNeill J.D. (1980b) *Applications of Transient Electromagnetic Techniques*, Technical Note, 7, Geonics.

Morris M., Rønning J.S. and Lile O.B. (1996) Geoelectric monitoring of a tracer injection experiment: modeling and interpretation. *European Journal of Environmental and Engineering Geophysics*, **1**, 15–34.

Nabighian M.N. (Ed.) (1991) *Electromagnetic Methods in Applied Geophysics,Application, Parts A and B. Vol. 2* Investigations in Geophysics, 3, Society of Exploration Geophysicists: Tulsa.

Nolen-Hoeksema R.C. (1993) Porosity and consolidation limits of sediments and Gassmann's elastic wave equation. *Geophysical Research Letters*, **20**, 847–850.

Oldenburg D.W. and Li Y. (1994) Inversion of induced polarization data. *Geophysics*, **59**, 1327–1341.

Oristaglio M. and Spies B. (Eds.) (1999) *Three-Dimensional Electromagnetics*, *Geophysical Developments, Vol. 7*, Society of Exploration Geophysicists: Tulsa.

Pape H., Riepe L. and Schopper J.R. (1987) Theory of self-similar network structures in sedimentary and igneous rocks and their investigation with microscopical and physical methods. *Journal of Microscopy*, **148**, 121–147.

Parasnis D.S. (1997) *Principles of Applied Geophysics*, *Fifth Edition*, Chapman & Hall: London.

Peterson J.E. Jr (2001) Pre-inversion corrections and analysis of radar tomographic data. *Journal of Environmental and Engineering Geophysics*, **6**, 1–18.

Peterson J.E. Jr Majer E.L. and Knoll M.D. (1999) Hydrogeological properties estimation using tomographic data at the Boise hydrogeophysical research site. *Proceedings of the Symposium on the Application of Geophysics to Engineering and Environmental Problems*, Environmental and Engineering Geophysical Society.

Puranen R., Säävuori H., Sahala L., Suppala I., Mäkilä M. and Lerssi J. (1999) Airborne electromagnetic mapping of surficial deposits in Finland. *First Break*, **17**, 145–154.

Purvance D.T. and Andricevic R. (2000) On the electrical-hydraulic conductivity correlation in aquifers. *Water Resources Research*, **36**, 2905–2913.

Reynolds J.M. (1997) *An Introduction to Applied and Environmental Geophysics*, John Wiley & Sons: Chichester.

Rink M. and Schopper J.R. (1974) Interface conductivity and its implication to electric logging. *Transactions of the SPWLA 15th Annual Logging Symposium*, Society of Professional Well Log Analysts: pp. 1–15.

Rubin Y. and Ezzedine S. (1997) The travel time of solutes at the Cape Cod tracer experiment: data analysis, modeling, and structural parameter inference. *Water Resources Research*, **33**, 1537–1547.

Rubin Y. and Hubbard S. (Eds.) (2005) *Hydrogeophysics*, Springer: Berlin.

Rubin Y., Mavko G. and Harris J. (1992) Mapping permeability in heterogeneous aquifers using hydrologic and seismic data. *Water Resources Research*, **28**, 1809–1816.

Sauck W.A., Atekwana E.A. and Nash M.S. (1998) Elevated conductivities associated with an LNAPL plume imaged by integrated geophysical techniques. *Journal of Environmental and Engineering Geophysics*, **2**, 203–212.

Schmalholz J., Stoffregen H., Kemna A. and Yaramanci U. (2004) Imaging of water content distribution inside a lysimeter using GPR tomography. *Vadose Zone Journal*, **3**, 1106–1115.

Schmalz B., Lennartz B. and Wachsmuth D. (2002) Analysis of soil water content variations and GPR attribute distributions. *Journal of Hydrology*, **167**, 217–226.

Schön J.H. (1996) *Physical Properties of Rocks, Fundamentals and Principles of Petrophysics*, *Handbook of geophysical exploration, Seismic Exploration, Vol. 18*, Elsevier Science.

Scott B.T. and Barker R.D. (2003) Determining pore-throat size in Permo-Triassic sandstones from low-frequency electrical spectroscopy. *Geophysical Research Letters*, **30**, 1450, doi:10.1029/2003GL016951.

Sen P.N., Scala C. and Cohen M.H. (1981) A self-similar model for sedimentary rocks with application to the dielectric constant of fused glass beads. *Geophysics*, **46**, 781–795.

Shen L.C., Savre W.C., Price J.M. and Athavale K. (1985) Dielectric properties of reservoir rocks at ultra-high frequencies. *Geophysics*, **50**, 692–704.

Slater L., Binley A., Daily W.D. and Johnson R. (2000) Cross-hole electrical imaging of a controlled saline tracer injection. *Journal of Applied Geophysics*, **44**, 85–102.

Slater L. and Lesmes D.P. (2002) Electrical-hydraulic relationships observed for unconsolidated sediments. *Water Resources Research*, **38**, 1213–1225.

Telford W.M., Geldart L.P. and Sheriff R.E. (1990) *Applied Geophysics*, *Second Edition*, Cambridge University Press: Cambridge.

Tillmann A. (2001) *Lösung der Grenzflächenproblematik bei der gemeinsamen Inversion geoelektrischer und seismischer Daten von oberflächennahen, porösen Schichten*, Ph.D. thesis, Bochum University (available at **http://www-brs.ub.ruhr-uni-bochum.de/netahtml/HSS/Diss/TillmannAxel/**).

Tillmann A. and Stöcker T. (2001) A new approach for the joint inversion of seismic and geoelectric data. Presented at the 63rd *EAGE Conference*, Amsterdam, 11–15 June 2001, European Association of Geoscientists and Engineers.

Titov K., Kemna A., Tarasov A. and Vereecken H. (2004) Induced polarization of unsaturated sands determined through time-domain measurements. *Vadose Zone Journal*, **3**, 1160–1168.

Toksöz M.N., Cheng C.H. and Timur A. (1976) Velocities of seismic waves in porous rocks. *Geophysics*, **41**, 621–645.

Topp G.C., Davis J.L. and Annan A.P. (1980) Soil water content: measurements in coaxial transmission lines. *Water Resources Research*, **16**, 574–582.

Ulrich C. and Slater L.D. (2004) Induced polarization measurements on unsaturated, unconsolidated sands. *Geophysics*, **69**, 762–771.

Vacquier V., Holmes C.R., Kintzinger P.R. and Lavergne M. (1957) Prospecting for groundwater by induced electrical polarization. *Geophysics*, **22**, 660–687.

Vanderborght J. and Vereecken H. (2001) Analyses of locally measured bromide breakthrough curves from a natural gradient tracer experiment at Krauthausen. *Journal of Contaminant Hydrology*, **48**, 23–43.

Van Overmeeren R.A. (1994) Georadar for hydrogeology. *First Break*, **12**, 401–408.

Vereecken H., Döring U., Hardelauf H., Jaekel U., Hashagen U., Neuendorf O., Schwarze H. and Seidemann R. (2000) Analysis of solute transport in a heterogeneous aquifer: the Krauthausen field experiment. *Journal of Contaminant Hydrology*, **45**, 329–358.

Vereecken H., Yaramanci U. and Kemna A. (2002) Non-invasive methods in hydrology. *Journal of Hydrology*, **267**(3–4), Special Issue, 125–297.

Vinegar H.J. and Waxman M.H. (1984) Induced polarization of shaly sands. *Geophysics*, **49**, 1267–1287.

Waxman M.H. and Smits L.J.M. (1968) Electrical conductivities in oil-bearing shaly sands. *Society of Petroleum Engineers Journal*, **8**, 107–122.

White P.A. (1994) Electrode arrays for measuring groundwater flow direction and velocity. *Geophysics*, **59**, 192–201.

Wyllie M.R., Gregory A.R. and Gardner G.H.F. (1956) Elastic wave velocities in heterogeneous and porous media. *Geophysics*, **21**, 41–70.

Wynn J. (2002) Evaluating groundwater in arid lands using airborne magnetic/EM methods: an example in the southwestern U.S. and northern Mexico. *The Leading Edge*, **21**, 62–66.

Yamamoto T., Nye T. and Kuru M. (1994) Porosity, permeability, shear strength: crosswell tomography below an iron foundry. *Geophysics*, **59**, 1530–1541.

Yaramanci U., Kemna A. and Vereecken H. (2005) Emerging technologies in hydrogeophysics. In *Hydrogeophysics*, Rubin Y. and Hubbard S. (Eds.), Springer.

Yeh T.-C.J., Liu S., Glass R.J., Baker K., Brainard J.R., Alumbaugh D. and LaBrecque D. (2002) A geostatistical based inverse model for electrical resistivity surveys and its applications to vadose zone hydrology. *Water Resources Research*, **38**, 1278, doi:10.1029/2001WR001204.

149: Hydrodynamics of Groundwater

FRITZ STAUFFER

Swiss Federal Institute of Technology, Zurich, Switzerland

The mathematical formulation of problems of flow in aquifers is usually based on a formulation of the fluid mass balance, which has to be fulfilled at any location in the domain under consideration. A knowledge of the fluid mass flux through the saturated aquifer material, or the motion equation, is required. The resulting partial differential equation can be solved for a chosen flow domain, given parameters, and initial and boundary conditions. The solution results in the steady state or transient field of the piezometric head, from which the flux and the velocity fields may be expressed.

MOTION EQUATION FOR SATURATED FLOW, DARCY'S LAW

The Experimentally Derived Law

In 1856, Henri Darcy published the results of flow experiments with filter sands ('Les fontaines publiques de la ville de Dijon'). The experimental layout is shown, in a slightly modified form, in Figure 1.

A homogeneous sand column is built in a vertical tube with cross-sectional area A. The packing is completely saturated with water. Water flows with the discharge Q through the sand packing. At measurement stations P1 and P2, with elevations z_1 and z_2, the water pressure p_1 and p_2 is measured with the help of water manometers. The corresponding pressure heads are $p_1/(\rho g)$ and $p_2/(\rho g)$, where ρ is the water density and g is the gravitational acceleration. The experimental results by Darcy yielded the following empirical relationship (*Darcy's law*):

$$q = \frac{Q}{A} = K\frac{\Delta h}{\Delta l} \tag{1}$$

The area A comprises the complete cross-sectional area including the column. The symbol q is the specific discharge, which is the (volumetric) discharge per unit area, K is the hydraulic conductivity (that depends on the permeability of the material), h is the piezometric head, which can be interpreted as sum of pressure energy and potential energy per unit weight of water, Δh is the head difference, and Δl is the distance between the two locations P1 and P2. The piezometric head h is defined as:

$$h = z + \frac{p}{\rho g} \tag{2}$$

where z is level of the considered measurement location relative to an arbitrary datum. The water density ρ is taken as constant.

Darcy's experiment is the phenomenological foundation of the motion equation of a fluid in porous media (equation 1). It can be shown that the relationship holds even for upward flow, or when the flow direction is arbitrary. To relate the flux to a finite area A is equivalent to an averaging of the microscopic velocity over many pores. In fact, A has to be large enough in order to get representative results for the porous medium. The same principle holds for the measurement of pressure using manometers, since the result is an average over the contact area between the manometer and the porous medium, which comprises many pores. The length Δl is not the real flow path of a microscopic fluid particle, but the length of the trajectory of the specific flux. Obviously, postulating Darcy's law signifies the adoption of a macroscopic approach (to be presented below).

For a confined aquifer, the concept of piezometric head (equation 2) is illustrated in Figure 2. Consider a vertical piezometer tube with opening at the lower end, at location $\mathbf{x} = (x, y, z)$, where the prevailing water pressure is $p(\mathbf{x})$. The air pressure, p_a, within the piezometer, above the liquid level, is at zero atmospheric pressure, $p_{a,at}$. Assuming hydrostatic conditions, the water level in the piezometer

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

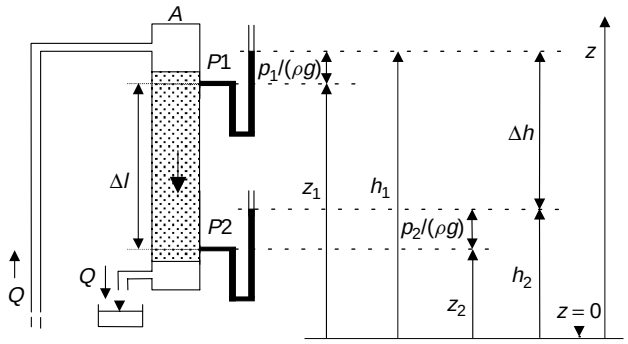


Figure 1 Schematic layout of darcy's experiment in a sand column (gray) with liquid manometers P1 and P2

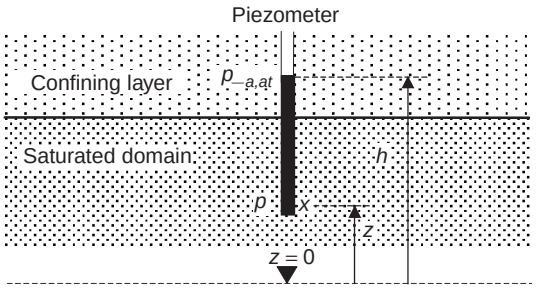


Figure 2 Concept of piezometric head h

reaches $h(\mathbf{x})$ according to equation 2, which holds for any liquid with the density ρ.

Typical values of hydraulic conductivity K (orders of magnitude) for unconsolidated sediments are listed in Table 1.

Although, as explained above, Darcy's law was proposed by Henri Darcy on the basis of experiments conducted in sand columns (thus limiting the law to one-dimensional flow of a homogeneous fluid in homogeneous porous media), it is possible to derive the motion equation from first principles of fluid mechanics. To achieve this goal, we start by considering the motion equation at a microscopic level, that is, at a point within the fluid that occupies the void space (section on Microscopic Consideration below), and then deriving the macroscopic law, that describes the flow in a porous medium domain, by averaging the microscopic motion equation over a representative elementary volume (REV, section on Macroscopic Consideration). An advantage of such an approach is that the derived law is applicable to the general case of three-dimensional flow in anisotropic inhomogeneous material of a variable density fluid.

Table 1 Typical values of hydraulic conductivity K and permeability k for various unconsolidated sediments

Sediment	K[m s^{-1}]	k[m^2]
Coarse gravel	10^{-2}	10^{-9}
Sandy gravel	10^{-3}	10^{-10}
Sand	10^{-4}	10^{-11}
Silt	10^{-6}	10^{-13}
Clay	10^{-9}	10^{-16}

Microscopic Consideration

Towards a physically based foundation and generalization of Darcy's law, a discussion may start with microscopic considerations (Figure 3). Theoretically, the fluid motion equation may be formulated for a microscopically small fluid volume within a pore. The flow problem may be formulated for the (complex) pore system with solid impermeable boundaries due to the grains. Inflow and outflow boundaries of the flow domain would be represented by prescribed head conditions.

Accordingly, a "microscopically" small fluid element in a pore is considered (Figure 3). On the basis of hydrodynamic considerations, the momentum balance equation for this fluid element is:

$$\rho\frac{\partial \mathbf{v}}{\partial t} + \rho\mathbf{v}\nabla\cdot\mathbf{v}\left(\equiv \rho\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}}t\right) = \rho\mathbf{g} - \nabla p + \nabla\cdot\boldsymbol{\tau} \qquad (3)$$

where $\mathbf{v}$ is the fluid's velocity, $\boldsymbol{\tau}$ is the viscous stress tensor. The symbol ∇p of a scalar quantity p means gradient of p; this is a vector with components $\partial p/\partial x$, $\partial p/\partial y$, $\partial p/\partial z$, and $\nabla\cdot\mathbf{v}$ of a vector quantity $\mathbf{v}$ means the divergence, $\nabla\cdot\mathbf{v} = \partial v_x/\partial x + \partial v_y/\partial y + \partial v_z/\partial z$. The terms on the left-hand side in equation 3 signify the momentum increase per unit time and unit volume. They can be interpreted as inertia forces, caused by the temporal and the convective fluid acceleration. The terms on the right-hand side are the gravity force, the pressure gradient (= force due to pressure), and the friction, all related to a unit volume.

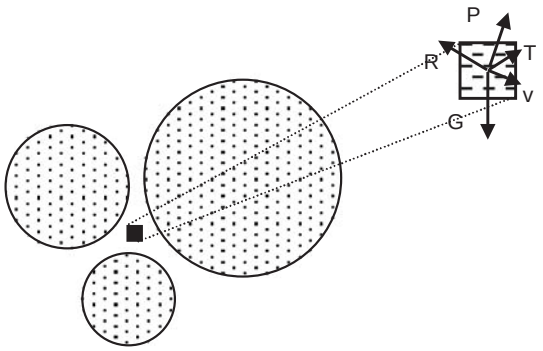


Figure 3 Microscopic approach: A small fluid element in a pore, with the acting forces per unit fluid volume inertia **T**, gravitation **G**, pressure gradient **P**, friction **R**, and the fluid velocity vector **v**

The relation between $\boldsymbol{\tau}$ and $\mathbf{v}$ is usually approximated by Newton's law:

$$\nabla \cdot \boldsymbol{\tau} = \mu \nabla^2 \mathbf{v} \tag{4}$$

where μ is the dynamic viscosity, and $\nabla^2\mathbf{v}$ of a velocity $\mathbf{v} = (v_x, v_y, v_z)$ is expressed as $\nabla^2\mathbf{v} = (\partial^2 v_x/\partial x^2 + \partial^2 v_x/\partial y^2 + \partial^2 v_x/\partial z^2, \partial^2 v_y/\partial x^2 + \partial^2 v_y/\partial y^2 + \partial^2 v_y/\partial z^2, \partial^2 v_z \partial x^2 + \partial^2 v_z/\partial y^2 + \partial^2 v_z/\partial z^2)$.

In cases with variable fluid viscosity μ, equation 4 may be valid approximately in an averaged sense. A property of the momentum equation 3 is that the variables $\mathbf{v}$, p, $\boldsymbol{\tau}$, and may be also the parameters ρ and μ are subject to more or less strong spatial variability within the pore space. The boundary condition along the solid boundaries within the complex pore system is $\mathbf{v} \cdot \mathbf{n} = 0$, where $\mathbf{n}$ denotes the unit outward normal vector normal to the solid walls, as is commonly assumed for flow near solid walls.

Macroscopic Consideration

The aim of a macroscopic consideration is to express averaged values of a microscopic quantity within a finite control volume (Figure 4), the size of which is chosen according to the requirement of obtaining stable averages. By averaging the spatial variability of the quantity within the control volume, we obtain a smoothed variation. Such a control volume is called *representative elementary volume, REV* (Bear, 1979). The average value is attributed to the center of the REV. The averaged form the microscopic momentum equation 3 is:

$$\left\langle \rho \frac{\partial \mathbf{v}}{\partial t} \right\rangle_{\mathrm{REV}} + \langle \rho \mathbf{v} \nabla \mathbf{v} \rangle_{\mathrm{REV}} = \langle \rho \mathbf{g} \rangle_{\mathrm{REV}} - \langle \nabla p \rangle_{\mathrm{REV}} + \langle \nabla \boldsymbol{\tau} \rangle_{\mathrm{REV}} \tag{5}$$

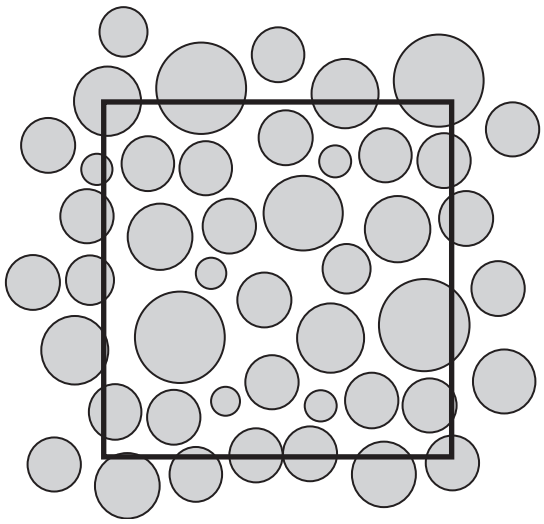


Figure 4 Concept of the representative elementary volume REV (within frame) with solid (gray) and fluid (white)

where $\langle\,\rangle_{\mathrm{REV}}$ means averaging over the REV. Equation 5 may be the starting point for the formulation of theories towards a theoretical foundation of Darcy's law, for example, after Bear and Bachmat (1991). After Matheron, (de Marsily, 1986), the linear form of Darcy's law is the consequence of the linear momentum balance provided the inertia terms are neglected.

Generalized Darcy's Law

The generally accepted result of the averaging procedure over the REV yields the generalized form of Darcy's law (Bear, 1979), which is the motion equation in a three-dimensional porous medium:

$$\mathbf{q} = -\frac{\mathbf{k}}{\boldsymbol{\mu}} \cdot [\nabla p - \rho \mathbf{g}] \tag{6}$$

The coefficient $\mathbf{k}$ is the permeability, which in general is a second order tensor. Typical values of permeability are listed in Table 1. Both fluid density ρ and dynamic viscosity μ may be space dependent. The values of p, ρ, and μ are averages over the REV.

Equation 6 allows a description of variable density flow (*see* **Chapter 157, Sea Water Intrusion Into Coastal Aquifers, Volume 4**), which may arise due to large temperature or large concentration gradients within the flow domain, where density differences may strongly affect the flow field. The latter is of special importance near coastal regions, or during the injection of water with high solute concentration into groundwater.

If the fluid density ρ is taken as constant within the considered flow domain, equation 6 can be expressed in the usual form, referred to as *Darcy's law*:

$$\mathbf{q} = -\mathbf{K} \nabla h \tag{7}$$

with the hydraulic conductivity tensor, $\mathbf{K}$, which relates to the permeability $\mathbf{k}$ as follows:

$$\mathbf{K} = \mathbf{k} \frac{g}{\nu} \tag{8}$$

and the piezometric head h (equation 2). The parameter ν is the kinematic viscosity with $\nu = \mu/\rho$. Given $\mathbf{k}$ and ν, the hydraulic conductivity $\mathbf{K}$ can be calculated, for example, for various temperatures affecting ν, using equation 8.

Applicability of Darcy's Law

Neglecting the inertia term is justified in most cases of ground water flow, due to the small groundwater velocities. A rough estimate of the ratio between the convective inertia term and the friction term within a single pore with scale d_{P} can be expressed by the Reynolds' number Re:

$$\frac{\rho \mathbf{v} \nabla \mathbf{v}}{\mu \nabla^2 \mathbf{v}} \approx \frac{v \dfrac{v}{d_{\mathrm{P}}}}{\nu \dfrac{v}{d_{\mathrm{P}}^2}} = \frac{v d_{\mathrm{P}}}{\nu} = Re \tag{9}$$

We may neglect the inertia terms compared to the friction term if *Re* is small. This means that creeping laminar flow conditions are required. The pore scale d_P might be best expressed by:

$$d_P = (k)^{1/2} \tag{10}$$

as suggested by Ward (Bear, 1979). For practical reasons, *Re* is usually taken as follows:

$$Re = \frac{qd}{\nu} \tag{11}$$

where d is a readily determined characteristic length of the pore space, for example, from sieve analysis of unconsolidated aquifer material as d_{10}, the grain size at 10% of the cumulative weight. The criterion for the applicability of Darcy's law is that *Re* is smaller than about 1 to 10 (Bear, 1979). In most ground water flow situations, this requirement is fulfilled. Exceptions may occur for flow at high gradient in aquifers, for example, near wells, or in macropore material.

Flow at higher Reynolds numbers is usually modeled by adding a correction term to the head gradient of the form aq^2 in Darcy's law according to Forchheimer (see Bear, 1979). A first order approximation of a more general motion equation was suggested by Chen *et al.* (2001) in the form:

$$-\rho J(\mathbf{q})\mathbf{q} + \mu\mathbf{q} = -\mathbf{k} \cdot [\nabla p - \rho\ \mathbf{g}] \tag{12}$$

where J is an operator. For $J(\mathbf{q}) = 0$ Darcy's law is obtained in its original form. For isotropic porous media, the first term reduces to $-\rho J(\mathbf{q}) = aq$.

Darcy's Law for Compressible Fluids

If variations of the fluid density, ρ, are only due to compressibility of the fluid, a potential after Hubbert (Bear, 1979) can be defined as follows:

$$h(\mathbf{x}) = z + \int_{p_0}^{p} \frac{\mathrm{d}p}{\rho(p)g} \tag{13}$$

with $p = p(\mathbf{x})$. This represents a generalization of the piezometric head for compressible fluids. Consider a piezometer according to Figure 2. Integrating the pressure profile $p(z)$ with fluid density $\rho(p)$, yields $h(\mathbf{x})$. Accordingly:

$$\nabla h = \frac{\partial h}{\partial p}\nabla p;\ \text{or}: \nabla p = \rho g \nabla h \tag{14}$$

and Darcy's law for compressible fluids becomes:

$$\mathbf{q} = -\frac{\mathbf{k}}{\boldsymbol{\mu}}\nabla h \tag{15}$$

Therefore, the relationship formally corresponds to equation 7.

The tensorial property of hydraulic conductivity, **K**, and permeability, **k**, is due to anisotropy effects. Anisotropy can arise, for example, due to stratification of heterogeneous aquifer.

Hydraulic Conductivity Tensor

Why are the tensors **K** and **k** symmetrical? This is shown for the example of **K**. Consider flow through the circular unit area with origin O (Figure 5) and normal unit vector $\mathbf{N} = (N_x, N_y, N_z)$. A hydraulic gradient $\mathbf{i} = (i_x, i_y, i_z)$ is present. Discharge is considered through the half-spherical area S, again with origin O, with an arbitrary surface element dS on S with normal unit vector $\mathbf{n} = (n_x, n_y, n_z)$. The hydraulic gradient in direction of **n** is $I(\mathbf{n}) = \mathbf{i} \cdot \mathbf{n} = n_x i_x + n_y i_y + n_z i_z$. Hydraulic conductivity in the direction of **n** is $K(\mathbf{n})$. The specific discharge d**q** through the surface element $dS(\mathbf{n})$ becomes:

$$\mathrm{d}\mathbf{q}(\mathbf{n}) = K(\mathbf{n})I(\mathbf{n})\mathbf{n}\,\mathrm{d}S(\mathbf{n}) \tag{16}$$

The component in direction of **N** is:

$$\mathrm{d}q(\mathbf{N}, \mathbf{n}) = K(\mathbf{n})I(\mathbf{n})\mathbf{n} \cdot \mathbf{N}\,\mathrm{d}S(\mathbf{n}) \tag{17}$$

The total discharge $q(\mathbf{N})$ through the half-spherical surface S is:

$$q(\mathbf{N}) = \int_S K(\mathbf{n})I(\mathbf{n})\mathbf{n} \cdot \mathbf{N}\,\mathrm{d}S \tag{18}$$

Choosing for **N** and **n,** the Cartesian directions x, y, and z and $i = 1$, the components of **K** are according to Table 2.

The integrals in Table 2 have to be evaluated over the surface area S. The expressions represent the discharge per unit area at a hydraulic gradient of one, and, therefore, the components of the tensor **K**. As shown in Table 2, the tensor is symmetric, provided $K(\mathbf{N}) = K(-\mathbf{N})$. This physical property of central symmetry is certainly fulfilled for porous media.

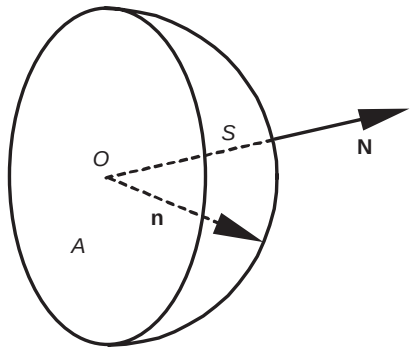


Figure 5 Circular unit area *A* with normal unit vector **N**, and half-spherical area *S* for direction-dependent hydraulic conductivity *K*

Table 2 Components of tensor **K**

	$\mathbf{i} = (1, 0, 0)$	$\mathbf{i} = (0, 1, 0)$	$\mathbf{i} = (0, 0, 1)$
$\mathbf{N} = (1, 0, 0)$	$K_{xx} = \int K(\mathbf{n}) n_x n_x \, dS(\mathbf{n})$	$K_{xy} = \int K(\mathbf{n}) n_x n_y \, dS(\mathbf{n})$	$K_{xz} = \int K(\mathbf{n}) n_x n_z \, dS(\mathbf{n})$
$\mathbf{N} = (0, 1, 0)$	$K_{yx} = \int K(\mathbf{n}) n_y n_x \, dS(\mathbf{n})$	$K_{yy} = \int K(\mathbf{n}) n_y n_y \, dS(\mathbf{n})$	$K_{yz} = \int K(\mathbf{n}) n_y n_z \, dS(\mathbf{n})$
$\mathbf{N} = (0, 0, 1)$	$K_{zx} = \int K(\mathbf{n}) n_z n_x \, dS(\mathbf{n})$	$K_{zy} = \int K(\mathbf{n}) n_z n_y \, dS(\mathbf{n})$	$K_{zz} = \int K(\mathbf{n}) n_z n_z \, dS(\mathbf{n})$

A symmetric tensor can be expressed by a symmetrical matrix, as for **K**:

$$\mathbf{K} = \begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix}; \qquad \text{with } K_{ij} = K_{ji} \tag{19}$$

According to the theorem for principal axes, a symmetric matrix can be transformed into a diagonal matrix $\mathbf{K}'$ by an orthogonal transformation, that is, by rotation of the coordinate system to the so called critical or principal axes x', y' and z':

$$\mathbf{K}' = \begin{bmatrix} K_{x'} & 0 & 0 \\ 0 & K_{y'} & 0 \\ 0 & 0 & K_{z'} \end{bmatrix} \tag{20}$$

The components $K_{m'}$ are the principal values of the tensor. The problem is identical to express the eigenvalues of the matrix requiring:

$$(\mathbf{K} - K'\mathbf{1}) \cdot \mathbf{e} = 0 \tag{21}$$

where **1** is the unit matrix, **e** is the unit eigenvector. The homogeneous Equation system 21 can be fulfilled if the determinant $det(\mathbf{K} - K'\mathbf{1})$ vanishes. The requirement leads to a characteristic polynomial of third degree with zero values equal to the eigenvalues $K_{x'}$, $K_{y'}$, and $K_{z'}$. These eigenvalues need to be positive for physical reasons. Furthermore, the sum of the diagonal components is constant in both coordinate systems:

$$K_{xx} + K_{yy} + K_{zz} = K_{x'} + K_{y'} + K_{z'} \tag{22}$$

Given the principal values $K_{x'}$, $K_{y'}$, and $K_{z'}$ the coefficients of the tensor **K** can be calculated as follows:

$$K_{ij} = K_{x'} \cos\alpha_{ix'} \cos\alpha_{x'j} + K_{y'} \cos\alpha_{iy'} \cos\alpha_{y'j} + K_{z'} \cos\alpha_{iz'} \cos\alpha_{z'j}; \quad i, j = x, y, z \tag{23}$$

K_{ij} can be interpreted as specific flux in direction i for a hydraulic gradient of 1 in direction of j. The functions $\cos(\alpha_{mn})$ are the direction-cosines between the directions m and n.

In the two-dimensional case, the directions of the principal axes x' and y' are determined by a single angle α (Figure 6). Together with the two principal components $K_{x'}$ and $K_{y'}$, the tensor is characterized by three parameters. In the three-dimensional case, we have three K-components and three angles.

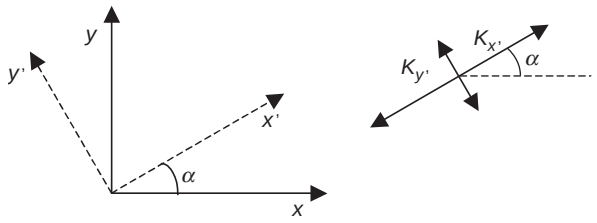


Figure 6 Coordinate system for the principal values for the hydraulic conductivity tensor $\mathbf{K}'$ in the two-dimensional case: Rotation by the angle α

Velocity

Based on the specific discharge $\mathbf{q}(\mathbf{x})$, the velocity vector $\mathbf{v}(\mathbf{x})$ can be determined by:

$$\mathbf{v} = \frac{\mathbf{q}(\mathbf{x})}{\phi(\mathbf{x})} \tag{24}$$

where ϕ is the porosity. The vector **v** is the average fluid velocity within the REV.

MASS BALANCE EQUATION FOR SATURATED FLOW

The formulation of the mass balance equation for saturated porous medium starts again with a microscopic consideration. For a microscopically small fluid volume without internal sources or sinks, it is required that (Figure 7):

$$\nabla \cdot (\rho \mathbf{v}) = -\frac{\partial \rho}{\partial t} \tag{25}$$

Basic element of the mass balance is the fluid mass flux $\mathbf{J} = \rho\mathbf{v}$ (= mass per unit area of fluid). The divergence of the mass flux equals the mass outflow minus the inflow per unit volume and unit time, and equals at the same time the change of fluid mass per unit time. Without sources and sinks, this mass change is due to fluid (and solid matrix) compressibility. Averaging equation 25 over an REV (Figure 4) yields:

$$\langle \nabla \cdot (\rho \mathbf{v}) \rangle_{\text{REV}} = \left\langle -\frac{\partial \rho}{\partial t} \right\rangle_{\text{REV}} \tag{26}$$

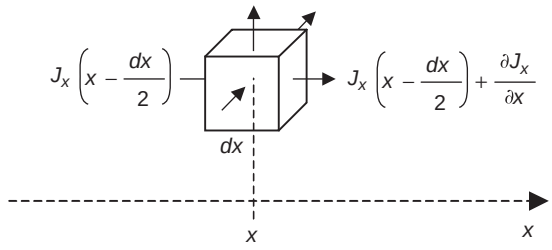


Figure 7 Microscopic volume element for the fluid mass balance with mass flux J_x in x-direction

Applying the averaging theorem (Whitaker, 1967), the balance equation is:

$$\langle \nabla \cdot (\rho \mathbf{v}) \rangle_{\mathrm{REV}} = \nabla \cdot \langle (\rho \mathbf{v}) \rangle_{\mathrm{REV}} + \frac{1}{V_0} \int_{A_{sf}} \mathbf{n} \cdot \mathbf{v} \, \mathrm{d}A \quad (27)$$

The surface area A_{sf} is the boundary between solid and fluid phase. The unit vector $\mathbf{n}$ is normal to this surface area at a particular point. Along the boundary, the velocity $\mathbf{v}$ vanishes. If the averaging is directly applied to the fluid phase f, the balance equation is:

$$\langle \nabla \cdot (\rho \mathbf{v}) \rangle_{\mathrm{REV}} = \nabla \cdot (\phi \langle \rho \mathbf{v} \rangle_{\mathrm{f}}) \quad (28)$$

The average mass flux $\langle \rho \mathbf{v} \rangle_{\mathrm{f}}$ can be written as:

$$\langle \rho \mathbf{v} \rangle_{\mathrm{f}} = \langle \rho \rangle_{\mathrm{f}} \cdot \langle \mathbf{v} \rangle_{\mathrm{f}} + \mathbf{J_d} \quad (29)$$

where $\mathbf{J}_{\mathrm{d}}$ is a dispersive mass flux caused by averaging the product of two variable quantities ρ and $\mathbf{v}$ within the REV. If the fluid density is constant, or can be considered as constant, at least within the REV, $\mathbf{J}_{\mathrm{d}}$ vanishes. The dispersive mass flux is usually neglected. The right-hand term of equation 26 can be written as:

$$\left\langle \frac{\partial \rho}{\partial t} \right\rangle_{\mathrm{REV}} = \frac{\partial}{\partial t} \langle \rho \rangle_{\mathrm{REV}} = \frac{\partial}{\partial t} (\phi \langle \rho \rangle_{\mathrm{f}}) \quad (30)$$

Equations 29 and 30 inserted in 26, with $\mathbf{q} = \phi \mathbf{v}$ and $\langle \rho \rangle_{\mathrm{f}} = \rho$ yields the macroscopic mass balance equation for a fluid of variable density in porous media:

$$\nabla \cdot (\rho \mathbf{q}) = -\frac{\partial (\phi \rho)}{\partial t} \quad (31)$$

For constant fluid density ρ the mass balance equation is:

$$\nabla \cdot \mathbf{q} = -\frac{\partial \phi}{\partial t} \quad (32)$$

and for constant density and porosity, it reduces to:

$$\nabla \cdot \mathbf{q} = 0 \quad (33)$$

If the fluid density depends on temperature T and solute concentration c, with $\rho(T, c)$, the mass change per unit time for constant porosity ϕ is:

$$\frac{\partial (\rho \phi)}{\partial t} = \phi \left(\frac{\partial \rho}{\partial T} \frac{\partial T}{\partial t} + \frac{\partial \rho}{\partial c} \frac{\partial c}{\partial t} \right) \quad (34)$$

If density variation is only due to fluid compressibility, the mass change per unit time is:

$$\frac{\partial (\phi \rho)}{\partial t} = \left(\rho \frac{\partial \phi}{\partial p} + \phi \frac{\partial \rho}{\partial p} \right) \frac{\partial p}{\partial t} \quad (35)$$

According to Hubbert's potential (equation 13) the change in pressure over time is:

$$\frac{\partial p}{\partial t} = \rho(p) g \frac{\partial h}{\partial t}$$

and therefore:

$$\frac{\partial (\phi \rho)}{\partial t} = \rho S_0 \frac{\partial h}{\partial t} \quad (36)$$

where S_0 is the called *specific storativity*, with:

$$S_0 = \left(\rho \frac{\partial \phi}{\partial p} + \phi \frac{\partial \rho}{\partial p} \right) g \quad (37)$$

Physically the specific storativity S_0 is the volume change of the fluid per unit volume of porous medium due to a unit increase or decrease in the piezometric head. It consists of two parts. The first part is due to a change in the pore volume, caused by deformation of the porous matrix. The second part is due to fluid compressibility. The latter can easily be assessed with the help of the fluid compressibility (compressibility of water: 4.810^{-10} Pa^{-1} for normal atmospheric conditions). In general, S_0 depends on fluid pressure, p. In practice, a constant value is usually applied for S_0.

The divergence of fluid mass flux can be further simplified by neglecting density differences within the REV, by assuming that at the macroscopic level, spatial changes are smaller than the temporal ones, by:

$$\nabla \cdot (\rho \mathbf{q}) = \rho \nabla \cdot \mathbf{q} \quad (38)$$

Therefore, the mass balance equation for compressible fluids and elastic matrix and an additional source/sink term, P, representing, for example, pumping, is, with Darcy's law (equation 7 inserted):

$$\nabla \cdot (\mathbf{K} \nabla h) + P = S_0 \frac{\partial h}{\partial t} \quad (39)$$

or in detail:

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}+K_{xy}\frac{\partial h}{\partial y}+K_{xz}\frac{\partial h}{\partial z}\right)+\frac{\partial}{\partial y}\left(K_{yx}\frac{\partial h}{\partial x}+K_{yy}\frac{\partial h}{\partial y}+K_{yz}\frac{\partial h}{\partial z}\right)+\frac{\partial}{\partial z}\left(K_{zx}\frac{\partial h}{\partial x}+K_{zy}\frac{\partial h}{\partial y}+K_{zz}\frac{\partial h}{\partial z}\right)+P=S_0\frac{\partial h}{\partial t} \tag{40}$$

For incompressible fluid and rigid matrix, it becomes (with $P = 0$):

$$\nabla \cdot (\mathbf{K}\nabla h) = 0 \tag{41}$$

If the porous medium is also isotropic, the balance equation is:

$$\nabla \cdot (K\nabla h) = 0 \tag{42}$$

If it is also homogeneous, we get the Laplace equation:

$$\nabla^2 h = \frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}+\frac{\partial^2 h}{\partial z^2} = 0 \tag{43}$$

The specific storativity S_0 is of the order of $S_0 = 10^{-2}\,\mathrm{m}^{-1}$ for highly compressible clay material, and $S_0 = 10^{-7}\,\mathrm{m}^{-1}$ for rock material with small porosity (de Marsily, 1986). For sandy gravel aquifer material, S_0 is near $S_0 = 10^{-5}\,\mathrm{m}^{-1}$ (Domenico and Schwartz, 1990). The specific storativity is clearly increased if the porous medium contains a residual amount of trapped air bubbles.

FORMULATION OF FLOW PROBLEMS

The formulation of a flow problem consists of:

1. specifying the flow domain D, with boundary B;
2. specifying the appropriate fluid mass balance equation (continuity equation) with the dependent variable, for example, $h(\mathbf{x},t)$;
3. specifying the coefficients, for example, $\mathbf{K}(\mathbf{x})$, $S_0(\mathbf{x})$;
4. specifying the initial condition for transient flow problems;
5. specifying the boundary conditions at the boundary of the domain D.

Altogether, this defines a complete flow problem or flow model.

Initial Condition

The initial condition consists of the state of the dependent variable within the domain D at time $t = 0$:

$$h(\mathbf{x}, t=0) = h_0(\mathbf{x}); \quad \text{or} \quad p(\mathbf{x}, t=0) = p_0(\mathbf{x}); \quad \mathbf{x} \in D \tag{44}$$

Boundary Conditions

Along the complete boundary B, the flow conditions have to be formulated (specified or prescribed). Owing to the continuum principle, since the REV is not well defined directly on the boundary, the use of a finite REV raises a conceptual problem near the boundary. As an idealization, it is generally adopted that the average over the REV can be extrapolated to the boundary in order to obtain a continuous variable.

Direct Contact with Surface Water (Boundary B_1)

If an aquifer is in direct contact with a surface water body (like river, creek, canal, lake, etc.), the pressure can be expressed for the surface water body along the boundary B_1 (Figure 8). Assuming static conditions in the surface water body, the condition for the piezometric head is:

$$h(\mathbf{x}, t) = h_{\mathrm{B}_1}(\mathbf{x}, t); \qquad \mathbf{x} \in B_1 \tag{45}$$

The parameter $h_{\mathrm{B}_1}(\mathbf{x},\mathrm{t})$ is the water level of the surface water body, which can be transient. Mathematically speaking, this is a first type boundary condition (Dirichlet boundary condition).

Prescribed Flux Through Boundary (Boundary B_2)

If a prescribed flux $q_n(\mathbf{x},t)$, with $\mathbf{x} \in \mathrm{B}_2$, through the boundary is present at boundary B_2(see Figure 9), we have:

$$q_n(\mathbf{x}, t) = q_{\mathrm{B}_2}(\mathbf{x}, t); \quad \mathbf{x} \in B_2 \tag{46}$$

This corresponds to a second type boundary condition (Neumann boundary condition).

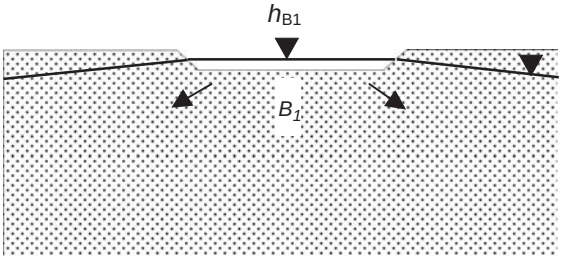


Figure 8 Boundary B_1: Direct contact of aquifer with a surface water body (schematic vertical cut)

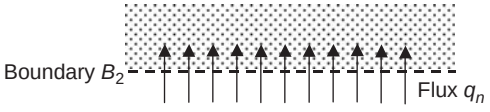


Figure 9 Boundary B_2: Prescribed flux q_n through a boundary

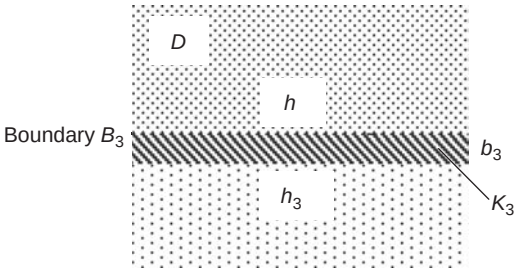


Figure 10 Semipermeable boundary B_3

A special case is the impermeable boundary, with $q_n = 0$ at boundary B_2, with the condition:

$$q_n(\mathbf{x}, t) = \mathbf{q}(\mathbf{x}) \cdot \mathbf{n} = \mathbf{K}(\mathbf{x}) \nabla h(\mathbf{x}) \cdot \mathbf{n} = 0; \qquad \mathbf{x} \in B_2 \tag{47}$$

which, for an isotropic hydraulic conductivity K, becomes:

$$\frac{\partial h}{\partial n}(\mathbf{x}) = 0; \qquad \mathbf{x} \in B_2 \tag{48}$$

Semipermeable Boundary (Boundary B_3)

Here, the flow domain is confined by a semipermeable layer of a porous medium with reduced permeability (Figure 10). Beyond this layer the hydraulic head is h_3. The hydraulic conductivity of the semipermeable layer is K_3, and the layer thickness is b_3.

According to Darcy's law, the specific flux through the boundary is:

$$q_n = K_3(\mathbf{x}) \frac{h_3(\mathbf{x}) - h(\mathbf{x})}{b_3}; \qquad \mathbf{x} \in B_3 \tag{49}$$

or:

$$(-\mathbf{K} \nabla h) \cdot \mathbf{n} = \frac{h_3(\mathbf{x}) - h(\mathbf{x})}{\sigma_3(\mathbf{x})}; \qquad \mathbf{x} \in B_3 \tag{50}$$

where $\mathbf{n}$ is a unit vector normal to the boundary. The factor $\sigma_3 = b_3/K_3$ is called *leakage coefficient*. The boundary condition (third type) is of mixed type (Cauchy boundary condition), with elements of first and second type. For isotropic hydraulic conductivity K, it becomes:

$$q_n = -K(\mathbf{x}) \frac{\partial h}{\partial n}(\mathbf{x}) = \frac{h_3(\mathbf{x}) - h(\mathbf{x})}{\sigma_3(\mathbf{x})}; \qquad \mathbf{x} \in B_3 \tag{51}$$

If the layer gets very permeable, $\sigma_3 \to 0$, and the present boundary condition reduces to one of prescribed head. For $\sigma_3 \to \infty$, the boundary is impermeable.

Groundwater Table as Boundary (Boundary B_4)

At the groundwater table, the pressure is, by definition, equal to atmospheric pressure, or relative pressure $p = 0$, or $h = z$. In the general case, the groundwater table is transient with $z_{GWT}(\mathbf{x},t)$ with $\mathbf{x} \in B_4$. Additionally, an areal recharge

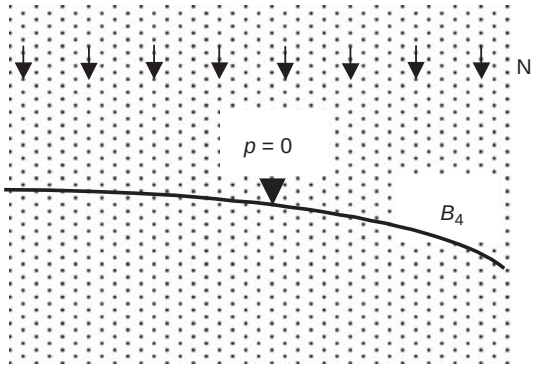


Figure 11 Groundwater table as boundary B_4 (schematic vertical section)

rate (fluid volume per unit area and unit time) $\mathbf{N}(\mathbf{x},t) = (0,0,-N)$ with $\mathbf{x} \in B_4$ may also be present (see Figure 11).

The problem consists of the fact that the boundary condition is known, but the location of the boundary is not known. Moreover, the boundary may be transient ("moving" boundary). Additional requirements are needed. The property that the relative water pressure at the groundwater table is zero, due to contact with air, can be represented by a function $F(\mathbf{x},t)$,

($F(\mathbf{x},t) = 0$, can be considered as the equation that describes the geometry of the boundary, Bear, 1979) with the condition for the water table:

$$F(x, y, z, t) = h - z = \frac{p}{\rho g} = 0; \qquad \mathbf{x} \in B_4 \tag{52}$$

The formulation of the condition $F(\mathbf{x},t) = 0$ can be accomplished by the material derivative:

$$\begin{aligned} \frac{DF}{Dt} &= \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial t} = 0 \\ &= \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x}u_x + \frac{\partial F}{\partial y}u_y + \frac{\partial F}{\partial z}u_z = 0 \\ &= \frac{\partial F}{\partial t} + \mathbf{u_F} \cdot \nabla F \end{aligned} \tag{53}$$

where $\mathbf{u_F}$ is the velocity of the groundwater table. The fluid mass balance condition at the moving groundwater table relative and normal to the surface $F(\mathbf{x},t) = 0$ requires (Figure 12):

$$\mathbf{q} \cdot \mathbf{n} - \mathbf{N} \cdot \mathbf{n} = \mathbf{u_F} \cdot \mathbf{n} \cdot \phi_e \tag{54}$$

where $\mathbf{n}$ is the unit vector normal to the surface $F = 0$, and ϕ_e is the effective (drainable) porosity. Therefore, the velocity of the groundwater table is expressed by:

$$\mathbf{u_F} = \frac{1}{\phi_e} \cdot (\mathbf{q} - \mathbf{N}) \tag{55}$$

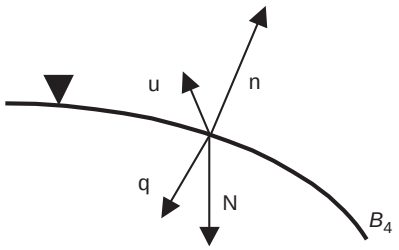


Figure 12 Specific flux vectors at the groundwater table (Boundary B_4)

Inserted in equation 53, the motion equation for the groundwater table is:

$$\frac{\partial F}{\partial t} + \frac{1}{\phi_e}[\mathbf{q} - \mathbf{N}]\nabla F = 0 \tag{56}$$

Together with equation 52 and Darcy's law for the principle axes of the tensor **K,** the condition for the 'moving' boundary B_4 gets:

$$\phi_e \frac{\partial h}{\partial t} = K_x \left(\frac{\partial h}{\partial x}\right)^2 + K_y \left(\frac{\partial h}{\partial y}\right)^2 + K_z \left(\frac{\partial h}{\partial z}\right)^2 - \frac{\partial h}{\partial z} \cdot (K_z + N) \tag{57}$$

This condition is nonlinear and has to be fulfilled simultaneously with the fluid mass balance equation.

Seepage Face Boundary (Boundary B_5)

A seepage face forms along an outlet surface (Figure 13), where water leaves the porous medium, gets in contact with the atmosphere, and seeps along the outlet surface. Assuming that the water thickness is negligible, the boundary condition at the seepage face between points D and E is:

$$p(\mathbf{x}) = 0; \quad \text{or:} \quad h(\mathbf{x}) = z; \quad \mathbf{x} \in B_5 \tag{58}$$

Contrary to the groundwater table (boundary B_4), the location of the seepage face is in principle known for a given boundary, with the exception of the upper point D, which is at the same time part of the free surface (boundary B_4).

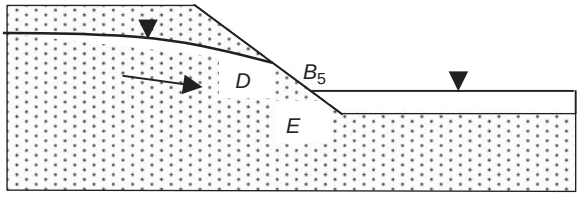


Figure 13 Seepage face along outlet surface (Boundary B_5)

FLOW IN EXTENDED THIN AQUIFERS

Flow of water in extended thin aquifers can be modeled as two-dimensional horizontal flow if the horizontal extension of the flow domain is much larger than the vertical. Usually, the following assumptions are adopted:

- The flow is horizontal. This means that vertical flux components are neglected.
- Water and porous matrix can be compressible. Therefore, the piezometric head is defined according to Hubbert's potential (equation 13).
- In the flow domain, no density effects due to temperature or concentration differences are considered.

By this formulation, a considerable simplification of the formulation of the flow problem is achieved. The flow domain not only reduces to the horizontal plane, but also the complicated motion equation 57 vanishes.

Flow in Confined Aquifers

The starting point for the formulation of the (macroscopic) mass balance equation within the aquifer is equation 39. For extended thin aquifers, the fluid mass balance is performed for a control volume (Figure 14), which extends over the complete thickness, H, of the aquifer. Aquifer bottom and top are taken as impermeable. Taking into account the fact that the aquifer may be more or less strongly heterogeneous, which leads to a spatial variability of the parameters **K**, and S_0, and the variable h. The size of the control volume has to that large, that stable average values are obtained. The integration of equation 39 over the control volume can be formulated as follows, with $P = 0$:

$$\iiint_{KV} \nabla \cdot (\mathbf{K}\nabla h)\, \mathrm{d}V = \iiint_{KV} S_0 \frac{\partial h}{\partial t}\, \mathrm{d}V \tag{59}$$

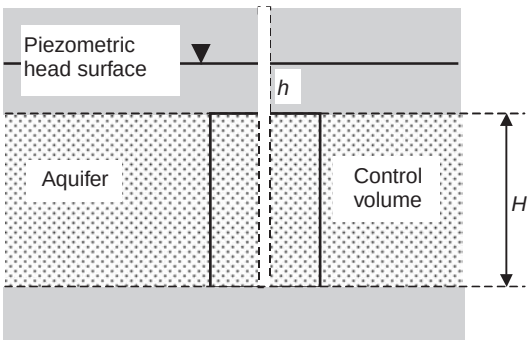


Figure 14 Control volume for an extended thin aquifer

Applying Gauss' law to the left-hand side of equation 59 leads to:

$$\iiint_{KV} \nabla \cdot (\mathbf{K}\nabla h)\,\mathrm{d}V = \iint_{A} (\mathbf{K}\nabla h)\mathbf{n}\,\mathrm{d}A = Q_{x2} - Q_{x1} + Q_{y2} - Q_{y1} \quad (60)$$

where $\mathbf{n}$ is the unit vector normal to the surface A. The surface A is the complete surface of the control volume. The terms Q_i are the horizontal fluxes through the vertical surfaces A_i. Vertical fluxes are not present. Assuming a horizontal bottom and top surface of the control volume leads to:

$$H \cdot \langle \nabla \cdot (K\nabla h) \rangle = \nabla \cdot (H\mathbf{q}) = H \cdot \left\langle S_0 \frac{\partial h}{\partial t} \right\rangle \quad (61)$$

where $\langle\rangle$ means averaging over the control volume. The term $\mathbf{q}H = Q'$ is the horizontal flux per unit width, with components in the direction $\mathbf{i}$:

$$Q'_i = H\ q_i = \frac{H}{A_i} \iiint_{A_i} (\mathbf{K}\nabla h)\mathbf{i}\,\mathrm{d}A \quad (62)$$

The horizontal flux $\mathbf{Q}'$ per unit width may be modeled by the approach in analogy to Darcy's law:

$$\mathbf{q}H = -H\mathbf{K}_{\mathbf{eq}}\nabla\langle h \rangle \quad (63)$$

where K_{eq} is the equivalent hydraulic conductivity, valid for the control volume. The expression $H\mathbf{K}_{\mathrm{eq}} = \mathbf{T}$ is called *transmissivity* $\mathbf{T}$. This coefficient is, in general, again a symmetrical second order tensor. For isotropic condition in the horizontal plane, it reduces to a scalar quantity $T = KH$. Assuming, furthermore, that the temporal volume changes within the control volume can be written as:

$$H \left\langle S_0 \frac{\partial h}{\partial t} \right\rangle = H S_0 \frac{\partial \langle h \rangle}{\partial t} = S \frac{\partial \langle h \rangle}{\partial t} \quad (64)$$

we get the flow balance equation for an extended thin aquifer, including a source/sink term, $P(x, t)$, with $h = \langle h \rangle$ of the form:

$$\nabla \cdot (\mathbf{T}\nabla h) + P = S \frac{\partial h}{\partial t} \quad (65)$$

The coefficient $S = HS_0$ is the storativity for confined aquifers. The storativity, S, is dimensionless and can be interpreted as *water volume change in the control volume per unit horizontal area of the aquifer at a unit increase of the piezometric head*. Typical values are around $S = 10^{-5}$ for sandy gravel material. If water and the porous matrix are taken as incompressible, it reduces to $S = 0$. In this case, the equation describes a pseudosteady state. In transient flows, time dependence is introduced via the boundary conditions.

For anisotropic porous media, $\mathbf{T}$ is a tensor, and the flow balance equation becomes (in detail):

$$\frac{\partial}{\partial x}\left(T_{xx}\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial x}\left(T_{xy}\frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial y}\left(T_{yx}\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(T_{yy}\frac{\partial h}{\partial y}\right) + P = S\frac{\partial h}{\partial t} \quad (66)$$

and for isotropic T:

$$\frac{\partial}{\partial x}\left(T\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(T\frac{\partial h}{\partial y}\right) + P = S\frac{\partial h}{\partial t} \quad (67)$$

and for constant isotropic T:

$$T\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right) + P = S\frac{\partial h}{\partial t} \quad (68)$$

If flow is additionally at steady state, we get:

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} + \frac{P}{T} = 0 \quad (69)$$

The flow balance equation can easily be extended to include leaky aquifers (Figure 15). A confined aquifer is connected with a lower (l) and/or an upper (u) aquifer by semipermeable layers of hydraulic conductivity K_{l} and K_{u} and of thickness b_{l} and b_{u}. The piezometric head in the lower and upper aquifers is h_{l} and h_{u}.

The vertical fluxes q_{vl} and q_{vu} can be expressed by Darcy's law and by introducing leakage coefficients, σ_{l} and σ_{u}:

$$q_{\mathrm{vl}} = K_{\mathrm{l}} \cdot \frac{h_{\mathrm{l}} - h}{b_{\mathrm{l}}} = \frac{h - h_{\mathrm{l}}}{\sigma_{\mathrm{l}}}$$
$$q_{\mathrm{vu}} = K_{\mathrm{u}} \cdot \frac{h_{\mathrm{u}} - h}{b_{\mathrm{u}}} = \frac{h - h_{\mathrm{u}}}{\sigma_{\mathrm{u}}} \quad (70)$$

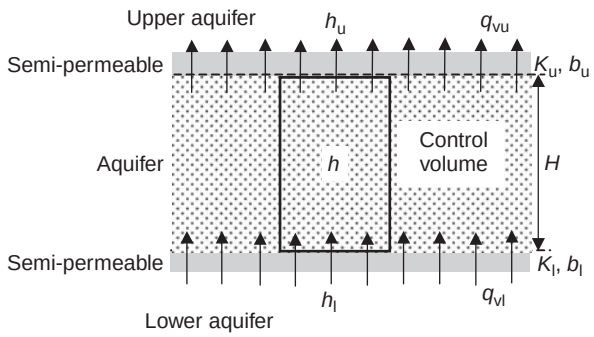


Figure 15 Leaky confined aquifer: control volume

Therefore the flow balance equation for a leaky confined aquifer is:

$$\nabla \cdot (\mathbf{T}\nabla h) + \frac{h_\mathrm{l} - h}{\sigma_\mathrm{l}} + \frac{h_\mathrm{u} - h}{\sigma_\mathrm{u}} = S\frac{\partial h}{\partial t} \qquad (71)$$

Flow in Unconfined Aquifers

In unconfined aquifers (Figure 16), the upper limit of the flow domain is given by the groundwater table, whose location may be transient. The bottom of the aquifer is at the level $z_\mathrm{l}(\mathbf{x})$, which does not need to be horizontal. The flow balance equation is formulated for a control volume with its center at location $\mathbf{x}$; it extends from the bottom to the groundwater table. Therefore, the control volume is transient. According to Dupuit's assumption, the elevation of the groundwater table can be approximated by the average piezometric head $h(\mathbf{x})$ within the control volume.

Contrary to the confined case, an additional volume change occurs within the control volume due to the vertical movement of the groundwater table and due to a possible areal replenishment rate N (Volume per unit horizontal area and unit time). A rise in the groundwater table elevation, h_{GWT}, causes an increase in the water volume by ΔV, and *vice versa*. This can be stated, per unit horizontal area and unit time, as follows:

$$\phi_\mathrm{e}\frac{\partial h_{GWT}}{\partial t} \approx \phi_\mathrm{e}\frac{\partial h}{\partial t} \qquad (72)$$

where ϕ_e is the effective (drainable) porosity, which is the part of the total pore volume, which can be filled or drained. Thus the flow balance equation for unconfined aquifers becomes, including a source/sink term P:

$$\nabla \cdot (\mathbf{T}\nabla h) + N + P = S\frac{\partial h}{\partial t} \qquad (73)$$

where S is the storativity of the unconfined aquifer. It is composed of a portion due to compressibility of water and porous matrix, and the effective porosity, which is in fact dominant. The storativity, S, for unconfined aquifer is the volume change of water per unit horizontal area and unit rise of the groundwater table. Typical values are around $S = 0.1$ to 0.2 for sandy gravel aquifers. Obviously the values are larger than those for confined conditions.

The transmissivity depend on the thickness $H(x, t)$, which, for unconfined aquifers, is:

$$H(\mathbf{x}, t) = h(\mathbf{x}, t) - z_\mathrm{l}(\mathbf{x}) \qquad (74)$$

Therefore, the transmissivity $T(\mathbf{x}, h(\mathbf{x}, t), t)$ depends on the level $h(\mathbf{x}, t)$ of the groundwater table, and the flow balance equation becomes nonlinear. If the transient changes in the piezometric head, $h(\mathbf{x}, t)$, can be neglected, compared to the thickness, H, the balance equation becomes linearized.

A leaky unconfined aquifer can be considered in a fashion similar to that of the confined case. The flow balance equation for a lower (l) leaky aquifer with leakage coefficient σ_l becomes:

$$\nabla \cdot (\mathbf{T}\nabla h) + N + \frac{h_\mathrm{l} - h}{\sigma_\mathrm{l}} = S\frac{\partial h}{\partial t} \qquad (75)$$

A frequent application of this concept is the consideration of the interaction between surface water and groundwater (Figure 17). Let an unconfined aquifer be directly connected with a surface water body. Clogging of the bottom of the surface water body (river, creek, lake, etc.) by fine sediments may cause the establishment of a semipermeable layer. This relates to the part of the aquifer beneath a (rather wide) river and an underlying (relatively) thin aquifer, as Figure 17 shows portions of the phreatic aquifer with no direct contact with surface water. Moreover, for this part, there is no replenishment rate N where there is a semipermeable layer. Such a situation can be treated using equation 75 by replacing h_l and σ_l by h_SW and σ_SW., and by setting $N = 0$. Even without a clogging layer, a leakage coefficient may be present. This may be due to the vertical flow components, which lead to a level difference between h (average head in control volume below surface water body) and h_SW (surface water level). Moreover it can happen that the leakage coefficient depends on the surface water level, $h_\mathrm{SW}(t)$. This may be the case when the lateral banks of the surface water body are more permeable than the bottom, thus leading to higher infiltration or exfiltration rates for higher surface water levels, h_SW.

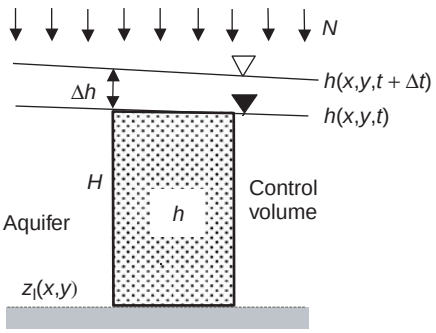


Figure 16 Unconfined aquifer: Control volume

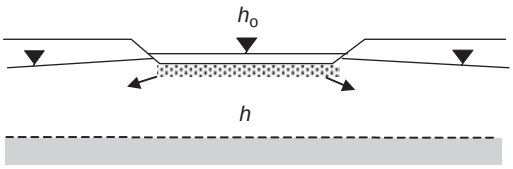


Figure 17 Direct infiltration from surface water to groundwater

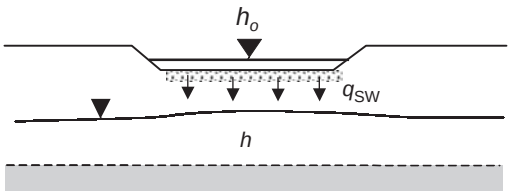


Figure 18 Groundwater table below the bottom of a surface water body

If the hydraulic conductivity, K_{SW}, of the semipermeable layer of the surface water decreases, the leakage coefficient of this layer increases and the groundwater table may, finally, fall below the bottom of the surface water body (Figure 18). An unsaturated domain establishes itself between surface water and the groundwater table. The flux through the bottom of the surface water body is now determined by h_{SW}, K_{SW} and the water pressure below the semipermeable layer. Such a situation is best modeled by replacing N in equation 75 by an areal recharge rate $q_{SW}(\mathbf{x}, t)$ within the surface water area.

Initial and Boundary Conditions

The formulation of the flow problem in extended thin aquifers, again, needs the definition of the flow domain D, of the flow balance equation for the dependent variable $h(\mathbf{x}, t)$, of the coefficients (e.g. T, S, $N(\mathbf{x}, t)$), of the initial condition $h(\mathbf{x}, t) = 0 \in D$, and of the boundary conditions. The coefficients $T(\mathbf{x})$, $S(\mathbf{x})$, and $N(\mathbf{x}, t)$ have to be defined within the complete flow domain D. Additionally, the levels of the aquifer bottom, $z_l(\mathbf{x})$, and top $z_u(\mathbf{x})$ have to be specified.

Direct Contact with Surface Water (Boundary B_1)

The aquifer is in direct contact with a surface water body along the boundary B_1. This leads to the boundary condition (first type) similar to equation 45:

$$h(\mathbf{x}, t) = h_{B_1}(\mathbf{x}, t); \qquad \mathbf{x} \in B_1 \tag{76}$$

Prescribed Flux Q' Through Boundary (Boundary B_2)

The prescribed flux $\mathbf{Q}'$ per unit width (second type boundary condition) is similar to equation 46:

$$\mathbf{Q}'(\mathbf{x}, t) \cdot \mathbf{n}(x, t) \equiv Q_n(\mathbf{x}, t) = -\mathbf{T} \cdot \nabla h \cdot \mathbf{n} = Q_{n,B_2}(\mathbf{x}, t); \quad \mathbf{x} \in B_2 \tag{77}$$

where $\mathbf{n}$ is the unit vector normal to the boundary. For an impermeable boundary, we get for isotropic T:

$$Q_n(\mathbf{x}, t) = -T \frac{\partial h}{\partial n} = 0; \text{ or : } \quad \frac{\partial h}{\partial n} = 0; \quad \mathbf{x} \in B_2 \tag{78}$$

Semipermeable Boundary (Boundary B_3)

If a surface water body represents a boundary, and is connected via a semipermeable layer with an aquifer, it can be idealized for a vertically integrated flow according to Figure 19:

$$Q_n(\mathbf{x}, t) = -\mathbf{T} \cdot \nabla h \cdot \mathbf{n} = \frac{h_3(\mathbf{x}, t) - h(\mathbf{x}, t)}{\sigma_3(\mathbf{x}, t)}; \qquad \mathbf{x} \in B_3 \tag{79}$$

where σ_3 is again a leakage coefficient.

Simple Example for Flow Problem Formulation

Consider a flow domain in an unconfined isotropic aquifer, according to Figure 20. The flow balance equation is

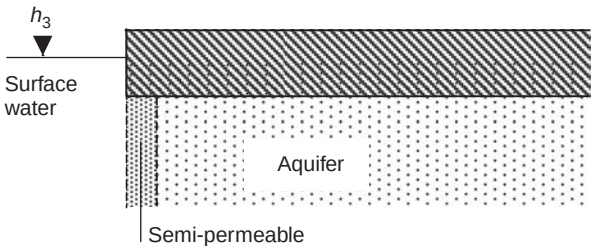


Figure 19 Semipermeable boundary B_3: Idealization

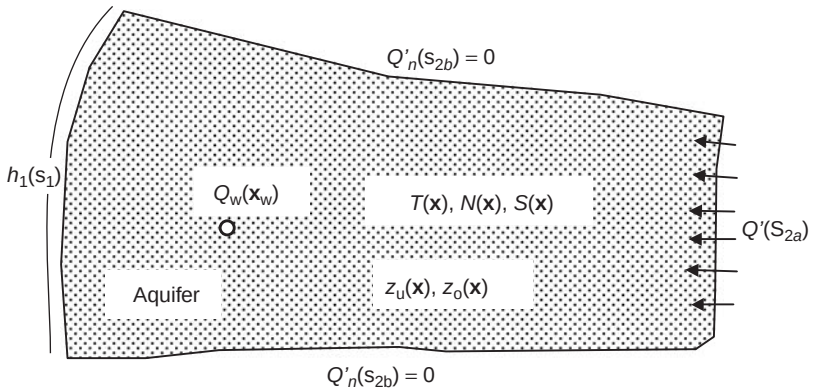


Figure 20 Example of an extended unconfined thin aquifer with a pumping well

given by equation 67. The transmissivity T is calculated with the help of the equivalent hydraulic conductivity, $K_{eq}(\mathbf{x})$, and the thickness, $H(\mathbf{x}, t) = h(\mathbf{x}, t) - z_l(\mathbf{x})$. Therefore, the aquifer bottom levels, $z_l(\mathbf{x})$, have to be provided as well. The values for $K_{eq}(\mathbf{x})$ may be determined by pumping tests in boreholes or wells. Either interpolated, distributed, or geologically motivated zonal values for K_{eq} can be provided. The replenishment rate is $N(\mathbf{x}, t)$. This rate is essentially the precipitation rate minus the evapotranspiration rate and the surface runoff and can be estimated by using hydrological methods. The boundary conditions along the boundaries are as follows. The aquifer is in direct contact with a river with a water level $h_1(\mathbf{s}_1)$. This defines a first type boundary condition along the river. At the opposite boundary, a prescribed distributed flux, $Q'(\mathbf{s}_{2a})$, is given. The lateral boundaries are impermeable, with $\partial h/\partial n(\mathbf{s}_{2b}) = 0$. The effect of a pumping well with pumping rate Q_w at location $\mathbf{x}_w$ is considered in the source/sink term P. The initial condition is $h(\mathbf{x}, t = 0) = h_0(\mathbf{x})$, which may be based on interpolated head measurements at a given time.

REFERENCES

Bear J. (1979) *Hydraulics of Groundwater*, McGraw-Hill: New York.

Bear J. and Bachmat Y. (1991) *Introduction to Modeling of Transport Phenomena in Porous Media*, Kluwer Academic Publishers: Dordrecht.

Chen Z., Lyons S.L. and Qin G. (2001) Derivation of the Forchheimer law via homogenization. *Transport in Porous Media*, **44**, 325–335.

de Marsily G. (1986) *Quantitative Hydrogeology*, Academic Press: Orlando.

Domenico P.A. and Schwartz F.W. (1990) *Physical and Chemical Hydrogeology*, Wiley: New York.

Whitaker S. (1967) Diffusion and dispersion in porous media. *American Institute of Chemical Engineers Journal*, **13**, 1066–1071.

150: Unsaturated Zone Flow Processes

JOHN R NIMMO

United States Geological Survey, Menlo Park, CA, US

Water flow in the unsaturated zone is greatly influenced by unsaturated hydrostatics (water content, energy, pressure, and retention) and by unsaturated hydrodynamics (diffuse flow and preferential flow). Important multiphase processes include the transport of gases, nonaqueous liquids, and solid particles. Numerous means are available for determination of unsaturated conditions and properties, both measurement (of moisture state, water retention, and dynamic characteristics) and through various formulas and models that are mostly empirical in nature, but in some cases incorporating insight into unsaturated-zone physical processes. Applications to practical problems include models and techniques relating to distributions of water and energy, fluxes at the land surface, inputs, outputs, and fluxes within the unsaturated zone, all of which are frequently complicated by heterogeneity and preferential flow. Further scientific advance requires new measurement techniques and theoretical constructs that more adequately represent the important physical processes within practical modeling schemes.

INTRODUCTION

The unsaturated zone, sometimes called the *vadose zone* or *zone of aeration*, plays several critical hydrologic roles. As a storage medium, it is a zone in which water is immediately available to the biosphere. As a buffer zone between the land surface and aquifers below, the unsaturated zone is a controlling agent in the transmission of contaminants and aquifer-recharging water. As an accessible body of material in which physical and chemical processes may be relatively slow, it is a place where wastes are emplaced to isolate them from significant exchange with other environmental components. Thus, the flow processes that occur in the unsaturated zone substantially contribute to a wide variety of hydrologic processes.

Scientifically, the unsaturated zone is highly complex and must be studied with an interdisciplinary approach. There is much variety in its natural constituents: soils, rocks, water, air, plants, animals, and microorganisms. Modern hydrology must consider interactions not only among these constituents themselves, but also with a wide variety of contaminants, including pesticides, fertilizers, irrigation wastewater, manure, sewage, toxic chemicals, radioactive substances, bacteria, mine wastes, and organic liquids.

This article first describes in a fundamental way the physical basis of phenomena that strongly relate to unsaturated flow. The next section presents techniques for obtaining quantitative values of properties that influence unsaturated flow, by direct measurement and by indirect means. The third major section describes some of the main hydrologic applications related to flow in the unsaturated zone.

PHENOMENA OF UNSATURATED ZONE FLOW

Water resides in an unsaturated porous medium along with air and solids, as Figure 1 illustrates. The usual tendency is for water to cling to solid surfaces, in films, and in curved air–water interfaces as shown. Hydrological processes involve movement of any of these materials. Thus, transport in an unsaturated medium is always a case of multiphase transport, though the term "multiphase" is used mainly for cases where gas, solid, or multiple-liquid phases are considered.

Various materials comprise the solid fabric of the unsaturated zone, including soil, stones, porous rock, and organic matter. A common distinction is between particulate, or granular, media in which particles are separate from each other, and consolidated or lithified media in which they are joined.

Basic features of soil and rock relevant to flow in the unsaturated zone include texture, structure, and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

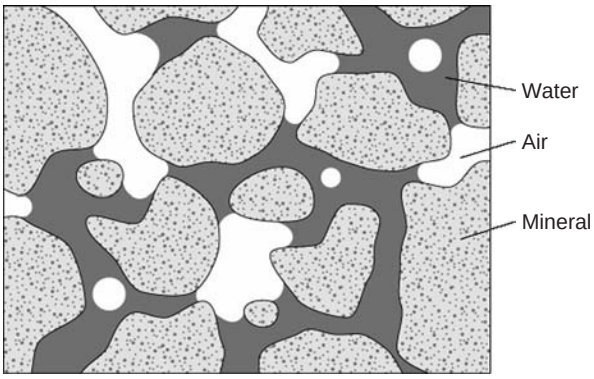


Figure 1 Microscopic cross-sectional view of a hypothetical unsaturated medium. Few intergrain contacts appear because these contacts are essentially points in three-dimensional space, and mostly do not lie in the two-dimensional plane of this figure

the content of mineral and nonmineral constituents (*see also* **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**). Texture, applied to a granular medium, refers to particle-size distribution. Structure refers to the arrangement of the solid components of the medium. Essential structural considerations include porosity, aggregation, cementation, and macropores. Structure is often considered alongside texture as a primary determinant of the pores of the medium.

This section considers unsaturated-zone flow phenomena through progressive increases in complexity as elaborated in Table 1. It concludes with a discussion of multiphase flow, with explicit attention to the transport of materials other than liquid water.

Unsaturated Hydrostatics

Water Content, Energy, and Pressure

The most basic measure of the water is volumetric water content, symbolized θ or θ_w, defined as the volume of water per bulk volume of the medium. An alternative is the mass-basis (or gravimetric) water content θ_m, the mass of water per mass of solid. These are related by the formula

$$\theta_w = \frac{\rho_b}{\rho_w}\theta_m \quad (1)$$

where ρ_b is the bulk density and ρ_w is the density of water. In general, the volumetric water content is most useful because its range has a clearly defined maximum at the medium's porosity, ϕ, and it relates easily to visualizations of intrapore geometry as in Figure 1.

Water is held in an unsaturated medium by forces whose effect is expressed in terms of the energy state of the water, force being the negative gradient of energy. The energy is usually expressed as a potential, taken as energy per unit volume. For an incompressible bulk material like water, the energy per unit volume can equivalently be considered as a pressure. Gravitational energy can be treated essentially as it is for saturated modes of fluid transport. The matric potential or pressure, which often is the only other significant type of energy determining the chief water transport processes in an unsaturated medium, arises from the interaction of water with a rigid matrix ("matric" being the adjective form of "matrix"). Sometimes this quantity is called *capillary potential* or *pressure*. Matric pressure may be thought of as the pressure of the water in a pore of the medium relative to the pressure of the air, in other words, the pressure difference across an air–water interface. When a medium is unsaturated, the water generally is at lower pressure than the air, so the matric pressure is negative. Another term is "suction", the negative of matric pressure. Many problems are simpler with the expression of matric pressure in head units.

Matric pressure is often thought of in relation to surface tension and capillary phenomena, especially in the range near saturation. A pore with water can be compared to a thin capillary tube with one end immersed, as in Figure 2.

Table 1 Levels of complexity considered in unsaturated flow

Type of flow	Phenomena	Mathematical description	Relevant features and properties of medium	Typical applications
Static (no flow)	All forces (on water) balance	Hydrostatic equation	Water retention	Available water for plants; basis for understanding
Steady	Flow driven by unchanging force field	Darcy's law	Hydraulic conductivity	Long-term averages
Unsteady – diffuse	Continuity, and force field affected by dynamic conditions	Darcy's law and continuity equation (combined as Richards' equation)	Hydraulic conductivity and water retention	Stable flow in media whose pore sizes approximate grain sizes
Preferential	Flow concentrated in pathways that differ in character from the bulk medium	Unknown; diverse alternatives in current use	Macropores, layer contrasts, propensity for unstable flow	Flow in media with significant nonuniformities; unstable flow

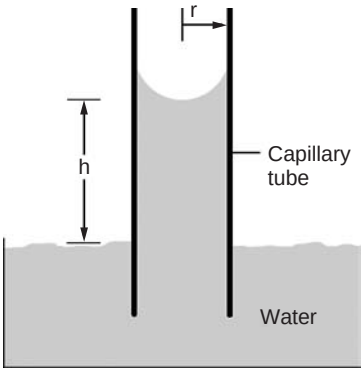


Figure 2 Capillary rise in a tube of circular cross section

The curvature of the air–water interface is inversely related to the water pressure: tighter curvature is associated with smaller pores and with more negative matric pressure. To quantify the relation between the tube radius and matric pressure ψ, consider the pressure in the water just below the curved air–water interface, which equals ψ. For the static situation, upward and downward forces balance. The downward force is the pressure acting on the area of the air–water interface, $\psi\pi r^2$. The upward force results from surface tension σ, the force per unit length of the air–water interface, acting on a circumference of the inner wall of the tube. For the case where the air–water interface is perfectly tangent to the tube wall, this force is $2\pi r\sigma$. Equating these forces gives

$$\psi = -\frac{2\sigma}{r} \tag{2}$$

Pores in natural media are not perfectly cylindrical, but the same relation applies if it is understood that r represents an effective pore radius. The relation given in equation (2) is useful in picturing what happens within a pore and in modeling unsaturated hydraulic properties.

Where the three phases come together, the angle between the air–water interface and the air–solid interface, measured through the water, is the contact angle. This angle depends on the materials. If the solid has much greater attraction to the liquid than to the air, it is highly wettable and the contact angle can be essentially zero, as assumed in the derivation of equation (2). If the solid has much greater attraction to the air, the material is nonwettable and would have a contact angle near 180°. Intermediate cases have intermediate values of contact angle, and some materials are difficult to classify as wettable or nonwettable. For a nonzero contact angle, formulas like (2) are modified by multiplying the force associated with surface tension by the cosine of the contact angle.

When the medium is so dry that water does not fill pores, but adheres in thin films to the solid matrix, the concepts of surface tension and capillarity are not so directly applicable, and the forces of adhesion establish the matric pressure. One can think of a capillary component of matric pressure that dominates when the medium is wet, and an adhesive component that dominates when it is dry.

Water Retention

If the matric pressure is close to zero, air–water interfaces are broadly curved, nearly all pores are filled, and the water content is high. If matric pressure is much less than zero, the interfaces are more tightly curved, they can no longer go across the largest pores, and the pores have less water in them. Thus, greater water content goes with greater (less strongly negative) matric pressure.

The relation between matric pressure and water content, called a *water retention* curve, depends on the medium. Larger pores empty first as the water content decreases. A medium with many large pores will have a retention curve that drops rapidly to low water content at high matric pressures. Conversely, a fine-pored medium will retain much water even at low matric pressures, and so will have a flatter retention curve. Figure 3 shows a typical retention curve for a sandy soil. The curve is far from linear and covers 5 orders of magnitude in ψ. This enormous range is difficult to work with and requires multiple measurement techniques. In most cases, investigators measure and plot only a portion of the range, usually at the wet end.

Considering the drying of soil from saturation, θ in Figure 3 stays high until a particular ψ value where it starts to decline. This ψ is called the *air-entry value*. By the capillary hypothesis, it exists because the largest

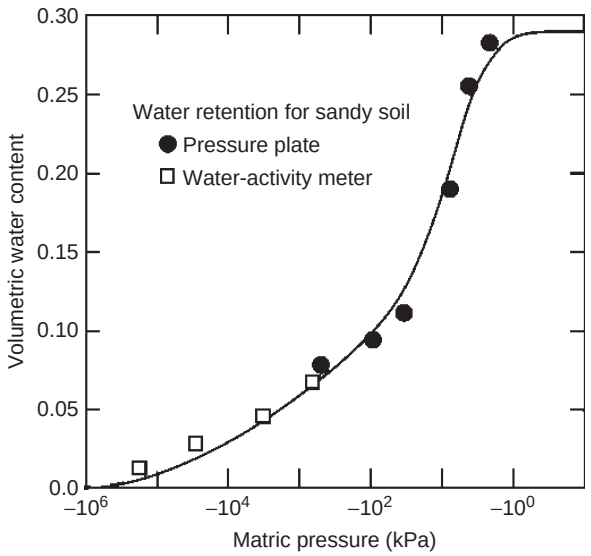


Figure 3 A retention curve for a sandy soil from the amargosa desert research site (Andraski, 1996). The points are measurements by two different methods and the smooth curve is a fit of the model of Rossi and Nimmo (1994)

fully wet pore of the medium will stay filled until the air–water pressure difference exceeds the equivalent ψ value of capillary rise. In natural media the air-entry value is usually poorly determined, as the decline in θ with ψ starts gradually, beginning at ψ nearly equal to zero. Artificial porous media, however, can be made in such a way that many pores are close to the size of the largest pore, so that air entry is a sharp and sudden phenomenon.

The maximum θ value of the curve usually is not equal to ϕ but rather something less, because at $\psi = 0$, trapped air occupies some of the pore space. At the other end, the curve goes to $\theta = 0$ at a ψ value of about -10^6 kPa.

In a granular medium, the particle-size distribution or texture relates in a general way to the pore-size distribution, as larger particles may have larger pores between them. Texture thus is a major influence on the retention curve. Additionally, the structure of the medium, especially as related to such features as aggregation, shrinkage cracks, and biologically generated holes, substantially influences the retention curve.

At low matric pressures, few pores are filled and a large fraction of the total water is in thin films. These films are thinner at lower matric pressures, with less energy for holding water onto the solid medium. At high matric pressures thicker films can be important, as when the medium is nearly saturated except for fractures and other large pores. If the films are thick enough, water in them may be free to flow. Tokunaga and Wan (1997) have measured film thicknesses of tens of microns at matric pressures between -0.1 kPa and 0.

The retention relation is strongly hysteretic: when measured as the medium wets, water content is less for a given matric pressure than it is when measured as the medium dries. Figure 4 shows a typical example. The outer curves, starting from extreme wet or extreme dry conditions, are called *main drying* and *main wetting* curves. The curves

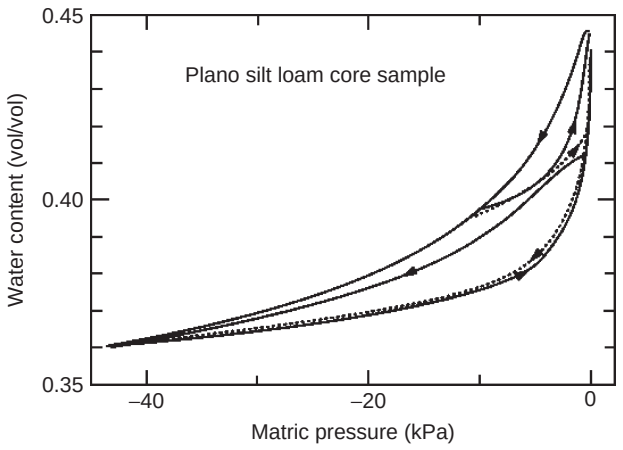


Figure 4 Hysteretic water retention in a soil core sample, data of Nimmo and Miller (1986)

starting from intermediate water contents are called *drying* and *wetting scanning* curves. Of course, there are whole families of possible scanning curves, starting from different points. Scanning curves that start on main curves are called *primary scanning* curves, those that start on primary scanning curves are secondary scanning curves, and so forth. Unsaturated zone investigations frequently neglect hysteresis, not always justifiably.

Several mechanisms cause $\theta - \psi$ hysteresis, the main one in fairly wet media being the Haines jump mechanism, illustrated in Figure 5. Unlike capillary tubes, pores in natural media are not uniform in effective diameter. The pore throats, which control the ψ at which pores empty and therefore determine the drying curve, are smaller than the pore bodies, which control the ψ at which pores fill and determine the wetting curve. As the medium dries and ψ decreases, water retreats gradually as the air–water interface becomes more curved. At the narrowest part of the pore throat, this surface can no longer increase curvature by gradual amounts, so in what is called a *Haines* jump, it retreats suddenly to narrower channels elsewhere in the nearby pore space. An analogous phenomenon occurs during wetting, when the decreasing interface curvature cannot be supported by the diameter of the pore at its maximum. Hysteresis occurs because the drying and wetting Haines jumps occur at different ψ values. Some of the pore space, where the movement of the interface is always gradual, is not subject to hysteresis. The amount of such nonhysteretic pore space varies with the pore geometry of the medium, and thus is one phenomenon that makes the degree of hysteresis vary among media. Other mechanisms include contact angle hysteresis, which is not well understood because contact angles within pores are difficult to measure, and adsorptive hysteresis, which may be quite significant near the dry end of the moisture range.

As explained below in connection with gas transport, natural media do not usually become fully saturated, even at $\psi = 0$. Some amount of air normally gets trapped in the form of bubbles enclosed by water, typically occupying 10 to 30% of the pore space. Thus, the main drying curves in Figures 3 and 4 have a maximum θ that is less than ϕ. Sometimes, though, the medium may be saturated long enough for all trapped bubbles to dissolve. On drying, the retention curve would then start from $\theta = \phi$, as illustrated by the data in Figure 6. This is called a *first* (or initial or primary) *drying* curve. One natural phenomenon likely to involve a first drying curve is the decline of a water table after a long time at a high level.

Temperature, because it affects surface tension and other relevant properties, significantly affects the retention relation. Increasing temperature means that less water will be held at a given matric pressure. This effect can be quantified using the gain-factor model of Nimmo and

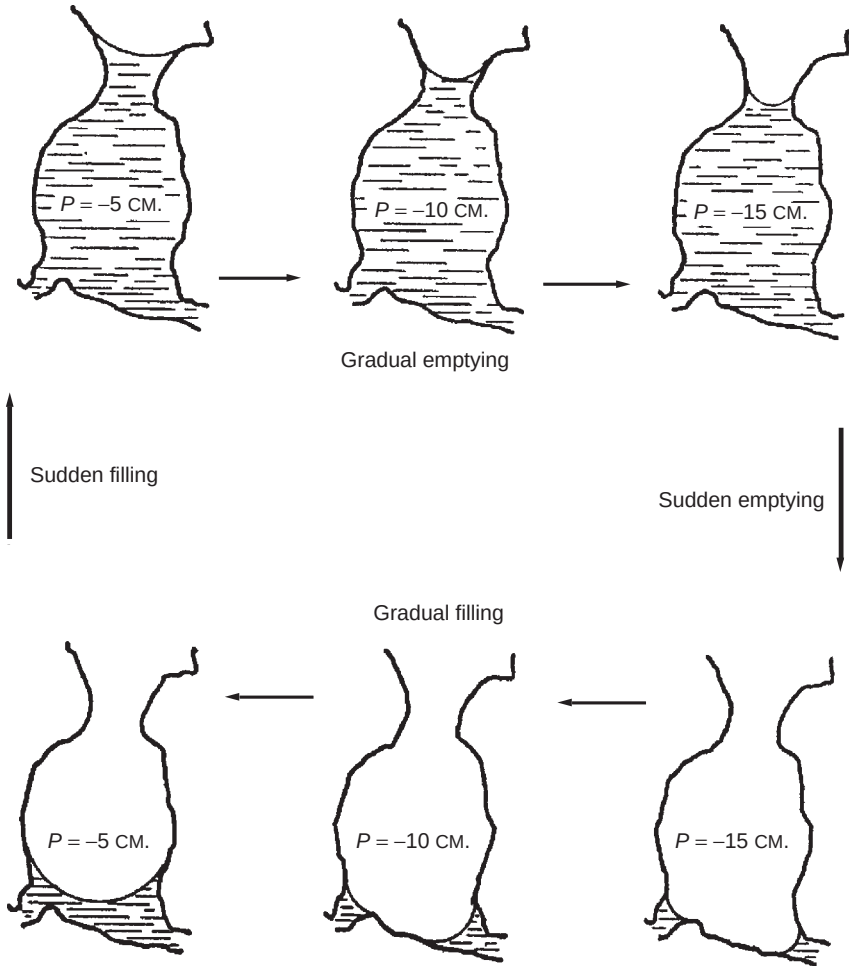


Figure 5 Haines jumps in a natural pore, illustration from Miller and Miller (1956) (Reprinted with permission from Miller, E. E. and Miller, R. D. (1956). Physical theory for capillary flow phenomena. *Journal of Applied Physics*, **27**, 324–332. © 1956 American Institute of Physics.)

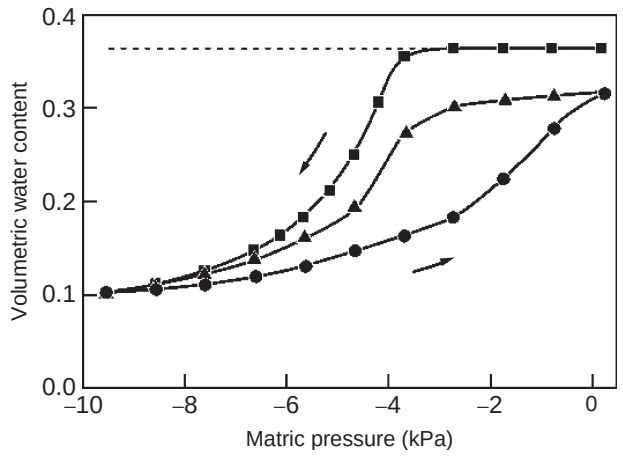


Figure 6 A first drying curve, shown with main wetting and drying curves, measured for a sandy soil (Oakley sand). The dashed line is at the value of θ equal to the porosity. Data of Stonestrom and Rubin (1989a)

Miller (1986) or the enthalpy-based model of Grant and Salehzadeh (1996).

Basic Unsaturated Flow

In conventional unsaturated flow theory, two types of factors determine water flux: driving forces (chiefly gravity and matric pressure gradients) and properties of the medium. The matric forces sometimes greatly exceed the gravitational force. Other forces may be significant for driving flow under some conditions, as when temperature gradients are significant, or when there are physical barriers to the movement of solutes. The medium properties of chief importance are the water retention relation and the hydraulic conductivity.

Unsaturated flow has its basic mathematical expression in Darcy's law, which states that the flux density q is proportional to the driving force. The proportionality constant is the hydraulic conductivity K. For the case of

one-dimensional flow driven by gravity and matric pressure gradients, Darcy's law can be expressed as

$$q = -\frac{K(\theta)}{\rho g}\left[\frac{\mathrm{d}\psi}{\mathrm{d}z} + \rho g\right] \tag{3}$$

where ρ is the density of water, g is the acceleration of gravity, and z is upward distance. The conversion factor $1/\rho g$ is shown here explicitly so that this expression can be used directly with ψ in SI pressure units such as kPa, and K in velocity units such as m/s. In head units, ρg (the weight of water per unit volume) equals dimensionless unity and ψ takes dimensions of length.

K of the medium depends on the whole set of filled pores, especially the size, shape, and connectedness of filled channels. The retention relation and the history of the moisture state determine what pores are filled. In unsaturated media, as illustrated by the measurements in Figure 7, K depends very strongly on the water content. As water content decreases, the large pores, which make by far the greatest contribution to K, empty first. Then, not only are there fewer filled pores to conduct water, but they are smaller and therefore less conductive because there is more viscous friction. With fewer pores filled, the paths of water flowing though the medium become more tortuous. When the soil is quite dry, few pores are filled, and the water moves mainly through poorly conducting films adhering to particle surfaces. These factors combine to reduce hydraulic conductivity by several orders of magnitude as the soil goes from saturation to typical field-dry conditions.

Other factors also can influence hydraulic conductivity. Matric pressure is relevant though its main effect is indirect, through influence on θ. Temperature affects K through its influence on viscosity and other factors. Microorganisms can reduce K by constricting or obstructing channels though the pores. Chemical activity can similarly reduce K if precipitates form. It can also influence K by affecting the cohesion of particles and hence the structure of the medium.

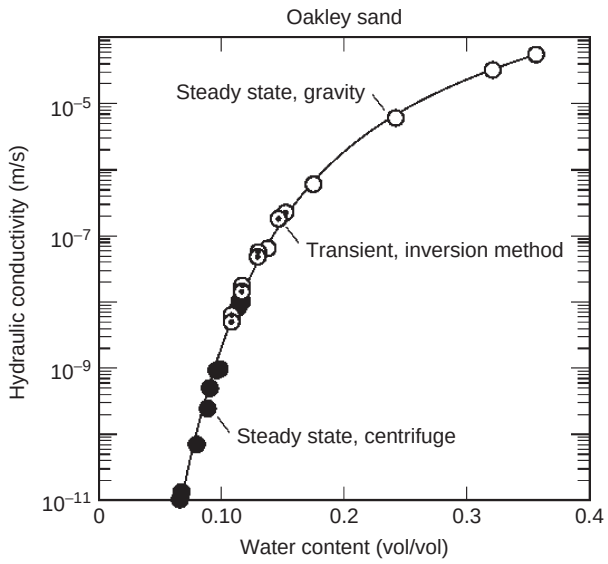


Figure 7 Hydraulic conductivity for a sandy soil (Oakley sand), measured by three methods

Unsteady Diffuse Flow

When unsaturated flow is transient (nonsteady), as it generally is, the flow itself causes the water content to change throughout the medium, which leads to continuously changing hydraulic conductivity and driving forces. These interacting processes, first described by Buckingham (1907), can be accommodated mathematically by combining the equation of continuity

$$\frac{\partial\theta}{\partial t} = -\frac{\partial q}{\partial z} \tag{4}$$

with Darcy's law (3) to get Richards' (1931) equation, which for one-dimensional vertical flow within a medium in earth gravity can be written

$$C\frac{\partial\psi}{\partial t} = \frac{1}{\rho g}\frac{\partial}{\partial z}\left[K\frac{\partial\psi}{\partial z}\right] + \frac{\partial K}{\partial z} \tag{5}$$

where C is the differential water capacity, a property of the medium defined as $\mathrm{d}\theta/\mathrm{d}\psi$. It is also possible to formulate this equation in terms of θ (Sposito, 1986). The equation can be solved numerically in the general case (e.g. Lappala *et al.*, 1987). Analytical solutions have also been developed (e.g. Salvucci, 1996), though these require simplifying assumptions that are not directly applicable to most situations of unsaturated flow. Richards' equation does not adequately represent all circumstances of unsaturated flow, for example, at wetting fronts and flow instabilities (Stonestrom and Akstin, 1994; DiCarlo, 2004). Some of these are discussed below in connection with unstable flow.

An alternative formulation of unsteady unsaturated flow depends on a property called *hydraulic diffusivity* or *soil-water diffusivity*. This is based on the fictional but useful assumption that the flow is driven by gradients of water content rather than potential. Richards' equation (5) without the gravitational term can be transformed into

$$\frac{\partial\theta}{\partial t} = \frac{\partial}{\partial z}\left[D(\theta)\frac{\partial}{\partial z}(\theta)\right] \tag{6}$$

where

$$D(\theta) = K(\theta)\frac{\mathrm{d}\psi}{\mathrm{d}\theta} \tag{7}$$

is the hydraulic diffusivity (Childs and Collis-George, 1950). $D(\theta)$ is a property of the medium dependent on θ and its history. Advantages of this formulation include that: (i) $D(\theta)$ can be easier to measure than $K(\theta)$ and

$\theta(\psi)$; (ii) the typical increase of $d\psi/d\theta$ with decreasing θ compensates in part for the decline in K with decreasing θ, so that in the field D usually varies less than K; and (iii) many mathematical techniques for solving equation (6) have previously been developed for diffusion applications.

Preferential Flow

Preferential Paths

Flowpaths that permit fast movement, whether because of their character (e.g. large pore diameter) or their present state (e.g. high water content) are called *preferential paths*. Such a path may be a single pore (i.e. a macropore), a connected series of pores, or a group of adjacent pores acting in parallel. These paths are common in natural porous media with significant heterogeneity, such as surface soils and fractured rock. Flow in preferential paths transports water and contaminants much faster than would be predicted from bulk medium properties and Richards' equation. Another important effect of preferential flow is that a relatively small fraction of the subsurface medium interacts with contaminants, which limits adsorption and other attenuating processes. Three basic types of preferential flow (Figure 8) are (i) macropore flow, caused by flow-enhancing features of the medium; (ii) funneled (or deflected or focused) flow, caused by flow-impeding features of the medium; and (iii) unstable (or fingered) flow, caused by temporary flow-enhancing conditions of parts of the medium.

Macropores, distinguished from other pores by their larger size, greater continuity, or other attributes, conduct preferential flow under conditions such as extreme wetness. Common macropores include wormholes, root holes, and fractures. Where macropore flow occurs, flow through the remainder of the medium may be called *matrix flow*. When macropores are filled, flow through them is fast. When macropores are empty, they constitute a barrier to matrix flow and there may be essentially no flow through the macropores themselves. In some conditions, however, there may be significant film flow along macropore walls (Su *et al.*, 2003). Macropores that are partly filled provide a variety of possibilities for the configuration and flow behavior of water.

Funneled flow commonly occurs with contrasting layers or lenses, where flow deflected in direction becomes spatially concentrated. The local increase in water content causes a corresponding increase in hydraulic conductivity and flux. In field experiments in a sandy soil, Kung (1990) found that the flow became more preferential with depth (Figure 9). At about 6-m depth, the flow was moving through less than 1% of the whole soil matrix. Although this medium had no significant observable macropores, preferential flowpaths were the dominant pattern. The main feature causing this preferential flow was "... an interbedded soil structure with textural discontinuities and inclined bedding planes". Considering Kung's work and others', Pruess (1998) noted that because funneling results from horizontal impediments and can produce very rapid flow "... we have the remarkable situation that unsaturated seepage can actually proceed faster in a medium with lower average permeability".

Unstable variations in flow and water content, even within a uniform portion of the medium, can increase flow rates considerably (Wang *et al.*, 2003). A typical case has a layer of fine material above coarse material. Downward-percolating water does not immediately cross

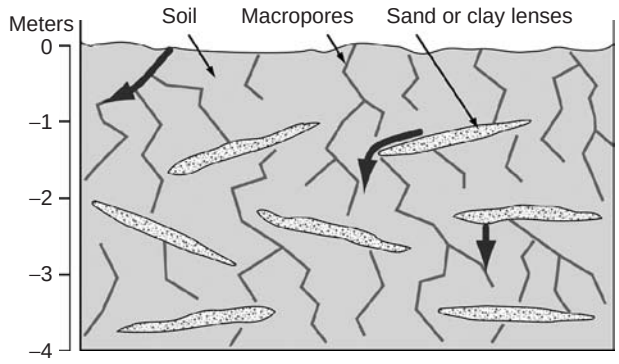


Figure 8 Three basic types of preferential flow. Arrows indicate narrow regions of faster flow than their surroundings. Macropore flow occurs through channels created by aggregation, biotic activity, or similar causes. Funneled flow occurs when flow is deflected by heterogeneities of the medium so as to create zones of higher water content and greater K. Unstable flow can be generated at layer boundaries such as the bottom of a sand lens at right, where flow into the lower layer moves in the form of highly wetted fingers separated by regions of relatively dry soil. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

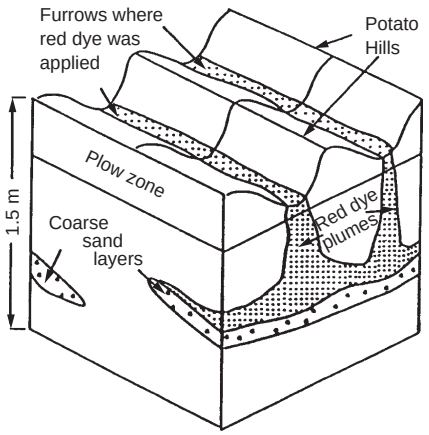


Figure 9 Funneled flow investigated using a dye tracer (Reprinted from Kung *et al.* 1990, © 1990, with permission from Elsevier)

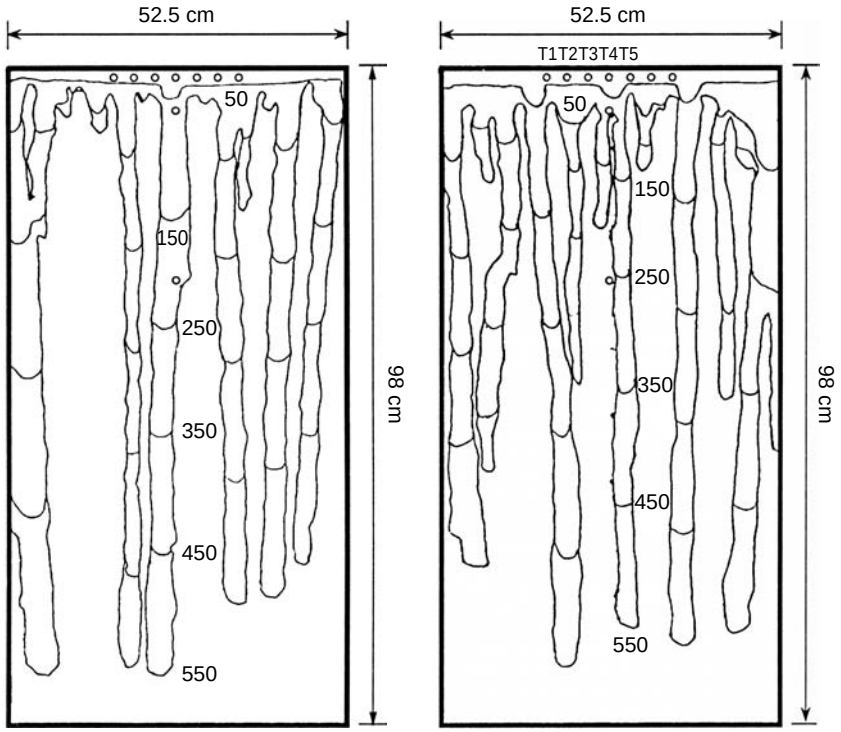


Figure 10 Fingers generated in unstable flow in a laboratory investigation (Reproduced from Selker *et al.,* 1992, by permission of American Geophysical Union)

the interface into the coarse material. When water pressure builds up significantly at the interface, water may break through into the coarse medium at only a few points. The material near individual points of breakthrough becomes wetter and hence much more conductive. For some time thereafter, additional flow into the coarse material moves in the few "fingers" that are already wet (Figure 10). Between fingers, the medium can be relatively dry. In addition to textural contrasts, hydrophobicity (water repellency) and air trapping may contribute to flow instability. Layer boundaries may interrupt preferential flowpaths. There may also be processes that homogenize preferential flow, though such effects have been little studied.

In virtually every unsaturated-zone transport problem, it is essential to assess the prevalence of preferential flow and the flow mechanisms that might be active. One approach is to evaluate the features of a particular site that might cause macropore, funneled, or unstable flow. Another is to collect and evaluate evidence from observed water or solute distributions that cannot easily be explained without hypothesizing preferential flow.

Quantification of Preferential Flow

One straightforward way of quantitatively treating preferential flow is by representing preferential flowpaths with discrete pathways whose geometry and water content, with appropriate laminar-flow expressions, predict the flow rate through the part of the medium they occupy. Usually this is impossible because the position, number, shape, orientation, and connectedness of pathways are unknown.

Conceptually opposite to the discrete pathway approach is the widely used equivalent-medium approach. The key assumption is that the effective hydraulic properties of a large volume of the medium are equivalent to the average properties of a homogeneous granular porous medium. The effective hydraulic properties then can be applied directly in numerical simulators employing Darcy's law and Richards' equation. The main advantage of the equivalent-medium approach is that the many existing theories, models, and techniques developed for diffuse flow in granular media can be applied to preferential flow. This approach is only valid if certain conditions are met. The medium must be representable as a continuum (Bear, 1979, p. 28–31), but this requirement is difficult to satisfy in macroporous media with flow inhomogeneities at relatively large scales. Representative volumes may necessarily be so large that only a small fraction of the assumed volume is participating in the flow. Modeled transport velocities then may greatly underestimate the actual velocities. In practice, an adjustment of the effective porosity of the equivalent granular medium is commonly done to compensate for such effects, though this adjustment will not produce accurate predictions if, as expected, the degree of preferential flow depends strongly on the degree of saturation. The

equivalent-medium approach also is inadequate where it is essential to have knowledge of the individual flowpaths.

Unstable flow complicates the quantification in at least two ways that do not apply to macropore or funneled flow: (i) unstable flow is not tied to particular permanent features of the medium; and (ii) the preferentiality of unstable flow changes dynamically (e.g. unstable flowpaths commonly grow wider as flow progresses through them). Theories of unstable flow in terms of scaling and other concepts, have been developed for example, by Raats (1973), Parlange and Hillel (1976), Glass *et al.* (1989), Selker *et al.* (1992), Wang *et al.* (1998), Eliassi and Glass (2002), and Jury *et al.* (2003).

Multimodality

A family of approaches that rely on a conceptual partitioning of water or pore space into portions with different flow rates and behaviors may more realistically represent flow that includes preferential paths than do traditional unsaturated-flow models. Models used in these approaches have names such as "dual-porosity" and "dual-modality". The concept of mobility, that is, how easily certain system components (in particular, water and contaminants) move within different parts of the medium, is frequently used in defining and characterizing these approaches.

The simplest of these models assume matrix flow to be negligible, so that all flow is preferential flow. Given the nonlinear nature of unsaturated flow, the difference in conductance between, say, a wormhole, and an interparticle space between clay or silt grains, may be several orders of magnitude. For practical purposes this may constitute a mobile–immobile distinction. Other models assume the matrix to be permeable but with different properties and possibly different modes of flow than the portion of the medium that has preferential flow.

The degrees of possible mobility cover a continuum, and truly immobile water is unlikely. Some models postulate three degrees of effective mobility, for example, mesopores in addition to micropores and macropores (Luxmoore, 1981). A closer approach to a continuum of mobility is that of Griffioen *et al.* (1998).

Solute transport occurs in both diffuse and preferential modes. **Chapter 152, Modeling Solute Transport Phenomena, Volume 4** describes important phenomena, many of which are as relevant in unsaturated as saturated media.

Multiphase Flow

The transport of water and solutes is often considered independently of other fluids or solids. But normally, more than one phase is transported in an unsaturated medium. Besides the liquid water, air and other gases, there may be nonaqueous liquids that need to be considered as separate phases, and there may be solid particles that are free to move. Multiphase flow is a common element of contamination problems because many contaminants are nonaqueous, for example, volatile vapors, organic liquids (oils, solvents, etc.), colloids, and microorganisms.

For gases and nonaqueous liquids, it is often helpful to use the concept of permeability to represent the flow-influencing property of the solid matrix independently of the fluid. This relates to K according to

$$k = \frac{K\mu}{\rho g} \tag{8}$$

where k (dimensions L^2) is the permeability, μ is the viscosity (dimensions $ML^{-1}T^{-1}$), and ρ (ML^{-3}) is the density of the fluid. Darcy's law (3) takes the form

$$q = -\frac{kk_r(\theta)}{\mu}\left[\frac{d\psi}{dz} + \rho g\right] \tag{9}$$

where k_r is relative permeability, a dimensionless factor needed to account for the θ dependence in an unsaturated medium. Ideally, a medium has the same k for any fluid. This independence is not absolute because fluid properties other than μ and ρ, such as those related to slip effects (described below) or other non-Newtonic phenomena, can affect the flow, and some fluids cause structural changes that affect k.

Transport of Gases

Lacking the strong intermolecular forces of a liquid, individual gas molecules move nearly independently. Gas in the unsaturated zone commonly extends over large distances without being blocked by liquid, forming a continuous gas phase. Unlike solutes in liquid, the minority components within a gaseous mixture tend to mix completely at the molecular level, and to move in the same way as any other gas molecules. The more independent motion of the molecules also makes gases less viscous than liquids. Gas is highly compressible, which complicates some transport phenomena while making possible some useful measurement techniques. Gas flow exhibits slip, also called the *Klinkenberg* effect. Even the gas molecules closest to surfaces are free to move, and, unlike the essentially stationary boundary molecules of liquids, will have a net motion in the direction of the bulk flow of gas. This makes the permeability to gas somewhat greater than one would expect from measurements of the liquid permeability. Gas molecules dissolve in liquids. Like other solutes, gases differ in solubility. Carbon dioxide is much more soluble than nitrogen or oxygen. Gases also interact in chemical or biological reactions, of great importance in some contamination and restoration problems.

Strictly speaking, there is viscous friction between gas and adjacent liquid, so moving liquid tends to drag gas with it, and vice versa. Usually this friction is so much smaller than other driving forces that it can be ignored.

Gas flow is then independent of liquid flow, and often described by a separate implementation of Darcy's law. The gas pressure takes the place of matric pressure. Gas flow is often slow because the driving forces are small. When components of the gas are of comparable density, gravity is not a major driving force, and, unlike water driven by differences in matric pressure, in gases there normally is no mechanism to sustain large pressure gradients. Therefore convection is often not as important as other mechanisms, especially diffusion. In a gas, the molecules of a minority component move more easily than they would in a liquid, because of the lesser intermolecular cohesion. This causes greater diffusion. Dispersion, on the other hand, still depends largely on convective flow rates, and so is not necessarily greater in a gas. In contrast to solute transport, therefore, diffusion of gases is normally more important than dispersion. The diffusion coefficient for two gases in a porous medium is directly related to the diffusion coefficient for the gases in a free space, but also depends on such factors as pore size, tortuosity, and continuity.

Gas is trapped if it resides in bubbles or pockets from which every possible path to the outside of the medium goes through liquid or solid. Air easily becomes trapped during wetting, wherever a pore is slower than its neighbors to fill with water. Once the pore is surrounded by water, air within it is trapped. Stonestrom and Rubin (1989b) found experimentally during the drying and wetting of two different soils that some amount of air was trapped whenever θ was greater than about 70% of the porosity. A medium that has been "saturated" by ordinary means usually has a significant fraction of its pore space occupied by trapped bubbles of air within certain pores. Sometimes the word "satiated" is used for this condition, to reserve "saturated" for the case when there is strictly no gas in the pores. In general, though, "saturation" means that the medium has finished absorbing water by the wetting process it has been subjected to. This normally results in a matric pressure of zero, with trapped air. Trapped air bubbles change size with matric pressure in accordance with Boyle's law. They also shrink as gas dissolves or expand as it comes out of solution. The quantity of trapped air is easy to determine at saturation because all air present is trapped; measurement of volumetric water content subtracted from the porosity will indicate volumetric trapped air content. For the amount of air trapped at saturation, Mualem (1974) proposed a rule of thumb based on empirical observations, that 10% of the pore space will be occupied by trapped air. In general, this will vary with the medium, rate of wetting, water content before wetting, and other factors, and can far exceed the 10% guideline.

Transport of Nonaqueous Liquids

Liquids such as oils and organic solvents that do not easily dissolve in water are retained and transported within a porous medium in some of the same ways as water, but with crucial differences. Being a phase separate from water, they are often called *nonaqueous phase liquids*, or NAPLs. Hess *et al.*, (1992) and Essaid *et al.*, (1993) give examples of how fluid contents can be observed and measured and their distributions simulated at a site of oil contamination of groundwater.

To a first approximation, nonaqueous liquids obey the laws of surface tension and viscous flow, just like water. Liquids such as oils typically have a weaker tendency to cling to solid surfaces than does water. Thus, they are a relatively nonwetting phase, and they exist mainly in blobs or interconnected shapes within the pores of the medium, separated from the particle surfaces by a layer of water (Figure 11). With enough nonaqueous liquid present, it also can form a continuous phase across significant distances within the medium. Then its transport may be described using the permeability form of Darcy's law (9). Otherwise, the nonaqueous liquid is present mainly as isolated blobs, around which water flows as it flows around solid particles. Dillard *et al.*, (1997) show how such factors as the nature of the K distribution can affect oil transport in the subsurface.

Nonaqueous liquid transport often must be considered in combination with other modes. Many liquids volatilize significantly, so the resulting vapor moves in gaseous form. Nonaqueous liquids normally dissolve significantly and undergo transport also as a solute. These modes of transport are less affected by the degree of continuity of the nonaqueous phase than is the liquid transport mode.

Transport of Particles

Small solid particles can move through the subsurface in response to a complex set of factors. These particles include bacteria and other microorganisms as well as nonliving colloids such as rock fragments, mineral precipitates such

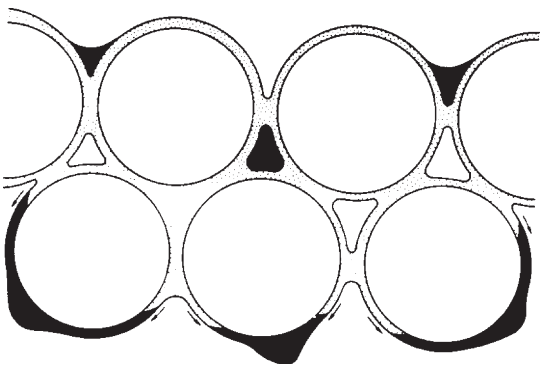


Figure 11 Four phases present in an artificial medium made of glass spheres. In this cross-sectional diagram, stippled areas are water, black areas are nonaqueous liquid, and white areas within the circles are solid glass and elsewhere are air (Reproduced from Schwille, 1988 by permission of Lewis Publishers (Chelsea, MI, USA))

as iron oxides and carbonates, weathering products such as clay minerals, macromolecular components of dissolved organic carbon such as humic substances, and microemulsions of nonaqueous liquids.

Mobile particles play several roles in practical unsaturated zone problems. (i) They themselves are sometimes contaminants. (ii) They act as carriers of any contaminants that sorb onto them, overriding the intrinsic transport characteristics of those contaminants. This complicates solute transport problems, especially when the solutes of interest adsorb onto the colloids and the colloids are more mobile than the solute. (iii) They can be naturally present or artificially introduced agents (typically bacteria) that break down organic contaminants and other compounds into other substances whose presence may be more desirable. (iv) Their presence and possible redistribution within the pore space can affect hydraulic conductivity or other transport properties.

Mechanisms of particulate transport include convection, advection, and adsorption, as for solutes, and other mechanisms (Harvey and Garabedian, 1991). Adsorption depends strongly on the composition of the medium, generally increasing with additional clay and decreasing with additional organic matter. Another mechanism of great importance is straining, the blocking of further motion by pores that are smaller than the particles themselves. For microbiological organisms such factors as death and reproduction also must normally be considered.

In unsaturated media, there are additional mechanisms that can make particle transport more complex than in saturated media (DeNovio *et al.*, 2004). The movement is strongly retarded as water content decreases, leading to a sometimes useful assumption that through the soil there is little motion except when the medium is nearly saturated. Particles can be removed from mobile fluid through mechanical filtration by films as well as by small pore spaces. Besides the reasons for water itself and solutes to move much more slowly at lower water contents, there is also a strong tendency for solid particles to adsorb onto the air–water interface (Wan and Wilson, 1994), as illustrated in Figure 12. This adsorption may be stronger than the adsorption onto the soil matrix itself. This effect also means that when the pores are all filled with water except for some amount of trapped air, that air may significantly retard the transport.

The mobility of inorganic colloids in the subsurface is controlled by chemical interactions between colloids and immobile matrix surfaces, and by hydrological and physical factors. Changes in aqueous chemistry can cause colloids to aggregate or to disaggregate. Higher ionic strengths, for example, usually favor colloid aggregation. This affects mobility because in general larger, aggregated particles are less mobile. Particles can be released into subsurface water as a result of mechanical grinding of mineral surfaces. Hydrodynamic forces associated with increases in flow rate can dislodge colloids.

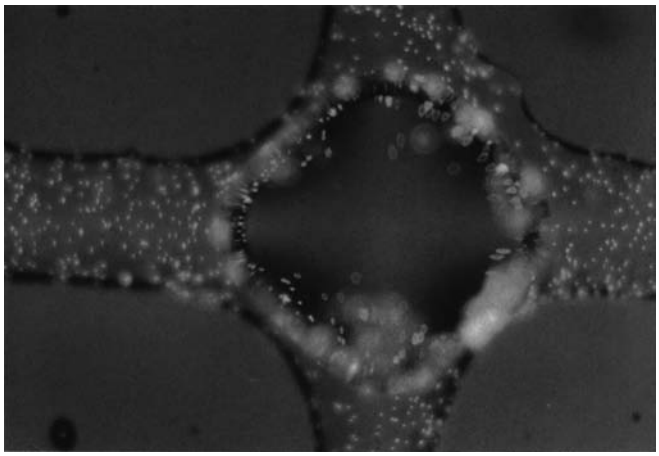


Figure 12 Microscopic cross-sectional view of hydrophobic latex particles in water in artificial pores (Wan and Wilson, 1994). The main pore body, the intersection between the two perpendicular pores, is about 200 μm across. The latex particles (light-colored dots) are 1 μm in diameter. A large air bubble occupies most of the main pore. Some of the particles are adsorbed on the air-water and water-solid interfaces (Reproduced from Wan and Wilson, 1994 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

DETERMINATION OF UNSATURATED CONDITIONS AND PROPERTIES

Moisture State

Water has many distinctive properties – volatility, density, molecular and nuclear structure, and electrical and thermal properties – that lend themselves to measurements indicating the amount of water present in a porous medium. The basic gravimetric method is to dry soil in an oven until the weight is constant, so that the difference between that weight and the initial wet weight indicates how much water was in the soil. This is often used as the standard for calibrating other methods.

Several methods in widespread use are minimally disruptive. X-ray or γ-ray attenuation can indicate water content in space and time: the more water, the less the beam intensity coming out of the medium. This effect is sometimes used tomographically to produce a two-dimensional map of water content in a cross section of the medium. Neutron scattering is commonly used for monitoring water content as a function of depth in the field. This method is based on the relative effectiveness of the various components of the wet soil in slowing neutrons. Because of the conservation of energy and momentum, a neutron passing through matter

is slowed down most effectively by collisions with particles that are about the same size as itself. In soil, essentially the only particles that are the size of a neutron are the hydrogen nuclei in water molecules. A probe that includes both a source of fast neutrons and a detector of slow neutrons registers more counts in wetter soil. Commercially available equipment has a neutron source and detector housed in a cylindrical probe that can be lowered to various depths in a lined hole, to obtain measurements as a function of depth. Another way to monitor water content over a period of time in the lab or field is by measurement of the dielectric constant of the medium, usually by time-domain reflectometry (TDR) (Topp *et al.*, 1980). Liquid water has a much greater dielectric constant than other constituents of soil or rock, so this effect can indicate the amount of water present within the volume sensed. For most applications, TDR electrodes in the form of metal rods are inserted into the soil. Various geometries of electrodes are possible, including a coaxial cylinder for laboratory use with core samples. A less common method is to measure electrical conductivity (e.g. Sheets and Hendricks, 1995), which increases with water content. This principle can be applied tomographically (e.g. Daily *et al.*, 1992) for observing two- or three-dimensional details of changing water distributions in the field. Ground penetrating radar (e.g. Eppstein and Dougherty, 1998) can also be applied directly or tomographically.

The most direct measurement of matric pressure is by a tensiometer. In firm contact with the porous medium, this device allows for equilibration of pressure between the water in unsaturated pores and the water in a larger chamber where a gauge or transducer reads the pressure (Figure 13). The key feature is the porous membrane that contacts the soil. In order to assure a continuous liquid water pathway from the pore water to the chamber water, this membrane must remain totally saturated. Thus, it must have an air-entry value beyond any ψ value to be measured, and must not be allowed to dry out. There is a basic ψ limit of about -80 kPa, below which ordinary tensiometers fail because of runaway bubble formation, though Miller and Salehzadeh (1993) have shown that devices that remove dissolved air can extend this range.

Other methods are available for media drier than -80 kPa and for easier application when less accuracy is acceptable. Some of these are based on the humidity of the air in soil pores. A low (strongly negative) matric pressure increases the pore water's effectiveness for absorbing water molecules out of the vapor in the soil air, resulting in a lower relative humidity. The effect is slight, however; a 0 to -1500 kPa matric pressure range corresponds to a 100 to 99% range in relative humidity. Psychrometers (Andraski and Scanlon, 2002) and chilled-mirror devices (Gee *et al.*, 1992) are both used to measure humidity for this purpose. Another class of methods uses an intermediary porous medium of known retention properties. Examples include

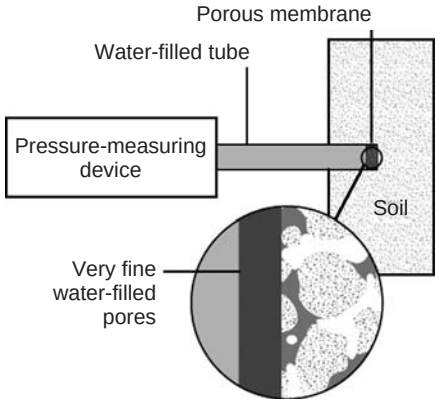


Figure 13 Schematic diagram of a tensiometer in contact with soil for matric pressure measurements. The pressure measuring device can be an electrical transducer, a manometer, or other pressure gauge. The porous membrane is often porous ceramic or sintered metal, or one of various other materials intended for use as filters. Especially in field applications, the membrane is often shaped as a tube or cup rather than a disk, to have greater contact area with the soil

gypsum blocks, nylon fabric, and filter paper (Scanlon *et al.*, 2002). This medium is placed in contact with the medium to be measured so that the matric pressure becomes equal in both. Then the water content of the intermediary medium is measured by other means (usually electrical conductivity, thermal diffusivity, or mass) and translated into a matric pressure using the known properties. Often the retention curve of the intermediary medium is not known directly, but rather a calibration relation between matric pressure and the measured quantity (e.g. electrical resistance) is known.

Water Retention

Measurement

Any system that makes independent simultaneous measurements of water content and matric pressure can indicate water retention relations. Additionally, there are methods specifically intended to measure this property. Many of these use a large porous membrane, often ceramic, to permit equilibration of water pressure between the porous medium on one side of the membrane and bulk water on the other. The pressure of this bulk water (and hence the pore water pressure) is controlled, as is the air pressure in the medium, in order to control its matric pressure. Pressure or suction chambers that do this use the principle of water-equilibration through a membrane of the sort used in tensiometers as in Figure 13. The pressure, or less commonly the volume of water, is adjusted through a planned sequence, and paired values of matric pressure and water content (one of them controlled and the other measured) represent the retention curve.

Estimation

Because various nonhydraulic properties of a medium, especially particle-size distribution, correlate in some way with water retention but are considerably easier to measure, models have been developed for estimating water retention from these. The basis is that the water retention curve of a medium depends directly on its pore-size distribution, which is related to the particle-size distribution. In a randomly packed medium, intergrain pores are expected to be larger where the particles are larger. Arya and Paris (1981) developed a widely used model based on this principle, using capillary theory and particular assumptions about the effective radii of capillaries to be associated with particular sizes. Models of this type often work reasonably well for sandy media. In general, however, the correlation between particle and pore size does not hold. For example, increasing fine-particle content, especially with clay particles, usually correlates in soils with increasing aggregation and more numerous macropores. Adding clay to sand can increase rather than decrease the number of large pores. More recent models (e.g. Rieu and Sposito, 1991; Nimmo, 1997) extend their applicability to more types of media with the use of additional information such as aggregate-size distributions.

Another way of estimating water retention without measuring it is with statistically calibrated pedotransfer functions such as the Rosetta model (Schaap, 1999). The basis for these is not a principle like the correlation of pore and particle size, but rather a database of measured water retention and other properties for a wide variety of media. Given a medium's particle-size distribution and other properties such as organic matter content, a model of this type can estimate a retention curve with good statistical comparability to known retention curves of other media with similar nonhydraulic properties. The foundation in empirical data helps to compensate for such effects as the increase in large pores with increasing clay. The accuracy and reliability of these models are limited, however. They sometimes, but not always, produce acceptable estimates. Without any retention measurements for the medium in question, it is usually impossible to know if the model result is a good representation of the retention curve or not.

Hysteresis models have been developed to represent the entire set of hysteretic moisture relations from a partial data set. The most widely used is that of Mualem (1974), which yields a complete set of scanning curves given both drying and wetting main curves. Other models require greater or lesser amounts of data. The model of Nimmo (1992), for example, requires the main drying curve and two points of the main wetting curve. In general, less stringent data requirements lead to modeled results of less reliability.

Empirical Formulas for Water Retention

Whatever method is used to determine a water retention curve, it is frequently convenient to express it as a parametric empirical formula. This helps in providing interpolation or extrapolation of data, in giving a smooth, continuous form that is easy to work with mathematically, and in representation of the curve with a few (typically 2 to 5) parameter values. Among the most widely used empirical formulas are that of Brooks and Corey (1964)

$$\theta = (\theta_{\max} - \theta_{\min}) \left[\frac{\psi}{\psi_b} \right]^b + \theta_{\min} \tag{10}$$

where ψ_b, b, and $\theta_{\min}$ are fitted empirical parameters and $\theta_{\max}$ is the maximum value of θ; and that of van Genuchten (1980)

$$\theta = (\theta_{\max} - \theta_{\min}) \left[\frac{1}{1 + \left(\frac{\psi}{\psi_c} \right)^{\nu}} \right]^{\mu} + \theta_{\min} \tag{11}$$

where ψ_c, ν, μ, and $\theta_{\min}$ are fitted empirical parameters. Representations as simplified as these have shortcomings. For example, the Brooks and Corey curve works poorly in the wettest portion of the retention curve, and the van Genuchten in the driest. Numerous alternatives have been published, with advantages for particular applications or types of media. Some of these are based on different mathematical formulas, for example, the lognormal distribution (Kosugi, 1994), which has advantages in understandability of the significance of its parameter values. Others are formed by combining formulas by addition (Ross and Smettem, 1993) or by joining different functions that each apply in a different portion of the range. For example, Ross *et al.* (1991) developed an equation for realistic representation in the dry range, Rossi and Nimmo (1994) for the whole range from oven-dryness to saturation, and Durner (1994) for bimodality of pore-size distribution.

Dynamic Characteristics

Measurement of K and D

The most accurate measurements of hydraulic conductivity are by steady-state methods. One way is to establish constant (though not necessarily equal) pressures of water at two opposing faces of a porous medium, measure the flux density, and calculate K using Darcy's law (Mualem and Klute, 1984). Another is to force water through at a constant and known flux density, which lets the matric pressure become uniform in part of the sample, then to compute K from the known flux density and force of gravity (Childs and Collis-George, 1950). With gravity as the main driving force, steady-state measurements are possible only for the high K values of fairly wet soil. Centrifugal force makes possible the accurate measurement of K at low water contents (Nimmo *et al.*, 2002b).

There are many techniques for measuring unsaturated hydraulic conductivity using unsteady flow. One of these is the instantaneous-profile or unsteady drainage flux method (Hamilton *et al.*, 1981). It is useful in both the laboratory and the field. This method is based on a determination, within a medium in which unsteady flow has been established, of both the flux density and the matric pressure gradient, from water contents in space and time, and from matric pressure measurements at a given instant of time after the application of water. Another alternative for laboratory applications uses flow driven by evaporation. There are various indirect and inverse methods – a wide variety of situations where available data describing water flow over time can provide information for an estimation of K or D (Hopmans *et al.*, 2002). The tension infiltrometer method is in widespread use for field applications. This method uses the measured infiltration rate as a function of time for water applied at controlled ψ values as data to calculate the unsaturated hydraulic properties. It is often implemented as an inverse method, as discussed further below. Garnier *et al.* (1997) have developed a transient flow method for the more difficult case of soils that swell as water content increases. Infiltrometer methods are useful in field applications (e.g. Smettem *et al.*, 1995; Vandervaere *et al.*, 2000a; Vandervaere *et al.*, 2000b). Water is supplied at a known or measured rate, generally through a flow-restricting membrane or crust to establish unsaturated flow, and measurements of the changing moisture state in the soil are used with the flow rate to estimate unsaturated K.

Diffusivity methods also give K if the water retention relation is known. Typically, these methods are easier but less accurate than direct K measurements. One family of techniques examines the water content distribution within a medium whose water content is changing with inflow. Another popular technique for diffusivity is the one-step outflow method (Gardner, 1956; Doering, 1965) in which a sample in a pressure chamber with an outflow membrane is suddenly exposed to a step increase in air pressure. This forces water out through the membrane as the matric pressure and water content decrease. Measurements of the rate of outflow permit an estimation of $D(\theta)$, normally to within about 1 order of magnitude. Various refinements of the method and alternative methods of calculation have been developed (e.g. Passioura, 1977; Valiantzas and Kerkides, 1990; Etching and Hopmans, 1993). Multistep techniques have also been developed in recent years (van Dam *et al.*, 1994). These resemble the one-step outflow method, but are conducted through a series of small pressure-change steps instead of one large one.

Determination of saturated K can be important in the unsaturated zone because frequently there are localized zones of saturation and also because this property is sometimes used in estimating other hydraulic properties. A straightforward method uses the flow rate measured with a constant applied head gradient, much like the analogous method for unsaturated K. It is also possible to supply water to a soil column from a source with a level that falls as the water is used up, the falling-head method for saturated K. A falling head, where the water level can be related to the cumulative flux, is more useful for saturated than for unsaturated properties because the water content does not change with the applied pressure. Saturated K is easier to measure than unsaturated K but it has unique difficulties. One is the fact that gaps between a soil sample and a retainer wall will fill with water, possibly becoming rapid-flow channels that lead to an indication of K much greater than the actual K of the soil. This is not usually a problem for unsaturated measurements because large gaps desaturate when the matric pressure is even slightly less than zero. Another problem in a saturated K measurement is that because of air trapping the soil is not really saturated and its water content may be poorly determined. For these and other reasons, measured values of saturated K frequently have poor reliability and may have little relation to observed field phenomena that would seem to occur at saturation.

To determine permeability for a gas, a straightforward flow-measuring apparatus can be used with soil samples in a laboratory for accurate results but with considerable effort (Stonestrom and Rubin, 1989a). Field methods (Weeks, 1978; Baehr and Hult, 1991; Shan, 1995) normally rely on measurements of pressure variations at points within the medium in response to pressure changes imposed elsewhere in the medium.

Estimation of K

Property transfer models can be useful for estimating K, as for water retention. Usually, these use water retention, not particle-size distribution, as the more easily measured type of data from which unsaturated K is calculated. If a transfer from particle size to K is needed, such a model may be combined with a water retention property transfer model like that of Arya and Paris (1981), though reliability is likely to be markedly reduced because the particle-size distribution is less directly related to K.

Capillary theory provides an interpretation of the pores in the medium that relates to both K and retention. Purcell (1949) was the first to use it to quantitatively relate these two properties, using the assumption that pores are equivalent to a bundle of capillary tubes of different sizes, with a particular size distribution for a given medium. The drying retention curve relates to the effective radius at which pores empty and the capillary formula (2) quantifies this radius in terms of ψ. Because larger pores empty first, for a given ψ, pores smaller than the radius corresponding to ψ are considered filled. Poiseuille's law gives the effective conductance of a filled capillary, and an integration of such conductances for all filled pores leads to an estimate of K for that ψ value (and its corresponding θ). In effect, this is a summation of contributions from

each size of filled pore, weighted by the abundance of that size, to give a number assumed proportional to K. Multiplication by a matching factor, quantified using a single known value of K, gives the $K(\theta)$ or $K(\psi)$ curve. In practice, the matching factor is often based on saturated K, though in general this is a poor choice because saturated K depends primarily on the very largest pores of the medium, which have little relevance to $K(\theta)$ over most of the θ or ψ range (Nimmo and Akstin, 1988). Choices of how the model handles issues such as pore length, connectedness, and tortuosity lead to different versions of this sort of model (e.g. Childs and Collis-George, 1950; Burdine, 1953). The version of Mualem (1976) has become widely used because it is mathematically easy to work with and gives estimates as good as or better than others. Mualem and Dagan (1978) and Hoffmann-Riem *et al.*, (1999) have developed schemes for generalizing this class of models. More recent developments include the incorporation of angular pore geometry (Tuller *et al.*, 1999), which allows a pore of given size to form a continuum of sizes of effective water flow conduit, rather than restricting it to the circular-capillary states of complete fullness or complete emptiness. Network and percolation-theory models have also been developed but have not come into widespread use.

As in the case of water retention, completely empirical formulas can represent unsaturated K. Gardner (1958), for example, used a formula of the form

$$K(\theta) = A\exp(-\alpha\theta) \tag{12}$$

where A and α are fitted empirical parameters. Such formulas have greater simplicity and sometimes lead to more realistic curve shapes than formulas developed for combined representation of K and water retention as described below. The α parameter in (12) is widely used in developing and applying other models, such as analytical solutions of Richards' equation.

Combined Estimation of K and Water Retention

A direct combination of an empirical formula for water retention into a capillary theory formulation of unsaturated K can yield a convenient analytical formula for $K(\theta)$, and facilitate the combined treatment of water retention and unsaturated K. For example, the Brooks and Corey (1964), van Genuchten (1980), or lognormal (Kosugi, 1999) model can be inserted into the Burdine (1953) or Mualem (1976) model. The combination of the van Genuchten and Mualem models results in the formula

$$K(\psi) = K_{\mathrm{m}}\frac{\left\{1-\left(\frac{\psi}{\psi_c}\right)^{\nu-1}\left[1+\left(\frac{\psi}{\psi_c}\right)^{\nu}\right]^{-\mu}\right\}^2}{\left[1+\left(\frac{\psi}{\psi_c}\right)^{\nu}\right]^{\mu/2}} \tag{13}$$

where K_{m} is the matching factor computed from a known (ψ, K) point. Leij *et al.* (1997) give a tabulation and evaluation of formulas that result from such combinations. Use of such formulas is central to pedotransfer-function models for coordinated estimation of water retention and K. As in the water-retention application of pedotransfer functions described above, these provide a particularly easy way to estimate the hydraulic properties needed for a Richards-equation analysis, though likewise with problems of poor and unknown reliability.

Translation of parameter values from one model to another, that is, for a given medium finding the values for one model's parameters directly from the parameter values previously estimated for a different model (Morel-Seytoux *et al.*, 1996; Morel-Seytoux and Nimmo, 1999), is a convenience where the retention curve is known only in terms of one particular parameterization, and an application requires another. In order to give a unique conversion, a specific equivalence criterion must be chosen. This can, for example, be the invariance of a relevant property like capillary drive (a single-number combination of retention and conductive properties related especially to infiltration rates).

Inverse approaches can yield hydraulic property values in a variety of situations and experiments, as explained in detail in **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**. The basic idea is that a forward model (e g. Richards' Equation), which can compute output conditions (θ or ψ as a function of space and time) using given properties (water retention and unsaturated K), can be used in an algorithm that does the inverse of this. That is, given a set of measured output conditions, the algorithm determines values of the properties that would most closely simulate the measured conditions. Those values then have effectively been determined, and can be used in the forward model. Conditions such as water content as a function of space and time frequently are easier to measure than the properties of the medium. One way of inverse modeling is to use a forward model with manual or automated trial-and-error: guess the property values, compute outputs, compare with observation, revise estimated property values, and try again. Many particular field or lab situations, for example, the one-step or multistep outflow method, are suitable for unsaturated hydraulic property estimation by inverse approaches (Hopmans *et al.*, 2002). Frequently, the transient ψ values from one or more tensiometers installed in the sample are used as data in addition to the outflow rate as a function of time.

Scaling provides another way of determining unsaturated properties from limited data, relating them from place to place or from one medium to another (Miller, 1980). In a wet medium, surface tension is the main phenomenon associated with retention, and viscosity with conductivity. These are drawn together in the surface-tension viscous-flow

(STVF) similitude model of Miller and Miller, (1956). One way for two media to be similar is for one of them to be geometrically a direct magnification of the other by a scale factor λ. For natural media, the point-to-point correspondence is not exact; the particles of Miller-similar media can be repositioned, and they do not have to be exact scale replicates, as long as they are statistically equivalent for the purpose at hand. If two media are Miller-similar, if surface tension controls water retention, and if viscous flow laws determine K for a given geometrical configuration, the unsaturated hydraulic properties of one medium can be directly determined from those of the other using the known value of λ. The value of ψ in one medium differs by a factor of λ from that in the other. Water content, being a ratio of volumes, is the same in each. Thus, the retention curve $\theta(\psi)$ of one similar medium equals $\theta(\lambda\psi)$ of the other. Because of the radius-squared factor for conductivity in Poiseuille's law, $K(\theta)$ for similar media is different by a factor of λ^2. Warrick *et al.*, (1977) relaxed some of the Miller–Miller criteria, opening up new field applications by showing the relaxed criteria to be useful for relating properties from one sample to another. Porosity, for example, does not have to be identical for Warrick-similar soils. Whenever a set of media can be considered similar with known values of λ, knowledge of hydraulic properties for one of the media permits calculation of the properties of all. Recent applications of this sort include those of Rockhold *et al.*, (1996) and Nimmo *et al.*, (2002a).

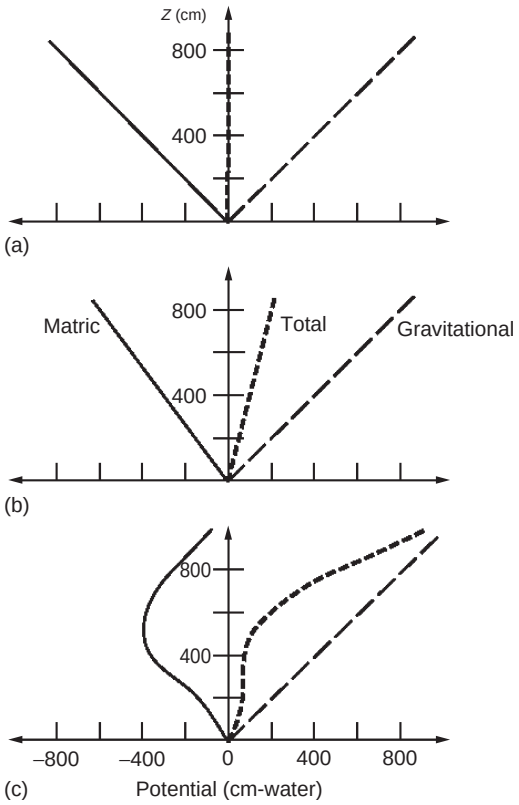


Figure 14 Profiles of matric, gravitational, and total potential for idealized situations of (a) static water, (b) steady downward flow, and (c) unsteady flow. The water table is at $z = 0$

APPLICATIONS TO DISTRIBUTIONS AND FLUXES

This article emphasizes the presence and movement of water within the unsaturated zone. **Chapter 146, Aquifer Recharge, Volume 4** considers the fluxes from the unsaturated zone into the saturated zone. **Chapter 152, Modeling Solute Transport Phenomena, Volume 4** and **Chapter 153, Groundwater Pollution and Remediation, Volume 4** consider the fluxes of other substances, for the evaluation of which the flux of water is usually the first consideration.

Distributions of Water and Energy

The distribution of water with depth at a given time depends on the energy state (based on components including matric and gravitational potential), wetting/drying history, and dynamics of the water itself. If there is no flow, one can infer that the gradient of total potential is zero, so if the matric and gravitational components are the only significant ones, they add to a constant total potential. Figure 14(a) shows this type of hydrostatic profile for the case where a water table is present. Since the matric pressure in this case is linear with depth and the water content is controlled by the water retention properties of the medium, for a uniform unsaturated zone, the water content profile (not shown in Figure 14) then mimics the shape of the water retention curve. The previous history of the unsaturated zone would dictate whether a wetting or drying retention curve would apply. In general, there is significant trapped air when the medium is close to saturation; thus, it frequently is appropriate to consider a repeat-cycle curve rather than a first drying curve. Given time, approximate versions of such a hydrostatic profile may develop in portions of a profile where water movement is negligible. If water flows vertically downward at a steady rate in a homogeneous medium, the total gradient must be constant, but the matric pressure does not cancel out the gravitational potential, as illustrated in Figure 14(b). At many locations this situation probably never develops, but some locations, especially where the unsaturated zone is thick, may commonly have constant potential or an approximation to it. Darcy's law applies directly in field situations where flow is steady, though steadiness in nature is often temporary or approximate. In the general case of unsteady flow, the matric pressure profile cannot be determined so simply, and may take on an irregular form like the example in Figure 14(c).

The uppermost part of the water distribution profile is sometimes described in relation to field capacity, defined as the water content of a soil profile when the rate of downward flow has become negligible two or three days after a major infiltration (SSSA, 1997). An essential assumption for use of the field capacity concept is that soil, when wetted to a high water content, loses water by drainage at a declining rate that does become negligible. This concept is used in agriculture to indicate the wettest soil conditions that need to be considered for plant growth, and sometimes is mentioned in hydrologic investigations, for example, as related to soil moisture storage. Elements of the definition of field capacity are imprecise and require subjective judgment, for example, in deciding what is negligible. Field capacity is implicitly associated with the entire soil profile through the root zone, including preferential flow characteristics and, especially, flow-retarding layers that enable layers above them to retain a high water content. Thus, it is not appropriate to consider the field capacity of an individual soil sample or a point within the unsaturated zone. Because it is based on distinct intervals of major infiltration and major drainage, the concept of field capacity is most applicable in field plots that are periodically wetted to a high degree, as by annual snowmelt, flood, or monsoons; especially in applications where only approximate quantification or rules of thumb are required. This concept does not work well in rocky or fracture-dominated media that retain significant water only briefly, on land where water application is consistently erratic, or in applications requiring exactness, repeatability, or validity at multiple scales of observation.

In a portion of the unsaturated zone immediately above the water table, it may happen that all pores are filled with water, held by capillary forces. The depth interval that is saturated but above the water table is called a *capillary fringe*. In a hydrostatic profile, this corresponds to a flat portion of the retention curve between saturation and an air-entry pressure. As in other cases of saturation within the unsaturated zone, there is likely to be significant trapped air, though not an air phase that is continuous through the medium. Some media will not have a significant capillary fringe because their retention characteristics have the air-entry pressure at essentially zero. These media, including materials with significant fractures or other macropores, are more common than is apparent from databases and literature surveys of measured water retention, because, historically, most measurements have been done on samples that have been repacked from mechanically disturbed material. Repacking often destroys the large pores that would otherwise lead to a zero air-entry pressure. Caution must also be invoked in applying the capillary fringe concept where the water table fluctuates. The hydrostatic equilibrium required for a capillary fringe may take considerable time to establish, given typically small values of unsaturated K. Furthermore, soil-water hysteresis would make for a different capillary fringe with a falling than with a rising water table.

Fluxes at the Land Surface

Input

Infiltration is the downward movement of water through the land surface. Because osmotic, thermal, and other modes of driving water flow are usually negligible, water is driven into the soil mainly by gravity and ψ gradients. If the soil is dry near the surface the ψ gradients usually dominate gravity. When the soil is very wet to some depth, gravity may become the major driving force. The usual case is that water infiltrates faster at the start and slows down as a zone of increased water content develops at the surface and grows. Figure 15 shows actual infiltration rates i [$\mathrm{L\,T^{-1}}$] varying over time in four soil columns. If water at the surface is abundantly available but not under significant pressure, infiltration occurs at the infiltration capacity, a rate determined only by the soil, not by the rate of application or

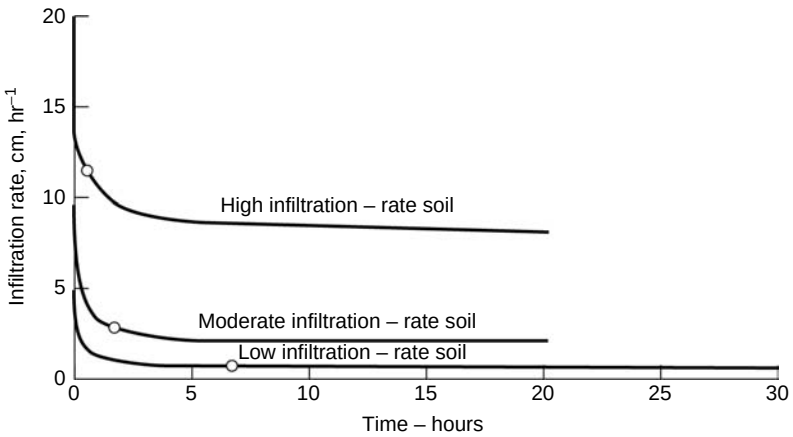


Figure 15 Measured infiltration rates over time for three soil columns

other factors. If water arrives at the surface faster than the infiltration capacity, excess water ponds or runs off. Like hydraulic conductivity, infiltration capacity is not single-valued for a given medium but varies with water content and other conditions. If water arrives at a rate less than the infiltration capacity, the water may immediately infiltrate. The infiltration rate then is determined by the rate of application. This article addresses the issue of what happens when the soil limits the infiltration rate, requiring attention to the infiltration capacity not only as the key element for quantifying i under that condition but also as the criterion that determines whether that condition exists. Conditions that complicate the ideal conception of infiltration include: variation of application rate with time, spatial variability of soil and surface properties, water repellency of the soil, air trapping, and variations of temperature.

Richards' equation (5) can be applied to infiltration processes. But the phenomena involved, especially with water contents, pressures, and fluxes sharply changing in time and space, make this difficult or inappropriate. Most studies use instead a formulation specifically tailored for infiltration, generally with more specific assumptions. Some of these formulations are based on a conceptualization of the physics of infiltration while others are more purely empirical.

The Green and Ampt model of infiltration assumes uniform properties of the medium, a perfectly sharp wetting front at depth L, a constant matric pressure ψ_o at the land surface, a uniform initial water content θ_i, a uniform water content θ_w behind the wetting front, and a value of matric pressure ψ_{wf} at the wetting front that remains fixed even as the front moves downward. Gravity may be included, or, in a simplified form of the model, assumed negligible. The mathematical formulation of the model comes from expressing the infiltration rate in two different ways – by Darcy's law and by the computed change in water storage – and setting them equal to each other. This leads to the depth of the wetting front being

$$L(t) = \left(2K\frac{\Delta\psi}{\Delta\theta}t\right)^{1/2} \tag{14}$$

and the infiltration rate

$$i = \Delta\theta\frac{\mathrm{d}L}{\mathrm{d}t} = \Delta\theta\left(\frac{K\dfrac{\Delta\psi}{\Delta\theta}}{2t}\right)^{1/2} \tag{15}$$

where $\Delta\theta = \theta_w - \theta_i$ and $\Delta\psi = \psi_o - \psi_{wf}$. Analogous formulas can be derived for the case where gravity is not negligible. The inverse proportionality of i to the square root of time appears frequently in infiltration formulas. The natural decline of the infiltration rate as the soil gets wetter correlates in the model with the declining driving force as the constant $\Delta\psi$ is stretched over an increasing depth interval. Mathematically, this works out to square-root-of-time behavior. The Green–Ampt model gives a direct quantification of infiltration with simple formulas for the basic character of infiltration under common circumstances. But, of course, it does not give specific details of the infiltration process. Generally, one must infer a value of ψ_{wf} from other known facts concerning the infiltration; the fact that ψ_{wf} cannot be measured severely limits the use of this model in a predictive manner. The Green–Ampt assumptions are most likely to be satisfied in uniform, coarse-textured soil with simple structure.

Philip (1957) presented mathematical solutions to the problem of infiltration under fewer restrictions than required by Green and Ampt. Philip assumed that the medium properties and the initial water content are uniform, and that the water content at the land surface is held at a fixed water content greater than the initial water content. Starting with a form of Richards' equation, he derived

$$i = \frac{1}{2}\mathrm{S}t^{-1/2} \tag{16}$$

where S is a soil hydraulic property called the *sorptivity* (dimensions $\mathrm{L\,T^{-1/2}}$), which indicates the relative tendency for the medium to absorb infiltrating water. For the case where gravity is significant, Philip developed a series solution that can be approximated for short times by the truncation

$$i = \frac{1}{2}\mathrm{S}t^{-1/2} + A \tag{17}$$

where A is an empirical parameter whose value approaches saturated K after a long time. Philip's model has been tested against measurements, notably by Davidson *et al.* (1963), who found good agreement with the model in terms of both preservation of the shape of the wetted profile, and of the rate of its downward movement. Yet, it would be exactly true only for certain unlikely restrictions on the unsaturated hydraulic conductivity and the ponding of infiltrating water. Furthermore, most cases of infiltration are in heterogeneous soils with nonuniform water content and variable water application, in which the Philip model would give approximate results at best.

Empirical formulas for infiltration have been used where the relevant soil hydraulic properties are not necessarily known. Gardner and Widtsoe (1921) used an exponential formula expressed by Horton (1940) as

$$i = i_c + (i_o - i_c)\exp(-\alpha t) \tag{18}$$

where i_o is the initial infiltration rate, i_c is the rate of infiltration after it becomes steady, and α is an empirical constant that depends on the soil. Holtan (1961) proposed an empirical formula useful for the early stages of infiltration, based on a power law function of the effective yet-unused water-holding capacity of the upper part of the soil profile:

$$i = i_c + a\left[ML_l - \int_0^t i\,dt\right]^n \qquad (19)$$

where i_c, a, and n are constants, and M is the air-filled soil porosity down to a specified depth L_l (ideally to a known impeding layer).

Output

The term exfiltration is occasionally used to indicate water fluxes out of the soil at the land surface, though mostly these processes are discussed in terms of evapotranspiration. As a whole, evapotranspiration is usually treated as a surface or micrometeorological problem, though certain critical unsaturated-zone processes are involved, especially for the evaporation, or bare-soil, component.

When the soil is wet enough, plenty of water is available at the surface so that atmospheric conditions control the evaporation rate. When the soil is too dry to supply water at the maximum rate the atmosphere will absorb it, the soil properties will control the evaporation rate. Thus, there are at least two cases to consider: the atmosphere-dominated "constant-rate" phase during which the transport mechanisms of the soil are ignored, and the soil-dominated "declining-rate" phase during which atmospheric effects are ignored. If the second phase persists long enough, the soil may become so dry that water transport within it is primarily in the form of vapor rather than liquid. This defines a third phase, in which the evaporation rate is slow and may be nearly constant. The initial atmosphere-dominated phase typically lasts a few hours to a few days after irrigation or rainfall. As with infiltration, there is a moisture state criterion based on local conditions and properties that separates atmosphere-dominated and soil-dominated flow. Gardner and Hillel (1962) have quantitatively treated the question of when the atmosphere-dominated phase ends. The overall reduction of evaporation rate with time can be considered in terms of self-mulching; after a bare-surface soil has dried to a certain extent, a layer with very small θ forms at the surface. The low hydraulic conductivity of this layer limits the rate of flow of water from the soil below it to the land surface, thus limiting evaporation (Buckingham, 1907).

During the soil-dominated phase, to roughly estimate the average evaporation rate, the problem can be approached with Richards' equation, assuming that the system is isothermal, vapor does not flow, gravity is negligible, diffusivity is constant, and the soil is a semi-infinite slab uniformly wet at the beginning. The problem then falls into the same class as the gravity-free infiltration problem. The solution for evaporative flux density goes as the inverse square root of time, just as in typical assessments of infiltration rate.

Water resource applications often require values of total evapotranspiration, including treatment of atmospheric conditions in addition to the unsaturated-flow phenomena considered here. Infiltration minus evapotranspiration is often taken as the net input of water to the subsurface. A second major hydrologic application is the evaporation of water from the saturated zone. Capillary forces can draw water up from the water table to depths from which it supplies the evaporative process. This can be a substantial loss mechanism from a water-table aquifer, especially where the unsaturated zone is thin. Various quantitative treatments of these effects have been developed, for example, by Ripple *et al.* (1972).

Fluxes Within the Unsaturated Zone

Redistribution

After water has infiltrated, it redistributes, driven by gravity, matric pressure gradients, and possibly other forces. Figure 16 illustrates water content distributions at various times during and after infiltration, in a mechanically disturbed soil and in a soil with intact natural structure. Redistribution continues until conditions are such that all forces balance out. Equivalently, the water may be considered to progress toward a state of minimal (and uniform) total energy of the earth–water–air system, that is, equilibrium.

In an idealized case of homogeneous soil and perfectly sharp wetting fronts, the redistributing water content profile may take the shape of an elongating rectangle. If there are no losses to evaporation or anything else, the rectangle will maintain the same area as it evolves. This picture of redistribution may be approximated to various degrees in real media, for example, as in Figure 16(d). Many factors cause deviations from the ideal, including layers that retard the flow, and preferential flow and lateral heterogeneity that advance the downward movement of a portion of the water, thereby blunting the rectangular sharpness of the wetting front. Greater K above than below the wetting front may cause water to accumulate in the lower portions of the wetted zone, distorting the rectangle laterally.

Normally, hysteresis strongly influences redistribution because the wetting front progresses downward according to the wetting curves of water retention and conductivity, whereas θ in the upper portions of the wetted zone decreases according to the drying curves. Because a drying retention curve has greater θ for a given ψ, water contents remain higher in the upper portions than they would if there were no hysteresis. Thus, one important consequence of hysteresis is to hold more water near the land surface where it is accessible to plants.

After unsteady infiltration, a pattern of variations in water content with depth becomes established as alternations of surface wetness and surface dryness move downward by gravity. Matric pressure gradients move water from wet to dry, both upward and downward, so as the pattern of θ variation moves downward, wet zones become drier and dry zones wetter. Thus, variations are damped out with depth.

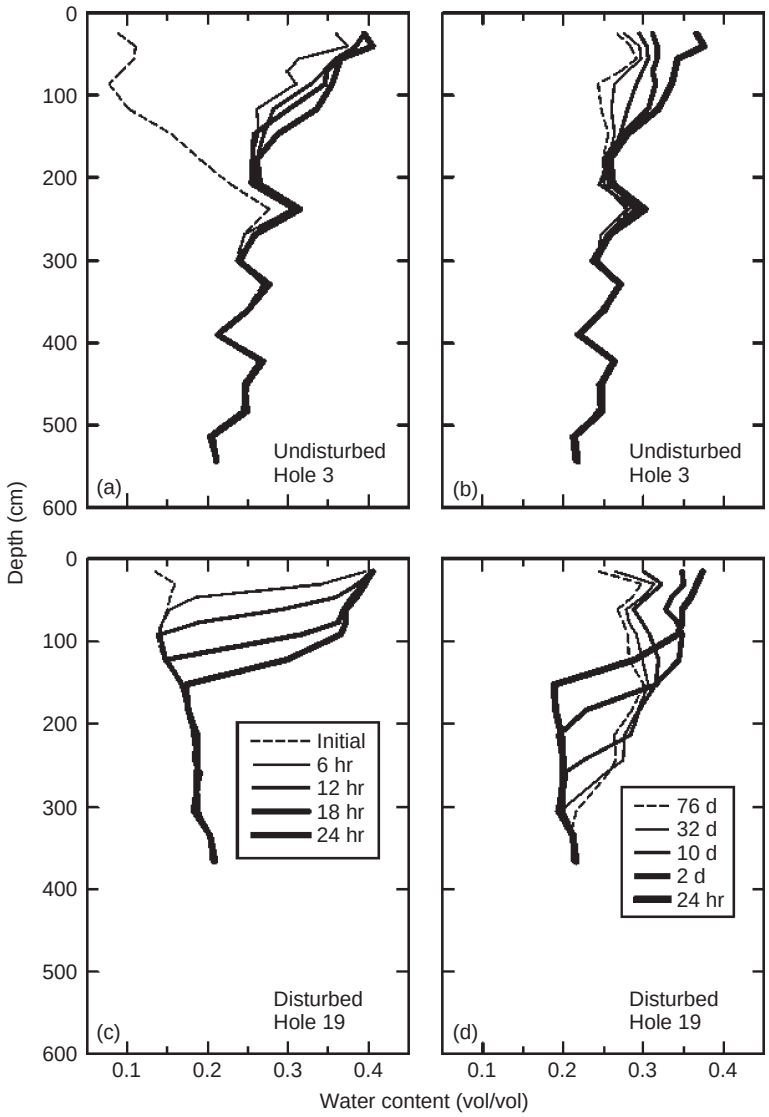


Figure 16 Measured water distributions during and after 24 h of flood infiltration in (c,d) a disturbed, relatively homogeneous soil in a landfill and (a,b) an adjacent undisturbed soil, on the snake river plain in Idaho (Nimmo *et al.,* 1999). Evaporation was inhibited by an impermeable cover at the land surface

Gravity takes on greater relative importance as the variations decrease. In a thick unsaturated zone, a region may develop in which gravity is the only significant driving force, as illustrated in Figure 2 of **Chapter 146, Aquifer Recharge, Volume 4**. Approximating the periodic infiltration of water with a sinusoidally varying flux, Gardner (1964) estimated the depth at which θ variations would become negligible. For a periodicity of one year and typical values of D, the profile would become damped within a few meters of the surface. An important application of this is in Darcian methods for estimating aquifer recharge rates as discussed in **Chapter 146, Aquifer Recharge, Volume 4**.

Usually the above considerations would need to be adjusted or reinterpreted with attention to preferential flow. There is not yet a widely accepted quantitative theory for this. Qualitatively, a major effect of preferential flow is to permit more rapid movement of water to significant depths. This would occur primarily under very wet conditions, and would be followed in the redistribution process by a slower flow of water into the regions between preferential flow channels.

Effects of Layering

Layers that contrast in hydraulic properties impede vertical flow by various mechanisms. (i) When water moves down from a coarse to a fine layer, as from coarse sand to silt, if both layers are near saturation, the fine layer has smaller hydraulic conductivity; therefore, flow slows when it reaches the fine layer. If, however, the coarse layer is

nearly saturated but the fine layer is initially fairly dry, at first the flow may be temporarily accelerated while the flow is dominated by the sorptive nature of the fine medium, which tends to suck water out of the coarse material. (ii) Where a fine layer overlies a coarse layer, water moving downward is impeded under many conditions. When coarse material is dry, it has an extremely small hydraulic conductivity; thus it tends not to admit water into the pores and exhibits a somewhat self-perpetuating resistance to flow. Water breaks into the coarse layer if the pressure at the layer contact builds to the point that the water-entry pressure (the minimum water pressure needed to fill an empty pore) of some of the large pores is exceeded. This can generate flow instabilities, as discussed previously. Stable or not, water flow into the pores of the coarse medium increases that medium's hydraulic conductivity. With equal ψ values across the layer boundary, which is necessary for water not to accumulate at the boundary, unsaturated K of the coarse layer is often less than that of the fine layer. Stable or diffuse flow through layers where fine overlies coarse is slower than it would be if both layers had the properties of the fine medium. Miller and Gardner (1962) demonstrated this effect experimentally.

Layer thickness can influence the long-term tendency of downward flow in other ways. An important case is where a granular medium overlies fractured bedrock. If the fractures are not microscopically narrow and the rock is otherwise impermeable, the bedrock admits water only under nearly saturated conditions. The thinner the layer of granular material, the more easily it becomes saturated down to the depth of the bedrock and hence the more frequently it generates deep percolation. This more frequent saturation at the layer boundary increases the fraction of average precipitation that flows into the bedrock and possibly into the aquifer as recharge.

A common phenomenon in layered media is the accumulation of water in a region of the unsaturated zone to the point where it becomes saturated, even though there is unsaturated material between that region and the saturated zone. Because this phenomenon usually results from an impeding layer on which excess water is perched, it is called *perching*. It may be a temporary or a nearly permanent feature, depending on the nature of the medium, the prevailing hydrologic conditions, and the effect of artificial modifications. The high water content of a perched zone causes greater hydraulic conductivity and potentially faster transport through the three-dimensional system. The main effect is not a direct increase in vertical flow because the increase in effective vertical hydraulic conductivity is offset by a diminished vertical hydraulic gradient within the perched water. (Vertical flow within and below the perched water cannot be faster than the vertical flow above the perched water or the perched water would have drained.) Horizontally, however, there may be greatly increased flow (e.g. Nimmo *et al.*, 2002c). New and different processes may significantly affect contaminant transport in a perched zone. Reduced aeration, for example, may affect biochemical processes. At the scale of the entire stratified vadose zone, perching may significantly increase anisotropy. For considerable horizontal distances the hydraulic conductivity might be as great as $1\,\mathrm{cm\,s^{-1}}$ or more, as for a saturated gravel. At the same time, however, vertical flow might be limited by an unsaturated layer having vertical hydraulic conductivity 10 or more orders of magnitude less, as for rock with unfilled fractures.

Layering strongly influences the direction of flow within the unsaturated zone. The flow is often assumed to be predominantly vertical because (i) with a continuous air phase in the pores, buoyancy does not counteract gravity as it does in the saturated zone, and (ii) whatever the effect of other forces, gravity acts vertically. Some vertical flow continues even where impeding layers cause substantial perching and horizontal diversion. Horizontal flow can be especially important in sloping or laterally heterogeneous media. Horizontal flow under nearly saturated conditions, called *interflow*, may be substantial near the land surface during storms. In thick unsaturated zones where adjacent layers contrast sharply, horizontal flow can have great importance even in arid climates.

CONCLUSION

Unsaturated flow is complicated by the significance of multiple phases. At least three drastically different substances – water, air, and solid mineral – are critical to its nature and quantification. Unsaturated flow phenomena are extremely sensitive to the proportions of those phases, especially the fluid phases, as natural variations in the proportion of water and air can cause a property like hydraulic conductivity to vary over many orders of magnitude.

Many of the features that make saturated zone hydrology difficult – for example, opacity of the subsurface, heterogeneity, complex geometry on small and large scales – are all the more influential when properties vary so drastically, for example, K at a given point in space varying by orders of magnitude. As a result of these complications, unsaturated zone hydrology is subject to a greater degree of inexactness than most fields of quantitative physical science. Often, even factor-of-10 accuracy is difficult to achieve. This means that in many presentations of analyses and model predictions, the precision of the mathematical results goes far beyond what is justified by sparseness or poor quality of the data that support those results, and also beyond what is justified by applicability of the models and concepts. Evaluation of uncertainty depends much on hard-to-quantify conceptual uncertainty.

Given only 100 years since quantitative physical theory was first applied to unsaturated flow, understanding has

developed considerably. In the early twenty-first century, it is still developing rapidly and maturing. Further advancement in response to scientific and societal needs requires new measurement techniques to obtain more and higher quality data, and new theoretical constructs that more adequately represent the important physical processes within practical modeling schemes.

REFERENCES

Andraski B.J. (1996) Properties and variability of soil and trench fill at an arid waste-burial site. *Soil Science Society of America Journal*, **60**, 54–66.

Andraski B.J. and Scanlon B.R. (2002) Thermocouple psychrometry. In *Methods of Soil Analysis, Part 4 – Physical Methods*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 609–642.

Arya L.M. and Paris J.F. (1981) A physicoempirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data. *Soil Science Society of America Journal*, **45**, 1023–1030.

Baehr A.L. and Hult M.F. (1991) Evaluation of unsaturated zone air permeability through pneumatic tests. *Water Resources Research*, **27**, 2605–2617.

Bear J. (1979) *Hydraulics of Ground Water*, McGraw-Hill: New York.

Brooks R.H. and Corey A.T. (1964) *Hydraulic Properties of Porous Media*, *Colorado State University Hydrology*, p. 27.

Buckingham E. (1907) *Studies on the Movement of Soil Moisture*, USDA Bureau of Soils: Washington, Bulletin 38.

Burdine N.T. (1953) Relative permeability calculations from pore size distribution data. *Transactions AIME*, **198**, 71–77.

Childs E.C. and Collis-George N. (1950) The permeability of porous materials. *Proceedings Royal Society London, Series A*, **201**, 392–405.

Daily W., Ramirez A., LaBrecque D. and Nitao J. (1992) Electrical resistivity tomography of vadose water movement. *Water Resources Research*, **28**, 1429–1442.

Davidson J.M., Biggar J.W. and Nielsen D.R. (1963) Gamma-radiation attenuation for measuring bulk density and transient water flow in porous materials. *Journal of Geophysical Research*, **68**, 4777–4783.

DeNovio N.M., Saiers J.E. and Ryan J.N. (2004) Colloid movement in unsaturated porous media: recent advances and future directions. *Vadose Zone Journal*, **3**, 338–351.

DiCarlo D.A. (2004) Experimental measurements of saturation overshoot on infiltration. *Water Resources Research*, **40**, W04215, doi:10.1029/2003WR002670.

Dillard L.A., Essaid H.I. and Herkelrath W.N. (1997) Multiphase flow modeling of a crude-oil spill site with a bimodal permeability distribution. *Water Resources Research*, **33**, 1617–1632, 1997–069078.

Doering E.J. (1965) Soil-water diffusivity by the one-step method. *Soil Science*, **31**, 2491–2495.

Durner W. (1994) Hydraulic conductivity estimation for soils with heterogeneous pore structure. *Water Resources Research*, **30**, 211–223.

Eliassi M. and Glass R.J. (2002) On the porous-continuum modeling of gravity-driven fingers in unsaturated materials – extension of standard theory with a hold-back-pile-up effect. *Water Resources Research*, **38**, 1234 doi:10.1029/2001WR001131.

Eppstein M.J. and Dougherty D.E. (1998) Efficient three-dimensional data inversion; soil characterization and moisture monitoring from cross-well ground-penetrating radar at a vermont test site. *Water Resources Research*, **34**, 1889–1900.

Essaid H.I., Herkelrath W.N. and Hess K.M. (1993) Simulation of fluid distributions observed at a crude oil spill site incorporating hysteresis, oil entrapment, and spatial variability of hydraulic properties. *Water Resources Research*, **29**, 1753–1770 1995–002840.

Etching S.O. and Hopmans J.W. (1993) Optimization of hydraulic functions from transient outflow and soil water pressure data. *Soil Science Society of America Journal*, **57**, 1167–1175.

Gardner W.R. (1956) Calculation of capillary conductivity by the one-step method. *Soil Science Society of America Proceedings*, **20**, 317–320.

Gardner W.R. (1958) Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science*, **85**, 228–232.

Gardner W.R. (1964) Water movement below the root zone. *Transactions, 8th International Congress of Soil Science*, International Society of Soil Science: Bucharest, pp. 63–68.

Gardner W.R. and Hillel D.I. (1962) The relation of external evaporative conditions to the drying of soils. *Journal of Geophysical Research*, **67**, 4319–4325.

Gardner W. and Widtsoe J.A. (1921) The movement of soil moisture. *Soil Science*, **11**, 215–232.

Garnier P., Rieu M., Boivin P., Vauclin M. and Baveye P. (1997) Determining the hydraulic properties of a swelling soil from a transient evaporation experiment. *Soil Science Society of America Journal*, **61**, 1555–1563.

Gee G.W., Campbell M.D., Campbell G.S. and Campbell J.H. (1992) Rapid measurement of low soil water potentials using a water activity meter. *Soil Science Society of America Journal*, **56**, 1068–1070.

Glass R.J., Steenhuis T.S. and Parlange J.-Y. (1989) Mechanism for finger persistence in homogeneous unsaturated, porous media: theory and verification. *Soil Science*, **148**, 60–70.

Grant S.A. and Salehzadeh A. (1996) Calculation of temperature effects on wetting coefficients of porous solids and their capillary pressure functions. *Water Resources Research*, **32**, 261–270, 1996–067403.

Griffioen J.W., Barry D.A. and Parlange J.Y. (1998) Interpretation of two-region model parameters. *Water Resources Research*, **34**, 373–384, 1998–033167.

Hamilton J.M., Daniel D.E. and Olson R.E. (1981) Measurement of hydraulic conductivity of partially saturated soils. In *Permeability and groundwater contaminant transport*, Zimmie T.F. and Riggs C.O. (Eds.), ASTM Special Technical Publication No. 746, American Society for Testing and Materials: pp. 182–196.

Harvey R.W. and Garabedian S.P. (1991) Use of colloid filtration theory in modeling movement of bacteria through a contaminated sandy aquifer. *Environmental Science and Technology*, **25**, 178–185, 1991–019144.

Hess K.M., Herkelrath W.N. and Essaid H.I. (1992) Determination of subsurface fluid contents at a crude-oil spill site. *Journal of Contaminant Hydrology*, **10**, 75–96.

Hillel D. and Gardner W.R. (1970) Transient infiltration into crust-topped profiles. *Soil Science*, **109**, 69–76.

Hoffmann-Riem, H., van Genuchten, M.T. and Fluehler, H. (1999) General model of the hydraulic conductivity of unsaturated soils. *Proceedings of the International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, pp. 31–42.

Holtan H.N. (1961) *A Concept for Infiltration Estimates in Watershed Engineering*, USDA ARS.

Hopmans J.W., Simunek J., Romano N. and Durner W. (2002) Inverse methods [for simultaneous determination of water transmission and retention properties]. In *Methods of Soil Analysis, Part 4 – Physical Methods*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 963–1008.

Horton R.E. (1940) An approach toward a physical interpretation of infiltration capacity. *Soil Science Society of America Proceedings*, **5**, 399–417.

Jury W.A., Wang Z. and Tuli A. (2003) A conceptual model of unstable flow in unsaturated soil during redistribution. *Vadose Zone Journal*, **2**, 61–67.

Kosugi K. (1994) Three-parameter lognormal distribution model for soil water retention. *Water Resources Research*, **30**, 891–901.

Kosugi K. (1999) General model for unsaturated hydraulic conductivity for soils with lognormal pore-size distribution. *Soil Science Society of America Journal*, **63**, 270–277.

Kung K.J.S. (1990) Preferential flow in a sandy vadose zone -1. Field observation. *Geoderma*, **46**, 51–58 1990–049956.

Lappala, E.G., Healy, R.W. and Weeks, E.P. (1987) *Documentation of Computer Program VS2D to Solve the Equations of Fluid Flow in Variably Saturated Porous Media*, U.S. Geological Survey Water Resources Investigations Report 83–4099, U.S. Geological Survey Water Resources Investigations, Denver.

Leij F.J., Russell W.B. and Lesch S.M. (1997) Closed-form expressions for water retention and conductivity data. *Ground Water*, **35**, 848–858.

Luxmoore R.J. (1981) Micro-,meso-, and macroporosity of soil. *Soil Science Society of America Journal*, **45**, 671–672.

Miller E.E. (1980) Similitude and scaling of soil-water phenomena. In *Applications of Soil Physics*, Hillel D. (Ed.), Academic Press: New York, pp. 300–318.

Miller E.E. and Miller R.D. (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics*, **27**, 324–332.

Miller E.E. and Salehzadeh A. (1993) Stripper for bubble-free tensiometry. *Soil Science Society of America Journal*, **57**, 1470–1473.

Miller D.E. and Gardner W.H. (1962) Water infiltration into stratified soil. *Soil Science Society of America Proceedings*, **26**, 115–119.

Morel-Seytoux H.J., Meyer P.D., Nachabe M., Touma J., Van Genuchten M.T. and Lenhard R.J. (1996) Parameter equivalence for the Brooks-Corey and van Genuchten soil characteristics: preserving the effective capillary drive. *Water Resources Research*, **32**, 1251–1258.

Morel-Seytoux H.J. and Nimmo J.R. (1999) Soil water retention and maximum capillary drive from saturation to oven dryness. *Water Resources Research*, **35**, 2031–2041.

Mualem Y. (1974) A conceptual model of hysteresis. *Water Resources Research*, **10**, 514–520.

Mualem Y. (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 513–522.

Mualem Y. and Dagan G. (1978) Hydraulic conductivity of soils – unified approach to the statistical models. *Soil Science Society of America Journal*, **42**, 392–395.

Mualem Y. and Klute A. (1984) Predictor-corrector method for measurement of hydraulic conductivity and membrane conductance. *Soil Science Society of America Journal*, **48**, 993–100.

Nimmo J.R. (1992) Semiempirical model of soil water hysteresis. *Soil Science Society of America Journal*, **56**, 1723–1730.

Nimmo J.R. (1997) Modeling structural influences on soil water retention. *Soil Science Society of America Journal*, **61**, 712–719.

Nimmo J.R. and Akstin K.C. (1988) Hydraulic conductivity of a sandy soil at low water content after compaction by various methods. *Soil Science Society of America Journal*, **52**, 303–310.

Nimmo J.R., Deason J.A., Izbicki J.A. and Martin P. (2002a) Evaluation of unsaturated-zone water fluxes in heterogeneous alluvium at a Mojave Basin site. *Water Resources Research*, **38**, 1215, doi:10.1029/2001WR000735, 33-1 - 33-13.

Nimmo J.R. and Miller E.E. (1986) The temperature dependence of isothermal moisture-vs.-potential characteristic of soils. *Soil Science Society of America Journal*, **50**, 1105–1113.

Nimmo, J.R., Perkins, K.S. and Lewis, A.M. (2002b) Steady-state centrifuge [simultaneous determination of water transmission and retention properties]. In *Methods of Soil Analysis, Part 4 – Physical Methods*, Dane, J.H. and Topp, G.C. (Eds.), Soil Science Society of America: Madison, pp. 903–916 933–936.

Nimmo J.R., Perkins K.S., Rose P.A., Rousseau J.P., Orr B.R., Twining B.V. and Anderson S.R. (2002c) Kilometer-scale rapid transport of naphthalene sulfonate tracer in the unsaturated zone at the Idaho national engineering and environmental laboratory. *Vadose Zone Journal*, **1**, 89–101.

Nimmo, J.R., Shakofsky, S.M., Kaminsky, J.F. and Lords, G.S. (1999) Laboratory and field hydrologic characterization of the shallow subsurface at an Idaho National Engineering and Environmental Laboratory waste-disposal site, U.S. Geological Survey Water-Resources Investigations Report 99–4263, U.S. Geological Survey Water-Resources Investigations, Idaho Falls.

Parlange J.Y. and Hillel D.E. (1976) Theoretical analysis of wetting front instability in soils. *Soil Science*, **122**, 236–239.

Passioura J.B. (1977) Determining soil water diffusivities from one-step outflow experiments. *Australian Journal of Soil Research*, **15**, 1–8.

Philip J.R. (1957) The theory of infiltration – 1. The infiltration equation and its solution. *Soil Science*, **83**, 345–357.

Pruess K. (1998) On water seepage and fast preferential flow in heterogeneous, unsaturated rock fractures. *Journal of Contaminant Hydrology*, **30**, 333–362.

Purcell W.R. (1949) Capillary pressures – their measurement using mercury and the calculation of permeability therefrom. *Transactions AIME*, **186**, 39–46.

Raats P.A.C. (1973) Unstable wetting fronts in uniform and nonuniform soils. *Soil Science Society of America Proceedings*, **37**, 681–685.

Richards L.A. (1931) Capillary conduction of liquids through porous materials. *Physics*, **1**, 318–333.

Rieu M. and Sposito G. (1991) Fractal fragmentation, soil porosity, and soil water properties – I. Theory. *Soil Science Society of America Journal*, **55**, 1233–1238.

Ripple, C.D., Rubin, J. and Van Hylckama, T.E.A. (1972) *Estimating Steady-State Evaporation Rates from Bare Soils under Conditions of High Water Table*.

Rockhold M.L., Rossi R.E. and Hills R.G. (1996) Application of similar media scaling and conditional simulation for modeling water flow and tritium transport at the Las Cruces trench site. *Water Resources Research*, **32**, 595–609.

Ross P.J. and Smettem K.R.J. (1993) Describing soil hydraulic properties with sums of simple functions. *Soil Science Society of America Journal*, **57**, 26–29.

Ross P.J., Williams J. and Bristow K.L. (1991) Equation for extending water-retention curves to dryness. *Soil Science Society of America Journal*, **55**, 923–927.

Rossi C. and Nimmo J.R. (1994) Modeling of soil water retention from saturation to oven dryness. *Water Resources Research*, **30**, 701–708.

Salvucci G.D. (1996) Series solution for Richards equation under concentration boundary conditions and uniform initial conditions. *Water Resources Research*, **32**, 2401–2407 1996–066284.

Scanlon B.R., Andraski B.J. and Bilskie J. (2002) Miscellaneous methods for measuring matric or water potential. In *Methods of Soil Analysis, Part 4–Physical Methods*, Dane J.H. and Topp G.C. (Eds.), Soil Science Society of America: Madison, pp. 643–670.

Schaap M.G., Leij F.J. and van Genuchten M.T. (1998) Neural network analysis for hierarchical prediction of soil hydraulic properties. *Soil Science Society of America Journal*, **62**, 847–855.

Schwille F. (1988) *Dense Chlorinated Solvents in Porous and Fractured Media*, Lewis Publishers: Chelsea.

Selker J., Parlange J.-Y. and Steenhuis T. (1992) Fingered flow in two dimensions 2. predicting finger moisture profile. *Water Resources Research*, **28**, 2523–2528.

Shan C. (1995) Analytical solutions for determining vertical air permeability in unsaturated soils. *Water Resources Research*, **31**, 2193–2200.

Sheets K.R. and Hendricks J.M.H. (1995) Noninvasive soil water content measurement using electromagnetic induction. *Water Resources Research*, **31**, 2401–2409.

Smettem K.R.J., Ross P.J., Haverkamp R. and Parlange J.Y. (1995) Three-dimensional analysis of infiltration from the disk infiltrometer. 3. Parameter estimation using a double-disk tension infiltrometer. *Water Resources Research*, **31**, 2491–2495.

Sposito G. (1986) The "physics" of soil water physics. *Water Resources Research*, **22**, 83S–88S.

SSSA (1997) *Glossary of Soil Science Terms 1996*, Soil Science Society of America: Madison.

Stonestrom D.A. and Akstin K.C. (1994) Nonmonotonic matric pressure histories during constant flux infiltration into homogeneous profiles. *Water Resources Research*, **30**, 81–91.

Stonestrom D.A. and Rubin J. (1989a) Air permeability and trapped-air content in two soils. *Water Resources Research*, **25**, 1959–1969.

Stonestrom D.A. and Rubin J. (1989b) Water content dependence of trapped air in two soils. *Water Resources Research*, **25**, 1947–1958.

Su G.W., Nimmo J.R. and Dragila M.I. (2003) Effect of isolated fractures on accelerated flow in unsaturated porous rock. *Water Resources Research*, **39**, 1326, doi:10.1029/2002WR001691.

Tokunaga T. and Wan J. (1997) Water film flow along fracture surfaces of porous rock. *Water Resources Research*, **33**, 1287–1295.

Topp G.C., Davis J.L. and Annan A.P. (1980) Electromagnetic determination of soil water content; measurements in coaxial transmission lines. *Water Resources Research*, **16**, 574–582.

Tuller M., Or D. and Dudley L.M. (1999) Adsorption and capillary condensation in porous media: liquid retention and interfacial configurations in angular pores. *Water Resources Research*, **35**(7), 1949–1964.

Valiantzas J.D. and Kerkides P.G. (1990) A simple iterative method for the simultaneous determination of soil hydraulic properties from one-step outflow data. *Water Resources Research*, **26**, 143–152.

van Dam J.C., Stricker J.N.M. and Droogers P. (1994) Inverse method to determine soil hydraulic functions from multistep outflow experiments. *Soil Science Society of America Journal*, **58**, 647–652.

van Genuchten M.T. (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

Vandervaere J.-P., Vauclin M. and Elrick D.E. (2000a) Transient flow from tension infiltrometers: I. The two-parameter equation. *Soil Science Society of America Journal*, **64**, 1263–1272.

Vandervaere J.-P., Vauclin M. and Elrick D.E. (2000b) Transient flow from tension infiltrometers: II. Four methods to determine sorptivity and conductivity. *Soil Science Society of America Journal*, **64**, 1272–1284.

Wan J. and Wilson J.L. (1994) Colloid transport in unsaturated porous media. *Water Resources Research*, **30**, 847–864.

Wang Z., Feyen J. and Elrick D.E. (1998) Prediction of fingering in porous media. *Water Resources Research*, **34**, 2183–2190.

Wang Z., Wu L., Harter T., Lu J. and Jury W.A. (2003) A field study of unstable preferential flow during soil water redistribution. *Water Resources Research*, **39**, 1075, doi:10.1029/2001WR000903.

Warrick A.W., Mullen G.J. and Nielsen D.R. (1977) Scaling field-measured soil hydraulic properties using a similar media concept. *Water Resources Research*, **13**, 355–362.

Weeks, E.P. (1978) *Field Determination of Vertical Permeability to Air in the Unsaturated Zone*, US Geological Survey Professional Paper 1051.

151: Hydraulics of Wells and Well Testing

PHILIPPE RENARD

University of Neuchâtel, Neuchâtel, Switzerland

Well testing and well hydraulics play a major role in applied hydrogeology. While well hydraulics aims at modeling the groundwater behavior in response to a perturbation – such as pumping – in a well, well testing aims at using these models in an inverse approach to infer the properties of the aquifer or of the well itself. The history of well hydraulics and well testing started in 1863 with Dupuit, who developed the first analytical solution to model radial flow to a well in steady state. In 1935, Theis published the most important analytical solution. The Theis solution assumes that the aquifer is confined, bidimensional, homogeneous, and isotropic. Subsequently, numerous models have been developed with the aim to enlarge the domain of applications. This article presents a brief review of these various models and describes their behavior in terms of drawdown and log derivative of the drawdown. The log derivative is used as a tool to help in the identification of the most appropriate model when analyzing field data.

INTRODUCTION

Well hydraulics is the part of groundwater hydraulics (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**) that deals specifically with aquifer response to hydraulic perturbations in a well. Well tests are field tests in which the hydraulic response is measured and analyzed. Well hydraulics and well testing are intimately related since they respectively solve the direct and the inverse problem. As any inverse problem (*see* **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**), the well-test interpretation suffers from nonuniqueness. Usually this is solved by assuming a very simple model with a very small number of parameters as compared to the number of field observations. The problem is then over-determined and has a unique solution when a model has been chosen. The difficulty is to identify the model that best represents reality. Often, different models will show the same type of response and they will not be distinguishable (e.g. unconfined aquifer and double porosity aquifer). In addition, the measurements may not be sufficient to see all the typical phases of a given model, either because the very early phase is too rapid and is not recorded or conversely the late phase is not recorded because the test does not last long enough. This long introductory remark highlights the fact that despite the relatively large number of models that are available, the practitioner will always be confronted with the dilemma of a nonunique interpretation.

Since groundwater is mainly extracted by wells and well testing represents the main field method for determining the hydraulic properties of the subsurface, much research has been conducted in this area. We can trace the first publication on well hydraulics to Dupuit (1863), who proposed the first series of analytical solutions to steady-state flow to a well in idealized confined or unconfined aquifers. Since then, there has been a continuous and parallel development of analytical models and well-testing procedures. This is reflected by the numerous books published, both in the fields of hydrogeology (Batu, 1998; Butler, 1998; Dawson and Istok, 1991; Hantush, 1964; Kruseman and Ridder, 1992; Lebbe, 1999; Lee, 1999; Walton, 1996) and petroleum engineering (Bourdet, 2002; Earlougher, 1977; Horne, 1995; Raghavan, 1993; Streltsova, 1988).

The aim of this article is to provide an introduction to this field. The article is not intended to be used as a manual or to be exhaustive. Rather, it includes some of the most important historical findings with sufficient details to permit the reader to apply the techniques and understand the basic concepts and equations. Additional aspects are covered in a cursory manner, simply to direct the reader to pertinent sources of information. Some aspects have arbitrarily been

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

left aside such as anisotropy, pulse tests, sinusoidal tests, horizontal wells, and so on.

The article is organized in five sections. The first section presents an overview of well-testing procedures and interpretation methodologies. The second introduces the basic equations of groundwater flow to a well in an ideal confined aquifer as well as the steady-state solution for such a case. The third considers the Theis solution for the transient regime in the ideal confined aquifer. The fourth reviews different nonidealities that can affect the response of an aquifer to pumping in a well. The last briefly considers some other types of hydraulic perturbations.

WELL TESTING

Well tests are conducted with two major objectives. One is to determine the properties of geological formations for a broad range of applications from groundwater exploration to waste-disposal-site evaluation. The other objective is to evaluate the hydraulic properties of the production well itself in order to design the exploitation scheme (depth of the pump, pumping rate) or to evaluate the well efficiency.

The fundamental principle of well testing is the imposition of a hydraulic perturbation in the well and monitoring the aquifer response. Testing procedure may be classified according to the type of perturbation, and to the type and location of response monitoring. Note that the typology is not mutually exclusive.

- *Single well test*: The perturbation and the monitoring are conducted in the same borehole.
- *Interference test*: The perturbation and the monitoring are conducted in separate boreholes.
- *Pumping test*: The aquifer is perturbed by pumping. It can either be a single well test or an interference test. Generally the pumping rate is constant, but variable-pumping-rate tests can also be interpreted. An *injection test* is similar to a pumping test, but water is injected rather than being extracted.
- *Step-drawdown test*: A single well test with a series of successive constant pumping rates.
- *Buildup or recovery test*: It follows a pumping test. After the pump has been stopped, the recovery to the initial level is observed, either in the pumping well or in observation wells.
- *Constant head test*: The head is maintained constant and the water discharge is recorded in the perturbation well. Head changes can be recorded in observation boreholes.
- *Slug test*: The perturbation is a sudden modification of the head in the well, the response is the head variation in the well itself or in observation boreholes.
- *Packer test*: It can be any of the above tests, but it is conducted in an interval of the well isolated with the help of packers. The packers are inflatable or mechanical and allow testing a distinct zone within a well.

Once data – time series of pressures or heads and/or discharge rates – have been recorded from one of the above tests, the interpretation procedure follows fours steps.

- *Data preprocessing*. The data are converted into adequate units, outliers and trends are removed.
- *Model identification*. A conceptual and mathematical model is chosen based on the geological information available and based on a qualitative analysis of the data.
- *Parameter identification*. The physical parameters of the model are obtained by fitting the theoretical response to the observed response. This is a typical inverse problem (*see* **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**).
- *Quality control*. The adequacy of the model must be checked. A statistical analysis of the discrepancies between the model and the data is conducted in order to test the validity of the interpretation.

While standard well-test analysis used to be based on manual type curve matching or straight-line analysis, the methodology is now mostly computerized for all the four above introduced steps. As a consequence, new methods have been intensively developed in the last twenty years. The benefits of interactive graphics and data handling with computers are obvious. Computerized analysis has allowed handling large data sets directly recorded in the field with electronic pressure gauges and data acquisition systems. Moreover, it has allowed the introduction of the systematic use of the logarithmic derivative of the drawdown as a tool for model identification (Bourdet *et al.*, 1989). Numerical Laplace inversion has broaden the range of models that would be tedious to program otherwise (Dougherty, 1989; Moench and Ogata, 1984). Numerical convolution has allowed interpretation of continuously variable pumping rate test. Automatic model fitting with different optimization algorithms became also possible (McElwee, 1980; Rosa and Horne, 1991), opening the way to a more objective model fitting as well as to a statistical analysis of the results. Finally, computerized data analysis has also allowed the application of general purpose or specifically developed numerical models (*see* **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**) to interpret field data (Lavenue and de Marsily, 2001; Lebbe, 1999; Pinder and Bredehoeft, 1968), or to forecast the impact of wells on aquifers.

With the expansion of personal computing technology, a large number of well test analysis software has been developed both in the groundwater and petroleum engineering fields. Some of them are freely available through various institutional or academic internet sites (`http://water.usgs.gov/nrp/gwsoftware/`,

http://water.usgs.gov/software/, **http://www.kgs.ukans.edu/Datasale/suprpump.html**, **http://home.olemiss.edu/~acheng/software/index.html**) or under commercial licenses **http://www.scisoftware.com/products/cat_pump_test/cat_pump_test.html**. Amongst other commercial software, and based on the author's experience, one of the best available software package from the petroleum engineering field is Saphir (**http://www.kappaeng.com/Saphir/index.asp**).

THE IDEAL CONFINED AQUIFER

Groundwater-Flow Equation Around a Well

Following Dupuit (1863), we consider an idealized aquifer (Figure 1a). The aquifer is assumed to be infinite in lateral extent, fully confined (no recharge or leakage), two dimensional (large extension compared to its thickness), having a homogeneous transmissivity and storativity. At time zero, pumping starts at a constant rate in the well. If we assume that the heads are constant in space prior to pumping, the flow field will be radial (Figure 1b). All flow lines will converge toward the well as straight lines. If we consider a cylinder of radius r centered on the well, the total water flux $Q(r)$ [L^3T^{-1}] flowing through the cylinder is given by multiplying the area of the cylinder by the specific discharge calculated according to Darcy's law. This yields the following:

$$Q(r) = -2\pi r T \frac{\partial h}{\partial r} \tag{1}$$

where T [L^2T^{-1}] is the transmissivity of the aquifer (hydraulic conductivity multiplied by the aquifer thickness). Since the aquifer is fully confined, and since the water is only slightly compressible, the mass conservation of water between two cylinders during a time interval can be expressed as

$$\frac{\partial Q}{\partial r} = -2\pi r S \frac{\partial h}{\partial t} \tag{2}$$

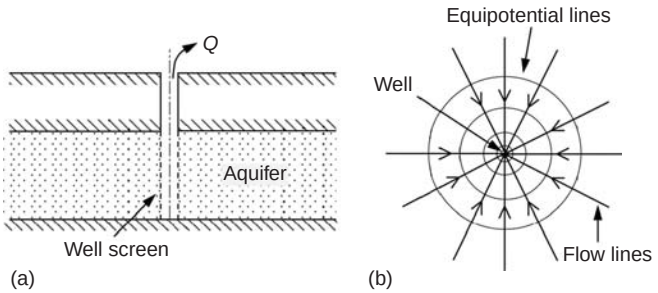


Figure 1 Schematic illustration of an idealized confined aquifer (a) and radial converging flow (b) to a fully penetrating pumping well in such aquifer

with S[−] representing the storage coefficient and t the time. Combining equations (1) and (2) yields

$$\frac{1}{r}\frac{\partial h}{\partial r} + \frac{\partial^2 h}{\partial r^2} = \frac{S}{T}\frac{\partial h}{\partial t} \tag{3}$$

Equation (3) is linear, and has been derived by assuming for didactic reasons that the heads were constant prior to pumping. But this is not a necessity. In general, equation (3) will be written instead in terms of drawdown s [L], defined as the difference between the head h_0 as it would have been without pumping minus the actual head h in the aquifer during pumping (Figure 2).

$$s(t) = h_0(t) - h(t) \tag{4}$$

While h is the solution to equation (3) with complex boundary and initial conditions, s is another solution with much simpler initial and boundary conditions.

It is important to note that when an aquifer is unconfined, the previous equations are not rigorously valid. They can only be applied if the saturated thickness is large compared to the drawdown. Other equations, which are nonlinear, were derived by Dupuit, under additional simplifying assumptions but are of limited applicability. In the following, we restrict ourselves to the case of confined aquifers unless we specify otherwise.

The Dupuit–Thiem Solution in Steady State

Once again, following Dupuit (1863), we consider the case in which flow to a well is in dynamic equilibrium, that is, steady state. The pumping rate Q is constant and the head in the aquifer varies in space but not in time. Under this assumption, the flow $Q(r)$ through any cylinder must be identical to the pumping rate Q. Hence, we can directly integrate equation (1), and making use of the definition of the drawdown to find

$$s(r) = -\frac{Q}{2\pi T}\ln(r) + A \tag{5}$$

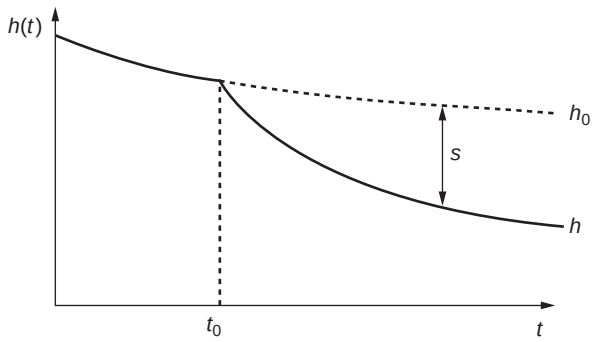


Figure 2 Definition of the drawdown

A is a constant of integration. To eliminate A, Dupuit defines rather arbitrarily a radius R at which the head h_0 is not affected by the pumping. A better technique was proposed by Thiem (1906) who considered the drawdown difference $s_1 - s_2$ between two observations points located at distance r_1 and r_2 from the pumping well. This yields

$$T = \frac{Q}{2\pi(s_2 - s_1)} \ln\left(\frac{r_1}{r_2}\right) \tag{6}$$

Equation (6), known as the *Thiem's formula*, allows determination of the transmissivity T of an aquifer from a constant pumping rate field experiment. Thus, one only has to measure the drawdown difference between two piezometers, and make sure that the difference remains constant and to calculate T with equation (6). Despite that this equation has been derived under steady state assumptions, we will show in Section Relation between Jacob and Thiem solutions that it is also valid for transient conditions for homogeneous aquifers. However, in practice, it should be applied with great care, as the calculated transmissivity is strongly affected by heterogeneities as we will discuss in Section Heterogeneous aquifers.

THE THEIS SOLUTION

The Theis (1935) solution considers the same geometry as the Dupuit–Thiem solution, but under transient-flow regime. It is the most important analytical solution for well hydraulics because most other transient models tend towards it either for early or late times. Therefore, it can often be used even if the underlying assumptions are not fully satisfied.

Assumptions and Solution

The Theis model assumes that the aquifer is confined, infinite, homogeneous, and isotropic; the well is fully penetrating the aquifer, the well has a negligible radius and is 100% efficient; the pumping rate Q is constant. The initial condition is set to zero drawdown everywhere within the aquifer prior to pumping.

$$s(r, t = 0) = 0 \tag{7}$$

The boundary conditions consist of zero drawdown at infinity and a fixed flux equal to the pumping rate at the well.

$$\lim_{r\to\infty} s = 0 \tag{8}$$

$$\lim_{r\to 0} 2\pi T r \frac{\partial s}{\partial r} = -Q \tag{9}$$

The solution of equation (3) subject to initial and boundary conditions (7–9) yields the Theis solution

$$s = \frac{Q}{4\pi T} E_1\left(\frac{r^2 S}{4tT}\right) \tag{10}$$

E_1 is the exponential integral function. Written in dimensionless form the Theis solution is

$$s_D = \frac{1}{2} E_1\left(\frac{r_D^2}{4t_D}\right) \tag{11}$$

where s_D, r_D, and t_D represent respectively the dimensionless drawdown, dimensionless radius, and dimensionless time defined as

$$s_D = \frac{2\pi T}{Q} s, \quad r_D = \frac{r}{r_W}, \quad t_D = \frac{Tt}{r_W^2 S} \tag{12}$$

with r_W being the radius of the well. Note that the logarithmic derivative of the drawdown is as follows (Chow, 1952):

$$\frac{\partial s_D}{\partial \ln(t_D)} = t_D \frac{\partial s_D}{\partial t_D} = \frac{1}{2} \exp\left(-\frac{r_D^2}{4t_D}\right) \tag{13}$$

Equation (11) describes how the drawdown evolves in time and space as a function of transmissivity and storativity of the aquifer and as a function of the pumping rate. Since Bourdet *et al.* (1983), equation (11) together with its logarithmic derivative, equation (13), are usually plotted in log–log scale as a function of t_D/r_D^2. It shows that the drawdown increases with time and decreases with radial distance (see Figure 3a). The derivative tends toward a constant value of 0.5 for late time. The spatio-temporal behavior of the cone of depression shows that the shape of the cone is created very quickly, and then the cone moves downward.

Compression Zone and Radius of Investigation

An important question in well hydraulics and well testing is the definition of the volume of aquifer that is affected by pumping and influences the drawdown behavior. For that purpose, it is useful to calculate the flux of groundwater through a cylinder centered on the well, having a radius r_D. This flux, normalized by the pumping rate, is

$$q_D = \frac{q(r_D)}{Q} = \exp\left(-\frac{r_D^2}{4t_D}\right) \tag{14}$$

Figure 4 shows the behavior of equation (14), three zones are distinguishable and evolve with time. At small radial distances, the normalized flux is close to 1. This indicates that the flux is equal to the pumping rate; the aquifer simply transfers water to the well. At infinity, the flux is close to

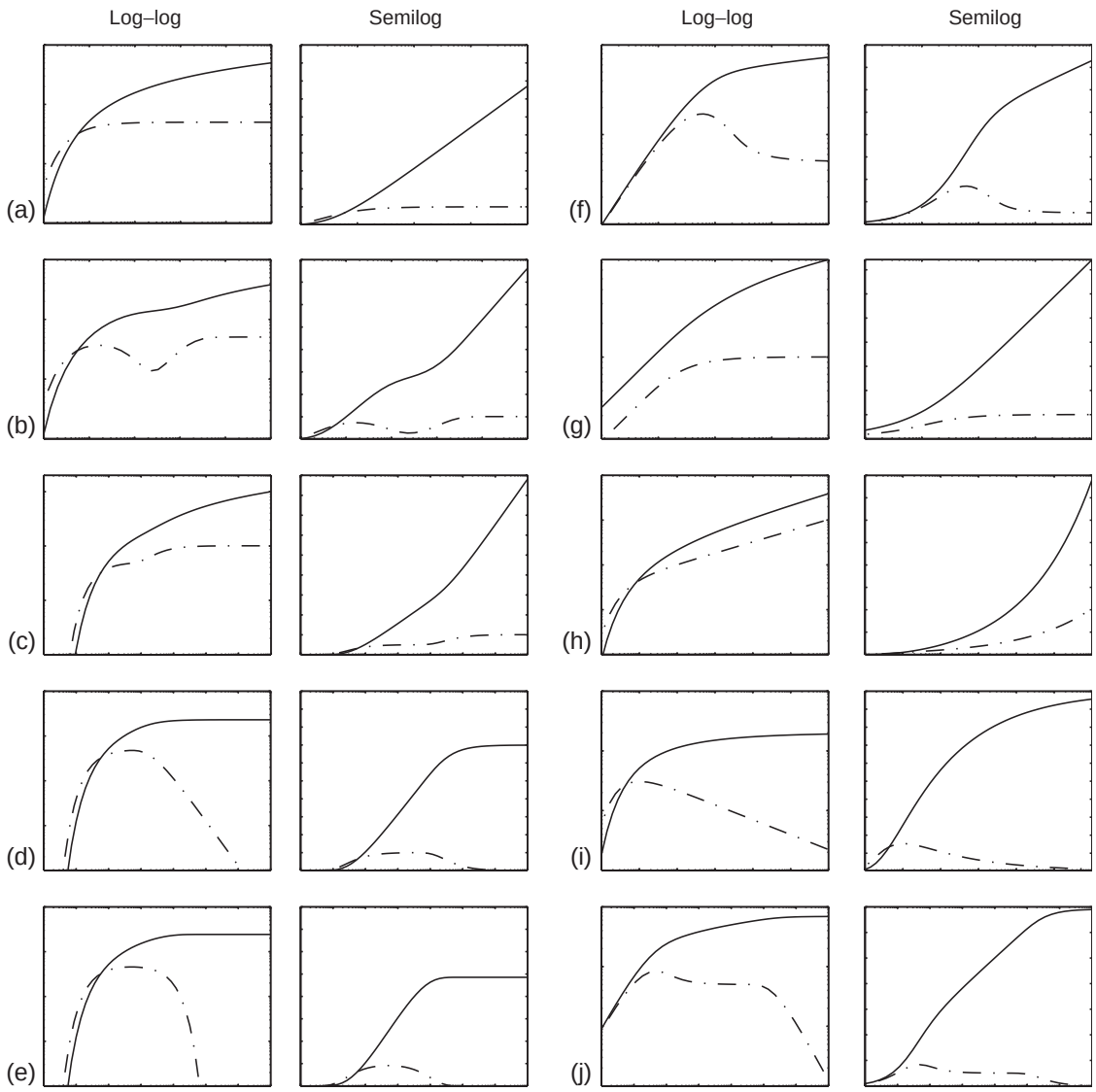


Figure 3 A synthesis of some typical drawdown behaviors in response to constant pumping rate. The drawdown (solid line) and the log derivative (dashed line) are plotted as a function of time in double-logarithmic or semilogarithmic scale. (a) Theis model: confined ideal aquifer. (b) Unconfined, or double porosity aquifer. c) Confined aquifer with a no-flow boundary. (d) Confined aquifer with a constant head boundary. (e) Leaky aquifer: Hantush and Jacob (1955) model. (f) Single well test with well-bore storage and possibly skin effects. (g) Single vertical fracture having an infinite conductivity. (h) General Radial-Flow model with $n < 2$. (i) General Radial-Flow model with $n > 2$. (j) Single well test with well-bore storage, infinite acting radial flow and constant head boundary (Reproduced from Gringarten *et al.* (1974), by permission of American Geophysical Union)

zero. There, the aquifer is not yet active. At intermediate distances, the fluxes vary from 0 to 1; water is mobilized from the compaction of aquifer and elastic expansion of the water itself. This is the so called *compression zone*, which propagates with time.

The radius of investigation of a pumping test is a concept introduced by Dupuit. It was defined as the radius beyond which the drawdown is zero. For the Theis solution, which never predicts drawdown to be strictly zero, there are several possible and arbitrary definitions depending on what is considered as being a negligible drawdown. An elegant possibility is to define the radius of investigation as the radius such that the rate of increase in drawdown with time is maximal (Van Poolen, 1964). Using this definition and calculating the distance at which the second temporal derivative of the drawdown is zero, one finds the simple result:

$$r_{\mathrm{D}}^{i} = 2\sqrt{t_{\mathrm{D}}} \quad \text{or} \quad r^{i} = 2\sqrt{\frac{Tt}{S}} \tag{15}$$

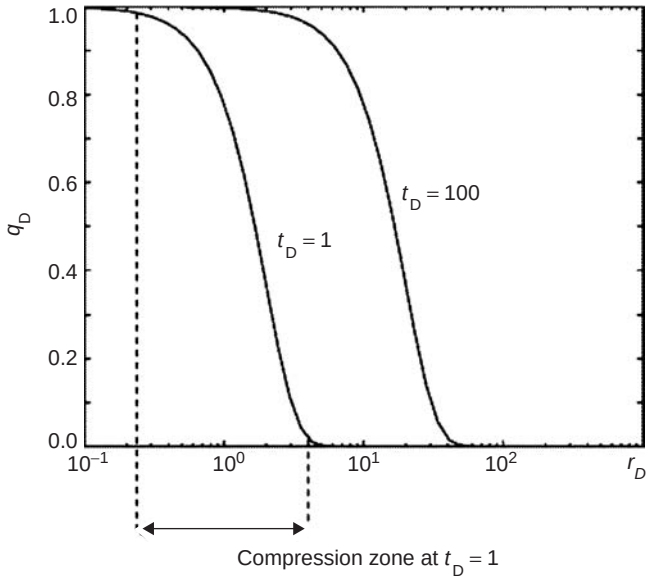


Figure 4 Normalized groundwater flux as a function of radial distance from the well and time

Most alternative definitions of the radius of investigation yield similar equations, being functions of the square root of dimensionless time, but with scaling factors varying from 1.5 to 4.3. All these definitions illustrate the arbitrariness of the concept, but interestingly they all locate the radius of investigation within the compression zone.

More recently, the question of the extension of the investigation domain during a pumping test has been formulated in the framework of spatial filtering functions and heterogeneous transmissivity fields (Beckie, 2001; Oliver, 1993). These authors found that the investigation area is more or less an ellipse that encloses the pumping and the observation wells, but the influence of the heterogeneities within this ellipse is not spatially uniform.

Approximation of the Exponential Integral Function

Applying equations (10) or (11) for well-test interpretation or to design exploitation schemes requires the use of the exponential integral function. It is generally directly available in most mathematical software but, if necessary, it can be calculated with a polynomial and rationale approximations (Abramowitz and Stegun, 1970). For small x, the approximation is

$$\begin{aligned} &0 \le x \le 1 \\ &E_1(x) = -\gamma - \ln(x) + a_1 x + a_2 x^2 \\ &\qquad + a_3 x^3 + a_4 x^4 + a_5 x^5 + \varepsilon \\ &|\varepsilon| < 210^{-7} \end{aligned} \tag{16}$$

with γ being the Euler constant and

$$\begin{aligned} &\gamma = 0.57721566 \quad a_2 = -0.24991055 \quad a_4 = -0.00976004 \\ &a_1 = 0.99999193 \quad a_3 = 0.05519968 \quad\quad a_5 = 0.00107857 \end{aligned} \tag{17}$$

For large x, the approximation is

$$\begin{aligned} &1 \le x \le \infty \\ &E_1(x) = \frac{1}{xe^x}\left(\frac{x^4 + a_1x^3 + a_2x^2 + a_3x + a_4}{x^4 + b_1x^3 + b_2x^2 + b_3x + b_4}\right) + \varepsilon \\ &|\varepsilon| < 210^{-8} \end{aligned} \tag{18}$$

with

$$\begin{aligned} a_1 &= 8.5733287401 \quad & a_2 &= 18.0590169730 \\ a_3 &= 8.6347608925 \quad & a_4 &= 0.2677737343 \\ b_1 &= 9.5733223454 \quad & b_2 &= 25.6329561486 \\ b_3 &= 21.0996530827 \quad & b_4 &= 3.9584969228 \end{aligned} \tag{19}$$

Asymptotic Behavior: the Jacob Solution

Equation (16) shows that the exponential integral $E_1(x)$ tends toward $-\gamma - \ln(x)$ when x tends toward 0. In practice, when the dimensionless number t_D/r_D^2 is greater than 10, the Theis solution can be approximated by a simple logarithmic function. This is the so-called Jacob's approximation of the Theis solution (Cooper and Jacob, 1946).

$$s_D \approx \frac{1}{2}\left[\ln\left(\frac{4t_D}{r_D^2}\right) - \gamma\right] \quad \text{or} \quad s \approx \frac{2.30Q}{4\pi T}\log\left(\frac{2.25tT}{r^2S}\right) \tag{20}$$

Note that when equation (20) is valid, the logarithmic derivative of the drawdown becomes constant.

$$\frac{\partial s_D}{\partial \ln(t_D)} = t_D\frac{\partial s_D}{\partial t_D} = \frac{1}{2} \quad \text{or} \quad \frac{\partial s}{\partial \ln(t)} = \frac{Q}{4\pi T} \tag{21}$$

This relation is used to analyze data sets. One calculates the logarithmic derivative of the measured drawdown and plots it as function of time. When the derivative becomes constant, it indicates that the drawdown has reached a logarithmic asymptote. In petroleum engineering literature, the interval of the data set that shows a constant derivative is identified as the *Infinite Acting Radial Flow* (IARF). During this period of time, one can use safely the logarithmic approximation (18) to interpret the data. If the derivative does not remain constant, the logarithmic approximation cannot be used.

Relation Between Jacob and Thiem Solutions

When Jacob's approximation is valid in two observation boreholes located at distances r_1 and r_2 from the pumping well, we can use equation (20) to calculate the drawdown

difference between the two piezometers. The result is a constant value

$$s_2 - s_1 = \frac{Q}{2\pi T} \ln\left(\frac{r_1}{r_2}\right) \tag{22}$$

which is identical to what Thiem predicted using the steady-state assumption (equation 6).

Interpretation of a Pumping Test with the Theis Solution

The original method was based on the graphical superposition of the Theis type curve with the observed data on two log–log sheets of paper. Nowadays, many software will contain algorithms that apply the Theis method, but we believe that, here, it is useful to illustrate how the interpretation procedure can easily be implemented with any spreadsheet, graphical, or mathematical software. As a preliminary remark, note that equation (20) can be written as

$$s = a \log\left(\frac{t}{t_0}\right) \tag{23}$$

with

$$a = \frac{2.30Q}{4\pi T} \tag{24}$$

$$t_0 = \frac{r^2 S}{2.25T} \tag{25}$$

a has the dimension of length [L], t_0 has the dimension of time [T]. A plot of the measured drawdown s as function of the log of time will show a straight line with a slope a for late time. t_0 will be the value of t for which $s = 0$. We can therefore very easily read the values of a and t_0 from the graph (Figure 5a). A simple interpretation procedure is then:

1. Plot the drawdown measured in the field and its logarithmic derivative as a function of time on the same semilog scale graph.
2. Estimate roughly on the graph the slope a of the straight line, or read the value of the derivative when it stabilizes, it is $a/2.3$. In the example shown in Figure 5(a), we find a slope of about 1.75 m.
3. Read the time t_0 by extrapolating the straight line to the intercept with the horizontal axis, that is, where $s = 0$. In the example, we find $t_0 = 3.5\,10^2$ s.
4. Calculate with the spreadsheet or mathematical software the theoretical drawdown and derivative using the values of a and t_0.

$$s = \frac{a}{2.3} E_1\left(\frac{0.5625t_0}{t}\right) \tag{26}$$

$$\frac{\partial s}{\partial \ln(t)} = \frac{a}{2.3} \exp\left(-\frac{0.5625t_0}{t}\right) \tag{27}$$

   If the exponential integral is not directly available you can use equations (16) and (18). Superpose the theoretical drawdown and derivative on the graph previously done.
5. Improve the fit iteratively by modifying the values of a and t_0 and by visual inspection on your graph or by using a nonlinear least-square algorithm. It is possible to switch (by using the graphical options of the plotting software) from a semilog to a log–log plot if one wants to represent the final plot in log–log scale but this is not required. In our example, the final fit is shown in Figure 5(b).

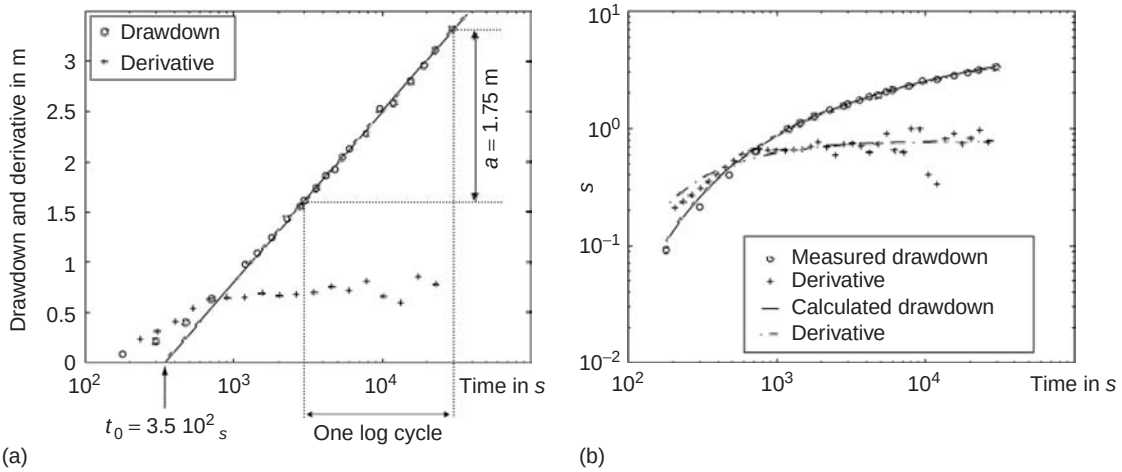


Figure 5 Example of an interpretation of a data set with Theis method. (a) semilog plot of the drawdown and log derivative of a data set, identification of the constant a and t_0. (b) Superposition of the data with the model after automatic fitting

6. Should the fit be acceptable (both drawdown and derivative), the transmissivity T and the storativity S of the aquifer can be estimated with:

$$T = \frac{2.30Q}{4\pi a} = 0.183\frac{Q}{a} \tag{28}$$

$$S = \frac{2.25Tt_0}{r^2} \tag{29}$$

NONIDEALITIES

Heterogeneous Aquifers

What is the validity of the assumption of homogeneity? What is the meaning of the transmissivity obtained from a pumping test interpreted with Theis or Jacob solutions when the aquifer is heterogeneous? Meier *et al.* (1998) conducted a numerical study to investigate these questions. They simulated numerically the flow to a well in a series of heterogeneous aquifers. They considered that the transmissivity field is heterogeneous (Figure 6a) while the storativity is kept constant. They imposed a constant rate in the well and calculated the drawdowns in several piezometers (Figure 6b) using a finite element code. The numerical experiment is repeated for different types of transmissivity fields.

The main conclusion from this study is that the apparent transmissivity, estimated from the slope of the Jacob's straight line, of the transmissivity fields that they investigated is almost identical in all observation wells even if the heterogeneity in transmissivity is important. In these cases, the apparent transmissivity is very close to the uniform-flow effective transmissivity of the heterogeneous media (*see* **Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). However, the straight lines are shifted from one observation well to the next (Figure 6b), and therefore the estimated storativities vary within several orders of magnitude. The apparent heterogeneity in storativity is only a consequence of the heterogeneity in transmissivity. When interpreting a well test, such apparent heterogeneity in estimated storativity may be used as an indicator of the degree of heterogeneity of the transmissivity field. Furthermore, when many observation wells are available the geometric mean of the estimated storativity can be used as an estimator of the real storativity (Sánchez-Vila *et al.*, 1999). A consequence of the shift of the straight lines between observation wells due to heterogeneity, is that the Thiem method that only accounts for drawdown difference between two points will be strongly biased depending on the location of the points. This effect is illustrated by a laboratory experiment in a sand tank filled with a spatially variable pattern of different sands (Silliman and Caswell, 1998). They show that the application of the Thiem formula provides estimates of hydraulic conductivities that are extremely variable and that can be significantly lower (even negative) or higher than the local hydraulic conductivities.

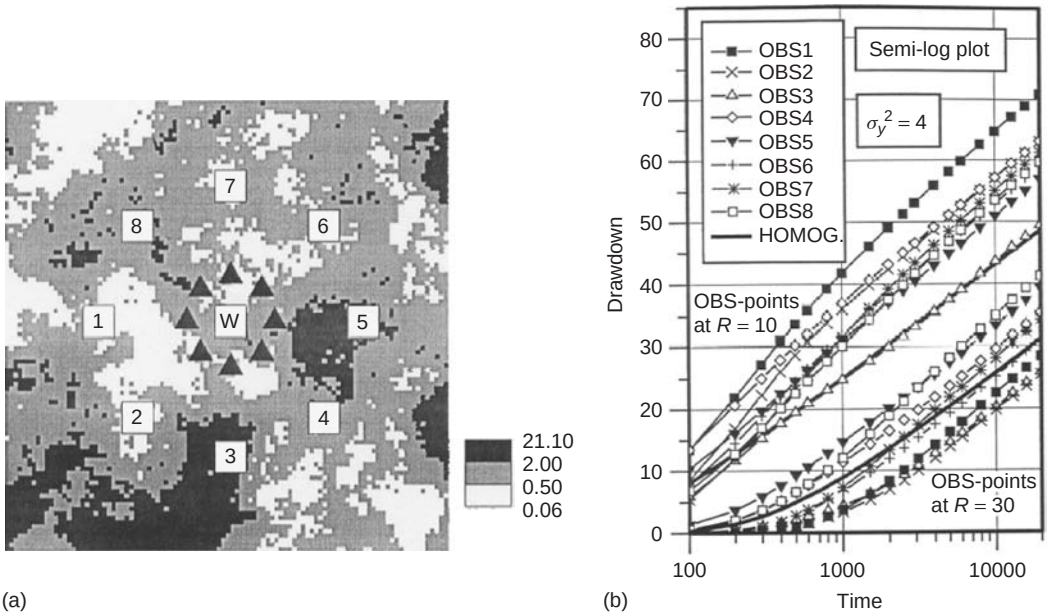


Figure 6 Effect of heterogeneity. (a) Example of a simulated map of transmissivity for a multilognormal field. The W indicates the location of the pumping well. The triangles and the squares indicate the location of two series of observation wells. (b) Calculated drawdown in all the observation points. The line labeled HOMOG represents the calculated drawdown for the equivalent homogeneous medium (Reproduced from Meier *et al.,* 1998 by permission of American Geophysical Union)

In steady state, an interesting result is provided by Dagan (1982), who shows that the expected value of the hydraulic head in a heterogeneous aquifer follows the usual Dupuit solution (equation 5). Dagan also provides approximations for the variance of the head and shows how it is influenced by the heterogeneity of transmissivity and pumping rate.

Partially Penetrating Wells

When the well does not fully penetrate the aquifer (Figure 7a), the flow field around the well has a vertical component. Hantush (1961) showed that in such a case the drawdown tends toward the classical Theis solution when

- the piezometer is located at a distance which is larger than one and a half times the thickness of the aquifer;
- or the observation well is screened over the complete thickness of the aquifer.

When the piezometer is close to the pumping well, the drawdown is affected by the partial penetration depending on the respective location of the well screen and the piezometer (Figure 7b). The late time asymptote is a Jacob straight line whose slope is the same as for a fully penetrating well. A common mistake is to use the screened interval thickness to obtain the hydraulic conductivity, while the real thickness of the aquifer should be used (if it is known).

Bounded Aquifers

When the aquifer is not infinite, the drawdown is affected by the presence of boundaries. There are two typical cases: the presence of a river recharging the aquifer (constant head boundary) and the presence of an impervious geological boundary (no flow boundary).

If we assume that all the other assumptions from Theis are still valid, then the solution to these two problems is obtained by the application of the superposition principle and the theory of images. The concept is that the boundary is mathematically equivalent to the presence of an imaginary well located on the opposite side of the boundary. For the case of a no flow boundary, the image well is a pumping well, while for the case of a constant head boundary, the image well is an injection well (Figure 8).

Using this theory, the solution for both cases is

$$s = \frac{Q}{4\pi T} E_1\left(\frac{r^2 S}{tT}\right) + \beta \frac{Q}{4\pi T} E_1\left(\frac{r_i^2 S}{tT}\right) \qquad (30)$$

with r_i being the distance between the observation well and the imaginary well, $\beta = 1$ for a no-flow boundary and $\beta = -1$ for a constant head boundary.

For the case of a no-flow boundary, the behavior of the drawdown is illustrated in Figure 3(c). It is characterized by a doubling of the late time slope. Two segments of straight line in semilog scale can be seen. The first one corresponds to the Jacob approximation before the drawdown is affected by the boundary. The second straight line corresponds to the superposition of the pumping well and the boundary. Figure 3(d) illustrates the constant head boundary case. The drawdown stabilizes because of recharge from a river; the derivative decreases continuously and follows a straight line on a log–log plot.

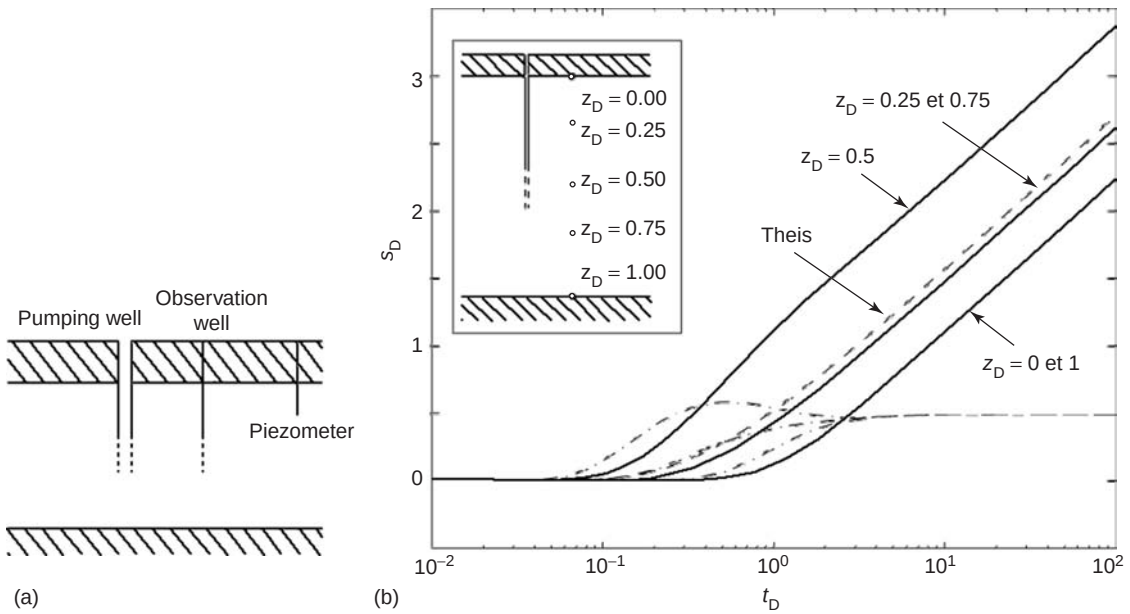


Figure 7 Schematic section through a confined aquifer partially penetrated by a pumping well. (b) Type curves of Hantush's solution for partial penetration in semilog scale for the geometrical configuration illustrated in the figure

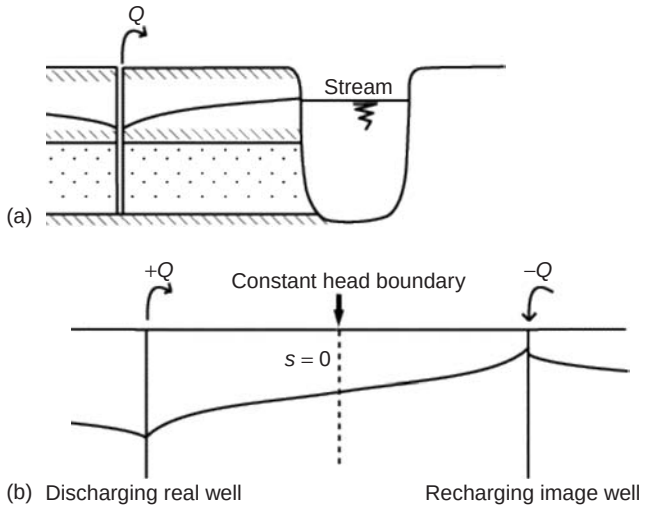


Figure 8 Schematical representation of a confined aquifer bounded by a constant head boundary

Equation (30) implicitly assumes that the constant head is imposed throughout the whole thickness of the aquifer (Figure 8). A recent model considers a more realistic setup that accounts for the geometry of the riverbed (width compared to aquifer thickness) and an imperfect hydraulic connection between the river and the aquifer (Butler *et al.*, 2001). This model allows propagation of drawdown beneath a partially penetrating stream. The type curves of predicted drawdown range from the fully penetrating stream case (equation 30) to the Theis solution (equation 11) as a function of aquifer geometry and stream leakance: streambed hydraulic conductance multiplied by the square of the streamed width divided by the transmissivity of the aquifer below the stream.

Equation (30) can be extended for multiple boundaries. When the boundaries fully penetrate the aquifer, analytical solutions are obtained by the application of image theory. The analytical solution is then the sum of a series of drawdowns resulting from pumping and injection wells located according to the geometry of the boundaries. If one of these boundaries is a constant head boundary, the drawdown stabilizes to a constant value and a steady state is reached for late times. Closed expressions for these asymptotic values are useful, for example for the design of pumping schemes. For a single constant head boundary, the late time asymptote is

$$s = \frac{Q}{2\pi T} \ln\left(\frac{2L}{r_0}\right) \tag{31}$$

with L being the distance between the well and the boundary and r_0 the radius of the well.

In the case of the closed rectangular system shown on Figure 9, the late time asymptote is obtained with the

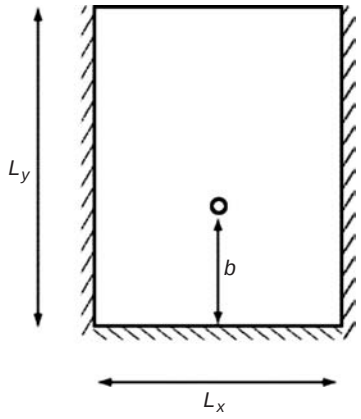


Figure 9 Geometry of a simplified rectangular system with a pumping well, three no flow boundaries and one constant head boundary

Perrochet approximation

$$s = \frac{Q}{2\pi T} \ln\left(\frac{e^{2\pi\frac{L_x - b}{L_y}}}{2\pi}\frac{L_y}{r_0}\right), \frac{L_x}{L_y} > \frac{1}{2}, \frac{3L_x}{4} > b > \frac{L_x}{4} \tag{32}$$

with relative errors less than 1%.

When b = 0, the asymptote becomes

$$s = \frac{Q}{\pi T} \ln\left(\frac{e^{\pi\frac{L_x}{L_y}}}{2\pi}\frac{L_y}{r_0}\right) \tag{33}$$

Variable Pumping Rate

As a starting point, suppose that the pumping rate is constant with a value Q_1 for $0 < t < t_1$ and then the pumping rate is suddenly increased to a value Q_2 at the time $t = t_1$. Assuming that all the other Theis assumptions are valid, the drawdown during the first phase of the test can be expressed with the Theis solution:

$$s = \frac{Q_1}{4\pi T} E_1\left(\frac{r^2 S}{4Tt}\right) t < t_1 \tag{34}$$

Because of the linearity of the groundwater flow equation for a confined aquifer, we can apply the principle of superposition in time and the drawdown in the second phase is simply the sum of the drawdown due to pumping Q_1 from $t = 0$ and the drawdown due to the pumping rate difference $Q_2 - Q_1$ starting at $t = t_1$.

$$s = \frac{Q_1}{4\pi T} E_1\left(\frac{r^2 S}{4Tt}\right) + \frac{Q_1 - Q_2}{4\pi T} E_1\left(\frac{r^2 S}{4T(t - t_1)}\right) t > t_1 \tag{35}$$

We can generalize this idea to a continuously varying pumping rate $Q(t)$:

$$s = \frac{Q(0)}{4\pi T} E_1\left(\frac{r^2 S}{4Tt}\right) + \frac{1}{4\pi T}\int_0^t \left.\frac{\partial Q}{\partial t}\right|_\tau E_1\left(\frac{r^2 S}{4T(t-\tau)}\right) \mathrm{d}\tau \qquad (36)$$

Leakage Through the Confining Layer

Often aquifers are not fully confined, and receive a significant inflow from adjacent beds. Hantush and Jacob (1955) developed the first analytical solution for this situation. Their model considers a confined aquifer overlain by an aquitard and another aquifer (Figure 10). They assume that the pumped aquifer is recharged from the unpumped aquifer through the aquitard. The pumped aquifer is an ideal homogeneous isotropic and infinite two-dimensional aquifer. The flow is assumed to be vertical in the aquitard, there is no storage in the aquitard, the head remains constant in the unpumped aquifer, and the flow remains horizontal in the aquifer. The analytical solution to this problem contains a new dimensionless number related to the aquitard property:

$$\frac{r}{B} = r\sqrt{\frac{k'}{Te'}} \qquad (37)$$

with e' the thickness of the aquitard and k' its hydraulic conductivity. Figure 3(e) shows the typical behavior of the drawdown calculated with the Hantush and Jacob model. After following the Theis solution at early time, the drawdown stabilizes and the derivative drops very fast. If the aquitard is impermeable or very thick or if the observation point is very close to the pumping well, that is, r/B is small, then the solution reaches a plateau at very late time. Conversely, if the aquitard is highly permeable or very thin, or if the observation well is located at a great distance from the pumping well, that is, r/B is large, then the drawdown rapidly stabilizes to a constant value:

$$s_D = K_0\left(\frac{r}{B}\right) \qquad (38)$$

with K_0 the modified Bessel function of the second kind and of the zeroth order.

Later on, improved solutions accounting for storage in the aquitard, complex multilayer systems were developed (Hantush, 1960; Neuman and Witherspoon, 1972). These solutions provide the basic theory allowing indirect testing of low permeability formations.

Unconfined Aquifer

Unconfined aquifers are commonly encountered in well-test analysis. However, radial flow to a well in such an aquifer is a complex physical and mathematical problem. The water table will decline during pumping and therefore the domain for which the equations have to be solved is not constant. Furthermore, the saturated zone is in direct hydraulic relation with the unsaturated zone (Figure 11) and finally, even if the aquifer is horizontal, there is a vertical component of the flow in the vicinity of the pumping well. A complete analytical solution for the saturated–unsaturated system was derived by Kroszynski and Dagan (1975). The main conclusion of their work is that, even if it is more exact to make a complete saturated–unsaturated analysis, the difference in the position of the calculated water table is small if we compare the prediction made by the complete model with the prediction made by a model that does not take into account the unsaturated zone.

The approach most often used is based on the concept of a delayed water-table response. It was initiated by Boulton (1954) and developed by Neuman (1972, 1974). The typical behavior of the drawdown is shown in Figure 3(b). There are three typical stages. In the early time, the drawdown follows a Theis type curve corresponding to the release of water from elastic storage. Then, there is a transition with a flattening of the curve and a hole in the derivative.

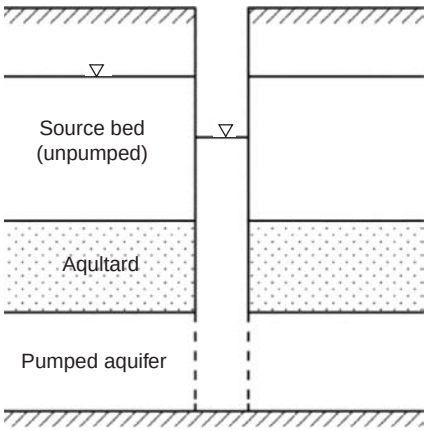


Figure 10 Schematic of Hantush–Jacob model

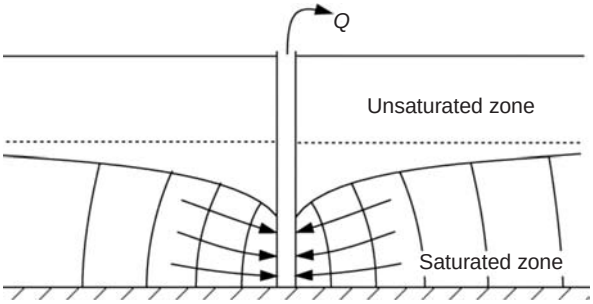


Figure 11 Radial flow to a well in an unconfined aquifer

For late time, the drawdown follows a second Theis curve corresponding to the release of water from the drainage of the unsaturated zone. The derivative becomes a constant. One of the latest model developed for unconfined aquifers is an extension of the Boulton and Neuman models for a large diameter well and partially penetrating the aquifer (Moench, 1997, 1998).

Well-bore Storage Effect

When the radius of the pumping well is large it contains a significant amount of fluid (Figure 12a). During early times, the aquifer's contribution to the total discharge is small (Figure 12b). Most of the water is pumped from the well itself. Later, the aquifer's contribution becomes dominant. Papadopulos and Cooper (1967) developed a model of this situation. Their model is identical to the Theis model for the governing equation, initial condition, and boundary condition at infinity. The difference is the boundary condition at the well that account for finite well radius and water storage in the well. This additional concept requires the definition of an additional dimensionless parameter, the well-bore storage coefficient:

$$C_{\mathrm{D}} = \frac{r_c^2}{2r_{\mathrm{W}}^2 S} \tag{39}$$

with r_c being the radius of the casing and r_{W} the radius of the well screen. Figure 3(f) shows a typical behavior of the drawdown in the pumping well. The main characteristic of this type curve is that for early times, the asymptote is a straight line with unit slope in the log–log plot:

$$s_{\mathrm{w}} = \frac{Q}{\pi r_c^2} t \tag{40}$$

The logarithmic derivative follows the same straight line of unit slope. During this period, when the drawdown and derivative follow the same line, the well-bore storage effect is dominant. This is followed by a transition where the derivative departs from the straight line and makes a hump. For late times, the derivative stabilizes and the behavior is dominated by the aquifer contribution. The drawdown tends toward the Jacob's solution. Note that later on, boundary effects can affect the solution. For example Figure 3(j) shows the same type of behavior with an additional constant head boundary that stabilizes the drawdown for the late time and shows a drop of the derivative.

Skin Effect

The drawdown in the pumping well can be strongly affected by the skin effect, that is, the existence of a zone of lower hydraulic conductivity than the aquifer in the immediate vicinity of the well (Figure 13). This zone of low hydraulic conductivity may be due to poor well development, deposition of particles or development of bacterial films. Also, increased hydraulic conductivity in the immediate vicinity of the well is possible and is accounted for in the theory but have less impact on the drawdowns in the well.

The concept of skin effect was introduced by van Everdingen (1953). Later, Agarwal *et al.* (1970) developed a solution including both well-bore storage and skin effect. To quantify the skin effect and assuming steady state, it is simple to show that the additional drawdown s_{D} in the well due to a cylinder of radius r_s and having a transmissivity T_s is

$$s_{\mathrm{D}} = \frac{Q}{2\pi T}\sigma \tag{41}$$

with σ being the skin factor:

$$\sigma = \left(\frac{T - T_S}{T}\right) \ln\left(\frac{r_s}{r_{\mathrm{w}}}\right) \tag{42}$$

Note that σ is positive if the well is clogged ($T_s < T$) and negative for the opposite case. The early time

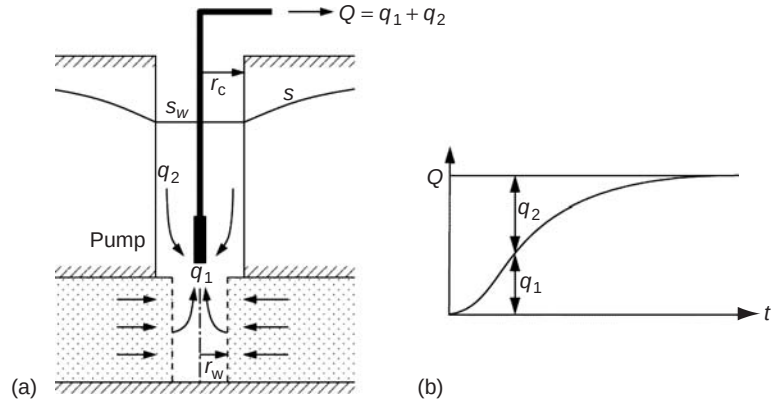


Figure 12 Schematic of the well-bore storage effect

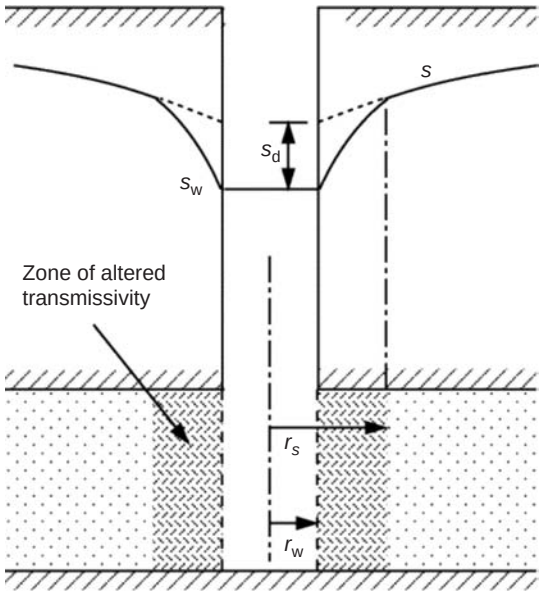


Figure 13 Finite thickness skin model

behavior of the solution of Agarwal *et al.* is identical to the Papadopulos–Cooper solution (1967). The late time asymptote is a straight line (on semilog plot) parallel to Jacob's straight line but shifted by σ.

$$s_D = \tfrac{1}{2}[\ln(4t_D) - \gamma] + \sigma \tag{43}$$

As with the Theis model, the late time data allows unique identification of the transmissivity of the aquifer from the slope of the straight line, or from the value of the logarithmic derivative. However, the position of the line is both a function of storativity and of the skin effect. It is therefore impossible to determine both the storativity and the skin effect with a single well test.

An other important feature of the Agarwal solution is that the shape of the type curves are identical for all type curves having the same value of the product $C_D e^{2\sigma}$ when $\sigma > 0$, and $C_D e^{2\sigma} > 10^3$. Furthermore the shape of these type curves (see Figure 3f) is identical to the original Papadopulos–Cooper type curves only accounting for well-bore storage effects. In practice, this means that the effect of well-bore storage and skin effect are not distinguishable.

Quadratic Head Losses

In the model of Agarwal *et al.* the drawdown in the pumped well increases linearly with Q. However, based on field observations, Jacob (1947) indicates that the drawdown in a pumping well generally increases as a function of the square of the flow rate (Figure 14a).

$$s = AQ + BQ^2 \tag{44}$$

This additional head loss can strongly affect the drawdown in the borehole and therefore its efficiency. The additional drawdown is attributed to inertial or turbulent flow occurring in the zone just outside the well, through the well screen and in the casing (Rorabaugh, 1953). This additional drawdown is commonly referred to as the quadratic head losses or the nonlinear head losses in the well. The *nonlinear head-loss coefficient* B allows quantifying the importance of this effect. In petroleum engineering, this phenomenon is mainly described for gas well testing and is modeled with a rate dependent skin effect. In practice, to evaluate the importance of these head losses, a step-drawdown test is conducted. At first, the well is pumped at a given flow rate for a given amount of time (Figure 14b); subsequently the flow rate is modified and applied for another given duration; then again the flow rate is changed and and so on. Usually the minimum number of steps is three, allowing to identify the nonlinear head-loss coefficient and to check the adequacy of equation (44). The interpretation of such step-drawdown tests relies on the late time asymptote and on the superposition principle (Eden and Hazel, 1973; Hantush, 1964). Some attempts have been made to include the inertial term in a transient solution

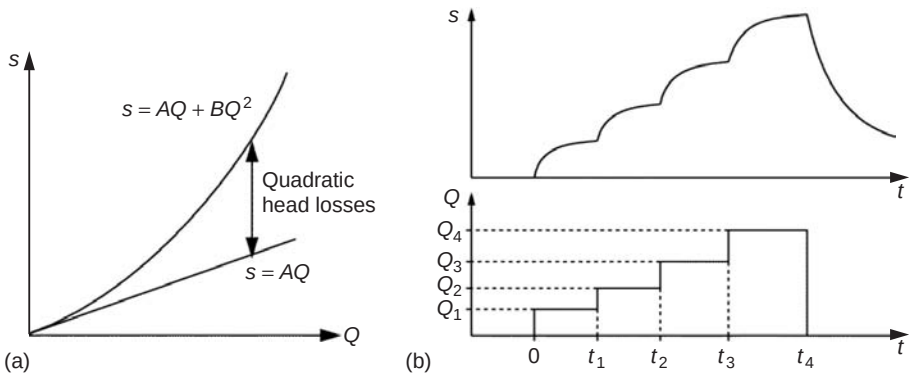


Figure 14 (a) Drawdown in the well as a function of the discharge rate. (b) A step-drawdown test

(Chachadi and Mishra, 1992), but the analytical approach is not yet fully convincing as it does not respect the continuity of heads throughout the well screen. The most promising approach used to date is numerical. It includes the inertial term within the aquifer and has been applied to a deep fractured aquifer (Kohl *et al.*, 1997).

The Double Porosity Model

In 1960, Barenblatt and his coworkers introduced a revolutionary concept:

> "Unlike in classical fluid flow theory, for each point in space, not one hydraulic head but two, h and h', are introduced. The head h represents the average head in the fractures in the neighborhood of the given point, whereas [...] h' is the average head in the matrix in the neighborhood of the given point" (Barenblatt *et al.*, 1960). This concept forms the foundation of the double porosity approach. In their article, Barenblatt *et al.* assumed that the storativity of the fractures was negligible. Warren and Root (1963) introduced a storativity for the fractures and developed a model, which is still used today.

The formulation of the double porosity models (see Figure 15) is based on two coupled standard groundwater flow equations (one for the fracture, one for the matrix) with an exchange term. The hydraulic conductivity and the specific storage coefficient are defined separately for each media. To simplify the system of equations, Warren and Root assume that the water moves from matrix blocks to fractures, but not from matrix block to matrix block. It is furthermore assumed that the flow rate between the matrix and the fractures is proportional to the hydraulic conductivity of the matrix, to the hydraulic head differences between the two systems, and to a geometrical factor depending on the size and the shape of the matrix blocks. This is the so-called pseudosteady state assumption of Warren and Root.

A typical drawdown curve is shown in Figure 3(b). It has a sigmoidal shape. During early time, the water is pumped from storage in the fractured system, the matrix does not affect the flow. In the intermediate times, water is released from the matrix while the drawdown in the matrix is small compared to drawdown in the fractures. During the late time, the drawdown in the matrix approaches the drawdown in the fractures and the aquifer behaves like a single porosity aquifer with the combined property of the matrix and the fractures. The log derivative shows a typical depression before it stabilizes to the constant value indicating that the drawdown reached the late time Jacob's straight line.

More recently, the model of Warren and Root has been extended to account for well-bore storage and skin effect (Bourdet and Gringarten, 1980). Another modification of the model is to account for transient flow within the matrix to the fractures (Boulton and Streltsova, 1977). In this model, the head distribution is solved within the blocks and therefore it is necessary to assume a given shape for the blocks. Boulton and Streltsova consider a representation of the fractured medium as alternating layers of matrix rocks and fractures. Once again, the resulting type curves show the typical sigmoidal shape and depression in the logarithmic derivative.

General Radial-Flow Model

The General Radial-Flow (GRF) model (Barker, 1988) is a generalization of radial-flow equations to any flow dimension n that includes the specific cases of linear flow ($n = 1$), usual cylindrical flow ($n = 2$), and spherical flow ($n = 3$), it extends as well to noninteger-flow dimensions for intermediate cases (see Figure 16). The beauty of this model is that it provides a unique solution for a large range of possible behaviors with only one additional parameter (the flow dimension n). In subsequent publications, the GRF model was extended to include well-bore storage, skin effects, and double porosity (Hamm and Bideaux, 1996). The solution of the standard GRF model (without skin, well-bore storage and double porosity) with a constant pumping rate in an infinite medium is

$$s_D = \frac{r_D^{2-n}}{\Gamma\left(\frac{n}{2}\right)}\Gamma\left(\frac{n}{2} - 1, \frac{t_D}{4r_D^2}\right) \tag{45}$$

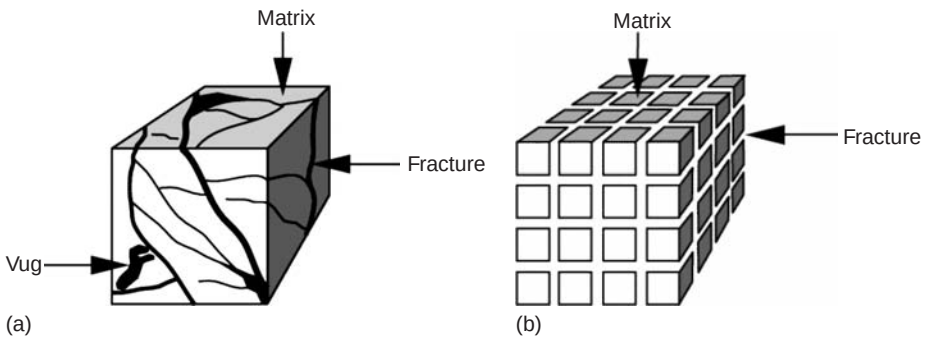


Figure 15 A fractured block illustrating the double porosity concept (Adapted from Cinco-Ley, 1996)

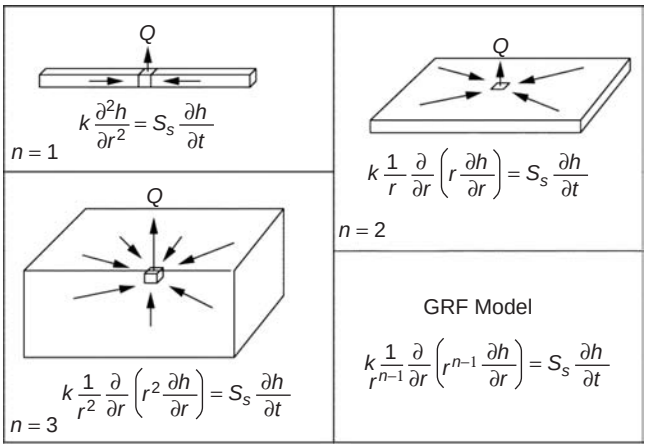


Figure 16 Concept of radial flow in 1, 2, 3, and generalization to n dimensions

with n being the flow dimension, $\Gamma(x)$ the gamma function, and $\Gamma(a, x)$ the incomplete gamma function. Figure 3(h) and Figure 3(i) show two examples of drawdown for $n < 2$ and $n > 2$ respectively. A typical characteristic of the GRF model is that the log derivative of the drawdown follows a straight line with a slope of $1 - n/2$ in the diagnostic plot. When $n = 2$, the GRF model converges toward the Theis solution, and the log derivative becomes constant for late time (0 slope). Barker suggests that the fractional-flow dimension may be related to the observed fractal properties of fracture networks. However, the diffusivity of the GRF model remains equal to k/S_s even when the flow dimension is fractional. An alternative analytical model based on a fractal network of fractures has been proposed by Chang and Yortsos (1990). Both models have been compared with numerical simulations in fracture networks. Acuna and Yortsos (1995) show that the model of Chang and Yortsos agrees with drawdowns simulated on artificially generated fractal networks (e.g., Sierpinski carpets). More recently, Jourde *et al.* (2002) show that fractional-flow dimension may even appear on a Euclidian network by using a network of pipes simulating a network of pseudorandom orthogonal fractures.

Individual Fractures

A more specific situation is the case of an individual fracture intersected by the well and acting as a drain in a larger porous aquifer. Several models have been developed for this type of configuration (Gringarten, 1982). The case of the drawdown in a well intersecting a vertical fracture of infinite conductivity and finite length is illustrated in Figure 3(g) (Gringarten *et al.*, 1974). A 0.5 slope for the drawdown and the derivative characterizes this solution at early time on a log–log plot.

OTHER HYDRAULIC PERTURBATIONS

Recovery Test

Head recovery after a pumping test can be modeled using the superposition principle. When the aquifer is confined, from a mathematical point of view, the end of pumping is equivalent to continuing pumping and simultaneously injecting the same amount. This implies that the drawdown can be calculated by adding drawdown due to pumping plus drawdown due to a simultaneous injection at the same location. In the case of the Theis assumptions, the solution for the recovery is therefore

$$s_{\mathrm{D}} = \frac{1}{2}E_1\left(\frac{r_{\mathrm{D}}^2}{4(t_{\mathrm{P}} + t_{\mathrm{R}})}\right) - \frac{1}{2}E_1\left(\frac{r_{\mathrm{D}}^2}{4t_{\mathrm{R}}}\right) \qquad (46)$$

where t_{R} is the dimensionless time since the recovery started and t_{P} is the dimensionless production time. Equation (46) is exact and can be used to analyze a data set and to estimate both T and S. An asymptotic solution is however often used. When both exponential integrals, in equation (46), can be approximated by a logarithmic function, one finds the late time asymptote:

$$s_{\mathrm{D}} = \frac{1}{2}\ln\left(\frac{t_{\mathrm{R}} + t_{\mathrm{P}}}{t_{\mathrm{R}}}\right) \quad \text{or} \quad s = \frac{2.30Q}{4\pi T}\log\left(\frac{t_{\mathrm{r}} + t_{\mathrm{p}}}{t_{\mathrm{r}}}\right) \qquad (47)$$

with t_{r} being the time since the beginning of the recovery and t_{p} the production time. This approximation allows rapid interpretation of late time recovery by constructing a semilog plot of the drawdown versus the log of the ratio $(t_{\mathrm{r}} + t_{\mathrm{p}})/t_{\mathrm{r}}$, denoted as the Horner time in the oil industry. Although, the slope is then proportional to the transmissivity, the storativity cannot be estimated. Another alternative to interpret recovery data has been proposed by Agarwal (1980). He found that if the recovery time is replaced by a corrected time t_a:

$$t_a = \frac{t_{\mathrm{p}}t_{\mathrm{r}}}{t_{\mathrm{p}} + t_{\mathrm{r}}} \qquad (48)$$

The interpretation is thus greatly facilitated. One can apply the standard interpretation models above described for constant rate pumping tests, including the log derivative for model identification. Note, however, that the approach is not valid for bounded aquifers.

Constant Head Test

In some situations, the artificial perturbation imposed in the well is not a constant discharge or recharge rate but a constant head. Under these circumstances, the aquifer responds by a groundwater discharge at a variable rate into the well, and a variation of head within the aquifer.

Both perturbations can be modeled. Considering an ideal confined aquifer, the standard analytical solution for the discharging rate in the well is the solution of Jacob and Lohman (1952). Mishra and Guyonnet (1992) indicate that using the drawdown normalized by the discharge rate allows field data to be interpreted with the usual Theis model. The discharge rate in the well can be evaluated with excellent accuracy using the Perrochet approximation:

$$q_D = \frac{q}{2\pi T s_0} = \frac{1}{\ln(1 + \sqrt{\pi t_D})} \tag{49}$$

with s_0 being the imposed drawdown at the well and with relative error less than 1% over the range $10^{-4} < t_D < 10^{12}$.

Additional solutions for the constant head case accounting for boundaries and transient effects are available in Murdoch and Franco (1994).

Constant head situations are naturally encountered in artesian aquifers. Constant head tests offer a useful alternative to pumping tests since they minimize the effect of well-bore storage and therefore reduce the required duration of a test. Another advantage is that constant head tests can be carried out when the maximum admissible drawdown may be limited. This can be the case when the hydraulic conductivity is small.

Slug Test

The slug test method (Hvorslev, 1951) was developed to rapidly estimate aquifer transmissivity. The principle involves instantaneously perturbing the water level in a well by an injection or an extraction of a known volume of water (Figure 17). After the perturbation, the water levels are recorded in the well or in a piezometer until they stabilize.

Depending on the geometry and type of aquifer, several analytical solutions can be used for slug-test data interpretation. Practical recommendations concerning the design and the performance of slug tests as well as a compendium of available analytical solutions is available in Butler (1998). Recent contributions include a detailed analysis of inertial effects, given the rapid and often oscillatory behavior of the head in the borehole itself (McElwee and Zenner, 1998; Zurbuchen *et al.*, 2002). From a general point of view, slug tests are very interesting because they take much less time to carry out than pumping tests or constant head tests and require less equipment. Consequently, they are cheaper to conduct. One can therefore perform a large number of slug tests in an aquifer in order to characterize the spatial distribution of hydraulic conductivities (see e.g. Mas-Pla *et al.*, 1997). Slug tests also allow testing low permeability formations that may not be tested within a reasonable period of time using pumping test techniques. The main disadvantage of slug tests is that the radius of investigation is much smaller than for a pumping test. As a consequence, they are much more strongly influenced by skin effects or local heterogeneities around the well (Butler and Healey, 1998). Another drawback of this technique is that it is difficult to estimate the storativity coefficient with sufficient certainty.

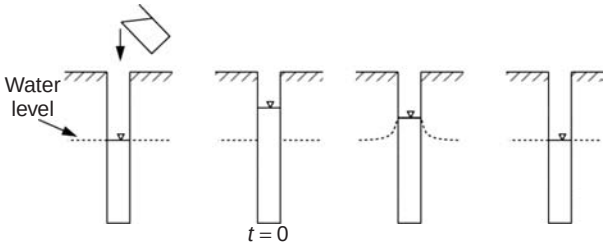


Figure 17 Slug-test procedure

REFERENCES

Abramowitz M. and Stegun I.A. (1970) *Handbook of Mathematical Functions*, National Bureau of Standards: Washington.

Acuna J.A. and Yortsos Y.C. (1995) Application of fractal geometry to the study of networks of fractures and their pressure transient. *Water Resources Research*, **31**, 527–540.

Agarwal R.G. (1980) A new method to account for producing time effects when drawdown type curves are used to analyze pressure buildup and other test data. Paper presented at the *55th Annual Fall Technical Conference and Exhibition of the Society of Petroleum Engineers of AIME*, Dallas.

Agarwal R.G., Al-Hussainy R. and Ramey H.J.J. (1970) An investigation of wellbore storage and skin effect in unsteady liquid flow. *SPE Journal*, **10**, 279–290.

Barenblatt G.I., Zheltov I.P. and Kochina I.N. (1960) Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks (strata). *Journal of Applied Mathematics and Mechanics*, **24**, 1286–1303.

Barker J.A. (1988) A Generalized radial flow model fro hydraulic tests in fractured rock. *Water Resources Research*, **24**, 1796–1804.

Batu V. (1998) *Aquifer Hydraulics: A Comprehensive Guide to Hydrogeologic Data Analysis*, John Wiley & Son.

Beckie R. (2001) A comparison of methods to determine measurement support volumes. *Water Resources Research*, **37**, 925–936.

Boulton N.S. (1954) Unsteady radial flow to a pumped well allowing for delayed yield from storage. *AIHS Publications*, **37**, 472–477.

Boulton N.S. and Streltsova T.D. (1977) Unsteady flow to a pumped well in a fissured water-bearing formation. *Journal of Hydrology*, **35**, 257–269.

Bourdet D. (2002) *Well Test Analysis*, Elsevier.

Bourdet D., Ayoub J.A. and Pirard Y.M. (1989) Use of pressure derivative in well-test interpretation. *SPE Formation Evaluation*, **4**(2), 293–302.

Bourdet D. and Gringarten A. (1980) Determination of fissured volume and block size in fractured reservoirs by type-curve analysis. Paper presented at the *55th Annual Technical Conference and Exhibition of the Society of Petroleum Engineers*, Dallas.

Bourdet D., Whittle T.M., Douglas A.A. and Pirard Y.M. (1983) A new set of type curves simplifies well test analysis. *Word Oil*, **196**, 95–106.

Butler J.J. (1998) *The Design, Performance, and Analysis of Slug Tests*, Lewis Publishers.

Butler J.J. and Healey J.M. (1998) Relationship between pumping-test and slug-test parameters: Scale effect or artifact? *Ground Water*, **36**, 305–313.

Butler J.J., Zlotnik V.A. and Tsou M.-S. (2001) Drawdown and stream depletion produced by pumping in the vicinity of a partially penetrating stream. *Ground Water*, **39**, 651–659.

Chachadi A.G. and Mishra G.C. (1992) Analysis of unsteady flow to a large well experiencing well loss. *Ground Water*, **30**, 369–375.

Chang J. and Yortsos Y.C. (1990) Pressure transient analysis of naturally fractured reservoirs. Paper presented at the *SPE Annual Fall Meeting, Society of Petroleum Engineering*, Washington.

Chow V.T. (1952) On the determination of transmissibility and storage coefficients from pumping test data. *Transactions of the American Geophysical Union*, **33**, 397–404.

Cinco-Ley H. (1996) Well-test analysis for naturally fractured reservoirs. *Journal of Petrolem Technology*, 51–54.

Cooper H.H.J. and Jacob C.E. (1946) A generalized graphical method for evaluating formation constants and summarizing well fiel history. *Transactions of the American Geophysical Union*, **27**, 526–534.

Dagan G. (1982) Stochastic modeling of groundwater flow by unconditional and conditional probabilities: 1. Conditional simulation and the direct problem. *Water Resources Research*, **18**, 813–833.

Dawson K.J. and Istok J.D. (1991) *Aquifer Testing*, Lewis Publishers.

Dougherty D.E. (1989) Computing well hydraulics solutions. *Ground Water*, **27**, 564–569.

Dupuit J. (1863) *Etude Théoriques et Pratiques Sur le Mouvement Des Eaux Dans Les Canaux Découverts et à Travers Les Terrains Perméables*, Dunot: Paris.

Earlougher R.C.J. (1977) *Advances in Well Test Analysis*, Society of Petroleum Engineers of AIME.

Eden R.N. and Hazel C.P. (1973) Computer and graphical analysis of variable discharge pumping test of wells. Paper presented at the *Institute of Engineers Australia, Civil Engineering Transactions*, Vol. 15, no. (1-2), pp 5–10.

Gringarten A. (1982) Flow-test evaluation of fractured reservoirs. *Geological Society of America*, **Special Paper 189**, 237–263.

Gringarten A., Ramey H.J.J. and Raghavan R. (1974) Unsteady-state pressure distributions created by a single infinite conductivity vertical fracture. *Society of Petroleum Engineers Journal*, **14**, 347.

Hamm S.-Y. and Bideaux P. (1996) Dual-porosity fractal models for transient flow analysis in fractured rocks. *Water Resources Research*, **32**, 2733–2745.

Hantush M.S. (1960) Modification of the theory of leaky aquifers. *Journal of Geophysical Research*, **65**, 3713–3725.

Hantush M.S. (1961) Aquifer tests on partially penetrating wells. *Proceedings of the American Society of Civil Engineers*, **87**, 171–195.

Hantush M.S. (1964) *Hydraulics of Wells*, Academic Press.

Hantush M.S. and Jacob C.E. (1955) Nonsteady radial flow in an infinite leaky aquifer. *Transactions of the American Geophysical Union*, **36**, 95–100.

Horne R. (1995) *Modern Well Test Analysis*, *Second Edition*, Petroway, Inc.: Palo Alto.

Hvorslev M.J. (1951) Time lag and soil permeability in ground water observations. *U.S. Army Corps of Engineers Waterway Experimentation Station Bulletin*, **36**, 1–50.

Jacob C.E. (1947) Drawdown test to determine effective radius of artesian well. *American Society of Civil Engineers, Transactions*, **112**, 1047–1064.

Jacob C.E. and Lohman S.W. (1952) Non steady flow to a well of constant drawdown in an extensive aquifer. *Transactions of the American Geophysical Union*, **33**, 559–569.

Jourde H., Pistre S., Perrochet P. and Drogue C. (2002) Origin of fractional flow dimension to a partially penetrating well in stratified fractured reservoirs. New results based on the study of synthetic fracture network. *Advances in Water Resources*, **25**, 371–387.

Kohl T., Evans K.F., Hopkirk R.J., Jung R. and Rybach L. (1997) Observation and simulation of non-Darcian flow transients in fractured rocks. *Water Resources Research*, **33**, 407–418.

Kroszynski U.I. and Dagan G. (1975) Well pumping in unconfined aquifers: The influence of the unsaturated zone. *Water Resources Research*, **11**, 479–490.

Kruseman G.P. and Ridder N.A.d (1992) *Analysis and Evaluation of Pumping Test Data*, Vol. 47, ILRI publication.

Lavenue M. and de Marsily G. (2001) Three-dimensional interference test interpretation in a fractured aquifer using the pilot point inverse method. *Water Resources Research*, **37**, 2659–2675.

Lebbe L.C. (1999) *Hydraulic Parameter Identification*, Springer Verlag.

Lee T.-C. (1999) *Applied Mathematics in Hydrogeology*, Lewis Publishers.

Mas-Pla J., Yeh T.-C.J., Williams T. and McCarthy J.F. (1997) Analyses of slug test and hydraulic conductivity variations in the near field of a two-well tracer experiment site. *Ground Water*, **35**, 492–501.

McElwee C.D. (1980) Theis parameter evaluation from pumping tests by sensitivity analysis. *Ground Water*, **18**, 56–60.

McElwee C.D. and Zenner M.A. (1998) A nonlinear model for analysis of slug-test data. *Water Resources Research*, **34**, 55–66.

Meier P.M., Carrera J. and Sánchez-Vila X. (1998) An evaluation of Jacob's method for the interpretation of pumping tests in heterogeneous porous media. *Water Resources Research*, **34**, 1011–1025.

Mishra S. and Guyonnet D. (1992) Analysis of observation-well response during constant-head testing. *Ground Water*, **30**, 523–528.

Moench A.F. (1997) Flow to a well of finite diameter in a homogeneous, anisotropic water table aquifer. *Water Resources Research*, **33**, 1397–1407.

Moench A.F. (1998) Correction to "Flow to a well of finite diameter in a homogeneous, anisotropic water table aquifer". *Water Resources Research*, **34**, 2431–2432.

Moench A.F. and Ogata A. (1984) Analysis of constant discharge wells by numerical inversion of Laplace transform. In *Ground-Water Hydraulics*, Vol. 9, Rosensheim J.S. and Bennets G.D. (Eds.), American Geophysical Union, pp. 146–170.

Murdoch L.C. and Franco J. (1994) The analysis of constant drawdown wells using instantaneous source functions. *Water Resources Research*, **30**, 117–124.

Neuman S.P. (1972) Theory of flow in unconfined aquifers considering delay response of the watertable. *Water Resources Research*, **8**, 1031–1045.

Neuman S.P. (1974) Effect of partial penetration on flow in unconfined aquifers considering delayed gravity response. *Water Resources Research*, **10**, 303–312.

Neuman S.P. and Witherspoon P.A. (1972) Field determination of the hydraulic properties of leaky multiple aquifer systems. *Water Resources Research*, **8**, 1284–1298.

Oliver D.S. (1993) The influence of nonuniform transmissivity and storativity on drawdown. *Water Resources Research*, **29**, 169–178.

Papadopulos I.S. and Cooper H.H.J. (1967) Drawdown in a well of large diameter. *Water Resources Research*, **3**, 241–244.

Pinder G.F. and Bredehoeft J.D. (1968) Application of digital computers for aquifer evaluation. *Water Resources Research*, **4**, 1069–1093.

Raghavan R. (1993) *Well Test Analysis*, Prentice Hall.

Rorabaugh M.I. (1953) Graphical and theoretical analysis of step-drawdown test of artesian well. *Proceedings of the American Society of Civil Engineers*, **362**, 23.

Rosa A.J. and Horne R.N. (1991) Automated well test analysis using robust (LAV) non linear parameter estimation. Paper presented at the *66th Annual Technical Conference and Exhibition of the Society of Petroleum Engineers*, Dallas.

Sánchez-Vila X., Meier P.M. and Carrera J. (1999) Pumping tests in heterogeneous aquifers: An analytical study of what can be obtained from their interpretation using Jacob's method. *Water Resources Research*, **35**, 943–952.

Silliman S.E. and Caswell S. (1998) Observations of measured hydraulic conductivity in two artificial, confined aquifers with boundaries. *Water Resources Research*, **34**, 2203–2213.

Streltsova T.D. (1988) *Well Testing in Heterogeneous Formation*, John Wiley & Sons.

Theis C.V. (1935) The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using groundwater storage. *Transactions of the American Geophysical Union*, **2**, 519–524.

Thiem G. (1906) *Hydrologische Methoden*, Gebhardt: Leipzig, p. 56.

van Everdingen A.F. (1953) The skin effect and its influence on the productive capacity of a well. *Transactions of AIME*, **198**, 171–176.

Van Poolen H.K. (1964) Radius-of-drainage and stabilization-time equation. *The Oil and Gas Journal*, **62**, 138–146.

Walton W.C. (1996) *Aquifer Test Analysis with Windows Software*, Lewis Publishers.

Warren J.E. and Root P.J. (1963) The behaviour of naturally fractured reservoirs. *Society of Petroleum Engineers Journal*, **3**, 245–255.

Zurbuchen B.R., Zlotnik V.A. and Butler J.J.J. (2002) Dynamic interpretation of slug tests in highly permeable aquifers. *Water Resources Research*, **38**. 10.1029/2001WR000354.

152: Modeling Solute Transport Phenomena

JACOB BEAR

Department of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa, Israel

The article presents the conceptual and well-posed mathematical models that describe the transport of solutes, or chemical species, in porous media, as encountered when considering the transport of dissolved salts and contaminants in aquifers. The various modes of transport: advection dispersion and diffusion are discussed and incorporated in the models. Solute sources are also incorporated in the models, including adsorption and an introduction to reactive transport. We shall emphasize that the fluid's velocity, and in unsaturated flow, also the fluid's saturation is required as input information to any solute transport problem. However, the model of flow problem that provides this information is discussed in ***Chapter 149, Hydrodynamics of Groundwater, Volume 4****. Unsaturated flow models are discussed in* ***Chapter 150, Unsaturated Zone Flow Processes, Volume 4****. The case of solute transport with variable density is not included, as it is presented in* ***Chapter 157, Sea Water Intrusion Into Coastal Aquifers, Volume 4****. More about chemical and biological reactions is presented in* ***Chapter 153, Groundwater Pollution and Remediation, Volume 4****.*

INTRODUCTION

The water *quality* aspect is essential in the management of any water resources system, in parallel to that of water *quantity*. In fact, in many parts of the world, with the increased demand for water and with the intensification of water utilization, the quality of water in the sources often deteriorates to the extent that it becomes the limiting factor in the development and management of the water resources system. Even under pristine conditions, ground water in springs, rivers, and aquifers always contains dissolved matter, especially dissolved salts. Sometimes, the concentration of solutes is such that the water is unfit for certain purposes, for example, for irrigation or for drinking. Unfortunately, too often, water contains (dissolved) *pollutants, or contaminants,* which are produced and released to the environment by human activities at ground surface, whether intentionally or not. Note that to emphasize the negative, often hazardous to human health, aspects of certain dissolved materials, or *chemical species* in solution, we usually refer to them as *pollutants* or *contaminants* rather than as merely *solutes.* In this article, we shall use these terms interchangeably. In **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**, we consider salinization of ground water associated with the phenomenon of seawater intrusion into coastal aquifers.

Although the processes of water *salinization* and *contamination* occur in both surface and ground water, special attention should be devoted to their occurrence in aquifers, because of the very slow velocity of ground water, and the capacity of the solid matrix to adsorb and absorb pollutants, and then release them over prolonged periods of time. Cleanup of a contaminated aquifer (and the vadose zone) is much more lengthy and costly.

Certain chemical species in groundwater may chemically interact with other species, and/or with the solid matrix. Others may undergo biological transformations. In this article, these processes are briefly introduced.

The issue of ground water pollution, including sources of pollutants, quality standards, and regulations, is presented in **Chapter 153, Groundwater Pollution and Remediation, Volume 4**. Here, we shall focus only on the models that describe the *transport*, that is, the movement, accumulation, and transformation of solutes in ground water.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

The entire discussion will be at the *macroscopic* level, that is, the level at which the porous medium comprising the aquifer is regarded as a *continuum*, and the behavior of the aquifer system is described in terms of state variables that represent *intrinsic phase averages* of their corresponding *microscopic* ones over a Representative Average Volume (REV) (e.g. Bear, 1972; Bear and Bachmat, 1990). The objective of constructing and solving such models is to enable planners and managers to predict future changes in the concentration of solutes, in response to planned management decisions. Such predictions are required in connection with aquifer management, in dealing with aquifer pollution and remediation techniques (*see* **Chapter 153, Groundwater Pollution and Remediation, Volume 4**), and in the use of tracers. The subjects of protecting ground water resources against pollution, and monitoring the movement of solutes in aquifers are discussed in **Chapter 153, Groundwater Pollution and Remediation, Volume 4**.

FLUXES

We consider the transport of a contaminant (= a chemical species) within a fluid phase that occupies the entire void space. Three modes of transport will be considered: advection, diffusion, and dispersion.

The quantity of the chemical species in solution will be represented as concentration (expressed as mass of species per unit fluid phase volume).

Advective Flux – Macroscopic Level

The (macroscopic) *advective flux*, $\mathbf{J}^{\gamma}_{\text{adv}}$, of a chemical γ-species (= mass of γ per unit area of fluid per unit time), is expressed in the form:

$$\mathbf{J}^{\gamma}_{\text{adv}} = \mathbf{V}c^{\gamma} \tag{1}$$

where c^{γ} denotes the concentration of the considered γ-species, and $\mathbf{V}$ denotes the (mass-averaged) velocity of the fluid phase, described by Darcy's law (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4** for saturated flow and **Chapter 150, Unsaturated Zone Flow Processes, Volume 4** for unsaturated and multiphase flow).

Per unit area of porous medium, the macroscopic advective flux is expressed by $\theta\mathbf{J}^{\gamma}_{\text{adv}} = \theta\mathbf{V}c^{\gamma}$, where θ denotes the volumetric fraction of the fluid (with $\theta = \phi =$ porosity for saturated flow).

Diffusive Flux

Diffusive Flux – Microscopic Level

We consider the *(microscopic) diffusive flux*, $\mathbf{j}^{\gamma}_{\text{diff}}$ of a γ-species. A fluid phase is, usually, composed of a number of *chemical species*, each made up of a large number of identical molecules (ions, atoms, etc.) that are in constant random motion. The *solvent* is considered as one of the chemical species. At the *microscopic level*, that is, at a point *within* the fluid that occupies the void space (or part of it), each extensive quantity (mass, momentum, energy) of a chemical species, or of a phase, present in a given domain, may be regarded as a *continuum*. At the microscopic level, an extensive quantity of a chemical species at a (microscopic point) is the average of that quantity over the individual molecules of that species within a microscopic representative elementary volume centered at that point.

Let $\rho^{\gamma}(\equiv c^{\gamma})$ and $\mathbf{V}^{\gamma}$ denote the density and velocity of a γ-species *at the microscopic level*. The flux of the γ-species (= mass per unit area of fluid per unit time) is given by $\rho^{\gamma}\mathbf{V}^{\gamma}$. Because $\mathbf{V}^{\gamma}$ is not known, we express this flux as the sum of two fluxes:

$$\rho^{\gamma}\mathbf{V}^{\gamma} \equiv \rho^{\gamma}\mathbf{V} + \rho^{\gamma}(\mathbf{V}^{\gamma} - \mathbf{V}) \equiv \rho^{\gamma}\mathbf{V} + \mathbf{j}^{\gamma}_{\text{diff}},$$
$$\mathbf{j}^{\gamma}_{\text{diff}} \equiv \rho^{\gamma}(\mathbf{V}^{\gamma} - \mathbf{V}) \tag{2}$$

that is, an advective flux (at the microscopic fluid velocity $\mathbf{V}$), $\rho^{\gamma}\mathbf{V}$, and a *diffusive flux*, $\mathbf{j}^{\gamma}_{\text{diff}}$, at a velocity relative to the latter, both at the microscopic level.

A *binary fluid*, is composed of two chemical species, a *solute*, γ, and a *solvent*, δ. In such a fluid, the (microscopic mass) flux of molecular diffusion is expressed by *Fick's law of molecular diffusion*:

$$\mathbf{j}^{\gamma}_{\text{diff}} \equiv \rho^{\gamma}(\mathbf{V}^{\gamma} - \mathbf{V}) = -\rho\mathcal{D}^{\gamma\delta}\nabla\omega^{\gamma} \tag{3}$$

where $\omega^{\gamma}(\equiv\rho^{\gamma}/\rho)$ denotes the mass fraction of γ, and $\mathcal{D}^{\gamma\delta}$ (a scalar often assumed to be independent of c^{γ}) is the *coefficient of molecular diffusion* (dims. L^2/T) of a γ-species in a fluid phase containing a second species, δ. The diffusive flux of the δ-species, is given by

$$\mathbf{j}^{\delta}_{\text{diff}} = \rho^{\delta}(\mathbf{V}^{\delta} - \mathbf{V}) = -\rho\mathcal{D}^{\delta\gamma}\nabla\omega^{\delta} \tag{4}$$

with $\mathcal{D}^{\gamma\delta} = \mathcal{D}^{\delta\gamma}$. For a homogeneous fluid, $\nabla\rho = 0$, and, therefore,

$$\mathbf{j}^{\gamma}_{\text{diff}} = -\mathcal{D}^{\gamma\delta}\nabla c^{\gamma} \tag{5}$$

Another form of the diffusive flux is

$$\mathbf{j}^{\gamma}_{\text{diff}} = -\frac{\hat{\rho}^2}{\rho}M^{\gamma}M^{\delta}\mathcal{D}^{\gamma\delta}\nabla n^{\gamma} \tag{6}$$

where $\hat{\rho}$ is the total *molar density* of the fluid (= number of moles of fluid per unit volume of fluid), n^γ is the *molar fraction* of γ, $\rho^\gamma/\hat{\rho} = n^\gamma M^\gamma + n^\delta M^\delta$, and $\omega^\gamma = n^\gamma M^\gamma/(n^\gamma M^\gamma + n^\delta M^\delta)$. More generally, the diffusive flux is driven by a gradient in the *chemical potential* (Denbigh, 1981), $\mu^\gamma = \mu^\gamma(p, n^\gamma, T)$. It then takes the form:

$$\mathbf{j}^\gamma_{\text{diff}} = -\frac{\hat{\rho}^2}{\rho RT} n^\gamma \mathcal{D}^{\gamma\delta} \nabla \mu^\gamma \big|_{p,T} = \left(\frac{\hat{\rho}^2}{\rho RT} n^\delta \mathcal{D}^{\delta\gamma} \nabla \mu^\delta \big|_{p,T} \right) \tag{7}$$

where R and T denote the universal gas constant and the temperature, respectively. Typical values of $\mathcal{D}^{\gamma\delta}$ at 25 °C, for a solute in an aqueous phase, are in the range of $5\text{–}100 \times 10^{-6}\,\text{cm}^2\,\text{s}^{-1}$. For example, for Ca^{2+} : $\mathcal{D}^{\gamma\delta} = 7.9 \times 10^{-6}\ \text{cm}^2\,\text{s}^{-1}$, for K^+ : $\mathcal{D}^{\gamma\delta} = 30 \times 10^{-6}\,\text{cm}^2\,\text{s}^{-1}$, for Cl^- : $\mathcal{D}^{\gamma\delta} = 20.3 \times 10^{-6}\,\text{cm}^2\,\text{s}^{-1}$. Typical values for a dilute chemical component in the air are: for water vapor: $\mathcal{D}^{\gamma\delta} = 2.2 \times 10^{-1}\ \text{cm}^2\,\text{s}^{-1}$, for TCE vapor: $\mathcal{D}^{\gamma\delta} = 7.8 \times 10^{-2}\text{cm}^2\,\text{s}^{-1}$. Correlations for $\mathcal{D}^{\gamma\delta}$ for a broad range of compounds, as a function of temperature and pressure, are given by Poling *et al.* (2000). Fick's law, equation (3), is also valid, as an approximation, for the diffusive flux of a γ-species in a *multicomponent system*, as long as the other component, δ, is the solvent, and all components, except δ and γ, are dilute. Another case for which Fick's law holds is when all components are dilute, except for the δ-component. Then, since $\Sigma_{(\gamma)} \mathbf{j}^\gamma_{\text{diff}} = 0$, we have

$$\begin{aligned} \mathbf{j}^\lambda_{\text{diff}} &= -\rho \mathcal{D}^{\lambda\delta} \nabla \omega^\lambda, \lambda \neq \delta \\ \mathbf{j}^\delta_{\text{diff}} &= -\sum_{\gamma(\neq\delta)} \mathbf{j}^\gamma_{\text{diff}} = \sum_{\gamma(\neq\delta)} \rho \mathcal{D}^{\gamma\delta} \nabla \omega^\gamma \end{aligned} \tag{8}$$

Lichtner (1996) presents an expression for the diffusive mass flux of an ionic species in a dilute aqueous solution that is electrically neutral; it is used for modeling the transport of multiple ionic species.

Diffusive Flux – Macroscopic Level

The expression for the macroscopic diffusive flux of a species is obtained by averaging the microscopic one over the appropriate REV, or by employing some *homogenization* technique. In the passage from the microscopic equation to its macroscopic counterpart, the configuration of the solid–fluid interface surface, and conditions on it, affect the transformation of the (local) gradient of concentration into a gradient of the average concentration (which is the state variable at the macroscopic level).

Bear and Bachmat (1990), who use volume (REV) averaging, obtained an expression for the *macroscopic* form of Fick's law for a chemical γ-species in a constant density fluid that occupies part of the void space (volumetric fraction θ), in the form:

$$\mathbf{J}^\gamma_{\text{diff}} \equiv \overline{\mathbf{j}^\gamma_{\text{diff}}} = -\mathcal{D}^\gamma \mathbf{T}^*(\theta) \cdot \nabla c^\gamma = -\boldsymbol{\mathcal{D}}^{*\gamma}(\theta) \cdot \nabla c^\gamma \tag{9}$$

where $c^\gamma (\equiv \overline{c^\gamma})$ denotes the component's concentration at the macroscopic level, and $\boldsymbol{\mathcal{D}}^{*\gamma}(\theta) (\equiv \mathcal{D}^\gamma \mathbf{T}^*(\theta))$, a second rank symmetric tensor, is the coefficient of molecular diffusion within a fluid phase *in a porous medium*. Note that $\mathbf{J}^\gamma$ denotes the flux of γ per *unit area of fluid within a porous medium cross-section*. For brevity, we have dropped the superscript δ in $\mathcal{D}^\gamma(\theta)$, and the overbar in the averaged concentration. The symbol $\mathbf{T}^*(\theta)$, a second rank symmetric tensor, represents the *tortuosity* of the porous medium (e.g. Bear and Bachmat, 1990). In an isotropic porous medium, the components, T^*_{ij} of the tortuosity tensor, may be represented as $T^*\delta_{ij}$, in which $T^*(<1)$ is a scalar tortuosity, and δ_{ij} is the ijth component of the *Kronecker delta,* with $\delta_{ij} = 1$ for $i = j$, $\delta_{ij} = 0$ for $i \neq j$.

For a fluid of variable density, the macroscopic diffusive flux of a γ-component is

$$\mathbf{J}^\gamma_{\text{diff}} = -\rho \boldsymbol{\mathcal{D}}^{*\gamma}(\theta) \cdot \nabla \omega^\gamma \tag{10}$$

where all variables and the coefficient are at the macroscopic level and the flux is per unit area of fluid in the porous medium cross-section.

The tortuosity of a phase is a macroscopic geometrical coefficient that expresses the effects of the microscopic surface that bounds that phase on the diffusive flux. As such, it depends on the configuration of the phase within the void space. In unsaturated flow, each of the tortuosity components is a function of the saturation. In an isotropic porous medium, Millington (1959) suggests:

$$\mathrm{T}^*(\theta) = \frac{\theta^{7/3}}{\phi^2} \tag{11}$$

Dispersive Flux

We start from the empirical observation (in field and laboratory experiments) that as an initially sharp interface between two miscible fluids (i.e. the same fluid, but with a jump in concentration of dissolved matter) is being displaced, say, in uniform flow, a transition zone is created between the two fluids. Across it, the concentration of the solute gradually varies from that of one fluid to that of the other (Figure 1). As flow continues, the width of this transition zone increases. Similarly, if a certain porous medium sub-domain initially contains a certain amount of a solute that can be used as a (conservative) *tracer*, and flow takes place in that porous medium domain, the zone occupied by the tracer will grow continuously, both in the direction of the flow, and transversely, that is, normal to it. These spreading phenomena cannot be predicted by using Darcy's law that describes the *averaged* flow, especially noting the spreading perpendicular to the direction of the (local) averaged flow, and the ever-growing zone occupied by the solute-labeled fluid.

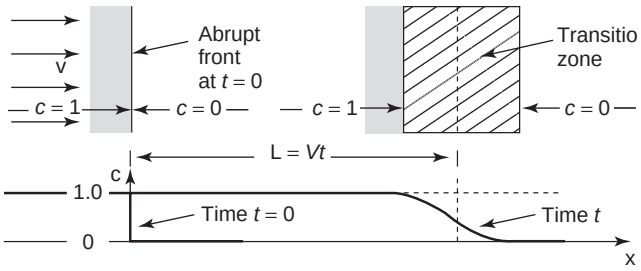


Figure 1 Longitudinal spreading of an initially sharp interface

The spreading phenomenon in a porous medium domain, as described above, is called *hydrodynamic dispersion* (or *miscible displacement*). It is an unsteady, *irreversible process* (in the sense that the initial solute distribution cannot be obtained by reversing the direction of the flow) in which the mass of a solute continuously "mixes" with the nonlabeled portion of the moving fluid.

The phenomenon of dispersion may be demonstrated also by a laboratory column experiment in which a steady flow of water takes place at a constant discharge, Q. The column contains a homogeneous porous material. At $t = 0$, tracer-marked water, at $c = 1.0$ (e.g. water with NaCl at a concentration that is sufficiently low so that the effect of density variations on the flow pattern is negligible), starts to displace the indigenous unmarked water ($c = 0$) in the column. Figure 2, called a *breakthrough curve*, shows the measured tracer concentration, $c = c(t)$, in the effluent leaving the column. In the absence of dispersion, the breakthrough curve would take the form of the dashed line shown in this figure. This would be indicative of the movement of a persistent sharp front between the labeled and unlabeled fluids. In reality, due to hydrodynamic dispersion, the breakthrough curve will take the form of the *S*-shaped curve shown as a solid line in the figure.

The above observations cannot be explained on the basis of the average flow velocity. We must refer to what happens at the *microscopic level*, namely, inside the REV. There, we observe velocity variations in both magnitude and direction across any pore cross-section (even when the averaged flow

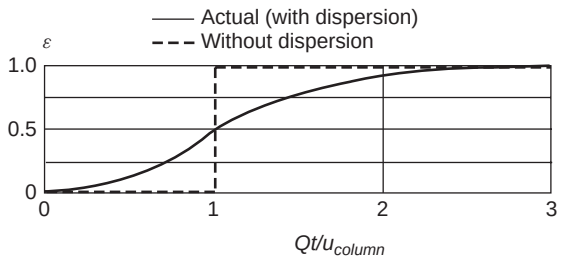


Figure 2 Breakthrough curve in one-dimensional flow in a column of homogeneous porous material

is uniform), and between flow paths. We recall that even in a straight circular capillary tube, we have a parabolic distribution of fluid velocity (see any text on fluid mechanics), with zero velocity at the (stationary) solid surface, and a maximum velocity at the center of the tube. Because of the shape of the interconnected pore space, the (microscopic) streamlines deviate from the mean direction of flow. Altogether, the velocity at the microscopic level varies in magnitude and direction from point to point within the fluid present in the void space. As a consequence, any initial cloud of closely spaced tracer particles will spread out, with each fluid particle traveling along its own microscopic streamline, and the shape and of the initial cloud will gradually change; the fluid volume occupied by the cloud will continuously grow. This phenomenon is referred to as *mechanical dispersion*, where the term "mechanical" is used to remind us that this part of the spreading is due to fluid mechanical phenomena, and "dispersion" is just another word for "spreading". The two basic factors that produce mechanical dispersion are, therefore, (i) *flow* and (ii) the *presence of a pore system* through which the flow takes place. Obviously, to observe spreading, we need a variation in concentration of the tracer (= identifiable particles).

Velocity variations alone cannot explain the ever-growing width (normal to the direction of flow) of a plume of tracer-labeled fluid particles originating at a point source. In order to explain the observed spreading, especially transverse to the flow direction, we must refer to an additional phenomenon that takes place in the void space, namely, *molecular diffusion.* Molecular diffusion of a chemical species in a fluid (see section "Diffusive Flux"), due to concentration gradients inside the void space, produces an additional flux of the species' particles (at the microscopic level) from regions of higher concentrations to those of lower ones, especially normal to the microscopic streamlines. This flux is relative to the advective one, produced by the fluid's velocity. This diffusive flux tends to equalize the concentration along every microscopic stream tube and (mainly) normal to it, between adjacent stream tubes. The phenomenon of diffusion, combined with the randomness of the streamlines, explains the observed ever-growing extent of transverse dispersion.

Thus, the deviations in solute concentration within a fluid from those obtained by assuming advection only (at the average velocity), are due to two simultaneous phenomena: (i) variations in the microscopic velocity of the phase, with respect to the averaged velocity, accompanied by variations in (microscopic) concentration, and (ii) molecular diffusion. In this way, molecular diffusion contributes to the dispersive flux. This contribution is in addition to the diffusive flux *at the macroscopic level*, as described by (the averaged) Fick's law, say equation (9). The latter is the only flux that takes place when the averaged velocity is zero.

Molecular diffusion turns dispersion, even in purely laminar flow, into an irreversible phenomenon. Irreversibility is exhibited, for example, by the growing width of a transition zone around an initially sharp tracer front in uniform flow, as the direction of the flow is reversed.

We use the term *hydrodynamic dispersion* to denote the spreading (at the macroscopic level) that results from both mechanical dispersion and molecular diffusion. Because molecular diffusion is a relatively slow process, its overall effect on dispersion is more significant at low velocities. We refer to the flux that causes mechanical dispersion (of a chemical species) as *dispersive flux*. It is a *macroscopic flux* that expresses the effect of the microscopic variations of velocity and solute concentration in the vicinity of a considered point.

Bear (1972, p. 232) showed that the average (over an REV), macroscopic flux (in terms of mass per unit area of fluid in the porous medium cross-section per unit time), $\overline{c\mathbf{V}}$, of the local, microscopic, advective flux of a chemical species, $c\mathbf{V}$, at a (macroscopic) point in a porous medium domain, can be obtained from (Figure 3):

$$\overline{cV} \equiv \overline{(\overline{c} + c^o)\left(\overline{\mathbf{V}} + \mathbf{V}^o\right)} = \overline{\overline{c}\,\overline{\mathbf{V}}} + \overline{c^o\overline{\mathbf{V}}} + \overline{\overline{c}\mathbf{V}^0} + \overline{c^o\mathbf{V}^0} \tag{12}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ denoting the center of the REV and a point within it, respectively, $\overline{\mathbf{V}}(\mathbf{x}, t)$ and $\overline{c}(\mathbf{x}, t)$ denote the (intrinsic phase) averages of the fluid's velocity and of the concentration, assigned to the center, $\mathbf{V}^o(\mathbf{x}', t; \mathbf{x})$ and $c^o(\mathbf{x}', t; \mathbf{x})$, denote the deviations from the above average values, respectively, at a point $\boldsymbol{x}'$, with $\overline{\mathbf{V}^o} = 0$ and $\overline{c^o} = 0$. The intrinsic phase average is defined as:

$$\overline{e}(\mathbf{x}, t) = \frac{1}{\mathcal{U}_{0\alpha}} \int_{\mathcal{U}_{0\alpha}(\mathbf{x},t)} e(\mathbf{x}', t; \mathbf{x})\, \mathrm{d}\mathcal{U}_\alpha(\mathbf{x}'),$$

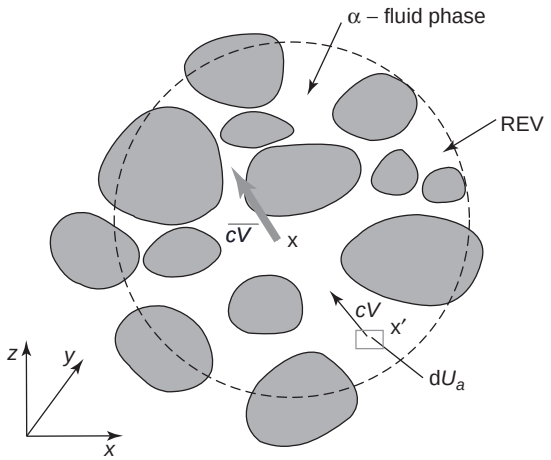


Figure 3 Nomenclature for averaging over an REV. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

$$\mathcal{U}_{0\alpha} = \theta_\alpha(\boldsymbol{x}, t)\mathcal{U}_0 \tag{13}$$

From equation (12), we obtain:

$$\overline{cV} = \overline{c}\,\overline{\mathbf{V}} + \overline{c^o\mathbf{V}^0} \tag{14}$$

From this equation, it follows that the average flux of a component at a point in a porous medium domain (= center of an REV) is equal to the sum of two macroscopic fluxes:

- An *advective flux*, $\overline{c}\,\overline{\mathbf{V}}$, expressing the mass of the component carried by the fluid, at the latter's average velocity, $\overline{\mathbf{V}}(\mathbf{x}, t)$.
- A flux, $\mathbf{J}_{\text{disp}}$ ($\equiv\overline{c^o\mathbf{V}^0}$), that results from the variation of c and $\mathbf{V}$ within the REV for which the considered point, $\boldsymbol{x}$, serves as a center. This flux is the flux referred to earlier as "mechanical dispersion". We refer to it as *dispersive flux*.

Investigations over a period of about five decades, starting around the mid-1950s (e.g. de Josselin de Jong, 1958; Saffman, 1959; Nikolaevskii, 1959; Bear, 1961; Scheidegger, 1961; Bear, 1972; Bear and Bachmat, 1990), led to the conclusion that the dispersive flux (per unit area of fluid) can be expressed as a Fickian-type law in the form:

$$\mathbf{J}_{\text{disp}}(\equiv\overline{c^o\mathbf{V}^0}) = -\mathbf{D}\cdot\nabla\overline{c}, \quad J_i^* = -D_{ij}\frac{\partial\overline{c}}{\partial x_j} \tag{15}$$

where $\mathbf{D}$, is the *coefficient of mechanical* (or *advective*) *dispersion*, and $\mathbf{J}_{\text{disp}}$ $^\circ\mathbf{J}$.

The coefficient of dispersion, $\mathbf{D}$, is a second rank symmetrical tensor, that is, $D_{ij} = D_{ji}$, $i, j = 1, 2, 3$. As such, it has *three principal directions*. Using these directions as Cartesian coordinate axes, x_1, x_2, x_3, the coefficient of dispersion takes the matrix form:

$$\mathbf{D} = \begin{pmatrix} D_{x_1x_1} & 0 & 0 \\ 0 & D_{x_2x_2} & 0 \\ 0 & 0 & D_{x_3x_3} \end{pmatrix} \tag{16}$$

In equation (16), $D_{x_1x_1}$ is called *coefficient of longitudinal dispersion*, while $D_{x_2x_2}$ and $D_{x_3x_3}$ are called *coefficients of transverse dispersion*.

Several authors (e.g. Nikolaevskii, 1959; Bear, 1961; Scheidegger, 1961; Bear and Bachmat, 1967, 1990) have derived the following expression for the relationship between the components D_{ij} of $\mathbf{D}$ and microscopic porous medium configuration, flow velocity, and molecular diffusion:

$$D_{ij} = a_{ijkm}\frac{\overline{V}_k\overline{V}_m}{\overline{V}}f(Pe, \delta), \quad Pe = \frac{L\overline{V}}{D_{\text{diff}}} \tag{17}$$

where $\overline{V}(\equiv|\overline{\mathbf{V}}|)$ is the magnitude of the average velocity, and Pe is a Peclet number, with L denoting a characteristic

length of the pores, for example, hydraulic radius, D_{diff} denoting the coefficient of molecular diffusion, and δ is some geometrical pore characteristics. The Peclet number expresses the ratio between the rates of transport of the considered component by advection and by diffusion. Bear and Bachmat (1967) suggested relationships for $f(Pe, \delta)$. In practice, we use $f(Pe, \delta) \simeq 1$. Henceforth, we shall omit the averaging symbols on c and $\mathbf{V}$.

The coefficients a_{ijkm} (dims. L) are components of a fourth rank tensor, **a**, called the *dispersivity of the porous medium*. It expresses the effect, on solute transport, of the microscopic configuration of the interface between the considered fluid phase and all other phases within the REV. In a saturated system, this interface is that between the fluid and the solid. When a fluid occupies only part of the void space, each of the dispersivity components depends on the volumetric fraction of the fluid, θ.

In a three-dimensional space, the dispersivity, **a**, has 81 a_{ijkm}-components. However, because of various symmetry considerations (Bear, 1961; Scheidegger, 1961), all but 36 components are zeros. In an *isotropic porous medium*, only 21 nonzero components remain. They, in turn, depend only on *two* parameters: a_L and a_T, called the *longitudinal* and *transversal dispersivities* of the porous medium, respectively. The value of a_L is of the order of magnitude of a length that characterizes the heterogeneous configuration of the phase within the REV (e.g. a typical grain size in saturated flow). Laboratory column experiments (e.g. de Josselin de Jong, 1958; Bear, 1961) have shown that a_T is 8 to 24 times smaller than a_L.

For an *isotropic porous medium*, we can express the components of the dispersivity tensor in terms of a_L and, a_T in the form:

$$a_{ijkm} = a_T \delta_{ij} \delta_{km} + \frac{a_L - a_T}{2} \left(\delta_{ik} \delta_{jm} + \delta_{im} \delta_{jk} \right) \quad (18)$$

where δ_{ij} are components of the Kronecker delta.

For an *anisotropic porous medium*, for example, a medium made up of a large number of thin layers normal to the axis of symmetry, the dispersivity can be expressed in the form:

$$\begin{aligned} a_{ijkm} = {} & a_I \delta_{ij} \delta_{km} + a_{II} \left(\delta_{ik} \delta_{jm} + \delta_{im} \delta_{jk} \right) \\ & + a_{III} \left(\delta_{ij} h_k h_m + \delta_{km} h_i h_j \right) \\ & + a_{IV} \left(\delta_{ik} h_j h_m + \delta_{jk} h_i h_m + \delta_{im} h_j h_k + \delta_{jm} h_i h_k \right) \\ & + a_V h_i h_j h_k h_m \end{aligned} \quad (19)$$

where a_I, a_{II}, a_{III}, a_{IV}, a_V are five independent parameters and $\boldsymbol{h}$ is a unit vector directed along the axis of symmetry.

By combining equations (17) and (18), we obtain

$$a_{ijkm} = a_T V \delta_{ij} + (a_L - a_T) \frac{V_i V_j}{V} \quad (20)$$

In the special case of uniform flow, say $V_x \equiv V_1 = V \neq 0$, $V_y = V_z = 0$, equation (20) reduces to $D_{11} = a_L V$, $D_{22} = D_{33} = a_T V$, $D_{ij} = 0$, for $i \neq j$.

In equations (12–15), we have shown how the dispersive flux and the coefficient of dispersion are obtained by averaging the advective flux over the domain occupied by a fluid within an REV. Usually we assume that within the void space, the fluid adheres to the wall, a condition that determines the velocity distribution within the fluid. However, under certain conditions, for example, when the solute is an anion or a cation repelled from a charged solid surface, near which the fluid's velocity is small, the average velocity of the fluid that carries the solute is, higher than when the solute is nonionic. As a consequence, the average advective flux of the solute will be higher, and so will the coefficient of dispersion, which is proportional to the average velocity. This phenomenon is called *ion exclusion* (Gvirtzman and Gorelick, 1991).

Another phenomenon (*size exclusion*) that may affect the magnitude of the coefficient of dispersion is when the molecules comprising the solute are large, to the extent that they are excluded from entering certain pores, or from getting close to the solid. As a consequence, they are carried by a higher average velocity than that of the fluid; this results in a higher coefficient of dispersion.

Total Flux

By combining the three modes transport – advection, dispersion, and diffusion – we can write the total macroscopic flux (*per unit area of a fluid phase*) of a γ-species as

$$\begin{aligned} \mathbf{J}^{\gamma}_{\text{total}} &= c^{\gamma} \mathbf{V} + \mathbf{J}^{\gamma}_{\text{diff}} + \mathbf{J}^{\gamma}_{\text{disp}} = c^{\gamma} \mathbf{V} - \mathbf{D}_h \cdot \nabla c^{\gamma}, \\ \mathbf{D}_h &= \mathbf{D} + \mathcal{D}^{*\gamma} \end{aligned} \quad (21)$$

where $\mathbf{D}_{\text{h}}$ is called the *coefficient of hydrodynamic dispersion*.

Macrodispersion

The phenomenon of dispersion was shown to be a consequence of the heterogeneity of the porous medium *at the microscopic level*, due to the presence of a solid matrix and a void space within the REV. A grain or pore diameter, or the hydraulic radius of the pore space, may serve as examples of a scale of heterogeneity at such level. This heterogeneity produces the velocity variations that take place at the microscopic level. The macroscopic level of description, obtained by averaging over an REV, and the dispersive flux, were introduced as a means for circumventing the need to know the details of the velocity distribution and of other transport features at the microscopic level.

Phenomena of transport take place within a *macroscopic* porous medium domain that may be homogeneous or heterogeneous with respect to the relevant macroscopic geometrical parameters of the porous medium, for example, porosity and permeability. In fact, *all subsurface domains encountered in practice are very heterogeneous*. This heterogeneity plays a dominant role in subsurface transport. We face a situation similar to that which is encountered at the microscopic level, namely, that the detailed information about the spatial variation of the relevant parameters is not available. At the microscopic level, we overcome the lack of information about the (pore scale) heterogeneity by averaging, employing the REV concept. Dispersion was introduced as a consequence of this averaging. An averaging, or smoothing, approach may also be applied (conceptually) to macroscopic level heterogeneities. As a result of this averaging, a new continuum is obtained for describing the porous medium and the phenomena that take place in it. We refer to it as the megascopic *level*.

As in the passage from the microscopic level to the macroscopic one, a new representative elementary volume is needed in order to perform the averaging/smoothing. We shall refer to such a volume as the Representative Macroscopic Volume (abbreviated as RMV). The characteristic size, l^*, of this volume, is constrained by $d^* \ll l^* \ll L$, where d^* is a length characterizing the macroscopic heterogeneity that we wish to smooth out, and L is a length characterizing the considered porous medium domain. The length scale of heterogeneity at the megascopic level is much larger than that corresponding to the macroscopic one.

Similar to what happens during microscopic-to-macroscopic smoothing, the (unknown) information about the heterogeneity at the macroscopic level appears at the megascopic one in the form of various coefficients that reflect the effect of the spatial variability of macroscopic level geometrical parameters (e.g. permeability and porosity) on the transport phenomena.

The total flux of a chemical component at the megascopic level can be obtained by averaging the macroscopic total flux of that component over an RMV. We obtain:

$$\overline{\overline{\theta \mathbf{J}^{\gamma}_{\text{total}}}} = \overline{\overline{c^{\gamma}\mathbf{q}}} + \overline{\overline{\theta\left(\mathbf{J}^{\gamma}_{\text{diff}} + \mathbf{J}^{*\gamma}\right)}} \approx \overline{\overline{c^{\gamma}}}\,\overline{\overline{\mathbf{q}}} + \overline{\overline{\hat{c}^{\gamma}\hat{\mathbf{q}}}}$$

$$\overline{\overline{\hat{c}^{\gamma}\hat{\mathbf{q}}}} \gg \overline{\overline{\theta\left(\mathbf{J}^{\gamma}_{\text{diff}} + \mathbf{J}^{*\gamma}\right)}}$$

$$\overline{\overline{e}}(\mathbf{x}, t) = \frac{1}{\overline{\overline{\mathcal{U}_0}}} \int_{\overline{\overline{\mathcal{U}_0}}} \overline{e}(\mathbf{x}', t; \mathbf{x})\, \mathrm{d}\overline{\overline{\mathcal{U}_0}}\left(\mathbf{x}'\right) \tag{22}$$

where the double bar over a macroscopic value indicates an average over an RMV (= megascopic volume, $\overline{\overline{\mathcal{U}_0}}$), and the hat symbol (ˆ) denotes the deviation of a macroscopic value of a quantity at any point within an RMV, from its average over the RMV. We note that the flux on the left-hand side of equation (22) (and hence all other fluxes) is *per unit area of porous medium*. The total megascopic flux is made up of two fluxes: an advective flux that carries the macroscopic concentration at the megascopic velocity, and a dispersive flux, due to the variations of the macroscopic specific discharge within an RMV. We have assumed that this latter flux is much larger than the average of the flux of hydrodynamic dispersion, that is, sum of the dispersive and diffusive fluxes at the macroscopic level.

Altogether, the total flux is, again, the sum of an advective flux and a dispersive one. There is no analogy here to the diffusive flux, as we have neglected it. At very low velocities, we may not neglect the average of the macroscopic diffusive flux. We usually *assume* that a Fickian-type dispersion law can be employed to describe also the dispersive flux at the megascopic level. A *macrodispersivity*, A_{ijkm}, can be defined in the same way as the dispersivity that was defined in equation (16). Bear (1979), and Bear and Verruijt (1987), while developing the vertically integrated mass balance equation for a component of a phase within an aquifer, suggested an expression for macrodispersivity. In an isotropic porous medium, the macrodispersivity reduces to a scalar. Gelhar (1976) and Gelhar *et al.* (1979) analyzed the dependence of macrodispersion on permeability variations.

To summarize, dispersion and macrodispersion are analogous phenomena, in that both are consequences of velocity variations that are due to heterogeneity, but at different scales. Dispersion arises from velocity variations within the void space, caused by the presence of the solid surfaces. Macrodispersion is caused by variations in the permeability and porosity at the macroscopic level. In both cases, the total flux can be expressed as a sum of an advective flux and a (mechanical) dispersive one, written at the respective levels. The structure of the coefficient of dispersion is the same in both cases, and so is the relationship between the coefficient of dispersion, the dispersivity, and the average velocity. The selection of which one to use in a particular case depends on the scale of the problem, that is, the *scale of heterogeneity*. The latter depends on the *scale of the problem*. Field experiments and model calibration seem to indicate that the longitudinal dispersivity is of the order of magnitude of 1/10 of the size of the *domain of interest,* although it should level off at some sufficiently large domain size. In practice, the transverse dispersivity is usually approximated as about 1/10 of the longitudinal one. Obviously, these are merely orders of magnitude and may be used as initial or preliminary estimates. In each particular case, the actual value should be determined by some model calibration, or parameter estimation procedure.

Apart from the different magnitude of the dispersivity to be employed, the expressions for advective and dispersive fluxes, equations (1) and (15–20) are assumed to remain

valid also when we model field conditions. The species mass balance equation discussed below is also the same at both scales.

MASS BALANCE EQUATION

Each of the flux equations discussed in the section "Fluxes", say, equation (21) involves two variables: the flux, $\mathbf{J}^{\gamma}_{\text{total}}$, and the (average) concentration, c^{γ}. To solve for these variables, the required additional equation is the balance equation for the mass of the considered species. We shall continue to discuss a fluid phase at a volumetric fraction θ; for saturated flow, we replace θ by the porosity, ϕ.

Fundamental Mass Balance

The general balance equation and the balance equation for the mass of a fluid phase are discussed in detail in **Chapter 149, Hydrodynamics of Groundwater, Volume 4**. To facilitate the discussion here, let us summarize this subject by introducing the general mass balance equation for any extensive quantity, E, *considered as a continuum*, having a density (= mass per unit volume of a continuum) e. For a finite domain, $\mathcal{U}_0$, bounded by a surface S, and for a finite period of time, Δt, this balance takes the form:

$$\left\{\begin{array}{c}\text{Quantity of } E\\ \text{accumulating}\\ \text{in } \mathcal{U}_o\\ \text{during } \Delta t\end{array}\right\} = \left\{\begin{array}{c}\text{Net quantity of}\\ E \text{ entering } \mathcal{U}_o\\ \text{through } \mathcal{S}\\ \text{during } \Delta t\end{array}\right\} + \left\{\begin{array}{c}\text{Net production}\\ \text{of } E\\ \text{in } \mathcal{U}_o\\ \text{during } \Delta t\end{array}\right\} \tag{23}$$

When writing the above equation for a small volume around a point in a fluid continuum, and letting both $\mathcal{U}_0$ and Δt shrink to zero, we obtain the balance equation for E in the form of the *partial differential equation*:

$$\frac{\partial e}{\partial t} = -\nabla \cdot e\mathbf{V}^E + \rho\Gamma^E \tag{24}$$

in which ρ denotes the mass density of that continuum, $\mathbf{V}^E$ denotes the velocity of the E-continuum, and Γ^E denotes the rate of net production of E, per unit mass of the continuum. Another form of equation (24) is

$$\frac{\partial e}{\partial t} = -\nabla \cdot (e\mathbf{V} + \mathbf{J}^E) + \rho\Gamma^E, \mathbf{J}^E = e(\mathbf{V}^E - \mathbf{V}) \tag{25}$$

in which $\mathbf{V}$ and $\mathbf{J}^E$ are the (mass averaged) velocity and the diffusive flux of E, respectively.

Equation (25) may be regarded as the most general balance equation for any extensive quantity E of a phase in a porous medium *at the microscopic level*. To obtain its *macroscopic* counterpart, we average it over the phase occupying part (volumetric fraction θ) of an REV. We obtain (Bear and Bachmat, 1990):

$$\frac{\partial \theta\overline{e}}{\partial t} = -\nabla \cdot \theta(\overline{e}\,\overline{\mathbf{V}} + \mathbf{J}^E + \mathbf{J}^{*E}) - \frac{1}{\mathcal{U}_o}\int_{\mathcal{S}} [\mathbf{e}(\mathbf{V} - \mathbf{u}) + \mathbf{j}^E] \cdot \mathbf{n}\, \mathrm{d}S + \theta\overline{\rho\Gamma^E} \tag{26}$$

where $\mathcal{U}_o$ is the volume of the REV, S is the (microscopic) surface that bounds the phase inside $\mathcal{U}_o$, $\boldsymbol{u}$ is the velocity of S, $\boldsymbol{n}$ is the outward vector normal to S, and $\mathbf{J}^E$ and $\mathbf{J}^{*E}$ denote the dispersive and (macroscopic) diffusive fluxes of E.

For example, for $e = c^{\gamma}$ = (mass) concentration of a γ-species, equation (26) becomes

$$\frac{\partial \theta\overline{c^{\gamma}}}{\partial t} = -\nabla \cdot \theta(\overline{c^{\gamma}}\,\overline{\mathbf{V}} + \mathbf{J}^{\gamma} + \mathbf{J}^{*\gamma}) - f^{\gamma} + \theta\overline{\rho\Gamma^{\gamma}}$$
$$\mathbf{J}^{*\gamma} = -\rho\mathbf{D} \cdot \nabla\overline{\omega^{\gamma}}, \quad \mathbf{J}^{\gamma} = -\rho\boldsymbol{\mathcal{D}}^{*\gamma} \cdot \nabla\overline{\omega^{\gamma}} \tag{27}$$

in which $f^{\gamma}(= (1/\mathcal{U}_o)\int_{\mathcal{S}} [c^{\gamma}(\mathbf{V} - \mathbf{u}) + \mathbf{j}^{\gamma}_{\text{diff}}] \cdot \mathbf{n}\, \mathrm{d}S)$ is a *source* term that expresses the net rate of transfer of mass by advection and diffusion, of γ from the considered fluid phase, across S, to all other phases within the REV (e.g. by adsorption on the solid and/or diffusion into another fluid phase), per unit volume of porous medium. When the surface, S, is a *material surface* with respect to the fluid, the transfer is by diffusion only. The term $\theta\overline{\rho\Gamma^{\gamma}}$ denotes another source of γ mass, due to production of γ *within* the phase (e.g. by chemical reactions, decay, or degradation). In equation (27), we note the expressions for the dispersive and diffusive fluxes.

Note that the values of θ and $\overline{\mathbf{V}}$ are obtained from the solution of the corresponding flow model. When the fluid's density depends on concentration, the flow and the solute transport models have to be solved simultaneously.

In what follows, we shall discuss the two kinds of sources of γ that appear in the mass balance equation (27).

Sources Due to Injection and Pumping

Let the considered fluid be injected at M isolated *points,* $\mathbf{x}^m$, at rates R^m_{ext} (dims. L^3/T), and known concentrations c^m_{ext}, and let the fluid be pumped at N isolated *points* at rates Q^n (dims. L^3/T) at the unknown concentration, $\overline{c^{\gamma}}$. The net source of solute mass is then expressed as

$$\theta\overline{\rho\Gamma^{\gamma}} = \sum_{(m)} R^m_{\text{ext}}(\mathbf{x}^m, t)\delta(\mathbf{x} - \mathbf{x}^m)\mathrm{c}^{\gamma}_{\text{ext}}(\mathbf{x}^m, t) - \sum_{(n)} Q^n(\mathbf{x}^n, t)\delta(\mathbf{x} - \mathbf{x}^n)\overline{c^{\gamma}}(\mathbf{x}^n, t) \tag{28}$$

where $\delta(\mathbf{x} - \mathbf{x}^m)$ denotes the *Dirac delta function* (dims. L^{-3}), with $\delta = 1$ for $\mathbf{x} = \mathbf{x}^m$, $\delta = 0$ for $\mathbf{x} \neq \mathbf{x}^m$.

Sources Due to Decay and Degradation

For radioactive decay, with λ denoting the decay coefficient, the source term takes the form:

$$\theta\overline{\rho\Gamma^\gamma} = -\theta\lambda\overline{c^\gamma} \tag{29}$$

The minus sign indicates that we have here a sink.

For an adsorbed component (see discussion in the next section) that undergoes radioactive decay, the source term takes the form:

$$\theta\overline{\rho\Gamma_s^\gamma} = -\lambda_s F \tag{30}$$

where F denotes the mass of the component per unit mass of solid, and λ_s is the coefficient of decay on the solid.

For any other decay or degradation phenomena of a considered component, we can replace λ in equation (29) by the degradation rate constant, k_f. For an adsorbed component, we replace λ by k_s in equation (30).

Without the averaging symbols, and for a fluid of constant density, equation (27) can be rewritten in the form:

$$\begin{aligned}\frac{\partial\theta c^\gamma}{\partial t} &= -\nabla\cdot\theta(c^\gamma\mathbf{V} + \mathbf{J}^\gamma + \mathbf{J}^{*\gamma}) - f^\gamma + \theta\rho\Gamma^\gamma \\ \mathbf{J}^{*\gamma} &= -\mathbf{D}\cdot\nabla c^\gamma,\ \mathbf{J}^\gamma = -\boldsymbol{\mathcal{D}}^{*\gamma}\cdot\nabla c^\gamma\end{aligned} \tag{31}$$

Sources Due to Adsorption

Adsorption (the opposite of *desorption*): It is the phenomenon of accumulation of a chemical species (= *adsorbate*) present in a liquid that occupies the void space, or part of it, on the solid matrix (= *adsorbent*) at the liquid-solid interface. This phenomenon is caused by the attraction of a species to the surfaces of the solid, or by reactions of the species in the liquid with the solid.

When the considered adsorption process is sufficiently fast, so that *chemical equilibrium* may be assumed to prevail, we can use of the thermodynamic relationship, called *adsorption isotherm*. This is a function that relates the quantity of a species adsorbed on the solid, to its quantity in the liquid phase that occupies the void space (or part of it), *at a fixed temperature*, under conditions of (chemical) equilibrium between the two quantities.

The concept of *adsorption isotherm* assumes that the amount of the species on the solid is solely a function of the concentration of the adsorbed species in the fluid close to the solid surface. This assumption fails when there is only one involved species in the fluid. It is also valid when there are more species, but their concentrations do not change appreciably in time. Otherwise, a more complicated analysis is required to determine the concentrations of all of the various species that are involved. Here, we consider only surface complexation.

The primary driving forces for adsorption are (i) the *lyophobic* (= solute disliking) nature of the solute, relative to the solvent that hosts it, (ii) high *affinity* of the solute for the solid, due, for example, to electrical attraction, (iii) *van der Waals attraction,* that is, intermolecular *forces* of attraction between molecules of the solid and those of the adsorbed species, and (iv) *chemisorption*, that is, chemical interaction between the solid and the adsorbed species. However the most significant factor is the degree of solubility of the dissolved species.

In certain porous media, the solid matrix itself is porous, with *much* smaller pores (i.e. with very low permeability). A soil made up of porous aggregates may serve as an example. Under such conditions, an adsorbate must first diffuse *into* the small pores of the aggregates, and then adsorb on their (very large) surface area. We refer to such phenomenon as *absorption*.

When adsorption of a chemical species in solution takes place, the total mass of the former is *partitioned* between the solution and the solid matrix. Any increase in the quantity of the species on the solid is accompanied by an appropriate increase in its quantity in the liquid, and vice versa.

We use the symbol F to denote the mass of the species (= adsorbate) adsorbed on the solid (= adsorbent), per unit mass of the latter (units: kg kg^{-1}, moles kg^{-1}). The isotherm for a given adsorbate–adsorbent pair can be directly obtained by performing a batch adsorption experiment.

When adsorption is the only interphase transfer mechanism, the term f^γ in equation (27) takes the form:

$$f^\gamma \equiv f^\gamma_{f\to s} = -f^\gamma_{s\to f} = \rho_b \left.\frac{\mathrm{d}F}{\mathrm{d}t}\right|_{\text{ads}} \tag{32}$$

where $\rho_b(\equiv(1-\phi)\rho_s)$ denotes the *bulk density of the solid matrix*. However, this rate is unknown.

Under the equilibrium assumption, we can eliminate the rate of interphase exchange due to adsorption, by writing and summing up the balance equations for the considered species in the fluid phase and on the solid. The macroscopic balance equation for the considered species on the solid can be obtained from equation (31), in which θc^γ is replaced by $\theta_s\rho_s F(\equiv\rho_b F)$, and assuming that no flux of the considered species takes place within the solid phase and/or on its surface. The resulting mass balance for the species on the solid surface, takes the form:

$$\frac{\partial}{\partial t}\rho_b F = -f^\gamma_{s\to f} + \rho_b\Gamma^\gamma_s \tag{33}$$

where $\rho_b\Gamma^\gamma_s$ is the rate of production of the species mass on the solid surface, per unit volume of porous medium. By summing up the two balance equations for the considered component, we obtain the *mass balance equation for the*

component in the porous medium as a whole, in the form:

$$\frac{\partial(\theta c^\gamma + \rho_b F)}{\partial t} = -\nabla \cdot \theta(c^\gamma \mathbf{V} - \mathbf{D}_h \cdot \nabla c^\gamma) + \theta\rho\Gamma^\gamma + \rho_b \Gamma_s^\gamma \tag{34}$$

in which appropriate expressions should replace the two source terms, for example, equation (28) through equation (30).

Equation (34) contains the two variables, $c^\gamma(\mathbf{x}, t)$ and $F(\mathbf{x}, t)$. By assuming equilibrium conditions, we can make use of an appropriate *isotherm* that relates c^γ to F. For example, we may use the linear *Freundlich isotherm*:

$$F = K_d c^\gamma \tag{35}$$

to obtain the mass balance equation for the species in the porous medium as a whole, in terms of a single variable, c^γ, in the form:

$$\frac{\partial(\theta c^\gamma + \rho_b F)}{\partial t}\left(\equiv \frac{\partial \theta R_d c^\gamma}{\partial t}\right) = -\nabla \cdot \theta(c^\gamma \mathbf{V} - \mathbf{D}_h \cdot \nabla c^\gamma) + \theta\rho\Gamma^\gamma + \rho_b \Gamma_s^\gamma \tag{36}$$

in which $R_\mathrm{d} = 1 + \rho_b K_d/\theta$ is called the *retardation coefficient* (discussed below). We note that this coefficient (>1) indicates the partitioning between the solute in solution and that on the solid. For saturated flow, we replace θ by ϕ. Equation (36) is applicable also to water in the unsaturated zone, as we assume that water is the (= wetting fluid) that is everywhere adjacent to the solid surface, albeit at some places as a very thin film, and that diffusion of the species is possible even through this very thin film.

When equilibrium cannot be assumed, the rate of transfer in both *balance equations* is expressed in terms of an appropriate rate transfer expression. Typically, this expression takes the form:

$$f_{f\to s}^\gamma = \alpha(Ac^\gamma - BF), \quad A, B = \text{coefficients} \tag{37}$$

For example, we may write the model in the form of the three equations:

$$\begin{aligned} \frac{\partial \theta c^\gamma}{\partial t} &= -\nabla \cdot \theta(c^\gamma \mathbf{V} - \mathbf{D}_h \cdot \nabla c^\gamma) + \theta\rho\Gamma^\gamma - f_{f\to s}^\gamma \\ \frac{\partial \rho_b F}{\partial t} &= f_{f\to s}^\gamma + \rho_b \Gamma_s^\gamma \\ f_{f\to s}^\gamma &= Ac^\gamma - BF, \quad A, B = \text{coefficients} \end{aligned} \tag{38}$$

which have to be solved simultaneously for $f_{f\to s}^\gamma$, $c^\gamma(\mathbf{x}, t)$ and $F(\mathbf{x}, t)$.

Retardation To understand the concept of *retardation*, R_d, appearing in equation (36), we consider one-dimensional solute transport in a homogeneous saturated porous medium domain, with and without any sources, except adsorption. The corresponding balance equations, respectively, are

$$\phi\frac{\partial c^\gamma}{\partial t} = -\nabla \cdot \phi\left(c^\gamma \frac{\mathbf{V}}{R_d} - \frac{\mathbf{D}_h}{R_d} \cdot \nabla c^\gamma\right), \quad R_d = 1 + \frac{\rho_b K_d}{\phi}$$

$$\phi\frac{\partial c^\gamma}{\partial t} = -\nabla \cdot \phi(c^\gamma \mathbf{V} - \mathbf{D}_h \cdot \nabla c^\gamma) \tag{39}$$

We note that both equations are similar, except that in the first equation, the average fluid velocity *seems* to be $\mathbf{V}/R_d$, and the coefficient of hydrodynamic dispersion seems to be reduced to $\mathbf{D}_h/R_d$. We recall that (except when velocities are very low) the major component of $\mathbf{D}_h$ is the coefficient of mechanical dispersion, $\mathbf{D}$, which is proportional to $\mathbf{V}$. Thus, under the assumption of equilibrium adsorption, described by a linear isotherm, the effect of adsorption is to retard the advance of the component (as part of it is adsorbed to the solid); the mean movement of the contaminant is at the reduced, or retarded velocity, $\mathbf{V}/R_d$. At the same time, spreading occurs as if the coefficient of mechanical dispersion is also reduced by the factor R_d, along with the coefficient of molecular diffusion in a porous medium. Figure 4 shows the phenomenon of retardation.

The same phenomenon of retardation exists also in the more general relations between c^γ to F.

Sources Due to Volatilization

Although the focus in this article is on saturated flow, as we have done earlier, by using θ instead of ϕ to indicate the volumetric fraction of a phase in a multiphase system, we add now the case of the unsaturated zone with two fluid phases: a gas, g, and a liquid, l, in which a considered chemical species can volatilize through the fluid-gaseous phase (microscopic) interface. The reason is that volatile organic compounds (VOC's) are often encountered as contaminants in the unsaturated zone. The considered volatile (superscript v) contaminant is a VOC that is present

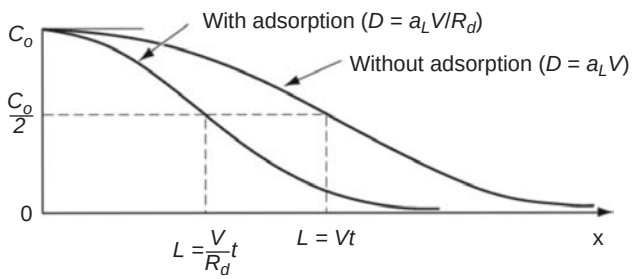


Figure 4 Retardation

as a vapor component in the gaseous phase (saturationS_g) at concentration c_g^v, in the aqueous liquid (saturationS_l) at concentration c_l^v, and as an adsorbate (linear isotherm, $F = K_d c_l^v$) on the solid at concentration F^v. The partitioning of v between the gas and the liquid is assumed to obey *Henry's law*

$$H \equiv \mathcal{H}_{g,l}^v = \frac{c_g^v}{c_l^v} \tag{40}$$

To generalize, we assume that the considered component also undergoes a first order decay, with rate coefficients, λ_g, λ_l, λ_s, in the gas, liquid, and solid phases, respectively. We shall assume that no sources and sinks are present in the considered domain, except that gas is extracted from the void space at a volumetric rate Q_g (= volume of gas per unit volume of porous medium). We shall disregard the presence of water vapor as a component in the gas, and of dissolved air as a component in the liquid. Because of the assumption of equilibrium partitioning, the three balance equations that describe the spatial and temporal concentration distributions of the VOC in the two fluid phases and on the solid, can be combined into a single mass balance equation for the component in the porous medium as a whole, in the form:

$$\begin{aligned}\frac{\partial}{\partial t}\left[\phi\left(S_l + S_g\mathcal{H} + \frac{\rho_b K_d}{\phi}\right)c_l^v\right] \\ = -\nabla\cdot\phi[S_g\mathcal{H}(c_l^v\mathbf{V}_g - \mathbf{D}_{gh}\cdot\nabla c_l^v) + S_l(c_l^v\mathbf{V}_l - \mathbf{D}_{lh}\cdot\nabla c_l^v)] \\ - \phi\left(\lambda_l S_l + \lambda_g S_g\mathcal{H} + \frac{\lambda_s\rho_b K_d}{\phi}\right)c_l^v - Q_g\mathcal{H}c_l^v\end{aligned} \tag{41}$$

This is a single mass balance equation, in the single variable, c_l^v.

Sources Due to Chemical Reactions

Chemical reactions may occur among species in solution in a fluid, that is, at the microscopic level. Each reaction has its own rate. When the characteristic time required for a reaction is much smaller than that for advection and diffusion, we consider the reaction as an *equilibrium reaction.* Otherwise, the reaction is referred to as a *kinetic reaction.* For homogeneous reactions, that is, ones that occur *within* a phase, it is usually assumed that the deviations in the thermodynamic state within the REV are sufficiently small such that the same form of the microscopic rate law can be used at the macroscopic level.

A chemical reaction can be described by a *stoichiometric* equation. The general form of a stoichiometric equation is

$$\nu^A\mathrm{A} + \nu^B\mathrm{B} + \cdots \longrightarrow \nu^P\mathrm{P} + \nu^Q\mathrm{Q} + \cdots \tag{42}$$

in which A,B,... denote *reactant* species, P,Q,...denote *product* species and the $\nu^{\gamma'}$s (>0) denote *stoichiometric coefficients*. For a reversible reaction, we replace the symbol $\rightarrow$ by $\rightleftharpoons$. For the reaction described by equation (42), the *rate of reaction,* R_r (in moles per second in a given volume of solution), is expressed in the form:

$$\begin{aligned}R_r &= -\frac{1}{\nu^A}\frac{\mathrm{d}[A]}{\mathrm{d}t}\bigg|_{c.r.} = -\frac{1}{\nu^B}\frac{\mathrm{d}[B]}{\mathrm{d}t}\bigg|_{c.r.} = \cdots \\ &= \frac{1}{\nu^P}\frac{\mathrm{d}[P]}{\mathrm{d}t}\bigg|_{c.r.} = \frac{1}{\nu^Q}\frac{\mathrm{d}[Q]}{\mathrm{d}t}\bigg|_{c.r.} = \cdots \\ \frac{\mathrm{d}[A]}{\mathrm{d}t}\bigg|_{c.r} &\equiv \frac{1}{M^A}\frac{\mathrm{d}c^A}{\mathrm{d}t}\bigg|_{c.r.}\end{aligned} \tag{43}$$

in which "c.r" denotes "chemical reaction", [A] denotes *molar concentration*, and M^A denotes the molar mass of A.

A reversible reaction can also be described by the stoichiometric equation

$$\sum_{(\gamma)}\nu^\gamma\mathcal{M}^\gamma \rightleftharpoons 0 \tag{44}$$

in which $\mathcal{M}^\gamma$ is a chemical symbols of the respective γ-species, and the ν^γ's are their corresponding stoichiometric coefficients. In this form of the stoichiometric equation, $\nu^\gamma < 0$ for a reactant and $\nu^\gamma > 0$ for a product. The rate of reaction is, then, given by:

$$R_r = \frac{1}{\nu^\gamma}\frac{\mathrm{d}[\gamma]}{\mathrm{d}t}\bigg|_{c.r.} \tag{45}$$

Often, the rate of the reaction for equation (42) is expressed as

$$R_r = k[A]^{\lambda^A}[B]^{\lambda^B}\cdots[P]^{\lambda^P}[Q]^{\lambda^Q}\cdots \tag{46}$$

with, in general, $\nu^\gamma \neq \lambda^\gamma$. For the reaction, $A + B \rightarrow C$, the reaction rate may be given, for example, by $R_r = -d[A]/\,\mathrm{d}t|_{c.r.} = k[A][B]$ – a *second order rate law*.

When a γ-species participates in several chemical reactions that cause its concentration within the fluid phase to increase (or decrease), we express the strength of the source (= rate of production) of that γ-species in the (macroscopic) balance equation, say equation (31), by:

$$\theta\rho\Gamma^\gamma = \sum_{(j)}\frac{\mathrm{d}c^\gamma}{\mathrm{d}t}\bigg|_{j\text{th hom.chem.react.}} = \theta M^\gamma\sum_{(j)}\nu_j^\gamma R_{rj} \tag{47}$$

in which R_{rj} is the reaction rate of the jth homogeneous chemical reaction in the fluid and ν_j^γ is the stoichiometric coefficient of the γ-species in the jth reaction. This rate of production is in addition to the rates of production resulting from other sources.

INITIAL AND BOUNDARY CONDITIONS

The solution of any of the balance equations for the mass of a solute requires initial and boundary conditions. Initial conditions involve the specification of the value of the concentration, $c = c(\mathbf{x}, t)$, at all points $\mathbf{x}$ of the domain at the initial time, $t = 0$.

General Boundary Conditions

The boundary conditions for solute transport models, at the macroscopic level, are based on the following two principles:

- At every point on a boundary, there exists no discontinuity in the concentration as the boundary is crossed from the considered domain to its vicinity. However, under certain conditions that are consequences of the various assumptions that underlie the model, we are forced to violate this condition in order to achieve a better approximation of a real situation.
- In the absence of sources and sinks of a considered solute on the boundary, which is the usual case, the component normal to the boundary of the total flux of that solute, with respect to the (possibly moving) boundary, undergoes no jump as the latter is crossed:

For a boundary surface that is also a *material surface* with respect to the solid, $(\mathbf{V}_{\text{solid}} - \mathbf{u}) \cdot \mathbf{n} = 0$, the condition of flux continuity can be written in the form:

$$[\![(c\mathbf{q}_r - \theta \mathbf{D}_h \cdot \nabla c)]\!]_{1,2} \cdot \mathbf{n} = 0, \mathbf{q}_r \equiv \theta(\mathbf{V}_{\text{fluid}} - \mathbf{V}_{\text{solid}}) \tag{48}$$

where $\mathbf{q}_r$ is the specific discharge of the fluid, relative to the solid, as described by Darcy's law, $\mathbf{n}$ denotes the unit vector normal to the boundary, moving at the velocity $\mathbf{u}$, and 1,2 denote the internal and external sides of the boundary, respectively. The values of $\mathbf{V}_{\text{fluid}}$ and θ are assumed known from the flow model.

For equation (48) to serve as a condition for c on a boundary, information on what happens on the external side of the latter (here, on c and ∇c) must be known as *a function of space and time*. Let us present a number of particular cases. We shall assume that all boundaries, except the phreatic surface (= water table), are stationary and material surfaces with respect to the solid.

Boundary of Prescribed Concentration

When the value of $c(\mathbf{x}, t)$ is prescribed as a known function, $f^1 = f^1(\mathbf{x}, t)$, at all points of a boundary segment, due to phenomena that take place on the external side of the latter, independent of what happens within the domain, the boundary condition is

$$c(\mathbf{x}, t) = f^1(\mathbf{x}, t) \tag{49}$$

This is a *first kind* or *Dirichlet boundary condition*.

Boundary of Prescribed Flux

When phenomena occurring on the external side of a boundary impose a known *total* flux of the considered component, $f^2 = f^2(\mathbf{x}, t)$, normal to a boundary segment, at all points of the latter, *regardless* of what happens within the considered domain itself, the condition takes the form:

$$(c\mathbf{q}_r - \theta \mathbf{D}_h \cdot \nabla c) \cdot \mathbf{n} = f^2(\mathbf{x}, t) \tag{50}$$

where the specific discharge (relative to the solid) $\mathbf{q}_r$ can be expressed by Darcy's law. As always, for saturated flow, we replace θ by ϕ. This is a *third kind*, or *Cauchy boundary condition*. A special case is an impervious boundary, where $f^2(\mathbf{x}, t) \equiv 0$, and $\mathbf{q}_r \equiv 0$. The boundary condition then reduces to a *second kind* or *Neumann boundary condition*.

Boundary Between Two Porous Media

Along such a boundary, we assume the existence of discontinuities in all solid matrix characteristics. Two conditions must be satisfied along such a boundary:

$$c|_1(\mathbf{x}, t) = c|_2(\mathbf{x}, t)$$
$$[c\mathbf{q}_r - \theta \mathbf{D}_h \cdot \nabla c]|_1 \cdot \mathbf{n} = [c\mathbf{q}_r - \theta \mathbf{D}_h \cdot \nabla c]|_2 \cdot \mathbf{n} \tag{51}$$

Since $[\mathbf{q}_r]|_1 \cdot \mathbf{n} = [\mathbf{q}_r]|_2 \cdot \mathbf{n}$, $[c]|_1 = [c]|_2$, the second equation can be written as

$$[\theta \mathbf{D}_h \cdot \nabla c]|_1 \cdot \mathbf{n} = [\theta \mathbf{D}_h \cdot \nabla c]|_2 \cdot \mathbf{n} \tag{52}$$

Boundary with a Body of Fluid

We consider a stationary boundary between a porous medium domain (denoted by *pm*) and a body of fluid (denoted by *fb*), for example, a lake or a river, assumed to be a "well mixed" domain, that is, at a *known uniform* concentration of the considered component, say, c''. The two domains are assumed to be in good hydraulic contact. The condition of no-jump (across the boundary) in the normal component of the total flux of the considered component, takes the form:

$$(c''\mathbf{V})_{fb} \cdot \mathbf{n} - [c\mathbf{q}_r - \phi \mathbf{D}_h \cdot \nabla c]|_{pm} \cdot \mathbf{n} = 0 \tag{53}$$

However, this boundary introduces an inconsistency in the case of no-flow across the boundary. An improvement,

introduced by adding a "buffer zone" (Bear and Bachmat, 1990), leads to the condition:

$$(c'' - c|_{pm})(\mathbf{q}_r \cdot \mathbf{n} + \alpha^*) = -\phi \mathbf{D}_h \cdot \nabla c|_{pm} \cdot \mathbf{n} \quad (54)$$

where α^* is a *transfer coefficient*, such that $\alpha^*(c'' - c)$ represents the sum of diffusive and dispersive fluxes through the transition zone. Equation (54) introduces a concentration jump on the boundary. This is a consequence of introducing the transition zone and the "well mixed zone" approximation.

Phreatic Surface with Accretion

The phreatic surface, defined in **Chapter 149, Hydrodynamics of Groundwater, Volume 4**, as the *surface at every point of which the water pressure is atmospheric*, serves as the upper boundary of the saturated domain. The condition for flow on this boundary is also presented there. Here, we shall discuss only the condition for the solute transport model.

The shape of the phreatic surface is defined by the equation $F = F(x, y, z, t) = 0$, with

$$F(x, y, z, t) \equiv h(x, y, z, t) - z = 0, \quad \mathbf{n} = \frac{\nabla F}{|\nabla F|},$$

$$\mathbf{u} \cdot \nabla F = -\frac{\partial F}{\partial t} \quad (55)$$

in which h denotes the piezometric head and $\mathbf{u}$ is the speed of displacement of the phreatic surface. Basically, the condition is one of continuity of the normal flux of the considered component across the boundary. As is usually done in groundwater hydrology, we neglect the details of the movement of water through the unsaturated zone above the phreatic surface, say at the *residual water volumetric fraction*, θ_{wr}, and consider only some mean value of natural replenishment, **N**, infiltrating at ground surface and reaching the phreatic surface as accretion, or *natural replenishment*. Let c' denote the concentration of the infiltrating water. The condition of equality of solute flux across the phreatic surface takes the form:

$$\phi[c(\mathbf{V} - \mathbf{u}) - \mathbf{D}_h \cdot \nabla c] \cdot \mathbf{n} = c'(\mathbf{N} - \theta_{wr}\mathbf{u}) \cdot \mathbf{n} \quad (56)$$

in which we may insert $F = h(x, y, z, t) - z$, and use equation (55) to express **n** and **u.**

Seepage Face

The seepage face as a boundary is discussed in **Chapter 149, Hydrodynamics of Groundwater, Volume 4**. The condition is based on equality of normal solute flux:

$$(\phi c\mathbf{V} - \phi \mathbf{D}_h \cdot \nabla c)|_{pm} \cdot \mathbf{n} = (c\mathbf{V})|_{env} \cdot \mathbf{n} \quad (57)$$

Since $c|_{pm} = c|_{env}$, $(\phi \mathbf{V})|_{pm} \cdot \mathbf{n} = \mathbf{V}|_{env} \cdot \mathbf{n}$, the condition of equation (57) reduces to

$$\mathbf{D}_h \cdot \nabla c|_{pm} \cdot \mathbf{n} = 0 \quad (58)$$

COMPLETE MATHEMATICAL MODEL

The mathematical statement, or *model*, of a multiphase flow problem, combined with the model of transport of a single chemical component, possibly with first order decay, adsorption, and volatilization, consists of the following parts:

- A mathematical description of the configuration of the surface that bounds the porous medium problem domain.
- A list of the dependent variables. These are the concentrations, c_α^γ of the considered γ-species within all fluid phases present in the system. In the case of adsorption, F^γ is included in the list of state variables. For the flow model, depending on the number of fluid phases that are in motion, we may add such variables as piezometric heads, pressures, saturations, and so on.
- Flux equations for the mass of the considered fluid phases. Darcy's law is often employed.
- Mass balance (partial differential) equations for the relevant fluid phases.
- Mass and momentum (partial differential) balance equations for the solid, when the latter is deformable.
- Mass balances (partial differential) equations that describe of the considered γ-species within all fluid phases present in the system and on the solid. These balance equations may contain source terms that correspond to decay, adsorption, and volatilization of the considered species.
- Dispersive and diffusive flux equations for the mass of the considered species
- Constitutive equations for the fluid phases, for the solid (in the case of a deformable solid), and for the γ-species. These include also thermodynamic relationships that describe the partitioning of the γ-species between adjacent phases under equilibrium conditions, or transfer functions for nonequilibrium conditions.
- Expressions for the various external sources and sinks (for the mass of fluid phases and the mass of the considered species).
- Statement of initial conditions for each of the relevant balance equations.
- Statement of boundary conditions for each of the relevant balance equations.
- Numerical values, or functional relationships for all the coefficients that appear in the various balance equations and constitutive relations included in the model.

The number of variables that describe the state of the system is usually large. To obtain a closed set of equations, within the framework of a *well-posed problem*, we need an equal number of equations. However, the number of *primary variables* (or *degrees of freedom*) of the problem, is much smaller. The number of partial differential equations of balance that has to be solved is then equal to the number of the primary variables. All other variables are obtained from the known values of the selected primary variables, using the remaining equations. Bear and Nitao (1995) present a detailed discussion on primary variables in modeling phenomena of transport in porous media.

The case of flow and solute transport with variable density is presented in **Chapter 157, Sea Water Intrusion Into Coastal Aquifers, Volume 4**, in connection with seawater intrusion into coastal aquifers.

REFERENCES

Bear J. (1961) On the tensor form of dispersion. *Journal of Geophysical Research*, **66**, 1185–1197.

Bear J. (1972) *Dynamics of Fluids in Porous Media*, American Elsevier: (also Dover Publications, 1988).

Bear J. (1979) *Hydraulics of Groundwater*, McGraw-Hill.

Bear J. and Bachmat Y. (1967) A generalized theory on hydrodynamic dispersion in porous media. *IASH Symposium on Artificial Recharge and Management of Aquifers*, **72**, 7–16.

Bear J. and Bachmat Y. (1990) *Introduction to Modeling of Transport Phenomena in Porous Media*, Kluwer Academic Publishers.

Bear J. and Nitao J.J. (1995) On equilibrium and primary variables in transport in porous media. *Transport in Porous Media*, **18**(2), 151–184.

Bear J. and Verruijt A. (1987) *Modeling Groundwater Flow and Pollution*, D. Reidel Publishing Company.

De Josselin de Jong G. (1958) Longitudinal and transverse diffusion in granular deposits. *Transactions, American Geophysical Union*, **39**(1), 67–74.

Denbigh K. (1981) *Principles of Chemical Equilibrium: With Applications in Chemistry and Chemical Engineering*, *Third Edition*, Cambridge University Press: Cambridge.

Gelhar L.W. (1976) Stochastic analysis of flow in aquifers. *AWRA Symp. on Advances in Groundwater Hydrology*, Chicago.

Gelhar L.W., Gutjhar A.L. and Naff R.L. (1979) Stochastic analysis of macrodispersion in a stratified aquifer. *Water Resources Research*, **15**, 1387–1397.

Gvirtzman H. and Gorelick S.M. (1991) Dispersion and advection in unsaturated porous media enhanced by anion exclusion. *Nature*, **352**, 793–795.

Lichtner P.C. (1996) Continuum formulation of multicomponent-multiphase reactive transport. In *Reactive Transport in Porous Media*, *Reviews in Mineralogy*, Vol. 34, Lichtner P.C., Steefel C.I. and Oelkers E.H. (Eds.), Mineralogical Soc. of Am.: Washington.

Millington R.J. (1959) Gas diffusion in porous media. *Science*, **130**, 100–102.

Nikolaevskii V.N. (1959) Convective diffusion in porous media. *Journal of Applied Mathematics and Mechanics (P.M.M.)*, **23**, 1042–1050.

Poling B.E., Prausnitz J.E. and O'Connell J.P. (2000) *Properties of gases and Liquids*, *Fifth Edition*, McGraw Hill.

Saffman P.G. (1959) A theory of dispersion in a porous medium. *Journal of Fluid Mechanics*, **6**(3), 321–349.

Scheidegger A.E. (1961) General theory of dispersion in porous media. *Journal of Geophysical Research*, **66**, 3273–3278.

153: Groundwater Pollution and Remediation

WALT MCNAB, FRED HOFFMAN AND BRENDAN DOOHER

Environmental Restoration Division, Lawrence Livermore National Laboratory, Livermore, CA, US

Contamination of groundwater resources by a variety of anthropogenic pollutants from both point and non-point sources represents a key global environmental problem. Contaminant species of concern include solvents, fuel hydrocarbons, heavy metals, pesticides, nitrate, and radionuclides. Groundwater contamination reflects local physical hydrogeological considerations but also the inherent properties of the contaminants. An understanding of the factors controlling contaminant behavior in the subsurface is a necessary prerequisite to exploring possible remedial solutions. These factors include the partitioning processes that govern the distribution of contaminant species between the aqueous phase, the solid phase (including adsorption onto soil particle surfaces), the gas phase, and, in some cases, nonaqueous phase liquids. Chemical and biological transformation processes also impact the behavior of different types of subsurface contaminants. The designs of passive and active groundwater remediation approaches often reflect consideration of one or more of these factors; however, the efficacy of most methods is often most closely dependent upon the physical characteristics of the subsurface environment, such as the distribution of hydraulic conductivity.

INTRODUCTION: SOURCES OF GROUNDWATER CONTAMINATION

Human interest in groundwater has historically reflected water supply concerns, which are widely recognized as a major problem facing human civilization as the global population continues to increase. In recent decades, groundwater contamination has been recognized as a substantial threat to the viability of this resource worldwide. Because of the relatively slow rates of recharge and fluid migration through the subsurface environment, and because of the vast potential for exposure of subsurface materials to sources of pollution, groundwater contamination represents a critical societal–environmental resource problem. Examples of typical groundwater contaminants include industrial chemicals (solvents, fuels, heavy metals, radionuclides), agricultural chemicals (fertilizers, pesticides, nitrate), and municipal wastes (fecal coliform, nutrients, nitrate, disinfectants, detergents). The 25 most frequently detected groundwater contaminants at hazardous waste sites in the United States as of the mid-1990s (National Research Council, 1994) are listed in Table 1. The list is composed of predominantly organic solvents and compounds associated with petroleum products and manufacturing and inorganic contaminants, such as heavy metals, from industry and mining. The list does not include what has recently become a ubiquitous groundwater contaminant in the United States, the gasoline additive methyl-*tert*-butyl-ether (MTBE), nor does it include perchlorate, ClO_4^-, used in the manufacture of rocket propellants and road flares, which is also only now gaining scrutiny as a contaminant of concern. Often, the emergence of such "new" groundwater contaminants reflects issues such as inadequate monitoring technologies and incomplete analysis of already available monitoring data, in addition to the simple introduction of new chemicals (Mackay and Smith, 1993).

The entry points of pollutants into the environment may be distinguished as point sources or nonpoint sources. Examples of point sources include leaking underground pipes and tanks, accidental or deliberate spill events, and seepage from disposal pits and landfills. Such sources threaten groundwater resources directly at the point of entry, where transport processes may subsequently facilitate migration. Nonpoint sources include such examples as aerial and ground spraying of pesticides, urban precipitation runoff, or the cumulative effect of numerous point sources, such as in large industrial areas. These nonpoint sources threaten regional water resources on a larger scale. In a

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

Table 1 The 25 most frequently detected ground-water contaminants at hazardous waste sites (Reprinted with permission from (Alternatives for Ground Water Cleanup). © (1994) by the National Academy of Sciences, courtesy of the National Academies Press, Washington, D.C)

Rank	Compound	Common sources
1	Trichloroethylene	Dry cleaning; metal degreasing
2	Lead	Gasoline (prior to 1975); mining construction material (pipes); manufacturing
3	Tetrachloroethlylene	Dry cleaning; metal degreasing
4	Benzene	Gasoline; manufacturing
5	Toluene	Gasoline; manufacturing
6	Chromium	Metal plating
7	Methylene chloride	Degreasing; solvents; paint removal
8	Zinc	Manufacturing; mining
9	1,1,1-Trichloroethane	Metal and plastic cleaning
10	Arsenic	Mining, manufacturing
11	Chloroform	Solvents
12	1,1-Dichloroethane	Degreasing; solvents
13	1,2-Dichloroethane	Transformation product of 1,1,1-trichloroethane
14	Cadmium	Mining; plating
15	Manganese	Manufacturing; mining; occurs in nature as oxide
16	Copper	Manufacturing; mining
17	1,1-Dichloroethene	Manufacturing
18	Vinyl chloride	Plastic and record manufacturing
19	Barium	Manufacturing; energy production
20	1,2-Dichloroethane	Metal degreasing; paint removal
21	Ethylbenzene	Styrene and asphalt manufacturing; gasoline
22	Nickel	Manufacturing; mining
23	Di(2-ethylhexyl)phthalate	Plastics manufacturing
24	Xylenes	Solvents; gasoline
25	Phenol	Wood treating; medicines

Note: This ranking was generated by the agency for toxic substances and disease registry using ground-water data from the National Priorities List of sites to be remediated under CERCLA. The ranking is based on the number of sites at which the substance was detected in ground water. (National Research Council, 1994).

1984 report, the United States Office of Technology Assessment (Office of Technology Assessment, 1984) produced a guide for categorizing the different potential sources of contamination:

- *Category I: Sources designed to discharge substances.* Subsurface percolation (e.g. septic tanks and cesspools); injection wells (for hazardous wastes and nonhazardous wastes, such as for brine disposal and drainage, and nonwastes, such as enhanced oil recovery, artificial recharge, solution mining, and *in situ* mining), and land applications (e.g. spray irrigation with wastewater and by-products such as sewer sludge).

- *Category II: Sources designed to store, treat, and/or dispose of substances discharged through unplanned release.* Landfills (industrial and municipal sanitary), open dumps including illegal dumping, residential (or local) disposal, surface impoundments, waste tailings, materials stockpiles, graveyards; animal burial sites, aboveground and underground storage tanks, open burning and detonation sites, and radioactive disposal sites.

- *Category III: Sources designed to retain substances during transport or transmission.* Pipelines, materials transport, and transfer.

- *Category IV: Sources discharging as consequences of other planned activities.* Irrigation practices (e.g. return flow), pesticide applications, fertilizer applications, animal feeding operations, de-icing salts applications, urban runoff, percolation of atmospheric pollutants, mining and mine drainage (both surface and underground mine-related).

- *Category V: Sources providing conduit or inducing discharge through altered flow patterns.* Production oil and gas wells, geothermal and heat recovery wells, water supply wells, and other types of wells (monitoring wells, exploration wells, construction/excavation borings).

- *Category VI: Naturally occurring sources whose discharge is created and/or exacerbated by human activity.* Groundwater–surface water interactions, natural leaching, saltwater intrusion/brackish water upconing (or intrusion of other poor-quality natural water).

Reducing the threats to groundwater resources posed by these sources of pollution is a major challenge facing hydrologists, hydrogeologists, engineers, and planners. An understanding of the problem of groundwater contamination – and the responses taken with respect to it – consists of three basic elements: the physical and chemical factors

that influence the behavior of contaminants in the subsurface, the risk posed by the contaminants to human health and the environment, and the steps that can be taken to mitigate that risk in some form. This article explores each of these elements in turn.

CONTAMINANT FATE AND TRANSPORT CONCEPTS AND PROCESSES

When liquid contaminants are released at the ground surface, they are transported primarily downward through the vadose (unsaturated) zone by a combination of gravity, density, and capillary forces. When contaminants reach the groundwater (the saturated zone), they may dissolve or be further diluted and are then transported by the moving groundwater (Figure 1) at rates dependent on the hydraulic gradient and the hydraulic conductivity and porosity of the subsurface geologic materials, as well as by physiochemical factors which may retard the movement of the contaminant. Because of the complex nature of the subsurface environment, groundwater contamination is a difficult and expensive problem to characterize and remediate. This is true for porous media, which this article addresses, as well as for the special cases of fractured rock and fractured porous rock media, which will not be considered further (refer to Sahimi, 1995) (*see* **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**).

In addition to physical heterogeneities in the subsurface environment that affect groundwater flow and hence contaminant migration pathways, chemical heterogeneities also exist within the subsurface that will profoundly affect the behavior of contaminants. In the case of organic solvents and fuels, these include the possible presence of separate-phase liquids, the preferential partitioning of contaminants onto solid-phase materials in the aquifer matrix such as natural organic matter, and the existence of distinct biogeochemical microenvironments where contaminants may undergo chemical or microbially mediated degradation reactions. For inorganic contaminants such as metals and radionuclides, precipitation reactions that occur in response to changes in pH or redox conditions can immobilize contaminants, along with partitioning of solutes onto aluminosilicate or metal oxide mineral surfaces.

Partitioning Processes

Mass Transfer Between Aqueous Phase and NonAqueous Phase Liquids

Organic groundwater contaminants range in solubility from being infinitely soluble in water (e.g. ethanol) to being essentially insoluble (e.g. dioxins, high molecular weight aliphatic hydrocarbons); many common organic contaminants are water-soluble to some extent (Table 2). Low-solubility compounds are often introduced into the subsurface as nonaqueous phase liquids, or NAPLs (otherwise referred to as separate-phase liquids). The partitioning of mass between the aqueous phase and NAPL is a critical issue because measurable health effects associated with consumption of some organic contaminants by humans, animals, and plants may occur at very low concentrations (part-per-billion levels), so that even sparingly soluble compounds may pose a substantial environmental threat, and because these contaminants entrapped in the subsurface may serve as continuous sources of aqueous-phase contamination over long periods. The primary factor controlling the solubility of NAPL components in water is the electrical charge distribution across the molecule, or polarity, quantified by the dipole moment. Water molecules, because of the geometric orientation of the constituent oxygen and hydrogen atoms, have a high dipole moment compared to other liquids of similar molecular weight and thus will have a much higher affinity for one another than for lower dipole moment molecules. This results in a segregation of water molecules from the lower dipole moment molecules.

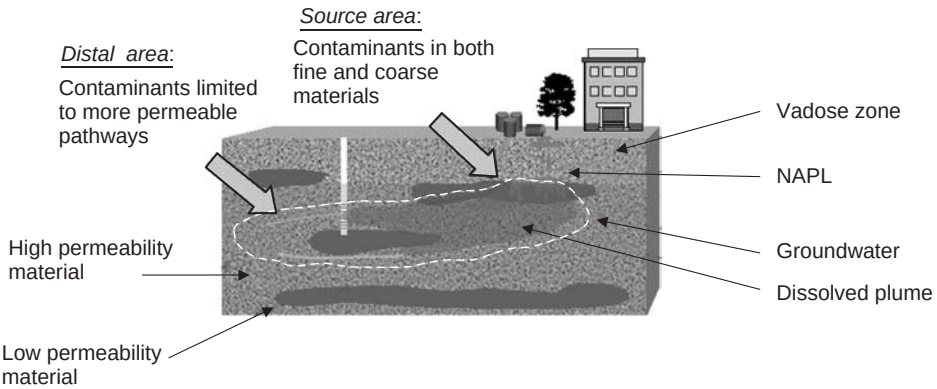


Figure 1 Conceptual release scenario for a NAPL penetrating the unsaturated zone and impacting underlying groundwater. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Table 2 Physical properties and cleanup levels of selected common organic groundwater contaminants

Compound	Density ($gm\,cm^{-3}$)[a]	Water solubility ($mg\,L^{-1}$)[b]	Vapor pressure (mm Hg)[b]	Henry's constant ($atm \cdot m^3\,mol^{-1}$)[b]	US EPA MCL ($\mu g\,L^{-1}$)[c]
Trichloroethylene	1.61	1100	58	9.1×10^{-3}	5
Tetrachloroethlylene	1.4	150	18	2.6×10^{-2}	5
1,1,1-Trichloroethane	1.33	1500	123	1.4×10^{-2}	200
1,2-Dichloroethane	1.25	8520	64	9.8×10^{-4}	5
Chloroform	1.48	8200	151	2.9×10^{-3}	80[d]
Benzene	0.88	1750	95	5.6×10^{-3}	5
Toluene	0.87	54	28	6.4×10^{-3}	1000
Chlorobenzene	1.11	466	117	3.7×10^{-3}	100
Pentachlorophenol	1.98	14	1.1×10^{-4}	2.8×10^{-6}	1
Napthalene	0.96	32	0.23	1.2×10^{-3}	–

[a]CRC Press (1991).
[b]Mercer *et al.* (1990).
[c]US EPA (2003b).
[d]Note: As total trihalomethanes.

Moreover, NAPLs cannot readily displace water in unsaturated soils by capillary effects because water generally exhibits a higher affinity for mineral surfaces than the organic liquids. As a result, NAPLs will often reside within larger pore spaces unoccupied by water. Under such circumstances, discontinuous bodies of NAPL may reside in the vadose zone, unable to migrate. Such NAPL bodies may slowly dissolve into the aqueous phase, acting as local contamination sources for a considerable period of time.

Aside from solubility considerations, the density of a NAPL, in comparison to that of water, is also an important consideration with regard to the fate of sparingly soluble contaminants in the environment. A number of common organic contaminants are characterized by densities greater than that of water (e.g. chlorinated halocarbons such as PCE and TCE), while others, particularly petroleum hydrocarbons, are significantly less dense than water (Table 2). Heavier-than-water organic liquids are referred to as DNAPLs (dense nonaqueous phase liquids), while lighter-than-water organic liquids are LNAPLs (light nonaqueous phase liquids). NAPLs will tend to float on top of the water table in an unconfined aquifer, whereas DNAPLs will tend to sink into the saturated zone. Different remediation strategies must be utilized for cleaning up these two classes of liquid-phase contaminants in the environment.

Mass Transfer Between Aqueous and Gas Phases

Lighter, volatile organic contaminants, either entrapped within the vadose zone or floating on the water table, may volatilize directly into the gas phase in unsaturated pore spaces. The volatility of a compound is indicated by its vapor pressure, which is the partial pressure of a vapor in equilibrium with its pure liquid state. Vapor pressures for some common organic contaminants are listed in Table 2. Many common organic contaminants are characterized by relatively high vapor pressures and are commonly referred to as volatile organic compounds (VOCs).

When organic compounds are dissolved in the aqueous phase (i.e. in soil water or groundwater), partitioning of mass between the aqueous phase and the soil gas phase is commonly modeled by Henry's law, which states that, at constant temperature, the vapor pressure of a compound is equal to the compound's mole fraction in the aqueous phase times a constant,

$$P_a = h_a X_a \tag{1}$$

where P_a is the vapor pressure of substance a, h_a the respective Henry's constant, and X_a the mole fraction of a in solution. This relationship is applicable in the ideal case of an infinitely dilute solute, where solute–solvent interactions are independent of concentration, so that Henry's law is most applicable when the solute concentration is fairly dilute. Henry's law constants for some common organic contaminants are given on Table 2.

Mass Transfer Between the Aqueous and Solid Phases

Mass transfer phenomena between the solid and aqueous phases include adsorption/desorption of dissolved species onto solid surfaces as well as dissolution or precipitation of mineral phases. These so-called water-rock interactions occur in response to a groundwater solution that is in chemical disequilibrium with solid materials in the aquifer, often the result of transport mechanisms bringing in a suite of solutes originally in equilibrium with a particular mineral assemblage into a portion of an aquifer system containing different minerals. Water-rock interactions are very significant with regard to groundwater contamination. For example, the continuous partitioning of solutes between the aqueous and solid phases can result in depletion of the aqueous solute at the leading edge of a migrating plume

and enrichment at the trailing edge, leading to an overall retardation of plume movement.

Adsorption reactions include a number of mechanisms, including complexation of metals onto mineral surfaces or the hydrophobic partitioning of organic contaminants onto organic materials present in the soil. Many common mineral phases found in geological materials tend to develop surface charges as a result of ionic substitution reactions in crystal lattices or crystal edge protonation/deprotonation reactions. These phenomena are especially prevalent in aluminosilicate clay minerals, which generally develop negative surface charges as a result of lattice substitution, and metal oxyhydroxides such as ferrihydrite, $Fe(OH)_3$, which develop variable surface charges (negative or positive) as a result of protonation/deprotonation reactions that are pH-dependent. Cations and anions in solution tend to redistribute in the vicinity of the mineral surface to satisfy charge imbalances. For ionic contaminants such as dissolved metals, this redistribution can take the form of chemical bonding to the mineral surface (either directly, or via hydrated complexes) or residence in a diffuse layer of ions near the surface (Sposito, 1989).

In contrast to ionic species, sparingly soluble (i.e. "hydrophobic") organic compounds generally exhibit a high affinity for naturally occurring solid-phase organic matter (e.g. cellulose, lignins, kerogen, and humic acids) present in soils rather than for charged mineral phases. Dissolved organic compounds tend to adsorb onto such materials either because of unfavorable free energy costs associated with remaining in solution or because of van der Waals or other weak electrostatic forces (Schwarzenbach *et al.*, 1993). In essence, adsorption of hydrophobic organic compounds to soil organic material is more a matter of the incompatibility of nonpolar compounds with water, rather than a direct attraction to the soil organic material (Westall, 1987). The tendency of an organic compound to partition onto natural organic matter may be gleaned from its organic carbon partition coefficient, or K_{oc}, which reflects the degree of nonpolarity, or hydrophobicity, of the compound and thus tends to correlate with the solubility of the compound (Table 2).

A simple model that is often employed for quantifying the adsorption of dissolved species onto a solid surface is based upon a distribution coefficient, or K_d:

$$K_d = \frac{C_s}{C_w} \tag{2}$$

where C_s is the sorbate concentration associated with the sorbent, typically in mol/kg, and C_w the aqueous concentration, typically in mol/L. For hydrophobic organic compounds in aquifers characterized by abundant solid-phase organic material, the K_d can be estimated as the product of the K_{oc} and the fractional organic carbon content of the sediments. For inorganic species, a variety of molecular adsorption models exist to quantify relationships between the adsorbate (the sorbed species) and the adsorbent (the solid surface), including the so-called diffuse double-layer model, which assumes that the adsorbent surface is a uniform plane of uniform charge density and that the adsorbate ions are point sources that interact with the adsorbent surface through electrostatic forces. Other approaches seek to quantify adsorption through the definition of conditional equilibrium constants (Sposito, 1989).

The same mechanisms responsible for the partitioning of contaminant species onto solid materials may also result in adsorption of these same species onto suspended colloidal particles, which are fine (nanometer to micron-size) particles of clay minerals, metal hydroxides, or organic material. This phenomenon is potentially important from an environmental viewpoint because some contaminants which would normally be expected to exhibit very low mobility in the subsurface as a result of adsorption might be transported over long distances (McCarthy and Zachara, 1989). For example, colloidal transport may be the likely explanation for the apparent migration of plutonium, a radionuclide that strongly sorbs to mineral surfaces, over hundreds of meters at the United States Department of Energy's Nevada Test Site (Kersting *et al.*, 1999).

In addition to adsorption reactions, the mobility of a variety of inorganic contaminants may be substantially influenced by precipitation and dissolution reactions (Stumm and Morgan, 1996; Parkhurst and Appelo, 1999). These types of reactions will occur when transport processes bring together a solution chemistry and a mineral assemblage that are not in mutual thermodynamic equilibrium. Disequilibrium provides the driving force for mineral precipitation reactions once a particular mineral phase becomes supersaturated in a solution. Conversely, a state of thermodynamic undersaturation in a solution with respect to a given mineral phase will provide a driving force for dissolution of mineral phases that may be present within aquifer material. For example, in the phenomenon of acid mine drainage, oxygenated rainwater in contact with iron sulfide-rich mine tailings oxidizes the iron and sulfide components, yielding a solution dominated by H^+, SO_4^{2-}, and metal cations. The effluent solution migrates into the subsurface where it reacts with carbonate minerals such as calcite, $CaCO_3$, that are unsaturated in acidic waters to produce secondary mineral phases not originally present in the aquifer, such as gypsum, $CaSO_4 \cdot 2H_2O$. Further downgradient, as carbonate dissolution reactions mitigate the acidity of the mine tailings effluent, iron hydroxide mineral phases such as ferrihydrite, $Fe(OH)_3$, become supersaturated and precipitate out of solution. This can provide sorption sites for dissolved metals such as lead, copper, or zinc.

A summary of the susceptibility of important metal and radionuclide groundwater contaminants to various mineral equilibria and adsorption constraints is shown in Table 3.

Table 3 Selected metals and radionuclides and potential water-rock interaction effects

Species	Water-Rock Interactions[a]
Lead	Low solubility of lead hydroxyl carbonates and phosphate, adsorption onto soil minerals and organic matter, coprecipitation with manganese oxide.
Chromium	Cr(VI), typically as CrO_4^{2-}, exists under oxidizing conditions and is highly mobile. Reducing conditions yield Cr(III), which is less mobile, generally precipitating as $Cr(OH)_3$ or coprecipitating with iron hydroxide phases.
Zinc	Mobility is limited at slightly to fairly alkaline pH values by formation of carbonate and hydroxide mineral phases and sorption to soil minerals.
Arsenic	Arsenate, As(V), exists under oxidizing conditions and is generally mobile, forming anionic $H_2AsO_4^-$ or $HAsO_4^{2-}$. Reducing conditions yield arsenite, As(III), which sorbs strongly onto iron oxyhydroxide minerals.
Cadmium	Mobility is limited at neutral to alkaline pH values by formation of cadmium carbonate and sorption to soil minerals; coprecipitation with manganese oxide.
Manganese	Very low solubility under oxidizing conditions where MnO_2 stable.
Copper	Cupric ion, Cu^{2+}, exists under oxidizing conditions; solubility is controlled by carbonate and hydroxide mineral phases and sorption to soil mineral phases above neutral pH. Reducing conditions yield cuprous ion, Cu^+, which forms low-solubility sulfide mineral phases.
Barium	Solubility controlled by the formation of $BaSO_4$; sorbs to metal oxides or hydroxides.
Nickel	Sorbs to soil mineral phases; may form nickel oxide or hydroxide mineral phases.
Uranium	U(IV) exists under oxidizing conditions and is generally mobile, forming species such as uranyl, UO_2^{2+} or carbonate or sulfate complexes. Reducing conditions yield U(IV), which typically forms the low-solubility mineral phase uranite, UO_2.
Plutonium	Generally sorbs strongly to soils, depending on pH, clay content, and presence of soil organic matter.[b]
Cesium	Sorbs to soil mineral phases.
Technetium	Pertechnate, TcO4−, is the dominant form under oxidizing conditions; sorbs weakly to mineral surfaces depending on pH and other anions. Reduced conditions produce sparingly soluble Tc(IV) which is readily sorbed by soil constituents and forms complexes with organic matter.[b]

[a]Hem (1985),
[b]Zhang *et al.* (2002).

Transformations

In contrast to mass exchange processes, which affect the distribution of contaminant mass between phases, transformation reactions affect the total mass of a particular species in the subsurface environment. Transformation processes include those mechanisms that act to destroy a contaminant altogether, converting it into another material. Examples include radioactive decay and the biotransformation of organic compounds by microorganisms.

Microbially Mediated Transformations of Organic Contaminants

Dissolved organic compounds in groundwater under ambient environmental conditions are inherently thermodynamically unstable and are energetically favored to degrade into stable species such as CO_2 and H_2O via oxidation, or, in certain instances, CH_4 via reduction. Oxidation and reduction reactions generally entail the breaking of covalent bonds, an energy-intensive process. The energy associated with the random, thermal agitation of molecules is often insufficient to allow such reactions to progress at measurable rates even when the reaction is thermodynamically favorable. However, numerous species of microorganisms, particularly bacteria (e.g. Bacillus, Pseudomonas), and yeasts possess enzymes that are capable of facilitating the chemical breakdown of organic contaminants through various reaction pathways. These microorganisms manufacture enzymes such as monooxygenase, dioxygenase, and dehydrogenase to assist, for example, in transferring oxygen atoms to a molecule or for removing hydrogen atoms, and, as such mediate both oxidative and reductive processes. Nevertheless, not all organic contaminants are easily oxidized or reduced via microbial activity; many artificially synthesized organic compounds are deliberately designed to be recalcitrant to transformations.

Oxidation reactions necessarily require a suitable electron acceptor to accommodate the electrons released:

$$C_6H_6 + 12H_2O \longrightarrow 6CO_2 + 30H^+ + 30e^- \qquad (3)$$

The primary electron acceptor in natural waters is dissolved oxygen. Reactions between O_2 and organic compounds are usually highly favorable thermodynamically. Under anoxic conditions (i.e. in the absence of dissolved O_2), other electron acceptors in the aqueous phase (e.g. NO_3^-, SO_4^{2-}) or the solid phase (e.g. $Fe(OH)_3$, MnO_2) may be utilized instead by microorganisms. Under extremely reducing conditions where all other electron acceptors have been depleted, CO_2 itself may act as the electron acceptor in reactions that yield methane as a

reaction product. Evidence of the associated changes in groundwater chemistry in response to the oxidation of organic contaminants has been most frequently encountered in instances of fuel hydrocarbon groundwater plumes. Fuel hydrocarbon compounds are common contaminants that are often released in high concentrations and are relatively susceptible to oxidation by populations of microorganisms utilizing a variety of electron acceptors. Laboratory and field studies have demonstrated the sequential microbial utilization of the common electron acceptors, beginning with dissolved O_2, followed, roughly in order of thermodynamic favorability, by MnO_2, NO_3^-, Fe(III)-oxyhydroxides, SO_4^{2-}, and occasionally CO_2, resulting in zonation of oxidation–reduction processes, frequently noted at field sites (Vroblesky and Chapelle, 1994).

Although many microbially mediated transformation processes involving organic contaminants are oxidative in nature, important reduction reactions also occur in certain circumstances. Reductive dehalogenation, which involves the replacement of halogen atoms (Cl, Br, I) in a halogenated hydrocarbon compound with protons, is the most familiar reductive transformation process. Because halogen atoms are generally characterized by negative electrical charges, carbon atoms associated with halocarbon compounds are relatively oxidized (i.e. less negatively charged) in comparison to their counterparts in nonhalogenated hydrocarbons. As a result, halogenated hydrocarbons are susceptible to reduction reactions, with the most heavily halogenated, and hence the most oxidized, the most susceptible to this process.

Many common halogenated hydrocarbons will undergo reductive dehalogenation under chemically reducing conditions (McCarty, 1996), including halogenated methanes, ethenes, ethanes, and aromatics. The reductive half-reaction of the dehalogenation reaction may be expressed generically as,

$$RX + H^+ + 2e^- \longrightarrow RH + X^- \qquad (4)$$

where R refers to the organic molecule and X the halogen atom. Reaction (4) implies that a suitable electron donor, typically organic carbon, must be readily available. As in the case of oxidative transformations, microbial mediation is usually necessary for reductive dehalogenation to proceed at significant rates (i.e. half-lives on the order of days to months). A well-known exception is in the presence of metallic iron, where the metal acts as both reducing agent and catalyst for the relatively rapid reductive dehalogenation of a number of common halogenated hydrocarbons, notably TCE.

Abiotic Transformations of Organic Contaminants: Some Examples

Oxidative and reductive pathways constitute an important means for transformations of organic contaminants in groundwater. However, other somewhat specialized mechanisms do exist in some instances; often they do not entail microbial mediation. For example, substitution reactions involve the replacement of an atom or functional group on a molecule by another. Hydrolysis reactions, a familiar class of substitution reactions, occur when an organic compound reacts directly with the water to produce an alcohol or a diol and a proton, such as the conversion of methyl chloride to methanol:

$$CH_3Cl + H_2O \longrightarrow CH_3OH + H^+ + Cl^- \qquad (5)$$

Beyond substitution reactions, dehydrohalogenation refers to a transformation mechanism specific to certain halogenated alkanes, entailing the simultaneous removal of a halogen atom and a proton from the molecule, along with the conversion of a single carbon–carbon bond to a double bond, forming an alkene. The one common example of this process is the dehydrohalogenation of 1,1,1-trichloroethane (TCA) to form 1,1-dichloroethene (DCE),

$$CCl_3–CH_3 \longrightarrow CCl_2{=}CH_2 + H^+ + Cl^- \qquad (6)$$

This reaction has been documented in laboratory experiments and observed at contaminated groundwater sites and may be responsible for much of the 1,1-DCE found in aquifers (McCarty, 1996). The half-life of this reaction is on the order of one to two years, depending in large part on temperature.

Many organic contaminants may also undergo reversible reactions that entail simple protonation or deprotonation, so that the compounds act as weak acids or bases in solution. These reactions typically involve functional groups (e.g. carboxyl, carbonyl, and phenyl groups) that exist as portions of a larger organic molecule. Protonation or deprotonation reactions are important in terms of the behavior of particular organic compounds in the environment. For example, charged species may interact with solid-phase mineral and organic surfaces through adsorption mechanisms that differ from uncharged equivalents and may also be characterized by different toxicities or biodegradability.

GROUNDWATER CONTAMINATION: A RISK-BASED PERSPECTIVE

An early introduction for the public to the looming danger that groundwater contamination poses to water quality was provided by the Love Canal hazardous waste site in Niagara Falls, New York, USA., which came to public attention in 1979. In that event, contaminated land had been turned over to the local school board with full deed restrictions but was subsequently used to build not only a school but a housing community and the associated sewage systems,

resulting in the piercing of the landfill's clay liner (Zuesse, 1981). By June 1994, Occidental Petroleum, the custodian of the site, had paid $98 million to cover the state of New York's cleanup costs and by December 1995 they had paid $129 million to cover the federal government's share (MacDonald, 2000). In 1996, a subsidiary of Occidental Chemical assumed full operation and maintenance of the chemical waste treatment plant at the Love Canal site, implying that the site was still not completely remediated at that time.

It was contamination at Love Canal, the associated human health consequences, and the cleanup costs that energized the creation of the Resource Conservation and Recovery Act (RCRA) and the Comprehensive Emergency Response, Compensation, and Liability Act (CERCLA) legislation in the United States to assure that such sites were properly addressed. Implicit in this legislation, which has driven a large portion of groundwater cleanup efforts in the United Sates, is the concept of the risks posed by groundwater contamination. Ultimately, the danger posed by contamination of groundwater resources to human health and the environment can ideally be quantified through an assessment of risk. Risk may be defined in terms of the potential for loss, harm, or damage, and the degree of probability of such harm (Kaplan and Garrick, 1981). Hazard represents the source of danger. In the context of environmental systems, hazards that may create risks are toxins, nuclear materials, and so on. Safeguards may include physical, chemical, or biological barriers to passage, regulatory action, or avoidance of the hazard. Quantitatively, risk is further defined succinctly as a series of three steps: (i) scenario identification or description (i.e. what can go wrong?), (ii) the probability of the scenario (i.e. how likely is it to happen?), and (iii) the consequence or measure of damage (i.e. what are the consequences if it does happen?). Defining the probabilities involved with the hazard, the transport of the hazard (or pathway), and the exposure parameters creates a broader understanding of the risk. For contaminants, risk may be described as the probability that a certain concentration of toxic material will reach a receptor and cause adverse health effects. In 1983, the United States National Academy of Sciences developed a risk assessment paradigm (National Research Council, 1983) that consists of four components:

- *Hazard identification:* The analysis of the physical and chemical properties of the contaminant, the metabolic properties and toxicological effects of the contaminant, including data from both long and short-term animal studies, and the routes of exposure.
- *Dose-response assessment:* The quantification of risk versus dose, confirmation of the effects of the toxin on humans, and extrapolation from the high dosages given to animal test subjects to the low dosages expected to be seen in humans.
- *Exposure assessment:* The analysis of the information for each chemical or mixture, possible sources, the exposure pathways and environmental fate, the measured or estimated concentrations, the exposed populations, and finally an integrated exposure analysis.
- *Risk characterization:* The numerical estimation of risk using the concept of unit risk, the dose corresponding to a given risk, the individual probability of the risk, and the framework to judge the significance of the risk.

Given this paradigm, the simplest approach to risk assessment entails establishing a proscribed concentration threshold for contaminant concentrations in groundwater that is protective of human health, by some definition, regardless of potential exposure pathways. For example, the United States EPA has proscribed Maximum Contaminant Levels (MCLs) for various chemical contaminants (Table 2) that are founded on a generalized risk assessment methodology, taking into account epidemiology studies performed on animals. Although there are advantages for using approaches such as MCLs to establish groundwater contaminant cleanup goals as opposed to site-specific risk assessments, such as consistency and simplicity, these advantages are fast disappearing as risk modeling methodologies improve, particularly through use of databases and geographic information system (GIS) approaches.

In contrast to the general applicability offered by the MCL approach or similar proscribed concentration thresholds, a major advantage for site-specific risk assessments is flexibility in considering the site conditions and the likelihood that receptors such as water resources or humans will be exposed to risk. For example, risk-based goals at groundwater contamination sites in the United States typically range from levels of 10^{-4} to 10^{-6} excess cancer risk over a lifetime from a continuous exposure to a suspected carcinogen contaminant. For noncarcinogens, a hazard index is often developed to quantify the potential for causing harm; cleanup goals are implemented to maintain a low hazard index. Other pseudo risk-based approaches include remediation to background levels or analytical detection limits of the contaminants of concern.

Although risk-based cleanup goals provide a structured framework for implementing remedial action plans, a number of factors may complicate application at any given groundwater contamination site. These may include taste and odor threshold concentrations, which may or may not correspond to toxicity effects, as well as economic and cost-benefit considerations for various remedial options. In cases where institutional controls such as land use restrictions are implemented as part of a remedial solution, the potential may exist for contaminant plumes to migrate under an adjoining property, resulting in property owners potentially

assuming an added liability, with a corresponding detrimental effect on property values, or the institutional controls my be simply ignored, as in the Love Canal example. At sites where analytical detection limits are established as the cleanup threshold, reductions in detection limits at some later date because of technological advances can potentially result in sites that have been declared "clean" being reopened. Finally, conventional risk-based cleanup strategies often fail to assess the costs and benefits of aquifer remediation in comparison with simple wellhead treatment options.

REMEDIATION TECHNOLOGIES

Ideally, strategies to remediate groundwater take into consideration an understanding of the physical and chemical processes that affect contaminant behavior as well as an assessment of the risks posed by contamination in terms of sources, exposure pathways, and receptors. Thus, in principle, many remediation approaches incorporate both technical and administrative components.

Remediation approaches may be grouped broadly into two general categories: active methods and passive methods. Active methods entail the application of engineered technology to physically remove contaminated groundwater for treatment or, alternatively, to actively destroy contaminants *in situ* (refer to Kobus *et al.*, 1996). Passive methods rely upon natural processes in groundwater to mitigate contamination without direct human intervention, other than monitoring. Some of the more widely used active and passive remediation approaches are discussed below.

ACTIVE METHODS

Pump and Treat

"Pump and treat" is the term used to describe the most common technology in use today to remediate contaminated groundwater (National Research Council, 1994). In the simplest case, extraction wells are installed with well screens at the depth of the groundwater contamination. The wellfield is designed such that the placement of the wells and the groundwater pumping rates provide a zone or zones of capture that will remediate the plume. The contaminated groundwater is pumped to a surface-based treatment facility designed to remove the particular contaminant or contaminants from the water. The treated effluent is then discharged in such a manner that it does not interfere with the pumping wellfield, or enhances the capture of the contaminated plume.

In designing an optimal remediation wellfield, it is critical to maintain and sample sufficient monitor wells to understand the effect of the pumping wellfield. Identification of capture zones and how they change with stress on the wellfield, identification of stagnation zones and how they change with individual well pumping rates, and monitoring of contaminant concentrations throughout the plume provide the necessary data to manage the efficient operation of the cleanup. In response to monitoring data, an efficient remediation scheme, designed to cleanup the groundwater as quickly as possible, will alter pumping rates, turn extraction wells on or off, and install new extraction wells at locations that are not being remediated (Hoffman, 1993). Methods for evaluating pump and treat system performance, and tools for creating optimal efficient pumping schemes are discussed in Gorelick *et al.* (1993) and Fetter (1993).

Because of the different physical and chemical characteristics of the individual areas of contaminant plumes and because of the varying public health and environmental risks posed by individual plumes, different pump and treat strategies may be applied (Hoffman *et al.*, 2003). Source areas often contain contaminants as separate-phase or high-concentration dissolved chemicals in both coarse and fine-grained sediments (high- and low permeability materials, respectively) or other aquifer materials. Since it is difficult to hydraulically stress low permeability geologic materials, a pump and treat strategy of hydraulic containment of the source area is often appropriate. Extraction wells are placed downgradient of the source area in such a fashion that the entire source area falls within the capture area of the extraction wellfield. Such a strategy prevents any more contaminant mass from migrating into the distal plume, away from the source area, remediates the contaminated permeable materials within the source area, and begins the remediation of the contaminants in the less permeable zones at a slower rate based on their diffusion rates from the low permeability zones into the permeable zones. For those plumes where it is necessary to prevent any further down-gradient movement, such as those in the vicinity of groundwater supply wells, an extraction wellfield at the leading edge of the plume is appropriate. Such a wellfield is designed to create a capture zone large enough to protect the groundwater resource from the encroaching contamination.

With the source area and leading edge hydraulically controlled, an extraction wellfield can be designed to cleanup the remainder of the distal plume expeditiously. Since this part of the plume is essentially limited to the permeable aquifer materials, a well-designed and monitored wellfield can be operated to remediate the distal plume to specified cleanup levels in much less time than it took the groundwater to be contaminated in the first place. Under this scenario, the source area will take much longer to remediate, but the hydraulic control will hold the contaminants in place pending the deployment, development, or improvement of a technology that will accelerate the extraction of the contaminants from the finer-grained materials. At the time of this writing, source

area remediation remains one of the leading challenges for groundwater cleanup.

Soil Vapor Extraction

Soil Vapor Extraction (SVE) is a cleanup technology that may be useful as part of the cleanup of source areas that are contaminated with volatile components. SVE entails the installation of wells, screened partially or fully within the vadose zone, designed to pump soil vapors, including the contaminants, out of the subsurface. Investigations of the probable source areas are conducted using soil vapor surveys (SVS). An SVS can include active pumping and analysis of vapors from vadose zone wells or passive surveys using devices that are placed in the subsurface to collect contaminants that are being transported by vapors through the vadose zone. These devices are collected and analyzed following a specified deployment period.

If soil vapor transport modeling, using the results of the SVS, indicates that the concentrations of contaminants in the soil vapor could cause an inhalation risk or could eventually result in the contamination of groundwater at concentrations that exceed specified action levels, then SVE may be appropriate for the site provided that the soil air permeability is sufficiently high to allow for efficient operation of the SVE system (Johnson *et al.*, 1990). Soil vapors are extracted from the subsurface, and contaminants are removed or destroyed by various treatment technologies such as activated carbon absorption or combustion.

Permeable Reactive Barriers

Permeable reactive barriers (PRB) involve the emplacement of a physical, biological, or chemical reagent in the pathway of a contaminant plume such that the contaminants are treated as they pass through the barrier. The transport of the contaminants to and through the PRB can be dependent on the natural gradient of the groundwater or enhanced by stressing the aquifer containing the plume through down gradient pumping. A physical reactive agent is designed to sorb the contaminant to prevent its continued movement, while the chemical and/or biological reactants may transform the particular contaminant into nontoxic, nonhazardous, or nonmobile components. A familiar example of the PRB application is the placement of walls of elemental iron filings in the subsurface to chemically reduce contaminants such as chlorinated hydrocarbons (reductive dehalogenation) or uranium (reduction and immobilization). Specific technical issues and deployment experiences with elemental iron-based PRBs are discussed by Gavaskar *et al.* (1998).

PRBs can be emplaced in trenches for shallow systems or injected into place for deeper systems. Injection systems vary from injections at moderate pressures to displace pore water with the reactant, to high-pressure injection systems that inject a combination of reactant and permeable materials. Some PRB designs use the "funnel and gate" technology. "Funnels" consisting of impermeable barriers, such as sheet pilings or slurry walls are installed to direct the contaminant plume to the "gate" where the PRB is emplaced. For shallow systems, the PRB can be installed in a removable cartridge or cartridges whose reactants can be replenished. Multiple cartridges containing different reactants can be installed to accommodate multicomponent plumes. Establishing a biological reactive barrier may include emplacement of nutrients to promote the growth of indigenous microbes or emplacement of colonies of nonnative microbes tailored to the particular contaminant.

The advantages of PRBs are that once they are emplaced, the surface can be restored to its previous beneficial use, and the system itself requires little if any maintenance, and does not consume any resources such as electric power. The disadvantages include the dependence on the slow movement of groundwater to deliver the contaminants to the reactant, and the finite life of the reactant (National Research Council, 1994).

Other Remediation Technologies

Other remediation technologies include chemical injection flooding, augmented bioremediation, thermal techniques, and surfactant flooding. Chemical injection and bioremediation techniques include flooding the area of contaminated plumes with reagents, nutrients, or microbes designed to degrade, destroy, or immobilize contaminants. Surfactants are chemical agents designed to increase the mobility of low-solubility contaminants so that other technologies like pump and treat can be made more effective. The success of this approach is dependent on achieving contact between the surfactant and the contaminant. Details of these and other technologies are provided in Pankow and Cherry (1996) and National Research Council (1999). Thermal techniques such as steam flooding to flush contaminants from one area to an area of extraction, and joule heating to volatilize and increase the mobility of some contaminants, have also been demonstrated (Newmark, 1994).

The uncertainties accompanying these various techniques include the physical mechanisms of ensuring that the emplaced materials will come into contact with the contaminants to allow them to do their work, and, in the case of thermal technologies, uncertainties related to controlling the system once it is implemented.

Passive Methods

Monitored Natural Attenuation

Monitored natural attenuation (MNA) is a broad remedy that may be useful (or acceptable) in those circumstances where natural existing processes in the subsurface are causing the contaminants to degrade, become immobilized, or in other ways not causing an unacceptable risk to human

health or the environment (National Research Council, 2000). The MNA remedy must include an understanding of the physical, chemical, and/or biological processes that reduce the risk of the contaminant and a monitoring network that will continue to ensure that the risk does not increase. MNA approaches are perhaps most familiar in the case of fuel hydrocarbon compounds, such as at leaking underground fuel tank sites, where numerous studies have demonstrated that natural biodegradation processes act to limit the downgradient movement of contaminant plumes (Rice *et al.*, 1995).

In considering instances when MNA may provide a viable remedial option, a number of factors must be taken into consideration. The technical issues that must be resolved include whether a contaminant is destroyed (e.g. fuel hydrocarbons) or merely immobilized by natural processes (e.g. metals), whether the rate of attenuation is sufficiently rapid, whether chemical or biological attenuation mechanisms could produce toxic by-products (e.g. vinyl chloride from the reductive dehalogenation of TCE or PCE), and whether a mixture of compounds may exist of varying susceptibility to attenuation mechanisms (e.g. fuel hydrocarbons and MTBE). Moreover, some decision-assisting protocol must be in place to provide standards for evaluating the validity of proposed attenuation processes at a given site and to determine when MNA provides a suitable reduction of risk. Finally, public concerns in regard to an MNA solution, such as perceptions of continued health risk or threat to property values, must also be addressed.

Institutional Controls

Institutional controls represent not a technology *per se* but rather an administrative approach to reducing risk. Instead of relying upon engineered or natural processes to mitigate contaminant sources or to interfere with contaminant travel pathways, institutional controls seek to remove the receptor from the risk equation. In practice, this is attempted through the use of devices such as deed restrictions, where a landowner is bound not to use a property in a manner that would expose humans to contaminants, or groundwater resource management plans that might control the beneficial use of the pumped water, control the location of extraction wells, or control pumping rates. Institutional controls are often used in tandem with MNA approaches as a means of insuring protection from contaminated groundwater in the event that the MNA solution, which could require years to achieve results, does not perform in an anticipated manner. However, it is important to recognize that institutional controls are only effective when properly recognized and enforced; the case of Love Canal exemplifies the danger of assuming that institutional controls intended to be protective of public health are easily enforced.

CONCLUSION

As a field of study, contaminant hydrogeology has come into its own over the last 50 years as groundwater investigations and new, more sensitive, chemical analytical procedures have been developed, and groundwater problems have been discovered. In this article, the extent of the problems of groundwater contamination around the globe has been discussed and some of the physical and chemical processes that govern transport of contaminants in the subsurface have been described. Many of the common technologies for investigating and cleaning up contaminated groundwater have been introduced and the strengths and limitations of each approach described. While great strides have been made in controlling groundwater contaminant plumes, the challenges of understanding the subsurface environment in chronically undersampled situations continue to provide room for considerable improvement in the field. In the last decade, much of the research in the field has focused on the *in situ* treatment of groundwater contaminants, and the next decade should provide full-scale field demonstrations of their efficacy. Another area of needed study is the behavior of contaminants in fine-grained geologic materials in contaminant source areas and the development of source area cleanup technologies. With the continued improvement of the efficiency of existing groundwater cleanup technologies and the development of new source area and *in situ* technologies, it is hoped that groundwater resources will be restored to a quality that allows it to take its place in responding to future crises in the growing global problems of water shortages.

REFERENCES

CRC Press (1991) In *Handbook of Chemistry and Physics*, *Seventy Second Edition*, D.R. Lide (Ed.) CRD Press: Boca Raton.

Fetter C.W. (1993) *Contaminant Hydrogeology*, *Second Edition*, Prentice Hall: Upper Saddle River.

Gavaskar A.R., Gupta N., Sass B.M., Janosy R.J. and O'Sullivan D. (1998) *Permeable Barriers for Groundwater Remediation: Design, Construction, and Monitoring*, Battelle Press: Columbus.

Gorelick S.M., Freeze R.A., Donohue D. and Keely J.F. (1993) *Groundwater Contamination: Optimal Capture and Containment*, Lewis Publishers: Boca Raton.

Hem, J.D., 1985 *Study and Interpretation of the Chemical Characteristics of Natural Water*, U.S. Geological Survey Water-Supply Paper 2254: Alexandria.

Hoffman F. (1993) Groundwater remediation using "Smart Pump and Treat". *Groundwater*, **31**(1), 98–106.

Hoffman F., Blake R.G., Demir Z., Gelinas R.J., McKereghan P.F. and Noyes C.D. (2003) A conceptual model and remediation strategy for volatile organic compounds in unconsolidated sediments: a Lawrence Livermore National Laboratory case study. *Environmental & Engineering Geoscience*, **9**(1), 83–94.

Johnson P.C., Stanley C.C., Kemblowski M.W., Byers D.L. and Colthart J.D. (1990) A practical approach to the design, operation, and monitoring of in situ soil-venting systems. *Groundwater Monitoring & Remediation*, **10**(2), 159–178.

Kaplan S. and Garrick B.J. (1981) On the quantitative definition of risk. *Risk Analysis*, **1**, 11–27.

Kobus H., Barczewski B. and Koschitzsky H.P. (Eds.) (1996) *Groundwater and Subsurface Remediation: Research Strategies for In-Situ Technologies*, Springer Verlag: New York.

Kersting A.B., Efurd D.W., Finnegan D.L., Rokop D.J., Smith D.K. and Thompson J.L. (1999) Migration of plutonium in groundwater at the Nevada Test Site. *Nature*, **397**, 56–59.

MacDonald, S., (2000) Project on Environmental Epidemiology – Love Canal. *Foundation for Blood Research*, University of Southern Maine: **www.usm.maine.edu/ams/envepi/curric.htm**.

Mackay D. and Smith L. (1993) Organic contaminants. In *Regional Ground-Water Quality*, Alley W.M. (Ed.) Van Norstrand Reinhold: New York.

McCarthy J.F. and Zachara J.M. (1989) Subsurface transport of contaminants. *Environmental Science and Technology*, **23**(5), 496–502.

McCarty P.L. (1996) Biotic and abiotic transformations of chlorinated solvents in groundwater. *Symposium on Natural Attenuation of Chlorinated Organics in Groundwater*, Dallas.

Mercer, J.W., Skipp D.C. and Giffin D. (1990) *Basics of Pump-and-Treat Ground-Water Remediation Technology*, Report EPA-600/8-90/003 U.S. Environmental Protection Agency.

National Research Council (1983) *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press: Washington.

National Research Council (1994) *Alternatives for Groundwater Cleanup*, National Academy Press: Washington.

National Research Council (1999) *Groundwater Soil & Cleanup: Improving Management of Persistent Contaminants*, National Academy Press: Washington.

National Research Council (2000) *Natural Attenuation for Groundwater Remediation*, National Academy Press: Washington.

Newmark, R.L. (Ed.) (1994) *Dynamic Underground Stripping Project: LLNL Gasoline Spill Demonstration Report*, UCRL-ID-116964, Lawrence Livermore National Laboratory, Livermore. A Web page listing the contents of the complete four-volume report, with links to all pdf files: **http://geosciences.llnl.gov/envtech/dynstrip/dus_summary.html**.

Office of Technology Assessment (1984) *Protecting the Nation's Groundwater from Contamination*, Vol. I, NTIS: Order #PB85-154201.

Pankow J.F. and Cherry J.A. (1996) *Dense Chlorinated Solvents and other DNAPLs in Groundwater: History, Behavior, and Remediation*, Waterloo Press: Portland.

Parkhurst, D.L., and Appelo C.A.J. (1999) *User's Guide to PHREEQC (Version2) – A Computer Program for Speciation, Batch-Reaction, One-Dimensional transport, and Inverse Geochemical Calculations*, Report 99-4259, U.S. Geological Survey Water-Resources Investigations.

Rice D.W., Grose R.D., Michaelson J.C., Dooher B.P., MacQuenn D.H., Cullen S.J., Kastenberg W.E., Everett L.E. and Marino M.A. (1995) *California Leaking Underground Fuel Tank (LUFT) Historical Case Analyses*, UCRL-AR-122207, Lawrence Livermore National Laboratory, Livermore.

Sahimi M. (1995) *Flow and Transport in Porous Media and Fractured Rock: From Classical Methods to Modern Approaches*, John Wiley & Sons: New York.

Schwarzenbach R.P., Gschwend P.M. and Imboden D.M. (1993) *Environmental Organic Chemistry*, John Wiley & Sons: New York.

Sposito G. (1989) *The Chemistry of Soils*, Oxford University Press: New York.

Stumm W. and Morgan J.J. (1996) *Aquatic Chemistry*, *Third Edition*, John Wiley & Sons: New York.

U.S. EPA (2003b) List of Contaminants & their MCLs, **http://www.epa.gov/safewater/mcl.html#mcls**.

Vroblesky D.A. and Chapelle F.H. (1994) Temporal and spatial changes of terminal electron-accepting processes in a petroleum hydrocarbon-contaminated aquifer and the significance for contaminant biodegradation. *Water Resources Research*, **30**(5), 1561–1570.

Westall J.C. (1987) Adsorption mechanisms in aquatic surface chemistry. In *Aquatic Surface Chemistry*, Stumm W. (Ed.) John Wiley & Sons: New York.

Zhang P.-C., Krumhansl J.L. and Brady P.V. (2002) Introduction to properties, sources, and characteristics of soil radionuclides. In *Geochemistry of Soil Radionuclides*, Brady P.V. (Ed.) Soil Science Society of America Special Publication 59, Soil Science Society of America Special Publication: Madison.

Zuesse, E. (1981) The truth seeps out, *Reason Magazine*, **reason.com/8102/fe.ez.the.shtml**.

154: Stochastic Modeling of Flow and Transport in Porous and Fractured Media

SOUHEIL M EZZEDINE

University of California, Lawrence Livermore National Laboratory, Livermore, CA, US

Estimation of various flow and mass transport parameters can be seen as a problem of spatial statistics. The definition of the properties of porous and fractured media in space and time using the concept of random functions provides means for (i) studying the inherent heterogeneity, (ii) evaluating the spatiotemporal variability of the properties, and (iii) assessing the uncertainty associated with their estimated values. In this chapter, the fundamentals of stochastic subsurface hydrology are presented. Applications include mapping hydrogeological properties, flow and transport in porous and fractured media, and inverse problems.

INTRODUCTION

Stochastic analysis has two main historic roots: One is in the analysis of the so-called *random errors* in measurements of physical entities, which has gradually embraced the study of irregularities and *uncertainties* observed in what were believed to be deterministic phenomena and has led to the discipline of *mathematical statistics*; the other root is in the study of *games of chance*, which has led to the *theory of probability*. Originally, both these branches of science were confined to the realm of mathematics and did not interfere much with the deterministic concept of the physical world governed by the laws of Newton's mechanics and uniquely describable by the differential calculus. At that instance, the only problems seemed to be (i) to develop the theory in sufficient detail, (ii) to ascertain the initial conditions, and (iii) to avoid measurement errors. All discrepancies between theory and observation and uncertainties in the prediction of physical phenomena were attributed to human limitations. All this changed with the establishment of the quantum theory, which introduces the revolutionary concept that uncertainty, apart from being the result of human limitations, is also an intrinsic feature of matter itself. This concept has transformed statistics and probability from a tool dealing with a *noise* in the laws of nature into a tool for the formulation of these laws themselves.

Undeniably, natural porous and fractured formations are heterogeneous, and display spatial variability of their petrophysical properties. This variability is of irregular and complex nature. It generally defies a precise quantitative description because of insufficient information at all relevant scales. In practice, however, only sparse scale-dependent measurements are available mainly due to limited costs. Owing to lack of detailed and exhaustive information, the higher the variability the higher is the uncertainty. For this very reason, stochastic approaches have been developed in subsurface hydrology to "fill in" or bridge the gap between the overall knowledge about the "physical law" and the strong spatial variation of the petrophysical properties, which are known through sparse tainted measurements. The theory of random, or stochastic, processes provides a natural framework for evaluating aquifer uncertainties. In the stochastic formalism, uncertainty is represented by probability or by related quantities like statistical moments. Boundary conditions, initial conditions, and parameters can be treated as random functions or fields whose values are determined by probability distributions conditional to sparse measurements or information.

The definition of the properties of porous media in space and time using the concept of random functions has two major advantages: (i) It conceptually defines the properties in space at a given point, without having to define a volume over which these properties must be integrated. (ii) It provides means for studying the inherent heterogeneity and variability of these properties in space, and for evaluating the uncertainty of any method

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

of estimation of their values. The pioneering works of Schvidler (1962) and Matheron (1965) were the backbone of the work of Gedeon Dagan and Lynn Gelhar who have shaped and popularized what stochastic subsurface hydrology is today. A few monographs have since been published to cover the richness, strengths, and weakness of the stochastic methods. These monographs include a wealth of references, and we urge readers to consult them and the references therein. In this article, we will cover the *fundamentals* or in other words what became the *need-to-know* in stochastic subsurface flow and transport. For more advanced topics, we urge interested readers to consult most recent publications.

DETERMINISTIC, RANDOM, AND STOCHASTIC CONCEPTS

Of these three, the term deterministic is probably least ambiguous and involves a one-to-one relationship among events. The term random is often used as a synonym for stochastic, however, statistically speaking, there is a difference. The term *random* will be used for a completely irregular order or arrangement of elements from a given set. Thus, for instance, a random sequence or series will mean one whose terms are mutually independent and therefore uncorrelated. In the literature, this notion is often termed *pure random*. The term *stochastic* will be used for a relationship, a variable, or a process, incorporating both an element of randomness and an element of determinism. Thus a stochastic entity Y, that is, the logarithm of the hydraulic conductivity K, will be viewed as a sum $Y = \langle Y \rangle + Y'$ where $\langle Y \rangle$ and Y' are the deterministic and random components, respectively. The deterministic component is often interpreted as an *overall trend* or *signal* and the random component as *error* or *noise*. In agreement with the original meaning of the Greek word $\sigma\tau\acute{o}o\zeta$ ([stokhos] target or aim), a stochastic relationship can be viewed as one *aiming at* the relationship indicated by the deterministic component. It follows that the two concepts, deterministic and random, are only special cases of stochastic and represent its two end limits approached as one or the other of the two components tends to zero. According to this interpretation, strictly deterministic relationships cannot be proved to exist in the real world since any verification involves measurement on the one hand and exact reproducibility on the other. Whether, due to the *limited accuracy of measurement* or due to the *intrinsic impossibility of exact* or *exhaustive* depiction of any specific petrophysical parameter, there is always a noise and purely deterministic relationship.

It may be noted that our intuitive perception of deterministic laws is in fact stochastic; we expect that actual measurements will agree only approximately with a given deterministic law. As a typical example from hydrogeology, one may take the concept of the pumping test. Though classified as a deterministic model, pumping tests in heterogeneous aquifer are expected to yield integrated and approximated value of the transmissivity (or hydraulic conductivity) and the storativity coefficient of the aquifer. Those values *minimize the square of the errors* between the theoretical model (the physical law) and the measurements (drawdown vs. time) and represent the deterministic component of the intuitively implied stochastic relationship between pumping rate and drawdown.

Similarly, any other deterministic law or model can be interpreted as a stochastic model in which the random component is not explicitly taken into account. It may be too small to be detected by the measuring apparatus or, though detectable, small enough in comparison with the deterministic component to be neglected; or, for the sake of simplicity, it may be *filtered* out. An important aspect of the above definition of *stochasticity* is that it does not imply any specific origin of the random element. It does not differentiate between a case where the latter accounts merely for the analyst's *limited knowledge* of some deterministic components and one where it might be an inherent feature of the analyzed phenomenon.

Limited knowledge of information in both observed data (i.e. hydraulic head, concentration), and measured petrophysical properties (i.e. hydraulic conductivity, porosity) and their spatial distribution causes uncertain predictions. *Spatial variability* and uncertainty have led engineers and geologists to use probabilistic theories that translate the uncertainty to a *random space function* (RSF) or a *random field*, consisting of an *ensemble* of "*infinite*" number of *equally probable realizations* of parameter values, all having the same spatial statistics, particularly correlation structure. Imbedded in this approach is a *geostatistical* model.

Geostatistics is commonly used to analyze and interpolate between measurements using methods such as *kriging* (pronounced as in brid*ging*), where the uncertainties in kriged (pronounced as in brid*ged*) values are also quantified. Usually, these data are collected on different scales that may differ from the required scale of predictions. The task of quantitatively relating measurements and properties on different scales is difficult and intriguing. We briefly present in the following section the essence of geostatistics.

GEOSTATISTICS

Geostatistics has its basis in Matheron's (1965) *Theory of Regionalized Variables*. A random variable is one that has a variety of values in accordance with a particular probability distribution (Journel and Huijbregts, 1978). If the random variable is distributed in space and/or time, we say that it is a regionalized variable. These variables, because of their spatial-temporal character, have a random as well as a

structural component. At first sight, a regionalized variable seems to be a contradiction. In one sense, it is a random variable that locally does not have any relation to the nearby variables. On the other hand, there is a structural aspect in the regionalized variable that depends on the distance of separation of the variables. Both characteristics can be described, however, using a random function for which each regionalized variable is a particular realization. By incorporating the random as well as the structural aspects of a variable in a simple function, the spatial variability can be addressed on the basis of the spatial structure shown by these variables. In this sense, a regionalized variable is a variable that qualifies a phenomenon that is distributed through space and/or in time and that presents a certain correlation structure. Following are useful applications of geostatistics:

- *Characterization of large data sets*: For spatially distributed data, the resolution of measurement points is mostly rather low. By contrast, time series often consist of thousands of measurements. Therefore, condensing the information may be necessary. In this context, the *probability density function* (pdf) of the measured quantity (which is characterized by its mean and variance if the pdf is Gaussian) and some information about the average persistence (like the auto-covariance function, variograms) may be adequate.
- *Interpolation*: Given point-measurements in space, one is asked to create maps of the measured quantity. This requires interpolation. In the geostatistical framework, spatial interpolation is identical to conditioning, that is, in the vicinity of a measurement point, the expected value of the unknown is, due to the spatial correlation, weighted toward the measured value, and the variance is decreased. Taking the mean of the conditional pdf as interpolated value is referred to as kriging. In contrast to other interpolation techniques, kriging allows to evaluate and assess the uncertainty of the interpolated value indicating also where an additional measurement would be most valuable.
- *Inverse modeling*: The geostatistical approach of inversing is essentially identical to that of interpolation. The difference is that cross-covariances describing the spatial correlation between different variables (such as hydraulic head and log-conductivity) are used rather than auto-covariance functions. This is referred to as cokriging. As will be shown later, cross-covariance functions can be calculated from the auto-covariance functions and the governing partial differential equation. Like kriging, cokriging yields not only the best estimate but also the spatial distribution of the uncertainty related to that estimate.
- *Monte Carlo simulations:* (*MCSs*): MCS is somewhat the experimental numerical apparatus of statisticians. By creating equally probable conditional realizations, preferably conditional to measured data, of the hydraulic conductivity field for example, one can simulate the flow and transport for each realization. All the results are then averaged to predict the probabilistic behavior of the flow and transport in the aquifer. In this context, using multiple realizations is justified by the uncertainty in the interpolation between measurements.

Basic Statistical Treatment

As a first step, the quality of the data is proved, and the basic statistical calculations from which the statistics or measurements are obtained are performed. These are the numerical values that enable us to characterize and compare the statistical distributions. These statistics can be made on the variable itself, or some transformation of the variable (e.g. logarithmic, log). They can be of several types, as follows:

1. *Measurements of centralization*: These measurements indicate a value around which the distribution values are distributed. They are generically known as *means*, and have the following forms. The arithmetic mean or mean, m. The median, M, is the value that divides the population into two equal parts. It is the quartile of 50% (Q_2). The lower quartile, Q_1, is the value with 25% of the population below it and the remaining 75% above it. The upper quartile, Q_3, is the value with 75% of the population below it and the remaining 25% above it. The mode, Mo, is the value of the greatest absolute frequency of a distribution. Other means such as harmonic and geometric means are of great interest. The geometric mean is defined as the mean of the log-transformed measurements, and the harmonic mean as the mean of the inverse of the measurements.
2. *Measurements of dispersion*: These measurements indicate the variability of dispersion of the values of a distribution. The variance, σ^2, is the average of the squares of the deviations with respect to the mean. The standard deviation, σ, is the square root of the variance and represents the margin of the variation or the error of estimation in which the data analyzed are included. It has the same units as the variable under consideration. The range, $\Re$, of a distribution is the difference between the extremes, maximum and minimum values. The coefficient of variation, CV, is the quotient of the standard deviation and the mean. It enables us to compare distributions that have different units. There is a direct relation between the value of this parameter and the dispersion of the distribution. The interquartile range, IR, is the difference between the upper quartile and the lower quartile and it indicates the range between which 50% of the central values of the population are distributed.
3. *Distribution moments*: These can be central or with respect to the origin. The central moment of order r,

Table 1 Basic Univariate statistics

$\{x_i \mid i = 1, \ldots, N\}$ being a set of data	
Observations #	N
Histogram and cumulative histogram	$F(x_i)$, $CF(x_i)$
Minimum	$\min/\forall x_i$, $\min \leq x_i$, $i = 1, \ldots, N$
25th% Percentile	Q_i
Mode	Mo
Median or 50th % Percentile	M
Mean	$m = \frac{1}{N}\sum_{i=1}^{N} x_i$
75th % Percentile	Q_3
Maximum	$\max/\forall x_i$, $\max \geq x_i$, $i = 1, \ldots, N$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$
Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$
Coefficient of variation (%)	$CV = \frac{\sigma}{m}$
Range	$R = \max - \min$
Interquartilic range	$IR = Q_3 - Q_4$
3rd Order central moment M_3	$M_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^3$
Coefficient of skewness	$CS = \frac{M_3}{\sigma^3}$
4th Order central moment M_4	$M_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^4$
Coefficient of Kurtosis	$CK = \frac{M_4}{\sigma^4}$

M_r, is the average of the deviation with respect to the mean taken to order r.

4. *Asymmetry or skewness measurements*: These provide an idea of the asymmetry of the distribution. A distribution is *symmetric* when the frequencies corresponding to equidistant values with respect to a central value are equal. In the ideal symmetry condition, the values of the mean, median, and mode coincide. A distribution has a *bias to the right* or is *positive* if the frequencies descend more slowly on the right of the histogram. In this case, the mean is greater than the mode. A distribution has its *bias to the left* or is *negative* if the frequencies descend more slowly on the left of the histogram. In this case, the mean is less than the mode. The *coefficient of skewness*, CS, calculated in terms of the moments, is the ratio between the moment of order 3 and the standard deviation cubed.
5. *Measurements of kurtosis*: These indicate the degree of kurtosis with respect to a normal or Gaussian distribution. The *coefficient of kurtosis*, CK, calculated in terms of the moments, is the ratio between the moment of order 4 and the standard deviation taken to the fourth power. In a Gaussian or normal distribution, this coefficient CK is 3 and its curve is normal or *mesokurtic*. If CK is greater than 3, the distribution has a *leptokurtic* curve, which is sharper than the Gaussian one. If CK is less than 3 the distribution has a *platikurtic* curve, which is flatter than the Gaussian one. The analytic expressions of the above concepts are given in Table 1.

Geostatistical Description of Spatially Variable Parameters

Depicting heterogeneity of aquifers is a daunting task and could be achieved through two approaches. On the one hand, solely on the basis of the geological deposition processes, one can create images (realizations) of the subsurface heterogeneity (Gelhar, 1992; Rubin, 2003). Unfortunately, the geology-based models are limited because it is difficult to integrate the geological observations. However, statistical and geostatistical methods honor the measured data regardless of their origin: *hard data* (direct measurements of the variable to be mapped, *primary variable*) or *soft data* (indirect measurements of the variable in consideration, *secondary variable*). Honoring data is usually

referred to as *conditioning*. Conditioning on soft data usually necessitates a physical law and/or a linear or nonlinear relationship or correlation that bind the secondary variable to the primary one (i.e. Ezzedine *et al.*, 1999).

A statistical description of hydraulic properties is equivalent to interpreting the distribution as an outcome of a random process. As a consequence, an infinite number of different distributions sharing the same statistical description are possible. Introducing conditioning decreases the variability but does not change the fact that an infinite number of realizations are equally probable. Our limited knowledge of the "true" setting through only contaminated sparse measurements of hydraulic properties made the stochastic techniques a *must* for interpolation and uncertainty assessment.

For example, one will take a certain number of core samples and analyze them to get point information about the hydraulic conductivity. On the basis of these data one may estimate the hydraulic conductivity distribution. At *every* location within the aquifer, the conductivity value is estimated by *interpolation* and so its level of *confidence* or *uncertainty*. It is rather unlikely that the interpolated value is the true one. Therefore, using the interpolated conductivity field as *unique* (single) realization in a flow and transport simulator may lead to biased outcome.

Statistics of Spatial Variables

In order to describe a spatial field, such as the hydraulic conductivity within an aquifer, by geostatistics, we must conduct as many point-measurements of the quantity of interest as possible. Without loss of generality, we focus on the log-hydraulic conductivity $Y = \ln(K)$. The simplest single-variate statistical description of the data set is the *mean* (1st moment) and the *variance* (2nd moment):

$$\mathbf{E}[Y] \cong \overline{Y} = \frac{1}{n}\sum_i Y_i = \int y f_Y(y)\ \mathrm{d}y \tag{1}$$

$$\mathbf{Var}[Y] = \mathbf{E}[(Y - \mathbf{E}[Y])^2] \cong \sigma_Y^2 = \frac{1}{n-1}\sum_i (Y_i - \overline{Y})^2 \tag{2}$$

where $Y_i(x)$ is the value of Y at a location x_i, and $\mathbf{E}[]$ and $\mathbf{Var}[]$ are the expected, and the variance *operator*, respectively. If Y is *normal* (*Gaussian*), then the first and second moment are sufficient to describe Y, otherwise higher statistical moments are needed. They further enhance the statistical characterization of Y, the hydraulic conductivity is considered as a *spatial random function* (SRF). Its value may differ from one location, $\mathbf{x}_1 = (x_1, x_2, x_3)_1$, to another, $\mathbf{x}_2 = (x_1, x_2, x_3)_2$. Consequently the *similarity* between the values of Y at different locations is expressed by the *covariance* function, which is defined by

$$\mathbf{Cov}[Y(\mathbf{x}_1)Y(\mathbf{x}_2)] = \mathbf{E}[(Y(\mathbf{x}_1) - \mathbf{E}[Y(\mathbf{x}_1)])(Y(\mathbf{x}_2) - \mathbf{E}[Y(\mathbf{x}_2)])] \tag{3}$$

Cov[] is the covariance operator. It is worth mentioning that the covariance does not always exist, given the type of variability of the variable. When dealing with a single data set, **Cov** is called *auto-covariance*. Otherwise, when dealing with multiple data sets, that is, the hydraulic conductivity and head, the similarity is termed as *cross-covariance*. Furthermore, when a SRF is said to be *stationary*, the **Cov** operator depends solely on the separation distance $\mathbf{h} = \mathbf{x}_1 - \mathbf{x}_2$, $\mathbf{Cov}[Y(\mathbf{x}_1)Y(\mathbf{x}_2)] = \mathbf{Cov}[Y(\mathbf{h})]$, regardless their locations, otherwise the SRF is location-dependent and called *nonstationary*. Stationarity has different levels, that is, *weak* and *strong* and we will discuss them later. It is worth noting that when h vanishes, the covariance reduces to σ_Y^2. Similarly, the *correlation function* or *correlogram* ρ of a SRF is also defined. The correlation function $\rho(\mathbf{h}) = \mathbf{Cov}[Y(\mathbf{h})]/\sigma_Y^2$ takes values in $[-1, +1]$. When $\rho = 0$ there is no correlation, the data sets are independent. Otherwise when $\rho = \pm 1$, there is a total (perfect) positive/negative (direct/indirect) correlation. The definition of the covariance assumes that the mean is known, which in reality may not be the case, and that a covariance exists. To overcome this limitation, Matheron suggested the use of yet another similarity measure, the variogram γ defined by

$$\gamma_{YY}(\mathbf{h}) = \frac{1}{2}\mathbf{E}[(Y(\mathbf{x}_i) - Y(\mathbf{x}_j)]^2 = \frac{1}{2n(\mathbf{h})}\sum_{\substack{i,j \\ h=x_i-x_j}}$$

$$(Y_i - Y_j)^2 = \sigma_Y^2 - \mathbf{Cov}[(Y(\mathbf{x}_i) - Y(\mathbf{x}_j)] \tag{4}$$

The word *semivariogram* stems from $1/2$ term in the above notation. The names variogram and semivariogram are used interchangeably. The variogram has two attractive properties: (i) γ is linked to the **Cov** through the variance σ_Y^2, (ii) it is defined even when the variance itself is not finite, that is, does not exist. A stationary SRF with a, not necessarily known, constant mean and a variogram is referred to as *intrinsic hypothesis*. Note that these statistical properties can also be defined in space and time, called *spatiotemporal* statistical tools and we urge the reader to consult Christakos (1992) for more details. These tools have become the main engine behind geology-related estimation problems since the 1960s. Matheron coined the word *geostatistics* since it relates the field of geology to the field of statistics as is in geophysics for example. In the next section we will tackle the variogram analysis, that is, the main step behind any geostatistical application.

Typical Covariances and Variograms

For the sake of clarity, we introduce only the typical covariances and variograms under stationarity hypothesis. The common models are (i) exponential, Gaussian, spherical,

and power-law models. The *exponential* model is given by

$$C_Y(\mathbf{h}) = \sigma_Y^2\, e^{-h}, \quad \mathbf{h} = \sqrt{\sum_{\mathbf{i}} \left(\frac{\mathbf{h_i}}{\boldsymbol{\lambda_i}}\right)^2},$$

$$\gamma(\mathbf{h}) = \sigma_Y^2(1 - e^{-h}) \tag{5}$$

where λ_i are correlation lengths (range, scaling parameters which can be different in different directions, in case of anisotropy of the variability, Figure 1). Because $\mathbf{h}$ is positive, the variogram slowly reaches a finite limit for large values of $\mathbf{h}$. This limit is called *sill* and, in this particular case, it is equal to the variance. Because the exponential model has very attractive mathematical properties, various analytical solutions for stationary Y random fields with an exponential covariance have been developed. One should notice that the derivative at the origin does not vanish. This causes small-scale variability. Such a behavior does not occur in the *Gaussian* model given by

$$C_Y(\mathbf{h}) = \sigma_Y^2\, e^{-h^2}, \quad \mathbf{h} = \sqrt{\sum_{i} \left(\frac{h_i}{\lambda_i}\right)^2},$$

$$\gamma(\mathbf{h}) = \sigma_Y^2(1 - e^{-h^2}) \tag{6}$$

Contrary to the exponential model, the Gaussian model converges faster to the sill (Figure 2). The Gaussian model is much smoother and gradual interpolator than the exponential model. This is different in the *spherical* model:

$$C_Y(\mathbf{h}) = \begin{cases} \sigma_Y^2\left[1 - \frac{3}{2}h + \frac{1}{2}h^3\right], & h \le 1 \\ 0, \text{otherwise} \end{cases}$$

$$\mathbf{h} = \sqrt{\sum_{i} \left(\frac{h_i}{\lambda_i}\right)^2}$$

$$\gamma(\mathbf{h}) = \begin{cases} \sigma_Y^2\left[\frac{3}{2}h - \frac{1}{2}h^3\right], h \le 1 \\ \sigma_Y^2, \text{otherwise} \end{cases} \tag{7}$$

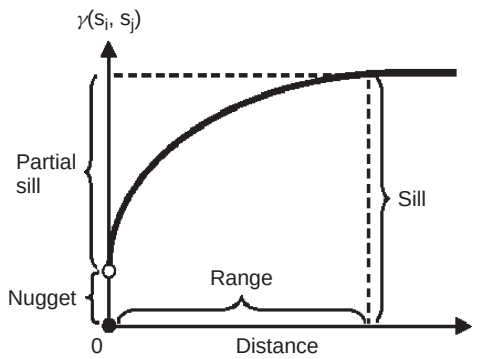


Figure 1 Anatomy of a variogram

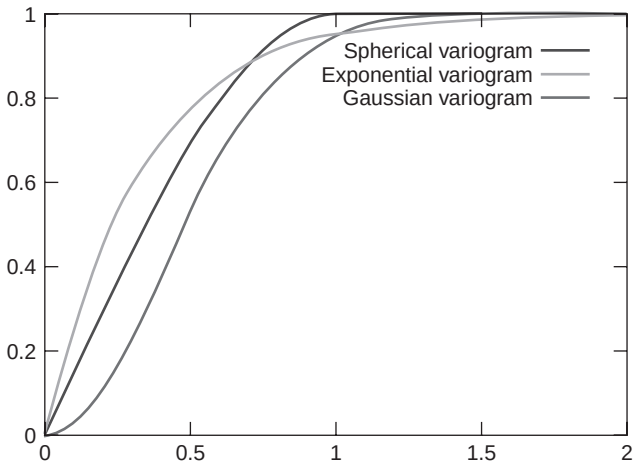


Figure 2 Different variograms. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The behavior of the spherical model near the origin is nearly linear. Unlike the exponential model, there is no correlation when h is greater than 1. The *power-law*, also called *self-similar* or *fractal*, model in the isotropic case is given by:

$$\gamma_Y(\mathbf{h}) = C\ \mathbf{h}^{2\mathrm{H}}, \quad \text{where} \quad 0 < \mathrm{H} < 1 \tag{8}$$

where C is a constant; H is called the *Hurst* number/exponent. For an exponent $\mathrm{H} = 1/2$, the power variogram reduces to the linear variogram. Y does not have a finite variance because the variability is unbounded, and a covariance function cannot be defined, only a variogram.

Theoretically, variograms vanish at the origin, however, very often, variograms exhibit a finite jump at the origin. This apparent jump at the origin is called *nugget effect* (originated from the mining industry). The nugget effect is taken into account by adding a nugget contribution, γ^0, to γ^t the variogram fitted to the data as if γ^0 were the origin

$$\gamma_Y(h) = \gamma^0(1 - \delta(\mathbf{h})) + \gamma^t(\mathbf{h}) \tag{9}$$

where $\delta(h)$ is the *Kronecker* or *delta function*. A nugget variogram expresses purely random phenomena without any spatial structure. The contribution of the nugget effect may be interpreted as a measurement error or as variability at scales smaller than the measurement scales. Any combination of the mentioned models is possible, and it is called a *nested model* or *nested structure*.

Variograms are direction-dependent; therefore it is useful to evaluate the variogram at different directions (i.e. North, East, West, South and any combinations). Generally, variograms do not show *anisotropy*, different behavior for different direction. If they do, then: (i) it may be a sign that

the assumption of stationarity does not hold, or (ii) the stationarity hypothesis is valid, and thus this anisotropy can be eliminated by a suitable linear transformation of the coordinate system. Otherwise, if the variograms do not show significant directional changes, the variogram is isotropic and *omnidirectional*.

The directional *integral scale* I_i of a covariance function is defined by

$$I_i = \frac{\int_0^\infty \mathbf{Cov}[h_i]\ \mathrm{d}h_i}{\mathbf{Cov}[0]} = \frac{1}{\sigma_Y^2}\int_0^\infty \mathbf{Cov}[h_i]\ \mathrm{d}h_i \qquad (10)$$

where $I = \lambda$ for the exponential model, $\lambda\sqrt{\pi}/2$ for the Gaussian and $3\lambda/8$ for the spherical model. When $I_1 = I_2 = I_3$ the field is isotropic, when $I_1 = I_2 \neq I_3$ the field is axisymmetric, otherwise it is anisotropic.

For sloped strata, a transformation of coordinates into the principal directions of the heterogeneities or a transformation into the stratigraphic coordinates may be necessary to assure that the measurement are correlated with respect to the geological depositions/layering and not to the geographical depth.

Calculation of Experimental Variograms

To estimate a variogram, we use the measurement points Y_i and assume *ergodicity* (i.e. space averages can be used to estimate the averages in the entire set of realizations). First, we define a certain number of distance classes between measurement points. Then, taking into account all possible pairs or points for each class of distances, we calculate: (i) the number of pairs present in that class, (ii) the average distance in the class, and (iii) the average square increments $1/2(Y_i - Y_j)^2$. This is referred to as *experimental variogram*. Generally, the number of pairs is not evenly distributed between the classes of distances due to the scattering of the measurements. There are more pairs at shorter than longer distances. The variogram becomes more uncertain as the separation distance h increases. Once the experimental variogram is calculated, one may fit one of the theoretical models to the experimental variogram. Curve fitting may be biased by the choice of classes and the averaging step. More details on the determination of a fitted variogram are given in Journel and Huijbregts (1978), Isaaks and Srivastava (1989), Kitanidis (1997) to name a few.

Kriging

Consider a set of n measurements of Y with Y_i measured at the location x_i. The variogram $\tilde{\gamma}_{YY}(h)$ is known. We want to estimate the value Y_0 at a location $\mathbf{x}_0$. A linear estimator is a linear combination of the measured data $Y_0^* = \sum_i^n \lambda_{0,i} Y_i$ in which λ_i is the *weighting* factor for measurement Y_i and the * indicates that Y_0 is estimated. If the true unknown value at x_0 is Y_0, the estimation is said to be an *optimal* estimate in the *minimum variance sense* if $\mathbf{E}[(Y_0^* - Y_0)^2]$ is minimal. We also require an *unbiased* estimator; in other words, the expected value of the estimation error is zero: $\mathbf{E}[(Y_0^* - Y_0)] = 0$. The unbiasedness condition yields a constraint on the choice of the weights, that is $1 = \sum_i^n \lambda_{0,i}$, which assumes a constant mean throughout out the domain $\mathbf{E}[Y(\mathbf{x})] = m$. The mean m is not necessarily known. The minimum variance conditions leads to

$$\begin{aligned} &\text{Min } \mathbf{E}[(Y_0^* - Y_0)^2] = \sigma_Y^{2*}(x_0) = -\sum_i \sum_j \lambda_i \lambda_j \\ &\gamma_{YY}(x_i - x_j) + 2\sum_i \lambda_i \gamma_{YY}(x_i - x_0); \\ &\text{subject to} \quad 1 = \sum_i^n \lambda_{0,i} \end{aligned} \qquad (11)$$

To minimization problem, an additional unknown referred to as *Lagrangian multiplier*, μ, (Matheron, 1965) is introduced to enforce the constraint. Taking the derivatives with respect to λ_i and μ, and finding the minimum by equating the derivatives to zero, leads to the following system of n + 1 linear equations:

$$\begin{aligned} &-\sum_j \lambda_j \gamma_{YY}(x_i - x_j) + \mu = \gamma_{YY}(x_i - x_0) \\ &\qquad \forall i = 1, n \quad \text{and} \quad 1 = \sum_i \lambda_i \end{aligned} \qquad (12)$$

This is known as *kriging* system named by Matheron after the South African mining engineer Krige. Because the variogram is a symmetric function, the *matrix* is *symmetric* also, however, it is not *positive definite*. One attractive property of Kriging is that the inverse of the matrix has to be done only once and stored since it does not depend on the location x_0; however, the right-hand vector does dependent on x_0 and has to be updated for each location. The kriging estimator is also referred to as the *Best Linear Unbiased Estimator* (BLUE). Expression for the estimation variance is as follows (Matheron, 1965, see also de Marsily, 1986):

$$\sigma_Y^{2*}(x_0) = \sum_j \lambda_j\ \gamma_{YY}(x_j - x_0) - \mu \qquad (13)$$

The kriging estimator with a constant unknown mean as described here is called *ordinary kriging*. When the mean is known it is called *simple kriging*. One may argue that the contributions of distance measurements are weak and therefore only the immediate measurements within a certain *neighborhood* of x_0, should be considered. In this case, the matrix to be inverted is neighborhood-dependent and has to

be assembled for each point-estimation. Readers are urged to look up Deutsch and Journel (1997), Kitanidis (1997), and Chiles and Delfiner (1999).

Cokriging

Hydraulic conductivity measurements are more intensive to obtain than hydraulic head and, therefore, expensive. Generally, K is estimated using pumping tests; and thus it is not a local measurement but rather an integrated measurement. Head measurements, however, are affordable and sometimes abundant in number. It is imperative then to benefit from the cross-correlation between both SRFs to improve point-estimation of either variable through a more generalized kriging process called *cokriging*. We illustrate it on an exhaustive set of N_Y and N_H measurements of hydraulic conductivity and head, respectively. Assuming that the both $E[Y]$ and $E[H]$ are known the simple Cokriging system for h reads (Rubin, 2003):

$$h(x_0) = E[h(x_0)] + \sum_{i=1}^{N_Y} \lambda_i (Y_i - E[Y]) + \sum_{i=N_Y+1}^{N_Y+N_H} \lambda_i (h_i - E[h]) \quad (14)$$

$$C_{YH}(x_l, x_0) = \sum_{i=1}^{N_Y} \lambda_i C_{YY}(x_l, x_i) + \sum_{i=N_Y+1}^{N_Y+N_H} \lambda_i C_{YH}(x_l, x_i); \quad \forall\, l = 1, N_Y \quad (15)$$

$$C_{HH}(x_0, x_l) = \sum_{i=1}^{N_Y} \lambda_i C_{YH}(x_i, x_l) + \sum_{i=N_Y+1}^{N_Y+N_H} \lambda_i C_{HH}(x_i, x_l); \quad \forall\, l = N_Y + 1, N_Y + N_H \quad (16)$$

and the simple Cokriging system for Y reads

$$Y(x_0) = E[Y] + \sum_{i=1}^{N_Y} \lambda_i (Y_i - E[Y]) + \sum_{i=N_Y+1}^{N_Y+N_H} \lambda_i (h_i - E[h]) \quad (17)$$

$$C_{YH}(x_0, x_l) = \sum_{i=1}^{N_Y} \lambda_i C_{YY}(x_i, x_l) + \sum_{i=N_Y+1}^{N_Y+N_H} \lambda_i C_{YH}(x_i, x_l); \quad \forall\, l = 1, N_Y \quad (18)$$

$$C_{HH}(x_0, x_l) = \sum_{i=1}^{N_Y} \lambda_i C_{YH}(x_i, x_l) + \sum_{i=N_Y+1}^{N_Y+N_H} \lambda_i C_{HH}(x_i, x_l); \quad \forall\, l = N_Y + 1, N_Y + N_H \quad (19)$$

where $E[h(x_0)]$ is the unconditional head mean; $E[Y]$ is the conditional mean of Y; C_{YH} is the $Y-H$ cross-correlation; C_{YY} and C_{HH} are the auto-correlations of Y and H, respectively. When C_{HH} is not defined (unbounded variability) γ_{HH} is used instead and an additional constraint on the unboundedness is added.

Closed form of C_{YH} and γ_{HH} for stationary Y field assuming an exponential covariance function C_{YY}, a uniform steady flow in an infinite domain have been obtained by Dagan (1989) using 2nd order perturbation of the flow field for 2D and 3D cases. The reader is referred to Dagan (1989) and Rubin (2003) for the analytical expressions.

The cokriging process can be generalized to N_Z different Z type of measurements. For illustrative applications of cokriging please (*see* Kitanidis, 1997) and the section on "Inverse Problems as Statistics". It is important to note that the back transform form Y to T is just $T = \exp(Y)$ and that no additional terms are justified, as some authors choose to correct for the biased mean and add a term related to the variance of Y (de Marsily, 1986).

Unconditional and Conditional Simulations

Matheron proposed and implemented the unconditional simulation called *Turning Bands method*. From the 1980s onward, many applications appeared, almost all of which were dedicated to geological mining. In the 1990s, many theoretical aspects have been developed, including new more efficient algorithms such as applications relevant to multiple fields, analysis of basins, treatment of images, simulation of porous and fractured media, simulations of geological lithofacies, and so on. This development has been aided by the widespread use of computers, making it possible to deal with the numerous calculations required. The values of the models obtained using geostatistical simulation agree with the experimental information and reproduce the observed variability. The fact that the variograms of the simulated values and the real values coincide implies that both sets of values have the same spatial and/or temporal variability. We will introduce two different ways on generation random fields. The Turning Bands method, which is an unconditional method, is presented. The Sequential Gauss Simulations (SGS) is then introduced as an example of conditional method. A more sophisticated Bayesian random field will be presented in the Section "Numerical methods – Monte Carlo solution".

Unconditional Simulations: Turning Bands Method

The Turning Bands method involves the simulation of isotropic random fields in two- or higher-dimensional space by using a sequence of one-dimensional processes along lines crossing the space. The algorithm can be described as follows (Figure 3):

1. Choose an arbitrary origin within or near the domain of the field to be generated.

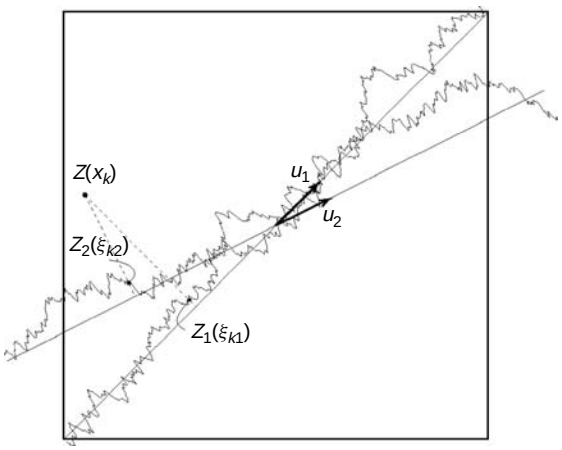


Figure 3 Illustration of the Turning Bands method

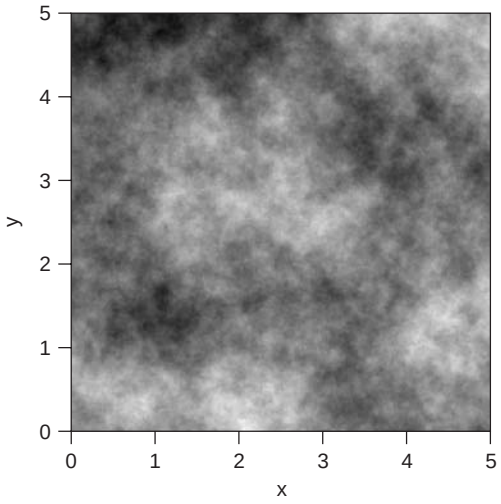


Figure 4 Example of 2D generated random field

2. Select a line i crossing the domain having a direction given by the unit vector $\mathbf{u}_i$, which may be chosen either randomly or from some fixed set.
3. Generate a realization of a one-dimensional process, $Z_i(l_i)$, along the line i having zero mean and covariance function $B_1(d_i)$ where l_i and d_i are measured along line i.
4. Orthogonally project each field point x_k onto the line i to define the coordinate $l_{ki} = x_k u_i$ of the one-dimensional process value $Z_i(l_{ki})$.
5. Add the component $Z_i(l_{ki})$ to the field value $Z(x_k)$, for each x_k.
6. Return to step (2) and generate a new one-dimensional process along a subsequent line until L lines have been produced.
7. Normalize the field $Z(x_k)$ by dividing through by the factor $\sqrt{L}$.

Essentially, the generating equation for the zero-mean discrete process $Z(x)$ is given by

$$Z(x) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L} Z_i(x_i \cdot u_i) \tag{20}$$

which can be an exceptionally fast algorithm, particularly as the number of dimensions of the process increases. It depends on the knowledge of the one-dimensional covariance function, $B_1(d)$. Once this is known, the line processes can be produced using some efficient $1-D$ algorithm such as autoregressive, moving average, or FFT techniques. The covariance function B_1 is chosen such that the multidimensional covariance structure B_n in R^n is reflected in each realization or over the ensemble. For two-dimensional isotropic processes Cressie (1993) gives the following relationship between B_2 and B_1 for $r = |d|$,

$$B_2(p) = \frac{2}{\pi} \int_0^r \frac{B_1(p)}{\sqrt{r^2 - p^2}} \, \mathrm{d}p \tag{21}$$

which is an integral equation to be solved for B_1. In three dimensions, the relationship between the isotropic B_3 and B_1 is particularly simple,

$$B_1(p) = \frac{\mathrm{d}}{\mathrm{d}p}(pB_3(p)) \tag{22}$$

Cressie (1993) supplied explicit solutions for either the equivalent one-dimensional covariance function or the equivalent one-dimensional spectral density function for a variety of common multidimensional covariance structures. In particular, for the exponential type covariance function,

$$B_2(p_1, p_2) = \sigma^2 \exp\left\{-\frac{2}{I}\sqrt{{p_1}^2 + {p_2}^2}\right\} \tag{23}$$

The line processes could be constructed using a $1-D$ FFT algorithm. Line lengths were chosen to be twice that of the field diagonal to avoid the symmetric covariance problem inherent with the FFT method. To reduce errors arising due to overly coarse discretization of the lines, the ratio between the incremental distance along the lines, Δu, and the minimum incremental distance in the field along any coordinate, Δx, is constrained by $\Delta u / \Delta x = 1/2$.

It is worth mentioning here that the unconditional random field can be made conditional with a simple double kriging. Otherwise conditional field can obtained using conditional sequential methods (Cressie, 1993). Figure 4 depicts an example of a 2D random field.

Conditional Simulations: Sequential Gaussian Simulations The procedure required producing a random field where, on one hand, at some specific locations some of the measurements Z are known and must be honored and, on the other hand, at the other locations, where measurements are lacking, must be estimated based on the inferred

correlation structure of the neighboring measurements. This can be accomplished through the kriging estimation. The interpolation by kriging leads to rather smooth distributions of the spatial variable, that is, the interpolated values themselves, although based on a certain variogram, do not satisfy the underlying variogram. This is due to the fact that at each point of interpolation the most likely value is chosen, neglecting the estimation variance. For many applications such as drawing a map of the spatial variable, the smooth estimate by kriging is adequate. However, using the distribution in numerical simulations of flow and transport processes one may undermine the heterogeneities effects. In contrast to the kriging interpolation, there are an unlimited number of possible realizations satisfying the underlying spatial variability variogram. These multiple realizations can then be used for MCSs. Several methods have been developed for the generation of geostatistical realizations (GSLIB library) (Deutsch and Journel, 1997). Here we will restrict our description to the sequential Gaussian simulation (SGSIM, or SISIM its analogue for indicator data). The basic principle of SGSIM is to perform multiple kriging, but rather than choosing the conditional mean at the point of interpolation, a random value is taken from the Gaussian distribution described by the estimation mean and variance. The algorithm consists of the following steps:

1. Choose a random sequence in which all points of interpolation will be considered.
2. For point i, calculate the kriging estimate and its variance using all measured values and the i already generated values for conditioning.
3. Choose a random value from the distribution described by the Gaussian distribution with the conditional mean and estimation variance calculated from kriging.
4. Go to step 2 unless all points have been considered.

The algorithm may be made more efficient if conditioning is restricted to points within a certain neighborhood (otherwise the computational effort for the kriging step increases linearly). An example of Sequential Indicator Simulation (SISIM) is given in Figure 5. As expected both realizations are similar around the boreholes.

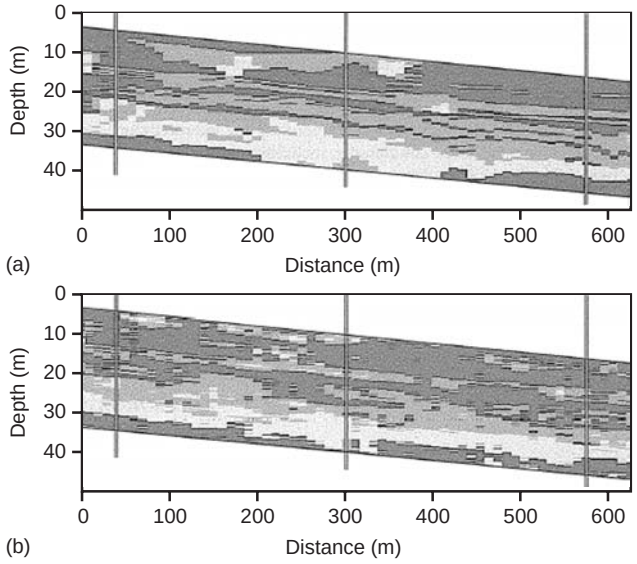


Figure 5 Two random realizations of lithologies using SISIM. The black, gray, and white colors represent shales, shally sandstone, and sandstone, respectively (Reproduced by permission of John Wiley & Sons, Inc. Modified from Chiles and Delfiner (1999))

STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS AND THEIR SOLUTIONS

Stochastic modeling deals with solving stochastic partial differential equations (SPDEs). To illustrate the concept of SPDE, we limit ourselves to the flow equation. The flow equation in porous media is based on the continuity of the mass balance concept. It is a partial differential equation (PDE) and, under transient conditions, is given by:

$$S_S \frac{\partial h}{\partial t} - \nabla.(K\ \nabla h) = 0 \tag{24}$$

If the hydraulic conductivity of the porous media K and the storativity S_S are assumed random functions, then the flow equation is called a *SPDE*. This means that the solution of the equation itself, the head h, is also a random function. Solving the flow equation implies determining the pdf of h, in particular, its first moments from the prescribed pdf of K and S_S. We introduce in the next section three ways for solving SPDEs, that is, the spectral, perturbation, and the Monte Carlo methods.

Spectral Methods

This method is applicable to the 2nd order stationary stochastic processes for both inputs and outputs. If *Y(x)* is a 2nd order stationary process, the spectrum (or spectral density) of Y is the Fourier transform and the inverse Fourier transform of its covariance function are given by:

$$\varphi(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{e}^{-i\omega s}\mathbf{Cov}[Y(x+s), Y(x)]\,\mathrm{d}s;$$
$$\mathbf{Cov}[Y(x+s), Y(x)] = C(s)$$
$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{e}^{iks}\varphi(\omega)\,\mathrm{d}\omega \tag{25}$$

The representation theorem (*Wiener-Kintchine*) states that if the 2nd order stochastic process *Y(x)* is of zero mean $\mathrm{E}[Y] = 0$ and of covariance *C(s)*, then there is an associated complex process Z (and Z^* its conjugate) that satisfies the

Fourier–Stieltjes integral:

$$Y(x) = \int_{-\infty}^{+\infty} e^{i\omega s}\, dZ(\omega)$$

$$E[\, dZ(\omega_1)\, dZ^*(\omega_2)] = 0 \quad \text{if} \quad \omega_1 \neq \omega_2 \quad \text{however}$$

$$E[\, dZ(\omega_1)\, dZ^*(\omega_2)] = \varphi(\omega_1) \quad \text{if} \quad \omega_1 = \omega_2 \qquad (26)$$

Gelhar (1992) gives a one-dimensional steady state flow in an infinite domain, as example, and it is reproduced here for completeness. The steady state flow equation reads:

$$\frac{d}{dx}\left[K(x)\ \frac{dh}{dx}\right] = 0 \qquad (27)$$

We assume that *K(x)* is a 2nd order stationary stochastic process, h is the head. Assuming a constant flow rate, q, throughout the domain, the integration of the flow equation once yields to:

$$\frac{dh}{dx} = -\frac{q}{K(x)} = -r(x)\ q \qquad (28)$$

because K is a nonzero finite physical property, one can define the hydraulic resistivity $r(x) = K(x)^{-1}$. The goal is then to determine the first two statistical moments of the SRF *h(x)* and *r(x)*. Each SRF can be decomposed in two terms: an expected value and a fluctuation around the expected value, $E[h]$ and h', and $E[r]$ and r' such that $E[h] = h - h';\ E[h'] = 0$ and $E[r] = r - r';\ E[r'] = 0$. Substituting these quantities in the above equation leads to:

$$\frac{dE[h]}{dx} = -E[r]\ q;\ \text{thus}$$

$$E[h] = -q\,E[r]\ x + \text{constant} \qquad (29)$$

$$\frac{dh'}{dx} = -q\ r' \qquad (30)$$

Assuming that h and r are 2nd order stationary processes and using the Wiener–Kintchine theorem, we introduce the following two complex stochastic processes:

$$h'(x) = \int_{-\infty}^{+\infty} e^{i\omega x}\, dZ_h(\omega);$$

$$r'(x) = \int_{-\infty}^{+\infty} e^{i\omega x}\ dZ_r(\omega) \qquad (31)$$

Substituting them in the perturbed equation, spectral density of h is given by

$$\varphi_h(\omega) = E[\, dZ_h(\omega)\ dZ_h^*(\omega)] = \left(\frac{q}{\omega}\right)^2 E[\, dZ_r(\omega)\, dZ_r^*(\omega)]$$

$$= \left(\frac{q}{\omega}\right)^2 \varphi_r(\omega) \qquad (32)$$

which gives the spectrum of h given a spectrum of r, and hence both moments are determined.

Perturbation Methods

Perturbation methods aim at rewriting a difficult mathematical problem into an infinite series of easy ones (terms). It is most useful when the first few terms of the series are pertinent to the solution of the original problem, the remaining of the series is assumed to be small. To illustrate this technique, let us apply it to the previous one-dimensional problem. First we assume that K is 2nd order stationary SRF with $E[K]$ its expected value and k' its fluctuation around the mean. Similarly, we assume that h is 2nd order stationary SRF, and $E[h]$ and h' are its mean and fluctuation, respectively. We seek a 1st order (ε^1) solution to the flow equation, thus we assume that K and h are expanded in term of small perturbations: $K = E[K] + \varepsilon k'$ and $h = E[h] + \varepsilon h'$. Substituting those into the flow equation and neglecting term of higher order (ε^2) yield to (de Marsily, 1986):

$$E[K]\frac{d^2E[h]}{dx^2} + \varepsilon\left(E[K]\frac{d^2h'}{dx^2} + \frac{dk'}{dx}\frac{dE[h]}{dx} + k'\frac{d^2E[h]}{dx^2}\right) = 0 \qquad (33)$$

If this holds for any small ε, each term of these two must vanish. Therefore,

$$E[K]\frac{d^2E[h]}{dx^2} = 0; \quad \text{thus} \quad E[h] = \frac{q}{E[K]}x + \text{constant} \qquad (34)$$

Substituting E[h] into the second term leads to

$$\frac{dh'}{dx} = \frac{q}{E[K]^2}k' + \text{constant}; \quad \text{and} \quad \mathbf{Cov}\left[\frac{dh'}{dx}\right]$$

$$= -\frac{d^2}{dx^2}(\mathbf{Cov}[h]) = \left(\frac{q}{E[K]^2}\right)^2 \mathbf{Cov}[K] \qquad (35)$$

After integrating twice one can determine the Cov function for a stationary SRF:

$$\mathbf{Cov}[h(x+s), h(x)] = \left(\frac{q}{E[K]^2}\right)^2 \int_{-\infty}^{s}\int_{-\infty}^{y} \times \mathbf{Cov}[K(x+w), K(x)]\, dw\, dy \qquad (36)$$

Thus, the first two moments of h are defined given the moment of K. It has been shown that the error involved in the method of perturbation compared to the spectral method is less than 10% if $\sigma_Y^2 \leq 1$.

All analytical and numerical solutions of the SPDEs involve the use of either technique (Gelhar (1992), Dagan (1989), Rubin (2003)).

Numerical Methods – Monte Carlo Solution

Given the petrophysical properties of an aquifer, one can, using geostatistical methods, create multiple conditional simulations/realizations of the spatial variability of those petrophysical parameters. These simulations are equally probable and the difference between them is a measurement of the estimation uncertainty. Contrary to the perturbation and spectral methods, the MCS necessitates a prior knowledge of the probability distribution and the covariance function of the parameter, that is, K. For each realization/simulation "i" of the field K_i, the flow equation is solved numerically and for instance, the solution h_i is obtained. It is then possible to statistically analyze the ensemble of computed solutions: expected value, variance, pdf, and so on, for each location x. Stationarity assumption is alleviated since statistics can be evaluated at every point of the field. Unfortunately, MCSs are time consuming for two main reasons. On the one hand, to capture heterogeneity, MCSs require solving numerically the flow and/or transport equation(s) in many detailed realizations of $K(x)$. This can be computationally expensive, especially under transient conditions. Each Monte Carlo grid must capture the detailed heterogeneity of parameter fields. On the other hand, to get meaningful statistics, a large number of realizations/simulations is necessary. It is worth mentioning that the experimental variance is proportional to the inverse of the square root of the number of realizations, and therefore the rate of convergence is slow.

An Application of SPDE to Effective Hydraulic Conductivity

Effective properties are of great interest for hydrogeological applications and especially for numerical simulations. For example, regional-scale models of groundwater flow and transport often discretize the domain into grid blocks larger than typical integral scales of field data (see Section "Typical covariances and variograms"). This process is purely numerical and it is intended to alleviate the complexity of problem and thus the number of unknowns. For heterogeneous formations, the difference between the sale (size) of heterogeneity and the discretization block size is often handled using effective (upscaled) hydrogeological parameters. For example, in the case of flow in large 3D heterogeneous media domain, the flux is given by Darcy's law: $-K\nabla H = \mathbf{q}$ and the effective hydraulic conductivity is defined by: $\langle \mathbf{q} \rangle = -K_{\text{eff}} \langle \nabla H \rangle$, where the brackets denote the expected values, boldface denotes vectors. Second-order analytical solutions were developed by Dagan (see Dagan (1989)) using perturbation methods for axisymmetric anisotropy and by Gelhar and Axness (1983) (see Gelhar (1992)) using spectral methods for fully anisotropic media. The effective tensor of hydraulic conductivity is given by

$$K_{\text{eff},\,ij} = K_G \left[\left(1 + \frac{\sigma_Y^2}{2} \right) \delta_{ij} - g_{ij} \right],$$
$$\text{and} \quad g_{ij} = \frac{1}{\sigma_Y^2} \int \frac{\omega_i \omega_j}{\omega^2} \varphi_Y(\omega) \; \mathrm{d}^3\omega \qquad (37)$$

where δ_{ij} is the Kronecker tensor, K_G is the geometric mean of the hydraulic conductivity, ω is the vector wave (Fourier space), and φ_Y is the spectral form of the covariance of the log-hydraulic conductivity, Y. More results are given in Dagan (1989) and Rubin (2003).

SOLUTE TRANSPORT IN STATIONARY HETEROGENEOUS MEDIA

A good understanding of the physical processes governing the transport of contaminants in the subsurface is crucial to decision makers and risk assessors. In the simplest case of nonreactive contaminants, the transport processes encompasses advection by regional flow and mixing or dispersion that is evident as the spreading of contaminant front in the subsurface. Early studies of the dispersion of the contaminants showed dispersion rates in the field that were much larger than the laboratory measured ones. The disparity between the laboratory and the field scale dispersivity is attributed to the velocity variation induced by heterogeneity in the hydraulic properties of the subsurface. This phenomenon is known as *macrodispersion*. The basis for calculating flow velocities for fluids in a heterogeneous porous medium is Darcy's law, which relates the specific discharge q with hydraulic conductivity K and the hydraulic gradient (∇H). Owing to the strong spatial variation of K and ∇H, which cannot be known in detail, stochastic approaches have been developed. Stochastic transport theories attempt to develop relationships between macrodispersivities and statistical parameters characterizing the hydraulic conductivity spatial variability. To develop such relationships, we need to quantify the statistical structure of the velocity variations resulting from spatial variation of the hydraulic conductivity. The resulting statistical structure of the velocity variations is then used for quantifying macrodispersivities.

Macrodispersion Coefficient

Characterization of the macrodispersion coefficient in terms of the specific discharge spectrum can be approached in two different frameworks. On the one hand, the Eulerian

approach: the solute transport equation serves as the basis for a perturbation approximation to the equation for the ensemble-average concentration field, Gelhar (1992). On the other hand, the Lagrangian approach that relies on the relationship between the dispersion of a passive tracer and the mean square displacement of a particle moving in a random velocity field, originally founded by Taylor (1921) and applied to subsurface hydrology by Matheron and de Marsily (1980) and Dagan (1989). Despite the differences, the results obtained by Gelhar (1992) are identical to those obtained by Dagan (1989). For brevity, we only summarize the main results; more details can be found in the monographs of the mentioned authors. Here we limit ourselves to the Lagrangian approach. Following Dagan (1989), let us assume that at $t = 0$ a body of tracer concentration C_0 is introduced in a fluid and spread subsequently due to both molecular diffusion and convection by the fluid. Taylor suggested that the solute body is regarded as a collection of invisible particles; the concentration is regarded as the relative number per unit volume. For a single particle the motion is described by the displacement:

$$X_i(t) = \int_0^t v_i(t')\ \mathrm{d}t' \tag{38}$$

where v is the Lagrangian velocity which is related to the Eulerian one through: $v(t) = u(X(t))$. The velocity is a random variable and is the sum of the two uncorrelated velocities: (i) a velocity due to the molecular diffusion and (ii) a fluid velocity (advective). The velocity vector is assumed to be a 2nd order stationary random process in time with the following statistics: $\mathrm{E}[v_i(t)]$, and $\mathrm{E}[(v_i(t) - \mathrm{E}[v_i(t)])(v_j(\tau) - \mathrm{E}[v_j(\tau)])] = \mathrm{Cov}[v_i, v_j](t - \tau)$. Thus the mean velocity is $\mathrm{E}[\mathrm{d}X_i/\mathrm{d}t] = \mathrm{E}[v_i(t)]$, and the fluctuation of the particle around the mean is $\mathrm{d}X_i'/\mathrm{d}t = v_i'(t)$. Therefore, the covariance of two Cartesian components of X' is given by

$$E[X_i' X_j'] = \int_0^t \int_0^t E[v_i(t')v_j(t'')]\,\mathrm{d}t'\,\mathrm{d}t''; \quad \forall i, j = 1, 2, 3 \tag{39}$$

The relationship between the dispersion coefficient and the Lagrangian velocity covariance is given by (Dagan, 1989):

$$D_{ij}(t) = \frac{1}{2}\frac{\mathrm{d}E[X_i' X_j']}{\mathrm{d}t} = \int_0^t [\mathbf{Cov}[v_i v_j](t - \tau) + \mathbf{Cov}[v_i v_j](\tau - t)]\,\mathrm{d}\tau; \quad \forall i, j = 1, 2, 3 \tag{40}$$

We present here a few solutions of the displacement tensor, X_{ii}.

Case 1 – 2D zero pore-scale dispersion: planar flow in the horizontal plane, thin aquifer with uniform mean head gradient, exponential covariance for Y, the dimensionless time is given by $t' = tU_1/I_Y$. The solution is given by

$$\frac{X_{11}(t')}{\sigma_Y^2 I_Y^2} = 2t' - 3\ln(t') + \frac{3}{2} - 3E + 3\left[Ei(-t') + \frac{e^{-t'}(1 + t') - 1}{t'^2}\right] \tag{41}$$

$$\frac{X_{22}(t')}{\sigma_Y^2 I_Y^2} = \ln(t') - \frac{3}{2} + E - Ei(t') + 3\left[\frac{1 - (1 + t')e^{-t'}}{t'^2}\right] \tag{42}$$

where E is the Euler constant.

Case 2 – 3D isotropic case, zero pore-scale dispersion. The longitudinal and transversal displacement tensor are given by

$$\frac{X_{11}(t')}{\sigma_Y^2 I_Y^2} = 2t' - 2\left[\frac{8}{3} - \frac{4}{t'} + \frac{8}{t'^3} - \frac{8}{t'^2}\left(1 + \frac{1}{t'}\right)e^{-t'}\right] \tag{43}$$

$$\frac{X_{22}(t')}{\sigma_Y^2 I_Y^2} = \frac{X_{33}(t')}{\sigma_Y^2 I_Y^2} = 2\left[\frac{1}{3} - \frac{1}{t'} + \frac{4}{t'^3} - \left(\frac{4}{t'^3} + \frac{4}{t'^2} + \frac{1}{t'}\right)e^{-t'}\right] \tag{44}$$

Case 3 – 3D anisotropic case, zero pore-scale dispersion. A more general flow case with an exponential axisymmetric covariance for Y, where e is the anisotropy ratio defined as the ratio between the vertical and the horizontal length scales, $e = I_v/I_h$. The displacements are given by (Figure 6):

$$\frac{X_{11}(t')}{\sigma_Y^2 I_Y^2} = 2t' + 2(e^{-t'} - 1) + 8e\int_0^\infty \frac{J_0(kt') - 1}{(1 + k^2 - (ek)^2)^2}\left[1 - \frac{ek}{\sqrt{1 + k^2}} - \frac{ek(1 + k^2 - (ek)^2)}{(1 + k^2)^{3/2}}\right]\mathrm{d}k - 2e\int_0^\infty \left\{J_0(kt') - \frac{J_1(kt')}{kt'} - \frac{1}{2}\right\}\left[-\frac{(ek)^3((ek)^2 - 5 - 5k^2)}{(1 + k^2 - (ek)^2)^3(1 + k^2)^{3/2}} + \frac{1 + k^2 - 5(ek)^2)}{(1 + k^2 - (ek)^2)^3}\right]\mathrm{d}k \tag{45}$$

$$\frac{X_{22}(t')}{\sigma_Y^2 I_Y^2} = -2e\int_0^\infty \left[\frac{J_1(kt')}{t'} - \frac{k}{2}\right]\left[-\frac{e(ek)^2((ek)^2 - 5 - 5k^2)}{(1 + k^2 - (ek)^2)^3(1 + k^2)^{3/2}} + \frac{1 + k^2 - 5(ek)^2)}{k(1 + k^2 - (ek)^2)^3}\right]\mathrm{d}k \tag{46}$$

$$\frac{X_{33}(t')}{\sigma_Y^2 I_Y^2} = -4e \int_0^\infty [J_0(kt') - 1] \left[\frac{1}{(-1 - k^2 + (ek)^2)^2} \left[\frac{1}{2} + \frac{2(ek)^2}{1 + k^2 - (ek)^2} - \frac{ek((ek)^2 + 3 + 3k^2)}{(1 + k^2 - (ek)^2)^3 (1 + k^2)^{3/2}} \right] \right] \mathrm{d}k \quad (47)$$

In the next section, we present how to estimate the displacement tensor from concentration measurements using the concept of statistical spatial moments. Reactive contaminant displacements in saturated and unsaturated zones are covered in Zhang (2001)

Statistical Spatial Moments

A contaminant plume is generally described using the statistical spatial moments. The pth ≤ 1 moment of the concentration distribution in space, $M_{p_1 \ldots p_n}(t)$, is defined here following Aris (1956) by

$$M_{p_1 \ldots p_n}(t) = \eta \int_\Omega C(x,t) \prod_i^n x_i^{p_i} \, \mathrm{d}\Omega \quad (48)$$

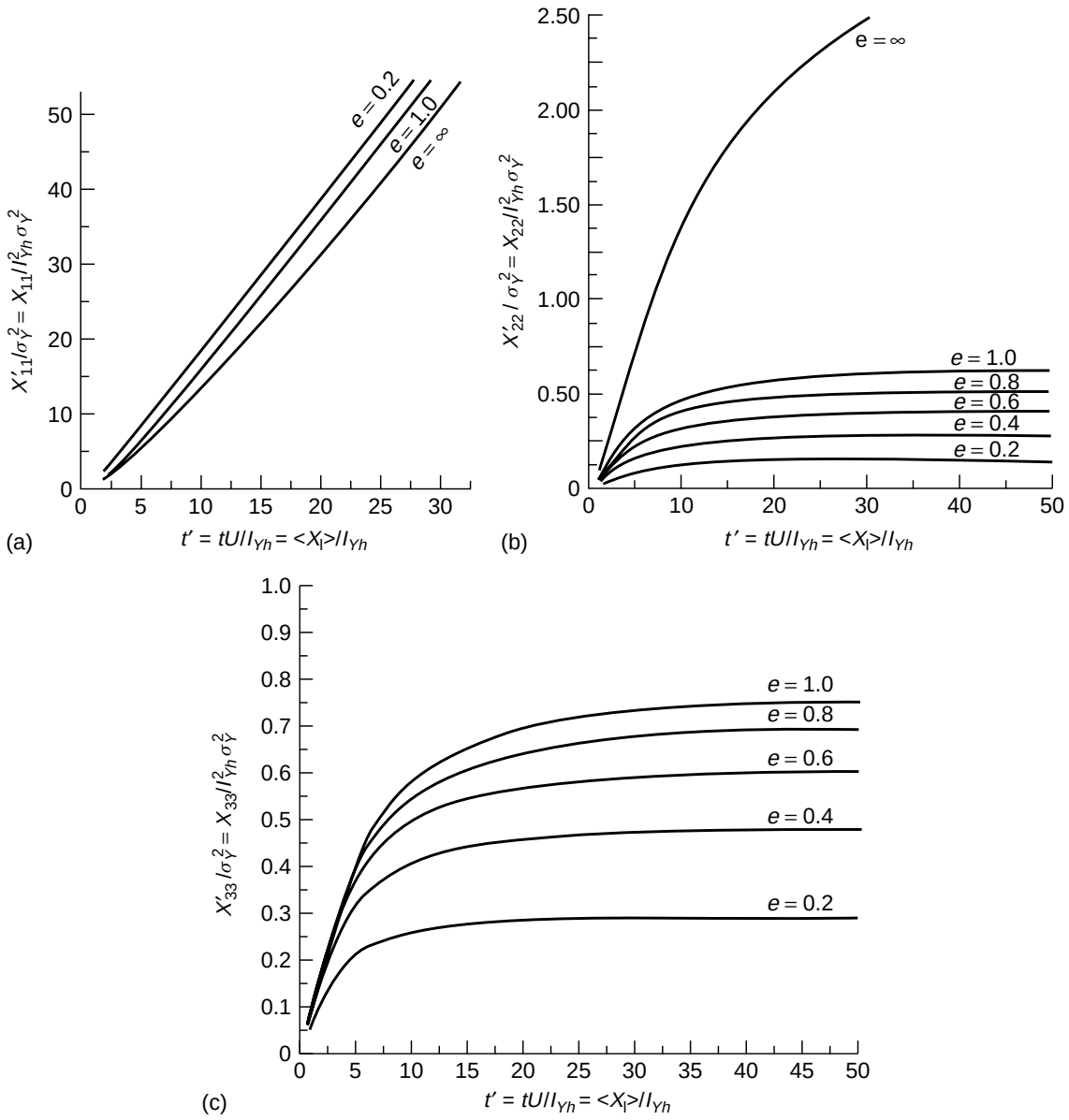


Figure 6 Longitudinal and lateral displacement as function of time (Reproduced from Dagan, 1988 by permission of American Geophysical Union)

where $C(x,t)$ is the concentration at point $x = (x_1, \ldots, x_n)$, and at time t; n denotes the space dimension and can be equal to 2 or 3, depending on the type of analysis being performed; η is the effective porosity; Ω denotes the aquifer domain occupied by the plume, and the order of the moment is given by $p = \Sigma p_i$. Here, we shall deal only with the zeroth, first, and second spatial moments of the plume. The 0th moment measures the mass of the plume, while the 1st moment provides a measure of the displacement of the centroid of the plume $\tilde{X}(t) = (\tilde{x}_{1,c}, \ldots, \tilde{x}_{n,c})$. In three-dimensional space, the centroid of the plume is given by:

$$\tilde{x}_{1,c} = \frac{M_{100}}{M_0}, \quad \tilde{x}_{2,c} = \frac{M_{010}}{M_0}, \quad \tilde{x}_{3,c} = \frac{M_{001}}{M_0} \tag{49}$$

For the case where $p \geq 2$ the central moments of the plume will be used, which are given by

$$M_{p_1 \ldots p_n}(t) = \eta \int_\Omega C(x,t) \prod_i^n (x_i{}^{p_i} - \tilde{x}_{i,c})^{p_i} \, \mathrm{d}\Omega \tag{50}$$

The 2nd moments provide a measure of the spread of the plume around its centroid. The terms of the displacement covariance tensor $\tilde{X}_{ij}(t)$are computed using the following expressions:

$$\tilde{X}_{11} = \frac{M_{200}}{M_0}, \quad \tilde{X}_{22} = \frac{M_{020}}{M_0}, \quad \tilde{X}_{33} = \frac{M_{002}}{M_0},$$
$$\tilde{X}_{12} = \frac{M_{110}}{M_0}, \quad \tilde{X}_{13} = \frac{M_{101}}{M_0}, \quad \tilde{X}_{23} = \frac{M_{011}}{M_0} \tag{51}$$

In applications, Ω is not known exactly, and in many cases it is only partly covered by samplers. Hence, prior to computing the moments, there is a need to estimate Ω. This problem is referred as the "null points" problem, and it has been approached in previous works using methods such as linear extrapolation or surface fitting using spline functions or kriging. A concise presentation of these methods is presented in Chapter **Chapter 147, Characterization of Porous and Fractured Media, Volume 4**.

We illustrate the application of the spatial statistical moments on the large-scale field experiment conducted at Cape Cod (Leblanc *et al.*, 1991) The Cape Cod large-scale natural gradient tracer test began in July 1985 with the injection of 7.6 m^3 of tracer solution in the aquifer through a volume of dimensions of $1.2 \times 3 \times 4\,m^3$. The tracer solution contained a nonreactive tracer, bromide, and reactive tracers, lithium and molybdate. For illustration purposes, we shall concentrate on the displacement of the bromide plume. The plume was sampled over a period of 511 days. In each of the 16 sampling sessions, concentration was measured using a large and dense three-dimensional array of samplers (see Figure 7). Zeroth, first-, and second-order moments were computed and given in Figures 8, 9(a, b), and 10(a–c) (Rubin and Ezzedine, 1997).

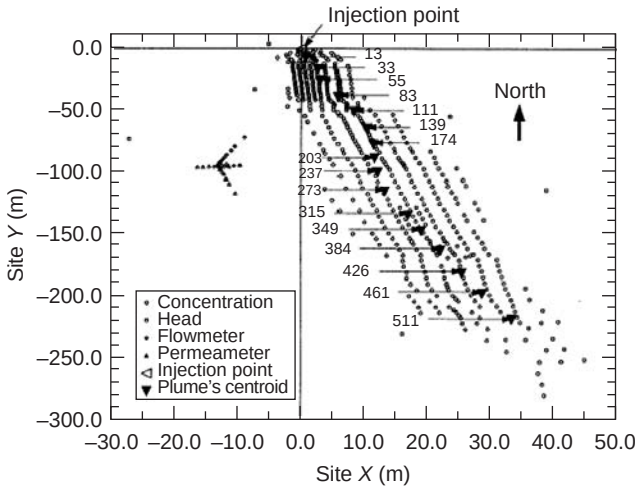


Figure 7 In situ locations of the concentration samplers, circles; the piezometers, squares; the hydraulic conductivity measurements using flowmeter, diamonds; and the permeameter, triangles. The centroids are shown and are referred to by elapsed travel time (in days) (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

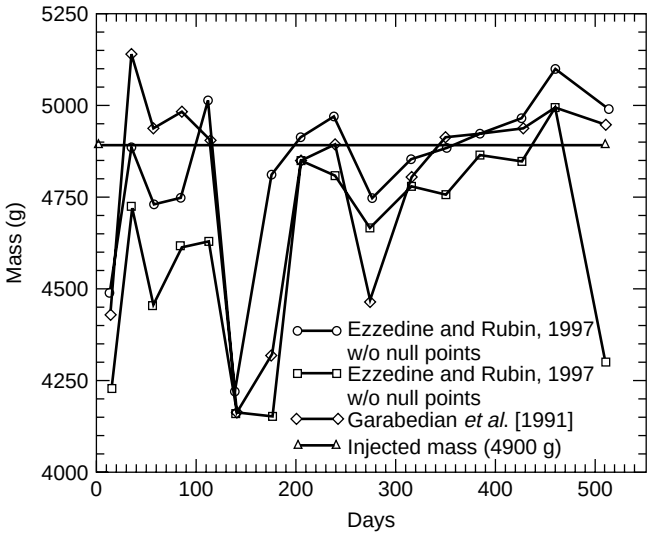


Figure 8 Mass recovery versus time: for (Rubin and Ezzedine, 1997), circles; original data without null points, squares; Garabedian *et al.* (1991) results, diamonds; and the injected mass (4900 g), triangles (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

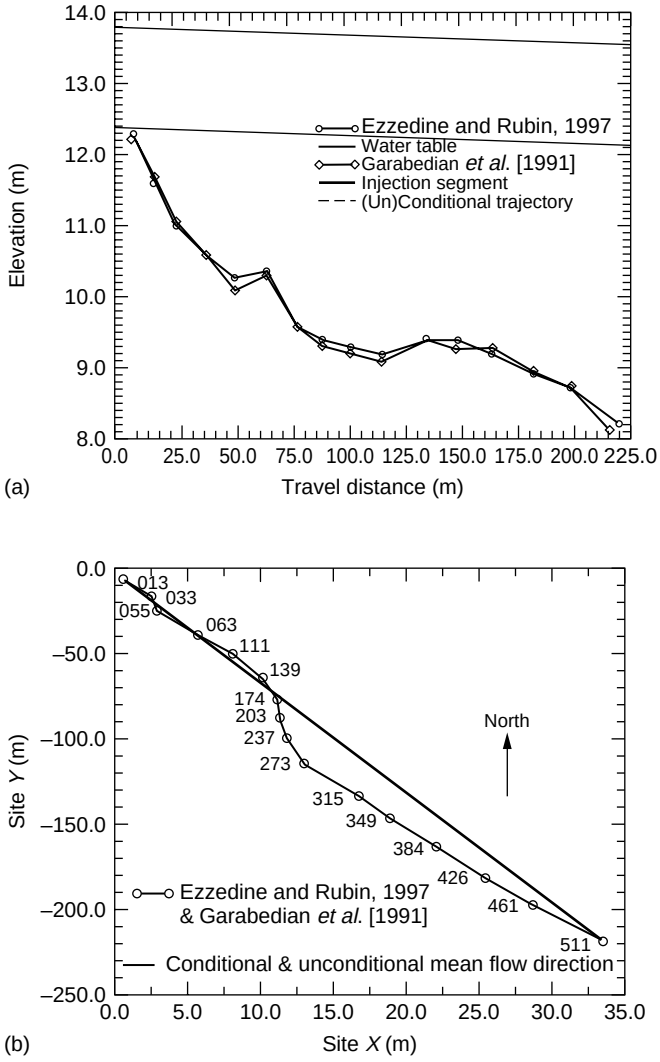


Figure 9 First moments: (a) Horizontal plume's centroid (2) plotted in original coordinates. The injection point is at (site $X = 0$, site $Y = 0$). Centroids are marked with the time elapsed since injection in days. The conditional and the unconditional trajectories are identical. Trajectories computed using the data and the Garabedian *et al.* (1991) results are the same. (b) Vertical plume's centroid versus travel distance (2). Elevation is above mean sea level. The conditional and the unconditional trajectories are identical (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

Ergodicity and Plume Size Effect on Displacement Tensor

A solute plume is ergodic if the mean and covariance of the displacement along the various streamlines are equal to the mean and covariance of the displacement of a single particle. In this case, the ensemble-average concentration plume dispersivity is expected to accurately predict the plume growth using single-particle displacement statistics only when the plume size is much larger than the integral scales of the hydraulic conductivity (Dagan, 1989). It is therefore important to investigate the applicability and limitation of the ergodic hypothesis. On the basis of the concept of relative dispersion, Dagan (1989) reached the following relationship:

$$E[S_{ij}(t)] = S_{ij}(t=0) + X_{ij}(t) - R_{ij}(t) \qquad (52)$$

where $S_{ij}(t)$ is the second spatial moment of the plume of a plume of arbitrary dimension as a random function of trajectories of the particles:

$$S_{ij}(t) = \frac{1}{|\Omega|}\int_{\Omega}[X_i(t|\Omega) - R_i(t)][X_j(t|\Omega) - R_j(t)]\ \mathrm{d}\Omega;$$

$$R_i(t) = \frac{1}{|\Omega|}\int_{\Omega} X_i(t|\Omega)\ \mathrm{d}\Omega \qquad (53)$$

where $R(t)$ is plume centroid. Because X_i is a random function, R is inherently so. $X_i(t|\Omega)$ is the trajectory of a particle emanating from a source of size Ω with and a volume $|\Omega|$. The variance of the centroid of the plume is given by

$$R_{ij}(t) = E[(R_i(t) - E[R_i(t)])(R_j(t) - E[R_j(t)])]$$
$$= \frac{1}{|\Omega|^2}\int_{\Omega}\int_{\Omega} E[X_i'(t|\Omega)X_j'(t|\Omega')]\ \mathrm{d}\Omega\,\mathrm{d}\Omega' \qquad (54)$$

If the plume spreads are much larger than the integral scale of Y, the statistical variability of R becomes narrow and R approaches the ensemble mean displacement. An elaborate discussion of ergodicity and operational ergodicity is given in Rubin (2003, Chapter 10). Ergodicity analysis was performed on the Cape Cod data and results are given in Figures 11(a–c).

Spatial moment concepts could be extended to temporal statistical moments. They are introduced in the following section.

Statistical Temporal Moments

The Lagrangian framework is most suitable for modeling and analyzing travel times. One compelling reason is the lack of discrepancy between the measurement scale in the field and the scale of numerical element used for modeling. Samplers in the field measure the concentration at a streamline, and the streamline is also the fundamental unit in the Lagrangian approach. However, the elegance of the Lagrangian-based solutions can sometimes be lost when confronted by the reality of field data analysis. To illustrate this matter, we proceed by presenting temporal moments statistics. Similar to spatial moments, the temporal moments are commonly defined by the following integrals:

$$\mu_i(x) = \int_{-\infty}^{+\infty} t^i C(x,t)\ \mathrm{d}t \qquad (55)$$

where $\mu_i(x)$ are the ith temporal noncentral moments. To demonstrate these concepts, let us consider the moments of

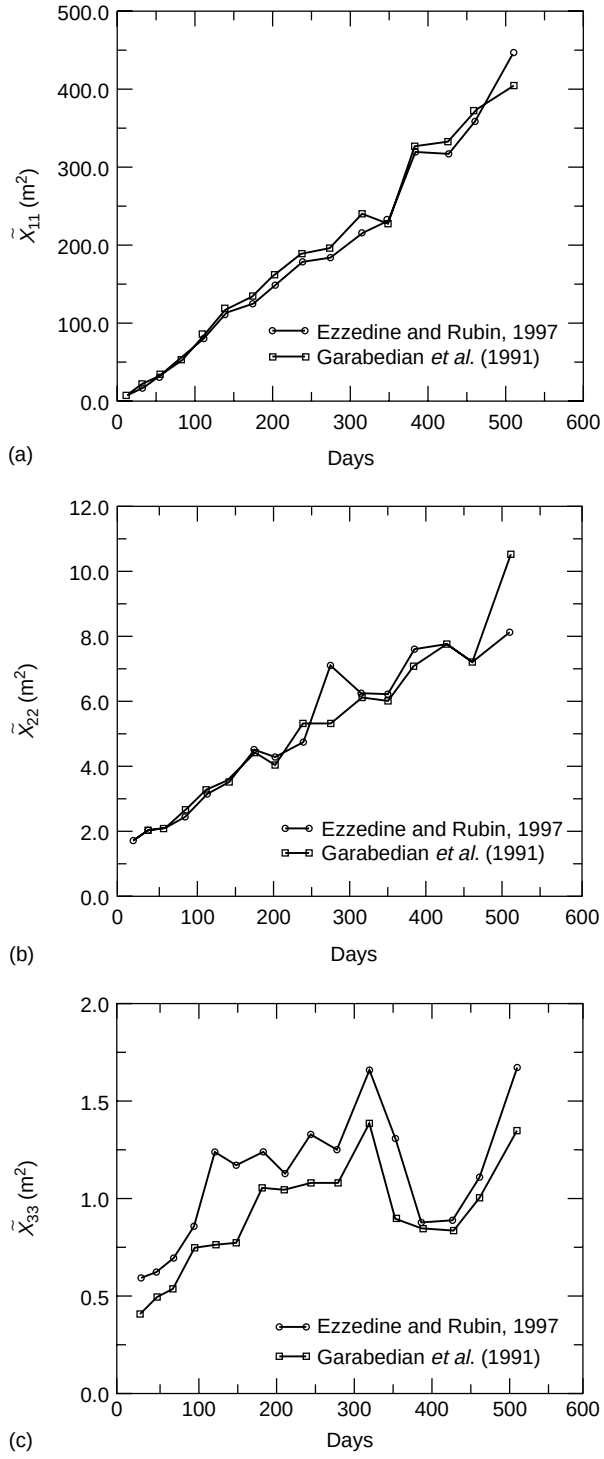


Figure 10 (a) The longitudinal central second moment $\tilde{X}_{11}$ as computed in the present study and by Garabedian *et al.* (1991), (b) The lateral second central moment $\tilde{X}_{22}$ as computed in the present study and by Garabedian *et al.* (1991), (c) The vertical second moment $\tilde{X}_{33}$ as computed in the present study and by Garabedian *et al.* (1991) (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

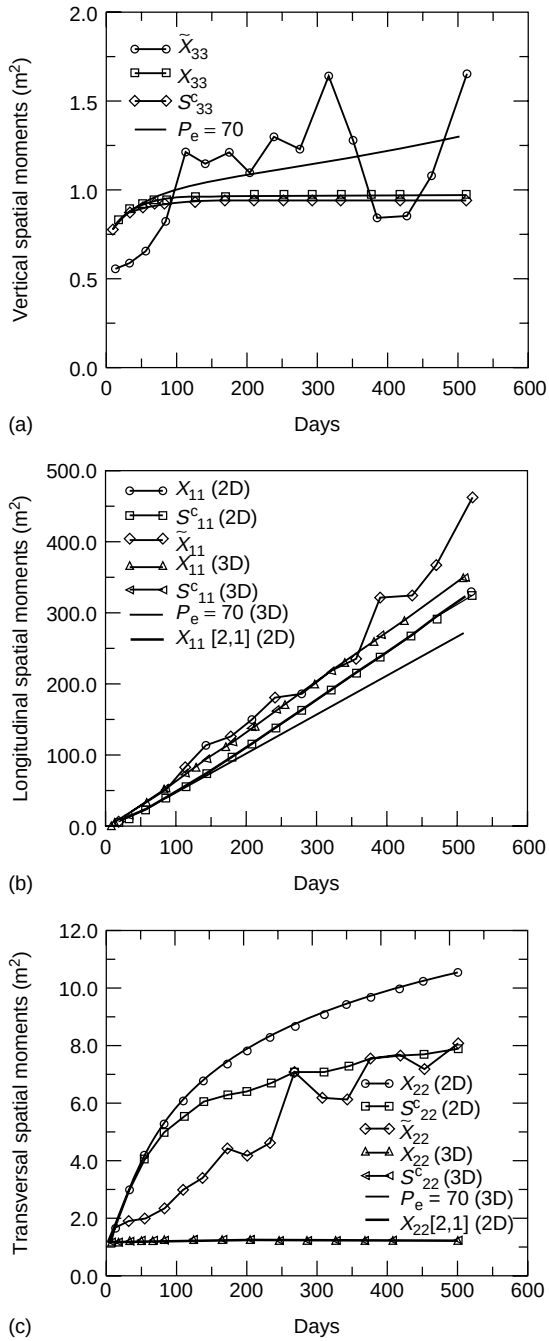


Figure 11 The 3D conditional S^c_{33} computed using (52), the conditional X_{33}, experimental results of vertical spread, and 3D pore scale dispersion (($P_e = 70$), Fiori, 1996). b) the 2D and 3D S^c_{11} computed using (52), unconditional X_{11}, 2D unconditional higher-order approximation X_{11}[1,2] (Hsu *et al.*, 1996) and 3D X_{11} with pore scale dispersion (($P_e = 70$), Fiori, 1996). c) the 2D and 3D S^c_{22} computed using (52), unconditional X_{22}, 2D unconditional higher-order approximation X_{22}[1,2] (Hsu *et al.*, 1996) and 3D X_{22} with pore scale dispersion (($P_e = 70$), Fiori, 1996) (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

the travel time at *s(L)* where L is a distance downstream from the injection source located at x_0, and s is a streamline coordinate system. Moreover, we assume that the tracer is injected at x_0 over a short period of time T_0, such that the following boundary condition for the advection dispersion equation holds: $C(x_0, t) = C_0$ for $t \in [-T_0/2, T_0/2]$ and $C(x_0, t) = 0$ otherwise. Neglecting pore-scale dispersion, the concentration distribution along a streamline is given by the square wave

$$C(t, \tau) = C_0 \left[H\left(t - \tau + \frac{T_0}{2}\right) - H\left(t - \tau - \frac{T_0}{2}\right) \right] \tag{56}$$

where H is the Heaviside step function, and τ, is the time that it takes for a particle emanating from x_0 to reach *s(L)* and is given by

$$\tau(L) = \int_{s(x_0)}^{s(L)} \frac{\mathrm{d}s}{V(s, x_0)} \tag{57}$$

where V denotes velocity. Applying the travel time to the square wave solution leads to the first three travel time moments:

$$\mu_0 = C_0 T_0, \mu_1 = C_0 T_0 \tau,$$
$$\text{and} \quad \mu_2 = C_0 T_0 \left(\tau^2 + \frac{T_0^2}{12}\right)$$
$$\text{and thus} \quad \tau = \frac{\mu_1}{\mu_0} \quad \text{and} \quad \tau^2 = \frac{\mu_2}{\mu_0} - \frac{T_0^2}{12} \tag{58}$$

The spatial average of τ over a control plane of area A perpendicular to the mean flow direction at L distance from the source is given by

$$\overline{\tau}(L) = \frac{1}{A} \int_A \frac{\mu_1(x_0 + L, y, z)}{\mu_0(x_0 + L, y, z)} \mathrm{d}y\, \mathrm{d}z;$$
$$\text{and} \quad \sigma_\tau^2(L) \equiv (\tau^2) - (\overline{\tau})^2 = \frac{1}{A} \int_A \frac{\mu_2(x_0 + L, y, z)}{\mu_0(x_0 + L, y, z)} \mathrm{d}y\, \mathrm{d}z - \frac{T_0^2}{12} - (\overline{\tau})^2 \tag{59}$$

In the Lagrangian approach, mass transfer across streamlines cannot be modeled, yet this aspect of the transport problem must be dealt with in order to make the travel time analysis wieldy. Hereafter, an approach is presented that will allow one to retain the simplicity of the Lagrangian approach yet account realistically for the effects of pore-scale dispersion.

Inspection of the time records of concentration at various samplers reveals that the above theoretical development is somewhat limited (Figure 12). Rather than a single square pulse, or even a diffused, Gaussian-like pulse, we observe trains of signals of different modes, stretching over periods of time much longer than the injection period.

Since the Cape Cod plume was injected over a period of 18 h, and the time-concentration records show pulses that take much longer to travel across individual samplers, it becomes clear that computing the travel time moments based on the entire time-concentration records is the culprit for the mismatch between the theoretical and experimental travel time variances (Figure 13).

To use the time records for inference, the effects of pore-scale dispersion need to be removed nonarbitrarily, and the geometry of the advective pulse reconstructed. To do so, it is suggested that the travel time t to each sampler be taken as the time at which the maximum concentration $C_{\max}$ is detected. We shall refer below to that time as $tc_{\max}$. The parameter $tc_{\max}$ is taken as the most likely estimator of τ, instead of using the Aris estimator. The justification for choosing the peak arrival time as the travel time least affected by pore-scale dispersion is based on the observation that this travel time is associated with the plume centroid, where the concentration is generally the largest and the concentration gradients equal to zero. Additionally, if one considers an ensemble of streamlines with a solute pulse moving in any of them, possibly with different velocities, it can be expected that the concentration that develops in any streamline due to lateral mass transfer will be smaller than that of the pulse that was originally injected, at least over a long time, until the concentration variations smooth out. The first two moments of the travel time are now given by

$$\overline{\tau}(x) = \frac{1}{A} \int_A t_{C_{\max}}(x, y, z)\ \mathrm{d}y\ \mathrm{d}z;$$
$$\sigma_\tau^2(x) = \frac{1}{A} \int_A [t_{C_{\max}}(x, y, z) - \overline{\tau}(x)]^2\ \mathrm{d}y\ \mathrm{d}z \tag{60}$$

The assumption underlying both moments is that $C_{\max}$ is the concentration of the pulse's centroid and that by tracking the displacement of the pulse's centroid, it is possible to capture the effects of advection while removing the effects of pore-scale dispersion. The viability of the suggested approach predicates upon the existence of a distinguishable $C_{\max}$. The time $tc_{\max}$ will become nonidentifiable as the concentration field smoothes out with time due to the effects of pore-scale dispersion.

Figure 14(a) depicts the mean travel time as a function of distance from the source. The results obtained by averaging $tc_{\max}$ over the planes normal to the mean flow direction are very similar to those obtained using Aris moments of order 0, and 1, for both analyses yield results which are close to those obtained using the average velocity. These results suggest that removing the effects of pore-scale dispersion is inconsequential for first moments analysis

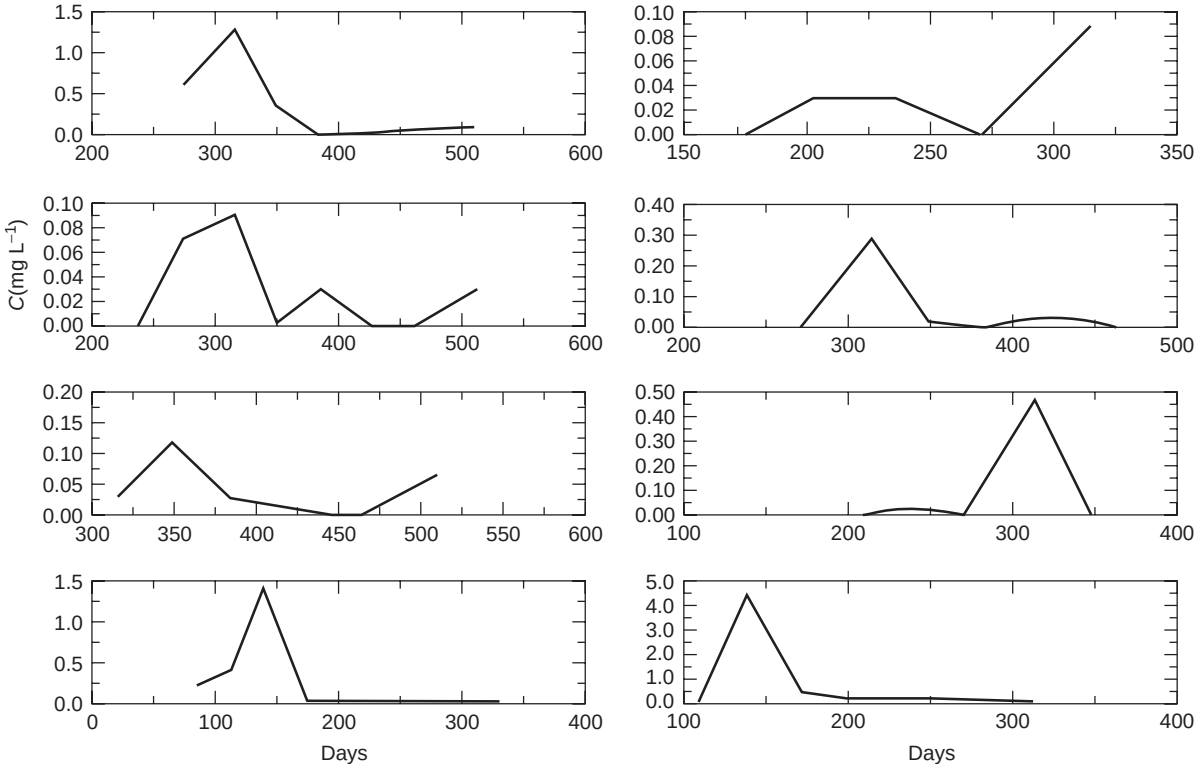


Figure 12 Examples of the concentration versus time records at several samplers. Typically, a high-level detect is followed or preceded by a long train of low-level detects (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

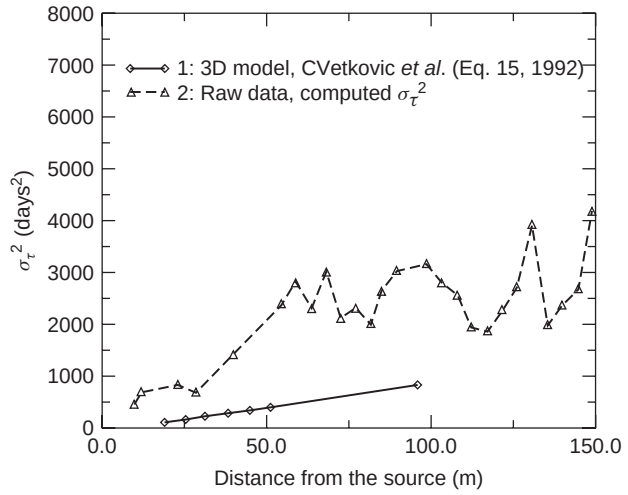


Figure 13 Theoretical and experimental travel time variance (Reproduced from Rubin and Ezzedine (1997) by permission of American Geophysical Union)

and that the large distribution of the pulses over the flow domain cancels out the effects of the low-level concentrations due to averaging. The $\langle\tau\rangle$ model following Rubin and Dagan (1992) is based on a transformation of the single-particle displacement variance, and it shows excellent agreement with the observed data. In analyzing the travel time variances, the situation becomes quite different. Figure 14(b) shows the travel time variance computed using the ensemble of the experimental $tc_{\max}$ over consecutive planes of multilevel samplers. The first and second moments were computed over consecutive planes and compared with the Aris-based experimental results. This comparison shows a dramatic reduction in the magnitude of the experimental variances. A comparison of the experimental results with theoretical results shows a good match.

INVERSE PROBLEMS AS STATISTICS

Much of the problem and initial cost of subsurface remediation concerns field site characterization. A full three-dimensional "picture" of the heterogeneous subsurface is needed in order to identify the key controls on the flow and contaminant transport processes. Natural heterogeneity and the large spatial variability of the permeability predominantly control the flow field and hence the transport. Moreover, natural heterogeneity exhibits variability over a wide range of scales and hence is difficult to characterize due to the scarcity of data and the costliness of

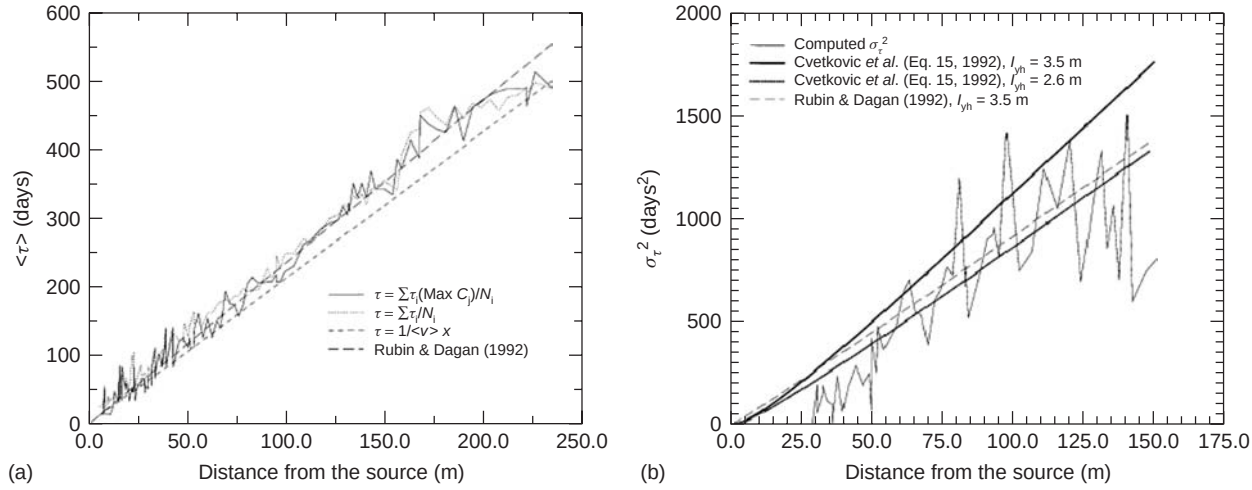


Figure 14 Theoretical and experimental (a) 1st moments (b) 2nd moments

conventional field sampling techniques such as drilling. With poor site characterization, remediation schemes are unnecessarily expensive, because costly overdesign may be required to compensate for uncertainty. Therefore, inverse problems were formulated in order to estimate the hydraulic properties from a spare data of measurements. This section describes how to determine model parameter values. Models are assumed to be valid; the only unknowns are parameter values that define the models. For completeness, we introduce some concepts and terminology commonly used in inverse/forward problem community. An elaborate discussion of the deterministic approach to inverse problems is given in Chapter **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**.

Well-posed versus Ill-posed Problem

Prediction based on a given set of parameter values is called "forward modeling". Determination of parameter value from observed data is called "inverse modeling". Inversion can be achieved in two ways. On one hand, a modeler iteratively modifies parameter values (such as hydraulic conductivity), run a forward model until attaining best "fit" or "match" between the measurements and the modeled variable, head for instance. This kind of process falls into the trial and error methods. Such forward modeling is sometimes tedious and time consuming. On the other hand, an inverse algorithm is adopted to automatically or semiautomatically obtain the parameter values from the observed data and an initial set of trial parameters values. The procedure also provides an estimate of parameter uncertainty. In both cases, inversion requires minimizing the discrepancy between predictions and observations (Sun, 1994).

A well-posed inverse problem requires "existence" of the problem, the "uniqueness" and the "stability" of the solution or algorithm. Obviously, in view of observed data and our understanding of a real-world physical system, a problem is presumed to exist, for example, detection of contaminant plume in groundwater suggests contamination must have happened in the past. The question then is how to relate the observation to the migration history of the contaminant. A cause generally has an effect. Can an effect result from different causes? Is it unique in theory or model? Even if it is, have we counted and resolved all parameters that define a model? Uniqueness has two levels: the model itself and the model-defining parameters. The latter is related to the stability of a solution algorithm. How sensitive are parameters to uncertainty of observed data? Are the errors amplified during inversion? Is the inversion algorithm efficient in terms of ease of usage and cost of running the inversion program (complexity)? Mathematically speaking, inverse problems are ill-posed problems.

Deterministic versus Stochastic Inversion

To address the inverse problem, two main frameworks have been developed and they are either deterministic or stochastic. In the deterministic framework, the structure of the spatial variability of the parameters is fixed. For example the aquifer is divided into a number of zones, and each zone is supposed to have a constant hydraulic conductivity; then, the algorithm seeks the best hydraulic conductivity values, for which the solution of the flow equation reproduces the observed data. In a stochastic framework, however, the spatial variability of the parameters is statistically mapped. For example, the overall mean and variance, and the variogram of the final realization are specified. This characterization is not enough to fully determine the parameter values at every cell. Thus, a spatial realization is sought, satisfying the statistical constraints and honoring the observed values. As it has been shown in Section "Geostatistics", an infinite

number of realizations can meet the statistical constraints and reproduce the observed data. To make these inferences quantitative in either deterministic or stochastic framework, one must answer three fundamental questions:

- How accurately is the data known, that is, what does it mean to "fit" the data?
- How accurately can we model the response of the aquifer system? In other words, have we included all the physics in the model that contribute significantly to the data?
- Finally, what is known about the system besides the data? This is called *a priori* information and is essential since for any sufficiently fine parameterization of an aquifer system there will be unreasonable models that fit the data too. Prior information is the means by which we reject or down-weight unreasonable estimation models.

There are a variety of "recipes" for constructing estimators; perhaps the most common in statistics are, minimum variance estimation, maximum likelihood estimation (MLE) and Bayesian estimation. The latters have asymptotically optimal properties under certain restrictive assumptions. Hereafter we present three inverse problems; they are introduced in order of complexity.

Illustrations and Examples

A Simple Statistical Inference Problem: Least-square Method

To illustrate the deterministic inversion process, let us consider a classic example of parameters estimation for a pumping test (readers are referred to **Chapter 151, Hydraulics of Wells and Well Testing, Volume 4** for more elaborate details). The solution of a pumping test in a confined aquifer is given by Jacob's semilog approximation for large time:

$$\delta h \approx \frac{Q}{4\pi T}\left(-0.5772 - \ln\left[\frac{r^2 S}{4T}\right] + \ln[t]\right) \tag{61}$$

where T is the transmissivity, S is the storativity, Q is the pumping rate, t is time, h is the hydraulic head and δh is the drawdown. The inverse problem is to determine T and S given a series of error-free observations of δh at different times. The relation (model) between the drawdown and time is not linear but logarithmic. However if $\ln[t]$ is treated as an independent variable the relation between δh and $\ln[t]$ is a linear one and the inversion in the least-square sense is straightforward. The sum of the square of discrepancy ε_i between each predicted δh_i and measured $\delta h_i{}^{\text{obs}}$ is minimized in least-square sense. If the above equation is rewritten as follows:

$$y = ax + b; \quad x \equiv \ln[t]; \quad a = \frac{Q}{4\pi T};$$
$$b = \frac{Q}{4\pi T}\left(-0.5772 - \ln\left[\frac{r^2 S}{4T}\right]\right); \; y \equiv \delta h \tag{62}$$

The goal is then to minimize the sum or the square of errors (minimum variance):

$$\text{Min } \delta^2 = \text{Min}\left[\sum_{i=1}^{n} \delta_i^2\right] = \text{Min}\left[\sum_{i=1}^{n} (y_i - y_i^{\text{obs}})^2\right] \tag{63}$$

The solution of the problem is readily available in textbooks and it is given by:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i^{\text{obs}} - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i^{\text{obs}}}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$$
$$b = \frac{n\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^{\text{obs}} - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i^{\text{obs}}}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{64}$$

a and b can be expressed in univariate statistics as shown on Table 1.

The expected transmissivity value is then determined by $E[T \mid \ln(t)] = a \ln[t] + b$.

Cokriging Methods

It has already been pointed out that the final hydraulic property map (realization) cannot disregard the measurement data; they are the only factual knowledge available about the aquifer. However, aquifers are systems, the state of which is described by the spatial distribution of piezometric heads, and by the concentration of the solutes dissolved in water. In general, there is more information about the state of the system than about the parameters that controls it. Therefore, it appears necessary to generate spatial distributions of the parameters that are not only conditional to parameter values, but also consistent with the (partial) knowledge about the state of the system. This matter has already been introduced in the Section "Geostatistics". We present here two examples for completeness.

Cokriging Head and Hydraulic Conductivity

Rubin (2003) used the analytical approach to solve the perturbed flow equation. Rubin calculated $h' = h - E[h]$ and $Y' = Y - E[Y]$ at the points of h, head, and Y, log of

transmissivity; and determined analytically the covariance function of h' and the cross-covariance (h', Y') as a function of the covariance of Y. The covariance of Y is function of a set of parameters θ (i.e. integral scale of Y and its variance). This is actually sufficient to estimate the transmissivity field by cokriging. The cokriging estimator then gives the optimal estimation of Y at any point as follows:

$$Y(x) = \sum_{i=1}^{n_Y} \lambda_i Y_i + \sum_{j=1}^{n_H} \upsilon_j (h_j - E[h_j]) \qquad (65)$$

where the λ_i and the υ_j are optimal weights, that depend on the position x. The cokriging equations that provide the value of the optimal weights simply require that the covariance functions of Y, of h and of h–Y be known. Rubin then calculate by cokriging all the values of Y at the measurement points where Y is known and where it is therefore possible to compare the known value with the one estimated by cokriging – without using the known value of this point in the cokriging equations. As the cokriging estimator is a function of the θ parameters, these parameters can thus be optimized to minimize the errors between the estimated and measured Y values. The Maximum Likelihood method was used for their optimization (*see* Chapter **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**). Once the θ parameters are known, the cokriging equations give an estimation of Y at all points and a map of Y is obtained (Figure 15).

Cokriging Head and Hydraulic Conductivity for Concentration Estimations The goal is to condition the first two moments of the concentration on measurements of conductivity K and hydraulic head H. In order to model their spatial variability, Y and H will be modeled here as space random functions, so that $Y(x) = m_Y + Y'(x)$, where $m_Y = \langle Y(x) \rangle$, and $H(x) = \langle H(x) \rangle + H'(x)$. Hence Y' and H' represent the fluctuations of Y and H about their respective means. The hydraulic head is assumed to be at steady state. Small variability of the log-conductivity is assumed, and the linearized steady state flow equation is used to relate between Y' and H' and other hydrogelogical variables. Under these conditions, H' becomes a linear function of Y', and their cross-covariance can be derived analytically. Similarly, the concentration C is also viewed as a space random function and we define $C(x) = \langle C(x) \rangle + C'(x)$, where $\langle C(x) \rangle$ is the unconditional mean of $C(x)$.

The goal can now be stated as follows: define $\langle C^c(\mathbf{x}, \tau) \rangle$ and $\sigma^{2,c}_{\langle C \rangle}$, the conditional mean of the concentration and its estimation variance, where conditioning is done over N measurements, M of which are log-conductivities, and the rest are head measurements. These estimates should also reflect the dependence of C on time, τ. A superscript "c" denotes estimators, which are conditional to measurements. This task is accomplished using the following relationships for the conditional mean:

$$\langle C^c(x, \tau) \rangle = \langle C(x, \tau) \rangle + \sum_{i=1}^{M} \lambda_i(x, \tau) Y'(x_i) + \sum_{j=M+1}^{N} \mu_j(x, \tau) H'(x_j) \qquad (66)$$

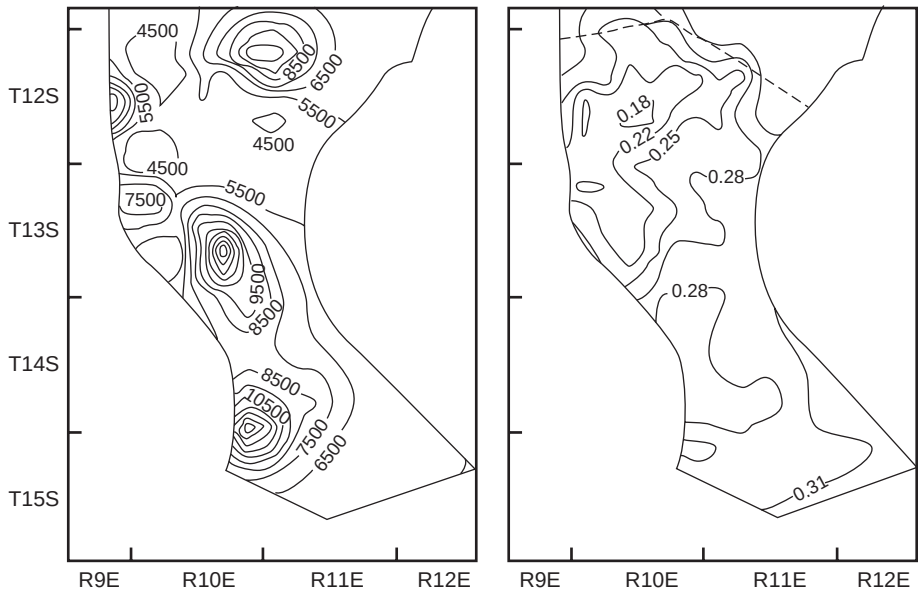


Figure 15 Estimated transmissivity and its variance (Reproduced from Rubin and Dagan, 1987 by permission of American Geophysical Union)

and the conditional estimation variance:

$$\sigma^{2,c}_{\langle C\rangle}(x,\tau) = \sigma^2_{\langle C\rangle}(x,\tau) - \sum_{i=1}^{M} \lambda_i(x,\tau) C_{CY}(x,\tau;x_i) - \sum_{j=M+1}^{N} \mu_j(x,\tau) C_{CH}(x,\tau;x_j) \quad (67)$$

where $\sigma^2_{\langle C\rangle}$ is the unconditional variance of the concentration. The above equations constitute the well-known geostatistical formalism of cokriging. Hence, the estimates of C will be unbiased and of minimum variance. A similar set of equations can be written for the conditional moments of Y and H, using the same cross-covariances, which appear in the above equations. The coefficients $\lambda_i(x,\tau)$ and $\mu_j(x,\tau)$ are the solutions of the following linear system (Journel and Huijbregts, 1978):

$$C_{CY}(x,\tau,x_k) >= \sum_{i=1}^{M} \lambda_i(x,\tau) C_Y(x_i,x_k) + \sum_{j=M+1}^{N} \mu_j(x,\tau) C_{YH}(x_j,x_k), \quad k = 1,\ldots,M \quad (68)$$

$$C_{CH}(x,\tau,x_k) >= \sum_{i=1}^{M} \lambda_i(x,\tau) C_{YH}(x_i,x_k) + \sum_{j=M+1}^{N} \mu_j(x,\tau) C_H(x_j,x_k), \quad k = M+1,\ldots,N \quad (69)$$

In the above equations the following definitions are used to denote spatial moments: $C_Y(x,x') = \langle Y'(x)Y'(x')\rangle$, $C_H(x,x') = \langle H'(x)H'(x')\rangle$, $C_{YH}(x,x') = \langle Y'(x)H'(x')\rangle$, $C_{CH}(x,\tau,x') = \langle C'(x\tau)H'(x')\rangle$ and $C_{CY}(x,\tau;x') = \langle C'(x,\tau)Y'(x')\rangle$. Given a set of N measurements, one can map the conditional expected concentration and its variance. C_{CY} and C_{CH} are depicted on Figure 16(a,b), respectively (Ezzedine and Rubin, 1996). Once these cross-covariances are analytically determined, one can proceed by mapping C conditional to Y and H measurements.

Cokriging Using Displacement and Travel Time for Concentration Estimations As part of the previous developments, the cross-covariances between Y and $X(\tau)$ were obtained. That suggests the possibility of using the time-displacement data to update the estimates of Y. Measured displacement can be coupled or perhaps used as an alternative to concentration measurements.

In fact, the use of displacements and travel times may be a better idea than using concentration data. For once, the concentration of a pulse is affected by pore-scale dispersion, while the travel times and displacements of the centroid of the pulse are affected to a much lesser extent. The concentration cross-covariances C_{CY} and C_{CH} were developed under the limiting assumption of negligible pore-scale dispersion, and by using travel times and displacements, we can avoid this limitation altogether.

In order to demonstrate the benefits of using travel time and displacements versus concentration, consider the following admittedly simplistic case. Solute of concentration C_0 is injected over a very short period. The displacement of the pulse is monitored. Let us assume that at time τ the pulse was located at $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. In the absence of pore-scale dispersion, the concentration is equal to the initial one, that is, C_0. It is clear that

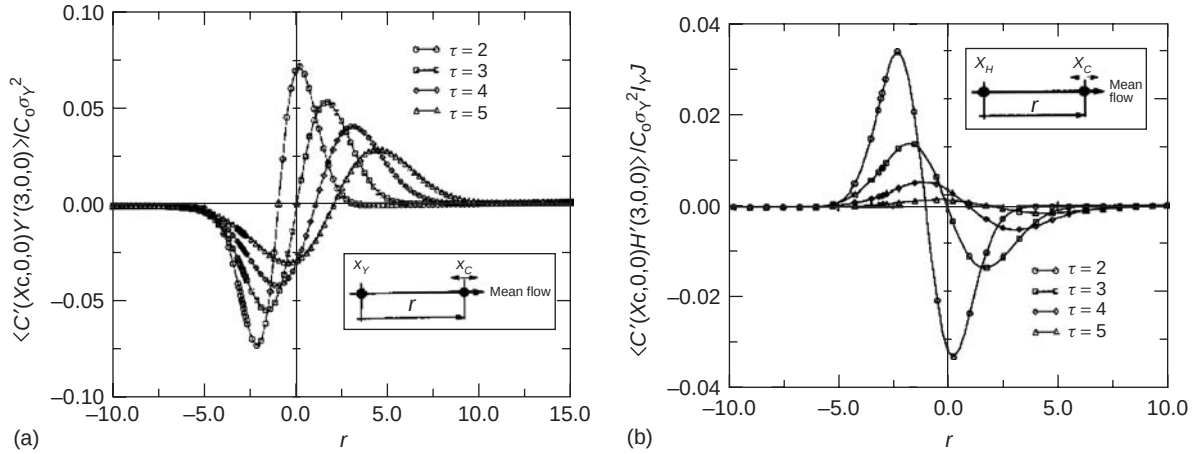


Figure 16 (a) Cross-covariances $\tilde{C}_{CY}(\mathbf{x}_C,\tau;\mathbf{x}_Y)$ for fixed $\mathbf{x}_Y = (3,0,0)$, $\mathbf{x}_C = (x_C,0,0)$ and τ are variables, and $r = x_C - 3$ (b) Cross-covariances $\tilde{C}_{CH}(\mathbf{x}_C,\tau;\mathbf{x}_H)$ for fixed $\mathbf{x}_H = (3,0,0)$, $\mathbf{x}_C = (x_C,0,0)$ and τ are variables, and $r = x_C - 3$ (Reproduced from (Ezzedine and Rubin, 1996) by permission of American Geophysical Union)

the magnitude of the displacement $X(\tau) = \alpha$ as well as the measurement $C(\alpha, \tau) = C_0$ are equivalent in term of the information content. Next we wish to compute the conditional mean and conditional variance of the pulse's longitudinal displacement. These moments can be related to the moments of the concentration (Ezzedine and Rubin, 1996). Conditioning on the concentration through kriging yields to $\langle X_1{}^c(\tau)\rangle = \langle X_1(\tau)\rangle + \lambda\ (C_0 - \langle C(\alpha, \tau)\rangle)$ and the conditional variance $X_{11}{}^c(\tau) = X_{11}(\tau) - \lambda\langle C'(\alpha, \tau)X_1'(\tau)\rangle$ with $\lambda = \langle C'(\alpha, \tau)X_1'(\tau)\rangle/\sigma_C^2(\alpha, \tau)$. The cross-covariance $\langle C'(\alpha, \tau)X_1'(\tau)\rangle$ is given by

$$\langle C'(\alpha, \tau)X_1'(\tau)\rangle = C_0(\alpha_1 - \langle X_1(\tau)\rangle)$$
$$\text{Prob}[\mathbf{X}(\tau) = \boldsymbol{\alpha}] = (\alpha_1 - \langle X_1(\tau)\rangle)\langle C\rangle \qquad (70)$$

we used $\text{Prob}[\mathbf{X}(\tau) = \alpha] = \langle C\rangle/C_0$. Recalling that $\sigma_C^2(\alpha, \tau) = \langle C\rangle(C_0 - \langle C\rangle)$ (Dagan, 1989), the conditional displacement is reduced to $\langle X_1{}^c(\tau)\rangle = \alpha_1$ and $X_{11}{}^c(\tau) = X_{11}(\tau) - (\alpha_1 - \langle X_1(\tau)\rangle)^2/(C_0 - \langle C\rangle)\langle C\rangle$. In the particular case of $\alpha_1 = \langle X_1(\tau)\rangle$ we get $\langle X_1{}^c(\tau)\rangle = \langle X_1(\tau)\rangle$ and $X_{11}{}^c(\tau) = X_{11}(\tau)$ which are the unconditional moments, that is, no reduction at all in the displacement variance is achieved: the result for the first moment is clearly right while the result for the variance is unsatisfactory because it does not show any benefit from the measurement. But this result is not wrong: it is a direct outcome from the $\langle C'(\alpha, \tau)X_1'(\tau)\rangle$ cross-covariance being zero for $\alpha_1 = \langle X_1(\tau)\rangle$: although C and X are not independent, they are not linearly correlated, and $C = 0$ may occur for X_1' either positive or negative (Figure 16a). Repeating this exercise but conditioning on the measured displacement instead will lead to $X_{11}{}^c(\tau) = 0$ due to the exactitude property of the kriging interpolator. The reason the displacement is more effective than the concentration is that unlike concentration, no ambiguity arises from its interpretation: a positive X_1', for example, can only arise from a positive Y'. Hence by a mere change of the interpretation we are able to make a better use of the data.

Bayesian Inversion

For a statistician, an inverse problem is an inference or estimation problem. The data are finite in number and contain errors, as they do in classical estimation or inference problems, and the unknown typically is infinite dimensional, as it is in nonparametric regression. The additional complication in inverse problem is that the data could be directly and indirectly related to the unknown. Bayesian techniques have become more attractive in the hydrogeological communities through the elegant work of Tarantola (1987). One of the fundamental tenets of Bayesian inference is that uncertainty always can be represented as a probability distribution; in particular, the Bayesian approach treats the model as the outcome of a random experiment. The essential defining property of a Bayesian is to talk about the probability $P(H|E)$ of a hypothesis H, given evidence E. Whether one adheres to a Bayesian view, estimators that arise from the Bayesian approach have an attractive property is that the posterior pdf is at least as informative as the prior one. In this case, the likelihood function is called *diffusive* or *totally noninformative*, and the prior estimates are exactly equal to the posterior estimates. It is emphasized that the method does not always guarantee better estimates for a couple of reasons. First, the Bayesian approach provides a pdf, not a single valued estimate. Second, the improvement achieved in the posterior pdf is dictated by the quality of external factors such as the accuracy of the likelihood function. Bayesian inversion is illustrated in the next Section "Joint geophysical-hydrogeological stochastic methods for subsurface characterization" particularly for combining geophysical survey and hydrogeological data for subsurface characterization.

JOINT GEOPHYSICAL-HYDROGEOLOGICAL STOCHASTIC METHODS FOR SUBSURFACE CHARACTERIZATION

Motivation

Combining ground-surface or cross-well geophysical surveys with well logs for enhancing the quality of subsurface characterization has been the goal of recent studies. The primary motivation has been the recognition that geophysical surveys offer unique opportunities for enhancing cross-well interpolation and are particularly promising in situations of data scarcity. Incorporation of two- and three-dimensional densely sampled geophysical data with conventional hydrological data increases the amount of data available for the characterization and thus has the potential to significantly improve the hydraulic parameter estimates over those obtained from borehole data alone.

The key problems are two: the nonuniqueness of the relationships between the hydrogeologic and geophysical rock/soil properties, and the absence of universal rock physics relationships that link the geophysical observable (i.e. electrical conductivity) to the hydrogeologic parameters (e.g. permeability, porosity). While these relations are sometimes difficult to obtain, it is critically important to recognize that even weak correlations can lead to a measurable improvement in estimation of the hydrogeologic variables. Integration of multisources of data is case and site-specific; however, the general framework is similar. The purpose of this section is to address the problem of joint hydrogeologic-geophysical site characterization following the study conducted by the author on Lawrence Livermore National Laboratory (LLNL) site data.

A few observations based on these studies are as follows: (i) No universal methods or petrophysical models are available for converting geophysical attributes to hydrogeological ones. (ii) The most challenging problem is tying well-logging measurements to the geophysical surveys. This issue involves problems of scale disparity and inconsistencies in the methods of data acquisition and interpretation. The last problem was studied extensively by Ezzedine *et al.* (1999) and can be demonstrated by the fact that resistivity at the LLNL site was measured along boreholes using several different tools, each characterized by a different support volume, sometimes leading to dramatically different results. Geophysical characterization is covered in **Chapter 148, Aquifer Characterization by Geophysical Methods, Volume 4**.

The main challenge that Ezzedine *et al.* (1999) had to face was in creating the bridge to link between ambiguously related geophysical surveys and well data. The second challenge was imposed by the disparity between the scale of the geophysical survey and the scale of the well logs. Ezzedine *et al.* (1999) approach is hierarchal and is intended to integrate and transform the well log data to a form where it can be updated by the geophysical survey, and this tends to be a convoluted process. In an ideal situation the geophysically measured attributes correlate well with the hydrogeological ones, for example, permeability, and the conversion of the geophysical survey to a hydrogeological distribution map is straightforward. In more realistic situations, such as the one described here, the conversion of the geophysical attributes to the hydrogeological ones is convoluted and nonunique. The difficulties faced in the implementation of the geophysical survey are several: the survey resistivity is expected to be of a relatively low resolution. At the same time, high-resolution permeability images are needed for flow and transport simulations. Thus the Bayesian framework has been chosen, which allows bridging between measurements of different resolutions.

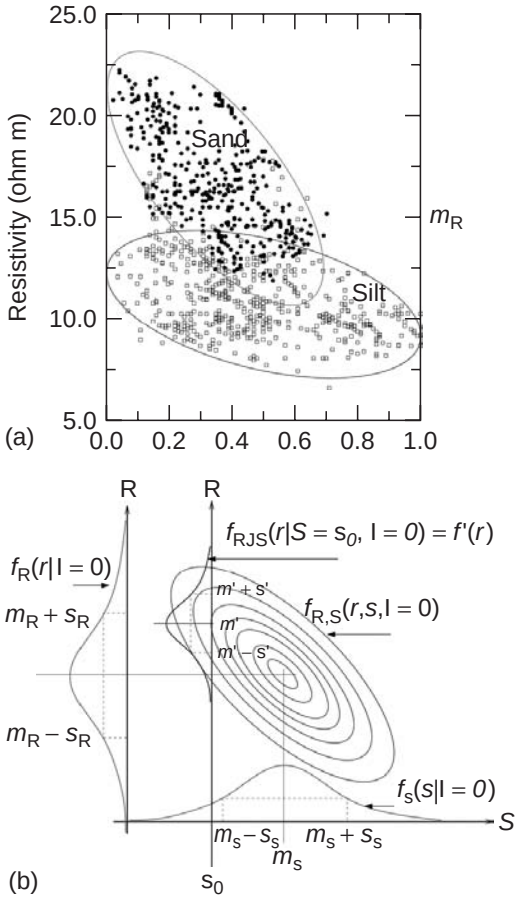


Figure 17 Shaliness versus resistivity cross-plot. (a) Two main clusters are identified: Silt and Sand. (b) Simplified pdf shown conditional and prior resistivity pdfs (Reproduced from Ezzedine *et al.,* 1999 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Bayesian Data Assimilation

Bayesian data assimilation is developed following a data-driven approach for lithology mapping based on the well log data. The proposed approach is general in its basic principles but at the same time is site specific since the petrophysical models employed are not universal. The general approach is stochastic. The choice is justified given the large uncertainty associated with cross-well interpolation, with the petrophysical models and with the interpretation of the geophysical surveys. The rationale for the approach is based on the following observations:

1. Resistivity and shaliness can be used for lithology identification through the cross-plot (Figure 17). Once a type of lithology is determined, further mapping of hydrogeological properties can be pursued.
2. Facies identification based on the shaliness-resistivity cross-plot is nonunique owing to some overlap between the sand and silt clusters.
3. Borehole resistivity measurements display short correlation range, and it is impractical to develop spatial images of the resistivity using cross-well geostatistical interpolation.
4. Shaliness displays a well-defined spatial correlation structure. It can be used for projecting resistivity measurements indirectly through a combination of geostatistical interpolation-simulation techniques, in conjunction with the nonlinear correlation structure it displays with the resistivity, as expressed through the cross-plot (Figure 17).

On the basis of these observations, the proposed approach consists of sequentially generating a series of collocated

attributes. At the basis of the hierarchy, images of the lithology are generated, conditional to well logs and possibly also to the survey resistivity. Each lithology image then serves as the basis for generating a series of shaliness images, again conditional to well data. The shaliness images are then used to correlate the survey resistivity with the hydrogeological attributes obtained experimentally. The series of generated images all have in common the well data and the same underlying spatial structure, and hence they are all physically plausible. The variations between the images constitute a measure of the spatial variability and estimation uncertainty. The focus is on estimating resistivity, but it can be converted to porosity and conductivity through well-known petrophysical models.

Bayesian Data Integration

The data integration encompasses the following steps:

Step 1: Generation of the lithology images using sequential indicator simulation (SISIM).

The lithology is defined through an indicator variable I according to: $i = 1$ if $\mathbf{x}$ is located in a silt body, 0 otherwise. Note that boldface letters denote vectors, that is, $\mathbf{x}$ is the location coordinates vector. Lowercase i is a realization of the spatial random function (SRF) I. I is characterized through its expected value conditional to the borehole data, $p^c = \mathbf{E}^c\{I\} = \mathbf{E}\{I|\text{measurements}\}$, with a superscript "c" denoting conditional. Since I is binary, p^c is statistically exhaustive. Its spatial variability is defined through the semivariogram and is shown in Figure 18(a,b). These statistics are the cornerstone of the SIS algorithm (Deutsch and Journel, 1997) adopted here.

Step 2: Generation of shaliness images.

This step is similar in principle to the previous one. The differences are in the fact that (i) the shaliness S is not a binary variable and (ii) the pattern of spatial variability of the shaliness may be different between the sand and silt lithologies, that is, $\gamma_S|i$, the semivariogram of the shaliness S, depends on the lithology $i = 0$ or 1. SGS algorithm (Deutsch and Journel, 1997) is adopted here to generate shaliness images. Shaliness S is defined by its mean $m_{S|i}$, its semivariogram $\gamma_{S|i}$ and its covariance, $\mathbf{Cov}_{S|i}$, for a given facies i (Figure 19a,b).

Step 3: Computing the resistivity prior pdf.

Once $\mathbf{x}$ is identified as being either sand or silt and is assigned a shaliness value, a prior pdf for the resistivity $f_{R(x)}(r|I = i, S = s)$ can be defined through Figure 17(a). R and S denote the SRF of the resistivity and the shaliness, respectively, and r and s denote their realizations. Figure 17(b) illustrates the joint pdf of R and S given $I = 0$ (i.e. sand lithology) and the marginals $f_R(r|I = 0)$ and $f_S(s|I = 0)$. Conditioning further on $S = s_0$ leads to $f_{R|S}(r|S = s_0, I = 0)$, which is the Bayesian prior. Scarcity of data leads to condition on ranges of S values rather than on single values. These pdfs are the Bayesian prior pdfs of the resistivity, and hence the stochastic estimation for the resistivity R at $\mathbf{x}$ in case no additional data become available through surveying.

Step 4: Updating $f_{R(x)}(r|I = i, S = s)$ based on cross-well electromagnetic resistivity survey $\rho(\mathbf{x})$.

Defining $f_{R(x)}(r|I = i, S = s) = f'_{R(x)}(r)$ for brevity, and given a collocated survey resistivity $\rho(x)$, the posterior pdf $f''_{R(x)}(r|\rho)$ can be defined through Bayes' rule:

$$f'_{R(x)}(r|\rho) = C_R L(\rho|r) f'_{R(x)}(r), \quad \text{and}$$

$$C_R = \left(\int_{-\infty}^{+\infty} L(\rho|r) f'_{R(x)}(r) dr \right)^{-1} \qquad (71)$$

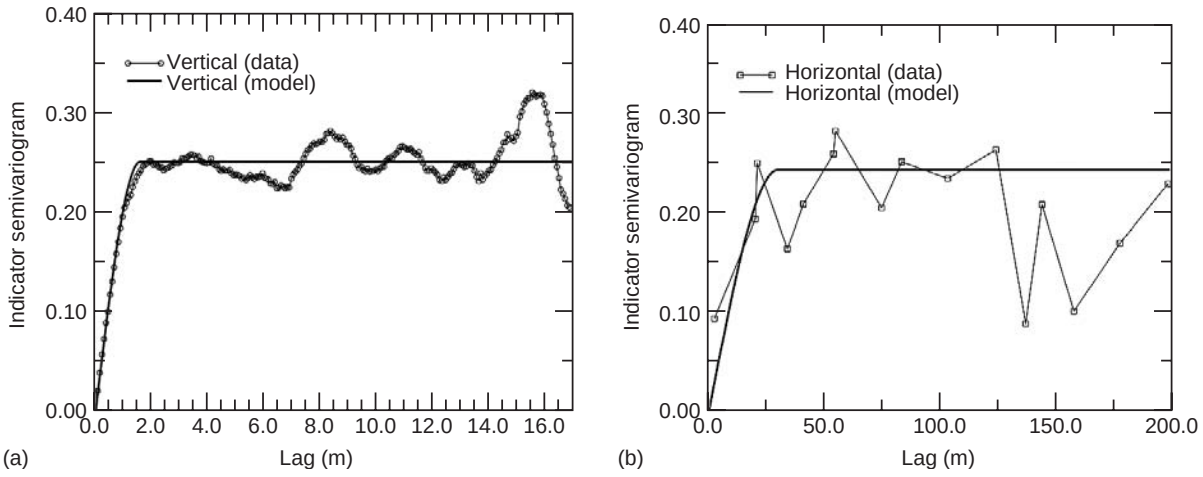


Figure 18 Experimental and theoretical indicator variograms: (a) vertical and (b) horizontal variograms (Reproduced from Ezzedine *et al.*, 1999 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

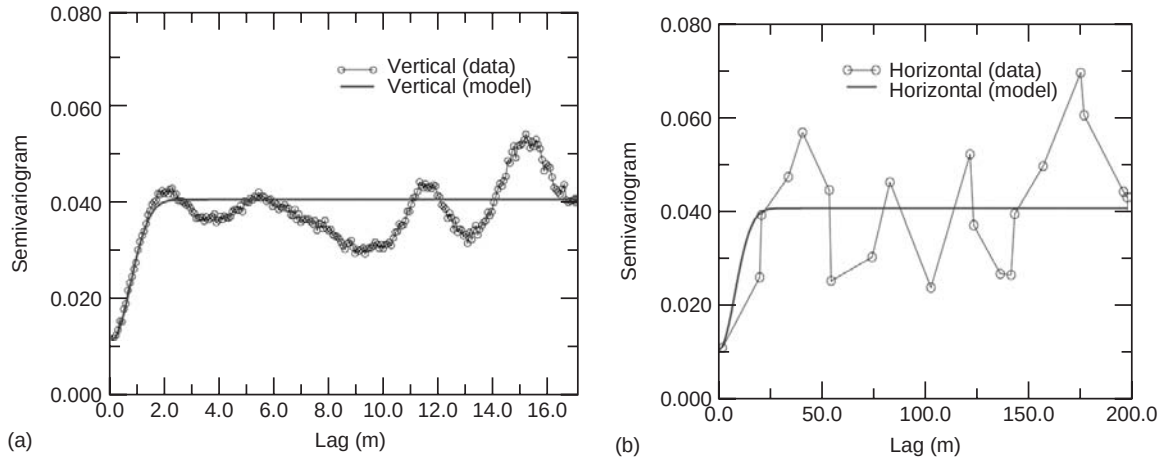


Figure 19 Experimental and theoretical variograms of shaliness: (a) vertical and (b) horizontal variograms (Reproduced from Ezzedine *et al.,* 1999 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

where $L(\rho|r)$ is the likelihood function, and C_R is a normalized factor. In general, ρ is defined over a support volume larger than the support volume of r. In the case of a high-resolution geophysical survey $\rho(\mathbf{x}) \rightarrow r(\mathbf{x})$ and Bayesian updating is unnecessary. This, however, is not generally the case and the alternative is to update $f'_{R(\mathbf{x})}(r)$ given ρ. Typically we are interested in R representative of a block of scale $\sim$1 m while ρ is defined by blocks of scale $\sim$3 m or greater. The inference of the likelihood function, $L(\rho|r)$, is critical for the successes of the updating process. Once $f''_R(r|\rho)$ is defined, a realization of R at $\mathbf{x}$ can be drawn. The whole process is repeated for all $\mathbf{x}$ until a complete image of the resistivity field is completed. Similarly, the lithology images can be improved through the resistivity survey despite the nonlinear and nonunique relationship displayed in the cross-plot. The approach calls also for Bayesian updating of p^c as well, through the relationship: $p^{c'} = C_I L(\rho|I) p^c$, where $L(\rho|I)$ is the likelihood function, of a similar nature to $L(\rho|r)$, only relating ρ to I rather than R. C_I is a normalized factor similar to C_R.

Step 5. Measure the effectiveness of the Bayesian updating.

To evaluate the effectiveness of the updating procedure, the following statistic were analyzed:

$$\Re_k = \frac{|r_k - m''|}{|r_k - m'|}; \quad \begin{cases} \Re_k \langle 1, & \text{Successful updating} \\ \Re_k = 1, & \text{Unsuccessful} \end{cases} \tag{72}$$

where k is a running index over all the points outside the wells, r is the actual resistivity, m'' is the mean of the posterior pdf $f''_{R(\mathbf{x})}(\mathbf{x})$, and m' the mean of the prior pdf $f'_R(\mathbf{x})(\mathbf{x})$. The ratio $\Re$ compares the performance of the posterior and the prior pdfs. $\Re$ smaller than 1 indicates a successful updating procedure. $\Re = 1$ is a diffuse likelihood and hence a noninformative survey. Figure 20 depicts the variation of $\Re$, as a function of the resolution of the survey. For completeness, statistics were also computed for resistivity surveys of (2×2) and (12×12) block resolution. As expected, Figure 20 shows that $\Re$ decreases with decrease in resolution.

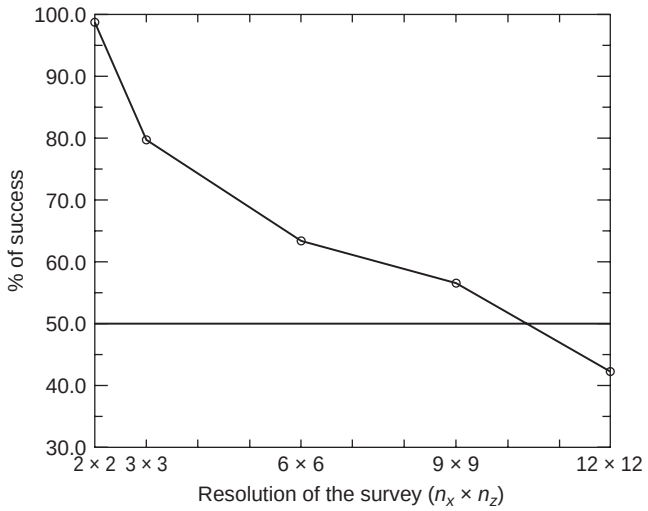


Figure 20 Effectiveness of the Bayesian updating (Reproduced from Ezzedine *et al.,* 1999 by permission of American Geophysical Union). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

THE USE OF STOCHASTIC CONCEPTS IN MODELING FRACTURED MEDIA

Introduction

Most fractured rock systems consist of rock blocks bounded by discrete discontinuities comprised of fractures, joints,

and shear zones, usually occurring in sets with similar geometries. Fractures may be open, mineral-filled, deformed, or any combination thereof. Open fractures may provide conduits for the movement of groundwater and contaminants through an otherwise relatively impermeable rock mass. Major factors affecting groundwater flow through fractured rock include fracture density, orientation, effective aperture width, and the nature of the rock matrix. Fracture density and orientation are important determinants of the degree of interconnection of fracture sets, which is a critical feature contributing to the hydraulic conductivity of a fractured rock system. Only interconnected fractures provide pathways for groundwater flow and contaminant transport. Fractures oriented parallel to the hydraulic gradient are more likely to provide effective pathways than fractures oriented perpendicular to the hydraulic gradient. The cross-sectional area of a fracture will have an important effect on flow through the fracture. Fracture-flux is proportional to the cube of the fracture aperture, this relationship is valid only for fractures with apertures greater than tens of microns (Bear *et al.*, 1993). Fracture apertures, and therefore flow through fractures, are highly stress-dependant, and generally decrease with depth. In recent years, a number of models have been developed to represent fracture flow. We introduce in the next section the most common geometric concepts of the models.

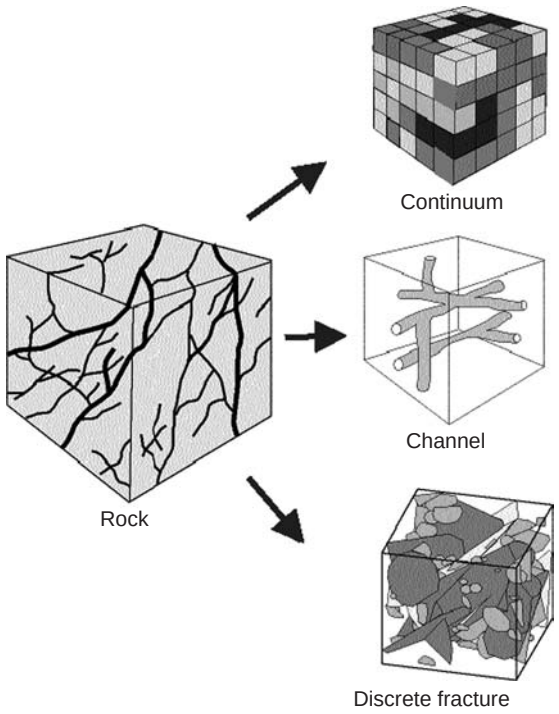


Figure 21 Different conceptual model geometry (Modified from Fractured Reservoir Discrete Feature Network Technologies, by Dershowitz *et al.,* 1998. **www.golder assosiates.com**). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Conceptual Model Geometry

Several different conceptual models have been used to describe flow and transport in fractured media. The most common conceptual picture of flow and contaminant transport in a fractured porous medium is that the advective flow of water and transport of pollutants is largely, or entirely, through the fractures. Water and contaminants may diffuse into and out of the porous rock matrix. This diffusion can act to spread out the contaminant plume in space and time, and to retard it.

In situations where transient water flow is involved, water may also be stored in, and released from, the rock matrix and the dead zones. To the extent that there is sufficient primary porosity in the matrix to allow advective flow and transport, as might be the case for a sandstone, this basic conceptual picture will be in error, as will any model that is based on it. If the rock matrix has very low porosity, such as would be the case for granite, then the role of the rock matrix can often be neglected.

Several different approaches, or concepts, have been used to describe the fractured mass (Figure 21). Models can be roughly classified as equivalent porous media models, multiple interacting continua (MINC) such dual porosity models, and stochastic fracture models. One could also develop models that overlap these categories.

The Equivalent Porous Medium

This approach treats the fractured rock system as if it were an unconsolidated porous medium. This approach is most likely to be successful when the spacing of the fractures is small compared to the scale of the system being studied, and the fractures are interconnected. The validity of using the Equivalent Porous Medium (EPM) approach to model pollutant transport in a fractured system is less well established. EPM could be extended to include models that assume an equivalent unique fracture intersecting the well, such fracture could be horizontal, vertical, and so on. This approach, generally used in the petroleum industry, only gives the local properties of a major fracture and does not necessarily assume that an EPM exists.

For the sake of simplicity, let us assume for a moment that fractures in a three-dimensional space are planes of finite extent, with the shape of irregular polygons or ellipses. If the average size of each fracture is small, and if its density is very low, a situation may arise in which none of the fractures intersects another. In this case, the global permeability of the system would be zero, since water cannot flow from one fracture to another. Whatever the size of the Representative Elementary Volume (REV), there is no EPM. A concept other than just size is thus required for the existence of an EPM; this other topological

characteristic is called the "*connectivity*" of the network and has been thoroughly studied in other areas of physics under the name of "percolation theory". An important concept in percolation theory is that of the "percolation threshold", defining a density of fractures above which the connectivity of the fractures is sufficient for flow to take place through a network, even an infinite one. Below this threshold, a few fractures may be connected, forming a "finite cluster" where flow can take place, but if the size of the domain of interest is increased, the system is globally unconnected, and only local pervious clusters exist.

At present, the best criterion for determining if a fracture network is above or below the percolation threshold is probably the one proposed by Charlaix (1984); the threshold is given by $nr^3 = 0.15$ to 0.30; n is the density of fractures (number of fractures per unit volume of rock); r^3 is the cube of the constant radius of the fractures, assuming that the shape of a fracture is a circular disc. This expression assumes that the orientation of the fractures is random. If fractures are not discs, r^3 should be taken as the average area of the fractures multiplied by their average half-perimeter divided by two.

Multiple Interacting Continua (MINC)

A modification of the EPM is to model the system as if it were composed of two overlapping continua with different porosities and permeabilities (Barenblatt *et al.*, 1960). Low porosity and high permeability are associated with the fractures and high porosity and low permeability are associated with the rock matrix. The model allows for the transfer of contaminants between the fractures and the rock matrix. This MINC approach requires that the fractures be closely spaced relative to the size of the system and that the fractures be frequently interconnected.

A simplified version of MINC is to represent the fractured system by a set of porous matrix blocks of well-defined geometry. The most common examples are parallel prismatic blocks (e.g. cubes, Figure 22) or spheres arranged in a regular array. The spaces between the blocks are the fracture channels. The blocks are assumed to be porous so that solutes can diffuse into and out of the matrix. This approach combines dual porosity with the discrete fracture approach. While no real aquifer has such a well-defined geometry, the model can provide insight into the important factors in solute transport in fractured porous media.

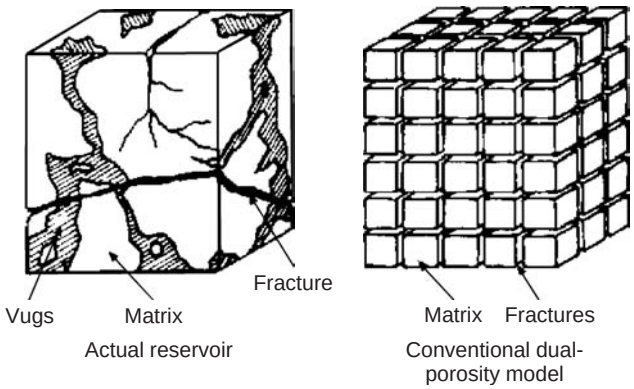


Figure 22 Double porosity model (Modified from Dershowitz *et al.*, 1998. **www.golderassosiates.com**)

Stochastic Discrete Fracture Networks (SDFN)

In the discrete fracture approach, the fracture geometry is explicitly included. Fractures are most often represented as channels with parallel sides, and the individual fractures are combined into fracture networks. The simplest network has a set of parallel fractures in what is basically a one-dimensional problem. A more complex network has two sets of parallel fractures oriented at some angle to each other in a two-dimensional array. A more complex, and one step closer to reality, is to allow the fractures to have varying lengths, locations, and orientations relative to one another (Figure 21).

One obvious problem in the practical application of discrete fracture models is that it is almost impossible to define the fracture system at a site in fine enough detail to apply the model. The best possibility for this approach seems to be through some sort of statistical modeling of the fracture system to duplicate the measured hydrology at the site. Most of the work on complex discrete fracture networks has been done in connection with the disposal of nuclear waste in crystalline rocks and has not included diffusion into the rock matrix. Real rock fractures may have rough surfaces (walls) that are not parallel to each other. The fracture may be partially filled with precipitated minerals or the walls could touch each other under mechanical stresses. In this case, it is better to describe the fracture system with flow through a series of tortuous intersecting channels.

When the SDFN may be simplified to single main channels connecting the center of the intersecting fractures and the center of their respective intersection areas, the SDFN is reduced to 3D network of 1D network of pipes or channels (Figure 21). Integrated hydraulic properties, for example, hydraulic conductivity, are then defined and the problem is solved through classical analogies to electric resistance network. This alternative is very useful when dealing with a large number of fractures and small computational capabilities.

None of the above conceptual pictures is "best" in an absolute sense. Rather, each may be appropriate for a particular situation. Models that are conceptually simpler have the advantage of being easier to implement as a rule, but they may also oversimplify the situation and miss important phenomena that are taking place. More complex models have the potential to provide a more detailed description of what is happening at the site being

modeled, but they are also likely to be more difficult to implement and may require data that cannot be collected with currently available techniques. Concepts introduced in the Section "Geostatistics" can be directly applied to the stochastic continuum approach where fractured mass is reduced to a Rubik's-cube-like equivalent heterogeneous "porous medium" where each block is defined by local EPM properties. Hereafter we focus only on SDFN models and their applications.

Stochastic Discrete Fracture Network Models

In this approach based on the model developed by Cacas (1989) and described in Ezzedine (1994), the fractures are treated as discs. The density and extensions are difficult to determine independently. There are relations between these magnitudes that can be used to test the connectivity of the medium. The hydraulic model is a bond model. The flow is not two-dimensional in the fracture plane but occurs in channels linking the centers of the connected discs. The walls are impervious. An additional parameter called *aperture* or "hydraulic thickness" is attributed to each disc in the model.

Characteristics of the Fracture Families

In natural fractured reservoirs, fractures can usually be grouped into families of directions using Schmidt or Rose diagram (Figure 23). For a single realization, a large number of fractures are drawn sequentially for one family of directions at a time. The families are then combined to

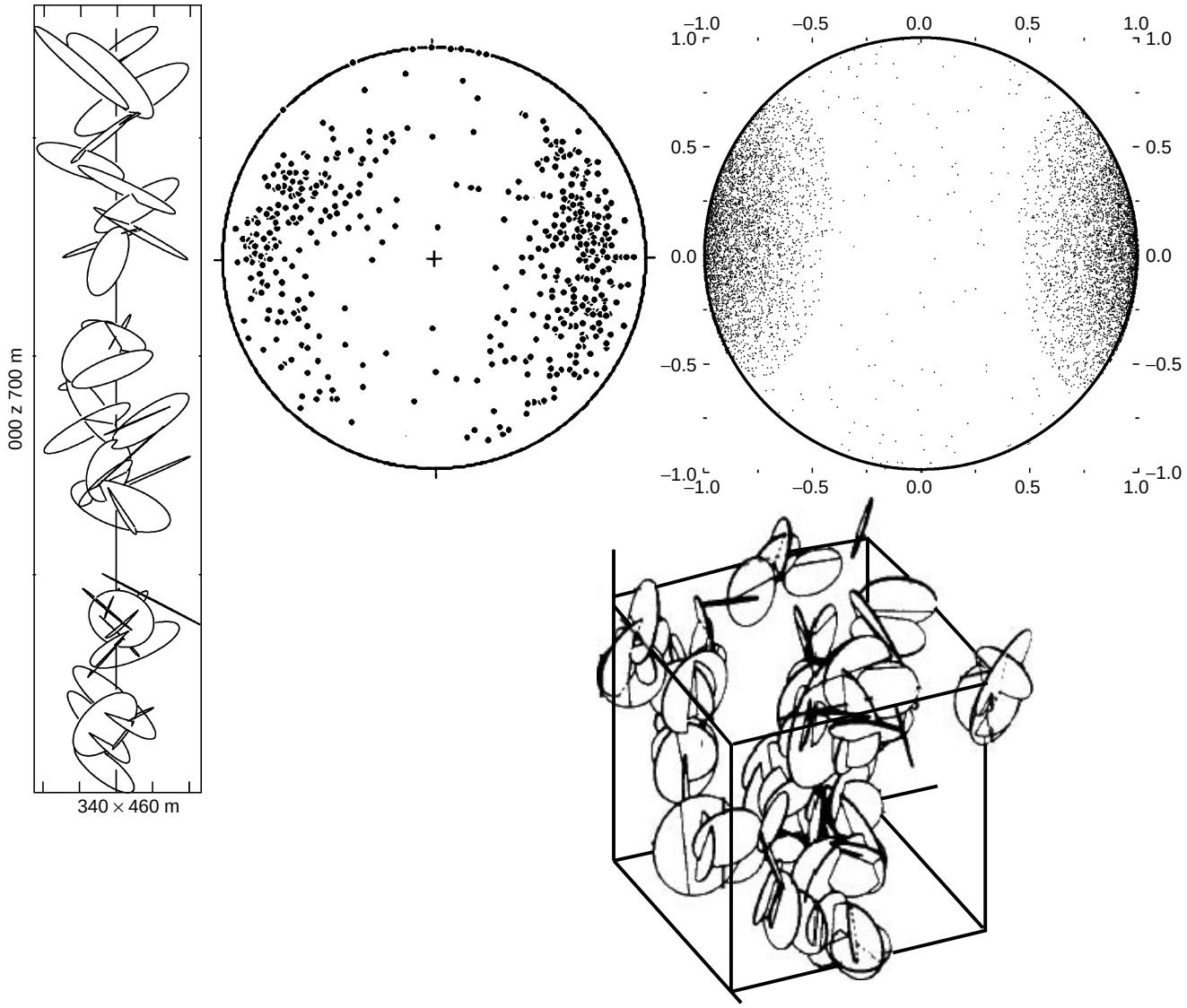


Figure 23 Stochastic discrete fracture network. Observed fracture in a borehole, observed fracture families (Rose diagram), simulated fracture families, and a 3D SDFN model (Reproduced with permission of Ezzedine, 1994)

Table 2 Example of probability density functions and their statistical moments

Law	Density f(x)	Mean E(X)	Variance V(X)
Log-normal	$\frac{1}{\sigma x\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right]$	$\exp\left[\mu+\frac{1}{2}\sigma^2\right]$	$\exp[2\mu+\sigma^2]\exp[\sigma^2-1]$
Exponential	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$e^{-\lambda x}\frac{(\lambda x)^{k-1}}{(k-1)!}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$

create the final network. For a given family, the draw is made as follows:

Number of Fractures A Poisson process is used to obtain the number of discs belonging to family i in the simulated volume $V(\mathrm{m}^3)$ as a function of the volumetric density λ_i (number of discs/m^3). The following expression gives the probability of having an N_i number of centers in the volume V.

$$p(N_i = k) = e^{-\lambda_i V}\frac{(\lambda_i V)^k}{k!} \tag{73}$$

Location of the Disc Centers The centers of the discs are independently distributed. The drawing of the Cartesian coordinates of the centers is done in a domain called the "*generation domain*". The (x_i, y_i, z_i) are coordinates of the centers are drawn according to a uniform law in the intervals $[x_{\min}, x_{\max}]$, $[y_{\min}, y_{\max}]$ and $[z_{\min}, z_{\max}]$.

Direction The direction of a fracture is defined by its normal unit vector **N**. The families of directions in a network consist of fractures with similar directions. Their normal vectors form an approximately cone-shaped segment, the axis of which is the "pole" of the family. Fisher's pdf is adopted and expressed as follows:

$$f_\kappa(\alpha) = C(\kappa)e^{\kappa\cos\alpha}\sin\alpha \quad \text{with} \quad C(\kappa) = \frac{\kappa}{4\pi\sinh(\kappa)} \tag{74}$$

where f represents the pdf of the angle α formed with the pole of the family (the pole is here the mean normal vector of the family); α is the angle between the fracture pole and the pole of the family; κ is the parameter of the law. For $\kappa \to +\infty$ the distribution is very concentrated around the mean direction. For $\kappa \to 0$, this distribution is closer to the uniform distribution.

Radius The radius is the parameter describing the extension of the fracture. Several statistical laws are available and given in Table 2 for completeness. Recently, de Dreuzy *et al.* (2001) suggested random fracture networks following a power-law length distribution.

Connectivity

The analysis of the connectivity and the search for continuous paths within such fracture networks use algorithms shared with the theory of graphs. It is, in fact, easy to ascertain whether two fractures are interconnected or not, but in a network of fractures, it becomes very costly to examine all the fracture couples. Therefore, the flow region is cut up into a certain number of blocks. A given disc intersects one or several of these elementary volumes. Two discs will not cut across each other unless both of them together cross at least one of the blocks. It is therefore necessary to test a possible intersection of two discs only if they are geographically fairly close to each other. This method makes it possible to limit the number of tests. Then, in 3D, the number of fracture couples is reduced from $1/2N(N-1)$ to $1/2N^2h^3k^2/R^3$ if the flow region is divided into cubes of volume h^3 and a fracture crosses k-cubes. Thus, the number of tests is reduced by $\sim h^3k^2/R^3$ provided that h is chosen as small as possible and k is of the order of 2 or 3 for one fracture.

Channeling, Integrated Hydraulic Conductivity

The "channeling hypothesis" is taken into account by defining permeability, said to be "integrated", into which the effective flow area is introduced. Once the interconnected geometry of the formation has been constructed, one defines the links connecting the centers of the two fractures that pass through the middle of the intersection segment. Thus, the geometry is reduced to a group of one-dimensional elements placed end to end in a three-dimensional domain

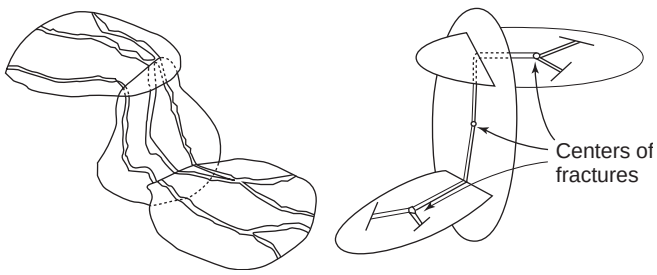


Figure 24 Channeling, ''true'' flow through fractures and simulated flow through pipe network (from Tsang *et al.*, 1991; Cacas *et al.*, 1989)

(Figure 24). Between two intersecting fractures, water circulates in a tube linking their centers and passing through the intersection segment. Assuming a fixed, straight, hollow cylinder of length l, large compared to the diameter $2R$, the volumetric flow in the tube is proportional to the fall of the pressure and to the fourth magnitude of the radius:

$$Q = \frac{\pi}{(8\mu)} R^4 \nabla h \tag{75}$$

μ is the kinematic viscosity and ∇h the hydraulic gradient between the ends of the tube. The "integrated hydraulic conductivity" is then defined as the ratio between the flow rate and the head gradient: $K = \pi/(8\mu)R^4$. The integrated hydraulic conductivity is the product of the "real" hydraulic conductivity (expressed in $\mathrm{m\,s^{-1}}$) multiplied by the cross-section of the flow, which is difficult to measure by hydrogeological tests. In this model, each disc is characterized by an integrated hydraulic conductivity, which is determined by the aperture (radius) according to a lognormal pdf. Effective integrated hydraulic conductivity between two intersecting fractures is estimated using a weighted harmonic average. The weights are the length of the two segments forming the pipe connection between the fractures (Figure 25). This magnitude represents the mean resistance, which prevents the water from circulating between two intersecting fractures. It integrates and lumps the whole flow geometry at the scale of the fracture: wall roughness, tortuosity, flow line lengths and channeling, and so on.

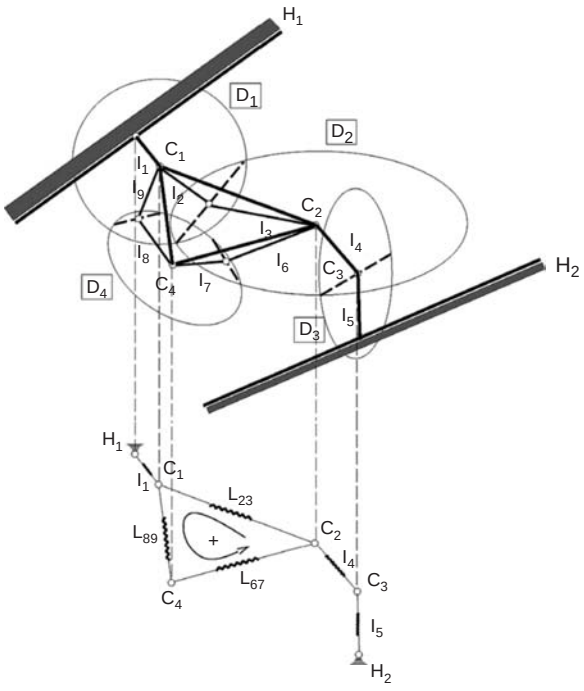


Figure 25 Analogy between flow in fracture network and electrical resistivity network (Reproduced with permission of Ezzedine, 1994). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Flow in Fractured Network

Similar to Kirchoffs law in electricity (Figure 25), when the flow is laminar and steady state in each fracture flowing following Poiseuille's law, at each fracture intersection the continuity of mass balance equation holds.

Consider a branch of a flow network (Figure 26) limited by two nodes i, and j, of length L_{ij} where the flow is governed by Darcy's law. This flow is characterized by a spatial variation of the conductivity k_{ij}. x is the distance between the observed section and the node i, measured along the branch. Because of the Darcy equation, the flow toward node i in branch ij is written:

$$q_{ij} = -k_{ij}(x) \left.\frac{\mathrm{d}h_{ij}}{\mathrm{d}x}\right|_{x=0} \quad \text{and} \quad q_{ji} = -k_{ji}(x) \left.\frac{\mathrm{d}h_{ji}}{\mathrm{d}x}\right|_{x=L_{ij}} \tag{76}$$

Assuming that the nodes have no capacity, the mass balance equation at each node i is $\sum_{j \neq i} q_{ij} = Q_i$, which is a system of linear equations in h_{ij} that can be solved readily using classic numerical techniques.

Modeling the flow through each fracture by a single conduit is restrictive and in some cases does not reflect the more complex reality, especially when the fracture network encompasses only a few fractures. To alleviate these restrictions, Ezzedine (1994) introduced "daughter fractures" that are imbedded into the "mother fractures" in order to create a 2D network of channels of interconnected 1D pipes. The introduction of a random perturbation to the normal vector of "daughter fractures" leads to different intersecting flow channels within the "mother fractures" (Figure 27). This technique allows enhancing the flow within each mother fracture from a 1D flow into 2D flow, which is more representative of the reality.

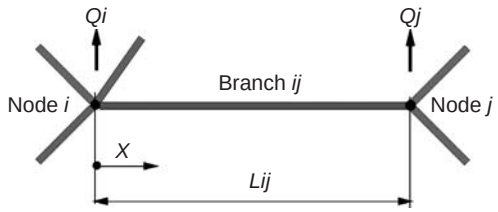


Figure 26 Single branch (pipe) representing a single fracture (Reproduced with permission of Ezzedine, 1994). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

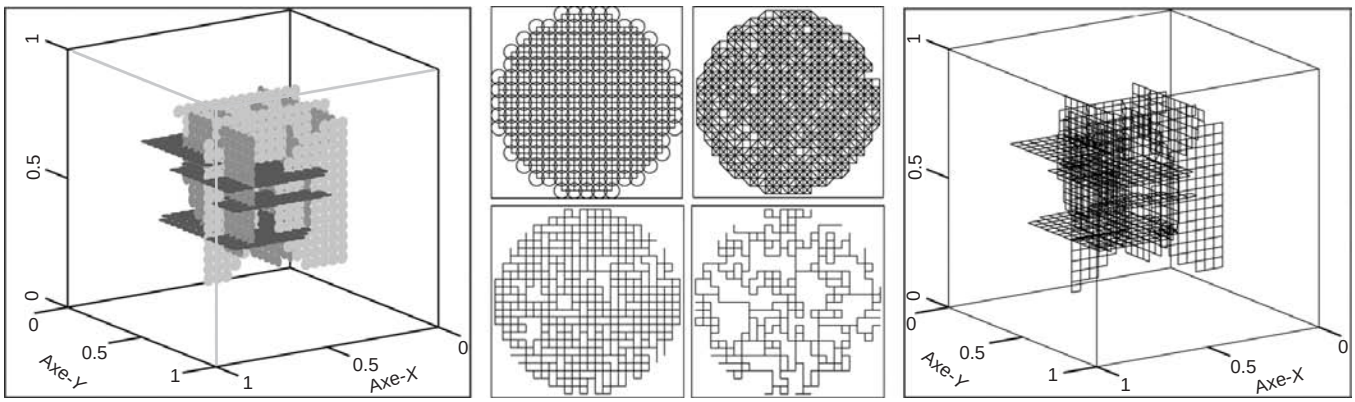


Figure 27 Simple 3D fracture network composed of 4 fractures in each direction (x,y,z). ''Mother'' fractures are mapped using ''Daughter'' fractures, four different cases are shown where the percolation network is depicted. The 3D fracture network is then reduced to a skeleton of 1D connected pipe network system (Reproduced with permission of Ezzedine, 1994)

Unsteady Flow and Transport in Fractured Network

In many applications, however, the assumption of steady flow is restrictive. Therefore, Ezzedine (1994) extended the model to simulate unsteady state flow. The flow at the fracture scale is derived analytically using Laplace transform. Flow at the formation level is then achieved by combining the solution of the generated individual fractures. A real flow injection test, performed at the Hot Dry Rock (HDR) site Soultz-sous-Forêts in Alsace/France, is simulated. Results show that the mechanical properties of the media influenced by the flow injection must be incorporated into the proposed model. In particular, changes in the hydraulic conductivity and the storage coefficient function of the effective stress appear to be quite significant. Consequently, the model is adjusted accordingly to include the mechanical effects that resulted in significant improvements. Heat extraction from Hot Dry Rock (HDR) is not limited to the hydrological phenomena discussed thus far but necessitates that thermal and geochemical processes must be addressed. Similar to fluid flow, a one-dimensional geochemical model was developed that takes into account the equilibrium and/or nonequilibrium transport of the chemical species and the water rock interaction. Thermal processes are modeled using a double porosity approach. The hydrothermal and chemical transport at the formation level is obtained by simultaneously combining the solution of the stochastically generated individual mother and daughter fractures.

Acknowledgments

The author would like to thank Ghislain de Marsily for his review. His comments and suggestions have improved the content and the clarity of this chapter.

REFERENCES

Aris R. (1956) On the dispersion of a solute in a fluid flowing through a tube. *Proceedings of the Royal Society of London Series B Biological Sciences*, **235**, 67–76.

Barenblatt G., Zheltov I.P. and Kochina I.N. (1960) Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. *Soviet Applied Mathematics And Mechanics (English Translation)*, **24**(5), 852–864.

Bear J., Tsang C.F. and deMarsily G. (1993) *Flow and Contaminant Transport in Fractured Rock*, Academic press: p. 560.

Cacas M-C.h (1989) *Développement d'un Modèle Tridimensionnel Stochastique Discret pour la Simulation de L'écoulement et des Transferts de Masse et de Chaleur en Milieu Fracturés*, PhD Ecole des Mines de Paris, p. 281.

Cacas M.-C. h, Ledoux E., de Marsily G., Tillie B., Barbreau A., Durand E., Feuga B. and Peaudecerf P. (1989) Modeling fracture flow with a stochastic discrete fracture network: calibration and validation - 1. *Water Resources Research*, **26**(3), 479–489.

Charlaix E. (1984) A criterion for percolation threshold in a random array of plates. *Solid State Communications*, **50**(11), 999–1002.

Chiles J.P. and Delfiner P. (1999) *Geostatistics, Modeling Spatial Uncertainty*, Wiley-Interscience: p. 695.

Christakos G. (1992) *Random Field Models in Earth Sciences*, Academic Press: p. 474.

Cressie N. (1993) *Statistics for Spatial Data*, Wiley-Interscience: p. 900.

Cvetkovic V., Shapiro A.M. and Dagan G. (1992) A solute flux approach to transport in heterogeneous formations. 2: uncertainty analysis. *Water Resources Research*, **28**(5), 1377–1388.

Dagan G. (1989) *Flow and Transport in Porous Formations*, Springer-Verlag. p. 465.

Dagan G. (1988) Time-dependent macrodispersion for solute transport in anisotropic heterogeneous aquifers. Water Resources Research, **24**, 1491–1500.

de Dreuzy J.-R., Davy P. and Bour O. (2001) Hydraulic properties of two-dimensional random fracture networks following a power law length distribution 2. Permeability of networks based on lognormal distribution of apertures. *Water Resources Research*, **37**(8), 2079–2095.

de Marsily G. (1986) *Quantitative Hydrogeology*, Academic Press: p. 440.

Dershowitz W.S., Foxford T., Doe T. (1998) Golder Associates Inc., Redmond: Washington, March 9, 963–1357.211.

Deutsch C.V., Journel A.G. (1997) *Gslib: Geostatistical Software Library and User's Guide*, Oxford University Press: 369.

Ezzedine S. (1994) *Modélisation des écoulements et Transports dans les Milieux Fractures*, PhD Ecole des Mines de Paris, p. 198.

Ezzedine S. and Rubin Y. (1996) A geostatistical approach to the conditional simulation of spatially distributed solute concentration and notes on the use of tracer data in the inverse problem. *Water Resources Research*, **32**(4), 853–862.

Ezzedine S., Rubin Y. and Chen J. (1999) Bayesian method for hydrogeological site characterization using borehole and geophysical survey data. *Water Resources Research*, **35**(9), 2671–2683.

Fiori A. (1996) Finite Peclet extensions of Dagan's solutions to transport in anisotropic heterogeneous formations. *Water Resources Research*, **32**(1), 193–198.

Garabedian S.P., Leblanc D.R., Gelhar L. and Celia M.A. (1991) Large scale natural gradient tracer test in sand and gravel, Cape Cod. *Water Resources Research*, **27**(5), 911–924.

Gelhar L. (1992) *Stochastic Subsurface Hydrology*, Prentice Hall: p. 390.

Gelhar L. and Axness C.L. (1983) Three dimensional stochastic analysis of macrodispersion in aquifers. *Water Resources Research*, **19**(1), 161–180.

Hsu K.-C, Zhang D. and Neuman S.P. (1996) Higher-order effects on flow and transport in randomly heterogeneous porous media. *Water Resources Research* **32**(3), 571–582.

Isaaks E.H. and Srivastava R.M. (1989) *An Introduction to Applied Geostatistics*, Oxford University press: p. 561.

Journel A.G. and Huijbregts J.C.h (1978) *Mining Geostatistics*, Academic press: p. 600.

Kitanidis P. (1997) *Introduction to Geostatistics*, Cambridge press: p. 249.

Leblanc D.R., Garabedian S.P. and Hess K.M. (1991) Large scale natural gradient tracer test in sand and gravel, Cape Cod, 1. experimental design and observed tracer movement. *Water Resources Research*, **27**(5), 779–910.

Matheron G. (1965) *Theory of Regionalized Variables*, Vols. 1 & 2, Masson & Cie.

Matheron G. and de Marsily G. (1980) is transport in porous media always diffusive? A counterexample. *Water Resources Research*, **16**(5), 901–917.

Rubin Y. and Ezzedine S. (1997) The travel times of solutes at the Cape-Cod tracer experiments. *Water Resources Research*, **28**(4), 1033–1547.

Rubin Y. and Dagan G. (1987) Stochastic identification of transmissivity and effective recharge in steady-groundwater flow. 2. Case Study. *Water Resources Research*, **23**(7), 1193–1200.

Rubin Y and Dagan G. (1992) Conditional estimation of solute travel time in heterogeneous formation: Impact of transmissivity measurements. *Water Resources Research*, **28**(4), 1033–1040.

Rubin Y. (2003) *Applied Stochastic Subsurface Hydrology*, Oxford press. p. 368.

Schvidler M.I. (1962) Flow in heterogeneous media (in Russian). *Izvestiia Akademi Nauk SSSR*, **3**, 185–190.

Sun N.Z. (1994) *Inverse Problems in Groundwater Modeling*, Kluwer: p. 337.

Tarantola A. (1987) *Inverse Problem Theory*, Elsevier: p. 630.

Tsang C.F., Tsang Y.W. and Hale F.V. (1991) Tracer transport in fractures: analysis of field data based on a variable-aperture channel model. *Water Resources Research*, **27**(12), 3095–3106.

Taylor G.I. (1921) Diffusion by continuous movements. *Proceedings of the London Mathematical Society*, **20**, 196.

Zhang D. (2001) *Stochastic Methods for Flow in Porous Media: Coping with Uncertainties*, Academic press: p. 336.

155: Numerical Models of Groundwater Flow and Transport

EKKEHARD HOLZBECHER[1] AND SHAUL SOREK[2]

[1] *Humboldt Universität, Inst. of Freshwater Ecology (IGB), Berlin, Germany*
[2] *Ben Gurion University of the Negev, J. Blaustein Institutes for Desert Research, Sede Boker, Israel*

The article gives an introduction to numerical modeling of flow and transport problems and to software tools that are currently in use for modeling such phenomena. Details are explained on numerical approximations leading to different numerical models. Extensions for reactive transport are mentioned. Basic guidelines and criteria are given that should be taken into account by the modeler in order to improve the accuracy of results. Inverse modeling is presented as an advanced feature. Some examples of pre- and postprocessing, as implemented in several codes, are given, in addition to fundamental properties of solution methods. Finally, most common codes are listed with basic features and web-sites.

INTRODUCTION

The forecast of state variables that describe flow and solute transport in a given aquifer can be obtained by solving the *mathematical model* that describes these phenomena. Such a model is based on a *conceptual* model that includes a set of verbal statements introducing a simplified version of the various physical, chemical, and biological aspects of the flow domain and the phenomena of transport that take place in it. Because, in most cases of practical interest, analytical solutions of the mathematical models are not possible, the mathematical models are transformed into *numerical models*, which, in turn, are solved by specially designed computer programs (= codes).

The codes account for physical aspects (e.g. multiphase/multicomponent, density-driven, chemical reactions, inertial/drag dominant, energy considerations, small/large matrix deformation, Newtonian/nonNewtonian fluids), modeling aspects (e.g. domain dimensionality, type of boundary conditions, Eulerian/Lagrangian formulation, deterministic/stochastic representations, lumped parameter/continuum/sharp interface approaches, phreatic/confined, unsaturated/saturated), and optimal management (i.e. mathematical procedures deriving the optimal extremum trajectories under different constraints and objectives). These are addressed by sensitivity and inverse methods, analytical and/or numerical (e.g. differences, virtual work/variational approaches) approximations, grid methods (e.g. finite differences, finite elements, volume elements, boundary elements, spectral elements), numerical algorithms (Eulerian/Lagrangian coordinates, particle tracking, explicit/implicit approximations, linear/nonlinear iteration procedures) and their associated reliability/efficiency measures (e.g. stability/monotonicity of the numerical scheme, Peclet/Courant grid based numbers). Such codes, which are the focus of this chapter, are applied by the modeler to set up models (in the literature sometimes the term model is used also for a code, which is not correct). Mostly the codes are executable software files, mostly equipped with user-friendly graphical user interfaces (GUIs). Sometimes source-codes are distributed, which have to be altered, compiled, and linked by the user.

A modeling task can be subdivided into several steps:

- Preprocessing (transformation of data into a format appropriate for the numerical algorithm, including grid generation)
- Numerical calculation (direct modeling)
- Calibration (inverse modeling)
- Postprocessing

Today's software packages or codes assist in all of these modeling steps.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Calibration as a task cannot be separated from the other tasks. Inverse modeling includes direct model runs, performed in order to determine one or more parameter values, which lead to an optimal approximation of measured results. Some pre- and postprocessing may be necessary during the inverse modeling procedure, when its not the standard in- and output variables, on which the calibration is based.

Many different software tools are today available to help users to set up their models. The aim of the models is to assist in the solution of practical problems, simulating processes in subsurface fluids and porous media. In the majority of cases, modeling serves to improve understanding of hydrogeological systems. Forecasting and thus studying the response due to different scenarios is the most ambitious goal of modeling efforts. In scientific literature, this is discussed under the term *validation*

- Validation is a process carried out by comparison of model predictions with independent field/experimental observations. A model cannot be considered validated until sufficient testing has been performed to ensure an acceptable level of predictive accuracy (IAEA International Atomic Energy Agency, 1988)
- Validation: the process of obtaining assurance that a model as embodied in a computer program is a correct representation of the process or system for which it is intended. (United States NRC – Nuclear Regulatory Commission, see Silling, 1983)

The task of code verification is a step towards validation in which numerical results are compared with analytical solutions or with well accepted published results.

Software tools can be subdivided into different classes, for which codes that perform numerical calculations are considered as the core software program. Around these packages have been developed for several pre- and postprocessing tasks. GMS, Visual MODFLOW, and PMWIN are examples, which are build around the MODFLOW code in the core – in most recent versions accompanied by other numerical codes. Other packages, like FEFLOW, embed all tasks in one package.

In judging a numerical code for simulating flow and transport scenarios imposed to a specific aquifer site, one should first verify what aspects are being addressed by the code. This in general accounts for:

- the theoretical and mathematical assumptions and other considerations;
- the numerical method, algorithm, and grid configuration;
- verifications of the code against analytical and numerical solutions as well as laboratory and field observations;
- performance under a variety of time and space increments;
- platforms on which the code can be run.

Input data that is required for code simulations can be classified into:

1. Geometry and topography issues
   (a) Site boundaries and dimensions
   (b) Surface topography (e.g. to detect zones with surface infiltration)
   (c) Location of streams, divides, ponds and so on
   (d) Land use (landfills, dikes, well locations, irrigation systems. . . .)
2. Geology and hydrology issues
   (a) Aquifers (stratification, depth, lithologic parameters, hydraulic conductivities, longitudinal and transversal dispersivities, storativities (i.e. matrix and water compressibilities), porosities)
   (b) Porous medium density
   (c) Water levels at surface reservoirs (rivers, ponds, etc.) compressing shallow aquifers
   (d) Pumping/recharge point sources (well depth, intensity, periodicity, and time of application)
   (e) Distributed sources of inflow, for example, rainfall and irrigation rates
   (f) Distributed sources of outflow, for example, evapotranspiration
   (g) Time dependent data at spatial points
3. Water and porous medium chemical properties
   (a) Sorption (adsorption and desorption) factors.
   (b) Electrical conductivities.
   (c) Temporal and spatial concentration of solutes in the water and the solid phases of the porous medium
   (d) Solutes associated with sources of recharge fluxes (rainfall, irrigation, etc.).
   (e) Concentration of stable isotopes and microelements.
4. Boundary and initial conditions
   (a) Initial field distribution of piezometric head and components concentration.
   (b) Pervious/impervious boundary segments with the ascribed flux conditions.
   (c) Piezometric heads and concentrations along boundaries.

The user has to take into account that codes often differ concerning the input parameters, an aspect that may make some codes more appropriate for a given task than others. As an example many codes allow conductivity anisotropy only in direction the principal axes of the coordinate system. When anisotropies in changing directions have to be taken into account, the modeler has to choose a code that is capable of handling such a situation.

MATHEMATICAL MODELS

Groundwater Flow

Groundwater flow models are based on the differential equations for groundwater flow. Such differential equations, as described in **Chapter 149, Hydrodynamics of Groundwater, Volume 4**, are usually based on Darcy's Law as the linear macroscopic fluid momentum balance equation, considering the drag terms of the Navier Stokes equation as dominant, and on the principle of the fluid mass conservation.

Depending on the special features of the situation to be modeled, different circumstances have to be taken into account. A model for a confined aquifer is different from that for an unconfined (phreatic) aquifer. The spatial dimensionality (1D or 2D or 3D) depends on the physical situation and the aim of modeling. Depending on the very same aspects, a decision about steady state versus unsteady simulations has to be taken, just to name the most basic properties of a model.

There are different formulations of the differential equations. Equation (1) states the mass balance in 3D:

$$\frac{\partial}{\partial t}(\phi\rho_f) = -\nabla\cdot(\rho_f \mathbf{v}) + Q \tag{1}$$

where ϕ denotes porosity, ρ_f fluid density, $\mathbf{v}$ the three-dimensional vector of Darcy velocity, that is specific discharge, and Q represents mass sources or sinks of whatever type. Most models work with a simplified version of equation (1), which is valid for constant density ρ_f. With the help of Darcy's Law, the equation can be reformulated in terms of hydraulic head h. Simplified 2D versions of equation (1) are used quite frequently, which are different for confined or unconfined aquifers. In the confined situation:

$$S\frac{\partial h}{\partial t} = -\nabla\cdot T\nabla h + P - Q \tag{2}$$

in which P and Q represent pumping and recharge rates, respectively, where S denotes the storativity and T the transmissivity. Usually the hydraulic head, h is the dependent prime variable, for which the differential flow equation is formulated and which is calculated by the model. In 2D problems the stream function can be used as an alternative (Holzbecher, 1996). For applications involving variable density, a generalized hydraulic head or pressure p have to be introduced (Holzbecher, 1998) as dependent prime variable.

The codes allow the specification of different boundary conditions, which are relevant for groundwater flow (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**). When the first-type (or Dirichlet) condition is used, h is specified at that location. For the second-type (or Neumann) condition, the normal velocity has to be specified. The often-used no-flow condition is a special case (zero velocity). Third type (or Cauchy) boundary conditions can be used, when a relation between flux and head has to be considered, for example, when an aquifer is connected to a surface water body.

Hydraulic conductivity is an input parameter for most models. In the case of the 2D flow in a confined aquifer, transmissivity (integration of the conductivity over the third spatial direction) is required instead. Unsteady models require specific storativity for unconfined aquifers, and storativity as product of specific storativity and layer thickness for confined aquifers (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**). Porosity is needed when real interstitial velocities have to be determined, for example, when a transport model is to be set up in addition or when travel times along flowpaths are required.

Transport (Mass and Heat)

Transport models are derived from the transport equation, which is the mass balance equation in terms of the concentration of a certain substance in case of solute transport (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**) through the aquifer system. In case of heat transport, it is a differential equation for temperature as the dependent prime variable (Holzbecher, 1998).

The generic form of the solute transport equation in porous media reads

$$\frac{\partial}{\partial t}(\phi c) = -\nabla(\mathbf{v}c - \phi\mathbf{D}\nabla c) + q \tag{3}$$

with porosity ϕ, combined coefficient of diffusion and dispersion tensor $\mathbf{D}$, specific discharge $\mathbf{v}$ and source/sink-term q. Equation (3) is a balance equation for component mass, which is valid for constant density fluids (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**). The dependent prime variable is the concentration c. The term on the left side of the equation represents storage in general (gains or losses). The term $\mathbf{v}c$ on the right side represents advection and $\phi\mathbf{D}\nabla c$ is the sum of diffusion and dispersion. The last term is for sources and sinks of all types. Within the heat transport equation, the same terms can be found with different parameters:

$$\frac{\partial}{\partial t}T = \nabla D_{\text{therm}}\nabla T - \kappa\mathbf{v}\nabla T \tag{4}$$

with thermal diffusivity D_{therm} the ratio κ between heat capacity of the fluid and the heat capacity of fluid and porous medium. Internal sources and sinks are neglected in equation (4), which results from the energy balance

equation by dividing through the heat capacity of the porous medium. The dependent prime variable is the temperature T (Holzbecher, 1998).

The types of boundary conditions in general are the same as in flow models. In the first type, a concentration or a temperature needs to be specified. The second type accounts in general for advective and dispersive fluxes. Frequently used is the no dispersive flux condition, where it is required that the derivative of the dependent variable normal to the boundary is zero. The latter condition is not only used at impermeable boundaries. For lack of alternative, it is most often also used at outflow edges. Obviously the condition is not correct for fronts crossing the boundary, or vice versa: it is only applicable when during the simulation period the concentration or temperature gradients are small.

Reactive Transport

Taking into account that only few biogeochemical species in the subsurface are independent from their biogeochemical surrounding, models couple physical transport processes and biogeochemical reactions. The basic features concerning the coupling of transport and reactions are outlined in **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**.

When several species, which are linked to each other by reactions, are modeled, the transport equation has to be solved for each of these species. Two types of reactions have to be distinguished: equilibrium or kinetics. If for an application the timescale of interest exceeds the typical reaction time, the reaction is fast and it can be assumed that the equilibrium is reached, when the reaction is reversible. Equilibria are usually described by the *Law of Mass Action*. For the reaction $A + B \leftrightarrow C$ between species A, B, and C, the equilibrium is given by:

$$\frac{a_C}{a_A a_B} = K \tag{5}$$

where a denotes the activity of the species, and K is the reaction specific equilibrium constant. The activity of a species is the product of the concentration and an activity factor γ, which depends on the charge of the species and the ionic strength of the solution (Krauskopf and Bird, 1995).

When for an application the reaction time is smaller than the time of interest, an explicit formula for the development of the reaction rate in time has to be used. Such a reaction is termed *kinetic*. The difference concerning the coupling with transport.

In general for a set of species, the coupled problem for reactive transport is given by:

$$\phi \frac{\partial}{\partial t}\mathbf{c} = \left(\frac{\partial}{\partial z} D \frac{\partial}{\partial z} - v\frac{\partial}{\partial z}\right)\mathbf{c} + \mathbf{S}_{\text{eq}}{}^{T}\mathbf{r}_{\text{eq}} + \mathbf{S}_{\text{kin}}{}^{T}\mathbf{r}_{\text{kin}} \tag{6}$$

(Saaltink *et al.*, 1998), where the vector $\mathbf{c}$ contains all species concentrations. The vectors $\mathbf{r}_{\text{eq}}$ and $\mathbf{r}_{\text{kin}}$ denote the reaction rates of equilibrium and kinetic reactions, and the matrices $\mathbf{S}_{\text{eq}}$ and $\mathbf{S}_{\text{kin}}$ relate reactions (in rows) and species (in columns) for equilibrium and kinetic reactions.

The problem with (6) is that the rates of the equilibrium reactions are not known beforehand. Thus the entire set of equations is manipulated by linear transformations in order to make the term corresponding with equilibrium reactions vanish. Such a transformation is always possible, but not unique. It can be described by the multiplication of the system (6) with another matrix $\mathbf{U}$ from the left (Saaltink *et al.*, 1998), which is equivalent to the transition from species concentrations to total concentrations, also called *components*.

The system for the total concentrations $\mathbf{u} = \mathbf{U}\cdot\mathbf{c}$ is given by

$$\varphi \frac{\partial}{\partial t}\mathbf{u} = \left(\frac{\partial}{\partial z} D \frac{\partial}{\partial z} - v\frac{\partial}{\partial z}\right)\mathbf{u} + \mathbf{U}\mathbf{S}_{\text{kin}}{}^{T}\mathbf{r}_{\text{kin}} \tag{7}$$

In addition the reaction equilibria have to be taken into account, which can be noted as:

$$\mathbf{S}_{\text{eq}}\cdot\mathbf{log(a)} = \mathbf{log(K)} \tag{8}$$

where the vector $\mathbf{a}$ denotes all activities, and the vector $\mathbf{K}$ all equilibrium constants. Altogether a mathematical system results in which transport differential equations (7) are combined with a set of algebraic equations (8), so-called algebraic differential equations.

Sorption

Sorption denotes a variety of phenomena and processes, which concern the interaction between fluid and solid phase. In the most general approach, sorption reactions can be treated within an extended concept of reactive transport as surface reactions (Parkhurst, 1995). More common are simplified approaches. Kinetic laws can be used to describe slow (nonequilibrium) sorption. The simplest approach is to take first-order kinetics, but this may not suffice for reactive transport, where more complex approaches (for example: Monod) may become necessary. The most common situation of fast (equilibrium) sorption is modeled by combining the transport differential equation with a sorption isotherm (linear sorption, Freundlich-, Langmuir) that describes the equilibrium between solid and fluid phase, leading to the concept of retardation. The details are outlined in **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**.

Different codes handle sorption differently. Some codes allow the direct input of retardation factors (FAST), while others require the specification of the isotherm and the corresponding parameters. Some codes require one isotherm

for the entire domain, others allow the sorption characteristic to change with layers, others by cell (MT3D96).

NUMERICAL MODELS

There are different numerical techniques by which computer algorithms are derived from equations that govern the model. In order to obtain a numerical model, the mathematical (e.g. differential) formulation for continuous variables has to be transformed into discrete form. The discrete variables (e.g. hydraulic head in flow models, concentration, or temperature in transport models) of the model are determined at nodes in the model domain, determined by a *grid*.

Finite Differences

The method of Finite Differences (FD) is derived as approximation of the differential equation. Derivatives (differential quotients) are replaced by difference quotients. For first and second order derivatives, the simplest central stencils (CIS = central in space) are given by:

$$\frac{\partial f}{\partial x} \approx \frac{f_{i+1} - f_{i-1}}{2\Delta x} \qquad \frac{\partial^2 f}{\partial x^2} \approx \frac{f_{i+1} - 2f_i + f_{i-1}}{\Delta x^2} \tag{9}$$

where the f-values denote function values at the grid nodes, that is, f_i is the approximate value of the function at node i, f_{i-1} at the previous node, and f_{i+1} at the following node (see Figure 1). This leads to a system of equations for the unknown values ($f_i, i = 1..N$), where N denotes the total number of nodes. For transport problems, the upwind scheme (BIS = backward in space) is important:

$$\frac{\partial f}{\partial x} \approx \frac{f_i - f_{i-1}}{\Delta x} \tag{10}$$

In two space dimensions (2D), the five-point stencils, describing finite differences, can be visualized as shown in Figure 1. For example, the Laplace-operator $(\partial^2/\partial x^2 + \partial^2/\partial y^2)f$ is represented by a stencil with factor -4 at the center and 1 in the other four nodes.

FD-grids are usually rectangular, and may be irregular, that is, each column, row, or layer may possess individual grid spacing. The values of the dependent variable are calculated at the nodes, while parameters are specified for the spacing between the nodes (node centered grid).

Finite Volumes

The method of Finite Volumes (FV) is derived from a mass or volume balance for all blocks of the model region. As visualized in Figure 2, the load (e.g. volume or mass) balance in block ij is obtained by:

$$\frac{\partial V}{\partial t} = Q_{i-} + Q_{i+} + Q_{j-} + Q_{j+} + Q \tag{11}$$

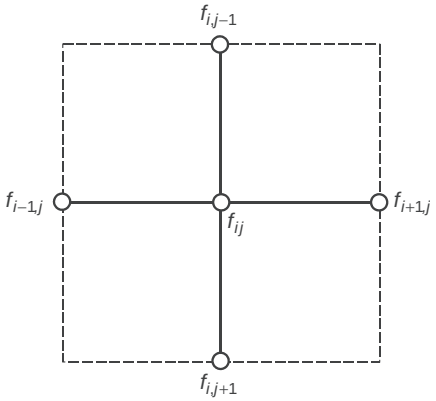


Figure 1 Finite difference stencil for function f in 2D

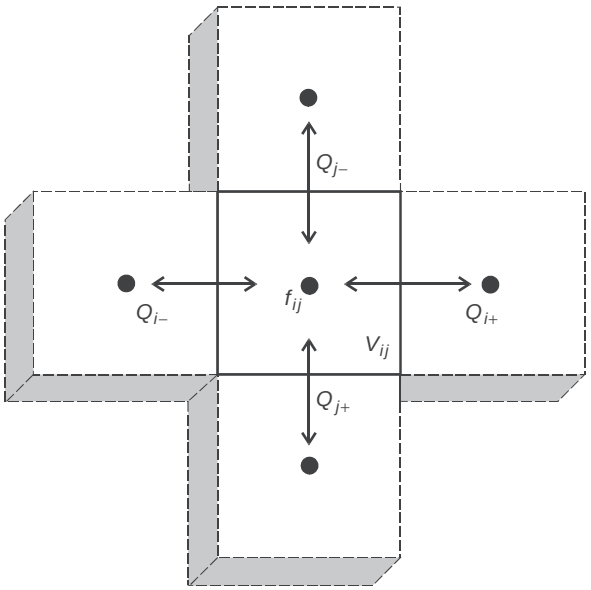


Figure 2 Quadratic finite volume in 2D

where V denotes the volume or mass (see Figure 2) in the block, Q_{i-}, Q_{i+}, Q_{j-}, Q_{j+} the fluxes across the block edges, and Q other sources or sinks for volume or mass.

A set of equations for the values of the unknown variables in the block centers is obtained by expressing the fluxes in terms of that variable. With the help of Darcy's Law, all volume balance equations come into a form that expresses relations between hydraulic heads $f_{ij} = h_{ij}$.

Similarly, Finite Volume grids may be of general form, as they are defined by the budget of fluxes across the boundaries of a block. The dependent variable is calculated at the center of the block (block centered grid). Parameters are specified for blocks, that is, around the block centers. The different form of grids used in FD and FV makes it difficult to compare results obtained by the different methods without interpolation.

Finite Elements

Finite elements grids can be of different shape, but very often they can be recognized by the simple triangular form of the single elements. The triangular shape is convenient to approximate arbitrary shaped regions with small deviations, where rectangular grids often show stairway structures at least at parts of the boundaries.

Using Finite Elements, the solution of the differential equation is found as a combination of shape (or Lagrangian)-functions. These functions are different for different elements. For the most common triangular elements, the prescribed Lagrangian functions are linear within each element (see Figure 3). Thus f is approximated, for example, in Cartesian coordinates, by

$$f(x, y) = a_{\alpha 0} + a_{\alpha 1}x + a_{\alpha 2}y \qquad \text{within element } \alpha \tag{12}$$

All coefficients $a_{\alpha j}$ for all elements are computed as solution of a linear or nonlinear system, which is derived from the so-called weak form of the differential equation (Huyakorn and Pinder, 1983).

Method of Characteristics / Lagrangian Methods

One approach that has been gaining popularity is the mixed Eulerian–Lagrangian method that combines the simplicity of the fixed Eulerian grid with the Lagrangian approach being especially effective in advective dominant regions. Following Neuman and Sorek (1982) for transport problems (e.g. see also Neuman, 1984; Sorek, 1988a,b), for flow problems (Sorek, 1985; Sorek and Braester, 1988) and a modified Eulerian–Lagrangian method for coupled flow and transport model (Bear *et al.*, 1997; Sorek *et al.*, 2000), a technique consisting of the following two steps is used:

1. Formal decomposition of the dependent variable into two parts, one controlled by pure Lagrangian advection and a residual governed by a combination of Euler–Lagrange approaches.

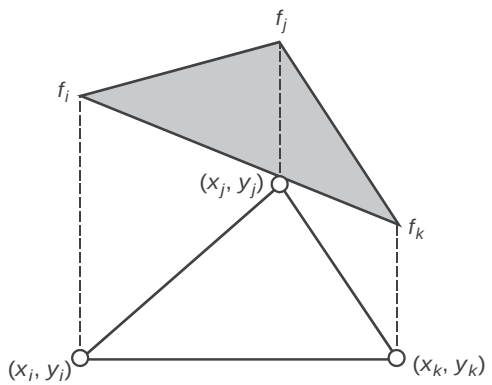


Figure 3 Finite element in 2D

2. Solution of the resulting advection problem by the method of characteristics for forward particle tracking. The residual problem is solved by, for example, an implicit finite element scheme on a fixed grid.

Information is projected back-and-forth between the Eulerian–Lagrangian and the Lagrangian schemes. The major problem of the decomposition strategy is the interpolation (coupling) between the residual and translation solutions as these are obtained at different spatial locations and should be performed in a manner that is mass conservative.

To understand the concept of the decomposition, consider a governing partial differential equation

$$\mathbb{L} f = Q \tag{13}$$

in which $\mathbb{L}$ denotes the differential operator, f denotes the dependent prime variable, and Q denotes the source term. We now decompose (13) into translation, $()_{\mathbb{T}}$, and residual, $()_{\mathbb{R}}$, terms to read

$$(\mathbb{L}_{\mathbb{T}} + \mathbb{L}_{\mathbb{R}})(f_{\mathbb{T}} + f_{\mathbb{R}}) = Q \tag{14}$$

One way to perform the decomposition is to allow

$$\mathbb{L}_{\mathbb{T}} f_{\mathbb{T}} = 0 \tag{15}$$

for which $f_{\mathbb{T}}$ is solved along a characteristic pathline defined by

$$\mathbb{L}_{\mathbb{T}} \mathbf{X} = \mathbf{V} \tag{16}$$

where $\mathbf{X}$ denotes the position spatial vector known only if $\mathbf{V}$ the velocity vector tangent to the characteristic pathline is provided, otherwise iterations are required. By virtue of (14) and (15), the next step is to solve

$$\mathbb{L} f_{\mathbb{R}} = Q - \mathbb{L}_{\mathbb{R}} f_{\mathbb{T}} \tag{17}$$

However, as we may expect the condition

$$\mathbb{L} f_{\mathbb{R}} \gg \mathbb{L}_{\mathbb{R}} f_{\mathbb{T}} \tag{18}$$

and in view of (17) and (18), we obtain

$$\mathbb{L} f_{\mathbb{R}} \cong Q \tag{19}$$

which is similar to a Poisson equation and is not expected to suffer from stability difficulties typical to advection dominant flow regimes. In cases with dominant source terms, we allow $f_{\mathbb{T}}$ to obey

$$\mathbb{L}_{\mathbb{T}} f_{\mathbb{T}} = Q \tag{20}$$

for which by virtue of (17) and (18), we solve $f_{\mathbb{R}}$ by a Laplacian equation in the form

$$\mathbb{L} f_{\mathbb{R}} \cong 0 \tag{21}$$

which is more stable than (19).

Variations of the method of characteristics are implemented in some of the most used transport codes, as MT3D. The user can choose between MOC (method of characteristics), Modified MOC (MMOC), and Hybrid MOC (Zheng and Bennett, 1995). While MOC uses forward tracing along the flowpaths (characteristics), in the MMOC the characteristics are traced backward from the nodes. The hybrid scheme combines both approaches (Neuman, 1984).

Time Stepping

The discretization of time is as a rule implemented as time stepping. The computed values at the previous time level t are used as initial condition for the calculation at the new time level $t + \Delta t$. The time-step Δt may vary. The entire algorithm starts in using the initial condition for the dependent variable of the problem setup, which has to be specified by the user in addition to the differential equation with its boundary conditions.

Simple time-stepping algorithms can be described by the following formula:

$$f(t + \Delta t) = f(t) + \Delta t[\kappa \cdot C(f(t + \Delta t) + (1 - \kappa)C(f(t))] \tag{22}$$

In the notation, the dependency of f on space variables is neglected. The operator C denotes the spatial discretization, which can be achieved using the finite differences (FD), finite volumes (FV), or finite elements (FE) method. κ is a parameter, which either has a fixed value for the implemented method or can be chosen by the modeler. For $\kappa = 1/2$, when both old and new time levels are equally weighted, one obtains the classical *Crank–Nicolson* procedure. For $\kappa = 1$, the totally implicit method results, where the spatial discretization is taken only at the new time level. For $\kappa = 0$, an explicit algorithm results, which is cheaper to solve (as the solution of a linear system is not required). But often the accuracy of the explicit algorithm is poor, or it even does not converge.

Advanced time-stepping algorithms are given by the so-called *Runge-Kutta* methods (Holzbecher, 1996), which can, for example, be found in the FEFLOW code. In mathematical toolboxes, which can also be applied for the solution of differential equations, often an automatic time stepping is implemented, where the accuracy is checked during the simulation and the time step is reduced, if necessary.

Mixing Cells

In contrast to the usual aforementioned procedure, in the mixing cell approach, grid spacing and time stepping are combined. For given velocity v, time-step Δt and grid spacing Δx are related by the formula

$$\Delta t = \frac{\Delta x}{v} \tag{23}$$

The approach is usually applied for constant velocity (1D). The technique can also be used for 1D flow fields with variable velocity, when Δx is varied along the flow paths. But the method is not applicable for general higher dimensional flow fields. Nevertheless, when transverse gradients are small, the mixing cell approach can be used for simulation of 1D transport along a flowpath.

It is obvious from formula (23) that advection with velocity v is modeled without any discretization error. It turns out to be advantageous to combine the mixing cell approach for advection with the conventional FD or FV approach for dispersion. Such operator splitting is implemented in the PHREEQC code (Appelo and Postma, 1993). The PHREEQC code was not originally intended to be combined with a transport model as velocity is not a prescribed input. Instead *lengths* (Δx), *time-step* (Δt), *cells* (number of blocks), and *shifts* (number of time steps) have to be specified as parameters (Parkhurst, 1995).

As for general time stepping, the discretization error, connected with the advection term, turns out to be the most severe in many applications, and the mixing cell approach is very competitive, concerning accuracy with other approaches, when it can be applied.

Reactive Transport

The simultaneous solution of general 3D transport and geochemical speciation is still a challenge for modelers nowadays, as there is a high demand on computer resources' time and space. In the simplest case, the equations for transport can be solved in a first step, delivering total concentrations. The second step is the speciation calculation based on the equilibrium equations (8) depending on the total concentrations available. Speciation calculations require the solution of a highly nonlinear problem, for which the Newton–Raphson algorithm is usually applied (Parkhurst, 1995).

Unfortunately with respect to the outlined solution strategy, the reaction terms for the kinetic reactions mostly depend on species concentrations. Thus the first step cannot be performed without the results of the second step. The problem is handled using three different strategies (Steefel and MacQuarrie, 1996):

- the sequential two step is performed (sequential noniterative approach – SNIA)

- the two-step method is iterated within each time step (sequential iterative approach – SIA)
- discretized differential and algebraic equations are gathered in one system and solved (direct solution approach – DSA)

Surely the SNIA is the cheapest of these methods with respect to the consumption of computer resources, but it delivers incorrect results when there is a strong coupling through the kinetics. Another reason for the popularity of the SNIA is that different available programs for transport and programs for speciation calculations can be loosely coupled, while both other approaches request an internal coupling during each time step.

The coupling of the popular MT3D-MS code for multi-species transport and PHREEQC2 is currently in development (Prommer *et al.*, 2003), for which the SNIA approach is applied.

Accuracy and Stability Criteria

The approximation of a differential equation by a numerical method is not exact, and yields discretization errors. When derivatives are replaced by finite differences, a truncation error results, for which formulae can be derived using the Taylor (or Lagrangian)-series representation. Local errors, due to truncation of derivatives or due to round-off of numbers, may be amplified with the further application of the algorithm. In such a case, the algorithm is called *unstable*.

Concerning accuracy and stability, three dimensionless numbers and three related criteria are relevant:

$$\text{Grid-Péclet number/criterion}: \quad Pe = \frac{v \cdot \Delta x}{D} \leq 2 \qquad (24)$$

$$\text{Courant number/criterion}: \quad Cou = \frac{v \cdot \Delta t}{\Delta x} \leq 1 \qquad (25)$$

$$\text{Neumann number/criterion}: \quad Neu = \frac{D \cdot \Delta t}{\Delta x^2} \leq \frac{1}{2} \qquad (26)$$

The grid-Péclet criterion is relevant for most numerical methods, although an explicit derivation is seldom found, except for finite differences. Figure 4 illustrates typical errors when the criterion is violated. Breakthrough curves from analytical and numerical solutions for 1D front propagation with constant parameters are depicted. The CIS method is burdened by overshooting. In contrast, the BIS method is burdened by enhanced dispersion, the so-called *numerical dispersion*. While the CIS algorithm is stable, when the grid-Péclet criterion is fulfilled the BIS still displays numerical dispersion for $Pe < 2$. An improvement of the BIS method can then be obtained, when the input value for dispersion is reduced by the numerical dispersion value, which according to the truncation error analysis is given by $D_{\text{num}} = v \cdot \Delta x/2$ (Lantz, 1971). The curve,

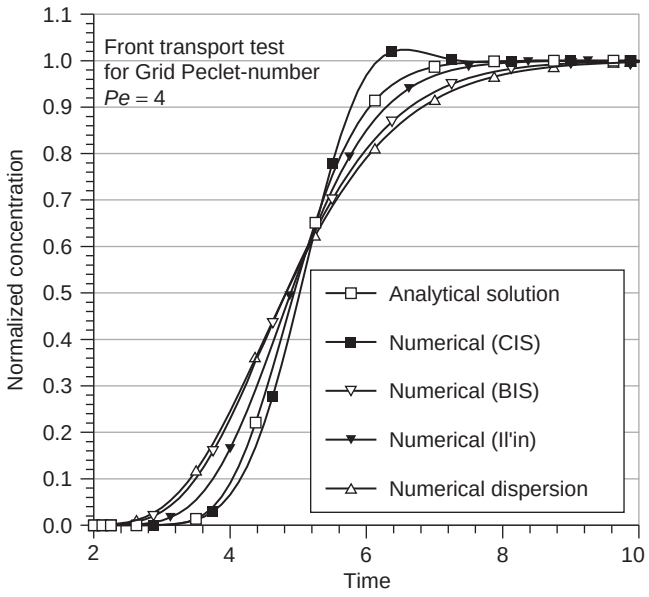


Figure 4 The performance of three numerical schemes for grid-Péclet number $Pe = 4$; all numerical results obtained using automatic time stepping with the Euler method (FIT)

referenced by 'numerical dispersion', depicts the analytical solution for a situation in which D_{num} increases D. The deviation of the BIS-result from that curve shows that the real error of BIS is slightly less than the one predicted by truncation error analysis.

The Il'in method (Il'in, 1969) can be described as a compromise between BIS and CIS. Tests show that it has advantages compared to the other methods only in the vicinity of the critical value for the grid-Péclet numbers, $Pe = 2$. For higher grid-Péclet numbers, the Il'in scheme suffers also from severe numerical dispersion.

The grid-Péclet criterion can be fulfilled when the grid is chosen fine enough, that is, for small Δx. For advection-dominated problems, this may lead to problems when the number of unknown becomes too high.

The Neumann criterion in the given form is valid for explicit algorithms and is less strict or can be completely neglected for implicit algorithms. Both Courant and Neumann criteria can be fulfilled when for a given grid the time step is chosen small enough. This strategy has its limits, as the execution time for a computer run increases with the number of time steps. Especially in 3D modeling, a reduction of the time step may lead to computation periods that are unacceptably high.

The mixing cell method has no discretization error for advection. In order to reduce the discretization error, an operator splitting approach can be applied. It is convenient to use a different time step for diffusion that is a fraction of the advection time step. The PHREEQC code uses a diffusion time-step Δt, which fulfills the condition:

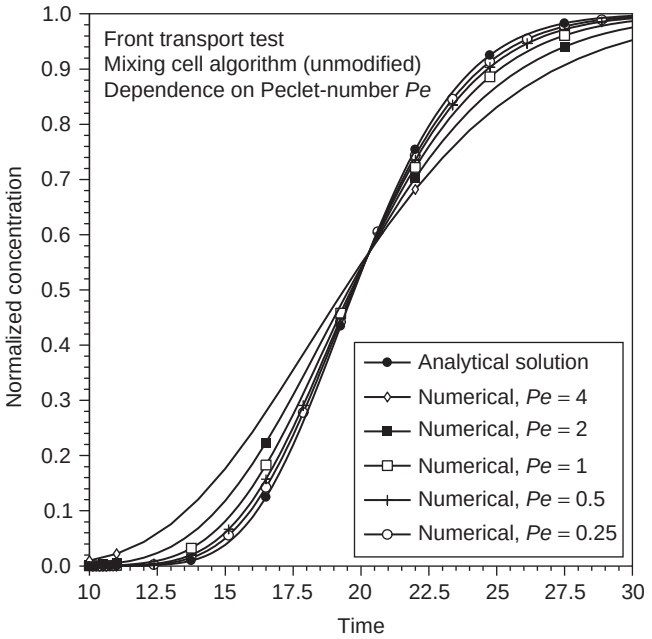


Figure 5 The performance of the mixing cell algorithm for different grid-Péclet numbers; all numerical results obtained using automatic time stepping with the Euler method (FIT)

$Neu \leq \frac{1}{3}$, and which reduces the numerical error very effectively (Appelo and Postma, 1993). Figure 5 shows the performance of a mixing cell method for the 1D front propagation testcase. For given velocity and diffusivity, the grid-spacing and corresponding time step have been reduced, leading to decreasing grid-Péclet numbers *Pe*. The breakthrough curves illustrate how the error is reduced upon reducing of the grid-Péclet number.

INVERSE MODELING

It is the usual task of modeling to determine the dependent variable while the user provides the parameters as input. In groundwater flow models, values for the piezometric head at grid nodes result from using input module for hydraulic conductivities and boundary conditions. But in practical problems, it is often the piezometric heads for which field measurements are available, while hydraulic conductivities are uncertain by at least 1 order of magnitude. In such a situation, the model is used in a different manner. Conductivities as input parameters are varied in order to obtain an optimal match between numerical and measured values of heads.

The illustrated strategy is termed *inverse modeling*, as the role of parameters and variables is exchanged. Inverse modeling is treated in more detail in **Chapter 156, Inverse Methods for Parameter Estimations, Volume 4**. The mathematical procedure is the same for calibration or parameter estimation.

The task of inverse modeling is mostly not implemented in a groundwater modeling code itself. For example, in the ProcessingMODFLOW or FEFLOW, the software package PEST is referred for parameter estimation. PEST or UCODE (alternative in ProcessingMODFLOW) perform all operations concerning the parameter estimation task, that is, new parameter datasets are calculated that are candidates for a model with a better match with the observed data. Then a direct model run is started and finally the results of that run are checked.

Modern GUIs allow the modeler to input measured datasets, to select parameters to be estimated, and to select options for the parameter estimation package. When there are also options for the comparison of measured and modeled data, there is no need to use any other tool for direct and inverse modeling.

PREPROCESSING

Preprocessing is the transformation of data into a form appropriate for the calculating code. Often geologic data are available in a database that cannot be accessed by the modeling program directly. Sometimes special transformation routines are needed to perform this task. Sometimes GIS software for the visualization of geo-data is used.

All operations that are concerned with the selection of the model region, which means especially the location of the boundaries, belong to preprocessing. The modeler can load a background map from some file and construct the model region on the screen using the mouse. Parts of the boundary can be selected for the definition of boundary conditions. Input boxes that require the relevant parameters to be specified open immediately.

Preprocessing includes the choice of model characteristics, such as: confined/unconfined/partially confined, 1D/2D/3D, flow/transport/flow and transport, transient/steady state, solute/heat/solute and heat transport. Nowadays using GUIs most of these options can be selected by mouse-click on a virtual button on the display.

Preprocessing includes the representation of data in a finite grid. Most codes allow external files to be linked with the model. An interpolation routine is then started internally, which delivers parameter values for all elements, blocks, or volumes (see FEFLOW for example). Some codes have options to define parameters as random variable with given statistical properties (see FAST for example).

Finite Element Codes are usually equipped with a *grid generator*. Based on some basic options concerning the approximate number of elements and the type of elements, the GUI usually allows an FE-grid to be created by the push of a button when the model region is defined.

When the modeler is not content with the grid, it can be refined entirely or in parts only. The part to be refined is selected by simple mouse operation. For FD or FV codes that use rectangular grids, such a gridding is usually not necessary.

SOLUTION

The solver is the code that really solves the set of equations for the discrete variable. From GUIs, the solver is started using a submenu entry (ProcessingMODFLOW, Visual MODFLOW, GMS) and for standard situations works with default values for solution parameters. In the case of problems with the solution, that is, no or poor convergence to the numerical solution, or when the results are not accurate enough, the solver or/and parameters should be changed.

Internally mostly a sparse linear or nonlinear system has to be solved. Often there are some options concerning the solution of such a system. Direct solvers can only be recommended for relatively small number of unknowns, that is, for coarse grids. An iterative solver usually is the default. Preconditioned conjugate gradient solvers are mostly used because they are relatively robust, that is, with their standard parameters they work well for a wide range of different problems. For details the reader should consult textbooks like Barrett *et al.* (1994). In some codes, multigrid solvers are included, which are expected to deliver faster convergence.

POSTPROCESSING

After the run of the solver, the discrete representation of the dependent variable is calculated. In order to get specific information from a huge array of numbers, the modeler can perform various postprocessing tasks.

A major postprocessing tool for groundwater flow models is the budget calculator. Fluxes of different type (e.g. inflow, outflow, well recharge, well discharge, groundwater recharge) can be calculated for the entire model region or for user-specified parts of it. For most models such a tool enables quantitative information on the basic fluxes. That is not only a basic information from a completed model, before the completion it is also a basic indicator for errors.

Contour lines (for 2D models) or contour surfaces (for 3D models), which can be plotted for the dependent variable, provide a picture of the variables distribution. Isopleths for hydraulic head, isobars for pressure, isotherms for temperature, isohalines for salt concentration, and streamlines for streamfunction visualize the results of a model run. They not only show where the variable is on an equal level, the minima and maxima also can clearly be identified when the contour levels are chosen appropriate. The different GUIs offer different options for the user to make such a choice.

Contour plots for hydraulic head provide additional information on the flow because flow is normal to the contour lines. When contour levels are equidistant, the density of the isolines gives an impression of the relative amount of the velocity. Regions with dense isolines have higher velocities than regions with less isoline density if the compared regions have the same hydraulic conductivity. Arrow plots give a direct impression of the velocity distribution and flow direction. Long thick arrows represent high velocities in contrast to short thin arrows representing small velocities. In groundwater, the velocities within a model region usually differ by at least 1 order of magnitude, which causes some problem of the arrow plot representation.

Prominent postprocessing tools are codes for particle tracking. For MODFLOW, several of such codes are available (MODPath, PMPath, PATH3D) that visualize streamlines and flowpaths. Some tracking software is designed for steady-state flow fields only, while others can be used for transient flow fields also. For a given flow field, particles can be traced forward or backward in time. There are various different options to set clusters of starting points for the algorithm at inflow or outflow boundaries and/or in the vicinity of wells. Backward particle tracking from positions around a well enables the visualization of the catchments. Similarly the watersheds of groundwater lakes can be determined (Holzbecher, 2001).

Time markers on streamlines or flowpaths indicate isochrones. Most tracking software provides several options to select time markers appropriately, concerning the time levels and the outlook of the marker. In order to calculate traveltimes for a flow field, the porosity has to be specified by the modeler.

Figure 6 depicts a typical output of a groundwater flow model worked out with several postprocessing tools. Added to a background map are well galleries, hydraulic head contour lines, and flowpaths with time markers.

SOFTWARE OVERVIEW

The most prominent code for groundwater modeling is MODFLOW. The most recent version is MODFLOW2000, described by Harbaugh *et al.* (2000). The origins of MODFLOW can be traced back to the beginning of the 1980s. An overview on the history of MODFLOW is given by McDonald and Harbaugh (2003).

Today there are several graphical user interfaces that are 'built around' MODFLOW, like ProcessingMODFLOW, Visual MODFLOW, GMS, and MODFLOW-GUI. These programs assist the user with various pre- and postprocessing tasks; they call MODFLOW for groundwater flow

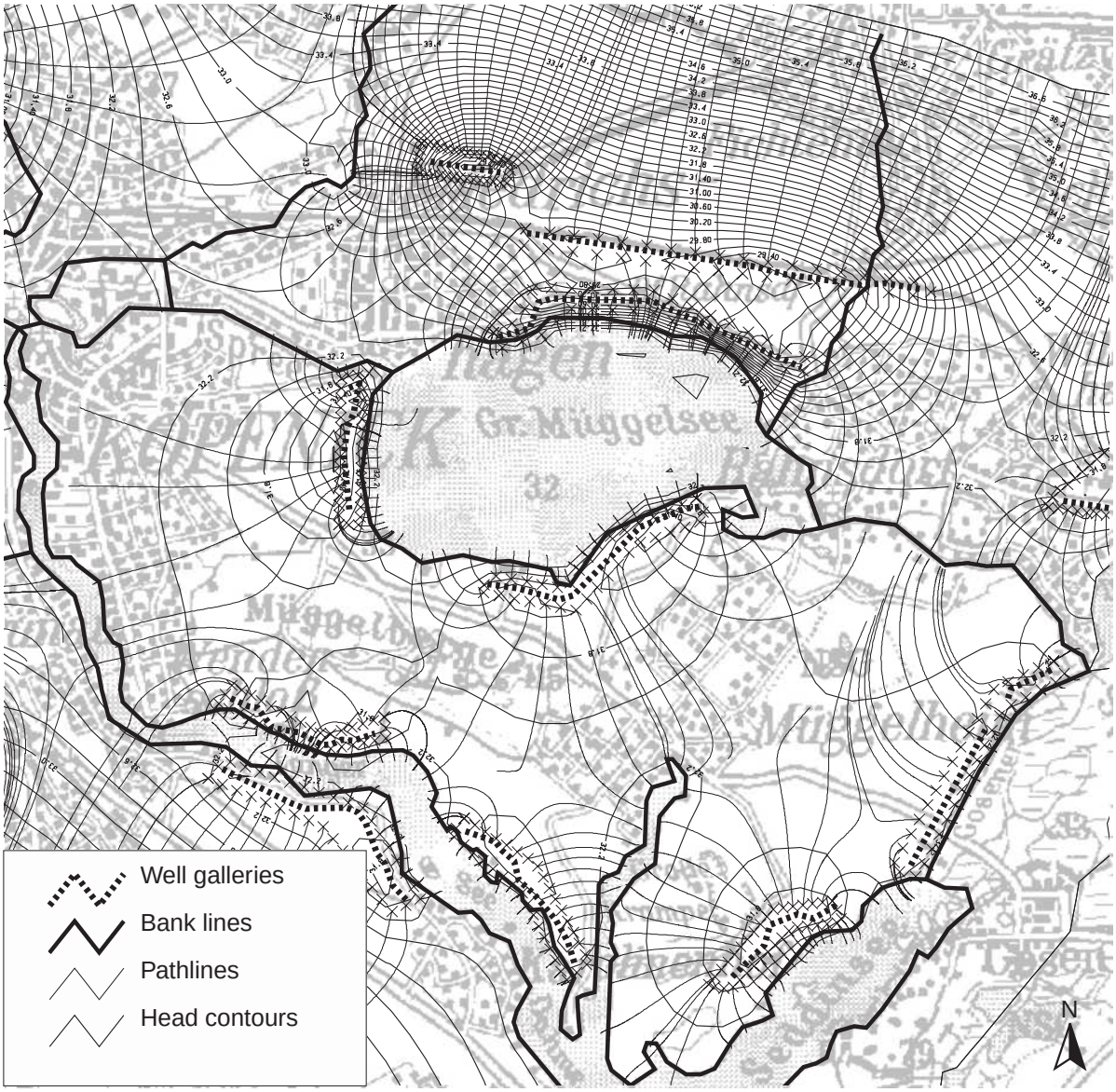


Figure 6 A typical output from a groundwater flow model, depicting background map with well galleries, and the output from several postprocessing tasks: head contours and flowpaths with time markers for illustration of isochrones. The model was setup by M. Zippel using FEFLOW code, postprocessing, and GIS-connection. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

models or other codes for transport simulations or other tasks. Moreover, it is possible to use the GUI for direct and inverse modeling.

A surely incomplete list of groundwater modeling codes is given at the end of this chapter. The second column explains if the code solves the resulting systems of equations itself or if it is build around another solver code as a GUI. Some programs are simply tools for pre- and postprocessing. GUIs include tools. The third column indicates the numerical method used for the discretization (FD = Finite Differences, FE = Finite Elements, FV = Finite Volumes). GUIs, which are connected to different solvers, can have more than one entry. For tools, the discretization method does not have to be given as it depends on the solver.

Acknowledgments

The authors are grateful to Matthias Zippel for the presentation of results of the FEFLOW model of well galleries around Lake Müggelsee, which is to be published as part of his doctorate thesis.

SOFTWARE LINKS

Code name	Solver GUI Tool	FD FE FV	Commercial PD	Short Description	Internet Link
CFEST	S	FE		Coupled fluid-energy-solute transport	http://db.nea.fr/abs/html/nesc9537.html
CHAIN-2D	S	FE	PD	2D transport of decay chain	http://www.ussl.ars.usda.gov/MODELS/CHAIN2D.HTM
FAST	S, GUI	FV		3D flow, 2D transport, 2D density driven	http://www.igb-berlin.de/abt1/mitarbeiter/holzbecher/index_e.shtml
FEFLOW	S, GUI	FE	Com.	3D flow, solute, and heat transport	http://www.wasy.de/english/produkte/feflow/index.html
FEMWATER	S, GUI	FE		3D groundwater flow	http://www.scisoftware.com/products/gms_fem/gms_fem.html
GMS	GUI	FE, FV	Com.	for FEMWATER, MODFLOW, MT3D, RT3D, SEEP2D, SEAM3D	http://www.ems-i.com
HST3D	S	FD	PD	3D flow, solute, and heat transport	http://water.usgs.gov/software/hst3d.html
MOC3D	S		PD	3D method of characteristics (flow and transport)	http://water.usgs.gov/software/moc3d.html
MODFLOW	S	FV	PD	3D flow	http://water.usgs.gov/software/modflow.html
MODFLOW-GUI	GUI	FV	PD	for MODFLOW and MOC3D, works under ARGUS ONE only	http://water.usgs.gov/nrp/gwsoftware/mfgui4/modflow-gui.html
MODPATH	Tool	–	PD	Particle tracking for MODFLOW	http://water.usgs.gov/software/modpath.html
Model Viewer	Tool	–	PD	Visualization of 3D model results	http://water.usgs.gov/nrp/gwsoftware/modelviewer/ModelViewer.html
MT3D	S	FV	PD	3D transport	http://hydro.geo.ua.edu/
MT3D-MS	S	FV	PD	3D multiple species transport	http://hydro.geo.ua.edu/
PATH3D	Tool	–		Particle tracking for MODFLOW	http://hydro.geo.ua.edu/mt3d/path3d.htm
PHREEQC	S	Cells[a]	PD	Geochemistry and 1D Transport	http://water.usgs.gov/software/phreeqc.html
PEST	Tool	–	Com.[b]	Parameter Estimation	http://www.parameter-estimation.com
PMWIN	GUI	FV	Com.[b]	for MODFLOW, MOC, MT3D, PEST and UCODE	http://www.scisoftware.com/products/pmwin_details/pmwin_details.html

(*continued*)

Code name	Solver GUI Tool	FD FE FV	Commercial PD	Short Description	Internet Link
PORFLOW		FD	Com.	3D flow, solute, and heat transport	http://www.acri.fr/English/Products/PORFLOW/porflow.html
ROCKFLOW		FE		3D flow, solute, and heat transport	http://www.hydromech.uni-hannover.de/Projekte/Grundwasser/misc/news.html
RT3D		FV	PD	Reactive transport based on MT3D-MS	http://bioprocess.pnl.gov/rt3d.htm
SEAM3D	GUI	FV		Reactive transport based on MT3D-MS	http://modflow.bossintl.com/html/seam3d.html
SUTRA	S, GUI	FE	PD	Flow and Transport	http://water.usgs.gov/software/sutra.html
SWIFT	S	FV	Com.	3D fluid, solute, and heat transport	http://www.scisoftware.com/products/swift_overview/swift_overview.html
TBC	S	FE, FV	PD	Transport, Biochemistry, and Chemistry	http://www.iwr.uni-heidelberg.de/~Wolfgang Schafer/tbc201.pdf
UCODE	Tool	–	PD	Parameter Estimation	http://water.usgs.gov/software/ucode.html
Visual MODFLOW	GUI	FV	Com.	for MODFLOW, MT3D, RT3D and PEST	http://www.flowpath.com/software/visualmodflow/visualmodflow.html

[a]Can be regarded as special type of Finite Volumes, for 1D only.
[b]Limited version is freeware.

FURTHER READING

Holzbecher E. (2002) *Groundwater Modeling – Computer Simulation of Groundwater Flow and Pollution*, FiatLux Publications: Fremont, **http://envirocomp.org**

REFERENCES

Appelo C.A. and Postma D. (1993) *Geochemistry, Groundwater and Pollution*, Balkema: Rotterdam.

Barrett R., Berry M., Chan T.F., Demmel J., Donato J.M., Dongarra J., Eijkhout V., Pozo R., Romine C. and van der Vorst H. (1994) *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM: Also available on the World Wide Web: **http://netlib2.cs.utk.edu/linalg/html_templates/Templates.html**

Bear J., Sorek S. and Borisov V. (1997) On the Eulerian-Lagrangian formulation of balance equations in Porous media. *Numerical Methods for Partial Differential Equations*, **13**(5), 505–530.

Harbaugh A.W., Banta E.R., Hill M.C. and McDonald M.G. (2000) *MODFLOW-2000, the U.S. Geological Survey Modular Ground-water Model – User Guide to Modularization Concepts and the Ground-Water Flow Process*, U.S. Geological Survey: Open-File Report 00-92.

Holzbecher E. (1996) *Modellierung Dynamischer Prozesse in Der Hydrologie: Grundwasser und Ungesättigte Zone*, Springer Publication: Heidelberg.

Holzbecher E. (1998) *Modeling Density-Driven Flow in Porous Media*, Springer Publication: Heidelberg.

Holzbecher E. (2001) The dynamics of subsurface water divides. *Hydrological Processes*, **15**, 2297–2304.

Huyakorn P.S. and Pinder G. (1983) *Computational Methods in Subsurface Flow*, Academic Press: New York.

IAEA International Atomic Energy Agency (1988) *Radioactive Waste Management Glossary, IAEA TECDOC-447, Second Edition*, IAEA: Vienna.

Il'in A.M. (1969) Differencing scheme for a differential equation with a small parameter effecting the highest derivative. *Mathematical Notes of the Academy of Sciences USSR*, **6**, 596–602.

Krauskopf K.B. and Bird D.K. (1995) *Introduction to Geochemistry*, McGraw Hill: New York.

Lantz R.B. (1971) Quantitative evaluation of numerical diffusion (truncation error). *Transactions Society of Petroleum Engineers*, **251**, 315–320.

McDonald M.G. and Harbaugh A.W. (2003) The history of MODFLOW. *Groundwater*, **41**(2), 280–283.

Neuman S.P. (1984) Adaptive Eulerian-Lagrangian finite element method for advection-dispersion. *The International Journal for Numerical Methods in Engineering*, **20**, 321–337.

Neuman S.P. and Sorek S. (1982) Eulerian-Lagrangian methods for advection dispersion, *Proceedings of 4th International Conference on Finite Elements in Water Resources*, Hannover.

Parkhurst D.L. (1995) *PHREEQC: A Computer Program for Speciation, Reaction-Path, Advective Transport, and Inverse Geochemical Calculations*, Water-Resources Investigation Report 95-4227, U.S. Geological Survey, Lakewood.

Prommer H., Barry D.A. and Zheng C. (2003) MODFLOW/MT3DMS-based reactive multicomponent transport modelling. *Groundwater*, **41**(2), 247–257.

Saaltink M.W., Ayora C. and Carrera J. (1998) A mathematical formulation for reactive transport that eliminates mineral concentrations. *Water Resources Research*, **34**(7), 1649–1656.

Silling S.A. (1983) *Final Technical Position on Documentation of Computer Codes for High-Level Radioactive Waste Management*, NUREG-0856.

Sorek S. (1985) Eulerian-Lagrangian formulation for flow in soils: theory. *Advances in Water Resources*, **8**, 118–120.

Sorek S. (1988a) Two-dimensional adaptive Eulerian-Lagrangian method for mass transport with spatial velocity distribution. *Transport in Porous Media*, **3**, 473–489.

Sorek S. (1988b) Eulerian-Lagrangian method for solving transport in aquifers. *Advances in Water Resources*, **11**(2), 67–73.

Sorek S. and Braester C. (1988) Eulerian-Lagrangian formulation of the equations for groundwater denitrification using bacterial activity. *Advances in Water Resources*, **11**(4), 162–169.

Sorek S., Borisov V. and Yakirevich A. (2000) Numerical modeling of coupled hydrological phenomena using the modified Eulerian-Lagrangian method. In *Theory, Modeling and Field Investigation in Hydrogeology: A Special Volume in Honor of Shlomo P. Neuman's 60th Birthday*, Zhang D. and Winter C.L. (Eds.), Geological Society of America: Special Paper 348, pp. 151–160.

Steefel C.I. and MacQuarrie K.T.B. (1996) Approaches to modeling of reactive transport in porous media. In *Reactive Transport in Porous Media*, Vol. 34, Lichtner P.C., Steefel C.I. and Oelkers E.H. (Eds.), Mineralogical Society of America: pp. 83–129.

Zheng C. and Bennett G.D. (1995) *Applied Contaminant Transport Modeling*, Van Norstrand Reinhold: New York.

156: Inverse Methods for Parameter Estimations

NE-ZHENG SUN[1] AND ALEXANDER Y SUN[2]

[1] *Department of Civil and Environmental Engineering, University of California, Los Angeles, CA, US*
[2] *CNWRA, Southwest Research Institute, San Antonio, TX, US*

Estimation of various flow and mass transport parameters can be seen as the inverse problem of groundwater modeling. In this article, current methodologies for parameter estimation are classified according to whether they consider the errors in observation data, in parameter structure, and in model applications. The structures of hydraulic parameters are usually very complex and unknown. This article gives an overview on adaptive parameterization methods for simultaneously identifying the parameter values, the complexity level, and the pattern of parameter structure. Methods for model reliability analysis are described. Finally, a generalized inverse problem is introduced that can find a representative parameter for a given model application.

INTRODUCTION

The Forward Problem

Mathematical models of groundwater flow and groundwater quality (**Chapter 149, Hydrodynamics of Groundwater, Volume 4**, **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**, **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**, **Chapter 150, Unsaturated Zone Flow Processes, Volume 4** and **Chapter 157, Sea Water Intrusion Into Coastal Aquifers, Volume 4**) can be represented in the following general form:

$$L(u; q; p; b; x) = 0 \tag{1}$$

where L is a set of partial differential operators, u a set of state variables, q a set of control variables, p a set of model parameters, b a set of initial and boundary conditions, and x a set of spatial and time variables. When $(q; p; b)$ are given, the problem of solving the state variables u from (1) is called the *forward problem* (FP). The general form of FP solution can be represented by

$$u = M(q; p; b; x) \tag{2}$$

Various analytical methods (**Chapter 151, Hydraulics of Wells and Well Testing, Volume 4**) and numerical methods (**Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**) have been developed for solving FP. Presently, the solution of FP is rather established in geohydrology. In the real world, however, $(q; p; b)$ cannot be determined completely by geological information and direct measurements. When incorrect $(q; p; b)$ are used in a simulation model, the result of model prediction is unreliable and the model may become useless, no matter how accurate the FP solution is. The determination of correct $(q; p; b)$ is, therefore, the key to successful groundwater modeling. Bear (1979) pointed out that there are three basic problems in groundwater modeling: the prediction problem, the calibration problem, and the management problem. A model, including its structure and parameters, must be carefully calibrated by observations of state variables and other correlated information. We must make sure that a model can give reliable prediction results before it is used, for example, in a management problem, for finding the optimal decisions. This article describes how model parameters are calibrated and how the reliability of model prediction can be guaranteed.

The Inverse Problem

For a model given by (2), the observed values of state variables, u^{obs}, at observation locations and times x^{obs} can

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

be expressed by the following *observation equation*:

$$u^{\text{obs}} = M(q; p; b; x^{\text{obs}}) + \varepsilon \tag{3}$$

where ε contains both observation and model errors. It is well known that a useful model must be very carefully calibrated with all available observations, $\{u^{\text{obs}}\}$. The problem of model calibration can be seen, in a certain sense, as the inverse problem (IP) of partial differential equations (Isakov, 1998). IP seeks model parameters (q, p, b) when state variables (u) are measured, while FP predicts state variables (u) when model parameters (q, p, b) are given. In their original meaning, FP finds "results" from "causes", whereas IP finds "causes" on the basis of "results". Because the same "results" may be caused by different "causes", IP is usually ill-posed, that is, its solution (the identified parameters) may be nonunique and discontinuously dependent on data. Examples of ill-posed IPs in groundwater modeling can be found in Sun (1994).

IP can also be considered in statistical frameworks as the parameter estimation problem. With different assumptions on the probability distribution of observation errors, different criteria of parameter estimation can be derived. The statistical method gives not only the estimated parameters, but also the reliability of the estimation.

The Study of IP in Groundwater Modeling

In groundwater modeling, the history of studying IP can be traced all the way back to the beginning. For example, in 1950s and 1960s, analytical solutions were used to identify the hydraulic conductivity and storage coefficient around a well through fitting a curve to aquifer testing data (**Chapter 151, Hydraulics of Wells and Well Testing, Volume 4**). In 1970s, the "history matching" methods developed in petroleum engineering were used to identify hydraulic parameters by numerical optimization (Chavent *et al.*, 1975). Hydrogeologists then developed their own methods for calibrating groundwater models (e.g. Carrera and Neuman, 1986; Yeh, 1986; Sun, 1994; McLaughlin and Townley, 1996; etc.). Currently, many groundwater packages include modules for solving IP. Hydrogeologists solve IP to identify hydraulic parameters, boundary conditions, pollution sources, dispersivities, adsorption kinetics, and reaction coefficients (**Chapter 149, Hydrodynamics of Groundwater, Volume 4**, **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**, **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**, **Chapter 150, Unsaturated Zone Flow Processes, Volume 4**). The IP-based sensitivity analysis and reliability analysis play very important roles in optimal management of groundwater resources, in optimal design of observation networks, and in feasibility study of remediation plans (**Chapter 153, Groundwater Pollution and Remediation, Volume 4**).

There are two inherent difficulties in solving IP for groundwater modeling: (i) the geological structure of a real aquifer is usually very complex and unknown, and (ii) data that can be used for model calibration are usually very limited, in both quantity and quality. These two difficulties are intertwined in the inverse solution process. A simple model structure (or parameter structure when the model structure is parameterized) may not fit the observed data well. On the other hand, a complicated model structure may cause over-parameterization when data are limited. Attempting to find the real structure of an aquifer and accurate parameter values by solving IP is cost-ineffective and may lead to a useless model. We must systematically consider the relationship between the complexity of model structure, the identifiability of model parameter, the sufficiency of data, and the reliability of model application. A new methodology for parameter structure reduction and robust design (Sun, 2005) may help us to find a way to construct cost-effective and reliable groundwater models.

Parameterization

Hydraulic parameters are usually functions of location and/or time. In this article, we use $\theta(x)$ to represent a distributed parameter (function). Obviously, it is impossible to identify a distributed parameter with infinite degrees of freedom by a finite set of observation data. *Parameterization* refers to approximately representing a distributed parameter by a function with low degrees of freedom. A general representation of parameterization is

$$\theta(x) \approx \sum_{j=1}^{m} \theta_j \phi_j(x, v) \tag{4}$$

where m is called the *dimension of parameterization*, θ_j $(j = 1, 2, \ldots, m)$ are *weighting coefficients*, $\{\phi_j(x, v)\}$ is a set of *basis functions* with a set of shape parameters v. Various parameterization methods, such as the *zonation* method, the finite element method, the spline method, the geostatistical method, the wavelet method, and the natural neighboring method (Sun, 1994; Liu, 1994; Sambridge, 2001), can be seen as special cases of (4). We will use (S, θ) to denote a *parameterization representation* (PR) of a distributed parameter $\theta(x)$, where S represents a parameter structure determined by m basis functions, and the vector

$$\theta = (\theta_1, \theta_2, \ldots, \theta_m)^T \tag{5}$$

consists of parameter values associated with the structure. The same distributed parameter may have different PRs when it is approximated by different structures.

In Sun and Sun (2002), three kinds of inverse problem are identified, namely, the *classical inverse problem* (CIP), the extended inverse problem (EIP), and the *generalized inverse problem* (GIP). In CIP, the structure S is assumed to be known and only the weighting coefficients (5) need to be identified. In EIP, all components in (4) (m, θ_j and $\phi_j(x, v)$, i.e. both S and θ) are identified simultaneously by data. In GIP, these components are identified on the basis of the reliability of model application.

In the next section, CIP is defined in both deterministic and statistical frameworks. Various criteria solution methods for optimal parameter estimation are described. The Section "Parameter structure identification" considers the function identification problem or EIP. All components of PR are optimized by data. Statistical criteria for complexity selection and various algorithms for pattern identification are reviewed. The geostatistical inverse solution is discussed in this section as a special method of function identification. The Section "Structure reduction and experimental design" is dedicated to the solution of the GIP, through which the simplest PR is found and is used to replace the true parameter for model applications. Concepts and algorithms on robust design are discussed. The Section "Software and applications" is a short review on available software packages and application of parameter estimation in groundwater modeling.

THE CLASSICAL PARAMETER ESTIMATION PROBLEM

Inverse Problems as Statistics

In the statistical framework, the estimated parameter, θ, is regarded as a random vector. The problem of parameter estimation can be considered as a procedure that transfers information from observations (u^{obs}) to the unknown parameter (θ) through a model ($u = M(\theta; x)$), and thus reduces the uncertainty of the estimated parameter. The best method of parameter estimation should extract information from data as much as possible, and thus reduce the uncertainty of the estimated parameter as much as possible.

Let $p(\theta)$ be the *joint probability density distribution* (pdf) of a parameter vector θ. The uncertainty associated with $p(\theta)$ is measured by its entropy (Bard, 1974):

$$H(p) = -E(\log p) = -\int_{(\Omega)} p(\theta) \log p(\theta)\, \mathrm{d}\theta \qquad (6)$$

where $E(\log p)$ is the mathematical expectation of $\log p(\theta)$, and (Ω) is the whole distribution space. The negative value of $H(p)$, that is, $-H(p) = E(\log p)$, is defined as the *information content* of the distribution $p(\theta)$. The prior information on parameters θ can be described by a pdf, $p_0(\theta)$, which is called the *prior distribution* of θ. After transferring the information from the observation data to the estimated parameter(s), we obtain a new pdf, $p_*(\theta)$, which is called the *posterior distribution* of θ. The information contents contained in the prior and posterior distributions are $-H(p_0)$ and $-H(p_*)$, respectively. Their difference, $I(p_0, p_*) = H(p_0) - H(p_*)$, measures the information content transferred from the observation data.

The posterior distribution $p_*(\theta)$ is the pdf of θ conditioned by observation data u^{obs}, that is, the conditional pdf, $p(\theta|u^{\text{obs}})$. On the other hand, the conditional pdf, $p(u^{\text{obs}}|\theta)$, is the pdf of u^{obs} conditioned by parameter θ. If there is no model error, $p(u^{\text{obs}}|\theta)$ is equal to $p(\varepsilon|\theta)$ – the pdf of observation errors when θ is given. The conditional pdf, $p(u^{\text{obs}}|\theta)$, is usually called the *likelihood function* of observations and is denoted by $L(\theta)$. According to *Bayes' theorem*, we have

$$p_*(\theta) = cL(\theta)p_0(\theta), \text{ where constant } c = \left(\int_{(\Omega)} p(u^{\text{obs}}|\theta)p_0(\theta)\, \mathrm{d}\theta\right)^{-1} \qquad (7)$$

Equation (7) facilitates the transfer of information from observation data to the estimated parameter. The *maximum a posterior* estimation that maximizes the information transferred from observation data and prior information is given by

$$\hat{\theta} = \arg \min_{\theta}\{-p_*(\theta)\} \qquad (8)$$

Substituting (7) into (8) and using $\ln p_*(\theta)$ to replace $p_*(\theta)$, we have

$$\hat{\theta} = \arg \min_{\theta}\{-\ln L(\theta) - \ln p_0(\theta)\} \qquad (9)$$

The estimated $\hat{\theta}$ thus depends on both the likelihood function and the prior distribution.

Criteria for Optimal Parameter Estimation

Different criteria for optimal parameter estimation can be obtained on the basis of different assumptions on the prior distribution and the distribution of observation error. When $p_0(\theta)$ is a uniform distribution and the admissible region is a multidimensional box $\Theta_{ad} = (\theta_L, \theta_U)$, where θ_U and θ_L are the upper and lower bounds of the estimated parameter vector, we have the following maximum likelihood estimator (MLE):

$$\hat{\theta} = \arg \min_{\theta}\{-\ln L(\theta)\}, \text{ s.t. } \theta \in \Theta_{ad}. \qquad (10)$$

When the observation error vector is normally distributed with zero mean and constant covariance matrix V_ε (an $n \times n$ matrix, where n is the number of observation data), the likelihood function can be specified as

$$L(\theta) = (2\pi)^{-n/2}(\det V_\varepsilon)^{-1/2} \exp\left\{-\tfrac{1}{2}[u^{\text{obs}} - u^{\text{cal}}(\theta)]^T V_\varepsilon^{-1}[u^{\text{obs}} - u^{\text{cal}}(\theta)]\right\} \quad (11)$$

where $u^{\text{cal}}(\theta) = M(\theta, x^{\text{obs}})$, and the MLE (10) reduces to the *generalized least squares estimator* (GLSE):

$$\hat{\theta} = \arg \min_\theta [u^{\text{obs}} - u^{\text{cal}}(\theta)]^T V_\varepsilon^{-1}[u^{\text{obs}} - u^{\text{cal}}(\theta)], \text{ s. t.} \theta \in \Theta_{ad} \quad (12)$$

Furthermore, if all components of ε are independent from each other, V_ε^{-1} reduces to a diagonal matrix, $\mathbf{W}$, and GLSE becomes the *weighted least squares estimator* (WLSE):

$$\hat{\theta} = \arg \min_\theta [u^{\text{obs}} - u^{\text{cal}}(\theta)]^T \mathbf{W}[u^{\text{obs}} - u^{\text{cal}}(\theta)], \text{ s.t. } \theta \in \Theta_{ad} \quad (13)$$

When all observation errors have the same variance, $\mathbf{W}$ reduces to a unit matrix.

When $p_0(\theta)$ is normally distributed with mean θ_0 and covariance matrix V_θ (an $m \times m$ matrix, where m is the dimension of θ), the above estimators must be changed according to (9), that is, adding $-\ln p_0(\theta)$ to the objective function of minimization. For example, the GLSE becomes

$$\hat{\theta} = \arg \min_\theta \{[u^{\text{obs}} - u^{\text{cal}}(\theta)]^T V_\varepsilon^{-1}[u^{\text{obs}} - u^{\text{cal}}(\theta)] + [\theta - \theta_0]^T V_\theta^{-1}[\theta - \theta_0]\} \quad (14)$$

When $V_\varepsilon = \sigma^2\mathbf{I}$ and $V_\theta = \tau^2\mathbf{I}$, where $\mathbf{I}$ denotes the unit matrix, (14) reduces to the following *regularized least squares estimator* (RLSE) with $\lambda = \tau^2/\sigma^2$:

$$\hat{\theta} = \arg \min_\theta \{[u^{\text{obs}} - u^{\text{cal}}(\theta)]^T [u^{\text{obs}} - u^{\text{cal}}(\theta)] + \lambda[\theta - \theta_0]^T[\theta - \theta_0]\} \quad (15)$$

In the *deterministic framework*, when observation error exists, we can only find an approximate inverse solution. IP is thus transformed into an optimization problem. After parameterization, the unknown parameter vector θ can be estimated by minimizing the "distance" (or a misfit function), d_D, between the observed data and the model output measured in the observation space D, that is, let

$$\hat{\theta} = \arg \min_\theta d_D[u^{\text{obs}}, u^{\text{cal}}(\theta)], \text{ s.t.} \theta \in \Theta_{ad} \quad (16)$$

where Θ_{ad} is the admissible region of the unknown parameter that can be estimated by prior information (**Chapter 151, Hydraulics of Wells and Well Testing, Volume 4**, **Chapter 148, Aquifer Characterization by Geophysical Methods, Volume 4**). A norm defined in the observation space can be used to measure the "distance" d_D. When the weighted L-2 norm is used, (16) is the same as the WLSE given in (13) but the weights may have different meanings. For example, we may assign larger weights to those observation wells that are more important for model prediction. Generally, the statistical criteria given above can still be used in the deterministic framework regardless of the statistical assumptions on observation error and prior information. When the observation error is not normally distributed, it is better to use L-1 norm to measure the "distance" in (16):

$$d_D(u^{\text{obs}}, u^{\text{cal}}) = \sum_{i=1}^{n} w_i |u_i^{\text{obs}} - u_i^{\text{cal}}| \quad (17)$$

where $\{w_i\}$ is a set of weighting coefficients. L-1 norm is less sensitive to observation error than L-2 norm and is thus more robust. The following Kullback–Leibler (KL) misfit function provides another robust criterion for parameter identification:

$$d_D(u^{\text{obs}}, u^{\text{cal}}) = \sum_{i=1}^{n} u_i^{\text{obs}}[\ln u_i^{\text{obs}} - \ln u_i^{\text{cal}}] \quad (18)$$

No matter what misfit functions or criteria are used for parameter estimation, the formulated optimization problem might have multiple local minima and its solution might be unstable due to the ill-posed nature of IP and the nonlinearity of the misfit function. The *regularization theory* (Tikhonov and Arsenin, 1977) may significantly improve the stability of the inverse solution with

$$\hat{\theta} = \arg \min_\theta \{d_D[u^{\text{obs}}, u^{\text{cal}}(\theta)] + \lambda R(\theta)\} \quad (19)$$

In (19), $\lambda R(\theta)$ is called a *regularization term*, $R(\theta)$ the *regularization function* and λ the *regularization factor*. For example, we may choose $R(\theta) = d_P(\theta, \theta_0)$, where $d_P\ (\theta, \theta_0)$ is a norm defined in the parameter space that measures the "distance" between the estimated parameter θ and its prior estimation θ_0. When L-2 norm is used, (19) has the same form as the statistical criterion (16). But, regularization is more general and robust. The value of λ represents an appropriate compromise between the misfit of observations and the misfit of prior information. From this point of view, the regularization term can be considered as a penalty function, (19) can be seen as a vector optimization problem, and λ as the weighting coefficient. Finding the optimal regularization factor is still an open problem in

real case studies (Vogel, 2002). Besides the identified parameter, the regularization term may contain also the state variable and its derivatives. In general, we may add more regularization terms to (19) if other kinds of information or measurement are available. When we have the measurements of k state variables $v_1, v_2, \ldots, v_k$, the *coupled inverse problem* (Sun, 1994) can be solved by:

$$\hat{\theta} = \arg \min_{\theta} \left\{ d_D[u^{\text{obs}}, u^{\text{cal}}(\theta)] + \sum_{j=1}^{k} \lambda_j d_{D,j}[v_j^{\text{obs}}, v_j^{cal}(\theta)] + \lambda_{k+1} R(\theta) \right\} \quad (20)$$

In the field of groundwater modeling, (20) has been used to identify the hydraulic conductivity by coupling the measurements of head, concentration, water temperature, water content, water age, flow velocity, as well as geophysical data (e.g. Sun and Yeh, 1990; Portniaguine and Solomon, 1998; Bravo *et al.*, 2002; Lin and Zhang, 2004).

Optimization Algorithms

The criteria for parameter estimation, in both statistical and deterministic frameworks, can be seen as a *constrained optimization problem*

$$\hat{\theta} = \arg \min_{\theta} E(\theta), \theta \in \Theta_{ad} \quad (21)$$

where the objective $E(\theta)$ is a misfit function. Various numerical methods have been developed for solving (21) (Kelley, 1999). An iterative procedure for solving (21) includes: (1) choose a starting point θ_0; (2) designate a way to generate a search sequence: $\theta_0, \theta_1, \theta_2, \ldots, \theta_k, \theta_{k+1}, \ldots$; and (3) specify a termination criterion. The search sequence has the following general form $\theta_{k+1} = \theta_k + \lambda_k \mathbf{d}_k$, where vector $\mathbf{d}_k$ is called a *displacement direction*, and λ_k is a *step size* along the direction. Different optimization methods use different algorithms to generate $\mathbf{d}_k$ and λ_k in each iteration. This iteration process usually leads to or terminates at a local minimum. When L-2 norm is used for parameter estimation, the objective $E(\theta)$ has the form of the sum of squares of functions. In this case, the *Levenberg-Marqardt algorithm,* a modified *Gauss-Newton* method, is often used to generate the search sequence:

$$\theta_{k+1} = \theta_k - (\mathbf{J}_k^T \mathbf{J}_k + \lambda \mathbf{I})^{-1} \mathbf{J}_k^T \mathbf{f}_k \quad (22)$$

where $\mathbf{f}_k = \mathbf{f}(\theta_k)$, $\mathbf{J}_k = \mathbf{J}(\theta_k)$, $\mathbf{f}$ has the components $f_i(\theta) = w_i[u_i{}^{\text{obs}} - u_i{}^{\text{cal}}(\theta)]$, $\mathbf{J}$ is the Jacobian matrix $[\partial f_i / \partial \theta_j]_{n \times m}$, $\mathbf{I}$ is the identity matrix, and λ is a variable coefficient. Starting from $\lambda = 0$, if the condition $E(\theta_{k+1}) < E(\theta_k)$ is satisfied, move to the next iteration, otherwise, increase the value of λ and try again. When the value of λ is large enough, the displacement direction approaches the steepest descent direction (the negative gradient direction) and the step size becomes very small. Therefore, we can expect the condition $E(\theta_{k+1}) < E(\theta_k)$ to be satisfied eventually.

In most real cases of groundwater modeling, a search sequence may be trapped to a *local minimum,* and the value of $E(\theta)$ may stay large, and is insensitive to the change of parameter values. *Genetic algorithm* (GA) is a kind of evolution algorithm that can find the *global optimum* for complex multimodal functions. Useful information on GA and its implementations can be found on the Internet, for example, **www.geneticprogramming.com**. The use of GA for parameter identification consists of the following steps:

1. Define an initial population. According to the prior distribution, we choose a set of possible solutions $\{\theta_1^0, \theta_2^0, \ldots, \theta_s^0\}$ from the admissible set Θ_{ad} as the initial population. For a homogeneous prior distribution, the initial population can be selected randomly.
2. Evaluate the fitness of the initial population. The simulation model is used to calculate the values of $\{E(\theta_1^0), E(\theta_2^0), \ldots, E(\theta_s^0)\}$.
3. Use genetic operators to create newer and "fitter" populations. The most common genetic operators are the selection, crossover, and mutation.
4. Evaluate the fitness of the new generation and test if a near-optimal solution has been found. If not, return to step 3 to create the next generation.

In the above procedure, all units in a generation must be encoded for genetic operations and decoded for fitness evaluation. The fitness of a parameter unit θ is evaluated by the value of $E(\theta)$. When significant computation effort is needed for solving FP, as in the case of groundwater modeling, GA could become very inefficient. An *artificial neural network* (ANN), after being trained, can approximately represent a complicated input-output relationship. Using ANN to replace the solution of the FP for fitness evaluation will significantly increase the effectiveness of GA (Morshed and Kaluarachchi, 1998). Solomatine *et al.* (1999) compared several global optimization algorithms for groundwater model calibration. Giacobbo *et al.* (2002) used GA to estimate the dispersivity parameters in a mass transport model. Shigidi and Garcia (2003) explained how to determine the optimal node number and training patterns of a neural network for transmissivity estimation.

Sensitivity Coefficients

When a gradient-based optimization method is used to solve the inverse problem, the gradient $g = [\partial E / \partial \theta]$ must be calculated in each iteration. When Levenberg–Marqardt method is used, the Jacobian, $\mathbf{J} = [\partial u^{\text{cal}} / \partial \theta]$, must be

calculated. In mathematical modeling, the first-order derivatives of dependent variables with respect to parameters, such as the elements of g and $\mathbf{J}$, are called *sensitivity coefficients*. Sensitivity coefficients are important not only for parameter estimation, but also for model reliability analysis and experimental design. From the analysis of sensitivity coefficients, we can understand the model behavior. There are three calculation methods.

The *perturbation method* uses the finite difference approximation of derivatives. For each component θ_j of θ, we assign an increment (perturbation) $h_j = \lambda\theta_j$, where $\lambda = 1\% \sim 10\%$, and run the simulation model to find the corresponding increment of objective function E (or the increments of the dependent variables u). The sensitivity coefficients are then calculated approximately as the proportion of increments. If the perturbation method is used to obtain the gradient g or the Jacobian $\mathbf{J}$ for an updated θ, we need to run the simulation model $m + 1$ times when the one-side finite difference approximation is used, and $2m + 1$ times when the two-side finite difference approximation is used. The perturbation method is the simplest one for sensitivity analysis and has been extensively used in groundwater modeling software, but its accuracy is low and is dependent on the value of perturbation.

The *sensitivity equation method* differentiates the governing equation as well as its subsidiary conditions with respect to m components of θ to obtain m sensitivity equations and then solves these equations to obtain sensitivity coefficients. For the general model given in (1), that is, $L(u, \theta) = 0$, its sensitivity equations with respect to θ_j are given by

$$\nabla_u L \frac{\partial u}{\partial \theta_j} + \nabla_\theta L \frac{\partial \theta}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, m \tag{23}$$

The gradient operator $\nabla_u L = [\partial L / \partial u]$ generally has the same structure as the original operator L, and, thus, (23) can be solved the same as the original problem. If the sensitivity equation method is used to obtain the gradient g or the Jacobian $\mathbf{J}$ for an updated θ, we need to solve the m sensitivity equations in (23), in addition to the original equation. The total computation effort is equal to $m + 1$ simulation runs.

The *adjoint state method* uses the variational form of the original model and solves the adjoint state equations to obtain the sensitivity coefficients. By defining the scalar product of two vector functions u and v as $(u, v)_\Omega = \int_\Omega (u \cdot v)\, d\Omega$, where Ω is the time/space definition domain, for a vector operator matrix $\mathbf{A}$, we may find a transposed adjoint operator matrix $\mathbf{A}^+$, such that the following identity stands:

$$(\mathbf{A}u, v)_\Omega = (u, \mathbf{A}^+ v)_\Omega \tag{24}$$

Now consider a general performance criterion $E = \int_\Omega f(u, \theta)\, d\Omega$, where $f(u, \theta)$ is a user-chosen function. The variation of E can be represented by

$$\delta E = \left(\delta u, \frac{\partial f}{\partial u}\right)_\Omega + \left(\delta\theta, \frac{\partial f}{\partial \theta}\right)_\Omega \tag{25}$$

Using (24) and the variational form $\nabla_u L \delta u + \nabla_\theta L \delta\theta = 0$ of the original problem (1), for any vector function ψ, we can rewrite (25) as

$$\delta E = \left(\delta u, \nabla_u^+ L\psi + \frac{\partial f}{\partial u}\right)_\Omega + \left(\delta\theta, \nabla_\theta^+ L\psi + \frac{\partial f}{\partial \theta}\right)_\Omega \tag{26}$$

Let ψ be the adjoint state of the original state variable u, that is, the solution of the following adjoint equation

$$\nabla_u^+ L\psi + \frac{\partial f}{\partial u} = 0 \tag{27}$$

with zero-boundary condition and zero-final condition, the first term on the right-hand side of (26) vanishes. For any component θ_j, we then have:

$$\frac{\partial E}{\partial \theta_j} = \int_\Omega \left[\nabla_\theta^+ L\psi + \frac{\partial f}{\partial \theta}\right]_j d\Omega, \quad j = 1, 2, \ldots, m \tag{28}$$

In the above equation, $[\cdot]_j$ represents the j-th component of vector $[\cdot]$. By selecting different $f(u, \theta)$, E can be a misfit function of parameter estimation or a model output corresponding to an observation. Using the adjoint state method to calculate the gradient $g = [\partial E / \partial \theta]$, we only need to solve the original problem (1) once and the adjoint problem (27) once. To calculate the Jacobian $\mathbf{J} = [\partial u^{\text{cal}} / \partial \theta]$, n adjoint problems need to be solved, where n is the number of observations. In Sun (1994), adjoint problems are derived for various groundwater flow, mass transport, and coupled problems.

When we use the sensitivity equation method (or the adjoint state method), different sensitivity equations (or different adjoint state equations) must be derived for different problems under consideration, the corresponding codes must be developed, and the correctness of each code must be validated. To avoid this manual work, a new approach called automatic differentiation (AD) is being developed that can directly differentiate a Fortran or a C++ code of forward solution to obtain the code for calculating sensitivity coefficients (Elizondo *et al.*, 2002). A complete list on AD software packages can be found from **http://www-unix.mcs.anl.gov/autodiff/AD_Tools**. The AD-based sensitivity analysis has been used in parameter estimation, uncertainty analysis, and experimental design (for example, Barhen and Reister, 2003).

The Reliability of Parameter Estimation

Linearization is the simplest method to estimate the reliability of the estimated parameter. During the estimation procedure, model output $u^{\text{cal}}(\theta)$ can be approximately regarded

as a linear function. In k-th iteration, we have $u^{\text{cal}}(\theta) \approx \mathbf{J}(\theta_k)(\theta - \theta_k) + u^{\text{cal}}(\theta_k)$, where $\mathbf{J} = [\partial u^{\text{cal}}/\partial\theta]$. From the theory of linear regression, when GLSE (12) is used, the covariance matrix of the estimated parameter is given by

$$Cov(\theta, \theta) = (\mathbf{J}^T V_\varepsilon^{-1} \mathbf{J} + V_\theta^{-1})^{-1} \tag{29}$$

Since the diagonal of the matrix gives the variances of estimation, the confidence region can then be calculated. Finding the confidence region for the estimated parameter, however, is not the sole purpose of reliability study. What we need to know is the reliability of model application when the identified parameter is used in the model. Let $g(\theta)$ be a set of given model application. There are several methods that can be used to assess the uncertainty of $g(\theta)$ on the basis of the uncertainty of the identified parameter θ.

When $g(\theta)$ is close to a linear function, we may use the *first-order approximation*:

$$\hat{g} = g(\hat{\theta}), Cov(\hat{g}, \hat{g}) = \mathbf{J}_g Cov(\hat{\theta}, \hat{\theta}) \mathbf{J}_g^T \tag{30}$$

where $Cov(\theta, \theta)$ is given by (29) and $\mathbf{J}_g = [\partial g/\partial\theta]$ can be obtained by the methods of sensitivity analysis. A recent example of using (30) is the work of Kunstmann *et al.* (2002), in which the reliability of groundwater head and concentration predictions is assessed. When $g(\theta)$ is not close to a linear function, we may use the *perturbation method* to find the variation δg from the variation $\delta\theta$ by solving both mean and perturbation equations (**Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). For example, the uncertainty associated with the model-predicted concentration distribution C is measured by its variance, σ_C^2, which can be directly solved from the perturbation equation (Tang and Pinder, 1979).

In Section "Inverse problems as statistics", we only find the maximum of the posterior $p_*(\theta)$ for parameter estimation. If we can find the whole distribution $p_*(\theta)$, the mean and variance of any model application $g(\theta)$ can be estimated by the following integrals:

$$E(g) = \int g(\theta) p_*(\theta)\, \mathrm{d}\theta$$

$$\text{and} \qquad Var(g) = \int [g(\theta) - E(g)]^2 p_*(\theta)\, \mathrm{d}\theta \tag{31}$$

The above equation is generally applicable for any nonlinear model and the unknown parameter may have any type of distribution. Letting $g(\theta) = \theta$ in (31), we can obtain the mean and variance of the estimated parameter. Moreover, with the whole distribution $p_*(\theta)$, we can search for the global optima for parameter estimation rather then a local one. To obtain $p_*(\theta)$, we can use a *sampling method* to find the values of $p(u|\theta)$ and then use Bayes' theorem to find the values of $p_*(\theta) = p(\theta|u)$. The *Markov Chain Monte Carlo* (MCMC)-based sampling approaches, such as the Metropolis algorithm and Gibbs sampler algorithm, use a constructed random walk procedure to sample the posterior distribution. Readers may refer Gilks *et al.* (1996) for a detailed discussion on MCMC. The sampling method is applicable for solving IP and for uncertainty analysis when the dimension m of θ is not high. Sohn *et al.* (2000) gave an example of using the sampling method in groundwater modeling for uncertainty assessment.

All of the above-mentioned methods of uncertainty analysis are for CIP, and thus they assume there is no parameter structure error. In practice, however, the structure error often dominates the parameter value error and is the cause of model failure. We will discuss this problem in the following sections.

PARAMETER STRUCTURE IDENTIFICATION

Parameter Structure Complexity

The statistical approach can be used not only for parameter estimation and reliability analysis, but also for determining the model complexity level. The classical model selection criteria, such as the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC), tend to select the simplest model if it can fit the observations in the same extent as complex models do (Akaike, 1974; Schwarz, 1978). For example, the Schwarz criterion is given by

$$d = d_D(u^{\text{obs}}, u^{\text{cal}}) + \frac{m}{n-m}\left(\ln\frac{n}{2}\right) \tag{32}$$

where n is the number of data, m the dimension of the unknown parameter. On the right-hand side of (32), the first term measures the fitting error, the second term is a penalty to the complexity of parameter structure. Another alternative of model complexity selection is resampling (Lahiri, 2003). In this approach, the prediction risk is estimated via cross-validation, and the model providing lowest estimated risk is chosen.

In the statistical learning theory, the following "ε-*insensitive misfit function*" is used:

$$d_{D,\varepsilon}(u^{\text{obs}}, u^{\text{cal}}) = \sum_{i=1}^{n} w_i d_{i,\varepsilon}^2 \tag{33}$$

where $d_{i,\varepsilon} = 0$, when $|u_i^{\text{obs}} - u_i^{\text{cal}}| \le \varepsilon$; otherwise, $d_{i,\varepsilon} = |u_i^{\text{obs}} - u_i^{\text{cal}}| - \varepsilon$. This definition means that the difference between observed and model calculated values is replaced by zero if it is less than ε (the upper bound of observation error, for example). The use of (33) (or its regularization form) in groundwater modeling is strongly suggested because it can avoid fitting noised observations with an overly complex structure. The *structure risk minimization*

(SRM) criterion provides a very general framework for complexity control (Vapnik, 1998). Under SRM, a set of possible models is ordered according to their complexity, the theoretical upper bounds of prediction risk can be found under certain assumptions, and a structure that provides the minimal prediction risk is chosen. Cherkassky and Ma (2003) showed that SRM outperforms AIC and BIC for their testing data sets. In groundwater modeling, we need to identify not only the complexity level but also the pattern (the pattern of hydraulic conductivity, for example). No analytical expression for the upper bound of prediction risk is available in this case.

Identification of the Optimal Parameterization

With different parameter structures for parameter estimation, we will obtain different parameter values and different fitting residuals. When the dimension of parameterization m is too small (under-parameterized), the fitting residual may have a large value. On the other hand, when m is too large (over-parameterized), the model prediction becomes unreliable. Moreover, even if we can find an appropriate m through a model selection criterion, the identified parameter may be still very different from the real one if the structure pattern of the unknown parameter is not correctly assigned. In EIP, structure S and parameter values θ are identified simultaneously by solving the following optimization problem

$$(\hat{S}, \hat{\theta}) = \arg \min_{(S,\theta)} \{d_D[u^{\text{obs}}, u^{\text{cal}}(S, \theta)] + \lambda d_P[(S, \theta) - (S_0, \theta_0)]\} \tag{34}$$

where $u^{\text{cal}}(S, \theta) = M(S, \theta; x^{\text{obs}})$. The second term on the right-hand side of the above equation is a regularization term, in which (S_0, θ_0) is a prior guess of the estimated parameter representation (S, θ). This term may be replaced by a constraint $(S, \theta) \in \Theta_{ad}$. (39) is a combinatorial optimization problem and is very difficult to solve because the dimension of the shape vector may be high. In practice, we can only identify an approximation of the true structure because of data limitation.

Sun and Yeh (1985) was the first to solve problem (34) to identify the hydraulic conductivity of a heterogeneous aquifer without knowing its structure. The shape vector that determines the parameter structure consists of the coordinates of a set of moving basis points and two variable parameters used for determining the type of basis functions, from discrete to continuous. Problem (34) is solved as a min−min optimization problem. The number of basis points is increased one by one. For each increase in basis points, the outer minimization loop finds the optimal structure (shape vector); and for each structure, the inner minimization loop finds the optimal parameter values associated with the structure. The same problem (34) was solved by Zheng and Wang (1996) with simulated annealing to find the global optimization of structure parameters. Heredia *et al.* (2000) presented the following criterion to replace the objective function given in (34):

$$d_D[u^{\text{obs}}, u^{\text{cal}}(S, \theta)] + m \ln\left(\frac{n}{2\pi}\right) + \ln\left|\frac{\det \mathbf{F}}{n}\right| - 2 \ln p_0(S, \theta) \tag{35}$$

where m is the dimension of θ, n the number of data, $\det \mathbf{F}$ is the determinant of the information matrix, $\mathbf{F} = (\mathbf{J}^T \mathbf{J})^{-1}$, $p_0(S, \theta)$ contains the prior information transferred from geology or geophysics on both parameter structure and parameter values. The first term of (35) measures the data fitting error, the second term is the penalty to the structure complexity, the third term is the penalty to the parameter uncertainty, and the final term is the prize to the prior information. In Heredia *et al.* (2000), the identified zonation pattern is obtained by gradually modifying an initially guessed pattern from the available prior information. Recently, Ameur *et al.* (2002) presented an adaptive procedure for pattern identification. In each iteration, a refinement indicator and a coarsening indicator are calculated to determine if a zone should be refined or coarsened. In Tsai *et al.* (2003), a genetic algorithm (GA), a grid search method, and a quasi-Newton algorithm are combined to find the optimal pattern of parameter structure in three dimensions. These optimization processes need huge computational effort.

In order to make the solution of EIP practical, Sun and Sun (2002) presented a simple *tree-regression procedure* that can find a nearly optimal parameter structure with less computational effort. In this procedure, the following nested structure sequence is formed by gradually increasing the complexity:

$$S_1 \subset S_2 \subset S_3 \subset \cdots \subset S_m \subset S_{m+1} \subset \cdots \tag{36}$$

In the above equation, S_1 is a homogeneous structure, S_{m+1} is generated by dividing a selected zone of S_m into two zones with a linear boundary. The parameter value associated with the selected zone is the most sensitive one to the misfit function. The location of the linear boundary and the parameter values associated with all $m + 1$ zones are then determined by the method of Sun and Yeh (1985). This sequence is terminated when one of the following criteria is satisfied: (1) the value of $E(\hat{\theta}_m)$ becomes small compared with the observation error; (2) the difference between $E(\theta_m)$ and $E(\theta_{m-1})$ becomes small; (3) there are no significant changes in parameter values when one zone is divided into two zones; and (4) $E(\theta_m)$ becomes insensitive to all zones. In this case, no more information can be extracted from the observation data. With this tree-regression algorithm, the over-parameterization problem

can be avoided, the dimension of the shape vector is minimized, and all parameters involved in the optimization procedure are sensitive to the observation data.

The Geostatistical Method for Parameter Estimation

We treat the unknown parameter as a random variable because it is identified from the observation equation that contains uncontrollable observation and model errors. From the point of view of geostatistics (Wackernagel, 2003), a distributed physical parameter itself, such as the hydraulic conductivity, may be regarded as a *spatial random field* because of the *variability* of natural formations at different scales (**Chapter 154, Stochastic Modeling of Flow and Transport in Porous and Fractured Media, Volume 4**). Usually, a random field is described approximately by its first two moments. A geostatistical inverse method finds the *mean* and *covariance* functions for the estimated random field $\theta(x)$ using the measurements of the fields (*kriging*) and the measurements of correlated fields (*co-kriging*). In groundwater modeling, for example, the estimated random field is often the log-hydraulic conductivity $\theta(x) = \ln K(x)$, and the correlated fields may be hydraulic head, concentration, water temperature, water content, and so on. The log-hydraulic conductivity is often assumed to be normally distributed with an exponential type of covariance function. For any two points x_i and x_j, we assume

$$Cov_{\theta\theta}(x_i, x_j) = \psi_1 \delta_{ij} + \psi_2 \exp\left(\frac{-d_{ij}}{\psi_3}\right) \qquad (37)$$

where δ_{ij} is the Kronecker delta, d_{ij} is the distance between x_i and x_j, $\psi = (\psi_1, \psi_2, \psi_3)$ are statistical parameters to be identified. The value of the distributed parameter estimated by the geostatistical method at any unmeasured location x is a linear combination of all available measurements:

$$\hat{\theta}(x) = \sum_{i=1}^{m} \lambda_i(x)\theta^*(x_i) + \sum_{k=1}^{K}\sum_{l=1}^{nk} \mu_{kj}(x)u^*_{kj}(x_{kj}) \qquad (38)$$

where $\{\theta^*(x_i)|i = 1, 2, \ldots, m\}$ is a set of parameter measurements, K is the number of correlated fields, $\{u^*_{kj}(x_{kj})|j = 1, 2, \ldots, nk\}(k = 1, 2, \ldots, K)$ are K sets measurements of the correlated fields, $\{\lambda_i(x)\}$ and $\{\mu_{kj}(x)\}$ are kriging and co-kriging coefficients associated with x, respectively. These coefficients can be obtained by solving the kriging and co-kriging equations. The coefficient matrices of these equations depend on the auto-covariance and cross-covariance functions between the parameter field and the correlated fields. The solution of a geostatistical inverse problem consists of two stages. In the first stage, the unknown mean function (characterized by parameters β) and the covariance function (characterized by parameters ψ) are estimated by the MLE on the basis of all sets of measurement data. In the second stage, the same sets of measurement data are used to obtain the co-kriging estimation (38) as the most possible realization. The geostatistical inverse solution can provide distributed values of the unknown parameter for describing the spatial variability. Moreover, it can also give the variance of estimation:

$$Var[\hat{\theta}(x) - \theta(x)] = \sigma^2_\theta - \sum_{i=1}^{m} \lambda_i(x)Cov_{\theta\theta}(x, x_i) - \sum_{k=1}^{K}\sum_{j=1}^{nk} \mu_{kj}(x)Cov_{\theta u_k}(x, x_{kj}) \qquad (39)$$

On the right-hand side of (39), the first term is the unconditional uncertainty, that is, ψ_2 in (37), the second term is the reduction in uncertainty after conditioned by the parameter measurements, and the third term is the reduction in uncertainty after conditioned by the measurements of correlation fields.

There are several problems associated with the geostatistical inverse method described above. First, the estimation variance (39) is not reliable because it does not take account of the structure error. The form of the covariance structure may not be correctly assigned and the statistical parameters in the mean and covariance functions may not be well calibrated by the measured data. Second, although the identified parameter is distributed, its degree of freedom is low. Only a few statistical parameters (β and ψ) are adjusted to fit all measured data. As a result, the fitting residual of the MLE may be significant. To increase the degree of freedom of the parameter structure, a so-called *"pivot point" method* was presented (Ahmed and de Marsily, 1987). In this method, a term $\sum_{p=1}^{np} \lambda_p(x)\hat{\theta}(x_p)$ is added to the right-hand side of (38), where $\{x_p|p = 1, 2, \ldots, np\}$ is a set of pivot points, $\{\lambda_p(x)\}$ the kriging coefficients. Parameter values at these pivot points, $\{\hat{\theta}(x_p)\}$, are identified together with the statistical parameters by the MLE. Lavenue and Pickens (1992) included optimizing the locations of pivot points in the parameter identification procedure. This can be considered as an alternative method of adaptive parameterization.

Gomez–Hernandez *et al.* (1997) presented a similar methodology called "*the self-calibrating approach*". The unconditioned random field is first conditioned by m measurements $\{\theta^*(x_i)|i = 1, 2, \ldots, m\}$ to obtain a seed field. After a set of master locations $\{x_p|p = 1, 2, \ldots, np\}$ is assigned, a method for solving CIP is then used to estimate the unknown statistical parameters as well as the parameter values at the master locations. Assume that the best fitting parameter values at the master locations are $\tilde{\theta}(x_p) = \theta^*(x_p) + \Delta\theta(x_p)$, where $\theta^*(x_p)$ is the value of the original seed field. The seed field is then updated by kriging with all m measurement points $\{x_i\}$ and np master locations

$\{x_p\}$as the basis points, that is, let

$$\theta^*(x) = \sum_{i=1}^{m} \lambda_i(x)\theta^*(x_i) + \sum_{j=1}^{np} \lambda_j(x)\tilde{\theta}(x_p) \qquad (40)$$

There is no essential difference between the revised pivot point approach and the self-calibrating approach. When the unknown parameter is hydraulic conductivity, its measured values (obtained by fitting pumping test data with an analytical solution) may contain significant error. Because the parameter measurement values are regarded as "hard" data, their errors will be finally transferred into the estimated parameter structure. Moreover, when measurement data are limited and the number of pivot points is not appropriately controlled, the over-parameterization problem may occur. Kitanidis (1999) suggested using the generalized covariance functions for structure description that can produce the flattest estimate of a distributed parameter. The geostatistical method was presented in the 1980s for steady-state flow condition. After 20 years of development, it has been extended to transient flow and unsaturated flow conditions. Besides using head measurements, velocity, concentration, arriving time, water age, geophysics, and geology measurements have been also used as correlated data to condition the log-K field (Rubin *et al.*, 1992; Harvey and Gorelick, 1995; Yeh and Zhang, 1996; Hubbard and Rubin, 2000; Medina and Carrera, 2003).

Uncertainty Caused by Structure Error

The error of an identified PR (S, θ) consists of two parts: the structure error in S and the value error in θ. In Section "The reliability of parameter estimation", we only considered the effect of the value error. Now, we understand that the error of parameterization can never be avoided when the true parameter structure is complex and unknown. Without considering the effect of structure error, we may seriously underestimate the uncertainty of model prediction or the risk of model application. In the practice of groundwater modeling, we often see that there are different PRs (or more generally, there are different conceptual models) that can fit the existing data to the same extent (Sambridge, 2001). If there are K parameter structures or models $(S_1, S_2, \ldots, S_K)$ that cannot be rejected by the existing data, we may consider their reliability one by one using the methods mentioned in Section "The reliability of parameter estimation" and find the worst-case estimation. Recently, Neuman (2003) presented a *Maximum Likelihood Bayesian Model Averaging* (MLBMA) method to assess the uncertainty of model prediction in the statistical framework. Assume that the valid structure is one of the K possible structures, the posterior distribution of a prediction Δ for given data D is defined as the weighted average of the posterior distributions with respect to all models (Hoeting *et al.*, 1999):

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|S_k, D)p(S_k|D) \qquad (41)$$

From Bayesian theorem, the weights $p(S_k|D) = p(D|S_k)p(S_k)/\sum_{l=1}^{K} p(D|S_l)p(S_l)$, in which $p(D|S_k)$ is the likelihood of structure S_k. Once the posterior distribution (41) is obtained, the posterior mean $E(\Delta|D)$ and the posterior variance $Var(\Delta|D)$ can be calculated, and the latter gives the reliability estimation of model prediction. Neuman (2003) presented effective methods to calculate $p(\Delta|S_k, D)$ and $p(D|S_k)$ through the maximum likelihood estimation. When the number K is large or infinite as often seen in the parameter structure identification, the calculation of (41) will be infeasible.

STRUCTURE REDUCTION AND EXPERIMENTAL DESIGN

Identification of Representative Parameters

From a practical point of view, making the identified PR (S, θ) of an unknown parameter to be close to its true structure and its true values is generally impossible, cost-ineffective, and unnecessary. Instead of the trueness, we should emphasize on the reliability of model application and the cost-effectiveness of model construction. The GIP defined in Sun and Sun (2002) aims at finding a representative parameter (or a model after it is parameterized) that has the simplest structure, and yet can produce reliable results for given model applications. The basic idea is to offset the error in parameter structure by the error in parameter values. Let $g_E(M)$ be a set of given model applications (predictions, managements, or decisions), where M represents a model or a PR (S, θ). The reliability of model application can be represented by

$$\|g_E(M) - g_E(M_t)\|_E < \varepsilon \qquad (42)$$

where $\|\cdot\|_E$ is a norm defined in the model application space, M_t is the true parameter, ε is an accuracy requirement. The GIP requires finding the *simplest* parameter structure S^* and its associated model parameter θ^* in the admissible region such that (42) is satisfied when PR $M^* = (S^*, \theta^*)$ is used as $M = (S, \theta)$ in that equation. Therefore, the parameter identified by solving GIP is a representative one for the specified model application. The so-defined GIP has the following advantages: first, the reliability of model application is incorporated into the identification procedure; second, the parameter identification problem is replaced by a weak requirement (42). This requirement may be satisfied by parameters that are not close to the true parameter in the

parameter space; third, the data requirement is minimized because GIP attempts to find the simplest representative parameter structure. The complexity of parameter structure is determined by the requirement of model application. Once the complexity is determined, the sufficiency of existing data can be assessed.

The true parameter is data-independent, but the solutions of CIP and EIP discussed before are data-dependent, that is, different parameter structures and values may be obtained for the same parameter when different data sets are used for parameter identification. To make the identified parameter to be independent of data, the cross-validation method is often used: using one set of data for parameter identification and another set of data for validation. This method is often ineffective in groundwater modeling because of data limitation. As a result, most of groundwater models used in real case studies are data-dependent models. Their reliability for different model applications is not guaranteed. The solutions of GIP, on the other hand, are application dependent. The following is a brief description of the stepwise regression method presented by Sun *et al.* (1998) for solving the GIP. Using the nested structure sequence defined in (36), for each complexity level m, we first solve the EIP to find the optimal PR and its fitting residual:

$$RE_m = \min_{S_m,\theta_m} \{\|u_D^{\text{obs}} - u_D^{\text{cal}}(S_m,\theta_m)\|_D + \lambda\|(S_m,\theta_m) - (S_m^0,\theta_m^0)\|_P\} \tag{43}$$

where u_D^{obs} denotes the observed state values based on an experimental design D. We then calculate the maximum model application error introduced by replacing S_m with S_{m-1}, which is defined by

$$AE_m = \max_{\theta_m} \min_{\theta_{m-1}} \|g_E(S_m,\theta_m) - g_E(S_{m-1},\theta_{m-1})\|_E, \quad \text{s.t. } \theta_{m-1} \in \Theta(S_{m-1}), \theta_m \in \Theta(S_m) \tag{44}$$

Let us consider three cases. (a) If both AE_m and RE_m are large, we increase m to $m+1$. S_{m+1} is obtained by dividing a zone of S_m, which is the most sensitive one to the specified model applications, into two zones. The boundary between the two zones is determined by minimizing the fitting residual RE_{m+1}. (b) If AE_m is small, stop and use the EIP solution $(\hat{S}_m, \hat{\theta}_m)$ as the identified parameter. (c) If AE_m is large but RE_m is small, new data need to be collected.

The maximum application error defined in (44) is a special case of the structure error defined in Sun (1994) and Sun *et al.* (1998). Letting (S_A,θ_A) and (S_B,θ_B) be two different PRs of a distributed parameter $\theta(x)$ in its admissible region Θ, the distance between them can be measured in parameter, observation, and model application spaces by d_p, d_D, and d_E, respectively. The overall distance d between the two PRs is defined by $d = d_E + \mu d_D + \lambda d_p$, where μ and λ are weighting coefficients. A PR (S_B,θ_{AB}) is called a *projection* of (S_A,θ_A) onto the structure S_B, when

$$\theta_{AB} = \arg \min_{\theta_B} d(S_A,\theta_A;S_B,\theta_B), \text{ s.t.}\theta_B \in \Theta(S_B) \tag{45}$$

Finding θ_{AB} from (45) is equivalent to solving a classical inverse problem. The *structure error* $SE(S_A,S_B)$ resulted from using parameter structure S_B to replace parameter structure S_A is defined by the following max-min problem:

$$SE(S_A,S_B) = \max_{\theta_A} \min_{\theta_B} d(S_A,\theta_A;S_B,\theta_B), \quad \text{s.t. } \theta_A \in \Theta(S_A), \theta_B \in \Theta(S_B) \tag{46}$$

If we take $\lambda = 0$ and $\mu = 0$ in the definition of distance, $SE(S_m,S_{m-1})$ reduces to the AE_m in (44). Let us introduce the following concept (Sun, 2005): A PR $(S_A,\tilde{\theta}_A)$ is called the *worst-case parameter* (WCP) for simplifying a structure S_A to a structure S_B, if

$$SE(S_A,S_B) = \min_{\theta_B} d(S_A,\tilde{\theta}_A;S_B,\theta_B), \text{ s.t. } \theta_B \in \Theta(S_B) \tag{47}$$

When we know WCP, the structure error can be obtained by solving a min problem (47) rather than a max-min problem (46). WCP is a parameter that is associated with the maximal error when the model structure is reduced. It cannot be located in any inner point of the admissible region $\Theta(S_A)$. *If $\Theta(S_A)$ is a multidimensional box defined by the upper and lower bounds of each parameter component, then WCP must be located at a vertex of the multidimensional box* (Sun, 2005). With this proposition, the genetic algorithm (GA) becomes very suitable for searching the WCP because each vertex can be encoded as a binary string with 0 corresponding to the lower bound and 1 to the upper bound of a parameter component. Note that WCP depends on the flow conditions and may not be unique. Once we have an effective method for calculating the structure error, the stepwise regression procedure for solving the GIP becomes more effective.

Robust Experimental Design

The objective of experimental design is either to provide the maximum information with a certain budget, or to provide certain information with the minimum budget. For groundwater modeling, the decision variables of design include (i) how to excite the system (pumping locations and rates), and (ii) how to observe the responses of the system (observation locations and frequencies). In statistics, the theory of experimental design has been well developed for linear models for which the observation equation can be represented by $u_D^{\text{obs}} = \mathbf{A}_D\theta + \varepsilon_D$, where $\mathbf{A}_D$ is an $n \times m$ matrix. If the observation error ε_D is

normally distributed with zero mean and variance σ^2, the covariance matrix of the estimated parameter $\hat{\theta}$ is given by $Cov(\hat{\theta}) = \sigma^2(\mathbf{A}_D^T\mathbf{A}_D)^{-1}$. The determinant of the covariance matrix is a measurement of the *uncertainty* of the estimated parameter. The matrix $\mathbf{A}_D^T\mathbf{A}_D$ is called the *information matrix*. To maximize a norm of the information matrix means to minimize the uncertainty of the estimated parameter. For example, the *D-optimal design criterion* finds a design D^* from all admissible designs $\{D_{ad}\}$ such that the determinant of $\mathbf{A}_D^T\mathbf{A}_D$ is maximized (or the volume of the corresponding confidence ellipsoid is minimized). Therefore, for a linear model, the optimal design decision variables can be obtained by solving an optimization problem.

A nonlinear model can be linearized by first-order approximation in the neighborhood of the unknown parameter. The sensitivity matrix $\mathbf{J}_D = [\partial u_{\mathbf{D}}^{\text{obs}}/\partial\theta]$ is used to approximate $\mathbf{A}_D$ and the matrix $\mathbf{J}_D^T\mathbf{J}_D$ is used to approximate the information matrix. All the elements of the sensitivity matrix should be evaluated at the true parameter. As a result, for nonlinear problems, the D-optimal design depends on the parameter to be estimated. In this case, a sequential design process is needed. The initial design is based on the initial guess of the unknown parameter, after the design is conducted in the field and the data are collected, a new design is obtained on the basis of the updated parameter. Note that a design obtained by the D-optimal criterion is neither necessary nor sufficient for the purpose of model application.

We have learned that the problem of parameter identification actually is a problem of how to transfer information from observation data to the estimated parameter. No identification method is effective if the quantity and quality of data are not sufficient. But how sufficient is sufficient? For a distributed parameter, the parameter structure error caused by parameterization can never be avoided. A parameter with more complicated structure needs more data to identify and vice versa. Therefore, the identification of parameter structure must be considered in the design stage. The problem of observation design for parameter structure identification is a new and interesting topic. Without knowing the complexity of the estimated parameter, we cannot decide whether the data provided by a design is sufficient. If we do not know how to determine the sufficiency of a design, how can we say a design is the optimal one? The solution of GIP provides important insight into experimental design because it can determine whether the existing data are sufficient for identifying a useful representative model. If the existing data are insufficient, we have to collect more data. A more challenging problem is thus presented: can we determine the sufficiency of an experimental design before it is actually conducted in the field? An experimental design that can provide sufficient information for identifying all parameters in the admissible region is called a *robust design.* We can prove that *if a design is sufficient for a WCP, it must be a robust one*. Once a WCP is found by GA, a robust design can be found through a heuristic procedure, in which the pumping rates, numbers of observation wells, and frequencies are increased gradually until the design becomes sufficient for identifying a representative parameter of the WCP. A numerical example of robust design can be found in Sun (2005).

SOFTWARE AND APPLICATIONS

Available Software Packages

The Levenberg–Marqardt algorithm is used in most parameter identification codes of groundwater modeling, such as PEST developed by Doherty (2000), UCODE developed by USGS (Poeter and Hill, 1999), and the coupled inverse solution code developed by Sun (1994). PEST has been incorporated into several packages of groundwater modeling (GMS, PMWIN, VMF, and others, see `www.scisoftware.com`), in which MODFLOW developed by USGS is used to solve the FP. UCODE has been incorporated into MODFLOW2000 and PMWIN. Parameters used in the flow model, such as hydraulic conductivity, storage coefficient, and recharge rates can be identified separately or simultaneously. In PEST, the sensitivity coefficients are calculated by the finite difference approximation, while in UCODE, the sensitivity equation method is used. In all the above-mentioned software packages, values of objective function, fitting curves, Jacobian, and correlation between parameters can be displayed during the inverse solution procedure. The code developed by Sun (1994) can identify both flow and mass transport parameters simultaneously for groundwater quality modeling. All of these packages can only solve the CIP, that is, the unknown parameter is parameterized by the zonation method with a given pattern. Changing the zonation pattern by hand is a very time-consuming work. Therefore, codes that can identify the parameter pattern automatically are badly needed.

Applications

Most studies of inverse solution in groundwater modeling and all the methods reviewed in this article are aimed at identification of hydraulic conductivity in confined and unconfined aquifers. After 40 years of study, this is still an open problem for highly heterogeneous formations. The methodologies for solving EIP and GIP have not been extensively used in real case studies. During the last decade, using software to solve CIP for model calibration, however, has become more and more popular. Besides hydraulic conductivity, we can identify storage coefficient and specific yield when transient head measurements are

available. Boundary inflow, infiltration, and recharge rates, leaky conductance, and other sink and source terms in the flow equations can also be identified. Generally speaking, any property or any part of a model can be identified provided (i) the property is appropriately parameterized, and (ii) there are data available that are sensitive to the property.

In an unsaturated zone, the hydraulic conductivity is a function of pressure head ψ or a function of water content θ. It is generally expressed by an empirical formula other than parameterized by a set of basis functions. The most often used expression is

$$K(\theta) = K_s S_e^{\lambda}[1 - (1 - S_e^{1/m})^m]^2 \tag{48}$$

where $S_e = (\theta - \theta_r)/(\theta_s - \theta_r)$, $\theta(\psi) = \theta_r + (\theta_s - \theta_r)[1 + |\alpha\psi|^n]^{-m}$, and $m = 1 - 1/n$ (**Chapter 150, Unsaturated Zone Flow Processes, Volume 4**). The function identification problem is then reduced to identifying six constant parameters: $p = (\theta_s, \theta_r, \alpha, n, K_s, \lambda)$. When concentration measurements are available, the dispersion and reaction parameters can be identified simultaneously with the hydraulic parameters. Abbasi *et al.* (2003) shows that n and θ_s in p are the most sensitive parameters to water content, while θ_s and the longitudinal dispersion coefficient D_L are the most sensitive parameters to concentration. Because of the nonlinearity of the unsaturated flow model, the least squares criterion may have many local minima, and it is better to use global optimization algorithms for inverse solution. Altmann–Dieses *et al.* (2002) presented an optimal experimental design method for identifying the parameters in unsaturated zones.

To study groundwater contamination caused by organic compounds, we need to develop models to simulate various reactive transport processes in multiphase flow. The governing equations for this case are a set of partial differential equations that describe the mass balance of each component in each phase. There are a lot of physical, chemical, and biological parameters that must be identified before a model can be used for prediction and design purposes. In recent years, many authors used the least squares criterion to find the best fitting between model outputs and concentration data obtained from experiments (Gramling *et al.*, 2002). The fitting procedure usually contains two steps: first, using tracer test data to identify flow and dispersion parameters, and then fitting the reactive transport data to identify mass exchange, reaction, and decay term. If the model outputs cannot fit the observed data well, more complicated expressions containing more unknown parameters may be used in the reaction model. For a nonlinear kinetic process, mass exchange, reaction, and decay terms are unknown functions of concentrations. The solution of EIP for identifying the functional structure, thus, should be considered. When the number of unknown parameters increases, we must consider the identifiability problem caused by possible correlation between parameters. For example, Sun *et al.* (2001) shows that only two of four parameters characterizing the attachment/detachment process of colloids are identifiable. The study for identifying complex chemical and biological parameters is being developed.

Locating contaminant sources and recovering their release history from the measurements of contaminant plumes are critical for remediation design and environmental litigation support. Three problems were considered in groundwater references: (i) find the release history when a source location is known; (ii) find the location of a source when its release history is known; and (iii) find both source locations and release history. Pollution source identification is a kind of inverse problem. Therefore, all the criteria and methodologies reviewed in the previous sections can be used once the identified source is parameterized. The unknown parameter vector θ associated with a source may consist of its coordinates (x_s, y_s) and the released concentrations $\{C_s(t_j)|j = 1, 2, \ldots, m\}$, where m is the number of times or time periods. When the weighted least squares criterion is used, the objective function of optimization is:

$$E(\theta) = \sum_{k=1}^{K}\sum_{l=1}^{L} w_{kl}[C_{kl}^{\mathrm{cal}}(\theta) - C_{kl}^{\mathrm{obs}}]^2, \theta = (x_s, y_s, C_s) \tag{49}$$

where $\{C_{kl}^{\mathrm{obs}}\}$ are observed values of the plume at K observation times and L locations, $\{C_{kl}^{\mathrm{cal}}(\theta)\}$ are the corresponding model output. The objective function based on WLSE criterion is highly nonlinear and usually nonconvex. Many authors have designed different optimization algorithms, including evolution algorithms, to find the minimum of E(θ) (Mahar and Datta, 2001; Aral *et al.*, 2001). By considering the ill-posed nature of the source identification problem, Skaggs and Kabala (1994) added a regularization term to the objective function to stabilize the inverse solution. Atmadja and Bagtzoglou (2001) used the backward beam equation for source identification, in which the advection transport is reversed in time but the dispersion transport is kept positive because that cannot be reversed. This method can assess the relative importance of each potential source. Neupauer and Wilson (1999) obtained the backward-in-time distribution of a potential source by solving the adjoint state equation that, actually, is equivalent to the solution of the backward beam equation. Snodgrass and Kitanidis (1997) and Michalak and Kitanidis (2003) used the geostatistical method for contaminant history recovery that can give also the variance of estimation error. Up to date, development of accurate and effective methods for source identification is still an open problem.

CONCLUSION

Inverse problems in groundwater modeling can be defined in both deterministic and statistical frameworks. Bayesian inference provides the most general method for parameter estimation. It can incorporate not only prior information but also observation and model errors into the estimation process. Several software packages are available for solving the classical inverse problems in groundwater modeling. Methods that can automatically identify both parameter structure and values from prior information and observed data are badly needed. Because the structure of a real aquifer is usually very complex and unknown, and the available data are always very limited in both quantity and quality, finding a representative model is the only feasible way in groundwater modeling. The generalized inverse problem does not require the uniqueness of inverse solution. Instead, it requires the reliability of model application. Collecting sufficient data is the key of successfully solving the inverse problem. The problem of identifying hydraulic parameters of highly heterogeneous aquifers is still not well resolved. The identification of chemical and biological reaction functions and the identification of pollution sources are being studied.

REFERENCES

Abbasi F., Jacques D., Simunek J., Feyen J. and van Genuchten M.T. (2003) Inverse estimation of soil hydraulic and solute transport parameters from transient field experiments: heterogeneous soil. *Transactions of the American Society of Civil Engineers*, **46**(4), 1097–1111.

Ahmed S. and de Marsily G. (1987) Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, **23**(9), 1717–1737.

Akaike H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Altmann-Dieses A.E., Schloder J.P., Bock H.G. and Richter O. (2002) Optimal experimental design for parameter estimation in column outflow experiments. *Water Resources Research*, **38**(10), 1186, doi:10.1029/2001WR001149.

Ameur H., Chavent G. and Jaffre J. (2002) Refinement and coarsening indicators for adaptive parameterization: application to the estimation of hydraulic transmissivities. *Inverse Problems*, **18**, 775–794.

Aral M.M., Guan J.B. and Maslia M.L. (2001) Identification of contaminant source location and release history in aquifers. *Journal of Hydrologic Engineering*, **6**(3), 225–234.

Atmadja J. and Bagtzoglou A.C. (2001) Pollution source identification in heterogeneous porous media. *Water Resources Research*, **37**(8), 2113–2125.

Bard Y. (1974) *Nonlinear Parameter Estimation*, Academic: San Diego.

Barhen J. and Reister D.B. (2003) Uncertainty analysis on sensitivities generated using automatic differentiation. *Computational Science and its Applications*, ICCSA: PT 2, pp. 70–77.

Bear J. (1979) *Hydraulics of Groundwater*, McGraw-Hill: New York.

Bravo H.R., Jiang F. and Hunt R.J. (2002) Using groundwater temperature data to constrain parameter estimation in a groundwater flow model of a wetland system. *Water Resources Research*, **38**(8), 1153, doi:10.1029/2001WR000496.

Carrera J. and Neuman S.P. (1986) Estimation of aquifer parameters under transient and steady state conditions, 1. Maximum likelihood method incorporating prior information. *Water Resources Research*, **22**(2), 199–210.

Chavent G., Dupuy M. and Lemonnier P. (1975) History matching by use of optimal control theory. *Society of Petroleum Engineers Journal*, **15**(1), 74–86.

Cherkassky V. and Ma Y.Q. (2003) Comparison of model selection for regression. *Neural Computation*, **15**(7), 1691–1714.

Doherty J. (2000) PEST-model-independent parameter estimation. *User Manual*, Watermark Computing.

Elizondo D., Cappelaere B. and Faure C.h (2002) Automatic versus manual model differentiation to compute sensitivities and solve non-linear inverse problems. *Computers and Geosciences*, **28**(3), 309–326.

Giacobbo F., Marseguerra M. and Zio E. (2002) Solving the inverse problem of parameter estimation by genetic algorithms: the case of groundwater contaminant transport model. *Annals of Nuclear Energy*, **29**(8), 967–981.

Gilks W.R., Richardson S. and Spiegelhalter D.J. (Eds.) (1996) *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.

Gomez-Hernandez J.J., Sahuquillo A. and Capilla J.E. (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 1. Theory. *Journal of Hydrologic Engineering*, **203**(1–4), 162–174.

Gramling C.M., Harvey C.F. and Meigs L.C. (2002) Reactive transport in porous media: a comparison of model prediction with laboratory visualization. *Environmental Science and Technology*, **36**(11), 2508–2514.

Harvey C.F. and Gorelick S.M. (1995) Mmapping hydraulic conductivity – sequential conditioning with measurements of solute arrival time, hydraulic-head, and local conductivity. *Water Resources Research*, **31**(7), 1615–1626.

Heredia J., Medina A. and Carrera J. (2000) Estimation of parameter geometry. In *Computational Methods for Flow and Transport in Porous Media*, Crolet J. (Ed.) Kluwer Academic Publishers.

Hoeting J.A., Madigan D., Raftery A.E. and Volinsky C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, **14**(4), 382–417.

Hubbard S.S. and Rubin Y. (2000) Hydrogeological parameter estimation using geophysical data: a review of selected techniques. *Journal of Contaminant Hydrology*, **45**(1–2), 3–34.

Isakov V. (1998) *Inverse Problems for Partial Differential Equations*, Springer-Verlag: New York.

Kelley C.T. (1999) *Iterative Methods for Optimization*, SIAM: Philadelphia.

Kitanidis P.K. (1999) Generalized covariance functions associated with the laplace equation and their use in interpolation and inverse problems. *Water Resources Research*, **35**(5), 1361–1367.

Kunstmann H., Kinzelbach W. and Siegfried T. (2002) Conditional first-order second-moment method and its application to the quantification of uncertainty in groundwater modeling. *Water Resources Research*, **38**(4), 15, Art. No. 1035.

Lahiri S.N. (2003) *Resampling Methods for Dependent Data*, Springer-Verlag: New York.

Lavenue A.M. and Pickens J.F. (1992) Application of a coupled adjoint-sensitivity and kriging approach to calibrate a groundwater flow model. *Water Resources Research*, **28**(6), 1543–1569.

Lin J. and Zhang G.Q. (2004) Transmissivity estimation for a two-dimensional aquifer by regularization of potential and streamline functions. *Inverse Problems*, **20**(2), 331–346.

Liu J. (1994) A sensitivity analysis for least-squares ill-posed problems using the haar basis. *Siam Journal on Numerical Analysis*, **31**(5), 1486–1496.

Mahar P.S. and Datta B. (2001) Optimal identification of groundwater pollution sources and parameter estimation. *Journal of Water Resources Planning and Management-Asce*, **127**(1), 20–29.

McLaughlin D. and Townley L.R. (1996) A reassessment of the groundwater inverse problem. *Water Resources Research*, **32**(5), 1131–1162.

Medina A. and Carrera J. (2003) Geostatistical inversion of coupled problems: dealing with computational burden and different types of data. *Journal of Hydrology*, **281**(4), 251–264.

Michalak A.M. and Kitanidis P.K. (2003) A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification. *Water Resources Research*, **39**(2), 1033, doi:10.1029/2002WR001480.

Morshed J. and Kaluarachchi J.J. (1998) Parameter estimation using artificial neural network and genetic algorithm for free-product migration and recovery. *Water Resources Research*, **34**(5), 1101–1113.

Neuman S.P. (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, **17**, 291–305.

Neupauer R.M. and Wilson J.L. (1999) Adjoint method for obtaining backward-in-time location and travel time probabilities of a conservative groundwater contaminant. *Water Resources Research*, **35**(11), 3389–3398.

Poeter E.P. and Hill M.C. (1999) UCODE, a computer code for universal inverse modeling. *Computers & Geosciences*, **25**(4), 457–462.

Portniaguine O. and Solomon D.K. (1998) Parameter estimation using groundwater age and head data, cape cod, Massachusetts. *Water Resources Research*, **34**(4), 637–645.

Rubin Y., Mavko G. and Harris J. (1992) Mapping permeability in heterogeneous aquifers using hydrological and seismic data. *Water Resources Research*, **28**(7), 1792–1718.

Sambridge M. (2001) Finding acceptable models in nonlinear inverse problems using a neighborhood algorithm. *Inverse Problems*, **17**, 387–403.

Schwarz G. (1978) Estimating the dimension of a model. *Annals of the Institute of Statistical Mathematics*, **6**, 461–464.

Shigidi A. and Garcia L.A. (2003) Parameter estimation in groundwater hydrology using artificial neural networks. *Journal of Computing in Civil Engineering*, **17**(4), 281–289.

Skaggs T.H. and Kabala Z.J. (1994) Recovering the release history of a groundwater contaminant. *Water Resources Research*, **30**(1), 71–79.

Snodgrass M.F. and Kitanidis P.K. (1997) A geostatistical approach to contaminant source identification. *Water Resources Research*, **33**(4), 537–546.

Sohn M.D., Small M.J. and Pantazidou M. (2000) Reducing uncertainty in site characterization using Bayes Monte Carlo method. *Journal of Environmental Engineering, ASCE*, **126**, 893–902.

Solomatine D.P., Dibike Y.B. and Kukuric N. (1999) Automatic calibration of groundwater models using global optimization techniques. *Hydrological Sciences Journal*, **44**(6), 879–894.

Sun N.-Z. (1994) *Inverse Problems in Groundwater Modeling*, Kluwer Academic Publishers.

Sun N.-Z. (2005) Structure reduction and robust experimental design for distributed parameter identification. *Inverse Problems*, **21**(4), 739–758.

Sun N.-Z. and Sun A.Y. (2002) Parameter identification of environmental systems, chapter 9. In *Environmental Fluid Mechanics: Theories and Applications*, Shen H.H. *et al.* (Ed.), ASCE: Reston, Virginia.

Sun N., Sun N.-Z., Elimelech M. and Ryan J.N. (2001) Sensitivity analysis and parameter identifiability for colloid transport in geochemically heterogeneous porous media. *Water Resources Research*, **37**(2), 209–222.

Sun N.-Z., Yang S.-L. and Yeh W.W.-G. (1998) A proposed stepwise regression method for model structure identification. *Water Resources Research*, **34**(10), 2561–2572.

Sun N.-Z. and Yeh W.W.-G. (1985) Identification of parameter structure in groundwater inverse problems. *Water Resources Research*, **21**(6), 869–883.

Sun N.-Z. and Yeh W.W.-G. (1990) Coupled inverse problem in groundwater modeling, 1, sensitivity analysis and parameter identification. *Water Resources Research*, **26**(10), 2507–2525.

Tang D.H. and Pinder G.F. (1979) Analysis of mass transport with uncertain physical parameters. *Water Resources Research*, **15**(5), 1147–1155.

Tikhonov A.N. and Arsenin V.Y. (1977) *Solution of Ill-Posed Problems*, Winston: New York.

Tsai F.T.-C., Sun N.-Z. and Yeh W.W.-G. (2003) Global-local optimization for parameter structure identification in three-dimensional groundwater modeling. *Water Resources Research*, **39**(2), 1043.

Vapnik V. (1998) *Statistical Learning Theory*, Wiley: New York.

Vogel C.R. (2002) *Computational Methods for Inverse Problems*, *Frontiers in Applied Mathematics: 23*, SIAM: Philadelphia.

Wackernagel H. (2003) *Multivariate Geostatistics: An Introduction with Applications*, *Third Edition*, Springer-Verlag.

Yeh W.W.-G. (1986) Review of parameter identification procedures in groundwater hydrology: the inverse problem. *Water Resources Research*, **22**(2), 95–108.

Yeh J.T.-C. and Zhang J. (1996) A geostatistical inverse method for variably saturated flow in the vadose zone. *Water Resources Research*, **32**(9), 2757–2766.

Zheng C. and Wang P. (1996) Parameter structure identification using tabu search and simulated annealing. *Advances in Water Resources*, **19**(4), 215–224.

157: Sea Water Intrusion Into Coastal Aquifers

JACOB BEAR

Department of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa, Israel

The objective of this article is to present the problem of seawater intrusion into a coastal aquifer and the conceptual and complete mathematical models that describe it. Three seawater intrusion models are presented: two- and three-dimensional sharp interface models and a three-dimensional transition zone model.

INTRODUCTION

In many parts of the world, coastal aquifers constitute an important source of water. Often, coastal areas are also heavily populated, a fact that makes the demand for freshwater even more acute. However, the proximity and contact with the sea requires special attention when planning the management of such aquifers, and the use of models to facilitate decision-making. In fact, in many coastal aquifers, intrusion of seawater has become one of the major constraints affecting groundwater management. As seawater intrusion progresses, the part of the aquifer close to the sea becomes saline. As a consequence, pumping wells that exist close to the coast become saline and have to be abandoned. Also, the area above the intruding seawater wedge is lost as a source of freshwater (by natural replenishment).

Since the famous works of Badon-Ghyben (1888) and Herzberg (1901), and the less-known work of Du Commun (Mentioned in Konikow, L. F. and Reilly, T. E. Seawater intrusion in the united states, in Bear *et al.*, 1999) (1828), extensive research has been carried out, leading to the understanding of the mechanisms that govern seawater intrusion. The dominant factors are the flow regime in the aquifer above the intruding seawater wedge, the variable density, and hydrodynamic dispersion. Reviews of the phenomenon of seawater intrusion, and of the research that has been carried out on this subject, both theoretical work, and field and laboratory investigations, may be found in many books and publications, and will not be repeated here (e.g. Bear, 1972, 1979; Bear and Verruijt, 1987; Reilly and Goodman, 1980; and Bear *et al.*, 1999).

Briefly, under normal conditions in a coastal aquifer, the sea is the recipient of freshwater excess, that is, natural and artificial recharge minus pumping. This means that a seaward hydraulic gradient exists in the aquifer. Due to the presence of seawater in the aquifer under the sea, a zone of contact is formed between the lighter freshwater flowing toward the sea and the heavier seawater in the aquifer. Typical cross sections, with interfaces under natural conditions are shown in Figure 1. In this figure, the surface indicated as an "interface" represents an "interface zone" (Figure 2). The detailed shape of the transition zone depends also on whether this zone is advancing inland or retreating. In all cases, the domain in the aquifer that is occupied by seawater has the form of an advancing or receding wedge. One should note that, like all figures that describe aquifers, these are also highly distorted figures, not drawn to scale.

Seawater and freshwater are often referred to as "miscible liquids", although, actually, both constitute a *single* liquid phase – water – with different concentrations of *total dissolved salt* (salt, TDS). For the sake of simplicity, we shall continue to refer to them as two liquids – fresh water, and seawater. Hence, the passage from the portion of the aquifer that is occupied by the former to that occupied by the latter, takes the form of a *transition zone*, rather than a *sharp interface*. Under certain circumstances, depending on the extent of seawater intrusion, and on certain aquifer properties, this transition zone, which is, primarily, a result of *hydrodynamic dispersion* of the dissolved matter, may be rather wide. Under other conditions, it may be rather narrow, relative to the aquifer's thickness, and the passage from the zone occupied by freshwater to that occupied by

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

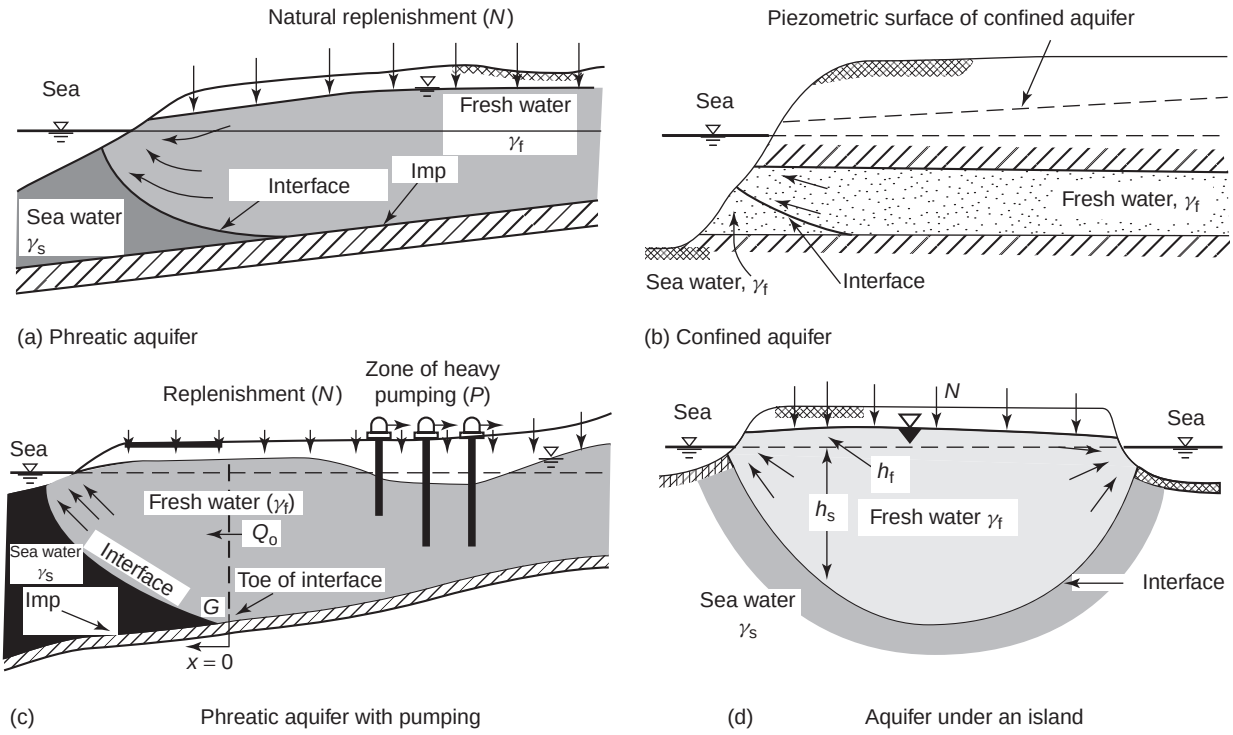


Figure 1 Typical cross sections of a seawater intrusion in coastal aquifers (Bear, 1979). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

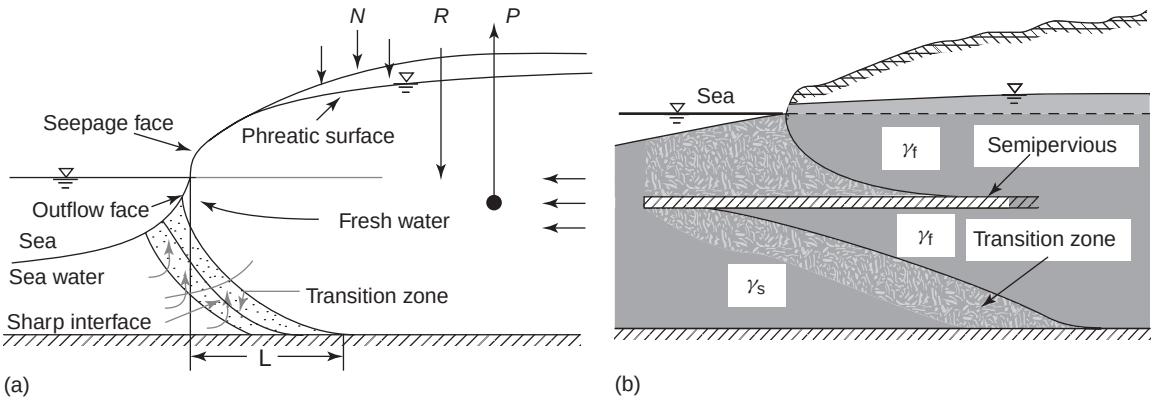


Figure 2 Transition zones in a single layer and in a multilayered coastal aquifer. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

seawater may be *approximated* as a *sharp interface*. Often, the term "interface" is used for the iso-density surface that is midway between fresh and seawater. In this article, the term "interface" will sometimes be used interchangeably with "transition zone".

Under natural undisturbed conditions in a coastal aquifer, a state of equilibrium is maintained, with a stationary interface zone and with a zone of freshwater flow above it. Under these (unrealistic) assumptions of a stationary sharp interface, the seawater is immobile, while the freshwater is flowing seaward above the interface. In reality, the transition zone is fed from below by seawater (Figure 2). When water is pumped from a coastal aquifer, the discharge to the sea is reduced, water levels (or the piezometric head in a confined aquifer) close to the sea are lowered and the transition zone rises. The entire seawater and transition zone wedge advance landward, until a new equilibrium is reached. Wells that operate within the wedge zone pump saline water and have to be abandoned. When pumping takes place in a well located above the transition zone,

the latter *upcones* toward the well. Unless the well is at a sufficient distance above this zone and/or the rate of pumping is sufficiently small, the well will eventually pump sea, or saline water.

As already indicated above, the extent of seawater intrusion, measured, say, by the length of the intruding wedge on the aquifer's bottom, is related to the rate of freshwater drained to the sea. *As this rate is reduced (by pumping), the extent of seawater intrusion increases.* This makes the problem of seawater intrusion into a coastal aquifer a *management problem.* In such a problem, we seek to determine the maximum annual net rate of pumping (i.e. pumping minus artificial recharge) from the aquifer such that a certain goal is achieved (e.g. maximize net benefits), without violating specified (legal, hydrologic, technical, etc.) constraints. Thus, in managing a coastal aquifer, the issue is not *to stem seawater intrusion*, but to determine the *optimal extent of seawater intrusion*. We wish to maximize pumping, in order to achieve certain, say, economic goals, without violating specified constraints. Models of flow, solute transport, and optimization are used in planning aquifer management. An essential constraint to be satisfied, as part of the optimization model, is the mathematical flow and solute transport model, which takes into account the dependence of water density on salt concentration (= variable density flow).

In the past, the *sharp interface approximation* was introduced, primarily, to enable relatively easy solutions, both analytical and numerical, of certain simple seawater intrusion problems of practical interest. However, nowadays, with the availability of new improved numerical techniques, including methods for coping with nonlinearities that are inherent in the transition zone model, and with fast and large memory computers (even PCs), numerical solutions of models that take the transition zone into account should not pose special difficulties. There is also no reason to limit the models to (vertical) two-dimensional flow domains. Indeed, a number of models and computer codes that consider seawater intrusion as a solute transport problem have already been developed (e.g. Konikow *et al.*, 1996). Plenty of information on seawater intrusion into coastal aquifers, the management problem, modeling with the sharp interface approximation, analytical solutions, modeling as a variable density flow and transport problem, numerical solutions, and discussions on specific numerical codes can be found in Bear *et al.* (1999).

THE 3-D SHARP INTERFACE MODEL

As emphasized earlier, except for cases in which the transition zone is relatively narrow, this is an unrealistic model. However, we present it because it gives some insight into certain features of the seawater intrusion problem.

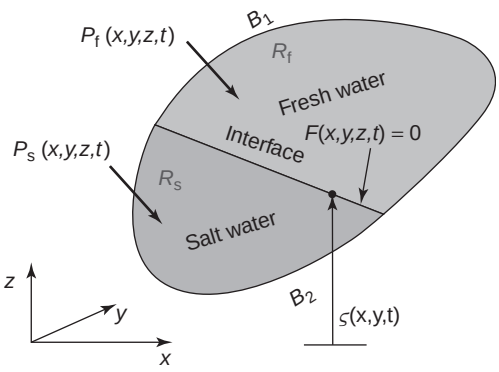


Figure 3 A sharp interface between regions occupied by freshwater and seawater. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The underlying conceptual model is that the freshwater and the seawater are two liquids of different density and viscosity, each occupying a *distinct* portion of the flow domain. The two portions (R_s and R_f in Figure 3) are separated by a sharp, possibly moving, interface. Sources and sinks of liquid (i.e. artificial recharge and pumping) may exist in both regions. In each of the two regions, we define a *piezometric head*, h (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**) in the form

$$h_\alpha = z + \frac{p}{g\rho_\alpha}, \text{ or, for a compressible fluid:}$$
$$h_\alpha = z + \int_{p_0}^{p} \frac{\mathrm{d}p}{g\rho_\alpha(p)} \tag{1}$$

where α denotes either fresh (f) or seawater (s), p denotes pressure, p_o denotes a reference pressure, ρ denotes the fluid's density, z denotes elevation above some datum level, and g denotes gravity acceleration.

Subject to certain simplifying assumptions, the mathematical model of the two-liquid flow problem can be stated as:

Determine h_f in R_f and h_s in R_s, such that they satisfy the two mass balance equations

$$(a)\quad S_{0f}\frac{\partial h_f}{\partial t} = -\nabla\cdot\mathbf{q}_f - P_f, \quad \mathbf{q}_f = -\mathbf{K}_f\cdot\nabla h_f, \quad \text{in } R_f;$$
$$(b)\quad S_{0s}\frac{\partial h_s}{\partial t} = -\nabla\cdot\mathbf{q}_s - P_s, \quad \mathbf{q}_s = -\mathbf{K}_s\cdot\nabla h_s, \quad \text{in } R_s \tag{2}$$

in which $P_\alpha = P_\alpha(x, y, z, t)$ is a symbol that represents sinks of the α-phase, $\mathbf{K}_\alpha \equiv \mathbf{k}\rho_\alpha g/\mu_\alpha$ is the hydraulic conductivity, $\mathbf{k}$ is the permeability (both tensors), μ is the dynamic viscosity, and $S_{0\alpha}(x, y)$ is the specific storativity, all of the α-phase (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4** for the development of (2) and the underlying assumptions). In addition, we have to specify

initial conditions. Boundary conditions for h_f on B_1 and h_s on B_2 are the usual ones encountered in the flow of a single fluid. However, the boundary condition on the interface separating the two fluids requires special attention. Moreover, similar to the case of a phreatic surface (**Chapter 149, Hydrodynamics of Groundwater, Volume 4**), the location and shape of the interface, which is expressed in the form $F = F(x, y, z, t) = 0$, is *a priori* unknown (until the problem is solved).

Denoting the elevation of points on the interface by $\zeta = \zeta(x, y, t)$, the relationship for F becomes

$$z = \zeta(x, y, t), \quad \text{or} \quad F \equiv z - \zeta(x, y, t) = 0 \tag{3}$$

The pressure at a point $P(x, y, \zeta)$ on the interface is the same when the latter is approached from both sides. Hence, from the definitions of h_f and $h_{s'}$, we obtain

$$\gamma_f(h_f - \zeta) = \gamma_s(h_s - \zeta), \quad \gamma_\alpha = \rho_\alpha g,$$
$$\zeta(x, y, t) = h_s(1 + \delta) - h_f\delta; \quad \delta = \frac{\gamma_f}{\gamma_s - \gamma_f} \tag{4}$$

Hence, once we know $h_f = h_f(x, y, z, t)$ and $h_s = h_s(x, y, z, t)$, in their respective domains, (4) becomes the sought equation for the shape of the interface, in the form

$$F(x, y, z, t) = z - h_s(1 + \delta) + h_f\delta = 0 \tag{5}$$

The boundary conditions on the interface defined by (5) – one for each side – are obtained from the fact that the interface is a *material surface* with respect to each of the fluids; no fluid mass crosses it. The two conditions are

$$(\mathbf{q}_{rf} - \phi\mathbf{u}) \cdot \mathbf{n} = 0, \quad (\mathbf{q}_{rs} - \phi\mathbf{u}) \cdot \mathbf{n} = 0, \quad (\mathbf{V}_\alpha - \mathbf{u}) \cdot \mathbf{n} = 0 \tag{6}$$

in which ϕ denotes the porosity, $\mathbf{q}_r$ denotes the specific discharge relative to the solid, $\mathbf{u}$ denotes the speed of displacement of the interface, and $\mathbf{n}$ is the unit vector normal to it, with

$$\frac{dF}{dt} \equiv \frac{\partial F}{\partial t} + \mathbf{u} \cdot \nabla F = 0, \quad \mathbf{n} = \frac{|\nabla F|}{\nabla F}, \quad \mathbf{u} \cdot \nabla F = -\frac{\partial F}{\partial t}$$

In (6), $(\mathbf{V}_\alpha - \mathbf{u}) \cdot \mathbf{n} = 0$, and $\mathbf{q}_{r\alpha} = \phi\,(\mathbf{V}_\alpha - \mathbf{V}_{\text{solid}})$, $\alpha = f, s$, is the specific discharge obtained from Darcy's law (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**). The conditions in (6) can be rewritten in the form

$$\phi\delta\frac{\partial h_f}{\partial t} - \phi(1 + \delta)\frac{\partial h_s}{\partial t} = (\mathbf{K}_f \cdot \nabla h_f) \times [\nabla z - (1 + \delta)\nabla h_s + \delta\nabla h_f] \tag{7}$$

for the R_f-region, and as an analogous equation for the R_s-region. These, *nonlinear* conditions indicate the coupling that exists between the two regions.

In principle, once the problem is solved for h_f in R_f and h_s in R_s, the shape of the interface can be determined by (5).

THE GHYBEN–HERZBERG APPROXIMATION

Consider the interface shown in Figure 4(a). The U-tube superimposed on this figure is intended to demonstrate the conceptual model proposed by Du Commun (1828), Badon-Ghyben (1888), and Herzberg (1901). Essentially, they assumed that a static equilibrium exists under steady-state conditions, with stationary seawater and a hydrostatic pressure distribution in the seaward flowing freshwater zone. This means that the flow is (essentially) horizontal

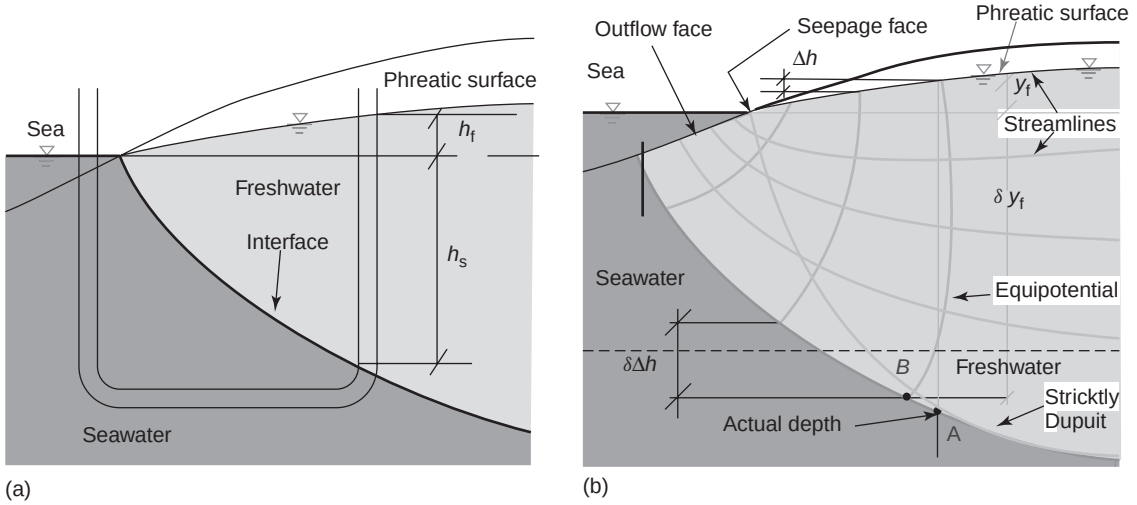


Figure 4 (a) The Ghyben–Herzberg approximation and (b) actual interface. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and equipotentials (= surfaces of equal piezometric head) are vertical lines or surfaces. This, in fact, is identical to the Dupuit assumption (*see* **Chapter 149, Hydrodynamics of Groundwater, Volume 4**). With the notation of Figure 4(a), we have

$$h_s = \delta h_f, \quad \delta = \frac{\rho_f}{\rho_s - \rho_f} \tag{8}$$

This means that at any distance from the sea, the depth of an *assumed stationary interface* below sea level is δ times the height of the freshwater table above it. Obviously, as the sea is approached, the assumption of essentially horizontal flow is no longer valid. Moreover, this assumption does not provide for an outflow surface through which freshwater can be drained to the sea. Figure 4(b) shows the actual flow conditions near the sea. We note the difference between the actual point (A) and the point (B) determined from the equipotential (= δh_f.), which is the depth predicted by the Ghyben–Herzberg relationship (8). The difference stems from the difference between the actual shape of the equipotentials and the vertical ones assumed by the Dupuit approximation. In a confined aquifer, h_s in (8) is the depth of a point on the interface below sea level, while h_f is the freshwater piezometric head. Bear and Dagan (1964) suggested that for steady flow, (8) gives an error that is less than 5% provided $\pi KB/\delta Q > 8$, where Q is the freshwater discharge to the sea and B is the (constant) thickness of a confined aquifer. In the case of a phreatic aquifer, the Ghyben–Herzberg approximation also overlooks the presence of the *seepage face* shown in Figure 4(b). Note that the interface always intersects the sea bottom as a line that is tangent to the vertical.

ESSENTIALLY HORIZONTAL FLOW MODEL

We continue to assume the existence of a sharp interface, but we introduce the *hydraulic approach*, on the basis of the (Dupuit) assumption of *essentially horizontal flow*. The flow equations are obtained by integrating the 3-D flow model presented earlier in the section 'The 3-D sharp interface model', separately for each region, over the vertical thickness of the respective region (Figure 5a).

For the freshwater region, we integrate 2(a) from the interface at $\zeta_1(x, y, t)$ to the phreatic surface with accretion, at $\zeta_2(x, y, t)$. For the salt water region, we integrate 2(b) from the aquifer's (impervious) bottom at $\zeta_0(x, y)$ to the interface, at $\zeta_1(x, y, t)$. The integration is based on the Leibnitz rule (Bear, 1979; Bear, 1999; Bear and Bachmat, 1990), which takes into account the conditions on the (possibly moving) boundary of integration. By integrating 2(a) and 2(b), we obtain for the freshwater region above the interface and for the seawater region below it, respectively,

$$\begin{aligned}\int_{\zeta_1}^{\zeta_2} \left(\nabla \cdot \mathbf{q}_f + S_{0f}\frac{\partial h_f}{\partial t} + P_f\right) dz = \nabla' \cdot B_f \widetilde{\mathbf{q}}_f' \\ + \mathbf{q}_f|_{F_2} \cdot \nabla F_2 - \mathbf{q}_f|_{F_1} \cdot \nabla F_1 \\ + S_{0f}\left(B_f \frac{\partial \widetilde{h}_f}{\partial t} + \widetilde{h}_f \frac{\partial B_f}{\partial t} + h_f|_{F_2} \cdot \frac{\partial F_2}{\partial t} - h_f|_{F_1} \cdot \frac{\partial F_1}{\partial t}\right) \\ + B_f \widetilde{P}_f = 0\end{aligned} \tag{9}$$

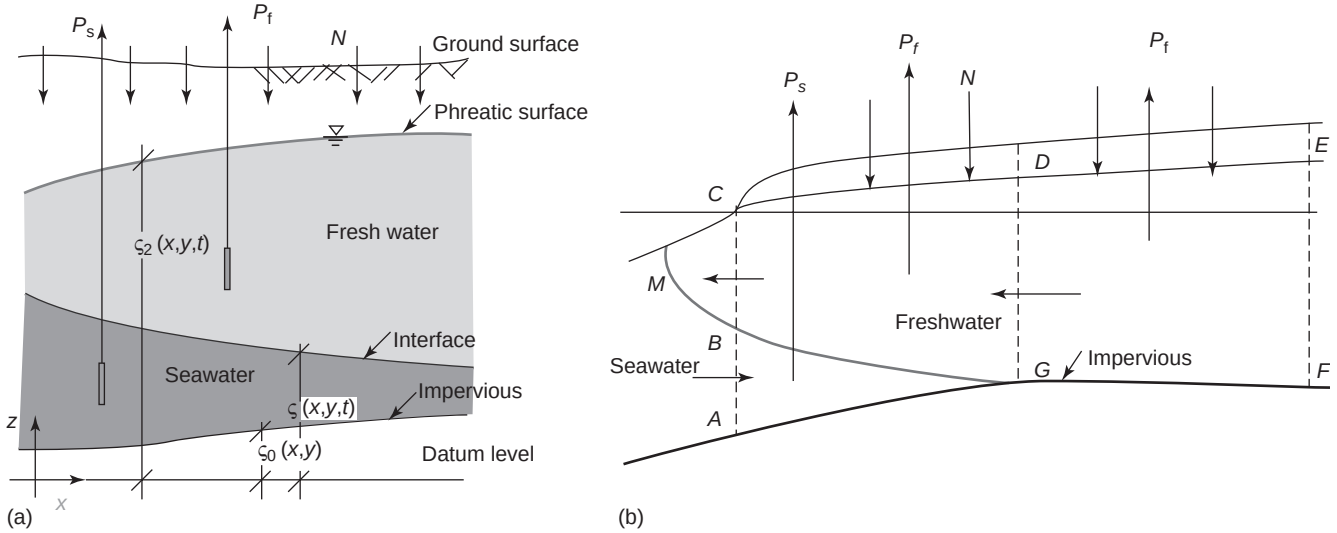


Figure 5 Nomenclature (a) for integrating over the thickness of sub-regions and (b) for boundary conditions. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

$$\int_{\zeta_0}^{\zeta_1}\left(\nabla\cdot\mathbf{q}_s+S_{0s}\frac{\partial h_s}{\partial t}+P_s\right)dz=\nabla'\cdot B_s\widetilde{\mathbf{q}}_s'+\mathbf{q}_s|_{F_1}\cdot\nabla F_1-\mathbf{q}_s|_{F_o}\cdot\nabla F_0+S_{0s}\left(B_s\frac{\partial\widetilde{h}_s}{\partial t}+\widetilde{h}_s\frac{\partial B_s}{\partial t}+h_s|_{F_1}\cdot\frac{\partial F_1}{\partial t}-h_s|_{F_0}\cdot\frac{\partial F_0}{\partial t}\right)+B_s\widetilde{P}_s=0 \tag{10}$$

in which $F_2=z-\zeta_2(x,y,t)$, $F_1=z-\zeta_1(x,y,t)$, $F_0=z-\zeta_0(x,y)$, ∇' denotes the gradient operator in the xy-plane, the prime in $\boldsymbol{q}'$ indicates the specific discharge vector in the xy-plane, $B_f=\zeta_2-\zeta_1$, $B_s=\zeta_1-\zeta_0$, and the tilde ($\widetilde{\ }$) symbol indicates the average over the relevant vertical length.

With $h_f|_{F_2}\simeq h_f|_{F_1}\simeq\widetilde{h}_f$, $h_s|_{F_o}\simeq h_s|_{F_1}\simeq\widetilde{h}_s$, which expresses the assumption of essentially horizontal flow in both domains, we obtain for the freshwater and saltwater domains, respectively,

$$\begin{aligned}\nabla'\cdot B_f\widetilde{\mathbf{q}}_f'+\mathbf{q}_f|_{F_2}\cdot\nabla F_2-\mathbf{q}_f|_{F_1}\cdot\nabla F_1+S_{of}B_f\frac{\partial\widetilde{h}_f}{\partial t}+B_f\widetilde{P}_f=0,\\ \nabla'\cdot B_s\widetilde{\mathbf{q}}_s'+\mathbf{q}_s|_{F_1}\cdot\nabla F_1-\mathbf{q}_s|_{F_o}\cdot\nabla F_0+S_{os}B_s\frac{\partial\widetilde{h}_s}{\partial t}+B_s\widetilde{P}_s=0\end{aligned} \tag{11}$$

We now introduce the conditions on the top and bottom bounding surfaces. On the impervious bottom, $F_0(x,y,z)=0$, the condition is

$$\mathbf{q}_s|_{F_0}\cdot\nabla F_0=0.$$

On the interface, $F_1(x,y,z,t)=0$, the condition is obtained from (5) and (6), that is,

$$\mathbf{q}_f|_{F_1}\cdot\nabla F_1\equiv-\phi\delta\frac{\partial\widetilde{h}_f}{\partial t}+\phi(1+\delta)\frac{\partial\widetilde{h}_s}{\partial t}=\phi\mathbf{u}\cdot\nabla F_1=-\phi\frac{\partial F_1}{\partial t} \tag{12}$$

in which $h_f|_{F_1}\simeq\widetilde{h}_f$, $h_s|_{F_1}\simeq\widetilde{h}_s$. The magnitude of the error introduced by these assumptions depends on the deviation of the actual flow from the assumed horizontal one in the two domains.

On the phreatic surface, $F_2(x,y,z,t)=0$, with accretion at the rate $\mathbf{N}=-N\nabla z$, we obtain

$$F_2(x,y,z,t)\equiv z-\zeta_2(x,y,t)=z-h_f|_{F2}\simeq z-\widetilde{h}_f=0,$$
$$\mathbf{q}_f|_{F_2}\cdot\nabla F_2=\mathbf{N}\cdot\nabla F_2-\phi_{\text{eff}}\frac{\partial F_2}{\partial t}=-N+\phi_{\text{eff}}\frac{\partial\widetilde{h}_f}{\partial t} \tag{13}$$

With these conditions, we obtain the flow equations for the two domains, as

$$\begin{aligned}-\nabla'\cdot B_f\widetilde{\mathbf{q}}_f'+N-B_f\widetilde{P}_f&=(\phi_{\text{eff}}+S_{0f}B_f+\phi\delta)\frac{\partial\widetilde{h}_f}{\partial t}-\phi(1+\delta)\frac{\partial\widetilde{h}_s}{\partial t},\\ -\nabla'\cdot B_s\widetilde{\mathbf{q}}_s'-B_s\widetilde{P}_s&=\{S_{0s}B_s+\phi(1+\delta)\}\frac{\partial\widetilde{h}_s}{\partial t}-\phi\delta\frac{\partial\widetilde{h}_f}{\partial t}\end{aligned} \tag{14}$$

in which we, usually, assume that

$$B_fS_{0f}\ll\phi_{\text{eff}}\approx\phi,\ B_fS_{0f}\ll\phi,$$
$$\text{and}\quad B_f\widetilde{\mathbf{q}}_f'=-B_f\widetilde{\mathbf{K}}_f'\cdot\nabla\widetilde{h}_f,\ B_s\widetilde{\mathbf{q}}_s'=-B_s\widetilde{\mathbf{K}}_s'\cdot\nabla\widetilde{h}_s$$

Then, in terms of interface and phreatic surface elevations, we obtain the set of equations

$$\begin{aligned}-\nabla'\cdot\mathbf{Q}_f'+N-B_f\widetilde{P}_f&=\phi\frac{\partial(\zeta_2-\zeta_1)}{\partial t},\\ -\nabla'\cdot\mathbf{Q}_s'-B_s\widetilde{P}_s&=\phi\frac{\partial\zeta_1}{\partial t},\\ \mathbf{Q}_f'\equiv B_f\widetilde{\mathbf{q}}_f'&=-B_f\mathbf{K}_f\cdot\nabla\widetilde{h}_f,\\ \mathbf{Q}_s'\equiv B_s\widetilde{\mathbf{q}}_s'&=-B_s\mathbf{K}_s\cdot\nabla\widetilde{h}_s,\\ \zeta_2=\widetilde{h}_f,\ \zeta_1&=(1+\delta)\widetilde{h}_s-\delta\widetilde{h}_f\end{aligned} \tag{15}$$

For a confined aquifer, we delete N and ϕ_{eff} in the first equation in (14).

Since we have assumed "essentially horizontal flow", the flow domain, for which we have to solve (15) for $\widetilde{h}_f(x,y,t)$ and $\widetilde{h}_s(x,y,t)$, is bounded by vertical surfaces that pass through the toe of the interface (= point G in Figure 5b) and through the coast (point C). We add a vertical surface (FE) to complete the delineation of the considered aquifer domain (Figure 5b). Appropriate boundary conditions have to be specified on all these boundaries. The conditions on FE are the common ones (of specified flux, or specified head), and need not be considered here. Along the boundary of GD, the considered freshwater domain becomes an aquifer without the seawater wedge, that is, $\zeta_1\equiv\zeta_0$, $F_1\equiv F_0$. In the case of a phreatic aquifer, the flow equation in (15) reduces for the domain without the interface to (Bear, 1979)

$$-\nabla'\cdot B_f\widetilde{\mathbf{q}}_f'+N+B_f\widetilde{P}_f=(\phi_{\text{eff}}+S_{0f}B_f)\frac{\partial\widetilde{h}_f}{\partial t} \tag{16}$$

in which we usually neglect the effect of elastic storativity, as it is much smaller than the specific yield, ϕ_{eff}. As the interface advances or retreats, the boundary between the two aquifer freshwater domains is also moving. On

the common boundary (DG), we have to maintain the same piezometric head and the same normal flux. On the surface that passes through the coastline, we have two parts: the freshwater part, CB, and the seawater one, BA. The boundary between these two parts (B) is not fixed, but varies as the interface advances or retreats. For the freshwater portion, we assume (Bear, 1979) that the aquifer sub-domain BCM acts as a resistance to the flow, so that across it the head on BC is reduced to that dictated by the sea level. This condition is expressed as

$$Q_{fo} = \frac{\widetilde{h}_{\mathrm{f}}|_{\mathrm{BC}}}{\text{Resist.}}, \quad Q_{fo} = -K_{\mathrm{f}}\overline{\mathrm{BC}}\left.\frac{\partial \widetilde{h}_{\mathrm{f}}}{\partial x}\right|_{\mathrm{BC}}$$
$$\Rightarrow \frac{\widetilde{h}_{\mathrm{f}}}{\alpha} + \overline{\mathrm{BC}}\frac{\partial \widetilde{h}_{\mathrm{f}}}{\partial x} = 0 \text{ on } \overline{\mathrm{BC}},$$

which is a third type boundary condition; α is a coefficient. A similar condition, but with a different coefficient, can be written for the seawater portion of this boundary (AB).

To demonstrate the use of the above equations, we consider the simple case of essentially horizontal steady flow normal to the coast in a confined aquifer, with the stationary interface shown in Figure 6(a). For this case, the freshwater flow equation reduces to

$$\frac{\partial Q_{\mathrm{f}}}{\partial x} \equiv \frac{\partial[-K_{\mathrm{f}}h_{\mathrm{s}}\partial h_{\mathrm{f}}(x)/\partial x]}{\partial x} = 0, \Rightarrow -K_{\mathrm{f}}h_{\mathrm{s}}\frac{\partial h_{\mathrm{f}}(x)}{\partial x} = Q_0,$$
$$\delta h_{\mathrm{f}} = h_{\mathrm{s}} + d, \quad \delta h_{f0} = B + d,$$
$$x = 0,\ h_{\mathrm{f}} = h_{f0}, \quad x = L,\ h_{\mathrm{s}} = 0 \tag{17}$$

with the solution

$$Q_0 x = \frac{K_{\mathrm{f}}[B^2 - h_{\mathrm{s}}^2(x)]}{2\delta},$$
$$Q_0 L = \frac{K_{\mathrm{f}}B^2}{2\delta} \tag{18}$$

This equation, which describes a *parabola*, clearly shows the relationship between the length of seawater intrusion, L, the discharge to the sea, Q_0, and the piezometric head, h_{fo}, above the toe (G).

For the phreatic aquifer shown in Figure 6(b), the flow equation, under the same conditions, takes the form

$$-\frac{\partial Q_{\mathrm{f}}}{\partial x} + N = 0,$$
$$Q_0 + Nx = -K_{\mathrm{f}}(h_{\mathrm{s}} + h_{\mathrm{f}})\frac{\partial h_{\mathrm{f}}(x)}{\partial x}$$
$$= -K_{\mathrm{f}}(1+\delta)h_{\mathrm{f}}\frac{\partial h_{\mathrm{f}}(x)}{\partial x},$$
$$\delta h_{\mathrm{f}} = h_{\mathrm{s}}, \delta h_{f0} = B,$$
$$x = 0,\ h_{\mathrm{f}} = h_{f0}, \quad x = L,\ h_{\mathrm{f}} = 0 \tag{19}$$

Its solution is

$$h_{f0}^2 - h_{\mathrm{f}}^2 = \frac{2Q_0x + Nx^2}{K_{\mathrm{f}}(1+\delta)}, \quad h_{f0}^2 = \frac{2Q_0L + NL^2}{K_{\mathrm{f}}(1+\delta)}$$
$$Q_0 = \frac{K_{\mathrm{f}}B^2}{2L}\frac{1+\delta}{\delta} - \frac{NL}{2} \tag{20}$$

Another possible condition at the sea is $x = L$, $h_{\mathrm{f}} = \beta\delta Q_0 K_{\mathrm{f}}$ (Bear and Verruijt, 1987, p. 207).

Again, we see here the relationship between the drainage of freshwater to the sea and the length of seawater intrusion. *Reducing Q_{f0} means an increased L*. By controlling h_{fo}, say by means of artificial recharge, the water table may be lowered landward of the toe, without any additional seawater intrusion.

When water is pumped above the interface, the latter will rise. This phenomenon is referred to as *upconing* (Bear, 1972, 1979). In considering this phenomenon, we should take into account the existence of the transition zone; as it rises, the pumped water gradually become saline.

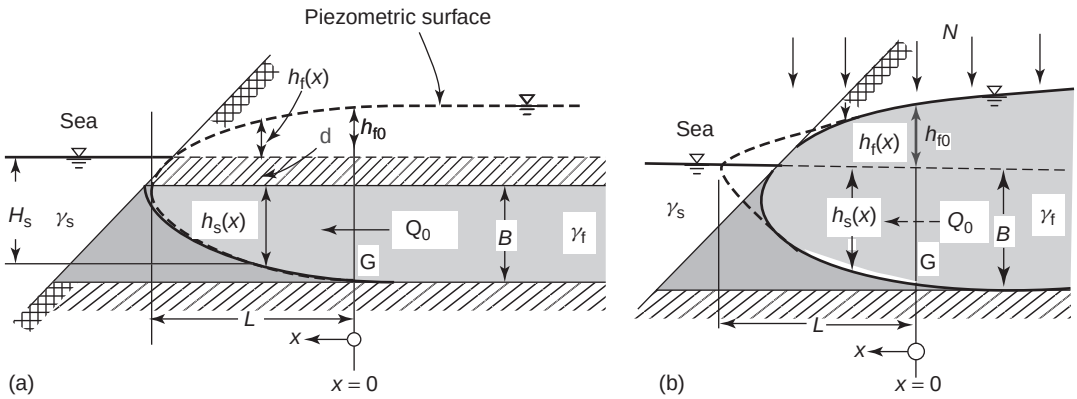


Figure 6 The interface (a) in a confined and (b) in a phreatic coastal aquifer. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

THE TRANSITION ZONE BETWEEN FRESH AND SEAWATER

As already mentioned in the introduction, in reality, the aquifer domain occupied only by seawater, and the aquifer domain occupied only by freshwater are separated by a transition zone. This is a consequence of the fact that the two ("miscible") liquids are actually a single liquid – water – with different concentrations of dissolved salts. The width of the transition zone is dictated by three phenomena: (a) advection of the water – fresh and mixed – toward the sea (or, under certain conditions, landward), (b) dispersion, and (c) molecular diffusion. These phenomena are discussed in (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**) and need not be discussed here. Suffice it to summarize that a transition zone exists; across it, the salinity of the water varies from that of freshwater to that of seawater. The width of this zone grows as it is being displaced in response to changes in the flow regime and in the discharge of water flow to the sea. The transition zone is also fed by a flux of salt from the seawater zone.

Figure 7 shows a phreatic coastal aquifer with a transition zone between seawater and freshwater. The considered flow domain is ABCDEMFA. The detailed conceptual and mathematical models of flow and solute transport are presented in **Chapter 149, Hydrodynamics of Groundwater, Volume 4** and **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**, respectively, and will not be repeated here. The new feature here is that the flow and the solute transport models are *coupled*, as the density of the liquid continuously varies in response to the changes in dissolved salt concentration. We usually refer to such a model as "density variable flow and transport model". In what follows, we shall present this model, assuming isothermal conditions.

The mathematical model describing seawater intrusion in a coastal aquifer consists of: (a) mass balance equation for the water, (b) flux equation (Darcy's law) for the water, (c) mass balance equation for the dissolved salts, and (d) flux equation for the dissolved salts. The first two equations are often combined as a single *flow equation* for the water. The last two equations may be combined to form a single mass balance equation for the dissolved salts, often referred to as the *advection–dispersion equation*, or *the transport equation*. In addition, the complete model includes: (a) constitutive equations that relate the liquid's density and dynamic viscosity to the total dissolved salt concentration and (b) initial and boundary conditions (Bear, 1999; Zhou, 1999).

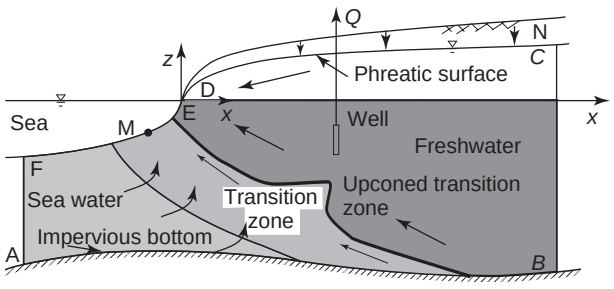


Figure 7 The transition zone with upconing in a simplified vertical cross section of a coastal aquifer, normal to the coastline. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Although, in principle, the flow model can be written in terms of pressure, p, as a state variable, a more (numerically) efficient model is obtained by expressing the flow model in terms of a *reference piezometric head*, h'. Accordingly, we introduce a reference piezometric head, $h'(x, y, z, t)$, associated with a reference water density. We often select ρ_{fw} (= density of freshwater) as a reference density. Thus,

$$h' = \frac{p}{\rho_{fw} g} + z \tag{21}$$

in which z and g denote the vertical coordinate and gravity acceleration, respectively.

To enable the use of the mathematical model to beyond the case of seawater intrusion, discussed here, we consider TDS (= Total Dissolved Solids) in the range where the volume of water may vary with concentration. To facilitate the discussion, we introduce a *normalized salt mass fraction, C(x,y,z,t),* defined by

$$C = \frac{\omega - \omega_{fw}}{\omega_{sw} - \omega_{fw}}, \quad 0 \le C \le 1.0 \tag{22}$$

in which ω and ω_{sw} denote the salt (or TDS) mass fraction in water (= salt solution) and in seawater, respectively, and ω_{fw} is the reference (freshwater) mass fraction. Note that $c = \rho\omega$, in which c denotes the TDS concentration (= mass of TDS per unit volume of fluid).

The constitutive equation that expresses the relationship, $\rho = \rho(p, \omega)$, between the fluid density, the pressure, and the salt mass fraction, ω (mass of dissolved salt per unit mass of fluid), is

$$\rho = \rho_0 \exp[\beta'_p(p - p_0) + \beta'_\omega(\omega - \omega_0)] \tag{23}$$

where p_0 and ω_0 are reference values of pressure and salt mass fraction, respectively, $\beta'_p = (1/\rho)\partial\rho/\partial p$ is the coefficient of water compressibility (at constant salt concentration), and $\beta'_\omega = (1/\rho)\partial\rho/\partial\omega$ is a coefficient that introduces the effect of change in salt mass fraction on the fluid's density (at constant pressure). We usually select $\omega_0 = \omega_{fw}$, that is, equal to the mass fraction of freshwater. The reference pressure is such that for freshwater at $p = p_0$, the density is $\rho = \rho_{fw}$, as the density of freshwater is selected as *a reference density*.

The linearized approximation of (23) is

$$\rho = \rho_0[1 + \beta''_p(p - p_0) + \beta''_\omega(\omega - \omega_0)] \qquad (24)$$

where $\beta''_p = (1/\rho_0)\partial\rho/\partial p$ and $\beta''_\omega = (1/\rho_0)\partial\rho/\partial\omega$. In what follows, we shall assume that for the range of pressures considered here, $\beta''_\omega|\Delta\omega| >> \beta''_p|\Delta p|$, so that we may employ the approximation

$$\rho = \rho_{\mathrm{fw}}(1 + \beta_c C), \quad \beta_c = \beta''_\omega(\omega_{\mathrm{sw}} - \omega_{\mathrm{fw}}) \qquad (25)$$

where β_c may be referred to as a *density difference factor* (dimensionless). In spite of the above assumption, we do take into account the effect of pressure in the expression for the specific storativity appearing in the mass balance equation.

To introduce the effect of concentration on the fluid's dynamic viscosity, we may use the constitutive relationship for dynamic viscosity in the form (Lever and Jackson, 1985)

$$\mu = \mu_{\mathrm{fw}}\mu_r = \mu_{\mathrm{fw}}(1 + 1.85\,\omega - 4.1\,\omega^2 + 44.50\,\omega^3) \qquad (26)$$

in which μ_{fw} is the viscosity that corresponds to $\omega = 0$, and μ_r is defined by (26).

The *specific discharge relative to the solid*, $\mathrm{q}_r \equiv \phi(\mathbf{V} - \mathbf{V}_{\mathrm{s}}) = \mathbf{q} - \phi\mathbf{V}_{\mathrm{s}}$, is expressed by Darcy's law

$$\mathbf{q}_r = \phi(\mathrm{V} - \mathrm{V}_{\mathrm{s}}) = -\frac{\mathbf{k}}{\mu}(\nabla p + \rho g \nabla z),$$

in which $\mathbf{V}$ and $\mathbf{V}_{\mathrm{s}}$ denote the velocity vectors of the water and of the solid matrix, respectively, z denotes the vertical coordinate, positive upward, g denotes gravity acceleration, and the second rank tensor, $\mathbf{k}$, denotes the permeability. We usually, assume, also here, that $\mathbf{V}_{\mathrm{s}} \approx 0$, $\mathbf{q}_r \approx \mathbf{q}$. In terms of h' and C, Darcy's law can be rewritten in the form

$$\mathbf{q}_r = -\frac{\mathbf{k}}{\mu}(\nabla p + \rho g \nabla z) = -\frac{\mathbf{K}^{\mathrm{o}}}{\mu_r}(\nabla h' + \beta_c C \nabla z)$$
$$\mathbf{K}^{\mathrm{o}} = \frac{\rho_{\mathrm{fw}} g \mathbf{k}}{\mu_{\mathrm{fw}}} \qquad (27)$$

with $\mathbf{K}^{\mathrm{o}}$ denoting the reference hydraulic conductivity.

For a single fluid phase, for example, water, of variable density, Bear (1972, 1979) and Bear and Bachmat (1990), as many others, present the general mass balance equations in the form

$$\frac{\partial\phi\rho}{\partial t} = -\nabla\cdot(\rho\mathbf{q}) + (\rho_R Q_R - \rho Q_P) \qquad (28)$$

in which ϕ denotes porosity, Q_R and Q_P, denote, symbolically, the rates of injection of water into and withdrawal of water from the aquifer, respectively (dims. 1/T), ρ_R denotes the density of the injected water, and $\mathbf{q}$ denotes the specific discharge vector. When water is withdrawn and injected through (point) wells, the source terms in (28) may, symbolically, be written as

$$\rho_R Q_R = \Sigma_{(n)}\rho_{Rn} Q_{Rn}(\mathrm{x}_n, t)d(\mathrm{x} - \mathrm{x}_n)$$
$$\rho Q_P = \Sigma_{(m)}\rho_{\mathrm{Pm}} Q_{\mathrm{Pm}}(\mathrm{x}_{\mathrm{m}}, t)d(\mathrm{x} - \mathrm{x}_{\mathrm{m}})$$

in which Q_{Rn} and ρ_R are the injection rate, and fluid's density of water injected through a well at point x_n, respectively, and Q_{Pm} and ρ_{Pm} are the pumping rate and density of water pumped through a well at point x_{m}.

Equation (28) is based on the *assumption that the dispersive flux of the total fluid mass is much smaller than the advective flux*. However, in the case of a fluid of variable density, significant density gradients may develop. For example, it is possible that a rather narrow transition zone will develop, with a relatively large density gradient across it, such that with flow that is more or less normal to such a gradient, significant lateral dispersion may take place. Bear and Bachmat (1990, p. 290) present and discuss a method, using appropriate Peclet numbers (that define the ratio between advective and dispersive fluxes), for examining the conditions under which the dispersive flux of the total mass may be neglected as being much smaller than the advective one. Note that *there is no diffusive flux of the total mass*. In the case of seawater intrusion, we may encounter a rather narrow transition zone, with flow parallel to it. The appropriate Peclet number may be less than or not much larger than one, so that we cannot conclude that advection dominates over dispersion.

Whenever we conclude that the dispersive flux of the total fluid mass cannot be neglected, the mass balance equation (28) should be replaced by

$$\frac{\partial\phi\rho}{\partial t} = -\nabla\cdot(\rho\mathbf{q} - \phi\mathbf{D}\cdot\nabla\rho) + (\rho_R Q_R - \rho Q_P) \qquad (29)$$

in which the dispersive flux of the total fluid (mass per unit area of fluid) is expressed by

$$\mathbf{J}^{*\rho} = -\mathbf{D}\cdot\nabla\rho = -\rho_{\mathrm{fw}}\beta_c\mathbf{D}\cdot\nabla C \qquad (30)$$

In view of the relationship $\rho = \rho(p, C)$, we make use of (27) and (30) to modify (29), rewriting it in terms of the reference piezometric head, h', and C, for example, in the form

$$S_0\frac{\partial h'}{\partial t} + \phi\beta_c\frac{\partial C}{\partial t}$$
$$= \nabla\cdot\left[(1 + \beta_c C)\frac{\mathbf{K}^{\mathrm{o}}}{\mu_r}\cdot(\nabla h' + \beta_c C\nabla z) + \phi\beta_c\mathbf{D}\cdot\nabla C\right]$$
$$+ \frac{\rho_R}{\rho_{\mathrm{fw}}}Q_R - (1 + \beta_c C)Q_P \qquad (31)$$

in which (Bear, 1979) $S_0 = g\rho(\alpha + \phi\beta'_p)$ denotes the *aquifer's elastic storativity*, with α denoting the aquifer's (vertical) compressibility. Other forms are also possible.

For a fluid of constant density, the general mass balance equations for the TDS in the water in the form (*see* **Chapter 152, Modeling Solute Transport Phenomena, Volume 4**):

$$\frac{\partial \phi c}{\partial t} = -\nabla \cdot (c\mathrm{q} - \phi \mathbf{D}_h \cdot \nabla c) + (c_R Q_R - c Q_P) \quad (32)$$

in which $\mathbf{D}_h$ denotes the *coefficient of hydrodynamic* dispersion (a second order symmetric tensor), $\mathbf{D}_h = \mathbf{D} + \mathbf{D}^*_{\mathrm{diff}}$, $\mathbf{D}$ denotes the *coefficient of dispersion* (dims. L^2/T), and $\mathbf{D}^*_{\mathrm{diff}}$ denotes the *coefficient of molecular diffusion in a porous medium*. In an *isotropic* porous medium, the *coefficient of dispersion* is related to the fluid's velocity and to the *longitudinal* and *transversal dispersivities* (dims. L), α_{L} and α_T, respectively, by

$$D_{ij} = \alpha_T V \delta_{ij} + (\alpha_{\mathrm{L}} - \alpha_T)\frac{V_i V_{\mathrm{j}}}{V} \quad (33)$$

in which δ_{ij} denotes the *Kronecker delta*.

For a variable density fluid, with $c = \omega\rho$, the salt balance equation is (e.g. Bear, 1972, 1979)

$$\frac{\partial \phi \rho \omega}{\partial t} = -\nabla \cdot \phi(\rho\omega \mathbf{V} - \mathbf{D}_h \cdot \nabla \rho\omega) + \omega_R \rho_R Q_R - \omega\rho Q_P \quad (34)$$

or, with $\omega_{\mathrm{fw}} = 0$, in terms of C (Zhou, 1999; Bear, 1999)

$$\frac{\partial \phi \rho C}{\partial t} = -\nabla \cdot (\rho C \mathbf{q} - \phi \mathbf{D}_h \cdot \nabla \rho C) + (\rho_R C_R Q_R - \rho C Q_P) \quad (35)$$

When combined with (29), the last equation can be rewritten in the forms

$$\phi\rho \frac{D\omega}{Dt} = \nabla \cdot (\phi\rho \mathbf{D}_h \cdot \nabla\omega) + \phi \nabla \cdot (\omega \mathbf{D} \cdot \nabla\rho) + \rho_R(\omega_R - \omega) Q_R \quad (36)$$

or

$$\phi\rho \frac{DC}{Dt} = \nabla \cdot (\phi\rho \mathbf{D}_h \cdot \nabla C) + \phi \nabla \cdot C(\mathbf{D} \cdot \nabla\rho) + \rho_R(C_R - C) Q_R \quad (37)$$

Altogether, we have to solve the variable density flow equation (31) and the salt transport equation (37) simultaneously, for the two primary variables $h'(x, y, z, t)$ and $C(x, y, z, t)$, making use of (25) to express ρ and (27) to express $\mathbf{V}$ or $\mathbf{q}$.

The initial conditions for the variable density flow and salt transport equations are

$$h'(x, y, z, 0) = h'_0(x, y, z)$$
$$c(x, y, z, 0) = c_0(x, y, z) \quad (38)$$

where $h'_0(x, y, z)$, and $c_0(x, y, z)$ are known distributions.

The conditions on the boundary segments shown in Figure 7, neglecting the dispersive flux of the total fluid mass, are

$$\mathbf{q} \cdot \mathbf{n} = 0, \quad \text{on AB (impervious bottom)}$$
$$h' = h'_p, \text{ or } \mathbf{q} \cdot \mathbf{n} = q_{np}, \quad \text{on BC (land-side lateral boundary)}$$
$$\rho \mathbf{q} \cdot \mathbf{n} = -\left(\rho_N N - (\phi\rho - \theta_{w0}\rho_N)\frac{\partial h'}{\partial t}\right) n_z, \quad \text{on CD (phreatic surface)}$$
$$h' = \varsigma_{\mathrm{DE}}, \quad \text{on DE (seepage face)}$$
$$h' = \beta_c H_{\mathrm{sea}}, \quad \text{on EF (sea bottom)}$$
$$h' = \beta_c H_{\mathrm{FA}}, \quad \text{on FA (sea-side lateral boundary)} \quad (39)$$

where h'_p is the prescribed reference head on a Dirichlet-type boundary, q_{np} is the prescribed fluid flux on a second-type boundary, $\mathbf{n}$ is the unit outward normal vector, with the components n_x, n_y, n_z in three dimensions, N is the replenishment, ρ_N is the fluid's density in the replenishment, θ_{w0} is the irreducible water content assumed to prevail in the unsaturated zone, ς_{DE} is the elevation at a point on the seepage face above a datum, H_{sea} and H_{FA} are the depth of seawater at a point on the sea bottom and the seaside lateral boundaries, respectively. If we wish to take into account the dispersive flux of the fluid's mass, we replace $\rho\mathbf{q}$ by $\rho\mathbf{q} - \phi\rho\mathbf{D} \cdot \nabla\rho$.

The boundary conditions for the salt transport equation are

$$\begin{array}{ll} q^d_n \equiv -\phi D_{ij}\frac{\partial C}{\partial x_{\mathrm{j}}} n_i = 0, & \text{on AB and DE} \\ C = 0, & \text{on BC} \\ q^d_n = \left(\frac{\rho_N C}{\rho} - c_N\right)\left(N + \theta_{w0}\frac{\partial h'}{\partial t}\right) n_z, & \text{on CD} \\ C = 1.0, & \text{on FA} \end{array} \quad (40)$$

where q^d_n is the hydrodynamic dispersive flux normal to a second-, or third-type boundary, and C_N is the salt concentration in the replenishment water.

On the sea bottom boundary, EF, we may have either only landward flow from the sea, in which case the condition there is the same as on FA, or we may have an inflow portion, AM, and outflow portion, MF, separated by a point with zero-normal fluid flux. For the inflow portion, we usually employ either a third-type condition

$$q_n^d \equiv -\mathbf{D}_h \cdot \nabla C = (1.0 - C)q_n \tag{41}$$

or a Dirichlet condition

$$C = 1.0 \tag{42}$$

For the outflow portion, the assumption is often made that the fluid concentrations are identical (i.e. continuous) on both sides of this boundary. Since the fluid fluxes are also identical, the condition becomes a second-type condition:

$$q_n^d \equiv -\mathbf{D}_h \cdot \nabla C = 0 \tag{43}$$

The (moving) point M between the inflow and outflow portions of the boundary is unknown *a priori*. In a numerical solution, during every iteration, we check whether flow along the sea bottom is directed inward or outward, and then assign the appropriate boundary condition accordingly.

We note an *inconsistency* between the head condition on MF, based on the assumption of seawater on the seaside of EM and the assumption of equality of concentration, leading to the flux condition, on this segment. As a consequence, a large error in the salt mass balance may occur in a numerical simulation, when such an inconsistency occurs in the specified flow and transport boundary conditions. To overcome this inconsistency, between the specified flow and transport conditions on the outflow portion of sea bottom, we may assume the presence of a *buffer zone* on the sea bottom, which contains outflowing water. We then take the density of this fluid into account when determining the head condition on this boundary. Iterations may be required in a numerical solution. The thickness of this buffer zone is a calibration parameter (Bear *et al.*, 2001).

A number of computer codes for solving the above set of nonlinear coupled equations are presented in **Chapter 155, Numerical Models of Groundwater Flow and Transport, Volume 4**.

MANAGEMENT OF A COASTAL AQUIFER

The concepts and methodologies for the management of a coastal aquifer are not much different from those of any other aquifer and will not be considered here. The main difference stems from the presence of the sea and the continuous threat of seawater intrusion, with all its negative consequences. In general, as demonstrated by (18) and (20), increasing the net pumping from the aquifer, or reducing freshwater discharge to the sea, will cause an increase in the extent of seawater intrusion. Thus, optimization in this case, or sustainable yield, means finding the right balance between the benefits from increased pumping and the damages or costs incurred by the increased extent of seawater intrusion. These damages will include salinization of wells to the extent that wells have to be abandoned and water has to be supplied from sources that are farther inland. Here, *sustainable yield* of an aquifer is the annual quantity of water that can be withdrawn from the aquifer year after year, as a constant volume, or as one that varies according to some rule, without violating specified constraints, especially **that the source will be preserved (quality and quantity) forever**. The models described above can be used to predict the extent of seawater intrusion in response to proposed management alternatives. Following are a number of additional management considerations.

We have demonstrated that the position of the interface toe (in the sharp interface model) can be controlled by controlling water levels (or piezometric heads) at a desired distance from the sea. In this way, it is possible to create a *barrier* to seawater intrusion. The elevation of the barrier can be achieved either by controlling pumping farther inland and/or by implementing artificial recharge techniques.

In general, pumping above the interface will cause upconing of the interface (or transition zone) toward the pumping well, causing salinization of that well. However, it is possible to avoid salinization, even when wells are operated close to the coast, by controlling their rates and by locating their screens sufficiently high above the transition zone. In Israel, this technique is implemented as a "coastal collector", in the form of an array of shallow wells along a line close to the sea. The objective is to intercept part of the drainage of freshwater to the sea.

Under certain circumstances, it is possible to control seawater intrusion by creating (by pumping) a "trough" parallel to the coast (Huisman and Olsthoorn, 1983), or by constructing a physical barrier in the form of an impervious or semi-pervious barrier wall parallel to the coast (Todd, 1980). The latter will be effective only if it extends over a significant length.

In many parts of the world, the coastal aquifer, which is an important source of freshwater, is plagued by two additional problems. The first is that these aquifers are usually heavily populated. This means that large portions of the ground surface are covered by concrete and asphalt, thus reducing natural replenishment over these areas. The second is that human activity in such heavily populated areas increase the danger of contamination of the aquifer by pollutants originating at ground surface and traveling through the unsaturated zone to the underlying aquifer. Special means and regulations are required, as part of aquifer management, to combat these threats to the sustainable yield of the aquifer.

REFERENCES

Badon-Ghyben W. (1888) *Nota in Verband met de Voorgenomen Putboering Nabij Amsterdam*, (Notes on the probable results of well drilling near Amsterdam), Tijdschrift van het Koninklijk Institut van Ingenieurs: The Hague pp. 8–22.

Bear J. (1972) *Dynamics of Fluids in Porous Media*, American Elsevier: p. 764.

Bear J. (1979) *Hydraulics of Groundwater*, McGraw-Hill: p. 569.

Bear J. (1999) *Conceptual and Mathematical Modeling*, Chap. 5 Bear J., Cheng A.H-D., Sorek S., Ouazar D. and Herrera I. (Eds.), Kluwer Academic Publishers.

Bear J. and Bachmat Y. (1990) *Introduction to Modeling of Transport Phenomena in Porous Media*, Kluwer Academic Publishers: p. 553.

Bear J., Cheng A.H-D., Sorek S., Ouazar D. and Herrera I. (Eds.) (1999) *Seawater Intrusion in Coastal Aquifers – Concepts, Methods and Practices*, Kluwer Academic Publishers: Dordrecht, p. 625.

Bear J. and Dagan G. (1964) Some exact solutions of interface problemsby means of the hodograph method. *Journal of Geophysical Research*, **69**(2), 1563–1572.

Bear J. and Verruijt A. (1987) *A Modeling Groundwater Flow and Pollution, D*, Reidel Publishing: p. 414.

Bear J., Zhou Q. and Bensabat J. (2001) Three dimensional simulation of seawater intrusion in heterogeneous aquifers, with application to the coastal aquifer of Israel. *Proceedings of the First International Conference on Saltwater Intrusion and Coastal Aquifers – Monitoring, Modeling and Management*, Essaouira: Morocco.

Du Commun J. (1828) On the cause of fresh water springs, fountains, etc. *American Journal of Science and Arts*, **14**, 174–175.

Herzberg A. (1901) Die Wasserversorgung einiger Nordseebaden. (The water supply on parts of the North Sea coast in Germany) *Journal fuer Gasbeleuchtung und Verwandete Beleuchtungsarten sowie fuer Wasserversorgung*, **44**, 815–819 824–844.

Huisman L. and Olsthoorn T.N. (1983) *Artificial Groundwater Recharge*, Pitman Advanced Publishing Program: p. 320.

Konikow L.F., Goode D.J. and Hornberger G.Z. (1996) *A Three-Dimensional Method-of-Characteristics Solute-Transport Model*, (MOC3D), U.S. Geological Survey Water Resources Investigations: Report 96–4267.

Lever D.A. and Jackson C.P. (1985) *On the Equations for the Flow of a Concentrated Salt Solution through Porous Medium*, Harwell Report AERE-R, 11765, Her majesty's Stationary Office, London.

Reilly T.E. and Goodman A.S. (1980) Quantitative analysis of saltwater-freshwater relationships in ground-water systems – a historical perspective. *Journal of Hydrology*, **80**, 125–160.

Todd D.K. (1980) *Groundwater Hydrology*, *Second Edition*, John Wiley & Sons: p. 535.

Zhou Q.L. (1999) *Modeling Seawater Intrusion in Coastal Aquifers*, Ph.D. thesis, Technion-Israel Institute of Technology.

158: Anthropogenic Land Subsidence

GIUSEPPE GAMBOLATI, PIETRO TEATINI AND MASSIMILIANO FERRONATO

University of Padova, Padova, Italy

Fluid removal from subsurface reservoirs, in the form of gas, oil, groundwater, geothermal water, and brine, produces a compaction of the depleted formations which migrates totally or partially to the ground surface thus inducing anthropogenic land subsidence. The paper presents: (i) a list of the major subsiding areas worldwide, (ii) a review of the mechanism which causes a measurable settlement above aquifer systems and gas/oil fields, (iii) a description of the currently available techniques to measure land subsidence and in situ rock compaction, (iv) a brief description of some mathematical models to predict the magnitude of subsidence, and (v) a description of a few remedial options that are available to control the event and mitigate the related environmental impact.

INTRODUCTION

One major environmental consequence of groundwater pumping and gas/oil/geothermal water production is anthropogenic land subsidence. To be of major concern, subsurface fluid withdrawal must occur in densely populated and highly developed areas located close to the sea or a lagoon or a delta, and take place from unconsolidated geological basins of alluvial, lacustrine or shallow marine origin, formed typically, although not exclusively, in the Quaternary period. Quite often, especially at the onset of the occurrence, land settlement goes unnoticed, to be later discovered when severe damages have already been experienced. At this stage, undertaking effective remedial measures to mitigate the associated environmental and socioeconomical impact may prove tremendously expensive. However, in recent times, the awareness about the damages caused by potential anthropogenic land subsidence has significantly grown at both the political and the general public level, thus contributing to lower the alarm threshold. As a major result, the newest plans for subsurface resource development are usually complemented by a study of the related environmental impact which may include, wherever appropriate, numerical predictions of the expected land settlement above (and close to) the exploited system.

The first observation concerning land subsidence due to fluid removal dates back to the beginning of the twentieth century with the first scientific report on the event written by two geologists (Pratt and Johnson, 1926). Their conclusion was that "*the cause of the subsidence is to be found in the extensive extraction of oil, water, gas, and sand from beneath the affected area*", which was located above the oil field of Goose Creek, S. Jacinto Bay, Texas. The early conjecture of Pratt and Johnson was to be later confirmed and reconfirmed by countless examples of anthropogenic land subsidence (see Table 1), and supported by geomechanical theory as well. The maximum recorded settlement amounts to as much as 14 m, while the depth of pumping wells may range from those tapping very shallow water table aquifers just close to the ground surface to those tapping very deep (4000–5000 m) gas/oil reservoirs.

Over gas/oil fields, the subsidence usually takes on a bowl-shaped appearance with the largest downward displacement occurring near the center of the field. The border of the bowl may roughly resemble the shape of the field although it may extend up to twice or more the area encompassed by the outline of the underlying reservoir. In the case of extensive pumped aquifer-aquitard systems, the overall extent of the sinking area can be much larger, totaling as much as 13 500 km^2 in the case of the S. Joaquin Valley, California (Poland and Lofgren, 1984), and 12 000 km^2 in the Houston–Galveston area, Texas (Gabrysch, 1984). By distinction, subsiding areas over gas/oil fields never reach such a large size.

The analysis and the prediction of the expected anthropogenic land subsidence due to fluid pumping is not

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Selected areas of major anthropogenic land subsidence worldwide

Location	Depth of pumping (m)	Maximum subsidence (m)	Area of subsidence (km^2)	Time of main occurrence	Principal reference
Ravenna, Italy	80–450[a]	1	400	1955–1985	Gambolati *et al.* (1991)
	1700–4000[d]	0.3	80	1960 to present	Baú *et al.* (2000)
Venice, Italy	70–350[a]	0.12	150	1952–1973	Gambolati *et al.* (1974)
Po River delta, Italy	0–600[e]	3.5	2600	1938–1961	Gambardella *et al.* (1991)
Groningen, Netherlands	2800–3000[d]	0.23	700	1959 to present	Kenselaar and Martens (2000)
Ekofisk, North Sea	3000–3200[c]	6.7	40	1970 to present	Hermansen *et al.* (2000)
Houston, Texas	60–900[a]	3	12 000	1906–1995	Gabrysch and Neighbors (2000)
San Joaquin V., California	60–900[a]	9	13 500	1930–1975	Galloway and Riley (1999)
Wilmington, California	600–1200[c]	9	70	1926–1968	Allen (1969a)
Las Vegas, Nevada	200–300[a]	2	250	1935 to present	Amelung *et al.* (1999)
Mexico City, Mexico	0–50[a]	9 (1978)	225 (1978)	1891 to present	Figueroa–Vega (1984)
Maracaibo lake, Venezuela	500–1000[c]	4.5 (1986)	500 (1986)	1928 to present	Finol and Sancevic (1995)
Wairakei, New Zealand	250–800[b]	14	30	1952 to present	Allis (2000)
Bangkok, Thailand	70–250[a]	1.6 (1988)	4550 (1988)	1933 to present	Nutalaya *et al.* (1996)
Tianjin, China	100–300[a]	3	8000	1959 to present	Hu *et al.* (2002)
	1900–2000[b]	0.2		1985 to present	Yao and Bi (2000)
Ping–Tung County, Taiwan	70–180[a]	3.1	105	1972 to present	Liu *et al.* (2000)
Tokyo, Japan	0–400[a]	4.6	3400	1918–1978	Yamamoto (1995)

[a]groundwater.
[b]thermal water.
[c]oil.
[d]gas.
[e]gas-bearing water.

an easy task. A careful reconnaissance study of the area of interest is required with the detailed recognition of the basin geology and geometry and the reconstruction of its past history. Geomechanical and hydraulic properties are of the utmost importance. Preconsolidation stress, zones of overpressure, and faults must all be reliably identified in formations located at a great burial depth. Advanced technology (2-D and 3-D seismic surveys, *in situ* geophysical measurements, explorative boreholes, field tests, laboratory analyses) can be of great help. Much progress has also been made in accurately recording and monitoring the ground surface movements from the traditional spirit leveling. New techniques include DGPS (Differential Global Positioning System) and InSAR (Interferometric Synthetic Aperture Radar) by which land subsidence is measured from space with a very high precision. Advances have also been accomplished in measuring aquifer system and reservoir compaction by borehole extensometers and radioactive markers, respectively.

We must mention other types of anthropogenic land subsidence that are not addressed by the present contribution, most of which are less important in terms of socioeconomical and environmental impacts. These include underground mining, carbonate rock solution, subsurface erosion, surface loading, land drainage and reclamation, histosoil (peat) oxidation, dissolution of soil carbon and water application (Allen, 1984).

The present paper is organized as follows. The mechanics of rock compaction due to fluid removal and resulting land subsidence is described. The most advanced tools to record and monitor the ground deformations are reviewed. The mathematical models to simulate and predict the occurrence are addressed. Methods to control, mitigate or stop anthropogenic land settlement are finally discussed.

MECHANICS OF LAND SUBSIDENCE DUE TO FLUID WITHDRAWAL

Basic Mechanisms

The mechanism that relates anthropogenic land subsidence to fluid withdrawal is that of subsurface sediment compaction caused by changes of the stress distribution within the solid skeleton. Stresses change because pumping induces a modification of the fluid-dynamic field. The fluid flow system is controlled by a potential field with the fluid head acting as the potential quantity. The introduction of a production well into a natural fluid flow system produces a disturbance that propagates its effect in space and time through the geological medium. Around the well, a cone of depression in the fluid head in the pumped formation develops and expands laterally, and to a minor extent also vertically. The intensity of the head drop at any point of the porous medium and the time lag between the inception of withdrawal and the arrival of the effect at that point depend on the distance of the point from the well field, on the geometric and geologic configuration of the subsurface basin, on its boundary conditions, and on the fluid-dynamic and geomechanical properties of both fluid and formation, specifically fluid density and viscosity, and medium intrinsic permeability, porosity, and compressibility.

The mechanism by which rock deforms and compacts under the influence of a fluid head change is well understood. The total geostatic load σ_t, acting on the aquifer, reservoir, confining beds and caprock, is balanced by the pore pressure p and the effective vertical and horizontal stresses σ_{ev} and σ_{eh} (Figure 1). As the fluid head declines, p declines too and can no longer support as large a percentage of the load of the overlying formations. Therefore, more of this load must now be borne by the grain-to-grain contacts of the geological material itself with a stress transfer from the fluid to the solid phase, and a consequent increase of effective stress in both the pumped units and the adjacent formations (i.e. confining beds, caprock, bottom and lateral aquifer) which are progressively drained, and hence compact, with the amount of compaction primarily related to the compressibility of the compacting layers. The resulting cumulative compaction of subsurface layers extends its effect to the ground surface, which therefore subsides (Figure 1). If the depleted units are seated deeply into the basin, as in the case of a typical gas/oil reservoir (Figure 2), the land surface behaviour resembles that caused by a finite source of stress in a semi-infinite medium. The surrounding rocks absorb part of the loss of support due to the local pressure drawdown, and the actual land settlement depends primarily on depth, volume, and compressibility of the reservoir and adjacent formations.

Typically settlement above gas/oil fields is smaller than the reservoir compaction but it spreads over a larger area than the extent of the field itself. Conversely, aquifer systems are generally shallower and have a much larger areal

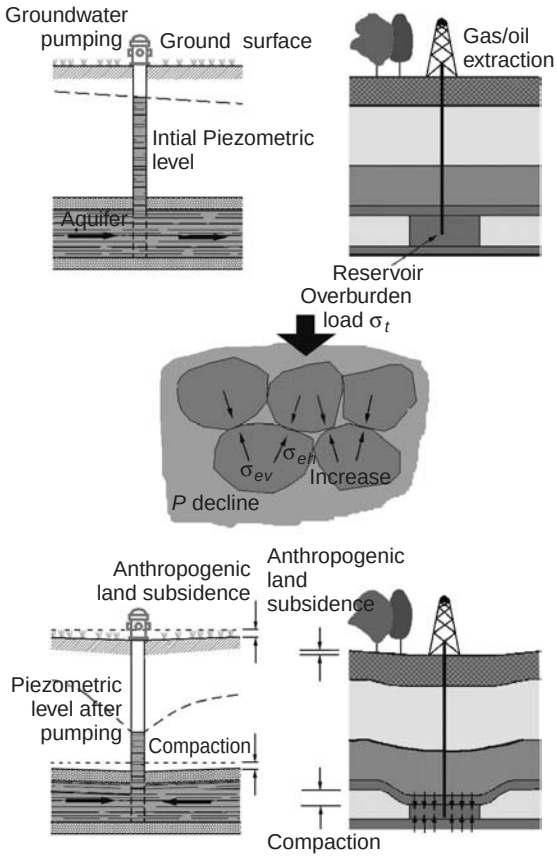


Figure 1 Mechanics of land subsidence due to groundwater and gas/oil withdrawal. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

extent than gas/oil fields. For these systems, the existence of a central zone may be conceived where rock compaction is not contrasted by the overburden and simply migrates to ground surface with a subsidence spreading factor equal to one. Hence, such stratified systems behave mechanically as if they were 1-D structures, and, although fluid flow may be 3-D (e.g. vertical in the confining beds, or aquitards, and horizontal in the aquifers, Figure 3), land displacement occurs mostly in the downward vertical direction.

In addition to dimensionality, other factors differentiate the mechanism of gas/oil field compaction from that of aquifer/aquitard compaction. Usually, both subsurface environments consist of a sequence of sands and clays (aquifers), and sandstones and shales (gas/oil reservoirs). Sandstones are cemented sands, whereas shales are clays that have undergone extensive mineralogical changes in the burial process that associated them with the hydrocarbon bearing strata. These changes may have had a pronounced influence on the shale compaction properties. Freshwater aquifer systems are usually normally-consolidated and normally-pressurized, or only slightly overpressurized, and lack important faults due to the typical

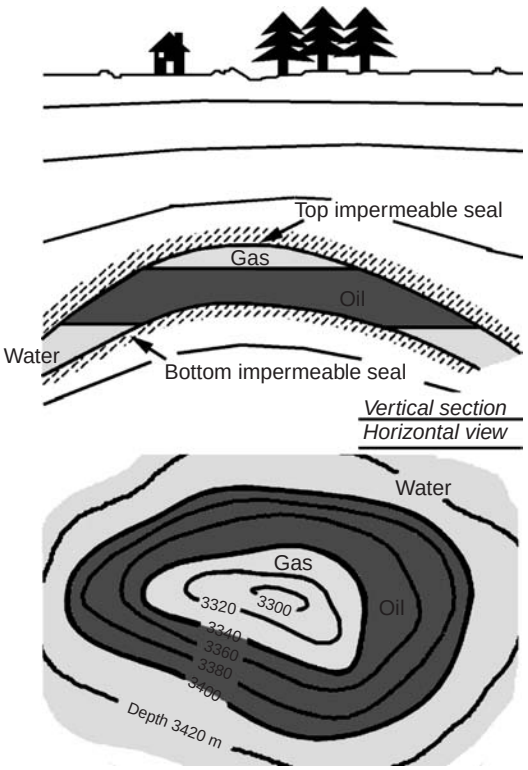


Figure 2 A typical gas/oil reservoir. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

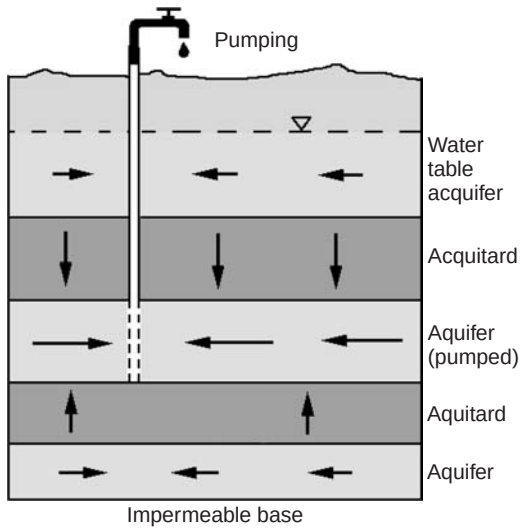


Figure 3 Flow in a typical pumped multiaquifer-aquitard system. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

formation mechanism, that is, a depositional environment without significant interfering tectonic movements. However, the geomechanical simplicity may be partially offset by the litho-stratigraphic complexity related to the distribution of clayey, silty, and sandy soils within the compacting system. It is well known that clay is up to two orders of magnitude more compressible than sand at shallow depth (Chilingarian and Knight, 1960). Hence, land subsidence of a freshwater system is very dependent on the clayey and silty fraction within the system in the form of confining beds, intervening aquitards, and interbedded lenses. Moreover, drainage from these beds can lag behind drainage from the producing sand, thus causing a delayed land subsidence which may manifest itself after wells shut down. By distinction, in deeply seated gas/oil fields, clay (shale) and sand (sandstone) tend to exhibit the same mechanical properties irrespective of lithology (Poland, 1984a; Finol and Sancevic, 1995), and this further differentiates the occurrence above pumped aquifer systems from that above productive gas/oil fields.

When a porous body experiences a change in the internal flow and stress fields, due to, for example, a sedimentation process which yields a σ_t increase, or a fluid pumping which causes a p decrease, the incremental effective stress and the fluid-dynamic gradient that develop are intimately connected. This connection was first recognized by Biot (1941) who developed the coupled theory of consolidation (and hence the coupled theory of land subsidence) where fluid flow influences the porous medium deformation which in turn affects the flow field. Groundwater hydrologists and petroleum engineers, who are mainly concerned with the fluid-dynamic aspects of this coupled interrelation, have advanced the uncoupled flow theory, which is based on the so-called diffusion equation. Theis (1935) solved the single-phase flow of groundwater incorporating the rock structural properties into a lumped geomechanical parameter (i.e. the elastic storage coefficient S_s). Theis' solution to the diffusion equation is calculated separately and independently of the medium structural solution to provide the pore pressure p distribution. Once obtained, p is used as the external driving force to predict the medium deformation, in particular, the vertical displacement at the ground surface, that is, the land subsidence. This is the uncoupled, or two step approach usually implemented for predicting land settlement due to fluid removal (Gambolati and Freeze, 1973; Gambolati *et al.*, 1974, 1991, 1998, 1999; Geertsma, 1973; Helm, 1975, 1976; Brutsaert and Corapcioglu, 1976; Harris Galveston Coastal Subsidence District, 1982; Narasimhan and Goyal, 1984; Martin and Serdengecti, 1984; Rivera *et al.*, 1991; Doornhof, 1992; Gonella *et al.*, 1998; Baú *et al.*, 1999, 2000; Teatini *et al.*, 2000).

Gambolati (1992) and Gambolati *et al.* (1996a,b) have shown that in regional geological settings, the coupled and uncoupled flow solutions are virtually indistinguishable on any timescale of practical interest, with the related land subsidence predictions practically the same. Minor differences have been obtained only in the transient behaviour

of horizontal surface ground displacements (Gambolati *et al.*, 2000).

A few authors (e.g. Cui and Delage, 1996; Delage *et al.*, 1996; Bolzon and Schrefler, 1997; Schrefler *et al.*, 2000) have recently advanced a new mechanism of compaction of gas/oil fields based on capillary suction theory as this relates to porous medium mechanical resistance. However, the capillary theory is not supported by adequate field evidence and the importance of suction on reservoir compaction has not yet been demonstrated using both undisputed physical arguments and *ad hoc* experimental tests.

In summary, a realistic representation of the mechanism of anthropogenic land subsidence due to fluid withdrawal involves a two-step process. The first step addresses the fluid-dynamic part of the occurrence where the substantial impact of the geomechanical porous medium behaviour on flow may be quite effectively accounted for by the elastic storage S_s. The second step solves the structural problem using the fluid pore pressure p distribution calculated in the first step as a driving force within the medium (possibly over or under-consolidated and faulted) to provide the solid skeleton deformation and the subsidence (i.e. the vertical displacement) at the land surface.

Controlling Factors

Four factors may partially combine to produce measurable settlement records: (i) shallow burial depth of the depleted layers; (ii) highly compressible rock laid down in alluvial or shallow marine environments; (iii) large pore pressure decline; and (iv) large thickness of the depressurized fluid bearing sediments. Unless the gas/oil fields are overpressurized, factors (i) and (iii) are mutually exclusive, while they can be both associated with factors (ii) and (iv). For a large subsidence to occur, however, a soft compacting rock is needed and/or a large pressure drawdown. To give a few examples, Mexico City sank by 9 m with a maximum pressure decline of only 0.7 MPa because of the extremely soft high-porosity soils of the compacting shallow formations located within the upper 50 m, while the 9 m and the 6.7 m settlement reported from Wilgminton (Allen, 1969a) and Ekofisk (Zaman *et al.*, 1995; Hermansen *et al.*, 2000) oil fields, respectively, are accounted for by the pronounced pore pressure drop (exceeding 20 MPa in the latter) combined with the large thickness of the compacting units. The Ekofisk case is a very interesting one in that the reservoir rocks have exhibited a sudden increase in compressibility at some stage of the field development with a large irreversible deformation defined as "pore collapse", which is believed to be the main reason for the increased subsidence over the field (Zaman *et al.*, 1995).

Some reservoirs and aquifers may be overconsolidated (e.g. the Venezuela fields, Holzer, 1981; Finol and Sancevic, 1995). Overconsolidation tends to reduce the early subsidence rate with a sudden unexpected growth at some stage of production, when the *in situ* stress exceeds the preconsolidation stress. If the water or gas/oil bearing sediments are preconsolidated, it may be very difficult to predict anthropogenic land subsidence prior to the field/aquifer development. A preconsolidation effect might have been caused in the geological past by uplift followed by erosion of the sediments overlying the fluid bearing layers or by fluid overpressure or both. The aforementioned processes may have led to a reservoir/aquifer system expansion which was much smaller than the original virgin, mostly unrecoverable compaction. When pore pressure lowers due to fluid production, a reloading of the depleted formations takes place. Initially compaction, and hence land settlement, are small. However, as soon as the maximum experienced load is overcome, rock compression occurs on the virgin loading curve with a sudden increase of compressibility, and of the resulting subsidence rate. If there is an important compressibility contrast between the reservoir and the caprock, land may even rise when fluid is removed (Ferronato *et al.*, 2001). This might occur in case the reservoir is overconsolidated and stiffer than the confining bed (and this further adds to the complexity of a reliable land subsidence prediction above overconsolidated reservoirs).

Another factor that may influence the occurrence is the presence of faults within the developed system and the overburden, as in the case, for instance, of Las Vegas (Amelung *et al.*, 1999). Faults may weaken the porous medium structure and make both the analysis and the prediction more difficult.

When a gas/oil field is considered, an additional source of complexity is the lateral/bottom aquifer (called waterdrive). Pore-pressure drawdown may extend to the waterdrive as well, and induce further land subsidence as a consequence of the waterdrive compaction, even after the wells shut down and the field is abandoned. It is not uncommon that land settlement around a field still continue for several years after the cessation of pumping as the result of the residual waterdrive compaction (Baú *et al.*, 2000).

MEASURING AND MONITORING LAND SUBSIDENCE

The occurrence of land subsidence due to subsurface fluid withdrawal quite often goes unnoticed, as it typically progresses at a slow rate and involves large areas. In the absence of observational evidence, such as protrusion of wells or infrastructures (e.g. bridges), submersion of offshore platform facilities, break of pipelines, drainage reversals, or failure of well casings, repeated measurements of ground surface elevation are needed to reveal anthropogenic land subsidence. The problem of monitoring the settlement over large areas is compounded by the requirement for vertically stable reference benchmarks located outside the region affected by subsidence to be used as control points.

Spirit Leveling

Spirit (differential) leveling is the traditional method of determining ground elevation changes, and despite its simplicity, can be very accurate (Table 2). Equipment and procedures are described in detail in several manuals, for example, in Rappleye (1948) and Floyd (1978). The technique, developed in the nineteenth century and much used at present as well, allows the surveyors to carry an elevation from a known reference point to other geodetic marks by the use of a precisely leveled telescope and a pair of graduated vertical rods. For each survey, the elevation difference e_d between two benchmarks is recorded twice by accumulating the elevation differences between a series of temporary turning points. The discrepancy between e_d measured in the forward and backward directions (relative to the direction of the traverse) must not exceed $1.5\sqrt{D}$ mm for a "very high (first order)" precision survey, and $3\sqrt{D}$ mm for a "high (second order)" precision survey, D being the length of the benchmark line in kilometers. Typically, benchmarks are spaced 1 km apart and turning points are 20 to 100 m apart. Once a network of benchmarks has been established and surveyed by precise leveling, a further survey at some later date shows whether vertical movements have occurred and quantifies their magnitude.

Digital levels and invar rods help fulfill the selected accuracy, eliminating the human errors and increasing the measurement speed. To ensure long-term time series, the benchmarks, usually consisting of a brass cup or headed bolt, are grouted into massive artificial structures, such as bridge abutments, or bedrocks, or attached at the top of a bar 5 to 10 m long driven into the ground and protected by an outer sleeve. Examples of leveling networks established to control land subsidence due to fluid withdrawal are published by Poland and Yamamoto (1984) and Tosi *et al.* (2000).

Differential GPS

Large regional benchmark networks warrant the use of the less time-consuming Global Positioning System (GPS) surveying method. USAF NAVSTAR (Navigation Signal Timing and Ranging) GPS uses earth-orbiting satellites to obtain accurate positions based on the time required by radio signals transmitted from the satellites to reach a receiving antenna. Since 1995, NAVSTAR has been operational with a constellation of 24 satellites at an average orbit altitude of 20 200 km, arranged in six orbital planes 55° inclined relative to earth's equator (US Coast Guard Navigation Center, 1996). The satellites, positioned in the orbit planes in such a way as to permit simultaneous vision of at least four of them, continuously transmit ranging signals on two carrier frequencies known as L1 (1574.42 MHz) and L2 (1227.60 MHz) on which a navigation message, including satellite clock data, satellite ephemeris (precise orbit) data, and ionospheric signal-propagation correction data, is superimposed.

Absolute positioning by GPS, that is, the determination of a single point on the earth surface collecting data from multiple satellites, is not sufficiently accurate for precise surveying because of satellite ephemeris, clock errors, and signal path delays through the atmosphere. In land subsidence surveys, Differential GPS (DGPS) technique is used instead to enhance elevation accuracy (Table 2). DGPS determines the relative position of two points on which two GPS receivers are placed to receive signals simultaneously from the same set of satellites. The system takes advantage of the fact that close receivers simultaneously experience common errors, which can be (partially) cancelled by observation data post-processing. When the same points are reoccupied after some time interval, the relative motion that has occurred between the points during the time interval can be measured. Geodetic networks of benchmarks, at least one of them located in a reference stable position, can be surveyed in this way. Since the measurement accuracy primarily depends on the relative benchmark distance (baseline) and baseline occupation time, a proper network design and survey planning are of paramount importance (US Army Corps of Engineers, 1996). DGPS measurement strategy currently

Table 2 Main features of the earth observation techniques presently used to record land subsidence

Method	Spirit leveling	DGPS (Differential Global Positioning System)	DInSAR (Diff. Interferometric Synthetic Aperture Radar)	PS – IPTA (Permanent Scatterers – Interf. Point Target Analysis)
Accuracy[a]	1 mm/year	2–5 mm/year	1–2 mm/year	1–2 mm/year
Spatial scale	line – network	network	map pixel	permanent reflector
Spatial resolution	0.5–1 km	5–10 km	25 m	25–1000 m
Temporal sampling[b]	5–10 years	2–3 years	5 years	2–3 years
Time-cost per survey[c]	2–3 months	1 month	1 week	2–3 weeks

[a]subsidence rate accuracy attainable under optimum conditions.
[b]indicative values for subsidence rate less than 1 cm/year, otherwise frequency must be increased.
[c]indicative values to survey a 50 × 50 km area, including data processing, with state-of-the-art methodologies.

consists of few expensive stations in permanent acquisition characterized by an extensive baseline (e.g. Sato *et al.*, 2003), or closer benchmarks for which the survey is carried on in "static" method, that is, by simultaneously placing on two/three of them time mobile antennas that record the satellite signals for a time ranging from few hours to few days per survey, or a combination of the two (e.g. Bitelli *et al.*, 2000).

InSAR

Interferometric Synthetic Aperture Radar (InSAR) is a powerful new tool that uses radar signals to measure deformation of the earth surface at an unprecedented level of spatial detail (Table 2). Radar transmits an electromagnetic microwave signal that, reflected off the ground surface, is received by the antenna producing a digital image of the scanned ground. Each image pixel carries amplitude and phase information of the microwave field back-scattered by all the scatterers (rocks, vegetation, buildings, etc.) located within the cell. Currently, operational satellite-borne radar systems are carried on the Canadian RADARSAT, the Japanese JERS, and the European ERS-2 and ENVISAT spacecrafts. SAR images are available since 1992. ERS-2 moves along a quasi-polar orbit with a pass every 35 days and an elevation of about 800 km. Its antenna footprint sweeps a 100-km-wide strip on the earth surface with a spatial resolution of about 25×25 m (Table 2).

In the Differential InSAR approach, used in geophysical sciences since the late '80s (Gabriel *et al.*, 1989), two (or more) SAR images taken from very close orbital positions at different times are combined (interferometric processing) to exploit the phase difference of the signals which is related only to the earth surface displacement occurring between the acquisition of the image pair, once the surface topography contribution is removed.

The interferogram phase noise ("decorrelation") restricts the DInSAR use. Major contributions to the phase noise are the temporal and the spatial decorrelation, the former due to the temporal change of the scatterers, as is the case of densely vegetated areas, and the latter to the slightly different viewing positions (interferometric baseline) related to the orbits. The phase noise is estimated by means of the local coherence L, a cross-correlation coefficient of the SAR image pair ranging from 0 to 1, when the interferometric phase is pure noise and phase noise is absent, respectively. The conventional DInSAR approach is performed by analyzing 2 to 10 SAR images characterized by a very small baseline (generally less $\leq$200 m). Requiring $L > 0.3$ on several adjacent pixels, its applicability is restricted to urban, built-up, and not-vegetated areas on which land subsidence is uniformly mapped (Figure 4).

More recently, an alternative approach known as PS (Permanent Scatterers) or IPTA (Interferometric Point Target Analysis) has been developed taking advantage from the

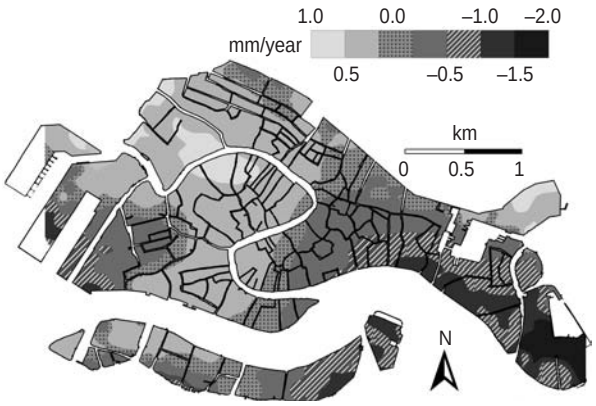


Figure 4 Vertical movement rate (mm/year) at Venice from 1992 to 1996 as detected by DInSAR with a 25-m resolution. Positive values indicate uplift, negative land subsidence [Reproduced by permission of AGU from Tosi L, Carbognin L, Teatini P, Strozzi T and Wegmüller U (2002) *Geophysical Research Letters*, 29, 10.1029/2001GL013211]. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

fact that, when the dimension of a scatterer is smaller than the image resolution cell, the coherence is good irrespective of the image pair baseline. Consequently, more observations are available, thus allowing for a reduction of temporal decorrelation errors. PS/IPTA interferometric phases are thus interpreted only for a number of selected single pixels that are coherent ($L > 0.7$) over long-time intervals and cover the monitored area as a sparse "natural" benchmark net. By the PS/IPTA technique, the interferometric SAR-based subsidence survey can be performed in rural areas as well using existing pointwise reflectors (e.g. buildings, bridges, etc.), which require a permanent scatterer/point target density larger than 1 km^{-2}. The advantages offered by the PS/IPTA technique in comparison with the conventional DInSAR justifies to a large extent the need (hence the cost) for a large number of SAR images (usually more than 30). Examples along with a detailed description of the techniques are given by Ferretti *et al.* (2000) and Strozzi *et al.* (2001).

In order to turn the displacement phase term into an absolute displacement velocity, a reference point is required by both DInSAR and PS/IPTA. Quite often the scene content allows for the specification of stable points. Sometimes, leveling or DGPS surveys may provide the reference point as the present tendency in monitoring land subsidence involves an integrated monitoring approach where all (or more than one) of the surveying techniques described above are simultaneously used to cross-validate the measurements and improve the accuracy of the measurements (e.g. Strozzi *et al.*, 2003).

Finally, it is worth mentioning that natural land subsidence occurring over centuries or millennia can be

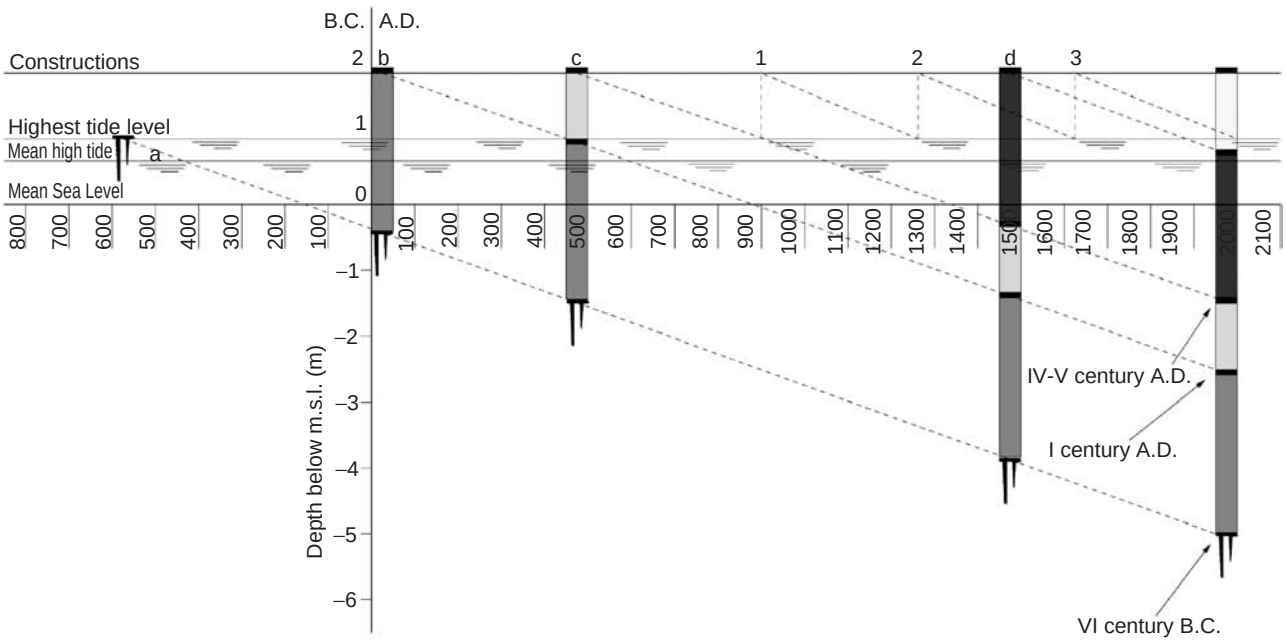


Figure 5 Land settlement at Ravenna in historical times as derived from archeological finds. The main rebuildings of the city are shown: wood on palafittes (a); port development from I century B.C. to II century A.D. (b); empire capital (V century A.D.) and capital of barbarian reigns (VI century A.D.) (c); Venetian domination (XV century A.D.) (d); 1, 2, 3: time of the floor upheaval of the paleochristian basiliques. Reproduced from Gambolati G, Teatini P, Tomasi L and Gonella M (1999) *Water Resources Research*, 35, 163–184, originally drawn by Roncuzzi (1986), by permission of American Geophysical Union

estimated by relating the age of archeological finds to their depth below the present ground surface. An example of such an approach is offered by the natural settlement of the Ravenna area, Italy, as shown in Figure 5, suggesting an average rate of 2.5 mm/year.

IN SITU COMPACTION/EXPANSION MEASUREMENTS

A major task in managing anthropogenic land subsidence due to groundwater pumping or gas/oil production is the *in situ* measurement of the related medium deformation. First, a continuous monitoring of the ongoing compaction of the depleted formations provides useful information about the environmental consequences of the extraction, and also gives a response to some important technical and operational issues, such as the break or damage of well casings (Friedrich *et al.*, 2000) or the decrease of the formation porosity and permeability with the consequent reduction of the field production life (Ruistuen *et al.*, 1999). Second, the availability of *in situ* compaction data, together with piezometric or reservoir pressure measurements, allows for a realistic evaluation of the geomechanical properties of the producing units, which is essential for a reliable prediction of the expected land subsidence. The main techniques currently available to measure the *in situ* deformation of depleted formations are borehole extensometers for shallow aquifer-aquitard systems and radioactive markers for deep gas/oil reservoirs.

Borehole Extensometry

The first borehole extensometer was installed in 1955 by the US Geological Survey in the San Joaquin Valley, California (Poland and Yamamoto, 1984). Since then, the extensometry technology has experienced a progressive development, playing an important role in relating land subsidence to the compaction of confined aquifer systems. For a review of early and recent developments of this methodology, see Lofgren (1961), Riley (1986), and Heywood (1995). The measuring equipment is schematically shown in Figure 6. A typical extensometer tool consists of a balance beam carrying a cable or a pipe, which is fastened at one end to an anchor weight at the bottom of the compacting system, and at the other end to a counterweight so as to keep the cable at a constant tension. A computer-controlled system records the compaction data versus time. The instrumental precision is very much dependent on the actual extensometer implementation, but a nominal strain resolution of a few parts out of 10^7 can be achieved over a depth from 200 down to 1000 m (Riley, 1986; Heywood, 1995). Recent applications of borehole extensometry for the compaction measurement of shallow sediments have been made in the

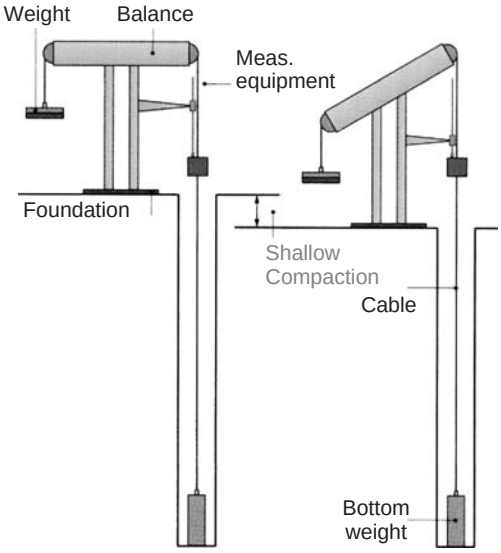


Figure 6 Schematic representation of a borehole extensometer. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Houston area(down to a depth of 936 m) and in the El Hueco basin, Texas, the Albuquerque basin, New Mexico, and above the Groningen gas field, Netherlands.

To be of major interest, borehole extensometers are often coupled to piezometers which record the hydraulic head variations in the depleted units. In the case of multiaquifer systems, several extensometers are usually installed at different depths, so as to monitor the compaction of each single formation. Plotting the vertical strain data versus the pore pressure decrease measured by piezometers during short pumping tests provides a direct *in situ* estimate of the porous medium vertical compressibility, and hence allows for the assessment of the specific storage of the aquifer system. Because of the small impact of water table lowering on the effective stress during a short pumping test, a unit decrease in measured pore pressure corresponds to a unit increase in effective stress, and a stress-strain plot can be drawn giving an estimate of both elastic and inelastic storage coefficients (Riley, 1969). In addition, the time-response characteristics of the system can be used to derive an average vertical permeability for the aquitards. The values of specific storage and hydraulic conductivity thus obtained can be used in numerical simulations to provide a more accurate prediction of land subsidence due to groundwater pumping.

Radioactive Markers

One of the most promising techniques for *in situ* monitoring of deep reservoir compaction relies on the use of weakly radioactive markers. After some pioneering applications performed as early as 1949 in the Wilmington oil field, California (Allen, 1969b), the radioactive marker technique was established 30 years ago in the Groningen gas field (de Loos, 1973) and since then continuously improved (Mobach and Gussinklo, 1994). A schematic representation of the measuring procedure is shown in Figure 7. The marker technique is based on the regular monitoring of the distances between a number of low-emission isotopes (usually ^{137}Cs or ^{60}Co), incorporated into bullet-shaped leak-proof steel containers and shot at fixed intervals along the wall of a generally nonproductive vertical well. The location of each bullet after the porous medium has deformed is obtained by drawing a specific tool bearing γ-ray detectors up through the well from the bottom. As currently implemented by Schlumberger and Western Atlas, the initial spacing between two adjacent radioactive sources is dictated by the logging tool geometry and is approximately 10.5 m, with the vertical strain measurements attaining a nominal accuracy of 10^{-4} (i.e. 1 mm). For a review of the technical and operational details of this technique, see Mobach and Gussinklo (1994), and Macini and Mesini (2000). Radioactive markers have been successfully implemented in a number of gas/oil fields worldwide, including the Ekofisk field, North Sea (Menghini, 1989), the

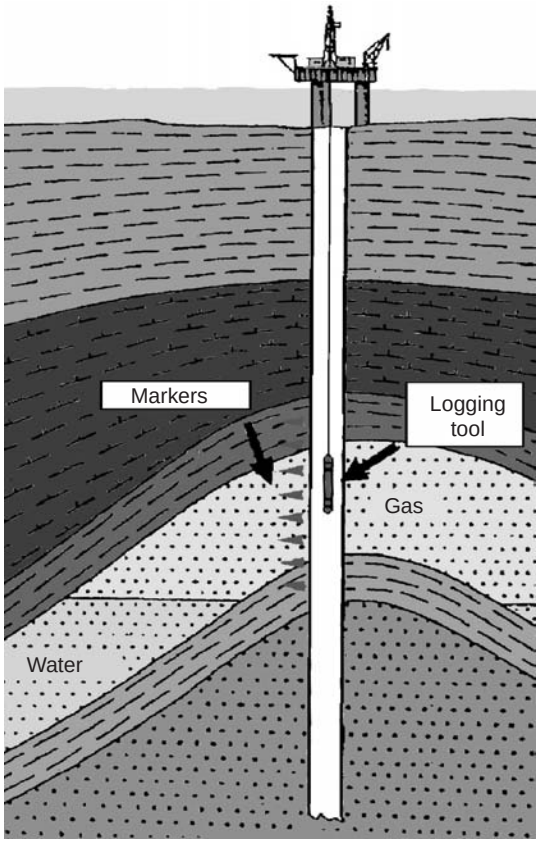


Figure 7 Schematic representation of the radioactive marker technique. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Champion oil field off Brunei (Schmidt, 1996), the Gulf of Mexico area (de Kock *et al.*, 1998), and the Northern Adriatic gas fields, Italy (Cassiani and Zoccatelli, 2000; Baú *et al.*, 2002).

Thin gas/oil reservoirs mainly compact in the vertical direction with a negligible lateral deformation, so that the uniaxial compressibility c_M of the medium encompassed between two consecutive markers can be readily derived as:

$$c_M = \frac{\Delta h}{h_0 \Delta p} \tag{1}$$

where Δh is the measured shortening, h_0 the initial spacing, and Δp the pore pressure variation provided by the available logging tools. The c_M values can be quite scattered, because of the uncertainty affecting each measurement, so that statistical procedures are often required to homogenize the data. Baú *et al.* (2002) suggest using the moving weighted average technique and regressing c_M versus the corresponding vertical effective stress σ_{ev} to obtain an average representative constitutive relationship for the monitored rock formations. It is usually recognized that *in situ* measurements offer more reliable estimates of the actual rock compressibility than laboratory core measurements, mainly because of the preservation of the *in situ* stress field in the tested formation and the elimination of any disturbance and damage during coring and laboratory operations.

Equation (1) provides a straightforward c_M evaluation if Δh indicates compaction and Δp a pore pressure decrease, as expected in producing reservoirs, or if Δh denotes expansion and Δp a pore pressure rise, as is the case in the postproduction recovery phase. In expansion, a smaller c_M is usually found, corresponding to unloading-reloading conditions. However, expansions with a decreased pore pressure, or compactions with zero Δp can be recorded in the field (Ferronato *et al.*, 2004), and of course these measurements should not be processed to get the c_M estimate. Such occurrences can be caused by the lithostratigraphy of the monitored rock units, and, to a lesser extent, by the hydro-mechanical coupling between fluid flow and the porous medium deformation. Ferronato *et al.*'s (2004) study of the Northern Adriatic marker data shows that a geologic configuration with thin producing layers (1–2 m thick or less) incorporated within a low permeable shale matrix can make the marker interpretation quite difficult. This difficulty can be remedied by an *ad hoc* nonstandard analysis using advanced numerical models as is done, for example, by Ferronato *et al.* (2003).

MODELING AND PREDICTING LAND SUBSIDENCE

As mentioned before, it has long been understood that land subsidence is best analyzed with reference to Biot's (1941) theory of consolidation, wherein it is recognized that consolidation itself may represent the response of a compressible porous medium to changes of the fluid flow operating within it. Denote E and ν as the Young and the Poisson moduli of porous medium (the bulk), c_{bm}, c_{br}, and β as the volumetric compressibility of the medium, solid matrix and water, respectively, n as the porosity, k as the tensor of the anisotropic hydraulic conductivity of medium saturated with water, and ε as the medium volumetric dilatation (or volume strain). Note that $c_{bm} = 3(1-\nu)/E$. Biot's (1941) stress and flow coupled equations, as modified by van der Knaap (1959) and Geertsma (1966) for a porous medium saturated with groundwater, read:

$$\begin{aligned}
(\lambda + G)\frac{\partial \varepsilon}{\partial x} + G\nabla^2 u_x &= \alpha \frac{\partial p}{\partial x} + X \\
(\lambda + G)\frac{\partial \varepsilon}{\partial y} + G\nabla^2 u_y &= \alpha \frac{\partial p}{\partial y} + Y \\
(\lambda + G)\frac{\partial \varepsilon}{\partial z} + G\nabla^2 u_z &= \alpha \frac{\partial p}{\partial z} + Z
\end{aligned} \tag{2}$$

$$\frac{1}{\gamma}\nabla(k\nabla p) = [n\beta + c_{br}(\alpha - n)]\frac{\partial p}{\partial t} + \alpha \frac{\partial \varepsilon}{\partial t} \tag{3}$$

where $\alpha = 1 - c_{br}/c_{bm}$ is the Biot coefficient, $G = E/2(1+\nu)$ is the shear modulus of the porous medium, $\lambda = \nu E/(1-2\nu)(1+\nu)$ is the Lamé constant, γ is the specific weight of water, ∇ is the gradient operator, p is the incremental pore pressure induced by pumping, and u_x, u_y, and u_z are the components of the incremental displacement $\boldsymbol{u}$. In (2), X, Y, and Z are the external incremental distributed load per unit volume. Slightly modified versions of (3), developed by Gambolati *et al.* (1996b) and Bai and Abousleiman (1997), but not addressed in the present analysis, are briefly discussed by Gambolati *et al.* (2000).

In the uncoupled formulation, most appropriate for predicting anthropogenic land subsidence due to fluid removal, the term $\partial\varepsilon/\partial t$ is replaced by $\alpha c_M \partial p/\partial t$. Defining with S_s the specific elastic storage coefficient:

$$S_s = n\beta + c_{br}(\alpha - n) + \alpha^2 c_M \tag{4}$$

where

$$c_M = \frac{(1-2\nu)(1+\nu)}{E(1-\nu)} \tag{5}$$

is the vertical medium compressibility, that is, the rate of volume change of a rock sample subject to a unit change of effective stress σ_{ev} along the vertical axis with lateral expansion precluded, equation (3) becomes:

$$\frac{1}{\gamma}\nabla(k\nabla p) = S_s \frac{\partial p}{\partial t}. \tag{6}$$

In the upper freshwater aquifer system, c_{br} is much smaller than c_{bm}. Hence $\alpha = 1$, and equation (6) takes on the form that is most familiar to hydrogeologists:

$$\frac{1}{\gamma}\nabla(k\nabla p) = (n\beta + c_M)\frac{\partial p}{\partial t}. \tag{7}$$

For an elegant discussion of some theoretical assumptions leading to (7), see Verruijt (1969). If flow is multiphase, as is the case with gas/oil fields, equation (7) is replaced by more complex mass conservation equations describing the fluid-dynamics of each single phase (e.g. Aziz and Settari, 1979) and p in equations (2) is a weighted mean of the pressures of each phase. It should be noted that the uncoupled formulation, equations (2) and (6) or (7), does not neglect the influence of deformation on fluid flow but lumps it into a single parameter, namely, the elastic storage coefficient, which accounts for a local pointwise deformation and discards the strength source originating from the remainder of the porous medium and generating the theoretical coupling (Gambolati, 1974, 1977). Gambolati *et al.* (2000) have shown that the separate use of S_s is quite appropriate in problems of land subsidence caused by fluid withdrawal. Also note that uncoupling does not imply a 1-D vertical deformation, as is erroneously assumed by some authors (e.g. Lewis and Schrefler, 1998). Rather, uncoupling simply postulates that fluid-dynamics is well represented by equations of mass conservation where the structural effects on fluid flow are accounted for by S_s. Generally, under the influence of the incremental pore pressure p, the medium undergoes the fully 3-D deformation described by equations (2).

A frequent geological setting of an alluvial aquifer system is that where horizontal or subhorizontal aquifers and aquitards alternate (Figure 3). In this case, flow within the aquifers may be described by a 2-D horizontal equation of the form:

$$\frac{\partial}{\partial x}\left(T_{xi}\frac{\partial p_i}{\partial x}\right) + \frac{\partial}{\partial y}\left(T_{yi}\frac{\partial p_i}{\partial y}\right) = S_i\frac{\partial p_i}{\partial t} + q_j - q_{j+1} \tag{8}$$

where T_{xi} and T_{yi} are the components of the anisotropic hydraulic transmissivity (equal to $b_i k_{xi}$ and $b_i k_{yi}$, respectively, with b_i the ith aquifer thickness and k_{xi} and k_{yi} the anisotropic hydraulic conductivity), $S_i = S_{si} b_i$, and q_j and q_{j+1} are leakage from the overlying j and underlying $j+1$ aquitards. Flow within aquitards j is governed by the 1-D vertical flow equation:

$$\frac{\partial}{\partial z}\left(k_{zj}\frac{\partial p_j}{\partial z}\right) = S_{sj}\frac{\partial p_j}{\partial t} \tag{9}$$

Equations (8) and (9) are hydraulically coupled, as the solution to (9) depends on p_i and p_{i+1}, that is, the pore pressure in the aquifers above and below aquitard j, and in turn q_j and q_{j+1} in equation (8) are calculated from the solutions p_j and p_{j+1} of the aquitards j and $j+1$ overlying and underlying aquifer i. For the implementation of the solution to equations (8) and (9) in a multiaquifer-aquitard system, see Gambolati *et al.* (1986) and Gambolati and Teatini (1996).

Once the pressure variations p_i $(i = 1, \ldots, N)$ and $p_j (j = 1, \ldots, N+1)$, with N the number of individual aquifers within the system, have been predicted by solving equations (8) and (9) for all i and j over the entire multiaquifer, land subsidence $\eta(t)$ at time t since pumping started is obtained simply as the cumulative units compaction:

$$\sum_{j=1}^{N+1}\sum_{k=1}^{M} 0.434\ \gamma(1-n_j)\frac{C_c}{\sigma_{\mathrm{ev}j}} p_{kj}(t)\Delta b_{kj} + \sum_{i=1}^{N} c_{Mi}\, p_i(t) b_i \tag{10}$$

where M is the number of vertical elements into which each aquitard is discretized, C_c is the aquitard compression index, and p_{ki} and Δb_{ki} are the pore pressure decline and the length of the vertical spacing, respectively, in element k of aquitard j. By distinction, c_{Mi}, p_i and b_i are the ith aquifer compressibility, pressure drawdown, and thickness. C_c is calculated from oedometric tests as the slope in a semilogarithmic diagram of the void ratio versus σ_{ev}. Equation (10) is used when aquitard j compacts on the virgin compaction curve with $\sigma_{\mathrm{ev}j} > \sigma_{\mathrm{ev}j,\mathrm{prec}}$, where $\sigma_{\mathrm{ev}j,\mathrm{prec}}$ is the preconsolidation stress. Porosity n_j, jth aquitard permeability k_{zj} and specific elastic storage S_{sj} can be updated by the empirical nonlinear relationships (Rivera *et al.*, 1991):

$$\begin{aligned} n_j &= n_{0j} - 0.434\ C_c(1-n_j)^2\frac{\Delta\sigma_{\mathrm{ev}j}}{\sigma_{\mathrm{ev}j}} \\ &\qquad (\sigma_{\mathrm{ev}j} > \sigma_{\mathrm{ev}j,\ \mathrm{prec}}) \\ k_{zj}(n_j) &= k_{z0j}\left[\frac{n_j(1-n_{0j})}{n_{0j}(1-n_j)}\right]^m \\ S_{sj} &= \gamma\left(0.434\ C_c\frac{1-n_j}{\sigma_{\mathrm{ev}j}} + n_j\beta\right) \\ &\qquad (\sigma_{\mathrm{ev}j} > \sigma_{\mathrm{ev}j,\ \mathrm{prec}}) \end{aligned} \tag{11}$$

with n_{0j} and k_{0j} the initial porosity and hydraulic conductivity, respectively, and m a material-dependent coefficient. C_c must be replaced by the swelling index C_s for $\sigma_{\mathrm{ev}j} < \sigma_{\mathrm{ev}j,\mathrm{prec}}$.

MITIGATION OF LAND SUBSIDENCE BY FLUID INJECTION

Assessing alternative strategies to mitigate anthropogenic land subsidence requires a level of technical support and

a cost-benefit analysis, which are outside the scope of the present paper. Moreover, each individual damage-prevention strategy also comes with an associated set of socioeconomic issues that influence its financial, legal, and political feasibility. The present review is restricted to the most straightforward and apparently simplest action for mitigating land subsidence caused by fluid withdrawal: artificial fluid injection. It goes without saying that other strategies can contribute to the prevention of land subsidence, and these include the policy for groundwater and gas/oil pumping exerted by the central and local authorities such as, withdrawal limits, permits, fees, taxes, metering, and enforcement control activities. Freeze (2000) conveys the general recommendation that land subsidence should be included as a basic driver when defining a fluid exploitation management strategy along with more traditional drivers, namely, water table decline, saltwater intrusion, and avoidance of groundwater contamination.

Generally speaking, when land subsidence has occurred and/or is still occurring, methods to control or mitigate or arrest it include reduction of pumping rates, artificial aquifer recharge from land surface, repressurization of depleted layers through wells, creation of a hydraulic barrier to stop the advancement of the cone of depression, and the generation of an overpressure in geological units that are not affected by pumping in order to build a structural obstacle to the migration of in-depth compaction to ground surface. A combination of any of the above methods can be used as well consistent with the outcome of a cost-benefit analysis. An example of conservative mitigation strategy is one whereby the effective stress within the depleted formation is not further increased beyond the stress level experienced to date. A more aggressive strategy could dictate a decrease of the effective stress and/or the active involvement of overlying formations through the use of fluid injection.

The local geologic conditions suggest whether artificial recharge, which is usually implemented on the basis of a conservative action, may be accomplished by application at the land surface or by repressurizing the aquifers through injection wells. The mitigation program in the Santa Clara Valley (CA) is an example of controlled percolation released from artificial surface reservoirs built close to the larger streams (Poland, 1984c). To reduce the pumping rate, one or more of the following solutions must be implemented (Poland, 1984b): import of substitute surface water, conservation in application and use of water (for instance, by improving the existing irrigation methods and/or changing crops with less-water demanding ones), reuse of treated water by industrial plants, development of well fields in more permeable deposits, as has been done in Shanghai (Liu, 2002), or enforcement of legal control.

Artificial replenishment of a confined aquifer system by application of water at the ground surface directly above the pumped system may not be practicable in many places. However, streams may be used which are connected to the aquifer system by creating surface reservoirs somewhere upstream along the watercourse. In some areas, the aquifer may crop out at or near the margins of the groundwater basin. So, recharging artificially on the outcrop area may be effective in replenishing the depleted formations with a mitigation of the downstream groundwater level decline, and hence of the related land settlement.

Repressurizing confined aquifers through artificial wells may prove a most effective and practical way to cope with anthropogenic land subsidence, especially in coastal areas where recharging reservoirs are difficult to build. The Wilmington oil field in Southern California offers an excellent example of land subsidence mitigation and control through injection wells (Colazas and Strehle, 1995). The water injection started on a major scale in 1958. Eleven years later when $2\,m^3/s$ was being pumped into the oil field, the settling area had been reduced from 58 to $8\,km^2$ with local land surface rebound equal to 0.3 m. However, the maximum subsidence had achieved peak values of 9 m (Table 1). Replenishing groundwater supplies by artificial recharge through wells and pits has been practiced in many sites in California and in Long Island, NY, with generally satisfactory results. In a few cases, problems of wells and aquifer clogging were reported, depending on the actual chemical and physical properties of the injected water. Injection of treated freshwater into a confined aquifer system to create a hydraulic barrier against seawater intrusion has been implemented successfully in Southern California. A paper by Rancilio (1977) describes in detail the typical design of an efficient injection well, cost, rates and head, clogging problems, and operating conditions. The operational injection heads ranged from 9 to 61 m, depending on injection depth, with injection rates from 6 to 28 l/s.

In Italy, ENI-E&P has recently implemented a project of reinjection of formation water below 3500-m depth to generate a barrier against the propagation onshore of the depressurization in the deep Angela–Angelina gas field which partly underlies the coastline. The initial injection rate from the borehole Angelina 1 is planned at $150\,m^3/d$ with the understanding that it can be appreciably increased using seawater, should it prove necessary to obtain a more important mitigation of the coastline subsidence induced by the ongoing gas production.

A numerical simulation of seawater injection beneath the lagoon has recently been performed to raise Venice (Comerlati *et al.*, 2003). As is widely known, Venice has experienced a relative sinking (natural and anthropogenic land subsidence plus average sea level rise) equal to 23 cm over the last century (Carbognin *et al.*, 2005). This has enhanced the Adriatic seawater ingress into the lagoon with a significant increase in the frequency of periodic flooding,

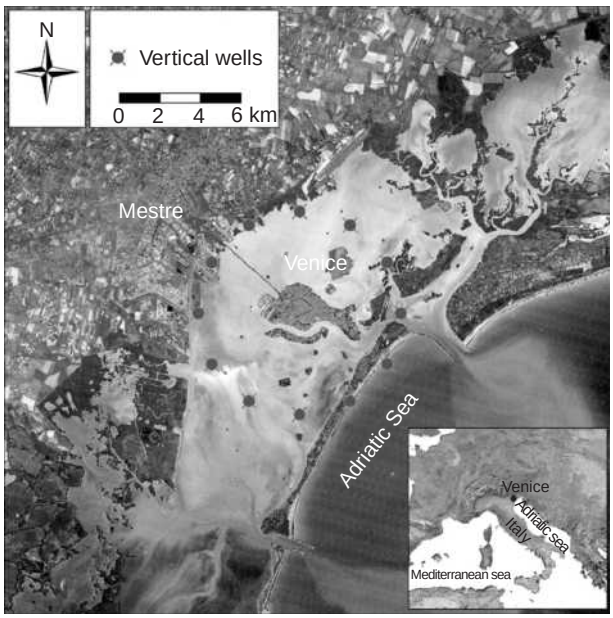


Figure 8 Space view of the Venice Lagoon with the location of the injection wells (Reproduced by permission of AGU from Comerlati *et al.,* 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the so-called "acqua alta", over the last 50 years. Having identified a suitable geologic formation lying between 600 and 800-m depth beneath the city, saturated with brackish water, an injection rate of 23 Mm^3/y of seawater has been simulated through 12 injection wells properly located around Venice (Figure 8) with the purpose to uplift the lagoon bottom so as to make up for the 23-cm loss experienced by Venice, and thus mitigate the city exposure to the "acqua alta". Based on new information and results from recent studies on the evolution of the Northern Adriatic coastline (Gambolati, 1998), and prediction of land settlement above Adriatic gas fields (Gambolati *et al.*, 1999; Baú *et al.*, 2000; Teatini *et al.*, 2000), which have promoted an in-depth knowledge of the subsurface near the Venice Lagoon, numerical analyses have been carried out with the aid of advanced fluid-dynamic and structural finite element models. The simulations aim at evaluating the amount of rebound Venice might experience because of the pore overpressure induced by seawater injection. Essentially new geomechanical data for a realistic prediction of the surface rebound have become available quite recently (Baú *et al.*, 2002; Ferronato *et al.*, 2003, 2004). Geology and hydrology of the upper freshwater aquifer system are known from previous investigations (Gambolati and Freeze, 1973; Gambolati *et al.*, 1974). Figure 9(a) shows the predicted overpressure distribution at 700-m depth 10 years after the beginning of injection. Figure 9(b) provides the expected rebound of Venice at the same time.

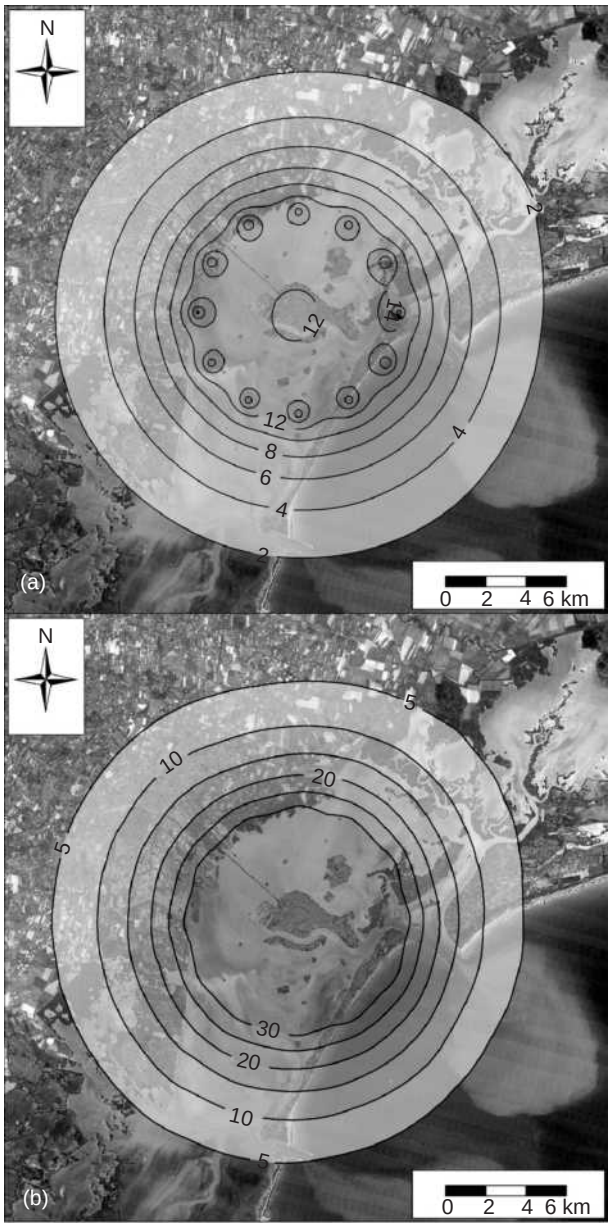


Figure 9 (a) Overpressure (MPa) and (b) expected land uplift (cm) in the Venice Lagoon 10 years after the beginning of injection (Reproduced by permission of AGU from Comerlati *et al.,* 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Note the uniform distribution of the rebound that makes the proposal an extremely appealing solution to the life-long problem of the city protection from flooding. Of course, the practical implementation of such a project requires much more information including an in-depth reconnaissance study of the geology and litho-stratigraphy below the lagoon and a full cost analysis of the engineering project needed for constructing the water pipelines and injection boreholes.

REFERENCES

Allen D.R. (1969a) *The Mechanics of Compaction and Rebound, Wilmington Oil Field, Long Beach, California*, Technical Report, Department of Oil Properties: City of Long Beach.

Allen D.R. (1969b) Collar and radioactive bullet logging for subsidence monitoring. *Transactions 10th Annual Logging Symposium*, Society of Professional Well Log Analysts: pp. G1–G19.

Allen S.A. (1984) Types of land subsidence. In *Guidebook to Studies of Land Subsidence Due to Groundwater Withdrawal*, Poland J.F. (Ed.), UNESCO: Paris, pp. 133–142.

Allis R.G. (2000) Review of subsidence at Wairakei field, New Zealand. *Geothermics*, **29**, 455–478.

Amelung F., Galloway D., Bell J.W., Zebker H.A. and Laczniak R.J. (1999) Sensing the ups and downs of Las Vegas: InSAR reveals structural control of land subsidence and aquifer-system deformation. *Geology*, **27**, 483–486.

Aziz K. and Settari A. (1979) *Petroleum Reservoir Simulation*, Applied Science Publishers: London.

Bai M. and Abousleiman Y. (1997) Thermoporoelastic coupling with application to consolidation. *International Journal for Numerical and Analytical Methods in Geomechanics*, **21**, 121–132.

Baú D., Gambolati G. and Teatini P. (1999) Residual land subsidence over depleted gas fields in the Northern Adriatic basin. *Environmental & Engineering Geoscience*, **5**, 389–405.

Baú D., Gambolati G. and Teatini P. (2000) Residual land subsidence near abandoned gas fields raises concern over Northern Adriatic coastland. *EOS Transactions-American Geophysical Union*, **81**, 245–249.

Baú D., Ferronato M., Gambolati G. and Teatini P. (2002) Basin-scale compressibility of the Northern Adriatic by the radioactive marker technique. *Géotechnique*, **52**, 605–616.

Biot M.A. (1941) General theory of three-dimensional consolidation. *Journal of Applied Physics*, **12**, 155–164.

Bitelli G., Bonsignore F. and Unguendoli U. (2000) Leveling and GPS networks to monitor ground subsidence in the southern Po valley. *Journal of Geodynamics*, **30**, 355–369.

Bolzon G. and Schrefler B.A. (1997) Compaction in gas reservoir due to capillary effects. In *Computational Plasticity, Fundamentals and Applications*, Owen D.R.J., Oñate E. and Hinton E. (Eds.), CIMNE Publishers: Barcelona, pp. 1625–1630.

Brutsaert W. and Corapcioglu M.Y. (1976) Pumping of aquifer with visco-elastic properties. *Journal of the Hydraulics Division-ASCE*, **102**, 1663–1675.

Carbognin L., Teatini P. and Tosi L. (2005) Land subsidence in the venetian area: known and recent aspects. *Italian Journal of Engineering Geology and Environment*, in press.

Cassiani G. and Zoccatelli C. (2000) Towards a reconciliation between laboratory and In Situ measurements of soil and rock compressibility. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 3–15.

Chilingarian G.V. and Knight L. (1960) Relationship between pressure and moisture content of kaolinite, illite and montmorillonite clays. *Bulletin of the American Association of Petroleum Geologists*, **44**, 101–106.

Colazas X.C. and Strehle R.W. (1995) Subsidence in the Wilmington oil field, Long Beach, California, USA. In *Subsidence due to Fluid Withdrawal*, Chilingarian G.V., Donaldson E.C. and Yen T.F. (Eds.), Elsevier Science B V: Amsterdam, pp. 2857–2334.

Comerlati A., Ferronato M., Gambolati G., Putti M. and Teatini P. (2003) Can CO_2 help save Venice from the sea? *EOS Transactions of the American Geophysical Union*, **84**(546), 552–553.

Cui Y.J. and Delage P. (1996) Yielding and plastic behaviour of unsaturated compacted silt. *Géotechnique*, Barla G. **46**, 291–311.

de Kock A.J., Johnson J., Hagiwara T., Zea H.A. and Santa F. (1998) Gulf of Mexico subsidence monitoring project with a new formation-compaction monitoring tool. *SPE Drilling & Completion*, **1**, 223–230.

Delage P., Cui Y.J. and Schroeder C. (1996) Subsidence and capillary effects in chalks. In *EUROCK'96*, Barla G. (Ed.), A. A. Balkema, Rotterdam, pp. 1291–1298.

de Loos J.M. (1973) In Situ compaction measurements in Groningen observation wells. *Verhandelingen van het Koninklijk Nederlands geologisch mijnbouwkundig Genootschap*, **28**, 79–104.

Doornhof D. (1992) Surface subsidence in the Netherlands: the Groningen gas field. *Geologie En Mijnbouw*, **71**, 119–130.

Ferretti A., Prati C. and Rocca F. (2000) Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 2202–2212.

Ferronato M., Gambolati G., Teatini P. and Baú D. (2001) Land surface uplift above compacting overconsolidated reservoirs. *International Journal of Solids and Structures*, **38**, 8155–8169.

Ferronato M., Gambolati G., Teatini P. and Baú D. (2004) Radioactive marker measurements in heterogeneous reservoirs: a numerical study. *International Journal of Geomechanics*, **4**, 79–92.

Ferronato M., Gambolati G., Teatini P. and Baú D. (2003) Interpretation of radioactive marker measurements to evaluate compaction in the Northern Adriatic gas fields. *SPE Reservoir Evaluation & Engineering*, **6**, 401–411.

Figueroa-Vega G.E. (1984) Case History No. 9.8, Mexico, D.F., Mexico. In *Guidebook to Studies of Land Subsidence due to Groudwater Withdrawal*, Poland J.F. (Ed.), UNESCO: Paris, pp. 217–232.

Finol A. and Sancevic Z.A. (1995) Subsidence in Venezuela. In *Subsidence due to Fluid Withdrawal*, Chilingarian G.V., Donaldson E.C. and Yen T.F. (Eds.), Elsevier Science B V: Amsterdam, pp. 337–372.

Floyd R.P. (1978) *Geodetic Bench Marks*, NOAA Manual NOS NGS1: Rockville.

Freeze R.A. (2000) Social decision making and land subsidence. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 1, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 353–384.

Friedrich J.T., Arguello J.G., Deitrick G.L. and de Rouffignac E.P. (2000) Geomechanical modeling of reservoir compaction, surface subsidence, and casing damage at the Belridge diatomite field. *SPE Reservoir Evaluation & Engineering*, **3**, 348–359.

Gabriel A.K., Goldstein R.M. and Zebker H.A. (1989) Mapping small elevation changes over large areas: differential

radar interferometry. *Journal of Geophysical Research*, **94**, 9183–9191.

Gabrysch R.K. (1984) The Houston-Galveston region, Texas, U.S.A. In *Guidebook to Studies of Land Subsidence due to Groundwater Withdrawal*, Poland J.F. (Ed.), UNESCO: Paris, pp. 253–262.

Gabrysch R.K. and Neighbors R.J. (2000) Land-surface subsidence and its control in the Houston-Galveston region. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 81–92.

Galloway D. and Riley F.S. (1999) San Joaquin valley, California: largest human alteration of the earth's surface. In *Land Subsidence in the United States*, Galloway D., Jones D.R. and Ingebritsen S.E. (Eds.), U S Geological Survey Circular 1182: pp. 23–34.

Gambardella F., Bortolotto S. and Zambon M. (1991) The positioning systems GPS for subsidence control of the terminal reach of the Po river. In *Land Subsidence. Proceedings of 4th International Symposium on Land Subsidence*, Johnson A.I. (Ed.), IAHS Publication No. 200: pp. 433–441.

Gambolati G. (1974) Second-order theory of flow in three-dimensional deforming media. *Water Resources Research*, **10**, 1217–1228.

Gambolati G. (1977) Deviations from Theis solution in aquifers undergoing three-dimensional consolidation. *Water Resources Research*, **13**, 62–68.

Gambolati G. (1992) Comment on coupling versus uncoupling in soil consolidation. by Lewis RW, Schrefler BA and Simoni L *International Journal for Numerical and Analytical Methods in Geomechanics*, **16**, 833–837.

Gambolati G. (Ed.) (1998) *CENAS, Coastline Evolution of the Upper Adriatic Sea due to Sea Level Rise and Natural and Anthropogenic Land Subsidence*, Water Science & Technology Library, No.28, Kluwer Academic Publishers: Dordrecht.

Gambolati G. and Freeze R.A. (1973) Mathematical simulation of the subsidence of Venice. 1. Theory. *Water Resources Research*, **9**, 721–733.

Gambolati G., Gatto P. and Freeze R.A. (1974) Mathematical simulation of the subsidence of Venice. 2. Results. *Water Resources Research*, **10**, 563–577.

Gambolati G., Putti M. and Teatini P. (1996a) Coupled and uncoupled poroelastic solutions to land subsidence due to groundwater withdrawal. In *ASCE Engineering Mechanics Conference, Mini Symposium on Poroelasticity*, LinY.K. and Su T.C. (Eds.), ASCE: New York, pp. 483–486.

Gambolati G., Putti M. and Teatini P. (1996b) Land subsidence. In *Hydrology of Disasters*, Singh V.P. (Ed.), Kluwer Academic Press: Dordrecht, pp. 231–268.

Gambolati G., Ricceri G., Bertoni W., Brighenti G. and Vuillermin E. (1991) Mathematical simulation of the subsidence of Ravenna. *Water Resources Research*, **27**, 2899–2918.

Gambolati G., Sartoretto F. and Uliana F. (1986) A conjugate gradient finite element model of flow for large multiaquifer systems. *Water Resources Research*, **22**, 1003–1015.

Gambolati G. and Teatini P. (1996) A block iterative finite element model for non-linear leaky aquifer systems. *Water Resources Research*, **32**, 199–204.

Gambolati G., Teatini P. and Bertoni W. (1998) Numerical prediction of land subsidence over Dosso degli Angeli gas field, Ravenna, Italy. In *Land Subsidence - Current Research and Case Studies. Proceedings of the Joseph F. Poland Symposium on Land Subsidence, Sacramento (CA), October 1995*, Borchers J. (Ed.), Star Publishing Company: Belmont, pp. 229–238.

Gambolati G., Teatini P., Baú D. and Ferronato M. (2000) Importance of poro-elastic coupling in dynamically active aquifers of the Po river basin, Italy. *Water Resources Research*, **36**, 2443–2459.

Gambolati G., Teatini P., Tomasi L. and Gonella M. (1999) Coastline regression of the Romagna region, Italy, due to sea level rise and natural and anthropogenic land subsidence. *Water Resources Research*, **35**, 163–184.

Geertsma J. (1966) Problems of rock mechanics in petroleum production engineering. *Proceedings of 1st Congress International Society of Rock Mechanics*, Lisbon, pp. 585–594.

Geertsma J. (1973) Land subsidence above compacting oil and gas reservoirs. *Journal of Petroleum Technology*, **25**, 734–744.

Gonella M., Gambolati G., Giunta G., Putti M. and Teatini P. (1998) Prediction of land subsidence due to groundwater withdrawal along the Emilia-Romagna coast. In *CENAS, Coastline Evolution of the Upper Adriatic Sea due to Sea Level Rise and Natural and Anthropogenic Land Subsidence*, Gambolati G. (Ed.), Kluwer Academic Publishers: Dordrecht, pp. 151–168.

Harris Galveston Coastal Subsidence District (1982) *Water Management Study: Phase 2 and Supplement 1*, Houston, p. 68.

Helm D.C. (1975) One-dimensional simulation of aquifer system compaction near Pixley, California. 1. Constant parameters. *Water Resources Research*, **11**, 465–478.

Helm D.C. (1976) One-dimensional simulation of aquifer system compaction near Pixley, California. 1. Stress-dependent parameters. *Water Resources Research*, **12**, 375–391.

Hermansen H., Landa H.A., Sylte J.E. and Thomas L.K. (2000) Experiences after 10 years of waterflooding the Ekofisk Field, Norway. *Journal of Petroleum Science and Engineering*, **26**, 11–18.

Heywood C. (1995) Investigation of aquifer-system compaction in the Hueco Basin, El Paso, Texas, USA. In *Land Subsidence*, Barends F.B.J., Brouwer F.J.J. and Schröder F.H. (Eds.), IAHS Publication No. 234: pp. 35–45.

Holzer T.L. (1981) Preconsolidation stress of aquifer systems in areas of induced land subsidence. *Water Resources Research*, **17**, 693–704.

Hu R.L., Wang S.J., Lee C.F. and Li M.L. (2002) Characteristics and trends of land subsidence in Tanggu, Tianjin, China. *Bulletin of Engineering Geology and the Environments*, **61**, 213–225.

Kenselaar F. and Martens M. (2000) Spatial temporal modelling of land subsidence due to gas extraction. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 383–396.

Lewis R.W. and Schrefler B.A. (1998) *The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media*, John Wiley: Chichester.

Liu Y. (2002) A strategy of groundwater distribution exploitation to mitigate the magnitude of land subsidence. *Proceedings of Annual Meeting of the Geological Society of America (October 27–30)*, Denver.

Liu C.H., Tu F.L., Huang C.T., Ouyang S., Chang K.C. and Tsai M.C. (2000) Current status of land subsidence in Taiwan. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 205–212.

Lofgren B.E. (1961) Measurement of compaction of aquifer systems in areas of land subsidence. *U.S. Geological Survey Professional Paper*, Vol. 424-B, B49–B52, Geological Survey Research.

Macini P. and Mesini E. (2000) Compaction monitoring from radioactive marker technique. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 43–55.

Martin J.C. and Serdengecti S. (1984) Subsidence over oil and gas fields. In *Man-Induced Land Subsidence, Reviews in Engineering Geology*, Vol. 6, Holzer T.L. (Ed.), Geological Society of America: Boulder, pp. 23–34.

Menghini M.L. (1989) Compaction monitoring in the Ekofisk area chalk fields. *Journal of Petroleum Technology*, **41**, 735–739.

Mobach E. and Gussinklo H.J. (1994) In Situ reservoir compaction monitoring in the Groningen field. *EUROCK 94: Rock Mechanics for Petroleum Engineering*, AA Balkema Publishers: Rotterdam, pp. 535–547.

Narasimhan T.N. and Goyal K.P. (1984) Subsidence due to geothermal fluid withdrawal. In *Man-Induced Land Subsidence, Reviews in Engineering Geology*, Vol. 6, Holzer T.L. (Ed.), Geological Society of America: Boulder, pp. 35–66.

Nutalaya P., Yong R.N., Chumnankit T. and Buapeng S. (1996) Land subsidence in Bangkok during 1978–1988. In *Sea-Level Rise and Coastal Subsidence: Causes, Consequences, and Strategies*, MillimanJ.D. and Haq B.U. (Eds.), Kluwer Academic Publisher: Dordrecht, pp. 105–130.

Poland J.F. (1984a) Mechanics of land subsidence due to fluid withdrawal. *Guidebook to Studies of Land Subsidence due to Groundwater Withdrawal*, UNESCO: Paris, pp. 37–54.

Poland J.F. (1984b) Review of methods to control or arrest subsidence. *Guidebook to Studies of Land Subsidence due to Groundwater Withdrawal*, UNESCO: Paris, pp. 127–130.

Poland J.F. (1984c) Santa Clara valley, California, U.S.A. *Guidebook to Studies of Land Subsidence due to Groundwater Withdrawal*, UNESCO: Paris, pp. 127–130.

Poland J.F. and Lofgren B.E. (1984) San Joaquin valley, California, U.S.A. In *Guidebook to Studies of Land Subsidence due to Groundwater Withdrawal*, Poland J.F. (Ed.), UNESCO: Paris, pp. 263–277.

Poland J.F. and Yamamoto S. (1984) Field measurement of deformation. In *Guidebook to Studies of Land Subsidence due to Groundwater Withdrawal*, Poland J.F. (Ed.), UNESCO: Paris, pp. 17–35.

Pratt W.E. and Johnson D.W. (1926) Local subsidence of the goose creek oil field. *Journal of Geology*, **34**, 577–590.

Rancilio J.A. (1977) *Injection Well Operation and Maintainance*, in IAHS Publication No. 121: pp. 325–333.

Rappleye H.S. (1948) *The Manual of Geodetic Leveling*, NOAA Special Publication No. 239: Rockville.

Riley F.S. (1969) Analysis of borehole extensometer data from central California. In *Land Subsidence*, Vol. 2, Tison L.J. (Ed.), IAHS Publication No. 89: pp. 423–431.

Riley F.S. (1986) Developments in borehole extensometry. In *Land Subsidence*, Johnson A.I., Carbognin L., Ubertini L. (Eds.), IAHS Publication No. 151: pp. 169–186.

Rivera A., Ledoux E. and de Marsily G. (1991) Nonlinear modeling of groundwater flow and total subsidence of the Mexico city aquifer-aquitard system. In *Land Subsidence. Proceedings of 4th International Symposium on Land Subsidence*, Johnson A.I. (Ed.), IAHS Publication No. 200: pp. 44–58.

Roncuzzi A. (1986) Ravenna nei tempi antichi. *Classe e Ravenna*, **III**(1–2), 2–4.

Ruistuen H., Teufel L.W. and Rhett D.W. (1999) Influence of reservoir stress path on deformation and permeability of weakly cemented sandstone reservoirs. *SPE Reservoir Evaluation & Engineering*, **2**, 266–272.

Sato H.P., Abe K. and Ootaki O. (2003) GPS-measured land subsidence in Ojiya City, Niigata Prefecture, Japan. *Engineering Geology*, **67**, 379–390.

Schmidt T. (1996) The measurement of the subsidence of geological formations with position sensitive detectors: application to the oil fields. *Proceedings of 5th European Union Hydrocarbon Symposium*, Edinburgh, pp. 1028–1061.

Schrefler B.A., Simoni L. and Zhang H.W. (2000) Capillary effects in reservoir compaction and surface subsidence. In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 267–280.

Strozzi T., Wegmüller U., Tosi L., Bitelli G. and Spreckels V. (2001) Land subsidence monitoring with differential SAR interferometry. *Photogrammetric Engineering & Remote Sensing*, **67**, 1261–1270.

Strozzi T., Tosi L., Wegmüller U., Werner C., Teatini P. and Carbognin L. (2003) Land subsidence monitoring service in the lagoon of Venice. *Proceedings of International Geoscience and Remote Sensing Symposium*, Toulose, France CD-ROM.

Teatini P., Baú D. and Gambolati G. (2000) Water-gas dynamics and coastal land subsidence over Chioggia Mare field, Northern Adriatic sea. *Hydrogeology Journal*, **8**, 462–479.

Theis C.V. (1935) The relationship between the lowering of the piezometric surface and the rate and duration of discharge of a well using groundwater storage. *EOS Transactions of the American Geophysical Union*, **16**, 519–524.

Tosi L., Carbognin L., Teatini P., Rosselli R. and Gasparetto-Stori G. (2000) The ISES project subsidence monitoring of the catchment basin south of the Venice lagoon (Italy). In *Proceedings of 6th International Symposium on Land Subsidence*, Vol. 2, Carbognin L., Gambolati G. and Johnson A.I. (Eds.), La Garangola: Padova, pp. 113–126.

Tosi L., Carbognin L., Teatini P., Strozzi T. and Wegmüller U. (2002) Evidence of the present relative land stability of Venice, Italy, from land, sea, and space observation. *Geophysical Research Letters*, **29**, doi 10.1029/2001GL013211.

U.S. Army Corps of Engineers (1996) *NAVSTAR Global Positioning System Surveying. Engineer Manual 1110-1-1003*.

U.S. Coast Guard Navigation Center (1996) *NAVSTAR GPS User Equipment Introduction*.

van der Knaap W. (1959) Nonlinear behaviour of elastic porous media. *Petroleum Transactions of the AIME*, **216**, 179–187.

Verruijt A. (1969) Elastic storage of aquifers. In *Flow Through Porous Media*, De Wiest R.J.M. (Ed.), Academic Press: New York, pp. 331–376.

Yamamoto S. (1995) Recent trend of land subsidence in Japan. In *Land Subsidence*, Barends F.B.J., Brouwer F.J.J. and Schröder F.H. (Eds.), IAHS Publication No. 234: pp. 487–492.

Yao Z. and Bi E. (2000) Environmental imprints on the temporal changes of the chemistry of geothermal water from some low-temperature geothermal fields in the North China plain.*Proceedings of World Geothermal Congress*, Kyushu-Tohoku: Japan, pp. 1979–1984.

Zaman M.M., Abdulraheem A. and Roegiers J.C. (1995) Reservoir compaction and surface subsidence in the north sea Ekofisk field. In *Subsidence due to Fluid Withdrawal*, Chilingarian G.V., Donaldson E.C. and Yen T.F. (Eds.), Elsevier Science B V: Amsterdam, pp. 373–423.

PART 14

Snow and Glacier Hydrology

159: Snow Cover

ROSS D BROWN[1] AND BARRY E GOODISON[2]

[1] *Meteorological Service of Canada, Dorval, QC, Canada*
[2] *Meteorological Service of Canada, Downsview, Canada*

Snow cover is encountered over most of the Northern Hemisphere (NH) mid- and high-latitudes during the winter season and over many mountainous regions of the world for extended periods. Snow cover is an important component of the climate system through its role in modifying energy and moisture fluxes between the surface and the atmosphere, and through its role as a water store in hydrological systems. Snow also plays critical roles in ecological and biological systems, and in nutrient and carbon cycling. This article provides an introductory overview of snow cover including: definition of terms (snowfall, solid precipitation, snow depth, snow density, snow water equivalent); a review of methods for measuring snow cover properties; a discussion of the processes and concepts involved in the spatial and temporal variability of snow cover; a look at avalanches and recent efforts to model these; and finally, a discussion of how snow cover is changing in response to global warming.

INTRODUCTION

Snow cover refers to the blanket of snow covering the ground, and includes the concepts of depth and areal extent (Sturm *et al.*, 1995). Snow cover is encountered over most of the Northern Hemisphere (NH) mid- and high latitudes during the winter season, and over many mountainous regions of the world for extended periods. Figure 1 shows the mean seasonal range in snow cover over the NH from satellite data covering the 1971 to 1995 period (data from NSIDC, 1996). The temporal variability is dominated by the seasonal cycle with average NH snow cover extent ranging from an average minimum extent of 3.6 million km^2 in August to an average maximum extent of 46.8 million km^2 in late-January. Most of the global snow cover extent is located over the NH: with the Antarctic and Greenland landmasses excluded, the Southern Hemisphere (SH) mean maximum terrestrial snow cover extent is less than 2% of the corresponding winter maximum snow cover extent over the NH. In the SH exclusive of Antarctica, most of the seasonal snow cover extent is located over South America.

Snow cover is a key component of the global climate system through its role in modifying energy and moisture fluxes between the surface and the atmosphere, and through its role as a water store in hydrological systems. Snow is a highly reflective material; the reflectivity or albedo of new snow is 0.8 to 0.9, which means that 80–90% of the incident solar energy is reflected away from the snow surface. Densely forested regions, however, will exhibit an albedo as low as 0.2 even with complete ground snow cover. A high albedo, combined with the excellent insulating characteristics of a snow cover dramatically reduces the energy exchange between the surface and the atmosphere. Empirical studies have shown that mean surface air temperatures are typically 5 °C colder when a snow cover is present (Groisman and Davies, 2001). This positive feedback is responsible for the rapid expansion of NH snow cover extent in October–November and is manifest in a typically close negative relationship between snow cover and air temperature (Figure 2).

The larger-scale climatic significance of the snow-albedo feedback is modulated by cloud cover and by the smaller amount of total solar radiation received in high latitudes during winter months. Groisman *et al.* (1994a) observed that snow cover exhibited the greatest influence on the Earth's radiative balance in the NH spring (April to May) period when incoming solar radiation was greatest over snow-covered areas. The impact of snow on surface reflectivity is an example of a direct feedback to the climate system. Snow is also involved in a number of indirect feedbacks through its thermal insulating properties (e.g. influence on sea ice growth and ground temperatures),

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

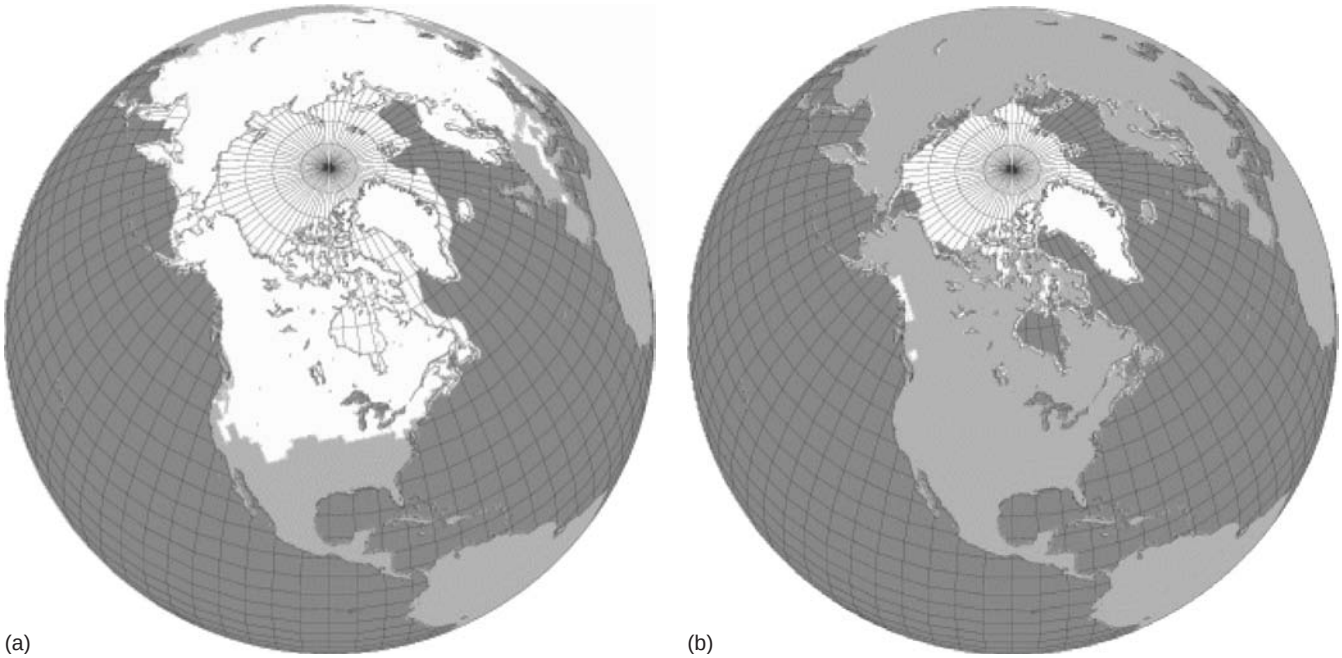


Figure 1 Mean seasonal variation in northern hemisphere snow and sea ice extent between (a) January and (b) August as derived from satellite data. Data are from Weekly Snow Cover and Sea Ice Extent, National Snow and Ice Data Center, 1996

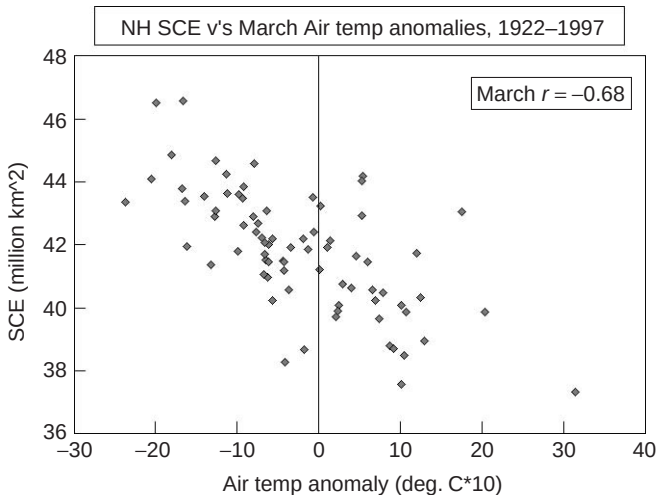


Figure 2 Scatterplot of reconstructed (1922–1971) and satellite-observed (1972–1997) NH snow cover extent from Brown (2000) versus NH midlatitudinal (40–60°N) land surface temperature anomalies for March. Air-temperature anomalies were computed from the Jones (1994) gridded land temperature dataset. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and snow's role in soil moisture recharge which generates indirect feedbacks to cloud cover and surface temperature (Cess *et al.*, 1991; Randall *et al.*, 1994). The influence of snow on the freeze/thaw status of ground temperatures is particularly important for the hydrology of cold regions. There is an extensive body of literature linking snow cover feedbacks to monsoon circulations (e.g. Vernekar *et al.*, 1995; Gutzler and Preston, 1997). However, recent research suggests snow's role in monsoon circulations may be more limited (Robock *et al.*, 2003). Groisman and Davies (2001) provide detailed reviews of the climatic significance of snow.

MEASUREMENT OF SNOWFALL AND SNOW COVER

Because of the profound influence snow has on natural ecosystems (Jones *et al.*, 2001) and the human environment, snow measurement science has a long and rich history extending from Chinese snow cages in the first millennium (Biswas, 1970) to ultrasonic snow sensors employed on today's automatic weather stations. Reviews of *in situ* snow observing methods are provided by Goodison *et al.* (1981), Pomeroy and Gray (1995), and Doesken and Judson (1997). The main properties of a snow cover that are routinely monitored *in situ* are: snow depth, snow water equivalent, and snowfall (amount and water equivalent). Strictly speaking, snowfall is not a property of a snow cover, but it is included here because of its essential role in the initiation and accumulation of a snowpack. In addition, the solid/liquid fraction of precipitation and the density and crystal structure of new snowfall have important consequences for snowpack evolution and internal structure.

Prior to the late 1800s, there were few systematic measurements of snow depth. However, routine observations were made of snow cover-related variables such as the number of days with snowfall, number of days with snow on the ground, or first/last dates of snow on the ground. The presence of snow on the ground is one of the earliest regularly recorded snow cover observations.

Snowfall and Solid Precipitation

Accurate precipitation data (adjusted for systematic errors) are essential to balance the energy and water cycles in the climate system, for climate monitoring, for determining the global and regional hydrological balance and for understanding key components of the cryosphere, such as, snow-covered area, snow water equivalent and glacier mass balance. Precipitation is expected to increase in response to global warming (IPCC, 1996) and there is evidence (Bradley *et al.*, 1987; Vinnikov *et al.*, 1990; Groisman and Easterling, 1994; Mekis and Hogg, 1999) that precipitation has indeed exhibited a significant upward trend over the last ~100 years. Recent warming has also been associated with a reduction in the solid:liquid ratio of precipitation in shoulder seasons (Karl *et al.*, 1993; Zhang *et al.*, 2000). Accurate information on precipitation intensity, timing, and solid/liquid fractions are essential to correctly simulate snowpack development and snowmelt with detailed energy balance models such as SNTHERM (Jordan, 1991).

It is important to make clear the distinction between snowfall and solid precipitation. Snowfall is the depth of freshly fallen snow that accumulates during the observing period and has been traditionally measured with a ruler. Solid precipitation is the equivalent liquid water of the snowfall intercepted by a precipitation gauge and is influenced by the catch characteristics of the particular gauge. At voluntary observing stations in Canada and the United States, the depth of new snowfall is measured once or twice per day using a ruler and snowboard. In many countries, solid precipitation is estimated from daily total snowfall assuming a fresh snowfall density of $100\,\mathrm{kg\,m^{-3}}$. However, the density of freshly fallen snow varies widely from 10 to $30\,\mathrm{kg\,m^{-3}}$ for dry, cold "wild snow" (Seligman, 1980) to more than $150\,\mathrm{kg\,m^{-3}}$ for warm, wet snow. Most fresh snowfall densities fall within a range of 50 to$120\,\mathrm{kg\,m^{-3}}$ (Pomeroy and Gray, 1995), but commonly exhibit large temporal and spatial variations (Goodison *et al.*, 1981). Manual ruler observations of snowfall are subject to numerous sources of error, the most important being the blowing and drifting of snow, and the melting or rapid settling of snow when it reaches the ground. For example, Goodison *et al.* (1981) monitored average density increases between 8 and $13\,\mathrm{kg\,m^{-3}\,h^{-1}}$ during snow storms with duration less than 12 h. Doesken and Judson (1997) provide a good overview of some of the practical problems associated with observing snowfall.

Solid precipitation is measured with some form of precipitation gauge that can range from a standard rain gauge to a specially shielded snow gauge. A fundamental problem of gauge measurement of snow precipitation is that most precipitation gauges catch less falling snow than the "true" amount because accelerated wind flow over the top of the gauge reduces the number of snowflakes able to enter the orifice. This effect increases with wind speed, and for a relatively modest wind speed of $15\,\mathrm{km\,hr^{-1}}$, an unshielded US standard 8" precipitation gauge has been found to catch only ~50% of the "true" snowfall (Yang *et al.*, 1998). This effect can be reduced by using shielding devices such as the Alter shield, which will increase the catch by 20 to 70% (Goodison *et al.*, 1998; Yang *et al.*, 1999).

Precipitation gauges, shields, and observing practices vary considerably from country to country, and over time (Goodison *et al.*, 1998; Sevruk and Klemm, 1989 and for USSR – Groisman *et al.*, 1991; USA – Groisman and Easterling, 1994; Canada – Metcalfe *et al.*, 1997). Other important systematic sources of error include instrument siting, trace precipitation amounts and wetting loss (the amount of water adhering to the inside of a gauge and not measured each time it is emptied). Metcalfe and Goodison (1993) showed that when trace precipitation amounts were adjusted for wetting loss, they could account for a significant increase in corrected precipitation totals (~30% for a prairie site). This correction is particularly important in the Arctic where some stations report over 80% of precipitation observations as trace amounts (Metcalfe and Goodison, 1993). Guidelines for the correction of commonly used precipitation gauges are provided by Goodison *et al.* (1998).

The recent trend toward increased automation of climate observations has important consequences for the homogeneity of precipitation measurement series. Goodison *et al.* (1998) recommended that heated tipping-bucket rain gauges not be used for measuring snowfall due to excessive evaporation loss. Weighing gauges were found to be the most practical but these can introduce a "timing" error due to snow or freezing precipitation sticking to the inside of the gauge and melting at some later time. Automated gauges, like manual ones, can also catch blowing snow and provide no information on the liquid/solid fractions of precipitation. These problems complicate the real-time interpretation of the data as well as the application of procedures to adjust for systematic errors.

Snow Depth

Snow depth is the most obvious property of a snowpack. It is, however, less useful from a water and energy budget perspective; snow depth can change independently from snow mass due to processes such as metamorphism and melt/refreeze events (Fitzharris *et al.*, 1992). Manual observations of daily snow depth have been carried out in association with regular meteorological observation programs

at principal climate stations in most countries with a seasonal snow cover. Typically, these networks have evolved in response to needs for weather and climate information in support of economic activities, and the spatial distribution of stations tends to follow the population distribution with poor spatial coverage in mountainous and remote areas. In general, there are few stations reporting snow depths before the early 1900s, after which the number of stations increases to a maximum during the 1970s followed by a rapid reduction in the 1990s, often associated with budget reductions and automation. A number of countries have systematic snow depth observations going back to the late 1800s for example, Switzerland (Foehn, 1990), USA (Easterling *et al.*, 1997), the former Soviet Union (Armstrong, 2001) and Finland (Kuusisto, 1984). Canadian snow depth data are available from the mid-1950s (MSC, 2000). When working with these data sets, it is important to remember that snow depth observations are subject to the same sources of inhomogeneity as other climatological elements, for example, changes in station location, changes in observing time, changes in measuring units, changes in observers, and urban effects (warming and dirtying of snow from pollution). Quality control procedures to check the internal consistency of daily snow depth observations were developed for US data by Robinson (1989). It is also important to note that many *in situ* snow depth observing sites are in open locations (e.g. airports) and may not be representative of snow conditions over the surrounding area, especially in surrounding vegetated areas. This is an issue when comparing *in situ* snow depth observations to snow cover information derived from remotely sensed sources.

Automated *in situ* snow depth measurements can be made with ultrasonic snow depth sensors which compute the distance to the snow surface from the time taken for a pulse of sound to reach the snow surface and be reflected back (Goodison *et al.*,1985). Snow depth can be measured with an accuracy of 1 cm, and the sensor can be interrogated at frequent regular time intervals (e.g. hourly) to obtain a detailed history of snow depth changes from settling, wind erosion, and melt that is particularly useful for validation of physical snow process models (Metcalfe *et al.*, 1987). Anomalous measurements may occur during falling or blowing snow, and the ultrasonic sensor may underestimate the depth of very low density freshly fallen snow (Goodison *et al.*, 1985). This requires the application of quality control logic to remove spurious values. An advantage of the automated depth sensor is that the snow layer is undisturbed, and the measurements reflect changes at that measurement point (manual snow depth measurements involve some subjectivity for drifting or patchy snow covers). A limitation of the auto sensor is that it only measures snow depth at a single point (a circular area with diameter 0.2 to 2 m depending on the height of the sensor above the snow surface – Pomeroy and Gray, 1995). This makes careful site selection essential. Another technique for *in situ* measurement of snow depth is the use of vertical thermistor arrays to deduce the location of the snow/air boundary from the observed temperature gradient. Portable Ground Penetrating Radar (GPR) systems linked to GPS systems provide the capability to map snow depth over a range of scales and import the information into GIS systems. These systems are used operationally for snowpack management at some ski centers.

Several attempts have been made to develop global snow depth climatologies from *in situ* observations (US Army Corps of Engineers, 1954; Schutz and Bregman, 1975; Foster and Davy, 1988). The Foster and Davy (1988) snow depth climatology provides global monthly mean snow depth at a ~50 km resolution and has been used extensively in the validation of GCM snow simulations (e.g. Foster *et al.*, 1996). Brown *et al.* (2003) developed a 37 -km gridded snow depth climatology for North America based on a modified version of the operational snow depth analysis scheme used at the Canadian Meteorological Centre (Brasnett, 1999). This climatology was based on extensive *in situ* snow depth observations from the United States and Canada covering snow seasons 1979/80 to 1996/97, and exhibited a number of improvements over the earlier Foster and Davy (1988) product. Figure 3 shows the monthly mean snow depth over NA in February.

A number of remotely sensed methods can be applied to retrieve information on snow depth (for a discussion of the theory of remote sensing of snow see Hall and

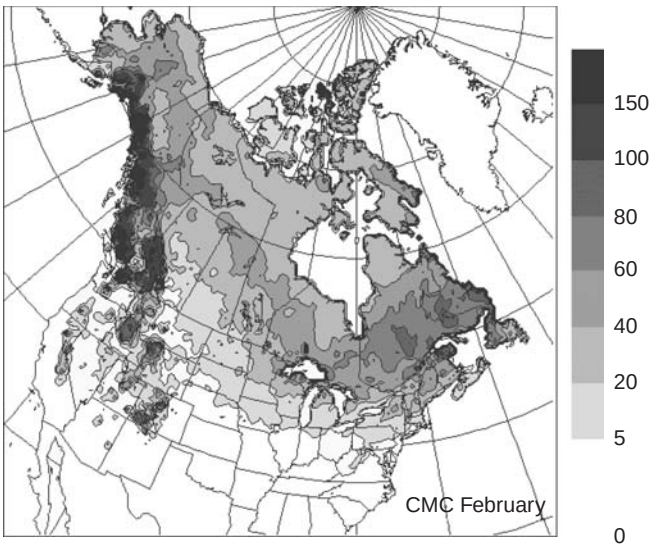


Figure 3 February monthly mean snow depth (cm) over North America from a blend of historical daily snow depth observations and snow model simulations over the period 1979–1997 (Brown *et al.*, 2003). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Martinec, 1985). More recent developments in the remote sensing of snow are covered in detail in Hall *et al.* (2005) (*see* **Chapter 55, Estimation of Snow Extent and Snow Properties, Volume 2**). Snow depth information has been estimated from passive microwave satellite data (e.g. Chang *et al.*, 1987), which has the advantage of all-weather coverage. Passive microwave data are not routinely used for snow depth mapping as they are sensitive to liquid water in the snow, vegetation cover, and seasonal changes in snowpack structure. The derivation of snow depth from passive microwave data also requires that certain assumptions be made about snow density and grain size. For example, the global SMMR-derived monthly snow depth climatology used by Foster *et al.* (1996) assumed a mean snow density of $300\,\mathrm{kg\,m^{-3}}$ and an average snow grain size of 0.3 mm (Chang *et al.*, 1987). High-resolution airborne laser altimetry can provide information on snow depth information in forested terrain through its ability to see between trees (Hopkinson *et al.*, 2001), but this is a relatively expensive mapping tool, and requires one presnow season survey to map the surface elevation without snow. Hopkinson *et al.* (2001) determined that the method was only suitable for deeper snowpack conditions $>\sim$50 cm. Other remotely sensed methods for mapping snow depth include GPR and SAR interferometry, but these are mainly used in research.

Snow Water Equivalent (SWE)

SWE is defined in the International Classification for Seasonal Snow on the Ground (Colbeck *et al.*, 1990) as the depth of water if a snow cover is completely melted, expressed in millimeters, on a corresponding horizontal surface area. SWE is related to snow depth and density by the following relationship

$$\mathrm{SWE(mm)} = 0.01\, h_s * \rho_s \qquad (1)$$

where h_s is the depth of snow (cm) and ρ_s is the density of snow ($\mathrm{kg\,m^{-3}}$). The conversion from a mass of snow ($\mathrm{kg\,m^{-2}}$) to a depth of water (mm) is based on the fact that 1 mm of water spread over an area of $1\,\mathrm{m^2}$ weighs 1 kg.

The most commonly used approach for determining SWE is the gravimetric method which involves taking a vertical core through the snowpack, and weighing or melting the core to obtain the SWE. A variety of coring and weighing systems have been used around the world with varying lengths and diameters depending on measurement units and local snow conditions (see Sevruk, 1992). One of the earliest national SWE observing networks was established in Finland in 1909 (Kuusisto, 1984). However, systematic observation of SWE was not widespread until the middle of the twentieth century. In order to obtain representative values of SWE, measurements are often carried out at regular marked intervals along a permanently marked transect or "snow course". There are many factors involved in the design of a snow course (e.g. purpose, accessibility, terrain, vegetation) and the reader is referred to Goodison *et al.* (1981) for a detailed discussion. The length of a snow course and the number of sampling points depends on the desired level of accuracy and the spatial variability of snow cover. There is extensive literature on this subject for example, Goodison *et al.* (1981), WMO (1981), Sevruk (1992), Pomeroy and Gray (1995).

The accuracy of manual SWE measurements from snow samplers is discussed in detail in Sevruk (1992). The main systematic error (due to instrumentation) is related to a tendency for additional snow to be forced up into a tube as it is pushed through a snowpack. Hence, the design of the cutter and its sharpness is a main contributor to performance and accuracy of a snow sampler (Goodison *et al.*, 1981). Random errors associated with observers and snow conditions include the difficulty in keeping loose granular snow in the sampler, drainage of water from very wet snow, ice crusts, and sampling very shallow, patchy snow cover. Snow course measurements are often carried out on a weekly or biweekly basis, although not all courses have regular measurements throughout the snow cover season. In many cases, measurements made by operational agencies or utilities for runoff management are confined to the late winter and early spring period as the main interest is in determining the peak SWE prior to melt. This can limit the usefulness of some snow course data for climate-related studies. Other limitations for using SWE surface observations in climate-related studies are: uneven spatial distribution of data, relatively short periods of continuous observations, data quality (see Schmidlin, 1990), and a general lack of easily accessible data. Snow survey data sets have been published for Canada (MSC, 2000) and the former Soviet Union (Krenke, 1998).

Automated surface-based observations of SWE are possible from devices such as snow pillows that measure the mass of snow over a small area from displaced fluid or a pressure transducer. Snow pillows are usually octagonal or circular in shape, with an area of $\sim$5–10 $\mathrm{m^2}$, and are most effective for monitoring relatively deep snowpacks in sheltered environments. Interpretation of data can be complicated by "bridging" (from ice or hard snow layers in the snowpack) or by the draining of wet snow (Pomeroy and Gray, 1995). Snow pillows are usually hooked to a land-line or satellite data transmission systems to provide real-time information such as the US Department of Agriculture (USDA) Snow Telemetry (SNOTEL) network over the western United States (Rallison, 1981). An advantage of these automated systems is that they can be interrogated on a daily basis to provide more detailed information during the melt season than regular weekly or biweekly snow course observations. Daily values of SWE from over 600 snow pillow sites in the western United States from 1979

are available from the USDA National Water and Climate Center. A comprehensive analysis of these data for the 1980 to 1998 period was provided by Serreze *et al.* (1999).

Other less commonly employed methods for *in situ* measurement of SWE or snow density include gamma ray attenuation and microwave radar (see Pomeroy and Gray, 1995). The Swiss Federal Institute for Snow and Avalanche Research are developing a system to automatically measure the density, the water equivalent and the liquid water content of a snow cover over a distance of 20 to 30 m using electrical impedance ("SNOWPOWER"). The sensor consists of long flat-band cables along which the dielectric constant is measured (in the time domain) and the impedance is measured at different low frequencies.

Accurate information on the amount and spatial distribution of SWE is essential for water resource management (e.g. agriculture, hydro-electric power generation, flood forecasting). Water authorities and utilities make use of surface-based snow surveys (and snow pillows in mountainous terrain) to monitor peak SWE values prior to snowmelt, as well as satellite imagery to map snow cover extent. In mountainous areas, the elevation of the snowline is particularly important information for water resource planners. SWE information can also be derived from a variety of aircraft or satellite sensors with varying degrees of success (Rango *et al.*, 2000). The most commonly used operational systems are airborne gamma surveys (Foster *et al.*, 1987), passive microwave (Foster *et al.*, 1984) and SAR (Bernier and Fortin, 1998; Baghdadi *et al.*, 2000). Each system has its own particular set of advantages and limitations (see Rango *et al.*, 2000) and in some case these can be exploited by using multiple sensors, for example, wet snow is not seen by passive microwave but is seen by SAR. Reviews of satellite remote sensing of SWE are provided by Rango (1996) and Hall *et al.* (2005) (*see* **Chapter 55, Estimation of Snow Extent and Snow Properties, Volume 2**). The derivation of consistent spatial SWE information over a range of land cover surfaces and terrain types requires the blending of *in situ*, satellite- and model-derived information. Considerable progress has been made in this area in recent years (Hartman *et al.*, 1995; Carroll *et al.*, 2001).

SNOW COVER INTERANNUAL VARIABILITY

Regional and continental snow cover exhibit large interannual variability in response to atmospheric circulation patterns that influence temperature and precipitation. There is an extensive body of literature documenting significant correlations between Eurasian winter snow cover extent and the North Atlantic Oscillation (NAO) and the related Arctic Oscillation (AO) mode of atmospheric variability (e.g. Gutzler and Rosen, 1992; Bamzai, 2003), and recent work (Gong *et al.*, 2003; Bojariu and Gimeno, 2003; Saito and Cohen, 2003) suggests that snow cover variations play a role in exciting and maintaining dominant NH winter patterns such as AO-NAO. Over North America, snow cover interannual variability is most closely related to the Pacific–North America (PNA) pattern (Gutzler and Rosen, 1992; Serreze *et al.*, 1998). The snow cover response to ENSO (El Niño–Southern Oscillation) is less pronounced and exhibits large regional variability (e.g. Cayan, 1996; Clark *et al.*, 2001; Hsieh and Tang, 2001) with a tendency for El Niño events to be associated with greater snow cover over Eurasia and less snow cover over North America (Groisman *et al.*, 1994b). A detailed review of snow cover–atmosphere relationships is provided by Groisman and Davies (2001).

DISTRIBUTION OF SNOW COVER

As outlined in Pomeroy and Gray (1995), the areal distribution of snow cover reflects processes and terrain/vegetation influences that occur over a range of spatial scales. At the regional-continental scale ($\sim$10–1000 km) factors such as proximity to moisture sources, location of winter storm tracks, orography, elevation, and latitude are dominant. At the local scale, ($\sim$100 m–10 km) terrain features (e.g. ridge/valley) and vegetation exert important influences, while at the microscale ($\sim$10–100 m), small-scale variations in topography (slope, aspect, surface roughness) and vegetation (height, density, interception efficiency) affect snow accumulation and ablation patterns through their influence on drifting and blowing snow and sublimation, canopy interception, and the surface energy budget. Extensive reviews of terrain and vegetation influences on snow cover are provided in McKay and Gray (1981), Kind (1981, 1986), Woo (1982), Sevruk (1992), Pomeroy and Gray (1995), and Marsh (1999). The understanding and parameterization of wind redistribution, blowing snow sublimation and canopy interception/sublimation processes is critical for accurate simulation of the energy and water balances in cold regions over a range of scales. For example, Pomeroy and Gray (1995) estimated sublimation loss over prairie environments to be 15 to 41% of annual snowfall precipitation. In forested environments, they observed that approximately one-third of total snowfall falling on spruce and pine was lost due to canopy sublimation.

The physical understanding of blowing snow and blowing snow sublimation has been incorporated into a number of models, notably: PBSM (Pomeroy *et al.*, 1993), PIEKTUK (Déry *et al.*, 1998), and SnowTran-3D (Liston and Sturm, 1998). When coupled with simple models for wind flow over complex terrain, these models are capable of generating realistic patterns of snow accumulation in various environments (e.g. Essery *et al.*, 1999; Greene *et al.*, 1999). One area of uncertainty is the sensitivity of the blowing snow sublimation process to relative humidity.

Time-dependent diffusion models such as PIEKTUK generate an increase in relative humidity and decrease in air temperature in the surface layer during blowing snow events, so that sublimation becomes self-limiting over longer fetches. Humidity observations during blowing snow events (Mann *et al.*, 2000) confirm that humidity does in fact increase, but not to the extent predicted by PIEKTUK. Déry and Yau (2001) concluded that entrainment and advective processes reduce the effect of the self-limiting feedback, and that blowing snow sublimation rates were 1.8 times larger than those predicted by the stand-alone application of PIEKTUK. There are still a number of uncertainties in the theory of turbulent transport of snow particles (e.g. their effective settling velocity, and the appropriate specification of particle concentrations and size distributions at or near the ground surface).

Forest canopy interception and sublimation are important processes for the water and energy budgets of large regions of the northern hemisphere with dense coniferous forests. Reviews of snow interception process and modeling research are provided by Hedstrom and Pomeroy (1998) and Marsh (1999). A coupled forest interception–sublimation model was developed by Pomeroy *et al.* (1998). As noted by Marsh (1999), there are numerous physically based algorithms controlling snow interception, sublimation, melt, and unloading. However, the challenge is to incorporate this knowledge into regional and global climate models, many of which have only rudimentary treatment of snow-canopy processes. Another challenge facing the modeling community is the availability of detailed datasets that can be used to assess the performance of snow-vegetation models.

Statistical Characterization of Snow Cover

The parameterization of the important physical processes, and topographic and land-cover factors affecting the spatial distribution of snow cover and its properties is a major challenge for climate and hydrological models. In the case of climate models, computational overheads are an important concern, and surface processes must be handled on grid dimensions on the order of 100 to 200 km. Realistic representation of snow processes into this framework requires an understanding of the scale that processes operate on, and the development of techniques to parameterize important subgrid scale processes. Knowledge of scale is also important when comparing datasets such as *in situ* observations and remotely sensed data.

A comprehensive review of the theory of scaling issues in snow hydrology was provided by Bloschl (1999). Bloschl (1999) defined three different types of scale: process scale, measurement scale, and model scale, and defined the term "scaling" to denote a change in scale which happens whenever interpolation, extrapolation, aggregation, and disaggregation are carried out. The "variogram" is fundamental to the concept of scale. The characteristic length of a variable can be estimated from the variogram which is essentially the between-data variance computed as a function of data separation (distance). Fitting the experimental variogram with an analytical function (often an exponential function) yields the spatial variance and the correlation length. For linear processes, (i.e. where simple averaging can be applied) such as mapping SWE over a basin, the variogram is used in standard geostatistical approaches such as kriging to interpolate data. For nonlinear variables and processes (e.g. the depletion of a patchy snow cover), alternative approaches are required. For example, Woo and Steer (1986) used Monte Carlo methods to account for vegetation effects on snow depth, while Walland and Simmonds (1996) used the cumulative frequency distribution to incorporate subgrid-scale topographic variability into a GCM. Shook and Gray (1996) discussed the application of fractals to describe the spatial distribution of snow depth and SWE in open terrain, and Pomeroy and Schmidt (1993) discussed their use for modeling intercepted snow on vegetation.

AVALANCHES

Avalanches are a well-documented hazard in most mountainous regions of the world, and a hazard that more people are being exposed to with increasing interest in back-country ski-touring and heli-skiing. An extensive review of avalanche theory and management practice is compiled in *The Avalanche Handbook* (McClung and Schaerer, 1993). Considerable progress has been made in recent years simulating the one-dimensional complex layered structure of a snowpack and the internal metamorphic processes such as depth hoar development that contribute to enhanced avalanche risk, for example, the French model CROCUS (Brun *et al.*, 1992) and the Swiss model SNOWPACK (Bartelt and Lehning, 2002; Lehning *et al.*, 2002). CROCUS has been coupled with a mesoscale model for the Alps and an expert system linking avalanche risk to simulated snowpack stability to form a comprehensive tool for regional avalanche risk assessment (Durand *et al.*, 1999). Extensive historical records of avalanches have been compiled in several countries such as Canada (Fitzharris and Shaerer, 1980), the United States (Mock and Birkeland, 2000) and Switzerland (Laternser and Scheebeli, 2002). There is some evidence that avalanche frequency is linked to atmospheric teleconnection patterns such as the PNA in the Rocky Mountains (Mock and Birkeland, 2000) and the NAO in Iceland (Keylock, 2003). However, many of the factors involved in avalanche formation and release operate at daily to weekly timescales, so the potential for seasonal prediction is limited. As noted by Fitzharris and Schaerer (1980), major avalanche winters are not necessarily related to large snowfall winters; it is the formation of weak layers and the

superposition of heavy snowfall and/or rapid advection of warm air that trigger major avalanche episodes. A number of dynamical models of avalanches have been developed to predict avalanche impact pressures on structures and runout length (Naaim and Gurer, 1998; Bartelt *et al.*, 1999; McClung, 2001). These provide quite realistic results but modeling the initiation of snow failure is still an area of ongoing research (Schweizer, 1999; Louchet, 2000).

SNOW COVER AND CLIMATE CHANGE

Snow cover extent (SCE) exhibits significant negative correlations to air temperature in many regions of the NH (see the "temperature response regions" in Groisman *et al.*, 1994b), and for the hemisphere as a whole (Robinson and Dewey, 1990; Karl *et al.*, 1993; Brown, 2000) (also see Figure 2). This negative relationship reflects the positive snow–albedo feedback. Analysis of NH snow cover changes over the recent period of satellite data (since ~1970) has shown that the most significant decreases have occurred in the second half of the snow year when the snow albedo feedback is strongest (Groisman *et al.*, 1994a; Pielke *et al.*, 2000). Satellite records indicate that the Northern Hemisphere annual SCE has decreased by about 10% since 1966 largely due to decreases in spring and summer since the mid-1980s over both the Eurasian and American continents (Robinson, 1999). Analysis of reconstructed SCE since 1915 (Brown, 2000) showed that most of the observed reduction occurred during the second half of the twentieth century. This period has been characterized by widespread trends toward earlier snowmelt (Brown and Braaten, 1998; Stone *et al.*, 2002; Dye, 2002) and earlier snowmelt runoff, with important implications for water resources. For example, reduced storage of water in the snowpack and earlier melt translate to a lower freshwater pulse for recharge of soil moisture and reservoirs, and increased potential for evaporation loss. Global Climate Model (GCM) simulations suggest widespread reductions in snow cover over the next 50 to 100 years in response to global warming. However, there is considerable uncertainty in model-projected regional patterns of snow cover change (Frei and Robinson, 1998) particularly in mountainous regions. A discussion of the impacts of reductions in snow cover is provided in Fitzharris (1996). A fundamental problem for realistic simulation of snow cover by GCMs is the realistic simulation of precipitation by atmospheric models (amount and phase). Evaluation of GCM simulations of winter regional precipitation (IPCC, 1996) documented errors ranging from −25 to +125% of observed precipitation over central North America, and major discrepancies have also been documented in modeled precipitation over high elevation and high latitude regions (e.g. Walsh *et al.*, 1998).

SUMMARY

Snow cover is a key component of the global climate system through its role in modifying energy and moisture fluxes between the surface and the atmosphere, and through its role as a water store in hydrological systems. Snow also plays critical roles in ecological and biological systems, and in nutrient and carbon cycling. The ability to describe the spatial and temporal variability of a snow cover (and its time- and depth-varying physical properties) is fundamental for many scientific needs and applications. This starts with accurate precipitation information for example, precipitation type, solid–liquid fraction, and water equivalent adjusted for systematic errors. Accurate measurement of solid precipitation is a major challenge, however, both in terms of instrumentation and in sampling. Recent trends toward automation of precipitation measurements have further complicated the situation with loss of information on precipitation type and the introduction of timing errors from snow or freezing precipitation sticking to the side of gauges.

Systematic *in situ* observations of snow depth and snow water equivalent are available from manual observations and snowpillows for various periods and locations. In general, the spatial distribution of *in situ* snow observations over the NH is concentrated to the populated midlatitudes (with major gaps in northern latitudes and mountains) and to the period after WWII. Only a few countries (e.g. Finland, Switzerland, former Soviet Union, and USA) have systematic snow depth observations for more than 100 years. Remotely sensed information on snow cover extent and SWE is available from various sources from the late-1960s onward with each sensor having its own particular set of advantages and disadvantages. Recent advances in physical snowpack modeling and the availability of reanalysis datasets have provided a means for systematically assimilating snow cover information from these various sources.

The ability to specify the areal distribution of snow cover and its physical properties at the local scale (~100 m −10 km) has received increasing attention in recent years. In particular, understanding and parameterizating wind redistribution, blowing snow sublimation, and canopy interception/sublimation processes has been determined to be critical for accurate simulation of the energy and water balances in cold regions. This requires an understanding of the fundamental scales on which these processes operate (e.g. Bloschl, 1999).

Considerable progress has been made in simulating the one-dimensional complex layered structure of a snowpack and the internal metamorphic processes such as depth hoar development. These advances have contributed to enhanced capabilities for forecasting avalanche risk in Switzerland and France. A number of dynamical avalanche

models have also been developed in recent years to predict avalanche impact pressures on structures and runout length. These provide quite realistic results but there are still major uncertainties in modeling the initiation of snow failure.

In situ data and satellite observations confirm significant reductions in NH snow cover extent and a trend toward earlier spring snowmelt in many regions of the NH since the 1970s. These changes are consistent with Global Climate Model (GCM) simulations which suggest widespread reductions in snow cover over the next 50 to 100 years in response to global warming. However, there is still considerable uncertainty in model-projected changes in precipitation and snow cover in mountainous regions.

FURTHER READING

Gray D.M. and Male D.H. (Eds.) (1981) *Handbook of Snow*, Pergamon Press Canada Ltd.

Hardy J.P., Albert M.R. and Marsh P. (Eds.) (1999) Snow hydrology – the integration of physical, chemical and biological systems. *Hydrological Processes*, **13**, 2117–2482.

REFERENCES

Armstrong, R. (2001) *Historical Soviet Daily Snow Depth Version 2 (HSDSD)*, National Snow and Ice Data Center: Boulder: CD-ROM. `http://nsidc.org/data/g01092.html`.

Baghdadi N., Gauthier Y., Bernier M. and Fortin J.P. (2000) Potential and limitations of RADARSAT SAR data for wet snow monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 316–320.

Bamzai A.S. (2003) Relationship between snow cover variability and Arctic oscillation index on a hierarchy of time scales. *International Journal of Climatology*, **23**, 131–142.

Bartelt P. and Lehning M. (2002) A physical SNOWPACK model for the Swiss avalanche warning Part I: numerical model. *Cold Regions Science and Technology*, **35**, 123–145.

Bartelt P., Salm B. and Gruber U. (1999) Calculating dense-snow avalanche runout using a Voellmy-fluid model with active/passive longitudinal straining. *Journal of Glaciology*, **45**, 242–254.

Bernier M. and Fortin J.P. (1998) The potential of time series of C-band SAR data to monitor dry and shallow snow cover. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 1–18.

Biswas A.K. (1970) *History of Hydrology*, Elsevier: Amsterdam, pp. 336.

Bloschl G. (1999) Scaling issues in snow hydrology. *Hydrological Processes*, **13**, 2149–2175.

Bojariu R. and Gimeno L. (2003) The role of snow cover fluctuations in multiannual NAO persistence. *Geophysical Research Letters*, **30**, 1156.

Bradley R.S., Diaz H.F., Eischeid J.K., Jones P.D., Kelly P.M. and Goodess C.M. (1987) Precipitation fluctuations over Northern Hemisphere land areas since the mid-19th Century. *Science*, **237**, 171–275.

Brasnett B. (1999) A global analysis of snow depth for numerical weather prediction. *Journal of Applied Meteorology*, **38**, 726–740.

Brown R.D. (2000) Northern Hemisphere snow cover variability and change, 1915–1997. *Journal of Climate*, **13**, 2339–2355.

Brown R.D. and Braaten R.O. (1998) Spatial and temporal variability of Canadian monthly snow depths, 1946–1995. *Atmosphere-Ocean*, **36**, 37–45.

Brown R.D., Brasnett B. and Robinson D. (2003) Gridded North American monthly snow depth and snow water equivalent for GCM evaluation. *Atmosphere-Ocean*, **41**, 1–14.

Brun E., David P., Sudul M. and Brunot G. (1992) A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting. *Journal of Glaciology*, **38**, 13–22.

Carroll T., Cline D., Fall G., Nilsson A., Li L. and Rost A. (2001) NOHRSC operations and the simulation of snow cover properties for the coterminous U.S.. *Proceeding of the 69th Western Snow Conference*, Sun Valley, April 16–19, 2001, 1–10.

Cayan D.R. (1996) Interannual climate variability and snowpack in the western United States. *J. Climate*, **9**, 928–948.

Cess R.D., Potter G.L., Zhang M.-H., Blanchet J.-P., Chalita S., Colman R., Dazlich D.A., Del Genio A.D., Dymnikov V., Galin V., *et al.* 23 others (1991) Interpretation of snow-climate feedback as produced by 17 general circulation models. *Science*, **253**, 888–892.

Chang A.T.C., Foster J.L. and Hall D.K. (1987) Nimbus-7 SMMR derived global snow cover parameters. *Annals of Glaciology*, **9**, 39–44.

Clark M.P., Serreze M.C. and McCabe G.J. (2001) Historical effects of El Nino and La Nina events on the seasonal evolution of the montane snowpack in the Columbia and Colorado River Basins. *Water Resources Research*, **37**, 741–757.

Colbeck S., Akitaya E., Armstrong R., Gubler H., Lafeuille J., Lied K., McClung D. and Morris E. (1990) The International Classification for Seasonal Snow on the Ground. *International Commission on Snow and Ice (IAHS)*, World Data Center A for Glaciology, U. of Colorado: Boulder, pp. 23.

Déry S.J., Taylor P.A. and Xiao J.B. (1998) The thermodynamic effects of sublimating, blowing snow in the atmospheric boundary layer. *Boundary – Layer Meteorology*, **89**, 251–283.

Déry, S.J. and Yau, M.K. (2001) Simulation of an Arctic ground blizzard using a coupled blowing snow-atmosphere model. *Journal of Hydrometeorology*, **2**, 579–598.

Doesken N.J. and Judson A. (1997) *The SNOW Booklet: A Guide to the Science, Climatology and Measurement of Snow in the United States*, Colorado State University: p. 86.

Durand Y., Giraud G., Brun E., Merindol L. and Martin E. (1999) A computer-based system simulating snowpack structures as a tool for regional avalanche forecasting. *Journal of Glaciology*, **45**, 469–484.

Dye D.G. (2002) Variability and trends in the annual snow-cover cycle in Northern Hemisphere land areas, 1972–2000. *Hydrological Processes*, **16**, 3065–3077.

Easterling D.R., Jamason P., Bowman D., Hughes P.Y. and Mason E.H. (1997) *Daily Snow Depth Measurements from 195*

Stations in the United States, Dataset NDP-059, Carbon Dioxide Information Analysis Center: **http://cdiac.esd.ornl.gov**.

Essery R., Li L. and Pomeroy J. (1999) A distributed model of blowing snow over complex terrain. *Hydrological Processes*, **13**, 2423–2438.

Fitzharris B.B. (1996) The Cryosphere: Changes and their Impacts. *Intergovernmental Panel on Climate Change, WGII Report*, Cambridge University Press: pp. 241–265.

Fitzharris B.B., Owens I. and Chinn T. (1992) *Snow and glacier hydrology*, Chap. 5, Waters of New Zealand, New Zealand Hydrological Society Inc.: Wellington, pp. 76–94.

Fitzharris B.B. and Schaerer P.A. (1980) Frequency of major avalanche winters. *Journal of Glaciology*, **26**, 43–52.

Foehn P.M. (1990) Schnee und Lawinen. *Schnee, Eis und Wasser in der Alpen in einer waermeren a Atmosphaere*, Mitteil: VAW/ETH Zuerich, Nr. 10-8, S. pp. 33–48.

Foster J.L., Hall D.K. and Chang A.T.C. (1984) An overview of passive microwave snow research and results. *Reviews of Geophysical and Space Physics*, **22**, 195–208.

Foster J.L., Hall D.K. and Chang A.T.C. (1987) Remote sensing of snow. *EOS Transaction American Geophysical Union*, **68**, 681–684.

Foster, D.J. Jr. and Davy, R.D. (1988) *Global Snow Depth Climatology*, USAFETAC/TN-88/006, USAF Environmental Technical Applications Center, p. 48.

Foster J., Liston G., Koster R., Essery R., Behr H., Dumenil L., Verseghy D., Thompson S., Pollard D. and Cohen J. (1996) Snow cover and snow mass intercomparisons of general circulation models and remotely sensed datasets. *Journal of Climate*, **9**, 409–426.

Frei A. and Robinson D.A. (1998) Evaluation of snow extent and its variability in the Atmospheric Model Intercomparison Project. *Journal of Geophysical Research Atmospheres*, **103**(D8), 8859–8871.

Gong G., Entekhabi D. and Cohen J. (2003) A large-ensemble model study of the wintertime AO-NAO and the role of interannual snow perturbations. *Journal of Climate*, **15**, 3488–3499.

Goodison B.E., Ferguson H.L. and McKay G.A. (1981) Measurement and Data Analysis. In *Handbook of Snow*, Gray D.M. and Male D.H. (Eds.), Chap. 6, Pergamon Press Canada Ltd, pp. 191–274.

Goodison, B.E., Louie, P.Y.T. and Yang, D. (1998) WMO Solid Precipitation Measurement Intercomparison, WMO Instruments and Observing Methods Report No. 67, WMO/TD No. 872.

Goodison B.E., Wilson B. and Metcalfe J. (1985) An inexpensive remote snow depth gauge. *Third WMO Technical Conference on Instruments and Methods of Observation (TECIMO-III)*, World Meteorological Organization: Instruments and observing methods Report No. 22 and WMO/TD No. 50, Geneva, pp. 111–116.

Greene E.M., Liston G.E. and Pielke R.A. (1999) Simulation of above treeline snowdrift formation using a numerical snow-transport model. *Cold Regions Science and Technology*, **30**, 135–144.

Groisman P.Y. and Davies T.D. (2001) Snow cover and the climate system. In *Snow Ecology – an interdisciplinary examination of snow-covered ecosystems*, Jones H.J., Pomeroy J., Walker D.A. and Hoham R. (Eds.), Cambridge University Press, pp. 1–44.

Groisman P.Y. and Easterling D. (1994) Variability and trends of total precipitation and snowfall over the United States and Canada. *Journal of Climate*, **7**, 184–205.

Groisman P.Y., Karl T.R. and Knight R.W. (1994a) Observed impact of snow cover on the heat balance and the rise of continental spring temperatures. *Science*, **263**, 198–200.

Groisman P.Y., Karl T.R. and Knight R.W. (1994b) Changes of snow cover, temperature and radiative heat balance over the Northern Hemisphere. *Journal of Climate*, **7**, 1633–1656.

Groisman P.Y., Koknaeva V.V., Belokrylova T.A. and Karl T.R. (1991) Overcoming biases of precipitation measurement: a history of the USSR experience. *Bulletin of the American Meteorological Society*, **72**, 1725–1733.

Gutzler D.S. and Preston J.W. (1997) Evidence for a relationship between spring snow cover in North America and summer rainfall in New Mexico. *Geophysical Research Letters*, **24**, 2207–2210.

Gutzler D.S. and Rosen R.D. (1992) Interannual variability of wintertime snow cover across the Northern Hemisphere. *Journal of Climate*, **5**, 1441–1447.

Hall, D.K., Kelly, R.E., Foster, J.L. and Chang, A.T.C. (2005) Hydrological application of remote sensing: surface states - snow. Contribution hsa062, Encyclopedia of Hydrological Sciences, John Wiley and Sons.

Hall D.K. and Martinec J. (1985) *Remote Sensing of Ice and Snow*, Chapman & Hall: London, pp. 189.

Hartman R.K., Rost A.A. and Anderson D.M. (1995) Operational processing of multi-source snow data. *Proceeding of the 63rd Western Snow Conference*, Sparks, April, 1995, pp. 147–151.

Hedstrom N.R. and Pomeroy J.W. (1998) Measurements and modelling of snow interception in the boreal forest. *Hydrological Processes*, **12**, 1611–1625.

Hopkinson C., Sitar M., Chasmer L., Gynan C., Agro D., Enter R., Foster J., Heels N., Hoffman C., Nillson J. and St. Pierre R. (2001) Mapping the spatial distribution of snowpack depth beneath a variable forest canopy using airborne laser altimetry. *Proceeding of the 58th Eastern Snow Conference*, Ottawa, Ontario. May 14 – 18, 2001, pp. 253–264.

Hsieh W.W. and Tang B. (2001) Interannual variability of accumulated snow in the Columbia basin, British Columbia. *Water Resources Research*, **37**, 1753–1759.

IPCC (1996) *Climate Change 1995: The Science of Climate Change*, Houghton J.T., Meira Filho L.G., Callander B.A., Harris N., Kattenberg A. and Maskell K. (Eds.), Contribution of WGI to the Second Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press: Cambridge, p. 572.

Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.) (2001) *Snow Ecology*, Cambridge University Press: Cambridge, p. 378.

Jones P.D. (1994) Hemispheric surface air temperature variations: a reanalysis and an update to 1993. *Journal of Climate*, **7**, 1794–1802.

Jordan, R. (1991) *A One-Dimensional Temperature Model for a Snow Cover*, Technical documentation for SNTHERM.89 Special Report 91-16, U.S. Army Corps of Engineers. Cold Regions Research & Engineering Laboratory, Hanover, p. 49.

Karl T.R., Groisman P.Y., Knight R.W. and Heim R.R. Jr (1993) Recent variations of snow cover and snowfall in North America and their relation to precipitation and temperature variations. *Journal of Climate*, **6**, 1327–1344.

Keylock C.J. (2003) The North Atlantic Oscillation and snow avalanching in Iceland. *Geophysical Research Letters*, **30**, 1254.

Kind R.J. (1981) Snow Drifting. in *Handbook of Snow*, Gray D.M. and Male, D.H. (Eds.), Chap. 8, Pergamon Press Canada Ltd, pp. 338–359.

Kind R.J. (1986) Snowdrifting: a review of modelling methods. *Cold Regions Science and Technology*, **12**, 217–228.

Krenke A., NSIDC (1998) *Former Soviet Union Hydrological Snow Surveys*, National Snow and Ice Data Center/World Data Center for Glaciology: Boulder. Digital media, updated 2003. `http://nsidc.org/data/g01170.html`.

Kuusisto E. (1984) *Snow Accumulation and Snowmelt in Finland*, Publications of the Water Research Institute, No. 55, National Board of Waters: Helsinki, p. 149.

Laternser M. and Schneebeli M. (2002) Temporal trend and spatial distribution of avalanche activity during the last 50 years in Switzerland. *Natural Hazards*, **27**, 201–230.

Lehning M., Bartelt P., Brown B., Fierz C. and Satyawali P. (2002) A physical SNOWPACK model for the Swiss avalanche warning Part II: snow microstructure. *Cold Regions Science and Technology*, **35**, 147–167.

Liston G.E. and Sturm M. (1998) A snow-transport model for complex terrain. *Journal of Glaciology*, **44**, 498–516.

Louchet F. (2000) A simple model for dry snow slab avalanche triggering. *Comptes Rendus de L'Académie Des Sciences Série II Fascicule A – Sciences de La Terre et Des Planétes*, **330**, 821–827.

Mann G.W., Anderson P.S. and Mobbs S.D. (2000) Profile measurements of blowing snow at Halley, Antarctica. *Journal of the Geophysical Research – Atmospheres*, **105**(D19), 24491–24508.

Marsh P. (1999) Snowcover formation and melt: recent advances and future prospects. *Hydrological Process*, **13**, 2117–2134.

McClung D.M. (2001) Superelevation of flowing avalanches around curved channel bends. *Journal of the Geophysical Research – Solid Earth*, **106**(B8), 16489–16498.

McClung D. and Schaerer P. (1993) *The avalanche handbook*, The Mountaineers: Seattle, p. 271.

McKay G.A. and Gray D.M. (1981) The distribution of snowcover. In *Handbook of Snow*, Gray D.M. and Male D.H. (Eds.), Chapter 5, Pergamon Press Canada Ltd., 153–190.

Metcalfe J.R. and Goodison B.E. (1993) Correction of Canadian winter precipitation data. *Proceeding of the 8th Symposium on Meteorological Observations and Instrumentation*, 17–22 January 1993, Anaheim, pp. 338–343.

Mekis E. and Hogg W. (1999) Rehabilitation and analysis of Canadian daily precipitation time series. *Atmosphere-Ocean*, **37**, 53–85.

Metcalfe J.R., Routledge B. and Devine K. (1997) Rainfall measurement in Canada: changing observational methods and archive adjustment procedures. *Journal of Climate*, **10**, 92–101.

Metcalfe J.R., Wilson R.A. and Goodison B.E. (1987) The use of acoustic ranging devices as snow depth sensors: an assessment. *Proceeding of the 44th Eastern Snow Conference*, Fredericton, June 3–4, 1987, pp. 203–207.

Mock C.J. and Birkeland K.W. (2000) Snow avalanche climatology of the western United States mountain ranges. *Bulletin of the American Meteorological Society*, **81**, 2367–2392.

MSC (2000) *Canadian Snow Data CD-ROM* CRYSYS Project, Climate Processes and Earth Observation Division, Meteorological Service of Canada: Downsview, Ontario, January 2000. (online copy available at `http://www.crysys.ca`).

Naaim M. and Gurer I. (1998) Two-phase numerical model of powder avalanche theory and application. *Natural Hazards*, **17**, 129–145.

NSIDC (1996) *Northern Hemisphere EASE-Grid Weekly Snow Cover and Sea Ice Extent Version 1*, National Snow and Ice Data Center: Boulder, CD-ROM.

Pielke R.A., Liston G.E. and Robock A. (2000) Insolation-weighted assessment of Northern Hemisphere snow-cover and sea-ice variability. *Geophysical Research Letters*, **27**, 3061–3064.

Pomeroy J.W. and Gray D.M. (1995) *Snowcover – Accumulation, Relocation and Management*, Report No. 7, National Hydrology Research Institute Science, Saskatoon, p. 144.

Pomeroy J.W., Gray D.M. and Landine P.G. (1993) The Prairie Blowing Snow Model: characteristics, validation, operation. *Journal of Hydrology*, **144**, 165–192.

Pomeroy J.W., Parviainen J., Hedstrom N. and Gray D.M. (1998) Coupled modelling of forest snow interception and sublimation. *Hydrological Processes*, **12**, 2317–2337.

Pomeroy J.W. and Schmidt R.A. (1993) The use of fractal geometry in modeling intercepted snow accumulation and sublimation. *Proceeding of the 50th Eastern Snow Conference*, Quebec City, June 8–10, 1993, pp. 1–10.

Rallison R.E. (1981) Automated system for collecting snow and related hydrological data in the mountains of the western United States. *Hydrological Sciences Bulletin*, **26**, 83–89.

Randall D.A., Cess R.D., Blanchet J.-P., Chalita S., Colman R., Dazlich D.A., Del Genio A.D., Keup E., Lacis A., Le Treut H., *et al.* 17 others (1994) Analysis of snow cover feedbacks in 14 general circulation models. *Journal of the Geophysical Research*, **99**(D10), 20757–20771.

Rango A. (1996) Spaceborne remote sensing for snow hydrology applications. *Hydrological Sciences*, **41**, 477–494.

Rango A., Walker A.E. and Goodison B.E. (2000) Snow and Ice. *Remote Sensing in Hydrology and Water Management* Schultz G.A. and Engman Edwin T. (Eds.), Springer, Berlin, pp. 231–270.

Robinson D.A. (1989) Evaluation of the collection, archiving and publication of daily snow data in the United States. *Physical Geography*, **12**, 120–130.

Robinson D.A. and Dewey K.F. (1990) Recent secular variations in the extent of Northern Hemisphere snow cover. *Geophysical Research Letters*, **17**, 1557–1560.

Robinson D.A. (1999) Northern Hemisphere snow extent during the satellite era. *Preprints 5th Conference on Polar Meteorology and Oceanography*, American Meteorological Society: Dallas, pp. 255–260.

Robock A., Mu M.Q., Vinnikov K. and Robinson D. (2003) Land surface conditions over Eurasia and Indian summer monsoon rainfall. *Journal of the Geophysical Research – Atmospheres*, **108**(D4), 4131.

Saito K. and Cohen J. (2003) The potential role of snow cover in forcing interannual variability of the major Northern Hemisphere mode. *Geophysical Research Letters*, **30**, 1302.

Schmidlin T.W. (1990) A critique of the climatic record of "Water equivalent of snow on the Ground" in the United States. *Journal of the Applied Meteorology*, **29**, 1136–1141.

Schutz, C. and Bregman, L.D. (1975) *Global Snow Depth Data: A Monthly Summary*, The Rand Corporation: Santa Monica, `http://www-nsidc.colorado.edu/data/g00788.html`.

Schweizer J. (1999) Review of dry snow slab avalanche release. *Cold Regions Science and Technology*, **30**, 43–57.

Seligman G. (1980) *Snow Structure and Ski Fields*, International Glaciological Society: Cambridge, p. 555.

Serreze M.C., Clark M.P., Armstrong R.L., McGinnis D.A. and Pulwarty R.S. (1999) Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. *Water Resources Research*, **7**, 2145–2160.

Serreze M.C., Clark M.P., McGinnis D.L. and Robinson D.A. (1998) Characteristics of snowfall over the eastern half of the United States and relationships with principal modes of low-frequency atmospheric variability. *Journal of Climate*, **11**, 234–250.

Sevruk B. (1992) *Snow cover measurements and areal assessment of precipitation and soil moisture*, Operational Hydrology Report No. 35, World Meteorological Organization, Geneva, p. 283.

Sevruk B. and Klemm S. (1989) *Catalogue of national standard precipitation gauges*, Instrument and observing methods Report No. 39 and TD No. 313, World Meteorological Organization, Geneva, p. 50.

Shook K. and Gray D.M. (1996) Small-scale structure of shallow snowcovers. *Hydrological Processes*, **10**, 1283–1292.

Stone R.S., Dutton E.G., Harris J.M. and Longenecker D. (2002) Earlier spring snowmelt in northern Alaska as an indicator of climate change. *Journal of the Geophysical Research – Atmospheres*, **107**(D10), 4089.

Sturm M., Holmgren J. and Liston G.E. (1995) A seasonal snow cover classification system for local to global applications. *Journal of Climate*, **8**, 1261–1283.

U.S. Army Corps of Engineers (1954) *Depth of Snow Cover in the Northern Hemisphere*, Arctic Construction and Frost Effects Laboratory, New England Division: Boston.

Vernekar A.D., Zhou J. and Shukla J. (1995) The effect of Eurasian snow cover on the Indian Monsoon. *Journal of Climate*, **8**, 248–266.

Vinnikov K.Y., Groisman P.Y. and Lugina K.M. (1990) Empirical data on contemporary global climate changes (temperature and precipitation). *Journal of Climate*, **3**, 662–677.

Walland D.J. and Simmonds I. (1996) Sub-grid-scale topography and the simulation of Northern Hemisphere snow cover. *International Journal of Climatology*, **16**, 961–982.

Walsh J.E., Kattsov V., Portis D. and Meleshko V. (1998) Arctic precipitation and evaporation: model results and observational estimates. *Journal of Climate*, **11**, 72–87.

WMO (1981) *Guide to Hydrological Practices*, *Fourth Edition*, Vol. 1, WMO-No. 168, World Meteorological Organization, Geneva.

Woo M.K. (1982) Snow hydrology of the high arctic. *Proceeding of the 39th Eastern Snow Conference, Joint Meeting of Eastern and Western Snow Conferences*, Reno, April 19–23, 1982, pp. 63–74.

Woo M.-K. and Steer P. (1986) Monte Carlo simulation of snow depth in a forest. *Water Resources Research*, **22**, 864–868.

Yang D., Goodison B.E., Metcalfe J.R., Golubev V.S., Bates R., Pangburn T. and Hanson C. (1998) Accuracy of NWS 8" standard nonrecording precipitation gauge: results and application of WMO intercomparison. *Journal of Atmospheric and Oceanic Technology*, **15**, 54–68.

Yang D., Goodison B.E., Metcalfe J.R., Louie P., Leravesley G., Emerson D., Hanson C.L., Golubev V.S., Elomaa E., Gunther T., *et al.* (1999) Quantification of precipitation measurement discontinuity induced by wind shields on national gauges. *Water Resources Research*, **35**, 491–508.

Zhang X., Vincent L.A., Hogg W.D. and Niitsoo A. (2000) Temperature and precipitation trends in Canada during the 20th Century. *Atmosphere-Ocean*, **38**, 395–429.

160: Energy Balance and Thermophysical Processes in Snowpacks

MICHAEL LEHNING

WSL, Swiss Federal Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland

This contribution discusses heat and mass fluxes relating to snow. The first process treated is heat flux through snow. Snow can be described as a bulk material or as consisting of different phases. When a separate water vapor phase is considered, vertical mass flux due to vapor pressure differences can be treated. A significant amount of heat is transported along with the vapor fluxes because of the phase changes occurring when water molecules enter the vapor phase at one point and deposit back onto the ice matrix somewhere else. The vapor fluxes in snow also cause snow metamorphism changing the crystals' form and size. Equilibrium metamorphism dominates when weak, large-scale temperature gradients exist, and water molecules are mainly rearranged locally by surface tension differences. Metamorphism is called kinetic when vertical vapor fluxes due to a large-scale temperature gradient lead to a snow crystal re-formation.

Mass- and energy fluxes in the snow cover are driven by surface exchange. The surface turbulent fluxes of sensible heat and moisture are derived from atmospheric surface layer similarity theory. The long-wave radiation balance leads to a strong surface cooling especially during cold nights. Short-wave radiation penetrates the snow cover and deposits energy at greater depths. Finally, the surface mass transport process of snow redistribution is treated with its subprocesses, saltation and suspension.

INTRODUCTION

On a yearly average, approximately 14% of the earth's surface is covered by snow (Brown, 2000 and references therein). The snow cover is thus an important part of the world's climate and weather system. The role of the snow cover is particularly important because of the strong snow – albedo positive feedback mechanism. Additionally, the seasonal snow cover is a major economic factor in tourism, transport, and water management. Therefore, the complex processes in the snow cover and at the snow – atmosphere interface need to be understood with enough detail so that we can assess and predict the role of snow in: (i) weather and climate development, (ii) natural hazard generation such as avalanches, floods, or mud flows, (iii) water resource management, and (iv) transportation maintenance and winter tourism.

Snow differs from other land surfaces in several aspects: First, snow has the highest albedo of all major land surfaces. This means that over snow a large part of the incoming solar radiation is reflected and not converted to heat. A positive feedback mechanism results because part of this reflected energy is subsequently lost to space, which leads to a cooling of the atmosphere and possibly to more snow. Of course, this feedback mechanism also means that if the average global snow coverage decreases, more short-wave radiation is absorbed from space, which is of great concern in recent years of rising temperatures. The high albedo of fresh natural snow makes it important to look at rapid changes of the albedo due to the presence of dust, sand, organic matter (such as lichen or snow algae), and metamorphism. A further particularity of snow is that it can be transported and redistributed by wind. The transport of snow changes the local mass and energy balance and may also alter the overall mass and energy balance over snow because the turbulent fluxes of heat and moisture are influenced. The turbulent fluxes are additionally influenced by the change of surface roughness associated with seasonal snow.

In this contribution, the relevant processes associated with the snow cover are treated. First, an overview on the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

processes and their interaction is given. Second, snow is introduced as a material and two complementary ways of snow description are treated. As a next step, the energy and mass exchanges at the snow surface are treated with an emphasis on the important role of the radiation balance and the turbulent fluxes. The energy and mass-transfer processes within the snow cover are treated and linked to snow metamorphism in the subsequent section. The final section is devoted to drifting snow.

OVERVIEW OF PROCESSES IN AND OVER SNOW

In the following, we concentrate on those snow processes that are of a thermophysical or mechanical nature. We do not treat snow chemistry or snow biology (*see* **Chapter 163, Hydrochemical Processes in Snow-covered Basins, Volume 4**). We briefly touch upon snow-water transport but redirect the interested reader to (*see* **Chapter 161, Water Flow Through Snow and Firn, Volume 4**). Figure 1 gives a schematic overview of snow processes.

Of particular importance are the mass and energy exchanges with the atmosphere. Snow cover dynamics are almost entirely driven by atmospheric forcing, except for a small heat flux contribution from below. First of all there is snow precipitation, which leads to the buildup of the snow cover. The buildup of the snow cover is very much dependent on ambient meteorological conditions: snow that falls at cold temperatures and low wind speeds differs significantly from snow that is deposited during a snow storm at high temperatures. Furthermore, initial grain shape is dependent on the cloud temperature and moisture conditions (Nakaya, 1954). During high wind conditions, snow is transported, creating – particularly in complex terrain – a very uneven snow distribution. A further important influence of precipitation comes from rain on an existing snow cover, which adds energy to the snow. If the snow is still at subfreezing temperatures, a large amount of energy is released by freezing the rain water. If the rain is warmer than the melting temperature, an additional but much smaller amount of energy is released by cooling the water to freezing temperature.

For snow-covered surfaces, radiation is the most important energy transfer process. Usually, two types of radiation transfer processes are distinguished: short-wave and long-wave radiation energy transfers. During clear nights (or during polar winter) the snow surface loses much energy by long-wave radiation. If the sun is shining, the snow cover gains energy through short-wave radiation that penetrates the snow to some depth. The penetration of short-wave radiation depends on the snow density, its impurity content, and the grain shape. It also depends on wavelength: the shorter the wavelength, the deeper the penetration. The penetration of radiation into the snow cover is important not only for the overall energy balance and thus for subsurface metamorphism but also for growth of organic material such as plants or snow algae and photochemical processes (Albert *et al.*, 2002).

In addition to the radiative fluxes, the turbulent fluxes of heat and moisture play an important role in the snow-surface energy balance. It is important to note that moisture transport is not only an energy-relevant process but also

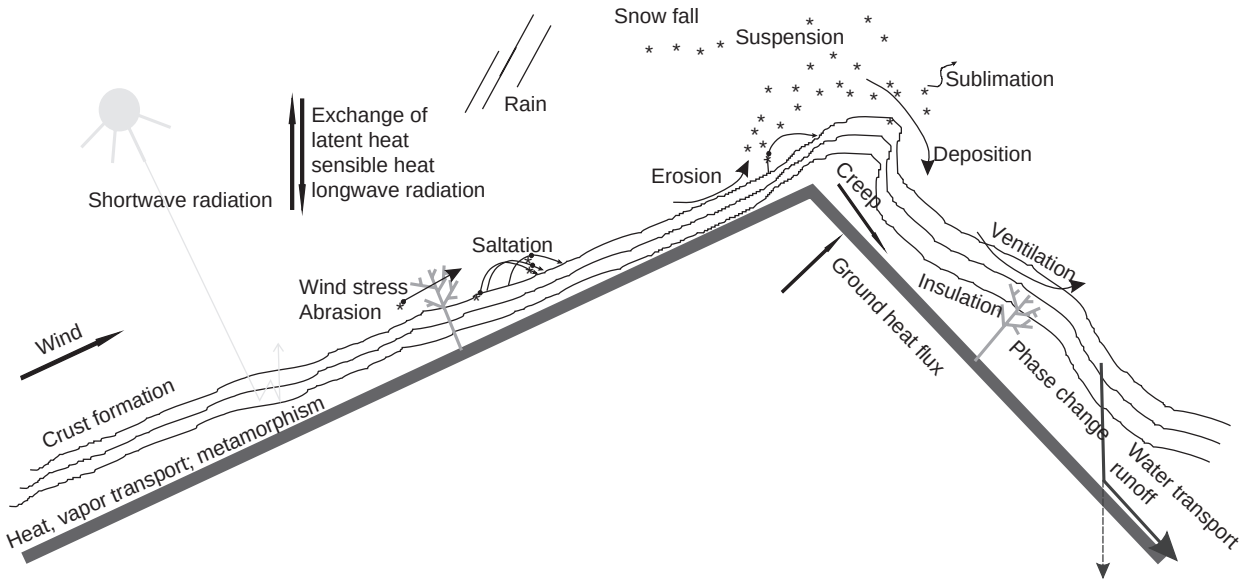


Figure 1 Schema of important physical processes in and over the snow cover. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

brings in or removes mass from the snow cover. In case of surface hoar, this mass frequently develops into a weak layer that may cause release of avalanches. In the case of warm winds such as Föhn, significant mass can be lost from the snow cover by sublimation. The question is still not answered as to what extent the turbulent moisture flux and therefore also the local mass balance is influenced by drifting snow. A further complication is the fact that the airflow (wind) penetrates the uppermost snow layers under certain conditions (ventilation), potentially altering the turbulent heat fluxes. While the process is not completely understood yet, its influence on snow metamorphism and energy exchange is significant.

Once the snow has settled on the ground, the crystals start to change shape and size immediately. A fast and important process is sintering, whereby the individual crystals or flakes form connecting bridges (bonds) to their neighbors. This process goes along with the so-called decomposing metamorphism, which tends to break the crystals down into smaller and rounder grains. The decomposing metamorphism dominates under conditions of weak temperature gradients in the snow cover. If there is no large-scale temperature gradient, water vapor is generated only by local differences in surface energy and as a result, only the edges of dendritic crystals disappear and round grains and bonds develop and grow. It is also interesting to note that the bonds between the grains will also grow in conjunction with snow settling, when the individual grains move closer together as the snow densifies. During the long time in a typical winter, the snow surface is cold but the insulating snow cover causes temperatures close to 0 °C at the snow – soil interface. This creates a strong temperature gradient in the snow, with the result that faceted crystals develop. This type of metamorphism is a real recrystallization and new crystals form. This recrystallization is the result of strong vapor fluxes that take place in the snow matrix (Akitaya, 1974). Because of the temperature gradient, vapor is generated at the upper part of existing snow grains, which have a relatively higher vapor pressure than the pore, diffuses upwards, and again undergoes a phase change to ice when it encounters an ice surface of lower vapor pressure. This recrystallization is only active in temperature gradient condition when vapor flows vertically over some distance larger then the grain size.

The macroscopic vapor fluxes also transport heat, because heat is consumed in the vapor generation process, while it is released when the vapor is deposited to form the new ice crystal. This is among the reasons why the heat conductivity of snow depends on snow microstructure: In snow with small pores and accordingly small grains (e.g. decomposed snow), the vapor flow is inhibited and can contribute less to heat transport. In snow of the same density but with larger pores and grains (e.g. depth hoar) much more heat can be transported by vapor. Therefore the bulk or macroscopic heat conductivity of snow is not a function of density only.

Phase change processes affect the development of the snow cover during its entire life cycle. When there is enough energy available, snow will start to melt and produce liquid water. Additional liquid water may come directly from rain. The water is retained up to a certain amount in the snow matrix and excess water moves downward under the influence of gravity. This behavior is similar to that of soil. However, the exact amount retained and the velocity of water movement in the snow matrix are difficult to predict because the water itself rapidly changes the snow matrix. First, snow grains develop into large round grains and form clusters. Second, preferential flow paths develop, where water moves rapidly downward, independent of a mean matrix potential controlling water movement in the remaining pore space. Water movement in snow governs water release for runoff and soil and is therefore an important subject, for example, for runoff prediction. It is treated in detail in (*see* **Chapter 161, Water Flow Through Snow and Firn, Volume 4**).

SNOW DESCRIPTION

Snow is a mixture of primarily water and air. Because it is – at environmental conditions – at a high homologous temperature, the water phase of snow undergoes rapid phase changes. In snow, at least two phases of water are present at all times: solid water (ice) and water vapor. Very often, water is present also in its liquid form. In addition to water and air, many other substances usually exist in natural snow. Primarily, organic matter and inorganic dust are present. Important from an environmental point of view are chemicals such as nitrate or other nutrients. Because of its high surface area, snow is also a good chemical reactor, in particular, since a strong exchange of trace gases between the snow and the atmosphere exists. The multiphase material snow can be described in a variety of ways. Traditionally, snow is described as a homogeneous material with its bulk properties given as density, effective thermal conductivity, viscosity, or elasticity. The microstructure of snow is neglected, as well as its composition of different phases. A simple bulk approach is, for example, to parameterize the other snow properties as a function of bulk snow density. However, it must be mentioned that, for example, effective thermal conductivity can vary up to an order of magnitude for different snow of identical density (Sturm *et al.*, 1997). Alternatively to this simple bulk approach, snow, or any other porous material (Hassanizadeh and Gray, 1990) can be treated as a mixture. The conservation laws of mass, momentum, and energy are applied to the mixture. This

results in a complex mathematical treatment of snow (e.g. Brown *et al.*, 1999; Morland *et al.*, 1990).

Mixture theory may be formulated to account for snow microstructure. The bulk properties such as thermal conductivity and viscosity can now be parameterized as a function of snow microstructure. As an example, the SNOWPACK model (Bartelt and Lehning, 2002) uses this approach. In SNOWPACK, snow is modeled as a three-component mixture:

$$\theta_i + \theta_a + \theta_w = 1. \tag{1}$$

The volumetric fractions of ice, θ_i, moist air, θ_a, and liquid water, θ_w, are required to always add up to 1 for all settling, phase change, or mass transport processes considered. Other common bulk snow properties can then be defined as a function of the volumetric fractions, for example, snow density, ρ:

$$\rho = \theta_i \rho_i + \theta_a \rho_a + \theta_w \rho_w. \tag{2}$$

Here, the material densities of ice, ρ_i, liquid water, ρ_w, and air, ρ_a, ($\mathrm{kg\,m^{-3}}$) are also used. The use of a bulk approach basically means that the complex geometry of the phase distributions in snow is collapsed to a simple volumetric description such as sketched in Figure 2b.

In order to improve the snow description, it is possible to parameterize the effect of snow microstructure. Snow properties, which depend on snow microstructure, can be treated in the form of microstructure parameters of grain size, bond size, dendricity, and sphericity. These four independent parameters have their own rate equations describing how they develop as a function of environmental conditions. More detail is given in the section on "Snow metamorphism" in the following text.

ENERGY AND MASS TRANSFER IN SNOW

Energy and mass exchanges in snow are governed by heat conduction in the ice and pore matrices, vapor transport, short-wave radiation absorption, water transport, and phase changes. Phase changes can take place between water and ice, water and vapor, and ice and vapor. They are highly relevant in terms of both energy and mass transport.

Heat Conduction

The energy conservation law applied to a snow cover in its bulk form gives the balance between the snow temperature change and the energy fluxes entering or leaving the snow cover. In a very simplified form, where the effect of water vapor transport is lumped into the effective conductivity, k_e, ($\mathrm{J\,m^{-1}\,K^{-1}}$) (Jordan, 1991) or into the phase change source or sink term, Q_{pc}, ($\mathrm{J\,m^{-3}}$) the equation in its one-dimensional form reads:

$$\rho_s c_p \frac{\partial T}{\partial t} + k_e \frac{\partial^2 T}{\partial z^2} = Q_{sw} + Q_{pc} \tag{3}$$

Here, ρ_s is the bulk density of snow and $c_p(\mathrm{J\,kg^{-1}K^{-1}})$ the bulk heat capacity at constant pressure. T is the bulk snow temperature, z stands for the vertical (surface perpendicular) coordinate and lateral heat fluxes are neglected. Q_{sw} is the short-wave radiation balance entering the snow cover and Q_{pc} contains the phase change contributions from both water vapor transport and melt-freeze phase changes. A form of equation (3) is common to the three most detailed snow models currently in use, namely, CROCUS (Brun *et al.*, 1989), SNTHERM (Jordan, 1991) and SNOWPACK (Bartelt and Lehning, 2002).

Transport of Water Vapor

Since water vapor transports significant amounts of heat through phase changes at the ice surface and drives snow metamorphism, it is of advantage to treat it explicitly. Under the assumption of Fickian diffusion of water vapor in the pore space, mass continuity leads to the following partial differential equation:

$$\theta_a \frac{\partial \rho_v}{\partial t} - \frac{\theta_a D_v}{\tau} \frac{\partial^2 \rho_v}{\partial z^2} = \frac{Q_{pc}^{s/v}}{L^{s/v}} \tag{4}$$

Here, $\rho_v(\mathrm{kg\,m^{-3}})$ is water vapor density in the pore space, D_v is the water vapor diffusivity in air ($2.2\ 10^{-5}$ $\mathrm{m^2\ s^{-1}}$), which has a small temperature and pressure dependence (Massman, 1998), and τ is the snow tortuosity, which is of the order of 2. Q_{pc} and L at the right-hand side of equation (4) give the source term through phase change and the latent heat, respectively. The superscript $^{s/v}$stands for sublimation or vaporization, depending on whether the phase change is between vapor and liquid or vapor and solid water. Note that the latent heat values are different ($L^s = 2.83\ 10^6 \mathrm{J\,kg^{-1}}$; $L^v = 2.50\ 10^6 \mathrm{J\,kg^{-1}}$) for both processes and further have a small dependence on temperature and pressure.

Discussion

Equations (3) and (4) are a coupled set of partial differential equations, which can be solved numerically to describe vertical mass and energy transport in a phase changing snow cover. If both equations are solved, the vapor-ice or vapor-water phase change is the coupling term. In SNOWPACK, only equation (3) is solved, and vapor transport is parameterized by using an appropriate formulation for the effective thermal conductivity, k_e

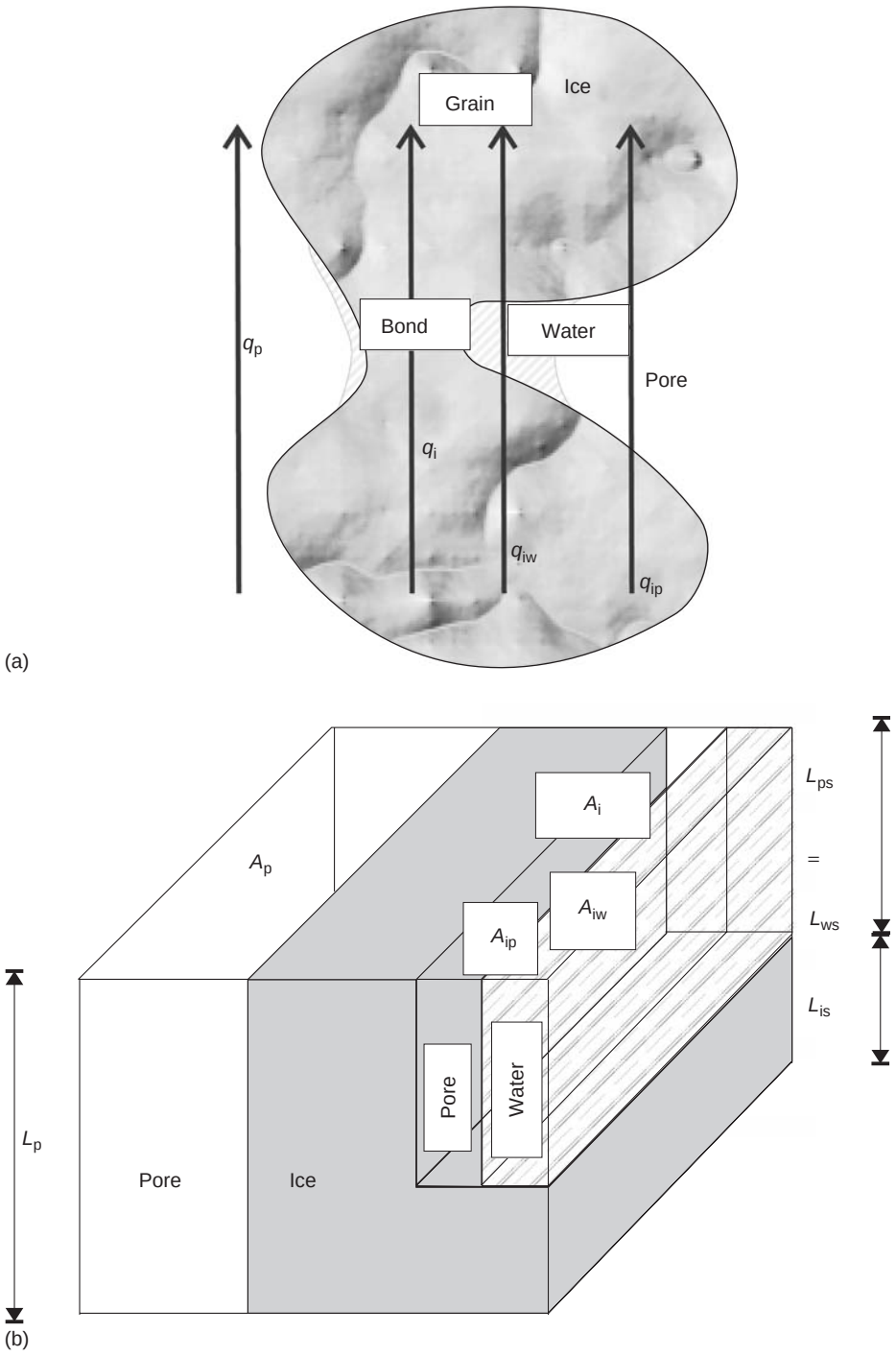


Figure 2 (a)Definition of ice, water, and air phases in snow and possible heat flux situations; (b)schematic representation of the phases in a bulk model

(Lehning *et al.*, 2002a; Fierz and Lehning, 2001). More recently, advanced versions of SNOWPACK (Bartelt *et al.*, 2004) have been developed. They have separate temperature equations for the air and ice phases and solve the vapor transport equation explicitly. However, the separate equations for the air and ice phases now contain heat exchange terms, which are used as fitting parameters in Bartelt *et al.* (2004). In addition, this separate treatment needs separate boundary conditions, which are not available from routine measurements. In fact, they cannot be measured or predicted with current techniques at all.

Not all transport processes in the snow cover are diffusive at all times. In addition to water flow in snow (*see* **Chapter 161, Water Flow Through Snow and Firn, Volume 4**), evidence also exists for convective air movements under certain circumstances. Those air movements, which are driven by atmospheric pressure fluctuations from above, are commonly called *ventilation* (Albert and McGilvary, 1992). If they are driven by buoyancy (which requires lateral differences), they are rather called *convective movements* (Sturm and Johnson, 1991). A correct description of these effects requires, in any case, that snow is treated in at least two dimensions. Therefore, equations (3) and (4) do not contain an airflow velocity. However, it is possible to include the effect of ventilation or convection even in a one-dimensional model description. This can be done by parameterizing the effect on the diffusivities of heat, k_e, or water vapor, D_v. An example of such a parameterization is given by Lehning *et al.* (2002b), while Bartelt *et al.* (2004) allow explicitly for a vertical velocity in their equations, which are driven by arbitrary pressure fluctuations.

Snow Settling

A final advective mass transport occurring in snow is snow settling. Fresh snow settles very fast initially, but settlement slows down very quickly. Snow is sometimes treated as a visco-elastic material that undergoes large irreversible deformations. For one dimension, a simple rheological Maxwell law relates the viscous strain rate, $d\varepsilon_v/dt$, to the mean stress in the snow, σ_s, where η_s is the snow viscosity.

$$\frac{d\varepsilon_v}{dt} = \frac{-\sigma_s}{\eta_s} \tag{5}$$

This law is applicable when elastic deformations are small, as is the case for snow settling under its own weight. As shown by Mellor (1975), the viscosity of snow can vary by several orders of magnitude over the density range of natural snow covers. Since snow viscosity is highly dependent on microstructure, it is very challenging to find good formulations of snow viscosity. An attempt to formulate snow viscosity as a function of snow microstructure by considering processes such as linear and nonlinear flow laws for ice and pressure sintering has been described by Lehning *et al.* (2002a). However, more work is needed to reduce the need for empirical adjustments.

SNOW METAMORPHISM

The primary driving mechanism behind dry snow metamorphism is the phase change associated with water vapor flux. Figure 3 shows a simplified scheme of snow metamorphism. The process starts immediately after snow has been deposited on the ground. The dendritic crystals have a high surface energy and vapor is created at the edges and deposited in the crystal cavities. In this way, a first snow texture consisting of grains and bonds is created. Further development strongly depends on the temperature conditions. Traditionally, kinetic and equilibrium metamorphism

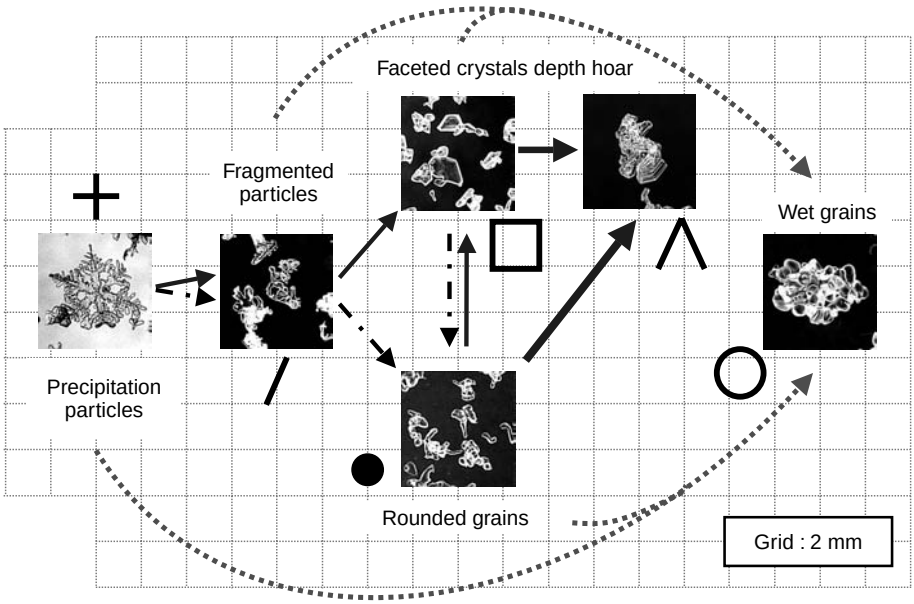


Figure 3 Schema of snow metamorphism; the snow grain types are shown as photographs and as the standard international symbols. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

are distinguished. They are treated in more detail in the following text. Another process is melt metamorphism, when large and round grains develop, which may also form clusters (Colbeck, 1997). They develop under the influence of liquid water.

Equilibrium Metamorphism

Equilibrium metamorphism takes place when no "global" vapor flux exists, that is, vapor molecules are relocated, but the source and sink grain may be the same. The driving force behind equilibrium metamorphism is differences in vapor pressure due to convex and concave surfaces as described by Kelvin's equation (Colbeck, 1980):

$$p_s = p_s^0 e^{\frac{2\sigma}{RT_i\rho_i r_s}} \tag{6}$$

Here, p_s is the saturation vapor pressure (Pa) over the curved ice surface at the temperature T_i (K), $p_s{}^0$ is the saturation pressure (Pa) over a flat ice surface, which can be obtained from an empirical form of the integrated Clausius-Clapeyron equation, R is the water vapor gas constant ($461\,\mathrm{J\,kg^{-1}K^{-1}}$), σ is the surface tension of ice ($\mathrm{N\,m^{-1}}$), and r_s is a mean ice surface radius of curvature (m). In order to apply equation (6) to describe snow metamorphism, the snow pore and ice geometry must be known and described to a high degree of detail. Therefore, it is still common to use semiempirical laws (Brown *et al.*, 2001) to describe equilibrium snow metamorphism. Snow microstructure development is modeled using rate equations for structure parameters. For example, the development of grain and bond size is most important for equilibrium metamorphism in the snow cover model SNOWPACK (Lehning *et al.*, 2002a). The bonds between the snow grains grow by vapor deposition and to a lesser degree by pressure sintering, which is related to snow settling. The end form of equilibrium metamorphism is fine grained snow with rather strong bonds, which is mechanically stable.

Kinetic Metamorphism

Kinetic metamorphism starts when macroscopic temperature gradients (i.e. temperature gradients over distances larger than the grain size) lead to vapor pressure differences that are not only local. If the pore space is sufficient, water vapor will then move. Small grains may disappear completely at certain locations in the ice matrix (source grains), while at other locations, where vapor gets deposited, grains develop into large faceted crystals. Kinetic metamorphism is therefore sometimes called *snow recrystallization*. Because of the high growth rate, these crystals develop as faceted hexagonal crystals (Petrenko and Whitworth, 1999). Depending on temperature conditions, prismlike structures, plates, or cups develop. If enough pore space is available, the end form of kinetic metamorphism, under high temperature gradient conditions, is cup-shaped crystals, also called *depth hoar*.

Model descriptions of kinetic metamorphism are predominantly empirical (Marbouty, 1980) or semiempirical (Lehning *et al.*, 2002a). In SNOWPACK, rate equations are developed for grain growth, bond growth, and change in sphericity. A classification routine is then used to recover conventional grain types. Figure 4 shows an example of grain type development as calculated by SNOWPACK for the infamous avalanche winter of 1999 at the study plot site Weissfluhjoch Versuchsfeld.

Discussion

A correct model of snow metamorphism is crucial for a correct snow description because the snow thermal and mechanical properties strongly depend on the snow microstructure. Therefore, a more physical model description of snow metamorphism is desirable. Miller *et al.* (2003), have presented a quasi two-dimensional model of snow metamorphism, based on the physical principles of vapor and heat transport as discussed previously. Introducing growth laws for faceted crystals, they can successfully describe equilibrium and kinetic metamorphism with their model. They can also predict growth dependency on available pore space, temperatures, and temperature gradients.

SURFACE ENERGY AND MASS EXCHANGE

The energy conservation law (equation 3) must be solved numerically with a correct set of boundary conditions. For all cases, Neumann boundary conditions can be employed for the upper boundary:

$$k_e \frac{\partial T}{\partial z} = q_{sh} + q_{lh} + q_{lw} + q_{rr} \tag{7}$$

Here, the four surface energy flux terms, sensible heat, q_{sh}, latent heat, q_{lh}, net long-wave radiation (incoming minus outgoing), q_{lw}, and energy through liquid precipitation, q_{rr}, ($\mathrm{W\,m^{-2}}$) determine the heat flux through the snow surface. They will be treated in more detail in the following text. While the surface of the snow cover is a fractal and therefore physically not precisely defined, latent and sensible heat fluxes as well as the net long-wave radiation balance are usually treated as surface energy sources. Note that implicitly the control surface contains no mass or volume, so equation (7) has no storage term. Short-wave radiation penetrates the snow and is therefore considered to be a volume source (equation 3). Short-wave radiation penetration is also treated in the following text.

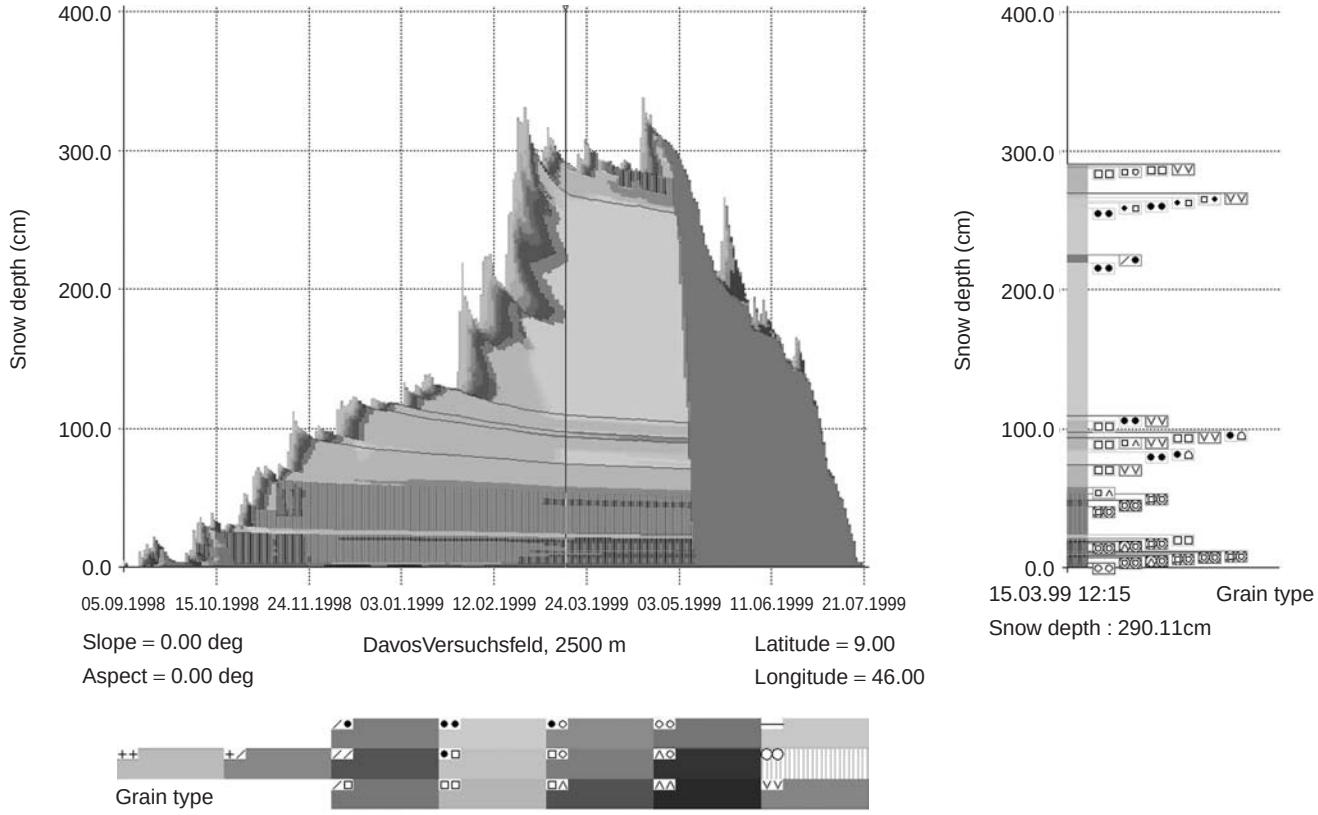


Figure 4 Development of snow grain types at Klosters. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Instead of Neumann boundary conditions, Dirichlet boundary conditions can also be used. This means measuring the temperature at the snow surface. This is a very good way to solve equation (3), as long as no phase change processes are taking place, and the sum of the surface energy fluxes is uniquely determined by the surface temperature. For all practical applications, one is allowed to neglect sublimation phase changes and work with a Dirichlet boundary condition as long as no melting (or refreezing) occurs. As soon as melt processes are involved, the surface temperature does not determine the surface energy exchange any more. The lower boundary condition can be formulated in the same way as the upper ones. Note that then the soil heat flux replaces the sum of the surface heat fluxes.

For a numerical solution of equation (4), boundary conditions must be formulated as well. The latent heat flux treated in the following discussion serves as an upper boundary condition. For all practical applications, a zero-flux condition can be used at the bottom.

Sensible Heat Exchange

From similarity theory for the atmospheric surface layer (e.g. Stull, 1988), the gradient of the scalar quantity air temperature, T_a, takes the form (Note here that since we are only concerned with minor changes in altitude within the surface layer, we do not have to use potential temperature instead of temperature.)

$$\frac{\partial T_a}{\partial z} = \frac{-0.74}{k\ z\ u_f} \frac{q_{sh}(z)}{\rho_a\ c_p} \tag{8}$$

Here k is the Karman constant (0.4) and u_f the friction velocity. In the surface layer (constant flux layer), the fluxes at any height are approximately equal to the surface flux, $q_{sh}(z) = q_{sh}(0)$.

The empirical constant 0.74 comes from data in the neutral atmospheric boundary layer and is related to the fact that the turbulent exchange coefficients for heat and momentum differ (e.g. Stull, 1988). The surface is not entirely smooth, and the roughness length, z_0, is a parameter describing the aerodynamic surface roughness. It is also known that the roughness length is always smaller than the height of the individual roughness elements. Thus, the measured surface temperature (of snow) is postulated to be equal to the air temperature at $z_0 : T_a(z_0) = T(0)$. We can now integrate equation (8) from z_0 to the height of the

measured air temperature to receive:

$$T_a(z) - T(0) = \frac{-0.74}{k\ u_f} \frac{q_{sh}(z)}{\rho_a\ c_p} \ln \frac{z}{z_0} \tag{9}$$

and the surface sensible heat flux, $q_{sh}(0)$ can be obtained from the measurement of wind speed and temperature at any height, z, and the surface temperature, $T(0)$:

$$q_{sh}(0) = \frac{-k\ u_f}{0.74 \ln \frac{z}{z_0}} \rho_a\ c_p (T_a(z) - T(0)) \tag{10}$$

It is easily seen that equation (10) is analogous to the form of the usual bulk transfer equation for surface fluxes and suggests an expected value for the kinematic transfer coefficient, C ($m\,s^{-1}$):

$$C = \frac{k\ u_f}{0.74 \ln \frac{z}{z_0}} \tag{11}$$

Latent Heat Exchange

The development for the latent heat flux, q_{lh}, is identical to the one for sensible heat and we derive using equation (11):

$$q_{lh} = -C \frac{0.622 L^{s/v}}{R_a T(z)} \left[e_s^w (T_a(z))\, rH - e_s^i (T(0)) \right] \tag{12}$$

R_a ($287\,J\,kg^{-1}\,K^{-1}$) is the gas constant for air, $e_s^{w/i}$ is the saturation vapor pressure (Pa) over water or ice respectively, and rH (1) is the relative humidity. The constant 0.622 is the ratio of the gas constant for dry air over that for water vapor. At the snow surface we use the saturated vapor pressure over ice. The saturation vapor pressure is approximated by common semiempirical integrations of the Clausius-Clapeyron equation (e.g. Magnus formula).

Discussion

The above formulations are very robust descriptions of surface turbulent flux exchange over snow. They should be used if measured meteorological variables are available at one height only and the snow surface temperature is known or can be estimated. When meteorological variables are only available at one height above the surface, the roughness length of the site must be estimated *a priori*. Note that usually the roughness length is not determined by the snow grain size but is a function of surface structures such as sastrugi, topography, or buildings. For a small flat field surrounded by mountainous terrain, a typical value for the roughness length is 7 mm. With the known roughness length, the friction velocity can be calculated from the logarithmic wind profile, u:

$$u(z) = \frac{u_f}{k} \ln \frac{z}{z_0} \tag{13}$$

The development above is stable and simple. However, especially over flat terrain, atmospheric stability effects may become important. The above development can be done in an identical way when stability correction functions are required (Essery, 1997). Another simplification is the assumption implicitly invoked that roughness lengths are equal for momentum, heat, and moisture fluxes. Measurements show that this is not the case, in particular, for snow surfaces (Calanca, 2001). Andreas (1987) describes a theory to include different roughness lengths for heat and moisture into the flux formulations.

Long-wave Radiation Balance:

The term, q_{lw}, in equation (7) is the net surface long-wave radiation balance source term, that is, long-wave radiation received from above at the snow surface minus the long-wave radiation emitted by the snow surface. Long-wave radiation is emitted from each body as described by the Stefan-Boltzmann law. The law can be applied directly to the snow surface, to determine the energy emitted from the snow surface. Snow has an emissivity (and absorptivity) very close to one. The long-wave radiation received from the sky is more difficult to determine, because dry air has a low emissivity/absorptivity and the snow surface sees a mix of radiation emitted from the air close to the surface, from clouds above (when there are any) and from space. Therefore, for a correct description of the surface energy balance, an accurate measurement of incoming long-wave radiation is advantageous. Alternatively, if the cloud cover, the air temperature, and the relative humidity are known, it is also possible to parameterize the incoming long-wave radiation quite accurately (Konzelmann *et al.*, 1994).

Rain Energy

The energy added to the snow cover by (warm) rain, q_{rr}, is treated as a flux boundary condition. Assuming that rain has the same temperature as the air, the flux of energy due to rain is given by:

$$q_{rr} = r_{rr} c_p^w (T_a(z) - T(0)). \tag{14}$$

Here r_{rr}($kg\,s^{-1}$) is the rain rate and $c_p{}^w$ the specific heat of water ($4190\,J\,kg^{-1}K^{-1}$). The rain mass is added to the uppermost element of the snow cover and the water will then undergo phase changes or will be transported downwards. Note that in the case of snow temperatures below 0 °C most of the effect of rain on snow will be due to the phase change energy when the water freezes.

Short-wave Radiation Balance

A photon that hits a snow crystal has a very low probability to be absorbed. A high fraction of the incoming direct and indirect short-wave radiation hitting the snow surface will be scattered back from crystals near the snow surface and become a part of the upwelling diffuse short-wave radiation. This is the reason for the high albedo of snow. This also means that a good radiative transfer description in snow will allow prediction of the snow surface albedo. On the basis of measurements made in Greenland, (Meirold and Lehning, 2004) such a transfer model has been developed, which takes into account snow density, snow grain size and snow grain form. Figure 5 shows spectrally resolved measurements of radiation penetration in snow. With a model based on ray tracing (Macke, 1993) and the delta Eddington approximation a good description of this transfer is possible (Figure 6). It is important to take the impurity content of snow and the influence of different grain shapes into account. Because snow crystals have a high refractive index, small amounts of impurities can lead to a significantly enhanced absorption.

For determining the volume source term, Q_{sw}, in equation (3), a simple extinction formulation with a density-dependent extinction coefficient has proven to satisfactorily calculate the energy penetration, provided that the albedo is known (Lehning *et al.*, 2002b). However, future snow models will certainly make use of the more complex techniques described above for two main reasons. First, the complex technique will provide an estimate for snow albedo and second, more and more applications such as snow chemistry will require that a spectrally resolved radiation flux in snow is known.

SNOW TRANSPORT AND THREE-DIMENSIONAL EFFECTS

Snow transport by wind is a spectacular and important phenomenon (Figure 7). It crucially influences the seasonal buildup of the snow cover and avalanche activity in Alpine terrain. It also influences the growth of vegetation and the storage of water and pollutants. Snow transport has been studied extensively over the last few decades, and much progress has been made in modeling and understanding of

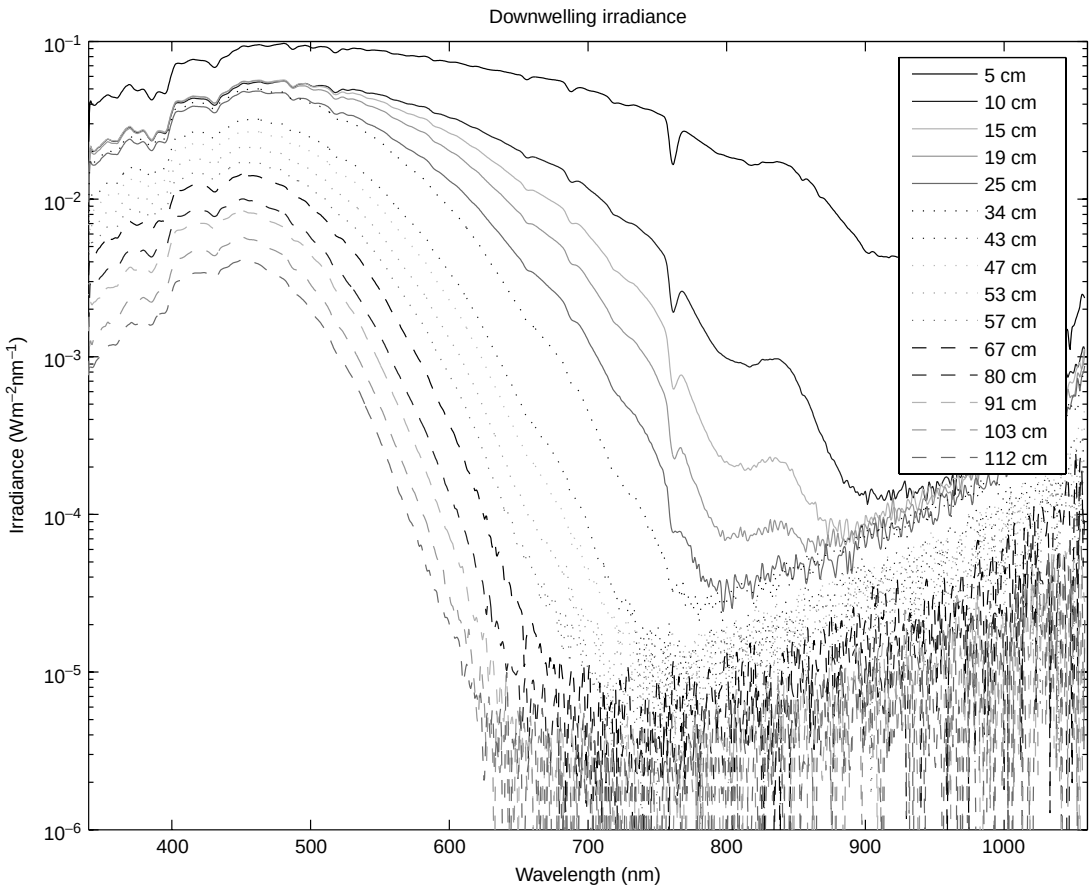


Figure 5 Spectrally resolved measurements of radiation transfer as a function of depth in snow at Summit, Greenland. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

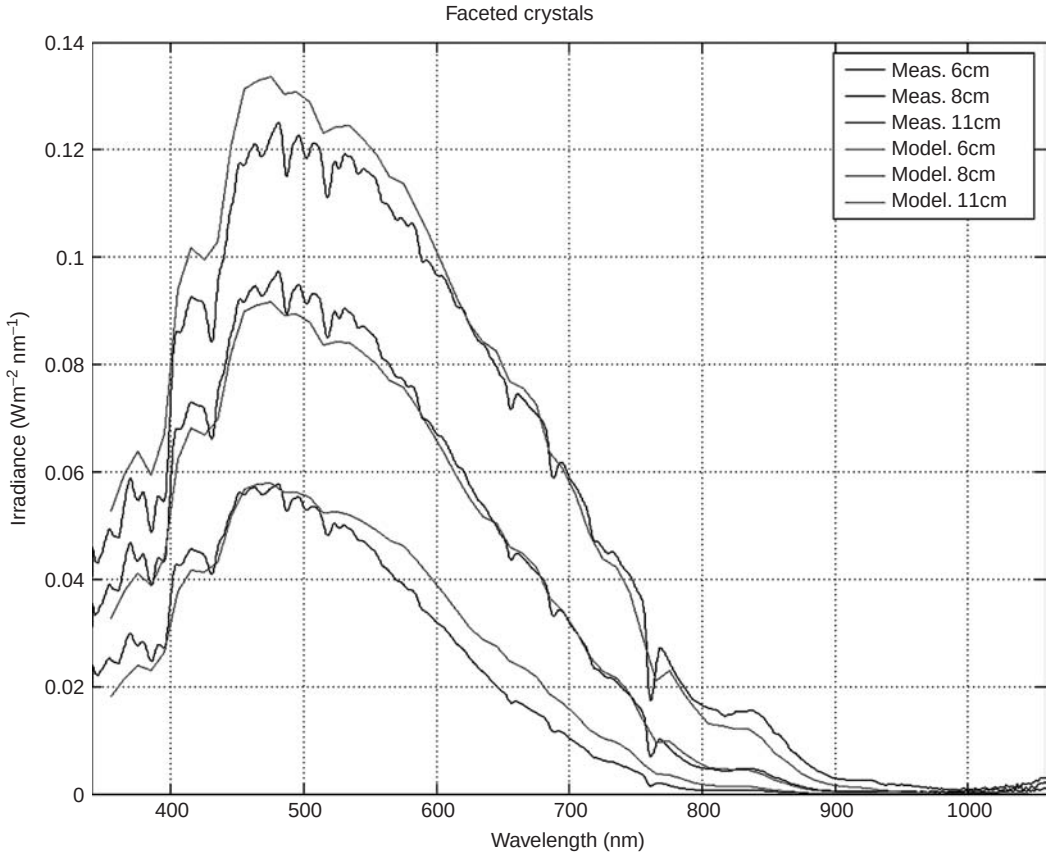


Figure 6 Comparison of measured and modeled radiation fluxes in snow. A good agreement is reached by accounting for snow impurities, snow-grain size, density, and snow-grain shape using a ray tracing technique. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

snow drift. Despite these efforts, due to the complexity of the physical processes involved, complete descriptions of snow redistribution by the wind hardly exist.

Traditionally, snow drift research has been motivated by engineering efforts to mitigate the adverse effects of blowing and drifting snow or to assess the mass and energy balance of snow and ice surfaces in high latitudes (Andreas, 1995; Mann *et al.*, 2000; Gallée *et al.*, 2001). Recently, modeling efforts have been intensified, in an attempt to describe the effects of drifting snow on the atmospheric boundary layer (Bintanja, 2000; Déry *et al.*, 2001). Snowdrift is also a notorious problem for avalanche warning (e.g. Gauer, 2001), because certain slopes may be loaded by huge amounts of additional snow and therefore, redistribution changes snow characteristics in a way that leads to a higher probability of avalanche release.

Snow Saltation

We usually distinguish between two major transport modes. The first is snow saltation (drifting snow), where grains hop over the snow surface following ballistic trajectories. The second transport mode is snow suspension (blowing snow), when grains become suspended by the airflow and follow the streamlines of the flow. In principle, an additional transport process, namely creep or reptation exists, where grains roll along the snow surface. This process only transports minor amounts of snow, however.

Snow saltation is believed to be the dominant transport mode in flat terrain. By definition, saltation can only take place over an existing snow cover, when wind speeds are high enough to lift grains from the ground. The exact mechanism of saltation initiation is not completely clarified. However, newer results suggest that a certain ratio of aerodynamic entrainment to rebound or splash dominates the saltation cloud (Doorschot and Lehning, 2002). Aerodynamic entrainment is the process by which grains are lifted because of the shear stress of the airflow, while rebound or splash is the process whereby an already saltating grain interacts with the surface and rebounds or ejects other snow grains. In any case, the initial velocity of the grain and the "ejection" angle determine to a significant degree the track of the snow grain. The flow conditions determine whether the energy gain of the particle during its jump is sufficient to counterbalance the energy loss at

Figure 7 Suspension cloud over a mountain ridge. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the next impact with the surface and therefore whether saltation will continue. Obviously, the energy lost with each impact is also a strong function of the material properties of the grain and the surface. This is the reason why snow saltation is more complicated than sand saltation, in which the elasticity of the grains is much higher and no strong adhesive forces act between the grains on the ground. The state-of-the-art model of snow saltation has been presented by Nemoto *et al.* (2004), where individual trajectories are calculated to predict an overall snow mass flux. The model includes also the transition from saltation to suspension. However for an application-oriented use, those models are computationally too expensive. Therefore, various empirical formulations for saltation transport rates remain important (e.g. Pomeroy and Gray, 1990). Most of them have in common the fact that the snow transport rate, Q ($\mathrm{kg\,m^{-1}\ s^{-1}}$) is assumed to be proportional to the flow mean shear stress, u_f, or the flow velocity, u raised to some power, x, which is often exactly or approximately 3. A good model of this type is that of Sorensen (1991), which also takes into account the dependence on the threshold for snow saltation initiation. It must be emphasized, however, that none of these simple formulations is valid over sloping terrain, where the effects of drifting snow are most important. In fact, to our knowledge no data set exists which describes snow saltation on slopes. In the following, we present results from a 3-D snowdrift model (Drift@ALPINE3D), which uses an equilibrium saltation model in the form described by Doorschot *et al.* (2004). It should, in principle, also work for steep slopes although it has not yet been validated for that case. The equilibrium model is simpler than the fully dynamic model of Nemoto *et al.* (2004) but is still based on physical principles and is computationally much less expensive.

Snow Suspension

Snow in suspension can originate from the snow cover or directly from precipitation. When wind speeds increase, saltating particles will be picked up at some point by turbulent eddies and transported away over longer distances. Frequently, suspension clouds can be observed over steep mountain ridges (Figure 7). Over flat terrain, on the other hand, suspension is less important and is most commonly described assuming an equilibrium concentration profile coupled to some concentration derived from the saltation model (e.g. Liston and Sturm, 1998):

$$c(z) = c(z_{\mathrm{ref}})\frac{z}{z_{\mathrm{ref}}}^{\frac{-s}{ku_{\mathrm{f}}}} \tag{15}$$

Here, c is the concentration (volume fraction), z_{ref} is a reference height, s is the settling velocity of snow flakes in still air ($\mathrm{m\,s^{-1}}$), k is the Karman constant (0.4), and u_{f} is the friction velocity ($\mathrm{m\,s^{-1}}$). This concentration profile is based on the assumption that the snow settling velocity

is counterbalanced by turbulent diffusion. The suspension mass balance can then be obtained by integrating the concentration profile multiplied by the mean wind-speed profile with height. More interesting is the case for complex terrain, where suspension is often described with a common advection–diffusion equation:

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} + (w - w_s)\frac{\partial c}{\partial z} - K\frac{\partial^2 c}{\partial z^2} = 0 \qquad (16)$$

Here w_s is the settling velocity in still air and K is a diffusion coefficient ($m^2 s^{-1}$). While the applicability of equation (16) is given, the numerical implementation is very difficult because the transition to the saltation layer below cannot be resolved with current numerical meshes with a reasonable computational effort. For the results of Figure 8, we use a numerical solution of equation (16) but parameterize the effect of atmospheric turbulence into a modified settling velocity and set K to zero. Furthermore, we use equation 15 as a wall function to bridge the distance between the reference height determined from the saltation and the lowest grid level of the atmospheric model, which is also the lowest grid level for the solution of the suspension equation. The mass balance for the saltation layer includes then the contribution from suspension and gives as a result the local erosion or deposition values.

Figure 8 shows the final snow distribution for the Gaudergrat ridge after the major drift period of the avalanche winter 1999. Simulated is a drift period of 5 days. The wind fields have been calculated with the ARPS mesoscale

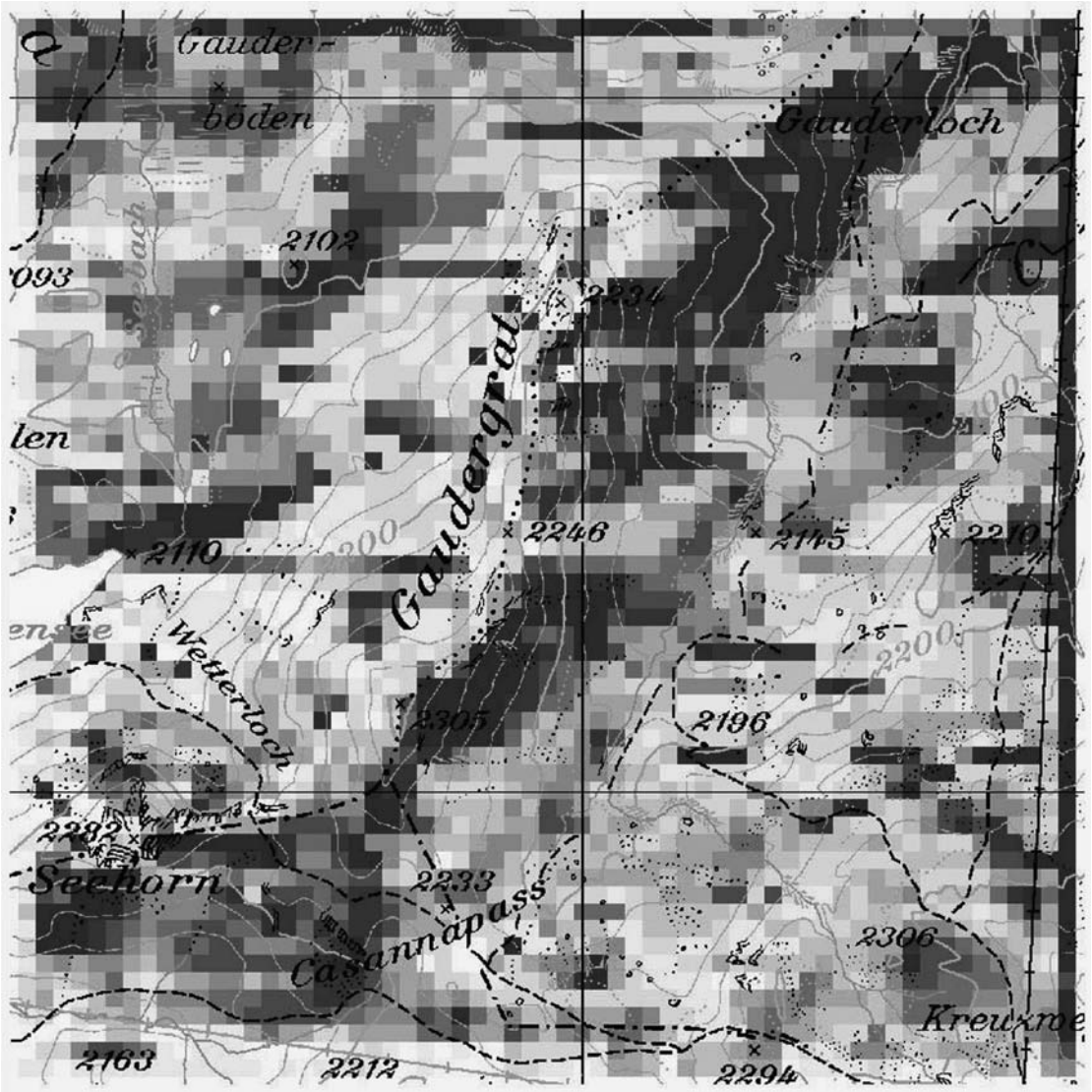


Figure 8 Snow distribution over the Gaudergrat ridge after the five-day storm period of the avalanche in the winter of 1999. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

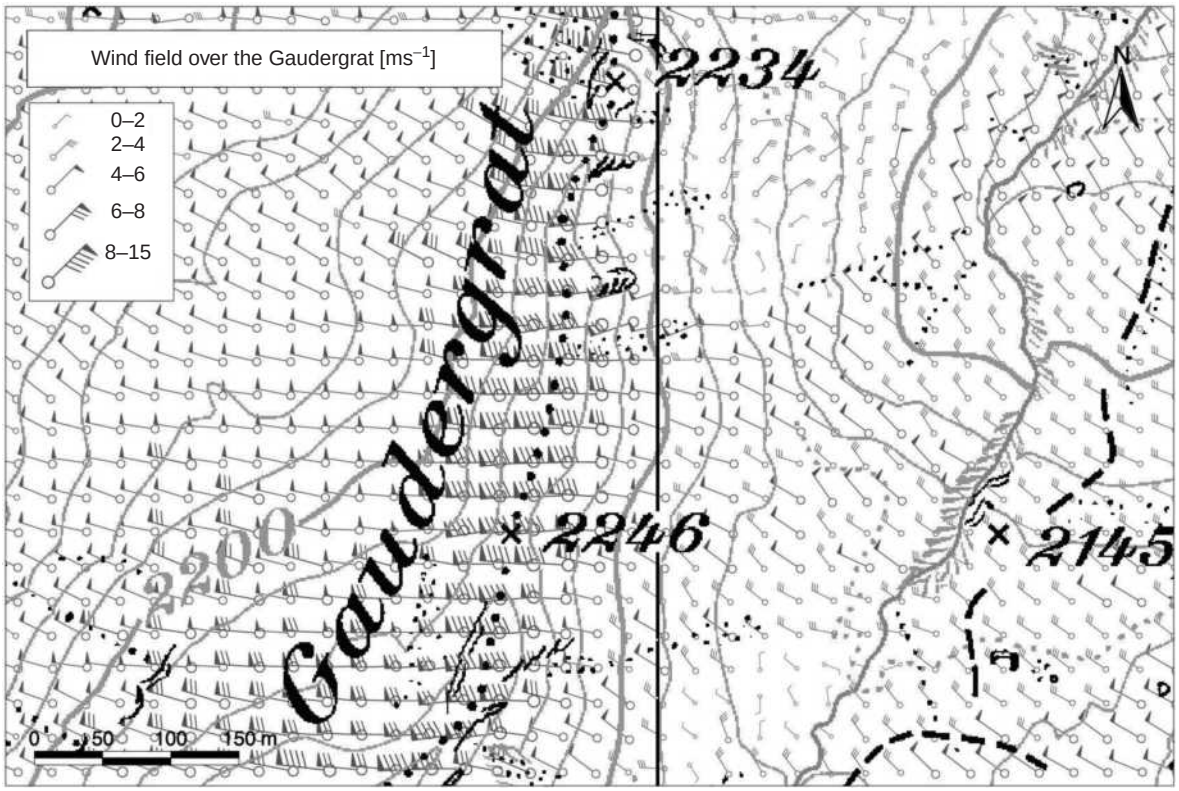


Figure 9 Sample wind field used to calculate the snow distribution of Figure 8. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

atmospheric model. One example of such a wind field is Figure 9. It is clear that zones of very low wind speed but high turbulence are present in the lee of the ridge. These are the zones of maximum snow deposition. The grid size for wind and transport simulations was 25 m.

Radiation Balance in Steep Terrain

Traditionally, snow cover models have been one-dimensional and this contribution largely describes processes occurring vertically in and over the snow cover. This chapter has been an exception in that three-dimensional snow distribution is treated. With increasing computer power, future generations of hydrological and meteorological models will have a resolution that is high enough that three-dimensional effects in steep terrain (such as snow transport) become important. For a correct description of the energy balance and therefore mass balance in snow-covered steep terrain, additional processes need to be considered. These processes include shadowing from mountains, multiple reflections of short-wave radiation at slopes, and long-wave radiation emitted by neighboring slopes or other objects. Recently, first treatments of such three-dimensional surface processes have become available (Fierz *et al.*, 2003).

The SNOWPACK software package

The processes discussed above are available from SLF as an integrated software package programmed in the C/C++ programming language and with a JAVA user interface. More information and downloads are possible from the SNOWPACK home page at **`www.slf.ch/snowpack`**.

Acknowledgments

The author acknowledges figure contributions from Ingo Meirold-Mautner, Charles Fierz and Norbert Raderschall. He also thanks Charles Fierz and Schagg Rhyner for comments and suggestions on the manuscript.

FURTHER READING

Gray D.M. and Male D.H. (1981) *Handbook of Snow*, Pergamon Press: Ontario, p. 776.

Jones H.G., Pomeroy J., Walker D. and Hoham R. (2001) *Snow Ecology*, Cambridge University Press, p. 378.

Morris E.M. (1991) Physics-based models of snow. In *Recent Advances in the Modeling of Hydrologic Systems*, Bowles D.S. and O'Connell P.E. (Eds.), Kluwer: The Netherlands, pp. 85–112.

REFERENCES

Albert M.R., Grannas A.M., Bottenheim J., Shepson P.B. and Perron F.E. (2002) Processes and properties of snow-air transfer in the high Arctic with application to interstitial ozone at Alert, Canada. *Atmospheric Environment*, **36**, 2779–2787.

Albert M.R. and McGilvary W.R. (1992) Thermal effects due to air flow and vapor transport in dry snow. *Journal of Glaciology*, **38**(129), 273–281.

Akitaya E. (1974) Studies on depth hoar. *Contributions from the Institute of Low Temperature Science, Series A*, **26**, 67.

Andreas E.L. (1987) A theory for the scalar roughness and the scalar transfer coefficient over snow and sea-ice. *Boundary-Layer Meteorology*, **38**, 159–184.

Andreas E.L. (1995) Air-ice drag coefficients in the western Weddell Sea 2. A model based on form drag and drifting snow. *Journal of Geophysical Research*, **100**, 4833–4843.

Bartelt P., Buser O. and Sokratov S. (2004) A non-equilibrium treatment of heat and mass transfer in alpine snowcovers. *Cold Regions Science and Technology*, **39**(2–3), 219–242.

Bartelt P. and Lehning M. (2002) A physical SNOWPACK model for the Swiss avalanche warning. Part I: Numerical Model. *Cold Regions Science and Technology*, **35**(3), 123–145.

Bintanja R. (2000) Snowdrift suspension and atmospheric turbulence. Part I: Theoretical background and model description. *Boundary-Layer Meteorology*, **95**, 343–368.

Brown R.D. (2000) Northern hemisphere snow cover variability and change. 1915–1997. *Journal of Climate*, **13**, 2339–2355.

Brown R.L., Edens M.Q. and Barber M. (1999) Mixture theory of mass transfer based upon microstructure. *Defense Science Journal*, **49**(5), 363–370.

Brown R.L., Satyawali P.K., Lehning M., Bartelt P. (2001) Modeling the changes in microstructure of snow during metamorphism. *Cold Regions Science and Technology*, **33**(2–3), 91–101.

Brun E., Martin E., Simon V., Gendre C. and Coléou C. (1989) An energy and mass model of snow cover suitable for operational avalanche forecasting. *Journal of Glaciology*, **35**, 333–342.

Calanca P. (2001) A note on the roughness length for temperature over melting snow and ice. *Quarterly Journal of the Royal Meteorological Society*, **127**, 255–260.

Colbeck S.C. (1997) *A Review of Sintering in Seasonal Snow*, CRREL Special Report 97–10.

Colbeck S.C. (1980) Thermodynamics of snow metamorphism due to variations in curvature. *Journal of Glaciology*, **26**(94), 291–301.

Déry S.J. and Yau M.K. (2001) Simulation of blowing snow in the Canadian Arctic using a double moment model. *Boundary-Layer Meteorology*, **99**, 297–316.

Doorschot J. and Lehning M. (2002) Equilibrium saltation: mass fluxes, aerodynamic entrainment and dependence on grain properties. *Boundary-Layer Meteorology*, **104**(1), 111–130.

Doorschot J., Lehning M. and Vrouwe A. (2004) Field measurements of snow drift threshold and mass fluxes and related model simulations. *Boundary-Layer Meteorology* **113**(3), 347–368.

Essery R. (1997) Modelling fluxes of momentum, sensible heat and latent heat over heterogeneous snow cover. *Quarterly Journal of the Royal Meteorological Society*, **123**(543), 1867–1883.

Fierz C. and Lehning M. (2001) Assessment of the microstructure-based snow-cover model SNOWPACK: thermal and mechanical properties. *Cold Regions Science and Technology*, **33**, 123–131.

Fierz C., Riber P., Adams E.A., Curran A.R., Föhn P.M.B., Lehning M. and Plüss C. (2003) Evaluation of snow-surface energy balance models in alpine terrain. *Journal of Hydrology*, **282**, 76–94.

Gallée H., Guyomarc'h G. and Brun E. (2001) Impact of snow drift on the Antarctic ice sheet surface mass balance: possible sensitivity to snow-surface properties. *Boundary-Layer Meteorology*, **99**, 1–19.

Gauer P. (2001) Numerical modelling of blowing and drifting snow in Alpine terrain. *Journal of Glaciology*, **47**(156), 97–110.

Hassanizadeh S. and Gray W. (1990) Mechanics and thermodynamics of multiphase flow in porous media including interphase boundaries. *Advances in Water Resources*, **13**(4), 169–186.

Jordan R. (1991) *A One Dimensional Temperature Model for a Snowcover*, CRREL Special Report, 91–16, U.S. Army Corps of Engineers.

Konzelmann T., van de Wal R.S.W., Greuell W., Bintanja R., Henneken E.A.C. and Abe-Ouchi A. (1994) Parameterization of global and longwave incoming radiation for the Greenland ice sheet. *Global and Planetary Change*, **9**, 143–164.

Lehning M., Bartelt P., Brown B. and Fierz C. (2002b) A physical SNOWPACK model for the Swiss avalanche warning. Part III: Meteorological forcing, thin layer formation and evaluation. *Cold Regions Science and Technology*, **35**(3), 169–184.

Lehning M., Bartelt P., Brown B., Fierz C. and Satyawali P. (2002a) A physical SNOWPACK model for the Swiss avalanche warning. Part II: snow microstructure. *Cold Regions Science and Technology*, **35**(3), 147–167.

Liston G.E. and Sturm M. (1998) A snow-transport model for complex terrain. *Journal of Glaciology*, **44**, 498–516.

Macke A. (1993) Scattering of light by polyhedral ice crystals. *Applied Optics*, **32**, 2780–2788.

Mann G.W., Anderson P.S. and Mobbs S.D. (2000) Profile measurements of blowing snow at Halley, Antarctica. *Journal of Geophysical Research*, **105**(D19), 24,491–24,508.

Marbouty D. (1980) An experimental study of temperature-gradient metamorphism. *Journal of Glaciology*, **26**(94), 303–312.

Massman W. (1998) A review of the molecular diffusivities of H_2O, CO_2, CH_4, CO, O_3, SO_2, NH_3, N_2O, NO and NO_2 in air, O_2 and N_2 near STP. *Atmospheric Environment*, **32**(6), 1111–1127.

Meirold I. and Lehning M. (2004) Measurements and simulation of radiation transfer in snow at Summit, Greenland. *Annals of Glaciology*, **38**, 279–284.

Mellor M. (1975) A review of basic snow mechanics, *IAHS-AISH Publication*, **114**, 251–291.

Miller D.A., Adams E.E. and Brown R.L. (2003) A microstructural approach to predict dry snow metamorphism in generalized thermal conditions. *Cold Regions Science and Technology*, **37**(3), 213–226.

Morland L.W., Kelly R.J. and Morris E.M. (1990) A mixture theory for a phase-changing snowpack. *Cold Regions Science and Technology*, **17**(3), 271–285.

Nakaya U. (1954) Snow crystals. *Natural and Artificial*, Harvard University Press: Cambridge, p. 510.

Nemoto M., Nishimura K., Kobayashi S. and Izumi K. (2004) Numerical study of the time development of drifting snow and its relation to the spatial development. *Annals of Glaciology*, **38**, 343–350.

Petrenko V.F. and Whitworth R.W. (1999) *Physics of Ice*, Oxford University Press, p. 373.

Pomeroy J.W. and Gray D.M. (1990) Saltation of snow. *Water Resources Research*, **26**, 1583–1594.

Sorensen M. (1991) An analytical model of wind-blown sand transport'. *Acta Mechanica*, **1**, (Suppl) 67–81.

Stull R.B. (1988) *An Introduction to Boundary Layer Meteorology*, Kluwer Academic Publishers: Dordrecht, p. 666.

Sturm M., Holmgren J. and König M. (1997) The thermal conductivity of seasonal snow. *Journal of Glaciology*, **43**(143), 26–41.

Sturm M. and Johnson J.B. (1991) Natural convection in the subarctic snow cover. *Journal of Geophysical Research*, **97**(B7), 11657–11671.

161: Water Flow Through Snow and Firn

PHILIP MARSH

NWRI Saskatoon, Saskatchewan, SK, Canada

Water flux through snow and firn has important implications for the timing and magnitude of runoff from seasonal snowpacks, glaciers, and ice caps. Studies conducted over the last few decades have greatly improved our understanding of water flux through snow and firn, including knowledge of the development of wetting fronts as meltwater enters dry snow, preferential flow paths, grain growth of wet snow, and the growth of ice layers and ice columns within the snowpack. Such studies have resulted in the development of a variety of analytical and numerical models for calculating runoff from melting snow and firn. Although our understanding of key processes continues to increase, there are major gaps in our knowledge that limit our ability to formulate comprehensive physically based models of snow metamorphism and water flux through snow and firn. Part of the problem is a lack of data, which is due to the great difficulty of conducting experiments on snow near 0°C and our inability to measure the properties of snow required for estimating the value of parameters needed for numerical simulations. In order to overcome these problems, there is a need for the development of new measurement techniques and carefully designed field and laboratory experiments. Comprehensive models are required for predicting not only melt runoff, but also the release of pollutants from the snowcover, the large-scale properties of the snowcover necessary for understanding large-scale energy-balance and global-change problems, and the release of meltwater from cold glaciers and icecaps under future climate scenarios.

INTRODUCTION

Water flux through snow and firn has important implications for the timing and magnitude of runoff from seasonal snowpacks, glaciers, and ice caps. In addition, it has implications for ecology (Jones *et al.*, 2001), solute release from snowpacks (Tranter and Jones, 2001), and snow avalanches (McClung and Schaerer, 1993). Previous reviews of water flow through snow have been provided by Colbeck (1978), Wankiewicz (1979), Marsh (1991), and Marsh (1999).

Although our understanding of snowmelt runoff processes continues to increase, there are major gaps in our knowledge that limit our ability to model all aspects of water flux through snow and firn. These have been noted by Marsh (1991), Waldner *et al.* (2004), and Gustafsson *et al.* (2004). The most important of these include (i) the relationship between water saturation and grain growth rates; (ii) the effect of grain properties on the water saturation, pressure, and permeability relationships; (iii) the physical properties of ice layers/columns and the mechanisms controlling their formation and persistence; (iv) the mechanisms controlling preferential flow or flow fingers; and (v) the relationship between flow fingers, grain growth, and internal flow systems as revealed by surface rill patterns, for example. Although some aspects of these have been considered over the last decade, our lack of understanding still limits the application of numerical models derived from basic physics (i.e. Tseng *et al.*, 1994; Sellers, 2000; Gray, 1996). A better understanding of the physical processes controlling water flux through snow and firn is required before a complete physically based model of heat and mass fluxes through melting snow covers can be developed.

This article will outline our knowledge of vertical, unsaturated water flux into both dry and wet snow. In addition, the advantages and limitations of the physically based models that are currently available will be discussed. Although important, solute release from snowpacks and the occurrence of snow avalanches are beyond the scope of this article. In addition, as few studies have considered water flux through firn, this article will concentrate on snow, but will discuss firn when possible.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

PROPERTIES OF SNOW AND FIRN

The physical properties of snow and firn influence water movement through these porous media. As detailed information is readily available in other articles of this Encyclopedia (*see* **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**; **Chapter 159, Snow Cover, Volume 4**; **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**) and elsewhere (Gray and Male, 1981; Singh and Singh, 2001; Fountain and Walder, 1998), we will only provide a brief overview of these properties.

Snow covers, either seasonal snow over terrestrial surfaces, glaciers, or multiyear snow found on perennial snow patches or glaciers, are layered. Each layer is formed by snowfall, wind drift, or a combination of these two processes (Figure 1), and modified by compaction, surface melting, and the refreezing of water within the snowcover (Figure 2) (Gustafsson *et al.*, 2004). Each layer varies in terms of grain size, shape, and density (less than $550\,\mathrm{kg\,m^{-3}}$), and resulting liquid permeability, depending on metamorphic changes once the snow is deposited on the ground (Colbeck, 1983, 1986a). Atmospheric processes control the nature of the falling snow (Magono and Lee, 1966; Nakaya, 1954), and therefore the properties of the upper layer of the snowcover. However, metamorphism is sufficiently rapid that snow deposited for more than a day, bears little resemblance to the initial form. When added to the snowcover, liquid water results in large changes in snow properties, including an increase in mean grain size, with the larger grains growing at the expense of the smaller grains (Colbeck, 1974b, 1986b).

Firn occurs in the accumulation zones of glaciers, and is found below the seasonal snowcover and above the glacial ice. It is defined as rounded, well-bonded snow that is older than one year, has partially closed air-pores, and has a density greater than $550\,\mathrm{kg\,m^{-3}}$. It is sometimes also referred to as *névé*. Snow with similar properties is also found in perennial snow patches. In temperate regions, firn develops because of the melting and pronounced settling. In polar regions, settling dominates. Firn formation is the intermediate step in the transformation of snow into glacial ice.

Figure 1 Multiple layers in a premelt snowcover. Each layer has a different hardness and can easily be highlighted by smoothing each layer with a brush. Thermistors are located at each layer. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

WET SNOW/FIRN

Definitions

The liquid-water content of snow/firn has important implications for the metamorphism of the snowpack and runoff from it. The liquid water content (Θ_w) expressed as a fraction of the total volume, is related to liquid saturation (S_w) expressed as a fraction of the pore volume, by

$$S_w = \frac{\Theta_w}{\phi} \tag{1}$$

where ϕ is the snow porosity. For calculations of water flux in snow, two other liquid-water measurements are commonly used. These include the irreducible water saturation (S_i) or irreducible water content (Θ_i), which are defined as the maximum liquid water which is held by capillary forces and is not available for flow, and the effective saturation (S^*), which is defined as the fraction of the pore volume occupied by mobile liquid. S^* is calculated as

$$S^* = \frac{(S_w - S_i)}{(1 - S_i)} \tag{2}$$

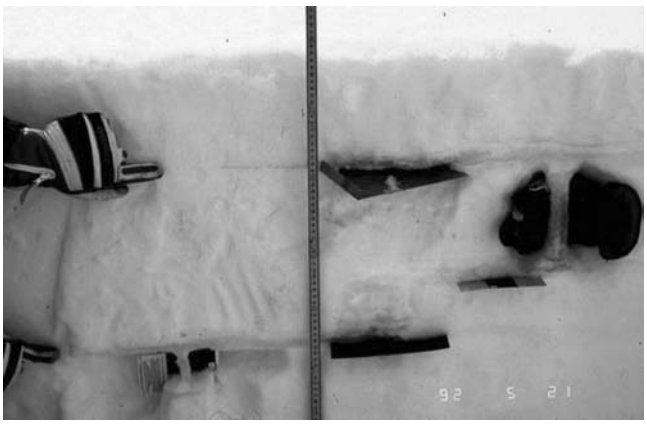


Figure 2 Horizontal ice layers and vertical ice columns are shown in this vertical profile through a melting snowcover. These ice bodies are formed by the freezing of meltwater at premelt horizons (as shown in Figure 1) and by freezing of water in vertical flow fingers. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

S^* is commonly used when considering water flow through snow because snow permeability, which is a function of water saturation, is equal to zero when S^* is equal to zero.

The liquid water content of snow can be measured by a number of techniques, including calorimetry and capacitance probes (Denoth *et al.*, 1984). The measurement of liquid water content in snow is most precise using a capacitance probe and concurrent density measurement. Combined determinations of water content and density using the imaginary part of the dielectric number of ice are less reliable. *In-situ* Time Domain Reflectometry (Schneebeli, 1998) was successfully used to determine the relative changes in water content, but absolute values must be corrected using density from other measurements (Stähli *et al.*, 2004). Brzoska *et al.* (1998) developed a flash-freezing/thick sectioning method to visualize the location of water within the ice matrix. This method is not suitable for field use.

Liquid Water Regimes

Wet snow/firn consists of a sintered ice matrix, enclosed by liquid water of variable thickness, and air. Only at water contents exceeding approximately 30% does the ice matrix dissolve into separated ice grains. The distribution of liquid water in snow is highly variable, with three distinct regimes (Colbeck, 1973). These include (i) the pendular regime, which occurs at low water saturations, where water rings the contact points of the snow grains, with air occurring in a continuous path through the snow. Only a thin liquid layer covers the surfaces of the grains. At very low water saturations the capillary forces are sufficiently large that water flow does not occur; (ii) the funicular regime, which occurs at higher water saturations, where the air becomes isolated as individual bubbles; and (iii) the saturated regime, which occurs when the pore volume is completely filled with water.

Freely draining snowpacks are in the pendular regime. Water saturations are less than 7% for old, coarse-grained snow, but typically range from 7% to 15% depending on the snow type (Colbeck, 1973; Denoth, 1980; Ambach and Denoth, 1980; Denoth, 1999). Typical values for the irreducible saturation (S_i) range from 2% to 7%, depending on snow properties (Lemmela, 1973; Ebaugh and DeWalle, 1977; Colbeck, 1974a; Gerdel, 1948). Field and laboratory measurements show that, as with other porous media, the irreducible saturation (S_i) decreases with increasing grain size. Since snow grains grow in size with the introduction of liquid water, S_i decreases with time after snow is first wetted. As a result, water is initially stored in the pack and is then gradually released as the snow grains grow in diameter.

Relationships Between Water Saturation and Permeability

Marsh (1991) reviewed a number of relationships between water saturation, water pressure, and water permeability of snow. However, the most commonly used relationship relating saturated permeability (κ_s) to grain size and porosity (represented by density) was provided by Shimizu (1970) using both air and kerosene permeameters:

$$\kappa_s = 0.077 d^2 \exp\left[-7.8\frac{\rho_d}{\rho_w}\right] \tag{3}$$

where d is the grain size, ρ_d and ρ_w are the dry snow density and liquid density respectively. Shimizu (1970) found that this equation was applicable to new snow, fine-grained compact snow, and coarse-grained snow. For snow density between 300 and 500 kg m^{-3} and grain size between 0.4 and 2.0 mm, κ_s ranges from 1 to 20×10^{-9} m^2. Equation (3) has often been used for both snow (Marsh, 1991) and firn (Ambach *et al.*, 1981).

Sommerfeld and Rocchio (1993) provided measurements of the relationships between snow density, grain size, and permeability, and suggested the following:

$$\kappa_s = 1.096x10^{-8} \exp[-9.57\rho_d] \tag{4}$$

Major differences noted by Sommerfeld and Rocchio (1993) between equations (3) and (4) include (i) the inclusion of the grain-size term did not improve the relationship; and (ii) permeability at any density averages about half of that suggested by equation (3). Sommerfeld and Rocchio (1993) suggested a number of reasons for these differences, including differences in snow type. They felt that, because

of the small sample size, their results were not definitive and that further studies are required.

Jordan *et al.* (1999), considered the relationship between permeability, capillary rise, and pore structure. Their results show a linear relationship between permeability and the ratio of porosity and the square of capillary rise, and suggest that this is in agreement with Shimizu's equation above.

Saturated (κ_s) and unsaturated permeability (κ) for varying water saturation are related by:

$$\kappa = \kappa_s (S^*)^{\varepsilon} \tag{5}$$

where $\varepsilon = \eta/\lambda$, and η and λ are exponents in relationships between either S^* or κ, and water pressure (Marsh, 1991). Unfortunately there are few estimates of the values of these parameters. However, Wankiewicz (1979) suggested the values for both η and λ from a limited number of samples. Substituting his values for η ($= 13$) and λ ($= 4$) gives $\varepsilon = 3.3$. This value was suggested by Colbeck (1978) for natural snow covers and is similar to the reported values of approximately 2.5 to 5.5, with a median value of 3.5, for sand (Mualem, 1978).

Grain Growth in Wet Snow/Firn

One of the major differences between water movement through typical porous media such as sand and through snow is that snow particles grow rapidly in size with the introduction of liquid water, with the resulting increasing liquid permeability (see Section "Relationships between water saturation and permeability").

The processes responsible for grain growth in wet snow, which at the larger scale is isothermal at 0 °C, have been described in a series of articles by Colbeck (1973, 1974b, 1979a, 1980). Grain growth occurs because of the effect of grain size on the equilibrium temperature among the solid, liquid, and vapor phases of water at the scale of individual grains, with the melting temperature increasing with increasing grain size. The resulting temperature gradient between particles of different radii, produces a heat flux from the larger to the smaller particles, with the larger grains growing at the expense of smaller grains, smaller grains disappearing and the mean grain size increasing with time. Since the heat flux is predominantly through the liquid/solid interface, the heat flux and therefore growth rate is greater with larger water saturations.

Observations of grain growth under saturated conditions in the laboratory (Wakahama, 1968, 1974; Colbeck, 1986b) and in the field (Marsh, 1987) have agreed with theoretical predictions showing (i) the rate of grain growth declines with time; (ii) grain size slowly approaches infinity; and (iii) growth rate increases with increasing water saturation. Colbeck (1986b) suggested that the rate of grain growth could be estimated from:

$$d = d_o + A\ t^{B} \tag{6}$$

where d is the grain diameter (mm), d_o is the initial grain diameter (mm), t is the time in days since the snow was first wetted, and A and B are coefficients. For saturated snow, Colbeck (1986b) suggested that $A = 0.42$ and $B = 0.36$, while Colbeck (1976) estimated $A = 0.1$ and $B = 0.5$ for pendular saturations, from the laboratory data of Wakahama (1968). Marsh (1987) used field observations to suggest $A = 0.44$ and $B = 0.89$.

Brun (1989) provided the first laboratory measurements of grain growth at low water contents typical of natural snowcovers. These results agreed with earlier studies, and confirmed that the mean volume of snow crystals increases at a constant rate, that the rate of increase increases with water content, and that the rate of increase can be expressed as a power function of the water content.

Pressure Gradients in the Snowcover

Assuming that the lower ($S_w = 0.07$) and upper ($S_w = 0.14$) boundaries of the pendular regime are representative of the water-saturation regime of natural snowpacks, then the resulting water pressures vary from approximately -500 to near $-4000\,\mathrm{N\,m^{-2}}$ (Marsh, 1991; Wankiewicz, 1979; Jordan, 1983a). As a result, the pressure gradients in freely draining snowpacks are relatively small (Colbeck, 1978; Wankiewicz, 1979; Pfeffer and Humphrey, 1996) and can often be ignored. However, there are a number of situations where the pressure gradient may be large. These include water flux at very low-flow rates, in the zone immediately above wetting fronts (Pfeffer and Humphrey, 1996), in snowpacks with closely spaced laminations, and above interfaces either within the snow or at the snow pack base. For example, Marsh (1991) showed a hypothetical case where the pressure gradient was $-180\,\mathrm{N\,m^{-2}\,m^{-1}}$ for the gravity-flow zone, and $-8500\,\mathrm{N\,m^{-2}\,m^{-1}}$ in a 0.07 m zone above an impermeable layer.

WATER FLUX INTO DRY SNOW/FIRN

Melting at the snow surface results in the percolation of liquid water into the underlying snow/firn. This movement of water into dry snow/firn is dominated by: (i) the development of a wetting front that separates wet snow/dry snow; and (ii) if the lower layer is below 0 °C, the refreezing of water within the snow. The resulting time lag between the melt and runoff is dependent on whether the snow was dry or already wetted, the snow temperature and snow depth. For seasonal arctic snowpacks, runoff to the stream channel can be delayed by up to 10 days (Marsh and Woo, 1984b), while for deep temperate snowpacks and cold glaciers and ice caps, the delay between melt and runoff may be longer. For glaciers and icecaps, these processes have significant implications for increased runoff and sea level rise under potential climate warming (Pfeffer *et al.*, 1990). Pfeffer

et al. (1991), for example, suggested that refreezing may reduce the runoff contribution from Greenland under a warmer climate by up to 0.4 mm year^{-1} sea level equivalent. Pfeffer *et al.* (1991) reviewed the literature on the influence of freezing on numerous aspects of cold glaciers, including glacier mass balance, changes in snow/firn density and structure, and the interpretation of ice cores.

Field observations of wetting fronts and freezing are difficult to obtain as even slight disturbance of the snow can change the thermal regime of the snow substantially because of the large variability in stratigraphy over small spatial scales and the lack of appropriate measurement techniques. Numerous studies have mapped the location of ice layers and pipes in snow pits (Marsh, 1991; Bøggild, 2000), while Marsh and Woo (1984a) used covered snowpits and excavated the snowpit wall daily to map sequential changes in wetting front locations, ice layers, and columns. Other studies have used tracers (Schneebeli, 1995), snow thick sections (McGurk and Marsh, 1995), buried temperature sensors (Conway and Benedict, 1994; Pfeffer and Humphrey, 1996), heat-flux transducers (Sturm and Holmgren, 1993), and Frequency Modulated Continuous Wave (FMCW) radar (Albert *et al.*, 1999) to infer water movement through the snowpack, and subsequent changes to the snowpack.

Wetting Fronts

The movement of liquid water into dry snow is controlled by capillary forces, which raise the liquid water content from zero to the irreducible water saturation (S_i) (Pfeffer and Humphrey, 1996). The resulting wetting front splits the snow into distinct zones (Yosida, 1973): (i) an upper layer of wet snow that is isothermal at 0 °C (noting that snow is not truly isothermal because of grain-scale temperature gradients that result in grain growth in wet snow) and (ii) a lower dry zone where the liquid water content is zero and the temperature is less than or equal to 0 °C. Pfeffer *et al.* (1990) describe a third zone between these two zones. This dynamic zone is composed of subfreezing snow grains surrounded by water at 0 °C. Unlike pores in soils, snow pores are too large to allow supercooling of water, and as a result, this dynamic zone in snow/firn is in thermodynamic disequilibrium (Pfeffer *et al.*, 1990). This dynamic wetting zone is typically small (less than one cm) and can often be ignored.

The heat and mass transfer in these zones can be described as follows. In the wet isothermal zone, mass and latent heat are transferred by water flow. No sensible heat or conducted heat is transferred as the zone is isothermal at 0 °C. In the dry zone, energy transfer is only by conduction along temperature gradients. In the dynamic zone, energy is transferred by mass transfer and by the release of latent heat of freezing. This energy is transferred from the 0 °C water to the subfreezing snow grains by conduction.

In temperate snowpacks, the snow often reaches an isothermal state before meltwater moves through the snowpack, while in cold snowcovers, the lower portions of the snowpack have a temperature below 0 °C at the start of melt. In this situation, the snowpack is warmed to 0 °C by the refreezing of water within the snowpack, and the subsequent conduction of heat through the dry snow.

Calculating Uniform Wetting Fronts without a Dynamic Zone

Colbeck (1974a) suggested that capillarity diffuses the leading edge of wetting fronts, but that the effect is small compared to the diffusive effect of flow fingering. As a result, and in order to simplify wetting front calculations for estimating the delay between melt and runoff, it is possible to ignore the dynamic zone and to assume that the wetting front is a discontinuous front where the water content jumps from zero to irreducible.

Given these assumptions, the downward rate of wetting front movement is controlled by (i) filling the irreducible water saturation (S_i) and raising the water saturation (S_w) to a level greater than S_i that is sufficiently large to transport the water flux (U) immediately above the wetting front; (ii) if the snow has a temperature below 0 °C, the refreezing of meltwater onto the snow grains at the leading edge of the wetting front. This results in the release of latent heat of fusion, and the raising of the snow temperature immediately below the wetting front to 0 °C. It is assumed that the rate of heat conduction away from the wetting front is sufficiently rapid that it can transfer all of the heat released by the freezing process. In most cases of low water flux and uniform wetting fronts, this assumption is reasonable. When these liquid and thermal storages are filled, additional water is available to move the wetting front downward, but the rate of movement of the wetting front is always slower than that of the meltwater above the wetting front (Yosida, 1973). The downward rate of movement of a horizontal, one-dimensional wetting front ($d\xi/dt$) is then given by (Colbeck, 1976):

$$\frac{d\xi}{dt} = \frac{U}{\phi S_w + (\phi - 1)\dfrac{T_s \rho_i}{(L_f/C_i)\rho_w}} \tag{7}$$

where U and S_w are the melt-water flux and water saturation immediately above the wetting front, ϕ is the snow porosity, T_s is the snow temperature immediately below the wetting front, ρ_i and ρ_w are the ice and water densities respectively, L_f is the latent heat of fusion, and C_i is the specific heat of ice.

For temperate snow with a temperature above approximately −4 °C, the liquid requirements dominate the wetting front advance, while for regions with lower snow temperatures the thermal requirements dominate (i.e. Marsh and Woo, 1984a).

Preferential Flow Paths

Preferential flow paths occur during fluid flow in many types of porous media (Glass *et al.*, 1989) due both to inhomogeneities in the porous media and to flow instabilities (e.g. Beven and Germann, 1982) that are triggered in both uniform and nonuniform media (Hill and Parlange, 1972; Raats, 1973; Glass *et al.*, 1989). In snow, these features are referred to as *flow fingers* (Colbeck *et al.*, 1990). One of the effects of flow fingers is to concentrate water into a smaller cross-sectional area than would occur with a uniform wetting front. For example, Marsh and Woo (1984a) and Kattelmann (1995) showed that flow fingers carry upwards of 70% of the total flow through only 20% of the horizontal cross-sectional area. In a laboratory experiment, Waldner *et al.* (2004) found that preferential flow paths occupied 10% of the horizontal cross-sectional area. Owing to their effect on flow, flow fingers have implications for solute release from snow (Marsh and Pomeroy, 1999), and they reduce the lag between melt and runoff (Marsh and Woo, 1984b; Bøggild, 2000). Examples of flow fingers in snow have been provided by Seligman (1936), U.S. Army (1956), Marsh and Woo (1984a), Kattelmann (1995), Albert *et al.* (1999), Schneebeli (1995), and Waldner *et al.* (2004). Waldner *et al.* (2004) found that flow fingers may have nearly or fully saturated conditions and are sharply delineated from the surrounding dry snow. This results in increased grain growth and therefore higher permeability and water flux.

Glass *et al.* (1989) noted that finger width increased with flow in porous media, and Kattelmann (1995) suggested that with low melt rates, flow fingers are small in size (up to 10 cm across) and occur at densities of up to $100\,m^{-2}$. These sizes are comparable to the mean finger diameter of 4 cm reported by Hill and Parlange (1972) for laboratory soils. Examples have been provided by Wankiewicz (1979), Colbeck (1979b), Marsh and Woo (1984a), Marsh (1988), Kattelmann (1995), McGurk and Marsh (1995), Albert *et al.* (1999), and Williams *et al.* (2000). With high flow rates, and especially with fresh snow, Kattelmann (1995) suggested that larger drains (approximately 30 cm in diameter) commonly occur. Examples have been provided by (Wakahama, 1968; Woo *et al.*, 1982; Kattelmann, 1995). Marsh (1991) suggested that these larger flow paths may also be related to dendritic rill patterns and the related internal flow channels described in a number of environments (e.g. Higuchi and Tanaka, 1982).

Although considerable research has been conducted on flow fingers in porous media (e.g. Glass and Yarrington, 2003), the processes controlling the development of preferential flow paths in snow are not well known. An early attempt to explain these processes was provided in Wankiewicz's (1979) "Flow Impeding, Neutral, or Accelerating"(FINA) conceptual model. FINA describes the interaction of water flux and snow layers as: (i) Impeding: flow is impeded if the snow water pressure required for gravity flow to occur is greater in the upper layer than the lower layer. In this situation, water accumulates above the snow layer boundary until the water pressure in the adjacent part of the upper snow layer increases to a value needed to force penetration of the lower layer. If the boundary is sloping, the stored water will move laterally downslope. (ii) Accelerating: flow is accelerated if the snow water pressure in the lower layer is less than that in the upper layer. In this situation the snow layer boundary will accelerate the flow as the wetting front becomes unstable at this boundary and fingering occurs in the lower layer. (iii) Neutral: layers have no effect on the flow if water pressures are similar on either side of the snow layer boundary. Wankiewicz (1979) hypothesized that the permeability – water-pressure relation for snow exhibits a crossover at some pressure, and as a result any given layer may be impeding, accelerating, or neutral depending entirely on the value of the flux.

On the basis of field observations, Pfeffer and Humphrey (1996) noted that impeding flow occurs most frequently in fine-to-coarse transitions in the snowpack. Waldner *et al.* (2004) described capillary barrier effects or impeding effects in a laboratory experiment with fine/coarse horizons. They also noted that coarse/fine horizons had little effect on water flow.

Waldner *et al.* (2004) carried out one of the few laboratory experiments that considered flow fingers in snow. From this work they suggested that the persistence of flow fingers is affected by two offsetting conditions. First, increasing grain size and permeability and the blocking of pores at the edge of flow fingers result in the persistence of flow fingers. In contrast to this, increasing pore size and high water contents result in gradual lateral movement of the flow fingers. Waldner *et al.* (2004) also noted that preferential flow is extremely nonlinear, and that the use of continuum concepts to model flow through snow using the Darcy–Buckingham law is inadequate for modeling preferential flow.

In addition, the flux value in flow fingers may be sufficiently large that the assumption for equation 7 may not be valid, and a dynamic zone of sufficient thickness forms. The implications of this are discussed in the following section.

Implications of a Dynamic Zone

As described in the previous section, a dynamic zone can develop between the upper wet layer, and the lower dry zone. With larger flux values, especially in flow fingers, this dynamic zone can be sufficiently large to affect the wetting-front advance.

The dynamic zone has two distinct temperatures: water at 0 °C and snow grains with a temperature below 0 °C. Heat fluxes include both intergrain conduction and latent heat from freezing of water onto the grains. Pfeffer *et al.* (1990) suggested that the latent heat fluxes are much larger than the intergrain fluxes, which could therefore be ignored.

Although the dynamic zone is small and can be ignored in simple models of wetting fronts, it is necessary to consider this process in numerical models for use in cold snow (Pfeffer *et al.*, 1990).

Including the dynamic zone in numerical models requires detailed knowledge of the temperature profile, the 3-D snow grain geometry, and the pore scale flow paths. Since such information is not easily obtained, Pfeffer *et al.* (1990) described two possible approaches using simple snow pack geometries. One of these represented snow grains as isolated spheres, and the other considered the flow paths as capillary tubes in ice. Model results suggest that the time required for snow grains to be warmed to 0 °C is of the order of 0.5 s and, as a result, can be ignored in simulations considering the time delay between melt and runoff. However, for detailed models of the physics of water flow, the dynamic zone must be considered.

Refreezing Within Snow/Firn

As noted previously, freezing of meltwater occurs at wetting fronts in order to raise the snow temperature to the freezing point. Depending on a number of factors, the freezing may be sufficient to block the pore spaces, resulting in ice bodies. For uniform wetting fronts, freezing does not fill all pore spaces. However, where the snow is sufficiently cold, freezing may fill all the pore spaces, resulting in an ice body (Colbeck, 1978).

Marsh and Woo (1984a) noted that all refreezing occurs within a zone of mixed wet and dry snow (this is not equivalent to the dynamic zone). This mixed wet and dry snow zone, with wet areas at 0 °C and dry areas with a temperature below 0 °C, occurs between the uniform wetting front and above the deepest flow finger. Within this zone, refreezing of meltwater occurs in various locations. First, as given in equation 7, freezing occurs at the leading edge of the wetting front. This adds mass to the ice grains, but typically is not sufficient to block the snow pores. Second, under certain conditions, the heat flux from the wet snow to the dry, cold snow is large enough to freeze sufficient water to fill the snow pores, resulting in ice bodies. This commonly happens (i) within flow fingers to form vertical ice columns, (ii) at premelt stratigraphic layers to form horizontal ice layers, and (iii) at the base of the snowcover to form an ice layer overlying the underlying impermeable surface. For terrestrial snowcover, this ice layer is referred to as *basal ice* (Woo *et al.*, 1982; Gardner, 1983), while on glaciers it is called *superimposed ice* (Wakahama *et al.*, 1976; Wadham and Nuttall, 2002). Depending on the amount of freezing, ice bodies may be composed of large polycrystalline melt-freeze particles (3 – 5 mm in diameter), which are frozen together into a permeable, honeycomb structure, or if freezing continues, low-permeability ice with a density of over 800 kg m^{-3} forms (Marsh and Woo, 1984a). Albert and Perron (2000) used a permeameter to measure permeabilities for ice layers between 1 and 3 mm in thickness. Typical values were between 1×10^{-10} and 19×10^{-10} m^2, a value considerably higher than for typical snow (equation 3). As noted by Langham (1974b) and Albert and Perron (2000), the permeability of ice layers often doubled within 24 h early in the melt period.

Ice columns form because of the heat flux from the flow finger to the surrounding dry, subzero snow. Growth ceases either when the snow surrounding the ice columns is warmed to 0 °C by conduction of the latent heat released by the freezing of water in the ice column, or when the ice column is overtaken by the background wetting front. Ice columns are similar in size and distribution to flow fingers (Marsh and Woo, 1984a; Kattelmann, 1995), with both small, evenly spaced columns and irregularly spaced, large columns found in both seasonal and perennial snow (Sharp, 1951).

When flow fingers reach a premelt impeding horizon, water spreads laterally along the horizon, with dry, subzero snow both above and below the horizon. This wet layer freezes to form a horizontal ice layer. The rate of ice-layer growth and the total thickness of the layer are related to the heat flux into the dry snow above and below it. Ice layers cease growing for the same reasons as noted for ice columns. Ice layers are commonly of 1 mm to 200 mm in thickness, and vary in horizontal extent from less than 0.2 m to more than 3 m (Langham, 1974a; Marsh and Woo, 1984a).

Basal/superimposed ice forms when the substrate is impermeable and has a temperature below 0 °C and therefore a negative heat flux from the snow to the substrate during snowmelt. Freezing continues until the substrate heat flux is reduced to zero, or the snowcover is removed by melting. Basal ice layers may grow for longer periods of time, and therefore develop greater thicknesses (up to 0.71 m reported by Woo *et al.*, 1982 for a seasonal snowcover) than ice layers within the snowcover. Longer periods of basal ice formation may occur at the base of thick perennial snowpacks. The development of superimposed ice on glaciers may also continue for many weeks, and results in the development of very thick superimposed ice layers. For example, Wadham and Nuttall (2002) describe superimposed ice formation on a high Arctic glacier in both early winter and summer, with the summer formation occurring for over five weeks. The formation of such ice layers plays an important role in the mass balance of such glaciers.

The relative importance of freezing at the wetting front as ice columns or ice layers is not well known. For a cold arctic snowcover, Marsh and Woo (1984a) reported that ice columns were a minor component, with the majority of freezing split between the wetting fronts and ice layers. In this case, the wetting fronts accounted for approximately

12% and ice layers 88% of the total refreezing within the snowpack.

The total volume of water that refreezes is determined by the initial snow temperature and the soil heat flux. In temperate areas, the soil heat flux is negligible, so that refreezing in the snow is controlled entirely by the cold content of the snow. When the snow is underlain by frozen soil, however, the soil heat flux can be large. For example, Marsh and Woo (1987) reported a soil heat flux of $-80\,\mathrm{W\,m^{-2}}$ (i.e. from the snow to the soil). This increased the refreezing within the snow by up to four times compared to that due to the snow cold content alone.

Modeling Wetting Fronts and Refreezing

Analytical, Quasi-two-dimensional Models

Marsh and Woo (1984b) described a quasi-two-dimensional model (MW model) that included the effect of flow fingers, the growth of ice layers, heat flux within the combined snow/soil system, and soil infiltration. The MW model considered the fingering phenomenon by incorporating a two-component wetting front, with one section representing the background front and the other depicting the finger front. It did not, however, consider the development of ice columns. The relative width and the volume of water reaching each front was determined from field observations. The water fluxes at the finger front (U_f) and background front (U_b) were then calculated as:

$$U_\mathrm{f} = U\left(\frac{F_\mathrm{f}}{A_\mathrm{f}}\right) \tag{8}$$

$$U_\mathrm{b} = U\left[\frac{(1-F_\mathrm{f})}{(1-A_\mathrm{f})}\right] \tag{9}$$

where U is the mean surface flux due to either snowmelt or rainfall. U_f and U_b can then be substituted into equation 7 to calculate the velocity of each front.

Freezing of water in the snow was calculated both directly at the wetting fronts from equation (7) and at ice strata within the snowpack. This occurred when the ice stratum was located between the two wetting fronts, with the rate of ice-layer growth controlled by the temperature gradient above and below the ice layer. As long as the water supply from the flow fingers was greater than the rate of freezing, the rate of heat released by freezing equaled the rate of heat conduction away from the ice layer. The water flux reaching the finger front was reduced by the rate of growth of the ice layers, and as a result, the finger front only advanced when the water flux in the fingers was greater than the rate of ice-layer growth. If there was no excess water, then the finger front did not move downward until either the rate of ice-layer growth declined, or the water flux increased.

The MW model also included the heat flux between the snow and soil, the infiltration of meltwater into the soil, and the freezing of this infiltrated water. In addition, if the soil infiltration capacity was exceeded and the soil heat flux was negative (i.e. from the snow into the soil), then a basal ice layer was allowed to form at the snow/soil interface. If a basal ice layer formed, it continued to grow until the underlying soil was warmed sufficiently to decrease the soil heat flux to zero. The growth of the basal ice layer limited the volume of water immediately available to runoff.

The MW model simulated a number of important features of wetting front and icelayer interaction. First, it showed that for cold arctic snow covers the rate of ice-layer growth could be sufficiently large to stop the finger-front advance for short periods of time. Second, ice layers were primarily responsible for warming the snow to near 0 °C, such that by the time wetting fronts entered dry snow, the snow temperature had been raised to near 0 °C. A third result was that the velocity of the background front was not affected by the initial snow temperature, while the finger-front velocity varied greatly with the initial snow temperature. This was not due to freezing at the wetting front, but due to the abstraction of water from the finger to grow ice layers.

VERTICAL, UNSATURATED WATER FLUX IN WET SNOW

Given typical low melt and rainfall rates, and the coarse-grain size and high permeability of wet snow, the vertical flux of water through freely draining, wet snow that is isothermal at 0 °C, is always nonsaturated. Various studies, including Colbeck (1972) and Jordan (1983a), for example, have considered water flux in such conditions. Owing to the complexities of metamorphic changes to wet snow, however, it is usually assumed that such changes are sufficiently slow that they can be ignored, and that the snowcover can be regarded as a simple, rigid porous medium. The following provides a brief overview of the processes controlling water flux through wet snow and various methods for modeling it.

Darcian Flow

Assuming that the snow is within the pendular regime and blockages do not occur, the downward water fluxes are balanced by an equal upward flux of air. The downward, nonsaturated flux of water (U) may be written as

$$U = -\left[\frac{\kappa_\mathrm{s}}{\mu_\mathrm{w}}\right]\left[\frac{\partial p_\mathrm{w}}{\partial z} + \rho_\mathrm{w} g\right] \tag{10}$$

where z is the elevation above a datum, and the permeability (κ) is related to hydraulic conductivity by

$$\mathrm{K} = \kappa \frac{\rho_\mathrm{w} g}{\mu_\mathrm{w}} \tag{11}$$

where ρ_w and μ_w are the density and dynamic viscosity of the water respectively and g is the acceleration due to gravity.

Modeling Vertical, Unsaturated Water Flux

The use of Darcy's Law for modeling unsaturated flow in snow has generally involved either the gravity-flow solution derived by Colbeck (1972, 1978) (i.e. Dunne *et al.*, 1976; Marsh and Woo, 1985), or the application of numerical techniques (i.e. Jordan, 1983b; Bengtsson, 1982; Illangasekare *et al.*, 1990).

Gravity-flow Model

For vertical percolation, where the water pressure gradients are normally less than 2% of the gravitational gradient, the effects of the pressure gradients are usually ignored, and equation (10) simplifies to

$$U = -\frac{\kappa}{\mu_\mathrm{w}} \rho_\mathrm{w} g \tag{12}$$

Colbeck (1972, 1978) used Darcy's Law and an analytical solution of the continuity equation to calculate water flux through wet, ripe snow. Colbeck assumed that the permeability followed a power relationship between the saturated permeability and the effective water saturation, and that the downward flux of water was balanced by the upward flux of air. Using a continuity equation and solving by the method of characteristics, Colbeck showed that any value of flux U propagates through the snow at the speed ($[\mathrm{d}z/\mathrm{d}t]_U$) given by

$$\left[\frac{\mathrm{d}z}{\mathrm{d}t}\right] U = \frac{\varepsilon \alpha^{1/\varepsilon} \kappa^{1/\varepsilon}}{\phi^*} U^{(\varepsilon-1)/\varepsilon} \tag{13}$$

where ϕ^* is the effective porosity calculated as $\phi^* = \phi(1 - S_\mathrm{i})$, and ϕ is the snow porosity.

Equation (13) shows that for constant snowcover properties, the rate of movement of flux U increases with increasing melt-water flux. As a result, during diurnal snowmelt events slow-moving fluxes in the morning are overtaken by faster-moving afternoon fluxes, and a flow discontinuity or shock front develops which is bounded by large fluxes ($U+$) above the front and smaller fluxes ($U-$) below the front. The rate of propagation of this shock front ($\mathrm{d}\zeta/\mathrm{d}t$) can be calculated as (Colbeck, 1978):

$$\frac{\mathrm{d}\zeta}{\mathrm{d}t} = \alpha^{1/\varepsilon} \frac{{\kappa_s}^{1/\varepsilon}}{\phi} [U+^{(\varepsilon-1)/\varepsilon} + U+^{1/\varepsilon} + U-^{1/\varepsilon} + U-^{(\varepsilon-1)/\varepsilon}] \tag{14}$$

Equations (13) and (14) may be used to predict the water flux through unsaturated, wet, ripe snow. The snowpack parameters required to use these equations include κ_s, ε, ϕ, and S_i.

The surface melt data used as the input are normally obtained from calculation of surface melt, which may be determined from either temperature-index or energy-balance methods (Male and Gray, 1981). The output is the water flux at any depth within the cover. This technique has been successfully applied to seasonal snow (Marsh and Woo, 1985) and firn (Ambach *et al.*, 1981).

Albert and Krajeski (1998) provided an analytical solution to water flow through snow – a method that is more easily implemented in a computer model than Colbeck's equations (13) and (14), but which provides very similar results.

Gravity-flow Methods for Uniform Snow/Firn

The gravity-flow equations may be used to reconstruct the melt-flux hydrograph at any depth in the snowpack. Dunne *et al.* (1976) described a graphical method for applying equations (13) and (14). Tucker and Colbeck (1977) developed a Fortran program that utilizes these two equations, and Albert and Krajeski (1998) used another analytical approach. Either approach provides satisfactory results, reproducing the major components of melt-water flux waves that have been observed in snow covers. For example, with increasing snow depth the rise to the peak steepens, the peak flow decreases in volume and arrives later, and the recession limb of the hydrograph lengthens and the minimum flow increases.

The gravity-flow model predicts the general timing of the rise to the peak and the recession from the peak in both homogeneous and heterogeneous snow. A common problem, however, is modeling the rising limb of the flux wave and the magnitude of the peak flux. In homogeneous snow, this problem is related entirely to the effect of pressure gradients that diffuse the wetting front. As a result, a shock front does not form. In heterogeneous snow, the problem with the rising limb is even more pronounced because of the additional effect of flow fingers, ice layers, and pressure gradients at interface boundaries. The effects of flow fingers and ice layers are believed to dominate, increasing the flow in some sections of the pack and decreasing it in others. This highly variable flow over short horizontal distances results in flow velocity being increased in some sections and decreased in others. The overall effect is that, on a sufficiently large scale, the rising limb is diffused. The relative importance of the pressure gradients versus flow fingers and ice layers in diffusing the rising limb is not well known.

Two methods have been used to overcome these problems. The first considers the snowpack as an equivalent uniform medium which accurately represents a natural, heterogeneous porous medium (Freeze, 1975; Colbeck, 1975,

1979b). This approach uses a snowpack equivalent permeability that is obtained from lysimeter measurements of water flux through snow (McGurk and Kattelmann, 1986; Marsh, 1991). The second approach, which is described in the following section, uses a multiple flow path model to estimate the effects of such variable flow in heterogeneous snow/firn.

Heterogeneous Flow in Snow/Firn

Water flux through wet snow is usually highly variable over short horizontal distances. Marsh and Woo (1985) used lysimeters to demonstrate flow variability at the sub-1 m^2 scale. They measured flow variability with standard 0.25 and 1.0 m^2 lysimeters and a 0.25 m^2 multicompartment lysimeter with 16 separate compartments. They found that between lysimeters, the flow was relatively uniform, always being within approximately 10% of the mean flow. Within a lysimeter, however, the flow varied from 0 to 240% of the mean flow. Although the multicompartment lysimeter was composed of discrete compartments, Marsh and Woo (1985) suggested that in reality the flow varied continuously from areas of high flow to areas of low flow, so that the flow could be characterized by a smooth curve. The degree of variability was dependent on flow volume, being more variable on low-flow days and less variable on high-flow days. In addition, they noted that the variability did not seem to be related to the number of ice layers in the snowcover or the snowcover depth. These characteristics of the variation in flow were caused by several factors: (i) the ice layers did not have distinct drains, but instead had continually varying permeability, such that the flow was concentrated beneath some areas and depleted between others with a smooth variation between the extremes; (ii) the principal effect of multiple ice layers is not to increase the flow variability, but instead to redistribute the location of high- and low-flow zones; and (iii) with increasing flow, lateral redistribution both at ice layers and within the flow helps to even out the flow. Kattelmann (2000) used lysimeters up to 10 m^2 in area to consider the spatial variability of flow in deeper, mountain snowpacks. He suggested that the rule of thumb that the flow of water is uniform over areas greater than the square of the snowpack depth is reasonable (Male and Gray, (1981)).

McGurk and Kattelmann (1988) and McGurk and Marsh (1995) used thick-section photography to demonstrate the existence of areas of larger grain size, which they interpreted as flow fingers with higher flow rates. These studies showed flow finger spacing at a scale similar to that of Marsh and Woo (1985).

Sommerfeld *et al.* (1994) and Williams *et al.* (1999) reported integrated studies from lysimeters, aerial photographs, and surface wetness measurements to demonstrate that flow variability also occurred at a larger scale than showed by Marsh and Woo (1985). They demonstrated that a correlation length of 5 to 7 m was typical. Other studies by Schneebeli (1998) also showed preferential flowpaths at a scale of 5 m.

Multiple Flow-path Models

Colbeck (1979b) suggested that variations in flow could be modeled by extending the "gravity-flow theory ... to describe simultaneous flow along different paths", including flow along fingers and flow that passes directly through ice layers. Colbeck described a technique that accomplished this by including the generation of fingers at each ice layer, with the flow in each finger being larger than the background flow. The application of this model was limited by the need to know (i) the number of ice layers; (ii) the number of fingers generated per unit area at each ice layer; and (iii) the flow in each finger. Marsh and Woo (1985) took a different approach in developing a multiple flow-path model based on Colbeck's gravity-flow theory. Instead of modeling the interaction with each ice layer, they made direct measurements of the flow variability in natural snow covers and then used this information to scale the flow in each flow path. Using this approach there is no need to know whether flow fingers, ice layers, or grain growth in wetter parts of the pack are responsible for variations in flow. Instead, all that is required is knowledge of the distribution of flows. To be applied to different snow covers, therefore, it is necessary to either measure the flow variability or to develop relationships between snow properties and flow variability. A major assumption of this model is that each flow path extends the complete depth of the snowcover and that the lag time due to lateral flow along ice layers or the residence time at each ice layer can be ignored.

In the examples given by Marsh and Woo (1985), 10 flow paths were used as a compromise between accuracy in representing the continuum of flow conditions and increased computation time. The flux in each path is then calculated as:

$$U_{\mathrm{F}}(j) = U(j)\left(\frac{V}{S}\right) \tag{15}$$

where $U_{\mathrm{F}}(j)$ is the flow in each path for hour j, $U(j)$ is the mean surface flux, S is the fraction of unit area covered by each path ($S = 0.1$ for 10 flow paths) and V is the percent of flow occurring within each flow path. This was determined from the multicompartment lysimeter measurements. For example, using measured flow data, the path (10% of the total area) with the least water would carry approximately 6% of the total flow, while the path (10% of the total area) with the most water would carry 22% of the total flow.

Since the arctic snowpacks studied by Marsh and Woo (1985) did not exhibit large lateral variations in snow properties, each flow path was assumed to have the same snow properties. The snow permeability was calculated

from equation (3), the water flux in each path was then calculated from equations (13) and (14) using the program of Tucker and Colbeck (1977) and the total flow from all flow paths was calculated by summing all the individual paths. The results of this model predicted the time of the hydrograph rise to within 1 h without any systematic error in flow magnitude. The rate of rise to the peak, the peak flux and the recession were also in good agreement with the observed.

In order to apply the Marsh and Woo (1985) multi flow-path model to different environments it is necessary to know the relationship between flow variability and snowpack properties in different environments. Do different snow covers respond differently in terms of flow variability, or are the flow variability data presented by Marsh and Woo (1985) representative of snow in other environments? Unfortunately the answer to this question is not well known. Examples of other environments are shown by Kattelmann (1995), Sommerfeld *et al.* (1994), and Williams *et al.* (1999).

NUMERICAL MODELS

Various methods have been used to develop numerical models of water movement through snow (Illangasekare *et al.*, 1990; Tseng *et al.*, 1994; Gray, 1996; Sellers, 2000). Each has various advantages and disadvantages, but all are currently limited by basic unknowns of the physical properties of snow. Given these limitations and the extensive nature of the equations, details of various numerical models will not be given in this article. Instead, readers are directed to the original articles.

Illangasekare *et al.* (1990) and Tseng *et al.* (1994) presented a numerical model of melt-water infiltration into subfreezing snow. This model was developed to consider the physics that govern wetting front advance into snow, and to allow a variety of numerical experiments to consider the importance of various processes described earlier in this article. It was hoped that such an approach would overcome limitations caused by the assumptions required in the above-mentioned models, including the consideration of the dynamic zone, capillary pressure gradients, and 1-D flow, for example.

The model of Illangasekare *et al.* (1990) and Tseng *et al.* (1994) is a two-dimensional, finite element, numerical model that includes an explicit coupling between water flow through partially saturated snow, phase change through the refreezing of meltwater in subfreezing snow, and heat conduction. As noted by Tseng *et al.* (1994) water flow in subfreezing snowpacks is highly complex, and as a result, there was a need to make a number of assumptions for want of experimental data. Some of these include (i) no fingering at the wetting front; (ii) no snowpack settling; (iii) condensation, sublimation, and evaporation within the snowpack are ignored; (iv) heat flux due to convection of air or water is ignored; (v) the dynamic, or nonequilibrium zone is neglected; (vi) hysteresis is ignored; (vii) freezing point depression is ignored; and (viii) there is no transfer of momentum due to viscous or frictional forces. Although these assumptions affect the generality of the model, they still allow the model to be used for investigating various aspects of melt-water infiltration into subfreezing snow. The authors also note that there are severe limitations on the use of this type of model due to uncertainties in the snow-water characteristic curves, which describe the pressure-water saturation relationship and the uncertainties in the water saturation-permeability relationships.

Gray (1996), who also considered various numerical approaches to numerical modeling of water flux in wet snow, noted that these models encounter severe difficulties at low saturation and fail completely when saturation approaches zero. Solution of these problems requires studies into alternate capillary pressure relations. Sellers (2000), presented a numerical model solution based on Colbeck's theory of water transport through snow, which is extended to account for the movement of the snow surface due to surface melting.

SATURATED FLOW

Saturated conditions only occur at impeding layers within the snowcover and at the base of the snowpack if the snow is overlying an impermeable surface. If this saturated layer is sloping, then horizontal, saturated flow occurs along the slope. Flow within this layer can be calculated following the methods of (Colbeck, 1974c). Wankiewicz (1979) suggested that laminar flow occurs only if slopes are less than 10°, and grain size and density are less than 0.5 mm and 300 kg m^{-3} respectively. The existence of nonlaminar, or turbulent flow in snow is suggested by the occurrence of open conduits or macropores. They are common in both ice and firn in glaciers and they have also been reported in ice layers at the base of arctic snow cover (Woo *et al.*, 1982). Kattelmann (1985) reported circular openings of between 5 mm and 3 cm in diameter as well as larger features (from 3 cm in diameter up to 1000 cm^2 in cross section), which were oval in cross section with the width exceeding the height. These macropores were located in the bottom one-half meter depth of the snowpack, with grain diameters of about 2 mm and on slopes of temperature less than 10°.

CONCLUSION

A number of studies conducted over the last few decades have greatly improved our understanding of water flux through snow and firn. As a result, there is considerable

knowledge of the development of wetting fronts in dry snow, the growth of ice layers and columns within the snowpack, and a variety of models for calculating runoff from melting snow covers. Even though there have been a number of attempts to develop numerical models of water flux through cold snow, these have had limited success. The primary deficiency with these models is the lack of understanding of critical processes. Part of the problem is a lack of data, which is due to the great difficulty of conducting experiments on snow near 0 °C. There are deficiencies in our ability to measure the properties of snow required for estimating the value of parameters needed for numerical simulations. The changes in hydraulic properties (porosity and permeability) are not currently measurable, and the ever-changing properties of snow make comparable experiments difficult. In order to overcome these problems, there is a need for the development of new measurement techniques and carefully designed field and laboratory experiments to provide the necessary information required to formulate comprehensive physically based models of snow metamorphism and water flux. These models are required for predicting not only melt runoff, but also the release of pollutants from the snowcover, the large-scale properties of the snowcover necessary for understanding large-scale energy-balance and global-change problems, and the release of meltwater from cold glaciers and icecaps under future climate scenarios.

REFERENCES

Albert M., Koh G. and Perron F. (1999) Radar investigations of melt pathways in a natural snowpack. *Hydrological Processes*, **13**, 2991–3000.

Albert M.R. and Krajeski G. (1998) A Fast, Physically-Based Point Snow Melt Model for Use in Distributed Applications. *Hydrological Processes*, **12**(11), 1809–1824.

Albert M. and Perron F.E. Jr (2000) Ice layer and surface crust permeability in a seasonal snow pack. *Hydrological Processes*, **14**, 3207–3214.

Ambach W., Blumthaler M. and Kirchlechner P. (1981) Application of the gravity flow theory to the percolation of melt water through firn. *Journal of Glaciology*, **27**(95), 67–75.

Ambach W. and Denoth A. (1980) *The Dielectric Behaviour of Snow: A Study Versus Liquid Water Content*, Conference Publication 2153, Rango A. (Ed.), NASA: pp. 69–92.

Bengtsson L. (1982) Percolation of meltwater through a snowpack. *Cold Regions Science and Technology*, **6**, 73–81.

Beven K. and Germann P. (1982) Macropores and water flow in soils. *Water Resources Research*, **18**(5), 1311–1325.

Bøggild C.E. (2000) Preferential flow and melt water retention in cold snow packs in West-Greenland. *Nordic Hydrology*, **31**, 287–300.

Brun E. (1989) Investigation on wet-snow metamorphism in respect of liquid-water content. *Annals of Glaciology*, **13**, 22–26.

Brzoska J.-B., Coleou C. and Lesffre B. (1998) Thin-sectioning of wet snow after flash-freezing. *Journal of Glaciology*, **44**, 54–62.

Colbeck S.C. (1972) A Theory of water percolation in snow. *Journal of Glaciology*, **11**(63), 369–305.

Colbeck S.C. (1973) *Theory of Metamorphism of Wet Snow*, Research Report 313, Cold Regions Research and Engineering Laboratory, p. 11.

Colbeck S.C. (1974a) The capillary effects on water percolation in homogeneous snow. *Journal of Glaciology*, **13**(67), 85–97.

Colbeck S.C. (1974b) Grain and bond growth in wet snow. *Snow Mechanics Symposium*, Publication No. 114, International Association of Hydrological Sciences: pp. 51–61.

Colbeck S.C. (1974c) Water flow through snow overlying an impermeable boundary. *Water Resources Research*, **10**(1), 119–123.

Colbeck S.C. (1975) A theory for water flow through a layered snowpack. *Water Resources Research*, **11**(2), 261–266.

Colbeck S.C. (1976) An analysis of water flow in dry snow. *Water Resources Research*, **12**(3), 523–527.

Colbeck S.C. (1978) The physical aspects of water flow through snow. *Advances in Hydroscience*, **V.II.**, 165–206.

Colbeck S.C. (1979a) Grain clusters in wet snow. *Journal of Colloid and Interface Science*, **72**, 371–384.

Colbeck S.C. (1979b) Water flow through heterogeneous snow. *Cold Regions Science and Technology*, **1**, 37–45.

Colbeck S.C. (1980) Thermodynamics of snow metamorphism due to variations in curvature. *Journal of Glaciology*, **26**(94), 291–301.

Colbeck S.C. (1983) Snow particle morphology in the seasonal snow cover. *Bulletin of the American Meteorological Society*, **64**(6), 602–609.

Colbeck S.C. (1986a) Classification of seasonal snow cover crystals. *Water Resources Research*, **22**(9), 595–705.

Colbeck S.C. (1986b) Statistics of coarsening in water-saturated snow. *Acta Metallurgica*, **34**(3), 347–352.

Colbeck S.C., Akitaya E., Armstrong R., Gubler H., Lafeuille J., Lied K., McClung D. and Morris E. (1990) *The International Classification for Seasonal Snow on the Ground*, International Commission for Snow and Ice (IAHS), World Data Center A for Glaciology, University of Colorado: Boulder, p. 23.

Conway H. and Benedict R. (1994) Infiltration of water into snow. *Water Resources Research*, **30**(3), 641–649.

Denoth A. (1980) The pendular-funicular liquid transition in snow. *Journal of Glaciology*, **25**(91), 93–97.

Denoth A. (1999) Wet snow pendular regime: the amount of water in ring-shaped configurations. *Cold Regions Science and Technology*, **30**, 13–18.

Denoth A., Foglar A., Weiland P., Matzler C., Aebischer H., Tiuri M. and Sihvola A. (1984) A comparative study of instruments for measuring the liquid water content of snow. *Journal of Applied Physics*, **56**, 2154–2159.

Dunne T., Price A.G. and Colbeck S.C. (1976) The generation of runoff from subarctic snowpacks. *Water Resources Research*, **12**(4), 677–685.

Ebaugh W.P. and DeWalle D.R. (1977) Retention and transmission of liquid water in fresh snow. *Proceedings 2nd Conference of Hydrometeorology*, American Meteorological Society: Toronto, pp. 255–260.

Fountain A.G. and Walder J.S. (1998) Water flow through temperate glaciers. *Reviews of Geophysics*, **36**, 299–328.

Freeze R.A. (1975) A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. *Water Resources Research*, **11**, 725–741.

Gardner J.S. (1983) Observations on erosion by wet snow avalanches, Mount Rae Area, Alberta, Canada. *Arctic and Alpine Research*, **15**(2), 271–274.

Gerdel R.W. (1948) Physical changes in snow-cover leading to runoff, especially floods. *Proceedings of General Assembly of Oslo*, Vol. 2, International Association of Hydrological Sciences, Commission of Snow and Ice: pp. 42–51.

Glass R.J., Parlange J.-Y. and Steenhuis T.S. (1989) Wetting front instability 1. Theoretical discussion and dimensional analysis. *Water Resources Research*, **25**, 1187–1194.

Glass R.J. and Yarrington L. (2003) Mechanistic modelling of fingering, nonmonotonicity, fragmentation, and pulsation within gravity/buoyant destabilized two-phase/unsaturated flow. *Water Resources Research*, **39**, 1058, doi:10.1029/2002WRR1542.

Gray J.M.N.T. (1996) Water movement in wet snow. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, **354**(1707), 465–500.

Gray D.M. and Male D.H. (1981) *Handbook of Snow: Principles, Processes, Management and Use*, Pergamon Press: Toronto, p. 776.

Gustafsson D., Waldner P.A. and Stähli M. (2004) Factors governing the formation and persistence of layers in a subalpine snowpack. *Hydrological Processes*, **18**, 1165–1183.

Higuchi K. and Tanaka Y. (1982) Flow pattern of meltwater in mountain snow cover. *Hydrological Aspects of Alpine and High Mountain Areas*, Publication No. 138, International Association of Hydrological Sciences: pp. 63–69.

Hill D.E. and Parlange J.Y. (1972) Wetting front instability in layered soils. *Soil Science Society of America Proceedings*, **36**(5), 697–702.

Illangasekare T.H., Walter R.J. Jr, Meier M.F. and Pfeffer W.T. (1990) Modelling of meltwater infiltration in subfreezing snow. *Water Resources Research*, **26**(5), 1001–1012.

Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (2001) *Snow Ecology: An Interdisciplinary Examination of Snow-Covered Ecosystems*, Cambridge University Press: Cambridge, p. 378.

Jordan P. (1983a) Meltwater movement in a deep snowpack 1. Field observations. *Water Resources Research*, **19**(4), 971–978.

Jordan P. (1983b) Meltwater movement in a deep snowpack 2. Simulation model. *Water Resources Research*, **19**(4), 979–985.

Jordan R.E., Hardy J.P., Perron F.E. Jr and Fisk D.J. (1999) Air permeability and capillary rise as measures of the pore structure of snow: an experimental and theoretical study. *Hydrological Processes*, **13**, 1733–1753.

Kattelmann R. (1985) Macropores in snowpacks of Sierra Nevada. *Annals of Glaciology*, **6**, 272–273.

Kattelmann R.C. (1995) *Water Movement and Ripening Processes in Snowpacks of the Sierra Nevada*, Ph.D. Thesis, University of California, Santa Barbara. 94 pp.

Kattelmann R. (2000) Snow melt lysimeters in the evaluation of snowmelt models. *Annals of Glaciology*, **31**, 406–410.

Langham E.J. (1974a) Phase equilibria of veins in polycrystalline ice. *Canadian Journal of Earth Sciences*, **11**, 1280–1287.

Langham E.J. (1974b) The mechanism of rotting of ice layers within a structured snowpack. *Snow Mechanics*, Publication No. 144, International Association of Scientific Hydrology: pp. 73–81.

Lemmela R. (1973) Measurements of evaporation-condensation and melting from a snow cover. *Proceedings, the Role of Snow and Ice in Hydrology*, Unesco-WMO-IASH, pp. 670–679.

Magono C. and Lee C.W. (1966) Meteorological classification of natural snow crystals. *Journal of Faculty of Science, Hokkaido University, Series*, **7**(2), 321–335.

Male D.H. and Gray D.M. (1981) Snowcover ablation and runoff. In *Handbook of Snow: Principles, Processes, Management and use*, Gray D.M. and Male D.H. (Eds.), Pergamon Press: Toronto, pp. 360–436.

Marsh P. (1987) Grain Growth in a Wet Arctic Snow Cover. *Cold Regions Science and Technology*, **14**, 23–31.

Marsh P. (1988) Flow fingers and ice columns in a cold snow-cover, *Proceedings 56th Western Snow Conference*, Kalispell, pp. 105–112.

Marsh P. (1991) Water flux in melting snow covers. *Advances in Porous Media*, Vol. 1, Elsevier: Amsterdam, pp. 61–124.

Marsh P. (1999) Snowcover formation and melt: recent advances and future prospects. *Hydrological Processes*, **13**, 2117–2134.

Marsh P. and Pomeroy J.W. (1999) Spatial and temporal variations in snowmelt runoff chemistry. *Water Resources Research*, **35**, 1559–1567.

Marsh P. and Woo M.K. (1984a) Wetting front advance and freezing of meltwater within a snow cover 1. Observations in the Canadian Arctic. *Water Resources Research*, **20**(12), 1853–1864.

Marsh P. and Woo M.K. (1984b) Wetting front advance and freezing of meltwater within a snow cover 2. A simulation model. *Water Resources Research*, **20**(12), 1865–1874.

Marsh P. and Woo M.K. (1985) Meltwater movement in natural heterogeneous snow covers. *Water Resources Research*, **21**(11), 1710–1716.

Marsh P. and Woo M.K. (1987) Soil heat flux, wetting front advance and ice layer growth in cold, dry snow covers. *Workshop on Snow Property Measurements*, National Research Council of Canada: Technical Memorandum 140, pp. 497–524.

McClung D. and Schaerer P. (1993) *The Avalanche Handbook*, The Mountaineers: Seattle, p. 271.

McGurk B.J. and Kattelmann R.C. (1986) Water flow rates, porosity and permeability in snowpacks in the Central Sierra Nevada. *Cold Regions Hydrology Symposium*, American Water Resources Association: pp. 359–366.

McGurk B.J. and Kattelmann R.C. (1988) Transport of liquid water through Sierran snowpacks: flow finger evidence from thick section photography. *American Geophysical Union, EOS*, **69**, 1204.

McGurk B.J. and Marsh P. (1995) Flow finger continuity in serial thick sections in a melting Sierran snowpack. In *Biogeochemistry of Seasonally Snow-Covered Catchements*, IAHS Publication No. 2289, IAHS: pp. 81–88.

Mualem Y. (1978) Hydraulic conductivity of unsaturated porous media: generalized macroscopic approach. *Water Resources Research*, **14**(2), 325–334.

Nakaya U. (1954) *Snow Crystals: Natural and Artificial*, Harvard University Press: Cambridge, p. 510.

Pfeffer W.T. and Humphrey N.F. (1996) Determination and timing and location of water movement and ice-layer formation by temperature measurements in sub-freezing snow. *Journal of Glaciology*, **42**(141), 292–304.

Pfeffer W.T., Illangasekare T.H. and Meier M.F. (1990) Analysis and modeling of melt-water refreezing in dry snow. *Journal of Glaciology*, **36**(123), 238–246.

Pfeffer W.T., Meier M.F. and Illangasekare T.H. (1991) Retention of Greenland runoff by refreezing: Implications for projected Future Sea Level Change. *Journal of Geophysical Research*, **96**(C12), 22, 117–122, 124.

Raats P.A.C. (1973) Unstable wetting fronts in uniform and nonuniform soils. *Soil Science Society of America Proceedings*, **37**, 681–685.

Schneebeli M. (1995) Development and stability of preferential flow paths in a layered snowpack. In *Biogeochemistry of Seasonally Snow-Covered Catchments*, Vol. 228, Tonnessen K.A., Williams M.W. and Tranter M. (Eds.), International Association of Hydrological Sciences: Wallingford, pp. 89–95, July.

Schneebeli M. (1998) Measurement of density and wetness in snow using time-domain-reflectometry. *Annals of Glaciology*, **26**, 1–10.

Seligman G. (1936) *Snow Structure and Ski Fields*, International Glaciological Society: Cambridge, p. 555.

Sellers S. (2000) Theory of water transport in melting snow with a moving surface. *Cold Regions Science and Technology*, **31**, 47–57.

Sharp R.P. (1951) Features on the firn on Upper Seward Glacier St. Elias Mountains, Canada. *Journal of Geology*, **59**, 599–621.

Shimizu H. (1970) Air permeability of deposited snow. *Low Temperature Science, Series A*, **22**, 1–32.

Singh P. and Singh V.P. (2001) *Snow and Glacier Hydrology*, Kluwer Academic Publishers: Dordricht, p. 742.

Sommerfeld R.A., Bales R.C. and Mast A. (1994) Spatial statistics of snowmelt flow: data from lysimeters and aerial photos. *Geophysical Research Letters*, **21**(25), 2821–2824.

Sommerfeld R.A. and Rocchio J.E. (1993) Permeability measurements on new and equitemperature snow. *Water Resources Research*, **29**, 2485–2490.

Stähli M., Stacheder M., Gustafsson D., Schlaeger S., Schneebeli M. and Brandelik A. (2004) A new in-situ sensor for large-scale snow cover monitoring. *Annals of Glaciology*, **38**, 273–278.

Sturm M. and Holmgren J. (1993) Rain-induced water percolation in snow as detected using heat flux transducers. *Water Resources Research*, **29**(7), 2323–2334.

Tranter M. and Jones H.G. (2001) The chemistry of snow: processes and nutrient cycling. In *Snow Ecology: An Interdisciplinary Examination of Snow-Covered Ecosystems*, Jones, H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 127–167, 378.

Tseng P.-H., Illangasekare T.H. and Meier M. (1994) Modelling of snow melting and uniform wetting front migration in a layered subfreezing snowpack. *Water Resources Research*, **30**(8), 2636–2376.

Tucker W.B. and Colbeck S.C. (1977) *A Computer Routing of Unsaturated Flow Through Snow*, Special Report 77-10, Cold Regions Research and Engineering Laboratory.

U.S. Army (1956) *Snow Hydrology: Summary Report of the Snow Investigations*, U.S. Army Corps of Engineers: Portland, p. 437.

Wadham J.L. and Nuttall A.-M. (2002) Multiphase formation of superimposed ice during a mass-balance year at a maritime high-Arctic glacier. *Journal of Glaciology*, **48**, 545–551.

Wakahama G. (1968) *The Metamorphism of Wet Snow*, Publication 79, International Association of Scientific Hydrology: pp. 370–379.

Wakahama G. (1974) *The role of meltwater in densification processes of snow and firn*. International Association of Hydrological Sciences, Publication No. 114, 66–72.

Wakahama G., Kuriowa D., Hasemi T. and Benson C.S. (1976) Field observations and experimental and theoretical studies on the superimposed ice of McCall Glacier, Alaska. *Journal of Glaciology*, **16**(74), 135–149.

Waldner P.A., Schneebeli M., Schultze-Zimmerman U. and Flühler H. (2004) Effect of snow structure on water flow and solute transport. *Hydrological Processes*, **18**, 1271–1290.

Wankiewicz A. (1979) A review of water movement in snow. *Proceedings Modelling Snowcover Runoff*, Cold Regions Research and Engineering Laboratory: Hanover, pp. 222–252.

Williams M.W., Rikkers M. and Pfeffer W.T. (2000) Ice columns and frozen rills in a warm snowpack, Green Lakes Valley, Colorado, U.S.A. *Nordic Hydrology*, **31**(3), 169–186.

Williams M.W., Sommerfeld R., Massman S. and Rikkers M. (1999) Correlation lengths of meltwater flow through ripe snowpacks, Colorado Front Range, USA. *Hydrological Processes*, **13**, 1807–1826.

Woo M.K., Heron R. and Marsh P. (1982) Basal ice in High Arctic snowpacks. *Arctic and Alpine Research*, **14**, 251–260.

Yosida Z. (1973) Infiltration of thaw water into a dry snow cover. *Low Temperature Science, Series A*, **31**, 117–133.

162: Hydrology of Snowcovered Basins

BRUCE DAVISON AND ALAIN PIETRONIRO

Environment Canada, National Hydrology Research Centre, National Water Research Institute, Saskatoon, SK, Canada

Snowcovered catchments are an integral part of the earth's physical, chemical, and biological regulatory processes. Catchment-based assessments of snow are typically derived from snow surveys and snow courses. Remote sensing is also a valuable tool for use in understanding the nature and distribution of snowcover in a catchment. Snowcovered basin classifications are founded on the source and timing of runoff and the characteristics of the snow. In order to understand the hydrology of snowcovered catchments, it is important to understand the physics of snowmelt and other related phenomena, such as canopy effects that influence the temporal and spatial distribution of snow within a catchment, including both the energy and mass balance components. Operational temperature index models for basin snowmelt and statistical models for horizontal snowcover distribution help approximate some of the physical processes that are not (or cannot be) directly measured. Hydrological models are an attempt to articulate an understanding of the complex and nonlinear interactions within the hydrological cycle using a rational mathematical approach such as an index method or the energy balance approach. A number of projects have been completed or are underway to compare these models. Hydrological applications, avalanche forecasting, weather forecasting, and climate modeling have all driven snow research, with much cross fertilization between these fields.

INTRODUCTION

Large areas of the earth are covered by seasonal or permanent snow. Catchments located at higher altitudes and/or latitudes tend to experience the most snow. In the Southern Hemisphere, less land at lower latitudes limits the amount of snow on the surface of the earth, but results in larger areas covered by sea ice. Figure 1 illustrates the spatial and temporal distributions of snowcover in the Northern Hemisphere for four time periods (Groisman *et al.*, 1994).

In many countries, snow accounts for the largest portion of the water budget. The spring runoff event (freshet) generally produces the annual maximum flow in all but the smallest of rivers. The hydrology of snowcovered basins has long been a focus of study by the hydrological sciences community. As noted in Gray and Male (1981), snow can be a pervasive element that temporarily paralyzes a community. It may have adverse economic impacts when it manifests itself in an extreme form and it has been responsible for many military and socioeconomic failures. Nonetheless, snowcovered basins provide the world with much of its hydroelectric capacity, irrigation and drinking water, and countless recreational opportunities.

Historically, operational hydrologists studied snowcovered basins to determine the snowmelt runoff in these basins. Runoff estimates were used in reservoir management and planning for consumptive use, power generation, public works, and land development. As a result, a number of snowmelt hydrological models were developed to predict snowmelt runoff with a focus on capturing or predicting peak flows and volumes for engineering design and reservoir management purposes. There is also a great need to understand the hydrology of snowcovered catchments for protection against snowmelt-induced floods.

The influence of snow in flooding depends on a number of factors, including basin size, the amount of snow in the basin prior to spring breakup, ice jamming, and the ability of the basin to store melt and rainfall. The importance of these factors is different in different regions. At one extreme, small urbanized catchments with little basin storage experience their annual maximum flood

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

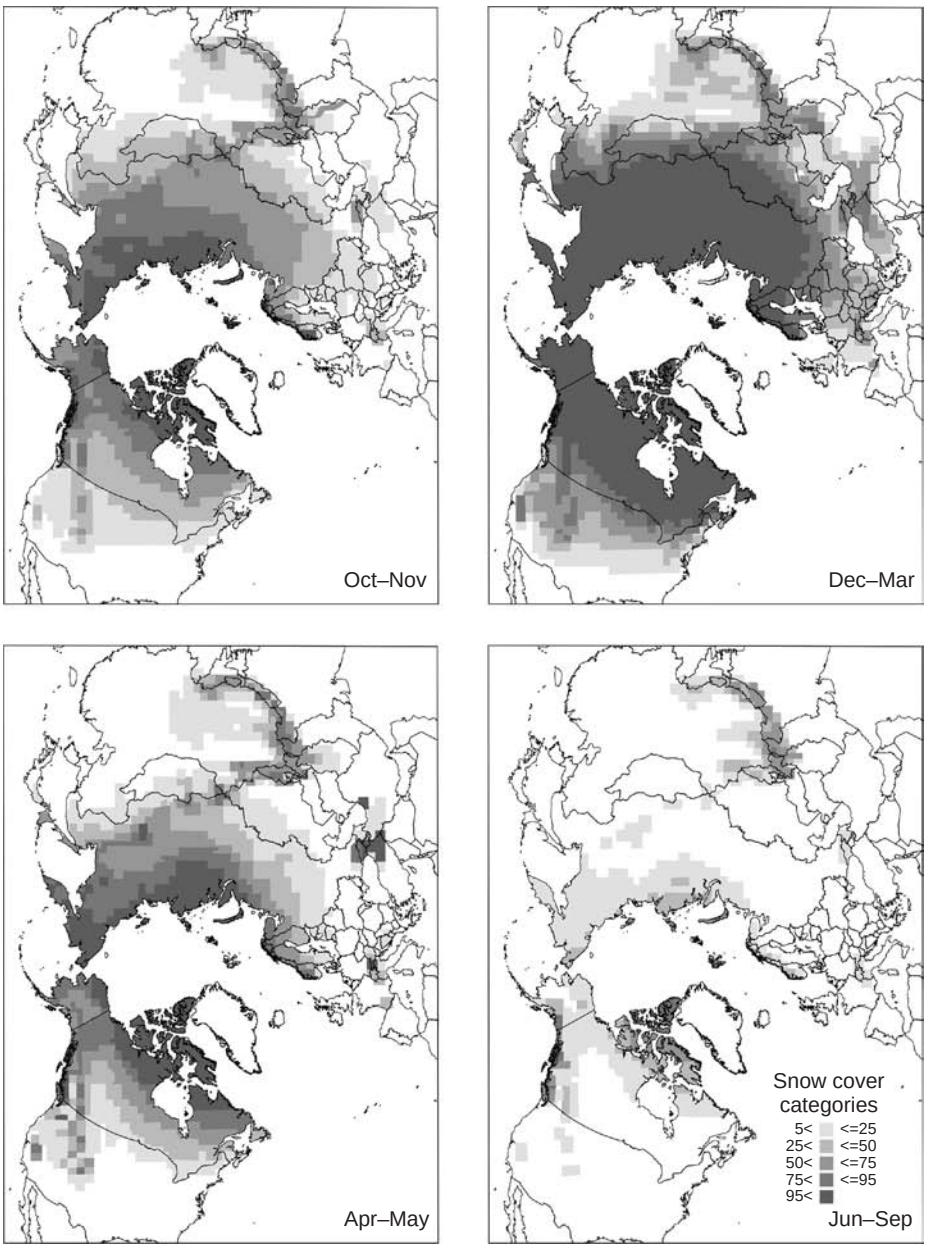


Figure 1 Spatial and temporal distributions of snowcover extent in the Northern Hemisphere for four seasons. Percentage time with snowcover in each season for the period 1973–1992

because of rainfall, regardless of the return period. At the other extreme, large northern watersheds, such as the Mackenzie river basin in Canada or the Lena river basin in Russia, experience their annual maximum flood because of snowmelt. In between these extremes, floods are caused by both snowmelt and rainfall, which is of particular interest to operational hydrologists concerned with the immediate needs of society.

Within this context, four possible cases of frequency relations with respect to snow on the ground and rain can be considered. This assumes two separate and independent probability distributions for the snowmelt and rainfall return periods. The four cases are described by Stoddart and Watt (1970). These four cases depend on the relative sizes of the means and standard deviations of the snowmelt (plus any rainfall) floods and rainfall (without any snowmelt) floods (Stoddart and Watt, 1970; NRC, 1989a). A special case of flooding in snowcovered basins results from ice jamming (NRC, 1989b). Ice jams occur on small and large rivers and have been a concern in many populated areas. The two general kinds of ice jams are freeze-up jams and breakup jams and these events can occur at any time of the year but

are typically coincident with the spring freshet. Attempting to predict ice-jam-induced floods is very difficult. The interested reader is referred to **Chapter 171, River-Ice Hydrology, Volume 4** for a detailed description of ice-covered rivers and ice jams.

While the primary impetus for much of the research on snowcovered catchments has been the importance of snow as a resource, snow also has an important role to play in the energy balance of the Earth's surface as a major influence on climate and weather. Snowcover delays air temperature warming in the spring, intensifies seasonal cooling in the fall (Cohen, 1994), and has considerable local, regional, and global influence on energy exchange through its high surface albedo and low thermal conductivity (Walsh, 1987). Snow also affects the energy balance because of its high latent heat. During melt or freezing, the exchange of latent heat within the snow pack is considerable. During sublimation or condensation, the exchange of latent heat with the atmosphere can also be important. The spatial distribution of snow determines the extent of its impact on the energy balance. Figure 2 illustrates one estimate of the distribution of snow depth over the Northern Hemisphere for December–January–February and March–April–May (from Marshall *et al.*, 1994, based on Foster and Davy, 1988).

The ecology of areas where snow occurs is intimately associated with the nature of the snow pack, and snow plays an important role in the ecological balance of many high latitude and high altitude catchments. As an example, snowmelt is the most significant hydrological event in the boreal forest biome, as its duration and timing exert strong controls over net primary production and the exchange of energy and carbon (Metcalfe and Buttle, 1998). Of all biomes, however, snow has the strongest influence in tundra regions, strongly affecting the vegetation patterns in these regions (Walker *et al.*, 2001). A sufficient pack will provide insulation for plants, protecting their root systems, and will also provide cover for mammals that do not migrate (Groisman and Davies, 2001). Many microorganisms such as bacteria, algae, and fungi have adapted to the extreme fluctuations of energy, water, and nutrients often found in snow environments, setting the stage for nivian (snow) food webs and other biogeochemical processes (Hoham and Duval, 2001). All animals that come into contact with snow are affected by the physical, chemical, and microbiological processes of the snow, possibly influencing the ability of many animals to survive (Aitchison, 2001).

Clearly, snowcovered catchments are an integral part of the earth's physical, chemical, and biological regulatory processes. They play an important role in global water and energy balance. They also effect an important ecological control in cold regions' ecosystems. This article briefly describes the hydrology of snowcovered catchments, with an emphasis on understanding the hydrological response of these cold-region basins. In order to understand the hydrological response of snowcovered catchments, knowledge of the temporal and spatial distribution of snow within the catchment is also very important. Snow pack accumulation and ablation, especially during the spring thaw, are significant inputs into daily hydrological forecasting systems that are in turn extremely useful for flood prevention and hydroelectric generation. Hence, the measurement and characterization of the distribution of snow within a catchment are critical to the prediction of subsequent melt.

MEASUREMENT OF SNOW

In Situ Estimates of Catchment Snowcover

The snowfall within a catchment provides a natural storage element that can release water to a stream at variable rates over time, depending on snow distribution and the physiographic features of the basin in question. Snow water equivalent (SWE) is a fundamental hydrological property that describes the amount of water present in a snow pack and is typically expressed in units of equivalent water depth. The relationship between SWE and snow depth is simply a density conversion. Pomeroy and Gray (1995) noted that, although newly fallen snow is often assumed to have a density of $100\,\mathrm{kg\,m^{-3}}$, densities in the range of $50\,\mathrm{kg\,m^{-3}} - 120\,\mathrm{kg\,m^{-3}}$ are often observed. Temporal increases in density have been measured by Goodison (1981), while Gray *et al.* (1970) noted increases in windpack snow density of up to six times the original density (Pomeroy and Gray, 1995). Seasonal changes in SWE show gradual increases with the density of melting snow, reaching values of $500\,\mathrm{kg\,m^{-3}}$. Clearly, the density of snow is variable in both space and time, making the true estimate of the amount of snow in a catchment difficult to obtain from point measurements. Nonetheless, because of the importance of catchment-wide snow assessments, several methods of measurement of depth, density, and SWE are used routinely in hydrology.

The WMO Guide to Hydrological Practices (1994a) notes that catchment-based assessments of snow are typically derived from snow surveys and snow courses. Snow courses are permanently marked lines where a snow survey, defined as a series of snow depth (and density) measurements, is taken. Surface-based depth and density measurements are obtained from simple ruler measurements of depth and density using a snow core sampler (WMO, 1994b). Density measurements along these transects are usually made with a graduated hollow tube and a cutter, where both the weight and volume of the snow sample are estimated. Snow courses must be carefully selected so that the resulting SWE estimates are representative of the basin. As a general rule, the Guide to Hydrological Practices (WMO, 1994c) recommends that in high-relief regions, snow courses are at

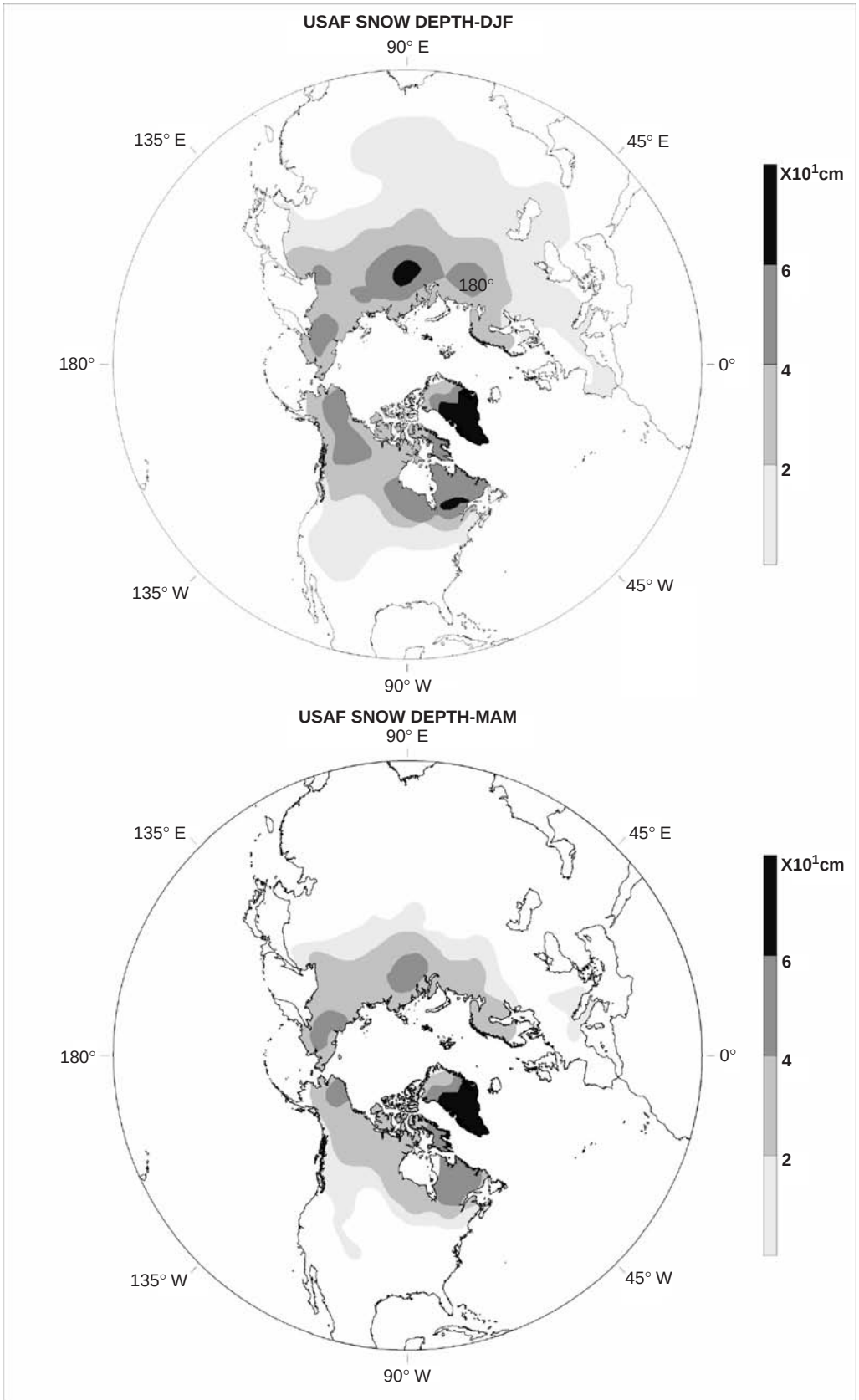


Figure 2 One estimate of the distribution of snow depth for December–January–February and March–April–May

elevations and exposures where there is little or no melting until peak accumulation is achieved. In mild to low-relief regions, these surveys need to represent the average snow conditions within a given catchment, and should be carried out on a variety of landscapes in order to properly depict the natural variability of the landscape.

Point estimates of snow mass can be obtained using a snow pillow apparatus. This method is common in many western states of the United States where traditional snow course methods are too costly or too difficult. Snow depth can also be estimated on tower-mounted instrumentation using ultrasonic depth sounding, where the sensor is placed above the snow pack and the distance to the snow pack is used to calculate snow depth. These point measurements, however, only provide an index of mass or depth at a fixed location in space, and are often supplemented by traditional snow courses.

Remote Sensing of Snowcovered Catchments

Remote sensing is a valuable tool for use in understanding the nature and distribution of snowcover in a catchment. To date, snow can be readily identified and mapped with the visible bands of satellite imagery, and the use of satellite data to map snowcover extent has become operational in several regions of the world. For example, digital snowcover maps of about 4000 basins in North America are produced from NOAA-AVHRR (National Oceanic and Atmospheric Administration – Advanced Very High Resolution Radiometer) images by the US National Weather Service National Operational Hydrologic Remote Sensing Center on a weekly basis and data distribution is possible in real time (Carroll, 1995).

Applications of microwave remote sensing to snow hydrology are also very promising. As an example, since the early 1980s, the Meteorological Service of Canada (MSC) has developed expertise on the use of passive microwaves for estimating SWE for dry snow (Figure 3). Maps of SWE for the Canadian Prairies are produced on a weekly basis using Special Sensor Microwave/Imager (SSM/I) data and are distributed to operational hydrological forecasters (Goodison *et al.*, 1990). Poor spatial resolution (25 km) limits the use of SSM/I to very large areas. Active

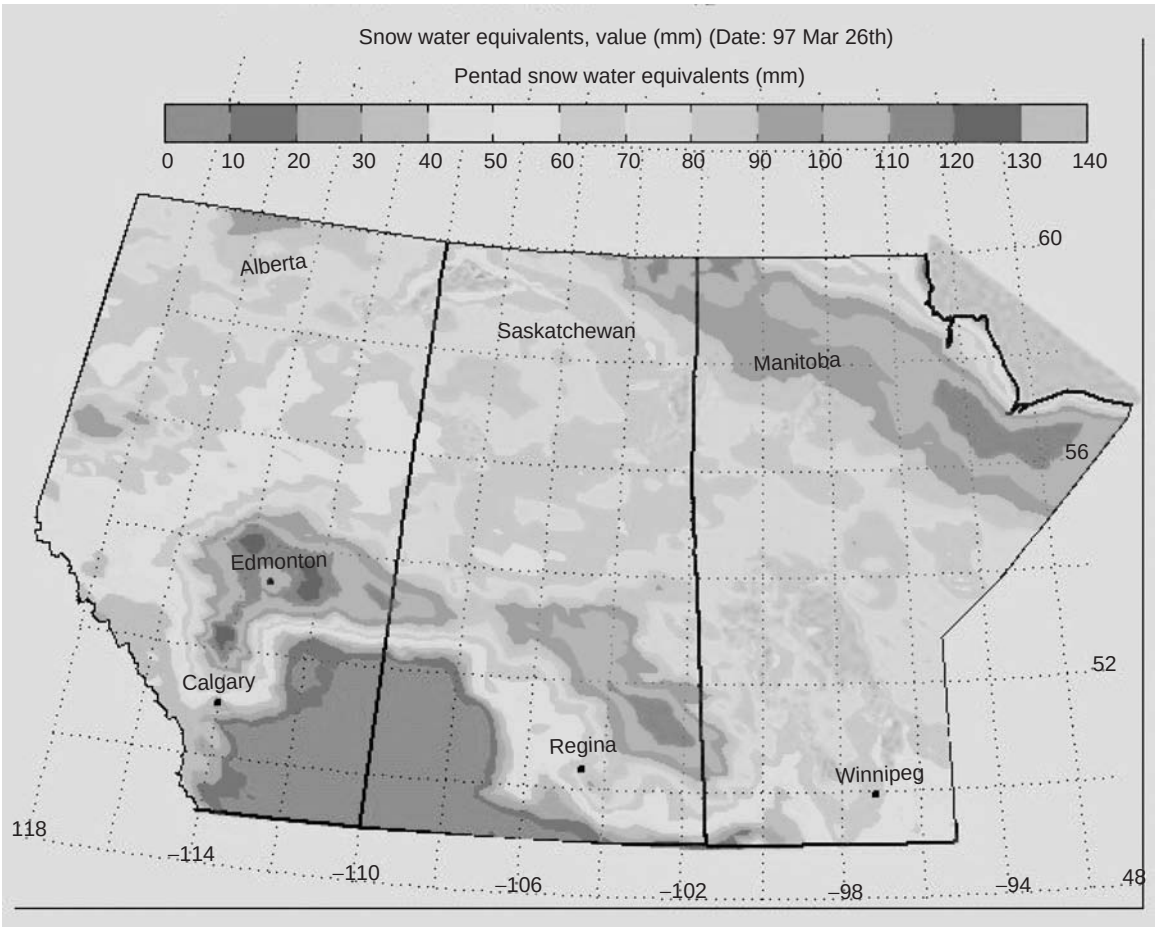


Figure 3 SWE estimates from SSM/I for the Canadian Prairies: March 26th–March 30th, 1997. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

microwave sensors, such as SAR (Synthetic Aperture Radar), can discriminate between wet and dry snow (Shi *et al.*, 1994) and have spatial resolutions (10–100 m) that make hydrological studies at the watershed scale possible.

Visible/near-infrared applications for mapping snowcover are widely accepted (Rango, 1996). For example, Lemieux *et al.*, (1995) have shown that daily remote sensing of snowcover NOAA-AVHRR images is efficient and could become operational in the context of hydropower production. Other promising applications include Landsat Thematic Mapper (TM) derived reflectance and albedo values of snowcovered surfaces (Crevier and Duguay, 1995), and mapping surface snow grain size from Landsat TM data (Brugman *et al.*, 1996). Landsat TM imagery has also been used to derive a canopy closure index, which, combined with ground-based canopy density measurements, allowed the study of spatially distributed snowmelt rates in a boreal forest basin (Metcalfe and Buttle, 1998).

A number of operational remote sensing products for snowcover are available from the United States. The National Weather Service's National Operational Hydrologic Remote Sensing Center maintains time series and spatial products of modeled and observed snow data, including national and regional snow analyses and airborne gamma snow surveys (NOHRSC, 2004). The National Aeronautics and Space Administration's (NASA's) Moderate Resolution Imaging Spectroradiometer (MODIS) satellite produces 500 m and 1 km resolution global snow and ice products within the United States, including daily and 8-day composite snowcover maps (NASA, 2004). In addition, NASA's Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) instrument on the NASA EOS Aqua satellite provides global passive microwave estimates of daily SWE at a 25 km resolution (NSIDC, 2004).

The interested reader is referred to **Chapter 159, Snow Cover, Volume 4** for a more detailed description of the measurement of snow.

CLASSIFICATION OF SNOWCOVERED AREAS

Two kinds of snowcovered basin classifications can be found in the literature. One is a classification founded on the source and timing of runoff and the other is a classification founded on the characteristics of the snow. In the first case, the classification is watershed-based because it is determined from runoff. In the second case, the classification is not watershed-based because it is determined from meteorological and snow conditions.

Basin Classification

Church (1974) has classified northern hydrological regimes on the basis of the source and timing of runoff. The four categories are: subarctic nival, arctic nival, proglacial, and muskeg. The subarctic nival regime exhibits a snowmelt flood in the spring, usually followed by low flows throughout the summer. The low summer flows can be augmented by rainstorm floods that are occasionally higher than the spring flood, especially if mountain snowmelt and rainfall contributions are high. Winter dry subarctic nival regimes, usually larger rivers in the north, freeze entirely and maintain no flow at all in the winter. Perennial flow subarctic nival regimes maintain low flow levels throughout the winter because of groundwater seepage from unfrozen gravels along the rivers.

Arctic nival regimes occur in continuous permafrost zones, usually eliminating the possibility of groundwater flow. The spring melt produces the most streamflow, except for intense rainfall events in very small watersheds. Proglacial regimes are influenced by glaciers that contribute to runoff throughout the summer as higher parts of the glacier experience melt. As a result, the spring melt does not represent as significant a portion of annual streamflow. Arctic and subarctic nival regimes can also mimic proglacial regimes if late-lying snowbanks contribute to streamflow into the summer (Marsh and Woo, 1981). Muskeg regimes are characterized by poor drainage because of shallow slopes, irregular landscapes, and an ability to retain large amounts of water. In tundra zones, the vegetation and hummocky surface restricts flow while the numerous small bodies of water in lowland tundra zones attenuate flow (Prowse, 1990). Hydrologically speaking, the end result is an attenuation of flood flows similar to the effect of lakes.

Figure 4 illustrates Trail Valley Creek (TVC; 68° 45′N, 133° 30′W), a typical arctic nival streamflow regime. The first peak in each year is the result of snowmelt runoff. The summer peaks in each year are the result of convective rainstorms. This 63 km^2 basin is located 40 km northeast of Inuvik, Northwest Territories, Canada (Marsh and Pomeroy, 1996).

Snow Classification

The natural variability of snow has given rise to a number of efforts to classify the physical nature of seasonal snowcover (Formosov, 1946; Richter, 1954; Benson, 1969; McKay and Gray, 1981; Pruitt, 1984; Sturm *et al.*, 1995). The classification by Sturm *et al.* (1995) as described by Groisman and Davies (2001) consists of seven classes of snowcover, namely: tundra, taiga, alpine, maritime, ephemeral, prairie, and mountain.

Tundra snow is a thin, cold, windblown snow usually found above the tree line. Taiga snow is a thin to moderately deep, low-density, cold snowcover found in cold forest climates with low winds, low initial snow densities, and low average winter temperatures. Alpine snow is deep, with intermediate to cold temperatures and often contains

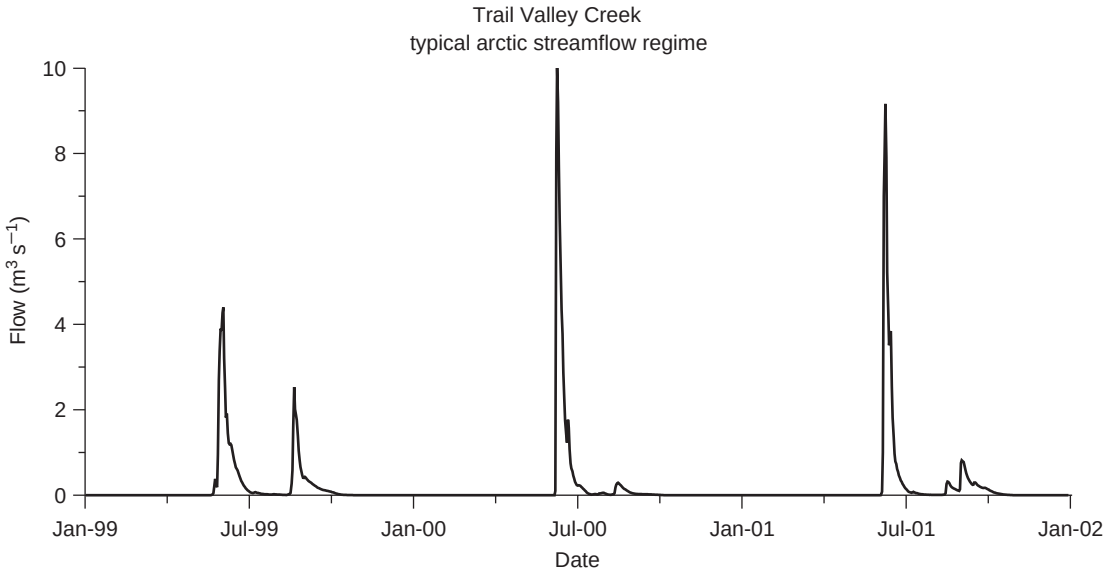


Figure 4 Typical arctic streamflow regime just North of Inuvik, Canada

alternating thick and thin layers. Maritime snow is warm and deep, typically characterized by common melt features such as ice layers and percolation columns. Maritime snow is usually course-grained due to its high moisture content. Ephemeral snow is a thin, extremely warm snowcover exemplified by one snowfall that melts away. Prairie snow is generally thin (except in drifts) and moderately cold, commonly with wind slabs and drifts. Mountain snow is highly variable, depending upon solar radiation effects and local wind patterns. Figure 5 (Sturm *et al.*, 1995) illustrates an estimate of the distribution of these snow classes for the Northern Hemisphere.

SNOW PHYSICS CONCEPTUALIZATION

The Conservation of Energy and Mass

In order to understand the hydrology of snowcovered catchments, it is important to understand the physics of snowmelt and other related phenomena that influence the temporal and spatial distribution of snow within a catchment. The snow pack energy balance and thermophysical process are thoroughly described in **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**, and some key points are described in the following paragraph.

Melting of a snow pack is driven by the energy balance. Conservation of energy dictates that the change in snow temperature is balanced with the energy fluxes entering or leaving the pack on the basis of a system of partial differential equations and appropriate boundary conditions. Shortwave radiation penetration, water transport, phase changes resulting from water transport, and melt–freeze phase changes represent the energy fluxes within the pack. Sensible heat, latent heat, net long-wave radiation, and energy from liquid precipitation represent the energy fluxes at the upper boundary of the pack and ground heat flux represents the energy flux at the lower boundary. Most models assume that the pack is one-dimensional, considering only vertical transfers of energy and water. In reality, horizontal transfers also occur.

The canopy plays an important role in both the energy and mass balances of a snow pack on the ground. Figure 6 illustrates the energy balance under the influence of the canopy. The mass balance implications of a canopy are discussed later.

The canopy intercepts precipitation and changes the relative importance of the energy balance components. Incoming shortwave radiation is reflected or absorbed by the canopy, allowing less shortwave radiation to reach the ground surface. The absorbed shortwave radiation increases the temperature of the canopy, which in turn increases the emission of long-wave radiation from the trees, some of which reaches the snow surface. Trees and other flora alter the vapor gradient through transpiration and the physical presence of gaseous, liquid, or solid water in the canopy, thereby affecting the exchange of latent heat. Wind attenuation reduces the eddies responsible for turbulent mixing, reducing both the latent and sensible heat fluxes at the snow surface. As a result of the reduced latent and sensible heat fluxes in a forest, the turbulent exchanges of heat and water have often been ignored (Male and Granger, 1981). During the melt period, however, these fluxes can have a significant impact if melt is initiated by the movement of a warm air mass

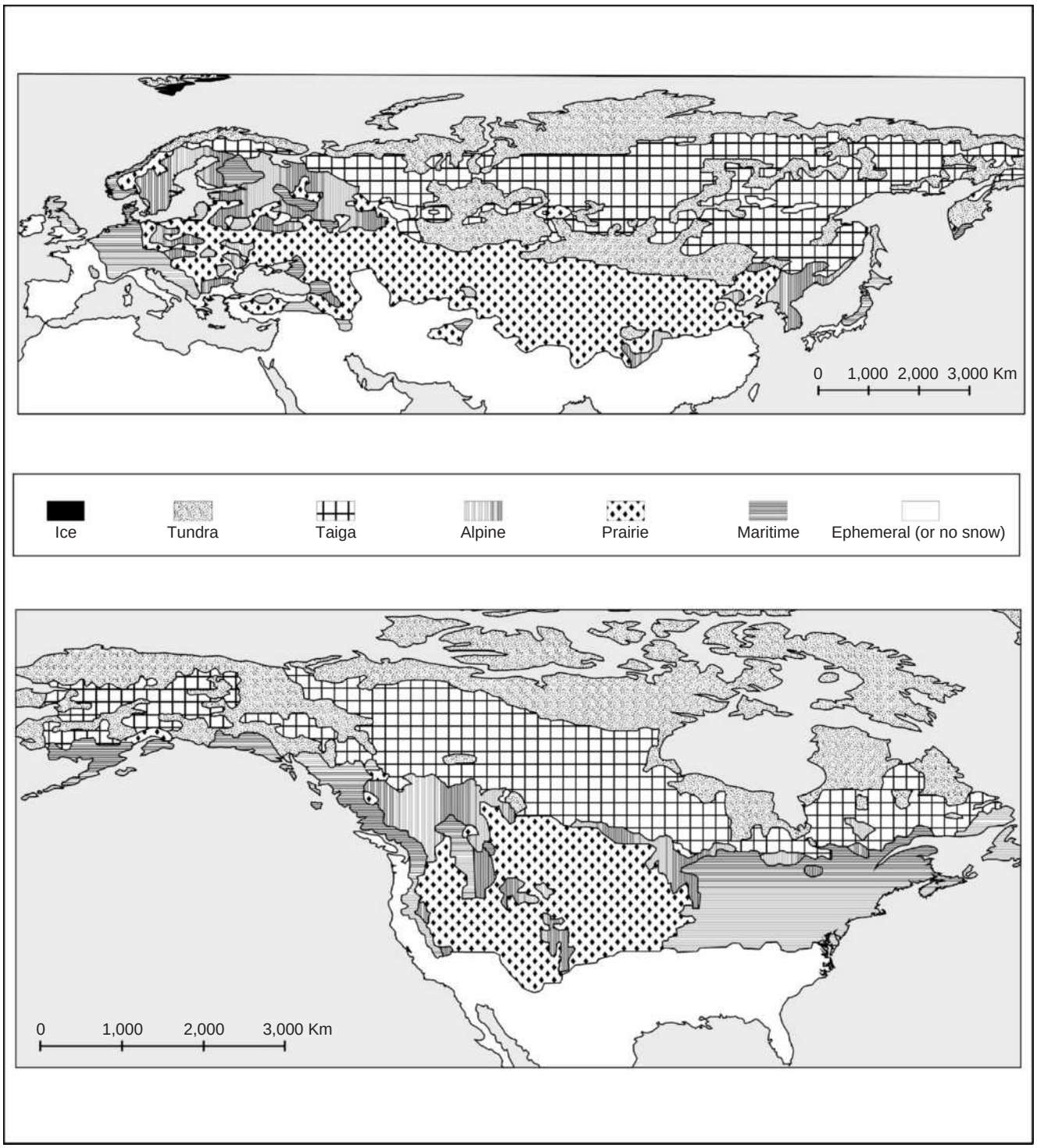


Figure 5 Estimated snow class distribution in the Northern Hemisphere

into an area (Male and Granger, 1981). The relative impact of the turbulent fluxes can also be greater due to the decrease in incoming shortwave radiation, which is only partially offset by the increase in incoming longwave radiation.

The conservation of mass within a snow pack can be described by equation (1).

$$I - O = \frac{dS}{dt} \tag{1}$$

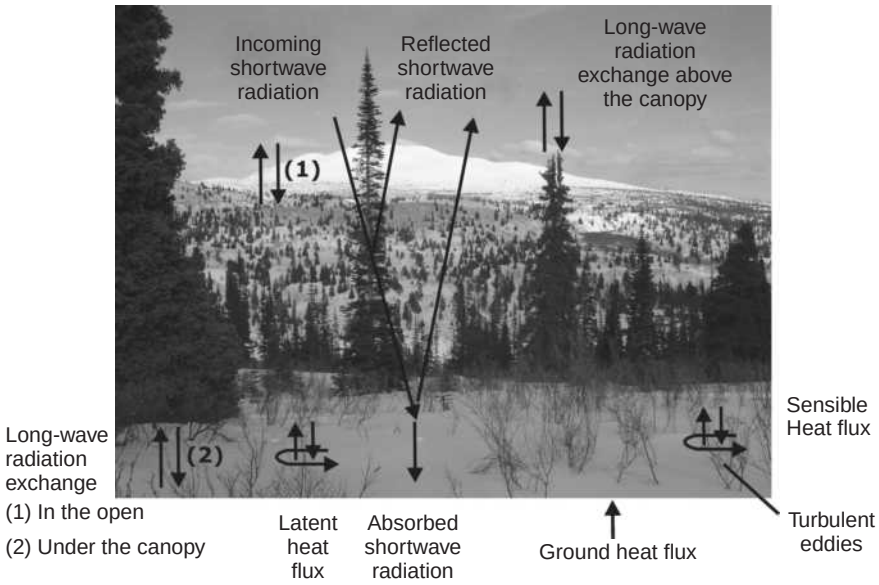


Figure 6 Snowpack energy balance including the canopy

As with any mass balance, equation (1) completely describes the mass regime. 'I' represents the inputs; 'O' represents the outputs and $\mathrm{d}S/\mathrm{d}t$ represents the instantaneous change in storage. Operational hydrologists usually work in units of millimeters or inches for all terms in equation (1) until the water reaches the stream, at which point $m^3\ s^{-1}$ (or $ft^3\ s^{-1}$) are used. Vapor and solid phases of water are always present in a pack and liquid can also be present. The inputs and outputs considered in equation (1) depend upon the purpose of the study. At the most basic level, the inputs are precipitation, condensation, and freezing surface water, while the outputs are sublimation and runoff. Greater detail is required when considering wind redistribution and canopy processes.

Wind redistribution is discussed in **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4** and will not be considered in detail in this article. Equation (1) can, however, be reformulated as in equation (2) to incorporate blowing snow and snow accumulation in wind-swept environments (Pomeroy and Brun, 2001).

$$Q_{\mathrm{surface}} = Q_{\mathrm{snowfall}} - \frac{\mathrm{d}Q_{\mathrm{T}}}{\mathrm{dx}}(\mathrm{x}) - Q_{\mathrm{E}} \qquad (2)$$

where Q_{surface} is the snow accumulation in ($\mathrm{kg\,m^{-2}\ s^{-1}}$); Q_{snowfall} is the snowfall rate ($\mathrm{kg\,m^{-2}\ s^{-1}}$); Q_T is the blowing snow horizontal transport flux ($\mathrm{kg\,m^{-1}\,s^{-1}}$); x is some distance along a fetch (m); and Q_{E} is the sublimation flux ($\mathrm{kg\,m^{-2}\ s^{-1}}$).

To relate the units of equations (1) and (2), 1 mm of water is the equivalent of $1\,\mathrm{kg\,m^{-2}}$. This relationship between mm and $\mathrm{kg\,m^{-2}}$ holds true for water because $1\,\mathrm{m^3}$ of water has a mass of 1000 kg. In addition, equation (2) is represented as a rate.

Snowfall canopy interception describes the process where snowfall is caught by the branches and foliage of a canopy. Some of the snow may get blown off the trees and reach the ground, but much of it sublimates. Snow can also drip from the trees as meltwater (Storck *et al.*, 2002). Hedstrom and Pomeroy (1998) estimated snowfall canopy interception (I) in equation (3) as:

$$\mathrm{I} = \mathrm{c_{suc}}(\mathrm{I^*} - \mathrm{I_0})\left(1 - \mathrm{e}^{-\frac{\mathrm{C_{can}R_s}}{\mathrm{I^*}}}\right) \qquad (3)$$

where c_{suc} is a dimensionless snow unloading coefficient; I^* is the maximum snow load in $\mathrm{kg\,m^{-2}}$; I_0 is the initial snow load in $\mathrm{kg\,m^{-2}}$; R_{s} is the snowfall for a unit of time in mm SWE or $\mathrm{kg\,m^{-2}}$; and C_{can} is the canopy coverage fraction.

Hedstrom and Pomeroy (1998) estimated c_{suc} to be 0.697 and the maximum snow load with equation (4) where S_{p} is a tree species coefficient; *LAI* is the leaf area index; and $\rho_{\mathrm{s}}(\mathrm{fresh})$ is the fresh snow density in $\mathrm{kg\,m^{-3}}$.

$$\mathrm{I^*} = \mathrm{S_p\ LAI}\left(0.27 + \frac{46}{\rho_{\mathrm{s}}(\mathrm{fresh})}\right) \qquad (4)$$

The above discussion and equations represent some of the complex physical constraints that control changes in the snow pack at a single point or within a small region. At the catchment scale, the complexity of these constraints increases and is reflected in the difficulties in capturing the movement, sublimation, and redistribution of snow, and in estimating the radiation balance at all points within the catchment. Because of these complexities,

conceptual methods for describing snow physical properties and subsequent melt at the catchment scale have been developed. These include temperature index melt models and snowcover depletion curves (SDC's), both of which are used in many operational hydrological systems to describe and predict the hydrological response of snow-covered catchments.

TEMPERATURE INDEX MODELS

The temperature index snowmelt algorithm (Anderson, 1973) is a well known algorithm used in many operational models and is given by equation (5):

$$\mathrm{M} = \alpha(\mathrm{T_a} - \mathrm{T_{base}}) \tag{5}$$

Where M is the daily snowmelt depth in mm; α is the melt factor (rate of melt per degree per unit time); T_a is the 2 m air temperature in °C; and T_{base} is the temperature at which the snow begins to melt (assumed to be fixed at 0 °C). This algorithm can be used on a daily or hourly time step. Some authors have suggested that hourly time increments should not be used for temperature index models as the hour-to-hour fluctuations in melting conditions are controlled largely by the radiation component of the energy budget (Rango and Martinec, 1995). However, recent studies using the temperature index model have shown that good results can be obtained in temperate climates such as southern Ontario, Canada, using landscape-based indices (see Donald, 1992; Seglenieks, 1994). The transferability of these parameters in time and space can be problematic, often leading to poor validation results, particularly in regions where radiation melt may dominate (Pietroniro *et al.*, 1996).

In addition to the standard temperature index algorithm, many of these index models also contain a snow pack heat balance accounting system that determines the snowmelt phase of the average snowcover (i.e. warming, ripening, or melt output) for each time step. Such algorithms were developed by Anderson (1973) for the NWSRFS (National Weather Service River Forecast System) Snow Accumulation and Ablation model (Figure 7) and take into account energy storage within the snow pack.

The radiation-temperature indexed model (equation 6) is a combination of the temperature index and the surface radiation budget as proposed by Martinec and de Quervain (1975), Ambach (1988), and Martinec (1989).

$$\mathrm{M} = \alpha(\mathrm{T_a} - \mathrm{T_{base}}) + rn \cdot \mathrm{R} \tag{6}$$

where rn is the conversion factor for energy flux density to mm of snowmelt ($\mathrm{kg\,W^{-1}}$); and R is the net all-wave radiation acting on the snow pack ($\mathrm{W\,m^{-2}}$). The first term in equation (6) attempts to conceptualize the turbulent energy components of the energy budget, namely, the sensible and latent heat exchanges through a simple index. The second term incorporates the surface radiation budget similar to that used in energy balance models. The WMO Guide to Hydrological Practices gives examples of basin-wide melt factors that range from $2\,\mathrm{mm\,C^{-1}}$–$7\,\mathrm{mm\,C^{-1}}$ for mountainous catchments in North America. A similar range of values has been estimated for lowland catchments in the former Soviet Union (WMO, 1994c).

In addition to the above discussion, the reader is referred to Hock (2003) for a comprehensive review of temperature index melt factors.

Snowcover Depletion Curves (SDCs)

The spatial distribution of snowcover is described by SDC's that summarize the percent areal coverage of the snow pack as it increases in average depth. Watershed-wide SDC relationships are currently used in lumped hydrological models such as NWSRFS (Anderson, 1973) to describe the snowcover distribution as the snowcover melts. These relationships are difficult to obtain and require calibration for each specific watershed. Despite the recognized limitations in the application of SDC's to large areas, such as elevation zones or entire watersheds, little progress has been made in the development and application of more stable snowcover representations for use in hydrological modeling. Research on the effect of surface cover on snow pack distribution in lowland regions (McKay and Gray, 1981; Adams and Roulet, 1982; Goodison, 1981; Schroeter, 1988) shows that vegetation roughness tends to determine snowcover distribution. Burkard *et al.* (1991) found that the response of the snow pack distribution to surface cover type could be summarized into three vegetation classes: little to no vegetation, such as ploughed field or pasture grassland; low vegetation such as corn stubble and marsh grasses; and high vegetation such as forests.

The simplest representation of snowcover is uniform snowcover, which is of constant depth and complete areal coverage. This ignores snowcover variability and results in a simulated snowcover that melts uniformly, creating an "instant bare ground" effect when the snowcover disappears. In reality, snowcover is highly variable within a response unit and the nature of the snowcover must be adequately characterized if distributed models are to improve hydrological modeling through improved representation of the state variables (Beven, 1989). The importance of different landscape units and topographical features in estimating the areal distribution of the snowcover is well established (McKay and Gray, 1981; Goodison, 1981; Adams and Roulet, 1982). Knowledge of the areal distribution of the snowcover within and between land units is required to make reasonable estimates of the total

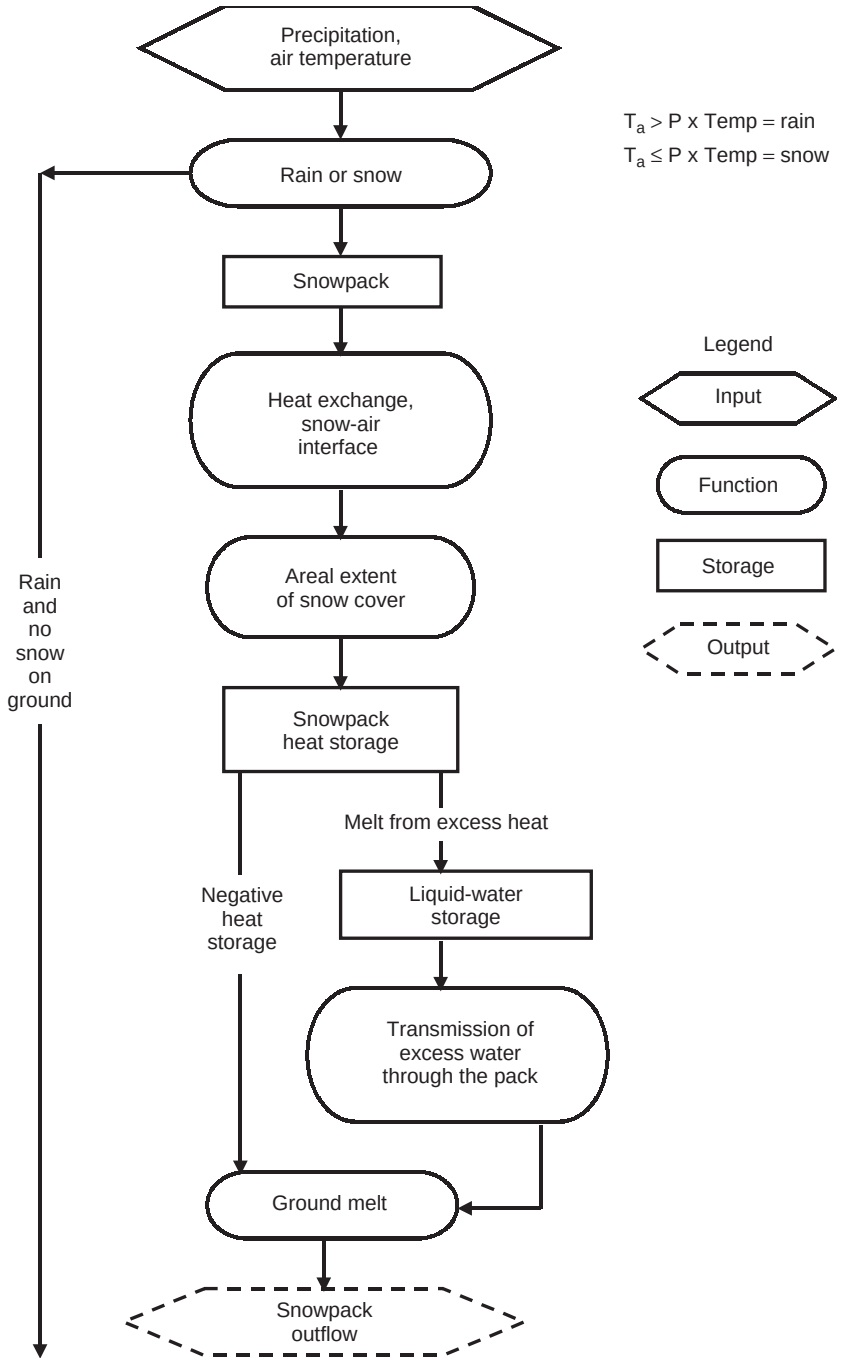


Figure 7 Flow chart of snow accumulation and ablation model

water available in the snowcover of a watershed (Goodison *et al.*, 1987). The areal distribution of the snowcover within a land unit type can be summarized in the form of an Areal Distribution Curve (ADC). An ADC is a summary of the state of the snowcover at a given time. Intense sampling programs are required to develop data sets for quantifying the snow distribution in the form of ADC's.

ADC's are frequency or cumulative frequency distributions of the occurrence of a given snowcover property, usually depth or water equivalent (see Goodison, 1981; Burkard *et al.*, 1991). An important observation in studies of the areal distribution of snowcover (e.g. Schroeter, 1988; Adams and Roulet, 1982; Willis, 1978) is the tendency for the snowcover to become aerodynamically shaped as snow accumulates due to the redistribution of the snowcover by

wind during and after initial deposition. This aerodynamic shape results in a limited capacity for ADC's in different cover types to be estimated from observation of the snowcover at maximum accumulation. The maximum accumulation depth is a function of vegetation height and topography. Another important observation in these studies is the tendency of the snow pack to follow consistent patterns from year to year. The spatial stability of the distributions at different locations within landcover units has not been as extensively studied. Burkard *et al.* (1991), however, indicate that no significant difference exists in land unit snowcover distributions from site to site in southern Ontario.

As noted, for hydrological purposes, it is not practical to physically model the distribution of snowcover (McKay and Gray, 1981). A generalized physically based mathematical model for areal snowcover accumulation and ablation is not yet formulated due to the complexity of the transport process in the natural environment. The development of statistical or empirical distribution relationships based on landcover types is a sensible approach to the problem, and again is accomplished by the use of the SDC, which summarizes the necessary information to model the snowcover depletion.

The scientific literature describes a number of different kinds of SDCs. In their discussion of the necessary resolution for the remote sensing of snowcover, Rango *et al.* (1983) presented a depletion curve where the percentage of snowcovered area (SCA) is plotted against time (see Figure 8). Conceptually, this approach is very easy to understand. In seasonal snow packs, the trend is for snowcovered areas to decrease from 100% to 0% over a period of time. Plotting this depletion curve with sufficiently regular satellite overpasses can result in an accurate calculation of the evolution of basin-wide SCA.

In work concerning the sub-grid parameterization of snow distribution for a lumped model, a second basin-wide SDC is presented by Luce *et al.* (1999) and Luce and Tarboton (2004) (see Figure 9). In this SDC, the SCA is plotted against the basin or element average SWE. The SCA fraction is assumed to be 1 at maximum SWE for the season. As a result, the SDC varies from year to year and from basin to basin with maximum seasonal SWE.

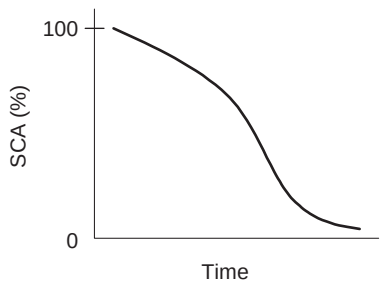


Figure 8 SCA versus Time SDC

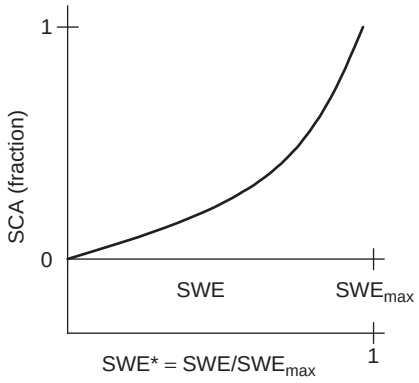


Figure 9 SCA versus. SWE SDC

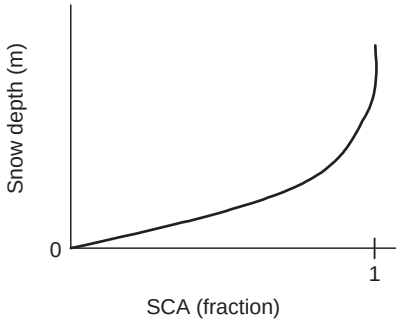


Figure 10 Snow Depth versus SCA SDC

Annual variability is addressed through the development of a dimensionless depletion curve by dividing the SWE by the maximum seasonal SWE. This assumes that the shape of the dimensionless curve is consistent for different values of maximum seasonal SWE. Luce and Tarboton (2004) show this assumption to be reasonable, implying that one basin-wide SDC is consistent from year to year.

A third method of representing snowcover depletion is described in detail by Donald (1992) and Donald *et al.* (1995) (see Figure 10). For this SDC, the average snow depth for the element is plotted against the SCA. The relationship between the SDC and the ADC is also described by Donald (1992) and Donald *et al.* (1995). Mathematically, the SDCs of Donald (1992) and Luce *et al.* (1999) are very similar. The differences between the two curves are subtle. Donald's curve is based on snow depth versus SCA, whereas Luce *et al.*'s is based on a dimensionless SWE term versus SCA. In addition, Donald's curve is landcover-based while Luce *et al.*'s curve is catchment-based. Both methods have their advantages and disadvantages, depending on the problem being solved.

Anderson (1973) discusses a number of ways that snowcover depletion curves can be developed for an entire watershed. In the context of the SCA versus Time SDC, the basin areal snow coverage is measured over a number

of years using aerial photography and related-to-average areal SWE estimated from ground surveys. Anderson recognized that such information was generally not available, nor could it be easily obtained. The same is true today.

Another suggested method, particularly suitable for the SCA versus average snow depth SDC (see Figure 10), is to make periodic measurements of the water equivalent at representative sites within each reasonably homogeneous geographical subarea of the watershed. Vegetation type, elevation, and aspect would be used for subarea selection criteria. Point estimations on the SDC would be possible as the subareas become bare because the percentage of bare area and basin-wide water equivalent would be known. If no data exist, as is often the case, the shape of the SDC must be arbitrarily selected. This SDC could also be modified by calibration during hydrograph simulation. Such watershed SDC's are watershed specific: they represent the characteristic response of the watershed to snow melt, which is a function of the specific surface cover and elevation characteristics within the basin.

The most recent work related to the topic of snowcover depletion curves can be found in Liston (2004), Essery and Pomeroy (2004), and Luce and Tarboton (2004).

ESTIMATING CATCHMENT SNOWMELT

Hydrological models, like all models, are an attempt to articulate an understanding of the complex and nonlinear interactions within the hydrological cycle using a rational mathematical approach. Since the advent of computers, there has been a dramatic explosion of modeling approaches in the literature, representing varying degrees of complexity and producing variable results. Landcover, hydraulic conductivity, surface soil moisture, and snowcover all vary tremendously at different spatial and temporal scales. For example, soil moisture can vary dramatically from one centimeter to the next and from one minute to the next. This has forced hydrological model developers to approach model development from a more conceptual paradigm, while retaining as much physics as possible. Landscape heterogeneity has forced hydrologists to conceptualize the physics and to seek effective parameter values (Pietroniro and Soulis, 2003). Hydrologists have traditionally used the integrated basin response of the hydrograph to test results and have used simple objective functions to describe complex terrain. However, as noted by Beven (1989), the relatively simple hydrologic comparison between observed and modeled streamflow can often lead to reasonable simulations with poor physical representation and/or parameterization. Within the context of snow models, Essery and Etchevers (2004) discuss this problem of parameterization with respect to a simple snow model. Nonetheless, it is useful to examine a number of snowmelt runoff models in the literature that are used operationally, as well as to highlight new approaches to dealing with snowmelt in models, particularly the more complex land-surface models currently being developed in atmospheric models.

Snowmelt runoff simulation models generally consist of a snowmelt model and a transformation function. The snowmelt model generates liquid water from the snow pack that is available for runoff and the transformation model is an algorithm that converts the liquid output at the ground surface to runoff at the basin outlet (Donald *et al.*, 1995). The snowmelt and transformation models can be lumped or distributed in nature. Lumped models use one set of parameter values to define the physical and hydrological characteristics of a watershed. Distributed models attempt to account for spatial variability by dividing the basin into subareas and computing snowmelt runoff for each subarea independently with a set of parameters corresponding to each of the subareas (WMO, 1994d). Snowmelt models generally include a snowcover representation that can range from a simple single-layered snow pack (e.g. Anderson, 1973) to a multilayer conceptual snow pack (e.g. Brun *et al.*, 1989). The snow pack representation has implications for the timing of the snowmelt runoff because of its ability to store water.

Many hydrological models were developed over the last 30 years in an attempt to characterize the behavior of snowcovered basins. Two of the most significant of these were the US NWSRFS (Anderson, 1973; Larson, 2002) and theStreamflow Simulation and Reservoir Regulation Model (SSARR) developed by the US Army Corps of Engineers (Singh, 1995; Croley, 2002). Other important models include the Martinec (1975) snowmelt runoff models developed for the Swiss Federal Institute for Snow and Avalanche Research and the Swedish HBV model (Bergström, 1975) developed at the Swedish Meteorological and Hydrological Institute.

This proliferation of models resulted from developments in computer technology and from a recognition of the increasing importance of managing storage of spring runoff for hydroelectric production, irrigation, water supply, or other consumptive or nonconsumptive uses. As computational efficiencies were realized and geospatial technologies such as geographic information systems and remote sensing were developed, snowmelt modeling was reinvented and several methods evolved to incorporate the distributed nature of a watershed and associated state variables into hydrological models (Donald *et al.*, 1995). The underlying concept in these methods is to discretize the watershed into zones, integrating hydrologically similar units into one larger unit for modeling purposes. An important function of watershed discretizations is to separate the runoff generation processes such as snowmelt, infiltration, and surface flow from stream flow routing. The distributed formulation of the watershed then accounts for the routing of water from unit to unit. Improvement in modeling is possible

by improving the characterization of the runoff generation processes that occur within each response unit. This can be accomplished through proper representation of the distributed nature of the hydrological state variables within the response units. Therefore, the approach to watershed discretization has important implications for snowcover representation.

The distributed representation used in most snowmelt runoff simulation models is the separation of the watershed into distinct elevation bands (WMO, 1986; Liston *et al.*, 1999; Essery, 2003). Elevation bands are appropriate in high-relief regions where snowmelt gradients are strongly related to elevation. For low-relief regions, elevation bands are not significant and for watershed discretization in distributed snowmelt models, it is important to emphasize physiographic differences such as vegetation that can significantly affect the snowcover. Collectively, the SCA and SDC are an attempt to conceptualize the snowmelt representation in both distributed and lumped models.

Catchment runoff is estimated using a number of possible algorithms that represent the physics of a melting snow pack. In many ways, melted snow is treated the same as rainfall, and infiltrated into the soil matrix using a number of possible infiltration algorithms. However, a melting snow pack can contain between 2 and 10 % of liquid water within the pack, resulting in a significant lag in meltwater production at the stream (WMO, 1994c). The melting of the snow pack is driven by the energy budget on the snow pack as described earlier. Although point estimates of all the forcing variables and parameters required to complete the energy budget and estimate the subsequent melt can be obtained through experimentation, in practice such complex treatments are often intractable. The integration of a reasonable snowmelt transfer function over relatively large catchments is extremely difficult without a highly sophisticated monitoring system. In light of these realities, the most widely applied algorithm for determining the melt rate of snow pack in a basin is the degree-day method, also known as the *temperature index method*.

INDEX METHODS FOR ESTIMATING BASIN RUNOFF

The success of a forecast is highly dependent upon how accurately the snow accumulation represents the basin conditions. According to WMO (1994d), there are at least two additional factors that may have some influence on runoff: antecedent groundwater storage and the amount of precipitation during the snowmelt season.

In high-relief areas, when sufficiently detailed measurements at multiple altitudes of the basin are available, the following formula can be used to calculate the weighted mean snow-accumulation index, I_n (WMO, 1994d):

$$\mathrm{I}_n = \frac{\mathrm{A}_1}{\mathrm{A}}\mathrm{w}_{n_1} + \frac{\mathrm{A}_2}{\mathrm{A}}\mathrm{w}_{n_2} + \cdots + \frac{\mathrm{A}_N}{\mathrm{A}}\mathrm{w}_{n_N} \qquad (7)$$

where $\mathrm{w}_{n1}, \mathrm{w}_{n2}, \ldots, \mathrm{w}_{nN}$ are the mean precipitation or SWE in millimeters at various altitudes, and $A_1, A_2, \ldots, A_N$ are the areas at these altitudes, and A is the total area.

Snow courses, which are located at various altitudes, are used to obtain data to establish a relationship between the SWE and the altitude, such that $\mathrm{w} = f(\mathrm{z})$. A different relationship is obtained for each year. Catchment runoff is then estimated using a number of possible algorithms that represent the physics of a melting snow pack.

As noted in the previous section, many snowmelt runoff models use some form of temperature index (degree-day) method to determine when snowmelt occurs and how much snowmelt may occur in a specific period of time. In the case of operational forecasts, the same approaches are used. An example of a well-established conceptual forecast model is the Snowmelt Runoff Model (SRM). This model has been applied in a range of conditions and for catchments of almost any size ($0.76–1\,22\,000\,\mathrm{km}^2$) and elevation range (305–7690 m a.s.l.) (Martinec and Rango, 1997). A model run starts with a known or estimated discharge value and can proceed for an unlimited number of days, as long as the input variables of temperature, precipitation, and snowcovered area are provided (Martinec and Rango, 1986).

Daily forecasts are estimated for the basin according to equation (8):

$$Q_{n+1} = [\mathrm{c_{Sn}}\cdot\mathrm{a}_n(\mathrm{T}_n + \Delta\mathrm{T}_n)\mathrm{S}_n + \mathrm{c_{Rn}}\cdot\mathrm{P}_n]\frac{A\cdot 10\,000}{86\,400}(1 - \mathrm{k}_{n+1}) + Q_n\mathrm{k}_{n+1} \qquad (8)$$

where Q is the average daily discharge ($\mathrm{m}^3\,\mathrm{s}^{-1}$); c is the runoff coefficient, expressing the losses as a ratio (runoff/precipitation), with $\mathrm{c_S}$ referring to snowmelt and $\mathrm{c_R}$ referring to rain; a is a degree-day factor ($\mathrm{cm}\,^\circ\mathrm{C}^{-1}\,\mathrm{d}^{-1}$), indicating the snowmelt depth resulting from a 1 degree-day; T is the number of degree-days ($^\circ$C d); ΔT is the adjustment by temperature lapse rate when extrapolating the temperature from the station to the average hypsometric elevation of the basin or zone ($^\circ$C d); S is the ratio of the snowcovered area to the total area; P is the precipitation contributing to runoff (cm), where a preselected threshold temperature, T_{crit}, determines whether this contribution is rainfall and immediate; A is the area of the basin or zone (km^2); k is the recession coefficient, indicating the decline of discharge in a period without snowmelt or rainfall: $k = (Q_{m+1})/(Q_m)$, (where m and $m + 1$ are the sequence of days during a true recession flow period); n is the sequence of days during the discharge computation period;

and 10 000/86 400 is the conversion from (cm km^2 d^{-1}) to (m^3 s^{-1}). Equation (8) is written for a time lag between the daily temperature cycle and the resulting discharge cycle of 18 h. In this case, the number of degree-days measured on the nth day corresponds to the discharge on the $n + 1$ day. Various lag times can be introduced by a subroutine. T, S, and P are variables to be measured or determined each day. C_r, C_s, delta T, T_{crit}, k, and the lag time are parameters that are characteristic for a given basin or, more generally, for a given climate. If the elevation range of the basin exceeds 500 m, it is recommended that the basin be subdivided into elevation zones of about 500 m each.

ENERGY BALANCE METHODS FOR ESTIMATING BASIN RUNOFF

In addition to lumped, temperature index models, detailed distributed models have also been developed and assessed. These Digital Elevation Model (DEM) grid-based models use the energy balance to calculate snowmelt and to predict runoff from input data on snow properties, terrain and region characteristics, precipitation, and climate. This approach is similar to those used by Anderson (1976) and Morris (1982), but designed to run on simpler, more generalizable inputs. A model of this type was initially presented by Marks (1988), described conceptually by Marks *et al.* (1992) and Marks and Dozier (1992), and then described in detail by Marks *et al.* (1998). The model approximates the snowcover as being composed of two layers and solves for snowmelt for each element of the DEM, taking into account local slope and aspect, and using estimated radiation inputs at the surface to close the energy balance (Susong *et al.*, 1999).

A recent addition to our ability to simulate snowcovered basins has come from the atmospheric research community. The first Global Circulation Models (GCMs) created by atmospheric scientists required some representation of land and ocean processes to provide the appropriate inputs to simulate the atmosphere. These early representations of nonatmospheric hydrologic subcycles first appeared in the late 1960s (Manabe, 1969) and were very simple for reasons of computational efficiency. Snow was lumped with soil, providing extra water storage and altering the energy balance through differing albedo and surface heat capacity. These land-surface schemes (LSS's) are often referred to as the first generation LSS's and many changes have been made in the second and third generation LSS's. With respect to snow, the work is occurring on two fronts. One approach uses the incorporation of better 1-D vertical component models into the LSS's, while the other method involves the incorporation of better 2-D process approximations. Many of the 1-D vertical point processes have been studied extensively in the scientific hydrology community. Work is underway to understand and describe the 2-D physical snow processes such as wind redistribution. Meanwhile, existing methods developed by the operational hydrology community are being incorporated into the LSS's.

SNOW MODEL INTERCOMPARISONS

A number of snow model intercomparison projects have been completed or are underway. In the mid-1980's, the WMO surveyed 11 snowmelt runoff models (WMO, 1986). Since that time, many more snow models have been developed and a more recent intercomparison is taking place with 30 land-surface models driven with meteorological data. This intercomparison is occurring through the Project for Intercomparison of Land-surface Parameterization Schemes (PILPS 2e) Arctic Model Intercomparison Study (Lettenmaier and Bowling, 2001; Bowling *et al.*, 2003). The PILPS 2d also considered a site with seasonal snowcover (Slater *et al.*, 2001). In addition, the International Association of Hydrological Sciences – International Commission on Snow and Ice (IAHS/ICSI) is helping facilitate an intercomparison (SnowMIP) of 40 snow models (Essery and Yang, 2001). These snow models are used for a wide variety of purposes, ranging from climate modeling and hydrology to research in snow physics, snow stability, and avalanche hazard forecasting. Results of this model intercomparison are discussed by Etchevers *et al.* (2004) and Essery and Etchevers (2004). In addition, ICSI has commissioned SnowMIP2 to assess the performance of snow models in forested environments.

CONCLUSIONS

Though the hydrology of snowcovered catchments has long been studied, the nature and character of the snow accumulation and ablation processes are still being explored. Progress in understanding the physical processes that control snowcover changes within a catchment is well in-hand, although there is still much uncertainty in the role of interception and sublimation in the overall water balance. It is clear that in order to understand catchment behavior, appropriate representation of SWE at the catchment scale is essential. Statistical methods, along with appropriate remote sensing techniques, may provide the only practical solutions to accurately estimating catchment SWE. Conceptual snowmelt models such as the SRM have provided the operational community with the necessary tools for robust forecasting in snowcovered catchments. These models, however, cannot explicitly resolve snow pack physics and are of limited use when assessing the energy balance of a system or the hydrology of smaller catchments. Detailed distributed models, using a

more physically based understanding of the melt and accumulation processes have yet to find their way into the operational community and are limited to small basin applications. The statistical characterization of snowcovered area and depletion provides an acceptable compromise to the detailed physical models, and allows for reasonable physics to be incorporated in the snowmelt models. These distributions are incorporated into current LSSs such as the Canadian Land Surface Scheme (CLASS; Verseghy, 2000), and may provide the best means for assessing both the water and energy balance of large catchments while retaining a physical basis for simulating snow accumulation and ablation.

REFERENCES

Adams W.P. and Roulet N.T. (1982) Areal differentiation of land and lake snowcover in a small sub arctic drainage basin. *Nordic Hydrology*, **13**, 139–156.

Aitchison C.W. (2001) The effect of snow cover on small animals. In *Snow Ecology: An Interdisciplinary Examination of Snow-covered Ecosystems*, Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 229–265.

Ambach W. (1988) Interpretation of the positive degree-day factor by heat balance characteristics - west Greenland. *Nordic Hydrology*, **19**, 217–224.

Anderson E.A. (1973) *National Weather Service River Forecast System - Snow Accumulation and Ablation Model*, NOAA Technical Memorandum, NWS HYDRO-17, November, 1973.

Anderson E.A. (1976) *A Point Energy and Mass Balance Model of a Snow Cover*, NWS Technical Report 19, National Oceanic and Atmospheric Administration, Washington, p. 150.

Benson C.S. (1969) *The Seasonal Snow Cover in Arctic Alaska*, Report 51, The Arctic Institute of North America, Calgary.

Bergström S. (1975) The development of a snow routine for the HBV-2 model. *Nordic Hydrology*, **6**, 73–92.

Beven K. (1989) Changing ideas in hydrology - the case of physically-based models. *Journal of Hydrology*, **105**, 157–192.

Bowling L.C., Lettenmaier D.P., Nijssen B., Graham L.P., Clark D.B., Maayar M.E., Essery R., Goers S., Gusev Y.M., Habets F., van den Hurk B., Jin J., Kahan D., Lohmann D., Ma X., Mahanama S., Mocko D., Nasonova O., Niu G.Y., Samuelsson P., Shmakin A.B., Takata K., Verseghy D., Viterbo P., Xia Y., Xue Y. and Yang Z.L. (2003) Simulation of high latitude hydrological processes in the Torne-Kalix basin, PILPS Phase 2e, 1: Experiment description and summary intercomparisons. *Journal of Global and Planetary Change*, **38**(1–2), 1–30.

Brugman M.M., Pietroniro A. and Shi J. (1996) Mapping alpine snow and ice using Landsat TM and SAR imagery at Wapta Icefield. *Canadian Journal of Remote Sensing*, **22**(1), 127–136.

Brun E., Martin E., Simon V., Gendre C., Coleou C. (1989) An energy and mass model of snow cover suitable for operational avalanche forecasting. *Journal of Glaciology*, **35**, 121, 333–342.

Burkard M.B., Schroeter H.O., Whitely H.R. and Donald J.R. (1991) Snow depth/area relationships for various landscape units in southwestern Ontario, *Proceedings of the 48th Annual Eastern Snow Conference*, Guelph, pp. 51–65.

Carroll T.R. (1995) Remote sensing of snow in the cold regions. *First Moderate Resolution Imaging Spectroradiometer (MODIS) Snow and Ice Workshop*, NASA/Goddard Space Flight Center: Greenbelt, pp. 3–14.

Church M. (1974) Hydrology and permafrost with reference to northern North America. *Permafrost Hydrology, Proceedings of Workshop Seminar 1974*, Canadian National Committee for the International Hydrological Decade: Ottawa, pp 7–20.

Cohen J. (1994) Snow cover and climate. *Weather*, **49**(5), 150–156.

Crevier Y. and Duguay C.R. (1995) Albedo of snow in the Kluane range, Yukon Territory: an estimating method using Landsat-5 TM data and terrain modelling, *Second International Workshop on Application of Remote Sensing in Hydrology*, Saskatoon, pp. 159–170 18–20 October 1994.

Croley T.E. II (2002) Large basin runoff model. In *Mathematical Models of Small Watershed Hydrology and Applications*, Singh V.P. and Frevert D.K. (Eds.), Water Resources Publications, LCC, Colorado, pp 717–770.

Donald J.R. (1992) *Snowcover Depletion Curves and Satellite Snowcover Estimates for Snowmelt Runoff Modelling*, PhD Thesis, Civil Engineering, University of Waterloo.

Donald J.R., Soulis E.D., Kouwen N. and Pietroniro A. (1995) A land cover-based snow cover representation for distributed hydrologic models. *Water Resources Research*, **31**(4), 995–1009.

Essery R. (2003) Aggregated and distributed modelling of snowcover for a high-latitude basin. *Global and Planetary Change*, **38**, 115–120.

Essery R.L.H. and Etchevers P. (2004) Parameter sensitivity in simulations of snowmelt. *Journal of Geophysical Research*, **109**, D20111, doi:10.1029/2004JD005036.

Essery R.L.H. and Pomeroy J.W. (2004) Implications of spatial distributions of snow mass and melt energy on snowcover depletion: theoretical considerations. *Annals of Glaciology*, **38**, 261–265.

Essery R. and Yang Z.-L. (2001) An overview of models participating in the Snow Model Intercomparison Project (SNOWMIP), *SnowMIP Workshop, 11 July 2001. 8th Scientific Assembly of IAMAS*, Innsbruck.

Etchevers P., Martin E., Brown R., Fierz C., Lejeune Y., Bazile E., Boone A., Dai Y.-J., Essery R., Fernandez A., Gusev Y., Jordan R., Koren V., Kowalczyk E., Nasonova O., Pyles R.D., Schlosser A., Shmakin A.B., Smirnova T.G., Strasser U., Verseghy D., Yamazaki T. and Yang Z.-L. (2004) Validation of the surface energy budget simulated by several snow models. *Annals of Glaciology*, **38**, 150–158.

Formosov A.N. (1946) Snow cover as an integral factor of the environment and its importance in the ecology of mammals and birds. *Materials for Fauna and Flora of the USSR. New Series, Zoology*, Boreal Institute, University of Alberta: Edmonton, pp. 51–152.

Foster D.J. Jr and Davy R.D. (1988) *Global Snow Depth Climatology*, USAFETAC/TN-88/006, USAF Environment Technical Applications Center: Scott Air Force Base.

Goodison B.E. (1981) Compatibility of Canadian snowfall and snow cover data. *Water Resources Research*, **17**(4), 893–900.

Goodison B.E., Glynn J.E., Harvey K.D. and Slater J.E. (1987) Snow surveying in Canada: a perspective. *Canadian Water Resources Journal*, **12**(2), 27–42.

Goodison B.E., Walker A.E. and Thirkettle F.W. (1990) Determination of snow water equivalent on the Canadian prairies using near real-time passive microwave data. In *Workshop on Applications of Remote Sensing in Hydrology*, Kite G.W. and Wankiewicz A. (Eds.), NHRI: Saskatoon, pp. 297–309.

Gray D.M., Norman D.I. and Dyck G.E. (1970) Measurement of prairie snowpacks, *Proceedings of the 38th Annual Meeting of the Western Snow Conference*, Victoria, April 21–23, 1970.

Gray D.M. and Male D.H. (Eds.) (1981) *Handbook of Snow: Principles, Processes, Management & Use*, Pergamon Press: Toronto.

Groisman P.Y. and Davies D.D. (2001) Snow cover and the climate system. In *Snow Ecology: An Interdisciplinary Examination of Snow-covered Ecosystems*, Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 1–44.

Groisman P.Y., Karl T.R., Knight R.W. and Stenchikov G.L. (1994) Changes of snow cover, temperature, and the radiative heat balance over the Northern Hemisphere. *Journal of Climate*, **7**, 1633–1656.

Hedstrom N.R. and Pomeroy J.W. (1998) Measurements and modelling of snow interception in the boreal forest. *Hydrological Processes*, **12**(10–11), 1611–1625.

Hock R. (2003) Temperature index melt modelling in mountain areas. *Journal of Hydrology*, **282**, 104–115.

Hoham R.W. and Duval B. (2001) Microbial ecology of snow and freshwater ice with emphasis on snow algae. In *Snow Ecology: An Interdisciplinary Examination of Snow-covered Ecosystems*, Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 168–228.

Larson L. (2002) National Weather Service River Forecast System (NWSRFS). In *Mathematical Models of Small Watershed Hydrology and Applications*, Singh V.P. and Frevert D.K. (Eds.), Water Resources Publications, LCC: Colorado, pp. 657–703.

Lemieux G.-H., Brisson C., Bégin D., Bégin R. and Gignac C. (1995) Using NOAA satellites for daily dynamic mapping of spring snowcover for hydroelectric production of lac Saint-Jean watershed, Québec. *Second International Workshop on Application of Remote Sensing in Hydrology*, NHRI: Saskatoon, pp. 171–186, 18–20 October 1994.

Lettenmaier D.P. and Bowling L.C. (Eds.) (2001) *Report of the ACSYS/GEWEX-GLASS PILPS 2e Stage 1 Arctic Hydrological Model Intercomparison Study Workshop*, World Climate Research Programme Informal Report No. 15/2001, World Climate Research Programme.

Liston G.E. (2004) Representing subgrid snow cover heterogeneities in regional and global models. *Journal of Climate*, **17**, 1381–1397.

Liston G.E., Pielke R.A. and Greene E.M. (1999) Improving first-order snow-related deficiencies in a regional climate model. *Journal of Geophysical Research*, **104**, 19559–19567.

Luce C.H. and Tarboton D.G. (2004) The application of depletion curves for parameterization of subgrid variability of snow. *Hydrological Processes*, **18**, 1409–1422.

Luce C.H., Tarboton D.G. and Cooley K.R. (1999) Sub-grid parameterization of snow distribution for an energy and mass balance snow cover model. *Hydrological Processes*, **13**, 1921–1933.

Male D.H. and Granger R.J. (1981) Snow surface energy exchange. *Water Resources Research*, **17**(3), 609–627.

Manabe S. (1969) Climate and the ocean circulation, I. The atmospheric circulation and the hydrology of the Earth's surface. *Monthly Weather Review*, **97**, 739–774.

Marks D. (1988) *Climate, Energy Exchange, and Snowmelt in Emerald Lake Watershed, Sierra Nevada*, PhD Thesis, Departments of Geography and Mechanical Engineering, University of California, Santa Barbara, p. 158.

Marks D. and Dozier J. (1992) Climate and energy exchange at the snow surface in the alpine region of the Sierra Nevada: 2. Snow cover energy balance. *Water Resources Research*, **28**(11), 3043–3054.

Marks D., Dozier J. and Davis R.E. (1992) Climate and energy exchange at the snow surface in the alpine region of the Sierra Nevada: 1. Meteorological measurements and monitoring. *Water Resources Research*, **28**(11), 3029–3042.

Marks D., Kimball J., Tingey D. and Link T. (1998) The sensitivity of snowmelt processes to climate conditions and forest cover during rain-n-snow: A study of the 1996 Pacific Northwest flood. *Hydrological Processes*, **12**(10–11), 1569–1587.

Marsh P. and Woo M.-K. (1981) Snowmelt, glacier melt, and high arctic streamflow regimes. *Canadian Journal of Earth Sciences*, **18**(8), 1380–1384.

Marsh P. and Pomeroy J.W. (1996) Meltwater fluxes at an arctic forest-tundra site. *Hydrological processes*, **10**, 1383–1400.

Marshall S., Roads J.O. and Glatzmaier G. (1994) Snow hydrology in a general circulation model. *Journal of Climate*, **7**, 1251–1269.

Martinec J. (1975) Snowmelt-runoff model for stream flow forecasts. *Nordic Hydrology*, **6**(3), 145–154.

Martinec J. (1989) Hour-to-hour snowmelt rates and lysimeter outflow during an entire ablation period. *Snow Cover and Glacier Variations*, IAHS: *Proceedings of the Baltimore Symposium*, IAHS: Maryland, IAHS Publication No. 183, pp. 19–28, May 1989.

Martinec J. and de Quervain M.R. (1975) The effect of snow displacement by avalanches on snowmelt and runoff, *Proceedings on Snow and Ice Symposium*, IAHS: Moscow, IAHS Publication No. 104, pp. 364–377.

Martinec J. and Rango A. (1986) Parameter values for snowmelt runoff modeling. *Journal of Hydrology*, **84**, 197–219.

Martinec J. and Rango A. (1997) The Snowmelt Runoff Model (SRM) User's Manual - Section 3, **http://dude.uibk.ac.**

at/Projects/HydAlp/demonstrator/hydalp/external dispframe, **html?http://hydrolab.arsusda.gov/cgi-bin/srmhome**. Accessed November 15, 2004.

McKay G.A. and Gray D.M. (1981) The distribution of snowcover. In *Handbook of Snow - Principles, Processes, Management & Use*, Gray D.M. and Male D.H. (Eds.), Pergamon Press: Oxford, pp. 153–190.

Metcalfe R.A. and Buttle J.M. (1998) Statistical model of spatially distributed snowmelt rates in a boreal forest basin. *Hydrological Processes*, **12**(10–11), 1701–1722.

Morris E.M. (1982) Sensitivity of the European Hydrological System snow models. In *Hydrological Aspects of Alpine and High-Mountain Areas*, Glen J.W. (Ed.), International Association of Hydrological Sciences: Wallingford, IAHS-AIHS Publication 138, pp. 221–231.

National Aeronautics and Space Administration (2004) **http://modis-snow-ice.gsfc.nasa. gov/intro.html**, accessed November 15, 2004.

National Operational Hydrologic Remote Sensing Center (2004) **http://www.nohrsc. nws.gov**, accessed November 15, 2004.

National Snow and Ice Data Centre (2004) **http://nsidc.org/data/ae_dysno.html**, accessed November, 15, 2004.

NRC (1989a) Chapter 7 - Snowmelt contributions. *Hydrology of Floods in Canada, a Guide to Planning and Design*, National Research Council Canada - Associate Committee on Hydrology: NRCC no. 29734, pp. 95–110.

NRC (1989b) Chapter 10 - Ice jam floods. *Hydrology of Floods in Canada, a Guide to Planning and Design*, National Research Council Canada - Associate Committee on Hydrology: NRCC no. 29734, pp. 169–184.

Pietroniro A., Prowse T., Hamlin L., Kouwen N. and Soulis R. (1996) Application of a Grouped Response unit Hydrological Model to a Northern Wetland Region. *Hydrological Processes*, **10**, 1245–261.

Pietroniro A. and Soulis R. (2003) A hydrology modelling framework for the Mackenzie GEWEX programme. *Hydrological Processes*, **17**, 673–676.

Pomeroy J.W. and Brun E. (2001) Physical Properties of Snow. In *Snow Ecology: An Interdisciplinary Examination of Snow-covered Ecosystems*, Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 45–126.

Pomeroy J.W. and Gray D.M. (1995) *Snowcover Accumulation, Relocation and Management*, National Hydrology Research Institute Science Report No. 7, NHRI, Environment Canada, Saskatoon, p. 144.

Prowse T.D. (1990) Northern hydrology: an overview. *Northern Hydrology - Canadian Perspectives*, Prowse, T.D. and Ommanney C.S.L. (Eds.), NHRI: NHRI Science Report No. 1, pp 1–36.

Pruitt W.O. (1984) Snow and living things. In *Northern Ecology and Resource Management*, Olson R. (Ed.), University of Alberta Press: Edmonton, pp. 51–77.

Rango A. (1996) Spaceborne remote sensing for snow hydrology applications. *Hydrological Sciences Journal*, **41**(4), 477–494.

Rango A. and Martinec J. (1995) Revisiting the degree-day method for snowmelt computations. *Water Resources Bulletin*, **31**(4), 657–669.

Rango A., Martinec J., Foster J. and Marks D. (1983) Resolution in operational remote sensing of snow cover. *Hydrological Applications of Remote Sensing and Remote Data Transmission*, *Proceedings of the Hamburg Symposium*, IAHS: Hamburg, IAHS Publication no. 145, August 1983.

Richter G.D. (1954) *Snow Cover, Its Formation and Properties*, U.S. Army Cold Regions Research and Engineering Laboratory: Transl. 6, NTIS AD 045950, Hanover.

Schroeter H.O. (1988) *An Operational Snow Accumulation Ablation Model for Areal Distribution of Shallow Ephemeral Snowpacks*, PhD Thesis, University of Guelph.

Seglenieks F.R. (1994) *Application of Remote Sensing and Ground Measurements to Calibrate the Hydrological Model WATFLOOD*, M.A.Sc. Thesis, Department of Civil Engineering, University of Waterloo, Waterloo, p. 162.

Shi J.C., Dozier J. and Rott H. (1994) Snow mapping in alpine regions with synthetic aperture radar. *IEEE Transactions on Geoscience and Remote Sensing*, **31**(1), 152–158.

Singh V.P. (1995) *Computer Models of Watershed Hydrology*, Water Resource Publications: pp. 367–394.

Slater A.G., Schlosser C.A., Desborough C.E., Pitman A.J., Henderson-Sellers A., Robock A., Vinnikov K.Y., Mitchell K., Boone A., Braden H., *et al.* (2001) The representation of snow in land surface schemes: results from PILPS 2(d). *Journal of Hydrometeorology*, **2**, 7–25.

Stoddart R.B.L. and Watt W.E. (1970) *Flood Frequency Prediction for Intermediate Drainage Basins in Southern Ontario*, Research Report No. 66, Department of Civil Engineering, Queen's University, Kingston, July 1970.

Storck P., Lettenmaier D.P. and Bolton S. (2002) Measurement of snow interception and canopy effects on snow accumulation and melt in mountainous maritime climate, Oregon, US. *Water Resources Research*, **38**(11), 1223–1238, doi:10.1029/2002WR001281.

Sturm M., Holmgren J. and Liston G.E. (1995) A seasonal snow cover classification system for local to global applications. *Journal of Climate*, **8**, 1261–1283.

Susong D., Marks D. and Garen D. (1999) Methods for developing time-series climate surfaces to drive topographically distributed energy- and water-balance models. *Hydrological Processes*, **13**, 2003–2021.

Verseghy D.L. (2000) The Canadian land surface scheme (class): its history and future. *Atmosphere-Ocean*, **38**(1), 1–13.

Walker D.A., Billings W.D. and de Molenaar J.G. (2001) Snow-Vegetation Interactions in Tundra Environments. In *Snow Ecology: An Interdisciplinary Examination of Snow-covered Ecosystems*, Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 266–324.

Walsh J.E. (1987) Large-scale effects of seasonal snow cover. *Large Scale Effects of Seasonal Snow Cover*, IAHS: Vancouver, IAHS Publication no. 166, pp. 3–14.

Willis W.O. (1978) Snow on the Great Plains. *Proceedings on the Modeling of Snow Cover Runoff*, U.S. CRREL: Hanover, pp. 56–62.

WMO (1986) *Intercomparison of Models of Snowmelt Runoff*, Publication no. 646, World Meteorological Organization: Geneva.

WMO (1994a) *Guide to Hydrological Practices*, *Fifth Edition*, Publication no. 168, World Meteorological Organization: Geneva.

WMO (1994b) Chapter 8: Snow Cover *Guide to Hydrological Practices*, *Fifth Edition*, Publication no. 168, World Meteorological Organization: Geneva.

WMO (1994c) Chapter 31: Snow-Melt Runoff Analysis. *Guide to Hydrological Practices*, *Fifth Edition*, Publication no. 168, World Meteorological Organization: Geneva.

WMO (1994d) Chapter 45: Snow-Melt Forecasts. *Guide to Hydrological Practices*, *Fifth Edition*, Publication no. 168, World Meteorological Organization: Geneva.

163: Hydrochemical Processes in Snow-covered Basins

JOHN W POMEROY[1], H GERALD JONES[2], MARTYN TRANTER[3] AND GRO LILBÆK[1]

[1] *Centre for Hydrology, Department of Geography, University of Saskatchewan, Saskatoon, SK, Canada*
[2] *INRS-ETE, Environnement, Université du Quebec, Sainte-Foy, QC, Canada*
[3] *Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol, UK*

This article reviews several aspects of snow hydrochemistry: the chemistry of snowfall including chemical incorporation in snowfall and snowfall chemistry variability, the chemistry of cold, dry snowcovers including snow redistribution, snow–atmosphere chemical exchange and in-pack chemical transformations, the chemistry of wet and melting snowcovers including solute leaching, particulate interactions and microbial activity, and snow-covered basin hydrochemistry with an emphasis on nutrient chemistry. The emphasis is on the processes of chemical transformation in seasonal snowpacks and meltwaters with strong attention to the broad ecosystem view of snow chemistry rather than solely focusing on acidification effects from snowmelt. The seasonal snowcover is shown to be a dynamic hydrochemical system with strong ecological interactions. Besides wet deposition by snowfall and rain, the processes of wind redistribution, dry deposition, volatilization, crystal metamorphism, photolysis, microbial uptake and release, solute elution, and meltwater movement strongly affect the chemistry of both the snowpack and meltwaters. Snowmelt chemistry alone is rarely directly responsible for major chemical fluctuations in water bodies, but meltwater has an important role in transporting ions from soils and organic material to water bodies.

INTRODUCTION

The seasonal snowcover plays a unique role in basin hydrochemistry by collecting and transforming chemicals over the snow accumulation season and then releasing them suddenly during melt. Chemicals in the snowcover are derived from dry and wet deposition and are preferentially flushed out, early in the melt season. When industrial pollutants are scavenged and deposited to snowcover, the resulting meltwaters provide a low pH contribution to aquatic and soil chemistry (Hendershot *et al.*, 1992; Galloway *et al.*, 1987). Tranter and Jones (2001) organized the processes, which influence the chemical composition of snow and meltwaters, into those involving

1. heat and mass fluxes that occur during sublimation and melting; and
2. chemical transformations.

During sublimation and melting, chemical species are considered conservative (excluding migration) in that they are not transformed. The physical properties of chemical species (e.g. solubility, vapor pressure) are therefore important in how these processes alter the chemical load of, and release from, the snow cover (Brimblecombe and Shooter, 1991). Chemical transformations occur by chemical reactions such as oxidation (Bales *et al.*, 1987) and photolysis (Beine *et al.*, 2002), or from microbiological activity (Hoham *et al.*, 1989). This review will consider hydrochemical processes in snow from snowfall formation to infiltration and runoff.

The Chemistry of Snowfall

Wet deposition to snowcovers primarily occurs through snowfall. The reader is referred to sections on rainfall chemistry for a discussion of the chemical inputs that can derive from rain-on-snow events (Stumm and Morgan, 1996).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Chemical Incorporation in Solid Precipitation

Solid precipitation particle formation and fall through the atmosphere incorporate atmospheric chemicals via three main processes:

1. imprisonment during the initial formation of ice crystals;
2. capture of gases, aerosol, and larger particulates within clouds; and
3. scavenging of these materials below the cloud layers during snowfall (Barrie, 1991).

Cloud water droplets contain solute as a result of aerosol scavenging, and the diffusion of atmospheric gases into solution. The soluble species are mainly NH_4, SO_4, NO_3, Ca, K, and Mg, derived from natural and anthropogenic emissions, in addition to Na and Cl from sea-salt aerosol. In addition to solute, droplets contain sea-salt aerosols, particulate organic debris, and/or fine particulate clays (Kamai, 1976) that can form ice nuclei for freezing. The chemical content of the center of falling snow grains is often characterized by these compounds. During the droplet freezing process, however, most solutes are rejected to the outer edges of the crystal as they do not fit well into the ice crystal lattice (Colbeck, 1987).

Ice crystals then grow by vapor transfer from supercooled water droplets because of small differences in saturation vapor pressure over ice and water. Very little solute is transported by this process to the growing ice crystals (Hewitt and Cragin, 1994), however, small amounts of HNO_3 and HCl are absorbed. Direct collision between growing crystals and supercooled droplets results in rimed crystals and graupel, which contain relatively high solute concentrations compared to crystals formed mainly by vapor transfer (Cerling and Alexander, 1987).

Further scavenging of aerosols and other particulates occurs below clouds by adsorption and impact; snowfall more efficiently scavenges particulates than does rain because of its higher surface area to mass ratio, and lower terminal fall velocity (Raynor and Haynes, 1983; Nicholson *et al.*, 1991).

Relationships between crystal form and chemistry have been found because of the association between genesis and form, and the strong differences in chemical incorporation during various genetic processes. Lamb *et al.* (1986) found that the highest concentrations of solute were associated with the smallest unrimed crystals. Borys *et al.* (1983) showed the solute content of snow crystals to increase with the degree of riming. Hewitt and Cragin (1994) found that dendritic and stellar plates had similar chemical content for most species, except for Cl, which showed much higher concentrations in stellar plates than dendritic crystals.

Variability of Snowfall Composition

Snowfall chemical composition depends on factors such as the air mass origin, altitude, and meteorological conditions (Colin *et al.*, 1989; Davies *et al.*, 1992). Maritime air masses will give rise to snow containing mostly Na and Cl (Tranter *et al.*, 1986), while polluted air masses from industrial areas will deposit snow that is highly acidic because of the presence of strong acid anions (NO_3, SO_4) from fossil fuel combustion (Davies *et al.*, 1984; Landsberger *et al.*, 1989).

The spatial variability of snowfall chemistry shows length scales from meters to hundreds of kilometers (Tranter *et al.*, 1987; Pomeroy *et al.*, 1995; Turk *et al.*, 2001; de Caritat *et al.*, 2005), reflecting factors such as the proximity to pollution or ocean sources, the back trajectory of the air masses associated with the snowfall, and wind mixing of different snowfalls on the ground. Generally, snowfall at high altitudes contains lower chemical concentrations than at lower altitudes because of the shorter air column for the falling snow crystals to scavenge from. In addition, chemical concentrations during a snowfall event often decrease exponentially with time, as the store of species available for scavenging in the atmosphere depletes with cumulative scavenging. As a result, persistent storm tracks (e.g. lake effect, coastal) have characteristic spatial trends in snowfall chemistry. For instance, as air masses rise up the windward side of mountains, they generate heavy orographic snowfall. This snowfall progressively depletes aerosols over the increasing distance, resulting in more dilute snowfall chemistry with increasing altitude (Lyons *et al.*, 1991). Postdepositional changes and changes due to redistribution of snow are dealt with in the next section.

Chemistry of Cold, Dry Snowcover

Snowcovers at temperatures below 0 °C are defined as "cold", and because of low liquid water contents are termed "dry". The temporal and spatial variation in the chemical composition of snowfall and subsequent redistribution processes usually produce a snowcover that is chemically heterogeneous. The main processes of transformation (see Figure 1) are

1. redistribution by wind and vegetation;
2. surface-exchange at the snow-atmosphere interface (dry deposition and volatilization);
3. surface and subsurface chemical reactions;
4. snow-grain metamorphism within the pack; and
5. basal-exchange processes at the snow-soil interface (gaseous emissions from soil).

Snow Redistribution Processes

Snow is redistributed by wind via blowing snow transport (Pomeroy *et al.*, 1991), and by vegetation via snow

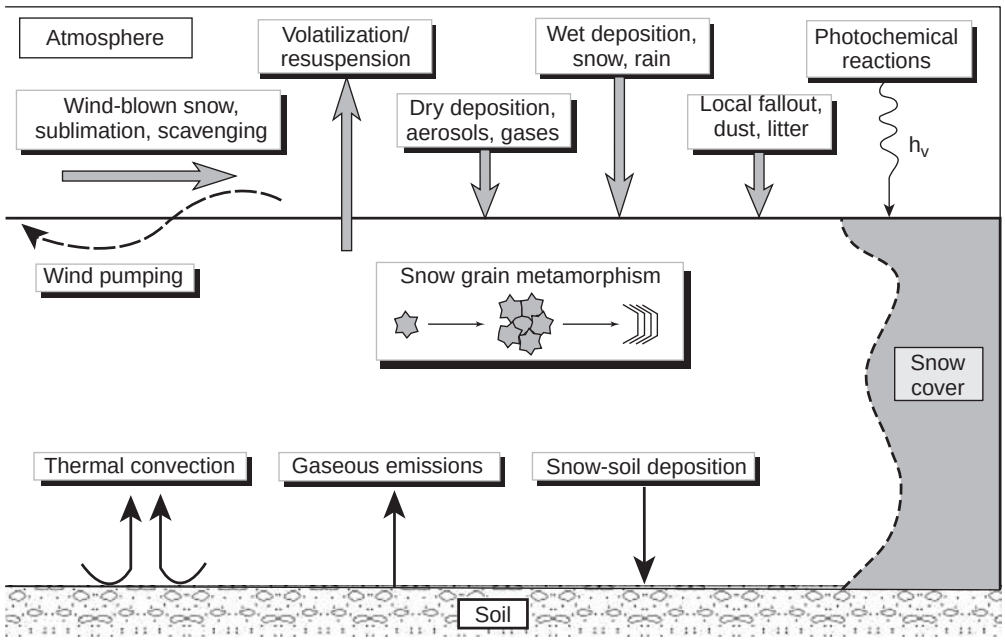


Figure 1 The main physical and chemical processes that influence the chemical composition of cold, dry snow cover during the accumulation season (After Tranter and Jones, 2001)

interception (Pomeroy *et al.*, 1999). Wind transport has the potential to change the chemical composition of snow due to three main physical processes, namely, sublimation of water vapor, scavenging of aerosols and gases from the atmosphere, and volatilization (Pomeroy and Jones, 1996). Direct wind redistribution of snow chemicals can move chemical species between basins, and both transport and sublimation increase the spatial variability of snow chemistry. In the Cairngorm Mountains of Scotland, Tranter *et al.* (1987) found that snowfall chemical concentrations had coefficients of variation (CV) from 0.03 to 0.04. However, CVs for ion concentration in wind redistributed snowcovers at this site varied from 0.1 to 0.74 (Pomeroy *et al.*, 2000). At the arctic treeline in NW Canada, Pomeroy *et al.* (1995) found that blowing snow redistribution was associated with a fivefold difference in chemical loading in snow within a 70 km^2 basin; this greatly exceeded the differences due to ion concentration for which the CVs ranged from only 0.06 to 0.12.

Interception by evergreen canopies can store over half the cumulative snowfall in midwinter (Pomeroy and Gray, 1995). Where chemical species are conserved, redistribution from trees occurs at scales of only a few meters and is generally unimportant. Dry deposition and volatilization are also affected by snow interception, and are discussed in the following section. Sublimation of intercepted snow increases the concentration of conserved ionic species up to sixfold, according to the loss of ice from intercepted snow clumps (Pomeroy *et al.*, 1999).

Processes at the Atmosphere-snow Surface Interface

Dry Deposition

Dry deposition is the direct deposition of chemical species from the atmosphere to the snow surface (Cadle, 1991). Aerosols and particulates may be directly deposited, while gaseous species may be adsorbed (Conklin, 1991). Because the aerodynamic surface roughness of snow is low (Cadle *et al.*, 1985), and liquid water layers are small to nonexistent in cold snow (Choi *et al.*, 2000), dry deposition to snow cover is much lower than to surfaces without snow or to forest canopies, and is small in relation to wet deposition from snowfall.

Bales *et al.* (1987) found that dry deposition rates to new snow were higher than those to old snow as a result of a reduction in the area of crystal surfaces during the metamorphism of snow. The importance of crystal form in dry deposition has also been reported by Ibrahim *et al.* (1983), who suggested that the interception of aerosols by ice needles in relatively fresh snow contributed significantly to the measured rates of dry deposition. Albert and Shultz (2002) show the effect of within-snowpack ventilation via diffusion and advection in enhancing dry deposition to snow.

Cadle *et al.* (1985) estimated deposition rates for HNO_3 to be one order of magnitude larger than those for SO_2. The difference was attributed to the relative solubility of the two gases in the liquid layers around the crystals, and also to other factors such as relative diffusion rates into

the ice lattice, and the rate of oxidation of SO_2 to SO_4 at the air–crystal interface (Bales *et al.*, 1987; Bales, 1991). The rate of deposition of HNO_3 is much higher to wet snow than to cold, dry snow (Cadle, 1991). Field studies in North America suggest that dry deposition contributes approximately 20–25% of the chemical composition of SO_4 and NO_3 in snowcovers (Cadle and Dasch, 1987; Barrie and Vet, 1984; Cadle, 1991).

Snow covers under forest canopies may receive greater contributions of dry deposition. Pomeroy *et al.* (1999) showed greater dry deposition of aerosols to intercepted snow, and subsequent unloaded snow in forest canopies. Dry deposition of gases and aerosols to forests is higher than to open snowfields because of enhanced aerodynamic roughness and absorptive needle/bark surfaces (Höfken *et al.*, 1983; Dasch, 1987). Species deposited to the canopy are subsequently redistributed to the snowcover via intercepted snow unloading, or rainfall interception and drip.

Aeolian dust and other particulate matter are continuously being deposited on snow covers. The result of dust deposition is usually a reduction in the acidity of the snow (Sequeira, 1991), particularly during melt periods (Clow and Ingersoll, 1994; Delmas *et al.*, 1996). Pomeroy *et al.* (1991) found enhancements of aerosol concentration in postdepositional and in-transit wind-blown snow during periods in which blowing snow particles developed strong electrical charges, and suggested that electrophoresis could attract small aerosols to blowing snow particles before deposition. The net effect of changes in aerodynamic roughness on dry deposition of gaseous species and distance from source of aerosols can be difficult to distinguish. An example of loadings of ions to basins with and without trees is shown in Figure 2.

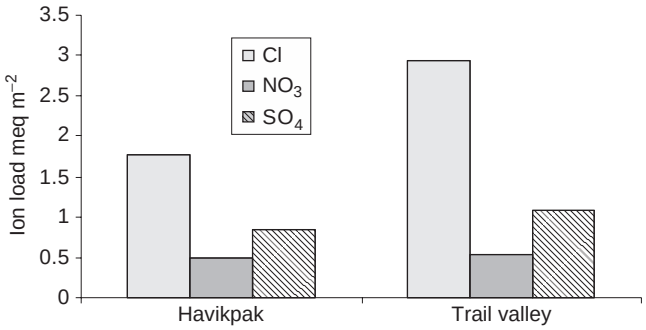


Figure 2 Ion loadings in snow to two basins in NW Canada; Havikpak Creek is sparsely forested and further away from the Arctic Ocean, Trail Valley Creek is dominated by tundra and 50 km closer to the sea. Differences in dry deposition of SO_4 and Cl are evident and because of the increased roughness of the Havikpak forests, and to the different availability of aerosol from the ocean (After Pomeroy *et al.*, 1995)

Snow will also accumulate nutrients by the deposition of biological debris, either as fallout from above the snow surface, or by direct incorporation into the snowpack. In forests, much of the deposition arises as litterfall from the canopy (Jones and Debois, 1987) and mammalian excrement (Jones, 1991). Invertebrate fallout from wind-borne arthropods and winged invertebrates may also contribute nutrients to snow, while vertebrate and invertebrate activity under and within the snow cover transforms and redistributes nutrients from the soil, upwards into the snow matrix. Jones (1991) attempted to calculate nutrient inputs to snow from different vertebrates by considering the spatial distribution of population densities, and the amount and chemical composition of animal excrement. Persistent spatial distributions of excrement were important to basin-scale estimates of N deposition. In the case of animals that herd (e.g. deer), deposition rates may be up to 200 g N ha^{-1} day^{-1} in areas of congregation. Solitary mammals, such as moose and hares, deposit only up to 2.5 g N ha^{-1} day^{-1}.

Volatilization

The dry deposition of aerosols to snow is often considered to be irreversible. However, some species may volatilize and be lost back into the atmosphere. Postdepositional loss of NO_3 has been observed in surface snow (Neubauer and Heumann, 1988; Beine *et al.*, 2002), and sublimating intercepted and blowing snow (Pomeroy *et al.*, 1991; Pomeroy and Jones, 1996; Pomeroy *et al.*, 1999). Volatilization of NO_3 in redistributed snow was proportional to ice sublimation losses over the winter, which is on the order of 35% in the boreal forest, 20% in arctic tundra, and 20–40% in the northern steppes. Figure 3 shows ion concentration enrichment as a function of sublimation loss of snow mass for NO_3, SO_4, and Cl in a boreal forest. SO_4 and Cl are conserved, as snow mass is decreased by sublimation; the loss of NO_3 is roughly proportional to the loss of snow mass, suggesting an association between volatilization and sublimation. These results are consistent with the observations of Stottlemyer and Troendle (1999) who observed increases in NO_3 in snowpacks and meltwater runoff, when evergreen forest basins were clear-cut in the Rocky Mountains, Colorado.

Oxidation and Photochemical Reactions in Surface Snowcover

Oxidation of certain species by atmospheric oxidants may take place on cold snow-grain surfaces, if a liquid film is present (Conklin and Bales, 1993). Bales (1991) has modeled the chemical oxidation of S(IV), SO_2, to S(VI), SO_3/SO_4, on the basis of the known oxidation rates by H_2O_2, O_3, and O_2. Whilst of great interest to ice core interpretations, this also has relevance to understanding S uptake in snow-covered catchments.

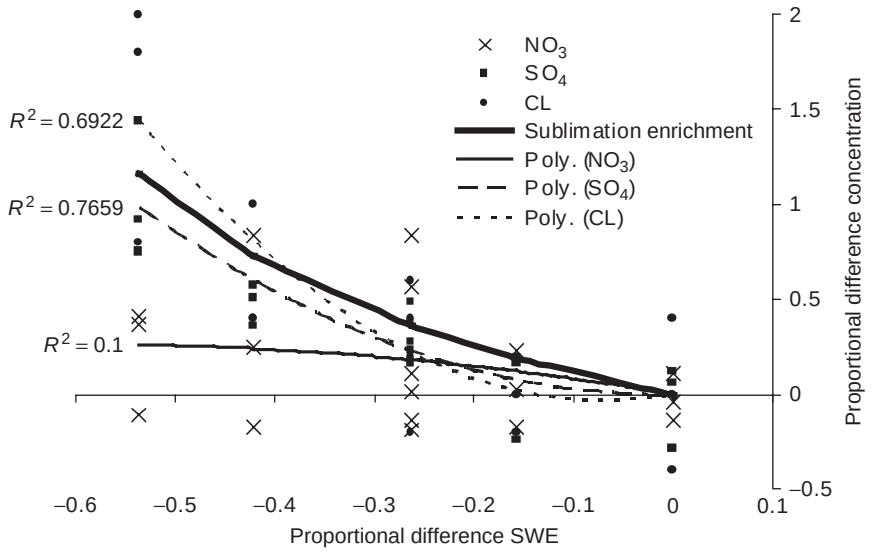


Figure 3 Enrichment of ion concentration because of sublimation of intercepted snow in the boreal forest. SO_4 and Cl follow the expected enrichment assuming conservation of ion, whilst NO_3 is depleted during sublimation. Snow mass is shown as snow water equivalent (SWE), and both SWE and ion concentration are referenced to levels that in a clearing experienced neither melt nor snow redistribution (After Pomeroy *et al.*, 1999)

Sigg *et al.* (1987) measured the decrease in concentrations of H_2O_2 in surface alpine snows, and proposed that photolysis was the primary mechanism. Neubauer and Heumann (1988) suggested that the apparent loss of NO_3 from Antarctic snow was due either to the photodegradation of HNO_3 to NO_2 by solar radiation, and/or to the volatilization of HNO_3 from snow during metamorphism, but were unable to distinguish between the two mechanisms. Beine *et al.* (2002) found that photolysis on a high Arctic snowpack caused HNO_3 transport to the snow surface.

In-pack Processes

Metamorphism

The processes of snow metamorphism are described in **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**. It is currently accepted that the solute becomes redistributed to, and concentrated on, the snow-grain surfaces or in snow particle bonds during dry snow metamorphism, although direct observation of this process has not yet been made. The net effect of weak temperature gradient metamorphism is to concentrate the solute onto or near the surfaces of ice crystals (Bales, 1991; Colbeck, 1987; Davis, 1991). The solute may be located in a "quasi-liquid" surface layer, as discrete aerosol or as concentrated, "doped ice" pockets (Davis, 1991). Strong temperature or kinetic metamorphism is also accompanied by loss of ions. Ion losses are likely due to some form of transport, either to the base of the pack or to adjacent snow strata and the atmosphere. Laberge and Jones (1991) found that SO_4 was lost during depth hoar formation. In contrast, Pomeroy *et al.* (1993) found that SO_4 and Cl concentrations in depth hoar increased in proportion to the overall loss of water vapor from the depth hoar to adjacent snow layers. However, NO_3 concentrations remained approximately constant, indicating a concomitant loss of the species.

Chemistry of Wet Snow and Snow-meltwater Systems

The percolation of meltwaters through the snowcover (*see* **Chapter 161, Water Flow Through Snow and Firn, Volume 4**) causes the chemical composition of both the snow matrix and the meltwaters to change. The concentration and distribution of solutes in the snow-meltwater system is controlled by a wide variety of physical and biological processes (see Figure 4). These processes include

1. solute leaching from snow grains;
2. meltwater-particulate interactions; and
3. microbiological activity.

In addition, snow-atmosphere exchange is another factor as dry deposition rates of certain species (e.g. SO_2, HNO_3, HCl) to wet snow crystals increase significantly because of their solubility in water (Cadle, 1991). Rain on snow will also influence the chemistry of meltwaters due to its own chemical composition (Tranter *et al.*, 1992).

Solute Leaching

Fractionation of solute species between snow grains and meltwater occurs because of leaching of the melting snow

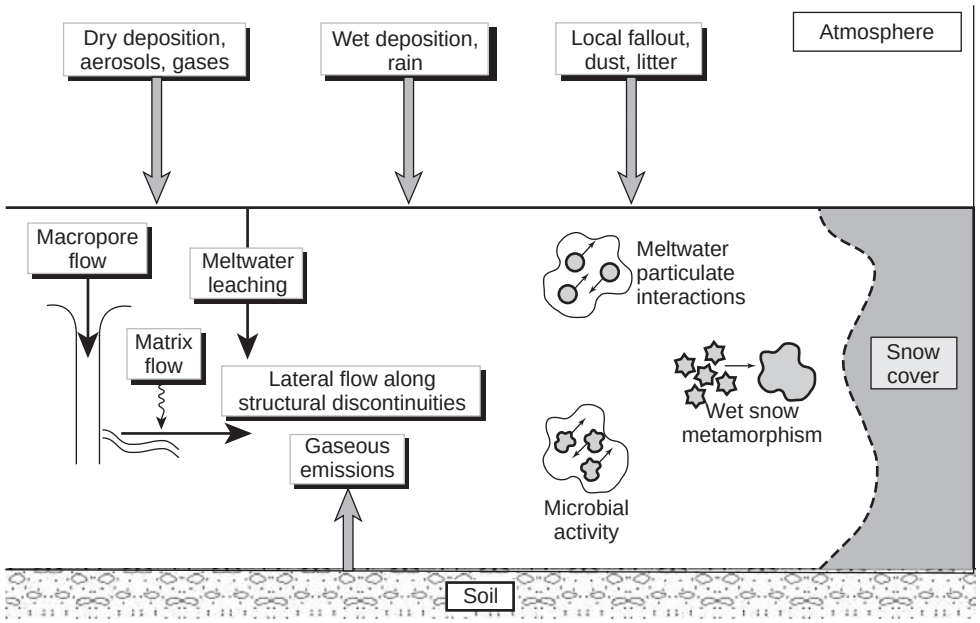


Figure 4 The main physical, chemical, and biological processes that influence the chemical composition of snow cover during melt (After Tranter and Jones, 2001)

grains. The result is that the meltwater front becomes progressively more concentrated as it moves through the pack (Johannessen and Henriksen, 1978; Colbeck, 1981). The degree of fractionation of any solute species, x, between snow and meltwater can be described by a nondimensional "concentration" factor, CF,

$$\mathrm{CF} = \frac{C_\mathrm{m}}{C_\mathrm{p}} \qquad (1)$$

where C_m is the ion concentration in any meltwater fraction, and C_p is the concentration in the parent snow prior to melt. Values of CF during the initial stages of meltwater discharge may range from 1 to 50, but a more typical range is 2 to 7 (Tranter, 1991). CF decreases with increasing melt and cumulative leaching to values of <0.1 in the final meltwaters. The efficiency of meltwater leaching (i.e. higher values of CF per volume of initial meltwater discharge) depends on the distribution of solute in snow grains, snowpack, and meltwater flow paths.

Wet snow metamorphism in the melt period is more rapid than dry snow metamorphism. It causes large grains to grow at the expense of small grains and solute to diffuse from grain boundaries into meltwater (Davis, 1991). The solute scavenging that results in fractionation is related to the rate of melt (Tsiouris *et al.*, 1985; Tranter *et al.*, 1988b) because it is affected by diffusion rates and the duration of snow-meltwater interaction. For example, Marsh and Webb (1979) reported the approximate doubling of initial snowmelt concentrations with the doubling of snow depth. Conversely, at high melt rates, solute scavenging is minimized. Diurnal changes in melt rate affect the CF of meltwaters, with the highest concentrations being found in the morning and evening, or during periods of shading, when melt rates are lowest.

Because of solute scavenging, meltwater flowing rapidly through snow macropores or "flow fingers" is more dilute than melt flowing through the snow matrix. The effects of heterogeneous flow on the spatial variability of fractionation are illustrated in Figure 5, where the CF is shown for two flow paths, one with the lowest (matrix) and one with the highest (macropore) measured flow rate (Marsh and Pomeroy, 1999). The CFs of both flow paths gradually converged over time, until all flow paths had similar values.

The mesoscale distribution of solute in snow cover will also affect the concentration of meltwaters. Discrete snowfalls or redistribution events cause snow strata to have differing composition. Solute-rich bands can arise from the exclusion of solute from ice lenses formed by the refreezing of meltwater, or rainwater in cold snow. The result of several diurnal melt-freeze cycles is often to increase the ionic concentrations in the first meltwaters issuing from the snowpack (Bales *et al.*, 1989; Williams and Melack, 1993). Both laboratory and field experiments have shown that solute-rich layers give rise to more concentrated meltwaters (Colbeck, 1981; Tranter *et al.*, 1986; Marsh and Pomeroy, 1999).

Modeling solute leaching and meltwater composition is extremely difficult because of uncertain processes, variable location of solutes in snow, and complex meltwater

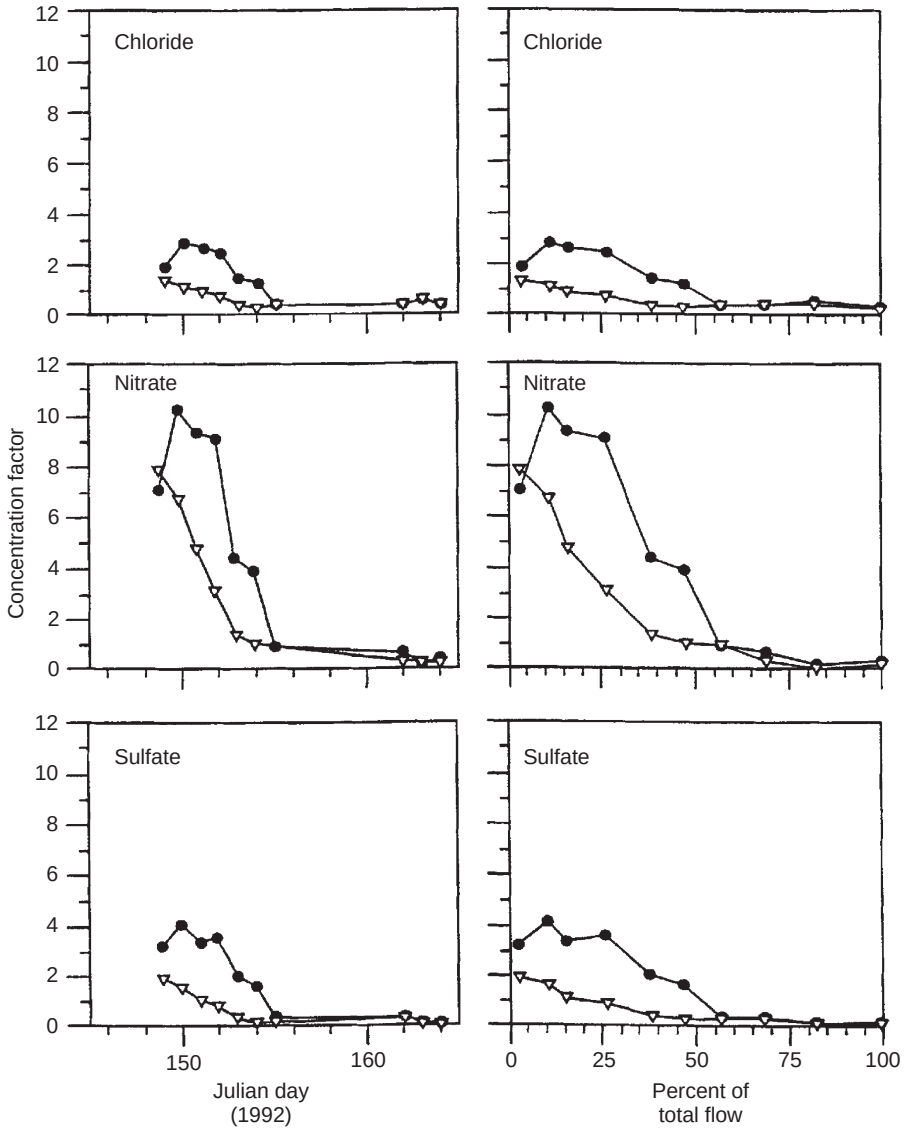


Figure 5 The impact of flow rate on the concentration factor of SO42-, SO3- and Cl- in snowmelt. The data is collected from a multiple compartment lysimeter under a tundra snowpack that samples both high and low meltwater flow zones of the melting snowpack. Open symbols and closed symbols denote the concentration factors from high and low flow chambers respectively

dynamics. Many early elution models were based on bulk transport, which in turn was based on snow depth and a bulk leaching coefficient (Stein *et al.*, 1986). The leaching coefficients were based on first-order removal of the solute from snow by meltwaters in advection-dispersion calculations (Hibberd, 1984). Recent models include metamorphism, preferential flow paths, and the solution of ions in snow, and are being used to explore the spatial and temporal distributions of water and solute flux (Iida *et al.*, 2000; Stagnitti *et al.*, 1999). Results suggest that solute should be partitioned into mobile and immobile fractions, whose interaction controls solute release into the meltwater flow (Feng *et al.*, 2001).

Snowmelt-particulate Interactions

Chemical reactions between meltwater and inorganic/organic particles can affect the concentration of solute in meltwater. Many studies have observed the neutralization of snow acidity by carbonaceous dusts from a variety of sources of either local (Colin *et al.*, 1987) or remote origin (Loye-Pilot *et al.*, 1986). Delmas *et al.* (1996) determined that the rate of chemical weathering of dusts in meltwaters depended on the location of the dust in the snow cover; dust in the lower strata of snow showed the highest rates of weathering due to increased partial pressures of CO_2 that arise during dust-meltwater interaction (see Figure 6).

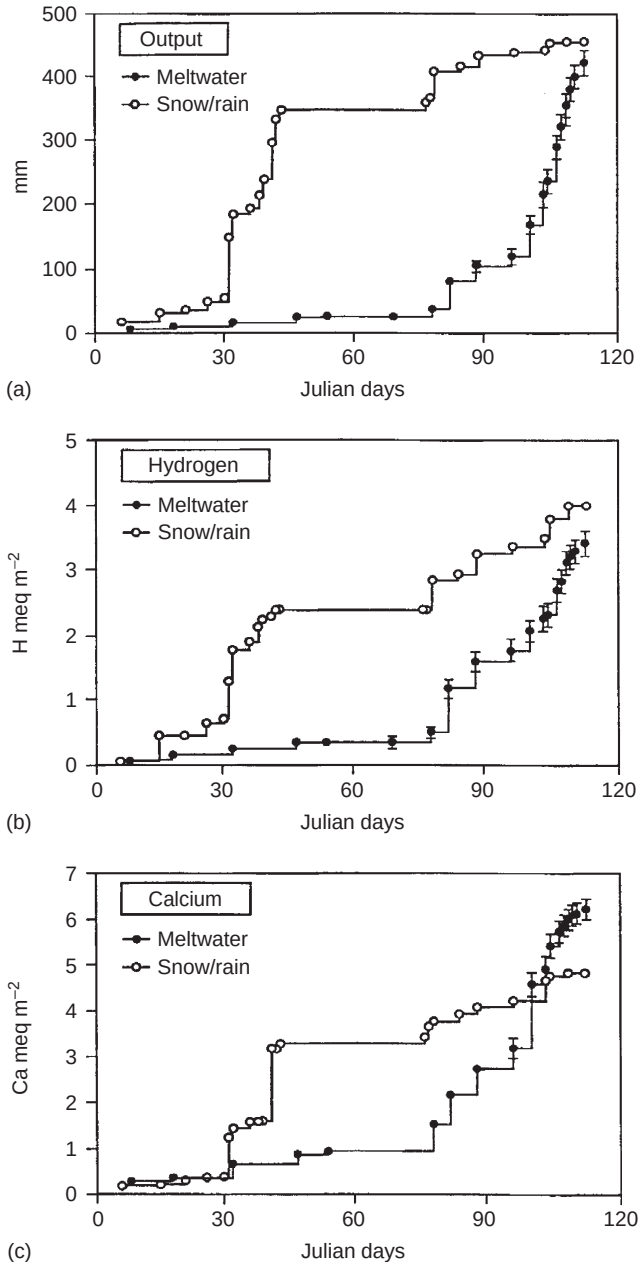


Figure 6 The neutralization of meltwater by calcareous dust in the French Alps. (a) The cumulative input of snow water equivalents and rain to the snowcover, and the cumulative output of meltwater. The water balance is approximately equal. (b) The cumulative input of H^+ to the snow cover, versus the cumulative output. There is a net loss of H^+ during the thaw. (c) The cumulative input of Ca^{2+} to the snow cover, versus the cumulative output. There is a net increase in Ca^{2+} during the thaw (After Delmas *et al.*, 1996)

These lower strata dust particles were, thus, the most efficient at neutralizing acidic meltwaters.

The leaching of litter in snow by meltwaters removes soluble organics and other chemical species (Jones and Sochanska, 1985; Stottlemyer, 1987). Surficial ionic exchange may also take place between meltwaters and the organic debris (Cronan and Reiner, 1983). Leaching experiments (Courchesne and Hendershot, 1988) show that large amounts of PO_4, K, Mn, Ca, and Mg are discharged from litter-laden snow covers, and a decrease in the acidity of meltwaters may arise from cation exchange.

Microbial Activity

During spring melt, the presence of liquid water and the increase in solar radiation stimulate microbiological and invertebrate activity in snow cover. Jones and Debois (1987) showed that meltwater production increased microbiological activity on canopy fallout in forest snowcover. The presence of meltwater also results in photosynthetic activity of truly motile algal populations within the snow cover (Hoham, 1987). Photosynthesis results in an increase in algal biomass at the expense of nutrient concentrations in the meltwaters. Decreases in the concentrations of NH_4 and NO_3 are particularly noted during the growth of algal populations, and may be of the order of 0.67 eq[N] ha^{-1} day^{-1} and 1.05 eq[N] ha^{-1} day^{-1} respectively (Jones, 1999). The loss of nutrients in snow meltwaters over the whole melt season may be appreciable, approaching 20–30% in some years (Jones, 1991).

Snow Nutrient Fluxes and Basin Budgets

Direct acidification from snowmelt remains a concern in many catchments with poor buffering capacity and proximity to certain industrial sources (Tranter *et al.*, 1988a). It is now realized, however, that many episodic stream and lake acidifications that had been associated with the melt of seasonal snowcover were, in fact, due to the mobilization of soil water that carried high ionic loads during the snowmelt period (e.g. Peters and Driscoll, 1987). The role of low ionic strength snowmelt waters is apparently to mobilize geochemical transport from soils in the basin (Hendershot *et al.*, 1992). The exception is where saturated frozen ground with macropores permits runoff with minimal soil interactions (e.g. Jones and Pomeroy, 2001; Quinton and Pomeroy, 2005). The direct delivery of nutrients from snow is now recognized as the primary snow-derived geochemical impact on basin hydro-ecology (Tranter and Jones, 2001). In basins that sustain a long snow-covered period, the major input of N and S to soils and water bodies arises in snow meltwaters in spring. The major output may also occur during the same period when export from the basin by streams is the greatest because of the meltwater runoff (Brooks and Williams, 1999; Brooks *et al.*, 1999). Where basins are poorly buffered, there is a dramatic drop in pH during snowmelt because of acids released directly from snow, and/or mobilized from soil. In either case, runoff during snowmelt has been linked to severe stream and

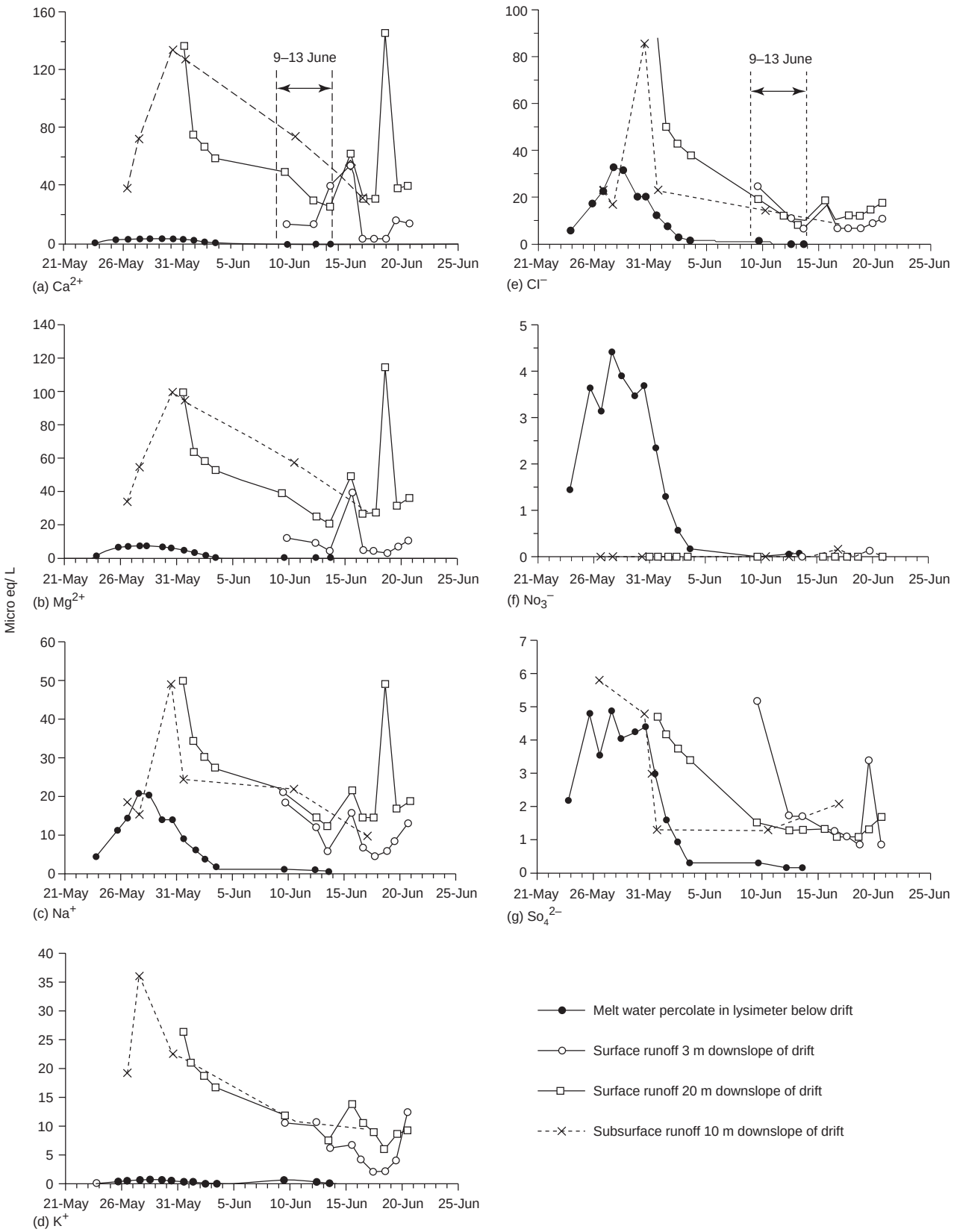


Figure 7 Ion concentrations in (i) meltwater, (ii) surface runoff 3 m downslope, (iii) surface runoff 20 m downslope, and (iv) subsurface runoff through 10 m of peat (After Quinton and Pomeroy, 2005)

lake acidification (Galloway *et al.*, 1987) (*see* **Chapter 95, Acidic Deposition: Sources and Effects, Volume 3**).

On an annual basis, the input and/or output of inorganic nutrients are relatively small, relative to the amount of nutrient that is being recycled within the basin itself. For example, values for the export of NO_3 by streamwater during melt in a boreal forest range from $0.65\,kg\,N\,ha^{-1}$ to $1.7\,kg\,N\,ha^{-1}$ (Tranter and Jones, 2001), which is small in comparison to organic pools of N. However, inorganic N has an extremely important role in primary productivity and is linked directly to carbon uptake by ecosystems. The contribution of inorganic N during snowmelt can be one of the largest inputs of this nutrient during the year (Jones, 1991)

The export of N as NO_3 originates both from the solute in the meltwaters (Williams *et al.*, 1993), and/or from the leaching of the species from soil after over-wintering nitrification of organic matter (Peters and Driscoll, 1987; Rascher *et al.*, 1987). In a study of a Colorado watershed, Lewis and Grant (1980) found that hydrological export of N as NO_3 increased significantly after winters over which the soil had frozen. This is supported by the work of Groffman *et al.* (1999) in a northern hardwood forest. The studies of Stottlemyer and Toczydlowski (1990), Stottlemyer and Toczydlowski (1999) indicated that nitrification occurs throughout the winter in organic soils when frozen soil conditions do not occur. Slow sustained meltwater discharge in winter, moves the mineralized N to lower inorganic horizons, where it cannot be taken up by microbiological activity. Some of the NO_3 can then be removed during the main melt period by macropore flow into surface water channels. Heuer *et al.* (1999) show that N dynamics in high altitude-basins are controlled by soil infiltration and transformation of N, and that soils can act as sources (alpine) or sinks (subalpine) of N during snowmelt.

Brooks *et al.* (1996) estimated that the winter/spring N losses by denitrification (N_2, N_2O) from an alpine basin were equal to the NO_3–N input by snowmelt, and represented 50% of the annual gaseous N loss. In this particular system, the loss of N via runoff was negligible during snowmelt. This latter result is in contrast with the hydrologic losses measured by Williams *et al.* (1993) at another alpine site, by Peters and Driscoll (1987) at a hardwood forest site, and by Jones and Roberge (1992) at a coniferous boreal site. These studies show the differing response of ecosystems to the input of N by snowmelt. The factors controlling the hydrologic export of N as NO_3 remain poorly understood.

An example of the effect of a nutrient-poor basin on the chemistry of snowmelt water, as it follows flow pathways to a stream is given by Quinton and Pomeroy (2005) for a frozen soil tundra site (Figure 7). Over flow pathways of from 3 to 20 m from a snow patch, enrichment of Ca, Mg, K, Cl, and Na was substantial. Concentrations increased from one to two orders of magnitude, indicating leaching from soil particles and surface vegetation, and possibly mixing with and flushing of soil water. SO_4 concentrations in meltwater were not strongly modified by hillslope flow, whilst NO_3 concentrations dropped to below detection limits shortly after exposure to the soil, indicating rapid microbiological uptake in the near-surface soil zones, despite below 0 °C soil temperatures.

Summary

Studies of the chemistry of snow-covered basins show that snow is not a passive reservoir of chemical species. Snow-atmosphere exchange by wind redistribution, dry deposition, and volatilization accompanied by various physical processes such as air and water movement, and snow metamorphism within the pack can increase or decrease the quantity of certain species, and/or lead to a redistribution of species within the snow cover. Chemical reactions such as photolysis also take place, while the presence of microorganisms in the snowpack will influence nutrient concentrations during the melt period. Inputs of inorganic nutrients to snow can be an important part of the annual cycling in the basin, particularly for N. Snowmelt chemistry alone is rarely directly responsible for major chemical fluctuations in water bodies, but the meltwater flux is important in mobilizing soil constituents and relocating soil water to the stream.

FURTHER READING

Cragin J.H. and McGilvary R. (1995) Can inorganic species volatilize from snow. In *Biogeochemistry of Seasonally Snow-covered Catchments*, Tonnessen K.A., Williams M.W. and Tranter M. (Eds.), IAHS Publication 228, IAHS, pp. 11–16.

Jeffries D.S. (1990) Snowpack storage of pollutants, release during melting, and impact on receiving waters. *Acidic Precipitation 4: Soils, Aquatic Processes, and Lake Acidification*, Advances in Environmental Science, Springer-Verlag: New York, pp. 107–132.

Valdez M.P., Bales R.C., Stanley D.A. and Dawson G.A. (1987) Gaseous deposition to snow: I. Experimental study of SO_2 and NO_2 deposition. *Journal of Geophysical Research*, **92**, 9779–9789.

REFERENCES

Albert M.R. and Shultz E.F. (2002) Snow and firn properties and air-snow transport processes at Summit, Greenland. *Atmospheric Environment*, **36**(15–16), 2789–2797.

Bales R.C. (1991) Modeling in-pack chemical transformations. In *Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D.,

Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 139–163.

Bales R.C., Davis R.E. and Stanley D.A. (1989) Ionic elution through shallow, homogeneous snow. *Water Resources Research*, **25**, 1869–1877.

Bales R.C., Valdez M.P., Dawson G.A. and Stanley D.A. (1987) Physical and chemical factors controlling gaseous deposition to snow. *Seasonal Snowcovers: Physics, Chemistry, Hydrology*, *NATO ASI Series V*, Reidel: Dordrecht, Netherlands, pp. 289–298.

Barrie L.A. (1991) Snow formation and processes in the atmosphere that influence its chemical composition. In *Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 1–20.

Barrie L.A. and Vet R.J. (1984) The concentration and deposition of acidity, major ions and trace metals in the snowpack of the eastern Canadian shield during the winter of 1980–1981. *Atmospheric Environment*, **18**, 1459–1469.

Beine H.J., Dominé F., Simpson W., Honrath R.E., Sparapani R., Zhou X., and King M. (2002) Snow-pile and chamber experiments during the polar sunrise experiment 'Alert 2000': exploration of nitrogen chemistry. *Atmospheric Environment*, **36**(15–16), 2707–2719.

Borys R.D., Demott P.J., Hindman E.E. and Feng D. (1983) The significance of snow crystal and mountain-surface riming to the removal of atmospheric trace constituents from cold clouds. In *Precipitation Scavenging, Dry Deposition and Resuspension. Precipitation Scavenging*, Pruppacher H.R., Semonin R.G. and Slinn W.G.N. (Eds.), Vol 1, Elsevier: New York, pp. 181–190.

Brimblecombe P. and Shooter D.S. (1991) Chemical change in snowpacks. In *Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 165–172.

Brooks P.D. and Williams M.W. (1999) Snowpack controls on nitrogen cycling and export in seasonally snow-covered catchments. *Hydrological Processes*, **13**, 2177–2190.

Brooks P.D., Williams M.W. and Schmidt S.K. (1996) Microbial activity under alpine snowpacks, Niwot Ridge, Colorado. *Biogeochemistry*, **32**, 93–113.

Brooks P.D., Campbell D.H., Tonnessen K.A. and Heuer K. (1999) Natural variability in N export from headwater catchments: snow cover controls on ecosystem N retention. *Hydrological Processes*, **13**(14–15), 2191–2201.

Cadle S.H. (1991) Dry deposition to snowpacks. In *Seasonal Snowpacks: Processes for Compositional Change. Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences, 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 21–66.

Cadle S.H. and Dasch J.M. (1987) The contribution of dry deposition to snowpack acidity in Michigan. In *Seasonal Snowcovers: Physics, Chemistry, Hydrology*, *NATO-ASI Series C: Mathematical and Physical Sciences*, Jones H.G. and Orville-Thomas W.J. (Eds.), Vol. 221, D. Reidel Publishing Company: pp. 299–230.

Cadle S.H., Dasch J.M. and Mulawa P.A. (1985) Atmospheric concentrations and deposition velocity to snow of nitric acid, sulfur dioxide and various particulate species. *Atmospheric Environment*, **19**(11), 1819–1827.

Cerling T.E. and Alexander A.J. (1987) Chemical composition of hoarfrost, rime and snow during a winter inversion in Utah,U.S.A. *Water Air and Soil Pollution*, **35**, 373–379.

Choi J., Conklin M.H., Bales R.C. and Sommerfeld R.A. (2000) Experimental investigation of SO_2 uptake in snow. *Atmospheric Environment*, **34**(5), 793–801.

Clow D.W. and Ingersoll G.P. (1994) Particulate carbonate matter in snow from selected sites in the south-central Rocky Mountains. *Atmospheric Environment*, **28**(4), 575–584.

Colbeck S.C. (1981) A simulation of the enrichment of atmospheric pollutants in snow cover runoff. *Water Resources Research*, **17**(5), 1383–1388.

Colbeck S.C. (1987) Snow metamorphism and classification. *Seasonal Snowcover: Physics, Chemistry, Hydrology*, *NATO ASI Series V*, Reidel: Dordrecht, Netherlands, Vol. 211, pp. 1–35.

Colin J.L., Jaffrezo J.L., Pinart J. and Roulette-Cadene S. (1987) Sequential sampling of snow in a rural area. Experimentation and identification of the acidifying agents. *Atmospheric Environment*, **21**, 1147–1157.

Colin J.L., Renard D., Lescoat V., Jaffrezo J.L., Gros J.M. and Strauss B. (1989) Relationship between rain and snow acidity and air mass trajectory in eastern France. *Atmospheric Environment*, **23**, 1487–1498.

Conklin M.H. (1991) Dry deposition to snowpacks. In *Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 67–70.

Conklin M.H. and Bales R.C. (1993) SO_2 Uptake on ice spheres: liquid nature of the ice-air interface. *Journal of Geophysical Research*, **98**(D9), 16 851–16 855.

Courchesne F. and Hendershot W.H. (1988) Cycle annuel des elements nutritifs dans un bassin-versant forestier: contribution de la litiere fraiche. *Canadian Journal of Forest Research*, **18**, 930–936.

Cronan C.S. and Reiner W.A. (1983) Canopy processing of acidic precipitation by coniferous and hardwood forests in New England. *Oecologia*, **59**, 216–223.

Dasch J.M. (1987) Measurement of dry deposition to surfaces in deciduous and pine canopies. *Environmental Pollution*, **44**, 261–277.

Davies T.D., Abrahams P.W., Tranter M., Blackwood I., Brimblecombe P. and Vincent C.E. (1984) Black acid snow in the remote Scottish Highlands. *Nature*, **312**, 58–61.

Davies T.D., Tranter M., Jickells T.D., Abrahams P.W., Landsberger S., Jarvis K. and Pierce C.E. (1992) Heavily contaminated snowfalls in the remote Scottish Highlands: a consequence of regional-scale mixing and transport. *Atmospheric Environment*, **26A**, 95–112.

Davis R.E. (1991) Links between snowpack physics and snowpack chemistry. In *Proceedings of the NATO Advanced*

Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 115–138.

de Caritat P., Hall G., Gìslason S., Belsey W., Braun M., Goloubeva N.I., Olsen H.K., Scheie J.O. and Vaive J.E. (2005) Chemical composition of arctic snow: concentration levels and regional distribution of major elements. *Science of the Total Environment*, **336**, 183–199.

Delmas V., Jones H.G., Tranter M. and Delmas R. (1996) The chemical weathering of aeolian dusts in alpine snows. *Atmospheric Environment*, **30**, 1317–1325.

Feng X., Kirchner J.W., Renshaw C.E., Osterhuber R.S., Klaue B. and Taylor S. (2001) A study of solute transport mechanisms using rare earth element tracers and artificial rainstorms on snow. *Water Resources Research*, **37**(5), 1425–1435.

Galloway J.N., Henrey J.R., Schofield C.L., Peters N.E. and Johannes H. (1987) Processes and causes of lake acidification during spring snowmelt in the west central Adirondack Mountains, New York. *Canadian Journal of Fisheries and Aquatic Sciences*, **44**, 1595–1602.

Groffman P.M., Hardy J.P., Nolan S., Fitzhugh R.D., Driscoll C.T. and Fahey T.J. (1999) Snow depth, soil frost and nutrient loss in a northern hardwood forest. *Hydrological Processes*, **13**, 2275–2286.

Hendershot W.H., Mendes L., Lalande H., Courchesne F. and Savoie S. (1992) Soil and stream water chemistry during spring snowmelt. *Nordic Hydrology*, **23**(1), 13–26.

Hewitt A.D. and Cragin J.H. (1994) Determination of ionic concentrations in individual snow crystals and snowflakes. *Atmospheric Environment*, **28**(15), 2545–2547.

Hibberd S. (1984) A model for pollutant concentrations during snow melt. *Journal of Glaciology*, **30**, 58–65.

Höfken K.D., Meixner F.X. and Ehhalt D.H. (1983) Deposition of atmospheric trace constituents onto different natural surfaces. In *Precipitation Scavanging, Dry Deposition and Resuspension*, Pruppacher H.R., Semonin R.G. and Slinn W.G.N. (Eds.), Elsevier, Science Publishing Co.: New York, pp. 825–835.

Hoham R.W. (1987) Snow algae from high-elevation temperate latitudes and semi-permanent snow: their interaction with the environment. In: Proceedings of the Eastern Snow Conference, 44th Annual meeting, Fredericton, N.B. June 3 & 4, pages: 73–79.

Hoham R.W., Yatsko C., Germain L. and Jones H.G. (1989) Recent discoveries of snow algae in upstate New York and Que'bec Province and preliminary reports on related snow chemistry. *Proceedings of the Eastern Snow Conference, 46th Annual meeting*, Quebec City, June 8 & 9, 1989.

Heuer K., Brooks P.D. and Tonnessen K.A. (1999) Nitrogen dynamics in two high elevation catchments during spring snowmelt 1996, Rocky Mountains, Colorado. *Hydrological Processes*, **13**(14–15), 2203–2214.

Ibrahim M., Barrie L.A. and Fanaki F. (1983) An experimental and theoretical investigation of the dry deposition of particles to snow, pine trees and artificial collectors. *Atmospheric Environment*, **17**(4), 781–788.

Iida T., Ueki K., Tsukahara H. and Kajihara A. (2000) Point physical model of movement of ions through natural snow cover. *Journal of Hydrology*, **235**(3–4), 170–182.

Johannessen M. and Henriksen A. (1978) Chemistry of snow meltwater: changes in concentration during melting. *Water Resources Research*, **14.**, 615–619.

Jones G. (1999) The ecology of snow-covered systems: a brief overview of nutrient cycling and life in the cold. *Hydrological Processes*, **13**(14–15), 2135–2147.

Jones H.G. (1991) Snow chemistry and biological activity: a particular perspective of nutrient cycling. In *Seasonal Snowpacks: Processes for Compositional Change. Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 21–66.

Jones H.G. and Debois C. (1987) Chemical dynamics of N-containing ionic species in a boreal forest snowcover during the spring melt period. *Hydrological Processes*, **1**, 271–282.

Jones H.G. and Pomeroy J.W. (2001) Early spring snowmelt in a small boreal forest watershed: influence of concrete frost on the hydrology and chemical composition of streamwaters during rain-on-snow events. *Proceedings of the Eastern Snow Conference*, **58**, 209–218.

Jones H.G. and Roberge J. (1992) Nitrogen dynamics and sub-ice meltwater patterns in a small boreal lake during snowmelt. In *Proceedings of the Eastern Snow Conference*, Ferrick M. and Pangburn T. (Eds.), 49, pp. 169–180.

Jones H.G. and Sochanska W. (1985) The chemical characteristics of snow cover in a northern boreal forest during the spring run-off. *Annals of Glaciology*, **7**, 167–174.

Kamai M. (1976) Identification of nucleii and concentrations of chemical species in snow crystals sampled at the South Pole. *Journal of the Atmospheric Sciences*, **33**, 833–841.

Laberge C. and Jones H.G. (1991) A statistical approach to field measurements of the chemical evolution of cold ($<0\,^{\circ}C$) snow cover. *Environmental Monitoring and Assessment*, **17**, 211–216.

Lamb D.S., Mitchell D. and Blumernstein R. (1986) Snow chemistry in relation to precipitation growth forms. *Proceedings of the 23rd Conference on Radar Meteorology and the Conference on Cloud Physics. Snowmass. Colorado, September 23–26, 1986*, American Meteorological Society: Boston, pp. 77–80.

Landsberger S., Davies T.D., Tranter M., Abrahams P.W. and Drake J.J. (1989) The solute and particulate chemistry of background snowfall on the Cairngorm Mountains, Scotland: a comparison with a black acid snowfall. *Atmospheric Environment*, **23**, 395–401.

Lewis M.J. and Grant M.C. (1980) Relationships between snow cover and winter losses of dissolved substances from a mountain watershed. *Arctic and Alpine Research*, **12**, 11–17.

Loye-Pilot M.D., Martin J.M. and Morelli J. (1986) Influence of Saharan dust on the rain acidity and atmospheric input to the mediterranean. *Nature*, **321**, 427–428.

Lyons W.B., Wake C. and Mayewski P.A. (1991) Chemistry of snow at high altitude, mid/low latitude glaciers. In *Seasonal Snowpacks: Processes of Compositional Change*, *NATO ASI Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Heidelburg, pp. 359–384.

Marsh A.R.W. and Webb A.H. (1979) *Physico-chemical Aspects of Snow-melt*, Central Electricity Generating Board Reports, RD/L/N, 60/79, London, p. 12.

Marsh P. and Pomeroy J.W. (1999) Spatial and temporal variations in snowmelt runoff chemistry, Northwest Territories, Canada. *Water Resources Research*, **35**(5), 1559–1567.

Neubauer J. and Heumann K.G. (1988) Nitrate trace determinations in snow and firn core samples of ice shelves at the Weddel Sea, Antarctica. *Atmospheric Environment*, **22**, 537–545.

Nicholson K.W., Branson J.R. and Giess P. (1991) Field measurements of the below-cloud scavenging of particulate material. *Atmospheric Environment*, **25A**, 771–777.

Peters N.E. and Driscoll C.T. (1987) Sources of acidity during snowmelt in the west-central Adirondack Mountains, New York. In *Forest Hydrology and Watershed Management*, Swanson R.H., Bernier P.Y. and Woodward P.D. (Eds.), IAHS Publication 167, IAHS, pp. 99–108.

Pomeroy J.W. and Gray D.M. (1995) *Snow Accumulation, Relocation and Management*, Report No. 7, National Hydrology Research Institute Science, Environment Canada: Saskatoon, p. 144.

Pomeroy J.W. and Jones H.G. (1996) Wind-blown snow: sublimation and changes to polar snow. In *Processes of Chemical Exchange Between the Atmosphere and Polar Snow*, *NATO-ASI Series I, 43*, Wolff E. and Bales R.C. (Eds.), Springer Verlag: New York, pp. 453–490.

Pomeroy J.W., Brown G., Davies T.D., Jones H.G., Tranter M. and Peters N.E. (2000) Micro-scale variation in the deposition of sea-salt components in snow in Celtic mountains. In *Water in the Celtic World: Managing Resources for the 21st Century*, Jones J.A.A., Gilman K., Jigorel A. and Griffin J. (Eds.), BHS: Wallingford, pp. 205–210,British Hydrological Society Occasional Paper No. 11.

Pomeroy J.W., Davies T.D. and Tranter M. (1991) The impact of blowing snow on snow chemistry. In *Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 71–114.

Pomeroy J.W., Davies T.D., Jones H.G., Marsh P., Peters N.E. and Tranter M. (1999) Transformations of snow chemistry in the boreal forest: accumulation and volatilisation. *Hydrological Processes*, **13**, 2257–2273.

Pomeroy J.W., Marsh P. and Lesack L. (1993) Relocation of major ions in snow along the Tundra-Taiga ecotone. *Nordic Hydrology*, **24**, 151–168.

Pomeroy J.W., Marsh P., Jones H.G. and Davies T.D. (1995) Spatial distribution of snow chemical load at the tundra-taiga transition. In *Biogeochemistry of Seasonally Snow-covered Catchments*, Tonnessen K.A., Williams M.W. and Tranter M. (Eds.), IAHS Publication No. 228, IAHS Press: Wallingford, pp. 191–206.

Quinton W.L. and Pomeroy J.W. (2005) Transformations of runoff chemistry in an Arctic tundra catchment. *Hydrological Processes*. in press.

Rascher C.M., Driscoll C.T. and Peters N.E. (1987) Concentration and flux of solutes from snow and forest floor during snowmelt in the west-central adirondack region of New York. *Biogeochemistry*, **3**, 209–224.

Raynor G.S. and Haynes J.V. (1983) Differential rain and snow scavenging efficiency implied by ionic concentration differences in winter precipitation. In *Precipitation Scavenging, Dry Deposition, and Resuspension*, Pruppacher H.R., Semonin R.G. and Slinn W.G.N. (Eds.), Vol. 1,Elsevier: New York, pp. 249–264.

Sequeira R. (1991) A note on the consumption of acid through cation exchange with clay minerals in atmospheric precipitation. *Atmospheric Environment*, **25A**, 487–490.

Sigg A., Neftel A. and Zircher F. (1987) Chemical transformation in a snowcover at Weissfluhjoch, Switzerland. Situated at 2 500 m.a.s.l. In *Seasonal Snowcovers: Physics, Chemistry, Hydrology; a NATO.ASI Les Arcs, France, 13–24 July 1986*, Jones H.G. and Orville-Thomas W.J. (Eds.), Reidel Publishing Company: pp. 269–280.

Stagnitti F., Li L., Allinson G., Phillips I., Lockington D., Zeiliguer A., Allingson M., Lloyd-Smith J. and Xie M. (1999) A mathematical model for estimating the extent of solute- and water-flux heterogeneity in multiple sample percolation experiments. *Journal of Hydrology*, **215**(1–4), 59–69.

Stein J., Jones H.G., Roberge J. and Sochanska W. (1986) The prediction of both runoff quality and quantity by the use of an integrated snowmelt model. *Modelling Snowmelt-induced Processes*, IHAS publication 155, IHAS, pp. 347–358.

Stottlemyer R. (1987) Snowpack ion accumulation and loss in a basin draining to lake superior. *Canadian Journal of Fisheries and Aquatic Sciences*, **44**(11), 1812–1819.

Stottlemyer R. and Toczydlowski D. (1990) Pattern of solute movement from snow into an upper Michigan stream. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 290–300.

Stottlemyer R. and Toczydlowski D. (1999) Seasonal change in precipitation, snowpack, snowmelt, soil water and streamwater chemistry, northern Michigan. *Hydrological Processes*, **13**(14–15), 2215–2231.

Stottlemyer R. and Troendle C. (1999) Effect of subalpine canopy removal on snowpack, soil solution and nutrient export, fraser experimental forest, co. *Hydrological Processes*, **13**, 2287–2300.

Stumm W. and Morgan J.J. (1996) *Aquatic Chemistry*, *Third Edition*, J Wiley & Sons: New York.

Tranter M. (1991) Controls on the chemical composition of snowmelt. In *Proceedings of the NATO Advanced Research Workshop on Processes of Chemical Change in Snowpacks, Maratea, Italy, July 1990*, *Series G: Ecological Sciences 28*, Davies T.D., Tranter M. and Jones H.G. (Eds.), Springer-Verlag: Berlin, pp. 241–270.

Tranter M. and Jones H.G. (2001) "Snow chemistry". In *Snow Ecology: an Interdisciplinary Examination of Snow-covered Ecosystems*, Jones H.G., Pomeroy J.W., Walker D.A. and Hoham R.W. (Eds.), Cambridge University Press: Cambridge, pp. 118–140.

Tranter M., Abrahams P.W., Blackwood I.L., Brimblecombe P. and Davies T.D. (1988a) The impact of a single, black, snowfall on streamwater chemistry in upland Britain. *Nature*, **332**, 826–829.

Tranter M., Brimblecombe P., Davies T.D., Vincent C.E., Abrahams P.W. and Blackwood I. (1986) The composition of snowfall, snowpack and meltwater in the Scottish Highlands –

Evidence for preferential elution. *Atmospheric Environment*, **20**, 517–525.

Tranter M., Davies T.D., Brimblecombe P. and Vincent C.E. (1988b) The composition of acidic meltwater during snowmelt in the Scottish Highlands. *Water, Air and Soil Pollution*, **36**, 75–90.

Tranter M., Davies T.D., Brimblecombe P., Abrahams P.W., Blackwood I. and Vincent C.E. (1987) Spatial variability of the chemical composition of snowcover in a small, remote Scottish catchment. *Atmospheric Environment*, **21**, 853–862.

Tranter M., Tsiouris S., Davies T.D. and Jones H.G. (1992) A laboratory investigation of the leaching of solute from snowpack by rainfall. *Hydrological Processes*, **6**, 169–179.

Tsiouris S., Vincent C.E., Davies T.D. and Brimblecombe P. (1985) The elution of ions through field and laboratory snowpacks. *Annals of Glaciology*, **7**, 196–201.

Turk J.T., Taylor H.E., Ingersoll G.P., Tonnessen K.A., Clow D.W., Mast M.A., Campbell D.H. and Melack J.M. (2001) Major-ion chemistry of the Rocky Mountain snowpack, USA. *Atmospheric Environment*, **35**(23), 3957–3966.

Williams M.W., Brown A. and Melack J.M. (1993) Geochemical and hydrological controls on the composition of surface waters in a high-elevation basin, Sierra Nevada. *Limnology and Oceanography*, **38**, 775–797.

Williams M.W. and Melack J.M. (1993) Solute chemistry of snowmelt and runoff in an alpine basin, Sierra Nevada. *Limnology Water Resources Research*, **27**, 1575–1588.

164: Role of Glaciers and Ice Sheets in Climate and the Global Water Cycle

MARTIN J SIEGERT

Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol, UK

Glaciers and ice sheets store nearly 70% of the world's freshwater and would raise sea level by around 70 m if they melted. They are, thus, an important component of the global water cycle. Their role in climate can be assessed through current responses to global warming and their influence on the environment during periods of glaciation. At the Last Glacial Maximum, ~21 000 years ago, the glacial cover caused sea level to be 115 to 130 m lower than at present. The climate of Ice Age Earth was affected in many ways by ice sheets. First, the expansion of ice sheets caused huge changes to the surface albedo and, hence, the radiation budget of the planet. Second, the orography of North America and Eurasia was altered significantly by large ice sheets, which had implications for atmospheric flow. Third, as the ice sheets grew and decayed, the influx of freshwater to the oceans, and in particular the North Atlantic, affected the ocean circulation, which in turn affected climate. The result of this ice–ocean–atmosphere interaction was large-scale short-term oscillations in air temperatures, which have been recorded in Greenland ice cores. The ice and climate processes identified for the glacial cycle are relevant today. For example, the Greenland ice sheet is currently melting at its southern margin, and releasing freshwater to the North Atlantic. Importantly, global temperatures were a few degrees higher in the previous interglacial, and the Greenland ice sheet was about half its current size. As global warming continues, the response of the Greenland ice sheet could be to return to its Eemian configuration. Although far smaller in volume, glaciers respond much quicker to climate change than do ice sheets. In the last century, most have experienced decay, and the runoff generated has been estimated to account for at least 10% of the 15 to 20 cm of sea-level rise measured (see ***Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4****). In combination, glacier and ice-sheet melting may send large quantities of water into the oceans, some of it rapidly. This could have serious implications for the climate of the North Atlantic and, through teleconnection, the world.*

INTRODUCTION

The global water cycle can be defined as the exchange of water between the land (including the biosphere), the hydrosphere, the atmosphere, and the cryosphere (Figure 1). Climate is inextricably linked to the water cycle, and glaciers and ice sheets play a central role within this cycle in defining the environmental conditions of the planet.

Processes critical to the water cycle are the evaporation of water into the atmosphere (predominantly from the oceans), atmospheric circulation, oceanic circulation, terrestrial runoff of fresh water, and processes leading to sea level and salinity changes. Glaciers and ice sheets, and their behavior in space and time, are key elements in each of these processes. They are essentially enormous stores of freshwater and their melting contributes to the freshwater input to the oceans. They have high surface albedos and so influence the surface heat balance, and they can grow to such a degree that they can change the hypsometry of a continent, which has implications for atmospheric circulation.

The global water cycle controls the amount of freshwater available for human use. As climate alters so too will the global water cycle and, thus, the availability of freshwater. It is therefore essential to ascertain how the global water cycle operates. This paper examines the role of glaciers

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

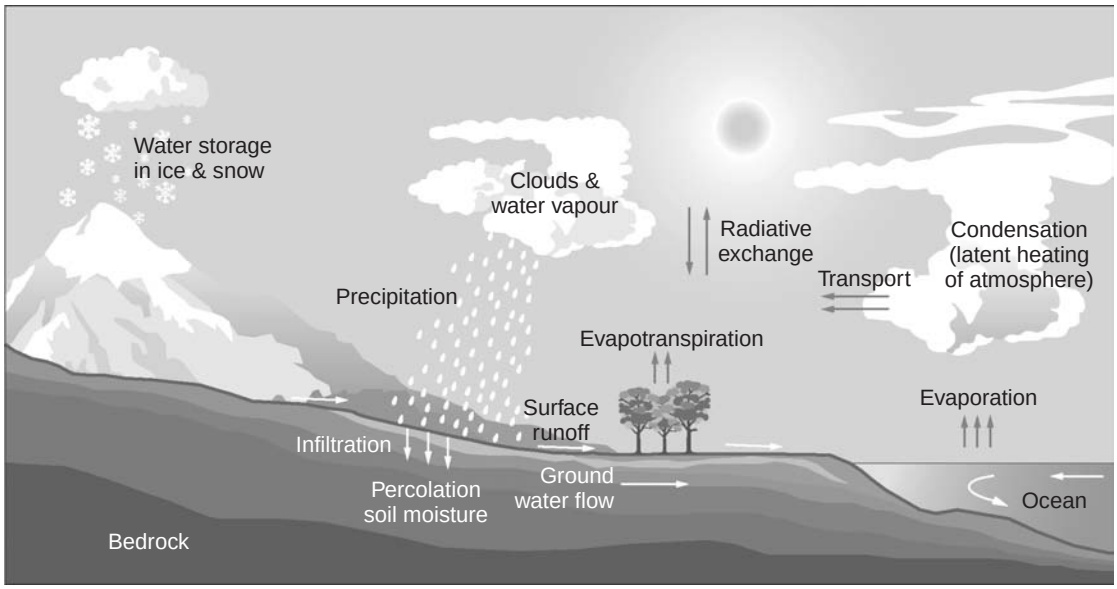


Figure 1 Schematic representation of the global water cycle

and ice sheets in the global water cycle, shows how past ice-sheet changes have had a marked influence on climate and water transfer processes, and discusses how ice sheets may respond to and influence climate today and in the future.

GLACIER VOLUMES

The oceans hold around 97% of all water on Earth (Gleick, 1996). Of the remaining 3%, around two thirds is held as ice within glaciers and ice sheets (the other third being in lakes, soils, rivers, and the atmosphere). Glaciers thus hold nearly 70% of the planet's freshwater. The distribution of glacier cover at present is provided in Table 1. By far the greatest ice volume is in Antarctica, where two ice sheets (the East and West Antarctic ice sheets) provide more than 25 million km^3 of ice, which, if melted, would raise global sea level by over 60 m. The next largest ice sheet is in Greenland, and this ice mass holds enough water to increase sea level by about 6 m. The remaining ice caps and glaciers in the world, if they too melted, would raise sea level by less than 0.5 m.

At the Last Glacial Maximum (LGM) at about 21 000 years ago, the spatial distribution of ice was much different from today (Table 2). Ice sheets at the LGM captured so much water from the oceans that global sea level was 115 to 130 m lower than its present-day level. Glacier expansion occurred predominantly in the high- to midlatitude Northern Hemisphere, where continental-scale ice sheets grew across North America and Europe. In North America, the Laurentide and Cordilleran ice sheets held as much ice as East Antarctica today. In some regions, such as Greenland and Antarctica, LGM ice sheets expanded from ice masses that had survived the preceding interglacial. The growth of ice was also witnessed in the Southern Hemisphere midlatitudes (e.g. Patagonia and New Zealand).

Table 1 Present-day volume of glaciers and ice sheets. The ice volumes for Greenland and Antarctica are for grounded ice only, and exclude ice shelves. As ice shelves are floating they have already displaced their weight in water and so if they melted sea level would not change as a consequence. The potential rise in sea level from ice sheets was determined by calculating the isostatic uplift that would occur if the ice were taken off, and allowing for the volume of seawater that would replace the ice in some areas. The sea level is therefore less than its ice volume equivalent (IPCC, 2001)

Geographic region	Volume (km^3)	Percent	Potential sea level rise (m)
Ice caps, ice fields, glaciers, and so on.	180 000	0.6	0.45
Greenland	2 850 000	10.0	7.2
Antarctica	25 710 000	89.4	61.1
Totals	28 740 000	100	68.75

Source: Adapted from Swithinbank (1985) and Williams and Ferrigno (2002).

Table 2 Ice sheets and glaciers at the LGM, and their likely contribution to sea-level lowering

Geographic region	Volume (km^3)	Percent	Contribution to LGM sea-level fall (m)
Antarctica[a]	37 000 000	45	14–18
Greenland[a]	4 000 000	5	2–3
North America[b]	34 000 000	41	78–88
Eurasian Arctic[c]	5 500 000	6	10–14
All others[d]	2 300 000	3	6
Totals	82 800 000	100	110–129

[a]From Huybrechts (2002).
[b]From Marshall *et al.* (2002).
[c]From Siegert *et al.* (1999).
[d]From Clark and Mix (2002).

However, outside of Antarctica, the lack of land surface (and continental shelf) limited the growth of ice in this half of the globe. The formation of Late Quaternary ice sheets caused alteration to the chemistry and circulation of the oceans, the flow of air (and moisture) within the atmosphere, the reflection of sunlight from the Earth's surface, and sea level. The environment of the planet was therefore very different at the LGM compared with today, and the global water cycle was an important control on this change.

GLACIER FLOW, MASS BALANCE, AND RESPONSE TIMES

Glaciers are fed with snow that, after a number of years of compression and firnification, turns to dense ice. The region in which ice accumulates is known as the *accumulation zone*, and is usually located in the highest, coldest regions of the ice mass. At the other end of the glacier, at lower elevations where it is warmer, ice is lost due to surface melting and, if the glacier terminates in water, iceberg calving. This region of net ice loss is known as the *ablation zone*. The line between these zones is where there is neither net ablation nor net accumulation of ice, and is called the *equilibrium line*. The altitude of this line is referred to as the "equilibrium line altitude" or, more commonly, the ELA. The flow of ice within glaciers acts to transport mass from the accumulation zone to the ablation zone (Figure 2). Thus, when the volume of ice accumulated equals the volume of ice ablated and the volume of ice transported from the accumulation zone to the ablation zone, the glacier is said to be "in balance". When a glacier is in balance, it remains in a stable, steady position, neither growing nor shrinking. In this case, the ELA is in a position that permits this "steady state" to exist. When the climate changes, the position of the ELA and the mass balance may alter. Glacier growth occurs when the annual mass balance is positive for a number of years. As a rule, ice accumulates in the

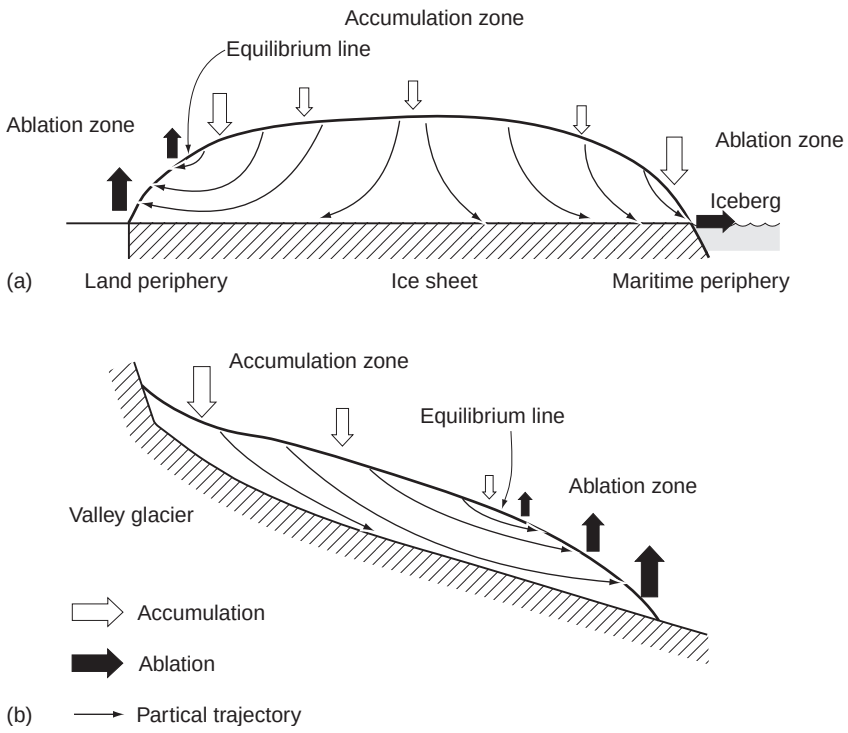


Figure 2 Glacier flow and the surface mass balance for (a) ice sheets and (b) glaciers. Reproduced from Siegert (2001) adapted from Sugden and John (1976) © Nature Publishing Group

winter and ablates in the summer. Generally, for glacier mass balance to become positive, the summer melt season needs to be cooler, resulting in less melting of ice. The net result of summer cooling is a lowering of the ELA. In this situation, the accumulation area is increased and the ablation area is decreased. This causes more flow of ice into the ablation region than gets ablated and, so, the glacier advances. Conversely, warmer conditions in the summer season result in a raising of the ELA, and more melting across a wider area of the glacier surface. If this happens for several years, the glacier will retreat.

This simple relationship between summer air temperatures, the ELA, and mass balance is true for large terrestrial ice sheets as much as it is for small glaciers. As the ELA is lowered, more and more land is available to be glacierized. It is therefore easy to comprehend how a reduction in summer air temperatures across a wide mountainous area would result in the large-scale expansion of glaciers. Further reduction in the ELA may cause these glaciers to coalesce and build up as single, large bodies of ice. And so it is that terrestrial ice sheets are initiated across mountain regions, when the ELA is forced down by a reduction in air temperature. Increase in summer air temperatures can cause an increase in the ELA and, therefore, an increase in net ablation. Thus, the growth and decay of ice sheets on land may be related to the position of the ELA and, because of this, to surface air temperatures. It should be noted, however, that in some cases, such as glaciers in Scandinavia, net accumulation occurs under warmer conditions because of the associated increase in precipitation. It should also be remembered that there is hardly any melting in Antarctica. Here, loss of ice is controlled predominantly by iceberg calving and sub-ice shelf melting.

If circumstances occur in which increased rate of ice flow causes more ice to be transported from the accumulation zone to the ablation zone than is supplied by accumulation, the advance of the glacier margin will eventually result in net ablation and subsequent retreat. This leaves the glacier with less ice than it had prior to the increase in ice flow. This unsteady flow of glaciers is called glacier surging and has been observed to occur in many glaciers and, recently, sections of relatively large ice caps (e.g. Sharp, 1988; Joughin *et al.*, 1996; Dowdeswell *et al.*, 1999).

Glaciers and ice sheets respond to climate change at different rates. The response time is "the time it takes a glacier to adjust to a change in its mass balance" (Paterson, 1994). The time taken for ice to flow from the accumulation zone to the ablation zone dictates the "response time" of an ice mass to climate change and is controlled by two parameters: the flow speed of ice and the glacier size. Fast-flowing small glaciers have a very low response time, while large slow-flowing ice sheets have much greater values. An estimate of the response time of a glacier (in which the deformation of ice dominates the total ice flow) is: $t = H/a_0$, where t is the time, H is the maximum ice thickness (at the divide if the base is flat) and a_0 is the ablation rate at the glacier margin (Paterson, 1994). For small temperate glaciers in maritime conditions, the response time is between 15 and 60 years, while for ice caps in the Canadian Arctic, which are larger and have a much cooler climate, the response time is between 250 and 1000 years. The greatest response times are in ice sheets, however. For example, in Greenland, Paterson (1994) estimates a response time of 3000 years. Thus, while ice sheets house the majority of the world's ice and freshwater, they have the largest response times. This makes them more robust in terms of withstanding short-term environmental variations than small temperate glaciers.

AGE OF GLACIERS AND ICE SHEETS

The response time of an ice mass is related generally to the age of ice within the ice mass. Near the center of a large ice sheet, or within the accumulation zone of a glacier, each annual layer of snow will lie on top of the previous layer. The deformation of this ice by vertical compression results in the thinning of ice layers. Thus, the age of ice becomes a function of depth and the accumulation rate of ice. An ice core taken from the center of an ice sheet will, therefore, comprise an ordered succession of ice from which climate records can be measured. Air bubbles within the ice trap the gaseous components of the atmosphere shortly after the time at which the original snow fell on the ice surface. Ice cores can be used to measure this gas and form a record of past concentrations of atmospheric constituents, such as carbon dioxide and methane. In addition, isotopes of oxygen and hydrogen can be measured from the ice itself, with values related to those of the snow from which the ice originates, to provide a proxy record of past air temperatures. The limit to the age of ice-core information is the age of the base of the ice sheet or glacier. Analysis of the Vostok ice core in central East Antarctica has provided knowledge not only of the last glacial climate, but of the previous four glaciations, spanning over 420 000 years. The oldest ice recovered from Antarctica is from the EPICA site at Dome C, where the ice is reported to be as old as 1 Ma at the base. The oldest water in Antarctica will be held as subglacial lakes, such as Lake Vostok (Siegert *et al.*, 2001), where water several million years of age could be present. For Greenland, ice cores such as GRIP and GISP2 provide a high resolution record of the glacial–interglacial cycle, and show major environmental instabilities during the last deglaciation (e.g. GRIP Project members, 1993; Grootes *et al.*, 1993; Stuiver and Grootes, 2000), but they rarely sample ice older than the last interglacial. Ice cores from other ice masses do not generally penetrate back to the LGM. They can often, however, provide detailed records of regional climate change over the last few thousand years.

Ice sheets, therefore, hold the majority of the planet's ice, have the greatest response times, and contain the oldest ice. They also have great potential for affecting climate and the hydrological cycle. Glaciers, on the other hand, contain far less ice, but they react quickly to climate change and so can be used as a means of assessing change.

GLACIERS, ICE SHEETS, AND CLIMATE CHANGE

Geological and glaciological data concerning the distribution of ice sheets at the LGM provide an excellent opportunity to assess how ice sheets both respond to and affect climate change. By examining the processes responsible for the buildup of ice at the LGM, we will uncover the ice-sheet processes critical to the global water cycle today.

Milankovitch Theory and the Forcing of Glaciations

By far the most cited theory relating climate change with ice-sheet growth was discussed initially by Croll (in the late nineteenth century) and later elaborated by Milankovitch in 1930. The theory is often referred to as the Milankovitch Theory (Figure 3). In this theory, cyclical changes in three of the Earth's orbital variants, namely, (i) orbital eccentricity, (ii) tilt of the axis (obliquity), and (iii) precession of the equinoxes, cause an alteration in solar radiation received at the Earth's surface and, hence, are thought to force air temperature changes.

The orbital eccentricity of the Earth varies in time between a highly elliptical orbit and a circular orbit. The periodicity of the cycle is 95 800 years (~100 ka). This is the most contentious of the orbital variations. Proponents of the theory claim it has the largest effect on ice-sheet growth, but it is actually responsible for only very small changes in solar insolation.

The tilt of the Earth's spin axis varies between 21.39° and 24.36°. The present value is 23.44°. The periodicity of this cycle is 41 000 years. Decreases in the axial tilt result in cooler summers in the more polar regions (high and midlatitudes), but do not change significantly the amount of solar radiation at low latitudes. Tilt variations do affect the degree of seasonal contrast at high latitudes, however, and higher obliquity results in increased seasonality.

The third orbital variation, caused by the conical rotation of the Earth's spin axis, affects the timing of perihelion and aphelion (the position where the Earth is closest and furthest from the Sun). The period of this movement is 21 700 years, over which time the Northern Hemisphere is tilted towards the Sun at successively different points in the Earth's elliptical orbit. At present, aphelion takes place during the Northern Hemisphere summer. This orbital variation causes an out of phase association in the degree of seasonality experienced by the two hemispheres.

Milankovitch analyzed, through the mathematics that describes the Earth's orbit around the Sun, the change in solar insolation caused by each of the orbital variations described above. These individual changes combine to yield the total modification in solar insolation due

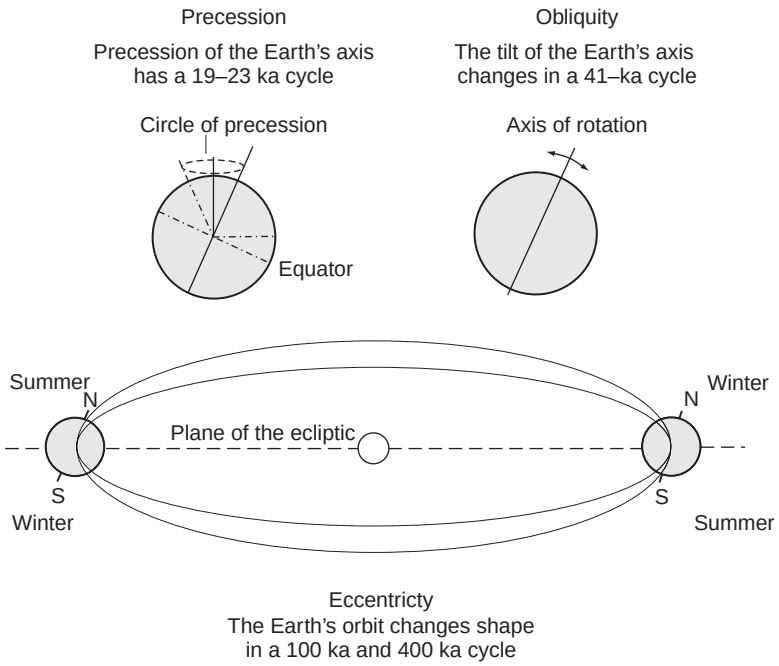


Figure 3 Milankovitch theory of orbital variations. Adapted from Siegert (2001)

to time-dependent variations in Earth's orbit. Ice growth (glaciation) occurs when a reduction in solar insolation in the Northern Hemisphere summer causes cooler temperatures and so less melting of snow and ice. This may be followed by a warm winter, when increased precipitation from the oceans falls onto high latitudes as snow. Ice recession (deglaciation) occurs when an increase in solar insolation in the summer causes relatively warm conditions, which result in an increase in the rate of ablation.

The most important season is the summer, when most melting can occur. Orbital calculations provide information on the variation of incident solar radiation as a function of season and latitude. From this, a complex relation between latitude and solar insolation through the last 130 000 years has been developed (Muller and MacDonald, 2000). On their own, the solar insolation changes predicted by Milankovitch are small, and not generally considered significant enough to initiate glaciations. Some sort of amplification of the climate response to orbital variation is therefore required. This is where ice sheets and glaciers can play a part in the global water cycle, climate modification and, consequently, their own existence.

There are several feedback mechanisms linking ice sheets, the oceans, and the atmosphere (Figure 4). A very good summary of the nature of these interactions is provided by Clark *et al.* (1999). They contend that ice sheets modulate ocean surface temperatures, the ocean circulation, the biosphere, surface albedo, and the continental water balance. In turn, these changes "feed back" and cause changes in ice-sheet configurations. Several of these feedback mechanisms are described below.

Amplifying Factors and Feedback Mechanisms

Sea Level

One important feedback mechanism allowing ice sheets to expand involves the link between sea level and ice-sheet volume. In this mechanism, ice growth in terrestrial regions causes sea-level reduction and, so, enables shallow marine regions such as the Barents Sea in the Northern Hemisphere or the Ross Shelf in Antarctica to become fully glaciated. The growth of ice in turn lowers the sea level further allowing additional expansion of marine-based ice. In addition, as sea level is lowered, the position of the ELA with respect to sea level will remain unchanged. However, the relative elevation of the land surface will increase. This can result in more land above the ELA, and an increase in the accumulation area of some glaciers. The subsequent growth of ice can then lead to further sea-level fall and the development of a simple feedback system. On the other hand, if accumulation decays with increasing distance from the ocean, then a reduction in sea level may result in less accumulation on certain glaciers.

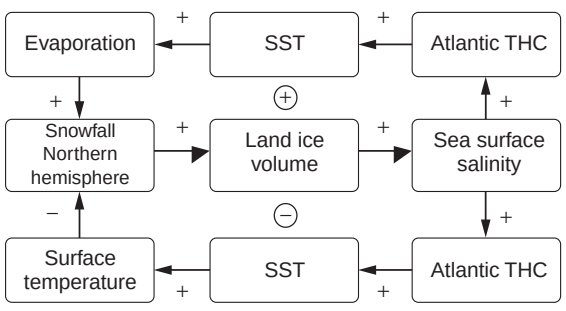


Figure 4 Flow diagram to describe the feedback processes associated with ice sheets, the oceans, and the atmosphere. Adapted from the IPCC (2001)

The decay of ice sheets during the last deglaciation was clearly associated with sea-level change. The melting of ice sheets into the oceans caused sea level to rise by 115 to 130 m. This may have caused marine-based ice sheets to become less stable since they would be more buoyant and more susceptible to basal motion. Ice-sheet melting would also have affected the salinity of the oceans, which would have implications for global climate. Ice–ocean interactions may have caused rapid changes in the climate during deglaciation; the likely processes involved are discussed in the Section "Evidence for rapid climate changes".

Surface Albedo

A second feedback system involves the reflective properties of the ice surface and the radiation budget of the Earth. The reflection of the Sun's radiation from the Earth's surface is related to the surface albedo. If the albedo is high, more radiation is reflected back into space. If the albedo is low, more radiation is absorbed by the planet's surface. Greater surface reflection of radiation results in a reduction in the amount of radiation absorbed. Snow and ice have extremely high albedo values. Therefore, as snowfields and ice sheets expand as a result of solar-insolation-induced adjustments in the air temperature, the increased surface albedo will cause an increase in the reflection of solar radiation and a further reduction in air temperature. Since this process will reduce the amount of summer warmth over snow-covered continents, it acts to reduce the rate of melting of ice. Cooler conditions also lead to the expansion of sea ice over the oceans and the length of time during the summer for which it persists. Sea-ice expansion lowers the albedo of the ocean surface considerably, and this contributes significantly to the feedback system. At the LGM, when large sections of the Northern Hemisphere were permanently covered with snow and ice and sea-ice extent was at a maximum, the surface albedo of the planet was drastically higher than it is today (although it is difficult to quantify by how much, a UK Meteorological Office model predicts a global rise in albedo of greater than 50%, from 0.15 at present to 0.26 at the LGM; Dan Lunt, personal communication, 2004).

As ice sheets decayed due to global warming, the albedo of the planet's surface decreased, causing the retention of heat at the Earth's surface and, hence, further warming.

Carbon Dioxide

Atmospheric carbon dioxide (CO_2) has an impact on climate because it absorbs/reradiates long-wave radiation emitted from the Earth's surface (i.e. the Greenhouse effect). For reasons not yet fully understood, when it gets colder, atmospheric CO_2 is depleted (Archer *et al.*, 2000). As the atmosphere cools in summer at the beginning of a glacial cycle (due to solar insolation changes), the CO_2 level drops and, because of this, so too does the absorption of long-wave radiation. Hence heat is lost from the Earth yielding further cooling. The influence of CO_2 on Late Quaternary ice sheets is currently not well understood. However, there are several potential climatic feedback mechanisms that involve CO_2 that may be relevant to the growth of ice sheets.

Levis *et al.* (1999) showed how CO_2 and surface albedo feedbacks may be linked, by running a model of the biosphere and climate at the LGM. The model results showed that CO_2 changes over a glacial cycle have an effect on the biospheric surface albedo. Their work indicated how a positive feedback, caused by changes to the gross albedo of the Earth's vegetation cover, might have contributed to cooling at the mid- and high latitudes at the LGM.

In addition, Antarctic sea ice would have covered a larger portion of the southern ocean during glacial times than today (Armand, 2000). Stephens and Keeling (2000) suggest that an expansion of southern ocean sea ice would reduce deep-water ventilation and inhibit the exchange of CO_2 between the ocean and atmosphere. They calculate that this process may be responsible for about 80% of the total CO_2 drop recorded in atmospheric proxy records during the last glacial.

Ice-sheet Elevation

As ice sheets grow, the surface area and elevation on which snow accumulates both increase. For example, the mean elevation of central East Antarctica is in excess of 2.5 km above sea level, as was the elevation of the Laurentide ice sheet in North America at the LGM. Because of this, ice growth can result in an increase in the area of the ice sheet above the ELA. The increase in the accumulation zone will result in further glacier growth. Such a change to the hypsometry of a continent will have a dramatic effect on the atmospheric circulation. This process was particularly relevant at the LGM across the Northern Hemisphere, when large ice sheets altered the elevation of the continents at mid to high latitudes. Kageyama and Valdes (2000) calculated the effect of the orography of the Laurentide ice sheet on the climate of the LGM. They used an Atmospheric General Circulation Model (GCM) with imposed ice-sheet dimensions. An additional experiment was undertaken where full LGM conditions were accounted for without any ice sheets present. By comparing the results of these two experiments, the effect of the Laurentide ice sheet topography on atmospheric flow and climate was evaluated. The height of the ice sheet was found to have a major effect on the climate over North America, the North Atlantic, and western regions of Europe. By subtracting the calculated rates of precipitation determined from the ice-free experiment from those from the full-LGM experiment, the influence of the Laurentide ice sheet on precipitation was identified. These model results suggest that the ice sheet was responsible for an increase in precipitation over Northern Canada, the northern North Atlantic, and Western Europe.

Ocean Circulation

Ocean circulation is important to glacier growth and climate change for three reasons. First, oceans are capable of transporting heat between latitudes. In the North Atlantic, for example, heat is transferred from low to high latitudes, providing moisture sources for northern hemisphere ice masses. Second, ocean thermal conditions are one of several factors affecting the growth of sea ice, which is an important albedo feedback. Third, ocean temperatures are a control on the rate of ice shelf melting, and possibly iceberg calving. Ocean circulation can be split into two components: surface currents and oceanic conveyor circulation.

Ocean surface currents are controlled by the direction and gradient of the prevailing winds, the rotation of the Earth (the Coriolis force) and by internal viscous dissipation giving rise to Ekman transport. As a consequence, gyres are set up with flow directions relating to their position on the globe, which can be disrupted (to a degree) by changes to wind field vectors and, in extreme polar environments, by changes in sea-ice extent (that can separate the ocean surface and the atmosphere).

The oceanic conveyor involves the formation, sinking, transfer, and upwelling of deep water due to changes in water density. In recent years, there have been several advances in the understanding of the global flow of the oceans (Figure 5). Oceanic convection is driven from the polar regions. Cold, salty water sinks to depth and moves equatorward. There are two main locations in the North Atlantic where this happens: the Labrador Sea between Greenland and America (known as the *Boreal region*) and the Norwegian–Greenland Sea (the Nordic region) (Imbrie *et al.*, 1992). The processes involved are evaporation of seawater, cooling of surface water through radiative and sensible heat loss, and formation of sea ice (e.g. the formation of the Odden Ice Tongue in the Norwegian–Greenland Sea), causing salty, high-density water to develop at the upper boundary, which then sinks

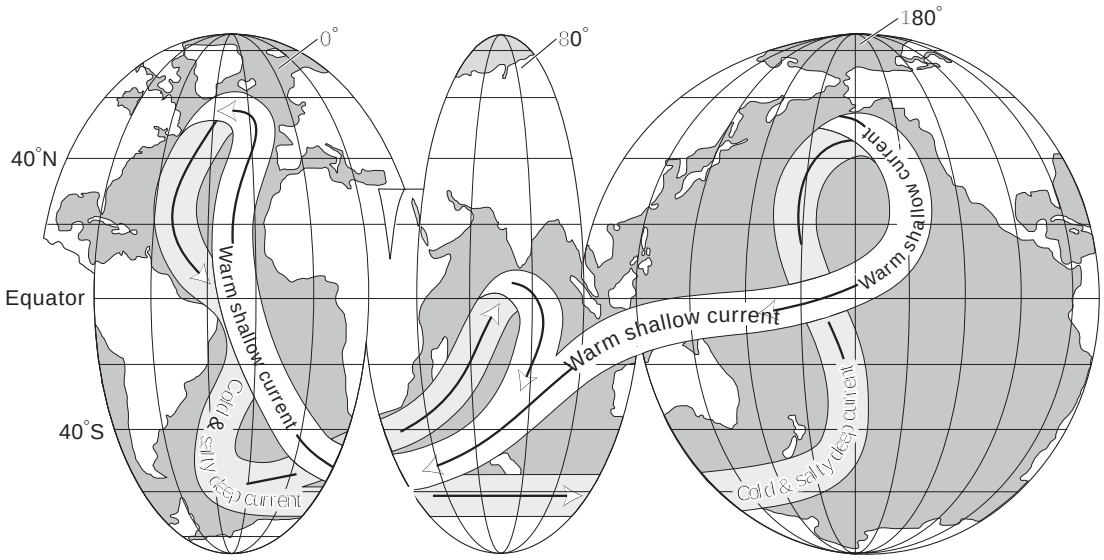


Figure 5 The global ocean conveyor system. Adapted from Siegert (2001)

to form "deep water". At present, North Atlantic Deep Water (NADW) meets the denser Antarctic Bottom Water (AABW) in the South Atlantic. These water masses then combine (at least in part) to flow into the Indian and Pacific oceans, where upwelling occurs and surface flow of water back to the Atlantic is initiated. The result is known as the *ocean conveyor belt*, which transports a flux of water more than 20 times greater than that of all the rivers combined (Figure 5). In the northern latitudes, oceans warm up between March and September and release heat in the rest of the year. In total, the amount of heat stored and released in the oceans is equivalent to about one-third of the heat received from the Sun. If the northward flow of warm water within the North Atlantic is halted, the ocean will cool and the heat supplied to Western Europe will be taken away. This would result in colder conditions over Europe and, hence, a reduction in the ELA.

One way of disrupting the ocean conveyor is to change the salinity of the ocean at the sites of NADW formation. For example, a decrease in sea-surface salinity caused by enhanced inputs of freshwater from rivers and glaciers may result in a reduction in NADW formation and, in turn, the effectiveness of deep ocean circulation (see the Section on "Evidence for rapid climate changes").

Imbrie *et al.* (1992) propose that the oceans were instrumental in the last deglaciation. In their model, the increase in summer insolation at the beginning of deglaciation caused warming of the atmosphere and oceans, which led to the snowfields shrinking. As the ELA increased, the Laurentide ice sheet melted in the south. The consequent effect on the wind field was to reestablish the Gulf Stream, and transport warm saline waters northward, increasing the exchange of waters within the Nordic regions. This heat helped to warm the atmosphere and, in turn, led to the melting of ice across Scandinavia and Eurasia. Deglaciation induced sea-level rise and, hence, the decay of marine ice sheets and then terrestrial ice sheets in the extreme north. The extra heat in the ocean increased the rate of evaporation, which in turn enhanced the effectiveness of the thermohaline system in the Norwegian–Greenland Sea.

Glaciers and Climate

Once a glaciation has started, a number of feedback processes can act to maintain the ice cover, regardless of the solar radiation inputs. It is interesting to note that, at the LGM, solar insolation values were similar to those at present (i.e. a deglacial period). It can thus be concluded that the climate at the LGM was greatly influenced, through feedback mechanisms detailed above, by the ice-sheet cover at that time. Moreover, as the planet warmed in the last deglaciation, ice sheets played an important role in regulating their own demise through their interaction with the atmosphere and oceans.

EVIDENCE FOR RAPID CLIMATE CHANGES

Greenland's GRIP and GISP2 cores hold a record of climate change as far back as the previous (Eemian) interglacial. The two cores, drilled about 30 km apart close to the summit of the Greenland ice sheet, yield a very similar palaeoenvironmental picture of the last 90 000 years or so (GRIP Project members, 1993; Grootes *et al.*, 1993). The ice-core records, with their very high temporal resolution, show that climate during the last deglaciation has been particularly variable and that alterations between warmer

and colder, and wetter and drier, conditions have been rapid. Fluctuations in the cores' oxygen isotope ratios suggest mean annual temperature variations of 10 to 15 °C, with shifts taking place rapidly over time periods on the order of decades. A doubling of accumulation at the beginning of the present interglacial also appears to have taken place over as little as three years. By contrast, the isotopic records from the Greenland ice cores demonstrate that climate variability during the present (Holocene) interglacial has been subdued by comparison, with changes of only a few degrees at most.

Ice sheets and glaciers are thought to be critical to rapid climate change of the last deglaciation. They can influence climate quickly by releasing large quantities of water, *via* melting or iceberg calving, into the oceans, so affecting ocean circulation and, thus, climate. Clark *et al.* (1999) focus on the causes of two periodic oceanic phenomena that may be related to cryospheric change: Dansgaard–Oeschger cycles and Bond cycles. Bond cycles are associated with a prolonged decrease in temperature (as recorded in Greenland ice cores) followed by an abrupt warming phase, and have a periodicity of around 7000 years (Alley, 1998). Dansgaard–Oeschger cycles, on the other hand, are millennial-scale periods of cooling and warming (Bond and Lotti, 1995; Bond *et al.*, 1997). Dansgaard–Oeschger cycles act in phase with Bond cycles during the transition between the cooling and warming, when large volumes of ice-rafted debris (IRD) are deposited across the North Atlantic by icebergs calved from the Laurentide ice sheet (so-called Heinrich events, see below) (Alley, 1998).

Alley and Clark (1999) suggest that changes in the ocean circulation caused by a Dansgaard–Oeschger event could be responsible for the Younger Dryas climate reversal at about 13 000 to 11 000 years ago. The idea is that a sudden switching off of the ocean conveyor caused sea-surface temperatures in the North Atlantic to become much colder, resulting in colder conditions over the adjacent land masses (e.g. Imbrie *et al.*, 1992; Clark *et al.*, 1999). This had the effect of reversing the deglaciation that had been going on since the LGM, and glaciers across most of Europe began to readvance. There is plenty of field evidence to show that this happened in Europe and many other parts of the Northern Hemisphere.

Heinrich Events and the Ice-sheet Binge-purge Theory

One theory fingers ice-sheet dynamics in North America as a possible cause of rapid ocean circulation and climate changes during periods of glaciation. The northeastern margin of the Laurentide ice sheet terminated as a calving ice wall along the Hudson Strait. Icebergs formed here would have drifted out into the North Atlantic where their sediments would have dropped to the ocean floor as IRD. Recent sedimentological investigations have revealed a series of IRD layers, corresponding to the 7000-year periodic production of huge volumes of icebergs. The sediments are known as *Heinrich layers* (see Figure 6). The cause of Heinrich layers is still debated. However, a commonly held view is that they were formed by periodic unstable dynamics of the Laurentide ice sheet (MacAyeal, 1993), the so-called binge-purge theory. The explanation for this theory is quite simple. The ice sheet slowly builds up over an essentially frozen base (the binge phase). Eventually, the basal temperatures reach the pressure melting value that initiates rapid basal motion (the purge phase). As ice flows out of the Hudson Strait it calves, and is advected across the North Atlantic. The drainage of so much ice depletes the reserves in the parent ice sheet such that it becomes thinner and eventually refreezes to the base, at which time the flux of ice to the Hudson Strait is decreased, and iceberg volume is reduced. This is then followed by ice regrowth. The cycle of binge and purge has been modeled by MacAyeal (1993) to be around 7000 years, which is a remarkably similar periodicity to that measured from IRD in the North Atlantic. The purge (surging) phase lasts for only a short time (around 700 years).

Meltwater Pulse 1A and the Bølling-Allerød Warm Period (14 600 years ago)

During the last deglaciation, the meltwater from decaying ice sheets and glaciers flowed to the ocean thus raising sea level. At around 14 600 years ago, so much water was transferred from the land that sea level increased by around 20 m in only a few hundred years. This huge, sudden input of freshwater to the ocean is referred to as "meltwater pulse 1A" (mwp-1A). The source of mwp-1A has been the subject of considerable debate. The traditional view is that the bulk of it came from the Laurentide ice sheet, with contributions from the Eurasian ice sheet. The cause of the decay was increased surface melting during the Bølling–Allerød warm period in North America and Europe. However, a rival theory has emerged in recent years linking mwp-1A to Antarctica (Weaver *et al.*, 2003). Both ideas are open to question. If the Laurentide ice sheet was responsible, it is not clear how a huge influx of cold freshwater to the North Atlantic is compatible with the Bølling–Allerød warm period as it may have led to a reduction in sea-surface temperatures (see the Section on "Proglacial lake outbursts and the 8200-year event" for an example of how freshening of the North Atlantic can result in climate cooling). There are two problems associated with the Antarctic ice sheet as the source. First, the ablation mechanism that would cause ice to melt rapidly is not obvious. Surface melting is an unrealistic candidate as there is virtually no runoff today. Temperatures would have to increase by several tens of degrees to cause enough water to melt from Antarctica to yield 20 m of sea-level rise, and there are no climate proxies to support such an event.

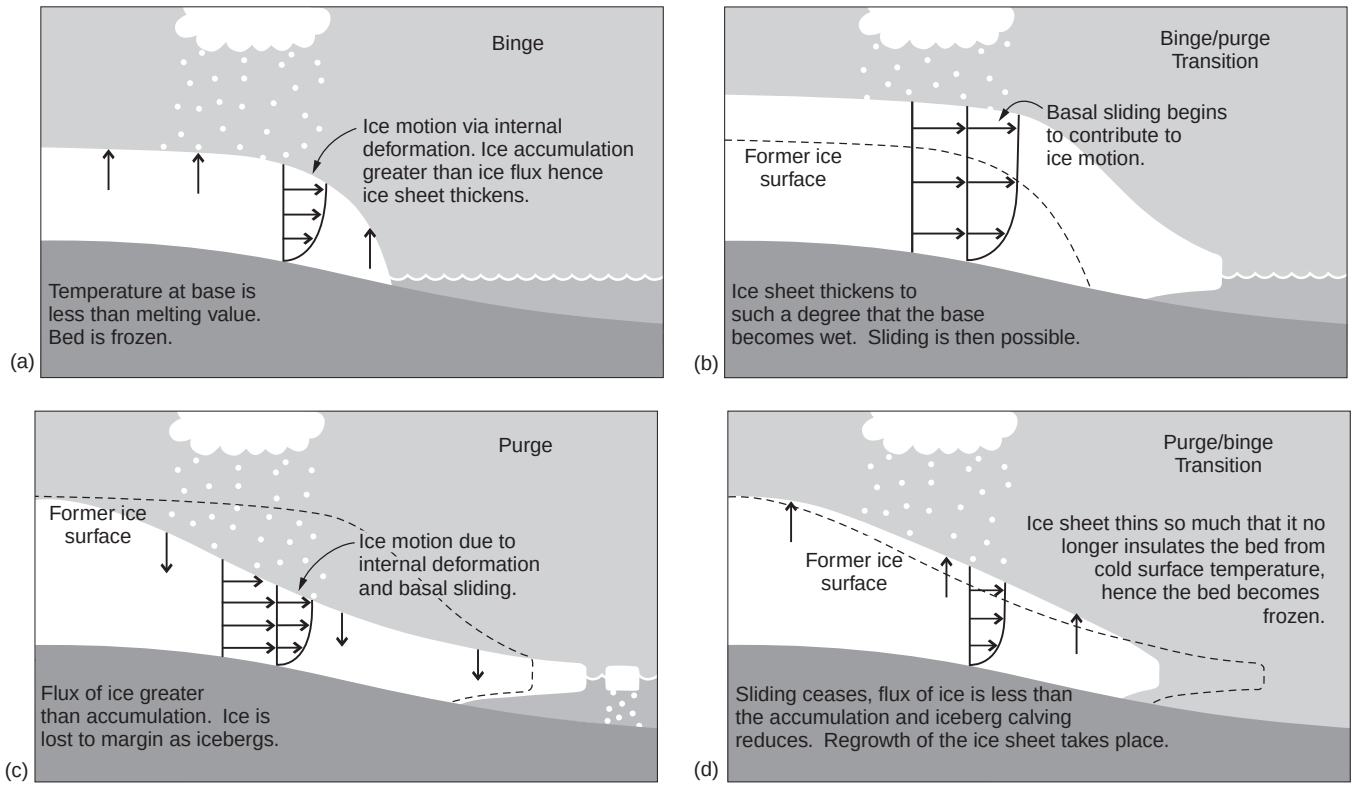


Figure 6 Binge-purge cycles of an ice sheet. (a) The binge phase. Ice-sheet growth occurs as a result of ice accumulation exceeding the flux of ice to the margin. Basal temperatures are cold as the only heat supply is from geothermal sources. (b) The binge/purge transition. As ice-sheet growth continues the base of the ice sheet warms due to enhanced heating from ice deformation and extra insulation from the ice above, until a critical threshold is reached at which the base becomes wet. (c) The purge phase. Once the subglacial conditions are warm basal sliding takes place (and this supplied more heat to the ice base), which leads to more ice transported to the margin than is replaced by falling snow and, hence, the ice sheet thins. (d) The purge/binge transition. As the ice sheet thins, the insulation effect of the ice diminishes and the base becomes cold. Once freezing conditions are reintroduced, sliding is no longer possible and the purge phase ends. As the accumulation of ice is once again greater than the flux of ice, the ice sheet begins to thicken and the binge phase begins again

Iceberg calving could be responsible, but only if this was associated with an increase in the flux of ice across the grounding line, which supposes a change to the interior ice-sheet dynamics. Second, numerical reconstructions of the Antarctic ice sheet show that its expansion involved an additional volume of ice less than that required to affect sea level by 20 m (Huybrechts, 2002). It is therefore unclear, how sea level could have risen by 20 m if the sole origin of mwp-1A was from Antarctica.

Proglacial Lake Outbursts and the 8200-year Event

The progressive retreat and thinning of the Laurentide ice sheet caused huge proglacial lakes to form along its southern margin. The largest of these was called *Lake Agassiz*. This massive lake fed freshwater into the Mississippi River between about 12 500 and 11 000 years BP (as is evident in the Gulf of Mexico isotope record). However, after 11 000 years BP lake waters were rerouted through the Great Lakes into the St. Lawrence. During its 4000-year history, Lake Agassiz played a significant role in the deglaciation of the Laurentide ice sheet. The lake reached its maximum extent between 9900 and 9500 years ago when it occupied an area of 350 000 km^2, within large regions of Ontario and Manitoba. At 8200 years ago, the ice sheet was too small to hold the lake water and so, catastrophically, Lake Agassiz drained into the Hudson Bay. Estimates put the volume of water involved in the instantaneous discharge at between 70 000 and 150 000 km^3, which would have been associated with an increase in global sea level of 0.2 to 0.4 m within a matter of days. Barber *et al.* (1999) match the sudden drainage of this lake through the Hudson Strait to a widespread cool period found in Greenland ice cores 8200 calendar years ago. Such an amount of freshwater must have affected the salinity of surface waters in the North Atlantic. In turn this

would have been disruptive to the formation of NADW and, thus, the thermohaline circulation. A decrease in the amount of warm northward-flowing mid-Atlantic surface water would have resulted in cooling of air temperatures in the higher latitudes and thus, a rapid climate change.

STABILITY OF ICE SHEETS AND IMPLICATIONS FOR FUTURE CLIMATE

The lessons being learned from the last deglaciation concerning ice sheets and their effect on climate are relevant today. The two largest ice sheets, in Greenland and Antarctica, have the potential to cause huge changes to the climate system, affecting the water cycle and society as a consequence.

Greenland

Palaeoclimatic records indicate that significant fluxes of freshwater as glacial runoff have caused dramatic cooling in the North Atlantic (Clark *et al.*, 2001) due to the freshening of Atlantic surface waters in regions of NADW formation. However, despite recognition from ocean models that the thermohaline circulation is sensitive to freshwater perturbations of as little as ~0.1 Sv (Manabe and Stouffer, 1994; Rahmstorf and Ganopolski, 1999), these models do not currently include the possible effects of significant freshwater input arising from changes in land ice sheets, which might well lead to bigger reductions in the thermohaline circulation (IPCC, 2001, p. 562).

The Greenland ice sheet is the largest store of freshwater in the Northern Hemisphere. The position, size, and topography of the southern Greenland ice sheet make it critical to future meltwater fluxes into the North Atlantic. It is well recognized that predicted global warming would likely be enhanced in high latitudes, with warming over Greenland likely to be up to three times the global average (IPCC, 2001). Runoff currently contributes ~50% of Greenland's annual ablation (Janssens and Huybrechts, 2000), concentrated along the southern margin of the ice sheet. The magnitude of runoff could be greatly enhanced in a warming world (Zwally *et al.*, 2002). The location of southern Greenland is particularly significant because it delivers runoff to key areas of NADW formation. The topography of Greenland also makes it highly susceptible to enhanced melting (Cuffey and Marshall, 2000). Greenland's ablation zone is currently relatively narrow (typically <100 km) but its topography (steep margins, shallow interior gradients) means that a modest climatic warming could rapidly expand areas of melt across the ice sheet surface. Since current temperatures already cause significant melting, any future warming will have a large impact on runoff, with the southern margin of Greenland being especially sensitive. Significantly, ice-sheet modeling and ice-core analyses suggest that the southern section of the GIS "collapsed" during the last interglacial period (Cuffey and Marshall, 2000), when temperatures were just a few degrees higher than at present.

As global temperatures rise to Eemian values, the Greenland ice sheet may return to its former configuration because of enhanced surface melting. Such change to the ice sheet would result in the direct input of huge volumes (2 to 3 m of global sea-level equivalent) of freshwater to the North Atlantic, with potentially significant consequences for the stability of the thermohaline circulation. The timescale over which such change might operate depends to a degree on the rate of meltwater supply. During the last deglaciation, when the rate of meltwater input to the oceans was at a maximum, such change occurred over decades. Gregory *et al.* (2004) predict that the Greenland ice sheet is likely to decay completely over the next 1000 years if the mean temperature of the region increases by 3 °C. Such change, if it happens, is bound to affect ocean conditions and, in turn, climate in a manner unprecedented in the last 8000 years.

Antarctica

In the 1970s, glaciologist John Mercer made two important observations. He discussed the prospect that the West Antarctic ice sheet was in danger of collapsing (Mercer, 1978). He also pointed out that the geography of West Antarctica was strikingly similar to the Eurasian Arctic: both involve a large continental shelf sea no more than a few hundred meters deep, and both have proximity to a pole and to a continental margin (Mercer, 1970). The major difference is that West Antarctica has a 2.5-km thick ice sheet over it, whereas the Eurasian Arctic is now comparatively free of grounded ice, having deglaciated after the LGM. The processes responsible for the decay of the Eurasian ice sheet are, thus, potentially important to understanding the stability of the present-day West Antarctic ice sheet.

In the last glaciation, the West Antarctic ice sheet was considerably larger than it is today. Grounded ice was probably in place across the whole continental shelf, just as it was in the Barents Sea. Ice decay in West Antarctica (from its maximum state to its present condition) was different to the decay of ice in the Eurasian Arctic in two ways. First, deglaciation began much later than in the Eurasian Arctic. Second, ice decay resulted in the formation of large floating ice shelves between the open ocean and the grounded ice sheet (i.e. the Filchner–Ronne and Ross ice shelves, each about 500 000 km^2 in area).

These differences suggest that two conclusions can be drawn about the stability of the West Antarctic ice sheet. First, the ice shelves may be influential in maintaining the stability of the ice sheet because they act as a buttress to support the grounded margin of the ice sheet, whereas in the Eurasian Arctic ice shelves were absent and the

grounded margin was actively involved in iceberg calving. Second, given the ice shelf buttressing effect on the ice sheet, the present sea level is not high enough to encourage ice decay in West Antarctica to the extent witnessed in the Eurasian Arctic.

One reassuring note relating to the latter conclusion is that during the Eemian interglacial, even though sea level was several meters higher than at present, the West Antarctic ice sheet did not decay. The majority of the water responsible for the Eemian sea-level rise probably came from Greenland (see above). The West Antarctic ice sheet is clearly capable of resisting substantial rises in sea level. The reason for this apparent stability could well be that the floating ice shelves in West Antarctica are key in maintaining grounded ice upstream. We should therefore be concerned with the stability of the ice shelves in West Antarctica. If these decay, the West Antarctic ice sheet will look much more like the former Eurasian ice sheet just before it broke up. In this situation, the processes that caused the break up of the marine portions of the Eurasian ice sheet may act over West Antarctica.

The oceans are critical to the stability of the ice shelves. Much of the mass loss from ice shelves is by subglacial melting (Jacobs *et al.*, 1992). If the ocean around Antarctica warmed, the ice shelf melt rates would increase, and the ice shelves could thin and ultimately disappear (provided melting exceeds surface accumulation).

The consequence of West Antarctic ice-sheet collapse would be a 6-m rise in global sea level, resulting in the inundation of many coastal and lowland areas of the world. Mercer called the situation a "threat of disaster". The glacial history of the Eurasian Arctic is an indicator that such ice-sheet collapse has occurred in the past.

West Antarctica and Heinrich Events

Although the southern ocean floors have not been cored as much as the North Atlantic, there is some IRD evidence from the southeast Atlantic to suggest that the West Antarctic ice sheet released regular pulses of icebergs during the last glacial cycle (Kanfoush *et al.*, 2000). Three sea-floor sediment cores reveal that IRD was deposited up to six times between 20 000 and 60 000 years ago, with a periodicity of between 6000 and 10 000 years. There are two main explanations for an IRD pulse in the southern oceans. The first is that the IRD was deposited by an increased number of icebergs. The second is that cooling of sea-surface temperatures may have allowed icebergs to survive transport across the southern ocean thus depositing their material in distal locations. However, one of the cores showing the IRD signal is taken from far north of the position of the LGM polar front, which seems to suggest that the former explanation may be relevant. Much work has still to be done to determine the true origin and significance of the southern ocean IRD pulses. However, there is a theory that the IRD reflects the instability of the West Antarctic ice sheet during the last 100 000 years (Kanfoush *et al.*, 2000).

GLACIER AND ICE SHEET RESPONSE TO RECENT CLIMATE CHANGE

Cogley (*see* **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**) reviews the various techniques by which the mass balance of large ice sheets can be assessed, and provides a summary of the results of mass balance investigations for a variety of glaciers and the ice sheets in both Greenland and Antarctica. A very brief overview of glacier and ice sheet response to recent climate change is provided below.

Glaciers

Global sea level has been rising by 1.5 mm yr^{-1} for the past hundred years or so, giving a total rise of approximately 15 cm (Warrick *et al.*, 1993). There are three main factors that affect global sea levels on timescales of decades to centuries. These are, first, the net loss of mass from glaciers and ice sheets to the oceans, second, the continued abstraction of groundwater aquifers, with the consequence that this water enters the global hydrological cycle and, third, the physical principle that the oceans will expand in a world, warmed by about one half of a degree over the last century. It is thought that a tenth or so of the observed annual rise in sea level could be a result of ablation from glaciers and ice sheets (IPCC, 2001).

The short response times of glaciers make it possible to assess their reaction to climate change by surveying them over a few decades. Net mass balance measurements from glaciers around the world show for the most part a consistent trend of negative mass balance over the last 30 years (Figure 7) and, in some examples where records permit, over the last 100 years (Haeberli *et al.*, 1996; Braithwaite, 2002). In Svalbard, for example, only four of the 44 measurements ($< 10\%$) of net mass balance show a positive annual increment. In the Canadian Arctic islands, over 70% of balance years are negative. However, it should be noted that these mass balance datasets have a relatively high interannual variability, although a number of the time series are statistically different from zero and indicate a negative mass balance. From the mass balance measurements of Arctic glaciers, distributed relatively widely across the polar North, it can be concluded that, during the last 30 to 40 years, glaciers from most areas of the High Arctic have shown no sign of building up, and have experienced either a negative or near-zero average mass balance. For glaciers at lower Arctic latitudes, the interannual scatter in the data is greater. This is presumably because of the higher turnover of mass, associated with greater precipitation and

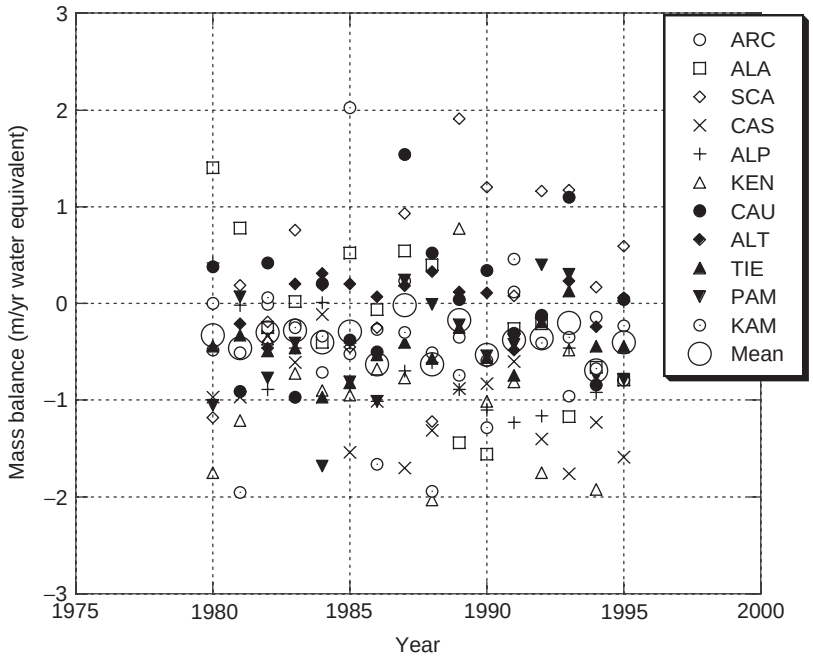


Figure 7 Glacier mass balance data from 11 locations over the 15 years (between 1980 and 1995). Data are from Braithwaite (2002) and Haeberli *et al.* (1996). The locations are denoted as follows: ARC (Arctic), ALA (Alaska), CAS (Cascades), SCA (Scandinavia), ALP (alps), KEN (Kenya), CAU (Caucasus), ALT (Altai), TIE (Tienshan), PAM (Pamir), and KAM (Kamchatka). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

stronger summer melting. Like their high latitude counterparts, most glaciers lower in the Arctic have experienced either a negative or a near-zero mass balance over the time period of observations, which is a maximum of 50 years. In Alaska, almost all the glaciers observed are in a state of negative mass balance and a few have mass balances close to zero over the period of observation (Arendt *et al.*, 2002).

Glaciers in maritime western Norway, such as Nigardsbreen, have had a clearly positive mean mass balance since measurements started in 1962 (Laumann and Reeh, 1993). A recent shift towards more positive annual balances has also been observed over almost the whole of Scandinavia since 1988 (Braithwaite, 2002). This region is a noticeable exception to the global trend of glacier retreat. It is thought that the region is influenced strongly by warming in North Atlantic, the positive phase of the North Atlantic oscillation and the northward deflection of winter storm tracks. Such conditions result in additional snowfall on the neighboring glaciers, which more than counters the negative effect of enhanced melting.

Glaciological investigations from a variety of glaciers within the tropics show most of them to be in a perilous state of decay (Thompson *et al.*, 2002; Kaser and Osmaston, 2001). In fact, the mass balance of glaciers such as those on Kilimanjaro in Tanzania is so negative that the glaciers could disappear entirely in as little as 20 to 30 years.

Ice Sheets

The long response times of ice sheets mean that it has traditionally been quite difficult to determine whether they are in positive or negative balance (*see* **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**). For example, Jacobs *et al.* (1992) estimated a negative balance for Antarctica of 469 Gt year^{-1}, based on measurements of accumulation, iceberg calving, sub-ice shelf melting, and runoff. However, the uncertainties associated with some of these measurements were between 20 and 50%.

Rignot and Thomas (2002) point out that there are three general ways of defining the mass balance of a large ice sheet. The first is by an assessment of the mass budget, as Jacobs *et al.* (1992) did for Antarctica. The second is from airborne and satellite measurements of the ice-sheet surface elevation (e.g. Wingham *et al.*, 1998), which, if data are acquired over a period of a few years, allow changes in ice thickness to be measured to centimeter accuracy. The third is by using satellite-based gravity measurements, since ice volume changes will have a measurable gravitational consequence (e.g. NASA's Gravity Recovery and Climate Experiment).

These techniques have been used to understand the current balance of the large ice sheets. In Greenland, NASA's Program for Arctic Regional Climate Assessment (PARCA) provided unprecedented data about the ice sheet's

state of health. This work showed that Greenland is melting, and thinning, across its southern margin. Importantly the surface meltwater finds its way to the ice base and acts as a lubricant to ice flow, so encouraging the ice to move faster, thus resulting in further thinning (Zwally *et al.*, 2002).

In Antarctica, the problem of mass balance remains hard to solve. Satellite altimeter data show that certain parts of West Antarctica are thinning (e.g. Pine Island and Thwaites glaciers). At the same time, however, the Siple Coast ice streams appear to be gaining mass. There is a connection between ice thickness changes and ice dynamics (i.e. as ice flow increases so the ice thickness will decrease), and this could explain the difference in elevation changes observed from different parts of the ice sheet. In the Siple Coast, for example, Ice Stream C has stopped flowing rapidly, and Whillans Ice Stream (formerly known as *Ice Stream B*) is slowing down (Joughin and Tulacyk, 2002). Similarly, the ice loss across the Pine Island sector could be result of increased velocities here, rather than an alteration to the surface balance.

SUMMARY

Glaciers and ice sheets comprise over two thirds of the world's freshwater and, if melted, could raise sea level by about 70 m. The distribution of ice around the planet is a function of today's climate and glacial history. Glaciers and ice sheets are thus intrinsically linked with climate.

Most glaciers are in a state of negative mass balance in response to climate warming over the last 100 years. Norway is an exception to this trend, where warmer conditions have led to enhanced rates of precipitation, which counter the extra melting experienced in summer. Glaciers have a short response time, and so are good barometers of climate change. Ice sheets, on the other hand, have much longer response times, and so their condition is not necessarily affected so much by recent changes.

Glaciers and ice sheets play a crucial role in the global water cycle as stores for freshwater on land. In the last glaciation, so much water was locked within ice sheets that global sea level fell by 115 to 130 m. Ice sheets at the LGM had a strong influence on the climate, in terms of their high albedo affecting the radiation budget, their orography affecting atmospheric circulation, and their runoff affecting the chemistry and circulation of the oceans.

During deglaciation the climate changed in a sporadic and rapid manner, switching from cold dry to warm wet conditions and back again within a few decades. Ice sheets were largely responsible for this rapid change in two ways. First, huge quantities of icebergs were released from ice sheets such as the Laurentide ice sheet in North America. These icebergs melted, causing IRD layers on the ocean floor (known as *Heinrich layers*) and the freshening of seawater. The reduction in salinity reduced the NADW formation and, hence, the thermohaline circulation which, in turn, cooled the climate of the North Atlantic region. A similar process may have been responsible for the Younger Dryas climate reversal at around 11 000 years ago that saw a regrowth of many Northern Hemisphere glaciers. Second, as the ice sheets melted large proglacial lakes were formed at their margins. One proglacial lake, named Lake Agassiz, outburst dramatically into the North Atlantic at 8200 years ago. The consequent release of freshwater directly into the Labrador Sea temporarily slowed the ocean circulation and a short cool period, recorded in Greenland ice cores, ensued.

The processes involving ice and climate during the last deglaciation could be relevant today. In particular, the southern margin of the Greenland ice sheet appears vulnerable to climate for two reasons. First, it is melting and retreating under the current climate. Second, in the previous (Eemian) interglacial, when air temperatures were only a few degrees warmer than at present, the Greenland ice sheet was far smaller than today. Hence, a future climate scenario could involve a much-reduced Greenland ice sheet. If so, the water melted from Greenland would be directed into the North Atlantic with potentially serious consequences for ocean circulation and climate.

Greenland is not the only ice mass capable both of affecting, and being affected by, future climate change. The majority of glaciers are currently in negative mass balance and are thought to be contributing at least 10% of the current rise in sea level (Meier, 1984; *see* **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**). Given their short response times to warming, they too could be influential to future climate and ocean processes.

The world's climate system is controlled to a large degree by the global water cycle. Glaciers and ice sheets are central to this cycle, and have the potential to affect it markedly over glacial–interglacial cycles. Moreover, their interaction with oceans and climate has often been an unsteady one and the processes involved in this interplay are relevant to our climate-changing world.

Acknowledgments

This chapter was improved greatly by the review of an anonymous referee and the comments of Martin Sharp (the associate editor).

FURTHER READING

Alley R.B. (2000) *The Two-mile Time Machine*, Princeton University Press: Princeton, p. 229.

Alley R.B., Mayewski P.A., Sowers T., Stuiver M., Taylor K.C. and Clark P.U. (1997) Holocene climatic instability: a prominent widespread event 8200 yr ago. *Geology*, **25**, 483–486.

Clark P.U., Pisias N.G., Stocker T.F. and Weaver A.J. (2002) The role of the thermohaline circulation in abrupt climate change. *Nature*, **415**, 863–869.

Dowdeswell J.A., Hagen J.O., Björnsson H., Glazovsky A.F., Harrison W.D., Holmlund P., Jania J., Koerner R.M., Lefauconnier B., Ommanney C.S.L., *et al.* (1997) The mass balance of circum-arctic glaciers and recent climate change. *Quaternary Research*, **48**, 1–14.

Grove J.M. (1988) *The Little Ice Age*, Methuen: London.

Muller R.A. and MacDonald G.J. (1997) Glacial cycles and astronomical forcing. *Science*, **277**, 215–218.

REFERENCES

Alley R.B. (1998) Iceing the North Atlantic. *Nature*, **392**, 335–336.

Alley R.B. and Clark P.U. (1999) The Deglaciation of the Northern Hemisphere: A Global Perspective. *Annual Reviews of Earth and Planetary Sciences*, **27**, 149–182.

Archer D., Winguth A., Lea D. and Mahowald N. (2000) What caused the glacial/interglacial pCO_2 cycles? *Reviews of Geophysics*, **38**, 159–189.

Arendt A.A., Echelmeyer K.A., Harrison W.D., Lingle C.S. and Valentine V.B. (2002) Rapid wastage of Alaskan glaciers and their contribution to rising sea level. *Science*, **297**, 382–386.

Armand L.K. (2000) An ocean of ice – advances in the estimation of past sea ice in the southern ocean. *Geological Society of America, Today*, **10**, 1–7.

Barber D.C., Dyke A., Hillaire-Marcel C., Jennings A.E., Andrews J.T., Kerwin M.W., Bilodeau G., McNeely R., Southon J., Morehead M.D., *et al.* (1999) Forcing of the cold event of 8,200 years ago by catastrophic drainage of Laurentide lakes. *Nature*, **400**, 344–348.

Bond G.C. and Lotti R. (1995) Iceberg discharges into the north Atlantic on millennial time scales during the last deglaciation. *Science*, **267**, 1005–1010.

Bond G., Showers W., Cheseby M., Lotti R., Almasi P., deMenocal P., Priore P., Cullen H., Hajdas I. and Bonani G. (1997) A pervasive millennial-scale cycle in North Atlantic Holocene and Glacial climates. *Science*, **278**, 1257–1266.

Braithwaite R.J. (2002) Glacier mass balance: the first 50 years of international monitoring. *Progress in Physical Geography*, **26**, 76–95.

Clark P.U., Alley R.B. and Pollard D. (1999) Northern Hemisphere ice-sheet influences on global climate change. *Science*, **286**, 1104–1111.

Cuffey K.M. and Marshall S.J. (2000) Substantial contribution to sea-level rise during the last interglacial from the Greenland ice sheet. *Nature*, **404**, 591–594.

Clark P.U., Marshall S.J., Clarke G.K.C., Hostetler S.W., Licciardi J.M. and Teller J.T. (2001) Freshwater forcing of abrupt climate change during the last glaciation. *Science*, **293**, 283–228.

Clark P.U. and Mix A.C. (2002) Ice sheets and sea level of the Last Glacial Maximum. *Quaternary Science Reviews*, **21**, 1–7.

Dowdeswell J.A., Unwin B., Nuttall A.M. and Wingham D.J. (1999) Velocity structure, flow instability and mass flux on a large arctic ice cap from satellite radar interferometry. *Earth and Planetary Science Letters*, **167**, 131–140.

Drewry D.J. (Ed.), (1983) *Antarctica – Glaciological and Geophysical Folio*; Cambridge England, Scott Polar Research Institute, University of Cambridge, Sheet 4.

Gleick P.H. (1996) Water resources. In *Encyclopedia of Climate and Weather*, Vol. 5, Schneider S.H. (Ed.) Oxford University Press: New York, pp. 817–823.

Gregory J.M., Huybrechts P. and Raper S.C.B. (2004) Threatened loss of the Greenland ice-sheet. *Nature*, **428**, 616–617.

GRIP Project members (1993) Climate instability during the last interglacial period recorded in the GRIP ice core. *Nature*, **364**, 203–207.

Grootes P.M., Stuiver M., White J.W.C., Johnsen S. and Jouzel J. (1993) Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature*, **366**, 552–554.

Haeberli W., Hoelzle M. and Suter S. (Eds.) (1996) Glacier mass balance bulletin. *World Glacier Monitoring Service Bulletin No. 4 (1994–1995)*.

Huybrechts P. (2002) Sea-level changes at the LGM from ice-dynamic reconstructions of the Greenland and Antarctic ice sheets during the glacial cycles. *Quaternary Science Reviews*, **21**, 1–3, 203–231.

Imbrie J., Boyle E.A., Clemens S.C., Duffy A., Howard W.R., Kukla G., Kutzbach J., Martinson D.G., McIntyre A., Mix A.C., *et al.* (1992) On the structure and origin of major glaciation cycles. 1. Linear responses to Milankovitch forcing. *Paleoceanography*, **7**, 701–738.

IPCC (2001) *The Scientific Basis Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J. and Xiaosu D. (Eds.), Cambridge University Press: p. 944.

Jacobs S.S., Hellmer H.H., Doake C.S.M., Jenkins A. and Frolich R.M. (1992) Melting of ice shelves and the mass balance of Antarctica. *Journal of Glaciology*, **38**, 375–387.

Janssens I. and Huybrechts P. (2000) The treatment of meltwater retention in mass balance parameterisations of the Greenland ice sheet. *Annals of Glaciology*, **31**, 133–140.

Joughin I., Slawek T., Fahnestock M. and Kwok R. (1996) A mini surge on the Ryder Glacier, Greenland, observed by Satellite Radar Interferometry. *Science*, **274**, 228–230.

Joughin I. and Tulacyk S. (2002) Positive mass balance of the Ross ice streams, West Antarctica. *Science*, **295**, 476–480.

Kageyama M. and Valdes P.J. (2000) Impact of the North American ice-sheet orography on the last glacial maximum eddies and snowfall. *Geophysical Research Letters*, **27**, 1515–1518.

Kanfoush S.L., Hodell D.A., Charles C.D., Guilderson T.P., Mortyn P.G. and Ninnemann U.S. (2000) Millennial-scale instability of the Antarctic ice sheet during the last glaciation. *Science*, **288**, 1815–1818.

Kaser G. and Osmaston H. (2001) *Tropical Glaciers*, Cambridge University Press: p. 228.

Laumann T. and Reeh N. (1993) Sensitivity to climate change of the mass balance of glaciers in southern Norway. *Journal of Glaciology*, **39**, 656–665.

Levis S., Foley J.A. and Pollard D. (1999) CO_2 climate, and vegetation feedbacks at the last glacial maximum. *Journal of Geophysical Research*, **104**, D24, 31, 191–31, 198.

MacAyeal D.R. (1993) Binge-purge oscillations of the Laurentide ice sheet as a cause of the North Atlantic's Heinrich events. *Paleoceanography*, **8**, 775–784.

Manabe S. and Stouffer R.J. (1994) Multiple-century response of a coupled ocean-atmosphere model to an increase of atmospheric carbon dioxide. *Journal of Climate*, **7**, 5–23.

Marshall S.J., James T.S. and Clarke G.K.C. (2002) North American ice sheet reconstructions at the last glacial maximum. *Quaternary Science Reviews*, **21**, 175–192.

Mercer J.H. (1970) A former ice sheet in the Arctic Ocean? *Palaeogeography, Palaeoclimatology, Palaeoecology*, **8**, 19–27.

Mercer J.H. (1978) West Antarctic ice sheet and CO2 greenhouse effect: a threat of disaster. *Nature*, **271**, 321–325.

Meier M.F. (1984) Contribution of small glaciers to global sea level. *Science*, **226**, 1418–1421.

Muller R.A. and MacDonald G.J. (2000) *Ice Ages and Astronomical Causes*. Springer Praxis publishing: Chichester, UK.

Paterson W.S.B. (1994) *The Physics of Glaciers*, *Third Edition*, Pergamon Press: p. 480.

Rahmstorf S. and Ganopolski A. (1999) Long-term global warming scenarios computed with an efficient coupled climate model. *Climatic Change*, **43**, 353–367.

Rignot E. and Thomas R.H. (2002) *Mass Balance of Polar Ice Sheets*. *Science*, **297**, 1502–1506.

Sharp M.J. (1988) Surging glaciers: behaviour and mechanisms. *Progress in Physical Geography*, **12**, 349–370.

Siegert M.J. (2001) *Ice Sheets and Late Quaternary Environmental Change*, John Wiley: Chichester, p. 231.

Siegert M.J., Dowdeswell J.A. and Melles M. (1999) Late Weichselian glaciation of the Eurasian high arctic. *Quaternary Research*, **52**, 273–285.

Siegert M.J., Ellis-Evans J.C., Tranter M., Mayer C., Petit J.-R., Salamatin A. and Priscu J.C. (2001) Physical, chemical and biological processes in Lake Vostok and other Antarctic subglacial lakes. *Nature*, **414**, 603–609.

Stephens B.B. and Keeling R.F. (2000) The influence of Antarctic sea ice on glacial-interglacial CO_2 variations. *Nature*, **404**, 171–174.

Stuiver M. and Grootes P.M. (2000) GISP2 Oxygen isotope ratios. *Quaternary Research*, **53**, 277–284.

Sugden D. and John B.S. (1976) *Glaciers and Landscape*, Edward Arnold: London.

Swithinbank C.W.S. (1985) A distant look at the cryosphere. *Advances in Space Research*, **5**, 263–274.

Thompson L.G., Mosely-Thompson E., Davis M.E., Henderson K.A., Brecher H.H., Zagorodnov V.S., Mashiotta T.A., Lin P.-N., Mikhalenko V.N., Hardy D.R. and Beer J. (2002) Kilimanjaro ice core records: evidence of Holocene climate change in tropical Africa. *Science* **298**, 589–593.

Warrick R.A., Barrow E.M. and Wigley T.M.L. (Eds.) (1993) *Climate and Sea Level Change: Observations, Projections and Implications*, Cambridge University Press: Cambridge, p. 424.

Weaver A.J., Saenko O.A., Clark P.U. and Mitrovica J.X. (2003) Meltwater pulse 1A from Antarctica as a trigger of the Bølling-Allerød warm interval. *Science*, **299**, 1709–1713.

Williams R.S. Jr and Ferrigno J.G. (Eds.) (2002) *Satellite Image Atlas of Glaciers of the World*, US Geological Survey Professional Paper 1386-A.

Wingham D.J., Ridout A.J., Scharroo R., Arthern R.J. and Shum C.K. (1998) Antarctic elevation change from 1992 to 1996. *Science*, **282**, 456–458.

Zwally H.J., Abdalati W., Herring T., Larson K., Saba J. and Steffen K. (2002) Surface melt-induced acceleration of Greenland ice-sheet flow. *Science*, **297**, 218–222.

165: Mass and Energy Balances of Glaciers and Ice Sheets

J GRAHAM COGLEY

Department of Geography, Trent University, Peterborough, ON, Canada

Glaciers exchange energy and mass with the rest of the hydrosphere by snowfall, melting, vapor transfer, and the calving of icebergs. Melting and vapor transfer are significant in both the energy balance and the mass balance, which in consequence are intimately coupled. Glacier energy balances differ from those of other natural surfaces in having small or even negative net radiation. Emission of terrestrial radiation is limited, the surface temperature being no greater than the freezing point, but the surface albedo is always high. The limit on surface temperature, and the year-round tendency for net radiative cooling, means that sensible heat transfer is generally downward, while vapor transfer may be either upward or downward. Once conduction has raised a surface layer to the freezing point, further energy surpluses are used to melt snow or ice. In winter, the energy balance is dominated by radiative cooling. Apart from its close connection with the energy balance, the mass balance is also influenced strongly by glacier dynamics. Glaciers and the flowlines of which they are composed exhibit vertical zonation, with net accumulation at higher and net ablation (mass loss) at lower elevations. This imbalance drives, and is corrected by, the ice flow. The leading methods for the measurement of mass balance are the direct, geodetic, and kinematic methods. Direct measurement involves determining the accumulation and ablation in situ or by equivalent remote sensing, with separate treatment of calving where it occurs. Geodetic measurements require the determination of glacier thickness at two epochs; the change of thickness, approximately equal to the change in surface elevation, gives a volume balance that may be converted to a mass balance if the density of the mass gained or lost can be supplied accurately. In the direct and geodetic approaches, the ice flow is assumed to integrate to zero over any one flowline (correctly, if the entire flowline is measured). Kinematic methods are free of this restriction. They involve measurement of all of the terms in the balance and are therefore more difficult. The need for better understanding of mass balance, at socioeconomic scales from local to global, has stimulated intense study of ways to improve the measurements. Recent and impending methodological advances are coming from radar altimetry, laser altimetry, gravimetry, passive-microwave remote sensing, and interferometry using synthetic aperture radar. A subject requiring increased attention, as the measurements improve in precision and coverage, is improved quantification of the measurement errors. The best current estimates of global average mass balance are equivalent to $0.14–0.44\,\mathrm{mm\,a^{-1}}$ of sea-level rise, to be compared with the inferred total rate of about $1.9\,\mathrm{mm\,a^{-1}}$. This figure is a composite of estimates for "small" glaciers (those other than the ice sheets), whose balance has been growing more negative since the 1960s; the Greenland Ice Sheet, which seems to have a negative balance; and the Antarctic Ice Sheet, for which the sign of the mass balance remains in doubt although its magnitude is probably within a few $\mathrm{kg\,m^{-2}\,a^{-1}}$ ($\mathrm{mm\,a^{-1}}$ water-equivalent) of zero.

INTRODUCTION

Glaciers exchange energy with the atmosphere overlying them and with the earth or ocean beneath. While the surface energy balance of a glacier is not fundamentally different from that of a drainage basin or other hydrological unit, the fact that glaciers have a basal energy balance and a (typically small) internal energy balance sets them apart. A more obvious distinguishing feature of glaciers is that, because they are made of frozen water which is apt to

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

melt, their energy and mass balances are very intimately coupled. Mass balance is the glaciological analog of the *water balance* in hydrology. Most glaciers gain mass by snowfall and lose mass by melting, although for some glaciers, including the largest, the calving of icebergs is an important term in the balance.

Glacier energy and mass balances are important in the following unranked respects:

- Glacier meltwaters are dominant components of the water balance of semiarid regions downstream of glacierized mountain ranges (e.g. Su and Shi, 2002). Such regions include the Prairies of Canada, Central Asia, and the Himalaya, and much of Andean South America. They also constitute an important resource for hydroelectric power generation, notably in Norway; a source of revenue from tourism; and hazards (Richardson and Reynolds, 2000).
- On the global scale, glaciers exchange mass with the ocean. As it is currently understood, the water balance of the ocean fails to add up and an accurate knowledge of glacier mass balance is required if we are to explain the observed contemporary rise of sea level (Church *et al.*, 2001; Munk, 2003). Glacier mass balance also affects the salt balance of the ocean.
- Glaciers play a part both in bringing about climatic change and in helping us to document it. They are highly reflective and so reduce the magnitude of net radiation at the Earth's surface, and as their extents change so does their influence on the global energy balance and the general circulation. As independent sources of information about environmental change, they are a valuable supplement to the weather stations from which we derive information about temperature and other leading climatic variables.
- Gains and losses of glacier mass imply redistribution of the mass of the Earth, altering its moments of inertia with consequences for the evolution of such geophysical quantities as length-of-day, true polar wander and the geoid, and with implications for understanding of the viscosity profile of the Earth's mantle (Peltier, 1998).

DEFINITION OF TERMS

A Column of Ice

In Figure 1(a), a column extends through ice at the Earth's surface. Ice, a soft solid, deforms readily under stress, so we orient the column with respect to the resulting flow. We assume that net exchanges of energy and mass through the side-walls are negligible, an idealization which is acceptable for balance studies if not for studies of dynamics. The lower surface makes contact with either the solid earth or water. At its upper surface, the column will ordinarily be exposed to the atmosphere, but there may be a complicating mantle of rocky debris (Nakawo *et al.*, 2000). Melting and freezing are regarded as loss and gain respectively; that is, liquid water is "outside" the column, being assumed to run off or refreeze in a time shorter than the span over which we compute the mass balance (of ice). The column also has internal energy and mass balances; for example, changes of water phase are not restricted to the surface and the base.

Flowline

A *flowline* (Figure 1b) is a sequence of ice columns of infinitesimal cross section arranged so that each column gains mass by flow from an up-ice neighbor and loses mass to a down-ice neighbor. To a good approximation, flowlines may be identified by beginning at any point where either the slope changes sign – at a flow divide – or the ice thickness drops to zero, and following the direction of steepest ascent or descent to another such point. The first column in the sequence has zero flow through one boundary. Most importantly, the integral of the mass flux divergence over the entire flowline is zero: a loss by flow from one part of the flowline must be compensated by a gain somewhere else, which means that we can neglect the flow when estimating the mass balance of the flowline.

Glacier

A *glacier* is a collection of contiguous complete flowlines through snow and ice that persists on the Earth's surface for more than one year (Figure 1c). The two largest glaciers are called *ice sheets*: the Greenland Ice Sheet and the Antarctic Ice Sheet. An *ice shelf* consists of the floating parts of two or more glaciers. There are small ice shelves in northernmost Canada and Greenland, but otherwise ice shelves are found only in Antarctica. Ice shelves differ from sea ice, which is a few meters thick, in being tens to thousands of meters thick.

Glacier Types and Glacier Zones

Glaciers are at or below their *freezing point* T_f, which in the absence of impurities increases from 273.16 K at the surface at a rate of about $0.67\,\mathrm{K\,km^{-1}}$ of ice overburden. *Cold* or *polar glaciers* are those in which temperature T is below T_f except possibly in a surface layer, up to 10–15 m thick, during summer. *Temperate glaciers* are at $T = T_f$ throughout, except in the surface layer during winter. *Polythermal glaciers* have, in addition to the surface layer of seasonal fluctuations, a basal layer at $T = T_f$ and an intermediate layer in which $T < T_f$. Cold glaciers are also *dry-based glaciers*, while polythermal and temperate glaciers are, at least locally, *wet-based glaciers*. These types

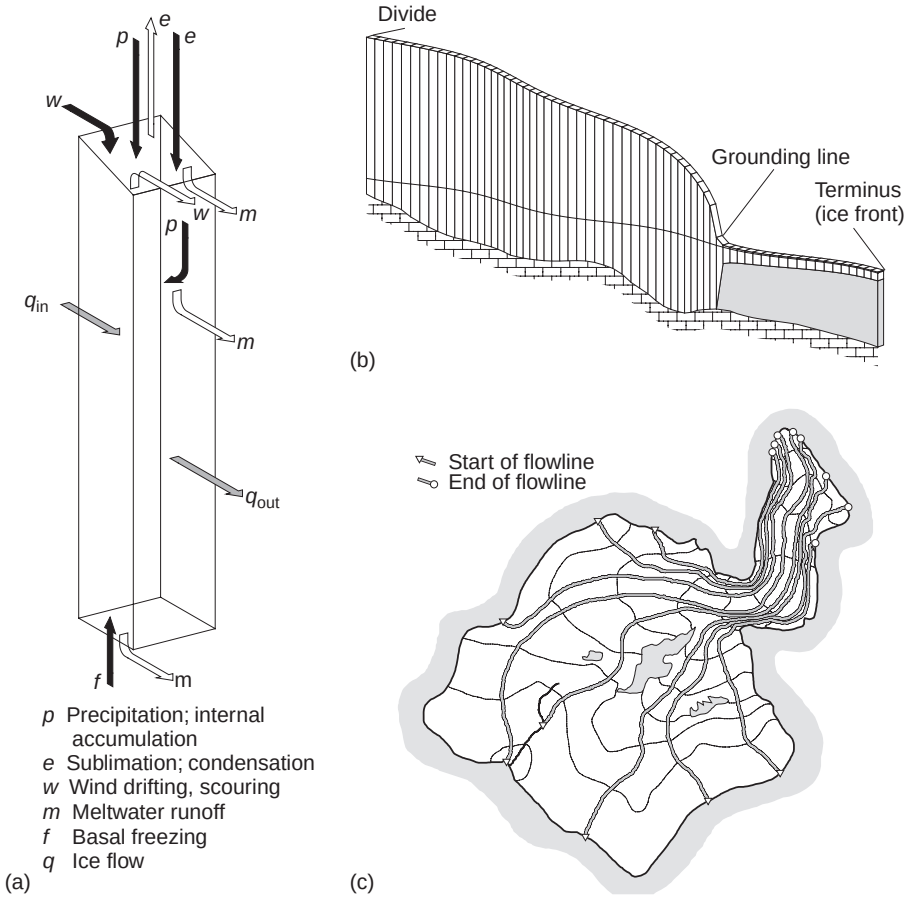


Figure 1 (a) A column of ice, showing leading mass-balance terms. Black arrows: accumulation (gain of mass); white arrows: ablation (loss of mass); grey arrows: throughflow. (b) A flowline considered as a sequence of ice columns. This flowline happens to have a floating terminal section. (c) Plan view of selected flowlines (thick) on a real glacier. Thin lines: contours (100-m interval)

can be misleading, for strictly the adjectives apply only to ice columns. Nevertheless, they are useful when considering energy and mass balances because the type determines whether, and if so where, melting and freezing may occur beneath the surface.

Because temperature decreases with increasing elevation at the surface, glaciers exhibit vertical zonation (Figure 2). This concept (Shumskiy, 1955; Benson, 1959) is the basis of most remote-sensing studies of energy and mass balances. *Snow* is solid condensates and precipitation, including freezing rain, added to the glacier during the current year. *Firn* is snow added in previous years. *Glacier ice* is firn that has been compacted to a density near that of pure ice. In the *dry-snow zone*, the temperature never rises to the freezing point and there is no melting. In the *upper percolation zone*, melting occurs at the surface in summer but the meltwater remains within the snow, while in the *lower percolation zone* some meltwater penetrates to the underlying firn before refreezing. This constitutes *internal accumulation*, which should be accounted for in the mass balance. The *slush zone* (Müller, 1962) is that part of the percolation zone in which at least some of the added snow is lost from the column, either as meltwater runoff or by slush avalanching. The *superimposed ice zone* represents either exposed, refrozen percolating meltwater or the product of slush avalanching. Ice is also found at the surface of the lowest zone, the *ablation zone*, but superimposed ice is newly gained mass while exposed glacier ice implies a loss of mass.

The surface volume balance of any column of the flowline is proportional to $z_1 - z_0 = z(t_1) - z(t_0)$ and, when we can neglect flux divergence and internal accumulation, the mass balance is simply this elevation difference multiplied by average density. The difference, and therefore the mass balance, is zero at the *equilibrium line* over the balance year (see Figure 2) or possibly some longer span. More generally, the whole flowline or glacier is said to be *in equilibrium* if the sum of its column balances is zero. Figure 2 hints, unrealistically, that the flowline will grow continually thicker in the accumulation zone and thinner in

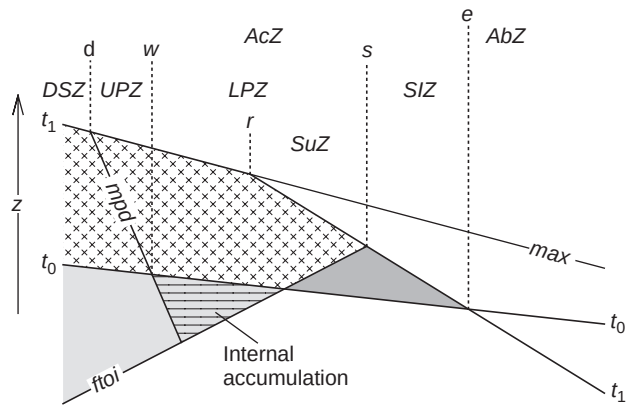


Figure 2 Cross section of a flowline, illustrating glacier zonation in elevation *z*. Broken cross-hatching indicates snow; firn is grey (the part vulnerable to internal accumulation being hatched), and superimposed ice is dark grey; glacier ice is unshaded. t_0, t_1: glacier surface at the start and end of the balance year; *max*: maximum elevation reached by transient glacier surface between t_0 and t_1; *ftoi*: boundary between firn and ice; *mpd*: maximum depth to which meltwater percolates before refreezing or reaching the effectively impervious (but not necessarily impermeable) barrier *ftoi*; d: dry-snow line (surface outcrop of *mpd*); *w*: wet snow line; *r*: runoff limit; *s*: snowline; *e*: equilibrium line. *DSZ*: dry-snow zone; *UPZ, LPZ*: upper and lower percolation zones; *SuZ*: slush zone; *SIZ*: superimposed ice zone; *AbZ*: ablation zone; *AcZ*: accumulation zone (the set of all zones above *e*)

the ablation zone. In fact, this pattern of thickness change and differential loading is what drives the glacier flow. The strongest definition of *equilibrium* is that it is the state that prevails when the flow is exactly that required to preserve the shape of the glacier unchanged, that is, for surface elevation z to remain constant everywhere. Figure 2 "works" only because of the hidden assumption that the surface z_1 transforms into the surface z_0 at the first instant of each new balance year, when snow turns into firn and superimposed ice into glacier ice.

Units

Glacier energy balances are usually reported in $\mathrm{W\,m^{-2}}$, which is the standard in climatology.

Mass balance is a rate per unit of projected horizontal area. It is nearly always reported for *balance years* (beginning at the start of winter) or their winter and summer components, so the appropriate units are $\mathrm{kg\,m^{-2}\,a^{-1}}$. Several other units are in common use. The most common are mm water-equivalent $\mathrm{a^{-1}}$, numerically identical with $\mathrm{kg\,m^{-2}\,a^{-1}}$ because 1 kg of water, with a density of $\rho_\mathrm{w} = 1000\,\mathrm{kg\,m^{-3}}$, is 1 mm deep when distributed over $1\,\mathrm{m^2}$. Care is needed when mass and volume balances are discussed together, for example when the measured quantity is a change of elevation or thickness. The thing to avoid is confusing ice thickness h (m) and water-equivalent "thickness" $(\rho_\mathrm{i}/\rho_\mathrm{w})h$ ($\mathrm{kg\,m^{-2}}$ or equivalently mm w.e.), $\rho_\mathrm{i} = 917\,\mathrm{kg\,m^{-3}}$ being the density of pure ice.

When discussing glacial contributions to changes of sea level a natural unit to use is mm sea-level-equivalent $\mathrm{a^{-1}}$. A glacier mass balance of B $\mathrm{kg\,m^{-2}\,a^{-1}}$ is equal to $-1000(S/S_\mathrm{o})B/\rho_\mathrm{w}$ mm s.l.e. $\mathrm{a^{-1}}$, S being the area of the glacier and $S_\mathrm{o} = 362.0\,\mathrm{Mm^2}$ the area of the ocean and ice shelves.

ENERGY BALANCE

Surface Energy Balance

The energy balance at the surface of a glacier is

$$K_\mathrm{s} + L_\mathrm{s} + H + \lambda E + G_\mathrm{s} + \mu M_\mathrm{s} = 0 \qquad (1)$$

where all the quantities are flux densities in $\mathrm{W\,m^{-2}}$. $K_\mathrm{s} = (1-\alpha)K\downarrow$ is net solar radiation, $K\downarrow$ being the incident solar irradiance and α the surface albedo. $L_\mathrm{s} = L\downarrow + L\uparrow$ is the net terrestrial radiation, the balance of gains from the upper hemisphere and the surface emittance $L\uparrow = -\varepsilon\sigma T_\mathrm{s}^4$. The surface emissivity, ε, is the ratio of surface emittance to that of a black body at the same temperature T_s, and $\sigma = 5.68 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ is the Stefan–Boltzmann constant. Snow and ice are usually treated as black bodies, that is, $\varepsilon = 1$. H and λE are turbulent fluxes of sensible and latent heat from above respectively, and G_s is the conductive flux of heat from below; $\lambda = 2.835 \times 10^6\ \mathrm{J\,kg^{-1}}$ is the latent heat of sublimation. Finally, μM_s is the energy used for surface melting, with $\mu = 0.335 \times 10^6\ \mathrm{J\,kg^{-1}}$ the latent heat of fusion and M_s the surface meltwater flux ($\mathrm{kg\,m^{-2}\,s^{-1}}$).

General methods for the measurement of terms in the surface energy balance are described by Oke (1987), and adaptations to the peculiarities of glacier surfaces by Oerlemans (2001). Accurate microclimatological measurements are demanding, and much of the effort in glacier climatology is devoted to parameterizing the results of research campaigns so that they may be used in wider contexts. Some important generalizations may however be made readily.

First, the net solar radiation on glaciers is usually small because the albedo is large (Table 1). At optical wavelengths snow is the brightest of natural surfaces, although its albedo can in fact vary greatly with age, grain size, the abundance of impurities and of liquid water, the incidence angle of the irradiant flux, and other factors (Warren, 1982). Exposed glacier ice is always darker than snow. Fresh snow may absorb three times less solar radiation than the ice that it covers, and its disappearance is followed by a marked shift in the energy balance to a more absorbent regime in which, other things being equal, melting is accelerated.

Table 1 Typical observed energy balances of glacier surfaces

Locality	Elevation (m)	Period	Type	α	R_s	H	λE	G_s	μM_s	References
QML	1180	37d (s)	B	0.58	50	0	−34	−16	0	Bintanja and Reijmer, 2001
QML	1170	37d (s)	S	0.79	2	9	−11	−1	0	Bintanja and Reijmer, 2001
QML	34	2a	S	0.87	−1.5	2.3	−0.8	−0.1	0	Reijmer and Oerlemans, 2002
QML	1420	2a	S	n/a	−26	26	0	0	0	Reijmer and Oerlemans, 2002
QML	2892	2a	S	0.84	−1.4	2.2	−0.7	−0.1	0	Reijmer and Oerlemans, 2002
W Greenland	1155	60d (s)	S	0.77	28	16	−6	−8	−30	Ohmura *et al.* 1994
SW Greenland	790	512d (s)	G	~0.30	103	62	−6	n/a	−161	Braithwaite and Zhang, 1999
N Greenland	540	35d (s)	G	~0.48	84	27	−24	−18	−71	Braithwaite and Zhang, 1999
Illimani, Bolivia	6340	21d (w)	S	0.82	−12	12	−22	22	0	Wagnon *et al.* 2003
Pasterze, Austria	2205	47d (s)	G	0.20	180	51	11	0	−242	van den Broeke, 1997
Pasterze, Austria	3225	47d (s)	S	0.59	65	22	1	0	−89	van den Broeke, 1997
Peyto, Alberta	2240	17d (s)	G	0.36	96	51	5	0	−152	Munro, 2001
Peyto, Alberta	2510	17d (s)	S	0.73	28	32	5	0	−63	Munro, 2001

Fluxes, in $W\,m^{-2}$, are positive towards the surface (R_s is net radiation, $K_s + L_s$; melting is negative). Error estimates vary from a few to several tens of $W\,m^{-2}$. QML: Queen Maud Land, East Antarctica. Period: s, w denote winter and summer. Type: B, blue ice; G, glacier ice; S, snow.

The longwave (terrestrial) balance is constrained by the fact that $T_s \leq T_f$, which limits $L\uparrow$ to magnitudes no greater than about $-316\,W\,m^{-2}$.

Because the air above glaciers is often warmer than the freezing point in summer, and is a heat source fueling intense radiative cooling in winter and at night, the sensible heat flux H is generally directed downwards. The latent heat flux λE is often directed downwards also because, even when liquid water is present, the vapor pressure at the surface will be appropriate to saturation at a temperature near T_f. On the lower parts of glaciers, the turbulent fluxes are enhanced by katabatic drainage of cooled air from high elevations. The katabatic wind, as well as being persistent and directionally constant, can be extremely strong.

The heat exchanged with the interior of the glacier drives an annual variation of temperature which is confined to the upper 10–15 m. However, in summertime, once an isothermal surface layer at the freezing point has been established, the heat flux G_s must dwindle to zero, and any surplus from the atmospheric terms in equation (1) will be used for melting. This surplus is responsible for most of the ablation on most glaciers, exceeding -10 m w.e. a^{-1} (about $-100\,W\,m^{-2}$ over the year) at lower latitudes. It may also be responsible for advective heat transfer to the interior of the glacier if the meltwater refreezes at depth instead of running off.

Some representative energy balances are summarized in Table 1. Blue ice is glacier ice exposed at the surface because snow fails to accumulate. Apart from scouring (removal as blowing snow), the principal reason for this in Antarctica is sublimation. Bintanja (1998) estimates by modeling that sublimation of blowing snow may reach $-15\,W\,m^{-2}$, twice the rate of sublimation of *in situ* snow and ice, near the Antarctic coast.

Internal Energy Balance

The energy balance of a small volume within a glacier may be understood (Paterson, 1994) in terms of thermal diffusion, advection of heat by the ice flow, and energy sources due to strain heating, including the compaction of firn, and the refreezing of meltwater. The strain-heating terms are of order $10^{-4}\,W\,m^{-3}$ or less, and are negligible for balance purposes even when integrated over typical column thicknesses, but the refreezing of meltwater can be significant. It may be expressed as $\mu f c/Z$, where c is the surface accumulation rate, f is the fraction of c that refreezes in the firn, and Z the thickness over which it refreezes. $\mu f c$ is of order $0.1-1\,W\,m^{-2}$ over the year, but Z is at most 10 m so that detectable summertime warming of the firn is possible.

Basal Energy Balance

Glaciers

The energy sources at the bed of a glacier are frictional heat and geothermal heat. If the basal temperature is below

the freezing point, the heat is conducted upwards into the body of the glacier. The opposite situation, heat flow from the glacier into its bed, is possible, for example, when temperate ice advances over permafrost, but very unusual. If the bed is at the freezing point the available energy is used to melt ice.

Pollack *et al.* (1993) have compiled measurements of geothermal heat flux. Averages are $0.09\,\mathrm{W\,m^{-2}}$ for Antarctica and $0.04\,\mathrm{W\,m^{-2}}$ for Greenland, with similar values for other glacierized regions. Only Iceland and the Rocky Mountains have geothermal fluxes above $0.10\,\mathrm{W\,m^{-2}}$, equivalent to $-10\,\mathrm{kg\,m^{-2}\,a^{-1}}$ of basal melting. Frictional heating derives from the loss of potential energy in the ice column as it moves downslope. Its rate can be expressed as the product of basal velocity times basal shear stress, yielding typical fluxes of $0.01-1\,\mathrm{W\,m^{-2}}$. Although the basal balance quantities are small, they are only marginally negligible given the current accuracy attainable in surface mass-balance calculations. They are also heat sources without compensating sinks, so they tend to make cold glaciers steadily warmer.

Ice Shelves

Assuming that thermodynamic equilibrium prevails (Doake, 1976; Holland and Jenkins, 1999), the contact between shelf ice and seawater must be at the freezing point of the seawater, which depends on the pressure of the overlying shelf ice and the salinity of the water. But in general the seawater and shelf ice at some distance from the contact will not be at T_f, so there is a heat source or sink, and therefore melting or freezing, at the contact. We can write

$$H_b + G_b + \mu M_b = 0 \tag{2}$$

where H_b is the sensible heat flux from the seawater, G_b the conductive flux from the shelf and μM_b the latent heat flux; fluxes are positive towards the shelf base and melting is negative. There are two complications. First, salt is coupled to heat because freezing increases and melting decreases the salinity of the seawater, altering both T_f and the buoyancy of the water. Second, the water flow itself is driven substantially by variations of temperature and salinity.

Holland and Jenkins (1999) envisage a boundary layer in the water flow beneath the shelf. At some elevation below the base, the water is at a temperature T_o and salinity determined by the mesoscale ocean circulation, and the sensible heat flux depends on the difference between T_f and T_o. Thus, the principal controls on the basal energy balance are the properties "imported" by the mesoscale water flow. Direct measurements are very difficult, but Rignot and Jacobs (2002) measured basal melting rates indirectly at 23 shelf grounding lines (Figure 1b). These rates are well correlated with an indirect estimate of T_o, and are extremely high. They pertain to areas of only a few tens of $\mathrm{km^2}$ (the square of glacier width at the grounding line), but the greatest magnitude, $-425\,\mathrm{W\,m^{-2}}$ or $-44\,\mathrm{m\ ice\ a^{-1}}$, at Pine Island Glacier, is a record. The meltwater is buoyant because it is fresh, and flows upwards along the base of the rapidly thinning ice shelf to where a lesser pressure implies a smaller sensible heat flux (higher T_f). Sometimes, it enters a regime in which it is actually colder than T_f and ice begins to form, accreting as "marine ice" at the base of the shelf. The latent heat flux averaged over all the Antarctic ice shelves, however, is believed to be negative. Jacobs *et al.* (1996) estimate it (with an uncertainty of ±50%) as $-5.4\,\mathrm{W\,m^{-2}}$, that is, $-500\,\mathrm{kg\,m^{-2}\,a^{-1}}$.

If there is net freezing at the base, it is reasonable (Holland and Jenkins, 1999) to set the temperature gradient in the shelf ice, and the implied heat flux, $G_b = k_\mathrm{i}(\partial T/\partial z)|_b$, to zero. Where there is net melting, the shelf temperature gradient is coupled to the dynamics of the shelf ice, but if we neglect this coupling a crude solution is available in terms of the temperature difference between surface and base. Taking typical values, $T_\mathrm{s} - T_b = -30\,\mathrm{K}$ and thermal conductivity $k_\mathrm{i} = 2\,\mathrm{W\,m^{-1}\,K^{-1}}$, we find that G_b ranges from -0.2 to $-0.02\,\mathrm{W\,m^{-2}}$ for ice shelves of thickness 300–3000 m.

MASS BALANCE

Methods of Measurement

The mass balance b of an ice column is

$$b = c + a + c_\mathrm{i} + a_\mathrm{i} + c_b + a_b + \Delta q \tag{3}$$

where the c are accumulation rates (black arrows in Figure 1a), the a are ablation rates (white arrows) and $\Delta q = q_\mathrm{in} - q_\mathrm{out}$ is the flux divergence. Subscripts i and b denote the interior and base of the column. Ablation by calving at the terminus is a special case of equation (3) in which a_i is equal to minus the entire mass of the column. The mass balance of a glacier or glacier flowline of area S is

$$B = \frac{1}{S}\int_\mathrm{s} b\ \mathrm{d}s \tag{4}$$

When b is assumed to vary only with elevation, as is usual on valley glaciers, the measurements are grouped into elevation bands and equation (4) becomes a sum of band averages, each average being weighted by the area of its band.

Direct Measurements

Direct measurements of column mass balance take the form

$$b = c + a = \frac{1}{\Delta t}\int_{z_0}^{z_1} \rho\,\mathrm{d}z \tag{5}$$

The flux divergence is ignored because the column balances are to be integrated over the glacier, and the other terms in equation (3) are either ignored or estimated as corrections. If it occurs, calving must be measured separately. Standard methods of measurement are described by Østrem and Brugman (1991), and Trabant and March (1999) give a detailed account of a careful protocol for fieldwork. Glaciers are dangerous places; safety in the field is discussed by Selters (1999).

A direct measurement of b involves emplacing a stake and/or digging a pit. If the stake is vertical, and does not tilt, bend, or settle, measurements of stake top height above the surface at t_0 and t_1 are proportional to z_0 and z_1, neither of which need be known in an independent coordinate system. In the ablation zone (Figure 2), the lost mass may be assumed to have a constant density $\rho = 900\,\mathrm{kg\,m^{-3}}$ (slightly less than ρ_i to allow for solid impurities, bubbles, intergranular voids, and macropores), so the mass balance is $b = \rho\Delta z/\Delta t$. In the accumulation zone, the density of the mass gained must be measured in the walls of snow pits, augmented with spatially extended surveys of the variability of Δz by probing or of b by shallow coring and weighing.

The column mass balance should be determinable with a standard error of the order of $\pm 50\,\mathrm{kg\,m^{-2}\,a^{-1}}$, and usually better. Except when the measurement network is very dense, an additional error is made by extrapolating from points to the whole glacier. Cogley *et al.* (1996) and Trabant and March (1999) both adopt a standard error of $\pm 200\,\mathrm{kg\,m^{-2}\,a^{-1}}$ for elevation-band averages of b, on the basis of the ability of single measurements to reproduce elevation-band averages determined with dense networks. Cogley (1999) showed that the uncertainty in B is not significantly less than this, because measurements of b at different elevations are nearly perfectly correlated.

Internal accumulation is a worrisome bias on any cold glacier with a percolation zone. It is impractical to measure it, and models for estimating it are as yet quite crude. Internal ablation (Mayo, 1992) occurs because of the conversion to heat of the potential energy of meltwater flowing down englacial channels. When the glacier is known to be wet-based, the basal ablation can be estimated as a function of the basal heat flux and frictional heating, although usually both it and internal ablation are neglected. Beneath polythermal glaciers, some of it is cancelled out when meltwater freezes on reaching cold parts of the bed, although the heat thus released helps to maintain the temperate ice at its melting point. Extensive basal accumulation occurs only beneath some parts of ice shelves.

The winter and summer balances, b_w and b_s, are measured separately on some glaciers. They are defined by equation (5) with z_0 and z_1 taken at the endpoints of the appropriate season, and in most climates they separate the two main controls, winter wetness and summer warmth, of the annual balance $b = b_\mathrm{w} + b_\mathrm{s}$. This is valuable because, for example, both controls are likely to involve more water than in neighboring unglacierized terrain, and relatively small changes in either can have substantial implications for the regional water balance.

Geodetic Measurements

Geodetic measurements of mass balance have until recently been used mostly as checks on the reliability of more frequent direct measurements. They rely on pairs of dated maps or other representations of $\Delta z/\Delta t$ to give a volume balance that may be converted to a mass balance by making correct assumptions about density. Geodetic measurements require the separate determination of z_1 and z_0 in geocentric coordinates, which introduces a quite different set of concerns about accuracy. For example, the quantity which should be measured is actually the rate of change of thickness, $h = z - z_b$, and changes in bed elevation, z_b, arising from glacial isostatic adjustment and other causes need to be allowed for.

In the accumulation zone, if the density profile remains unchanged between t_0 and t_1, then *Sorge's Law* is said to apply: Density is a function only of depth beneath the surface. However, the compaction rate varies with the rate of surface loading by new snow, the temperature and, possibly, the rate of internal accumulation. Wingham (2000) modeled the compaction of dry isothermal firn, finding that the spatial scale of fluctuations in accumulation is of critical importance. Zwally and Li (2002) modeled the effect of seasonal fluctuations of temperature and accumulation loading, reproducing observed fluctuations of z with fair accuracy. Where melting occurs and the resulting meltwater refreezes in the form of ice lenses, the situation is much more complicated, and at present it is necessary to invoke Sorge's Law arbitrarily.

Kinematic Measurements

In kinematic measurements, q_in or q_out or both are measured at "gates" (cross-sections), in combination with up-ice or down-ice measurements of accumulation and ablation. The advantage is that the complete-flowline assumption can be relaxed. The disadvantage is that q_in and q_out cannot be measured inexpensively. They require knowledge of ice thickness across the gate and of the vertical distribution of ice velocity. In practice, only the surface velocity is known, and either it must be assumed that the glacier moves entirely by basal sliding, or the velocity profile must be modeled (Kostecka and Whillans, 1988). Hubbard *et al.* (2000) modeled glacier flow to generate a map of the flux divergence. This method is not likely to be applied widely because of the amount of boundary-condition information required by the flow model. An important recent advance, discussed below, is the ability to measure q at grounding lines by radar interferometry.

Hydrological Measurements

In the hydrological approach, precipitation and evaporation over the glacier are estimated along with the runoff of meltwater, and the water balance is solved. This is only done routinely for Aletschgletscher, Switzerland. Bhutiyani (1999) has published hydrological estimates for Siachen Glacier in the Karakoram. The uncertainty in the hydrological method is in practice much greater than in a typical direct measurement. For example, glacier runoff must be separated from that contributed by unglacierized parts of the catchment tributary to the discharge measurement station.

Surrogate Observations

Glacier mass balance B is well correlated with the elevation e of the equilibrium line at the end of the balance year (Braithwaite, 1984). This offers a means of inferring B from less expensive observations of e, possibly from space, but, apart from glaciers on which B is already measured, the only regular reports of e are those of Chinn (1999) for New Zealand glaciers. B is equally well correlated with the *accumulation area ratio*, which is the area of the accumulation zone divided by the area of the glacier (Slupetzky, 1989). These economical surrogates are valuable for extending knowledge of mass balance variability, but they are necessarily quite uncertain.

It is easier to measure the position of a glacier's terminus than to measure its mass balance. Terminus fluctuations are reported annually for several hundred glaciers, as against fewer than 100 reports of mass balance (Haeberli *et al.*, 1998). Unfortunately, the link between mass balance and glacier length is indirect. An observed annual change of terminus position is a response to balance forcing integrated over the glacier and over some indefinite span much longer than one year.

Developing and Emerging Technologies

Conventional measurement methods are labor-intensive and uncertain, but the need for better and more comprehensive estimates of mass balance has stimulated vigorous research into alternatives. Some of these are discussed briefly here.

Radar Altimetry Radar altimeters, mounted on aircraft or satellites, emit pulses towards the nadir (the point on the surface vertically beneath the sensor) and "track" the waveforms of the return pulses. Tracking involves prediction of future ranges (distances to the surface, convertible to elevations z) from ranges detected in the immediate past. Poor predictions, as when the surface is steep or undulating, result in loss of "lock" on the surface. A cross-track interferometer on the future CryoSat radar-altimetry mission will address this problem, which at present compromises the glaciological use of radar altimetry outside the interiors of the two ice sheets. In those regions, however, repeated radar altimetry has transformed our knowledge of accumulation. Accurate identification of orbit crossover points is essential, so orbital errors, discussed in detail by Davis *et al.* (2000), become important. Wingham *et al.* (1998) give a long list of corrections needed before the range differences may be interpreted as elevation changes. It is also assumed that the radar interacts with the glacier by surface backscatter, with extinction and subsurface volume backscatter being unvarying. This appears to be a good description of the interaction in dry-snow zones, but melting and internal accumulation complicate matters. Bamber *et al.* (2001) showed that the root mean square (rms) error of radar altimetry with respect to collocated estimates from airborne laser altimetry was ~ 7 m. The laser-altimeter measurements have decimeter-level accuracy (Krabill *et al.*, 1995).

Laser Altimetry Laser altimetry is the measurement of surface elevation by measurement of the two-way travel time of a pulse of optical radiation. Repeated measurements yield the change of elevation, leaving the density of the ice column to be supplied, for example, by Sorge's Law. The position of the laser must be known precisely, and it must be possible to reoccupy horizontal positions with an accuracy comparable to the radius of the "footprint" of the laser pulse. These problems have been solved by the integration of GPS measurements into altimeter systems. Neither laser altimeters nor radar altimeters can estimate the flux divergence, so it is necessary to accumulate a glacier-wide coverage of column measurements.

The first satellite laser altimeter, GLAS (the Geoscience Laser Altimeter System), was launched on ICESat in January 2003. Airborne laser altimetry was used by Krabill *et al.* (2000), who found significant thinning at lower elevations and overall balance at higher elevations on the Greenland Ice Sheet over a five-year span. Arendt *et al.* (2002) presented balance estimates for 67 glaciers in Alaska, relying on maps drawn from aerial photography flown in the 1950s and on laser-altimetric surveys in the 1990s. A notable methodological conclusion is that errors in the geodetic balance are dominated by the errors in the old maps. The old maps offer the chance of extending the historical span of the measurements, but their inaccuracy limits significantly what can be achieved.

Gravimetry Velicogna and Wahr (2002) have studied the joint resolving power of GLAS, GPS measurements, and the Gravity Recovery And Climate Experiment (GRACE). Launched in March 2002, GRACE is a pair of satellites flying about 200–250 km apart along-track. Each satellite ranges to the other using microwave phase measurements. When nongravitational accelerations are accounted for, the residual fluctuations in range may be used to map the gravity field with a horizontal resolution of a few hundred kilometers once every 30 days. With the aid of GLAS and GPS estimates of change in elevation, GRACE geoid changes are interpreted as a result of glacial isostatic

adjustment and of changes in ice mass. Errors in the latter of the order of $\pm 16\,\mathrm{kg\,m^{-2}\,a^{-1}}$ may be expected over 250-km spatial and 5-year temporal scales. This is quite large by comparison with the minimum accumulation in the interior of East Antarctica, but GRACE offers a substantial improvement over current coverage.

Ice-penetrating Radar and Ice Cores Radars with wavelengths in the meter range (frequencies of 10–500 MHz) can yield information on the depth to reflective horizons within the ice. The vertical separation between any two such horizons, if both horizons are isochronous and of known age, is a measure of volume balance. The same reasoning applies to horizons identified in ice cores. Only the accumulation at the core site can be measured, but the resolution in time is likely to be much better, at least in the upper core where annual layers are recognizable. When a well-dated core record can be tied to extensive radar surveys, a great increase in coverage results. It would be valuable if this capacity, at present largely restricted to century and longer timescales (e.g. Siegert, 2003), could be extended to shorter and more recent time spans and to smaller glaciers. Here, the aim is to find summer surfaces – the crusts that separate balance years. Pälli *et al.* (2002) estimated errors for a traverse in Svalbard with a 50-MHz radar. Most of the uncertainty was due to reflector-tracking error. For 13-year-old (Chernobyl) and 36-year-old (nuclear test) layers, with accumulation rates of 670 and $580\,\mathrm{kg\,m^{-2}\,a^{-1}}$, errors were ± 134 and $\pm 49\,\mathrm{kg\,m^{-2}\,a^{-1}}$.

Microwave Emission and Backscatter Passive-microwave radiometers sense the emission of surfaces at which they are pointed. Microwave emission from glaciers (Mätzler, 1987) is always reported in terms of the brightness temperature $T_\mathrm{B} = \varepsilon T - T_a$, where ε is the emissivity of the medium (a mixture of ice, air, and possibly water), T is its physical temperature and T_a, often neglected, is the brightness temperature of the downwelling sky radiation. In dry snow, the emissivity is determined by volume scattering at interfaces such as grain boundaries and larger structures such as buried hoar layers. Grain growth rate depends on both temperature and accumulation rate, and it is possible to exploit this (Zwally, 1977) to model accumulation rate as a function of T_B and T. Estimates of c from the Zwally model compare well with estimates from surface measurements (Vaughan *et al.*, 1999).

Scattering at air-water interfaces is much more effective than at air-ice interfaces, and, when melting occurs at the surface, volume scattering ceases to be significant and the emissivity approaches unity. Abrupt changes are observed in T_B as meltwater comes and goes, and have been used to estimate the extent and duration of melting over the ice sheets (Abdalati and Steffen, 2001; Mote, 2003). Ramage and Isacks (2003) have exploited the same behavior to determine the start and end of the melt season on Alaskan glaciers.

Active-microwave sensors (scatterometers and imaging radars) measure the backscatter from pulses which they themselves emit. Like radiometers, scatterometers have poor spatial resolution (tens of kilometers or worse), although it can be improved by temporal averaging. Drinkwater *et al.* (2001) reported standard errors of $50-70\,\mathrm{kg\,m^{-2}\,a^{-1}}$ for a linear regression between climatological estimates of accumulation and NSCAT scatterometer data in Greenland. In the percolation zone of western Greenland, Wismann (2000) showed an excellent correlation between seasonally integrated reduction of backscatter, with respect to a reference wintertime average, and positive degree-days. Because of its buried ice lenses, resulting from internal accumulation, the percolation zone is one of the most radar-bright surface types on Earth when cold and dry, and one of the darkest when wet.

Synthetic aperture radars (SARs), unlike scatterometers, can resolve surface features as small as 5–100 m. They cover much less ground per megabyte of data, so that complete one-time coverage of an ice sheet is a major undertaking (e.g. Jezek, 2002). In SAR studies of small glaciers, the emphasis, to date, has been on delineating zones and on searching for the equilibrium line (e.g. Demuth and Pietroniro, 1999). Cogley *et al.* (2001) demonstrated a different approach to SAR observations, relying on browse images to identify the seasonal course of melting as a function of elevation. There was a clear dependence of change in image brightness upon positive degree-days accumulated between same-day image pairs from summer days. In a similar way, Nghiem *et al.* (2001) were able to map the extent of melt and refreezing in Greenland, relying on diurnal variations measured by the Quikscat scatterometer.

Microwave estimates of accumulation in the dry-snow zone are already among the best available, and estimates of melting in the percolation zone have shown great promise. However, the ablation zone lacks a distinctive seasonal microwave signature.

Interferometric Synthetic Aperture Radar In a suitable configuration (e.g. Madsen and Zebker, 1998), two SAR images of a ground target can be used to estimate its elevation z by interferometry. More important payoffs of InSAR can be realized once the effect of topography is removed: the measurement of surface velocity and the mapping of grounding lines. When the images are separated by one day, as in the ERS tandem mission, 1995–1996, errors in surface velocity are of the order of meters/year. Vertical motions can be resolved as well, and near to grounding lines the contribution of tidal flexure to the vertical motion can be evaluated. The downstream change in flexure is largest just downstream of the grounding line, so the number of interference fringes is greatest there. Accurately located grounding lines are valuable because

at the grounding line the ice thickness is a function of z by hydrostatic equilibrium, and the discharge q may be estimated by integrating thickness times velocity across the gate. Speckle tracking (Gray *et al.*, 2001) is a less accurate but more robust way of measuring velocity. It is analogous to feature tracking in optical imagery, although the conditions for interferometry must still be satisfied.

Undersampling

Sparse information is a problem common to nearly all mass-balance methods. We wish to know the balance at M points in space during N spans of time, but have measurements only at a smaller number m of points for a smaller number n of spans. In whatever way m, n and M, N are related, interpolation or extrapolation will be required. How do spatial and temporal variability affect our ability to interpolate or extrapolate? This question can appear in different guises. Altimeter measurements of elevation change over a few years, and interferometric estimates of ice discharge over a few days, need to be compared with estimates of accumulation over much longer spans. In ice cores, measurements can be well resolved in time, but they may measure spatial as well as temporal variability if sastrugi and snow dunes have been migrating across the core site. Global estimates of small-glacier mass balance rely on observations spanning 1–50 years from at most a few hundred glaciers, and some of the unmeasured glaciers are thousands of kilometers from the nearest measured glacier. Information on time and space scales of variability is essential in each of these contexts if the measurements are to be interpreted meaningfully, that is, if they are to be given accurate error bars. This statistical problem is comparable in magnitude with the observational challenges being addressed by new technologies.

Results

Small Glaciers

Measurements of mass balance are reported to the World Glacier Monitoring Service (WGMS), Zurich, which publishes biennial bulletins (Haeberli *et al.*, 2001) and quinquennial summary volumes (Haeberli *et al.*, 1998). Not all measurements find their way into the WGMS database. Cogley and Adams (1998) and Dyurgerov (2002) are among those who have published more complete compilations. Dyurgerov's is the most comprehensive, and

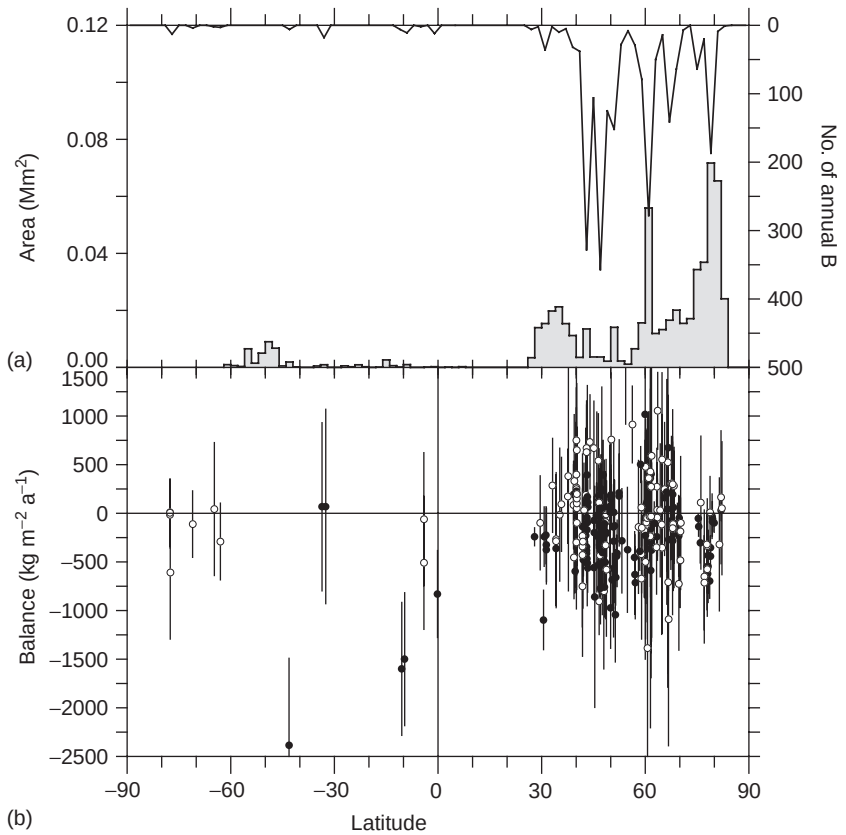


Figure 3 (a) Zonal distribution of small glaciers (shaded histogram, left axis; Antarctica excluded) and annual mass-balance measurements (thick line, right axis). (b) Average annual balances of measured glaciers, 1961–1990 (open circles: record length $n < 5$ years; solid circles: $n \geq 5$ years)

is available in spreadsheet form on CD-ROM. That of Cogley and Adams (1998; "CA" hereafter), covering only annual mass balance, may be obtained from **http://www.trentu.ca/geography/glaciology.htm.** Most reported measurements are direct measurements. Conventional geodetic measurements are unlikely to alter the picture greatly, although recent laser-altimetric measurements (Arendt *et al.*, 2002) have established more confidently that Alaska is a substantial contributor to sea-level rise.

Small glaciers occupy between 0.6 and 0.7 Mm^2, consisting of 0.539 Mm^2 plus about 0.070 Mm^2 in Greenland and an undetermined extent in Antarctica not belonging to the ice sheet. At present (2004) the CA dataset contains measurements from 310 glaciers dating back to 1885, although continuous measurements began only in the 1940s and a worldwide picture is only available after 1960. The spatial distribution of measured glaciers is uneven (Figure 3a), with high northern latitudes, Patagonia, and Tibet underrepresented at the expense of less remote regions (Europe and western North America). Glaciers with calving terminuses (19 of the 310) are not well represented. Most records are short. The modal length is 1 year, only 51 are 20 years or longer, and in no one year have as many as 100 glaciers been measured (Figure 4c). To set against this evidence of sparse coverage, there are about 2500 km^2 of ice per measured glacier, to be compared with about 30 000 km^2 of land per station for temperature climatologies. Moreover, CA note that three quarters of the small-glacier ice has at least one annual balance measurement within 400 km, and that the decorrelation distance for balance time series is about 600 km.

Spatial variations in mass balance are difficult to identify, at least at zonal resolution (Figure 3b). Allowing for uncertainties, most measurement series have averages indistinguishable from zero, but together they give a global arithmetic average of $-165 \pm 34\,kg\,m^{-2}\,a^{-1}$ for the reference period 1961–1990. (Here, and below, uncertainties are twice the standard error.) The error estimate is somewhat optimistic, and the balance estimate is biased by the neglect of internal accumulation in cold glaciers and of internal ablation in temperate glaciers, and other factors including possibly the spatial unevenness of the measurement network.

As records grow longer, however, the evolution of mass balance presents an increasingly coherent picture (Figure 4b). The world's small glaciers were close to equilibrium in the 1960s and have been losing mass since then at a growing rate. When spatial bias is corrected with an interpolation algorithm, the global average for 1961–1990 increases to $-123\,kg\,m^{-2}\,a^{-1}$, which is a best estimate.

Mass balance is well correlated with temperature (Figure 4a; $r = -0.79$ for the spatially corrected balance). The association would no doubt be closer if regional balances were matched to regional anomalies. Figures 4(a) and 4(b) help to justify the modeling of mass balance as a function of temperature (Wild *et al.*, 2003), and more importantly show that two independent measures agree in identifying the late twentieth century as a period of significant global change.

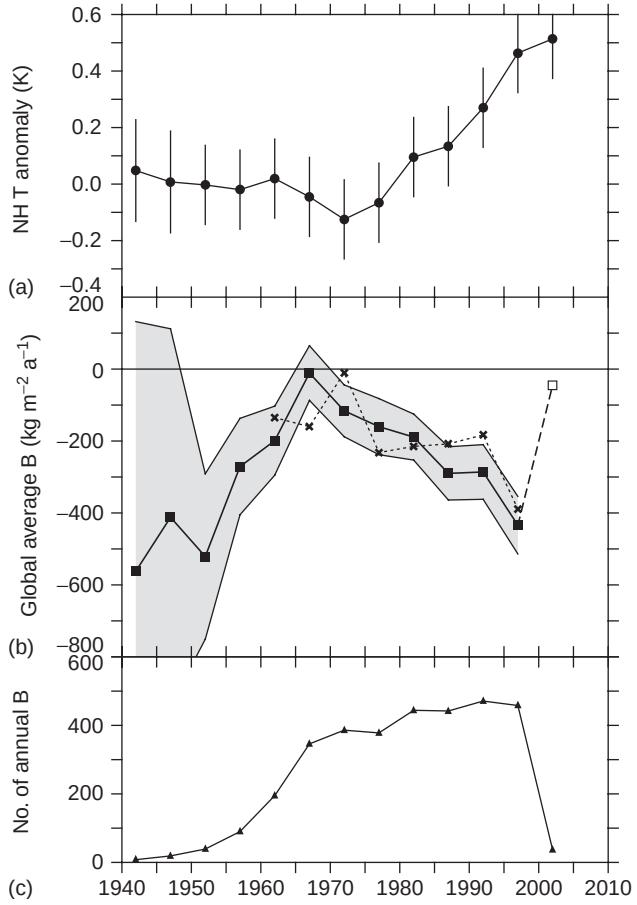


Figure 4 (a) Northern Hemisphere surface temperature anomalies; (b) small-glacier mass balance from the CA dataset, with shaded confidence region; crosses show effect of correcting for spatial bias; and (c) annual measurements contributing to each pentadal average balance (Jones and Moberg, 2003. © American Meteorological Society. **http://www.cru.uea.ac.uk/cru/data/temperature/**)

Greenland Ice Sheet

The ice sheets are too big for an integrated measurement of mass balance to be practical. Instead the aim is to compile the results of separate evaluations of each component.

For Greenland, accumulation is the best-known component. There are up to 400 column measurements, and three recent analyses interpolate to unmeasured parts of the ice sheet in different ways, hand contouring (Ohmura *et al.*, 1999) and kriging (Calanca *et al.*, 2000; Bales *et al.*, 2001). Figure 5 (Cogley, 2004) is constructed with another

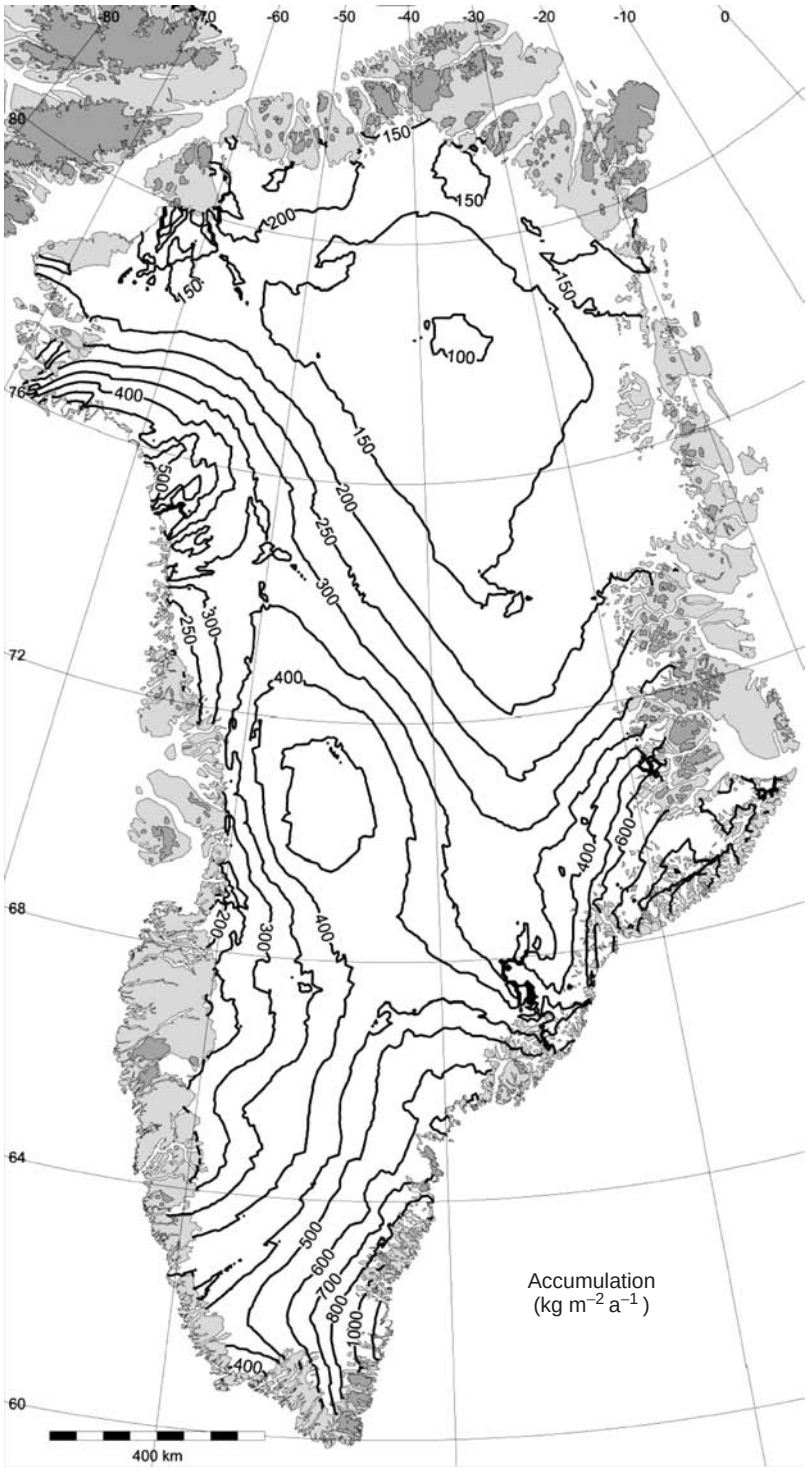


Figure 5 Accumulation on the Greenland Ice Sheet. The rate is below 100 kg m^{-2} a^{-1} in the northern interior and approaches 1400 kg m^{-2} a^{-1} in the southeast (Reproduced from Cogley (2004) by permission of American Geophysical Union)

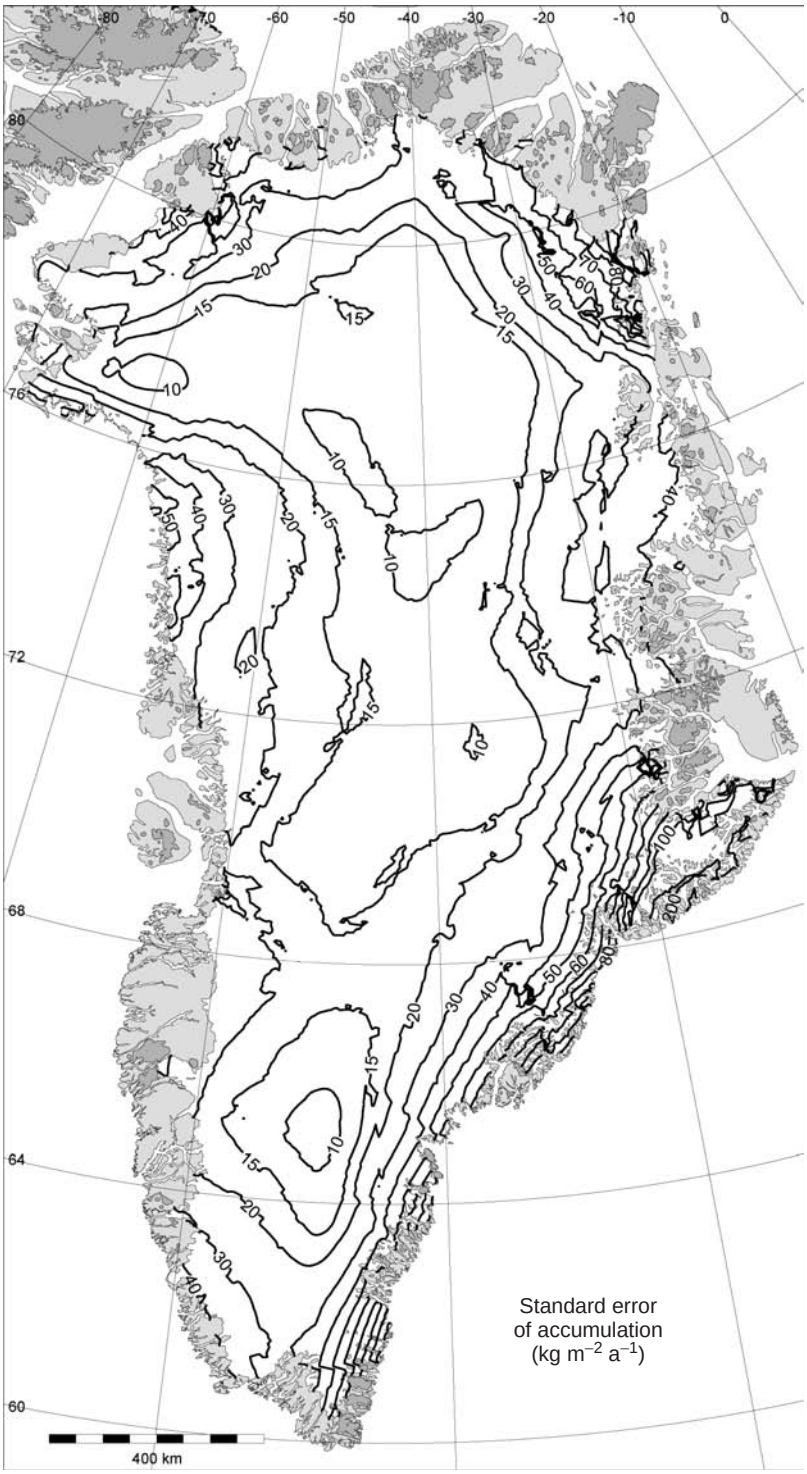


Figure 6 Standard error of accumulation on the Greenland Ice Sheet. Errors (for an assumed 30-year span) are as low as 8 kg m^{-2} a^{-1} in the interior and reach several hundred kg m^{-2} a^{-1} in places near the margin, where the interpolation algorithm falters for lack of information (Reproduced from Cogley (2004) by permission of American Geophysical Union)

interpolation algorithm. It is broadly similar to the other maps, but the algorithm has the advantage of generating formal estimates of the error at each interpolation point (Figure 6). The result is an accumulation estimate of $299 \pm 23\,\mathrm{kg\,m^{-2}\,a^{-1}}$, very close to the Ohmura, Calanca, and Bales estimates, 297, 290, and $305\,\mathrm{kg\,m^{-2}\,a^{-1}}$ respectively. Uncertainty is small in the interior of the ice sheet, where measurements are relatively abundant, and grows rapidly towards the edge where measurements are few (or nonexistent because the balance is negative and no snow survives).

The spatial and temporal variability of accumulation has of late received considerable attention (e.g. McConnell *et al.*, 2001), stimulated by advances in geodetic measurement technology which have led to improved estimates of surface elevation change (e.g. Krabill *et al.*, 2000; Davis *et al.*, 2000). The variability is such that the patterns of elevation change determined by altimetry can be understood mostly in terms of short-term fluctuations of the compaction rate and the accumulation rate (Braithwaite and Zhang, 1999) and spatial "glaciological noise".

According to the altimetry, average elevation change in the interior of the ice sheet is close to zero, implying $b = c + \Delta q \simeq 0$ because ablation a is near to zero. The kinematic measurements of Thomas *et al.* (2001) support this conclusion. The altimetry shows, however, that parts of the ablation zone are thinning rapidly.

The ablation zone occupies 10–15% of the Greenland Ice Sheet. Surface measurements are too few for a coherent picture to be drawn from them, and the most comprehensive current understanding of surface ablation derives from observations of melt extent and duration (Abdalati and Steffen, 2001) and from modeling. Mote's model (2003), which builds on passive-microwave observations of melt duration, yields a 12-year average of $-155\,\mathrm{kg\,m^{-2}\,a^{-1}}$ for meltwater runoff from the whole ice sheet. Wild *et al.* (2003) parameterized ablation as a function of temperature and surface elevation to obtain an equivalent estimate of $-152\,\mathrm{kg\,m^{-2}\,a^{-1}}$.

The estimates given so far imply a net surface mass balance of about $150\,\mathrm{kg\,m^{-2}\,a^{-1}}$. Zwally and Giovinetto (2000), using passive-microwave and thermal infrared satellite observations and a parameterization of meltwater runoff, estimated this quantity as $128\,\mathrm{kg\,m^{-2}\,a^{-1}}$, and summarized earlier estimates ranging between 97 and $211\,\mathrm{kg\,m^{2}\,a^{-1}}$.

It remains to evaluate the losses due to calving, or preferably the discharge q_{out} across the grounding line. Many outlet glaciers, particularly in north Greenland, have floating terminal sections. We would prefer the grounding-line flux because the floating ice has already made its contribution to sea-level rise, and because the basal mass balance of the floating terminuses is difficult to evaluate. Bigg (1999) estimated ablation due to calving using empirical relationships to predict calving velocity as a function of ice thickness or water depth. The latter were taken mostly from bathymetric charts. The estimates range from -73 to $-132\,\mathrm{kg\,m^{-2}\,a^{-1}}$ with an uncertainty estimated as $\pm 70\%$. Rignot *et al.* (1997) used InSAR to locate the grounding lines of 14 outlet glaciers of the northern Greenland Ice Sheet and to measure q_{out} as $136\,\mathrm{kg\,m^{-2}\,a^{-1}}$ from an area of $0.332\,\mathrm{Mm^2}$. The surface mass balance, estimated from a map of accumulation and a simple degree-day model of melting, was $113\,\mathrm{kg\,m^{-2}\,a^{-1}}$, so the total balance was $B = -25\,\mathrm{kg\,m^{-2}\,a^{-1}}$. The calving flux at the ice front was several times smaller than the grounding-line flux, requiring that basal melting of the floating ice be of the order of thousands of $\mathrm{kg\,m^{-2}\,a^{-1}}$. Rignot *et al.* (2001) found $B = -2\,\mathrm{kg\,m^{-2}\,a^{-1}}$ for a different but overlapping set of north Greenland glaciers. On 8 of 12 outlet glaciers, the grounding line retreated inland over 1992–1996. This is consistent with the laser-altimetric observations of near-coastal thinning, which also imply that the margin should be retreating where it is on land. There is some limited evidence for this (Sohn *et al.*, 1998).

Not all of the balance components are accompanied by detailed error estimates, but it seems likely that the mass balance of the Greenland Ice Sheet still cannot be distinguished reliably from zero. Krabill *et al.* (2000) estimated it as $-27\,\mathrm{kg\,m^{-2}\,a^{-1}}$, but this figure awaits analysis of errors and more complete documentation. Nevertheless priorities for future study are clear. Radar interferometry and laser altimetry both suggest that the closest attention should be given to lower altitudes of the ice sheet, where surface measurements are fewest and the energetics of melting and the dynamics of thinning need to be better understood.

Antarctic Ice Sheet

The Antarctic Ice Sheet is seven times the size of the Greenland Ice Sheet: $12.3\,\mathrm{Mm^2}$ of conterminous grounded ice as against $1.7\,\mathrm{Mm^2}$, plus about $1.6\,\mathrm{Mm^2}$ of ice shelf and ice rises.

Accumulation reaches less than $25\,\mathrm{kg\,m^{-2}\,a^{-1}}$ in the interior of East Antarctica and exceeds $2500\,\mathrm{kg\,m^{-2}\,a^{-1}}$ in the mountains of the Antarctic Peninsula (Turner *et al.*, 2002). Vaughan *et al.* (1999) compiled surface observations of accumulation and extrapolated them using a model based on microwave brightness temperature (Zwally, 1977) as a background field. They estimated accumulation to be $149\,\mathrm{kg\,m^{-2}\,a^{-1}}$ over the grounded ice and $166\,\mathrm{kg\,m^{-2}\,a^{-1}}$ over the entire ice sheet. Giovinetto and Zwally (2000) contoured the surface observations by hand and constructed from the contour map a grid of interpolates by eye. The resulting estimate of accumulation over the entire ice sheet was $159\,\mathrm{kg\,m^{-2}\,a^{-1}}$, to which they applied somewhat conjectural bulk adjustments totalling $-10\,\mathrm{kg\,m^{-2}\,a^{-1}}$ for loss by melting and deflation (wind scouring and sublimation) in coastal areas.

Although ablation by meltwater runoff is small in Antarctica, not all of the ice sheet has a positive surface balance. Bintanja (1999) reviewed the widely distributed areas of blue ice, where there is no snow at the surface. Mass balance minima range from $-350\,\mathrm{kg\,m^{-2}\,a^{-1}}$ in northern coastal areas where melting is significant to smaller magnitudes ($\sim -50\,\mathrm{kg\,m^{-2}\,a^{-1}}$) at higher elevations where sublimation is low because temperatures are low. Winther *et al.* (2001) estimated the extent of blue ice to be $0.12–0.24\,\mathrm{Mm^2}$. Over about half of this extent of the negative balance is due to melting and over the other half it is due to scouring and sublimation. These results suggest that blue-ice areas reduce the mass balance of the grounded ice sheet by perhaps $0.5–4\,\mathrm{kg\,m^{-2}\,a^{-1}}$, which is negligible given the present accuracy of accumulation estimates at the ice sheet scale. Surface ablation may not be negligible, however, in some of the smaller basins.

Rignot and Thomas (2002) reported ice fluxes obtained at grounding lines by InSAR and coupled these estimates with the Giovinetto–Zwally estimates of accumulation to yield the nearest approach to date to a whole-glacier estimate of B for catchments in Antarctica. The 33 catchments, labelled in Figure 7, cover $7.2\,\mathrm{Mm^2}$. Their ice discharges range from a negligible $3\,\mathrm{kg\,m^{-2}\,a^{-1}}$ (from the inactive Ice Stream C) to more than $1000\,\mathrm{kg\,m^{-2}\,a^{-1}}$ (Smith and DeVicq Glaciers in the Amundsen sector). Their balances (accumulation minus discharge) range from $131\,\mathrm{kg\,m^{-2}\,a^{-1}}$ (Ice Stream C) to well below $-500\,\mathrm{kg\,m^{-2}\,a^{-1}}$ (the small outlet glaciers in the Amundsen sector). The spatially variable picture thus presented is not the least significant contribution made by this work.

When pooled, the Rignot–Thomas measurements give a balance of $-3 \pm 4\,\mathrm{kg\,m^{-2}\,a^{-1}}$. If the Vaughan accumulation field is used in place of the Giovinetto–Zwally field, the balance becomes $6 \pm 4\,\mathrm{kg\,m^{-2}\,a^{-1}}$. It is not clear how much weight should be given to the reservations that led Rignot and Thomas to prefer the Giovinetto–Zwally field, and, by analogy with the error analysis for Greenland (Figure 6), the standard error of accumulation may not be as small as their chosen 5%. If it is larger, or if equal weight is given to the two accumulation fields, then the Rignot–Thomas results resemble all earlier estimates of Antarctic mass balance in showing no significant difference from zero. A further source of uncertainty, not considered by Rignot and Thomas, is the difference in time span between the accumulation maps, based on measurements covering up to several decades, and the quasi-instantaneous 1996 InSAR measurements of discharge. Too little is known about the temporal variability of the discharge of ice streams and outlet glaciers for us to be confident that the InSAR snapshots are good estimators of multidecadal averages; this is a question requiring more systematic attention.

Radar-altimetric surveys of thickness change (Wingham *et al.*, 1998; Davis *et al.*, 2001) agree with the InSAR/accumulation estimates in finding no significant change in the interior of East Antarctica. (This is a puzzle awaiting an explanation, for in a warmer world, as in Figure 4(a), the atmosphere should deliver more snow to Antarctica.) In the Amundsen sector of West Antarctica, the

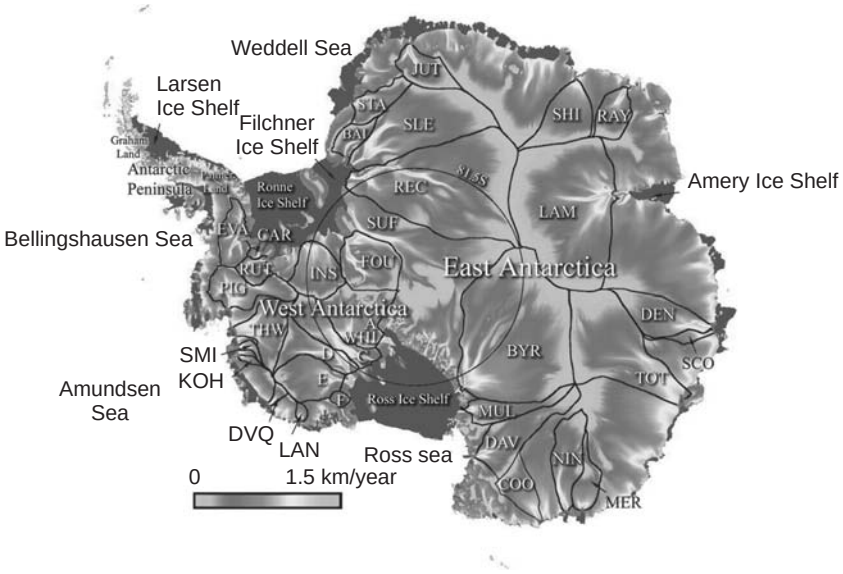


Figure 7 Major basins of Antarctica. The underlying field is the balance velocity (the column-averaged velocity $q_{\mathrm{out}}/\rho_{\mathrm{i}}$ which makes $b + q_{\mathrm{out}}$ zero in equation 3) (Reprinted from Rignot and Thomas, 2002. Mass balance of polar ice sheets. *Science*, **297**, 1502–1506. © 2002 American Association for the Advancement of Science). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

two methods also agree in finding a substantial negative balance. Shepherd *et al.* (2002) measured inland migration of the grounding lines and rapid thinning of Pine Island, Thwaites, and Smith Glaciers (Figure 7). They estimated the balances of Thwaites and Smith Glaciers as $-22 \pm 8\,\mathrm{kg\,m^{-2}\,a^{-1}}$ and $-233 \pm 34\,\mathrm{kg\,m^{-2}\,a^{-1}}$ respectively for 1991–2001, while Rignot and Thomas (2002) gave $-123 \pm 92\,\mathrm{kg\,m^{-2}\,a^{-1}}$ and $-698 \pm 222\,\mathrm{kg\,m^{-2}\,a^{-1}}$. The discrepancies obviously require investigation, but may be consistent with other evidence suggesting that the thinning began rather abruptly at some time during the 1990s. An abrupt onset would invite the speculation that the thinning is due to changes in ocean circulation leading to increased basal melting near the grounding line (Rignot and Jacobs, 2002).

The mass balance of the ice shelves is difficult to assess, mainly for lack of complete estimates of its components. The grounding-line flux estimates of Rignot and Thomas (2002) sum to $754\,\mathrm{Gt\,a^{-1}}$. If they account for the same fraction of shelf nourishment as the fraction of grounded ice from which they come, 59.3%, the total grounding-line flux would be $1271\,\mathrm{Gt\,a^{-1}}$. Vaughan *et al.* (1999) estimated the surface accumulation on the ice shelves as $478\,\mathrm{Gt\,a^{-1}}$. Several of the most northerly shelves have disintegrated in recent years, and surface melting has been implicated in these events, but quantitatively it makes little contribution to the balance; Jacobs *et al.* (1992) gave a crude estimate of $-36\,\mathrm{Gt\,a^{-1}}$. They also gave $-2016\,\mathrm{Gt\,a^{-1}}$ for the rate of calving, based mainly on iceberg censuses. From oceanographic observations, Jacobs *et al.* (1996) estimated the basal balance to be $-756\,\mathrm{Gt\,a^{-1}}$. The two inputs and the three outputs sum to $-1059\,\mathrm{Gt\,a^{-1}}$, or $-708\,\mathrm{kg\,m^{-2}\,a^{-1}}$.

Although there are scattered *in situ* and remotely sensed measurements, at present the only realistic way to get a broad view of the basal mass balance of an ice shelf is to model it using an ocean circulation model. For example, Williams *et al.* (2001) simulated the basal balance of Amery Ice Shelf to be $-254\,\mathrm{kg\,m^{-2}\,a^{-1}}$, which is substantially less negative than the Jacobs estimate of $-596\,\mathrm{kg\,m^{-2}\,a^{-1}}$. Nevertheless, the appropriate conclusion from the incomplete evidence appears to be that the ice shelves of Antarctica are losing mass at a rate that is very uncertain.

SUMMARY

Imprecise measurements, with insufficient spatial density and coverage, are a significant constraint upon what can be said about glacier mass balance in its global context. The cost of improving the methods and extending the scope of the measurements is high, but so are the probable costs of failure to understand mass and energy exchange between glaciers and the rest of the hydrosphere. For this reason, measurement technology is the subject of vigorous research and is improving rapidly. In the near future, the mass balance of the two ice sheets will become known with an uncertainty small enough to state with confidence whether they are growing or shrinking. At present the weight of evidence, some of it circumstantial, implies that both have negative mass balance, but the errors are not adequately quantified and for the Antarctic Ice Sheet, in particular, a conclusion about the sign of the mass balance would be premature. For the small glaciers, the picture is more clear. Their balances are negative on average and have been growing more negative since the 1960s, but here also there is scope for a more thorough assessment of errors.

The best balance estimates, in sea-level equivalents, are: for small glaciers, as calculated above, $0.21\,\mathrm{mm\,s.l.e.\,a^{-1}}$ with a poorly quantified uncertainty of the order of $0.06\,\mathrm{mm\,s.l.e.\,a^{-1}}$; for the Greenland Ice Sheet (Krabill *et al.*, 2000), $0.13\,\mathrm{mm\,s.l.e.\,a^{-1}}$ with no estimate of uncertainty; and for the Antarctic Ice Sheet (Rignot and Thomas, 2002; Vaughan *et al.*, 1999), between 0.10 and $-0.20\,\mathrm{mm\,s.l.e.\,a^{-1}}$ with an uncertainty of at least $0.13\,\mathrm{mm\,s.l.e.\,a^{-1}}$. The sum of these estimates, 0.14 to $0.44\,\mathrm{mm\,s.l.e.\,a^{-1}}$, is a small proportion of the contemporary rate of sea-level rise, 1.84–1.91 mm s.l.e. $\mathrm{a^{-1}}$ (Peltier, 2001; but see also Miller and Douglas, 2004, and references cited therein).

REFERENCES

Abdalati W. and Steffen K. (2001) Greenland ice sheet melt extent: 1979–1999. *Journal of Geophysical Research*, **106**(D24), 33983–33988.

Arendt A.A., Echelmeyer K.A., Harrison W.D., Lingle C.S. and Valentine V.B. (2002) Rapid wastage of Alaska glaciers and their contribution to rising sea level. *Science*, **297**, 382–386.

Bales R.C., McConnell J.R., Mosley-Thompson E. and Csatho B. (2001) Accumulation over the Greenland ice sheet from historical and recent records. *Journal of Geophysical Research*, **106**(D24), 33813–33825.

Bamber J.L., Ekholm S. and Krabill W.B. (2001) A new, high-resolution digital elevation model of Greenland fully validated with airborne laser-altimeter data. *Journal of Geophysical Research*, **106**(B4), 6733–6745.

Benson C.S. (1959) *Physical Investigations on the Snow and Firn of Northwest Greenland 1952, 1953, and 1954*, Research Report 26, U.S. Army Corps of Engineers Snow, Ice and Permafrost Research Establishment: Wilmette, p. 62.

Bhutiyani M.R. (1999) Mass-balance studies on Siachen glacier in the Nubra valley, Karakoram Himalaya, India. *Journal of Glaciology*, **45**(149), 112–118.

Bigg G. (1999) An estimate of the flux of iceberg calving from Greenland. *Arctic, Antarctic and Alpine Research*, **31**(2), 174–178.

Bintanja R. (1999) On the glaciological, meteorological and climatological significance of Antarctic blue ice areas. *Reviews of Geophysics*, **37**(3), 337–359.

Bintanja R. (1998) The contribution of snowdrift sublimation to the surface mass balance of Antarctica. *Annals of Glaciology*, **27**, 251–259.

Bintanja R. and Reijmer C.H. (2001) Meteorological conditions over Antarctic blue-ice areas and their influence on the local surface mass balance. *Journal of Glaciology*, **47**(156), 37–50.

Braithwaite R.J. (1984) Can the mass balance of a glacier be estimated from its equilibrium-line altitude? *Journal of Glaciology*, **30**(106), 364–368.

Braithwaite R.J. and Zhang Y. (1999) Relationships between interannual variability of glacier mass balance and climate. *Journal of Glaciology*, **45**(151), 456–462.

Calanca P., Gilgen H., Ekholm S. and Ohmura A. (2000) Gridded temperature and accumulation distributions for Greenland for use in cryospheric models. *Annals of Glaciology*, **31**, 118–120.

Chinn T.J. (1999) New Zealand glacier response to climate change of the past 2 decades. *Global and Planetary Change*, **22**, 155–168.

Church J.A., Gregory J.M., Huybrechts P., Kuhn M., Lambeck K., Nhuan M.T., Qin D. and Woodworth P. (2001) Changes in sea level. In *Climate Change 2001: The Scientific Basis*, Houghton J.R., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: New York, pp. 639–693.

Cogley J.G. (1999) Effective sample size for glacier mass balance. *Geografiska Annaler*, **81A**(4), 497–507.

Cogley J.G. (2004) Greenland accumulation: an error model. *Journal of Geophysical Research*, **109**(D18), D18101, doi:10.1029/2003JD004449.

Cogley J.G., Ecclestone M.A. and Andersen D.T. (2001) Melting on glaciers: environmental controls examined with orbiting radar. *Hydrological Processes*, **15**, 3541–3558.

Cogley J.G. and Adams W.P. (1998) Mass balance of glaciers other than the ice sheets. *Journal of Glaciology*, **44**(147), 315–325.

Cogley J.G., Adams W.P., Ecclestone M.A., Jung-Rothenhäusler F. and Ommanney C.S.L. (1996) Mass balance of White Glacier, Axel Heiberg Island, N.W.T., Canada, 1960–91. *Journal of Glaciology*, **42**, 548–563.

Davis C.H., Belu R.G. and Feng G. (2001) Elevation-change measurement of the East Antarctic Ice Sheet, 1978 to 1988, from satellite radar altimetry. *IEEE Transactions on Geoscience and Remote Sensing*, **39**(3), 635–644.

Davis C.H., Kluever C.A., Haines B.J., Perez C. and Yoon Y.T. (2000) Improved elevation-change measurement of the southern Greenland Ice Sheet from satellite radar altimetry. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(3), 1367–1378.

Demuth M.N. and Pietroniro A. (1999) Inferring glacier mass balance using RADARSAT: results from Peyto glacier, Canada. *Geografiska Annaler*, **81A**(4), 521–540.

Doake C.S.M. (1976) Thermodynamics of the interaction between ice shelves and the sea. *Polar Record*, **18**(112), 37–41.

Drinkwater M.R., Long D.G. and Bingham A.W. (2001) Greenland snow accumulation estimates from satellite radar scatterometer data. *Journal of Geophysical Research*, **106**(D24), 33935–33950.

Dyurgerov M.B. (2002) *Glacier Mass Balance and Regime: Data of Measurements and Analysis*, Occasional Paper 55, Institute of Arctic and Alpine Research, University of Colorado: Boulder, p. 87, 4 Appendices and CD-ROM.

Giovinetto M.B. and Zwally H.J. (2000) Spatial distribution of net surface accumulation on the Antarctic ice sheet. *Annals of Glaciology*, **31**, 171–178.

Gray A.L., Short N., Mattar K.E. and Jezek K.C. (2001) Velocities and ice flux of the Filchner Ice Shelf and its tributaries determined from speckle tracking interferometry. *Canadian Journal of Remote Sensing*, **27**(3), 193–206.

Haeberli W., Hoelzle M., Suter S. and Frauenfelder R. (1998) *Fluctuations of Glaciers 1990–1995*, Vol. VII, International Commission on Snow and Ice of International Association of Hydrological Sciences/UNESCO: Paris.

Haeberli W., Frauenfelder R. and Hoelzle M. (2001) *Glacier Mass Balance Bulletin No. 6 (1998–1999)*, International Commission on Snow and Ice of International Association of Hydrological Sciences/UNESCO: Paris.

Holland D.M. and Jenkins A. (1999) Modeling thermodynamic ice-ocean interactions at the base of an ice shelf. *Journal of Physical Oceanography*, **29**(8), 1787–1800.

Hubbard A., Willis I., Sharp M., Mair D., Nienow P., Hubbard B. and Blatter H. (2000) Glacier mass balance determined by remote sensing and high-resolution modelling. *Journal of Glaciology*, **46**(154), 491–498.

Jacobs S.S., Hellmer H.H. and Jenkins A. (1996) Antarctic ice sheet melting in the southeast pacific. *Geophysical Research Letters*, **23**(9), 957–960.

Jacobs S.S., Hellmer H.H., Doake C.S.M., Jenkins A. and Frolich R.M. (1992) Melting of ice shelves and the mass balance of Antarctica. *Journal of Glaciology*, **38**(130), 375–387.

Jezek K.C. (2002) RADARSAT-1 Antarctic Mapping Project: change-detection and surface velocity campaign. *Annals of Glaciology*, **34**, 263–268.

Jones P.D. and Moberg A. (2003) Hemispheric and large-scale surface air temperature variations: extensive revisions and an update to 2001. *Journal of Climate*, **16**(2), 206–223.

Kostecka J.M. and Whillans I.M. (1988) Mass balance along two transects of the west side of the Greenland Ice Sheet. *Journal of Glaciology*, **34**(116), 31–39.

Krabill W., Abdalati W., Frederick E., Manizade S., Martin C., Sonntag J., Swift R., Thomas R., Wright W. and Yungel J. (2000) Greenland Ice Sheet: high-elevation balance and peripheral thinning. *Science*, **289**, 428–430.

Krabill W.B., Thomas R.H., Martin C.F., Swift R.N. and Frederick E.B. (1995) Accuracy of airborne laser altimetry over the Greenland Ice Sheet. *International Journal of Remote Sensing*, **16**(7), 1211–1222.

Madsen S.N. and Zebker H.A. (1998) Imaging radar interferometry. In *Principles and Applications of Imaging Radar, Manual of Remote Sensing*, *Third Edition*, Vol. 2, Henderson F.M. and Lewis A.J. (Eds.), John Wiley: New York, pp. 359–380.

Mätzler C. (1987) Applications of the interaction of microwaves with the natural snow cover. *Remote Sensing Reviews*, **2**, 259–387.

Mayo L.R. (1992) Internal ablation – an overlooked component of glacier mass balance. *Eos*, **73**(43), 180, (abstract).

McConnell J.R., Lamorey G., Hanna E., Mosley-Thompson E., Bales R.C., Belle-Oudry D. and Kyne J.D. (2001) Annual

net snow accumulation over southern Greenland from 1975 to 1998. *Journal of Geophysical Research*, **106**(D24), 33827–33837.

Miller L. and Douglas B.C. (2004) Mass and volume contributions to twentieth-century global sea-level rise. *Nature*, **428**, 406–409.

Mote, T.L. (2003) Estimation of runoff rates, mass balance, and elevation changes on the Greenland Ice Sheet from passive-microwave observations. *Journal of Geophysical Research*, **108**(D2), 4056, doi:10.1029/2001JD002032.

Müller F. (1962) Zonation in the accumulation area of the glaciers of Axel Heiberg Island, N.W.T., Canada. *Journal of Glaciology*, **4**(33), 302–313.

Munk W. (2003) Ocean freshening, sea level rising. *Science*, **300**, 2041–2043.

Munro D.S. (2001) *Linking the Weather to Glacier Hydrology and Mass Balance at Peyto Glacier*. In Science Report 8, National Hydrology Research Institute, Environment Canada: Saskatoon, pp. 135–175.

Nakawo M., Raymond C.F. and Fountain A. (Eds.) (2000) Debris-covered glaciers. *International Association of Hydrological Sciences Publications*, Vol. 264, IAHS Press: Wallingford, p. 288.

Nghiem S.V., Steffen K., Kwok R. and Tsai W.-Y. (2001) Detection of snowmelt regions on the Greenland Ice Sheet using diurnal backscatter change. *Journal of Glaciology*, **47**(159), 539–547.

Oerlemans J. (2001) *Glaciers and Climate Change*, Balkema: Lisse, p. 148.

Ohmura A., Calanca P., Wild M. and Anklin M. (1999) Precipitation, accumulation and mass balance of the Greenland Ice Sheet. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **35**, 1–20.

Ohmura A., Konzelmann T., Rotach M., Forrer J., Wild M., Abe-Ouchi A. and Toritani H. (1994) Energy balance for the Greenland Ice Sheet by observation and model computation. *International Association of Hydrological Sciences Publications*, **223**, 85–94.

Oke T.R. (1987) *Boundary Layer Climates*, *Second Edition*, Routledge: New York, p. 435.

Østrem G. and Brugman M.M. (1991) *Glacier Mass-Balance Measurements: A Manual for Field and Office Work*, Science Report 4, National Hydrology Research Institute, Environment Canada: Saskatoon, p. 224.

Pälli A., Kohler J.C., Isaksson E., Moore J.C., Pinglot J.F., Pohjola V.A. and Samuelsson H. (2002) Spatial and temporal variability of snow accumulation using ground-penetrating radar and ice cores on a Svalbard glacier. *Journal of Glaciology*, **48**(162), 417–424.

Paterson W.S.B. (1994) *The Physics of Glaciers*, *Third Edition*, Elsevier Science: Tarrytown, p. 480.

Peltier W.R. (2001) Global glacial isostatic adjustment and modern instrumental records of relative sea level history. In *Sea-Level Rise*, Douglas B.C., Kearney M.S. and Leatherman S.P. (Eds.), Academic Press: San Diego, pp. 65–95.

Peltier W.R. (1998) Postglacial variations in the level of the sea: implications for climate dynamics and solid-Earth geophysics. *Reviews of Geophysics*, **36**(4), 603–689.

Pollack H.N., Hurter S.J. and Johnson J.R. (1993) Heat flow from the earth's interior: analysis of the global data set. *Reviews of Geophysics*, **31**(3), 267–280, [Data set available from `http://www.heatflow.und.edu/data.html`.].

Ramage J.M. and Isacks B.L. (2003) Interannual variations of snowmelt and refreeze timing on southeast-Alaskan icefields, U.S.A. *Journal of Glaciology*, **49**(164), 102–116.

Reijmer C.H. and Oerlemans J. (2002) Temporal and spatial variability of the surface energy balance in Dronning Maud Land, East Antarctica. *Journal of Geophysical Research*, **107**(D24), 4759, doi:10.1029/2000JD000110.

Richardson S.D. and Reynolds J.M. (2000) An overview of glacial hazards in the Himalayas. *Quaternary International*, **65**, 31–47.

Rignot E. and Jacobs S.S. (2002) Rapid bottom melting widespread near Antarctic ice sheet grounding lines. *Science*, **296**, 2020–2023.

Rignot E. and Thomas R.H. (2002) Mass balance of polar ice sheets. *Science*, **297**, 1502–1506.

Rignot E., Gogineni S., Joughin I. and Krabill W. (2001) Contribution to the glaciology of northern Greenland from satellite radar interferometry. *Journal of Geophysical Research*, **106**(D24), 34007–34019.

Rignot E.J., Gogineni S.P., Krabill W.B. and Ekholm S. (1997) North and north-east Greenland ice discharge from satellite radar interferometry. *Science*, **276**(5314), 934–937.

Selters A. (1999) *Glacier Travel and Crevasse Rescue*, *Second Edition*, The Mountaineers: Seattle, p. 143.

Shepherd A., Wingham D.J. and Mansley J.A.D. (2002) Inland thinning of the Amundsen Sea sector, West Antarctica. *Geophysical Research Letters*, **29**(10), doi:10.1029/2001GL014183.

Shumskiy P.A. (1955) *Osnovy Strukturnogo Ledovedeniya*, Izdatel'stvo Akademiy Nauk SSSR: Moscow, p. 492; Translated by Kraus D. (Ed.) (1964) as *Principles of Structural Glaciology*, Dover: New York, p. 497.

Siegert M.J. (2003) Glacial-interglacial variations in central East Antarctic ice accumulation rates. *Quaternary Science Reviews*, **22**(5–7), 741–750, doi:10.1016/S0277-3791(02)00191-9.

Slupetzky H. (1989) Die massenbilanzmessreihe vom Stubacher Sonnblickkees 1958/59 bis 1987/88. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **25**, 69–89.

Sohn H.-G., Jezek K.C. and van der Veen C.J. (1998) Jakobshavn Glacier, west Greenland: 30 years of spaceborne observations. *Geophysical Research Letters*, **25**(14), 2699–2702.

Su Z. and Shi Y. (2002) Response of monsoonal temperate glaciers to global warming since the Little Ice Age. *Quaternary International*, **97**, 123–131.

Thomas R.H., Csatho B., Davis C., Kim C., Krabill W., Manizade S., McConnell J. and Sonntag J. (2001) Mass balance of higher-elevation parts of the Greenland Ice Sheet. *Journal of Geophysical Research*, **106**(D24), 33707–33716.

Trabant D.C. and March R.S. (1999) Mass-balance measurements in Alaska and suggestions for simplified observation programs. *Geografiska Annaler*, **81A**(4), 777–789.

Turner J., Lachlan-Cope T.A., Marshall G.J., Morris E.M., Mulvaney R. and Winter W. (2002) Spatial variability of Antarctic Peninsula net surface mass balance. *Journal of Geophysical Research*, **107**(D13), 4173, doi:10.1029/2001JD000755.

van den Broeke M.R. (1997) Structure and diurnal variation of the atmospheric boundary layer over a mid-latitude glacier during summer. *Boundary-Layer Meteorology*, **83**(2), 183–205.

Vaughan D.G., Bamber J.L., Giovinetto M., Russell J. and Cooper A.P.R. (1999) Reassessment of net surface mass balance in Antarctica. *Journal of Climate*, **12**(4), 933–946.

Velicogna I. and Wahr J. (2002) A method for separating Antarctic postglacial rebound and ice mass balance using future ICESat Geoscience Laser Altimeter, Gravity Recovery and Climate Experiment, and GPS satellite data. *Journal of Geophysical Research*, **107**(B10), 2263, doi:10.1029/2001JB000708.

Wagnon P., Sicart J.-E., Berthier E. and Chazarin J.-P. (2003) Wintertime high-altitude surface energy balance of a Bolivian glacier, Illimani, 6340 m above sea level. *Journal of Geophysical Research*, **108**(D6), doi:10.1029/2002JD002088.

Warren S.G. (1982) Optical properties of snow. *Reviews of Geophysics and Space Physics*, **20**, 67–89.

Wild M., Calanca P., Scherrer S.C. and Ohmura A. (2003) Effects of polar ice sheets on global sea level in high-resolution greenhouse scenarios. *Journal of Geophysical Research*, **108**(D5), doi:10.1029/2002JD002451.

Williams M.J.M., Grosfeld K., Warner R.C., Gerdes R. and Determann J. (2001) Ocean circulation and ice-ocean interaction beneath the Amery Ice Shelf, Antarctica. *Journal of Geophysical Research*, **106**(C10), 22383–22399.

Wingham D.J. (2000) Small fluctuations in the density and thickness of a dry firn column. *Journal of Glaciology*, **46**(154), 399–411.

Wingham D.J., Ridout A.J., Scharroo R., Arthern R.J. and Shum C.K. (1998) Antarctic elevation change from 1992 to 1996. *Science*, **282**, 456–458.

Winther J.-G., Jespersen M.N. and Liston G.E. (2001) Blue-ice areas in Antarctica derived from NOAA AVHRR satellite data. *Journal of Glaciology*, **47**(157), 325–333.

Wismann V. (2000) Monitoring of seasonal snowmelt on Greenland with ERS scatterometer data. *IEEE Transactions on Geoscience and Remote Sensing*, **38**(4), 1821–1826.

Zwally H.J. (1977) Microwave emissivity and accumulation rate of polar firn. *Journal of Glaciology*, **18**(79), 195–215.

Zwally H.J. and Giovinetto M.B. (2000) Spatial distribution of net surface mass balance on Greenland. *Annals of Glaciology*, **31**, 126–132.

Zwally H.J. and Li J. (2002) Seasonal and interannual variations of firn densification and ice-sheet surface elevation at the Greenland summit. *Journal of Glaciology*, **48**(161), 199–207.

166: Surface and Englacial Drainage of Glaciers and Ice Sheets

PETER NIENOW[1] AND BRYN HUBBARD[2]

[1] *Institute of Geography, University of Edinburgh, Edinburgh, UK*
[2] *Institute of Geography and Earth Sciences, University of Wales, Aberystwyth, UK*

Supraglacial drainage occurs wherever snowpack, firn, or ice at the glacier surface is at the pressure melting point and supplied with additional energy, thereby generating melt water. Energy sources vary, but net radiation is usually the dominant source, although inputs of rainwater can also provide large volumes of surface runoff. The surface melt water is routed through the snowpack, firn, and across the glacier surface according to the local hydraulic gradient. In general, routing of water through snow and firn is slow (10^{-4}–10^{-5} m s^{-1}), contributing to a significant lag between melt water production and runoff at the glacier snout, whereas flow across exposed ice surfaces is typically 3–5 orders of magnitude faster. Under certain conditions where parts of the snowpack, firn, and/or ice surface are below the pressure melting point, the melt water will refreeze. Otherwise, the melt water will be routed supraglacially to the glacier margin unless intersected by a pathway from the glacier surface to the glacier interior. Such pathways include crevasses and moulins, and in temperate glaciers, microscale englacial veins. Flow rates through the englacial system vary considerably according to the hydraulic efficiency of the route taken. The routing of melt waters through both supraglacial and englacial drainage systems therefore affects the runoff response of an ice mass to rain and melt water inputs. During the course of a melt season, the efficiency of routing through both systems evolves, thereby altering the runoff response time to input variations.

INTRODUCTION

The routing of melt water through a glacial system is controlled by three interlinked subsystems – the surface, englacial, and subglacial drainage systems. The structure within each of these subsystems determines the efficiency and the route by which waters originating at the glacier surface are delivered to the glacier snout. The runoff response of a given ice mass to an externally forced melt or precipitation event is, therefore, controlled by the structure of these subsystems, which can vary both between ice masses and at a given ice mass over time and space.

The glacial drainage system can be envisaged as a cascade whereby surface melt waters are delivered to the glacier terminus either directly, via marginal streams, or via englacial and subglacial pathways. The distribution of snow, firn, and exposed ice and the hydraulic properties of each of these components determine the structure and efficiency of the surface or *supraglacial* drainage system. Flow rates vary substantially, from slow percolation through the snowpack and firn, to rapid channelized and sheet flow across exposed glacier ice (Krimmel *et al.*, 1973). The presence of a snowpack or firn layer, therefore, dampens the delivery of surface melt water to the glacier terminus (Fountain, 1989; Fountain and Walder, 1998). If the supraglacial melt water encounters a permeable pathway from the surface into the glacier interior, it will enter the englacial system. The rate of flow through the englacial system varies dramatically between the vertical shafts or "moulins" that can deliver water quickly to the glacier bed and the vein network in temperate ice where melt waters may flow at only a meter per year (Wakahama *et al.*, 1973) or less (Raymond and Harrison, 1975). In addition, water may be stored for significant periods in crevasses and englacial voids.

The flow of water through supraglacial and englacial drainage systems is important for a number of reasons. The structure within both systems critically affects the rate at which waters at the glacier surface are delivered to

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the glacier snout. Both systems, therefore, modify the link between surface melt (and liquid precipitation) and runoff (Jansson *et al.*, 2003). This is clearly important where melt waters from glaciated catchments are used for hydroelectric power and irrigation or where flood prediction is desirable. The catastrophic release of large volumes of supraglacially stored water has caused substantial loss of human life, and the mechanisms responsible for these flood events need to be better understood (Richardson and Reynolds, 2000). Whether melt water generated in a given year at the glacier surface is routed to the terminus or refreezes within the snowpack/firn has significant implications for glacier mass balance and for long-term sea-level fluctuations (Pfeffer *et al.*, 1991). Finally, the proportion of surface-derived melt waters that access the englacial drainage system and the pattern by which these waters are subsequently delivered to the glacier bed will impact on the structure of the subglacial drainage system and therefore on ice dynamics, sediment evacuation, and solute acquisition.

Firnification

The presence of firn plays a critical role in the storage and release of surface-derived glacial melt waters (Fountain, 1996; Jansson *et al.*, 2003) because of its porosity. Firn represents an intermediate or metamorphic state between "first-year" snow and glacier ice, and the rate and processes controlling the transformation from snow to ice are determined by the local accumulation rate and thermal regime. Firn becomes ice when the density reaches $830\,\mathrm{kg\,m^{-3}}$ (Paterson, 1994) and the air spaces become isolated, but the time taken for and the depth at which this transition occurs varies dramatically between ice masses. Thus, in the dry-snow zone of ice sheets such as at the Vostok Station in Antarctica, the firn-ice transition occurs after $\sim$2500 years at a depth of 95 m (Barnola *et al.*, 1987). The extremely slow transition is the result of both very low temperatures and accumulation rates. By contrast, the firn-ice transition in temperate glaciers can take place in less than 10 years at depths of $<$15 m (Sharp, 1951).

One of the key factors controlling the rate of firnification is the presence of water in the snowpack. In all snowpacks, compaction of individual crystals occurs during the initial stages of transformation, resulting in a more efficiently packed structure and greater density. Also, '*pressure-sintering*' increases compaction and reorganization of the crystal structure with a resultant reduction in porosity and the associated increase in density. In snowpacks where melting occurs, the presence of water dramatically increases rates of firnification through processes of regelation, diurnal freeze-thaw cycles, and the refreezing of percolating melt waters. The presence of water also implies that the snowpack is warmer than dry snow, so all processes of metamorphism occur faster. The most rapid transformation of snow to ice occurs when melt waters from a single summer percolate to the base of the snowpack and refreeze at the cold ice surface as an annual layer of superimposed ice. This process is most important in High Arctic ice masses such as Greenland in the "superimposed ice zone" immediately above the equilibrium line (Benson, 1962).

Because of its porous nature, the presence and thickness of the firn in the accumulation area of an ice mass will affect the potential for water storage. The movement of melt water through the glacier system will be delayed by temporary storage within the firn, as outlined in the following section.

Water Flow and Storage in Snow and Firn Reservoirs

The supraglacial snowpack and firn are aquifers that store water and delay runoff from glaciers. The role played by these aquifers in the hydrology of glaciers has been discussed at length (Fountain, 1996; Fountain and Walder, 1998; Schneider, 2000; Jansson *et al.*, 2003).

Water Flow and Storage in Snow

Movement of water through snowpacks is often complex because of the heterogeneity of snowpack conditions and varies temporally and spatially on a glacier and between glaciers. The processes controlling snowpack hydrology are generally the same in glacial and extraglacial snowpacks – the one common difference being the presence beneath supraglacial snowpacks of glacier ice that presents either a virtually impermeable or completely impermeable surface at temperate and polythermal/cold glaciers respectively. Marsh (*see* **Chapter 161, Water Flow Through Snow and Firn, Volume 4**) provides a more comprehensive summary of water flow through snowpacks.

Supraglacial snowpacks are often highly heterogeneous because of the complex stratigraphy associated with microscale variations in ice grain structure and the presence of macroscopic layering, both of which affect the porosity and hydraulic conductivity. The presence of impermeable ice layers, in particular, can significantly inhibit infiltration (Lang *et al.*, 1977; Echelmeyer *et al.*, 1992). Surface melt waters therefore percolate through the snowpack at different rates according to the route taken (Colbeck, 1991). Percolation rates are controlled by gravity and local degree of water saturation (in unsaturated conditions) and approximate to Darcy's Law (Colbeck, 1973) with velocities typically in the range 10^{-4}–$10^{-5}\,\mathrm{m\,s^{-1}}$ (Wakahama, 1968). These slow percolation rates through the snowpack mean that its presence delays the delivery of surface-derived melt waters to the englacial drainage system when compared to conditions where snow is absent. The seasonal thinning and removal of the supraglacial snowpack in the ablation zone therefore plays an important role in decreasing flow times from the glacier surface to the snout.

Melt waters reaching the base of the snowpack in the accumulation area typically percolate into the unsaturated firn whilst, in the ablation area, waters may: (i) enter the englacial system directly through a crevasse or moulin; (ii) flow over the ice surface; or (iii) become ponded in a saturated aquifer. The thickness of the aquifer depends on snow thickness, ice surface slope, melt water flux from the surface and the distribution of moulins and crevasses that act as sinks. Where surface gradients are very low and crevasses infrequent, the snowpack may become fully saturated and supraglacial "swamps" or slush zones can develop (Holmes, 1955; Hagen *et al.*, 1991). Stenborg (1970) observed slush layers up to 2-m deep at Mikkaglaciären in Sweden. Saturated snowpacks can destabilize, generating slush flows or avalanches that remove large volumes of supraglacially stored melt water and create levee-bounded channels. Such flows can be particularly spectacular at cold glaciers where reduced surface macropermeability can result in the development of vast slush fields.

Water Flow and Storage in Firn

In the accumulation zone, melt water percolates through the snowpack and into the unsaturated firn before building up as a saturated layer above the underlying low permeability glacier ice. The saturated layer develops after the onset of spring melt, and the thickness reached is dependent on the flux of waters from surface melt and rainfall, the hydraulic gradient in the firn, and the spacing and, to some degree, size of crevasses.

The thickness of the saturated layer varies both across and between glaciers with maximum values, for example, of 2 m and 7 m observed at South Cascade Glacier (Fountain, 1989) and Aletschgletscher (Schommer, 1978) respectively. The volumes of melt water stored within the firn can be significant and reduce the amplitude of diurnal runoff variations at the glacier snout. At South Cascade Glacier, Fountain (1989) found that storage in the firn accounted for $\sim$12% of the total spring storage, whilst at Storglaciären, 44% of total water storage was in the firn (Schneider, 2000). The hydraulic structure of the firn results in percolation rates with velocities in the order of $10^{-5}\,\mathrm{m\,s^{-1}}$ (Schneider, 2000). These velocities, and the subsequent temporary storage of waters in the firn acquifer, can induce a lag of days to weeks between supraglacial melt water generation in the accumulation area and its arrival at the glacier snout (Lang *et al.*, 1979; Behrens *et al.*, 1982). The percolation rates also tend to dampen the effect of the surface melt-signal so that diurnal water level variations in the firn are limited (Jansson *et al.*, 2003). Once melt water inputs to the snowpack and firn cease at the end of the melt season, the saturated firn layer drains, contributing to runoff or storage elsewhere in the glacier in the early winter. The firn layer is typically fully drained by late November (Fountain, 1996).

Supraglacial Channel Systems and Lakes

Supraglacial Channel Systems

In the ablation zone, seasonal removal of the supraglacial snowpack exposes glacier ice that is essentially impermeable at the microscale. Melt water flows across exposed ice surfaces at much higher flow velocities than those that are observed in the firn or snowpack. Initially, waters may percolate slowly along crystal boundaries in the weathered surface ice, but they are then routed quickly across the surface in either a thin film akin to sheet flow or in supraglacial channels. Melt waters follow the line of steepest gradient and are routed to the ice margin, unless they first drain into crevasses or moulins. In temperate glaciers, a very small proportion of surface melt waters may drain into the englacial vein network.

The size and discharge of a given supraglacial stream is dependent on the surface melt rate and the area of the catchment basin, which itself is controlled by the surface topography and the distribution of crevasses and moulins. On temperate glaciers, supraglacial streams rarely flow for more than a kilometer before intersecting a crevasse. On cold or polythermal ice masses, large deeply incised (>3 m) supraglacial streams (Thomsen *et al.*, 1988; Bingham *et al.*, 2003) may flow for many kilometers because of the lack of crevasses, and individual catchment basins may reach up to several hundred square kilometers (Sugden and John, 1976). Because stream location is controlled by the surface topography of the ice mass, the position of the larger streams is relatively constant between melt seasons (Hagen *et al.*, 1991; Kohler, 1995). However, stream routing may change during a melt season if melt waters initially routed over crevasses or snow-plugged moulins are subsequently "captured" when englacial connections enable melt waters to drain into the glacier. This can result in the abandonment of down-glacier sections of the supraglacial drainage system.

Supraglacial channel networks often develop arborescent drainage patterns similar to terrestrial river catchments (Hagen *et al.*, 1991; Knight, 1999), and rectilinear patterns are also common resulting from structural control imparted by crevasse traces and longitudinal foliation (Sugden and John, 1976). Some workers suggest stream density decreases upglacier because of the decrease in ablation and thus runoff with increasing elevation and the up-glacier reduction in slope angles away from the glacier margin (Sugden and John, 1976). However, the spatially irregular removal of surface melt waters into crevasses and moulins ensures that there is no simple relationship between drainage density and location on most ice masses.

Supraglacial streams usually meander (Knighton, 1972) and have similar hydraulic geometries to those observed in terrestrial river channels (Ferguson, 1973) with some longitudinal sections analogous to pool-riffle sequences (Carver *et al.*, 1994). Channel roughnesses are very low

with Manning's "n" of 0.01–0.012 (Chow, 1973) reflecting the low friction of the smooth ice-walls, and flow velocities are correspondingly high at about $1-2\,m\,s^{-1}$. The rapid velocities in supraglacial melt water channels ensure that melt water fluxes into moulins and crevasses have the same diurnal pattern as supraglacial melt-rates (Kohler, 1995; Nienow *et al.*, 1996). As the melt season progresses and the thinning snowpack reveals increasing areas of glacier ice, the time between peak melt rate and peak runoff from the glacier snout decreases (Elliston, 1973). In addition, as the snowpack is removed, the impact of rainfall events on glacier runoff is enhanced because of the rapid routing of water over the exposed impermeable ice surfaces (Østrem *et al.*, 1967).

Supraglacial Lakes

Supraglacial lakes form where melt waters gather within topographic depressions on the ice surface. These depressions and associated lakes are most common on large polythermal or "cold" glaciers where the stiffness of cold ice promotes flow-induced surface undulations or topographic ramps in response to bedrock obstacles (Rabus and Echelmeyer, 1997) and the surface ice is largely impermeable. In addition, since average surface gradients typically decrease as ice mass/glacier length increases (Raper, 2001), the incidence of depressions occupied by supraglacial lakes increases with ice mass size (as long as melt waters are available to fill the lakes). Thus, in Greenland where surface gradients away from the ice-sheet margin are very shallow, supraglacial lakes are common (Holmes, 1955), especially in the "slush" zones around the equilibrium line (Greuell, 2000; Greuell and Knap, 2000). In this zone, individual lakes can cover several square kilometers and may survive between melt seasons. However, supraglacial lakes are often transient features formed during the early melt season before draining either supraglacially, as melt waters overtop their retaining topographic barrier, or englacially as crevasses or snow-plugged moulins open (Liestøl *et al.*, 1980; Hagen *et al.*, 1991; Bingham *et al.*, in press). Ice-marginal lakes may also drain across glacier surfaces if the water level overtops the retaining topographic ice barrier or if snow-plugged supraglacial outlet channels become reactivated when the temporary snow dam becomes saturated and fails. Ice-marginal lakes that drain onto glacier surfaces are most commonly found beside polythermal glaciers (Hambrey, 1994) where the ice margin is frozen to the bed, thereby inhibiting drainage of the lake waters by subglacial routes.

Surface melt waters often fill crevasses during the early melt season, prior to the development of drainage links into the englacial system. Ponding of melt waters in crevasses exerts extreme forces at the base of the crevasse (Van der Veen, 1998) and may induce hydrofracture and crack propagation to the glacier bed. Work at John Evans Glacier in the Canadian High Arctic suggested that over a period of days, water ponded in a crevasse induced a series of fractures before the crack reached the ice-bed interface, whereupon the ponded melt waters drained rapidly (Boon and Sharp, 2003).

In areas where debris-rich glaciers stagnate (including the "stagnating" snout of surge type glaciers immediately postsurge), differential surface melt often results in the formation of numerous supraglacial lakes or "glacier karst" (Clayton, 1964; Reynolds, 2000). In recent years, warming trends and glacier recession in alpine glaciated landscapes have resulted in a dramatic increase in the occurrence of such lakes, particularly in parts of the Himalayas and Andes (Ageta *et al.*, 2000; Richardson and Reynolds, 2000). Catastrophic drainage of these lakes as glacier lake outburst floods (or GLOFs) can lead to severe loss of life and destruction in downstream areas. Richardson and Reynolds (2000) estimate that 32 000 people have been killed by GLOFs in Peru alone in the last century. The precise mechanisms that initiate GLOFs remain poorly understood and consequently, their occurrence is hard to predict. Mitigation of the hazard potential of GLOFs is typically predicated on on-site monitoring programs (Richardson and Reynolds, 2000). However, advances in remote sensing have demonstrated the potential for remotely monitoring the evolution of these supraglacial lakes (e.g. Wessels *et al.*, 2002).

Controls on the Character of Supraglacial Drainage Systems

The character of supraglacial drainage systems is principally determined by the permeability of the glacier surface and the climate that controls the development of the snowpack, the thermal regime of the glacier, and the volume of melt water generated during a melt season. Crevasses and moulins are the most efficient way of draining melt waters from the glacier surface. The distribution of crevasses therefore controls the extent to which supraglacially derived melt waters are routed across the glacier surface prior to entering the englacial drainage system. Glaciers containing a cold surface ice layer typically have streams flowing over the surface for greater distances before the melt waters are routed to the glacier margin or access the englacial system via crevasses or moulins (Iken, 1972; Bingham *et al.*, 2003; Boon and Sharp, 2003).

The character of supraglacial drainage systems is temporally and spatially variable and critically dependent on when and where on an ice mass surface runoff is generated. The presence or absence of a snowpack exerts a fundamental control on supraglacial drainage characteristics, as do variations in hydrological conditions within the snowpack which itself typically undergoes a seasonal evolution. Seminal work by Benson (1962) described distinct snowpack facies in the accumulation area of the Greenland

Ice Sheet, which can be characterized by their physical and thermal properties. Benson's classification is applicable to other ice masses (e.g. Müller, 1962) and the different facies are characterized by different supraglacial drainage conditions because of variations in both the availability of melt water and the possible hydrological flow paths in each zone. The typical hydrological conditions present in each of Benson's zones at the height of the summer melt season are outlined in the following sections.

Dry-snow Facies This zone is generally only found high in the interior of the Greenland and Antarctic ice sheets and on some very high mountains at lower latitudes. In this zone, no melting occurs and there is thus no supraglacial drainage.

Percolation Facies At lower elevations, some melting occurs for a brief period during summer. The melt water generated percolates into the snowpack or firn, where it refreezes as ice lenses or in layers at a particular depth. Whilst refreezing releases latent heat, not enough melt water is generated to saturate the whole snowpack and raise the temperature to 0 °C. There is therefore no runoff from the percolation zone.

Wet-snow Facies Within this zone, summer melt is sufficient to raise the whole of the snowpack to 0 °C for at least part of the melt season. Melt water flow is characterized by vertical percolation through the snowpack and subsequent lateral flow across impermeable ice layers in the snowpack or across glacier ice. The volume of runoff increases with decreasing elevation as less of the melt water refreezes within the snowpack and firn. In this zone, extensive slush fields can develop (Greuell and Knap, 2000), and the depth of saturation is dependent on the volume of melt, the surface gradient, and the prevalence of crevasses. This zone is also characterized by large supraglacial lakes and surface streams that incise down into the underlying snowpack and firn, and the melt waters usually drain englacially via moulins and crevasses. Thick layers of ice commonly form through refreezing at the end of the melt season and if this occurs at the snow-ice interface, superimposed ice is formed.

Superimposed Ice Facies This zone is located where all of the winter snowpack is melted, but some of the melt water refreezes as superimposed ice when it reaches the cold glacier ice. The lower edge of the superimposed ice zone represents the equilibrium line. Hydrological conditions are very similar to the wet-snow facies until all of the snow is removed.

Ice Facies In the ablation zone, warming of the winter snowpack as the melt season progresses raises the whole snowpack to 0 °C. Percolation with some refreezing is characteristic of the early melt season, and as the snowpack becomes increasingly saturated, slush flows can play an important role in removing significant volumes of snow. Once the snowpack is removed, runoff follows the steepest gradient over the bare ice surface, and supraglacial stream systems rapidly route the melt water to crevasses and moulins or the glacier margin.

In temperate glaciers, only the wet-snow and ice facies occur because summer melting warms the entire snowpack to melting point. The extent of ice (ablation zone) and wet-snow (accumulation zone) facies at the end of a given melt season is clearly controlled by variations in annual mass balance. As the proportion of exposed ice increases during a melt season, the lag-time between peak melt rate and peak runoff decreases, as the efficiency of the hydrological pathways increases.

Water Supply to, and Flow through, the Englacial Drainage System

Supraglacial water will enter a glacier and flow englacially upon encountering a permeable pathway. At the largest scale, such pathways are provided by surface crevasses that open up and close as ice passes through a crevasse field. Once exploited by melt water, however, notches incised by melting into the up-glacier edges of crevasses may be transformed into vertical tubes or *moulins*, which can remain as flow pathways long after the crevasse itself has closed (Stenborg, 1969). Consequently, moulins frequently occur as linear trains reflecting the orientation of the original crevasse from which they developed. Individual moulins can vary in size from under a meter to over 10 m across. Subhorizontal englacial channels may also form as a result of the incorporation of water present along the base of crevasses into the underlying englacial ice (Fountain and Walder, 1998). Similarly, basal fractures and crevasses can also propagate upwards from the glacier bed, providing effective transport pathways for the interchange of englacial and basal melt waters (e.g. Röthlisberger and Lang, 1987). Although supraglacial melting has the capacity to supply large fluxes of water to this channelized englacial system, water can also be produced within the body of the glacier itself. Melt water, for example, can be generated by frictional heating resulting from ice deformation and, nearer the glacier bed, basal sliding. Further, sliding by regelation, whereby temperate ice melts and refreezes in response to the difference in pressure melting point around a bedrock hummock, can induce melting some distance above the glacier bed (Lliboutry, 1993).

Polycrystalline ice is permeable at the microscale because individual ice crystals are separated by a network of water-filled veins. These veins form where three adjacent crystals intersect (so-called *triple-grain junctions*) and the veins themselves meet at *nodes*, formed at the intersections of four crystals (Nye, 1989) (Figure 1). In general, the larger the ice crystals (and these can range from less than 1 mm to

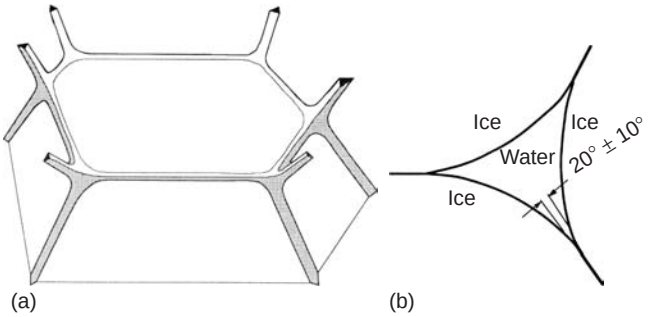


Figure 1 Ice crystals and inter-granular water veins and nodes illustrated conceptually (a) in three-dimensions and (b) in cross section (after Nye and Frank (1973) with the permission of the IASH)

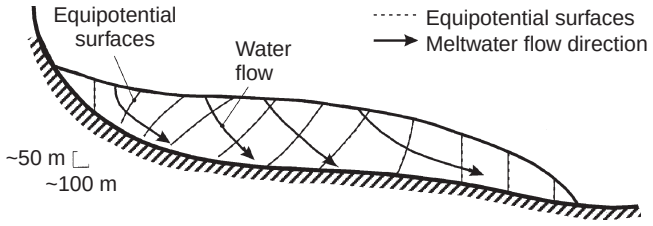


Figure 2 A conceptual glacier long section illustrating the general direction of englacial drainage (solid line) predicted from (Shreve, 1972) and hydraulic equipotentials (dashed lines) (after Paterson (1994) with the permission of Elsevier)

greater than 10 cm in diameter), the larger the veins between them. Veins are typically a few microns to tens of microns across. Laboratory-based research has been carried out on ice to measure the bulk permeability of vein-water systems at realistic temperatures. These studies generally indicate englacial flow rates in the region of 10^{-5} to 10^{-9} m s^{-1} (Wakahama *et al.*, 1973; Raymond and Harrison, 1975; Berner *et al.*, 1977). However, Lliboutry (1996) pointed out that these speeds are likely to be too high, since high confining pressures, the presence of debris and gas bubbles, and ice recrystallization will impede intergranular water flow in real glaciers relative to those measured in the laboratory. However, faster vein-water flow may occur near the glacier surface (where ice may be partially disintegrated because of melting and, below this, fine-grained) and near the glacier bed, where vein waters may be more saline and pressurized. For example, the bulk ionic concentrations measured in ice and vein water recovered from near the bed of Tsanfleuron Glacier, Switzerland, was ~11 times that typical of the overlying ice (Hubbard *et al.*, 2003). Further, relatively clear, bubble-free ice has been observed in cores recovered from close to the bed of many temperate glaciers. This *clear facies* ice is interpreted to result from gas expulsion along the intergranular vein system under the influence of intense deformation, indicating more active vein-water flow in this zone (Hubbard and Sharp, 1995). Hantz and Lliboutry (1983) identified an ~40-m thick layer of bubble-free "blue" ice in cores drilled to the base of Glacier d'Argentière, France. These authors also concluded from borehole water levels that bulk ice permeability increased substantially near to the glacier bed.

Despite the slow transfer speed of vein waters, this system is dense and widespread. Thus, although crevasses and moulins have the capacity to deliver large fluxes of melt water to the englacial drainage system, a more ubiquitous, but slower, delivery is achieved through the intergranular vein–water system. Indeed, the latter process may act as the only englacial meltwater source at glaciers on which little or no surface melting occurs.

Once present within the englacial system, melt waters generally flow towards the glacier bed and the glacier terminus. Arcone and Yankeilun (2000), for example, identified an englacial channel at Black Rapids Glacier, USA, which dipped consistently down-glacier at an angle of ~8° relative to the ice surface slope. Shreve (1972) argued that such flow is driven by two gravity-induced forces that can be combined to define a *water pressure potential* (ϕ). These forces are (i) the effect of gravity on the water itself (which drives "normal", unpressurized water flow) and (ii) the effect of gravity on the overlying ice which pressurizes the water flowing within the englacial channel. Application of this theory allows the 3-D equipotential surface, down which water is driven, to be defined from simple calculations based on glacier geometry (e.g. Sharp *et al.*, 1993) (Figure 2). One consequence of this analysis is that small surface englacial flow pathways may progressively capture water and enlarge into macroscopic flow pathways with depth (Raymond and Harrison, 1975).

If a glacier bed is overdeepened (i.e. the bed slope is reversed), the local pressure field may drive water away from the ice-bed interface and into an englacial position at the upglacier end of the overdeepening. This water would then reintersect the bed on the down-glacier lip of the overdeepening. In these cases, which are by no means rare (Fountain and Walder, 1998), basal water may become supercooled as it rises up (and possibly away from) the reverse slope of the overdeepening, resulting in the formation of refrozen basal ice (Alley *et al.*, 1998; Lawson *et al.*, 1998).

Although markedly more water is present in temperate ice than in cold ice, liquid water is by no means completely absent from the latter. Water, for example, is present in the intergranular vein system of ice whose bulk temperature is well below 0 °C (e.g. Mulvaney *et al.*, 1988). This is because the freezing point of vein waters is lowered as a result of the high concentration of soluble impurities within them. However, veins do close up as ice temperature decreases (e.g. Mader, 1992), and the porosity and permeability of cold ice are substantially lower than those of warm ice. One consequence of this contrast is that melt water flow

is inhibited at the boundary between warm ice and cold ice at polythermal glaciers. This thermal and hydrological contrast has been identified by airborne and surface radar (e.g. Bamber, 1987; Moore *et al.*, 1999) and may be closely associated with the location and propagation of surge waves at polythermal surging glaciers (e.g. Murray *et al.*, 2000).

Englacial Water Storage and Release

Clearly, any temperate or polythermal glacier's englacial drainage system is capable of storing a large amount of water at any given time. Our understanding of storage in intergranular veins and channels indicates that this is probably between 1 and 5% of the total volume of temperate ice within a glacier (Murray *et al.*, 2000; Moore *et al.*, 1999). In addition, evidence for temporal variations in storage is provided by water balance studies, which compare water inputs to a glacier with the outputs from it. Stenborg (1970), for example, calculated by this method that water was stored at Mikkaglaciären, Sweden, during the early summer and released during mid summer. Similar studies at other glaciers (e.g. Tangborn *et al.*, 1975; Willis *et al.*, 1993) also indicate early summer or summer storage. Importantly, Fountain and Walder (1998) calculated that subglacial storage could not account for the total volume of water stored at South Cascade Glacier, USA. These authors concluded that a substantial proportion of the total volume of water stored within the glacier is stored englacially. Such stored water is generally released through the winter or spring, sustaining outflow at times of negligible surface melting. The delivery of such water may be important not only hydrologically but also for glacier dynamics. Lingle and Fatland (2003), for example, argued that surges may be initiated by the delivery of stored water to a constricted basal drainage system during late winter or spring.

Methods of Investigating Englacial Drainage

Investigating englacial drainage is severely hampered by the limited access to the glacier interior. Thus, very little is known about the sizes and directions of englacial flow pathways or of the manner in which those pathways interact. However, a variety of techniques has contributed to our understanding of the character of englacial drainage.

Direct Observations

At the glacier surface, melt water can be observed to flow over the ice surface and to descend into crevasses and moulins (Figure 3). Thus, we know that these features transport large fluxes of water rapidly during the melt season. The integrity of a glacier's surface ice layer can also be observed to disintegrate to a depth of some tens to hundreds of millimeters under the influence of summer melting thereby bringing large volumes of melt water into contact with the upper surface of the englacial vein system.

Figure 3 Supraglacial melt water entering one of a train of moulins formed along a closed transverse crevasse on the surface of Mont Miné Glacier, Switzerland

However, while the capacity of the underlying vein system to transport this water has been measured on ice samples in the laboratory, it has not been measured directly in the field.

Moulins investigated by Holmlund (1988) descended almost vertically for some tens of meters below the glacier surface. Here, they ended in plunge pools and deviated laterally. Further exploration, however, was not possible as the channels narrowed. On the other hand, stones dropped from the glacier surface, particularly into large moulins, commonly indicate considerable depths of >100 m. The overall degree of lateral deviation of a moulin system was investigated by Iken (1972) on White Glacier, Axel Heiberg Island, Canada. The author lowered a pressure sensor into the water column in a moulin and simultaneously recorded both the length of line and the head of water above the sensor, and determined that the moulin dipped at a fairly consistent angle of $\sim 25^\circ$ from the vertical between the depths of 40 and 80 m below the ice surface.

Finally, in the ablation zone of glaciers, englacial channels occasionally intersect the ice surface as a result of ice ablation. Channels revealed in such situations are commonly some decimeters in diameter (Figure 4).

Figure 4 Englacial melt water conduit exposed in cross section by ablation at the surface of Llaca Glacier, Peru

Borehole-based Investigations

Over the past 30 years, numerous glaciological research teams have drilled boreholes into the glacier bed – allowing simultaneous access to the englacial and basal zones at multiple locations. Several techniques have been employed down such boreholes to investigate the character of the englacial drainage system.

Borehole Video TV footage has been recorded from video cameras moved along glacier boreholes. Pohjola (1994) reported on the basis of such a study at Storglaciären, Sweden, that ~1.3% of the glacier's thickness was composed of decimeter- to meter-scale englacial channels and voids. Later, Harper and Humphrey (1995), and Copland *et al.* (1997) drew similar conclusions on the basis of borehole video observations at Worthington Glacier, Alaska, and Haut Glacier d'Arolla, Switzerland, respectively.

Borehole Hydrology Many borehole-based studies used measurements of the head of water within individual boreholes to draw inferences about the water pressure in the subglacial drainage system to which it is connected (using the borehole as an open manometer). If these pressure data are associated with repeated electroconductivity (EC) profiles along the borehole water columns being measured, then englacial water sources and sinks can be identified and quantified. This is because water that has entered the borehole and englacial system from the glacier bed is generally of high EC while that entering the borehole and englacial system from the ice surface is generally of low EC. Gordon *et al.* (2001) used this approach at Haut Glacier d'Arolla, Switzerland, identifying complex patterns of interaction between the englacial and subglacial drainage systems. As a result of this study, the authors were able to identify five different types of borehole hydrological behaviour, or "borehole plumbing". Borehole behaviour patterns identified ranged from those that were characterized by a static, unchanging water column ("unconnected") to those that were simultaneously hydraulically connected to both the subglacial and the (channelized) englacial drainage systems ("basally connected").

Electrical Resistivity Tomography Electrical resistivity tomography (ERT) enables the internal resistivity distribution of a medium to be reconstructed from multiple resistance measurements made at its boundaries. Although the technique is best known for medical (Webster, 1990) and hydrogeological applications (e.g. Daily *et al.*, 1992), it has been used to image englacial drainage. Thus, Hubbard *et al.* (1998) identified an englacial hydraulic connection between two boreholes at Haut Glacier d'Arolla, Switzerland, by ERT.

Ice Radar

Ice radar involves transmitting waves of radio frequency into the glacier and investigating the properties of the waves that return, having been reflected off boundaries between materials with different *dielectric constants* or *relative electrical permittivities* within the ice. While ice has a permittivity of 3 to 4, freshwater has a permittivity of 80, making it is possible for ice radar to be used to locate major englacial water bodies and flow pathways. However, the presence of water throughout temperate glaciers means that radar signal penetration into such ice is poor relative to cold ice, although a small number of studies have successfully used ice radar to locate englacial (and subglacial) channels. For example, Jacobel and Anderson (1987) used stationary ice surface radar to identify temporal changes in the behaviour of englacial channels at Variegated Glacier, USA, while Arcone and Yankeilun (2000) identified and traced a single englacial channel using high frequency airborne radar at Black Rapids Glacier, USA.

Variations in the velocity of a propagating radar wave can be interpreted in terms of the water content of the ice through which that wave passes – allowing the bulk vein-water content of ice to be calculated. Murray *et al.* (2000) combined borehole and ice surface radar measurements at Falljökull, Iceland, and identified four ice layers with bulk water contents that varied between 0.2 and ~4%.

Tracing

Tracing studies involve injecting a tracer (normally fluorescent dye, although salt can be used) into the glacier's drainage system and recording its delivery from the glacier in its proglacial stream. If the recorded tracer concentration is plotted against time, the resulting *concentration breakthrough curve* yields information on the character of the flow pathways followed by the tracer. Properties of breakthrough curves include the time of arrival of peak tracer concentration and tracer dispersion. The former yields a travel time, which can be used to calculate average water velocity. The latter yields a measure of the

nature of the flow pathways followed, with more dispersed and multiple-peaked breakthrough curves reflecting more complex and/or multiple flow pathways. Many field-based studies have investigated these properties, but isolating the properties of the englacial drainage system from the subglacial drainage system is problematic. However, Stenborg (1969) and Behrens *et al.* (1975) isolated velocities of some tens of centimeters per second through the englacial drainage component, indicating the presence of channelized flow pathways from the glacier surface to the bed.

Ice Analysis

The physical character of ice recovered from either field cores or formed in the laboratory can also provide useful information relating to the nature of englacial drainage. For example, Wakahama *et al.* (1973), Raymond and Harrison (1975), and Berner *et al.* (1977) all carried out laboratory tests on ice to investigate the rate at which water flows through the intergranular vein system (reported previously). At a larger scale, Raymond and Harrison (1975) observed tubular conduits of some millimeters diameter in core samples recovered from 10 to 20 m beneath the surface of Blue Glacier, USA. Unfortunately, no further research of this kind has been carried out since this pioneering study.

Character of Englacial Drainage: Summary

As a result of combining the techniques and approaches outlined in this article, it is possible to draw some inferences about the character of englacial drainage systems. At the broadest scale, these systems can be classified in a manner similar to subglacial drainage – as discrete or distributed. Wherever water enters the ice surface through discrete channels, it continues to flow englacially at speeds similar to those of open-channel flow ($\sim 10^{-1}$ m s^{-1}). Much of this water enters the englacial drainage system during the melt season via moulins. Some of these form splash pools at a depth of some tens of meters and deviate laterally, while others appear to dip more regularly down-glacier. Little is known about the dynamics of englacial channels, but it is likely that they shrink as a result of ice deformation and grow (and tend towards the vertical) as a result of thermal erosion by melt water. The nature of individual englacial channels may therefore depend principally on local glacier dynamics and melt water flux. Since both controls vary seasonally at many glaciers, it is quite possible that moulin-fed englacial channels do so as well. A third probable control over the direction taken by englacial channels is that of the internal structure of the glacier. However, the association remains unknown.

Distributed englacial drainage comprises water present everywhere within the intergranular vein network, possibly supplemented by microscale channels at depth within the glacier. The vein-water system itself cannot transport water faster than some tens of centimeters per year, and cannot therefore be responsible for removing the bulk of surface-generated melt water at temperate glaciers. Although microscale channels have been observed in ice cores and decimeter-sized englacial voids have been identified from borehole studies, the coalescence of vein passages to form increasingly large channels with depth remains a tantalizing, but unproven hypothesis. Deep in the glacier, evidence suggests that waters can move away from a basal location into an englacial location through englacial channels, moulins and crevasses, and through the intergranular vein–water system. The latter system may open up in this zone as a result of the presence of pressurized water and higher ionic concentrations, both of which are strongly controlled by the character of the subglacial drainage system (*see* **Chapter 167, Subglacial Drainage, Volume 4**).

Acknowledgment

We would like to thank Martin Sharp and an anonymous referee for comments that resulted in significant improvements to the manuscript.

FURTHER READING

Behrens H., Bergmann H., Moser H., Rauert W., Stichler W., Ambach W., Eisner H. and Pessl K. (1971) Study of the discharge of alpine glaciers by means of environmental isotopes and dye tracers. *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **7**, 79–102.

Fowler A.C. (1984) On the transport of moisture in polythermal glaciers. *Geophysical and Astrophysical Fluid Dynamics*, **28**, 99–140.

Moore J.C., Mulvaney R. and Paren J.G. (1989) Dielectric stratigraphy of ice: a new technique for determining total ionic concentrations in polar ice cores. *Geophysical Research Letters*, **16**, 1177–1180.

Vaughan D.G. (1993) Relating the occurrence of crevasses to surface strain rates. *Journal of Glaciology*, **39**(132), 255–266.

REFERENCES

Ageta Y., Iwata S., Yabuki H., Naito N., Sakai A., Naruma C. and Karma (2000) Expansion of glacier lakes in recent decades in the Bhutan Himalayas. *International Association of Hydrological Sciences Publication*, **264**, 165–175.

Alley R.B., Lawson D.E., Evenson E.B., Strasser J.C. and Larson G.J. (1998) Glaciohydraulic supercooling: a freeze-on mechanism to create stratified, debris-rich basal ice: II. Theory. *Journal of Glaciology*, **44**(148), 563–569.

Arcone S.A. and Yankeilun N.E. (2000) 1.4 GHz radar penetration and evidence of drainage structures in temperate ice: black rapids Glacier, Alaska, USA. *Journal of Glaciology*, **46**(154), 477–490.

Bamber J.L. (1987) Internal reflecting horizons in Spitsbergen glaciers. *Annals of Glaciology*, **9**, 5–10.

Barnola J.M., Raynaud D., Korotkevich Y.S. and Lorius C. (1987) Vostok ice core provides 160,000-year record of atmospheric CO_2. *Nature*, **329**, 408–414.

Behrens H., Bergmann H. and Moser H. (1975) On the water channels of the internal drainage system of the Hintereisferner, Ötztal Alps, Austria. *Journal of Glaciology*, **14**(72), 375–382.

Behrens H., Oerter H. and Reinwarth O. (1982) Results from tracer experiments with fluorescent dyes on Vernagtferner (Oetztal Alps, Austria) from 1974–1982. *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **18**(1), 65–83.

Benson C.S. (1962) *Stratigraphic Studies in the Snow and Firn of the Greenland Ice Sheet*, Research Report 70, U.S. Snow, Ice and Permafrost Research Establishment.

Berner W., Stauffer B. and Oeschger H. (1977) Dynamic glacier flow model and the production of internal melt water. *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **13**, 209–217.

Bingham R., Nienow P. and Sharp M. (2003) Intra-annual and intra-seasonal flow dynamics of a High arctic polythermal valley glacier. *Annals of Glaciology*, **37**, 181–188.

Bingham R., Nienow P., Sharp M. and Boon S. (in press) Subglacial drainage processes at a high arctic polythermal valley glacier. *Journal of Glaciology*.

Boon S. and Sharp M. (2003) The role of hydrologically-driven ice fracture in drainage system evolution on an arctic glacier. *Geophysical Research Letters*, **30**(18), 1916, No. DOI: 10.1029/2003GL018034.

Carver S., Sear D. and Valentine E. (1994) An observation of roll waves in a supraglacial melt water channel, Harlech Gletscher, East Greenland. *Journal of Glaciology*, **40**(134), 75–78.

Chow V.T. (1973) *Open-channel Hydraulics*, McGraw Hill: London.

Clayton K.M. (1964) Karst topography on stagnant glaciers. *Journal of Glaciology*, **5**, 107–112.

Colbeck S.C. (1973) Effects of stratigraphic layers on water flow through snow, CRREL Research Report 311, U.S. Army Cold Regions Research and Engineering Laboratory.

Colbeck S.C. (1991) The layered character of snow covers. *Reviews of Geophysics*, **29**(1), 81–96.

Copland L., Harbor J. and Sharp M. (1997) Borehole video observation of englacial and basal ice conditions in a temperate valley glacier. *Annals of Glaciology*, **24**, 277–282.

Daily W., Ramirez A., LaBrecque D. and Nitao J. (1992) Electrical resistivity tomography of vadose water movement. *Water Resources Research*, **28**(5), 1429–1442.

Echelmeyer K., Harrison W.D., Clarke T.S. and Benson C. (1992) Surficial glaciology of Jakobshavn Isbrae, West Greenland: Part II. Ablation, accumulation and temperature. *Journal of Glaciology*, **38**(128), 169–181.

Elliston G.R. (1973) Water movement through the Gornergletscher. *International Association of Hydrological Sciences Publication*, **95**, 79–84.

Ferguson R.I. (1973) Sinuosity of supraglacial streams. *Geological Society of America Bulletin*, **84**, 251–256.

Fountain A.G. (1989) The storage of water in, and hydraulic characteristics of, the firn of South Cascade Glacier, Washington State, U.S.A. *Journal of Glaciology*, **39**, 143–156.

Fountain A.G. (1996) Effect of snow and firn hydrology on the physical and chemical characteristics of glacial runoff. *Hydrological Processes*, **10**(4), 509–521.

Fountain A.G. and Walder J.S. (1998) Water flow through temperate glaciers. *Reviews of Geophysics*, **36**(3), 299–328.

Gordon S., Sharp M., Hubbard B., Willis I., Smart C., Copland L., Harbor J. and Ketterling B. (2001) Borehole drainage and its implications for the investigation of glacier hydrology: experiences from Haut Glacier d'Arolla, Switzerland. *Hydrological Processes*, **15**(5), 797–813.

Greuell W. (2000) Melt-water accumulation on the surface of the Greenland ice sheet: effect on albedo and mass balance. *Geografiska Annaler*, **82**(A), 489–498.

Greuell W. and Knap W.H. (2000) Remote sensing of the albedo and detection of the slush line on the Greenland ice sheet, *Journal of Geophysical Research*, **105**(D12), 15567–15576.

Hagen J.O., Korsen O.M. and Vatne G. (1991) Drainage pattern in a subpolar glacier: Brøggerbreen, Svalbard. In *Arctic Hydrology: Present and Future Tasks*, Gjessing Y., Hagen J.O., Hassel K.A., Sand K. and Wold B. (Eds.), Norwegian National Committee for Hydrology: Oslo, pp. 121–131.

Hambrey M.J. (1994) *Glacial Environments*, UCL Press: London.

Hantz D. and Lliboutry L. (1983) Waterways, ice permeability, and water pressures at Glacier d'Argentière, French Alps. *Journal of Glaciology*, **29**(102), 227–239.

Harper J. and Humphrey N. (1995) Borehole video analysis of a temperate glacier's englacial and subglacial structure: implications for glacier flow models. *Geology*, **23**, 901–904.

Holmes G.W. (1955) Morphology and hydrology of the Mint Julep area, southwest Greenland. *Mint Julep Reports, Part II*, *Arctic Desert Topic Information Center U.S. Air University*: Pub. A-104-B.

Holmlund P. (1988) Internal geometry and evolution of moulins, Storglaciären, Sweden. *Journal of Glaciology*, **34**(117), 242–248.

Hubbard B., Binley A., Slater L., Middleton R. and Kulessa B. (1998) Inter-borehole electrical resistivity imaging of englacial drainage. *Journal of Glaciology*, **44**(147), 429–434.

Hubbard B., Hubbard A., Tison J.-L., Mader H.M., Nienow P. and Grust K. (2003) Spatial variability in the water content and rheology of temperate glaciers: Glacier de Tsanfleuron, Switzerland. *Annals of Glaciology*, **37**, 1–6.

Hubbard B. and Sharp M. (1995) Basal ice facies and their formation in the western Alps. *Arctic and Alpine Research*, **27**(4), 301–310.

Iken A. (1972) Measurements of water pressure in moulins as part of a movement study of the White Glacier, Axel Heiberg Island, Northwest Territories, Canada. *Journal of Glaciology*, **11**(61), 53–58.

Jacobel R.W. and Anderson S.K. (1987) Interpretation of radio-echo returns from internal water bodies in Variegated Glacier, Alaska, USA. *Journal of Glaciology*, **33**(115), 319–323.

Jansson P., Hock R. and Schneider T. (2003) The concept of glacier storage: a review. *Journal of Hydrology*, **282**, 116–129.

Knight P.G. (1999) *Glaciers*, Stanley Thornes: Cheltenham.

Knighton D. (1972) Meandering habit of supraglacial streams. *Geological Society of America Bulletin*, **83**, 201–204.

Kohler J. (1995) Determining the extent of pressurized flow beneath Storglaciaren, Sweden, using results of tracer

experiments and measurements of input and output discharge. *Journal of Glaciology*, **41**(138), 217–231.

Krimmel R.M., Tangborn W.V. and Meier M.F. (1973) Water flow through a temperate glacier. *International Association of Hydrological Sciences Publication*, **107**, 401–416.

Lang H., Leibundgut C.h and Festel E. (1979) Results from tracer experiments on the water flow through the Aletschgletscher. *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **15**, 209–218.

Lang H., Schadler B. and Davidson G. (1977) Hydroglaciological investigations on the Ewigschneefeld – Gr. Aletschgletscher. *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **13**, 119–124.

Lawson D.E., Strasser J.C., Evenson E.B., Alley R.B., Larson G.J. and Arcone S.A. (1998) Glaciohydraulic supercooling: a freeze-on mechanism to create stratified, debris-rich basal ice: I. field evidence. *Journal of Glaciology*, **44**(148), 547–562.

Liestøl O., Repp K. and Wold B. (1980) Supra-glacial lakes in Spitsbergen. *Norsk Geografisk Tidsskrift*, **34**, 89–92.

Lingle C.S. and Fatland D.R. (2003) Does englacial water storage drive temperate glacier surges? *Annals of Glaciology*, **36**, 14–20.

Lliboutry L. (1993) Internal melting and ice accretion at the bottom of temperate glaciers. *Journal of Glaciology*, **39**(131), 50–64.

Lliboutry L. (1996) Temperate ice permeability, stability of water veins and percolation of internal melt water. *Journal of Glaciology*, **42**(141), 201–211.

Mader H.M. (1992) The thermal-behavior of the water-vein system in polycrystalline ice. *Journal of Glaciology*, **38**(130), 359–374.

Moore J.C., Palli A., Ludwig F., Blatter H., Jania J., Gadek B., Glowacki P., Mochnacki D. and Isaksson E. (1999) High-resolution hydrothermal structure of Hansbreen, Spitsbergen, mapped by ground-penetrating radar. *Journal of Glaciology*, **45**(151), 524–532.

Müller F. (1962) Zonation in the accumulation area of the glaciers of Axel Heiberg Island, N.W.T., Canada. *Journal of Glaciology*, **4**(33), 302–313.

Mulvaney R., Wolff E.W. and Oates K. (1988) Sulfuric-acid at grain-boundaries in Antarctic ice. *Nature*, **331**(6153), 247–249.

Murray T., Stuart G.W., Fry M., Gamble N.H. and Crabtree M.D. (2000) Englacial water distribution in a temperate glacier from surface and borehole radar velocity analysis. *Journal of Glaciology*, **46**(154), 389–398.

Nienow P.W., Sharp M.J. and Willis I.C. (1996) Velocity-discharge relationships derived from dye tracer experiments in glacial melt waters: Implications for subglacial flow conditions. *Hydrological Processes*, **10**(10), 1411–1426.

Nye J.F. (1989) The geometry of water veins and nodes in polycrystalline ice. *Journal of Glaciology*, **35**(119), 17–22.

Nye J.F. and Frank F.C. (1973) Hydrology of the intergranular veins in a temperate glacier. *International Association of Hydrological Sciences Publication*, **95**, 157–161.

Østrem G., Bridge C.W. and Rannie W.F. (1967) Glacio-hydrology, discharge and sediment transport in the Decade glacier area, Baffin Island, NWT. *Geografiska Annaler*, **49A**, 268–282.

Paterson W.S.B. (1994) *The Physics of Glaciers*, *Third Edition*, Pergamon: Oxford.

Pfeffer W.T., Meier M.F. and Illangasekare T.H. (1991) Retention of Greenland runoff by refreezing: implications for projected future sea level change. *Journal of Geophysical Research*, **96**, 22117–22124.

Pohjola V.A. (1994) Tv-video observations of englacial voids in Storglaciaren, Sweden. *Journal of Glaciology*, **40**(135), 231–240.

Rabus B.T. and Echelmeyer K.A. (1997) The flow of a polythermal glacier: McCall Glacier, Alaska, U.S.A.*Journal of Glaciology*, **43**(145), 522–536.

Raper S. (2001) On factors that effect glacier response time and hence dynamic response, *Presented at the International Glaciological Society British Branch Meeting*, Cambridge, September 2001.

Raymond C.F. and Harrison W.D. (1975) Some observations on the behaviour of the liquid and gas phases in temperate ice. *Journal of Glaciology*, **14**(71), 213–233.

Reynolds J.M. (2000) On the formation of supraglacial lakes on debris covered glaciers. *International Association of Hydrological Sciences Publication*, **264**, 153–161.

Richardson S.D. and Reynolds J.M. (2000) An overview of glacial hazards in the Himalayas. *Quaternary International*, **65/66**, 31–47.

Röthlisberger H. and Lang H. (1987) Glacial hydrology. In *Glacio-fluvial Sediment Transfer – An Alpine Perspective*, Gurnell A.M. and Clark M.J. (Eds.), John Wiley and Sons: Chichester, pp. 207–256.

Schommer P. (1978) Rechnereische nachbildung von Wasserspiegelganglinien im Firn und Vergleich mit feldmessungen im Ewigschneefeld (Schweizer Alpen). *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **14**, 173–190.

Schneider T. (2000) Hydrological processes in the wet-snow zone of a glacier: a review. *Zeitschrift fur Gletscherkunde und Glazialgeologie*, **36**, 89–105.

Sharp R.P. (1951) Features of the firn on Upper Seward Glacier, St Elias Mountains, Canada. *Journal of Geology*, **59**, 599–621.

Sharp M., Richards K., Willis I., Arnold N., Nienow P., Lawson W. and Tison J.-L. (1993) Geometry, bed topography and drainage system structure of the Haut Glacier d'Arolla, Switzerland. *Earth Surface Processes and Landforms*, **18**, 119–134.

Shreve R.L. (1972) Movement of water in glaciers. *Journal of Glaciology*, **11**, 205–214.

Stenborg T. (1969) Studies of the internal drainage of glaciers. *Geografiska Annaler*, **51A**(1–2), 13–41.

Stenborg T. (1970) Delay of run-off from a glacier basin. *Geografiska Annaler*, **52**(A), 1–30.

Sugden D.E. and John B.S. (1976) *Glaciers and Glaciation*, Arnold: London.

Tangborn W.V., Krimmel R. and Meier M. (1975) A comparison of glacier mass balance by glaciological, hydrological, and mapping methods, South Cascade Glacier, Washington. *International Association of Hydrological Sciences Publication*, **104**, 185–196.

Thomsen H.H., Thorning L. and Braithwaite R.J. (1988) *Glacier-hydrological conditions on the Inland Ice North-East of*

Jakobshavn/Illulisat, West Greenland, Report 138, Grønlands Geologiske Undersøgelse.

Van der Veen C.J. (1998) Fracture mechanics approach to penetration of surface crevasses on glaciers. *Cold Regions Science and Technology*, **27**, 31–47.

Wakahama G. (1968) Infiltration of melt water into snowcover. *Low Temperature Science A*, **26**, 77–86.

Wakahama G., Kuroiwa D., Kobayashi O., Tanuma K., Endo Y., Mizuno Y. and Kobayashi S. (1973) Observations of permeating water through a glacier body. *Low Temperature Science A*, **31**, 217–219.

Webster J.G. (1990) *Electrical Impedance Tomography*, Adam Hilger.

Wessels R.L., Kargel J.S. and Kieffer H.H. (2002) ASTER measurement of supraglacial lakes in the Mount Everest region of the Himalaya. *Annals of Glaciology*, **34**, 399–408.

Willis I., Sharp M.J. and Richards K.S. (1993) Studies of the water balance of Midtdalsbreen, Hardangerjokulen, Norway: 2. Water storage and runoff prediction. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **27–28**(1991–1992), 117–138.

167: Subglacial Drainage

MARTIN SHARP

Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, AB, Canada

Subglacial drainage can occur wherever ice at a glacier bed reaches the pressure melting point. The subglacial drainage system is fed from a mixture of surface, englacial, subglacial, and groundwater sources that differ in terms of their spatial distribution and characteristic patterns of temporal variability. Subglacial drainage systems are not readily accessible, and knowledge of their characteristics is derived from a range of indirect methods including radio-echo sounding, the use of artificial tracers, monitoring, and manipulation of subglacial conditions via boreholes, and monitoring of glacial runoff properties. Subglacial water flow is driven by gradients in hydraulic potential, and occurs through either fast/channelized or slow/distributed systems located at the ice-bed interface, or through subglacial aquifers. Water can be stored subglacially in cavities, in the pore space of subglacial sediments, or in subglacial lakes. Drainage system structure evolves continually on various timescales in response to changing water inputs, evolving glacier geometry, and changes in glacier flow dynamics. Major hydrological events in such systems include the spring and fall transitions, outburst floods, and structural changes related to glacier advances and glacier surging. In the absence of variable water inputs from the glacier surface, subglacial drainage systems may be sensitive to forcing by earth, ocean, and atmospheric tides.

INTRODUCTION

Whenever ice at the base of a glacier or ice sheet reaches the pressure melting point, water may be present and a subglacial drainage system can develop. Subglacial water flow is of interest for a number of reasons, including its role in modulating glacier runoff, its potential as a source of geo-hazards, its influence on ice flow dynamics, and its role in chemical weathering (*see* **Chapter 168, Hydrology of Glacierized Basins, Volume 4** and **Chapter 169, Sediment and Solute Transport in Glacial Meltwater Streams, Volume 4**). If meltwater produced at the glacier surface penetrates to the glacier bed, the rate at which it is transported to the glacier terminus is determined in part by the character of the subglacial drainage system, which therefore mediates the relationship between surface melting and runoff. This is an important issue where glacier runoff has economic value (e.g. for the production of hydroelectricity) since efficient harvesting of the resource depends upon the ability to forecast runoff (Willis and Bonvin, 1995). Subglacial drainage systems may receive sudden inputs of large volumes of water from the drainage of supraglacial, ice-marginal, and subglacial lakes (Nye, 1976). The escape of this water from the glacier can be a significant hazard to population centers. Subglacial drainage can also exert a strong influence on glacier flow dynamics. Basal water reduces the friction at the ice-bed interface, affecting the rate of glacier flow by sliding (Iken, 1981) and the strength and deformation rate of unconsolidated subglacial sediments (Clarke, 1987a). Spatial variations in subglacial drainage system character may help explain the location of ice streams within large ice sheets (Bentley, 1987), and temporal variations have been implicated in the mechanics of glacier surging (Kamb, 1987). Subglacial water has access to abundant supplies of freshly ground and highly reactive "rock flour", creating a unique chemical weathering environment (Raiswell, 1984; Tranter *et al.*, 1993) (*see* **Chapter 169, Sediment and Solute Transport in Glacial Meltwater Streams, Volume 4**). The investigation of subglacial drainage has therefore become a major research focus in glacier hydrology.

SOURCES, SINKS, AND DIRECTION OF SUBGLACIAL DRAINAGE

Water Sources

Subglacial water has four main sources: surface, englacial, and basal melt, and subglacial groundwater. The relative

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

importance of these sources depends upon the climatic regime at the ice surface, the temperature of the glacier ice, ice flow dynamics and the nature of the glacier bed. On annual timescales, surface melt is usually the dominant source in "temperate" glaciers, where ice is almost all at the pressure melting point. Basal melt may, however, be the dominant subglacial water source in large areas of Antarctica where surface air temperatures never rise above the freezing point but ice temperatures reach the pressure melting point near the glacier bed. It may also be important in areas of high geothermal heat flux, such as Iceland. Englacial melt due to strain heating may be a significant water source in fast flowing glaciers such as Jakobshavn Isbrae, Greenland. Groundwater inputs have received little study, but they may be significant where a glacier rests on an underlying aquifer, especially in winter when surface melt ceases.

The Distribution and Magnitude of Water Inputs

The distribution of water inputs to a glacier bed depends upon the water source and thermal structure of the glacier. Basal melting (driven by geothermal heat and by frictional heat produced as the glacier slides over its bed) occurs throughout warm-based areas of a glacier, where ice at the bed is at the pressure melting temperature. This usually results in a relatively uniform distribution of melt inputs, because spatial variations in the geothermal heat flux usually occur on length scales greater than the size of individual glaciers. In some areas, like Iceland, however, there can be large spatial variations in the basal melt rate related to the distribution of active geothermal areas. There can also be significant spatial gradients in basal melt rate due to sliding friction or strain heating. Water produced within the glacier may drain to the glacier bed via a network of veins located at crystal boundaries in temperate ice (Nye and Frank, 1973). This also results in a relatively uniform distribution of water inputs to the bed. Most surface melt, however, becomes channelized subaerially and penetrates to the bed via crevasses or vertical shafts known as *moulins* (Holmlund, 1988) (*see* **Chapter 166, Surface and Englacial Drainage of Glaciers and Ice Sheets, Volume 4**). These localized inputs of water may be restricted to limited regions of the bed, depending upon the distribution and density of crevasses and moulins. This is especially true for nontemperate glaciers, where crevasses are often rare due to low rates of ice flow (Hodgkins, 1997).

The magnitude and variability of surface water inputs will depend on the density and exposure of input sites and on the nature of the melting surface. Surface snow and firn reservoirs have considerable potential to store water, and they act to delay melt-induced runoff and damp diurnal and meteorologically driven variations in meltwater flux (Fountain, 1996). These reservoirs produce slowly varying water inputs. Ice surfaces have comparatively little storage capacity, such that water inputs to the glacier bed from bare ice surfaces track the surface melt rate (Willis *et al.*, 2002). An exception to this generalization occurs on some cold glaciers, where there can be significant surface storage in lakes and channel systems (Liestøl *et al.*, 1980). This stored water may drain rapidly to the glacier bed as a result of water-pressure induced propagation of water-filled crevasses (Van der Veen, 1998; Boon and Sharp, 2003).

The Location and Direction of Subglacial Drainage

Subglacial drainage can occur either at the interface between ice and the underlying substrate, or within subglacial sediments. The pattern of water flow is influenced by the topography of the subglacial surface, and by horizontal and vertical gradients in water pressure. In general, this pattern is strongly influenced by the geometry of the overlying ice because basal water often supports a substantial fraction of the ice overburden. Where the substrate is permeable, vertical gradients in water-pressure influence the magnitude and sign of water exchange between substrate and ice-substrate interface. For the case of a glacier resting on an impermeable substrate, water flow will be perpendicular to contours of equal hydraulic potential at the glacier bed. One simplified view of subglacial drainage assumes that the drainage system is a hydraulically connected water sheet, in which water pressure is equal to the ice overburden pressure (Shreve, 1972). In this case, the hydraulic potential can be written:

$$\Phi = \Phi_0 + \rho_I g(z_s - z_b) + \rho_w g\ z_b \qquad (1)$$

Where Φ_0 is a reference potential, ρ_I and ρ_w are the densities of ice and water, g is the acceleration due to gravity, and z_s and z_b are the elevations of the ice surface and glacier bed, respectively. Shreve's model predicts that the slope of subglacial equipotential surfaces is $\sim$11 times the ice surface slope, but in the opposite direction. This implies that subglacial water can flow uphill out of overdeepened sections of glacier bed, so long as the magnitude of the water-pressure gradient exceeds that of the elevation potential gradient. Bedrock obstacles that are extensive in a direction transverse to ice flow will, however, likely cause substantial lateral diversions of water flow (Flowers and Clarke, 2002a), and water may refreeze if it emerges supercooled and quickly enough from over-deepenings (Alley *et al.*, 1998). Subglacial equipotential surfaces may contain depressions that define a closed interior drainage basin and become a focus of drainage, allowing the formation of subglacial lakes (Björnsson, 2002). Such depressions may be associated with localized sources of basal melting such as volcanoes. Changes in glacier geometry change the form of the subglacial hydraulic potential surface, allowing drainage capture and restructuring of subglacial drainage

catchments (Fountain and Vaughn, 1995). The assumption on which Shreve's formulation of the hydraulic potential is based may, however, break down in practice because subglacial drainage systems do not consist of hydraulically connected sheets. In reality, they contain discrete elements, such as channels, in which water pressure may differ from the ice overburden pressure and vary with discharge (Röthlisberger, 1972; Walder, 1986).

THE INVESTIGATION OF SUBGLACIAL DRAINAGE SYSTEMS

Direct observations of active subglacial drainage systems are limited because these systems are typically located beneath tens of meters to kilometers of glacier ice. A variety of remote sensing and other approaches has therefore been employed to determine their character and behavior.

Characteristics of Recently Deglaciated Surfaces

Observations of recently deglaciated glacier beds have shown that glaciers can rest on both rigid bedrock and unconsolidated, permeable, and potentially deformable sediments that vary substantially in terms of key hydrologic parameters such as thickness, continuity, hydraulic conductivity, and porosity. Studies of deglaciated carbonate surfaces (Figure 1) have provided information on the size, morphology, density, and network characteristics of formerly subglacial channel systems eroded into bedrock (Walder and Hallet, 1979; Hallet and Anderson, 1980; Sharp *et al.*, 1989). They have provided evidence for extensive ice-bed separation (cavity formation) on the downglacier side of bedrock bumps and steps (Figure 1), and revealed the widespread presence of laminated crusts composed of secondary calcite (Figure 1). These crusts are typically found on the lee side of small bedrock protuberances and are believed to have precipitated from a thin water film as a result of either refreezing of film waters or degassing of CO_2 rich solutions in the low-pressure region downstream of bed obstacles. The maximum size ($\sim$50 μm) of bedrock fragments trapped within individual laminae in the crusts sets an upper limit on the likely thickness of the water film (Hallet, 1976, 1979).

Such observations have motivated the development of several theoretical models of subglacial drainage (Walder, 1982, 1986; Kamb, 1987; Flowers and Clarke, 2002a,b). They are, however, limited in that they provide a temporally integrated view of drainage system structure and little insight into interactions between the various components of the drainage system (e.g. channels, cavities, film). Furthermore, the high solubility of carbonate minerals may mean that the structure of drainage systems developed on carbonate substrates is very different to that of systems developed on more resistant rock types.

Figure 1 Recently deglaciated carbonate bedrock surface at Glacier de Tsanfleuron, Switzerland, showing: (1) Roughened surface associated with an area of ice-bed separation (or cavity) on the downstream side of a bedrock obstacle. (2) Calcite deposits precipitated from the subglacial regelation water film. (3) A small Nye or ''N'' channel with solution scallops on its walls. (4) A polished and striated surface indicative of intimate ice-bedrock contact. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Radio-echo Sounding

Given the inaccessibility of active glacier beds, there has been considerable interest in remote sensing approaches to mapping their characteristics. Radio-echo sounding (radar) has been used to map the distribution of warm and cold-based ice using variations in the strength of the radar echo from the bed. The echo strength depends upon the dielectric contrast between ice and substrate, and is much greater for an ice-water interface than for an interface between ice and rock or dry, unfrozen sediment (Bentley *et al.*, 1998). Within warm-based areas of a glacier, the spatial pattern of bed reflection power may be indicative of the relative abundance of water at the bed (Copland and Sharp, 2001). To date, attempts to locate and map the distribution of individual channels at the glacier bed have only been successful in relatively thin ice (Moorman and Michel, 2000).

Radar has played a key role in mapping the size, shape, and distribution of subglacial lakes in Antarctica (Robin *et al.*, 1970; Oswald and Robin, 1973). Lakes are identifiable on radar records from a very strong, mirrorlike bed reflection that is constant in strength along track, suggesting a very smooth basal interface. In Iceland, radar mapping of ice thickness and bedrock topography

has allowed the reconstruction of subglacial equipotential surfaces beneath ice caps (Björnsson, 1982). Subglacial lakes, such as that in the Grimsvötn caldera beneath Vatnajökull, are located within internal drainage basins that can be identified from these surfaces.

Investigations with Artificial Tracers

Various tracers have been used to characterize subglacial drainage systems. These include salt solutions and fluorescent dyes, such as rhodamine or fluorescein (Burkimsher, 1983; Kamb *et al.*, 1985; Seaberg *et al.*, 1988; Willis *et al.*, 1990; Hock and Hooke, 1993; Nienow *et al.*, 1998). Where multiple streams enter and leave a glacier, the drainage catchment structure of the glacier can be mapped by delineating those areas of the glacier surface from which tracer inputs are recovered in each individual outflow stream (Sharp *et al.*, 1993). Measures of drainage system efficiency include the time of travel from input point to the glacier terminus, and the travel distance divided by the travel time (an index of flow velocity). If the discharge during the tracer test is known, the ratio of discharge to tracer velocity provides an index of the cross-sectional area of the drainage system. Since neither the channel sinuosity nor the water fluxes through individual channels are usually known, however, it is not possible to derive true measures of velocity or cross-sectional area. Nevertheless, the relationship between the indices of velocity and cross-sectional area may provide insight into the dynamics of the drainage system because the timescale of geometric adjustment of channels can be longer than that of water input variations. Thus, in a permanently water-filled conduit, variations in water flux on short timescales would be accommodated almost entirely by variations in flow velocity, while cross-sectional area would be relatively constant. By contrast, in a channel that was not full at low flows, the flow cross section would increase as discharge rose. Water flux increases would therefore be accommodated by a combination of increases in flow velocity and the cross-sectional area of the flow, resulting in a less sensitive relationship between flow velocity and discharge (Nienow *et al.*, 1996).

Continuous monitoring of dye concentrations in an outflow stream allows the construction of a dye return curve. This provides additional information about the character of the drainage system through which the tracer has passed. Curves with a single well-defined peak indicate that most dye was advected through a single major channel, whereas multipeaked curves are indicative of a more complex, multipath configuration (Willis *et al.*, 1990). Often, there is a correlation between peak shape and transit velocity, with single-peaked curves being associated with rapid transport and multipeaked curves with slower transport. Integration of the dye flux as a function of time allows calculation of the fraction of injected tracer that was recovered at the point of outflow. Typically, this fraction is much less than 1, indicating either significant retention of tracer within the glacier drainage system, or that much of the tracer emerged at concentrations below the fluorimetric detection limit. Retention may reflect diversion of water into subglacial storage or sorption of dye to fine-grained sediment (Bencala *et al.*, 1983).

As a tracer cloud passes through a drainage system, it becomes dispersed due to vertical and horizontal variations in water velocity and diversion of packets of tracer into storage sites such as backwater eddies. Dispersion results in reduced peak heights and increased peak widths. The rate of dispersion is described by the *dispersion coefficient*, D ($m^2\,s^{-1}$), but a more useful descriptor of peak characteristics is the *dispersivity* d(m), which is the ratio of the dispersion coefficient to the tracer velocity (Fischer, 1968). The dispersivity describes the rate of peak spreading relative to the rate of peak advection, and provides a characteristic length scale for the drainage system. Flow through major channels typically results in $d < \sim 10$ m, while more complex, less efficient systems result in much higher values ($> \sim 50$ m) (Seaberg *et al.*, 1988; Nienow *et al.*, 1998).

Borehole Manipulation and Monitoring

Boreholes, drilled with high-pressure hot water, provide direct access to glacier beds and are routinely employed in investigations of subglacial drainage. This has allowed measurements of the minimum thickness of subglacial sediment layers (by penetrometry; Hooke *et al.*, 1997; Harbor *et al.*, 1997), and *in situ* estimation of parameters that are important determinants of the hydraulic behavior of these sediments. These parameters include the porosity, hydraulic diffusivity and conductivity, and compressibility of the sediments. The techniques employed to measure them include analysis of the horizontal and vertical propagation and attenuation of natural pressure waves through subglacial sediments (Fountain, 1994; Hubbard *et al.*, 1995; Fischer *et al.*, 1998) and impulse-response tests (Stone and Clarke, 1993; Stone *et al.*, 1997; Kulessa and Murray, 2003).

Boreholes also provide an opportunity to measure water pressure and water quality within subglacial drainage systems. Boreholes sealed by freezing allow direct measurement of water pressure in the subglacial drainage system connected to the borehole. Water levels in open boreholes have also been used as a measure of subglacial water pressure on the assumption that the borehole functions as a manometer. Although there are potential problems with this assumption, a number of studies have recorded spatially coherent patterns of water level variation across arrays of open boreholes, suggesting that at least some boreholes do behave approximately as manometers (Hubbard and Nienow, 1997).

The behavior of borehole water levels during and after the time when the drill contacts the glacier bed has

been used to characterize different morphologies of the subglacial drainage system. Boreholes differ in terms of their base water levels, the amplitude, and timescale of water level oscillations, and the phase relationship between these oscillations and glacier runoff (Murray and Clarke, 1995; Hubbard *et al.*, 1995; Smart, 1996; Gordon *et al.*, 1998). Boreholes that are close to drainage channels often display low minimum water levels and high amplitude diurnal water level fluctuations that are almost in phase with runoff. Boreholes sampling hydraulically resistant elements of the drainage system typically exhibit high background water levels with minimal variability. Water level fluctuations in such boreholes may be out of phase with those in boreholes located close to channels, indicating diurnal transfer of mechanical support for the glacier between areas close to and remote from major channels as channel water pressure varies (Murray and Clarke, 1995; Gordon *et al.*, 1998).

In situ measurements of basal water quality have focused on the electrical conductivity (EC) and turbidity of borehole waters (Stone *et al.*, 1993; Hubbard *et al.*, 1995; Stone and Clarke, 1996; Kavanaugh and Clarke, 2001). High conductivity is usually associated with waters that have a long subglacial residence time, while low conductivity and low turbidity are associated with surface meltwater. High turbidity indicates basal water and/or high flow velocities. Temporal variations in conductivity and turbidity can thus provide insight into interactions between drainage system components (Hubbard *et al.*, 1995) and processes of drainage reorganization (Gordon *et al.*, 1998; Kavanaugh and Clarke, 2001). For open boreholes in particular, interpretation of measurements made at the base of boreholes demands an understanding of the system geometry and patterns of water exchange between the borehole and the subglacial drainage system (Gordon *et al.*, 2001). There have also been attempts at *in situ* hydrochemical characterization of borehole waters (Tranter *et al.*, 1997, 2002). Such measurements provide an alternative means of identifying different components of the subglacial drainage system, on the assumption that different modal water chemistries reflect differences in parameters such as the degree of water-rock contact and access to supplies of dissolved gases.

Properties of Glacial Runoff

Several studies have used the characteristics of glacial runoff to make deductions about subglacial drainage systems. The form of the diurnal discharge hydrograph from a glacier changes systematically over the melt season (Röthlisberger and Lang, 1987). Advances in the timing of the daily discharge peak and increases in the amplitude of the diurnal cycle reflect increased efficiency of water transfer through the glacier (Richards *et al.*, 1996). In part, this is a consequence of removal of the supraglacial snowpack and its delaying effect on runoff of surface meltwater, but it may also reflect the development of large drainage channels at the glacier bed. In some cases, the two processes may be closely connected (Nienow *et al.*, 1998) (*see* **Chapter 168, Hydrology of Glacierized Basins, Volume 4**).

Glacial runoff typically shows strong seasonal variations in solute concentration and composition (Tranter *et al.*, 1993). Solute concentrations in water leaving glaciers are enhanced relative to concentrations in the snowpack and supraglacial streams, indicating significant solute acquisition from water-rock interaction at or near the glacier bed. The nature and extent of this interaction depend upon the duration of water-rock contact, the rock:water ratio in the drainage environment, the extent to which water has access to supplies of gaseous O_2 and CO_2, and the availability of other potential sources of protons such as organic carbon and sulfide minerals (Raiswell, 1984; Tranter *et al.*, 1993). Glacial runoff appears to acquire solute in two distinct types of environment. Environments characterized by long transit times, high rock:water ratios, and proton supply from oxidation of sulfides and/or organic carbon result in water with high solute concentrations. This water type dominates runoff in the winter, the early melt season, and periods of recession flow in summer, and may also be characteristic of sudden outbursts of subglacially stored water at any stage of the melt season (Anderson *et al.*, 1999; Wadham *et al.*, 2001). Such waters are inferred to have drained via a predominantly distributed drainage system (see below; Tranter *et al.*, 1993). More dilute waters that have acquired significant amounts of solute by rapid reactions with suspended sediment are characteristic of periods of high flow later in the melt season (Brown *et al.*, 1994). These waters appear to have drained primarily and rapidly through large subglacial channels, though they may have mixed with and diluted waters of the more concentrated type in the process (*see* **Chapter 169, Sediment and Solute Transport in Glacial Meltwater Streams, Volume 4**).

The suspended sediment content of meltwaters also provides insight into subglacial drainage system structure. Waters that pass slowly through distributed drainage systems are unable to entrain significant amounts of sediment even though it may be abundant within the drainage environment. Water passing through large channels may have more limited contact with sediment but is better able to entrain what is available. Thus, sediment evacuation increases as channelized drainage becomes established in summer, but sediment exhaustion may occur towards the end of the melt season (Swift *et al.*, 2002). Sudden releases of sediment may result from outburst events and reorganization of drainage systems at the glacier bed (Humphrey and Raymond, 1994; Wadham *et al.*, 2001) (*see* **Chapter 169, Sediment and Solute Transport in Glacial Meltwater Streams, Volume 4**).

TYPES OF SUBGLACIAL DRAINAGE SYSTEM

Two types of subglacial drainage system have been widely recognized, though each can assume a diversity of forms. These have been characterized as *distributed* or *slow* systems, and *channelized* or *fast* systems (Fountain and Walder, 1998; Raymond *et al.*, 1995). In addition, some glaciers rest on permeable sediments that function as a groundwater aquifer (Boulton *et al.*, 1995; Flowers and Clarke, 2002b).

In distributed/slow systems, drainage pathways are tortuous, anastomosing, and widely distributed across the bed. Such systems are hydraulically resistant, and changes in water flux result in large changes in flow cross section but little change in flow velocity, which is generally low. By contrast, in channelized/fast systems most of the water flux passes through a few major channels that have an arborescent structure and occupy only a small fraction of the bed. Such systems offer much less resistance to water flow, and changes in water flux result in major changes in flow velocity. In the steady state, distributed systems exhibit an inverse relationship between water pressure and discharge, while the reverse is true for channelized systems (Walder, 1986; Kamb, 1987; Walder and Fowler, 1994; Röthlisberger, 1972). Both drainage morphologies can exist beneath the same glacier, and they will interact with each other (Hubbard *et al.*, 1995; Alley, 1996). The relative importance of the two systems in transmitting the imposed water flux may change on timescales ranging from hours to decades (Kamb *et al.*, 1985; Nienow *et al.*, 1998).

Distributed/Slow Subglacial Drainage Systems

One of the earliest configurations proposed for a subglacial drainage system was a pervasive *water film* (Weertman, 1964). Such a film might transport water produced at the bed by geothermal heating and sliding friction, or waters involved in regelation (pressure-induced phase change) around small obstacles on the glacier bed (Hallet, 1979). Weertman (1972) argued that the film might be the primary means of subglacial drainage because the pressure distribution around incipient channels at the glacier bed would impede their ability to capture water from the film. Since this argument only holds for a bed that is planar, impermeable, and free of debris, however, it is probably not applicable in reality. Thin water films are inherently unstable in the face of perturbations in film thickness because rates of viscous energy dissipation increase with film thickness, enhancing the local melt rate of the overlying ice, and causing a tendency towards channelization of the flow (Nye, 1976; Walder, 1982). Thus, while films are likely components of many subglacial drainage systems, their role in water transport is probably minor.

Where a glacier slides over a rough substrate, ice may separate from the bed to form cavities downstream of bedrock obstacles (Lliboutry, 1968) (Figure 1). Two types of cavity have been recognized: *autonomous cavities*, which contain ponded meltwater but transmit no water flux, and *interconnected cavities* which form an active part of the subglacial drainage system (Lliboutry, 1976). The status of individual cavities probably changes over time in response to changes in the imposed water flux and ice dynamics (Iken and Truffer, 1997). Nonarborescent cavity networks, linked by small bedrock (or *Nye*) channels oriented parallel to ice flow and incised up to 0.2 m into the glacier bed, have been mapped on recently deglaciated bedrock surfaces (Figure 1) (Walder and Hallet, 1979; Hallet and Anderson, 1980; Sharp *et al.*, 1989). Analyses of the hydraulics of linked-cavity systems demonstrate that cavities open by sliding of the overlying ice and close by ice deformation (Walder, 1986; Kamb, 1987). Melting of the ice roof by energy dissipated by flowing water makes only a minor contribution to the enlargement of cavities. The major stimulus to cavity growth is a discharge-induced rise in water pressure within the cavity system. The positive relationship between water flux and water pressure in cavity systems (and other forms of distributed system) has the consequence that water tends to be at higher pressure in larger cavities than in smaller cavities. It thus tends to drain from large to small cavities. This equalizes water pressures and ensures that there is no tendency for master channels to develop within the cavity network, thus acting as a negative feedback on the development of efficient (fast) drainage systems. The tendency for cavities to grow in response to rising water fluxes means that cavity-based drainage systems have significant potential to store water.

Glaciers often rest on unconsolidated sediments, rather than bedrock. The subglacial sediment layer may vary substantially in thickness and continuity, and the sediment characteristics can range from coarse gravel to fine-grained silt and even clay. Porous and permeable subglacial sediment functions as a confined aquifer that can drain of some of the water that penetrates to or is produced at the glacier bed. Measured hydraulic conductivities of *in situ* subglacial sediments are in the range 10^{-4} to 10^{-9} m s^{-1} (Fountain and Walder, 1998). With these hydraulic conductivities, most calculations suggest that Darcian flow through aquifers of thickness $\sim$0.1 m under a hydraulic gradient of $\sim$0.1 is unable to transport discharges of the magnitudes typical of the summer melt season, or even the winter period (Boulton and Jones, 1979; Alley, 1989; Fountain and Walder, 1998). Thus, the aquifer may become saturated and a drainage system will develop at its upper surface. Elution of fines from the upper layers of the sediment (Hubbard *et al.*, 1995) may form a *macroporous horizon* at the surface of the sediment layer (Stone and Clarke, 1993; Fischer and Clarke, 1994). Alternatively, a network of interconnected *microcavities* may form on the downstream sides of larger particles protruding up through the sediment surface (Kamb, 1991).

Such a system could have an effective hydraulic conductivity several orders of magnitude higher than the values quoted above (Fountain and Walder, 1998), and would be capable of adjusting its porosity by dilation in response to water flux variations (Clarke, 1987a).

Alternatively, a system of *canals* may coexist with the subglacial aquifer (Walder and Fowler, 1994). It may be fed directly by englacial conduits, in which case much of the surface-derived runoff would drain through the canal system without entering the aquifer and the aquifer would play a limited role in overall drainage. The canals would have an ice roof and a sediment floor, and would grow by roof melting or basal erosion, and close by deformation of ice or sediment. Canal systems could adopt two very different configurations: an arborescent network incised upwards into the ice, or a nonarborescent network incised into the sediment. The type of system that develops depends upon the magnitude of the effective pressure within the canal system ($P_e = P_{ice} - P_{water}$) relative to some critical value (P_{crit}) which depends upon the hydraulic gradient and the relative creep properties of ice and sediment. Nonarborescent networks are expected when the hydraulic gradient is low (as under ice sheets and ice streams), while arborescent networks are more likely beneath valley glaciers where hydraulic gradients are high. A canal network may exist beneath Ice Stream B, Antarctica (Engelhardt and Kamb, 1997).

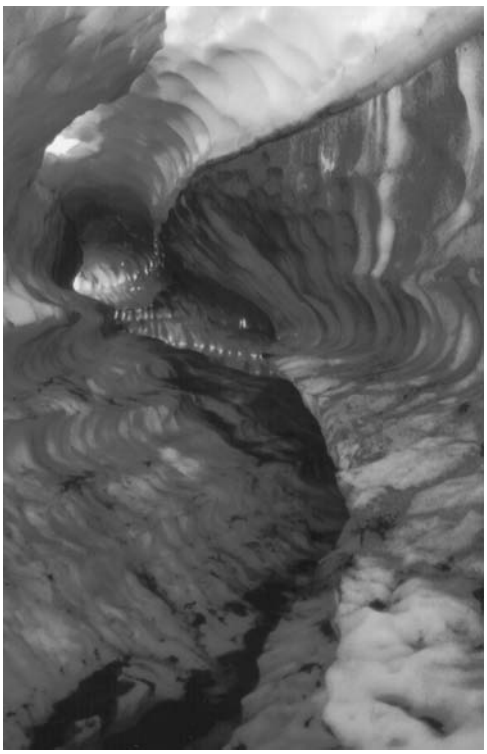


Figure 2 Röthlisberger or ''R'' channel at Matanuska Glacier, Alaska. Note scalloping and sediment deposits on the channel walls. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Channelized/Fast Subglacial Drainage Systems

Large ice-walled conduits (*Röthlisberger or "R" channels)* are the major component of channelized or fast subglacial drainage systems (Figure 2). They form where the rate of wall melting due to viscous dissipation of energy in flowing water outstrips the rate of creep closure of the overlying ice. The original analysis of the hydraulics of R-channels assumes steady water flow through conduits of semicircular cross section (Röthlisberger, 1972). This analysis predicts an inverse relationship between the water flux through an R-channel and the pressure gradient that drives the water flow. This implies that when two parallel conduits transporting different water fluxes coexist, the water pressure at a given distance from the glacier terminus along the conduits will be greater in the conduit with the lower water flux. Water from the smaller conduit will therefore tend to be captured by the larger conduit, a process that can account for the arborescent character of R-channel networks.

In reality, flow through R-channels fed by surface melt is unlikely to be steady as discharge varies with surface melt rates on timescales that are shorter than those on which channel cross section can adjust by a combination of wall melting and creep closure. Thus, R-channels may cease to be water-filled at low flows, and water pressures may rise to values in excess of the local ice overburden pressure during peak discharges (Hubbard *et al.*, 1995). Although there have been attempts to calculate the conditions under which open channel flow might occur within subglacial conduits (Hooke, 1984), there are few data with which to validate the calculations. In general, however, open channel flow is most likely where ice is thin and steeply sloping, and at times when discharge has dropped substantially from a preceding period of high flow.

There is also evidence that the cross-sectional shape of R-channels may not be semicircular. Attempts to forecast the distribution of water pressure within subglacial conduits using Röthlisberger's (1972) analysis have often predicted water pressures substantially below measured values. This might be explained if channels are broad and low in cross section, since such channels would tend to close more readily than semicircular channels, resulting in higher average water pressures (Hooke *et al.*, 1990). Conduits with this geometry might be expected because ice melt would be concentrated low down on channel walls at times of low discharge. In addition, creep of ice into partially full channels would be most rapid at the tunnel ceiling (especially when channels are only partly full), rather than at the base of the tunnel walls where creep is resisted by friction with the glacier bed. These arguments are supported by modeling of the evolution of channel cross section in response to time-varying water fluxes (Cutler, 1998).

Subglacial Groundwater Systems

In some cases, a complex hydrostratigraphy exists beneath a glacier, consisting of interbedded layers of contrasting grain size and hydraulic conductivity. Some of these layers may function as aquifers and others as aquitards. Water flow will tend to be predominantly horizontal within the aquifers and vertical across the aquitards (Flowers and Clarke, 2002a). In such cases, groundwater flow may represent an important sink or source for subglacial water.

A common case is that of a low conductivity till aquitard overlying a higher conductivity sand or gravel aquifer. For such a system, the pressure and elevation gradients across the aquitard drive the exchange between the glacier bed and the aquifer. Upward flow of groundwater commonly occurs where the glacier bed profile is concave (Tulaczyk *et al.*, 2000) and at the glacier margin (Flowers and Clarke, 2002a). The magnitude of the water flux into the aquifer is very sensitive to the aquifer conductivity, and it decreases as the aquitard and aquifer conductivities decrease. The mean water pressure and water fluxes at the glacier bed are therefore strongly influenced by the hydraulic conductivities of both the aquitard and the aquifer, decreasing as the conductivities rise. Thus, these properties of the subglacial sediments act as valves that control the routing of subglacial water. For most reasonable values of till conductivity, fluctuations in water pressure and flux recorded at the ice-bed interface will be transmitted to the aquifer, but with a discernible lag (Flowers and Clarke, 2002a).

Subglacial Lakes

Subglacial lakes can represent a significant form of subglacial water storage. 77 subglacial lakes have been identified beneath the Antarctic Ice Sheet (Siegert *et al.*, 1996). These are mostly located beneath regions of low surface slope close to ice divides, with major clusters in the Dome C and Ridge B regions. Those located away from ice divides are found close to regions where enhanced ice flow begins. The largest lake so far identified is the 230-km long Lake Vostok in East Antarctica (Kapitsa *et al.*, 1996), which has an estimated depth of at least 510 m, an area of $\sim$14 000 km^2, and a volume of 1500 to 7000 km^3. The mean observed length of the 77 lakes identified is, however, only 10.8 km (Siegert, 2000), and the total water volume stored in Antarctic subglacial lakes is likely in the range 4000–12,000 km^3 (Dowdeswell and Siegert, 1999). It is not known whether these lakes formed *in situ* beneath the ice sheet as a result of the progressive accumulation of water produced by basal melting, or whether they represent preglacial water bodies that were overridden by the growing ice sheet in the past. Little is known about whether and how these lakes drain under present-day conditions.

Subglacial lakes are common beneath ice caps in the volcanically active region of Iceland (Björnsson, 2002). High rates of heat flow associated with subglacial hydrothermal systems result in rapid melting of basal ice and a local depression in the ice cap surface. This can create closed depressions in the subglacial equipotential surface in which water can accumulate. Some lakes, such as Grimsvötn beneath Vatnajökull, fill caldera depressions in the bedrock, but the lake water level may rise above the caldera rim so long as the potential gradient continues to direct water flow towards the lake from surrounding areas. Other lakes form more rapidly as a result of extreme rates of basal melting during subglacial volcanic eruptions. These Icelandic subglacial lakes drain by episodic outburst floods (*see* **Chapter 168, Hydrology of Glacierized Basins, Volume 4**).

MAJOR EVENTS IN SUBGLACIAL DRAINAGE SYSTEMS

Spring Transitions

Where surface melting represents a significant input to subglacial drainage systems, there tends to be a strong seasonality to the magnitude of the water input. Subglacial drainage systems adjust to these variations in water flux and may thus have very different configurations in summer and winter. Thus, the subglacial drainage system will undergo major periods of evolution in spring and fall.

The driving force behind drainage evolution in spring is the progressive increase in the amount of surface meltwater delivered to the glacier bed. The temporal pattern of water inputs depends upon the end of winter snowpack distribution, meteorological conditions at the glacier surface, and the ease with which water can penetrate to the bed. Snow, with its high albedo, melts relatively slowly in comparison to glacier ice, and it acts as a potential storage site for water. As snow is removed from the glacier surface, a flood can result from the release of water stored in the snowpack (Flowers and Clarke, 2002b). Where drainage pathways from surface to bed are not immediately available, water is stored in supraglacial lakes, channels, and crevasses at the start of the melt season, and may be delivered to the bed very suddenly (Flowers and Clarke, 2000; Boon and Sharp, 2003). As glacier ice is exposed and subjected to surface melting, both the total daily runoff and the amplitude of the diurnal variation in runoff increase (Willis *et al.*, 2002). These changes in water input drive the evolution from a slow/distributed subglacial drainage system that is characteristic of winter, to a system that is usually dominated by fast/channelized components (Nienow *et al.*, 1998).

Drainage development in spring seems to occur in two phases. Because the drainage system is initially underdeveloped, the first water to penetrate to the bed forces

increased separation between ice and substrate not exclusively, but primarily, downstream from points where it reaches the bed (Mair *et al.*, 2002). This generates an increase in the subglacial drainage system volume that acts as a preferred pathway for subglacial drainage, and may allow significant water storage at the bed (Nienow *et al.*, 1998). As ice is exposed and surface meltwater production increases, water fluxes to the bed increase, water backs up in the drainage system, water pressures tend to rise, and major subglacial channels develop rapidly within regions of preferred drainage (Nienow *et al.*, 1998; Gordon *et al.*, 1998; Cutler, 1998). Rainstorms during this period may be a major stimulus to channel growth (Cutler, 1998; Gordon *et al.*, 1998), which is accompanied by the onset of strong diurnal water-pressure cycling in areas close to channels (Gordon *et al.*, 1998). Locally, channel growth is probably directed down-glacier from each point of water input. At a larger scale, however, the subglacial channel network extends upglacier over time at a rate that mirrors the retreat of the transient snowline. Where crevasses and moulins are distributed widely across a glacier, this can be a relatively smooth process. Where they are sparsely distributed, however, the process can be more episodic, with large areas of glacier bed developing channelized drainage in relatively short periods of time.

Where a frozen glacier margin effectively dams the outflow of basal water, a wave of high subglacial water pressures may spread upglacier from the margin in response to surface water input. Eventually, these high pressures are dissipated by either increased water flux into the subglacial groundwater system (Flowers and Clarke, 2002b), or by a sudden release of water at the glacier terminus (Skidmore and Sharp, 1999; Flowers and Clarke, 2000). In this case, channel development may follow the outburst.

End of Summer Transitions

The end of summer is marked by a decline in available melt energy. As water inputs decline, channels cease to be water-filled and begin to contract by ice creep (Cutler, 1998). Initially, water pressures tend to fall, allowing drainage of water previously stored in other components of the subglacial drainage system (Flowers and Clarke, 2002b). Eventually, the drainage system (both fast and slow components) contracts to the point that water pressures begin to rise again (Hubbard and Nienow, 1997), and the distributed/slow components come to dominate the system. In some areas, there is evidence that both the upstream and downstream ends of major channel systems may freeze shut very rapidly when surface melt ceases (Skidmore and Sharp, 1999; Copland *et al.*, 2003). This can preclude further input of surface water and trap water at the glacier bed. In other areas, including Svalbard, drainage continues throughout the winter (Hodgkins, 1997).

Hydrological Changes During Glacier Surges

Glacier surges involve periodic, abrupt one to three order of magnitude increases in glacier flow velocity that are sustained for months to years and separated by periods of quiescence lasting years to decades (Meier and Post, 1969; Raymond, 1987). Ice builds up in a reservoir area during quiescence and then is transferred rapidly down-glacier during surges. Rapid flow during surges is usually attributed to an increase in the basal velocity of the glacier. This results from a reduction in ice-bed coupling due to the buildup of high-pressure water at the bed (Clarke, 1987b). Two models have been proposed to explain how changes in ice flow dynamics might be associated with a reorganization of subglacial drainage (Clarke *et al.*, 1984; Kamb, 1987; Raymond, 1987), though available observations are currently insufficient to validate either model. In both models, transitions in drainage system character between the quiescent and surge phases are linked to the evolution of glacier geometry over a surge cycle. The buildup of ice in the reservoir area during quiescence results in local thickening of the glacier, an increase in the mean surface slope, and a rise in shear stress at the glacier bed (Raymond and Harrison, 1988).

One model of surging assumes the glacier rests on a largely impermeable substrate, and suggests that rapid surge phase motion occurs primarily by basal sliding. This model proposes that during surges, the basal drainage system is distributed in character and may consist of linked cavities (Kamb, 1987). During quiescence, this distributed system is replaced by a low-pressure system dominated by major subglacial channels. This temporal switching of drainage morphologies is analogous to seasonal changes but operates on longer timescales. This model of surging proposes that thickening of the reservoir area during quiescence increases the rate of closure of subglacial channels by ice creep. This drives up the channel water pressure (especially over winter when water fluxes are low) until it exceeds the mean water pressure in the surrounding drainage system. Once this happens, the channels lose water to the cavity system, initiating a positive feedback whereby decreasing water flux results in a further increase in water pressure and further water loss from the channels (Raymond, 1987). This process may be facilitated if the glacier contains sufficient englacial water storage capacity that downward drainage of englacially stored water over winter eventually overwhelms the constricted subglacial system and initiates rapid sliding (Lingle and Fatland, 2003). If the cavity system is sufficiently stable to withstand the following melt season's discharge perturbation, rapid sliding may be sustained (Kamb, 1987). If the surge is propagating down-glacier into thinner ice, water inputs into the drainage system under the surging region increase, and episodic releases of water occur (Kamb *et al.*, 1985). Initially, rapid ice creep shuts down channels that start to form during

these releases, but eventually the ice becomes too thin to allow this. Permanent drainage of water stored beneath the surging region occurs and the channel system becomes reestablished. Subglacial water pressure is lowered and the surge phase of motion comes to an abrupt end (Kamb *et al.*, 1985).

The second model assumes the glacier rests on an unconsolidated, permeable bed, and that rapid motion is due in part to deformation of water-saturated sediments (Clarke *et al.*, 1984). Sediment deformation may be negligible during quiescence if the sediment is drained by canals at the till surface, or a network of pipes within the sediments. Such types of system promote more efficient drainage, a reduction of sediment pore water pressure, and an increase in sediment strength. In this model, the rise in basal shear stress due to the evolution of glacier geometry during quiescence may initiate or increase the rate of sediment deformation. Although this might induce dilation of the sediment, increasing its porosity and hydraulic conductivity, it could also destroy pipes and surface canals, reducing the bulk transmissivity of the drainage system, driving up pore water pressures and weakening sediments (Clarke *et al.*, 1984). As the sediments weaken, they will deform more rapidly, initiating the surge phase of motion. Eventually, the down-glacier transfer of ice by the surge results in thinning of ice and a reduction in surface slope in the reservoir region, and a decrease in the basal shear stress acting on subglacial sediments. The basal shear stress may become too low to deform the sediments, and the cessation of deformation may allow drainage pipes and canals to become reestablished. This will further reduce pore water pressures and increase sediment strength, bringing an end to the surge motion.

Glacial Outbursts

Sudden releases of stored water are a common feature of the discharge regimes of glacier-fed rivers. The most common water sources for these "glacier outbursts" are ice-dammed lakes along glacier margins, but releases from supraglacial, subglacial, and englacial sources are also known. Often, these outbursts are routed through the subglacial drainage system and can be a major stimulus for changes in its character (Nye, 1976; Stone and Clarke, 1996; Flowers and Clarke, 2000; Björnsson, 2002). For the case of drainage of an ice-marginal lake, drainage commonly begins when the lake water level approaches that required to float the ice dam. Once water starts to leak into the subglacial drainage system, channel development likely occurs in a manner similar to that described for the spring transition. In some cases, channel development is initiated close to the ice dam and channels grow in a downstream direction over a period of days to weeks. In other cases, initial drainage may take the form of a turbulent sheet flood and channel development first occurs near the glacier margin. This facilitates drainage in this region, and provides a positive feedback for channel growth by reducing back-pressure in the system (Flowers *et al.*, 2004). When there is insufficient water in the lake to keep the channel full, the channel begins to shrink and eventually closes. This terminates the flood, which has a characteristically asymmetric hydrograph, with a gradual rise to peak discharge and abrupt termination. The flood characteristics depend upon the volume of water in the lake, the lake hypsometry, lake water temperature, ice overburden pressure, hydraulic gradient, and wall roughness of the tunnel (Clarke, 1982).

Outbursts from supraglacial and englacial water pockets may have a very different form, with a more abrupt and symmetric hydrograph (sudden break outbursts, Haeberli, 1983). This may reflect the sudden delivery of water to the subglacial drainage system as a result of water pressure-induced propagation of fractures from the void that contains the water pocket (such as a crevasse) to the glacier bed (Van der Veen, 1998). Abrupt increases in water delivery to the bed result in over-pressuring of the subglacial drainage system and hydraulic uplift of the surrounding ice. This in turn permits the disturbance to propagate away from its point of origin (Engelhardt and Kamb, 1997). This may activate new areas of the bed, increase the connectivity of the drainage system and allow flushing of stored water. Such events tend to propagate down-glacier and will eventually culminate as an outburst at the glacier margin. Connections created by these events may persist after the outburst, resulting in a more efficient drainage system (Flowers and Clarke, 2000). In some polythermal Arctic glaciers, such outbursts are the normal means by which subglacial outflow begins in spring (Skidmore and Sharp, 1999).

Tidal Forcing of Subglacial Drainage

Variability in subglacial water pressure and electrical conductivity on diurnal timescales is typically attributed to forcing by inputs of meltwater from the glacier surface. Such variability has, however, also been recorded in winter and in environments that are too cold for surface melting to occur. Semidiurnal fluctuations have also been observed in these environments. This may result from forcing of the subglacial drainage system by earth, atmospheric, or ocean tides. Diurnal variations in basal water pressure of Haut Glacier d'Arolla, Switzerland, in winter resulted from water flow between a borehole and the subglacial drainage system that was forced by deformation of the glacier substrate and the glacier ice induced by the luni-solar diurnal earth tide (Kulessa *et al.*, 2003). Diurnal velocity variations of Ice Stream D, Antarctica, appear to be driven by the ocean tide beneath the Ross Ice Shelf, into which the ice stream flows (Anandakrishnan *et al.*, 2003).

SUMMARY

Subglacial drainage can occur wherever ice at a glacier bed reaches the pressure melting point. The subglacial drainage system is fed from a mixture of surface, englacial, subglacial, and groundwater sources that differ in terms of their spatial distribution and characteristic patterns of temporal variability. Subglacial drainage systems are not readily accessible, and knowledge of their characteristics is derived from a range of indirect methods including radio-echo sounding, the use of artificial tracers, monitoring, and manipulation of subglacial conditions via boreholes, and monitoring of glacial runoff properties. Subglacial water flow is driven by gradients in hydraulic potential, and occurs through either fast/channelized or slow/distributed systems located at the ice-bed interface, or through subglacial aquifers. Water can be stored subglacially in cavities, in the pore space of subglacial sediments, or in subglacial lakes. Drainage system structure evolves continually on various timescales in response to changing water inputs, evolving glacier geometry, and changes in glacier flow dynamics. Major hydrological events in such systems include the spring and fall transitions, outburst floods, and structural changes related to glacier advances and glacier surging. In the absence of variable water inputs from the glacier surface, subglacial drainage systems may be sensitive to forcing by earth, ocean, and atmospheric tides.

Acknowledgments

I am indebted to Gwenn Flowers and Peter Nienow for constructive and thoughtful reviews of an earlier version of this article.

REFERENCES

Alley R.B. (1989) Water-pressure coupling of sliding and bed deformation, I. Water system. *Journal of Glaciology*, **35**, 108–118.

Alley R.B. (1996) Towards a hydrologic model for computerized ice sheet simulations. *Hydrological Processes*, **10**, 649–660.

Alley R.B., Lawson D.E., Evenson E.B., Strasser J.C. and Larson G.J. (1998) Glaciohydraulic super-cooling: a freeze-on mechanism to create stratified, debris-rich ice. II. Theory. *Journal of Glaciology*, **44**, 563–569.

Anandakrishnan S., Voigt D.E., Alley R.B. and King M.A. (2003) Ice stream D flow speed is strongly modulated by the tide beneath the Ross ice shelf. *Geophysical Research Letters*, **30**, 1361, doi:10.1029/2002GL016329.

Anderson S.P., Fernald K.M.H., Anderson R.S. and Humphrey N.F. (1999) Physical and chemical characterization of a spring flood event, Bench Glacier, Alaska: evidence for water storage. *Journal of Glaciology*, **45**, 177–189.

Bencala K.E., Rathburn R.E., Jackman A.P., Kennedy V.C., Zellweger G.W. and Avanzino R.J. (1983) Rhodamine WT dye losses in a mountain stream environment. *Water Resources Bulletin*, **19**, 943–950.

Bentley C.R. (1987) Antarctic ice streams: a review. *Journal of Geophysical Research*, **92**, 8843–8858.

Bentley C.R., Lord N. and Liu C. (1998) Radar reflections reveal a wet bed beneath stagnant Ice Stream C and a frozen bed beneath ridge BC, West Antarctica. *Journal of Glaciology*, **44**, 149–156.

Björnsson H. (1982) Drainage basins on Vatnajökull mapped by radio echo soundings. *Nordic Hydrology*, **13**, 213–232.

Björnsson H. (2002) Subglacial lakes and jökulhlaups in Iceland. *Global and Planetary Change*, **35**, 255–271.

Boon S. and Sharp M. (2003) The role of hydrologically-driven ice fracture in drainage system evolution on an Arctic glacier. *Geophysical Research Letters*, **30**, 4, doi:10.1029/2003GL018034.

Boulton G.S., Caban P.E. and van Gijssel K. (1995) Groundwater flow beneath ice sheets, part 1, Large scale patterns. *Quaternary Science Reviews*, **14**, 545–562.

Boulton G.S. and Jones A.S. (1979) Stability of temperate ice caps and ice sheets resting on beds of deformable sediment. *Journal of Glaciology*, **24**, 29–43.

Brown G.H., Sharp M.J., Tranter M., Gurnell A.M. and Nienow P.W. (1994) The impact of post-mixing chemical reactions on the major ion chemistry of bulk meltwaters draining the Haut Glacier d'Arolla, Switzerland. *Hydrological Processes*, **8**, 465–480.

Burkimsher M. (1983) Investigations of glacier hydrological systems using dye-tracer techniques: observations at Pasterzengletscher, Austria. *Journal of Glaciology*, **29**, 403–416.

Clarke G.K.C. (1982) Glacier outburst floods from "Hazard Lake", Yukon, and the problem of flood magnitude prediction. *Journal of Glaciology*, **28**, 3–21.

Clarke G.K.C. (1987a) Subglacial till: a physical framework for its properties and processes. *Journal of Geophysical Research*, **92**, 9023–9036.

Clarke G.K.C. (1987b) Fast glacier flow: ice streams, surging and tidewater glaciers. *Journal of Geophysical Research*, **92**, 8835–8841.

Clarke G.K.C., Collins S.G. and Thompson D.E. (1984) Flow, thermal structure and subglacial conditions of a surge-type glacier. *Canadian Journal of Earth Sciences*, **21**, 232–240.

Copland L. and Sharp M. (2001) Mapping thermal and hydrological conditions beneath a polythermal glacier with radio-echo sounding. *Journal of Glaciology*, **47**, 232–242.

Copland L., Sharp M. and Nienow P. (2003) Links between short-term velocity variations and the subglacial hydrology of a predominantly cold polythermal glacier. *Journal of Glaciology*, **49**, 337–348.

Cutler P.M. (1998) Modeling the evolution of subglacial tunnels due to varying water input. *Journal of Glaciology*, **44**, 485–497.

Dowdeswell J.A. and Siegert M.J. (1999) The dimensions and topographic setting of Antarctic subglacial lakes and implications for large scale water storage beneath continental ice sheets. *Geological Society of America Bulletin*, **111**, 254–263.

Engelhardt H. and Kamb B. (1997) Basal hydraulic system of a West Antarctic ice stream: constraints from borehole observations. *Journal of Glaciology*, **43**, 207–230.

Fischer H.B. (1968) *Methods for Predicting Dispersion Coefficients in Natural Streams, with Applications to the Green and Duwamish Rivers*, Vol. 582(A) United States Geological Survey Professional Paper: Washington.

Fischer U.H. and Clarke G.K.C. (1994) Ploughing of subglacial sediment. *Journal of Glaciology*, **40**, 97–106.

Fischer U.H., Iverson N.R., Hanson B., Hooke R.L.eB. and Jansson P. (1998) Estimation of hydraulic properties of subglacial till from ploughmeter measurements. *Journal of Glaciology*, **44**, 517–522.

Flowers G.E., Björnsson H., Pálsson F. and Clarke G.K.C. (2004) A coupled sheet-conduit mechanism for jökulhlaup propagation. *Geophysical Research Letters*, **31**, L05401, doi:10.1029/2003GL019088.

Flowers G.E. and Clarke G.K.C. (2000) An integrated modeling approach to understanding subglacial hydraulic release events. *Annals of Glaciology*, **31**, 222–228.

Flowers G.E. and Clarke G.K.C. (2002a) A multicomponent coupled model of glacier hydrology. 1. Theory and synthetic examples. *Journal of Geophysical Research*, **107**, 2287, doi:10.1029/2001JB001122.

Flowers G.E. and Clarke G.K.C. (2002b) A multicomponent coupled model of glacier hydrology. 2. Application to Trapridge Glacier, Yukon, Canada. *Journal of Geophysical Research*, **107**, 2288, doi:10.1029/2001JB001124.

Fountain A.G. (1994) Borehole water-level variations and implications for the subglacial hydraulics of South Cascade Glacier, Washington State, U.S.A. *Journal of Glaciology*, **40**, 293–304.

Fountain A.G. (1996) Effect of snow and firn hydrology on the physical and chemical characteristics of glacial runoff. *Hydrological Processes*, **10**, 509–521.

Fountain A.G. and Vaughn B.H. (1995) *Changing drainage patterns within South Cascade Glacier*, Vol. 228, International Association of Hydrological Sciences Publication: Washington, 1964–1992, pp. 379–386.

Fountain A.G. and Walder J.S. (1998) Water flow through temperate glaciers. *Reviews of Geophysics*, **36**, 299–328.

Gordon S., Sharp M., Hubbard B., Smart C., Ketterling B. and Willis I. (1998) Seasonal reorganization of subglacial drainage inferred from measurements in boreholes. *Hydrological Processes*, **12**, 105–133.

Gordon S., Sharp M., Hubbard B., Willis I., Smart C., Copland L., Harbor J. and Ketterling B. (2001) Borehole drainage and its implications for the investigation of glacier hydrology: experiences from Haut Glacier d'Arolla, Switzerland. *Hydrological Processes*, **15**, 797–813.

Haeberli W. (1983) Frequency and characteristics of glacier floods in the Swiss Alps. *Annals of Glaciology*, **4**, 85–90.

Hallet B. (1976) The effect of subglacial chemical processes on glacier sliding. *Journal of Glaciology*, **17**, 209–221.

Hallet B. (1979) Subglacial regelation water film. *Journal of Glaciology*, **23**, 321–334.

Hallet B. and Anderson R.S. (1980) Detailed glacial geomorphology of a proglacial bedrock area at Castleguard Glacier, Alberta, Canada. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **16**, 171–184.

Harbor J.M., Sharp M., Copland L., Hubbard B., Nienow P. and Mair D. *et al.* (1997) Influence of subglacial drainage conditions on the velocity distribution within a glacier cross section. *Geology*, **25**, 739–742.

Hock R. and Hooke R.L.eB. (1993) Evolution of the internal drainage system in the lower part of the ablation area of Storglaciären, Sweden. *Geological Society of America Bulletin*, **105**, 537–546.

Hodgkins R. (1997) Glacier hydrology in Svalbard, Norwegian high Arctic. *Quaternary Science Reviews*, **16**, 957–973.

Holmlund P. (1988) Internal geometry and evolution of moulins, Storglaciären, Sweden. *Journal of Glaciology*, **34**, 242–248.

Hooke R.L.eB. (1984) On the role of mechanical energy in maintaining subglacial water conduits at atmospheric pressure. *Journal of Glaciology*, **30**, 180–187.

Hooke R.L.eB., Hanson B., Iverson N.R., Jansson P. and Fischer U.H. (1997) Rheology of till beneath Storglaciären, Sweden. *Journal of Glaciology*, **43**, 172–179.

Hooke R.L.eB., Laumann T. and Kohler J. (1990) Subglacial water pressures and the shape of subglacial conduits. *Journal of Glaciology*, **36**, 67–71.

Hubbard B. and Nienow P. (1997) Alpine subglacial hydrology. *Quaternary Science Reviews*, **16**, 939–955.

Hubbard B.P., Sharp M.J., Willis I.C., Nielsen M.K. and Smart C.C. (1995) Borehole water level variations and the structure of the subglacial drainage system of Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **41**, 572–583.

Humphrey N.F. and Raymond C.F. (1994) Hydrology, erosion and sediment production in a surging glacier: Variegated Glacier, Alaska. *Journal of Glaciology*, **40**, 539–552.

Iken A. (1981) The effect of the subglacial water pressure on the sliding velocity of a glacier in an idealized numerical model. *Journal of Glaciology*, **27**, 407–421.

Iken A. and Truffer M. (1997) The relationship between subglacial water pressure and velocity of Findelengletscher, Switzerland, during its advance and retreat. *Journal of Glaciology*, **43**, 328–338.

Kamb W.B. (1987) Glacier surge mechanism based on linked cavity configuration of the basal conduit system. *Journal of Geophysical Research*, **92**, 9083–9100.

Kamb W.B. (1991) Rheological nonlinearity and flow instability in the deforming bed mechanism of ice-stream motion. *Journal of Geophysical Research*, **96**, 16585–16595.

Kamb B., Raymond C., Harrison W., Engelhardt H., Echelmeyer K., Humphrey N., Brugman M. and Pfeffer T. (1985) Glacier surge mechanism: 1982–1983 surge of Variegated Glacier, Alaska. *Science*, **227**, 469–479.

Kapitsa A., Ridley J.K., Robin G.deQ., Siegert M.J. and Zotikov I. (1996) A large deep freshwater lake beneath the ice of central East Antarctica. *Nature*, **381**, 684–686.

Kavanaugh J.L. and Clarke G.K.C. (2001) Abrupt glacier motion and reorganization of basal shear stress following the establishment of a connected drainage system. *Journal of Glaciology*, **47**, 472–480.

Kulessa B., Hubbard B., Brown G.H. and Becker J. (2003) Earth tide forcing of glacier drainage. *Geophysical Research Letters*, **30**, 1011, doi:10.1029/2002GL015303.

Kulessa B. and Murray T. (2003) Slug-test derived differences in bed hydraulic properties between a surge-type and a non-surge-type Svalbard glacier. *Annals of Glaciology*, **36**, 103–109.

Liestøl O., Repp K. and Wold B. (1980) Supra-glacial lakes in Spitsbergen. *Norsk Geografisk Tidsskrift*, **34**, 89–92.

Lingle C.S. and Fatland D.R. (2003) Does englacial water storage drive temperate glacier surges? *Annals of Glaciology*, **36**, 14–20.

Lliboutry L. (1968) General theory of subglacial cavitation and sliding of temperate glaciers. *Journal of Glaciology*, **7**, 21–58.

Lliboutry L. (1976) Physical processes in temperate glaciers. *Journal of Glaciology*, **16**, 151–158.

Mair D.W.F., Sharp M.J. and Willis I.C. (2002) Evidence for basal cavity opening from analysis of surface uplift during a high velocity event, Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **48**, 208–216.

Meier M.F. and Post A. (1969) What are glacier surges? *Canadian Journal of Earth Sciences*, **6**, 807–817.

Moorman B.J. and Michel F.A. (2000) Glacial hydrological system characterization using ground penetrating radar. *Hydrological Processes*, **14**, 2645–2667.

Murray T. and Clarke G.K.C. (1995) Black-box modeling of the subglacial water system. *Journal of Geophysical Research*, **100**, 10231–10245.

Nienow P.W., Sharp M. and Willis I.C. (1996) Velocity-discharge relationships derived from dye tracer experiments in glacial meltwaters: implications for subglacial flow conditions. *Hydrological Processes*, **10**, 1411–1426.

Nienow P.W., Sharp M. and Willis I.C. (1998) Seasonal changes in the morphology of the subglacial drainage system, Haut Glacier d'Arolla, Switzerland. *Earth Surface Processes and Landforms*, **23**, 823–845.

Nye J.F. (1976) Water flow in glaciers: Jökulhlaups, tunnels and veins. *Journal of Glaciology*, **17**, 181–207.

Nye J.F. and Frank F.C. (1973) The hydrology of the intergranular veins in a temperate glacier. International Association of Scientific Hydrology Publication, 95, International Association of Scientific Hydrology: pp. 157–161.

Oswald G.K.A. and Robin G.deQ. (1973) Lakes beneath the Antarctic ice sheet. *Nature*, **245**, 251–254.

Raiswell R. (1984) Chemical models of solute acquisition in glacial meltwaters. *Journal of Glaciology*, **30**, 49–57.

Raymond C.F. (1987) How do glaciers surge ? A review. *Journal of Geophysical Research*, **92**, 9121–9134.

Raymond C.F., Benedict R.J., Harrison W.D., Echelmeyer K.A. and Sturm M. (1995) Hydrological discharges and motion of Fels and Black Rapids Glaciers, Alaska, USA: implications for the structure of their drainage systems. *Journal of Glaciology*, **41**, 290–304.

Raymond C.F. and Harrison W.D. (1988) Evolution of Variegated Glacier, Alaska, U.S.A. prior to its surge. *Journal of Glaciology*, **34**, 1–16.

Richards K.S., Sharp M., Arnold N.S., Gurnell A.M., Clark M.J., Tranter M., Nienow P., Brown G.H., Willis I.C. and Lawson W. (1996) An integrated approach to modeling hydrology and water quality in glacierised catchments. *Hydrological Processes*, **10**, 479–508.

Robin G.deQ., Swithinbank C.W.M. and Smith B.M.E. (1970) *Radio Echo Exploration of the Antarctic Ice Sheet*. International Association of Hydrological Sciences Publication, 86, International Association of Hydrological Sciences: pp. 97–115.

Röthlisberger H. (1972) Water pressure in intra- and subglacial channels. *Journal of Glaciology*, **11**, 177–203.

Röthlisberger H. and Lang H. (1987) Glacial Hydrology. In *Glaciofluvial Sediment Transfer: An Alpine Perspective*, Gurnell A.M. and Clark M.J. (Eds.), John Wiley & Sons: New York, pp. 207–284.

Seaberg S.Z., Seaberg J.Z., Hooke R.L.eB. and Wiberg D. (1988) Character of the englacial and subglacial drainage system in the lower part of the ablation area of Storglaciären, Sweden, as revealed by dye trace studies. *Journal of Glaciology*, **34**, 217–227.

Sharp M.J., Gemmell J.C. and Tison J.-L. (1989) Structure and stability of the former subglacial drainage system of the Glacier de Tsanfleuron, Switzerland. *Earth Surface Processes and Landforms*, **14**, 119–134.

Sharp M.J., Richards K.S., Willis I.C., Arnold N.S., Nienow P., Lawson W. and Tison J.-L. (1993) Geometry, bed topography and drainage system structure of the Haut Glacier d'Arolla, Switzerland. *Earth Surface Processes and Landforms*, **18**, 557–571.

Shreve R.L. (1972) Movement of water in glaciers. *Journal of Glaciology*, **11**, 205–214.

Siegert M.J. (2000) Antarctic subglacial lakes. *Earth-Science Reviews*, **50**, 29–50.

Siegert M.J., Dowdeswell J.A., Gorman M.R. and McIntyre N.F. (1996) An inventory of Antarctic subglacial lakes. *Antarctic Science*, **8**, 281–186.

Skidmore M. and Sharp M. (1999) Drainage system behaviour of a High Arctic polythermal glacier. *Annals of Glaciology*, **28**, 209–215.

Smart C.C. (1996) Statistical evaluation of glacier boreholes as indicators of basal drainage systems. *Hydrological Processes*, **10**, 599–613.

Stone D.B. and Clarke G.K.C. (1993) Estimation of subglacial hydraulic properties from induced changes in basal water pressure: a theoretical framework for borehole response tests. *Journal of Glaciology*, **39**, 327–340.

Stone D.B. and Clarke G.K.C. (1996) In situ measurements of basal water quality and pressure as an indicator of the character of subglacial drainage systems. *Hydrological Processes*, **10**, 615–628.

Stone D.B., Clarke G.K.C. and Blake E.W. (1993) Subglacial measurement of turbidity and electrical conductivity. *Journal of Glaciology*, **39**, 415–420.

Stone D.B., Clarke G.K.C. and Ellis R.G. (1997) Inversion of borehole response test data for estimation of subglacial hydraulic properties. *Journal of Glaciology*, **43**, 103–113.

Swift D.A., Nienow P.W., Spedding N. and Hoey T.B. (2002) Geomorphic implications of subglacial drainage configuration: rates of basal sediment evacuation controlled by seasonal drainage system evolution. *Sedimentary Geology*, **149**, 5–19.

Tranter M., Brown G.H., Raiswell R., Sharp M.J. and Gurnell A.M. (1993) A conceptual model of solute acquisition by Alpine glacial meltwaters. *Journal of Glaciology*, **39**, 573–581.

Tranter M., Sharp M., Brown G.H., Willis I.C., Hubbard B.P., Nielsen M.K., Smart C.C., Gordon S., Tulley M. and Lamb H.R. (1997) Variability in the chemical composition of in situ subglacial meltwaters. *Hydrological Processes*, **11**, 59–77.

Tranter M., Sharp M.J., Lamb H.R., Brown G.H., Hubbard B.P. and Willis I.C. (2002) Geochemical weathering at the bed of Haut Glacier d'Arolla, Switzerland – a new model. *Hydrological Processes*, **16**, 959–993.

Tulaczyk S., Kamb B. and Engelhardt H. (2000) Basal mechanics of ice stream B, West Antarctica, II, Undrained plastic bed model. *Journal of Geophysical Research*, **105**, 483–494.

Van der Veen C.J. (1998) Fracture mechanics approach to penetration of surface crevasses on glaciers. *Cold Regions Science and Technology*, **27**, 31–47.

Wadham J.L., Hodgkins R., Cooper R.J. and Tranter M. (2001) Evidence for seasonal subglacial outburst events at a polythermal glacier, Finsterwalderbreen, Svalbard. *Hydrological Processes*, **15**, 2259–2280.

Walder J.S. (1982) Stability of sheet water flow beneath temperate glaciers and implications for glacier surging. *Journal of Glaciology*, **28**, 273–293.

Walder J.S. (1986) Hydraulics of subglacial cavities. *Journal of Glaciology*, **32**, 439–445.

Walder J.S. and Fowler A.C. (1994) Channelized subglacial drainage over a deformable bed. *Journal of Glaciology*, **40**, 3–15.

Walder J.S. and Hallet B. (1979) Geometry of former subglacial water channels and cavities. *Journal of Glaciology*, **23**, 335–346.

Weertman J. (1964) The theory of glacier sliding. *Journal of Glaciology*, **3**, 287–303.

Weertman J. (1972) General theory of water flow at the base of a glacier or ice sheet. *Reviews of Geophysics and Space Physics*, **10**, 287–333.

Willis I.C., Arnold N.S. and Brock B.W. (2002) Effect of snowpack removal on energy balance, melt and runoff in a small supraglacial catchment. *Hydrological Processes*, **16**, 2721–2749.

Willis I.C. and Bonvin J.-M. (1995) Climatic change in mountain environments. *Geography*, **80**, 247–261.

Willis I.C., Sharp M.J. and Richards K.S. (1990) Configuration of the drainage system of Midtdalsbreen, Norway, as indicated by dye-tracing experiments. *Journal of Glaciology*, **36**, 89–101.

168: Hydrology of Glacierized Basins

IAN WILLIS

Department of Geography, University of Cambridge, Cambridge, UK

The presence of snow and ice and the lack of vegetation and soil make the hydrology of glacierized basins very different to that of equivalent nonglacierized basins. This article discusses how the characteristics of glacierized basins control their runoff regime and its interannual, intraannual, and intraseasonal variability. The magnitudes and patterns of seasonal water storage and release associated with the snowpack, firn layers, and englacial and subglacial drainage systems are also considered. The controls on flooding in glacierized catchments are discussed with specific reference to the catastrophic drainage of ice-dammed subglacial and ice-marginal lakes, moraine-dammed proglacial lakes, and drainage of water from the englacial/subglacial drainage network that may be triggered by high surface melting or rainstorms. Finally, the effects of climate change on runoff regimes, water storage, and floods are assessed. The similarities and differences between the hydrological regimes of mid-, high-, and low-latitude basins containing glaciers with different thermal regimes are highlighted.

INTRODUCTION

The hydrology of glacierized basins differs from that of equivalent nonglacierized basins because of the presence of snow and ice and lack of vegetation and soil. Whereas rainfall is often the dominant water input to nonglacierized basins, snow and ice melt provide important inputs to glacierized basins (Röthlisberger and Lang, 1987; Chen and Ohmura, 1990). Vegetation interception and evapotranspiration are important hydrological processes in many nonglacierized catchments, but are of negligible importance in glacierized basins. Similarly, soil porosity and permeability are usually important influences on runoff from nonglacierized basins, while the porosity and permeability of snow and firn are key in glacierized basins. The routing of water vertically and laterally through snow and firn and laterally across ice surfaces are important processes in glacierized catchments, but soil throughflow and overland flow play more dominant roles in nonglacierized basins. Furthermore, englacial routing of water through ice in small veins and pipes or, more importantly, along crevasses and large conduits, is unique to glacierized basins. Subglacial routing of water at the base of ice through thin films, sediment pores, linked cavities, or channels incised upward into ice and/or downward into rock or sediment, is a unique feature of glacierized basins.

The distinctive characteristics of glacierized catchments have important implications for (i) runoff regimes; (ii) water storage and release; (iii) the incidence of flooding; and (iv) the effects on these of climate variability and change. The hydrology of glacierized basins has been studied mainly in midlatitudes in catchments containing temperate glaciers (especially in the European Alps; Norway; Iceland; New Zealand; USA and Canada) or mainly temperate polythermal glaciers (e.g. northern Sweden). Less work has been carried out in glacierized basins at lower latitudes (e.g. Central and South Asia; Africa; South America). Only recently has the hydrology of high-latitude basins containing cold or mainly cold polythermal glaciers begun to be explored (e.g. Nunavut, Canada; Spitsbergen, Svalbard; McMurdo Dry Valleys, Antarctica). This bias toward temperate glaciers in the midlatitudes is necessarily reflected in this review. However, reference is made to basins in other areas where appropriate, and similarities and differences between these catchments and those in the more commonly studied regions are highlighted.

INTERANNUAL VARIATIONS IN GLACIER WATER (MASS) BALANCE

Although they constitute only ~0.4% of all freshwater on the Earth, glaciers and ice caps respond rapidly to climate

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

change, made an important contribution to sea-level rise over the last century (Meier, 1984) and will continue to do so in the future (Houghton *et al.*, 2001). Glaciers and ice caps are often regarded as natural reservoirs, storing water as snow and ice during wet, cool years and releasing it during dry, warm years. Thus, a glacier's water (mass) balance is more positive during wetter, cooler years, and more negative during drier, warmer years.

Several authors have synthesized the many mass balance records from around the world (Dyurgerov and Meier, 1997; Cogley and Adams, 1998; Haeberli *et al.*, 2000; Braithwaite, 2002; see also **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**). During the latter part of the twentieth century, most small glaciers and ice caps shrank in response to global climate change, although there were marked differences between regions. Thus, between 1980 and 1995, glaciers in the Cascades and Kenya underwent almost continuous and rapid shrinkage, glaciers in the Alps, Arctic, and Tienshan underwent fairly continuous but moderate shrinkage, and those in Alaska, the Pamirs, and Kamchatka had years of both positive and negative mass balance with a moderate shrinkage overall. Glaciers in the Caucasus and Altai, however, had periods of positive and negative mass balance but a slight increase in size overall, while those in Scandinavia also had periods of positive and negative mass balance and a moderate increase in overall size.

Global climate change is manifested as regional changes to atmospheric circulation patterns, which may be described by negative and positive atmospheric pressure anomalies (e.g. Leather *et al.*, 1991). Regional scale climate controls glacier mass balance through its effects on local scale precipitation, cloudiness, humidity, and air temperature. Therefore, teleconnections exist between atmospheric circulation patterns or pressure anomalies and glacier mass balance, terminus position, or equilibrium line altitude (Table 1). The increased rate of glacier shrinkage since the mid-1970s coincided with shifts in atmosphere/ocean, temperature/pressure patterns such as El Niño Southern Oscillation (ENSO) (e.g. Francou *et al.*, 2000) and Pacific Decadal Oscillation (PDO) (e.g. Moore and Demuth, 2001).

Detailed understanding of the links between local scale climate and glacier mass balance is best achieved through the use of numerical models in which ablation is calculated using energy balance techniques (e.g. Oerlemans and Fortuin, 1992; Oerlemans, 1993; Hofinger and Kuhn, 1996). Alternatively, where the energy balance components cannot be determined, ablation may be calculated using simpler approaches involving "seasonal sensitivity characteristics" (e.g. de Ruyter de Wildt *et al.*, 2003) or "positive degree days" (e.g. Laumann and Reeh, 1993; Rabus and Echelmeyer, 1998; Braithwaite and Zhang, 2000). All these modeling approaches show that glacier mass balance is dependent on both air temperature and precipitation but that the sensitivity of mass balance to these variables varies with climatic setting. For example, glaciers in maritime settings are more sensitive to air temperature changes than those in continental locations (e.g. Oerlemans and Fortuin, 1992; Laumann and Reeh, 1993). This is because higher air temperatures not only increase ablation but also decrease accumulation by reducing the proportion of precipitation falling as snow. This lowers mass balance both directly and indirectly because reduced snow accumulation allows glacier ice to be exposed for a longer period in summer. This increases ablation because the albedo of ice is lower than that of snow. Energy balance modeling studies show that glacier mass balance may be sensitive to variables other than precipitation and air temperature. For example, Oerlemans (1993) showed that a decrease in glacier albedo of just 0.03 might have the same effect on glacier mass balance as an increase in air temperature of 1 °C.

INTERANNUAL VARIATIONS IN GLACIER RUNOFF

Year-to-year variations in regional climate control not only variations in glacier mass balance, but also runoff from glacierized basins (Ribstein *et al.*, 1995; Depetris and Pasquini, 2000; Lawler *et al.*, 2003). Years with a strongly negative Southern Oscillation (SO) index are associated with high annual runoff from Zongo Glacier, Bolivia because of their association with warm dry weather and high melt rates (Ribstein *et al.*, 1995). Runoff from three glacierized basins in Iceland decreased between 1973 and 1992, particularly in spring, because of marked cooling and reduced melting, and in autumn, because of fewer heavy rainstorms (Lawler *et al.*, 2003). Spring cooling was associated with an increase in the North Atlantic Oscillation (NAO) Index and the weakening of southwesterly circulation over Iceland, whereas autumn rainstorm reduction was linked to a decrease in the NAO index and reduction in zonal airflow. Depetris and Pasquini (2000) analyzed interannual variations in stream flow between 1956 and 1994, downstream of Lake Argentino, an upstream arm of which is periodically dammed by the Perito Moreno Glacier, Argentina. Spectral analysis of the discharge series revealed significant peaks in the 33–36-month and the 42–58-month ranges, corresponding with similar periodicities in the position of subtropical Pacific anticyclones and the intensity of ENSO respectively. Further analysis of the 42–58-month periodicity showed that anomalously low and high discharges were associated with La Niña and El Niño years respectively. Furthermore, the high discharges associated with El Niño years were due to lake damming by the Perito Moreno Glacier. El Niño conditions led to glacier advance (due to positive mass balance) and high rainfall, both of which led to larger lake

Table 1 Teleconnections between global atmospheric circulation and glaciers

Atmospheric index	Glacier index	Correlation	Glacier	Year(s)	Cause(s)	Reference
Mean sea level pressure & 500 hPa height over North Atlantic (winter and summer)	Winter and summer mass balance	High (low) winter accumulation = strong (weak) Azores to Iceland pressure gradient & weak (strong) Siberian high; High (low) summer ablation = low (high) pressure from Southern Greenland to Gulf of Bothnia & high (low) pressure over Barents Sea	Storglaciären, Sweden	1946 to 1992	(a) Strength of prevailing westerlies; (b) winter precipitation; (c) summer temperature	Pohjola and Rogers (1997)
SOI	Advanced terminus position	Negative	Franz Josef Glacier, New Zealand	1954 to 1994	(a) Pressure anomalies over New Zealand in ablation season; (b) prevailing wind direction; (c) precipitation; (d) summer temperature	Hooker and Fitzharris (1999)
Decadal ENSO-like phenomena & PDO	Winter and net mass balance	Negative with three glaciers, positive with one	Four glaciers, Washington, USA & BC, Canada		(a) Precipitation & storminess; (b) temperature	Bitz and Battisti (1999)
Decadal ENSO-like phenomena & PDO	Winter and net mass balance	Positive	Wolverine Glacier, Alaska		(a) Precipitation & advected moisture flux; (b) winter temperature	Bitz and Battisti (1999)
Multivariate ENSO index (monthly)	Monthly mass balance	Negative	Chacaltaya Glacier, Bolivia & Antizana Glacier, Ecuador	1991 to 1999	Summer ablation	Francou *et al.* (2000)

(*continued overleaf*)

Table 1 (*continued*)

Atmospheric Index	Glacier index	Correlation	Glacier	Year(s)	Cause(s)	Reference
NAO (Oct–Apr) & (Dec–Mar)	Winter and net mass balance	Positive	Ten glaciers (especially maritime), southern Norway		Winter precipitation	Nesje *et al.* (2000)
SOI (Oct–Mar)	End of summer snowline altitude	Negative	Sajama Volcano Glacier, Bolivia	1963 to 1998	(i) Rainy season precipitation; (ii) temperature of warmest month	Arnaud *et al.* (2001)
PDO (Nov–Apr)	Winter mass balance	Negative	Place Glacier, British Columbia, Canada	1965 to 1999	Winter precipitation	Moore and Demuth (2001)
PDO (Nov–Apr) SOI (Nov–Apr)	Net mass balance	Negative Positive				
SOI	Net mass balance	Positive	Zongo Glacier, Bolivia	1991 to 1998	(i) precipitation; (ii) temperatures; (iii) sublimation; (iv) melting.	Wagnon *et al.* (2001)
PDO NP	Advanced terminus position	Negative Positive	Six glaciers, Mount Baker, North Cascades, Washington, USA	1940 to 1990	(i) Winter precipitation; and (ii) summer temperatures	Kovanen (2003)

volumes above the dam and bigger flood discharges when the dam broke.

As with glacier mass balance, regional scale climate controls annual runoff, largely through its effects on local precipitation and air temperature. However, the effect of local climate variability on annual river flows depends on the sensitivity of the runoff regime to both these variables. Thus, an increase in air temperature and decrease in precipitation may result in increased runoff in a glacier melt-dominated river but decreased runoff in a rainfall-dominated one. For example, annual runoff from glacierized catchments in central Tien Shan, China, increased by 7% between the 1950s and the 1990s corresponding with air temperature and precipitation increases, although it was not obvious whether the runoff increases were due to increased melt, rainfall, or both (Jingshi and Fukushima, 1999).

Interannual runoff variability also depends on the percentage of the catchment that is glacierized (Kasser, 1973; Fountain and Tangborn, 1985; Braithwaite and Olesen, 1988; Chen and Ohmura, 1990; Moore, 1992; Pelto, 1996; Hopkinson and Young, 1998). Variability is at a minimum for catchments that are ~40% glacierized and goes up as the glacierized percentage both increases and decreases (Figure 1). The reason is that in catchments where runoff is controlled by both rainfall and glacier melt, low rainfall in warm, dry years is offset by high melt, while low melt in cold, wet years is offset by high rainfall. This "compensating effect" is greatest where catchment runoff is controlled approximately equally by rainfall and glacier melt (Röthlisberger and Lang, 1987). For example, in the period 1985–1993, summer stream flows in the North Cascades, Washington, USA, were lower than their long-term averages. The biggest reductions occurred in the least

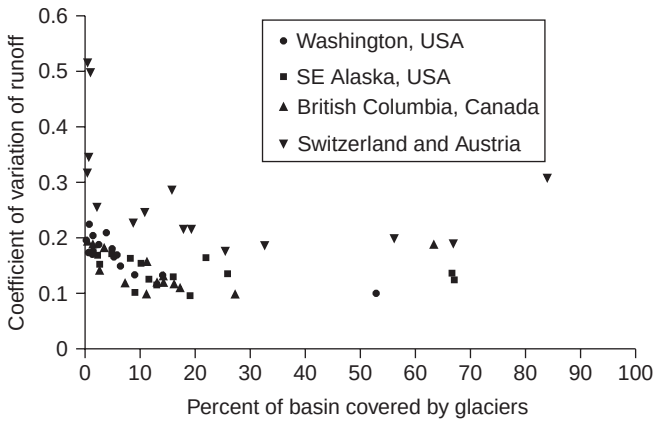


Figure 1 Coefficient of variation of runoff versus the percentage of the catchment that is glacierized. Data shown are annual runoff variability for Washington and Alaska, USA (Fountain and Tangborn, 1985), and British Columbia, Canada (Moore, 1992), and August runoff variability for Switzerland and Austria (Röthlisberger and Lang, 1987)

glacierized catchments, where glacier ice melt compensated to some extent for the lower runoff associated with reduced summer snowcover and rainfall (Pelto, 1996). Similarly, Hopkinson and Young (1998) showed that the "compensating effect" is measurable even in catchments with a low percentage glacier cover. Working in the 3% glacierized Bow River Basin, Alberta, Canada, where the runoff regime is dominated by snow melt and rain inputs, they compared annual water yields with annual glacier volume changes over the period 1952 to 1993. Glacier volumes tended to increase during high-flow years with high snowmelt and rainfall, and decreased during low-flow years. On average, glacier melt contributed ~2% of basin yield, but this increased to ~13% for the lowest-flow year. Interestingly, there were several years when this general trend was disrupted, showing that glaciers do not always augment river discharge during low-flow years. In an extreme case, basin yield during a very low-flow year was reduced by up to 4% because of an increase, rather than a decrease in glacier volume.

Interannual runoff variability depends not only on the magnitude of air temperature and rainfall changes but also on their timing during the year. For example, Liu *et al.* (1999) studied air temperature, precipitation, and runoff changes in eight glacierized catchments in western China between 1953 and 1993. Mean annual air temperatures were 0.5 °C higher and precipitation was 10–20% higher after the late 1970s than before. However, annual stream flow decreased over time because rises in temperature were concentrated in winter and spring, while precipitation increases were spread throughout the year. The effects were to increase winter accumulation but not summer melt. Summer runoff increased slightly, because of increases in rain rather than melt.

Secular trends in climate can affect annual river flows indirectly through their effects on glacier mass balance and geometry. Thus, an increase in air temperature may increase runoff in a glacier melt–dominated stream in the short term but decrease it over the longer term because of glacier shrinkage. Pelto (1996) analyzed stream flow variations in five glacierized basins in the North Cascades, Washington, USA. Between 1985 and 1993, mean summer temperatures were higher, mean winter precipitation totals were lower, and mass balances were more negative than their long-term means. Summer runoff was also lower by an average of ~23% because of lower winter snowcover, earlier spring snow melt, and reduced glacier area.

Changes in the relative areas of snow, firn, and ice on a glacier surface can also affect river flows. For example, Braun and Escher-Vetter (1996) examined year-to-year variations in proglacial stream discharge below Vernagtferner, Austria, in relation to documented changes in mass balance. Between 1979 and 1990, the total glacier area shrank by ~0.46 km^2 but summer discharges increased.

This was attributed to an increase in the area of ice relative to snow and firn, which reduced the overall glacier albedo and increased melt rates. Conversely, a similar study of Place Glacier, BC, Canada showed that the glacier shrank by $\sim 0.24\,km^2$ between 1970 and 1984, that August discharges were negatively correlated with winter snow accumulation and positively correlated with August air temperatures, and that once these controls had been accounted for, August discharges showed a negative trend through time (Moore and Demuth, 2001). The areas of snow and firn remained fairly constant at Place Glacier so the area of ice relative to snow and firn decreased, glacier albedo increased, and melt rates decreased.

INTRAANNUAL VARIATIONS IN RUNOFF

The annual runoff regime of glacierized basins depends largely on the annual distribution of radiation, air temperature, humidity, cloudiness, and precipitation and therefore varies with geographical setting.

Midlatitudes

The midlatitudes have marked annual variations in solar radiation and air temperature and small to moderate variations in precipitation (Figure 2a and b). Consequently, in midlatitude glacierized catchments (e.g. Alps, North America, Scandinavia, Iceland, New Zealand, Greater Himalayas, India), the year can be divided into a winter accumulation season and a summer ablation season. (Catchments in the Greater Himalayas are included here as most of the precipitation falls during the winter months and the summer monsoon has little influence (Singh *et al.*, 1995). Catchments in the Outer and Middle Himalayas, where most of the precipitation falls during the monsoon with only negligible amounts during the winter, are included in the low latitude section). During winter, precipitation falls mainly as snow and glacier melt is negligible. During summer, precipitation falls mainly as rain, especially at lower elevations, and glacier melt is high. Such glaciers are known as *winter accumulation type* (Ageta and Higuchi, 1984), and tend to store water as snow and ice during the winter, and release it as water during the summer. Consequently, there are three important characteristics of the runoff regimes of midlatitude glacierized basins (Figure 3a). First, there is a long summer melt season, coinciding approximately with the period when mean daily temperatures are above $0\,^{\circ}C$. Second, there is a large annual variation in runoff as summer melting coincides with rainfall across the lower elevations of the catchment. For example, ~90% of the annual runoff from Vernagtferner, Austria, occurs between June and October (Escher-Vetter and Reinwarth, 1994). Scandinavian glaciers discharge ~85% of their annual runoff between June and August (Østrem, 1973). Approximately 70% of the runoff in the Satluj River, Western Himalayas occurs between June and September (Singh and Jain, 2002). Third, there is a delay in the timing of maximum annual flow relative to equivalent nonglacierized basins. This is caused by temporary storage of spring meltwater, and by peak meltwater production in midsummer (Fountain and Tangborn, 1985).

High Latitudes

The high latitudes have even greater annual variations in solar radiation and air temperature than midlatitudes, with long periods of the year when air temperatures are below freezing point (Figure 2c and d). They also have moderate variations in precipitation, but lower precipitation totals than midlatitudes (Figure 2c and d). High-latitude glaciers in maritime areas (e.g. Svalbard) may be described as *winter accumulation type*, but those in more continental regions (e.g. Arctic Canada) may be thought of as *summer accumulation type* (Ageta and Higuchi, 1984). All these glaciers release water in the summer but the melt season is shorter than in midlatitudes (Figure 3b and c). For example, ~97% of the annual runoff from Austre Broggerbreen, Svalbard, occurs from June to August (Repp, 1988), while ~94% of the annual runoff from Quviagivaa and Nirukittuq Glaciers, Ellesmere Island, Canada, occurs in July and August (Wolfe and English, 1995). The lower radiation receipts and air temperatures mean that the annual variation in runoff is, typically, less than in equivalent midlatitude basins (Figure 3b and c).

An extreme case is Canada Glacier, Taylor Valley, Antarctica, where the entire annual runoff is typically confined to a six to eight week period in December and January (Lewis *et al.*, 1999) (Figure 3c). Runoff in the ice-marginal streams occurs only when air temperatures near the glacier terminus are above freezing point. Melting occurs at other times, because of high radiation energy, particularly on the vertical cliffs of the terminus, but the water refreezes before it reaches the streams. For their area (~2% of the ablation zone), the vertical ice cliffs are more important than the horizontal glacier surfaces in supplying meltwater to the streams (~15–20% of the annual glacier runoff). This is because the cliffs often have a higher net short-wave radiation flux than the horizontal surfaces, as they cannot accumulate snow and therefore have a lower albedo. They also have a higher net long-wave radiation flux because of radiation from the bare ground beneath the cliffs. Finally, the lower elevation of the cliffs means that air temperatures are higher, specific humidity is higher, and sublimation rates are negligible. Thus, virtually all the energy receipt on the cliffs is used for melting, whereas only ~20–60% of the energy is used for melting on the horizontal surfaces, the rest being used for sublimation (Lewis *et al.*, 1999).

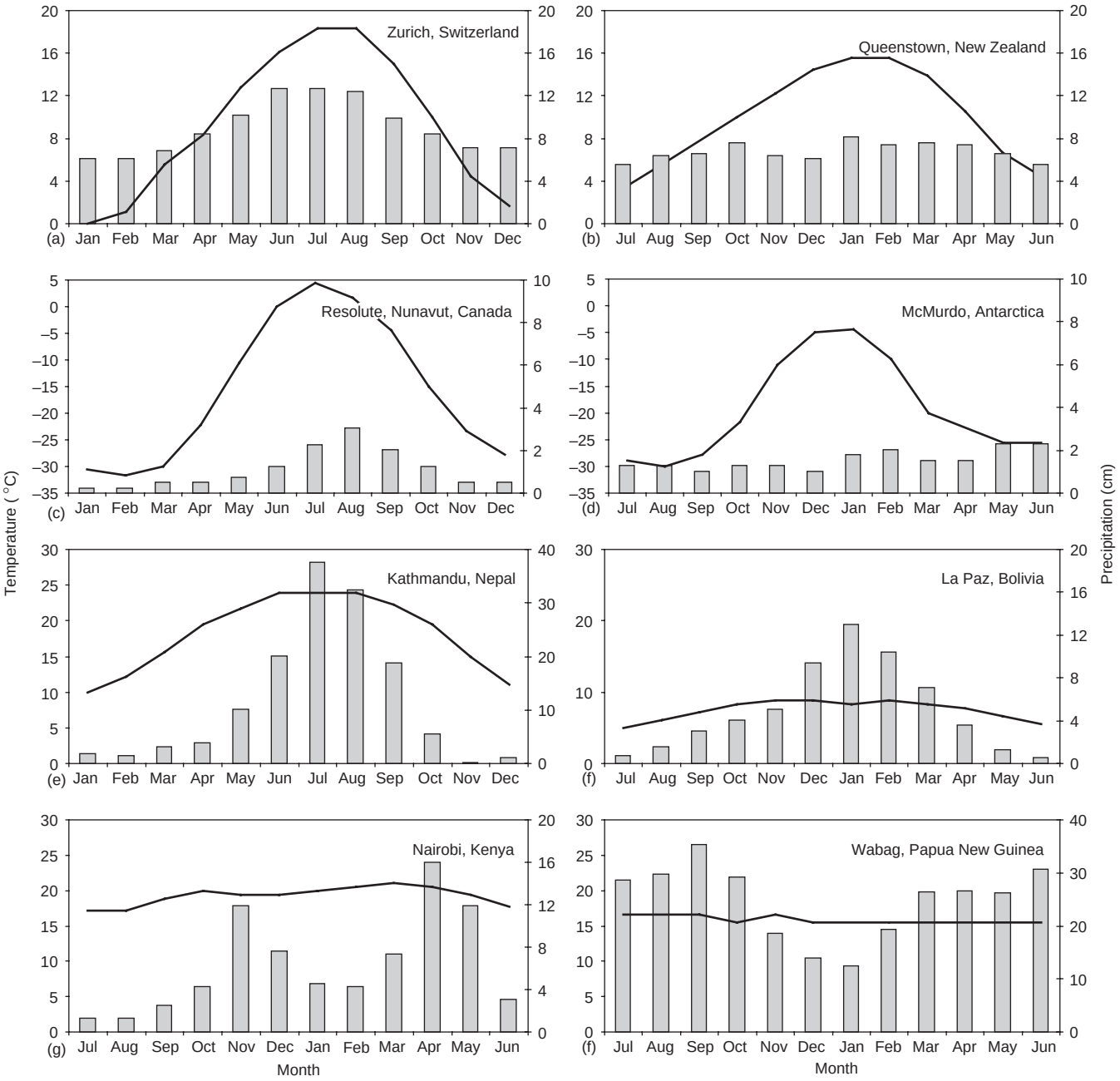


Figure 2 Monthly air temperature (line) and precipitation (bars) near glacierized basins in midlatitudes (a and b), high latitudes (c and d), low latitudes with ''summer accumulation'' regime (e and f), and low latitudes with ''year-round ablation'' regime (g and h)

Low Latitudes

The low latitudes have relatively small annual variations in upper atmosphere solar radiation receipt and air temperature, but annual fluctuations in the Inter Tropical Convergence Zone produce large variations in humidity, cloudiness, and precipitation (Figure 2e–h). Low latitude glaciers can be described as either "summer accumulation type" or "year-round ablation type" (Ageta and Higuchi, 1984).

Summer accumulation type glaciers occur at high altitudes, with cold, dry winters and marked summer precipitation maxima. Such areas include the monsoon-dominated parts of the Outer and Middle Himalayas in India and Nepal (Higuchi *et al.*, 1982); the Pamiro-Altai in Tadjikistan (Konovalov and Shchetinnicov, 1994); the Karakoram, Pamirs, Tien Shan, Altai, Kunlun and Qilian mountains of Pakistan, Tadjikistan, Kazakhstan, and China (Xie and Lui, 1993; Aizen *et al.*, 1995; Xie *et al.*, 1999; Liu *et al.*,

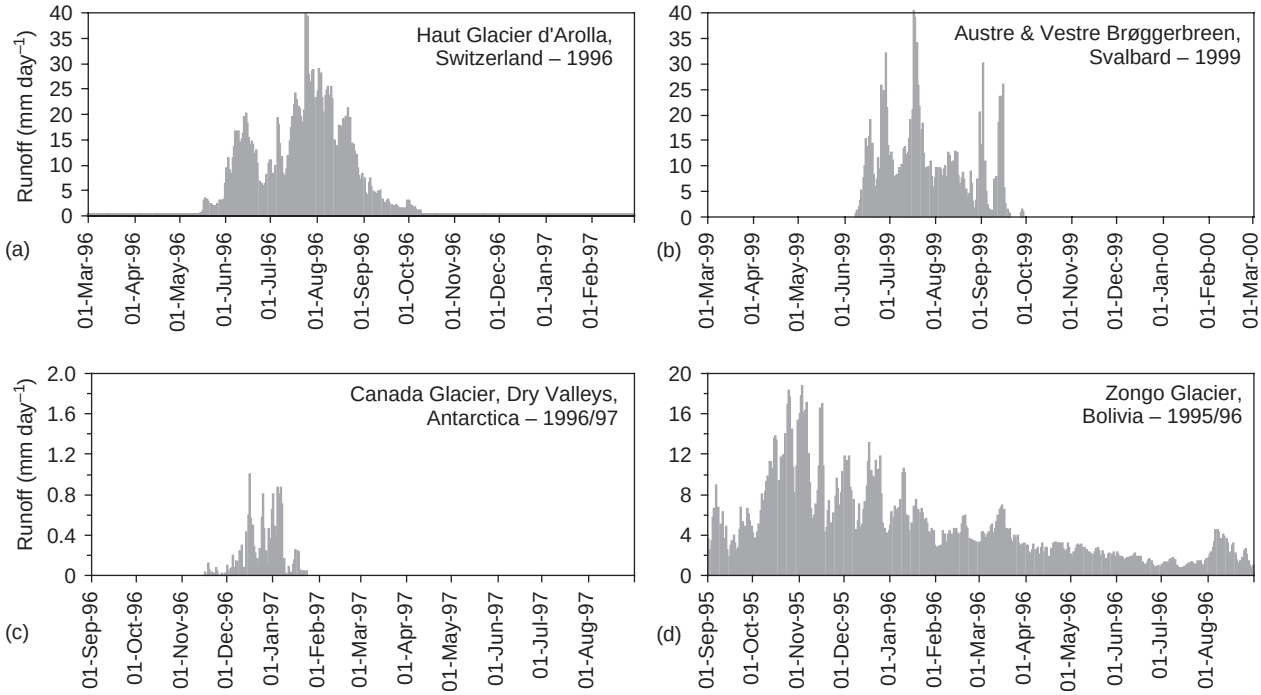


Figure 3 Daily specific runoff from (a) Haut Glacier d'Arolla, Switzerland; (b) Austre and Vestre Brøggerbreen, Svalbard; (c) Canada Glacier, Taylor Valley, Antartcia; (d) Zongo Glacier, Bolivia

1999); and the high Andes of Peru and Bolivia in the outer tropics (Francou *et al.*, 1995; Hastenrath and Ames, 1995; Ribstein *et al.*, 1995; Ames and Hastenrath, 1996; Kaser, 1999, 2001; Kaser and Osmaston, 2001). In these regions, there is very little winter precipitation and in the summer, precipitation falls mainly as snow at high elevations and rain at lower elevations (Figure 2e and f). Upper atmosphere solar radiation receipt and air temperatures are fairly constant throughout the year (Figure 2e and f), and annual variations in ablation are controlled mainly by variations in cloudiness and humidity (Kaser and Georges, 1999; Wagnon *et al.*, 1999a,b; Kaser and Osmaston, 2001). Ablation occurs throughout the year, but the majority of winter ablation is by sublimation due to the dry air. Conversely, the humid summer air means that sublimation is negligible, and most of the ablation is by melting.

As both accumulation and melting are concentrated in the summer, low latitude glaciers tend not to store and release water with the same seasonal rhythm as mid- to high-latitude glaciers. However, glacierized catchments in these regions are subject to large annual variations in runoff, with high runoff corresponding with the wet season when both rainfall and glacier melting are high (Fukushima *et al.*, 1987; Collins, 1999; Hasnain and Thayyen, 1999). For example, virtually all of the annual runoff from Langtang Glacier, Nepal occurs between May and October (Tangborn and Rana, 2000), and 70–90% of runoff in glacierized basins in western China occurs between June and September (Liu *et al.*, 1999). Similarly, ~60% of the annual runoff from Lirung Glacier, Nepal occurs during the monsoon period (mid-June–mid-September) (Bhatt *et al.*, 2000), and ~65% of the May–October runoff occurs during the monsoon months of July and August (Hasnain and Thayyen, 1999). At Zongo Glacier, Bolivia, runoff is more evenly distributed throughout the year. Only ~50% of annual runoff occurs during the warm summer period (December–March), and precipitation-induced runoff associated with the passage of cold fronts is not uncommon during the winter months of June and July (Ribstein *et al.*, 1995) (Figure 3d).

Year round ablation type glaciers occur in high-altitude areas in the inner tropics where radiation and air temperatures are high and their annual variations are negligible (Figure 2g and h). Precipitation also occurs throughout the year, although distinct peaks may coincide with one or two wet seasons. Glaciers in these regions are subject to fairly even accumulation and ablation throughout the year although, as with summer accumulation type glaciers, both accumulation and melting may be slightly higher during the wetter months. Glaciers with one wet month are found on Mount Jaya, Papua New Guinea, Mount Kilimanjaro, East Africa, and the volcanoes of Ecuador (Hope *et al.*, 1976; Hastenrath, 1981) (Figure 2g). Those with two wet months are located in the Ruwenzori Mountains and on Mount Kenya, East Central Africa (Whittow *et al.*, 1963; Hastenrath, 1984; Hastenrath and Kruss, 1992) (Figure 2h). As accumulation and ablation occur fairly evenly throughout the

year, runoff also tends to be distributed throughout the year.

Effect of Catchment Size and Percentage Glacier Cover

The annual runoff regime of glacierized catchments depends not only on geographical setting but also on catchment size, hypsometry, and percentage glacier cover. In the midlatitudes (Kasser, 1973; Fountain and Tangborn, 1985; Röthlisberger and Lang, 1987; Young and Hewitt, 1993) and low latitudes (Kaser *et al.*, 2003), annual runoff variation is at a minimum for basins with ~30–40% glacier cover. Higher annual runoff variation occurs in basins with both lower and higher glacier cover. Catchments with a high percentage glacier cover tend to be small upland catchments with a low elevation range. Liquid water inputs are very low during winter (due to precipitation falling as snow and negligible melt) but very high during summer (due to rainfall and high melt). Consequently, annual runoff variability is high, as discussed earlier. Conversely, catchments with a low percentage glacier cover tend to be large, extend into the lowlands, and have a high elevation range. Here, liquid water inputs are highest during the winter and lowest during the summer (due to the majority of precipitation falling across the catchment as rain, and the relatively insignificant effect of glacier accumulation and high melt in the catchment). Thus, annual runoff variability is again high but with maximum discharges in the winter and low flows in the summer. Between these extremes, in catchments with ~30–40% glacier cover, melt-dominated glacier runoff from the upper parts of the catchment compensates for the low rain-induced runoff from lower elevations during the summer. Conversely, reduced runoff from the upper parts offsets higher rain-induced runoff from lower elevations in winter. The net effect is a relatively even distribution of runoff throughout the year.

Finally, in midlatitude catchments, the time of maximum summer runoff is increasingly delayed with increasing glacier cover (Meier and Tangborn, 1961; Stenborg, 1970, Fountain and Tangborn, 1985; Pertziger, 1990). Furthermore, the relationship between runoff peak timing and percentage glacier cover appears to be nonlinear. For example, in the North Cascades, Washington, USA, maximum runoff was delayed by 1 month from mid-June to mid-July as glacier cover increased from ~5% to ~25%. It was delayed by a further 2 weeks to late July as glacier cover increased to ~50%, but by only 1 more week as cover increased to ~100% (Fountain and Tangborn, 1985).

INTRASEASONAL VARIATIONS IN RUNOFF

Glacierized basins not only have pronounced year-to-year and season-to-season runoff variations, but major intraseasonal runoff fluctuations too. This is especially true during the spring and summer months for midlatitude glaciers and, to a lesser extent, for high-latitude glaciers. It appears to be less true for low-latitude glaciers.

Midlatitude Glaciers

Midlatitude glaciers undergo major changes in their hydrology during the spring and summer months because of seasonal changes in both water inputs and water routing.

Seasonal Changes in Water Inputs

Over the ablation season, water inputs increase as the transient summer snowline migrates upglacier largely because of increases in net short-wave radiation but also increases in sensible and latent heat (Arnold *et al.*, 1996; Fountain, 1996; Willis *et al.*, 2002). As the snowpack melts, metamorphoses, and thins, its albedo drops because of (i) higher water content; (ii) larger snow crystals; (iii) higher dust concentration at the snow surface; and (iv) the presence of underlying ice, which, typically, has a lower albedo than snow and influences the albedo of the overlying snowpack (Wiscombe and Warren, 1980; Brock *et al.*, 2000). Once ice is exposed, the albedo of the glacier surface often drops dramatically, mainly because ice has larger crystal sizes and higher debris concentrations than snow (Mattson *et al.*, 1993; Brock *et al.*, 2000). Thus, as albedo falls during spring and summer, surface melt rises because of increases in net short-wave radiation. Furthermore, ice surfaces are typically rougher than snow surfaces (van de Wal *et al.*, 1992; Duynkerke and van den Broeke, 1994). Surface roughness scales directly with the aerodynamic roughness lengths for wind, air temperature and vapor pressure (Munro, 1989). Thus, as surface roughness increases during spring and summer, surface melt rises because of increases in sensible and latent heat transfer between the atmosphere and the glacier surface.

Seasonal Changes in Water Routing

The speed at which water is routed through glacierized catchments increases over the spring and summer (Hock and Noetzli, 1997; Arnold *et al.*, 1998; Willis *et al.*, 2002). There are three aspects to this: (i) supraglacial routing (vertically through snow, laterally through snow, and across ice); (ii) routing through firn; and (iii) englacial/subglacial routing (down crevasses and moulins, through pipes and conduits, and along the subglacial drainage system).

As the snowpack melts, metamorphoses and thins, the hydraulic conductivity (velocity of water movement) of the upper unsaturated layers increases for four main reasons. First, the snow crystals enlarge, increasing the saturated permeability (Shimuzu, 1970). Second, the melt rate and water content increase, raising the percentage water saturation (Colbeck, 1978). Since the hydraulic conductivity of

snow depends on its effective permeability, which depends on both its saturated permeability and its percentage water saturation, bigger crystal sizes and more melt mean higher vertical velocities (Colbeck and Anderson, 1982). Third, the development of preferential flow paths due to lateral heterogeneity (e.g. discontinuous layers or ice lenses), or heterogeneous infiltration patterns (e.g. heterogeneous surface melt) will speed up water movement (Marsh and Woo, 1984). Fourth, a thinning snowpack means decreasing vertical distance to travel, which also reduces the travel time of water movement to the base of the snowpack.

The speed at which water is routed laterally in a saturated layer at the base of the snowpack depends on snow crystal size, water content, and surface slope (Colbeck, 1978). As these are fairly constant during the ablation season, the lateral hydraulic conductivity of saturated snow may not change significantly. However, lateral water velocities at the base of supraglacial snowpacks may depend less on grain scale phenomena and more on larger scale processes. Observations at Haut Glacier d'Arolla, Switzerland suggest that water flows in a thin layer at the base of the snowpack along preferential flowpaths in rills and small channels incised into the ice. If the hydraulic efficiency of these increases over time, the speed at which water moves laterally at the base of the snowpack may also increase. Also, if the depth of saturated snow increases in some areas more than others, hydraulic gradients may increase lateral water velocities beyond those determined by surface slopes alone.

Vertical and lateral routing of water through firn is analogous to vertical and lateral movement through snow (Fountain, 1996; Schneider, 2000). Water tends to drain from firn layers during the winter. During spring and summer, water reaching the base of a snowpack overlying firn infiltrates the firn, moving at speeds determined by crystal size and water content. Once the water reaches the firn-ice transition, a firn water table builds up and lateral water movement is governed by elevation and pressure head gradients. Average vertical and lateral water velocities through firn vary over similar ranges of $\sim 0.1 - 0.35\,\mathrm{m\,h^{-1}}$ (Schneider, 2000). Seasonal changes in water velocities through firn have not been measured, although they may be expected to increase during spring and summer as the water content of the unsaturated zone increases, and hydraulic gradients in the saturated aquifer rise.

Seasonal changes in englacial and subglacial drainage pathways also increase the speed at which water is routed through glacierized catchments (*see* **Chapter 167, Subglacial Drainage, Volume 4**). Englacial pipes and conduits, and subglacial channels tend to close during the winter as the discharge of water flowing through them drops. In spring, once water reaches the base of the winter snowpack, the flux of water through crevasses and englacial pipes and conduits increases, and they enlarge through melting and become more hydraulically efficient. Eventually, water reaches a distributed hydrological system at the bed, which enlarges mechanically because of sliding and hydraulic jacking (Iken, 1981), and thermally because of ice melting (Kamb, 1987). If the flux of water reaching the bed is large enough, a distributed hydrological system will be unable to transmit the water, and parts of the system will enlarge to form channels, which have a greater hydraulic efficiency (Kamb, 1987; Nienow *et al.*, 1998). This switch from predominantly distributed to channelized drainage may be triggered by high melt rates or rainstorms, and may be associated with the release of water stored englacially and/or subglacially (see the following section).

Seasonal Changes in Runoff

Seasonal changes in water inputs and routing in midlatitude glacierized basins have a big impact on the shape, peak magnitude, and timing of diurnal discharge variations in proglacial streams (Elliston, 1973; Collins, 1982; Röthlisberger and Lang, 1987; Fountain, 1992; Gurnell *et al.*, 1992, 1994; Willis and Bonvin, 1995; Hannah *et al.*, 1999, 2000). Figure 4 shows how diurnal discharge hydrographs change through an ablation season. Discharges increase from May to August, and then drop during September because of variations in water inputs driven by changing patterns of net short-wave radiation and sensible and latent heat fluxes. Between May and September, the diurnal discharge range increases progressively, the lag between minimum and maximum discharge steadily declines, and the gradients of the rising and falling limbs of the hydrographs systematically increase. These changes reflect increasing hydraulic efficiency of supraglacial, englacial, and subglacial pathways associated with the gradual depletion of the snowpack and the evolution of the englacial and subglacial drainage system.

Seasonal changes in surface conditions and water routing can also affect a glacier's response to rainfall events (Collins, 1995; Denner *et al.*, 1999). A snowpack and a hydraulically inefficient englacial/subglacial drainage system in early summer can significantly dampen the effects of rainstorms on proglacial stream discharges. In late summer, large areas of ice and hydraulically efficient englacial/subglacial drainage mean that proglacial discharges respond much more rapidly to rainstorms.

Several recent studies have used statistical techniques to divide melt seasons into distinct periods on the basis of proglacial stream hydrograph characteristics (Hannah *et al.*, 1999, 2000; Lafreniere and Sharp, 2003). Hannah *et al.* (1999) used Principal Component Analysis and Cluster Analysis to identify diurnal hydrographs with distinct shapes and magnitudes during the 1995 and 1996 summers at the Taillon Glacier, France. The temporal sequencing of hydrographs with distinct characteristics was used to divide the melt seasons into different periods. The differences

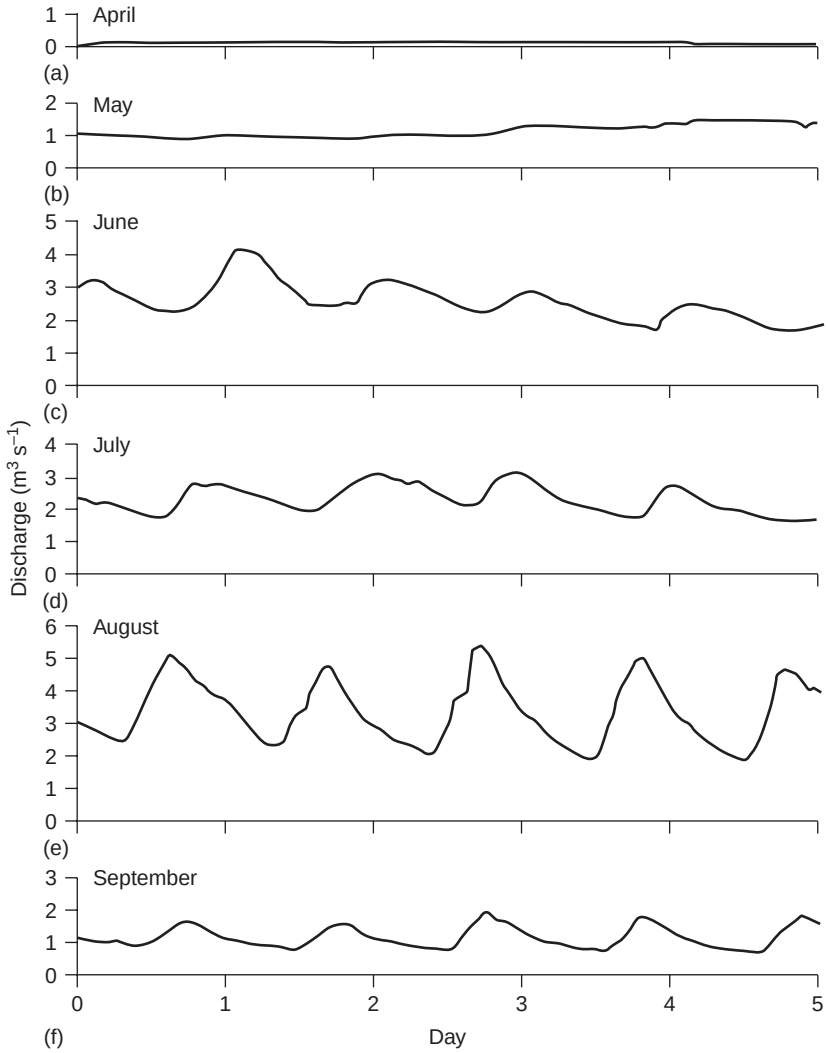


Figure 4 Changing shape of proglacial stream hydrographs from Haut Glacier d'Arolla, Switzerland in 1993 (from Willis and Bonvin, 1995)

in hydrograph characteristics, together with differences in energy receipt, ablation, and rainfall between the periods were interpreted in terms of temporal changes in both surface water inputs and the characteristics of the glacier's drainage system. During the 1995 summer, there was a progressive change from "building", "low magnitude" hydrographs, representing early season snowmelt and slow routing and high storage in the snowpack, through "low-intermediate peaked" hydrographs, reflecting increasing exposure of ice, rising melt rates and more rapid glacial routing of water, to "high-intermediate" and "high peaked" hydrographs signifying maximum exposure of ice, high melt and rainfall inputs, fast glacial routing, and low water storage. During the 1996 summer, however, there was a less obvious progressive change in hydrograph characteristics, reflecting a greater dominance of rainfall inputs and a less distinct evolution of the glacier's drainage system.

Lafreniere and Sharp (2003) used Wavelet Analysis to determine individual and cross wavelet spectra for air temperature, rainfall, and discharge time series from glacierized and nonglacierized catchments feeding Bow Lake, Alberta, Canada, during four summer melt seasons. The differences between the catchments were most marked in 1998, a strong El Niño year with low winter precipitation and high spring and summer air temperatures. In this year, the 24-h power in the discharge series was particularly low for the nonglacierized catchment and unusually high for the glacierized catchment, reflecting the lower snow melt contribution in the former and higher ice melt contribution in the latter compared with other years. In the glacierized catchment, there was a strong and persistent short scale ($\sim$12 h) power in the discharge series during July and August and a shift in the phase lag from $\sim$7 h to 4.5 h in early July, suggesting snowpack depletion and the evolution

of the glacier's drainage system. A low phase difference between air temperature and discharge also showed that the routing of diurnal melt cycles through the glacier was more hydraulically efficient in 1998 than in the other years. In the nonglacierized catchment, the relatively weak 24-h power in the discharge series was maintained throughout the summer in 1998 but largely disappeared during July and August in other years. This suggests that patches of ice were exposed in 1998 but not in normal years. The variability in discharge at short scales (<24 h) and rapid changes in the temperature–discharge phase difference suggest that precipitation was a more dominant component of runoff and was routed more quickly through the nonglacierized catchment than through the glacierized one.

High-latitude Glaciers

Compared with midlatitude temperate glaciers, very few studies have investigated seasonal variations in water inputs, routing, and runoff from high-latitude cold and largely cold polythermal glaciers (Hodgkins, 1997).

Seasonal Changes in Water Inputs and Routing

The few studies from the Arctic confirm that, as at midlatitude glaciers, water inputs increase over the summer largely because of decreases in albedo and increases in net short-wave radiation (e.g. Arendt, 1999). However, there are important differences in the way water is routed through catchments containing cold and polythermal glaciers and the ways in which this routing changes over the summer.

First, the end-of-winter snowpack is typically much colder than midlatitude glaciers. As water infiltrates the snow, large volumes refreeze to form ice lenses and superimposed ice layers (Repp, 1988; Bøggild, 2000; Wadham and Nuttall, 2002). The associated release of latent heat is the main way in which the snowpack warms up. Once the snowpack reaches 0 °C, some water is used to supply the irreducible water saturation (water retained by capillary forces). A depth of water, equivalent to >50% of the winter snowpack depth may be needed to raise the temperature to 0 °C and satisfy the irreducible water saturation (Marsh and Woo, 1984; Reeh, 1991). The effect is to increase the delay between initial melt and arrival of meltwater at the snowpack base compared with midlatitude glaciers. On John Evans Glacier, Ellesmere Island, Nunavut, Canada, for example, the delay is about 2 weeks on the lower glacier and over 3 weeks on the upper glacier where snow depths are only ~0.1 m w.e. (Boon *et al.*, 2003).

Second, once water has reached the snowpack base, it tends to pond in surface depressions, since cold ice is largely impermeable (Liestøl *et al.*, 1980; Hagen *et al.*, 1991; Boon *et al.*, 2003). On Austre Brøggerbreen, Svalbard, lakes appear between late May and early June, last between 2 and 10 weeks, reach volumes of up to ~9500 m^3, empty quickly over 1 or 2 days as supraglacial streams filled by snow or superimposed ice become unblocked, and then drain to ice-marginal channels (Liestøl *et al.*, 1980; Hagen *et al.*, 1991). Similar lakes appear and drain on John Evans Glacier, although others drain after moulins open up ~4 km from the glacier terminus, allowing water to reach the bed (Boon *et al.*, 2003).

Third, cold glaciers such as Austre Brøggerbreen and Scott Turnerbreen, Svalbard, are essentially impermeable and surface water flows largely in supraglacial and ice-marginal streams (Hagen *et al.*, 1991; Hodgkins, 1997; Hodson and Ferguson, 1999). Tracer tests from a meandering supraglacial stream and a boulder-strewn marginal stream at Hannabreen produced slow, dispersed returns, suggesting that routing of water at cold glaciers is likely to be hydraulically inefficient (Vatne *et al.*, 1995). Moulins feeding englacial and ice-marginal channels may be present, although these may be incised supraglacial streams or crevasses, which have been enclosed by ice creep but remain open due to heat advection and dissipation of flowing water. Tracer tests conducted from such moulins on Austre Brøggerbreen had high throughflow velocities, suggesting that englacial routing of water may be hydraulically efficient (Hagen *et al.*, 1991). However, cold glaciers do not have much of an englacial drainage system and they lack a subglacial drainage system altogether. Thus, the drainage system cannot evolve through the summer as it can on temperate glaciers.

By contrast, largely cold polythermal glaciers with a warm-based core and cold-based margins do have an englacial/subglacial drainage system. Such glaciers include Midre Lovénbreen, Erikbreen, Finsterwalderbreen, Kongsvegen, Hannabreen, and Bakaninbreen, Svalbard (Vatne *et al.*, 1992, 1996; Hagen *et al.*, 1993; Hodson and Ferguson, 1999; Wadham *et al.*, 2001; Murray and Porter, 2001; Rippin *et al.*, 2003) and John Evans Glacier, Ellesmere Island, Nunavut, Canada (Skidmore and Sharp, 1999; Boon *et al.*, 2003). Very little is known about this system or how it evolves through time. Flow separation techniques using proglacial stream discharges and electrical conductivities suggest that 30–40% of water is routed slowly, possibly subglacially, through Hannabreen (Vatne *et al.*, 1996). Furthermore, concentrations of sulfate in the proglacial stream showed no change through time, implying a relatively stable subglacial drainage system. Tracer tests from moulins on Erikbreen had low throughflow velocities and high dispersivities, suggesting that the glacier is underlain by a distributed drainage system which does not evolve significantly over the summer (Vatne *et al.*, 1995). One of the most important aspects of the hydrology of many largely cold polythermal glaciers is the sudden emergence of subglacially routed water at the glacier terminus in the form of pressurized fountains on the glacier or, more commonly,

Figure 5 Artesian fountain observed ~0.5 km from the terminus of John Evans Glacier, Ellesmere Island, Nunavut between 29th June & 5th July 1998 (Reproduced from Copland *et al.,* 2003 by permission of International Glaciological Society; Photo credit: Martin Sharp)

pressurized proglacial upwellings (Figure 5). These demonstrate that a subglacial drainage system is present beneath the warm-based core of the glacier and that this is able to extend beneath the cold-based margins of the glacier at some point during the summer. At John Evans Glacier, the breaching of the "thermal dam" occurs after hydrofracture events allow the sudden transfer of supraglacially stored water into the subglacial drainage system (Copland *et al.*, 2003). This phenomenon also involves the storage and release of subglacial water and small floods in proglacial streams (see later sections).

Seasonal Changes in Runoff

The three important differences between the hydrology of midlatitude and high-latitude glaciers produce distinct differences in the runoff regimes of their proglacial streams (Hodson *et al.*, 1998; Hodgkins, 2001; Irvine-Fynn *et al.*, 2005). First, the delay between the onset of melt and the rise in proglacial discharge is often greater at high-latitude glaciers because of the greater volumes of water stored supraglacially, ice-marginally and at polythermal glaciers, englacially and subglacially. Second, diurnal variations in proglacial discharge are often absent or more subdued at high-latitude glaciers, particularly early in the summer, because of the delaying and dampening effects of supraglacial slush and ponds, hydraulically inefficient meandering supraglacial and boulder-strewn marginal channels, and at mainly cold polythermal glaciers, subglacial water storage above the thermal dam. Third, the relatively sudden release of water from supraglacial ponds and, on mainly cold polythermal glaciers, from subglacial locations above the thermal dam produces a marked flood hydrograph, often lasting a few days.

Low-latitude Glaciers

There have been very few investigations into the hydrology of low latitude glaciers. Tracer experiments and flow separation techniques based on proglacial stream discharge and electrical conductivity measurements led Wagnon *et al.* (1998) to conclude that Zongo Glacier, Bolivia had a more "highly developed" englacial/subglacial drainage system than similar sized midlatitude glaciers. They suggested that year round ablation maintained the hydrological system, preventing conduits and channels from closing down, as occurs on mid- and high-latitude glaciers during the winter. Dye tracer experiments at Dokriani Glacier, India, suggest that the glacier is underlain by efficient channelized drainage in July, distributed drainage in August, and channelized drainage again in September (Hasnain *et al.*, 2001). However, the results are equivocal since they are based on only 10 experiments from three locations.

SEASONAL WATER STORAGE

As mentioned previously, glaciers in mid- and high latitudes store water as snow in winter, and release it as snow and ice melt during the summer. However, glaciers can also store and release liquid water on a variety of timescales (Jansson *et al.*, 2003). On several glaciers, seasonal variations in liquid water storage and release have been examined using a water balance approach, by comparing time series of surface water inputs (from melt and rainfall) with proglacial stream outputs (Figure 6). Most studies show that water is stored during spring and early summer and then released. The exception is Scott Turnerbreen, which showed initial release followed by storage, although this might reflect the short period of measurement. The maximum amount of water stored varies between glaciers, from <0.1 m (averaged over glacier area) at Haut Glacier d'Arolla, Scott Turnerbreen, and Midre Lovenbreen to ~0.4 m at Unteraargletscher (Figure 6). The timing of maximum storage also varies between glaciers, from early June on South Cascade Glacier to late August on Midtdalsbreen, although this may, at least in part, reflect the different measurement start dates. Several studies (Mikkaglaciären, South Cascade Glacier, Haut Glacier d'Arolla, Midre Lovénbreen) show that the amount of water released is greater than that stored, implying that additional water must have been stored before measurements began. Others (Midtdalsbreen, Unteraargletscher) show that the amount released is less than that stored, suggesting that more may have been released after measurements finished or that there are significant errors in the calculations. Only Storglaciären showed an approximate net balance over the entire measurement period. Unfortunately, none of these studies has covered an entire hydrological year, and none has examined interannual water storage variability by investigating more than one season (Jansson *et al.*, 2003).

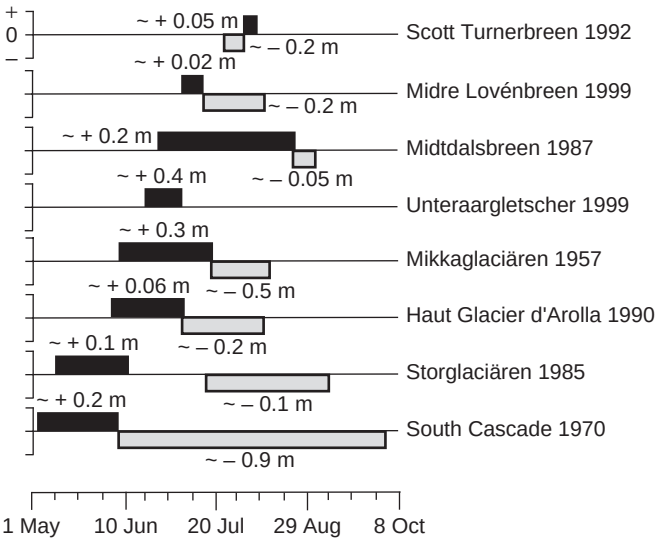


Figure 6 Patterns of storage gain (positive, black) and loss (negative, white) for several glaciers discussed in the text. (based on Figure from Jansson *et al.*, 2003)

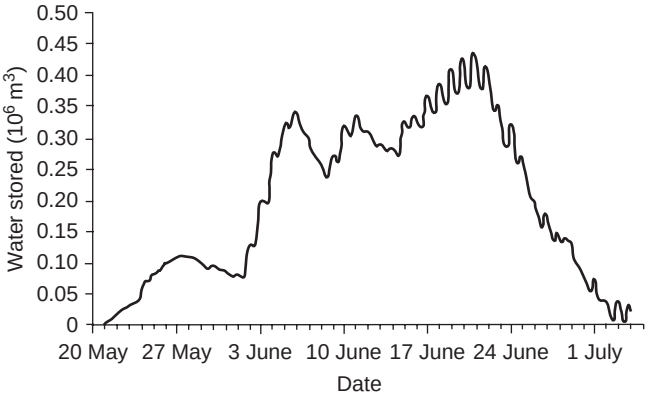


Figure 7 Cumulative total snowpack water storage calculated using a physically based distributed model for Haut Glacier d'Arolla, Switzerland in the year 2000 (Courtesy of Andrew Fox)

All the studies listed above have calculated bulk water storage and release for entire glaciers by comparing water input and water output time series. Such an approach does not reveal where the water is stored, and there is still uncertainty about the relative importance of supraglacial, englacial, subglacial, and ice-marginal storage, and whether the relative importance of these stores varies through the summer. These issues are discussed further in the following sections.

Supraglacial Storage

Large volumes of water may be stored in snow, particularly at the start of the melt season (Willis *et al.*, 1993; Fountain, 1996). This is particularly important on high-latitude glaciers where cold ice is largely impermeable and large slush zones develop (Liestøl *et al.*, 1980; Hagen *et al.*, 1991; Hodgkins, 2001; Boon *et al.*, 2003). This water tends to be released gradually during the summer as the snowpack melts, although it may drain rapidly in the form of slush avalanches (Smart *et al.*, 2000). On cold and polythermal glaciers, water may also be released rapidly as supraglacial and ice-marginal channels are unblocked, or as moulins open up (Boon *et al.*, 2003). This has important implications for both seasonal runoff (see the earlier section) and flooding (see the following section) in high-latitude glacierized basins. Some water may continue to drain from the remaining snowpack during the early winter. Some may refreeze during the winter to form superimposed ice, which accounts for 6–25% of the annual accumulation on Midtre Lovénbreen, Svalbard (Wadham and Nuttall, 2002).

Recent hydrological modeling work at Haut Glacier d'Arolla has calculated the temporal patterns of water storage in the supraglacial snowpack from late May to early July 2000 (Fox, 2003). A distributed snow energy and mass balance model was used to calculate spatial patterns of melt and vertical routing through the unsaturated snowpack. A modified groundwater model was then used to determine spatial patterns of lateral water routing and water storage in the saturated snowpack. Figure 7 shows the cumulative volume of water stored in the saturated snowpack. The daily water balance is usually positive between late May and mid-June with short periods of negative balance that last a few days. The cumulative balance reaches a maximum on 21 June when ~0.1 m of water averaged over the glacier is stored in the snowpack. This represents ~17% of the total meltwater production up to 21 June, and is more than the maximum volume stored seasonally in 1990 (Richards *et al.*, 1996). This water then drains from the snowpack over the next 2 weeks as the snowpack thins, and the snowline migrates rapidly upglacier. Figure 8 shows the spatial patterns of water storage in the saturated snowpack on 10 June, shortly before the time of maximum storage. Thus, most water is stored along valleys in the glacier surface where water tends to pond.

Firn Storage

Appreciable volumes of water can also be stored in and released from firn (Fountain, 1996; Fountain and Walder, 1998; Schneider, 2000; Jansson *et al.*, 2003). The firn zone is essentially an unconfined aquifer. The water table height depends on the surface recharge rate, the hydraulic properties of the firn, and the rate of drainage from the base, which is governed by the surface slope and the distribution of crevasses. Measurements on three glaciers, Aletschgletscher, Switzerland, South Cascade Glacier, Washington, USA, and Storglaciären, Sweden, show that the firn aquifer reaches thicknesses of

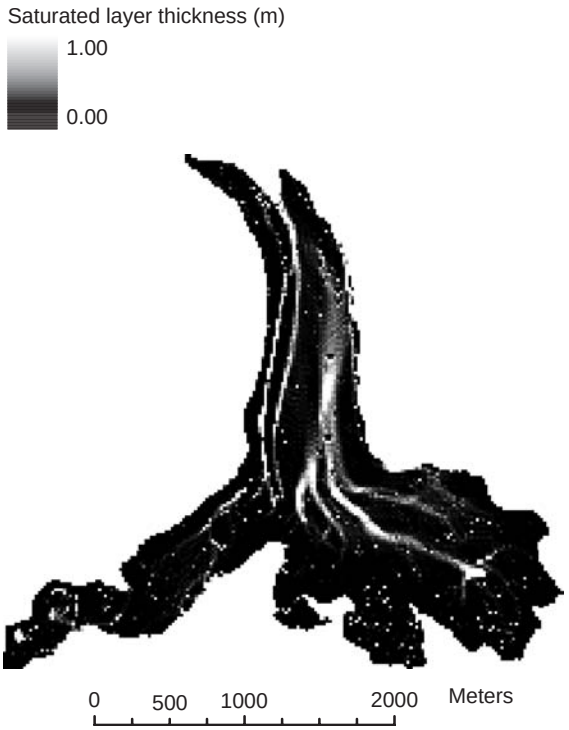


Figure 8 Spatial distribution of modeled snowpack water storage at Haut Glacier d'Arolla, Switzerland on 10th June 2000 (Courtesy of Andrew Fox)

~3–7 m during the summer. The firn tends to gain water during spring and drain during late summer. Day-to-day variations in melt and rainfall cause damped water-level fluctuations of 1–2 m lasting days to weeks, superimposed on the seasonal trends. As with water in the snowpack, some water in firn may refreeze during the winter, accounting for 7–64% of annual accumulation for different glaciers in Alaska (Trabant and Mayo, 1985), and ~4% of annual accumulation on Storglaciären (Schneider, 2000). Measurements of firn area, porosity, and water saturation at South Cascade Glacier suggested that ~12% of the maximum volume stored seasonally at this glacier may be stored in the firn (Fountain, 1989). Similar measurements suggest the equivalent figure maybe ~44% at Storglaciären (Östling and Hooke, 1986).

Englacial Storage

In addition to firn storage, Östling and Hooke (1986) suggested that englacial reservoirs such as crevasses might be important storage sites on Storglaciären. Fountain and Walder (1998) argue that the majority of water storage on temperate glaciers is probably englacial. They refer to borehole-video studies, which show that englacial voids and conduits represent a macroporosity of ~1%, and to radar studies and work measuring water levels in moulins, which show how englacial passageways fill and drain in response to variations in water inputs. This macroporosity is easily able to accommodate the maximum depths of seasonal water storage measured on several glaciers (Figure 6).

Subglacial Storage

Glacier surfaces often rise and fall in response to weather-related variations in surface water inputs and/or changes in the morphology of the subglacial drainage system (Iken *et al.*, 1983; Iken and Bindschadler, 1986; Kamb and Engelhardt, 1987; Jansson and Hooke, 1989; Raymond *et al.*, 1995; Mair *et al.*, 2002). Although some of this movement may be due to strain-induced variations in ice thickness, some is also due to the growth and shrinkage of subglacial cavities and associated changes in subglacial water storage. Fountain and Walder (1998) estimate that linked cavities may account for ~10–30% of the volume of water stored seasonally in several glaciers (Figure 6). They also estimate that subglacial till may account for up to ~40%, but probably much less. Tangborn *et al.* (1975) suggested that most water is stored in subglacial locations beneath South Cascade Glacier, although they had no direct evidence for this. They suggested that a subglacial conduit system closes down in places during periods of low or zero surface water inputs during the winter, and then fills with water during the spring and early summer before reopening and releasing the stored water later in the summer. A similar mechanism was advocated to explain patterns of water storage and release at Haut Glacier d'Arolla (Richards *et al.*, 1996). These authors linked storage/release patterns to rates of water input and the evolution of the subglacial drainage system as determined by dye tracing experiments (Nienow *et al.*, 1998).

Similarly, Schuler *et al.* (2002) observed that over most of Unteraargletscher, the snowpack was already saturated in mid-June when water balance calculations began. They concluded that between mid-June and early July, water storage occurred within and at the base of the glacier. This was supported further by observations of a rainfall-induced rise in water pressure measured in a borehole ~3 km from the glacier terminus, followed by a flood in the proglacial stream. The rainstorm caused more water to back up in the englacial/subglacial drainage system causing water pressures to rise. This triggered a reorganization of the subglacial drainage system, leading to the release of water stored in and beneath the glacier.

Rippin *et al.* (in press) showed how the dynamics of the mainly cold polythermal Midtre Lovénbreen are affected by patterns of water storage and release. The upper tongue accelerated during times of water storage but the lower tongue accelerated during a time of water release, which also coincided with the emergence of a pressurized proglacial upwelling. This implied that at least some

of the water was stored subglacially behind the "thermal dam", marking the boundary between warm-based and cold-based ice. The water was then released as it forced its way beneath the cold-based lower tongue. Similar behavior was observed at the mainly cold polythermal John Evans Glacier (Copland *et al.*, 2003). Early season velocity events affected large areas of the glacier tongue, and occurred when supraglacially stored waters connected to the subglacial drainage system, and triggered the release of subglacially stored water. High pressures above the thermal dam were shown by the emergence of an artesian fountain on the glacier surface (Figure 5) in addition to a pressurized proglacial upwelling.

Summary

Previous research into seasonal patterns of water storage suggests that most glaciers store water in spring and early summer and release it during the late summer. Snow is an important location for stored water, especially early in the melt season, and may account for up to 100% of the stored water in the spring. As the snowpack melts, the stored water is released. It may drain directly from cold glaciers or may enter firn, englacial, or subglacial locations on temperate and polythermal glaciers. Firn may account for ~10–40% of the maximum volume of stored water. Similarly, ~10–40% may be stored subglacially. However, especially on temperate glaciers, englacial water storage in pipes, conduits, crevasses, and moulins is likely to be more important than subglacial storage during mid- to late summer. Recent work from two polythermal glaciers suggests that water is stored subglacially beneath the warm-based core in early summer and then released as the water penetrates the cold-based margin.

Reviewing several studies of glacier water storage, Willis *et al.* (1993) suggested that there may be two main controls on seasonal patterns of water storage and release depending on the relative importance of snow/firn storage verses englacial/subglacial storage. On Midtdalsbreen, and possibly Storglaciären, most storage was in snow/firn, and release from storage tended to occur in response to declining water inputs. By contrast, on South Cascade Glacier and possibly Mikkaglaciären, significant storage occurred in englacial/subglacial locations, and the seasonal evolution of the glacier's drainage system dominated the pattern of water release. Recent work at Haut Glacier d'Arolla and Unteraargletscher suggests that they too fit into this latter category. Recent work on Arctic glaciers suggests that two more categories may be needed. On cold glaciers (e.g. Scott Turnerbreen), the release of water may occur relatively early in the melt season because of the depletion of the supraglacial snowpack and the water it contains. The release of water from mainly cold-based polythermal glaciers (e.g. Midre Lovénbreen and John Evans Glacier) may be driven largely by the timing of the breakthrough of subglacially stored water through the cold margin.

FLOODS IN GLACIERIZED BASINS

Floods are a common feature of many glacierized basins and a major natural hazard responsible for damage to human structures and loss of human life (Tufnell, 1984). A rare type of flood occurs when volcanic eruptions beneath or adjacent to glaciers produce hot gases, ash, or lava, causing exceptionally rapid melting and generating large volumes of water that immediately produce runoff (Trabant *et al.*, 1994; Smellie, 2002). For example, severe flooding in the proglacial stream of Drift Glacier, Alaska occurred when the 1989–1990 eruption of Redoubt Volcano melted over $100 \times 10^6\,m^3$ of snow and ice on the glacier, and a further $\sim 35 \times 10^6\,m^3$ of snow in the valley downstream (Trabant *et al.*, 1994). The absence of high-water marks in depressions and of ice-collapse features in the glacier showed that the melted water ran off immediately and was not stored in lakes.

A more common type of flood arises from the sudden drainage of lakes. The "catastrophic" floods produced by lake drainage are usually referred to by the Icelandic term jökulhlaups, although they are also known as débâcles in parts of Europe, aluviones in South America and chhugyümha in Nepal (Tweed and Russell, 1999). The most common types of lake producing jökulhlaups are either ice-dammed (subglacial or ice-marginal) or moraine-dammed (usually proglacial).

A final type of flood, usually smaller than those mentioned above, can be triggered by high meteorologically driven melt rates, rainfall, or the release of water stored in a glacier's englacial/subglacial drainage system.

Ice-dammed Subglacial Lakes

The most dramatic and best-documented subglacial lake floods involve the sudden drainage of lakes beneath the Icelandic ice caps of Vatnajökull (Thorarinsson, 1953; Björnsson, 1974, 1975, 1992, 1997; Guðmundsson *et al.*, 1995, 1997; Snorrason *et al.*, 1997; Roberts *et al.*, 2000, 2001) and Myrdalsjökull (Thorarinsson, 1957; Tómasson, 1996; Sigurðsson, 1999; Björnsson *et al.*, 2000; Roberts *et al.*, 2002, 2003). The elevated geothermal heat fluxes in the Grimsvötn and Katla calderas and occasional volcanic eruptions produce high subglacial melt rates. Hydraulic potential gradients cause water to collect in large subglacial lakes known as *cupolas*. A surface depression often forms over the cupolas, causing supraglacial meltwater to collect there too. Jökulhlaups associated with the drainage of Grimsvötn are particularly well studied. Water from the cupola drains beneath Skeiðarárjökull, an outlet glacier on the south side of Vatnajökull, every 6 years or so,

releasing up to 4.5 km^3 of water at maximum discharges of up to 50 000 $m^3 s^{-1}$ (Guðmundsson *et al.*, 1995). Less predictable floods are produced during times of particularly intense volcanic activity such as the fissure eruption in 1996 at Gjalp, just north of Grimsvötn (Snorrason *et al.*, 1997). Flood volumes and maximum discharges tended to decrease during the twentieth century because of declining geothermal and volcanic activity (Guðmundsson *et al.*, 1995), although such activity has been increasing recently and so jökulhlaups may become bigger once again (Björnsson, 1997).

Subglacial lakes are traditionally assumed to drain through a single subglacial conduit that exits the glacier terminus. According to the most widely cited model (Clarke, 1982; based on earlier work by Spring and Hutter (1981) and Nye (1976)), rising lake levels cause water pressures to approach ice overburden (~91% of ice depth), allowing the ice dam to be breached. Once water starts to leak through the dam, the advection of heat from the lake water and the dissipation of potential energy allow a conduit to develop. Initially, rapid melting causes the conduit to enlarge, allowing more water to escape by a positive feedback process. As the lake level drops, the conduit effective pressure (ice overburden minus water pressure) declines. Conduit closure due to ice creep exceeds enlargement due to melting, and eventually the conduit closes, the ice dam is reestablished, and the flood stops.

A limitation of the model is that subglacial lake drainage often starts before water levels reach ice overburden pressure at the dam. Grimsvötn floods usually start when lake levels are 20–50 m less than ice overburden pressures (Björnsson, 1992), although the 1996 jökulhlaup was one of the few triggered at ice overburden pressures (Björnsson, 2003). Several explanations have been advanced:

- Crevasses reduce the ice density above the dam (Thorarinsson, 1939).
- Drainage occurs through hydrofractures created in the ice above the dam (Glen, 1954; Higgins, 1970).
- Ice flexure means ice acts like a cantilever over the dam (Nye, 1976).
- A conduit always links the lake to the subglacial drainage system below the lake, but hydraulic gradients route water to the lake at low lake levels. Once the lake level is high enough and if the pressure drops low enough in the drainage system below the lake, a siphoning mechanism pulls water out of lake, although this mechanism can only pull ~10 m of water (Whalley, 1971; Fisher, 1973; Knudsen and Theakstone, 1988; Knight and Russell, 1993; Schöner and Schöner, 1997).
- Subglacial drainage below the dam switches from flow through sediment pores to flow in a broad low conduit (canal) as the lake level rises; this happens at a level less than the ice overburden pressure (Fowler and Ng, 1996).
- The pressure at which the dam breaks depends on the hydraulic gradient at the caldera rim and the speed at which the lake level rises (Fowler, 1999). A conduit always links the lake to the subglacial drainage system below the lake. At low lake levels, the dam occurs along the conduit some distance below the caldera rim and the hydraulic gradient routes water upglacier from the dam to the lake. As the lake level rises, the dam migrates upglacier toward the lake. If the lake level rises slowly (weeks–months) and if the dam is "strong" (large hydraulic gradient at rim), the lake level reaches ice overburden before the dam reaches the lake, and a flood is initiated due to flotation. If the lake level rises slowly but the dam is "weak" (small hydraulic gradient at rim), then the dam reaches the lake while the lake level is still below ice overburden, and a flood is initiated by conduit enlargement at the dam. Finally, if the lake level rises quickly (days–weeks), the dam migrates upglacier more slowly, so overburden pressures are more likely to be reached before a flood is initiated.

Hydrographs of Subglacial Lake Drainage Floods

According to Clarke's (1982) model, the flood hydrograph shape and its peak magnitude and timing are determined by three factors: (i) the temperature, volume, and hypsometry of the lake; (ii) the ice overburden pressure where the lake meets the dam/conduit; and (iii) the roughness and hydraulic gradient of the conduit. Largest peaks occur when the lake temperature is high, channel enlargement is dominated by the lake's thermal energy, and frictional dissipation is small. Floods draining through a subglacial conduit should have hydrographs with a long gentle rising limb (as the conduit gradually enlarges) and a steep abrupt falling limb (as the conduit closes or the lake is emptied). These characteristics match many observed flood hydrographs (Thorarinsson, 1957; Whalley, 1971; Mathews, 1973; Mottershead, 1975; Theakstone, 1978; Clarke and Waldron, 1984; Sugden *et al.*, 1985; Russell, 1989; Björnsson, 1992) (Figure 9a).

However, some jökulhlaup hydrographs are very different in shape and peak magnitude and timing from the predicted ones. Recent observations in Iceland provide a possible explanation. Björnsson (1998) compared jökulhlaup hydrographs before, during, and after the 1991 Skeiðarárjökull surge. After the surge, the hydrograph showed a gradual rise to a peak discharge of ~2000 $m^3 s^{-1}$ and a rapid decline, matching the pattern of lake water outflow calculated from measurements of ice subsidence above it (Figure 10a). During the surge, however, the hydrograph showed a gradual rise to a peak discharge of just ~300 $m^3 s^{-1}$, and a very gradual decline (Figure 10b). It did not match the pattern of lake outflow, and it involved the subglacial storage and release of up to ~200 $\times 10^6$ m^3 of water as it travelled toward the terminus (Figure 10b).

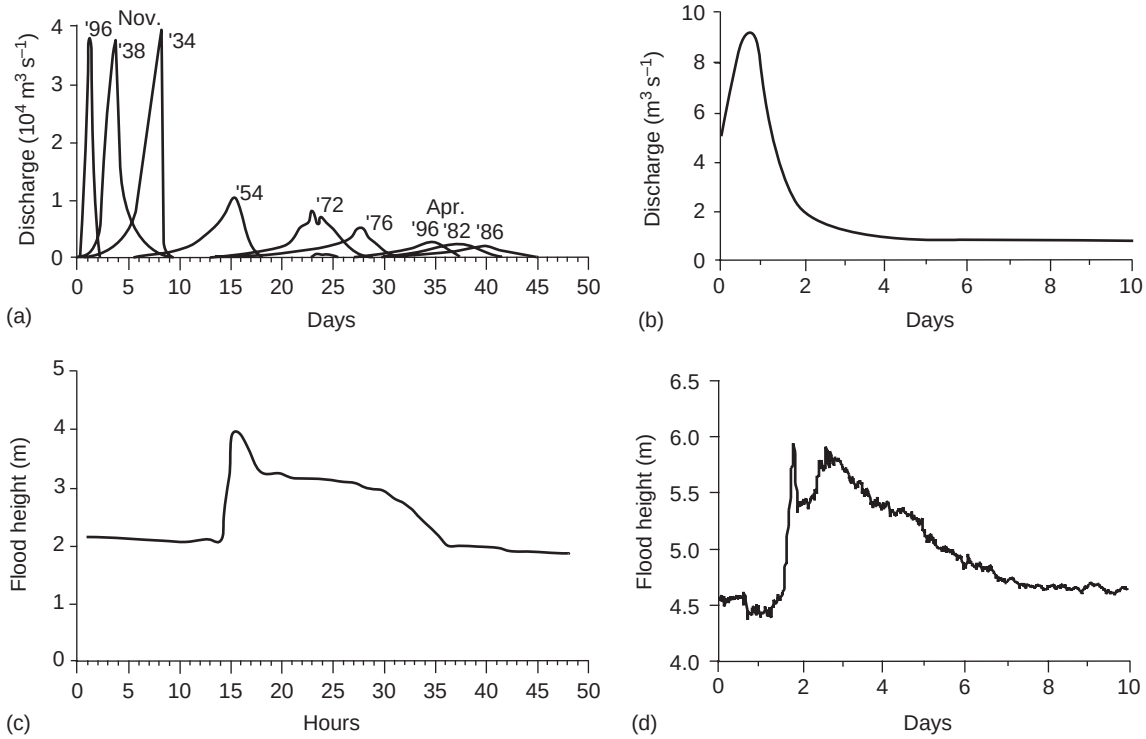


Figure 9 Hydrographs of floods from (a) Grímsvötn, Vatnajökull, Iceland in 1934, 1938, 1954, 1972, 1976, 1982, 1986 and 1996 (from Björnsson, 2003); (b) Aurora Lake, Black Rapids Glacier, Alaska in June 1993 (from Raymond and Nolan, 2000); (c) Luggye Tsho, Bhutan in October 1994 (from Richardson and Reynolds, 2000a); d) Franz Josef Glacier, New Zealand in February 2003 (from Goodsell *et al.*, 2005). All hydrographs are measured except for (b) which is modeled to fit daily measurements

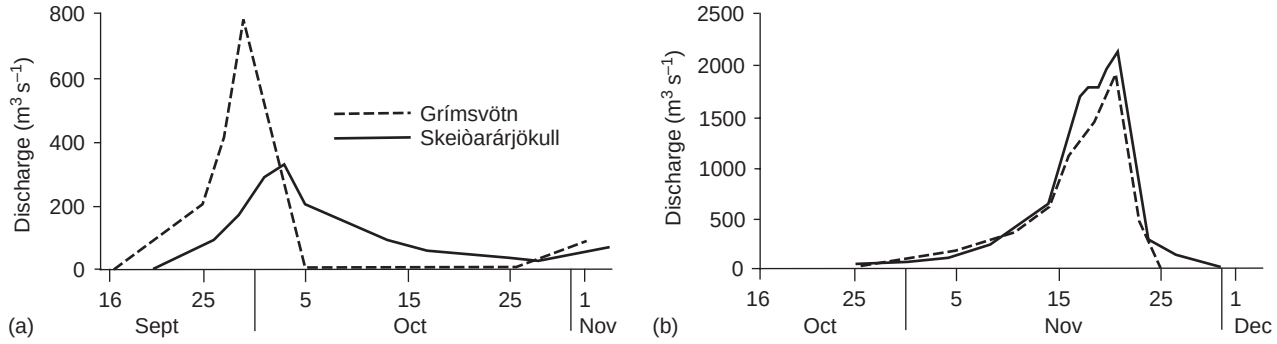


Figure 10 Drainage of outburst flood water from Grimsvotn and Skeiðarárjökull (a) during and (b) after the 1991 surge of Skeiðarárjökull (after Björnsson, 1998)

Since surging of at least some glaciers is associated with distributed linked cavity drainage (Kamb *et al.*, 1985), Björnsson (1998) concluded that the flood during the surge was routed through linked cavities, not a single conduit.

Recent observations during and after jökulhlaups at both Skeiðarárjökull and Sólheimajökull suggest that distributed drainage may occur beneath glaciers that are not surging, at least around glacier margins (Roberts *et al.*, 2000). Floodwater is partially evacuated from numerous ice-marginal and supraglacial outlets that change position over time, indicating hydraulic jacking and ice fracturing as subglacial water pressures vastly exceed ice overburden pressures. Thus, floodwater is distributed widely over the bed, as subglacial drainage pathways enlarge mechanically and not just thermally. Water may also be stored and released in locations other than subglacial cavities during flooding. During the 1999 Sólheimajökull jökulhlaup, subglacial water purged to ice-marginal locations and "retro-filled"

relict ice-dammed lake basins, before draining back into the glacier (Roberts *et al.*, 2003).

Flowers and others (2004) have developed a new model that accounts for some of the discrepancies between some observed jökulhlaups and calculations made with the original Clarke (1982) model. Floodwater initially propagates as a turbulent subglacial sheet, which feeds an enlarging system of conduits. The model was compared with the observed flood hydrograph in Skeiðarárjökull associated with the 1996 Gjalp eruption (Figure 9a). It successfully accounts for the ~10.5-h lag between lake drainage onset and the flood water arrival at the terminus, the lag between maximum rate of lake drainage and maximum stream discharge, and the high subglacial water pressures in the early stages of the flood.

Ice-dammed Marginal Lakes

Ice-marginal lake floods are widely reported in the literature (Thorarinsson, 1939; Liestøl, 1956, 1977; Stone, 1963; Maag, 1969; Björnsson, 1976; Young, 1980; Hewitt, 1982; Dawson, 1983; Clement, 1984; Sturm and Benson, 1985; Sugden *et al.*, 1985; Hagen, 1987; Desloges *et al.*, 1989; Mayo, 1989; Russell, 1989; Liu, 1992; Zhang, 1992; Del Valle *et al.*, 1995). Ice-marginal lakes form in three locations: (i) where a trunk valley glacier blocks the drainage of a glacier-free side valley; (ii) where a side-valley glacier extends across a glacier-free trunk valley; and (iii) at the confluence of two valley glaciers (Benn and Evans, 1998). They are especially common next to cold or mainly cold polythermal glaciers, where ice is frozen to its bed and water cannot drain englacially/subglacially.

Ice-marginal lakes are traditionally assumed to drain through either a subglacial conduit (see the earlier section) or a subaerial breach between the glacier and the valley side. It was originally thought that subaerial breach drainage would be uncommon since flotation and subglacial drainage would occur before lake levels reached the height of the glacier surface to initiate a breach (Miller, 1952; Liestøl, 1956). However, subaerial breach drainage is now recognized as an important mechanism for lakes next to cold or mainly cold polythermal glaciers where ice is frozen to the bed and flotation is more difficult (Howarth, 1968; Maag, 1969; Costa and Schuster, 1988). It may also be important on glaciers with high debris contents, as this will also make flotation more difficult (Tweed and Russell, 1999). Drainage may begin in a valley-side conduit where ice is thin, but a subaerial breach may develop subsequently if the conduit roof collapses or is melted completely by flowing water (Fountain and Walder, 1998).

Hydrographs of Ice-marginal Lake Drainage Floods

Walder and Costa (1996) produced a model for breach drainage outburst floods, assuming that channel widening occurs by sideward ice melting (ignoring the possible role of ice calving) and ignoring the negligible role of ice creep on channel closure. Like Clarke's (1982) model for subglacial conduit drainage (see the earlier section), the flood hydrograph shape depends mainly on the temperature, volume, and hypsometry of the lake and the roughness and hydraulic gradient of the channel. Again, like Clarke's (1982) model, largest peaks result when the lake temperature is high, channel enlargement is dominated by the lake's thermal energy, and frictional dissipation is small.

Raymond and Nolan (2000) described the drainage of an ice-marginal lake at Black Rapids Glacier, Alaska, that drained across the surface of the glacier through a supraglacial ice spillway. Their spillway drainage model accounts for spillway lowering by ice melt. A jökulhlaup (rapid unstable discharge) occurs if the spillway elevation melts down faster than the lake level falls, otherwise stable discharge occurs. According to the model, a jökulhlaup occurs when the lake area exceeds some critical value, governed by the lake temperature and the slope and width of the spillway as it leaves the lake. For typical spillway slopes of ~0.1 and widths of ~3 m, a lake area of $>1\,\mathrm{km}^2$ is required to initiate a jökulhlaup if lake temperatures are 0 °C. If temperatures are slightly higher than 0 °C, a jökulhlaup can occur for much smaller lake areas.

Floods along ice-marginal channels or supraglacial spillways typically possess a more symmetrical steep rise to and fall from peak than subglacial conduit floods. They tend to last just a few minutes to hours, but may last days if the lake volume is large (Thorarinsson, 1939; Haeberli, 1983; Dawson, 1983; Raymond and Nolan, 2000) (Figure 9b).

Moraine-dammed Lakes

The drainage of moraine-dammed proglacial lakes has been widely reported in the literature (Lliboutry *et al.*, 1977; Haeberli, 1983; Buchroithner *et al.*, 1982; Hewitt, 1982; Clague *et al.*, 1985; Fushimi *et al.*, 1985; Ives, 1986; Vuichard and Zimmermann, 1986, 1987; Carling and Glaister, 1987; Costa, 1988; Costa and Schuster, 1988; Clague and Evans, 1994, 2000; Ding and Liu, 1992; Yongjian and Jingshi, 1992; Mathews and Clague, 1993; O'Connor and Costa, 1993; Yamada and Sharma, 1993; Mool, 1995; Popov, 1997; Reynolds, 1998; Yamada, 1998; Cenderelli, 2000; Richardson and Reynolds, 2000a; Cenderelli and Wohl, 2001, 2003; Haeberli *et al.*, 2001; Huggel *et al.*, 2002). Such lakes form where a glacier terminates on ground that slopes back toward the glacier. A typical situation is where retreat of a glacier allows a lake to form behind a terminal moraine.

Proglacial lakes may drain catastrophically through a subglacial conduit or subaerial breach next to the glacier, in which case they behave like ice-marginal lakes described previously. However, floods from proglacial lakes usually

occur through the moraine dam. This may involve progressive groundwater flow and piping through the dam leading to simple mechanical failure, but it often involves the melting of ice if the moraine is ice cored (Buchroithner *et al.*, 1982; Fushimi *et al.*, 1985; Richardson and Reynolds, 2000b). One of the most common triggering mechanisms involves lake water overtopping part of the dam, initiating rapid and accelerating incision. This may be caused by catastrophic drainage into the lake from ice-marginal or subglacial sources upglacier, by rapid glacier advance, or by large waves caused by glacier calving or snow, ice, or rock avalanching (Lliboutry *et al.*, 1977; Grabs and Hanisch, 1993; Ives, 1986; Vuichard and Zimmermann, 1986, 1987; Clague and Evans, 1994).

Hydrographs of Moraine-dammed Lake Drainage Floods

There are virtually no recorded hydrographs associated with moraine-dammed lake floods. An exception is the flood from Luggye Tsho, Bhutan, in 1994, for which a complete stage hydrograph measured ~100 km downstream is available (Richardson and Reynolds, 2000a) (Figure 9c). It shows a rapid one-hour rise to peak stage, followed by a moderate fall and then a more gradual fall. The entire flood lasted ~22 h. Another fairly well-documented moraine-dammed lake flood occurred in 1985, when the terminus of Langmoche Glacier, Nepal, collapsed into a proglacial lake. A wave travelled along the lake, and overtopped the moraine dam, initiating dam erosion and ultimately failure. The initial flood discharge was ~2000 $m^3 s^{-1}$, with an average discharge of ~500 $m^3 s^{-1}$ over 4 h, draining a total volume of ~6–10 $\times 10^6$ m^3 of water (Vuichard and Zimmermann, 1986, 1987). This 1985 flood was not the first from the lake. Cenderelli and Wohl (2001) used a step-backwater model together with geomorphic evidence downstream of the Langmoche Glacier to estimate the seasonal peak discharges associated with rainfall and snow and ice melt, and those associated with the repeated lake drainage. Peak discharges from rainfall and melt ranged from 7 to 205 $m^3 s^{-1}$; those from the proglacial lake floods were up to 60 times greater and ranged between 400 and 2350 $m^3 s^{-1}$.

Because of their damage potential, mitigation schemes have been introduced at some lakes known to be prone to such floods in an attempt to keep lake levels below critical thresholds (e.g. Grabs and Hanisch, 1993; Ames, 1998; Reynolds, 1998; Reynolds *et al.*, 1998; Haeberli *et al.*, 2001).

Drainage from the Englacial/Subglacial Drainage System

Although the most dramatic outburst floods are caused by catastrophic drainage of subglacial, ice-marginal, or proglacial lakes, floods can also be caused by the release of water stored in the englacial/subglacial drainage system and are often associated with high melt rates and/or rainstorms. Such floods have been described from both temperate and mainly cold polythermal glaciers.

Temperate Glaciers

On temperate glaciers, floods are most prevalent in spring, shortly after melt onset (Beecroft, 1983; Haeberli, 1983; Driedger and Fountain, 1989; Warburton and Fenn, 1994; Walder and Driedger, 1995; Barrett and Collins, 1997; Collins, 1998; Anderson *et al.*, 1999; Schuler *et al.*, 2002). If the water is stored englacially, the release mechanism may involve the forced connection of isolated englacial voids, pipes, and conduits to the subglacial drainage system (Fountain and Walder, 1998). However, there is strong evidence to suggest that at least some of the water is stored subglacially and that its release involves the rapid conversion of a distributed subglacial drainage system with low transmissivity and high storage potential to a channelized system that transmits water more rapidly and stores less water.

Rapid surface water inputs cause high subglacial water pressures, initiate instabilities at the glacier bed, and lead to a reorganization of the subglacial drainage system (Stone and Clarke, 1996; Barrett and Collins, 1997; Gordon *et al.*, 1998; Kavanaugh and Clarke, 2001). The conversion may involve the unstable enlargement of links between basal cavities through melting (Walder and Driedger, 1995; Nienow *et al.*, 1998), a mechanism initially advocated to explain the termination of some glacier surges (Kamb, 1987). However, many spring floods are also associated with increases in glacier surface velocities, surface uplift, reductions in basal drag, increases in basal sliding and the evacuation of subglacial sediment (Collins, 1989; Iken *et al.*, 1983; Hooke *et al.*, 1989; Anderson *et al.*, 1999; Mair *et al.*, 2001, 2002, 2003; Swift *et al.*, 2002). Taken together, these phenomena are referred to as *spring events* (Röthlisberger and Lang, 1987). Thus, conversion from distributed to channelized drainage must also involve mechanical enlargement of cavities through hydraulic jacking (Röthlisberger and Iken, 1981), in addition to thermal enlargement of the links between them.

Several mechanisms appear to explain a series of rainfall-induced floods observed at Franz Josef Glacier, New Zealand since the 1920s (Davies *et al.*, 2003; Goodsell *et al.*, 2003, 2005). There appear to be four distinct types of flood, which are not restricted to the spring and early summer. The first three occur at or near the glacier snout and involve the blockage of a main subglacial conduit by coarse sediment (Davies *et al.*, 2003). Depending on the conduit gradient, water discharge, and sediment characteristics, this may lead to (i) overflow of water in ice-marginal moulins and then along ice-marginal channels; (ii) the bursting of water from the glacier portal as the blockage is removed;

(iii) widespread uplift of the glacier margin and the discharge of water along the width of the terminus. The fourth type of flood involves the blockage of a subglacial conduit by ice at the base of an icefall near the centerline of the glacier ~2 km from the terminus (Goodsell *et al.*, 2003). Water is forced to flow supraglacially and then ice-marginally to the glacier snout. The first three flood types may be more common when the glacier is in an advanced or advancing position since this would promote the blockage of subglacial conduits near the glacier snout (Davies *et al.*, 2003). The flood described by Goodsell *et al.* (2003, 2005) occurred while the glacier was retreating, and this flood type may be favored by glacier recession, which might promote ice collapse around the icefall.

Very few floods associated with drainage from englacial/subglacial stores have been documented. Those that have suggest hydrographs can be highly variable. The flood from Bas Glacier d'Arolla, Switzerland, was symmetrical and lasted 2 days (Warburton and Fenn, 1994). The flood from Bench Glacier, Alaska, involved a ramped increase in flow with a declining limb lasting several days with erratic short-term pulses of very high flow superimposed on top (Anderson *et al.*, 1999). The hydrographs associated with the fourth flood type at Franz Josef Glacier also had a very rapid rise to peak within a few hours, followed by a more gradual decline lasting several days (Goodsell *et al.*, 2005) (Figure 9d).

Polythermal Glaciers

Small floods associated with drainage from the englacial/subglacial drainage system appear to be a common feature of mainly cold polythermal glaciers that have a warm-based core and cold-based margins (Baranowski, 1973; Skidmore and Sharp, 1999; Hodson and Ferguson, 1999; Wadham *et al.*, 2001; Rippin *et al.*, 2003). Water is trapped beneath the warm-based core by the "thermal dam" at the boundary with the cold-based margin. Once pressures reach some critical level behind the thermal dam, water escapes from beneath the cold-based margin producing small floods associated with supraglacial geyser-like spouts (Baranowski, 1973; Skidmore and Sharp, 1999), proglacial upwellings (Skidmore and Sharp, 1999; Hodson and Ferguson, 1999; Rippin *et al.*, 2003), or enhanced flow in ice-marginal streams (Wadham *et al.*, 2001). The precise mechanisms whereby water escapes through the margin are not known, but possibilities include flow through fissures within the sediments beneath the permafrost, through hydrofractures in the basal ice, or by hydraulic jacking between the basal ice and frozen sediments (Rippin *et al.*, 2003).

On John Evans Glacier, burst-through events occur in late June/early July and are initiated when supraglacial streams connect to moulins located in a major crevasse field ~4 km from the glacier terminus (Boon *et al.*, 2003; Copland *et al.*, 2003). This allows water to drain from the surface to the bed, and the burst-through event usually starts within 24 h of the moulins opening. Initially, outflow consists of highly solute-rich waters that have likely been stored subglacially over winter, but solute concentrations drop rapidly within 3 to 4 days as these "old" waters are diluted by the new season's melt (Skidmore and Sharp, 1999). In the relatively warm summers of 1998–2000, subglacial outflow was continuous once initiated, similar to the situation that has been observed on other mainly cold polythermal glaciers (Hodson and Ferguson, 1999; Rippin *et al.*, 2003). However, during the cooler summers of 1994 and 1996, subglacial outflow occurred as a series of floods, with the subglacial drainage system closing down in intervening periods (Skidmore and Sharp, 1999; Boon *et al.*, 2003). Wadham *et al.* (2001) report similar findings from Finsterwalderbreen, Svalbard. Floods were observed in mid-August 1995 and mid-July 1999, once the snowline had retreated, allowing surface meltwaters to reach a subglacial reservoir beneath the upper glacier. The chemistry of the floodwaters suggested that they consist mainly of subglacially stored snowmelt. The floodwaters were routed to an existing ice-marginal stream rather than an existing proglacial upwelling, and there was no evidence for supraglacial fountains.

CONCLUSIONS

Summary

The presence of snow and ice, and the lack of vegetation and soil make the hydrology of glacierized basins very different to that of equivalent nonglacierized basins. These differences have important implications for the runoff regime, water storage and release, and incidence of flooding in glacierized catchments.

Many studies have shown how regional and local climates control individual glacier mass (water) balance. Studies are starting to identify how regional and local climate influence year-to-year variations in runoff from glacierized catchments. Interannual runoff variability depends on the sensitivity of catchments to different climatic variables, particularly air temperature and rainfall. This sensitivity partly depends on the percentage glacier cover. Interannual runoff variability depends on the seasonality and magnitude of climate variability. Finally, the nature of interannual runoff variability may change over the long-term in response to changes in snow, firn, and ice extent.

Regional and local climate affect season-to-season variations in runoff from glacierized catchments with big differences between catchments in mid-, high, and low latitudes. Glaciers in midlatitudes have pronounced accumulation and ablation seasons with low runoff in the winter and high runoff in the summer. Glaciers in high latitudes also have pronounced accumulation and ablation seasons, but

have shorter ablation seasons with high summer runoff in maritime areas and low summer runoff in continental regions. Low latitude glaciers usually have accumulation and ablation that are distributed more evenly throughout the year and show less marked runoff variability between the seasons.

There are also significant differences in the behavior of glaciers in mid-, high, and low latitudes in terms of their runoff regime during the spring and summer. Temperate glaciers in mid-latitudes show marked changes in the magnitude and shape of diurnal hydrographs over spring and summer in response to seasonal changes in meltwater production, snowpack cover and supraglacial, englacial, and subglacial routing. Cold and mainly cold polythermal glaciers in high latitudes show less marked diurnal hydrograph changes due to lower melt rates and fewer changes in glacial routing. Similarly, temperate glaciers in low latitudes also appear to undergo fewer changes in meltwater production due to the occurrence of both snowfall and melting throughout the year, maintaining a relatively stable englacial and subglacial drainage system.

Glaciers in mid- and high latitudes store and release water on a seasonal basis. Water is stored in snow, firn, and within the englacial and subglacial drainage system. Temperate glaciers that develop a hydraulically efficient subglacial drainage system store water in the spring and early summer and release it as the snow melts and the subglacial drainage system develops during the summer. Those with a less well-developed subglacial drainage system store and release water in response to weather-related inputs. Mainly cold polythermal glaciers store water in spring and early summer. This water may be released dramatically in the form of ice-marginal fountains or pressurized proglacial upwellings as the water is forced from the warm- based core through the cold-based terminus. Cold glaciers have little capacity for storing water, but store water in the snowpack in spring, release it relatively quickly in early summer as the snow melts, and may store water in ice-marginal channels during mid- to late summer.

Floods are prevalent in glacierized catchments. Most large floods result from the catastrophic drainage of ice-dammed subglacial or ice-marginal lakes, or moraine-dammed proglacial lakes. Smaller floods can occur in response to rapid surface melting or rainstorms, and may be associated with the release of water stored in the englacial / subglacial drainage system.

Effects of Future Climate Change

Global climate change has the potential to alter the hydrology of glacierized basins in major ways. The effect of global climate change on glacier mass balance is difficult to quantify. Most estimates involve sensitivity tests of local glacier mass balance models to uniform changes in air temperature and precipitation. But global climate change may have more subtle effects on regional and local climate through, for example, changes in the timing and intensity of weather systems. Glaciers may be more sensitive to these changes than to a uniform increase in annual air temperature or decrease in annual precipitation. It is also, important to recognize the different mass balance regimes discussed by Ageta and Higuchi (1984). As a rise in temperature will decrease the amount of precipitation falling as snow, midlatitude glaciers will be more sensitive to temperature increases in the winter months (Oerlemans, 1993), whereas low latitude glaciers may be more sensitive to changes in summer months (Ageta and Kadota, 1992). Similarly, low latitude glaciers may be more sensitive to changes in humidity than midlatitude glaciers, especially during dry winter months, since humidity controls the partitioning between sublimation and melting.

The effect of climate change on runoff from glacierized basins is even more difficult to quantify. Annual or melt-season runoff may either increase or decrease in the short-term (years) depending on the magnitude and timing of climate change, particularly air temperature and rainfall, and the effects this has on patterns of snow accumulation and snow and ice melt. In the longer term (decades), the effects of changes in glacier geometry will play an increasingly important role. There are very few modeling studies that have tried to quantify the effects of climate change on glacierized basin river flow. Singh and Kumar (1997) used the UBC watershed model to examine the effects of air temperature and precipitation changes on snowmelt runoff, glacier melt runoff, and total stream runoff in the 2.5% glacierized Spiti River catchment, western Himalayas. Annual snowmelt, glacier melt, and total runoff increased linearly with increases in temperature between 1 °C and 3 °C, but the effects were greatest on glacier melt runoff. For example, a temperature increase of 2 °C increased annual snowmelt, glacier melt, and total runoff by an average of 12%, 35%, and 9% respectively. Changes in precipitation produced linear changes in runoff, but the relationships were direct for snowmelt and total runoff, but inverse for glacier melt runoff. Braun *et al.* (2000) modeled the effects of declining glacier cover on the runoff from Alpine basins. Larger discharge peaks will occur when higher melt rates coincide with more intense rainstorms. Boon *et al.* (2003) discuss how global climate change may increase the incidence of synoptic conditions favorable to extreme melt events, which will increase the incidence of flooding in Arctic streams, particularly if they occur during late summer.

Climate change may also alter the patterns of annual water storage and release, if the thickness and extent of the snowpack and firn layers change. Furthermore, increased melting and rainstorms may encourage the development of hydraulically efficient englacial and subglacial drainage

pathways, which will also alter a glacier's water storage capacity. Finally, significant changes in water storage and release may occur in the Arctic if changes in climate and glacier geometry alter the distribution of cold and polythermal glaciers.

Floods associated with the drainage of ice-dammed or moraine-dammed lakes are likely to increase in number, frequency, and size at least in the short to medium term as glaciers retreat in response to climate change (Haeberli, 1983; Fujita *et al.*, 1997; Ageta *et al.*, 2000; Richardson and Reynolds, 2000a). However, many potentially hazardous lakes have not yet been identified as they are situated in remote locations. Thus, increased use is being made of remote sensing techniques to map the distribution of proglacial lakes so that their hazard potential can be identified. For example, Huggel *et al.* (2002) use a combination of satellite imagery to map lake areas, aerial photographs, and analytical photogrammetry to derive digital elevation models (DEMs), and empirical equations to derive lake volume and potential flood discharges, and run out distances. They applied their approach to modeling the 1993 flood associated with the drainage of Sirwolte Lake, Switzerland.

Acknowledgments

I thank Jean-Michel Bonvin, Grande Dixence SA, Switzerland; Lars-Evan Pettersson, NVE, Norway; Andrew Fountain, Portland State University, USA; and Pierre Ribstein, Université Paris, France for the data shown in Figure 3. This paper was written while I was on sabbatical leave at the Department of Geography, University of Canterbury, New Zealand, and I thank the staff there, especially Wendy Lawson and Ian Owens, for their hospitality. Extra special thanks go to Ragazza Glaciale. The paper benefited from my useful discussions with Luke Copland and Becky Goodsell, the constructive reviews of two anonymous referees, and editorial help from Martin Sharp. Phil Stickler and Andy Fox helped produce the Figures.

REFERENCES

Ageta Y. and Higuchi K. (1984) Estimation of mass balance components of a summer-accumulation type glacier in the Nepal Himalaya. *Geografiska Annaler A*, **66**, 249–255.

Ageta Y., Iwata S., Yabuki H., Naito N., Sakai A., Narama C. and Karma C. (2000) Expansion of glacier lakes in recent decades in the Bhutan Himalayas. In *Debris-Covered Glaciers*, Nakawo M., Raymond C.F. and Fountain A. (Eds.), IAHS Publication, 264, IAHS: pp. 165–175.

Ageta Y. and Kadota T. (1992) Predictions of changes of glacier mass balance in the Nepal Himalayas and Tibetan Plateau: a case study of air temperature increase for three glaciers. *Annals of Glaciology*, **16**, 89–94.

Aizen V.B., Aizen E.M. and Melack J.M. (1995) Characteristics of runoff formation at Kirgizskiy Alatoo, Tien Shan. In *Biogeochemistry of Seasonally Snow-Covered Catchments*, Tonnessen K.A., Williams M.W. and Tranter M. (Eds.), IAHS Publication 228, IAHS: pp. 3–30.

Ames A. (1998) A documentation of glacier tongue variations and lake development in the Cordillera Blanca, Peru. *Zeitschrift fur Gletscherkunde and Glazialgeologie*, **34**, 1–36.

Ames A. and Hastenrath S. (1996) Mass balance and iceflow of the Uruashraju Glacier, Cordillera Blanca, Peru. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **32**, 83–89.

Anderson S.P., Fernald K.M.H., Anderson R.S. and Humphrey N.F. (1999) Physical and chemical characterization of a spring flood event, Bench Glacier, Alaska, U.S.A.: evidence for water storage. *Journal of Glaciology*, **45**, 177–189.

Arendt A. (1999) Approaches to modelling the surface albedo of a high Arctic glacier. *Geografiska Annaler A*, **81**, 477–487.

Arnaud Y., Muller F., Vuille M. and Ribstein P. (2001) El Nino-Southern Oscillation (ENSO) influence on a Sajama volcano glacier (Bolivia) from 1963 to 1998 as seen from Landsat data and aerial photography. *Journal of Geophysical Research D*, **106**, 17773–17784.

Arnold N., Richards K., Willis I. and Sharp M. (1998) Initial results from a semi-distributed, physically-based model of glacier hydrology. *Hydrological Processes*, **12**, 191–219.

Arnold N.S., Willis I.C., Sharp M.J., Richards K.S. and Lawson W.J. (1996) A distributed surface energy-balance model for a small valley glacier. I. Development and testing for Haut Glacier d'Arolla, Valais, Switzerland. *Journal of Glaciology*, **42**, 77–89.

Baranowski S. (1973) Geyser-like water spouts at Werenskioldbreen, Spitsbergen. In *Hydrology of Glaciers*, Glen J.W., Adie R.J. and Johnson D.M. (Eds.), IAHS Publication 95, IAHS.

Barrett A.P. and Collins D.N. (1997) Interaction between water pressure in the basal drainage system and discharge from an Alpine glacier before and during a rainfall-induced subglacial hydrological event. *Annals of Glaciology*, **24**, 288–292.

Beecroft I.R. (1983) Sediment transport during an outburst from Glacier de Tsidiore Nouve, Switzerland, 16–19 June 1981. *Journal of Glaciology*, **29**, 185–190.

Benn D.I. and Evans D.J.A. (1998) *Glaciers and Glaciation*, Arnold: London.

Bhatt M.P., Masuzawa T., Yamamoto M., Sakai A. and Fujita K. (2000) Seasonal changes in dissolved chemical composition and flux of meltwater draining from Lirung Glacier in the Nepal Himalayas. In *Debris-Covered Glaciers*, Nakawo M., Raymond C.F. and Fountain A. (Eds.), IAHS Publication 264, IAHS: pp. 277–288.

Bitz C.M. and Battisti D.S. (1999) Interannual to decadal variability in climate and glacier mass balance in Washington, western Canada and Alaska. *Journal of Climate*, **12**, 3181–3196.

Björnsson H. (1974) Explanation of jökulhlaups from Grímsvötn, Vatnajökull, Iceland. *Jökull*, **24**, 1–26.

Björnsson H. (1975) Subglacial water reservoirs, jökulhlaups and volcanic eruptions. *Jökull*, **25**, 1–14.

Björnsson H. (1976) Marginal and supraglacial lakes in Iceland. *Jökull*, **26**, 40–50.

Björnsson H. (1992) Jökulhlaups in Iceland: prediction, characteristics and simulation. *Annals of Glaciology*, **16**, 95–106.

Björnsson H. (1997) Grímsvatnahlaup Fyrr og Nú. In *Vatnajökull: Gos Og Hlaup*, Haraldsson H. (Ed.), Reykjavík: Vegager Din, pp. 61–77.

Björnsson H. (1998) Hydrological characteristics of the drainage system beneath a surging glacier. *Nature*, **395**, 771–774.

Björnsson H. (2003) Sub glacial lakes and jokulhlaups in Iceland. *Global and Planetary Change*, **35**, 255–271.

Björnsson H., Pálsson F. and Guðmundsson M.T. (2000) Surface and bedrock topography of the Myrdalsjökull ice cap, Iceland: the Katla caldera eruption sites and routes of jökulhlaups. *Jökull*, **49**, 29–46.

Bøggild C.E. (2000) Preferential flow and melt water retention in cold snow packs in West-Greenland. *Nordic Hydrology*, **31**, 287–300.

Boon S., Sharp M. and Nienow P. (2003) Impact of an extreme melt event on the runoff and hydrology of a high Arctic glacier. *Hydrological Processes*, **17**, 1051–1072.

Braithwaite R.J. (2002) Glacier mass balance: the first 50 years of international monitoring. *Progress in Physical Geography*, **26**, 76–95.

Braithwaite R.J. and Olesen O.B. (1988) Effect of glaciers on annual runoff, Johan Dahl Land, South Greenland. *Journal of Glaciology*, **34**, 200–207.

Braithwaite R.J. and Zhang Y. (2000) Sensitivity of mass balance of five Swiss glaciers to temperature changes assessed by tuning a degree-day model. *Journal of Glaciology*, **46**, 7–14.

Braun L.N. and Escher-Vetter H. (1996) *Glacial Discharge as Affected by Climate Change*, Interpraevent 1996 – Garmisch Partenkirchen, pp. 65–74.

Braun L.N., Weber M. and Schulz M. (2000) Consequences of climate change for runoff from Alpine regions. *Annals of Glaciology*, **31**, 19–25.

Brock B.W., Willis I.C. and Sharp M.J. (2000) Measurement and parameterization of albedo variations at Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **46**, 675–688.

Buchroithner M.F., Jentsch G. and Waniverhaus B. (1982) Monitoring of recent geological events in the Khumbu area (Himalaya, Nepal) by digital processing of Landsat MSS data. *Rock Mechanics*, **15**, 181–197.

Carling P.A. and Glaister M.S. (1987) Reconstruction of a flood resulting from a moraine-dam failure using geomorphological evidence and dam-break modeling. *The Binghamton Symposia in Geomophology International Series*, Allen and Unwin: London and Boston, Vol. 18, pp. 181–200.

Cenderelli D.A. (2000) Floods from natural and artificial dam failures. In *Inland Flood Hazards: Human, Riparian, and Aquatic Communities*, Wohl E.E. (Ed.), Cambridge University Press: New York, pp. 73–103.

Cenderelli D.A. and Wohl E.E. (2001) Peak discharge estimates of glacial-lake outburst floods and "normal" climatic floods in the Mount Everest region, Nepal. *Geomorphology*, **40**, 57–90.

Cenderelli D.A. and Wohl E.E. (2003) Flow hydraulics and geomorphic effects of glacial-lake outburst floods in the Mount Everest region, Nepal. *Earth Surface Processes and Landforms*, **28**, 385–407.

Chen J. and Ohmura A. (1990) Estimation of Alpine glacier water resources and their change since the 1870 s. In *Hydrology of Mountainous Regions*, Lang H. and Musy A. (Eds.), IAHS Publication 193, IAHS: pp. 127–136.

Clague J.J. and Evans S.G. (1994) Formation and failure of natural dams in the Canadian Cordillera. *Bulletin of the Geological Survey of Canada*, **464**, 35.

Clague J.J. and Evans S.G. (2000) A review of catastrophic drainage of moraine-dammed lakes in British Columbia. *Quaternary Science Reviews*, **19**, 1763–1783.

Clague J.J., Evans S.G. and Blown I.G. (1985) A debris flow triggered by the breaching of a moraine-dammed lake, Klattasine Creek, British Columbia. *Canadian Journal of Earth Sciences*, **22**, 1492–502.

Clarke G.K.C. (1982) Glacier outburst floods from 'Hazard Lake', Yukon Territory, and the problem of flood magnitude prediction. *Journal of Glaciology*, **28**, 3–21.

Clarke G.K.C. and Waldron D.A. (1984) Simulation of the August 1979 sudden discharge of glacier-dammed Flood Lake, British Columbia. *Canadian Journal of Earth Sciences*, **21**, 502–504.

Clement P. (1984) The drainage of a marginal ice-dammed lake at Nordbogletscher, Johan Dahl Land, south Greenland. *Arctic and Alpine Research*, **16**, 209–16.

Cogley J.G. and Adams W.P. (1998) Mass balance of glaciers other than the ice sheets. *Journal of Glaciology*, **44**, 315–325.

Colbeck S.C. (1978) The physical aspects of water flow through snow. In *Advances in Hydroscience*, Chow V.T. (Ed.), Academic Press: New York.

Colbeck S.C. and Anderson E.A. (1982) The permeability of a melting snow cover. *Water Resources Research*, **18**, 904–908.

Collins D.N. (1982) Water storage in an Alpine glacier. In *Hydrological Aspects of Alpine and High-Mountain Areas*, Glen J.W. (Ed.), IAHS Publication **138**, IAHS: pp. 113–122.

Collins D.N. (1989) Seasonal development of subglacial drainage and suspended sediment delivery to meltwaters beneath an Alpine glacier. *Annals of Glaciology*, **13**, 45–50.

Collins D.N. (1995) Rainfall-induced high magnitude runoff events in late summer in highly glacierised Alpine basins. *BHS 5th National Hydrology Symposium, Edinburgh, 1995*, British Hydrological Society: pp. 3.55–3.59.

Collins D.N. (1998) Outburst and rainfall-induced peak runoff events in highly glacierized Alpine basins. *Hydrological Processes*, **12**, 2369–2381.

Collins D.N. (1999) Solute flux in meltwaters draining from a glacierized basin in the Karakoram mountains. *Hydrological Processes*, **13**, 3001–3015.

Copland L., Sharp M. and Nienow P. (2003) Links between short-term velocity variations and the subglacial hydrology of a predominantly cold polythermal glacier. *Journal of Glaciology*, **49**, 337–348.

Costa J.E. (1988) Rheologic, geomorphic and sedimentologic differentiation of water floods, hyperconcentrated flows and debris flows. In *Flood Geomorphology*, Baker V.R., Kochel K.C. and Patton P.C. (Eds.), Wiley: Chichester.

Costa J.E. and Schuster R.L. (1988) The formation and failure of natural dams. *Geological Society of America Bulletin*, **100**, 1054–68.

Davies T.R.H., Smart C.C. and Turnbull J.M. (2003) Water and sediment outbursts from advanced Franz Josef Glacier, New Zealand. *Earth Surface Processes and Landforms*, **28**, 1081–1096.

Dawson A.G. (1983) Glacier-dammed lake investigations in the Hullet Lake area, south Greenland. *Meddelelser om Grønland Geoscience*, **11**, 3–22.

de Ruyter de Wildt M.S., Klok E.J. and Oerlemans J. (2003) Reconstruction of the mean specific mass balance of Vatnajokull (Iceland) with a Seasonal Sensitivity Characteristic. *Geografiska Annaler A*, **85**, 57–72.

Del Valle R.A., Skvarca P., Mancini M.V. and Lusky J. (1995) A preliminary study of sediment cores from Lago Argentino and fluctuations of Moreno Glacier, Patagonia. *Bulletin of Glacier Research*, **13**, 121–126.

Denner J.C., Lawson D.E., Larson G.J., Evenson E.B., Alley R.B., Strasser J.C. and Kopczynski S. (1999) Seasonal variability in hydrologic-system response to intense rain events, Matanuska Glacier, Alaska, U.S.A. *Annals of Glaciology*, **28**, 267–271.

Depetris P.J. and Pasquini A.I. (2000) The hydrological signal of the Perito Moreno Glacier damming of Lake Argentino southern Andean Patagonia: the connection to climate anomalies. *Global and Planetary Change*, **26**, 367–374.

Desloges J.R., Jones D.P. and Ricker K.E. (1989) Estimates of peak discharge from the drainage of ice-dammed Ape Lake, British Columbia, Canada. *Journal of Glaciology*, **35**, 349–54.

Ding Y. and Liu J. (1992) Glacial lake outburst flood disasters in China. *Annals of Glaciology* **16**, 180–184.

Driedger C.L. and Fountain A.G. (1989) Glacier outburst floods at Mount Rainier, Washington State, USA. *Annals of Glaciology*, **13**, 51–55.

Duynkerke P.G. and van den Broeke R. (1994) Surface energy-balance and katabatic flow over glacier and tundra during GIMEX-91. *Global and Planetary Change*, **9**, 17–28.

Dyurgerov M. and Meier M.F. (1997) Mass balance of mountain and sub-polar glaciers: a new global assessment for 1961–1990. *Arctic and Alpine Research*, **29**, 379–391.

Elliston G.R. (1973) Water movement through the Gornergletscher. In *Hydrology of Glaciers*, Glen J.W., Adie R.J. and Johnson D.M. (Eds.), IAHS Publication 95, IAHS: pp. 79–84.

Escher-Vetter H. and Reinwarth O. (1994) Two decades of runoff measurements (1974 to 1993) at the Pegelstation Vernagtbach/Oetztal Alps. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **30**, 53–98.

Fisher D. (1973) Subglacial leakage of Summit Lake, British Columbia, by dye determinations. In *Hydrology of Glaciers*, Glen J.W., Adie R.J. and Johnson D.M. (Eds.), IAHS Publication 95, IAHS: pp. 111–16.

Fountain A.G. (1989) The storage of water in, and hydraulic characteristics of, the firn of South Cascade Glacier, Washington State, USA. *Annals of Glaciology*, **13**, 69–75.

Fountain A.G. (1992) Subglacial water flow inferred from stream measurements at South cascade Glacier, Washington, USA. *Journal of Glaciology*, **38**, 51–64.

Fountain A.G. (1996) Effect of snow and firn hydrology on the physical and chemical characteristics of glacial runoff. *Hydrological Processes*, **10**, 509–521.

Fountain A.G. and Tangborn W.V. (1985) The effect of glaciers on streamflow variations. *Water Resources Research*, **21**, 579–586.

Fountain A.G. and Walder J.S. (1998) Water flow in glaciers. *Reviews of Geophysics*, **36**, 299–328.

Fowler A.C. (1999) Breaking the seal at Grímsvötn. *Journal of Glaciology*, **45**, 506–516.

Fowler A.C. and Ng F.S.L. (1996) The role of sediment transport in the mechanics of jokulhlaups. *Annals of Glaciology*, **22**, 255–259.

Fox A.M. (2003) *A Distributed, Physically Based Snow Melt and Runoff Model for Alpine Glaciers*, Unpublished PhD Thesis, University of Cambridge, October 2003.

Francou B., Ramirez E., Caceres B. and Mendoza J. (2000) Glacier evolution in the tropical Andes during the last decades of the 20th century: Chacaltaya, Bolivia, and Antizana, Ecuador. *Ambio*, **29**, 416–422.

Francou B., Ribstein P., Saravia R. and Tiriau E. (1995) Monthly balance and water discharge of an inter-tropical glacier: Zongo Glacier, Cordillera Real, Bolivia, 16 degrees S. *Journal of Glaciology*, **41**, 61–67.

Fujita K., Nakawo M., Fujii Y. and Paudyal P. (1997) Changes in glaciers in Hidden Valley, Mukut Himal, Nepal Himalayas, from 1974 to 1994. *Journal of Glaciology*, **43**, 583–588.

Fukushima Y., Kawashima K., Suzuki M., Ohta T., Kubota H., Yamada T. and Bajracharya O.R. (1987) Runoff characteristics in three glacier-covered watersheds of Langtang Valley, Nepal Himalayas. *Bulletin of Glacier Research*, **5**, 11–18.

Fushimi H., Ikegami K., Higuchi K. and Shankar K. (1985) Nepal case study: catastrophic floods. In *Techniques for Prediction of Runoff from Glacierized Areas*, Young G.J. (Ed.), IAHS Publication 149, IAHS: pp. 125–30.

Glen J.W. (1954) The stability of ice-dammed lakes and other water-filled holes in glaciers. *Journal of Glaciology*, **2**, 316–318.

Goodsell B., Anderson B. and Lawson W. (2003) Correspondence: supraglacial routing of subglacial water at Franz Josef Glacier, South Westland, New Zealand. *Journal of Glaciology*, **49**, 469–470.

Goodsell B., Anderson B., Lawson W. and Owens I.F. (2005) Outburst flooding and associated events at Franz Josef Glacier, South Westland, New Zealand. *New Zealand Journal of Geology and Geophysics*, **48**, 1–10.

Gordon S., Sharp M., Hubbard B., Smart C., Ketterling B. and Willis I. (1998) Seasonal reorganization of subglacial drainage inferred from measurements in boreholes. *Hydrological Processes*, **12**, 105–133.

Grabs W.E. and Hanisch J. (1993) Objectives and methods for glacier lake outburst floods (GLOF's). In *Snow and Glacier Hydrology*, Young G.J. (Ed.), IAHS Publication 218, IAHS: pp. 341–352.

Guðmundsson M.T., Björnsson H. and Palsson F. (1995) Changes in jökulhlaup sizes in Grimsvötn, Vatnajökull, Iceland, 1934-91, deduced from in-situ measurements of subglacial lake volume. *Journal of Glaciology*, **41**, 263–272.

Guðmundsson M.T., Sigmunðsson F. and Björnsson H. (1997) Ice- volcano interaction of the Gjálp subglacial eruption, Vatnajökull, Iceland. *Nature*, **389**, 954–957.

Gurnell A.M., Clark M.J. and Hill C.T. (1992) Analysis and interpretation of patterns within and between hydroclimatological time series in an Alpine glacier basin. *Earth Surface Processes and Landforms*, **17**, 821–839.

Gurnell A.M., Hodson A.J., Clark M.J., Bogen J., Hagen J.O. and Tranter M. (1994) *Water and Sediment Discharge from Glacier Basins: An Arctic and Alpine Comparison*, IAHS Publication 224, IAHS: pp. 325–334.

Haeberli W. (1983) Frequency and characteristics of glacier floods in the Swiss Alps. *Annals of Glaciology*, **4**, 85–90.

Haeberli W., Frauenfelder R., Hoelzle M. and Maisch M. (2000) On rates and acceleration trends of global glacier mass changes. *Geografiska Annaler A*, **81**, 585–591.

Haeberli W., Kaab A., Vonder Muhll D. and Teysseire P. (2001) Prevention of outburst floods from periglacial lakes at Gruben Glacier, Valais, Swiss Alps. *Journal of Glaciology*, **47**, 111–122.

Hagen J.O. (1987) Glacier surge at Usherbreen, Svalbard. *Polar Research*, **5**, 239–252.

Hagen J.O., Korsen O.M. and Vatne G. (1991) Drainage pattern in a sub-polar glacier, Brøggerbreen, Svalbard. In *Arctic Hydrology. Present and Future Tasks*, Gjessing U., Hagen J.O., Hassel K.A., Sand K. and Wold B. (Eds.), Norwegian National Committee for Hydrology Report No. 23, NNCH: Oslo. pp. 121–131.

Hagen J.O., Liestol O., Roland E. and Jorgensen T. (1993) *Glacier Atlas of Svalbard and Jan Mayen*, Norsk Polarinstitutt: Oslo.

Hannah D.M., Gurnell A.M. and McGregor G.R. (1999) A methodology for investigation of the seasonal evolution in proglacial hydrograph form. *Hydrological Processes*, **13**, 2603–2621.

Hannah D.M., Smith B.P.G., Gurnell A.M. and McGregor G.R. (2000) An approach to hydrograph classification. *Hydrological Processes*, **14**, 317–338.

Hasnain S.I., Jose P.G., Ahmad S. and Negi D.C. (2001) Character of the subglacial drainage system in the ablation area of Dokriani Glacier, India, as revealed by dye-tracer studies. *Journal of Hydrology*, **248**, 216–223.

Hasnain S.I. and Thayyen R.J. (1999) Discharge and suspended-sediment concentration of meltwaters draining from the Dokriani Glacier, Garhwal Himalaya, India. *Journal of Hydrology*, **218**, 191–198.

Hastenrath S. (1981) *The Glaciation of the Ecuadorian Andes*, Balkema: Rotterdam.

Hastenrath S. (1984) *The Glaciers of Equatorial East Africa*, D. Reidel Publishing: Dordrecht.

Hastenrath S. and Ames A. (1995) Recession of Yanamarey Glacier in Cordillera Blanca, Peru, during the 20th century. *Journal of Glaciology*, **41**, 191–196.

Hastenrath S. and Kruss P. (1992) The dramatic retreat of Mount Kenya's glaciers 1963–1987. *Annals of Glaciology*, **16**, 127–133.

Hewitt K. (1982) Natural dams and outburst floods of the Karakoram Himalaya. In *Hydrological Aspects of Alpine and High Mountain Areas*, Glen J.W. (Ed.), IAHS Publication, 138, IAHS: pp. 259–269.

Higgins A.K. (1970) On some ice-dammed lakes in the FrederikshÅb district, south-west Greenland. *Meddelelser Fra Dansk Geologisk Forening*, **19**, 378–97.

Higuchi K., Ageta Y., Yasunari T. and Inoue J. (1982) Characteristics of precipitation during the monsoon season in high-mountain areas of the Nepal Himalaya. In *Hydrological Aspects of Alpine and High-Mountain Areas*, Glen J.W. (Ed.), IAHS Publication 138, IAHS: pp. 21–30.

Hock R. and Noetzli C. (1997) Areal melt and discharge modelling of Storglaciären, Sweden. *Annals of Glaciology*, **24**, 211–216.

Hodgkins R. (1997) Glacier hydrology in Svalbard, Norwegian high Arctic. *Quaternary Science Reviews*, **16**, 957–973.

Hodgkins R. (2001) Seasonal evolution of meltwater generation, storage and discharge at a non-temperate glacier in Svalbard. *Hydrological Processes*, **15**, 441–460.

Hodson A.J. and Ferguson R.I. (1999) Fluvial suspended sediment transport from cold and warm-based glaciers in Svalbard. *Earth Surface Processes and Landforms*, **24**, 957–974.

Hodson A.J., Gurnell A.M., Washington R., Tranter M., Clark M.J. and Hagen J.O. (1998) Meteorological and runoff time-series characteristics in a small, high-Arctic glaciated basin, Svalbard. *Hydrological Processes*, **12**, 509–526.

Hofinger S. and Kuhn M. (1996) Reconstruction of the summer mass balance of Hintereisferner since 1953. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **32**, 137–149.

Hooke R.L.eB., Calla P., Holmlund P., Nilsson M. and Stroeven A. (1989) A 3 year record of seasonal variations in surface velocity, Storglaciären, Sweden. *Journal of Glaciology*, **35**, 235–247.

Hooker B.L. and Fitzharris B.B. (1999) The correlation between climatic parameters and the retreat and advance of Franz Josef Glacier, New Zealand. *Global and Planetary Change*, **22**, 39–48.

Hope G.S., Peterson J.A., Radok U. and Allison I. (Eds.) (1976) *The Equatorial Glaciers of New Guinea*, Balkema: Rotterdam.

Hopkinson C. and Young G.J. (1998) The effect of glacier wastage on the flow of the Bow River at Banff, Alberta, 1951–1993. *Hydrological Processes*, **12**, 1745–1762.

Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J. and Xiaosu D. (Eds.) (2001) *Climate Change 2001: The Scientific Basis*, Intergovernmental Panel on Climate Change (IPCC) Cambridge University Press: p. 944.

Howarth P.J. (1968) A supra-glacial extension of an ice-dammed lake, Tunsbergdalsbreen, Norway. *Journal of Glaciology*, **7**, 414–19.

Huggel C., Kaab A., Haeberli W., Teysseire P. and Paul F. (2002) Remote sensing based assessment of hazards from glacier lake outbursts: a case study in the Swiss Alps. *Canadian Geotechnical Journal*, **39**, 316–330.

Iken A. (1981) The effect of the subglacial water pressure on the sliding velocity of a glacier in an idealized numerical model. *Journal of Glaciology*, **27**, 407–421.

Iken A. and Bindschadler R.A. (1986) Combined measurements of subglacial water pressure and surface ice velocity of Findelengletscher, Switzerland: conclusions about drainage system and sliding mechanism. *Journal of Glaciology*, **32**, 101–117.

Iken A., Röthlisberger H., Flotron A. and Haeberli W. (1983) The uplift of Unteraargletscher at the beginning of the melt season - a consequence of water storage at the bed? *Journal of Glaciology*, **29**, 28–47.

Irvine-Fynn T.D.L., Moorman B.J., Willis I.C., Sjogren D.B., Hodson A.J., Mumford P.N., Walter F.S.A. and Williams J.L.M.

(2005) Geocryological processes linked to High-Arctic proglacial stream suspended sediment dynamics: examples from Bylot Island, Nunavut and Spitsbergen, Svalbard. *Hydrological Processes*, **19**, 115–135.

Ives J.D. (1986) *Glacial Lake Outburst Floods and Risk Engineering in the Himalaya*, Occasional Paper no. 5, ICIMOD: Kathmandu, p. 42.

Jansson P., Hock R. and Schneider T. (2003) The concept of glacier storage: a review. *Journal of Hydrology*, **282**, 116–129.

Jansson P. and Hooke R.L.eB. (1989) Short-term variations in strain and surface tilt on Storglaciären, Kebnekaise, Northern Sweden. *Journal of Glaciology*, **35**, 201–208.

Jingshi L. and Fukushima Y. (1999) Recent change and prediction of glacier-dammed lake outburst floods from Kunmalik River in southern Tien Shan, China. In *Hydrological Extremes: Understanding, Predicting, Mitigating*, Gottschalk L., Olivry J.C., Reed D. and Rosbjerg D. (Eds.), IAHS Publication 255, IAHS: pp. 99–107.

Kamb B. (1987) Glacier surge mechanism based on linked cavity configuration of the basal conduit system. *Journal of Geophysical Research*, **92**, 9083–9100.

Kamb B. and Engelhardt H.F. (1987) Waves of accelerated motion in a glacier approaching surge: the mini-surges of Variegated Glacier, Alaska, U.S.A. *Journal of Glaciology*, **33**, 27–46.

Kamb B., Raymond C.F., Harrison W.D., Engelhardt H.F., Echelmeyer K.A., Humphrey N., Brugman M.M. and Pfeffer T. (1985) Glacier surge mechanism: 1982–1983 surge of Variegated Glacier, Alaska. *Science*, **227**, 469–479.

Kaser G. (1999) A review of the modern fluctuations of tropical glaciers. *Global and Planetary Change*, **22**, 93–103.

Kaser G. (2001) Glacier-climate interaction at low latitudes. *Journal of Glaciology*, **47**, 195–204.

Kaser G. and Georges C. (1999) On the mass balance of low latitude glaciers with particular consideration of the Peruvian Cordillera Blanca. *Geografiska Annaler A*, **81**, 643–652.

Kaser G. and Osmaston H. (2001) *Tropical Glaciers*, International Hydrology Series, UNESCO and Cambridge University Press.

Kaser G., Juen I., Georges C., Gomez J. and Tamayo W. (2003) The impact of glaciers on the runoff and the reconstruction of mass balance history from hydrological data in the tropical Cordillera Blanca, Perú. *Journal of Hydrology*, **282**, 130–144.

Kasser P. (1973) Influence of changes in the glacierized area on summer runoff in the Porte du Scex Drainage Basin of the Rhone. In *Hydrology of Glaciers*, Glen J.W., Adie R.J. and Johnson D.M. (Eds.), IAHS Publication 95, IAHS: pp. 221–225.

Kavanaugh J.L. and Clarke G.K.C. (2001) Abrupt glacier motion and reorganization of basal shear stress following the establishment of a connected drainage system. *Journal of Glaciology*, **47**, 472–480.

Knight P.G. and Russell A.J. (1993) Most recent observations of an ice- dammed lake at Russell Glacier, west Greenland, and a new hypothesis regarding the mechanisms of drainage initiation. *Journal of Glaciology*, **39**, 701–703.

Knudsen N.T. and Theakstone W.H. (1988) Drainage of the Austre Okstindbreen ice-dammed lake, Okstindan, Norway. *Journal of Glaciology*, **34**, 87–94.

Konovalov V.G. and Shchetinnicov A.S. (1994) Evolution of glaciation in the Pamiro-Alai mountains and its effect on river run-off. *Journal of Glaciology*, **134**, 149–157.

Kovanen D.J. (2003) Decadal variability in climate and glacier fluctuations on Mt Baker, Washington, USA. *Geografiska Annaler A*, **85**, 43–55.

Lafreniere M. and Sharp M. (2003) Wavelet analysis of inter-annual variability in the runoff regimes of glacial and nival stream catchments, Bow Lake, Alberta. *Hydrological Processes*, **17**, 1093–1118.

Laumann T. and Reeh N. (1993) Sensitivity to climate change of the mass balance of glaciers in southern Norway. *Journal of Glaciology*, **39**, 656–665.

Lawler D.M., McGregor G.R. and Phillips I.D. (2003) Influence of atmospheric circulation changes and regional climate variability on river flow and suspended sediment fluxes in southern Iceland. *Hydrological Processes*, **17**, 3195–3223.

Leather D.J., Yarnal B. and Palecki M.A. (1991) The Pacific/NorthAmerican teleconnection pattern and United States climate. PartI. Regional temperature and precipitation associations. *Journal of Climate*, **4**, 517–528.

Lewis K.J., Fountain A.G. and Dana G.L. (1999) How important is terminus cliff melt?: a study of the Canada Glacier terminus, Taylor Valley, Antarctica. *Global and Planetary Change*, **22**, 105–115.

Liestøl O. (1956) Glacier dammed lakes in Norway. *Norsk Geografisk Tidsskrift*, **15**, 122–49.

Liestøl O. (1977) *Setevatnet, a Glacier Dammed Lake in Spitsbergen*, Norsk Polarinstitutt Årbok 1975, pp. 31–35.

Liestøl O., Repp K. and Wold B. (1980) Supra-glacial lakes in Spitzbergen. *Norsk Geografisk Tidsskrift*, **34**, 89–92.

Liu J. (1992) Jökulhlaups in the Kunmalike River, southern Tien Shan mountains, China. *Annals of Glaciology*, **16**, 85–88.

Liu J., Fukushima Y. and Hiyama T. (1999) Hydrological response of meltwater from glacier covered mountain basins to climate change in northwest China. In *Interactions between the Cryosphere, Climate and Greenhouse Gases*, Tranter M., Armstrong R., Brun E., Jones G., Sharp M. and Williams M. (Eds.), IAHS Publication 256, IAHS: pp. 193–207.

Lliboutry L., Morales B., Pautre A. and Schneider B. (1977) Glaciological problems set by the control of dangerous lakes in Cordillera Blanca, Peru. I: Historical failure of morainic dams, their causes and prevention. *Journal of Glaciology*, **18**, 239–254.

Maag H. (1969) Ice-dammed lakes and marginal glacial drainage on Axel Heiberg Island. *Axel Heiberg Island Research Report*, McGill University: Montreal.

Mair D., Nienow P., Willis I. and Sharp M. (2001) Spatial patterns of glacier dynamics during an early melt-season high velocity event: Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **47**, 9–20.

Mair D., Sharp M. and Willis I. (2002) Evidence for basal cavity opening from analysis of surface uplift during an early melt-season high velocity event: Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **48**, 208–216.

Mair D., Willis I., Hubbard B., Fischer U., Nienow P. and Hubbard A. (2003) Hydrological controls on patterns of surface, internal and basal velocities during three "spring events":

Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **49**, 555–567.

Marsh P. and Woo M. (1984) Wetting front advance and freezing of meltwater within a snowcover 1. Observations in the Canadian Arctic. *Water Resources Research*, **20**, 1853–1864.

Mathews W.H. (1973) Record of two jökulhlaups. In *The Hydrology of Glaciers*, Glen J.W., Adie R.J. and Johnson D.M. (Eds.), IAHS Publication 95, IAHS: pp. 99–110.

Mathews W.H. and Clague J.J. (1993) The record of jökulhlaups from Summit Lake, northwestern British Columbia. *Canadian Journal of Earth Sciences*, **30**, 499–505.

Mattson L.E., Gardner J.S. and Young G.J. (1993) Ablation on debris covered glaciers: an example from the Rakhiot Glacier, Punjab, Himalayer. In *Snow and Glacier Hydrology*, Young G.J. (Ed.), IAHS Publication 218, IAHS: pp. 289–296.

Mayo L.R. (1989) Advance of Hubbard Glacier and 1986 outburst of Russell Fiord. Alaska, U.S.A. *Annals of Glaciology*, **13**, 189–194.

Meier M.F. (1984) Contributions of small glaciers to global sea level. *Science*, **226**, 1418–21.

Meier M.F. and Tangborn W.V. (1961) *Distinctive characteristics of glacier runoff*, U.S. Geological Survey Professional Paper 424-B, pp. B14–B16.

Miller M.M. (1952) Preliminary notes concerning certain structures and glacial lakes on the Juneau ice field. *American Geographical Society Juneau Ice Field Research Report*, **6**, 49–86.

Mool P.K. (1995) Glacier lake outburst floods in Nepal. *Journal of Nepal Geological Society*, **11**, 273–280.

Moore R.D. (1992) The influence of glacial cover on the variability of annual runoff, Coast Mountains, British Columbia, Canada. Canada. *Water Resources Journal*, **17**, 101–109.

Moore R.D. and Demuth M.N. (2001) Mass balance and streamflow variability at Place Glacier, Canada in relation to recent climate fluctuations. *Hydrological Processes*, **15**, 3473–3486.

Mottershead D.N. (1975) Observation of a temporary ice-dammed lake, Brimkjelen, southern Norway. *Norsk Geografisk Tidsskrift*, **29**, 69–74.

Munro D.S. (1989) Surface roughness and bulk heat transfer on a glacier: comparison with eddy correlation. *Journal of Glaciology*, **35**, 343–348.

Murray T. and Porter P.R. (2001) Basal conditions beneath a soft-bedded polythermal surge-type glacier: Bakaninbreen, Svalbard. *Quaternary International*, **86**, 103–116.

Nesje A., Lie Ø. and Dahl S.O. (2000) Is the North Atlantic Oscillation reflected in Scandinavian glacier mass balance records? *Journal of Quaternary Science*, **15**, 587–601.

Nienow P.W., Sharp M.J. and Willis I.C. (1998) Seasonal changes in the morphology of the subglacial drainage system, Haut Glacier d'Arolla, Switzerland. *Earth Surface Processes and Landforms*, **23**, 825–843.

Nye J.F. (1976) Water flow in glaciers - jökulhlaups, tunnels and veins. *Journal of Glaciology*, **17**, 181–88.

O'Connor J.E. and Costa J.E. (1993) Geologic and hydrologic hazards in glacierized basins in North America resulting from 19th and 20th century warming. *Natural Hazards*, **8**, 121–140.

Oerlemans J. (1993) A model for the surface balance of ice masses: Part 1. Alpine Glaciers. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **27/28**, 63–83.

Oerlemans J. and Fortuin J.P.F. (1992) Sensitivity of glaciers and small ice caps to greenhouse warming. *Science*, **258**, 115–117.

Östling M. and Hooke R.L.eB. (1986) Water storage in Storglaciären, Kebnekaise, Sweden. *Geografiska Annaler A*, **68**, 279–290.

Østrem G. (1973) Runoff forecasts for highly glacierized basins. *The Role of Snow and Ice in Hydrology*, IAHS Publication 107, IAHS: pp. 1111–1129.

Pelto M.S. (1996) Changes in glacier and alpine runoff in the North Cascade Range, Washington, USA 1985–1993. *Hydrological Processes*, **10**, 1173–1180.

Pertziger F.I. (1990) Role of glacier and snow cover melting in runoff variations from the small basins in Pamir and the Alps. In *Hydrology in Mountainous Regions*, Lang H. and Musy A. (Eds.), IAHS Publication 193, IAHS: pp. 189–196.

Pohjola V.A. and Rogers J.C. (1997) Coupling between the atmospheric circulation and extremes of the mass balance of Storglaciären, northern Scandinavia. *Annals of Glaciology*, **24**, 229–233.

Popov N. (1997) Glacial debris-flows mitigation in Kazakhstan: assessment, prediction and control. In *Proceedings of the 1st International Conference on Debris-Flow Hazard Mitigation: Mechanics, Prediction and Assessment, San Francisco, California, 7–9 Aug. 1997*, Chen C.-L. (Ed.), ASCE: New York, pp. 113–122.

Rabus B.T. and Echelmeyer K.A. (1998) The mass balance of McCall Glacier, Brooks Range, Alaska, U.S.A: its regional relevance and implications for climate change in the Arctic. *Journal of Glaciology*, **44**, 333–351.

Raymond C.F., Benedict R.J., Harrison W.D., Echelmeyer K.A. and Sturm M. (1995) Hydrological discharges and motion of Fels and Black Rapids Glaciers, Alaska, U.S.A.: implications for the structure of their drainage systems. *Journal of Glaciology*, **41**, 290–304.

Raymond C.F. and Nolan M. (2000) Drainage of a glacial lake through an ice spillway. In *Debris-Covered Glaciers*, Nakawo M., Raymond C.F. and Fountain A. (Eds.), IAHS Publication, 264, IAHS: pp. 199–207.

Reeh N. (1991) Parameterization of melt rate and surface temperature on the Greenland Ice Sheet. *Polarforschung*, **59**, 113–128.

Repp K. (1988) The hydrology of Bayelva, Spitsbergen. *Nordic Hydrology*, **19**, 259–268.

Reynolds J.M. (1998) High-altitude glacial lake hazard assessment and mitigation: a Himalayan perspective. In *Geohazards in Engineering Geology*, Maund J.G. and Eddleston M. (Eds.), Engineering Geology Special Publications 15, Geological Society: London, pp. 25–34.

Reynolds J.M., Dolecki A. and Portocarrero C. (1998) Construction of a drainage tunnel as part of glacial lake hazard mitigation at Hualcan, Cordillera Blanca, Peru. In *Geohazards in Engineering Geology*, Maund J.G. and Eddleston M. (Eds.), Engineering Geology Special Publications 15, Geological Society: London, pp. 41–48.

Ribstein P., Tiriau E., Francou B. and Saravia R. (1995) Tropical climate and glacier hydrology: a case study in Bolivia. *Journal of Hydrology*, **165**, 221–234.

Richards K., Sharp M., Arnold N., Gurnell A., Clark M., Tranter M., Nienow P., Brown G., Willis I. and Lawson W. (1996) An integrated approach to modelling hydrology and water quality in glacierized catchments. *Hydrological Processes*, **10**, 368–497.

Richardson S.D. and Reynolds J.M. (2000a) An overview of glacial hazards in the Himalayas. *Quaternary International*, **65–66**, 31–47.

Richardson S.D. and Reynolds J.M. (2000b) Degradation of ice-cored moraine dams: implications for hazard development. In *Debris-Covered Glaciers*, Nakawo M., Raymond C.F. and Fountain A. (Eds.), IAHS Publication, 264, IAHS: pp. 187–197.

Rippin D., Willis I., Arnold N., Hodson A. and Brinkhaus M. Spatial and temporal variations in surface velocity and basal drag across the tongue of the polythermal Midre Lovénbreen, Svalbard. *Journal of Glaciology*, in press.

Rippin D., Willis I., Arnold N., Hodson A., Moore J., Kohler J. and Björnsson H. (2003) Changes in geometry and subglacial drainage of Midre Lovénbreen, Svalbard determined from digital elevation models. *Earth Surface Processes and Landforms*, **28**, 273–298.

Roberts M.J., Russell A.J., Tweed F.S. and Knudsen O. (2000) Ice fracturing during jökulhlaups: implications for englacial floodwater routing and outlet development. *Earth Surface Processes and Landforms*, **25**, 1429–1446.

Roberts M.J., Russell A.J., Tweed F.S. and Knudsen O. (2001) Controls on englacial sediment deposition during the November 1996 jökulhlaup, Skeidararjökull, Iceland. *Earth Surface Processes and Landforms*, **26**, 935–952.

Roberts M.J., Russell A.J., Tweed F.S. and Knudsen O. (2002) Controls on the development of supraglacial floodwater outlets during jökulhlaups. In *The Extremes of the Extremes: Extraordinary Floods*, Snorrason Á., Finnsdóttir H.P. and Moss M.E. (Eds.), IAHS Publication 271, IAHS: pp. 71–76.

Roberts M.J., Tweed F.S., Russell A.J., Knudsen O. and Harris T.D. (2003) Hydrologic and geomorphic effects of temporary ice-dammed lake formation during jökulhlaups. *Earth Surface Processes and Landforms*, **28**, 723–737.

Röthlisberger H. and Iken A. (1981) Plucking as an effect of water-pressure variations at the glacier bed. *Annals of Glaciology*, **2**, 56–62.

Röthlisberger H. and Lang H. (1987) Glacial hydrology. In *Glacio-fluvial Sediment Transfer, an Alpine Perspective*, Gurnell A.M. and Clark M.J. (Eds.), John Wiley and Sons: Chichester, pp. 207–284.

Russell A.J. (1989) A comparison of two recent jökulhlaups from an ice-dammed lake, Søndre Strømfjord, West Greenland. *Journal of Glaciology*, **35**, 157–62.

Schneider T. (2000) Hydrological Processes in the wet-snow zone of glaciers - a review. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **36**, 89–105.

Schöner W. and Schöner M. (1997) Effects of glacier retreat on the outbursts of Goësvatnet, southwest Spitsbergen, Svalbard. *Journal of Glaciology*, **43**, 276–82.

Schuler T., Fischer U.H., Sterr R., Hock R. and Guðmundsson G.H. (2002) Comparison of modeled water input and measured discharge prior to a release event: Unteraargletscher, Bernese Alps, Switzerland. *Nordic Hydrology*, **33**, 27–46.

Shimuzu H. (1970) Air permeability of deposited snow. *Contributions from the Institute of Low Temperature Science, Series A*, **22**, 1–32.

Sigurðsson O. (1999) Jökulhlaup úr Sólheimajökull 17–18 Júlí 1999. *Jöklarannsóknafélag Íslands*, **74**, 6–7.

Singh P. and Jain S.K. (2002) Snow and glacier melt in the Satluj River at Bhakra Dam in the Western Himalayan region. *Hydrological Sciences Journal*, **47**, 93–106.

Singh P. and Kumar N. (1997) Impact assessment of climate change on the hydrological response of a snow and glacier melt runoff dominated Himalayan river. *Journal of Hydrology*, **193**, 316–350.

Singh P., Ramasastri K.S. and Kumar N. (1995) Topographical influence on precipitation distribution in the different ranges of western Himalayas. *Nordic Hydrology*, **26**, 259–284.

Skidmore M. and Sharp M. (1999) Drainage system behaviour of a high Arctic polythermal glacier. *Annals of Glaciology*, **28**, 209–215.

Smart C.C., Owens I.F., Lawson W. and Morris A.L. (2000) Exceptional ablation arising from rainfall-induced slushflows: Brewster Glacier, New Zealand. *Hydrological Processes*, **14**, 1045–1052.

Smellie J.L. (2002) The 1969 subglacial eruption on Deception Island (Antarctica): events and processes during an eruption beneath a thin glacier and implications for volcanic hazards. In *Volcano-Ice Interactions on Earth and Mars*, Smellie J.L. and Chapman M.G. (Eds.), Geological Society Special Publications, 202, Geological Society: pp. 59–79.

Snorrason Á., Jónsson P., Pálsson S., Árnason S., Sigurðsson O., Víkingsson S., Sigurðsson Á. and Zóphóníasson S. (1997) Hlaupið á Skeiðarársandi haustið 1996: ú tbreiðsla, rennsli og aurburður. In *Vatnajökull: Gos Og Hlaup*, Haraldsson H. (Ed.), Vegagerðin: Reykjavík, pp. 79–137.

Spring U. and Hutter K. (1981) Numerical studies of jökulhlaups. *Cold Regions Science and Technology*, **4**, 227–244.

Stenborg T. (1970) Delay of run-off from a glacier basin. *Geografiska Annaler A*, **52**, 1–30.

Stone K.H. (1963) Alaskan ice-dammed lakes. *Annals of the Association of American Geographers*, **53**, 332–49.

Stone D.B. and Clarke G.K.C. (1996) In situ measurements of basal water quality and pressure as an indicator of the character of subglacial drainage systems. *Hydrological Processes*, **10**, 615–628.

Sturm M. and Benson C.S. (1985) A history of jökulhlaups from Strandline Lake, Alaska, USA. *Journal of Glaciology*, **31**, 272–80.

Sugden D.E., Clapperton C.M. and Knight P.G. (1985) A jökulhlaup near Søndre Strømfjord, West Greenland, and some effects on the ice sheet margin. *Journal of Glaciology*, **31**, 366–68.

Swift D.A., Nienow P.W., Spedding N. and Hoey T.B. (2002) Geomorphic implications of subglacial drainage configuration: rates of basal sediment evacuation controlled by seasonal drainage system evolution. *Sedimentary Geology*, **149**, 5–19.

Tangborn W.V., Krimmel R.M. and Meier M.F. (1975) A comparison of glacier mass balance by glacier hydrology and mapping methods, South Cascade Glacier. *Snow and Ice*, IAHS Publication 104, IAHS: pp. 185–196.

Tangborn W. and Rana B. (2000) Mass balance and runoff of the partially debris-covered Langtang Glacier, Nepal. In *Debris-Covered Glaciers*, Nakawo M., Raymond C.F. and Fountain A. (Eds.), IAHS Publication, 264, IAHS: pp. 99–108.

Theakstone W.H. (1978) The 1977 drainage of Austre Okstindbreen ice-dammed lake, its causes and consequences. *Norsk Geografisk Tidsskrift*, **32**, 159–71.

Thorarinsson S. (1939) The ice-dammed lakes of Iceland, with particular reference to their value as indicators of glacier oscillations. *Geografiska Annaler A*, **21**, 216–42.

Thorarinsson S. (1953) Some new aspects of the Grímsvötn problem. *Journal of Glaciology*, **2**, 267–74.

Thorarinsson S. (1957) The jökulhlaup from the Katla area in 1955 compared with other jökulhlaups in Iceland. *Jökull*, **7**, 21–25.

Tómasson H. (1996) The jökulhlaup from Katla in 1918. *Annals of Glaciology*, **22**, 249–254.

Trabant D.C. and Mayo L.R. (1985) Estimation and effects of internal accumulation on five different glaciers in Alaska. *Annals of Glaciology*, **6**, 113–117.

Trabant D.C., Waitt R.B. and Major J.J. (1994) Disruption of Drift glacier and origin of floods during the 1989–1990 eruptions of Redoubt Volcano, Alaska. *Journal of Volcanology and Geothermal Research*, **62**, 369–385.

Tufnell L. (1984) *Glacier Hazards*, Longman: London.

Tweed F.S. and Russell A.J. (1999) Controls on the formation and sudden drainage of glacier-impounded lakes: implications for jökulhlaup characteristics. *Progress in Physical Geography*, **23**, 79–110.

van de Wal R.S.W., Oerlemans J. and van der Hage J.C. (1992) A study of ablation variations on the tongue of Hintereisferner, Austrian Alps. *Journal of Glaciology*, **38**, 319–324.

Vatne G., Etzelmüller B., ØdegÅrd R. and Sollid J.L. (1992) Glaciofluvial sediment transfer of a subpolar glacier, Erikbreen, Svalbard. *Stuttgater Geographische Studien*, **117**, 253–266.

Vatne G., Etzelmüller B., ØdegÅrd R.S. and Sollid J.L. (1996) Meltwater routing in a high Arctic glacier, Hannabreen, northern Spitsbergen. *Norsk Geografisk Tidsskrift*, **50**, 67–74.

Vatne G., Etzelmüller B., Sollid J.L. and ØdegÅrd R.S. (1995) Hydrology of a polythermal glacier, Erikbreen, Northern Spitsbergen. *Nordic Hydrology*, **26**, 169–190.

Vuichard D. and Zimmermann M. (1986) The Langmoche flash-flood, Khumbu Himal, Nepal. *Mountain Research and Development*, **6**, 90–93.

Vuichard D. and Zimmermann M. (1987) The 1985 catastrophic drainage of a moraine-dammed lake, Khumbu Himal, Nepal: cause and consequences. *Mountain Research and Development*, **7**, 91–110.

Wadham J.L., Hodgkins R., Cooper R.J. and Tranter M. (2001) Evidence for seasonal subglacial outburst events at a polythermal glacier, Finsterwalderbreen, Svalbard. *Hydrological Processes*, **15**, 2259–2280.

Wadham J.L. and Nuttall A.M. (2002) Multi-phase formation of superimposed ice during a mass balance year at a high Arctic glacier. *Journal of Glaciology*, **48**, 545–551.

Wagnon P., Ribstein P., Francou B. and Pouyaud B. (1999a) Annual cycle of energy balance of Zongo Glacier, Cordillera Real, Bolivia. *Journal of Geophysical Research D*, **104**, 3907–3923.

Wagnon P., Ribstein P., Francou B. and Sicart J.E. (2001) Anomalous heat and mass budget of Glaciar Zongo, Bolivia, during the 1997/98 El Niño year. *Journal of Glaciology*, **47**, 21–28.

Wagnon P., Ribstein P., Kaser G. and Berton P. (1999b) Energy balance and runoff seasonality of a Bolivian glacier. *Global and Planetary Change*, **22**, 49–58.

Wagnon P., Ribstein P., Schuler T. and Francou B. (1998) Flow separation on Zongo Glacier, Cordillera Real, Bolivia. *Hydrological Processes*, **12**, 1911–1926.

Walder J.S. and Costa J.E. (1996) Outburst floods from glacier-dammed lakes: the effect of mode of lake drainage on flood magnitude. *Earth Surface Processes and Landforms*, **21**, 701–723.

Walder J.S. and Driedger C.L. (1995) Frequent outburst floods from South Tahoma Glacier, Mount Rainier, U.S.A.: relation to debris flows, meteorological origin and implications for subglacial hydrology. *Journal of Glaciology*, **41**, 1–10.

Warburton J. and Fenn C.R. (1994) Unusual flood events from an Alpine glacier: observations and deductions on generating mechanism. *Journal of Glaciology*, **40**, 176–186.

Whalley W.B. (1971) Observations of the drainage of an ice-dammed lake - Strupvatnet, Troms, Norway. *Norsk Geografisk Tidsskrift*, **25**, 165–74.

Whittow J.B., Shepherd A., Goldthorpe J.E. and Temple P.H. (1963) Observations on the glaciers of the Ruwenzori. *Journal of Glaciology*, **4**, 581–616.

Willis I.C., Arnold N.S. and Brock B.W. (2002) Effect of snowpack removal on energy balance, melt and runoff in a small supraglacial catchment. *Hydrological Processes*, **16**, 2721–2749.

Willis I.C. and Bonvin J.-M. (1995) Climatic change in mountain environments: hydrological and water resource implications. *Geography*, **80**, 247–261.

Willis I.C., Sharp M.J. and Richards K.S. (1993) Studies of the water balance of Midtdalsbreen, Hardangerjokulen, Norway. II. Water storage and runoff prediction. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **27/28**, 117–138.

Wiscombe W.J. and Warren S.G. (1980) A model for the spectral albedo of snow I: pure snow. *Journal of Atmospheric Science*, **37**, 2712–2733.

Wolfe P.M. and English M.C. (1995) Hydrometeorlogical relationships in a glacierized catchment in the Canadian high Arctic. *Hydrological Processes*, **9**, 911–921.

Xie Z., Han J., Liu C. and Liu S. (1999) Measurement and estimative models of glacier mass balance in China. *Geografiska Annaler A*, **81**, 791–796.

Xie Z. and Lui C. (1993) Measurement method and main characteristics of the glacier mass balance in Asia. In *Application of Geographic Information Systems in Hydrology and Water Resources Management*, Kovar K. and Nachtnebel H.P. (Eds.), IAHS Publication 211, IAHS: pp. 453–459.

Yamada T. (1998) *Glacier Lake and its Outburst Flood in the Nepal Himalaya*, Monograph No. 1, Data Centre for Glacier Research, Japanese Society of Snow and Ice: p. 96.

Yamada T. and Sharma C.K. (1993) Glacier lakes and outburst floods in the Nepal Himalaya. In *Snow and Glacier Hydrology*, Young G. (Ed.), IAHS Publication 218, IAHS: pp. 319–330.

Yongjian D. and Jingshi L. (1992) Glacier lake outburst flood disasters in China. *Annals of Glaciology*, **16**, 180–184.

Young G.J. (1980) Monitoring glacier outburst floods. *Nordic Hydrology*, **11**, 285–300.

Young G.J. and Hewitt K. (1993) Glaciohydrological features of the Karakoram Himalaya: measurement possibilities and constraints. In *Snow and Glacier Hydrology*, Young G.J. (Ed.), IAHS Publication 218, IAHS: pp. 273–283.

Zhang X. (1992) Investigation of glacier bursts of the Yarkant River in Xinjiang, China. *Annals of Glaciology*, **16**, 135–139.

169: Sediment and Solute Transport in Glacial Meltwater Streams

MARTYN TRANTER

Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol, UK

This review documents the suspended sediment and major ion concentrations of glacial runoff, with a view to understanding the controls on sediment and solute fluxes from glaciated catchments. The sediment yield from glaciated catchments (those with >30% glacial cover) is often an order of magnitude greater than those of similar, nonglaciated catchments. Remarkably, sediment yield from glacier basins scales with basin area, most likely as a consequence of thicker ice thicknesses and ice fluxes, which increase the rate of subglacial erosion. The rate of glacial erosion usually exceeds the rate of transport of material from the catchment, so that following deglaciation sediment yield remains high for periods of 10^3–10^5 years. The concentration of ions in glacial runoff is typically less than that of temperate river waters, yet the solute yield from glaciated catchments is similar to that of temperate catchments when scaled by specific runoff. Glaciers are particularly efficient in dissolving carbonates, sulfides, fluid inclusions, and the surfaces of silicate minerals. This is a consequence of the production of fine-grained, geochemically reactive sediments in subglacial environments and the high water flux through many glaciated catchments. Subglacial chemical weathering is microbially mediated. Hence, glacier beds act as refugia for microbial life, which spans the range from fully oxic to fully anoxic conditions.

INTRODUCTION

Glaciers are the largest freshwater reservoir on earth (Knight, 1999), containing some $28 \times 10^6\,km^3$ of water (Table 1), and are acknowledged as powerful agents of physical erosion (Paterson, 1994; Hallet *et al.*, 1996). It is known that globally significant volumes of sediment are transferred by glacial runoff into the oceans (Syvitski *et al.*, 1987; Jones *et al.*, 2002), although the annual runoff of glacial meltwaters to the oceans is poorly known, particularly the contribution made by valley glaciers. Similarly, glacial chemical erosion is also poorly documented (Sharp *et al.*, 1995; Anderson *et al.*, 1997), but is of less significance to global cycles than the products of physical erosion. Globally significant quantities of solute may be transported by glacial runoff only during relatively brief periods ($\sim 10^2$ years) during deglaciation (Tranter *et al.*, 2002a). This review expands on these statements, commencing with a discussion of the annual runoff of glaciers, since this has a first-order control on the quantity of both the sediment and the solute that is exported from glaciated catchments. Summary of current physical and chemical erosion rates in glacial catchments (*see* **Chapter 168, Hydrology of Glacierized Basins, Volume 4**) follows, together with a discussion of the principal controls on these values. This review concludes with an assessment of the role of glaciers in global sediment and chemical cycles, and how chemical and physical erosion by glaciers impacts on certain types of ecosystems.

GLACIER HYDROLOGY, HYDROGRAPHS, FLUXES, AND SPECIFIC RUNOFF

Interest in the magnitude, duration, and variability of glacier runoff arose for two main reasons. First, glacial runoff is an important source of water to a variety of environments and anthropogenic activities (Röthlisberger and Lang, 1987; Fountain and Walder, 1998), and second, there is an intrinsic link between glacier hydrology and glacier dynamics (Copland *et al.*, 2003; Mair *et al.*, 2003; Paterson, 1994). For example, water derived from glaciers is used for agricultural irrigation (e.g. in NW China; Liu *et al.*, 1999), is fundamental to the longevity of certain

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Table 1 Volumes of terrestrial ice contained in ice sheets and other ice masses (after Knight, 1999)

Region	Current volume (10^6 km^3)	Volume at LGM (10^6 km^3)
Antarctica	26	26
Greenland	2.6	3.5
Laurentide	–	30
Cordilleran	–	3.6
Scandinavian	–	13
Other ice masses	0.2	1.1
Total	28.4	77.2

semiarid and arid ecosystems (e.g. the polar deserts of the Dry Valleys, Antarctica; McKnight *et al.*, 1998, and is stored for hydroelectric power generation (e.g. in the European Alps; Knight, 1999). Glacial runoff may be an environmental hazard, since glaciers often flood with a range of timescales from hours to weeks (Björnsson, 1998; Knight, 1999). The magnitude of runoff can be very high. For example, jökulhaups from Vatnajökull, in SE Iceland, discharge up to 4 km^3 over 15 days, with a peak discharge of 10^4 m^3 s^{-1} (Rist, 1955; Drewry, 1986), and during deglaciation in North America, a super flood from glacial Lake Agassiz is believed to have discharged 23 000 km^3 via the St Lawrence at 11 000 BP, and 75–150 000 km^3 via the Hudson Strait, contributing 6.4 and 19–42 cm equivalent sea level rise respectively (Dyke and Prest, 1987; Dawson, 1992).

Most of the work on the hydrology of glaciers has been conducted on smaller valley glaciers rather than ice sheets because of logistical and technical constraints. For example, the peak discharge from many outlet glaciers on the Greenland Ice Sheet may be of the order of 10^2–10^3 m^3 s^{-1}, and many discharge directly into coastal waters. Any attempt to construct an annual hydrograph from these outlet glaciers will be both expensive and reliant on specialist technology. By contrast, it is a more practical challenge to monitor small valley glaciers with peak discharges of the order of 10 m^3 s^{-1}. Figure 1 shows that the hydrographs of glaciers in the Alps, the High Arctic and Greenland exhibit characteristic midablation season peaks, when the receipt of solar radiation is greatest and the snow line is at its highest. Ice has a lower albedo than snow, and so there is more melt per unit incoming solar radiation. An example of the classical hydrograph of an Alpine valley glacier is shown in Figure 1(a), where prominent diurnal variations in discharge during the ablation season are evident. Diurnal variations are less evident in the hydrograph of the valley glacier in the High Arctic (Figure 1b), where there is midnight sun during summer, and less diurnal variation in the receipt of solar radiation. The basic hydrograph shape is more sensitive to the passage of cold and warm air masses across the catchment (Hodson *et al.*, 1998a,b).

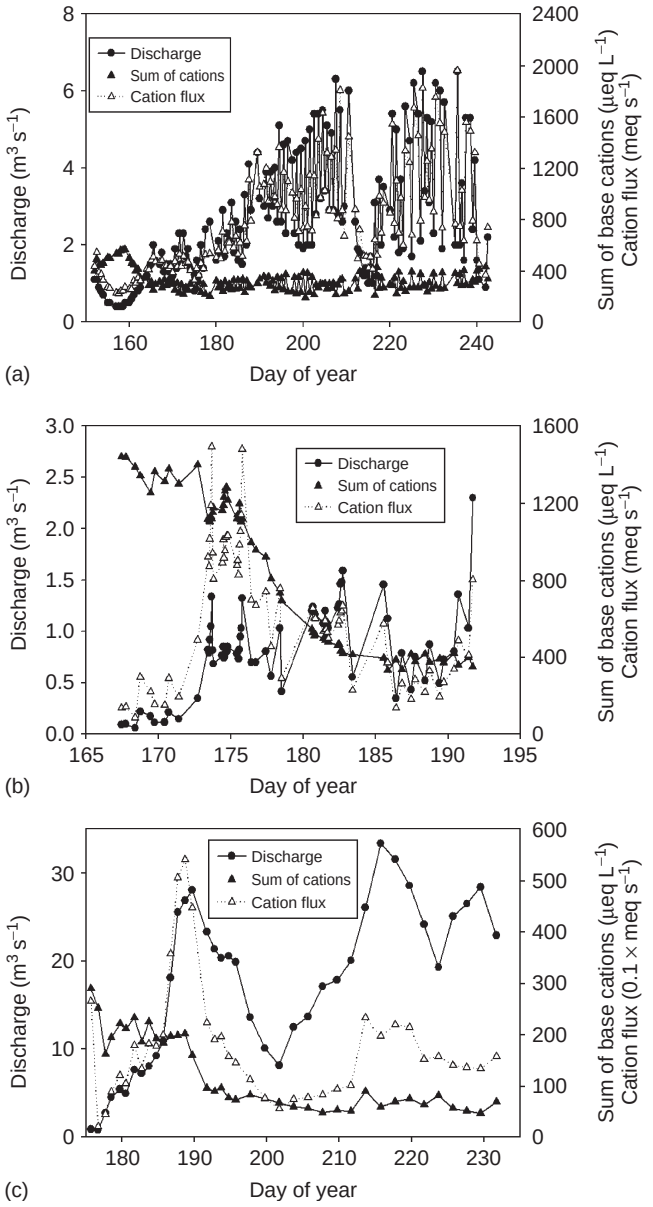


Figure 1 Hydrographs and time series of the sum of base cations and the cation flux for three glaciers with different thermal regime. (a) Haut Glacier d'Arolla, a small warm-based, valley glacier in the Swiss Alps (Brown *et al.*, 1993). Data are for 1989. (b) Scott Turnerbreen, a small, cold-based valley glacier on Svalbard (Hodgkins *et al.*, 1997). Data are for 1993, and only cover the early part of the ablation season. (c) Manitsoq Glacier, a small, polythermal-based outlet glacier to the SW of the Greenland Ice Sheet (Skidmore *et al.*, in press). Data are for 1999

The hydrograph of the outlet glacier from the Greenland Ice Sheet (Figure 1c) is dominated by an early season flood of basally stored water and the midablation season high. The large size of the potential drainage system,

which can extend back some 50 km from the ice margin, mitigates against large diurnal variations in discharge. It is evident from Figure 1 that glacial runoff exhibits considerable variation on intra-annual timescales. There may be considerable inter-annual variations also, arising from factors such as climate, precipitation regime, and glacier dynamics (Drewry, 1986).

Hydrographs for small valley glaciers are often derived from water depth monitored by pressure transducers, which are sampled at a resolution of 15–30 min. By contrast, samples for the determination of suspended sediment and solute concentrations are often collected at much longer intervals (i.e. with poorer resolution). Annual sediment and solute budgets are therefore constructed from estimates of how sediment and solute concentrations map onto the hydrograph. For example, the annual discharge can be multiplied by the volume-weighted solute concentration to give an annual load (Wadham *et al.*, 1997), or regression of discharge versus solute load can be used to convert the hydrograph into an estimate of solute load variations throughout the ablation season (Sharp *et al.*, 1995). Summation of the latter gives the annual load.

It is evident from the brief summary above that there are considerable variations in both the size of glacial catchments and the magnitude and variability of annual runoff. This could create fundamental problems in comparison of the voracity of physical and chemical erosion in different glaciated catchments. Comparison of annual solute yields among temperate catchments is often achieved by determining the specific runoff, which is the total annual runoff divided by the catchment area (Holland, 1978). The specific runoff of glacial catchments is also a useful comparative index (Anderson *et al.*, 1997; Hodson *et al.*, 2000), but it should be noted that a considerable area of the glacier bed may be hydrologically inactive beneath glaciers that are polythermal- or cold-based (since, by definition, all or a proportion of the glacier bed is frozen to the bedrock). Additionally, the catchment areas of outlet glaciers from ice sheets are hard to define, since the extent to which water is sourced from the ice sheet interior is difficult to assess (Skidmore *et al.*, in press). The specific runoff of glaciers varies from a few cm $year^{-1}$ in Antarctica (Fountain *et al.*, 1999) to several m $year^{-1}$ in lower latitudes (Hodson *et al.*, 2000), and, despite the potential drawbacks identified above, the specific runoff is currently one of the more useful metrics with which to compare solute yields from different glacial catchments (Anderson *et al.*, 1997; Hodson *et al.*, 2000). Figure 2(a) gives an illustration of the association between solute yield and specific runoff for glaciers worldwide. By contrast, Figure 2(b) shows that basin area is a poor metric with which to compare solute yields, contrasting sharply with the excellence of this metric when comparing sediment yields (Figure 3).

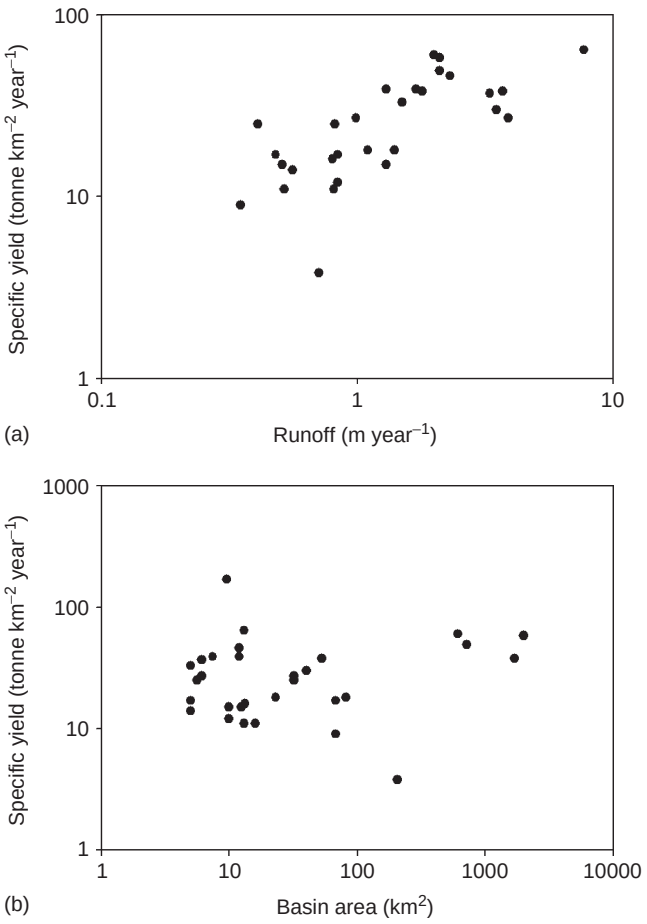


Figure 2 Scatter plots of the specific yield of solute versus (a) specific runoff and (b) basin area for worldwide glacier basins (data from Hodson *et al.,* 2000)

SUSPENDED SEDIMENT CONCENTRATIONS, BED LOAD, SEDIMENT FLUXES, AND YIELDS

Mechanisms of glacial erosion are dominated by abrasion, which produces fine sediment, and fracture/traction, which includes plucking, quarrying, crushing, and joint block removal and produces coarser sediment in general (Knight, 1999). Much of the finer material produced by subglacial erosion is transported by meltwater (Hallet *et al.*, 1996) for glaciers terminating on land. As a consequence, glacial meltwater streams are characteristically turbid, containing high concentrations of predominantly silt-sized sediment. Suspended sediment concentrations typically range from 0.2–10 kg m^{-3} (Gurnell, 1987). There is often a crude increase in suspended sediment concentration with discharge (Swift *et al.*, 2002), as shown in Figure 4, although the specific association between these variables is often complex (Gurnell, 1987; Lawson, 1993; Hodgkins, 1999). Suspended sediment concentrations are influenced on diurnal timescales by factors such as the type of subglacial

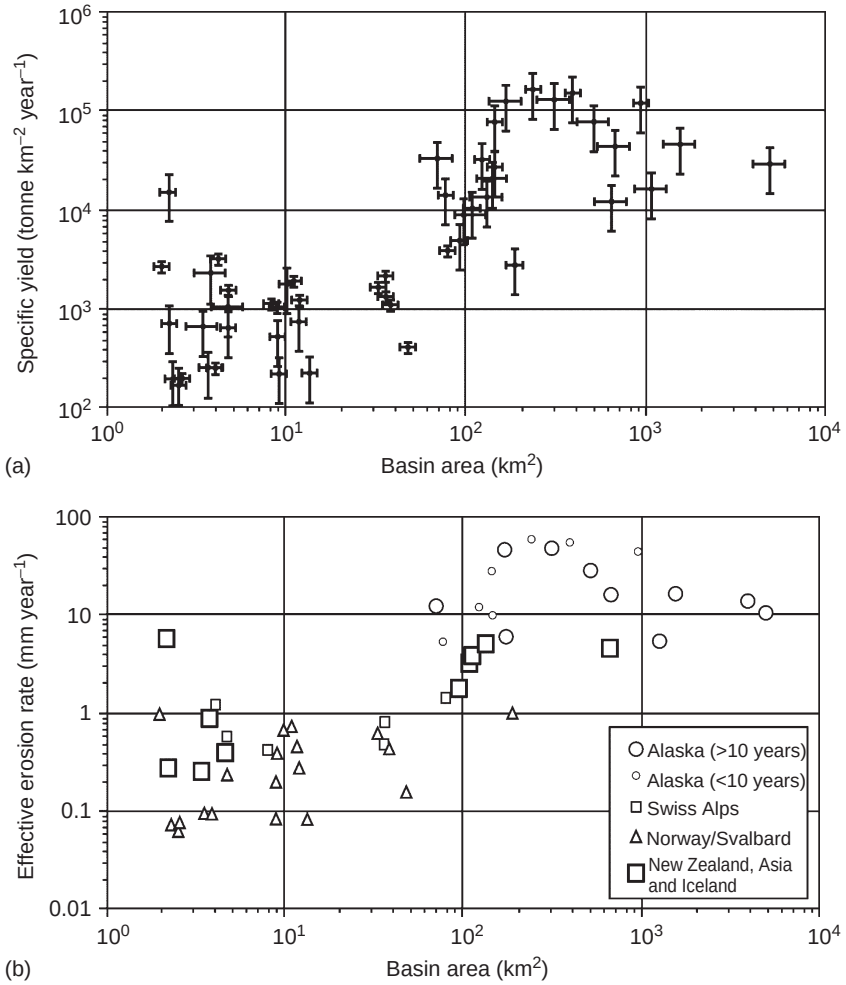


Figure 3 Scatterplots of (a) specific yield of sediment and (b) effective erosion rate versus basin area for worldwide glacier basins (after Hallet *et al.,* 1996)

drainage system (Swift *et al.*, 2002; Hodgkins, 1999), channel bank collapse, rainstorms, and periods of high snowmelt (Orwin and Smart, 2004). Very high concentrations of suspended sediment occur during outburst floods (Hodgkins *et al.*, 2003) and jökulhaups (Knight, 1999). Usually, sediment concentrations are greater on the rising limb of a hydrological event than the falling limb, since readily available sediment sources are used up or exhausted by the rising discharge, and there is often hysteresis in time series of suspended sediment concentration and transport versus discharge (Gurnell, 1987; Lawson, 1993). Figure 5(a) shows examples of hysteresis in runoff over timescales of weeks and Figure 5(b) shows hysteresis during a flood from small valley glaciers in Norway.

Bed load is seldom measured in glacial runoff channels, because of the logistical problems associated with channel migration and large seasonal changes in discharge. Where bed load has been measured, the flux of sediment at the bed is of a similar magnitude (~20–80%; Lawson, 1993)

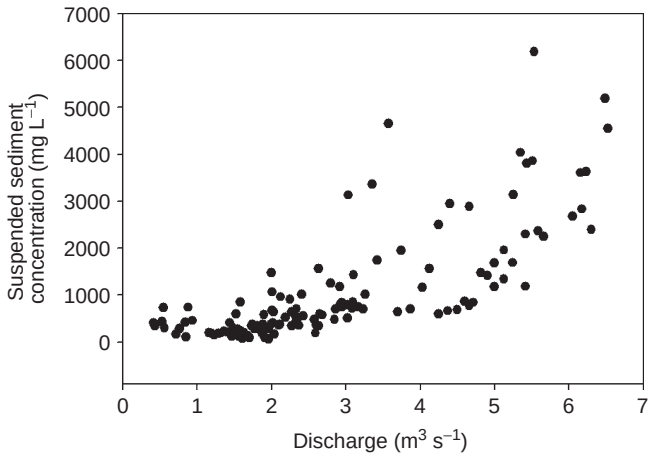


Figure 4 Scatterplot of suspended sediment concentration versus discharge for Haut Glacier d'Arolla. Data are for 1989

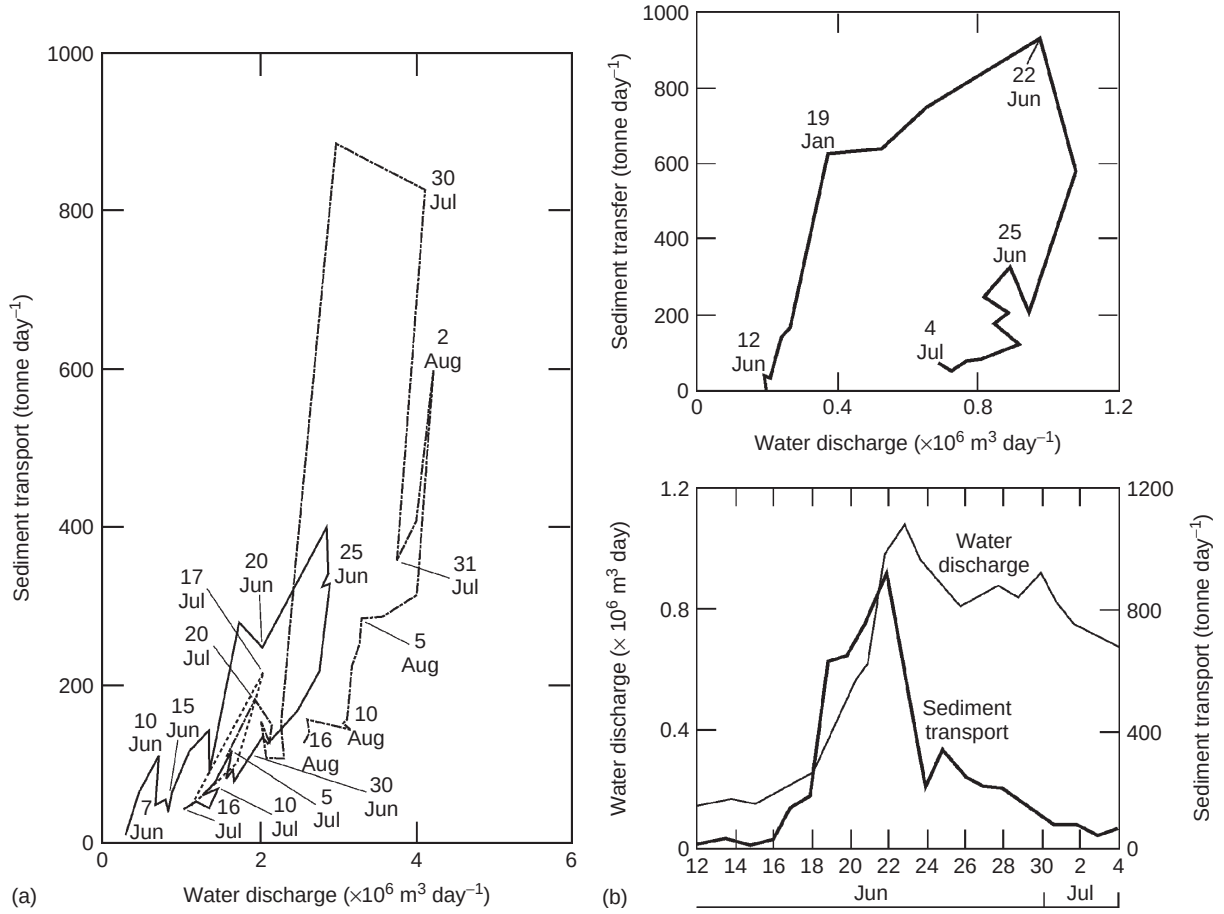


Figure 5 Hysteresis in time series of suspended sediment concentrations versus discharge. (a) Nigardsbreen, Norway, in 1993 (after Lawson, 1993). (b) Erdalsbreen, Norway in flood during 1969 (after Lawson, 1993)

to that in suspension. Bed load fluxes are greater than those in suspension at Nigardsbreen, Norway, during glacial advance (Figure 6), although the reasons for this are not clear-cut. Presumably, it is a consequence of the physical disruption of subglacial and proglacial sediments as ice thicknesses and velocities in the vicinity of the glacier margin increase.

Sediment fluxes from glaciated catchments can be obtained by a variety of means. Direct measurements usually concentrate on suspended sediment concentrations, largely ignoring bed load (Hallet *et al.*, 1996). The frequency of sampling is on timescales ranging from hours to days. Estimates of suspended sediment concentrations can also be made more frequently from continuous measurements of turbidity, although frequent calibration of the turbidity probes is necessary (Gurnell, 1987). These latter studies obtain estimates of suspended sediment concentrations with a resolution similar to that of discharge, but most studies use rating curves, which establish an association between variations in suspended sediment concentration and discharge for specific intervals of time. This approach gives rise to estimates of annual sediment yield that are of order of magnitude accuracy (Lawson, 1993). Longer-term sediment fluxes may be calculated from sedimentation rates in ice-proximal environments, such as fjords and lakes. The shortcomings with these estimates include the difficulty of defining sediment sources with certainty and problems of associated with cyclical erosion and deposition of sediment (Hallet *et al.*, 1996).

There is a positive association between the specific sediment yield and basin area in glaciated terrain (Figure 3). The largest basins are in Alaska, while those of intermediate and smallest size are from New Zealand, Asia, Iceland, Norway, Svalbard, and the Swiss Alps. There is some scatter in the data, but the data for each region show a positive linear trend, and data from different regions overlap. The overall association is that, to a first approximation, the specific sediment yield increases by approximately one order of magnitude for each order of magnitude increase in basin area. This remarkable association is believed to be the consequence of greater ice thicknesses and fluxes through the equilibrium line in larger basins, which in turn impacts on

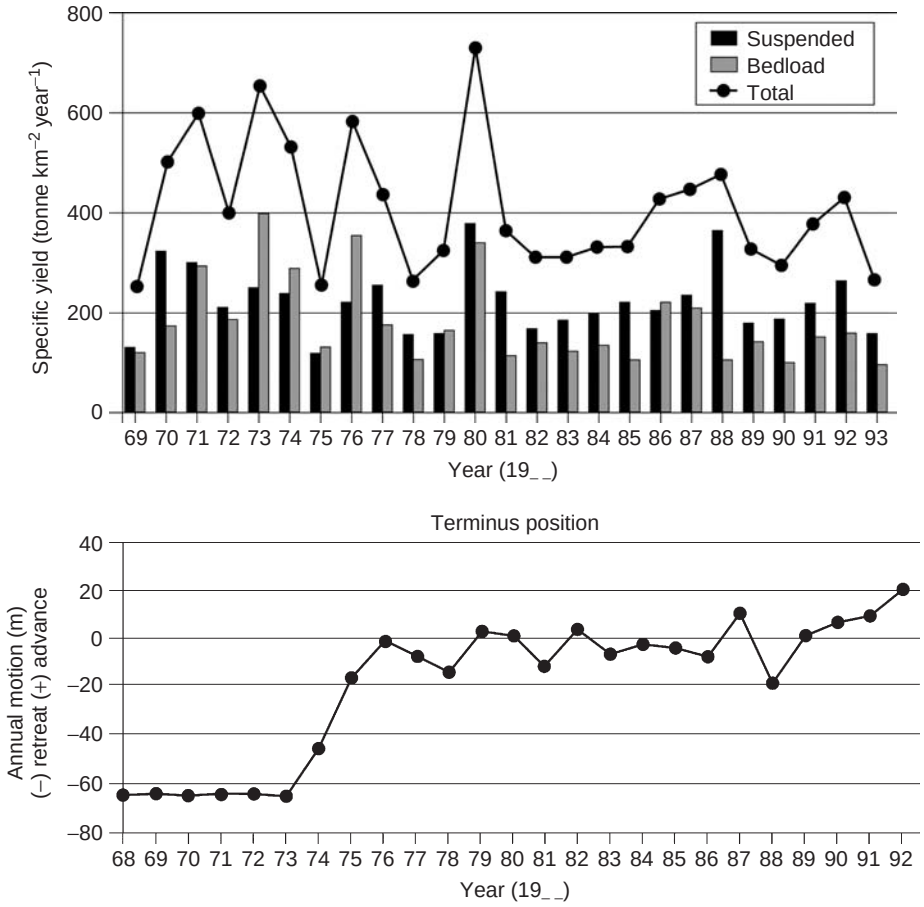


Figure 6 Suspended and bed load sediment yield for Nigardsbreen, Norway, in relation to the advance of the terminus (after Lawson, 1993)

the potential physical erosion that the glacier can undertake (Hallet *et al.*, 1996). Basins with >30% glacial cover have sediment yields that are an order of magnitude greater than those of proximal unglaciated basins (Hallet *et al.*, 1996).

It follows that current estimates of effective erosion rates (mm year^{-1}) vary by several orders of magnitude (Figure 3). Closer examination of effective rates reveals the influence of factors such as the type of glacier, the bedrock lithology, the climate, and the tectonic setting (Hallet *et al.*, 1996). The lowest effective erosion rates are for cold-based (i.e. frozen to the bed) polar glaciers and for those glaciers eroding crystalline bedrock of glaciers, and are comparable to those of the continental average (Table 2). By contrast, the highest rates are three or four orders of magnitude higher, occurring in SE Alaska, where the temperate valley glaciers are large, fast moving, and occupy tectonically active mountain ranges.

Glacial erosion rates do not stay constant during a glacial cycle, since the rate of sediment production and transport are not directly linked (Hallet *et al.*, 1996). Effective erosion increases during glacial advance, as sediment is mobilized from the proglacial zone and rates of subglacial erosion increase as a consequence of increasing ice thicknesses and fluxes. Effective erosion remains at relatively high values during the early stages of retreat, as sediment sources are plentiful and specific runoff is high. The relaxation time to background effective erosion rates following glaciation can be on the order of 10^3–10^5 years (Church and Slaymaker, 1989).

Table 2 Effective erosion in different glaciated regions (Hallet *et al.*, 1996)

Region	Effective erosion (mm a^{-1})
Svalbard	0.27–1.0
Norway	0.08–0.96
European Alps	0.41–1.7
SE Alaska	5.2–60
Canada	0.03–0.86
Central Asia	0.24–5.0
Greenland	0.01–0.04
Continental average	0.008[a]

[a] After Holland (1978).

CHEMICAL CONCENTRATIONS, SUBGLACIAL CHEMICAL WEATHERING AND CHEMICAL WEATHERING FLUXES

The major ions in glacial runoff are Ca^{2+}, HCO_3^-, and SO_4^{2-} (Table 3). The dominant non-seasalt cation is Ca^{2+}, irrespective of bedrock lithology (Raiswell, 1984), and HCO_3^- and SO_4^{2-} are the dominant non-seasalt anions because glaciers preferentially dissolve carbonates and sulfides. Glaciers are effective at promoting the solubilization of trace reactive components in the bedrock, which include carbonates, sulfides, and fluid inclusions. Laboratory experiments and direct sampling of waters from the glacier bed (Tranter *et al.*, 1997, 2002b) show that the initial reactions to occur when dilute snow and ice-melt first access glacial flour are carbonate and silicate hydrolysis (equations 1 and 2; Figure 7). These reactions raise the pH to high values (>9), lower the PCO_2 (to $\sim 10^{-6}$ atms), and maximize the potential of the water to adsorb CO_2. Carbonate hydrolysis produces a solution with a Ca^{2+} concentration of $\sim 200\,\mu eq\,L^{-1}$, with HCO_3^- the dominant anion.

$$\underset{\text{calcite}}{Ca_{1-x}(Mg_x)CO_3}\ (s) + H_2O(l) \leftrightarrow (1-x)Ca^{2+}(aq) + x\ Mg^{2+}(aq) + HCO_3^-(aq) + OH^-(aq) \quad (1)$$

$$\underset{\text{K-feldspar}}{KAlSi_3O_8(s)} + H_2O(l) \leftrightarrow \underset{\text{weathered feldspar surface}}{HAlSi_3O_8} + K^+(aq) + OH^-(aq) \quad (2)$$

Relatively dilute meltwater in contact with fine-grained glacial flour promotes the exchange of divalent ions from solution for monovalent ions on surface exchange sites. Hence, some of the Ca^{2+} and Mg^{2+} released from carbonate

Table 3 The concentration of major ions in glacial runoff from different regions of the world (after Brown, 2002). Concentrations are reported in $\mu eq\,L^{-1}$

Region	Ca^{2+}	HCO_3^-	SO_4^{2-}
Greenland	130–170	220–340	90–200
Antarctica	72–1300	91–1600	34–1200
Iceland	110–350	190–570	26–130
Alaska	550	430	260
Canadian High Arctic	260–2600	210–690	59–3900
Canadian Rockies	960–1100	890–920	380–520
Cascades	35–80	83–100	7.9–29
European Alps	20–640	11–400	10–240
Himalayas	75–590	200–730	160–410
Norway	8.8–620	1.4–680	7–140
Svalbard	120–1000	110–940	96–760
Global mean runoff	670	850	170

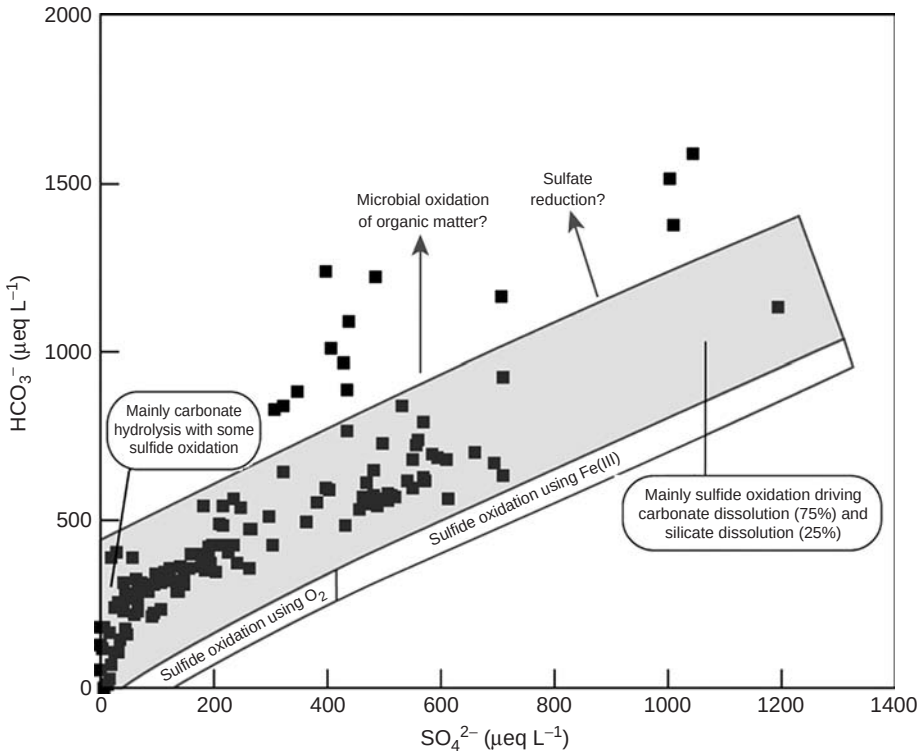


Figure 7 Scatterplot of HCO_3^- versus SO_4^{2-} for waters sampled from the bed of Haut Glacier d'Arolla (after Tranter *et al.*, 2002b). Microbial oxidation of organic matter and sulfate reduction in water-filled subglacial environments effect drive water compositions toward relatively high HCO_3^- concentrations with respect to SO_4^{2-}

and silicate hydrolysis is exchanged for Na^+ and K^+ (Figure 8).

$$\text{glacial flour} - Na_{(2-z)}, K_z(s) + (1-x)Ca^{2+}(aq) + xMg^{2+}(aq) \leftrightarrow \text{glacial flour} - Ca_{(1-x)}, Mg_x(s) + (2-z)Na^+(aq) + zK^+(aq) \quad (3)$$

The high pH derived from hydrolysis enhances the dissolution of aluminosilicate lattices, since Al and Si become more soluble at pH > 9 (Wollast, 1967). Hydrolysis of carbonates results in a solution that is near saturation with calcite. It is only in these types of waters that aluminosilicate dissolution is greater than carbonate dissolution. The influx of gases (including CO_2 and O_2), either from the atmosphere or from basal ice, and CO_2 produced by microbial respiration (see the following text) both lowers the pH and the saturation with respect to carbonates. In addition, sulfide oxidation produces acidity (see the following text). Hence, almost all subglacial meltwaters are undersaturated with respect to calcite. The rapid dissolution kinetics of carbonates with respect to silicates means that carbonate dissolution continues to have a large impact on meltwater chemistry despite carbonates being present often in only trace concentrations in the bedrock. For example, Haut Glacier d'Arolla has a bedrock that is composed of metamorphic silicate rocks. Carbonates and sulfides are present in trace quantities in bedrock samples (0.00–0.58% and <0.005–0.71% respectively). There are also occasional carbonate veins present in the schistose granite. Despite the bedrock being dominated by silicates, sulfide oxidation in subglacial environments dissolves carbonate to silicate in a ratio of ~5:1 (Tranter *et al.*, 2002b), compared to the global average of ~1.3:1 (Holland, 1978).

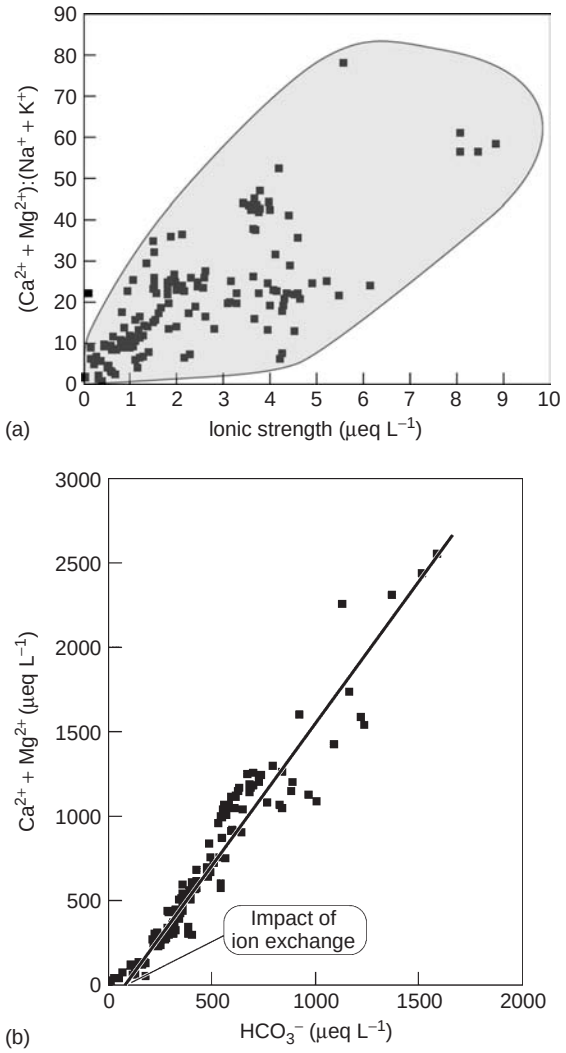


Figure 8 (a) Scatterplot of the ratio of divalent:monovalent base cations versus ionic strength for waters sampled from the bed of Haut Glacier d'Arolla (after Tranter *et al.*, 2002b) (b) Scatterplot of divalent base cations versus HCO_3^- for waters sampled from the bed of Haut Glacier d'Arolla (after Tranter *et al.*, 2002b)

The acid hydrolysis of silicates and carbonates (equations 4 and 5) that arises from the dissociation of CO_2 in solution is known as *carbonation*. Carbonation occurs in a restricted number of subglacial environments because ingress of atmospheric gases to these water-filled environments is restricted. It largely occurs in the major arterial channels at low flow, particularly near the terminus, and at the bottom of crevasses and moulins that reach the bed. Fine-grained sediment is flushed rapidly from these environments, and there is little time for the formation of secondary weathering products, such as clays. Hence, silicates dissolve incongruently (Figure 9), as crudely represented by equation 4.

$$\underset{\text{anorthite}}{CaAl_2Si_2O_8(s)} + 2CO_2(aq) + 2H_2O(l) \leftrightarrow Ca^{2+}(aq) + 2HCO_3^-(aq) + \underset{\text{weathered feldspar surfaces}}{H_2Al_2Si_2O_8(s)} \quad (4)$$

$$\underset{\text{calcite}}{Ca_{1-x}(Mg_x)CO_3(s)} + CO_2(aq) + H_2O(l) \leftrightarrow (1-x)Ca^{2+}(aq) + xMg^{2+}(aq) + 2HCO_3(aq) \quad (5)$$

There is a limited body of evidence that suggests that microbial oxidation of bedrock kerogen occurs (Wadham *et al.*, 2004), and if this is the case, carbonation as a consequence of microbial respiration may occur in debris-rich environments, such as in the distributed drainage system and the channel marginal zone.

$$C_{org}(s) + O_2(aq) + H_2O(l) \leftrightarrow CO_2(aq) + H_2O(l) \leftrightarrow H^+(aq) + HCO_3^-(aq) \quad (6)$$

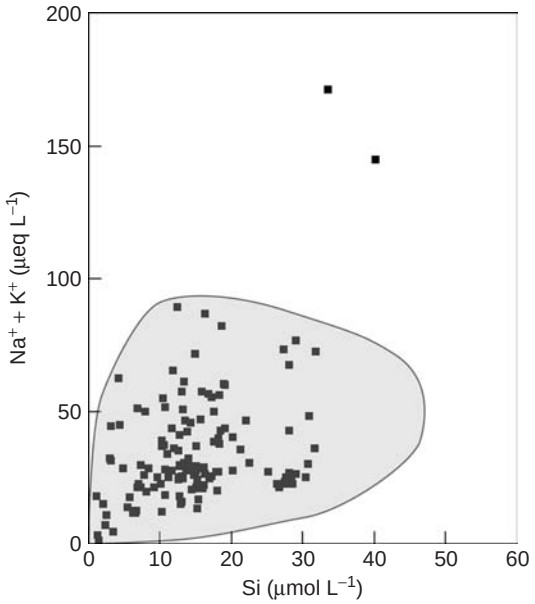


Figure 9 Scatterplot of monovalent base cations versus dissolved Si for waters sampled from the bed of Haut Glacier d'Arolla (after Tranter *et al.*, 2002b)

The dominant reaction in subglacial environments is sulfide oxidation, since, following hydrolysis, this is the major reaction that provides protons to solution thereby lowering the pH, decreasing the saturation index of carbonates, and allowing more carbonate dissolution (equation 7; Figure 7). Sulfide oxidation occurs predominantly in debris-rich environments where comminuted bedrock is first in contact with water. It is microbially mediated, occurring several orders of magnitude faster than in sterile systems (Sharp *et al.*, 1999). It consumes oxygen, driving down the PO_2 of the water.

$$\underset{\text{pyrite}}{4FeS_2(s)} + 16Ca_{1-x}(Mg_x)CO_3(s) + 15O_2(aq) + 14H_2O(l) \longleftrightarrow 16(1-x)Ca^{2+}(aq) + 16xMg^{2+}(aq) + 16HCO_3^-(aq) + 8SO_4^{2-}(aq) + \underset{\text{ferric oxyhydroxides}}{4Fe(OH)_3(s)} \quad (7)$$

Earlier studies suggested that the limit on sulfide oxidation was the oxygen content of supraglacial melt, since subglacial supplies of oxygen are limited to that released from bubbles in the ice during regelation, the process of basal ice-melting and refreezing as it flows around bedrock obstacles. However, studies of water samples from boreholes drilled to the glacier bed show that the SO_4^{2-} concentrations may be two or three times that allowed by the oxygen content of supraglacial meltwaters (Tranter *et al.*, 2002b). This suggests that oxidizing agents other than oxygen are present at the glacier bed. It seems very likely that microbially mediated sulfide oxidation drives certain sectors of the bed toward anoxia, and that in these anoxic conditions, Fe(III), rather than O_2, is used as an oxidizing agent (equation 8; Figure 7). Sources of Fe(III) include the products of the oxidation of pyrite and other Fe(II) silicates in a previous oxic environment, as well as that found in magnetite and hematite.

$$FeS_2(s) + 14Fe^{3+}(aq) + 8H_2O(l) \leftrightarrow 15Fe^{2+}(aq) + 2SO_4^{2-}(aq) + 16H^+(aq) \quad (8)$$

Support for anoxia within subglacial environments comes from the $\delta^{18}O\text{-}SO_4$, which is enriched in ^{16}O when sulfide is oxidized in the absence of oxygen (Bottrell and Tranter, 2002).

The predominance of carbonate hydrolysis, carbonation, and sulfide oxidation in subglacial weathering reactions on aluminosilicate/silicate bedrock is also found on carbonate bedrock. However, the balance between carbonate dissolution and sulfide oxidation depends on the spatial distribution of sulfides in the bedrock and basal debris (Fairchild *et al.*, 1999). Noncongruent dissolution of Sr and Mg from carbonate is also observed in high rock–water weathering environments, such as the distributed drainage system, in which water flow is also low (Fairchild *et al.*, 1999).

To date, there are few studies of glacial chemical weathering on bedrock with a significant evaporitic content. Work at John Evans Glacier in the Canadian High Arctic has shown that gypsum is dissolved in some areas of the bed, and that mixing of relatively concentrated Ca^{2+}–SO_4^{2-} waters with more dilute Ca^{2+}–HCO_3^-–SO_4^{2-} waters results in $CaCO_3$ precipitation due to the common ion effect (Skidmore, personal communication). Kennicott Glacier, Alaska, is underlain by a sabkha facies limestone, which contains trace quantities of halite. Waters accessing sites of active erosion readily acquire Na^+ and Cl^- (Anderson *et al.*, in press).

A key feature of the above chemical weathering scenarios is that relatively little atmospheric or biogenic CO_2 is involved. Hence, whereas ~23% and ~77% of solutes, excluding recycled sea salt, found in global mean river water is derived from the atmosphere and rock respectively (Holland, 1978), atmospheric sources account for a maximum of 3–11% of solute in glacial runoff (after Hodson *et al.*, 2000).

Typically, glacial runoff contains a few tens to a few hundreds of $\mu eq\,L^{-1}$ of positive charge. Higher concentrations may be found in the concentrated runoff that occurs at low discharge (Figure 1), but the discharge-weighted concentration is always much lower. Hence, glacial runoff is usually more dilute than global mean river water. The major ions are often 2–10 times higher in global mean riverine runoff than in glacial runoff (Table 3). Glacial runoff usually contains proportionally higher concentrations of K^+

and SO_4^{2-}, and lower concentrations of Si (Anderson *et al.*, 1997; Tranter, 2003).

Holland (1978) showed that specific annual discharge is the most significant control upon chemical erosion in temperate catchments. The same is true in glacierized basins. The lithology of the catchment is an important secondary control (Figure 10), with carbonate-rich and basaltic lithologies exhibiting the highest cationic denudation rates (Anderson *et al.*, 1997; Hodson *et al.*, 2000). Chemical erosion rates in glaciated catchments are usually near to or greater than the continental average (see Table 4) because glaciated catchments usually have high specific runoff. The rapid flow of water over fine-grained, recently crushed, reactive mineral surfaces maximizes both the potential rates of chemical weathering and chemical erosion.

Two studies have attempted to measure the enhancement of chemical erosion by reactions within the proglacial zone (Anderson *et al.*, 2000; Wadham *et al.*, 2001). Both studies noted an enhancement of chemical weathering rate in the proglacial zone relative to that of the glacier by a factor of 3–4. Colonization of the proglacial zone by plants is likely to further increase the rate of chemical weathering of silicates (Anderson *et al.*, 2000). However, the flux of water through sediments within the proglacial zone relative to the water flux from glaciers is small, and hence the net impact on solute export from glacial basins is likely to be of the order of 10%.

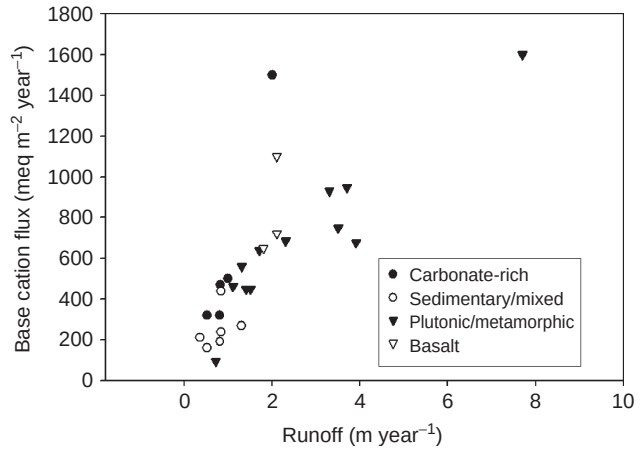


Figure 10 Scatter plot of base cation yield versus specific runoff for worldwide glacier basins of differing lithology (data from Hodson *et al.*, 2000)

Table 4 Specific runoff and cationic denudation rates for glaciers in different regions (after Hodson *et al.*, 2000)

Region	Specific runoff ($m\,a^{-1}$)	Cationic denudation rate ($\Sigma meq^{+}\,m^{-2}\,a^{-1}$)
Svalbard	0.35–1.5	190–560
European Alps	1.4–2.3	450–690
North America	0.7–7.7	94–1600
Iceland	1.8–2.1	650–1100
Asia	1.1–3.5	460–1600
Continental average	0.31	390

SIGNIFICANCE OF GLACIAL FLUXES OF SEDIMENT AND SOLUTE IN GLOBAL WATER, SEDIMENT, AND GEOCHEMICAL CYCLES

Approximately 10% ($\sim 15.9 \times 10^6\,km^2$; Knight, 1999) of the earth's surface is presently covered by glaciers, with most being concentrated in Antarctica and Greenland. Estimates of the current annual runoff from ice sheets range from $0.3–1 \times 10^{12}\,m^3\,year^{-1}$ (Jones *et al.*, 2002). Some $0.3 \times 10^{12}\,m^3\,year^{-1}$ is from Greenland (Oerlemans, 1993) and $0.4–50 \times 10^9\,m^3\,year^{-1}$ is from Antarctica (Jacobs *et al.*, 1992). The amount of runoff from smaller terrestrial ice masses is not well known. However, the potential water flux may be significant. A crude potential value can be obtained from the average residence time of ice in the smaller ice masses. Table 1 shows that the current volume of terrestrial ice that exists in locations other than the Antarctic and Greenland ice sheets is of the order of $0.2 \times 10^{15}\,m^3$. If the runoff from these glaciers comes from ice that has a volume-weighted glacier residence time of the order of $10^2–10^4$ years (Warrick and Oerlemans, 1990), then runoff from these smaller ice masses is of the order of $0.02–2.0 \times 10^{12}$. Hence, glaciers may currently contribute ~0.7–7% to current global annual runoff ($\sim 46 \times 10^{12}\,m^3\,year^{-1}$; Holland, 1978).

The volume of terrestrial ice approximately trebled at the LGM, when some ~22–29% of the present-day land surface was covered by glaciers. Table 1 shows that much of the additional ice volume was in the two great ice sheets of the northern hemisphere, the Laurentide Ice Sheet that grew over North America, and the Scandinavian Ice Sheet that grew over Europe. The average runoff from ice sheets during the last 100 kyr was $1.3 \times 10^{12}\,m^3\,year^{-1}$, approximately three times the current ice sheet runoff (Tranter *et al.*, 2002b). Average global glacial runoff might be double this value, dependent on the volume-weighted average residence time of ice in smaller ice masses. There were several occasions between 80–10 kyr BP when ice sheet runoff approached 20–30% of current global annual runoff to the oceans, but the most sustained period of relatively high ice sheet runoff was between 15–5 kyr BP, when ice sheet runoff was of the order of $2–10 \times 10^{12}\,m^3\,year^{-1}$. The impact of runoff from smaller ice masses has not been documented.

To date, modeling that attempts to determine the impact of physical and chemical erosion on atmospheric CO_2 concentrations is still at an early stage of development. Significant perturbations have not yet been found (Ludwig *et al.*, 1998; Jones *et al.*, 2002; Tranter *et al.*, 2002b).

Models of water fluxes from glaciers over cycles of advance and retreat have focused on ice sheet runoff. There is the potential for significant underestimation of total global glacial runoff if the volume-weighted runoff from smaller ice masses is calculated assuming that the mean residence time of ice in these smaller masses is the order of 10^3 years. The impact of glacial solute fluxes on ocean chemistry may become more significant as a consequence. Some 25% of all suspended sediment from rivers has been trapped in fjords over the last 100 kyr (Syvitski *et al.*, 1987), testimony to the direct and indirect impact that glaciers have had on marine sedimentation. As yet, the global impact of glacio-marine sedimentation on organic carbon burial has not been fully documented.

SIGNIFICANCE OF GLACIAL WEATHERING AND EROSION TO SUBGLACIAL ECOSYSTEMS

The realization that there is microbial mediation of certain chemical weathering reactions in subglacial environments (Sharp *et al.*, 1999; Skidmore *et al.*, 2000; Bottrell and Tranter, 2002) has resulted in a paradigm shift, since the types of reactions that may occur in anoxic sectors of the bed include the common redox reactions that occur, for instance, in lake or marine sediments (Drever, 1988). A key difference in glacial systems is that the supply of new or recent organic matter is limited to that in-washed from the glacier surface, such as algae, insects, and animal feces, or overridden soils during glacier advance. By contrast, the supply of old organic matter from comminuted rocks is plentiful. Given the thermodynamic instability of organic matter in the presence of O_2 or SO_4^{2-}, it seems likely that microbes will have evolved to colonize subglacial environments and utilize kerogen as an energy source. The first data to support this assertion is stable isotope analysis from Finsterwalderbreen, a small polythermal-based glacier on Svalbard, which has shale as a significant component of its bedrock (Wadham *et al.*, 2004). The δ^{18}O-SO_4 of waters upwelling from subglacial sediments are very enriched in ^{18}O, and the δ^{34}S is enriched in ^{34}S, which suggests that cyclical sulfate reduction and oxidation has been occurring. The δ^{13}C of DIC (dissolved inorganic carbon) is highly negative, consistent with the assertion that organic matter has been oxidized. Mass balance calculations suggest that a possible source of organic matter is kerogen, but the necromass of dead bacteria cannot be discounted. Whatever is the source of organic matter, sectors of the bed at Finsterwalderbreen are so anoxic that sulfate reduction is occurring (equation 9).

$$\underset{\text{organic carbon}}{2CH_2O(s)} + SO_4^{2-}(aq) \leftrightarrow 2HCO_3^{-}(aq) + H_2S(aq) \quad (9)$$

It is possible that methanogenesis occurs under certain ice masses, since methanogens have been isolated from subglacial debris (Skidmore *et al.*, 2000). The low δ^{13}C-CH_4 and high concentration of methane found in gas bubbles within the basal ice of the Greenland Ice Sheet are consistent with there being methanogenesis within the basal organic-rich paleosols.

The colonization of subglacial environments by microbes suggests that both energy and nutrient sources are readily available. Energy sources, such as sulfides and kerogen, have been discussed previously. Comminuted bedrock may also provide a source of nutrient. Average crustal rock contains 1050 ppm of P. Typically, this is contained in sparingly soluble minerals such as apatite, and calcium, aluminum, and ferrous phosphates (O'Neill, 1985). Comminuted bedrock and basal debris provides a renewable source of P on mineral surfaces, and it is likely that uptake of P by microbes maximizes the extraction of P from these activated surfaces. Hodson *et al.* (in press) suggest that $1-23\ \mu gP\,g^{-1}$ is present as readily extractable P on the surface of glacial flour. Sources of N may also be derived from comminuted rock (Holloway and Dahlgren, 2002). The N content of rocks is typically 20 ppm (Krauskopf, 1967), but may exceed 1000 ppm in some sedimentary and metasedimentary rock (Holloway and Dahlgren, 2002). For example, bedrock has been shown to be a source of NH_4^+ from schists in the Sierra Nevadas, California (Holloway *et al.*, 1998), and there may be appreciable concentrations of NH_4^+, which substitutes for K^+, in biotite, muscovite, K-feldspar, and pagioclase (Mingram and Brauer, 2001). It follows that glacial comminution of bedrock and basal debris maximizes the likelihood that N-producing surfaces are exposed to meltwaters and microbes, and given that bedrock in the Sierra Nevadas can act as an N source, it is likely that comminuted glacial debris is also a potential source of N.

CONCLUSIONS

This review has attempted to quantify the typical concentrations of suspended sediment and solute in glacial meltwaters, and shown that sediment and solute yield scale with basin area and specific runoff respectively. Globally significant quantities of sediment are transported by glacial runoff to the oceans, but significant quantities of the major ions are only transported during deglaciation, when large quantities of dilute meltwater are discharged into the oceans. That glaciers comminute bedrock and allow the transit of water at their beds provides a set of physical and chemical conditions that allow glacier beds to act as refugia for microbial life. Comminuted subglacial flour appears to contain sufficient energy sources and nutrient to sustain a range of microbes that inhabit oxic

through anoxic environments. Hence, even waters draining glaciers and ice sheets, in common with waters in other rivers at the earth surface, contain fingerprints of biological activity.

REFERENCES

Anderson S.P., Drever J.I., Frost C.D. and Holden P. (2000) Chemical weathering in the foreland of a retreating glacier. *Geochimica et Cosmochimica Acta*, **64**, 1173–1189.

Anderson S.P., Drever J.I. and Humphrey N.F. (1997) Chemical weathering in glacial environments. *Geology*, **25**, 399–402.

Anderson S.P., Longacre S.A. and Kraal E.R. (2003) Patterns of water chemistry and discharge in the glacier-fed Kennicott River, Alaska: evidence for subglacial water storage cycles. *Chemical Geology*, **202**, 297–312.

Björnsson H. (1998) Hydrological characteristics of the drainage system beneath a surging glacier. *Nature*, **395**, 771–774.

Bottrell S.H. and Tranter M. (2002) Sulphide oxidation under partially anoxic conditions at the bed of Haut Glacier d'Arolla, Switzerland. *Hydrological Processes*, **16**, 2363–2368.

Brown G.H. (2002) Glacier meltwater hydrochemistry. *Applied Geochemistry*, **17**, 855–883.

Brown G.H., Tranter M., Sharp M.J. and Gurnell A.M. (1993) The impact of post-mixing chemical reactions on the major ion chemistry of bulk meltwaters draining the Haut Glacier d'Arolla, Valais, Switzerland. *Hydrological Processes*, **8**, 465–480.

Church M. and Slaymaker O. (1989) Disequilibrium of Holocene sediment yield in glaciated British Columbia. *Nature*, **337**, 452–454.

Copland L., Sharp M.J. and Nienow P. (2003) Links between short-term velocity variations and the subglacial hydrology of a predominantly cold polythermal glacier. *Journal of Glaciology*, **49**, 337–348.

Dawson A.G. (1992) *Ice Age Earth. Late Quaternary Geology and Climate*, Routledge: 293 pp.

Drever J.I. (1988) *The Geochemistry of Natural Waters*, *Second Edition*, Prentice Hall: 437 pp.

Drewry D. (1986) *Glacial Geologic Processes*, Edward Arnold: 276 pp.

Dyke A.S. and Prest V.K. (1987) Late Wisconsian and Holocene history of the Laurentide Ice Sheet. *Geographie Physique et Quaternaire*, **41**, 237–264.

Fairchild I.J., Killawee J.A., Sharp M.J., Hubbard B., Lorrain R.D. and Tison J.-L. (1999) Solute generation and transfer from a chemically reactive alpine glacial-proglacial system. *Earth Surface Processes and Landforms*, **4**, 1189–1211.

Fountain A.G. and Walder J.S. (1998) Water flow through temperate glaciers. *Reviews of Geophysics*, **36**, 299–328.

Fountain A.G., Lyons W.B., Burkins M.B., Dana G.L., Doran P.T., Lewis K.J., McNight D.M., Moorhead D.L., Parsons A.N., Priscu J.C., Wall D.H., Wharton Jr. R.A. and Virginia R.A. (1999) Physical controls on the Taylor Valley ecosystem, Antarctica. *Bioscience*, **49**, 961–971.

Gurnell A.M. (1987) Suspended sediment. In *Glacio-Fluvial Sediment Transfer*, Gurnell A.M. and Clarke M.J. (Eds.), Wiley: Chichester, pp. 305–354.

Hallet B., Hunter L. and Bogen J. (1996) Rates of erosion and sediment evacuation by glaciers: A review of field data and their implications. *Global and Planetary Change*, **12**, 213–235.

Hodgkins R. (1999) Controls on suspended-sediment transfer at a High-Arctic glacier, determined from statistical modelling. *Earth Surface Processes and Landforms*, **24**, 1–21.

Hodgkins R., Cooper R., Wadham J. and Tranter M. (2003) Suspended sediment fluxes in a high-Arctic glacierised catchment: implications for fluvial sediment storage. *Sedimentary Geology*, **162**, 105–117.

Hodgkins R., Tranter M. and Dowdeswell J.A. (1997) Solute provenance, transport and denudation in a high Arctic glacierised catchment. *Hydrological Processes*, **11**, 1813–1832.

Hodson A.J., Gurnell A.M., Tranter M., Bogen J., Hagen J.O. and Clark M. (1998a) Suspended sediment yield and transfer processes in a small High-Arctic glacier basin, Svalbard. *Hydrological Processes*, **12**, 73–86.

Hodson A.J., Gurnell A.M., Washington R., Tranter M., Clark M.J. and Hagen J.O. (1998b) Meteorological and runoff time-series characteristics in a small, high-Arctic glaciated basin, Svalbard. *Hydrological Processes*, **12**, 509–526.

Hodson A.J., Mumford P. and Lister D. (2004) Suspended sediment and phosphorus in proglacial rivers: bioavailability and potential impacts upon the P status of ice-marginal receiving waters. *Hydrological Processes*, **18**, 2409–2422.

Hodson A.J., Tranter M. and Vatne G. (2000) Contemporary rates of chemical weathering and atmospheric CO_2 sequestration in glaciated catchments: an Arctic perspective. *Earth Surface Processes and Landforms*, **25**, 1447–1471.

Holland H.D. (1978) *The Chemistry of the Atmosphere and Oceans*, Wiley: New York, 351 pp.

Holloway J.M. and Dahlgren R.A. (2002) Nitrogen in rock: occurrences and biogeochemical implications. *Global Biogeochemical Cycles*, **16**, 1118, doi:10.1029/2002GB001862.

Holloway J.M., Dahlgren R.A., Hansen B. and Casey W.H. (1998) Contribution of bedrock nitrogen to high nitrate concentrations in stream water. *Nature*, **395**, 785–788.

Jacobs S.S., Helmer H.H., Doake C.S.M., Jenkins A. and Frolich R.M. (1992) Melting of ice shelves and the mass balance of Antarctica. *Journal of Glaciology*, **38**, 375–387.

Jones I.W., Munhoven G., Tranter M., Huybrechts P. and Sharp M.J. (2002) Modelled glacial and non-glacial HCO_3^-, Si and Ge fluxes since the LGM: little potential for impact on atmospheric CO_2 concentrations and the marine Ge:Si ratio. *Global and Planetary Change*, **33**, 139–153.

Knight P.G. (1999) *Glaciers*, Stanley Thornes: Cheltenham, 261 pp.

Krauskopf K.B. (1967) *Introduction to Geochemistry*, McGraw-Hill: 721 pp.

Lawson D.E. (1993) *Glaciohydrological and Glaciohydraulic Effects on Runoff and Sediment Yield in Glacierized Basins*, US Army Corps of Engineers, Cold Regions Research and Engineering Laboratory: Monograph, 93–2.

Liu J., Yoshihiro F. and Tetsuya H. (1999) Hydrological response of meltwater from glacier covered mountain basins to climate change in northwest China. *IAHS Publication*, **256**, 193–207.

Ludwig W., Amiotte-Suchet P., Munhoven G. and Probst J.-L. (1998) Atmospheric CO_2 consumption by continental erosion:

present-day controls and implications for the last glacial maximum. *Global and Planetary Change*, **16-17**, 107–120.

McKnight D.M., Alger A., Tate C.M., Shupe G. and Spauldinget S. (1998) Longitudinal patterns in algal abundance and species distribution in meltwater streams in Taylor Valley, Southern Victoria Land, Antarctica. *Antarctic Research Series*, **72**, 109–127.

Mingram B. and Brauer K. (2001) Ammonium concentration and nitrogen isotope composition in metasedimentary rocks from different tectonometamorphic units of the European Variscan Belt. *Geochimica et Cosmochimica Acta*, **65**, 273–287.

Mair D., Willis I., Fisher U.H., Hubbard B., Nienow P. and Hubbard A. (2003) Hydrological controls on patterns of surface, internal and basal motion during three "spring events". Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, **49**, 555–567.

Oerlemans J. (1993) Evaluating the role of climate cooling in iceberg production and Heinrich events. *Nature*, **364**, 783–786.

O'Neill P.O. (1985) *Environmental Chemistry*, George Allen and Unwin: 232 pp.

Orwin J.F. and Smart C.C. (2004) Short-term spatial and temporal patterns of suspended sediment transfer in proglacial channels, Small Rive Glacier, Canada. *Hydrological Processes*, **18**, 1521–1542.

Paterson W.S.B. (1994) *The Physics of Glaciers*, *Third Edition*, Pergamon: 480 pp.

Raiswell R. (1984) Chemical models of solute acquisition in glacial meltwaters. *Journal of Glaciology*, **30**, 49–57.

Rist S. (1955) Skeidarahlaup 1954. The hlaup of Skeidara 1954. *Jökull*, **5**, 30–36.

Röthlisberger H. and Lang H. (1987) *Glacial hydrology*. In *Glacio-Fluvial Sediment Transfer*, Gurnell A.M. and Clarke M.J. (Eds.), Wiley: Chichester, pp. 207–284.

Sharp M., Parkes J., Cragg B., Fairchild I.J., Lamb H. and Tranter M. (1999) Bacterial populations at glacier beds and their relationship to rock weathering and carbon cycling. *Geology*, **27**, 107–110.

Sharp M., Tranter M., Brown G.H. and Skidmore M. (1995) Rates of chemical denudation and CO_2 drawdown in a glacier-covered alpine catchment. *Geology*, **23**, 61–64.

Skidmore M.L., Foght J.M. and Sharp M.J. (2000) Microbial life beneath a High Arctic glacier. *Applied and Environmental Microbiology*, **66**, 3214–3220.

Skidmore M., Jackson A., Grust K., Nienow P. and Tranter M. (in prep.). Seasonal suspended sediment dynamics at a polythermal outlet glacier of the Greenland Ice Sheet. *Annals of Glaciology*.

Swift D.A., Nienow P.W., Spedding N. and Hoey T.B. (2002) Geomorphic implications of subglacial drainage configuration: rates of basal sediment evacuation controlled by seasonal drainage system evolution. *Sedimentary Geology*, **149**, 5–19.

Syvitski J.P.M., Burrell D.C. and Skei J.M. (1987) *Fjords*, Springer-Verlag: 379 pp.

Tranter M. (2003) Chemical weathering in glacial and proglacial environments. In *Treatise on Geochemistry*, Vol. 5, Holland H.D. and Turekian K.K. (Eds.), Elsevier: Pergamon; *Surface and Ground Water, Weathering, Erosion and Soils*, Drever J.I. (Ed.), Elsevier: Pergamon, pp. 189–205.

Tranter M., Huybrechts P., Munhoven G., Sharp M.J., Brown G.H., Jones I.W., Hodson A.J., Hodgkins R. and Wadham J.L. (2002a) Glacial bicarbonate, sulphate and base cation fluxes during the last glacial cycle, and their potential impact on atmospheric CO_2. *Chemical Geology*, **190**, 33–44.

Tranter M., Sharp M.J., Brown G.H., Willis I.C., Hubbard B.P., Nielsen M.K., Smart C.C., Gordon S., Tulley M. and Lamb H.R. (1997) Variability in the chemical composition of in situ subglacial meltwaters. *Hydrological Processes*, **11**, 59–77.

Tranter M., Sharp M.J., Lamb H.R., Brown G.H., Hubbard B.P. and Willis I.C. (2002b) Geochemical weathering at the bed of Haut Glacier d'Arolla, Switzerland – a new model. *Hydrological Processes*, **16**, 959–993.

Wadham J.L., Bottrell S., Tranter M. and Raiswell R. (2004) Stable isotope evidence for microbial sulphate reduction at the bed of a polythermal high Arctic glacier, *Earth and Planetary Science Letters*, **219**, 341–355.

Wadham J.L., Cooper R.J., Tranter M. and Hodgkins R. (2001) Enhancement of glacial solute fluxes in the proglacial zone of a polythermal glacier. *Journal of Glaciology*, **47**, 378–386.

Wadham J.L., Hodson A.J., Tranter M. and Dowdeswell J.A. (1997) The rate of chemical weathering beneath a quiescent, surge-type, polythermal based glacier, southern Spitsbergen. *Annals of Glaciology*, **24**, 27–31.

Warrick R. and Oerlemans J. (1990) Sea level rise. In *Climate Change. The IPCC Scientific Assessment*, Houghton J.T., Jenkins G.J. and Ephraums J.J. (Eds.), Cambridge University Press: Cambridge, pp. 257–281.

Wollast R. (1967) Kinetics of the alteration of K^+-feldspar in buffered solutions at low temperature. *Geochimica et Cosmochimica Acta*, **31**, 635–648.

170: Modeling Glacier Hydrology

REGINE HOCK AND PETER JANSSON

Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden

Modeling glacier hydrology is essential for many aspects of glacier research and water resources planning. Model complexity has developed in parallel with our increasing understanding of the glacier system. Different models can be used depending on the purpose of the modeled results. Stochastic models constitute the first attempts to model glacier discharge. Use of such models has declined because they require extensive site-specific calibrations and lack a conceptual basis, which makes them cumbersome to use. Conceptual models are the most widely used models for estimating discharge from glacierized basins. Commonly, the concept of linear reservoirs is applied to route melt and rain water through the glacier. Despite the wide variety of models of differing complexity, common to most conceptual models is that they omit many of the physical processes. Therefore attempts have been made to build physically based models that account for all processes in the glacier hydrological system. Because of the complexity of the modeled system and the wide variety of conditions on, in and under glaciers, the first models are relatively site-specific. However, despite the inherent difficulties, they still yield good results. Physically based models will likely play an increasingly important role in glaciohydrological research while the conceptual models will continue to dominate the monitoring and forecasting modeling in applied hydrology because of their modest data requirements and ease of use.

INTRODUCTION

Modeling glacier hydrology is of great scientific and practical interest. Modeling efforts have contributed substantially to our understanding of the processes and mechanisms of glacier hydrology. Increasing human use of areas influenced by glacier runoff has nourished the development of models to predict glacier runoff. Such models are widely used for all aspects of watershed management, including optimization of hydroelectric power schemes, reservoir operation, water supply, and flood forecasting, but also for assessing the contribution of glaciers to sea-level change. The need for modeling is particularly emphasized in the face of global warming and expected enhanced glacier retreat (IPCC, 2001). In the longer term, continued glacier mass loss will invoke a risk of low flow in for example, semiarid areas since water amounts currently delivered by glacier melt will diminish as glaciers become smaller. On the other hand, short-term effects of enhanced glacier melt will lead to an increased risk for floods in the vicinity of glaciers, as peak flows increase dramatically, mostly due to faster runoff generation when snow and firn cover vanish (Braun *et al.*, 2000; Hock *et al.*, 2005).

From a hydrological perspective, glaciers represent important water resources contributing substantial quantities of water to streamflow in many areas of the world, even in lowlands beyond the source areas in mountain ranges. Glacier hydrology differs from conventional hydrology since glaciers significantly modify streamflow in quantity, timing, and variability by temporarily storing water as snow and ice on many different timescales (Jansson *et al.*, 2003). Dominant characteristics of glacier discharge include pronounced melt-induced diurnal cyclicity and a concentration of annual runoff during the melt season (Escher-Vetter and Reinwarth, 1994). Many areas benefit from this specific seasonal runoff variation characteristic for glaciers in mid- and high-latitudes since ice meltwater is typically released during periods of otherwise low flow conditions. Total annual runoff is enhanced or decreased in years of negative or positive mass balances, respectively. In addition, glaciers tend to act as streamflow regulators reducing year-to-year variability to a minimum at moderate fractions

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

(often 10–40%) of catchment glacierization (Röthlisberger and Lang, 1987).

The seasonal change in glacier discharge hydrographs reflects fluctuations in meltwater inputs but also an increase in efficiency of the glacier drainage system as it evolves during the melt season. Hence, modeling glacier hydrology consists of two principal steps: (i) estimation of water input to the glacier, (ii) discharge routing through the glacier, that is, the transformation of rain- and meltwater into a runoff hydrograph. Melt modeling is relatively advanced with a hierarchy of ice and snow melt models ranging from simple *temperature index models*, which relate melt in empirical expressions to one or more variables including air temperature (Hock, 2003) to more physically based *energy balance models*, which compute melt from an assessment of the energy fluxes to and from the surface (*see* **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**). The processes of water movement in and under a glacier, however, are intrinsically complex and far less understood. Temperate glaciers consist of spatially zoned aquifers with distinctly different hydraulic properties. Water movement through firn and snow occurs by slow percolation (Schneider, 2000), whereas the bulk of water in and under the ice body generally travels along a well-defined passage system with considerably larger travel velocities. A variety of models of subglacial drainage have been invoked, including channels incised into ice or underlying bedrock, sheet flow, linked-cavity system (tortuous system of cavities connected by small channels or orifices), flow through subglacial unconsolidated sediments and "canals" in deformable beds (see reviews in e.g. Hubbard and Nienow, 1997; Fountain and Walder, 1998; *see also* **Chapter 167, Subglacial Drainage, Volume 4**). Channels are subject to change from both creep processes and phase changes (Röthlisberger, 1972; Hooke, 1984). Various modes of flow have been shown to coexist but switch locally from one to another during the melt season and between years (Nienow *et al.*, 1998). The shape of discharge hydrographs changes over the season owing to increased efficiency of the glacier drainage system as the winter snow cover melts away and the internal system develops. The continuous process of snow metamorphism, the high deformability of ice under pressure and phase transformations, since the aquifer consists of a mixture of water in its liquid and solid phase, contribute to the complexity of glacial drainage processes. The system is hence dynamic at all timescales from minutes to years, rendering modeling inherently difficult.

While models of internal glacier drainage are dealt with extensively by **Chapter 166, Surface and Englacial Drainage of Glaciers and Ice Sheets, Volume 4** and **Chapter 167, Subglacial Drainage, Volume 4**, this review focuses on modeling of glacier hydrology from a water resources perspective, hence on models predicting glacier runoff, and representation of internal flow routing in such models. Such work has to date focused on predominantly temperate glaciers. Very few operational models have been developed exclusively for glacier melt-runoff estimation (Fountain and Tangborn, 1985; Singh, 1995). Mostly, snowmelt runoff routines in existing runoff models are used or adjusted for glacier-runoff modeling. Generally, models fall into three categories: stochastic, conceptual, and physically based models.

STOCHASTIC MODELS

Stochastic models relate glacier runoff directly to meteorological variables using multiple regression techniques (Lang, 1968; Jensen and Lang, 1973; Østrem, 1973). Such models were widely developed in the 1960s and 1970s prompted by a need for seasonal and short-term runoff forecasts in dimensioning and efficient operation of hydroelectric facilities. These models are typically adjusted and calibrated to specific glacier basins. Results have often shown that meteorological variables like air temperature, wind speed, and vapor pressure contribute significantly to the overall correlation, while the contribution of radiative fluxes is surprisingly small, despite their general dominance in the heat budget (Østrem, 1973).

Through stepwise multiple regression of up to 10 observed variables in various Norwegian glacierized catchments, Østrem (1973) found that discharge could be well explained by only two to three meteorological variables, usually air temperature, precipitation, and wind speed. Inclusion of additional variables did not enhance correlation. Best results were often obtained when products of meteorological variables, instead of single variables, were used as independent variables. In addition models including the meteorological averages for the preceding two days were established to account for any time lag between meteorological variables and runoff. Lang (1968) and Jensen and Lang (1973) developed sets of multiple regression equations for daily and hourly discharge for several glaciers in the Swiss Alps to serve primarily for operational purposes. Forecasts were based on air temperature, precipitation, global radiation, and vapor pressure and included data from up to five preceding days. In addition, antecedent discharge turned out to be a powerful predictor. In contrast to Østrem (1973), the effect of wind speed appeared negligible, while vapor pressure was an important variable. To accommodate the glacier drainage system evolution, the melt season was subdivided into three intervals based on observed runoff characteristics employing different sets of equations for each interval.

In spite of accurate forecasts in practical operational hydrology, stochastic models lack a conceptual basis and general applicability since they require detailed regression analysis and calibration of coefficients at each site. Their

calibration is based on observational records, and they thus fail to provide reliable results in assessing the response of glacier discharge to future climate changes.

CONCEPTUAL MODELS

Conceptual models are most widely used for both practical and scientific studies of glacier hydrology. They generally consist of series of numerical steps representing the known physical processes in a simplified manner (Kachroo, 1992). Owing to generally limited knowledge of the large heterogeneity of physical processes in a glacier basin, some spatial variability of the input data is ignored by replacing spatially variable functions by their areal means (lumping). Nevertheless, the strength of conceptual models is moderate data requirements, and hence widespread applicability in hydrological modeling.

Linear Reservoir Models

In models considering internal flow routing, the most widely adopted concept to route water through glaciers is the concept of *linear reservoirs* (Chow *et al.*, 1988). A linear reservoir can be visualized as a water-filled container with a tap at the bottom, where outflow, $Q_{(t)}$, is proportional to the stored water volume, $V_{(t)}$. The factor of proportionality is the *storage constant*, k, with units of time, representing the time shift between the centroid of the inflow and that of the outflow. The use of linear reservoirs thus accommodates the storage effect and resulting delay of water in the glacier hydrological system. Storage and continuity equations for the reservoir are

$$V_{(t)} = kQ_{(t)} \tag{1}$$

and

$$\frac{\mathrm{d}V}{\mathrm{d}t} = I_{(t)} - Q_{(t)} \tag{2}$$

where $I_{(t)}$ is the inflow, here the sum of melt- and rainwater. In the case of constant $Q = I$, k corresponds to the time required for water entering the reservoir to flow out at the bottom, and thus to the residence time of the water in the reservoir. Combining both equations gives

$$k\frac{\mathrm{d}Q}{\mathrm{d}t} = I_{(t)} - Q_{(t)} \tag{3}$$

Integrating equation (3) yields

$$Q_{(t)} = Q(t_0)e^{-(t-t_0)/k} + I_{(t)}\left[1 - e^{-(t-t_0)/k}\right] \tag{4}$$

where t_0 is the time step preceding t.

The glacier is usually subdivided into one or several linear reservoirs with different storage constants (Figure 1): one reservoir (NAM-model: Gottlieb, 1980; HYMET model: Tangborn, 1984; Moore, 1993), two reservoirs (UBC model: Quick and Pipes, 1977; Van de Wal and Russell, 1994; Hannah and Gurnell, 2001) or three reservoirs (Baker *et al.*, 1982; Hock and Noetzli, 1997; Escher-Vetter, 2000). The reservoirs are coupled either in series with outflow of one reservoir providing the inflow to another reservoir (e.g. Van de Wal and Russell, 1994), or in parallel, whereby outflow of each individual reservoir is summed up for each time step to yield total glacier runoff (e.g. Baker *et al.*, 1982). Escher-Vetter (2000) used three linear reservoirs but added a constant groundwater flow which is equivalent to a fourth linear reservoir with an infinite storage constant. By assigning different storage constants to different reservoirs, this model formulation accommodates the markedly different through-flow velocities characteristic of the different reservoirs of a glacier.

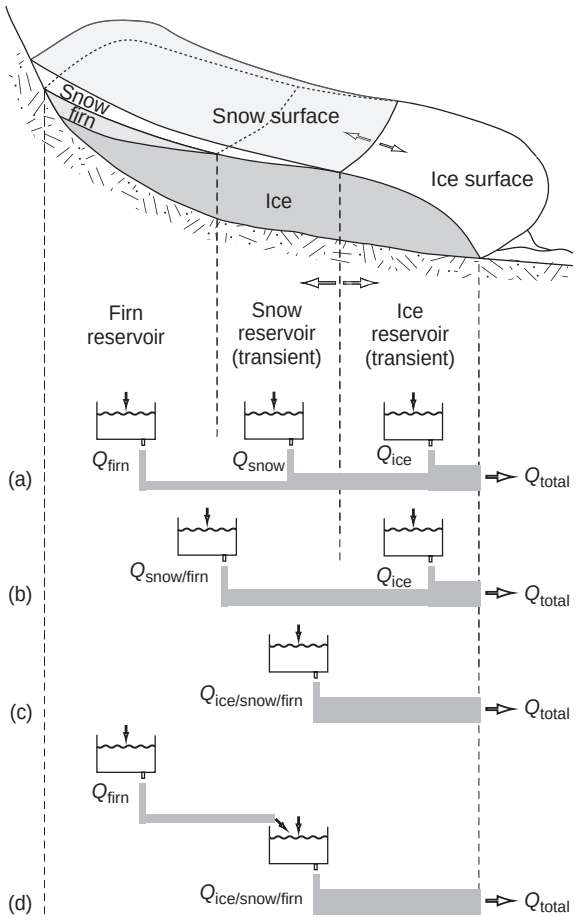


Figure 1 Concept of linear reservoirs as applied to glaciers using one to three (c–a) different linear reservoirs. Reservoirs are coupled in parallel in (a–b), and in series in (d). Exact delineation of reservoirs varies between studies. Q is outflow from the reservoirs

For example, Hock and Noetzli (1997) elaborate the work of Baker *et al.* (1982) by subdividing the glacier into three reservoirs discriminated by their surface characteristics: a firn reservoir defined by the area above the previous year's equilibrium line; a seasonally variable snow reservoir defined as the snow-covered area outside the firn reservoir; and an ice reservoir, defined as the area of exposed ice (Figure 1a). A large and small storage constant are assigned to the firn and ice reservoir, respectively, with an intermediate value for the snow reservoir, thus accounting for the different hydraulic properties of these media with respect to water transit times. Since the area of exposed ice grows at the expense of the snow-covered area as the melt season proceeds, increasing portions of water are routed through the "faster" ice reservoir, generating more peaked hydrographs (Figure 2). In this way, seasonal changes resulting from areal increases and decreases in firn, snow, and ice reservoirs are accounted for. However, most studies assume time-invariant storage constants, thus ignoring any temporal changes within each reservoir, such as enhanced drainage efficiency of the subglacial channel system or thinning of the snowpack during the course of the melt season.

Moore (1993) applied a single glacier reservoir and allowed the storage coefficient to vary on a daily basis according to an Antecedent Flow Index (AFI), which is computed as a function of the inflow to the reservoir for the current day. Although in his study the model performed better compared to employment of a constant storage coefficient, the approach does not account for the influence of any other factors on the evolution of the drainage system and also fails to capture any subdaily variations. Generally, storage constants are obtained either by tuning, that is, by maximizing the agreement between modeled and observed glacier discharge (e.g. Hock and Noetzli, 1997; Klok *et al.*, 2001), or by recession analysis (e.g. Gurnell, 1993) discussed below. Values used in some melt-runoff modeling studies are summarized in Table 1.

Models are validated primarily by comparing modeled and observed hydrographs and usually expressing agreement by the efficiency criterion R^2 ($-\infty$ to 1; Nash and Sutcliffe, 1970). Despite their simplicity, conceptual models are known to be robust and powerful tools for daily and hourly glacier melt-runoff modeling, often yielding $R^2 > 0.8$. Disagreement between simulations and observations often occurs at the onset and end of the ablation season and during peak (flood) flows (e.g. Moore, 1993; Hock and Noetzli, 1997; Escher-Vetter, 2000). The former is partly attributed to difficulties in modeling whether precipitation falls as rain or snow, while the latter can result from uncertainties in precipitation data, unreliable discharge measurements under floodlike conditions or neglect of variability of

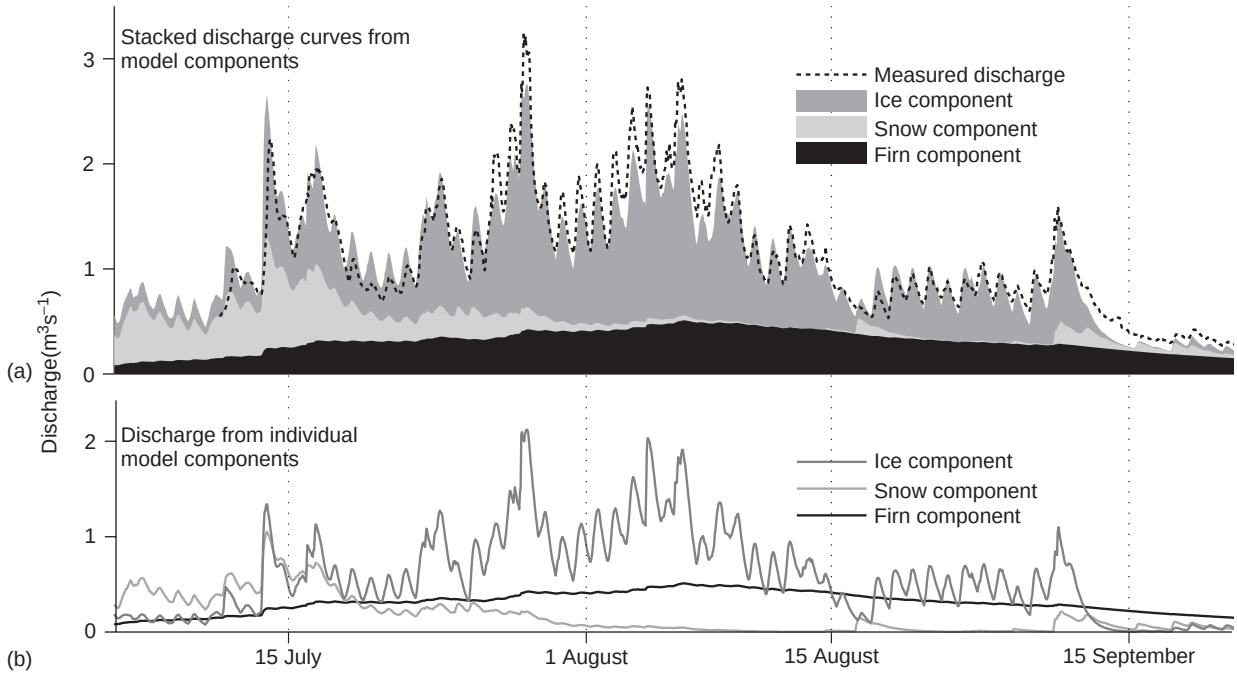


Figure 2 (a) Simulated and measured hourly discharge at Storglaciären 1994 applying three parallel linear reservoirs with storage constants $k_{firn} = 350$ h, $k_{snow} = 30$ h, $k_{ice} = 16$ h. The shaded areas denote the stacked contributions of the firn, snow, and ice reservoirs to total discharge. Contributions from the snow reservoir decline as the snowline retreats. Melt is computed from a distributed energy balance model. (b) Simulated ice, firn, and snow reservoir discharges (Reprinted from Jansson *et al.*, 2003, The concept of glacier storage: a review, *Journal of Hydrology* **282**:116–129. © 2003, with permission from Elsevier)

Table 1 Storage constants used in various melt-runoff modeling studies on glaciers, applying the concept of three parallel linear reservoirs for water routing through the glacier. Storage constants, k, are given in hours (h) for the firn reservoir (slow), snow (intermediate) and ice reservoir (fast)

Site (basin size, glacierization)	k_{firn} (h)	k_{snow} (h)	k_{ice} (h)	Source	Reference
Storglaciären[a] (4.4 km[b], 70%)	350	30	16	Tuning	Hock and Noetzli (1997)
Vernagtferner[b] (11.1 km[b], 81%)	430	30	4	Recession analysis	Baker *et al.* (1982)
Vernagtferner[b] (11.1 km[b], 81%)	430	40–80	6	Tuning	Escher-Vetter (2000)
Rhoneglacier at Gletsch[a] (38.9 km[b], 48%)	350	120	45	Tuning	Klok *et al.* (2001)

[a]Delineation of reservoirs as given by Hock and Noetzli (1997).
[b]The firn reservoir is defined as the uppermost ca. 10% of the firn area. The remaining firn area is classified as snow reservoir.

the internal drainage system in model formulation. Despite good model performance regarding discharge, good fits can be achieved for the wrong physical reasons, leading to false impressions of model accuracy (Burlando *et al.*, 2002). An underestimation of snow precipitation can be compensated for by excess ice melt and vice versa (Braun and Aellen, 1990). Hence, models are preferably tested against additional data, such as glacier mass balance or snow line retreat, the latter greatly facilitated by advances in remote sensing techniques (e.g. Turpin *et al.*, 1997). Willis *et al.* (2002) test performance of their distributed glacier hydrology model against hourly point measurements of ablation and supraglacial discharge measurements. Internal model validation of intermediate results will gain importance as melt-runoff models are used for simulation of variables other than total streamflow.

Recession Analysis

Recession analysis is based on hydrograph-recession curves allowing storage constants, k, to be estimated from the slope of a semilogarithmic plot of discharge versus time (Figure 3). Recessions from a linear reservoir will conform as straight lines since outflow Q_t at time t during periods of no recharge is given by

$$Q_t = Q_0 e^{-(t-t_0)/k} \quad (5)$$

where Q_0 is preceding recession flow at time t_0. Identification of more than one linear component is interpreted as recessions from different reservoirs, thus allowing the number of reservoirs that are operating to be determined, although such interpretation is questionable if reservoirs are partly or entirely linked in series (Gurnell, 1993).

Analyzing recession limbs at Haut Glacier d'Arolla (glacier area 6.3 km^2), Gurnell (1993) identified four reservoirs, but recession coefficients varied as a function of discharge and time of the season. For the same glacier, Richards *et al.* (1996) inferred three reservoirs with approximate average reservoir constants of 12, 27, and 72 h, but emphasized nonlinear behavior of the reservoirs. Hannah and Gurnell (2001) identified two linear reservoirs for a small cirque glacier in the Pyrénées (glacier area 0.2 km^2).

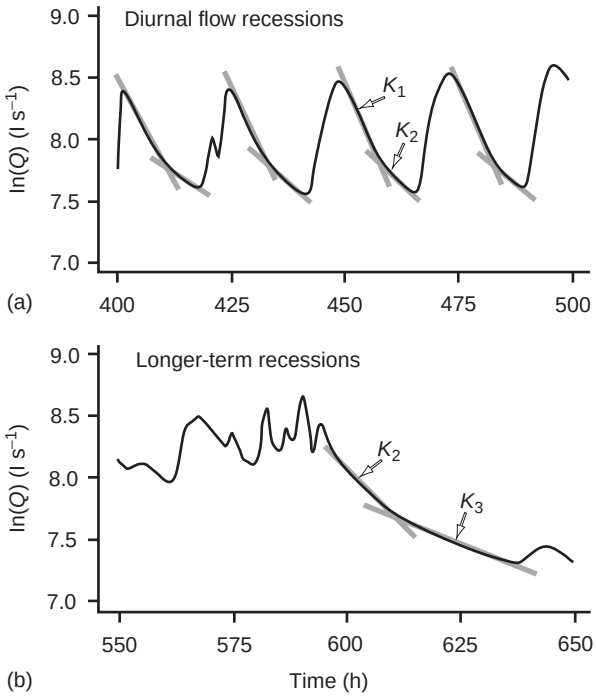


Figure 3 Semilogarithmic recession curves for diurnal and longer term discharge variations (Q), illustrating the separation of flow components from different reservoirs (modified after Gurnell, 1993) (Reproduced with modifications from Gurnell, 1993, by permission of the International Glaciological Society)

Storage coefficients from the "fast" reservoir declined from ~13 to 5 h and for the "slow" reservoir from ~45 to 19 h over the melt season. Later in the season, the drainage system was best represented as a single reservoir. Although recession analysis assuming linear reservoirs is recognized as useful to provide insight into glacier hydrological processes (Nilsson and Sundblad, 1975; Hannah and Gurnell, 2001), Gurnell (1993) suggests that use of a single but nonlinear reservoir might be more adequate for operational modeling purposes.

It is obvious that much uncertainty remains as to the use of recession analysis to provide input to conceptual glacier-runoff modeling. Despite drawbacks of linear

reservoir models to account for observed nonlinearity of storage coefficients, linear reservoir models have generally performed well (Figure 2). This might suggest that such models are sufficiently robust while accurate estimation of melt water input, including modeling the aggregational state of precipitation are more important for hydrograph modeling than deficiencies arising from the assumption of linear reservoirs in subsequent water routing. The differences in hydraulic properties between the "faster" ice reservoir and the "slower" snow/firn reservoirs are probably sufficiently pronounced that effects from seasonal evolution of drainage efficiency are subdued.

PHYSICALLY BASED MODELS

Modeling the physics of glacier hydrology is complex since it involves the liquid phase, namely water, moving through the solid phase, namely ice, at the melting temperature. Furthermore, ice is deformable under relatively low stresses, which allows channels and conduits in glaciers to change size and shape much more rapidly than channels eroded in rock or sediment. The glacier system is thus transient on all timescales and over all spatial scales.

The first physically based models in glaciology were made to reproduce simple observable phenomena in glacier hydrology. Röthlisberger (1972) and Shreve (1972) presented simple numerical models (Röthlisberger: conduit flow; Shreve: potential flow) for water flow in glaciers. The assumptions made were that the rate of melt-enlargement is balanced by creep closure (Röthlisberger) and that water pressure was in balance with ice pressure (Shreve). Nye (1976) and Spring and Hutter (1981) presented a physically based theory of unsteady flow through intraglacial channels. Hooke (1984) investigated the effects of nonsteady conditions, in particular, open channel flow, on the Röthlisberger system. This was elaborated by Kohler (1995) who investigated the ratio of open-to-filled channel flow in a subglacial system. Most models have assumed circular cross-section pipe flow. However, discrepancies between measurements and model calculations led Hooke *et al.* (1990) to propose low broad tunnels in place of the semicircular types in order to be able to apply Glen's flow law deformation with reasonable viscosity parameter values. Conduit flow was found not to reproduce water pressure and fluxes observed during the surge of Variegated Glacier. This prompted Kamb (1987) to propose the linked-cavity system which combined large cross section with small flow velocities. Walder and Fowler (1994) extended Röthlisberger's model to include "canals" cut into deformable till.

A first approach to produce a complete physically based model for a glacier was reported in Arnold *et al.* (1998; henceforth referred to as the Arnold model). Their model contains three components. A surface energy balance model calculates distributed glacier surface melt water production. A routing model takes rain- and meltwater and either delivers it to *a priori* prescribed moulins, where it enters the glacier interior, or supraglacially to the ice margin. The third and final component is a subglacial model which involves a conduit system. The englacial drainage is treated as part of the subglacial conduit system and simulated analogous to a sewage pipe system. Subglacial conduits are fed by a network of "drains", which represent moulins where water can enter or overflow from the system. Subglacial conduits can enlarge and contract in response to changes in the rates of wall melting and creep closure associated to changes in water inputs. In addition, the configuration of the system can change between a distributed system and a channelized system represented by different predefined number and geometries of conduits. "Distributed" links were changed to "channelized" flowpaths as the modeled snowline passed each moulin. From this perspective, the subglacial model is rigid since switching between both systems does not permit smooth transitions with the bed being partially drained by each system. The model was applied to Haut Glacier d'Arolla, Switzerland, and performed well in comparison with proglacial stream discharge, but limitations are indicated by discrepancies between model outputs and field observations of subglacial water pressure and water velocities, although the substantial features of these records could be reproduced.

Flowers and Clarke (2002a; 2002b; henceforth referred to as the Flowers model) developed a different approach to a complete physically based model. The model comprises coupled surface runoff, englacial, subglacial, and groundwater systems. Each of the four components is represented as a two-dimensional, vertically integrated layer that communicates with its neighbors through water exchange (Figure 4). Melt is computed by a distributed temperature

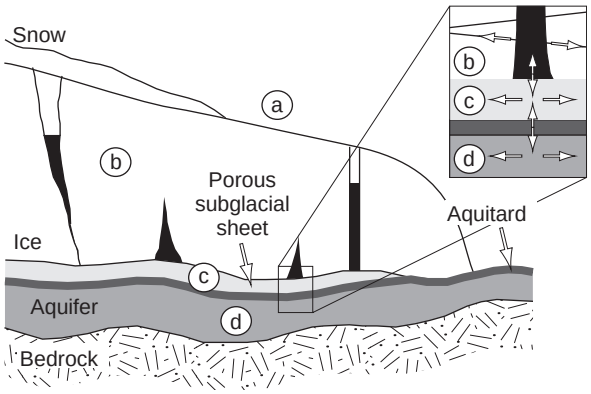


Figure 4 Schematic presentation of the glacier hydrology components in the Flowers model. Circled letters refer to (a) surface ablation and runoff, (b) englacial storage and transport, (c) flow in a porous subglacial layer and (d) subsurface groundwater flow (Reproduced with modifications from Flowers and Clarke, 2002a, by permission of the American Geophysical Union)

index model. Englacial hydrology is represented by describing a variety of bulk storage elements and allowing water transport between them in a system of cracks. Hence, the englacial system is treated as fracture-connected crevasses and pipes. The Flowers model was built to model conditions at Trapridge Glacier in Canada which is underlain by thick porous sediments. Their subglacial model therefore comprises flow through macroporous glacier sediments at the ice-bed interface and subsurface aquifers of buried sediment layers not directly exposed to the base of the glacier. Hence there are no explicit tunnels as in the Arnold model. The model reproduced well diurnal cycles of subglacial water pressure as measured in boreholes.

The present state of physically based models is still crude, in that these models have been tailored to specific glaciers or glacier-types and are not necessarily easily and generally transferable. Melt water production is handled comparatively well; it is probably also the simplest part to model since the entire system is easily observable. Surface energy balance models and temperature index melt models are readily available and have proven to perform very well, even at relatively uncomplicated levels (Hock, 2005; *see also* **Chapter 165, Mass and Energy Balances of Glaciers and Ice Sheets, Volume 4**). The internal and subglacial system is more complicated, characterized by large variability in possible drainage system configurations. Much uncertainty yet remains as to the exact geometry of drainage systems and how these systems can coexist or switch from one state to the other, thus hampering physically based modeling. A crevasselike englacial network, as adopted in the Flowers model, has been observed on Storglaciären, Sweden (Fountain *et al.*, 2005), countering previous notions of englacial water flow through few melt-enlarged conduits. If subglacial sediments are present, groundwater flow may be a significant part of the system, as in the Flowers model, or flow may occur in channels eroded into the subglacial sediments. On harder beds or beds of less permeable sediments such as tills, water flow may occur in conduit systems melted into the ice or in linked-cavity systems. Hence, a general physically based model would have to accommodate all these possibilities and even coupled different types of systems beneath different parts of a glacier. Such a model would be inherently complex.

Another issue which is not satisfactorily met is the time-transgressive development of subglacial systems. In the Flowers model, this is not necessary since flow through porous sediments does not involve significant time-dependent changes, except possible changes such as development of piping or siltation of the porous media from fines produced through sediment deformation. In the case of the Arnold model, it is widely reported (e.g. Nienow *et al.*, 1998) that the subglacial drainage system changes both in sizes of conduits and in complexity of the network of channels through the course of a season. The subglacial system is very dynamic and it seems as if a complete and accurate model description of this system may be distant. However, most changes in such a system occur in response to rapid changes, both increases or decreases in water inputs, so a first-order approximation may be to switch between a series of systems prompted by key events in the forcing.

CONCLUDING REMARKS

An increasing practical and scientific need to model glacier water resources in light of globally increasing freshwater demand and potential impacts of climate change has fostered a large variety of glacier hydrology models varying greatly in scope and sophistication. A fundamental question concerns the level of complexity which models should employ. Stochastic models, simple but of limited value if transferred to different areas or climate conditions, have often been replaced by more process-orientated models following considerable advancements in computer power and process-understanding over the last few decades. Sophisticated physically based models including surficial, englacial, and subglacial components and their variability in space and time are hampered by complexity and large variability of the drainage system. Such models most accurately describe the processes involved, but limits are set by availability, accuracy, and representativeness of input data and also by spatial variability of processes. Although latest attempts are tailored to specific sites (Arnold *et al.*, 1998; Flowers and Clarke, 2002), and thus may lack direct transferability to other glaciers, physically based models are important research tools for process studies and should contribute to advancing simpler conceptual models.

Lumped conceptual models will most certainly retain their prominent position in operational glacier-runoff modeling. Commonly, the concept of linear reservoirs is invoked to route water through the glacier. Despite moderate data requirements and simplification of processes involved, such as neglect of nonlinearity of storage coefficients, such models have proven to provide robust tools for predictive purposes. Many widely used runoff models (Singh, 1995) include some kind of glacier melt routine, but neglect explicit consideration of the presence of glaciers in runoff-routing. Considering the large effect of glaciers on basin hydrology even at glacierization levels of only a few percent due to the specific characteristics of glacier drainage (Hock *et al.*, 2005), glacier-runoff models need to include specific routines for internal glacier drainage. This is necessary in particular if long- and short-term effects of climate change on glacier hydrology are to be assessed since partitioning of "slow" and "fast" flow components will change in response to changes in snow and firn cover (Braun *et al.*, 2000). Much effort has been devoted to glacier melt modeling, while relatively little development of conceptual models of internal flow routing has occurred. Further

research should explore possibilities to consider nonlinear reservoirs for flow routing, while retaining low data input requirements.

Generally speaking, both stochastic and conceptual models have been used widely in operational forecasting of glacier runoff, while physically based models are yet sparse and have generally been tailored to scientific interests and specific glaciers. Merging both strategies – aiming at robust and easy-to-operate conceptual models while enhancing their physical base to better represent the large spatial and temporal variability in glacier hydrology – provides the challenge for future model developments and quantitative assessment of possible future evolution of glacial water resources.

Acknowledgments

Gwenn Flowers is gratefully acknowledged for providing original digital files for Figure 4 and for carefully scrutinizing the manuscript. Thomas Schuler and two anonymous reviewers made useful comments on the paper.

REFERENCES

Arnold N., Richards K., Willis I. and Sharp M. (1998) Initial results from a distributed, physically based model of glacier hydrology. *Hydrological Processes*, **12**(2), 191–219.

Baker D., Escher-Vetter H., Moser H., Oerter H. and Reinwarth O. (1982) A glacier discharge model based on results from field studies of energy balance, water storage and flow. In *Hydrological Aspects of Mountain Areas*, Glen J. (Ed.), IAHS Publication No. 138, IAHS: pp. 103–112.

Braun L.N. and Aellen M. (1990) Modelling discharge of glacierized basins assisted by direct measurements of glacier mass balance. In *Hydrology of Mountainous Regions. I*, Lang H. and Musy A. (Eds.), IAHS Publication No. 193, IAHS: pp. 99–106.

Braun L.N., Weber M. and Schulz M. (2000) Consequences of climate change for runoff from Alpine regions. *Annals of Glaciology*, **31**, 19–25.

Burlando P., Pellicciotti F. and Strasser U. (2002) Modelling mountainous water systems between learning and speculating looking for challenges. *Nordic Hydrology*, **33**(1), 47–74.

Chow V., Maidment D.R. and Mays L.W. (1988) *Applied Hydrology*, *Civil Engineering Series*, McGraw-Hill International Editions: New York, p. 572.

Escher-Vetter H. (2000) Modelling meltwater production with a distributed energy balance method and runoff using a linear reservoir approach – results from Vernagtferner, Oetztal Alps, for the ablation seasons 1992 to 1995. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **36**, 119–150.

Escher-Vetter H. and Reinwarth O. (1994) Two decades of runoff measurements (1974 to 1993) at the Pegelstation Vernagtbach/Oetztal Alps. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **30**, 53–98.

Flowers G.E. and Clarke G.K.C. (2002a) A multicomponent coupled model of glacier hydrology. 1. Theory and synthetic examples. *Journal of Geophysical Research*, **107B**(11), 2287, doi:10.1029/2001JB001122.

Flowers G.E. and Clarke G.K.C. (2002b) A multicomponent coupled model of glacier hydrology. 2. Application to Trapridge Glacier, Yukon, Canada. *Journal of Geophysical Research*, **107B**(11), 2288, doi:10.1029/2001JB001124.

Fountain A.G. and Tangborn W.V. (1985) Overview of contemporary techniques. In *Techniques for Prediction of Runoff from Glacierized Areas*, Young G. (Ed.), IAHS Publication No. 149, IAHS: pp. 1–41.

Fountain A.G. and Walder J.S. (1998) Water flow through temperate glaciers. *Reviews of Geophysics*, **36**(3), 299–328.

Fountain A.G., Jacobel R.W., Schlichting R.B. and Jansson P. (2005) Fractures as the main pathways of water flow in temperate glaciers. *Nature*, **433**(7026), 618–621.

Gottlieb L. (1980) Development and applications of a runoff model for snowcovered and glacierized basins. *Nordic Hydrology*, **11**, 255–272.

Gurnell A.M. (1993) How many reservoirs? An analysis of flow recessions from a glacier basin. *Journal of Glaciology*, **39**, 409–414.

Hannah D. and Gurnell A.M. (2001) A conceptual, linear reservoir runoff model to investigate melt season changes in cirque glacier hydrology. *Journal of Hydrology*, **246**, 123–141.

Hock R. (2003) Temperature index melt modelling in mountain regions. *Journal of Hydrology*, **282**(1–4), 104–115. doi:10.1016/S0022-1694(03)00257-9.

Hock R. (2005) Glacier melt: A review on processes and their modelling. *Progress in Physical Geography*, **29**(4), in press.

Hock R. and Noetzli C.h (1997) Areal mass balance and discharge modelling of Storglaciären, Sweden. *Annals of Glaciology*, **24**, 211–217.

Hock R., Jansson P. and Braun L. (2005) Modelling the response of mountain glacier discharge to climate warming. In *Global Change and Mountain Regions – A State of Knowledge Overview*, *Advances in Global Change Series*, Huber U.M., Reasoner M.A. and Bugmann H. (Eds.), Springer: Dordrecht, 243–252.

Hooke R.LeB. (1984) On the role of mechanical energy in maintaining subglacial water conduits at atmospheric pressure. *Journal of Glaciology*, **30**(105), 180–187.

Hooke R.LeB., Laumann T. and Kohler J. (1990) Subglacial water pressures and the shape of subglacial conduits. *Journal of Glaciology*, **36**(122), 67–71.

Hubbard B. and Nienow P. (1997) Alpine subglacial hydrology. *Quaternary Science Reviews*, **16**, 939–955.

IPCC (Intergovernmental Panel on Climate Change) (2001) In *Climate Change 2001: The Scientific Basis*, Houghton J.T., Ding Y., Griggs D.J., Noguer M., Van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: p. 881.

Jansson P., Hock R. and Schneider T. (2003) The concept of glacier water storage – a review. *Journal of Hydrology*, **282**(1–4), 116–129. doi:10.1016/S0022-1694(03)00258-0.

Jensen H. and Lang H. (1973) Forecasting discharge from a glaciated basin in the Swiss Alps. *Role of Snow and*

Ice in Hydrology, IAHS Publication No. 107(2), IAHS: pp. 1047–1054.

Kachroo R.K. (1992) River flow forecasting. Part 1. A discussion of the principles. *Journal of Hydrology*, **133**, 1–15.

Kamb B. (1987) Glacier surge mechanism based on linked cavity configuration of the basal water conduit system. *Journal of Geophysical Research*, **92**(B9), 9083–9100.

Klok E.J., Jasper K., Roelofsma K.P., Gurtz J. and Badoux A. (2001) Distributed hydrological modelling of a heavily glaciated Alpine river basin. *Hydrological Sciences Journal*, **46**(4), 553–570.

Kohler J. (1995) Determining the extent of pressurized flow beneath Storglaciären, Sweden, using results of tracer experiments and measurements of input and output discharge. *Journal of Glaciology*, **41**(138), 217–231.

Lang H. (1968) Relations between glacier runoff and meteorological factors observed on and outside the glacier. *Snow and Ice. Reports and Discussions*, IAHS Publication No. 79, IAHS: pp. 429–439.

Moore R.D. (1993) Application of a conceptual streamflow model in a glacierized drainage basin. *Journal of Hydrology*, **150**, 151–168.

Nash J.E. and Sutcliffe J.V. (1970) River flow forecasting through conceptual models. Part I – A discussion of principles. *Journal of Hydrology*, **10**(3), 282–290.

Nienow P., Sharp M. and Willis I. (1998) Seasonal changes in the morphology of the subglacial drainage system, Haut Glacier d'Arolla, Switzerland. *Earth Surface Processes and Landforms*, **23**, 825–843.

Nilsson J. and Sundblad B. (1975) The internal drainage of Storglaciären and Isfallsglaciären described by an autoregressive model. *Geografiska Annaler*, **57A**(1–2), 73–98.

Nye J.F. (1976) Water flow in glaciers: Jökulhlaups, tunnels and veins. *Journal of Glaciology*, **17**(76), 181–207.

Østrem G. (1973) Runoff forecasts for highly glacierized basins. *Role of Snow and Ice in Hydrology*, IAHS Publication No. 107(2), IAHS: pp. 1111–1132.

Quick M.C. and Pipes A. (1977) U.B.C. watershed model. *Hydrological Sciences Bulletin*, **22**(1), 153–161.

Richards K., Sharp M., Arnold N., Gurnell A., Clark M., Tranter M., Nienow P., Brown G., Willis I. and Lawson W. (1996) An integrated approach to modelling hydrology and water quality in glacierized catchments. *Hydrological Processes*, **10**(4), 479–508.

Röthlisberger H. (1972) Water pressure in intra- and subglacial channels. *Journal of Glaciology*, **11**(62), 177–203.

Röthlisberger H. and Lang H. (1987) Glacial hydrology. In *Glacio-Fluvial Sediment Transfer: An Alpine Perspective*, Gurnell A.M. and Clark M.J. (Eds.), John Wiley & Sons: pp. 207–284.

Schneider T. (2000) Hydrological processes in the wet-snow zone of glaciers – a review. *Zeitschrift für Gletscherkunde und Glazialgeologie*, **36**(1), 89–105.

Shreve R.P. (1972) Movement of water in glaciers. *Journal of Glaciology*, **11**(62), 205–214.

Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Water Resources Publications: Colorado. ISBN-Number 0-918334-91-8.

Spring U. and Hutter K. (1981) Numerical studies of Jökulhlaups. *Cold Regions Science and Technology*, **4**, 227–244.

Tangborn W.V. (1984) Prediction of glacier-derived runoff for hydroelectric development. *Geografiska Annaler*, **66A**(3), 257–265.

Turpin O.C., Ferguson R.I. and Clark C.D. (1997) Remote sensing of snowline rise as an aid to testing and calibrating a glacier runoff model. *Physics and Chemistry of the Earth*, **22**(3–4), 279–283.

Van de Wal R.S.W. and Russell A.J. (1994) A comparison of energy balance calculations, measured ablation and meltwater runoff near Søndre Strømfjord, West Greenland. *Global and Planetary Change*, **9**(1/2), 29–38.

Walder J.S. and Fowler A. (1994) Channelized subglacial drainage over a deformable bed. *Journal of Glaciology*, **40**(134), 3–15.

Willis I.C., Arnold N.S. and Brock B.W. (2002) Effect of snowpack removal on energy balance, melt and runoff in a small supraglacial catchment. *Hydrological Processes*, **16**, 2721–2749.

171: River-Ice Hydrology

TERRY D PROWSE

Department of Geography, National Water Research Institute of Environment Canada, University of Victoria, Victoria, BC, Canada

The hydrology of almost 60% of rivers in the Northern Hemisphere is affected by ice for at least a portion of the winter. At the higher latitudes, river ice is the major controller of most fluvial processes. Although traditionally the pursuit of hydraulic engineers, hydrologic studies of river ice have broadened greatly over the last decade. Initially, this was driven by recognition that many hydrologic extremes produced by river ice, such as floods and low flows, exceed those of the more intensively studied, open-water period. Added recognition that river ice strongly influenced other fluvial disciplines, such as geomorphology and aquatic ecology, increased its scientific visibility. This manuscript reviews the basic physics of river ice; evaluates its effect on various hydrologic processes and events from freeze-up to breakup; and discusses its extended role in affecting sediment transport/erosion, river morphology and a number of key ecological processes including geochemical mixing and habitat modification. Arguments are made for increased focus on river-ice hydrology given its significance to physical and ecological processes, and the growing need to understand the effects of climate change and flow regulation on cold regions rivers.

INTRODUCTION

Ice is a dominant feature of the hydrologic cycle in the near-polar latitudes and at higher elevations. For example, almost 25% of the land-mass or 29% of the total river length in the Northern Hemisphere are found north of the mean-annual 0 °C isotherm (Bennett and Prowse, 2004) and thus experience significant ice effects. Notably, this area includes the downstream reaches of four of the world's 15 largest rivers (by discharge or drainage area), including the Lena, Hob, Mackenzie, and Yenisei. For almost half the year, the hydrologic conditions in rivers of this immense area are influenced by the formation, growth and breakup of freshwater ice. Considering the total hydrologic zone experiencing even transient ice effects (i.e. climates where the mean January air temperature is <0 °C), the land-mass and river-length percentages of the Northern Hemisphere rise to 53 and 57, respectively. Moreover, this zone extends as far south as 33 °N in North America and 26 °N in Eurasia, and encompasses all of the above four large Arctic rivers and increases the listing from the top 15 rivers to include the Mississippi, Yangtze, and even the Ganges, which have cold regions headwaters.

Traditionally, the study of river ice has been the domain of hydraulic engineers focused largely on finding solutions to localized river-development problems. Over the last two decades, however, the subfield of river-ice hydrology has emerged (Beltaos, 2000; Morse and Hicks, 2004). To a large part, this has been driven by the recognition that in-channel ice effects are frequently more important than landscape processes, or even open-water flow, in producing hydrologic extremes such as low flows and floods (e.g. Gerard, 1990; Prowse, 1994; Prowse and Beltaos, 2002) and related, costly infrastructure damage (e.g. Gerard and Davar, 1995). Given the importance of river ice to water levels and discharge, there has been further recognition that river ice controls many related fluvial processes, such as sediment transport/deposition, bio-chemical exchanges and aquatic productivity/diversity (e.g. Prowse, 2001a,b).

Achieving an understanding of river-ice hydrology requires some additional scientific fundamentals that are not normally a part of hydrologic studies of open-water conditions, such as mechanics and thermodynamics. To provide an introduction to the field, the next section of this chapter reviews basic river-ice physics, explaining how ice forms, grows, decays, and breaks up in fluvial systems.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

For additional background about river-ice hydraulics and physics, interested readers are referred to other texts, such as Ashton (1986), Beltaos (1995) and Davar *et al.* (1996; which includes worked examples) and Gray and Prowse (1993; which focuses on hydrologic issues and discusses problems of ice-affected hydrometric measurements). The next section considers the major stage, storage, and flow extremes that are produced by river ice from first freeze-over in the autumn to final ice clearance in the spring. Additional subsections review the effects of river ice on fluvial geomorphology and aquatic ecology.

RIVER-ICE PROCESSES

Autumn Cooling

Unlike on lakes, flowing water in rivers is constantly mixed and, in the absence of thermal stratification, the entire water column cools during the autumn months at approximately the same rate. Changes in water temperature are controlled by the net heat flux to the water column (Φ_*), which comprises a number of heat fluxes that occur at the water surface and riverbed, and within the water column:

$$\phi_* = \phi_S + \phi_L + \phi_H + \phi_E + \phi_P + \phi_G + \phi_B + \phi_F \quad (1)$$

where Φ_S and Φ_L are the net fluxes of short- and long-wave radiation; Φ_H, $\Phi_{E,}$ and Φ_P are the sensible, latent, and precipitation heat fluxes from the overlying atmosphere; Φ_G and Φ_B are the groundwater and bed (geothermal and sediment) heat fluxes, and Φ_F is the heat flux to the water from flow friction (positive values representing a heat flux to the water). Under open-water conditions, the last two terms are relatively minor, although heat supplied from the bed can become important once an ice cover eliminates direct water column-atmosphere heat exchanges. The magnitude of Φ_B declines as the heat stored in the bed during the warmer months is gradually conducted to the river. The significance of Φ_G as a major heat source depends on the temperature of the groundwater and its flow rate relative to the river discharge. Groundwater is sometimes the major source of baseflow and can be responsible for keeping river reaches open throughout the winter, even at high northern latitudes (e.g. Prowse, 2001b).

Shortwave radiation supplies most heat prior to freeze-over, although much of it can be reflected away by surface fog that frequently forms when air temperatures are significantly below those of the water. Most cooling of the water column results from negative values of Φ_L (negative long-wave radiation budget), Φ_H (convection of heat upwards to the atmosphere) and Φ_E (evaporation) – the strength of all three controlled by the water–air temperature difference and, for the latter two, windspeed.

The temperature of the precipitation relative to that of the river water and whether it is in liquid or solid form determine the magnitude and direction of Φ_P. Snow is most effective in cooling river water because of the heat required to raise its temperature and, more importantly, the large amounts of latent heat required to melt it.

Efficiency of river cooling depends on surface area to volume ratios. Thus, for similar discharge and meteorological conditions, wide shallow rivers cool most rapidly and deep narrow ones relatively slowly. Small tributary streams also cool more quickly than larger rivers carrying greater flow. Assuming that cooling is sufficient for the water column to reach 0 °C and permit subsequent ice formation, the same sequence among rivers of varying geometry and size characterizes the freeze-up pattern. Such contrasts can cause, for example, freeze-up delays of up to a month or more between small and large rivers in the same climatic region.

Ice Production

Once water temperatures reach 0 °C, a river is capable of producing a variety of ice forms. As on lakes, the first ice to form is usually "border ice" along the banks and around islands or other material (e.g. large boulders) protruding above the water level. Ice develops here because the bank material is usually colder than the main river flow and effectively chills water in close contact with it. The lateral growth of border ice, however, depends on flow turbulence and the rate of cooling at the growing edge. In pool sections, for example, quiescent flow conditions combined with a high rate of cooling can permit border ice to cover the entire river width. Under similar conditions, large moving ice sheets can develop from vertical ice growth without attachment to the border ice, such as on the St. Lawrence River where sheets several hundred meters across have been observed (e.g. Marcotte, 1984).

Quite a different type of river ice is formed within the more turbulent portions of a river, and is generally referred to as "frazil ice", a word derived from the old French word "fraisil" meaning an accumulation of cinders (Michel, 1981) – a suitably descriptive term for the initial stages of frazil generation. Frazil ice particles nucleate when the water experiences a slight degree of supercooling (typically < -0.05 °C) and can cause further rapid multiplication of additional frazil particles with concentrations reaching as high as 10^6 m^{-3} (Gilfilian *et al.*, 1972). Although having an initial discoid shape, they become more angular as their size increases to typical diameters of 0.1 to 5 mm. Under intense water-to-air heat loss, frazil may be generated during the daytime but, because of warming from solar radiation, it often follows a diurnal cycle characterized by rapid growth at night and cessation during the day.

While frazil particles remain supercooled, they may adhere to and coat other objects within the flow (e.g.

trash racks at hydroelectric plants) or on the riverbed (e.g. macrophytes, rocks, and boulders). Rarely do they coat fine-grained material because they are too easily buoyed by even small amounts of ice and/or because the stronger bed heat flow to such fine-grained smooth surfaces prevents a strong ice-bed bond. In very turbulent conditions, even supercooled water can be carried to the bed, where it nucleates and adheres to the bed. Both of these bed-ice forms are termed "anchor ice", the nucleation-anchor ice having the smoother surface of the two (Tsang, 1982). Where flow and cold conditions permit continued anchor-ice growth, substantial accumulations can develop. These will either ablate during warming periods or release, either when they are sufficient to buoy the bed material or when they suddenly lose their bond during a period of rapid warming. Marcotte (1984), for example, reports large blocks 6 m in length being released from a 20-km^2 zone of the St. Lawrence River, where anchor ice grows up to 1.5 m thick.

Frazil crystals that remain within the main flow gradually grow and agglomerate into small clusters and larger flocs (5–100 mm). Eventually, they reach sufficient size and buoyancy that they periodically float to the surface where they are exposed to direct surface cooling, begin to grow and become even more buoyant. Accumulation of large amounts of frazil flocs at the surface is referred to as "frazil slush". Such material also tends to adhere to border ice, successive "buttering" of the edges producing bands that demark cycles in frazil-ice production. Eventually, interstitial spaces in the frazil slush become sufficiently frozen that they take on a more rigid surface form called "frazil pans".

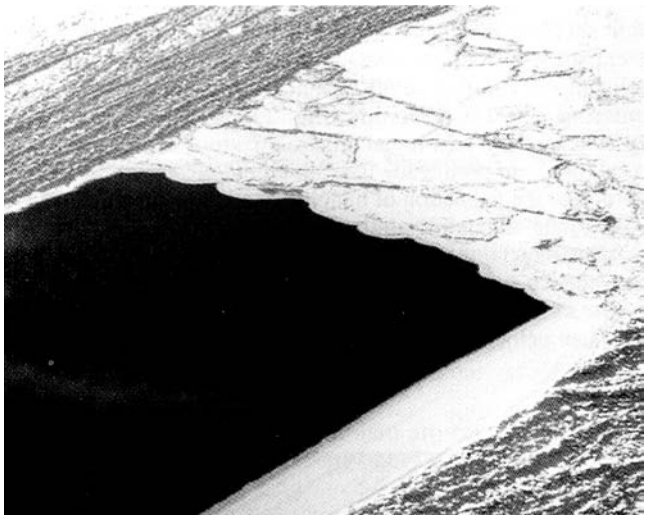
Swept around by surface currents and ground against one another, the frazil pans gradually take on a circular form and are referred to as "pancake ice" (Figure 1a). As these frazil products become more congested and in contact with one another for long enough periods, they freeze-bond into large frazil floes. Although the above has described frazil evolution from initial crystal to large floe, all forms can be present within a reach at the same time.

Highly turbulent sections, such as rapids, can generate tremendous amounts of frazil ice; much more than would be produced if the section developed an *in situ* ice cover. An estimate of the total potential ice production in an open reach can be obtained by simply integrating the water surface to atmosphere heat loss over time (when the water temperature = 0 °C):

$$V_i = \frac{1}{\rho_i \lambda_i} \int_{t_2}^{t_1} A_o \phi_* \, dt \qquad (2)$$

where V_i is the volumetric production of ice (m^3 s^{-1}), ρ_i is the density of ice ($\sim$920 kg m^{-3}), λ_i is the latent heat of fusion for ice (3.34×10^5 J kg^{-1}), A_o is the area of the open-water reach that may also change with time from ice accumulation and border-ice growth (m^2), and t is time (s). This equation assumes that all ice is produced under open-water conditions, such as the typical case for frazil production. It does not account for ice growth at the base of an *in situ* ice cover, which can be much slower because of the insulation to air temperatures created by the overlying ice and snow.

(a)



(b)

Figure 1 (a) Border and pancake ice. Note circular form to frazil pans produced by interfloe contact during travel. (Photograph: North Saskatchewan River, Canada C.R. Onclin); (b) Frazil floes accumulating between border ice. Note layering of border ice. Flow is from left to right. (Photograph: F. Parkinson). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Freeze-up

Eventually, enough ice is produced and transported along a river that floes become sufficiently large and congested to bridge across the river and obstruct the downstream passage of other upstream ice (Figure 1b; see, e.g. Shen (2000) for a review of ice transport theories). Bridging locations are difficult to predict but reach geometry and flow characteristics that favor congestion, such as a narrowing reach or a concentration of bridge piers, are particularly susceptible. Whether additional upstream ice will accumulate behind, or pass beneath the arched floes depends primarily on ice thickness, flow depth, and river velocity. Flow velocities greater than $2\,\mathrm{m\,s^{-1}}$, for example, are required to submerge and entrain large ice pans (Michel, 1971), whereas much lower velocities are sufficient to transport less-developed frazil forms.

Travelling beneath an ice cover, transported ice will melt, reemerge at a downstream open-water section, or deposit underneath where local flow conditions reduce transport capacity. Large accumulations of such deposited ice are referred to as hanging dams. Common locations for their formation include deep pools adjoining upstream rapids and confluences of rivers and lakes. Some massive hanging dams have been reported. On the La Grande River in Canada, for example, one was reported to contain $56 \times 10^6\,\mathrm{m^3}$ of ice over a 16-km long reach (Tsang, 1982). Such large obstructions in the channel cross-section can concentrate flow and cause localized bed erosion. During periods of rising discharge, particularly during the spring, they can also create serious backwater flooding problems before being flushed downstream or thermally eroded.

Winter Ice Growth

Initially, the surface layer of river ice will comprise a myriad of accumulating and/or compressed frazil forms. Subsequent growth results from vertical downward freezing of the underlying water column and the incorporation of any frazil ice deposited along the ice bottom. In the absence of frazil, the ice cover can be highly transparent, similar to the "black" ice (or blue ice having a large-grained columnar structure, often transparent) that is typical of lake ice (Adams, 1981; Prowse, 1995). The rate of growth depends on heat exchanges at both the ice-atmosphere and ice-water boundaries:

$$\frac{\mathrm{d}h_\mathrm{i}}{\mathrm{d}t}\rho_\mathrm{i}\lambda_\mathrm{i} = \frac{(T_\mathrm{b} - T_\mathrm{a})}{(t_\mathrm{i}/k_\mathrm{i}) + (t_\mathrm{s}/K_\mathrm{s})(1/C_\mathrm{a})} - C_\mathrm{w}(T_\mathrm{w} - T_\mathrm{b}) \quad (3)$$

where t_i and t_s are ice and snow thicknesses (m), k_i and k_s are the thermal conductivities of ice ($2.2\,\mathrm{W\,m^{-1}/°C}$) and snow (0.1 to $0.5\,\mathrm{W\,m^{-1}/°C}$ for snow densities of 150 to $500\,\mathrm{kg\,m^{-3}}$); T_a, T_b and T_w are the air, basal ice and water temperatures (°C); and, C_a and C_w are the heat transfer coefficients from ice to air and water to ice ($\mathrm{W\,m^{-2}/°C}$). The latter two coefficients are highly variable primarily depending on ice roughness and wind and current velocity. See Ashton (1986) and Prowse (1995) for further explanation.

When the ice cover is rapidly growing in near 0 °C-water, the right-hand term in equation (3) – the water-ice heat transfer is often considered to be relatively insignificant. With other relatively minor simplifications (see Prowse, 1995), it is possible to predict ice growth using the Stefan formula that uses a degree-day index for the major heat fluxes:

$$t_\mathrm{i} = \kappa(D_\mathrm{f})^{1/2} \quad (4)$$

where D_f is accumulated freezing degree-days below 0 °C and κ is a coefficient (which has a theoretical maximum of $[2k_\mathrm{i}/\rho_\mathrm{i}\lambda_\mathrm{i}]^{1/2}$ or $0.035\,\mathrm{m/°C^{1/2}\ d^{1/2}}$) that accounts for differing surface insulation as supplied by overlying snow and varying degrees of exposure to atmospheric heat fluxes. Although equation (4) tends to overpredict growth during the initial stages, it gives reasonably reliable results once the cover is greater than approximately 0.1 m. Typical values of κ for a river with snow are $0.014–0.017\,\mathrm{m/°C^{1/2}\ d^{1/2}}$. Insulation provided by snow can dramatically slow ice growth. For relatively thick ice covers, a snow depth equal to approximately half the ice thickness will effectively halt further ice growth (Michel, 1971).

When significant accumulations of snow load an ice cover, it can depress the ice below the hydrostatic water level such that water rises into the overlying snow, "slushes" it and, following a subsequent refreezing of this snow-water mixture, leaves a surface stratum of "white-ice" (or snow-ice having a typical small-grain polycrystalline structure, generally translucent; e.g. Adams, 1981; Prowse, 1995). The small-grain physical structure of this ice is not too dissimilar from many frazil-ice forms. Such ice growth occurs at an accelerated rate since the freezing occurs in more direct contact with the cold atmosphere. Thick layers of frazil deposited at the base of an ice cover can also accelerate total growth since the downward freezing will occur much more rapidly into an ice-water mixture than into water alone (Calkins, 1979; Timco and Goodrich, 1988).

Icings, aufeis, or naleds are another form of river ice but are most typical of the colder subarctic and arctic climates. Somewhat similar to the formation of snow-ice, icings are surficial accumulations formed by the flow of water onto the surface of an ice cover, usually the result of some localized flow resistance in a channel that forces water to the surface or, in the case of smaller streams, from groundwater seepage (e.g. Carey, 1973; Kane, 1981; van Everdingen, 1990). They can accumulate as a widespread sheet of ice or as a localized mound depending on the rate of flow versus surface freezing. In cases where the process continues throughout the winter and where an icing deposit

may be fed by a number of seepage points, entire channels and even floodplains can be coated to thicknesses several times the open-water channel depth or the ice thickness that would occur from normal freezing of the underlying flow. So large are some accumulations that they control the routing of flow during spring melt and sustain flow during the warmer summer months (e.g. Shumskii, 1964; Gray and MacKay, 1979). Icings can create a flood risk where they plug culverts and, by their mass, decrease the flow required to overtop banks.

Ice Breakup

Breakup of a river-ice cover can occur at anytime during the ice-covered season. Within the more temperate climates, the winter period may actually comprise a series of freeze-up/breakup cycles, whereas in colder climates, breakup is typically a spring event. In either case, breakup is initiated when the forces driving breakup, primarily resulting from the spring flood wave produced by snowmelt and sometimes assisted by rainfall, exceed the resisting forces or those operating to keep the ice cover intact, such as ice thickness, internal mechanical strength, and bonds to the bank and bed.

During the prebreakup period, as the snowcover on the landscape begins to ripen and melt, so does the river-ice cover both at the surface and at the bottom. The ice bottom melts when the water-ice heat transfer exceeds the rate at which heat can be conducted upward into the ice sheet – most rapid melt typically occurring after the ice sheet has warmed to a 0 °C isothermal state. Although the temperature of the river water may only be a fraction of a degree above freezing, even moderate flow velocities can cause a large and rapid heat transfer to the ice bottom (Marsh and Prowse, 1987). The rate is also influenced by the bottom-ice roughness, which tends to increase during the prebreakup period from the formation of ice ripples. Heat transfer greater than 50 to 60% has been reported for rippled compared to smooth ice bottoms (Ashton and Kennedy, 1972; Gilpin *et al.*, 1980).

Obtaining relevant data to calculate the combined surface and bottom ablation of a river-ice cover is extremely difficult but moderate success has been achieved by relating total thinning with accumulated melting degree-days (Bilello, 1980):

$$t_i = t_o - \varepsilon\beta_m \tag{5}$$

where t_o is the maximum ice thickness prior to decay (m), β_m is accumulated melting degree-days (–5 °C base temperature for river locations) and ε is a site dependent, empirical constant (for values typical of specific regions see Bilello, 1980; Beltaos, 2005).

Melt at the snow and ice surface is driven by a combination of sensible, latent and radiative heat transfers but it is only shortwave radiation that produces major changes to the mechanical strength of an ice cover. Once the highly reflective surface-ice forms such as surface frazil and snow-ice are ablated, significant amounts of radiation can penetrate the ice sheet where it is attenuated at ice-crystal boundaries producing localized melt (Ashton, 1985; Prowse *et al.*, 1990). As this process continues, the structural integrity of the ice sheet reduces and poses an ever-decreasing resistance to the forces impinging on it by the rising flow.

As flows and stage increase, the ice cover eventually fractures in patterns determined by its spatial heterogeneity in thickness, strength, and bond to the bed and banks. In particular, upward loading by the rising water causes cracks to develop parallel to and in close proximity to the banks (Beltaos, 1997). Eventually, the main ice sheet will detach from the banks completely and an open-water shore lead will exist. With such detachment and a wider channel produced by the rising stage, the ice cover is able to shift within its channel, eventually leading to the formation of transverse cracks (Shulyakovskii, 1972; Beltaos, 1997). These are commonly found at acute changes in channel geometry such as at meander bends (Beltaos, 1984a).

Hydrometeorological conditions leading to the final breakup are typically defined into two general states: over-mature and premature. In the first case, which resembles ice breakup conditions on lakes, the ice cover continues to be thermally and mechanically weakened while river runoff remains relatively small. The low discharge is usually because of a small winter snowpack and/or slow protracted snowmelt. The ice cover is eventually so weakened that it offers minimal resistance to even low flow and is easily swept downstream. This type of event is usually referred to as a "thermal" breakup. For the premature case, conditions are reversed to include a large flood wave being driven into a mechanically strong, largely competent ice cover.

Eventually, the rising discharge becomes sufficient to dislodge, further fracture, and drive the ice cover downstream. This is usually referred to as a "dynamic" or "mechanical" breakup.

Breakup timing at a site is a complex process dependent on a suite of variables including, cover thickness (ice and snow), ice strength and structural variability, river geometry, flow velocity, and stage. The simplest approach used in predicting breakup timing has relied on a single heat index such as accumulated degree-days. Only limited success has been achieved with this approach because it does not treat separately changes in the components that comprise the driving and resisting forces. Another simplified approach has focused on the rise in stage relative to the freeze-up level and the peak ice thickness. It assumes that breakup should not occur until the stage has surpassed the original freeze-up level, at which time the ice cover has been freed from most bed and bank constraints. The stage

at which breakup occurs is then measured relative to the ice thickness. A rise in level equivalent of two to four times the ice thickness is a typical value (e.g. Beltaos, 1984b, 1990; Donchenko, 1975), although this relates to a fairly broad range in water levels and rises in stage are also difficult to predict. A refinement of this approach (Shulyakovskii, 1966; Beltaos, 1990) also considers the decreases in ice thickness and strength that occur during the prebreakup period by including an accumulated heat flux term, usually calculated from local climate-station data. A premature (mature) breakup, for example, would be one with small (large) amounts of prebreakup warming but large (small) rises in the stage. Beltaos (1997) reviews more physically based equations for predicting the initiation of mechanical breakups using variables such as river surface width, channel curvature, freeze-up stage, and ice competence.

ICE EFFECTS ON WATER LEVELS AND DISCHARGE

Ice-covered conditions on rivers create water level and flow conditions quite dissimilar from those for the open-water period. Such conditions begin with the first formation of an ice cover and its elevation of water levels. For steady flow in a wide rectangular channel the water depth, H (m), is given as

$$H = 1.32\left[\frac{Qn_c}{B\sqrt{S}}\right]^{3/5} + \frac{\rho_i t_i}{\rho_w} \tag{6}$$

where Q is discharge ($m^3\ s^{-1}$); B is the cover accumulation width (m); S is water surface slope; and n_c is the composite Manning roughness coefficient. The depth of flow beneath the ice, as derived from the uniform-flow resistance equation, is represented by the first term on the right-hand side of equation (6). The second term represents the submerged portion of the ice cover, assuming free floatation without snow loading. The composite Manning roughness coefficient is given by (e.g. Ashton, 1986)

$$n_c = \left[\frac{n_b^{3/2} + n_i^{3/2}}{2}\right]^{2/3} \tag{7}$$

where n_b and n_i are the hydraulic roughness coefficients for the riverbed and ice bottom respectively. For typical values see Ashton (1986) and Gray and Prowse (1993). Much of the increase in stage under ice conditions results from the increase in wetted perimeter produced by the ice cover. Ice roughness plays an additional role depending on whether it is rougher or smoother than that of the bed. The ratio of ice to open-water depth for varying ice roughnesses can be obtained from the simplification of Beltaos (1982):

$$\frac{H_i}{H_o} = \left[1 + \left(\frac{n_i}{n_b}\right)^{3/2}\right]^{2/5} + 0.92\frac{t_i}{H_o} \tag{8}$$

where H_i and H_o are the water depth (m) for ice and open-water conditions, respectively.

Freeze-up Levels

As frazil pans and floes begin to accumulate in a river reach, it is their thickness and sub-surface roughness that cause a rise of the water level above the open-water value. Even when the ice thickness is relatively small, their roughness alone can cause an appreciable increase in stage. For example, when n_i is equal to n_b (equation 8), the flow depth will be approximately one-third greater than that for equivalent discharge under open-water conditions. Much larger backwater effects can be created when the ice cover thickens or collapses to form a relatively rough, freeze-up ice jam.

The initial thickness of an accumulating freeze-up cover depends on flow depth and velocity, and ice floe characteristics such as porosity (e.g. see reviews by Beltaos, 1995; Gray and Prowse, 1993). This type of cover development assumes that the internal strength of the accumulation can withstand the increases in forces created by its own gravitational-weight component and fluid shear stress at its base, as it grows upstream. For wide river channels, however, such "deposition" covers lack the initial strength to resist these forces and tend to collapse and telescope down into a thickness with sufficient internal resistance. These thicker and usually rougher accumulations create substantially more backwater and in extreme cases can pose a significant flood hazard. For example, with $n_i = 3n_b$ and a jam thickness approximately equal to the open-water depth, water depth would approximately triple. Over time, however, the roughness of such accumulations tends to reduce, either from thermal smoothing of the ice bottom or from additional infilling of ice from frazil deposition in concave portions (Ashton and Zufelt, 1991). As this occurs, the backwater decreases. Large additions of frazil, however, could also increase roughness and raise water levels.

Flow Storage and Low Flows

As a freeze-up cover develops, it can also affect discharge through the storage of river water either in the form of ice or in hydraulic storage. The first process most strongly affects shallow low-velocity streams, and is most pronounced during the initial rapid stage of ice formation, but continues throughout the winter as the ice cover continues to grow. Its total effect on flow is relatively minor. A much larger effect on discharge is created by the hydraulic storage of

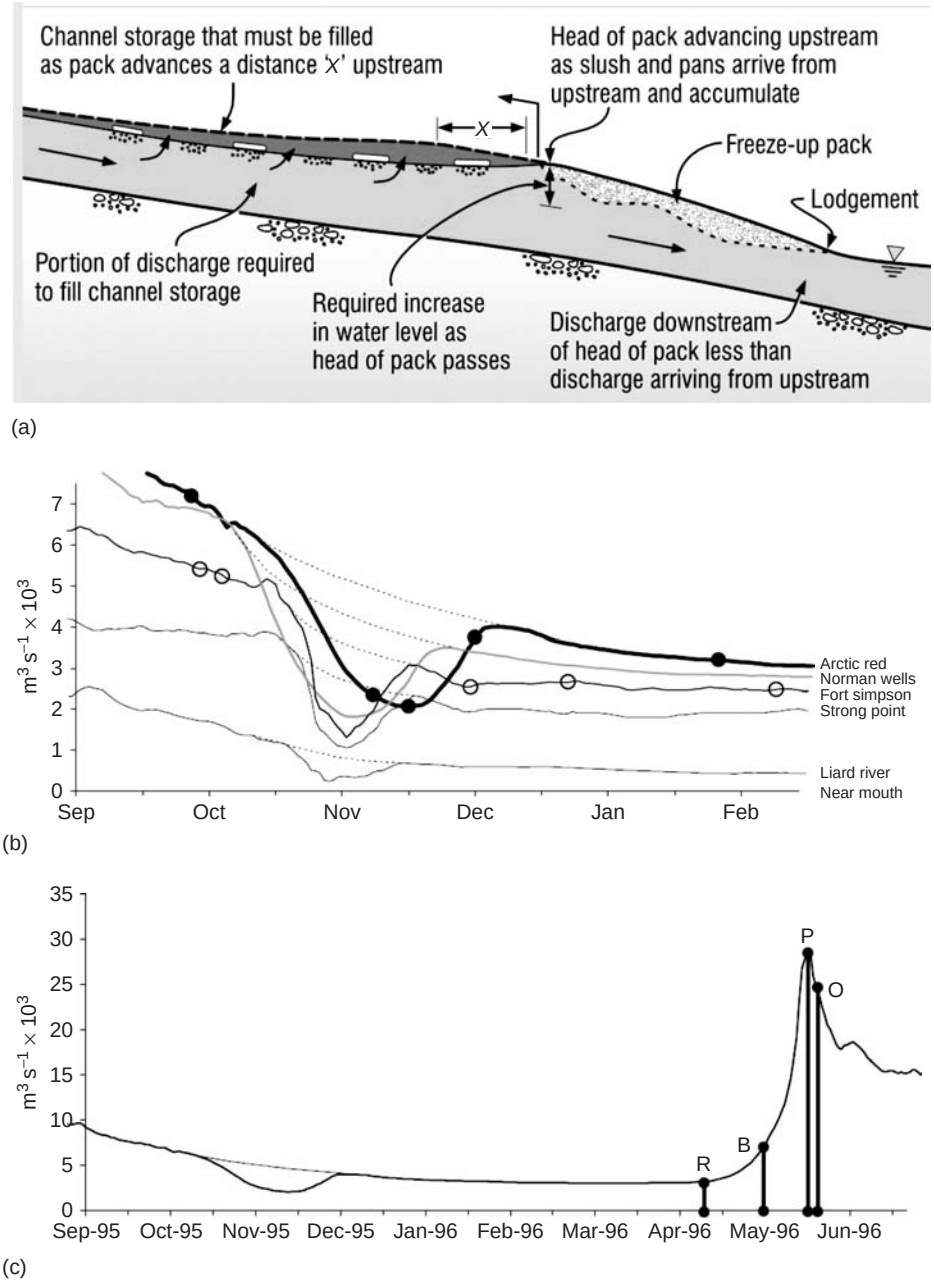


Figure 2 (a) Loss of storage to advancing freeze-up front (after Gerard, 1990); (b) Depressions in fall discharge on various sections of the Mackenzie River, Canada, 1995. Circles indicate field hydrometric measurements. (c) Same depression is evident in the seasonal hydrograph, which contrasts the size of the spring freshet. Most of the freeze-up storage is released between interval labeled ''R'' and ''O'' ((b) and (c) Reproduced from Prowse and Carter, 2002 by permission of John Wiley & Sons Ltd.). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

river water as it is abstracted to satisfy the stage increase at freeze-up (Figure 2a). The amount required depends on equation (8) and the upstream geometry of the channel. Huge storage requirements exist, for example, in cases where a lake or wide-channel system exist upstream of an ice-jam location. During the time that water is being removed from the flow, a period of reduced discharge exists downstream. As shown in Figure 2(a), the low flow at the trough of the discharge depression can be less than half that which would occur without ice effects (i.e. dashed line). Moreover, it is lower than that occurring in late winter when runoff from the snow-covered landscape is at a minimum. Flow abstraction ceases when the ice cover stops advancing upstream. Although this effect has been documented on a

variety of systems ranging from small temperate streams to large subarctic rivers (e.g. Conly and Prowse, 1997; Gerard, 1981; Gerard, 1990; Gray and Prowse, 1993; Hamilton and Moore, 1996; Moore *et al.*, 2002; Prowse and Carter, 2002), it is most pronounced and long-lasting on large rivers, with extensive upstream storage requirements. On the lower portions of the Mackenzie River, for example, Prowse and Carter (2002) found the period of freeze-up flow abstraction to last 60 days and to account for approximately 27% of the total expected flow during this period (Figure 2b).

Flow Augmentation

All the water abstracted from flow by river ice, either through hydraulic or frozen storage, is eventually released back into the flow. One of the more significant flow additions occurs when water in hydraulic storage is "discharged" at breakup. Even prior to this, however, volumes of water can be released as the ice-cover roughness changes. Thermal smoothing of initially rough freeze-up ice covers, for example, will lead to a release of some stored water although additional water can be placed back into storage as the bottom roughness increases due to thermal roughening (development of waves or ripples in the ice bottom) in early spring. Final release of this stored water only occurs at breakup when the ice cover is removed and the flow resistance of the ice cover eliminated. Studies of the Mackenzie River have shown that such seasonal transfer of water (from freeze-up to breakup) *via* hydraulic storage effects is rarely considered in flow calculations and can lead to an overestimate of the size of the spring snowmelt runoff (Figure 2c; Prowse and Carter, 2002).

Accompanying the release of hydraulically stored water at breakup is the melt of the accumulated winter ice cover. Depending on the intensity of the spring heating from atmospheric and hydrothermal sources, ice melt can be very protracted or produce a rapid addition of water to river flow. The localized addition of flow can be especially pronounced where breakup has led to a large thick accumulation of ice. Prowse (1990), for example, showed that the intense ablation of an ice jam produced a flow addition of approximately $100\,m^3\,s^{-1}$ over a three-day period on a large northern river, a value roughly equal to that contributed by spring melt from major tributary streams. Significant additions to flow from melting of in-channel ice also occur on systems characterized by large and/or numerous icings. These can be quite extensive on some northern rivers and produce a significant effect on the magnitude of flow on downstream larger-order rivers (e.g. Gerard, 1981). Of particular importance are contributions to summer flow on high-latitude continental rivers, which are typically low because of small summer precipitation and runoff (van Everdingen, 1987; Prowse, 1994).

Floating ice covers on lakes can assist in augmenting flow to downstream rivers at various times throughout the ice-covered season. This is particularly the case in regions or seasons characterized by significant winter snowfall. Loading of an ice sheet by a snowpack depresses it into the water column and displaces the related volume of water into the downstream flow system. Over 10% of the total winter flow in areas of northern Quebec, Canada, has been attributed to snow loading effects (Jones, 1969). Moreover, a snow load of only 50 mm (water equivalent) has been calculated to produce a 20% increase in discharge from some Finnish lakes (Kuusisto, 1984). Such flow additions are important modifiers of the winter flow regime, particularly in areas of low winter discharge. Periodic snowfall events are reflected as flow pulses in river hydrographs.

Breakup Ice Jams, Flooding and Surges

The progression of river-ice breakup is not continuous and is frequently interrupted by the formation of ice jams, the initiation of which is the focus of considerable current research (e.g. Beltaos, 1995; Hicks *et al.*, 1997; White, 2003). Although ice jams may develop at almost any location, especially when initially developed from the final runout of a breakup surge (Figure 3a), sites of regular major ice-jam formation are those where ice conveyance is significantly reduced. Channel constrictions, sharp bends, island complexes, bridge piers (Figure 3b), and reaches with significant decreases in river slope and/or widening of the channel, such as rivers entering larger rivers or lakes, are sites especially prone to ice-jam development. Unfortunately, an ability to forecast whether and where a jam will develop remains elusive (Beltaos, 2000).

A shoving and thickening of fragmented floes into a thickness with sufficient internal strength to resist the downstream forces exerted on it usually forms major breakup ice jams. At the downstream end or "toe", ice thickness is greatest and can even extend to the bed, thereby greatly increasing the flow resistance and strength of the accumulation. Water levels increase rapidly through the toe, typically creating a hydraulic slope greater than anything possible under normal open-water conditions. Assuming that the ice jam is of sufficient length, the central portion is occupied by an equilibrium reach characterized by maximum water levels (for this jam at this site) as well as uniform ice thickness and flow conditions (Beltaos, 1983, 1995). Further upstream within the ice jam head, the ice thicknesses gradually decrease into the upstream open-water zone. Backwater effects, however, can be experienced significant distances upstream beyond the ice-covered zone. On large northern rivers, they can easily extend 20 or more river widths upstream. The distance depends on the magnitude of the ice-jam water-level increase and the slope of the upstream channel.

(a)

(b)

Figure 3 (a) Breakup front advancing into downstream intact ice cover, Liard River, Canada. Depression of intact ice cover by advancing front has produced temporary surface ponding of water (Photograph: T.D. Prowse). (b) Ice jam formed against road bridge and supporting piers, St. John River, Canada (Photograph: S. Beltaos). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Ice jams exceeding 5 to 10 m in thickness are common on large northern rivers and the n_i created by the rough ice in such accumulations can be 0.05 to 0.10 (Gray and Prowse, 1993; Beltaos, 2001). Given such high thicknesses and that roughness values are several times those for even gravel or cobble bed rivers (e.g. 0.02–0.03; Shen and Julien, 1993), it is evident from equation (8) that the largest backwater levels are possible under ice-jam conditions.

Maximum water levels from an ice jam exist within the equilibrium reach but not all ice jams reach an equilibrium state. Regardless, however, the water levels that they create still exceed those under open-water conditions at equivalent discharge. It is the equilibrium condition that is typically used to assess maximum flood risk from ice jams (Watt, 1989). As for open-water conditions, higher discharge equates to increased equilibrium ice-jam water levels. Figure 4(a) illustrates a case for a large northern river including a theoretical maximum ice-jam curve, records of observed ice-jam water levels, and a rating curve derived for open-water conditions. Notably, in this case, many of the jam-induced water levels exceed those produced by the highest open-water flood ever recorded on the river. Moreover, most were produced at discharges only one–third to one–fourth as great even without reaching an equilibrium state.

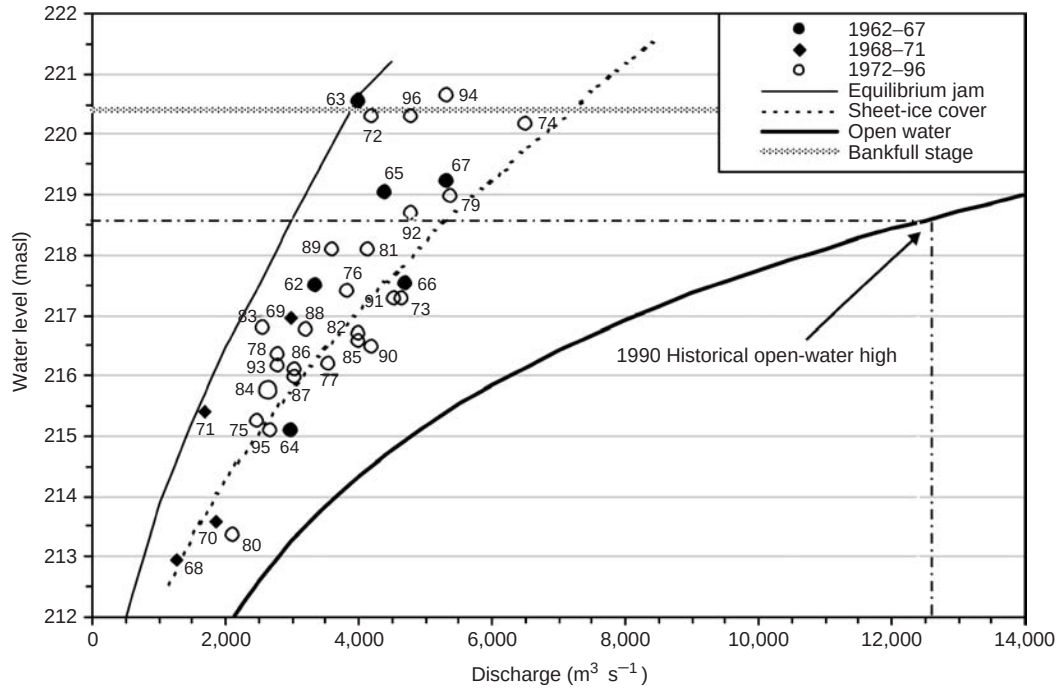
Despite the strong effect that ice has on water levels, probability/recurrence interval analyses of ice-induced flooding are relatively rare (e.g. Gerard and Karpuk, 1979; Gerard and Calkins, 1984; Watt, 1989; Grover *et al.*, 1999). This is more surprising from a flood-damage reduction perspective since data (e.g. Beltaos, 1982; Beltaos and Burrell, 2000a; Gerard and Karpuk, 1979) suggest that the larger floods have lower recurrence intervals under breakup than open-water conditions (e.g. Figure 4b).

Even when ice jams do not form in a particular reach, the large roughness associated with the rubbled ice in the advancing breakup front can produce dramatic increases in stage. Some of the largest and most rapid increases can be produced by surges associated with the upstream release of an ice jam (e.g. Beltaos and Krishnappan, 1982; Henderson and Gerard, 1981; Blackburn and Hicks, 2003). Breaking fronts associated with such surges have been documented to travel at over $5\,\mathrm{m\,s^{-1}}$ over long distances (e.g. Jasek, 1999, 2003). Rapid releases from flow-control structures have produced similar responses and offer opportunities for operational control of breakup conditions (Ferrick and Mulherin, 1989).

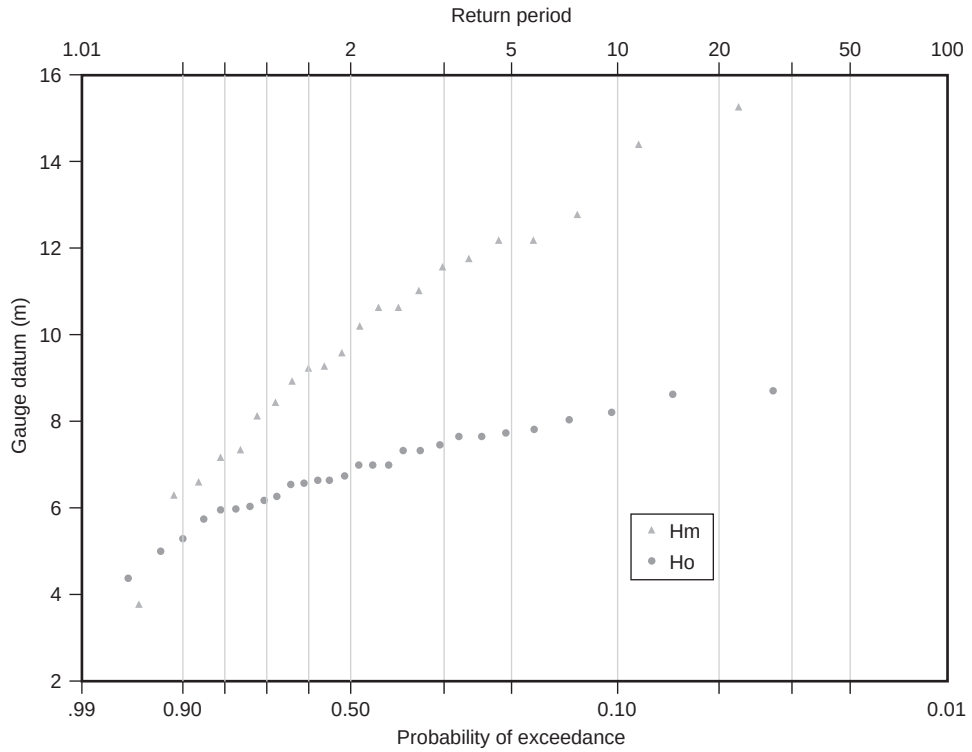
Sediment Transport and River Morphology

River ice is an important agent of geomorphological change because of its effects on erosional and depositional processes. The additional resistance to flow created by an ice cover leads to an increase in flow depth and a decrease in average flow velocity. A slower rate of energy dissipation is therefore required because the channel is conveying a deeper, slower flow. In general, therefore, the overall capacity of the channel to transport sediment and channel shaping is reduced. This can, for example, trigger changes to thalweg alignment and/or concentration. In the case of sinuous braided channels, for example, ice-induced reductions in the energy gradient can cause flow to concentrate in a single thalweg of greater sinuosity than the open-water thalweg (Ettema, 2002).

Variations in an ice cover can also lead to the redistribution of flow and changes in the position and/or depth of the thalweg (e.g. Hicks, 1993; Ettema and Zabilansky, 2004). In the special case of hanging ice dams, for example, their



(a)



(b)

Figure 4 (a) Rating curves showing water level/discharge relationships for different flow conditions and comparison with peak levels recorded during breakup. Uppermost line: calculated maximum stage that would occur from an equilibrium ice jam. Intermediate line: sheet-ice cover just prior to breakup, assumed thickness of 0.6 m and Manning roughness coefficient of 0.03; lowermost line: open-water condition (from Prowse *et al.*, 2002b); (b) Return period for flood levels produced under open-water (Ho) and ice breakup (Hm) conditions, Liard River, Canada (Reproduced from Prowse *et al.*, 2001 by permission of Water International). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

(a)

(b)

Figure 5 (a) Ice scour of high-elevation bank sediments, Mackenzie River, Canada (Photograph: T.D. Prowse); (b) Liard River (right hand of photograph) joining the Mackenzie River, Canada. Flow is from background to foreground for both rivers. Note that a dynamic breakup is affecting the Liard River and has produced large amounts of suspended sediment (dark tint to open water). The upstream portion of the Mackenzie (left background) is experiencing a thermal breakup, characterized by the movement of large sheets rather than fragmented ice as on the Liard, and remains relatively clear. (Photograph: T.D. Prowse). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

large thickness can cause a concentration of flow and a pronounced local scour of the riverbed (e.g. Sui *et al.*, 2000).

Many morphological features of an ice-covered river have developed over multiple years, although ice is also capable of producing some very dramatic changes within a single season. The most dramatic of these occur during the dynamic periods of freeze-up and particularly breakup. In contrast to the general reduction in sediment transport capacity that occurs during the stable ice period, breakup can produce dramatic increases in suspended sediment concentration, largely as a result of the intense ice scour of the bed and banks (Figure 5a and b). Field studies of two quite different Canadian rivers found that although sediment concentrations were quite low during the late-winter prebreakup period ($<10\,\mathrm{mg\,L^{-1}}$), they gradually increased by an order of magnitude or more just prior to breakup, and then peaked at over $1000\,\mathrm{mg\,L^{-1}}$ during active breakup (Prowse, 1993; Beltaos *et al.*, 1994; Beltaos and Burrell, 2000b; Milburn and Prowse, 1996, 2002). Peak concentrations can be several times that which would occur under open-water conditions for equivalent discharge (Prowse, 1993), pointing to the weakness of using any open-water derived sediment rating curves during these dynamic periods (Milburn and Prowse, 1996).

For many northern rivers, breakup can also be the period of maximum annual sediment transport. Walker (1969), for example, found that 75% of the annual sediment load was transported on a large Alaskan River during the three to four weeks period accompanying and immediately following breakup (also associated with 50% of the annual runoff). Flow surges during breakup are also believed to be effective transporters of bed-load material. Beltaos (1993), for example, estimated that a $5\text{-m\,s}^{-1}$ breakup surge would produce sufficient shear to move bed material as large as 0.2 m in diameter.

Although ice can accelerate erosion of soft-rock material, such as shale and limestone (Dionne, 1974; Danilov, 1972), ice-scour effects are most pronounced in alluvial rivers. Examples of erosional features include high-level benches (Smith, 1980) and undercut banks (e.g. Marusenko, 1956); the oversteepening of the latter being a major contributing factor in the initiation of landslides (Code, 1973). Slope failures can also result from sudden changes in effective (inter-granular) pressure and shearing resistance of the bank material produced by rapid fluctuations in water level produced by dynamic ice events. Over the long term, breakup has been proposed as being a significant factor in overall channel enlargement (Smith, 1980; Martinson, 1980). This issue remains unresolved (e.g. Kellerhals and Church, 1980; Church and Miles, 1982; Koutaniemi, 1984) and some evidence exists to suggest that overbank losses of flow due to ice jams may even promote channel narrowing (Ettema, 2002).

In regions dominated by permafrost or well-frozen near-surface material, the effects of ice scour are usually limited to the initial layer of unfrozen bank material (e.g. Eardley, 1938; Gill, 1972; Outhet, 1974), although near the ice-water-bank interface, thermo-erosional niches can be common (e.g. Lawson, 1983). Where such ice and water action exposes ice-rich permafrost, major bank destabilization can occur (Newbury and McCullough, 1983; Scott, 1978; Walker and Arnborg, 1966; Walker, 1969).

Ice-induced erosion can be especially effective in modifying channel morphology at river bends and meanders (e.g. Martinson, 1980). Under open-water conditions, the

outside of meander bends is normally a site of sediment erosion and the inside, sediment deposition. Under ice-jam conditions, however, the congestion of ice in the main channel often leads to a diversion of flow across a meander bend (e.g. Mackay *et al.*, 1974). If severe enough and/or if a meander-loop neck is relatively short, flow can be redirected into ancillary channels or a new dominant channel can develop through the neck thereby reducing local channel sinuosity (Williams and MacKay, 1973; MacKay *et al.*, 1974). Redirection of flow is common in river deltas, where ice-induced flooding frequently causes the shifting of distributaries through channel avulsion.

Although most ice-induced erosional features have been observed above the ice-water level, breakup ice scour is also known to affect the channel bed (e.g. Collinson, 1971; Danilov, 1972; King and Martini, 1984). Similarly, high-water velocities produced by ice-jam surges and found beneath ice jams can produce significant local bed erosion (e.g. Mercer and Cooper, 1977; Wuebben, 1988).

In addition to erosional features, river ice also leaves some unique depositional features varying from thick layers of sediment to localized boulder accumulations. In the meander case note above, for example, where a meander-loop neck is too wide and/or not easily eroded, overbank deposition of large amounts of sediment can raise the bank height and reinforce the meander loop (Ettema, 2002). For river regimes characterized by a decline in discharge following breakup, thick coats of sediment often blanket the low-velocity portions of the channel bed. Similar coatings can be found on the bank and floodplains, their depth being exceptionally large where breakup flow velocities, and hence sediment transport capacity, are dramatically reduced compared to those experienced in the main channel. Alluvium depths from a single event of up to 1 m, for example, have been reported on the Yukon River where successive accumulation are believed to have raised portions of the floodplain by as much as 6 m (Eardley, 1938). Sediment accumulation on the floodplain combined with low-level ice scour can accentuate the form of the above-noted high-level benches. The term "bechevniks" (portion of a bank suitable for rope towing of river craft) has been used to describe some of these ridge or bench features on Russian rivers, especially those repetitively worked by ice push (Hamelin, 1979). Sediment deposited by backwater flooding from ice jams in river deltas has also been noted to assist in the development of levees (Henoch, 1960) and the buildup of land between channels (King and Martini, 1984; Ritchie and Walker, 1974).

Direct ice-shove can also produce some localized large-scale depositional features, especially at outside bends, upstream ends of islands and banks opposite major tributaries. Although layers of light alluvium are often bulldozed into ramparts at such locations, they are usually washed away by subsequent surface runoff and high channel discharge (e.g. Bird, 1974). Where heavier till exists, the regular bulldozing of material and the subsequent washing away of finer materials can lead to the formation of more stable features such as boulder buttresses, barricades, ridges, and pavements (e.g. Mackay and MacKay, 1977; Rosen, 1979; Barnes, 1982). The first three are relatively localized features (Figure 6a) but wide stretches of the latter (Figure 6b) can be found covering hundreds of kilometers of river bank on large northern rivers (Kindle, 1918; Wentworth, 1932).

River Ecology

Dissolved Oxygen and Mixing Processes

Hydrologic processes affect numerous bio-geochemical systems in rivers. Over the winter months, isolation of the water column from direct atmospheric exchanges through

(a)

(b)

Figure 6 (a) Boulder barricade and (b) Boulder pavement, Mackenzie River, Canada (Photograph: T.D. Prowse). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the development of an ice cover produces some dramatic changes in stream-water chemistry. A change in dissolved oxygen (DO) levels is one of the most significant and has the largest potential to affect the entire river ecology.

As water temperatures decline during the autumn, rivers typically experience a rise in DO, often reaching a peak just prior to freeze-up, largely because of the increasing solubility of oxygen with decreasing water temperatures. Following complete freeze-over, however, DO can drop dramatically (e.g. within days; Whitfield and McNaughton, 1986). The rate of decline is largely dependent on the oxygen demand of the water column and channel bed (i.e. from respiration and decomposition), and the background DO level of the waters comprising the late-autumn discharge. With the elimination of direct water–air reaeration, DO decline can be quite precipitous for rivers where poorly oxygenated groundwater is the major, autumn flow component. Subsequent reductions tend to be much more gradual, reflecting the increasing proportion of poorly oxygenated groundwater in the flow (e.g. Schreier *et al.*, 1980), such as that originating from extensive wetlands, peatlands, bogs and/or fens. The oxidation of organic carbon flowing from such environments can further reduce DO levels. Similar DO depression can result from the input of oxygen-consuming material such as from industrial and municipal sources (e.g. Chambers *et al.*, 1997). In all these cases, it is important to remember that the presence of an ice cover effectively increases the residence time of the flow (by increasing stage and slowing velocity) exposed to various oxygen exchange processes (Ferrick and Calkins, 1996). Furthermore, since cold water temperatures slow decomposition and winter low flows reduce dilution, biological oxygen demand (BOD) can also be the highest during the ice-covered period (Hou and Li, 1987).

Although many rivers exhibit a gradual downstream decline in DO, this effect is not universal (Chambers *et al.*, 1997). Downstream gradients in DO are rarely simple and reflect a complex mixture of climatic, hydrologic, hydraulic, geochemical, and anthropogenic chemical conditions. Often, severe depletion can be eliminated by reaeration from a reach of open-water rapids or from highly oxygenated flow introduced from a lake-fed tributary. Moreover, although ice-covered conditions favor low DO, even larger depressions can occur under summer conditions with higher water temperatures and greater inputs of organic material.

Similar to the controlling conditions in lakes (e.g. Prowse and Stephenson, 1986), DO in rivers can rise in the spring because of the seasonal increase in available radiation, a decline in reflective losses as the snow is ablated exposing lower-albedo ice surfaces, and the resultant increase in photosynthetic activity (Whitfield and McNaughton, 1986; Cheng *et al.*, 1993). Over much of the winter, reflective losses are controlled by the overlying snowcover, which can have an albedo as high as 0.95 but decreases to the 0.4–0.6 range as the snow ages and ripens (e.g. Gray and Prowse, 1993). Much lower albedos characterize the various underlying white and black ice forms produced under varying combinations of precipitation input, thermal conditions, energy gradients and flow turbulence. Most bare river-ice surfaces are typically formed of white ice with albedos in the 0.3–0.4 range, although some black ice can develop, such as in calm windblown reaches. Highly transparent black ice can have an albedo similar to that for open water, although such conditions are more typical on lakes.

Conditions conducive to increasing the prebreakup photosynthetic production of oxygen are poorest on deep northern rivers characterized by thick ice covers and minimal under-ice flora and greatest on shallower, temperate rivers with thinner ice covers and larger plant biomass. Once the turbulent phase of breakup commences, direct reaeration resumes and DO levels can rapidly increase (e.g. Tilsworth and Bateman, 1982), even to the point of brief supersaturation (Milburn and Prowse, 2000). By contrast, breakup can also promote dramatic spring declines in DO. This can occur when significant amounts of organic material are resuspended by ice-induced breakup turbulence and scour and/or collected into the main flow by backwater floods sweeping organically rich riparian zones (Hynes, 1970; Hou and Li, 1987). High water temperatures that often immediately follow ice clearance (e.g. Prowse, 1990; Prowse and Marsh, 1989) can cause a rapid metabolism of such material and produce a marked reduction in DO.

Reductions in flow velocity and shear-stress parameters by an ice cover also directly affect a number of river-mixing parameters such as vertical diffusivity, transverse mixing and longitudinal dispersion. Vertical spreading of dissolved and suspended particles is largely controlled by the degree of turbulence in the flow. Although some questions remain about the vertical distributions of ice-affected flow velocities, it can be generally expected that an ice cover reduces vertical diffusive capacity (Beltaos, 1993). Similarly, although ice is known to change radial velocities (e.g. Zufelt, 1988; Urroz and Ettema, 1992), its effect on transverse mixing remains poorly defined. Most data, however, indicate that similar values of the dimensionless mixing coefficient (relative to total flow depth, not hydraulic radius) can be used for ice and open-water conditions (e.g. see Lau, 1985). Similarly, knowledge is relatively meager about ice effects on longitudinal dispersion, but the downstream spread of materials is expected to be smaller primarily because flow velocities are lower under ice than open-water conditions at equivalent flow depths (e.g. Beltaos, 1993).

Habitat Modification and Disturbance

Ice is capable of creating and destroying the quantity and quality of freshwater habitat on river systems. This

occurs throughout the entire winter season from initial ice growth to final breakup, and the effects are quite different from those normally associated with open-water hydrologic processes.

With the initial formation of border ice, new low-velocity zones protected from terrestrial predators are created along the channel margins (e.g. Power *et al.*, 1993; Maciolek and Needham, 1952). By contrast, the frazil-ice generation that accompanies freeze-up can cause some fish species to avoid main flow sections (e.g. Armstrong, 1986). Such avoidance is probably related to negative effects of frazil on gills and associated breathing (e.g. Brown *et al.*, 1994, 1999; Tack, 1938).

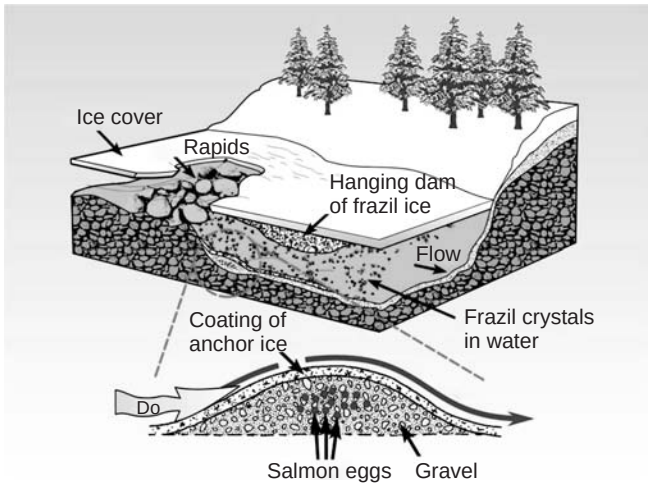
Deposition of frazil ice on the bed and the formation of anchor ice can also cause significant mortality of benthic fauna, ranging from invertebrates to fish embryo and even parr (e.g. McNeil, 1966; Reiser and Wesche, 1979; Calkins and Brockett, 1988; Walsh and Calkins, 1986). Impermeable coats of anchor ice severely restrict the flow of oxygenated water to substrate habitat and prevent the removal of waste products, thereby significantly decreasing the quality of substrate habitat (Figure 7a and b; e.g. Power *et al.*, 1993).

Frazil ice that accumulates on the bottom of an ice cover can, when in large quantities, significantly decrease the amount of available underwater fisheries habitat. Moreover, most deposition occurs in low-velocity zones and pools, the preferred winter habitat for fish to minimize energy expenditure (Brown and Mackay, 1995; Cunjak and Caissie, 1994). As a result of changing winter ice conditions, fish can be forced to emigrate into less-desirable winter habitat. Some fish species, however, depend on frazil masses as an incubating medium, such as the Atlantic tomcod (Fortin *et al.*, 1992; Power *et al.*, 1993). Prowse (2001b) reviews other effects of frazil and anchor ice on winter habitat.

Once a surface ice cover is established, increases in river-water stage and groundwater levels (created either by backwater staging or from direct increases associated with anchor ice buildup) can effectively recharge aquatic habitat in riparian zones that become dewatered by the autumn–winter decline in discharge (e.g. Burn, 1993; Paschke and Coleman, 1986). Many such side-channels and ponds even become preferred winter habitat for some fish species. Decreases in flow resulting from freeze-up staging, however, can also temporarily dewater downstream zones of shallow habitat.

Gradual declines in winter discharge on most ice-covered rivers can also create some unique near-bank habitat. As water levels decline, the original ice cover hinged along the banks drops to lower levels leaving open spaces beneath. These are known to provide access for aquatic mammals that forage bank vegetation and aquatic biota (Calkins, 1993). A similar under-ice habitat can be created in the middle portions of streams where the ice cover

(a)



(b)

Figure 7 (a) Frazil and anchor ice beneath a floating ice cover, Dutch Creek, Alberta, Canada (Photograph: R.S. Brown); (b) Schematic of frazil production beneath an ice cover, hanging dam formation and coating of gravel bed. Anchor ice makes the gravel bed relatively impermeable to flow and DO, thereby being detrimental to fish egg survival (from Prowse, 2000). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

becomes suspended on high-elevation points of the bed as flows decline (e.g. Calkins and Brockett, 1988; Power *et al.*, 1993).

As ice continues to grow over the winter period, it can freeze directly to the substrate, usually along the banks and in shallow zones. This can produce mortality of benthic organisms depending on their ability to withstand freezing effects and cold temperatures (e.g. Olsson, 1981; Oswood *et al.*, 1991). Extensive reaches of some shallow streams and larger rivers at high latitudes can freeze completely to the bed leaving only isolated pools for overwintering habitat. These are often sustained by groundwater flow and, for some arctic coastal rivers devoid of other unfrozen freshwater habitat (e.g. deep lakes), provide critical overwintering

and/or spawning sites for fish (Craig, 1989; Craig and Poulin, 1975; West and Smith, 1992).

Extreme hydrologic events create a large disturbance to the ecological form and function of freshwater ecosystems. Only of late has the role of river-ice hydrology, and specifically breakup, been identified as a neglected factor in ecological theory of rivers, particularly as it affects disturbances that control biological productivity and diversity (Prowse and Culp, 2003). One of the most obvious effects of the disturbance created by breakup is the gradient in riparian vegetation, whose floristic composition varies by exposure to ice action (e.g. Uunila, 1997). Where relatively frequent dynamic breakups occur, there typically exists a down-bank transition from upper woody species to lower herbaceous plants. Scour from moving or shoved ice usually determines the lower limit of tree growth. Lower sections become denuded of vegetation by breakup action but are usually rapidly recolonized by fast-growing, highly productive pioneer species, usually with root structures that permit a resumption of growth after breakup (e.g. willows (*Salix* spp.)). Banks lacking such vegetation gradients and species diversity are usually indicative of a system dominated by thermal breakups with correspondingly minimal ice scour. Ice scars on trees provide an excellent source of dendrochronological information for examining historical trends in breakup flooding (e.g. Gerard, 1981; Smith and Reynolds, 1983).

Although field observations are relatively scarce, breakup is also believed to be a major modifier of aquatic systems below the water line. Submerged macrophytes, for example, are believed to experience a similar fate to lower-elevation bank vegetation (e.g. Nichols *et al.*, 1989). Breakup intensity has also been suggested to affect the timing of peak algal biomass, species abundance, and community composition (e.g. Scrimgeour *et al.*, 1994). Moreover, since breakup floodwaters reach the elevations of organically rich riparian zones more often than open-water floods, they are likely to be the major supplier of allocthonous organic material and associated nutrients to the river channel and ultimately downstream lakes, deltas and estuaries (Lesack *et al.*, 1991; Scrimgeour *et al.*, 1994). This is especially important in high-latitude cold regions rivers, where biological productivity is commonly limited by nutrient supply (Peterson *et al.*, 1993). Breakup flooding has also been noted as being an essential water-balance component for maintaining water within the myriad of "perched" basins (elevated above the main flow network) that typify northern river deltas (e.g. Marsh and Lesack, 1996; Peters *et al.*, 2004). Drying because of the lack of breakup flooding has been noted to lead to major habitat modification and affect populations of, for example, waterfowl, aquatic mammals, and even large terrestrial mammals (PADIC, 1987; Prowse and Conly, 2001).

Although breakup has been shown to cause major disturbance to macroinvertebrates, their communities seem to recolonize quickly afterward and remain relatively stable from year to year. Interestingly, however, some invertebrate and fish species also exhibit specific behavioral responses, usually involving some type of shelter seeking, to avoid the effects of breakup and ice scour (Brown *et al.*, 2000; Olsson and Söderström, 1978; Prowse and Culp, 2003). In severe breakup that scours large portions of the channel, however, it may not be possible to avoid negative effects. Cunjak *et al.* (1998), for example, found that a major breakup event led to significant declines in the survival of various life stages of Atlantic salmon, especially those in the immobile egg stage at the time of the disturbance. Importantly, this event occurred in the midwinter when some biota tend to be less mobile and more susceptible to impact. Midwinter breakups, often initiated by sudden warming and/or rain-on snow events (e.g. Lawford *et al.*, 1995; Beltaos, 2002) are most characteristic of the climatic margins of the cold regions (Prowse and Bonsal, 2004), the ones most difficult to predict, and with the greatest potential to cause extensive damage to natural and built environments.

SUMMARY

Although past water resource studies of river ice have been limited primarily to site-specific hydraulic issues, the last decade has seen a rapidly growing recognition that river ice can affect a much broader realm of hydrologic processes. Moreover, its effects on basic hydrologic variables can be quite broad in time and space, varying from the short-term generation of localized floods to seasonal redistribution of discharge in major northern basins. Because of its often-dominant role in controlling discharge, velocity and water levels as well as energy and mass exchanges between the atmosphere and water column, it is also a major controller of other geomorphologic and ecologic processes. Despite its recent recognition as an important subfield of hydrology, the study of river ice remains a nascent and slow-growing science. This can partly be explained by (i) a reluctance by many researchers to shift energies away from the more readily studied open-water period, (ii) the greater practical difficulties of conducting field research into river ice as compared to, for example, cold regions landscape processes, and (iii) the broad scientific background required to tackle many river-ice hydrologic problems, such as ice mechanics, hydraulics and thermodynamics. However, considering that the major northward flowing rivers in the cold regions of the Northern Hemisphere are least affected by flow regulation and that they also offer some of the greatest potential for future development (e.g. Dynesius and Nilsson, 1994), it would seem scientifically prudent to focus more attention on advancing our understanding of river-ice hydrology. Further justification is provided by

the fact that ice-covered rivers at high latitudes and the cold regions/temperate margins are the ones expected to be most significantly affected by future climate change (Anisimov *et al.*, 2001; Gitay *et al.*, 2001; Prowse *et al.*, 2002a; Wrona *et al.*, 2004).

Acknowledgments

The author is indebted to S. Beltaos, M. Lacroix, M. Sharp, L. de Rham, D.L. Peters, and the anonymous reviewers for comments on earlier drafts of this manuscript. Production of this manuscript was supported by the National Water Research Institute of Environment Canada and the Department of Geography, University of Victoria, Canada.

REFERENCES

Adams W.P. (1981) Snow and ice on lakes. In *Handbook of Snow Principles, Processes, Management & Use*, Gray D.M. and Male D.H. (Eds.), Pergamon Press: pp. 437–466.

Anisimov O.A., Fitzharris B., Hagen J.O., Jeffries R., Nelson F., Prowse T.D. and Vaughan D. (2001) Polar Regions (Arctic and Antarctic). *Climate Change 2001, Impacts, Adaptation, and Vulnerability*, Chap. 16, Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press: Cambridge, pp. 801–841.

Armstrong R.H. (1986) A review of Arctic grayling studies in Alaska, 1952–1982. *Biological Papers of the University of Alaska*, **23**, 3–17.

Ashton G.D. (1985) Deterioration of floating ice covers. *Journal of Energy Resources Technology*, **107**, 177–182.

Ashton G.D. (Ed.) (1986) *River and Lake Ice Engineering*, Water Resources Publications: Littleton, p. 485.

Ashton G.D. and Kennedy J.F. (1972) Ripples on underside of river ice covers. *Journal of the Hydraulics Division, ASCE*, **98**, 1603–1624.

Ashton G.D. and Zufelt J. (1991) Evolution of ice cover roughness. In: *Proceedings of the 6th International Specialty Conference on Cold Regions Engineering (ASCE)*, Sodhi D.S. (Ed.) West Lebanon: 294–305.

Barnes P.W. (1982) Marine ice-pushed boulder ridge, Beaufort Sea, Alaska. *Arctic*, **35**(2), 312–316.

Beltaos S. (1982) *Hydraulics of Ice Covered Rivers*. In notes from Inland Waters Directorate Seminar A *Hydraulics of ice covered rivers and ice jam analysis*, Hull, Quebec.

Beltaos S. (1983) River ice jams: theory, case studies and applications. *Journal of Hydraulic Engineering, ASCE*, **109**(10), 1338–1359.

Beltaos S. (1984a) A conceptual model of river ice breakup. *Canadian Journal of Civil Engineering*, **11**, 516–529.

Beltaos S. (1984b) Study of river ice breakup using hydrometric stations records. *Proceedings of Workshop on the Hydraulics of River Ice*, June 20–21, Fredericton, pp. 41–59.

Beltaos S. (1990) Fracture and breakup of river ice cover. *Canadian Journal of Civil Engineering*, **17**(2), 173–183.

Beltaos S. (1993) Transport and mixing processes. In *Environmental Aspects of River Ice*, Sec. 2.5, NHRI Science Report No. 5: Prowse T.D., Gridley N.C. (Eds.), Environment Canada, National Hydrology Research Institute: Saskatoon, pp. 31–42.

Beltaos S.. (Ed.) (1995) Ice jam processes. In *River Ice Jams*. Water Resources Publications, Littleton, p. 372.

Beltaos S. (1997) Onset of river ice breakup. *Cold Regions Science and Technology*, **25**(3), 183–196.

Beltaos S. (2000) Advances in river ice hydrology. *Hydrological Processes*, **14**, 1613–1625.

Beltaos S. (2001) Hydraulic roughness of breakup ice jams. *Journal of Hydraulic Engineering ASCE*, **127**(8), 650–656.

Beltaos S. (2002) Effects of climate on mid-winter ice jams. *Hydrological Processes*, **16**, 789–804.

Beltaos S. and Burrell B.C. (2000a) Ice jam floods along the Saint John River, New Brunswick, Canada. *Proceedings of Extremes 2000 Symposium Iceland*, IAHS Publication No. 271, IAHS: pp. 9–14.

Beltaos S. and Burrell B.C. (2000b) Suspended sediment concentrations in the Saint John River during ice breakup. *Proceedings, 2000 Annual Conference of the Canadian Society for Civil Engineering*, London, Ontario, 235–242.

Beltaos S., Burrell B. and Ismail S. (1994) Ice and sedimentation processes in the Saint John River, Canada. *IAHR Ice Symposium*, Vol. 1, Trondheim, Norway The Norwegian Institute of Technology, pp. 11–21.

Beltaos S. and Krishnappan B.G. (1982) Surges from ice jam releases: a case study. *Canadian Journal of Civil Engineering*, **9**(2), 276–284.

Beltaos S., Prowse T.D. and Carter T. (2005) Ice regime of the lower peace river and ice-jam flooding of the Peace-Athabasca delta. *Hydrological Processes*, (submitted).

Bennett K. and Prowse T.D. (2005) *Defining the Geography of Ice-Covered Rivers*, Environment Canada, National Water Research Institute: NWRI Contribution No.05-305 Burlington/Saskatoon.

Bilello M.A. (1980) *Maximum Thickness and Subsequent Decay of Lake, River and Fast Sea Ice in Canada and Alaska*, Report 80-6, U.S. Army, Cold Regions Research and Engineering Laboratory, Hanover, p. 160.

Bird J.B. (1974) Geomorphic processes in the Arctic, in *Arctic and Alpine Environments*, Chap. 12(A), Ives J.D. and Barry R.G. (Eds.), London, pp. 703–720.

Blackburn J. and Hicks F. (2003) Suitability of dynamic modelling for flood forecasting during ice jam release surge events. *Journal of Cold Regions Engineering*, **17**(1), 18–36.

Brown R.S. and Mackay W.C. (1995) Fall and winter movements of and habitat use by cutthroat trout in the ram river, Alberta. *Transactions of the American Fisheries Society*, **124**(6), 873–885.

Brown R.S., Brodeur J.C., Power G., Daly S.F., White K.D. and McKinley R.S. (1999) Blood chemistry and swimming activity of rainbow trout exposed to supercooling and frazil ice. *Proceedings of the 10th Workshop on the Hydraulics of Ice Covered Rivers*, Winnipeg, pp. 97–110.

Brown R.S., Power G., Beltaos S. and Beddow T.A. (2000) Effects of hanging ice dams on winter movements and swimming activity of fish. *Journal of Fish Biology*, **57**, 1150–1159.

Brown R.S., Stanislawski S.S. and Mackay W.C. (1994) Effects of frazil ice on fish. In *Proceedings of the Workshop on Environmental Aspects of River Ice*, Prowse T.D. (Ed.), Environment Canada, Subcommittee on Hydraulics of Ice Covered Rivers: pp. 261–278.

Burn C.R. (1993) Stage-discharge relations in the Mackenzie Delta during winter and development of intrusive ice in lake-bottom sediments. *Proceedings of the VI International Permafrost Conference*, Beijing, pp. 60–65.

Calkins D.J. (1979) *Accelerated Ice Growth in Rivers*, Report 79-14, U.S. Army CRREL, Hanover, p. 4.

Calkins D.J. (1993) Physical effects of river ice. In *Environmental Aspects of River Ice*, Sec. 2.2, Prowse T.D. and Gridley N.C. (Eds.), Environment Canada, National Hydrology Research Institute, Saskatoon, NHRI Science Report No. 5, pp. 4–11.

Calkins D.J. and Brockett B.E. (1988) Ice cover distribution in Vermont and New Hampshire Atlantic salmon rearing streams. *Proceedings of the Fifth Workshop on Hydraulics of River Ice/Ice*Winnipeg, Manitoba, Canada*Jams*, pp. 85–93.

Carey K.L. (1973) *Icings Developed from Surface Water and Ground Water*, Monograph III – D3, Cold Regions Research and Engineering Laboratory: Hanover, p. 65.

Chambers P.A., Scrimgeour G.J. and Pietroniro A. (1997) Winter oxygen conditions in ice-covered rivers: the impact of pulp mill and municipal effluents. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 2796–2806.

Cheng H., Leppinen S. and Whitley G. (1993) Effects of river ice on chemical processes. In *Environmental Aspects of River Ice*, Prowse T.D. and Gridley N.C. (Eds.), Environment Canada, National Hydrology Research Institute: Saskatoon. NHRI Science Report No. 5, pp. 75–96.

Church M. and Miles M.J. (1982) (Discussion of Chapter 9): Processes and mechanisms of bank erosion. In *Gravel-bed Rivers*, Hey R.D., Bathurst J.C. and Thorne C.R. (Eds.), John Wiley & Sons: New York, pp. 259–271.

Code J.A. (1973) *The Stability of Natural Slopes in the Mackenzie Valley*, Report no. 73-9, Task Force on Northern Oil Development, p. 18.

Collinson J.D. (1971) Some effects of ice on a river bed. *Journal of Sedimentary Petrology*, **41**(2), 557–564.

Conly F.M. and Prowse T.D. (1997) Hydrologic response to freeze-up on large northern rivers (case study). In *Winter Environments of Regulated Rivers, Proceedings of the 8th Workshop on the Hydraulics of Ice Covered Rivers*, Andres D.D. Committee on River Ice Processes and the Environment, Hydrology Section, Canadian Geophysical Union(Ed.) pp. 23–42.

Craig P.C. (1989) An introduction to anadromous fishes in the Alaskan Arctic. *University of Alaska Biological Papers*, **24**, 27–54.

Craig P.C. and Poulin V.A. (1975) Movements and growth of arctic grayling (*Thymallus arcticus*) and juvenile arctic char (Salvelinus alpinus) in a small arctic stream, Alaska. *Journal of the Fisheries Research Board of Canada*, **32**, 689–697.

Cunjak R.A. and Caissie D. (1994) Frazil ice accumulation in a large salmon pool in the Miramichi River, New Brunswick: ecological implications for overwintering fishes. In *Proceedings of the Workshop on Environmental Aspects of River Ice*, NHRI Symposium No. 12, Prowse T.D. (Ed.) Environment Canada, National Hydrology Research Institute: Saskatoon, pp. 279–295.

Cunjak R.A., Prowse T.D. and Parrish D.L. (1998) Atlantic salmon in winter: "The season of parr discontent". *Canadian Journal of Fisheries and Aquatic Sciences*, **55**(Suppl. 1), 161–180.

Danilov I.D. (1972) River ice as a factor of relief formation and sedimentation. *Problemy Kriolitologii*, (Problems of Cryolithology), Vol. 2 Moscow University Press: (English translation), pp. 137–143.

Davar K.S., Beltaos S. and Pratte B. (Eds.) (1996) *A Primer on Hydraulics of Ice Covered Rivers. Canadian Committee on River Ice Processes and the Environment*, Canadian Geophysical Union, Hydrology section, p. 176.

Dionne J.-C. (1974) How drift ice shapes the St. Lawrence. *Canadian Geographical Journal*, **88**(2), 4–9.

Donchenko R.V. (1975) *Conditions for Ice Jam Formation in Tailwaters*, U.S. Army Cold Regions Research and Engineering Laboratory: Hanover, Draft Translation 669.

Dynesius M. and Nilsson C. (1994) Fragmentation and flow regulation of river systems in the northern third of the world. *Science*, **266**, 753–762.

Eardley A.J. (1938) Yukon channel shifting. *Bulletin of the Geological Society of America*, **49**, 343–358.

Ettema R. (2002) Review of alluvial-channel responses to river ice. *Journal of Cold Regions Engineering*, **16**(4), 191–217.

Ettema R. and Zabilansky L. (2004) Ice influences on channel stability: insights from Missouri's Fort Peck reach. *Journal of Hydraulic Engineering*, **130**(4), 279–292.

Ferrick M.G. and Calkins D.J. (1996) (Discussion of) Risk-equivalent seasonal discharge programs for ice-covered rivers. *Journal of Water Resources Planning and Management*, **122**(6), 442–444.

Ferrick M.G. and Mulherin N.D. (1989) *Framework for Control of Dynamic Ice Break-Up by River Regulation*, Report 89-12, U.S. Army CRREL, p. 14.

Fortin R., Léveillé M., Guénette S. and Laramée P. (1992) Contrôle hydrodynamique de l=avalaison des oeufs des larves de poulamon atlantique (*Microgadus tomcod*) sous le couvert de glace de la rivière Saine-Anne, Québec. *Aquatic Living Resources*, **5**, 127–136.

Gerard R. (1981) Regional analysis of low flows: a cold region example. In *Proceedings of the 5th Canadian Hydrotechnical Conference*, Canadian Society for Civil Engineering: pp. 95–113.

Gerard R. (1990) Hydrology of floating ice. In *Northern Hydrology, Canadian Perspectives*, Prowse T.D. and Ommanney C.S.L. (Eds.), Environment Canada, National Hydrology Research Institute: Saskatoon.NHRI Science Report No. 1, pp. 103–134 + ref.

Gerard R. and Calkins D.J. (1984) Ice-related flood frequency analysis: applications of analytical estimates. *Proceedings of Cold Regions Specialty Conference*, *Montreal, Quebec*, pp. 85–101.

Gerard R. and Davar K.S. (1995) Introduction. In *River Ice Jams*, Chap. 1, Beltaos S. (Ed.) Water Resources Publications: Highlands Ranch.

Gerard R. and Karpuk E. (1979) Probability analysis of historical flood data. *Journal of the Hydraulics Division, ASCE*, **105**,(HY9) 1153–1165.

Gilfilian R.E., Kline W.L., Osterkamp T.E. and Benson C.S. (1972) Ice formation in a small Alaskan stream. *Proceedings of Banff Symposia on the Role of Snow and Ice in Hydrology*, Banff, pp. 505–513.

Gill D. (1972) Modification of levee morphology by erosion in the Mackenzie River Delta, Northwest Territories, Canada. *Polar Geomorphology*, Institute of British Geographers: pp. 123–138.

Gilpin R.R., Hiarata T. and Cheng K.C. (1980) Wave formation and heat transfer at an ice-water interface in the presence of turbulent flow. *Journal of Fluid Mechanics*, **99**, 619–640.

Gitay H., Brown S., Easterling W., Jallow B., Antle J., Apps M., Beamish R., Chapin T., Cramer W. and Frangi J. *et al.* (2001) Ecosystems and their goods and services. *Climate Change 2001, Impacts, Adaptation, and Vulnerability*, Chap. 5, Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press: Cambridge, pp. 235–342.

Gray B.J., MacKay D.K. (1979) Aufeis (overflow ice) in rivers. *Proceeding of the Canadian Hydrology Symposium 79*, Vancouver, NRCC No. 17834, pp. 139–163.

Gray D.M. and Prowse T.D. (1993) Snow and floating ice, In *Handbook of Hydrology*, Chap. 7, Maidment D. (Ed.) McGraw-Hill: New York, pp. 7.1–7.58.

Grover P., Vrkljan C., Beltaos S. and Andres D. (1999) Prediction of ice jam water levels in a multi-channel river: fort Albany, Ontario. *Proceedings of the 10th Workshop on River Ice*, Winnipeg, pp. 15–29.

Hamelin L.-E. (1979) The *Bechevnik:* a river bank feature from Siberia. *The Musk-Ox*, **25**, 70–72.

Hamilton A.S. and Moore R.D. (1996) Winter streamflow variability in two groundwater-fed sub-Arctic rivers, Yukon Territory, Canada. *Canadian Journal of Civil Engineering*, **12**, 1249–1259.

Henderson F.M. and Gerard R. (1981) Flood waves caused by ice jam formation and failure. *Proceedings of the IAHR International Symposium on Ice*, Quebec, Vol. 1, pp. 277–287.

Henoch W.E.S. (1960) Fluvio-morphological features of the Peel and lower Mackenzie Rivers. *Geographical Bulletin*, **15**, 31–45.

Hicks F.E. (1993) Ice as the geomorphologic agent in an anastomosing river system. *Proceeding of the 7th Workshop on the Hydraulics of Ice Covered Rivers*, NHRI Symposium Series No. 12, Environment Canada, National Hydrology Research Institute: Saskatoon, pp. 3–20.

Hicks F., Cui W. and Andres D. (1997) Modelling thermal breakup on the Mackenzie River at the outlet of Great Slave Lake, N.W.T. *Canadian Journal of Civil Engineering*, **24**, 570–585.

Hou R. and Li H. (1987) A Modelling of BOD-DO dynamics in an ice-covered river in northern China. *Water Resources*, **21**(3), 247–251.

Hynes H.B.N. (1970) *The Ecology of Running Waters*, Liverpool University Press: p. 555.

Jasek M. (1999) Analysis of ice jam surge and ice velocity data. *Proceedings of the 10th Workshop on the Hydraulics of Ice covered Rivers*, Winnipeg, pp. 174–184.

Jasek M. (2003) Ice jam release surges, ice runs, and breaking fronts: field measurements, physical descriptions, and research needs. *Canadian Journal of Civil Engineering*, **30**, 113–127.

Jones J.A.A. (1969) The growth and significance of white ice at Knob Lake, Quebec. *Canadian Geographer*, **13**(4), 354–372.

Kane D.L. (1981) Physical mechanics of aufeis growth. *Canadian Journal of Civil Engineering*, **8**, 186–195.

Kellerhals R. and Church M. (1980) (Comment on) Effects of channel enlargement by river ice processes on bankfull discharge in Alberta, Canada. *Water Resources Research*, **16**(6), 1131–1134.

Kindle E.M. (1918) Notes on sedimentation in the Mackenzie River basin. *Journal of Geology*, **26**, 341–360.

King W. and Martini I.P. (1984) Morphology and recent sediments of the lower anastomosing reaches of the Attawapiskat River, James Bay, Ontario, Canada. *Sedimentary Geology*, **37**(4), 295–320.

Koutaniemi L. (1984) The role of ground frost, snow cover, ice break-up and flooding in the fluvial processes of the Oulanka River, NE Finland. *Fennia*, **162**(2), 127–161.

Kuusisto E. (1984) *Snow accumulation and snowmelt in Finland*, Publications of the Water Research Institute, National Board of Waters, Finalnd, No. 55.

Lau Y.L. (1985) Mixing coefficient for ice-covered and free-surface flows. *Canadian Journal of Civil Engineering*, **12**(3), 521–526.

Lawford R.G., Prowse T.D., Hogg W.D., Warkentin A.A. and Pilon P.J. (1995) Hydrometeorological aspects of flood hazards in Canada. *Atmosphere-Ocean*, **33**(1), 303–328.

Lawson D.E. (1983) *Erosion of perennially frozen streambanks*, Report 83-29, U.S. Army Corps of Engineers CRREL.

Lesack L.F.W., Hecky R.E. and Marsh P. (1991) The influence of frequency and duration of flooding on the nutrient chemistry of Mackenzie Delta lakes. In *Mackenzie Delta, Environmental Interactions and Implications of Development*, Marsh P. and Ommanney C.S.L. (Eds.), Environment Canada, National Hydrology Research Institute: Saskatoon, pp. 19–36.

Maciolek J.A. and Needham P.R. (1952) Ecological effects of winter conditions on trout and trout foods in Convict Creek, California. *Transactions of the American Fisheries Society*, **18**, 202–217.

MacKay J.R. and Mackay D.K. (1977) The stability of ice-push features, Mackenzie River, Canada. *Canadian Journal of Earth Sciences*, **14**(10), 2213–2225.

MacKay D.K., Sherstone D.A., Arnold K.C. (1974) Channel ice effects and surface water velocities from aerial photography of Mackenzie River break-up. *Hydrological Aspects of Northern Pipeline Development*, Task Force on Northern Oil Development, Environmental-Social Program, Northern Pipelines. Report No. 74-12.

Marcotte N. (1984) Anchor ice in Lachine Rapids, results of observations and analysis. *Proceedings of the IAHR International Symposium on Ice*, Vol. 1, Hamburg, 151–159.

Marsh P. and Lesack L.F.W. (1996) The hydrologic regime of perched lakes in the Mackenzie Delta: potential responses

to climate change. *Limnology and Oceanography*, **41**(5), 849–885.

Marsh P. and Prowse T.D. (1987) Water temperature and heat flux at the base of river ice covers. *Cold Regions Science and Technology*, **14**, 33–50.

Martinson C. (1980) *Sediment displacement in the Ottauquechee River – 1975–1978*, CRREL Special Report 80-20 US Army, Cold Regions Research and Engineering Laboratory, Hanover, NH, USA.

Marusenko Y.I. (1956) The action of ice on river banks (English translation of Russian text provided by Secretary of State, Canada, Multilingual Services Division, Ottawa.). *Priroda (Nature)*, **45**(12), 91–93.

McNeil W.J. (1966) Effect of the spawning bed environment on reproduction of pink and chum salmon. *Fishery Bulletin*, **65**(2), 495–523.

Mercer A.G., Cooper R.H. (1977) River bed scour related to the growth of a major ice jam. *Proceedings of the Third National Hydrotechnical Conference*, May 30–31, Université Laval, Laval Québec. Canadian Society for Civil Engineering: 291–308.

Michel B. (1971) *Winter regime of rivers and lakes*. *Cold Regions Science and Engineering Monograph III-Bla*, Cold Regions Research and Engineering Laboratory, U.S. Army: Hanover, p. 131.

Michel B. (1981) History of research on river and lake ice in Canada. *Proceedings of the IAHR International Symposium on Ice*, Vol. 1, Quebec, pp. 1–10.

Milburn D. and Prowse T.D. (1996) The effect of river-ice break-up on suspended sediment and select trace-element fluxes. *Nordic Hydrology*, **27**,(1/2) 69–84.

Milburn D. and Prowse T.D. (2000) Observations on some physical-chemical characteristics of ice breakup. *Journal of Cold Regions Engineering*, **14**(4), 214–223.

Milburn D. and Prowse T.D. (2002) Under-ice movement of cohesive sediments before river-ice breakup. *Hydrological Processes*, **16**(4), 823–834.

Moore R.D., Hamilton A.S. and Scibek J. (2002) Winter streamflow variability, Yukon Territory, Canada. *Hydrological Processes*, **16**, 763–778.

Morse B. and Hicks F.E. (2005) Advances in river ice hydrology 1992-2003. *Hydrological Processes*, **19**, 247–263.

Newbury R.W. and McCullough G.K. (1983) Shoreline erosion and restabilization in a permafrost-affected impoundment. *Proceedings 4th International Conference on Permafrost*, National Academy Press: pp. 918–923.

Nichols S.J., Schloesser D.W. and Hudson P.L. (1989) Submersed macrophyte communities before and after an episodic ice jam in the St. Clair and Detroit rivers. *Canadian Journal of Botany*, **67**, 2364–2370.

Olsson T.I. (1981) Overwintering of benthic macroinvertebrates in ice and frozen sediment in a North Swedish river. *Holarctic Ecology*, **4**(3), 161–166.

Olsson T.I. and Soderstrom O. (1978) Springtime migration and growth of *Parameletus chelifer* (Ephemeroptera) in a temporary stream in northern Sweden. I. *Oikos*, **31**(3), 284–289.

Oswood M.W., Miller L.K., Irons J.G. III (1991) Overwintering of freshwater benthic macroinvertebrates. In *Insects at Low Temperatures*, Chap. 15, Lee R.E. and Denlinger D.L. (Eds.), Wiley: New York, pp. 360–375.

Outhet D.N. (1974) Progress report on bank erosion studies in the Mackenzie River Delta, Northwest Territories, Canada. In: *Hydrologic Aspects of Northern Pipeline Development*, Vol. 74-12, Task force on northern oil development, environmental-social program, Northern Pipelines, pp. 297–345.

PADIC (Peace-Athabasca Delta Implementation Committee) (1987) *Peace-Athabasca Delta Water Management Works Evaluation, Final Report* under the Peace-Athabasca Implementation Agreement. Hydrological Assessment; Appendix B, Biological Assessment; Appendix C, Ancillary Studies. p. 63 and Appendix A.

Paschke N.W. and Coleman H.W. (1986) Forecasting the effects on river ice due to the proposed Susitna hydroelectric project. *Cold Regions Hydrology Symposium*, University of Alaska, American Water Resources Association: Fairbanks, pp. 557–563.

Peters D.L., Prowse T.D., Marsh P., Lafleur P.M. and Buttle J.M. (2005) Estimating the persistence of water within perched basins of the Peace-Athabasca Delta, Northern Canada. *Wetlands Ecology and Management*, (submitted).

Peterson B.J., Deegan L., Helfrich J., Hobbie J.E., Hullar M., Moller B., Ford T.E., Hershey A., Hiltner A. and Kipphut G. *et al.* (1993) Biological responses of a tundra river to fertilization. *Ecology*, **74**, 653–672.

Power G., Cunjak R., Flannagan J. and Katopodis C. (1993) Biological effects of river ice. In *Environmental Aspects of River Ice*, NHRI Science Report No 5, Prowse T.D. and Gridley N.C. (Eds.), Environment Canada, National Hydrology Research Institute: Saskatoon, pp. 97–119.

Prowse T.D. (1990) Heat and mass balance of an ablating ice jam. *Canadian Journal of Civil Engineering*, **17**(4), 629–635.

Prowse T.D. (1993) Suspended sediment concentration during river ice break-up. *Canadian Journal of Civil Engineering*, **20**(5), 872–875.

Prowse T.D. (1994) The environmental significance of ice to cold-regions streamflow. *Freshwater Biology*, **32**(2), 241–260.

Prowse T.D. (1995) River ice processes. In *River Ice Jams*, Beltaos S. (Ed.) Water Resources Publications, LLC: pp. 29–70.

Prowse T.D. (2000) *River Ice Ecology*, Monograph, Environment Canada, National Water Research Institute: Saskatoon, p. 64.

Prowse T.D. (2001a) River-ice ecology: Part A) Hydrologic, geomorphic and chemical aspects. *Journal of Cold Regions Engineering*, **15**(1), 1–16.

Prowse T.D. (2001b) River-ice ecology: Part B) Biological aspects. *Journal of Cold Regions Engineering*, **15**, 17–33.

Prowse T.D. and Beltaos S. (2002) Climatic control of river-ice hydrology: a review. *Hydrological Processes*, **16**(4), 805–822.

Prowse T.D. and Bonsal B.R. (2004) Historical trends in river-ice break-up: a review. *Nordic Hydrology*, **35**(4), 281–293.

Prowse T.D., Bonsal B.R., Lacroix M.P. (2002a) Trends in river-ice breakup and related temperature controls. *Proceedings 16th IAHR International Symposium on Ice*, Dunedin, December 2–6, Vol. 3, pp. 64–71.

Prowse T.D. and Carter T. (2002) Significance of ice-induced hydraulic storage to spring runoff: a case study of the Mackenzie River. *Hydrological Processes*, **16**(4), 779–788.

Prowse T.D. and Conly M. (2001) Multiple-hydrologic stressors of a northern delta ecosystem. *Journal of Aquatic Ecosystem Stress and Recovery*, **8**(1), 17–26.

Prowse T.D. and Culp J.M. (2003) Ice breakup: a neglected factor in river ecology. *Canadian Journal of Civil Engineering*, **30**, 128–144.

Prowse T.D., Demuth M.N. and Chew H.A.M. (1990) The deterioration of freshwater ice due to radiation decay. *Journal of Hydraulic Research*, **28**(6), 685–697.

Prowse T.D., Lacroix M.P. and Beltaos S. (2001) Flood frequencies on cold-regions rivers. *Abstracts of the 27th Scientific Meeting of the Canadian Geophysical Union*, May 14–17, Ottawa.

Prowse T.D. and Marsh P. (1989) Thermal budget of river ice covers during break-up. *Canadian Journal of Civil Engineering*, **16**(1), 62–71.

Prowse T.D., Peters D., Beltaos S., Pietroniro A., Romolo L., Töyrä J. and Leconte R. (2002b) Restoring ice-jam floodwater to a drying delta ecosystem. *Water International*, **27**(1), 58–69.

Prowse T.D. and Stephenson R.L. (1986) The relationship between winter lake cover, radiation receipts and the oxygen deficit in temperate lakes. *Atmosphere-Ocean*, **24**(4), 386–403.

Reiser D.W. and Wesche T.A. (1979) In situ freezing as a cause of mortality in brown trout eggs. *The Progressive Fish Culturist*, **41**(2), 58–60.

Ritchie W. and Walker H.J. (1974) Riverbank forms of the Colville River delta. In *The Coast and Shelf of the Beaufort Sea*, Reed J.C. and Sater J.E. (Eds.), The Arctic Institute of North America: pp. 545–562.

Rosen P.S. (1979) Boulder barricades in central Labrador. *Journal of Sedimentary Petrology*, **49**(4), 1113–1124.

Schreier H., Erlebach W. and Albright L. (1980) Variations in water quality during winter in two Yukon rivers with emphasis on dissolved oxygen concentration. *Water Research*, **14**(9), 1345–1351.

Scott K.M. (1978) *Effects of Permafrost on Stream Channel Behaviour in Arctic Alaska*, USGS Professional Paper No. 1068, p. 19.

Scrimgeour G.J., Prowse T.D., Culp J.M. and Chambers P.A. (1994) Ecological effects of river ice break-up: a review and perspective. *Freshwater Biology*, **32**, 261–275.

Shen H.T. (2000) River ice transport theories: past, present and future. *Proceedings IAHR Ice Symposium*, Gdansk, Vol. 2: pp. 29–40.

Shen H.W. and Julien P.Y. (1993) Erosion and sediment transport. In *Handbook of Hydrology*, Maidment D. (Ed.), McGraw-Hill: New York, pp. 12.1–12.61.

Shulyakovskii L.G. (1966) *Manual of Forecasting Ice-Formation for Rivers and Inland Lakes*, Gidrometeorologicheskoe Izdatel=stvo, Israel Program for Scientific Translations: Leningrad.

Shulyakovskii L.G. (1972) On a model of the break-up process. *Soviet Hydrology: Selected Papers*, **1**, 21–27.

Shumskii P.A. (1964) *Principles of structural glaciology*, Dover Publications: New York, pp. 497.

Smith D.G. (1980) River ice processes: thresholds and geomorphologic effects in northern and mountain rivers. In *Thresholds in Geomorphology*, Chap. 15, Coates D.R. and Vitek J.D. (Eds.), George, Allen, and Unwin: London, pp. 323–343.

Smith D.G. and Reynolds D.M. (1983) Tree scars to determine the frequency and stage of high magnitude river ice drives and jams, Red Deer, Alberta. Canadian. *Water Resources Journal*, **8**(3), 77–94.

Sui J., Wang D. and Karny B.W. (2000) Suspended sediment concentration and deformation of riverbed in a frazil jammed reach. *Canadian Journal of Civil Engineering*, **27**(6), 1120–1129.

Tack E. (1938) Trout mortality from the formation of suspended ice crystals. *Fischerei-Zeitun*, **41**(4), 42;Reviewed in Wolf L.E. (1938) *The Progressive Fish Culturist*, **37**, 26.

Tilsworth T. and Bateman P.L. (1982) Changes in Chena river water quality. *The Northern Engineer*, **14**(3), 29–37.

Timco G.W. and Goodrich L.E. (1988) Ice rubble consolidation. *Proceedings of the 9th IAHR International Symposium on Ice*, Sapporo, Aug 23–27, Vol. I, pp. 427–438.

Tsang G. 1982 *Frazil and Anchor Ice B A Monograph*, National Research Council of Canada: Ottawa, Ontario: Subcommittee on Hydraulics of Ice-covered Rivers. p. 90 and Appendix.

Urroz G.E. and Ettema R. (1992) Bend ice jams: laboratory observations. *Canadian Journal of Civil Engineering*, **19**, 855–864.

Uunila L.S. 1997 Effects of river ice on bank morphology and riparian vegetation along Peace River, Clayhurst to Fort Vermilion. In: *Proceedings of the 9th Workshop on River Ice*. Sept. 24–26, Fredericton, NB. Canadian Geophysical Union – Hydrology Section, Committee on river Ice Processes and the Environment: 315–334.

van Everdingen R.O. (1987) The importance of permafrost in the hydrological regime. In *Canadian Aquatic Resources*, Healey M.C. and Wallace R.R. (Eds.), Canadian Bulletin of Fisheries and Aquatic Sciences No. 215: Department of Fisheries and Oceans, Ottawa, Canada pp. 243–276.

van Everdingen R.O. (1990) Ground-water hydrology. In *Northern Hydrology, Canadian Perspectives*, NHRI Science Report No. 1, Prowse T.D. and Ommanney C.S.L. (Eds.), Environment Canada, National Hydrology Research Institute: Saskatchewan 77–101 p + ref.

Walker H.J. (1969) Some aspects of erosion and sedimentation in an arctic delta during breakup. *Symposium on the Hydrology of Deltas*, Publication 90, International Association of Scientific Hydrology: Bucharest, May 6–14, pp. 209–219.

Walker H.J. and Arnborg L. (1966) Permafrost and ice-wedge effects on riverbank erosion. *Proceedings of the First International Conference on Permafrost*, Publication No. 1287, National Research Council: Lafayette, November 11–15, pp. 164–171.

Walsh M., Calkins D. (1986) River ice and salmonics. *Fourth Workshop on Hydraulics of River Ice*, US Army Cold Regions Research and Engineering Laboratory: Montreal, June 19–20, D4.1–D4.26.

Watt W.E. (Ed.) (1989) *Hydrology of Floods in Canada: A Guide to Planning and Design*, National Research Council of Canada: Ottawa, p. 245.

Wentworth C.K. (1932) The geologic work of ice jams in subarctic rivers. *Contributions in Geology and Geography*, Thomas L. F., Washington University Studies New Series, *Science and Technology*, **7**, 49–82.

West R.L. and Smith M.W. (1992) Autumn migration and overwintering of Arctic grayling in coastal streams of the Arctic National Wildlife Refuge, Alaska. *Transactions of the American Fisheries Society*, **121**, 709–715.

Williams G.P. and MacKay D.K. (1973) The characteristics of ice jams. *Seminar on Ice Jams in Canada*, University of Alberta, National Research Council of Canada: Edmonton, pp. 17–35.

White K.D. (2003) Review of prediction methods for breakup ice jams. *Canadian Journal of Civil Engineering*, **30**, 89–100.

Whitfield P. and McNaughton B. (1986) Dissolved-oxygen depressions under ice cover in two Yukon rivers. *Water Resources Research*, **22**(12), 1675–1679.

Wrona F.J., Prowse T.D. and Reist J. (2004) *Freshwater Arctic Ecosystems*, Chap. 7 Arctic Climate Impact Assessment.

Wuebben J.L. (1988) A preliminary study of scour under an ice jam, *Proceedings of the 5th Workshop on Hydraulics of River Ice/Ice Jams*, Winnipeg, June 21–24, pp. 177–190.

Zufelt J.E. (1988) Transverse velocities and ice jamming potential in a river bend, *Proceedings of the 5th Workshop on Hydraulics of River Ice/Ice Jams*, Winnipeg, pp. 193–207.

172: Permafrost Hydrology

LARRY D HINZMAN[1], DOUGLAS L KANE[1] AND MING-KO WOO[2]

[1]Water and Environmental Research Center, Institute of Northern Engineering, University of Alaska, Fairbanks, AK, US
[2]School of Geography and Geology, McMaster University, Hamilton, Canada

Permafrost is earth material that has temperature at or below 0°C for at least two consecutive summers. Above the permafrost is the active layer, a zone that freezes in winter and thaws in summer. Even though the principles governing water movement in permafrost areas are the same as those in more temperate regions, interactions of extremes in climate and the land surface characteristics render permafrost hydrology different from the hydrology of temperate latitudes. Ice-rich permafrost prevents infiltration of rainfall or snowmelt water, often maintaining a moist to saturated active layer where the permafrost table is shallow. Most hydrologic activities are confined above ground or in the thin active layer, which supplies summer moisture to plants and for evaporative flux. Limited storage capacity of the thawed zone does not support extended baseflow in a stream, though the proportion of baseflow increases as the percentage of permafrost extent decreases. In areas where permafrost is discontinuous or where it has thawed substantially near the surface, local hydrology may display a marked different character as there are stronger exchanges between the surface water and the ground water system, or water may drain laterally resulting in drier surface conditions. Runoff paths range from interhummock cracks, soil pipes, water tracks to distinct channels; and permafrost affects the ground water and contaminant migration pathways. Understanding the interdependence of permafrost, hydrology, and ecosystems is critically important to enable accurate projections of future conditions in the high latitudes.

INTRODUCTION

Hydrology in high latitudes and elevations is uniquely dominated by snow, ice, and frozen ground. Permafrost establishes a unique discipline of hydrologic science where the hydrologic regime is intimately coupled with the thermal regime to the extent that one may not be completely understood without correct characterization of the other. In permafrost regions, material properties may change drastically on a scale of centimeters to meters, particularly in the vertical dimension due to distinct changes in soil and thermal characteristics. Properties may vary just as dramatically in the horizontal dimension across the boundary of discontinuous permafrost. The active layer, the thin layer of soil (on the order of cm to m) above the permafrost, thaws and freezes annually, creating a system of constantly changing storage capacity and thermal and hydraulic properties. Although the spatial extent of permafrost changes on relatively slow timescales in response to disturbance or a changing climate, this too introduces an added level of complexity, as some watershed characteristics may not be considered constant on short or long timescales. Permafrost may nearly eliminate the interactions between near-surface and subpermafrost aquifers, which in essence defines the hydrologic response of every watershed that is directly influenced by permafrost. This text will explore the dynamic control that permafrost exerts upon hydrological processes including distribution, movement, and storage of water that is directly or indirectly influenced by permafrost.

PERMAFROST EXTENT AND CHARACTERISTICS

Permafrost, or perennially frozen ground (*see* **Chapter 71, Freezing and Thawing Phenomena in Soils, Volume 2**), is common throughout the high latitudes and is estimated

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

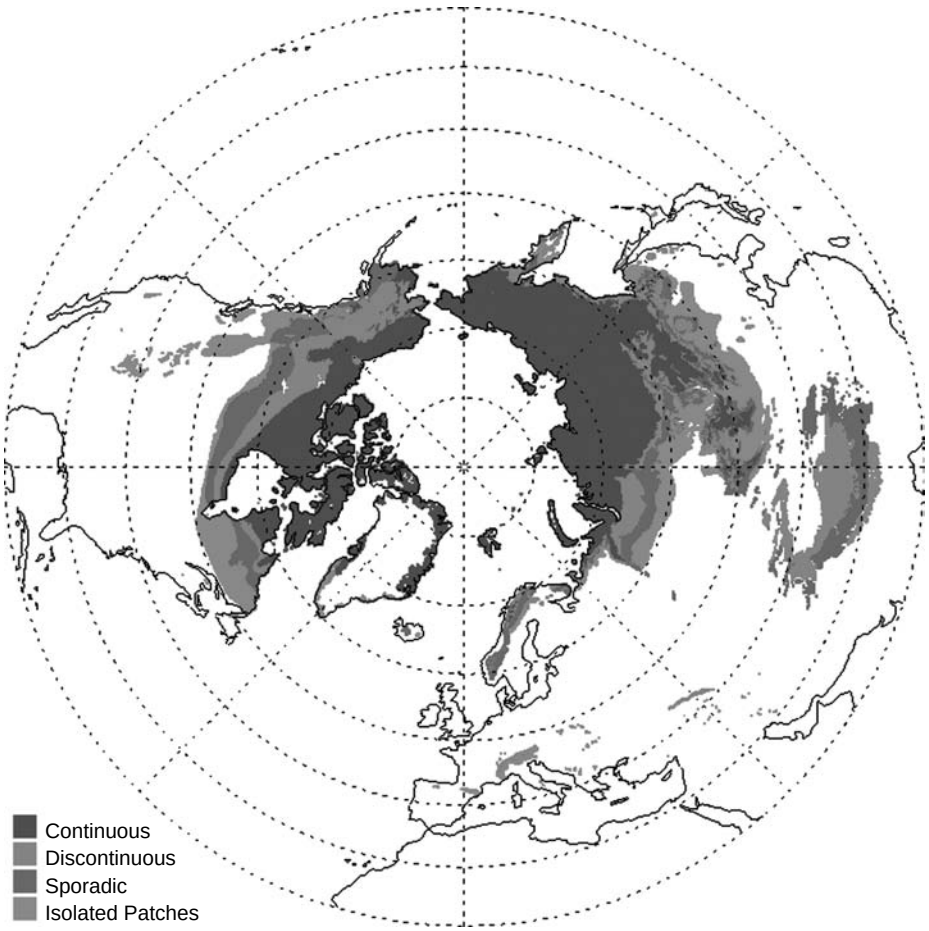


Figure 1 North Polar projection of permafrost distribution in the NorthernHemisphere (Zhang *et al.,* 1999). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

to underlie about 18–24% of the terrestrial surface of the Northern Hemisphere (Brown, 1970; Brown *et al.*, 1998; Zhang *et al.*, 2001) (Figure 1) and exerts tremendous influences upon the ecological, climatological, and hydrological processes occurring in those areas. Permafrost is not defined by geology or region, but is characterized only by a description of the thermal regime. If the ground temperature (be it ice, peat, mineral soil, or rock) remains below 0 °C for two or more consecutive years, it is considered permafrost.

Permafrost may be continuous, discontinuous, or sporadic under the ground surface (Figure 2), as controlled by the surface energy balance, soil thermal and hydraulic properties, terrain, and snow cover. Permafrost is generally continuous in the more northerly reaches of North America and Eastern Russia, but discontinuous permafrost does extend as far south as 50° north latitude in Canada and Russia (Brown, 1970; Romanovskii, 1985, *see* **Chapter 25, Global Energy and Water Balances, Volume 1**). Discontinuous permafrost is also extensive in the Tibetan Plateau due to the high elevations (Wang and French, 1995; Jin *et al.*, 2000). Alpine permafrost exists in high mountainous regions throughout the world, even in the tropical latitudes (Paepe and Van Overloop, 2001). Permafrost is also common in the Southern Hemisphere in the nonglaciated areas of Antarctic and in the high elevations of the South American Andes (Trombotto *et al.*, 1997). In the nonglaciated regions of Russia, permafrost may penetrate as deep as 1500 m (Grigoriev and Sokolov, 1994), while the thickest permafrost in northern Alaska is estimated at about 600 m (Osterkamp *et al.*, 1985) and 1000 m in northern Canada (Brown, 1978). Even within these bounds, small areas of unfrozen ground, called *taliks,* exist beneath large lakes and rivers or near springs (Permafrost Subcommittee, 1988). Permafrost boundaries in the discontinuous regions are primarily controlled by physiographic factors such as slope, aspect, and elevation; however, anthropogenic and natural disturbances such as wildfires can have substantial effects on the permafrost thermal status and extent (Yoshikawa *et al.*, 2002). The predominant control of permafrost on hydrology is most apparent when comparisons are made across a permafrost boundary where regional differences in climate may be eliminated. Although we

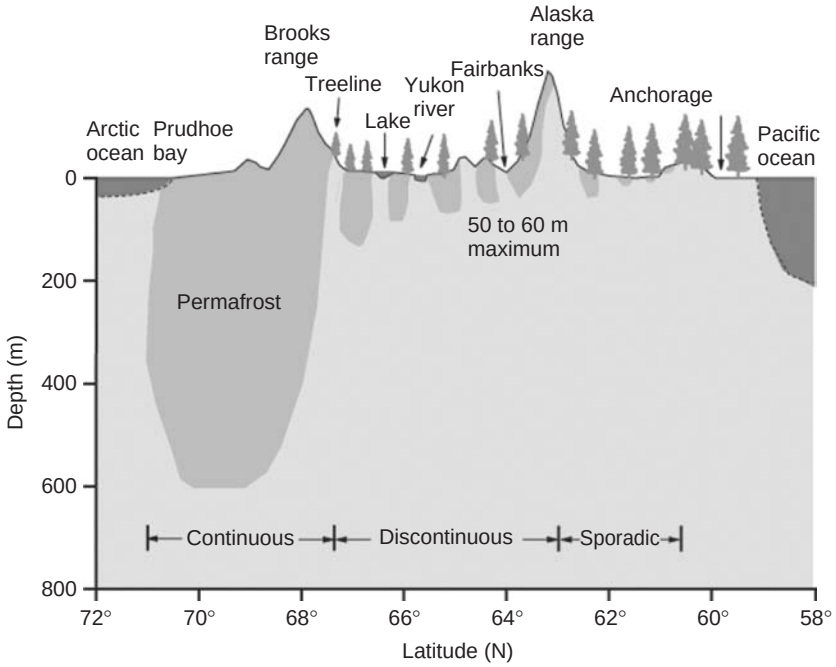


Figure 2 In Alaska, permafrost is generally continuous and quite thick north of the Brooks Range. It is discontinuous between the Brooks Range and Alaska Range and becomes sporadic south of the Alaska Range. The extent and thickness of the permafrost strongly influences the surface and subsurface hydrologic processes (taken from Kane and Boike, 2002). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

often observe changes in vegetation and other ecosystem processes occurring there as well (*see* **Chapter 101, Ecosystem Processes, Volume 3**), such indirect impacts of permafrost also influence the hydrology.

The characteristics of hydrology in permafrost regions vary greatly with region, being strongly influenced by regional climate, topography, and geology (Prowse, 1990; Kane and Yang, 2004). In general, evapotranspiration rates decrease with increasing latitude. The ratios of runoff to precipitation (R/P) tends to increase with increasing latitude; this is probably related to lower losses to evaporation and more limited storage in the thinner active layers of the higher latitude watersheds. However, in regional and local perspectives, hydrological processes are greatly influenced by permafrost characteristics that may be unique to a particular area. For example, hydrology in the Low Arctic of Alaska and Canada is strongly influenced by the thick organic soils (Dingman, 1973; Quinton and Gray, 2003) that have accumulated in some cases since the Pleistocene (Mann *et al.*, 2002). The polar deserts of the Canadian High Arctic and Antarctica yield markedly different hydrologic responses due to the thinner active layers, more undeveloped soils, and shorter thaw seasons (Edlund *et al.*, 1990; Woo and Young, 2003; Lyons *et al.*, 1997).

There are also several important characteristics shared by watersheds influenced by permafrost. Snowmelt is typically the dominant annual hydrologic event (*see* **Chapter 114, Snowmelt Runoff Generation, Volume 3**) usually yielding both the peak runoff magnitude and much of the annual runoff volume (Woo, 1986; Marsh, 1990; Ishii *et al.*, 2001). However, for small watersheds ($< \sim 1000\,\text{km}^2$) the floods of record are rainfall generated (Kane *et al.*, 2003). The low permeability zone created by permafrost functions as an aquiclude that effectively divides the ground water system into sub- and suprapermafrost components (Woo, 1990). In the zone of continuous permafrost, the predominance of hydrologic processes occurs in the seasonally frozen and thawed zone above the permafrost table, and processes associated with deep ground water are usually absent (Woo, 1990). Typically, during the winter season, the only active hydrologic processes are freezing of soil moisture in the active layer, redistribution of snow by wind and, in some locations, development of icings (also called *aufeis*) on the surface (Washburn, 1979; Kane, 1981; Grigoriev and Sokolov, 1994). There is a close association between energy and water fluxes and freeze-thaw events in water storage and redistribution as snow and ice melt. There is also a threshold change in hydrologic processes.

Perhaps the most important characteristic of permafrost is the ice content, impacting surface terrain, the geotechnical properties, and hydrologic properties and processes. Permafrost may be ice rich or lack significant amounts of ice, and the hydraulic properties can vary dramatically depending upon the ice-filled pore volume. Kane (1980)

and Kane and Stein (1983a) looked at infiltration rates into ice-rich frozen soils and demonstrated that no significant ground water recharge occurs in areas with shallow permafrost due to reduced permeability of the frozen soils where pores are completely blocked with ice. Kane and Stein (1983b) found that the hydraulic conductivity of ice-rich frozen soil can be several orders of magnitude less than its unfrozen counterpart and that permafrost acts as a relatively impermeable confining layer (*see* **Chapter 74, Soil Hydraulic Properties, Volume 2**). Woo (1986), investigating hydrologically similar areas in Canada, stated that hydraulic activity occurs almost exclusively in the taliks and the active layer. Percolation of soil water to permafrost can effectively eliminate flow to subpermafrost aquifers as the percolated water freezes to seal the soil pore space.

The amount of ice contained within the permafrost is an important condition and influences hydrological processes in several ways. Ground ice ranges in size from the microscopic pore ice (Williams and Smith, 1989) to large ice wedges (Brown, 1967) and on to the massive ice-rich yedoma complexes (Popov, 1983). Segregated ice forms as water migrates across a thermal gradient to form ice lenses within the active layer (Guymon *et al.*, 1984). These ice lenses form parallel to the freezing front, which is, in general, parallel to the ground surface but not necessarily horizontal. Ice wedges are larger ice structures (Figure 3a), which initiate within the active layer but may penetrate tens of meters into the permafrost. These form over many years as soil contracts during cold winters, causing cracks to form within existing ice wedges. As the snow melts during the following spring, water percolates and refreezes in the crack, expanding and pushing the adjacent soil apart (Harry, 1988). Networks of ice wedges are commonly found in marshy floodplains, dry surfaces of alluvial, aeolian, and glaciofluvial sands, or on peat bogs (Romanovskii, 1985) and in fact are found throughout the arctic landscape, on plateaus and hillslopes as well as the low plains. These networks of ice wedges create the characteristic low- or high-centered polygons common throughout the Arctic and occasionally found in the Subarctic. Polygonal networks with the ridges raised relative to the polygon center (low-centered polygons) indicate recently active ice wedges as the soil expands during the summer and experiences upward displacement adjacent to the ice wedge. Polygonal networks where ice wedges have begun to degrade due to disturbance or climate warming will display polygon centers markedly higher than the edges (high-centered polygons). In many low-gradient areas, the microtopography of the polygonal network controls the surface hydrological regime, typically remaining wet to saturated immediately above the ice wedge and in the middle of the low-centered polygons. The polygon rims and middle of high-centered polygons may become quite dry during extended periods of no precipitation.

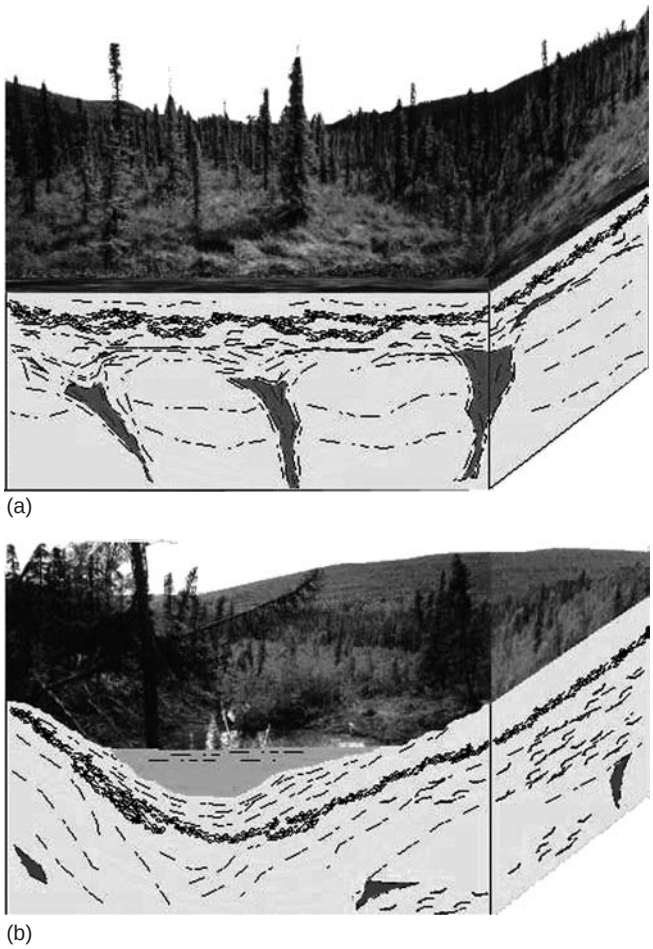


Figure 3 (a,b.) Thermokarst topography forms as ice-rich permafrost thaws, either naturally or anthropogenically, and the ground surface subsides into the resulting voids. The important and dynamic processes involved in thermokarsting include thaw, ponding, surface and subsurface drainage, surface subsidence, and related erosion. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Substantial subsidence of the surface or thermokarst may occur following disturbance of the surface if buried ice is exposed to warmer air temperatures or running water (Figure 3b). Thermokarst lakes are a common feature throughout the Arctic, often forming in response to some disturbance at the surface, such as a change in the surface drainage, removal of vegetation, wildfire, or warming climate. Many of these lakes exist because the talik, common below lakes, does not completely penetrate the permafrost, preventing the lakes from draining vertically. However, if the talik does thaw to the extent that infiltration to the subpermafrost ground water does occur, then the water balance of the lake may change, perhaps shrinking the pond surface area or drying completely (Yoshikawa and Hinzman, 2003) or conversely permitting lake recharge from ground

water (Kane and Slaughter, 1973). Thermokarst can also impact surface hydrological processes even if the talik does not completely penetrate the permafrost. Thermokarsts also improve surface drainage, creating drier conditions at least adjacent to the disturbance, promoting introduction of invasive species (Lloyd *et al.*, 2003) and enabling establishment of different community types according to tolerance of soil saturation (Drury, 1956).

ACTIVE LAYER HYDROLOGY

The soil, peat, or rock above the permafrost that experiences freeze and thaw annually is called the *active* layer (*see* **Chapter 71, Freezing and Thawing Phenomena in Soils, Volume 2**). Almost all of the biological and hydrological processes occur within this thin layer of soil, which ranges from centimeters to a few meters in thickness depending upon local conditions. Throughout the summer and early winter, the active layer progresses from frozen to thawed states or vice versa; consequently the soil volume available for liquid moisture storage is continuously changing. This is an important consideration as one attempts to calculate water balance of permafrost-dominated watersheds. It is sometimes feasible to conduct comparative water balance studies over time periods where the active layer is completely saturated, allowing the soil moisture storage term to be eliminated from the water balance equation (Kane *et al.*, 2000). However, it has been demonstrated that the short-term storage of the active layer can vary if autumn rains do not replenish the deficit that develops because of evapotranspiration through the summer (Woo *et al.*, 1983).

The active layer begins to thaw during the spring snowmelt event and continues to thicken until reaching maximum thickness in about late August or September (Figure 4). In much of the Low Arctic of Alaska and Canada, depending upon local physiography, a porous organic soil overlies a less porous mineral soil. Over the Canadian Shield, this organic layer may be quite thin or absent on the uplands, giving way to bedrock, except for the soil-filled valleys (Spence and Woo, 2002). In many parts of the High Arctic of Canada or Russia, the organic layer may not have developed due to low rates

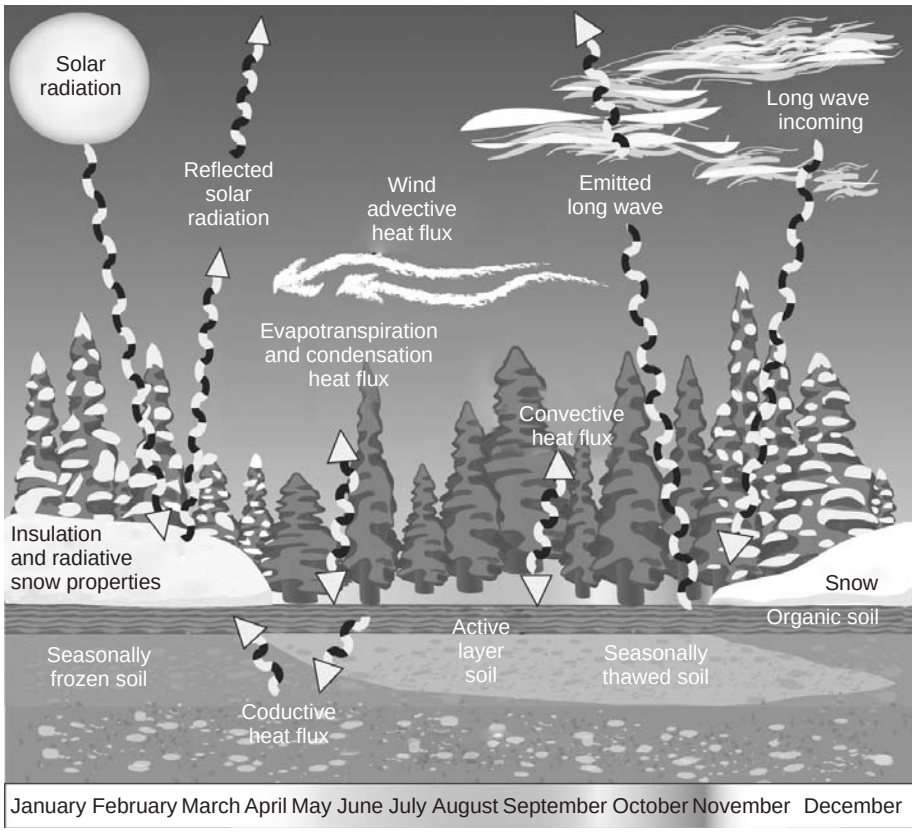


Figure 4 In regions of continuous permafrost, all hydrological processes occur in the active layer, the thin layer of soil above the permafrost that thaws during the summer months. The rate of thawing and refreezing is controlled by the surface energy balance and the thermal properties of the soil and the overlying snow. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of vegetative growth or frequent wildfire occurrence. In the vast peat lands of central Siberia, Subarctic Canada, and Alaska, the organic layers may be several meters thick (Racine and Walters, 1991; Gibson *et al.*, 1993; Smith *et al.*, 2000). The low thermal conductivity and relatively high hydraulic conductivity of the near-surface organic soils (as compared to the underlying mineral soils) exerts a controlling influence upon the thermal and hydrologic regime of the active layer (Hinzman *et al.*, 1991). Spring thawing of the surficial organic layer is initially quite rapid, due primarily to the open pores caused by desiccation of ice during the previous winter. The rate of thawing generally follows a curve that may be approximated by a square root of time (Terzaghi, 1952). Soil freezing primarily occurs from the surface downward, but can also freeze from the permafrost upwards, giving rise to two-sided freezing (Mackay, 1983). Clear, segregated ice lenses may form within the active layer as water migrates to meet the advancing freezing front (Ad Hoc Study Group on Ice Segregation and Frost Heaving, 1984). Likewise, water migration towards the permafrost table during the summer months maintains an ice-rich layer of very low permeability. In the sandy soils of southern Siberia, the active layer can reach several meters in thickness, permitting good lateral drainage in spite of underlying permafrost.

The surface organic layer exerts a strong influence upon hillslope hydrology in the continuous permafrost regions of the low Arctic and the Subarctic. During spring runoff events, when the active layer is still very thin, the rapid component of hillslope runoff occurs either as overland flow, inter-tussock flow, or as pipeflow and matrix flow in the highly organic near-surface soil (Carey and Woo, 2000). Runoff occurring during this period is quite flashy, typically with fast responses and recessions following rain and snowmelt events (Kane *et al.*, 1991). As the active layer thaws throughout the summer, a slower component of subsurface flow occurs in the mineral soil above the ice-rich permafrost table (Hinzman and Kane, 1991). The characteristic hydrographs of streams in permafrost areas display a rapid rise to peak and rapid fall from the peak as near-surface runoff occurs through highly porous organic layers, followed by a more prolonged recession of flow through the deeper layers (Dingman, 1973; McNamara *et al.*, 1998). In regions of continuous permafrost, although an extensive subpermafrost aquifer does exist (Deming *et al.*, 1992), most ground water activity occurs above the permafrost in the suprapermafrost zone. As the active layer continues to develop throughout the summer season, greater volumes of liquid soil water may be temporarily held in storage; as a result, flashy responses are slightly less, peak flows may decrease somewhat, and hydrograph recessions become slightly longer (McNamara *et al.*, 1998; Bolton *et al.*, 2000).

It is difficult to generalize the role of hydrological processes in the dynamics of the thermal regime because of the regional variations in geology, vegetation, and permafrost and because of the low density of stations collecting the extensive data required to complete an energy balance. Ablation is usually the most important hydrologic event annually in permafrost regions, but it is also a threshold event for the surface energy balance (Boike *et al.*, 2003). Dramatic changes occur as the albedo decreases from about 0.8 for the snowpack to 0.2 for the snow-free surface. The active layer begins to thaw immediately after snowmelt (Boike *et al.*, 1998). The surface soil moisture content is a very important parameter controlling many of the processes associated with the surface energy balance. Unfortunately, this is not an easy variable to quantify in the spatial and temporal scales needed to accurately characterize thermal and hydrological interactions over large areas. Surface soil moistures will affect albedo, long-wave emissivity, evapotranspiration, and subsurface heat conduction.

The surface energy balance may be written as

$$Q_h + Q_e + Q_{net} + Q_c + Q_a + Q_m = 0 \qquad (1)$$

where Q_h is sensible heat flux between the surface and the air associated with convection, Q_e is latent heat flux associated with evaporation (sublimation or condensation) (*see* **Chapter 45, Actual Evaporation, Volume 1**), Q_{net} is energy transferred at the surface by net radiation (*see* **Chapter 39, Surface Radiation Balance, Volume 1**), Q_c is energy flux via conduction between the soil surface and the subsurface, Q_a is the energy convected in running water or rainfall, and Q_m is energy utilized for melting of the snowpack (*see* **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**). Heat transfer within ice-rich permafrost masses is almost completely due to conduction, as liquid water and vapor fluxes are generally limited (Romanovsky and Osterkamp, 2000). However, vapor flux from near-surface soil throughout the winter can leave the surface layer quite desiccated (Woo, 1982), and liquid water migration across thermal gradients can create ice lenses within permafrost (Perfect and Williams, 1980). In cases where liquid moisture movement can occur (such as springs, or cracks with flowing water), advective heat transfer will usually dominate (Kane *et al.*, 2001) and can initiate severe permafrost degradation and erosion (Brown and Grave, 1979).

SURFACE WATER HYDROLOGY IN PERMAFROST REGIONS

The connection between surface water and deeper ground water is perhaps the primary distinction between watershed

processes in continuous and discontinuous or permafrost-free regions (Hinzman *et al.*, 2003). The proportion of baseflow in a stream increases as the percentage of permafrost extent decreases (Bolton *et al.*, 2000). Watersheds with larger proportions of permafrost tend to display flashier responses to precipitation events, rising quickly to higher specific discharge peaks and falling rapidly after reaching the peak. Basins with less permafrost allow a greater percentage of precipitation to infiltrate to deeper ground water reservoirs, thus providing a more stable base streamflow between precipitation events or during winter. Additionally, streams in basins with larger proportions of permafrost tend to have greater contributing areas (i.e. areas near the stream contributing water directly to streamflow during storms) as compared to watersheds with smaller amounts of permafrost (Petrone *et al.*, 2000). Although occasionally relatively warm springs do penetrate thick layers of permafrost (Williams and van Everdingen, 1973), the ice-rich permafrost typically isolates surface water from deeper ground water. Consequently, in areas of continuous permafrost, river base flow is extremely limited and small streams often cease flow completely during prolonged periods of low rainfall. Of those rivers with watersheds completely in regions of continuous permafrost, all but the largest typically have zero or very limited discharge during the late winter months. Several of the major north flowing rivers in the Russian Arctic (Ob, Lena, etc.) and the Mackenzie in the Canadian Arctic have their headwaters in permafrost-free areas, providing a stable source of baseflow and thus allowing year round river flow. The smaller coastal rivers in the Alaskan, Canadian, and much of the Siberian Arctic originate within the zone of continuous permafrost, and although several are fed by subpermafrost springs, discharge is quite low during the winter.

Thaw lakes are an important characteristic of surficial hydrology in some areas underlain by continuous permafrost (Carson, 2001). These lakes occupy up to 25% of the Arctic coastal plain of Alaska (Sellmann *et al.*, 1975) and are also common in coastal plain and the silty-clay flat lowlands in the northern territories of Canada. Thaw lakes initially form in thermokarsts when subsidence at the surface creates depressions where water can become ponded (Hopkins, 1949). These ponds absorb more solar radiation and may accumulate heat through advective heat transfer, accelerating the thawing of permafrost and further subsidence. Additionally, the ponds may be enlarged by wave action, which prompts thawing and caving or erosion along the margins. As the lakes grow in size, they occasionally coalesce as lakes intersect each other. Thaw lakes seldom become filled by sediment, but often drain catastrophically when an ice wedge becomes exposed to running water and rapidly melts. In areas where the thaw lake cycle is very active, the age of the lakes is usually less than several thousand years (Sellmann *et al.*, 1975). These lakes may be called *oriented* lakes if the primary axis of the lake length is nearly normal to the dominant wind direction forming a characteristic pattern in response to thermal and mechanical erosion along the edges of the longer ends (Black and Barksdale, 1949). Typically, the lakes are quite shallow, usually less than 2 m and most freeze completely to the bottom (Jeffries *et al.*, 1996) each winter. As the active layer and lakes freeze during the winter, solute exclusion causes the liquid water below the ice to become increasingly brackish and unpotable (Hinzman *et al.*, 1998).

The storage capacity of these lakes plays an important role in summer hydrology of permafrost regions. Pond levels typically drop throughout the summer as evaporation usually exceeds precipitation, although late summer storms occasionally replenish water levels to some degree. In low-gradient watersheds, lateral flows essentially cease when water levels drop below lake outlets (Bowling *et al.*, 2003). Rivers such as those completely within the Arctic Coastal Plain of Alaska transmit little, if any, discharge after snowmelt is complete, as surface outflow from the lakes diminishes because ice-rich permafrost restricts subsurface drainage to rivers. These lakes are replenished annually during spring melt when much of the Coastal Plain is inundated. Thaw ponds of the Mackenzie Delta are similarly replenished during spring melt when the river level rises above the banks and floods the adjacent delta lakes (Marsh, 1986).

Typically in areas of thick, relatively impermeable permafrost, the surface water is effectively isolated from subpermafrost ground water processes. However, in some isolated locations, springs extend through the permafrost to release ground water at the surface (Nelson and Munter, 1990; van Everdingen, 1990). Ground water springs may penetrate thick permafrost; the limiting factors of permafrost thickness and ground temperature must be offset by spring flow rate and water temperature (Liestøl, 1977; Yoshikawa, 1998). In zones of thinner permafrost (tens to hundreds of meters) taliks may extend completely through permafrost allowing ground water recharge or discharge. These taliks allow connection of deeper subpermafrost ground water with surficial water bodies. In some cases (depending upon the thickness of the permafrost), they allow ground water recharge as the water from rivers and lakes infiltrates to the subpermafrost aquifer (Yoshikawa and Hinzman, 2003) or in reverse as ground water discharge (Jorgenson *et al.*, 2001).

GROUND WATER HYDROLOGY IN CONTINUOUS AND DISCONTINUOUS PERMAFROST

Permafrost is the primary factor influencing ground water supply, availability, quality, and distribution (Prowse, 1990; *see* also **Chapter 145, Groundwater as an Element in**

the Hydrological Cycle, Volume 4). Ground water flow in arctic and subarctic regions is usually classified as suprapermafrost or subpermafrost depending upon whether it is flow above or below permafrost (van Everdingen, 1990; *see* also **Chapter 149, Hydrodynamics of Groundwater, Volume 4**). The suprapermafrost ground water is usually quite limited in volume and often subject to surface contamination from anthropogenic compounds or naturally occurring organics. In areas where permafrost is very thick, this source of water can be quite important. Suprapermafrost ground water is responsible for sustaining wetlands and may supply lakes at certain times of the year. A dynamic equilibrium of suprapermafrost ground water exists between wetlands and lakes in the Arctic, sustaining life in both ecosystems (Rovansek *et al.*, 1996; *see* also **Chapter 108, Lake Ecosystems (Stratification and Seasonal Mixing Processes, Pelagic and Benthic Coupling), Volume 3**). Taliks are zones of unfrozen ground that forms below lakes and rivers. These unfrozen islands in permafrost can become water sources for the growth of pingos and are critical for supporting fish habitat in deep pools that do not freeze solid (Nelson and Munter, 1990). These unfrozen zones may extend completely through permafrost under very large lakes. A talik has been used for water supply after it was enlarged by stripping vegetation and creating an impoundment (Feulner and Williams, 1967). If a talik in direct contact with a deep pool that supports fish is used for water supply, it may result in death of the fish unless the water is replenished by flowing water.

Subpermafrost ground water is generally not present in large quantities in the areas in the Arctic where the soils are alluvium (Nelson and Munter, 1990). However, in permafrost areas underlain by carbonate rocks, extensive ground water flow can occur through springs (Sloan and van Everdingen, 1988). Significant springs exist on many of the major river systems in the Alaskan, Canadian, and Russian Arctic. The quality of subpermafrost ground water ranges from relatively fresh ($400\,\mu S\,cm^{-1}$) (Childers *et al.*, 1977) to brine. The temperatures range from hot to cold (Sloan and van Everdingen, 1988). Nelson and Munter state that our knowledge of subpermafrost ground water is poor because: "(1) the population density in the Arctic and Subarctic is relatively low; (2) the expense of drilling through thick permafrost is high; (3) the thickness of permafrost is poorly known in many areas; and (4) nonpotable water occurs at depth in some places."

In the Northern Hemisphere, as one travels south of the zone of continuous permafrost, the permafrost becomes thinner and discontinuous below the ground surface, occurring primarily on north-facing slopes and in valley bottoms. The dominating influence of permafrost on hydrological processes is particularly evident in the zone of discontinuous permafrost. If the active layer is less than approximately one meter, then ice-rich permafrost can maintain wet soils in the rooting zone by limiting vertical percolation. In the discontinuous permafrost zone of Interior Alaska and Canada, distinct differences in ecology and hydrology are often apparent as one crosses a permafrost boundary (Brown, 1969; Viereck *et al.*, 1992; *see* also **Chapter 103, Terrestrial Ecosystems, Volume 3**). Although the rainfall rates are low, ice-rich permafrost maintains moist to saturated conditions in near-surface soils. The organic layer is typically thicker as low temperatures and anaerobic conditions slow decomposition rates. In valley bottoms, typical vegetation includes thick moss mats and other hydrophilic vegetation. On hillslopes underlain by shallow permafrost, thick organic layers with stunted black spruce indicate moist soils. Permafrost-free hillslopes often support deciduous vegetation with only a thin duff layer 10–20 cm in thickness, overlying a mineral soil. The runoff patterns from these contrasting slopes are markedly different (Slaughter and Kane, 1979), and consequently reflect strongly contrasting water chemistry signatures (Figure 5, Petrone *et al.*, 2000; *see* also **Chapter 91, Water Quality, Volume 3**). Soil water that migrates to a stream above a permafrost table is usually enriched in dissolved organic carbon (DOC) and nitrogen (DON). Soil water percolating into permafrost-free slopes infiltrates into a deeper ground water system, where organic carbon, and nitrogen are often sequestered and water becomes enriched with inorganic forms of carbon (DIC) and nitrogen (DIN).

Permafrost not only controls the pathways of ground water flow, but also dictates the zones of recharge. In areas of shallow, discontinuous permafrost, surface water will infiltrate to the permafrost table at which point it will flow along the top of the permafrost until it reaches a permafrost-free area (Hinzman *et al.*, 2000). The ground water recharge occurs in these permafrost-free zones. It is also possible that velocities in these zones may be higher. An understanding of where ground water recharge occurs is perhaps more important than the rate of recharge to successfully quantify ground water flow in discontinuous permafrost regions. Snowmelt is the primary source of ground water recharge in Subarctic Alaska as summer rainfalls typically only meet evapotranspiration losses (Kane and Stein, 1983a; Gieck and Kane, 1986).

Water that is able to infiltrate to deeper ground water in the permafrost-free hillslopes is often older and considerably more mineralized (i.e. higher electrical conductivity) when it enters the stream, as compared to the water following shorter and quicker flowpaths over permafrost or through surface organic layers. The ground water below the permafrost may have increased mineralization due to longer residence times; however, water quality is often greatly influenced by local conditions and it is difficult to generalize (Williams and van Everdingen, 1973). Water below very thick permafrost is usually of lower quality compared to shallow ground water. Ground water pumped

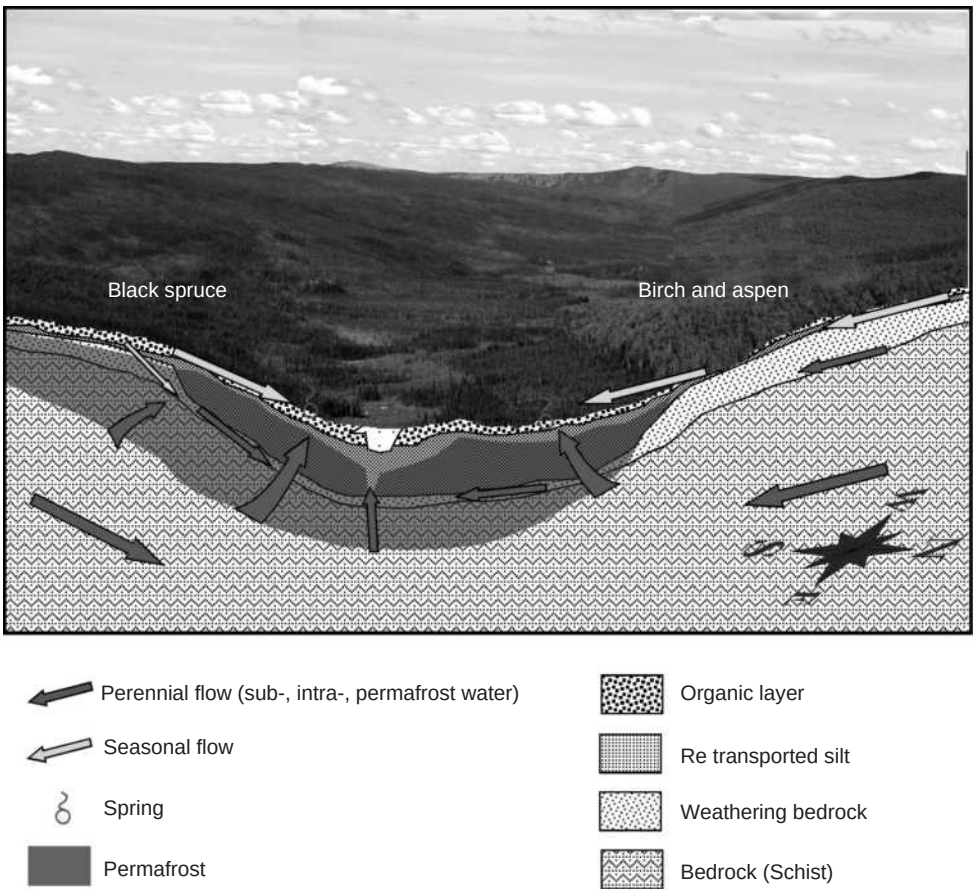


Figure 5 The various pathways of surface and subsurface flow in regions of discontinuous permafrost will create distinctly different water chemistry signatures as influenced by the flow media (*see* **Chapter 91, Water Quality, Volume 3**). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

from alluvium often contains high concentrations of iron and manganese, high organic content (with brownish color and objectionable odor) often with excess hardness. Permafrost may serve to shield ground water in coastal areas from excess salinity and hardness. Wells in regions of continuous permafrost are usually located in the talik of larger rivers, although surface water from lakes, rivers, or streams is usually the predominant source of water for most communities in these areas. In most cases where ground water is extracted from a talik or thaw bulb under a river, the availability may be of limited extent (Isaacs, 2000). In large floodplains, wells may be drilled to a depth below the permafrost with some success (Trimble *et al.*, 1983). Recently, some Alaskan villages along the major river systems are utilizing the relatively new technology of directional drilling to establish wells directly beneath a river ensuring ample supply of good quality water.

Ground water springs are frequently found near the upslope boundary of discontinuous permafrost where it interrupts the subsurface flow paths (Sloan *et al.*, 1975). Ground water may also be forced to the surface when the seasonally frozen layer freezes down to depths where it meets the permafrost, again blocking subsurface flow paths (van Everdingen, 1988). In both situations, large icings (also called *aufeis* or *naleds*) can form during the winter (Washburn, 1979; Kane, 1981). These icings may become a significant source of flow augmentation during the summer and may provide stable base flow in some basins (Reedyk *et al.*, 1995). Their locations are controlled by local geology, terrain, and thermal characteristics of discharge water. In some cases, aufeis deposits source from near-surface water supplies (such as in a talik above permafrost) or form at the permafrost boundary. In other cases, these aufeis fields develop from springs that perforate the permafrost. These subterranean channels remain open (unfrozen) only because of the combination of adequate flow rates and warm temperature. The areal extent of the aufeis is directly related to the ground water discharge rates (Hall and Roswell, 1981).

Extensive layers of permafrost in valleys can act as confining layers for ground water aquifers occasionally creating artesian conditions (Figure 6). Although this is true for continuous and discontinuous permafrost regions, it is more problematic in the warmer discontinuous permafrost of the

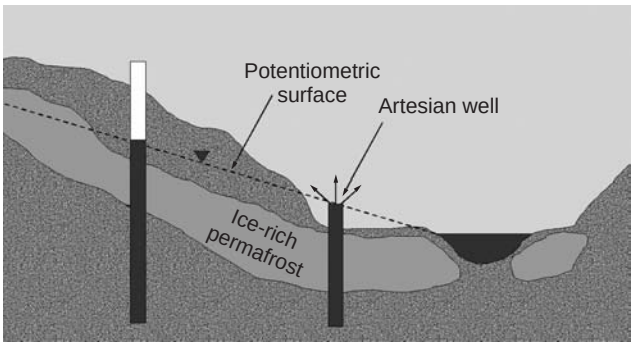


Figure 6 Permafrost often acts as an impermeable or confining layer that may induce natural spring formation. Wells drilled through substantial permafrost in valley bottoms frequently encounter artesian conditions. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Subarctic. This is partially due to the fact that few ground water wells penetrate the continuous permafrost but also because ground water in the discontinuous permafrost zone experiences far greater recharge. Soil moisture may readily percolate through the permafrost-free south-facing slopes to encounter a confining layer along north-facing slopes (White *et al.*, 2002). Wells penetrating the permafrost in such valley bottoms may encounter substantial positive water pressures. A well drilled in a valley near Fairbanks, Alaska yielded artesian pressures 8.45 m above the ground surface (Linell, 1973). Lakes and large streams lower in the valley enabled development of taliks that penetrated the permafrost and allowed natural discharge of ground water to the surface (Kane and Slaughter, 1973).

Movement of water within permafrost masses is restricted to very thin films along soil grains, even at temperatures well below freezing, the amount being a function of soil type and temperature (Anderson and Morgenstern, 1973). Although this water movement can, over long time periods, be important in its effect upon roads or houses, from a watershed water balance viewpoint, ice-rich permafrost is usually considered impermeable to water. Dry permafrost does exist and may be very permeable to contaminants such as hydrocarbons or other chemicals.

PERMAFROST CONTROL ON WATERSHED MORPHOLOGY

Permafrost exerts a dominant control over the physical character of arctic watersheds by greatly restricting fluvial erosion. The drainage networks of some basins in the Alaskan Arctic tend to be less developed as compared to those in more temperate regions (McNamara *et al.*, 1999). Watersheds in temperate regions tend to evolve towards a dendritic drainage pattern; however, headwater watersheds in regions of continuous permafrost seem to be locked in a less-developed state. Small drainage features called *water tracks* form on hillsides to rapidly transport water straight down the hillside directly to the stream (Hastings *et al.*, 1989). These water tracks may be quite subtle in their general morphology, often being distinguishable only by slight changes in dominant vegetation type. Numerous water tracks are generally parallel to each other, perpendicular to the stream channel, and spaced on the order of tens of meters apart (Hinzman *et al.*, 1993). The channel network may more closely resemble a comb, as opposed to the treelike networks typically encountered. Interhummock flow (Quinton and Marsh, 1998, 1999), often observed in areas of tussock tundra, behaves in a similar manner yielding large drainage densities and surface storage, but less runoff efficiency and longer recession periods following rain events. Pipeflow, or runoff within small tunnels between the surface organic soil and underlying mineral soils, has been observed in highly organic soils of Subarctic Canada (Carey and Woo, 2000). This runoff mechanism is highly efficient, but ephemeral, as this flow only occurs when the water table is above the level of the pipes. As the soils are usually frozen during the annual peak flow events (spring melt), and the high density of channels prevents large runoff volumes from causing significant erosion, these channels may not develop incised channels nor yield significant sediment if not disturbed. Conversely, if the permafrost on hillsides or stream banks is disturbed and experiences additional thawing, then thermal and mechanical erosion may be rapid and substantial (Viereck, 1982; Gautier *et al.*, 2003). Other features of channelized water conveyance, including gullies, rills, the patterned ground called *stripes* (Hall, 1994), slump scars, and thermokarst subsidence conduits are observed throughout the Arctic, particularly in areas without a vegetation cover. These drainage features form in response to erosional processes similar to those that occur in nonpermafrost areas, but thermal processes also play a role in controlling mechanical properties. Running water can accelerate thermal and mechanical erosion of frozen soils, rapidly accelerating sediment transport and substantial morphological changes (Gautier *et al.*, 2003). It has also been noted that permafrost-dominated areas may be particularly prone to increased sedimentation following wildfire or under a warmer climate (Huscroft *et al.*, 2004).

Headwater streams in the low Arctic regions over continuous permafrost typically freeze solidly to the bottom each winter. This bedfast ice offers rigid protection from the spring melt flood event; however, some rafting of cobbles and pebbles can occur as the ice breaks free of the bed. As one moves out of the headwaters to channels that drain a larger watershed area, eventually channels become deep enough to not completely freeze during the winter. These channels are particularly susceptible to erosive forces and mechanical damage of large ice floes during the spring melt

event (McNamara, 2000; Prowse and Ferrick, 2002). Additionally, the lack of complete freezing in the river (or lakes) maintains taliks above the permafrost (Arcone, 1998). In the high Arctic though, with the exception of larger rivers, most arctic streams do not have any ice in the channels because the limited drainage from the thin active layer and the absence of baseflow leaves channels dry in autumn. Most arctic streams would be snow-filled, often with snow dams, and their breakup follows the fashion described in Woo and Sauriol (1980, 1981).

CONTAMINANT TRANSPORT

Permafrost, particularly discontinuous permafrost, makes hydrologic analyses of contaminant transport challenging (*see* **Chapter 91, Water Quality, Volume 3**), both in field studies and in computer model simulations. Permafrost was shown to clearly affect the water and contaminant migration pathways in a study of contaminant transport in a region of discontinuous permafrost near Fairbanks, Alaska (Hinzman *et al.*, 2000; *see* also **Chapter 153, Groundwater Pollution and Remediation, Volume 4**). Thawed channels within permafrost fields existed due to the presence of a talik that was formed under an ancient river slough. The down-gradient concentration of contaminants is not necessarily diminished by the permafrost masses, but rather is more channelized. This is important in field investigations with a limited number of wells. If the permafrost configuration is not understood, the contaminant plumes may be missed because of being channeled in directions different from those indicated by area-wide hydraulic gradients. Numeric simulations of contaminant transport modeling in areas of discontinuous permafrost can lead to model instabilities due to the drastic changes in hydraulic properties over short distances (Johnson *et al.*, 1996). Ice-rich permafrost may protect subpermafrost aquifers from contamination at the point of impact (*see* **Chapter 91, Water Quality, Volume 3**) (Collins *et al.*, 1994); however, downslope migration of fuel spills may allow such contamination if the spill is not effectively contained. Soil structure and soil moisture content greatly influence contaminant infiltration into frozen ground (White and Williams, 1999). Dry permafrost may not act as a barrier to infiltration or migration by contaminants such as fuels with low freezing points.

CONCLUSIONS

Permafrost exerts a dominant influence on hydrologic processes in arctic, subarctic, and high alpine regions. The predominant role of permafrost is acting as a relatively impermeable layer; in some cases, enhancing near-surface soil moisture, sometimes creating perched lakes, frequently confining deeper ground water aquifers that discharge through taliks developed under lakes and rivers. In continuous permafrost regions, hillside hydrology is dominated by the properties of the shallow active layer, generally characterized by limited storage that increases as it thaws throughout the summer. River discharges in the continuous permafrost regions are generally flashy with rapid rise and falls from peak flows and long recessions to limited or no base flow. Snowmelt typically produces the dominant runoff event each year. Winter discharge in rivers that source completely within continuous permafrost is almost zero. The influence of permafrost is perhaps most apparent in the discontinuous zones due to the sharp contrasts between permafrost-free areas and those underlain with shallow permafrost. These differences become less distinct as the active layer becomes thicker, particularly when a talik or unfrozen layer is maintained above the permafrost throughout the winter. This unfrozen layer permits subsurface drainage throughout the year and initiates drier conditions at the surface. River discharges in the discontinuous permafrost regions are generally characterized by considerable baseflow, typically flowing throughout the winter.

Permafrost binds fine-grained soils together, particularly where organic matter is incorporated into the frozen ground, lowering the erodibility of fine-grained soils. Sediment transport in permafrost regions depends on the type of soil and the interaction among ground ice, soil materials, and thermoerosional processes, in some cases limiting bank erosion and channel network development. Running water can accelerate thermal and mechanical erosion of frozen soils, rapidly accelerating sediment transport and substantial morphological changes.

In the Subarctic to low Arctic, the cold, moist conditions maintained by permafrost often creates conditions unfavorable to aerobic decomposition resulting in accumulation of organic matter ranging in thickness from centimeters to meters. This organic layer acts as an insulative mat, thermally protecting the underlying permafrost and maintaining many of the unique characteristics of permafrost hydrology. The hydraulic conductivity of the unfrozen porous organic soils often contrast markedly with the underlying mineral soils, which, when unfrozen, also differ greatly in properties from permafrost. These sharp distinctions in properties shape much of the hydrologic response of basins influenced by permafrost. In the high Arctic and Antarctic, growing seasons are more limited and development of organic soils is greatly diminished. There, hillslope hydrology during the thaw season is strongly influenced by limited storage and melting of active layer ice.

REFERENCES

Ad Hoc Study Group on Ice Segregation and Frost Heaving (1984) *Ice Segregation and Frost Heaving*, U.S. National Research Council, National Academy Press: Washington, p. 72.

Anderson D.M. and Morgenstern N.R. (1973) *Physics, Chemistry, and Mechanics of Frozen Ground: A Review. Permafrost: North American Contribution to the Second International Conference*, National Academy of Sciences: Washington, pp. 257–288.

Arcone S.A. (1998) Seasonal structure of taliks beneath Arctic streams determined with ground-penetrating radar. In *Permafrost: Seventh International Conference, June 23-27, 1998: Proceedings*, Lewkowicz A.G. and Allard M. (Eds.), Centre d'études nordiques, Laval University, Yellowknife, pp. 19–24.

Black R.F. and Barksdale W.L. (1949) Oriented lakes of northern Alaska. *Journal of Geology*, **51**, 105–118.

Boike J., Hinzman L.D., Overduin P.P., Romanovsky V., Ippisch O. and Roth K. (2003) A comparison of snow melt at three circumpolar sites: Spitsbergen, Siberia, Alaska. *Proceedings of 8th International Conference on Permafrost*, Zurich, pp. 79–84, 21–25 July 2003.

Boike J., Roth K. and Overduin P.P. (1998) Thermal and hydrologic dynamics of the active layer at a continuous permafrost site (Taymyr Peninsula, Siberia). *Water Resources Research*, **34**(3), 355–363.

Bolton W.R., Hinzman L.D. and Yoshikawa K. (2000) Stream flow studies in a watershed underlain by discontinuous permafrost. In *Water Resources in Extreme Environments*, Kane D.L. (Ed.), American Water Resources Association Proceedings: Anchorage, pp. 31–36, 1–3 May 2000.

Bowling L.C., Kane D.L., Gieck R.E., Hinzman L.D. and Lettenmaier D.P. (2003) The role of surface storage in a low-gradient arctic watershed. *Water Resources Research*, **39**(4), 1087, doi:10.1029/2002WR001466.

Brown J. (1967) Tundra soils formed over ice wedges, Northern Alaska. *Soil Science Society of America*, **31**(5), 86–691.

Brown J., Ferrians O.J. Jr, Heginbottom J.A. and Melnikov E.S. (1998) *Circum-Arctic Map of Permafrost and Ground-Ice Conditions*, National Snow and Ice Data Center/World Data Center for Glaciology: Boulder, Digital Media.

Brown R.J.E. (1969) *Factors Influencing Discontinuous Permafrost in Canada*, Péwé T.L. (Ed.), The periglacial environment, past and present: Arctic Institute of North America: Ottawa, pp. 11–53.

Brown R.J.E. (1970) *Permafrost in Canada*, University of Toronto Press, p. 231.

Brown R.J.E. (1978) *Permafrost*, Hydrological Atlas of Canada. Department of Fisheries and Environment: Ottawa, Plate 32. Map.

Brown J. and Grave N.A. (1979) *Physical and Thermal Disturbance and Protection of Permafrost*, CRREL Special Report 79–5, U.S. Army Cold Regions Research and Engineering Laboratory, p. 42.

Carey S.K. and Woo M.K. (2000) The role of soil pipes as a slope runoff mechanism, subarctic Yukon, Canada. *Journal of Hydrology*, **233**, 206–222.

Carson C.E. (2001) *The Oriented Thaw Lakes: A Retrospective*, Norton D.W. (Ed.), Fifty more years below zero; tributes and meditations for the Naval Arctic Research Laboratory's first half century at Barrow: Alaska, pp. 129–137.

Childers J.M., Nauman J.W., Kernodle D.R. and Doyle P.F. (1977) *Water Resources Along the Taps Route, 1970-74*, Open-File Report 78-137, U.S. Geological Survey, p. 136.

Collins C.M., Racine C.H. and Walsh M.E. (1994) The physical, chemical, and biological effects of crude oil spills after 15 years on a black spruce forest, interior Alaska. *Arctic*, **47**(2), 164–175.

Deming D., Sass J.H., Lachenbruch A.H. and De Rito R.F. (1992) Heat flow and subsurface temperature as evidence for basin-scale ground-water flow, North Slope of Alaska. *Geological Society of America. Bulletin*, **104**(5), 528–542.

Dingman S.L. (1973) Effects of permafrost on stream flow characteristics in the discontinuous permafrost zone of central Alaska. *Permafrost: North American Contribution to the Second International Conference*, National Academy of Sciences: Washington, pp. 447–453.

Drury W.H. (1956) *Bog Flats and Physiographic Processes in the Upper Kuskokwim River Region, Alaska*, Gray Herbarium of Harvard University: Cambridge, pp. 130, Contribution No.178.

Edlund S.A., Woo M.-k and Young K.L. (1990) Climate, hydrology and vegetation patterns, hot weather creek, Ellesmere Island, Arctic Canada. *Nordic Hydrology*, **21**, 273–286.

Feulner A.J. and Williams J.R. (1967) *Development of a Ground-Water Supply at Cape Lisburne, Alaska, by Modification of the Thermal Regime of Permafrost*, U.S. Geological Survey: Professional Papers, No.575-B, pp. 199–202.

Gautier E., Brunstein D., Costard F. and Lodina R. (2003) Fluvial dynamics in a deep permafrost zone: the case of the Middle Lena river (central Siberia). *Eighth International Conference on Permafrost*, Zurich, pp. 271–275, July 21–25, 2003.

Gibson J.J., Edwards T.W.D. and Prowse T.D. (1993) Runoff generation in a high boreal wetland in northern Canada. *Nordic Hydrology*, **24**(2/3), 213–224.

Gieck R.E. Jr and Kane D.L. (1986) Hydrology of two subarctic watersheds. In *Proceedings of the Cold Regions Hydrology Symposium*, Kane D.L. (Ed.), American Water Resources Association: Fairbanks, pp. 283–291.

Grigoriev V.Yu. and Sokolov B.L. (1994) Northern hydrology in the former Soviet Union. In *Northern Hydrology: International Perspectives*, Prowse T.D., Ommanney C.S.L. and Watson L.E. (Eds.), NHRI Science Report No. 3. NHRI, pp. 147–179.

Guymon G.L., Hromadka T.V. and Berg R.L. (1984) Two-dimensional model of coupled heat and moisture transport in frost heaving soils. *Proceedings of the Third International Offshore Mechanics and Arctic Engineering Symposium*, American Society of Mechanical Engineers: New York, pp. 91–98; *Journal of Energy Resources Technology*.

Hall D.K. and Roswell C. (1981) The origin of water feeding icings on the eastern north slope of Alaska. *Polar Record*, **20**(128), 433–438.

Hall K. (1994) Some observations regarding sorted stripes, Livingston Island, South Shetlands. *Permafrost and Periglacial Processes*, **5**(1), 119–126.

Harry D.G. (1988) Ground ice and permafrost. *Advances in Periglacial Geomorphology*, John Wiley & Sons: Chichester, pp. 113–149.

Hastings S.J., Luchessa S.A., Oechel W.C. and Tenhunen J.D. (1989) Standing biomass and production in water drainages of the foothills of the Philip Smith Mountains, Alaska. *Holarctic Ecology*, **12**(3), 304–311.

Hinzman L.D., Johnson R.A., Kane D.L., Farris A.M. and Light G.J. (2000) Measurements and modeling of benzene transport in

a discontinuous permafrost region. In *Contaminant Hydrology: Cold Regions Modeling*, Alex Iskandar and Steven Grant (Eds.), Lewis Publishers, pp. 175–237.

Hinzman L.D. and Kane D.L. (1991) Snow hydrology of a headwater arctic basin 2, conceptual analysis and computer modeling. *Water Resources Research*, **27**(6), 1111–1121.

Hinzman L.D., Kane D.L., Benson C.S. and Everett K.R. (1991) Hydrologic and thermal properties of the active layer in the Alaskan Arctic. *Cold Regions Science and Technology*, **19**(2), 95–110.

Hinzman L.D., Kane D.L. and Everett K.R. (1993) Hillslope hydrology in an Arctic setting. *Sixth International Conference on Permafrost*, Beijing, 5–9 July, 1993, pp. 267–271.

Hinzman L.D., Kane D.L., Yoshikawa K., Carr A., Bolton W.R. and Fraver M. (2003) Hydrological variations among watersheds with varying degrees of permafrost, 2003. *Proceedings of 8th International Conference on Permafrost*, Zurich, pp. 407–411, 21–25 July 2003.

Hinzman L.D., Robinson D.W. and Kane D.L. (1998) *A Biogeochemical Survey of an Arctic Coastal Wetland*, Seventh International Conference on Permafrost. Yellowknife, Canada. June 1998. pp. 459–464.

Hopkins D.M. (1949) Thaw lakes and thaw sinks in the imuruk lake area Seward Peninsula, Alaska. *Journal of Geology*, **57**, 119–131.

Huscroft C.A., Lipovsky P. and Bond J.D. (2004) Permafrost and landslide activity: case studies from southwestern Yukon Territory. In *Yukon Exploration and Geology 2003*, Emond D.S. and Lewis L.L. (Eds.), Yukon Geological Survey, pp. 107–119.

Isaacs J. (2000) A comprehensive approach to planning for sanitation facilities in Buckland, Alaska. In *Research and Development Conference on Rural Sanitation*, Woolard C.R., Woolard L.A. and White D.M. (Eds.), Alaska Water and Wastewater Management Association: Fairbanks.

Ishii Y., Ishikawa N., Kobayashi D., Kodama Y., Nomura M. and Sato N. (2001) *Characteristics of Summer Water Balance in Eastern Siberian Tundra Watershed*, 23rd Symposium on Polar Meteorology and Glaciology: Nov. 29-30, 2000: Polar Meteorology and Glaciology, 15, pp. 91–106.

Jeffries M.O., Morris K. and Liston G.E. (1996) A method to determine lake depth and water availability on the north slope of Alaska with spaceborne imaging radar and numerical ice growth modelling. *Arctic*, **49**(4), 367–374.

Jin H., Li S., Cheng G., Wang S. and Li X. (2000) Permafrost and climatic change in China. *Global and Planetary Change*, **26**(4), 387–404.

Johnson R.A., Kane D.L., Hinzman L.D., Light G. and Farris A.M. (1996) Modeling of contaminant transport in ground water in regions of discontinuous permafrost. In *Proceedings of the Eighth International Conference on Cold Regions Engineering*, Carlson R.F. (Ed.), American Society of Civil Engineers, pp. 82–93.

Jorgenson M.T., Racine C.H., Walters J.C. and Osterkamp T.E. (2001) Permafrost degradation and ecological changes associated with a warming climate in central Alaska. *Climatic Change*, **48**, 551–579.

Kane D.L. (1980) Snowmelt infiltration into seasonally frozen soils. *Cold Regions Science and Technology*, **3**, 153–161.

Kane D.L. (1981) Physical mechanics of aufeis growth. *Canadian Journal of Civil Engineering*, **8**, 186–195.

Kane D.L. and Boike J. (2002) Permafrost: soil temperature/special problems. *Encyclopedia of Soil Science*, Marcel Dekker: pp. 972–975.

Kane D.L., Hinkel K.M., Goering D.J., Hinzman L.D. and Outcalt S.I. (2001) Non-conductive heat transfer associated with frozen soils. *Global and Planetary Change*, **29**, 275–292.

Kane D.L., Hinzman L.D., Benson C.S. and Liston G.E. (1991) Snow hydrology of a headwater arctic basin 1, physical measurements and process studies. *Water Resources Research*, **27**(6), 1099–1109.

Kane D.L., Hinzman L.D., McNamara J.P., Zhang Z. and Benson C.S. (2000) An overview of a nested watershed study in Arctic Alaska. *Nordic Hydrology*, **31**(4/5), 245–266.

Kane D.L., McNamara J.P., Yang D., Olsson P.Q. and Gieck R.E. (2003) An extreme rainfall/runoff event in Arctic Alaska. *Journal of Hydrometeorology*, **4**(6), 1220–1228.

Kane D.L. and Slaughter C.W. (1973) Recharge of a central Alaska Lake by subpermafrost ground water. *Permafrost: North American Contribution to the Second International Conference*, National Academy of Sciences: Washington, pp. 458–462.

Kane D.L. and Stein J. (1983a) *Field Evidence of Ground Water Recharge in Interior Alaska in Proceedings of Permafrost: 4th International Conference*, National Academy Press: Washington, pp. 572–577.

Kane D.L. and Stein J. (1983b) Water movement into seasonally frozen soils. *Water Resources Research*, **19**, 1547–1557.

Kane D.L. and Yang D. (2004) Overview of water balance determinations for high latitude watersheds. *Northern Research Basins Water Balance*, (*Proceedings of a Workshop held at Victoria*, Canada, March 2004), IAHS Publication No. 290, IAHS.

Liestøl O. (1977) Pingos, Springs, and Permafrost in Spitsbergen. *Norsk Polarinstitut Arbok*, 7–29.

Linell K.A. (1973) Risk of uncontrolled flow from wells through permafrost. *Permafrost: North American Contribution to the Second International Conference*, National Academy of Sciences: Washington, pp. 462–468.

Lloyd A.H., Yoshikawa K., Fastie C.L., Hinzman L.D. and Fraver M. (2003) Effects of permafrost degradation on woody vegetation at arctic treeline on the Seward Peninsula, Alaska. *Permafrost and Periglacial Processes*, **14**(2), 93–102.

Lyons W.B., Welch K.A., Nezat C.A., Crick K., Toxey J.K., Mastrine J.A., McKnight D.M. (1997) Chemical weathering rates and reactions in the Lake Fryxell Basin, Taylor Valley: comparison to temperate river basins. Lyons W.B., Howard-Williams C. and Hawes I. (Eds.), *Ecosystem Processes in Antarctic Ice-free Landscapes*, *Proceedings of an International Workshop on Polar Desert Ecosystems*, Christchurch, pp. 147–154, 1–4 July 1996.

Mackay J.R. (1983) *Active Layer Growth, Illisarvik Experimental Drained Lake Site, Richards Island, NWT*, Geological Survey of Canada, Paper, No. 82-1A.

Mann D.H., Peteet D.M., Reanier R.E. and Kunz M.L. (2002) Responses of an arctic landscape to late glacial and early holocene climatic changes: the importance of moisture. *Quaternary Science Reviews*, **21**(8-9), 997–1021.

Marsh P. (1986) Modelling water levels for a lake in the Mackenzie Delta. In *Proceedings of Cold Regions Hydrology Symposium*, Kane D.L. (Ed.), American Water Resources Association: Fairbanks, pp. 23–29.

Marsh P. (1990) Snow hydrology. In *Northern Hydrology: Canadian Perspective*, Prowse T.D. and Ommanney C.S.L. (Eds.), NHRI Science Report 1.NHRI, pp. 37–61.

McNamara J.P. (2000) Bankfull flow, hydraulic geometry, and river ice in a northern river. In *Water Resources in Extreme Environments*, Kane D.L. (Ed.), American Water Resources Association Proceedings: Anchorage, pp. 191–196, 1–3 May 2000.

McNamara J.P., Kane D.L. and Hinzman L.D. (1998) An analysis of streamflow hydrology in the Kuparuk river basin, Arctic Alaska: a nested watershed approach. *Journal of Hydrology*, **206**, 39–57.

McNamara J.P., Kane D.L. and Hinzman L.D. (1999) An analysis of an arctic channel network using a digital elevation model. *Geomorphology*, **29**, 339–353.

Nelson G.L. and Munter J.A. (1990) Ground Water. In *Cold Regions Hydrology and Hydraulics*, Technical Council on Cold Regions Engineering Monograph, Ryan W.L. and Crissman R.D. (Eds.), American Society of Civil Engineers: New York, pp. 317–348.

Osterkamp T.E., Peterson J.K. and Collet T.S. (1985) Permafrost thickness in the Oliktok Point, Prudhoe Bay, and Mikkelson Bay area of Alaska. *Cold Regions Science and Technology*, **11**, 99–105.

Paepe R. and Van Overloop E. (2001) Permafrost equivalents from boreal to tropical zones. *NATO Advanced Research Workshop on Permafrost Response on Economic Development, Environmental Security and Natural Resources: November 12–16, 1998: Novosibirsk, Russian Federation*, NATO Science Series. Partnership Sub-Series 2, Environmental Security. 76; Kluwer Academic Publishers: Dordrecht, pp. 151–184.

Perfect E. and Williams P.J. (1980) Thermally induced water migration in frozen soils. *Cold Regions Science and Technology*, **3**(2 and 3), 101–109.

Permafrost Subcommittee (1988) *Glossary of Permafrost and Related Ground-Ice Terms*, Technical Memorandum No. 142, National Research Council of Canada, pp. 156.

Petrone K.C., Hinzman L.D. and Boone R.D. (2000) Nitrogen and carbon dynamics of storm runoff in three sub-arctic streams. In *Water Resources in Extreme Environments*, Kane D.L. (Ed.), American Water Resources Association Proceedings: Anchorage, pp. 167–172, 1–3 May 2000.

Popov A.I. (1983) Genesis and conditions of formation of the yedoma cryogenic sedimentary complex on the coastal plains of the sub-arctic. *Polar Geography and Geology*, **7**(4), 281–289.

Prowse T. (1990) Northern hydrology: an overview. In *Northern Hydrology: Canadian Perspective*, Prowse T. and Ommanney S. (Eds.), NHRI Science Report 1, NHRI, pp. 1–36.

Prowse T.D. and Ferrick M.G. (2002) Hydrology of ice-covered rivers and lakes: scoping the subject. *Hydrological Processes*, **16**(4), 759–762.

Quinton W.L. and Gray D.M. (2003) Subsurface drainage from organic soils in permafrost terrain: the major factors to be represented in a runoff model. *Eighth International Conference on Permafrost*, Vol. 2 Zurich, pp. 917–922, July 21–25, 2003.

Quinton W.L. and Marsh P. (1998) The influence of mineral earth hummocks on subsurface drainage in the continuous permafrost zone. *Permafrost and Periglacial Processes*, **9**(3), 213–228.

Quinton W.L. and Marsh P. (1999) *A Conceptual Framework for Runoff Generation in a Permafrost Environment*, Hydrological Processes 13, pp. 2563–2581.

Racine C.H. and Walters J.C. (1991) *Ground Water-discharge Wetlands in the Tanana Flats, Interior Alaska*, CRREL Report (U.S. Army Cold Regions Research and Engineering Laboratory) No. 91-14, U.S. Army Cold Regions Research and Engineering Laboratory, p. 10.

Reedyk S., Woo M.K. and Prowse T.D. (1995) Contribution of icing ablation to streamflow in a discontinuous permafrost area. *Canadian Journal of Earth Sciences*, **32**(1), 1–12.

Romanovskii N.N. (1985) Distribution of recently active ice and soil wedges in the USSR. In *Field and Theory; Lectures in Geocryology*, Church M.A. (Ed.), University of British Columbia Press: Vancouver, pp. 154–165.

Romanovsky V.E. and Osterkamp T.E. (2000) Effects of unfrozen water on heat and mass transport processes in the active layer and permafrost. *Permafrost and Periglacial Processes*, **11**(3), 219–239.

Rovansek R.J., Hinzman L.D. and Kane D.L. (1996) Hydrology of a tundra wetland complex on the Alaskan Arctic coastal plain. *Arctic and Alpine Research*, **28**(3), 311–317.

Sellmann P.V., Brown J., Lewellen R.I., McKim H.L. and Merry C.J. (1975) *Classification and Geomorphic Implications of Thaw lakes on the Arctic Coastal Plain, Alaska*, CRREL Research Reports, No. 344, U.S. Army Cold Regions Research and Engineering Laboratory, p. 21.

Slaughter C.W. and Kane D.L. (1979) Hydrologic role of shallow organic soils in cold climates. *Proceedings of the Canadian Hydrology Symposium: 79–Cold Climate Hydrology*, Associate Committee on Hydrology, National Research Council: Vancouver, pp. 380–389, May 10–11, 1979.

Sloan C.E. and van Everdingen R.O. (1988) Region 28, permafrost region. In *Hydrogeology*, *The Geology of North America, Vol 0–2*, Back W., Rosenshein J.S. and Seaber P.R. (Eds.), Geological Society of America, pp. 263–270.

Sloan C.E., Zenone C. and Mayo L.R. (1975) *Icings Along The Trans-Alaska Pipeline Route*, Open-file Report 75–87, U.S. Geological Survey, p. 39.

Smith L.C., MacDonald G.A., Frey K.E., Velichko A., Kremenetski K., Borisova O., Dubinin P. and Forster R.R. (2000) U.S.-Russian venture probes Siberian peatlands' sensitivity to climate. *Eos*, **81**(43), 497–504.

Spence C. and Woo M.K. (2002) Hydrology of subarctic Canadian shield: bedrock upland. *Journal of Hydrology*, **262**(1–4), 111–127.

Terzaghi K. (1952) Permafrost. *Journal Boston Society of Civil Engineers*, **39**, 319–368.

Trimble J., Grainger J. and Glen P. (1983) A sub-permafrost ground water supply for the community of Old Crow, Yukon. In *Proceedings of the First Conference on Cold Regions Environmental Engineering*, Tilsworth T. and Smith D.W. (Eds.), University of Alaska: Fairbanks, pp. 149–179.

Trombotto D., Buk E. and Hernández J. (1997) Monitoring of mountain permafrost in the Central Andes, Cordon del Plata,

Mendoza, Argentina. *Permafrost and Periglacial Processes*, **8**(1), 123–129.

van Everdingen R.O. (1988) Perennial discharge of subpermafrost ground water in two small drainage basins, Yukon, Canada. *Proceedings of the Fifth International Conference on Permafrost*, Tapir Publishers: Trondheim, pp. 639–643, August 2–5, 1988.

van Everdingen R.O. (1990) Ground-water hydrology. In *Northern Hydrology: Canadian Perspectives*, *Proceedings of the Northern Hydrology Symposium 10–12 July 1990*, Prowse T.D. and Ommanney C.S.L. (Eds.), National Hydrology Research Institute: Saskatoon, pp. 77–101.

Viereck L.A. (1982) Effects of fire and firelines on active layer thickness and soil temperatures in interior Alaska. In *The Roger J.E. Brown Memorial Volume: Proceedings of the Fourth Canadian Permafrost Conference*, Calgary, Alberta, March 2–6, 1981 / Edited by No. 20124, French H.M. (Ed.), National Research Council of Canada, pp. 123–135.

Viereck L.A., Dyrness C.T., Batten A.R. and Wenzlick K.J. (1992) *The Alaska Vegetation Classification*, Pacific Northwest Research Station, Portland, p. 278.

Wang B.L. and French H.M. (1995) Permafrost on the Tibet Plateau, China. *Quaternary Science Reviews*, **14**(3), 255–274.

Washburn L.A. (1979) Geocryology. *A survey of periglacial process and environments*, E. Arnold: London.

White T.L. and Williams P.J. (1999) The influence of soil microstructure on hydraulic properties of hydrocarbon-contaminated freezing ground. *Polar Record*, **35**(192), 25–32.

White D.M., Yoshikawa K. and Garland D.S. (2002) Use of dissolved organic matter to support hydrologic investigations in a permafrost-dominated watershed. *Cold Regions Science and Technology*, **35**, 27–33.

Williams P. and Smith M. (1989) *The Frozen Earth*, Cambridge University Press, p. 306.

Williams J.R. and van Everdingen R.O. (1973) Ground water investigations in permafrost regions of North America: a review. *Permafrost: North American Contribution to the Second International Conference*, National Academy of Sciences: Washington, pp. 435–446.

Woo M.K. (1982) Upward flux of vapor from frozen materials in the high Arctic. *Cold Regions Science and Technology*, **5**, 269–274.

Woo M.K. (1986) Permafrost hydrology in North America. *Atmosphere-Ocean*, **24**(3), 201–234.

Woo M.K. (1990) Permafrost hydrology. In *Northern Hydrology: Canadian Perspective*, Prowse T.D. and Ommanney C.S.L. (Eds.), NHRI Science Report 1, NHRI, pp. 63–76.

Woo M., Marsh P. and Steer P. (1983) Basin water balance in a continuous permafrost environment. *Fourth International Conference on Permafrost*, National Academy Press: Fairbanks, pp. 1407–1411, July 17–22, 1983.

Woo M.K. and Sauriol J. (1980) Channel development in snow-filled valleys, Resolute, Northwest Territories, Canada. *Geografiska Annaler*, **62A**, 37–56.

Woo M.K. and Sauriol J. (1981) Effects of snow jams on fluvial activities in the high arctic. *Physical Geography*, **2**, 83–98.

Woo M.K. and Young K.L. (2003) Hydrogeomorphology of patchy wetlands in the high Arctic, polar desert environment. *Wetlands*, **23**(2), 291–309.

Yoshikawa K. (1998) The ground water hydraulics of open system pingos. In *Proceedings of the Seventh International Conference on Permafrost*, Lewkowicz A.G. and Allard M. (Eds.), Yellowknife, pp. 1177–1184, June 23–27, 1998.

Yoshikawa K., Bolton W.R., Romanovsky V.E., Fukuda M. and Hinzman L.D. (2002) Impacts of wildfire on the permafrost in the boreal forests of interior Alaska. *Journal of Geophysical Research*, **107**, 8148, doi:10.1029/2001JD000438, 2002. [printed 108(D1), 2003].

Yoshikawa K. and Hinzman L.D. (2003) Shrinking thermokarst ponds and ground water dynamics in discontinuous permafrost. *Permafrost and Periglacial Processes*, **14**(2), 151–160.

Zhang T., Barry R.G., Knowles K., Heginbottom J.A. and Brown J. (1999) Statistics and characteristics of permafrost and ground-ice distribution in the Northern Hemisphere. *Polar Geography*, **23**(2), 132–154.

Zhang T., Heginbottom J.A., Barry R.G. and Brown J. (2001) Further statistics on the distribution of permafrost and ground ice in the Northern Hemisphere. *Polar Geography*, **24**(2), 126–131.

Encyclopedia of
Hydrological Sciences

Encyclopedia of
Hydrological Sciences

Editor-in-Chief

Malcolm G Anderson
Department of Geography, University of Bristol, Bristol, UK

Senior Advisory Editor

Jeffrey J McDonnell
Department of Forest Engineering, Oregon State University, Corvallis, OR, US

5

WILEY

PART 15

Global Hydrology

173: Global Water Cycle (Fundamental, Theory, Mechanisms)

THOMAS C PAGANO[1] AND SOROOSH SOROOSHIAN[2]

[1]Department of Agriculture, Natural Resources Conservation Service, Portland, OR, US
[2]Department of Civil and Environment Engineering, University of California, Irvine, CA, US

Water cycle is the never-ending movement of water on Earth. Water continuously cycles through various reservoirs in the ocean, sky, and soil. A variety of measurement systems are necessary to quantify the fluxes between reservoirs, ranging from simple buckets to measure rainfall to sophisticated satellites orbiting the Earth. The hydrologic cycle is mainly characterized by its variability in space and time. The Earth contains an incredible diversity in climates, including rainforests, parched deserts and frozen tundra. Water cycle at any given location changes both rapidly and slowly, with the passing of flash floods to the decadal shifts in ocean patterns. Anticipated impacts of human-induced climate change on the hydrologic cycle are as yet unknown, but the consequences are potentially severe. Nonetheless, humans have a long history of altering the hydrologic cycle including the diversion and impoundment of streamflow, pumping of groundwater, irrigation of fields and management of forests. While the water cycle has been recognized for ages, many interesting fundamental research questions remain.

INTRODUCTION

Water is critical to the habitability of our environment and planet as a whole. It is both a subtle agent and a powerful geologic force, shaping landscapes and influencing climate. On a wide range of time scales, water is in eternal movement, cycling through various reservoirs in the ocean, sky, and soil. This unending circulation of the Earth's moisture is called the water cycle or hydrologic cycle.

As with all cycles, the hydrologic cycle is ongoing and continuous, with no specific start or end point; however, by far, the greatest reservoir of water is the ocean, covering about three fourth of the Earth's surface. Water from the oceans evaporate into the atmosphere. The atmosphere then releases this water vapor primarily as precipitation in the form of rain, snow, sleet, or hail. During precipitation, some of the moisture evaporates back to the atmosphere before reaching the ground, some water is intercepted by vegetation, a portion infiltrates the ground, and the remainder flows off the land into lakes, rivers, or back to the ocean. The moisture on and beneath the Earth's surface is of particular importance to humans and society. Water cycle is also intricately intertwined with many other environmental cycles, such as the transportation of energy, chemicals, and sediments.

About half a million km^3 of water evaporate from the ocean's surface each year. Approximately the same amount falls back as precipitation across the globe, only one fifth of which falls on land. The water falling on land would equal a depth of more than 300 m over an area of the size of Germany.

Hydrologic cycle is mainly characterized by its variability in space and time. For example, water continuously evaporates from the surfaces of water bodies (such as oceans, lakes, and rivers). Similarly, precipitation that is intercepted by plants and other surfaces often evaporates in a matter of hours. Once evaporated, it takes an average of about ten days for a parcel of water to cycle through the atmosphere again. If the precipitation infiltrates into the water table, or falls on a perennial glacier, it may reside there for hundreds of years before moving on to the next step in the hydrologic cycle. In addition to variable residence times, the processes associated with the hydrologic cycle are not evenly distributed over the globe; and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

vary by climatic regions. For example, evapotranspiration occurs readily in semiarid regions, but subsequent precipitation may not occur within the same basin or region. The dramatic differences in the revolutions of water are especially evident when one evaluates the hydrologic cycle at the catchment scale.

Water in the hydrologic cycle can be both a benefit as well as a hazard, with its extreme variations being particularly dangerous. Societies can thrive in otherwise hostile climates by drawing supplemental water from the ground or diverting it from rivers, but rapid and intense precipitation or snowmelt can also cause devastating floods and contribute to soil erosion. However, more detrimental are the extended periods without precipitation causing severe droughts. Such droughts cause hardships today, but have contributed to the collapse of civilizations in the past.

Additional, varying and detailed discussions of the water cycle may be found in Freeze and Cherry (1979), Driscoll (1986), Chahine (1992), Maidment (1993), Horden (1998), and Bonan (2002). Shiklomanov (1997, 1999) is an excellent source of quantitative data about the availability of global freshwater resources.

RESERVOIRS AND FLUXES OF THE WATER CYCLE

The fluxes and reservoirs of the hydrologic cycle are depicted graphically, schematically, and conceptually in Figures 1–3. The relative size and importance of the different components ultimately depend on the scale of interest. However, on a global level, the major reservoirs for water are the ocean, atmosphere, cryosphere (snow and ice), lithosphere (surface and groundwater), and biosphere (Table 1). Water is transferred between reservoirs primarily via four fluxes: precipitation, evapotranspiration, sublimation, and runoff (Table 2). Additionally, there are fluxes that transfer water within a reservoir, such as advection of moisture in the air, percolation in soils, and the so-called Thermohaline Circulation, which conveys water to and from the ocean's surface and its depth.

Reservoirs

As mentioned earlier, oceans are the largest reservoirs of moisture on Earth, containing approximately 96.5% of the total available water. The oceans store and circulate an enormous amount of water and energy. Varying patterns of ocean surface temperatures can exert strong influence on circulation patterns in the atmosphere. The most well known of such a pattern is El Niño, characterized by unseasonably warm or cool surface temperatures in the equatorial Pacific Ocean. On average, water resides in the ocean for about 4000 years before evaporating. This residence time varies with ocean depth, with waters of ancient time residing at the bottom of the ocean.

Although the volume of water contained in the atmosphere is small (only 0.001% of the Earth's total reserves), it is a very quickly cycling reservoir. On average, water resides in the atmosphere eight to ten days before falling back to Earth. Water is stored in the atmosphere in liquid and solid forms in clouds or as water vapor. While the oceans circulate both water and energy, water regulates the radiation budget of the atmosphere. Water absorbs and reflects incoming short-wave solar radiation and absorbs long-wave radiation emitted from the Earth's surface. Atmospheric water is the most significant contributor to the natural greenhouse effect; without it, the Earth would be on average about 30 °C colder than it is today.

The *cryosphere* includes all portions of the climatic system, consisting of the world's ice masses and snow

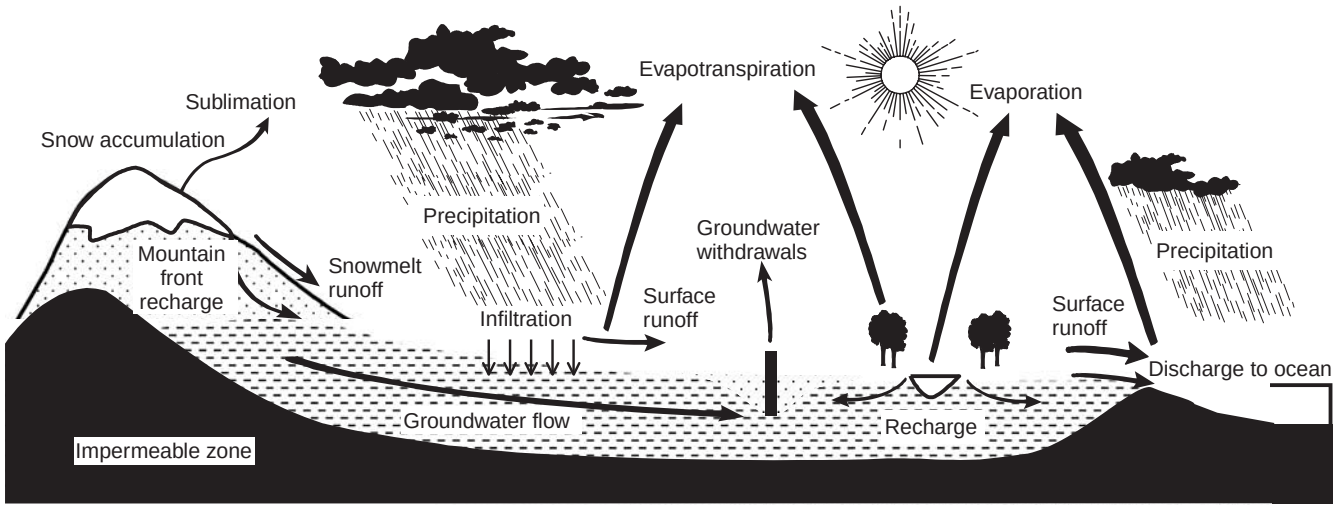


Figure 1 A schematic diagram of various fluxes within the hydrologic cycle (courtesy B. Imam)

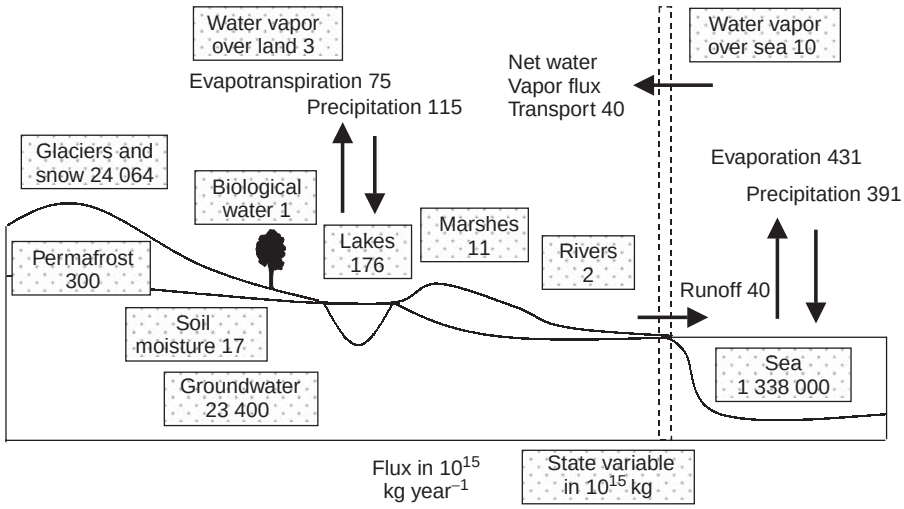


Figure 2 A diagram of the various fluxes and reservoirs within the hydrologic cycle with their yearly average magnitudes (Oki, 1999). The magnitudes given are approximate, and differ from other authors. For example, *see* Chahine (1992); Figure 1 for comparison

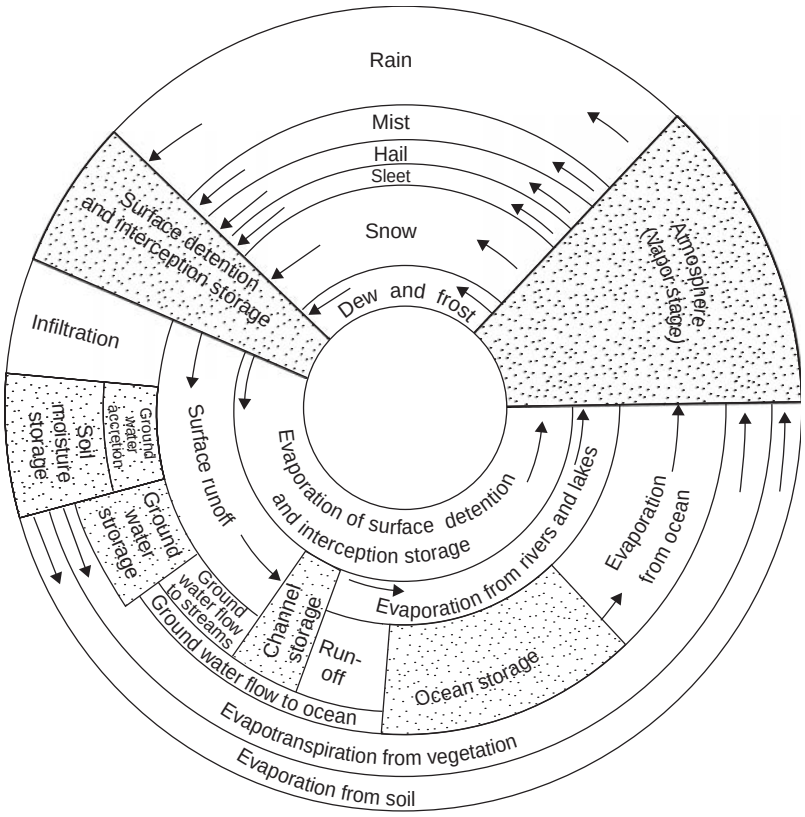


Figure 3 A conceptual diagram of the hydrologic cycle (after Wisler and Brater, 1959)

deposits. This comprises the ice sheets, ice shelves, ice caps and glaciers, sea ice, seasonal snow cover, lake and river ice, seasonally frozen ground, and permafrost. In many mid-latitude regions, winter precipitation falls as snow and remains on the ground for months. During spring and summer, melting snow sustains river flow during much of the period of greatest human demand. Nature's mountain snow is a water storage facility greater than any storage ever

Table 1 Distribution of water on Earth (from Korzun, 1978). The individual numbers and column totals may not exactly agree due to rounding

Form of water	Area covered ($1000\,km^2$)	Volume ($1000\,km^3$)	Share of world reserves (%)	
			Total water reserves	Freshwater reserves
Oceans	361 300	1 338 000	96.5	–
Groundwater	134 800	23 400[a]	1.7	–
Fresh groundwater	134 800	10 530	0.76	30.1
Soil moisture	82 000	16.5	0.001	0.05
Glaciers and permanent snow cover	16 000	24 000	1.74	68.7
Antarctica	14 000	22 000	1.56	61.7
Greenland	1800	2300	0.17	6.68
Arctic islands	230	83.5	0.006	0.24
Mountainous areas	220	40.6	0.003	0.12
Ground ice in zones of permafrost strata	21 000	300	0.022	0.86
Water reserves in				
Lakes	2000	180	0.013	–
Freshwater	1240	91	0.007	0.26
Salt water	820	85.4	0.006	–
Marsh water	2700	11.47	0.0008	0.03
Water in rivers	148 800	2.12	0.0002	0.006
Biological water	510 000	1.12	0.0001	0.003
Atmospheric water	510 000	12.9	0.001	0.04
Total water reserves	510 000	1 390 000	100	–
Freshwater	148 800	35 000	2.35	100

[a]Not including groundwater reserves in Antarctica.

Table 2 Estimates of average annual precipitation (P), evaporation (E), runoff rate (P-E), and runoff ratio ((P-E)/P). Over land, the runoff ratio represents the fraction of precipitation that contributes to runoff. (Table modified from Piexoto and Oort (1992))

Region	Surface area ($10^6\,km^2$)	P ($mm\,year^{-1}$)	E ($mm\,year^{-1}$)	P-E ($mm\,year^{-1}$)	(P-E)/P
Europe	10.0	657	375	282	0.43
Asia	44.1	696	420	276	0.40
Africa	29.8	696	582	114	0.16
Australia	8.9	803	534	269	0.33
North America	24.1	645	403	242	0.38
South America	17.9	1564	946	618	0.40
Antarctica	14.1	169	28	141	0.83
All land areas	148.9	746	480	266	0.36
Arctic ocean	8.5	97	53	44	0.45
Atlantic ocean	98.0	761	1133	−372	−0.49
Indian ocean	77.7	1043	1294	−251	−0.24
Pacific ocean	176.9	1292	1202	90	0.07
All oceans	361.1	1066	1176	−110	−0.10
Globe	510.0	973	973	0	0

made or conceived by man. While representing only 1.7% of the Earth's total water supply, the cryosphere comprises almost 67% of the Earth's freshwater reserves.

The cryosphere further has a powerful impact on the Earth's energy cycle, due to the high albedo (reflective ability) of ice and snow. Dark, snow-free surfaces absorb more solar radiation, which in turn increases the surface air temperature, leading again to more warming and less snow due to the positive feedback. At the opposite extreme, it has been hypothesized that, if the polar icecaps grow beyond approximately 30° latitude, a "runaway ice-albedo" effect would result in the Earth being entirely covered in ice.

The low thermal inertia of high latitude and high altitude regions also makes these reservoirs particularly vulnerable to changes in climate.

Rain that reaches the soil surface is wholly or partially absorbed by the soil in the process of infiltration. Water can be conveyed across or stored on the land surface by lakes, wetlands, rivers, surface soil moisture, and in biota. Like the atmosphere, the content of water in this reservoir is relatively small, but its fluxes are very rapid. The relevance of surface water to human activities is paramount. Society has a long history of altering the surface component of the hydrologic cycle by building aqueducts, digging irrigation canals, and diverting and damming rivers. The character of vegetation on the land's surface influences both its hydrologic and thermal properties. Barren soils or grasslands have dramatically different evapotranspiration and radiation regimes than tropical or boreal forests.

Water beneath the land surface, or groundwater, can be classified in a variety of ways. Typically, above a zone of saturation is a zone of aeration, which itself can be divided into a region of near-surface soil moisture, an intermediate zone, and a capillary fringe. The boundary between the saturated and unsaturated parts of the soil is called the water table. While the near-surface soil moisture is important for vegetation and agriculture, the saturated zone or "aquifer" supplies water for wells and baseflow for streams and springs. The movement of water within the groundwater system is typically very slow, with some reservoirs effectively considered ancient and nonrenewable. On a global scale, groundwater contains a small percentage of the world's total water, yet it accounts for 30% of the Earth's freshwater reserves.

The issue of scale is important when considering the various reservoirs of the hydrologic cycle. If all the water on Earth were to fit proportionately within a 55-gallon drum, groundwater and water in the cryosphere would fill less than one gallon each. Water reserves in lakes, including the Great Lakes, would fill less than two tablespoons. The water in the atmosphere would amount to three drops, and all of the Earth's rivers; from the Nile to the Amazon, the Yangtze, and the Mississippi, would fit into the eye of a needle.

Fluxes

Precipitation is the process by which water, in the form of rain, snow, sleet, and hail, falls from the atmosphere to the Earth's surface. The occurrence of precipitation over land is typically cited as the driving force of the hydrologic cycle, because it triggers the commencement of other fluxes (evapotranspiration, runoff, infiltration) by providing a new source of moisture to the system. There are three general storm types that bring about precipitation: cyclonic, convectional, and orographic. *Cyclonic storms* are generally widespread in nature and form at the boundaries between cold polar air masses and warm tropical air. These storms are in contrast to the small *convective storms* that result from uneven heating of the Earth's surface. Unstable air rises, forming towering cloud masses that deliver brief but intense rainfall. Finally, when warm moist air masses interact with topography such as mountains, the air is lifted and cooled, causing the air parcels to saturate and precipitate moisture via *orographic lifting*. Of course, individual storms such as thunderstorms that cluster around mountains, can display more than one characteristic of the three general storm types. Evaporation is the return of water from bare soil or open bodies of water (mainly the ocean surface) to the atmosphere. Transpiration is the transfer of water to the atmosphere through the stomata of vegetation. Except on the ocean surface, it is difficult to distinguish evaporation and transpiration, and the two terms are often combined into *evapotranspiration* when vegetation is present. Evaporation rates are determined by the availability of energy, primarily solar radiation, and the capacity of air to carry away moisture. Strong warm, dry winds are very efficient at wicking water into the atmosphere. Soil conditions and characteristics also influence evaporation rates. Potential evaporation is the rate at which water could be evaporated, were it available, such as from a completely saturated surface. Actual evaporation is limited by the rate at which unsaturated surfaces can supply available moisture to the atmosphere. Actual evaporation is the minimum of atmospheric demand and land-surface supply. If the soil surface is free of plants and covered with a thick insulating layer of litter, evaporation will be much less than if the soil is bare. Agriculturists combat soil-moisture loss by laying protective mulches on the soil.

During photosynthesis, plants absorb CO_2 from the atmosphere through stomata or tiny holes. When these stomata are open, moisture escapes to the atmosphere. This moisture is often drawn up through the roots of plants and to the leaves, desiccating soils in the process. Many of the same factors that influence evaporation such as radiation and dry winds, also influence transpiration. However, the type, condition, and quantity of vegetation are also important. Generally, agricultural crops have greater transpiration rates than natural vegetation under equivalent conditions. When plants are dormant, transpiration rates are low. Plants transpire until the soil-moisture level falls to the wilting point, at which the plant cannot extract moisture from the soils. A variety of sophisticated methods for calculating evapotranspiration exist.

Similar to evaporation, *sublimation* is the loss of snow or ice on the land surface to vapor in the atmosphere. Warm, dry winds favor sublimation. The transformation of solid water directly to water vapor requires over eight times the energy required for evaporation. While sublimation in nature is not well understood or measured, it can be of serious importance to water managers. During a 5-day period in April 2003 in northcentral Utah, a warm air

mass with sustained $50-100\,\mathrm{km\,h^{-1}}$ winds sublimated an estimated 7.5 cm of snow water equivalent. This disappearance of 30% of the existing mountain snowpack rapidly darkened the water supply outlook for this region. This exceptional event had sublimation rates much higher than typical values.

While snow can sublimate from the land surface, it can also be lost while resting on vegetation. Interception of precipitation by vegetation and throughfall to the surface is a small but important component of the hydrologic cycle. Leaves can typically store 1.3 mm of liquid water, but forests can store 3.8 mm of water in snow. Water leaves the plant surface by evaporation, by throughfall when incoming precipitation exceeds leaf storage ability, or by mechanical removal when trees sway in the wind.

As water is intercepted by vegetation, it also falls on other permeable and impermeable surfaces. This water can gather into small rivulets that carry this overland flow into small channels, then onto larger channels, and finally as streamflow in rivers.

Of the water that permeates into the land surface, subsurface flow is the runoff that travels through the upper soil layers laterally toward streams. A part of the subsurface flow may enter the stream quickly, while the remaining part may take a long time before becoming streamflow. *Hortonian overland flow* occurs when rainfall intensity exceeds the rate at which the soils can absorb the water. In contrast, *saturation excess flow* results when new precipitation is added to already saturated soils. The final source of runoff is water that travels to the deep soils, joins the groundwater and contributes to baseflow. The entry of water into soils is called infiltration, and the movement of water underneath the land surface is called percolation or subflow.

Many factors contribute to the character and variability of runoff. The most influential factors are, of course, precipitation and evaporation, specifically their amount, seasonality, intensity, and sequencing. Land-surface characteristics, as well as basin geology and geography, also play an important role. Soil compaction, trampling, intense forest fires, and urbanization can reduce infiltration rates, which increase overland flow and decrease recharge. Soil porosity and soil type influence the landscape's water transmission and water-holding capacity.

Vegetation and forests have a complex relationship with runoff. The effect of ground litter and vegetation on soil structure generally helps infiltration. Vegetation transpires, and canopies can both inhibit and foster sublimation. In the past, land managers have viewed deforestation and removal of riparian vegetation as a way to increase water yields. However, these increases generally do not offset the costs of mechanical removal or the secondary consequences of erosion or increased runoff "flashiness".

DATA AND THE HYDROLOGIC CYCLE

Timely and accurate hydrologic data are critical for human activities ranging from irrigation water management and reservoir control to developing and improving fundamental understanding of the Earth's system and its reactions to past and future forces of climatic change. There are many ways to measure different aspects of the hydrologic cycle and, because these approaches are always changing, only a limited subset of the major techniques will be addressed in this section.

Precipitation is measured primarily by gauges, radar, and satellites. In the United States, precipitation has been measured on the ground at over 18 000 locations. Typically, rainfall is funneled into a small-diameter measuring tube, where its depth is amplified so that hundredths of an inch can be measured accurately by the human eye. Many automatic systems also exist, such as those that funnel the rainfall into a bucket that is weighed, or a small "tipping bucket" of a known volume that empties when filled (precipitation rate is measured by the frequency of tipping). In cold regions where precipitation often falls as snow, the gauge can be charged with antifreeze and capped with an oil film to minimize evaporation.

Although ground-based gauge measurements are accurate, they lack the complete spatial coverage and continuous measuring capability of the radar. An antenna transmits radiation of a specific frequency at defined intervals and listens for an echo. Heavier rain reflects more energy back to the radar than lighter rain. However, rain that is more distant also gives a weaker, delayed signal. After accounting for the time between sending the transmission and receiving the echo, and the intensity of the echo, one can measure precipitation rate relatively accurately. This system is particularly useful because it also provides information about storm direction, speed, and whether the precipitation is falling as liquid or solid.

Because radar cannot travel through the Earth's surface, there can be gaps in coverage where the radar beam is blocked by mountains. Satellites provide a "bird's-eye" view of the surface. Tall clouds are often associated with vigorous precipitation; satellites measure cloud-top temperatures to infer precipitation rates. Hybrid systems exist, taking advantage of the strengths of multiple sources of information. Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN), (Sorooshian *et al.*, 2000) specifically uses cloud top-temperature and texture from geosynchronous satellite images to estimate surface rainfall rates. The parameters of the model are updated using gauge and radar data. Several other similar satellite precipitation estimation systems exist, as described by Kuligowski (2002), Ba and Gruber (2001) and Scofield (1987).

Snow depth can be sampled using a measuring stick or a permanently mounted stake. Newer automatic measurements echo ultrasonic waves off the snow surface. However, as snow density can vary widely, snow water equivalent measurements are more accurate indications of total water supply. At one time, snow water equivalent measurements involved extracting a core from the snowpack, originally using a stovepipe, and melting it in the field to its liquid form. After the turn of the century, metal snow samplers were developed so that one only needed to obtain a core and weigh the entire apparatus. In the 1970s, over 1500 locations (snow courses) were being routinely measured near the first of each month throughout the western United States to support seasonal water supply forecast activities. Shortly after, many of these sites were converted to Snow Telemetry (SNOTEL) sites that measure snowpacks by the pressure they exert on a large liquid-filled bladder resting on the ground. State-of-the-art satellites also measure snow-covered area with reasonable accuracy.

Atmospheric humidity is measured by a variety of methods, including the sling psychrometer. Two thermometers, one of which is wrapped in damp gauze, are attached to a sling and swung in a circle. This motion causes the moisture in the gauze to evaporate, releasing latent energy, and thereby lowering the temperature of the thermometer. The final difference in the temperatures is proportional to the absolute humidity. Chilling a mirror until dew condenses on its surface and using the data so obtained works on a similar principle. Many measurements are taken by electronic means, such as how humidity changes the capacitance of a thin gold-plated plastic film transducer. Sensors are routinely attached to balloons, and the humidity is measured at several layers of the atmosphere. The atmospheric advection component of a local water budget is sometimes calculated using a battery of such radiosondes launched in unison around the perimeter of a region. Satellite measurements of humidity exist, yet the capacity of satellite-based sensors to probe beneath cloud levels remains an issue.

Aside from groundwater recharge, the most uncertain component of the water budget is evaporation. Stainless steel pans are filled with water, and the lowering of water level, minus precipitation, is proportional to potential evaporation. This method does not give an estimate of the actual evaporation, because it does not account for unsaturated soils and vegetation. Often, all other components of the water budget are measured, and evaporation is "measured" by calculating the residual of the other terms.

Streamflow in rivers is measured in two steps. First, a current meter is dropped into the stream at fixed spatial intervals. The earlier current meters involved rotating cups that spun as water passed through the meter, measuring water velocity in much the same way that wind speed is measured. Newer technologies involve sonic sensors. Measurements are taken at fixed depths and intervals from one riverbed to the other. The velocity of the river multiplied by the area equals the total flow. Because the depth of a river is proportional to its flow, routine, automatic measurements typically measure only the river depth, which is later converted to flow by an equation specific to the river and location being measured. Depth is measured by observing the water level in a stilling well. If possible, a controlled structure such as a weir or a flume can be placed in the streambed and the flow calculated directly by measuring the water depth in the structure.

MODELING OF THE WATER CYCLE

Mathematical Water Balance

The most simplistic formulation of a water balance is denoted by the elementary continuity equation (1) that conveys the notion that "input to a hydrologic system equals the output from the system, plus or minus any changes in storage":

$$I = O \pm \Delta S \tag{1}$$

where, for a given domain, I is the total inflow, O is the outflow, and ΔS is the change in storage. Figure 4 shows a conceptual model of a water balance at a watershed scale. Note that the figure does not represent changes in storage caused by anthropogenic activities such as groundwater pumping, artificial recharge, and other modifications.

In a parcel of the atmosphere, for example, the inputs would be evapotranspiration from below and advection of moisture from nearby parts of the atmosphere. Outputs would be precipitation as well as advection out of the parcel. The change in storage is the change in the specific humidity.

The water balance beneath the soil surface is the mirror image of the atmosphere. Precipitation (infiltration) is now an input, evapotranspiration is an output, along with runoff. Lateral groundwater flow and subsurface flow is the equivalent of atmospheric advection. The change in storage is measured by the change in near-surface soil moisture and water content of the aquifer.

On a global basis, the Earth is effectively a closed system, and the amount of water present remains relatively constant (i.e., $\Delta S \cong 0$). However, input and output rates of the hydrologic cycle vary regionally and on a wide range of time scales (see Figure 5). Describing, quantifying, and predicting these variations are, in essence, major tasks in contemporary hydrology.

Climate and Hydrology Models

To describe and predict variations within the hydrologic cycle, considerable effort has been invested in developing computer-based numerical models of hydrology and climate. Every component of the climate system has its own models (from groundwater to oceanography and the

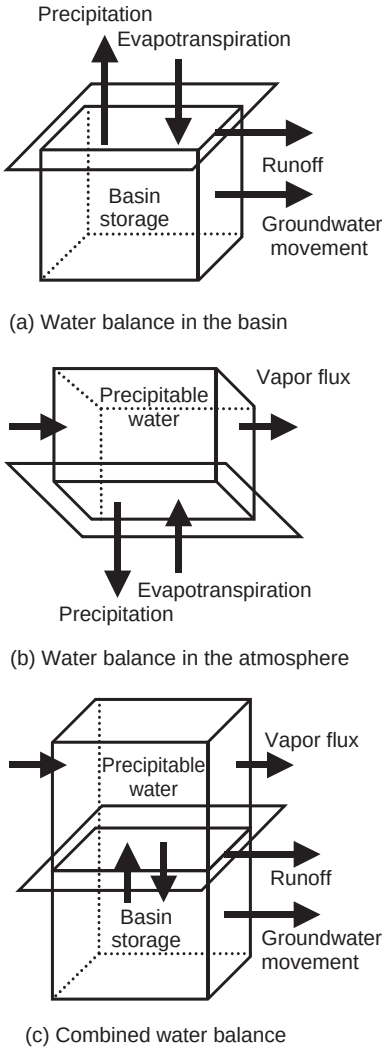


Figure 4 Mathematical schematic diagram of water balance for (a) the atmosphere (b) the land surface, and (c) the atmosphere and land surface combined (Reproduced from Oki, 1999 with permission of Cambridge University Press)

atmosphere to the land surface) and, within disciplines, there are too many different models to be completely described here.

Simple surface hydrologic models are often statistically based or regression models. For example, operational hydrologists within the United States commonly use the principal components-based multiple linear regression procedure developed by Garen (1992). At lead times of several months, precipitation, snowpack, and antecedent streamflow data are used to predict the total amount of flow that will occur over a season. Other simple models include the Natural Resources Conservation Service Curve Number, Rational Method, or the Antecedent Precipitation Index.

In a coarse sense, simple conceptual hydrologic models describe the land surface like a leaky bucket of water. A very wet watershed releases water to the streams at a much greater rate than a parched watershed would, much like how a full bucket leaks water faster than a bucket that is half full. Similarly, when a watershed is completely saturated, all additional moisture will go directly to the streams, like a bucket overtopping when it is completely full.

More complex hydrologic models, such as the Sacramento Soil Moisture Accounting Model and the Precipitation-Runoff Modeling System, often comprises a series of interlinked leaky buckets. Various buckets represent various parts of the soil profile, from the near-surface soil moisture to the deep water contributing to baseflow. Each model has a number of parameters to describe the size of each bucket and the rate at which water travels between reservoirs. These parameters are often related to physical characteristics of the watershed, such as soil types, vegetation cover, geology, basin size, and so on. Some models may treat an entire basin as a single bucket; increasingly complex models have high spatial resolution and describe the movement of water between grid cells across the landscape.

The spatial and temporal scales of their applications vary from model to model, but range from tens to thousands of km^2 and from minutes and hours to days and years (Sorooshian *et al.*, 1997 and Singh, 1995). The structure of a model also reflects its intended purpose. Some models have the ability to describe the coupling of hydrology with other natural cycles (e.g., erosion or water quality).

The most complex models available describe nearly all of the major natural cycles in concert. These General Circulation Models (GCMs) have full global representations of the ocean, atmosphere, cryosphere, and land surface. Most of the early work on GCMs related to refining the treatment of the ocean-atmosphere interface. Recently, more emphasis has been put upon accurately describing the land surface-atmosphere interface. Such land surface models include the Biosphere Atmosphere Transfer Scheme and the Simple Biosphere Model. Lau *et al.* (1995) compared the ability of 29 GCMs in simulating various aspects of regional hydrologic processes and found them insufficient for use in climate studies related to continental scale water balance. One of the primary impediments is the fine scale of hydrologic processes and the coarse resolution of climate models. Regardless, this is an area of very active research and, as computing power increases rapidly in the near future, one can expect these models to improve.

HUMANS AND THE HYDROLOGIC CYCLE

Climate Change

The primary motivation for improving the representation of land-surface processes in GCMs is to gain a better understanding of the anticipated impacts of climate change

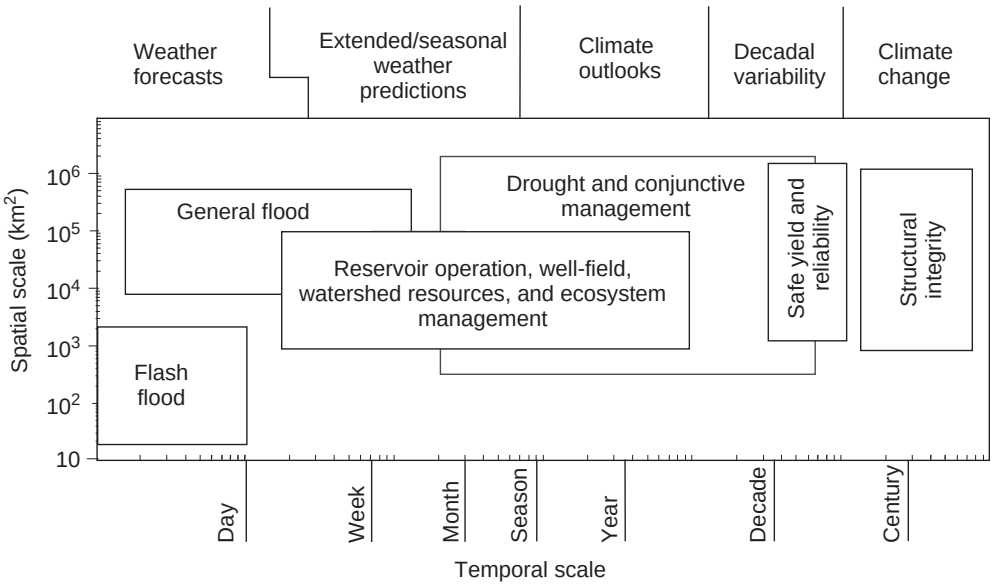


Figure 5 Space time issues in hydrology

on the hydrologic cycle. Humans are currently engaged in what has been described as a grand irreversible and uncontrolled experiment, the environmental consequences of which could be second only to global nuclear war. Since World War II, the emission of CO_2 into the atmosphere has increased almost sevenfold due to the burning of fossil fuels. This rate could increase dramatically during the impending modernization and industrialization of developing nations. Carbon dioxide in the atmosphere contributes to the "greenhouse effect" which could raise the Earth's temperature at an unprecedented rate. The resulting changes on the remainder of the climate system are yet unknown (see Table 3). Table 4 shows some of the trends in hydrology observed in the recent past.

While there is some uncertainty and debate concerning climate change and global warming, the theory as to why the hydrologic cycle may change is relatively straightforward. Greenhouse gases in the atmosphere reflect and reradiate outbound long-wave radiation back to the Earth's surface, increasing the surface temperature. Carbon dioxide is one such gas, while methane and nitrous oxide are among some of the others. As mentioned previously, atmospheric water vapor is the most significant contributor to the natural greenhouse effect.

Coincident variations in greenhouse gases and air temperatures have been thoroughly documented on geologic time scales. For example, during the last glacial period when atmospheric CO_2 concentrations were closer to 200 parts per million by volume (ppmv), the Earth was 6–8 °C cooler than during preindustrial conditions of 280 ppmv. The current concentration as of 2005 is approximately 380 ppmv, and the predictions of future concentrations range from 540

Table 3 Best estimates of climate change projections over the next 50–100 years (from Schneider *et al.* (1992) and S. Schneider (personal communication, July 27, 2000))

		Distribution of changes				Confidence of projection	
Indicators	Annual average change	Regional average	Change in seasonality	Interannual variability	Significant transients	Global average	Regional average
Temperature	+1 to +3.5 °C	−3 to +10 °C	Yes	Down?	Yes	High	Medium
Sea level	+15 to +95 cm	–	No	?	Unlikely	High	Medium
Precipitation	+7 to +15%	−20 to +20%	Yes	Up	Yes	High	Low
Direct solar radiation	−10 to +10%	−30 to +30%	Yes	?	Possible	Low	Low
Evapo-transpiration	+5 to +10%	−10 to +10%	Yes	?	Possible	High	Low
Soil moisture	?	−50 to +50%	Yes	?	Yes	?	Medium
Runoff	Increase	−50 to +50%	Yes	?	Yes	Medium	Low
Severe storms	?	?	?	?	Yes	?	?

Table 4 Observed trends in the hydrologic cycle (from IPCC, 1995)

Variable	Observed trend	Confidence
Ocean		
High clouds	Increase 1951–1981	Low
Mid-level clouds	Increase in Northern Hemisphere, mid-latitude 1951-81	Low
Convective clouds	Increase 1951–1981	Low
Fair weather cumulus clouds	Decrease 1951–1981	Low
Water vapor	Increase 1973–1988	Low
Evaporation in tropics	Increase 1949–1989	Medium
Land		
Mid- to high-latitude clouds	Increasing 1900–1980s	Medium
Mid- to high-latitude precipitation	Increasing since 1900	Medium
Northern hemisphere subtropical precipitation	10% decrease since 1970	Medium
Evaporation in USA and FSU[a]	Decreasing since 1950	Low
Soil moisture in FSU[a]	Increasing 1970s-1990s	Low
Runoff	Pattern consistent with precipitation changes	Medium

[a]Former Soviet Union.

to 970 ppmv by the year 2100. The anticipated increase in global surface temperatures over this period is expected to be between 1.4 and 5.8 °C.

If the amount of long-wave radiation reflected back to the Earth's surface is increased, the potential evaporation from the ocean and the land surface may also increase. The amount of water required to saturate a parcel of air increases exponentially in direct proportion to its temperature; because warmer air can potentially hold more water, humidity may increase. If the increase in water vapor storage does not balance the increased evaporation, then precipitation rates, on average, could increase.

Beyond these first few concepts, the vision of the future of the hydrologic cycle becomes murky. The Earth's climate contains many complex feedbacks, some of which are self-regulating and others reinforcing. For example, a more vigorous atmospheric water cycle may involve increased low clouds that might reduce the amount of incoming solar radiation, leading to a negative feedback and cooling. An increase in high clouds, which are efficient at reflecting long-wave radiation back to the Earth's surface while letting short-wave radiation pass through would reinforce positive feedback with further warming.

The most difficult issue concerning global warming is translating changes in climate to impacts on societies. Humans are most sensitive to extremes in hydrologic variations, and the sequences of events it causes. Such questions as, "How will climate change affect the occurrence of drought in a particular region?" remain open to debate. However, considerable effort has been devoted to researching these questions through analyses of historical records and computer simulations of the future (National Assessment Synthesis Team, 2001).

It is possible that the clearest changes associated with global warming could be in the occurrences of catastrophic events such as heavy floods, extended droughts, and hurricanes. If a climatological variable has a Normal (bell-shaped) distribution, changes in the mean will bring proportionally greater shifts in intensity and frequency at the tails of the distribution (Trenberth, 1999; *see also* Easterling *et al.*, 2000 for an excellent discussion of the observed and anticipated changes in extreme events induced by climate change). Unfortunately, because of its rarity, it is very difficult to estimate the change in frequency of extreme events. The climate may change irreversibly by the time that undeniable evidence of change accumulates.

The atmosphere is not the only component of the hydrologic cycle that may be affected by climate change; the ocean and cryosphere may be impacted as well. Computer simulations of climate change predict that the greatest warming will occur at high latitudes and altitudes. Locally, this poses a threat to regions with permafrost, where melting can cause land-surface subsidence and damage to structures. For the first time in its 30-year

history, the start of the 2003 Iditarod sled dog race had to be moved north because of exceptionally warm weather and a lack of snowpack in Alaska. From October to April 2003, warm temperatures, heavy rains, and the lack of a winter freeze caused a near-complete reversal of the seasonal cycle for the Kenai River at Cooper Landing in Alaska. Although record-high streamflows occurred during the normally frozen winter, sparse snowpack remained to carry the flow through the summer when the surge of melt usually occurs. There is concern that a similar change in the seasons may occur in the mountainous western United States, particularly the Cascades and Sierra Nevada mountain ranges.

If temperatures increase sufficiently, glacial melt and thermal expansion of the oceans may also cause sea-level changes over the next 100 years. A 1% decrease in glacial water content translates to a 30-cm rise in sea level. A 2-meter rise in sea level could inundate the Republic of the Maldives, making an entire nation a casualty of climatic change. Changes in surface albedo associated with the change from surface-ice to bare ground or open ocean is a positive feedback, and associated with increased absorption, higher temperatures, and possibly increased melting. The latest projection for global mean sea-level rise between 1990 and 2100 is between 9 and 88 cm, with significant regional variations (IPCC, 2001). Such a change could have dire consequences for coastal ecosystems.

Not all of the changes in the hydrologic cycle may be gradual or linear. Warm temperatures at high latitudes, increased precipitation, and glacial melt may alter the vertical density profile of the North Atlantic Ocean. Warm and relatively freshwater in the North Atlantic could slow or halt the sinking of colder, saltier water to the bottom of the ocean. A disruption in this cycle would greatly impact the climate of Europe, as was seen during the "Younger Dryas" period 13 000 years ago, when temperatures plunged several degrees Celsius in the course of a few decades. Model simulations to study the effects of doubling CO_2 concentrations in the atmosphere predict a disruption in the Northern Atlantic circulation with eventual recovery. Quadrupling concentrations lead to a nearly irreversible collapse of the circulation. These simulations are highly model-dependent, but they do suggest the presence of "thresholds" and "triggers" within the climate system.

Land-use and Hydrology Changes

By far, global climate change associated with fossil fuel burning will not be the first interaction humans have had with the hydrologic cycle. Most of history is linked to redirecting and damming rivers, cultivating land, moving mountains, and pushing the bounds of habitability on Earth (Figure 6). Construction of water supply systems, drainage systems, and irrigation canals extend back to ancient civilizations. In several basins across the globe, surface-water resources have been so extensively developed that major rivers (e.g., the Colorado River in the United States and the Yellow River in China) periodically cease flowing into the ocean.

Water is both a necessity and a resource for financial gain. However, while it is beneficial, it can also turn into a hazard. On average, over $8 billion (£5 billion) worth of damages have resulted annually from flooding and hurricanes in the United States alone (Kunkel *et al.*, 1999). Often, such damages arise partly from unwise human choices. The above figure does not include the damages caused by pandemic health hazards created by poor water quality.

The most pervasive change to the hydrologic cycle due to human activities is caused by land-use change. According to van Dam (1999, pp. xiii), "The effects of climate variability and change on the hydrological cycle will be coincident with those of changes in land use, which could be of the same order of magnitude." The various types of land-use changes range from deforestation, agricultural use, and urbanization to draining of swamplands for various purposes. The impact of these land uses on streamflow is presented in Table 5. The impacts arise from changes in surface albedo, surface roughness, surface permeability (the ability of water to pass through a surface, such as concrete), and the ability of the surface to intercept and evaporate moisture.

These impacts are inherently scale-dependent, and most local land-use change will not have a major impact on the continental and global hydrologic cycle. However, the extent of land-use change in total is considerable; Flohn (1973) suggested that, over the last 8000 years, approximately 11% of the land surface has already been converted to arable land, and 31% of forests have been modified from their original condition. Additionally, certain regions are poised to have a disproportionately strong impact on global circulation. For example, there is debate as to whether Amazon deforestation will have a remote impact on tropical and extratropical climate. Although this is being actively researched, the expected impacts from Amazon deforestation outside the region remain unclear (Gash and Nobre, 1996).

While global fluxes or distributions of water may not be influenced by water quality, there is significant impact of humans on water quality at every step of the hydrologic cycle. Since the 1970s, the primary atmospheric water quality concern has been acid rain. Acid rain damages trees at high elevations and contributes to the acidification of lakes and streams. Regions already affected include North America and northern Europe.

Contamination of water at and below the land surface poses a significant threat to potable water supplies (*see* Fetter, 1998; Bedient *et al.*, 1999 for further reading on groundwater contamination). Water quality can be affected

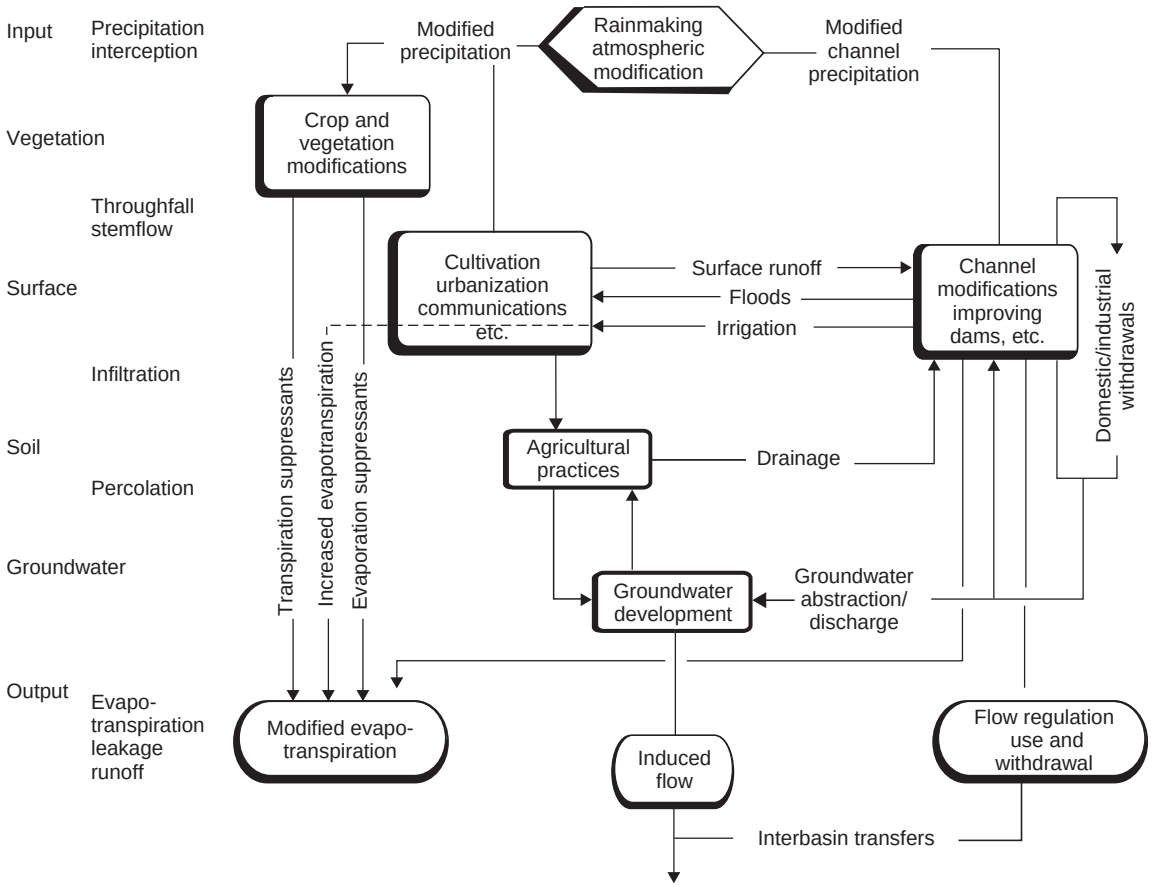


Figure 6 Systems diagram of the impacts of human activities on streamflow (From Ward, 1990 reproduced with kind permission of the Open University/McGraw-Hill Publishing Company)

by human activities in a multitude of ways, including release of effluents, leeching from landfills, industrial and mining activities, and runoffs from fertilizers and pesticides used in agriculture among others. In particular, the long-term isolation of hazardous radiological byproducts from water supplies poses a special challenge.

TOWARD IMPROVED UNDERSTANDING OF THE HYDROLOGIC CYCLE

Although the hydrologic cycle has been studied for well over a century, with considerable work being done in recent decades, many unresolved questions in the expanding frontiers of research into the water cycle remain. In particular, understanding the impacts of human activities on the hydrologic cycle is an area of active research. Considerable effort is also being devoted to understanding the linkages between the ocean and atmosphere in the tropics (such as El Niño), as well as at mid-latitudes and on longer time scales. Increased attention has been devoted to land-surface and hydrologic representation in large-scale computer models, the intent of which is to improve predictions of variations in the hydrologic cycle on interannual and decadal horizons. Medium-range weather predictions can also improve by having a better understanding and model representation of soil-moisture processes (Chahine, 1992). Finally, groundwater and surface interactions, ranging from the amount of water recharged into an aquifer from snowmelt to the hydrology of natural springs, are some of the less understood components of basin-scale water balances.

Improved understanding of the hydrologic cycle is not limited to the understanding contained within the scientific research community. It is also that of the operational water management community, the public, and other stakeholders of water resources. Concepts such as climate stationarity (the belief that the statistics of the climate of a particular region are constant in time) are key assumptions for most design and planning of water management, although the research community has long recognized the flaws in this assumption. For example, for structural design purposes, relatively brief historical records are used to estimate the magnitude of the flood that would happen once in a 100 years. If the historical period being considered contains

Table 5 Summary of impacts of land-use changes. Table modified from Calder (1993) with additional data from Urbonas and Roesner (1993)

Land-use change	Component affected	Principal hydrologic process involved	Geographic scale and likely magnitude of effect
Afforestation (deforestation has converse effect, except where disturbance caused by forest clearance may be of overriding importance)	Annual flow	Increased interception in wet periods. Increased transpiration in dry periods through increased water availability to deep root systems	Basin scale; magnitude proportional to forest cover. World average is 34 mm $year^{-1}$ reduction for 10% increase in forest cover
	Seasonal flow	Increased interception and increased dry period transpiration will increase soil-moisture deficits and reduce dry season flow	Basin scale; can be of sufficient magnitude to stop dry season flows
		Drainage activities associated with planting may increase dry season flows through initial dewatering and also through long-term effects of the drainage system	Basin scale; drainage activities will increase dry season flows
		Cloud water (mist or fog) deposition will augment dry season flows	High-altitude basins only; increased cloud water deposition may have a significant effect on dry season flows
	Climate	Increased evaporation and reduced sensible heat fluxes from forests affect climate	Micro, meso, and global scale; forests generally cool and humidify the atmosphere
Agricultural intensification	Water quantity	Altering of transpiration rates affect runoff	Basin scale; effect is marginal
		Timing of storm runoff altered through land drainage	Basin scale; significant effect
Draining wetlands	Annual flow	Initial dewatering following drainage will increase annual flow	Basin scale; effect may last from 1 to 2 years to decades
		Afforestation following drainage will reduce annual flow	Basin scale; effects as for afforestation
	Seasonal flow	Upland peat bogs, groundwater fends, and African dambos have little effect in maintaining dry season flows	Basin scale: drainage or removal of wetland will not reduce and may increase dry season flows
		Lowering of the water table may induce soil-moisture stress, reduce transpiration, and increase dry season flows	Basin scale; a reduction of water-table depth to a minimum of 30 cm below surface is required
		Initial dewatering following drainage will increase dry season flows	Basin scale; effect may last from 1–2 years to decades
		The deeper flow outlet of the drainage system will lead to increased dry season flows	Basin scale; effects will be long-term
Urbanization	Runoff volume	Impervious surfaces such as paved roads, roofs, and parking lots increase surface runoff during storm events and decrease groundwater recharge	Basin scale; magnitude of effect depends on extent of urbanization

unusually wet or dry spells, this approach will not give a representative estimate of what may occur in the future. Many reservoirs are operated based on expected inflow resulting from snowmelt runoff. If temperatures warm to the point where more winter precipitation falls as rain than snow, then the reservoir operator and water users will come up short in late summer months. Likewise, most regions lack legal recognition of the connection between the groundwater and surface-water components of the hydrologic cycle. The impacts of excessive groundwater withdrawals on streamflow have been fairly well understood and developed within the scientific research community, although few, if any, regions have laws that reflect this understanding.

There are several major research programs designed to develop an increasingly more sophisticated understanding of the hydrologic cycle. One such program, the Global Energy and Water cycle EXperiment (GEWEX), was initiated in 1988 by the World Climate Research Programme (WCRP) 1990; (Chahine, 1992). Part of GEWEX is designed to observe and model the hydrologic cycle, with the ultimate goal of predicting global and regional climate change. The program includes large-scale field activities, intensive measurements as well as modeling and research. GEWEX has contributed to the development of improved numerical models and the creation of state-of-the-art climate data sets, and it is helping to achieve its goals of improving resource management through scientific outreach to the engineering as well as other user communities.

Acknowledgments

We gratefully acknowledge the support provided by SAHRA ("Sustainability of Semiarid Hydrology and Riparian Areas"), an National Science Foundation (NSF) Science and Technology Center at the University of Arizona, as well as the Global Energy and Water cycle EXperiment (GEWEX). Our sincere gratitude is also extended to Terri Hogue, Martha Whitaker, and Corrie Thies for their thoughtful comments and editing of this document in its various stages of completion.

REFERENCES

Ba M. and Gruber A. (2001) GOES Multispectral Rainfall Algorithm (GMSRA). *Journal of Applied Meteorology*, **29**, 1120–1135.

Bedient P.B., Rifai H.S. and Newell C.J. (1999) *Ground Water Contamination: Transport and Remediation*, Prentice-Hall: Englewood Cliffs.

Bonan G. (2002) *Ecological Climatology*, Cambridge University Press: Cambridge.

Calder I.R. (1993) Hydrologic effects of land use change. In *Handbook of Hydrology*, Maidment D.R. (Eds.), McGraw-Hill: New York, pp. 13.1–13.50.

Chahine M.T. (1992) The hydrological cycle and its influence on climate. *Nature*, **359**, 373–380.

Driscoll F.G. (Ed.) (1986) *Groundwater and Wells*, Johnson Division: St Paul.

Easterling D.R., Evans J.L., Groisman P.Y., Karl T.R., Kunkel K.E. and Ambenje P. (2000) Observed variability and trends in extreme climate events: a brief review. *Bulletin of the American Meteorological Society*, **81**, 417–425.

Fetter C.W. (1998) *Contaminant Hydrogeology*, Prentice-Hall: Upper Saddle River.

Flohn H. (1973) Globale energiebilanz und klimaschwankungen. *Bonner Meteorologische Abhandlungen*, Westdeutscher Verlog: pp. 75–117.

Freeze R.A. and Cherry J.A. (1979) *Groundwater*, Prentice-Hall: Englewood Cliffs.

Garen D.C. (1992) Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management*, **118**(6), 654–670.

Gash J.H.C. and Nobre C.A. (Eds.) (1996) *Amazonian Deforestation and Climate*, John Wiley & Sons: New York.

Horden R.H. (1998) The hydrologic cycle. In *Encyclopedia of Hydrology and Water Resources*, Herschy R.W. and Fairbridge R.W. (Eds.), Kluwer Academic: Dordrecht, pp. 400–404 803.

IPCC (1995) Climate change 1995: the science of climate change. *Intergovernmental Panel on Climate Change*, Cambridge University Press: Cambridge.

IPCC (2001) IPCC third assessment report – climate change 2001. *Intergovernmental Panel on Climate Change*, Cambridge University Press: Cambridge.

Korzun V.I. (1978) *World Water Balance and Water Resources of the Earth*, UNESCO.

Kuligowski R.J. (2002) A self-calibrating real-time GOES rainfall algorithm for short-term rainfall estimates. *Journal of Hydrometeorology*, **3**, 112–130.

Kunkel K.E., Pielke R.A.J. and Changnon S.A. (1999) Temporal fluctuations in weather and climate extremes that cause economic and human health impacts: a review. *Bulletin of the American Meteorological Society*, **80**, 1077–1098.

Lau W.K.-M., Sud Y.C. and Kim J.-H. (1995) *Intercomparison of Hydrologic Processes in Global Climate Models*, Technical Memorandum 104617, NASA: Washington.

Maidment D.R. (1993) Hydrology. In *Handbook of Hydrology*, Chap. 1, Maidment D.R. (Ed.) McGraw-Hill: New York.

National Assessment Synthesis Team (2001) *Climate Change Impacts on the United States: The Potential Consequences of Climate Variability and Change: Foundation Report*, Cambridge University Press: New York.

Oki T. (1999) The global water cycle. In *Global Energy and Water Cycles*, Browning K.A. and Gurney R.J. (Eds.), Cambridge University Press: Cambridge, pp. 10–29.

Piexoto J.P. and Oort A.H. (1992) *Physics of Climate*, American Institute of Physics: New York.

Sorooshian S., Gupta H. and Rodda J.C. (1997). Global environmental change and land surface processes in hydrology: the trials and tribulations of modelling and measuring. In *Proceedings of the NATO Advanced Workshop*, Tucson, Arizona, May 17–21, 1993. NATO Advanced Science Institute Series. Series 1: Global Environmental Change,

Vol. 46, Springer-Verlag, Berlin, Heidelberg, New York, XVII, p. 497.

Schneider S.H., Mearns L.O. and Gleick P.H. (1992) Climate-change scenarios for impact assessment. In *Global Warming and Biological Diversity*, Peters R. and Lovejoy T. (Eds.), Yale University Press: New Haven, pp. 38–55.

Scofield R.A. (1987) The NESDIS operational convective precipitation estimation technique. *Monthly Weather Review*, **115**, 1773–1792.

Shiklomanov I.A. (1997) *Comprehensive Assessment of the Freshwater Resources of the World: Assessment of Water Resources and Water Availability in the World*, Report 556.18 SHI, World Meteorological Organization: Geneva.

Shiklomanov I. (1999) *World Water Resources and their Use*, Russian Federation, State Hydrological Institute/UNESCO: St Petersburg, [CD-ROM]; Singh V.P. (Ed.) (1995) *Computer Models of Watershed Hydrology*, Water Resource Publications, Highlands Ranch, CO, US, pp. 1130.

Sorooshian S., Hsu K., Gao X., Gupta H.V., Imam B. and Braithwaite D. (2000) Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bulletin of the American Meteorological Society*, **81**(9), 2035–2046.

Trenberth K.E. (1999) Conceptual framework for changes of extremes of the hydrological cycle with climate change. *Climate Change*, **42**, 327–339.

Urbonas B.R. and Roesner L.A. (1993) Hydrologic design for urban drainage and flood control. In *Handbook of Hydrology*, Maidment D.R. (Ed.) McGraw-Hill: New York, pp. 28.1–28.52.

van Dam J.C. (Ed.) (1999) *Impacts of Climate Change and Climate Variability on Hydrological Regimes*, Cambridge University Press: Cambridge.

Ward R.C. (1990) *Principles of Hydrology*, McGraw-Hill: London.

WCRP (1990) *Scientific Plan for the Global Energy and Water Cycle Experiment*, WCRP-40, World Climate Research Program: Geneva.

Wisler C.O. and Brater E.F. (1959) *Hydrology*, John Wiley & Sons: New York.

174: Global Water Budgets – Fundamental Theory and Mechanisms

C ADAM SCHLOSSER

Joint Program on the Science and Policy of Global Change Massachusetts Institute of Technology, Cambridge, MA, US

The global water cycle represents the transport and transformation of water within the Earth system, and, in doing so, distributes freshwater over the Earth's surface. The cycling of water in the Earth's system employs water in all three of its phases: solid, liquid, and gaseous. In its coarsest and most conventional depiction, the Earth system's cycling of water is viewed as the continual displacement of water taken from the ocean, transported through the atmosphere, deposited over land, and ultimately fed back to the ocean. However, many processes and pathways are responsible for this global transit and cycling. Therefore, to monitor the global water cycle in the context of all of these key interactions with the global environment requires a comprehensive representation of the budgets and cycling of all phases of water storage. An analytic discussion is provided to elucidate the various mechanisms that contribute to the global transport of water within and between the Earth's atmosphere, land, and ocean systems.

INTRODUCTION

The global water cycle (Figure 1) represents the transport and transformation of water within the Earth system, and, in doing so, distributes freshwater over the Earth's surface. The cycling of water in the Earth's system employs water in all three of its phases: solid, liquid, and gaseous. The global water cycle operates on a continuum of time and spatial scales and exchanges large amounts of energy as water undergoes phase changes and is moved from one part of the Earth system to another. In its coarsest and most conventional depiction, the Earth system's cycling of water is viewed as the continual displacement of water taken from the ocean, transported through the atmosphere, deposited over land, and ultimately fed back to the ocean. However, many processes and pathways are responsible for this global transit and cycling (as depicted in Figure 1). Water is, indeed, abundant in our environment, but it is unevenly distributed in its forms that are amenable for sustaining life on our planet. Therefore, to monitor the global water cycle in the context of all of these key interactions with the global environment requires a comprehensive representation of the budgets and cycling of all phases of water storage.

WATER BUDGETS

Global Budget

Two conventional perspectives are typically used in global water budget analyses to analytically represent the global transport and storage of water: from the atmosphere or from the surface of the earth. From an atmospheric perspective, the *global* water budget can be described as

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = E - P \tag{1}$$

where Q is the *total* (i.e. vapor, liquid, and solid) amount of water stored by the *entire* atmosphere and P and E are the corresponding global fluxes of precipitation and evapotranspiration respectively. Note that, implicit in this and all equations that follow, all storage terms are in units of mass and fluxes are mass per unit time, but these equations can also be viewed as volumetric budgets. Strictly speaking, the expression given by equation (1) ignores any loss (e.g. molecular diffusion) or gain (e.g. comet material) of water to/from outer space. These source and sink terms are typically regarded as insignificant when compared with the magnitude of the variation of global Q (on the order

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

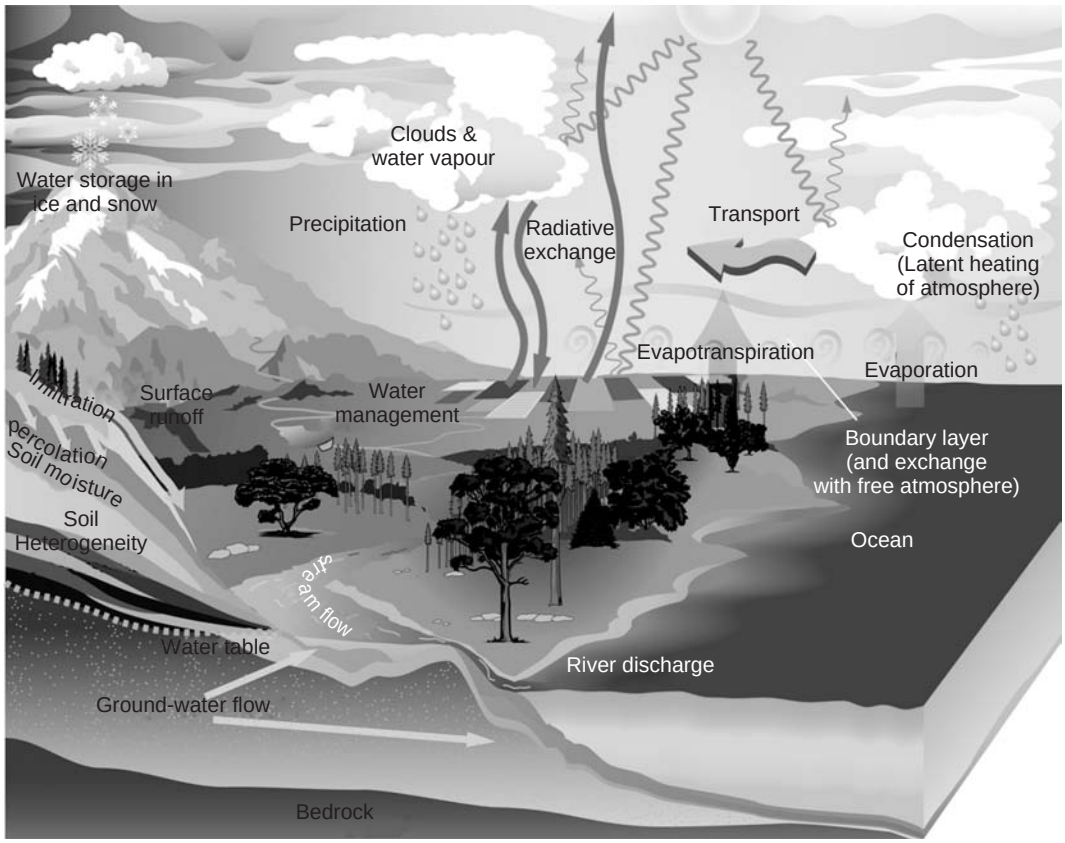


Figure 1 Conceptualization of the global water cycle. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of 10^{14} kg of water, (c.f. Peixoto and Oort, 1992). When viewed from the perspective over the entire Earth's surface (i.e. ocean and land), the water budget can also be expressed as

$$\frac{\mathrm{d}W}{\mathrm{d}t} = P - E \tag{2}$$

where W is the *total* (i.e. vapor, liquid, and solid) change of storage of water over and within the Earth's continents and oceans (surface and subsurface). By combining the two equations above, the expression

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\frac{\mathrm{d}W}{\mathrm{d}t} \tag{3}$$

states simply that for the globe, over any time interval, a loss in total water storage in either the atmosphere or over the Earth's surface must be balanced by a gain in the other. The expression given by equation (3) is quite impractical in terms of gaining any salient scientific applications or insights into the processes and/or mechanisms that are embedded within the global water cycle – which can vary in strength quite substantially in space and time. As a result, more descriptive and comprehensive versions of equations (1) and/or (2) must be used to depict the processes and mechanisms that comprise the global water cycle.

Atmosphere

When considering the atmospheric column and taking equation (1) over a unit area (or any subdomain of the globe), the water budget must take into account the net horizontal transport of water into and out of the domain's horizontal boundaries. By assuming that diffusive processes of water transport are negligible, and using Gauss' theorem (otherwise known as the *divergence theorem*), the atmospheric water budget is expanded to (c.f. Peixoto and Oort, 1992)

$$\frac{\partial Q}{\partial t} = E - P - \frac{\oint (A_{\mathrm{Q}} \cdot n)\,\mathrm{d}s}{D} \tag{4}$$

in which the rightmost term represents the net total-column (i.e. vertically integrated) horizontal flux of atmospheric water, A_{Q}, around the domain of area D (note that for the unit area this becomes = 1). The unit vector n points outward and normal to the domain boundaries. The horizontal flux term, A_{Q}, is a vector quantity:

$$A_{\mathrm{Q}} = \int_0^{z_{\mathrm{top}}} q\vec{\mathbf{V}}\,\mathrm{d}z \tag{5}$$

where q is the vertical profile of *total* atmospheric water storage (vapor, solid, and liquid) surrounding the given domain. However, for most practices, the solid and liquid components of A_Q are usually ignored (e.g. Peixoto and Oort, 1992). Nevertheless, the dot product will result in an integration of the advective flux that is everywhere normal around the horizontal boundaries. Note also that the time derivative on the left-hand side of equation (4) now reflects the local (i.e. Eulerian) change in atmospheric water storage.

In practice, the expression given by equation (4) is widely used as the basis to calculate regional or zonal accountings of atmospheric water budgets. The flux term in equation (4) is viewed as the analytic aggregation of all the mechanisms that contribute to the "transport" (depicted in Figure 1) of water. The interpretation as to what particular atmospheric processes and/or systems comprise the flux term will depend largely on a few factors that include spatial characteristics of the chosen domain (i.e. size, location, orientation, etc.) and temporal resolution of calculation (length, time of year, etc.).

For the purposes of studying convection and cloud processes (both major facets of the global water cycle, i.e. Figure 1), the vertical transport of water within the atmospheric column and additional (microphysical) processes must be considered, such as phase changes of water and entrainment and detrainment. In such a cloud or convective system, discrete vertical layers of the atmosphere (or horizontal "slices"), and the transport between these layers, are necessarily considered. Therefore equation (4) is further expanded to represent storage changes within a given atmospheric layer, a, in order to comprehensively include vertical transport processes (Emanuel, 1994):

$$\frac{\partial Q_a}{\partial t} = E_a - \left.\frac{\partial (wQ)}{\partial z}\right|_a - P_a - \frac{\oint_a (A_Q \cdot n)\,\mathrm{d}s}{D} \qquad (6)$$

The first two terms on the right-hand side of equation (6) collectively represent the transport of water via turbulent fluxes (i.e. evapotranspiration from the surface), convective processes, and larger-scale motions (i.e. synoptic weather systems). The subscript, a, is applied to E and P to denote the *net effect* of these terms to water storage in the layer. E_a will only be nonzero when the bottom of the atmospheric layer considered is bounded by the Earth's surface and P_a will be nonzero when precipitation forms *within* the layer or the process of collision-coalescence increases the size of precipitation droplets falling through the layer. The subscript a in the rightmost term denotes that the vertical integration (equation 5) is applied only through the thickness of the atmospheric layer considered. The vertical flux divergence, $\left.\frac{\partial (wQ)}{\partial z}\right|_a$, is usually determined via a discrete method, which calculates the difference between the value of wQ at the top and bottom of the layer, normalized by the thickness of the layer. As shown, the water budget implicitly includes evapotranspiration, convection, precipitation, and advection of all phases of water storage. Nevertheless, the form of equation (6) represents the fundamental mechanisms of the atmosphere for the large-scale water transport: precipitation, horizontal transport, convection, and large-scale vertical transport.

Land

For the entire Earth's surface, the cycling of water could be more explicitly expressed as

$$\frac{\mathrm{d}W_\mathrm{l}}{\mathrm{d}t} + \frac{\mathrm{d}W_\mathrm{o}}{\mathrm{d}t} = P_\mathrm{l} + P_\mathrm{o} - E_\mathrm{l} - E_\mathrm{o} \qquad (7)$$

where the l and o subscripts refer to the land and ocean components respectively. Typically, this budget is treated separately for the land and ocean components. When considering only land surfaces, the loss of water storage to the ocean due to runoff (ultimately river discharge) must then be taken into account. Therefore, after pooling all land terms in equation (7) and dropping the subscripts for clarity, the land portion of the equation becomes

$$\frac{\mathrm{d}W}{\mathrm{d}t} = P - E - R \qquad (8)$$

For all global land surfaces, the total runoff term, R, represents the *total* (i.e. surface and subsurface) lateral flow of water that reaches the oceans, and this is primarily in the form of global river discharge. However, when considering a water budget over smaller horizontal domains (i.e. river basins, catchments, etc.), both surface (i.e. stream/river flow and/or surface) and subsurface lateral flow (i.e. groundwater and aquifer flow) would contribute more equally to the runoff term. Surface runoff, which ultimately results as river/stream flow, is caused by a few possible conditions: extreme precipitation events in which the rate of precipitation falling is much faster than the ground can absorb, and precipitation falling atop a saturated and/or impermeable ground surface. Lateral subsurface runoff primarily results from groundwater and/or aquifer flow. In addition, evapotranspiration is the aggregate of two distinct biogeophysical subprocesses: *transpiration* – defined as the water transport and ultimately water vapor loss through the root/stem/leaf/stomata system of the canopy as a result of plant photosynthesis; and *evaporation* – this includes vaporization/sublimation of liquid/frozen water from the soil and snow surface as well as from the vegetation canopy surface (i.e. standing liquid/frozen water that vaporizes/sublimates directly from leaves, branches, etc.). Moreover, for high latitude regions, the storage of frozen water (i.e. snow and ice) and phase changes between

solid and liquid water storage are considerable and represent a considerable water resource as well as influence of climate variations. Thus, in considering all these issues above, the global land water budget equation (8) may then be more comprehensively written as two budget equations:

$$\frac{dL}{dt} = R + C - E^{i} - E^{t} - E^{s} - R^{s} - R^{u} \quad (9)$$

$$\frac{dF}{dt} = S - E_{f} - C \quad (10)$$

where the superscripts i, t, and s refer to the canopy evaporation, transpiration, and soil-surface evaporation respectively. The terms L and F refer to the liquid and frozen storages and fluxes of water respectively. Rainfall and snowfall rates are given by R and S respectively. For the case of frozen water budget, the term E_f then represents the process of sublimation. The superscripts s and u refer to the surface and subsurface (i.e. underground) fluxes. Note that equations (9) and (10) can be applied to either subsurface or surface (i.e. snow and/or canopy) water budgets, noting that for the case of the latter, E^t is zero (i.e. transpiration from the canopy root-system extracts water from the soil subsurface) and C represents the phase change between solid and liquid storage (i.e. melting/freezing). The budgets as given by equations (9) and (10) omit any explicit consideration of water vapor storage within the soil/canopy column. Although the total storage of water vapor is typically considered small compared to the liquid and solid counterparts, the *transport* of water vapor within the soil column can be, at times, important and the process is primarily controlled by diffusion (e.g. Hillel, 1982).

Similar to the case for the atmosphere, the vertical transport (and resulting profile) of water storage is an important issue for land–atmosphere interactions as well as deep-soil hydrologic variations. To represent the relevant processes, a vertical discretization is taken (as done for the atmosphere), in which changes of soil-water storage are considered within a soil layer $-$ l. In this case, vertical water flux, H, must be considered and is typically described by the effects of gravity and mass (i.e. hydraulic) flow (e.g. Hillel, 1982). Thus, changes of soil-water storage within l are given by

$$\frac{dL_l}{dt} = R - E^{i} - \eta_l E^{t} - E^{s} - R^{s} - R^{u} + \left.\frac{\partial H}{\partial z}\right|_l + C_l \quad (11)$$

$$\frac{dF_l}{dt} = S - E_f - C_l \quad (12)$$

Similar to the atmospheric layer water budget, the precipitation (R and S), evaporation (E^i and E^s) terms, and surface runoff terms are nonzero only for a soil layer whose top is at the surface (and thus the soil layer feels the impact of these surface fluxes). The transpiration term, E^t, is nonzero when the soil layer contains roots, and water is thus extracted by plant-system demand. The amount of water extracted by roots within a given soil layer will be a fraction (depicted by η_l in equation 11) of the total transpiration rate, and is controlled by a complex combination of factors that primarily include the thickness of the soil layer considered, the root density in the soil layer, the amount of liquid water storage in the layer, soil type, and plant type. The divergence of vertical water flux, $\left.\frac{\partial H}{\partial z}\right|_l$, is largely the *net* effect of gravitational drainage and diffusive transport of water through the soil layer. The frozen water budget as given by equation (12) crudely captures the fundamental governing mechanisms over the land (i.e. sublimation, phase changes, and frozen precipitation) by treating frozen soil water and snow pack as a single lumped storage. In nature, however, these storages have a complicated vertical and horizontal structure, and, thus, to accurately track the changes of these storages in time, the frozen soil water and snow processes must be accounted in substantially greater detail.

Ocean

Perhaps ironically for the oceans, the change in storage of (fresh) water becomes a rather impractical and untenable variable to diagnose. Therefore, tracking the variations in sea-surface salinity is commonly used as the surrogate to an explicit water budget, and subsequently underscores an important linkage between the global water cycle and Earth's climate system. Given that the input (i.e. precipitation) and release (i.e. evaporation) of freshwater from the ocean, as well as the influx of freshwater from the land's river basins will have a substantial impact on the salinity, S, of ocean water. The expression

$$h\frac{\partial S}{\partial t} = S[E_o - P_o - R] - \upsilon_H \left.\frac{\partial S}{\partial z}\right|_h \quad (13)$$

describes analytically the major processes influencing the (Eulerian) evolution of ocean salinity for a well-defined, homogeneous mixed layer (e.g. Haidvogel and Bryan, 1986) that is free of sea ice. In order to highlight the water cycle processes that play a controlling role in equation (13), the depth of the mixed layer, h, is assumed time invariant (and thus changes in the mixed layer depth are ignored), and υ_H is the turbulent viscosity coefficient. When considering subglobal domains (i.e. ocean or sea basins), R will only be nonzero when the domain is bounded by a land region, and would then be equal to the sum of all river-basin discharge across the land/ocean border. The critical link shown in equation (13) is the vertical and lateral fluxes of (fresh) water directly contributing to change in S, and, in turn, variations in salinity controlling many of the ocean's

major overturning currents/circulations (e.g. thermohaline circulation and bottom-water development). As such, it becomes quite salient that the global water cycle and ocean circulation operate in close harmony, and therefore changes in one system will potentially solicit a response in the other.

In addition to salinity, an equally important form of "storage" of ocean water is sea ice, W_i, as it plays an important role in modulating climate variations and changes. Taking the ocean components of equation (7), and noting that the relevant vertical fluxes would be incident frozen precipitation, P_i, and sublimation, E_i, over sea ice, the budget for sea-ice systems can be given simply as

$$\frac{dW_i}{dt} = P_i - E_i - M + G \tag{14}$$

The form of equation (14) as given, while analytically accurate, masks a number of important, but very complex, processes that necessarily govern sea-ice evolution. For example, the melting term, M_i, would be influenced by such processes as ablation via incident rainfall, radiant energy, and/or heat-flux from the surrounding/underlying ocean or overlying atmosphere, and would also be dependent on salinity. Similarly, the rate of sea-ice growth, G, that is independent of incident precipitation would be a function of the environment and its ability to freeze saline ocean water. Moreover, the budget given by equation (12) omits any consideration of the geometry and spatial distribution of the sea ice, and treats the frozen storage of water as one lumped store.

REFERENCES

Emanuel K.A. (1994) *Atmospheric Convection*, Oxford University Press: New York, p. 580.

Haidvogel D.B. and Bryan F.O. (1986) Ocean circulation modeling. In *Climate System Modeling*, Trenberth K.E. (Ed.), Cambridge University Press: New York, p. 788.

Hillel D. (1982) *Introduction to Soil Physics*, Academic Press: Boston, p. 364.

Peixoto J.P. and Oort A.H. (1992) *Physics of Climate*, American Institute of Physics: New York, p. 520.

175: Observations of the Global Water Cycle – Global Monitoring Networks

DENNIS P LETTENMAIER

Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, US

Until the last few decades, in situ networks were the primary source of information for the estimation of the global water cycle. Such networks provide reasonably good climatological estimates of the major terms in the water cycle over the industrialized parts of the world, but are much less adequate in the lesser-developed countries and in sparsely populated areas. Furthermore, in situ networks have been in the decline over the last several decades. An ongoing trend is toward greater reliance on satellite remote sensing. In some cases (e.g. soil moisture), satellite estimates have characteristics that cannot reasonably be matched by in situ networks. In others, it is less likely that the needs for global observations can be met from remote sensing. Furthermore, there will remain a need for high quality in situ data for algorithm testing and evaluation, and for evaluation of differences between historic in situ-based, and satellite-based, observations. The characteristics of in situ and satellite-based estimates of the dominant fluxes and storages in the land surface and atmospheric branches of the global water cycle are reviewed.

BACKGROUND

"Earth is the water planet." So begins a recent report of the US Global Change Research Program (Hornberger *et al.*, 2001). To be convinced of this, one only needs to look at a globe or any image of Earth from space. Given that over 70% of the Earth's surface is ocean, it is no surprise that the oceans dominate this first impression. Yet, to humans, the ocean arguably is less important a component of the global water cycle than those fluxes – like precipitation and streamflow – and stores of water (like lakes, reservoirs, and groundwater) over land that provide most of the water required for human existence.

While hydrology is an ancient field, dating at least to the first civilizations in the Middle East that developed methods to divert water for irrigation and potable supplies, hydrologists have tended to have a local focus. Recently, however, there has evolved a growing attention to the water cycle as a global system. While it is straightforward to write water balance equations for the land surface, the atmosphere, and the oceans, estimating the terms remains difficult. Legates and Mather (1992) summarize seven previous estimates of global and continental water balances. The range in estimates of the various water balance terms from these studies can be considered to give an indication of the lower bound of the actual errors (because many of the studies used variations of the same data, the estimates cannot be considered independent, and in any event, biases, such as those due to precipitation undercatch errors, are not accounted for). In any event, the normalized midrange of the values from the different studies is about 8% for precipitation, 14% for evapotranspiration, and 16% for runoff. For the individual continents, the errors are considerably larger. This suggests that even for long-term averages over global land areas, most terms are probably not known to an accuracy greater than about $\pm 10\%$ (and errors are most likely larger over the oceans). The reasons for these relatively large uncertainties are twofold. First, over the oceans, measurements of fluxes like precipitation and evaporation are difficult, and the available observations using *in situ* platforms like ships and buoys are sparse. Second, over land, measurement networks, as well as instrumentation, vary across political boundaries, and network density is usually affected by political instability, and national wealth, as well as population density. For the most part, observation densities for key terms in the water

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

budget over land are greatest in the industrialized countries of the Northern Hemisphere.

SCOPE

The water balance equations relevant to this article are:

Land surface (over a river basin):

$$\overline{P} - \overline{E} = Q_s + Q_g + \frac{d\overline{S}}{dt} \quad (1)$$

where $\overline{E}, \overline{P}$ are basin-averaged evapotranspiration and precipitation, Q_s is river discharge, Q_g is groundwater discharge across the basin boundary, total surface and subsurface storage is $S = S_{sm} + S_{sn} + S_{lw} + S_v + S_g + S_{gi}$, with S_{sm} = soil moisture, S_{sn} = snow water content, S_{lw} = lakes and wetlands storage, S_v = vegetation water content, S_g = groundwater storage, S_{gi} = storage in glaciers and ice sheets.

Atmosphere (over an arbitrary land area):

$$\overline{E} - \overline{P} = \overline{\nabla \cdot \vec{Q}} + \frac{\partial \overline{(W + W_c)}}{\partial t} \quad (2)$$

where W is the vertically averaged water vapor and W_c the vertically averaged cloud (liquid) water content.

The oceanic freshwater balance is not treated here in detail. However, it should be noted that the key terms in the global ocean balance are precipitation and evaporation over the ocean surface and river and groundwater flux from the continents. The latter two terms appear also in the land surface water balance (equation 2) and are discussed in Sections "Scope" and "*In situ* methods". Some brief comments are made in the Section "Satellite methods" about the estimation of precipitation and evaporation over the oceans.

Clearly, equations (1) and (2) reflect different perspectives. For the land surface, the implied control volume is from the land surface to an indefinite depth. For the atmosphere, the control volume is from the top of the atmosphere to the land surface. For the oceans, the control volume is from the ocean surface to the ocean bottom. However, as noted earlier, this article considers only the fluxes across the ocean–atmosphere surface and from the land to the ocean. Changes in storage of freshwater and ocean fluxes are not treated – not because they are not important but because the topic and issues are so broad that they cannot reasonably be covered in an article of this length.

Given the topic of the article, we assume that the objective of the networks discussed is to estimate the terms in the various water balance equations at the global scale, or, over land, at the scale of the continents or major divisions thereof, like major continental river basins. This in turn implies timescales of the order of monthly to seasonal, notwithstanding that considerably shorter observation or accumulation intervals might be required to provide accurate estimates at the monthly to seasonal timescale.

IN SITU METHODS

Precipitation

Most nations have precipitation gage networks intended for a variety of purposes. Often there are two classes of networks with some overlap. Real-time observations are made for weather-related purposes like flood forecasting. Some of these observations are archived and are used for climate-related purposes (like estimation of agricultural water requirements). Climate networks, which do not require real-time information, may include more stations that do not report in real time. This is the case in the United States where there are about 5000 operationally reporting precipitation gauges of which about 3500 report through the Global Telecommunications System (GTS) and are archived at the National Climatic Data Center (NCDC) in real time. A much denser observation network of precipitation from accumulation gauges (often daily) operated by unpaid cooperative observers produces data that are also archived (with a delay of several weeks to months or more) at NCDC. In the United States, the primary reporting method for cooperative observer stations is via mailing of summary forms to NCDC once per month, where they are keypunched and archived. NCDC is currently in the process of a modernization program that is upgrading some cooperative observer stations to have a capability for real-time data transmission and archiving.

NCDC also operates the World Data Center for Meteorology (WDCM) that archives precipitation (among other variables) in near real time. Because these data are acquired in near real time, they generally do not meet the quality required for climate and water balance studies. The Global Precipitation Climatology Center (GPCC) archives precipitation data for the land areas of the globe on a delayed basis. The data archived by GPCC come from a variety of sources, including the GTS (in this sense the real-time data should be similar to those received by WDMC), but also receives climate network data from many countries, in some cases with delays (relative to real time) of several years or more. Figure 1 indicates a key problem with gage-based estimates of precipitation over land – the spatial coverage is highly nonuniform. The densities shown in Figure 1 reflect maxima in the industrialized parts of Europe and North America and correspond to about one station per 400 km^2.

Evapotranspiration

Measurement of actual evaporation is difficult, particularly over large areas. Historically, attention has been

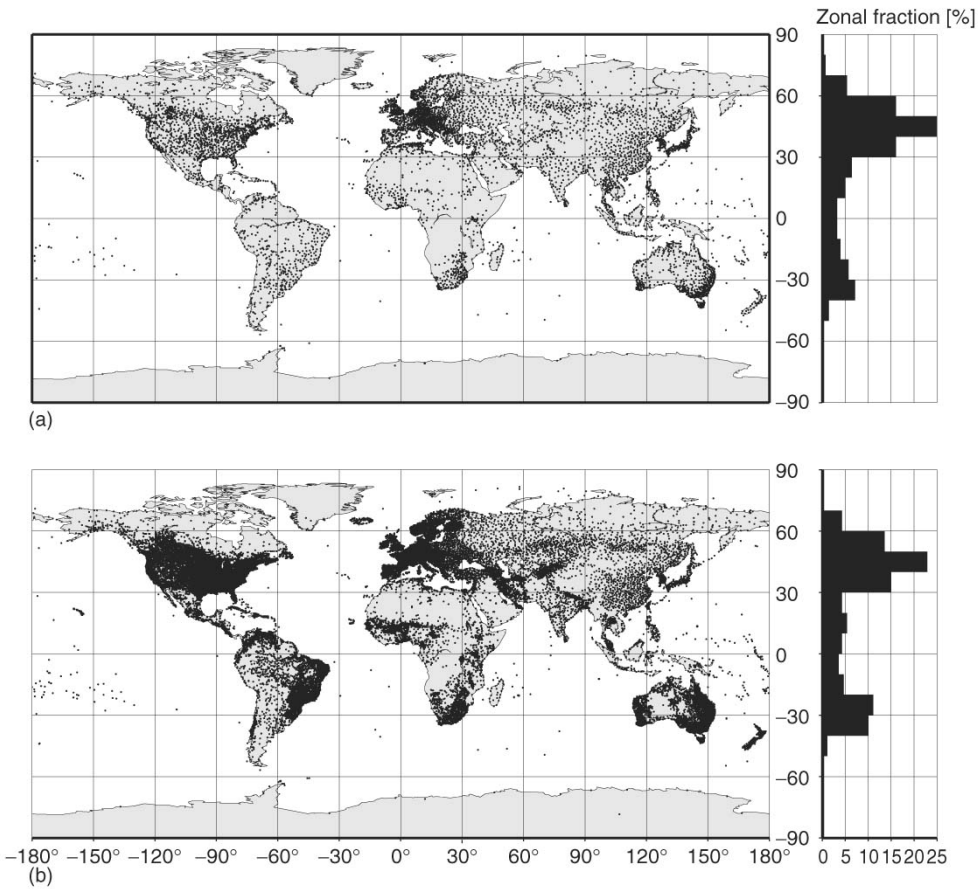


Figure 1 Archived precipitation stations reporting to the Global Precipitation Climatology Centre. (a) shows stations reporting in real time (total approximately 7000); (b) shows climate stations with delayed reporting (total approximately 40 000) (Reproduced from Rudolf, B. and F. Rubel (2005) Global Precipitation. Chapter 11 of Hantel, M. (Ed.). Observed Global Climate. Landolt-Börnstein by permission of Springer)

given mostly to estimation of the evaporative demand of crops, especially in irrigated areas. For this purpose, many countries have established pan evaporation networks that use somewhat standardized procedures, for instance, in the United States, measurements are taken using what is termed the National Oceanic and Atmospheric Administration (NOAA) Class A pan, which is accompanied by wind speed measurements, typically at a height of 2 m above the pan. The main complication with evaporation pan measurements is that they are related more closely to potential than to actual evaporation, where potential evaporation is defined as the evaporation that would occur from a fully wetted surface. Even as an index of potential evaporation, however, interpretation of pan measurements is complicated by the facts that (i) the pan surface is open water, as opposed to a" freely evaporating" land surface consisting of wet bare soil or vegetation that is fully wetted and (ii) there are aerodynamic and thermodynamic effects associated with the physical characteristics and area surrounding the pan.

Weighing lysimeters are a more accurate (but costly) method of measuring evapotranspiration (actual, as opposed to potential). Weighing lysimeters monitor the weight of a soil column, including plants. By monitoring rainfall, evapotranspiration can be computed by difference. If carefully installed, measurements can be quite accurate, although weighing lysimeters are generally most appropriate for agricultural conditions, where the physical setting (minimal lateral moisture flux, for instance) is more faithfully replicated. Dingman (1994; Chapter 7) contains a more detailed discussion of weighing lysimeters.

An indirect method of estimating evapotranspiration over short vegetation is the Bowen ratio method. In this method, gradients of near-surface temperature and specific humidity are measured, and the Bowen ratio (ratio of sensible to latent heat flux, which is proportional to the ratio of the near-surface temperature and specific humidity gradients, assuming equal eddy diffusivities for mass and heat) is used to partition the available energy, Rn – G (where Rn is the net radiation, and G is the ground heat flux, both of which are measured directly), into latent and sensible heat. While neither the latent nor the sensible heat flux is measured directly, the Bowen ratio allows the available energy to be

partitioned into these terms. Standard Bowen ratio systems are available commercially and generally perform best for short vegetation.

An alternate approach that provides direct measurements of evapotranspiration is the eddy correlation method. In the eddy correlation method, high frequency measurements of air temperature, specific humidity, and the vertical component of wind are made near the surface. Averaged overtime, the product of the vertical component of wind and specific humidity represents the transport of moisture from the surface (evapotranspiration). Similarly, when high frequency measurements of carbon dioxide are made, eddy correlation estimates of CO_2 flux can be produced, and in fact many, if not most, eddy correlation systems are installed primarily for that purpose. Among these are regional and global networks such as AmeriFlux, EuroFlux, and Fluxnet (Gu and Baldocchi, 2002). Eddy correlation measurements are usually made from towers, the height of which is proportional to vegetation height, and for forests typically yield "footprints" (the area that affects the measurements) of one to several square kilometers. Over short vegetation, tower heights may be only a few meters, while over forests, tower heights of several tens of meters or more are typical. In any event, while the number of flux towers globally is increasing, routine measurements are probably made at considerably fewer than 1000 sites globally, which is inadequate to provide a coherent spatial picture of evapotranspiration at any scale. The most common use of these towers is to validate model estimates, which then can be applied more widely (see e.g. Betts *et al.*, 1998).

Streamflow

Streamflow is usually measured by continuous recorders of river stage that are converted to discharge via stage–discharge relationships constructed using limited coincident measurements of stage and discharge. Depending on the hydraulic characteristics of the stream cross section and the stability of the rating curve, instantaneous discharge estimates can be accurate to within 5% (larger errors are typical at extreme low and high flows), and considerably lower errors are possible for (e.g. monthly or annual) time averages (Winter, 1981). Because stream discharge provides an areally integrated measurement, for time averages it is one of the most (if not the most) accurately measured of the terms in the water cycle.

Most developed countries have networks of stream gauges that are operated for a variety of purposes, including water management, flood forecasting, and scientific purposes. In the United States, the US Geological Survey maintains a stream gauging network with over 7000 currently active gauges, most of which are telemetered to a central archiving facility. In lesser-developed countries, however, gauge networks are much sparser. The Global Runoff Data Centre in Koblenz, Germany maintains an archive of stream discharge measurements globally. Participation is voluntary, however, and time lags of 10 years or longer are not uncommon for receipt of the data. Furthermore, many countries impose restrictions on release of the data, especially for river basins where there is perceived proprietary interest in the data. In addition to the time lag in acquiring data, many countries (including developed nations like the United States) have reduced support for stream gauging networks over the last two decades.

The result of the time lag and reduction in network support is shown in Figure 2. As shown in the figure, the location of stations globally is highly nonuniform. Also, as shown in the inset, there has been a dramatic reduction in station availability in the last decade. This is of particular concern for studies that make use of remote sensing data, as the period of overlap for many sensors with globally available stream gauge information is minimal.

Lake and Reservoir Storage

Measurement of storage in lakes and reservoirs (and hence storage change, the term of interest for water balance purposes) is straightforward – storage is simply the integral of elevation times the elevation-dependent surface area. Even in the United States, however, there is no integrated network for measurement of lake and reservoir storage. The US Geological Survey maintains some reservoir and lake stage gauges from which storage is estimated, but this is primarily for purposes of water management on a cost reimbursable basis. Other agencies, like the US Army Corps of Engineers, and the US Bureau of Reclamation, observe reservoir storage for water management purposes. There is, however, no central location where these data are archived. Globally, the situation is even more difficult – there is no central archive for lake and reservoir storage data, and even if there were, there are few sources from which the data might be drawn. Lake and reservoir storage is, however, a variable that is amenable to satellite observation as indicated in the Section "Lake and reservoir storage", and as indicated in that section, even now some satellite-based archives are beginning to appear.

Wetlands

At present there is no global observation system for water stored in wetlands, nor are there good estimates of this term in the water budget. In some cases, for instance, the Sud Swamps of Africa, and the Pantanal of South America, the amount of water stored and its seasonal variability can be large. In the Section "Wetlands", options for remote sensing of lake, reservoir, and wetland storage are discussed.

Groundwater

Many countries have groundwater observation networks, although most are associated closely with groundwater

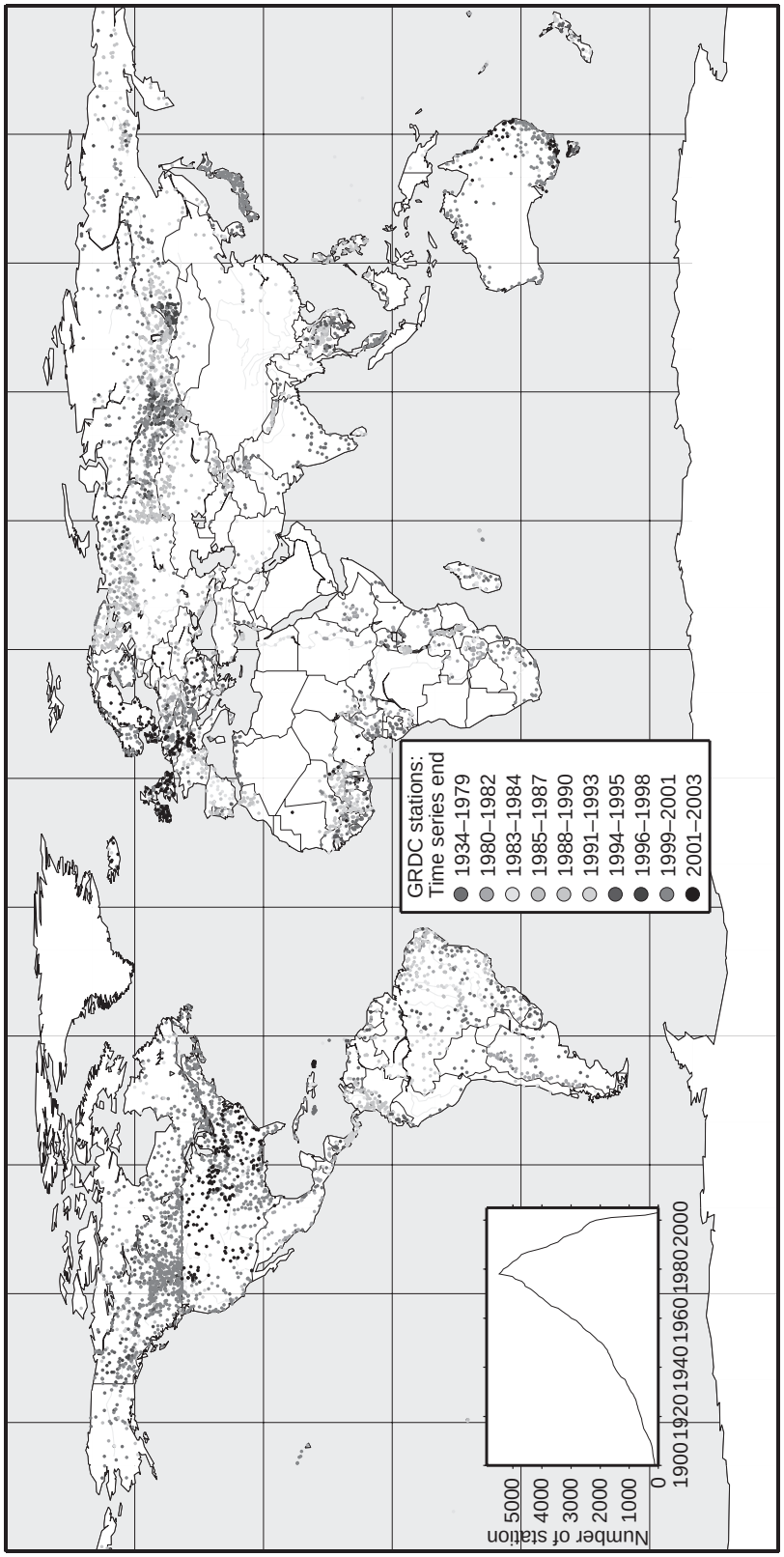


Figure 2 Location and length of period of record of stream discharge records archived by the Global Runoff Data Centre. Inset shows the number of stations archived per year; rapid drop off after about 1990 is due to combination of loss of stations and time lags in data acquisition and archiving. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

management efforts. A significant difficulty in constructing estimates of changes in subsurface storage from well level data is that pumped wells (or wells close to wells that are pumped) are not well suited to providing regional groundwater storage estimates. Unlike precipitation, streamflow, and other surface variables, there is no international archive center for groundwater data, and global data sets are generally sparse.

Soil Moisture

Soil moisture has historically been monitored primarily in association with agricultural concerns, so observations are poorly distributed in time and space. For water balance applications, it is important to monitor changes in soil moisture over a column, typically to a depth of one to several meters. This is most accurately done with weighing lysimeters that are constructed to weigh the soil column to a given depth. However, weighing lysimeters are fairly costly and exist mainly at research sites. Other methods that have been used include neutron probes and time domain reflectometry (TDR), as well as methods based on differential electrical resistance and thermal signatures of soil moisture. Neutron probes generally involve manual readings and hence do not produce time-continuous data, and for that reason have been less used recently. New methods, like TDR, require installation of probes at multiple depths if depth profiles are desired. Neutron probe measurements form the basis for a network of about 20 stations distributed over the state of Illinois, where weekly soil moisture observations have been collected for about 20 years. TDR is used in the Oklahoma Mesonet, which provides, for a shorter period, observations across Oklahoma.

Globally, soil moisture measurements have been collected in a number of countries, notably the Former Soviet Union and Mongolia. These observations, and others from many countries around the globe, have been archived in the Global Soil Moisture Data Bank (**`http://climate.envsci.rutgers.edu/soil_moisture`**; see Robock *et al.*, 2000 for a description). As in the United States, observations are generally poorly distributed in time and space, but nonetheless have proved useful for evaluation of model predictions of soil moisture temporal variability (see e.g. Nijssen *et al.*, 2001, who constructed a transect across Eurasia from the Global Soil Moisture Data Bank for comparison with model simulations). Recently, the United States has installed a skeletal network of soil moisture sensors (Soil Climate Analysis Network, SCAN) that use a frequency-shift dielectric measuring device at multiple depths of 1 m to provide column soil moisture profiles. These sensors provide time-continuous measurements.

A major concern with point measurements of soil moisture is that soil moisture is highly variable in space due to heterogeneity in soil properties at sub-meter spatial scales. However, point observations have nonetheless been shown to have some value for evaluation of temporal variability of area-average soil moistures, primarily because soil moisture is an integrator of precipitation, and therefore its temporal evolution (especially when integrated over depth) tends to respond to long-term (e.g. monthly to seasonal scale) variations in precipitation that usually have relatively long (spatial) correlation lengths. Although networks like SCAN are much too coarse to provide useful information about the continental water balance (other than for local evaluation of models intended for use at continental scales), it does appear that useful information can be provided about larger scales from networks with station separations of 100 km or greater, like those of the Oklahoma network (Vinnikov *et al.*, 1996).

Snow

Snow depth is a relatively easily made measurement, and many precipitation observations networks (e.g. the Cooperative Observer network in the United States) also make routine measurements of snow depth. In Canada, snow depth measurements are made in many parts of the country in lieu of direct measurement of solid precipitation to avoid difficulties with wind catch deficiencies in gauge observations of solid precipitation. The water equivalent of falling (or accumulated) snow is the variable of primary interest for water balance estimates, however, and is more difficult to measure. The most widely used approach has been the use of snow coring devices that are then weighed to estimate the water content of the snowpack. Snow course observations (that usually consist of multiple cores taken in a fixed location) have been widely used to estimate snowpacks in the western United States and are the basis for estimates of spring snowmelt runoff. Because snow course observations are labor intensive, they usually are only made a few times a year, typically near the time of maximum snow accumulation (usually early spring in mountainous areas). In areas with thin snowpacks, high spatial variability of snow depth makes accurate measurement of snow water equivalent difficult.

Over the last 20 years, automated weighing devices have replaced most manual snow courses in the United States. The device used by the Natural Resources Conservation Service in its SNOpack TELemetry (SNOTEL) network monitors pressure in a glycometh-filled bladder ("snow pillow") that reflects the weight, and hence water equivalent, of the overlying snowpack. These devices provide a near-continuous measurement of snow water equivalent (that is telemetered from the usually remote sites to a central data archiving facility). Problems with snow pillows include the influence of local effects that may not be averaged as well as by snow courses, and differential snow accumulation and ablation on and around the pillow due to thermal and other effects. Also, "bridging" of snow across the pillow to the surrounding snowpack, which reduces the snow load on

the pillow and leads to spurious measurements, can occur in some cases. Nonetheless, the US SNOTEL network has been fairly successful and provides the best time-continuous data available on the current status of snowpack in the western United States. Similar networks are used elsewhere, for example, in the mountainous areas of Canada.

Other approaches to measuring snow properties are available as well. Commercially available sonic sensors can provide continuous measurements of the distance from the sensor (that typically is suspended so as to look downward from a fixed measurement point) to the snow surface. Such measurement devices are well suited to remote installations from which the data may be telemetered to a central location. Snow water equivalent can be estimated by measuring the attenuation of naturally emitted gamma radiation through the snowpack. This requires a background measurement (typically performed shortly before the onset of the snow season) to which the measurements over snow are compared. The US National Weather Service operates an aircraft-based gamma system that is routinely flown over the upper Midwest and elsewhere. Their system provides transect estimates of snow water equivalent throughout the winter. There are some limitations to this method in deep snowpacks, where the signal can be too attenuated to provide accurate estimates, and over mountain topography where high spatial variability complicates the interpretation of measurements.

Glaciers and Ice Sheets

Estimates of global frozen water storage as alpine glaciers are based on glacier inventories of glacier (number and area of glaciers) and area-volume scaling methods to estimate total volume. Estimates of the increase or decrease in glacier storage rely on site-specific mass balance studies that make use of direct measurements of snow accumulation, ice and melt loss, and measurements of changes in glacier extent. The World Glacier Monitoring Service (WGMS) publishes estimates of changes of individual glaciers over time, as well as glacier inventories (Haeberli, 1998). WGMS maintains two data sets: a World Glacier Inventory (WGI) that documents the spatial variability of glaciers over the entire globe and the Fluctuations of Glaciers data set that provides a history of a larger number of glaciers, including changes in length, area, elevation, and mass. These data are especially useful for reconstructing the history of "indicator" glaciers that ideally represent the glacial changes in the surrounding region. Unfortunately, these indicator glaciers constitute only a very small fraction of all glaciers globally and their locations are not well distributed. Global monitoring of glacier change is arguably better accomplished by satellite observations (see Section "Glaciers and ice sheets"), and are the only feasible way to measure changes in ice storage in the two great global ice sheets, Greenland and Antarctica.

Atmospheric Moisture Storage and Fluxes

The two terms on the right-hand side of equation 2 are the atmospheric moisture storage and transport respectively. Radiosondes provide a means of directly estimating both of these terms. They consist of a small instrument pack that records and transmits atmospheric pressure, temperature, and humidity from a weather balloon (usually manually launched from a surface observing station) to a receiver at the observing station. The radio signal also allows the position of the balloon to be tracked, and hence its velocity. Usually, profiles of atmospheric moisture are reported at fixed pressure heights, which are specified through knowledge of the atmospheric pressure observations. Integration over the balloon profile provides an estimate of the total column moisture. By tracking the position of the balloon, it is also possible to estimate wind velocity profiles, and hence moisture transport, as well as the total column moisture content, which is the first term on the right-hand side of equation 2. Radiosondes observe moisture as vapor only, and not cloud liquid water or precipitation. Hence radiosonde-based estimates of both atmospheric moisture content and transport are most accurate during nonprecipitating periods (however, Ropelewski and Yarosh (1998) show that the contribution of cloud liquid water to seasonal variations in continental scale moisture budgets is small). Radiosondes are fairly costly (around $200 each), and hence operational launches are usually made 2–4 times per day (twice per day is the United States standard). Station separation is typically several hundred kilometers in parts of the globe that are well observed (e.g. the most populous parts of North American and Europe) and can be much larger elsewhere (e.g. the operational radiosonde network in South America has about 20 stations). Because there can be strong diurnal variations in atmospheric moisture content and low level winds, especially in the interior of the continents in summer, infrequent radiosonde launches can lead to considerable errors in estimates of the atmospheric moisture budget. For instance, Yarosh *et al.* (1999) found that long-term estimates of mean atmospheric convergence from twice-daily radiosondes over central United States resulted in underestimates of as much as 60% when compared with mean river discharge, and in general atmospheric water budget estimates based on radiosonde data alone are not as accurate as those based on atmospheric model analysis fields that are effectively used in the merger of models and observations, both *in situ* and satellite (see e.g. Roads *et al.* (2003)).

SATELLITE METHODS

Precipitation

Satellite estimation of precipitation is based on one or more of several general approaches. The first uses infrared

sensing of cloud-top temperatures to infer precipitation rates. A desirable aspect of these approaches is that the infrared channels on geostationary satellites provide relatively high-resolution (several km spatial resolution) continuous measurements of cloud-top temperatures. Among the most widely used infrared methods is the GOES Precipitation Index (GPI) of Arkin and Meisner (1987), which is based on the number of GOES pixels within a larger area (typically of order 50 km × 50 km) for which the cloud-top temperature is less than a threshold. In general, the GPI method is most successful for deep convection in the tropics. Numerous refinements of this approach have been reported, but all suffer from the disadvantage that the measurement is indirect and is only applicable to convective storms. Furthermore, even for convective storms, rain rates can vary widely for the same cloud-top temperature.

The primary current satellite-based alternative to infrared methods is based on passive microwave remote sensing, and utilizes the principle that the combined microwave emission signature of a precipitating atmosphere and the surface background differs from that of a dry atmosphere. The advantage of this method is that it is more closely related to the precipitation process than are methods based on cloud-top temperature. However, these methods suffer from the fact that all of the candidate sensors are flown in low earth orbit (LEO), and even so, spatial resolutions of the retrievals are quite coarse, typically several tens of kilometers. Furthermore, LEO implies instantaneous "snapshots", rather than continuous observations, and hence relatively large sampling errors. For this reason, most passive microwave products are only usable when aggregated to relatively long time intervals, unless some additional information source (e.g. *in situ* observations) are used to create a merged product. Recent work by Joyce *et al.* (2004) has attempted to combine passive microwave and infrared measurements so as to take advantage of the superior accuracy of passive microwave with the superior spatial resolution and temporal sampling of infrared algorithms.

The Tropical Rainfall Measurement Mission (TRMM) satellite includes three instruments that are targeted toward rainfall estimation: a precipitation radar, a microwave imager, and a near-infrared visible scanner. The precipitation radar provides valuable information about precipitation microphysics, and hence better rainfall rate information, than is possible from either infrared or passive microwave sensors alone. The main disadvantages of TRMM-based products are that the overpass is only roughly once per day, so sampling errors are severe, and the orbit goes only as far north and south of the equator as 35°. The planned Global Precipitation Measurement mission (GPM) is expected to alleviate both of these problems by providing, via a set of "drone" satellites with passive microwave sensors, three-hourly passive microwave signatures and daily radar overpasses. The orbit is expected to extend to roughly 60–65°. Nonetheless, sampling errors will remain a significant issue (Nijssen and Lettenmaier, 2004), as will errors associated with information transfer from the precipitation radar to the microwave drones.

Evapotranspiration

All satellite-based methods of estimating evapotranspiration are indirect. Most of them are based loosely on the principle that over an area, there will be some mixture of vegetation and bare soil, and that the evaporating vegetation will have a different thermal signature than bare soil. Hence, the most successful algorithms use visible and near-infrared sensors to extract a vegetation index (VI), and from the surface radiative temperature to an inference about the corresponding surface temperature T_s (Jiang and Islam, 2003). The MODIS algorithm, for instance, (Nishida *et al.*, 2003) is based on estimation of the evaporative fraction of a mixture of bare soil and vegetated surfaces, via isolation of end members of a vegetation index over a scene with a large enough number of pixels to allow estimation of the endpoint surface temperature signature. Most of these so-called VI–Ts (vegetation index–surface temperature) methods work best when they are used in combination with some *in situ* data (typically including one or more of surface wind, vapor pressure deficit, and net radiation and ground heat flux), although Nishida *et al.* show how some of the requirements for *in situ* data can be relaxed via various assumptions. Satellite evapotranspiration algorithms generally require some calibration and are susceptible to instrument change. On the other hand, they work with visible and near-IR data, of which there are a number of current (and past) candidate sensors.

Streamflow

Birkett *et al.* (2002) have demonstrated the capability to measure the stage of very large (width $>\sim$1–2 km) rivers using an existing radar altimeter (TOPEX/Poseidon) designed primarily for ocean altimetry. Figure 3, for example, taken from Birkett *et al.* (2002) shows that the observed river stage at selected locations along the Amazon River could be reproduced to within several tens of centimeters under most conditions. Alsdorf and Lettenmaier (2003) argue that the potential exists using current technology to provide stage estimates with accuracy well under 10 cm and for rivers with widths of order 100 m (vs. several kilometers for the Amazon cross sections analyzed by Birkett *et al.*). The main "weak link" in satellite estimation of river discharge is the inability to measure river velocity remotely. However, Birkett *et al.* (2002) show that surface slope is potentially measurable, which offers some hope that discharge might be derived via a combination

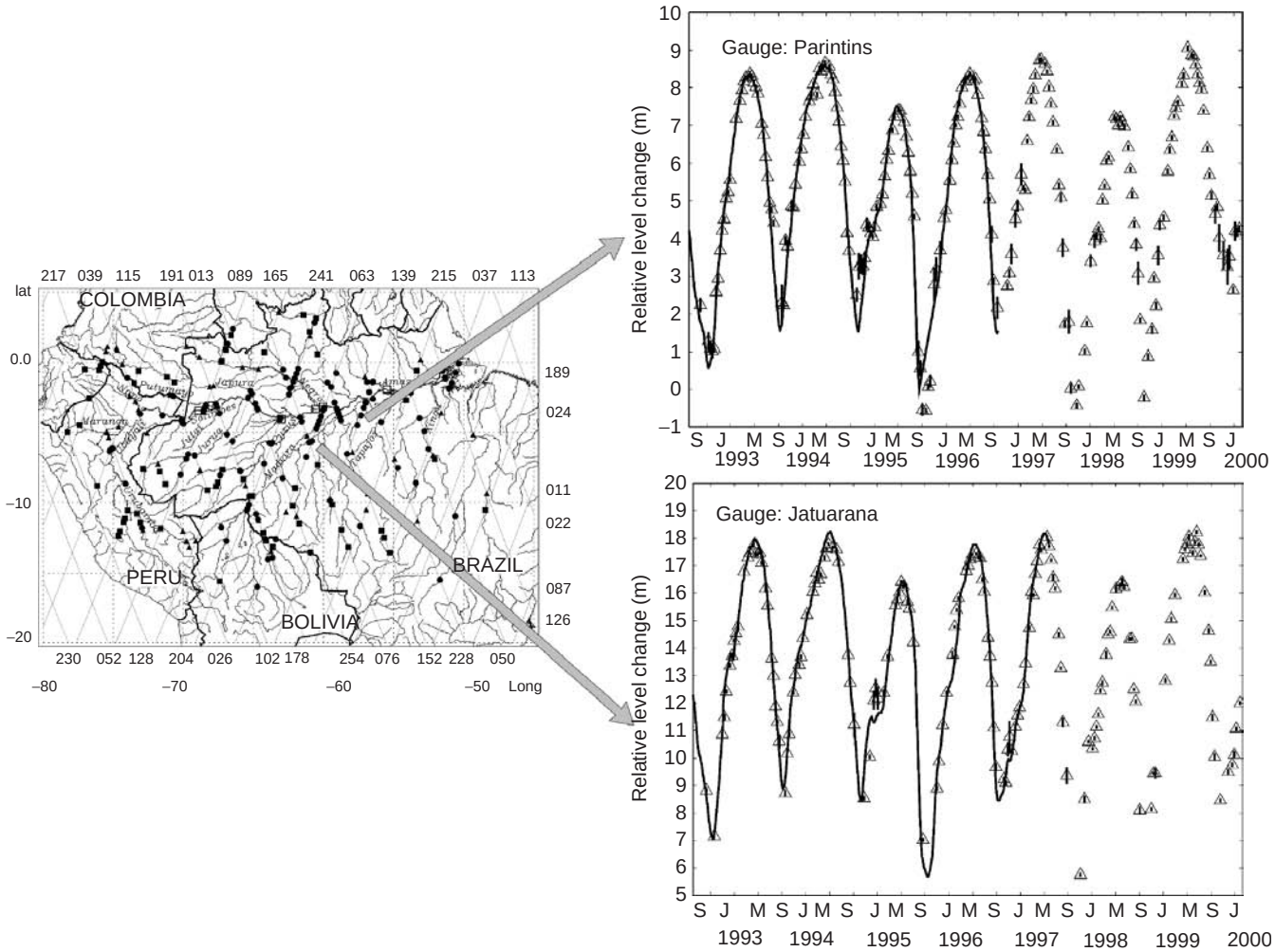


Figure 3 Remotely sensed (from TOPEX-POSEIDON; triangles) and *in situ* (solid black) measurements of river stage at two locations in the Amazon basin (Reproduced from Birkett *et al.* (2002) by permission of American Geophysical Union (AGU))

of stage measurements and hydraulic calculations for discharge, although complications remain in determining the channel cross section and its hydraulic characteristics (e.g. roughness).

Lake and Reservoir Storage

Radar altimeters have the capability to measure the stage of very large lakes and wetlands. Birkett (1995) have shown the capability to observe the stage of large water bodies such as large lakes in Africa to accuracies of about 10 cm using existing satellite radar altimetry. Existing radar altimeters (TOPEX/Poseidon and Jason-1) are designed to provide accurate measurements of the ocean surface, but over relatively large footprints. However, technology exists via hardware and signal processing to provide estimates of water surface elevation for much smaller (e.g. inland) water bodies, with spatial dimensions well less than 1 km (Alsdorf *et al.*, 2003). The potential also exists to provide sub-1 cm accuracy altimetry over relatively small footprints using laser altimetry (e.g. ICESAT).

Absent vegetation, lake, and reservoir extent can be measured by a variety of visible band sensors, or in the presence of flooded vegetation, Synthetic Aperture Radar (SAR) with relatively long wavelength (L band) has been demonstrated. While open water extent of large water bodies is measurable using operational satellites (LandSat, AVHRR), the accuracy of horizontal extent measurements is determined by sensor spatial resolution, and detection of storage change for shallow water bodies is generally less accurate than for deep ones.

Wetlands

Satellite-based methods for detection of change of storage in wetlands are essentially the same as those that are

applicable to lakes and reservoirs. Radar altimeters are sensitive to flooded vegetation, so an adequate open water footprint (currently order of 10 km) is essential, although there is the potential to reduce this considerably. Feasibility of detecting changes in storage in the Sud Swamps, for instance, has been demonstrated by Birkett (1995). The same complications noted in Section "Lake and reservoir storage" for lakes and reservoirs pertain to the estimation of the extent of wetlands, which usually involve determination of the extent of flooding of vegetation, which is best accomplished by L-band radar.

Groundwater

There is no existing remote sensing method that is able to measure directly (or indirectly for that matter) groundwater levels. However, the microgravity sensor onboard the Gravity Recovery and Climate Experiment (GRACE) has the potential to provide indirect large area ($o(10^6\ km^2)$) estimates of subsurface moisture variations in time. The microgravity sensor records variations in the total gravity field "felt" by the satellite. Over relatively short time intervals (order of a month or so), these result primarily from variations in moisture content of the atmosphere, the land surface (lakes, wetlands, snow), and the subsurface (soil moisture, groundwater). Atmospheric moisture is relatively well observed (or estimated by weather prediction models that combine observations and model predictions via data assimilation) to the accuracies required to adjust for the atmospheric influence. The other terms – snow, surface water, and soil moisture – can be estimated to varying accuracy via land surface modeling and land data assimilation, although the supporting models and observations are only now evolving. Swenson *et al.* (2003) have performed synthetic experiments to estimate the accuracy with which GRACE data will be able to estimate variations in total (surface plus subsurface) water storage at seasonal time intervals and at the scale of large continental river basins. Discussions have begun on a "next generation" microgravity sensor that might have an effective footprint of the order of $10^4\ km^2$. At this spatial resolution, the estimation of the dynamics of regional groundwater systems will become possible.

Soil Moisture

The microwave emission brightness temperature from the land surface varies strongly with soil moisture content to a depth of approximately one-quarter of the wavelength. Thus passive microwave measurements at lower frequencies (longer wavelengths) provide better estimates of the near-surface soil moisture and are less influenced by overlaying vegetation. Additionally, the sensitivity of the microwave brightness temperature to soil moisture is larger at lower frequencies. Collectively, these factors dictate that sensors that measure brightness temperatures at low microwave frequencies are preferred for estimating surface soil moisture. This desire to measure at the lowest possible frequency is countered by the fact that the antenna size (e.g. diameter) increases with wavelength for a fixed spatial resolution. Because of technological limitations of antenna size (and hence spatial resolution with which soil moisture variations can be inferred), the lowest frequency passive microwave sensor that has reasonable spatial resolution is L-band (~1.4 GHz). At this frequency, soil moisture can be retrieved reliably from vegetated surfaces with a vegetation water content of roughly $5\ kg\ m^2$ (corresponding approximately to dense cropland or shrubs, but not to forests).

Several passive microwave instruments have been flown with somewhat shorter wavelengths that have facilitated proof of concept studies targeted at areas with relatively sparse vegetation. For instance, the Scanning Multichannel Microwave Radiometer (SMMR) instrument was flown on the SeaSat and Nimbus-7 satellites from the 1970s to mid-1987. Its lowest frequency was 6.8 GHz at which the footprint was about 150 km. Owe *et al.* (1992) evaluated the feasibility of estimating surface soil moisture over a relatively sparsely vegetated (shrub/grass savannah) area of southern Africa. Some attempts have been made to use the higher frequency 19 GHz channel of the operational Defense Meteorological Satellite Program (DMSP) Special Sensor Microwave/Imager (SSM/I) instruments. While the footprint at this frequency is about 25 km, the frequency is so high that except in areas of very sparse vegetation, soil moisture products derived from SSM/I primarily reflect vegetation moisture. An intermediate frequency (10.7 GHz) is available from the (TRMM) microwave imager (TMI). Although the TRMM orbital inclination of about 35° precludes observations over temperate and high latitudes, some success has been demonstrated with TMI-based soil moisture products in areas of sparse vegetation, like the US Southern Great Plains (Gao *et al.*, 2004). The Advanced Microwave Sounding Radiometer (AMSR) was launched on the NASA EOS Earth Observing System Aqua satellite in mid-2002. Its lowest frequency (6.9 GHz) is close to that as SSMR, but the spatial resolution is considerably higher due to a larger antenna. However, early attempts to derive soil moisture products from the AMSR-E 6.9 GHz channel have been confounded by radio interference, especially over North American and northern Europe.

Both the European Space Agency and NASA have now planned launches of L-band passive radiometers designed specifically for soil moisture measurement. The European Soil Moisture and Ocean Salinity (SMOS) instrument will most likely be launched first. It is a two-dimensional instrument and will use a Y-shaped antenna that will support a product spatial resolution of about 50 km. The NASA Earth System Pathfinder HYDROS mission uses a somewhat different approach. It will consist of a multipolarization L-band radiometer capable of producing a 40-km spatial

resolution product, and an unfocused L-band Synthetic Aperture Radar that will allow disaggregation of the passive microwave product to a spatial resolution of 1–3 km. It will have a revisit time of 2–3 days over all global land areas. It is expected to be launched in about 2010.

Snow

Snow areal extent can be observed via visible sensors and a variety of algorithms have been developed. In the United States, the NOAA Operational Hydrologic Remote Sensing Center produces an operational product on the basis of daily GOES and AVHRR data at approximately 1 km spatial resolution. Coarser resolution global products that use both AVHRR and SSM/I passive microwave sensors are produced for operational use by the National Environmental Satellite Data Service (NESDIS). Higher resolution products based on LandSat TM and other high-resolution sensors have also been developed. These high-resolution products are better able to resolve partial area coverage of snow. Among the key problems with satellite snow areal extent from visible band sensors are discrimination of clouds from snow and detection of snow under forest cover. More recent MODIS-based products at 500 m resolution are available daily in near real time, and are better able to discriminate snow from clouds and to detect snow under forest canopies than AVHRR-based products (Maurer *et al.*, 2002).

Passive microwave algorithms can be used to estimate snow water equivalent (SWE). An operational product based on SSM/I 37 and 85 GHz channels developed by the Meteorological Service of Canada is used operationally in Canada over areas with thin, cold snowpacks (prairies and tundra). The spatial resolution of this product is relatively coarse (approximately 25 km). AMSR-based SWE products are currently becoming available. Problems with all passive microwave SWE products include dependence of brightness temperature, and hence SWE estimates, on grain size, saturation of the signal from deep snowpacks, and inability to provide measurements for wet snow (although wet and dry snow can be discriminated). Radar-based algorithms have also been researched and show some promise, although they generally require more surface information (e.g. surface roughness) than passive microwave methods.

Glaciers and Ice Sheets

The Glacier Land Ice Measurements from Space (GLIMS) program is an effort to estimate changes in glacier extent and storage using satellite remote sensing such as interferometric SAR (various instruments) and lidar altimetry (ICESAT). However, this effort has been underway for only a few years. Use of older visible band satellite data (covering the last three decades) has facilitated estimates of change of glacier extent, but these data are not able to provide the surface topographic measurements that are needed to estimate changes in glacier mass storage accurately. For the two great ice sheets, Antarctica and Greenland, satellite measurements are the only feasible ways to measure change in storage. Although changes in ice sheet volume and area are increasingly easier to obtain, changes in mass continue to be a problem due to a lack of information on density over the ice sheet and submarine mass loss though bottom melting of the ice shelves.

Atmospheric Moisture Storage and Fluxes

Because of the high cost of radiosonde launches (around $200 per instrument package, not including labor costs) and the sparseness of the global network, there has long been an interest in remote sensing alternatives. Passive microwave sounders have been used since the 1970s and are able to retrieve profiles of atmospheric moisture and temperature at horizontal resolutions that have improved to less than 1° latitude–longitude, and vertical resolutions that now approach 1 km. The most recent series of operational instruments are the Advanced Microwave Sounding Unit (AMSU) that flies on the NOAA operational satellites. These data are assimilated into the global weather prediction models run by most major weather centers. Microwave sounding units do not supply information about the atmospheric wind profile, which is a major source of error in atmospheric moisture flux estimates (second term on the right-hand side of equation 2) globally. Furthermore, estimates of moisture flux and total column moisture content based on microwave soundings are impeded by the lack of detailed information in the lower troposphere. On the other hand, analysis fields of moisture profiles and fluxes produced via data assimilation in global models are considerably more accurate than estimates derived directly from observations, which implies that despite the absence of wind information, there is considerable information content in the satellite profiles.

Recently, options for retrieving atmospheric moisture information from global positioning satellites have been pursued. The basis of the method is that the deformation in the signal path (hence time delay) through the atmosphere between two GPS satellites is a function primarily of the moisture content of the atmosphere along the path. Hence, it is possible to infer an effective moisture content along the path. Furthermore, because many pairs of satellites are available, it is possible at any time to retrieve moisture information along many paths at a range of heights in the atmosphere. This information is fundamentally different from that obtained from sounding units, though it does not produce a consistent profile, but rather information for a series of nearly random paths. On the other hand, there is no constraint on vertical resolution; the path depends only on the location of the two satellites at the time the signal was sent and received. Like sounding data, the power of these observations will lie in the development

of data assimilation methods that can effectively exploit the information. In common with microwave soundings, no information is provided about wind.

SUMMARY

Understanding the global water cycle demands measurements that allow estimation of the dominant hydrologic storages and fluxes at spatial scales from local to continental and global, and at temporal scales sufficient to resolve the diurnal variations that are driven by Earth's ultimate source of energy, the Sun. The arrival of the satellite era over 30 years ago has lead to an earth system science approach to observations that is perhaps most apparent in the genesis of NASA's EOS satellites. Nonetheless, satellites are unable to measure, or to measure accurately, a number of the key water cycle storage and flux terms globally. For instance, while great strides have been made in satellite estimation of precipitation, and atmospheric moisture and moisture fluxes, key land surface terms like streamflow and lake and wetland storage, soil moisture, and groundwater are poorly observed, or not at all, via remote sensing. The global networks of *in situ* observations that are needed to provide estimates of these water cycle terms generally are unable to do so – in some cases because the observations are so sparse as to provide adequate global coverage (as in the case of soil moisture), and in others because of data access problems and/or lack of coordination among nations (as in the case of streamflow data), or due to a combination of these and other limitations (as in the case of groundwater data).

Generally, *in situ* networks of water cycle and other climate-related data have been in the decline over the last several decades, due in large part to cost, especially in remote areas, but for other reasons that are more difficult to ascertain as well. It appears that the trend with respect to global observations is toward greater reliance on satellite remote sensing, and in some cases, improved technology (e.g. as in the case of soil moisture) will doubtless lead to global observations with characteristics that cannot reasonable be matched by *in situ* networks. In others, it is less likely that the needs for global observations can be met from remote sensing. Furthermore, even in cases where remote sensing can eventually provide good estimates of water cycle variables, there will remain a need for high quality *in situ* data for algorithm testing and evaluation, and for evaluation of differences between historic *in situ*-based and satellite-based observations.

REFERENCES

Alsdorf D.A. and Lettenmaier D.P. (2003) Tracking fresh water from space. *Science*, **301**, 1491–1494.

Alsdorf D.E., Lettenmaier D.P., Vörösmarty C. and The Nasa Surface Water Working Group. (2003) The need for global, satellite-based observations of terrestrial surface waters, *EOS Transactions American Geophysical Union*, **84**, 269–276.

Arkin P.A. and Meisner B.N. (1987) The Relationship between large-scale convective rainfall and cold cloud over the western hemisphere during 1982–84. *Monthly Weather Review*, **115**, 51–74.

Betts A.K., Viterbo P., Beljaars A., Pan H.L., Hong S.Y., Goulden M. and Wofsy S. (1998) Evaluation of land-surface interaction in ECMWF and NCEP/NCAR reanalysis models over grassland (FIFE) and boreal forest (BOREAS). *Journal of Geophysical Research*, **103**, 23079–23085.

Birkett C.M. (1995) The contribution of TOPEX/POSEIDON to the global monitoring of climatically sensitive lakes. *Journal of Geophysical Research*, **100**, 25179–25204.

Birkett C.M., Mertes L.A.K., Dunne T., Costa M.H. and Jasinski M.J. (2002) Surface water dynamics in the Amazon basin: application of satellite radar altimetry. *Journal of Geophysical Research*, **107**, Art. No. 8059.

Dingman S.L. (1994) *Physical Hydrology*, Prentice Hall: p. 575.

Gao H., Wood E.F., Drusch M., McCabe M., Jackson T.J. and Bindlish R. (2004) Using TRMM/TMI to retrieve soil moisture over the southern United States from 1998 to 2002, paper presented at *European Geophysical Union Annual Meeting*, Nice, April, 2004.

Gu L.H. and Baldocchi D. (2002) Fluxnet 2000 synthesis – foreword. *Agricultural and Forest Meteorology*, **113**, 1–2.

Haeberli W. (1998) Historical evolution and operational aspects of worldwide glacier monitoring. In *Into the Second Century of World Glacier Monitoring - Prospects and Strategies*, Haeberli W., Hoelzle M. and Suter S. (Eds.), UNESCO publishing: Paris, pp. 35–51.

Hornberger G.M., Aber J.D., Bahr J., Bales R.C., Beven K., Foufoula-Georgiou E., Katul G., Kinter J.L., Koster R.D., Lettenmaier D.P., McKnight D., Miller K., Mitchell K., Roads J.O., Scanlon B.R. and Smith E.A. (2001) *A Plan for a New Science Initiative on the Global Water Cycle*, U.S. Global Change Research Program: Washington.

Jiang L. and Islam S. (2003) An intercomparison of regional latent heat flux estimation using remote sensing data. *International Journal of Remote Sensing*, **24**, 2221–2236.

Joyce R.J., Janowiak J.E., Arkin P.A. and Xie P. (2004) CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydromet.*, **5**, 487–503.

Legates D.R. and Mather J.R. (1992) An evaluation of the average annual global water balance. *Geographical Review*, **82**, 253–267.

Maurer E.P., Wood A.W., Adam J.C., Lettenmaier D.P. and Nijssen B. (2002) A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States, *Journal of Climate* **15**, 3237–3251.

Nijssen B. and Lettenmaier D.P. (2004) Effect of precipitation sampling error on simulated hydrological fluxes and states: anticipating the global precipitation measurement satellites. *Journal of Geophysical Research-Atmospheres*, **109**, Art. No. D02103.

Nijssen B., Schnur R. and Lettenmaier D.P. (2001) Global retrospective estimation of soil moisture using the variable infiltration capacity land surface model, 1980–1993. *Journal of Climate*, **14**, 1790–1808.

Nishida K., Nemani R.R., Running S.R. and Glassy J.M. (2003) An operational remote sensing algorithm of land surface evaporation. *Journal of Geophysical Research*, **108**, Art. No. D4270.

Owe M., Van de Griend A.A. and Chang A.T.C. (1992) Surface moisture and satellite microwave observations in semiarid southern Africa. *Water Resources Research*, **28**, 829–839.

Roads J., Lawford R., Bainto E., Berbery E., Chen S., Fekete B., Gallo K., Grundstein A., Higgins W., Kanamitsu M., *et al.* (2003) GCIP Water and Energy Budget Synthesis (WEBS). *Journal of Geophysical Research*, **108**, Art. No. D002583.

Robock A., Vinnikov K.Y., Srinivasan G., Entin J.K., Hollinger S.E., Speranskaya N.A., Liu S. and Namkhai A. (2000) The global soil moisture data bank. *Bulletin of the American Meteorological Society*, **81**, 1281–1299.

Ropelewski C.F. and Yarosh E.S. (1998) The observed mean annual cycle of moisture budgets over the central United States (1973–92). *Journal of Climate*, **11**, 2180–2190.

Rudolf B. and Rubel F. (2005) Global Precipitation. Chapter 11 of Hantel M. (Ed.). *Observed Global Climate*. Springer Verlag, Landolt-Börnstein.

Swenson S., Wahr J. and Milly P.C.D. (2003) Estimated accuracies of regional water storage variations inferred from the Gravity Recovery and Climate Experiment (GRACE). *Water Resources Research*, **39**, Art. No. 1223.

Vinnikov K.Y., Robock A., Speranskaya N. and Schlosser C.A. (1996) Scales of temporal and partial variability of midlatitude soil moisture. *Journal of Geophysical Research*, **101**, 7163–7174.

Winter T.C. (1981) Uncertainties in estimating the water balance of lakes. *Water Resources Bulletin*, **17**, 82–115.

Yarosh E.S., Ropelewski C.F. and Berbery E.H. (1999) Biases of the observed atmospheric water budgets over the central United States. *Journal of Geophysical Research*, **104**, 19349–19360.

176: Observations of the Global Water Cycle – Satellites

FRANKLIN R ROBERTSON

NASA/Marshall Space Flight Center, Huntsville, AL, US

Satellite observations of Earth's climate system have evolved dramatically since the first crude video images of cloud cover taken over 40 years ago. Today, spaceborne measurements of precipitation, cloud attributes, radiative fluxes, and surface properties make crucial contributions to monitoring key processes and the evolving state of the planet's hydrologic cycle. We illustrate how this maturing capability has enabled study of the hydrologic cycle in a global context that goes beyond consideration of local or regional water balance. Remote sensing has become a key technology for addressing such interdisciplinary questions as whether the global hydrologic cycle is changing and if so how? What key processes linking the atmosphere, biosphere, and hydrosphere are at play, and what are the prospects for predicting extreme hydrologic events of significant societal impact? The discussion is motivated first by considering these key science problems. Following a brief overview of key remote sensing concepts, we then survey existing capabilities to determine various quantities such as precipitation, evaporation, soil moisture, turbulent fluxes, cloud cover, water vapor, and the links to radiative fluxes. Both successes and persisting challenges are discussed. An assessment of where we stand in developing an integrated picture of the hydrologic cycle and climate is presented, along with anticipated improvements in the future measurements.

In the context of climate, we demonstrate that in spite of substantial improvements in determining hydrologic variables and processes from space, some key challenges remain. Our examination here of signals in precipitation, evaporation, and radiative fluxes suggests that at present we can detect regional patterns of interannual variability with considerable skill; we can resolve globally integrated signals with only modest certainty; and for the most part, we have little confidence yet in decadal changes or trends. Nevertheless, the situation is encouraging because in many cases these retrievals have been done without regard to intercalibration of sensor data streams, and in the case of operational retrievals, algorithm modifications which result in spurious trends. Careful reprocessing is likely to yield substantial benefits. Future planned missions with active remote sensing at 94 GHz (CloudSat) and interferometric measurements at 1.4 GHz promise to expand our knowledge of hydrologic processes of crucial importance to climate.

INTRODUCTION: WHY A REMOTE SENSING APPROACH?

In 2003, the National Aeronautics and Space Administration (NASA) launched the Aqua spacecraft with the mission of monitoring the global hydrologic cycle. The diverse sensors aboard, spanning the visible to microwave portion of the spectrum, represent an amazing evolution of remote sensing from 1960, when the first grainy video pictures of clouds were beamed back to Earth from the weather satellite Television Infrared Operational Satellite (TIROS) I. (See IEEE Transactions on Geoscience and Remote Sensing, February 2003, for a Special Issue on the EOS Aqua mission). Aqua joins Terra, TRMM (Tropical Rainfall Measuring Mission), Envisat and satellite platforms from various other international efforts that now provide monitoring and research capability for the global water cycle. Motivation for this impressive expansion of measurement capabilities for viewing our planet from space has come from the challenge of developing a globally extensive yet locally resolved description of the planet's water cycle, as well as its connections to Earth's radiative balance and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

biogeochemical processes. Water is dispersed over atmosphere, ocean, and land/biosphere. It exists in three phases: in the atmosphere and ocean, as fluid parts of the earth system, as life-sustaining liquid in the biosphere, and as seasonal and "permanent" ice in the cryosphere. For various cost and technological reasons, comprehensive *in situ* measurements of these reservoirs and the fluxes of water and energy between them are not feasible. A prominent deficiency is the lack of measurements in the vast, uninhabited regions of the globe such as the oceans, polar regions, deserts, and tropical rainforests. Aside from the existence of large data-void areas, even existing available point measurements of rainfall, snow accumulation, and turbulent heat fluxes are often difficult to interpret. Sampling biases and representativeness of point measurements (especially for rainfall) are difficult to contend with, when correlation length and timescales of hydrologic processes are far smaller than instrument spacing.

Remote sensing advances have allowed us to relax some of these sampling uncertainties and obtain a more complete picture of the distribution of water over the planet, documenting, for example, how exchanges of water between the atmosphere and surface are related to radiative, chemical, and biological processes. Global mapping of cloud properties, precipitation, water vapor, sea ice, snowcover, ocean topography and heat content, and other variables is now routinely done on a daily basis. Much progress has been made in determining the optical and microphysical properties of clouds, their bulk mass of liquid and ice water, some coarse vertical structure of water substance, and how winds redistribute water (and heat) on scales as small as individual storm systems that affect our weather. At the same time, satellite platforms have provided the ability to relate these atmospheric properties to sea-surface temperature and near-surface soil-moisture variations during the past two decades and thereby improved our understanding of year-to-year climate disruptions (e.g. *El Nino*) and even trends in certain variables (sea-ice distributions, surface temperature).

To be sure, there are significant remaining challenges for improving remote sensing of the global water cycle. The error properties of current sensors and their algorithms are far larger and less well understood than desirable. Historically, remote sensing has been done primarily with passive instruments, interpreting Earth's emitted radiation in terms of geophysical variables. Passive systems have certain limitations, most notably, the inability to provide detailed atmospheric vertical profiling. The inability to directly sense soil moisture several centimeters below the surface even with today's advanced passive microwave systems is another example. Technological considerations also currently restrict geostationary satellite measurements to visible and infrared (IR) wavelengths, which provide only cloud-top temperature and brightness information. This renders monitoring the life cycle of rainfall and the evolution of cloud systems from space problematic. Even the stability of satellite orbits over time induces diurnal sampling biases that, over time, may contribute substantially to retrieval uncertainties. However, these technology challenges are not insurmountable and are currently being addressed. Active systems that emit and capture reflected radiation are the promise of the future. Radar measurements from TRMM have provided unprecedented surveys of vertical profiling of precipitation.

As we begin this survey, we need to clarify just what we mean by the global water cycle. The most inclusive description of the hydrologic cycle of the planet encompasses not only the mass cycling of water but also connections to heat and radiation balance, as well as the role of water in biological processes like net primary productivity. In a sense, the water cycle is a key signature of the "metabolic rate" of the planet. While this integrated view is extremely important, we will, for the purposes of a tractable discussion, limit our focus to the physical components of the global water cycle and to some key connections to radiation and the planetary energy balance. Consider, for example, that although the climate system is very near radiative balance at the top of atmosphere (TOA), the atmosphere itself and remaining land/ocean/ice complex are not. Balancing the global mean $168\,W\,m^{-2}$ of solar energy absorbed at the earth's surface are a net cooling of $66\,W\,m^{-2}$ from net long-wave radiative flux, $24\,W\,m^{-2}$ by sensible heat flux, and $78\,W\,m^{-2}$ by latent heat flux (Kiehl and Trenberth, 1997). Realization of evaporation as latent heat of condensation in the atmosphere has major consequences and, in fact, determines, in complex fashion, how radiative fluxes are partitioned through the earth's climate system. Paradoxically, while only a miniscule amount of water on this planet resides in the atmosphere, its controlling effect on habitability of the planet through impacts on the heat budget is prominent. Our discussion focuses primarily on remote sensing capabilities from space that are most relevant to the perspective of climate. We shall emphasize certain science questions that drive observational requirements and examine the current capabilities and weaknesses of the observational strategy. We conclude with some comments on expected future observational systems.

MOTIVATING/DRIVING QUESTIONS

Improved satellite observations have stimulated research on interdisciplinary science topics that span meteorology and climate, hydrology, oceanography, and other Earth sciences. In part, this has been because space-based remote sensing requires that algorithms and retrieval methodologies contend simultaneously with atmospheric and surface effects on upwelling radiation. Solving this issue physically yields information about the surface geophysical and biophysical

properties and, to a large extent, the turbulent and radiative energy fluxes that link the atmosphere to the varied underlying surface. So besides the added benefit of global coverage, an improved determination of the physical processes that link water and energy reservoirs is now available to complement the historical point flux measurements gathered during field campaigns. With the growing length of these data sets, we are now able to ask questions about the climate system as a whole – its natural modes of variability and possible signatures of anthropogenic impacts. The global hydrologic cycle is at the center of many of these issues that ultimately have great practical concern in agriculture, land use practices, transportation, health and public policy. These scientific issues have been articulated by a number of intergovernmental agencies and panels [the Global Water and Energy Cycle Experiment (GEWEX) and the Climate Variability and Predictability Programme (CLIVAR), both supported under the World Climate Research Program (WCRP), the Intergovernmental Panel on Climate Change (IPCC), and national space agencies (NASA), the European Space Agency (ESA), and the Japan Aerospace Exploration Agency (JAXA)]. We briefly examine some of these questions here in order to provide context to subsequent discussion of remote sensing challenges.

- *What is magnitude of hydrologic cycle and is it changing with time?* How do fluxes of water and energy and their stores change in response to natural interannual to decadal variations, and are there signals of anthropogenic change? How, if at all, are rainfall and evaporation changing with time? While measurements of rainfall for many decades and even longer exist at specific points around the world, the variability of rainfall is poorly sampled even in densely populated countries, and conventional measurements are virtually absent over the ocean and other data-sparse regions. Evaporation measurements are even less robust in coverage, particularly outside agricultural regions. Satellite estimates, though not as mature as those for precipitation, hold significant promise here.
- *How do clouds and surface processes control mass cycling of water and associated energy exchanges?* Clouds affect surface radiative fluxes directly but, at the same time, respond to surface energy fluxes. Water and energy availability controls evaporation and transpiration in determining the surface energy balance and thermodynamic energy in the boundary layer. Atmospheric transients tap and release this energy, producing clouds that have systematic organization reflecting atmospheric dynamical scales. These processes are so complex over land that even from space not all the components are adequately observed. Land data assimilation techniques to merge observations with physical models are a fundamental strategy to help answer this question.
- *How are extreme hydrologic events related to climate variability?* Atmospheric transients that induce local hydrologic events (either droughts or floods) are organized by, and, in turn, influence, large-scale atmospheric flow features (e.g. jet streams). Anomalous atmospheric planetary wave configurations and circulation changes may arise from internal atmospheric instabilities, from SST forcing associated with the *El Nino*-Southern Oscillation (ENSO) or perhaps in response to low frequency ocean circulations (e.g. Atlantic meridional overturning). Remote sensing helps monitor the energy exchanges between the atmosphere and surface, thereby sorting out precursors to hydrologic events. Over longer periods, this information aids in developing statistical models of how extreme events are associated with disruptions or anomalies in average climatic conditions. The extent to which anthropogenic changes to the atmosphere and land surface might influence the frequency and intensity of extreme events is also an important issue.
- *What potential exists for predictability and what do we need to realize it?* If we can understand and quantify these processes, can we then determine their predictability, and can the physics in our numerical models and the forcing observations from space be used to realize this predictability to its inherent deterministic and stochastic limits? Because the capacity to store heat energy in the oceans, and, to a lesser extent, in soil, is far greater than in the atmosphere, these reservoirs of energy are thought to contain "memory" or boundary conditions that are useful beyond the theoretical two-week limit to deterministic weather forecasting. The slow evolution of SSTs or soil moisture and snowcover may, in a sense, "load the dice" such that seasonal or multiyear predictions may be statistically much more useful than climatology for many applications. Use of remote sensing to build a quantitative picture of covariability of the atmosphere, land surface, and oceans is central to this capability.

This list is clearly not exhaustive but represents the intersection of the water cycle with weather, hydrology, and the physical climate system. These questions require that in studying the water cycle we maintain focus on the associated radiative processes, on transports of energy by the atmosphere and ocean, and on energy exchanges between the atmosphere and underlying ocean, land/ocean, and cryosphere. Not only is it important to resolve and retrieve the physical attributes of processes and variables (e.g. cloud microphysical and radiative properties and soil-moisture availability) but we must also construct and maintain sufficiently long records to understand variability on interannual to decadal scales. Many of our evolving data sets are just beginning to meet this last criterion, making observational study of these questions now accessible.

A BRIEF BACKGROUND OF THE REMOTE SENSING PROBLEM

Remote sensing is a classic inverse problem in that measurements are collected, say, of upwelling radiation at different frequencies, representing the convolved effects of many geophysical signals. These measurements do not by themselves define a unique solution to the retrieval of a given variable and typically are supplemented by *a priori* constraining "guesses". However, one key concept to be noted is that remote sensing of water using different regions of the electromagnetic spectrum reveals different physical attributes of clouds, water vapor or surface water, and energy states. Visible and IR measurements from space provide key measurements for two reasons. First, it is in these wavelengths that most of the Sun's energy is intercepted by the earth system and then reradiated to space. Thus, we are measuring the direct effects of clouds and moisture on radiant energy exchanges that drive weather, climate, and the hydrologic cycle. Because these wavelengths are shorter than or nearly the same as the characteristic dimensions of cloud particles, they are strongly reflected and absorbed. Therefore, when viewing cloudy atmospheres, we are typically only seeing into the top portions of clouds – in a sense we are seeing the optical and thermal morphology of the cloud tops. But it is these very energy fluxes that help connect the global radiation budget to the earth's hydrologic cycle and heat budget. Second, because of the wavelength, visible/IR spaceborne detectors can provide very high spatial resolution for imaging and sounding – typically footprints of a few kilometers or less. Cloud fields and surface vegetation have much fine structure so that if we wish to monitor precipitating clouds, soil-moisture gradients, or snow cover in elevated terrain, these measurements provide a unique way of sampling these features globally.

In contrast, microwave remote sensing measures radiation at wavelengths of several millimeters to centimeters, which is larger than characteristic sizes of cloud droplets and even most precipitating raindrops or frozen particles. Consequently, there is a weaker response to condensate in the atmosphere that allows microwave sensors to sense upwelling radiation from much deeper in clouds and frequently from the surface below. Instead of just seeing the hydrometeors at cloud top, the entire mass of condensate in the clouds affects the upwelling radiation, either absorbing and re-emitting energy or scattering it from the field of view depending on the wavelength and size of the particles. Wavelength-dependent sensitivity to the *entire mass of hydrometeors* by upwelling radiation from the ocean or land-surface backgrounds gives radiative flux measurements that are much more closely related to bulk cloud water and ice mass and surface precipitation rate.

One might gain the impression that with these two classes of measurements – visible/IR and microwave – quantifying the hydrologic cycle is relatively straightforward. Unfortunately, there are key uncertainties that limit us. Size, shape, and density of hydrometeor populations are each important to varying degrees, as are the atmospheric thermodynamic profiles. Surface characteristics of emissivity and roughness respond to soil moisture, vegetative cover, sea roughness, and wind speed in a dramatic fashion. This provides both problems and opportunities. Some additional *a priori* information is needed about these factors if one is trying to produce retrievals of atmospheric variables. On the other hand, by viewing the surface through an undisturbed (cloud free) atmosphere, the frequency dependence, in the way these geophysical characteristics present themselves, can yield much information about soil moisture, surface temperature, emissivity, ocean surface wind stress, and other factors important to surface hydrologic processes.

CURRENT CAPABILITIES

We examine now the major components of the hydrologic cycle, beginning with precipitation and evaporation. As primary source and sink terms of the global water balance, these variables are treated in some depth. Evapotranspiration and soil moisture are also discussed in the overall context of land-surface data assimilation strategies. Clouds and water vapor are treated largely in relation to their important effects on radiative fluxes. Finally, we turn our focus on the cryosphere, noting some of the challenges in this more slowly evolving component of the global water cycle.

Precipitation

Attempts to estimate precipitation from spaceborne remote sensing have a long history. Most early studies were indirect methods that inferred rain accumulation on the basis of the presence of cloud cover alone (Woodley and Sancho, 1971; Martin and Suomi, 1972; Griffith *et al.*, 1978). The physics behind these approaches was that cold IR and bright (visible) clouds were deep and thick and so offered the most probable location of rainfall. The IR technique was investigated systematically by Arkin (1979) and Richards and Arkin (1981) in studies of tropical rainfall in the GATE (Global Atmospheric Research Program Atlantic Tropical Experiment) region. They showed that under appropriate time- and space-averaging, precipitation estimates useful for climate diagnostics studies could be achieved using geostationary IR imagery. This approach, applied to the global suite of geostationary imagers, eventually became the basis for the initial Global Precipitation Climatology Project (GPCP) on rainfall climatology. While these cloud-top temperature statistics do not measure rainfall *per se*, their statistical relationship to rainfall and their high temporal availability (as frequently as every hour) have

made them an integral component to the mix of information used to monitor global rainfall even today. The historical record of these observations stretches back to the mid-1970s, making these measurements a valuable complement to microwave data, which have poor sampling before late 1987.

The sensing of actual precipitating hydrometeors from space began with the Nimbus-5 Electrically Scanning Microwave Radiometer (ESMR), a passive microwave imaging sensor (Wilheit *et al.*, 1977). These measurements clearly demonstrated the radiometrically warm signature of raindrops (and cloud water droplets) over cold ocean backgrounds. Because the ocean surface is highly polarized, its emissivity is small and the ocean looks "cold" radiometrically. In contrast, liquid water in the atmosphere emits unpolarized microwave radiation and the liquid hydrometeors appear as "warm" regions against the cold ocean background. Over land, surface emissivity is much larger and more variable, so the liquid water emission signature is largely indistinguishable against the land background. However, in a series of papers (Wilheit *et al.*, 1982; Spencer *et al.*, 1983; Spencer *et al.*, 1989) it was shown that an alternative technique that recognizes the scattering of upwelling radiation by large raindrops and precipitating ice hydrometeors present above the freezing level can be effective in detecting rainfall.

Over the past two decades these complementary methodologies of passive microwave emission and scattering and cloud-top temperature and optical brightness have been expanded, refined, and verified for the purposes of climate and weather studies. In various forms, they presently constitute the basis for international efforts to systematically monitor rainfall from space. The most significant recent advance, however, has come with active radar remote sensing from TRMM, (Simpson *et al.*, 1988) jointly developed by the NASA and JAXA. Using a single frequency radar at 13.8 GHz, this joint US/Japanese mission has yielded unprecedented vertical structure of precipitation reflectivity and rainfall. Range-gating the energy returned from precipitating clouds allows vertical resolution of better than 250 m. Instantaneous spatial resolution of about 4 km exceeds that of conventional passive microwave sensors. For a more extensive discussion of remote sensing techniques and data set production the reader is referred to **Chapter 64, Satellite-based Estimation of Precipitation Using Microwave Sensors, Volume 2**. In what follows here, we focus more on how these methodologies have provided signals of climate variability.

Regime Dependencies in Tropical Precipitation

One of the major illuminating results from the TRMM precipitation radar (PR) has been the confirmation of differing precipitation regimes over the tropical regions. From studies combining TRMM PR reflectivities, TRMM Microwave Imager (TMI) 85-GHz-deep convective ice scattering, and Lightning Imaging Sensor (LIS) lightning flash densities (Boccippio *et al.*, 2000; Toracinta and Zipser, 2001; Petersen and Rutledge, 2001), there has emerged confirmation of distinct variations in vertical structure and intensity of convection, particularly over tropical "continental" locations. On a seasonal mean basis, distinct differences over the Amazon basin are noted with a more "maritime-like" weak reflectivity regime during the wet season and more strongly electrified, strong ice-scattering regime during the dry season (Petersen and Rutledge, 2001). There is also evidence of differing relationships between deep convective ice frequency and surface rain rate in the Amazon as compared to equatorial Africa (McCollum *et al.*, 2000). Although convection over tropical land tends to have higher vertical penetrations of a given reflectivity intensity (e.g. 30 dBZ), much higher lightning density (10 to $10^2 \times$ frequency) and more intense 85-GHzice scattering (Toracinta *et al.*, 2002), there frequently exist transitions to periods of "maritime-like" character in precipitation (e.g. low reflectivities above the freezing level, reduced lightning, and ice scattering), which seem to have variability on intraseasonal timescales. These variations contrast with precipitation systems over oceanic convergence zones, which are generally more muted in variability because boundary-layer moisture and temperature are typically nearer equilibrium with the ocean surface. Here, large accumulations of convective, available potential energy to fuel strong updrafts and produce hail or large graupel are quite rare. Even for these regimes, however, there is evidence that intraseasonal modulation of low-level moisture can serve as an agent to alter the hydrometeor characteristics within organized, deep precipitating systems.

Vertical profiles of hydrometeors (raindrops, graupel, cloud, water, and ice) are closely linked to atmospheric condensational heating profiles. One TRMM algorithm, the Goddard Profiling Algorithm, GPROF (Kummerow *et al.*, 2001), derives hydrometeor profiles by statistically associating TMI-measured brightness temperatures with *a priori* information from cloud model predictions of hydrometeors and associated modeled brightness temperatures. These estimates also take into account cloud water and ice and thus complement the direct measurements of precipitation from PR. Along with maturing TRMM algorithms for diabatic heating, (e.g. Tao *et al.*, 2001) these global, tropical retrievals will be useful in studies of how atmospheric heating affects atmospheric dynamics and how precipitation is organized on regional and global scales. Follow-on missions to measure precipitation and vertical profile information at higher latitudes are expected in the future.

Climate Variations (ENSO Signals)

It is well documented that El Nino/Southern Oscillation (ENSO) events, with marked SST changes over the tropical

oceans, produce significant regional changes in precipitation, water vapor, and radiative fluxes in the tropics (e.g. Ropelewski and Halpert, 1987; Sun and Trenberth, 1998; Curtis and Adler, 2000). Berg *et al.* (2002) have recently documented the distinct differences between precipitation structure over the eastern and western Pacific ITCZ and noted how various satellite precipitation algorithms may respond quite differently to ENSO modulations of these precipitation regimes. However, quantifying the associated *net integrated changes* to water and heat balance over the entire tropical oceanic or land sectors remains an observational challenge.

The GPCP data set (Huffman *et al.*, 1997; Adler *et al.*, 2003) shows consistent increases (decreases) in precipitation associated with ENSO warm events when integrated over tropical ocean (land) regions. The reverse tends to occur during ENSO cold SST events. These variations are typically of order 0.30 mm per day, less than 10% of climatological mean values. Over the tropics as a whole, these perturbations in rainfall essentially vanish.

Why should we be interested in these large regional averages? ENSO events offer a natural experiment in terms of perturbations to the tropical heat balance. Changes in fluxes averaged over this system or its land and ocean components thus represent restoring mechanisms and involve many of the same feedback processes relevant to climate change – changes in evaporation, cloud and water vapor forcing, and energy transport. It is important to remember that since ENSO is a natural climate variation the internal responses or feedbacks are different from those related to anthropogenic forcing (Bony *et al.*, 1997). Nevertheless, the ENSO 2–7 year periodicity offers an accessible "experiment" involving changes in air–sea interaction processes, precipitation, radiative fluxes, and surface temperature – changes in water and energy fluxes that are highly coupled (Sun and Trenberth, 1998). These changes constitute a necessary test that climate models must successfully navigate.

Precipitation Accuracies

Estimating error in precipitation retrievals is sensitive to the density of rain gauges, the character of the precipitation, the topography, and spatial scale of the satellite and gauge averaging. Adler *et al.* (2003) estimate, overall, the errors for the monthly GPCP product to range between 10% and 30% at rain amounts above 100 mm per month (3 mm per day). This percentage error is for a 2.5° latitude × 2.5° longitude square, and time- and space-averaging will significantly reduce the random error. Over land, comparison to independent gauge data over the Oklahoma Mesonet indicates that GPCP monthly amounts correlate to the observations at 0.74 (0.94) before (after) global adjustments of the satellite estimates to rain gauges. Globally, the GPCP data set has similar magnitudes and zonal mean distributions as the historical Jaeger (1976) and Legates and Willmott (1990) values. In the tropics, values are systematically smaller than Legates and Willmott values due to reliance by the latter on Pacific atoll data whose magnitude is still in question.

For interannual and longer signals, error inferences are complicated by the stability of the satellite sensors and subtleties in retrieval physics assumptions and how they respond to precipitation physics. In a paper that documented the uncertainties in detecting interannual variations of precipitation, Robertson *et al.* (2001) critically examined six satellite-derived precipitation data sets over the tropical oceans. When averaged over the domain 30° N/S, only two of these, the Wentz and Spencer (1998) Special Sensor Microwave Imager (SSM/I) algorithm and a deep convective index (DCI) from Microwave Sounding Unit Channel 2 (MSU2) showed strong correlations between each other and with tropically averaged SST variations. The implication of these results was that for the purposes of climate diagnostics work and model validation, the precipitation data sets commonly used in the research community have significant observational uncertainties. The expectation was that measurements from TRMM PR and TMI would reveal the physical and quantitative reliability of these algorithms.

Interestingly enough, a definitive statement regarding interannual variability of oceanic precipitation has not yet been provided by TRMM, even though it sampled perhaps the largest warm SST event in decades. Figure 1 shows differences between the PR and various passive algorithms from TMI in both the sense and magnitude of interannual precipitation variations averaged over the tropical oceans.

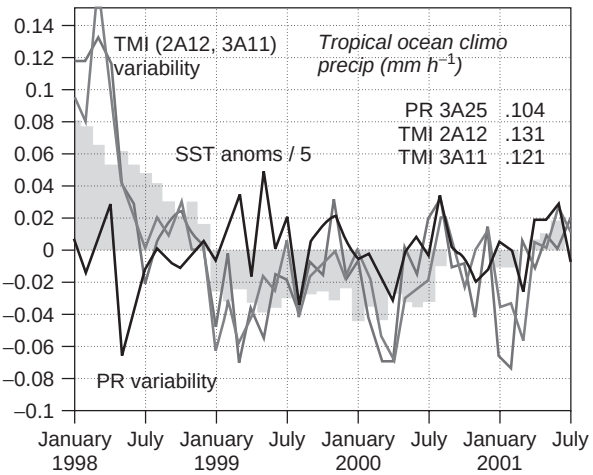


Figure 1 Time series of precipitation anomalies averaged over the tropical oceans from the TRMM Version 5 PR 2A25 and TMI passive microwave algorithms 2A12 and 3A11. Time series have been normalized by their respective climatological values to yield fractional variability. SST anomalies (°C) over the tropical oceans scaled by 1/5 are shaded. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The two TRMM TMI algorithms, in agreement with the previous DCI and Wentz/Spencer results, indicate a positive correlation with SST. The TRMM PR time series is notably different, exhibiting correlations with the TMI algorithms of only 0.12. At first blush, one might regard the TRMM PR result as definitive because the reflectivity measurements respond directly to precipitation size hydrometeors. However, the TRMM PR operates with a single frequency (13.8 GHz), an attribute necessitating significant microphysical assumptions regarding drop-size distributions and parameterizations of reflectivity/attenuation/rainfall relationships. In a recent paper, Robertson *et al.* (2003) found evidence that it is the uncertainties in these microphysical assumptions needed for the primary TRMM algorithm (2A25) that are problematic in trying to isolate interannual variations integrated over the tropical oceans. They found that a direct rainfall attenuation measurement made from the return signal of the TRMM PR (the so-called Surface Reference Technique) in fact shows strong similarities to those interannual variations detected from passive microwave estimates from TMI. This is consistent with the agreement found by Robertson *et al.* (2001) between the MSU precipitating ice and SSM/I depictions of rainfall variations. It is also consistent with the results of Adler *et al.* (2003). New versions of the TRMM algorithms are encouraging for the reconciliation of active and passive precipitation estimates for climate work over the tropical oceans.

We note in passing that a number of satellite precipitation estimates (GPCP, TRMM, TMI, and PR) seem to agree somewhat better over land areas. Whether this has to do with the influence of rain gauges on the GPCP product or the dominance of rainfall by stronger, deeper rain systems more easily and reliably detected by precipitating ice-scattering algorithms used over land has yet to be determined. Finally, in understanding the differences between passive and active results, it is important to keep in mind that both approaches measure only the physical attributes of hydrometeors – neither is a direct measurement of rain flux at the earth's surface.

Oceanic Turbulent Heat Fluxes

The counterpart of precipitation in the hydrologic cycle is, of course, surface latent heat flux or evaporation. In this section we consider processes over the global oceans. Because it is impractical with *in situ* eddy correlation techniques to measure latent (and sensible) heat fluxes over large regions, the complexities of this process are typically approached from bulk aerodynamic methods to which conventional data can be applied. Efforts to infer surface turbulent fluxes from space have also followed this tack. The availability of global observations of wind stress with the Seasat and Nimbus 7 (Scanning Microchannel Microwave Radiometer (SMMR)) instruments inspired the development of techniques over oceans using bulk formulae of the form

$$E = \rho C_{\mathrm{E}} U (q_{\mathrm{s}} - q) \tag{1}$$

where E is evaporation, U is surface wind speed at 10m, q is 2m specific humidity and q_{s} is the saturation q of the ocean surface SST, ρ is air density, and C_{E} is the transfer coefficient for the reference height. Estimates of SST from *in situ* and satellite (e.g. Reynolds and Smith, 1994) provide a means of obtaining q_{s}. For many applications, C_{E} for a neutral atmosphere over the oceans is a reasonable first estimate. Situations such as cold, dry air outbreaks over warm middle latitude waters are a typical exception. Liu *et al.* (1979) provided a methodology for estimating variations in exchange coefficients. The other remaining unknown, near-surface humidity, q, was estimated by Liu (1984) and Liu and Niiler (1984) on the basis of its statistical relationship with column-integrated water vapor, W, which is also retrieved from microwave satellites.

This basic approach, with variations and improvements, has formed the cornerstone of virtually all succeeding efforts to develop latent heat-flux data sets. Validation and improvement of bulk flux models are receiving considerable emphasis in recent years (Chou, 1993; Fairall *et al.*, 1996b, 2003; Clayson *et al.*, 1996; Zeng *et al.*, 1998; Brunke *et al.*, 2003). As in the case of precipitation, the growing archive of SSM/I data has provided the extensive data set needed to produce global evaporation estimates. Several near-global data sets of retrieved evaporation and other air–sea flux parameters have been derived. The Hamburg Ocean–Atmosphere Parameters and Fluxes from Satellite Data, HOAPS dataset, (Schulz *et al.*, 1997; Grassl *et al.*, 2000) provides gridded fluxes on various scales (0.5 to 2.5 degrees, daily to monthly averaging). In attempting to improve upon Liu's methodology, HOAPS uses estimates of the integrated water vapor content of the lowest 500 m of the atmosphere, W_{b} (Schulz *et al.*, 1993; Schlüssel, 1996) and relates this statistically to specific humidity near the oceans surface. The atmospheric stability to compute the transfer coefficient is estimated with a fairly simple technique. Following Smith (1988) it is assumed that the atmosphere has a relative humidity (RH) of 80% at any time. Thus, the near-surface air temperature can be computed from the measured q. The errors in the transfer coefficient that occur if this assumption is wrong are small at high wind speeds and unstable conditions ($\sim$2%) but can approach 50% at low wind speeds and strong stable conditions (Schulz *et al.*, 1997).

An alternative but similar methodology is the Goddard Satellite-based Surface Turbulent Fluxes data set, Version 2, GSSTF2, (Chou *et al.*, 2003). These flux retrievals use Wentz (1997) Version-4 SSM/I surface (10-m) wind speeds and total precipitable water, and SSM/I antenna temperatures, as well as the SST, 2-m air temperature and sea-level

pressure of the National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCEP/NCAR) reanalysis. The methods for deriving the GSSTF2 daily turbulent fluxes and input parameters essentially follow those of Chou *et al.* (1997). A fairly sophisticated bulk flux model is used. The methodology for retrieving daily q at 10 m height, q10 m, (Chou *et al.*, 1995, 1997) also uses the Schulz *et al.* (1993) method to estimate W_b but applies a vertical empirical orthogonal function methodology (Wagner *et al.*, 1990) to aid in inferring q10 m. Daily W_b is derived from SSM/I following Schulz *et al.* (1993), while daily precipitable water is taken from that of Wentz (1997) for the GSSTF2.

Turbulent Flux Accuracies

There is no doubt that satellite techniques are able to produce reasonable flux fields on a climatological basis. Annual climatological evaporation estimates for the Chou *et al.* (2003), da Silva *et al.* (1994) and NCEP/NCAR Reanalysis each show similar regional patterns of broad, elevated latent heat flux over subtropical oceans and maxima off the east coasts of continents in winter. Nevertheless, each of these estimates has significant uncertainties. For example, Gleckler and Weare (1997) estimated that monthly mean latent heat fluxes had uncertainties of approximately 35 $\mathrm{W m^{-2}}$. Estimates are that reanalysis evaporation also has a similar uncertainty. Aside from random error, regional biases are of particular concern. It has also been noted that in subtropical regions the satellite evaporation estimates are significantly larger than either the reanalysis or da Silva values. This type of bias was originally pointed out by Esbensen *et al.* (1993) and linked to biases in estimates of q inferred from column-integrated water. Varying scale heights of water depending on large-scale vertical motion can significantly affect near-surface q. For example, in subtropical regions sinking motion produces a drier-than-average atmosphere; however, in the planetary boundary layer, turbulent mixing still moistens the overlying air. Thus, estimates of q inferred from the column-integrated water vapor will systematically indicate a smaller q value and hence a larger bulk estimate of evaporation.

Wind speed accuracy is also a question. Mears *et al.* (2001) and Meissner *et al.* (2001) have shown generally very good correspondence with buoy winds and suggest rms errors of order 1.0 $\mathrm{m\,s^{-1}}$ and overall biases of order several tenths $\mathrm{m\,s^{-1}}$. There are also known wind speed biases along coastal areas where ocean currents and wave geometry relative to spacecraft orbital track produce systematic biases also of order 1.0 $\mathrm{m\,s^{-1}}$ or more.

What are the implications of these uncertainties for monitoring evaporation changes on interannual and longer timescales? Some idea of current capabilities is seen in Figure 2 in which tropical ocean evaporation estimates from two different SSM/I-based estimates and the NCEP/NCAR reanalysis are presented. Long-term trends in each of the three estimates are seen, but the two SSM/I estimates are significantly larger. On interannual timescales, the SSM/I estimates clearly show the influence of ENSO warm and cold events with time-lagged responses to SST, with reanalysis estimates being much weaker and noisy. Part of the large evaporation changes in the satellite estimates can be related to the effects of the q biases. But it is also true that the NCEP reanalysis, like any model, is driven strongly by model physical parameterizations in the tropical regions.

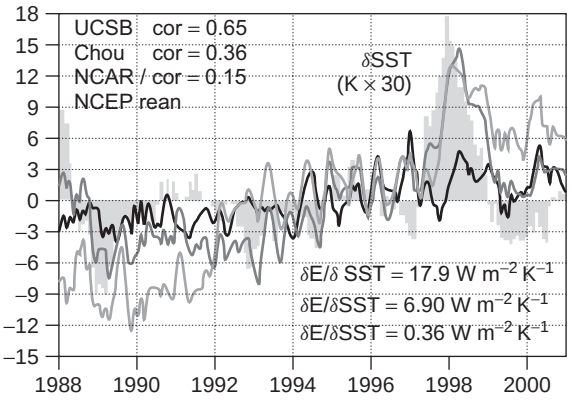


Figure 2 Bulk aerodynamic evaporation anomalies (30 °N/S average) from two different SSM/I passive microwave algorithms [University of CA Santa Barbara (Jones *et al.,* 1999) and GSSTF (Chou *et al.,* 2003)] show large flux trends and δE/δSST sensitivities compared to NCAR/NCEP Reanalysis. SST anomalies (°C) over the tropical oceans scaled by 30 are shaded. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Global coverage of ocean evaporative fluxes is steadily improving with clearer insight into bulk formula approaches and detection and understanding of systematic biases in satellite input data. A collaborative project, SEAFLUX, (Curry *et al.*, 2004) is serving as a forum for improving algorithms and recommending best practices for improving remotely sensed ocean fluxes. This includes moisture, heat, and momentum. It is expected that continued efforts will fulfill the need for high-resolution, accurate surface turbulent fluxes (heat, water vapor, and momentum) over the global ocean that has been articulated by numerous advocacy groups within the global climate community, including the WCRP Joint Steering Committee/Scientific Committee on Ocean Research (SCOR), Joint Working Group on Air/Sea Fluxes (JWGASF), the GEWEX Radiation Panel, and the CLIVAR Science Steering Group.

Land-surface Hydrology

Water and energy balance obviously differs fundamentally over land compared to ocean regions. Because mixing and dynamical transport can facilitate heat storage from solar

radiation in the ocean (Levitus *et al.*, 2000; Sun, 2003), timescales of near-surface temperature change can extend past seasons to interannual and decadal scales. Over land, the capacity for subsurface energy storage is diminished and so radiative, latent, and sensible heat fluxes can have strong variability on shorter timescales. The diurnal cycle in surface fluxes is a prominent example of this, as are frequent subseasonal anomalies associated with regional droughts or wet periods. The availability of near-surface water in soil and vegetation is of critical importance to the partitioning of energy flux among these three processes. In middle and higher latitudes where solar forcing is weak, soil-moisture anomalies can extend to seasonal and even interannual or decadal timescales (Delworth and Manabe, 1988). Further complicating the water balance and energy exchange, in comparison to oceans, is the extreme heterogeneity in vegetative cover, soil characteristics, topography, and drainage characteristics. Surface emissivities are constantly changing owing to soil-moisture changes. The diurnal growth and collapse of planetary boundary-layer structure and sharp vertical gradients in thermodynamic quantities and wind shear also present extreme challenges for passive remote sensing, which is not adept at capturing vertical structure. All these distinguishing attributes of the water cycle over land make remote sensing of hydrologic processes more problematic.

Given these challenges, how do we best use the remotely sensed data to determine the water and energy balance over the varied terrestrial regions of the globe? Beginning with GEWEX in the early 1990s, a scientific consensus emerged for evolving hydrologic models into systems that can assimilate, or be constrained by, observations from space in such away that *a priori* knowledge on soils characteristics, vegetation, drainage, and so on are optimally combined with observed radiative fluxes and atmospheric retrievals of near-surface meteorology. This assimilation process allows the information we gain from space-based observations to build physical consistency into inferences of surface water, heat, and radiative fluxes.

Added degrees of freedom due to inclusion of such land-surface schemes require the specification of additional parameters within the model system such as vegetative resistances, green vegetation fraction, leaf area index, soil physical and hydraulic characteristics, stream flow, runoff, and the vertical distribution of soil moisture, all of which are difficult to measure over large regions. This issue is even more problematic in areas where the land-surface characteristics are extremely heterogeneous. Consequently, some attempts to use remotely sensed data have emphasized simplified treatments of surface energy balance. For example, McNider *et al.* (1994) developed a procedure based on the work of Wetzel (1984) and Carlson (1986), which takes a practical approach toward assimilating space-based observations of land-surface temperature (LST) by recognizing that adjustments in this parameter must be consistent with other components of the surface energy budget. The technique requires the use of Geostationary Operational Environmental Satellite (GOES) derived IR LST *tendencies* and GOES-derived net solar radiation over the time period of interest. It is based upon adjusting the model's bulk moisture availability so that the model's *rate of change* of LST agrees more closely with that observed from the satellite. Therefore, the simulated latent heat flux, which is highly correlated with surface moisture availability during this time period, is adjusted based upon differences between the modeled and satellite-derived LST tendencies.

Lapenta *et al.* (1999) tested the technique over the southeastern US where the vegetation is highly variable. Twenty-five case days were initialized at 12 UTC from the NCEP Early Eta and run through 12 h.A 25-km nested grid centered over the southeastern US was employed, and satellite data were assimilated between 1400 and 1700 UTC (forecast time of 2–5 h). The remainder of the model run was executed in a free forecast mode. Verification statistics for the 2-m air temperature were computed and are displayed in Figure 3. The control run (marked SLAB) was performed with no assimilation and a second run (marked ASSIM) assimilated the satellite data. The results demonstrate that short-term simulations of near-surface temperature were generally improved on a daily basis. Another nonassimilation run was made with the Oregon

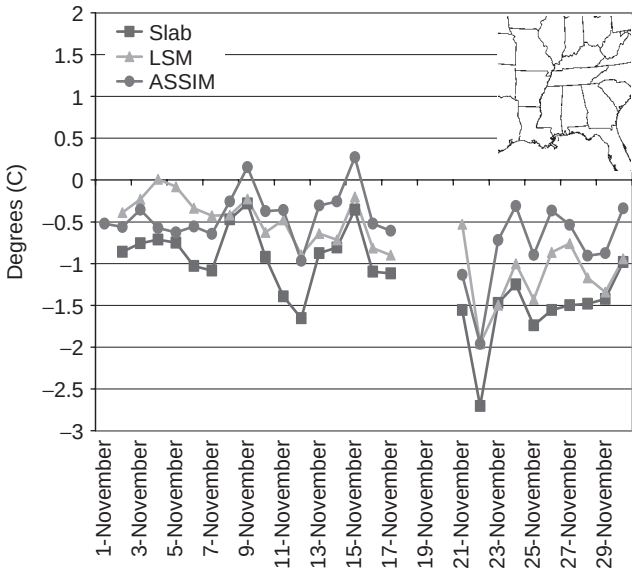


Figure 3 Mean absolute error (°C) of simulated 2-m air temperature for the control (SLAB), assimilation (ASSIM), and land-surface model (LSM) runs for November 1998. Hourly error statistics were averaged over the 12-h forecast period. The geographic coverage of the model domain is shown in the inset. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

State University Land-surface Model, LSM, (marked LSM in Figure 3), to determine how the assimilation technique performs relative to a more sophisticated approach. Early in the month, the LSM performs better than control and assimilation runs. However, this pattern is reversed from the 9th to the end of the month. Examination of satellite data revealed that the poor performance of the assimilation run occurred during a cloudy period. The later half of the month when the assimilation had the most positive impact was mostly cloud free.

In addressing the more complex task of driving fully functional hydrologic models with space observations, a number of internationally coordinated efforts have arisen. For example, the International Satellite Land Surface Climatology Project (ISLSCP) has served as a pathfinder effort to determine just what types of surface and near-surface satellite measurements are relevant to climate and global change studies. It facilitated improvements in algorithms for the interpretation of satellite measurements of land-surface features and developed methods to validate area-averaged quantities derived from satellite measurements for climate simulation models. The First ISLSCP Field Experiment (FIFE) and Boreal Ecosystem–Atmosphere Study (BOREAS) field programs were middle- and high-latitude field studies for addressing these problems.

Under the ISLSCP Initiative I, the first global land cover, hydrometeorology, radiation, and soils data sets regridded to a common 1×1-degree format for 1987–1988 were produced. The Global Soil Wetness Project-1, GSWP-1, (Dirmeyer *et al.*, 1999) has used these data sets in conjunction with an international suite of land-surface models to produce state-of-the-art global data sets of land-surface fluxes, surface state variables, and related hydrologic quantities. Key components of the forcing data have been derived from satellite. Among these are the surface radiation budget components from the NASA/GEWEX Surface Radiation Budget Project. In this effort, thermodynamic soundings and satellite measurements of surface properties have been combined with the International Satellite Cloud Climatology Project (ISCCP) cloud data to constrain radiative models and yield surface radiative fluxes. Precipitation forcing to drive soil moisture has been taken from the GPCP merged gauge/satellite product with temporal disaggregation estimated from 6h resolution rainfall from the European Centre for Medium-Range Weather Forecasts (ECMWF) analysis. These products are being extended with ISLSCP-2 forcing, which now covers the period from late 1987 through 1995.

In support of first GEWEX Continental-Scale Experiment, the North American Land Data Assimilation System (NLDAS) (; see Mitchell *et al.*, 1999) has provided a useful regional focus for extending GSWP capabilities to finer scales in an operational setting and provided impetus for development of ground-based radar estimates of precipitation as well. A Global Land Data Assimilation System (GLDAS) that uses various new satellite- and ground-based observation systems within a land data assimilation framework is being developed by NASA and various collaborators (Houser *et al.*, 2001; Rodell *et al.*, 2004) and provides an extension of GSWP products to near real time. This integrated system assimilates land-surface and diverse satellite observations (including precipitation, solar radiation, snow cover, surface temperature, and soil moisture) globally at 1/4 degree resolution in near real time. This real-time aspect is crucial if the hydrologic state variables and fluxes are to be used for global weather forecast models and for initializing ensemble model experiments for seasonal prediction. Figure 4 depicts representative fields of evapotranspiration and soil-moisture content produced from this system.

Accuracy of Hydrologic Quantities

Determining the accuracy and sources of errors in surface water and energy balance components is at present very difficult. There are uncertainties in the model physics, the accuracy of forcing data, the specified fields (e.g. root zone properties, leaf area index, soil properties, vegetation type etc.), and in verifying field data. In fact, the whole question of how to compare point measurements to retrievals of fluxes and surface state supposedly representative of model grid points or entire watersheds is far from settled. Some perspectives on the accuracy and physical realism of satellite-constrained surface retrievals are emerging nonetheless.

The GSWP-1 intercomparison effort (Dirmeyer *et al.*, 1999) showed significant spread among the participating land-surface schemes in terms of their partitioning of surface energy between latent and sensible heat flux, and of water between runoff and evapotranspiration. When averaged globally, the ensemble mean Bowen ratio was 0.95 with a standard deviation of 0.18. The runoff mean of 228.6 mm underestimated basin-scale measurements, possibly due to the GSWP specification of the treatment of convective precipitation. The standard deviation of 39.6 mm suggests that subgrid-scale variability in infiltration due to heterogeneity in soil properties and the uncertainties in distribution of rainfall within a grid box have significant impacts on the simulation of runoff. Most notably, the comparison of the LSM simulations against point data showed poor and inconsistent performance, raising the question of whether the model physics is adequate or simultaneously appropriate at both point and grid scales.

Robock *et al.* (2003) studied the performance of four state-of-the-art land-surface models participating in the NLDAS Project (Mitchell *et al.*, 2003). The models were forced by observed precipitation and solar insolation for the periods May–September of 1998 and 1999 and validated with observations taken at Oklahoma Mesonet and

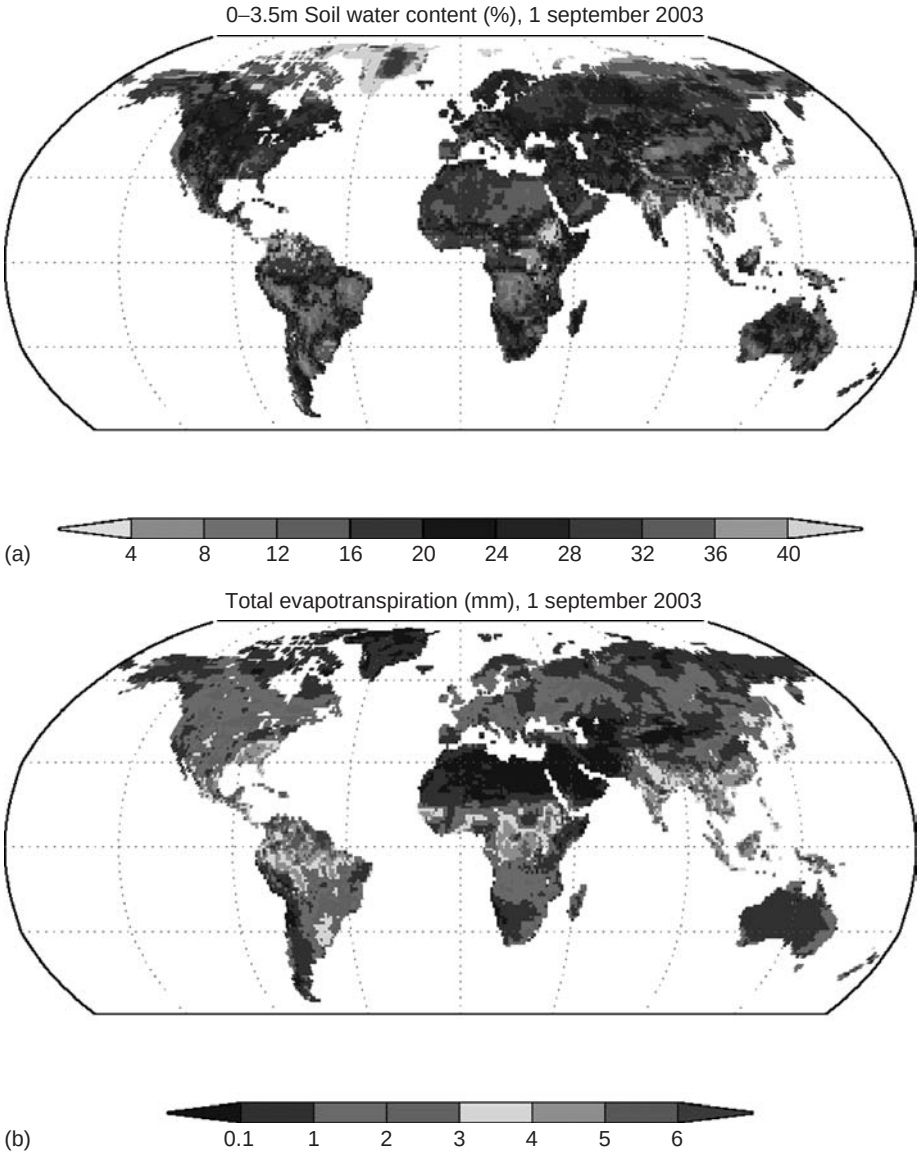


Figure 4 Total evapotranspiration (a) and soil water content in the top-most 3.5 m (b) for September 2003 produced by the NASA global land data assimilation system. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the Department of Energy Atmospheric Radiation Measurement (ARM)/Cloud and Radiation Testbed (CART) site in the southern Great Plains region. While the NLDAS models captured broad features of soil-moisture variations in the top 40 cm and detected interannual differences, intermodel variability was still substantial. Monthly departures of turbulent fluxes were typically several tens of $W\,m^{-2}$ from the sensible and latent heat-flux observations of 35 and 70 $W\,m^{-2}$, respectively. These systematic biases also varied with season and between years.

These early comparisons suggest that physically based land-surface models do respond to atmospheric forcing in realistic fashion, though the quantitative accuracy remains to be increased across regional to global scales. The transition from statistical modeling of hydrologic processes to a physically based methodology means that, while substantial work remains, the linking of surface states to the atmosphere and the derivation of associated energy fluxes is rapidly being put on a sound physical basis needed for coupled global prediction.

Clouds and Water Vapor

Traditional studies of the hydrologic cycle have tended to emphasize the mass balance of water yet, as noted in the Introduction, clouds and water vapor are tightly coupled

to the surface temperature and the overall heat balance of the planet. Monitoring the radiative aspects of this coupling and the effect of clouds and water vapor on TOA fluxes are particularly amenable to remote sensing by satellite.

A detailed compilation of cloud fractional cover and morphology in both the visible and IR regions has been sustained under the ISCCP since 1983 (Rossow and Schiffer, 1991; 1999). Available at 280 km resolution, these cloud statistics derived from combined geostationary and polar-orbiting satellites have provided diurnally resolved, monthly mean statistics on cloud type, frequency, cloud-top temperature and pressure, optical depth, cloud water content, and other physical attributes. Because ISCCP cloud properties are derived from a downward looking perspective, this data set is complementary to historical ground-based climatologies (e.g. Warren and Hahn, 1986; Warren *et al.*, 1988). These data have been instrumental in evaluating and improving the representation of clouds in global models (Yu *et al.*, 1996; Webb *et al.*, 2001). The ISCCP clouds have also served as input to a global data set of radiative flux retrievals, ISCCP-FD, (Zhang *et al.*, 1995, 2004) and to the GEWEX Surface Radiation Budget products (Gupta, *et al.*, 1999).

Subsequent to the start of ISCCP, more sensitive methods for detecting upper-tropospheric clouds have been demonstrated using IR channels on atmospheric sounding instruments (Susskind *et al.*, 1987). The IR channels from 13–15 μm, which are operational on all polar-orbiting National Oceanic and Atmospheric Administration (NOAA) satellites provide a means of detecting optically thin cloud that the visible and IR window channels used by ISCCP may misinterpret. These measurements, though unable to resolve the diurnal cycle since they are made from polar-orbiting platforms, provide an independent and complementary set of data regarding cloud frequency (Wylie and Menzel, 1999; Stubenrauch *et al.*, 1999). With the recent availability of data from the Moderate Resolution Imaging Spectroradiometer (MODIS) on Terra and Aqua (Platnick *et al.*, 2003), significant improvements to cloud detection and retrieval of microphysical retrievals are possible. MODIS is a 36-channel whiskbroom scanning radiometer with bands distributed between 0.415 and 14.235 μm and nadir spatial resolutions of 250 m to 1 km. The products consist of a cloud mask for detection of clear skies, cloud-top properties (temperature, pressure, effective emissivity), cloud thermodynamic phase, and cloud optical thickness and microphysical properties (effective radius, water path).

Microwave remote sensing offers yet another measurement of cloud properties (e.g. Ferraro *et al.*, 1996; Wentz, 1997). Over oceans, these frequencies can retrieve column-integrated cloud water, which, with visible and IR methods, can only be done reliably for thin clouds. Aside from snow and graupel, ice clouds are largely transparent to most microwave frequencies flying on current sensors. Imaging and sounding at submillimeter wavelengths are needed. The cloud radar on CloudSat (Stephens *et al.*, 2002) operating at 94 GHz will offer an opportunity to directly determine cloud vertical structure, a measurement that is largely lacking today except for limited nadir views from the lidar flying on ICESat.

Atmospheric water vapor obviously plays a direct role in the global water cycle since it is maintained by surface evapotranspiration and provides a source for precipitation as air parcels are lifted and cooled. However, far from being a passive constituent, atmospheric moisture is intertwined with dynamics and radiation. Convective available potential energy depends crucially on the amount of water vapor in the planetary boundary layer and has a strong control on the occurrence and intensity of atmospheric convection. Because of its IR opacity, water vapor modulates atmospheric long-wave radiation and enables a far warmer surface temperature than would be possible if Earth were a dry planet.

Despite considerable focus on how water vapor would change radiative fluxes and climate in a CO_2-enriched world, we are still constructing an observational framework for understanding the full effects. Long-term satellite retrievals of water vapor over the globe date back to the NOAA Operational Polar-Orbiting Satellite Sounders, beginning with TIROS-N in 1978. Geostationary sensors (e.g. GOES and Meteosat) have provided complementary high time resolution of this quantity beginning in the early 1980s. These IR measurements have provided insight on the variations of atmospheric moisture on regional to global scales unavailable from widely spaced radiosonde reports. IR retrievals cannot however be made in cloudy regions, so sampling bias is a factor in using these data. Beginning with SSM/I microwave imager data in late 1987, the ability to retrieve column-integrated water vapor over oceans under most cloudy conditions has been a critical supplementary source of water vapor data, especially in providing information of the bulk of water vapor in the lower troposphere where IR methods are most suspect. The NASA/GEWEX Water Vapor Project, NVap, (Randel *et al.*, 1996; Vonder Haar *et al.*, 2003) has been instrumental in merging these varied satellite measurements, along with radiosonde data, to provide global gridded time series of water vapor needed to study atmospheric hydrologic processes.

While it is generally agreed that water vapor in the lower troposphere varies approximately at constant RH in response to surface temperatures, there exists continuing debate as to whether this is true in the mid- to upper troposphere where the radiative effects of water vapor are strongest. The importance of humidity variations in modulating OLR over the dry subtropical "oceanic deserts" has been pointed out by Spencer and Braswell (1997). Using RH estimates from the HIRS-2 6.7 μm channel, Soden *et al.* (2002) found a roughly constant RH response

to global atmospheric cooling produced in the two years following the eruption of Mount Pinatubo. Stratospheric aerosol from this event reduced solar insolation at the surface, surface temperature, and atmospheric convection intensity. Bates *et al.* (2001) have used a longer time series of HIRS data to show that interannual variations of upper-tropospheric humidity (UTH) are strongly modulated by the dynamical flow changes during ENSO events and that the relationship between UTH and SSTs involves both seasonal and interannual timescales. On the other hand, findings by Minschwaner and Dessler (2004), using Microwave Limb Sounder (MLS) data, have suggested that UTH variations with SST over the tropics are significant, but with water vapor changing at a rate somewhat smaller than that of constant RH.

Accuracy of Cloud and Water Vapor Retrievals

In part, our current ability to determine many attributes of clouds stems from the need to account for their effects on retrieving temperature soundings whose accuracy is crucial for numerical weather prediction. Inversion techniques have followed the philosophy that accounting for the effects of clouds across channels with varying sensitivity to moisture and temperature yields more physically based, accurate retrievals (Susskind *et al.*, 1987). Limiting temperature retrievals to purely clear scenes would be much less productive. However, like other hydrologic quantities, clouds vary greatly over a broad range of scales, diurnal to interannual, and from small fair weather cumulus to extensive multilayered clouds in baroclinic zones, thus making validation efforts exceedingly difficult. Virtually all clouds have significant spatial variability below the resolution of the High-resolution Infrared Sounder (HIRS) and even of the new Atmospheric Infrared Sounder (AIRS) instrument flying on Aqua. Geostationary visible and IR techniques used by ISCCP have higher spatial resolution, but at the expense of spectral diversity that is needed to accurately determine cloud microphysical attributes.

Since no single current strategy is optimum, our assessment of cloud cover, height, and optical properties has significant sensor dependence. Cloud overlap issues from satellite are a significant uncertainty, particularly for maritime stratocumulus. Weare (2000) has estimated that the magnitudes of annual-mean low cloud amount in the ISCCP and Hahn *et al.* (1994) climatologies differ by up to about 0.4 for a number of locations. Surface-based observations have the inverse problem that satellites have, recording low clouds well but missing overlapping middle and high clouds. Surface cloud observations are scarce over oceanic regions and cannot adequately monitor interannual variability. High clouds are more readily measured by satellite, but the pervasiveness of optically thin cirrus is a challenge. Stubenrauch *et al.* (1999) have shown that nighttime thin cirrus detection with the HIRS multispectral information is substantially improved over infrared window methods of ISCCP.

One of the principal challenges in retrieving water vapor concerns the small-scale height (several km) that results from the strong combined effects the atmospheric lapse rate and the Clausius-Clapeyron constraint on saturation vapor pressure. Since passive remote sensing has very poor vertical resolution, accuracies in RH retrievals at specific pressure levels can often be 30–50% for operational soundings. The inability of IR techniques to deal effectively with cloud contamination is another source of error. For climate studies, a more tenable approach has been to interpret water vapor sounding channel brightness temperatures (either the 6.7 μm or 183.3 GHZ regions) as proportional to log(RH) weighted over the effective width of the channel response function; typically, pressure layers several hundred hPa thick (Soden and Bretherton, 1996; Spencer and Braswell, 1997). On the basis of the work of Escoffier *et al.* (2001), one can estimate biases among different satellite sensors to be of order 20 to 30%. These disparities are traceable in large part to processing differences. Despite these biases space/time correlations for 5-day mean RH estimates between sensors are quite high (order of 0.80 to 0.90) indicating that the variability in humidity retrieved by this approach is quite accurate for climate variability studies.

Cryospheric Issues

Our examination of remote sensing and the hydrologic cycle has focused so far on tropical and middle latitude regions. Many of the greatest challenges lie in fact with detecting ice mass balance trends of the polar regions, changes in sea-ice extent, seasonal snowcover amount, and trends in alpine glaciers. Issues of climate change have strong intersections here since climate models tend to project largest CO_2-related warmings to occur in the cryosphere due to ice-albedo feedback. This process, in which melting ice from a CO_2-warmed planet lowers albedos in the cold regions and results in even more solar absorption, is an important yet poorly quantified feedback. Sea-level rise with melting in polar regions is one possible, yet still highly speculative, outcome of anthropogenic effects. In general, remote sensing uncertainties over ice surfaces are some of the most difficult challenges.

Perhaps the most straightforward measurements in the cryosphere are simple ice-extent estimates. Both visible band data from NOAA polar-orbiting satellites and microwave measurements from SMMR and SSM/I instruments have been used to track changes in snowcover over the past 20+ years (Armstrong and Brodzik, 2001). Interannual variations have been found to be as large as long-term trends. The latter show a decrease in snow extent of approximately 0.2% per year, which is consistent with recent surface warming. Determining snow water equivalent is

much more difficult since ice age strongly affects this variable. New estimates of this quantity from the Advanced Multichannel Scanning Radiometer for the Earth Observing System (AMSR-E) on Aqua are expected to be greatly improved over historical estimates from SMMR and SSM/I.

Sea ice is important because it regulates energy exchanges and ocean salinity in the polar oceans. The extent of cracks or leads in the sea ice is important because these are the only regions where cold, dry polar air can gain access to the underlying seawater. Tremendous latent and sensible heat fluxes here affect the production of atmospheric clouds and affect precipitation. Charting sea-ice variations has led to the present understanding that, whereas Arctic sea ice has been decreasing over the past three decades, Antarctic ice has increased after previous decreases (Zwally *et al.*, 2002). Trends estimated from these data suggest a net decrease in Arctic ice extent of about 2.9% per decade while Antarctic ice extent increased by 1.3% per decade. Again, the SMMR and SSM/I estimates of sea-ice coverage have been crucial to making these estimates. A variety of remote sensing instruments have been used successfully to map sea-ice conditions. However, because of frequent cloud cover in the polar regions and the fact that the sun remains below the horizon for continuous periods in winter, microwave sensors are the most commonly used instruments for ice-cover mapping.

Global sea level is currently rising, mainly because of ocean thermal expansion and glacier melt, both thought to result from recent increases in global surface temperature. Will ice sheet melting contribute to this significantly in the future? By far, most of ice water is contained in the Antarctic and Greenland ice sheets. Current estimates, however, indicate that mass balance for these ice sheets is in approximate equilibrium and therefore they are not currently affecting sea level to any significant extent. The NASA ICESat (Ice, Cloud, and land Elevation Satellite) launched in early 2003 is using a laser altimeter to precisely map global ice sheet topography. Repeated flights of this type of sensor in the decades to come will enable accurate measurements of how these huge stores of freshwater are changing.

SYNTHESIS AND ASSESSMENT

The inventory of remote sensing accomplishments and data set generation over recent years represents an impressive advance in observational capability and understanding of water on this planet. We now possess global records of precipitation, evaporation, clouds, and radiative fluxes extending over more than two decades. These time series have delineated regional ENSO responses to SST fluctuations and ocean heat content. Methodologies have been developed to assimilate surface soil and vegetative state, skin temperature, and precipitation with physically based models, an achievement that, in parallel with ocean modeling advances, has laid the groundwork for true Earth system models. Cloud, water vapor, and radiative measurements have provided a basis for understanding connections between convection, radiation and atmospheric dynamics. These data sets have also guided the improvement of atmospheric hydrologic processes in regional and global models. Now these models typically carry conservation equations for all three phases of water as opposed to statistically inferring cloud and precipitation from water vapor amounts alone.

At this point we return to the science questions posed in the Section "Motivating/driving questions" and ask how these pieces now fit into the overall goal of understanding the global hydrologic cycle. What can we say about the nature of the water and energy cycles as viewed from space, and do there seem to be changes afoot? What have we learned about variability in the climate system as viewed from space? It is here, when we consider integrated signals and lower frequency signals, that the remaining challenges become more obvious. The precipitation and evaporation uncertainties related to ENSO suggest that natural variations in the water and heat balance of the tropics are clearly present, but their amplitude remains in question. We know that tropical and global SST signals may reach nearly 0.5 K for a large ENSO event. Signatures of increased (decreased) precipitation over ocean (land) emerge from GPCP. Suggestions of increased evaporation over oceans are also in evidence from microwave measurements. Yet these results are not yet definitive. Uncertainties and artifacts in the time series are significant compared to the signals of hydrologic processes averaged over large global regions.

In considering decadal scale signals we appeal to the TOA radiative flux record that is sensitive to cloud and water vapor distributions. The series of global satellite measurements begun by the Earth Radiation Budget Experiment has been continued under NASA's Clouds and Earth's Radiant Energy System (CERES) sensors on TRMM, Aqua, and Terra and has provided a stable, precise monitoring capability. From a TOA perspective the net radiative flux appears nearly invariant with net shortwave and outgoing long-wave fluxes of approximately 235 $\mathrm{Wm^{-2}}$ globally; clouds provide a small net cooling effect of about 20 $\mathrm{Wm^{-2}}$ (Ramanathan *et al.*, 1989; Kiehl and Trenberth, 1997). However, the ERBE/CERES measurement stream suggests the emergence during the 1990s of decreased shortwave reflection of several $\mathrm{W\,m^{-2}}$ over the tropical region 20° N/S and nearly offsetting increases in outgoing long-wave radiation (OLR) (Wielicki *et al.*, 2002a). Chen *et al.* (2002) have interpreted these radiative signatures as indications of strengthened tropical convection and vertical Hadley and Walker circulations. The net effect of these flux changes

is small, about 1.5 W m^{-2} increased radiation to the tropical earth/atmosphere system, and the quantitative accuracy of these changes is still under debate (Trenberth, 2002; Wielicki *et al.*, 2002b) with data set revisions in progress. These changes are of the same amplitude as higher frequency responses signatures associated with ENSO events but substantially smaller than the effects of the eruption of Mt. Pinatubo in June 1991.

Independent ISCCP cloud cover retrievals show a reduction in cloud cover of 3–5% over the 1983–2001 period and the trends in calculated fluxes from the ISCCP-FD data set match the ERBE/CERES quite well, yet mid- and high level cloud cover from the HIRS data set (Wylie and Menzel, 1999) gives little indication of the trend so these results remain to be reconciled. From the uncertainties discussed in the Section "Current capabilities", it is clear that our ability to detect changes (or confirm the lack thereof) in precipitation, evaporation, surface radiation, and land-surface signals is not yet at a stage where regional to global variations of order 1.0 W m^{-2} can be isolated.

Observational verification of water vapor feedback is complicated by the difficulty in isolating radiative/convective processes from dynamical effects (e.g. ENSO-related changes in subtropical jet structure) and to a lesser degree by uncertainties of cloud contamination in IR retrievals. However, with ever-lengthening time series and sensor improvements we can expect more robust time series measurements and perhaps a more thorough assessment of humidity changes over a greater variety of dynamical flow regimes. Improved cloud penetration via microwave sounding channels (e.g. 183.3 GHz) starting with the Special sensor Microwave Imager-2 (SSM/T-2) and now with the equivalent frequencies on the Advanced Microwave Sounding Unit-B (AMSU-B) sensors will be an important measurement series. Complementary vertical moisture structure from MLS, which will be continued on the Aura spacecraft, will certainly assist in a more robust understanding of UTH and the important intersection of the water cycle and radiative balance of the planet.

What is clearly needed at this point is increased emphasis on assessing the error characteristics of historical satellite retrievals, reprocessing and integrating those retrievals with contemporary and near-term planned improvements in remote sensing, and reconciling these with existing conventional *in situ* data. Our examination here of signals in precipitation, evaporation, and radiative fluxes suggests that at present we can detect regional patterns of interannual variability with considerable skill; we can resolve globally integrated signals with only modest certainty, and for the most part, we have little confidence yet in decadal changes or trends. Nevertheless, the situation is encouraging because in many cases these retrievals have been done without regard to intercalibration of sensor data streams and, in the case of operational retrievals, algorithm changes and improvements that result in spurious changes. A near-term challenge will be to reprocess many of these data streams to enforce more stringent consistency requirements.

We touched only briefly here on assimilation of energy flux quantities into numerical models. Data assimilation strategies are likely the most objective and statistically optimal means of blending diverse measurements, but considerable work is still needed. In the past, data assimilation has been driven by requirements for operational numerical weather prediction with emphasis on use of state-variable information (temperature, wind, and moisture). Blending in water and energy flux data is much more of a challenge because these variables are closely related to model physics parameterizations – traditionally the weakest parts of these tools. There is currently much emphasis on improving the physical basis of model moist processes and radiation so they are more physically consistent and are amenable to accepting constraints by remotely sensed fluxes. The assimilation of TRMM precipitation into the NASA Goddard Earth Observing System (GEOS) model (Hou *et al.*, 2001) is an example of this ongoing work.

A LOOK TO THE FUTURE

One can usefully regard the past 15 years of progress as the first maturation of remote sensing and a period characterized by major advances in retrieval accuracy and breadth of physical processes measured. We are seeing increased integration of measurements from multiple sensors in retrieval algorithms and steady improvement in data assimilation techniques as tools for fitting models with data. Much of the impetus for this has come from the WCRP GEWEX and CLIVAR programs, coordinated national responses to the IPCC process, and other international advocacy efforts that have served as organizing entities for setting scientific and observational remote sensing requirements. These requirements include priorities of quantities to be measured, required new retrieval technologies, analysis strategies, and associated accuracy needs.

What improvements can we expect new remote sensing strategies to contribute? One of the attributes of future missions will be to place more reliance on active remote sensing. In the near-term, the CloudSat mission with its profiling 94 GHz radar to be launched in 2005 will provide unprecedented vertical structure of cloud hydrometeors over the globe. In addition to demonstrating the value of this active remote sensing technology, CloudSat will provide new ways of examining relationships between clouds and other properties of the atmosphere that are important for understanding the earth's hydrological cycle and how cloud feedbacks are established within the climate system (Stephens *et al.*, 2002).

With respect to precipitation and land-surface controls on energy fluxes, several planned efforts are of special note. The Global Precipitation Mission, GPM, is currently envisioned as an expansion of TRMM concepts to a continuous, long-term precipitation monitoring capability. A dual-frequency radar will be at the core of this mission and will include a 35 GHz measurement to improve detection of light rainfall amounts. The concept for GPM involves use of this facility instrument in a high inclination orbit to calibrate a more extensive constellation of passive microwave sensors flying on operational and dedicated research platforms. This strategy allows for ~3 h temporal resolution over much of the globe needed for many hydrometeorological applications. As a true international effort, JAXA and NASA will provide the core platform with other international partners contributing passive instruments, ground validation, and other ancillary processing. The earliest expected launch date for the core platform is currently 2011, although passive components of this system will be in place as extensions of current operational meteorological satellites.

Scheduled for launch in early 2007, ESA's Soil Moisture and Ocean Salinity (SMOS) mission () is to further the development of climatological, meteorological, and hydrological models by observing soil moisture over the Earth's landmasses and sea-surface salinity over the oceans for a period of at least 3 years. SMOS will carry the first ever polar-orbiting satellite-borne 2-D interferometric radiometer MIRAS (Microwave Imaging Radiometer by Aperture Synthesis). From an altitude of 763 km, the novel MIRAS instrument has been designed to capture images of microwave radiation emitted from the surface of the Earth at L-band (1.4 GHz). The NASA Hydrospheric State (HYDROS) mission, projected for launch in 2010 will provide global mapping of soil moisture and freeze/thaw state. HYDROS will combine the attributes of L-band active and passive microwave sensing to meet the resolution (order of tens of kilometers) and accuracy requirements while minimizing undesirable vegetation and atmospheric effects. This mission should help assess how well key hydrologic processes are being represented in models and assimilation systems.

These missions are but a sampling of how continued international efforts to insure remote sensing from space serves the broader earth science community in which global hydrology takes its place. As we consider how far remote sensing of the water cycle has come in recent decades, prospects for the future seem bright. A scientific quest, which was once data limited, is now becoming one of synthesizing consistency from an abundance of detailed, yet admittedly still in many ways indirect, satellite measurements. Consistency of interrelationships between water and energy processes is bound to be a prominent theme as time series of water and energy fluxes are examined. The advances in quantifying hydrologic processes across scales to produce a truly integrated understanding will surely be a continuing challenge, but one which will have considerable scientific and practical yield.

Acknowledgments

The author expresses his appreciation to Dr. William Lapenta for discussions and insight regarding land-surface data assimilation and to Dr. M. Rodell for the GLDAS data presented in Figure 4. This work was supported by the NASA Global Water and Energy Cycle Program, Dr. Jared Entin being the Program Manager.

REFERENCES

Adler R.F., Huffman G.J., Chang A., Ferraro R., Xie P.-P., Janowiak J., Bruno R., Schneider U., Curtis S., Bolvin D., Gruber A., Susskind J., Arkin P. and Nelkin E. (2003) The Version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, **4**, 1147–1167.

Arkin P.A. (1979) The relationship between fractional coverage of high cloud and rainfall accumulations during GATE over the B-Scale array. *Monthly Weather Review*, **107**(10), 1382–1387.

Armstrong R.L. and Brodzik M.J. (2001) Recent northern hemisphere snow extent: a comparison of data derived from visible and microwave satellite sensors. *Geophysical Research Letters*, **28**, 3673–3676.

Bates J.J., Jackson D.L., Breon F.-M. and Bergen Z.D. (2001) Variability of tropical upper tropospheric humidity 1979–1998. *Journal of Geophysical Research*, **106**(D23), 32271–32281.

Berg W., Kummerow C. and Morales C.A. (2002) Differences between east and west pacific rainfall systems. *Journal of Climate*, **15**, 3659–3672.

Boccippio D.J., Goodman S.J. and Heckman S. (2000) Regional differences in tropical lightning distributions. *Journal of Applied Meteorology*, **39**(12), 2231–2248.

Bony S., Lau K.-M. and Sud Y.C. (1997) Sea surface temperature and large-scale circulation influences on tropical greenhouse effect and cloud radiative forcing. *Journal of Climate*, **8**, 2055–2077.

Brunke M.A., Fairall C.W., Zeng X., Eymard L. and Curry J.A. (2003) Which bulk aerodynamic algorithms are least problematic in computing ocean surface turbulent fluxes? *Journal of Climate*, **16**, 619–635.

Carlson T.N. (1986) Regional scale estimates of surface moisture availability and thermal inertia using remote thermal measurements. *Remote Sensing Review*, **1**, 197–246.

Chen J., Carlson B.E. and Del Genio A.D. (2002) Evidence for strengthening of the tropical general circulation in the 1990s. *Science*, **295**, 841–844.

Chou S.-H. (1993) A comparison of airborne eddy correlation and bulk aerodynamic methods for ocean–air turbulent fluxes during cold-air outbreaks. *Boundary-Layer Meteorology*, **64**, 75–100.

Chou S.-H., Atlas R.M., Shie C.-L. and Ardizzone J. (1995) Estimates of surface humidity and latent heat fluxes over oceans from SSM/I data. *Monthly Weather Review*, **123**, 2405–2425.

Chou S.-H., Shie C.-L., Atlas R.M. and Ardizzone J. (1997) Air–sea fluxes retrieved from special sensor microwave imager data. *Journal of Geophysical Research*, **102**, 12705–12726.

Chou S.-H., Shie C.-L., Atlas R.M. and Ardizzone J. (2003) Air–sea fluxes retrieved from special sensor microwave imager data. *Journal of Geophysical Research*, **102**, 12705–12726.

Clayson C.A., Fairall C.W. and Curry J.A. (1996) Evaluation of turbulent fluxes at the ocean surface using surface renewal theory. *Journal of Geophysical Research*, **101**, 28503–28513.

Curry J.A., Bentamy A., Bourassa M.A., Bourras D., Bradley E.F., Brunke M., Castro S., Chou S.H., Clayson C.A., Emery W.J., *et al.* (2004) SEAFLUX. *Bulletin of the American Meteorological Society*, **85**, 409–424.

Curtis S. and Adler R. (2000) ENSO indices based on patterns of satellite-derived precipitation. *Journal of Climate*, **13**, 2786–2793.

da Silva A., Young C.C. and Levitus S. (1994) *Algorithms and Procedures*, *Vol. 1, Atlas of Surface Marine Data 1994*, NOAA Atlas NESDIS 6, NOAA, NESDIS, p. 83.

Delworth T.L. and Manabe S. (1988) The influence of potential evaporation on the variabilities of simulated soil wetness and climate. *Journal of Climate*, **1**, 523–547.

Dirmeyer P.A., Dolman A.J. and Nobuo S. (1999) The pilot phase of the global soil wetness project. *Bulletin of the American Meteorological Society*, **80**, 851–878.

Esbensen S.K., Chelton D.B., Vickers D. and Sun J. (1993) An analysis of errors in special sensor microwave imager evaporation estimates over the global oceans. *Journal of Geophysical Research*, **98**(C4), 7081–7101.

Escoffier C., Bates J.J., Chédin A., Rossow W.B. and Schmetz J. (2001) Comparison of upper tropospheric humidity retrievals from TOVS and Meteosat. *Journal of Geophysical Research*, **106**, 5227–5238.

Fairall C.W., Bradley E.F., Hare J.E., Grachev A.A. and Edson J.B. (2003) Bulk parameterization of air-sea fluxes: updates and verification for the COARE algorithm. *Journal of Climate*, **16**, 571–591.

Fairall C.W., Bradley E.F., Rogers D.P., Edson J.B. and Young G.S. (1996b) Bulk parameterization of air–sea fluxes for tropical ocean global atmosphere coupled Ocean–Atmosphere response experiment. *Journal of Geophysical Research*, **101**(C2), 3747–3764.

Ferraro R.F., Grody N.C., Weng F. and Basist A. (1996) An eight-year (1987–1994) time series of rainfall, clouds, water vapor, snow cover, and sea ice derived from SSM/I measurements. *Bulletin of the American Meteorological Society*, **77**, 891–905.

Gleckler P.J. and Weare B.C. (1997) Uncertainties in global ocean surface heat flux climatologies derived from ship observations. *Journal of Climate*, **10**(11), 2764–2781.

Grassl H., Jost V., Kumar R., Schulz J., Bauer P., Schluessel P., (2000) *The Hamburg Ocean-Atmosphere Parameters and Fluxes from Satellite Data (HOAPS): A Climatological Atlas of Satellite-derived Air-sea-Interaction Parameters Over the Oceans*, Report No. 312, ISSN 0937–1060, Max Planck Institute for Meteorology, Hamburg.

Griffith C.G., Woodley W.L., Grube P.G., Martin D., Stout J. and Sikdar D. (1978) Rain estimation from geosynchronous satellite imagery – visible and infrared studies. *Monthly Weather Review*, **106**(8), 1153–1171.

Gupta S.K., Ritchey N.A., Wilber A.C., Whitlock C.H., Gibson G.G. and Stackhouse P.W. Jr (1999) A climatology of surface radiation budget derived from satellite data. *Journal of Climate*, **12**, 2691–2710.

Hahn C.J., Warren S.G. and London J. (1994) *Climatological Data for Clouds over the Globe from Surface Observations, 1982–1991: The Total Cloud Edition.*, Technical Report No. NDP026A, p. 42 [Available from Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge.].

Hou A.Y., Zhang S.Q., da Silva A.M., Olson W.S., Kummerow C. and Simpson J. (2001) Improving global analysis and short-range forecast using rainfall and moisture observations derived from TRMM and SSM/I passive microwave sensors. *Bulletin of the American Meteorological Society*, **82**, 659–679.

Houser P.R., Rodell M., Jambor U., Gottschalck J., Cosgrove B., Radakovich J., Arsenault K., Bosilovich M., Entin J.K. and Walker J.P. (2001) The Global Land Data Assimilation Scheme (GLDAS). *BAHC-GEWEX News Joint Issue*, **11**(2), 11–13.

Huffman G.J., Adler R.F., Arkin P., Chang A., Ferraro R., Gruber A., Janowiak J., McNab A., Rudolf B. and Schneider U. (1997) The Global Precipitation Climatology Project (GPCP) combined precipitation dataset. *Bulletin of the American Meteorological Society*, **78**, 5–20.

Jaeger L. (1976) Monatskarten des Niederschlags fu¨ r die ganze Erde. *Bericht des Deutschen Wetterdienstes*, **139**, 33, Offenbach a.M., pp and plates.

Jones C., Peterson P. and Gautier C. (1999) A new method for deriving ocean surface specific humidity and air temperature: an artificial neural network approach. *Journal of Applied Meteorology*, **38**, 1229–1246.

Kiehl J.T. and Trenberth K.E. (1997) Earth's annual global mean energy budget. *Bulletin of the American Meteorological Society*, **78**, 197–197.

Kummerow C., Hong Y., Olson W.S., Yang S., Adler R.F., McCollum J., Ferraro R., Petty G., Shin D.-B. and Wilheit T.T. (2001) The evolution of the Goddard Profiling Algorithm (GPROF) for rainfall estimation from passive microwave sensors. *Journal of Applied Meteorology*, **40**, 1801–1820.

Lapenta W.M., Suggs R., Jedlovec G.J. and McNider R.T. (1999) Impact of assimilating GOES-derived land surface variables into the PSU/NCAR MM5. Preprints, Workshop on *Land-Surface Modeling and Applications to Mesoscale Models*, NCAR: Boulder, pp. 65–68.

Legates D.R. and Willmott C.J. (1990) Mean seasonal and spatial variability in gauge-corrected, global precipitation. *International Journal of Climatology*, **10**, 111–127.

Levitus S., Antonov J.I., Boyer T.P. and Stephens C. (2000) Warming of the world ocean. *Science*, **287**, 2285–2289.

Liu W.T. (1984) Estimation of latent heat flux with Seasat-SMMR, a case study in N. Atlantic. *Large-Scale Oceanographic Experiments and Satellites*, Gautier C. and Fieus M. (Eds.), Reidel Publishing: Norwell, 205–221.

Liu W.T., Katsaros K.B. and Businger J.A. (1979) Bulk parameterization of air-sea exchanges of heat and water vapor including the molecular constraints at the interface. *Journal of the Atmospheric Sciences*, **36**, 1722–1735.

Liu W.T. and Niiler P.P. (1984) Determination of monthly mean humidity in the atmospheric surface layer over oceans

from satellite data. *Journal of Physical Oceanography*, **14**, 1451–1457.

Martin D.W. and Suomi V.E. (1972) A satellite study of cloud clusters over the tropical north Atlantic Ocean. *Bulletin of the American Meteorological Society*, **53**, 135–156.

McCollum J.R., Gruber A. and Ba M. (2000) Discrepancy between gauges and satellite estimates of rainfall in equatorial Africa. *Journal of Applied Meteorology*, **39**, 666–679.

McNider R.T., Song A.J., Casey D.M., Wetzel P.J., Crosson W.L. and Rabin R.M. (1994) Toward a dynamic-thermodynamic assimilation of satellite surface temperature in numerical atmospheric models. *Monthly Weather Review*, **122**, 2784–2803.

Mears C., Smith D. and Wentz F.J. (2001) Comparison of SSM/I and buoy-measured wind speeds from 1987–1997. *Journal of Geophysical Research*, **106**, 11719–11729.

Meissner T., Smith D. and Wentz F.J. (2001) A 10-year intercomparison between collocated SSM/I oceanic surface wind speed retrievals and global analyses. *Journal of Geophysical Research*, **106**, 11731–11742.

Minschwaner K. and Dessler A.E. (2004) Water vapor feedback in the tropical upper troposphere: model results and observations. *Journal of Climate*, **17**, 1272–1282.

Mitchell K., Houser P., Wood E., Schaake J., Tarpley D., Lettenmaier D., Higgins W., Marshall C., Lohmann D., Ek M., *et al.* (1999) GCIP Land Data Assimilation System (LDAS) project now underway. *GEWEX News*, **9**(4), 3–6.

Mitchell K.E., Lohmann D., Houser P.R., Wood E.F., Schaake J.C., Robock A., Cosgrove A.B., Sheffield J., Duan Q., Luo L., *et al.* (2003) Multi-institution North American Land Data Assimilation System (NLDAS) project: utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, **108**(D22), 8841, doi:10.1029/2003JD003823.

Petersen W.A. and Rutledge S.A. (2001) Regional variability in tropical convection: observations from TRMM. *Journal of Climate*, **14**, 3566–3586.

Platnick S., King M.D., Ackerman S.A., Menzel W.P., Baum B.A., Riédi J.C. and Frey R.A. (2003) The MODIS cloud products: algorithms and examples from Terra. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 459–472.

Ramanathan V.R., Cess D., Harrison E.F., Minnis P., Barkstrom B.R., Ahmad E. and Hartmann D. (1989) Cloud-radiative forcing and climate: results from the Earth radiation budget experiment. *Science*, **243**, 57–63.

Randel D.L., Vonder Haar T.H., Ringerud M.A., Stephens G.L., Greenwald T.J. and Combs C.L. (1996) A new global water vapor data set. *Bulletin of the American Meteorological Society*, **77**, 1233–1246.

Reynolds R.W. and Smith T.S. (1994) Improved global sea surface temperature analyses. *Journal of Climate*, **7**, 929–948.

Richards F. and Arkin P. (1981) On the relationship between satellite-observed cloud cover and precipitation. *Monthly Weather Review*, **109**(5), 1081–1093.

Robertson F.R., Fitzjarrald D.E. and Kummerow C.D. (2003) Effects of uncertainty in TRMM precipitation radar path integrated attenuation on interannual variations of tropical oceanic rainfall. *Geophysical Research Letters*, **30**(4), 10.1029/2002GL016416.

Robertson F.R., Spencer R.W. and Fitzjarrald D.E. (2001) A new satellite deep convective ice index for tropical climate monitoring: possible implications for existing oceanic precipitation data sets. *Geophysical Research Letters*, **28**, 251–254.

Robock A., Lifeng L., Wood E.F., Wen F., Mitchell K.E., Houser P.R., Schaake J.C., Lohmann D., Cosgrove B., Sheffield J., *et al.* (2003) Evaluation of the North American land data assimilation system over the southern great plains during the warm season. *Journal of Geophysical Research*, **108**(D22), 8846. No. doi:10.1029/2002JD003245.

Rodell M., Houser P.R., Jambor U., Gottschalck J., Mitchell K., Meng C.-J., Arsenault K., Cosgrove B., Radakovich J., Bosilovich M., *et al.* (2004) The global land data assimilation system. *Bulletin of the American Meteorological Society*, **85**(3), 381–394.

Ropelewski C.F. and Halpert M.S. (1987) Global and regional scale precipitation patterns associated with the El Niño/Southern oscillation. *Monthly Weather Review*, **115**, 1606–1626.

Rossow W.B. and Schiffer R.A. (1991) ISCCP cloud data products. *Bulletin of the American Meteorological Society*, **72**, 2–20.

Rossow W.B. and Schiffer R.A. (1999) Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, **80**, 2261–2287.

Schlüssel P. (1996) Satellite remote sensing of evaporation over sea. In *Radiation and Water in the Climate System: Remote Measurements*, *NATO ASI Series, Vol. 145*, Raschke Erhard (Ed.), Springer Verlag: Heidelberg, pp. 431–461.

Schulz J., Meywerk J., Ewald S. and Schlüssel P. (1997) Evaluation of satellite-derived latent heat fluxes. *Journal of Climate*, **10**, 2782–2795.

Schulz J., Schlüssel P. and Grassl H. (1993) Water vapour in the atmospheric boundary layer over oceans from SSM/I measurements. *International Journal of Remote Sensing*, **14**, 2773–2789.

Simpson J., Adler R.F. and North G.R. (1988) A proposed tropical rainfall measuring mission (TRMM) satellite. *Bulletin of the American Meteorological Society*, **69**, 278–278.

Soden B.J. and Bretherton F.P. (1996) Interpretation of TOVS water vapor radiances in terms of layer-average relative humidities: method and climatology for the upper, middle, and lower troposphere. *Journal of Geophysical Research*, **101**, 9333–9343.

Soden B.J., Wetherald R.T., Stenchikov G.L. and Robock A. (2002) Global cooling following the eruption of Mt. Pinatubo: A test of climate feedback by water vapor. *Science*, **296**, 727–730.

Smith S.D. (1988) Coefficients for sea surface wind stress, heat flux, and wind profiles as a function of wind speed and temperature. *Journal of Geophysical Research*, **93**, 2859–2874.

Spencer R.W. and Braswell W.D. (1997) How dry is the tropical troposphere? Implications for global warming theory. *Bulletin of the American Meteorological Society*, **78**, 1097–1106.

Spencer R.W., Goodman H.M. and Hood R.E. (1989) Precipitation retrieval over land and ocean with the SSM/I: identification and characteristics of the scattering signal. *Journal of Atmospheric Oceanic and Technology*, **6**, 254–273.

Spencer R.W., Olson W.S., Wu R., Martin D.W., Weinman J.A. and Santek D.A. (1983) Heavy thunderstorms observed over land by the nimbus 7 scanning multichannel microwave radiometer. *Journal of Applied Meteorology*, **22**(6), 1041–1046.

Stephens G.L., Vane D.G., Boain R.J., Mace G.G., Saasen K., Wang Z., Illingworth A.J., O'Connor E.J., Rossow W.B., Durden S.L., Miller S.D., Austin R.T., Benedetti A., and Mitrescu C., The CloudSat Science Team (2002) The CLOUDSAT mission and the A-Train. *Bulletin of the American Meteorological Society*, **83**, 1771–1790.

Stubenrauch C.J., Rossow W.B., Scott N.A. and Chédin A. (1999) Clouds as seen by satellite sounders (3I) and imagers (ISCCP). Part III: spatial heterogeneity and radiative effects. *Journal of Climate*, **12**, 3419–3442.

Sun D.-Z. (2003) A possible effect of an increase in the warm-pool SST on the magnitude of El Niño warming. *Journal of Climate*, **16**, 185–205.

Sun D.-Z. and Trenberth K.E. (1998) Coordinated heat removal from the equatorial pacific during the 1986-86 El Nino. *Geophysical Research Letters*, **25**, 2659–2662.

Susskind J., Reuter D. and Chahine M.T. (1987) Clouds fields retrieved from HIRS/MSU data. *Journal of Geophysical Research*, **92**, 4035–4050.

Tao W.-K., Lang S., Olson W.S., Meneghini R., Yang S., Simpson J., Kummerow C., Smith E. and Halverson J. (2001) Retrieved vertical profiles of latent heat release using TRMM rainfall products for February 1998. *Journal of Applied Meteorology*, **40**, 957–982.

Toracinta E.R., Cecil D.J., Zipser E.J. and Nesbitt S.W. (2002) Radar, passive microwave, and lightning characteristics of precipitating systems in the tropics. *Monthly Weather Review*, **130**, 802–824.

Toracinta ER and Zipser EJ (2001) Lightning and SSM/I-ice-scattering mesoscale convective systems in the global tropics. *Journal of Applied Meteorology*, **40**, 983–1002.

Trenberth K.E., (2002) Comments on changes in tropical clouds and radiation. *Science*, **296**, 2095–2096.

Vonder Haar T.H., Forsythe J.M., McKague D., Randel D.L., Ruston B. and Woo S. (2003) *Continuation of the NVAP Global Water Vapor Data Sets for Pathfinder Science Analysis*, STC Technical Report 3333, Science and Technology Corporation,. Hampton, 10 Basil Sawyer Drive, 23666-1393.

Wagner D., Ruprecht E. and Simmer C. (1990) A combination of microwave observations from satellites and EOF analysis to retrieve vertical humidity profiles over the ocean. *Journal of Applied Meteorology*, **29**, 1142–1157.

Warren S.G. and Hahn C.J. (1986) *Global Distribution of Total Cloud Cover and Cloud Type Amounts Over Land*, NCAR Technical Note TN-2731STR, National Center for Atmospheric Research, Boulder, p. 209 [NTIS DE87-006903.]

Warren S.G., Hahn C.J., London J., Chervin R.M. and Jenne R.L. (1988) *Global Distribution of Total Cloud over and Cloud Type Amounts over the Ocean*, NCAR Technical Note TN-3171STR, National Center for Atmospheric Research, Boulder, p. 212 [Available from UCAR Publications, P.O. Box 3000, Boulder,CO 80307.]

Weare B.C. (2000) Near-global observations of low clouds. *Journal of Climate*, **13**, 1255–1268.

Webb M., Senior C., Bony S. and Morcrette J.-J. (2001) Combining ERBE and ISCCP data to assess clouds in the Hadley centre, ECMWF and LMD atmospheric climate models. *Climate Dynamic*, **17**, 905–922.

Wentz F.J. (1997) A well calibrated ocean algorithm for special sensor microwave/imager. *Journal of Geophysical Research*, **102**, 8703–8718.

Wentz F.J. and Spencer R.W. (1998) SSM/I rain retrievals within a unified all-weather ocean algorithm. *Journal of Atmospheric Science*, **55**, 1613–1627.

Wetzel P.J. (1984) Determining soil moisture from geosynchronous satellite infrared data: a feasibility study. *Journal of Climate Applied Meteorology*, **23**, 375–391.

Wielicki B.A., Del Genio A.D., Wong T., Chen J., Carlson B.E., Allan R.P., Robertson F., Jacobowitz H., Slingo A., Randall D., Kiehl J.T., *et al.* (2002b) Response to changes in tropical clouds and radiation. *Science*, **296**, 2095A.

Wielicki B.A., Wong T., Allan R., Slingo A., Kiehl J.T., Soden B.J., Gordon C.T., Miller A.J., Yang S.-K., Randall D., *et al.* (2002a) Evidence for large decadal variability in the tropical mean radiative energy budget. *Science*, **295**, 841–844.

Wilheit T.T., Chang A.T.C., King J.L., Rodgers E.B., Nieman R.A., Krupp B., Milman A.S., Stratigos J.S. and Siddalingaiah H. (1982) Microwave radiometric observations near 19.35, 92 and 183 GHz of precipitation in tropical storm cora. *Journal of Applied Meteorology*, **21**(8), 1137–1145.

Wilheit T.T., Chang A.T.C., Rao M.S., Rodgers E.B. and Theon J.S. (1977) A satellite technique for quantitatively mapping rainfall rates over the oceans. *Journal of Applied Meteorology*, **16**, 551–560.

Woodley W.L. and Sancho B. (1971) A first step toward rain estimation from satellite photographs. *Weather*, **26**, 279–289.

Wylie D.P. and Menzel W.P. (1999) Eight years of high cloud statistics using HIRS. *Journal of Climate*, **12**, 170–184.

Yu W., Doutriaux M., Sèze G., Le Treut H. and Desbois M. (1996) A methodology study of the validation of clouds in GCMs using ISCCP satellite observations. *Climate Dynamics*, **12**, 389–401.

Zeng X., Zhao M. and Dickinson R.E. (1998) Intercomparison of bulk aerodynamic algorithms for the computation of sea surface fluxes using TOGA COARE and TAO data. *Journal of Climate*, **11**, 2628–2644.

Zhang Y.-C., Rossow W.B. and Lacis A.A. (1995) Calculation of surface and top of atmosphere radiative fluxes from physical quantities based on ISCCP datasets, 1. Method and sensitivity to input data uncertainties. *Journal of Geophysical Research*, **100**, 1149–1165.

Zhang Y.-C., Rossow W.B., Lacis A.A., Oinas M. and Mishchenko M.M. (2004) Calculation of radiative flux profiles from the surface to top-of-atmosphere based on ISCCP and other global datasets: refinements of the radiative transfer model and the input data. *Journal of Geophysical Research* **109**, D19105, doi:10.1029/2003JD004457.

Zwally H.J., Comiso J.C., Parkinson J.C., Cavalieri D.J. and Gloersen P. (2002) Variability of antarctic sea ice 1979–1998. *Journal of Geophysical Research*, **107**(C5), 3041–3060.

177: The Role of Large-Scale Field Experiments in Water and Energy Balance Studies

W JAMES SHUTTLEWORTH[1] AND JOHN HC GASH[2]

[1]*Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ, US*
[2]*Centre for Ecology & Hydrology, Wallingford, Oxfordshire, UK*

Over the last two decades, increased awareness of the potential importance of anthropogenic drivers of "global change" and growing interest in predicting climate variability at seasonal to interannual timescales has stimulated the development of new experimental techniques and new modeling methods to upscale conventional observations of land-surface exchanges to area scales consistent with the grid scales used in General Circulation Models, and large-scale experiments in important global biomes that had previously been neglected. The nature, purpose, and scale of such large-scale field experiments, many fostered by international programs such as the World Climate Research Programme (WCRP) and the International Geosphere-Biosphere Programme (IGBP), has evolved with time. Initial studies focused on exploring the value of remotely sensed data when upscaling or provided data at several sites that sampled the most important vegetation covers across large biomes. One important class of experiments, the Mesoscale Field Experiments, laid emphasis on gathering data over an area comparable with that of a Global Circulation Model (GCM) grid square and explored the nature and significance of mesoscale coupling between the land surface and the overlying atmosphere and, in particular, the response of the atmosphere to heterogeneity in vegetation cover. Subsequently, the spatial scale of studies increased in a suite of Continental-Scale Experiments (CSEs) which explored the extent to which the atmospheric component of energy and water budgets can be reconciled with the equivalent surface budgets, fostered improvement in the performance of the regional scale–coupled hydrometeorological models, and investigated "land memory" processes that operate at regional scale and at the seasonal timescales. Recently the emphasis has been on seeking to link several CSEs together in a Coordinated Enhanced Observing Period (CEOP) using satellite systems to provide observations at the global scale. This article describes the evolving purpose and developing nature of large-scale field experiments over the period 1984–2004 and their role in providing improved knowledge and modeling of land-surface energy exchanges.

INTRODUCTION

In the late 1970s and early 1980s there was a rapid increase in awareness of the large-scale changes taking place in the global environment and interest in predicting climate variability. Anthropogenic drivers such as deforestation, desertification, and the buildup of greenhouse gases in the atmosphere became recognized as having the potential to alter the climate of the whole planet and as prospective contributors to what came to be collectively known as "global change". The development of Global Circulation Models (GCMs, *see* **Chapter 178, Modeling of the Global Water Cycle: Numerical Models (General Circulation Models), Volume 5**) and their use for predicting the climatic effects of large-scale land-cover change (e.g. Taylor *et al.*, 2002) and increased radiative forcing by greenhouse gases (e.g. Cox *et al.*, 2000) together with an emerging capability in predicting climate variability at all timescales, including seasonal to interannual, created a need for improved knowledge and modeling of land-surface energy exchange and evaporation at the scale of the grids at which GCMs operate. Typically, the grids used in GCMs have a length scale of hundreds of kilometers and cover an area which in most parts of the world comprises a patchwork of soils and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

vegetation types. This posed a new challenge to scientists previously used to thinking at the plot or field scale. How can land-surface interactions be adequately represented at the scale of a grid square, biome, or large river basin? To address this question, new experimental techniques had to be created and new modeling methods developed to upscale conventional observations. From the outset, it was recognized that remote sensing would likely play a major role in doing this. At the same time, there was also recognition that field studies of important global biomes had been neglected. Previous observations of energy exchange and evaporation had largely been limited to temperate latitudes. Similar studies were therefore needed in the key biomes important in the context of global change: in the vast, remote, and sparsely populated areas of the tropics and high latitudes; for rainforest, savanna, tundra, and boreal forest biomes. For such biomes, climatic conditions and difficult access combine to make the logistics of experimental field studies challenging and expensive.

To meet the needs of global change science, international programs such as the World Climate Research Programme (WCRP) and the International Geosphere–Biosphere Programme (IGBP) were created and set research agendas to better understand the interactions and controls that define the global environment. A series of large-scale field experiments were initiated in which all the components of the energy, water, and latterly carbon balance were measured simultaneously over large areas.

A CATALOGUE OF EXPERIMENTS AT INCREASING SPATIAL SCALE

Table 1 draws heavily on Dirmeyer and Hoff (2004) and, although not exhaustive, lists most of the major international large-scale field experiments over the period 1984 to 2004. It does not include large-scale field experiments for which the focus was on improving understanding of ocean–atmosphere exchanges or of atmospheric chemistry and aerosols. The most noticeable feature of the experiments included in Table 1 is that they have progressively evolved towards documenting land–surface interactions over increasing spatial scales over this period. Some early experiments, notably those stimulated under the International Satellite Land Surface Climatology Project (ISLSCP), gave emphasis to exploring the use of remotely sensed data to upscale land–surface interactions. In these experiments, data was typically gathered over 100–400 km^2 (e.g. First ISLSCP field experiment (FIFE): Sellers and Hall, 1992; Hall and Sellers, 1995) in some cases simultaneously at more than one location (e.g. Boreal ecosystem-atmosphere study (BOREAS): Sellers *et al.*, 1995; Hall, 1999). Other experiments provided simultaneous data at several sites which sampled the most important vegetation covers at several places across large biomes (e.g. Anglo-Brazilian Climate Observation Study (ABRACOS): Gash and Nobre, 1997).

One important class of experiments, the *Mesoscale Field Experiments* (e.g. HAPEX-MOBILHY: André *et al.*, 1988, 1989. EFEDA: Bolle *et al.*, 1993. HAPEX-Sahel, Goutorbe *et al.*, 1994, 1997), laid emphasis on gathering data over an area comparable with that of a GCM grid square. Subsequently, the spatial scale of studies increased in a suite of Continental-Scale Experiments (CSEs: e.g. GCIP, Coughlan and Avissar, 1996; Lawford, 1999: BALTEX, Raschke *et al.*, 1998: LBA, LBA SCIENCE Planning Group, 1996: GAME, Yasunari, 1993: MAGS: Stewart *et al.*, 1998). Recently the emphasis has been on seeking to link several CSEs together in a Coordinated Enhanced Observing Period (CEOP) using satellite systems to provide observations at the global scale (see Koike, 2002).

MESOSCALE FIELD EXPERIMENTS

The design of Mesoscale Field Experiments owed much to the ocean–atmosphere experiments carried out in the 1960s and 1970s (see Garstang and Fitzjarrald, 1999) which involved intensive measurement campaigns with multiple teams of researchers making coordinated measurements from ships, buoys, balloons, and aircraft. In the equivalent land-atmosphere studies, such as HAPEX-MOBILHY (Hydrological-Atmosphere Pilot Experiment – Modélisation du Bilan Hydrique; André *et al.*, 1988) in France, there was emphasis on simultaneous measurements to sample the surface exchanges of different vegetation covers and soils. Indeed, in the case of HAPEX-MOBILHY, the energy balance, evaporation, and soil moisture for all the major vegetation types in a 10^4 km square were monitored (*see* **Chapter 39, Surface Radiation Balance, Volume 1** and **Chapter 40, Evaporation Measurement, Volume 1** for a description of the techniques). Airborne remote sensing techniques and large-scale flux measurements by aircraft were also added. A critique of how these techniques have developed and how land–surface experiments have contributed to understanding is given by Gash and Kabat (2004) and elsewhere in Kabat *et al.* (2004; Part B). In HAPEX-MOBILHY it was found that, while the mixed agricultural land could be modeled as an aggregated surface, it was necessary to differentiate between agricultural land and forest. The marked difference in surface exchanges between these dissimilar vegetation types was found to result in different rates of convective boundary-layer development and, on occasions, also in different cloud cover (e.g. Shuttleworth, 1988).

HAPEX-Sahel (Goutorbe *et al.*, 1994, 1997) provides a good example of a Mesoscale Field Experiment held as an intensive field campaign. This experiment was designed

Table 1 List of most of the major international large-scale field experiments over the period 1984 to 2004

Name	Location	Period	References
Hydrological and Atmospheric Pilot Experiment – Modelization du Bilan Hydrique (HAPEX-MOBILHY)	Southern France	1985–1987	André *et al.* (1989)
First ISLSCP Field Experiment (FIFE)	Central Kansas, USA	1987–1989	Sellers and Hall (1992), Hall and Sellers (1995)
Regio-Klima-Projekt (REKLIP)	Middle and southern upper Rhine Valley	1989	Parlow (1996)
Anglo-Brazilian Climate Observation Study (ABRACOS)	Manaus, Ji-Paraná and Marabá, Brazil	1990–1995	Gash and Nobre (1997)
Hydrological and Atmospheric Pilot Experiment in the Sahel (HAPEX-Sahel)	Western Niger	1991–1993	Goutorbe *et al.* (1994)
European International Project on Climatic and Hydrological Interactions between Vegetation, Atmosphere, and Land Surface (ECHIVAL) Field Experiment in Desertification Threatened Areas (EFEDA)	South-eastern Spain	1991–1995	Bolle *et al.* (1993)
Hei Ho River Basin Field Experiment (HEIFE)	Gansu Province, China	1992–1993	Wang *et al.* (1993)
Atmospheric Radiation Measurement Southern Great Plains Site	Kansas and Oklahoma, USA	1992	**http://www.arm.gov/docs/sites/sgp/sgp.html**
Boreal Ecosystem-Atmosphere Study (BOREAS)	Central Canada	1993–1996	Sellers *et al.* (1995), Hall (1999).
Mackenzie GEWEX Study (MAGS)	Mackenzie River basin, Canada	1994	Stewart *et al.* (1998)
Observation at Several Interacting Scales (OASIS)	Murray-Darling basin, Australia	1994–1995	**http://www.clw.csiro.au/research/waterway/interactions/oasis**
Northern Hemisphere Climate Processes Land Surface Experiment (NOPEX)	Central Sweden	1994–1996	Halldin *et al.* (1999)
Baltic Sea Experiment (BALTEX)	Baltic Sea basin	1994–2001	Raschke *et al.* (1998)
Monitoring the Usable Soil Reservoir Experimentally (MUREX)	South-western France	1995–1997	Calvet *et al.* (1999)
GEWEX Continental-scale International Project (GCIP)	Mississippi River basin, USA	1995–2000	Coughlan and Avissar (1996), Lawford (1999)
GEWEX Asian Monsoon Experiment (GAME)	Siberia, Tibet, Thailand, Huaihe River basin, China	1996	Yasunari (1993)
Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA)	Amazon region of South America	1996	LBA Science Planning Group (1996)
Inner Mongolia Grassland Atmosphere-Surface Study (IMGRASS)	Xilinhot, Inner Mongolia	1997–2000	**http://www.iap.ac.cn/english/iap/Divisions/LAGEO.htm**
Southern Great Plains (SGP)	Oklahoma and Kansas, USA	1997, 1999, 2001	**http://hydrolab.arsusda.gov/sgp97/ http://daac.gsfc.nasa.gov/CAMPAIGN_DOCS/SGP99/**
Semiarid Land-Surface-Atmosphere Program (SALSA)	Upper San Pedro River basin, Mexico, and USA	1997–1998	Goodrich *et al.* (2000)
Couplage de l'Atmosphère Tropicale et du Cycle Hydrologique (CATCH)	Niger, Benin	2000	**http://www.lthe.hmg.inpg.fr/catch/**
GEWEX Americas Prediction Project (GAPP)	USA	2000–2005	**http://www.ogp.noaa.gov/mpe/gapp/gapp/index.htm**
Coordinated Enhanced Observing Period (CEOP)	Worldwide	2001–2004	**http://www.ceop.net/**

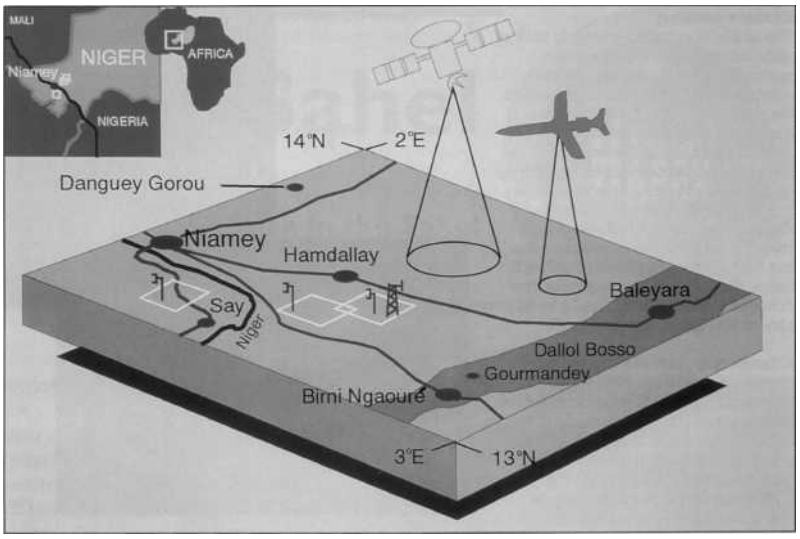


Figure 1 The HAPEX-Sahel experimental square in Niger, West Africa. In August and September 1992, some 180 scientists collected soil moisture and hydrological data, vegetation and flux measurements, and atmospheric boundary-layer and remote sensing data. HAPEX-Sahel is typical of other Mesoscale Field Experiments and was designed to investigate how to represent surface interactions at the GCM grid square scale for a landscape of mixed vegetation

to address the issue of aggregation, that is, how to represent surface interactions at the grid square scale for a landscape of mixed vegetation, but it was also significant in being located in an important area of the world where potential desertification is an issue. Figure 1 shows the experimental layout of the HAPEX-Sahel experiment. Over a period of two months, intensive measurements of energy surface exchange, water, and carbon fluxes, and soil moisture were made at three "super-sites" and, at each of these, for the three main vegetation types present in the landscape. Aircraft were used to make flux remote sensing measurements, tethered balloons, and frequent ascents by free flying radiosondes were used to monitor the atmospheric boundary layer and the atmosphere above, and many plot-scale process studies investigated how the individual water pathways contributed to the overall hydrological functioning of the area. Mesoscale meteorological models were used as the framework for bringing the results together. As was the case for many large-scale field experiments, a special issue of a scientific journal, in this case a 50-paper issue of the Journal of Hydrology (Goutorbe *et al.*, 1997) was used to document many of the results of the study.

One of the key results from HAPEX-Sahel was the discovery that in sparse vegetation, with much bare soil exposed, there is rapid evaporation from the soil in the days after rain. Modeling this flux requires two-source models (*see* **Chapter 45, Actual Evaporation, Volume 1**). Importantly, it also results in instantaneous spatial patterns of evaporation which are controlled by the variability in rainfall rather than vegetation. However, over a season, systematic differences in evaporation otherwise masked by large day-to-day variation emerged for the different vegetation types (Gash *et al.*, 1997). These data, when applied in a GCM modeling experiment (Taylor *et al.*, 2002) in which the improved representation of the land surface was coupled with realistic estimates of demographic and land use change, provided a convincing demonstration of how changes in the land surface may affect climate in the Sahel. Surprising feedbacks between the land surface and the atmosphere were discovered at the meso- (1–10 km) scale. These feedbacks, again resulting from the evaporation from bare soil after rainfall, result in persistent rainfall patterns where rain falls preferentially over areas where it has fallen before (Taylor and Lebel, 1998; Clark *et al.*, 2003).

As previously stated, most of the Mesoscale Field Experiments typically carried out at 10^4 km study sites were primarily focused on investigating how best to represent area-average surface exchanges at grid scales. However, in the case of the ongoing Atmospheric Radiation Measurement Southern Great Plains (ARM-SGP) study (DOE, 1996), although many of the field systems used are similar, the main motivation is to investigate how best to parameterize atmospheric (cloud and radiation) processes in models with grid scales of 100 km or larger. Consequently, substantial additional sensing of the lower atmosphere is included. As part of its purpose, ARM-CART also explores the nature and significance of mesoscale coupling between the land surface and the overlying atmosphere and, in particular, the response of the atmosphere to heterogeneity in vegetation cover.

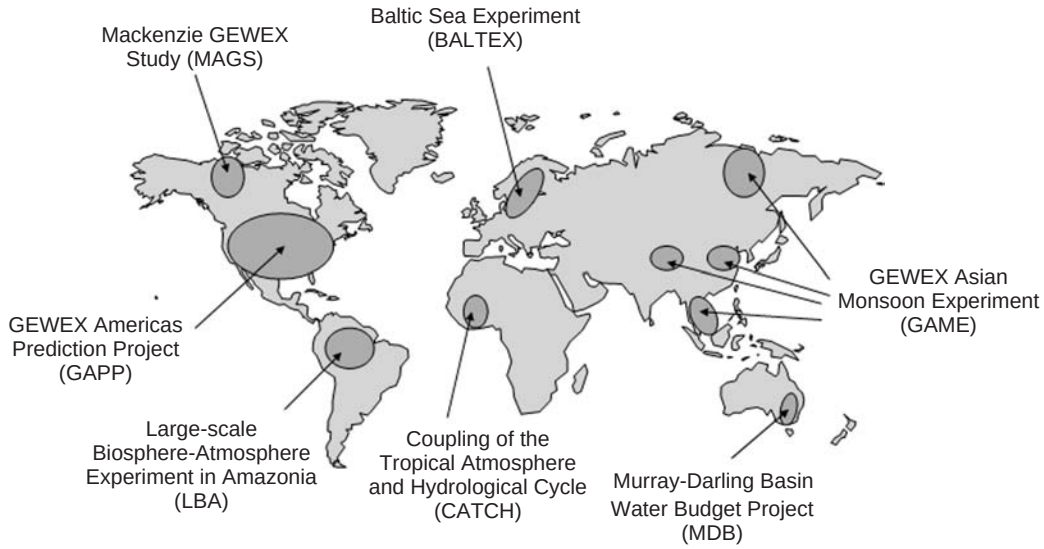


Figure 2 The location of the seven continental-scale experiments CSEs (GAPP, LBA, BALTEX, GAME, MAGS; MDB, and CATCH) coordinated by the GEWEX hydrometeorology panel and selected to sample important hydroclimatic regions across the globe. CATCH is still in development at the time of writing. CSEs operate over study areas substantially greater than Mesoscale Field Experiments, typically with length scales of 1000 km, or greater

CONTINENTAL-SCALE EXPERIMENTS

Over the last decade, WCRP has established several Continental-scale Experiments (CSEs) coordinated by the Global Energy Water-cycle Experiment (GEWEX) Hydrometeorology Panel (**http://ecpc.ucsd.edu/projects/ghp/ghp.html**). CSEs operate over study areas substantially greater than Mesoscale Field Experiments, typically with length scales of 1000 km, or greater. Figure 2 shows the location of seven CSEs (GAPP, LBA, BALTEX, GAME, MAGS; MDB, and CATCH) which were selected to sample important hydroclimatic regions across the globe.

An important motivation for CSEs is to explore the extent to which the atmospheric component of energy and water budgets documented can be reconciled with the equivalent surface budgets at an area scale of 10^6 km^2 or greater, areas this large being necessary for atmospheric estimates to have reasonable accuracy. In practice, however, their most direct and immediate benefit has arguably been to foster a marked improvement in the performance of the regional scale–coupled hydrometeorological models with which each CSE is associated, and in Numerical Prediction Models at several forecast centers around the world. Such improvement was achieved by providing better representation of surface energy and water exchange processes and atmospheric processes through the provision of relevant calibration and validation data. Associated with this improvement in the predictive ability of regional models operating at timescales up to seasonal, there has been a worthwhile improvement in the quality of the model-calculated fields derived by assimilating surface, atmospheric, and remotely sensed data into such models.

The simultaneous availability of improved regional models and continental-scale measurements of geophysical variables in CSEs also allows exploration of the "land memory" processes that operate at regional scale and at the seasonal timescales associated with, for example, soil moisture, snow, and ice cover, and vegetation growth and senescence. In parallel with the improvement in the regional and global models themselves, there has been improvement in the ability to initiate such models, most significantly through the development of Land Data Assimilation Systems (LDAS). A substantial additional benefit from CSEs has been a marked improvement in the interaction between hydrologists and atmospheric scientists, with the result that weather and climate predictions can receive more effective and immediate hydrological application, and the significance of hydrological processes in meteorological forecasting is better understood and represented in models.

COORDINATED ENHANCED OBSERVING PERIOD

It is likely that in the future remote sensing by air- and space-borne platforms of various kinds will be increasingly relied upon as the source of data for documenting energy and water related processes and phenomena. However, many key observational issues currently remain uncertain, such as which types of observations are most useful, what are the preferred horizontal, vertical, and time resolutions of these observations, and how can calibration best

be addressed. Recognizing these issues, the Coordinated Enhanced Observing Period (Koike, 2002) has brought together ongoing simultaneous data gathering by the several CSEs motivated by a desire to stimulate and test the application of a new generation of simultaneously available satellite sensors (TERRA, AQUA, ENVISAT, and ADEOS-II in addition to TRMM, Landsat-7, the NOAA-K series and several other operational satellites).

In practice, the simultaneous gathering of data from several CSEs also provides a new opportunity to test the transferability of understanding and models between CSEs in different climatic regions. The observation and data collection phase of CEOP extends from 1 July 2001 to 30 September 2004 with primary focus on the collection of a 2-year dataset beginning October 2002. Data collection includes not only basic satellite data, but also continuous observations from many CEOP reference sites in each of the CSEs, and data derived by assimilating remotely sensed, atmospheric sounding, and *in situ* climate data into models at many of the worlds leading numerical weather prediction centers. Perhaps CEOP can be viewed as a precursor to a true global scale observational study to document the Earth system as a whole by assimilating data from satellites, aircraft, and balloon systems, ships and bouys, and *in situ* observations over land surfaces into coupled ocean–atmosphere–land models operating at global scale.

REFERENCES

André J.-C., Bougeault P., Mahfouf J.-F., Mascart P., Noilhan J. and Pinty J.-P. (1989) Impact of forests on mesoscale meteorology. *Philosophical Transactions of the Royal Society of London Series B*, **324**, 407–422.

André J.-C., Goutorbe J.P., Perrier A., Becker F., Bessemoulin P., Bougeault P., Brunet Y., Brutsaert W., Carlson T., Cuenca R., *et al.* (1988) HAPEX-MOBILHY: first results from the special observing period. *Annales Geophysicae*, **6**, 477–492.

Bolle H.J., Oliver H.R. and Shuttleworth W.J. (1993) EFEDA: European field experiment in a desertification-threatened area. *Annales Geophysicae*, **11**, 173–189.

Calvet J.-C., Bessemoulin P., Noilhan J., Berne C., Braud I., Courault D., Fritz N., Gonzalez-Sosa E., Goutorbe J.-P., Haverkamp R., *et al.* (1999) MUREX: a land-surface field experiment to study the annual cycle of the energy and water budgets. *Annales Geophysicae*, **17**, 838–854.

Clark D.B., Taylor C.M., Thorpe A.J., Harding R.J. and Nicholls M.E. (2003) The influence of spatial variability of boundary-layer moisture on tropical continental squall lines. *Quarterly Journal of the Royal Meteorological Society*, **129**, 1101–1121.

Coughlan M. and Avissar R. (1996) The Global Energy and Water Cycle Experiment (GEWEX) Continental-Scale International Project (GCIP): an overview. *Journal of Geophysical Research*, **101**, 7139–7148.

Cox P.M., Betts R.A., Jones C.D., Spall S.A. and Totterdell I.J. (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, **408**, 184–187.

Dirmeyer P.A. and Hoff H. (2004) Motivation for data consolidation. In *Vegetation, Water, Humans and the Climate*, Kabat P., Claussen M., Dirmeyer P.A., Gash J.H.C., Bravo de Guenni L., Meybeck M., Pielke R.A. Sr, Vörösmarty C.J., Hutjes R.W.A. and Lütkmeier S. (Eds.), Springer: Heidelberg, pp. 247–253.

DOE (1996) *Science Plan for the Atmospheric Radiation Measurement (ARM) Program*, U. S. Department of Energy: Washington, DOE/ER-0670.

Garstang M. and Fitzjarrald D.R. (1999) *Observations of Surface to Atmosphere Interactions in the Tropics*, Oxford University Press: New York, p. 405.

Gash J.H.C. and Kabat P. (2004) Further insight from large-scale observational studies of land/atmosphere interactions. In *Vegetation, Water, Humans and the Climate*, Kabat P., Claussen M., Dirmeyer P.A., Gash J.H.C., Bravo de Guenni L., Meybeck M., Pielke R.A. Sr, Vörösmarty C.J., Hutjes R.W.A. and Lütkmeier S. (Eds.), Springer: Heidelberg, pp. 229–233.

Gash J.H.C., Kabat P., Monteny B.A., Amadou M., Bessemoulin P., Billing H., Blyth E.M., deBruin H.A.R., Elbers J.A., Friborg T., Harrison G., Holwill C.J., Lloyd C.R., Lhomme J.-P., Moncrieff J.B., Puech D., Sögaard H., Taupin J.D., Tuzet A. and Verhoef A., (1997) The variability of evaporation during the HAPEX Sahel Intensive Observation Period. *Journal of Hydrology*, **188–189**, 385–399.

Gash J.H.C. and Nobre C.A. (1997) Climatic effects of Amazonian deforestation: some results from ABRACOS. *Bulletin of the American Meteorological Society*, **78**, 823–830.

Goodrich D.C., Chehbouni A., Goff B., MacNish B., Maddock T., Moran S., Suttleworth W.J., Williams D.G., Watts C., Hipps L.H., *et al.* (2000) Preface paper to the Semi-Arid Land-Surface-Atmosphere (SALSA) program special issue. *Agricultural and Forest Meteorology*, **105**, 3–20.

Goutorbe J.-P., Lebel T., Dolman A.J., Gash J.H.C., Kabat P., Kerr Y.H., Monteny B., Prince S.D., Stricker J.N.M., Tinga A., *et al.* (1997) An overview of HAPEX-Sahel: a study in climate and desertification. *Journal of Hydrology*, **188–189**, 4–17.

Goutorbe J.-P., Lebel T., Tinga A., Bessemoulin P., Brouwer J., Dolman A.J., Engman E.T., Gash J.H.C., Hoeppfner M., Kabat P., *et al.* (1994) HAPEX-Sahel: a large scale study of land-atmosphere interactions in the semi-arid tropics. *Annales Geophysicae*, **12**, 53–64.

Hall F.G. (1999) BOREAS in 1999: experiment and science overview. *Journal of Geophysical Research-Atmospheres*, **104**(D2), 27627–27639.

Hall F.G. and Sellers P.J. (1995) First International Satellite Land Surface Climatology Project (ISLSCP) Field Experiment (FIFE) in 1995. *Journal of Geophysical Research*, **100**(D12), 25383–25395.

Halldin S., Gryning S.-E., Gottschalk L., Jochum A., Lundin L.-C. and Van de Griend A.A. (1999) Energy, water and carbon exchange in a boreal forest landscape – NOPEX experiences. *Agricultural and Forest Meteorology*, **98/99**, 5–29.

Kabat P., Claussen M., Dirmeyer P.A., Gash J.H.C., Bravo de Guenni L., Meybeck M., Pielke R.A. Sr, Vörösmarty C.J., Hutjes R.W.A. and Lütkmeier S. (2004) *Vegetation, Water, Humans and the Climate*, Springer: Heidelberg.

Koike T. (2002) CEOP starts – a step for predictability improvement of the water cycle and water resources. *CEOP Newsletter*, (**www.ceop.net**), **1**, 1.

Lawford R.G. (1999) A mid-term report on the GEWEX Continental-Scale International Project (GCIP). *Journal of Geophysical Research*, **104**(D16), 19279–19292.

LBA Science Planning Group (1996) *The Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA): Concise Experiment Plan*. Available from the LBA Project Office, CPTEC/INPE, Rodovia Presidente Dutra, km 40, Caixa Postal 01, 12630-000 Cachoeira Paulista, SP, p. 41.

Parlow E. (1996) The regional climate project REKLIP – an overview. *Theoretical and Applied Climatology*, **53**, 3–7.

Raschke E., Karstens U., Nolte-Holube R., Brandt R., Esemer H.-J., Lohmann D., Lobmeyr M., Rockel B. and Stuhlmann R. (1998) The Baltic Sea experiment BALTEX: a brief overview and some selected results of the authors. *Surveys in Geophysics*, **19**, 1–22.

Sellers P.J. and Hall F.G. (1992) FIFE in 1992: results, scientific gains, and future research directions. *Journal of Geophysical Research*, **97**(D17), 19091–19019.

Sellers P.J., Hall F.G., Margolis H., Kelly R.D., Baldocchi D., den Hartog G., Cihlar J., Ryan M.G., Goodison B., Crill P., *et al.* (1995) The Boreal Ecosystem-Atmosphere Study (BOREAS): an overview and early results from the 1994 field year. *Bulletin of the American Meteorological Society*, **76**, 1549–1577.

Shuttleworth W.J. (1988) Macrohydrology–the new challenge for process hydrology. *Journal of Hydrology*, **100**, 31–56.

Stewart R.E., Leighton H.G., Marsh P., Moore G.W.K., Ritchie H., Rouse W.R., Soulis E.D., Strong G.S., Crawford R.W. and Kochtubajda B. (1998) The Mackenzie GEWEX study: the water and energy cycles of a major North American river basin. *Bulletin of the American Meteorological Society*, **79**, 2665–2683.

Taylor C.N., Lambin E.F., Stephenne N., Harding R.J. and Essery R.L.H. (2002) The influence of land use change on climate in the Sahel. *Journal of Climate*, **15**, 3615–3629.

Taylor C.M. and Lebel T. (1998) Observational evidence of persistent convective scale rainfall patterns. *Monthly Weather Review*, **126**, 1597–1607.

Wang J., Gao Y. and Hu Y. (1993) An overview of the HEIFE experiment in the people's republic of China. Exchange processes at the land surface for a range of space and time scales. In *Proceedings of a Symposium Held During the Joint Meeting of the International Association of Meteorology and Atmospheric Physics and IAHS at Yokohama*, Bolle H.-J., Feddes R.A. and Kalma J.D. (Eds.), IAHS Publication 212, IAHS: July 1993, pp. 397–406.

Yasunari T. (1993) GEWEX-related Asian monsoon experiment (GAME). *Advances in Space Research*, **14**, 161–165.

178: Modeling of the Global Water Cycle: Numerical Models (General Circulation Models)

RANDAL D KOSTER

NASA/Goddard Space Flight Center, Greenbelt, MD, US

Atmospheric general circulation models (AGCMs) are powerful tools for exploring climate sensitivity and the global water cycle. The wealth of data provided by an AGCM allows a complete statistical characterization of its water cycle, a far more complete picture than would be allowed from real-world observations. AGCM numerical experiments have been used to isolate the mechanisms that control water cycle variability, illustrate the potential for water cycle prediction, and quantify water cycle changes resulting from human activities. The AGCM fields can also be merged with observations (via data assimilation) to provide the best estimates possible of the true water cycle. Unfortunately, biases in simulated fields resulting from coarse resolution and inadequate physical parameterizations continue to plague AGCMs, forcing climate modelers to temper their faith in the models' results.

INTRODUCTION: ATMOSPHERIC GENERAL CIRCULATION MODELS

Atmospheric general circulation models (AGCMs), usually coupled to land-surface models (LSMs) and sometimes to ocean models (OGCMs), are numerical representations of the atmosphere that allow the simulation of global weather patterns. The atmosphere in AGCMs is either explicitly (for grid point models) or effectively (for spectral models) discretized into elements with a length scale of 1–5 degrees in the horizontal and hundreds of meters in the vertical; the discretized time step is typically of the order of minutes. The equations of motion determine transports of mass and energy between the elements during each time step. Meanwhile, various physical parameterizations compute the sources and sinks of energy and water within all elements – they determine how much water vapor is converted to rain, how much radiation is absorbed by the atmosphere, and so on. Simulated weather is averaged over long periods of time to reveal the AGCM's inherent "climate".

AGCMs are used extensively to study and predict variability in climate and the global water cycle. For such applications, these models have some powerful advantages and some severe limitations. One key advantage is the ability of an AGCM to provide complete, global fields of self-consistent data over a wide variety of timescales, ranging from hours to decades or centuries. Parallel ensembles of decadal simulations can provide enough data to characterize completely an AGCM's simulated water cycle, though not at scales below the model's resolution. For some applications, observed data can be combined with model data in an optimal way to produce an enhanced product (see section Data assimilation below). AGCM output includes values for many quantities (e.g. evaporation from land) that simply cannot be measured at the global scale, either now or in the foreseeable future.

Another key advantage of an AGCM is its ability to host sensitivity studies. Modifications of model boundary conditions in the atmosphere (e.g. an increase in CO_2), at the land surface (e.g. projected deforestation), or the ocean surface (e.g. the onset of El Niño conditions) are typically straightforward to implement, and resulting impacts on simulated water cycle components can be quantified and analyzed. Such sensitivity experiments are arguably the best approach we have for predicting the likely impacts of human activities on climate. Sensitivity of modeled climate to changes in model parameterizations can identify the physical mechanisms that control variations in the water cycle in the model and thus, by extension, in the real world.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

The AGCM's main limitation is the inaccuracy of its simulated fields. The aspect of the AGCM that makes it useful for climate studies in the first place, namely, the ability of various model states to feed back on one another, also allows the climate of the model to drift into unrealistic situations – to generate unrealistic annual precipitation totals, for example, or an unrealistic radiation balance. The unwanted drift stems largely from inadequacies in the treatments of the many physical processes at work in the atmosphere, land, and ocean. Because many processes occur at spatial scales well below the resolution of the AGCM, overly simple, inaccurate parameterizations have to be used (section Model parameterizations and limitations), parameterizations that are further hampered by a lack of the global-scale measurements needed to define suitable parameter values. Climate scientists who use AGCMs must constantly qualify their results in light of the AGCM's deficiencies.

MODELED WATER CYCLE COMPONENTS

Together, AGCMs and their coupled LSMs simulate most of the components of the global water cycle. Modeled atmospheric processes include water vapor transport (both vertical and horizontal), cloud formation, and rainfall generation (both convective and nonconvective). Modeled land processes include evapotranspiration, canopy interception, snowpack growth and ablation, runoff generation, and subsurface soil moisture transport. Typically missing from the coupled AGCM-LSM system is a representation of deep groundwater; the modeled soil moisture variables typically represent no more than the top few meters of soil. Glacier and ice sheet dynamics are also typically missing. Oceanic transports of water could be provided through a coupled ocean model.

Our ability to test an AGCM's simulated global water cycle against observations is limited by a paucity of relevant global measurements. Large-scale direct measurements of evaporation, for example, are essentially nonexistent, and direct measurements of soil moisture and atmospheric vapor transport are far from complete. Fortunately, precipitation, a critical component of the water cycle, has been estimated on the global scale by many groups. The Global Precipitation Climatology Project (GPCP), for example, generated a product based on a combination of *in situ* gauge measurements and satellite observations (Huffmann *et al.*, 1997). Figure 1 shows a comparison between observed precipitation, averaged into zonal means, and the corresponding values derived from various AGCMs. The errors for all models are large, particularly in the tropics and midlatitudes.

Even more difficult for AGCMs to capture are the variability characteristics of precipitation. Soden (2000), for example, shows that AGCMs substantially underestimate

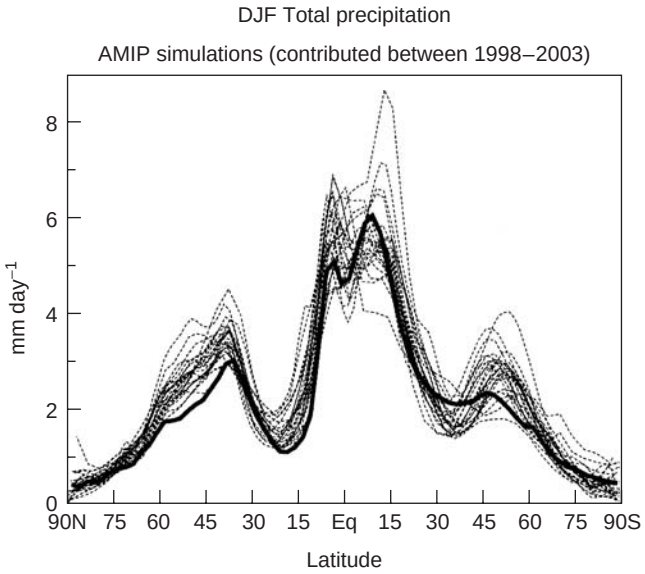


Figure 1 Zonal mean DJF precipitation totals from observations (heavy solid line), as derived from the dataset of Xie and Arkin (1998), and from AGCMs participating in the Atmospheric Model Intercomparison Project (dotted lines). Figure courtesy of Peter Gleckler

the observed interannual variability of precipitation in the tropics. In midlatitudes, on the other hand, an excessive land–atmosphere feedback can lead to overestimates of precipitation variance (Koster *et al.*, 2003). Although observations of year-to-year variability are certainly not perfect, the disagreement is most likely caused by AGCM error. The accurate reproduction of the observed year-to-year variability of the water cycle is typically given less attention in these models than the reproduction of observed means.

Surface and subsurface runoff (streamflow) generated by AGCMs can also be evaluated against observations. One comparison was provided by Russell and Miller (1990), who found, for example, that an early version of the Goddard Institute for Space Studies (GISS) AGCM strongly underestimated Amazon River outflow. Unfortunately, because the realism of simulated runoff depends in large part on the realism of simulated precipitation, and because errors in simulated precipitation abound (Figure 1), an evaluation of AGCM-simulated runoff may say little about the correctness of LSM runoff treatments. For a proper test of an LSM's runoff formulation, the LSM should be driven offline with the most realistic atmospheric forcing possible. Runoffs generated in this way through the Global Soil Wetness Project (GSWP) (Dirmeyer, 1999) tended to be significantly underestimated relative to observations in higher latitudes (Figure 2, from Oki *et al.*, 1999). Work is proceeding to determine if the LSMs are at fault or if strong measurement biases associated with the undercatch of snowfall are responsible.

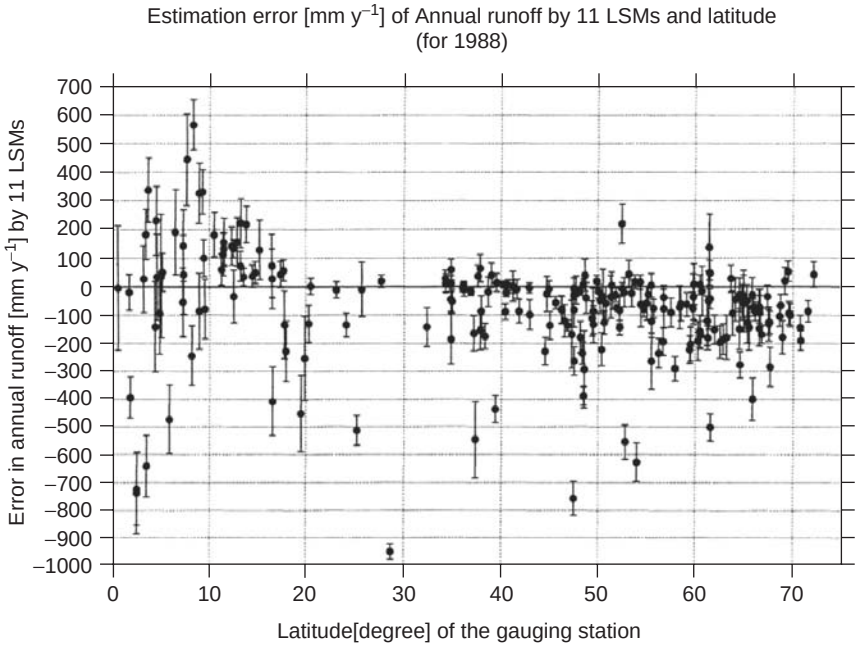


Figure 2 Errors in simulated runoff as a function of latitude, as determined from 11 land-surface models participating in the Global Soil Wetness Project (GSWP) (Reproduced from Oki *et al.*, 1999, by permission of the Meteorological Society of Japan)

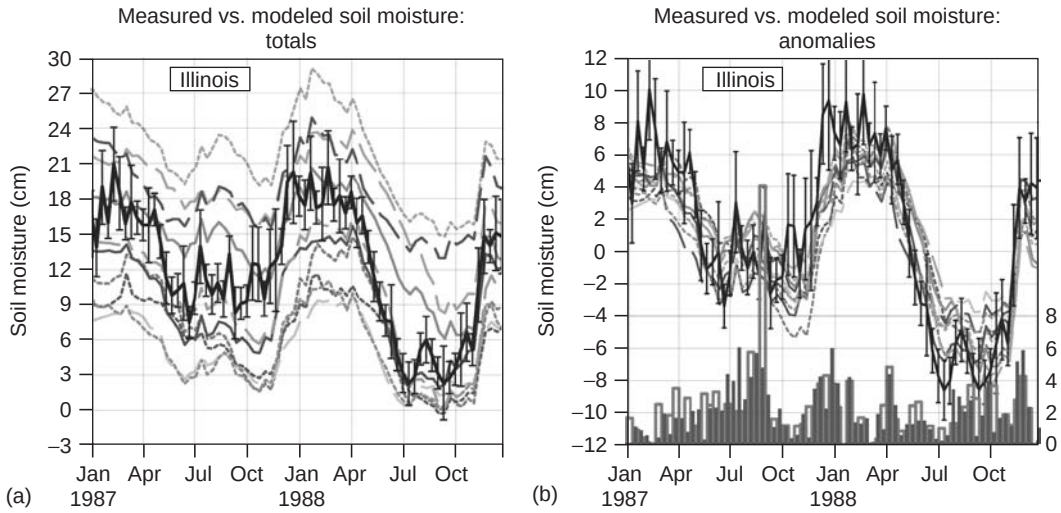


Figure 3 (a) Observed soil in Illinois (black line with error bars) and simulated soil moistures (other lines) in the same region, as determined by several models participating in the Global Soil Wetness Project (GSWP). (b) Same, but for the measured and simulated soil moisture anomalies relative to each dataset's own long-term mean. Histograms for precipitation are provided at the bottom of the graph (Reproduced from Entin *et al.*, 1999, by permission of the Meteorological Society of Japan). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

For similar reasons, while observed soil moisture contents can be compared to AGCM-generated values, a more useful exercise is to compare the observed values with those produced by the LSM running offline with realistic atmospheric forcing. Long-term observations are essentially limited to Asia and small parts of Illinois and Iowa (Robock *et al.*, 2000). These data were used by Entin *et al.* (1999) to examine the soil-moisture contents generated by the LSMs participating in GSWP. Figure 3 shows one of the results. The LSMs successfully capture the amplitude and phase of the seasonal cycle of observed soil moisture but fail to match the observed values' absolute magnitudes. This

is presumably due to the somewhat nebulous meaning of "soil-moisture" in LSMs. (See section Model parameterizations and limitations.)

Atmospheric vapor can be estimated from space. SMMI-derived estimates of total precipitable water (TPW), averaged over the year, are compared in Figure 4 to values simulated by the NASA Seasonal-to-Interannual Prediction Project (NSIPP) AGCM. The overall patterns are well reproduced. Observational estimates of atmospheric vapor transport, however, are more difficult to obtain; satellites cannot as easily sense wind direction or speed, and "*in situ*" radiosonde measurements are low in density in most parts of the world. Vapor transport is usually estimated via data assimilation, which involves the merging of an AGCM with radiosonde and satellite observations (see section Data assimilation below).

The examples above illustrate the sometimes large biases that pervade state-of-the-art AGCMs. Reductions of this bias can occur in two ways. First, the model can be "forced" to have more realistic fields through data assimilation, that is, by using observations to modify its state variables (section Data assimilation). Second, the model's representation of dynamics and physics can be improved through increased resolution and more powerful, realistic

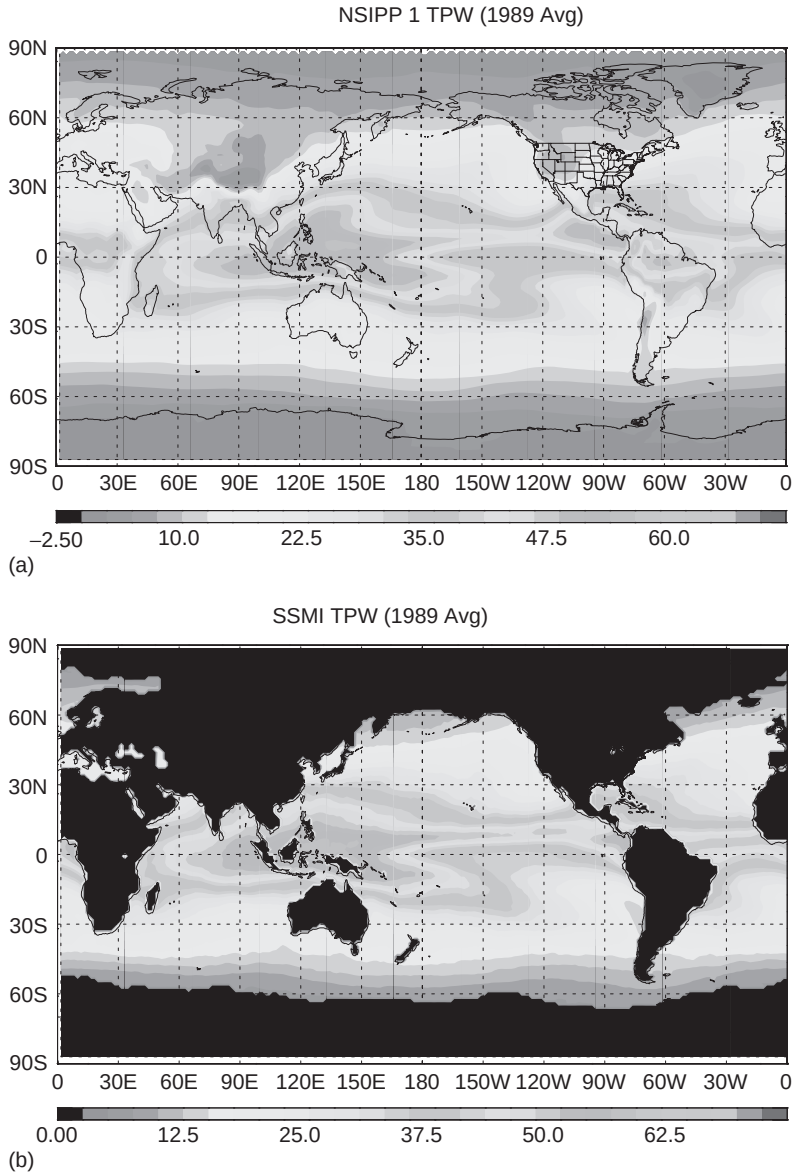


Figure 4 Comparison of total precipitable water (in mm) for 1989 as produced by an AGCM (a) and as measured by satellite (b). The satellite values are undefined over continents. Figure courtesy of Julio Bacmeister, UMBC, Maryland. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

parameterizations. This second approach is critical for climate sensitivity studies (section Variability analyses), which require a maximum number of unconstrained degrees of freedom and thus should not rely on ingested observations. Unfortunately, the pace of such improvements over the last few decades has been plodding at best. The search for reliable parameterizations of nonresolvable subgrid processes (section Model parameterizations and limitations) will continue to challenge modelers in the years to come.

Note that a philosophy adopted by many AGCM modelers is that inaccuracies in an AGCM's climate, if they are small enough, may be quite acceptable for sensitivity studies. The idea is simple – two simulations, one with and one without an imposed modification (e.g. a deforested surface), produce two model climates that can be directly compared. Assuming that the bias of concern appears in both climates, differencing the climates will tend to remove the bias and will thereby highlight climate differences associated with the imposed change. Of course, this approach is predicated on assumed linearities in the modeled system. This assumption is often untenable, particularly when the bias is large.

MODEL PARAMETERIZATIONS AND LIMITATIONS

As noted in section "Modeled water cycle components", an AGCM's simulation of the global water cycle is far from perfect. Much of the error is induced by the necessity of parameterizing physical processes that occur at scales too small to be resolved explicitly by the model. This is not just a temporary problem, one that will soon be eliminated by the advent of faster, more powerful computers. Increasing an AGCM's resolution by two requires about a 10-fold increase in computational power – a factor of two for both horizontal spatial dimensions and a factor of two for the time step. Thus, even if the most powerful computers today became a million times faster, the typical spatial resolution would reduce from about 200 km to about 3 km, still too large to resolve cloud updrafts, individual trees, the hillslope structure that controls runoff production, and other critical components of the Earth's system. Parameterizing the effects of these features with approximate bulk or statistical representations will remain an unavoidable strategy for the foreseeable future.

A cursory description of the evolution of land-surface process models should serve to illustrate some key limitations of parameterizations. The earliest LSMs simply prescribed soil moisture conditions, imposing, for example, dry conditions in deserts and wet conditions in tropical forests. Interactive land surface models were then introduced, the first and simplest being the "bucket" model of Manabe *et al.* (1969), which allows the water level in a soil moisture reservoir and thus the evaporation efficiency to increase during precipitation events and decrease as the water evaporates. In the mid-1980s, Sellers *et al.*, (1986) and Dickinson *et al.* (1986) introduced the "SVAT" (soil-vegetation-atmosphere-transfer) model, which allows vegetation to control the heat and moisture balances of a region. A simplified description of the basic SVAT transpiration calculation is provided in Figure 5.

Recently, various groups have extended the approach to the next level. The physics of photosynthesis is now explicitly included in many models (e.g. Bonan, 1995; Sellers *et al.*, 1996), using parameters that are strongly tied to satellite-based land-surface data. The idea is that because plants and trees open their stomata to maximize carbon uptake while minimizing water loss, an accurate representation of the physics of this uptake is needed to ensure realistic transpiration rates. Such emphasis on the carbon cycle is indeed noted by Sellers *et al.* (1997) as the next logical step (after the SVAT) in the evolution of LSMs. Related to carbon-cycle modeling are many recent efforts to model interactive vegetation phenology and/or species distribution, as noted in section Variability analyses.

While the latter advances are important, they represent an improvement over only one facet of the land surface. Other facets remain relatively simple, and this imbalance

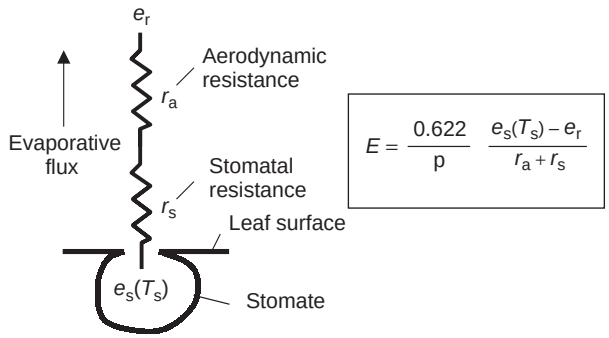


Figure 5 Basis of the "big leaf" transpiration calculation in the standard SVAT model. The vapor pressure within the representative stomate ($e_s(T_s)$, the saturated vapor pressure at the leaf temperature T_s) is separated from the vapor pressure in the air (e_r) by two resistances in series, one associated with the stomatal aperture (r_s) and one associated with transport through the overlying air (r_a, which accounts, for example, for wind speed and turbulent mixing effects). An Ohm's law analogy is applied: evaporation (E) is calculated as the difference between the vapor pressures (analogous to voltages) divided by the sum of the series of resistances. The constants ρ and p in the equation are the air density and air pressure, respectively. A key aspect of the SVAT model is the sensitivity of r_s to the environment – the modeled plants act to shut down transpiration by increasing r_s in times of stress

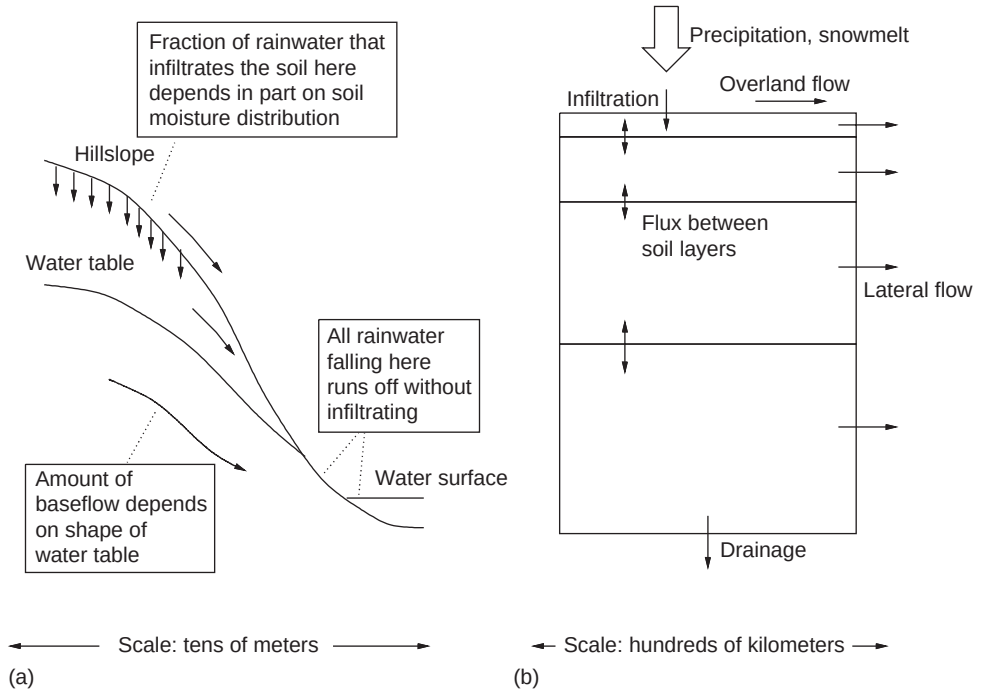


Figure 6 (a) Simple representation of the hillslope processes controlling runoff in nature. (b) Structural framework of a typical LSM. The focus on vertical physics prevents an explicit treatment of the processes highlighted in the left panel

in modeling complexity can lead to severe problems. For example, shown in Figure 6(a) is a representation of hillslope hydrology. The water table intercepts the land surface at the bottom of the hill. Rain falling at the top of the hill may infiltrate the soil, whereas rain hitting the bottom of the hill necessarily runs off. The shape of the water table controls the baseflow rate, an important component of the runoff. Because the soil is drier at the top of the hill, evaporation at the top may be less than that at the bottom.

The point to be made in Figure 6(a) is that subgrid soil moisture heterogeneity, as controlled by topography (among other things), determines in large part the average energy and water budget fluxes produced by a land surface. An explicit representation of soil moisture variability, however, is missing from the great majority of land-surface models – even those with the advanced treatments of photosynthesis and dynamic phenology. In most land-surface models, the soil is treated as shown in Figure 6(b), that is, as a column of vertically stacked soil layers, with soil moisture effectively assumed uniform in the horizontal over distances of hundreds of kilometers. Such a treatment arguably acts as an impassable barrier to a proper treatment of runoff generation. A few LSMs do try to come to grips with the subgrid soil moisture variability problem (e.g. Liang *et al.*, 1994; Famiglietti and Wood, 1994; Stieglitz, 1997; Koster *et al.*, 2000b), tying statistical distributions of soil moisture to observed topography. Such models, however, are arguably still immature.

One may argue that representing a grid cell's hydrology with a stacked layer model (as in Figure 6b), though not optimal for runoff, is acceptable in light of the improvements to evaporation afforded by the more advanced photosynthesis and phenology models. This argument, however, ignores the fact that evaporation and runoff processes are strongly interlinked through their mutual dependence on soil moisture (Koster and Milly, 1997). A poor representation of runoff, which is arguably inescapable with the layer approach, can be a "weak link" that reduces or even eliminates the benefits of an advancement in the treatment of evaporation.

The interaction between the evaporation and runoff schemes also has an impact on the interpretation of modeled soil moisture. The necessity of parameterizing the nonresolvable surface and subsurface physics leads to an LSM "soil moisture" that is arguably more an index of wetness than a physical quantity that can be measured in the field. This explains in large part the apparent poor simulation of soil moisture by various LSMs in Figure 3. The model-dependent nature of the soil moisture variable implies that soil moisture generated in one LSM should not be inserted blindly into another (Koster and Milly, 1997). In time, as representations of LSM physics improve, direct comparisons of simulated soil moisture between models and with observations should become more appropriate.

DATA ASSIMILATION

To quantify the global water cycle, a complete slate of observations, with minimal measurement error, is clearly desirable. A complete set of high-quality, high-resolution observations on the global scale, however, is logistically impossible to collect, both now and for the foreseeable future. For the desired global data, many rely instead on a technique known as data assimilation.

Data assimilation is the optimal marriage between measurements and modeling. In essence, the technique combines measurements and model predictions into a single data set that is superior to either standing alone. The ability of observations to improve on model results should not be surprising, given the model errors outlined in section Modeled water cycle components above. To see how model results could in turn improve on observations, consider the following simple thought experiment. Air temperature is measured in two adjacent regions. The observations in the western region are excellent, with significant spatial coverage and high-quality instruments. In the eastern region, however, observations are spotty, and the instruments used are poor. If we relied on observations alone, we would have only a rough, faulty estimate for the average temperature in the eastern region. If the wind blows from west to east, however, then the temperature in the west (measured extremely well) should have some bearing on the temperature in the east and can be used to improve its estimation. All that is needed for the correction is some representation of how the two regions are physically connected. A model (e.g. an AGCM) that is built upon well-established, fundamental equations of motion and state in the atmosphere can provide the needed representation.

The mathematics underlying data assimilation is complex and is outlined elsewhere in the encyclopedia. Put simply, a given variable (air temperature, humidity, soil moisture, etc.) may have both a measured value and a model-generated value (which reflects, in part, earlier measurements). The "merged" state produced by the assimilation system is a weighted average of these two values, the weights depending on the relative size of their errors. The error analysis – quantifying the measurement error for the observation systems and the evolution of the model errors associated with crude parameterizations, inappropriate parameter values, poorly defined boundary conditions, and so on – is typically key to the system's design.

Various "reanalyses" – data assimilation procedures performed with a single model ingesting decades of data in a consistent way – provide an optimal, though imperfect, view of the global water cycle (e.g. Roads *et al.*, 2002). Figure 7 shows moisture flux fields (July–September) generated with CDAS, the Climate Data Assimilation System of the National Center for Environmental Prediction (NCEP) in the United States. Such a detailed description of moisture transport could not be generated from observations

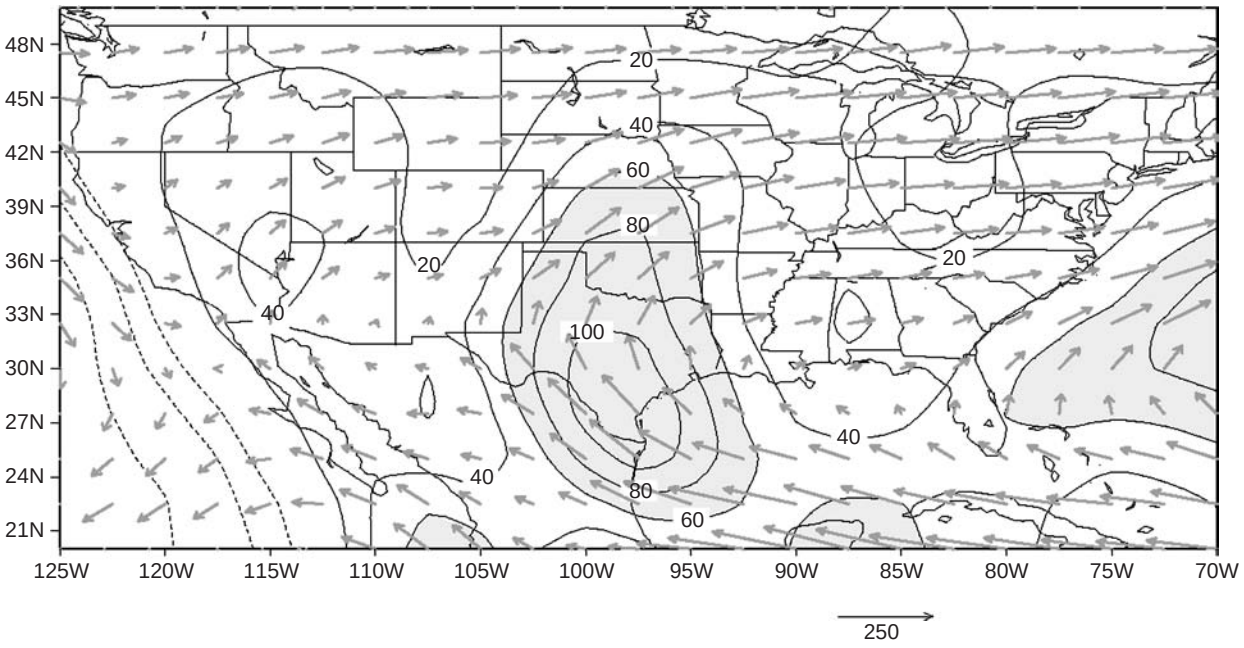


Figure 7 Vertically integrated moisture flux (vector) averaged for the period 1995 to 2000 for July–September from CDAS. The unit vector is 250 kg (ms)$^{-1}$. The meridional flux is shaded and contoured. The contour interval is 20 kg (ms)$^{-1}$. Values greater than 60 kg (ms)$^{-1}$ are shaded. (Figure courtesy of Kingtse Mo, NCEP.) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

alone. Note, though, that the meridional flux in Figure 7 is not perfect; it should have a maximum in Oklahoma and should be stronger coming out of the Gulf of California (K. Mo, personal communication, 2003).

Assimilating precipitation information into a system is challenging and has been avoided in the first reanalyses. As a result, the precipitation generated by a model in a reanalysis tends to show error relative to the independent observational datasets (Curtis *et al.*, 2001). Note further that fields based on little observational input are not necessarily sound, even if other fields are. Evaporation rates produced by data assimilation systems can be biased (e.g. Betts and Viterbo, 2000); they are arguably less reliable than those generated offline with observed precipitation forcing.

TRACER STUDIES

In addition to providing estimates of quantities that are very difficult to observe, such as evaporation, AGCMs can provide information on quantities that are utterly *impossible* to measure, even though they lie at the heart of the global water cycle. An important example is the distribution of evaporative sources for a given precipitation volume. Any molecule of water that falls as rain or snow in a given region must have had some prior evaporative source – it must have most recently evaporated from some location on the Earth's surface. Establishing the sources of a region's precipitation is important for establishing how that region's climate is "teleconnected" to that of other regions. Unfortunately, definitively determining from observations where water originates is impossible, since all water molecules look alike. Sources can only be inferred indirectly, for example, by examining the rainwater's isotopic content or by making assumptions regarding large-scale mixing of the atmosphere upwind of the precipitation.

AGCMs, however, can "tag" the water that evaporates from a given source region and diagnostically follow it until it precipitates. By dividing the earth's surface into source regions and then tagging the evaporated water from each as an independent tracer, the modeled rainfall at any location can be precisely separated into contributions from the source regions – the relative importance of each source region can be precisely quantified. The procedure can provide, for example, length scales of horizontal vapor transport and water recycling rates (Figure 8).

Tracer studies performed with free-running AGCMs (e.g. Joussaume *et al.*, 1986; Koster *et al.*, 1993) are, of course, subject to the many errors in the AGCM's climate. The results cannot be directly verified and are thus unavoidably questionable. Recently, however, a tracer diagnostic has been incorporated into a full data assimilation system (the NASA Data Assimilation Office system), implying a more reliable representation of tracer transport (Bosilovich *et al.*, 2003). Figure 9 shows an example calculation with this improved system.

VARIABILITY ANALYSES

As mentioned above, the AGCM allows sensitivity experiments that isolate and quantify the physical mechanisms that control the global water cycle. To some extent, any

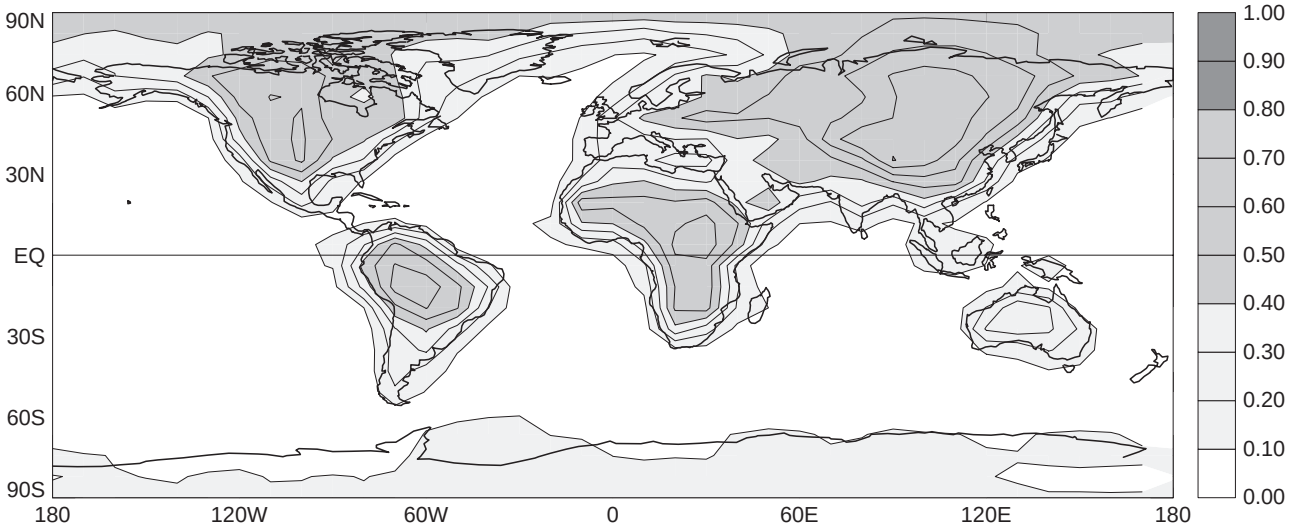


Figure 8 Recycling ratio: fraction of annual precipitation water derived from continental evaporative sources, as determined by a tracer water version of the Goddard Institute for Space Studies AGCM (Reproduced from Koster *et al.*, 1993, by permission of American Geophysical Union)

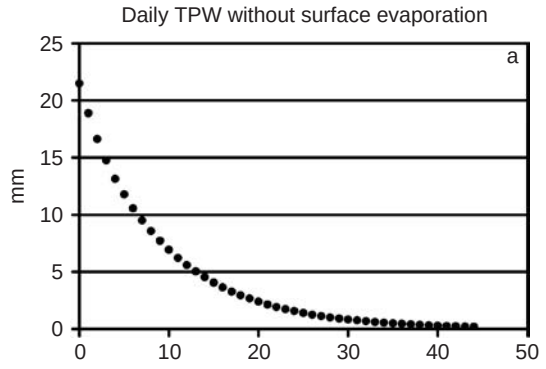


Figure 9 Results of a tracer model simulation in which the initial atmospheric water vapor is tagged as a particular tracer (Bosilovich *et al.*, 2003). The plot shows the reduction (due to rainout) of this tracer with time. The average residence time of water vapor in the atmosphere, as determined by fitting this curve with an exponential, is 9.27 days, which is significantly different from that determined by the more expedient but more approximate method of dividing the total precipitable water by the precipitation flux (7.55 days)

survey of these experiments will be incomplete, given the profusion of existing AGCM climate studies, and it will soon be dated, since new, innovative experiments are constantly being devised and performed. With this caveat in mind, a brief survey is provided here. The intent is to illustrate the type of studies being done without attempting to be comprehensive.

Contributions to Precipitation Variability

Delworth and Manabe (1988, 1989) pioneered the use of AGCMs to evaluate the impact of land variability on climate. Their basic experiment consists of two parallel AGCM simulations – one in which the land-surface soil moisture is free to interact with the atmosphere (e.g. rainfall wets the soil, leading to increased subsequent evaporation) and one in which the land surface is "fixed", meaning that soil moistures are forced to follow a climatological seasonal cycle, so that soil moisture and subsequent evaporation do not respond to anomalies in rainfall. A difference in the character of the atmospheric variability between the two simulations would indicate the existence of land–atmosphere feedback. Land–atmosphere feedback is the process by which rainfall-induced soil-moisture anomalies affect, for example, through modified evaporation, the subsequent evolution of the atmosphere's dynamics. Through land–atmosphere feedback, a period of anomalously heavy rainfall might act to sustain itself.

Delworth and Manabe (1989) found that coupled land processes produced, in their AGCM, a clear impact on near-surface air temperature and humidity. The impact on precipitation, however, was not significant. In a similar study with a different AGCM, Koster *et al.* (2000a) examined the joint impact of land-surface variability and ocean variability on the variability of precipitation. Their results supported a useful interpretation of the following simple tautology:

$$\sigma_{\mathrm{P}}^2 = \sigma_{\text{P-(no land)}}^2 [X_{\mathrm{O}} + (1 - X_{\mathrm{O}})] \frac{\sigma_P^2}{\sigma_{\text{P-(no land)}}^2} \qquad (1)$$

where σ_{P}^2 is the variance of precipitation at a point. According to the equation, this total precipitation variance is the product of the variance that would be obtained in the absence of land–atmosphere interaction ($\sigma_{\text{P-(no land)}}^2$) and an amplification factor ($\sigma_P^2/\sigma_{\text{P-(no land)}}^2$) associated with the interaction. The variance $\sigma_{\text{P-(no land)}}^2$ can in turn be separated into two parts: the fraction explained by ocean variability (X_{O}) and the fraction explained by chaotic atmospheric dynamics ($1 - X_{\mathrm{O}}$). Figure 10, from Koster *et al.* (2000a), shows the global distribution of $\sigma^2{}_{\text{P-(no land)}}$, X_{O}, $1 - X_{\mathrm{O}}$, and $\sigma_{\mathrm{P}}^2/\sigma_{\text{P-(no land)}}^2$, as established in several ensembles of multidecadal simulations using different combinations of enabled or disabled land and ocean variability. In agreement with other AGCM studies (Kumar and Hoerling, 1995; Shukla, 1998; Trenberth *et al.*, 1998), ocean impacts (X_{O}) are limited largely to the tropics; elsewhere, atmospheric chaos ($1 - X_{\mathrm{O}}$) controls precipitation variability. Land–atmosphere feedback amplifies the variability in various parts of the globe. Other studies in this genre include those by Dirmeyer (2001), who examined atmospheric response to prescribed time series of land-surface states, and Douville (2003), who examined the impact of relaxing soil-moisture states to a climatological mean. One study (Koster *et al.*, 2002) suggests a large amount of intermodel variation in the strength of simulated land–atmosphere feedback. Note that observed estimates of feedback strength are not available – our observational networks do not provide the necessary data.

Contributions to Precipitation Predictability

The potential impact of land–atmosphere feedback on precipitation has led to speculation that soil moisture can contribute to precipitation prediction. AGCM studies are uniquely suited to testing this hypothesis. Observed pluvial periods and droughts have been reproduced in AGCMs when soil moistures are maintained at suitably wet or dry levels during the period in question (S. Schubert, as cited in Entekhabi *et al.*, 1999; Hong and Kalnay, 2000). The positive impact of maintaining an observations-based soil-moisture field on precipitation has also been demonstrated in less extreme years (Dirmeyer, 2000). Figure 11, for example, shows where the prescription of land moisture affects precipitation in the Center for Ocean-Land-Atmosphere Studies AGCM. Some studies suggest that the prescription of soil moisture has the largest impact in transition zones between wet and arid regions (Koster *et al.*, 2000a).

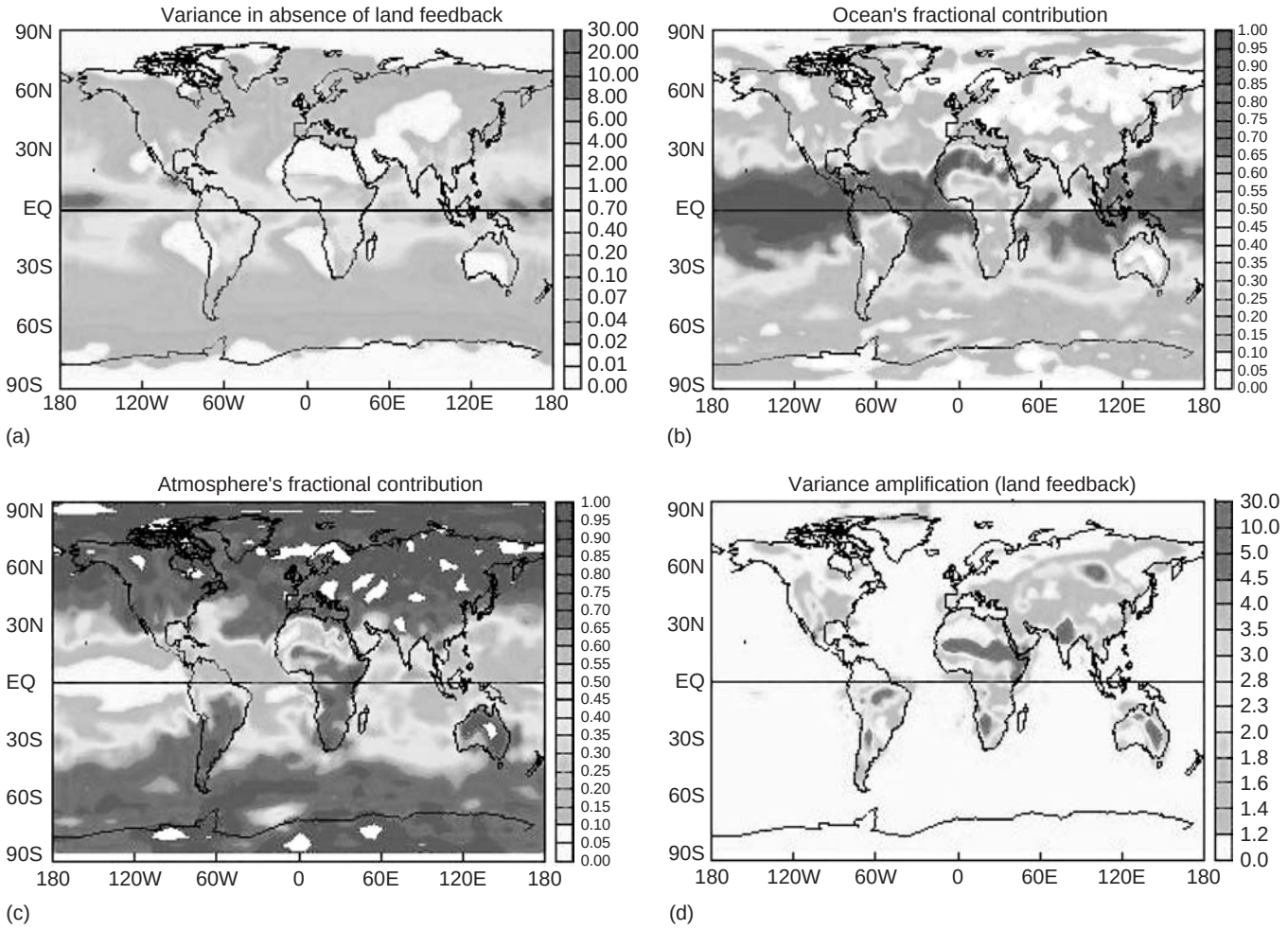


Figure 10 AGCM-generated estimates of contributions to precipitation variance. (a) $\sigma^2_{P\text{-(no land)}}$, the variance that would be achieved in the absence of land–atmosphere feedback. (b) X_O, the contribution of ocean (SST) variability to precipitation variability. (c) $1 - X_O$, the contribution of chaotic atmospheric dynamics to precipitation variability. (d) $\sigma^2_P/\sigma^2_{P\text{-(no land)}}$, the amplification of precipitation variance due to land–atmosphere feedback. (Koster *et al.*, 2000a © 2000 American Meteorological Society). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Of course, for true forecasts, soil moistures may not be prescribed throughout a simulation; they can only be initialized, with hopes that the initial anomaly will last well into the forecast period and that precipitation will continue to respond to the persisting anomaly. Initializing AGCMs with large, idealized soil-moisture anomalies does lead to changes in subsequent AGCM precipitation (Rind, 1982; Oglesby and Erickson, 1989; Beljaars *et al.*, 1996; Schar *et al.*, 1999). Initializing an AGCM with less extreme anomalies can also provide some predictability, though to a lesser level (Wang and Kumar, 1998; Schlosser and Milly, 2002; Oglesby *et al.*, 2002). Comparisons of AGCM-forecasted precipitation against *observed* precipitation allow a quantification of the skill associated with soil-moisture initialization. These studies are still in their infancy and are limited by the short data record. Nevertheless, existing results are suggestive and even encouraging (Fennessy and Shukla, 1999; Viterbo and Betts, 1999; Douville and Chauvin, 2000; Koster and Suarez, 2003; Koster *et al.*, 2004).

Snowpack represents another surface moisture reservoir with a memory long enough to contribute to seasonal prediction. Compared to soil-moisture studies, relatively few AGCM studies have focused on snow-climate feedbacks or on the climatic impacts of snow initialization or specification (e.g. Cohen *et al.*, 1991; Walsh and Ross, 1988; Schlosser and Mocko, 2003). The potential usefulness of the snow reservoir's memory is not yet fully known.

The ocean heat reservoir has a much stronger memory than the land moisture reservoirs. Thus, for long-term prediction, many consider the initialization of the ocean in a coupled atmosphere-ocean modeling system to be

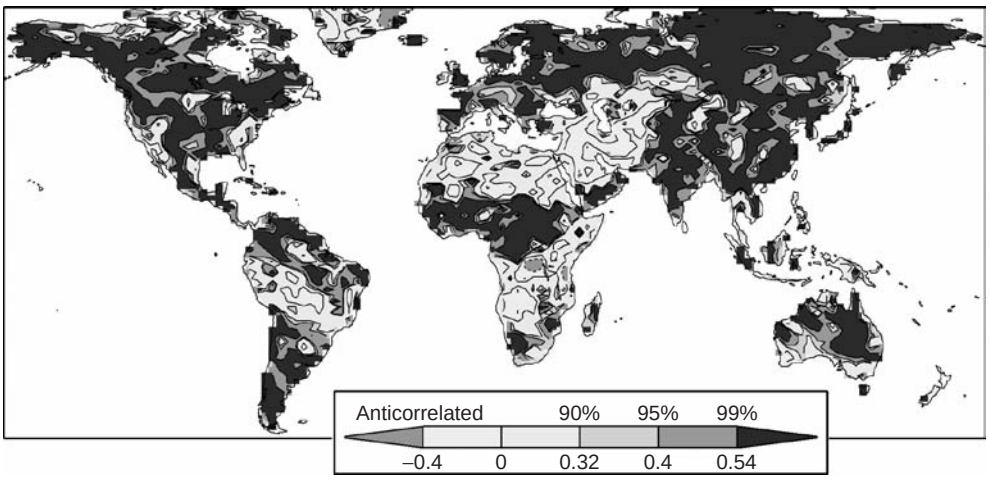


Figure 11 Shaded part indicates regions where a continually prescribed soil wetness has a direct impact, through evaporation, on precipitation simulated during boreal summer in a climate model. Percentages indicate the likelihood that simulated impacts are not the product of chance (Figure courtesy of Paul Dirmeyer, COLA.) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

most critical. The US National Center for Environmental Prediction (NCEP), the European Center for Medium-Range Weather Forecasts (ECMWF) and other institutions routinely issue seasonal precipitation predictions based largely on predicted sea surface temperatures (SSTs).

Impacts of Anthropogenic Change

Human-induced modification of the land surface – deforestation in the Amazon, for example, or the global-scale conversion of forests to cropland – may already have had a significant impact on the global water cycle. Because the land surface can be modified arbitrarily in AGCMs, these models are effective tools for examining such impacts.

Chase *et al.* (2001) and Pitman and Zhao (2000) incorporated estimates of the natural vegetation distribution (i.e. the vegetation distribution that would exist today if humans were not present) into AGCMs and showed that the climate impacts associated with human modification of land cover are comparable to those predicted under a doubling of CO_2. Amazonian deforestation and its impacts on climate and the water cycle have been the subject of a great many studies (e.g. Henderson-Sellers *et al.*, 1993; Polcher and Laval, 1994; Sud *et al.*, 1996), many of which suggest that deforestation leads to warmer local temperatures and reduced local precipitation (see review by Hahmann and Dickinson, 1997). Some deforestation studies suggest that impacts can also be felt remotely (Gedney and Valdes, 2000; Werth and Avissar, 2002). Other land impact studies have addressed, for example, North American land use change (Bonan, 1997) and desertification in the Sahel (Xue and Shukla, 1993).

The impact of increased greenhouse gases on global temperature has long been a subject of serious study. Their impact on the global water cycle is now being given considerable attention as well (e.g. Roads *et al.*, 1996; Zhang *et al.*, 2001; Costa and Foley, 2000; Pitman and McAvaney, 2002). At issue is whether a warmer atmosphere can induce significant changes in average regional rainfall or, just as important, significant changes in the variability of the hydrological cycle – in particular, an increased likelihood of a devastating drought or flood. AGCMs are the only tool we have for addressing such questions. Their limitations, however, are manifest when comparing the results of AGCMs performing the same climate change experiment (e.g. Gedney *et al.*, 2000). While all of the AGCMs indicate some global warming, they often disagree about the direction of regional hydrological change.

Additional studies

AGCMs are flexible enough for many additional types of water cycle studies. The following is just a sampling.

When evaluating an AGCM climate change experiment (such as one involving a doubling of CO_2), one should remember that parts of the AGCM were calibrated, or "tuned", to produce a reasonable representation of present-day climate. The relevance of such a calibration in a modified climate must clearly be considered. One way to evaluate the multiclimate appropriateness of calibrations is to simulate climates of the past, for which some validation data do exist. The Paleoclimate Modeling Intercomparison Project (PMIP) has organized a number of paleoclimatic simulations and has shown, for example, that the models qualitatively (if not quantitatively) reproduce the monsoon shifts responsible for the relative lushness of northern Africa 6000 years BP (Joussaume *et al.*, 1999).

The monsoon is indeed a unique and critical mechanism for bringing ocean-derived moisture onto continents. AGCM studies of monsoon mechanics have addressed,

for example, the impacts of land evaporation (Lau and Bua, 1998), SST distribution and vegetation representation (Laval *et al.*, 1996), and continental layout (Dirmeyer, 1998) on monsoon strength. Kang *et al.* (2002) describe a multi-AGCM intercomparison of monsoon simulation.

The relative control of energy availability and water availability on the annual water balance of a continental region was examined long ago by Budyko (1974). AGCMs provide a means of addressing such issues with comprehensive data (e.g. Milly and Dunne, 1994; Koster and Suarez, 1999).

Isotopes of water can be followed with the tracer diagnostics described in the section Tracer studies above (Joussaume *et al.*, 1984; Jouzel *et al.*, 1987). These heavy isotopes, which participate in every aspect of the global water cycle, behave just like regular water except for a tendency to fractionate during changes of phase – to favor slightly the condensed phase. The monitoring of isotopic tracers, in conjunction with *in situ* isotope measurements, may reveal important subtleties of the global water cycle (e.g. Hoffmann *et al.*, 2000).

In most AGCM water cycle studies, vegetation characteristics are prescribed a priori. These studies ignore changes in vegetation that occur on the seasonal-to-interannual scale, as manifested by variations in phenology (e.g., the density of leaves, as measured by the leaf area index), and on the decadal to century scale, as manifested by species succession and migration. The tide, however, is turning in this regard. Recent studies have examined the impacts of satellite-derived phenology variations on water cycle variations (e.g. Bounoua *et al.*, 2000; Guillevic *et al.*, 2002). More importantly, some groups have begun modeling phenology and species succession explicitly to allow a full consideration of climate-vegetation feedbacks (e.g., Foley *et al.*, 1998).

The modeling of the carbon cycle in conjunction with the water cycle is a natural step in earth system modeling (Sellers *et al.*, 1997). Variations in the carbon cycle could have a first-order impact on the water cycle – aside from the radiative impacts of increased CO_2, plants exposed to higher CO_2 concentrations tend to transpire less. Dynamic vegetation modules, of the type discussed above, provide a means for modeling the carbon cycle (Cox *et al.*, 2000) and will undoubtedly become more common in AGCMs in the years to come. Nitrogen cycle modeling (Dickinson *et al.*, 2002) is a further logical step in the analysis of the climate system.

SUMMARY

The AGCM and its coupled land and ocean models afford a unique means of quantifying and analyzing the global water cycle. The coupled models can produce a wealth of self-consistent global data across timescales ranging from minutes to centuries. The key benefit of these data, of course, is that they generally have no observational counterparts – much of the data relevant to the global water cycle can only be provided by an AGCM. The AGCM can be run in two modes for water cycle studies. In data assimilation mode, model data is merged optimally with existing observational data to produce a dataset superior to either on its own. In climate simulation mode, AGCMs can host sensitivity experiments that illustrate the mechanisms governing the variability of the water cycle and/or quantify the sensitivity of the cycle to human impacts.

Clearly, the chief limitation to AGCM studies of the global water cycle lies in the errors that continue to persist in the simulated climates. These errors will not simply disappear with the advent of greater computational power. Mitigating the effects of the errors will require the development of improved parameterizations for many nonresolvable physical processes.

The AGCM's limitations naturally temper one's belief in the model's interpretation of the global water cycle and its variability. Nevertheless, given the absence of any superior approach, and given the extensive (though qualified) success already achieved by AGCMs in revealing key mechanisms underlying the Earth's water cycle and climate system, climate scientists will continue to employ AGCMs in a wide variety of variability, predictability, and anthropogenic change studies.

Acknowledgments

Figures shown in this article were provided by Julio Bacmeister, Michael Bosilovich, Paul Dirmeyer, Jared Entin, Peter Gleckler, Kingtse Mo, and Alan Robock.

FURTHER READING

Gates W.L. (1992) AMIP: The atmospheric model intercomparison project. *Bulletin of the American Meteorological Society*, **73**, 1962–1970.

Lau K.M., Kim J.H. and Sud Y. (1996) Intercomparison of hydrologic processes in AMIP GCMs. *Bulletin of the American Meteorological Society*, **77**, 2209–2227.

REFERENCES

Beljaars A.C.M., Viterbo P., Miller M.J. and Betts A.K. (1996) The anomalous rainfall over the United States during July 1993: sensitivity to land surface parameterization and soil moisture anomalies. *Monthly Weather Review*, **124**, 362–383.

Betts A.K. and Viterbo P. (2000) Hydrological budgets and surface energy balance of seven subbasins of the Mackenzie river from the ECMWF model. *Journal of Hydrometeorology*, **1**, 47–60.

Bonan G.B. (1995) Land-atmosphere interactions for climate system models – coupling biophysical, biogeochemical,

and ecosystem dynamical processes. *Remote Sensing of Environment*, **51**, 57–73.

Bonan G.B. (1997) Effects of land use on the climate of the United States. *Climatic Change*, **37**, 449–486.

Bosilovich M.G., Schubert S.D. and Walker G.K. (2005) Global changes of the water cycle intensity. *Journal of Climate*, in press.

Bounoua L., Collatz G.J., Los S.O., Sellers P.J., Dazlich D.A., Tucker C.J. and Randall D.A. (2000) Sensitivity of climate to changes in NDVI. *Journal of Climate*, **13**, 2277–2292.

Budyko M.I. (1974) *Climate and Life*, Academic Press: New York, p. 508.

Chase T.N., Pielke R.A. Sr, Kittel T.G.F., Zhao M., Pitman A.J., Running S.W. and Nemani R.R. (2001) Relative climatic effects of landcover change and elevated carbon dioxide combined with aerosols: a comparison of model results and observations. *Journal of Geophysical Research*, **106**, 31685–31691.

Cohen J. and Rind D. (1991) The effect of snow cover on the climate. *Journal of Climate*, **4**, 689–706.

Costa M.H. and Foley J.A. (2000) Combined effects of deforestation and doubled atmospheric CO2 concentrations on the climate of Amazonia. *Journal of Climate*, **13**, 18–34.

Cox P.M., Betts R.A., Jones C.D., Spall S.A. and Totterdell I.J. (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, **408**, 184–187.

Curtis S., Adler R., Huffmann G., Nelkin E. and Bolvin D. (2001) Evolution of tropical and extratropical precipitation anomalies during the 1997–1999 ENSO cycle. *International Journal of Climatology*, **21**, 961–971.

Delworth T.L. and Manabe S. (1988) The influence of potential evaporation on the variabilities of simulated soil wetness and climate. *Journal of Climate*, **1**, 523–547.

Delworth T.L. and Manabe S. (1989) The influence of soil wetness on near-surface atmospheric variability. *Journal of Climate*, **2**, 1447–1462.

Dickinson R.E., Henderson-Sellers A., Kennedy P.J. and Wilson M.F. (1986) *Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Center Model*, Technical Note TN-275 + STR, National Center for Atmospheric Research, Boulder, p. 69.

Dickinson R.E., Berry J.A., Bonan G.B., Collatz G.J., Field C.B., Fung I.Y., Goulden M., Hoffmann W.A., Jackson R.B., Myneni R., Sellers P.J. and Shaikh M. (2002) Nitrogen controls on climate model evapotranspiration. *Journal of Climate*, **15**, 278–295.

Dirmeyer P.A. (1998) Land-sea geometry and its effect on monsoon circulations. *Journal of Geophysical Research*, **103**, 11555–11572.

Dirmeyer P.A. (2000) Using a global soil wetness dataset to improve seasonal climate simulation. *Journal of Climate*, **13**, 2900–2922.

Dirmeyer P.A. (2001) An evaluation of the strength of land-atmosphere coupling. *Journal of Hydrometeorology*, **2**, 329–344.

Dirmeyer P.A., Dolman A.J. and Sato N. (1999) The pilot phase of the global soil wetness project. *Bulletin of the American Meteorological Society*, **80**, 851–878.

Douville H. (2003) Assessing the influence of soil moisture on seasonal climate variability with AGCMs. *Journal of Hydrometeorology*, **4**, 1044–1066.

Douville H. and Chauvin F. (2000) Relevance of soil moisture for seasonal climate predictions: a preliminary study. *Climate Dynamics*, **16**, 719–736.

Entekhabi D., *et al.* Coauthors (1999) An agenda for land surface hydrology research and a call for the second international hydrological decade. *Bulletin of the American Meteorological Society*, **80**, 2043–2058.

Entin J.K., Robock A., Vinnikov K.Y., Zabelin V., Liu S., Namkhai A. and Adyasuren T. (1999) Evaluation of global soil wetness project soil-moisture simulations. *Journal of the Meteorological Society of Japan*, **77**, 183–198.

Famiglietti J.S. and Wood E.F. (1994) Multiscale modeling of spatially variable water and energy balance processes. *Water Resources Research*, **30**, 3061–3078.

Fennessy M.J. and Shukla J. (1999) Impact of initial soil wetness on seasonal atmospheric prediction. *Journal of Climate*, **12**, 3167–3180.

Foley J.A., Levis S., Prentice I.C., Pollard D. and Thompson S.L. (1998) Coupling dynamic models of climate and vegetation. *Global Change Biology*, **4**, 561–579.

Gedney N., Cox P.M., Douville H., Polcher J. and Valdes P.J. (2000) Characterizing GCM land surface schemes to understand their responses to climate change. *Journal of Climate*, **13**, 3066–3079.

Gedney N. and Valdes P.J. (2000) The effect of Amazonian deforestation on the northern hemisphere circulation and climate. *Geophysical Research Letters*, **27**, 3053–3056.

Guillevic P., Koster R.D., Suarez M.J., Bounoua L., Collatz G.J., Los S.O. and Mahanama S.P.P. (2002) Influence of the interannual variability of vegetation on the surface energy balance – a global sensitivity study. *Journal of Hydrometeorology*, **3**, 617–629.

Hahmann A.N. and Dickinson R.E. (1997) RCCM2-BATS model over tropical South America: applications to tropical deforestation. *Journal of Climate*, **10**, 1944–1963.

Henderson-Sellers A., Dickinson R.E., Durbidge T.B., Kennedy P.J., McGuffie K. and Pitman A.J. (1993) Tropical deforestation, modeling local-scale to regional-scale climate change. *Journal of Geophysical Research*, **98**, 7289–7315.

Hoffmann G., Jouzel J. and Masson V. (2000) Stable water isotopes in atmospheric general circulation models. *Hydrological Processes*, **14**, 1385–1406.

Hong S. and Kalnay E. (2000) Role of sea surface temperature and soil-moisture feedback in the 1998 Oklahoma-Texas drought. *Nature*, **408**, 842–844.

Huffmann G.J., Adler R.F., Arkin P.A., Chang A., Ferraro R., Gruber A., Janowiak J., Joyce R.J., McNab A., Rudolf B., Schneider U., *et al.* (1997) The global precipitation climatology project (GPCP) combined precipitation data set. *Bulletin of the American Meteorological Society*, **78**, 5–20.

Joussaume S., Sadourny R. and Jouzel J. (1984) A general circulation model of water isotope cycles in the atmosphere. *Nature*, **311**, 24–29.

Joussaume S., Sadourny R. and Vignal C. (1986) Origin of precipitating water in a numerical simulation of the July climate. *Ocean-Air Interactions*, **1**, 43–56.

Joussaume S., Taylor K.E., Braconnot P., Mitchell J.F.B., Kutzbach J.E., Harrison S.P., Prentice I.C., Broccoli A.J., Abe-Ouchi A., Bartlein P.J., Bonfils C. and 25 others (1999) Monsoon changes for 6000 years ago: results of 18 simulations from the Paleoclimate modeling intercomparison project (PMIP). *Geophysical Research Letters*, **26**, 859–862.

Jouzel J., Russell G., Suozzo R., Koster R., White J. and Broecker W. (1987) Simulations of the HDO and H_2 ^{18}O atmospheric cycles using the NASA GISS general circulation model: the seasonal cycle for present-day conditions. *Journal of Geophysical Research*, **92**, 14739–14760.

Kang I.S., Jin K., Wang B., Lau K.M., Shukla J., Krishnamurthy V., Schubert S.D., Wailser D.E., Stern W.F., Kitoh A., Meehl G.A., Kanamitsu M., Galin V.Y., Satyan V., Park C.K., Liu Y. (2002) Intercomparison of the climatological variations of Asian summer monsoon precipitation simulated by 10 GCMs. *Climate Dynamics*, **19**, 383–395.

Koster R.D., de Valpine D.P. and Jouzel J. (1993) Continental water recycling and H_2 ^{18}O concentrations. *Geophysical Research Letters*, **20**, 2215–2218.

Koster R.D., Dirmeyer P.A., Hahmann A.N., Ijpelaar R., Tyahla L., Cox P. and Suarez M.J. (2002) Comparing the degree of land-atmosphere interaction in four atmospheric general circulation models. *Journal of Hydrometeorology*, **3**, 363–375.

Koster R.D. and Milly P.C.D. (1997) The interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models. *Journal of Climate*, **10**, 1578–1591.

Koster R.D. and Suarez M.J. (1999) A simple framework for examining the interannual variability of land surface moisture fluxes. *Journal of Climate*, **12**, 1911–1917.

Koster R.D. and Suarez M.J. (2003) Impact of land surface initialization on seasonal precipitation and temperature prediction. *Journal of Hydrometeorology*, **4**, 408–423.

Koster R.D., Suarez M.J., Ducharne A., Stieglitz M. and Kumar P. (2000b) A catchment-based approach to modeling land surface processes in a general circulation model, 1, Model structure. *Journal of Geophysical Research*, **105**, 24809–24822.

Koster R.D., Suarez M.J. and Heiser M. (2000a) Variance and predictability of precipitation at seasonal-to-interannual timescales. *Journal of Hydrometeorology*, **1**, 26–46.

Koster R.D., Suarez M.J., Higgins R.W. and van den Dool H.M. (2003) Observational evidence that soil-moisture variations affect precipitation. *Geophysical Research Letters*, **30**, 1241, doi10.1029/2002GL016571.

Koster R.D., Suarez M.J., Liu P., Jambor U., Berg A., Kistler M., Reichle R., Rodell M. and Famiglietti J. (2004) Realistic initialization of land surface states: impacts on subseasonal forecast skill. *Journal of Hydrometeorology*, **5**, 1049–1063.

Kumar A. and Hoerling M.P. (1995) Prospects and limitations of seasonal atmospheric GCM predictions. *Bulletin of the American Meteorological Society*, **76**, 335–345.

Lau K.-M. and Bua W. (1998) Mechanisms of monsoon-southern oscillation coupling: insights from GCM experiments. *Climate Dynamics*, **14**, 759–779.

Laval K., Raghava R., Polcher J., Sadourny R. and Forichon M. (1996) Simulations of the 1987 and 1988 Indian monsoons using the LMD GCM. *Journal of Climate*, **9**, 3357–3371.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically-based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, **99**, 14415–14428.

Manabe S. (1969) Climate and the ocean circulation, I: The atmospheric circulation and the hydrology of the earth's surface. *Monthly Weather Review*, **97**, 739–774.

Milly P.C.D. and Dunne K.A. (1994) Sensitivity of the global water cycle to the water-holding capacity of land. *Journal of Climate*, **7**, 506–526.

Oglesby R.J. and Erickson D.J. III (1989) Soil moisture and the persistence of North American drought. *Journal of Climate*, **2**, 1362–1380.

Oglesby R.J., Marshall S., Erickson D.J. III, Roads J.O. and Robertson F.R. (2002) Thresholds in atmosphere-soil moisture interactions: results from climate model studies. *Journal of Geophysical Research*, **107**, 4224, doi:10.1029/2001JD001045.

Oki T., Nishimura T. and Dirmeyer P. (1999) Assessment of annual runoff from land surface models using total runoff integrating pathways (TRIP). *Journal of the Meteorological Society of Japan*, **77**, 235–255.

Pitman A.J. and McAvaney B.J. (2002) The role of surface energy balance complexity in land surface models' sensitivity to increasing carbon dioxide. *Climate Dynamics*, **19**, 609–618.

Pitman A.J. and Zhao M. (2000) The relative impact of observed change in land cover and carbon dioxide as simulated by a climate model. *Geophysical Research Letters*, **27**, 1267–1270.

Polcher J. and Laval K. (1994) A statistical study of the regional impact of deforestation on climate in the LMD GCM. *Climate Dynamics*, **10**, 205–219.

Rind D. (1982) The influence of ground moisture conditions in North America on summer climate as modeled in the GISS GCM. *Monthly Weather Review*, **110**, 1487–1494.

Roads J., Kanamitsu M. and Stewart R. (2002) CSE water and energy budgets in the NCEP-DOE Reanalysis II. *Journal of Hydrometeorology*, **3**, 227–248.

Roads J.O., Marshall S., Oglesby R. and Chen S.C. (1996) Sensitivity of the CCM1 hydrologic cycle to CO2. *Journal of Geophysical Research*, **101**, 7321–7339.

Robock A., Vinnikov K.Y., Srinivasan G., Entin J.K., Hollinger S.E., Speranskaya N.A., Liu S. and Namkhai A. (2000) The global soil-moisture data bank. *Bulletin of the American Meteorological Society*, **81**, 1281–1299.

Russell G.L. and Miller J.R. (1990) Global river runoff calculated from a global atmospheric general circulation model. *Journal of Hydrology*, **117**, 241–254.

Schar C., Luthi D., Beyerle U. and Heise E. (1999) The soil-precipitation feedback: a process study with a regional climate model. *Journal of Climate*, **12**, 722–741.

Schlosser C.A. and Milly P.C.D. (2002) A model-based investigation of soil-moisture predictability and associated climate predictability. *Journal of Hydrometeorology*, **3**, 483–501.

Schlosser C.A. and Mocko D.M. (2003) The impact of snow conditions in spring dynamical seasonal predictions. *Journal of Geophysical Research*, **108**, Art. No. 8616.

Sellers P.J., Dickinson R.E., Randall D.A., Betts A.K., Hall F.G., Berry J.A., Collatz G.J., Denning A.S., Mooney H.A. Nobre C.A., *et al.* (1997) Modeling the exchanges of energy,

water and carbon between continents and the atmosphere. *Science*, **275**, 502–509.

Sellers P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A simple biosphere model (SiB) for use within general circulation models. *Journal of the Atmospheric Sciences*, **43**, 505–531.

Sellers P.J., Randall D.A., Collatz G.J., Berry J.A., Field C.B., Dazlich D.A., Zhang C., Collelo G.D. and Bounoua L. (1996) A revised land surface parameterization (SiB2) for atmospheric GCMs, 1, model formulation. *Journal of Climate*, **9**, 676–705.

Shukla J. (1998) Predictability in the midst of chaos: a scientific basis for climate forecasting. *Science*, **282**, 728–731.

Soden B.J. (2000) The sensitivity of the tropical hydrological cycle to ENSO. *Journal of Climate*, **13**, 538–549.

Stieglitz M., Rind D., Famiglietti J. and Rosenzweig C. (1997) An efficient approach to modeling the topographic control of surface hydrology for regional and global climate modeling. *Journal of Climate*, **10**, 118–137.

Sud Y.C., Walker G.K., Kim J.H., Liston G.E., Sellers P.J. and Lau W.K.M. (1996) Biogeophysical consequences of a tropical deforestation scenario: a GCM simulation study. *Journal of Climate*, **9**, 3225–3247.

Trenberth K.E., Branstator G.W., Karoly D., Kumar A., Lau N.C. and Ropelewski C. (1998) Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *Journal of Geophysical Research*, **103**, 14291–14324.

Viterbo P. and Betts A.K. (1999) Impact of the ECMWF reanalysis soil water on forecasts of the July 1993 Mississippi flood. *Journal of Geophysical Research*, **104**, 19361–19366.

Walsh J.E. and Ross B. (1988) Sensitivity of 30-day dynamical forecasts to continental snow cover. *Journal of Climate*, **1**, 739–756.

Wang W. and Kumar A. (1998) A GCM assessment of atmospheric seasonal predictability associated with soil-moisture anomalies over North America. *Journal of Geophysical Research*, **103**, 28637–28646.

Werth D. and Avissar R. (2002) The local and global effects of Amazon deforestation. *Journal of Geophysical Research*, **107**, Art. No. 8087.

Xie P.P. and Arkin P.A. (1998) Global monthly precipitation estimates from satellite-observed outgoing longwave radiation. *Journal of Climate*, **11**, 137–164.

Xue Y.K. and Shukla J. (1993) The influence of land-surface properties on sahel climate, 1, desertification. *Journal of Climate*, **6**, 2232–2245.

Zhang H., Henderson-Sellers A. and McGuffie K. (2001) The compounding effects of tropical deforestation and greenhouse warming on climate. *Climatic Change*, **49**, 309–338.

179: Modeling of the Global Water Cycle – Analytical Models

YONGQIANG LIU[1] AND RONI AVISSAR[2]

[1] *Forestry Sciences Laboratory, USDA Forest Service, Athens, GA, US*
[2] *Department of Civil and Environmental Engineering, Duke University, Durham, NC, US*

Both numerical and analytical models of coupled atmosphere and its underlying ground components (land, ocean, ice) are useful tools for modeling the global and regional water cycle. Unlike complex three-dimensional climate models, which need very large computing resources and involve a large number of complicated interactions often difficult to interpret, analytical models are able to provide more direct and intuitive figures of variability and processes in a highly simplified system. They can be a good and efficient alternative modeling tool, especially for studying continental water cycle. This article describes the analytical models developed based on the soil and atmospheric water and energy conservation equations. A fourth-order model is used to illustrate the perturbation equation, solutions, and physical interpretation. Our understanding of some water cycle variability issues, including timescale, persistence, and major physical parameters and processes, obtained from the analytical models is presented.

INTRODUCTION

One of the major improvements in our understanding of the global and regional water cycle during the past three decades was the realization of the importance of soil memory (especially soil moisture) to seasonal and interannual variability of continental water cycle. Atmospheric manifestations of variability in the hydrological cycle occur at multiple timescales, including cloud and precipitation events related to synoptic systems (daily and weekly), floods and droughts (seasonal), the Southern Oscillation (interannual), and climate shifts (decadal). Because an anomaly signal in the atmosphere alone (i.e. isolated from the Earth's surface) can only last for a period of about two weeks, variability of the atmospheric hydrological processes at seasonal or longer scales has to be related in some way to the Earth's surface, which has longer memory.

Soil moisture controls water and energy exchanges by providing available water for evapotranspiration and by determining the partition of radiative energy absorbed on the ground surface into sensible and latent heat fluxes (Avissar, 1995). Therefore, anomalies in soil moisture can result in significant changes in atmospheric hydrological and thermal processes by land–atmospheric interactions (e.g. Mintz, 1984; Avissar and Verstraete, 1990; Betts *et al.*, 1996). Furthermore, with the capacity of retaining anomalous signals over months to seasons (e.g. Delworth and Manabe, 1988; Vinnikov *et al.*, 1996), soil moisture can contribute to long-term atmospheric variability over land by passing its relatively slow anomalous signals to the atmospheric hydrological processes.

The importance of soil moisture in the continental water cycle indicates a fundamental difference with the oceanic water cycle, in which the surface fluxes and long-term anomalies are mostly affected by sea-surface temperature (Cane, 1992). This, of course, is due to the fact that the amount of water available for evaporation is unlimited in the ocean, and its thermal capacity is very large.

Three-dimensional general circulation models (GCMs) and regional climate models (RCMs) have been very useful tools for studying soil-moisture variability and its interactions with atmospheric processes. Coupled with detailed land-surface parameterizations (e.g. Dickinson *et al.*, 1993; Sellers *et al.*, 1986; Xue *et al.*, 1991), these models are able to simulate long-term variations of atmospheric and

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

soil states, hydrological and thermal processes, and water and energy fluxes at the ground surface. The simulation outputs are used to analyze the water cycle variability and its causes (e.g. Delworth and Manabe, 1989; Bonan, 1994; Koster and Suarez, 1995). Experiments of model's response to initial soil-moisture anomalies provide evidence for the possible roles of soil moisture in onset and/or intensification of floods and droughts (e.g. Giorgi *et al.*, 1996) and local water cycle processes (Avissar and Liu, 1996), and improvement of predictability of the water cycle variability (e.g. Schlosser and Milly, 2002; Dirmeyer, 2003).

Considering that three-dimensional climate models need very large computing resources and involve a large number of complicated interactions often difficult to interpret, analytical models can be a good and efficient alternative to study the mechanisms underlying the global water cycle, especially in continental areas. Many analytical models have been developed, based on soil-water balance equation (e.g. Delworth and Manabe, 1988; Rodriguez-Iturbe *et al.*, 1991; Entekhabi *et al.*, 1992; Huang *et al.*, 1996), complete water and energy balance equations of the land-atmosphere system (e.g. Brubaker and Entekhabi, 1996; Liu and Avissar, 1999b), or something in between (e.g. Liu *et al.*, 1992). Analytical models including interactions between soil moisture and the atmospheric planetary boundary layer (PBL) processes also have been developed (e.g. Eltahir, 1998; Findell and Eltahir, 2003).

Analytical models for the land-atmosphere water and energy system have varying levels of complexity. The most basic model is a first-order system composed of a conservation equation for water or heat of the soil or the atmosphere. The most important first-order system is the soil water conservation, which describes the internal variability of soil water and the effects of external processes. The most complete system is the forth-order one, which describes internal variability of each of the four system components and interactions between them. Liu and Avissar (1999b) discussed a forth-order system and compared it with various lower-order systems.

This article describes analytical models of the land-atmosphere energy and water balance and their applications to studying continental water cycle variability. The models are first introduced in the Section "The models", with a focus on the system described in LA99. The solutions and dependence on the water exchange are presented in the Sections "Solutions and Control of seasonal scale by water exchange", respectively. Applications of the models to a number of issues, including water cycle timescales, persistence, and physical mechanisms, are described in the Section "Understanding global water cycle using analytical models". A summary is given in the final section.

THE MODELS

Framework

Figure 1 shows a schematic representation of the continental water and energy cycle system. The atmospheric component is assumed to consist of an air column of height h_a, and the soil component consists of an active layer and a sublayer. Because the soil depths at which thermal and hydrological processes are active are not necessarily identical, they are differentiated between a thermally active layer of depth d_T and a hydrologically active layer of depth d_w. To account for possible diffusion of heat and moisture at the bottom of the active soil layer, soil sublayers of depth d_{To} and d_{wo} are defined for thermal and hydrological processes, respectively.

The atmosphere is characterized by its temperature, T_a, and its specific humidity, q_a. Similarly, the soil is characterized by its temperature, T_g, and its volumetric soil-water content, W_g, for the active layer, and by a constant temperature, T_{go}, and a constant volumetric soil-water content, W_{go}, for the sublayer. The four state variables are determined by the energy and water conservation equations of the soil and atmosphere, and interact with each other through fluxes of radiation, sensible heat flux, latent heat flux (evaporation), and precipitation. Exchanges through diffusion and runoff also affect temperature and moisture of the soil active layer.

Basic Equations

The fourth-order continental water and energy cycle (Liu and Avissar, 1999b, hereafter referred to as LA99) consists of the following set of heat and water conservation equations:

$$C_a \frac{dT_a}{dt} = R_a + H + LP \tag{1}$$

$$M_a \frac{dq_a}{dt} = E - P \tag{2}$$

$$C_g \frac{dT_g}{dt} = R_s - H - LE + K_T(T_{go} - T_g) \tag{3}$$

$$M_g \frac{dW_g}{dt} = P - E - F + K_w(W_{go} - W_g) \tag{4}$$

where $K_T = 2D_T C_g/[d_T(d_T + d_{T0})]$ and $K_w = 2D_w M_g/[d_w(d_w + d_{w0})]$ are forcing factors to restore thermal and hydrologic anomalies to the normal states, respectively; $C_a(= \rho_a c_p h_a)$ and $C_g(= \rho_g c_g d_T)$ are the heat capacity of the atmosphere and the soil, respectively (c_p is specific heat of air at constant pressure, c_g is specific heat of the ground, ρ is density, and subscripts a and g indicate atmosphere and ground); $M_a(= \rho_a h_a)$ and $M_g(= \rho_w d_w)$ are masses of the air column and of a column of water of depth d_w per unit

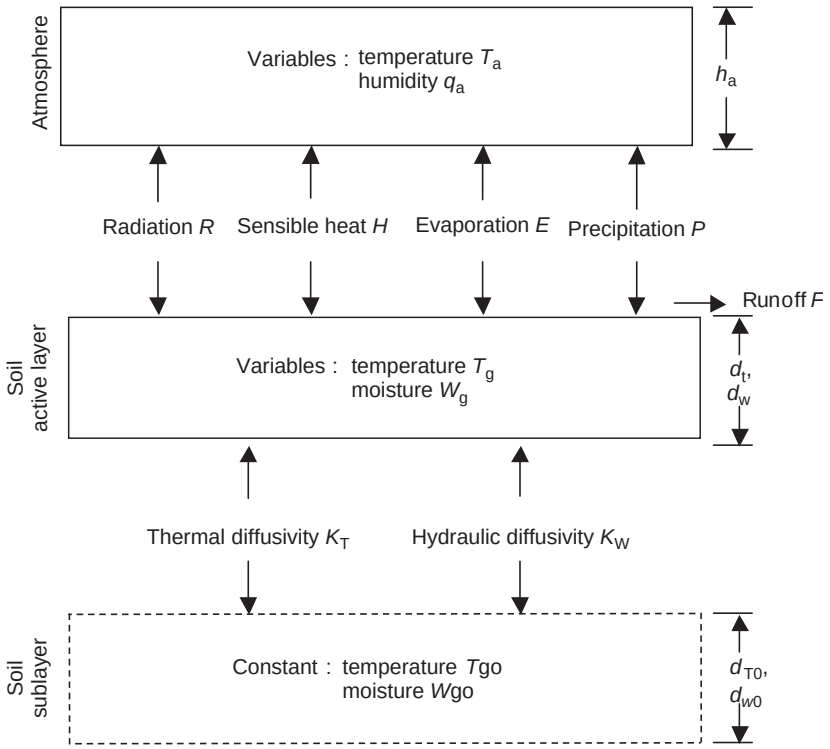


Figure 1 Schematic representation of the continental water and energy cycle system

area; R_a and R_s are the radiation balance of the atmosphere and the land surface; H and E are the sensible heat flux and evaporation on the land surface; P and F are precipitation and runoff; D_T and D_w are the soil thermal and hydraulic diffusivity; and L is the latent heat. It should be mentioned that this is a one-dimensional, vertical representation of land–atmosphere interaction, centered over a land surface, and vegetation and snow are neglected.

Fluxes

The fluxes in equations (1–4) are calculated by the following formulae:

$$R_s = [1 - \alpha_a(1-n) - \alpha_c n - A_a - C_r n]S_0 - \varepsilon\sigma T_g^4 + n\varepsilon\sigma T_a^4 \quad (5)$$

$$R_a = (A_a + C_r n)S_0 + n\varepsilon\sigma T_g^4 - n\varepsilon\sigma T_a^4 - n\varepsilon\sigma T_a^4 - (1-n)G\sigma T_g^4 \quad (6)$$

$$H = H_0(T_g - T_a) \quad (7)$$

$$E = \eta E_p = \eta E_0[q_s(T_g) - q_a] \quad (8)$$

$$P = C_1 W_a \exp(-C_2/R_h)^{C3} \quad (9)$$

$$F = \eta(P - E) \quad (10)$$

The radiative fluxes are calculated following Paltridge (1974). The first term on the right-hand side of equation (5) is the net shortwave radiative flux at the land surface, where S_0 is the incident shortwave radiation at the tropopause, α_a and α_c are the planetary albedo for clear sky and for cloudy atmosphere, A_a is the atmospheric absorption of shortwave radiation, and n is the cloud fraction. The next two terms together represent the net long-wave radiative fluxes. The atmosphere is assumed to be opaque to long-wave radiation at all wavelengths other than in the atmospheric window (7.5–12.5 μm), where absorption is assumed negligible. In this model, long-wave energy transfer occurs only in the atmospheric window and, therefore, it is the only flux that needs to be considered. The constant σ is the Stefan–Boltzmann constant, and ε is the fraction of total blackbody radiation at normal earth temperatures contained within the wavelengths of the atmospheric window. Note that this model represents an averaging condition over time and land area and, therefore, the cloud fraction, n, is never zero.

The first term on the right-hand side of equation (6) is the net shortwave radiative flux of the atmosphere. In this term, the factor C_r accounts explicitly for the extra shortwave absorption in clouds. The second and third terms represent the net long-wave absorption in clouds resulting from exchange of radiant energy with the Earth's surface. In the clear atmosphere, the absorption bands of water vapor become less opaque with increasing height and decreasing water vapor concentration. In effect, the

atmospheric window expands with increasing altitude so that each atmospheric level loses energy directly to space. Since the average temperature lapse to the tropopause is close to adiabatic as a result of mixing, on the average, the temperature at any level bears a fairly constant relation to the surface temperature. Thus, the total long-wave loss from the clear-sky troposphere is likely to be proportional to surface blackbody radiation. This concept is the basis of the last term in equation (6), which contains a simple constant of proportionality G, that was found to be about 0.38 (on average) for a large number of studied typical cases. A detailed discussion of the philosophy behind this scheme, as well as of the values assigned to each of its variables, are given in Paltridge (1974), and are not repeated here for brevity.

In the bulk flux formulae for sensible heat and evaporation (equations 7–8), $H_0 = \rho_a c_p C_{DT} V$ and $E_0 = \rho_a C_{DW} V$; C_{DT} and C_{DW} are the sensible heat and water drag coefficients, respectively, which are strongly dependent upon land surface, and atmospheric dynamic and thermal properties (e.g. vegetation cover and stability); V is the wind speed in the atmospheric surface layer; E_p is the potential evaporation; q_s (T_g) is the saturated air specific humidity at temperature T_g; and $\eta = W_g/W_s$ is the soil wetness, with W_s being the volumetric soil-water content at saturation.

Precipitation is arguably the most important and complex hydrological process in the water cycle and the land–atmosphere interactions system. Both atmospheric dynamics and local water and heat exchanges in the land-atmosphere system can contribute to the essential conditions for the formation of precipitation (i.e. ascending motion of air masses, stratified instability, and water vapor supply). Because the model proposed here does not simulate atmospheric dynamics, precipitation is parameterized with a simple formula, which only accounts for the variances of atmospheric moisture and sensible heat caused by land–atmosphere interactions. The precipitation parameterization (equation 9) is based on the premise that at a given geographic location and a given general circulation pattern, precipitation is mostly related to atmospheric water vapor content, W_a, and the relative humidity, R_h, of the entire air column. This formula is formed by applying dimensional analysis to relations between precipitation and these two factors. C_i ($i = 1, 2, 3$) are empirical constants. In this model, $C_2 = 0.5$ and $C_3 = 4.5$. With its relatively simple form, this parameterization certainly has a very limited capacity in reproducing observed precipitation statistics, which would affect the analysis with the fourth-order model. However, C_1 is derived from the ratio of climatological evaporation to precipitation, namely $r_{EP} = \overline{E}/\overline{P}$, which therefore adds a restraint in estimating average rainfall. Runoff (equation 10) is proportional to the net soil water gain, and to the ratio of soil moisture to its saturation value. The linear relation between runoff and both the net water gain and the ratio is valid mostly for calculation of runoff process at seasonal or longer scales.

Perturbation Equation

The perturbation approach (Holton, 1979) is used to linearize the equations of the model. Accordingly, any variable ϕ is separated into a mean $\bar{\phi}$ and a perturbation ϕ'. Assuming that the perturbations are small as compared to the means, any terms including the product of two or more perturbations are much smaller than a term that includes only one or no perturbation. Thus, for instance, the radiation emitted by a black body at a temperature T is given by $\sigma T^4 = \sigma(\overline{T} + T')^4 \approx \sigma(\overline{T}^4 + 4\overline{T}^3 T')$.

The perturbation equations for the sensible heat and evaporation fluxes are: $H' = H_0(T'_g - T'_a)$ and $E' = \bar{\eta} E_0(\Delta_g T'_g - q'_a)$, where $\Delta_g = \mathrm{d}q_s/\mathrm{d}T|_{T=Tg}$. To develop a perturbation equation for precipitation, the atmospheric water content in equation (9) is first substituted with the product of the specific humidity with the mass of the atmosphere, and the relative humidity by the ratio of the specific humidity to the specific humidity at saturation. Subsequently, we differentiate the expression and assume that the perturbation of a quantity is approximately equal to its differential value, and that the average is expressed by the nondifferential value. As a result, $P'/\overline{P} = P_1 q'_a - P_2 T'_a$, where $P_1 = [1 + C_3(C_2/\overline{R}_h)^{C_3}]/\bar{q}_a$, $P_2 = C_3(C_2/\overline{R}_h)^{C_3}\Delta_a/\bar{q}_{as}$, and $\Delta_a = \mathrm{d}q_s/\mathrm{d}T|_{T=Ta}$. The perturbation of cloud amount is related to the precipitation perturbation as follows: $n'/\bar{n} = P'/\overline{P}$. The perturbation of the radiative fluxes is given by: $Rs' = (\varepsilon\bar{n}\Delta_{La} + R_{s1}P_2\bar{n})T'_a - \varepsilon\Delta_{Lg}T'_g - R_{s1}P_1\bar{n}q'_a$ and $R'_a = -(R_{a1} + R_{a3}P_2\bar{n})T'_a + R_{a2}T'_g + R_{a3}P_1\bar{n}q'_a$, where $R_{s1} = (\alpha_c - \alpha_a - C_r)S_0 - \varepsilon\sigma\overline{T}_a^4$, $R_{a1} = (\varepsilon + \varepsilon')\bar{n}\Delta_{La}$, $R_{a2} = [(\varepsilon + G)\bar{n} - G]\bar{n}\Delta_{Lg}$, $R_{a3} = C_r S_0 - (\varepsilon + \varepsilon')\sigma\overline{T}_a^4 + (\varepsilon + G)\sigma\overline{T}_g^4$, and $\Delta_{Li} = \mathrm{d}(\sigma\overline{T}_i^4)/\mathrm{d}\overline{T}_i$, $(i = a, g)$.

Using these flux perturbations, the heat and water conservation equations (equations 1–4) can be written:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{Y} \tag{11}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_i) = (T'_a/\overline{T}_a,\ q'_a/\bar{q}_a,\ T'_g/\overline{T}_g,\ W'_g/\overline{W}_g)$ and $\boldsymbol{C} = (c_{ij}) = \boldsymbol{I}_1\boldsymbol{A}\boldsymbol{I}_2$. $\boldsymbol{I}_1$ is a diagonal matrix with elements $[(C_a\overline{T}_a)^{-1}, (M_a\bar{q}_a)^{-1}, (C_g\overline{T}_g)^{-1}, (M_g\overline{W}_g)^{-1}]$, $\boldsymbol{I}_2$ is a diagonal matrix with elements $[\overline{T}_a^{-1},\ \bar{q}_a^{-1},\ \overline{T}_g^{-1},\ \overline{W}_g^{-1}]$, and $A = (a_{ij})$ is a coefficient matrix. This equation represents a set of linear homogeneous differential equations with constant coefficients, whose linearly independent solutions can be written as

$$\boldsymbol{Y}_i(t) = \sum_{j=1}^{4} P_{ij}(t)e^{\lambda_j t} \tag{12}$$

where $\lambda_j = (\lambda_r)_j + i(\lambda_i)_j (j = 1, 4)$ are roots of the characteristic equation of equation (12) given by

$$\sum_{k=0}^{4} d_k \lambda^{4-k} = 0 \tag{13}$$

The parameter λ_r is the perturbation growth rate, and the reciprocal of its absolute value is the e-folding time. If $\lambda_r < 0$, the e-folding time is also called *damping time*. In equation (12), $P_{ij}(t)$ is a polynomial whose order is equivalent to the number of equal roots. In equation (13), d_k are constant coefficients, with $d_0 = 1$ and $d_1 = -\sum_{k=0}^{4} a_{kk}$.

Orders of Models

Equation (11) is a fourth-order analytical model of the continental water and energy cycle system. It contains all interactions (couplings) between the land and the atmosphere, and between hydrological and thermal processes. By assuming no disturbance with soil moisture in the fourth-order model, a third-order model can be obtained. There are no interactions between soil moisture and other model variables. Similarly, three other third-order models can be obtained by assuming no disturbance with soil temperature, air humidity, or air temperature in the fourth-order model. Four second-order models can be obtained by assuming no disturbance with any two of the four system variables. Finally, by assuming no interactions among the four system variables, the fourth-order model becomes four independent first-order models, each of which contains only one of the four variables. In these cases, the variation of the disturbance in the models is caused by self-feedback.

Statistical-Dynamical Models

Delworth and Manabe (1988) described a first-order model of soil water balance equation with a stochastic forcing. Precipitation usually has a much faster pace than soil moisture, suggesting that variation of precipitation could be expressed as "white noise", $z(t)$. Soil moisture can be expressed as "red noise", $y(t)$, which is the output of the following first-order Markov process:

$$\frac{dy(t)}{dt} = -\lambda y(t) + z(t) \tag{14}$$

The stochastic process $z(t)$ can be simulated with observed precipitation series. The solution of this equation indicates the role of the external forcing in variability of soil moisture.

In another first-order model of soil water balance described by Rodriguez-Iturbe *et al.* (1991), precipitation is expressed as

$$P(w) = P_a[1 + f(w, \Omega)] \tag{15}$$

where P_a is the advective precipitation resulting from the atmospheric horizontal transport of water vapor from outside, and Ω is a Gaussian noise of atmospheric processes related to the internal precipitation resulting from the vertical transport of water vapor from the evaporation on the ground. P is related to soil moisture w.

The fourth-order model described by Brubaker and Entekhabi (1996) consists of four stochastic ordinary differential equations in soil moisture, soil temperature, mixed-layer humidity, and mixed-layer potential temperature. The variations of the system are determined by deterministic steady forcing and white-noise perturbation processes. The solution is a physically consistent joint probability distribution.

SOLUTIONS

Table 1 lists the various atmospheric and soil conditions needed to calculate the elements of the matrix C. These values are assumed to be representative of annual conditions under current climate over land. The values for the radiative flux are adopted from Paltridge (1974). A few values of parameters related to other processes are taken from the Biosphere-Atmosphere Transfer Scheme (BATS, Dickinson *et al.*, 1993) and from the results of the simulation with the Community Climate Model (CCM2) (`http://www.cgd.ucar.edu/cms/ccm.html`) coupled with BATS (`http://www.atmo.arizona.edu/faculty/research/bats/batsmain.html`) performed by Bonan (1994). For instance, in BATS, the active soil layer is 1 m deep for 11 of the 16 land-cover types, and 1.5–2 m deep for the other types. The soil sublayer is 9 m deep for all land-cover types. Here these two layers are assumed to be 1 m and 10 m deep, respectively. The ratio of evaporation to precipitation (r_{EP}) in the CCM2-BATS simulation is 0.60 (Both evaporation and precipitation are spatially and temporally averaged before taking the ratio). The corresponding land surface sensible and latent heat fluxes in the CCM2-BATS simulation are about 33 and 61 W m^{-2}, respectively. On the basis of these values of the fluxes, as well as climatological values of T_a, T_g, q_a, and η provided in Table 1, one can obtain H_0 and E_0 from equations (7–8), respectively.

Equation (13) has four solutions. The number of solutions is reduced to three and two for the third- and second-order models, respectively. The solutions, given in Table 2, measure damping times or scales. Note that each solution (scale) represents a perturbation mode of the system, rather than a timescale of a variable. The damping times with the fourth-order model are of the orders of one day, one week, two months, and eight months. Thus, hereafter, they will be referred to as daily, weekly, monthly, and seasonal-scale processes, respectively. The monthly and seasonal scales, which represent long-term variations, will be referred to as long-term scales. The damping time points out how fast

Table 1 Parameters used in the continental water and energy cycle system

Parameter	Meaning	Unit	Mean	Range in FAST[a]
T_a	Air temperature	K	255.0	250.0–260.0
q_a	Air specific humidity	g/kg	5.0	3.0–7.0
T_g	Soil temperature	K	285.0	280.0–290.0
W_g	Soil volumetric water content	%	250.0	10.0–40.0
S_0	Solar radiation at tropopause	$W\,m^{-2}$	330.0	–
r_{EP}	Ratio of evaporation to precipitation	%	60.0	45.0–75.0
n	Cloud fraction	%	50.0	35.0–65.0
d_T	Depth of thermally active soil layer	m	10.0	7.5–12.5
d_{T0}	Depth of thermal soil sublayer	m	10.0	–
d_w	Depth of hydrologically active soil layer	m	1.0	0.75–1.25
d_{w0}	Depth of hydrological sublayer	m	10.0	–
h_a	Height of air column	km	10.0	9.0–11.0
D_T	Soil thermal diffusivity	$10^{-7}\,m^2\,s^{-1}$	5.0	2.5–7.5
D_w	Soil hydraulic diffusivity	$10^{-8}\,m^2\,s^{-1}$	5.0	2.5–7.5
C_g	Soil heat capacity	$JK^{-1}\,m^{-3}$	1.0	0.75–1.25
W_s	Saturated soil volumetric water content	%	40.0	30.0–50.0
H_0	Turbulent sensible heat exchange index	$W\,m^{-2}\,K^{-1}$	1.0	0.5–1.5
E_0	Turbulent water vapor exchange index	$10^{-3}\,kg\,m^{-2}\,s^{-1}$	4.5	3.5–5.5
ρ_a	Air density	$kg\,m^{-3}$	0.75	–

Source: (Liu and Avissar, 1999b. © 1999 American Meteorological Society).
[a]FAST stands for Fourier amplitude sensitivity test, see the Section on ''Control of seasonal scale by water exchange''.

Table 2 Damping times of the fourth-, third-, and second-order models (in days)

Model order	Variables	Scales 1	2	3	4
4th	T_a, q_a, T_g, W_g	231.5	57.5	5.7	1.2
3rd	q_a, T_g, W_g	229.4	7.7	1.2	
	T_a, T_g, W_g	195.6	18.7	1.3	
	T_a, q_a, W_g	235.6	29.1	5.8	
	T_a, q_a, T_g	58.1	5.7	1.2	
2nd	T_a, q_a	29.6	5.9		
	T_g, W_g	188.2	1.3		
	T_a, T_g	18.7	1.3		
	q_a, W_g	233.0	6.6		

Source: (Liu and Avissar, 1999b. © 1999 American Meteorological Society).

disturbances decay with time. Consequently, it provides essential information for understanding at which climatic scale a particular perturbation in the water cycle persists. A damping time of a few days implies that the considered perturbation is relevant to short-term, synoptic-scale processes, but is unlikely to have a significant impact on longer timescales. On the other hand, a damping time of several months suggests a potential impact on seasonal climatic processes.

CONTROL OF SEASONAL SCALE BY WATER EXCHANGE

1. Relative Importance of Model Parameters: The Fourier Amplitude Sensitivity Test (FAST) was used to identify which parameters mostly affect the damping times obtained from this fourth-order continental water and energy cycle model. A complete description of the theory and implementation of FAST and approximations used in computer implementation, mainly following Cukier *et al.* (1973), was given in Collins and Avissar (1994).

The number of parameters in the FAST computer program used in this analysis is limited to 15. The chosen parameters from the fourth-order model are listed in Table 1. Among the parameters which are included in the FAST analysis, the role of solar radiation in the damping times cannot be directly examined by the sensitivity analysis because the latent and sensible heat fluxes have been prescribed separately, without any constraint on the balance between the surface net radiation and the two fluxes (the role of solar radiation will be discussed below using a regular sensitivity analysis technique); d_{T0} and d_{w0} have similar roles to d_T and d_w; and ρ_a has a similar role to h_a. Normal distributions of the 15 input parameters were considered, and their ranges are also given in that table. Figure 2 depicts the partial variance of the damping times, which shows the sensitivity of model output parameters to the variation of individual input parameters in terms of a percentage of the variance. It appears that the longest damping timescale, which is about 9.5 months, is very sensitive to soil moisture, ratio of evaporation to precipitation, soil temperature, and soil moisture at saturation. Each of these parameters contributes more than 10% to the total variance. In addition, drag coefficient of water vapor flux, depth of hydrologically active soil layer, air humidity, and soil hydraulic diffusivity have some impact on the variance. It is interesting to note that all these parameters are related

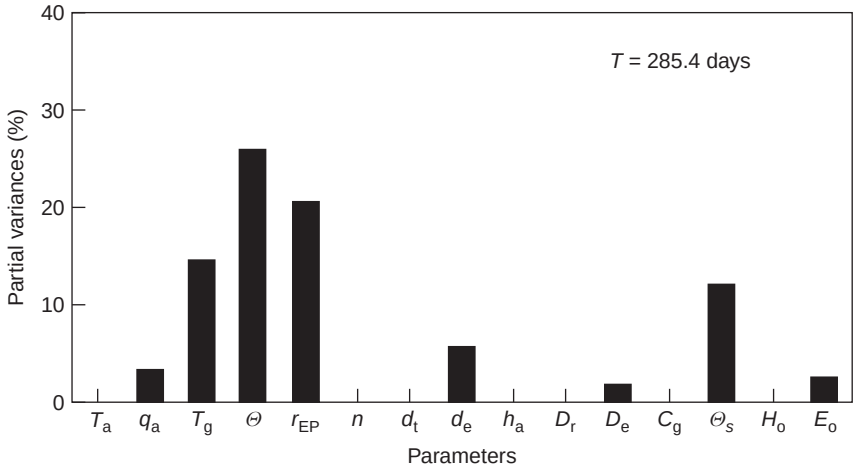


Figure 2 FAST analysis of the fourth-order model outputs for the seasonal scale indicated by the average damping time T obtained from all model runs. Partial variances reflect sensitivity of the damping times to the model inputs (Liu and Avissar, 1999b © 1999 American Meteorological Society)

to soil hydrological processes. Indeed, the ratio of W_g to W_s affects the actual evaporation, E_0, T_g, and q_a affect the potential evaporation, r_{EP} is related to evaporation and precipitation, and D_w and d_w affect the transfer of water between the active soil layer and the soil sublayer. Thus, clearly, the parameters associated with soil hydrology determine the length of the longest process in the water cycle.

2. Sensitivity to Model Parameters: On the basis of the FAST analysis, the effects of a number of the most important parameters on the seasonal damping timescale are examined using a regular sensitivity technique, that is, analyzing a perturbation in the timescale due to a change in a parameter. These parameters can significantly modify the magnitude of the damping timescale. For instance, an increase of E_0 (which is proportional to the turbulence activity in the atmospheric surface layer) from 3.5 to $5.5 \times 10^{-3}\,\mathrm{kg\,m^{-2}\,s^{-1}}$, results in a reduction of the damping time from 283 to 196 days. When r_{EP} and W_s increase, they result in a very significant increase of the damping time. A dry soil ($W_g = 0.1$) has a damping time of 485 days, but a wet soil ($W_g = 0.4$) has a damping time of only 153 days, emphasizing that a moist soil damps a perturbation much faster than a dry soil does.

The evaporation equation (equation 8) consists of two variables, namely, potential evaporation, E_p, and a fraction factor, η. E_p increases with T_g and E_o, and decreases with increasing q_a. Therefore, it can be derived from the relations between these three parameters and the damping time that the damping timescale decreases with increase in potential evaporation. The relations between the damping time and W_g and W_s indicate that the damping time decreases with increasing η. The combined effects of E_p and η result in decrease of the damping time with increase in actual evaporation.

The runoff parameterization (equation 10) consists of two variables as well, that is, η, and the difference between precipitation and actual evaporation $(P - E)$. The latter term is inversely proportional to r_{EP}. Figure 3 shows that the damping time increases from 138 to 393 days with variation of r_{EP} from 0.45 to 0.75 or, equivalently, decreases with $(P - E)$. Therefore, together with the variation of η, the damping time decreases with increase in runoff.

The last term in the right-hand side of equation (4) is the exchange of soil moisture between the active layer and the sublayer, which acts as a force-restore term for soil moisture disturbances. Its intensity is proportional to D_w, and inversely proportional to d_w. The damping time decreases with increasing D_w, and increases with d_w (not shown). Therefore, it decreases with a faster exchange between the two soil layers. d_w is also a measure of the total available water in the active layer. For a given forcing, as expressed by the right-hand side terms in equation (3), the larger d_w, the slower the variation rate of soil moisture, and the longer the damping time. This effect is somewhat similar to that of soil heat capacity in damping the variation of soil temperature. However, evaporation and runoff affect the damping time much more significantly than the exchange between the two soil layers does.

The above results indicate that the smaller the fluxes of water in the continental water and energy cycle system (i.e. evaporation, runoff, and underground diffusion), the longer the damping time is, and, therefore, the more significant persistence is.

By using the Bowen ratio to relate sensible to latent heat flux, the role of solar radiation in the seasonal scale can be examined. From such an analysis, it appears that the damping time decreases from 295 to 206 days as solar radiation increases from 300 to $380\,\mathrm{W\,m^{-2}}$.

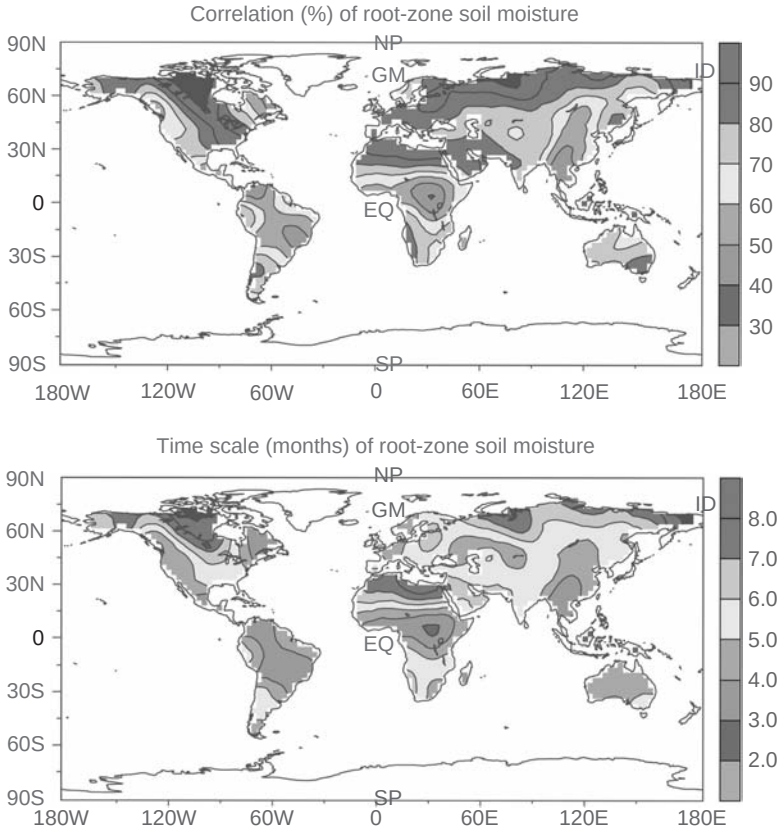


Figure 3 One-month lag autocorrelation coefficients (a) and perturbation timescales (b) of the root-layer soil moisture simulated with CCM2-BATS (Liu and Avissar, 1999a © 1999 American Meteorological Society). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

UNDERSTANDING GLOBAL WATER CYCLE USING ANALYTICAL MODELS

Timescales of Water Cycle

The timescale relevant to this analysis is the period spanned by an anomalous hydrological process. For nonoscillation variability, it is the time period during which anomalies maintain the same sign, and for oscillation variability, it is the half-time period of the oscillations. This analysis aims to identify the major scales at which land-surface processes can contribute to water cycle variability.

1. Interpretation of solutions in terms of timescales: The solutions of the analytical models measure timescales of variability in the water and energy cycle system. This can be illustrated by the first-order model of soil water balance equation (Delworth and Manabe, 1988):

$$\frac{\mathrm{d}w(t)}{\mathrm{d}t} = P - E - F \tag{16}$$

where $w = M_g W_g$. P is an external forcing and F is related to P. E can be calculated by

$$E = \lambda w(t) \tag{17}$$

where $\lambda = E_p / W_s$. Considering the perturbation form of equation (16) and assuming no perturbations in precipitation and runoff, the solution is

$$w(t) = w(t=0)\mathrm{e}^{-\lambda t} \tag{18}$$

indicating that initial soil-moisture perturbation damps exponentially with a rate of λ. The reciprocal of the rate is the e-folding time $= 1/\lambda = W_s/E_p$. This time is used to measure timescale of soil-moisture variability.

2. Scale: Delworth and Manabe (1988) obtained a globally averaged scale of soil-moisture variability of about one to two months based on the simulations with the Geophysical Fluid Dynamics Laboratory (GFDL) GCM. A value of three months was obtained based on the soil-moisture measurements in Russia (Vinnikov *et al.*, 1996) and about two and half months based on the soil-moisture measurements in North China (Entin *et al.*, 2000).

Perturbation strength, defined as the time period over which a perturbation maintains the same sign as an initial anomaly, is presented in Figure 3. It is calculated using a 10-year CCM2-BATS simulation conducted by Bonan

(1994) and it ranges between about two to eight months (Liu and Avissar,1999a).

These studies suggest that the major scales at which land-surface processes could affect the long-term atmospheric water cycle variability are the monthly and seasonal scales. Global climate models have provided some numerical evidence for these analytical results (e.g. Yeh *et al.*, 1984; Liu and Avissar, 1999a). Yeh *et al.* (1984) conducted a numerical experiment of initial soil moisture forcing with a simplified version of the GFDL GCM, and showed that the induced anomalies in evaporation, precipitation, and soil moisture last for three to five months.

3. Timescales in different-order models: Interactions between soil moisture and other variables included in some higher-order models could significantly increase the length of a scale, as indicated by the timescales obtained from the four solutions of equation (17). A seasonal scale as long as eight months is obtained (Table 2).

The damping times in the four third-order systems (Table 2) indicate that the seasonal damping timescale appears only in the third-order systems with disturbance of soil moisture. The longest damping time in the first three systems varies between 196 and 236 days. This emphasizes that the soil moisture feedback, and its interactions with the other variables, are the primary cause for the damping timescale. In the third-order system without physical processes related to soil-moisture disturbances, the maximum damping timescale is only of the order of two months. In addition, among the various interactions, the one between soil moisture and air humidity is the predominant: Excluding this, interaction results in a reduction of the damping time from 232 to 196 days. On the other hand, the exclusion of the interactions involving air and soil temperature has little impact on the damping timescales.

Four second-order systems are considered in Table 2. With System (i), which does not account for soil-variable perturbations, the disturbances sustain for a period of about one month. However, a damping time of about six months is obtained with System (ii), which does not account for air-variable perturbations. Considering these results and those obtained with the third-order systems, it seems that atmospheric disturbances can persist on the seasonal scale only if there are interactions between atmospheric variables and soil moisture. The other two second-order systems further illustrate the importance of the moisture processes. Clearly, thermal disturbances have a short-time impact on the water cycle, while moisture disturbances persist for as long as eight months.

Each of the four first-order models contains only one of the four perturbation variables. Obviously, in these cases, the variation of the disturbances in the systems is caused by self-feedback. The results show that the damping time is about three and a half months for the first-order system that accounts for soil moisture perturbations, and several days to three weeks for the other systems.

It is concluded from the above results that self-feedbacks are the primary factors affecting the timescale of perturbation. Specifically, soil-moisture feedback causes the seasonal-scale perturbation. The interactions between soil moisture and the other variables of the water and energy cycle system (mainly air humidity), cause a significant increase of the seasonal timescale. For atmospheric disturbances to persist at the seasonal scale, interactions between the atmosphere and soil moisture must be considered.

Persistence of Water Cycle

It has long been recognized that precipitation can persist over relatively long time periods (i.e. months to seasons), a feature known as *persistence of atmospheric disturbances*. This property indicates that if an anomaly occurs in a month or season, then there is a good chance for this anomaly to occur in the following month or season as well. Therefore, this property is quite useful for the predictability of the system. Using autocorrelations between adjacent monthly or seasonal rainfall and other atmospheric variables, many studies (e.g. Namias, 1952) demonstrated the existence of such persistence.

Land can contribute to precipitation persistence through its long memory and interaction with the atmosphere. For example, following a dry spring, soil would likely be desiccated during the summer. This would result in a relatively large sensible heat flux injected in the atmosphere from the ground surface, which could perhaps maintain anticyclonic circulations. Under such circumstances, one could expect reduced summer rainfall. A theoretical framework of the role of soil-moisture memory in precipitation statistics was presented by Koster *et al.* (2000), which indicates that a larger soil-moisture memory would lead to a larger correlation between initial precipitation and its subsequent variability.

1. Interpretation of model solutions in terms of persistence: The solutions of the analytical models also measure persistence of anomalies in the water and energy cycle system. Persistence can be estimated using autocorrelation of a variable. For soil moisture, persistence can be related to its timescale by (Delworth and Manabe, 1988)

$$r(\tau) = \mathrm{e}^{-\lambda\tau} \tag{19}$$

where r and τ are autocorrelation and time lag, respectively. This relation indicates that the larger the timescale is, the more significant the persistence of soil-moisture anomalies is.

2. Persistence signals: Persistence of the land-atmosphere system has been investigated with simulated data series.

Using multiyear simulations produced with the GFDL GCM, Delworth and Manabe (1988) analyzed the variability in soil moisture and found statistically significant persistence of soil moisture. Liu and Avissar (1999a) obtained similar results from an analysis based on a simulation with the CCM2-BATS, as shown in Figure 3. The one-month lag autocorrelation coefficients of soil moisture vary between about 30% and 90%, with a global average of 61%. Correlations are statistically significant at a 99.9% confidence level (critical value of 29.6%) almost everywhere. This emphasizes a strong persistence of the simulated soil moisture.

3. Control of persistence by water exchange: The solution of the first-order model (equation 16), λ, is determined by potential evaporation, with a fixed water field capacity. A larger potential evaporation leads to a faster decay rate and thus a less significant persistence of soil-moisture disturbances. Because potential evaporation is larger at low latitudes and in warm seasons, persistence of soil moisture is more significant at high latitudes and in winter (Delworth and Manabe, 1988). In Figure 4, autocorrelation coefficients tend to increase from about 30% in the equatorial regions to 80–90% at high latitudes.

Persistence also depends on climate regime; it is stronger in drier geographic regions (Liu and Avissar, 1999a). For soil moisture, the coefficients are less than 55% in the moist Southeast China, and over 80% in the dry Northwest China. Persistence of the other three variables (soil temperature, air humidity, and temperature) has the same dependence on climate regime as soil moisture. In addition, a global analysis (Liu and Avissar, 1999a) depicts relatively large autocorrelation coefficients in dry northern Africa and a strong contrast between this region and the moist tropics, where autocorrelation coefficients are relatively low.

Physical Mechanisms

1. Feedbacks: The evolution of the disturbances in the fourth-order continental water and energy cycle system is controlled by two mechanisms: self-feedbacks, which are measured by the four diagonal elements in the matrix **C** (equation 11, Table 3), and interactions, which are measured by the 12 other elements of this matrix. All the four diagonal elements are negative. Thus, the average of the roots, which is equal to the average of the four diagonal elements of the matrix **C**, is also negative. A negative value indicates a negative self-feedback, which results in disturbance decay.

Table 3 Values of matrix $\mathbf{C} = (c_{ij})(\times 10^{-6})$

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$i = 1$	−0.641	0.007	0.1171	0.0
$i = 2$	4.494	−1.708	19.79	0.6672
$i = 3$	1.673	0.0456	−9.178	−0.2195
$i = 4$	−0.264	0.9961	−1.113	−0.0883

Source: (Liu and Avissar, 1999b. © 1999 American Meteorological Society).

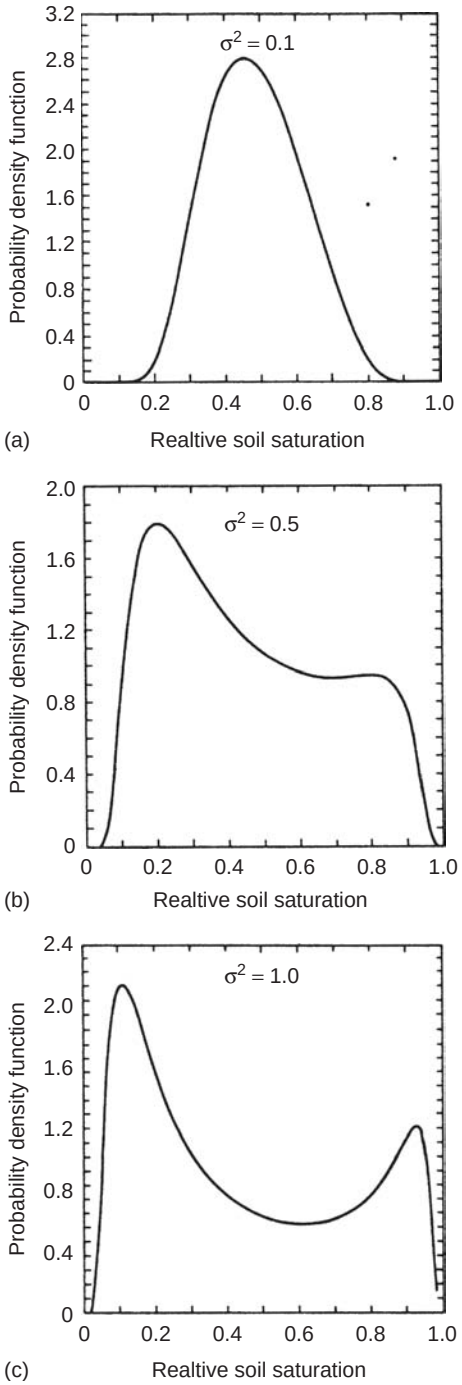


Figure 4 Steady state probability density functions of soil moisture under different atmospheric forcing (Reproduced from Rodriguez-Iturbe, ID *et al.* (1991) by permission of American Geophysical Union)

A physical explanation to the roles of these self-feedbacks can be given by using the perturbation equation (equation 11). For instance, a positive perturbation of air-temperature results in a decrease of the sensible heat released in the atmospheric surface layer from the land surface, and in an increase of the heat lost by long-wave radiation. In addition, one can expect a decrease in precipitation, which results in a reduced release of condensation latent heat, due to the corresponding lower relative humidity. These effects induce a negative air-temperature tendency. Similarly, a positive perturbation of soil temperature results in an increase in heat loss to the atmosphere by sensible heat flux, latent heat flux, and long-wave radiation, and into the soil sublayer by conduction, which leads to a decrease in soil temperature. A positive perturbation of air moisture results in a decrease of the amount of water evaporated from the land surface and, possibly, in an increase in condensation, which removes vapor from the atmosphere. Finally, a positive perturbation of soil moisture results in an increase in evaporation to the atmosphere, in runoff, and in the amount of water lost by percolation to the soil sublayer.

2. Interactions: There are also positive interactions in the system, as indicated by the positive elements of the matrix **C**, which can promote the growth of disturbances. For example, a positive perturbation of soil moisture results in a more humid atmosphere due to a stronger evaporation at the land surface, and a moister atmosphere is likely to produce more precipitation, which in turn increases soil moisture.

The overall evolution of the disturbances (i.e. amplification or decay with time) depends on the relative importance of the feedbacks and the interactions. Because all the roots of the water and energy cycle system (and their mean) are negative (Table 3), disturbances decay with time. Thus, one can expect that self-feedbacks have a predominant impact on this system.

In the fourth-order model developed by Brubaker and Entekhabi (1996), self-feedbacks of soil moisture (through control of infiltration and evaporation) and temperature (through dependence of surface specific humidity) serve to restore each state individually, and interactions between soil moisture and temperature (through soil moisture control of evaporation and the temperature dependence on surface specific humidity) act to reinforce anomaly of the other state.

3. External forcing: The soil water balance equation with the precipitation expressed by equation (15) (Rodriguez-Iturbe *et al.*, 1991) allows us to study variation in soil moisture in response to atmospheric forcing and feedback. The forcing was found to lead to transitions from one stable mode of soil moisture to another. Figure 4 shows soil moisture under different intensities of atmospheric forcing indicated by σ^2. For weak forcing (a), there is a steady state with soil moisture most probably occurring at 0.5. This is the value of an equilibrium solution of the water balance equation without atmospheric forcing (that is, $\sigma^2 = 0$). For moderate forcing (b) and strong forcing (c), the system has a quite different property. Two states of soil moisture are obtained with the most probable values of 0.2 and 0.8. The soil frequently experiences persistent dry and wet anomalies. Atmospheric forcing may lead to transitions from one to another state.

SUMMARY AND DISCUSSION

Analytical models for examining continental water cycle have been described and applied to studies of timescale, persistence, and physical mechanism of the water cycle variability. A fourth-order model of land-atmosphere energy and water balance was used as an example to show the procedure for obtaining the perturbation equation, its solutions, the impacts of the system properties, and the physical interpretation.

The analytical results indicated that the fourth-order model has seasonal, monthly, weekly, and daily damping timescales. The seasonal scale appears only in those third- or second-order models that include disturbance of soil moisture. This emphasizes that the soil moisture feedback, and its interactions with the other variables, are the primary cause for the scale. In the model without soil moisture disturbance, the maximum timescale is only about two months. In addition, among the various interactions, the one between soil moisture and air humidity is the predominant; excluding this interaction results in a reduction in the length of the seasonal scale by more than one month.

Perturbations in the land-atmosphere system could persist for seasons. Thus, persistence is an inherent property of the continental water and energy cycle. This result provides theoretical support to the studies of Namias (1952) and others, who indicated that atmospheric anomalies possess monthly-to-seasonal persistence. The timescale and persistence features were also obtained in a number of numerical simulation and observation studies (e.g. Walker and Rowntree, 1977; Rind, 1982; Rowntree and Bolton, 1983; Yeh *et al.*, 1984; Walsh *et al.*, 1985; Liu *et al.*, 1993; Gao *et al.*, 1996; Vinnikov *et al.*, 1996). In a study using the National Center for Atmospheric Research (NCAR) Community Climate Model(CCM2) coupled with the Biosphere-Atmosphere Transfer Schemes (BATS) and observations of soil and atmospheric variables in China, Liu and Avissar (1999a) showed that soil moisture variability has significant persistence, with timescales on the order of months to seasons.

The seasonal damping timescale is mostly affected by the physical factors related to soil moisture (i.e. evaporation, runoff, and soil-moisture diffusion), and the monthly damping timescale is mainly affected by the thermal characteristics of the system. The soil-moisture self-feedback is a primary cause for the seasonal damping timescale, and

interactions between soil moisture and the other variables of the system greatly increase length of the damping timescale. These results clearly emphasize the importance of soil moisture in the variability and the causes of water cycle at seasonal timescale and support findings from other investigations with more complex tools. For instance, Castelli and Rodriguez-Iturbe (1995) found that land–atmosphere interactions can influence local atmospheric processes through the modification of the vertical lapse rate, and large-scale processes through the global dynamics of baroclinic waves. The advection rates of mass and energy, and the strength of the ageostrophic frontal circulations are particularly important in that case. Betts *et al.* (1996) showed that the monthly precipitation pattern is quite sensitive to initial soil moisture in the FIFE and BOREAS experiments. They suggested that, due to the memory of the soil-moisture reservoir, some predictability exists at monthly and seasonal scales. Avissar (1995) emphasized the importance of an appropriate representation of soil moisture's role to global climate model. While the analytical models described in this article are able to provide significant insights on the continental water cycle, it is important to keep in mind that only small-amplitude disturbances in a linear system were considered here. One obvious limitation with such models is that they neglect scale interactions in the climate system. In the real world, short-time forcings can generate long-time fluctuations. The relations between the tropical convection activities, the 30–60 day low-frequency fluctuation, and the El Nino/Southern Oscillation are examples of such interactions.

The timescale of the land-atmosphere system obtained by the linear forth-order model is independent on the sign of the initial perturbation in soil moisture. Actual processes in the land-atmosphere system are nonlinear. For example, evapotranspiration, which is a major parameter for the seasonal timescale, is specified to be linearly proportional to soil moisture in the forth-order model. In other words, the increasing rate of evapotranspiration with soil moisture is constant. But evapotranspiration actually increases nonlinearly with soil moisture (Lowry, 1959); the increasing rate is smaller when the soil is dry and larger when the soil is wet for the bare soil. This suggests that the seasonal timescale or persistence time period of the land-atmosphere system should be different when the soil is abnormally dry and wet.

The fourth-order analytical model described here does not have positive growth modes. Two factors contribute to this. First, the model does not have external forcing related to, for example, atmospheric dynamics, which causes disturbances in the water cycle. Second, the nature of development of the system (i.e. growth or decay) depends very much on model parameters. In this study, climatological values have been given to these parameters. Thus, results presented here reflect the climatological behavior of the water cycle. However, the model could give solutions that indicate growing disturbances by selecting specific sets of parameters reflecting local, short-term conditions.

Another limitation with the analytical models is that it is unable to examine the interaction between the seasonal cycle and the processes involved in the models. For example, many parameters used in the fourth-order model of LA99 have significant seasonal cycle. Among those, which are important to the seasonal timescale and persistence, soil moisture, soil temperature (therefore, potential evaporation), and solar radiation are much smaller in winter than in summer. This suggests that the damping time of the seasonal scale should be longer in winter. In other words, perturbation in the land-atmosphere system described by the fourth-order system should last longer in winter. This feature was found in the analysis of a 3D climate modeling study (Delworth and Manabe, 1988) but cannot be showed with this analytical model in which the model parameters are annual averages.

Snowpack and vegetation are two other primary processes involved in the water cycle. Snow can significantly affect soil moisture and, therefore, affect persistence of the water cycle. Vegetation can influence long-term water cycle variability by modifying surface albedo, intercepting precipitation, extracting soil water from deep layers through transpiration, and resisting runoff. Vegetation dynamics have a significant influence on the seasonal and interannual water cycle (e.g. Dickinson *et al.*, 1998; Lu *et al.*, 2001). Thus, snow and vegetation processes, in addition to soil moisture, need to be included in analytical models to better understand seasonal and interannual variability of the water cycle.

SOFTWARE LINKS

1. NCAR CCM2
   The National Center for Atmospheric Research Community Climate Model version 2. It is used to perform simulation of global climate. (**http://www.cgd.ucar.edu/cms/ccm.html**)
2. BATS
   The Biosphere-Atmosphere Transfer Scheme. It is used to simulate land-surface processes. (**http://www.atmo.arizona.edu/faculty/research/bats/batsmain.html**)

REFERENCES

Avissar R. (1995) Recent advances in the representation of land-atmosphere interactions in global climate models. *Reviews of Geophysics*, **33**, 1005–1010.

Avissar R. and Liu Y.-Q. (1996) A three-dimensional numerical study of shallow convective clouds and precipitation induced by land-surface forcings. *Journal of Geophysical Research*, **101**, 7499–7518.

Avissar R. and Verstraete M.M. (1990) The representation of continental surface processes in atmospheric models. *Reviews of Geophysics*, **28**, 35–52.

Betts A.K., Ball J.H., Beljaars A.C.M., Miller M.J. and Viterbo P.A. (1996) The land-surface atmosphere interaction: a review based on observational and global modeling perspectives. *Journal of Geophysical Research*, **101**, 7209–7226.

Brubaker K.L. and Entekhabi D. (1996) Analysis of feedback mechanisms in land-atmospheric interaction. *Water Resources Research*, **32**, 1343–1357.

Bonan G.B. (1994) Comparison of the land surface climatology of the NCAR CCM2 at R15 and T42 resolutions with implications for sub-grid land surface heterogeneity. *Journal of Geophysical Research*, **99**, 10357–10364.

Cane M.A. (1992) Tropical pacific ENSO models: ENSO as a mode of the coupled system. In *Climate System Modeling*, Trenberth K.E. (Ed.) The Press of the University of Cambridge: p. 788.

Castelli F. and Rodriguez-Iturbe I. (1995) Soil moisture-atmosphere interaction in a moist semigeostrophic model of baroclinic instability. *Journal of the Atmospheric Sciences*, **52**, 2152–2159.

Collins C. and Avissar R. (1994) An evaluation with the fourier amplitude sensitivity test (FAST) of which land-surface parameters are of greatest importance in atmospheric modeling. *Journal of Climate*, **7**, 681–703.

Cukier R.I., Fortuin C.M., Shuler K.E., Petschek A.G. and Schaibly J.H. (1973) Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. Theory. *Journal of Chemical Physics*, **59**, 3873–3878.

Delworth T. and Manabe S. (1988) The influence of potential evaporation on the variabilities of simulated soil wetness and climate. *Journal of Climate*, **1**, 523–547.

Delworth T. and Manabe S. (1989) The influence of soil wetness on near-surface atmospheric variability. *Journal of Climate*, **2**, 1447–1462.

Dickinson R.E., Henderson-Sellers A. and Kennedy P.J. (1993) *Biosphere-Atmosphere Transfer Scheme* (*BATS*) *Version 1E as Coupled to the NCAR Community Climate Model*, NCAR Technical Note/TN-387, National Center for Atmospheric Research, Boulder, p. 72.

Dickinson R.E., Shaikh M., Bryant R. and Graumlich L. (1998) Interactive canopies for a climate model. *Journal of Climate*, **11**, 2823–2836.

Dirmeyer P.A. (2003) The role of the land surface background state in climate predictability. *Journal of Hydrometeorology*, **4**, 599–610.

Eltahir E.A.B. (1998) A soil moisture-rainfall feedback mechanism. 1. Theory and observations. *Water Resources Research*, **34**, 765–785.

Entekhabi D., Rodriguez-Iturbe I. and Bras R.I. (1992) Variability in large-scale water balance with land surface-atmosphere interaction. *Journal of Climate*, **57**, 798–813.

Entin J.K., Robock A., Vinnikov K.Y., Hollinger S.E., Liu S.X. and Namkai A. (2000) Temporal and spatial scales of observed soil moisture variations in the extratropics. *Journal of Geophysical Research*, **105**, 11865–11877.

Findell K.L. and Eltahir E.A.B. (2003) Atmospheric controls on soil moisture–boundary layer interactions. Part I: framework development. *Journal of Hydrometeorology*, **4**, 552–569.

Gao X.G., Sorooshian S. and Gupta H.V. (1996) A sensitivity analysis of the biosphere-atmosphere transfer scheme (BATS). *Journal of Geophysical Research*, **101**(D3), 7279–7290.

Giorgi F., Means L.O., Shields C. and Mayer L. (1996) A regional model study of the importance of local versus remote controls of the 1988 drought and the 1993 flood over the central united states. *Journal of Climate*, **9**, 1150–1162.

Holton J.R. (1979) *An Introduction to Dynamic Meteorology*, Academic Press: p. 391.

Huang J., van den Dool H.M. and Georgakakos K.P. (1996) Analysis of model calculated soil moisture over the united states (1931–1993) and applications to long-range temperature forecasts. *Journal of Climate*, **9**, 1350–1362.

Koster R.D. and Suarez M.J. (1995) Relative contributions of land and ocean processes to precipitation variability. *Journal of Geophysical Research-Atmospheres*, **100**(D7), 13775–13790.

Koster R.D., Suarez M.J. and Heiser M. (2000) Variance and predictability of precipitation at seasonal-to-interannual timescales on precipitation. *Journal of Hydrometeorology*, **1**, 26–46.

Liu Y.-Q. and Avissar R. (1999a) A study of persistence in the land-atmosphere system using a general circulation model and conservations. *Journal of Climate*, **12**, 2139–2153.

Liu Y.-Q. and Avissar R. (1999b) A study of persistence in the land-atmosphere system with a fourth-order analytical model. *Journal of Climate*, **12**, 2154–2168.

Liu Y.-Q., Ye D.Z. and Ji J.J. (1992) Influence of soil moisture and vegetation on climate. I: a theoretical analysis on persistence of short-term climatic anomalies. *Science in China*, **35**, 441–448.

Liu Y.-Q., Ye D.Z. and Jin J.J. (1993) Influence of soil moisture and vegetation on climate. II: numerical experiments on persistence of short-term climatic anomalies. *Science in China*, **36**, 102–109.

Lowry W.P. (1959) The falling rate phase of evaporative soil moisture loss: a critical evaluation. *Bulletin of the American Meteorological Society*, **40**, 605.

Lu L., Pielke R.A., Liston G.E., Parton W.J., Ojima D. and Hartman M. (2001) Implementation of a two-way interactive atmospheric and ecological model and its application to the central united states. *Journal of Climate*, **14**, 900–919.

Mintz Y. (1984) *The Sensitivity of Numerically Simulated Climate to Land-Surface Boundary Conditions*, The Global Climate, Houghton J.T. (Ed.) Cambridge University Press: pp. 79–105.

Namias J. (1952) The annual course of month-to-month persistence in climatic anomalies. *Bulletin of the American Meteorological Society*, **33**, 279–285.

Paltridge G.W. (1974) Global cloud cover and earth surface temperature. *Journal of the Atmospheric Sciences*, **31**, 1571–1576.

Rind D. (1982) The influence of ground moisture conditions in North America on summer climate as modeled in the GISS GCM. *Monthly Weather Review*, **110**, 1487–1494.

Rodriguez-Iturbe I., Entekhabi D. and Bras R.I. (1991) Nonlinear dynamics of soil moisture at climate scales: I. Stochastic analysis. *Water Resources Research*, **27**, 1899–1906.

Rowntree P.R. and Bolton J.A. (1983) Simulation of the atmospheric response to soil-moisture anomalies over Europe. *Quarterly Journal of the Royal Meteorological Society*, **109**, 501–526.

Schlosser C.A. and Milly P.C.D. (2002) A model-based investigation of soil moisture predictability and associated climate predictability. *Journal of Hydrometeorology*, **3**, 483–501.

Sellers P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A simple biosphere model (SiB) for use within general circulation models. *Journal of the Atmospheric Sciences*, **43**, 505–531.

Vinnikov K., Robock A., Speranskaya N.A. and Schlosser C.A. (1996) Scales of temporal and spatial variability of midlatitude soil moisture. *Journal of Geophysical Research*, **101**, 7163–7174.

Walker J.M. and Rowntree P.R. (1977) The effect of soil moisture on circulation and rainfall in a tropical model. *Quarterly Journal of the Royal Meteorological Society*, **103**, 29–46.

Walsh J.E., Jasperson W.H. and Ross B. (1985) Influence of snow cover and soil moisture on monthly air temperature. *Monthly Weather Review*, **113**, 756–768.

Xue Y.-K., Sellers P.J., Kinter J.L. and Shukla J. (1991) A simplified biosphere model for global climate studies. *Journal of Climate*, **4**, 345–364.

Yeh T.C., Wetherald R.T. and Manabe S. (1984) The effect of soil moisture on the short-term climate and hydrology change-a numerical experiment. *Monthly Weather Review*, **112**, 474–490.

180: Short-Term Predictions (Weather Forecasting Purposes)

GEORGE KALLOS AND IOANNIS PYTHAROULIS

Atmospheric Modeling and Weather Forecasting Group, School of Physics, University of Athens, Athens, Greece

In recent years there is a growing need in hydrology for high-resolution precipitation predictions. Accurate quantitative precipitation forecasts at various spatiotemporal scales can be made by the atmospheric numerical models that utilize sophisticated parameterization schemes to represent the various physical processes. Two soil parameterization schemes are presented here: a) the bucket model which is a simple single-layer scheme and was the first one to be used in numerical models, and b) the multilayer Oregon State University scheme which is currently utilized in operational and research mode. A number of successful short-term predictions of severe precipitation events are also presented. These predictions were produced by various full-physics numerical models, such as the SKIRON/Eta modeling system, RAMS and the UKMO Unified model, under different forcing mechanisms. The performance of the models in representing the spatiotemporal variability of precipitation is mainly discussed. In general, the atmospheric numerical models exhibited a skill in providing accurate short-term predictions of the spatiotemporal variability and severity of precipitation.

INTRODUCTION

In recent years, there is a growing appreciation of the importance of the surface conditions in controlling weather and climate on different temporal and spatial scales (Garratt, 1993). Several studies have shown the significance of the sea surface temperatures in forcing the atmosphere. Over land, the variability of soil properties affects the partition of the surface energy fluxes and it ultimately controls the near-surface atmospheric variables (Avissar and Pielke, 1989; Li and Avissar, 1994; Betts *et al.*, 1996). The evolution of the atmospheric boundary layer and the development of synoptic and mesoscale circulations, such as surface fronts, depend on soil hydrology (e.g. Segal *et al.*, 1989b; Chang and Wetzel, 1991; Fast and McCorcle, 1991).

The land surface and the atmosphere have been shown to interact through feedback mechanisms that range from diurnal to seasonal scales (Dooge, 1992; Betts *et al.*, 1996). In his pioneer work, Horton (1931) represented the hydrological cycle as a balanced closed cycle. Following Dooge (1992), Figure 1 presents a conceptual diagram of the interactions between the land surface and the atmosphere that control the soil water content. It describes the behavior of the soil and the fluxes of moisture during wet periods and dry periods. The circular diagram starts with the end of a long period of rain (point A). The upper soil layer is considered to be saturated and the evolution of the soil water is determined by evaporation. The rate of evaporation is at the potential rate (Shuttleworth, 1993) and the flux is controlled by the atmosphere. The combined effects of evaporation and subsurface outflow will limit the supply of water from the root soil layer into the atmosphere (point B), and evaporation will drop below its potential rate. In this case, the flux is controlled by the soil. In the beginning of the wet period (point C), precipitation starts falling on a dry enough soil. In the absence of heavy precipitation, the rate of infiltration into the surface of the dry soil will be almost equal to the rate of precipitation, and the evolution of soil water will be atmospheric-controlled through the precipitation rate. If the soil becomes moist enough that not all precipitation can percolate through the surface (point D), some of it goes into runoff. In this case, the regime is again controlled by the soil, and the rate of infiltration is dictated by the soil conditions.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

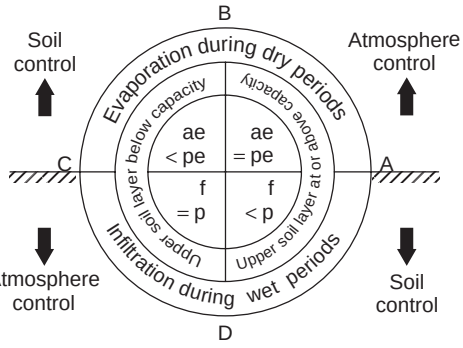


Figure 1 Diagram of the control of the surface fluxes in wet and dry periods (Dooge, 1992 © 1992 American Meteorological Society)

The above discussion indicates the need of hydrology for accurate precipitation forecasts. This quantity is a sink of atmospheric moisture while at the same time it moistens the surface of the earth and the soil. The natural environment is divided in catchments that range from a few square meters to millions of square kilometers. Therefore, the spatial and temporal scales of rainfall forecasts that are appropriate for water resource applications vary significantly. For example, flash floods are floods that occur suddenly, within 6 h of the event, on small streams and fast-moving rivers, while river floods occur on rivers and major streams because of prolonged heavy rain events and usually last over a relatively long period up to several days (LaPenta *et al.*, 1995). Nowadays, there is a large demand for quantitative precipitation forecasts with high spatial and temporal resolution by various sources, such as flood control agencies, water and emergency management agencies, and hydrologic forecasters. Opitz *et al.* (1995) argued that quantitative precipitation forecasts allow for longer lead times for river flood and crest stage forecasts, providing users additional time for preparedness.

Quantitative predictions of precipitation and soil conditions can only be made with numerical models that take into account the relevant laws of dynamics and thermodynamics and use realistic initial conditions. The state-of-the-art atmospheric models include detailed information about the characteristics and the state of the earth's surface, and sophisticated parameterization schemes for precipitation, cloud, radiation, boundary layer, surface, and soil processes. Thus, they are expected to exhibit a skill in predicting the spatiotemporal variability of intense precipitation events and their consequences.

The numerical models generally exhibit a high skill in predicting the state of the atmosphere in short ranges. However, the accuracy of the predictions falls rapidly in longer ranges, that is, in timescales longer than about 2–3 days, especially under certain weather conditions. The model resolution and the level of complexity of the various parameterization schemes are important issues for the forecast skill, especially in operational forecasting in which model run-time is a key parameter.

The model resolution also plays a critical role on whether explicit predictions can be made on the catchment size. Nowadays, most of the operational global modeling systems use a resolution of 25–50 km, while the operational regional models use meshes with a grid size down to about 10 km. In special cases, mesoscale numerical models used for nowcasting purposes employ a very high horizontal resolution of 1–2 km. This means that in operational weather forecasts, even at short ranges, the hydrology can only be represented in regional to large-scale catchments.

This article mainly deals with short-range weather forecasts produced by a number of state-of-the-art numerical models using various horizontal resolutions. Section "Parameterization of land-surface processes" describes two schemes widely used to represent the land-surface processes in operational atmospheric models. A number of case studies of severe precipitation events are discussed in the Section "Case studies of severe precipitation events". Finally, some concluding remarks are presented in the final section.

PARAMETERIZATION OF LAND-SURFACE PROCESSES

The land-surface parameterization schemes are an important constituent of the atmospheric models (Viterbo, 2002) because they define the lowest boundary conditions. Thus, they exert an important influence on both global climate and weather prediction models. Beljaars *et al.* (1993) demonstrated considerable sensitivity of the ECMWF forecast model to the complexity of the land-surface parameterization scheme. Sensitivity studies with numerical weather prediction models indicate that anomalously wet land surfaces tend to overestimate the precipitation amounts locally. On the other hand, unrealistic drying of the soil may lead to overestimation of the skin temperature. Henderson-Sellers *et al.* (1995) argued that soil moisture anomalies persist in weather forecast models because they prompt fairly substantial differences in the predicted rainfall during the first few days of simulation.

During the 1990s, a large number of land-surface schemes were evaluated and intercompared in the framework of the World Climate Research Programme Project for Inter-Comparison of Land Surface Parameterization Schemes (PILPS; see Table 1). PILPS was a project designed to improve the parameterization of the continental surface, especially the hydrological, energy, momentum, and carbon exchanges with the atmosphere (Henderson-Sellers *et al.*, 1993). The findings of this major project have been documented extensively in the literature (e.g.

Table 1 A list of models who contributed results to PILPS. Some basic information describing each scheme is included. (After Henderson-Sellers *et al.*, 1995; © 1995 American Meteorological Society)

Model	Canopy layers	Interception treated (Y/N)	Number of layers included for			Canopy	Soil moisture	Reference
			Temperature	Soil moisture	Roots			
BATSIE	1	Y	2	3	2	Penman/Monteith	Darcy's Law	Dickinson *et al.* (1986, 1993)
BEST	1	Y	3	2	2	Penman/Monteith	Philip-de Vries	Pitman *et al.* (1991) Cogley *et al.* (1990)
BUCKET	0	N	0	1	1	–	bucket + variation	Robock *et al.* (1995)
CLASS	1	Y	3	3	3	Penman/Monteith	Darcy's Law	Verseghy (1991) Verseghy *et al.* (1993)
CSIRO	1	Y	3	2	1	aerodynamic	force-restore	Kowalczyk *et al.* (1991)
GISS	1	Y	6	6	6	aerodynamic	Darcy's Law	Abramopoulos *et al.* (1988)
ISBA	1	Y	2–3	2	1	aerodynamic	force-restore	Noilhan and Planton (1989)
TOPLATS	1	Y	1	2	1	Penman/Monteith	Philip-de Vries	Famiglietti and Wood (1994)
LEAF	1	Y	7	7	3	Penman/Monteith	Darcy's Law	Avissar and Pielke (1989)
LSX	2	Y	6	6	6	Penman/Monteith	Philip-de Vries	
GFDL	0	N	1	1	1	–	bucket	Manabe (1969)
MILLY	0	N	1	1	1	–	bucket	Manabe (1969)
MIT	0	N	3	3	3	–	Darcy's Law	Abramopoulos *et al.* (1988) Entekhabi and Eagleson (1989)
MOSAIC	1	Y	2	3	2	Penman/Monteith	Darcy's Law	Koster and Suarez (1992)
NMC-MRF	1	Y	1	1	1	Lumped with soil	–	Pan (1990)
CAPS	1	Y	2	2	1	Penman/Monteith	diffusion	Mahrt and Pan (1984)
PLACE	1	Y	30	30	2	Ohm's law analogy	force-restore	Wetzel and Chang (1988)
RSTOM	–	N	0	1	1	–	bucket+variation	Milly (1992)
SPONSOR	1	Y	1	2	2	Penman/Monteith	variation	Shmakin (1998)
SECHIBA	1	Y	2	2	1	Penman/Monteith	Choisnel	Ducoudre *et al.* (1993)
SSIB	1	Y	2	3	1	Penman/Monteith	diffusion	Xue *et al.* (1991)
UKMO	1	Y	4	1	1	Penman/Monteith	diffusion	Warrilow *et al.* (1986)
VIC	1	Y	1	2	1	Penman/Monteith or full energy balance	Philip-de Vries	Liang *et al.* (1994)
BIOME	1	Y	1	1	1	Penman/Monteith		
UGAMP	1	Y	3	3	2	aerodynamic	Darcy's Law	
SIBJMA	2	Y	4	3	3	Penman/Monteith	diffusion	Sellers *et al.* (1986)
SECHIBA2	1	Y	2	2	1	Penman/Monteith	Choisnel	

Source: Reproduced from Henderson-Sellers *et al.* (1995) with the permission of Henderson-Sellers *et al.* (1995).

Henderson-Sellers *et al.*, 1993, 1995, 2003; Henderson-Sellers, 1996 and references therein) and they are of potential use to research that deals with both climate and numerical weather prediction models.

In the following sections, the hydrological considerations of two land-surface parameterization schemes will be described. The bucket model was chosen because of its simplicity and for historical purposes since it was the first attempt to parameterize land-surface processes in atmospheric numerical models. On the other hand, the Oregon State University (OSU) model is a multilayer scheme that represents a large number of hydrological processes and is widely used for research and operational purposes.

The Bucket Model

This scheme was introduced by Manabe (1969) and it was a first attempt to incorporate the effects of the earth's surface hydrology into a general circulation model. A number of numerical models have employed the scheme of Manabe or variations (e.g. Arakawa, 1972; Delworth and Manabe, 1988; Robock *et al.*, 1995; Pitman *et al.*, 2003).

The bucket model is a simple scheme that does not take into account the role of vegetation and soil type variability, and it does not include a canopy parameterization. The soil hydrology is represented in one layer that extends from the surface to 1 m depth. The choice of the soil layer of 1 m depth was due to the work of Romanova (1954) in the steppes and forest zones, and to the fact that most of the root system of a plant is concentrated in the top 1 m depth.

The amount of evaporation and the soil moisture content (SMC) are calculated following the concepts of Budyko (1956) that make use of a soil moisture availability parameter. The surface evaporation depends mainly on meteorological factors, and it is assumed to be equal to the "potential evaporation" when the soil moisture exceeds a certain critical amount. Otherwise, the evaporation is equal to a fraction of the "potential evaporation" that depends on the available soil moisture. In summary:

$$E = E_{\mathrm{o}}, \qquad \text{if } W \geq W_{\mathrm{K}} \tag{1}$$

$$E = E_{\mathrm{o}}\frac{W}{W_{\mathrm{K}}}, \quad \text{if } W < W_{\mathrm{K}} \tag{2}$$

where E_{o} is the "potential evaporation", W is the soil moisture contained in the soil layer from the surface to 1m depth, and W_{K} is a critical soil moisture value. According to Alpatev (1954), when soil moisture is not lower than 70–80% of its field capacity, evaporation from crop field is close to the value of "potential evaporation", that is, $W_K = 0.75 \cdot W_{\mathrm{FC}}$, where W_{FC} is the field capacity of moisture. For W_{FC}, a single value of 0.15 m was used.

In Manabe (1969), the evaporation (E_{o}) from a sufficiently wet land surface or ocean is calculated by:

$$E_{\mathrm{o}} = \rho(h) \cdot C_{\mathrm{D}}(h) \cdot |V(h)| \cdot (r_{\mathrm{ws}} - r(h)) \tag{3}$$

where ρ is the air density, h is the height of the lowest model level, C_{D} is the drag coefficient, V is the wind velocity, r is the mixing ratio, and r_{ws} is the saturation mixing ratio of water vapor.

In the absence of snow cover, the rate of change of soil moisture and surface runoff follow Budyko (1956) and they are expressed as:

$$\frac{\partial W}{\partial t} = R_{\mathrm{A}} - E, \qquad \text{if } W < W_{\mathrm{FC}} \tag{4}$$

$$\frac{\partial W}{\partial t} = 0 \;\&\; r_{\mathrm{f}} = R_{\mathrm{A}} - E_{\mathrm{o}}, \quad \text{if } W = W_{\mathrm{FC}} \text{ and } R_{\mathrm{A}} > E_{\mathrm{o}} \tag{5}$$

where R_{A} is the rainfall rate and r_{f} is the surface runoff, that is, runoff occurs when the soil exceeds its maximum moisture capacity. In the presence of snow, the above rates are modified taking into account the soil moisture source through snow melting and omitting the sink of soil moisture through evaporation.

The above assumptions lead to a simplified representation of the soil hydrological cycle since a large number of processes such as the interception of rainfall by the canopy, the plant evapotranspiration, the subsurface runoff, the variability in field capacity of moisture due to the soil type, the storage of infiltrated water, and so on are not considered. The bucket method has been shown to overestimate the latent heat fluxes. Pan (1990) argues that even when the soil is wet, there still exists some stomatal resistance in the plants that will reduce the evaporation from the potential value. On the other hand, in dry soils, the use of the surface temperature to infer the saturation-mixing ratio can lead to overestimation of evaporation.

In the previous decades, when the computer power was not sufficient to use advanced schemes, the bucket model was still in use and a number of authors presented various improvements to the original bucket model (e.g. Pan, 1990; Milly, 1992). Nowadays, complex multilayer schemes, which utilize all the available information about the canopy and soil characteristics, are employed in order to represent the surface and soil hydrology.

OSU Scheme

The OSU land-surface model is an advanced parameterization scheme that was originally developed by Mahrt and Pan (1984) and Pan and Mahrt (1987). A detailed description of the scheme is provided by Ek and Mahrt (1991). Modified versions of the scheme have been implemented in various widely used atmospheric numerical models including the operational NCEP-Eta model (Chen *et al.*, 1997), the SKIRON/Eta modeling system (Kallos, 1997), the weather forecasting system POSEIDON (Papadopoulos

et al., 2002), and the Penn State-NCAR MM5 model (Chen and Dudhia, 2001).

In this section, the methodology used for the hydrological processes in the SKIRON/Eta modeling system will be discussed. The SKIRON/Eta system is based on the NCEP-Eta model. A detailed description of its characteristics and configurations is to be found in Kallos (1997), Nickovic *et al.* (1998), Papadopoulos *et al.* (2002), and others. This model runs in operational and research mode in a large number of meteorological centers and institutes such as the University of Athens, the Hellenic Centre for Marine Research, the Hellenic National Meteorological Service, the Atmospheric Sciences Research Centre of the University of Albany, and others.

The land-surface scheme that represents the ground processes and calculates the energy exchanges in the soil is based on an active multilayer soil model (Mahrt and Pan, 1984) coupled to a plant canopy model (Pan and Mahrt, 1987). According to the original formulation of the scheme, the calculations for the soil temperature and moisture proceed at two layers, from the surface to 0.1 m and from 0.1 to 2 m. Storage of water is allowed in the underlying deeper layer. Papadopoulos (2001) showed that the behavior of the scheme improves significantly with the use of at least six vertical layers. This is because of the better representation of the second order differences in the soil moisture tendency equation. A schematic representation of the six-layer soil scheme currently used by the SKIRON/Eta modeling system and the corresponding depth of each layer are presented in Figure 2. It illustrates the main processes that determine the amount of soil moisture in each layer and the interactions between the atmosphere and the top layer, such as the evaporation (E), the interception and infiltration of precipitation, the vertical flux of moisture (F), and the runoff (R).

Evaporation

An important parameter for the calculation of the soil moisture and for the atmospheric forecasts in general is the evaporation from the surface of the land. The total evaporation (ET_a) from the soil and the canopy surface is obtained by adding the direct evaporation (E_d) from the top soil layer, the transpiration (E_t) via canopy, and

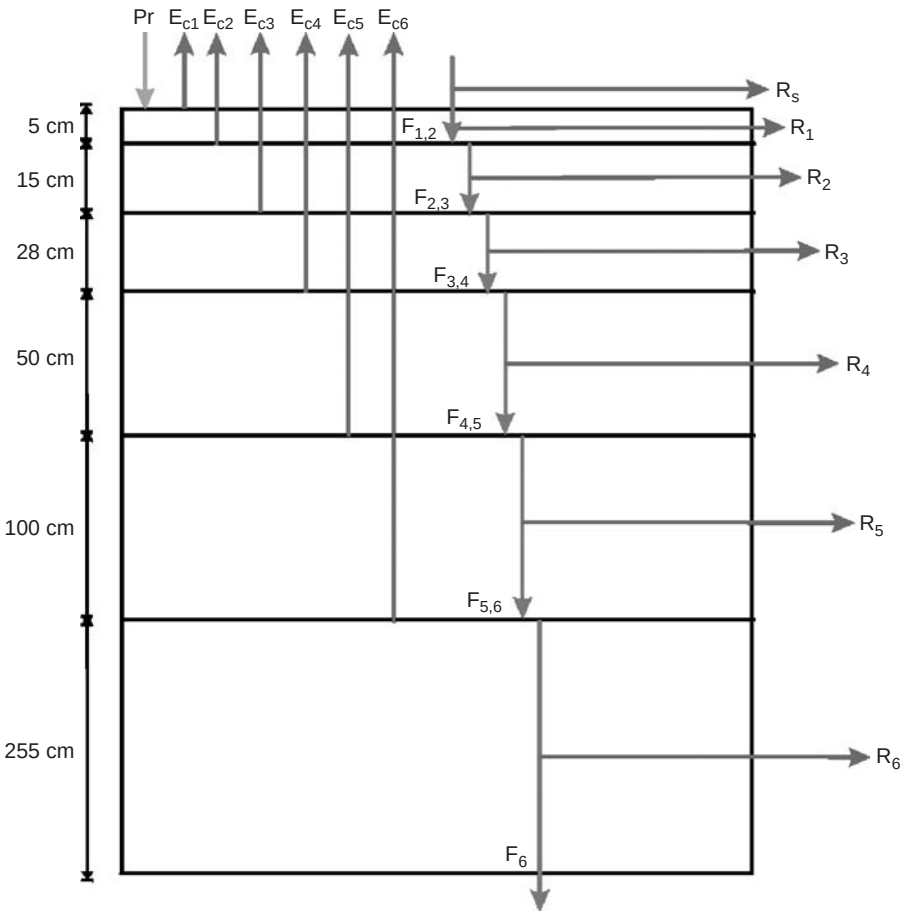


Figure 2 Schematic representation of the multilayer soil hydrology scheme used by the SKIRON/Eta modelling system (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the evaporation of the precipitation (E_c) intercepted by the canopy:

$$ET_a = E_d + E_t + E_c \tag{6}$$

The total evaporation cannot exceed the potential evaporation (ET_p).

1. *Potential evaporation*

The potential evaporation is used to compute the actual evaporation in the model. Each component of the right-hand side of the above equation can be also bounded by the potential evaporation ET_p. The usual Penman relationship has been modified (Mahrt and Ek, 1984), since the surface temperature is needed to compute the net radiation. The expression for the potential evaporation used in the model is:

$$ET_p = \left(\frac{A \cdot RR + RAD \cdot DELTA}{DELTA + RR}\right) \cdot \frac{\rho \cdot c_p \cdot c_h}{L_v} \tag{7}$$

where,

$$\begin{aligned} A &= (q_{0s} - q_0) \cdot \frac{L_v}{c_p} \\ RR &= 1 + \frac{4 \cdot \sigma \cdot T_{air}^4 \cdot R_d}{p_{sfc} \cdot c_p \cdot c_h} \\ RAD &= \frac{F_{net}}{\rho \cdot c_p \cdot c_h} + (\vartheta_{air} - T_{air}) \\ DELTA &= \frac{dq_s}{dT} \cdot \frac{L_v}{c_p} \\ F_{net} &= RSW_{in} - RSW_{out} \\ &\quad + RLW_{in} - \sigma \cdot T_{air}^4 - S \end{aligned}$$

where q_{0s} and q_0 are the saturation and the actual specific humidity at the first model level, respectively, L_v is the latent heat of evaporation ($2.5 \times 10^6\,J\,Kg^{-1}$), c_p is the specific heat of dry air ($1004.6\,J\,Kg^{-1}\,K^{-1}$), θ_{air} and T_{air} are the potential and actual air temperature at the first model level, respectively, R_d is the gas constant for dry air ($287.04\,J\,Kg^{-1}\,K^{-1}$), p_{sfc} is the surface pressure (Pa), c_h is the surface exchange coefficient for heat and moisture ($m\,s^{-1}$).

2. *Direct evaporation*

The direct evaporation from the top soil layer is calculated when the potential evaporation is greater than zero. At the air–soil interface, the direct evaporation is:

$$E_{dir} = \left(-D_{SMC} \cdot \left(\frac{\partial SMC}{\partial z}\right)_0 - K_{SMC_0}\right)(1 - \sigma_f) \tag{8}$$

where SMC is the soil moisture content, D_{SMC} is the soil water diffusivity ($m^2\ s^{-1}$), and K_{SMCo} is referred to as the hydraulic conductivity at surface ($m\,s^{-1}$), which are functions of the volumetric water content (Mahrt and Pan, 1984). The dimensionless parameter σ_f (ranges between 0 and 1) is the plant shading factor.

Chen *et al.* (1996) suggested an alternative way to calculate the direct evaporation using the equation:

$$E_{dir} = g(SMC) \cdot (1 - \sigma_f) \cdot ET_p \tag{9}$$

where the function *g (SMC)* is defined as

$$g(SMC) = \frac{SMC - SMC_{wilt}}{SMC_{ref} - SMC_{wilt}} \tag{10}$$

The soil moisture content limit SMC_{ref} refers to an upper reference value, which is the SMC value when gravity drainage of moisture in the viscous soil is allowed at a rate of $5.79 \times \times 10^{-9}\ m\,s^{-1}$ (equal to $0.5\,mm\,day^{-1}$). The SMC_{wilt} is the value of soil moisture content below which permanent wilting of the plant occurs and it corresponds to maximum potential energy needed to extract water from inside the soil. In the latter way of calculating the direct evaporation, the heterogeneity in the surface cover and the soil texture and composition are considered.

The direct evaporation (E_{dir}) can proceed at a potential rate (ET_p) when the apparent soil moisture at the surface (SMC_1) is greater than the air-dry value (SMC_d), that is, when the soil is sufficiently wet (demand control stage). When the soil dries out, the evaporation can only proceed at the rate by which the soil can diffuse water upward from below (flux control stage), in which case $SMC_1 = SMC_d$ and $E < ET_p$.

3. *Canopy evaporation*

The evaporation of the precipitation intercepted by the canopy (E_c) is calculated by:

$$E_c = ET_p \cdot \sigma_f \cdot \left(\frac{CMC}{CMC_{max}}\right)^{C_{factr}} \tag{11}$$

where CMC is the canopy water content (in m), CMC_{max} (in units of m) is the maximum (saturated) water capacity for a canopy surface (chosen equal to 0.005 m), and C_{factr} is a dimensionless constant that according to Pan and Mahrt (1987) is considered to be 0.5. The rate of change of the canopy water content is

$$\frac{dCMC}{dt} = \sigma_f \cdot PRCP - E_c \tag{12}$$

where $PRCP$ is the total precipitation intercepted by the canopy in the physics timestep. If the canopy moisture content (CMC) exceeds the CMC_{max}, the excess precipitation is considered to fall to the ground.

4. *Transpiration*

The soil water is transferred to the leaves of the plants through their roots and then to the atmosphere through plant transpiration. The model also incorporates plant transpiration (E_t) through the equation

$$E_t = \sum_{i=1}^{N} \left[ET_P \cdot \sigma_f \cdot PC \cdot \left[1 - \left(\frac{CMC}{CMC_{max}}\right)^{C_{factr}}\right] \cdot g(SMC_i) \cdot \frac{\Delta z_i}{z_N} \right] \tag{13}$$

where N is the number of soil layers, ET_p the potential evaporation, σ_f (ranges between 0 and 1) is the plant shading factor, plant coefficient (PC) is a dimensionless plant coefficient (ranging between 0 and 1), CMC and CMC_{max} are the actual and the saturated water capacity for a canopy surface, C_{factr} is equal to 0.5, and z_i is the depth of each soil layer.

The soil model uses the PC for the reduction in transpiration due to internal plant physiology. The PC is calculated using the formula

$$PC = \frac{RR + DELTA}{RR \cdot (1 + RC \cdot c_h) + DELTA} \tag{14}$$

where,
$$RR = 1 + \frac{4 \cdot \sigma \cdot T_{air}^4 \cdot R_d}{p_{sfc} \cdot c_p \cdot c_h}$$
$$DELTA = \frac{dq_s}{dT} \cdot \frac{L_v}{c_p}$$

RC is the canopy resistance that depends on the incoming solar radiation, air temperature, atmospheric water vapor pressure deficit at the lowest model level, and soil moisture availability and is calculated as (Noilhan and Planton, 1989; Jacquemin and Noilhan, 1990):

$$RC = \frac{RC_{min}}{RCS \cdot RCT \cdot RCQ \cdot RCSOIL} \tag{15}$$

where,
$$RCS = \frac{\frac{RC_{min}}{RC_{max}} + ff}{1 + ff}, \text{ with}$$
$$ff = 0.55 \cdot 2 \cdot \frac{RSW_{in}}{R_{gl}}$$
$$RCT = 1 - 0.0016 \cdot (T_{ref} - T_{sfc})^2$$
$$RCQ = \frac{1}{1 + HS \cdot (q_{0s} - q_0)}$$
$$RCSOIL = \sum_{i=1}^{N} \left(\frac{\Delta z_i}{z_N} \cdot g(SMC_i) \right)$$

where RC_{min} is the minimum stomatal resistance (depending on the vegetation type), RC_{max} is the maximum stomatal resistance and is set equal to $5000\,s\,m^{-1}$, RSW_{in} the incoming solar radiation, R_{gl} and HS are functions of vegetation type, T_{ref} is a reference temperature equal to 298 K, T_{sfc} is the skin surface temperature, q_{0s} and q_0 are the saturation and the actual specific humidity at the first model level, respectively, Δz_i is the thickness of the I soil layer, and z_N is the depth of the deepest sublayer.

Soil Hydrology

After obtaining the total evaporation, it is checked whether the CMC exceeds the maximum capacity. The canopy moisture excess is assumed to fall on the ground. The next step is the determination of the rainfall infiltration rate. The precipitation amount that cannot infiltrate or reevaporate (e.g. in the case of heavy precipitation) is specified to be surface runoff. The rest of the surface water is assumed to move in the soil. To calculate the vertical soil moisture gradient among the soil layers, the soil moisture tendency equation is solved using the integrated form of:

$$\frac{\partial SMC}{\partial t} = \frac{\partial}{\partial z} \left(D_{SMC} \cdot \frac{\partial SMC}{\partial z} \right) + \frac{\partial K_{SMC}}{\partial z} \tag{16}$$

where the soil water diffusivity, D_{SMC}, and the hydraulic conductivity, K_{SMC}, are functions of SMC (Clapp and Hornberger, 1978).

Runoff within layers is allowed if the SMC exceeds the maximum capacity of the specified soil type. The minimum allowable value of SMC is 0.02. The moisture loss due to gravitational percolation (i.e. drainage) occurs through the bottom of the lowest soil layer.

Snow Pack

The OSU scheme takes into account the effects of snow on the land surface and soil hydrology. In the presence of snowfall, the snow depth is calculated and is converted to a liquid equivalent snow depth. If the equivalent snow depth is greater than 10^{-6} m. (in a timestep), the actual evapotranspiration, the soil moisture, and the heat flux are calculated using the snow pack considerations.

If the snow layer is thick, it will evaporate/sublimate for the entire timestep at the potential evaporation rate. On the contrary, when the snow depth is thin, the evaporation cannot be maintained at the potential rate, and it is assumed that the snow will be evaporated completely over the timestep.

In the presence of snow, the snow melt is calculated and it is added to the effective precipitation that falls on the top soil layer. The soil hydrology operates as in soil without snow coverage, with the difference that the soil moisture is updated setting the effective evapotranspiration equal to zero.

CASE STUDIES OF SEVERE PRECIPITATION EVENTS

Extreme rainfall resulting from rapidly developing systems in the Mediterranean often lead to devastating floods causing significant loss of life and property damage. This kind of extreme weather phenomena is considered as the second most important natural disaster after the earthquakes in several Mediterranean countries. Since it is impossible to avoid them, a correct understanding of their principal cause can help in the early and accurate prediction and, therefore, trigger the necessary operations for the rescue of human lives and reduction of material damage.

A number of short-range predictions of severe precipitation events in the Mediterranean region will be illustrated in this section. These case studies correspond to real events,

well-documented in the literature, that were triggered by different forcing mechanisms and were predicted successfully by the utilized numerical models. The severity of these storms, their societal impacts, and the challenge they provide in order to examine the forecasting capabilities of the numerical models are the most important reasons for their choice.

The Storm of October 1994 over Greece

During October 21 1994, Greece was affected by a frontal depression that caused catastrophic floods and many casualties in urban areas as well as in the countryside (Lagouvardos *et al.*, 1996). The cold front passage was associated with heavy precipitation amounts that exhibited significant spatial variability. Eleven deaths were reported during this event, nine of them inside the Greater Athens area. The transportation, telecommunication, and energy supply networks were significantly damaged, especially in the eastern part of the country.

The observed total accumulated precipitation during a 24-h period (from 0600 UTC on 21 October to 0600 UTC on 22 October 1994) is presented in Figure 3. A band of heavy precipitation extended northward through Peloponnisos into the area of Athens. The maximum precipitation (96 mm) in Athens was reported near the city center, in the station of Nea Philadelfia, where most of the damages occurred. In central Greece, where important flooding was reported, the maximum precipitation amount was 149 mm. In northern Greece, heavy precipitation was observed in Trikala (east of the Pindos mountains) and in Katerini (east of mount Olympos) with maxima of 113 and 84 mm, respectively. These two regions of heavy precipitation were separated by a dry band. The precipitation pattern exhibited a substantial spatial variability that was investigated by model simulations.

Lagouvardos *et al.* (1996) simulated this event using two different numerical models. Their main aim was to

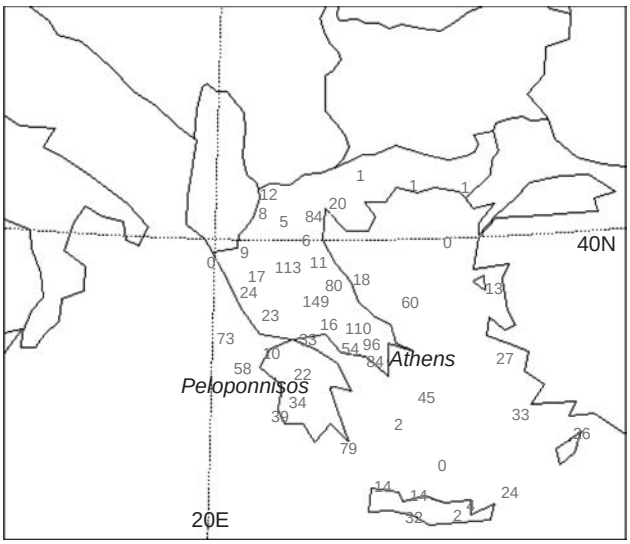


Figure 3 Total accumulated precipitation in mm from 0600 UTC 21 October to 0600 UTC 22 October 1994, reported by the surface stations (Reproduced from Lagouvardos *et al.,* 1996 by permission of American Geophysical Union (AGU)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

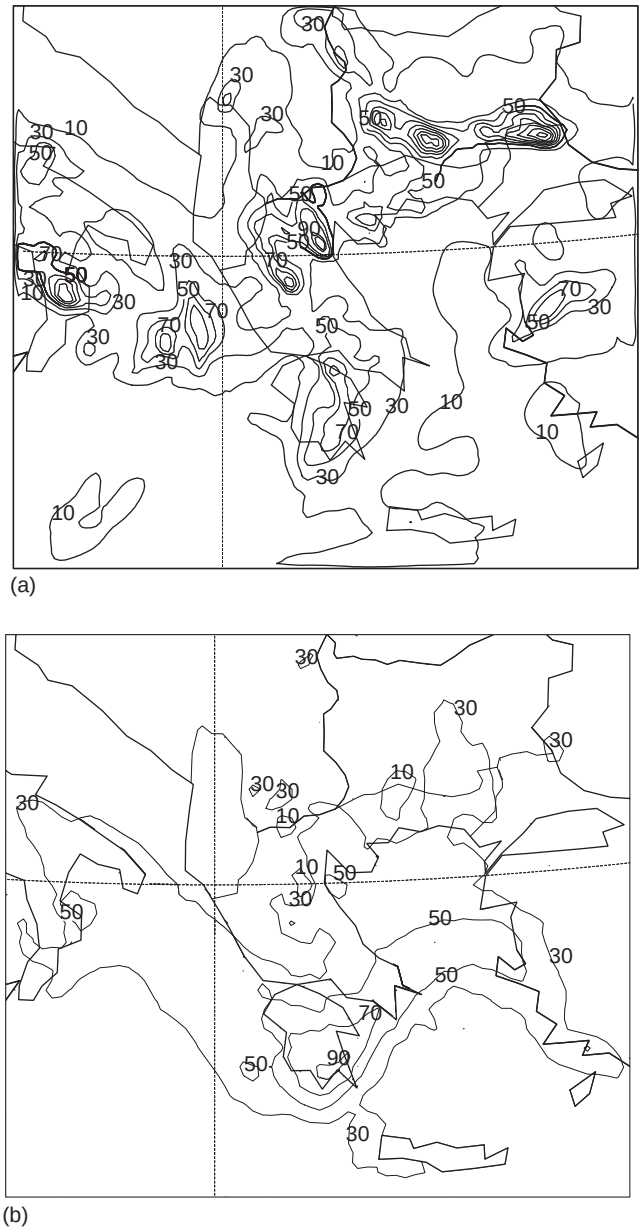


Figure 4 Total predicted accumulated precipitation from 0600 UTC 21 October to 0600 UTC 22 October 1994 (contour interval of 20 mm) from (a) RAMS inner grid and (b) SKIRON/Eta model (Reproduced from Lagouvardos *et al.,* 1996 by permission of American Geophysical Union (AGU))

investigate the capability of the numerical models to reproduce the spatial variability of precipitation at mesoscale. The two models were the Regional Atmospheric Modelling System (RAMS) and the SKIRON/Eta modeling system. RAMS used two nested grids: the outer encompassed eastern Mediterranean with a grid spacing of 40 km, and the inner covered Greece with a grid spacing of 10 km. SKIRON/Eta employed one nest that covered the central and eastern Mediterranean with a grid spacing of 25 km.

Both models were initialized at 0000 UTC, 21 October 1994, and they represented the mesoscale flow associated with the frontal depression in good agreement with the available analyses and observations. The 24-h total accumulated precipitation fields predicted from 0600 UTC on 21 October to 0600 UTC on 22 October, from the inner nest of RAMS and SKIRON/Eta models, are presented in Figure 4. The RAMS model simulations were considered as successful in describing the main mesoscale features of the precipitation pattern. The model predicted well the heavy precipitation observed over southern Greece as well as the two regions of heavy precipitation separated by a dry region over northern Greece. Also, the predicted precipitation amounts compared quite well with the observations. However, the spatial variability of precipitation within the Athens region was not represented adequately because of the model resolution. On the other hand, the SKIRON/Eta model succeeded in reproducing the main mesoscale features of the precipitation field, but its major shortcoming was the underestimation of the precipitation amounts especially in northern Greece. Finally, Lagouvardos *et al.* (1996) investigated the main sources and paths of air masses that affected Greece during this flood. Figure 5 depicts the air masses to originate in a number of areas around Greece (northern Africa, Sahara, Black Sea), to be enriched in moisture over eastern Mediterranean, and finally to move over Greece provoking catastrophic floods.

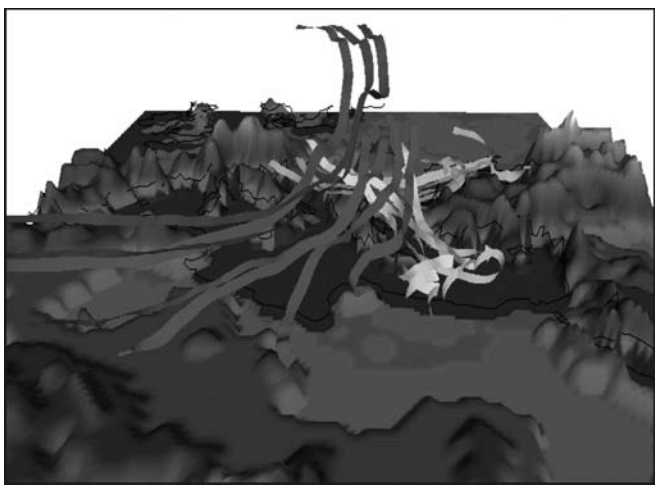


Figure 5 SKIRON/Eta trajectories showing the origin and the paths of the air masses that affected Greece during the storm of 20–21 October 1994. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

In summary, both numerical models exhibited a skill in reproducing the observed mesoscale pattern of the extreme precipitation event of 21 October 1994. The model resolution was a limiting factor in representing the details of the observed precipitation pattern and amounts in local scale. It is noted that it is very difficult to analyze the spatiotemporal variability of precipitation, especially in convective events, given the low-density of the rain gauge network.

Thunderstorm Activity Associated with Convergence Zones

A typical summer storm activity occurring over Greece was investigated by Kotroni *et al.* (1997). On 10 July 1994, a thermal low prevailed over the Anatolian plateau while central Europe and the Balkans were under the influence of a high-pressure system associated with a very weak pressure gradient over the Greek peninsula. A weak northeasterly flow resulted from this synoptic configuration. These conditions are favorable for the development of local thermal circulations and for the triggering of convergence. At the 500 hPa level, there was a trough oriented from north to south with its axis extending from the eastern part of central Europe to western Greece and the Ionian Sea. Thus, cold air advection was evident over the Balkans and Greece in the layer 850–500 hPa. The combination of the cold air advection with the low-level convergence resulted in convective activity over the Greek peninsula. Moreover, the radiosonde observations from Athens (Greece) and Sofia (Bulgaria) revealed the existence of potential instability in the middle and lower troposphere of both areas. At 1200 UTC, the CAPE reached a value of 2083 J kg^{-1} over Athens, showing that the air was potentially unstable to a high degree.

Thunderstorms developed over northern Greece in the morning of 10 July, and by 1200 UTC, the activity affected the whole peninsula (Figure 6a, b). Maximum convective activity was observed over central Greece at 1300 UTC. During the afternoon, the storms affected mainly the eastern part of the peninsula (Figure 6c, d), while after 1800 UTC, the activity gradually decayed.

The detailed structure of the atmosphere during this event, the characteristics and origin of the air masses which fed the storm activity, and the potential of predicting such events were examined through modeling. RAMS and HYPACT models were utilized in this study. Kotroni *et al.* (1997) initialized RAMS at 0000 UTC 10 July 1994 and the duration of the simulation was 24 h. Two nested grids were employed: (i) the outer, that covered the Balkans with a grid spacing of 20 km and (ii) the inner that covered the Greek peninsula with a 5-km horizontal grid spacing.

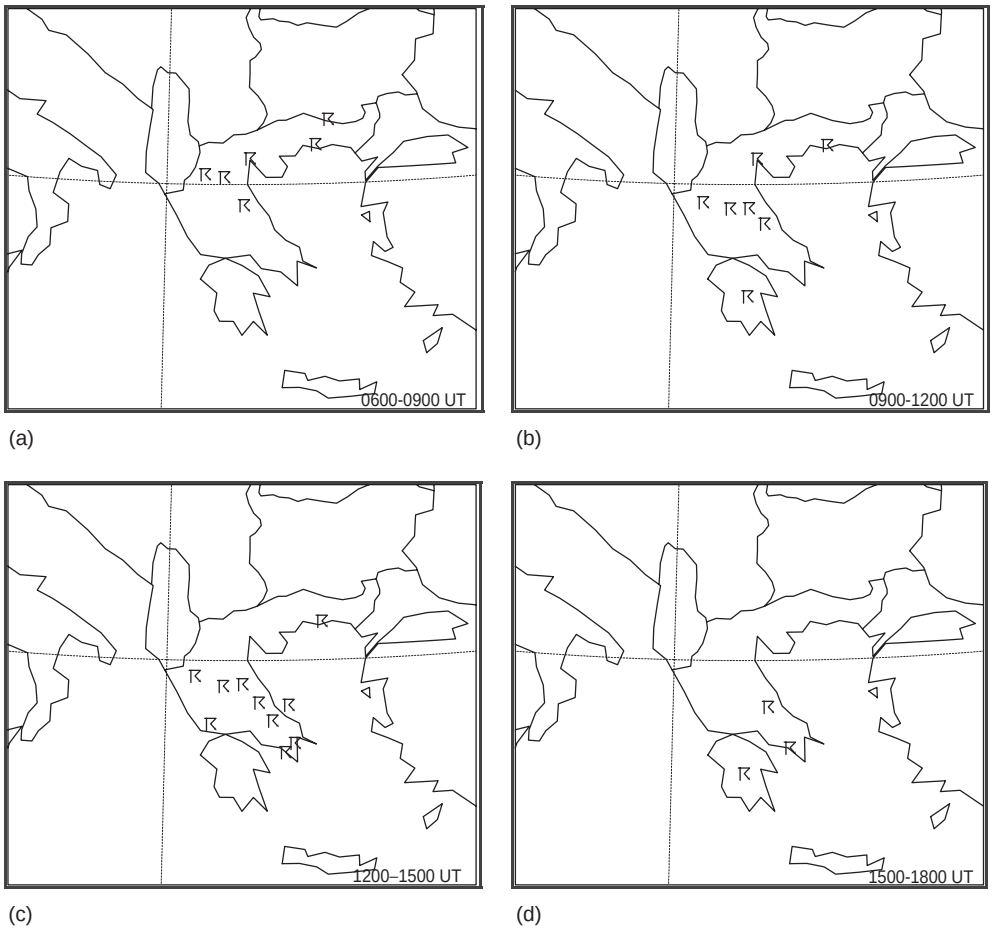


Figure 6 Reported thunderstorm activity during three hours intervals (a) from 0600 to 0900 UTC, (b) from 0900 to 1200 UTC, (c) from 1200 to 1500 UTC and (d) from 1500 to 1800 UTC, 10 July 1994 (Reproduced from Kotroni *et al.*, 1997 by permission of Royal Meteorological Society)

RAMS model predicted the synoptic conditions in good agreement with the surface synoptic network observations. Over the Greek peninsula, organized convergence was evident on an axis oriented NW to SE through central Greece, following the main axis of the topography, and over the southern part of the peninsula. The organized convergence along the main axis of the peninsula gave rise to a mechanical lifting of the air masses that met a colder air aloft.

Quite strong vertical motions were predicted over the Greek peninsula due to the differential heating and complex topography. Figure 7 presents a horizontal section of the vertical velocity field, inside the inner grid of RAMS at a height of 1500 m above ground at 1300 UTC. A band of significant upward vertical motion is associated with the convergence zone. At low levels, the strongest updrafts, exceeding $3.5\,\mathrm{m\,s^{-1}}$, were predicted over central Greece and Albania.

Regarding the time evolution of the preferred location for convective activity at the time of the initiation of thunderstorm activity at 1000 UTC, updrafts with a maximum of $2\,\mathrm{m\,s^{-1}}$ (confined to the first 4 km) are predicted in the western part of the peninsula. Later on, at the time of maximum activity (1300 UTC), the updrafts extend up to 10 km with a maximum of $5\,\mathrm{m\,s^{-1}}$ at about 6 km. The model has now shifted the convective shells by about 100 km towards the east, in agreement with the observed thunderstorm activity (Figure 6c).

Vertical sections of mixing ratio revealed the inflow of moist air masses from the maritime regions on both sides of the peninsula, which feed the convective activity and are injected aloft in the region of the strongest convection. The origin of the air masses feeding the convective activity with moisture was investigated using the HYPACT dispersion model. Particle releases parallel to the west and east coasts of the Greek peninsula were used as tracers of air masses originating from the marine boundary layer (Figure 8). The position of the sources and the distance from the coast were defined on the basis of inspection of the wind flow simulated by RAMS. Southward advection of the air

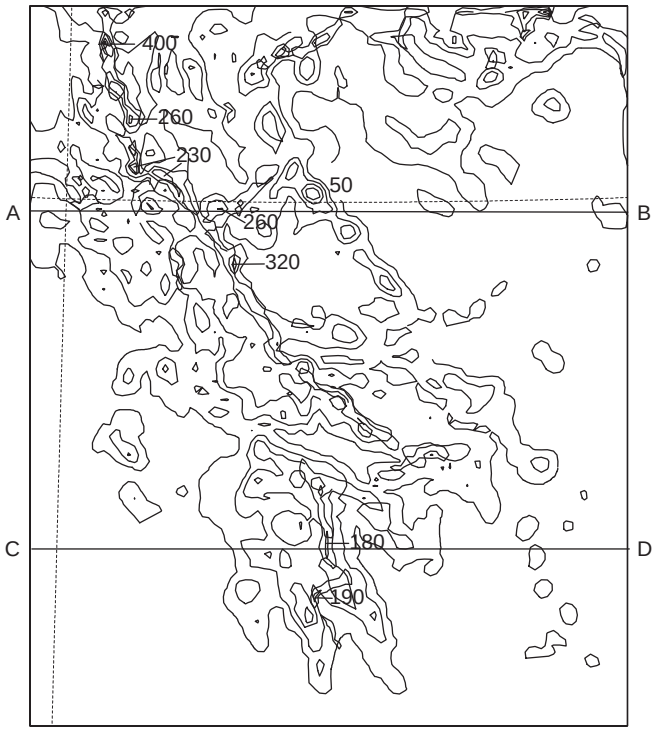


Figure 7 Horizontal section of RAMS inner grid at $z = 1500$ m AGL at 1300 UTC, 10 July 1994. Bold lines denote vertical velocity with 100 cm s^{-1} interval. Only values exceeding 50 cm s^{-1} are shown, while maximum values at each location are denoted with bold numbers (in cm s^{-1}). Topography is contoured every 500 m (Reproduced from Kotroni *et al.*, 1997 by permission of Royal Meteorological Society)

masses, away from the peninsula, was predicted during the morning hours of 10 July 1994. Later on, the development of thermal circulations begins to advect the air masses towards the coasts of Greece. These air masses enter

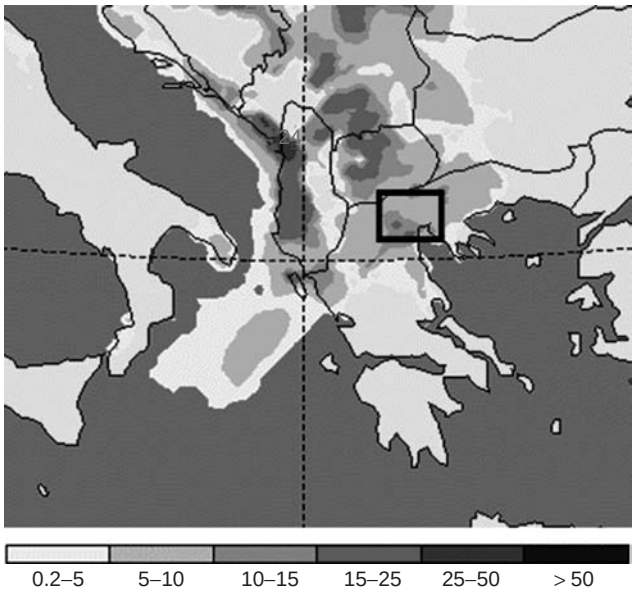


Figure 9 The 12-h predicted accumulated precipitation (mm) at 1800 UTC, 21 June 1999 (SKIRON/Eta modelling system). (Papadopoulos, 2001). The bold frame encompasses the domain of the radar image of Figure 10. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

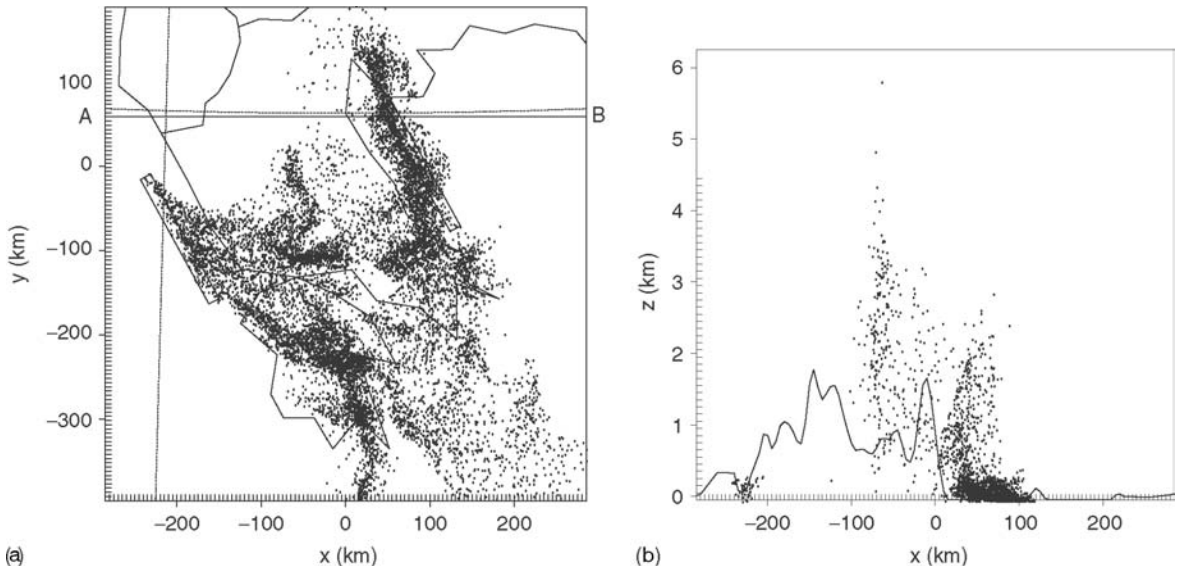


Figure 8 (a) Hypothetical moist particles position at 1400 UTC, 10 July 1994. Rectangles indicate the position of the hypothetical area source of moist particles. Line AB gives the location of the vertical cross section presented in (b). (b) Position of hypothetical moist particles, in a vertical cross section following the line AB in (a) (Reproduced from Kotroni *et al.*, 1997 by permission of Royal Meteorological Society)

within the mainland and they converge towards the main mountain axis.

A similar case of summer thunderstorm activity over Greece was studied by Papadopoulos (2001) with the use of SKIRON/Eta modeling system at a horizontal resolution of 0.125°. The accumulated precipitation predicted between 0600 and 1800 UTC on 21 June 1999 appears in Figure 9. Two areas with maximum precipitation amounts were predicted in northern Greece, in agreement with the maximum reflectivity image of the local radar (Figure 10). Moreover, Papadopoulos (2001) showed that the precipitation amounts predicted by the numerical model were in good quantitative agreement with the local rain gauge measurements.

In summary, the model simulations represented successfully the spatial and temporal evolution of the convection associated with typical summer storms occurring over Greece. Kotroni *et al.* (1997) concluded that the thunderstorm activity showed a selective spatial distribution because of the locality of thermal circulations. This could be correlated with the more vegetated and irrigated areas that supply the convective cells with moisture. The impact of vegetation and soil wetness on the moisture content of the boundary layer and the potential for development of convective clouds has been pointed out by Segal *et al.* (1989a,b). Moreover, Chen *et al.* (2001) showed that the landscape variability at small and large scales significantly affects the location and intensity of

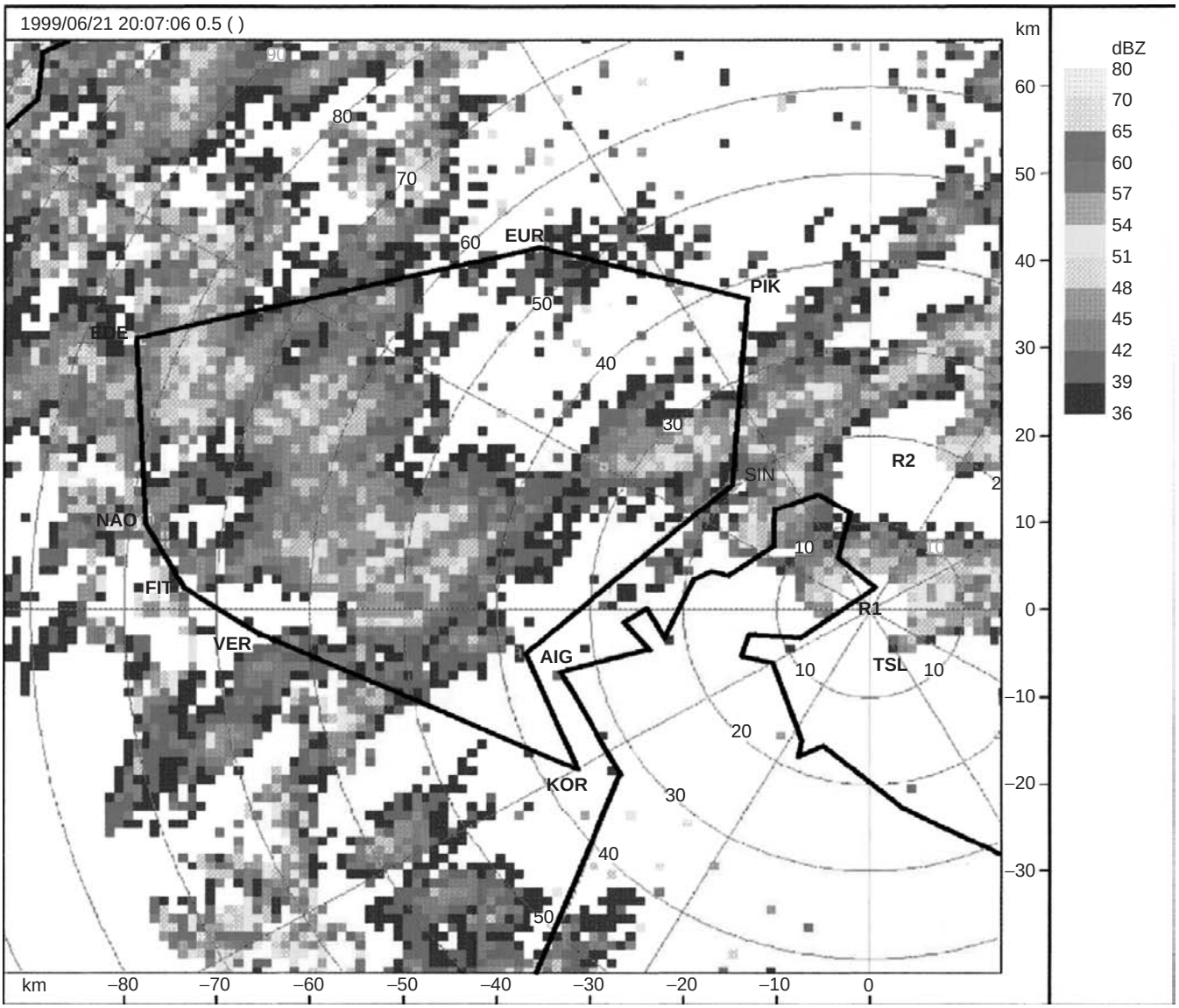


Figure 10 Maximum reflectivity image from the radar of the airport of Thessaloniki, Greece, on 21 June 1999. (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

moist convection, while Grasso (2000) suggested that the movement of a dryline and the magnitude of the low-level water vapor gradient are sensitive to changes in soil moisture.

The ''Hurricane-like'' Mediterranean Cyclone of January 1995

In mid January 1995, a small hurricane-like cyclone was detected by the Meteosat and the polar-orbiting NOAA satellites over the central Mediterranean Sea. Hereafter, it will be called *the Mediterranean cyclone*. Generally, most of the cyclogenesis occurring in the Mediterranean is caused by the lee effect of the surrounding mountains. Sometimes, small cyclones with intense convection and looking remarkably like tropical cyclones develop over the Mediterranean Sea. At least 10 times in the past 40 years, similar cyclones have been documented (e.g. Rasmussen and Zick, 1987; Reale and Atlas, 1998; Lagouvardos *et al.*, 1999; Pytharoulis *et al.*, 2000).

The Mediterranean cyclone formed at about 0330 UTC on 15 January 1995 over the sea between Greece and Sicily, close to western Greece. Before its formation, a synoptic scale low was present and strong convection was observed. Afterwards, the larger-scale parent low continued to move eastwards and decayed, while the Mediterranean cyclone remained in a wider low-pressure area. The latter system evolved for about three days, moving southwards, and finally it made landfall in northern Libya at about 1800 UTC on 17 January.

Deep convection was associated with the Mediterranean cyclone during almost all of its lifetime. The satellite images revealed a well-organized vortex with a clear "eye" (Figure 11) and several cumulonimbi. At 0600 UTC on 16 January, four ships provided reports from the vicinity of the disturbance. The ship nearest to the vortex center reported a 23 m s^{-1} (45 knots) surface wind. Lagouvardos *et al.* (1999), using data from the Special Sensor Microwave/Imager (SSM/I), concluded that surface winds of 28 m s^{-1} (54 knots) were associated with the vortex on 16 January. The threshold value of the maximum sustained surface winds for a tropical depression to be upgraded to a named tropical storm is 34 knots (Foley, 1995).

A number of modeling studies have been conducted on the Mediterranean cyclone (Blier and Ma, 1997; Lagouvardos *et al.*, 1999; Pytharoulis *et al.*, 2000), examining the predictability of its genesis and track, as well as the characteristics of its structure. The United Kingdom Meteorological Office (UKMO) Unified Model used by Pytharoulis *et al.* (2000) predicted successfully the formation of the Mediterranean cyclone and its track for the first 24–30 h. Figure 12 illustrates the predicted surface pressure and the instantaneous rate of total precipitation as well as the actual location of the system. The comparison of the location of the simulated vortex with the location of the actual system reveals that the model performed very well for the first 24–30 h. Afterwards, the model moved the cyclone in the correct path but faster than in reality, and important differences (6–8 hPa) appeared in the minimum surface pressure.

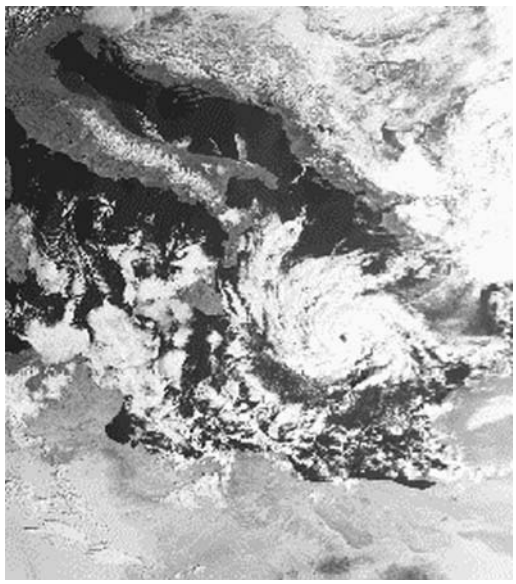


Figure 11 Visible MeteoSat satellite image showing the location and the structure of the Mediterranean cyclone at 0900 UTC on 16 January 1995

The numerical model almost always predicted a precipitation-free area at the center of the cyclone while the heaviest precipitation was predicted to fall in the area of the cyclone corresponding to the eyewall. In the area close to the cyclone, the largest part of the precipitation was produced by the convection scheme. Heavy precipitation rates were sometimes due to the large-scale precipitation scheme, indicating that there is grid-scale saturation and/or that resolved scale motions are dominating the production of precipitation in those areas. Unfortunately, there was no measurement of precipitation amount in the region of the vortex since it evolved over the sea. However, the simulated precipitation pattern, as illustrated in Figure 12, seems to be very reasonable. With one exception, precipitation was always reported from ships in the vicinity of the cyclone. No rainfall was reported at 1200 UTC on 15 January by a ship located south of the vortex. Figure 12 shows that the model predicted no precipitation south of the cyclone at that time, in agreement with this observation. Similar simulations produced with the SKIRON/Eta modeling system (using a horizontal resolution of 0.25 degrees) resulted in the development of a hurricane-like vortex with a cloud-free "eye" at the center (Figures 13, 14).

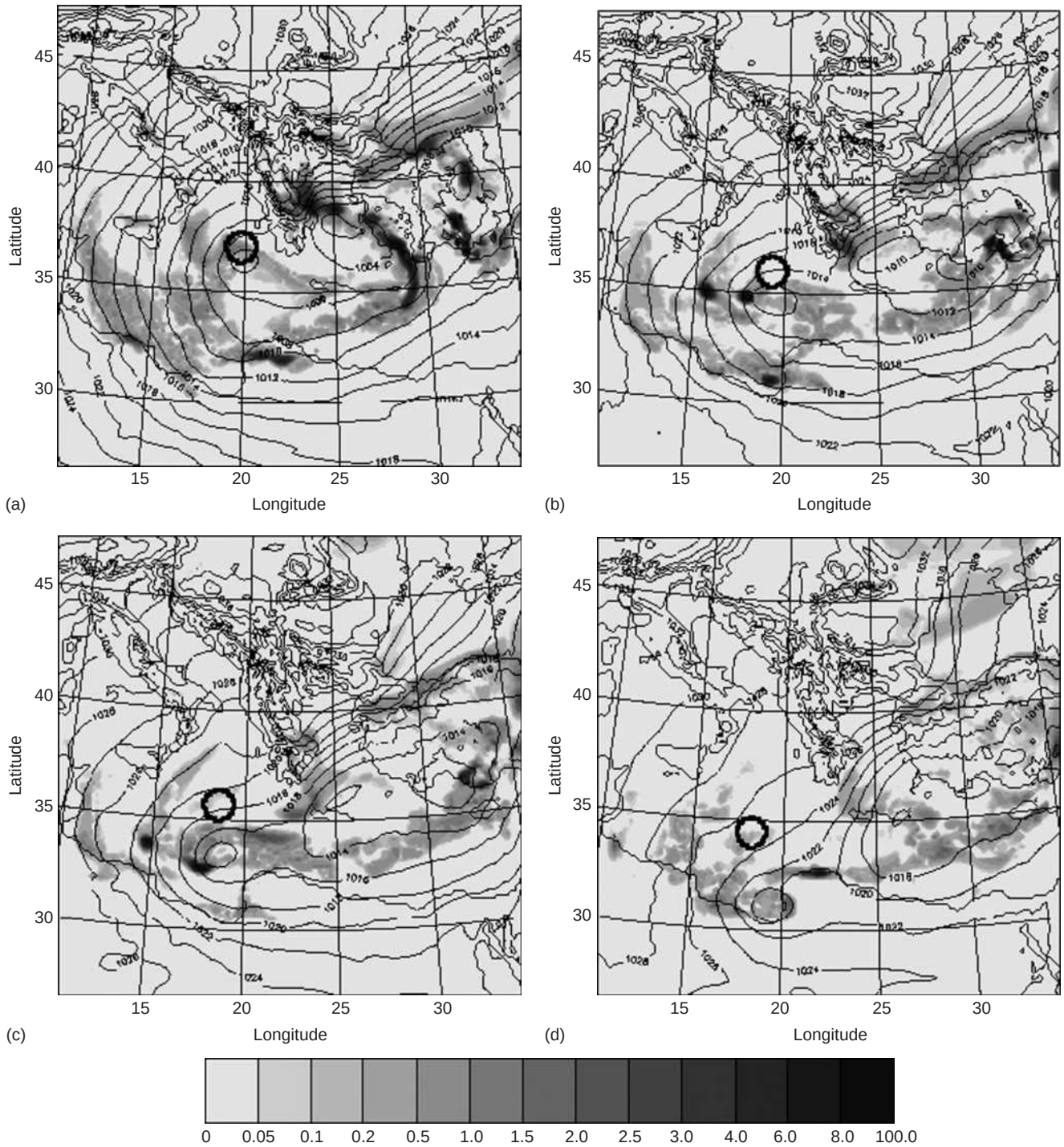


Figure 12 The mean sea-level pressure forecast and the instantaneous rate of total precipitation expressed in mm h^{-1}, at (a) T + 12, (b) T + 24, (c) T + 30 and (d) T + 42 (from 0000 UTC on 15 January 1995). The isobars are drawn every 2 hPa. The scale below the panels corresponds to the rate of precipitation and is expressed in mm h^{-1} (Unified Model). The actual location of the Mediterranean cyclone is indicated by the small cycle (Reproduced from Pytharoulis *et al.*, 2000 by permission of Royal Meteorological Society)

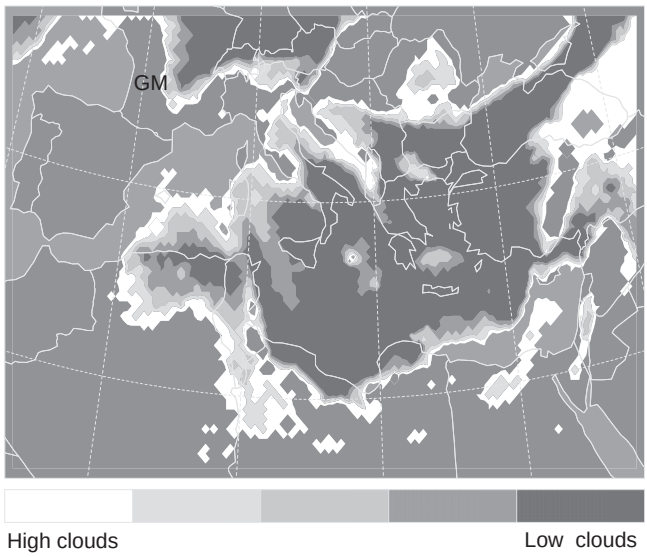


Figure 13 The SKIRON/Eta predicted cloud pattern at 1200 UTC, 15 January 1995. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

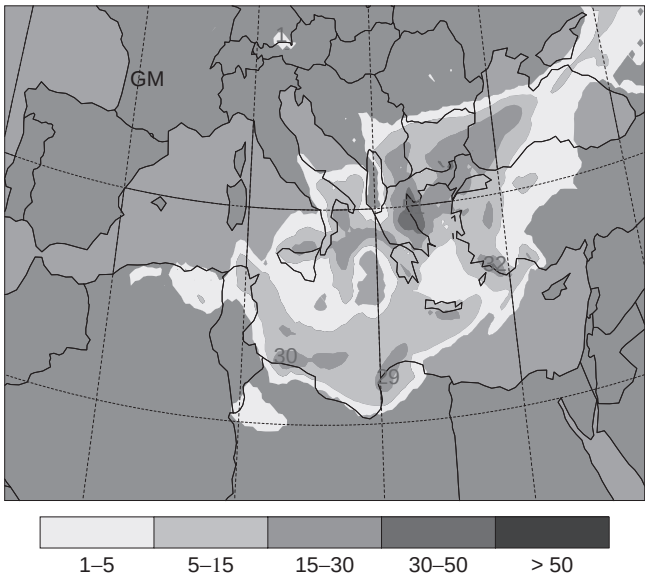


Figure 14 The 24-h SKIRON/Eta predicted accumulated precipitation (mm) at 1200 UTC, 15 January 1995. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

A similar hurricane-like cyclone appeared in the western Mediterranean basin on 6–7 October 1996 (Figure 15). Papadopoulos (2001) studied the system using the SKIRON/Eta modeling system with a horizontal resolution of 0.125°. The model successfully reproduced the track and the characteristics of the actual system, including a cloud-free "eye" surrounded by spiral cloud bands (Figure 16). Moreover, the model predicted the most intense convective activity to occur near the center of the storm in the region that corresponds to the eyewall (Figure 17).

In summary, the numerical models successfully predicted the formation and the track of these Mediterranean cyclones and the associated precipitation pattern. The strong winds, heavy precipitation, low visibility, and generally the severe weather associated with this kind of cyclones make them a great danger for the coastal areas. Many damages and casualties associated with actual tropical cyclones are not only due to their ferocity but also to the floods they cause at landfall (e.g. Hurricane Mitch in Central America in 1998). Certainly the "hurricane-like" cyclones

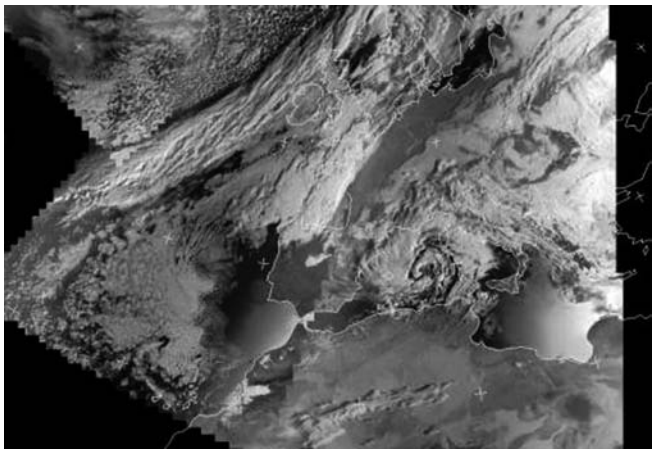


Figure 15 Composite image of successive NOAA satellite passes between 0600 and 1000 UTC on 7 October 1996. (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

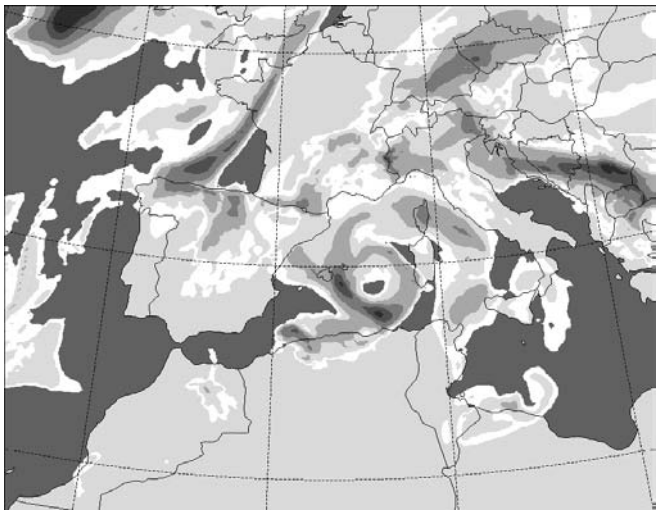


Figure 16 The SKIRON/Eta predicted cloud pattern at 0600 UTC, 7 October 1996. (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

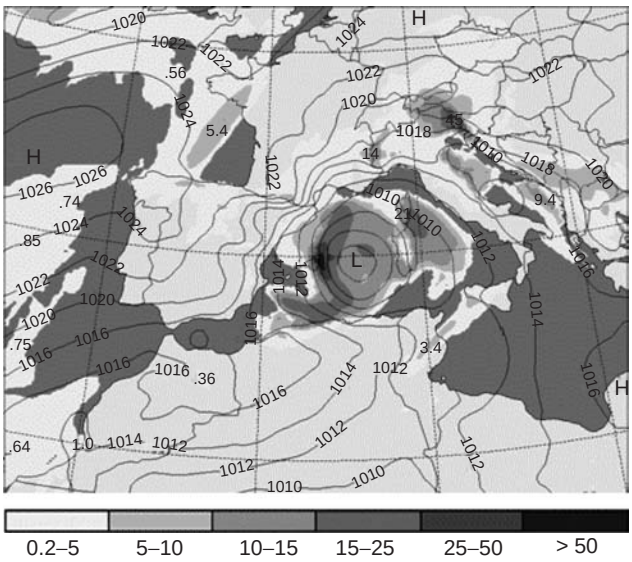


Figure 17 The 12-h SKIRON/Eta predicted accumulated precipitation (mm) at 0600 UTC, 7 October 1996. (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

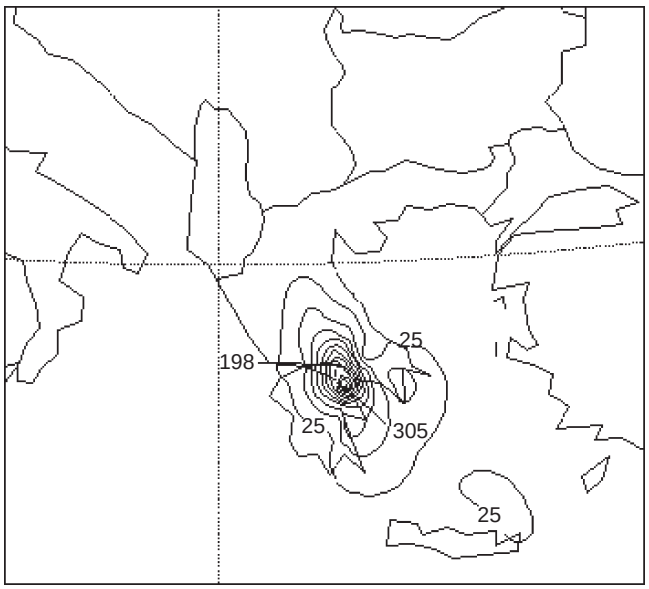


Figure 18 24-h accumulated precipitation ending at 1800 UTC 12 January 1997 (at 25 mm interval) from a network of 59 rain gauges. The 100 and 200 m isolines are in bold. Numbers indicate the accumulated precipitation at Korinthos and Desfina stations (305 and 198 mm respectively) (Reproduced from Kotroni *et al.,* 1999 by permission of Royal Meteorological Society)

that form in the Mediterranean cannot attain the strength of the Atlantic hurricanes because of a smaller reservoir of available energy from the sea and to the smaller size of the Mediterranean basin. Finally, the successful application of the numerical models for such events is very promising since they provide important guidance to the local forecasters.

The Floods over Central and Southern Greece in January 1997

On 11 and 12 January 1997, a major precipitation event occurred over the eastern Mediterranean and especially over central and southern Greece (Kotroni *et al.*, 1999) in association with the passage of a cold front. Heavy precipitation caused severe damages over southern Greece in Peloponnisos, where several cities experienced severe flooding. More than 300 mm of accumulated precipitation were measured at the station of Korinthos (Figure 18) within 24 h, ending at 1800 UTC, 12 January 1997. Flooding was not only restricted over Peloponnisos. Heavy precipitation (200 mm within 24 h) caused severe problems in the road network and some bridges collapsed in central Greece, while in Athens, flooding was reported in some suburban areas (more than 70 mm of rain within a few hours).

Kotroni *et al.* (1999) suggested that a significant factor that provoked the floods was not only the amount of the precipitation but also the fact that almost 50% was produced before the arrival of the cold front over Greece. It is speculated that the high precipitation amounts that fell before the passage of the cold front brought the soil conditions closer to saturation, reducing its infiltration capacity. In such cases, most of any additional rainfall becomes surface runoff contributing to floods, especially in heavy precipitation events such as the cold front passage that is discussed here.

This severe storm was studied by Kotroni *et al.* (1999) with the aid of the RAMS modeling system. The model was initialized at 1200 UTC on 11 January 1997 and the duration of the simulation was 36 h. Two nested grids were employed: (i) the outer, that covered southern Europe with a grid spacing of 40 km and (ii) the inner that covered the Greek peninsula with a 10-km horizontal grid spacing.

It is interesting to compare the 24-h accumulated precipitation pattern ending at 1800 UTC, 12 January 1997 (Figure 19), as it was predicted within the inner grid of RAMS, with the observed precipitation pattern (Figure 18). Figure 19 shows that most of the precipitation was accumulated along the eastern coasts of Greece. The orographic enhancement and lee side minima associated with the topography of the Greek peninsula appear to be simulated quite realistically. The model represented the amounts and the locations of the maximum precipitation amounts in good agreement with observations. The precipitation amount predicted by RAMS exceeded 350 mm within 24 h over northern Peloponnisos and central Greece, where severe

Figure 19 24-h accumulated precipitation ending at 1800 UTC 12 January 1997 (at 50 mm intervals), predicted inside the inner grid of RAMS. The 100 and 200 mm contour lines are in bold (Reproduced from Kotroni *et al.*, 1999 by permission of Royal Meteorological Society)

floodings occurred. Moreover, the spatial extent of the precipitation was predicted adequately by the numerical model.

Finally, the storm characteristics were used by Kotroni *et al.* (1998) in order to estimate the Probable Maximum Precipitation (PMP) value for the Sperhios river catchment in central Greece. Some bridges collapsed over Sperhios river catchment and the road network was damaged because of the heavy precipitation. The methodology of Collier and Hardaker (1996) was followed in this task. The PMP over a particular catchment is defined by the World Meteorological Organization as "the greatest depth of precipitation for a given duration meteorologically possible for a given size of storm area at a particular location at a particular time of year, with no allowances made for climatic trends".

In order to produce a fine resolution 3-D wind field to be assimilated into the storm model, a third inner grid was configured over Sperhios river catchment (with a horizontal resolution of 2 km). The 36-h PMP over Sperhios river for a storm with the characteristics of the one that affected Greece on 11–12 January 1997 appears in Figure 20. The storm model used to estimate PMP produced the maximum values over the Sperhios river, in agreement with the floods and the damages that occurred in the area. These results suggest that the combination of the numerical models with a number of tools, such as the above-mentioned storm model, can provide to the hydrologists very valuable information about the spatiotemporal variability of precipitation.

A Severe Frontal Cyclone over Greece in March 1998

On 25 and 26 March 1998, central and southern Greece were affected by a deep frontal cyclone associated with intense thunderstorm activity and heavy snowfall (Papadopoulos, 2001). The bad weather caused serious problems in the Greater Athens area, while the transportation was disrupted in many regions of central Greece. The observed rainfall amounts that appear in Figure 21 indicate that the strongest rainfall was due to mesoscale phenomena.

An analysis of this storm was provided by Papadopoulos (2001) who utilized the SKIRON/Eta modeling system at a horizontal resolution of 0.125°. On 25 March 1998, the cyclone was located over the sea south of Italy and it was developing in association with strong latent heat fluxes (not shown). At 0600 UTC on 26 March, the depression was located south of Greece. The southeasterly flow that was established in southern Greece transported moist air

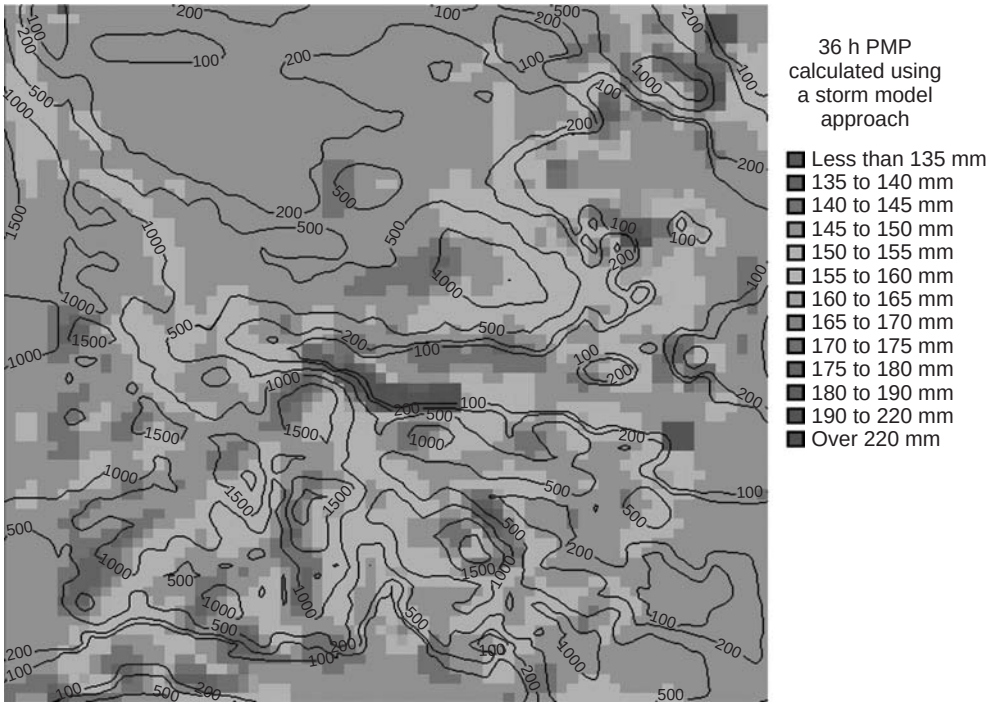


Figure 20 Contour plot of topography shaded by the 36-h Probable Maximum Precipitation for the storm of 11–12 January 1997 in the Sperhios river catchment. (Kotroni *et al.*, 1998). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

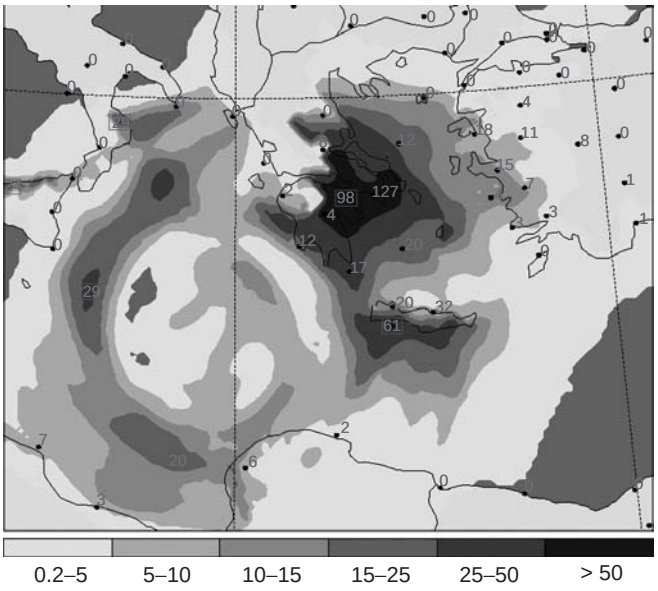


Figure 21 The accumulated rainfall (mm) predicted by SKIRON/Eta between 1800 UTC, 25 March 1998 and 0600 UTC, 26 March 1998. The framed numbers show the maximum predicted rainfall amount, while the numbers next to the black dots show the observed precipitation amount at each station. (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

masses from the Aegean Sea to continental Greece at low levels. Cold air masses were advected at higher levels by the prevailing northeasterly flow. The convergence of the warm and moist air masses at low levels (Figure 22) in combination with the presence of cold air aloft provoked intense thunderstorm activity in the Greater Athens area (Figure 21) and heavy snowfalls in central Greece (not shown). The continuous inflow of moist air at low levels contributed to the long duration of the system.

The comparison between the observed and predicted rainfall amounts (Figure 21) shows that the numerical model represented adequately the spatial distribution of rainfall. A finer horizontal resolution would be required in order to predict the magnitude of the maximum rainfall amounts in better agreement with observations and to analyze the structure of the phenomenon at the catchment scale.

CONCLUDING REMARKS

The case studies presented in this chapter revealed that the atmospheric numerical models exhibit a skill in providing accurate short-term weather forecasts. The predictions of the spatiotemporal variability and the amount of precipitation, especially in extreme weather events, are vital information for the hydrologists. However, nowadays the hydrology can be represented more accurately in regional to large-scale catchments, especially in an operational mode.

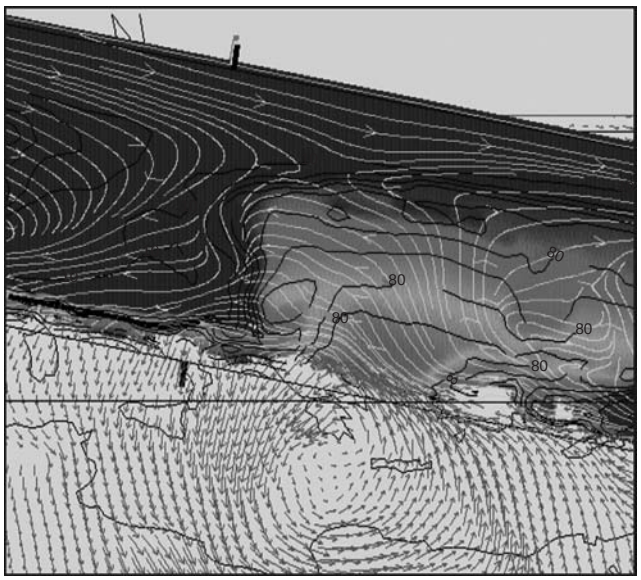


Figure 22 The SKIRON/Eta predicted flow at 400 m above sea level (arrows), streamlines and the relative humidity (contours) on a vertical section at 0600 UTC, 26 March 1998. (Papadopoulos, 2001). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The explanation of this behavior is twofold. Firstly, the numerical models generally lack the necessary resolution in order to make explicit predictions for small-scale catchments. Even for short-term nowcasting purposes, the model resolution is not finer than about 1–2 km. Secondly, the routinely available meteorological observations, even in dense networks such as central Europe, resolve the atmospheric conditions in meso-α and meso-β scales. The meso-γ scale circulations are very rarely analyzed by the observational network. Therefore, the initial conditions utilized by the numerical models generally lack the necessary spatial resolution. This means that the use of very high resolution of less than 1–2 km may not always provide better forecasts. Mass *et al.* (2002) showed that the transition from 36 to 12 km grid spacing produced a beneficial effect on the weather forecasts, while only small improvements appeared in the verification statistics as grid spacing decreased from 12 to 4 km. They concluded that the value of high resolution might be more obvious near substantial mesoscale terrain or in areas with a dense observation network.

On the other hand, the meteorological radars have the ability to scan weak precipitation and thunderstorms over large areas very quickly and make quantitative estimates of precipitation at very high resolution. However, even the most advanced advection methods that utilize the radar products cannot provide accurate forecasts for a period longer than one hour.

Finally, the analysis of the flood of January 1997 (Section "The floods over central and southern Greece in January 1997") showed that the hydrologists need information not only about the precipitation rates during an extreme event but also about the preexisting conditions of the soil. If the soil is already saturated, the falling precipitation will not be infiltrated but it will become runoff. This situation is likely to lead to flooding depending on the precipitation amounts and the topographic features. Thus, the operational short and medium range weather forecasting systems are invaluable tools for the hydrologists and the local authorities since they exhibit the necessary forecast skill and also they allow the issuance of warnings for the public and the farmers in a reasonable time frame.

REFERENCES

Abramopoulos F., Rosenweig C. and Choudhury B. (1988) Improved ground hydrology calculations for global climate models (GCMs): soil water movement and evapotranspiration. *Journal of Climate*, **1**, 921–941.

Alpatev A.M. (1954) *Vlagooborot Kul'turnykh Rastenil*, (Moisture Exchange in Crops) Gidrometeoizdat, Leningrad, p. 247.

Arakawa A. (1972) *Design of the UCLA General Circulation Model*, Technical Report No. 7, Department of Meteorology, University of California: Los Angeles, p. 116.

Avissar R. and Pielke R.A. (1989) A parameterization of heterogeneous land surfaces for atmospheric numerical models and its impact on regional meteorology. *Monthly Weather Review*, **117**, 2113–2136.

Beljaars A.C.M., Viterbo P., Miller M.J., Betts A.K. and Ball J.H. (1993) A new surface boundary layer formulation at ECMWF and experimental continental precipitation forecasts (July 1993). *GEWEX News*, **3**(3), 1, 5–8.

Betts A.K., Ball J.H., Beljaars A.C.M., Miller M.J. and Viterbo P. (1996) The land surface-atmosphere interaction: a review based on observational and global modelling perspectives. *Journal of Geophysical Research*, **101D**, 7209–7225.

Blier W. and Ma Q. (1997) A Mediterranean sea hurricane? *Proceedings of the 22nd Conference on Hurricanes and Tropical Meteorology*, American Meteorological Society: Fort Collins, pp. 592–595.

Budyko M.I. (1956) *Teplovl Balans Zemnol Poverkhnosti*, (Heat Balance of the Earth's Surface), Gidrometeoizdat, Leningrad, p. 255.

Chang J.-T. and Wetzel P.J. (1991) Effect of spatial variations of soil moisture and vegetation on the evolution of a prestorm environment: a numerical case study. *Monthly Weather Review*, **119**, 1368–1390.

Chen F. and Dudhia J. (2001) Coupling an advanced land surface-hydrology model with the Penn state-NCAR MM5 modeling system. Part I: model implementation and sensitivity. *Monthly Weather Review*, **129**, 569–585.

Chen F., Janjic Z. and Mitchell K. (1997) Impact of atmospheric-surface layer parameterizations in the new land-surface scheme of the NCEP mesoscale Eta numerical model. *Boundary-Layer Meteorology*, **85**, 391–421.

Chen F., Warner T.T. and Manning K. (2001) Sensitivity of orographic moist convection to landscape variability: a study of

the Buffalo Creek, Colorado, flash flood case of 1996. *Journal of the Atmospheric Sciences*, **58**, 3204–3223.

Chen F., Mitchell K., Schaake J., Xue Y., Pan H.-L., Karen V., Duan Q.Y., Ek M. and Betts A. (1996) Modelling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research*, **101**, 7251–7268.

Clapp R.B. and Hornberger G.M. (1978) Empirical equations for some soil hydraulic properties. *Water Resources Research*, **14**(4), 601–604.

Cogley J.G., Pitman A.J. and Henderson-Sellers A. (1990) *A Land Surface Scheme for Large Scale Climate Models*, Trent University Technical Note 90–1, p. 124 (Available from Trent University, Peterborough).

Collier C.G. and Hardaker P.J. (1996) Estimating probable maximum precipitation using a storm model approach. *Journal of Hydrology*, **183**, 277–306.

Delworth T.L. and Manabe S. (1988) The influence of potential evaporation on the variabilities of simulated soil wetness and climate. *Journal of Climate*, **1**, 523–547.

Dickinson R.E., Henderson-Sellers A. and Kennedy P.J. (1993) *Biosphere-Atmosphere Transfer Scheme (BATS) Version 1e as Coupled to the NCAR Community Climate Model*, NCAR/TN-387 STR, NCAR, p. 80.

Dickinson R.E., Henderson-Sellers A., Kennedy P.J. and Wilson M.F. (1986) *Biosphere Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model*, NCAR TN275 STR, NCAR, p. 69.

Dooge J. (1992) Sensitivity of runoff to climate change: a Hortonian approach. *Bulletin of the American Meteorological Society*, **73**, 2013–2024.

Ducoudre N.I., Laval K. and Perrier A. (1993) SECHIBA, a new set of parameterizations of the hydrologic exchanges at the land-atmosphere interface within the LMD atmospheric general circulation model. *Journal of Climate*, **6**, 248–273.

Ek M. and Mahrt L. (1991) *OSU 1-D PBL Model User's Guide*, Version 1.04, (Available from Department of Atmospheric Sciences, Oregon State University: Corvallis, 97331–2209) p. 120.

Entekhabi D. and Eagleson P.S. (1989) Land surface hydrology parameterization for atmospheric general circulation models including subgrid scale spatial variability. *Journal of Climate*, **2**, 816–831.

Famiglietti J. and Wood E.F. (1994) Multiscale modeling of spatially variable water and energy balance processes. *Water Resources Research*, **30**, 3061–3078.

Fast J.D. and McCorcle M.D. (1991) The effect of heterogeneous soil moisture on a summer baroclinic circulation in the Central United States. *Monthly Weather Review*, **119**, 2140–2167.

Foley G.R. (1995) *Observations and Analysis of Tropical Cyclones. Global Perspectives on Tropical Cyclones*, WMO Technical Document No. TCP-38, World Meteorological Organization.

Garratt J. (1993) Sensitivity of climate simulations to land surface and atmospheric boundary-layer treatments – a review. *Journal of Climate*, **6**, 419–449.

Grasso L.D. (2000) A numerical simulation of dryline sensitivity to soil moisture. *Monthly Weather Review*, **128**, 2816–2834.

Henderson-Sellers A. (1996) Soil moisture simulation: achievements of the RICE and PILPS intercomparison workshop and future directions. *Global and Planetary Change*, **13**, 99–115.

Henderson-Sellers A., Pitman A.J., Love P.K., Irannejad P. and Chen T.H. (1995) The project for intercomparison of land surface parameterization schemes (PILPS): phases 2 and 3. *Bulletin of the American Meteorological Society*, **76**, 489–503.

Henderson-Sellers A., Irannejad P., McGuffie K. and Pitman A. (2003) Predicting land-surface climates-better skill or moving targets? *Geophysical Research Letters*, **30**, 1777.

Henderson-Sellers A., Yang Z.-L. and Dickinson R.E. (1993) The project for intercomparison of land-surface parameterization schemes. *Bulletin of the American Meteorological Society*, **74**, 1335–1349.

Horton R.E. (1931) The field, scope and status of the science of hydrology. *Transactions-American Geophysical Union*, **12**, 189–202.

Jacquemin B. and Noilhan J. (1990) Sensitivity study and validation of a land surface parameterization using the HAPEX-MOBILHY data set. *Boundary-Layer Meteorology*, **52**, 93–134.

Kallos G. (1997) The regional weather forecasting system SKIRON. *Proceedings of the Symposium on Regional Weather Prediction on Parallel Computer Environments*, Athens, 15–17 October 1997.

Kotroni V., Kallos G. and Lagouvardos K. (1997) Convergence zones over the Greek Peninsula and associated thunderstorm activity. *Quarterly Journal of the Royal Meteorological Society*, **123**, 1961–1984.

Kotroni V., Lagouvardos K., Kallos G. and Ziakopoulos D. (1999) Severe flooding over central and southern Greece associated with pre-cold frontal orographic lifting. *Quarterly Journal of the Royal Meteorological Society*, **125**, 967–991.

Kotroni V., Lagouvardos K., Smith C., Hardaker P., Collier C. and Kallos G. (1998) Atmospheric modeling of a severe storm and PMP calculation over Sperhios river (Greece) catchment. *6th International Conference on Precipitation: Predictability of Rain at the Various Scales*, Hawaii, 29 June–01 July.

Koster R. and Suarez M. (1992) Modeling the land surface boundary in climate models as a composite of independent vegetation stands. *Journal of Geophysical Research*, **97**, 2697–2715.

Kowalczyk E.A., Garratt J.R. and Krummell P.B. (1991) *A Soil-Canopy Scheme for Use in a Numerical Model of the Atmosphere -1D Stand-Alone Model*, CSIRO Division of Atmospheric Research Technical Paper 23, p. 53.

Lagouvardos K., Kotroni V., Dobricic S., Nickovic S. and Kallos G. (1996) The storm of October 21–22 1994 over Greece: observations and model results. *Journal of Geophysical Research*, **101D**, 26217–26226.

Lagouvardos K., Kotroni V., Nickovic S., Jovic D. and Kallos G. (1999) Observations and model simulations of a winter sub-synoptic vortex over the Central Mediterranean. *Meteorological Applications*, **6**, 371–383.

LaPenta K.D., McNaught B.J., Capriola S.J., Giordano L.A., Little C.D., Hrebenach S.D., Carter G.M., Valverde M.D. and Frey D.S. (1995) The challenge of forecasting heavy rain and flooding throughout the Eastern region of the National

Weather Service. Part I: characteristics and events. *Weather and Forecasting*, **10**, 78–90.

Li B. and Avissar R. (1994) The impact of spatial variability of land-surface characteristics on land-surface heat fluxes. *Journal of Climate*, **7**, 527–537.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically based model of land surface water and energy fluxes for GCMs. *Journal of Geophysical Research*, **99D**, 14415–14428.

Mahrt L. and Ek M. (1984) The influence of atmospheric stability on potential evaporation. *Journal of Climate and Applied Meteorology*, **23**, 222–234.

Mahrt L. and Pan H. (1984) A two-layer model of soil hydrology. *Boundary-Layer Meteorology*, **29**, 1–20.

Manabe S. (1969) Climate and the ocean circulation. I: the atmospheric circulation and the hydrology of the earth's surface. *Monthly Weather Review*, **97**, 739–774.

Mass C.F., Ovens D., Westrick K. and Colle B.A. (2002) Does increasing horizontal resolution produce more skilful forecasts? *Bulletin of the American Meteorological Society*, **83**, 407–430.

Milly P.C.D. (1992) Potential evaporation and soil moisture in general circulation models. *Journal of Climate*, **5**, 209–226.

Nickovic S., Mihailovic D., Rajkovic B. and Papadopoulos A. (1998) *The Weather Forecasting System SKIRON. Volume II: Description of the Model*, (Available from Atmospheric Modelling and Weather Forecasting Group, School of Physics, University of Athens: Athens).

Noilhan J. and Planton S. (1989) A simple parameterization of land surface processes for meteorological models. *Monthly Weather Review*, **117**, 536–549.

Opitz H.H., Summer S.G., Wert D.A., Snyder W.R., Kane R.J., Brady R.H., Stokols P.M., Kuhl S.C. and Carter G.M. (1995) The challenge of forecasting heavy rain and flooding throughout the Eastern region of the National Weather Service. Part II: forecast techniques and applications. *Weather and Forecasting*, **10**, 91–104.

Pan H.L. (1990) A simple parameterization scheme for evapotranspiration over land for the NMC medium-range forecast model. *Monthly Weather Review*, **118**, 2500–2512.

Pan H.L. and Mahrt L. (1987) Interaction between soil hydrology and boundary-layer development. *Boundary-Layer Meteorology*, **38**, 185–202.

Papadopoulos A. (2001) *A Regional Numerical Model with Special Capabilities in the Use of the Initial and Boundary Conditions*, Ph.D. Thesis, School of Physics, University of Athens, Athens, (in Greek), p. 272.

Papadopoulos A., Katsafados P., Kallos G. and Nickovic S. (2002) The weather forecasting system for POSEIDON-an overview. *GAOS*, **8**, 219–237.

Pitman A.J., Yang Z.L., Cogley J.G. and Henderson-Sellers A. (1991) *Description of Bare Essentials of Surface Transfer for the Bureau of Meteorology Research Centre AGCM*, BMRC Research Report 32, BMRC, p. 117.

Pitman A.J., Xia Y., Leplastrier M. and Henderson-Sellers A. (2003) The Chameleon surface model: description and use with the PILPS phase 2(e) forcing data. *Global and Planetary Change*, **38**, 121–135.

Pytharoulis I., Craig G.C. and Ballard S.P. (2000) The hurricane-like Mediterranean cyclone of January 1995. *Meteorological Applications*, **7**, 261–279.

Rasmussen E. and Zick C. (1987) A subsynoptic vortex over the Mediterranean with some resemblance to polar lows. *Tellus*, **39A**, 408–425.

Reale O. and Atlas R. (1998) A tropical-like cyclone in the extratropics. *International Centre for Theoretical Physics*, Trieste, No. IC98007.

Robock A., Vinnikov K.Y., Schlosser C.A., Speranskaya N.A. and Xue Y. (1995) Use of midlatitude soil moisture and meteorological observations to validate soil moisture simulations with biosphere and bucket models. *Journal of Climate*, **8**, 15–35.

Romanova E.N. (1954) The influence of Forest Belts on the vertical structure of wind and on the Turbulent exchange. *Study of the Central Geophysical Observatory*, No. 44(106), Glavnaia Geofizicheskaia Observatoriia: Trudy, pp. 80–90.

Segal M., Schreiber W.E., Kallos G., Garratt J.R., Rodi A., Weaver J. and Pielke R.A. (1989a) The impact of crop areas in Northeast Colorado on midsummer mesoscale thermal circulations. *Monthly Weather Review*, **117**, 809–824.

Segal M., Garratt J.R., Kallos G. and Pielke R.A. (1989b) The impact of wet soil and canopy temperatures on daytime boundary-layer growth. *Journal of the Atmospheric Sciences*, **46**, 3673–3684.

Sellers P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A simple biosphere model (SiB) for use within general circulation models. *Journal of the Atmospheric Sciences*, **43**, 505–531.

Shmakin A.B. (1998) The updated version of SPONSOR land surface scheme: PILPS-influenced improvements. *Global and Planetary Change*, **19**, 49–62.

Shuttleworth W.J. (1993) Evaporation. In *Handbook of Hydrology*, Maidment D.R. (Ed.), McGraw-Hill: pp. 4.1–4.53.

Verseghy D. (1991) CLASS – a Canadian land surface scheme for GCMs. I: soil model. *International Journal of Climatology*, **11**, 111–133.

Verseghy D., McFarlane N.A. and Lazare M. (1993) CLASS – a Canadian land surface scheme for GCMs. II: vegetation model and coupled runs. *International Journal of Climatology*, **13**, 347–370.

Viterbo P. (2002) The role of the land surface in the climate system. *Meteorological Training Course Lecture Series*, ECMWF: Reading.

Warrilow D.A., Sangster A.B. and Slingo A. (1986) *Modelling of Land Surface Processes and their Influence on European Climate*, Technical Report, No. DCTN38, U.K.M.O.

Wetzel P.J. and Chang J.T. (1988) Evapotranspiration from nonuniform surfaces: a first approach for short-term numerical weather prediction. *Monthly Weather Review*, **116**, 600–621.

Xue Y., Sellers P.J., Kinter J.L. and Shukla J. (1991) A simplified biosphere model for climate studies. *Journal of Climate*, **4**, 345–364.

181: Long-Term Predictions (Climate Simulation and Analysis)

RICHARD A BETTS

Hadley Centre for Climate Prediction and Research, Exeter, UK

The global hydrological cycle is expected to change over the next century as a consequence of human perturbations to the Earth System. Human activities are modifying the composition of the atmosphere, in particular the concentrations of greenhouse gases such as CO_2 and the concentrations of aerosols. This exerts a radiative forcing of the climate system, which already appears to be warming the global climate. These changes are predicted to increase in the future, and are expected to modify precipitation and evaporation regimes, and hence impact on hydrology, with streamflow increasing in some regions and decreasing in others. Furthermore, changes in atmospheric composition may also affect ecosystems directly, with plant physiological responses to CO_2 increasing carbon storage, modifying biogeography, and decreasing evapotranspiration. This may have further effects on hydrology in addition to climate change. Human-induced changes in land cover, such as deforestation and afforestation, may also directly modify hydrology and climate by perturbing the surface water and energy budgets as well as by contributing to radiatively-forced climate change through changes in atmospheric composition. This article compares and contrasts the climatic, ecological, and hydrological changes predicted as a result of these three human perturbations to the Earth System, and discusses the key interactions between them.

INTRODUCTION

Human Perturbations to the Earth System

Human activities are perturbing the Earth System in multiple ways (*see* **Chapter 33, Human Impacts on Weather and Climate, Volume 1** and **Chapter 34, Climate Change – Past, Present and Future, Volume 1**). Concentrations of carbon dioxide, methane, and other key gases are increasing, causing an enhancement of the "greenhouse effect" in which heat radiation emitted by the Earth's surface is absorbed and reemitted by the atmosphere. This is the most likely cause of the rising temperatures observed worldwide over the last century (IPCC, 2001; *see* **Chapter 34, Climate Change – Past, Present and Future, Volume 1**). Continued increase of the greenhouse effect in this way is expected to cause significant changes in climate and weather patterns (IPCC, 2001), with major consequences for other components of the climate system such as the hydrological cycle and ecosystems (*see* **Chapter 200, Changes in Regional Hydroclimatology and Water Resources on Seasonal to Interannual and Decade-to-Century Timescales, Volume 5**).

As well as exerting a radiative forcing on the climate system, increasing concentration of atmospheric CO_2 may also exert a forcing through direct effects on plant physiology. Photosynthesis generally increases with CO_2 concentration, a mechanism often termed "CO_2 fertilization". Furthermore, a number of studies have shown that plant stomata open less under higher CO_2 concentrations (Field *et al.*, 1995), which directly reduces the flux of moisture from the surface to the atmosphere (Sellers *et al.*, 1996). If this were to occur on a large scale, this could have a significant impact on the surface water balance of the landscape, affecting runoff and the supply of moisture to the atmosphere. In regions where much of the moisture for precipitation is supplied by evaporation from the land surface, reduced stomatal opening may also contribute to decreased precipitation. Furthermore, a decrease in moisture flux modifies the surface energy balance, increasing the ratio of sensible heat flux to latent heat flux and therefore warming the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

air near the surface. A decrease in stomatal opening may therefore cause an increase in near-surface air temperature through this physiological forcing alone, independent of any radiative forcing.

Although this may be partly offset by increases in leaf area (Betts *et al.*, 1997), this offset may not be total. While it is yet to be established whether such responses are universal, it is therefore possible that rising CO_2 concentrations could exert two forcings on the climate system; (i) radiative forcing of global climate modifying atmospheric circulation patterns, and (ii) physiological forcing modifying near-surface temperature and also reducing the supply of moisture for precipitation.

Although the forcing by increasing greenhouse-gas emissions includes a contribution from deforestation, this is only one means by which land-cover change can influence climate. Vegetation influences the surface fluxes of radiation, heat, moisture, and momentum, so if the character of the vegetation cover is modified, changes to the climate can result. Conversion of forest to cropland or pasture can reduce the aerodynamic roughness of the landscape and decrease both the capture of precipitation on the canopy and the root extraction of soil moisture; these changes tend to decrease evaporation and hence the fluxes of moisture and latent heat from the surface to the atmosphere, which acts to increase the temperature near the surface. Also, a forested landscape generally has a lower surface albedo than open land, particularly in conditions of lying snow when shortwave radiation is trapped by multiple reflections within the forest canopy. Deforestation can therefore lead to increased shortwave reflection, which provides a cooling influence. The relative importance of these processes depends on local conditions and can vary with season and location.

Changes in global ecosystems under climate change may themselves exert further effects on climate. These feedback could be both global and local in nature. Changes in total carbon storage of an ecosystem will modify the concentration of CO_2 in the local atmosphere, and since CO_2 has a long residence time and is well mixed, this change will be spread throughout the global atmosphere. Changes in vegetation structure may also modify regional climates through the surface energy and water budgets. In many cases such feedback will be predominantly local, mainly affecting the environment of the ecosystem providing the feedback. In other cases, biotic feedback can exert significant remote effects through teleconnections, sometimes extending across the globe.

The various human perturbations to the climate system have the potential to produce major impacts either individually or in combination. The combined effect of these perturbations acting together may be very different to the linear sum of the individual perturbations acting in isolation. This article reviews the current state of the art in the use of numerical models for projections of anthropogenic changes to the climate system, and discusses potential interactions between the different perturbations.

Using Models for Long-term Projections

Guidance on the nature of future environmental changes can in principle be gained through several methods. One method is to refer directly to a past event or state similar to that expected for the future, that is, an analogue. For example, in order to assess the climatic effect of a doubling of the atmospheric CO_2 concentration, one could in principle refer to a period in the Earth's history when the CO_2 concentration was at this level and examine the differences in climate relative to the present day. In general, analysis of paleoclimate and paleoecological data can provide many useful insights into the operation of the Earth system and potentially provide opportunities for validation of climate models. However, the use of the past as a guide to specific states in the future relies on such an analog state having existed and having left useful evidence, and this is often not the case in practice. To continue the above example, atmospheric CO_2 seems not to have been at double the present-day level for more than twenty million years (IPCC, 2001), outside the period of extensive, reliable paleoreconstruction for many aspects of the climate system. Therefore our ability to assess the implications of doubling CO_2 through the analogue method is limited.

Another technique might be to examine current trends and make some assumptions regarding the continuation of these into the future, that is, an extrapolation. For example, one could correlate the observed changes in atmospheric CO_2 and climate and extrapolate these to infer the climatic effects of doubling CO_2. However, the changes observed to date are relatively small compared to the natural internal variability of the climate system, so many apparent climate changes cannot be reliably correlated with CO_2 rise.

A further problem with both the analogue and extrapolation techniques is the difficulty in accounting for multiple sources of perturbations acting together. Since the different human influences on the climate system may interact in complex, nonlinear fashions, past changes associated with any individual perturbation may not be a useful guide to their contributions to combined perturbations.

The most promising technique for assessing future changes is to use observations and experiments to gain understanding of the system and its components, and synthesize this in a model (*see* **Chapter 201, Land-Atmosphere Models for Water and Energy Cycle Studies, Volume 5**). This allows the modeler to potentially explore the properties of the system beyond those that have been directly observed. If the model validates well against reality in those aspects that can be directly observed, the model can then be used as a controllable "virtual reality" to provide information on the aspects of the system that cannot be directly observed. This can provide improved

understanding of the behaviour of the system, and allow projections of future behavior to be made.

While a perfect "virtual reality" would include every detail of the entire system, the level of detail is in practice limited by the knowledge of the modeler and the capacity of the computers on which the model simulations are performed. An appropriate level of detail is therefore chosen within pragmatic boundaries on the basis of the particular application of the model. For example, a model of interactions between particular species of plant will require treatment of detailed physiological processes at a level at which interspecies differences are significant. In contrast, a model of the interactions between large-scale vegetation patterns and atmospheric processes may assume interspecies differences to be negligible in comparison with the differences between functional types such as trees and grasses.

In modeling studies of climate change, the climate system is typically represented as a number of subsystems such as the atmosphere, the biosphere, and the land surface. The interactions between these subsystems can be treated with two different methods. The traditional, disciplinary method is to assume a linear cause-effect chain in which one subsystem is simulated first, and the resulting state is then applied as input to a second subsystem model, and so on. For example, projections of radiatively forced climate change and its impacts on ecosystems and hydrology generally consist of such a linear sequence of simulations with models of different components of the climate system. Scenarios of greenhouse-gas and aerosol emissions are derived with models that follow scenarios of future changes in population, technology, and economic state. These emissions scenarios are then translated into scenarios of greenhouse-gas concentrations in the atmosphere, using models that consider the processes acting to reduce concentrations either by chemical processes in the atmosphere or uptake at the land and ocean surfaces. These scenarios of greenhouse-gas rise are then applied as inputs to climate models such as General Circulation Models (GCMs), which compute the radiative forcing due to the greenhouse-gas rise and the consequent changes in climate. The resulting climate scenarios are then applied to further models that simulate the response of ecosystems, hydrology, and other aspects of the Earth system including human societies.

One advantage of using a chain of loosely coupled models is that is allows each subsystem model to be developed to best represent the particular subsystem, such as with appropriate spatial resolutions and time steps, without any constraint imposed by a need for compatibility with the other models. For example, coarse-resolution output from a global-scale climate model can be used in combination with statistics of fine-scale climatic features in order to provide input to a higher resolution hydrology model capturing catchment-scale topographic detail. A second advantage is that any biases in the output of one model can potentially be accounted for by adjustment of data prior to input to the next model. For example, such adjustment is frequently used when applying climate model output to climate impacts models, in order to remove systematic errors from the climate simulations. A third advantage is that analysis of causes and effects is relatively straightforward, as processes can be broken down into components.

However, the use of a linear cause-effect chain of subsystem models neglects the possibility of feedback between the subsystems. For example, if an ecosystem model simulates major changes in terrestrial carbon storage in response to a simulated climate change, this might imply a potential feedback to atmospheric CO_2 and hence climate, which was not accounted for in the original climate simulation. Furthermore, some subsystem models represent the same processes at the interface between the subsystems, but with inconsistent methods. Atmospheric models, for instance, include models of surface evaporation in order to simulate a conservative water cycle. Hydrological models also simulate surface evaporation, but often at a higher spatial resolution and with different treatments of aspects such as surface saturation. If the evaporation simulated by a hydrological model is different to that simulated by the atmospheric model from which its input data is derived, the input data is inconsistent with the simulation.

In the light of these drawbacks, an alternative modeling approach has come into use in the last decade, involving the coupling of models of the climate subsystems to allow two-way interactions. While this means that model features such as resolution may be an additional constraint on some of the subsystem models, and makes adjustment of biases and understanding of model behavior more difficult, it does ensure consistency between the models and allows the simulations to include feedback within the climate system.

This article describes a range of models used to assess potential future changes in the climate system. Both linear sequences of models and tightly coupled model systems are described, as both approaches provide insight into mechanisms of change in the Earth system.

Model assessments of future changes in the climate system typically fall into one of two categories. One category is sensitivity studies, in which the response of the system to some illustrative perturbation is examined. Examples of this are simulations of the climate sensitivity to doubled CO_2 or to a complete deforestation of a particular region. Such studies are used to assess and understand in principle the potential effects of perturbations, in a simple and clear manner. The other category is scenario studies, where the perturbation is intended to represent a particular expected change more closely. Scenario studies are typically more complex than sensitivity studies, so may be less easy to interpret although the final outcome may be more representative of an actual future change.

The following discussion will include both scenario and sensitivity studies.

MODELS OF THE CLIMATE SYSTEM AND ITS COMPONENTS

Atmosphere and Ocean General Circulation Models

Models of the atmosphere and oceans are fundamental tools in projecting climate changes. One widely used form of atmospheric model is the General Circulation Model (GCM), which simulates the global circulation of the atmosphere based on equations of fluid motion and thermodynamics along with other meteorological processes such as cloud formation, precipitation, and absorption and emission of solar and terrestrial radiation. Similarly, oceanic GCMs simulate the circulation of the oceans. Ocean and atmospheric GCMs are routinely used together as a coupled model, to allow for key interactions through evaporation, precipitation, fluxes of heat and transfers of momentum. Atmospheric GCMs also represent similar interactions with the land surface, and it is often the output of the GCM land surface schemes that are of greatest interest in projections of climate change from a human perspective. GCM and surface schemes take account of the characteristics of the land surface in terms of the topography and nature of the land cover. GCMs used for global simulations currently use spatial resolutions of approximately $1^\circ - 3^\circ$ in the horizontal and 15–40 levels in the vertical.

Regional Climate Models (RCMs) are essentially GCMs applied over a portion of the Earth's surface rather than the entire globe. The smaller area of coverage means that spatial resolution can be higher, typically 20 km–50 km in current RCMs, providing more precise regional detail and better representation of smaller-scale atmospheric phenomena such as storms or orographic rainfall. RCMs are typically nested within GCMs that provide the atmospheric conditions at the boundary of the RCM area, such as temperature, humidity, wind velocity, and pressure.

Similarly, mesoscale models operate on similar principles to GCMs but at a resolution higher still than RCMs (approximately 1 km–10 km). Mesoscale models can capture fine-scale details of the atmosphere and underlying surface, such as detailed orography or patterns of land cover. Again, mesoscale models are nested within a larger, more coarse-resolution model such as RCM or GCM in order to obtain boundary conditions.

Ecosystem Models

Biogeography models represent ecosystems at large scales. Instead of modeling individuals, biogeography models consider ecosystems as entities in themselves. Environmental factors such as climate and soil type are used to determine the character of an ecosystem through either (or both) mechanistic modeling of physiological processes or/and empirical relationships derived from observations. These models therefore simulate the potential natural vegetation that would exist under the applied conditions if given sufficient time for successional processes to complete. Input data typically consists of mean climate variables for each calendar month of the year, so the modeled ecosystems are effectively simulated under a single year of mean climate. Output data represent that mean ecosystem state for each grid cell, either in terms of structural or physiological characteristics such as leaf area index (LAI) or net primary productivity (NPP), or in terms of an ecosystem classification such as "evergreen needleleaf forest" or "savanna".

Biogeography models may be used in conjunction with GCMs, RCMs, or mesoscale models to directly investigate ecosystem responses to the simulated climate changes and for studies of feedback between ecosystems and climate. They do not include time-dependent responses of ecosystems to environmental changes, and assume ecosystems to be at equilibrium with the climate, so are best applied to sensitivity studies and simulations of slow, long-term changes, rather than scenarios of rapid change in which time-dependency is important. In the context of anthropogenic climate change over decadal or century timescale, biogeography models cannot make specific predictions but can be used as a guide to the potential sensitivity of ecosystems to projected changes.

Dynamic global vegetation models (DGVMs) combine the large-scale applicability of biogeography models with time-dependent vegetation dynamics. Terrestrial vegetation is modeled at the global scale including the dynamics of competition and succession. Plant physiological processes such as photosynthesis, respiration, and transpiration are modeled mechanistically, but the calculations are applied to large-scale ecosystems (e.g. GCM resolutions of approximately 3°) rather than specific individuals. DGVMs take time-dependent input data such as monthly climatic variables for many years or decades, and output large-scale ecosystem structure, which again varies in time. The inclusion of rates of growth and spreading therefore makes DGVMs suitable for time-dependent modeling of vegetation responses to transient climate change.

Different DGVM modeling groups employ a range of different approaches to representing vegetation dynamics at large scales. Since individual species would be too numerous to model explicitly, all DGVMs represent vegetation in terms of plant functional types (PFTs), which effectively group together species of plants according to functional similarities. The number of PFTs identified can vary between models. For instance, the VECODE model (Brovkin *et al.*, 1997) identifies only evergreen and deciduous trees and grass, whereas other models (LPJ, Sitch *et al.*, 2003) and (IBIS, Foley *et al.*, 1996) distinguish dry and

cold deciduousness, tropical, temperate, and boreal trees, and C3 and C4 grasses. Competition between PFTs is either represented in terms of individuals over several patches and scaled up to the grid cell (HYBRID; Friend *et al.*, 1997) or in terms of competition between populations (TRIFFID; Cox, 2001).

DGVMs are concerned with the general functional character of vegetation rather than the behavior and distribution of particular species, so cannot directly model biodiversity. However, DGVMs can be used to investigate a number of processes of relevance to climate change and biodiversity. Changes in the general viability of a PFT under a climate change will imply potential impacts on species within that PFT. Furthermore, changes in the abundance of species making up a particular PFT can impinge on the resources available to the species of another PFT, so changing competitive balances between PFTs suggests changes in ecosystem structure that have implications for biodiversity. If a PFT is simulated to disappear in a particular location, this may be taken to imply a threat of local extinction of all species within that PFT. However, it should be noted that the PFT by definition do not represent the diversity within a group of similar species, some of which may be more resilient than others.

Synergies between the direct effects of climate change and the indirect effects acting via competing PFTs can therefore be investigated with DGVMs. The extra effects of CO_2 fertilization, again acting both directly on a given PFT and indirectly through competition, can also be investigated.

Hydrological Models

Hydrological models simulate the land surface components of the water cycle, namely, evaporation and runoff, in response to meteorological inputs. The boundary conditions of hydrological models typically include landscape topography, soil characteristics (such as grain size and infiltration properties) and vegetation characteristics (such as rooting depth and aerodynamic roughness). The domains of hydrological models range from individual catchments to the globe, with horizontal resolutions typically decreasing with domain size.

The more sophisticated and/or higher resolution hydrological models are typically used "offline" in studies of the impacts of climate change on surface hydrology (e.g. Arnell *et al.*, 2001). However, climate models and ecosystem models generally include their own surface hydrology submodels as a necessary component of the climate or ecological system (e.g. Cox *et al.*, 1999). Such models are typically simpler than the "offline" hydrological models in terms of their representation of landscape- and soil-related processes such as horizontal transport and water table depth. Nevertheless, hydrological submodels within climate and ecosystem models allow for interactions and feedback between hydrology and climate or ecosystem processes.

Coupled Ecosystem-climate Models

Given the potential for major feedback from ecosystems, it is clear that predictions of future climate change should consider ecosystem responses and their effect on climate. This has led to the development of "Earth-system Models", which couple models of the physical climate system (the atmosphere and oceans) to models of the terrestrial and marine biosphere (Foley *et al.*, 1998; Cox *et al.*, 2000; Ganapolski *et al.*, 2001). The physical and biological models interact via biogeochemical cycles and through the impact of life on the physical properties of the Earth's surface. Some current models feature a DGVM and/or an interactive carbon cycle included within an existing GCM to provide detailed simulations with as much mechanistic process representation as possible. Other models use approximations in the more complex parts of the Earth System (e.g. atmospheric dynamics) in order to increase computational efficiency and facilitate a greater number of simulations – such models are termed *Earth-system Models of Intermediate Complexity* (EMICs).

The inclusion of DGVMs in GCMs allows climate prediction simulations to include feedback from ecosystems responding to climatic changes at global and regional scales (Cox *et al.*, 2000). Coupled GCM-DGVMs are therefore potentially valuable for understanding and predicting synergistic responses of ecosystems to climate change over timescales of centuries and spatial scales of hundred of kilometers. Betts *et al.* (2004) discuss the global vegetation dynamics simulated by the Hadley Centre coupled climate-vegetation model for a greenhouse-gas emissions scenario over the next century.

Earth-system Models of Intermediate Complexity (EMICs, e.g. Ganapolski *et al.*, 2001) use simpler representations of the atmosphere and biosphere to allow simulations over longer periods than those possible with GCMs. They can therefore be used to explore long-timescale global synergies over millennia. Their high computational efficiency allows multiple simulations exploring many combinations of drivers and internal feedback processes.

Ecosystem processes are therefore now beginning to be included in climate models. This new generation of climate system models provides great opportunities to investigate how ecosystems affect their own environmental conditions, and also how ecosystems interact with each other through the shared mediums of the atmosphere, oceans, and continental water flows.

RADIATIVE FORCING BY CHANGES IN ATMOSPHERIC COMPOSITION

Radiatively-forced Climate Change

Most recent work on climate projections has been centered around scenarios of greenhouse emissions and concentrations published by the Intergovernmental Panel on Climate

Change (IPCC) in the Special Report on Emissions Scenarios (SRES). These scenarios are based on a number of narratives characterizing global demographic, economic, and technological changes over the twenty-first Century, and provide a range of potential future trajectories for greenhouse-gas emissions. For example, the scenarios of CO_2 changes range from small decreases by 2100 to increases of around 400%. When these are translated into atmospheric concentrations, with feedback effects of climate change not considered, CO_2 concentrations (currently at approximately 370ppmv) are projected to rise to between 500 and 950 ppmv by 2100.

A large number of GCMs have been used to simulate climate change under scenarios of anthropogenic greenhouse-gas changes. GCMs vary in their sensitivity to a given radiative forcing, and in conjunction with the large number of scenarios this results in a wide range of results from the climate change simulations (Figure 1). While all models simulate a rise in global mean temperatures due to increasing greenhouse-gas concentrations, the extent of this rise ranges from 0.9 to 5.8 Ks relative to the preindustrial climate. However, despite the differences, the results of these simulations still have a number of key features in common. For example, the models show more rapid warming over the land surface than over the oceans, owing to the large heat capacity of water (Figure 2). The models also project more rapid warming at high latitudes than at low latitudes, owing to positive feedback on the warming arising from the melting of snow and ice. While snow and ice reflect a large proportion of solar radiation back to space, the underlying ground and ocean surface is generally darker and absorbs a greater fraction of the radiation. Melting of snow and ice therefore exposes the darker underlying surface, increasing the fraction of solar radiation absorbed and causing a further warming of these regions.

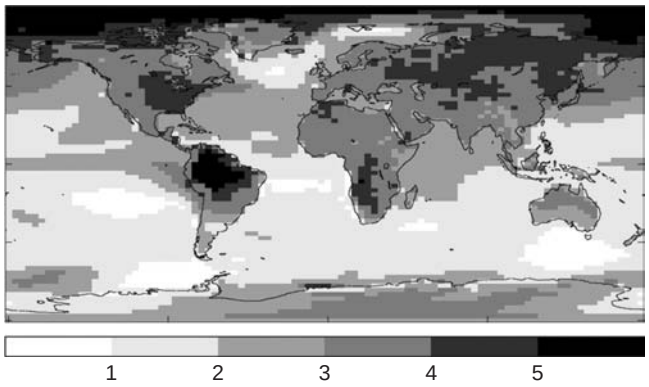


Figure 2 Changes in temperature (K) relative to 2000 simulated by the HadCM3LC GCM. 30-year mean centered around 2080. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

The different models show much less agreement on changes in precipitation with rising GHG concentrations, although there is still some general agreement at large scales. Global mean precipitation is generally simulated to increase, as a result of an increased supply of moisture to the atmosphere by evaporation from the oceans and continents in a warmer world (Figure 3). Regional-scale differences, however, differ greatly from model to model, with the differences often being in sign as well as magnitude

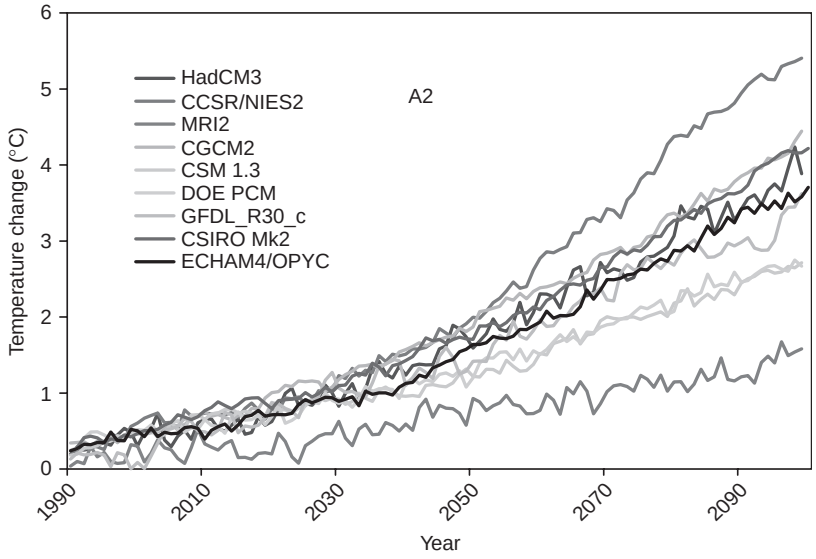


Figure 1 Projections of global mean near-surface air temperature rise by 9 different General Circulation Models (GCMs) under the same scenario of greenhouse-gas concentration changes. The scenario is SRES A2 (IPCC, 2001) (Reproduced by permission of Intergovernmental Panel on Climate Change (IPCC)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

(Figure 4). These variations between models are often due to differences in the simulated changes in atmospheric circulation. Demonstration of predictive skill in local precipitation remains one of the greatest challenges for GCMs.

One notable result in some models, but not all, is a reduction in precipitation over Amazonia (Figure 3), associated with more rapid warming of the ocean surface in the equatorial east pacific. This pattern of warming in the Pacific ocean is reminiscent of an El Nino event (Cai and Whetton, 2001), during which precipitation is also reduced over Amazonia. This increased warming is in a region of the ocean that is beneath a descending portion of the Hadley-Walker large-scale atmospheric circulation cell, and the warming ocean surface tends to oppose descending motion, which suppresses the overall circulation leading to less ascent of the atmosphere over Amazonia. Reduced ascent implies reduced convection and hence reduced precipitation, with crucial implications for this region, which is dependent on its currently high rainfall rate.

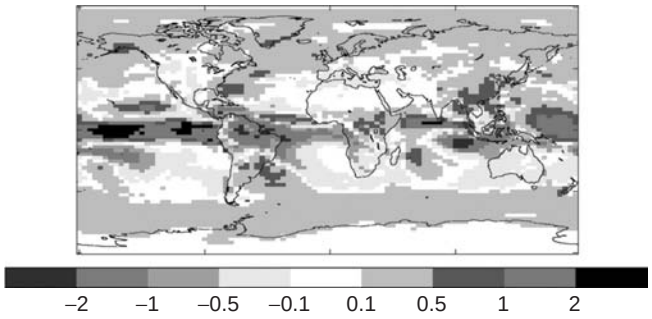


Figure 3 Changes in precipitation (mm day^{-1}) relative to 2000 simulated by HadCM3LC. 30-year mean centered around 2080. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Effects of Radiatively forced Climate Change on Global Ecosystems

Most studies of the impacts of radiatively forced climate change on global ecosystems also consider the effects of CO_2 fertilization. To gain insight into the roles of the two forcings, many such studies also consider the impacts of each forcing separately. This is useful as it allows the implications of climate change alone to be assessed, irrespective of any other properties of the gases responsible for the change. The results therefore apply to a given climate state whether it is caused by changes in CO_2 or any other greenhouse gas. This is important because other greenhouse gases are also increasing in concentration, and not all of these will also exert a fertilization effect.

Haxeltine (1996) and Betts *et al.* (2000) performed sensitivity studies with biogeography models under climate

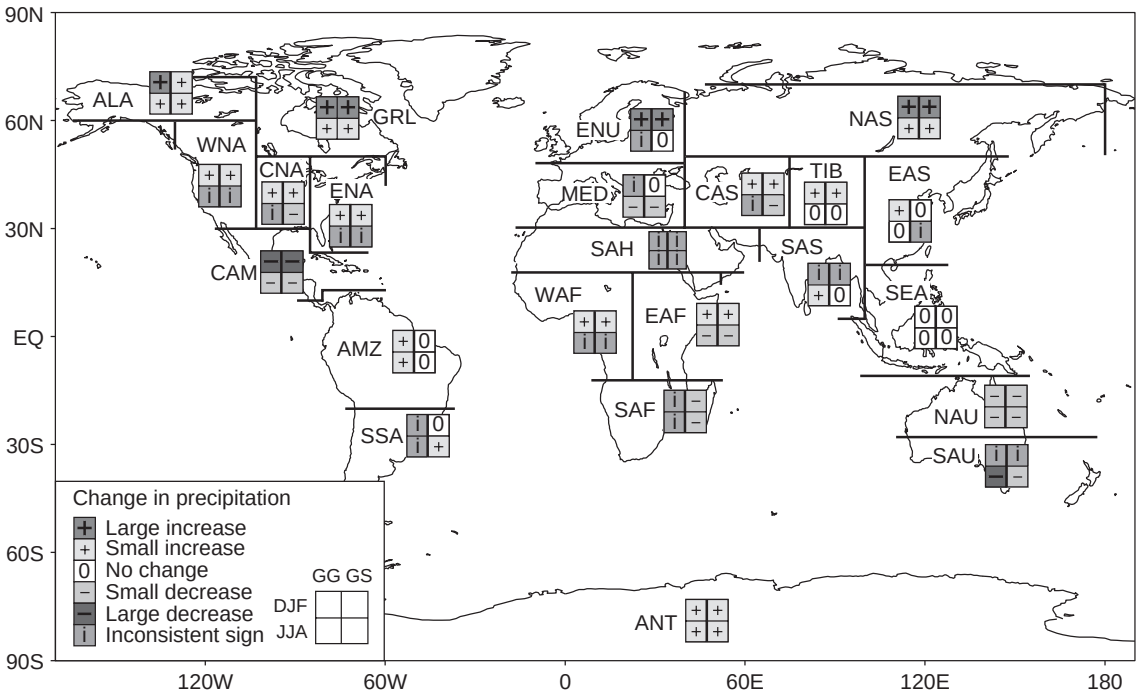


Figure 4 Illustration of differences in precipitation changes over the twenty-first Century simulated by different GCMs (Reproduced by permission of Intergovernmental Panel on Climate Change (IPCC)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

changes simulated by GCMs with greenhouse-gas radiative forcing doubled relative to the present-day. In both cases simulations both with and without CO_2 fertilization were performed. In a simulation without CO_2 fertilization, the BIOME3 model simulated and expansion of the boreal forests northwards into areas currently occupied by tundra (Haxeltine, 1996). A similar simulation with the DOLY model produced an increase in leaf area index, and further simulation showed that this was due mainly to the warmer temperatures (Betts *et al.*, 2000). As well as a general increase in high-latitude tree cover, BIOME3 simulated an increase in the ratio of evergreen needleleaf to deciduous needleleaf trees, and an increase in the ratio of broadleaf to needleleaf trees. This again reflects a warming climate.

BIOME3 simulated a decrease in forest cover in parts of the tropics, for example, central Africa, attributed to increased temperature and water stress favoring C4 grasses over C3 woody plants. DOLY simulated increases in LAI with temperature in moister regions, and decreases with temperature in drier regions. LAI was simulated to decrease in areas where precipitation decreased.

The global mean net primary productivity simulated by BIOME3 increased by 6% relative to the present-day as a result of climate change. It was found that the climatically induced redistribution of vegetation was necessary to realize this increase in NPP increase; when global vegetation patterns were fixed at the present-day state, global mean NPP increased by only 1%.

Cramer *et al.* (2001) used 6 DGVMs to make projections of global vegetation responses to transient climate change simulated with the HadCM2 GCM under the IS92a greenhouse-gas and sulphate aerosol concentration scenario. Vegetation simulations were performed with and without CO_2 fertilization. The DGVMs showed global vegetation responses to climate change alone, which were similar to those simulated by the biogeography models described above, with increasing forest cover in temperate and boreal regions but less in the tropics (Figure 5). A particular feature of this study was a major reduction in forest cover in the eastern half of Amazonia, as a consequence of a significant reduction in precipitation and increase in temperature.

The DGVMS simulated a range of changes in global mean NPP under climate change, some increasing and some decreasing. A key influence on decreasing NPP was the drying climate simulated in Amazonia. The DGVMs also simulated increases in soil respiration under the warming climate, which generally offset any increases in NPP. The simulated net atmosphere-land carbon flux was therefore negative in most regions in most models throughout the twenty-first Century, suggesting an almost universal release of carbon from terrestrial ecosystems as a consequence of climate change alone.

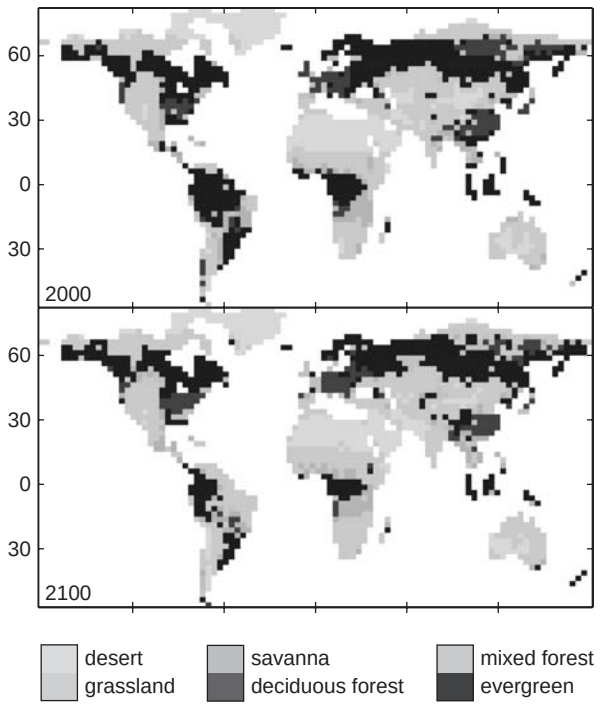


Figure 5 Dominant vegetation types simulated for 2000 and 2100, considering only changes in climate and neglecting CO_2 fertilization effects. Each data point shows the mode result of 6 Dynamic Global Vegetation Models (DGVMs). (Reproduced from Cramer *et al.*, 2001, with permission of Blackwell Science Ltd). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Effects of Radiatively Forced Climate Change on Global Hydrology

Many studies of the impacts of climate change on surface hydrology use catchment-scale models driven by GCM or RCM output, processed to correct biases and improve regional detail. When such models are applied to catchments in which a significant fraction of the precipitation currently falls as snow, winter runoff is generally simulated to increase under climate change while spring runoff decreases. This is due to the warmer climate causing more winter precipitation to fall as rain rather than snow. Rainwater contributes to runoff very rapidly, whereas snow is effectively stored until the spring thaw and only then does the water contribute to runoff. The exception to this is very cold regions where the warming is not sufficient to cause a significant shift from snow to rain.

In other regions, changes in runoff generally reflect changes in precipitation, with areas receiving more precipitation producing greater quantities of runoff. However, increases in temperature also exert an important influence in some regions, leading to increases in evaporation that are significant enough to offset the increases in precipitation

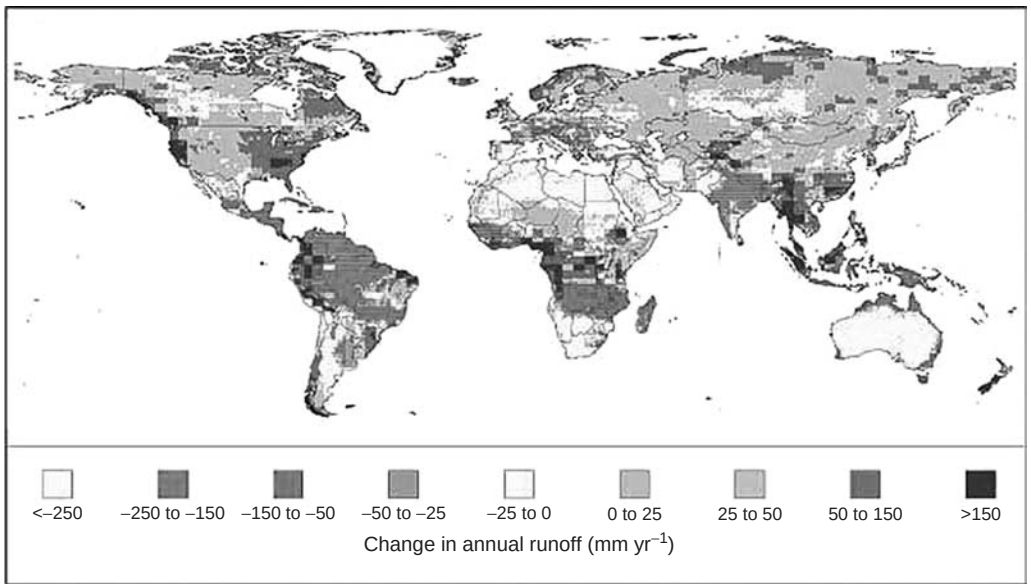


Figure 6 Changes in annual mean runoff by relative to the present day, simulated by a macroscale hydrology model under climate changes simulated by the HadCM2 GCM. Arnell, 1999. (Reproduced by permission of Intergovernmental Panel on Climate Change (IPCC)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

and hence cause a decrease in runoff. However, the local characteristics of catchments can also exert an important influence. For example, if groundwater forms a significant component of the catchment water budget, runoff in summer may be affected predominantly by precipitation the previous winter.

Arnell (1999) used a global-scale hydrological model with climate model projections from the HadCM2 GCM (Figure 6). Under the simulated 2050s climate, annual mean runoff was simulated to increase in most of the high and midlatitudes as a result of increases in precipitation, except in Europe and central Siberia where the warming caused sufficient increases in evaporation to offset precipitation increases. Increased runoff was also simulated in China and central Africa, again as a result of increased precipitation, but in large parts of the subtropics runoff was simulated to decrease. In some of these areas this is again a result of the warming increasing evaporation by more than the increase in precipitation, but in northeastern Amazonia and Southwest Africa the decreased runoff results from reductions in precipitation. Global mean runoff was simulated to increase by 6.5%.

The DGVMs used by Cramer *et al.* (2001) included hydrological submodels as part of the ecosystem models. When driven by output of the HadCM2 under the IS92a concentration scenario, without any direct effect of CO_2 on transpiration but with transpiration being modified by ecosystem responses to climate change, these models simulated global mean runoff changes from -1 to $+9\%$ by 2100 relative to the present.

PHYSIOLOGICAL FORCING OF THE CLIMATE SYSTEM BY CARBON DIOXIDE

Effects of Physiological Forcing on Ecosystems

Projections of the direct effects of rising CO_2 on vegetation are generally made using models incorporating mechanistic descriptions of plant physiological processes, based on individual-level observational and experimental studies. In particular, photosynthesis is generally observed to increase with rising CO_2, but the rate of this increase reduces with rising CO_2, implying a saturation of photosynthesis at high CO_2. Pot and plot-scale studies also indicate an increased efficiency of water use due to the reduced opening of stomata under higher CO_2. Biogeography models and DGVMs generally adopt some methodology for scaling up of these processes to generate projections at large scales.

Model simulations of global vegetation responses to increasing CO_2 (neglecting any associated climate change) show an increase in net primary productivity (NPP) and carbon storage. Cramer *et al.* (2001) simulated global mean NPP and carbon stocks to nearly double by 2100 when CO_2 effects were included without climate change. These simulations also featured some changes in global vegetation distribution, mainly due to increased water-use efficiency allowing more growth in semiarid regions. The African and South American equatorial forests expanded into savanna regions, and the relative viability of woody vegetation also increased in India. Australia became more grass-covered, and the boundary between the Sahara desert and the Sahel moved northwards.

Enhancement of photosynthesis by increased CO_2 plays an important part in the response of the climate system to anthropogenic carbon emissions. Along with increased uptake in the oceans, this increased uptake of CO_2 by enhanced photosynthesis provides a partial buffer against the anthropogenic emissions, such that the atmospheric CO_2 concentration currently rises at approximately only 50% of the rate of emissions (IPCC, 2001). This "airborne fraction" of emissions is a crucial component of projections of future CO_2 rises, and any changes in this fraction could significantly influence the rate of CO_2 rise. A saturation of photosynthesis at high CO_2 would contribute to an increase in the airborne fraction. This has major implications for climate change, as the current 50% airborne fraction may be providing only a temporary buffering of the full effect of anthropogenic CO_2 emissions.

Effects of Physiological Forcing on Hydrology

Reduced transpiration also implies an increase in the proportion of water remaining at the land surface. This implies an increase in runoff as a consequence of the physiological impact of rising CO_2 on vegetation (Wigley and Jones, 1985).

In simulations with vegetation-hydrology models under the IS92a scenario of CO_2 increase, ignoring any associated radiatively forced climate change, Cramer *et al.* (2001) found all but one of their models to simulate an increase in global runoff owing to the physiological effects of CO_2 rise. The simulated changes in runoff ranged from -3 to $+47\%$. It should be noted that the simulations included dynamic vegetation, so the physiological forcing could also exert an impact through changes in leaf area and vegetation distribution. This is important, because such changes in vegetation structure may affect climate through changes in land surface properties as well as through stomatal closure. The changes in runoff therefore arose from physiological impacts on plant stomata, leaf area, and vegetation distribution.

The Section "Effects of radiatively forced climate change on global hydrology" noted that Cramer *et al.* (2001) simulated a change in runoff of -1 to $+9\%$ with the DGVMs driven by climate change without CO_2 physiological forcing. When both climate change and physiological forcing were included, the changes in runoff were $+1$ to $+45\%$. This suggests that rising CO_2 could increase global mean runoff more through physiological forcing of transpiration than radiatively forced climate change. In regions where radiatively forced climate change does not significantly increase local precipitation, increased runoff may still occur as a result of physiological forcing.

Effects of Physiological Forcing on Climate

Effects of Physiological Forcing on Surface Temperature

The rate of transpiration influences the partitioning of energy fluxes between latent heat and sensible heat, and reduced transpiration implies a smaller latent heat flux and larger sensible heat flux, which will warm the air near the surface. If stomata open less wide under higher CO_2 concentrations, this would be expected to exert a further warming influence on climate in addition to that associated with radiative forcing. In a doubled-CO_2 sensitivity study with the SiB ecosystem model coupled to a GCM, Sellers *et al.* (1996) found the simulated surface air temperature in the tropics to increase by $0.4-0.7$K as a result of physiological forcing alone. By comparison, the warming was 1.7 K as a result of radiative forcing alone. So while physiological forcing exerted a smaller influence than the radiative forcing, the relative effect was not insignificant.

In a similar experiment with the HadAM3 GCM incorporating the MOSES ecosystem model, Cox *et al.* (1999) found global land temperatures to rise by 3.1 K owing to radiative forcing alone, and additional 0.4 K when physiological forcing was included. The land climate sensitivity to doubling CO_2 was therefore increased by 12% by the inclusion of physiological forcing.

Effects of Physiological Forcing on Precipitation

A decrease in transpiration also implies a decreased return of moisture to the atmosphere. In regions such as Amazonia where a large proportion of water for precipitation is supplied by evaporation from the land surface rather than advection from over the oceans (Salati and Vose, 1984), reduced transpiration by stomatal closure could lead to reduced precipitation. Betts *et al.* (2004) examined the possible effect of this in Amazonia, using a coupled climate-biosphere model. Two simulations were performed, both with increasing radiative forcing from CO_2 rise according to the IS92a scenario, but with only one simulation including CO_2-induced stomatal closure.

The two simulations showed marked differences in precipitation in many regions across the globe by the end of the twenty-first Century (Figure 7). Over Amazonia, for example, the simulation with stomatal closure featured a more drastic decrease in precipitation. In western Amazonia, this was due to a reduced evaporative return of water to the atmosphere through reduced canopy conductance (Cox *et al.*, 1999).

However, there were also large-scale patterns of precipitation increase and decrease across the globe, indicative of differences in the global atmospheric circulation. These are attributable to different rates of land surface warming in the two simulations as described above. In the simulation including CO_2-induced stomatal closure, the generally lower rate of transpiration leads to greater warming over

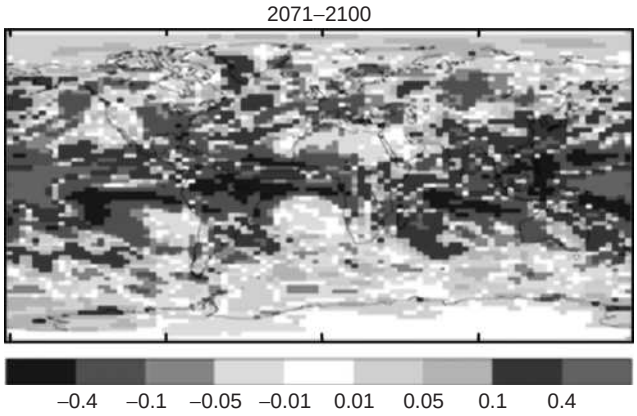


Figure 7 Effect of physiological forcing (reduced stomatal opening under higher CO_2) on global precipitation patterns. Difference in precipitation (mm day^{-1}) between simulations with and without physiological forcing. 30-year mean centered around 2080. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

land (Sellers *et al.*, 1996; Betts *et al.*, 1997; Cox *et al.*, 1999). Since the northern hemisphere includes more land than the southern hemisphere, a greater rate of warming over land leads to an interhemispheric difference in warming that would be expected to modify global atmospheric circulation. In particular, the Inter-Tropical Convergence Zone (ITCZ), an area of ascending air motion and hence precipitation, is drawn northwards by a warmer northern hemisphere. This changed the patterns of precipitation around the equator, bringing more rainfall to some regions but less to others.

LAND USE FORCING OF THE CLIMATE SYSTEM

The nature of vegetation cover is a major influence on the physical properties of the land surface. In general, a forested landscape exhibits a lower surface albedo, a larger aerodynamic roughness and a greater availability of moisture for evaporation than a nonforested landscape. Trees significantly increase evaporation by drawing-up soil moisture and transpiring it through their leaves, and also by enhancing turbulent transport by increasing the aerodynamic roughness of the landscape. Removal of forest cover would therefore be expected to lead to decreased evaporation and a reduction in the absorption of solar radiation. When resulting from a large-scale change, these changes in land surface properties can significantly influence regional climates and wider-scale atmospheric circulations. Since the extent of forest cover is changing rapidly around the world, especially in the tropics, which are now generally being deforested, this is another area in which long-term projections of climatic, ecological, and hydrological states are required.

A large number of modeling studies have examined the climate sensitivity to total deforestation in tropical regions such Amazonia. There is general agreement among these studies that complete deforestation would cause a warming of surface temperature and reduction in precipitation, owing to a reduced level of transpiration from the deforested landscape (*see*, for example, Lean and Rowntree, 1997 for summary). The smaller flux of moisture due to reduced transpiration causes a reduction in the ratio of latent to sensible heat fluxes, so the air near the surface is warmed. Since much of the rainfall in the Amazon basin relies on water transported from over the oceans through the repeated cycling of water through rainfall and evaporation (Salati and Vose, 1984), the reduced transfer of moisture to the atmosphere also decreases the recycling of moisture across the continent. Less moisture is therefore available for precipitation in the center and west of the Amazon basin. In addition, an increase in surface albedo results in a smaller heating of the surface by the net radiation balance, suppressing ascent and hence causing less moisture to be drawn into the region (Charney, 1975). A reduced frictional drag exerted by the deforested landscape will also act to reduce this moisture convergence. All these mechanisms will tend to reduce precipitation (Figure 8).

Large-scale deforestation in the tropics may also exert more far-reaching effects, with changes in the near-surface

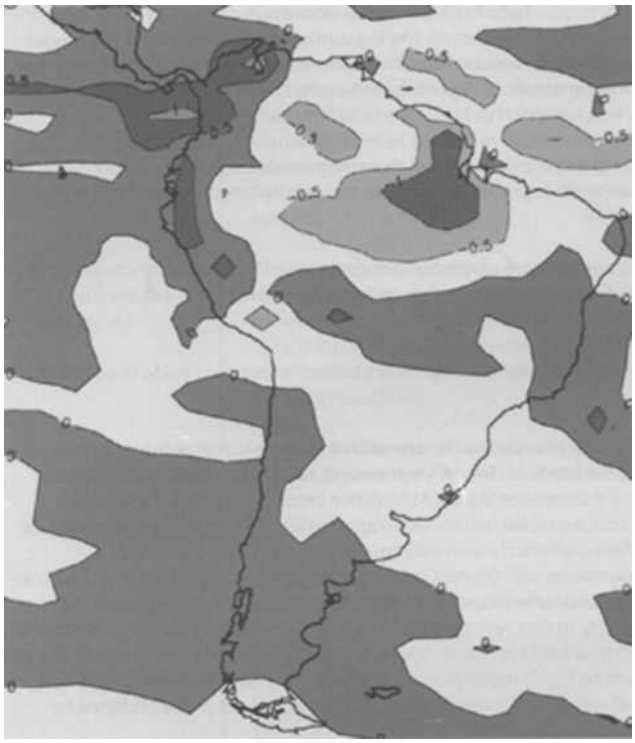


Figure 8 Simulated changes in annual mean precipitation as a result of complete deforestation of Amazonia (Lean and Rowntree, 1993). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

energy balance modifying the large cells of ascending and descending motion near the equator. Changes to these may cause shifts in the atmospheric circulation that propagate across the globe. Gedney and Valdes (2000) found that complete deforestation of Amazonia could lead to increased precipitation in Europe through this mechanism.

Feedback from the oceans may enhance the effects of land-cover change on climate. For example, Delire *et al.* (2001) found that in the Indonesian Archipelago, the impacts of deforestation on wind speeds may be sufficient to modify ocean up-welling and cause warming over the surrounding ocean surfaces, in addition to the warming caused over land by reduced evaporation. This amplified warming may impact global-scale atmospheric circulations.

Since changes in land cover such as deforestation have been shown to potentially impact evaporation and precipitation, it follows that such activities would also impact runoff. Lean and Rowntree (1993) found an 8% reduction in simulated runoff from Amazonia following complete deforestation, as a result of decreased precipitation outweighing the decrease in evaporation. However, surface runoff was simulated to increase by 27% at the expense of subsurface runoff, because infiltration rates were assumed to decrease after deforestation. Infiltration rates in forest soils were assumed to be high owing to the presence of litter and tree root systems, while infiltration in land deforested for pasture was assumed to be low owing to soil compaction during the deforestation process and by the presence of livestock.

The issue of partial deforestation may, however, be more complex. High-resolution mesoscale models that resolve the patterns of partial deforestation suggest that the fragmentation of the forest cover may induce small-scale atmospheric circulations that can increase atmospheric ascent and therefore enhance convection (Roy and Avissar, 2002). Therefore, partial deforestation may actually increase precipitation if the supply of moisture from advection remains unchanged. With current computing limitations, atmospheric models with resolution high enough to capture the effects of fine-scale forest fragmentation can only be applied over relatively small regions. Since actual patterns of deforestation often consist of fine-scale partial clearance over large areas, there is a need to investigate the interactions between mesoscale circulations and large-scale advection and evaporative recycling. There is therefore a requirement for large-domain modeling studies with either higher resolutions or adequate parameterizations of fine-scale circulations.

While tropical forests exert a cooling effect on their regional climates, forests in cold regions exert a warming influence through their large impact on surface albedo, which outweighs the influence of transpiration. Forests are generally darker than open land, particularly when snow is lying because trees generally remain exposed while cultivated land can become entirely snow-covered (Harding and Pomeroy, 1996). Snow-free foliage is considerably darker than snow, but even if large quantities of snow are held on the canopy, multiple reflections within the canopy scatter rather than reflect shortwave radiation, which also reduces the landscape albedo (Harding and Pomeroy, 1996). Models suggest that the resulting low surface albedo causes boreal and cool-temperate forests to exert a warming influence on climate relative to non-forested land (Bonan *et al.*, 1992; Thomas and Rowntree, 1992).

Deforestation of the boreal and temperate forests would therefore exert a cooling influence on climate (Douville and Royer, 1997), and indeed from the global perspective, most deforestation until the mid twentieth century had occurred in the temperate regions exerting an overall local cooling effect (Brovkin *et al.*, 1999; Betts, 2001; Govindasamy *et al.*, 2001). However, in more recent decades, land abandonment in western Europe and North America is leading to reforestation, which would cause a warming influence. This point is particularly important in the context of the potential use of afforestation and reforestation to slow radiatively forced climate change (UNFCCC, 1997). Since forest growth sequesters carbon dioxide from the atmosphere, this is one means by which the rise in atmospheric CO_2 could potentially be slowed providing a possible means of mitigating radiatively forced climate change. However, in the case of afforestation in cooler regions, the cooling influence of carbon sequestration could be partly or wholly offset by the warming influence of decreased surface albedo (Betts, 2000). Temperate and boreal afforestation and reforestation could therefore provide a less effective means of climate change mitigation, or even an unintentional means of exacerbating climate warming.

INTERACTIONS BETWEEN FORCINGS

The above discussion has considered three anthropogenic forcings of the climate system, and reviewed studies of them acting in isolation. However, in reality these would be expected to act together, so it is their combined effects that are of greatest interest when making projections of future changes in the climate system. While in some cases the combined effects may be similar to a simple sum of the individual effects, in other cases the forcings may interact nonlinearly causing more unexpected results.

The radiative and physiological forcings exerted by CO_2 rise may exert nonadditive impacts on local surface temperatures and precipitation through modifications of the atmospheric circulation. For example, Sellers *et al.* (1996) found the radiatively forced warming in high northern latitudes to be locally reduced when physiological forcing was included, as a consequence of circulation changes.

A major issue is the relative impact of radiatively forced climate change and CO_2 fertilization on terrestrial ecosystem, and in particular their total carbon storage. Section "Effects of radiatively forced climate change on global ecosystems" discussed the potential redistribution of ecosystem types and a projected release of carbon from global ecosystems as a result of a warmer climate, while Section "Effects of physiological forcing on ecosystems" discussed the projected increase in carbon storage due to physiological responses to increased CO_2 concentrations. A key question is therefore whether terrestrial carbon stocks will increase or decrease under scenarios of rising CO_2 and the associated climate change.

Haxeltine (1996) found that climate warming alone increased NPP by 6%, largely as a result of a net expansion of woody vegetation under climate change. The effect of CO_2 fertilization was to increase NPP by 30% in the absence of any forest expansion, but when climatically induced forest expansion was considered the overall increase in NPP was 43%. This suggests that the effects of climate change of biogeography may be a key factor in limiting NPP responses to increased CO_2 concentrations.

Considering the overall net exchange of carbon between the atmosphere and terrestrial biosphere, including soil respiration in addition to NPP, it would be expected conceptually that CO_2 fertilization would dominate initially and then be overtaken by the effects of warming, as CO_2 fertilization saturates at higher CO_2 concentrations while respiration is generally observed to increase with temperature. A number of model projections of ecosystem responses to scenarios of CO_2 rise and radiatively forced climate change are consistent with this expectation (e.g. Cao and Woodward, 1998; Cramer *et al.*, 2001), with the terrestrial biosphere simulated to be a sink of carbon initially but then becoming a weaker sink or even a carbon source by the end of the twenty-first century.

Cramer *et al.* (2001) found CO_2 fertilization and climate change to interact nonlinearly. By 2100, climate change alone generally led to a negative net atmosphere-land carbon flux or net ecosystem productivity (NEP), which is the difference between NPP and soil respiration. However, CO_2 increase alone led to positive NEP. The linear sum of these effects was a general increase in NEP, but in simulations driven by both CO_2 and climate change together, negative NEP was simulated in many parts of the tropics.

Betts *et al.* (2004) found the dieback of Amazonian forest under a drying climate to be reduced but not eliminated by CO_2 fertilization. This was despite the further reduction in local precipitation induced by stomatal closure. Physiological forcing therefore enhanced the drying of the climate but slowed the forest dieback due to the increased water-use efficiency.

ECOSYSTEM FEEDBACK ON CLIMATE CHANGE

Sections "Effects of radiatively forced climate change on global ecosystems" and "Effects of physiological forcing on ecosystems" showed that CO_2 rise and radiatively forced climate change may alter the character of global ecosystems. Conversely, Section "Land use forcing of the climate system" illustrated the influence of changes of the physical characteristics of ecosystems on the climate. Widespread increases or decreases in forest cover projected in response to CO_2 rise and climate change may therefore make further additional contributions to the regional and global climate changes through alterations to the land surface properties. Furthermore, the net changes in global carbon stocks discussed in Section "Interactions between forcings" may influence the rise in CO_2 itself. Ecosystems may therefore exert a number of feedback on climate change, both at the regional and global scale.

Feedback via the carbon cycle have been investigated by Cox *et al.* (2000), Friedlingstein *et al.* (2001) and more recently by other groups (e.g. Thompson *et al.*, 2004). Since these studies were similar in methodology and produced broadly similar results, at least in terms of the feedback on atmospheric CO_2 rise, only the study by Cox *et al.* (2000) will be discussed here in detail.

Cox *et al.* (2000) used the HadCM3LC coupled climate-carbon cycle model, which includes the TRIFFID DGVM (Cox, 2001) and HadOCC ocean carbon cycle model (Palmer and Totterdell, 2001) in a version of the HadCM3 GCM (Gordon *et al.*, 2000). Atmospheric CO_2 concentrations were calculated interactively within then model as opposed to being prescribed from an externally derived scenario.

A simulation of climate and ecosystem change from 1860 to 2100 was performed, with the input perturbation being the CO_2 emissions from the IS92a scenario. This contrasted with the usual approach to GCM climate simulations, which use CO_2 concentrations rather than emissions as the input driving data. Terrestrial ecosystems and the ocean carbon cycle to respond to the CO_2 rise and climate change and feed back on the CO_2, so the model simulated the changes in both climate and CO_2 simultaneously and consistently.

A second simulation was performed with the interactive carbon cycle but with the radiative effect of CO_2 excluded from the calculations. This therefore allowed the model to simulate the changes in CO_2 due to the anthropogenic emissions and the responses of the oceans and terrestrial biosphere to the CO_2 rise, but without any climatic effect inducing a feedback. Comparison between this simulation and the first simulation quantified the effect of climate change on the atmospheric CO_2 rise.

A third simulation was driven by prescribed CO_2 concentrations rather than emissions, with the CO_2 concentrations calculated without the effect of climate change as in

simulation 2. Comparison between this and the first simulation therefore showed the importance of the climate-carbon cycle feedback on the climate change itself.

In the second simulation, neglecting the effects of climate change on the carbon cycle, the terrestrial biosphere was projected to take up 620 Gigatonnes of carbon (GtC) by 2100, two-thirds of this in the soil. In contrast, the first simulation (including the effects of climate change) projected an overall loss of 90 GtC, owing to large losses from the soil and smaller uptake in vegetation (Figure 9a) Climate change therefore induced a deficit of 710 GtC in the terrestrial biosphere in comparison with the changes projected without the effects of climate change. With CO_2 uptake by the oceans also considered, this feedback resulted in CO_2 concentrations rising to 1000 ppmv by 2100 rather than 700 ppmv as projected without the feedback (Figure 9b).

This feedback was found to significantly increase the rate of global warming. In the third simulation, with a CO_2 rise of only 700ppmv, the global mean temperature rise over land was 5 K, whereas in the first simulation with all feedback included the warming over land was 8 K (Figure 9c) The inclusion of climate-carbon cycle feedback was therefore to enhance the rate of global land warming by 50%. This magnified the patterns of precipitation change across the world, with regions of drying generally become even drier while regions of wetting became wetter still (Figure 10).

Friedlingstein *et al.* (2001) and Thompson *et al.* (2004) also found the atmospheric CO_2 rise and global warming to be accelerated by positive feedback between climate change and the terrestrial biosphere. However, in contrast to Cox *et al.* (2000), the models used in these two studies simulated the terrestrial biosphere to merely weaken as a sink of carbon rather than actually becoming a source of carbon. The positive feedback were therefore weaker than that found by Cox *et al.* (2000).

As well as changes in carbon stocks, the simulations also featured changes in the character and distribution of global ecosystems (Figure 11) As the climate warming accelerated in the cold regions, shrub cover continues to expand in the arctic tundra and on the Tibetan plateau. The boreal forests continue to become more dense, and also spread towards the pole. In gridboxes on northern edge of boreal forest, tree cover becomes more dense, implying a northward movement of the treeline. A number

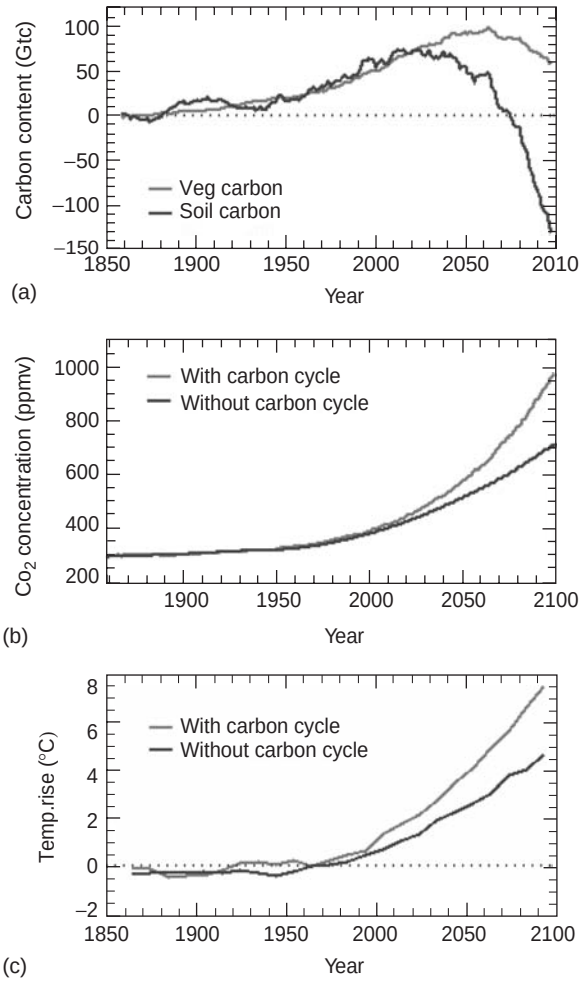


Figure 9 Results from a coupled climate-carbon cycle simulation with HadCM3LC. (a) Projected changes in global terrestrial carbon stores in a simulation including climate-carbon cycle feedbacks, relative to 1850. (b) Projected changes in atmospheric CO_2 concentration with and without climate-carbon cycle feedbacks. (c) Projected changes in global land temperatures with and without climate-carbon cycle feedbacks. Cox *et al.* (2000). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

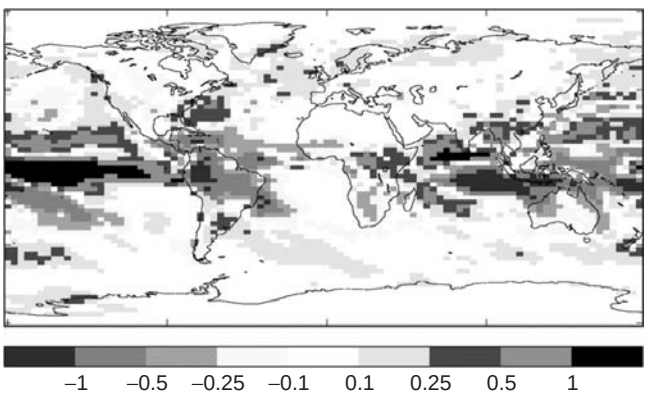


Figure 10 Effect of including carbon cycle feedbacks on global precipitation patterns in HadCM3LC. Difference in precipitation (mm day^{-1}) between simulations with and without carbon cycle feedbacks. 30-year mean centered around 2080. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

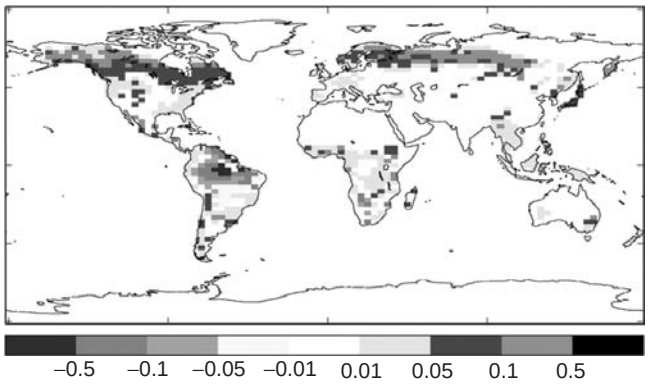


Figure 11 Changes in fractional cover of trees to 2000 simulated by HadCM3LC. 30-year means centered around (a) 2050 and (b) 2080. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

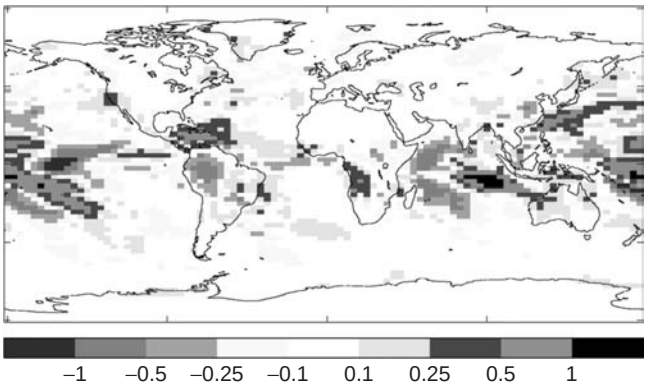


Figure 12 Effect of biogeophysical feedback on global precipitation patterns in HadCM3LC. Difference in precipitation (mm day^{-1}) between simulations with and without biophysical feedbacks. 30-year mean centered around 2080. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

of dry regions such as central Asia showed a greening trend, consistent with increased water-use efficiency but also partly a result of increased rainfall. Other dry regions such as Southwest Africa continue to lose tree and shrub cover – a decrease in rainfall is simulated in SW Africa. The forests of Southeast Asia and central Africa increased in density in the absence of future deforestation. However, the forests of Amazonia showed a very large reduction in tree cover as a result of decreased rainfall. Some signs of the beginning of this process were already simulated by 2000, with broadleaf tree cover reducing in the northeast of Amazonia in response to a drier climate than that simulated for 1860. The reduction in rainfall spreads towards the southwest through the twenty-first century, and the tree cover reduced until it is less than 1% in the northeast quarter of Amazonia by 2100. Almost all of the Amazon basin loses at least 50% of its tree cover by the end of the simulation, to be replaced mainly by C4 grass but also with large areas of bare soil. The general character of the region is therefore fundamentally changes from dense evergreen broadleaf forest to savanna, grassland, or even semidesert.

These changes in terrestrial ecosystems in these simulations exerted some significant feedback on regional climates through changes in the physical properties of the land surface. To examine the extent of these feedback, Betts *et al.* (2004) performed a fourth simulation with vegetation fixed at the preindustrial state, driven by the same CO_2 concentration scenario as the third simulation above. Comparison between the third and fourth simulations therefore reveals the extent of biogeophysical feedback on climate change.

The general global patterns of climate change were similar in the two simulations, with almost all changes in temperature and precipitation being of the same sign irrespective of the inclusion of vegetation feedback. This implies that vegetation feedback are not a significant influence on atmospheric circulation in comparison with the greenhouse-gas forcing. However, some of the regional climate changes were significantly affected by vegetation feedback (Figure 12) In particular, the precipitation reduction over Amazonia was found to be enhanced by 25% by feedback from the loss of forest cover. In the western part of the basin, the feedback was greater still, magnifying the precipitation reduction by over 30%. The larger precipitation decrease in western Amazonia was attributed to drought-induced dieback of the eastern forests contributing to further rainfall reductions in the west. The forest loss also increased surface albedo, which reduced convection and moisture convergence, providing a further positive feedback on rainfall reduction (Charney, 1975).

The expansion of the boreal forests and tundra shrub cover exerted a further feedback on climate warming. The more extensive vegetation cover resulted in a lower surface albedo, which increased the absorption of solar radiation. This provided a further 1 K of warming in addition to the 6 K simulated with fixed vegetation cover, a feedback of approximately 15%.

Friedlingstein *et al.* (2001) also found a reduction in precipitation to be simulated in Amazonia in their model, but the model did not include dynamic vegetation so there was no feedback on climate through biogeophysical effects. The model used by Thompson *et al.* (2004) included the IBIS2 dynamic vegetation model (Foley *et al.*, 1996; Kucharik *et al.*, 2000), but this model did not produce a drying in Amazonia. The carbon cycle and biogeophysical feedback simulated by HadCM3LC (Cox *et al.*, 2000; Betts *et al.*, 2004) are therefore more extreme than those simulated by other modeling groups. However, HadCM3LC has been shown to successfully reproduce the historical CO_2 rise and the observed interannual relationship between Pacific sea surface temperatures, variability in the CO_2

rise and precipitation in Amazonia (Jones *et al.*, 2001; Cox *et al.*, 2004).

UNCERTAINTIES IN PROJECTIONS: THE IMPLICATIONS OF SYNERGISMS AND FEEDBACK

Many of the projected changes discussed here have major implications for both human and ecological communities. It is therefore vital to consider the reliability of these model projections as indicators of actual future change. Such a consideration requires an understanding and of the processes involved in the changes in the climate system, so that the implications of uncertainties in the different components of the system can be recognized.

Uncertainties in all components of the climate change process are recognized. Uncertainties exist in the scenarios of socioeconomic change, the translation of these into GHG emissions scenarios, the subsequent responses of GHG concentrations, the implications of these for global and regional climates and the impacts on ecological and hydrological systems. A typical approach to quantifying uncertainty is to make projections with a large number of models, which are assumed to sample the range of plausible representations of processes. The IPCC (2001) illustrated this with a number of GCMs simulating climate changes under a given scenario of GHG concentrations. Global mean temperature was simulated to increase by between 1.3 and 4.5 K by 2100 under this scenario. When this was extended to more scenarios, with both GCMs and simple models approximating the GCM responses, the projections of global warming ranged from 0.9 to 5.8 K by 2100.

Uncertainty is typically viewed as increasing along the chain from socioeconomic scenarios to climate impacts, as each component in the process is subject to its own uncertainties and also those in the processes to which it responds. For example, while uncertainties in regional climate change provide a major source of uncertainty in projections of hydrological change (Arnell *et al.*, 2001), incomplete understanding of large-scale responses of transpiration to CO_2 introduces further significant uncertainties in potential changes in the surface water budget.

However, the potential existence of significant feedback within this chain may mean that the large uncertainties at the end of the chain propagate back towards the beginning. For example, it has been illustrated here that the regional climate changes simulated in some regions exert feedback on global climate change through the effects of forest dieback on atmospheric CO_2 rise. Since regional climate change predictions are subject to considerable uncertainty, the importance of the feedback to global climate implies that the uncertainties in global climate are also subject to uncertainties that depend on the regional-scale uncertainties. Betts *et al.* (2004), for example, showed that the uncertainties over possible Amazonian rainfall reductions lead to uncertainties in carbon cycle feedback, which imply a 10% uncertainty in global mean temperature rise.

Moreover, uncertainties in ecosystem responses to climate change may significantly contribute to uncertainties in the climate change itself. In particular, uncertainties in the response of soil respiration to warming (Jones and Cox, 2001) also lead to uncertainties in the rise of atmospheric CO_2 concentrations and hence global warming. The potential positive feedback found by Cox *et al.* (2000) suggest that uncertainties surrounding the temperature response of soil respiration imply a 50% uncertainty in the rate of global warming simulated by one climate model under a given emissions scenario. This is comparable with the uncertainties associated with the range of responses of different climate models to a given concentration scenario as described above, implying that the true uncertainty may be much greater owing to incomplete understanding of critical processes.

While this does not imply that the uncertainties in global climate change are greater than those in regional climate change or ecosystem response, it does imply that the uncertainties in the different components of climate change may be more similar than previously assumed. The "cascade of uncertainty" from higher confidence in emissions to lower confidence in impacts may not be as steep as formerly assumed.

SUMMARY AND CONCLUSIONS

Human-induced changes in atmospheric chemistry and land cover are projected to cause significant changes in climate, ecology, and hydrology around the globe. An enhancement of the greenhouse effect by increased concentrations of CO_2 and certain other gases is projected to increase global average temperatures by between 0.9 and 5.8 K by 2100. Warming in some regions may well be greater. Precipitation patterns are also projected to change, causing some regions to be drier and others wetter.

Direct physiological responses of vegetation to increased CO_2 may constitute an additional, secondary, forcing of the climate system, especially from hydrological and ecological perspectives. This forcing acts directly on ecosystems, enhancing photosynthesis and therefore modifying growth and development. Plants are also observed to reduce the opening of their stomatal apertures under increased CO_2, which reduces the loss of water by transpiration. Changes in transpiration at large scales would influence runoff, precipitation, and surface temperature.

Increased concentrations of CO_2 may therefore exert two forcings on the climate system, both resulting in impacts on ecosystems and hydrology. Radiative forcing affects ecosystems indirectly via climate change, while

physiological forcing may also affect plant functioning and the surface water budget directly through increased water-use efficiency and fertilization of photosynthesis.

Changes in land cover, such as the extent of forest cover, are projected to exert further impacts on climate through modifications to the surface energy and moisture budgets and the atmospheric circulation. Major changes in land cover are expected to occur as a direct result of human activities such as forest clearance. In addition, climate change may itself lead to shifts in the global patterns of ecosystems. As well as being critical to biodiversity and human well-being, such changes could exert significant feedback on climate change and potentially increase its magnitude.

Changes in climate, ecosystems, and hydrometeorology depend on complex interactions between the atmosphere, biosphere, and hydrological cycle. While numerical models remain the best tools for examining these interactions and assessing the implications of anthropogenic perturbations to the climate system, projections made with these models are nevertheless, subject to significant uncertainties. However, all models agree that significant changes, particularly a warming of climate and redistribution of rainfall and water resources, can be expected. This information in itself can provide a basis for preparing for the future.

Acknowledgments

The author thanks P. M. Cox and S. Sitch for valuable comments. The author's work forms part of the Climate Prediction Programme of the UK Department of Environment, Food and Rural Affairs (contract PECD 7/12/37).

REFERENCES

Arnell N., Chunzhen L., Compagnucci R., da Cunha L., Hanaki K., Howe C., Mailu G., Shiklomanov I. and Stakhiv E. (2001) Hydrology and water resources. *IPCC, 2001: Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Arnell N.W. (1999) Climate change and global water resources. *Global Environmental Change*, **9**, S31–S49.

Betts R.A. (2000) Offset of the potential carbon sink from boreal forestation by decreases in surface albedo. *Nature*, **408**, 187–190.

Betts R.A. (2001) Biogeophysical impacts of land use on present-day climate: near-surface temperature and radiative forcing. *Atmospheric Science Letters*, **2**, 39–51, doi:10.1006/asle.2000.0023.

Betts R.A., Cox P.M., Collins M., Harris P.P., Huntingford C. and Jones C.D. (2004) The role of ecosystem-atmosphere interactions in simulated Amazonian precipitation decrease and forest dieback under global climate warming. *Theoretical and Applied Climatology*, **78**, 157–175.

Betts R.A., Cox P.M., Lee S.E. and Woodward F.I. (1997) Contrasting physiological and structural vegetation feedbacks in climate change simulations. *Nature*, **387**, 796–799.

Betts R.A., Cox P.M. and Woodward F.I. (2000) Simulated responses of potential vegetation to doubled-CO_2 climate change and feedbacks on near-surface temperature. *Global Ecology and Biogeography*, **9**(2), 171–180.

Bonan G.B., Pollard D. and Thompson S.L. (1992) Effects of boreal forest vegetation on global climate. *Nature*, **359**, 716–718.

Brovkin V., Ganapolski A., Claussen M., Kubatzki C. and Petoukhov V. (1999) Modelling climate response to historical land cover change. *Global Ecology and Biogeography*, **8**, 509–517.

Brovkin V., Ganapolski A. and Svirezhev Y. (1997) A continuous climate-vegetation classification for use in climate-biosphere studies. *Ecological Modelling*, **101**, 251–261.

Cai W. and Whetton P.H. (2001) A time-varying greenhouse warming pattern and the tropical-extratropical circulation linkage in the Pacific Ocean. *Journal of Climate*, **14**, 3337–3355.

Cao M. and Woodward F.I. (1998) Dynamic responses of terrestrial ecosystem carbon cycling to global climate change. *Nature*, **393**, 249–252.

Charney J.G. (1975) Dynamics of deserts and droughts in the sahel. *Quarterly Journal of the Royal Meteorological Society*, **101**, 193–202.

Cox P.M. (2001) *Description of the TRIFFID Dynamic Global Vegetation Model*, Technical Note 24, Hadley Centre, Met Office.

Cox P.M., Betts R.A., Bunton C.B., Essery R.L.H., Rowntree P.R. and Smith J. (1999) The impact of new land surface physics on the GCM simulation of climate and climate sensitivity. *Climate Dynamics*, **15**, 183–203.

Cox P.M., Betts R.A., Collins M., Harris P.P., Huntingford C. and Jones C.D. (2004) Amazonian forest dieback under climate-carbon cycle projections for the 21st Century. *Theoretical and Applied Climatology*, **78**, 137–156.

Cox P.M., Betts R.A., Jones C.D., Spall S.A. and Totterdell I.J. (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, **408**, 184–187.

Cramer W., Bondeau A., Woodward F.I., Prentice I.C., Betts R.A., Brovkin V., Cox P.M., Fisher V., Foley J.A. Friend A.D., *et al.* (2001) Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. *Global Change Biology*, **7**(4), 357–374.

Delire C., Behling P., Coe M.T., Foley J.A., Jacob R., Kutzbach J., Liu Z.Y. and Vavrus S. (2001) Simulated response of the atmosphere-ocean system to deforestation in the Indonesian archipelago. *Geophysical Research Letters*, **28**(10), 2081–2084.

Douville H. and Royer J.-F. (1997) Influence of the temperate and boreal forests on the northern hemisphere climate in the Météo-France climate model. *Climate Dynamics*, **13**, 57–74.

Field C., Jackson R. and Mooney H. (1995) Stomatal responses to increased CO_2: implications from the plant to the global scale. *Plant Cell and Environment*, **18**, 1214–1255.

Foley J.A., Levis S., Prentice I.C., Pollard D. and Thompson S.L. (1998) Coupling dynamic models of climate and vegetation. *Global Change Biology*, **4**(5), 561–579.

Foley J.A., Prentice I.C., Ramankutty N., Levis S., Pollard D., Sitch S. and Haxeltine A. (1996) An integrated biosphere model of land surface processes, terrestrial carbon balance and vegetation dynamics. *Global Biogeochemical Cycles*, **10**(4), 603–628.

Friedlingstein P., Bopp L., Ciais P., Dufresne J., LeTreut H., Fairhead L., Monfray P. and Orr J. (2001) Positive feedback between future climate change and the carbon cycle. *Geophysical Research Letters*, **28**(8), 1543–1546.

Friend A.D., Stevens A.K., Knox R.G. and Cannell M.G.R. (1997) A process-based, terrestrial biosphere model of ecosystem dynamics (Hybrid v3.0). *Ecological Modelling*, **95**, 249–287.

Ganapolski A., Petoukhov V., Rahmstorf S., Brovkin V., Claussen M., Eliseev A. and Kubatzki C. (2001) CLIMBER-2: a climate system model of intermediate complexity. Part II: model sensitivity. *Climate Dynamics*, **17**, 735–751.

Gedney N. and Valdes P.J. (2000) The effect of Amazonian deforestation on the northern hemisphere circulation and climate. *Geophysical Research Letters*, **27**(19), 3053–3056.

Gordon C., Cooper C., Senior C.A., Banks H., Gregory J.M., Johns T.C., Mitchell J.F.B. and Wood R.A. (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dynamics*, **16**, 147–168.

Govindasamy B., Duffy P.B. and Caldeira K. (2001) Land use changes and northern hemisphere cooling. *Geophysical Research Letters*, **28**(2), 291–294.

Harding R.J. and Pomeroy J.W. (1996) The energy balance of the winter boreal landscape. *Journal of Climate*, **9**, 2778–2787.

Haxeltine A. (1996) *Modelling the Vegetation of the Earth*, PhD thesis, Lund University, Sweden.

IPCC (2001) *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment of the Intergovernmental Panel on Climate Change. Cambridge University Press: p. 881.

Jones C.D., Collins M., Cox P.M. and Spall S.A. (2001) The carbon cycle response to ENSO: a coupled climate-carbon cycle model study. *Journal of Climate*, **14**(21), 4133–4129.

Jones C.D. and Cox P.M. (2001) Constraints on the temperature sensitivity of global soil respiration from the observed interannual variability in atmospheric CO_2. *Atmospheric Science Letters*, **5**, 166–172, doi:10.1006/asle.2001.004.

Kucharik C.J., Foley J.A., Delire C., Fisher V.A., Coe M.T., Lenters J.D., Young-Molling C., Ramankutty N., Norman J.M. and Gower S.T. (2000) *Global Biogeochemical Cycles*, **14**(3), 795–825.

Lean J. and Rowntree P.R. (1993) A GCM simulation of the impact of Amazonian deforestation on climate using an improved canopy representation. *Quarterly Journal of the Royal Meteorological Society*, **119**, 509–530.

Lean J. and Rowntree P.R. (1997) Understanding the sensitivity of a GCM simulation of Amazonian deforestation to the specification of vegetation and soil characteristics. *Journal of Climate*, **10**(6), 1216–1235.

Palmer J.R. and Totterdell I. (2001) Production and export in a global ocean ecosystem model. *Deep-Sea Research*, **48**(5), 1169–1198.

Roy S.B. and Avissar R. (2002) Impact of land use/land cover change on regional hydrometeorology in Amazonia. *Journal of Geophysical Research (Atmospheres)*, **107**(D20), LBA 4-1–LBA 4-13.

Salati E. and Vose P.B. (1984) Amazon Basin: a system in equilibrium. *Nature*, **225**(4658), 129–138.

Sellers P.J., Bounoua L., Collatz G.J., Randall D.A., Dazlich D.A., Los S.O., Berry J.A., Fung I., Tucker C.J., Field C.B., *et al.* (1996) Comparison of radiative and physiological effects of doubled atmospheric CO_2 on climate. *Science*, **271**, 1402–1406.

Sitch S., Smith B., Prentice I.C., Arneth A., Bondeau A., Cramer W., Kaplan J.O., Levis S., Lucht W., Sykes M.T., *et al.* (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, **9**(2), 161–185.

Thomas G. and Rowntree P.R. (1992) The boreal forests and climate. *Quarterly Journal of the Royal Meteorological Society*, **118**, 469–497.

Thompson S.L., Govindasamy B., Mirin A., Caldeira K., Delire C., Milovich J., Wickett M. and Erickson D. (2004) Quantifying the effects of CO_2-fertilized vegetation on future climate and carbon dynamics. *Geophysical Research Letters*, **31**(23), doi:10.1029/2004GL021239.

UNFCCC (1997) *Kyoto Protocol to the United Nations Framework Convention on Climate Change*, FCCC/CP/1997/L.7 Add.1.

Wigley T.M.L. and Jones P.D. (1985) Influences of precipitation changes and direct CO_2 effects on streamflow. *Nature*, **314**(6007), 149–152.

182: The Hydrological Cycle in Atmospheric Reanalysis

ANTON BELJAARS, ULF ANDRAE, PER KALLBERG, ADRIAN SIMMONS, SAKARI UPPALA AND PEDRO VITERBO

Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Atmospheric (re-)analysis is a powerful tool to create data sets that provide a good description of the atmosphere including the hydrological cycle. Atmospheric analysis systems use a numerical weather prediction model to propagate the state of the atmosphere in time. Observations, which are irregular in space and time, are inserted into the system in such a way that an optimal blend is obtained between the background fields (forecast from e.g. 6 hours before) and the observations. Observations range from traditional surface observations and radiosondes, to commercial aircraft observations, cloud track winds from satellites and profiling measurements from infrared and microwave instruments in space. Re-analyses are performed to benefit from the most recent model improvements, data assimilation improvements and increased computer power and to obtain a time series that does not suffer from system changes as in operational numerical weather prediction. Current re-analyses have a spatial resolution of the order of 1 degree with a temporal sampling interval of 3 to 6 hours. Re-analysis products have the advantage of being projected on a global numerical grid and have no gaps.

The ECMWF 40-year Re-Analysis is described as an example. Its realism is discussed by comparing with observations of precipitation, vertically integrated water vapor, moisture convergence and surface fluxes over land and ocean. It is concluded that re-analysis products are very useful to study the hydrological cycle in the atmosphere. The synoptic variability is particularly good although biases may exist in some of the parameters.

INTRODUCTION

Data assimilation is a powerful tool to integrate observations of different types which are irregularly distributed in space and time, into a single analysis of the atmosphere. A forecast model is an integral part of data assimilation systems and is used to provide consistency between variables and to propagate the atmospheric state in time. Analyzed model variables in current numerical weather prediction (NWP) systems are surface pressure, the two horizontal wind components, temperature, and specific humidity. It also provides a series of derived parameters like cloud cover, precipitation, and turbulent and radiative fluxes. The ECMWF 40-year ReAnalysis (ERA-40) includes the analysis of ocean waves and atmospheric ozone (the actual length of ERA-40 is 44 years, from 1958 to 2001). Because surface-boundary conditions are needed, an externally provided sea surface-temperature analysis is used and land root zone soil moisture and soil temperatures are also analyzed with an indirect method. Future systems will include analysis of cloud variables and precipitation.

An important feature of data assimilation in NWP is that consistency of the fields in time and space is achieved by using a forecast model and a constraint (e.g. geostrophy) on corrections due to observations (analysis increments). A clear example is moisture. Moisture is correlated with temperature because it is constrained by its saturation value and carries the history of vertical motion, with subsidence leading to dry air and ascending motion leading to saturation. Therefore, even without moisture observations, the moisture fields look very realistic because they are controlled by horizontal and vertical advection and model physics.

A major benefit of data assimilation is that an uninterrupted time series (e.g. with time intervals of 3 or 6 h) of

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

global fields is obtained. The quality of these fields depends of course on the amount and quality of the data and on the quality of the assimilation model, but even in the case of missing data in a particular area, an analysis will be produced by the forecast model that is based on earlier data.

NWP centers produce analyses on a daily basis, because they need them as an initial state for weather forecasting (*see* **Chapter 180, Short-Term Predictions (Weather Forecasting Purposes), Volume 5**). These analyses are very useful and form the basis for many diagnostic studies. However, analysis and forecasting systems evolve continuously and are typically changed once or twice per year. It is well known that particularly the analysis of vertical motion and the hydrological cycle are to some extent system dependent. So these system changes lead to changes in the characteristics of the fields, clearly undesirable for applications where long time series of data are required. From the application point of view, it is desirable to have the best possible quality of analysis, which means that there is a preference for a recent system. Over the years, computer capacity has increased and data assimilation systems have improved. The consequence is that with current systems higher resolution analyses of better quality can be produced than the operational ones of say 10 years ago. On the other hand, reanalyses have less computer time available per analyzed day than operational systems because, to be practical, production rates have to be of the order of one month per day.

To get the best possible analyses without system changes, reanalysis projects have been launched. The principal global analyses that have been completed are the NCEP/NCAR 40 and 50-year reanalyses (Kalnay *et al.*, 1996; Kistler *et al.*, 2001), the NASA reanalysis (Schubert *et al.*, 1993), the ECMWF 15-year reanalysis ERA-15 (Gibson *et al.*, 1997) and the ECMWF 40-year reanalysis ERA-40 (Uppala *et al.*, 2003). All this work would not have been possible without a massive data gathering and archiving exercise (Kalnay and Jenne, 1991). For the earlier years, the meteorological data was not available in electronic form and had to be quality controlled and coded in a standard format for modern data assimilation systems. Much of this work was done by NCAR/NCEP and several conventional data sources were made available to the different reanalysis centers. The different sources were merged at ECMWF and satellite radiance data sets (VTPR, TOVS, SSM/I) were processed from their raw form.

Reanalyses have two basic products: (i) The analysis and short-range forecast fields on a standard grid and (ii) information on acceptance or rejection of observations and on the fit of the observations to the analysis and the first guess (analysis statistics). Both products are important for validation and verification studies which provide feedback for model and analysis system development. Also, the importance of analysis statistics on the quality of the observations should not be underestimated (Andersson and Järvinen, 1999). It can be used to improve the observations through bias adjustments, correction of coding mistakes and through flagging of unreliable data. The whole sequence of activities gives reanalysis a cyclic nature: The output of the project and the subsequent studies provide feedback for system development and improvement to the observation databases (Figure 1). These observation databases are shared among the different centers.

In this article, a short description of the ERA-40 system will be given as an example of a state-of-the-art reanalysis system. The accuracy of analysis products is still difficult to quantify. Kalnay *et al.* (1996) distinguish three levels of products. The A-level fields, that are most reliable because data are used directly for these fields. Examples are temperature, pressure, and the rotational part of the flow. B-level fields are indirectly analyzed with the divergent flow as an example. Finally, C-level fields are even more indirect as they are the result of a model simulation. Precipitation and turbulent fluxes are examples of level C data. It is therefore important to compare products from different analyses and to compare with independent observations. In this article, a few comparisons will be shown with *observations that have not been used by the system*. Emphasis will be on hydrology related parameters and hydrological applications. More information on ERA-40 can be found in Uppala *et al.* (2005) and at URL **`http://www.ecmwf.int/research/era/`**

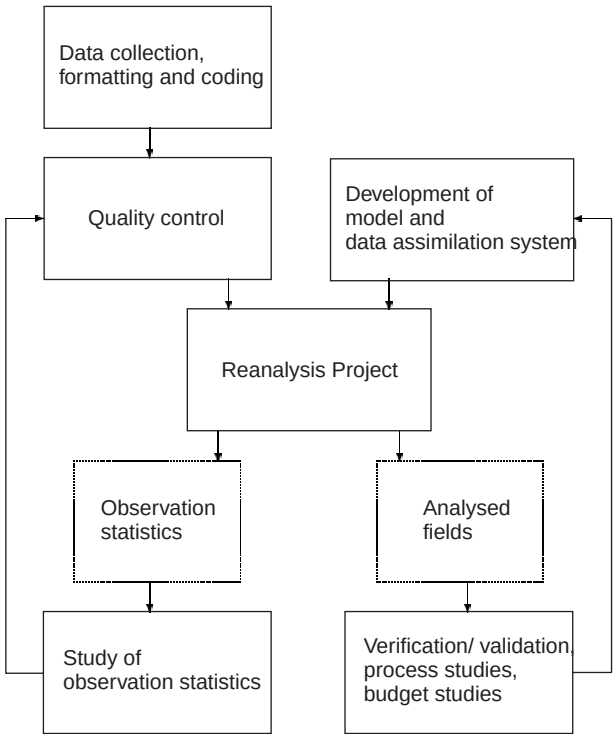


Figure 1 Flow diagram of reanalysis projects

DESCRIPTION OF THE ERA-40 SYSTEM

The ERA-40 system uses the ECMWF Integrated Forecasting System (IFS) developed in collaboration with Meteo France. The model applies a spectral representation with triangular truncation at wave number 159 (T159). The grid point spacing in physical space is about 125 km all over the globe due to a gradual decrease of grid points along a latitude circle towards the poles (a reduced Gaussian grid). The model has 60 "hybrid" levels in the vertical which are terrain following near the surface and pressure coordinates at the top of the model. The lowest model level is at about 10 m above the model surface; the top of the model is at 0.1 hPa. The model has prognostic variables for the two horizontal wind components, temperature, specific humidity, cloud cover, cloud water/ice, and ozone. From these variables, only the wind components, temperature, specific humidity, and ozone are constrained by observations. The model has a comprehensive physics package to describe radiative transfer, turbulence, subgrid orographic drag, moist convection, cloud processes including precipitation, land-surface processes, and simple ozone chemistry.

An analysis is generated every 6 h by adding increments to the background field (a short-range forecast, often called *the first guess*) to obtain a better fit to the available observations. The three-dimensional variational method (3DVAR, Courtier *et al.*, 1998; Rabier *et al.*, 1998; Andersson *et al.*, 1998) is used. It computes the equivalent of the observations, for example, through interpolation to radiosonde locations or through a radiative transfer model to find the radiance in a particular satellite channel. Then an adjoint model is used to find the sensitivity of these simulated observations with respect to the model fields. Finally, a minimization algorithm is applied to minimize the distance between the simulated and real observations and the analysis distance to the background by adjusting the model fields. Estimates of error variances and covariances of the background fields together with observation errors are used to achieve an optimal blend of first guess and observations and to obtain a proper spreading of observation information in the horizontal and vertical. Furthermore, a constraint is applied to obtain balance between mass and wind fields (Derber and Bouttier, 1999).

The ERA-40 intermittent analysis with a 6-hourly interval is consistent with the observation times of the main conventional observations (e.g. radiosondes and SYNOP's). However, many observations do not coincide with the main synoptic hours and, therefore, the observations are grouped in 6-hourly intervals from 3 h before to 3 h after analysis time. To account for the time evolution, the background field is compared to the observation at the appropriate time and for this purpose the first guess forecast runs up to 9 h. The departures of the observations from the background are, however, interpreted as if they apply to the synoptic time (see Figure 2 for a schematic representation of the analysis cycling).

Sea-surface temperature and sea ice cover are obtained from NCEP and specified as a daily updated fixed surface-boundary condition (Rayner, 2002). The surface-boundary condition over land is much more complicated. To provide a boundary condition for moisture and heat fluxes, the land-surface module has four soil layers to describe the evolution of soil moisture and soil temperature. The scheme relies on near-surface meteorology, radiative fluxes, and precipitation from the atmospheric model for its forcing and it returns sensible and latent heat flux. To represent surface heterogeneity, the model uses a tile approach in which six different subareas with different fractions have their own surface-energy balance (Van Den Hurk *et al.*, 2000). To prevent soil moisture and soil temperature from drifting, a soil analysis scheme is applied. It is based on the idea that temperature and moisture in the boundary layer will drift if soil moisture and/or soil temperature are biased, because the surface will provide biased fluxes to the boundary layer. The screen level temperature and specific humidity observations are used to quantify such drift and used to make corrections to soil moisture and/or soil temperature. The Optimum Interpolation (OI) method is used, as described by Douville *et al.* (2001). This method assumes a perfect model representation of the land surface which can be questioned because land-surface processes are complex and the model's representation of for example, soil type, and vegetation characteristics has limited accuracy. In spite of these shortcomings, the method will adjust the surface-soil moisture in such a way that the surface fluxes of sensible and latent heat are realistic.

DATA SOURCES

Conventional data from radiosondes and surface observations from land, ships, and buoys (so-called SYNOP reports) have traditionally been the backbone for

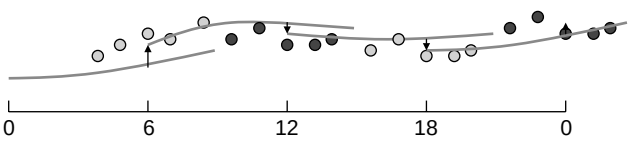


Figure 2 3DVAR cycling in ERA-40. The lines indicate the 9-h first-guess forecasts which are compared at the correct time with the observations (dots). The differences between observation and first guess are used at the synoptic times of 0, 6, 12, and 18 UTC to compute corrections to the first guess as indicated by the vertical arrows. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

atmospheric analysis. The radiosondes are obviously very important for the upper air and provide temperature, moisture, and wind information. SYNOP observations from land provide surface pressure and humidity information for the atmospheric analysis. SYNOP observations over the ocean are used for wind and surface pressure. Screen-level observations over land are also used to make a separate analysis of 2 m temperature and humidity, which is needed as input for the soil analysis scheme.

With the increase in commercial aviation, aircraft reports have become an important source of information, providing wind and temperature data for the upper air analysis.

A relatively new source of data are the meteorological satellites with good global coverage but limited vertical resolution. For most satellites, radiance observations are used directly in conjunction with a forward observation operator that simulates the measurement of the satellite channels using model fields as input. The cloud motion winds from geostationary satellites are an exception to this. For ERA-40, cloud imagery has been reprocessed by EUMETSAT (European Meteorological Satellite Agency) to obtain wind vectors, which are offered to the analysis system as wind observations.

Satellite coverage changes enormously over the ERA-40 period. During the early years from 1958 to 1972, no satellite radiance data was available and the upper air analysis relies predominantly on radiosonde data. From 1973 to 1978, there are infrared observations from the Vertical Temperature Profiler Radiometer (VTPR), which are for the first time used in ERA-40 as raw radiances. From 1979 onwards, infrared and microwave sounders are available from the Television and Infrared Observational Satellite Operational Vertical Sounder (TOVS) suite of instruments (High Resolution Infrared Sounder HIRS, Microwave Sounder Unit MSU, Stratospheric Sounder Unit SSU, Total Ozone Mapping Spectrometer TOMS). Introduction of the TOVS data had a big positive impact on the quality of the Southern Hemisphere analysis. The Special Sensor Microwave Imager (SSM/I) started in 1987 and is particularly important for atmospheric water vapor and surface-wind assimilation over the ocean (*see* **Chapter 65, Estimation of Water Vapor and Clouds Using Microwave Sensors, Volume 2**). Another instrument that is important for the ocean-surface wind is the scatterometer in operation from different platforms since 1991. The scatterometer is an active instruments that takes information from backscattered radar signals from the ocean surface.

Although the analysis system may be the same for the entire analysis period, the results are still affected by changes in the observing system. Figure 3 shows the number of reports from radiosondes, satellites, and surface observations over the years. It was a design choice of ERA-40 to make maximum use of all available observations. The biggest change over the ERA-40 period is in the satellite data, which is particularly relevant for the Southern Hemisphere where the number of conventional observations is limited.

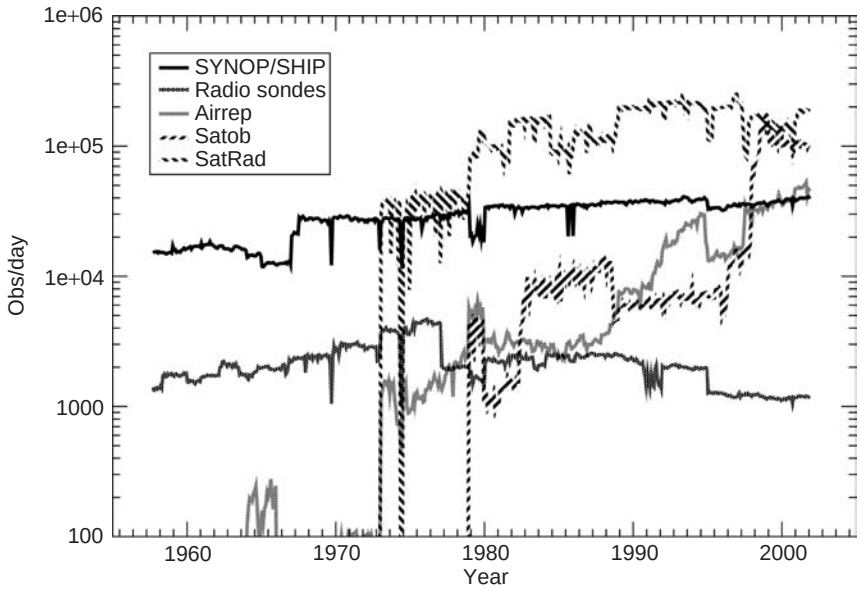


Figure 3 Number of observations of different type during the ERA-40 period in number of reports per day. SYNOP/SHIP's are the conventional surface observations over land and ocean, Radiosondes are conventional upper air observations, AIREP's are the aircraft reports, SATOB's are the geostationary satellite cloud track winds, and SATRAD are the satellite radiance observations. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

THE ATMOSPHERIC BRANCH OF THE HYDROLOGICAL CYCLE

Atmospheric moisture is obviously a key component of the hydrological cycle (*see* **Chapter 25, Global Energy and Water Balances, Volume 1**). It is a greenhouse gas, a major carrier of energy and highly relevant for precipitation. However, it is fair to say that the analysis of moisture in the atmosphere is still in its infancy as most of the development in data assimilation has been on temperature and wind. Routine moisture observations are difficult, often biased and one may wonder how accurate an atmospheric analysis of moisture is. Moisture is constrained by other variables and influenced by horizontal and vertical advection. Over the ocean, SSM/I radiances are used to control vertically integrated moisture. An example of vertically integrated water vapor is given in Figure 4. The model first guess is compared with an independent retrieval from SSM/I with an empirical algorithm (*see* **Chapter 65, Estimation of Water Vapor and Clouds Using Microwave Sensors, Volume 2**). The patterns of the vertically integrated moisture are remarkably similar, which suggests that the moisture variability is well captured by the analysis system. This is also clear from a 17-month-long time series of moisture difference between the sea surface and the 3.2 m level (model interpolated to the buoy observation level) compared to observations from the PACS buoy (not assimilated) in the Eastern Pacific (Figure 5). This difference is controlled by exchange with the surface, by advection, by evaporation of

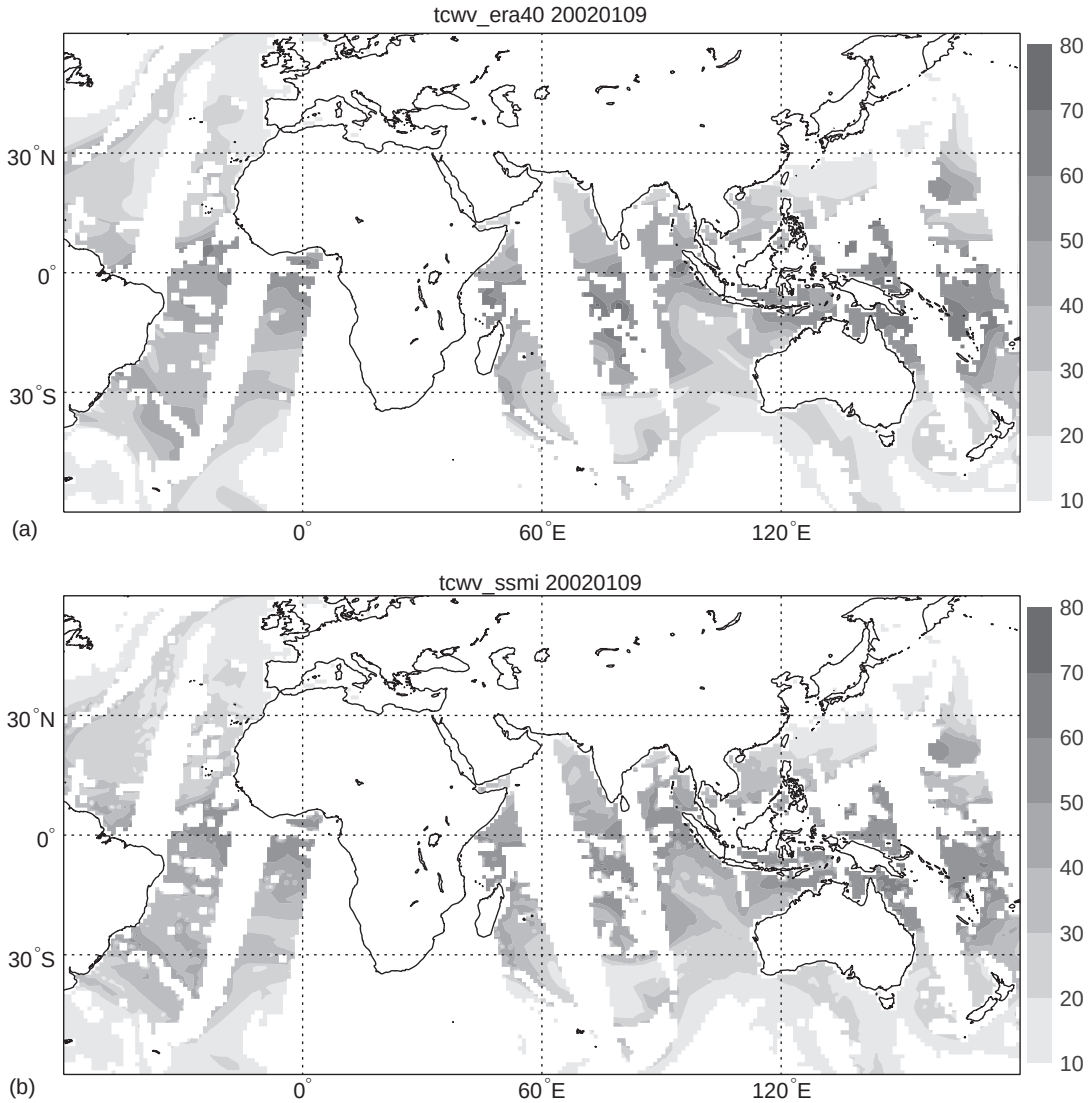


Figure 4 Total column water vapor in kg/m^2 from ERA-40 (a) and from an empirical algorithm using SSMI (b). For both plots, a mask has been applied that excludes missing SSMI data and areas where SSMI is contaminated by rain. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

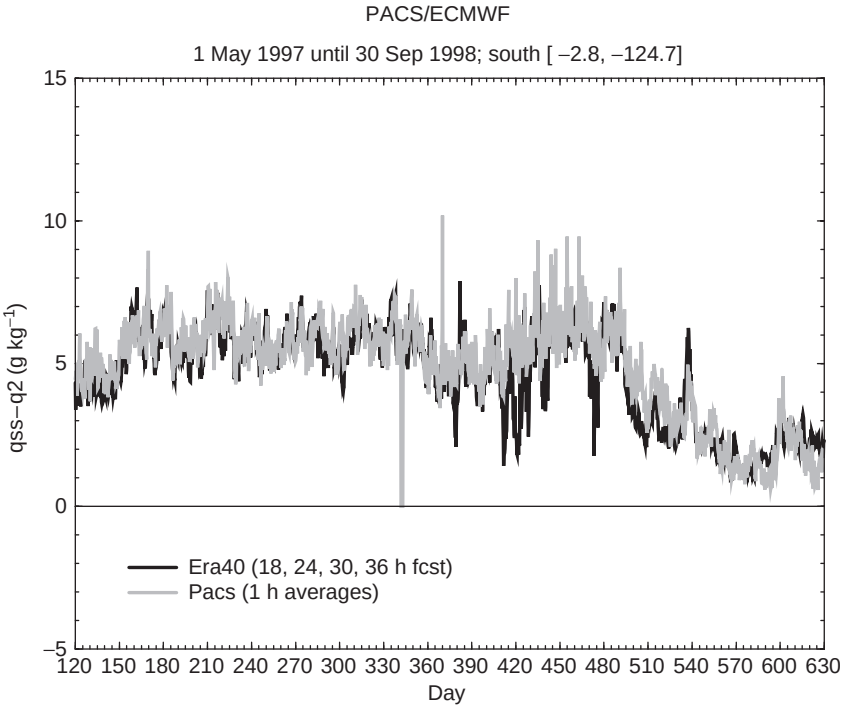


Figure 5 Time series of moisture difference between the sea surface and the observation level of 3.2 m in the Western Pacific (3 °S, 125 °W). ERA-40 data are black and PACS buoy observations are in grey (data are provided by the Woods Hole Institute of Oceanography; Weller and Anderson, 1996). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

precipitation, and mixing processes in the boundary layer that transport moisture away from the surface. A considerable amount of the variability as seen in the observations is captured by the short-range forecasts (18-, 24-, 30-, and 36-h forecasts). The last months of this period are particularly interesting because a cold SST anomaly develops, resulting in advection of warm moist air over a relatively cold surface which reduces the ocean/air moisture difference. This example clearly illustrates that particularly in data-sparse areas, the dynamical information is important for the moisture analysis.

Precipitation is a parameter that is not analyzed directly but it is simulated by the model physics during short-range forecasts. Two types of precipitation are distinguished: (i) convective precipitation due to condensation in convective updraught in moist unstable areas, and (ii) large-scale precipitation due to microphysical processes in clouds for example, in frontal areas with large-scale ascent, and also in convective areas where the convection scheme detrains ice to the cloud variables. Figure 6 shows a 3-h average of total precipitation from ERA-40, compared to retrievals of rainfall over the oceans from the TRMM Micro wave Instrument (TMI), which was not used in ERA-40. The main precipitation areas are well captured, although there are big differences in the details. This is partially because the ERA-40 precipitation chart represents the average over 3 h, whereas the TMI product is a snap shot. Furthermore, TMI is not capable of detecting the low precipitation rates. However, this Figure indicates that the synoptic variability is captured and that the ERA-40 system tends to smear out the precipitation fields due to its spatial and temporal averaging.

Figure 7 shows monthly averages of 12- to 24-h forecasts from ERA-40 as a scatter plot against Global Precipitation Climatology Project (GPCP) data (over land mainly based on gauge data (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**)). The correlation is rather good in all climatological regimes but systematic biases do exist and depend on the area. In a comparison with independent data over the Mackenzie catchment area, Betts *et al.* (2003) found a good correspondence for monthly averaged precipitation with errors of the order of 10% dependent on the period (i.e. dependent on the data that went into the system) and dependent on forecast range.

The largest bias is seen over the tropical oceans and is related to the so-called spinup problem. A time series of monthly tropical mean precipitation is shown in Figure 8 (c). The level of precipitation is substantially higher than for example, the GPCP estimate which is a little over 3 mm day^{-1}. In the ERA-40 system, some of the difference is related to an imbalance between model and satellite observations. The satellite observations add moisture to the

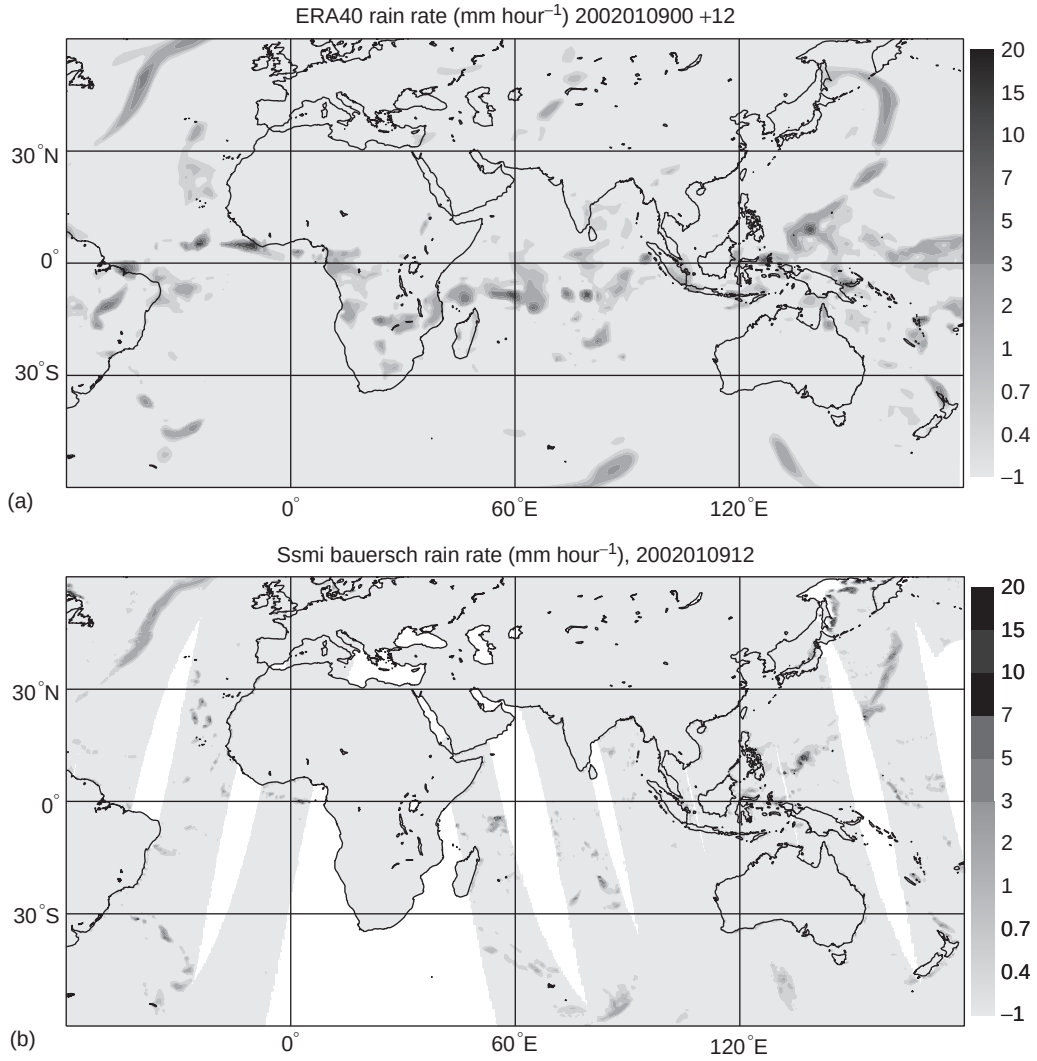


Figure 6 Example of a 3-hourly averaged precipitation field from ERA-40 (a) and a retrieval of instantaneous precipitation over the ocean from TMI with the Bauer and Schlüssel (1993) algorithm. (b) The 3-hourly period is from 9 to 12 UTC 20020109. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

system in a systematic way and this moisture rains out in the beginning of the forecast. This is illustrated in the part (b) and (c) of Figure 8. The amount of atmospheric moisture increases in the analysis with the introduction of satellite observations. Also, the increments (i.e. the amount of moisture that is added by the analysis to the first guess using observations) increase in time. Before the satellite era, there are virtually no increments. The introduction of VTPR in 1973 is clearly visible and later the increments increase with the growing number of satellite observations. The increased increments of total column water vapor applied 4 times per day, actually match the increased precipitation rather closely, suggesting that the added water vapor leads to additional but spurious rain during the early hours of the forecasts. The grey curve (from daily 12- to 36-h precipitation) and the black curve (from 0 to 6 h precipitation 4 times per day) in the part (c) of Figure 8, show a slight increase of precipitation with forecast range in the presatellite era and a systematic decrease later. The spin-down of moisture is only 2 to 5% and is within calibration accuracy of most instruments. However, in terms of precipitation, the additional amount of moisture put into the system every 6 h is very large and leads to substantial biases. The character of the precipitation spin-down is further illustrated in Figure 9, showing that it takes about three to four days for the model to reach its own equilibrium. The reason is that the excessive latent heat release during the 6-h data assimilation cycling maintains a Hadley circulation that is about 20% stronger than in the model climate (not shown). It is difficult to verify the strength of the Hadley circulation, but the spin-down over a period of a few days is obviously an artifact of the data assimilation system.

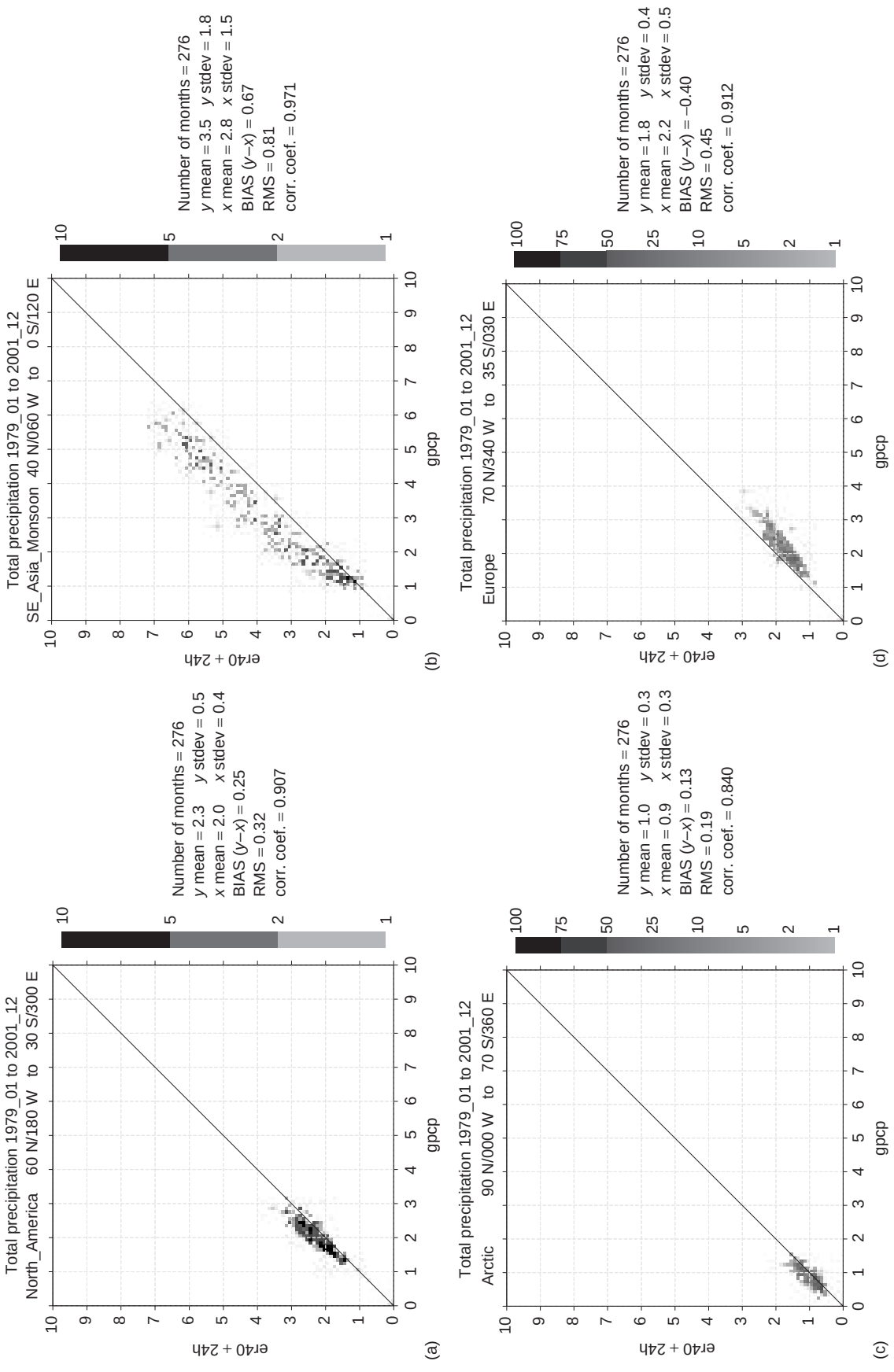


Figure 7 Monthly total precipitation over four different areas from ERA-40 1979–2001 compared to GPCP data. The units are mm day^{-1}; the number of points are color coded. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

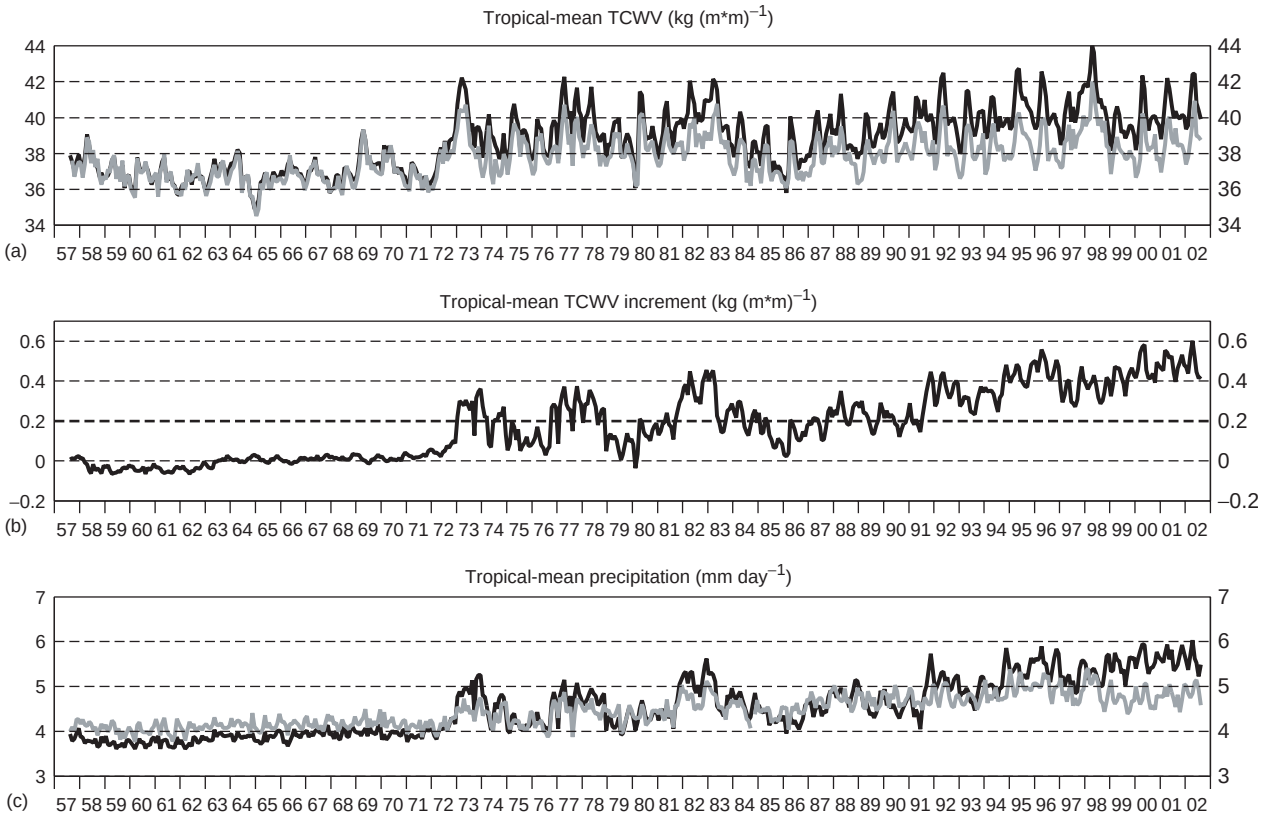


Figure 8 Time series of monthly tropical mean total column water vapor (TCWV, a), TCWV increment every 6-hourly analysis cycle (b), and precipitation (c). In part (a), the black curve is the analysis and the grey curve the 24-h forecast. In part (c), the black curve is from 0- to 6-h forecasts 4 times per day and the grey curve is from the 12 to 36-h daily forecasts. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

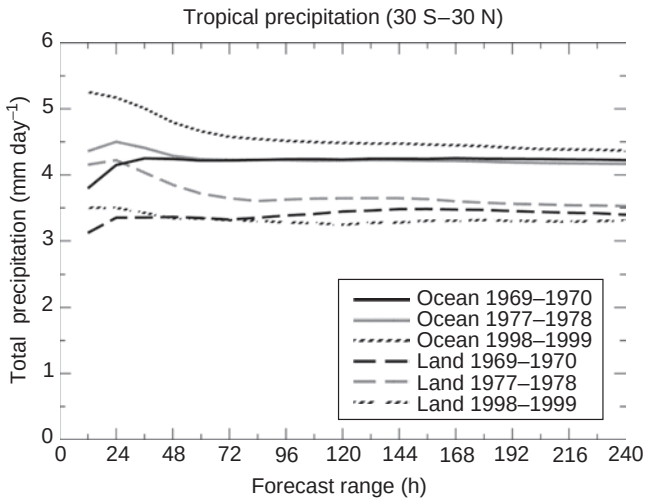


Figure 9 Evolution tropical precipitation averaged over the ocean (solid), land (dashed) during the 10-day forecast for 2-year periods with different observing systems. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

To summarize, ERA-40 was designed to use as much available data as possible, resulting in a good representation of synoptic variability. The side effect is that a too active hydrological cycle is maintained in the tropics due to a slight imbalance in moisture between observations and model. It is difficult to say whether the bias is due to model errors, due to observation biases or due to the way the observations are used. Precipitation over tropical oceans is highly biased compared to GPCP and drifts to higher values with the increasing number of satellite observations.

TURBULENT FLUXES OVER THE OCEAN

Ocean fluxes are important for a variety of applications for example, budget studies, ocean modeling, ocean wave simulations. Several climatologies of ocean fluxes exist, based on ship observations and/or satellite observations and NWP analysis products. A good review of these products and a comprehensive intercomparison study is given by Taylor (2000). In their basic form, most products do not satisfy the global energy and water-budget constraints, and therefore adaptations are often made. The resulting products that focus on say monthly climatology tend to have closed energy and water budgets by construction, but they do not have the day-to-day variability. NWP products are at the other extreme; they have the short timescale variability (diurnal, synoptic), but they do not (yet) satisfy large-scale budget constraints.

An example of a 10-year average of surface-latent heat flux of ERA-40 is shown in Figure 10. Such a field looks very similar to other climatologies (e.g. the da Silva *et al.*, 1994 climatology as shown in Figure 10), but it is very difficult to give estimates of uncertainty. Therefore, a

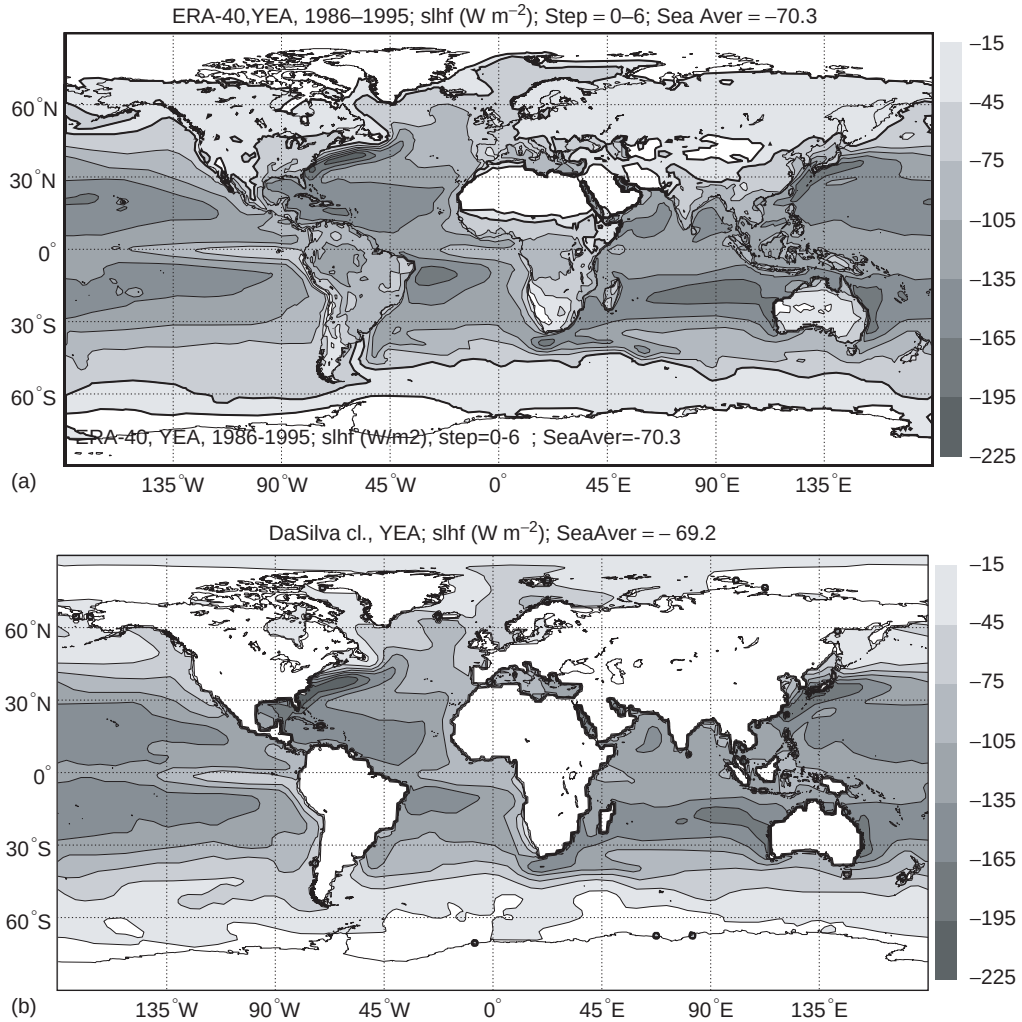


Figure 10 Surface-latent heat flux from ERA-40 averaged from 1986 to 1995 (a) and the climatology over the ocean by da Silva *et al.* 1994, (b). Negative fluxes are upward. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

comparison is made with research-quality buoy data that are not used in the ERA-40 assimilation. Point observations over the open ocean are representative for a large area because of the homogeneity of the ocean surface. Time series of 6-hourly fluxes of latent heat and net solar fluxes for the Atlantic IMET buoy (Weller and Anderson, 1996) are shown in Figure 11 in comparison with the ERA-40 fluxes. The reanalysis fluxes follow the observations very well, although small systematic biases exist. This Figure again illustrates that the analysis products do have a very realistic level of variability even in data-sparse areas over the ocean.

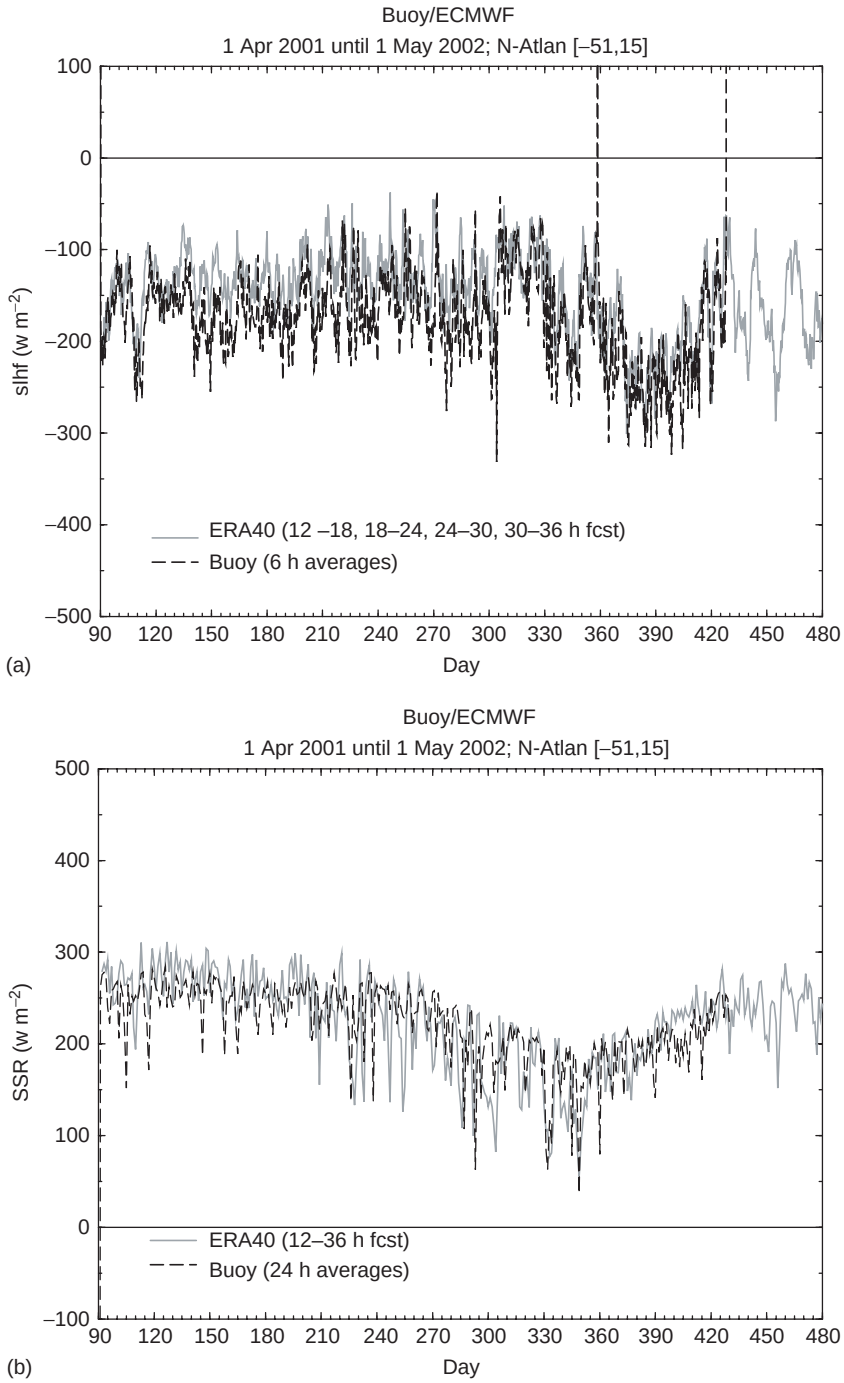


Figure 11 Time series of 6-hourly latent flux (a) and net solar radiation (b) from the Atlantic IMET buoy (51 °W/15 °N) and ERA-40. The time period is from 1 April 2001 to 1 May 2002 (data provided by the Woods Hole Institute of Oceanography; Weller and Anderson, 1996). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

TURBULENT FLUXES OVER LAND

Fluxes over land from reanalyses are more uncertain than fluxes over the ocean. Over the ocean, there is a fairly well-observed boundary condition, namely, the sea-surface temperature, which makes the turbulent fluxes only dependent on atmospheric variables (wind, temperature, and specific humidity) and on the air-sea transfer coefficients which are reasonably well known (Zeng *et al.*, 1998). Over land, the situation is much more complicated. There is a close balance between available energy at the surface (net radiation) and the sensible and latent heat fluxes, with the ground heat flux as a small residual (at least during day time). The partitioning between sensible and latent heat flux is controlled by soil moisture availability with near-surface soil moisture affecting bare soil evaporation and the root zone soil moisture controlling the plant transpiration (*see* **Chapter 70, Transpiration and Root Water Uptake, Volume 2**). The land-surface system is very complex and only represented in models in a very simplified way, so there is a real question about the accuracy of reanalysis fluxes over land.

On the other hand, climatology of turbulent surface fluxes over land is virtually nonexistent, because simple and robust observational methods are not available. So, reanalysis products and other model-derived products are well worth exploring. Verification is possible with new technology that allows for flux observations with reasonable accuracy at research sites (e.g. with eddy correlation equipment (*see* **Chapter 40, Evaporation Measurement, Volume 1**)). Many "flux towers" exist with most of them providing data during campaigns. This material is ideal for verification over land, because it is completely independent of the reanalysis. An example of a 5-year time series of 7-day averages of fluxes is presented in Figure 12 for the Cabauw location in the Netherlands. The part (a) of the Figure 12 shows that the downward short-wave radiation in the model has a good annual cycle, but also a very reasonable variability on the weekly timescale. The latter means that clouds are sufficiently well represented to modulate the solar radiation in a realistic way. The net radiation shows the same variability but has a positive bias, which is probably related to a difference in surface albedo between the observational site (0.23) and the model value for this area (0.18). The ratio of sensible and latent heat fluxes (Bowen ratio) is fairly small, because evaporation is seldom restricted by soil moisture. The consequence is that latent heat flux dominates and follows the net radiation and shows relatively small errors. The biggest errors are seen in the sensible heat flux with generally an overestimation of the upward flux (negative) in summer and also an overestimation of the downward heat flux (positive) in winter.

Comparisons as shown in this section are not without problems. Point observations may not be representative for a large area, and observations often have problems with closure of the surface-energy balance. However, with more observations becoming available and also observations over typical subareas of a grid box (e.g. over forest as well as low vegetation), it will be possible to reduce the uncertainty.

In spite of the uncertainties, reanalysis products are relevant for the energy and moisture budgets over land, because very little other information exists. Existing climatologies are based on long-term averages of observed radiation, precipitation and runoff, but are obviously limited to large areas and exclude the short timescales (Brutsaert, 1982). Reanalysis in conjunction with many *in situ* observations and budget constraints will hopefully lead to good documentation of biases and deficiencies. It may even be possible to improve the reanalysis products by applying corrections.

OPTIONS FOR RESEARCH

Reanalyses generate a wealth of information covering timescales from a few hours to the full length of a reanalysis with a spatial resolution of typically 100 km. This data is very useful in the study of atmospheric processes, and in the study of large-scale budgets. Parameters that are not analyzed directly are only indirectly constrained by observations and may be model dependent. It is therefore important to intercompare products and do validation studies using independent data. The Coordinated Enhanced Observing Period (CEOP) by the GEWEX Hydrometeorology Panel is such an initiative. CEOP gathers analysis products and research-quality observations at different locations to allow for validation and large-scale budget studies. Such studies will give important information on the quality of analysis products and will also suggest correction procedures for systematic errors.

An example of a large-scale budget study is given by Seneviratne *et al.* (2004). They use the vertically integrated atmospheric moisture field and its convergence from ERA-40 to determine the net moisture flux at the surface (precipitation minus evaporation) and compare it with the observed soil moisture budgets (soil moisture change minus runoff). Figure 13 shows a good correspondence between the atmospheric and soil budgets. It means that the analysis of moisture in this area is sufficiently accurate to give information about the soil water budget. It should be remembered that this is an area with a rather good radiosonde network to constrain the moisture field. The traditional way of doing moisture budgets is to use the radiosonde data only (Peixoto and Oort, 1992). Data assimilation has a clear advantage here, because: (i) the system provides consistency with other observations like temperature and wind, (ii) the system also uses other observations for example, surface observations, aircraft reports, and satellite data,

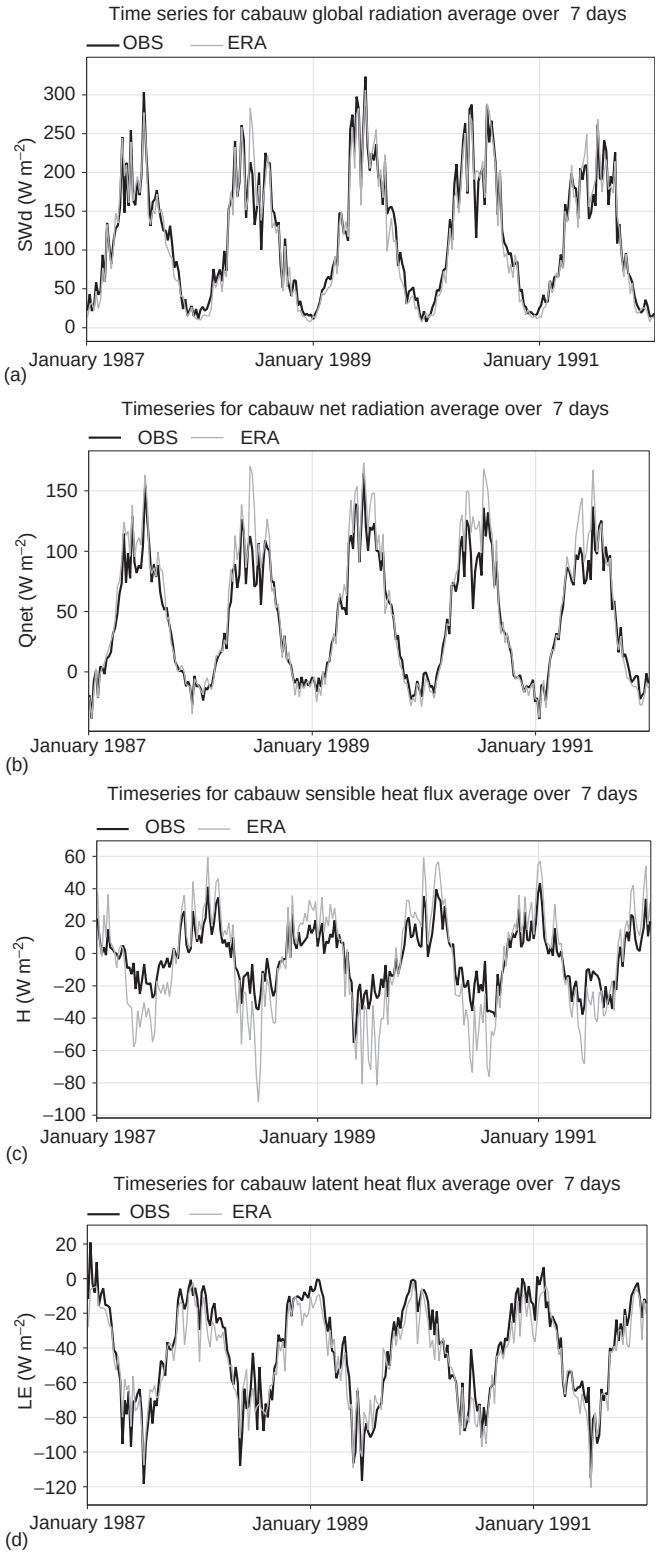


Figure 12 Time series of weekly averages of downward solar radiation (a), net radiation (b), sensible heat flux (c) and latent heat flux (d) from ERA-40 compared to observations at the Cabauw tower in the Netherlands (Beljaars and Bosveld, 1997). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

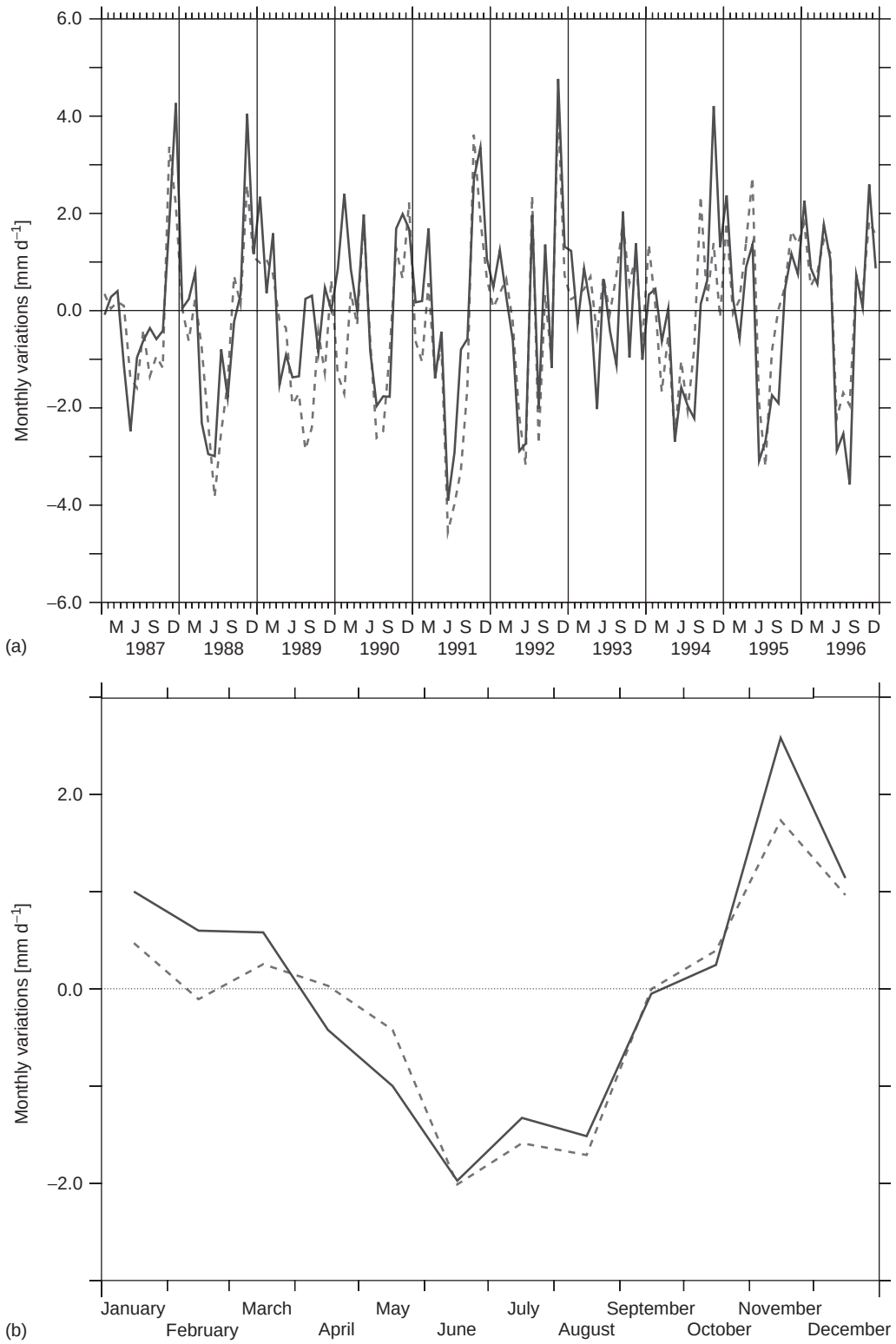


Figure 13 Atmospheric moisture convergence minus runoff minus the change of atmospheric moisture content (dashed) compared to the observed change of soil moisture content (solid) for a river catchment area in Illinois. The atmospheric moisture convergence and the atmospheric moisture content are from ERA-40. The figure (a) shows a 10-year time series of monthly averages. The figure (b) represents the mean annual cycle. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

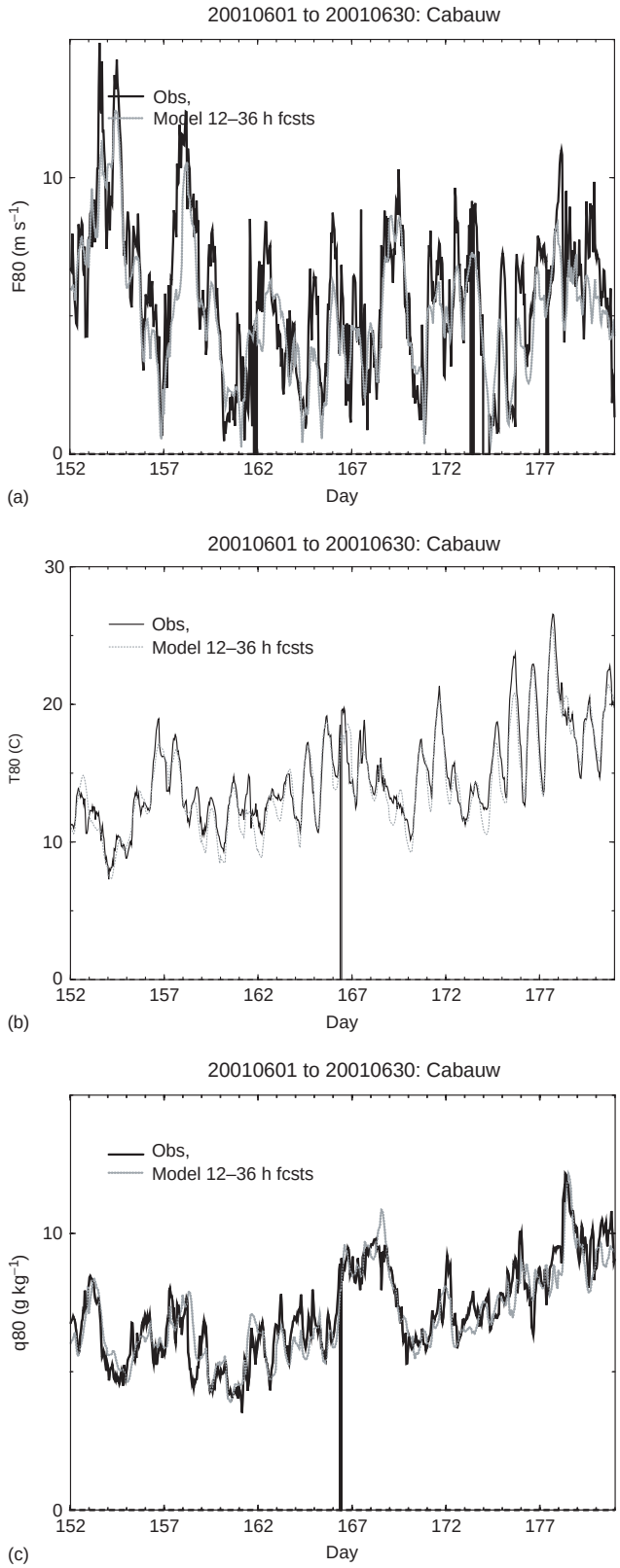


Figure 14 Time series for June 2001 of hourly observations of wind speed (a), temperature (b) and specific humidity (c) from the Cabauw tower in the Netherlands at a height of 80 m, together with hourly ERA-40 data at 100 m from the nearest grid point (the observations have been provided by Fred Bosveld from KNMI)

(iii) it is trivial to compute moisture convergence from the gridded results, and (iv) data assimilation has high time resolution.

Another example of an application is the analysis of soil moisture on a continental scale. Soil moisture observations are difficult and not available on a global scale. Therefore, land-surface models are used to simulate soil moisture using observed atmospheric forcing, that is, downward radiation, precipitation, wind, temperature, and moisture at some height above the surface. The resulting soil moisture products are model dependent and comparison of different models is used to gauge the uncertainty (Dirmeyer *et al.*, 1999). Of course, soil moisture from reanalysis can be used directly, but in this case precipitation and radiation are also simulated by a model rather than observed. To drive land-surface models with an atmospheric boundary condition, gridded fields of temperature, moisture, and wind are needed with sufficient time resolution to resolve the diurnal cycle. Reanalysis fields are highly suitable for such an application, and therefore ERA-40 data is one of the data sets to be used in the ISLSCP-II project (International Satellite Land Surface Climatology Project), which is a follow-up of the ISLSCP-I project to collect land-surface data (see Sellers *et al.*, 1995), but now over a 10-year period. The choice of the height above the surface for atmospheric forcing is not obvious, therefore, two options are provided: the lowest model level at a height of about 10 m and the 4^{th} level from the surface which is at a height of about 100 m. The latter is believed to be the most suitable, because it is still fairly close to the surface, but is less influenced by the surface-boundary condition as used in ERA-40 (e.g. roughness lengths). A one-month time series of wind, temperature, and moisture is shown in Figure 14 compared to observations from the Cabauw mast in The Netherlands (not used by ERA-40). The time series shows the high level of realism of the basic atmospheric parameters even in the boundary layer, where boundary layer and land-surface parameterization aspects may play an important role. The high level of realism makes them very suitable for applications as forcing in land-surface models.

CONCLUDING REMARKS

Reanalysis is a powerful way of integrating a large volume and a wide range of atmospheric observations into a synthesis of the atmosphere and land surface, using the same state-of-the-art method over the entire period of interest. Reanalyses provide a wealth of atmospheric fields in a gridded format at high time resolution for a long period of time, and therefore open many options for research. Some reanalysis fields are more directly constrained by observations than others. For instance, the large-scale atmospheric flow and temperature are well represented, but model-derived fields or less well-observed fields (e.g. precipitation, moisture, turbulent, and radiative fluxes) are less accurate. We focused here on the hydrological cycle and concluded that many derived fields show realistic synoptic variability. Clear statements about the accuracy of the different elements of the hydrological cycle are difficult to make mainly because the accuracy of the verification material is not known. The moisture fields in ERA-40 might well be within calibration accuracy of most routine observations. However, there is a systematic mismatch of a few percent between moisture observations and the model equilibrium, leading to moisture increments in the tropics. This results in excessive oceanic precipitation in the tropics during the first days of the forecast and to a too active hydrological cycle. On the other hand, the ocean evaporation might be as accurate as existing monthly climatologies with the advantage of high time resolution. More research is needed to establish the accuracy.

Preliminary studies have shown that over well-observed land areas (e.g. North America and Europe) moisture convergence from reanalyses is a valuable product in the context of budget studies. Accuracy of precipitation depends on the area and is often between 10 and 20% for monthly averages. The largest uncertainty is in evaporation because of the lack of ground truth. Verification for a single station shows that the latent heat flux follows the net available energy rather closely, because it is a location with virtually no water stress. Many more stations and more research is needed to establish how accurate these surface fluxes are. Such research is timely because observation technology has evolved and many flux towers exist.

Acknowledgment

The authors would like to thank Emmanuel Moreau, Philippe Lopez, and Sonia Seneviratne for making available the figures 4, 6 and 13. Alan Betts carefully reviewed the manuscript and provided valuable suggestions for improvement.

FURTHER READING

Arpe K., Klepp C. and Rhodin A. (2000) Differences in the hydrological cycles from different reanalyses – which one should we believe? *Proceedings of the 2nd WCRP International Conference on Reanalyses*, WCRP-109, WMO/TD-985, World Meteorological Organization: pp. 193–196.

Derber J. and Bouttier F. (1997) A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**, 195–222.

Derber J. and Bouttier F. (1999) A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**, 195–221.

Fisher M and Courtier P (1995) *Estimating the Covariance Matrices of Analysis and Forecast Error in Variational Data*

Assimilation, ECMWF Technical Memorandum 220, ECMWF, Reading, UK.

REFERENCES

Andersson E., Haseler J., Unden P., Courtier P., Kelly G., Vasiljevic D., Brankovic C., Cardinali C., Gaffard C., Hollingsworth A., *et al.* (1998) The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part III: experimental results. *Quarterly Journal of the Royal Meteorological Society*, **124**, 1831–1860.

Andersson E. and Järvinen H. (1999) Variational quality control. *Quarterly Journal of the Royal Meteorological Society*, **125**, 697–722.

Bauer P. and Schlüssel P. (1993) Rainfall, total water, ice water, and water vapor over sea from polarized microwave simulations and special sensor microwave/imager data. *Journal of Geophysical Research*, **98**, 20737–20759.

Beljaars A.C.M. and Bosveld F.C. (1997) Cabauw data for the validation of land surface parameterization schemes. *Journal of Climate*, **10**, 1172–1193.

Betts A.K., Ball J.H. and Viterbo P. (2003) Evaluation of the ERA-40 surface water budget and surface temperature for the Mackenzie river basin. *Journal of Hydrometeorology*, **4**, 1194–1211.

Brutsaert W. (1982) *Evaporation into the Atmosphere*, Reidel Publishers: Dordrecht.

Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljevic D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M. (1998) The ECMWF implementation of three dimensional variational assimilation (3D-Var) I: formulation. *Quarterly Journal of the Royal Meteorological Society*, **124**, 1783–1807.

da Silva A., Young C.C. and Levitus S. (1994) *Atlas of Surface Marine Data 1994*, Vol. 1: Algorithms and Procedures, NOAA Atlas NESDIS 6, US Department of Commerce: Washington, p. 83.

Dirmeyer P.A., Dolman A.J. and Sato N. (1999) The global soil wetness project: a pilot project for global land surface modeling and validation. *Bulletin of the American Meteorological Society*, **80**, 851–878.

Douville H., Viterbo P., Mahfouf J.-F. and Beljaars A.C.M. (2001) Evaluation of the optimum interpolation and nudging techniques for soil moisture analysis using FIFE data. *Monthly Weather Review*, **128**, 1733–1756.

Gibson J.K., Kallberg P., Uppala S., Nomura A., Hernandez A. and Serrano E. (1997) *ERA Description*, ECMWF Reanalysis Project Report Series nr. 1, ECMWF, Reading, UK.

Kalnay E. and Jenne R. (1991) Summary of the NMC/NCAR reanalysis workshop of April 1991. *Bulletin of the American Meteorological Society*, **72**, 1897–1904.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G.,Woollen J., *et al.* (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77**, 437–471.

Kistler R., Kalnay E., Collins W., Saha S., White G., Woollen J., Chelliah M., Ebisuzaki W., Kanamitsu M., Kousky V., *et al.* (2001) The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society*, **82**, 247–268.

Peixoto J.P. and Oort A.H. (1992) *Physics of Climate*, American Institute of Physics: New York.

Rabier F., McNally A., Andersson E., Courtier P., Unden P., Eyre J., Hollingsworth A. and Bouttier F. (1998) The ECMWF implementation of three dimensional variational assimilation (3D-Var). II: structure functions. *Quarterly Journal of the Royal Meteorological Society*, **124B**, 1809–1830.

Rayner N. (2002) HadISST1 and the Reynolds *et al.* analysis, Proceedings of Workshop on Reanalysis, 5–9 November 2001, ECMWF, Reading, UK.

Schubert S., Rood R. and Pfaendtner J. (1993) An assimilated data set for earth sciences applications. *Bulletin of the American Meteorological Society*, **74**, 2331–2342.

Sellers P.J., Meeson B.W., Hall F.G., Asrar G., Murphy R.E., Schiffer R.A., Bretherton F.P., Dickinson R.E., Ellingson R.G., Field C.B., *et al.* (1995) Remote sensing of the land surface for studies of global change: models – algorithms – experiments. *Remote Sensing of Environment*, **51**, 3–26.

Seneviratne S., Viterbo P., Lüthi D. and Schär C. (2004) Inferring changes to terrestrial water storage using ERA-40 reanalysis data: the Mississippi river basin. *Journal of Climate*, **17**, 2039–2057.

Taylor P.K. (Ed.) (2000) Intercomparison and Validation of Ocean-Atmosphere Energy Flux Fields, WCRP-12, WMO/TD-No. 1036, Geneva.

Uppala S.M., Kallberg P.W., Simmons A.J., Andrae U., da Costa Bechtold V., Fiorino M., Gibson J.K., Haseler J., Hernandez A., Kelly G.A., Li X., Onogi K., Saarinen S., Sokka N., Allan R.P., Andersson A., Arpe K., Balmaseda M.A., Beljaars A.C.M., van de Berg L., Bidlot J., Bormann N., Caires S., Dethof A., Dragasovac M., Fisher M., Fuentes M., Hagemann S., Holm E., Hoskins B.J., Isaksen L., Janssen P.A.E.M., McNally A.P., Mahfouf J.-F., Jenne R., Morcrette J.-M., Raynor N.A., Saunders R.W., Simon P., Sterl A., Trenberth K.E., Untch A., Vasiljevic D., Viterbo P. and Woollen J. (2005) The ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society*, in press.

Van Den Hurk B.J.J.M., Viterbo P., Beljaars A.C.M. and Betts A.K. (2000) *Offline Validation of the ERA-40 Surface Scheme*, ECMWF Technical Memorandum Nr. 295, ECMWF, Reading, UK.

Weller R.A. and Anderson S.P. (1996) Surface meteorology and air-sea fluxes in the western equatorial Pacific warm pool during the TOGA Coupled Ocean-Atmosphere Response Experiment (COARE). *Journal of Climate*, **9**, 1959–1990.

Zeng X., Zhao M. and Dickinson R.E. (1998) Intercomparison of bulk aerodynamic algorithms for the computation of sea surface fluxes using TOGA COARE and TAO data. *Journal of Climate*, **11**, 2628–2644.

183: Teleconnections in the Earth System

THOMAS N CHASE[1], ROGER A PIELKE SR[2] AND RONI AVISSAR[3]

[1] *Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, US*
[2] *Department of Atmospheric Science, Colorado State University, Fort Collins, CO, US*
[3] *Department of Civil and Environmental Engineering, Duke University, Durham, NC, US*

This article illustrates the large-scale connectivity of the atmosphere–ocean coupled system and generalizes the concept to regional scales and to other components of the earth system. Connections at a distance, or teleconnections, can occur by the direct transfer of mass by changes in regular circulations or by propagating waves initiated by a variety of mechanisms. Questions as to what extent recognized teleconnection patterns can be associated with identifiable forcing mechanisms, to what extent these patterns are interrelated and how they might cause, react to, or interact with changing forcing such as changes in atmospheric composition, land cover, or the distribution of sea ice to produce climate changes are examined.

INTRODUCTION

The term teleconnection is usually defined as a coherent atmospheric response to remote forcing such as particular sea surface temperature or atmospheric pressure patterns. The term is generally applied to a disturbance in the atmospheric circulation that is persistent and of large spatial scale (continental and above). However, a more complete definition should refer to a teleconnection as any transmission of a coherent effect beyond the location at which a forcing occurred. Seasonal weather forecasters noticed certain persistent atmospheric circulation features and were using these patterns for seasonal weather forecasts by the 1950s based on theoretical development by Bjerknes and Rossby in the previous decades (Namias, 1953, 1959).

Circulating fluids, such as the atmosphere and oceans, communicate information over large parts their volume and these teleconnections can be defined to occur in two ways. First, the atmosphere and oceans organize themselves into coherent circulations on a variety of time and spatial scales. These include the Hadley cell, subtropical jet streams, monsoons, sea and mountain breezes, and the oceanic thermohaline circulation. A change to the overall strength or position of the circulation will generally be noticeable over a wide area. Secondly, disturbances associated with these coherent circulations generate waves of several kinds that propagate in fluids in different ways and can be quite persistent. These waves do not necessarily follow the path of the coherent circulations mentioned above and can generate regional climate anomalies far from the source of the original disturbance.

The atmosphere–ocean system appears to oscillate in certain quasi-periodic teleconnection patterns that typically move between different states on a variety of timescales (Barnston and Livzey, 1987). These transitions between states can be quite abrupt or rather gradual, both indicative of the nonlinear character of the climate system (Rial *et al.* 2004). Recognizable climate anomalies associated with each phase of a particular oscillation are then transmitted over wide areas of the globe through the mechanisms mentioned above.

The simplest way to identify such patterns in observational data is to choose points on the globe and correlate that point with every other point. In this way, coherent regions of correlation and anticorrelation may become apparent. More sophisticated statistical methods, such as empirical orthogonal function (EOF) analysis or rotated principal component analysis (RPCA), seek to isolate independent patterns and to maximize the variability associated with the major patterns. Recognized teleconnection patterns seem to result from the internal dynamics of the atmosphere and/or ocean rather than from forcing from outside the

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

earth system and for this reason they are often referred to as natural modes of variability. Teleconnections offer potential long-term weather predictability based on their persistence and oscillatory behavior and potentially hold the key to understanding the relation between natural and anthropogenic climate change.

Several critical questions as to the nature and mechanics of teleconnections, however, remain. These questions reflect uncertainty as to how to define more-or-less independent patterns and how individual patterns are related to each other, to what extent teleconnection patterns can be associated with a response to an identifiable forcing mechanism, what physical mechanisms are involved in a propagating pattern, and finally, how might teleconnection patterns react to, or interact with, changing forcing such as a atmospheric composition changes, land-cover changes, or changes in the distribution of sea ice? The United States Climate Prediction Center actively monitors 13 separate teleconnection patterns in the Northern Hemisphere for weather and climate forecasting purposes. We will not discuss each pattern extensively, but instead use several of the major teleconnection patterns as illustrations of the remaining uncertainties.

We begin with the apparently more complicated teleconnections, those involving both the atmosphere and the oceans including El Niño/Southern Oscillation and the Madden–Julian Oscillation (MJO).

TYPES OF TELECONNECTIONS

Ocean to Atmosphere/Atmosphere to Ocean

El Niño/Southern Oscillation

El Niño/Southern Oscillation (ENSO) teleconnection patterns can be thought of as resulting from the interannual warming and cooling of equatorial Pacific sea surface temperatures (SSTs) and associated atmospheric circulation changes. This is an arbitrary starting point in the cycle as the changes in SST are likely themselves the result of distant atmospheric and oceanic forcing. ENSO appears to be the result of a series of internal interactions and feedbacks between ocean and atmosphere. The opposing phases of ENSO, the warm El Niño and the cold La Niña, though occurring quasi-periodically with roughly a 3–7 year cycle, have not proved to be highly predictable (Landsea and Knaff, 2000) despite considerable effort.

The ENSO cycle has weather and climate implications in the tropics and across the extra-tropics of both hemispheres. Climatologically, the warmest water in the equatorial Pacific occurs in the western Pacific warm pool (Figure 1a). The tropical signature of El Niño include large, persistent, warm SST anomalies in the eastern and central equatorial Pacific (Figure 1b), a relaxation of the easterly, near-surface winds; an anomalous tilting of the thermocline along the equator towards the east and an associated reduction in cold, nutrient-rich upwelling waters in the eastern equatorial Pacific which in turn affects fisheries production. Warm SST anomalies in the central and eastern Pacific are thought to be caused by an eastward traveling oceanic Kelvin wave following the relaxation of surface easterlies. This shift allows the main Pacific convective storm center from the western Pacific warm pool to shift to the east following the warm SST anomalies. The reduction in cold, upwelling waters in the eastern Pacific further enhances the warm SST anomaly. The La Niña pattern can be thought of as an amplification of the climatological SST patterns with unusually cold SSTs in the central and eastern Pacific and warm SSTs in the west (Figure 1c).

Rising motion due to convective storms in regions of high SST form the starting point for the entire large-scale, tropical circulation, including the north–south Hadley cell and the east–west Walker cells. Changes in the magnitude and spatial pattern of tropical convection therefore alter the magnitude and pattern of the Walker cells and affect the upper-level tropical outflow in the Hadley cell which feeds the higher latitude zonal jet (e.g. Bjerknes, 1969; Krishnamurti, 1961; Chen *et al.*, 1988; Oort and Yienger, 1996). The altered position of the Pacific Walker cell is such that large shifts in atmospheric mass occur with pressure drops in the eastern Pacific and increases to the west. This east-west change in pressure is the basis for the Southern Oscillation index (SOI), a measure of ENSO phase and strength.

Climatologically, rising air and therefore heavy precipitation occur in the western Pacific while the eastern Pacific is under the subsiding branch of the Walker circulation and so is relatively dry. El Niño causes shifts in tropical circulation, which generally create drier than average conditions in the western Pacific including Indonesia, Australia, and India and above average precipitation over parts of South America. El Niños also tend to cause warmer than average conditions over parts of the tropics and into the extra-tropics (Halpert and Ropelewski, 1992). El Niño patterns are so powerful that these events can generally be seen in a general warming of the area-averaged tropics. This warming can also be detected in the globally averaged temperature. For instance, 1998, a year of a very large El Niño event, is the warmest year of the satellite record in the global average. La Niña, on the other hand, is associated with enhanced rainfall in western Pacific regions and decreased rainfall in the central and east Pacific. La Niña is generally associated with cold regional anomalies that are less easily seen in the globally averaged temperature.

Individual El Niño events vary considerably from the average in terms of duration, time of onset and magnitude and appear to have certain longer-term fluctuations that hamper prediction. For instance, a strong and persistent correlation between reduced Indian monsoon rainfall

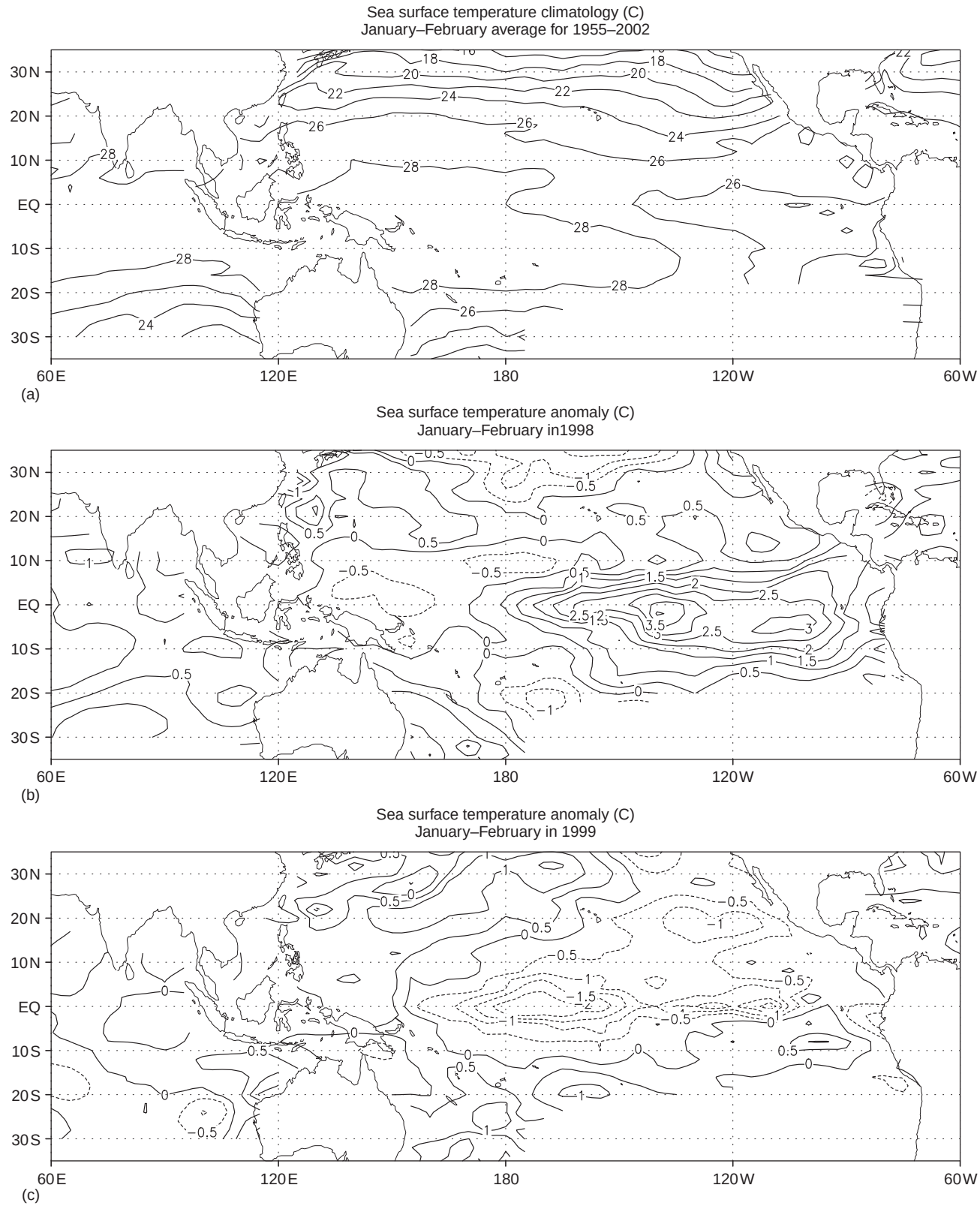


Figure 1 January and February tropical Pacific sea surface temperatures. (a) Climatology with the warmest temperatures in the western Pacific, (b) anomalies for 1998, a strong El Niño year with warmer than usual temperatures in the eastern Pacific, and (c) anomalies for 1999, a La Niña year with cool anomalies in the central and eastern Pacific

and El Niño has diminished in recent years for unknown reasons. Suggested mechanisms further illustrate the interconnectivity of the atmosphere–ocean system and include a higher latitude warming strong enough to enhance the monsoon and overcome the rain suppressing effects of El Niño or a small shift in El Niño convective anomalies, which in turn, shifts the downward branch of the Walker circulation away from India. (Kumar *et al.*, 1999).

Apart from affecting the mean zonal and meridional flow in the tropics, changes in upper-level outflow from tropical convection may also force anomalous atmospheric Rossby waves which can propagate to higher latitudes in a westerly background flow (e.g. Wallace and Gutzler, 1981; Tiedtke, 1984; James, 1994; Tribbia, 1991; Berbery and Nogues-Paegle, 1993). Such wave propagation out of the tropics into high latitudes in great arching patterns is a readily identifiable teleconnection pattern called the *Pacific North American Pattern* (PNA) which arches in a great circle northward from the tropical Pacific and eastward across North America and then southward towards the tropics. Discussion as to whether El Niño distinctly forces the PNA or simply modifies the statistics of an already existing mode of variability is ongoing (Straus and Shukla, 2002). Interestingly, details of each individual ENSO event may be very different in terms of season of onset, duration, strength, and exact location of convective anomalies. The extratropical patterns generated by Rossby waves tend to be roughly similar from event to event though the details of the pattern from event to event can be different (Hoerling and Kumar, 1997). This gave rise to the quite successful westerly duct hypothesis that Rossby waves can only escape the tropics at certain times of year and in certain locations because they cannot propagate in regions of ambient easterly winds. Winter season winds near Southeast Asia turn from easterly to westerly so waves generated within the tropics by shifts in convective activity can escape to higher latitudes from generally the same region (Webster, 1981; Hoskins and Karoly, 1981). It also appears that the extratropical waves may be more a function of the position and strength of the Southeast Asian jet on which the waves are excited, making the response less sensitive to the details of the tropical convective anomalies (Sardeshmukh and Hoskins, 1998).

Regionally, ENSO teleconnections can also be quite important.

In Northern Hemisphere winter, more intense storms occur farther north during El Niño years. Warmer than average conditions occur in the northern United States, eastern Canada, and near Japan. Extratropical effects for La Niña are, in some sense, the opposite of those caused by El Niño with warm conditions in the southern United States and cold, wet conditions in the northwest United States, Japan and southern Africa in northern winter. Correlations with ENSO in many climate variables have been reported across the globe though these correlations tend to be relatively weak and not highly statistically significant and therefore do not offer strong predictability in any single event. Strongly significant and repeatable teleconnections cover only a small portion of the area of the globe but are still highly important climatologically. In other regions, ENSO can offer probabilistic forecasting information.

While the chain of events constituting an El Niño or La Niña is well recognized and provides some statistical predictability in regions around the globe once the phase of ENSO is established (discussed in the Section "Predictability: numerical forecasts using teleconnection patterns"), actual understanding of the mechanisms which start the chain of events or stop it once started have been elusive, and the prediction of the onset of El Niños by both dynamical and physical models have yet to display skill relative to a simple climatology and persistence model of ENSO onset (Landsea and Knaff, 2000). One theory for ENSO is the delayed-oscillator theory (Suarez and Schopf, 1988; Battisti and Hirst, 1989) which posits an unstable atmosphere–ocean system where oceanic Rossby waves generated from previous El Niños or La Niñas act as the excitation for the next ENSO event. Other mechanisms such as monsoonal activity (Webster and Yang, 1992) and the MJO (discussed in the Section "Madden–Julian oscillation") have also been proposed as modulators of ENSO though such theories only lengthen the chain of causality as these phenomena are themselves even less understood than ENSO. Finally, Penland and Sardeshmukh (1995) have hypothesized that the tropical Pacific atmosphere/ocean system is not an unstable system waiting for triggering mechanisms but that the ENSO variability is best thought of as a response of the tropical Pacific system to stochastic climate noise.

Recent trends in ENSO and other teleconnection patterns will be discussed in the climate change section below.

Madden–Julian Oscillation

An example of the difficulties in associating an observed teleconnection pattern with physical mechanisms is the Madden–Julian Oscillation. First identified in the 1970s, the MJO is characterized by an observed, eastward moving atmospheric circulation anomaly and associated convection anomalies that can be identified in the wind, cloud, and outgoing long-wave radiation (OLR) fields along the equator with an approximate 40–50 day time period. The convective anomalies are strongest over the Indian Ocean and eastward over the west Pacific warm pool to the date line. Little sign of convective anomalies appear from the central through eastern Pacific. The MJO differs from the relatively spatially fixed teleconnection patterns such as the Arctic Oscillation (discussed below) in that it travels across the Pacific with a speed of approximately $5–10\,\mathrm{m\,s^{-1}}$.

The exact nature of the forcing of the MJO and its method of eastward propagation have eluded adequate theoretical

explanation as yet (Waliser *et al.*, 1999) and the oscillation is not well represented in model simulations (Slingo *et al.*, 1996). This is a problem in that the MJO dominates tropical climate variability at intra-annual timescales (while ENSO dominates interannual variability). Theories for the initiation and propagation of the MJO have centered on the wave convective instability of the second kind (CISK) mechanism (Lau and Peng, 1987) and the wind evaporation feedback (Emmanuel, 1987; Neelin *et al.*, 1987). However, both have failed to produce adequate representations of the MJO. Wave CISK theory typically produces oscillations with phase speeds of $15\,\mathrm{m\,s^{-1}}$ or greater, which is significantly faster than observed phase speeds. Wind-evaporative feedback mechanisms require easterly winds at the surface. While surface winds are easterly in much of the tropics, regions where the oscillation is most noticeable have climatological westerly winds at the surface.

The idea that the MJO is part of a coupled atmosphere–ocean oscillation, similar to ENSO, is the subject of active research (e.g. Waliser *et al.*, 1999; Woolnough *et al.*, 2000). Recent research (Seo and Kim, 2003) conclude that the MJO is a coupled oscillation of the ocean–atmosphere system and represents an interaction between two classes of waves, Rossby and Kelvin waves, leading to a self-generating and self-propagating disturbance.

The MJO is associated with the timing of the active and break period of both the Indian and Australian monsoons (Madden and Julian, 1994) and may have some role in triggering ENSO events (Kessler *et al.*, 1996; Zhang and Gottschalk, 2002) further complicating ENSO prediction. There does not appear to be a strong signal of the MJO in the extra-tropics (Madden and Julian, 1994).

Teleconnections in the Atmosphere

North Atlantic Oscillation-NAO/Arctic Oscillation-AO

Whether the major mode of mid and high Northern Hemisphere variability is better characterized as a regional oscillation known as the *North Atlantic Oscillation* (NAO) or a circumpolar mode referred to as the Arctic Oscillation (AO, Deser, Clara, 2000; Wallace and Thompson, 2002; Ambaum *et al.*, 2001) remains an open question illustrating the sometimes ambiguous way the atmospheric circulation organizes itself.

The NAO is a north–south oscillation of mass between the subtropical North Atlantic and Arctic. A measure of the NAO is an index generally defined as a pressure difference between a high-latitude station representative of the Icelandic low (Reykjavik or Stykkisholmur, Iceland) and a subtropical station (Lisbon or Gibraltar) representative of the other center of action in the Azores surface high-pressure system. A positive NAO index is an indication of more meridional flow across the Atlantic, which allows for warmer and more moist conditions in northwestern Europe while the negative phase is an indication of zonal flow and colder temperatures in western Europe and more moisture in southwestern Europe. The phase of the NAO also modulates climate in eastern North America (Wettstein and Mearns, 2002) and North Atlantic, particularly in winter months. The NAO, like ENSO, is an interannual oscillation with an irregular pattern of several years.

A related pattern, the AO, has been recently thought to be a more general mode of variability that actually includes the oscillation in the North Atlantic but expands it to a more symmetric annular mode, meaning that mass oscillates between the Arctic and lower latitudes in a giant ring around the circumference of the globe (Figure 2). The Arctic oscillation is seen in the first EOF in sea level pressure as three main centers of action, one in the Arctic, and two in the lower latitude centers in the North Atlantic and the North Pacific (the oscillation is not perfectly annular). The North Pacific center is substantially weaker than the North Atlantic center, giving rise to controversy as to which pattern, the AO or NAO, is actually climatologically more significant, and whether the AO is simply an artifact of statistical analysis (Deser, Clara, 2000; Ambaum *et al.*,

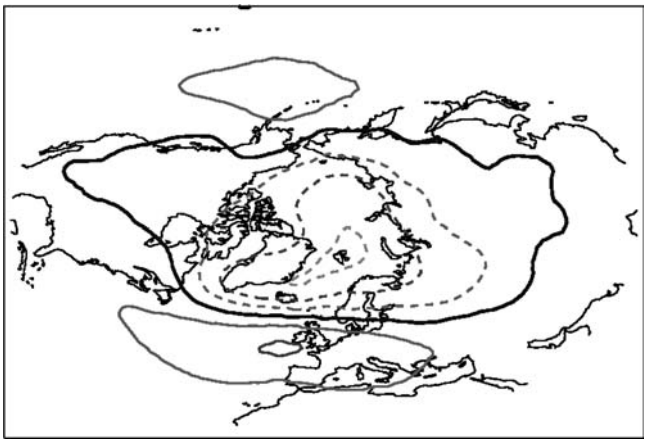


Figure 2 First EOF of sea level pressure from NCAR/NCEP Reanalysis showing the major mode of variability in the Northern hemisphere circulation including both the NAO and the AO. EOF analysis shows the preferred spatial organization of variability over time. In this case, three nodes are visible, one centered in the Arctic Ocean and two others at low latitudes in the Pacific and Atlantic oceans of opposite sign (thick dark contour is zero). The interpretation of this figure is that atmospheric mass oscillates back and forth between the high-latitude node and the two low latitude nodes. If one only looks at the Arctic-North Atlantic nodes this is the NAO. All three nodes comprise the AO and suggest a more annular character as the zero line extends around the globe. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

2001). Pressure in the two lower latitude centers of action is observed to be anticorrelated with pressure in the Arctic as would be expected in a coherent, annular movement of mass from high latitudes to lower latitudes. However, the two lower latitude centers of action are only very weakly correlated with each other suggesting that the AO is not really a globally coherent pattern. Wallace and Thompson (2002) suggest that other modes of variability are masking the coherence of the lower latitude pattern but it is still unclear which pattern will become favored.

In an effort to produce an idealized annular mode in a general circulation model, Cash *et al.* (2002) applied zonally symmetric surface boundary conditions and examined individual annular mode events (as opposed to a long-term average) and conclude that, even with no zonal changes in surface conditions, the primary mode is best conceived of as a series of regionally localized, north-south shifts in mass more in line with the NAO conception of Northern Hemisphere variability.

Monahan *et al.* (2001) suggest that the two patterns are not independent and the question may come down to the subjective judgment as to which paradigm organizes thinking most productively as suggested by Wallace (2000).

Teleconnection patterns are of interest because of the potential for long-term predictions if the source of the teleconnection can be understood. This is still not the case with the NAO/AO. Many studies have examined the factors involved in forcing variability in the AO/NAO and have found a variety of potential mechanisms within the earth system: SST (Rodwell *et al.*, 1999; Schneider *et al.*, 2003; Mehta *et al.*, 2000), snow cover (Gong *et al.*, 2002), volcanism (Stenchikov *et al.*, 2002), random stochastic variability (Schneider *et al.*, 2003; Tanaka, 2003; Wunsch, 1999), and stratospheric dynamics (Zhou *et al.*, 2001; Baldwin and Dunkerton, 1999; Black, 2002). Robertson (2001) concludes that the Arctic Oscillation is an inherent mode of atmospheric variability alone and that coupling a model to an interactive ocean does not change the simulated AO relative to a simulation with fixed SSTs suggesting limited predictability even with a knowledge of SST distribution.

Pacific Decadal Oscillation-PDO

The Pacific Decadal Oscillation (PDO) is a longer-term oscillation than those discussed previously. The average PDO phase persists for several decades though the period for an individual event is also quite a bit more variable than previously discussed patterns. PDO oscillations have energy peaks at both 15–25 year and 50–70 year timescales (Minobe, 1997) and is an example of interdecadal atmospheric variability. Such long-term patterns are of great interest because of the possibility of long-term predictive skill. However, as with other teleconnection patterns, the source of the PDO is not currently known and the predictability has not proved highly successful. It also remains unclear whether the PDO has been a consistently dominant mode of variability over long time periods. For instance, Gedalof and Mantua (2002) conclude that the PDO was less of a climatic factor in the nineteenth century than at present based on long-term proxy records.

While the PDO teleconnection pattern looks somewhat similar to El Niño, the strongest anomalies are in the extratropics with secondary signatures in the tropics, the opposite as observed with El Niño. The similarity in patterns may be explained by the observation that the PDO may actually be forced by ENSO anomalies, which are subsequently projected onto lower frequency variations by stochastic climate variability (Newman *et al.*, 2003). Such a forcing mechanism, if it proves robust, would make the PDO another coupled atmosphere–ocean mode of variability but we categorize it as atmospheric only until further evidence is accumulated. The PDO seems also to have utility in long-range forecasting. For example, Castro *et al.* (2001) have used a combination of ENSO and the PDO to diagnose summer precipitation patterns and temporal evolution in the western United States.

Teleconnections From Land to Atmosphere

The high interconnectivity of the atmosphere–ocean system suggests that other perturbations to the earth system might be reflected far from the original disturbance and may have some influence on previously discussed teleconnection patterns. Here, we examine the example of land-cover changes in detail but a variety of processes, such as perturbations due to atmospheric pollution including aerosol clouds, may be operating in very complicated ways which are only beginning to be appreciated or investigated.

Large-scale land-cover changes, particularly in the tropics appear to generate remote climatic effects and may interact with better-known teleconnection patterns. The three major tropical convective heating centers are associated with the land surfaces of Africa, Amazonia, and the maritime continent of Indonesia, Malaysia, New Guinea, and surrounding regions (e.g. Kreuger and Winston, 1973). Changes in this vegetation structure has major impacts on the momentum and radiant energy absorbed at the surface and its partitioning into latent and sensible forms which affect surface temperatures and the structure and strength of convective storms (e.g. Dickinson and Kennedy, 1992; Nobre *et al.*, 1991; Eltahir, 1996; Polcher and Laval, 1994; Baidya Roy and Avissar, 2002).

Teleconnections resulting from land-cover changes in climate models have been discussed by Franchito and Rao (1992), McGuffie *et al.* (1995), Chase *et al.* (1996) and Zhang *et al.* (1996). Others have also noted isolated extratropical effects due to simulated tropical vegetation changes (Sud *et al.*, 1996; Sellers *et al.*, 1996). Chase *et al.* (2000) examined general circulation model (GCM) model simulations of the effect of observed levels of land-cover change globally and found strong evidence of

changes in global scale circulations and for the propagation of Rossby waves into the mid latitudes. Pitman and Zhao (2000), Zhao *et al.* (2001), Bounoua *et al.* (2002) again demonstrated that the remote effects of observed levels of land-cover change were prevalent in a variety of models under a range of configurations and model assumptions and that remote temperature anomalies resulting from land-cover change could be similar in magnitude as effects of the historical increase of the radiative effect of increased CO_2 (Chase *et al.*, 2002). Gedney and Valdes (2000), also using a GCM, specifically examined the effects of a wholesale removal on the Amazonian rainforest on remote climates and found significant evidence for a reduction in large-scale circulations generated by tropical convection and for propagating Rossby waves which affected rainfall in Northern Hemisphere winter. Werth and Avissar (2002) and Avissar and Werth (2004) find statistically significant teleconnection patterns due to deforestation in Amazonia, Central Africa, and Southeast Asia (Figure 3). Defries *et al.* (2002) examined potential impacts of future land-use changes and found regional temperature anomalies of up to 1.5 °C in regions not directly affected by land-cover changes. Such teleconnection patterns due to human activity might be expected to interact with natural modes of variability and this is an ongoing area of research.

Teleconnection patterns also strongly affect the biosphere, which implies high levels of interaction and mutual self-adjustment between these components of the earth system. For instance, Asner *et al.* (2000) attribute changes in net primary production in Amazonian forests of up to 18% due to ENSO oscillations and found that El Niño years were responsible for large fluxes of CO_2 into the atmosphere in the tropics. Kitzberger *et al.* (2001) found that forest fires in the Southwest United States and in Patagonia, Argentina were related to phases of ENSO. El Niño years are wet in these regions allowing plant growth and accumulation of fuel. La Niña years are dry in these regions and so the high fuel loads become desiccated leading to high rates of burning. Additionally, the effects of teleconnection patterns are not limited to the primary producers. Nott *et al.* (2002) found that seasonal ENSO and NAO weather

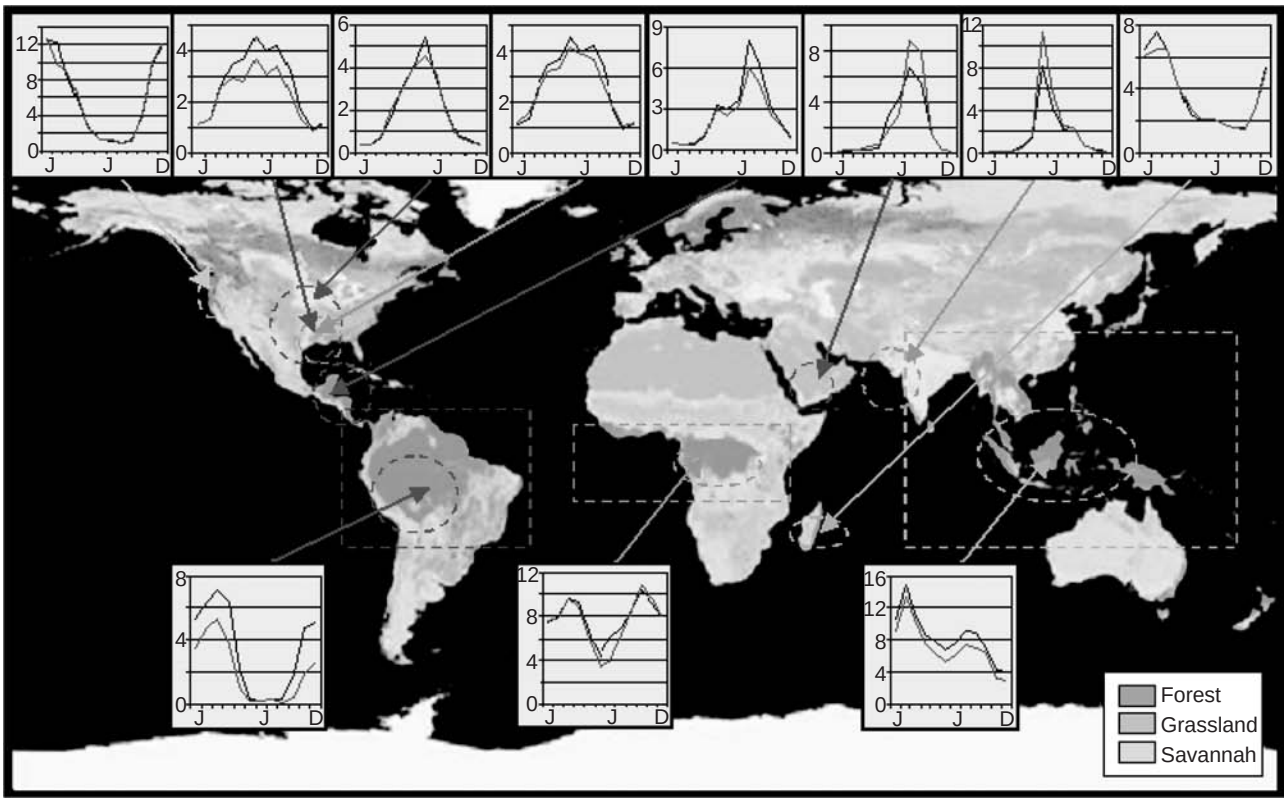


Figure 3 Annual cycle of precipitation (mm day^{-1}) in continental regions particularly affected by large-scale deforestation of Amazonia (dark grey), Central Africa (grey), and South-East Asia (light grey). The dark-grey curves show simulated mean monthly precipitation before massive tropical deforestation (i.e. the ''control'' case). whereas the grey curves show precipitation after deforestation. The size and location of the color-coded areas corresponding to the deforested regions are at scale. Ellipses indicate regions where tropical forest (dark grey on the 1-km resolution land-cover map used for the background) was replaced with a mixture of shrubs and grassland (derived from Avissar and Werth, 2004). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

changes resulted in strong effects on the productivity of a variety of bird populations.

Regional Teleconnections

While the term teleconnection is usually applied to patterns with continental-scale variations that are long lasting, the term can also be generalized to distant influences resulting from changes in smaller scale circulations, or from changes in ecosystem function. For example, Chase *et al.* (1999) found weather influences in the high Rocky Mountains due to the presence of irrigated farmland in the plains below. Irrigated regions affected the daily summer mountain-plains breeze by altering temperature patterns thereby allowing communication between the two regions. Such changes in local circulation regime would be expected to alter the transport of pollutants or atmospherically transported micronutrients and so may be of importance in many locales. Along similar lines, Eastman *et al.* (2001) described a "biological teleconnection" where changed ecosystem characteristics affected local weather such that effects were communicated to regions to regions distant from actual changes in surface characteristics.

A final example of the complex regional nature of some teleconnections and how human effects on climate might be difficult to pinpoint is an observational study of the effect of irrigation on downwind rainfall patterns. Moore and Rojstaczer (2002) found a maximum increase in summer precipitation some 90 km away from the irrigated region.

PREDICTABILITY: NUMERICAL FORECASTS USING TELECONNECTION PATTERNS

The atmosphere-ocean system is nonlinear making long-term prediction of individual weather events impossible. Even assuming a perfect climate model, small errors in the observed initial conditions invariably grow exponentially in magnitude and spatial scale. Therefore, long-term weather and climate forecasts use expected probabilities of certain types of weather based on statistical relationships from past observations. These relationships often take the form of teleconnection patterns.

One major advance in long-range forecasts is the realization that ENSO has documented effects in many parts of the globe. Many regions show a statistical tendency towards more or less precipitation or higher/lower temperatures depending on the phase of ENSO. ENSO has a fairly regular periodicity allowing for some skill in predicting changes in phase just from climatology or persistence. Several dynamical models also try to predict the future phase and resulting teleconnections of ENSO though these have not been dramatically more successful than knowledge of the climatology and persistence (Landsea and Knaff, 2000). The phase of ENSO is the single most important factor going into long-range forecasts today.

A statistical technique called a *canonical correlation analysis*, used in long-range climate forecasts, combines a series of indicators, including teleconnection indices, to infer possible preferred future patterns. This technique uses model-simulated weather patterns, global SST patterns, surface temperature, and precipitation for the past year to infer information about persistence and trends over the year. ENSO is emphasized in this analysis but other natural modes of variability such as the NAO are also accounted for. This analysis makes use of the forecast for La Niña, El Niño, or neutral conditions in the equatorial Pacific and then takes into account the confidence that this one phase of ENSO will exist.

Such information is used in a series of forecast simulations that differ slightly in order to estimate the envelope of possible atmospheric responses under such conditions. The relative occurrence of a particular climate pattern in these ensemble forecasts determines the probability of an individual event being forecast.

CLIMATE CHANGE

"Greenhouse Gas Warming" and Projection on Natural Modes

The major teleconnection patterns such as ENSO, NAO/AO, and the PDO occur in an irregularly periodic manner. Longer-term structure characterizes each of these patterns with periods of stronger or weaker intensity, altering periodicity and shifts favoring one phase of the oscillation over the other. The variability of climate due to these teleconnection patterns is therefore variable over longer timescales. Such longer-term variability has yet to be explained and appears to be particularly important in assessing present-day and future human impacts on climate.

For example, recent shifts favoring the warmer phase of two natural teleconnection patterns, ENSO and NAO/AO have been directly linked to a large portion of the observed Northern Hemisphere winter warming signal (Palecki and Leathers, 1993; Hurrell, 1996; Corti *et al.*, 1999). A trend in the NAO index toward more positive values since the early 1960s has been documented (Hurrell, 1996). Similarly, the observed SO index has shown a tendency towards more negative (El Niño-like) values since the middle of the century with a steep change to more negative values in the mid 1970s. Hurrell (1996) demonstrates that when these two natural circulation influences are removed from the time series, no discernible upward surface temperature trend remains (See Figure 4 in Hurrell, 1996). While such observations might be taken as an indication that recently observed warming is natural rather than a result of rising greenhouse gases, Corti *et al.* (1999) argue that greenhouse warming might be expressed in terms of changes in natural modes of variability. Stone *et al.* (2001) do find a general

projection of climate change on the most dominant modes of variability suggesting that these need to be actively monitored for future change.

Therefore, the question remains whether recent shifts towards more and larger El Niños is a reflection of natural variability or is forced by human activity. Cobb *et al.* (2003), examining isotope signatures in fossil corrals, conclude that past variability of ENSO cycles is unrelated to the mean temperature and that eras in the past 1100 years have seen ENSO cycles that rival present-day fluctuations. This observational study supports the statistical analyses of Rajagopalan *et al.* (1997) and Wunsch (1999) who both found the present trend toward more and larger El Niño events to be within the bounds of natural variability. Trenberth and Hoar (1996), however, in a differing statistical analysis, found recent trends to be statistically unusual and concluded that this was evidence for human influence on climate.

Reports from model simulations as to circulation changes due to increasing greenhouse gases are, at present, contradictory. There have been reports of changes, which favor a positive shift in the Southern Oscillation (more La Niña-like) (e.g. Timmerman *et al.*, 1999; Hu *et al.*, 2001) while others find a tendency for increasing negative phase (e.g. Meehl *et al.*, 2000; Collins, 2000). Still others find no change (e.g. Tett, 1995) or an increase in amplitude in both phases of the SO but no clear favoring of one phase over the other. Additionally, reported changes in the SO typically occur at CO_2 levels far above present levels of forcing and are therefore not entirely applicable to present-day conditions.

Simulated changes in the NAO/AO under increased greenhouse gases and/or aerosols also have a quite complicated response between simulations. Paeth *et al.* (1999) show a steadily increasing NAO index in climate change simulations starting at about the correct time but find no statistical significance. Shindell *et al.* (1999) show a positive trend in model-simulated NAO with present-day levels of CO_2 forcing, however, the trend between 1959 and 2000, the period of observed increase in the NAO index, is static (see Shindell *et al.*, 1999: Figure 2b). Fyfe *et al.* (1999), also show an increase in the NAO but only at much higher levels of CO_2 forcing than presently observed.

Further complicating the picture, Osborn *et al.* (1999) find the opposite effect with a decreasing NAO index in climate change simulations starting at present-day and continuing through the century. Zorita and Gonzalez-Rouco (2000), in a head-to-head comparison of two climate models, found highly variable results between the models. In the first model examined, they found positive AO trends in the first realization and negative AO trends in the second realization using different initial conditions. The second model had a clear positive trend in two realizations though at differing times in the simulation (i.e. under different greenhouse forcing) so that present observations are still not well explained.

Finally, it is unclear how robust the results from any single model are. For example, Collins (2000) found a shift towards a more El Niño-like state at four times natural CO_2 (approximately 12 times present levels) though when small details of the model formulation were changed the simulation produced the opposite change in circulation.

Other Human Influences on Teleconnection Patterns

There exists some evidence for a highly complicated human effect on natural climate variability. Chase *et al.* (1996) found that climate model-simulated tropical circulation shifts due to historical changes in vegetation were consistent with conditions favorable for inducing El Niño events at the expense of La Niña while Chung and Ramanathan (2003) found in model simulations that atmospheric haze originating in southern Asia and due mostly to human activity (such as clearing of agricultural land) also led toward conditions which favored El Niño at the expense of La Niña.

Human effects on teleconnection patterns may also be quite indirect.

For example, Alexander *et al.* (2004) find a NAO-like pattern in response to model-simulated changes in Arctic sea ice extent suggesting that should a warming climate cause large changes in sea ice, a large part of the effect would be seen in changes in circulation patterns. Similarly, changes in snow cover in a warming world would be expected to affect other large-scale circulations, such as the Indian monsoon system (Blanford, 1884; Fasullo, 2004a), which also interacts with ENSO. While such interactions require further confirmation, the potential effect of human activity on climate and circulation variability is apparently multifaceted and complex.

ISOLATED OR INTERCONNECTED?

A fundamental question concerning the major patterns of teleconnections is how independent they are. Because a definitive "cause" of any teleconnection pattern is elusive, it is possible that subsets of the patterns are in reality interrelated and part of a larger oscillatory phenomenon. Signs of interrelatedness are appearing more frequently in studies of teleconnection patterns.

For instance, there are indications that tropical Pacific SST patterns and hence ENSO may have some impact on the evolution of the NAO/AO pattern (Hoerling *et al.*, 2001; Schneider *et al.*, 2003). Further, it appears that South Pacific SST patterns may influence the development of tropical SSTs on decadal timescales (Bratcher and Giese, 2002). Hakkinen and Mo (2002) find that tropical

Atlantic Ocean temperature anomalies in boreal winter are related to forcings in the North Atlantic due to NAO fluctuations that generate equatorward propagating Kelvin waves, and are also influenced by teleconnections from the tropical Pacific, which can either work in conjunction or in opposition to each other. Gong and Ho (2003), find a significant relationship between the AO and East Asian summer monsoon rainfall due to a northward shift of the East Asian jet stream. As discussed previously, the Asian summer monsoon is known to interact with ENSO and the MJO. Further, Yang *et al.* (2002) find a teleconnection between the strength of the East Asian jet stream and weather downstream in East Asia, the Pacific, and North America, which appear distinct from ENSO patterns. Branstator (2002) further highlights the importance of remote effects due to the East Asian jet by showing that perturbations to the jet can be circumglobal in nature (an indication that the regional analysis of teleconnection patterns may be fundamentally misleading) and may, in part, be responsible for the patterns associated with the NAO. Kiladis and Weickman (1992) find forcing of El Niño variability by high-latitude storms. Xie and Tanimoto (1998) find a decadal teleconnection pattern spanning both hemispheres from the southern subtropics to the high northern latitudes apparently unrelated to the other patterns mentioned here but suggesting that the major teleconnection patterns may be linked by a variety of mechanisms. Miller *et al.* (2003) find a statistical association between the phases of the AO and the MJO further suggesting regular tropical/extratropical interactions between modes of atmospheric variability.

Therefore, it appears that evidence is emerging that the climate system is coupled in a variety of complicated ways and that conceiving of variability in terms of a series of isolated teleconnection patterns may give way to a view that each of the patterns is interrelated in some way, each forcing and being forced by the others. Long chains of causality linking some or all modes of variability might improve predictability if the chains of events are regular, though past experience indicates that relationships between the modes vary with time.

SUMMARY

This discussion illustrates the large-scale connectivity of the atmosphere–ocean coupled system and generalizes the concept to regional scales and to other components of the earth system. These connections at a distance, referred to as teleconnections, can occur by the direct transfer of mass by changes in regular circulations or by propagating waves initiated by a variety of mechanisms.

We have not discussed in detail several processes, which could rightfully be included in this section such as the regional monsoon systems, local winds, or the oceanic thermohaline circulation that, if changed, could have large climate repercussions all around the globe. We have, however, addressed the basic remaining uncertainties as to the nature of teleconnection patterns with prominent examples. Questions remain as to what extent recognized teleconnection patterns can be associated with an identifiable forcing mechanism, to what extent these patterns are interrelated and how they might cause, react to, or interact with changing forcing such as changes in atmospheric composition, land cover, or the distribution of sea ice to produce climate changes?

Acknowledgments

This work was supported under the following grants: NSF ATM-0001476, ATM-0346554; NASA NAG5-11400, NAG5-11402, NAG5-13781, NAG5-11370; NOAA NA17-RJ1228 Amendment 6. We thank Eungul Lee and Aaron Rivers for graphics assistance and two anonymous referees for their helpful comments.

FURTHER READING

Deweaver E. and Nigam S. (2002) Linearity in ENSO's atmospheric response. *Journal of Climate*, **15**, 2446–2461.

REFERENCES

Ambaum M.H.P., Hoskins B.J. and Stephenson D.B. (2001) Arctic oscillation or North Atlantic oscillation. *Journal of Climate*, **14**, 3495–3507.

Alexander M.A., Bhatt U.S., Walsh J.E., Timlin M.S., Miller J.S. and Scott J.D. (2004) The atmospheric response to realistic Arctic sea ice anomalies in an AGCM during winter. *Journal of Climate*, **17**, 890–905.

Asner G.P., Townsend A.R. and Braswell B.H. (2000) Satellite observation of El Niño effects on Amazon forest phenology and productivity. *Geophysical Research Letters*, **27**, 981–984.

Avissar R. and Werth D. (2004) Global hydroclimatological teleconnections resulting from tropical deforestation. *Journal of Hydrometeorology*, (in press).

Baidya Roy S. and Avissar R. (2002) Impact of land use/land cover on regional hydrometeorology in the Amazon. *Journal of Geophysical Research*, **107**, 8037, doi:10.1029/2000JD000266.

Baldwin M.P. and Dunkerton T.J. (1999) Propagation of the Arctic oscillation from the stratosphere to the troposphere. *Journal of Geophysical Research*, **104**, 30937–30946.

Barnston A.G. and Livzey R.E. (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, **115**, 1083–1126.

Battisti D.S. and Hirst A.C. (1989) Inter-annual variability in a tropical atmosphere-ocean model: influence of the basic state, ocean geometry and nonlinearity. *Journal of the Atmospheric Sciences*, **46**, 1687–1712.

Berbery E.H. and Nogues-Paegle J. (1993) Intraseasonal interactions between the tropics and extra-tropics in the southern hemisphere. *Journal of the Atmospheric Sciences*, **50**, 1950–1965.

Bjerknes J. (1969) Atmospheric teleconnections from the equatorial Pacific. *Monthly Weather Review*, **97**, 163–172.

Black R.X. (2002) Stratospheric forcing of surface climate in the Arctic oscillation. *Journal of Climate*, **15**, 268–277.

Blanford H.F. (1884) On the connection of the Himalaya snowfall with dry winds and seasons of drought in India. *Proceedings of the Royal Society of London*, **37**, 3–22.

Bounoua L., Defries R., Collatz G.J., Sellers P. and Khan H. (2002) Effects of land-cover conversion on surface climate. *Climate Change*, **52**, 29–64.

Branstator G. (2002) Circum-global teleconnections, the jet stream waveguide, and the North Atlantic oscillation. *Journal of Climate*, **15**, 1893–1910.

Bratcher A.J. and Giese B.S. (2002) Tropical Pacific decadal variability and global warming. *Geophysical Research Letters*, **29**, 1918, 10.1029/2002GL015191.

Cash B.A., Kushner P.J. and Vallis G.K. (2002) The structure and composition of the annular modes in an aquaplanet general circulation model. *Journal of the Atmospheric Sciences*, **59**, 3399–3414.

Castro C.L., McKee T.B. and Pielke R.A. Sr (2001) The relationship of the North American monsoon to tropical and North Pacific sea surface temperatures as revealed by observational analysis. *Journal of Climate*, **14**, 4449–4473.

Chase T.N., Pielke R.A. Sr, Kittel T.G.F., Baron J.S. and Stohlgren T.J. (1999) Potential impacts on Colorado rocky mountain weather and climate due to land use changes on the adjacent great plains. *Journal of Geophysical Research*, **104**, 16673–16690.

Chase T.N., Pielke R.A., Kittel T.G.F., Nemani R.R. and Running S.W. (1996) Sensitivity of a general circulation model to global changes in leaf area index. *Journal of Geophysical Research*, **101**, 7393–7408.

Chase T.N., Pielke R.A. Sr, Kittel T.G.F., Nemani R.R. and Running S.W. (2000) Simulated impacts of historical land-cover changes on global climate in northern winter. *Climate Dynamics*, **16**, 93–105.

Chase T.N., Pielke R.A. Sr, Kittel T.G.F., Zhao M., Pitman A.J., Nemani R.R. and Running S.W. (2002) Relative climatic effects of land-cover change and elevated carbon dioxide combined with aerosols: a comparison of model results and observations. *Journal of Geophysical Research*, **106**(31), 685–31, 691.

Chen T.-C., Tzeng R.-Y. and Van Loon H. (1988) Study on the maintenance of the winter subtropical jet streams in the Northern Hemisphere. *Tellus*, **40A**, 392–397.

Chung C.E. and Ramanathan V. (2003) South Asian haze forcing: remote impacts with implications to ENSO and AO. *Journal of Climate*, **16**, 1791–1806.

Cobb K.M., Charles C.D., Cheng H. and Edwards R.L., (2003) El Niño/Southern Oscillation and tropical Pacific climate during the last millennium. *Nature*, **424**, 271–276.

Collins M. (2000) Understanding uncertainties in the response of ENSO to greenhouse warming. *Geophysical Research Letters*, **27**, 3509–3512.

Corti S., Molteni F. and Palmer T.N. (1999) Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, **398**, 799–802.

Defries R.S., Bounoua L. and Collatz G.J. (2002) Human modification of the landscape and surface climate in the next fifty years. *Global Change Biology*, **8**, 438–458.

Deser, Clara (2000) On the teleconnectivity of the "Arctic Oscillation". *Geophysical Research Letters*, **27**, 779–782.

Dickinson R.E. and Kennedy P.J. (1992) Impacts on regional climate of Amazonian deforestation. *Geophysical Research Letters*, **19**, 1947–1950.

Eastman J.L., Coughenour M.B. and Pielke R.A. (2001) The effects of CO_2 and landscape change using a coupled plant and meteorological model. *Global Change Biology*, **7**, 797–815.

Eltahir E.A.B. (1996) Role of vegetation in sustaining large-scale atmospheric circulations in the tropics. *Journal of Geophysical Research*, **101**, 4255–4268.

Emmanuel K.A. (1987) An air-sea interaction model of intraseasonal oscillations in the tropics. *Journal of the Atmospheric Sciences*, **44**, 2324–2340.

Fasullo J. (2004a) A stratified diagnosis of the Indian monsoon-Eurasian snow cover relationship. *Journal of Climate*, **17**, 1110–1122.

Franchito S.H. and Rao V.B. (1992) Climatic change due to land surface alterations. *Climatic Change*, **22**, 1–34.

Fyfe J.C., Boer G.J. and Flato G.M. (1999) The Arctic and Antarctic oscillations and their projected changes under global warming. *Geophysical Research Letters*, **26**, 1601–1604.

Gedalof Z. and Mantua N.J. (2002) A multicentury perspective of variability in the Pacific decadal oscillation: new insights from tree rings and coral. *Geophysical Research Letters*, **29**, 2204, No. D24, doi:10:1029/2002GL015824.

Gedney N. and Valdes P.J. (2000) The effect of Amazonian deforestation on the Northern hemisphere circulation and climate. *Geophysical Research Letters*, **27**, 3053–3056.

Gong G., Entekhabi D. and Cohen J. (2002) A large ensemble model study of the wintertime AO-NAO and the role of interannual snow perturbations. *Journal of Climate*, **15**, 3488–3499.

Gong D.-Y. and Ho C.-H. (2003) Arctic oscillation signals in the East Asian summer monsoon. *Journal of Geophysical Research*, **108**(D2), 4066, doi:10.1029/2002JD002193.

Hakkinen S. and Mo K.C. (2002) The low-frequency variability of the tropical Atlantic. *Journal of Climate*, **15**, 237–250.

Halpert M.S. and Ropelewski C.F. (1992) Surface temperature patterns associated with the Southern Oscillation. *Journal of Climate*, **5**, 577–593.

Hoerling M.P., Hurrell J.W. and Xu T. (2001) Tropical origins for recent climate change. *Science*, **292**, 90–92.

Hoerling M.P. and Kumar A. (1997) Why do North American climate anomalies differ from one El Niño event to another? *Geophysical Research Letters*, **24**, 1059–1062.

Hoskins B.J. and Karoly D.J. (1981) The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of the Atmospheric Sciences*, **38**, 1179–1196.

Hu Z.-Z., Bengtsson L., Roeckner E., Christoph M., Bacher A. and Oberhurer J.M. (2001) Impact of global warming on the interannual and interdecadal climate modes in a coupled GCM. *Climate Dynamics*, **17**, 361–374.

Hurrell J.W. (1996) Influence of variations in extratropical wintertime teleconnections on Northern Hemisphere temperature. *Geophysical Research Letters*, **23**(6), 665–668, 20009.

James I.N. (1994) *Introduction to Circulating Atmospheres*, Cambridge University press: Cambridge.

Kessler W.S., McPhaden M.J. and Weickman K.M. (1996) Forcing of the intraseasonal Kelvin waves in the equatorial Pacific. *Journal of Geophysical Research*, **100**, 10613–10631.

Kiladis G.N. and Weickman K.M. (1992) Extratropical forcing of tropical Pacific convection during Northern winter. *Monthly Weather Review*, **120**, 1924–1937.

Kitzberger T., Swetnam T.W. and Veblen T.T. (2001) Interhemispheric synchrony of forest fires and El Niño-Southern Oscillation. *Global Ecology and Biogeography*, **10**, 315–326.

Kreuger A.F. and Winston J.S. (1973) A comparison of the flow over the tropics during two contrasting flow regimes. *Journal of the Atmospheric Sciences*, **31**, 358–369.

Krishnamurti T.N. (1961) The subtropical jet stream of winter. *Journal of Meteorology*, **18**, 172–191.

Kumar K.K., Balaji R. and Cane M.A. (1999) On the weakening relationship between the Indian monsoon and ENSO. *Science*, **284**, 2156–2159.

Landsea C.W. and Knaff J.A. (2000) How much skill was there in forecasting the very strong 1997–98 El Niño? *Bulletin of the American Meteorological Society*, **81**, 2107–2119.

Lau K.-M. and Peng L. (1987) Origin of low frequency (intraseasonal) oscillation in the tropical atmosphere. Part I: Basic theory. *Journal of the Atmospheric Sciences*, **44**, 465–472.

Madden R.A. and Julian P.R. (1994) Observations of the 40–50 day tropical oscillation–A review. *Monthly Weather Review*, **122**, 814–837.

McGuffie K., Henderson-Sellers A., Zhang H., Durbidge T.B. and Pitman A.J. (1995) Global climate sensitivity to tropical deforestation. *Global and Planetary Change*, **10**, 97–128.

Meehl G.A., Washington W.M., Arblaster J.M., Bettge T.W. and Strand W.G. Jr (2000) Anthropogenic forcing and decadal climate variability in sensitivity experiments of twentieth and twenty-first-century climate. *Journal of Climate*, **13**, 3728–3744.

Mehta V.M., Suarez M.J., Manganello J.V. and Delworth T.L. (2000) Oceanic influence on the North Atlantic oscillation and associated Northern hemisphere climate variations: 1959–1993. *Geophysical Research Letters*, **27**, 121–124.

Miller A.J., Zhou S. and Yang S.-K. (2003) Relationship of the Arctic and Antarctic oscillations to the outgoing longwave radiation. *Journal of Climate*, **16**, 1583–1592.

Minobe S. (1997) A 50–70 year climatic oscillation over the North Pacific and North America. *Geophysical Research Letters*, **24**, 683–686.

Monahan A.H., Pandolfo L. and Fyfe J.C. (2001) The preferred structure of variability of the Northern hemisphere atmospheric circulation. *Geophysical Research Letters*, **28**, 1019–1022.

Moore N. and Rojstaczer S. (2002) Irrigation's influence on precipitation: Texas high plains, U.S.A. *Geophysical Research Letters*, **29**(16), doi: 10.1029/2002GL014940.

Namias J. (1953) Thirty-day forecasting: a review of a ten-year experiment. *Meteorological Monographs. American Meteorological Society*, **2**(6), 83.

Namias J. (1959) Recent seasonal interactions between North Pacific waters and the overlying atmospheric circulation. *Journal of Geophysical Research*, **64**, 631–646.

Neelin J.D., Held I.M. and Cook K.H. (1987) Evaporation-wind feedback and low frequency variability of the tropical atmosphere. *Journal of the Atmospheric Sciences*, **44**, 2341–2348.

Newman M., Compo G.P. and Alexander M.A. (2003) ENSO-forced variability of the Pacific decadal oscillation. *Journal of Climate*, **16**, 3853–3857.

Nobre C.A., Sellers P.J. and Shukla J. (1991) Amazonian deforestation and regional climate change. *Journal of Climate*, **4**, 957–988.

Nott M.P., Desante D.F., Siegal R.B. and Pyle P. (2002) Influences of the El Niño/Southern Oscillation and the North Atlantic oscillation on avian productivity in forests of the Pacific Northwest of North America. *Global Ecology and Biogeography*, **11**, 333–342.

Oort A.H. and Yienger J.J. (1996) Observed inter-annual variability of the Hadley circulation and its connection to ENSO. *Journal of Climate*, **9**, 2751–2767.

Osborn T.J., Briffa K.R., Tett S.B.F., Jones P.D. and Trigo R.M. (1999) Evaluation of the North Atlantic oscillation as simulated by a coupled climate model. *Climate Dynamics*, **15**, 685–702.

Paeth H., Hense A., Glowienka-Hense R., Voss R. and Cubash U. (1999) The North Atlantic oscillation as an indicator for greenhouse-gas induced regional climate change. *Climate Dynamics*, **15**, 953–960.

Palecki M.A. and Leathers R.J. (1993) Northern hemisphere extratropical circulation anomalies and recent January land surface temperature trends. *Geophysical Research Letters*, **20**, 819–822.

Penland C. and Sardeshmukh P.D. (1995) The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, **8**, 1999–2024.

Pitman A. and Zhao M. (2000) The relative impact of observed changes in land cover and carbon dioxide as simulated by a climate model. *Geophysical Research Letters*, **27**, 1267.

Polcher J. and Laval K. (1994) Statistical study of the regional impact of deforestation on climate in the LMD GCM. *Climate Dynamics*, **10**, 205–219.

Rajagopalan B., Lall U. and Cane M.A. (1997) Anomalous ENSO occurrences: an alternate view. *Journal of Climate*, **10**, 2351–2357.

Rial J., Pielke R.A. Sr, Beniston M., Claussen M., Canadell J., Cox P., Held H., de Noblet-Ducoudre N., Prinn R., Reynolds J. and Salas J.D. (2004) Nonlinearities, feedbacks and critical thresholds within the Earth's climate system. *Climatic Change* **65**, 11–38.

Robertson A.W. (2001) Influence of ocean-atmosphere interaction on the Arctic oscillation in two general circulation models. *Journal of Climate*, **14**, 3240–3254.

Rodwell M.J., Rowell D.P. and Folland C.K. (1999) Oceanic forcing of the wintertime North Atlantic oscillation. *Nature*, **398**, 320–323.

Sardeshmukh P.D. and Hoskins B.J. (1998) The generation of global rotational flow by steady, idealized tropical divergence. *Journal of the Atmospheric Sciences*, **45**, 12.

Schneider E.K., Bengtsson L. and Hu Z.-Z. (2003) Forcing of Northern hemisphere climate trends. *Journal of the Atmospheric Sciences*, **60**, 1504–1521.

Sellers P.J., Bounoua L., Collatz G.J., Randall D.A., Dazlich D.A., Los S.O., Berry J.A., Fung I., Tucker C.J., Field C.B., *et al.* (1996) Comparison of radiative and physiological effects of doubled atmospheric CO_2 on climate. *Science* **271**, 1402–1406.

Seo K.-H. and Kim K.-Y. (2003) Propagation and initiation mechanisms of the Madden–Julian Oscillation. *Journal of Geophysical Research*, **108**(D13), 4384, doi:10.1029/2002JD002876.

Shindell D., Miller R.L., Schmidt G.A. and Pandolfo L. (1999) Simulation of recent northern winter climate trends by greenhouse-gas forcing. *Nature*, **399**, 452–455.

Slingo J.M., Sperber K.R., Boyle J.S., Ceron J.-P., Dix M., Dugas B., Ebisuzaki W., Fyfe J., Gregory D., Gueremy J.-F., Hack J., Harzallah A., Inness P., Kitoh A., Lau W.K.-M., McAvaney B., Madden R., Matthews A., Palmer T.N., Park C.-K., Randall D. and Renno N. (1996) Intraseasonal oscillations in 15 atmospheric general circulation models: results from an AMIP diagnostic subproject. *Climate Dynamics*, **12**, 325–357.

Stenchikov G., Robock A., Ramaswamy V., Daniel Schwartzkopf M., Hamilton K. and Ramachandran S. (2002) Arctic oscillation response to the 1991 Mount Pinatubo eruption: effects of volcanic aerosols and ozone depletion. *Journal of Geophysical Research*, **107**(D24), 4803, doi:10.1029/2002JD002090.

Straus D.M. and Shukla J. (2002) Does ENSO force the PNA? *Journal of Climate*, **15**, 2340–2358.

Stone D.A., Weaver A.J. and Stouffer R.J. (2001) Projection of climate change onto modes of atmospheric variability. *Journal of Climate*, **14**, 3551–3565.

Suarez M.J. and Schopf P.S. (1988) Delayed action oscillator for ENSO. *Journal of the Atmospheric Sciences*, **45**, 3283–3287.

Sud Y.C., Walker G.K., Kim J.H., Liston G.E., Sellers P.J. and Lau W.K.M. (1996) Biogeophysical consequences of a tropical deforestation scenario: a GCM simulation. *Journal of Climate*, **9**, 3225–3247.

Tanaka H.L. (2003) Analysis and modeling of the Arctic oscillation using a simple barotropic model with baroclinic eddy forcing. *Journal of the Atmospheric Sciences*, **60**, 1359–1379.

Tett S. (1995) Simulation of El Niño-Southern Oscillation-like variability in a global AOGCM and its response to CO_2 increase. *Journal of Climate*, **8**, 1473–1502.

Tiedtke M. (1984) The effect of penetrative cumulus convection on the large-scale flow in a general circulation model. *Beitrage zur Physik der Atmosphere*, **57**, 216–224.

Timmerman A., Oberhuber J., Bacher A., Esch M., Latif M. and Roeckner E. (1999) Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature*, **398**, 694–696.

Trenberth K.E. and Hoar T.J. (1996) The 1990–1995 El Niño event; longest on record. *Geophysical Research Letters*, **23**, 57–60.

Tribbia J.J. (1991) The rudimentary theory of atmospheric teleconnections associated with ENSO. In *Teleconnections Linking Worldwide Climate Anomalies*, Glantz M.H., Katz R.W. and Nicholls N. (Eds.), Cambridge University Press: pp. 285–307.

Waliser D.E., Lau K.M. and Kim J.-H. (1999) The influence of coupled sea surface temperatures on the Madden–Julian oscillation: a model perturbation experiment. *Journal of the Atmospheric Sciences*, **56**, 333–358.

Wallace J.M. (2000) North Atlantic oscillation/annular mode: two paradigms-one phenomenon. *Quarterly Journal of the Royal Meteorological Society*, **126**, 791–805.

Wallace J.M. and Gutzler D.S. (1981) Teleconnections in the geopotential height field during the Northern hemisphere winter. *Monthly Weather Review*, **109**, 784–812.

Wallace J.M. and Thompson D.W. (2002) The Pacific center of action of the northern hemisphere annular mode: real or artifact? *Journal of Climate*, **15**, 1987–1991.

Webster P.J. (1981) Mechanisms determining the atmospheric response to sea surface temperature anomalies. *Journal of the Atmospheric Sciences*, **38**, 554–571.

Webster P.J. and Yang S. (1992) Monsoon and ENSO: selectively interactive systems. *Quarterly Journal of the Royal Meteorological Society*, **118**, 877–926.

Werth D. and Avissar R. (2002) The local and global effects of Amazon deforestation. *Journal of Geophysical Research*, **107**(D20), 8087, doi:10.1029/2001JD000717.

Wettstein J.J. and Mearns L.O. (2002) The influence of the North Atlantic-Arctic oscillation on mean, variance, and extremes of temperature in the northeastern United States and Canada. *Journal of Climate*, **15**, 3586–3600.

Woolnough S.J., Slingo J.M. and Hoskins B.J. (2000) The relationship between convection and sea surface temperature on intraseasonal timescales. *Journal of Climate*, **13**, 2086–2104.

Wunsch C. (1999) The interpretation of short climate records, with comments on the North Atlantic and Southern oscillations. *Bulletin of the American Meteorological Society*, **80**, 245–255.

Xie S.-P. and Tanimoto Y. (1998) A pan-Atlantic decadal climate oscillation. *Geophysical Research Letters*, **25**, 2185–2188.

Yang S., Lau K.-M. and Kim K.-M. (2002) Variations of the east Asian jet stream and Asian-Pacific-American winter climate anomalies. *Journal of Climate*, **15**, 306–325.

Zhang C. and Gottschalk J. (2002) SST anomalies of ENSO and the Madden–Julian oscillation in the equatorial Pacific. *Journal of Climate*, **15**, 2429–2445.

Zhang H., Henderson-Sellers A. and McGuffie K. (1996) Impacts of tropical deforestation. Part II: the role of large scale dynamics. *Journal of Climate*, **10**, 2498–2521.

Zhao M., Pitman A.J. and Chase T.N. (2001) Influence of land-cover change on the atmospheric circulation. *Climate Dynamics*, **17**, 467–477.

Zhou S., Miller A.J., Wang J. and Angell J.K. (2001) Trends of NAO and AO and their association with stratospheric processes. *Geophysical Research Letters*, **28**, 4107–4110.

Zorita E. and Gonzalez-Rouco F. (2000) Disagreement between predictions of the future behavior of the Arctic oscillation as simulated in two different climate models: implications for global warming. *Geophysical Research Letters*, **27**, 1755–1758.

184: Global River Carbon Biogeochemistry

JEFFREY E RICHEY

School of Oceanography, University of Washington, Seattle, WA, US

The fluxes of dissolved and particulate carbon from land through fluvial systems to the oceans and to the atmosphere represent important pathways in the global carbon cycle. The processes controlling the distributions of solute species in river waters are established initially by weathering within the watershed, and physical transport via runoff. Superimposed on the underlying geochemical and physical processes is production and mineralization by terrestrial and aquatic biota. These factors play out differentially across the world's river basins, producing chemical signatures that vary from river to river. As a global aggregate, there would appear to be a net sink (between continental sedimentation and marine sedimentation and dissolution) of ~1 to 1.5 Pg year^{-1}. These sinks are partially compensated for by the outgassing. These processes are geographically very dispersed, with the continental sedimentation occurring in northern temperate regions, and much of the marine sedimentation and outgassing occurring in more tropical regions.

BIOGEOCHEMICAL DYNAMICS IN RIVER BASINS

A significant challenge for global biogeochemistry is to determine how the interaction of hydrological and biogeochemical cycles functions at the land surface, on regional to continental scales, producing the river flow and chemical load delivered to the oceans. Riverine transport represents a major link in the global cycles of bioactive elements, which modulates the biosphere over geological time (Meybeck, 1982). Dissolved and particulate organic matter (DOM and POM) in river systems serve as important heterotrophic substrates (Vannote *et al.*, 1980), provide an integrated continuous record of processes within drainage basins (Meybeck, 1982; Degens *et al.*, 1991), and constitute a major source of reduced carbon to the world ocean (Olson *et al.*, 1985; Schlesinger and Melack, 1981). The transfer of organic matter (OM) from the land to the oceans is the main pathway for the ultimate preservation of terrigenous production in modern environments, a key link in the global carbon cycle (Ittekkot and Haake, 1990; Hedges *et al.*, 1992). Terrestrial OM losses support significant heterotrophic activity within rivers, estuaries, and marine systems alike (Kaplan and Newbold, 1993; Mayer *et al.*, 1998), while natural and especially anthropogenic nutrient loading promote primary production. Where and when this river-borne OM is finally respired is of consequence to global carbon models (Stallard, 1998). Thus, understanding the processes that control the pathways from initial source to final mineralization of riverine organic matter is important on both regional and global scales.

At regional scales, river basins are natural integrators of surficial processes (Figure 1). Large rivers owe their flow and chemical loads to a much denser network of small rivers and streams bordered by areas of periodically inundated land, so that upland areas are dissected by corridors of wet soils and flowing water. Hence, understanding the hydrological and chemical patterns observed at the mouths of major rivers requires delineating the sequences of biogeochemical processes operating across multiple time and space scales. As embodied originally by the River Continuum Model, river properties should vary systematically downstream as processes affecting primarily the interactions of flowing water with the landscape give way to within river transport and processing (Vannote *et al.*, 1980; Minshall *et al.*, 1985). They should respond with differing magnitudes and lags to natural or man-made perturbations depending on the processes involved and the downstream transfer rates of their characteristic products. The central premise of a river basin model is that the constituents of river water provide a continuous, integrated record of upstream processes whose balances vary systematically

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

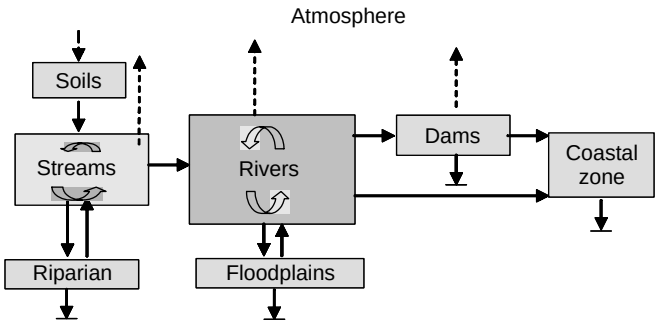


Figure 1 Schematic figure of the major reservoirs and pathways in fluvial systems. Inputs from land occur directly or pass through the riparian zone. Streams coalesce to form larger rivers that exchange with their floodplain. Rivers can pass directly to the coastal zone, or be retained behind dams. Dotted lines indicate exchange with the atmosphere, grounded areas indicate sinks, arrows within boxes indicate internal transformations (adapted from Richey, 2004)

depending upon changing interactions of flowing water with the landscape and the interplay of biological and physical processes (Karlsson *et al.*, 1988; Billen *et al.*, 1991). Hence, the chemical signatures of riverine materials can be used to identify different drainage basin source regions, reaches or stages and can be tied to landscape-related processes such as chemical weathering and nutrient retention by local vegetation (Meyer *et al.*, 1988). Because of the dynamic nature and abrupt moisture gradients of river corridors, the cumulative signal from a series of low-order streams may be manifest by higher-order rivers in a nonadditive manner.

The intent of this article is to develop a quantitative understanding through a heuristic and mass balance model of the sequence of processes from uplands to floodplains that produce the integrated hydrological, chemical, and biological signals at the mouths of the major tributary basins of large-scale river systems. The guiding questions to be addressed in this article include,

- *How does a large river system obtain and subsequently modify its biogeochemical composition?* This question can be separated into several parts.
- *How is the biogeochemical signature which persists through the river system imparted by the (aggregated) land surfaces?* That is, where do particular materials enter the river system, under the control of what process(es)? Observations of minimal changes in bioactive element forms and compositions within a study reach suggest that the mainstem of a major river is effectively transporting a complex compositional fingerprint that is imprinted somewhere upstream.
- *How is the land-derived signature modified through transit within the river system?* While the compositional fingerprint may be imprinted upstream, how do internal dynamics, such as production/respiration and sediment deposition/mobilization cycles modify that fingerprint, and in turn influence downstream export?
- *What are the time sequences of the sources, quantities, and chemical attributes of the river-borne fluxes of C, N, and P to the ocean? What are the physical and chemical controls operating on those fluxes at continental scales?* These questions in turn address a central problem in riverine ecology, which is to determine how the dynamics of flowing waters change as they increase in size from first-order springs and seeps to the world's great rivers.

RIVER BIOGEOCHEMISTRY: THE BASIC PARAMETERS AND PROCESSES ACROSS SCALES

A watershed, as the landscape through which all waters flow from their highest source before draining naturally to the sea, can be considered to be a fundamental organizing unit of the land surface. Rain or snow falls to the land surface, some of which is returned to the atmosphere (via evapotranspiration), some is stored in the soil (as soil moisture), and the balance drains into stream networks, mobilizing with it dissolved solutes and particulate materials from the landscape (Figure 2). As streams descend, tributaries and groundwater add to their volume, creating ever-larger rivers. Biological and chemical processes affect the materials in transport. As rivers leave the highlands, they slow down and start to meander and braid, and move between the river's main stream and its floodplain, modifying the flow regime and creating critical ecological (and biogeochemical) niches. The diversity of a river lies not only in the various types of land surfaces (or land uses) it flows through but also in the changing seasons and the differences between wet and dry years. Disruption of the linkages between the landscape and rivers and between rivers and

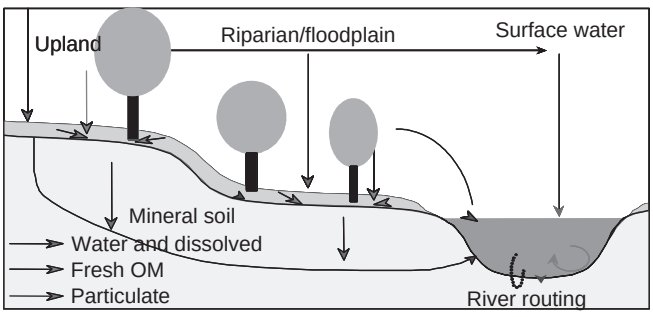


Figure 2 Grid-based view of land surface processes transferring water and its dissolved and particulate load to streams, where these constituents are subsequently routed downstream A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

their floodplains through human intervention fundamentally alters the nature of riverine ecosystems. The temporal scales of watersheds cover as much range as do the spatial scales. While the shape of the landscape itself (hills, valleys) will evolve (on geological timescales), these changes occur much more slowly than the seasonal and yearly evolution of the landcover and land use. These seasonal changes in turn occur much more slowly than a rainstorm, whose characteristic timescale is in the order of minutes to hours. The processes with which we are concerned, for example, the translation of rainfall into runoff across different types of landscapes, are themselves described differently at different space scales. The overall ensemble of these landscape features, which we can observe at multiple scales, can be thought of as the "physical template" upon which the more rapidly dynamic processes can operate.

Hence, the biogeochemistry of rivers is established as the interplay of terrestrial, biological, and geochemical weathering reactions that produce a suite of dissolved and particulate inorganic and organic compounds, via a series of pathways. The fundamental processes controlling the distributions of dissolved inorganic carbon (DIC) and the solute species in river waters are established initially by weathering within the watershed, and physical transport via runoff. Superimposed on the underlying geochemical and physical processes is production and mineralization by terrestrial and aquatic biota. These factors play out differentially across the world's river basins, producing chemical signatures that vary from river to river.

While matter transported by rivers ranges in size from molecules to trees, four broad size fractions are commonly used to characterize the bulk materials. Based operationally on filtration, the fractions include coarse particulate organic matter (CPOM, 63 microns to ~2 mm), fine particulate organic matter (FPOM, ~0.5 micron to 63 microns), dissolved organic matter (DOM, <0.5 microns) and dissolved nutrients, and dissolved gases (CO_2, CH_4, N_2O). As summarized (below) by Mayorga and Aufdenkampe (2002), these size classes exhibit very distinct transport dynamics, degradation patterns, and compositional characteristics (Figure 3).

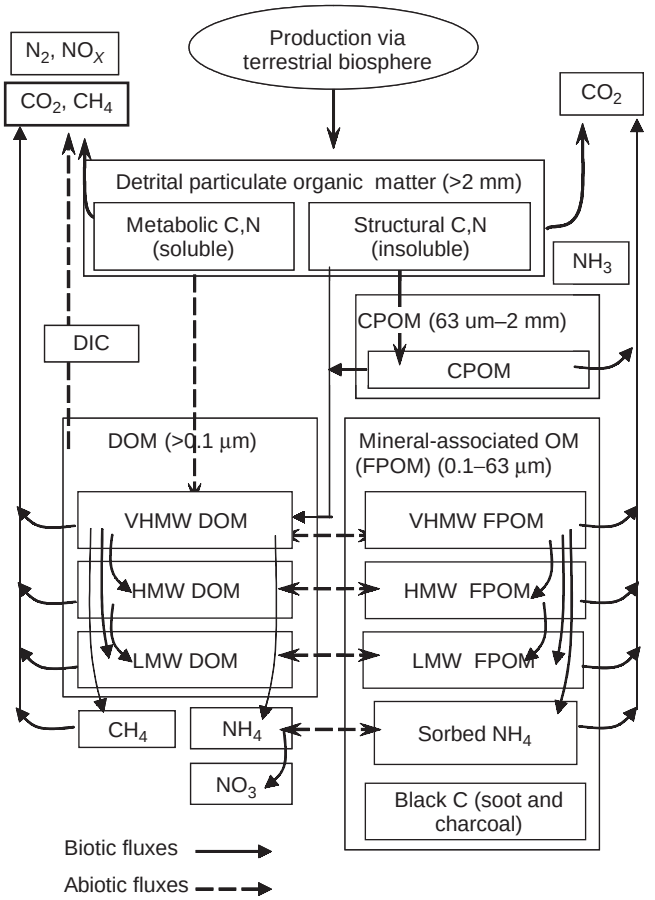


Figure 3 Fractions of the dissolved inorganic and organic components of the primary biogeochemical constituents of C and N in river water. Dissolved and particulate fractions can be sorted by molecular weight MW (V = very, H high, L low), VHMW DOC (>30 kDa), HMW DOC (1–30 kDa), LMW DOC (<1 kDa), VHMW DON (>30 kDa), LMW DON (200–1000 Da), and VLMW DON (<200 Da). The reason for such fractions is to break down bulk materials in a manner that can be related to system function

Dissolved Inorganic Species

Watersheds contain soils and rock that consist of a wide variety of different minerals. Weathering of the soil and rock minerals produces runoff containing the common ions (and many other ions in lesser concentration). The major inorganic ion composition of world rivers is controlled largely by geology and the weathering regime with minor inputs from precipitation (Stallard and Edmond, 1983, 1987). In general, weathering reactions can be characterized as a weak acid (carbonic acid) slowly dissolving basic minerals. Weathering reactions, such as sodium feldspar or calcite weathering, result in runoff containing dissolved ions, and commonly, undissolved particles of less soluble minerals such as clays. Many of the world's great rivers discharge water that can be characterized as rock-dominated (Gibbs, 1970). Water more or less in equilibrium with the materials in the drainage basin is characterized by higher concentrations of ions and an increased significance of Ca, Mg, and bicarbonate ions. Rivers that receive rainwater and snowmelt with little influence of rock weathering contain water with low concentrations of ions, reflecting rainwater that is (approximately) a dilute solution of carbonic acid with an admixture of a small amount of sea salt. Evaporation and fractional precipitation tend to dominate in more arid regions. The overall concentration of ions increases with evaporation until selected minerals begin to precipitate because their solubility product is exceeded.

The overall weathering regime in turn establishes the carbonate system, as the sum of the species of DIC, $DIC = [H_2CO_3^*] + [HCO_3^-] + [CO_3^{-2}]$. Natural waters are buffered with respect to pH mostly because of the content of inorganic carbon species. In turn, pH is an important controlling variable for many important geochemical reactions (e.g. solubility of carbonates). Many important biochemical reactions, such as photosynthesis and respiration, interact with pH and the carbonate system. Alkalinity, as the approximate imbalance of cations and anions, is a measure of the ability of a water system to resist changes in pH when acid is added to water. A stream that has a high alkalinity is well buffered so that large inputs of acid (from acid rain for instance) can be made with little effect on the stream pH. A stream that has a low alkalinity is poorly buffered and may undergo large, sudden drops in pH in response to acid inputs. Reduction/oxidation (redox) processes, which typically occur in oxygen-depleted zones near streams, will alter alkalinity. Examples of importance to rivers include oxidation of ammonia to nitrate (nitrification), oxidation of sulfide or reduction of sulfate and oxidation or reduction of iron. Precipitation or dissolution of minerals involving participating species will alter alkalinity. For example, the precipitation or dissolution of calcium carbonate can alter alkalinity. Other participating species, such as organic acids, may influence alkalinity, especially in low alkalinity environments. Dissolved inorganic nutrients are closely associated with the cycling of organic matter. Inorganic nitrogen compounds – NO_3^- and NH_4^+ – cycle rapidly via remineralization of organic matter and other microbial processes such as nitrification and denitrification. Of the bioactive compounds, PO_4^{3-} shows the least systematic variability, generally $0.4 - 2.0\,\mu M$ in turbid rivers as a result of buffering with larger mineral-bound reservoirs in the suspended sediments. On the other hand, NO_3^- and PO_4^{3-} are particularly subject to anthropogenic influences, via fertilizers and domestic and industrial wastes. Dynamics of dissolved gasses, such as O_2, CO_2, CH_4 are also largely controlled by the respiration of organic matter because river, floodplain, and lake waters are dominantly heterotrophic (Cole *et al.*, 1994; Cole and Caraco, 2001). Thus, the dissolved inorganic constituents of river waters are constantly evolving as a result of interactions with nondissolved phases within the river corridor.

Dissolved Organic Matter

Perhaps the most important characteristic of dissolved material is that it has the potential to be directly bioavailable. Microbial organisms, plant roots, and many animal tissues transport dissolved molecules across cellular membranes, both passively and actively. Likewise, contaminants exhibit their highest toxicity when in the dissolved phase. The dissolved fraction is particularly characterized by diversity and contrasts. Organics and inorganics exist in both truly dissolved and colloidal phases. DOM exists as a mixture of simple molecules, complex biomacromolecules, their partial degradation products and molecular assemblages or gels. This mixture contains the most labile material carried by the river (e.g. NH_4^+, free amino acids, etc.) and also relatively nonlabile weathering end-products (e.g. inorganic ions that determine alkalinity, low molecular weight (LMW), DOM, etc.). While the dissolved fraction is generally considered to be $<0.45\,\mu m$, to better understand the composition and dynamics of DOM, ultrafiltration techniques have employed membranes with pores as small as 1 nm to separate and concentrate DOM into various size fractions (Hedges *et al.*, 1994; Küchler *et al.*, 1994; Amon and Benner, 1996a,b; Mounier *et al.*, 1999; Patel *et al.*, 1999). Generally, ultrafiltered DOM (UDOM) refers to organic material with molecular weights $>1000\,g\,mol^{-1}$ or daltons (high molecular weight or HMW), including very high molecular weight subsets variously named very high molecular weight (VHMW) DOM or colloidal organic carbon (COC) by different research groups (1000 daltons is approximately equivalent to a molecule of 1-nm diameter).

Cumulative evidence suggests that DOM is produced largely from the degradation and/or leaching of leaf detritus similar to that in CPOM (Devol and Hedges, 2001). Once in solution, biomacromolecules such as proteins and carbohydrates are easily hydrolyzed (at least partially) by exoenzymes for subsequent microbial uptake. As a result, degradation tends to decrease both the size and bioavailability of HMW DOM to form the low molecular weight (LMW, 200–1000 daltons) fraction (Amon and Benner, 1996a). However, as all particulate and dissolved organic carbon fractions degrade, microbial activity and photochemistry can generate a pool of the smallest molecules (<200 daltons) – free amino acids, free sugars, and organic acids such as acetate and citrate (Amon and Benner, 1996b; Moran and Zepp, 1997). Despite the likelihood that this very low molecular weight (VLMW) DOM represents an exceedingly small proportion of total DOM in rivers, these compounds are generally extremely bioavailable and could drive significant biological fluxes. The relative proportions of VHMW, HMW, LMW, and VLMW DOM fractions would be expected to evolve down river as a result of their different degradation rates. Coagulation and disassociation of DOM in and out of colloidal gel phases or mineral surfaces complicates these size dynamics significantly however, as many of these processes respond to changes in pH and to ratios of polyvalent to monovalent ions in solution (such as Ca^{2+}/Na^+) (Chin *et al.*, 1998; Kaiser, 1998).

Particulate Organic Matter

Coarse suspended river sediments are a heterogeneous mixture of sand-sized mineral grains (coarse suspended sediments, (CSS)) and discrete plant fragments. Because CSS settles quickly to the streambed, suspended concentrations

are strongly dependent on streamflow velocities and increase substantially with depth in the river. Thus, CSS transport is highly episodic or seasonal, with most occurring during flood events. In large turbid rivers, the CPOM is a small but important portion of the CSS fraction, comprising only 0.6–3.3% by mass. CPOM is less dense than mineral grains of the same size, hence explaining its higher contribution to CSS in river channels under lowflow. Between 10 and 20% of CPOM can be identified biochemically as amino acids, carbohydrates, and lignins relative to 25–60% within biomass sources (Hedges *et al.*, 1986a; Hedges *et al.*, 1994; Hedges *et al.*, 2000). Biochemical source indicators, such as the carbon to nitrogen ratio and the ratio of cinamyl to vanilyl lignin phenols, all show that Amazon basin CPOM, for example, is primarily derived from tree leaves (Devol and Hedges, 2001). Biochemical indicators of degradation, such as the contribution of fucose and rhamnose sugars to total carbohydrates and acid to aldehyde ratios in lignin phenols, all support evidence from microscopic studies and major biochemical composition that CPOM is sparingly degraded and rather fresh. Radiocarbon analysis of CPOM and low density soil particulate organic matter confirms their recent origin (Hedges *et al.*, 1986b; Trumbore *et al.*, 1995). It is clear that CPOM is actively degrading and leaching, supplying microbes with substrate and releasing dissolved organic and inorganic compounds into the river. These rates have not been directly measured, but are likely to be quite high. Stream and river budgets suggest that CPOM continuously enters the river mostly from bank vegetation and detritus falling directly into the water (McClain *et al.*, 1997).

FPOM is essentially the product of diverse dissolved organic and inorganic compounds binding to fine suspended sediments (FSS, as clays and silt materials between 0.45 and 63 μm). The sorption of natural DOM to minerals is the primary pathway in which FPOM is formed (Mayer, 1994; Hedges and Keil, 1995), organic matter contributes only to a small fraction of FSS, generally. As FPOM is, typically, only a small weight-percent of FSS (typically 0.5–2.0% by weight), in the turbid rivers that transport most of the FPOM, the dynamics of FPOM have to start with an understanding of the sources and transport of FSS. FSS is mobilized initially into stream channels by erosion events (especially extreme events, resulting in hillslope failure). As a product of rock weathering, mineral diversity within FSS is very large and depends on the geology (source minerals) and climate (weathering rates) of the watershed. Clays and oxides are aggregated with the larger minerals in important quantities. The clay-size fraction (0.45–5 μm) is composed mostly of phyloalumino-silicate clay minerals, which are the weathering products of primary silicates. As a result, mineralogical compositions within large drainage basins are constantly evolving downstream, with inputs from tributary watersheds and from weathering during temporary storage on the floodplain. While maintained in suspension by the slightest turbulence, FSS is not transported conservatively downstream. Evidence suggests that within the Amazon, for example, a typical FSS particle passes through floodplain deposits several times between the Peruvian border and the Atlantic (Martinelli *et al.*, 1993; Dunne *et al.*, 1998). Given the patterns of channel migration, each cycle of floodplain deposition and resuspension requires a few thousand years (Mertes *et al.*, 1996). The presence of lakes and especially reservoirs created by dams trap sediment, thereby creating a very large anthropogenic impact on the downstream movement of suspended sediments.

From a biogeochemical perspective, the critical issue is, how does the FSS mineral particle acquire its organic matter composition? The major characteristics of FPOM, as tightly associated with the mineral phase, most likely form within soils prior to erosion into rivers. While mineralogically diverse, FSS is generally high in surface area and high in cation and anion exchange capacity. As such, significant quantities of certain inorganic ions (e.g. NH_4^+, PO_4^{3-}, and most metals) and organic compounds can potentially be adsorbed to the mineral surfaces of FSS. Thus, FSS actively interacts with the dissolved fraction, often acting as a buffer or reservoir for dissolved compounds. Sorbed organic matter affects mineral surface properties significantly by increasing cation exchange capacity and by offering hydrophobic phases into which organic contaminants (e.g. pesticides, petroleum products, etc.) and heavy metals readily sorb (Benedetti *et al.*, 1996). Biochemically, a smaller fraction of FPOM is identifiable as carbohydrates or lignins when compared to CPOM, but often a larger fraction is identifiable as amino acids. It appears that FPOM comes largely from leaf material as does CPOM, but often falls slightly outside of the biochemical range that is possible by mixing biomass alone, as might be expected of diagenetically altered material (Devol and Hedges, 2001). Specific degradation parameters show this clearly; however, riverine FPOM is still relatively fresh compared to riverine DOM or deep-sea sediments (Hedges *et al.*, 1994). Overall, 90% of the FPOM cannot be physically separated from the mineral material (Keil *et al.*, 1997; Mayer *et al.*, 1998). This appears to be the result of physical protection from microbial attack that is offered by the intimate association of FPOM with mineral surfaces (Keil *et al.*, 1994; Baldock and Skjemstad, 2000; Kaiser and Guggenberger, 2000). Whether an organic molecule sorbs or remains dissolved determines, in large part, its transport potential and susceptibility to degradation.

SCALE-DEPENDENT MODELING AND PROCESS DYNAMICS

Clearly, the biogeochemical signature of large-scale river systems is the composite of processes occurring across

multiple time and space scales. A parcel of river water passing a particular location in a network with a particular chemical signature represents the cumulative upstream history of that parcel. That chemical signal is the product of what occurred in the stream segment above that point (internal transformations, such as respiration or primary production, and lateral exchanges, such as riparian interactions or floodplains), plus the addition of mass from upstream segments. That is, the chemical signature at some point in a river is the product of both terrestrial and in-stream processes. To capture these dynamics, it is useful to organize our understanding into first heuristic, and ultimately, computer models. Ideally, we want to develop "first-principle" models, rather than the regression approaches commonly used, relating water flow across the landscape to river chemistry.

First, we must be able to describe the movement of water through the landscape, as a function of the specific physiochemical attributes of the landscape (as discussed qualitatively in the previous section). It is necessary to be able to explicitly represent infiltration, soil moisture, and runoff, and then the downstream routing of a water parcel. The emergence of geospatially explicit landscape and hydrology models is making it feasible to start developing coupled landscape/fluvial biogeochemistry models. At a macro scale, hydrology models such as the Variable Infiltration Capacity (VIC; Liang *et al.*, 1994; Liang *et al.*, 1996; Lohmann *et al.*, 1998) and others described elsewhere in this Encyclopedia describe a vegetation cover type (through parameters including leaf area index, stomatal resistance, root mass distribution with soil depth, and others) and soil attributes in multiple layers for each grid cell. The model solves for a water balance and energy balance simultaneously, producing fields of soil moisture and runoff, with the runoff routed down a river network. Models such as the Distributed Hydrology Soil Vegetation Model (DHSVM; Wigmosta *et al.*, 1994), can recognize the spatial heterogeneity of smaller watersheds (typically less than about 10 000 km^2), and compute cell-to-cell fluxes.

Such landscape/hydrology models can now be used to compute chemical properties, as the generation and routing of a chemical substance within the stream network. That is, a chemical substance can enter a river segment laterally or it can enter from an upstream reach (that previously had a lateral input). The combined inputs are then subject to advection (downstream flow) and chemical or physical reactions (respiration, sedimentation) during transport. Schematically, this is represented in Figure 4. Each discrete reach of the network can be represented by connected straight segments, where each segment lies within one of the grid cells used to run the hydrology model (e.g. VIC or DHSVM). The model is built in an Eulerian framework, where each model element is a stream segment.

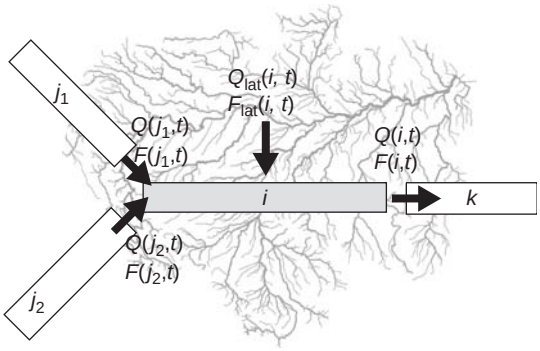


Figure 4 An overall river network (e.g. the Amazon), with the basic source/routing scheme superimposed. A representative segment i has two contributing segments, j_1 and j_2, and one outlet segment, k (Note that a segment can have anywhere from 0 to 7 input segments – more than 4 being rare – but it always has a single output segment). At time step t, input flows to i are $Q(j_1, t)$, $Q(j_2, t)$, and $Q_{lat}(i, t)$; and input mass fluxes are $F(j_1, t)$, $F(j_2, t)$, and $F_{lat}(i, t)$.$Q_{lat}(i, t)$ is the runoff rate entering segment i at time step t from the corresponding grid cell, and has previously been computed by a hydrology model. $F_{lat}(i, t)$ is the mass flux of species C entering segment i at time step t from the corresponding grid cell. At time step t, the output flow from i is $Q(i, t)$ and the output mass flux from i is $F(i, t)$. The travel time in i is $\Delta t(i)$. Input streamflows and mass fluxes at time t are used to compute the output streamflow and mass flux at time $t + \Delta t(i)$

Hence, there are as many model elements as there are stream segments and grid cells. Model elements are treated as control volumes with state variables that are updated at each model time step. There are only two state variables, $Q(i, t)$ (stream flow rate exiting element i at time step t, $m^3\ s^{-1}$) and $F(i, t)$ (mass flow rate of species C exiting element i at time step t, $\mu g\, s^{-1}$.

The first part of the problem is to compute the lateral inflow term. This is problematic, and depends very much on where in the drainage net the inflow is. The solute export module of a basin biogeochemistry model needs to estimate the amount and concentration of basin solutes (e.g. dissolved carbon and nitrogen species) exported to a stream via subsurface flow and in-stream concentrations (Figure 5). The control volumes in the basin are the soil solutions in each soil root zone and in the saturated lateral layer. The form of the specific equations can be derived from the relationships illustrated in Figure 3.

For the sake of illustration, let us now consider a hypothetical species C with concentration $[C]$ assumed to decay exponentially in time as it moves downstream, with decay constant k_1:

$$\frac{d[C]}{dt} = -k_1[C] \qquad (1)$$

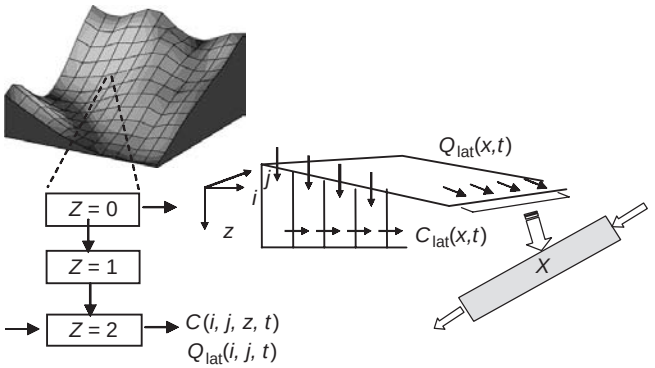


Figure 5 Schematic of lateral inflow (see Figure 4 for definitions) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Integrating equation (1) from time $t = 0$ to $t = T$

$$[C](T) = [C](0)\mathrm{e}^{-k_1 T} \tag{2}$$

The concentration of species C (given in $\mu\mathrm{g\,m^{-3}}$) in the streamflow exiting a model element i at time step t is obtained from these state variables, as.

$$[C](i,t) = \frac{F(i,t)}{Q(i,t)} \tag{3}$$

The equations used to update the state variables are:

$$F_{\mathrm{lat}}(i,t) = k_2 Q_{\mathrm{lat}}(i,t)^{k3} \tag{4}$$

$$\Delta t(i) = \frac{length(i)}{velocity(i)} \tag{5}$$

$$Q(i, t + \Delta t(i)) = \Sigma_j Q(j,t) + Q_{\mathrm{lat}}(i,t) \tag{6}$$

$$F(i, t + \Delta t(i)) = [\Sigma_j F(j,t) + F_{\mathrm{lat}}(i,t)]\,\mathrm{d}t\mathrm{e}^{-k1\Delta t(i)} \tag{7}$$

where:

$\Delta t(i)$ is the travel time in segment i,

length(i) is the length of segment i,

velocity(i) is the streamflow velocity in segment i (In this first model version, *velocity(i)* is assumed to be independent of streamflow, making $\Delta t(i)$ constant in time for any segment i – an assumption that can later be relaxed),

$\mathrm{d}t$ is the model time step,

k_1 is the parameter in equations (1) and (2) (with units day^{-1}),

$Q_{\mathrm{lat}}(i,t)$ is the runoff rate entering stream segment i, a value previously computed by VIC for the grid cell corresponding to i,

$Flat(i,t)$ is the rate at which mass of species C is entering stream segment i at time step t from the corresponding grid cell (here assumed to be a function of $Q_{\mathrm{lat}}(i,t)$ according to (5)),

k_2 (units $\mu\mathrm{g\,m^{-3}}$) and k_3 (nondimensional) are model parameters, and $Q(*,*)$ and $F(*,*)$ are the state variables.

Equation (7) is directly obtained from equation (2), with the simplifying assumption that the travel distance (and time) for the lateral inputs is the same as for the inputs from tributaries.

Equations (5), (6), and (7) are computed for each model time step.

From equations (3) and (4), we obtain:

$$[C]_{\mathrm{lat}}(i,t) = k_2 Q_{\mathrm{lat}}(i,t)^{k3-1} \tag{8}$$

From equation (8) we see that the concentration of species C in the runoff entering segment i from its corresponding grid cell will increase with the runoff rate for $k_3 > 1$, and will decrease with runoff rate for $k_3 < 1$. For $k_3 = 1$, this concentration will be constant (and equal to k_2), regardless of runoff rate.

More specifically, the downstream advection and reaction for the different species can be represented as follows, where the exact form of the equation depends on the degree to which a specific parameter is dependent on its own concentration:

For conservative species (alkalinity, Ca, Si, O^{18}-H_2O) : $[C](i,t) = C_o$

Add a constant internal, nonconcentration dependent source S (FSS as function of erosion/deposition, production of nitrate): $[C](i,t) = C_o + S\ length(i)/\ velocity(i)$

Source + concentration dependent sink (e.g. PO_4 mineralization and adsorption) $[C](i,t) = s/k + (C_o\text{-}s/k)\ e^{(-k\ length(i)/velocity(i))}$

where S represents the aggregate of internal processes (erosion/deposition, mineralization, gas exchange, adsorption)

Gas Exchange: $[C](i,t) = (C_{\mathrm{eq}} \pm R/(D/\mu z)) - (C_{\mathrm{eq}}\text{-}]\ C_o \pm R/(D/\mu z))\mathrm{e}^{(-(D/\mu z)\ length(i)/velocity(i))}$

where

k = gain/loss from mineralization of dissolved organics (respiration/production)

z = thickness of hypothetical stagnant boundary layer.

C_{eq} = gas equilibrium with the atmosphere

D = molecular diffusion coefficient

While the above framework is still very much under development, it does represent the direction that basin biogeochemical modeling is developing.

RIVER BASIN MASS BALANCES OF CARBON

We will now evaluate the biogeochemical fluxes through global rivers, in terms of the box model of Figure 1. While the primary emphasis will be on carbon, the current export of nitrogen and phosphorus will also be examined. Evaluation of this model can only be approximate; the dynamics are complex, and multiple time constants are involved. Data are scarce, particularly in many of the most anthropogenically impacted systems. The distribution of the constituent processes varies dramatically across the face of the globe (with some of the most important regions being the least measured). This discussion is derived primarily from Richey (2004).

Mobilization from Land to Water and Riparian Zones

The modern terrestrial sediment cycle is not in equilibrium (Stallard, 1998). Meade *et al.* (1990) estimated that agricultural land use typically accelerates erosion 10- to 100-fold, via both fluvial and Aeolian processes. Multiple other reports in the literature support this conclusion. With the maturation of farmlands worldwide, and with the development of better soil conservation practices, it is probable that the human-induced erosion is less than it was several decades ago. Overall, however, there has been a significant anthropogenic increase in the mobilization of sediments (and associated POC) through fluvial processes. The global estimates of the quantities, however, vary dramatically. Stallard (1998) poses a range of scenarios, from 24 to 64 Pg year^{-1} of bulk sediments (from 0.4 to 1.2 Pg year^{-1} of POC). Smith *et al.* (2001) estimate that as much as 200 Pg year^{-1} of sediment is moving, resulting in about 1.4 Pg year^{-1} of POC (using a lower %C than Stallard, 1998).

Where does this material go? Does it all go downstream via big rivers, ultimately to the ocean, or is it stored inland? Stallard (1998) argues that between 0 and 40 Pg year^{-1} of sediments (0 to 0.8 Pg C year^{-1} POC) is stored as colluvium and alluvium, and never makes it downstream. Smith *et al.* (2001), using a different approach, estimate that about 1 Pg C year^{-1} of POC is stored this way. If this movement is merely transferring POC from one reservoir to another, with the same residence times, there is no net change in the C cycle. Then the issue is, to what degree can the (remaining) soils sequester carbon by sorption to the newly exposed mineral soils? Both Stallard (1998) and Smith *et al.* (2001) argue that carbon is removed from the upper portion of the soil horizon, where turnover times are relatively rapid (decades, or shorter), into either of the two classes of environments with longer turnover times; wetlands and smaller, deeper depositional zones, coupled with new carbon accumulation at either erosional or depositional sites. Both assume that oxidation of organic C in transit is minimal, and both use quite conservative values for total suspended sediment (TSS) export. If true, this sequence of processes would result in a significant C sink, on the order of 1 Pg C year^{-1}.

Within-river Transport and Reaction Processes

Within-river transport processes carry these eroded materials downstream through the river network. Transport is not passive; significant transformations occur along the way. Rivers exchange with their floodplains (depending on how canalized and diked a river is). A significant process within flowing water significantly affects organic matter (OM) – the mineralization to pCO_2. Most river and floodplain environments maintain pCO_2 levels that are supersaturated with respect to the atmosphere. High partial pressures of CO_2 translate to large gas evasion fluxes from water to atmosphere. Early measurements in the Amazon suggested that global CO_2 efflux (fluvial export plus respiration) from the world's rivers could be on the order of 1.0 Pg C year^{-1}. Recent measurements of temperate rivers lead to estimates of global river-to-atmosphere (outgassing) fluxes of $\sim$0.3 Pg C year^{-1}, which is nearly equivalent to riverine total organic carbon (TOC) or DIC export (Cole and Caraco, 2001). Richey *et al.* (2002) computed that outgassing from the Amazon alone was about 0.5 Pg C year^{-1}. Assuming that the fluxes computed for the Amazon are representative of fluvial environments of lowland humid tropical forests in general, surface water CO_2 evasion in the tropics would be on the order of roughly 0.9 Pg C year^{-1} (three times larger than previous estimates of global evasion). Factoring in the recent Amazon results, a global flux of at least 1 Pg C year^{-1} directly from river systems to the atmosphere is likely.

What is the source of the organic matter being respired? Is it labile contemporary organic matter, recently fixed in the water by plankton or nearshore vegetation, or is it some fraction of the allochthonous (terrestrial) matter in transport? The prevailing wisdom is that river-borne organic matter is already very refractory and not subject to oxidation (after centuries on land). The "age" of riverine organic matter yields some important insights.

Thus, preaging and degradation may alter significantly the structure, distribution, and quantity of terrestrial organic matter before its delivery to the oceans. As noted by Ludwig (2001), the OM that runs from rivers into the sea is not necessarily identical to the OM upstream in river catchments. Cole and Caraco (2001) observe that the apparent high rate of decomposition of terrestrial organic matter in rivers may resolve the enigma of why OM that leaves the land does not accumulate in the ocean (*sensu* Hedges *et al.*, 1997). Overall, this sequence of processes suggests that the OM that is being respired is translocated in space and time from its points of origin, such that, over long times and large spatial scales, the modern aquatic environment may be connected with terrestrial conditions of another time.

Input to Reservoirs

Just because dissolved and particulate materials enter a river it does not mean that they reach the ocean; modern reservoirs have had a tremendous impact on the hydrologic cycle. Starting about 50 years ago, large dams were seen as a solution to water resource issues, including flood control, hydroelectric power generation, and irrigation. Now there are more than 40 000 large dams worldwide (World Commission on Dams, 2000). This has resulted in a substantial distortion of freshwater runoff from the continents, raising the "age" of discharge through channels from a mean between 16 and 26 days to nearly 60 days (Vörösmarty *et al.*, 1997). While erosion has clearly increased the mobilization of sediment off the land, the proliferation of dams has acted to retain those sediments. Vörösmarty *et al.* (2003) estimates that the aggregate impact of all registered impoundments is on the order of 4 to 5 Pg year^{-1} of suspended sediments (of the 15–20 Pg year^{-1} total that he references). Stallard (1998) extrapolates from a more detailed analysis of the coterminous US to an estimate of about 10 Pg year^{-1} worldwide (versus 13 Pg year^{-1} efflux to the oceans), for a storage of about 0.2 Pg C year^{-1} (which he includes as part of his overall calculation of continental sedimentation.

Export to the Coastal Zone

The conventional wisdom is that the flux of POC and DOC are each about 0.2 Pg C year^{-1}, and DIC is 0.4 Pg C year^{-1} (e.g. Schlesinger and Melack, 1981; Degens, 1982, Meybeck, 1982, 1994; Ittekkot, 1988; Ittekkot and Laane, 1991; Ludwig *et al.*, 1996; Ver *et al.*, 1999). That these analyses converge is not terribly surprising. They are all based on much of the same (very sparse) field data, and use variations of the same statistically based interpolation schemes. Let us evaluate these numbers.

Because direct measurements are few, POC flux estimations are typically a product of the flux of total suspended sediments (TSS) and the (estimated) weight-percent organic C (w%C) associated with the sediment (because the bulk of POC is organic C sorbed to mineral grains). The first problem is an adequate resolution of the TSS flux. Data on TSS are frequently poor and of unknown quality. Many reported data are surface samples, and the depth integrations necessary to accurately characterize sediment flux are on the order of 2–3x higher. Additionally, much sediment moves during episodic storm events, when measurements are almost never made.

Estimates of TSS transport to the oceans have ranged from 9 Pg year^{-1} to more than 58 Pg year^{-1}, with more recent studies converging around 15 – 20 Pg year^{-1}. These estimates are generally based on extrapolations of existing data, which are weighted to the large rivers of passive margins and temperate regions. Milliman and Syvitski (1992) called attention to the much higher yield rates from steep mountainous environments (without directly computing a global total). More recently, Milliman *et al.* (1999) estimated that the total sediment flux from the East Indies alone (the islands of Sumatra, Borneo, New Guinea, Java, Sulawesi, and Timor), representing about 2% of the global land mass, is about 4 Pg year^{-1}, or 20–25% of the current global values. This type of environment (steep relief, draining directly to the oceans) is found elsewhere in the world, so the results are not likely to be unique. New data from Taiwan support these high levels, with isotopic analyses of the C showing that a significant part of the flux is human-driven (Kao and Liu, 2002).

Then, to obtain POC flux estimates, these values (and their uncertainties) must be multiplied by w%C. Meybeck (1991) divided particulate carbon into inorganic (PIC) and organic (POC) phases, and assumed that high sediment rivers have very low carbon fractions (0.5 w%C), representative of shale; he essentially does not consider the latter to be "atmospherically derived" and hence discounts it from estimates of fluxes to the ocean. More recent values for w%C tend to be in the 1–2% range, and higher for organic-rich systems (Richey *et al.*, 1990; Stallard, 1998; Gao *et al.*, 2002).

To account for this range, POC flux can be computed as an ensemble based on different combinations of w%C and TSS fluxes, resulting in a range of 0.3 – 0.8 Pg C year^{-1}, with a "more likely" value of about 0.5 Pg C year^{-1} (depending on assumptions used). Therefore, it is possible that the common estimate of 0.2 Pg C year^{-1} is low, and that the overall value lies in the range of 0.2–0.5 Pg C year^{-1}. The common estimate for PIC of 0.2 Pg C year^{-1} (Meybeck, 1991; Ver *et al.*, 1999) may also be underestimated if sediment fluxes are higher.

The value of 0.2 Pg C year^{-1} for DOC export may also be low. DOC is also subject to sparse (and questionable) measurements, without the availability of a proxy like TSS for POC. Aitkenhead and McDowell (2000) developed a model of riverine DOC flux as a function of soil C:N. Using this model, they computed a global flux of 0.4 Pg C year^{-1}, or twice the common estimate. That is, the total organic C output from fluvial systems may well be approximately double the original estimates, in the ~0.8 Pg C year^{-1} range.

Anthropogenic Transient

To what degree have these fluxes been influenced or impacted by human activities, that is, how much of this carbon is an anthropogenic transient? The consensus is that human-induced erosion has dramatically accelerated the movement of sediments (and POC). While some of this material "hangs up" on land (sedimentation, reservoirs), some of it likely escapes to the sea (some of the regions with highest sediment yields have very few dams). There

is evidence that anthropogenic processes have an affect on DIC. Raymond and Cole (2003) report an increase in the alkalinity of the Mississippi, that implies an increase in the consumption of atmospheric CO_2 through weathering. However, Jones *et al.* (2003) reported a systematic decrease in pCO_2 in rivers across the United States, which they attributed to large-scale declines in terrestrial CO_2 production and import into aquatic ecosystems, and not to terrestrial weathering or in-stream processes.

In the case of DOC, there is simply not enough information available to make conclusions. Clair *et al.* (1999) suggested that DOC export from basins in Canada might increase by 14% with a doubling in atmospheric CO_2. An additional factor rarely addressed in rivers is direct loading from urban and industrial sources (Ver *et al.*, 1999; Abril *et al.*, 2002). In evaluating the consequences of continental sedimentation and the potentially higher fluxes of POC, a net transient exported from land of roughly $1\,\mathrm{Pg\,C\,year^{-1}}$ is possible, with perhaps half of that going to the sea, and the other half divided equally between outgassing and sedimentation.

Overall Fluvial System – Atmosphere Exchange

Considerable uncertainty remains in the assessment of the carbon cycle of fluvial systems, including the actual magnitude of fluxes, how to include processes not previously considered, and delineation of anthropogenic and natural processes; all with an explicit recognition of geography. Estimates of the bulk transfer of atmospheric C through the land to fluvial systems (assuming a steady state summation of downstream processes, in a non–steady state environment) ranges from about $0.6\,\mathrm{Pg\,C\,year^{-1}}$ to $2.6\,\mathrm{Pg\,C\,year^{-1}}$. Continental sedimentation results in a significant sink, but that sink is reduced with CO_2 outgassing (because of the way the sedimentation was computed). The inclusion of continental sedimentation, and then the larger export of OM to the sea (about twice conventional assumptions, under +POC, DOC), yields net sinks of atmospheric CO_2 of up to $1.6\,\mathrm{Pg\,C\,year^{-1}}$. However, if outgassing is included, then the fluvial net sink is reduced to $0.2\,\mathrm{Pg\,C\,year^{-1}}$. While partitioning the total fluvial fluxes into natural conditions and anthropogenic transients is problematic at best, there is substantial evidence that there has been a dramatic increase in the mobilization of sediments. While much of this material is captured in reservoirs, it is reasonable to expect that a considerable amount escapes to the sea (especially in nondeltaic regions with steep slopes and few dams).

Summarizing all the components of the riverine carbon cycle, several images emerge. As a global steady state aggregate, there appears to be a sink (between continental sedimentation and marine sedimentation and dissolution) on the order of $1-1.5\,\mathrm{Pg\,C\,year^{-1}}$, with a significant anthropogenically enhanced component. A return flux to the atmosphere, on the order of $1\,\mathrm{Pg\,C\,year^{-1}}$, reduces the net sink to about 0.2 or $0.3\,\mathrm{Pg\,C\,year^{-1}}$. There are significant regional implications in this analysis. With its preponderance of land mass, extensive reservoirs, and agriculture, the bulk of continental sedimentation (and its implications for C sink), is focused in the Northern Hemisphere, between $30^\circ - 50^\circ\mathrm{N}$. The C sequestration in paddy lands would be closer to the equator. The greatest amount of sediment flux to the ocean (and the greatest uncertainty) is in South and Southeast Asia and Oceania. Outgassing is a function of both pCO_2 concentrations (driven by *in situ* oxidation) and surface area of water. It is likely the most significant in the humid tropics, particularly during the peak of the wet seasons. The highly canalized temperate areas have less area available. The northern latitudes, particularly with warming, are liable to have significant fluxes.

Dissolved N and P

While evaluation of the overall anthropogenic impact on river basin carbon is complex, there is little ambiguity about impacts on dissolved N and P. To establish a baseline for evaluation of the N cycle, Lewis *et al.* (1999) evaluated yields of total fixed nitrogen and nitrogen fractions for watersheds across the world, over a broad range of watershed areas, elevations, and vegetation types, in which anthropogenic disturbance was small. They found that yields were substantially lower than previously estimated for background conditions, and that yields can be predicted on the basis of general environmental variables such as drainage area or the amount of runoff. However, the production of food and energy has markedly increased the amount of newly fixed nitrogen entering terrestrial and aquatic ecosystems during the last century. As of 1990, the amount of newly fixed N entering terrestrial systems annually had about doubled (Galloway *et al.*, 1995; Galloway and Cowling, 2002). The fraction of this N that reaches rivers and ultimately the coastal oceans has led to the enrichment of coastal ecosystems (e.g. increased phytoplankton production, increased turbidity with subsequent loss of submerged aquatic vegetation, oxygen deficiency, and decrease in biodiversity, and so on (Nixon, 1995; Vitousek *et al.*, 1997; National Research Council, 2000; Cloern, 2001; Rabalais, 2002). Under a "business as usual" scenario, Seitzinger *et al.* (2002) computed that future DIN loadings to coastal systems could approximately double yet again, by 2050.

Similarly, Harrison *et al.* (in press) used a spatially explicit, global model of river dissolved inorganic phosphorus (DIP) export to evaluate the relative magnitudes of sewage, fertilizer, manure, and weathering P sources, and the inclusion of reservoir retention and consumptive water use terms. Their model computed that of the $34\,\mathrm{Tg\ of\ P\,year^{-1}}$ loaded on watersheds by human activity

globally, approximately 2% (0.71 Tg year^{-1}) reaches river mouths as DIP; of the total predicted annual DIP export (1.1 Tg P year^{-1}), anthropogenic sources account for 65%, with the remaining 35% (0.38 Tg year^{-1}) attributable to natural weathering processes.

SUMMARY STATEMENT

The overall biogeochemistry of river basins at large scales is fundamental in establishing the overall dynamics of the landscape, its linkages to the atmosphere, and ultimately to the oceans. Changes in land use and climate will significantly impact the critical water resources that these basins represent. New tools enable us to make more rapid progress than ever in our understanding. A cautionary note, however, is that the fact remains that there are far too few field studies on the actual state of a surprisingly large number of these basins.

REFERENCES

Abril G., Nogueira M., Etcheber H., Cabecadas G., Lemaire E. and Brogueira M.J. (2002) Behaviour of organic carbon in nine contrasting European estuaries. *Estuarine Coastal and Shelf Science*, **54**, 241.

Aitkenhead J.A. and McDowell W. (2000) Soil C:N as a predictor of annual riverine DOC flux at local and global scales. *Global Biogeochemical Cycles*, **14**, 127–138.

Amon R.M.W. and Benner R. (1996a) Bacterial utilization of different size classes of dissolved organic matter. *Limnology and Oceanography*, **41**, 41–51.

Amon R.M.W. and Benner R. (1996b) Photochemical and microbial consumption of dissolved organic carbon and dissolved oxygen in the Amazon River system. *Geochimica et Cosmochimica Acta*, **60**, 1783–1792.

Baldock J.A. and Skjemstad J.O. (2000) Role of the soil matrix and minerals in protecting natural organic materials against biological attack. *Organic Geochemistry*, **31**, 697–710.

Benedetti M.F., Van Riemsdijk W.H., Koopal L.K., Kinniburgh D.G., Gooddy D.C. and Milne C.J. (1996) Metal ion binding by natural organic matter: from the model to the field. *Geochimica et Cosmochimica Acta*, **60**, 2503–2513.

Billen G., Lancelot C. and Meybeck M. (1991) N, P, and Si retention along the aquatic continuum from land to ocean. In *Ocean Margin Processes in Global Change*, Mantoura R.F.C., Martin J.-M. and Wollast R. (Eds.), John Wiley and Sons: New York, pp. 19–44.

Chin W.-C., Verdugo P. and Orellana M. (1998) Spontaneous assembly of marine dissolved organic matter into polymer gels. *Nature*, **391**, 568–572.

Clair T.A., Ehrman J.M. and Higuichi K. (1999) Changes in freshwater carbon exports from Canadian terrestrial basins to lakes and estuaries under a 2xCO_2 atmosphere scenario. *Global Biogeochemical Cycles*, **13**, 1091–1097.

Cloern J.E. (2001) Our evolving conceptual model of the coastal eutrophication problem. *Marine Ecology Progress Series*, **210**, 223–253.

Cole J.J. and Caraco N.F. (2001) Carbon in catchments: connecting terrestrial carbon losses with aquatic metabolism. *Marine and Freshwater Research*, **52**(1), 101–110.

Cole J.J., Caraco N.F., Kling G.W. and Kratz T.K. (1994) Carbon dioxide supersaturation in the surface waters of lakes. *Science*, **256**, 1568–1570.

Degens E.T. (1982) Riverine carbon – an overview. In *Transport of Carbon and Minerals in Major World Rivers, Pt 1*, Degens E.T. (Ed.), SCOPE/UNEP Sonderbd, Mitt. Geol.-Paläont. Inst. University: Hamburg, pp. 1–12.

Degens E.T., Kempe S. and Richey J.E. (1991) *Biogeochemistry of Major World Rivers*, John Wiley.

Devol A.H. and Hedges J.I. (2001) The biogeochemistry of the Amazon River mainstem. In *The Biogeochemistry of the Amazon Basin and its Role in a Changing World*, McClain M.E., Victoria R.L. and Richey J.E. (Eds.), Oxford University Press: UK.

Dunne T., Mertes L.A.K., Meade R.H., Richey J.E. and Forsberg B.R. (1998) Exchanges of sediment between the flood plain and channel of the Amazon River in Brazil. *Geological Society of America Bulletin*, **110**(4), 450–467.

Galloway J.N. and Cowling E.B. (2002) Reactive nitrogen and the world: two hundred years of change. *Ambio*, **31**, 64–77.

Galloway J.N., Schlesinger W.H., Levy H., Michaels A. and Schnoor J.L. (1995) Nitrogen fixation: atmospheric enhancement-environmental response. *Global Biogeochemical Cycles*, **9**, 235–252.

Gao Q., Tao Z., Shen C., Sun Y., Yi W. and Xing C. (2002) Riverine organic carbon in the Xijiang River (South China): seasonal variation in content and flux budget. *Environmental Geology*, **41**, 826.

Gibbs R.J. (1970) Mechanisms controlling world water chemistry. *Science*, **170**, 1088–1090.

Harrison J., Seitzinger S.P., Caraco N., Bouwman A.F., Beusen A. and Vörösmarty C. (In press) Soluble reactive phosphorus export to the coastal zone: results from a new, spatially explicit, global model (NEWS-SRP). *Global Biogeochemical Cycles*.

Hedges J.I., Clark W.A., Quay P.D., Richey J.E., Devol A.H. and Santos U.D.M. (1986b) Compositions and fluxes of particulate organic material in the Amazon River. *Limnology and Oceanography*, **31**, 717–738.

Hedges J.I., Cowie G.L., Richey J.E., Quay P.D., Benner R. and Strom M. (1994) Origins and processing of organic matter in the Amazon River as indicated by carbohydrates and amino acids. *Limnology and Oceanography*, **39**(4), 743–761.

Hedges J.I., Ertel J.R., Quay P.D., Grootes P.M., Richey J.E., Devol A.H., Farwell G.W., Schmitt F.W. and Salati E. (1986a) Organic carbon-14 in the Amazon River system. *Science*, **231**, 1129–1131.

Hedges J.I., Hatcher P.G., Ertel J.R. and Meyers-Schulte K.J. (1992) A comparison of dissolved humic substances from seawater with Amazon River counterparts by ^{13}C-NMR spectrometry. *Geochimica et Cosmochimica Acta*, **56**, 1753–1757.

Hedges J.I. and Keil R.G. (1995) Sedimentary organic matter preservation: an assessment and speculative synthesis. *Marine Chemistry*, **49**, 81–115.

Hedges J.I., Keil R.G. and Benner R. (1997) What happens to land-derived organic matter in the ocean? *Organic Geochemistry*, **27**, 195–212.

Hedges J.I., Mayorga E., Tsamakis E., Mcclain M.E., Aufdenkampe A.K., Quay P., Richey J.E., Benner R., Opsahl S., Black B., *et al.* (2000) Organic matter in Bolivian tributaries of the Amazon River: a comparison to the lower mainstem. *Limnology and Oceanography*, **45**, 1449–1466.

Ittekkot V. (1988) Global trends in the nature of organic matter in river suspensions. *Nature*, **332**, 436–438.

Ittekkot V. and Haake B. (1990) The terrestrial link in the removal of organic carbon in the sea. In *Facets of Modern Biogeochemistry*, Ittekkot V., Kempe S., Michaelis W. and Spitzy A. (Eds.), Springer Verlag: New York, pp. 318–325.

Ittekkot V. and Laane R.W.P.M. (1991) Fate of riverine particulate organic matter. In *Biogeochemistry of Major World Rivers*, Degens E.T., Kempe S. and ad Richey J.E. (Eds.), John Wiley & Sons: New York, pp. 233–243.

Jones J.B., Stanley E.H. and Mulholland P.J. (2003) Long-term decline in carbon dioxide supersaturation in rivers across the contiguous United States. *Geophysical Research Letters*, **30**, 2/1–2/4.

Kaiser K. (1998) Fractionation of dissolved organic matter affected by polyvalent metal cations. *Organic Geochemistry*, **28**, 849–854.

Kaiser K. and Guggenberger G. (2000) The role of DOM sorption to mineral surfaces in the preservation of organic matter in soils. *Organic Geochemistry*, **31**, 711–725.

Kao S.J. and Liu K.-K. (2002) Exacerbation of erosion induced by human perturbation in a typical Oceania watershed: Insight from 45 years of hydrological records from the Lanyang-Hsi River, northeastern Taiwan. *Global Biogeochemical Cycles*, **16**, 1–7.

Kaplan L.A. and Newbold J.D. (1993) Sources and biogeochemistry of terrestrial dissolved organic carbon entering streams. In *Aquatic Microbiology: An Ecological Approach*, Ford T.E. (Ed.), Blackwell Scientific: pp. 139–165.

Karlsson G., Grimvall A. and Lowgren M. (1988) River basin perspective on long-term changes in the transport of nitrogen and phosphorus. *Water Research*, **22**, 139–149.

Keil R.G., Mayer L.M., Quay P.D., Richey J.E. and Hedges J.I. (1997) Losses of organic matter from riverine particles in deltas. *Geochimica et Cosmochimica Acta*, **61**(7), 1507–1511.

Keil R.G., Montluçon D.B., Prahl F.G. and Hedges J.I. (1994) Sorptive preservation of labile organic matter in marine sediments. *Nature*, **370**, 549–552.

Küchler I.L., Miekeley N. and Forsberg B.R. (1994) Molecular mass distributions of dissolved organic carbon and associated metals in waters from Río Negro and Río Solimões. *Science of the Total Environment*, **156**, 207–216.

Lewis W.M., Melack J.M., Mcdowell W.H., Mcclain M.E. and Richey J.E. (1999) Nitrogen yields from undisturbed watersheds in the Americas. *Biogeochemistry*, **46**, 149–162.

Liang X., Lettenmaier D.P. and Wood E.F. (1996) One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model. *Journal of Geophysical Research*, **101**(D16), 21 403–21 422.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically based model of land surface, water, and energy fluxes for general circulation models. *Journal of Geophysical Research* **99**(D7), 14 415–14 428.

Lohmann D., Raschke E., Nijssen B. and Lettenmaier D.P. (1998) Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model. *Hydrological Sciences Bulletin*, **43**, 131–141.

Ludwig W. (2001) The age of river carbon. *Nature*, **409**, 466–467.

Ludwig W., Probst J.L. and Kempe S. (1996) Predicting the oceanic input of organic carbon by continental erosion. *Global Biogeochemical Cycles*, **10**(1), 23–41.

Martinelli L.A., Victoria R.L., Dematte J., Richey J.E. and Devol A. (1993) Chemical and mineralogical composition of Amazon River floodplain sediments, Brazil. *Applied Geochemistry*, **8**, 391–402.

Mayer L.M. (1994) Adsorptive control of organic carbon accumulation in continental shelf sediments. *Geochimica et Cosmochimica Acta*, **58**, 1271–1284.

Mayer L.M., Keil R.G., Macko S.A., Joye S.B., Ruttenberg K.C. and Aller R.C. (1998) Importance of suspended particulates in riverine delivery of bioavailable nitrogen to coastal zones. *Global Biogeochemical Cycles*, **12**, 573–579.

Mayorga E. and Aufdenkampe A.K. (2002) Processing of bioactive elements by the Amazon River system. In *The Ecohydrology of South American Rivers and Wetlands*, McClain M.E. (Ed.), IAHS Special Publication No. 6, IAHS: pp. 1–24.

McClain M.E., Richey J.E., Brandes J.A. and Pimentel T.P. (1997) Dissolved organic matter and terrestrial-lotic linkages in the central Amazon basin of Brazil. *Global Biogeochemical Cycles*, **11**, 295–311.

Meade R.H., Yuzyk T.R. and Day T.J. (1990) Movement and storage of sediment in rivers of the United States and Canada. In *Surface Water Hydrology*, Wolman M.G. (Ed.), Geol. of N. Am Geol. Soc. of America: Boulder, pp. 255–280.

Mertes L.A.K., Dunne T. and Martinelli L.A. (1996) Channel–floodplain geomorphology of the Solimões–Amazon River, Brazil. *Geological Society of America Bulletin*, **108**, 1089–1107.

Meybeck M. (1982) Carbon, nitrogen and phosphorus transport by world rivers. *American Journal of Science*, **282**, 401–425.

Meybeck M. (1991) C, N, P, and S in rivers: from soures to global inputs. In *Interactions of C, N, P and S, Biogeochemical Cycles*, Wollast R., MacKenzie F.T. and Chou L. (Eds.), Springer Verlag: Berlin.

Meybeck M. (1994) Origin and variable composition of present day riverborne material. In *Material Fluxes on the Surface of the Earth. Studies in Geophysics*, Council N.R. (Ed.), National Academy Press: Washington, pp. 61–73.

Meyer J.L., McDowell W., Bott T., Elwood J.W., Ishizaki C., Melack J.M., Peckarsky B.L., Peterson B.J. and Rublee P.A. (1988) Elemental dynamics in streams. *Journal of the North American Benthological Society*, **7**, 410–432.

Milliman J.D., Farnswortha K.L. and Albertin C.S. (1999) Flux and fate of fluvial sediments leaving large islands in the East Indies. *Journal of Sea Research*, **41**, 97–107.

Milliman J.D. and Syvitski J.P.M. (1992) Geomorphic tectonic control os sediment discharge to the ocean – the importance of small mountainous rivers. *The Journal of Geology*, **100**(5), 525–544.

Minshall G.W., Cummins K.W., Petersen R.C., Cushing C.E., Bruns D.A., Sedell J.R. and Vannote R.L. (1985) Developments in stream ecosystem theory. *Canadian Journal of Fisheries and Aquatic Science*, **42**, 1045–1055.

Moran M.A. and Zepp R.G. (1997) Role of photoreactions in the formation of biologically labile compounds from dissolved organic matter. *Limnology and Oceanography*, **42**, 1307–1316.

Mounier S., Braucher R. and Benaïm J.Y. (1999) Differentiation of organic matter's properties of the Río Negro basin by crossflow ultra-filtration and UV-spectrofluorescence. *Water Research*, **33**, 2363–2373.

National Research Council (2000) *Clean Coastal Waters. Understanding and Reducing the Effects of Nutrient Pollution*, National Academy Press: Washington.

Nixon S.W. (1995) Coastal marine eutrophication: a definition, social causes, and future concerns. *Ophelia*, **41**, 199–219.

Olson J.S., Garrels R.M., Berner R.A., Armentano T.V., Dyer M.I. and Taalon D.H. (1985) The natural carbon cycle. In *Atmospheric Carbon Dioxide and the Global Carbon Cycle*, Trabalka J.R. (Ed.), US Department of Energy.

Patel N., Mounier S., Guyot J.L., Benamou C. and Benaïm J.Y. (1999) Fluxes of dissolved and colloidal organic carbon, along the Purus and Amazonas Rivers (Brazil). *Science of the Total Environment*, **229**, 53–64.

Rabalais N. (2002) Nitrogen in aquatic ecosystems. *Ambio*, **31**, 102–122.

Raymond P.I. and Cole J.J. (2003) Increase in the export of alkalinity from North America's largest river. *Science*, **301**, 88–91.

Richey J.E. (2004) Pathways of atmospheric CO_2 through Fluvial systems. In *Toward CO_2 Stabilization: Issues, Strategies, and Consequences*, A SCOPE/GCP Rapid Assessment Project, Fields C. (Ed.), Island Press: pp. 329–340, 526.

Richey J.E., Hedges J.I., Devol A.H., Quay P.D., Victoria R., Martinelli L. and Forsberg B.R. (1990) Biogeochemistry of carbon in the Amazon River. *Limnology and Oceanography*, **35**, 352–371.

Richey J.E., Melack J.M., Aufdenkampe A.A.K., Ballester V.M. and Hess L. (2002) Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO_2. *Nature*, **416**, 617–620.

Schlesinger W.H. and Melack J.M. (1981) Transport of organic carbon in the worlds rivers. *Tellus*, **33**, 172–187.

Seitzinger S.P., Kroeze C., Bouwman A.F., Caraco N., Dentener F. and Styles R.V. (2002) Global patterns of dissolved inorganic and particulate nitrogen inputs to coastal systems: recent conditions and future projections. *Estuaries*, **25**, 640–655.

Smith S.V., Renwick W.H., Buddemeier R.W. and Crossland C.J. (2001) Budgets of soil erosion and deposition for sediments and sedimentary organic carbon across the conterminous united states. *Global Biogeochemical Cycles*, **15**(3), 697–707.

Stallard R.F. (1998) Terrestrial sedimentation and the carbon cycle: coupling weathering and erosion to carbon burial. *Global Biogeochemical Cycles*, **12**, 231–257.

Stallard R.F. and Edmond J.M. (1983) Geochemistry of the Amazon: 2. The influence of geology and weathering environment on the dissolved load. *Journal of Geophysical Research*, **88**, 9671–9688.

Stallard R.F. and Edmond J.M. (1987) Geochemistry of the Amazon. 3. Weathering chemistry and limits to dissolved inputs. *Journal of Geophysical Research*, **92**, 8293–8302.

Trumbore S.E., Davidson E.A., Camargo P., Nepstad D.C. and Martinelli L.A. (1995) Belowground cycling of carbon in forests and pastures of Eastern Amazonia. *Global Biogeochemical Cycles*, **9**, 515–528.

Vannote R.L., Minshall G.W., Cummings K.W., Sedell J.R. and Cushing C.E. (1980) The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 130–137.

Ver L.M., Mackenzie F.T. and Lerman A. (1999) Biogeochemical responses of the carbon cycle to natural and human perturbations: past, present and future. *American Journal of Science*, **299**, 762–801.

Vitousek P.M., Aber J.D., Howarth R.W., Likens G.E., Matson P.A., Schindler D.W., Schlesinger W.H. and Tilman D.G. (1997) Human alteration of the global nitrogen cycle: sources and consequences. *Ecological Applications*, **7**, 737–750.

Vörösmarty C.J., Meybeck M., Fekete B., Sharma K., Green P. and Syvitski J. (2003) Anthropogenic sediment retention: major global impact from registered river impoundments. *Global and Planetary Change*, **39**, 169–190.

Vörösmarty C.J., Sharma K., Fekete B., Copeland A.H., Holden J., Marble J. and Lough J.A. (1997) The storage and aging of continental runoff in large reservoir systems of the world. *Ambio*, **26**, 210–219.

Wigmosta M.S., Vail L.W. and Lettenmaier D.P. (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, **30**(6), 1665–1679.

World Commission on Dams (2000) *Dams and Development: A New Framework for Decision-making* Earthscan Publications: London and Sterling, p. 356.

PART 16

Land Use and Water Management

185: Integrated Land and Water Resources Management

IAN R CALDER

Centre for Land Use & Water Resources Research, University of Newcastle, Newcastle upon Tyne, UK

The need for Integrated Land and Water Resources Management (ILWRM), to meet and balance competing demands for food, timber and water, for conservation, amenity, recreation and the environment and for supporting people's livelihoods, is discussed. The principles of ILWRM including notions of sustainability, approaches to the economic valuation of land and water and demand management, are outlined. Also discussed are the differing public and science perceptions of the impacts of land use change, the interactions between land use, climate change, and water resources and how ILWRM can be implemented with stakeholder involvement.

THE NEED FOR INTEGRATED LAND AND WATER RESOURCES MANAGEMENT

There is increasing recognition for the need to better manage land use to meet and balance the competing demands being placed upon it. These demands can be in terms of production, for food, timber and water, for Conservation, Amenity, Recreation and Environment "products" (sometimes known as *CARE* products), and for supporting people's livelihoods. With projections of an increasing world population with increasing food requirements, there are concerns about where the extra water required to meet the increased agricultural production will come from (Falkenmark, 2003). There are also concerns that some of these increasing demands on land and water resources may be leading to degradations in quality and ecosystem function.

There is also the awareness that current land and water policies may not always be based on the best science, which may have led, in some instances, to perverse policy outcomes.

These are the drivers for improved integrated land and water resource management methodologies.

The Evolving Approach to Land and Water Management

The modern history of water resource and catchment management can be traced from its origins in the achievements of the nineteenth century engineers whose great civil engineering feats provided wholesome water to the world's growing and industrializing cities. Probably no other single factor contributed as much to the improved quality of life, and life expectancy, of city dwellers as this gift of safe water, literally available at the turn of a tap, and the provision of water-based sanitation systems, which then became possible. In those days, the surface water catchments or gathering grounds were managed to assure the pristine quality of water. Human occupation was regarded with distrust, at best a necessary evil, which had to be contained as far as possible. The success of the engineering approach was not just limited to water supply. The engineer had the ability to "tame the river". Through impoundments, barrages, and sluices, river flow from catchments could be regulated to reduce floods and to provide more water during times of drought.

More recently, the ethos of catchment and water resource management has shifted away from the tightly focused engineering viewpoint. At the same time, our perceptions of what we mean by a catchment have subtly changed. Originally, this may have meant just the headwaters where impoundments had been built to capture water for supply, irrigation, and hydropower purposes. Now, catchments are regarded more as the hydrological units that occupy the whole land surface of the globe.

As demands on water increase this has to be the case. Upland headwater catchments can no longer meet our

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

needs. The need to recycle water, together with the need to exploit groundwater means that more and more we regard every part of the land surface of the globe now as part of a catchment that can either supply water or receive our watery discharges.

The spatial linkages between upstream and downstream users and the "externalities" – the effect of one party's actions that impose a cost or benefit on another party without that cost or benefit being accounted for in the market – are becoming increasingly recognized. For example, land-based actions that alter catchment flows through alteration of the evaporative regime are now being considered and addressed through legislation. In South Africa, plantation forestry is designated as a "Stream Flow Reduction Activity". Soil water retention structures, ranging in scale from field bunds to check dams, farm dams, village dams through to large dams and reservoirs, may all have benefits in terms of storing or infiltrating water, but all have a cost in terms of increased evaporation from their water surfaces and their wet riparian zones. When soil water conservation measures are applied in excess within a catchment, they may have the same cumulative impact in terms of increased evaporation as that from a large reservoir. Together with high demands for irrigation water, the uncontrolled and unregulated implementation of such measures can contribute to the growing concern around the world arising from the phenomenon of "catchment closure", where, particularly in dry years, no flow emanates from a catchment.

The human and environmental dimension is achieving much greater prominence now in catchment and water resource management. Organizational and governance issues including institutional arrangements for collective action and community participation; decentralization of land and water management; laws, policies, and markets relating to watershed services and access to land and water; equity and poverty alleviation and are now central to what has been termed the "*Blue Revolution*" in land and integrated water resources management (Calder, 1999; Calder, 2005).

PRINCIPLES OF INTEGRATED LAND AND WATER RESOURCES MANAGEMENT

Concepts and Principles of Integrated Land and Water Resources Management

ILWRM involves the coordinated planning and management of land, water, and other environmental resources for their equitable, efficient, and sustainable use. ILWRM programs need to be developed alongside, and not in isolation from economic structural adjustment and other sectoral programs. For ILWRM strategies to be implemented, fragmentation of institutional responsibilities must be reduced.

ILWRM objectives encompass the United Nations Conference on Environment and Development (UNCED) principles:

- Water has multiple uses, and water and land must be managed in an integrated way.
- Water should be managed at the lowest appropriate level.
- Water allocation should take account of the interests of all who are affected.
- Water should be recognized and treated as an economic good.

ILWRM strategies seek to ensure

- a long-term viable economic future for basin dependants (both national and transnational);
- equitable access to water resources for basin dependants;
- the application of principles of demand management and appropriate pricing policies to encourage efficient usage of water between the agricultural, industrial, and urban supply sectors;
- in the short term, the prevention of further environmental degradation, and in the longer term, the restoration of degraded resources;
- the safeguarding of local cultural heritage and the local ecology as they relate to water management and the maintenance and encouragement of the potential for water-related tourism together with linkages between tourism and conservation.

ILWRM strategies should recognize that

- Solutions must focus on underlying causes not merely their symptoms;
- Issues must be approached in an integrated way;
- In general, development of sound resource management and collective responsibility for resources will take place at the subregional or village level.

History and the Dublin Principles

United Nations organizations have played a major role in the development of the new approach to water management. The United Nations Conference on the Human Environment in Stockholm in 1972 and the United Nations sponsored Conference on Water at Mar del Plata, Argentina, 1977 were instrumental in promoting the importance of water and water management to world governments. The delegates at the International Conference on Water and the Environment (ICWE) in Dublin, Ireland, in January 1992 called for a fundamental new approach to the assessment, development, and management of freshwater resources, which could only be brought about through political commitment and involvement from the highest levels of government to the smallest communities. They recognized that "commitment

will need to be backed by substantial and immediate investments, public awareness campaigns, legislative and institutional changes, technology development, and capacity building programmes. Underlying all these must be a greater recognition of the interdependence of all peoples, and of their place in the natural world".

The Dublin Conference Report, which was presented to the United Nations Conference on Environment and Development (UNCED) in Rio de Janeiro held in June 1992, sets out recommendations for action at local, national, and international levels based on four guiding principles (known now as the Dublin or UNCED principles).

Principle No. 1:

Freshwater is a finite and vulnerable resource, essential to sustain life, development, and the environment. Since water sustains life, effective management of water resources demands a holistic approach, linking social and economic development with protection of natural ecosystems. Effective management links land and water uses across the whole of a catchment area or groundwater aquifer.

Principle No. 2:

Water development and management should be based on a participatory approach, involving users, planners, and policy-makers at all levels. The participatory approach involves raising awareness of the importance of water among policy-makers and the general public. It means that decisions are taken at the lowest appropriate level, with full public consultation and involvement of users in the planning and implementation of water projects.

Principle No. 3:

Women play a central part in the provision, management, and safeguarding of water. This pivotal role of women as providers and users of water and guardians of the living environment has seldom been reflected in institutional arrangements for the development and management of water resources. Acceptance and implementation of this principle requires positive policies to address women's specific needs and to equip and empower women to participate at all levels in water resources programs, including decision making and implementation, in ways defined by them.

Principle No. 4:

Water has an economic value in all its competing uses and should be recognized as an economic good. Within this principle, it is vital to recognize first the basic right of all human beings to have access to clean water and sanitation at an affordable price. Past failure to recognize the economic value of water has led to wasteful and environmentally damaging uses of the resource. Managing water as an economic good is an important way of achieving efficient and equitable use, and of encouraging conservation and protection of water resources.

The Dublin Principles address questions of sustainability, equity, and economic efficiency. Originally, the term "sustainability" was used in the context of environmental and ecological sustainability, but more recently it has also been applied to peoples "livelihoods" and thereby includes the "equity" dimension.

NOTIONS OF SUSTAINABILITY

Tragedy of the Commons

Some of the early discussions and awareness of issues that we now conceptualize in terms of "sustainability" or "sustainable development" were provoked by Hardin (1968). Referring to the tradition of a community pasture, "the commons", Hardin gave the following example of a society that permitted freedom of action in activities that affected common property and was eventually doomed to failure.

> Picture a pasture open to all. It is to be expected that each herdsman will try to keep as many cattle as possible on the commons. Such an arrangement may work reasonably satisfactorily for centuries because tribal wars, poaching, and disease keep the numbers of both man and beast well below the carrying capacity of the land. Finally, however, comes the day of reckoning, that is, the day when the long desired goal of social stability becomes a reality. At that point the inherent logic of the commons remorselessly generates a tragedy. . . Ruin is the destination towards which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons.

The Limits to Growth

A team from MIT (Meadows *et al.*, 1972) presented in the book "The Limits to Growth", a crisis scenario of the depletion of the world's resources of fossil fuels, metals, timber, and fish and raised serious concerns about the sustainability of the world's ecosystem. It is now realized that many of the assumptions on which the computer predictions were based, particularly in relation to the extent of the reserves of fossil fuels and metals, were flawed, and consequently many of the dire forecasts that were made were erroneous. These failures have to some extent undermined the legitimacy of some very real concerns about the sustainability of land and water resources that are now faced in the twenty-first century.

The Brundtland Report

The United Nations World Commission on Environment and Development, chaired by Norwegian Prime Minister

Gro Harlem Brundtland, debated and investigated current concerns of the mid-1980s about the nature of development and the long-term consequences of certain development pathways. The report that was produced, titled "Our Common Future", but widely known as *the Brundtland Report* (United Nations World Commission on Environment and Development, 1987) reframed the environmental debate and laid the foundations for what was to become a new approach to development issues. The report provided a definition of sustainable development:

> Humanity has the ability to make development sustainable – to ensure that it meets the needs of the present without compromising the ability of future generations to meet their own needs.

Freshwater, together with food, energy, basic housing, and health are recognized as basic human needs. Sustainability introduces notions of sound management of the world's resources that leave them in as good a condition for the next generation as we find them today.

The report also advanced seven strategic imperatives and seven preconditions for sustainability to be achieved. The strategic imperatives advanced were as follows:

1. Growth be revived in the developing nations, to alleviate poverty and reduce pressure on the environment.
2. Notions of equity and nonmaterialistic values be included in the definition of growth.
3. Essential human needs for food, housing, and energy be met whilst accepting that this will necessitate changing patterns of consumption.
4. The issue of population growth be addressed particularly through reducing the economic pressures to have children.
5. The resource base be conserved and enhanced.
6. The necessary environmental risk-management technology be developed and also made available to the developing world.
7. Ecological as well as economic factors be taken into account in decision making.

Seven preconditions for these imperatives were identified as

1. responsive political decision making processes;
2. economic systems that make less resource demands;
3. responsive social systems that maintain union by redistributing both the costs and benefits of development;
4. production systems that can operate within ecological limits;
5. technology developments that support energy and resource efficient solutions;
6. international order that maintains cohesion globally;
7. responsive, flexible, self-correcting governments.

The report suggests that the shift to sustainable development must be powered by a continuous flow of wealth from industry but recognizes that future wealth creation will need to be much less environmentally damaging, more just, and more secure.

The report has been widely applauded for taking a long term and strategic, rather than piecemeal, approach to dealing with sustainability.

The Brundtland Report provided the stimulus for environmental organizations, industry, and development agencies to rethink their strategies and to develop the detail in the processes needed to address sustainability issues. Some of these detailed strategies, the "blueprints" for sustainability, have been defined, reflecting the needs of business, conservation, and development interests: The Natural Step, the Triple Bottom Line, Sustainable Livelihoods and the Pentagon, and Integrated Economic and Environmental Satellite Accounts.

The Natural Step

The concepts behind the Natural Step were originated by Dr Karl-Henrik Robert and are now being developed by the Natural Step Foundation (**www.naturalstep.org**). The Natural Step philosophy is based on four principles or "system conditions":

1. Substances from the earth's crust must not be extracted at a rate faster than their slow redeposit into the earth's crust.
2. Substances must not be produced by society faster than they can be broken down in nature or deposited into the earth's crust.
3. The physical basis for nature's productivity and diversity must not be allowed to deteriorate.
4. There must be fair and efficient use of energy and other resources to meet human needs.

The Natural Step Foundation claim that:

> The four system conditions provide a descriptive framework for a sustainable society. Participants on all levels – households, corporations, local authorities, nations – can systematically direct their activities to fit into this frame by requiring all secondary goals to function as natural steps in the process of achieving the four conditions of sustainability.

The World Business Council for Sustainable Development (WBCSD), a coalition of 125 international companies that claim a shared commitment to the environment and to the principles of economic growth and sustainable development, endorse the Natural Step approach. The approach is also supported by conservation organizations (IUCN, UNEP and WWF).

The Triple Bottom Line

Whereas the Natural Step focuses on the physical aspects of sustainability, the Triple Bottom-Line approach recognizes explicitly the social aspects. The ideas behind the Triple Bottom Line were developed by SustainAbility, (**www.sustainability.com**), a strategic management consultancy and think-tank concerned with foresight, agenda-setting, and change management. The elements of the "triple bottom line" are seen as representing society, the economy, and the environment (Elkington, 1997). Society is seen as dependent on the economy – and the economy dependent on the global ecosystem, whose health represents the ultimate "bottom line". The "bottom line" metaphor arose from business' use of the term to represent the profit figure in a company's earning-per-share statement. To arrive at this figure, accountants will, as part of standard accounting practice, collate, record, and analyze a wide range of numerical data that relates to economic performance. This approach is seen as the model for social and environmental accounting, to allow the calculation of the social and environmental "bottom lines".

The three bottom lines are not regarded as stable, but in constant flux, because of social, political, economic and environmental pressures, cycles, and conflicts. Elkington uses the example of plate tectonics to describe the movement of the lines:

> Think of each bottom line as a continental plate, often moving independently from the others. As the plates move under, over or against each other, 'shear zones' emerge where the social, economic, or ecological equivalents of tremors and earthquakes occur.

SustainAbility's Perceptions of the Shear Zone Interactions:

Economic/environmental

In the economic/environmental shear zone, some companies already promote eco-efficiency. But there are greater challenges ahead, for example, environmental economics and accounting, shadow pricing, and ecological tax reform.

Social/environmental

In the social/environmental shear zone, business is working on environmental literacy and training issues, but new challenges will be sparked by, for example, environmental justice, environmental refugees, and the intergenerational equity agenda.

Economic/social

In the economic/social shear zone, some companies are looking at the social impacts of proposed investment, but bubbling under are issues like business ethics, fair trade, human and minority rights, and stakeholder capitalism.

Sustainable Livelihoods and the Pentagon

The Department for International Development (DFID) White Paper on International Development (DFID, 1997), *Eliminating World Poverty: A Challenge for the twenty-first century*, presents the concept of the stewardship of natural resources so that the needs of both present and future generations can be met. Sustainability does not rely upon a "quick fix" solution becoming available in the future to reverse degradation. The White Paper also promotes the concept of "sustainable livelihoods" and this, together with the management of "the natural and physical environment" is expected to achieve the overall goal of poverty alleviation.

DFID provides the following definition for a sustainable livelihood:

> A livelihood comprises the capabilities, assets (including both material and social resources) and activities required for a means of living. A livelihood is sustainable when it can cope with and recover from stresses and shocks and maintain or enhance its capabilities and assets both now and in the future, while not undermining the natural resource base.

On the basis of the work of the Institute of Development Studies (Scoones, 1998), five types of assets, upon which individuals draw to build their livelihoods, are defined: natural capital, social capital, human capital, physical capital, and financial capital.

Capital Assets (Carney, 1998)
Natural capital

The natural resource stocks from which resource flows useful for livelihoods are derived (e.g. land, water, wildlife, biodiversity, environmental resources).

Social capital

The social resources (networks, membership of groups, relationships of trust, access to wider institutions of society) upon which people draw in pursuit of livelihoods.

Human capital

The skills, knowledge, ability to labor and good health, important to the ability to pursue different livelihood strategies.

Physical capital

The basic infrastructure (transport, shelter, water, energy, and communications), and the production equipment and means that enable people to pursue their livelihoods.

Financial capital

The financial resources that are available to people (whether savings, supplies of credit, or regular

remittance, or pensions), and that provide them with different livelihood options.

The consequences for natural resource management of applying this new "people-first" poverty-focused approach to sustainable development are not yet known. There are concerns that poverty-alleviation approaches focused at the microcatchment scale may result in "tragedy of the commons" type impacts at larger scales. Conversely, it could be argued that Integrated Water Resources Management (IWRM) approaches, although aiming to achieve net economic benefits to basin inhabitants, may not be taking sufficient account of the poorest in society.

Implementing the Ideals – the Blue Revolution

The different blueprints for obtaining sustainability objectives: the Natural Step, the Triple Bottom Line, the Pentagon, and Integrated Economic and Environmental Satellite Accounts (IEESAs), are not mutually exclusive. Nor are they definitive, and further efforts will be needed to achieve the Blue Revolution (Calder, 1999) in land use and integrated water resource management methodologies.

It is also recognized that successful implementation of integrated water resource management requires that the complementary elements of an effective water resources management system be developed and strengthened concurrently. These elements include

1. the enabling environment – the general framework of national policies, legislation and regulations, and information for water resources management stakeholders;
2. defined institutional roles and functions of the various institutions and stakeholders;
3. the management instruments, including operational instruments for effective regulation, monitoring, and enforcement that enable the decision-makers to make informed choices between alternative actions. These choices need to be based on agreed policies, available resources, environmental impacts, and the social and economic consequences.

The integrated water resource management process, which encompasses the ideals described above, has been defined by the Global Water Partnership as:

> "IWRM is a process which promotes the coordinated development and management of water, land and related resources in order to maximise the resultant economic and social welfare in an equitable manner, without compromising the sustainability of vital ecosystems" (GWP Technical Advisory Committee, 2000)

ECONOMICS OF LAND AND WATER

The Value of Water

Enshrined in the UNCED principles is the concept that water should be recognized and treated as an economic good, and economists are now able to devise methods for calculating the value of water as a commodity in its many uses.

The advantages to be gained by ascribing a value to water, and charging users a tariff related to this value include

1. provision of funds for water developments;
2. reducing demands on the public sector for the capital and recurrent cost of providing water;
3. releasing water for higher value use and assisting the prioritization of water allocation;
4. the resolution of resource conflicts between, for example, demands for water, power, fisheries, transport, and so on;
5. environmental benefits, through demand management, by releasing more water for environmental usage;
6. demand management, reducing the demand for water.

Provision of funds for water developments becomes increasingly important as cheaper sources of water become used up. The World Bank (Bhatia and Falkenmark, 1992) estimated that the cost of providing water from the "next" project was often two to three times that of providing water from the "current" project. Although US$10 billion was being spent each year on improving water supply and sanitation in the developing world in the early nineties, it was estimated by the World Bank that the investment would, even if costs were fixed, have to be five times this rate if reasonable water services were to be had by all by the year 2000. Cost recovery is, therefore, an important feature of a water pricing policy.

Reducing demands on the public sector and donor organizations as a result of cost recovery should, in theory, release funds for other forms of development. The World Bank estimated that in 1991, only 10% of the cost of water projects that it funded were financed by internal cash generation.

Releasing water for higher value uses will come about if water pricing is set at such a level that it discourages lower value uses. Water pricing can, therefore, be an effective instrument in water allocation.

The resolution of resource conflicts may be aided by realistic resource pricing instruments. Competition between agriculture, environmental, and hydropower requirements for water would be eased in many countries through water pricing, and agricultural demands for electric power for pumping groundwater for irrigation in countries such as India and China might be alleviated through realistic pricing of both water and electricity.

Environmental benefits. The environment as a valid user of water is becoming increasingly recognized. Also recognized is the increasing damage to the environment that comes through the growing supply and consumption of water. Abstractions for water supply, agriculture, and industry means less flow in rivers and the drawdown of water bodies. Lower flows in rivers also reduces the assimilative capacity and, with the discharge of pollutants, may lead to toxic conditions for wildlife and extra costs to public services for the treatment of water. Demand management, through curbing water use may have environmental benefits. In Australia, increasing charges for irrigation water is one of the instruments proposed to curb increasing salinization problems.

Demand management. Arguably, within the context of IWRM, the greatest benefits to be achieved from a pricing policy for water are in relation to controlling demands for water in situations where the water resource is close to being fully exploited.

DEMAND MANAGEMENT

> The equivalent of the Hippocratic Oath for water engineers is to promise to meet all reasonable needs for water without question by enlarging and improving supplies. – Winpenny (1992)

In many engineeringly developed countries, the ethos of the water engineer was to equate efficiency with maximum utilization of the water resource for the users. Taken to its limit, any residual flows in rivers were seen as wastage. With this ethos, environmental and ecological considerations and downstream users, especially if they were transnational users, were given little weight in the quest for "Development".

The trend away from supply-driven large water schemes toward greater demand management was well summarized by Arthur (1997); he recognized that "in the background of the differentiation between need and want, lies a much wider ideological conflict about the limits to growth".

In most countries, the accepted response to water supply shortages was to augment supplies. In the short-term demand management devices such as rationing, prohibited uses, public exhortation, and the use of standpipes, or the cutting off of supplies, are widely applied. But these nonprice devices, although often effective, are both costly and inconvenient to users and do not take account of the relative value of water to different consumers.

Where water is provided to users at a price less than the supply cost, a situation common in most parts of the world except the UK, the incentive for conservation and waste reduction is absent. This *negative* demand management leads to the paradox that in a situation when the water resource is already under stress, the subsidy is actually encouraging users to make additional demands upon it.

This situation is exemplified in India where subsidies to farmers in the form of free electricity encourages the pumping of irrigation water from boreholes from ever greater depths, leading to reduced environmental flows from catchments, and reduced access to water from downstream users.

The alternative approach, consistent with the concepts of IWRM, is to recognize that water resources are limited and new sources cannot be developed indefinitely. Demand management, in some form, must ultimately be applied, and the use of pricing is an instrument that can be both effective and can be defended by rational and objective arguments.

Although governments see many political hurdles in introducing water pricing policies, the principal concern being a short-term loss of votes, a secondary concern being the inflationary effect of charging for a commodity that is universally used, the benefits are increasingly being regarded as outweighing the disadvantages.

The development of a pricing policy for managing water demand does, however, require a methodology for deciding on the value of water to determine the price to be charged.

WATER AS AN ECONOMIC COMMODITY

Traditionally, water has been regarded as a "free" resource of unlimited supply with zero cost at the point of supply. Users have been charged for (often only a proportion of) the costs of transfer, treatment, and disposal. Opportunity costs for water are often ignored, and as a consequence, users have little incentive to ensure that water is used efficiently and not wasted.

The new (economic) approach to the allocation of water is to use prices and markets to ensure efficiency, and that water is supplied to its most valuable uses.

In an economic sense, efficiency requires that:

1. **Marginal Benefit of Use > Marginal Cost of Supply.** Which means that the users are able to derive greater economic benefit from the next unit of water supplied than the cost expended by the supplier in providing that extra unit of water.
2. **Marginal Benefit per unit of Resource is equal across all uses**. Where the supply of water is not unlimited, and the marginal units of supply have a positive cost, that is the supplier has to expend money to develop the next unit of supply water, the value of water consumption is maximized when net marginal benefits are equal in all uses. This implies that efficiency would be increased by transferring water between users until the marginal value of the water to the users is the same across all sectors. When equality of marginal values is achieved, further redistribution of water would make no sector better off without making another sector worse off.

In theory, it is possible to derive demand and supply curves that show both:

- the marginal benefit obtained from consumption and the willingness to pay; and
- the marginal cost of supply and the willingness to supply at given prices.

VALUING WATER

Various methods have been developed to determine the value of water in its many uses. Those, reviewed by Winpenny (1996) include

1. willingness to pay (WTP);
2. marginal cost analysis;
3. "netback";
4. WTP taking into account recycling and reuse costs;
5. in-stream value for pollution assimilation;
6. in-stream value for transport;
7. hydropower generation with and without capital costs included;
8. travel Cost and Contingent Valuation methods for amenity and recreational use.

For urban household consumers, the willingness-to-pay (WTP) approach has often been adopted to water pricing. This involves determining the demand curve, the relationship between the amount of water used by consumers and the charging price. This information can be obtained either by a survey, or from a knowledge of the change in consumption following a price change.

For agricultural use, the value of water is often determined in terms of marginal productivity. Here the economic benefits through increased yield or quality of the crop resulting from a unit addition of water are calculated under conditions when all other farm inputs are held constant. An alternative method *known as* "Netback" is related to the WTP principle. By taking the gross value of the crop per hectare and subtracting all production costs, and if required, a capital recovery cost and an acceptable profit margin, the remainder can be viewed as the maximum WTP for the water.

The same methodology can be applied to industrial water valuation, but as water is usually only a small part of total production costs, it would be misleading to attribute the whole of the remainder in the value and cost calculations to water valuation. For industrial use, another method is to regard the cost of recycling water, following water treatment as the upper limit on industrial WTP. The reasoning would be that if water pricing were greater than the treatment costs, industry would opt for treatment rather than buying water.

In-stream values for water arise through its natural capacity to assimilate waste and pollutants. This value can be compared with the alternative cost of reducing the pollutants at source or the additional costs incurred in water treatment. In-stream values for water for transport can be calculated as the cost advantage of water transport over the next cheapest form of transport (Winpenny, 1996).

Values of water for amenity and tourism have been calculated using the Travel Cost and Contingent Valuation methods. The travel cost method infers the amenity value from the travel costs of visitors to the site, which are then used to construct a hypothetical demand curve for the amenity or recreational value of the water-body. The Contingent Valuation method relies on opinion surveys to reveal the value that visitors derive from the water-body.

Winpenny (1996) has reviewed the values obtained for different uses of water in different parts of the world using the methods outlined earlier. Although he introduces the caveat that it is important to examine the fine print associated with each valuation before using the valuations as representative of the sector, there is a general picture that emerges: industrial uses are always high value, together with speciality crops and domestic uses (Table 1). Recreational uses can be high value whilst the majority of agricultural crops, often one of the largest overall consumers of water, are relatively low-value uses.

THE VALUE OF LAND USE

Whilst methods have been devised for valuing water in its many diverse uses including those for water supply, irrigation of crops, hydropower generation, industrial production, mining, amenity and recreational uses, and for the environment, there is also a need for methods for valuing land use. Within an IWRM context this is particularly true for land uses such as forests, which although having a high impact on water resources may have valuable multiple uses. These may include not only primary uses for timber production and other forest products but also secondary uses for recreation, conservation, and tourism. The valuation of land uses for primary production is inherently straightforward, although tax incentives and production subsidies may significantly effect the valuation. Valuation methods for secondary uses are less straightforward and more controversial but nevertheless important; in some circumstances, secondary land uses may be at least as valuable. Trade-offs need then to be considered between the value of water that may be forgone under a particular land use against the sum value of the different uses of the land.

Some of the methods that have been used for assessing the amenity and recreational value of land use are identical to those used for assessing the amenity and recreational value of water. Willis and Benson (1989) have discussed the use of travel cost and contingent valuation methods

Table 1 Estimated value of water for different uses, US$ per 1000 cubic meters

Use	USA Gibbons (1986)	China Dixon *et al.* (1994) Kutcher *et al.* (1992) Adams *et al.* (1994)	UK Rees *et al.* (1993) Bate and Dubourg (1994)	Zimbabwe Winpenny (1996)	South Africa Hassan *et al.* (1995)
Industrial process	180–800	500–4000		>230	
Speciality crop	100–800		80–140 (fruit) 1890 (potatoes)	100–150	
Domestic	20–360			>120	
Recreational	10–400				
Navigation	generally 0 but 370 in some stretches				
Intermediate value farm crops		average: 30 critical times: 90–100			100 (sugar)
Low value farm crops	10–60	<120	50–90 (field vegetables) 10–30 (cereals & grass)		100 (dryland crops) 0 (traditional livestock rearing)
Waste assimilation	0–20				
Industrial cooling	0–10				
Sediment prevention		0–20			
Power		20			

for assessing the recreational value of UK forests. They also discussed other "market related" methods including the Hedonic Price Method which links environmental assets with markets for private goods and services. The linkage of wages or house prices to environmental attributes were cited as examples that could be followed with this approach, but the practical difficulty in obtaining information on housing characteristics and sale prices of houses deterred the authors from pursuing this method for estimating the recreational value of forests. Using the travel cost method, Willis and Benson arrived at an average recreational value of UK £1.90 for one visit to a forest owned by the UK Forestry Commission. They estimated that the wildlife attributes of the forests contributed about 38% toward this value and showed that the total annual recreational benefit from the Forestry Commission's forests lay between £14 million and £45 million per year, later to be reassessed at £53 million (Benson and Willis, 1991).

Aylward and colleagues (Aylward *et al.*, 1998) have made an analysis of the use and nonuse values of different land uses on the catchment of Lake Arenal, the source area for Costa Rica's largest hydroelectric facility and irrigation scheme. This landmark study, entitled "Economic Incentives for Watershed Protection" takes into account both valuation, particularly non-use valuation related to watershed protection, and institutional analyses. Aware of modern research that has demonstrated that forests generally reduce water flows as compared with shorter crops, Aylward and colleagues state in their conclusions that "by examining the issue of externalities in detail the study shows that the crucial hydrological externality (water yield) and its relative direction (positive) are contrary to that expected by previous characterizations of the problem." They go on to say "while considerable variability can be expected in applying the valuation and institutional analyses to other sites and conditions, at a minimum this case study suggests the benefits of integrating the(se) two aspects of watershed analysis under a single framework. Additional case studies and more general theoretical work should assist in the development of a defensible consensus around rules of thumb, and shortcuts in such analyses that would contribute to better policy and project formulation. Such guidance is desperately needed given the current reliance on partial analyses and outdated conventional wisdom of what constitutes watershed protection, and watershed management in the humid tropics".

For the successful integration of land and water management, robust and accepted methods for valuing both water and land use will increasingly be required in the future. This is especially true where land and water are associated with environmental and recreational attributes, and where markets for environmental services are being developed (*see* **Chapter 193, Markets for Watershed Services, Volume 5**).

ENVIRONMENTAL ACCOUNTING

Environmental movements, including the IUCN, have long recognized that the System of National Accounts as defined by the United Nations, and implemented by governments worldwide does not accurately incorporate or take account the environment. National accounts are the economic data systems used to calculate familiar macroeconomic indicators such as gross national product (GNP), gross domestic product (GDP), savings rates, and income per capita. They are built and maintained by governments, following standard accounting practices defined largely through an international process coordinated by the United Nations.

Environmental accounting is seen as the mechanism by which national accounting systems can be modified to account for the economic role played by the natural environment. To increase international acceptance and implementation of environmental accounting, IUCN launched the first phase of its Green Accounting Initiative in 1996 (IUCN, 1998).

Several dozen countries have experimented with implementing environmental accounting methods to better understand economy-environment interactions and to test environmental strategies. Norway is one of the few countries to have institutionalized environmental accounting as a routine government activity.

Other countries have been active in developing environmental accounting methodologies. The US Bureau of Economic Analysis (BEA) has been responsible for developing the system of national economic accounts, which are used to produce the national income and product accounts, input–output accounts, and balance sheets for the US economy. Responding to the need for environmental accounting (Landefeld and Carson, 1994), the BEA produced a new accounting framework that covers the interactions of the economy and the environment.

It is expected that fully implemented IEESAs would allow the identification of the economic contribution of natural and environmental resources broken down by industry, by type of income, and by product.

A number of UK management agencies have commissioned and developed a variant of this approach as a basis for public participation. This approach was originally known as *Environmental Capital*, and more recently as "Quality of Life Capital". It is discussed more fully in **Chapter 192, Public Participation in River Basin Planning and Management: Quality-of-Life Capital as an Information Aid to Sustainable Decisions, Volume 5**.

IMPACTS OF LAND USE CHANGE, PUBLIC, AND SCIENCE PERCEPTIONS

Before we are in a position to devise and develop land and water policies that are aimed at improving either the water environment from the ecosystem perspective, water resources, economic returns, or for improving the livelihoods of poor people living within the catchment, it is important that we fully understand the biophysical impacts of any change in land use on the water regime. Arguably, the land use changes that have the largest impacts on water resources involve forestry and irrigated agriculture. It is clear that the "science" and "public" perceptions of the impacts of these activities are not always in agreement.

It is also important that the impacts of climate change, which may or may not occur at the same time as land use change, are also understood.

Forests and Hydrological Services

The conventional wisdom that forests are in all circumstances necessarily good for the water environment and, that they increase rainfall, increase runoff, regulate flows, reduce erosion, reduce floods, "sterilize" water supplies and improve water quality, has long been questioned by the scientific community. Although these issues have been debated since the nineteenth century (Saberwal, 1997), the "modern" science perception has been presented in a number of reviews by Bosch and Hewlett (1982), Hamilton and King (1983), Hamilton (1987), Bruijnzeel (1990), Calder (1992), particularly as regards tropical forests, and the more recent reviews, in the light of new studies by Calder (1999); Calder, 2000; Calder, 2005 and Bruijnzeel (2004), and the summary given in **Chapter 186, Water and Forests, Volume 5**.

On the basis of this conventional wisdom, in many countries of the world, forest and afforestation programs are still being promoted within watershed development programs on the basis of these "hydrological services" and "headwater conservation functions". In turn, the expectation is often that the increased "hydrological services" will benefit the livelihoods of poor people through increased access to water supplies. Planting trees to increase local rainfall and runoff are some of the misconceptions about water and land use that have influenced watershed development in India (KAWAD, 2001; Batchelor *et al.*, 2003; Calder and Gosain, 2003). It is believed that watershed management policies, which, together with increased afforestation, also promote increased irrigation from groundwater supplies as a means of improving the livelihoods of the poor, may be resulting in perverse outcomes (Calder and Gosain, 2003). In many Indian states, implementation of these policies has contributed to the lowering of groundwater tables to the extent that local people are now unable to access groundwater supplies through hand pumps and are having to purchase water from tanker deliveries.

Saberwal (1997) has reviewed the policy discourse in India, and traces the policy makers acceptance of the

positive link between water and forest to a "desiccationist" discourse promoted by Indian foresters since the middle of the twentieth century. The need to reconcile the public and science perceptions of forest and hydrological services has been outlined by Calder (2002), and is described in more detail in **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**.

Financing mechanisms designed to help conserve and protect indigenous forests, and to partly support the costs of reforestation programs have traditionally referred to, and are also often based on, this conventional wisdom. In Costa Rica, both the government led Payment for Environmental Services (PES) program, and all private agreements with the hydroelectric sector were initially based on this "universally accepted" knowledge, although currently it is defended as a "precautionary principle" in the absence of better – or more accepted – scientific knowledge. Further details on the development of markets for watershed services are given in **Chapter 193, Markets for Watershed Services, Volume 5**.

Land Use, Climate Change, and Water Resources

The interactions between land use, climate change, and water resources is an active research area, but one in which, as yet, there remains little consensus regarding the scale of the effects. Central to the research effort is the use of Global Circulation Models (GCMs) to represent the transfer of heat, water vapor, and momentum between the surface of the earth and the atmosphere. The GCMs require values for the parameters relating to the surface vegetation type and the availability of water to the vegetation, to allow the estimation of the heat and mass-transfer terms, under the atmospheric and rainfall conditions calculated by the GCM (*see* **Chapter 189, Land Use and Water Resources Under a Changing Climate, Volume 5**). Some of the limitations of the approach, particularly in relation to calculating the impacts of anthropogenic (greenhouse effect) climate change on water resources, have been identified by Bonell (1999). Bonell recognizes, quoting a recent International Hydrology Programme (IHP) expert group (Shiklomanov, 1999), that, even given the same initial starting conditions, different Global Climate Model (GCM)s lead to widely differing estimates of the extent of the climate change.

Nevertheless, at present GCMs represent the best technology available for predicting future climate change, and whilst it must be accepted that considerable uncertainty is attached to future GCM predictions, there is a need to understand how these climate change scenarios translate to future water resource scenarios. Further discussion on land use and water resource management under a changing climate are given in **Chapter 189, Land Use and Water Resources Under a Changing Climate, Volume 5**.

Agricultural Irrigation and Water Resources

To satisfy ambitions of food security and self-sufficiency, agricultural research for developing countries has traditionally focused, almost exclusively, on means of increasing productivity. As productivity and water use are intimately linked, agricultural and irrigation strategies, have often called for any remaining undeveloped water resources to be directed towards the agricultural sector. The requirements of other water users have not always been considered in these strategies. To support the water demands of this perceived priority user of water, Government Water Departments have placed great emphasis on developing the "supply side" infrastructure, activities associated with dam construction and the construction of transmission networks to irrigation schemes. For many years, the logic of this sequence of activities remained unquestioned. But with increasing support for the principles enunciated by the United Nations Conference on Environment and Development, which calls for greater recognition of equity, stakeholder involvement, the environment, and that water should be treated as an economic good, and with the publication of the report of the World Commission on Dams (2000), the past logic appears increasingly flawed.

Whereas in the past there would have been no doubt that agriculture should be regarded as the priority water user, questions regarding water priorities are now being posed in many countries. In Namibia, for example, although 43% of the country's water is used for irrigation, it contributes only 3% of GDP (Table 2).

The United Nations supported Consultative Group on International Agricultural Research (CGIAR) is now attempting to align strategies for increasing agricultural productivity within an environment of growing scarcity and competition for water through the System-Wide Initiative on Water Management (SWIM). The first paper produced under this initiative by Molden (1997) presents a conceptual framework for considering how water is used and recycled within a basin, and gives a useful definition of water accounting terms. Molden also makes the important point "that the portion of water diverted to an irrigation scheme that is not consumed, is not necessarily lost from

Table 2 Water use by sector and contribution to the Gross Domestic Product (GDP) in Namibia, from ''Sharing water in Southern Africa'', (Pallett, 1997)

Sector	Water used (%)	Contribution to GDP (%)
Irrigation	43.0	3
Cattle	25.3	8
Household + other	25.3	27
Mining	3.2	16
Tourism	0.4	4
Industry + Commerce	2.8	42

a river basin, because much of it may be reused downstream". An associated paper by Seckler (1996) discusses the concept of "dry" and "wet" water savings. Here "wet" savings are regarded as genuinely beneficial savings to the whole system that allow other, possibly downstream, users to make use of the savings. "Dry" savings occur if there are no "downstream" benefits from the savings, which might occur if the basin drained straight to the sea. Any texts such as these are surely valuable if they can introduce the ideas into the agricultural and, particularly, the irrigated agricultural community that there are downstream users that also need water, that allowing water to flow in a river and not use all of it for irrigated agriculture is not necessarily "a waste", and that there may, indeed, be valid environmental or downstream uses for that water.

There is no doubt that with an increasing world population and greater water requirements to produce the world's food needs, increasing efforts will be necessary to manage our land and water systems. To what extent the extra food needs should be met from increased irrigation or for the "horizontal expansion" of agricultural lands into grasslands and forests (Falkenmark, 2003) is the subject of current debates, and is also discussed in.

IMPLEMENTING INTEGRATED LAND AND WATER MANAGEMENT – STAKEHOLDER INVOLVEMENT

Increasingly, integrated land and water resources management is regarded as the philosophy, or process underlying the management of land and water resources. It is recognized that integrated land and water resources management must accommodate means of obtaining progressive commitments from stakeholders for new developments and initiatives.

Participatory Approaches

Principle number 2 of the Dublin Conference Report states:

> Water development and management should be based on a participatory approach, involving users, planners, and policymakers at all levels.

Whilst the participatory approach to water management and development was formalized in the Dublin Statement in 1992, participatory approaches had been advocated, and participatory methodologies had been developed, much earlier by social and management scientists in governmental and nongovernmental development organizations. The World Bank has compiled a sourcebook on participatory approaches that are outlined briefly below. The incorporation of vernacular knowledge in management approaches is also outlined below.

Glossary of Participatory Tools – from the World Bank Sourcebook on Participatory Approaches

- **Access to resources**. A series of participatory exercises that allows development practitioners to collect information and raises awareness among beneficiaries about the ways in which access to resources varies according to gender and other important social variables.
- **Analysis of tasks**. A gender analysis tool that raises community awareness about the distribution of domestic, market, and community activities according to gender, and familiarizes planners with the degree of role flexibility that is associated with different tasks.
- **Focus group meetings**. Relatively low-cost, semistructured, small group (4 to 12 participants plus a facilitator) consultations used to explore people's attitudes, feelings, or preferences, and to build consensus.
- **Force field analysis**. A tool similar to the one called "Story With a Gap," which engages people to define and classify goals and to make sustainable plans by working on thorough "before and after" scenarios.
- **Health-seeking behavior**. A culturally sensitive tool for generation of data about health care and health related activities.
- **Logical Framework or LogFRAME**. A matrix that illustrates a summary of project designs, emphasizing the results that are expected when a project is successfully completed. The logical framework approach to project planning, developed under that name by the US Agency for International Development has been adapted for use in participatory methods such as ZOPP and TeamUP. ZOPP, from the German term "Zielorientierte Projektplanung", is a project planning and management method that encourages participatory planning, and analysis throughout the project cycle. The TeamUP process assists stakeholders in planning and decision making and encourages stakeholders to collaborate as an effective working group.
- **Mapping.** A generic term for gathering in pictorial form, baseline data on a variety of indicators. This is an excellent starting point for participatory work because it gets people involved in creating a visual output that can be used immediately to bridge verbal communication gaps and to generate lively discussion.
- **Needs assessment**. A tool that draws out information about people's varied needs, raises participants' awareness of related issues, and provides a framework for prioritizing needs.
- **Participant observation** is a fieldwork technique used by anthropologists and sociologists to collect qualitative and quantitative data that leads to an in-depth understanding of people's practices, motivations, and attitudes.
- **Pocket charts**. Investigative tools that use pictures as stimuli to encourage people to assess and analyze a

given situation. Through a "voting" process, participants use the chart to draw attention to the complex elements of a development issue in an uncomplicated way.

- **Preference ranking**. Also called *direct matrix ranking*, an exercise in which people identify what they do and do not value about a class of objects. Ranking allows participants to understand the reasons for local preferences and to see how values differ among local groups.
- **Role playing**. Enables people to creatively remove themselves from their usual roles and perspectives to allow them to understand choices and decisions made by other people with other responsibilities.
- **Seasonal diagrams or seasonal calendars**. Show the major changes that affect a household, community, or region within a year, such as those associated with climate, crops, labor availability and demand, livestock, prices, and so on.
- **Secondary data review**. Also called *desk review*, an inexpensive, initial inquiry that provides necessary contextual background. Sources include academic theses and dissertations, annual reports, archival materials, census data, life histories, maps, project documents, and so on.
- **Semistructured interviews**. Also called *conversational interviews*, interviews that are partially structured by a flexible interview guide with a limited number of preset questions.
- **Sociocultural profiles**. Detailed descriptions of the social and cultural dimensions that in combination with technical, economic, and environmental dimensions serve as a basis for design and preparation of policy and project work.
- **Surveys**. A sequence of focused, predetermined questions in a fixed order, often with predetermined, limited options for responses.
- **Tree diagrams**. Multipurpose, visual tools for narrowing and prioritizing problems, objectives, or decisions. Information is organized into a treelike diagram that includes information on the main issue, relevant factors, and influences and outcomes of these factors.
- **Village meetings**. Meetings with many uses in participatory development, including information sharing and group consultation, consensus building, prioritization and sequencing of interventions, and collaborative monitoring and evaluation.
- **Wealth ranking**. Also known as *wellbeing ranking* or *vulnerability analysis*, this technique allows for the rapid collection and analysis of specific data on social stratification at the community level.
- **Workshops**. Structured group meetings at which a variety of key stakeholder groups, whose activities or influence affect a development issue or project, share knowledge, and work toward a common vision.

Indigenous and Vernacular Knowledge

For stakeholder involvement in environmental, land, and water management to be more than tokenism by planning and decision-making authorities, efforts are required to both structure the consultative process and incorporate stakeholder knowledge.

Newson *et al.* (1999) state:

> "Paradoxically, the planning ethos is open to all science, including "vernacular" science or the "common knowledge of ordinary folk". Adaptive planning, like public response to hazards, requires options and experience; whilst the scientist may set up the valid options, it remains the experience of (and "comfort" with) those options by the public which allow on-line adaptation of the plan to occur."

The incorporation of Indigenous Knowledge (IK) in development programs dealing with natural resource management is discussed by Barr (1998). Although the incorporation of IK is now becoming more common and methodological research on the incorporation of IK within natural resources research is now being developed, Barr warns of the difficulties associated with linking overlapping spheres of knowledge between local people at one end of a spectrum, with the applied and basic sciences at the other and with social scientists and anthropologists somewhere in the center. He argues that, in theory, natural resources IK has much to offer in tackling natural resource management problems, but the practical realities of operationalizing this type of interdisciplinary research are far from straightforward. Conflict, or at least disagreement, is recognized as a common feature of interdisciplinary research, especially where the disciplines are "closely guarded cabals".

The key importance of having appropriate and functioning links in place between the institutions involved in land and water management is discussed in **Chapter 194, Inter-Institutional Links in Land and Water Management, Volume 5**.

FURTHER READING

Bruijnzeel L.A. (2001) *Hydrology of Tropical Montane Cloud Forests: A Reassessment*, Land Use and Water Resources Research (LUWRR), `http://www.luwrr.com/`

United Nations Conference on Environment and Development (1992) *Agenda 21 & the UNCED Proceedings*, Proceedings of the UNCED Conference, Rio de Janeiro, Brazil. 1992, UNCED: New York, ISBN 0379103508.

REFERENCES

Adams B., Grimble R., Shearer T.R., Kitching R., Calow R., Chen D.J., Cui X.D. and Yu Z.M. (1994) *Aquifer Overexploitation in the Hangu Region of Tianjin, People's Republic of China*, British Geological Survey: Nottingham.

Arthur R.A.J. (1997) Water without limits. *Water and Environment*, 16–19.

Aylward B., Echeverria J., Fernandez Gonzalez A.F., Porras I., Allen K., and Mejias R. (1998) *Economic Incentives for Watershed Protection: A Case Study of Lake Arenal, Costa Rica*, Final report on a research project under the Program of Collaborative Research of Environment and Development (CREED), IIED: London.

Barr J.J.F. (1998) Use of indigenous knowledge by natural resources scientists: issues in theory and practice, Paper presented at the National Workshop on "The State of Indigenous Knowledge in Bangladesh", Dhaka, 6th-7th May.

Batchelor C.H., Rama Mohan Rao M.S. and Manohar Rao S. (2003) Watershed development: A solution to water shortages in semi-arid India or part of the problem. *Land Use and Water Resources Research*, **3**, 3.1–10, **http://www.luwrr.com**.

Bate R.N. and Dubourg W.R. (1994) *A Netback Analysis of Water Irrigation Demand in East Anglia*, CSERGE Discussion Paper WM94, University College London.

Benson J.F. and Willis K.G. 1991 The demand for forests for recreation. In: *Forestry Expansion: A study of Technical, Economic and Ecological Factors*, Forestry Commission: Edinburgh.

Bhatia R. and Falkenmark M. (1992) Water resource policies and urban poor: innovative thinking and policy imperatives, *Paper Presented to the Dublin International Conference on Water and the Environment*, January 1992.

Bonell M. (1999) Tropical forest hydrology and the role of the UNESCO International Hydrology Programme: some personal observations. Submitted to: *Hydrology and Earth System Sciences*, European Geophysical Society.

Bosch J.M. and Hewlett J.D. (1982) A review of catchment experiments to determine the effects of vegetation changes on water yield and evapotranspiration. *Journal of Hydrology*, **55**, 3–23.

Bruijnzeel, L.A. (1990) *Hydrology of Moist Tropical Forests and Effects of Conversion: A State of Knowledge Review*, UNESCO International Hydrological Programme, A publication of the Humid Tropics Programme, UNESCO: Paris.

Bruijnzeel L.A. (2004) "Hydrological functions of tropical forests: not seeing the soil for the trees?" *Agriculture, Ecosystems and Environments*, **104**(1), 185–228.

Calder I.R. (1992) The hydrological impact of land use change (with special reference to afforestation and deforestation), *Proceedings of the Conference on Priorities for Water Resources Allocation and Management, Southampton, July 1992*, Overseas Development Administration, London, pp. 91–101.

Calder I.R. (1999) *The Blue Revolution, Land Use and Integrated Water Resources Management*, Earthscan: London, ISBN 1 85383 634 6.

Calder, I.R. (2000) Land use impacts on water resources. Background paper 1. FAO Electronic Workshop on Land-Water Linkages in Rural Watersheds. 18 September–27 October 2000. **http://www.fao.org/ag/agl/watershed/**.

Calder I.R. (2002). Forest valuation and water; the need to reconcile public and science perceptions. In: Verweij P.A. (Ed.), *Understanding and Capturing the Multiple Values of Tropical Forests*. Tropenbos Proceedings, Tropenbos International, Wageningen, pp. 49–62.

Calder I.R. (2005) *Blue Revolution II, Integrated Land and Water Resources Management*, Earthscan: London.

Calder I.R. and Gosain A.K. (2003) Inter-relating resource management issues. *Proceedings of the KAWAD Conference on Watershed Development and Sustainable Livelihoods – past Lessons and Future Strategies, 16–17 January*, Karnataka Watershed Development Society: Bangalore.

Carney D. (1998) Implementing the sustainable rural livelihoods approach. In *Sustainable Rural Livelihoods, What Contribution can We Make?*, Papers presented at the Department for International Development's Natural Resources Advisers' Conference, July 1998. IBSN 1 86192 082 2, Carney D. (Ed.), DFID: London.

DFID (1997) *White Paper on International Development*, Department for International Development: London.

Dixon J.A., Scura L.F., Carpenter R. and Sherman P.B. (1994) *Economic Analysis of Environmental Impacts*, Earthscan Publications: London.

Elkington J. (1997) *Cannibals With Forks*, Capstone Publishing: Oxford, p. 402, ISBN 1 900961.

Falkenmark M. (2003) Water cycle and people: water for feeding humanity. *Land Use and Water Resources Research*, **3**, 1–3.4, **http://www.luwrr.com/**.

Gibbons D.C. (1986) The economic value of water. *Resources for the Future*, Environmental Economics Series 027. The World Bank, Washington, DC.

GWP Technical Advisory Committee. (2000) **http://www.gwpforum.org/**.

Hamilton L.S. (1987) Tropical watershed forestry – aiming for greater accuracy. *Ambio*, **16**, 372–373.

Hamilton L.S. and King P.N. (1983) *Tropical Forested Watersheds. Hydrologic and Soils Response to Major Uses or Conversions*, Westview Press: Boulder.

Hardin G. (1968) The tragedy of the commons. *Science*, **162**, 1243–1248.

Hassan R., Berns J., Chapman A., Smith R., Scott D. and Ntsaba M. (1995) *Economic Policies and the Environment in South Africa: the Case of Water Resources in Mpumalanga*, Division of Forest Science & Technology, CSIR: Pretoria.

IUCN. (1998) *The Green Accounting Initiative* **http://iucn.org/places/usa/gai-activitiespage.html**.

KAWAD (2001) *A Fine Balance: Managing Karnataka's Scarce Water Resources*, Karnataka Watershed Development Society: Bangalore.

Kutcher G., McGurk S. and Gunaratnam J. (1992) *China: Yellow River Basin*, Water investment planning study presented at a World Bank Irrigation and Drainage Seminar: Winpenny.

Landefeld J. and Carson C.S. (1994) Integrated Economic and Environmental Satellite Accounts. *Survey of Current Business*, April, pp. 33–49.

Meadows D.H., Meadows D.L., Randers J. and Behrens W.W. III (1972) *The Limits to Growth; a Report for the Club of Rome's Project on the Predicament of Mankind*, Universe Books: New York, p. 205, ISBN 0 87663 165 0.

Molden D. (1997) *Accounting for Water use and Productivity*, SWIM Paper 1. Colombo: International Irrigation Management Institute (IIMI), ISBN: 92 9090 349 X.

Newson M.D., Gardiner J. and Slater S. (1999) River catchment planning. In *Changing Hydrology of the UK Routledge*, Acreman M. (Ed.), Routledge, pp. 315–344.

Pallett J. (Ed.), (1997) *Sharing Water in Southern Africa*, Desert Research Foundation of Namibia: Windhoek.

Rees J.A., Williams S., Atkins J.P., Hammond C.J. and Trotter S.D. (1993) *Economics of Water Resource Management*, R&D Note 128, National Rivers Authority: Bristol.

Saberwal V.K. (1997) Science and the desiccationist discourse of the 20th century. *Environment and History*, **3**, 309–343.

Scoones I. (1998) *Sustainable Rural Livelihoods: A Framework for Analysis*, Institute of Development Studies: Brighton, Working Paper No. 72.

Seckler D. (1996) *The new era of water resources management*, research report 1, International Irrigation Management Institute, Colombo.

Shiklomanov I.A. (1999) Climate change hydrology and water resources: the work of the IPCC, 1988–1994, In: *Impacts of Climate Change and Climate Variability on Hydrological Regimes*, Van Dam J.C. (Ed.), Cambridge University Press-UNESCO International Hydrology Series: pp. 8–20.

Willis K.G. and Benson J.F. (1989) Recreational values of forests. *Forestry*, **62**(2), 93–110.

Winpenny J.T. (1992) Water as an economic resource Paper 4. *Proceedings of the Conference on Priorities for Water Resources Allocation and Management. Southampton, July 1992*, Overseas Development Administration: London, pp. 35–41, ISBN: 090 2500 49X.

Winpenny J.T. (1996) The value of water valuation. In *Water Policy: Allocation and Management in Practice, Proceeding of the International Conference on Water Policy*, Howsam P. and Carter R. (Eds.). Cranfield University, E & FN Spon, London, pp. 197–204.

World Commission on Environment and Development. (1987), (The Brundtland-Report) Our common future, Oxford University Press.

World Commission on Dams. (2000) Dams and Development: A New Framework for Decision-Making, Earthscan, London, ISBN 1 85383 797 0.

186: Water and Forests

GRAHAM JEWITT

School of Bioresources Engineering and Environmental Hydrology, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Years of research in various parts of the world has highlighted the interaction between forests and water. Forests impact on the hydrological cycle affecting processes such as interception, transpiration, infiltration, groundwater recharge, and runoff. These inevitably manifest themselves as impacts on the water resource, both positive and negative. In this article, the role of forests in the hydrological cycle is described, and the potential impacts for water resources are considered.

INTRODUCTION

Approximately 70% of the Earth is covered by water and of the remaining land surface, over 30% by forest. Humans, forests, and water have a relationship that dates back for millennia. Throughout these times, both forest and water resources have been utilized by humans, often with a focus on one to the detriment of the other. Forests produce many goods and services that are utilized by humans. These include "traditional" commercial functions such as production of wood and other raw materials, as well as more difficult-to-value functions such as "biodiversity" that are fundamental to the to ecosystem integrity. Calder (1996) suggested that on a global scale, the largest land-cover change in terms of area, and "arguably also in terms of hydrologic effects, is from deforestation and afforestation".

It is well accepted that forests influence the hydrological cycle affecting both the quantity and quality of a catchment's water resources. The degree to which forests influence the hydrological cycle, the need to supply water resources to fulfill environmental and societal needs, and the associated management issues for both resources has evoked much controversy. The removal of forests for agricultural and industrial purposes has been blamed for both flood and drought, whilst in other regions, the establishment of commercial forestry has been blamed for reducing flow in rivers, and its establishment is controlled by water-specific legislation. Perhaps, because of the historical relationship between trees, forests, and humans, strong perceptions, both public and scientific exist (*see* **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**).

In essence, two broad schools of thought developed around the relationships between forests and water. A notion that has become entrenched in some quarters is that the complex of forest soil, roots, and litter acts as a "sponge" soaking up water during rainy spells and releasing it evenly during dry periods. The proponents emphasize that forests are "good" for water supplies and reinforce perceptions that forests "improve" water resources by recharging groundwater, maintaining baseflows, improving water quality, and moderating floods. Despite insufficient experimental data to support these perceptions, and the publication of many findings to the contrary, these are entrenched in policies and operational frameworks of many development agencies, foresters, and water managers.

The contradictory view is probably most effectively stated by Hamilton and King (1983) who suggested that "roots" may be more appropriately labelled a pump rather than a "sponge" and that "roots certainly do not release water in the dry season but rather remove it from the soil in order that the trees may transpire and grow". They added that "major floods occur because too much rain falls in too short a time, or over too long a time. In either case, the rainfall exceeds the capacity of the soil mantle to store it and the stream channel to convey it". In effect, they provided a thorough synthesis of issues that had been debated for over a 100 years, but were pioneers in an era of important advances in forest hydrology that have occurred over the past 25 years and which have improved

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

the state of hydrological knowledge in the field. However, contradictory results and difficulties in the measurement of some variables mean that a conclusive interpretation of how forests influence water resources is not always possible.

The strong perceptions and misconceptions that exist are symptomatic of problems in natural resource management, where the use of scientific information can degenerate into simple application of generalized values and a widespread belief is allowed to pass as fact. The result is that many aspects of land cover and water resource decision making are based more on perceived wisdom (myth) than scientifically established reality (Calder, 1999). These generalizations are seldom explicitly tested and are often applied inappropriately to address specific problems at local scales over short time periods. The consequence is that a body of "pseudo-fact" is generated in a self-reinforcing mode that seldom challenges theory, and eventually forms a barrier to effective, scientifically based management and perpetuates inappropriate paradigms or myths (Gunderson *et al.*, 1995). The failure of many individuals and organizations to understand the relationships between forest and water has led to the "misuse of both money and land" (Cossalter and Pye-Smith, 2003).

The subdiscipline of "forest hydrology" considers the scientific basis for these perceptions, whilst many of these misperceptions are explored in **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**.

Forests Around the World

A recent United Nations Environment Programme (UNEP) estimates that worldwide, there is approximately 40 million km^2 of forest, with just over 3 million km^2 in International Union for Conservation (IUCN) recognized areas (Table 1). This encompasses 25 different forest types, divided into tropical and nontropical, and includes both indigenous and exotic forest plantations, as well as "sparse trees and parkland", in which canopy cover ranges from only 10–30%.

Forest Plantations

There is an ongoing discussion worldwide of the potential environmental impacts of the large-scale establishment of forest plantations. The growing of trees to provide for industrial needs, in particular, pulp and fiber for paper production and charcoal for steel production, has become a major business over the past 30 years. The expansion of so-called fast-wood plantations has been particularly rapid in a few key countries in both the developed and the developing world (FAO, 2003; Cossalter and Pye-Smith, 2003). Fast-wood plantations are defined as intensively managed commercial plantations, set in blocks of a single species that produce industrial round wood at high growth rates (mean annual increment of no less than 15 m^3 per hectare), and that are harvested in less than 20 years (Cossalter and Pye-Smith, 2003). The rapid growth rates of such plantations imply high water use by them, and the typically short (6–12 year) crop rotations are intended to optimize this.

Table 1 Estimated extent (1996) of the global forest area (`http://www.unep-dwcmc.org/forest/homepage.htm`)

Region	1996 Forest area (km^2)	1996 Protected forest area (km^2)
Africa	5 683 130	496 927
Australasia	1 493 234	125 619
Caribbean	53 847	7899
Central America	901 984	88 096
Continental S and SE Asia	1 707 679	192 461
Europe	1815396	144 832
Far East	1 456 027	77 401
SE Asia	1 468 360	247 497
Middle East	167 661	6386
North America	8 453 988	699 956
Russia	8 257 159	150 637
South America	8 429 459	874 924
World	39 887 924	3 112 835

It is estimated that forest plantations cover approximately 10 million ha worldwide and that this area is growing at a rate in the order of 1 million ha per year (UNEP, 1996; Cossalter and Pye-Smith, 2003).

Invasive Alien Plants

Invasive Alien Plants (IAPs) have been identified as a major biodiversity and water resources problem worldwide. In the case of stands of IAPs made up of tree species, these may have similar characteristics to forests, and particularly in semiarid regions they have often been shown to use more water than the land cover they invade (Calder and Dye, 2001). For example, it has been suggested that IAPs are causing the loss of almost 7% of the annual flow in South Africa's rivers each year (Versfeld *et al.*, 1998). Water resources issues are especially highlighted in semiarid regions and much of the focus in this regard has been on IAPS growing in riparian zones, where it is believed that plants transpire more water than their counterparts elsewhere in the catchment (Mack and d'Antonio, 1998; Dye *et al.*, 2000), especially where dense stands develop (Sala *et al.*, 1996). This has reportedly led to reduction of flow levels in places by up to 50% (Higgins and Richardson, 1996; Mack and d'Antonio, 1998).

A Brief History of Forest Hydrology

Although most discussions around the topic of forests and water credit the work of European and American researchers with the first comprehensive studies of the

effects of forests on the water cycle, the history of forest hydrological research in China can be traced back to the mid-period of the Ming Dynasty (1368–1644)(Xinxiao, 1991). The earliest forest hydrology studies in Europe are reported to have been undertaken in Switzerland, reported by Engler (1919), and cited by McCulloch and Robinson, 1993). A major work by Zon (1927), in which the available literature on forest influences was summarized, preceded a period of "unparalleled research" (Lee, 1980; p. 16) not only in Europe and America, but also in East Africa (Blackie *et al.*, 1979), South Africa (Nänni, 1970; Scott *et al.*, 2000), and Australia (Best *et al.*, 2003). Kittredge, (1948) published his highly influential book "Forest Influences" that became a major reference work. Bruijnzeel (2004) reports that a heated debate on the hydrological role of forests took place in the 1930s and 1940s in the forestry journal of the former Dutch East Indies (Tectona). In the past 40 years, major text book contributions to the subject have been made by Hewlett and Nutter (1969), which were later revised by Hewlett (1982) and Lee (1980) and more recently by Chang (2002), as well as chapters and sections in the myriad hydrological text books that exist. However, the work by Hewlett – Principles of Forest Hydrology (Hewlett, 1982) probably remains the most prescribed forest hydrology text, and the review of forested catchment experiments by Bosch and Hewlett (1982) remains one of the most cited hydrological publications. Forest hydrology studies have been at the forefront of development of many hydrological theories. Perhaps, the most well known of these is the "variable source area" concept of runoff generation (Hewlett and Hibbert, 1967) proposed in 1965 at the International Forest Hydrology Symposium held at Penn State University.

Prior to 1980, the vast majority of forest hydrology studies were focused on water balance approaches such as paired catchment or split sample experiments. These studies were inevitably "input–output" type studies, with the result that critical processes such as transpiration and soil-water uptake had to be inferred, rather than measured. Arguably, the most significant advance in the studies of forest hydrology over the past 25 years have been allied with the advances in computing power and microcomputing techniques, which have provided the opportunities for the development of techniques and methodologies which allow the study of these processes representing internal catchment storages and fluxes and a more sophisticated study of "whole tree" water use. By 1993, at the "Water Issues in Forests Today" symposium held in Canberra, Australia, it was felt that the need for such process studies had been widely accepted as a new paradigm in forest hydrology (O'Loughlin and Dunin, 1993), and recent literature shows that researchers have undertaken detailed hydrological process studies in both forested and nonforested catchments as part of this new paradigm. Energy balance approaches using Bowen Ratio and Eddy Covariance, and more recently scintillation techniques has allowed more spatially and temporally detailed measurements of total evaporation, whilst sophisticated heat pulse and isotope studies have allowed for direct estimation of the transpiration component of the hydrological cycle. Advances with soil-water measurement instruments, such as neutron moisture meter and time domain reflectrometry have informed scientists about root uptake patterns and groundwater recharge beneath forests. Furthermore, advances in tracer studies are providing the impetus for the development on new theories of soil water and runoff generation (Bonell, 1998), whilst advances in remote sensing techniques provide opportunities for large-scale estimates of total evaporation.

Some argue that these process studies have been undertaken at a vast expense, resulting in many publications, but have done little to enhance the applicability of the science in a water resources management context and they at least need revisiting to address issues that are now relevant (e.g. DeWalle, 2003). In particular, the complexities of scaling results from detailed hydrological process studies at an experimental site to the catchment have slowed some of the expected progress in this regard. However, there have been notable cases where such process studies have had a direct influence on water resources planning decisions as highlighted in Section "Forests and water resources" below.

THE IMPACT OF FORESTS ON THE HYDROLOGICAL CYCLE

The influences of land cover on the hydrological cycle are well known (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**–**Chapter 120, Land Use and Land Cover Effects on Runoff Processes: Fire, Volume 3**). Forests, as a tall and often evergreen land cover, typically with deep root systems, litter layers, and secondary canopies occupy a particular niche in studies of land cover on hydrological functioning. The subdiscipline of forest hydrology has as its focus the way in which forests influence the hydrological cycle. As highlighted above, "modern" scientific understanding on forest hydrology has evolved from almost 100 years of research from forests based on paired catchment, water balance, and more recently detailed hydrological process studies. The vast majority of these studies indicate decreased runoff from areas under forests as compared with areas under shorter land covers. Furthermore, research has highlighted the potential for differences in water use between different forest and tree types, and the rapid increase in the area of forest plantations since the 1960s has highlighted the importance of different growth stages in tree water use.

Partitioning of Precipitation in Forests

Figure 1 provides a forest hydrology perspective of the hydrological cycle. In particular, the key water partitioning points of the hydrological cycle are highlighted, that is, the points where land cover in general, and forest in particular, are likely to impact on the hydrological processes that affect the way in which incoming precipitation is partitioned:

1. At the canopy level, incoming precipitation is partitioned into vertically orientated fluxes, that is, upward represented by evaporating water (interception loss and transpiration from the stomata) and downward represented by throughfall, stemflow, and canopy drip ultimately forming net precipitation once passing through the litter layer.
2. At the soil surface, net precipitation is partitioned horizontally into runoff and vertically into infiltration, as well as water vapor through direct evaporation from the soil or litter layer.
3. At the root zone, upward water fluxes are generated, that is, direct evaporation from the soil, but more importantly uptake of water by the root system for transpiration. In this zone, water is also partitioned both into downward percolation, ultimately providing groundwater recharge, and horizontally into "interflow", provided by unsaturated flow that moves downslope to eventually form runoff.

A perspective of the hydrological cycle that is becoming increasingly popular when considering water in the context of food and fiber production classifies water vapor and liquid water as green water and blue water respectively (Falkenmark *et al.*, 1999). Blue water is the runoff originating from the partitioning of precipitation at the plant canopy and land surface (eventually reaching the river network and forming streamflow), and the partitioning of soil water (forming baseflow and groundwater recharge). Green water is water vapor and is represented by the flow of water to the atmosphere as evaporation from soil, lakes, and water intercepted by canopy surfaces and transpiration by vegetation i.e. Total Evaporation (ET). As highlighted in Figure 2, which illustrates the expected differences in ET from grassland and forested catchments on the basis of equations derived from over 250 experimental catchments worldwide, in most forested catchments, green water flow greatly exceeds blue water flow (Zhang, 1999).

To the landowner, the benefits most commonly associated with green water use are the production of biomass resulting from the movement of water through agricultural crops and timber by transpiration. To others, forests offer attractive recreational areas and secondary products such as firewood or honey production. Such benefits can be considered "ecosystem goods and services", which foresters in particular, and society in general, derive from the land that is utilized (Jewitt, 2002a). The role of land cover in

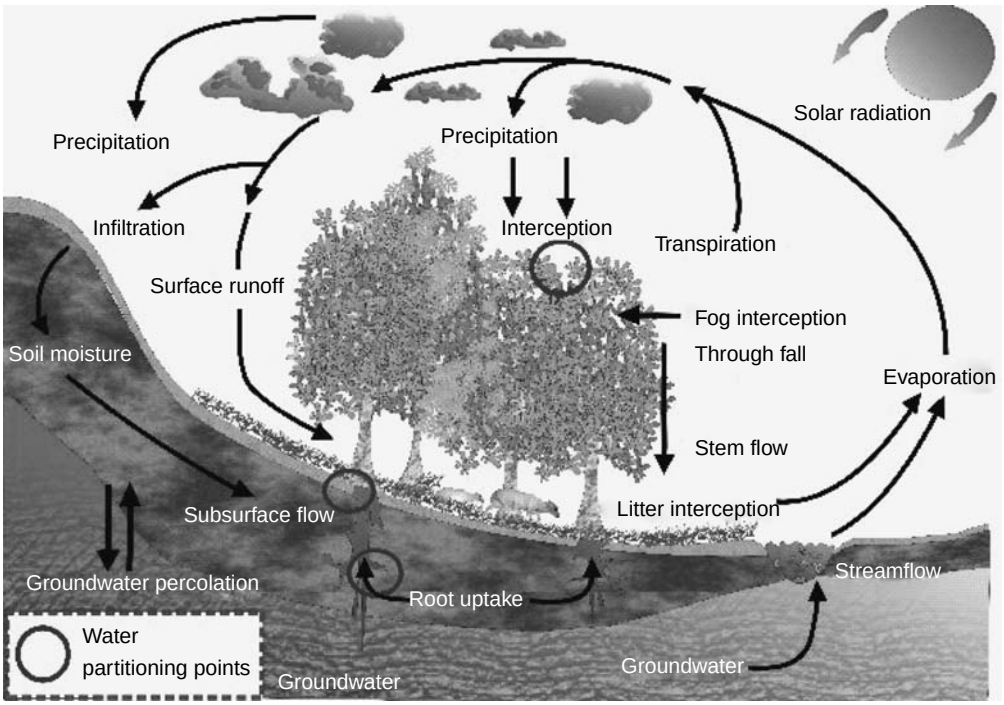


Figure 1 The key water partitioning points in the forest hydrology cycle. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

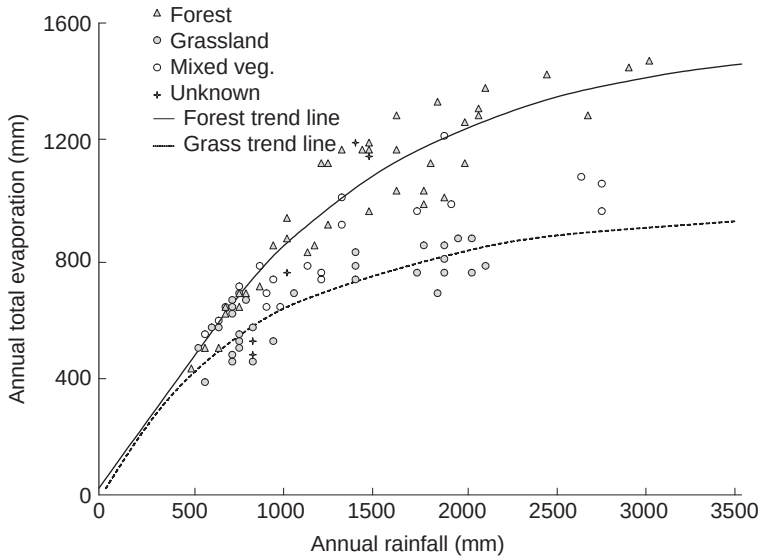


Figure 2 Relationship between annual total evaporation and rainfall for different vegetation (After Zhang, 1999, Reproduced by permission of the Cooperative Research Centre for Catchment Hydrology, CSIRO, Canberra, Australia). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the catchment is of critical importance in this regard, as it partitions rainfall between water vapor flows to the atmosphere as evaporation and transpiration (green water), and the flow of water to rivers and groundwater (blue water). In arid and semiarid areas, runoff to rivers (i.e. blue water production) is generally less than 10%. Thus, a small change in the partitioning of rainfall by land cover, which is typically the case when forest plantations are established, can have a relatively large impact on runoff.

The Soil–Vegetation–Atmosphere Continuum

Forests and water cannot be considered in isolation, as the influences of forestry on water are largely dependent on the soil and climate conditions as well as the physiological characteristics of the tree itself. Although it may be useful to isolate different components of forest water use, it must always be recognized that these processes are part of a continuum. The fundamental importance of the soil–vegetation–atmosphere relationship and the hydrological role thereof has been increasingly highlighted in recent years, as process studies in forests have gained impetus, leading some to state that "Climate-Soil-Vegetation dynamics is the core of hydrology" (Rodriguez-Iturbe, 2000). Tree water use is driven by above (leaf area and pattern) and below ground (rooting depth and pattern) structure, and the rate at which water is used (interception and transpiration) that is further controlled by atmospheric and soil conditions. The role of soil properties in controlling infiltration, root water uptake, and evaporation from the soil as well as percolation of water through the soil and groundwater recharge coupled with an increase in studies of "whole tree water use" and water use efficiency is probably the most topical research aspect in forest hydrology today.

The main reasons for water use differences between forests and other land covers and different tree genera and species are related to physiological differences, that is, canopy and root structure. In assessing water use by different forest genera and species, a number of general points emerge:

- Different tree types have varying rates of transpiration under conditions of adequate available soil water (Bosch and Smith, 1989; Wullschleger *et al.*, 1998; Landsberg, 1999).
- Some trees maintain near-optimal rates of transpiration, even at relatively low soil-water content (Russel, 1973; Pryor, 1976; Landsberg, 1999). Others, however, are generally more conservative water users and reduce transpiration rates at higher soil-water contents and different atmospheric demands than other species.
- The distribution of roots differs with both age and tree type (Sherry, 1971) and in the case of man-made forestry plantations, will be affected by the method of site preparation used (Boden, 1984).
- The density of the trees' canopy will affect the amount of rainfall intercepted at the canopy level (Crockford and Richardson, 1990a).
- The thickness and nature of the litter layer on the forest floor will differ according to climatic conditions as well as tree type and age. The older the forest, the greater the build up of litter, and the higher the litter interception is likely to be.

Total Evaporation

Total Evaporation (ET) can be considered the consumptive use of water by a forest. It consists of the transpiration of water by plants and the evaporation of water from the soil and of the rain water intercepted by the forest canopy. ET depends on both climatic factors such as solar radiation, temperature, air humidity, and wind velocity as well as plant physiology factors, such as leaf characteristics (leaf area, evergreen versus deciduous), albedo, tree canopy, and rooting characteristics. Current scientific understanding indicates that in both wet and dry climates, evaporation from forests is likely to be higher than that from shorter crops (Calder, 1998) Trees form aerodynamically rough surfaces to the wind providing a high degree of turbulent mixing within the forest canopy. This results in the maintenance of high vapor pressure deficits and correspondingly high rates of transport of water vapor from the leaf surface to the atmosphere, and may result in evaporation from the canopy at rates in excess of reference potential evaporation during wet conditions (Calder, 1990). In the dry season and periods of drought, several studies have highlighted the fact that transpiration from forests is likely to be greater than other catchment land covers (Calder, 1990; Dye, 1996a), because of the generally greater depth and extent of their roots compared to other shorter land covers and the resulting greater access to soil water. Furthermore, litter is an important factor when considering the water balance of a forest, since it will both prevent a certain fraction of rainfall from entering the soil by intercepting it, and simultaneously will reduce soil-water evaporation losses and prevent overland flow.

Because of their size, longevity, and the deep rooting nature of trees and the fact that many forests grow in inaccessible areas, direct estimation of total evaporation from forests is particularly difficult. Typically, ET is either determined directly on the basis of microclimatic studies above the canopy, or indirectly by solving the water balance at tree, plot, or catchment scale. Conversely, in many studies, catchment runoff is calculated as the difference between precipitation and ET, or measured at the level of the streams coming from the upstream catchment area. Calder (1998) has suggested a "limits approach" to focus attention on the different processes that are hypothesized to control evaporation in temperate and tropical climates and wet and dry areas within these, and thus aid selection of which measurements are undertaken or which models should be used to estimate these. These limits are presented in Table 2.

Table 2 Principal limits and controls on evaporation for different land uses in different climates (After Calder, 1998)

Principal limits on evaporation		
	Temperate climate	
Land use	Dry	Wet
Tall crop	Physiology Soil moisture	Advection
Short crop	Soil moisture Radiation	Radiation Physiology
	Tropical climate	
Land use	Dry	Wet
Tall Crop	Soil moisture Tree size	Raindrop size Physiology
Short crop	Soil moisture	Radiation

Interception

When the forest canopy is closed, the interception of precipitation by forests is a major component of the influence that forests exert on the hydrological cycle. Interception by forest canopies has been considered to be one of the most thoroughly researched topics in hydrologic literature (Gray, 1970; Crockford and Richardson, 1990a). A detailed description of the interception process is found in **Chapter 43, Evaporation of Intercepted Rainfall, Volume 1**.

Interception loss from forests depends upon the atmospheric condition driving the evaporation process and rainfall characteristics, but also on the density and nature of the forest stand. Usually, depending upon management practices, older trees have denser canopies, correspondingly higher canopy storage capacities, and higher interception loss from younger stands. It has also been noted that water droplet size affects the rate at which canopy storage capacity is reached. Thus, both raindrop size as well as canopy characteristics influence the drop size as it passes through the forest. In the case of forests with low leaf area, interception characteristics may be closely related to the drop size of the incident rain (Calder, 1996).

Significant differences in interception between evergreen and deciduous trees can be expected. For example, mature pines have a very dense canopy with high leaf area, are usually evergreen, have high surface tension between needles, and usually occur in temperate regions where rainfall of low intensity and small drop size is common and have been shown to have high values of interception loss (Crockford and Richardson, 1990b). Interception loss is estimated as 25–50% in conifer stands and 15–30% in deciduous forests. In semiarid regions and in forests of the tropical moist zone where high intensity convective storms are typical, the percentages of interception are variable, but are usually lower than in the temperate regions. In semiarid Mpumalanga, South Africa, Dye (1996a) believed that this reflects the less frequent, more intense rainfall characteristic of summer rainfall regions and concluded that transpiration, rather than interception from dry canopies was the dominant evaporation process in South African plantations. This conclusion can be extended to presume that the limit on ET in this situation is available soil moisture (Table 2).

Litter Interception Most forest stands have a forest floor of accumulated debris through which rainfall must pass before it enters the soil. Litter is usually regarded as beneficial to the forest system. Litter may play an important role in increasing infiltration rates in forest soils and may also protect the soil from temperature and moisture extremes as well as providing protection from erosional forces such as raindrop impact (Metz, 1958; Schutz, 1990). Litter also constitutes a critical source in the nutrient cycle and is a key to maintaining productivity in plantation systems. In many forests, but in particular in tropical and subtropical forests, tree roots use water and nutrients from the lower levels of the litter layer and as the forest stand matures, the importance of the litter layer in the nutrient cycle increases until it may eventually replace the mineral soil as the principal source of nutrients (Schutz, 1990).

As rainfall passes through the litter layer, some rainfall is retained and part later returned to the atmosphere by evaporation. The litter layer may have important effects on surface runoff, sediment yield and, as mentioned previously, infiltration into and evaporation from the soil. Researchers have shown that moisture intercepted by litter is governed primarily by moisture holding capacity and initial storage capacity of the litter, and by the evaporative demand following the rainfall event (Rowe, 1955; Helvey and Patric, 1965). Litter interception values range from 2% of mean annual precipitation (MAP) in temperate deciduous forests (Blow, 1955) to 4% in mature pine forests (Rowe, 1955; Bernard, 1963; Helvey, 1967). Pathak *et al.* (1985) estimated litter interception to be approximately 12% of MAP in a Himalayan forest.

Transpiration

The rate at which forests transpire is a function of the available radiant energy, atmospheric demand, windspeed, the nature of the rooting system, and the amount of water available in the soil as well as resistance imposed by the vegetation through stomatal conductance and the surface area of the leaves exposed to the atmosphere (Landsberg, 1999). In an undisturbed landscape, it is likely that the structure and leaf area of the vegetation will tend towards being optimal to utilize available water (Specht, 1972).

Essentially, transpiration is the process whereby water is evaporated through the stomata on the leaves of plants (*see* **Chapter 104, Satellite-Based Analysis of Ecological Controls for Land-Surface Evaporation Resistance, Volume 3**). Transpiration is both a result of, and a fundamental component of the photosynthesis process. In order to assimilate carbon, plants need to open their stomata. As the internal cells of the stomata are coated with water, the relative humidity within the stomata is assumed to be high, resulting in a vapor pressure deficit between the interior of the leaf and the surrounding atmosphere and thus, the movement of water from the leaf to the atmosphere. Thus, many of the tree and climatic characteristics that affect evaporation of water intercepted by the tree canopy also affect the transpiration process. Forest species vary considerably in physiological characteristics such as height, leaf area, stomatal resistance, whether deciduous or not, with the result that the amount of energy available to drive the evaporation and transpiration processes varies considerably between tree types. In this regard, researchers have highlighted the strong relationship that exists between leaf area and transpiration rates (Calder, 1998; Landsberg, 1999). Furthermore, the amount of water that is available for transpiration depends on the texture and depth of the soil and the ability of the plant to extract it. The drier the soil, the greater the tension between soil and water particles and the more difficult it is for trees to extract water resulting in reduced transpiration rates. The ability of plants to extract water from the soil varies according to root depth and pattern, as well as other physiological aspects, with the result that different tree types reduce their transpiration rates at different soil-water tensions (Russel, 1973; Pryor, 1976). Furthermore, the depth to which water is extracted corresponds to the rooting depth of the tree. In many circumstances, trees with deep roots can extract water that is unavailable to other plants. Initially, water is almost exclusively extracted from the surface layers; as the soil dries out, water is extracted from increasingly deeper levels of the soil profile (Richards and Caldwell, 1987).

Rates of transpiration from plants and the tree physiological factors that affect them are much debated. Coniferous vegetation has also been reported to close stomata when the surrounding atmosphere is dry, even if soil moisture is freely available. However, in the case of fast-growing eucalypt species, stomatal conductance and transpiration have been shown to be most strongly related to predawn leaf water potential and, thus, to soil moisture content (Mielke *et al.*, 1999; White *et al.*, 2000). Experiments and measurements based on porometer measurements of leaf conductance and rates of soil moisture depletion have highlighted differences in transpiration rates between species (Chang, 2002). In South Africa, Dye (Pers. Comm., 2003) have evidence that slow growing indigenous forest trees of the Podocarpus genus have much lower transpiration rates than commercially grown plantation species such as *Eucalyptis grandis*. However, although there may not be consensus regarding differences in stomatal conductance in trees, it is accepted that leaf area varies considerably and that these differences are likely to be more important than differences in stomatal responses (Landsberg, 1999). Periodic soil-water stress is also known to cause long-term adaptive reductions in leaf area, stomatal conductance, and sapwood structure and these may limit the rate of water use even in seasons when soil water is plentiful (Eamus *et al.*, 2000).

A phenomenon that has received some attention in recent years is that of "hydraulic lift", which is described as the movement of water from lower and wetter soil layers to upper drier layers. This results from the flow of water extracted from the deeper soil layers by the "trees" tap root, down the "trees" lateral roots, and extruded into the upper soil layers during the night when transpiration has ceased (Richards and Caldwell, 1987). This phenomenon is not typical of all tree types, but where it does occur, transpiration rates may increase as water becomes more readily available to the upper layers of the soil.

Infiltration, Drainage, and Recharge

Typically, forest soils have a high infiltration capacity. This is due both to the presence of above ground vegetation and litter that protect the soil from compaction caused by raindrop impact and because of the well-structured porous and highly permeable soils that are found to exist in forested areas. It is the forest floor rather than the canopy that is most important in controlling infiltration rates in forest stands (Hornberger *et al.*, 1997). However, it is also true that the forest canopy and litter reduce the mechanical effect of water drops that in turn protects and maintains the surface soil structure. High proportions of stemflow are common in many forest tree types and associated with this are high infiltration rates of the water flowing down the tree trunk to the soil surface.

In forest plantations, certain site preparation techniques, for example, deep soil ripping or the construction of terraces can substantially increase infiltration. However, the infiltration capacity can also be reduced by management practices associated with logging activities or fire.

Movement of Water in the Soil

The downslope movement of water in the unsaturated zone (often called *interflow*) is believed to play a very important role in the generation of both low flows and high flows (Rice and Hornberger, 1998; Lorentz *et al.*, 2003). It is clear from isotopic signatures that substantial portions of the storm hydrograph consist of pre-event water (Bonell, 1998). In addition, forest soils are often characterized by high levels of macropore development (Bonell, 1993; Lorentz *et al.*, 2003) and it is generally accepted among forest hydrologists that preferential flow within these macropores is a significant component of the hydrological functioning in forested catchments. Macropores form pathways where roots have rotted, or where cracks in the soil or animal burrows exist, and can form extensive networks, with the result that subsurface flow in these situations is transported in saturated form within these areas of the unsaturated soil matrix (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**–**Chapter 191, Environmental Flows: Managing Hydrological Environments, Volume 5**). Johnston (1987) reports that recharge of groundwater can also be positively affected by such preferential flow and provides evidence of recharge rates that increase from a range of 2.2–7.2 mm $year^{-1}$ to 50–100 mm $year^{-1}$.

Drainage and Groundwater Recharge

It is difficult to define the groundwater recharge component of the hydrological cycle, and often this perspective depends upon the scale at which associated processes are observed (Lorentz *et al.*, 2003). There has been much research on the hydrological functioning of the unsaturated zone, both in humid and arid regions. Important differences between them, such as the fact that the unsaturated zone is usually deeper in arid and semiarid areas lead to different interactions with land cover. Forests influence groundwater levels by direct abstraction of water from the phreatic surface where this is accessible to tree roots, but perhaps more importantly by impacting the rate of recharge of percolating water to the groundwater table, as tree roots abstract water from the unsaturated zone during the transpiration process. Whilst these two processes are probably most significant, the interaction of forests and groundwater is highly complex and depends upon the nature and type of vegetation, as well the position of the landscape as illustrated in Figure 3.

Forests are widely reported to have deep roots that can access groundwater directly. However, because hydrological studies tend to focus on the upper portion of the soil profile, there are very few studies that substantiate this perspective. Consequently, some authors believe that the ability of trees to dry out soil profiles to great depths or access groundwater directly is "not generally appreciated" (Le Maitre *et al.*, 1999).

FORESTS AND WATER RESOURCES

The processes associated with the key precipitation partitioning points highlighted above inevitably manifest themselves as impacts on the water resource. It is now generally accepted that forest cover has a negative impact on total water yield from a catchment. However, the impact of forests on low flows and on floods is highly controversial as has been highlighted by several authors (e.g. McCulloch and Robinson, 1993; DeWalle, 2003; Robinson *et al.*, 2003). The following sections serve to highlight the water resources impacts and benefits associated with forests.

Runoff and Catchment Yield

Many years of research in forested catchments worldwide has highlighted the major role that vegetative cover has on the total evaporation component of the hydrological cycle and the resulting catchment runoff. In some cases, this has resulted in the development of government policy specific to forests. For example, in South Africa, concern

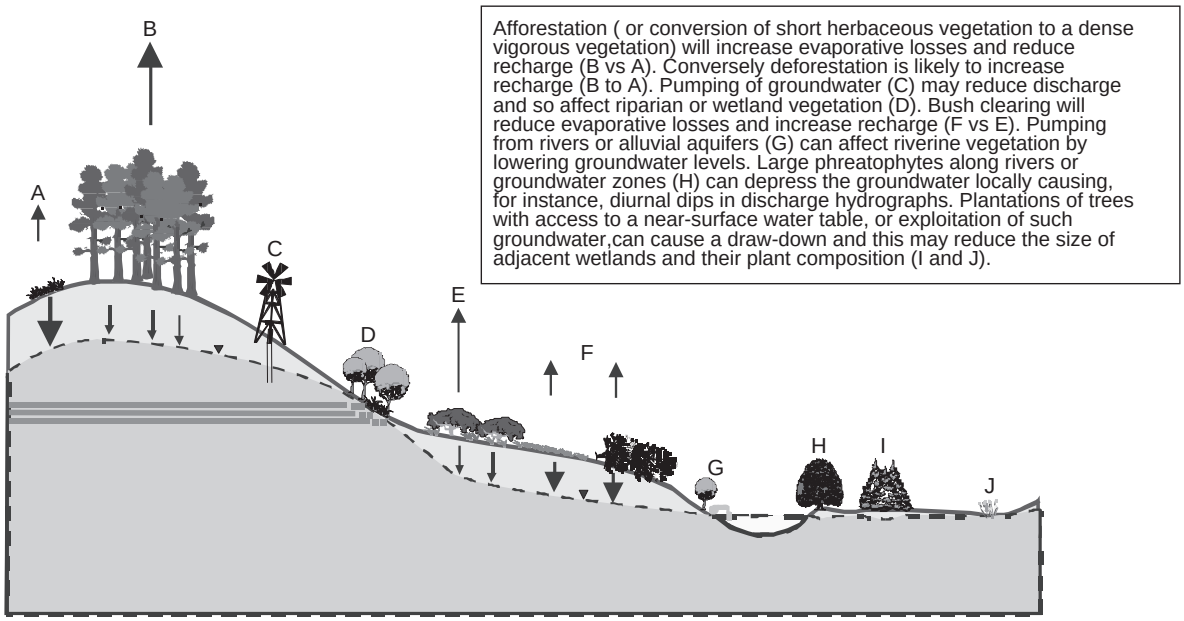


Figure 3 Some typical interactions between vegetation and groundwater (Drawing on ideas developed by Le Maitre *et al.,* 1999). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

regarding the impact of forest on available water resources has led to legislation controlling the establishment of commercial forest plantations since 1972. Under a new National Water Act promulgated in 1998, forest plantations have been declared "Stream Flow Reduction Activities", and as such landowners are required to apply for a water use licence before planting plantation forests. This legislation has been informed by many years of paired catchment and forest hydrology process studies, all of which have highlighted a reduction in runoff from areas converted to forest plantations, a summary of which is shown in Figure 4.

Peel *et al.* (2001) have highlighted the role that changes in ET may play in affecting variability in runoff. On the basis of an analysis of several catchments worldwide, they concluded that the variability of annual runoff is 1–9% higher for catchments with evergreen, as opposed to deciduous cover and presented the equation (for annual

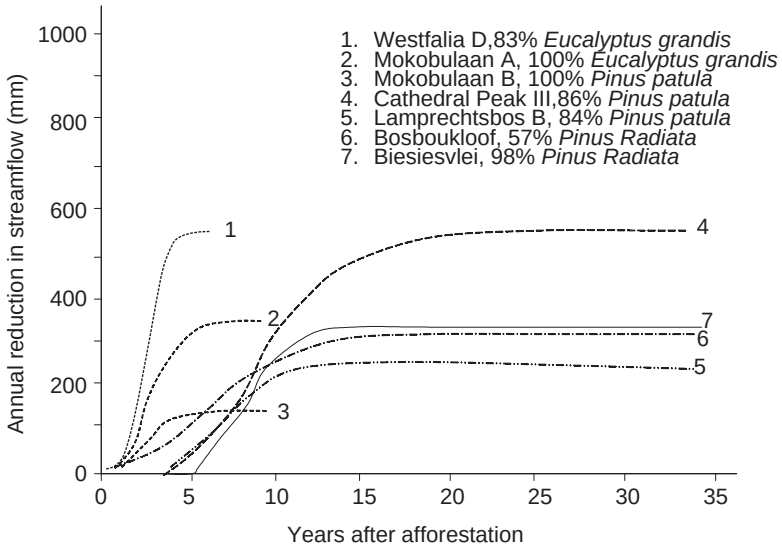


Figure 4 Streamflow reduction measured from six different paired catchment studies in South Africa with different extents of forest (Dye and Bosch, 2000, © South African Institute of Forestry)

values) below:

$$C_{vr} = C_{vp}\left(\frac{MAP}{MAP - AET}\right) \quad (1)$$

Where

C_{vr} – coefficient of variation of annual runoff,
C_{vp} – coefficient of variation of annual precipitation
MAP – mean annual precipitation
AET – annual total evaporation

The implications are that any change in ET in a catchment will alter the variability of runoff (assuming that precipitation is unchanged). Furthermore, if ET is increased, C_{vr} will also increase. As a change in land cover is the only significant change in ET that may be expected, Peel *et al.* (2001) concluded that any change in land cover that transpires more than the original will result in an increase in runoff variability, as well as a decrease in annual runoff. This is an issue of particular concern when tall evergreen tree plantations are planned in catchments with predominately grassland cover, and in addition to the well-known concerns in South Africa, which has led to calls for caution in the establishment of forest plantations elsewhere in the world (e.g. Vertessy *et al.*, 2003). The consideration of runoff reduction as well as increased variability is especially important when considering a potential reduction in water resources yield from a catchment. Peel *et al.* (2002) made use of the Gould–Gamma procedure to highlight the impact that such land-cover change could have on yield–storage relationships based on annual runoff from a catchment. According to this method, constant regulated yield (D), is a ratio of mean annual flow from a reservoir equal in size to the mean annual inflow into the reservoir with a reliability of 95% is defined as:

$$D = 1 - \frac{1.12}{1 - 1.67{C_{vr}}^{-2}} \quad (2)$$

Thus, storage yield is dependent upon both the amount and the variability of runoff with the implication that changes in catchment land cover that increase the variability of runoff must also reduce the yield and the reliability of that yield: a critical concern for water resources planners. In South Africa, the invasion of large areas of a catchment by forest species IAPs has been shown to reduce the yield from water supply systems considerably (Versfeld *et al.*, 1998; Gorgens and van Wilgen, 2004). Yield considerations provide a useful input to decision making. However, the provision of annual storage yield is only one of many considerations in water resources planning. In this regard, the consideration of periods of low flows when direct abstractions from a river, or when environmental minimum flows are specified, is critical.

Low Flows

Low flows are a normal, seasonal part of the natural hydrologic regime. Low flows are distinct from drought, though the effects are often similar (Smakhtin, 2001). In the absence of an adequate store of water, the low flow period is critical when considering competing water user needs, as it is in these periods that most conflict arises between different users, water supply schemes fail, river ecosystems are stressed, and water pollution issues are highlighted. As highlighted in the introduction to this article, it has often been suggested that forests behave as "sponges", retaining rainwater that is slowly released following a rainfall event or wet season and thus maintain or regulate river flows. However, recent advances in hydrological process studies, particular those which examine the interconnection between surface and ground water have highlighted the fallacy of this argument. Although many definitions of low flow exist, the key issue in the context of forest hydrology studies is an emerging body of work that highlights the role of accumulated soil moisture flow in the unsaturated zone, in addition to groundwater, in the generation of low flows (Kendall and McDonnell, 1998; Lorentz *et al.*, 2003). The basis for the assumption that forests "store water" is the high infiltration rate that is often found on forested soils, and the specific hydraulic properties of many organic rich forest soils. However, it is increasingly recognized that the transpiration demand of the forest and their deep roots that provide access to soil moisture deep in the profile mean that forests are more likely to have a negative, rather than positive impact on low flows from a catchment. Conversely, the most significant change following deforestation is an increase in "dry weather" streamflow (delayed flow) that arises from a decrease in total evaporation (Bonell, 1998).

Observations from South Africa and India indicate that increased dry period transpiration reduces low flows. Smith and Scott (1992), analyzing results from five of the South African catchment studies, concluded that percentage reductions in low (dry season) flow as a result of afforestation were actually greater than the reduction in annual flow. In South Africa, estimates of low flow reduction are considered more important in assessing applications for licences for commercial plantations than the estimated reduction in Mean Annual Runoff (MAR) (Jewitt, 2002b). In India, Sikka *et al.* (2003), focusing on the Nilgiri catchment in southern India, identified reductions in low flows in the driest months of the year when comparing streamflow from a grassland catchment with that from a catchment afforested with Eucalyptus globulus. In Fiji, the planting of Pinus radiata in a dry grassland area is reported to have resulted in a 65% reduction in dry season flows (Kammer and Raj, 1979) cited by FAO (1991). Conversely, observations from many parts of the world have highlighted the increase in dry season flows following deforestation (Bruijnzeel, 2004).

Authors such as Zadroga (1981) and Bruijnzeel (1990, 2001) have highlighted the hydrological significance of tropical montane cloud forests (TMCF) and their role in generating dry season flows through the interception of fog. However, Bruijnzeel (2001) concluded that it is the infiltration properties of the forest that are critical in determining how the available water is partitioned between runoff and recharge and that decreased dry season flows following deforestation in areas of TMCF may not necessarily result from the loss interception of fog, but possibly from decreased infiltration capacity of the soil once the forest cover has been removed. This highlights the importance of the partitioning points illustrated in Figure 1, and gives credence to Rodriguez-Iturbe's assertion that Climate-Soil-Vegetation dynamics is the core of hydrology (Rodriguez-Iturbe, 2000) with the recognition that different and often competing processes may dominate at different sites and that the direction, let alone the magnitude of the impact, may be difficult to predict for a particular site (Calder, 1998).

Groundwater Resources

Worldwide, forests have been reported to both increase and decrease groundwater depths (Chang, 2002). The combination of complex forest plant characteristics and subsurface stratigraphy and soil functioning mean that general statements in this regard are often not applicable at local scales. However, most reports on the influences of forests on groundwater suggest that the perception that forests are "good" for groundwater resources is false. Some authors have reported increases in available groundwater following deforestation (Hamilton, 1987; Ruprecht and Stoneman, 1993) whilst others, particularly in arid and semiarid regions, have reported a lowering in the groundwater surface following afforestation. Hamilton and King (1983) made a strong statement that "the overwhelming evidence from catchment research is that following reforestation, groundwater levels are lowered and stream yields are reduced, both effects being more pronounced during the dry season or growing season".

There is an extensive body of literature that highlights lowering of groundwater levels in response to afforestation (e.g. Bari and Schofield, 1992). In semiarid regions, most notably areas of Australia, there has been extensive research into the water resources problems associated with the rise of saline groundwater tables in response to removal of trees, particularly eucalyptus, to establish other forms of agriculture. In central England, a study by Calder (1999) showed there was no wetting of the soil profile beneath newly planted stands of Corsican pines, whereas wetting fronts were observed down the profile beneath stands of indigenous oak and heath vegetation. This suggests that large-scale establishment of Corsican pine in the area will have a detrimental affect on groundwater recharge, and thus water resources in the area.

Some forest species, by way of their deep roots are able to access the phreatic surface and directly abstract water, especially in cases where the groundwater table is shallow. Studies of the soil water use by trees of the Eucalyptus family under soil-water deficits in India (Calder *et al.*, 1993) and Australia (Ruprecht and Stoneman, 1993) have found that trees at some sites were obtaining most of their soil water from the groundwater, resulting in direct depletion of the groundwater. Conversely, there are many reports of rising water tables following the deforestation of parts of Australia, and reforestation is often seen as part of the solution to salinity problems that have arisen in these areas (Schofield, 1992.).

The issue of trees and groundwater in India has provoked much controversy. Many impoverished rural communities are dependent upon access to relatively shallow groundwater for their survival. However, a number of developments, including abstractions for irrigation and large-scale establishments of eucalypt plantations, have resulted in a lowering of the groundwater table, with concomitant impacts on rural livelihoods. Shiva and Bandyopadhyay (1983), focusing on the Karnataka region in southern India, described eucalyptus as a "disastrous tree for India" highlighting, *inter alia* the potential for the lowering of local groundwater tables through a reduction in the recharge of groundwater. The controversy has resulted in many publications, including reviews by the British Overseas Development Agency (ODA, now DFID) (Calder *et al.*, 1993; Calder, 1994) and the Food and Agriculture Organisation (FAO, 1989). Both reports confirm the potential for eucalypts to lower groundwater tables, but moderate many of the claims made by Shiva and Bandyopadhyay (1983). Calder *et al.* (1993) concluded that in three of the four sites studied eucalypts would not use more water than the indigenous vegetation, but would use more water than many of the locally produced crops. However, in one of the sites, water use was found to exceed rainfall, leading Calder *et al.* (1993) to conclude that it was possible that the trees were using groundwater directly. The FAO (1989) report recognized the potential for detrimental affects on local groundwater tables, but highlighted the water use efficiency by eucalypts suggesting that they consume less water for each ton of biomass produced compared to many other species.

Flood Peaks (Volume and Timing)

It is in analyzing the role of forest on floods that some of the most heated arguments surrounding water and forests arise. It is often claimed that forests prevent floods and that deforestation is a major cause of floods. A flood is the result of many complex catchment interactions. A catchment exhibits variable responses to similar magnitude rainfall events because of multiple interacting processes,

spatial variability in the susceptibility of the catchment to impacts, impacts of land cover and associated management practices, and amplification or suppression of response due to differing environmental conditions. With reference to Figure 1, it could be expected that interception of rainfall by forests may reduce the magnitude of flood peaks, delay their onset by intercepting a proportion of the storm rainfall and by creating soil moisture deficits prior to the event. Such effects would be most significant for small storms and least significant for the larger events. Furthermore, high infiltration rates under natural forests may also serve to reduce surface runoff and flood response.

However, forest hydrological studies carried out in many parts of the world: America (Hewlett and Helvey, 1970), South Africa (Hewlett and Bosch, 1984), United Kingdom (Kirby *et al.*, 1991), New Zealand (Taylor and Pearce, 1982), and Asia (Bruijnzeel and Bremmer, 1989) have all concluded that there is little linkage between land cover and storm flow. However, Fahey and Jackson (1997) have shown that significant decreases of 60% in mean flood peaks for different rainfall classes on small catchments in New Zealand, which were converted from indigenous tussock grasslands to pine plantations. More recently (La Marche and Lettenmair, 2001; Robinson *et al.*, 2003; Sikka *et al.*, 2003) have all concluded that the impacts of forest on floods are only likely to be noticeable for minor events and on small catchments. Bonell (1998) through a detailed study of catchment runoff mechanisms has shown that floods can occur from pristine forested drainage basins on the same scale as from disturbed areas. When unusually heavy storms form, floods are liable to occur whether forests are present or not.

However, field studies generally indicate that it is often the management activities associated with forestry such as cultivation, drainage, road construction (Jones and Grant, 1996), and soil compaction during logging that are more likely to influence, and may increase flood response, than the presence or absence of the forests themselves. Furthermore, Lee (1980) suggested that the most important influence of forest cover on floods and flood damage "has more to do with sedimentation and debris discharge than with absolute volume or rate of flow" (Lee, 1980 p. 280).

CONCLUDING REMARKS

Despite improved understanding of forest hydrology processes, there are still many calls for more research to address forest water use issues. Arguably, however, the area that needs most attention is the science-management interface. Too often, the complexities of the hydrological system have been simplified because of management approaches that cannot adequately consider complexity, and others whose goals are best served by promoting existing pseudofacts. The management of both water and forests requires a sound scientific underpinning. The investment in forest hydrology research worldwide has provided some of the best databases in the world in this regard. Ongoing innovations in hydrological process studies will continue to add value to this body of information. However, management of forests and water requires not only a sound scientific base, but also understanding, commitment, and collaboration of the responsible and affected organizations. As is the case in many aspects of natural resource management, scientists and researchers have failed to ensure that the results of their research are properly applied. The problems lie as much in scientists' reluctance to follow through to implementation of their research and present and promote their findings in management forums, as with the reluctance of management to deal with complexity. The interfaces between these stakeholders are still inadequate, and in particular, the satisfactory implementation of research findings into management operations is an area that needs much attention.

REFERENCES

Bari M.A. and Schofield N.J. (1992) Lowering of a shallow, saline water table by extensive eucalypt reforestation. *Journal of Hydrology*, **133**, 273–291.

Bernard J.M. (1963) Forest floor moisture capacity of the New Jersey pine barrens. *Ecology*, **44**, 574–576.

Best A., Zhang L., McMahon T., Western A. and Vertessy R. (2003) *A Critical Review of Paired Catchment Studies with Reference to Seasonal Flows and Climatic Variability*, CSIRO Land and Water Technical Report 25/03, CSIRO; CRC for Catchment Hydrology Technical Report 03/4, CRC.

Blackie J.R., Edwards K.A. and Clarke R.T. (1979) Hydrological research in East Africa. *East African Agricultural and Forestry Journal*, **45**, entire special edition, edited by Blackie.

Blow F.E. (1955) Quantity and hydrologic characteristics of litter and upland oak forests in eastern Tennessee. *J. Forestry*, **53**, 190–195.

Boden D.I. (1984) Early responses to different methods of site preparation for three commercial tree species. *Proceedings of the IUFRO Symposium on Site and Productivity of Fast Growing Plantations*, Pretoria and Pietermaritzburg.

Bonell M. (1993) Progress in the understanding of runoff generation dynamics in forests. *Journal of Hydrology*, **150**, 217–275.

Bonell M. (1998) Selected challenges in runoff generation research in forests from the hillslope to headwater drainage basin scale. *Journal of the American Water Resources Association*, **34**(4), 765–785.

Bruijnzeel L.A. (1990) *Hydrology of Moist Tropical Forests and Effects of Conservation: A State of Knowledge Review*, UNESCO International Hydrological Programme, UNESCO: Paris.

Bruijnzeel L.A. (2001) Hydrology of tropical montane cloud forests. *A Reassessment Landuse and Water Resources Research*, **1**, 1.

Bruijnzeel L.A. (2004) Hydrological functions of tropical forests, not seeing the soil for the trees? In *Environmental Services*

and Land Use Change: Bridging the Gap Between Policy and Research in Southeast Asia, Tomich T.P., van Noordwijk M. and Thomas D.E. (Eds.), A special issue of Agriculture, Ecosystems and Environment: Vol. 104/1, pp. 185–228, (September).

Bruijnzeel L.A. and Bremmer C.N. (1989) Highland-lowland interactions in the Ganges-Brahmaputra river basin: *A review of Published Literature*. ICIMOD Occasional Paper, No.11. ICIMOD.

Bosch J.M. and Hewlett J.D. (1982) A review of catchment experiments to determine the effects of vegetation changes on water yield and evapotranspiration. *Journal of Hydrology*, **55**, 3–23.

Bosch J.M. and Smith R.E. (1989) The effect of afforestation of indigenous scrub forest with Eucalyptus on streamflow from a small catchment in the Transvaal, South Africa. *South African Forestry Journal*, **150**, 7–17.

Calder I.R. (1990) *Evaporation in the Uplands*, Wiley: New York, p. 148.

Calder I.R. (1994) *Eucalyptus, Water and Sustainability: A Summary Report*, ODA Forestry Series No. 6, ODA: London.

Calder I.R. (1996) Water use by forests at the plot and catchment scale. *Commonwealth Forestry Review*, **75**(1), 19–30.

Calder I.R. (1998) Water use by forests: limits and controls. *Tree Physiology*, **18**, 625–631.

Calder I.R. (1999) *The Blue Revolution, Land Use and Integrated Water Resources Management*, Earthscan.

Calder I.R. and Dye P.J. (2001) Hydrological impacts of invasive alien plants. *Land Use and Water Resources Research*, **1**, `http://www.venus.co.uk/luwrr`.

Calder I.R., Hall R.L. and Prasanna K.T. (1993) Hydrological impact of Eucalyptus plantation in India. *Journal of Hydrology*, **150**(2–4), 635–648.

Chang M. (2002) *Forest Hydrology: An Introduction to Water and Forests*, CRC Press.

Cossalter C. and Pye-Smith C. (2003) *Fast-Wood Forestry, myths and realities*, Centre for International Forestry Research (CIFOR): Jakarta, p. 50.

Crockford R.H. and Richardson D.P. (1990a) Partitioning of rainfall in a eucalypt forest and pine plantation: i the effect of throughfall measurements in a eucalypt forest, effect of method and species composition. *Hydrological Processes*, **4**, 157–167.

Crockford R.H. and Richardson D.P. (1990b) Partitioning of rainfall in a eucalypt forest and pine plantation: IV The relationship of interception and canopy storage capacity, the interception of these forests, and the effect on interception of thinning the pine plantation. *Hydrological Processes*, **4**, 169–188.

DeWalle D.R. (2003) Forest hydrology revisited. *Hydrological Processes*, **17**(6), 1255–1256.

Dye P.J. (1996a) Climate, forest and streamflow relationships in South African afforested catchments. *Commonwealth Forestry Review*, **75**(1), 31–38.

Dye P.J. (2003) *Personal Communication*, Council for Scientific and Industrial Research: Pietermaritzburg.

Dye P.J. and Bosch J.M. (2000) Water, wetlands and catchments. *South African Forest Handbook* South African Institute of Forestry, Pretoria South Africa Vol. 2, pp. 567–573.

Dye P.J., Moses G., Vilikazi P., Ndlela R. and Royappen M. (2000) *A Comparison of the Water Use of Selected Invasive and Indigenous Riparian Plant Communities*, WRC Report K5/808, Water Research Commission, Pretoria.

Eamus D., O'Grady A.P. and Hutley L. (2000) Dry season conditions determine wet season water use in the wet-dry tropical savannas of northern Australia. *Tree Physiology*, **20**, 1219–1226.

Engler A. (1919) Untersuchungen u¨ ber den einfluss desWaldes auf den stand der gewasser. Mitt. Schweiz. anst. forst. *Versuchswes*, **12**, 636.

FAO (1989) *The Eucalypt Dilemna*, Food and Agriculture Organization of the United Nations.

FAO (1991) *Forestry and Food Security*, *Food and Agriculture Organization of the United Nations*: Forestry Paper 90. ISBN 92-5-102847-8.

FAO (2003) *Cross-sectoral policy impacts between forestry and other sectors*. FAO Forestry Paper No. 142. ISBN 9251049378.

Falkenmark M., Andersson L., Castensson R. and Sundblad K. (1999) *Water – A Reflection of Land Use. Options for Counteracting Land and Water Mismanagement*, Swedish Natural Science Research Council: Stockholm.

Fahey B. and Jackson R. (1997) Hydrological impacts of converting native forests and grasslands to pine plantations, South Island, New Zealand. *Agricultural and Forest Meteorology*, **84**, 69–82.

Gorgens A.H.M. and van Wilgen B.W. (2004) Invasive alien plants and water resources in South Africa: current understanding, predictive ability and research challenges. *South African Journal of Science*, **100**(1/2), 27–34.

Gray D.M. (1970) *Handbook on the Principles of Hydrology*, National Research Council of Canada, Water Information Center Inc.: Water Research Building, Manhasset Isle, Port Washington, p. 11050.

Gunderson L.H., Holling C.S. and Light S.S. (Eds.), (1995) *Barriers & Bridges to the Renewal of Ecosystems and Institutions*, Columbia University Press: New York.

Hamilton L.S. (1987) What are the impacts of deforestation in the Himalayas on the Ganges-Brahmaputra lowlands and delta? relations between assumptions and facts. *Mountain Research and Development*, **7**, 256–263.

Hamilton L.S. and King P.N. (1983) *Tropical Forested Watersheds: Hydrological and Soils Response to Major Uses or Conservation*, Westview Press: Boulder.

Helvey J.D. (1967) Interception by eastern white pine. *Water Resources Research*, **3**, 723–729.

Helvey J.D. and Patric J.H. (1965) Canopy and litter interception of rainfall by hardwoods of the eastern United States. *Water Resources Research*, **1**, 193–206.

Hewlett J.D. (1982) *Principles of Forest Hydrology*, University of Georgia Press.

Hewlett J.D. and Bosch J.M. (1984) The dependence of storm flows on rainfall intensity and vegetal cover in South Africa. *Journal of Hydrology*, **75**, 365–381.

Hewlett J.D. and Helvey J.D. (1970) Effects of forest clearfelling on the storm hydro-graph. *Water Resources Research*, **6**(3), 768–782.

Hewlett J.D. and Hibbert A.R. (1967) Factors affecting the response of small watersheds to precipitation in humid areas. In

Proceedings – International Symposium on Forest Hydrology, Sopper W.E. and Lull H.W. (Eds.), Pergamon: New York, pp. 275–290.

Hewlett J. and Nutter W. (1969) *An Outline of Forest Hydrology*, University of Georgia Press: Athens.

Higgins S.I. and Richardson D.M. (1996) A review of models of alien plant spread. *Ecological Modelling*, **87**, 249–265.

Hornberger G.M., Raffensperger J.P., Wiberg P.L. and Eshleman K.N. (1997) *Elements of Physical Hydrology*, Johns Hopkins University Press: Baltimore.

Jewitt G.P.W. (2002a) Can integrated water resources management sustain the provision of ecosystem goods and services? *Physics and Chemistry of the Earth*, **27/11-22**, 887–895.

Jewitt G.P.W. (2002b) The 8%-4% debate: commercial afforestation and water use in South Africa. *SA Forestry of Journal*, **194**(2), 4–6.

Johnston C.D. (1987) Preferred water flow and localised recharge in a variable regolith. *Journal of Hydrology*, **94**, 129–142.

Jones J.A. and Grant G.E. (1996) Peak flow responses to clear-cutting and roads in small and large basins, western Cascades, Oregon. *Water Resources Research*, **32**, 959–974.

Kammer R. and Raj (1979) *Preliminary Estimates of Minimum Flows in Varaciva Creek and the Effect of Afforestation on Water Resources*, Fiji Public Works Department: Technical Note 7911, Suva.

Kendall C. and McDonnell J.J. (Eds.) (1998) *Isotope Tracers in Catchment Hydrology*, Elsevier Science Publishers; p. 816.

Kirby C., Newson M.D. and Gilman K. (1991) *Plynlimon Research: The First Two Decades*, Institute of (Wallingford) Hydrology Report, No.109, Institute of (Wallingford) Hydrology, p. 188.

Kittredge J. (1948) *Forest Influences*, McGraw Hill Book Company: New York.

Landsberg J.J. (1999) Tree water use and its implications in relation to agroforestry systems. In *The Way Trees Use Water'*, Water and Salinity Issues in Agroferestry No.5. RIRDC Publication No. 99/37,Landsberg J.J. (Ed.), RIRDC, p. 92.

Le Maitre D.C., Scott D.F. and Colvin C. (1999) A review of information on interactions between vegetation and groundwater. *Water SA*, **25**, 137–152.

La Marche J. and Lettenmair D.P. (2001) Effects of forest roads on flood flows in the Deschutes River, Washington. *Earth Surface Processes and Landforms*, **26**, 115–134.

Lee R. (1980) *Forest Hydrology*, Columbia University Press: New York.

Lorentz S.A.L., Schulze R.E. and Hughes G. (2003) Techniques for estimating groundwater recharge at different scales in southern Africa. In *Groundwater Recharge Estimation in Southern Africa*, Xu. X. and Beekman H.E. (Eds.), UNESCO: Paris.

Mack M. and d'Antonio C.M. (1998) Impacts of biological invasions on disturbance regimes. *TREE*, **13**, 195–198.

McCulloch J.S.G. and Robinson M. (1993) History of forest hydrology. *Journal of Hydrology*, **150**, 189–216.

Metz L.J. (1958) Moisture held in pine litter. *Journal of Forestry*, **56**, 36–37.

Mielke M., Oliva M.A., Barros N., Penchel R., Martinez C. and Almeida A. (1999) Stomatal control of transpiration in the canopy of a clonal Eucalyptus grandis plantation. *Trees*, **13**, 152–160.

Nänni U.W. (1970) Trees, water and perspective. *South African Forestry Journal*, **75**, 9–17.

O'Loughlin E.M. and Dunin F.X. (1993) Water Issues in Forest Today. *Journal of Hydrology*, *(Special Issue)* **150**(2–4), 189–786.

Pathak P.C., Pandey A.N. and Singh J.S. (1985) Apportionment of rainfall in central Himalayan forests (India). *Journal of Hydrology*, **76**, 319–332.

Peel M.C., McMahon T.A., Finlayson B.L. and Watson F.G.R. (2001) Identification and explanation of continental differences in the variability of annual runoff. *Journal of Hydrology*, **250**, 224–240.

Peel M.C., McMahon T.A., Finlayson B.L. and Watson F.G.R. (2002) Implications of the relationship between catchment vegetation type and the variability of annual runoff. *Hydrological Processes*, **16**, 2995–3002.

Pryor L.D. (1976) *Biology of Eucalypts*. Edward Arnold: London, pp 51–58.

Rice K. and Hornberger G.M. (1998) Comparison of hydrochemical tracers to estimate source contributions to peak flow in a small, forested headwater catchment. *Water Resources Research*, **34**(7), 1755–1766.

Richards J.H. and Caldwell M.M. (1987) Hydraulic lift: substantial nocturnal water transport between soil layers by Artemisia tridentata roots. *Oecologia*, **73**, 486–489.

Robinson M., Cognard-Plancq A.L., Cosandey C., David J., Durand P., Fuhrer H.W., Hall R., Hendriques M.O., Marc V., McCarthy R., McDonnell M., Martin C., Nisbet T., O'Dea P., Rodgers M. and Zollner A. (2003) Studies of the impact of forests on peak flows and baseflows: a European perspective. *Forest Ecology and Management*, **186**, 85–97.

Rodriguez-Iturbe I. (2000) Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamics. *Water Resources Research*, **36**(1), 3–9.

Rowe P.B. (1955) Effects of the forest floor on disposition of rainfall in pine stands. *Journal of Forestry*, **53**, 342–348.

Ruprecht J.K. and Stoneman G.L. (1993) Water yield issues in the jarrah forest of south-western Australia. *Journal of Hydrology*, **150**, 369–391.

Russel E.W. (1973) *Soil Conditions and Plant Growth*, Longmans: London, p. 847.

Sala A., Smith S.D. and Devitt D.A. (1996) Water use by Tamarix ramosissima and associated phreatophytes in a Mojave Desert floodplain. *Ecological Applications*, **6**(3), 888–898.

Schofield N.J. (1992) Tree-planting for dryland salinity control in Australia. *Agroforestry Systems*, **20**, 1–23.

Schutz C.J. (1990) *Site relationships for Pinus Patula in the Easter Transvaal Escarpment*, Unpublished P.H.D. thesis. Department of Soil Science and Agrometeorolgy, University of Natal, PMB.

Smith R.E. and Scott D.F. (1992) The effects of afforestation on low flows in various regions of South Africa. *Water SA*, **18**(3), 185–194.

Scott D.F., Prinsloo F.W., Moses G., Mehlomakulu M. and Simmers A.D.A. (2000) *A Re-analysis of the South African Catchment Afforestation Experimental Data*, WRC Report 810/1/00, Water Research Commission, Pretoria.

Shiva V. and Bandyopadhyay J. (1983) Eucalyptus – a disastrous tree for India. *The Ecologist*, **13**, 184–187.

Sherry S.P. (1971) *The Black Wattle*, University of Natal Press: Pietermaritzburg.

Sikka A.K., Samra J.A., Sharda V.N., Samraj P. and Lakshmanan V. (2003) Low flow and high flow responses to converting natural grassland into bluegum (Eucalyptus globulus) in Nilgiris watersheds of South India. *Journal of Hydrology*, **270**(2003), 1–2.

Smakhtin V.U. (2001) Low flow hydrology: a review. *Journal of Hydrology*, **240**, 147–186.

Specht R.L. (1972) *The Vegetation of South Australia*, Government Printers: Adelaide.

Taylor C.H. and Pearce A.J. (1982) Storm runoff processes and sub-catchments characteristics in a New Zealand hill country catchment. *Earth Surface Processes and Landforms*, **7**, 439–447.

Versfeld D.B., Le Maitre D.C. and Chapman R.A. (1998) *Alien Invading Plants and Water Resources in South Africa:* A Preliminary Assessment, WRC Report No. TT99/98, Water Research Commission, Pretoria, ISBN 1 86845 360.

Vertessy R.A., Zhang L. and Dawes W.R. (2003) Plantations, river flows and river salinity. *Australian Forestry*, **66**(1), 55–61.

White D.A., Turner N.C. and Gailbraith J.H. (2000) Leaf water relations and stomatal behavior of four allopatric Eucalyptus species planted in Mediterranean southwestern Australia. *Tree Physiology*, **20**, 1157–1165.

Wullschleger S.D., Hanson P.J. and Tschaplinski T.J. (1998) Whole-plant water flux in understory red maple exposed to altered precipitation regimes. *Tree Physiology*, **18**, 71–79.

Xinxiao Y. (1991) Forest hydrologic research in China. *Journal of Hydrology*, **122**, 23–31.

Zadroga F. (1981) The hydrological importance of a montane cloud forest area of Costa Rica. In *Tropical Agricultural Hydrology*, Lal R. and Russell E.W. (Eds.), J. Wiley: New York, pp. 59–73.

Zhang L. (1999) *Predicting the Effect of Vegetation Changes on Catchment Average Water Balance*, Technical Report 99/12, Cooperative Research Centre for Catchment Hydrology, CSIRO, Canberra.

Zon R. (1927) Forests and water in the light of scientific investigation. *U.S. National Waterways Commission*, Final Report, Appendix V, 205–302. 62nd Congress, 2d session Senate Document No. 469.

187: Land Use Impacts on Water Resources – Science, Social and Political Factors

TIM FORSYTH

Development Studies Institute, London School of Economics and Political Science, London, UK

Public perceptions of land-use impacts on water resources are important because they influence formal environmental policies and popular attitudes about land use. For example, upland agriculture and deforestation are commonly blamed for reducing rainfall levels and for causing lowland water shortages, and some governments have consequently passed logging bans or restrictions on upland agriculture. However, there is intense controversy concerning whether these public perceptions are supported by scientific evidence; whether these policies may actually address underlying problems; and how far hydrological science is itself influenced by social and political factors. This article draws upon debates in social science (rather than hydrological science alone) to discuss linkages between public perceptions and scientific explanations of hydrological change. As examples, the article discusses the cases of so-called Himalayan environmental degradation; dryland desiccation in the Sahel; and debates concerning the impacts of deforestation and reforestation in watersheds. The article argues that hydrologists should not categorize public perceptions and formal science separately, but see both as mutually evolving. Dominant perceptions, or "narratives", of hydrological change may occur from various coincidences of historical research and public concern, and reflect the viewpoints and experiences of only selected social groups. It is suggested that hydrologists can help overcome apparent conflicts between public perceptions and scientific explanations by increasing public participation in describing and framing complex environmental problems, and using this information to make existing narratives more diverse and flexible. There is also a need to understand the political and institutional factors that lead to the persistence of contested narratives within different national or organizational contexts.

INTRODUCTION

Human impacts on hydrological systems are a growing source of public concern. Few days seem to pass without some kind of public debate about society's impacts on water resources, which are usually seen as negative. In 2000, for example, Britain's Prince Charles once blamed a period of unusually high rainfall in Britain on "Mankind's arrogant disregard of the delicate balance of nature" ("Storms are Man's fault, says Prince" by Charles Clover, The Daily Telegraph, 7 November 2000, pp. 1). In other countries, many activists warn that excessive agriculture or deforestation will lead to desertification. Around the world, various policies and land-management schemes restrict human activities because of their presumed impacts on water resources.

But are these statements and policies justified? Can we really blame human activities for apparent problems such as water shortages or changes in rainfall? New research is suggesting that human impacts on water resources are highly varied, and that many common perceptions, or generalizations, about human impacts of land use are simplistic. These findings are fueling a debate about the origin of public perceptions of human impacts on water resources. The debate includes asking "how and why do public perceptions become seen as 'fact'?" Or, "how can we make public perceptions more accurate?"

This article of the Encyclopedia may be slightly different to others because it summarizes social and political debates rather than hydrological science alone. The article adopts a "science-studies" approach, or a focus on the political and social contexts that shape hydrological science. This

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

approach does not suggest that hydrological science cannot help explain environmental change, but instead shows how current and historic social concerns have influenced explanations of complex environmental problems. Understanding these social influences can help make hydrological science more effective and reduce apparent conflicts between public perceptions and science.

THE SOCIAL SCIENCE DEBATE ABOUT ENVIRONMENTAL ''MYTHS'' AND PERCEPTIONS

What are "perceptions", and why are they important? At the most basic level, all humans have their own perceptions – or experiences of – environmental events or changes. In this sense, a "perception" may be one person's individual experience of, say, a flood, or knowledge of how hydrological events occur, or what may cause them.

But "perceptions" may also include the tendency for large numbers of people to make assumptions about the nature or causes of environmental problems or hydrological events. In this sense, perceptions are no longer controlled by individuals' experiences, but are influenced by wide-scale discourses, or "received wisdom", and which are commonly reported in books, popular discussions, and media as though they are unquestioned "facts". Some examples of common perceptions concerning hydrology include:

- Forests increase rainfall (and hence deforestation decreases rainfall).
- Agriculture in upland areas reduces the downstream supply of water from those areas.
- Agriculture in dry lands leads to desiccation and irreversible degradation of land.
- Forests in upland areas reduce erosion and prevent lowland flooding (and hence deforestation may cause erosion and flooding).

Such views are frequently adopted as "factual" by many formal organizations such as state forestry departments or international development agencies. Yet, an increasing number of researchers are now questioning these statements because they are challenged by a variety of research and other evidence (e.g. Calder, 1999).

This kind of conflict between public perception and scientific research is, of course, not new, nor restricted to hydrology. Some authors have discussed the mismatches between orthodox perceptions of environmental problems and scientific research for years, and have called perceptions "myths" (see Leach and Mearns, 1996; Forsyth, 2003 for summaries). Some physical scientists have used the word "myths" to indicate scientific falsehoods, or the persistence of ideas after research should have falsified them. For example, Thomas and Middleton (1994) called desertification a "myth" because they claim recent research has made the traditional meanings attributed to "desertification" inapplicable to newer understandings of desiccation and drought (see the discussion of dryland desiccation below).

However, many social scientists have suggested that environmental "myths" should not be seen as falsehoods, but as "truths" that uphold some essence of cultural belief or knowledge (similar to ancient myths or folklore). This definition of "myth" throws the debate on its head: rather than asking why apparently false explanations still persist, we should ask why these explanations are still seen as "true" in certain contexts.

So, where do "myths" come from? For many analysts, the existence of environmental "myths" is explained by insufficient research or poor communication between scientists and the public and policymakers. Environmental problems are usually highly complex, and information is difficult to achieve for long time or space scales. Furthermore, it is not surprising that government agencies or the public cannot keep abreast of all research.

Researchers in science studies, however, consider these explanations insufficient. Instead of seeing a linear relationship – of science feeding public perceptions or policy – analysts seek each as coevolving and mutually enforcing. Under science studies, researchers look more at the social and political factors that lead to the identification and research of distinct environmental problems in the first place, and then at the social barriers to communication between different disciplines, social groups, or organizations. They argue that all scientific research reflects different social framings in what is researched, and how, and how policy circles see different knowledge claims as authoritative or legitimate. This approach does not ask "why doesn't the public or policymakers listen to science?" but, "what social factors lead to different scientific explanations being considered meaningful and true?"

It is worth reviewing the two main approaches to explaining environmental myths within science studies. The earliest approach (Cultural Theory is an approach influenced by British sociologist, Mary Douglas. It distinguishes itself from other cultural studies by using a capital C and T) argued that all societies inevitably contain four distinct social groups with specific worldviews (Schwarz and Thompson, 1990). *Individualists* are those seeking to maximize personal gain, and who see little social responsibilities for their actions (e.g. transnational corporations). *Egalitarians* are those most worried about the communal implications of social actions (e.g. nongovernmental organizations). *Hierarchists* are those seeking to impose rules on society in order to manage social actions and accommodate all perspectives (e.g. governments). And, *fatalists* are those who feel powerless to affect any change on how society works (e.g. poor farmers). According to Cultural Theory,

each group will explain environmental change in characteristically optimistic, pessimistic, managerial, or fatalist ways. Consequently, effective environmental governance would not rely upon deciding which of these worldviews was correct, but in devising a system of government that acknowledged all environmental claims were filtered in these ways. As Thompson *et al.* (1986) claimed, governors should not ask the public "what are the facts?" but "what would you like the facts to be?"

Later analysts, however, have questioned whether these four groups are too reductionist. The narrative, or storylines, approach focuses instead on the historical evolution of environmental explanations as linguistically and culturally embedded conventions (see Hajer, 1995). Under this approach, environmental myths emerge historically as the result of how different social groups and conflicts have defined them. Over time, narratives are seen to be factual or universal, but reflect only a limited knowledge of biophysical events, and limited perspectives on how such processes can be perceived by different groups. Two processes are especially influential on narratives. First, narratives usually reflect a selective "problem closure", or framing of how hydrological events are seen to be problematic to selected groups of people. For example, in Bengal, India during the nineteenth century, formal forest policies prioritized teak and sal production, and the practices of shifting cultivators were seen to be inimical to the objectives of the foresters. Hill forest areas were identified as less valuable for timber production and therefore, were burnt to encourage the cultivation of less valuable products such as sabai grass (Sivaramakrishnan, 2000). A second process is the formation of "discourse coalitions", or where political negotiations between different actors may serve to reinforce a narrative. For example, historians of environmentalism in North America have claimed that industrialization was associated with a growing perception of "wilderness" as threatened, and the urgent need to protect areas such as watersheds. This may lead to a discourse coalition when conservationist NGOs and governments may disagree about levels of logging versus conservation, but agree on narratives concerning the impacts of upland agriculture on watersheds even if these impacts are uncertain. This was perhaps illustrated when the Indonesian Forum for the Environment (Walhi) brought an action against the Indonesian government in 2004 for allegedly being responsible for flash flooding in Sumatra in 2003 by allowing logging to take place in upland watersheds.

Narratives, or storylines, therefore reflect the political debates of their day, and are partly controlled by the degree of public participation in formulating them. These, in turn, have political impacts and purposes. Hajer (1995, pp. 64–65) wrote, "Storylines are devices through which actors are positioned, and through which specific ideas of 'blame' and 'responsibility' and 'urgency' and 'responsible behavior' are attributed". Other authors have suggested that narratives allow actors such as governments or development agencies to foreclose scientific debate in order to support specific policies. "Development narratives tell scenarios not so much about what should happen as about what will happen according to their tellers – if the events or positions are carried out as described" (Roe, 1991, p. 288). Accordingly, it is important not to take a narrative at face value, and to ask, instead, on whose information and framings they are based.

The following examples illustrate some well-known narratives concerning land-use impacts on water resources. This article cannot hope to summarize all aspects of hydrological research on these important topics (see Bonell and Bruijnzeel, 2004; Bruijnzeel, 2004; Chomitz and Kumari, 1998), but the purpose here is to show the links between public perceptions and hydrological science, and how changing the framings and public participation has affected explanations.

HIMALAYAN ENVIRONMENTAL DEGRADATION THEORY

One of the most famous debates relating to public perceptions of land-use impacts on water resources is the so-called "Himalayan theory of environmental degradation". This narrative was used to describe processes of land degradation in the wet Middle Hills of Nepal resulting from the effects of rapid population growth and modernization. Eckholm (1976, p. 77) summarized the "theory" when he wrote:

> "Population growth in the context of a traditional agrarian technology is forcing farmers onto even steeper slopes, slopes unfit for sustained farming even with the astonishingly elaborate terracing practiced there. Meanwhile, villagers must roam further and further from their houses to gather fodder and firewood, thus surrounding villages with a widening circle of denuded hillsides."

In turn, these hillside processes have also been linked to a further vicious circle of deforestation and soil erosion that, in turn, enhance lowland floods and water shortages and sedimentation as far east as Bangladesh. Much of this narrative has been based upon the idea that forests form a "sponge" to hold water in the Middle Hills, and release this slowly throughout the year. This narrative has influenced environmental policy in various countries, and formed part of general watershed policies that link upland agriculture to lowland impacts (also see the discussion of watersheds). Indeed, the ecologist Norman Myers (1986, p. 2) wrote:

> "The Himalayan forests normally exert a sponge effect, soaking up abundant rainfall and storing it before releasing it in regular amounts over an extended period. When the forest is cleared, rivers turn muddy and swollen during the wet season, before shrinking during drier periods... Flood disasters are becoming more frequent and more severe."

Much research, however, has questioned many of the assumptions in the Himalayan theory (see Thompson *et al.*, 1986; Ives and Messerli, 1989; Ives, 2004). First, not all empirical work supported the basic premises that erosion and deforestation were occurring at unprecedented rates. One survey revealed that the estimations of deforestation rates in Nepal between 1965 and 1981 varied by a factor of 67, even after excluding some apparent typing errors (Donovan, 1981). This variety indicates that biophysical processes are more complex than commonly claimed. Moreover, it suggested that different actors and organizations represented environmental change in terms to suit their own perceptions and objectives.

Second, various research has questioned whether the metaphor of the forest as "sponge" is accurate, as trees form important sources of water demand. Some authors have suggested the metaphor of "pump" is better (Hamilton and Pearce, 1988; Bruijnzeel, 2004). In order to save space, the discussion of the impacts of upland land-use change on water supplies is discussed in the later section concerning deforestation and reforestation.

Third, research from anthropologists and cultural ecologists suggested that biophysical processes were not always experienced as problematic by farmers, who either knew how to lessen them, or who developed mechanisms to protect themselves against them. For example, one study found that some farmers triggered landslides themselves in order to revitalize soil fertility and facilitate the construction of terraces (Kienholz *et al.*, 1984). The assumption that population growth encouraged cultivation of steep slopes was also challenged by research in Nepal and Thailand, which indicated that upland farmers actually avoided steeper slopes because they knew this would accelerate erosion (Thapa and Weber, 1995; Forsyth, 1996). Studies like these showed the diversity of framings of environmental problems (or "problem closures"), and the need to consider both local evaluations of, and reactions to, biophysical changes.

And fourthly, geomorphologists argued that many processes of so-called degradation were probably more influenced by long-term geological or climatic factors that predated agriculture. For example, much mass wasting on steep Himalayan slopes can be attributed to tectonic uplift. Many deep gullies – or so-called *pahiros* – incising Himalayan slopes, reflect long-term erosion. Agricultural practices will inevitably have some impact on these processes, but it is unclear how far agriculture in itself is the sole cause of erosion (see also Gilmour *et al.*, 1987).

Of course, none of these criticisms suggest that there are no environmental problems in the Himalayas, or that upland agriculture and population pressure will have no impacts. But increasingly, analysts are arguing that we need to understand the social and political factors that led to the idea of "one" single model of environmental degradation, and how this reflected western concerns about the ecological crisis during the 1960s and 1970s, rather than acknowledging the diversity of local perceptions and the complexity of biophysical processes (Gyawali, 2000; Ives, 2004).

DRYLANDS AND WATER SCARCITY: DESICCATION IN THE SAHEL

Desiccation in the context of land use refers to the drying out of land, including the possible reduction in rainfall, resulting from human actions (see Saberwal, 1997). It is usually related to the similar concept of desertification, which is land degradation in drylands. Public perceptions of desiccation and desertification usually include images of fragile land being rapidly degraded by cultivation, grazing, and urbanization. The cofounder of the Worldwatch Institute, and well-known environmentalist, Brown (2001, p. 8) wrote:

> "Easily a third of the world's cropland is losing topsoil at a rate that is undermining its long-term productivity. Fully 50% of the world's rangeland is overgrazed and deteriorating into desert."

Public perceptions of human despoliation of dryland ecosystems are not new. Scholars in the eighteenth century, for example, considered the Sahara desert to have been created by the Romans and Phoenicians as the result of deforestation, overgrazing, and overcultivation (Goudie, 2000). Such beliefs were strengthened by the apparent collapse of local empires in North Africa. In 1324, it was reported that the Emperor of Mali, Mansu Musa crossed the Sahara to Mecca with 500 slaves and 100 camels laden with gold (Bass, 1990, p. 13). The caravan's arrival en route in Egypt depreciated the precious metals market there by 12%, and spread rumors of the fabulous wealth of the empire's capital in Timbuktu. The empire declined, however, as the result of competition from new Portuguese and Spanish empires, and in 1738 half the population of Timbuktu died of famine. When the city was visited in 1828 by a French traveler, he wrote graphically of his shock at finding apparent evidence of human failure in a barren land:

> "I looked around and found that the sight before me did not answer my expectations… [The city] presented, at first view, nothing but a mass of ill-looking houses, built of earth. Nothing was to be seen in all directions but immense plains of quicksand of a yellowish-white color. The sky was a pale red as far as the horizon, all nature wore a dreary aspect; and the most profound silence prevailed; not even the warbling of a bird was to be heard" (Caillié, 1830, in Bass, 1990, p. 13).

In time, such views led to the narrative that local land uses, and particularly agricultural intensification and grazing, were responsible for desertification. This was especially true in the Sahel, or the strip of land immediately south

of the Sahara, and comprising some of Africa's poorest countries such as Chad and Burkina Faso, and which has experienced sever drought since the 1970s. One British colonial administrator wrote about East Africa in 1937:

> "Anyone possessing some knowledge of the desert-country types can come and study the stages, quite sufficiently clear-cut once the eye is attuned to discerning them, by which the desert has through the centuries, assisted by man (sic), advanced over rich and fertile regions" (Stebbing, 1937, p. 1).

And consequently, many land-management policies have sought to control the physical processes of desiccation – such as by using fences to restrict sand movement – or restrict activities such as grazing. But again, the narrative linking desiccation to land use alone has been questioned. As with Himalayan degradation, criticisms have highlighted the role of human impacts vis-à-vis preexisting long-term biophysical processes. At the most fundamental level, land uses have not caused the existence of deserts, but these have developed because of long-term climate patterns influenced by large volumes of rising hot air (Goudie, 2000). Other studies have pointed out that many dryland zones – such as the Namib Desert of southwest Africa – are subject to cycles of relative wet and dry periods of precipitation, and consequently these cycles need to be considered before any analysis of land-use impacts (Thomas and Middleton, 1994). Other biophysical changes are difficult to estimate. Some analysts have sought to model impacts of land use–cover change on desiccation (for example, see Sud and Fennessy, 1982; Nicholson, 1988; Xue, 1997). But these models may be questioned too. For example, it may be reasonable to propose – as some of these models do – that rainfall may decrease by 25% if (say) an acacia savanna ecosystem was replaced by grassland used for pastoralism. But there are (at present) insufficient measurements, for the critical times, to indicate whether such changes took place. Plus, such changes are reversible, and are constantly in flux (Hulme, 2001, p. 26).

Critics have also questioned the public perception of drought itself. Hulme (2001) argues that the primary problem concerning public perceptions about the Sahel is the belief in the concept of a uniform climate that is in equilibrium. Instead, the climate – and associated drought – is highly variable over space and time. This implies that different localities have diverse experiences of climatic change. Furthermore, the concept that there is necessarily a problem of drought in the Sahel implies that the region has departed from some "normal" level of precipitation that is clearly identifiable. The so-called Sahel drought became a topic of international concern in 1972 and 1973, which were the region's driest years on record. At this time, analysts tried to predict the length of the drought according to past periods of dryness. But, counter to expectations, the years 1983 to 1985 were then even drier. The long-term trend suggested further desiccation. For the period 1931 to 1960, the 30-year average Sahel rainfall was 520 mm. Between 1941 and 1970, it was 512 mm. 1951 to 1980 was measured at 488 mm and 1961 to 1990 was measured at 428 mm. The predicted level for 1970 to 2000 was 410 mm (Hulme, 2001 p. 24). On this basis, perhaps it is more important to see the current "drought" as normal, and instead seek to explain how the decades of the 1920s, 1930s, and 1950s were comparatively wet.

The implications of these concerns is to place less emphasis on "desertification" or "desiccation" as identifiable processes with known causes and effects, and instead to see each as more varied and governable in different ways. For example, recent research has suggested that increasing land use, combined with dust storms, may be responsible for maintaining high levels of atmospheric dust over the Sahel, which may be responsible for reducing rainfall. Indeed, the growth in off-road vehicles may disrupt surface crusts and lichens and hence release dust to the air (Goudie, 2004). Similarly, social science researchers have urged more attention to the socioeconomic causes of vulnerability to drought within Sahel communities (Batterbury and Warren, 2001). For example, it may be more effective to encourage local technologies of rainwater harvesting as a response to drought, rather than to address only the broad causes of drought. In both these physical and socioeconomic cases, the approaches allow more flexibility in explaining environmental problems in the Sahel, and allow research to identify both the causes and responses to desiccation in more locally grounded ways. This does not mean suggesting drought or desiccation do not occur, or that they are not problematic. But it means we should replace simple narratives focusing on single causes with those that incorporate greater awareness of the narrative's history, and more local participation about current experiences and exposures to risk.

DEFORESTATION AND REFORESTATION IN WATERSHED ZONES

The examples of Himalayan degradation and desiccation in the Sahel are examples of narratives linked to specific locations. But the impacts of deforestation and reforestation in watershed areas are equally controversial and more general. Public perceptions in various countries attribute various land-use impacts on water resources, notably that deforestation will decrease levels of rainfall, or that upland agriculture will cause lowland water shortages, especially during dry seasons. In turn, it is commonly believed that reforestation is an adequate solution to these problems.

Yet, these views are also questioned. Sometimes, critics include those who oppose forest protection in order to allow logging. But other critics instead ask whether these explanations of hydrological impacts are simplistic, and whether the proposed solutions will be effective, or whether they would unnecessarily interfere in farmers' livelihoods.

Again, this article cannot hope to discuss all aspects of these hydrological debates (see Bruijnzeel, 2004), but seeks to illustrate the coevolution of perceptions and scientific narratives. Four themes are discussed: the meanings and extent of deforestation and the impacts of deforestation or upland agriculture on rainfall, lowland water flows, and erosion.

Defining ''Deforestation''

For many analysts, "deforestation" is a narrative in itself (Fairhead and Leach, 1998). Public perceptions usually state that deforestation is universally damaging, even though the term comprises great diversity. The impacts of deforestation on hydrological resources depend on its extent, and on the local socioeconomic needs attributed to those resources.

First, public perceptions often consider deforestation a wide-scale and uniform phenomenon. Undoubtedly, in many locations, deforestation has occurred at alarming rates, and continues to do so. But our knowledge of deforestation rates is highly varied. Comparisons of satellite data and ground surveys of forests, for example, suggest that some estimates based upon satellite imagery alone exaggerate rates of forest loss because of the simplifications and assumptions used in making estimates at different spatial scales. For example, Fairhead and Leach (1998, p. 183) estimated that total forest loss in six West African countries since 1900 may reach 9.5 to 10.5 million hectares, rather than commonly discussed estimates of 25.5 to 30.2 million hectares. (Indeed, some agencies, such as the World Conservation Monitoring Center, have placed deforestation in this region even higher, at 48.6 million hectares). As with Himalayan environmental degradation, we need to acknowledge the diversity of these estimations and ask why statistical uncertainties are often not acknowledged, and, accordingly, become seen as "fact".

Moreover, there are diverse framings or classifications of "deforestation". The term often conjures images of clearfell logging, such as in the Amazon. But widespread deforestation for commercial logging should not be confused with selective logging, the harvesting of nontimber forest products, or the cyclical burning and regrowth under forms of shifting cultivation (sometimes known in more extreme forms as "slash and burn" agriculture). Under these diverse forms of forest usage, it is crucial to consider what is meant by "deforestation". Each will have different impacts on resources. As two critics wrote:

> "The generic term "deforestation" is used so ambiguously that it is virtually meaningless as a description of land-use change... It is our contention that the use of the term "deforestation" must be discontinued, if scientists, forestland managers, government planners, and environmentalists are to have meaningful dialogue on the various human activities that affect forests and the biophysical consequences of those actions" (Hamilton and Pearce, 1988, p. 75).

Similarly, there are various meanings of the word "reforestation". Afforestation is commonly meant to imply planting trees where there were previously no forests, whereas "reforestation" is planting trees in sites of previous forest. Yet, of course, it is important to distinguish between monoculture forest plantations, and carefully reseeded, long-term reconstructions of diverse ecosystems. Some outspoken public perceptions are strongly for or against plantations, either because they are seen to be effective ways of replacing forests (including for the purpose of carbon sequestration) or because they are seen to be ineffective for hydrological or biodiversity problems, and may also exclude local land users from agricultural land. Eucalyptus plantations have especially been criticized because they have been associated with commercial forestry with little impacts on local economic growth, yet their fast growth rates usually absorb much water from local fields. A full investigation into the nature and impacts of reforestation/deforestation is beyond this article, but it is important to note that these terms evoke important, and historical, meanings, and are highly varied.

Do Forests Increase Rainfall?

Many public perceptions assume that deforestation reduces rainfall. For example, a magazine for tourists in Thailand carried an article entitled "*There's no doubt – it's a drought!*" by Thammasak Thinnsawat and David Hardy, *Good Morning Chiang Mai*, vol. 4 no. 3, pp. 12–23. March 1999 which claimed: "the bottom line is forests decimated by excessive tree felling and land denuded by slash and burn agriculture severely reduce cloud formations – and thus rainfall, the main cause of Thailand's drought". Similarly, the leader of the Chipko environmental movement in India once claimed that cutting forest results in drought (World Water, 1981). The link between forests and rainfall is easy to make, because so many forests exist in rainy areas – notably rainforests and mountainous zones.

Research, however, has sought to redefine these narratives. At one level, the presence of forests will usually influence cloud formation because of higher levels of evapotranspiration and humidity than experienced under other land uses. But much research has suggested, instead, that many forest areas follow, rather than cause rainfall, and that long-term patterns of climate and topography have more influence on precipitation than forests. Moreover, the impacts of deforestation on albedo (or atmospheric reflectivity) are also considered less dramatic than originally thought because most replacement vegetation has similar reflective values.

Research into the impacts of deforestation on rainfall is usually hard to achieve because a temporally monitored experiment, using a large land area, is infeasible. Furthermore, these studies do not always use the same baseline or extent of "deforestation", or acknowledge the impacts of other climatological factors such as the El Nino

Southern Oscillation. But various studies using time-series data or simulations have suggested that there is little statistical evidence for forests causing rainfall. There are many examples. One of the earliest studies (Bernard, 1953) found no evidence for forests influencing rainfall in the Central Congo Basin (estimated one million km^2). Research in Russia, reviewed by Shpak (1968) indicated that forestland produces approximately 10% more rainfall than land without forest (in Hamilton, 1988, p. 103). In southern India, annual rainfall over the last 100 years has not decreased, even despite the general conversion of the dry deciduous forest to agriculture, although evidence suggests a decline in the number of rainy days (Meher-Homji, 1980). Tangtham and Sutthipibul (1989) found a negative correlation between 10-year moving averages of annual rainfall and remaining forest area in northern Thailand, 1951 to 1984, but a positive correlation between forest area and number of rainy days. A further study found no changes in rainfall totals or patterns on the 12 100 km^2 Nam Pong basin in northeastern Thailand between 1957 and 1995, despite a decrease in areas classified as forest from 80% to 27% since the 1970s (Wilk *et al.*, 2001). And in the Amazon basin, Global Circulation Models have suggested that removing even the entire Amazon basin would result in reduced rainfall, especially in the dry northeast of Brazil, of only an average of about 0.5 mm per day (Rowntree, 1988).

But, of course, these studies do not suggest that forests have no influence on rainfall. Some simulation studies suggest forest conversion over areas between 1000 and 10 000 km^2 may cause delays in the formation of clouds. This result was found, for example, in deforested areas of southwestern Amazonia (Cutrim *et al.*, 1995). Similarly, reductions in clouds was observed over deforested parts of Costa Rica during the dry season, but these reductions were not observed in neighboring Nicaragua in regions with good forest cover (Lawton *et al.*, 2001). Most importantly, there is strong evidence that some forests do influence rainfall in specific locations such as coastal or montane regions, where fog may coincide with tall vegetation. In such "cloud forests", cloud-water interception may account for some 5% to 20% of ordinary rainfall, and even more in exposed sites (Bruijnzeel, 2004, p. 194). In these locations, deforestation is likely to have notable impacts on local levels of precipitation and lowland supply of water.

The public perception that forests cause rainfall reflect a general trend to see deforestation as damaging to water resources, and is supplemented by popular beliefs about water supply and erosion.

Do Forests Increase Runoff?

A further public perception is that deforestation – and associated activities such as upland agriculture – may disrupt water supplies to the lowlands, notably by reducing supplies during dry seasons or by increasing flooding during peak flows. Linked to this, some organizations have called for reforestation in upland areas to prevent water shortages. Countries such as Thailand and China have also implemented logging bans partly for these reasons. One of the common beliefs underpinning these interpretations is that forests act as a "sponge" in watersheds, which hold water throughout the year. Converting the forests to agriculture may result in losing this sponge effect, and hence damage the water-holding properties of watersheds.

A variety of public experiences have confirmed or contradicted these statements. There are numerous reports within lowland sites of springs drying up after deforestation. But there are similar reports following reforestation. Some researchers have consequently argued that we need to distinguish more closely between total water yields, and the seasonal distribution of flows (Bruijnzeel, 1990). Similarly, it is difficult to generalize about land-use impacts on the basis of information from various sites, which vary in terms of underlying geology and climate, and, again, in terms of how deforestation is defined and experienced.

That said, many hydrologists agree that removing forest cover by a third or more results in significant increases in streamflow because of reductions in evapotranspiration. In effect, this is to suggest that forests do not necessarily act as "sponges" in watersheds, but as "pumps". As Bruijnzeel (1990, p. 84) wrote:

> "Removal of forest cover leads to higher streamflow totals and reforestation of open lands generally leads to a decline in overall streamflow."

And similarly, various studies suggest that upland agriculture, or limited forms of deforestation, may impact only marginally on lowland water supplies. Alford (1992) in Thailand, for example, concluded that there was no apparent connection between net runoff from river basins and shifting cultivation between 1976 and 1987 (although it is worth noting that much deforestation from shifting cultivation started in the nineteenth century, and so these results may not be fully representative). In Africa, two studies estimated the increase in water supply following the conversion of forests to crops was an estimated 140 mm per year in Nigeria (Lal, 1983) and 410 mm per year in Tanzania (Edwards, 1979). No declines in annual streamflow have been reported following lowland tropical forest removal (Bruijnzeel, 2004, p. 199). Such results have also been applied in reverse in relation to the impacts of reforestation. For example, Bosch (1979) found that afforestation of former grasslands with pine resulted in reductions in both annual streamflow (of 440 mm) and during the dry season (of 15 mm). These were supported by later research in South Africa by Scott and Smith (1997) who also found that the reduction in dry season flows were actually greater than the reduction in annual flow rates. In Kenya, water yields have been reported to return to original levels within

eight years where pine plantations replaced forests (Blackie, 1979). Bruijnzeel (2004, p. 209) wrote:

> "The conclusion that already diminished dry seas flows in degraded tropical areas may decrease even further upon reforestation with fast-growing tree species seems inescapable."

Yet, these findings are contingent upon a variety of local contexts, and particularly the protection of soils under and around forests. The case of tropical montane cloud forests generally experience low evapotranspiration, and removing these forests may also reduce the impacts of this vegetation on channeling water to the ground (see discussion of rainfall and deforestation). Different geological zones have varying runoff efficiencies (or the ratio between water input as precipitation and output as streamflow). Alford (1992), for example, estimated basins in Thailand to have a comparatively low efficiency of just 20%. This low efficiency, and the observed lack of responsiveness from rainfalls to lowland flows were attributed to the extraction of water by local irrigation and dams before they reached the main river channels (Walker, 2003). Different soil conservation measures may also enhance or reduce the ability for soils to hold water. (The topic of erosion is discussed in the Section "Do forests reduce erosion?").

Similarly, the influence of land-use changes on flooding is also controlled by local topographical and climatic factors such as rainfall events, hydraulic conductivity of soil, and slope morphology. It is therefore difficult to find clear relationships between simple criteria and flooding. In Nepal, for example, Hofer (1993) found no statistically significant relationships between river discharge, precipitation, and flooding in large watersheds between Nepal and the Gangetic plain. After analyzing some 40 years of data, he later wrote of the Ganga–Brahmaputra–Megha river system:

> "It can be inferred that floods are a normal process in the Ganga-Brahmaputra-Megha lowlands. Neither the frequency nor the magnitude of flooding has increased over the last few decades. Consequently there is no reason to believe that floods in the lowlands have intensified as a result of human impact in the highlands" (Hofer, 1998).

Earlier research in the Nepal Himalayan produced similar results at a smaller scale. Marston *et al.* (1996, p. 1) studied data at 22 stream crossings together with drainage basin morphometric data and information about forest cover to identify the controls on bank-full discharge from monsoon storms. They claimed: "results demonstrate that 82% of the variation in bank-full discharge can be explained as a function of drainage area alone; forest cover did not add explanatory power" (see also Calder and Aylward, 2002).

As with other biophysical events, the framing and perceived frequency of floods may reflect socioeconomic trends. The dramatic impacts associated with flooding may frequently be linked to the increased financial valuation of flood damage, and the trend towards locating housing and economic activities on floodplains, rather than a universal increase in the size and frequency of floods. Hamilton and Pearce (1988, p. 87) warned that public perceptions often misplace the causes of flooding:

> "These stormflow effects must not be extrapolated to support statements that appear in the press (and the misconception commonly held) that logging in upper watershed is the principle cause of serious and widespread flooding in the lower reaches of major river basins."

This article cannot hope to summarize the diverse and large literature concerning land-use impacts on water flows. Instead, the purpose is to illustrate the strength of public perceptions despite various hydrological studies that question them. These differences are discussed further after the final exploration of deforestation and erosion.

Do Forests Reduce Erosion?

It is generally agreed that severe soil erosion can seriously damage the ability of soils to hold water. Erosion also restricts the supply of nutrients and water to plants. Cultivation of steep slopes usually accelerates erosion, and the removal of tree canopies can accelerate splash erosion, at least in the short term. Many policy statements therefore state that upland cultivation is a primary cause of erosion, and that reforestation can help reduce erosion. Indeed, much statistical information about different rates of soil erosion show that sheet (or surface) erosion occurring under forests is much less than under agricultural plots.

Hydrological research does not question the general perception that erosion is problematic, or that agriculture can accelerate erosion. But there are various questions about how far upland agriculture is the primary cause of erosion, or how far reforestation is a solution. First, many authors have pointed out that much debate about erosion and watershed degradation overlooks the role of naturally occurring erosion, which can also occur under forests, although mainly in hilly areas. As discussed under Himalayan degradation, many deep gullies (called *pahiros* in Nepal) predate agriculture, and can occur in forestland. Similarly, gullies on agricultural plots in South Africa and Thailand should not be blamed on agriculture, but should instead be seen as characteristic of granite weathering (Twidale, 1982; Froehlich, and Starkel, 1993; Forsyth, 1996). A further category of erosion is mass wasting, or landslips, which have generally been linked more to climatic and geological factors (including tectonic uplift) rather than the impacts of agriculture or deforestation. Statistics that show less erosion under forests than on agricultural land are usually based on the measurement of sheet (or surface) erosion. Sheet erosion is still an important form of erosion and does reduce agricultural productivity and water-holding properties of soils. But full

measurements of erosion should also include other forms of erosion.

Second, some studies have also suggested that reforestation using plantations can accelerate erosion. Teak (*Tectona grandis*), for example, has been shown to increase splash erosion because of its large-sized leaves, and hence reforesting land with teak plantations in order to reduce erosion may have surprisingly counterresults (Calder, 1999, p. 30). Trees exposed to fire may also find that litter layers burn quickly, thus exposing soil to erosion.

Thirdly, various studies show that hill farmers have acknowledged the problem of erosion and taken steps to reduce it or avoid it. This topic also implies asking how different land users frame the problem of erosion, and this has been controversial. For example, many studies of farmers' perceptions of soil erosion have produced results that erosion is not perceived to be a problem by farmers, and this has occasionally reinforced narratives that farmers do not care about the impacts of upland agriculture on environmental problems. It is indeed likely that many farmers do not perceive impacts of erosion on watershed properties, and this may be especially true for the so-called "pioneer" form of shifting cultivators who historically relocated villages every 10 to 20 years in search of fertile ground (Harper and El-Swaify, 1988). But various social scientists have pointed out that the word "erosion" usually means different forms of degradation to different users, and that farmers are certainly more likely to perceive problems of declining soil fertility, of which erosion is a part, rather than erosion *per se*. On this basis, various studies have shown how farmers may avoid overall problems of soil degradation by adopting soil conservation measures, or organizing agriculture to avoid steep slopes (for example in Amazonia: Nortcliffe and Dias, 1988; in Papua New Guinea: Sillitoe, 1993; and in Sarawak: Hatch, 1983; Douglas *et al.*, 1993.)

Local land-use patterns and vegetation may also have an impact on how far soil erosion on slopes may result in sediment yields in rivers. Alford (1992) in Thailand, for example, found that suspended sediment transport in northern Thailand is comparable to the lowest values measured in river systems of the world to be between 44 t km^{-2} (Ping river) and 256 t km^{-2} (Nan river). For comparison, the more arid and tectonically active Hunza river of Pakistan has a sediment load of 13 200 t km^{-2}, while the Tamur river in eastern Nepal carries 5500 t km^{-2}. Moreover, the relationship between streamflow and sediment transport in Thailand was observed to be more constant on a year-to-year basis than streamflow and precipitation. Alford argued that these findings suggest that the sediment in Thailand's rivers is more likely to come from sources within the stream channel rather than erosion from slopes. This again suggests that local dams and rice fields trap both water and sediment from higher slopes, and hence that eroded material may not travel very far. Indeed, these results have been found in other cultivated slopes (Trimble, 1983).

Some studies have shown increases in sediment yield when farmers have removed logs or other barriers that act as barriers for soil movement (e.g. in West Java: Bons, 1990). The construction of roads, however, has been blamed for accelerating erosion. In Indonesia, one study suggested that rural roads were just 3% of the study area, but contributed disproportionately to the total basin sediment yield (Rijsdijk and Bruijnzeel, 1991). Similar results have also been found in Africa and South America (see Ziegler and Giambelluca, 1997).

Again, of course, the purpose of this discussion is not to suggest erosion is unproblematic or that upland agriculture does not cause erosion. The purpose is to challenge simplistic narratives that blame erosion on agriculture alone, or which suggest that reforestation is an adequate response. Frequently, actions by the state such as road construction or plantation forestry may even increase erosion. Reducing erosion from agricultural land is clearly necessary, but considering different framings and the variety of hydrological research on this theme will inevitably result in more complex and more effective explanations and policy responses.

Linking Public Perceptions and Science

Of course, there is insufficient space in this article to review all debates and information about watershed degradation. But it is clear that there are many differences between common public perceptions of environmental problems in watershed areas and scientific explanations. Why do these differences exist? What do they tell us about the relationship of science and public concern?

As discussed in the section on environmental "myths", environmental narratives arise from the unquestioned adoption of existing explanations for phenomena based on historic framings of events and partial public participation in the formulation of explanations. Some watershed narratives have clearly been constructed from the viewpoints and experiences of some, rather than all, social groups. For example, the perception that upland agriculture may cause lowland water shortages may reflect the experiences of many lowlanders that water shortages are increasing, and the framing that watershed zones should be managed to increase water supplies to the lowlands. But shortages may also occur because of growing lowland demand for water, and an alternative framing of watershed areas may also acknowledge the potential impacts of reforestation and resettlement on upland livelihoods. If a narrative is a succinct summary of cause, effect, blame, and responsibility, then it is not surprising, perhaps, that some actors may seek to find solutions for problems that do not involve changes to their activities, but instead seek to attribute causes elsewhere.

Similarly, dominant public perceptions or narratives may also emerge because of the great uncertainty and contestation of environmental policy in many locations, and the relative absence of inclusive and critical arenas for environmental policy formulation. Universal hydrological "laws", or statements of cause and effect, are difficult to achieve because of the variety of contexts in which land-use changes occur, and the diversity with which different users evaluate or frame impacts. It is difficult to generalize conclusions between different scales because of the growth in complicating factors and concurrent changes at larger scales. Furthermore, words such as "deforestation" and "upland agriculture" are often used without acknowledging the diversity they contain. Debates may be dominated by actors who fear that "deforestation" means logging, or that "agriculture" implies the most destructive forms of slash and burn, when in fact there are various less damaging forms of cultivation in-between.

The nascent character of many environmental policy-making arenas, and a relative lack of participation in the application formal expertise, may also contribute to the evolution and persistence of narratives. In rapidly developing countries such as Thailand or Nepal, environmental politics may also be dominated by new political allegiances between the state and domestic activists (see Ives, 2004). As discussed above, classically this trend has been associated with a desire of urban middle classes to perceive "wilderness" as threatened, and hence to protect forests or watersheds from the impacts of modernization or population growth. These desires may coincide with those of the state to assert ownership over forestland, or to place controls on mountainous or watershed zones for other reasons such as national security. As discussed in the section on "myths", these alliances may "discourse coalitions" in favor of reforestation and restrictions on agriculture, despite evidence questioning the hydrological assumptions of these policies. The lack of local arenas for discursive governance (where both local experiences and framings of environmental problems can occur), or the domination of scientific discourse by state agencies, may strengthen narratives.

But it is important to note that criticizing these narratives is not the same as suggesting that there should be no concern about watersheds, or that upland agriculture has no impact. As with debates about Himalayan degradation and dryland desiccation, the objective has been to assist scientific explanation by showing the simplicity of the narratives, and identifying more relevant and locally determined targets of research. This involves diversifying public participation in research and policymaking. Bruijnzeel (2004), for example, suggests that the debate about tropical forests and hydrology risks focusing too much on questions concerning forest protection, rather than highlighting the important supplemental role of soil protection. This suggestion is clearly justified for the purposes of securing lowland water supplies because evidence suggests that soils may be more accurately described as a "sponge" than forests. But protecting soils may also address upland framings of environmental problems by protecting agricultural productivity. Reframing environmental narratives to focus on new targets may allow the creation of more holistic and equitable forms of environmental policy. Protecting upland livelihoods as well as watershed properties may also reduce the socioeconomic driving forces behind some forms of agricultural expansion and deforestation. Furthermore, policymakers should investigate other explanations for water shortages, such as the role of increased water demand in the lowlands, or introducing ways to make lowland settlements less vulnerable to events such as floods.

CONCLUSION

Public perceptions of land-use impacts on water resources are often dramatic and simplistic. But it is wrong to see perceptions as "separate" from science. Instead, both public perceptions and scientific explanations coevolve historically, and many expert bodies such as government agencies or scientific organizations share these perceptions. Consequently, it may be better to call public perceptions "narratives" because they are succinct summaries of complex events that are taken as true by various actors. Certain narratives can dominate, or be seen as unquestioned "facts", when there are sufficient supporters of the explanation, or when alternative explanations are less publicized or seen to be less applicable by the people currently involved in the policymaking process.

This article has summarized three debates where such narratives exist: "Himalayan" environmental degradation, dryland desiccation in the Sahel, and impacts of deforestation and reforestation in watersheds. In these cases, complex biophysical processes involving various elements of serious land degradation have been linked to public perceptions – or narratives – of change that have been questioned by various types of hydrological research. Many physical scientists have suggested that the way to improve the accuracy of these narratives is to conduct more research, and to enhance communication between scientists, the public, and policymakers. While these actions are, of course, welcome, many social scientists suggest such strategies are insufficient because they do not acknowledge the social and political origins of narratives and their persistence. As an alternative, social scientists, and particularly science-studies scholars, argue that we should look at the mutual dependencies of science and public perceptions, and to use this knowledge to enhance hydrological research.

The first, and most important, point made under a narrative approach is that narratives serve to foreclose scientific and political debate. Believing, for example, that

upland agriculture always causes lowland water shortages (and that, consequently, upland reforestation will solve this problem) may be used to legitimize reforestation or the relocation of villages without public debate. A more critical approach would question the hydrological evidence for this presumed relationship of agriculture and water shortages, plus how far this relationship reflects framings of specific social groups (see sections above). For example, in Thailand, many government agencies, environmental activists, and media repeat the narrative that upland agriculture has to be restricted or partly replaced by upland reforestation in order to regulate dry season water flows and avoid lowland water shortages. Yet, much research suggests that reforestation (if done on a widespread level) may actually reduce flows, and that water shortages may be better addressed by managing lowland water demand. A more holistic approach might ask how far lowland water supplies, reforestation, and upland agricultural livelihoods can be enhanced concurrently. Diversifying the framing of problems – by seeing problems as both in lowland and uplands, rather than simply for the viewpoint of lowlanders – may therefore lead to more effective and more equitable solutions. Indeed, this approach may result in scientific progress by rejecting explanations that have been questioned by research, and instead refocusing research on topics that are considered more relevant. As discussed above, it may be necessary to see upland soil protection, rather than forests alone, as a means of protecting both upland livelihoods and lowland water supply.

Foreclosing of debate may also take place through the unquestioned used of language. For example, Blaikie and Brookfield (1987, p. 4) famously remarked that "one farmer's soil erosion is another's soil fertility". Of course, this statement does not suggest that erosion (or sedimentation) is unproblematic. But it shows how words such as "erosion", "deforestation", or "wilderness" evoke images of environmental cause and effect, or of the status of a particular problem. Adopting a narrative approach requires seeing the implicit meanings in these words, and the specific histories of how these meanings were identified, and with whose participation. A narrative approach acknowledges the cultural embeddedness of supposedly "factual" words and explanations in order to show how alternative framings and experiences may lead to different explanatory conclusions. There is also a need to see much more diversity – and contestation – within terms such as "deforestation" in order to indicate that this does not always imply "logging", or similarly that "upland agriculture" or "shifting cultivation" need not necessarily mean the most stereotypically destructive forms of "slash and burn".

The existence and persistence of narratives reflect the degree of public participation in formulating scientific explanations. Enhancing participation increases the range of information about biophysical events, and diversifies how events are framed. Many environmental "myths" remain unchallenged because alternative framings are not seen to be necessary, and because existing narratives do not interfere with current socioeconomic activities. For example, the old "myth" that the Sun rotates around the Earth has been challenged by science for centuries, and would make some activities – such as space travel – untenable. But for the majority of humans, there is no practical need to challenge this myth, and most people still say "the Sun rises in the morning", even though this statement is clearly inaccurate. In similar ways, many narratives of watershed degradation are unchallenged by lowlanders because they have not yet considered alternative framings of watershed policies from people living in highlands. But the increasing consideration of alternative framings, and the realization that old narratives do not explain growing problems such as water shortages, may gradually lead to a reevaluation of popular explanations. Indeed, some countries such as South Africa and New Zealand have replaced the old narratives of watershed degradation within water policies, and have adopted more diverse and holistic approaches to integrated forest and water management (Calder, 1999). It is worth asking why some countries have adopted this approach when others have not.

Of course, there are some important caveats. Adopting a narrative approach does not imply replacing one simplistic narrative with another. Bruijnzeel (2004), for example, has suggested that debates about upland watershed degradation are commonly divided between narratives for or against forest protection, and consequently, criticizing orthodox watershed explanations may enhance logging. This statement is misplaced. As discussed above, a narrative approach aims to show the social or political influences on any scientific explanation, and does not seek to dismiss forest protection or any other choice as policy options. Moreover, Bruijnzeel's statement confuses normative positions and statements of causality. Forest protection can be justified normatively on various social, political, or ecological reasons. But it is becoming increasingly difficult to justify forest protection on behalf of many of the watershed narratives discussed above. As Hamilton and Pearce (1988, pp. 92–93) noted some years ago, "Banning forest product harvesting on the basis of the basis of the harmful soil and water consequences of 'deforestation' is aiming at the wrong target". Rather than clinging onto such scientific explanations as justifications for normative positions, policymakers should ask how they can still achieve forest protection in more diverse and transparent ways.

That said, the analysis of narratives remains a politicized activity. The political implications of adopting or opposing environmental narratives need to be acknowledged. Many authors have argued that narratives persist for self-interested reasons, and some critics have suggested that criticizing narratives may also be to legitimize agendas. Roe

(1991), for example, claimed narratives allow development agencies to predefine "problems" in order to demonstrate "successful" aid work. Similarly, Thomas and Middleton (1994) suggested the government of Chad used narratives of desertification during the 1980s to avoid implementing democratization. Others have suggested state forestry departments also enforce narratives as ways to maintain control over forests (e.g. Fairhead and Leach, 1998). Pereira (1989, p. 1), for example, wrote:

> "The worldwide evidence that high hills and mountains usually have more rainfall and more natural forests than do the adjacent lowlands has historically led to confusion of cause and effect. Although the physical explanations have been known for more than 50 years, the idea that forests cause or attract rainfall has persisted. The myth was created more than a century ago by foresters in defense of their trees... The myth was written into the textbooks and became an article of faith for early generations of foresters."

Allegations like these should be made with are: they should not be interpreted as personal attacks on individuals, and many professionals acknowledge the controversies associated with some of these narratives. Yet, there is still a need to understand how narratives may persist despite the accumulating scientific evidence against them. In turn, this requires analyzing the relationships between public perceptions and scientific advice within specific networks or institutional contexts. For example, some critics have questioned whether the "Alternatives to Slash and Burn" (ASB) initiative of the Consultative Group on International Agricultural Research (CGIAR) may – by its very name – repeat narratives and foreclose scientific debate about the impacts of this kind of agriculture (see Forsyth, 2003, p. 146). According to the ASB: "the consequences of (slash and burn) are devastating, in terms of climate change, soil erosion and degradation, watershed degradation, and loss of biodiversity" (ICRAF, 1999). As discussed above, many studies have suggested these criticisms cannot be applied to all forms of shifting cultivation. Moreover, this statement seems to frame problems of climate change and biodiversity loss within debates about agricultural smallholdings, and hence may allocate notions of "blame" and "responsibility" for these problems in ways that some critics suggest should be applied primarily to more obvious targets such as industry or cities. More research is needed on how formal expert organizations frame and help shape public perceptions of hydrological problems, and why some countries and organizations persist in using narratives and others have stopped.

FURTHER READING

Chomitz K. and Kumari K. (1996) *The Domestic Benefits of Tropical Forests: A Critical Review Emphasizing Hydrological Functions*, World Bank Policy Research Working Paper 160, World Bank, Policy Research Department, Washington, DC.

REFERENCES

Alford D. (1992) Streamflow and sediment transport from mountain watersheds of the Chao Phraya basin, northern Thailand: a reconnaissance study. *Mountain Research and Development*, **12**(3), 257–268.

Bass T. (1990) *Camping with the Prince, and Other Tales of Science in Africa*, Latterworth Press: Cambridge.

Batterbury S. and Warren A. (Eds.) (2001) The African Sahel 25 years after the great drought: assessing progress and moving towards new agendas and approaches. Special Edition of *Global Environmental Change*, **11**(1), 1–8.

Bernard A. (1953) *L'evapotranspiration Annuelle de la Foret Equatoriale Congolaise et Sone Influence sur la Pluviosite*, IUFRO Congress: Comptes Rendus, Rome, pp. 201–204.

Blackie J. (1979) The water balance of the Kericho catchments. *East African Agriculture and Forestry Journal* **43**, 55–84.

Blaikie P. and Brookfield H. (1987) *Land Degradation and Society*, Methuen, London.

Bonell M. and Bruijnzeel L.A. (Eds.) (2004) *Forests, Water, and People in the Humid Tropics: Past, Present, and Future Hydrological Research for Integrated Land and Water Management*, Cambridge University Press: Cambridge.

Bons C. (1990) Accelerated erosion due to clearcutting of plantation forest and subsequent Taungya cultivation in upland West Java. *International Association of Hydrological Science*, Publication **192**, 279–288.

Bosch J. (1979) Treatment effects on annual and dry period streamflow at Cathedral Park. *South African Forestry Journal*, **108**, 29–38.

Brown L. (2001) *Eco-Economy: Building an Economy for the Earth*, Earthscan and Earth Policy Institute: London and Washington.

Bruijnzeel L. (1990) *Hydrology of Moist Tropical Forests and Effects of Conservation: A State of Knowledge Review*, UNESCO International Hydrological Programme, Humid Tropics Programme, and the Faculty of Earth Sciences, Free University: Amsterdam, Paris.

Bruijnzeel L.A. (2004) Hydrological functions of tropical forests: not seeing the soil for the trees? In *Environmental Services and Land Use Change: Bridging the Gap between Policy and Research in Southeast Asia*. Tomich T.P., van Noordwijk M. and Thomas D.E. (Eds.) A special issue of Agriculture, Ecosystems and Environment, Vol. 104/1. pp. 185–228.

Caillié R. (1830) Journal d'un voyage à Temboctou et à Jenné, dans l'Afrique Centrale, précédé d'observations faites chez les Maures Braknas, les Nalous et d'autres peuples; pendant les années 1824, 1825, 1826, 1827, 1828. Paris, Impr. par autorisation du roi a l'Imprimerie royale, 1830.

Calder I. (1999) *The Blue Revolution: Land Use and Integrated Water Resources*, Earthscan: London.

Calder I. and Aylward B. (2002) *Forest and floods: Perspectives on Watershed Management and Integrated Flood Management*, FAO, and Newcastle, University of Newcastle: Rome.

Chomitz M. and Kumari K. (1998) The domestic benefits of tropical forests: a critical review. *The World Bank Research Observer*, **13**(1), 13–35.

Cutrim E., Martin D. and Rabin R. (1995) Enhancement of cumulus clouds over deforested lands in Amazonia. *Bulletin of the American Meteorological Society*, **76**(10), 1801–1805.

Donovan D. (1981) Fuelwood: how much do we need? *Newsletter (DGD 14)*, Institute of Current World Affairs:.Hanover.

Douglas I., Greer T., Bidin K. and Spilsbury M. (1993) Impacts of rainforest logging on river systems and communities in Malaysia. *Global Ecology and Biogeography Letters*, **3**(4–6), 245–252.

Eckholm E. (1976) *Losing Ground: Environmental Stress and Food Problems*, Norton: New York.

Edwards K. (1979) The water balance of the Mbeya experimental catchments. *East Africa Agriculture and Forestry Journal*, **43**, 231–247.

Fairhead J. and Leach M. (1998) *Reframing Deforestation: Global Analysis and Local Realities: Studies in West Africa*, Routledge: London.

Forsyth T. (1996) Science, myth and knowledge: testing Himalayan environmental degradation in Thailand. *Geoforum*, **27**(3), 375–392.

Forsyth T. (2003) *Critical Political Ecology: The Politics of Environmental Science*, Routledge: London and New York.

Froehlich W. and Starkel L. (1993) The effects of deforestation on slope and channel evolution in the tectonically active Darjeeling Himalaya. *Earth Surface Processes and Landforms*, **18**, 285–290.

Gilmour D., Bonell M. and Cassells D. (1987) The effects of forestation on soil hydraulic properties in the middle Hills of Nepal: a preliminary assessment. *Mountain Research and Development*, **7**, 239–249.

Goudie A. (2000) *The Human Impact on the Natural Environment, Fifth Edition*, Blackwell: Oxford.

Goudie A. (2004) Dust storms in the global system. *Paper Presented at the International Geographical Union Conference*, Glasgow.

Gyawali D. (2000) *Water in Nepal*, Himal Books: Kathmandu.

Hajer M. (1995) *The Politics of Environmental Discourse*, Clarendon: Oxford.

Hamilton L. (1988) Forestry and watershed management. In *Deforestation: Social Dynamics in Watershed and Mountain Ecosystems*, Ives J. and Pitt D. (Eds.), Routledge: London, pp. 99–131.

Hamilton L. and Pearce A. (1988) Soil and water impacts of deforestation. In *Deforestation: Social Dynamics in Watershed and Mountain Ecosystems*, Ives J. and Pitt D. (Eds.), Routledge: London, pp. 75–98.

Harper D. and El-Swaify S. (1988) Sustainable agricultural development in north Thailand: conservation as a component of success in assistance projects. In *Conservation Farming on Steep Slopes*, Moldenhauer W. and Hudson N. (Eds.), Soil and Water Conservation Society of America: Ankeny, pp. 77–92.

Hatch T. (1983) Shifting cultivation in Sarawak. In *Proceedings of the Workshop on Hydrological Impacts of Forestry Practices and Reafforestation*, Kamis A., Lai F.S., Lee S.S. and Abdul Rahman Mohammed D. (Eds.), Faculty of Forestry, Universiti Pertanian Malaysia: Serdang, pp. 51–60.

Hofer T. (1993) Himalayan deforestation, changing river discharge, and increasing floods: myth or reality? *Mountain Research and Development*, **13**(3), 213–233.

Hofer T. (1998) Floods in Bangladesh. A highland-lowland interaction? *Geographica Bernensia*, **G48**, 171.

Hulme M. (2001) Climatic perspectives on sahelian desiccation: 1973–1998. *Global Environmental Change*, **11**(1), 19–29.

ICRAF (International Center for Agroforestry Research) (1999) *ICRAF in Southeast Asia*, Publicity Brochure, ICRAF:.Bogor.

Ives J. (2004) *Himalayan Perceptions: Environmental Change and the Well-Being of Mountain Peoples*, Routledge: London and New York.

Ives J. and Messerli B. (1989) *The Himalayan Dilemma: Reconciling Conservation and Development*, Routledge/UNU: London.

Kienholz H., Schneider G., Bichsel M., Grunder M. and Mool P. (1984) Mapping of mountain hazards and slope stability. *Mountain Research and Development*, **4**(3), 247–266.

Lal R. (1983) Soil erosion in the humid tropics with particular reference to agricultural land development and soil management. *International Association of Hydrological Science*, **140**, 221–239.

Lawton R., Nair U., Pielke R. and Welch R. (2001) Climatic impact of tropical lowland deforestation on nearby montane cloud forests. *Science*, **294**, 584–587.

Leach M. and Mearns R. (Eds.) (1996) *The Lie of the Land: Challenging Received Wisdom on the African Environment*, James Currey: Oxford.

Marston R., Kleinman J. and Miller M. (1996) Geomorphic and forest cover controls on monsoon flooding, central Nepal Himalaya. *Mountain Research and Development*, **16**(3), 257–264.

Myers N. (1986) Environmental repercussions of deforestation in the Himalaya. *Journal of World Forest Resource Management*, **2**, 63–72.

Meher-Homji V. (1980) Repercussions of deforestation on precipitation in Western Karnataka, India. *Aceh Met. Geogph Biokl Series B*, **28**, 385–400.

Nicholson S. (1988) Land surface atmosphere interaction: physical processes and surface changes and their impact. *Progress in Physical Geography*, **12**, 36–65.

Nortcliffe S. and Dias A. (1988) The change in soil physical conditions resulting from forest clearance in the humid tropics. *Journal of Biogeography*, **15**, 61–66.

Pereira H. (1989) *Policy and Practice in the Management of Tropical Watersheds*, Westview Press: Boulder.

Rijsdijk A. and Bruijnzeel L. (1991) *Erosion, Sediment Yield and Land-Use Patterns in the Upper Konto Watershed, East Java, Indonesia, Part III: Results of the 1989–1990 Measuring Campaign*, Project Communication no 18, Konto River Project, Ministry of Foreign Affairs.

Roe E. (1991) Development narratives, or making the best of blueprint development. *World Development*, **19**(4), 287–300.

Rowntree P. (1988) Review of general circulation models as a basis for predicting the effects of vegetation change on climate. In *Forests, Climate and Hydrology: Regional Impacts*, Reynolds E. and Thompson F. (Eds.), Kefford Press: pp. 162–193.

Saberwal V. (1997) Science and the desiccationist discourse of the 20th Century. *Environment and History*, **3**, 309–343.

Schwarz M. and Thompson M. (1990) *Divided We Stand: Redefining Politics, Technology and Social Choice*, Harvester Wheatsheaf: Hertfordshire.

Scott D. and Smith R. (1997) Preliminary empirical models to predict reduction in total and low flows resulting from afforestation. *Water SA*, **23**, 145–140.

Shpak I. (1968) *Effect of Forest on Water Balance Components of Drainage Basins*, Academy of Sciences of the Ukraine: Kiev, (translated from Russian, 1971, by the Israel Program for Scientific Translations, Jerusalem).

Sivaramakrishnan K. (2000) State sciences and development histories: encoding local forest knowledge in Bengal. *Development and Change*, **31**, 61–89.

Sillitoe P. (1993) Losing ground? Soil loss and erosion in the highlands of Papua New Guinea. *Land Degradation and Rehabilitation*, **5**(3), 179–190.

Stebbing E. (1937) The threat of the Sahara. *Journal of the Royal African Society*, **36**, 1–35.

Sud Y. and Fennessy M. (1982) A study of the influence of surface albedo on July circulation in semi-arid regions using the GLAS GCM. *Journal of Climatology*, **2**, 105–125.

Tangtham N. and Sutthipibul V. (1989) Effects of diminishing forest area on rainfall amount and distribution in northeastern Thailand. *Paper Presented at the FRIM–IHP–UNESCO Regional Seminar on Tropical Forest Hydrology*, Kuala Lumpur.

Thapa G. and Weber K. (1995) Status and management of watersheds in the upper Pokhara valley, Nepal. *Environmental Management*, **19**(4), 497–513.

Thomas D. and Middleton N. (1994) *Desertification: Exploding the Myth*, Wiley: Chichester.

Thompson M., Warburton M. and Hatley T. (1986) *Uncertainty on a Himalayan Scale: An Institutional Theory of Environmental Perception and a Strategic Framework for the Sustainable Development of the Himalayas*, Ethnographica, Milton Ash Publications: London.

Trimble S. (1983) A sediment budget for coon creek basin in the driftless area, Wisconsin 1853–1977. *American Journal of Science*, **283**, 454–474.

Twidale C. (1982) *Granite Landforms*, Elsevier: Oxford.

Walker A. (2003) Agricultural transformation and the politics of hydrology in northern Thailand. *Development and Change*, **24**(5), 941–964.

Wilk J., Andersson L. and Plermkamon V. (2001) Hydrological impacts of forest conversion to agriculture in a large river basin in northeast Thailand. *Hydrological Process*, **15**, 2729–2748.

World Water (1981) How trees can combat droughts and floods. *World Water*, **4**, 10.

Xue Y. (1997) Biosphere feedback on regional climate in tropical North Africa. *Quarterly Journal of the Royal Meteorological Society*, **123**, 1483–1515.

Ziegler A. and Giambelluca T. (1997) Hydrological change and accelerated erosion in northern Thailand: simulating the impacts of rural roads and agriculture. *Explorations in Southeast Asia Studies: A journal of the Southeast Asian Studies Student Association*, **1**(1), 1–18.

188: Land Use and Water Quality

ANDY BAKER

School of Geography, Earth and Environmental Sciences, The University of Birmingham, Birmingham, UK

The literature investigating the relationship between land use and water quality is synthesized. It is demonstrated that although some general correlations between land use and water quality can be observed, in general the relationship is complex with correlations in individual watersheds likely to be site or regionally specific. Land cover measurements discussed include land use, land cover, ecotone, buffer strip, and indices of landscape metrics, whereas water quality issues include spatial and temporal sampling strategies and the use of nonconservative tracers. Examples of land use–water quality relationships are given for both the impacts of urbanization and the intensification of agriculture. A paucity of studies in developing countries and the need to advance models of land use–water quality relationships are highlighted.

INTRODUCTION

Water quality is affected by a combination of natural and anthropogenic factors, whose relative influence changes with both time and space. Natural factors affecting water quality include precipitation intensity and amount, river discharge, geology, and soil type, topography (slope length and gradient), and vegetation cover. Meybeck *et al.* (1989) and **Chapter 91, Water Quality, Volume 3**, **Chapter 93, Effects of Human Activities on Water Quality, Volume 3**, **Chapter 94, Point and NonPoint Source Pollution, Volume 3**, **Chapter 96, Nutrient Cycling, Volume 3**, **Chapter 97, Urban Water Quality, Volume 3**, **Chapter 98, Pathogens, Volume 3**, and **Chapter 100, Water Quality Modeling, Volume 3** in the Water Quality and Biogeochemistry section in this Encyclopedia provide detailed reviews. Most of these factors can, and have, been affected by humans; for example, changes in river discharge due to abstraction, urbanization or impounding; polluted discharges from industry, agriculture, sewerage, and so on. Many of these anthropogenic influences are part of the larger process of catchment land use or land cover change that can affect water quality in rivers and lakes, as well as downstream estuarine and coastal waters. Investigation of the relationship between land cover and water quality is particularly useful when considering diffuse source pollution. Diffuse sources of suspended sediments, pathogens, nutrients, and pesticides in agricultural land uses; and suspended sediments, pathogens, nutrients, oxygen demanders, heavy metals, oil, and road salt in urban areas are often difficult to measure. One aim of understanding land use–water quality relationships would be to be able to use land use as a proxy for water quality in rivers suffering from diffuse pollution. Another would be to use satellite-derived land cover to predict water quality in unmonitored catchments; as for example, 80% of stream miles go unassessed in the United States with respect to the requirements for Section 305(b) of the Clean Water Act (GAO, 2000; cited in Griffith, 2002).

For a relationship between land use and water quality to be observed, a number of criteria have to be met. As in any water quality investigation, the determinand has to be relatively conservative, such that an upstream land-use fingerprint is preserved at the sampling location. For example, many studies have shown that chloride concentration correlates with urbanization at all scales and at all stages of urban development, and although natural chloride inputs need to be understood from sea spray/precipitation and in some cases geology (Smart *et al.*, 2001), this has the potential to be an excellent and easily measured fingerprint of human impact in part due to its conservative nature. Sampling location and frequency have

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

to be appropriate to the size of the catchment and the land-use impact to be monitored. For example, a baseflow survey of dissolved organic matter will not obtain much useful land use–water quality information for phosphorous which is less mobile in this form and more often associated with suspended sediments, but would be appropriate for an investigation of land use–water quality trends for nitrate. Similarly, the land use has to be determined for the catchment, and this is not necessarily a trivial task. For small catchments, this can often be obtained by ground survey or map data. However, for large catchments satellite imagery such as the Landsat TM has to be used to obtain the necessary spatial coverage, and classification schemes such as the Normalized Difference Vegetation Index (NDVI) used to allocate land-use classes (see review by Griffith, 2002). However, these normally contain some significant amount of error: this is a particular issue when trying to ascertain the source of errors in models of land use and water quality during their validation.

Given these requirements, it is useful to observe that the majority of studies have occurred on low order streams in small catchments with a wide variety of land uses; few are on large catchments with a similar variety of land use. Similarly, the largest body of research is on the temperate zone in the United States and European rivers, in predominantly agricultural and/or forested regions, and with relatively stable land uses. Fewer studies have been undertaken in tropical and developing countries with rapidly changing land use. The former regions also have, in general, moved through the transition from having predominantly point source pollution problems to a situation where diffuse pollution is more important; also where fertilizers are more common and frequently overapplied compared to developing countries; and finally, where land use is substantially different.

LAND-USE AND WATER QUALITY ISSUES

Land-use Pattern and Water Quality

The location of different types of land use within a catchment is the focus of considerable research, primarily driven by the buffer strip concept. However, the impact of buffer strips (riparian zones, land-water ecotones) on water quality appears unclear (Johnson *et al.*, 1997; Sliva and Williams, 2001; Jones *et al.*, 2001; Griffith, 2002). It is not always clear when, if, or how riparian vegetation maintains water quality, or whether the land use of the entire catchment has a greater effect. For example, Johnson *et al.* (1997) show that landscape characteristics explain some water quality variations in some seasons, but that catchment land use was just as important in predicting water quality. Similarly, too few studies have investigated the relationship between landscape pattern metrics and water quality. Wear *et al.* (1998) looked at a rural–urban gradient and suggested that the most remote portion of watershed and the outer edge of urban development have disproportionate influences on water quality. Other studies, however, have found little or no importance of landscape pattern metrics. As for the general relationship between catchment land use and water quality, different landscape pattern metrics and buffer strips will work in different catchments depending on land use, topography, climate, catchment size and soil, and geology. A comprehensive review can be found in Griffith (2002).

Water Quality Sampling Strategy

Sampling location and frequency have to be appropriate to the size of the catchment and the land-use impact to be monitored. Many studies have focused on simplified experimental conditions to test the land use–water quality relationship: low order streams in small catchments with a wide variety of land uses sampled over short time periods (<1 year). However land use–water quality issues of practical concern often occur in larger catchments, which are harder to sample at an appropriate spatial and temporal resolution. Land use–water quality relationships observed at the subcatchment scale may not be up scaleable, and any compromises in spatial and temporal sampling strategy means that the land use–water quality relationship obtained will be in part a function of the sampling strategy. For example, Johnson *et al.* (1997), in the 16 317 km^2 Saginaw basin in the United States, monitored 62 sites over two years with summer and autumn sampling. Summer water quality samples showed good positive correlations between row crop land cover and nutrients, total dissolved solids and alkalinity, but autumn water quality was poorly correlated with land use in comparison and a lack of relationship between phosphorous and row crop probably due to the limited sampling strategy.

Effect of Increased Urbanization on Water Quality

The effects of increasing urban land cover can be observed on a wide range of water quality determinands. Probably the most well known spatial study is the relationship between population density and annual nitrate concentration observed by Peierls *et al.* (1991) in a review of published data for 42 globally significant rivers that accounted for 37% of freshwater input to the ocean. A log–log regression of population density against mean annual nitrate concentration was shown to have an R^2 of 76%, with no differences between tropical and other rivers. One of the best historical records of the land use–water quality relationship is that of the River Seine, where river water quality data has been measured at Paris since 1886 (Meybeck, 1998). Here sodium, chloride, potassium, sulfate, and total dissolved solids were all observed to have increased

moderately, in part from urbanization (from increased industrialization, domestic wastes and deicers). Phosphate, ammonia, and total coliforms have all increased significantly from increasing sewerage inputs. A comparison with forested headwaters of the Seine demonstrated that peri-urban streams have a two times increase in sulfate, five times increase in sodium, chlorine and potassium, and a 100 times increase in ammonia compared to the headwaters. Ren *et al.* (2003) published a 49 year record that also demonstrated a long-term trend of decreasing water quality with increased urban land cover in the Shanghai region.

On a smaller scale, similar relationships can be observed. Within urbanized catchments correlations have been observed between percentage urban land use and fecal coliforms, dissolved copper, chloride, and ammonia (Sliva and Williams, 2001), and between population density and mean *Escherichia Coli* and *Faecal Coli* (Frenzel and Couvillion, 2002) – see Figure 1. In a review of water quality data from 56 catchments in Scotland with 0–40% urban land cover, Ferrier *et al.* (2001) demonstrated significant correlations between the percentage urban area and ammonia concentration ($R^2 = 44\%$), phosphate ($R^2 = 47\%$), suspended sediments ($R^2 = 50\%$), and biochemical oxygen demand ($R^2 = 38\%$). In diverse New England watersheds, correlations between dissolved nitrate concentration and percentage urban land use ($R^2 = 68\%$) and between dissolved chloride concentration increases and road density ($R^2 = 87\%$) have been observed (Rhodes *et al.*, 2001). However, what is still unclear is the percentage urban land cover necessary such that urbanization determines water quality over all other land-use types. For example, in an investigation of the 1295 km^2 Salt Fork watershed of eastcentral Illinois, two major urban areas (5% of cover) determined the majority of soluble reactive phosphate in all seasons and explained 50% of the variance of nitrate concentration, despite a 90% row crop land cover (Osborne and Wiley, 1988).

Agricultural Intensification and Water Quality

There is a clear relationship between agricultural land use and water quality (for a review see Hooda *et al.*, 2000), especially when urban land cover percentage is low (<5%) and the agricultural activity is intensive. Nitrate is the most common water quality determinand correlating with agricultural land cover, suspended solids, and pesticides are also frequently reported to correlate with agricultural land cover. At the smallest scale (40 sites in a 7.3 km^2 watershed that is 45% forested and 52% cropped), it has been demonstrated that dissolved nitrate concentration could be predicted from land use (in rank importance): percentage area of confined animals > percentage pasture > percentage corn > percentage rotation > percentage forest (Gburek and Folmar, 1999). At a watershed scale, the Chesapeake Bay region of the United States total dissolved nitrate correlated with the percentage cropland ($R^2 = 85\%$; Jordan *et al.*, 1997) with the calculation that only 10 – 30% of the nitrogen inputs in the catchments were discharged, and that the rest must be either stored in the catchment or evolved. Ferrier *et al.* (2001) demonstrated in their Scottish catchments study that arable land cover correlated with nitrate concentration ($R^2 = 44\%$), and that improved grassland land cover correlated primarily with

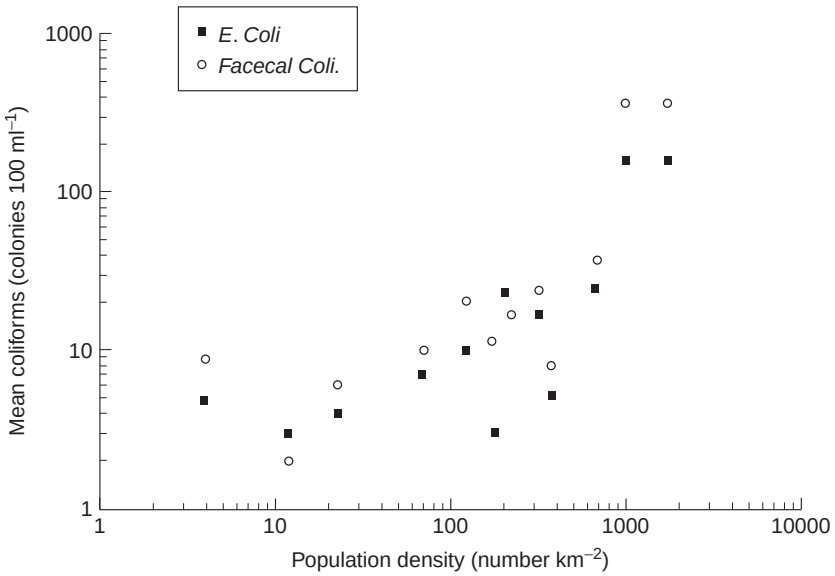


Figure 1 Graph of mean *E. Coli* and *Faecal Coli* against population density ($R^2 = 60\%$ and $R^2 = 64\%$ respectively). (After Frenzel and Couvillion (2002))

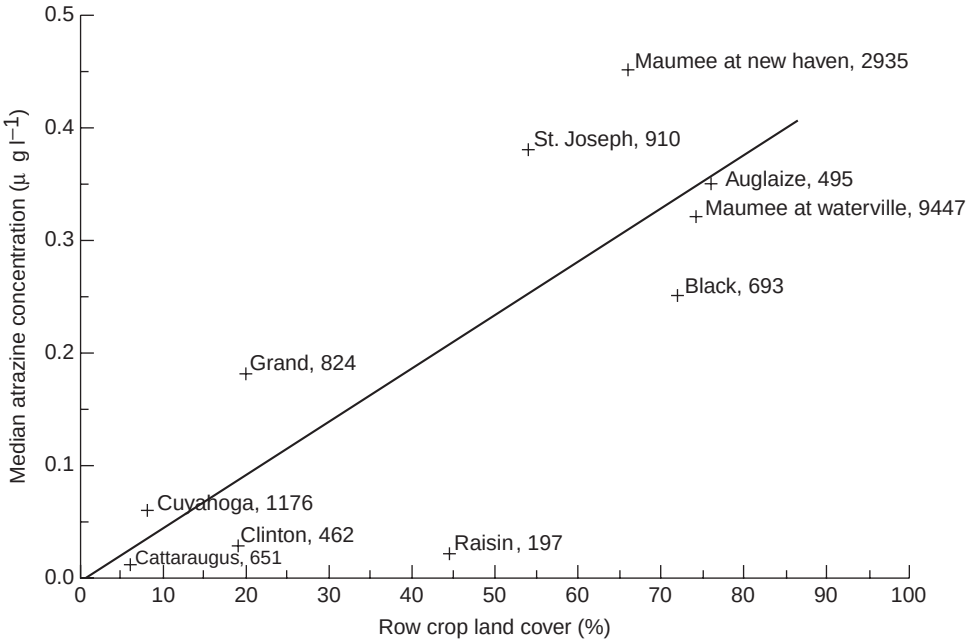


Figure 2 Relationship between atrazine and row crop percentage cover, adapted from Frey (2001). Data was collected in 1996–1998 as part of the US National Water Quality Assessment Program. Median concentrations are based on 24–36 samples per site over a wide range of flow conditions. Figures are the drainage areas in km^2

phosphate ($R^2 = 44\%$), and with weaker correlations with biochemical oxygen demand ($R^2 = 27\%$) and suspended solids ($R^2 = 27\%$). An example of the pesticide–land use relationship is given in Figure 2, from the Maumee River in Northeast United States. With subcatchments from 342 to 16 400 km^2, and where the primary land use is row crop (soybean and corn; 74%), herbicide concentrations were highest where land use was row crop, then next highest where land use was mixed-use and urban land, and finally lowest where there was pasture or forest (Myers *et al.*, 2000; Frey, 2001).

The relationship between water quality and agricultural land use, particularly as agricultural intensification occurs, are less well understood. In two small relatively unimproved agricultural catchments in upland United Kingdom, improved grassland was shown to correlate positively with mean fecal indicators at high discharge, whereas there was a negative correlation between fecal indicators and percentage unimproved land cover (moor land and rough grazing) (Crowther *et al.*, 2002). However, a paucity of published studies in developing countries (see Section "Land use–water quality studies in developing countries") limits our understanding of temporal trends in the agriculture land use–water quality relationship during intensification in these countries. Any shift towards organic farming, leading to a less intensive form of production, is likely to increase investigations of the water quality–organically farmed land cover relationship.

Land Use–Water Quality Studies in Developing Countries

The land use–water quality relationship needs investigation over a wider range of land uses and climate zones. A large body of research exists on temperate zone North American and European rivers, in predominantly agricultural and/or forested regions, the combination of which are by implication regions of relatively stable land uses. Fewer studies have been undertaken in tropical and developing countries, often with more rapidly changing land use. North American and European regions also have in general moved through the transition from having predominantly point source pollution problems to a situation where diffuse pollution is more important; fertilizer use is more common and frequently overapplied when compared to developing countries; and have land uses that are substantially different from tropical and arctic regions. Studies in temperate regions suggest that nitrate export has increased by 3 times to 20 times since industrialization owing to increased animal and human wastes and increased runoff (Peierls *et al.*, 1991). Large-scale land-use changes are occurring in tropical countries; will tropical deforestation lead to a similar phenomenon when tropical rivers are naturally high nitrate transport environments (Downing *et al.*, 1999)?

Two studies highlight the land use–water quality relationships observed in South America. Ometo *et al.* (2000) investigated 10–57 km^2 subcatchments in southeast Brazil that had been deforested in the 1850s and replaced by coffee

plantations and subsequently pasture and sugar cane. Strong correlations between land use and most water chemistry parameters were observed, probably due to the strong contrasts between catchments dominated by pasture or sugar cane agricultural land use, and urban versus nonurban areas. Dissolved calcium and magnesium ions, alkalinity and pH, all correlated with the addition of lime to leached soils in the catchments to improve agricultural productivity. Dissolved sodium and chloride concentrations correlated with urban land cover: given the lack of road salt used this is postulated to be from sewage sources. Nitrate showed no correlations with land use due to a lack of fertilizer use. Biggs *et al.* (2002) investigated part of the Amazon basin where 22% has been deforested with ~50% of deforested area currently in regrowth. A sampling strategy of sampling once at each of high and low flow at 49 sites from watersheds ranging in area from 19 to 12 500 km^2 was used. Sodium, chloride, and sulfate were all observed to increase with urbanization, with 40–58% of dry season and 83–89% of wet season chloride determined to be from urban sources when population is >250 per square kilometer. The effect of deforestation on water quality was shown to depend on soil type and geology, due to long hill slopes and thick soils which buffer the effects of this land-use change and a sampling strategy too long after deforestation and with inadequate temporal sampling.

CONCLUSION

The land use–water quality relationship is complex and, in general, any correlation observed in one catchment is likely to be site or region specific. However, some water quality determinands and land covers do show reproducible trends. Human influence in general, often determines chloride concentration (for example, see Herlihy *et al.*, 1998). Agricultural land use often determines nutrient concentrations in intensive agricultural regions, and if liming is widespread then lime-derived anions and cations can be observed to correlate with land use in both intensive and extensive agriculture. In general, as soon as urban land use increases beyond a small percentage of total land cover (~5%; the precise amount requires further research), then urban land use impacts on water quality dominate over agricultural. Undisturbed forest cover in all cases appears to have relatively little impact on river water quality compared to other land uses, despite the demonstration of enhanced nitrogen loss from some forests. However, land use is forever changing, and future land cover changes will be driven by the interplay between population growth, climate change, and policy makers, which will vary in different parts of the world. Understanding human land use and natural vegetation adaptations to climate change, together with river response, are essential; for example, see the assessment of the relative influences of land use, climate, and discharge changes on water quality in a Pennsylvanian river (Interlandi and Crockett, 2003). This will be achieved by improved understanding, and modeling, of the processes driving the observed land use–water quality relationships. Current models typically have a strong empirical component, either (i) using land use to calibrate empirically derived rating curves of water quality versus discharge to a mechanistic river flow model (such as SPARROW; Alexander *et al.*, 2004) or (ii) using Bayesian neural networks (such as Ha and Stenstrom (2003) who classify land use from water quality data). Both modeling approaches are limited to the watershed conditions for which they were calibrated or trained. Modeling approaches such as BASINS (Tong and Chen, 2002) have greater potential to improve our understanding of land use–water quality relationships under changing land use and climate conditions, due to its physical-process-based core, its incorporated point and nonpoint source models, and a GIS interface that permits simple input of land-use data.

Acknowledgments

Jeff McDonnell, Ian Calder, Malcolm Newson, and an anonymous referee provided useful comments on early drafts of this chapter.

REFERENCES

Alexander R.B., Smith R.A. and Schwarz G.E. (2004) Estimates of diffuse phosphorous sources in surface waters of the United States using a spatially referenced watershed model. *Water Science and Technology*, **49**, 1–10.

Biggs T.W., Dunne T., Domingues T.F. and Martinelli L.A. (2002) Relative importance of natural watershed properties and human disturbance on stream solute concentrations in the southwestern Brazilian Amazon Basin. *Water Resources Research*, **38**(8), doi:10.1029/2001WR000271, Art. No. 1150.

Crowther J., Kay D. and Wyer M.D. (2002) Faecal-indicator concentrations in waters draining lowland pastoral catchments in the UK: relationships with land use and farming practices. *Water Research*, **36**, 1725–1734.

Downing J.A., McClain M., Twilley R., Melack J.M., Elser J., Rabalais N.N., Lewis W.M. Jr, Turner R.E., Corredor J., Soto D., *et al.* (1999) The impact of accelerating land-use change on the N-cycle of tropical aquatic ecosystems: current conditions and projected changes. *Biogeochemistry*, **46**, 109–148.

Ferrier R.C., Edwards A.C., Hirst D., Littlewood I.G., Watts C.D. and Morris R. (2001) Water quality of Scottish rivers: spatial and temporal trends. *Science of the Total Environment*, **265**, 327–342.

Frenzel S.A. and Couvillion C.S. (2002) Fecal-indicator bacteria in streams along a gradient of residential development. *Journal of the American Water Resources Association*, **38**, 265–273.

Frey J.W. (2001) *Occurrence, Distribution, and Loads of Selected Pesticides in Streams in the Lake Erie-Lake St. Clair Basin, 1996–1998*, Water-Resources Investigations Report, USGS, 00–4169.

Gburek W.J. and Folmar G.J. (1999) Flow and chemical contributions to streamflow in an upland watershed: a baseflow survey. *Journal of Hydrology*, **217**, 1–18.

Griffith J.A. (2002) Geographic techniques and recent applications of remote sensing to landscape-water quality studies. *Water, Air and Soil Pollution*, **138**, 181–197.

Ha H. and Stenstrom M.K. (2003) Identification of land use with water quality data in stormwater using a neural network. *Water Research*, **37**, 4222–4230.

Herlihy A.T., Stoddard J.L. and Johnson C.B. (1998) The relationship between stream chemistry and watershed land cover data in the mid-Atlantic region, US. *Water, Air and Soil Pollution*, **105**, 377–386.

Hooda P.S., Edwards A.C., Anderson H.A. and Miller A. (2000) A review of water quality concerns in livestock farming areas. *Science of the Total Environment*, **250**, 143–167.

Interlandi S.J. and Crockett C.S. (2003) Recent water quality trends in the Schuylkill River, Pennsylvania, USA: a preliminary assessment of the relative influences of climate, river discharge and suburban development. *Water Research*, **37**, 1737–1748.

Johnson L., Richards C., Host G. and Arthur J. (1997) Landscape influences on water chemistry in Midwestern stream ecosystems. *Freshwater Biology*, **37**, 193–208.

Jones K.B., Neale A.C., Nash M.S., Van Remortel R.D., Wickham J.D., Ritters K.H. and O'Neill R.V. (2001) Predicting nutrient and sediment loadings to streams from landscape metrics: a multiple watershed study from the United States Mid-Atlantic Region. *Landscape Ecology*, **16**, 310–312.

Jordan T.E., Correll D.L. and Weller D.E. (1997) Effect of agriculture on discharges of nutrients from coastal plain watersheds of Chesapeake Bay. *Journal of Environmental Quality*, **26**, 836–848-.

Meybeck M. (1998) Man and river interface: multiple impacts on water and particulate chemistry illustrated in the Seine river basin. *Hydrobiologia*, **373**, 1–20.

Meybeck M., Chapman D.V. and Helmer R. (1989) *Global Freshwater Quality, a First Assessment*, WHO and UNEP, Blackwell: p. 306.

Myers D.N., Thomas M.A., Frey J.W., Rheaume S.J. and Button D.T. (2000) *Water Quality in the Lake Erie – Lake Saint Clair Drainages*, United States Geological Survey: United States Geological Survey Circular, 1203.

Ometo J.P.H.B., Martinelli L.A., Ballester M.V., Gessner A., Krusche A.V., Victoria R.L. and Williams M. (2000) Effect of land use on water chemistry and macroinvertebrates in two streams of the Piracicaba river basin, south-east Brazil. *Freshwater Biology*, **44**, 327–337.

Osborne L.L. and Wiley M.J. (1988) Empirical relationships between land use/cover and stream water quality in an agricultural watershed. *Journal of Environmental Management*, **26**, 9–27.

Peierls B.L., Caraco N.F., Pave M.L. and Cole J.J. (1991) Human influence on river nitrogen. *Nature*, **350**, 386–387.

Ren W., Zhong Y., Meligrana J., Anderson B., Watt W.E., Chen J. and Leung H.-L. (2003) Urbanisation, land use and water quality in Shanghai 1947–1996. *Environment International*, **29**, 649–659.

Rhodes A.L., Newton R.M. and Pufall A. (2001) Influences of land use on water quality of a diverse New England watershed. *Environmental Science and Technology*, **35**, 3640–3645.

Sliva L. and Williams D.D. (2001) Buffer zones versus whole catchment approaches to studying land use impact on river water quality. *Water Research*, **35**, 3462–3472.

Smart R., White C.C., Townend J. and Cresser M.S. (2001) A model for predicting chloride concentrations in river water in a relatively unpolluted catchment in north-east Scotland. *Science of the Total Environment*, **265**, 131–141.

Tong S.T.Y. and Chen W. (2002) Modeling the relationship between land use and surface water quality. *Journal of Environmental Management*, **66**, 377–393.

Wear D., Turner M. and Naiman R. (1998) Land cover along an urban-rural gradient: implications for water quality. *Ecological Applications*, **8**, 619–630.

189: Land Use and Water Resources Under a Changing Climate

THOMAS W GIAMBELLUCA

Geography Department, University of Hawaii at Manoa, Honolulu, HI, US

Water resources are becoming increasingly scarce because of the mounting demand for water associated with population growth and economic development. The availability of and need for water resources are strongly influenced by land use. Land use and land cover change can alter the regional hydrologic cycle by changing evapotranspiration (ET), runoff, soil moisture, and perhaps even precipitation, resulting in changes in water supply and demand. In addition to changes associated with land use, climate warming may significantly influence the global hydrological cycle, leading to a variety of regional hydrological changes. Higher atmospheric CO_2 levels and changing climate will influence land cover in various ways. Conversely, many types of land cover change disrupt the global carbon cycle, contributing to changes in atmospheric greenhouse gases. The drivers of climate and land cover change will be subject to complex feedbacks involving the hydrological cycle and water resources. Hence, it is important to consider the web of interacting factors affecting water resources in the context of climate and land use change. This article reviews the effects of land cover and land use change on water supply and demand, the predicted effects of global warming on water resources, the interaction between climate change and land cover change, and the combined effects of land use and climate change on water resources.

INTRODUCTION

Water resources are becoming increasingly scarce because of the mounting demand for water associated with population growth and economic development. Global water withdrawals are projected to increase by about 35% between 1995 and 2025 (Shiklomanov *et al.*, 2000) on the basis of population growth and changes in per capita usage.

The availability of and need for water resources are strongly influenced by land use. Land use and land cover change can alter the regional hydrologic cycle by changing evapotranspiration (ET), runoff, soil moisture, and perhaps even precipitation, resulting in changes in water supply and demand. In addition to changes associated with land use, climate warming may significantly influence the global hydrological cycle, leading to a variety of regional hydrological changes. Higher atmospheric CO_2 levels and changing climate will influence land cover in various ways. Conversely, many types of land cover change disrupt the global carbon cycle, contributing to changes in atmospheric greenhouse gases. The drivers of climate and land cover change will be subject to complex feedbacks involving the hydrological cycle and water resources. Hence, it is important to consider the web of interacting factors affecting water resources in the context of climate and land use change (Figure 1). This article will review the effects of land cover and land use change on water supply and demand, the predicted effects of global warming on water resources, the interaction between climate change and land cover change, and the combined effects of land use and climate change on water resources.

LAND USE AND LAND COVER EFFECTS ON WATER RESOURCES

Land Cover Effects on Supply

As Lettenmaier *et al.* (1999) emphasize, it is important to distinguish between water resources systems and hydrological systems when examining sensitivity to change. Nevertheless, the ultimate limits to water supply are determined by the rate and steadiness with which water flows through

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

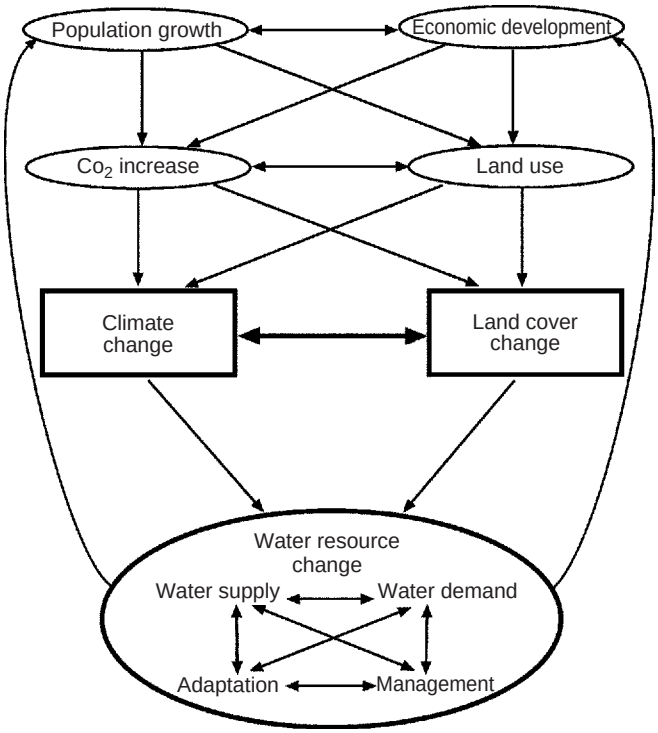


Figure 1 The effects of climate and land cover change on water resources are embedded within a web of interacting drivers. Feedbacks involving water resources may alter the drivers of change and influence future climate and land use

the hydrological cycle. Water cycling over land areas is influenced by soil and vegetation characteristics, and, therefore, is subject to modification from human activities that influence land cover (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**; **Chapter 118, Land Use and Land Cover Effects on Runoff Processes: Agricultural Effects, Volume 3**; **Chapter 119, Land Use and Landcover Effects on Runoff Processes: Forest Harvesting and Road Construction, Volume 3**; and **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**). Forest clearing, agriculture, and urbanization change the land surface characteristics and influence hydrological processes by changing energy and water flows between the surface and the atmosphere. Land cover affects how much incoming energy from the sun is absorbed by the surface, which is very important in determining how much water returns to the atmosphere via ET. Land cover change, therefore, changes the energy balance of the surface, affecting the water cycle. Deforestation, for example, alters the disposition of radiant energy by increasing surface albedo (the proportion of solar radiation reflected by the surface) and daytime long-wave radiation emission (loss of thermal infrared radiation) by the surface, resulting in lower net radiation, that is, less radiant energy retained by the surface (e.g. Bastable *et al.*, 1993; Culf *et al.*, 1995; Giambelluca *et al.*, 1997, 1999; Larkin, 2002). The characteristics and variable state of the land surface also control the proportions of net radiation used for ET and heating of the air. Replacing forest with other land covers changes energy partitioning because of differences in leaf area, aerodynamic roughness, root depth, and stomatal behavior. Field studies have verified that replacement of forest by pasture or nonirrigated crops reduces ET (Wright *et al.*, 1992; Jipp *et al.*, 1998, Giambelluca *et al.*, 2000), which results in more streamflow.

Land cover change can also alter soil hydraulic properties, resulting in reduced infiltration rates (e.g. Ziegler and Giambelluca, 1997). In drier climates, intermittent streamflow is often mainly dependent on overland flow produced when rainfall intensity exceeds infiltration capacity (e.g. Loague *et al.*, 1996), a process known as *Horton overland flow* (HOF). Reduced infiltration (increased HOF) can increase mean annual streamflow and lower the amount of water available for ET. In these environments, land cover change directly influences the water cycle that, in turn, affects the energy balance.

As Bruijnzeel (2000) emphasizes, the effects of land cover change on ET and streamflow are not uniform, and depend on the original vegetation type, characteristics of replacement land cover, climate, exposure, and soil depth. Much of the uncertainty regarding the hydrological impacts of deforestation stems from a failure to develop and use realistic land cover projections in climate simulations (Giambelluca *et al.*, 1996b), and a relative shortage of information about the hydrological characteristics of replacement land covers other than crops and pasture/grassland, especially secondary vegetation (Giambelluca, 2002).

In addition to local effects described above, land-atmosphere feedbacks may lead to regional changes in climate including rainfall. As much as 25 to 56% of the Amazon Basin rainfall is estimated to be derived from "recycling" of water evaporated within the basin (Eltahir and Bras, 1996). Basin rainfall, therefore, may decrease if evaporation is reduced by deforestation. A series of General Circulation Models (GCM) simulations (e.g. Nobre *et al.*, 1991; Henderson-Sellers *et al.*, 1996; Polcher and Laval, 1994; McGuffie *et al.*, 1995; Xue *et al.*, 1996) suggests that replacement of the entire Amazon rainforest with pasture could reduce evaporation and alter atmospheric circulation, decreasing basin rainfall by as much as 20%.

Past studies using GCMs (e.g. Henderson-Sellers *et al.*, 1993; McGuffie *et al.*, 1995) and Regional Climate Models (e.g. Kanae *et al.*, 2001; Sen *et al.*, 2004) suggest that land cover change in Southeast Asia has significant local and remote effects on the regional climate. Sen *et al.* (2004) found that the deforestation in the Indochina Peninsula increases rainfall downstream of the deforested area and

reduces it upstream during the summer months when the whole region is under the influence of the East Asia summer monsoon (EASM). Further, they find that deforestation on the peninsula causes significant changes in the EASM flow, which is the main moisture supplier to the higher latitudes over China, and thus, in its rainfall over eastern China. These findings are supported by observations of regional precipitation made during the second half of the 20th century.

Urbanization has profound effects on hydrological processes (*see* **Chapter 117, Land Use and Land Cover Effects on Runoff Processes: Urban and Suburban Development, Volume 3**) and, hence, water supply. Urban development generally involves the reduction of vegetation coverage, construction of impervious structures such as buildings and roads, compaction or removal of surface soil, and channelization of river and stream systems. These urban effects generally reduce infiltration of water and increase rates of overland flow, resulting in higher flood frequency, and increased sediment transport. Even in cases where urbanization increases groundwater recharge rates by concentrating water into pervious areas (Giambelluca, 1986), industry, transportation, and other human activities increase the risk of chemical and biological contamination, and water supply is usually negatively impacted.

Land Cover Effects on Demand

Water use is strongly tied to land use. Urban and agricultural development carry specific water demands, and are associated with certain land use changes. Therefore, we can associate land use changes with changes in water demand. Land use planning in some regions is increasingly concerned with the water demand implications of land use change. In parts of Hawaii, for example, private land development is constrained by limited water supply (Giambelluca *et al.*, 1996a). Recent expansion of residential development there was facilitated in part by demand savings from shifts of large land tracts out of irrigated sugar cultivation.

Globally, 66% of water withdrawals and 83% of consumptive use go to agriculture, mostly irrigation (Shiklomanov, 1997). As land use changes in response to continued population growth and economic development, the accompanying demands for freshwater will increase. Growth projections suggest that irrigated agriculture will have to expand in area by 20 to 30% between 1995 and 2025 to meet increasing food needs, although slowing rates of dam construction and irrigation development may limit increases in irrigated area to 5 to 10% (Cosgrove and Rijsberman, 2000). Estimations of future water withdrawal increases, taking into account likely land use changes accompanying population growth and economic development (but not the effects of climate change), range from 10.5% (Cosgrove and Rijsberman, 2000) to between 23 and 49% (Raskin *et al.*, 1997) for the period 1995 to 2025.

CLIMATE CHANGE EFFECTS ON WATER RESOURCES

Impacts on water resources may be the most important potential effect of global warming. Changes in land surface hydrology due to changing climate, such as changes in the discharge of large continental rivers, have potentially far-reaching implications for human populations. Many studies on climate change impacts on water resources have been reported (e.g. Gleick, 1999). Warming is expected to intensify the global hydrological cycle, that is, to increase global ET and precipitation (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5**). But increases in precipitation will be unevenly distributed with some regions projected to receive lower amounts as the climate warms. As a result, water supply and demand will change in different ways in different parts of the world. Some areas will experience improvements in water supply in relation to demand as a result of climate change. Other areas will see additional degradation of water resources because of warming.

Climate Change Effects on Supply

Climate change can affect water supply through changes in precipitation and ET. Changes in atmospheric circulation are predicted to result in changes in precipitation amounts, intensities, and patterns (e.g. Felzer and Heard, 1999). Trends in precipitation have been detected through analysis of historical data. In general, precipitation is increasing in the middle and higher latitudes of the Northern Hemisphere and decreasing in the lower latitudes of both hemispheres (Folland and Karl, 2001). Climate models predict that future global warming will be accompanied by an increase in global mean precipitation. Regional precipitation increases are predicted for the mid- and high-latitude land areas of the Northern Hemisphere, and over Antarctica and Africa. Regional precipitation is generally predicted to decrease in subtropical regions (Giorgi and Hewitson, 2001).

ET is affected by atmospheric demand, or potential evapotranspiration (PE), and moisture availability. Atmospheric demand is controlled by net radiation, air temperature, humidity, and wind speed. Projected air-temperature increases would generally result in higher PE, if other variables were unchanged. However, temperature changes might bring about shifts in other variables that could compensate for warming. For example, water vapor content of the air could increase, offsetting the effect of higher temperature. Cloud cover could increase, thus reducing net radiation. Studies do predict higher PE for some regions. Even where PE increases, ET will increase only if sufficient moisture is available. However, models consistently predict increases in mean global ET with greenhouse warming.

Because of increased ET, drought occurrence is predicted to increase in many regions, especially in summer, in some

cases even where precipitation will increase. Lower soil moisture and streamflows in those regions will stress water resource systems by reducing supply and increasing the demand for water.

Numerous studies have been done on the possible impacts of global warming on streamflow. Most of these studies have been based on output from GCM simulations of steady-state climate at some future condition of elevated greenhouse gases, usually double the preindustrial level. Predictions of precipitation and temperature from the model are downscaled to river basins, for which river discharge is estimated using hydrological models. More recently, several studies have been done using output from transient GCM runs, that is, estimates done using a projection of time-dependent change in greenhouse gas levels over the next 50 to 100 years. Arnell (1998) investigated how projected climate changes will affect streamflow on the European continent. The climate model output he used indicated that precipitation would increase in northern Europe and decrease in the south. Streamflow changes mirrored this regional pattern. Lettenmaier *et al.* (1999) also found that, in general, precipitation changes dominated projected streamflow changes in their study of six basins in the United States. In some basins, though, ET increases offset precipitation increases or amplified precipitation decreases. Nijssen *et al.* (2001) studied climate change implications for nine global rivers. They found that while precipitation was generally higher with warming, annual streamflow did not necessarily increase. For most tropical and midlatitude river basins, they projected annual streamflow declines because of higher ET.

The most unequivocal projections of streamflow change are seen in the middle and higher latitudes, where warming will influence both the proportion of precipitation falling as snow and the timing of snow melt. In general, these changes will result in greater flows during the winter and less in the spring.

Climate Change Effects on Demand

Climate change may influence the demand for water resources in several ways. Changes in either precipitation or ET will alter regional water budgets, affecting the demand for water, especially for agriculture. Perhaps most significant in this regard are projected changes in ET, especially in light of consistent model projections of global increases in ET rates. Many types of water use are linked to ET, especially irrigation. Because precipitation is forecast to increase in many regions, some of the effects of increased ET on water demand will be offset. Nevertheless, irrigation requirements are expected to be higher because of warming. Döll (2002) analyzed the effects of climate- change-induced shifts in growing season and changes in ET, finding that global net irrigation requirements will increase by 3 to 5% by the 2020s and 5 to 8% by the 2070s, over and above the increases due to population growth. Most climate projections suggest greater climate variability under global warming, with a higher frequency of extreme events including droughts. Long dry spells usually correspond with increases in demand, putting pressure on water resource systems. In Honolulu, Hawaii, for example, water shortages during droughts are primarily the result of increased demand, rather than reduced supply.

EFFECTS OF CLIMATE CHANGE ON FUTURE LAND COVER

Natural vegetation and agriculture will be altered by higher CO_2 levels and climatic changes. This may include both shifts in ecosystem range boundaries and structural changes within existing ecosystems. Observations and experiments show potential for climate change to disrupt natural and managed ecosystems, especially at high latitudes (Gitay, *et al.*, 2001).

Climate change effects on natural ecosystems may be difficult to distinguish from changes brought by other human influences, such as land use change. However, global warming will inevitably cause changes in vegetation in existing rangelands, forests, wetlands, and arctic and alpine systems; some of these changes are already being seen. CO_2 increases, for example, may favor C_3 plants over C_4 plants, which could give woody plants a competitive advantage over grasses in areas near current forest-grassland ecotones (Noble and Gitay, 1996). Forests will be affected by CO_2-concentration- and warming-related changes in productivity, with some regions gaining and others losing. Forests and other ecosystems may be impacted by changes in disturbance regimes, including fire, disease, and insects. Invasive species are likely to get a boost from the combined influences of climate change and increased disturbance (Loope and Giambelluca, 1998; Dukes and Mooney, 1999). Recent experience has demonstrated the importance of the interacting effects of climate and human activities in determining the impacts on natural vegetation.

The effects of global-warming-induced changes in natural ecosystems on water supply are likely to vary regionally. In areas where vegetative cover decreases, streamflow may increase because of ET reduction. However, increased sedimentation might negate such gains by lowering water quality. Conversely, where vegetative cover increases, water yield may suffer because of higher ET. Calder and Dye (2001) give evidence that invasive species tend to increase stand ET, which would reduce water supply in affected areas.

Climate change can also influence water demand through its effects on land use. For example, warming may influence crop selection that in turn may affect irrigation requirements. Higher CO_2 concentrations can influence crop water use by increasing water use efficiency

(WUE). Because plants regulate their stomata to optimize water loss and carbon uptake, increased CO_2 may translate into higher photosynthetic rates and lower transpiration for crops, that is, higher WUE (Field *et al.*, 1995; Morison, 1987). However, increases in plant growth may offset the WUE increases (Arnell and Liu, 2001), so that the impacts of higher WUE on water demand may be negligible.

HOW WILL LAND COVER–WATER RESOURCES RELATIONSHIP CHANGE WITH GLOBAL WARMING?

Given our knowledge of (i) the interactions between land cover and the hydrological cycle and the effects of those interactions on water resources, and (ii) the potential impacts of climate change on land cover and water resources, there arises a need to consider these two issues together. Most research to this point has been directed at either land use change impacts or climate change impacts. It seems obvious that we should now look at combined effects and ask questions such as: how will global warming affect the land cover–water resources relationship? Below are some issues that may be considered in this context.

Deforestation continues to be a dominant cause of land cover change, especially in developing countries in the tropics (FAO, 2003). Deforestation profoundly affects water supply by reducing ET and increasing overland flow and sedimentation. Much of the deforested land in the tropics is not permanently converted to alternative land uses, especially where land is cleared for shifting agriculture. Once abandoned, this deforested land experiences secondary vegetation succession leading eventually, in the absence of repeated clearing or other disturbances, to the development of vegetative cover resembling forest (see e.g. Moran *et al.*, 1994). The hydrologically sensitive land cover characteristics of deforested areas gradually revert back to those of forest as secondary vegetation matures (Giambelluca, 2002). The productivity of natural vegetation is forecast to increase in some regions because of increasing atmospheric CO_2 concentrations and warming (see e.g. Bugmann, 1997). We might infer, therefore, that the growth of secondary vegetation in some deforested areas will be faster, and, hence, deforestation-related ET reduction will be shorter lived. This would reduce the overall impacts of deforestation on water resources. However, in many regions, water and nutrient availability limits may actually result in lower productivity under warmer, higher CO_2 conditions. Even if deforested areas do recover faster, farmers might adapt to this change by cycling land more frequently in shifting cultivation systems.

Higher CO_2 concentrations and warming are also predicted to increase agricultural productivity in some areas (Gitay *et al.*, 2001). If this effect is realized, it may have implications for the amount of land devoted to agriculture in the future. Reduced land requirements for agriculture could have beneficial effects on water resources by reducing impacts on water supply and demand. Any gains in this regard, however, are likely to be overwhelmed by an increasing demand for agricultural products as the global population grows.

Adaptations by farmers to the changing environment will have important consequences for water resources. As mentioned, shifting cultivators might reduce the fallow period in response to higher growth rates. In areas where climate becomes drier, because of precipitation and ET changes, farmers may have to choose between using more irrigation and adopting drought resistant crops. Each of these choices would impact regional water resources.

Land management will certainly be affected in many ways by the problems of increasing greenhouse gases and global warming. Governments and industry are likely to invest heavily in afforestation and reforestation efforts as a means of sequestering atmospheric carbon. International negotiations on reducing emissions and stabilizing atmospheric levels of greenhouse gases make clear that future agreements will involve a large role for carbon credits through increasing terrestrial biomass. By increasing the forest cover, however, water resources may be negatively impacted (Calder, 1999). It is virtually certain that afforestation/reforestation will result in higher ET and reduced streamflow in the affected areas. This impact is often not considered in assessments of the costs and benefits of such efforts.

Consideration of land use change effects on water supply and demand may increasingly affect future land use policy (Giambelluca *et al.*, 1996a). Municipal water agencies are increasingly using integrated land and water management approaches for long-range planning (Calder, 1999; also *see* **Chapter 185, Integrated Land and Water Resources Management, Volume 5**). Adaptations by water and land managers and water users will strongly influence the effects of future climate on water resources (Arnell and Liu, 2001).

CONCLUSION

Climate change will affect water resources on a global scale. Regional-scale effects will be variable and will often be less important than effects of land cover change. Land cover change effects can be very important at smaller spatial scales. Studies of hydrologic impacts of land cover change and climate change have largely been done separately. Clearly, both sources of hydrologic change will be important in this century, and progress in understanding and predicting impacts requires that land cover and climate change be considered together.

REFERENCES

Arnell N.W. (1998) The effect of climate change on hydrological regimes in Europe: a continental perspective. *Global Environmental Change*, **9**, 5–23.

Arnell N. and Liu C. (2001) Hydrology and water resources. In *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, McCarthy J.J., Canziani O.F., Leary N.A., Dokken D.J. and White K.S. (Eds.), Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press: Cambridge, pp. 191–234.

Bastable H.G., Shuttleworth W.J., Dallarosa R.L.G., Fisch G. and Nobre C.A. (1993) Observations of climate albedo, and surface radiation over cleared and undisturbed Amazonian forest. *International Journal of Climatology*, **13**, 783–796.

Bruijnzeel L.A. (2000) Forest hydrology. In *The Forestry Handbook*, Evans J.C. (Ed.), Blackwell Scientific: Oxford.

Bugmann H.K.M. (1997) Sensitivity of forests in the European Alps to future climatic change. *Climate Research*, **8**, 35–44.

Calder I.R. (1999) *The Blue Revolution, Land Use and Integrated Water Resources Management*, Earthscan Publications: London.

Calder I. and Dye P. (2001) Hydrological impacts of invasive alien plants. *Land Use and Water Resources Research*, **1**, 7.1–7.12.

Cosgrove W.J. and Rijsberman F.R. (2000) *World Water Vision: Making Water Everybody's Business*, Earthscan: London.

Culf A.D., Fisch G. and Hodnett M.G. (1995) The albedo of Amazonian forest and ranchland. *Journal of Climate*, **8**, 1544–1554.

Döll P. (2002) Impact of climate change and variability on irrigation requirements: a global perspective. *Climatic Change*, **54**, 269–293.

Dukes J.S. and Mooney H.A. (1999) Does global change increase the success of geological invaders? *Trends in Ecology and Evolution*, **14**, 135–139.

Eltahir E.A.B. and Bras R.L. (1996) Precipitation recycling. *Reviews of Geophysics*, **34**, 367–378.

FAO (2003) *State of the World's Forests*, Food and Agriculture Organization of the United Nations: Rome.

Felzer B. and Heard P. (1999) Precipitation differences amongst GCMs used for the US National Assessment. *Journal of the American Water Resources Association*, **35**, 1327–1340.

Field C.B., Jackson R.B. and Mooney H.A. (1995) Stomatal responses to increased CO_2 – implications from the plant to the global scales. *Plant Cell an Environment*, **18**, 124–125.

Folland C.K. and Karl T.R. (2001) Observed climate variability and change. In *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge, pp. 99–182.

Giambelluca T.W. (1986) Land use effects on the water balance of a tropical island. *National Geographic Research*, **2**, 125–151.

Giambelluca T.W. (2002) Hydrology of altered tropical forest. *Hydrological Processes*, **16**, 1665–1669.

Giambelluca T.W., Fox J., Yarnasarn S., Onibutr P. and Nullet M.A. (1999) Dry-season radiation balance of land covers replacing forest in northern Thailand. *Agricultural and Forest Meteorology*, **95**, 53–65.

Giambelluca T.W., Hölscher D., Bastos T.X., Frazão R.R., Nullet M.A. and Ziegler A.D. (1997) Observations of albedo and radiation balance over post-forest land surfaces in eastern Amazon Basin. *Journal of Climate*, **10**, 919–928.

Giambelluca T.W., Nullet M., Ziegler A.D. and Tran L. (2000) Latent and sensible energy flux over deforested land surfaces in the eastern Amazon and northern Thailand. *Singapore Journal of Tropical Geography*, **21**, 107–130.

Giambelluca T.W., Ridgley M.A. and Nullet M.A. (1996a) Water balance, climate change, and land-use planning in the Pearl Harbor Basin, Hawaii. *International Journal of Water Resources Development*, **12**, 515–530.

Giambelluca T.W., Tran L.T., Ziegler A.L., Menard T.P. and Nullet M.A. (1996b) Soil-vegetation atmosphere processes: Simulation and field measurement for deforested sites in northern Thailand. *Journal of Geophysical Research (Atmospheres)*, **101**, 25,867–25,885.

Giorgi F. and Hewitson B. (2001) Regional climate information – evaluation and projections. In *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge, pp. 583–638.

Gitay H., Brown S., Easterling W. and Jallow B. (2001) Ecosystems and their goods and services. In *Climate Change 2001: Impacts, Adaptation, and Vulnerability, Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, McCarthy J.J., Canziani O.F., Leary N.A., Dokken D.J. and White K.S. (Eds.), Cambridge University Press: Cambridge, pp. 235–342.

Gleick P.H. (1999) Studies from the water sector of the National Assessment. *Journal of the American Water Resources Association*, **35**, 1297–1300.

Henderson-Sellers A., Dickinson R.E., Durbridge T.B., Kennedy P.J., McGuffie K. and Pitman A.J. (1993) Tropical deforestation: Modeling local- to regional-scale climate change. *Journal of Geophysical Research*, **98**, 7289–7315.

Henderson-Sellers A., Zhang H. and Howe W. (1996) Human and physical aspects of tropical deforestation. In *Climate change: Developing Southern Hemisphere perspectives*, Giambelluca T.W. and Henderson-Sellers A. (Eds.), John Wiley and Sons: Sussex.

Jipp P.H., Nepstad D.C., Cassel D.K. and Reis de Carvalho C. (1998) Deep soil moisture storage and transpiration in forests and pastures of seasonally-dry Amazonia. *Climatic Change*, **39**, 395–412.

Kanae S., Oki T. and Musiake K. (2001) Impact of deforestation on regional precipitation over the Indochina Peninsula. *Journal of Hydrometeorology*, **2**, 51–70.

Larkin E. (2002) *Radiation Balance of Forested and Agricultural Sites in Pang Khum, Thailand*. Master's thesis, Geography, University of Hawai'i at Manoa: Honolulu.

Lettenmaier D.P., Wood A.W., Palmer R.N., Wood E.F. and Stakhiv E.Z. (1999) Water resources implications of global warming: a U.S. regional perspective. *Climatic Change*, **43**, 537–579.

Loague K., Lloyd D., Giambelluca T.W., Nguyen A. and Sakata B. (1996) Land misuse and hydrologic response: Kahoolawe, Hawaii. *Pacific Science*, **50**, 1–35.

Loope L.L. and Giambelluca T.W. (1998) Vulnerability of island tropical montane cloud forests to climate change, with special reference to East Maui, Hawaii. *Climatic Change*, **39**, 503–517.

McGuffie K., Henderson-Sellers A., Zhang H., Durbidge T.B. and Pitman A.J. (1995) Global climate sensitivity to tropical deforestation. *Global and Planetary Change*, **10**, 97–128.

Moran E.F., Brondizio E., Mausel P. and Wu Y. (1994) Integrating Amazonian vegetation, land-use, and satellite data. *Bioscience*, **44**, 329–338.

Morison J.I.L. (1987) Intercellular CO_2 concentration and stomatal response to CO_2. In *Stomatal Function*, Zeigler E. and Farquhar G.D. (Eds.), Stanford University Press: Stanford, pp. 229–251.

Nijssen B., O'Donnell G.M., Hamlet A.F. and Lettenmaier D.P. (2001) Hydrologic sensitivity of global rivers to climate change. *Climatic Change*, **50**, 143–175.

Noble I.R. and Gitay H. (1996) Functional classifications for predicting the dynamics of landscapes. *Journal of Vegetation Science*, **7**, 329–336.

Nobre C.A., Sellers P.J. and Shukla J. (1991) Amazonian deforestation and regional climate change. *Journal of Climate*, **4**, 957–988.

Polcher J. and Laval K. (1994) The impact of African and Amazonian deforestation on tropical climate. *Journal of Hydrology*, **155**, 389–405.

Raskin P., Gleick P., Kirshen P., Pontius G. and Strzepek K. (1997) *Water Futures: Assessment of Long-Range Patterns and Problems*, Background report for the Comprehensive Assessment of the Freshwater Resources of the World, Stockholm Environment Institute: Stockholm.

Sen O.L., Wang Y. and Wang B. (2004) Impact of Indochina deforestation on the east Asian summer monsoon. *Journal of Climate*, **17**, 1366–1380.

Shiklomanov I.A. (1997) *Comprehensive Assessment of the Freshwater Resources of the World*, World Meteorological Organization: Stockholm.

Shiklomanov I.A., Shiklomanov A.I., Lammers R.B., Peterson B.J. and Vorosmarty C. (2000) The dynamics of river water inflow to the Arctic Ocean. In *The Freshwater Budget of the Arctic Ocean*, Lewis L.L., Jones E.P., Lemke P., Prowse T.D. and Wadhams P. (Eds.), Kluwer Academic Publishers: Dordrecht.

Wright I.R., Gash J.H.C., Da Rocha H.R., Shuttleworth W.J., Nobre C.A., Maitelli G.T., Zamparoni C.A.G.P. and Carvalho P.R.A. (1992) Dry season micrometeorology of central Amazonian ranchland. *Quarterly Journal of the Royal Meteorological Society*, **188**, 1083–1099.

Xue Y., Bastable H.G., Dirmeyer P.A. and Seller P.J. (1996) Sensitivity of simulated surface fluxes to changes in land surface parameterizations – A study using ABRACOS data. *Journal of Applied Meteorology*, **35**, 386–400.

Ziegler A.D. and Giambelluca T.W. (1997) Importance of rural roads as sources areas for runoff in mountainous areas of northern Thailand. *Journal of Hydrology*, **196**, 204–229.

190: Hydromorphological Quality – A Policy Template for Channel Design in River Restoration

ANDREW RG LARGE[1] AND MALCOLM D NEWSON[2]

[1]Centre for Land Use and Water Resources Research, University of Newcastle, Newcastle upon Tyne, UK
[2]School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne, UK

The European Union's Water Framework Directive (WFD) is the most substantial piece of European water legislation to date and will have a profound effect on how water is managed in Europe over the next 25 years. The term "hydromorphological" was introduced by the WFD to describe the quality of water bodies in addition to physicochemical and ecological elements, and has rapidly become a prominent descriptor. Reference conditions are described as "having no or only very minor alterations" resulting from human activity yet, at the same time, there is considerable confusion as to how far back we have to go to approximate such conditions. A core uncertainty or ambiguity concerns the possible differences between structural definitions of a reference condition and those that relate to processes, or "functionality". Whilst ecohydrology and hydroecology are clearly critical in the provision of a science foundation for the WFD, the introduction of hydromorphological quality as a measure of the condition of fresh waters requires new (or, at least, innovative) thinking.

A momentous challenge for those involved with the concept and realization of "good ecological quality" is to assess it, monitor it, and restore it within a basin-wide strategy, considering also the element of land use and management. The potential role of fluvial geomorphology in understanding, assessing, monitoring, and restoring hydromorphological quality in river systems is stressed.

BACKGROUND

The practical application of the strategic principles of sustainable development to river basin management requires a broad range of public and private policies, including regulatory and economic instruments. It also makes heavy demands upon new forms, sources and applications of scientific knowledge and information (Newson, 2000). A key element of a more integrated approach to the development of river basin resources is "catchment consciousness", the view that land and water must be considered together, the whole basin comprising an ecosystem capable, when intact, of supporting sustainable solutions to human needs (Newson, 1997; Calder, 1999).

The European Union (EU)'s Water Framework Directive (WFD) is the most substantial piece of European water legislation to date. With the exception of heavily modified water bodies, the WFD requires all inland and coastal waters to reach "good ecological status" by 2015, and will have a profound effect on how water is managed in Europe over the next 25 years. It seeks to do this by introducing a "river basin" scale management system, within which demanding environmental objectives will be set. This Directive includes ecological targets for surface waters, came into force on 22 December 2000 (European Commission, 2000), and Member States have had three years from that date to transpose it into national law. The process is being assisted by intensive scientific, technical and legal efforts at "Common Implementation Strategies": uniformity but including subsidiarity. Of the core environmental elements of the Directive, which address surface water, groundwater, inland waters, lakes, and transitional (estuarine) waters, coastal waters, and aquifers – coverage is restricted here to rivers. Overall, the legislation is

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

innovative in that it reflects a move from human health, already legislated for in earlier European legislation (e.g. Bathing Water Directive, 76/160/EEEC, and the Dangerous Substances Directive, 76/464/EEC) to *ecosystem health* as the driver.

New Terminology – ''Hydromorphology''

The term *hydromorphological* was introduced by the WFD to describe the quality of water bodies in addition to physicochemical and ecological elements, and has rapidly become a prominent descriptor. It does not yet figure in any recognized English dictionaries (unlike the words "fluvial" and "geomorphology", which together describe the most likely scientific discipline capable of describing the hydromorphological quality of rivers). The term is defined by NTNU (2002) as "the hydrological characteristics of rivers together with the physical structure that they create". A semantic extension from "morphology" (of river channels) to "hydromorphology" has occurred because of the need to consider the many natural and anthropogenic variants of river *flow regime* as well as *fluvial geomorphology* (both processes and landforms) in the description of river physical habitat (Figure 1).

The broad interest in river physical habitat (now virtually forced upon river basin planners and managers) has been brought about by the widespread achievements in controlling chemical pollution in rivers of the developed world. Much effort has been spent in defining river "health", and Europe is no exception, as evidenced by the range of EU Directives prior to the WFD. These achievements in water quality management coincided with the early period of incorporating the ideals of *sustainable development and integrated water resources* management into basin strategies and operations (Newson, 1997; Calder, 1999). Almost independently, a third dimension, the growth and early success of a widespread *river restoration* movement, has restarted the long-running debate in ecology about what structures and/or processes constitute "natural" conditions

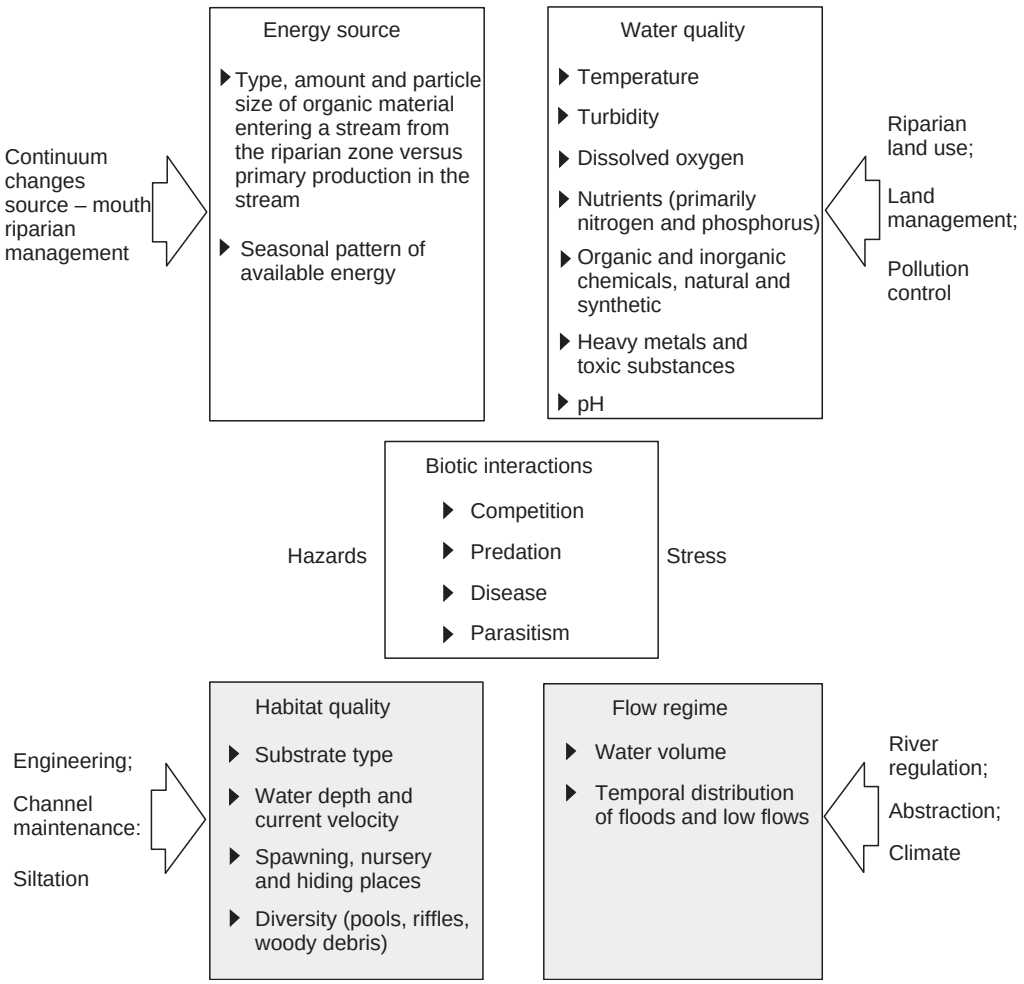


Figure 1 How river flow, sediment systems, and anthropogenic influences interact to create channel physical habitat (Reproduced from Newson and Newson, 2000 by permission of Edward Arnold)

in an ecosystem. To an extent, this debate has become one of many crises of inadequate or uncoordinated scientific support frustrating the common implementation of the WFD. Reference conditions are described as "having no or only very minor alterations" resulting from human activity yet, at the same time, there is considerable confusion as to how far back we have to go to approximate such conditions. In practice, there are extremely few such examples in existence today, and Wilby (1999) has concluded that there are few if any aquatic environments that have not been "marked by the human fingerprint".

Ecology, Hydrology, and Morphology: Interrelationships

Among the relevant scientific disciplines contributing to the hydromorphological discourse are hydrology (particularly hydraulics) and ecology. Over the last two decades, however, hydrologists have paid ever-increasing attention to the relationships between water and sediment in the catchment landscape and the ecological processes driven by its distribution and movement. In a similar way, ecologists have become more sophisticated in their appreciation of water storage and transfer processes in ecosystems. Ward *et al.* (2001) have summarized the evolution of our conceptual understanding of river systems over this period; the result of this process has been the development of two new subdisciplines termed *ecohydrology* and *hydroecology* respectively. *Ecohydrology* concerns the understanding of hydrological factors determining the natural development of wet ecosystems (Wassen and Grootjans, 1996) and river environments. For example, Large and Prach (1999) have described the ways in which plants exert a considerable influence on the hydraulic properties of channels, principally through their effect on channel roughness or friction to flow. At the same time, there is a growing belief that a deeper understanding of the functioning and significance of ecohydrological relations (e.g. Brunke and Gonser, 1997) can aid the rehabilitation or restoration of degraded riverine ecosystems.

Hydroecology, on the other hand (e.g. Amoros and Petts, 1993; Petts *et al.*, 1995), has as its basis the study of ecological and hydrological processes in rivers *and floodplains* and the development of models to simulate these interrelationships. In following this approach, hydrologists no longer regard ecological processes and the biota inhabiting fluvial hydrosystems to be static and reactive components of the hydrological landscape. For example, an increasing number of hydrologists are concerned with how flow velocities affect plant growth in channels and the relationship between flow regimes and ecological processes within channel and riparian habitats (Large and Prach, 1999).

Whilst ecohydrology and hydroecology are clearly critical in the provision of a science foundation for the WFD, the introduction of hydromorphological quality as a measure of the condition of fresh waters requires new (or, at least, innovative) thinking. This includes the deployment of all available knowledge about the flora and fauna of the water column and boundary layer, the geomorphology of bed, banks, floodplain, and the hydraulics of the full "natural" range of flows. The emerging field of *ecohydraulics* seeks to support the characterization, monitoring, and restoration of hydromorphological quality (Newson and Newson, 2000).

The WFD takes a brave political stance in setting an ecosystem concept for river basin management. The corollary is that ecosystem quality forms the basis of scaling existing conditions, monitoring change, and progressing improvements, including the broad group of activities manifesting river restoration (Figure 2). The sheer breadth of the freshwater restoration ideal creates semantic and legislative controversies; some hard choices now need to be made in the adoption by member nations of the WFD's stated objectives. The WFD required Member States to report on the impacts of human activity on all waters by the end of 2004, and much effort is

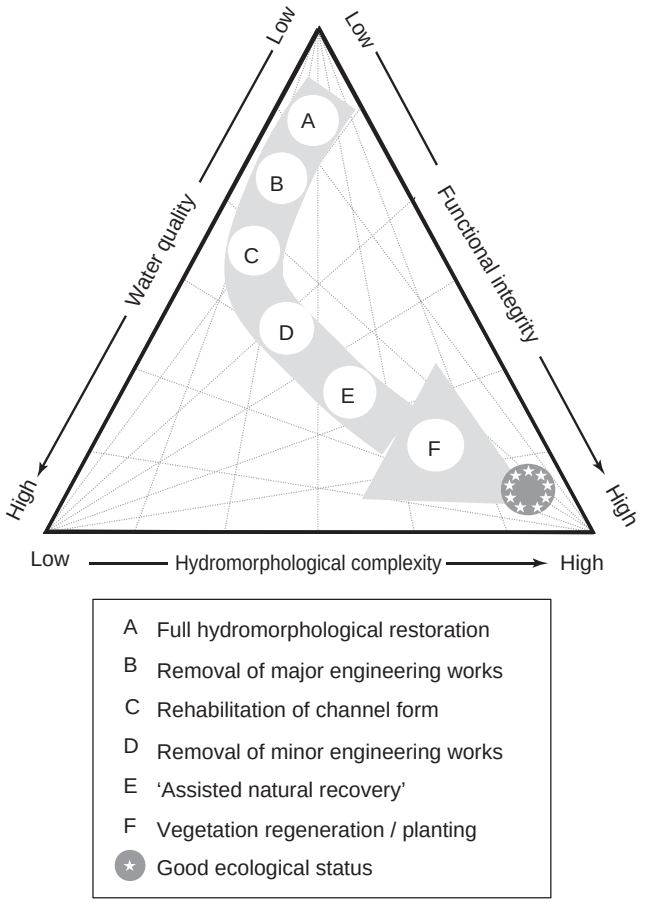


Figure 2 The ''arc of restoration'' approaches aimed towards good ecological status as defined by the EU water framework directive. Scenario A will also equate to HMWB condition

currently being directed at developing tools to measure the impact of hydromorphological changes on the ecological status of freshwaters. In order to measure and mitigate human impact, however, the notion of a natural state or other near-natural reference condition must first be defined.

DEBATES CONCERNING "NATURAL" HYDROMORPHOLOGY: REFERENCE CONDITIONS AND THEIR RESTORATION

Considerable uncertainty pervades the issues, but, in essence, the WFD judges hydromorphological quality against reference conditions that are deemed to be typical of a natural (or "little altered") river. Two definitions within Article 2 of the WFD are important in the context of water status (Chave, 2001): *artificial water body* means a body of surface water created by human activity, while *heavily modified water body* (HMWB) means a naturally occurring body of water that has been substantially changed from its natural condition as a result of *physical alterations* by human activity. As such, the latter situation has clear implications for the integration of hydromorphology into river restoration approaches.

A core uncertainty or ambiguity concerns the possible differences between *structural* definitions of a reference condition and those that relate to *processes*, or "functionality". There is also debate as to whether the definitions used should begin with *pristine* conditions, admit that *seminatural* conditions should be the starting point, or measure modifications in order to select "least modified". There are also metaethical views that advocate "natural" as meaning only the richly varied results of natural processes; Elliot (1997) uses the term *faking nature* for our often-partial attempts to keep faith with this fundamental meaning. There are clear dangers in a philosophy based simplistically upon putting river engineering into reverse gear. Geomorphologists have used the less disparaging term "mimicking" for the construction of river channel features, such as "riffles", that have the architecture of the naturally occurring form but which are designed to be stable and therefore have no function in the fluvial sediment system (e.g. Petersen *et al.*, 1992 for an example of this type of approach to stream rehabilitation). For a variety of reasons, therefore, including uncertainty about the outcome of our designs, *assisted natural recovery* has begun to gain ground as a compromise restoration strategy.

Under the WFD, the huge spatial scale of the desire to recover good ecological quality demands low-cost (or, in its place, low-maintenance) approaches. This philosophy, which intervenes only to encourage restorative fluvial processes (such as sedimentation to narrow over widened sections, and thus rehabilitate in terms of hydromorphology), has resonance with the longer timescales of concern inherent in the drive for sustainability. As Parker and Pickett (1997) suggest:

> *The goal of ecological restoration ostensibly is to return ecosystems to a state or condition from which they can be self-sustaining thereafter."* (p17)

Whatever the criteria used to classify (i) reference attributes of river systems and (ii) the success of measures put in place to rehabilitate these, the fact remains that the methodology and terminology of river protection and restoration have long been challenging and contentious. The reason, as Kentula (2000) points out, is that "success" in projects is an imprecise term that means different things to different audiences and in different situations. Kentula (2000) distinguishes between two sorts of success: *compliance success* and *functional success*. The former is determined by evaluating compliance with the terms of a regulatory framework such as the WFD, while the latter is determined by evaluating whether the "reference" ecological functions of the system have been restored (often extremely difficult to ascertain over short time periods). Realists suggest the need to consider contemporary reference conditions, chosen to view human action more inclusively because humans are bound to have an interest and involvement in restoration; few believe we are in the business of creating new wilderness! In the case of rivers, adherence to Parker and Pickett (1997) criterion may, however, still mean that catchment-scale processes remain heavily anthropogenic and the combination of modified flows and sediment supplies from upstream will create sustainable but "heavily modified" forms and hence habitats. Sear (1994) deals with the essential space/time relationships that may frustrate river restoration schemes that do not incorporate an appreciation of catchment-scale fluvial dynamics.

Selecting Reference Conditions for Hydromorphological Quality

The search for reference conditions of channel morphology and structure is frustrated by the rapidity and extent of human modifications to rivers throughout Europe. As Nienhuis and Leuven (1998) put it:

"good descriptions of pristine European rivers do not exist".

Thus, the most obvious route to achieving improved hydromorphological conditions is in mitigating the most obviously damaging impacts of basin development to date. Such an approach offers options to tackle both causative processes and their structural outcomes in river channel morphology. The process approach has the ability to cope with aspirations for "man-inclusive nature", or "nature

Table 1 The variety of terms used to describe works used to improve or maintain the hydromorphological quality of rivers and their associated habitats

Terms	Definition
Restoration	Complete structural and functional return to a predisturbance state (Perrow and Wightman, 1993)
	Return of a system to a close approximation of its condition prior to disturbance (National Research Council, 1992)
	Return to a system which closely resembles unstressed surrounding areas (Gore, 1985)
	Full, structural and functional return to a previous, natural, state (Wade *et al.*, 1993)
	The act of bringing back to a former state (Hayward and Sparks, 1990; Hawkins, 1991)
Rehabilitation	The partial return of some physical and biological components to achieve an incomplete functional return to an earlier condition (Wade *et al.*, 1993)
	Used primarily to indicate improvements of a visual nature to a natural resource: putting back into good condition or working order (National Research Council, 1992)
	Partial structural and functional return to a predisturbance state or putting back to good order (Perrow and Wightman, 1993)
	Reestablishment of character, bringing back to a good condition (Hayward and Sparks, 1990; Hawkins, 1991)
Enhancement	Any improvement of a structural or functional attribute (National Research Council, 1992; Perrow and Wightman, 1993)
	To increase in importance, quality (Hayward and Sparks, 1990; Hawkins, 1991)
	Improvements with no particular reference to a previous state (Wade *et al.*, 1993)
Creation	The birth or bringing into being new ecosystems that previously did not exist on the site (National Research Council, 1992; Perrow and Wightman, 1993)
	That which is brought into existence (Hayward and Sparks, 1990; Hawkins, 1991)
Mitigation	Actions taken to avoid, reduce or compensate for the effects of environmental damage, either potential or real (Perrow and Wightman, 1993). Among the broad spectrum of actions are those that restore, enhance, create or replace damaged ecosystems (National Research Council, 1992)
Amelioration	Avoiding damage and implies improvement is made (Perrow and Wightman, 1993)
	The act of making better, improvement (Hayward and Sparks, 1990; Hawkins, 1991)
Reclamation	The adaptation of a resource to fulfill a utilitarian human need (Perrow and Wightman, 1993)
	Putting a natural resource to a new or altered use. Often used to refer to processes that destroy natural ecosystems and convert them to agricultural or urban uses (National Research Council, 1992)
	The act of bringing back from error, to make usable (Hayward and Sparks, 1990; Hawkins, 1991)

in balance" (Lenders *et al.*, 1998), or design for "man-exclusive nature" or "nature in flux".

Table 1 summarizes commonly used definitions and the variety of terms used to describe works used to improve or maintain the hydromorphological quality of rivers and their associated habitats. All of these terms can have slightly different meanings to different people and user groups and have been interchangeable in the literature. Whisenant (1999) highlights the potential for confusion by means of a table defining 11 different descriptions for the act of "improving ecological conditions in damaged wildlands". Brooks (1995) and Brookes and Sear (1996) explore these semantic problems further, and echo the modeling approach of Cairns (1987) who advocates that, in a pure sense, the principal objective of river channel (hydromorphological) restoration should be for ecological improvement. Distinction between definitions of "restoration" and "rehabilitation" is essential, and there is little scope for supporters of "enhancement" because of its anthropocentric focus (Tapsell, 1995). Nevertheless, community-based schemes to improve the ambience of rivers and their corridors are growing in frequency, perhaps partly because of the commercial support they receive in return for adjacent investment opportunities.

Structure and Process, Catchment and Site – Further Debates

The core division of approaches to defining "natural" or excellent hydromorphological quality is that between reference conditions and the definition of natural process mechanisms. In practice, the separation is artificial. For example, physical processes of fluvial dynamics may be universal but are heavily influenced by local boundary conditions, that is, structure. However, there are also distinctions of detail between the two approaches, in that approaches *via* structural reference images tend to ignore present conditions (including anthropogenic influences on process and form). Kern (1992), in stating that rehabilitation must be based on a thorough knowledge of a river history and present condition, has outlined a number of key topics for consideration and espouses a conceptual approach representing the ideal solution for the river in question and for contemporary conditions (Table 2). This approach takes into account contemporary conditions and constraints (water rights, flood defence etc.), with the result being an optimal – not ideal – solution, as there are always political and economic aspects that influence the realization of any scheme. Lenders *et al.* (1998) suggest that reference conditions can emerge as

Table 2 Essential topics to be considered in river rehabilitation and desirable stream properties relating to the natural potential for restoration (after Kern, 1992)

Topic	Description
Land use	• Historical land use and former extend of floodplain • Present land use and future prospects • Flood defence
Floodplain habitats	• Valuable habitats (oxbows, forested wetlands etc.) • Natural capital (soil properties, groundwater levels, flora and fauna etc.)
Channel morphology and stream type	• Condition/stability of stream bed and bank • Sediment transport characteristics • Hydraulic properties
Hydrological and hydraulic data	• Alterations in discharge due to land use • Historical and present bankfull discharge
Limnology	• Physical and chemical properties • Pollution sources • Recolonization sources
Elements	**Description**
Natural stream properties	• River pattern • Morphodynamics • Natural flood dynamics • Floodplain morphology
Irreversible changes	• Abiotic and biotic factors • Changes in run-off regime • Changes in sediment transport regime • Mining of alluvial sediments • Species and habitat loss
Cultural ecology	• Loss of species diversity due to historical land use • Encouragement of biodiversity under traditional management

"palaeoreferences" (from past evidence), "actuoreferences" (from other similar sites before modification), or "system-theoretical references" (effectively, the process approach). System-theoretical or process approaches will tend to take as their conceptual base our best approximations of how the river flow and sediment systems interact to create channel physical habitat (Figure 3).

The process approach finds favor in both ecology and geomorphology because the integrity of process dynamics creates resilience and, hence, a sustainable project outcome. The concept is not far removed from *ecosystem health* (Rapport, 1995) in which ecosystem processes deliver environmental services (*sense stricto* Jewitt, 2002) such as flood protection, water supply, and water purification. Woodley *et al.* (1993) suggest that "integrity" corresponds with "self-correcting capability"; they hypothesize, however, that as ecosystems have both structural and functional integrity, the latter may not always be lost or damaged with the former because they are arranged in parallel, and, thus, resilience to all but catastrophic (or chaotic) processes is assured. The process approach also allows for *assisted natural recovery* of the system, a much cheaper form of restoration or rehabilitation.

Assessment and Monitoring "Towards Good Ecological Quality"

A final, momentous challenge for those involved with the concept and realization of "good ecological quality" is to assess it, monitor it, and restore it within a basin-wide strategy, considering also the element of land use and management. This is not the challenge faced by most community-based projects for enhancement or rehabilitation but is crucial to the river basin scale government institutions charged with sustainable development practice and implementation of the WFD. As Chave (2001) highlights, the four most important and innovative aspects of the WFD legislation are that it aims to:

- manage water as a whole *on a river basin basis*.
- use a combined approach for the control of pollution, setting both emission limits and water quality objectives.
- reflect the true cost of water, ensuring the user bears the cost of provision and use.
- involve the public in making decisions on water management.

Such an approach has been advocated for some time. Doppelt *et al.* (1993) discuss options available for the assessment of ecosystem status and suggest priorities for action at the basin scale, highlighting a prime need for data at the basin scale from which to select spatial, system-driven aspects of rehabilitation, and arguing for "a new synthesis of conservation biology, ecosystem science, and geomorphic knowledge". Key elements of the spatially planned approach include securing existing populations of desirable species, maintaining critical areas/functions, constructing functional recovery measures, and ensuring well-distributed biotic response.

In Europe, the catchment (watershed) approach is also implicit in the demands made by the EU WFD. Here, the need is for systems capable of assessing the ecological and hydromorphological quality of freshwaters. Article 8 of the WFD requires the establishment of monitoring and sampling programmes within six years of the date of entry into force of the Directive. Three types of monitoring activity are envisaged (Chave, 2001): surveillance, operational, and investigative (Table 5). Such systems must be

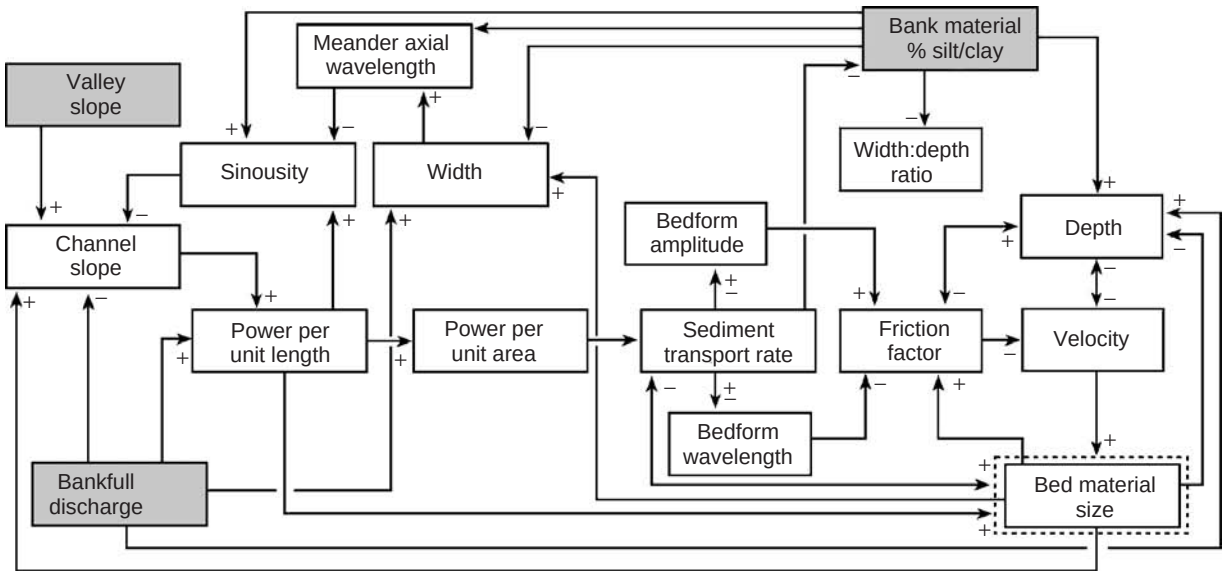


Figure 3 Systems approach to modeling the fluvial system. Independent variables are shaded, direct relationships are shown by +, inverse relationships by −. Reversible linkages are signified by double-headed arrows (From Richards, 1982, by permission of Methuen)

robust, repeatable, and harmonized; in the United Kingdom, a lead towards surveillance monitoring has been provided through the development of the River Habitat Surveys (RHS) methodology to describe and evaluate the physical characteristics of rivers (Raven *et al.*, 1997; Fox *et al.*, 1998, p458). Many European nations have analogous systems because all were aware of the impending and impressive database requirements of the WFD (Raven *et al.*, 2002). The RHS system uses survey lengths of 500 m, chosen by stratified random sampling, and including both transect-based and "sweep-up" (whole length) observations. The "hydromorphological" dimension is respected by the inclusion of both geomorphological features/dimensions and flow types (map data, valley form, channel features, bed and bank materials, vegetation, land use, profiles, dimensions, biotopes, and land management are addressed by the methodology). To date, 6000 sites have been covered by RHS for the 85 000 km of channel in England and Wales. Whilst the RHS include geomorphological variables sufficient to describe the main framework of physical habitat and the impacts thereon of channel and riparian development, the fact remains that a functional appreciation of the ecosystem in hydromorphological terms is beyond the present Survey methodology. Existing River Habitat Survey techniques have not, for example, covered the floodplain environment and physical aspects of the hydromorphology of rivers. The GeoRHS programme in the United Kingdom is aimed at enhancing RHS to develop appropriate indices to support a wide range of functional operations (e.g. flood defence, WFD, conservation, and restoration). Other innovative remote sensing techniques also offer potential in this regard (e.g. LiDAR as described by, among others, Charlton *et al.*, 2003).

Finally, into the vacuum between RHS and the preceding neglect of channel dynamics by river managers has come a range of geomorphological survey techniques, or "geomorphological audit" that allow a much greater emphasis on the often quite local processes of channel "stability" or adjustment (Newson, 2002).

THE GEOMORPHOLOGICAL BASIS FOR ''REFERENCE'' AND ''RESTORATION''

The potential role of fluvial geomorphology in understanding, assessing, monitoring, and restoring hydromorphological quality in river systems has already been stressed. There are, however, considerable limitations to the exploitation of this potential, notably in the relative lack of empirical data about channel form and process (compared, e.g. with hydrological data) and the relatively high levels of influence on channel forms and dynamics operated by site factors and the historical dimension. Accordingly, geomorphologists have been forced to adopt survey procedures and local calibrations for "universal" models in their applications to both the modified channel forms of flood-prone regions and the modified flows characteristic of dryland regions (Newson *et al.*, 2002).

Table 3 highlights the traditional preoccupation of fluvial geomorphologists with processes, centering on the composite process of "adjustment". A river channel slowly or episodically adjusts to the major drivers of river discharge (dominantly high flows, around bankfull) and fluvial

Table 3 Geomorphological description of terminology commonly used for UK rivers (after Newson, 2002)

Terminology	Suggested geomorphological description for UK rivers
Reach	Length of river in which channel dimensions and features relate characteristically to identifiable sediment sources and sinks. Reaches may be demarcated by tributary inputs under certain conditions of climate, river regulation and land use.
Stable (i.e. channel planform/elevation)	Essential to differentiate between engineering concepts of stability, legal/popular interpretation and natural resilience or ''robust'' behavior. Geomorphological stability incorporates adjustments, short of threshold behavior, that can be predicted from assessment of channel ''styles''.
Reference conditions (or ''natural'')	Rivers with planform/sectional geometry and features that represent the full interplay of water and sediment fluxes with local boundary conditions. ''Natural'' rivers are free to adjust their form and features (by aggradation/degradation and lateral migration across floodplain/valley floor) to both system-scale drivers and local conditions. ''Natural'' rivers, therefore, provide a diverse physical habitat in both space and time.
Heavily modified rivers	Rivers that, through human modification or repeated actions, are constrained in their direction/rate of adjustment and diversity of features, frequently to the extent that they create a geomorphological hiatus in the flow/sediment system, causing upstream or downstream impacts or both. Depending on system location and conditions they may recover if human action is ceased or modified.

sediment fluxes (supplied from catchment or the upstream channel). In engineering terms, adjustment is "instability", and so its constituent processes have been heavily controlled throughout the history of anthropogenic influences on rivers. In England and Wales, for example, two-thirds of all river channels sampled by the RHS have been modified in form or boundary conditions, mainly to control flooding or to reduce the risk of "instability". Many of the modifications of river channel cross sections carried out in the name of flood protection also modify the geomorphological effectiveness and work rate of the flow in that section, for example, by increasing channel depth and, hence, stream power. The flow regime, by contrast, is most heavily impacted by impoundment, regulation, and abstraction, although catchment land use, land cover, and land management are now also seen to have a subtle but important influence, especially on flow extremes.

The limitations posed to the application of geomorphological "tools" (Newson, 2002; Kondolf and Piegay, 2003a,b) by the need to conduct local surveys and calibrations is further complicated by the disparate scale of the demand for practical assessment, monitoring, and design: that is, the catchment scale. The procedures adopted by the Environment Agency in England and Wales are progressively scaled by level of detail (Newson *et al.*, 2002), but approaches elsewhere include geomorphic classification (e.g. Rosgen, 1996), hierarchical channel classification (Frissell *et al.*, 1986), and "river styles", an Australian approach also scaled through a spatial hierarchy (Brierley and Fryirs, 2000; Thomson *et al.*, 2001). River classifications attract scepticism (e.g. Wadeson, 1994; Kondolf, 1995), but Thomson *et al.* (2001) advocate the typology represented by "river styles" on the basis that it proceeds downwards in spatial scale from basin to hydraulic units, thus encapsulating system-scale information and a basin-wide long-term strategy in its local guidance. Similar approaches have been independently advocated by Van Niekerk *et al.* (1995) for South African rivers. By contrast, the Rosgen classification is of individual channel cross sections, perhaps more relevant to an engineering approach to local dynamics.

THE POLICY TEMPLATE: IMPLEMENTING AND COMPLYING WITH THE WFD

The WFD designates five classes of water body (in the case under discussion, these are typed river lengths, excluding those considered "heavily modified"), labeled as being of "high", "good", "moderate", "poor", or "bad" ecological status. These labels are based upon a comparison between actual and reference conditions for (i) biological quality elements, (ii) hydromorphological quality elements, and (iii) physicochemical quality elements. Only "high" ecological status requires reference conditions to exist for hydromorphological quality, whereas biological quality enters the description of each status. "By definition, rivers with undisturbed physical structure, and, therefore, of high hydromorphological status should, given good water quality, support aquatic communities of high ecological status" (Raven *et al.*, 2002). River Basin Management Plans are to form the strategic guidance to the achievement of "good" ecological status for all relevant water bodies by 2015. The environmental objective for heavily modified water bodies (HMWB) is to realize *good ecological potential* subject to the limitations posed by their use as a socioeconomic priority. HMWB, by definition therefore, are bodies of water that, as a result of physical alteration by human activity, are

Table 4 Hydromorphological reference conditions as specified by the WFD (after NTNU, 2002)

Component	Reference condition
Bed and bank character	Lacking any artificial in-stream and bank structures that disrupt natural hydromorphological processes and/or unaffected by any such structures outside the site Bed and banks composed of natural materials
Planform and river profile	Planform and river profile unmodified by human activities
Lateral connectivity and freedom of lateral movement	Lacking any structural modifications that hinder the flow of water between the channel and the floodplain, or prevent the migration of the river channel across the floodplain
Free flow of water and sediment in the channel	Lacking any structural modifications that affect the natural movement of sediment, water, and biota
Vegetation in the riparian zone	Having natural vegetation appropriate to the type and geographical location of the river

substantially changed in character and *cannot* meet "good ecological status".

Measurement of river ecosystem functioning requires integration of a wide variety of parameters (Lorenz, 1997), and of particular relevance here are stream water quality, hydrology (discharge and stream velocity), and geomorphology (describing the structure of the river channel, riparian zone, and floodplain), along with ecological information on species diversity and abundance. In addition to measurement of these parameters, it will be necessary to develop systems that allow for the definition of pristine (reference) conditions for rivers (Table 4), as well as defining "high", "good", and "moderate" status for hydromorphological elements insofar as they support the biotic components dependant on the parameters listed above (EPA, 2002a; NTNU, 2002). The latter is a requirement of the WFD allowing the quality of rivers to be compared in an equitable and ecologically meaningful way.

Each State is required to establish programmes for river monitoring in order to determine "water status" with River Basin Districts (EPA, 2002a). The identification of hydromorphological reference conditions is thus an essential prerequisite for assessing hydromorphological quality, and is a specific requirement of the WFD to enable adequate classification of other (i.e. high, moderate, and good) status levels. Furthermore, the Directive states that the objective of achieving good water status should be pursued in an integrated manner for each river basin, so that measures in respect of surface water and groundwaters belonging to the same ecological, hydrological, and hydrogeological system are coordinated. The WFD specifies a nested programme of surveillance (Table 5) with additional monitoring required for protected areas.

Biological Integrity

In many countries of the EU, there has been much attention paid recently in devising practical methodologies and evaluating ecological integrity. In the United States, the concept of "biological integrity" has been a basis for much legislation in relation to river environments over the last 20 years (Karr and Chu, 2000). In the United Kingdom and Europe, however, there has been considerable discussion but little practical progress in development of methods for assessing "integrity" as opposed to "quality" (Table 6) in running waters (Harper *et al.*, 2000) and, as Karr (1993) has pointed out, the concepts of biotic integrity *per se* are not new. As far back as 1949, Aldo Leopold argued that:

Table 5 Monitoring requirements of the water framework directive (after EPA, 2002b)

Operation	Objectives
Surveillance monitoring	• To assess overall surface water status within each catchment or subcatchment • To validate assessment of potential impacts on waters, assist in optimization of future monitoring, and aid detection of long-term changes • To include all of the biological, physicochemical and hydromorphological elements (see Table 6)
Operational monitoring	• To establish the status of waters that the surveillance monitoring indicates to be in danger of failing to meet quality objectives • To determine changes arising from remedial measures
Investigative monitoring	• Required in cases where the reason for failure to meet water quality objectives is unclear or unknown • Where surveillance monitoring indicates objectives are unlikely to be achieved, to determine the causes why • To assess the impact of accidental pollution

Table 6 Quality elements for the classification of ecological status in rivers under the WFD

Quality element	Description
Biological elements	• Composition and abundance of aquatic flora • Composition and abundance of benthic invertebrate fauna • Composition, abundance, and age structure of fish fauna
Hydromorphological elements supporting the biological elements	• Hydrological regime • Quantity and dynamics of water flow • Connection to groundwater bodies • River continuity • Morphological conditions • River depth and width variation • Structure and substrate of the river bed • Structure of the riparian zone
Chemical and physicochemical elements supporting the biological elements	• Thermal conditions • Oxygenation conditions • Salinity • Acidification status • Nutrient conditions • Specific pollutants • Pollution by all priority substances identified as being discharged into the body of water • Pollution by other substances identified as being discharged into the body of water

> "a thing is right when it tends to preserve the integrity, stability and beauty of the biotic community"

In the United States, the phrase "biological integrity" first appeared in a policy context under the Water Quality Amendments of 1972 (PL-92-500) where the legislation called for the restoration and maintenance of "the chemical, physical and biological integrity of the nation's waters" (Karr, 1993). Overall, the development of methods to assess the ecological integrity of running waters requires the integration of physical and chemical parameters, as well as their effects on biological diversity, structure, and process (Harper *et al.*, 2000).

Walker *et al.* (2002) take the practical, administrative, and consultative burdens of the WFD a stage further by describing a system of Physical Quality Objectives for rivers in England and Wales. The existence of the very large database from the RHS allows experimentation with scoring systems involving habitat quality and habitat modification (both judged against "reference sites" in the database). The system goes further by classifying river survey sites in England and Wales along the axes formed by the Habitat Quality Assessment and the Habitat Modification Score in order to indicate how progress in meeting the demands of WFD compliance might be publicized and restoration strategies debated.

FLOW RESTORATION

The above account has shown considerable bias away from "hydro-" towards "morphological" in the definition of supporting conditions for freshwater biodiversity; this is understandable in England and Wales where, as mentioned, two-thirds of all channel lengths surveyed by RHS have been structurally modified. Nevertheless, a high proportion of rivers also experience a flow regime that is regulated (Gustard *et al.*, 1987) by reservoir impoundment, abstractions (Petts, 1996), or wastewater discharges (more than 70% of gauged rivers, according to Petts, 1988). A pertinent example of the issues raised by the pending application of the WFD in Scotland is the highly (perhaps "heavily") modified flow regime necessitated by hydroelectric power generation that dominates the hydromorphological quality of otherwise little directly altered channels (Walker, 2002).

Assessment of the ecological impacts of regulated flow regimes in England and Wales has been driven more by episodes of drought or by poorly assessed and regulated abstraction volumes than by reservoir operation. It is very clear that there have been considerable geomorphological impacts from reservoir operations (Petts, 1984) and that subtle, progressive impacts may follow (e.g. Sear, 1992 for the effects of reservoir releases on riffle-pool structures). Reservoir construction has now abated in the United Kingdom, in favor of demand management and other water supply strategies; however, in future projects, the flow assessment technologies developed for conditions of drought and over-abstraction (e.g. Bragg *et al.*, 1999) are likely to be deployed as part of environmental impact assessment and as part of compliance with the WFD – if only to confirm a "heavily modified" status. A simple physical assessment is already part of the Environment Agency's Catchment Abstraction Management System (Dunbar *et al.*, 2002).

CONCLUSIONS AND REFLECTION

This article commenced with discussion of the hydromorphological quality of freshwaters (in support of ecological quality) with reference to the considerable progress made in solving problems of chemical water quality. The methods of assessment, monitoring, and regulation applied to water

pollution perhaps offer us, in view of their apparent success, a model (or at least an analogue) for the application of similar controls on hydromorphology.

In fact, at least two different schemes of water quality management existed in Europe prior to the WFD, the Water Quality Objectives (or ambient) approach of the United Kingdom and the continental model of emission controls (without reference to the capacity of the receiving environment). The WFD unites these two models, which, despite their differences, have relied on:

- Toxicological evidence for system degradation – the central concept of pollution impacts;
- Appropriately independent authorities capable of regulation within nation states;
- Legal registration of the problem sites and surveillance to assess compliance and permit public accountability;
- Legal sanctions to prevent degradation;
- Derivation of indices of water quality and monitoring to indicate improvement/deterioration.

Several major areas of contrast in the new policy drive to rehabilitate freshwater ecosystems in the interests of sustainable development through environmental protection have been outlined. In terms of a regulatory role for hydromorphological elements, it can be shown that:

- The equivalent "exact" definitions (*cf.* toxicology) do not yet exist;
- Available data and tools for river morphology are scarce compared to the (basin) scale of the demand;
- The most profound causes of harmful impacts (channelization, excessive flow regulation, hydropower) may persist unregulated because of their socioeconomic importance (cf. chemical pollution controls where "the polluter pays").

Despite optimism that the appropriate authorities are largely in place and that public pressures become part of participation in the WFD, there remain the considerable contextual and academic problems of moving from limits imposed by chemistry to those imposed by physics and biology. Implicit, therefore, in the contemporary approach to ecosystem dynamics is a requirement to understand process and context; a requirement emphasized by the approaching deadline for implementation of the WFD. Process here refers to system dynamics and the mechanisms underlying them, while context refers to both the "spatial influences on the system" (e.g. Parker and Pickett, 1997 p. 22) and the timescales across which they operate.

FURTHER READING

Baird A.J. (1999) Introduction. In *Ecohydrology: Plants and Water in Terrestrial and Aquatic Environments*, Baird A.J. and Wilby R.L. (Eds.), Routledge: London, pp. 1–10.

Davies B. and Day J. (1997) *Vanishing Waters*, University of Cape Town Press: Cape Town.

Leopold A. (1949) *A Sand County Almanac: and Sketches Here and There*, Oxford University Press: New York.

REFERENCES

Amoros C. and Petts G.E. (Eds.) (1993) *Hydrosystèmes Fluviaux*, Masson: Paris.

Bragg O.M., Black A.R. and Duck R.W. (1999) *Anthropogenic Impacts on the Hydrology of Rivers and Lochs*, SNIFFER Report W98(50)I1, University of Dundee.

Brierley G.J. and Fryirs K. (2000) River styles, a geomorphic approach to catchment characterization: implications for river rehabilitation in Bega Catchment, New South Wales, Australia. *Environmental Management*, **25**, 661–679.

Brookes A. and Sear D.A. (1996) Geomorphological principles for restoring channels. In *River Channel Restoration: Guiding Principles for Sustainable Projects*, Brookes A. and Shields F.D. Jr (Eds.), John Wiley & Sons: New York, pp. 75–101.

Brooks A. (1995) River channel restoration: theory and practice. In *Changing River Channels*, Gurnell A.M. and Petts G.E. (Eds.), John Wiley & Sons: Chichester, pp. 368–388.

Brunke M. and Gonser T. (1997) The ecological significance of exchange processes between rivers and groundwater. *Freshwater Biology*, **37**, 1–33.

Cairns J. (1987) Disturbed ecosystems as opportunities for research in restoration ecology. In *Restoration Ecology*, Jordan W.R., Gilpin M.E. and Aber J.D. (Eds.), Cambridge University Press: Cambridge, pp. 307–320.

Calder I.R. (1999) *The Blue Revolution. Land Use and Integrated Water Resources Management*, Earthscan: London.

Charlton M.E., Large A.R.G. and Fuller I.C. (2003) Application of LiDAR in river environments: the river Coquet, Northumberland, UK. *Earth Surface Processes and Landforms*, **28**, 299–306.

Chave P.A. (2001) *The EU Water Framework Directive: an Introduction*, IWA Publishing: London.

Doppelt B., Scurlock M., Frissell C. and Karr J. (1993) *Entering the Watershed*, Island Press: Washington.

Dunbar M.J., Acreman M. and Kirk S. (2002) Environmental flow setting in England and Wales. Current practice; future challenges, *International Conference on Environmental Flows for River Systems*, Cape Town.

Elliot R. (1997) *Faking Nature. The Ethics of Environmental Restoration*, Routledge: London.

EPA (2002a) *Outline of Research Priorities to support the Implementation of the EU Water Framework Directive*, Document C4/2002, National Development Plan, Environmental Protection Agency: Dublin.

EPA (2002b) *National Rivers Monitoring Programme (Incorporating Monitoring Requirements of the Water Framework Directive)*, Environmental Protection Agency: Dublin.

European Commission (2000) Directive 2000/60/EC. Establishing a framework for Community action in the field of water policy. *Official Journal of the European Communities*, **L327**, 1–72.

Fox P.J.A., Naura M. and Scarlett P. (1998) An account of the derivation and testing of a standard field method, River Habitat Survey. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **8**, 455–475.

Frissell C.A., Liss W.J., Warren C.E. and Hurley M.D. (1986) A hierarchical framework for stream classification: viewing streams in a watershed context. *Environmental Management*, **10**, 199–214.

Gustard A., Cole G., Marshall D. and Bayliss A. (1987) *A Study of Compensation Flows in the UK*, Report 99, Institute of Hydrology, Wallingford.

Gore J.A. (1985) Mechanisms of colonisation and habitat enhancement for Benthic macro-invertebrates in restored river channels. In *The Restoration of Rivers and Streams: Theories and Experiences*, Gore J.A. (Ed.), CRC Press: Boca Raton, pp. 23–38.

Harper D.M., Kemp J.L., Vogel B. and Newson M.D. (2000) Towards the assessment of 'ecological integrity' in running waters of the United Kingdom. *Hydrobiologia*, **422/423**, 133–142.

Hawkins J.W. (1991) *The Oxford Minidictionary*, Clarendon Press: Oxford.

Hayward A.L. and Sparks J.J. (1990) *The Concise English Dictionary*, New Orchard Editions: London.

Jewitt G. (2002) Can integrated Water Resources Management sustain the provision of ecosystem goods and services? *Physics and Chemistry of the Earth*, **27**, 887–895.

Karr J.R. (1993) Measuring biological integrity: lessons from streams. In *Ecological Integrity and the Management of Ecosystems*, Woodley S., Kay J. and Francis J. (Eds.), St Lucie Press: Canada, pp. 83–103.

Karr J.R. and Chu E.W. (2000) Introduction: sustaining living rivers. *Hydrobiologia*, **422–423**, 1–14.

Kentula M.E. (2000) Perspectives on setting success criteria for wetland restoration. *Ecological Engineering*, **15**, 199–209.

Kern K. (1992) Rehabilitation of streams in South-west Germany. In *River Conservation and Management*, Boon P.J., Calow P. and Petts G.E. (Eds.), John Wiley & Sons: Chichester, pp. 321–335.

Kondolf G.M. (1995) Geomorphological stream classification in aquatic habitat restoration: uses and limitations. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **5**, 127–141.

Kondolf G.M. and Piegay H. (Eds.) (2003a) *Tools in Fluvial Geomorphology*, Wiley & Sons: Chichester.

Kondolf G.M. and Piegay H. (2003b) Tools in fluvial geomorphology: problem statement and recent practice. In *Tools in Fluvial Geomorphology*, Kondolf G.M. and Piegay H. (Eds.), John Wiley & Sons: Chichester, pp. 3–22.

Large A.R.G. and Prach K. (1999) Plants and water in streams and rivers. In *Ecohydrology: Plants and Water in Terrestrial and Aquatic Environments*, Baird A.J. and Wilby R.L. (Eds.), Routledge: London, pp. 237–268.

Lenders H.J.R., Aarts B.G.W., Strijbosch H. and Van der Velde G. (1998) The role of reference and target images in ecological recovery of river systems: lines of thought in the Netherlands. In *New Concepts for Sustainable Management of River Basins*, Nienhuis P.H., Leuven R.S.E.W. and Ragas A.M.J. (Eds.), Backhuis: Leiden, pp. 35–52.

Lorenz C.M. (1997) Concepts in river ecology: implications for indicator development. *Regulated Rivers: Research and Management*, **13**, 501–516.

National Research Council, Committee on Restoration of Aquatic Ecosystems: Science, Technology and Public Policy, Water Science and Technology Board and Commission on Geosciences, Environment and Resources (1992) *Restoration of Aquatic Ecosystems: Science technology and Public Policy*, National Academy Press: Washington.

Newson M.D. (1997) *Land, Water and Development. Sustainable Management of River Basin Systems*, *Second Edition*, Routledge: London.

Newson M.D. (2000) Science and sustainability: addressing the world water 'crisis'. *Progress in Environmental Science*, **2**, 205–229.

Newson M.D. (2002) Geomorphological concepts and tools for sustainable river ecosystem management. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **12**, 365–379.

Newson M.D. and Newson C.L. (2000) Geomorphology, ecology and river channel habitat: mesoscale approaches to basin-scale challenges. *Progress in Physical Geography*, **24**, 195–217.

Newson M.D., Pitlick J. and Sear D.A. (2002) Running water: fluvial geomorphology and river restoration. In *Handbook of Ecological Restoration*, Perrow M.R. and Davy A.J. (Eds.), Cambridge University Press: Vol. 1, pp. 133–152.

Nienhuis P.H. and Leuven R.S.E.W. (1998) Ecological concepts for the sustainable management of lowland river basins: a review. In *New Concepts for Sustainable Management of River Basins*, Nienhuis P.H., Leuven R.S.E.W. and Ragas A.M.J. (Eds.), Backhuis: Leiden, pp. 7–33.

NTNU (2002) *A Guidance Standard for Assessing the Hydromorphological Features of Rivers*, Report CEN TC 230/WG 2/TG 5: N32, Norges-Teknisk-Naturvitenskapelige Universitet, **http://www.bygg.ntnu.no/~borsanyi/eamnweb/documents/CEN_TC230-WG2-TG5-N32_05-02.pdf**. Accessed 18/5/03.

Parker V. and Pickett S.T. (1997) Restoration as an ecosystem process: implications of the modern ecological paradigm. In *Restoration Ecology and Sustainable Development*, Urbanska, K.M., Webb N.R. and Edwards P.J. (Eds.), Cambridge University Press: pp. 17–32.

Perrow M.R. and Wightman A.S. (1993) *The River Restoration Project Phase 1: Feasibility Study*, ECON, University of East Anglia & River Restoration Secretariat, Oxford Polytechnic: Oxford.

Petersen R.C., Petersen L.B.-M. and Lacoursie J.L. (1992) A building-block model for stream restoration. In *River Conservation and Management*, Boon P.J., Calow P. and Petts G.E. (Eds.), John Wiley & Sons: Chichester, pp. 293–309.

Petts G.E. (1984) *Impounded Rivers: Perspectives for Ecological Management*, Wiley: Chichester.

Petts G.E. (1988) Regulated rivers in the UK. *Regulated Rivers: Research and Management*, **2**, 201–220.

Petts G.E. (1996) Water allocation to protect ecosystems. *Regulated Rivers: Research and Management*, **12**, 353–365.

Petts G.E., Maddock I., Bickerton M. and Ferguson A.J.D. (1995) Linking hydrology and ecology: the scientific basis for river management. In *The Ecological Basis for River Management*,

Harper D.M. and Ferguson A.J.D. (Eds.), John Wiley & Sons: Chichester, pp. 1–16.

Rapport D.J. (1995) Ecosystem health: more than a metaphor? *Environmental Values*, **4**, 287–309.

Raven P.J., Fox P.J.A., Everard M., Holmes N.T.H. and Dawson F.H. (1997) River Habitat Survey: a new system for classifying rivers according to their habitat quality. In *Freshwater Quality: Defining the Indefinable?* Boon P.J. and Howel D.L. (Eds.), The Stationery Office: Edinburgh, pp. 215–234.

Raven P.J., Holmes N.T.H., Charrier P., Dawson F.H., Naura M. and Boon P.J. (2002) Towards a harmonized approach for hydromorphological assessment of rivers in Europe: a qualitative comparison of three survey methods. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **12**, 405–424.

Richards K. (1982) *Rivers: Form and Process in Alluvial Channels*, Methuen & Co: London.

Rosgen D.L. (1996) *Applied River Morphology*, Wildland Hydrology, Pagosa Springs: Colorado.

Sear D.A. (1992) Impact of hydro-electric power releases on sediment transport processes in pool-riffle sequences. In *Dynamics of Gravel-bed Rivers*, Billi P., Hey R.D., Thorne C.R. and Tacconi P. (Eds.), Wiley: Chichester, pp. 630–650.

Sear D.A. (1994) River restoration and geomorphology. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **4**, 169–177.

Tapsell S.M. (1995) River restoration: what are we restoring to? A case study of the Ravensbourne River, London. *Landscape Research*, **20**, 98–111.

Thomson J.R., Taylor M.P., Fryirs K.A. and Brierley G.J. (2001) A geomorphological framework for river characterization and habitat assessment. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **11**, 373–389.

Van Niekerk A.W., Heritage G.L. and Moon B.P. (1995) River classification for management: the geomorphology of the Sabie River in the Eastern Transvaal. *South African Geographical Journal*, **77**, 68–76.

Wade P.M., de Waal L.C. and Haukeland A. (1993) Integrating river rehabilitation into the planning process – a comparison of Norway and England. In *Ökologische Gewässeranierung im Spannungsfeld zwischen Natur und Kultur, Wasser, Abwasser, Abfall*, 11, Wolf P. (Ed.), Schriftenreihe des Fachgebietes Siedlungswasserwirtschaft Universität: Gesamthochschule Kassel, pp. 85–92.

Wadeson R.A. (1994) A geomorphological approach to the identification and classification of instream flow environments. *South African Journal of Aquatic Science*, **20**, 38–61.

Walker S. (2002) Challenges to the implementation of the Water Framework Directive in Scotland: a personal view. *Journal of the Chartered Institution of Water and Environmental Management*, **16**, 277–281.

Walker J., Diamond M. and Naura M. (2002) The development of physical quality objectives for rivers in England and Wales. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **12**, 381–390.

Ward J.V., Tockner K., Uehlinger U. and Malard F. (2001) Understanding natural patterns and processes in river corridors as the basis for effective river restoration. *Regulated Rivers: Research & Management*, **17**, 311–323.

Wassen M.J. and Grootjans A.P. (1996) Ecohydrology: an interdisciplinary approach or wetland management and restoration. *Vegetatio*, **126**, 1–4.

Whisenant S.G. (1999) *Repairing Damaged Wildlands. a Process-Oriented Landscape Scale Approach*, Cambridge University Press.

Wilby R.L. (1999) The future of ecohydrology. In *Ecohydrology: Plants and Water in Terrestrial and Aquatic Environments*, Baird A.J. and Wilby R.L. (Eds.), Routledge: London, pp. 346–373.

Woodley S., Kay J. and Francis G. (Eds.) (1993) *Ecological Integrity and the Management of Ecosystems*, St Lucie Press, Canada.

191: Environmental Flows: Managing Hydrological Environments

CHRISTOPHER J GIPPEL

Fluvial Systems Pty Ltd, Stockton, Australia

Environmental flows refer to the facets of natural flow regimes that are required to sustain riparian and aquatic life in a healthy condition, plus provide for the needs of those human uses that rely on, and do not compromise, stream health values. Environmental flow assessments are required for regulated rivers, where the flow regime has been modified, or flow volumes significantly reduced. The methodological trend has been from narrow studies that catered for a single fish species at one critical life phase to a holistic approach that aims to restore natural river processes. Consideration has also been given to the specific issues of flood flows required for wetland inundation and channel formation, needs of groundwater-dependent ecosystems, the needs of estuarine ecosystems, and flows required for recreational activities. Holistic methodologies are particularly appropriate in developing countries with strong livelihood dependencies on the goods and services provided by aquatic ecosystems. The number of studies worldwide that report environmental flow recommendations greatly outweighs the number of studies that report scientific evaluation of the results of implementation of these recommendations. However, where implemented environmental flow regimes have been monitored, the results have generally been positive.

INTRODUCTION

Many studies conducted over the past two decades throughout the world have cited flow regulation (by dams, diversions, and direct pumping) as a major cause of degradation across the spectrum of river health (Gordon *et al.*, 2004, p. 286). It is often difficult to isolate observed changes in the aquatic environment that are due to flow regulation from those that are due to changes in other factors such as catchment land use, fishing pressure, introduced species, riparian vegetation cover, large woody debris distribution, and natural variations in flow regime (i.e. the pattern of floods, droughts, and seasonal flow variability). The difficulty arises from the fact that changes to these factors can produce similar ecological responses, and the changes in these factors are often overlapping or simultaneous with flow regulation. While data are often available to quantify the hydrological impact of regulation, in most situations, regulation began well before ecological monitoring programs were initiated.

Identification of flow regulation as one of the major causes of stream health degradation has led river managers to seek ways of reversing this effect, to the extent that this is currently one of the priority issues in stream management throughout the world (Gordon *et al.*, 2004, p. 287). Modifying regulated flow regimes and returning diverted water back to rivers to benefit wildlife is a highly contentious and controversial issue. Conflicts over water use are common, especially where water supplies are limited, as in arid and semiarid areas, where demands for water supplies are increasing, and where there are other in-stream uses such as navigation and recreation. It is necessary then to determine a satisfactory balance between competing uses. The process of achieving this balance is the topic of this article.

In-stream and Environmental Flows Defined

The term *in-stream flow* is used, particularly in the United States, to refer to a specific stream flow that is identified for the purposes of planning or management of a stream or river. In-stream flows are those that are retained in their natural setting, as opposed to those waters that are diverted for offstream users such as industry, agriculture, and town water supply. Flows influence adjacent groundwater levels,

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

as well as the hydrological status of foodplain wetlands. The level of flow in a stream also conditions aesthetic and scenic values. Navigation is affected by flows. For example, kayakers in mountain streams require high flows, while in large rivers like the Columbia River in Washington State, if flows are below a certain level, the river becomes impassable to commercial barges, tugs, and other watercraft because of the lack of draft (Rushton, 2000). In-stream flows are also used for the production of hydroelectricity and waste disposal. In some places, indigenous people live close to the river and their livelihoods depend on the river's resources, both clean water and biological resources (Quinn, 1991). Some rivers have valuable commercial and recreational fishing industries that depend on in-stream flows. The intrinsic wildlife values of rivers are also sustained by in-stream flows. The flows required to maintain or rehabilitate the habitat for riparian and aquatic life are the most difficult to quantify, and this is the main focus of the rest of this article. This subset of in-stream flows is known, particularly in Australia and southern Africa, as *environmental flows*, although in recent times, this term has been used more broadly to also include other related uses.

Brown and King's (2003) definition of environmental flows encompassed aesthetic, recreational, and cultural values as well as biophysical ones, and they also placed social and economic issues within the realm of environmental flows. In-stream flows for hydropower releases, irrigation releases, navigation, dilution of pollution, release of wastewater, and interbasin transfers were excluded from Brown and King's (2003) definition of environmental flows, because these uses traditionally work against achievement of their defined ecological and human needs goals. So, for the remainder of this article, the term environmental flows is used to refer to in-stream flows required to sustain riparian and aquatic life in a healthy condition, plus those human uses that rely on, and do not compromise, stream health values.

Three Basic Assumptions of Environmental Flows Assessment

Gordon *et al.* (2004, p. 288–290) stated three basic assumptions that underlie most environmental flow assessments:

- assessments are generally grounded on the natural flow paradigm, even though the final regime is likely to be a compromise on this;
- less than all the natural flow will maintain stream integrity (stream health); and
- stream health exists on a continuum

The natural flow paradigm states that discharge variability is central to sustaining and conserving biodiversity and ecological integrity (e.g. Poff *et al.*, 1997; Richter *et al.*, 1997; Puckridge *et al.*, 1998; Tharme and King, 1998). Given the currently limited understanding of flow-ecology relationships, and the improbability of ever being able to fully define the needs of the whole biological community, the conservative alternative is to assume that the natural flow regime is the best indicator of environmental needs. The idea of mimicking a natural system, also known as *physiomimesis* (Katopodis, 2003), appears as an objective in numerous publications in the environmental flow literature (e.g. Environment Protection Authority of NSW, 1997; Arthington, 1998; Tharme and King, 1998; Snowy Water Inquiry, 1998). While the natural flow regime paradigm has merit, translating it into practical recommendations for environmental flows can be problematic. In practice, the process of determining environmental flows does not involve attempting to devise a regulated flow regime that has a statistically defined variability (across all timescales) identical to that of the natural flow regime. For example, all of the environmental flow projects listed by Poff *et al.* (1997) involved only partial restoration of the natural flow regime. Under conditions of limited water resources, competing demands, and constraints on flow control imposed by river structures, some flow variability targets will be low priority and others will be impossible to implement.

There is an implicit understanding that environmental flow assessment is not about recommending the full natural flow regime as the ideal. Rather, it is about determining what lesser amount will maintain stream health in the desired condition. For streams regulated by dams and major diversions, the objective of environmental flow assessment is to set maintenance or rehabilitation targets (Gippel *et al.*, 2002), while for other streams, the objective is to set a sustainable limit to current and/or future pumping and diversions (Nathan *et al.*, 2002). The biotic and abiotic components of stream ecosystems can range in their sensitivity to water resource development (Figure 1). Estimates of how much flow must remain in a river before its health is impaired range from about 65% to 95% of natural flow, provided the natural pattern of flow is also retained (Dyson, 2003). The range of values for an acceptable level of hydrological modification given in the literature partly reflects the enormous variation in sensitivity and importance of the systems that have been assessed (Gordon *et al.*, 2004, p. 290).

When stream health is defined as ecosystem integrity (Karr and Dudley, 1981) it could be construed as an absolute term that cannot have degrees, that is, ecosystem integrity is intact (stream is healthy) or it is not (stream is not healthy). A healthy working river is one that is managed to provide a compromise, agreed to by the community, between the condition of the river and the level of human use (Gordon *et al.*, 2004, p. 239). This concept is implicit in stream health assessment methodologies that rate stream condition against a reference condition, but along a

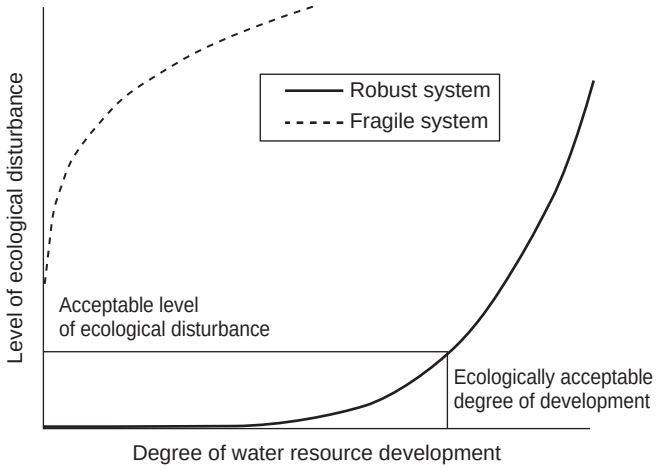


Figure 1 Conceptual diagram of the relation between the degree of water resource development and the level of disturbance. Source: Stewardson and Gippel (1997), Hydro Tasmania

grade, such as expressed by a range of observed/expected scores, or as a range of index scores compared against the highest possible score (Gordon *et al.*, 2004, p. 239–262). Also, grades of stream health are allocated to rivers in classification schemes (Gordon *et al.*, 2004, p. 262–286). In situations where the vision for the river, and hence the management goal, is a departure from the reference condition, the environmental flows assessment process seeks to find the balance between the desired ecosystem condition and other social and economic needs for water (Dyson, 2003). The flows allocated to achieve the chosen condition are the environmental flows.

FORMS OF ENVIRONMENTAL FLOW ASSESSMENT

Environmental flow assessments vary across a wide range of complexity and depth, as dictated by the level of funding, availability of data, technical capacity, time frame, priority of the site, or expected level of controversy. Environmental flows may be specified at several levels of resolution, from a single annual flow volume or a minimum flow limit below which diversions are not permitted, through to a comprehensive flow regime that specifies the distribution of a range of flows throughout the year. A comprehensive study might specify flows necessary to allow the passage of fish, flows to provide sufficient living space for biota or to ensure acceptable levels of temperature, dissolved oxygen or salinity, and higher flows (flushing flows or channel maintenance flows) to remove fine materials from the streambed, scour out encroaching vegetation, or flush anoxic or highly saline waters from stratified pools. The spatial scale of environmental flows assessment also varies widely, from whole of catchment to the river reach. Methodologies range from relatively simplistic, reconnaissance-level approaches to resource-intensive methodologies for detailed studies.

There have been numerous reviews and evaluations of environmental flows methodologies (e.g. Arthington and Zalucki, 1998; Espegren, 1998; Jowett, 1997; King *et al.*, 1999; Annear *et al.*, 2002; Caissie and El–Jabi, 2003; Tharme, 2003; Scatena, 2004), and a comprehensive bibliography is provided in Gordon *et al.* (2004, p. 291). Tharme's (2003) recent global review of the present status of environmental flow methodologies revealed the existence of some 207 individual methodologies, recorded for 44 countries within six world regions. A database of these methodologies is available online at **`www.lk.iwmi.org/ehdb/EFM/efm.asp`** (International Water Management Institute, 2003). This article provides only a sketch of the main methods currently in use. There is a vast literature on the topic of relationships between flow and ecological responses (e.g. Pusey *et al.*, 2000; Ward *et al.*, 2001; Bunn and Arthington, 2002), and while this knowledge forms the empirical basis of much of the environmental flow assessment process, it is not reviewed here. Some key river ecosystem processes are discussed in the article on river ecosystems (river continuum concept, flood-pulse dynamics, hyporheic zone processes). The focus of this article is environmental flow methodologies, with an emphasis on the hydrological, hydraulic, and geomorphological aspects.

Hydrological Methods

Hydrological methodologies use simple rules based on flow duration or mean discharge to scale down the natural flow regime. The Tennant method (Tennant, 1976), also referred to as the *Montana method*, is the most commonly applied hydrological methodology worldwide (Tharme, 2003). Recommended minimum flows are based on percentages of the average annual flow, with different percentages for winter and summer months. The Hoppe (1975) regional rule-of-thumb method utilizes flow duration curves ideally based on daily discharge data.

The main attraction of hydrological techniques is the fact that an answer can be obtained rapidly if gauged flow records are available, eliminating the time and cost of field data collection. The techniques tend to be site- and species-specific, and for them to be applicable in other situations, the relationship between habitat and discharge must be similar. Orth and Leonard (1990), Pusey (1998), Dunbar *et al.* (1998), and Tharme (2003) concluded that simple hydrological methods requiring little or no fieldwork are most appropriate for basin-wide planning purposes, or for providing preliminary estimates in uncontroversial situations. A number of studies have compared hydrological methods with more sophisticated approaches (e.g. Orth and Maughan, 1981; Orth and Leonard, 1990; Caissie *et al.*,

1998; Bureau of Reclamation, 1999). These comparisons generally found that there was no consistent or reliable pattern in the relative magnitude of flows recommended by these methods.

Hydraulic Rating Methods

Hydraulic rating methods utilize a quantifiable relationship between the quantity and quality of an in-stream resource, such as fishery habitat, and discharge, to calculate flow recommendations. Most emphasis has been placed on the passage, spawning, rearing, and other flow-related maintenance requirements of individual, economically or recreationally important fish species. This approach is sometimes known as a *transect method* or *wetted-perimeter method* because it involves measuring and interpolating changes in simple hydraulic variables, such as wetted perimeter, mean velocity, or maximum depth, usually measured across single river cross sections, as a surrogate for habitat factors known or assumed to be limiting to target biota (Figure 2). Simple wetted-perimeter methods (Nelson, 1980) (Figure 3) have been widely applied for many years (Tharme, 2003), and are still being used to make important environmental flow determinations (e.g. McCarthy, 2003). Gippel and Stewardson (1998) provided an application and evaluation of the hydraulic rating method for headwater streams in Victoria, Australia, and compared the results with those of hydrological methods, and macroinvertebrate and fish data. In that case, the wetted-perimeter breakpoints occurred at flows considered too low to adequately protect the biota.

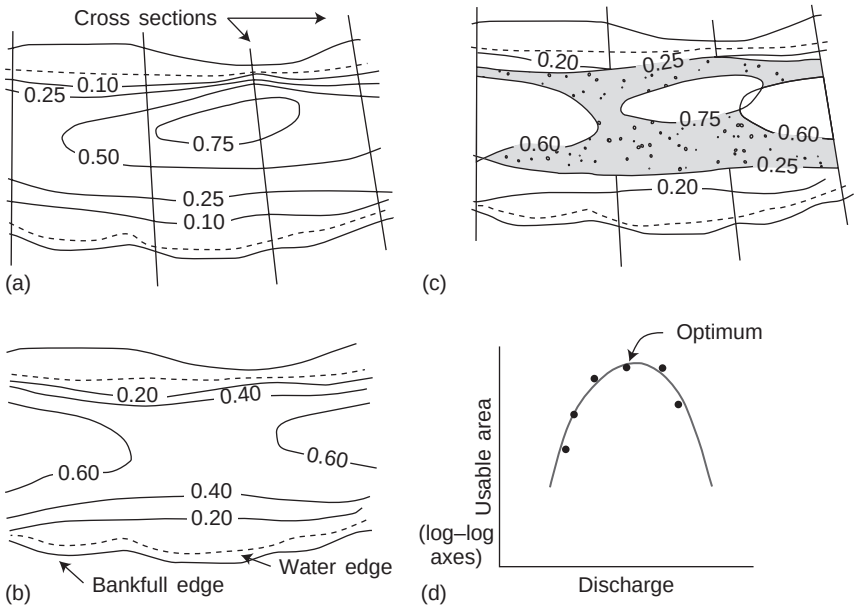


Figure 2 Washington method for determining preferred discharge. For this theoretical example, ''preferred habitat'' is: depth 0.25–0.75 m and velocity 0.20–0.60 m s^{-1}. Figures display: (a) depth contours in meters, (b) velocity contours in m s^{-1}, (c) a combination of maps (a) and (b) with shaded region showing usable area, and (d) trend-fitted curve derived from measurements (point shown) taken at several discharges. The ''optimum'' discharge corresponds with the greatest usable area. (Adapted from Collings (1972) © Washington State Department of Ecology)

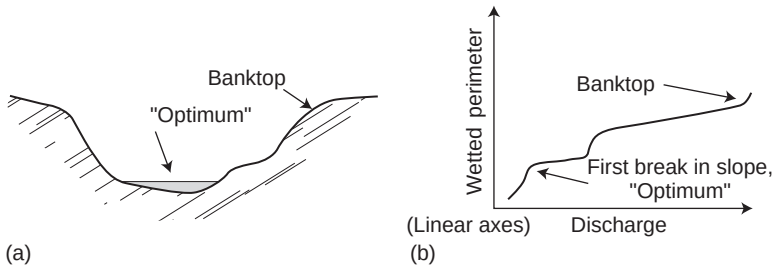


Figure 3 Wetted-perimeter method: (a) hypothetical channel cross section and (b) graph of wetted perimeter versus discharge. The first breakpoint in slope is used as an index of optimum available water (Reproduced from Gordon *et al.* (2004) by permissions of John Wiley & Sons.)

Habitat Rating Methods

Hydraulic rating methodologies were the precursors of more sophisticated habitat rating or simulation methodologies, also referred to as microhabitat or habitat modeling methodologies (Tharme, 2003). Habitat rating methods not only consider how physical habitat changes with stream flow but also combine this information with continuous habitat preference functions for given species to determine the amount of suitable habitat available over a range of stream flows. Results are normally in the form of a curve showing the relationship between suitable available habitat area and stream discharge. From this curve, the optimum stream flows for a number of individual species can be ascertained and the results used as a guide for recommending environmental flows. The most well known of the habitat rating methods is contained within the In-stream Flow Incremental Methodology (IFIM) (Bovee, 1982).

IFIM is the most commonly used flow assessment method worldwide (Tharme, 2003; King and Brown, 2003). In many States of the USA, IFIM is regarded as the most scientifically legally defensible method, so it has become the standard for large rivers, with hydraulic rating methods used on smaller streams (Espegren, 1998). There has been considerable development of IFIM since its introduction over 20 years ago (Bovee, 1982), with some of the earlier concepts becoming redundant. King and Tharme (1993) presented an updated step-by-step guide to applying the IFIM, including a review of the process.

The IFIM is much more than a habitat rating method. It is a problem-solving tool made up of a collection of analytical procedures and computer models. It was designed as a communication link between fishery biologists, hydrologists, and hydraulic engineers. "Incremental" means the slight or incremental modification of the problem or the perspective or view of the problem until a solution is found. It also refers to the ability to look at the effects of incremental changes in a variable (e.g. discharge) on available habitat. Rather than generating a single answer, the methodology produces a range of solutions that permit the evaluation of different alternatives.

The IFIM is grounded in the ecological niche concept of explaining community distribution, incorporating both macrohabitat and microhabitat concepts. Macrohabitat applies to large-scale longitudinal gradients in habitat characteristics along streams, and microhabitat is the precise location where an individual species is normally found. The microhabitat approach was originally justified by reference to two studies showing that competition between fish species was reduced by physical habitat isolation (Bovee and Milhous, 1978). The Weighted Useable Area (WUA) microhabitat model is the core habitat rating component of the IFIM. WUA is an indicator of the net suitability of use of a given reach by a certain life stage of a certain species. PHABSIM (Physical HABitat SIMulation) is a collection of computer programs used to model WUA. The IFIM habitat modeling process is conceptualized in Figure 4.

The IFIM/PHABSIM approach is not without its critics (Castleberry *et al.*, 1996; Van Winkle *et al.*, 1997; Pusey, 1998), with claims that its misuse has actually hastened the decline of rivers in the western United States, particularly the large alluvial rivers (Woo, 1999). Two major criticisms of PHABSIM are that it promotes the application of flows that are too low for sediment mobilization, and that it promotes relatively constant flows. Much of the criticism relates to the earlier applications of IFIM, when the focus was on a single species of fish, or on studies where PHABSIM was enthusiastically applied outside of the more

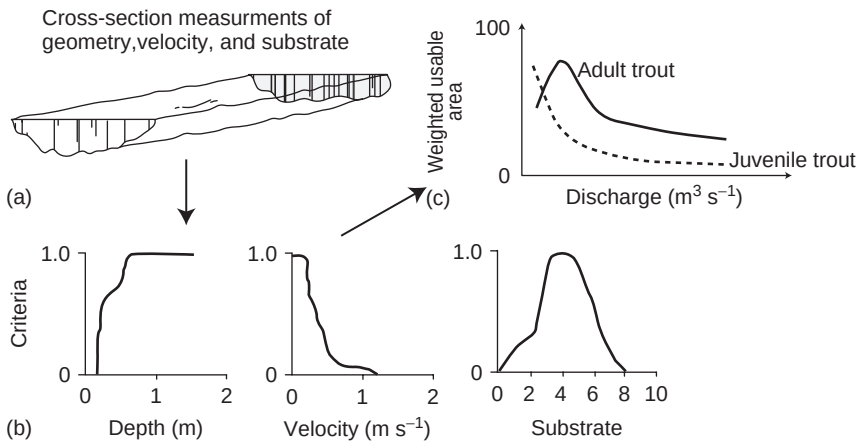


Figure 4 A conceptualization of the procedures in PHABSIM: a stream reach is selected and surveyed at a particular discharge (a), and the information is combined with habitat-Suitability Index (SI) curves defined for a particular species and life stage (b). The procedure is repeated for several discharges (using measured or modeled data) to obtain a curve describing the suitability of use of that reach as a function of discharge (c) (Reproduced from Gore and Nestler (1988) by permission of John Wiley & Sons, Ltd.)

holistic IFIM framework, and channel forming flows were ignored (Barinaga, 1996).

A critical limitation on the use of habitat-simulation models is the lack of well-defined habitat-suitability curves. As these curves are essentially empirical correlations, some authors (e.g. Nestler *et al.*, 1989) warn that the curves may not be transferable from one stream to another. The strongest criticism of PHABSIM has centered on the ecological interpretation of the WUA index. Gore and Nestler (1988) reviewed and commented on the criticisms put forward by a number of authors. The assumption is made in in-stream flow uses of PHABSIM that if the habitat is maintained, the fish population will be maintained. Some studies found evidence to support this assumption (Orth and Maughan, 1982; Stalnaker, 1979), while others did not (Irvine *et al.*, 1987; Mathur *et al.*, 1985; Scott and Shirvell, 1987). Other factors such as food supply, biological interactions (e.g. competition, predation), nutrients, dissolved oxygen, presence of ice cover, temperature, and flow regime (including the effect of floods) may be of greater importance than physical habitat in limiting species biomass or abundance (e.g. Pusey, 1998). Orth and Maughan (1982) concluded that IFIM is a useful framework for managing streams with altered flow regimes but it is not a panacea. Biological expertise is still needed in the interpretation of results. However, the model can provide a basis upon which a biologist may apply professional judgment (Mosley and Jowett, 1985) and a methodology for comparing the relative effects of different management decisions.

Stewardson and Gippel (2003) noted that because IFIM in its original form is grounded in ecological niche theory, it does not adequately account for disturbance, which is an important determinant of community structure. Their Flow Events Method extends the hydrological variability (Range of Variability Approach, or RVA) idea of Richter *et al.* (1996) to also consider temporal changes in physical habitat conditions at the appropriate spatial scales. The method characterizes flow variations for environmental flow studies using knowledge of the influence of flow events on biological and geomorphic processes. The Flow Events Method is not a new method, but an analytical tool that can be used within any framework, such as IFIM or various holistic approaches.

Holistic Methodologies

Holistic approaches are essentially frameworks for organizing and using flow-related data and knowledge (Arthington *et al.*, 1992; Brown and King, 2003; King *et al.*, 2003). Holistic approaches cover a wide range of related "methodologies", all listed in Gordon *et al.* (2004, p. 299) and online at **www.lk.iwmi.org/ehdb/EFM/efm.asp** (International Water Management Institute, 2003). Several holistic methodologies have been developed, primarily in Australia and South Africa, with the same fundamental characteristics, but different approaches to development of quantitative or semiquantitative relationships between flow modifications and geomorphological/ecological responses. All are reviewed in Tharme (2003) and Arthington and Pusey (2003). The holistic approaches are not constrained by the available analytical tools, and it is not unusual for these approaches to make use of several different techniques. The full IFIM process also falls within the realm of a holistic approach. While the hydrological, hydraulic rating, and habitat rating methods usually focus on key sport fish such as salmonids, or fish with a very high conservation value, the holistic approaches attempt to consider the entire ecosystem, using all available information, much of which may be little more than personal experience and working hypotheses.

Proponents of holistic methodologies take the view that, where possible, management strategies for rivers should be aimed at maintaining or rehabilitating as much as possible of the original, functional aquatic ecosystem (Arthington and Pusey, 1993; Arthington and Pusey, 2003). Holistic approaches are based on the natural hydrological regime and are intended to provide water required for the complete ecosystem, including the river channel, riparian zone, floodplain, groundwater, wetlands, and estuary. Explicit numerical models that relate discharge to aspects of the river's geomorphology, water quality, or ecology (i.e. hydraulic or habitat rating models) may be available, or even developed through the course of the investigation, but they are usually used as an aid to decision making, rather than as a numerical solution to the problem of defining a suitable regulated flow regime.

An important advance in holistic methodology has been the establishment, in Queensland, Australia, of the Benchmarking methodology, which undertakes basin-scale evaluation of the potential environmental impacts of future scenarios of water resource management (Arthington, 1998; Brizga *et al.*, 2002). Another significant development has been the incorporation of social, cultural, and economic components into environmental flow assessments. For example, the DRIFT (Downstream Response to Imposed Flow Transformation) method developed for use in southern Africa (Brown and King, 2000; King *et al.*, 2003) evaluates links between flow modifications, geomorphological and ecological responses and social consequences for subsistence users of river resources, and economic implications in terms of mitigation and compensation. Another example is The Living Murray initiative (**www.thelivingmurray.mdbc.gov.au**), which is an ongoing process that is examining the ecological, cultural, social, and economic consequences of flow scenarios for the entire 2530 km long River Murray, Australia (Gippel *et al.*, 2002).

A holistic approach may adopt a bottom-up philosophy (Arthington, 1998), whereby the regime is built up

from hydrological facets (usually a pattern of baseflows combined with various sized floods) that is assumed from experience or research to be the minimum combination required to achieve ecological sustainability. The BBM (Building Block Methodology) (King and Louw, 1998) builds the regime from three main groups of flow facets (Figure 5). The alternative is the top-down philosophy, whereby it is assumed that the entire natural flow regime is ecologically important, but that some facets of it can be modified or omitted without threatening ecological sustainability (Arthington, 1998). The top-down DRIFT process (Figure 6), which emerged from the foundations of the BBM, was developed for the assessment of environmental flows for the Lesotho Highlands Water Project (King and Brown, 2003; Arthington *et al.*, 2003).

Holistic approaches share four main assumptions regarding achievement, or maintenance, of ecological sustainability:

- some facets of the natural flow regime cannot be scaled down, and must be retained in their entirety
- some facets of the natural flow regime can be scaled down
- some facets of the natural flow regime can be omitted altogether
- variability of the regulated flow regime should mimic that of the natural flow regime, in certain respects

These assumptions arise from the notion that high- and low-flow events are more important than in-between conditions because of the stresses and opportunities they present to the biota (Poff *et al.*, 1997). Also, many geomorphic and ecological processes show nonlinear responses to flow, requiring a threshold to be exceeded before the process

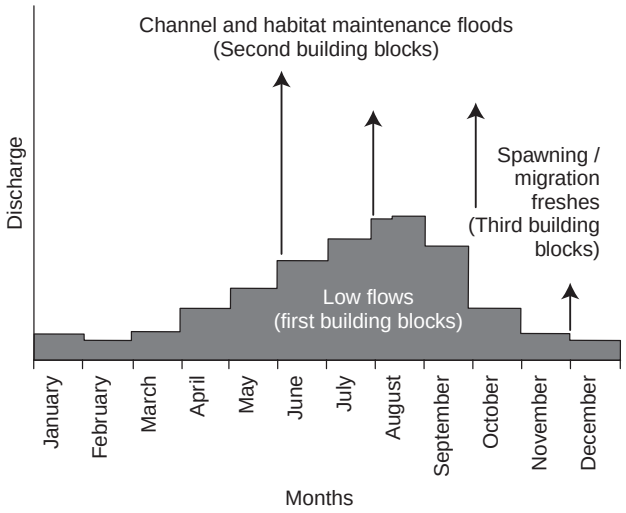


Figure 5 The ''building blocks'' of a hypothetical environmental flow regime created using the BBM approach. Source: Tharme and King (1998)

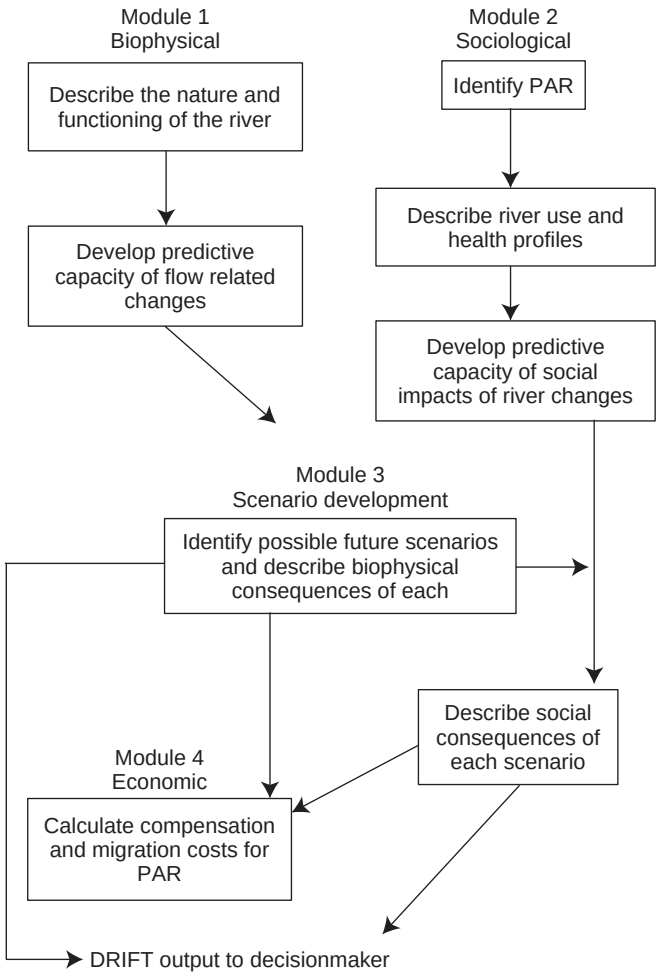


Figure 6 The four modules used in a DRIFT assessment. PAR is human ''populations at risk''. Source: King and Brown (2003), The World Bank, Washington DC

is activated (i.e. they cannot be scaled down) (Poff *et al.*, 1997). This approach attributes lower ecological importance to medium to high-level baseflows, and repeated small and medium-sized flood events, so they are sacrificed in the environmental flow regime. Drought events are normally retained in their entirety by adding the proviso that prevailing natural flows (e.g. as defined by unregulated dam inlet discharge) are applied when they fall below the baseflow level set for that month. Very large floods, although their ecological and geomorphological importance has been acknowledged, are not usually considered in environmental flow assessment because in most situations they cannot be managed and are likely to occur regardless of water infrastructure. Central to the holistic approach of addressing biophysical aspects is the idea of assembling a team of specialists from various disciplines to combine their expert knowledge to recommend a flow regime that satisfies objectives for the particular site or river system. The outcomes of

expert panel deliberations are conditioned by interpersonal dynamics, the potential for groupings to arise within the panel, the impact of a single dominant personality, and the possibility that consensus can be a product of collective bias (rather than independent assessments). For reviews of these issues, see Cottingham *et al.* (2002) and Arthington *et al.* (2004). Bishop (1996) statistically analyzed the results of a study of two expert panels that independently rated a range of flows for their suitability for native fish, invertebrates, and habitat quality at the same six sites (see Swales and Harris, 1995). Of the 18 assessments made, the two panels agreed on only one of them. It was suggested that the differences in the ratings were derived from differences in expertise, from the subjective nature of the assessment, and conflicts between the experts' knowledge of a system and the supplied data (Bishop, 1996; Pusey, 1998). An expert panel assembled to examine the flow needs of the Snowy River, Australia attempted to overcome a number of the typical shortcomings of the expert panel approach by (i) integrating available data with expert opinion in its assessments, and (ii) developing a composite River Condition Index underpinned by a conceptual framework that links habitat and biotic condition and consists of several subindices that translate qualitative assessments (relative to a predisturbance reference condition) into numeric values in a transparent and repeatable manner (Young *et al.*, 2004).

The little use of structured risk assessment procedures in the field of environmental flow assessment has been remarked by Arthington and Pusey (2003). The DRIFT process handles uncertainty in knowledge and data through the use of severity ratings, which allow scientists to indicate, within a coarse range, the likely biophysical response to a change in a single aspect of the hydrological regime. Environmental flow scenarios are thus illustrated as risk envelopes of predicted changes, with wider envelopes indicating greater uncertainty (King and Brown, 2003; King *et al.*, 2003). The Benchmarking methodology attempts to employ a risk assessment framework based on past evidence of the impacts of certain forms of flow regulation (Brizga *et al.*, 2002). The problem of uncertainty can also be handled by adopting a conservative "precautionary principle", and by recommending ongoing monitoring and adaptive management (e.g. Poff *et al.*, 1997; Richter *et al.*, 1997; Environment Protection Authority of NSW, 1997; Arthington and Pusey, 2003).

Prescriptive Versus Interactive Approaches

Brown and King (2003) placed all environmental flow assessment methods into two categories: prescriptive and interactive (Table 1). The prescriptive approach involves the study team spending considerable effort in justifying their derivation of the "best answer" to specific project objectives. The team does not necessarily examine alternative scenarios, and the outputs do not include functions that inform the process of compromise. Interactive approaches, on the other hand, develop relationships between changes in river flow and one or more aspects of the river, allowing multiple interpretations of possible future river health. Brown and King (2003) illustrated the characteristics of the interactive approach by comparing the IFIM and DRIFT methods. Both are essentially problem-solving tools with similar approaches. The output is a set of options, termed "alternatives" in IFIM terminology, and "scenarios" in DRIFT terminology.

SOFTWARE TO ASSIST ENVIRONMENTAL FLOW ASSESSMENTS

In detailed environmental flow assessments, relationships between discharge and certain thresholds, often morphologically defined, which have known or assumed ecological significance, are usually developed with the aid of a one-dimensional (1-D) hydraulic model (for detailed discussions on the use of numerical models in hydrology, *see* **Chapter 17, Hydrological and Environmental Modeling of Transport Processes in Rivers and Estuaries, Volume 1**, **Chapter 18, Shallow Water Models with Porosity for Urban Flood Modeling, Volume 1**). An example of a 1-D model is HEC-RAS, which is public domain software that can be downloaded free of charge from The Hydrologic Engineering Center, US Army Corps of Engineers (`http://www.hec.usace.army.mil/`) or via a link at the Scientific Software Group (`www.scisoftware.com`). The Hydraulic Analysis (HA) module of the Cooperative Research Centre for Catchment Hydrology River

Table 1 Relative data and time requirements of selected flow assessment methods. Source: Brown and King (2003)

Output	Method	Data and time requirements	Approximate duration of assessment	Relative confidence in output
Prescriptive	Tennant method	Moderate to low	2 weeks	Low
	Wetted- perimeter method	Moderate	2–4 months	Low
	Expert panel	Moderate to low	1–2 months	Medium
	Holistic method	Moderate to high	6–18 months	Medium
Interactive	IFIM	Very high	2–5 years	High
	DRIFT	High to very high	1–3 years	High

Analysis Package (RAP), which can be downloaded free of charge (**www.toolkit.net.au/rap**), also has 1-D hydraulic modeling capability, and it can be used to define habitat criteria and calculate the area of habitat at a range of discharges. The HA module can read HEC-RAS output files or users can enter their own channel geometry data. The hydrodynamic module of the DHI Water & Environment MIKE 11 software (**http://www.dhisoftware.com/mike11/**) can also be used to perform the hydraulic modeling for environmental flow assessment. The formulations can be applied to branched and looped networks and quasi 2-D flow simulation on floodplains. MIKE 11 also has advanced cohesive and noncohesive sediment transport modules that can be used in the geomorphological assessment. The Surface-Water Modeling System (SMS) software, available from the Scientific Software Group (**www.scisoftware.com**), provides a high quality graphical user interface for HEC-RAS (also for a number of 2-D models); it has tools for extracting cross sections from digital elevation data, defining and referencing stream reaches using digital maps, generating the HEC-RAS input file, running HEC-RAS directly within the interface, and integrating the HEC-RAS results with floodplain delineation or 2-D hydraulic analysis. Additional SMS modules that may be useful in environmental flows assessment include the modeling of contaminant migration, salinity intrusion, and sediment transport (scour and deposition).

The PHABSIM software and manual (Waddle, 2001), which is used with IFIM, can be downloaded free of charge from the United States Geological Survey (**http://www.fort.usgs.gov/products/software/PHABSIM/PHABSIM.asp**). A number of alternative software systems with the same or similar components as PHABSIM have been developed for use with IFIM. Details of these alternatives, plus other potentially useful software, are provided in Gordon *et al.* (2004, p. 298).

The HA module of the RAP (**www.toolkit.net.au/rap**) is based on the Flow Events Method of allocating environmental flows and can be downloaded free of charge. The Time Series Analysis (TSA) module of RAP has been developed to produce a graphical user interface that allows users to investigate time series data. TSA is especially useful as an interactive tool to assist the deliberations of environmental flow expert panels. A further two modules are under development for inclusion in RAP: the Rules Based Models (RBM) module and Quantitative Models (QM) module will allow prediction of biological responses to alternate flow scenarios.

The Index of Hydrological Alteration (IHA) software calculates a suite of more than 60 ecologically relevant statistics from a daily hydrologic data series. The software can be downloaded free of charge from The Nature Conservancy's Sustainable Waters Program (**http://www.freshwaters.org/tools/**). In an environmental flow assessment of a regulated river, these statistics can be used in the preliminary stages to measure the degree of hydrological change (Richter *et al.*, 1996). The IHA software also includes the Range of Variability Approach (RVA) to support ecologically based management of hydrological systems. This method can be used to help design adaptive management programs that use the quantified natural variation of a hydrological system as an interim management target (Richter *et al.*, 1997).

SOME SPECIFIC ENVIRONMENTAL FLOW ISSUES

The environmental flow issues discussed below are not presented as alternatives to the methodologies discussed previously. Rather, they can be integral components of holistic approaches. There are numerous specific issues that could be included in a discussion of environmental flow methodologies. Assessments of the in-channel flows required for maintenance of fish, macroinvertebrates, plants, and adequate water quality feature in most environmental flow studies – the issues discussed below will also require consideration in many cases. This is not meant to be an exhaustive list, and each case is likely to present some site-specific issues.

Managed Floods for Riparian, Floodplain, and Wetland Inundation

Acreman (2003a) described the concept of reinstatement of flood releases from reservoirs for floodplain inundation, and emphasized the need to ensure that managed floods are compatible with the livelihood strategies of the floodplain communities. Objectives of managed floods include flood recession agriculture, fishing, animal husbandry, groundwater recharge, or conservation of biodiversity and bioproductivity. Managed floods have been demonstrated to result in stimulation of fish breeding, provided they occur during the normal flood period (Welcomme, 1989). One of the main limitations of managed floods is technical feasibility. Many dams do not have adequate outlet structures or sufficient storage capacity, and water quality can be poor due to stratification. Another reality is that the practice of managed flood releases for wetland flooding will mostly involve enhancement (or boosting) of the peak magnitude or duration of naturally occurring floods, rather than generating the entire flood from a dam release (Gordon *et al.*, 2004, p. 303). This is simply a matter of the large volume of water required to achieve overbank flows.

While the flood-pulse concept emphasizes the ecological importance of floods, in practice it may be difficult to identify thresholds of flooding below which floodplain ecosystems cease to function adequately (Welcomme,

1976). Acreman (2003a) suggested that the alternative to attempting to recommend how much water the floodplain ecosystem requires is to set managed flood objectives in terms of specific targets, such as fish production, for which sufficient data may be available to develop a tentative predictive relationship. In other cases, flows necessary to maintain specific floodplain and riparian plant and animal communities are reasonably well known (e.g. Stromberg and Richter, 1991; Blanch *et al.*, 1999; Roberts *et al.*, 2000; Leslie, 2001; Hughes *et al.*, 2001). In some cases, lack of flood flows for plant recruitment is not the problem; rather it is lack of flows to keep riparian plant growth in check (Johnson, 1999).

Managed Flows for Channel Formation

There is overwhelming evidence in the literature that marked alteration of channel forming flow processes is associated with declining ecological health, or degradation of the physical channel attributes required for normal ecological functioning (e.g. Gippel and Collier, 1998; Tharme and King, 1998; Petts and Maddock, 1996; Gippel, 2000; Reiser *et al.*, 1989a,b; Rutherfurd, 2000). Channel forming flows include any flows that have a role in shaping the physical form of the channel, maintaining habitat forms, prevention of vegetation encroachment, and removal of fine sediment and detritus from the surface of the substrate (Reiser *et al.*, 1985, 1989a,b; Brookes, 1995). In environmental flow literature, the terms "flushing flows" and "channel maintenance flows" are also used to describe some of these processes (Reiser *et al.*, 1985; Reiser *et al.*, 1989a,b; Kondolf and Wilcock, 1996; Milhous, 1996; Milhous, 1998; Texas Parks and Wildlife Department, 2003). Channel forming flows can be specified in terms of magnitude, frequency, duration, and timing. Like many aspects of environmental flow assessment, the idea of providing special flows that maintain channel morphology has attracted some criticism. This in part reflects gaps in the knowledge of the way channels form, regional and site-specific differences in the way channels form, and indeterminacy of channel form (Gippel, 2002). Where flows for geomorphic processes have been recommended (e.g. Gippel and Stewardson, 1995; Wilcock *et al.*, 1996; Snowy Water Inquiry, 1998; Kondolf, 1998), they invariably represent only a small part of the natural medium and high flow regime. There is currently only limited understanding of how specific bed and bank features form and are maintained. Low flows can transport sand-sized sediment and cause in-fill of scour pools, so there is a need to also consider the impacts of the duration of low-flow events on river geomorphology (Brizga, 1998).

Reiser *et al.* (1985, 1989a,b), Milhous (1996), Brizga (1998), and Gippel (2002) reviewed the literature on channel forming flows as a component of environmental flow assessment. The approaches generally fell into three categories:

1. *Hydrological* – These methods use an index obtained from runoff records, such as the Tennant and Hoppe Methods.
2. *Morphological* – Channel characteristics such as some percentage of bankfull flow are used to obtain a flushing flow.
3. *Sedimentological* – A channel forming flow is determined by using a sediment transport equation to find the flow level needed to move particles of a given size.

Gippel (2002) suggested that geomorphologists should be prepared to apply a range of methods to the problem of understanding channel forming flows, and not expect a simple and consistent explanation to emerge from the analysis. Natural flow regimes are composed of numerous facets, or components, occurring as a complex time series, and not as discrete, predictable, and independent events. Scaled-down flow regimes, or regimes with certain components culled from the natural regime, should not be expected to produce the same morphology as a natural flow regime. Environmental (regulated) flows will nearly always be simpler and less variable than natural flow regimes, so the resulting geomorphology will probably be less complex (Gippel, 2002).

Assessing Water Requirements of Groundwater-dependent Ecosystems

Nearly all rivers have some level of groundwater interaction. Arthington *et al.* (1992) included groundwater in the original holistic approach and King *et al.* (2000) recognized the need to consider groundwater when doing surface water environmental flow assessments in South Africa, especially for nonperennial rivers, although only a few studies have done so. Cook and Lamontagne (2002) noted that while methods are available for assessing the groundwater dependency of ecosystems, water requirements are much more difficult to determine.

Recognition of the need to consider the water requirements of groundwater-dependent ecosystems is relatively recent (Hatton and Evans, 1998; Petts *et al.*, 1999). Clifton and Evans (2001) developed a conceptual framework for assessing the water requirements for groundwater-dependent ecosystems. The framework is divided into three main components. The first concerns determining the level of groundwater dependency of the ecosystem, which involves identifying the dependent elements of the ecosystem. Next, the key groundwater elements are described in terms of flux (rate of surface or subsurface discharge), level (depth of water table), pressure (applies to confined aquifers), and quality (salinity, nutrients, and contaminants). The second main component of the framework is to assess

the water regime in which the dependency operates. Like surface water assessments, this step involves studies of the ecology, habitat, consumptive use, and hydrological characterization in terms of timing, frequency, and duration. These models are used in the third component of the framework, which is to determine the environmental water requirements.

Assessing Water Requirements of Estuarine Ecosystems

Most of the research and application of environmental flows has been limited to the freshwater reaches of rivers. However, there is a reasonable basis of understanding of the role of freshwater in maintaining estuarine ecosystems [see papers in Montagna *et al.* (2002)]. The conceptual models of Kimmerer (2002) describe the consequences of freshwater inflows to estuaries in terms of physical, chemical, and biological components, or resources. The model of Sklar and Browder (1998) also included the effects of landscape modifications, tidal actions, and solar activity.

Freshwater inflow is known to be a strong determinant of fish abundance or fish catches (Loneragan and Bunn, 1999), and numerical modeling tools for maximizing estuarine fish harvest have been devised on this basis (Bao and Mays, 1994a,b). Salinity is an important variable used to define estuarine processes. The limits of oceanic saline influence have traditionally been defined by the 1 ppt (parts per thousand) salinity isohaline, because this is the upper limit for many agricultural and industrial applications, but a higher limit of 2 ppt (termed "X2"), measured 1 m from the bottom of the water column, is more useful as an ecological indicator (e.g. Kimmerer *et al.*, 1998; Kimmerer and Schubel, 1994; Jassby *et al.*, 1995; Gordon *et al.*, 2004, p. 308–309). In San Francisco Bay, the X2 index was found to be easy to measure, meaningful to non-scientists, and therefore, a good communication tool, it could be directly related to management actions, a historical record was available, and it was correlated with habitat conditions and ecosystem responses (Jassby *et al.*, 1995).

Coastal fisheries production appears to have strong potential as an indicator of freshwater flows and ecosystem health in estuaries (Loneragan and Bunn, 1999; Coastal, 2002). It is considered important by the general public, it has tangible economic value in the form of recreational and commercial fisheries, and time series of data are usually available (i.e. through commercial catch or landing records). Even though habitat availability and fishing pressure also affect fish catch, providing environmental flows to sustain fisheries production has become a key feature of many water management plans in Queensland, Australia (Arthington *et al.*, 2000).

One of the first frameworks for assessing the freshwater flow needs of estuaries was devised by Fruh and Lambert (1976) and Lambert and Fruh (1978) for the Corpus Christi Bay, Texas. In the United Kingdom, Binnie, Black, and Veath Engineering Consultants (1998) provided a systematic procedure for assessing the freshwater flow requirement for estuaries. Estuaries have had little attention compared to rivers when it comes to environmental flow assessments. One problem is that there is often a jurisdictional boundary at the tidal limit, so more attention could be directed to estuaries if they were made a compliance point in riverine environmental flow assessments. A general approach to determining the flow requirements of estuaries (based on a risk framework) has been proposed by Peirson *et al.* (2002). In the Republic of South Africa, the recommended framework for assessing the freshwater flow needs of estuaries is based on the same approach used for freshwater parts of rivers (Department of Water Affairs and Forestry, 1999). Alber (2002) suggested that the management approach can be inflow-based (i.e. there is a limit to which inflows can depart from the hydrologically natural inflows, under the assumption that too much alteration will result in decline in the resources), condition-based (inflow standards are set in order to maintain specified conditions in the estuary), or resource-based (inflow standards are set based on the requirements of specific resources), but each of these is carried out by regulating inflow. Montagna *et al.* (2002) contains several papers that detail methods to assess the freshwater inflow requirements of estuaries in the United States and South Africa.

Assessing Flow Needs for Recreation

Rivers are used for a range of water-based recreational activities, including fishing, kayaking/canoeing/rafting, boating/waterskiing, and associated land-based activities including sightseeing, walking/jogging, camping, picnicking, hunting, and bicycle riding. Recreational flow requirements have been largely ignored outside of the USA where, in certain states, laws specifically designate recreation or aesthetics as beneficial uses for which in-stream flows are protected (Shelby *et al.*, 1992a). The peak in recreational activities usually occurs in the warmer months, which coincides with the period of irrigation demand, potentially creating conflicts in resource management. Each recreation activity has different requirements, and recreation is an experience, not just an activity (Merrill and O'Laughlin, 1993). Thus, recreational users desire protection of natural stream values (namely, water quality, riparian vegetation, natural channel features, adjacent wetlands, and the opportunity to see and hear moving water), as well as provision of the hydraulic conditions that facilitate the particular activities of interest (Merrill and O'Laughlin, 1993; Brown, 2003).

Whittaker *et al.* (1993) evaluated the relationship between stream flows and recreational values, and concluded that one of the most effective methods for evaluating flows for recreation was surveys of users. Brown

and Daniel (1991) and Brown *et al.* (1991) found that recreational quality and positive reactions to scenic beauty increased with discharge up to a point, and then decreased with further flow increases. Suitability for paddling can be assessed by consulting river guides, seeking the opinion of recognized local expert kayakers and rafters, or by conducting a survey using a dedicated crew of paddlers (Shelby *et al.*, 1992b; Rood and Tymensen, 2001; Rood *et al.*, 2003).

Tennant (1976) and Corbett (1990) suggested particular proportions of mean annual flow that would offer suitable conditions for recreational boating and other uses, but this method is relatively crude (Whittaker *et al.*, 1993; Burley, 1990). The transect survey method can be used to identify the flow that will limit any particular recreational activity, which usually requires a minimum depth of water. Bureau of Reclamation (1999) suggested four classes of suitability of flow depth for general recreational uses: 0–45 cm (poor); 45–75 cm (fair); 75–105 cm (good); and 105–135 cm (excellent). For kayaking/canoeing/rafting, Rood and Tymensen (2001) suggested a minimum depth of 60 cm was required to immerse a paddle blade, and depths of 75 cm to 100 cm progressively improved the appeal of many hydraulic features, reduced the chances of hitting rocks, permitted the kayaker to perform a roll, and provided less obstructed conditions for a paddler who swims following a capsize.

IMPLEMENTATION AND EVALUATION OF ENVIRONMENTAL FLOWS

Methods of environmental flow assessment are at best indicative of the flow required to meet the environmental needs of rivers. The best way to test hypotheses regarding environmental flows is to trial full-scale implementation. Acreman (2003b) recommended that three types of responses be monitored: the river flow, the response of the ecosystem, and the social response to ecosystem change. The number of studies worldwide that report environmental flow recommendations greatly outweighs the number of studies that report scientific evaluation of the results of implementation of these recommendations. In fact, very few such studies can be found in the international literature (Annear *et al.*, 2002, p. 309). This is partly explained by the fact that most environmental flows have been implemented relatively recently, and there has been insufficient time to evaluate their effect. Another reality is that some implemented flows are not well monitored, or the results are not published in a readily accessible form. A recent review of cases of scientifically monitored environmental flow implementation (Gordon *et al.*, 2004, p. 315–318) revealed generally positive results. Where the ecological response was negative, neutral, or unexpected, the results of the studies allowed refinement of the environmental flow regimes.

Although adaptive management (Holling, 1978; Walters, 1997) has been used in resource management for around 30 years, it is now being widely promoted as a suitable model for management of river resources (e.g. Annear *et al.*, 2002, p. 309; Ontario Ministry of Natural Resources and Watershed Science Centre, 2001; Newson, 2002). Because it is scientifically based, includes stakeholders, requires attention to documentation, and is based on the idea of implementing interventions as field-scale experiments that generate knowledge useful to management, adaptive management offers a possible framework for implementing and evaluating environmental flows (e.g. Castleberry *et al.*, 1996; McBain and Trush, 1997; Reid and Brooks, 2000; U.S. Fish and Wildlife Service, 2000; Irwin and Freeman, 2002; Freeman, 2002; Arthington *et al.*, 2003; Poff *et al.*, 2003). Gippel (2003) undertook a comprehensive review of all implemented environmental flow-related actions on the River Murray, Australia. Overall, there was a high level of awareness of adaptive management, but no examples could be found where a formal adaptive management framework had been explicitly followed from inception to implementation and adjustment (Gordon *et al.*, 2004, p. 315).

FUTURE DIRECTIONS

The choice of environmental flow method depends on time and budget allocations, data availability, and the level of competition between in-stream and offstream water users. Answers will only be as good as the quality of information examined and the skill of the user in applying the method and interpreting the results. The major trend in in-stream-environmental flow assessment over the past 30 years has been a shift from narrow studies that catered for a single fish species at one critical life phase to a holistic approach that aims to restore natural river processes. For complex and controversial projects on large river systems, there has been a shift from prescriptive to interactive approaches and from bottom-up to top-down approaches

Proponents of the habitat-simulation methods feel that these are more attuned to biological principles than the hydrological index and hydraulic rating techniques, and are thus superior for evaluating biological impacts resulting from flow alterations. Whatever tools are used, it would appear that they are best applied within a holistic framework. None of the methods can, as yet, directly address the potential changes in biomass or populations resulting from altered flow regimes. Many scientists and practitioners recognize the need for more research on the effects of flow regimes on biota because of resultant changes in physical habitat as well as changes in water quality and biological interactions.

The field of in-stream flow assessment is still a dynamic, controversial, and evolving area, with plenty of scope for testing and refinement of methods. The well-established IFIM and the new DRIFT holistic frameworks can be considered the present "state-of-the-art" methods for decision making on large-scale, complex, and controversial projects, but other methods are more than adequate for simpler situations, especially where resources and time are limited. Tharme (2003) noted a widespread move towards hierarchical application of environmental flow methodologies, with at least two stages to the framework: (ii) reconnaissance-level assessment, primarily using hydrological methodologies; and (ii) comprehensive assessment, using either habitat-simulation or holistic methodologies. Holistic methodologies are particularly appropriate in developing countries, where the focus is on protection of the resource at an ecosystem scale, as well as the strong livelihood dependencies on the goods and services provided by aquatic ecosystems (Tharme, 2003).

Rivers often support economically significant industries (i.e. they are "working" rivers) and are important for recreation and navigation. A "healthy working river" is one that is managed to provide a compromise, agreed to by the community, between the condition of the river and the level of human use (Jones *et al.*, 2002; Gordon *et al.*, 2004, p. 239). This definition acknowledges the need for negotiation and compromise between the often-competing values and uses of the river. In situations where river management is based primarily on agreed human uses, the role of scientists would be to work out how to achieve the highest possible level of stream health for the designated class of stream use (Gordon *et al.*, 2004, p. 285). This is an alternative to the approach of determining the minimum flow regime required to maintain the ideal of ecological integrity, with an acceptable level of risk.

The search for an "ideal" universal flow allocation technique may be fruitless, as the practical reality is that environmental flow problems are highly diverse, in terms of the characteristics of the environment, the funds available, the time available, the scope of the study, and the potential cost of making an error in the allocation. Thus, while the importance of developing methods and testing hypotheses cannot be understated, there is a need to ensure that the industry develops the capacity to make prudent use of the outcomes of environmental flows research (Gippel, 2000). At the same time, it is vital that the stakeholders and wider community are informed of, and involved in, developments in the environmental flows field (this has not been a strong feature of progress to date). Lack of understanding can disempower community advisors and stakeholders – a problem that could threaten successful implementation of environmental flows (Gippel, 2001). Poff *et al.* (2003) called for innovative funding partnerships between government agencies, not-for-profit foundations, and the private sector in order to advance the scientific basis of water management. The weakest area of environmental flows endeavors is not the science of assessment, but in the transformation of recommendations into implemented flow regimes (Gordon *et al.*, 2004, p. 315–318). In an extreme example, it took fifty years of research and negotiation to agree on a flow regime that would restore fish populations in the Trinity River (U.S. Fish and Wildlife Service and Hoopa Valley Tribe, 1999; U.S. Fish and Wildlife Service, 2000). Implemented environmental flows require performance evaluation, ideally within an adaptive management framework, which encourages refinement of the flow recommendations and the flow assessment methodologies.

Acknowledgment

Fluvial Systems Pty Ltd funded preparation of this article. Ms Fatima Basic, School of Anthropology, Geography, and Environmental Science, The University of Melbourne prepared the figures.

REFERENCES

Acreman M.C. (2003a) Environmental flows: flood flows. In *Water Resources and Environment Technical Note C3*, Davis R. and Hirji R. (Eds.), The World Bank: Washington, Available on the World Wide Web: **http://lnweb18.worldbank.org/ESSD/ardext.nsf/18ByDocName/EnvironmentalFlowAssessment-NOTEC3EnvironmentalFlowsFloodFlows/$FILE/NoteC3EnvironmentalFlowAssessment2003.pdf** (accessed 1st November 2003).

Acreman M.C. (2003b) Defining water requirements. In *Flow: the Essentials of Environmental Flows*, Dyson M. Bergkamp G. and Scanlon J. (Eds.), Water and Nature Initiative, International Union for Conservation of Nature and Natural Resources: Gland, Switzerland and Cambridge, pp. 11–29, URL: **http://www.waterandnature.org/pub/FLOW.pdf** (accessed 1st November 2003).

Alber M. (2002) A conceptual model of estuarine freshwater inflow management. *Estuaries*, **25**, 1246–1261.

Annear T., Chisholm I., Beecher H., Locke A., Aarrestad P., Burkardt N., Coomer C., Estes C., Hunt J., Jacobson R., *et al.* (2002) *Instream Flows for Riverine Resource Stewardship*, Instream Flow Council: Cheyenne.

Arthington A.H. (1998) *Comparative Evaluation of Environmental Flow Assessment Techniques: Review of Holistic Methodologies*, LWRRDC Occasional Paper No. 26/98, Land and Water Resources Research & Development Corporation: Canberra, Australian Capital Territory.

Arthington A.H. and Pusey B.J. (1993) In-stream flow management in Australia: methods, deficiencies and future directions. *Australian Biology*, **6**, 52–60.

Arthington A.H. and Pusey B.J. (2003) Flow restoration and protection in Australian rivers. *River Research and Applications*, **19**(5–6), 37–395.

Arthington A.H., Brizga S., Bunn S.E. and Loneragan N.R. (2000) *Burnett Basin WAMP, Current Environmental*

Conditions and Impacts of Existing Water Resource Development, Appendix J, Estuarine and Marine Ecosystems, The State of Queensland, Department of Natural Resources: Brisbane, June, Available on the World Wide Web: **http://www.nrm.qld.gov.au/wrp/pdf/burnett/vol2/AppJ.pdf** (accessed 1st November 2004).

Arthington A.H., King J.M., O'Keefe J.H., Bunn S.E., Day J.A., Pusey B.J., Bluhdorn D.R. and Tharme R. (1992) Development of an holistic approach for assessing environmental flow requirements of riverine ecosystems. In *Proceedings of an International Seminar and Workshop on Water Allocation for the Environment*, Pigram J.J. and Hooper B.P. (Eds.), Centre for Water Policy Research, University of New England: Armidale, New South Wales, pp. 69–76.

Arthington A.H., Rall J.L., Kennard M.J. and Pusey B.J. (2003) Environmental flow requirements of fish in Lesotho rivers using the DRIFT methodology. *River Research and Applications*, **19**(5–6), 641–666.

Arthington A.H., Tharme R., Brizga S.O., Pusey B.J. and Kennard M.J. (2004) Environmental flow assessment with emphasis on holistic methodologies. In *Proceedings of LARS2 (Second Large Rivers Symposium on Fisheries)*, Welcomme R. and Petr T. (Eds.), Mekong River Commission: Phnom Penh, p. 24.

Arthington A.H. and Zalucki J.M. (Eds.) (1998) *Comparative Evaluation of Environmental Flow Assessment Techniques: Review of Methods*, LWRRDC Occasional Paper No. 27/98, Land and Water Resources Research & Development Corporation: Canberra, Australian Capital Territory.

Bao Y. and Mays L.W. (1994a) New methodology for optimisation of freshwater inflows to estuaries. *Journal of Water Resources Planning and Management, ASCE*, **102**(2), 199–217.

Bao Y. and Mays L.W. (1994b) Optimization of freshwater inflows to Lavaca-tres Palacios, Texas, Estuary. *Journal of Water Resources Planning and Management, ASCE*, **102**(2), 218–236.

Barinaga M. (1996) A recipe for river recovery? *Science*, **273**, 1648–1650.

Binnie, Black and Veath Engineering Consultants (1998) *Determining the Freshwater Flow Needs of Estuaries*, R&D Technical Report W113, Environment Agency, Bristol. Technical summary available on the World Wide Web: **http://www.eareports.com/ea/rdreport.nsf/0/f9b3b062221fd7cf802567e80054efd5?OpenDocument** (accessed 1st November 2003).

Bishop K. (1996) *Review of the "Expert Panel" (EPAM) Process as a Mechanism for Determining Environmental Flow Releases for Freshwater Fish*, Report to the Centre for Water Policy Research, University of New England, on behalf of the Snowy Mountains Hydro-Electric Authority: Cooma, New South Wales.

Blanch S.J., Ganf G.G. and Walker K.F. (1999) Tolerance of riverine plants to flooding and exposure indicated by water regime. *Regulated Rivers: Research & Management*, **15**(1–3), 43–62.

Bovee K.D. (1982) *A Guide to Stream Habitat Assessment Using the Instream Flow Incremental Methodology*, Instream Flow Information Paper 12, FWS/OBS-82/26, Cooperative Instream Flow Services Group, U.S. Fish and Wildlife Service: Fort Collins.

Bovee K.D. and Milhous R. (1978) *Hydraulic Simulation in Instream Flow Studies: Theory and Techniques*, Report FWS/OBS-78/33, U.S. Fish and Wildlife Service: Fort Collins.

Brizga S. (1998) Methods addressing flow requirements for geomorphological purposes. In *Comparative Evaluation of Environmental Flow Assessment Techniques: Review of Methods*, Occasional Paper No 27/98, Arthington A.H. and Zalucki J.M. (Eds.), Land and Water Resources Research & Development Corporation: Canberra, Australian Capital Territory, pp. 8–46.

Brizga S.O., Arthington A.H., Pusey B.J., Kennard M.J., Mackay S.J., Werren G.L., Craigie N.M. and Choy S.J. (2002) Benchmarking, a 'top-down' methodology for assessing environmental flows in Australian rivers. In *Proceedings, Environmental Flows in River Systems. An International Working Conference on Assessment and Implementation, incorporating the 4th International Ecohydraulics Symposium*, Cape Town, March 2002 (published on CD).

Brookes A. (1995) The importance of high flows for riverine environments. In *The Ecological Basis for River Management*, Harper D.M. and Ferguson A.J.D. (Eds.), John Wiley & Sons: Chichester, pp. 33–49.

Brown C.A. and King J.M. (2000) *A summary of the DRIFT process, environmental flow assessments for rivers. Southern Waters' Information* Report No. 01/00. Southern Waters, Ecological Research and Consulting, Mowbray, Republic of South Africa, August.

Brown C.A. and King J.M. (2003) Environmental Flows: Concepts and Methods. Water Resources and Environment Technical Note C1. In *The World Bank Water Resources and Environment Technical Note Series*, Davis R. and Hirji R. (Eds.), The World Bank: Washington, DC, pp. 28.

Brown T.C. (2003) Water availability and recreational opportunities. In *Riparian Areas of the Southwestern United States: Hydrology, Ecology, and Management*, Baker M.B., Ffolliott P.F., DeBano L.F. and Neary D.G. (Eds.), Lewis Publishers: New York, pp. 299–314.

Brown T.C. and Daniel T.C. (1991) Landscape aesthetics of riparian environments: relationship of flow quantity to scenic quality along a wild and scenic river. *Water Resources Research*, **27**(8), 1787–1795.

Brown C.A. and King J.M. (2003) Environmental flows: concepts and methods. In *Water Resources and Environment Technical Note C1*, Davis R. and Hirji R. (Eds.), The World Bank: Washington, Available on the World Wide Web: **http://lnweb18.worldbank.org/ESSD/ardext.nsf/18ByDocName/EnvironmentalFlowAssessment-NOTEC1EnvironmentalFlowsConceptsandMethods/$FILE/NoteC1EnvironmentalFlowAssessment2003.pdf** (accessed 1st November 2003).

Brown T.C., Taylor J. and Shelby B. (1991) Assessing the direct effects of streamflow on recreation: a literature review. *Water Resources Bulletin*, **27**(6), 979–989.

Bunn S.E. and Arthington A.H. (2002) Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. *Environmental Management*, **30**, 492–507.

Bureau of Reclamation (1999) Red River valley municipal, rural, and industrial water needs assessment Phase I, Part B, instream flow needs assessment. *Sheyenne River and Red River*

of the North, North Dakota and Minnesota, Final Appraisal Report, United States Department of the Interior, Bureau of Reclamation, Dakotas Area Office: Bismarck, August.

Burley J.B. (1990) Advancing recreation assessments. *Rivers*, **1**, 236–239.

Caissie D. and El–Jabi N. (2003) Instream flow assessment: from holistic approaches to habitat modelling. *Canadian Water Resources Journal*, **28**(2), 173–184.

Caissie D., El-Jabi N. and Bourgeois G. (1998) Évaluation du débit réservé par méthodes hydrologiques et hydrobiologiques. *Revue Des Sciences De l'Eau*, **11**(3), 347–364.

Castleberry D.T., Czech J.J., Erman D.C., Hankin D., Healey M., Kondolf G.M., Mangel M., Mohr M., Moyle P.B., Nielsen J., *et al.* (1996) Uncertainty and instream flow standards. *Fisheries*, **21**(8), 20–21.

Clifton C. and Evans R. (2001) *Environmental Water Requirements of Groundwater Dependent Ecosystems*, Environmental flows Initiative Technical Report, Report No. 2, Environment Australia, Commonwealth of Australia: Canberra, Australian Capital Territory. Available on the World Wide Web: **http://www.deh.gov.au/water/rivers/nrhp/groundwater/** (accessed 1st November 2003).

Coastal C.R.C. (2002) Flowing estuaries needed for healthy fisheries. *Exploring Coastal Science*, Cooperative Research Centre for Coastal Zone, Estuary and Waterway Management: Indooroopilly, May, Available on the World Wide Web: **http://www.coastal.crc.org.au/pdf/exploring_coastal_science/flow_estuaries.pdf** (accessed 1st November 2003).

Collings M.R. (1972) A methodology for determining instream flow requirements for fish. *Proceedings of Instream Flow Methodology Workshop*, Washington Department of Ecology: Olympia, pp. 72–86.

Cook P. and Lamontagne S. (2002) Assessing and protecting water requirements for groundwater dependent ecosystems. *Seminar Proceedings, The Science of Environmental Water Requirements in South Australia*, The Hydrological Society of South Australia: Adelaide, pp. 49–54, 24 September.

Corbett R. (1990) *A Method for Determining Minimum Instream Flow for Recreational Boating*, SAIC Special Report 1-239-91-01, Science Applications International Corporation: McLean.

Cottingham P., Thoms M.C. and Quinn G.P. (2002) Scientific panels and their use in environmental flow assessment in Australia. *Australian Journal of Water Resources*, **5**, 103–111.

Department of Water Affairs and Forestry (1999) *Resource Directed Measures for Protection of Water Resources*, *Vol. 5: Estuarine Ecosystems Version 1.0*: Pretoria, September, Available on the World Wide Web: **http://www.dwaf.gov.za/Documents/Policies/WRPP/Estuarine%20Ecosystems.htm** (accessed 1st November 2003).

Dunbar M.J., Gustard A., Acreman M.C. and Elliott C.R.N. (1998) *Review of Overseas Approaches to Setting River Flow Objectives*, Environment Agency R&D Technical Report W6B(96)4, Institute of Hydrology: Wallingford.

Dyson M. (2003) Getting started. In *Flow: The Essentials of Environmental Flows*, Dyson M. Bergkamp G. and Scanlon J. (Eds.), Water and Nature Initiative, International Union for Conservation of Nature and Natural Resources: Gland, Switzerland and Cambridge, pp. 1–8, Available on the World Wide Web: **http://www.waterandnature.org/pub/FLOW.pdf** (accessed 1st November 2003).

Environment Protection Authority of NSW (1997) *Proposed Interim Environmental Objectives for NSW Waters – Inland Rivers*, New South Wales EPA: Chatswood.

Espegren G.D. (1998) *Evaluation of the Standards and Methods Used for Quantifying Instream Flows in Colorado*, Colorado Water Conservation Board: Denver, November. Available on the World Wide Web: **http://cwcb.state.co.us/isf/Programs/Docs/evalstan.pdf** (accessed 1st November 2003).

Freeman R. (2002) Harnessing the restoration potential of artificial floods. *Conservation in Practice*, **3**(2), Society for Conservation Biology, [online] Available on the World Wide Web: **http://www.conbio.org/InPractice/article32HAR.html** (accessed 1st November 2003).

Fruh E.G. and Lambert W.P. (1976) Methodology to evaluate alternative coastal zone management policies: application in the Texas coastal zone. *Special Report III: A Methodology for Investigating Fresh Water Inflow Requirements of a Texas Estuary*, Centre for Research in Water Resources, University of Texas: Austin, p. 348.

Gippel C.J. (2000) Managing regulated rivers for environmental values: selected case studies from Southeastern Australia. In *River Management, the Australian Experience*, Brizga S. and Finlayson B. (Eds.), John Wiley & Sons: Chichester, pp. 97–122.

Gippel C.J. (2001) Australia's environmental flow initiative: filling some knowledge gaps and exposing others. *Water Science and Technology*, **43**(9), 73–88.

Gippel C.J. (2002) Geomorphic issues associated with environmental flow assessment in alluvial non-tidal rivers. *Australian Journal of Water Resources*, **5**(1), 3–19.

Gippel C.J. (2003) *Review of Achievements and Outcomes of Environmental Flow Initiatives Undertaken on the Extended River Murray System to August 2002*, Report by Fluvial Systems Pty Ltd: Stockton, to Murray-Darling Basin Commission: Canberra, Australian Capital Territory, March, Available on the World Wide Web: **http://www.thelivingmurray.mdbc.gov.au/content/item.phtml?itemId=10287&nodeId=file3efacce9b46f8&fn=Review_of_Achievements_OutcomesV10.pdf** (accessed 1st November 2003).

Gippel C.J. and Collier K.J. (1998) Degradation and rehabilitation of waterways in Australia and New Zealand. In *Rehabilitation of Rivers: Principles and Implementation*, DeWaal L.C., Large A.R.G. and Wade P.M. (Eds.), John Wiley & Sons: Chichester, pp. 269–300.

Gippel C.J., Jacobs T. and McLeod T. (2002) Determining environmental flow needs and scenarios for the River Murray System, Australia. *Australian Journal of Water Resources*, **5**(1), 61–74.

Gippel C.J. and Stewardson M.J. (1995) Development of an environmental flow management strategy for the Thomson River, Victoria, Australia. *Regulated Rivers: Research & Management*, **10**, 121–135.

Gippel C.J. and Stewardson M.J. (1998) Use of wetted perimeter in defining minimum environmental flows. *Regulated Rivers: Research & Management*, **14**, 53–67.

Gordon N.D., McMahon T.A., Finlayson B.L., Gippel C.J. and Nathan R.J. (2004) *Stream Hydrology: An Introduction for Ecologists*, *Second Edition*, John Wiley & Sons: Chichester.

Gore J.A. and Nestler J.M. (1988) Instream flow studies in perspective. *Regulated Rivers*, **2**, 93–101.

Hatton T. and Evans R. (1998) *Dependence of Ecosystems on Groundwater and its Significance to Australia*, LWRRDC Occasional Paper No. 12/98, Land and Water Resources Research & Development Corporation: Canberra, Australian Capital Territory.

Holling C.S. (Ed.) (1978) *Adaptive Environmental Assessment and Management*, John Wiley & Sons: Chichester.

Hoppe R.A. (1975) *Minimum Streamflows for Fish*, Paper distributed at the Soils-Hydrology Workshop, 26–30 January 1976, U.S. Forest Service, Montana State University: Bozeman.

Hughes F.M.R., Adams W.M., Muller E., Nilsson C., Richards K.S., Barsoum N., Decamps H., Foussadier R., Girel J., Guilloy H., *et al.* (2001) The importance of different scale processes for the restoration of floodplain woodlands. *Regulated Rivers: Research & Management*, **17**(4–5), 325–345.

International Water Management Institute (2003) *Environmental Flow Assessment for Aquatic Ecosystems: A Database of Methodologies*, Tharme R (Moderator), Ecohydrological Databases, IWMI, Headquarters: Colombo, Available on the World Wide Web: **http://www.lk.iwmi.org/ehdb/EFM/efm.asp** (accessed 1st November 2003).

Irvine J.R., Jowett I.G. and Scott D. (1987) A test of the instream flow incremental methodology for underyearling rainbow trout, *Salmo gairdnerii*, in experimental New Zealand streams. *NZ Journal of Marine and Freshwater Research*, **21**, 35–40.

Irwin E.R. and Freeman M.C. (2002) Proposal for adaptive management to conserve biotic integrity in a regulated segment of the Tallapoosa River, Alabama, U.S.A. *The Journal of the Society for Conservation Biology*, **16**(5), 1212–1222.

Jassby A.D., Kimmerer W.J., Monismith S., Armor C., Cloern J.E., Powell T.M., Schubel J.R. and Vendlinski T. (1995) Isohaline position as a habitat indicator for estuarine resources-San Francisco Bay-Delta, California, U.S.A. *Ecological Applications*, **5**, 272–289.

Johnson W.C. (1999) Tree recruitment and survival in rivers: influence of hydrological processes. *Hydrological Processes*, **14**(16–17), 3051–3074.

Jones G., Hillman T., Kingsford R., McMahon T., Walker K., Arthington A., Whittington J. and Cartwright S. (2002) *Independent Report of the Expert Reference Panel on Environmental Flows and Water Quality Requirements for the River Murray System*, Cooperative Research Centre for Freshwater Hydrology, Murray-Darling Basin Ministerial Council: Canberra, Australian Capital Territory, Available on the World Wide Web: **http://www.thelivingmurray.mdbc.gov.au/content/index.phtml/itemId/10313/fromItemId/4484** (accessed 28th October 2004).

Jowett I.G. (1997) Instream flow methods: a comparison of approaches. *Regulated Rivers: Research and Management*, **13**, 115–127.

Karr J.R. and Dudley D.R. (1981) Ecological perspectives on water quality goals. *Environmental Management*, **5**, 55–68.

Katopodis C. (2003) Case studies of instream flow modelling for fish habitat. *Canadian Water Resources Journal*, **28**(2), 199–216.

Kimmerer W.J. (2002) Physical, biological, and management responses to variable freshwater flow into the San Francisco estuary. *Estuaries*, **25**, 1275–1290.

Kimmerer W.J., Burau J.R. and Bennett W.A. (1998) Tidally-oriented vertical migration and position maintenance of zooplankton in a temperate estuary. *Limnology and Oceanography*, **43**, 1697–1709.

Kimmerer W.J. and Schubel J.R. (1994) Managing freshwater flows into San Francisco Bay using a salinity standard: results of a workshop. In *Changes in Fluxes in Estuaries: Implications from Science to Management*, Dyer K.R. and Orth R.J. (Eds.), Olsen and Olsen: Fredensborg, pp. 411–416.

King J.M. and Brown C.A. (2003) Environmental flows: case studies. In *Water Resources and Environment Technical Note C2*, Davis R. and Hirji R. (Eds.), The World Bank: Washington, Available on the World Wide Web: **http://lnweb18.worldbank.org/ESSD/ardext.nsf/18ByDocName/EnvironmentalFlowAssessment-NOTEC2EnvironmentalFlowsCaseStudies/$FILE/NoteC2EnvironmentalFlowAssessment2003.pdf** (accessed 1st November 2003).

King J., Brown C. and Sabet H. (2003) A scenario-based holistic approach to environmental flow assessments for rivers. *River Research and Applications*, **19**(5–6), 619–639.

King J.M. and Louw D. (1998) Instream flow assessments for regulated rivers in South Africa using the Building Block Methodology. *Aquatic Ecosystem Health and Restoration*, **1**, 109–124.

King J.M. and Tharme R.E. (1993) *Assessment of the Instream Flow Incremental Methodology and Initial Development of Alternate Instream Flow Methodologies for South Africa*, Water Research Commission Report No. 295/1/94, Freshwater Research Unit, University of Cape Town: Cape Town.

King J.M., Tharme R.E. and Brown C.A. (1999) Definition and implementation of instream flows. *Thematic Report for the World Commission on Dams*, Southern Waters Ecological Research and Consulting: Cape Town, Available on the World Wide Web: **http://www.dams.org/docs/kbase/contrib/env238.pdf** (accessed 1st November 2003).

King J.M., Tharme R.E. and De Villiers M.S. (Eds.) (2000) *Environmental Flow Assessments for Rivers: Manual for the Building Block Methodology*, WRC Report No. TT 131/00, Freshwater Research Unit, University of Cape Town. Water Research Commission: Pretoria, July.

Kondolf G.M. (1998) Development of flushing flows for channel restoration on Rush Creek, California. *Rivers*, **6**(3), 183–193.

Kondolf G.M. and Wilcock P.R. (1996) The flushing flow problem: defining and evaluating objectives. *Water Resources Research*, **32**(8), 2589–2599.

Lambert W.P. and Fruh E.G. (1978) A methodology for investigating freshwater inflow requirements for a Texas estuary. In *Estuarine Interactions*, Wiley M.L. (Ed.), Academic Press: New York, pp. 403–413.

Leslie D.J. (2001) Effect of river management on colonially-nesting waterbirds in the Barmah-Millewa Forest, south-eastern

Australia. *Regulated Rivers: Research and Management*, **17**, 21–36.

Loneragan N.R. and Bunn S.E. (1999) River flows and estuarine ecosystems: implications for coastal fisheries from a review and a case study of the Logan River, southeast Queensland. *Australian Journal of Ecology*, **24**(4), 431–440.

Mathur D., Bason W.H., Purdy E.J. and Silver C.A. Jr (1985) A critique of the instream flow incremental methodology. *Canadian Journal of Fisheries and Aquatic Sciences*, **42**, 825–831.

McBain & Trush (1997) *Trinity River Maintenance Flow Study Final Report*, Hoopa Valley Tribe Fisheries Department: Hoopa, CA, November.

McCarthy J.H. (2003) *Wetted Perimeter Assessment Shoal Harbour River, Shoal Harbour, Clarenville, Newfoundland*, Report by AMEC Earth & Environmental Limited: St. John's, to SGE-Acres: Clarenville, January, Available on the World Wide Web: **http://www.gov.nf.ca/env/Env/EA%202001/pdf%20files/1059%20-%20WettedPerimeterAssessment.pdf** (accessed 1st November 2003).

Merrill T. and O'Laughlin J. (1993) *Analysis of Methods for Determining Minimum Instream Flows*, Idaho Forest, Wildlife and Range Policy Analysis Group Report No. 9, College of Natural Resources Policy Analysis Group, University of Idaho: Moscow, March, Available on the World Wide Web: **http://www.cnr.uidaho.edu/pag/pag9es.html** (accessed 1st November 2003).

Milhous R.T. (1996) Modeling of instream flow needs: the link between sediment and aquatic habitat. In *Proceedings of the Second IAHR Symposium on Habitat Hydraulics, Ecohydraulics 2000*, Leclerc M., Capra H., Valentin S., Boudreault A. and Côté Y. (Eds.), Institute National de la Recherche Scientifique – Eau, co-published with FQSA, IAHR/AIRH: Ste-Foy, Québec, pp. A319–A330.

Milhous R.T. (1998) Modeling of instream flow needs: the link between sediment and aquatic habitat. *Regulated Rivers: Research and Management*, **14**, 79–94.

Montagna P.A., Alber M. and Doering P.H. (2002) Freshwater inflow: science, policy, management. Symposium papers from the 16th Biennial Estuarine Research Federation conference, St. Pete Beach, Florida. November 4–8, 2001. Dedicated issue of *Estuaries*, **25**(6B) 1243–1456.

Mosley M.P. and Jowett I.G. (1985) Fish habitat analysis using river flow simulation. *NZ Journal of Marine and Freshwater Research*, **19**, 293–309.

Nathan R., Doeg T. and Voorwinde L. (2002) Towards defining sustainable limits to winter diversions in Victorian catchments. *Australian Journal of Water Resources*, **5**(1), 49–60.

Nelson F.A. (1980) *Evaluation of four instream flow methods applied to four trout rivers in southwest Montana*. Final report to U.S. Fish and Wildlife Service, Contract No. 14-16-0006-78-48. Montana Department of Fish, Wildlife and Parks: Boseman, Montana. in Montana.

Nestler J.M., Milhous R.T. and Layzer J.B. (1989) Instream habitat modeling techniques. In *Alternatives in Regulated River Management*, Gore J.A. and Petts G.E. (Eds.), CRC Press: Boca Raton, pp. 295–315.

Newson M.D. (2002) Geomorphological concepts and tools for sustainable river ecosystem management. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **12**(4), 365–379.

Ontario Ministry of Natural Resources and Watershed Science Centre (2001) *Adaptive Management of Stream Corridors in Ontario*, Natural Channel Systems. Watershed Science Center, Trent University: Peterborough (CD-ROM).

Orth D.J. and Leonard P.M. (1990) Comparison of discharge methods and habitat optimization for recommending instream flows to protect fish habitat. *Regulated Rivers*, **5**, 129–138.

Orth D.J. and Maughan O.E. (1981) Evaluation of the "Montana Method" for recommending instream flows in Oklahoma streams. *Proceedings of the Oklahoma Academy of Science*, **61**, 62–66.

Orth D.J. and Maughan O.E. (1982) Evaluation of the Incremental Methodology for recommending instream flows for fishes. *Transactions American Fisheries Society*, **111**, 413–445.

Peirson W.L., Bishop K., Van Senden D., Horton P.R. and Adamantidis C.A. (2002) *Environmental Water Requirements to Maintain Estuarine Processes*, Environmental Flows Initiative Technical Report, Report No. 3, Environment Australia, Commonwealth of Australia: Canberra, Australian Capital Territory. Available on the World Wide Web: **http://www.deh.gov.au/water/rivers/nrhp/estuarine/index.html** (accessed 1st November 2003).

Petts G.E., Bickerton M.A., Crawford C., Lerner D.N. and Evans D. (1999) Flow management to sustain groundwater-dominated stream ecosystems. *Hydrological Processes*, **13**, 497–513.

Petts G.E. and Maddock I. (1996) Flow allocation for in-river needs. In *River Restoration*, Petts G.E. and Calow P. (Eds.), Blackwell Science: Oxford, pp. 60–79.

Poff N.L., Allan J.D., Bain M.B., Karr J.R., Prestegaard K.L., Richter B.D., Sparks R.E. and Stromberg J.C. (1997) The natural flow regime, a paradigm for river conservation and restoration. *BioScience*, **47**, 769–784.

Poff N.L., Allan J.D., Palmer M.A., Hart D.D., Richter B.D., Arthington A.H., Rogers K.H., Meyer J.L. and Stanford J.A. (2003) River flows and water wars: emerging science for environmental decision making. *Frontiers in Ecology and the Environment*, **1**(6), 298–306.

Puckridge J.T., Sheldon F., Walker K.F. and Boulton A.J. (1998) Flow variability and the ecology of large rivers. *Marine and Freshwater Research*, **49**, 55–72.

Pusey B.J. (1998) Methods addressing the flow requirements of fish. In *Comparative Evaluation of Environmental Flow Assessment Techniques: Review of Methods*, LWRRDC Occasional Paper No 27/98, Arthington A.H. and Zalucki J.M. (Eds.), Land and Water Resources Research & Development Corporation: Canberra, Australian Capital Territory, pp. 66–105.

Pusey B.J., Kennard M.J. and Arthington A.H. (2000) Discharge variability and the development of predictive models relating stream fish assemblage structure to habitat in north-eastern Australia. *Ecology of Freshwater Fishes*, **9**, 30–50.

Quinn F. (1991) As long as the rivers run: the impacts of corporate water development on native communities in Canada. *The Canadian Journal of Native Studies*, **11**(1), 137–154.

Reid M.A. and Brooks J.J. (2000) Detecting effects of environmental water allocations in wetlands of the Murray-Darling basin, Australia. *Regulated Rivers: Research & Management*, **16**, 479–496.

Reiser D.W., Ramey M.P. and Lambert T.R. (1985) *Review of Flushing Flow Requirements in Regulated Streams*, Pacific Gas and Electric Company, Department of Engineering Research: San Ramon, February.

Reiser D.W., Ramey M.P. and Wesche T.A. (1989a) Flushing flows. In *Alternatives in Regulated River Management*, Gore J.A. and Petts G.E. (Eds.), CRC Press: Boca Raton, pp. 91–138.

Reiser D.W., Wesche T.A. and Estes C. (1989b) Status of instream flow legislation and practise in North America. *Fisheries*, **14**(2), 22–29.

Richter B.D., Baumgartner J.V., Powell J. and Braun J.P. (1996) A method for assessing hydrological alteration within ecosystems. *Conservation Biology*, **10**(4), 1163–1174.

Richter B.D., Baumgartner J.V., Wigington R. and Braun D.P. (1997) How much water does a river need? *Freshwater Biology*, **37**, 231–249.

Roberts J., Young W.J. and Marston F. (2000) *Estimating the Water Requirements for Plants of Floodplain Wetlands: A Guide*, LWRRDC Occasional Paper No. 04/00, Land and Water Resources Research and Development Corporation: Canberra, Australian Capital Territory.

Rood S.B. and Tymensen W. (2001) *Recreational Flows for Paddling Along Rivers in Southern Alberta*, Chinook Environmental Resources and Department Biological Sciences, University of Lethbridge and Alberta Environment: Lethbridge, February, Available on the World Wide Web: **http://www3.gov.ab.ca/env/water/regions/ssrb/pdf_phase2/OldmanR%20Rec%20Flows%20report%20FINAL1.pdf** (accessed 1st November 2003).

Rood S.B., Tymensen W. and Middleton R. (2003) A comparison of methods for evaluating instream flow needs for recreation along rivers in southern Alberta, Canada. *River Research and Applications*, **19**(2), 123–135.

Rushton, C.D. (2000) *Instream flows in Washington State, Past, Present and Future (White paper)*, Water Resources Program, Department of Ecology: Olympia, July, Available on the World Wide Web: **http://www.olympus.net/community/dungenesswc/InstreamFlowversion12.PDF** (accessed 1st November 2003).

Rutherfurd I.D. (2000) Some human impacts on Australian stream channel morphology. In *River Management: The Australasian Experience*, Brizga S.O. and Finlayson B.L. (Eds.), John Wiley & Sons: Chichester, pp. 11–49.

Scatena F.N. (2004) A survey of methods for setting minimum instream flow standards in the Caribbean basin. *River Research and Applications*, **20**(2), 127–135.

Scott D. and Shirvell C.S. (1987) A critique of the instream flow incremental methodology and observations on flow determination in New Zealand. In *Regulated Streams – Advances in Ecology*, Kemper J.B. and Craig J. (Eds.), Plenum Press: New York, pp. 27–44.

Shelby B., Brown T.C. and Baumgartner R. (1992b) Effects of streamflows on river trips on the Colorado River in Grand Canyon, Arizona. *Rivers*, **3**(3), 191–201.

Shelby B., Brown T.C. and Taylor J.G. (1992a) *Streamflow and Recreation*, General Technical Report RM-209, USDA Forest Service Rocky Mountain Forest and Range Experiment Station: Fort Collins, p. 27.

Sklar F.H. and Browder J.A. (1998) Coastal environmental impacts brought about by alterations to freshwater flow in the Gulf of Mexico. *Environmental Management*, **22**(4), 547–562.

Snowy Water Inquiry (1998) *Final Report*, Snowy Water Inquiry: Sydney.

Stalnaker C.B. (1979) The use of habitat structure preferenda for establishing flow regimes necessary for maintenance of fish habitat. In *The Ecology of Regulated Streams*, Ward J.V. and Stanford J.A. (Eds.), Plenum Press: New York, pp. 321–337.

Stewardson M.J. and Gippel C.J. (1997) *In-Stream Environmental Flow Design: A Review*, Report to the Hydro-Electric Corporation: Tasmania, Department of Civil and Environmental Engineering, The University of Melbourne: Parkville, p. 99.

Stewardson M.J. and Gippel C.J. (2003) Incorporating flow variability into environmental flow regimes using the Flow Events Method. *River Research and Applications*, **19**(5–6), 459–472.

Stromberg J.C. and Richter B.D. (1991) Flood flows and dynamics of Sonoran riparian forests. *Rivers*, **2**(3), 221–235.

Swales S. and Harris J. (1995) The expert panel assessment method (EPAM): a new tool for determining environmental flows in regulated rivers. In *The Ecological Basis for River Management*, Harper D.M. and Ferguson A.J.D. (Eds.), Wiley: Chichester, pp. 125–134.

Tennant D.L. (1976) Instream flow regimens for fish, wildlife, recreation and related environmental resources. *Fisheries*, **1**(4), 6–10.

Texas Parks and Wildlife Department, Texas Commission on Environmental Quality and Texas Water Development Board (2003) *Texas Instream Flow Studies: Technical Overview*, Draft, Texas Water Development Board: Austin, August, Available on the World Wide Web: **http://www.twdb.state.tx.us/assistance/InstreamFlows/InstreamFlows-Draft-TechnicalOverviewForNAS.pdf** (accessed 1st November 2003).

Tharme R.E. (2003) A global perspective on environmental flow assessment: emerging trends in the development and application of environmental flow methodologies for rivers. *River Research and Applications*, **19**(5–6), 397–441.

Tharme R.E. and King J.M. (1998) *Development of the Building Block Methodology for Instream Flow Assessments and Supporting Research on the Effects of Different Magnitude Flows on Riverine Ecosystems, Freshwater Research Unit*, WRC Report No. 576/1/98, University of Capetown, Capetown.

U.S. Fish and Wildlife Service (2000) *Trinity River Mainstream Fishery Restoration Environmental Impact Statement/Report*, U.S. Fish and Wildlife Service: Sacramento, October, Available on the World Wide Web: **http://www.ccfwo.r1.fws.gov/treis/default.htm** (accessed 1st Nov 2003).

U.S. Fish and Wildlife Service and Hoopa Valley Tribe (1999) *Trinity River Flow Evaluation, Final Report*, Report to U.S. Department of Interior by USFWS, Arcata Fish and Wildlife Office: Arcata, and Hoopa Valley Tribe: Hoopa, June, Available on the World Wide Web:

http://arcata.fws.gov/fisheries/trfefinal.html (accessed 1st November 2003).

Van Winkle W., Coutant C.C., Jager H.I., Mattice J.S., Orth D.J., Otto R.G., Railsback S.F. and Sale M.J. (1997) Uncertainty and instream flow standards: perspectives based on hydropower research and assessment. *Fisheries*, **22**(7), 21–22.

Waddle T.J. (Ed.) (2001) *PHABSIM for Windows: User's Manual and Exercises*: U.S. Geological Survey, Fort Collins, CO, p. 288. Available on the World Wide Web: **http://www.fort.usgs.gov/products/Publications/15000/preface.html** (accessed 1st April 2005).

Walters, C. (1997) Challenges in adaptive management of riparian and coastal ecosystems. *Conservation Ecology* [online], **1**(2), 1. Available on the World Wide Web: **http://www.ecologyandsociety.org/vol1/iss2/art1/** (accessed 1st November 2003).

Ward J.V., Tockner U., Uehlinger U. and Malard F. (2001) Understanding natural patterns and processes in river corridors as the basis for effective river restoration. *Regulated Rivers Research and Management*, **117**, 311–323.

Welcomme R.L. (1976) Some general and theoretical considerations on fish yields of African rivers. *Journal of Fisheries Biology*, **8**, 351–364.

Welcomme R.L. (1989) Floodplain fisheries management. In *Alternatives in Regulated River Management*, Gore J.A. and Petts G.E. (Eds.), CRC Press: Boca Raton, pp. 209–223.

Whittaker D., Shelby B., Jackson W. and Beschta R. (1993) *Instream Flows for Recreation: A Handbook on Concepts and Research Methods*, U.S. Department of the Interior, National Parks Service, Alaska Region: Anchorage, p. 104.

Wilcock P.R., Kondolf G.M., Matthews W.V. and Barta A.F. (1996) Specification of sediment maintenance flows for a large gravel-bed river. *Water Resources Research*, **32**(9), 2911–2921.

Woo S. (1999) Habitat modeling not enough to save fish... or rivers. *Stream Notes*, Stream Systems Technology Center, Rocky Mountains Research Station, USDA Forest Service: Fort Collins, pp. 5–7, April.

Young W.J., Chessman B.C., Erskine W.D., Raadik T.A., Wimbush D.J., Tilleard J., Jakeman A.J., Varley I. and Verhoeven T.J. (2004) Improving expert panel assessments through the use of a composite river condition index – the case of the rivers affected by the Snowy Mountains hydroelectric scheme, Australia. *River Research and Applications*, **20**(6), 733–750.

192: Public Participation in River Basin Planning and Management: Quality-of-Life Capital as an Information Aid to Sustainable Decisions

MALCOLM NEWSON[1] AND LIZ CHALK[2]

[1]*School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne, UK*
[2]*Environment Agency, North East Region, York, UK*

The long-term benefits conveyed by intact ecosystems to human development can be considered as an innate "natural" capital investment. A number of UK environmental management agencies have commissioned a practical realization of this concept, partly as a basis for public participation, successively titled "Environmental Capital" and "Quality of Life Capital". Hydrological concepts are particularly appropriate to an evaluation of natural goods and services; intact natural systems provide flood and pollution controls which need costly artificial replacement when lost or damaged. There are obvious opportunities, therefore, to incorporate assessments of environmental capital as part of the public participation in river basin planning demanded by the European Water Framework Directive. This contribution examines the effectiveness of "Quality of Life Capital" as an information pool and decision-support mechanism in two catchment-scale initiatives (sponsored by the European Union), aimed at community involvement in integrated land and water management. Firstly the conceptual and practical justifications for considering river basins as "multistakeholder platforms" are reviewed, followed by a comparison of the two trial applications in the north of England – to the headwater catchments of the rivers Wharfe and Coquet. What emerges from the comparison and from evaluations of other integrated river basin and coastal zone projects is the importance of information (and its management), continuity and facilitation by project management and a lasting system of monitoring to support adaptive management. The spatial scale of most river basin units defined under the Water Framework Directive is possibly the biggest barrier to widespread adoption of the approach.

INTRODUCTION

Variants of the "natural capital" theme are popular amongst those environmentalists who see an effective dialogue between ecology and economics as a major contributor to sustainable development. Practically, sustainable development is also in need of innovative delivery mechanisms for two other vital planning/management components:

- deriving strategic visions and principles that act as enduring (but flexible) *templates for specific actions* and
- achieving *stakeholder participation* in both strategies and operations.

These components are vital in practical, community-based schemes ("multiple stakeholder platforms") which have become popular for promoting best practice in integrated (land and water) management of river basin units. Despite the reservations of some authors (see below), the scale of river basin management units in most of Europe makes it feasible to contemplate basin-scale authorities, using holistic approaches (with Germany a notable exception). The EU Water Framework Directive (WFD) renders integration of ecology and economics in the difficult arena of public participation a statutory requirement. A final practical need, implied but not stated in the WFD, is for communities and professionals to move towards adaptive

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

management through monitoring ever-changing physical and social contexts.

Implementation of sustainable development has come to rely on three main elements: ecosystem understanding (including the use of indigenous knowledge), economic evaluation of environmental assets, and widespread public participation. The WFD is an example of a policy resting on these three elements (European Commission, 2000). However, despite the EU's Common Implementation Strategy for public participation there is clearly much greater urgency in national policies to implement technical, biophysical aspects in Programmes of Measures.

Semantics are important as one seeks to detect the strength of commitment in the WFD. Article 14, dealing with this matter states that public information and consultation must be "ensured", whilst active involvement is to be "encouraged".

In interpreting the Directive for UK water industry readers, Chave (2001) variously lists the requirements as:

> "*to involve the public in making decisions on water management*" (p. 11);
> "*to consult with the local population*" (p. 13).

These statements clearly ignore the heroic challenge framed by Holzwarth (2002) who states the Directive will be the basis of:

> "*catchment-based governance*" (p. 105) and
> "*getting the citizen involved*" (p. 108).

Clearly, there are considerable doubts in this piece of legislation (and in the minds of many specialists who support integrated water resource management) about the "why, when, where, and how" of public participation and what levels are required. However, debates about *levels* of participation (see e.g. Arnstein, 1969; Priscoli, 1989) have now been joined by almost frantic attempts to evaluate empirical examples to discover *processes and outcomes* (e.g. Connick and Innes, 2003) of planning and management. In this context, that is the separation of strategy and operations, or planning and management, it is also worthy of note that the WFD reserves public participation mainly for the River Basin District *planning* process, rather than seeing a role for public participation in the more technical measures.

RIVER BASINS AND THEIR STAKEHOLDERS: A CONTEXT FOR QUALITY OF LIFE CAPITAL

A frequently neglected element of debate about public participation in environmental decision making is the scale and political identity of the geographical unit or polity upon which the "multiplestakeholder platform" is to be constructed.

The "closure" of open systems to form geographical units for environmental management makes many environmental scientists nervous. However, the "fitting" of sustainable management systems onto system-defined spatial units has become a major policy dimension in the last decade. Politicians, seeking to interpret the rise of new policies and forms of governance, find the concept of bioregions (McGinnis, 1999) both convenient and attractive (e.g. Gore, 1992; Taylor, 1992). However, there remain huge problems of mapping of environmental problems on to a coherent and active community to form the basis for multistakeholder platforms.

To hydrologist readers, the "catchment", "watershed", or "basin" may be a familiar and unchallenged (except in the case of groundwater) scale for research and management but the "fit" of the responsible agencies with their geographical territories and the "interplay" between the administrative functions of integrated management cannot be taken for granted (Moss, 2003).

Rhoades (1998) is explicitly skeptical:

> "*The assumption that a precisely defined geophysical unit also serves as a socio-political or economic unit for planning and management is clearly flawed. Watersheds as closed human management units are external bureaucratic or researcher fantasies, not indigenous ones*" (p. 5).

This skepticism is important because the indigenous perception is vital to successful public participation.

However, other authors, for example Clark and Richards (2002), suggest that at least some contribution to the failure of sustainable river basin management ideals comes from fissures within and between management institutions and styles. If we reform the *management*, the management *unit* becomes effective; there are many other arguments in favor of sustainable management units structured around river basins. Large basins may not correspond to the world's "bioregions", often incorporating many such zones from source to mouth, but the coherence of the freshwater ecosystem (land *and* water), the obvious global value of that system, and a general public/professional resonance with rivers, give river basins a considerable boost as planning and management units.

In terms of employing techniques to aid public involvement such as natural capital, a number of authors have tabulated the financial value of both the goods and services provided by intact river basin ecosystems (Costanza *et al.*, 1997) and their astonishing rate of loss (see Newson, 1997, for the United Kingdom). Certain elements of natural capital only become valued because they have been lost (Daily, 1997): "*Interestingly, the nature and value of Earth's life-support systems have been illuminated primarily by their disruption and loss*" (p. 5).

Perhaps the single most successful argument in favor of river basin units as "bioregional" units is that provided by the success of the science of hydrology in the last 50 years

and its transformation into a basis for decisions concerning large-scale land use and land management (Newson, 1997; Calder, 1999). In the engineered river basin system of dams, transfers, supplies, and discharges the disciplines of law and finance can (in theory – see Winpenny, 1994) structure a sustainable use of *water* resources. However, far more than 95% of the area of most basins comprises land, not water, and can therefore experience profound changes of use and management during conventional development processes. Basin land use seldom exercises as obvious a control on water quantity and quality as for example a major dam, but at the very least has an influence on the resilience of the river ecosystem as a whole and its sensitivity to climate variability.

Table 1(a) shows the elements of the hydrological and geomorphological systems within river basins that might be considered as natural capital, based on the composite outputs of international research. Table 1(b) demonstrates, for the United Kingdom, the spectacular reduction in this capital during the last century, or rather its substitution by human capital for example intensive food production systems.

In the formation of the Environment Agency in England and Wales, river basin or watershed ("catchment" in the UK terminology) units "won" over the claims of others that, local government boundaries are the best spatial unit for particular regulatory policies for example waste control. The EU has taken the concept further with the WFD which requires nation states to submit river basin plans by 2009, based upon full ecosystem appraisal and the setting of targets based upon shortfalls between actual conditions and ecosystem "quality" (European Commission, 2000; Williams, 2001; Holzwarth, 2002).

The Directive makes implicit demands on sciences (e.g. hydrology, ecology, geomorphology, economics) which admit their uncertain predictive power, especially under conditions of rapid environmental change. Public participation does not merely fulfill a democratic ideal but helps by configuring the risks resulting from predictive uncertainties (Clark, 2002). Public participation is also timely because of:

- changes in environmental science (a stronger attempt to couch predictions in terms of risk);
- changes in governance (particularly within new bureaucratic territories such as river basins, and through popular processes of localization) and
- changes in the philosophy of public agencies towards adaptive management as a means of coping with the very long timescales inherent in the sustainability vision.

Viable techniques of implementation *via* decision support must therefore deliver: "*the alluring prospect of combining the rigor of the scientific method with the contingent realities of policy and politics*" (Dovers and Mobbs, 1997). In a sense this reveals a considerable information agenda, confirmed by what follows, and demands that the information base facilitates a combination of traditional "top–down" (technocratic) approaches with reliance on the now questioned "bottom–up" strategies (Rockstrom, 2003). Sociopolitical analysis of such partnerships is a fast-growing field in planning (e.g. Davies, 2002), but here we concentrate on the simpler, empirical evaluation of an information core to the "social learning" element of river-basin scale public participation projects.

This contribution outlines justifications and practical experiences in applying the "Environmental Capital" (now

Table 1(a) River basin system ''natural capital''

''Spontaneous regulator'' (terminology: Marchand and Toornstra, 1986)	Environmental goods and services under ''capital'' scenarios
''Natural'' river channels	Regulation of sedimentation
Undeveloped floodplains	Ditto, plus attenuation of peak flows
''Natural'' forests	Ditto, plus contributions to snowmelt
Deep soils (not eroded by bad practices)	''Natural'' reservoir and flood control
Wetlands throughout the basin	Ditto, plus water quality benefits

Based upon (from Newson, 1997).

Table 1(b) Losses of ''spontaneous regulators'' in the United Kingdom during development

Element of loss	Extent of loss
Soil moisture storage	60.9% of agricultural land drained
Ponds	75% lost
Natural channels	35 000 km of channel modified/''maintained''
Floodplain loss	$>3000\,km^2$
Transformation to urban	1/3 of all recent land-use change

Based upon (from Newson, 1997).

"Quality of Life Capital": **www.qualityoflifecapital.org**) approach to two schemes of integrated land and water management in a general framework of restoring ecological quality (a WFD objective) to headwater catchments. We use a comparison of the relative success and failure of Quality-of-Life Capital (QoLC) in the two basins to derive general guidance for effective stakeholder participation.

"CAPITAL" ASSESSMENTS OF ECOSYSTEM ATTRIBUTES

It has frequently been argued that, whilst moral persuasion, regulation, and economic instruments are the major conventional tools available to national governments to achieve sustainable development through environmental protection, an holistic framework for both strategic planning and operations might best be provided by the concept of nature (and its anthropogenic transformations) as "capital" (Table 2(a)). This approach is much wider than the resource economics paradigm that dominated our concept of environmental limits for the second half of last century, driven partly by the strictures of classical economics. The major focus has now shifted from simple resource conservation to limits posed by environmental degradation and loss of biodiversity. Natural processes are seen as counteracting these anthropogenic impacts and are delivered to human societies as "services" by intact ecosystems. As Daily (1997) puts it:

> "*In addition to the production of goods, ecosystem services are the actual life-support functions, such as cleansing, recycling and renewal and they confer many intangible aesthetic and cultural benefits as well*" (p. 3).

Clearly, the concept has much resonance with definitions of sustainable resource use; Costanza *et al.*, (1991) have stated "*a useful definition is the amount of consumption that can be continued indefinitely without degrading capital stocks – including 'natural capital stocks'. This natural capital stock uses primary inputs (sunlight) to produce the range of ecosystem services and physical natural resource flows*" (p. 8).

Better management of complex ecosystems inevitably demands understanding at the system scale as well as requiring a myriad of local details. An inconvenient paradox of applying "systems thinking" to environmental management is the complexity inherent in our understanding of the system being managed. Whilst fundamental to the approach (Clayton and Radcliffe, 1997), its outputs require translation into "useable knowledge" by the assignment of values and risks to the management options it offers and the selection of priorities, before practical applications are possible. With this in mind, the concept of natural capital (later operationalized as Environmental Capital and finally QoLC in the United Kingdom) was introduced as a bridge between scientific assessments of environmental problems and the emerging body of economic techniques to aid decision making.

The basic analogue is between environmental assets and the economists' notion of capital; assets are deemed to be well used if they continue to supply "interest" income. Natural capital is not substituted into other forms of, for example, man-made capital without debate on the long-term significance of this trade-off; in this way the polity decides the relative "strength" of the sustainability model it adopts (Turner, 1993 and Table 2(b)). The central position of trade-offs in the natural capital argument makes stakeholder participation vital (Clark, 2002).

The vital question, however, remains that of whether such conceptually elegant procedures can be translated into simple, robust, and accessible public procedures, without preempting or replacing decision-making procedures. In 1997, four public agencies in England and Wales funded the development of a survey and appraisal technique known as

Table 2(a) "Natural Capital" and other relevant forms

Form of capital	Definition
Natural	Natural resources, living systems, ecosystem services. Enlarged by Pearce and Barbier as: • renewable/nonrenewable resources; • assimilation of waste; • amenity; • biogeochemical cycles – stabilization; • genetic blueprints, behavior.
Human	Labor, intelligence, culture, and organization. Includes social capital: • Membership of groups; • Relationships of trust; • Access to institutions.
Manufactured (or physical)	Infrastructure, machines, tools, factories, vehicles.
Financial	Cash, investments, economic instruments

Based upon (after Pearce and Barbier, 2000; Hawken *et al.*, 2000).

Table 2(b) Substitution of capital: defining the ''strength'' of sustainability (after Pearce and Barbier, 2000; Clayton and Radcliffe, 1997)

''Strength'', criticality	Substitution of capital between forms
Weak sustainability	Essentially no inherent difference is assumed between natural capital and other forms and substitution occurs if it moves the capital to more valuable human/physical forms
Strong sustainability	Environmental resources and ecological services that are essential for human welfare and cannot easily be substituted should be protected
Critical natural capital	Applied to attributes that are irreplaceable or irreparable and currently scarce; also applied to those attributes whose loss would promote further significant loss.

Based upon (after Pearce and Barbier, 2000; Clayton and Radcliffe, 1997).

"*Environmental Capital*". It was avowedly scientific, based on natural systems guidance (Clayton and Radcliffe, 1997) for sustainable development (CAG/LUC, 1997). However, the change of name from "Natural Capital" to "Environmental Capital" was based upon the inclusion, within the proposed survey technique, of all elements of environmental capital (Table 2), as well as natural capital. In this way, the ecosystem is seen as inclusive of anthropogenic processes rather than a separate pristine ideal. The technique set out to emphasize *functions*, as well as features, of a given geographical unit and the main themes of sustainable development were promoted by the pattern of the environmental capital enquiry: from attributes to importance and thence to trends/sufficiency and substitution or trade-offs (between forms and features of environmental capital).

A subsequent change to Quality-of-Life Capital or QoLC sought to widen its public appeal. An Overview Report (CAG/LUC, 2001a,b) suggests, *inter alia* that:

> "*Quality of life can be understood in terms of having a wide range of human needs satisfied*".
> "*A decision support tool that considered all these stocks or capitals in an explicit, transparent, coherent and integrated way might reasonably be hoped to help us reach better decisions than the current ramshackle, piecemeal, inconsistent and often disguised treatment of environmental, social and economic issues in decision and planning processes*" (p. 14).

The CAG/LUC report came too early to be matched to a growing orthodoxy of "multistakeholder processes" (Hemmati, 2002) and gives much of the responsibility for steering the technique to officers of the participating agencies. Yet, the ability to produce a shared understanding of environmental (river basin) needs by multiple stakeholder communities is a strength of QoLC; it is able to provide a cohesive look at catchment attributes, thus overcoming the current fragmented approach to management which does not mimic complex catchment-scale system interactions. However, the approach also has aspirations to become the basis of public participation (in e.g. Environmental Impact Assessment, planning, and within sectorial policy development):

> "*Professional judgement should be supported by assessment of public perceptions of the characteristics of the environment*" (CAG/LUC, 1997; p. 22).

The inclusion of the public voice at every stage of environmental decision-making has become a hall-mark of the ecological restoration movement, but now has powerful support within integrated water resource management as central to resource development projects, river/wetland restoration projects and projects to develop and encourage best practice in land/water management. Given proper attention to issues of subsidiarity and translation into a commonplace language, QoLC has the potential to deliver these ideals.

INTRODUCTION TO UK CASE STUDIES: UPPER WHARFEDALE AND UPPER COQUETDALE

It is currently useful to evaluate the few available reports of public participation in river basin management to elucidate any general principles lying behind the site-specific problem orientation and unique sociopolitical contexts of each case study; generally, failures of consultation and participation are not reported. We here report from two integrated (land and water) management projects in the north of England. In one QoLC, techniques were used as an information platform with apparent success, and in the other, the approach largely failed (but not because of technical weaknesses – see following text).

Common elements of the Upper Wharfedale and Upper Coquetdale projects were their EU funding *via* the "Objective 5b" rural development (Agricultural Adjustment) route (hence their socioeconomic core purpose) and their integrated land/water approach to restoring some of the environmental goods and services provided by in tact catchments. Another common thread was the prominent involvement of the Environment Agency, the statutory body for river management in England and Wales. However, the manner of that involvement, and finer details of project orientation and management, differed between the two projects, providing the basis of an experimental comparison.

An essential question surrounding the ecosystems approach to management is that of the costs surrounding data collection (Clark, 2002). A holistic survey system such

as QoLC is particularly demanding in this respect; environmental management professionals have gained a rather negative impression of survey costs from the scoping of Environmental Impact Assessments. In fact, our experience in both Upper Wharfedale and Coquetdale is that much of the essential information is already available in a developed country and therefore the basic requirement is for collation, within a new conceptual framework, rather than primary survey, even though the technique reveals some survey gaps. In addition, a major value of stakeholder involvement derives from the "indigenous knowledge", or "vernacular science" (Wynne, 1992) which becomes a genuinely interactive element of the social learning experience between experts and the community (Rolling, 2000).

The QoLC framework offers considerable opportunities to access the experiential knowledge of local people; however, this exposes managers to the need to reconcile popular impression with scientific "fact", for example concerning the seriousness of river erosion. This may ultimately enhance stakeholder understanding of catchment processes leading to a higher potential for adaptive community decisions in the longer term.

The Upper Wharfedale "Best Practice" Project The Upper Wharfedale Best Practice Project (UWBPP) was carried out between 1998 and 2002 in an upland river catchment area within the Yorkshire Dales National Park, with the highest point approximately 670 m above sea level. It is 100 km^2 in area and has a population of 5600, heavily inflated by visitor pressure in summer. The initial positive response from local organizations and the community to the idea of project on water and land management, was stimulated by brainstorming sessions of both interested professionals and local stakeholders, the latter at a public community meeting.

An initial Feasibility Study which accessed local opinion at its core stages (RKL-ARUP, 1998) listed the "main programme elements" which might feature in the Project as:

- river bank erosion;
- regeneration of "gill" (headwater valley-side stream) habitats;
- flood defence strategy;
- pollution control;
- biodiversity;
- moorland drainage;
- livestock farming practice;
- education.

The consultants also suggested delivery mechanisms for a "Best Practice" project in Upper Wharfedale:

- agricultural reforms and demonstrations;
- forestry restructuring
- fluvial management as a holistic concept;
- ecological improvements;
- recreational management;
- archaeological and landscape conservation;
- socioeconomic support.

The Project was developed and implemented through a partnership, led by the Environment Agency, and managed by a Steering Group. Day to day management on the ground was coordinated by a Project Officer with overall management coordinated by a Project Manager. Some issues were progressed using subgroups with themes such as sustainability, education, and habitats. Workshops, public meetings in the Parish Hall, individual meetings at the farmhouses, walk events at locations in the Dale and the production by local artists and the community of a floor-scale river game (We Have A River For Everyone) were some of the other methods used to disseminate information and seek public input to the core Environmental Capital survey, to which we now turn. Whilst exact, linear management principles are often inappropriate for participatory approaches, some choreography of the Environmental Capital approach is vital for all concerned (Table 3); the working "capital" features identified by professionals and the public in Upper Wharfedale are identified in Table 4 and Figure 1.

A series of Project leaflets have formed a principal information outlet; they cover sheep dip treatment, river erosion and the uplands, moorland gripping, river gravel management, the Environmental Capital Approach (its title at that time) and sustainable river management techniques (Environment Agency, 2003).

The River Coquet drains the southern flanks of the Cheviot Hills in Northumberland, UK; its upper part, known as *Upper Coquetdale*, has an area of 358 km^2, a population of 3200 and is similar in relief, climate, and land-use to Upper Wharfedale. The river has the status of an SSSI, part of the area is in the Northumbrian National Park and a number of agri-environment "Best Practice" policies have been executed in the catchment. The application of QoLC surveys did not, however, commence until the

Table 3 Steps involving public participation, worked around QoLC assessments in the Upper Wharfedale Best Practice Project

Step	Aims/objectives
1	Establish study area
2	Determine features
3	Understand features, their functions
4	Identify benefits and evaluate features
5	Produce management principles
6	Produce short-term objectives
7	Implement and monitor success of short-term objectives
8	Fill long-term information gaps
9	Produce long-term objectives
10	Implement and monitor success of long-term objectives

Table 4 Upper Wharfedale – evaluation of quality of life capital elements within the Upper Wharfe catchment

Feature	Attributes/ services	Why it matters	Scale which matters	Substitutability	Impor- tance	Trend	Management/ monitoring
Blanket peat **moorland**	Stores carbon from atmospheric pollution, stores water, key habitat	Environmental control (stores water, carbon) – vital to all	Global/catchment/ Wharfedale valley floor	Impossible to substitute: peat growth rates very slow	High	Eroding and thinning due to grazing/ gripping	Block drains, ''assisted recovery'', low intensity use. Monitor hydrological recovery, monitor access, grouse, cover/stocking etc.
Native **woodland**	Absorbs carbon from atmospheric pollution, ground stabilizer, habitat	Amenity, nature conservation, local economy	Local and regional/ catchment	Impossible to substitute, existing woodland can be managed/ extended	Medium/ high	Undermanaged – area below critical threshold	Proof against stock, regenerate, extend in upper catchment and riparian zone
Plantation **woodland**	Economic services, absorbs carbon from polluted air	Local economy and Conservation potential	Local/catchment	Restore where past practices unsustainable	Medium	Mitigation (e.g. habitat, runoff) welcome	Forestry Commission ''guidelines'' retroactive for existing plantations. Stream sampling and wildlife monitoring
Limestone features	Pavements, caves scientifically import, attract visitors	Part of ''spirit of place'' of the Dales, biodiversity/ tourism	Local (river regime, tourism, water supply) national, biodiversity	Impossible to recreate. Can be lost through quarrying etc	High	Pavement needs protection, caves: access, pollution protection	Related to peat moorland, but also to recreation management/pressure and farm pollution (cave ecosystems)
River channel and gills	Oxygenation, sediment storage, aesthetic attraction, angling, biota	Central to the catchment and to the identity of place. Local water supplies	Local, regional/ catchment	Enhancement of past flood works (e.g. gravel trap) to assist recovery	High	Works in the past created erosion and deposition problems	Research on past and present sediment and flow regime of the channel followed by restoration including flood protection

(continued overleaf)

Table 4 *(continued)*

Feature	Attributes/ services	Why it matters	Scale which matters	Substitutability	Impor- tance	Trend	Management/ monitoring
River floodplain: Hay meadows, wetland pastures, flood deposited material	Economic core, visitor focus, flood storage, wildlife habitat	Key element of farming economy but some services are to communities downstream	Local and regional/ catchment	We may make publicly funded substitutes (e.g. agri-environment incentives)	High	Flood defence and drainage schemes led to damage. Aims now are natural rivers	Hydraulic modeling and geomorphological/ ecological studies to plan sustainable floodplain environment – e.g. flood areas v improved areas
Farmland in general excluding moorland and floodplain (see below)	Basic natural resource for human occupancy, fabric of landscape	Provides food for farm stock and a habitat for wildlife	Local, national and international. Essential to livestock economy, landscape	Potentially improvable but intensification may be unsustainable or damaging	High	May degrade if farming income falls, best practice conservation strategy essential	Manage according to stocking density/ fertilizer in relation to ecology but in tune with farm finance. Monitor cover/ stocking
Built environment: villages, farmsteads, field barns, walls	Homes, fabric of the landscape, habitat, economic core. Major aesthetic appeal	Provides cultural, aesthetic and economic services to humans	Local, national and international	None, except *via* new livelihoods for locals – e.g. high technology cottage industries	Critical	House prices, loss of services, visitor pressure may threaten farm economy in crisis	Basic economic structure must be retained but perhaps subtly modified in response to changing agricultural activity/ roles
Routes: roads, lanes, tracks, paths	Add high value to landscape, limits and diversifies human exploitation	Cultural, aesthetic and economic services, set ''carrying capacity'' for dale	Local, national: access to care, accessibility to delivery vehicles and to scenery	Techno, substitutes for mobility make no impact yet in this economy and culture	Critical	Maintenance essential to sustainability at all scales	Conditions and status of infrastructure must be recognized, pressures must be monitored on a regular basis
Archaeology Prehistoric to industrial & agricultural	Context to human occupation, attractive to visitors	Provides cultural setting and sense of renewal, educational values	Local, national. Greater in the context of landscape	None	High	Limit visitor pressure, protect against vandalism, make accessible	Full survey, designation, sustainable exploitation, restoration where appropriate

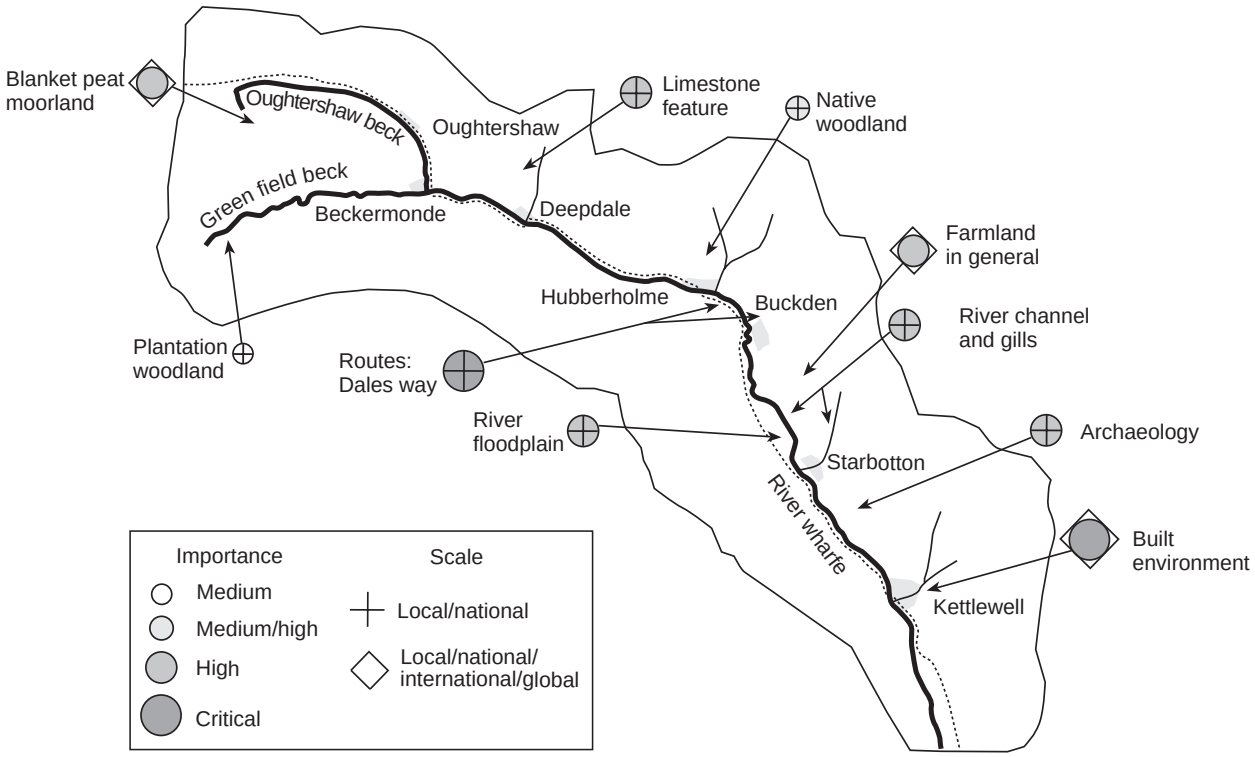


Figure 1 Upper Wharfedale – examples of QoLC mapped in the catchment

Northumbrian Rivers Project, also an EU "Objective 5b" scheme, chose the Coquet as one of its target catchments in 1997.

The Northumbrian Rivers Project's main targets were to improve riparian land management in order to conserve and extend the fishery interest, thus supporting both the farming community through grant aid and attracting important income from tourism. The Environment Agency, concerned that riparian initiatives did not conflict with their Flood Defence function, commissioned Newcastle University to apply QoLC to the river and its geomorphology (Newson, 1999); community issues and wider public participation did not occur until later (Fox, 2002). Fox gained access to a wide sample of Upper Coquetdale residents during a very detailed survey of QoLC attributes, revealing a number of tensions in the relationship between environmental managers and the assessment of environmental values based on indigenous perceptions. Broadly, local people viewed the "natural" environment of the valley as heavily anthropogenic and placed meaning and identity in relation to the degree to which the landscape is "worked". They did not lack insights into natural capital and some of the goods and services provided by intact ecosystems but they generally refused to participate in financial or trade-off exercises. There was a dominant popular impression that environmental management should be a means of representing indigenous pride in their "worked" landscape to institutions outside (to increase the financial support for its infrastructure), rather than as a delivery mechanism for national and transnational policies on sustainability, that is bringing the "outside" in.

LESSONS FROM EVALUATION: A COMPARATIVE ANALYSIS OF ''CAPITAL'' AS A PARTICIPATORY DEVICE

Whilst the two projects described briefly above (for more detail see Newson and Chalk, 2004) were not selected to provide an analytical comparison, they exhibit contrasting characteristics worthy of consideration in reforming and adapting the use of QoLC in the much broader application now sought for public consultation in river management.

However, before making the analysis, it is possible to gain further insights on other exercises in public participation from the small sample of published evaluations (e.g. Edwards-Jones, 1997a,b; Hillman *et al.*, 2003; Connick and Innes, 2003; King, 2003 – the latter example being from Integrated Coastal Zone Management).

Almost unanimously these evaluations assert the central role of decisions based upon a broad but systematic body of information whose compilation and shared value becomes part of the "social capital" of the participatory venture (Hillman *et al.*, 2003, p. 221).

King (2003) suggests that:

> "*An information strategy is vital for participation and the two aspects should be conceived together*" (p. 142).

Another common thread is to put emphasis on quality of process rather than on mechanistic measures of outcome; "collaborative practices are more fluid and less predictable than traditional forms of policy making" (Connick and Innes, 2003, p. 178). Perhaps this realization faced Professor Kader Asmal in his role as Chair of the World Commission on Dams (Asmal, 2000):

> "*the process becomes a messy, loose-knit, exasperating, sprawling cacophony. Like pluralist democracy, it is the absolute worst form of consensus-building except for all the others*."

This quotation underlines the complexity arising from flexibility but emphasizes a core need for structure! Information provides a neutral structure for participation, but this neutrality is, paradoxically, best delivered by firm, very energetic, project management (Edwards-Jones, 2003).

Thus, it is vital for the following comparison between our experiences in Upper Wharfedale and Upper Coquetdale to consider QoLC as a contender for structuring the vital information core of a participation exercise, but with the contexts for its success set by attitudes and social processes. Table 5 draws up a list of contrasts that have impinged upon the delivery of stakeholder participation through QoLC in the two small UK catchments. Virtually none relates to the technical proficiency of the QoLC technique; the openings into community views potentially provided by QoLC are undoubted, but the method's "success" is likely to be judged as that of the project itself.

Clearly, the prime contributors to success or failure were the spatial and temporal focus (and clarity!) of the project objectives, activity rates in both consultation and operations, coordination and facilitation by clearly identifiable leaders use of knowledge and information. There is considerable resonance with a list of criteria for collaborative dialogue presented by Connick and Innes (2003):

- includes representatives of all relevant interests;
- is driven by a practical purpose and task shared in the group;
- is self-organizing;

Table 5 Comparative elements of the UWBPP and the Upper Coquetdale applications of QoLC assessment techniques

Element	UWBPP	Upper Coquetdale
Spatial focus	Catchment wide, with prominent moorland and valley-side as well as riparian activities. Sole focus for financial aid.	Channel obstructions, bank erosion and riparian support to river ecosystem. Finance shared with three other rivers.
Temporal focus	Five-year Project lifetime followed by landowner responsibility for maintenance but with monitoring continuing.	Five-year Project lifetime followed by landowner responsibilities for maintenance.
Public consultation – timing and context	Began at feasibility stage and integral throughout. Project set in a geographical context of heavy visitor pressure and a plethora of institutional consultations.	Expert-led until quite late in Project lifetime. Experience of environmental consultation developed only by National Park and Catchment Management Plans.
Coordination and project management	''Expert'' and voluntary members of Steering Group in close coordination with community representatives. Clear responsibility of Project staff (from Environment Agency); no staff changes.	Identical ''expert'' (institutional) bodies represented on Steering Group but closest voluntary linkage to fishing community. Project officers from FFWAG. Several staff changes.
Operations	EA able to facilitate aspects of river management (machinery) and monitoring (e.g. hydrology), although majority undertaken by local contractors, landowners, tenants, and angling clubs.	Landowner responsibility – no Project labor force.
Knowledge and information management	Indigenous knowledge sought from feasibility, collated with ''expert'' knowledge in QoLC attributes and valuation. Dissemination at all levels including schools and events linked to best practice leaflets.	Catchment Management Plan and knowledge of anglers used to guide priorities (largely also fixed by *modus operandi* of Project). Public information events and river geomorphology training courses. Project followed, rather than led, best practice.
Dissemination and continuity postproject	Trials of best practice a specific aim – hence dissemination by variety of media: education and training. Monitoring and research tasks set for longer term.	Relatively little innovation: survey directed to problem sites, funding guided by opportunity. Publicity, not publications a project aim since heavy links to tourism economy (fishing).

- is encouraging to participants as they learn and interact;
- encourages challenges to assumptions and fosters creativity;
- incorporates many kinds of high-quality information;
- seeks consensus only after discussions have fully explored issues and interests and significant effort has been made to find creative responses to difference.

WIDER ISSUES: MAINTAINING/UPDATING THE "STAKEHOLDER PLATFORM" AND APPLYING ECONOMIC VALUATION

One of the major costs attending truly sustainable development is monitoring, demanded both by the uncertainties of the knowledge base applied to current management operations and by the needs of adaptive management systems to be aware of environmental changes. Following exercises in public consultation of the type described here, it is tempting for the appropriate agencies to return the task of monitoring to professional cadres in the sole interest of addressing legislative necessities. The UWBPP intends to leave the community in Upper Wharfedale with both a role and an interest in monitoring the "best practices" established during the life of the project. The example of the Australian "Catchment Health" project (Walker and Reuter, 1996) has been copied informally for future activity by for example schools, voluntary groups, and farming families; women and children have played particularly significant roles in adoption.

Additionally, the Project has identified longer-term research tasks for the academic community and these have been adopted and funded (McDonald *et al.*, 2003). Longer-term strategic options identified included the radical option of reconnecting the river with its floodplain. Current information was inadequate to achieve an integrated understanding of how the land is linked to the river and answer questions like how to manage gravels in the river and what is the impact downstream. The impacts of wider, less certain influences, such as climate change, also need to be factored in. Clearly, however, at significant points in the collection of new and relevant information it must be disseminated to allow the reconstruction of the basic QoLC matrices.

If longevity and community-based schemes of monitoring are an indicator of success in participatory schemes of river basin management, the QoLC attracts a further test – that of whether it is feasible to convert stakeholder values for environmental attributes from an ordinal scale to an interval, financial one. The apparent unwillingness of those surveyed in Upper Coquetdale to consider formal valuation of their QoLC matrix is a set-back, but it may be that in projects where there is a genuine chance of economic changes (gain) as variation to national or regional practice local people may be persuaded by the seriousness of the possible outcomes for their own interests.

A technique that helps to create multistakeholder platforms by incorporating the concept of capital should be capable of interfacing with the world of classical economics or at least with existing local manipulations of the macroeconomic framework (e.g. agricultural subsidies under the Common Agricultural Policy (CAP)). The early fiscal manifestations of "environmental economics", for example *via* the use of specific economic instruments to promote environmental values, would benefit from such a linkage. It would clearly be a considerable additional burden of survey to the existing QoLC techniques to undertake separate analysis of the multidimensional "values" of the "natural" and anthropogenic assets of a river basin. Techniques abound for such surveys and, indeed, have been applied to the Upper Wharfedale area (Yorkshire Dales) in exercises predating the UWBPP (Garrod and Willis, 1999).

However, these conventional valuation techniques do not directly embrace (because public questionnaire survey alone is inadequate as a basis) the concept of environmental services. Here QoLC should be capable, *via* its natural systems analytical approach, of introducing "hard" economic values by "parallel accounting". For example, if the volume of water passing through a specific floodplain wetland can be estimated and the flood protection or water quality benefits of this passage can be inferred from the hydrological literature, then the costs foregone to the local or regional or global economy from the environmental services provided by the wetland can be calculated (e.g. Bockstael *et al.*, 1997). Given a wider uptake for QoLC and similar techniques, the application of valuation techniques should now be attempted: "*All we need to do now is to apply accounting to environmental capital . . .attempting to bring measurable elements into the process as our knowledge improves. But to wait until everything falls properly into place will mean that we shall have to wait for ever.*" (El Sefary, 1991).

CONCLUSIONS

The QoLC methodology addresses many of the practical problems raised by those with empirical experience of integrated river basin planning and management (e.g. Edwards-Jones, 1997a,b, 2003) who converge on information, and its management through to decisions, as the core of successful projects. Whilst it is currently a less formal manifestation of a decision-support system for river management than the model SURCoMES, also calibrated to run in Upper Wharfedale (Clark and Richards, 2002), future opportunities afforded by minor reforms of the basic technique and the addition of improved communication technologies (e.g. Kingston *et al.*, 2000) may give it a prominent position as a foundation for multistakeholder platforms. Rule-based models with "fuzzy" decision-making circuits, such as SURCoMES, will inevitably play a part in future adaptive management scenarios. As Clark (2002) puts it:

"*decision support itself strengthens sustainability, which may be unachievable without it, and enables informed and aware stakeholder participation*" (Clark, 2002, p. 349).

The EU Water Framework Directive makes heavy demands for, but is currently in a learning phase for, both stakeholder participation and economic evaluation. QoLC techniques are certainly a potential for delivering a framework for both these innovative contributions to sustainable development. They act as "formal ways of integrating stakeholders and science" (Osidele *et al.*, 2000) even if they do not create multistakeholder platforms, defined by Rolling and Maarleveld (1999) as "devices or procedures for social learning and negotiation about effective collective action". To the authors, it is doubtful whether multistakeholder processes can be universally "designed" (as by Hemmati, 2002, who gives 15 design principles) but the characteristics of the two Projects used as case studies here, offer signposts about the general characteristics of "collective action to reduce a collective impact" (Rolling, 2000).

Most importantly Environmental (QoLC) Capital has been a successful way to involve a wide range of interests and provoke debate between specialists and people in the community and thus provide a catalyst for some visionary longer-term thinking and initiatives to further understanding. It has placed professional judgement alongside the concerns of local people, helping everybody to be aware of and own change.

Perhaps the most daunting barrier to the widespread use of information "hubs" to participatory approaches under the EU WFD, is the large spatial scale of the river basin districts to be planned. A possible remedy already adopted in Scotland is subdivision of the basin districts for the specific purpose of stakeholder analysis followed by participation.

Acknowledgments

Clearly, in studies of this type those who contributed are too many to name, in both professional and stakeholder camps. However, our colleagues Ivan Ingles (Environment Agency: Upper Wharfedale Project) and Helen Fox (Upper Coquetdale surveys) made prolonged contributions. We would also like to thank Ann Rooke for drawing Figure 1.

FURTHER READING

Buckingham-Hatfield S. and Percy S. (Eds.) (1999) *Constructing Local Environmental Agendas. People, Places + Participation*, Routledge: London.

Chambers N., Simmons C. and Wackernagel M. (2000) *Sharing Nature's Interest. Ecological Footprints as an Indicator of Sustainability*, Earthscan: London.

Costanza R. (Ed.) (1991) *Ecological Economics: the Science and Management of Sustainability*, Columbia University Press.

Foster C.H.W. (1984) *Experiments in Bioregionalism. The New England River Basins Story*, University Press of New England: Hanover and London.

Mason M. (1999) *Environmental Democracy*, Earthscan: London.

Warburton D. (1998) *Community and Sustainable Development. Participation in the Future*, Earthscan: London.

REFERENCES

Arnstein S. (1969) Continuum of involvement. *Journal of the American Institution of Planners*, **35**, 224.

Asmal K. (2000) *First World Chaos, Third World Calm: a Multi-Stakeholder Process to Test the Water in the Debate Over Dams*, Le Monde, 15th November.

Bockstael N., Costanza R., Strans I., Boynton W., Bell K. and Wanger L. (1997) Ecological economic modelling and the value of ecosystems. *Ecological Economics*, **14**, 143–159.

CAG/LUC (1997) *What Matters and Why. Environmental Capital: A New Approach. A Provisional Guide*, CAG/LUC: London.

CAG/LUC (2001a) Overview report. *Quality of Life Capital Managing Environmental, Social and Economic Benefits*, CAG/LUC: London.

CAG/LUC (2001b) *Quality of Life Capital Managing Environmental, Social and Economic Benefits*, `http://www.qualityoflifecapital.org.uk`.

Calder I.R. (1999) *The Blue Revolution: Land Use and Integrated Water Resources Management*, Earthscan: London.

Chave P. (2001) *The EU Water Framework Directive. An Introduction*, IWA Publishing: London.

Clark M.J. (2002) Dealing with uncertainty: adaptive approaches to sustainable river management. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **12**, 347–363.

Clark M.J. and Richards K.J. (2002) Supporting complex decisions for sustainable river management in England and Wales. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **12**, 471–483.

Clayton A.M. and Radcliffe N.J. (1997) *Sustainability: A Systems Approach*, Earthscan: London.

Connick S. and Innes J.E. (2003) Outcomes of collaborative water policy making: applying complexity thinking to evaluation. *Journal of Environmental Planning and Management*, **46**(2), 177–197.

Costanza R., Daly H.E. and Bartholomew J.A. (1991) Goals, agenda and policy recommendations for ecological economics. In *Ecological Economics: The Science and Management of Sustainability*, Costanza R. (Ed.), Columbia University Press: pp. 1–20.

Costanza R., D'arge R., de Grrot R., Faber S., Grasso M., Hannon B., Limburg K., Naeem S., O'Neill R.V. and Parvelo J. (1997) The value of the world's ecosystem services and natural capital. *Nature*, **387**, 253–260.

Daily G.C. (Ed.) (1997) *Nature's Services. Societal Dependence on Natural Ecosystems*, Island Press: Washington.

Davies A.R. (2002) Power, politics and networks: shaping partnerships for sustainable communities. *Area*, **34.2**, 190–203.

Dovers S.R. and Mobbs C.D. (1997) An alluring prospect? Ecology and the requirements of adaptive management. In

Frontiers in Ecology: Building the Links, Klomp N. and Lunt I. (Eds.), Elsevier Science: Oxford, pp. 39–52.

Edwards-Jones E.S. (1997a) The river valleys project – a participatory approach to integrated catchment planning in Scotland. *Journal of Environmental Planning and Management*, **40**(1), 125–141.

Edwards-Jones E.S. (1997b) The river valleys project: using integrated catchment planning to improve and enhance two Scottish rivers. In *Freshwater Quality: Defining the Indefinable*, Boon P.J. and Howell D.L. (Eds.), HMSO: Edinburgh, pp. 506–512.

Edwards-Jones E.S. (2003) *The Development of a Generic Methodology for Integrated Catchment Planning in the UK*, Unpublished Ph.D. Thesis, Herriot-Watt University.

El Sefary S. (1991) The environment as capital. In *Ecological Economics: the Science and Management of Sustainability*, Costanza R. (Ed.), Columbia University Press: pp. 168–175.

Environment Agency (2003) *Upper Wharfedale Best Practice Project 1998–2002, Information Series – Folder of Six Editions*, Environment Agency: New York.

European Commission (2000) *Directive 2000/60/EC, Establishing a Framework for Community Action in the Field of Water Policy*. European Commission, PE-CONS 3639/1/100, Luxemborg.

Fox H. (2002) *The People of Coquetdale, Northumberland, UK: Their Perceptions of the 'Environmental Capital' of the Dale*, School of Geography, Politics and Sociology, University of Newcastle upon Tyne.

Garrod G. and Willis K.G. (1999) *Economic Evaluation of the Environment. Methods and Case Studies*, Edward Elgar: Cheltenham.

Gore A. (1992) *Earth in the Balance: Forging a New Common Purpose*, Earthscan: London.

Hawken P., Lovins A.B. and Lovins L.H. (2000) *Natural Capitalism. The Next Industrial Revolution*, Earthscan: London.

Hemmati M. (2002) *Multi-Stakeholder Processes for Governance and Sustainability*, Earthscan: London.

Hillman M., Aplin G. and Brierley G. (2003) The importance of process in ecosystem management: lessons from the Lachlan catchment, New South Wales, Australia. *Journal of Environmental Planning and Management*, **46**(2), 219–237.

Holzwarth F. (2002) The EU water framework directive – a key to catchment-based governance. *Water Science and Technology*, **45**(8), 105–112.

King G. (2003) The role of participation in the European demonstration projects in ICZM. *Coastal Management*, **31**, 137–143.

Kingston R., Carver S., Evans A. and Turton I. (2000) Web-based public participation geographical information systems: an aid to local environmental decision-making. *Computers, Environmental and Urban Systems*, **24**, 109–125.

Marchand M. and Toornstra F.H. (1986) *Ecological Guidelines for River Basin Development*, Report 28, Centrum voor Milieukunde, Rijksuniversiteit: Leiden.

McDonald A., Lane S., Kirkby M., Holden J., Ashley D., Reid S., Tafeyi V. and Brookes C., (2003) *Information Requirements for the Integrated Management of Agricultural Areas in Sensitive River Basins*, R&D Technical Report E1-108/TR, Environment Agency, Bristol.

McGinnis M.V. (Ed.) (1999) *Bioregionalism*, Routledge: London.

Moss T. (2003) Resolving problems of spatial fit and institutional interplay with river basin management: an institutionalist appraisal of the EU WFD. *Royal Geographical Society – Institute of British Geographers Annual Conference*, 3rd–5th September. (Abstracts CD). RGS-IBG: London.

Newson M.D. (1997) *Land, Water and Development, Second Edition*, Routledge: London.

Newson M.D. (1999) *Environmental Capital: a Strategic and Operational Guide to the Management of River SSSIs*, Department of Geography, University of Newcastle upon Tyne.

Newson M.D. and Chalk L. (2004) Environmental capital: an information hub for public participation in river basin 'best practice' projects. *Journal of Environmental Planning and Management* **37**(6), 899–920.

Osidele O.O., Beck M.B. and Fath B.D. (2000) A case study in integrating stakeholder concerns with the water sciences. *7th National Hydrology Symposium, Newcastle*, Institution of Civil Engineers: London, pp. 1.53–1.59.

Pearce D. and Barbier E.B. (2000) *Blueprint for a Sustainable Economy*, Earthscan: London.

Priscoli J.D. (1989) Public involvement, conflict management: means to environmental quality and social objectives. *Journal of Water Resource Planning and Management, Proceedings of the American Society of Civil Engineers*, **115**(1), 31–42.

Rhoades R.E. (1998) *Participatory Watershed Research and Management: Where the Shadow Falls*. Gatekeeper Series number 81, International Institute for Environment and Development, London.

RKL-ARUP (1998) *Upper Wharfedale 'Best Practice' Project – Feasibility Study*, Leeds.

Rockstrom J. (2003) Managing rain for the future. In *Rethinking Water Management. Innovative Approaches to Contemporary Issues*, Figueres M., Tortajada C. and Rockstrom J. (Eds.), Earthscan: London, pp. 70–101.

Rolling N. (2000) Gateway to the global garden. *8th Annual Hopper Lecture*, University of Guelph.

Rolling N. and Maarleveld M. (1999) Facing strategic narratives: an argument for interactive effectiveness. *Agriculture and Human Values*, **16**, 295–308.

Taylor A. (1992) *Choosing Our Future. A Practical Politics of the Environment*, Routledge: London.

Turner R.K. (1993) *Sustainable Environmental Economics and Management*, Bellhaven Press: London.

Walker J., Reuter D.J. (1996) *Indicators of Catchment Health: a Technical Perspective*. CSIRO Publishing: Collingwood, Victoria, Autralia.

Williams K. (2001) The impact of the water framework directive on catchment management planning in the British Isles. *Water and Environmental Management*, **15**(2), 97–102.

Winpenny J.T. (1994) *Managing Water as an Economic Resource*, Routledge: London.

Wynne B. (1992) Uncertainty and environmental learning. Reconceiving science and policy in the preventive paradigm. *Global Environmental Change*, **2**, 111–127.

193: Markets for Watershed Services

SYLVIA S TOGNETTI[1], BRUCE AYLWARD[2] AND GUILLERMO F MENDOZA[3]

[1] *Consultant, Environmental Science and Policy*
[2] *Deschutes Water Exchange Program, Deschutes Resources Conservancy, Bend, OR, US*
[3] *New York Department of Environmental Protection, Flushing, NY, US*

Increasing degradation of watersheds has led to increased recognition of the services they provide in various forms of support for livelihoods and general well-being, as well as to a greater willingness to pay for them and to cooperate in initiatives to protect them. This is reflected in numerous initiatives, in which market-based instruments and other supporting institutional arrangements are used as a way to create incentives and to recover the costs of watershed protection, as well as to allocate water more efficiently among various uses. Many of these payment initiatives have focused narrowly on the role of forests in the hydrological regime as a way to justify funding for their conservation, but should be developed in the context of basin-wide management objectives, which provide a framework for considering the full range of interests that share a common river basin in the context of specific ecosystem functions that support them, and for identifying and quantifying trade-offs associated with various management options. In addition to forests, this would include consideration of the relative values of all types of land cover and land uses such as wetlands, riparian areas, steep slopes, roads, and management practices. It also requires an accounting for the role of human consumption in the modification of the hydrological cycle, so that these changes can be distinguished from natural variation, and so as to be able to distinguish biophysical from economic causes of scarcity. Economic justification for an initiative may also require that initiatives aimed at protection of freshwater supplies be part of a package of approaches designed to capture the value of multiple ecosystem services found in landscapes, and to resolve conflict among various uses. Given the heterogeneity of landscapes and the site-specific nature of ecosystem services, a key challenge is to develop the capacity for place-based approaches to monitoring and assessment. This article provides an overview of the types of economic instruments and institutional arrangements used to capture the value of watershed services, the assumptions on which they are based, institutional challenges faced in implementing them, and the kinds of scientific information needed to identify economic trade-offs, inform stakeholder negotiations, and support decision-making.

INTRODUCTION

The general decline in the capacity of watershed ecosystems to provide essential goods and services has led to an increased recognition of the ways in which they support human well-being. These include direct and indirect economic benefits, ranging from provision of freshwater for various consumptive and nonconsumptive uses, to those associated with regulation of the flow of water and sediment, and support for ways of life that have cultural value and that involve land use practices consistent with continued provision of services. However, unless those asked to pay the costs of practices necessary to insure continued provision of these services are also assured of benefits, there is little incentive for them to do so.

Regulation of land use practices alone is often ineffective because they tend to place a disproportionate share of the burden on upstream land users without giving them a corresponding access to benefits. For example, it is common for states to claim ownership of forested areas, and to protect watersheds through policies that exclude local populations from access to resources on which they have traditionally relied, which may lead them to occupy more marginal land areas (Tomich *et al.*, 2004). Those who

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

do practice farming in upland areas may barely be able to recover their costs of production even in the absence of conservation measures.

This situation has led to an increased interest in the use of market-based mechanisms as a way for upstream land users to recover the costs of protecting watershed services. A recent review by IIED identified 287 initiatives of payments for ecosystem services of forests, of which 61 were specifically for those associated with watersheds. The main concerns addressed in these initiatives have been maintenance of dry season flows, protection of water quality, and control of sedimentation (Landell-Mills and Porras, 2002).

In theory, market-based approaches can lead to more efficient allocation of resources and to more cost-effective solutions to the problem of watershed degradation. In practice, there are a number of scientific and institutional challenges encountered in their implementation, and little evaluation of the transaction costs of addressing these challenges. However, in evaluating these costs, it should be kept in mind that the development of institutional capacity needed to effectively respond to watershed degradation can also have other benefits. For example, it can lead to social cooperation in other matters, development of skills, opportunities for clarification of land titles, and increase of scientific understanding, and environmental education (Landell-Mills and Porras, 2002). It should also be kept in mind that regulatory and market-based approaches are not mutually exclusive, and often complement one another. Given also that ecosystem services tend to be very site-specific, there may be no clear-cut answer regarding *the* most efficient approach.

To the extent that there is rivalry over access to a limited supply of watershed services, a central challenge is to develop institutions or enforceable rules through which access can be limited to those who are entitled to them, and which also define responsibilities for actions needed to insure they are provided (Ostrom, 1990). To the extent that watershed goods and services have characteristics of public goods, which makes it difficult or expensive to limit access, the willingness of potential beneficiaries to pay for them, depends not only on demand but also on whether they have confidence in the effectiveness of proposed management actions needed to ensure that the service is actually delivered and that they will have access to the stream of benefits. In other words, the value of watershed services will depend on

- the integrity of ecosystem functions or processes that support service provision;
- the scale at which impacts or benefits have economic significance; and
- the effectiveness of institutional arrangements needed to insure provision and access.

All of the above are often assumed rather than assessed.

A fundamental paradox is that, given the complexity, natural variability, and stochastic nature of multiple interdependent and site-specific factors that ultimately determine outcomes, and the spatial and temporal separation of causes and effects between upstream and downstream, and between the present and the future, complete information is unobtainable and uncertainty is inherent. Market mechanisms, on the other hand, tend to be more effective when uncertainty is low, because buyers like to know if they are getting what they pay for. A precise determination of costs and benefits and their distribution, for purposes of establishing market values, presumes the ability to link actions and outcomes, so as to be able to demonstrate this. Making uncertainty explicit may be a harder sell, but is critical to managing buyer expectations and maintaining their cooperation in the long term. It is also a critical consideration in negotiating an equitable distribution of costs and benefits.

In the absence of an independent and transparent process of assessment, initiatives have often been based on myths about land and water relationships that can lead to inappropriate or partial solutions and also to placing a disproportionate share of the blame on marginal groups in remote upper watershed areas. Equally misleading are notions that science can provide certainty and that markets can solve all problems. However, science can allow a better approximation as to the magnitude and direction of impacts, monitoring, and more informed decision-making, while market-based instruments are critical to finding ways to cover the costs of providing valued services. Because of uncertainty, which is inherent in complex problems, to some extent, myths may be unavoidable and even useful as plausible scenarios, but need to be continuously questioned as new knowledge becomes available, and replaced when they are found to be implausible or irrelevant in a particular context.

This article provides an overview of the various kinds of potential watershed services, their biophysical characteristics, the kinds of instruments that have been used to create incentives for their provision, and the information used to support them. This is followed by discussion of their inherent assumptions, and challenges faced in implementing them. The article concludes by identifying the kinds of scientific information needed to identify economic trade-offs, to inform stakeholder negotiations, and to support decision-making.

DEFINING WATERSHED SERVICES

The ability to link payments to the level of service provision requires an identification and quantification of benefits actually provided, in a specific context so as to be able to determine their economic significance. This section outlines

the various types of watershed services and biophysical characteristics that need to be considered in doing this.

Watershed services are products of ecosystem functions or processes that provide different kinds of direct and indirect streams of benefits to humans in the following general categories:

1. Provision of freshwater for
   (a) consumptive uses (drinking, domestic, agricultural, and industrial),
   (b) nonconsumptive uses (hydropower generation, cooling water, and navigation).
2. Flow regulation and filtration, key aspects of which are the control of mean surface runoff, peak or flood flows, base or dry season flow, and erosion and sediment load, as well as recharge of groundwater and soil moisture (UN FAO, 2002). Benefits of these may include the following:
   (a) Water storage in soils, wetlands, and floodplains, which can buffer flood flows and drought.
   (b) Control of erosion and sedimentation which, in excess, can have adverse effects on aquatic life, irrigation canals, dams, and navigation. Below normal flows of sediment downstream from dams can have adverse effects on coastal areas where it provides protection from erosion and nourishes the development of mangroves, both of which can reduce storm damage.
   (c) Maintenance of river channels, wetlands, riparian habitats, fisheries, and other wildlife habitat that may be important for hunting, migratory birds, rice cultivation, and fertilization of floodplains.
   (d) Maintenance of mangroves, estuaries, and coastal zone processes, which often rely on seasonal pulses of freshwater inputs, and are critical habitats for fisheries as well as for other marine life.
   (e) Control of the level of groundwater tables that may have adverse effects on agriculture by bringing salinity to the surface.
   (f) Maintenance of water quality, which may be impacted by inputs of nutrients and organic matter, pathogens, pesticides, and other persistent organic pollutants, salinity, heavy metals, and changes in the thermal regime.
3. Supporting services that may include the following:
   (a) The maintenance of natural flow and disturbance regimes as drivers of ecosystem processes and which support ecosystem resilience. Resilience, in turn, provides some measure of insurance against uncertain effects of a change in conditions, for which thresholds are generally uncertain.
   (b) Support for cultural values which may include aesthetic qualities that support tourism and recreational uses, and support for ways of life.

These various kinds of services are interdependent, in that there is a trade-off between provision of freshwater for direct uses, and on the regulatory and supporting services that insure continued provision. Therefore, a fundamental objective should be to achieve an acceptable or optimal balance between these trade-offs.

The above list only represents the kinds of benefits that watersheds may provide. However, benefits cannot be considered "services" unless they also have economic significance. Therefore, site-specific assessments are necessary to identify benefits that are provided in a specific context, and the scale at which they can be detected. This then provides a basis for identifying the economic significance to various stakeholders, so as to be able to identify potential "buyers", actions required to insure that benefits continue to be provided, and the levels of compensation needed to create an economic incentive for potential "sellers". This is further discussed in the section on challenges to implementation.

In the absence of an independent, transparent, and place-based process of assessment, payment initiatives are often based on myths or general beliefs regarding watershed services and their values. These can, in turn, lead to ineffective or partial actions, inefficient allocation of resources, and often to placing an inappropriate share of the blame for watershed degradation and water scarcity on marginal groups in remote upland areas – all of which can exacerbate rather than solve problems, or create entirely new ones (Kaimowitz, 2001). Myths regarding land and water relationships fall into three general categories:

- *Inappropriate generalizations* – from one site to another, and, in particular, application of knowledge from wet temperate to arid and tropical zones.
- *Forests and water myths* – for example, that forests significantly reduce or prevent flooding and increase dry season flows. Whether or not either of these occur depends on numerous site-specific factors that determine the levels of evapotranspiration and infiltration, and, therefore, the quantity of water that is available to stream flow, and the scale at which increased or decreased flows of water are significant. For example, forests may significantly reduce flooding in a localized area, but tend to have an insignificant impact or to be averaged out at larger scales, where runoff is received at different rates from many different sources in the upper watershed. The types of vegetation that occur, the depth of roots relative to water tables, and whether there is ground cover will also play a significant role in the amount of evapotranspiration that occurs, and in the levels of stream flow. Soil that has been compacted as a result of previous management activities, the presence of roads, and other construction associated with development, can also reduce infiltration and

change drainage patterns, regardless of whether trees are planted (Calder, 1999).

- *Erosion myths* – that soil conservation practices in limited areas upstream can have a significant impact on downstream areas, particularly in arid areas with naturally high rates of erosion. For example, modifying land use practices in areas where erosion is naturally high will not prevent sedimentation of dams, though it may have on-site benefits for the farmers (UN FAO, 2002). Another fallacy regarding erosion at a landscape level is that it can be determined from plot level measurements, which assumes that all eroded soil has a negative impact, regardless of where it is deposited, and that agriculture is the main culprit (Swallow *et al.*, 2001). There is considerable evidence that erosion rates are significantly higher on roads and foot paths and marginal areas such as forest margins and steep slopes (Bruijnzeel, 2004). A recent study in an upper catchment in northern Thailand found that unpaved roads produce as much sediment as agricultural land despite the fact that these roads occupy less than one-tenth of the area occupied by agriculture (Ziegler and Giambelluca *et al.*, 2004).

In sum, land use and hydrology interact in complex ways, in which forests are one of several factors that should be considered in the context of the entire flow regime. Even in small subbasins, there may be complex interactions between various biophysical aspects and between them and human management systems that make it difficult to predict management outcomes. Whether any of the above presumptions hold therefore depends on the outcome of multiple interactions, and is therefore highly contextual or site-specific.

Not to be overlooked are increases in water scarcity brought about by increases in demand, both upstream and downstream. For example, a case study in Thailand suggests that dry season flows have diminished primarily because of a dramatic increase, both downstream and upstream, in dry season cultivation and irrigation of soybeans by those who own paddy fields. However, the focus of regulation intended to address the problem has been on the more vulnerable farmers who are dependent on rain-fed slopes in areas where significant forest cover remains, who have the least significant impacts on hydrology, and who are generally regarded as guardians of resources rather than as legitimate users (Walker, 2003).

Even when land and water links can be demonstrated and quantified, a recent literature review on the subject raises questions about whether the magnitude of damages or benefits is likely to be economically significant, when considering just the relationship between land use and hydrology. This will largely depend on downstream economic interests that rely on water, and the scale at which impacts are significant (Aylward, 2004). However, there are other off-site values associated with land use practices, in addition to those associated with freshwater flows that may have significance when combined (e.g. biodiversity protection, carbon storage, recreational values). Therefore, it is important to consider how watershed protection contributes to basin-wide management objectives.

TYPES OF MARKET-BASED INSTRUMENTS USED TO CREATE INCENTIVES FOR PROTECTING WATERSHED SERVICES

As mentioned above, initiatives to develop market-based incentives for the provision of ecosystem services associated with freshwater have been made possible by a general recognition that regulations alone are inadequate for achieving this. Also, the increased threats or the increased perception of threats has led to an increase in willingness to pay by beneficiaries of these services (Landell-Mills and Porras, 2002). The main concerns addressed in the initiatives reviewed by IIED have been maintenance of dry season flows, protection of water quality, and control of sedimentation (Landell-Mills and Porras, 2002). However, in the absence of the use of appropriate economic instruments and supporting institutional arrangements among beneficiaries and providers, that is, buyers and sellers of services, these values are hypothetical. This section describes the various kinds of economic instruments used to protect watershed services, discusses factors that need to be considered in their selection, and provides examples of their use.

Payment arrangements for watershed services may take various forms depending on the nature of the service, what is required to provide it, whether and how it is possible to limit access to benefits to those who pay the costs of provision, the scale of relevant ecosystem processes that support it, and, more generally, on the geographical and historical context. They may range from informal, community-based initiatives, to more formal contracts between individual parties, and to complex arrangements among multiple parties facilitated by intermediary organizations. They may also include a mix of complementary market-based, regulatory and policy incentives that are more likely to become necessary at larger scales, when threats are beyond the response capacity of individual communities (Rose, 2002), and when multiple services are involved. The government may play different kinds of roles depending on the type of arrangement, ranging from the enforcement of contractual agreements, to the creation of regulatory incentives, monitoring compliance, contracting with service providers, providing technical assistance, and identification of priority conservation areas. Some of these roles may also be filled by nongovernmental organizations, which often have the advantage of greater flexibility, and the ability to act more expediently. NGOs may also play the role of advocates on behalf of less powerful constituencies, so as to create

political pressure that may be necessary for governments to recognize their rights and respond to their concerns.

Actual design of payment arrangements will also reflect policy decisions as to who should pay costs and who is entitled to benefits. An example of a policy decision would be whether or not farmers should be entitled to the value of farmland when it is converted to urban uses, when this is driven by public investments in infrastructure, and the extent to which they should be compensated for any restrictions on land use conversion (Buist *et al.*, 1995).

Individual economic instruments can roughly be placed in the following general categories, and are often used in combination:

Voluntary Contractual Arrangements (VCA) – These typically involve the negotiation and agreement of a contract in which resource users, who benefit from watershed services, compensate upstream landowners for the costs of adopting management actions needed to insure provision. Intermediary organizations, such as landowner associations, are often necessary as a way to reduce transaction costs associated with the need for agreement and collaboration among numerous downstream beneficiaries and landowners dispersed over large upper watershed areas. An important consideration is the relative power of stakeholders at the bargaining table, what information is available to them, and whether any significant stakeholders have been excluded. Another is whether land users have some form of property right or tenure security, without which they may not have authority to enter into contractual agreements and therefore have access to the benefits. VCAs are more straightforward when negotiated among individual parties, such as the case of an agreement between the La Esperanza Hydropower Company and the Monteverde Conservation League in Costa Rica, which is the sole owner of the forested area upstream from the plant (Rojas and Aylward, 2002). Agreements among numerous parties will also require that more consideration be given to the establishment of decision-making entities for purposes of allocating funds to priority conservation measures. This may take the form of a trust fund, such as FONAG (Fondo del Agua), a trust fund established in Quito Ecuador to protect 2 upstream ecological reserves. This fund is overseen by a stakeholder board, and allocates pooled funds and in kind support received from municipal entities that provide water and electric services, from NGOs, and from private sources (Echavarria, 2002a; Echavarria, 2003).

Transfer Payments (TP) – These are payments made to land owners as compensation for the costs of adhering to specified management practices. This is a hybrid approach, in that payments are usually made on the basis of VCAs with landowners, but differ in that the contract is normally made between landowners and the government, for purposes of achieving broader policy objectives rather than directly with downstream users, in response to their specific concerns. When voluntary, landowner participation will depend on whether payments are sufficient to offset their opportunity costs, which may be revealed by allocating them on the basis of bidding systems. It is also a mechanism often used to allocate funds collected through various sources, which may include User Fees, conservation donors, and general tax revenues. Transfer Payments may also be funded through the sale of marketable permits (described below) to those who have higher costs of regulatory compliance. The best-known examples of TPs are the United States Conservation Reserve Program and similar initiatives in some countries in the European Union, in which farmers are compensated for conservation measures based on a number of criteria that include water quality (USDA, 2000). An arrangement established by irrigation farmers in the Cauca Valley in Colombia, in which they voluntarily pay additional water-use fees to support watershed conservation, could also be considered a system of transfer payments. This is because the farmers themselves do not have the authority to directly fund watershed management activities. Instead, the fees are paid to a governmental entity – the Corporación Autonóma Regional del Valle del Cauca – which uses them to implement existing watershed management plans (Echavarria, 2002b).

Acquisition – This approach encompasses many forms, ranging from the acquisition of rights to water and land parcels, to the acquisition of partial rights through easements that restrict uses, and leasing, in which rights are acquired for a specific period of time. In the context of watershed protection, it is common for governments or nongovernmental entities, such as land trusts, to acquire development rights on particular parcels of land so as to prevent their conversion to other uses, while owners may retain rights to occupancy and/or other specified uses. An example is the purchase of land and conservation easements in critical areas that cover the Edwards Aquifer, which supplies drinking water to 1.5 million people in Austin and San Antonio Texas (US) (Trust for Public Land, 2001). It is also one of the instruments used to implement the New York City Watershed Agreement, in which the city invests in upstream watershed protection measures (Perrot-Maître and Davis, 2001; Catskill Corporation, 2001; New York Department of Environmental Protection, 2001).

Tradable Development Rights (TDRs) – This is a form of acquisition that shifts the cost of acquiring development rights to developers and future residents. Using this approach, developers who acquire the rights from areas designated for conservation are granted permits to build at higher densities than would otherwise be allowed, in areas specifically designated for development. The more successful initiatives are generally part of a comprehensive regional plan that justifies the designation of conservation and development areas. The best-known example of the

use of TDRs for the purpose of insuring the provision of freshwater may be the New Jersey Pinelands Development Credit Program in the United States. A key concern is that the Pinelands occupy an area of sandy soil that covers a very large aquifer, and where massive development is otherwise expected (Collins and Russell, 1988) (New Jersey Pinelands Commission, 2004). A related approach is the US Wetland Mitigation Banking program, in which wetland restoration may be funded through the sale of credits to developers, who may be required to purchase these as compensation for development impacts on wetlands that cannot be mitigated. The sale of credits also provides a way of concentrating wetlands restoration efforts in areas where they can be most beneficial (Liebesman and Plott, 1998; Salzman and Ruhl, 2002). A survey conducted in 2000 identified over 230 operating wetland mitigation banks, over half of which were commercial, and which involved the restoration of 16 500 ha in exchange for the development of 9500 ha (IWR, 2000; National Research Council, 2001).

Marketable Permit Systems (MPS) – These are similar in concept to TDRs, but, in this context, are applied to trades between point and nonpoint sources of pollution. Once allowable levels of pollution and resource use have been established as a matter of policy and regulation, MPSs provide a way to reduce the overall costs of compliance by allowing those who have higher compliance costs to purchase credits from those whose costs are lower, usually through intermediary organizations that register and verify credits. They may also be structured such that proceeds from the sale of permits are used to fund transfer payments, thereby shifting the cost of payments from the public to the private sector. They may be in the form of a cap-and-trade or credit programs. The first requires establishing an aggregate limit on resource use or emission of pollutants. The second instead offers credits to those who reduce emissions or resource use beyond legal requirements or engage in specific conservation practices. Among the key policy issues is the choice of method for allocating initial rights to shares of the total allowable emissions, and rules regarding the transfer of those rights (Tietenberg, 2002). For example, rights may be allocated on the basis of existing uses, lotteries, auctions, or rules that define who is eligible. Limits on transferability may be used to prevent concentration of rights in the hands of a few, or to maintain rights within a particular community, but may also reduce efficiency by reducing the pool of buyers and sellers. A third policy issue is for stakeholders to reach an agreement as to what constitutes an equitable allocation of the burden of emissions reduction among point and nonpoint sources, so as to avoid violating the "polluter pays" principle (King, 2003).

In the context of watersheds, several programs are under development in the United States, though very few trades have actually taken place to date. Those few include one among nonpoint sources in the Dillon Reservoir, and one in Minnesota, in which the Minnesota Pollution Control Agency permitted the Rahr Malting Company to construct a downstream facility that would increase biological oxygen demand, in exchange for funding reductions of upstream nonpoint sources of phosphorus. However, trades are expected in a number of other basins as estimates of Total Maximum Daily Loads (TMDLs) are completed (Nutrient Net, 2004). TMDLs define permissible levels of emissions that are consistent with achieving water quality standards, and provide a basis for allocation of the burden of emission reductions. In the Tar Pamlico basin in the state of North Carolina, an association of point source dischargers has purchased credits from the state that have been banked for use when their cap is exceeded, and that are intended for future use in funding Best Management Practices (NC NPSMP, 2004). A credit-trading program is also being used for salinity reduction in New South Wales in Australia. In this case, an association of farmers purchases salinity credits from the State Forests agency of NSW, which, in turn, contracts with upstream landholders to plant trees, which also provide carbon storage benefits (Perrot-Maître and Davis, 2001; State Forests of New South Wales, 2001; Sundstrom, 2001).

Certification and Labeling (CL) – This creates an incentive to adhere to specified management practices by providing consumers with the information necessary to make their choices of products more consistent with the values they place on ecosystem services. This may increase the market share of a product, and/or result in a price premium. However, producers are not always the ones who benefit from price premiums, given the number of actors between them and those who ultimately purchase the products. This arrangement requires intermediary organizations to establish standards for labeling and to certify practices, all of which has a cost. An example is the Salmon Safe initiative, which certifies and promotes wines and other agricultural products from Oregon farms and vineyards that have adhered to management practices designed to protect water quality and salmon populations (Salmon Safe, 2003).

In addition to the sale of permits and voluntary payments based on contractual agreements are a number of different kinds of sources of funds, ranging from user fees, to taxes, donations, and earmarked proceeds from sales of specific products. The appropriateness of particular sources will depend on whether or not it is possible to limit benefits to those who pay for them, and the scale at which benefits are detectable and significant.

For example, *User Fees* (UF) involve charges to users of a resource, based on the principle of "User Pays", which is only feasible when it is possible to limit benefits to those who pay for them. To be considered an instrument for protecting services, funds must also be specifically designated for conservation purposes. Intermediaries usually play a

role in the collection of funds and their disbursement to landowners. These are often added to the fees already paid by users for delivery of water. Examples of this approach include the cases of New York City, Quito Ecuador, Cauca Valley Colombia, where charges are added to existing water charges, and are designated specifically for funding upstream conservation measures (New York Department of Environmental Protection, 2001; Perrot-Maître and Davis, 2001; Echavarria, 2002b; Echavarria, 2002a). Fees may also be differentiated so as to support policy objectives of giving priorities to particular uses. For example, under the South African water law, water licensing charges vary by sector and do not apply to a certain amount of water that is reserved for basic human and ecosystem needs (DWAF, 1999a).

When it is not possible to limit access to benefits to specific users, *taxes* may be more appropriate. Transfer payments, such as the United States Conservation Reserve program, tend to be funded through general tax revenue. In Costa Rica, the National Fund for Forest Financing (FONAFIFO), a program of payments for ecosystem services that includes protection of watersheds, is in part funded by a fuel tax, in addition to payments from beneficiaries. In Colombia, watershed management is in part funded through a 6% tax on the revenue of large hydroelectric plants. This tax also supports political decentralization, as it is a key source of funding for regional authorities, which have the primary authority for watershed management. This revenue is divided among the autonomous regional corporations and municipal governments. A portion of the allocation to the autonomous regional corporations goes to a general fund that supports management of watersheds in which there is no hydropower facility. In addition, 1% of funds invested by towns in water projects must be invested in watershed protection.

Donations from external sources may be important for addressing impacts associated with actions of external actors, such as those associated with timber concessions and other uses of government owned land, and with macroeconomic policies. Funding from external donors may also be necessary to finance the high transaction costs faced up front, in the beginning stages of an initiative, associated with feasibility studies, assessment activities, and capacity building, generally done through intermediary organizations.

Earmarked proceeds from the sale of specific products that are associated with special places and symbols may simply be designated for the support of conservation objectives that have broad benefits, which may also influence consumer purchase decisions and have other benefits for the company or organization. For example, in the state of Maryland (in the United States), for an additional fee, car owners have an option of obtaining Chesapeake Bay license plates, which supports the Chesapeake Bay Trust – an entity created by the state legislature to support Bay restoration. The University of Maryland also recently began to designate 2% of the proceeds from the sale of items associated with the "Fear the Turtle" slogan toward the protection of the terrapin – the school mascot and official state reptile (Hsieh, 2003). The terrapin is generally regarded as an "ambassador" for the Chesapeake Bay ecosystem, as it relies on its water quality, shorelines, salt marshes, and tidal rivers, and has provided the basis for extensive public education and outreach regarding comprehensive approaches to conservation (Dunlap, 2003; Golder *et al.*, 2000).

Most arrangements will consist of a package of various instruments. For example, TPs may be made through VCAs with sellers, using funds derived from UFs. Individual arrangements may also be part of a comprehensive plan intended to protect multiple services, which, in addition to those associated with freshwater, may include other benefits of forests, such as carbon storage, aesthetic values, and biodiversity. These are the four services for which forest owners who adhere to approved management plans are compensated in Costa Rica's FONAFIFO program, which is then able to sell them to different kinds of buyers. For example, hydroelectric companies and municipalities may pay for watershed benefits, tourism agencies for landscape beauty, and foreign energy companies may purchase carbon offsets. State Forests of New South Wales has also initiated an Environmental Services Scheme in which landowners are to be compensated through a credit scheme for multiple benefits of forests, including carbon sequestration, biodiversity, soil conservation, in addition to protection of water quality and offsetting the rise in salinity levels (State Forests of New South Wales, 2004). This approach may also be necessary when watershed benefits alone are not sufficient to offset the opportunity costs of forgone land uses (Landell-Mills and Porras, 2002).

Fundamental to all of these instruments is the need for a consistent set of criteria and a transparent process for decision-making with respect to the establishment of priorities for allocation of the funds. It is also important to keep in mind that market-based mechanisms are not a substitute for regulations, which are often necessary to create an incentive to find more cost-effective ways of achieving compliance through the use of market mechanisms. For example, in the well-known case of New York City, the incentive for the city to negotiate with upper watershed communities and to pay for upper watershed conservation measures was created by new regulations that would have otherwise required much higher expenditures to construct a filtration plant (Perrot-Maître and Davis, 2001). Trading schemes rely directly on regulatory standards or caps, without which there would be no incentive for using market-based approaches.

In general, benefits will be more tangible, and contractual agreements more feasible, at smaller scales, where links between causes and effects can be more readily established, and where property rights and stakeholders can be better defined. At larger scales, where it is harder to link causes and effects, and rights and responsibilities are harder to define because of public good or common pool characteristics of the service, or of multiple services, there will be a greater need for government involvement. Larger scales also offer a larger pool of buyers and sellers, but are harder to tailor to local conditions (Rose, 2002).

CHALLENGES TO IMPLEMENTATION

Implementation of market-based payment arrangements for watershed services will be difficult unless beneficiaries are both able and willing to pay for them. Willingness to pay will generally depend on perceptions and beliefs regarding the benefits of ecosystem services. In the long if not the short term, this may require a demonstration that there are measurable benefits that support livelihoods and overall well-being. It will also require the development of institutional arrangements that insure access to benefits by those who pay for them, over the relevant period of time. By providing compensation for inevitable trade-offs among conflicting interests and objectives, market-based instruments will often be one element in a more comprehensive management strategy. This section begins with an overview of the difficulties of demonstrating measurable benefits, and the use of scientific and other information to evaluate trade-offs. It concludes with a discussion of the various kinds of institutional arrangements needed to support payment initiatives.

Demonstrating Measurable Benefits

Assessment and verification of the benefits of ecosystem services, and the effectiveness of management actions taken to insure continued provision, are critical to building and maintaining stakeholder confidence and willingness to pay for them. In general, this aspect has received less attention than the identification of able buyers and systems for collecting payments (Pagiola *et al.*, 2002). However, it can require gathering of data over time and significant commitments of funding and expertise that are not available until an initiative is well established. Rules of thumb may be sufficient in the initial phase, provided initiatives are designed to include monitoring and assessment that allows them to be improved over time.

Key questions that should be answered in assessments, or for which they should at least provide a working hypothesis are as follows:

1. What ecosystem functions support the provision of specific benefits and what are their key parameters, or, how can they be measured or approximated? How does land use in the watershed affect these functions?
2. What is the direction and magnitude of changes in parameters of interest?
3. At what spatial and temporal scales can these changes be detected?

Investigation of changes in the water balance and the impacts of these changes on the flow regime provide a framework for quantifying changes in watershed conditions relative to management objectives. These objectives, or points of interest, may include the following:

- Total flow yield
- Attenuation of peak or flood flows
- Maintenance of minimum low or dry season flows
- Protection of water quality
- Recharge of groundwater
- Protection of biodiversity.

(Tognetti *et al.*, 2004).

The most significant changes in the water balance are linked to extreme and randomly timed events. For example, most sediment and pollutants are transported during storms, which also produce the periodic high flows necessary for maintenance of channels, riparian areas, wetlands, and coastal mangroves. Such events are the dominant process in upstream hill-slope areas, which make up 70 to 80% of catchment areas, have tightly coupled land and water interactions, and have greater variability in discharge. This is in contrast with downstream fluvial processes that are more continuous and are dominated by regular flood pulses and movement of bedload within the channel as well as cumulative impacts of hill-slope processes (Gomi *et al.*, 2002). Therefore, it is more important to identify known ranges and patterns of variability and uncertainty than to identify average values. These ranges can then be used to put lower bounds on unknown thresholds of resilience – which refers to the range of variability within which changes can be adjusted to (Tognetti *et al.*, 2004).

Human uses also play a significant role in the hydrological cycle, and should be accounted for so that their cumulative impacts can be distinguished from natural variation, and so as to be able to distinguish physical from economic causes of scarcity – which has implications for the kinds of measures that will be effective. Economic causes of scarcity imply the need for changes in allocation among human uses. Physical causes of scarcity imply the need for changes in the flow regime, which refers to allocation between ecosystems and human uses, and changes in land use practices that affect the timing and routing of flows. This should be accounted for on a seasonal basis – to the extent water is limiting during dry periods, there will generally be a greater willingness to pay for it (Tognetti *et al.*, 2004).

Basin-wide seasonal Actual Evapotranspiration (AET) is a principal component of the water balance. It is also a key source of uncertainty because it is a function of numerous variables that include climatic factors, vegetation, and land use. Given the heterogeneity of landscapes, this is difficult to estimate. However, it is important to consider land use practices and landscape characteristics such as riparian zones, which have disproportionate impacts on levels of runoff and sedimentation and which may account for significant differences in AET (van Noordwijk *et al.*, 2004).

Because of such heterogeneity, land use impacts on flows of water and sediment are generally best addressed at small scales, at level of hillslopes and patches, at which they can be detected. With some exceptions (e.g. suspended sediment and microbial contamination), water quality impacts, diversions, and impacts of large infrastructure are more appropriately addressed at basin-wide scales, which allows for consideration of a broader range of trade-offs.

The measurement of benefits should ultimately provide a basis for estimating relationships between the extent of changes in land use and the amount of benefits provided, so as to provide justification for payment amounts. Given inherent uncertainties, it may be sufficient to rank the relative values of land uses and practices for achieving desired outcomes, and to identify opportunity costs and conflicts associated with changes in land use practices. As a general rule, greater precision and more data will be needed when it is unclear whether or not benefits exceed costs. However, this also depends on what costs and benefits are considered – which is largely a political and institutional question.

Identifying and Evaluating Trade-offs

Identifying the economic significance of changes in watershed processes provides a basis for identifying actual or potential trade-offs between meeting various objectives. This will require consideration of downstream land uses and stakeholder vulnerability, which depends on what options they have, and on their opportunity costs. Just as with biophysical processes in watersheds – vulnerability, options available to stakeholders, and conflicts become more transparent under extreme conditions. Impacts may be both positive and negative, depending on what is measured and valued. For example, whether floods increase or decrease welfare will depend on the presence of development in flood plains, and valued habitats that rely on periodic floods.

An example of an extensive assessment that was done regarding relationships between land use and hydrology is a study of the Arenal basin of Costa Rica, which examined the marginal values of changes in flows of water and sedimentation for a downstream hydroelectric plant. In this particular case, it was found that, while sediment from pasture compared with forested areas did have a cost in terms of lost hydroelectric production, ranging from $35 to $75/ha, this was exceeded by the benefits to the facility of increased water yield from pasture areas, which ranged from $250 to $1100, depending on the type of forest area cleared – the greatest yield of water appeared to be associated with fragmented cloud forest areas that have the highest rates of interception of precipitation (Aylward and Echeverria, 2001). This (counterintuitive) result occurs in part because the Arenal reservoir is an interannual regulation reservoir, in which hydroelectric production depends on total flows, and is therefore largely independent of dry season flows. Given that the reservoir also contains a large amount of dead storage space, sedimentation also provides a benefit when it fills this space and displaces water upward, which makes more water available for production.

Given the high economic values associated with forest clearing in the Arenal study, ranching was also found to produce higher net present values than was offered by the government for reforestation and thus to be more economically efficient. On the other hand, these costs and benefits were not uniformly distributed. A subsequent companion study that examined costs and benefits from the perspective of major stakeholders, and which made distinctions among various kinds of landholders, found that the higher return per hectare depended in part on location in the catchment, that they accrue primarily to large landholders, and that incentives that were being offered for conservation may still appear attractive to small landholders who, not coincidentally, also tend to disproportionately occupy the steepest slopes (Aylward and Fernández González, 1998). The scale of the study did not permit it to answer questions regarding the more localized impacts of forest clearing on landowners residing within the upstream area.

In contrast with the results at Arenal are those of another Costa Rican reservoir that is part of the La Esperanza hydroelectric facility, which is smaller, depends on dry season flows, and has higher costs associated with sedimentation. In that case, payments from the hydroelectric company to the Monteverde Conservation League, the sole owner of the upper watershed area, are simply based on the value to the company of reducing uncertainty that would accompany any change in upstream land use (Rojas and Aylward, 2002).

Because of natural variability, landscape heterogeneity, and human activities, it will be difficult or impossible to unequivocally establish links between land use and water yield at the watershed scale, even with a full analysis of land and water relationships. For example, in the case of Arenal, there are a number of land uses and vegetation types that include pasture, forest with different canopy heights, cloud forest with different degrees of fragmentation, all of which are changing over time, are a key parameter in estimating water balances, and are therefore a major

source of uncertainty. A field study in the Arenal basin found the highest capture of precipitation in fragmented primary cloud forest, of 12% compared to a negative gain of 11%, in intact high primary forest. However, during the dry season, yields were positive for all types of forest, when water availability is most limiting and therefore has higher economic value, with gains ranging from 15 to 53% (Fallas, 1996). Other potentially significant sources of variability in the interception of precipitation in cloud forests, often not accounted for, are the position of the slope in relation to moist winds and storm intensity, which affects the amount of water held in the forest canopy (Bruijnzeel, 2001). Although water quantity impacts associated with forest clearance are often ambiguous, water quality impacts can generally be expected to be negative (Aylward, 2004).

Recent reviews of the scientific knowledge base that has supported many initiatives in Costa Rica suggests that management is generally limited by the lack of reliable and precise information on forest water linkages (Pagiola, 2002; Rojas and Aylward, 2003). Instead, most are based on conventional wisdom, secondary sources of information, and also on selective references to more balanced literature reviews on forest hydrology. In other words, regardless of the source material, information used in decision-making tends to invariably support statements that protection of forests will increase water yields (Rojas and Aylward, 2003). In some cases, such as in the Arenal basin (Castro and Barrantes, 1998) and in Heredia (Castro and Salazar, 2000), the values of watershed protection are calculated on the basis of the opportunity cost of returning cleared land to forest cover, with no attempt to model and assess links between land use and hydrology, and to estimate the marginal values of water in specific consumption and production activities. In other words, estimates of opportunity costs of land under alternative uses is used to infer the benefits of forest hydrological services. In the latter case, payments are also justified on the basis of the statement that "Costa Rican society positively correlates the presence of forest with the supply of hydrological services".

Faced with these and other kinds of uncertainties, such as data that is rarely if ever complete, reliance of models on average values, the random timing and disproportionate significance of extreme events – decisions to protect watershed services are often based on the costs of alternative courses of action rather than on the value of changes in land use practices. For example, in the New York City case, the decision to invest in upstream conservation and upgrading of infrastructure was based on the avoided cost of building a filtration plant that would have otherwise been required to meet a new regulatory standard designed to insure safe drinking water. However, the relationship between changes in land use practices and resulting water quality remains the subject of ongoing research.

Other factors that have been used to inform the analysis of trade-offs and to justify decisions and influence WTP have included

- costs of implementing management plans;
- regulatory costs avoided;
- reduction of uncertainty associated with proposed changes in land use;
- individual WTP in the form of user fees and purchase of certified products;
- political WTP as indicated in national budget allocations of tax revenue;
- landowner WTA compensation or cost of supplying the service; and
- consistency with comprehensive management plans.

Given the period of time that may be required to detect the results of changes in management practices, and the impossibility of obtaining complete information, it will usually be necessary to begin with rules of thumb. These can provide a basis for constructing plausible scenarios, or serve as working hypotheses, to be tested in an adaptive approach to management, in which they are continuously questioned and revised as new information becomes available and in response to unanticipated consequences. Initiatives are likely to be more effective in the long term when uncertainty is made explicit and initiatives are also informed by a process of monitoring and assessment, which are essential as a feedback component in the design of any initiative.

Ultimately, the extent of information needed for decision-making will depend on what is required to justify an expenditure in a particular case, or, in other words, to convince the buyers. To the extent that uncertainty regarding links between actions and outcomes can be reduced – or at least made transparent so that it can be factored into decision-making, stakeholders will have greater economic incentive to participate in these initiatives and in helping to insure that objectives are actually achieved. A key challenge then is to develop greater capacity for conducting site-specific assessment of functions that support services, of threats they will be degraded or disrupted, and to use these to develop feasible and effective response options.

Development of Appropriate Institutional Arrangements

As discussed in the introduction, willingness to pay is inextricably linked to confidence in the effectiveness of management actions as well as of the institutional arrangements needed to insure access to benefits by those who pay the costs of management actions. In the absence of such arrangements, economic value is no more than hypothetical, as there would be no incentive to take actions needed to insure provision of the service. Institutional arrangements are essentially the "rules-of-the-game" that are needed to

resolve conflicts among competing demands on any limited resource so as to avoid its depletion or degradation. They may take several different forms, addressed in this section, chief among which are property rights, which define the rights and responsibilities of all players, and have implications on the distribution of costs and benefits. The second is the decision-making process, which may or may not allow for effective participation of all concerned, and the process of assessment – which includes both the gathering and dissemination to users, of information needed to support decision-making. The feasibility of these arrangements will depend largely on their transaction costs, and on the broader social and economic context in which they are embedded.

Institutional arrangements also serve the purpose of reducing transaction costs by facilitating transactions among numerous buyers and sellers, and, where necessary, collective action. For example, given that areas of high erosion tend to be marginal areas such as steep slopes, or common areas such as roads and paths, individual landowners will have little incentive to invest in improvements, which makes these areas *de facto*, open access (Swallow *et al.*, 2001). In this kind of situation, provision of services may require incentives for collective action by communities. When there are numerous stakeholders dispersed over large upper watershed areas, the establishment of intermediary organizations, who can negotiate on their behalf, serve as advocates for the rights of marginalized stakeholders, and provide technical assistance, which is an important way of reducing transaction costs and making an initiative feasible.

Rights to Watershed Services and Responsibilities for Their Provision

Property rights play an important role in creating appropriate economic incentives because they determine who has access to benefits and also define responsibilities for costs or actions needed to insure the provision of benefits. For example, in the absence of a clear land title, upper watershed land users will lack the authority to enter into contractual agreements and therefore be unable to benefit from payments. They may also risk eviction as values are placed on services to which they lack recognized rights (Landell-Mills and Porras, 2002). Property rights may take different forms, ranging from informal rights or norms recognized by users, to various forms of formally recognized public and private ownership by individuals, groups, or government entities. Failure to control access is often mistakenly referred to as a "common property" situation, but is actually an "open access" situation in which no property rights are in effect (Ostrom *et al.*, 1994).

Appropriateness of property regimes depends on whether the incentives they create are consistent with social objectives. For example, rights to water based on historic use or "prior appropriation", which usually also require that the water be used in ways that are considered socially beneficial, was consistent with social objectives in the 1800s of promoting development in the western United States. However, it creates a disincentive for reducing consumption as this would lead to a reduction of the amount of water a user may claim in the future, and is inconsistent with uses associated with emerging social objectives of conservation, such as in-stream flow, that are not legally defined as "beneficial" (Wilkinson, 1992). Rights to water based on possession of adjacent land or "riparian rights" allow reasonable use that does not interfere with the reasonable use by others, and may allow communities to control access and exercise customary rights. The latter, however, may limit the ability to transfer the water and to develop water markets, which can provide incentives for more efficient allocation among various uses (Meinzen-Dick and Bruns, 2000). In an open access situation, the incentive is simply to consume resources before someone else does.

Appropriate or not, property rights do not change easily or quickly in the absence of political momentum generated by events such as the end of the cold war, as their purpose is to provide some security without which there is little incentive for investment. Thus, they cannot be arbitrarily changed. However, they do tend to change over time to reflect changes in social values, as new problems emerge, and as technological improvements bring down the transaction costs of controlling access to particular resources – and are not always compensated. For example, development of hydropower at the beginning of the industrial era led to a change in rights to the natural flow of water because it was considered to be of greater value to society, and continues to lead to widespread displacement of communities. Similarly, as a consequence of the growth of urban areas, rivers became more highly valued for sewage disposal than for fisheries and recreational values. Just as changes in rights have been implicit in the development of physical infrastructure (hydropower, dams, irrigation, and navigation), the rise of values placed on freshwater services implies the negotiation and definition of new rights and responsibilities in which uses of land and water are limited to those that do not impair ecosystem functions that support valued services (Sax, 1993).

Typically, different claims and sources of authority will tend to overlap and conflict in a process referred to as "legal pluralism" (Meinzen-Dick and Pradhan, 2002). Therefore, changes in property rights tend to come about through a contested process that can lead to the development of new and more appropriate institutional arrangements. Any initiative to protect downstream water supplies or biodiversity either by providing compensation to upstream landowners for altering land use practices, or by attempting to hold them responsible for damages, in effect involves negotiating new and appropriate forms of property rights that resolve conflicts between these objectives and existing practices.

Allocation of Costs and Benefits

Equitable distribution of costs and benefits is an important aspect of feasibility in that, if significant stakeholders are excluded or disadvantaged, and regard existing rights as inequitable, they will have little incentive to cooperate in their enforcement. On the buyer side, some studies have found differences in willingness to pay that depended on the protection mechanism suggested and on the distribution of property rights. In some cases, stakeholders are unwilling to pay, not because they are unaware of ecosystem values, but simply because they do not feel that it should be their responsibility to do so (O'Connor, 2000).

Direct payments for environmental services raise fundamental questions as to who *should* pay and how much, and the extent to which providing these services should simply be regarded as an obligation inherent in the responsibility not to harm others. In some cases, transfer payments to upstream areas could be seen as violating the principle of "polluter pays", unless accompanied by sanctions on pollution (UN FAO, 2002). However, given the low prices paid for agricultural commodities, direct payments for providing services of maintaining the landscape and water quality may also be seen simply as recognition of the values of environmental services in addition to those of agricultural commodities.

Equity issues may also present an obstacle to nutrient trading between emitters of nutrients from point and nonpoint sources, as it has been suggested that these may not be regarded as fair by point source emitters who have already invested in point source reductions (King, 2003). Agreement on TMDLs that equitably allocate load restrictions between point and nonpoint sources, as a basis for the initial allocation of permits, may help overcome this.

In a number of cases that have been evaluated, it was found that payments for watershed have disproportionately benefited those who own larger tracts of forests or forest plantations, excluding smaller and marginalized landholders, who tend to occupy the steepest slopes, who do not have large forest areas that they can be compensated for protecting, and may not have the property rights needed to gain access to benefits. They also tend to exclude activities such as agroforestry and organic farming. This suggests the need for broader criteria for allocation of payments, so as to include smaller landholders and better support livelihood strategies (Rosa *et al.*, 2003).

Participation in Decision-making

Acceptance and cooperation of both buyers and sellers may ultimately depend on whether all of the relevant stakeholders cooperate, how funds will be spent, and whether stakeholders are able to effectively participate in allocation decisions, all of which are issues of governance. For example, in Brazil, which adopted a nationwide river basin management policy, domestic water users were found to be willing to pay more for water when the revenue from water fees is invested in the basin where the funds are generated, and when users are able to participate in decisions as to how the revenue is spent (Porto *et al.*, 1999). Other studies have found differences in WTP that depended on the protection mechanism suggested, and whether it was regarded as fair and effective (O'Connor, 2000). A study of water resources management in Cyprus found a higher WTP, even for less tangible values such as protection of wetlands along an international bird migration route, under scenarios in which all of the relevant stakeholders participate; in this case, all countries along the bird migration route (Koundouri *et al.*, 2003).

A key obstacle to effective participation of stakeholders in decision-making has been the association of large water resource infrastructure with the need for highly centralized management authority. In the case of large infrastructure projects, decision-making takes place at national levels, and is largely driven by geopolitical considerations in which local stakeholders have little in any voice. Because of environmental and institutional heterogeneity, highly centralized authorities tend to have a limited capacity to respond to livelihood concerns. The site-specific characteristics and variability of freshwater ecosystems and other natural resources imply the need for detailed local knowledge, discretionary powers, and also greater representation and accountability, which can increase the capacity to respond to factors such as variations in rainfall and crises associated with extreme events, as well as to mediate conflicts. Transfer of rights and sharing of benefits can also provide a stream of revenue to local governments that can be used to build and sustain capacity for resource management. Provision of watershed and other ecosystem services may therefore be inextricably linked with efforts to achieve democratic forms of decentralization, or to "pry open... local democratic space" (Kaimowitz and Ribot, 2002).

A second obstacle to effective participation is the gathering and dissemination of information needed to support decision-making. Given the site-specific nature of watershed services, this presents the challenge of developing an integrated and place-based approach to assessment. Given that adequate information seldom exists in advance of an initiative, and that complete information is unobtainable, this is particularly important in the implementation phase, so as to identify gaps between policies and practices. This provides a basis for course correction, as new information becomes available and as lessons are learned, and for engaging stakeholders in identifying feasible options.

Socioeconomic and Political Context

Ultimately, the development of market-based arrangements for watershed ecosystem services also needs to be

considered in the context of a broader trend of institutional changes in water resource management. Among the key elements of this trend are (Saleth and Dinar, 1999; Bauer, 2004)

1. a shift from the development of new water supplies through infrastructure development, to the reallocation of existing ones, and
2. efforts to improve cost recovery both for
   (a) operations and maintenance of infrastructure, and
   (b) to cover the costs of conservation management and research activities.

In theory, recovery of these costs could increase the capacity of governments to deliver basic water supplies and sanitation, and the capacity of ecosystems to provide freshwater supplies. However, cost recovery faces a major obstacle in that major water users, particularly in agriculture, which accounts for approximately 70% of water withdrawals worldwide (United Nations, 2003), are also accustomed to high subsidies and pay only a fraction of the costs of operations and maintenance alone. In urban areas, inability to recover costs from those who are served reduces the capacity of the government to extend services to the poor, who must then pay more to obtain water from trucks. Reallocation of existing supplies therefore faces significant political constraints, and may need to be linked to broader macroeconomic reforms that tend to be associated with crises or sweeping political changes, such as the end of apartheid in South Africa which, among other things, made it possible to make significant reforms in the country's water law (see box 1).

Conversely, given the role of water in most, if not all, sectors of the economy, the development of payment arrangements can have implications for broader policy reforms. For example, potential spin-off benefits associated with the development of markets for watershed services include clarification of property rights, stakeholder cooperation in other areas important to livelihoods as a result of strengthened institutions, technological transfer and skill development, development of market infrastructure, contributions toward the protection of other ecosystem services not traded in markets, improved scientific understanding, and environmental education (Landell-Mills and Porras, 2002). Payment arrangements for freshwater and other ecosystem services is therefore a long-term process of institutional development that needs to be considered in the context of broader issues of democratic governance. For markets to work, democratic institutions and equity are essential because there needs to be trust that people will obey rules and abide by agreements made, which may not occur unless arrangements are regarded as fair (Lipton, 1985). This is a continuing challenge in developed and developing countries alike.

Although development of appropriate supporting institutional arrangements can have high transaction costs, it should be kept in mind that this is no different from costs that have been and continue to be incurred in the development and maintenance of institutions that support existing markets and that are generally not paid for in the prices of private goods and services.

Box 1 National Water Act of South Africa

The South Africa National Water Act is among the more innovative in that it explicitly seeks to recover costs both for operation and maintenance and for conservation, management and research, as well as for meeting basic human subsistence needs. Therefore, water charges do not apply to minimal levels of water for ecosystem and human subsistence needs (DWAF, 1997). Beyond these basic needs, it requires the registration of water uses. Licensing and payment of fees are required for water service authorities, industrial uses, irrigation, and for activities that reduce streamflow, such as tree plantations. Licensing applications are evaluated for consistency with a number of criteria including individual catchment management strategies that are prepared in accordance with a national water strategy (DWAF, 1999b). Implementation has not been without problems. Among these are tensions with existing riparian rights holders, hardships associated with the withdrawal of subsidies from irrigators, and the inability of many of the poor to pay even minimal amounts for water delivery. Whether these problems can be resolved, and whether licensing can truly serve as a reallocation mechanism, may depend less on formal changes in the law than on the strength of public participation, and the bargaining power of the poor (Schreiner and van Koppen, 2000).

CONCLUSION

The specific services provided by watersheds will depend on how they are managed, and whether impacts have economic significance. This, in turn, depends on downstream uses of land and water, on the scale at which impacts can be detected, and on what options are available to stakeholders for maintaining their livelihoods. Therefore, market-based instruments will not solve all problems of watershed degradation, but are important tools that should be used as components of broader management strategy.

In the past, efforts to manage freshwater ecosystems to meet multiple objectives has been driven by the more dominant and tangible economic interests, typically navigation and hydropower (Barrow, 1998). More recently, in a somewhat similar fashion, efforts to protect watersheds using market-based instruments have been driven by the interests of those most able to pay such as municipal water suppliers and hydroelectric facilities, and have placed emphasis on protection of forests and protected areas. This approach tends to overlook management practices that may have greater impacts than forest clearing, and to exclude the most

vulnerable populations from access to benefits. A significant obstacle to cost recovery is that the most significant users, for example, agriculture, are accustomed to subsidized prices.

The use of market-based instruments to recover the costs of watershed management activities needs to be considered in the context of basin-wide management objectives, and should be part of a package of approaches constructed to identify the relative values of multiple ecosystem services of watersheds as a basis for setting conservation priorities, and to resolve conflict among multiple uses. This can provide a basis for engaging stakeholders in identifying a broader range of management options that are consistent with meeting the broader objectives of ecosystem management to support human livelihoods and general well-being. It may also increase their confidence and willingness to pay and to cooperate in management activities.

Experience to date with the use of market-based instruments to cover the cost of actions needed to insure continued provision of watershed services suggests that little is known about their effectiveness for actually delivering ecosystem services. This is, in part, because of the time lag between management activities and their outcomes in a watershed context, and, in part, because less attention has been given to development of an independent and transparent process of assessment.

Given the heterogeneity and constant change in ecosystems and in human institutions, the site-specific nature of watershed processes – which are dominated by randomly timed and extreme events, and the difficulty of linking multiple causes and effects, or predicting outcomes, an adaptive approach to management is required. Ongoing assessment to support decision-making is a critical component of such initiatives. However, assessments based on generalizations can only provide some rules of thumb and working hypotheses from which to begin. Perhaps the most significant challenge therefore is to develop the capacity for a place-based approach to assessment, which is necessary to identify ecosystem functions that support provision of valued ecosystem services in a specific context, and to select payment and institutional arrangements that are feasible and appropriate to that context. To be effective, market-based initiatives also need to be viewed as part of a long-term process of building appropriate institutions, and in the context of broader issues of structural reform.

Acknowledgment

This paper is in part based on a "Knowledge and Assessment Guide to Support the Development of Payment Arrangements for Watershed Ecosystem Services" prepared for the World Bank Environment Department, with support from the Bank-Netherlands Watershed Partnership Program, by Sylvia Tognetti, Guillermo Mendoza, Bruce Aylward, Douglas Southgate, and Luis Garcia. It also benefited from comments and suggestions provided by Benjamin Kiersch.

FURTHER READING

Blaikie P.M. and Muldavin J.S.S. (2004) *Upstream, Downstream, China, India: The Politics of Environment in the Himalayan Region*. *Annals of the Association of American Geographers*, **94**, 520–548.

Echavarria M., Vogel J., Montserrát A. and Meneses F. (2004) *The impacts of payments for watershed services in Ecuador: Emerging Lessons from Pimampira and Cuenca (pdf)*, International Institute for Environment and Development, Environmental Economics Programme, London. **[http://www.iied.org/docs/eep/MES Series/MES4EcuadorWatersheds.pdf]**.

Miranda M., Porras I.T. and Moreno M.L. (2003). *The social impacts of payments for environmental services in Costa Rica. A quantitative field survey and analysis of the Virilla watershed*. International Institute for Environment and Development, London. **[http://www.iied.org/eep/pubs/documents/MES1.pdf]**.

The Ecosystem Marketplace, **www.ecosystemmarketplace.com** (Sponsored by the Katoomba Group/Forest Trends).

The Flows Bulletin of News on Payments for Watershed Services. **www.flowsonline.net** (International Institute for Environment and Development and The World Bank).

Tomich T.P., van Noordwijk M. and Thomas D.E. (Eds.) *Environmental Services and Land Use Change: Bridging the Gap between Policy and Research in Southeast Asia*. A special issue of Agriculture, Ecosystems and Environment, Vol. 104/1 (September, 2004).

UN FAO (2004) *Payment Schemes for Environmental Services in Watersheds*, Land and Water Discussion Paper 3. UN Food and Agriculture Organization: Rome.

REFERENCES

Aylward B. (2004) Land-use, hydrological function and economic valuation. In *Forest-Water-People in the Humid Tropics*, Bonnell M. and Bruijnzeel L.A. (Eds.), Cambridge University Press: Cambridge.

Aylward B. and Echeverria J. (2001) *Environment and Development Economics*, **6**, 39–382.

Aylward B. and Fernández González A. (1998) *Institutional Arrangements for Watershed Management: A Case Study of Arenal, Costa Rica. CREED Working Paper No. 21*, International Institute for Environment and Development, London.

Barrow C. (1998) River basin development planning and management: a critical review. *World Development*, **26**, 171–186.

Bauer C.J. (2004) *Siren Song: Chilean Water Law as a Model for International Reform*, Washington, Resources for the Future.

Bruijnzeel L.A. (2001) *Land Use and Water Resources Research*, **1**, 1.1–1.18.

Bruijnzeel L.A. (2004) *Agriculture Ecosystems and Environment*, **104**, 185–228.

Buist H., Fischer C., Michos J. and Tegene A. (1995) *Purchase of Development Rights and the Economics of Easements*, Agricultural Economic Report 718, Natural Resources and Environment Division, Economic Research Service, U.S. Department of Agriculture, Washington.

Calder I.R. (1999) *The Blue Revolution: Land Use and Integrated Water Resources Management*, Earthscan: London.

Castro E. and Barrantes G. (1998) *Area de Conservación Arenal (ACA)*, Ministerio de Ambiente y Energía (MINAE), Heredia, Costa Rica.

Castro E. and Salazar S. (2000) SEED. Documento preparado para la Empresa de Servicios Publicos de Heredia S.A., Guapiles, Costa Rica.

Catskill Corporation (2001) **http://www.cwconline.org/about/ab_hist.htm**.

Collins B.R. and Russell E.W.B. (Eds.) (1988) *Protecting the New Jersey Pinelands: A New Direction in Land Use Management*, Rutgers University Press: New Bruswick and London.

Dunlap J. (2003) Bringing back the terrapins. *What's up?* 9–11. **http://www.whatsupmag.com/oct03/terrapins.shtml**.

DWAF (1997) *White Paper on a National Water Policy for South Africa*, Department of Water Affairs and Forestry: Pretoria.

DWAF (1999a) *Establishment of a Pricing Strategy for Water Use Charges in Terms of Section 56(1) of the National Water Act, 1998*, Government Notice 1353, Department of Water Affairs and Forestry: Pretoria.

DWAF (1999b) *Water-Use Licensing: the Policy and Procedure for Licensing Stream Flow Reduction Activities*, Department of Water Affairs and Forestry: Pretoria.

Echavarria M. (2002a) Financing watershed conservation: the FONAG water fund in Quito, Ecuador. In *Selling Forest Environmental Services: Market-Based Mechanisms for Conservation and Development*, Pagiola S., Bishop J. and Landell-Mills N. (Eds.), Earthscan: London.

Echavarria M. (2002b) *Water User Associations in the Cauca Valley, Colombia: A Voluntary Mechanism to Promote Upstream-Downstream Cooperation in the Protection of Rural Watersheds*, Land-Water Linkages in Rural Watersheds Case Study Series Food and Agriculture Organization of the United Nations: Rome.

Echavarria M. (2003) *Algunas lecciones sobre la aplicación de pagos por la protección del agua con base en experiencias en Colombia y Ecuador*, Tercer Congreso Latinoamericana de Manejo de Cuencas Hidrográficas, Foro Regional sobre Sistemas de Pago por Servicios Ambientales: Arequipa.

Fallas J. (1996) *Cuantificación de la Intercepción en un Bosque Nuboso*, Mte. de los Olivos, Cuenca del Río Chiquito, Guanacaste, Costa Rica. CCT/CINPE/IIED, San José, Costa Rica.

Golder W., Lee D. and Whilden M. (2000) Terrapin conservation efforts. *Turtle and Tortoise Newsletter*, 2. **http://www.chelonian.org/ttn/archives/ttn2/**.

Gomi T., Sidle R.C. and Richardson J.S. (2002) Understanding processes and downstream linkages of headwater systems. *BioScience*, **52**, 905–916.

Hsieh J. (2003) Terrapin conservation in trouble due to state cuts. In *The Diamondback*, College Park. **http://www.diamondbackonline.com/News/Diamondback/archives/2003/10/01/news8.html**.

IWR (2000) *Existing Wetland Mitigation Bank Inventory*, **http://www.iwr.usace.army.mil/iwr/regulatory/banks.pdf**.

Kaimowitz D. (2001) *Useful Myths and Intractable Truths: The Politics of the Link between Forests and Water in Central America*, Working paper Center for International Forest Research (CIFOR): San Jose.

Kaimowitz D. and Ribot J.C. (2002) Services and infrastructure versus natural resource management: building a base for democratic decentralization, *Conference on Decentralization and the Environment*, Bellagio, 18–22 February 2002.

King D.M. (2003) Will nutrient credit trading ever work? An assessment of supply and demand problems and institutional obstacles. *ELR*, **33**, 10352–10368.

Koundouri P., Pashardes P., Swanson T.M. and Xepapadeas A. (2003) *Economics of Water Management in Developing Countries: Problems, Principles and Policies*, Edward Elgar: Cheltenham.

Landell-Mills N. and Porras I.T. (2002) *Silver Bullet of Fools' Gold? A Global Review of Markets for Forest Environmental Services and Their Impact on the Poor*, International Institute for Environment and Development: London.

Liebesman L.R. and Plott D.M. (1998) The emergence of private wetlands mitigation banking. *Natural Resources and Environment*, **13**, 341–371.

Lipton M. (1985) The Prisoners' Dilemma and Coase's theorem: a case for democracy in less developed countries? In *Economy and Democracy.*, Matthews R.C.O. (Ed.), St. Martin's Press: New York.

Meinzen-Dick R.S. and Bruns B.R. (2000) Negotiating water rights: introduction. In *Negotiating Water Rights*, Bruns B.R. and Meinzen-Dick R.S. (Eds.), Intermediate Technology Publications and the International Food Policy Research Institute: London.

Meinzen-Dick R.S. and Pradhan R. (2002) *Legal Pluralism and Dynamic Property Rights*. CAPRI Working Paper 22, CGIAR Systemwide Program on Collective Action and Property Rights, Washington DC.

National Research Council (2001) *Compensating for Wetland Losses under the Clean Water Act*, National Academy Press: Washington.

NC NPSMP (2004) *Tar Pamlico Nutrient Strategy*. **http://h2o.enr.state.nc.us/nps/tarpam.htm**.

New Jersey Pinelands Commission (2004) *A Summary of the New Jersey Pinelands Comprehensive Management Plan*. **http://www.state.nj.us/pinelands/cmp.htm**.

New York Department of Environmental Protection (2001) **http://www.ci.nyc.ny.us/html/dep**.

Nutrient Net (2004) *Program Summaries*, **<http://www.nutrientnet.org/prototype/html/program-summary.html>**.

O'Connor M. (2000) Pathways for environmental evaluation: a walk in the (Hanging) Gardens of Babylon. *Ecological Economics*, **34**, 175–194.

Ostrom E. (1990) *Governing the Commons. The Evolution of Institutions for Collective Action*, Cambridge University Press: Cambridge.

Ostrom E., Gardner R. and Walker J. (1994) *Rules, games, and common-pool resources*, University of Michigan Press, Ann Arbor.

Pagiola S. (2002) In *Selling Forest Environmental Services: Market-based Mechanisms for Conservation and Development* Pagiola S., Bishop J. and Landell-Mills N. (Eds.), Earthscan: London.

Pagiola S., Landell-Mills N. and Bishop J. (2002) Making Market-based mechanisms work for forests and people. In *Selling Forest Environmental Services: Market-based Mechanisms for Conservation and Development*, Pagiola S., Bishop J. and Landell-Mills N. (Eds.), Earthscan: London.

Perrot-Maître D. and Davis P. (2001) *Case Studies: Developing Markets for Water Services from Forests*, Forest Trends: Washington.

Porto M., Porto R.L. and Azevedo L.G.T. (1999) A participatory approach to watershed management: the Brazilian system. *Journal of the American Water Resources Association*, **35**, 675–684.

Rojas M. and Aylward B. (2002) *The Case of La Esperanza: A Small Private, Hydropower Producer and a Conservation NGO in Costa Rica*, Land and Water Linkages in Rural Watersheds Case Study Series FAO: Rome.

Rojas M. and Aylward B. (2003) *What are we Learning from Experiences with Markets for Environmental Services in Costa Rica? A Review and Critique of the Literature. Working Paper*, International Institute for Environment and Development: London.

Rosa H., Kandel S., Dimas L. and Méndez W.C.F.N.C.A.E. (2003) *Compensation for Environmental Services and Rural Communities: Lessons from the Americas and Key Issues for Strengthening Community Strategies*, PRISMA: San Salvador.

Rose C.M. (2002) Common property, regulatory property, and environmental protection: comparing community-based management and tradable environmental allowances. In *The Drama of the Commons*, Ostrom E., Dietz T., Dolšak N., Stern P., Stonich S. and Weber E.U. (Eds.), National Academy Press: Washington.

Saleth R.M. and Dinar A. (1999) *Water Challenge and Institutional Response: A Cross-Country Perspective*, Policy Research Working Paper 2045, The World Bank Development Research Group Rural Development and Rural Development Department: Washington.

Salmon Safe (2003) **www.salmonsafe.org**.

Salzman J. and Ruhl J.B. (2002) Paying to protect watershed services: Wetland banking in the United States. In *Selling Forest Environmental Services: Market-based Mechanisms for Conservation and Development.*, Pagiola S., Bishop J. and Landell-Mills N. (Eds.), Earthscan: London.

Sax J.L. (1993) Property rights and the economy of nature: understanding *Lucas v. South Carolina coastal council*. *Stanford Law Review*, **45**, 1433–1455.

Schreiner B. and Van Koppen B. (2000) *From Bucket to Basin: Poverty, Gender, and Integrated Water Management in South Africa*. Integrated Water Management in Water-Stressed River Basins in Developing Countries: Strategies for Poverty Alleviation and Agricultural Growth, Loskop Dam, South Africa, 16–21 October 2000.

State Forests of New South Wales (2001) *The War Against Salinity – Could Forests be the Answer?* **http://www.forest.nsw.gov.au/_navigation/active_frame.asp?bodypath=/bush/feb00/feature/page14.asp**.

State Forests of New South Wales (2004) *Environmental Services Scheme*, **http://www.forest.nsw.gov.au/env_services/ess**.

Sundstrom A. (2001) *Salinity Control Credits: a Comment*, **http://www.nccnsw.org.au/veg/reference/salt.credits.html**.

Swallow B.M., Garrity D.P. and van Noordwijk M. (2001) The Effects of Scales, Flows and Filters on Property Rights and Collective Action in Watershed Management. CAPRI Working Paper 16, IFPRI, CGIAR Systemwide Program on Collective Action and Property Rights, Washington, D.C.

Tietenberg T. (2002) The tradable permits approach to protecting the commons: what have we learned? In *The Drama of the Commons*, Ostrom E., Dietz T., Dolšak N., Stern P., Stonich S. and Weber E.U. (Eds.), National Academy Press: Washington.

Tognetti S.S., Mendoza G., Aylward B., Southgate D. and Garcia L. (2004) *A Knowledge and Assessment Guide to Support the Development of Payment Arrangements for Watershed Ecosystem Services (PWES)*. Prepared for the World Bank Environment Department with support from the Bank-Netherlands Watershed Partnership Program, Washington.

Tomich T.P., Thomas D.E. and van Noordwijk M. (2004) Environmental services and land use change in Southeast Asia: from recognition to regulation or reward? *Agriculture Ecosystems and Environment*, **104**, 1.

Trust for Public Land (2001) *Land & People: Source of Inspiration*, **http://www.tpl.org**.

UN FAO (2002) *Land-Water Linkages in Rural Watersheds Electronic Workshop, 18 September–27 October 2000*, Land and Water Bulletin 9, UN Food and Agriculture Organization: Rome.

United Nations (2003) *Water for People, Water for Life: UN World Water Development Report (WWDR)*, UNESCO Publishing, Paris.

USDA (2000) *Farm Service Agency Online: The Conservation Reserve Program*, **http://www.fsa.usda.gov/dafp/cepd/12logocv.htm**.

van Noordwijk M., Poulsen J. and Ericksen P. (2004) Quantifying off-site effects of land use change: filters, flows and fallacies. *Agriculture Ecosystems and Environment*, **104**, 1.

Walker A. (2003) Agricultural transformation and the politics of hydrology in Northern Thailand. *Development and change*, **34**, 941–964.

Wilkinson C.F. (1992) *Crossing the Next Meridian: Land, Water, and the Future of the West*, Island Press: Washington.

Ziegler A.D., Giambelluca T.W., Sutherland R.A., Nullet M.A., Yarnasarn S., Pinthong J. and Jaiaree S. (2004) *Agriculture Ecosystems and Environment*, **104**, 145–158.

194: Inter-Institutional Links in Land and Water Management

JAIME M AMEZAGA

Centre for Land Use and Water Resources Research, Institute for Research on the Environment and Sustainability, University of Newcastle, Newcastle upon Tyne, UK

This article reviews the institutional aspects of Integrated Land and Water Resources Management. It explores the definition of what constitutes an institution in the land and water sector and presents several approaches to institutional analysis applicable in this context. It examines land and water interinstitutional linkages at several levels: international, national, river basin, and upstream/downstream. For each of these levels, theoretical approaches and examples are discussed. While institutional constraints such as cross-sectoral coordination and the conflict between different land-based administrative structures are still common, there are an increasing number of instruments applicable to this problem at all levels. Ensuring basic human needs and rights associated with water and land and establishing institutional systems to benefit disadvantaged people and improve livelihoods remain the greatest challenges to integrated management.

FROM WATER TO LAND AND WATER MANAGEMENT

The twentieth century saw the evolution of increasingly sophisticated approaches to integrated water management (White, 1998). At the beginning of the century, the prevailing approach was single purpose, with water harnessed independently for power, irrigation, or storage for drinking water purposes. By the third decade of the century, engineering developments made possible the multipurpose development of water resources, with the emergence of dams able to generate electricity and store water for other aims, as well as the exploitation of deep wells and systems for the treatment of wastewater. By the sixth decade, questions about the full social and environmental implications of these increasingly complex water systems started to arise. The global institutions created after the Second World War began to pay explicit attention to integrated water management in a number of conferences and reports. A critical landmark was the UN Water Conference in Mar del Plata in 1977, which sought to formulate general principles for national and international agencies. Those principles were updated in an International Conference in Dublin in 1992, which shaped the water policies and programs adopted at the UN Conference on Environment and Development in Rio de Janeiro that same year, known as the *Agenda 21* (Sitarz, 1993). The Agenda 21 was an additive document with limited efforts devoted to integration. While there was considerable attention paid to water, this was sectorized, and the stress on integration was not accompanied by indications on how to achieve it. The land-use chapters were almost completely "dry" (Falkenmark, 2001). There was neither indication of the vertical presence of water in the saturated and root zones, nor of the implications of land use for water management beyond irrigation. After Rio, it became increasingly clear that there was a need to develop the ability to integrate land/water/ecosystems management and properly incorporate land use; "a *land use decision is also a water decision*" (Falkenmark *et al.*, 1999; Newson, 1997; Calder, 1999; Calder, 2005). This integrated approach had to be translated into policy and strong institutions able to handle the potential conflicts between incompatible land and water uses.

The awareness of the land and water link is explicit in the definition of integrated water resources management (IWRM) popularized by the Global Water Partnership after the second World Water Forum in The Hague (2000):

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

"IWRM is a process which promotes the coordinated development and management of water, land and related resources, in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems" (GWP TAC, 2000).

This definition emphasizes that IWRM is about coordination (Jønch-Clausen and Fugl, 2001); it is the integrative handle that can lead from fragmented subsectoral to cross-sectoral water management. A key concept to be grasped is that IWRM is not a spontaneous phenomenon, but rather an institutionalized process facilitated or constrained by a particular constellation of policies and actors working in a certain time and space. It requires an enabling environment of polices and legislation from the international to the local level. There must be an institutional framework with clear demarcation of responsibilities between actors, adequate coordination mechanisms, the filling of jurisdictional gaps, and capacity to action.

The different uses of water for people, food, nature, and industry are intrinsically linked to land use. The importance of this link has seen the emergence of the concept of Integrated Land and Water Resources Management (ILWRM) as the ideal for coordination (*see* **Chapter 185, Integrated Land and Water Resources Management, Volume 5**). This article reviews how the current institutional frameworks for water management relate to this ideal.

INSTITUTIONAL FRAMEWORKS FOR WATER MANAGEMENT

Defining Institutions

Understanding institutional frameworks requires definition of what is meant by the term institutions. Ostrom (1999), one of the originators of the Institutional Analysis and Development (IAD) framework, points out that this is one of the main difficulties involved in studying institutions. The term refers to many different types of entities, including both organizations and the rules used to structure patterns of interaction within and across organizations.

The International Water Management Institute (IWMI) and the Rural Development Department of the World Bank have developed concepts and methods for institutional analysis specifically devoted for the water sector. In an IWMI study on a framework for institutional analysis for water resources management in a river-basin context, Bandaroga (2000) reviews the concepts of institution and organization. Institutional economics has defined institutions as the rules of the game in a society, or more formally, the humanly devised constraints that shape human action (North, 1990). Similarly, the IAD framework sees institutions as the shared concepts used by humans in repetitive situations organized by rules, norms, and strategies (Ostrom, 1999). Meanwhile, organizations are groups of individuals with defined roles and bound by some common purpose and some rules and procedures to achieve objectives (Bandaroga, 2000). Both concepts are linked in at least two ways. Firstly, how organizations come into existence and evolve is influenced by the surrounding institutional framework. So, for example, irrigation departments are created following the enunciation of water-related policy, and organizations such as the World Bank are the product of international agreements. The second route is that established organizations represent a set of norms and behaviors that persist because they are valued and accepted as useful. In this case, organizations can also be conceptualized as institutions or as "institutionalized" organizations. This second definition applies to river-basin organizations, water companies, or ministerial departments.

In summary, water institutions are a combination of:

- policies and objectives,
- laws, rules, and regulations,
- organizations, their bylaws and core values,
- operational plans and procedures,
- incentive mechanisms,
- accountability mechanisms, and
- norms, traditions, practices and customs.

Considering both the core meaning of the term "institutions", and the emphasis in the popular usage of the term associated with "organizations", Bandaroga (2000) proposes a broad interpretation of the term in the context of river-basin management, which sees the institutional framework for water resources as consisting of established rules, norms, practices, and organizations that provide a structure to human actions related to water management. Notably, the established organizations are to be considered here as a subset of institutions. For practical purposes, the overall institutional framework is considered in three broad categories: policies, laws, and administration, all of which are related in some way to water resources management in a river-basin context.

Institutional Analysis

Bandaroga's interpretation covers the three important elements in the institutional framework, namely, policies, laws, and organizations, identified by Saleth and Dinar in their comprehensive evaluation of water institutions and water sector performance (Saleth and Dinar, 1999; Saleth and Dinar, 2000). These authors have proposed a research paradigm for the strategic analysis of water institutions based on two complementary analytical and theoretical components, the "Institutional Decomposition and Analysis" (IDA) framework, and the "Institutional Transaction Cost" theory (Saleth and Dinar, 2004; Saleth, 2004). The theoretical basis for this paradigm is institutional economics and has analytical similarities to the IAD framework. It

sees institutions as entities defined by a configuration of legal, policy and organizational rules, conventions and practices that are structurally linked, and operationally embedded with a well-specific environment. IAD distinguishes the *institutional structure* (or governance structure) from its *institutional environment* (or governance framework). While the institutional environment is characterized by the overall physical, cultural, historic, socioeconomic and political milieu of a country or region, the institutional structure is defined by the interactive effects of the legal, policy, and organizational or administrative components and their constituent aspects. The institutional structure is embedded within the institutional environment and both elements influence each other.

The water institutional environment is characterized, amongst others, by the water resources condition and critically by other related sectors linked to land use such as agriculture, environment, forestry, and urban development. The water institutional structure can be broadly decomposed in three interrelated components: water law, water policy, and water administration. These *institutional components* cover not only the formal and macrolevel arrangements but also the informal and microlevel arrangements. Examples of the latter are those reflected in local customs, conventions, and informal contracts. The formal and informal institutional components can also be decomposed further to highlight some *institutional aspects*. Water law, for example, can be separated into intergovernmental issues, water rights, and accountability provisions. Water policy can distinguish between policy principles, project-selection criteria, pricing and cost recovery, and user and private sector participations. In water administration, one can highlight organizational structures, roles of the different layers of government, financing and management, regulatory mechanisms, and conflict resolution arrangements.

In general, the IDA framework falls within the rational choice paradigm based on methodological individualism and the concept of self-interested behavior. Mollinga (2001) has reported critical analyses of the preponderance of this type of approach in the land and water sector. The main critiques reported focus on the deficiencies of the concept of human agency employed in rational choice. There are other approaches that start from a broader concept of human agency, contextualize and emphasize the importance of history, and examine the social construction of rationality, interests, and identity. Cultural Theory (Douglas and Wildavsky, 1982) provides one counterpoint to rational choice theory that has been applied in land and water resources studies (*see* **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**). Cultural theory is concerned with the socioinstitutional styles, such as hierarchy, egalitarianism, individualism and fatalism, which define people's, groups' and organizations' behavior and strategic choices. Mollinga (2001) quotes a study by Gyawali in 1998, showing how, in the debates around embankment works in the Ganges plain, the government department concerned worked from hierarchical premises, the social and environmental activist from egalitarian principles, while there are fatalist masses in the background and individualistic contractors involved.

LAND AND WATER INSTITUTIONAL LINKAGES

Interinstitutional Linkages

One fundamental feature of institutional analysis is that the decomposed institutional components and aspects should be treated independently only for analytical convenience. In reality, all these elements are functionally nested and interlinked both within and across institutional components. So, water policy can be seen as the political translation of water law, and water law as the legal representation of policies, while water administration is influenced by the implementation of both law and policy. Within one component such as water law, water rights, conflict resolution, and accountability are intrinsically related. Institutions are also path-dependent. History matters, and the present status and future direction of equilibrium cannot be divorced for its earlier course (Saleth, 2004). In particular, there is a history of multisectoral relationships and the different ways in which they have been shaped by the institutional environment. These are very important concepts for understanding the institutional links in land and water management.

Interinstitutional linkages are also affected by three clusters of factors: problems of fit, problems of interplay, and problems of scale (Young, 1999). *Problems of fit* are related to the fact that the effectiveness of social institutions is a function of the match between the characteristics of institutions themselves and the characteristics of the biogeophysical systems with which they interact. This problem arises, for instance, when land and water institutions do not cover the entire geographical range of land and water units. The *problem of interplay* refers to the fact that the effectiveness of institutions often depends not only on their own features but also on their interactions with other institutions. Young (1999) recommends that in order to achieve integrated land and water management, land and water institutions must interact with other institutions dealing with activities that impact on these resources. Finally, *problems of scale* refer to the transferability of both empirical generalizations and causal inferences from one level to another in the dimensions of space and time. What works on institutional arrangements at the microlevel may not work at the macrolevel and vice versa. This is a common problem in the generalization of participatory approaches from small communities to whole river basins.

The strength of integration within land and water institutions can be formalized in terms of the strength of

institutional linkages (Saleth and Dinar, 2003), and the way in which the problems of fit, interplay, and scale are overcome by interinstitutional coordination. For instance, the way water sources as well as their relationships with land and environmental resources are treated within water law can influence water policy aspects such as priority settings for water uses and project-selection criteria. Thus, if the water law recognizes the ecological linkages between water and other resources, it is more likely to give higher priority to environmental imperatives when allocating water (see the example of the concept of the reserve in South Africa in **Chapter 186, Water and Forests, Volume 5**). The allocation of water has to happen at the catchment level and take into account all uses and integrate the logics of different scales from national economic interests to local livelihood priorities.

International Level

At an international level, the land and water institutional framework is characterized by a number of global organizations, mainly UN- and World Bank related, including, on the research side, the international and regional organizations known as the *Consultative Group of International Agricultural Research* (CGIAR) to which IWMI belongs. There are also an increasing number of binding and nonbinding instruments such as treaties, protocols, and agreements (Hannam, 2003). The international conventions form a fragmented framework, with four single-purpose international conventions directly related to crises affecting land and water resources: the United Nations Convention on Biological Diversity, the United Nations Framework Convention of Climate Change, the United Nations Convention to Combat Desertification, and the United Nations Convention on the Law of the Sea. On balance, the attention is favored more toward land-related issues than water-related issues. No freshwater related convention has entered into force that outlines the behavior of parties for balancing uses of water within their sovereign jurisdiction or for linking land management with water management (Duda, 2003).

One particularly important international institution is the Global Environmental Facility (GEF). The GEF was established in 1991 as a pilot multilateral financial mechanism to test new approaches and innovative ways to respond to global environmental challenges in the four focal areas of climate change, biodiversity conservation, ozone depletion, and international waters (Duda, 2003). The GEF builds in partnership among the United Nations Development Program (UNDP), the United Nations Environment Program (UNEP), and the World Bank. The Operational Strategy of the international waters' focal area recommends several tools that take an integrated approach to land and water resources. These are the Transboundary Diagnostic Analysis (TDA), Strategic Action Programs (SAP), national interministerial committees, and local demonstration activities. The strategic approach is to start with the entire multicountry basin, move into single country tributary basins for demonstration activities, and then assist the nations in undertaking institutional reforms to integrate land and water resources management, often in concert with investment lending from multilateral finance institutions or donor organizations. An important lesson from the GEF experience is that science-based and joint fact-finding among participant nations can serve as an important catalytic tool for developing political buy-in. This may be enhanced if official interministerial committees of national and subnational governments are involved in the joint institutions, instead of developing a supranational international organization to do the work for the countries. The GEF has stressed the development of institutional structures for land and water-related policy reforms including water rights, water allocation systems, land tenure reforms, pricing policy reforms, and creation of basin management organizations (Duda, 2003).

The development of institutional frameworks for the management of transboundary water resources has received considerable attention in the international literature. For instance, Kliot *et al.* (2001) present a comparative analysis of institutions in nine international river basins divided according to their level of cooperation and commitment in three categories: the highly committed (Colorado, Niger, Rio Grande and Senegal); the least cooperative (Ganges-Brahmaputra and Indus); and the intermediate level of cooperation (Danube, Elbe, and Mekong). Very few of the examined institutions corresponded to the theoretically ideal model of institutions for management of water resources, namely, a basin-wide multipurpose institution that treats the whole basin as one unit and integrates all riparians in an equitable manner (see section "River Basin Level"). The institutions that worked better in the examples above were those able to contend with overt conflict over the use of resources.

National Level

Looking at the national level, Hannam (2003) has developed a methodology to identify and evaluate the features of the legal and other elements of the institutional framework for the management of water and land, and applied it to four countries in the Asian region (Bangladesh, Laos, Philippines, and China). In this methodology, the capacity of a legal/institutional framework is determined by the number and type of essential legal and institutional elements present in a legal instrument in a format that enables the key issues of the sustainable use of water and land to be identified, and with the legal, administrative, and technical capability in the particular instrument to take some form of effective action. Most primary water and land management issues are multifactorial. They include a sociological, a legal, and a technical component. Therefore, generally, more than one piece of environmental legislation will be

needed for a particular nation to manage effectively individual issues.

Essential elements integrating land and water that should be present are shown in Table 1.

The general framework includes international treaties and other agreements, and national legislation. In summary, the following areas of national environmental law are applicable (see Table 2).

The methodology proposed by Hannam identifies the primary water and land management issues in a region, the operational environment, and the policy instruments at several levels. These elements are then analyzed to see if they had the standard essential elements integrating land and water in order to determine their capacity to achieve sustainable land and water management.

A critical issue at the level of national integration is the problem of interplay between policy instruments in different sectors. For instance, Wilchens (2003) points out that most of the contributions of water resources professionals to policy discussions regarding agricultural productivity are focused on increasing the output generated with limited water supplies, either at the field level or throughout a river basin. However, policy discussions can be enhanced by placing greater emphasis on the roles of nonwater inputs and resource constraints in farm-level production and marketing decisions. These could include policies that modify farm-level input and output prices directly; international trade policies; and policies that modify key institutions, such as land tenure and the sources of investment funds. Most importantly, agricultural polices have traditionally driven the need for large dams and water transfers. The construction of these expensive water systems has not always taken into account the hydrological limitations, or the impacts on environment and society (WCD, 2000; Getches, 2003). Sometimes cross-sectoral linkages are prevented by the set of beliefs prevalent in the institutional environment. One sector where policies have traditionally ignored hydrological evidence of negative impacts is forestry. Deeply engrained beliefs on the positive links of forest and water are still evident in most of the forestry policies in the world (e.g. Calder, 1999; Calder *et al.*, 2004). Saberwall (1997) has studied the historical origin of these myths and their influence on current Indian forestry policy. Another sector that has been particularly reluctant to regulate its impact on the water environment is mining. In spite of the well-documented links between mining and water (Younger *et al.*, 2002), the economic and strategic importance of the mining industry has made governments very reluctant to intervene in this sector. The global mining industry is now fully aware of the need to implement water strategies, but their application in developing countries is still deficient (MMSD, 2002). One limitation is the lack of awareness of policy-makers of the links between this particular industrial sector and water. Even in an advanced environmental regulatory system like the European Union (EU), it has taken a succession of mining disasters with high public impact (Aznalcóllar in 1998 and Baia Mare in 2000) to enact specific environmental regulation for the extractive industry. However, the chosen policy instrument has been waste legislation, with a limited coverage of mine water issues (Amezaga and Kroll, 2004). In conservation policy, the management of wetlands is particularly dependent on land and water integration. Most of the conservation approaches are limited to specific protected areas with a high local level of protection; but wetlands are connected physically and socially with processes occurring in a much wider territory. Changes on water quantity and quality generated by urban, agricultural, and industrial development can be critical for the survival of protected areas (Amezaga and Santamaría, 2000). Particularly relevant in the problem of interplay is the policy gap between water management and land use and development planning (Newson, 1997). Urban development, use of agricultural land and forestry, location of industry, and demarcation of protected areas are all determined by the planning system. Frequently, planning

Table 1 Essential elements integrating land and water according to Hannam (2003)

General intent	Education
Jurisdiction	Research and investigation
Responsibility	Community participation
Goals and objectives	Water and land planning
Definitions	Water and land management
Duty of care	Financing
Hierarchy of responsibility	Enforcement
Institutional (administration)	Dispute resolution
Policy	

Table 2 Areas of national law for the management of land and water (Hannam, 2003)

Constitutional law	Mining law
Environmental planning law	Indigenous people's and customary law
Pollution control law	Agricultural land-use law
Forestry law	Agricultural reform
Plantation and reafforestation law	Protected area law
Soil conservation law	Protection of the marine environment
Water conservation law	Land administration and tenure law
Water use law	Law for women's rights, poverty alleviation, and financial management
Environmental protection law	Criminal law and law courts

is completely independent of water management and does not take into account the impact of land-use decisions on the water environment.

In general, and most importantly in developing countries, there are two important issues that need to be developed to improve the integration of land and water management frameworks (Hannam, 2003):

- Basic human needs and rights associated with water and land.
- Legislative/institutional systems to benefit disadvantaged people (the minorities, and poverty-stricken) and improve livelihoods.

River-basin Level

The most important planning and management unit for the integration of land and water is the river basin (also referred to with similar meanings as watershed or catchment) (Newson, 1997; Falkenmark *et al.*, 1999). The river basin represents a hydrological unit that integrates land and water relationships. The challenge is to achieve a similar level of integration in the social, economic, legal, and political relationships affecting land and water. These relationships are of two types: intrabasin, including upstream-downstream relations and conflicts between water users; and the interbasin relations which represent local, regional, and national administrative levels. Ideally, planning and decision making in a river basin or management unit should be placed in an institutional system in which all components are related. However, the mismatch between river basin and land-use administrative units presents a difficult challenge.

Jaspers (2003) has reviewed the institutional arrangements necessary for integrated river basin management with a comparative assessment of the situation in Zimbabwe, South Africa, Tanzania, Turkey, Indonesia, France, and The Netherlands. The key basic institutional elements required are:

- the functioning of a platform for stakeholders involved in decision making,
- water resources management on hydrological boundaries,
- an organizational setup in river basin and subbasin authorities with their respective bylaws to incorporate decision making at the lowest appropriate level,
- a planning system oriented at the production of integrated river-basin plans,
- the introduction of a system of water pricing and cost recovery.

Tasks and competencies of river-basin organizations may differ substantially from country to country; and the distribution of tasks between basin and subbasin organizations will represent a variety of strategic and operational functions. A key function of river-basin organizations is the development of plans setting out a common vision for the river to influence private and public sector initiatives. These river-basin plans should include specific evaluations of land-use options within the river basin.

The Murray–Darling Basin in Australia is frequently presented as a paradigmatic example of a river-basin management institution (Falkenmark *et al.*, 1999) with strong coordination of governments and communities. The basin is an area of 1 000 000 km^2 with 3700 km of major rivers traversing four independent states. Each state is responsible for the management of its environmental resources. The basin is also Australia's main agricultural area, with 75% of the country's irrigated land. From the mid-1960s, governments and communities in the downstream segment were concerned about deteriorating water quality because of salinity. Meanwhile, upstream communities were calling for action to control shallow water tables that were causing waterlogging and salination. Changes in the institutional setup with the establishment of the Murray–Darling Basin Ministerial Council to promote effective planning and management for the equitable, efficient and sustainable use of the water, land, and environmental resources allowed the development of integrated solutions to these problems. Several institutional lessons have been drawn from this experience:

- the need for government leadership involving policymakers at the higher level and trade-offs of sovereignty rights;
- the need for community leadership committed to developing and implementing action plans;
- the need for technical knowledge to provide a strong knowledge base of the causes, effects, and impacts of the various land and water management options;
- the use of market instruments to ensure that the cost of resource exploitation is included in day-to-day decision making.

The Water Framework Directive (WFD) of the EU is a remarkable attempt to introduce integrated river-basin planning and management at a continental level (EC, 2000). All countries of the EU had to adopt this Directive in their national legislation by 2003 and create river-basin administrations, which will produce river-basin plans. The river-basin plans will have to take into account all the impacts and pressures from human activities on the water environment and develop a program of measures to improve its condition. The objective is to achieve good surface and groundwater status in all European river basins by 2015. But the WFD is also an excellent example of the difficulties to integrate land and water management. A good deal of the necessary measures in a river-basin plan, and in particular, those related to flooding and diffusion control, are intrinsically linked to land use. However, the Directive does not give direct powers to intervene on land uses. The

Treaty of the EU, on which all European law is based, distinguishes clearly between legislation dealing with water quality and legislation dealing with water quantity and land use (Krämmer, 1998). Water quantity and land use are considered Member State issues and can only be legislated at the EU level by unanimous decisions, which is why the EU has only adopted guidance documents on land use planning and not binding legislation. So although the WFD assumes the need for integrated land and water management, it is only the Member States that can make it legally binding in their own legislation.

Institutional inertia, partly explained by the high transactional cost of institutional reform and partly by a reluctance to tie development planning to compulsory environmental objectives, has meant that most of the EU countries have not yet introduced this legally binding link in their national legislations (2005). An interesting example is the United Kingdom, where England and Scotland have different legal systems. While England has transposed the WFD with minimal legislation and only compulsory consultation between water and land use planners, Scotland has approved primary legislation making the WFD plans legally binding for planners. Germany (Moss, 2003) has also followed the path of minimizing institutional reforms. Here the challenge is even greater because water management in Germany has traditionally followed political boundaries and not catchment boundaries. So river-basin administration is a relatively weak institution compared with local and State governments within the German federal system. By contrast, The Netherlands provides a good example of integration between environmental planning and water resources planning (Jaspers, 2003). The Netherlands is a very developed, small country with a high density of population. In order to deal with the intensive pressures on land use, the country has developed a sophisticated planning system where separate plans allocate guidelines or tasks to one another, and every plan indicates how issues earmarked by the other plan are dealt with. Every four years plans are revised in alternating sequence.

The usefulness of the river-basin administration in developing countries has been questioned because of the high transaction cost of institutional building in those conditions (Shah, 2005). However, more and more countries are trying to implement it. This is partly due to donor's pressure, who see it as best practice. Thus, South Africa has introduced via their progressive National Water Act (1998), the creation of Catchment Management Agencies for the management of water resources (Thompson *et al.*, 2001). India has also seen the emergence of some weak river-basin organizations. The Indian National Water Policy (1987) states that water resource planning should be done with reference to a hydrological unit such as a basin or a subbasin. However, the effective implementation of this policy is hampered by the fragmentation of the water sector due to constitutional constraints and the lack of administrative capacity. Iyer (2003) argues that the idea of basin planning and management, which seems eminently sound, contains within itself the seeds of centralization and gigantism. Subject to that caution, it is still necessary, and indeed logical, to take a comprehensive view of a river system as a whole.

Upstream-downstream Linkages

An integral subset of linkages within a river basin is the upstream-downstream relationship. In the year 2000, the Food and Agriculture Organization (FAO) of the United Nations Land and Water Development Division involved a number of international experts in the definition of land and water linkages in rural watersheds through an electronic conference. In a discussion paper prepared for the 2000 FAO conference, Kiersch (FAO, 2002) reviewed the instruments and mechanisms available for upstream-downstream linkages:

- regulatory instruments (command and control measures) widely used in developed countries to protect water resources from agricultural land-use practices;
- economic instruments to distribute benefits and costs resulting from land use impacts on water resources, including direct or indirect subsidies, taxes, and transferable property or use rights for land, water, and emissions;
- education and awareness building to encourage farmers to switch to less polluting practices, usually coupled with an incentive program to reduce pollution risks, and to improve the economic performance of the farm;
- market support, improving access of upstream farmers to downstream markets within the framework of a watershed agreement;
- organizational development for a successful implementation of instruments establishing linkages between upstream land users and downstream water users. Organizations can provide a forum of exchange between upstream and downstream stakeholders and a forum to consolidate the interests and opinions of scattered groups of users such as upstream farmers;
- participatory approaches. Usually participatory watershed planning and management projects focus on the community level and encompass only very small land units.

Constraints for the implementation of benefit-sharing mechanisms range from the need to find a compromise among conflicting interests over the distribution of cost and benefits to institutional challenges and up-front costs of engaging stakeholders in initial planning stages. Constraints include weak or nonexistent property rights that can provide some assurance to the people that they can reap the benefits, poverty, and a lack of perception that there is a

problem to begin with. At the most general level, there is often a conflict between objectives of sustaining livelihoods in the short term, and protection of resources. For benefit-sharing arrangements to be successful, stakeholders must at least have a common understanding and agreement about the nature of the expected impact, the approximate magnitude of cost and benefits, and the areas of uncertainty (*see* **Chapter 193, Markets for Watershed Services, Volume 5**).

A remarkable example of successful upstream/downstream collaboration is the 1997 Memorandum of Agreement (MOA), which protects the water supply of New York in the United States (Pires, 2004). This document outlining a watershed management plan was signed by the City and State of NewYork, the EPA, 73 local municipalities and eight counties in the watersheds, and five environmental organizations. The MOA contains both structural and nonstructural strategies for water quality protection. Some of the nonstructural mechanisms are directly related to land use policy. The most important ones are land acquisition, conservation easements, setbacks and buffer zones, and land trusts. Perhaps this is a best practice example of the way forward for ILWRM.

CONCLUSIONS

The institutional dimension is critical for the implementation of ILWRM approaches. Institutional development to support these approaches must take into account both institutional environments and institutional structures. While institutional constraints such as cross-sectoral coordination and the conflict between different land-based administrative structures are still common, there are an increasing number of instruments applicable to this problem at the global, regional, national, and river-basin level. Challenges remain; primary to these are ensuring basic human needs and rights associated with water and land, and establishing institutional systems to benefit disadvantaged people and improve livelihoods.

REFERENCES

Amezaga J.M. and Kroll A. (2004) European Union policies and mine water management. *Mine Water and the Environment*, **4**, 162–164.

Amezaga J.M. and Santamaría L. (2000) Wetland connectedness and policy fragmentation: steps towards a sustainable european wetland policy. *Physics and Chemistry of the Earth (B)*, **25**, 635–640.

Bandaroga D.J. (2000) *A Framework for Institutional Analysis for Water Resources Management in a River Basin Context*, International Water Management Institute: Colombo, Working Paper 5.

Calder I.R. (1999) *The Blue Revolution. Land Use and Integrated Water Resources Management*, Earthscan: London.

Calder I.R. (2005) *The Blue Revolution II*, Earthscan: London.

Calder I., Amezaga J., Bosch J., Fuller L., Gallop K., Gosain A., Hope R., Jewitt G., Miranda M., Porras I. and (2004) Forest and water policies – The need to reconcile public and science perceptions, *Geologica Acta*, **2**, 157–166.

Douglas M. and Wildavsky A. (1982) *Risk and Culture*, University of California Press: Berkeley.

Duda A.M. (2003) Integrated management of land and water resources based on a collective approach to fragmented international conventions. *Philosophical transactions of the Royal Society of London. Series B*, **358**, 2051–2062.

European Community (EC) (2000) Directive 2000/60/EC of the European parliament and of the council of 23 October 2000 establishing a framework for community action in the field of water policy, *Official Journal of the European Communities*, **L327**, 1–72.

Falkenmark M. (2001) The greatest water problem: the inability to link environmental security, water security and food security. *Water Resources Development*, **17**, 539–554.

Falkenmark M., Andersson L., Castensson R., Sundblad K., Batchelor C., Gardiner J., Lyle C., Peters N., Pettersen B., Quinn P. (1999) *Water- a Reflection of Land us. Options for Counteracting Land and Water Mismanagement*, Swedish Natural Science Research Council: Stockholm.

FAO (2002) *Land-water Linkages in Rural Watersheds. Proceedings of the Electronic Workshop Organized by the FAO Land and Water Development Division 18 September-27 October 2000*, FAO Land and Water Bulletin 9, FAO: Rome.

Getches D.H. (2003) Spain's Ebro River transfers: test case for water policy in the European Union. *Water Resources Development*, **19**, 501–512.

Global Water Partnership Technical Advisory Committee (GWP TAC) (2000) *Integrated Water Resources Management*, TAC Background Paper No. 4. GWP: Stockholm.

Hannam I. (2003) *A Method to Identify and Evaluate the Legal and Institutional Framework for the Management of Water for the Management of Water and Land in Asia: The Outcome of a Study in Southeast Asia and the People's Republic of China*, Research Report 73, International Water Management Institute: Colombo.

Iyer R.R. (2003) *Water- Perspectives, Issues, Concerns*, SAGE: New Delhi.

Jaspers F.G.W. (2003) Institutional arrangements for integrated river basin management. *Water Policy*, **5**, 77–90.

Jønch-Clausen T. and Fugl J. (2001) Firming up the conceptual basis for integrated water resources management. *Water Resources Development*, **17**, 501–510.

Kliot N., Shmueli D. and Shamir U. (2001) Development of institutional frameworks for the management of transboundary water resources. *International Journal of Global Environmental Issues*, **1**, 306–328.

Krämmer L. (1998) *E.C. Treaty and Environmental Law*, *Third Edition*, Sweet & Maxwell: London.

MMSD (2002) *Breaking New Ground: The Report of the Mining, Minerals and Sustainable Development Project*, Earthscan: London.

Mollinga P.P. (2001) Water and politics: levels, rational choice and South Indian canal irrigation. *Futures*, **33**, 733–752.

Moss T. (2003) The governance of land use in river basins: prospects for overcoming problems of institutional interplay with the EU Water Framework Directive. *Land Use Policy*, **21**, 85–94.

Newson M.D. (1997) *Land, Water and Development. Sustainable Management of River Basin Systems*, *Second Edition*, Routledge: London.

North D.C. (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge University Press: NewYork.

Ostrom E. (1999) Institutional rational choice. An assessment of the institutional analysis and development framework. In *Theories of the Policy Process*, Sabatier P.A. (Ed.), Westview Press: Boulder.

Pires M. (2004) Watershed protection for a world city: the case of New York. *Land Use Policy*, **21**, 161–175.

Saberwall V.K. (1997) Science and the desiccationist discourse of the 20th century. *Environmental History*, **3**, 309–343.

Saleth R.M. (2004) Strategic Analysis of Water Institutions in India: Applications of a New Research Paradigm, Research Report 79, International Water Management Institute: Colombo.

Saleth R.M. and Dinar A. (1999) *Evaluating Water Institutions and Water Sector Performance*, World Bank Technical Paper No. 447, World Bank: Washington.

Saleth R.M. and Dinar A. (2000) Institutional changes in global water sector: trends, patterns, and implications. *Water Policy*, **2**, 175–199.

Saleth R.M. and Dinar A. (2003) *Water Institutional Reforms in Developing Countries: Insights, Evidences and Case Studies*, Working Papers, Initiative for Police Dialogue, Columbia University.

Saleth R.M. and Dinar A. (2004) *The Institutional Economics of Water: A cross-country Analysis of Institutions and Performance*, Edward Elgar: Cheltenham.

Sitarz D. (Ed.) (1993) *Agenda 21: the Earth Summit Strategy to Save Our Planet*, Westview Press: Boulder.

Shah T. (2005) The new institutional economics of India's water policy, *Presented at the International Workshop on 'African Water Laws: Plural Legislative Frameworks for Rural Water Management in Africa'*, Gauteng, 26–28 January 2005.

Thompson H., Stimie C.M., Ritchers E. and Perret S. (2001) *Policies, Legislation and Organizations Related to Water in South Africa, with Special Reference to the Olifants River Basin*, Working Paper 18, International Water Management Institute: Colombo.

WCD (2000) *Dams and Development: A New Framework for Decision-Making- The Report of the World Commission on Dams*, Earthscan: London.

White G.F. (1998) Reflections on the 50-year international search for integrated water management. *Water Policy*, **1**, 21–27.

Wilchens D. (2003) Enhancing water policy discussions by including analysis of non- water inputs and farm-level constraints. *Agricultural Water Management*, **62**, 93–103.

Young O. (1999) *Institutional Dimensions of Global Environmental Change- Science Plan*, IHDP Report No. 9, IHDP.

Younger P.L., Banwart S.A. and Hedin R.S. (2002) *Mine Water: Hydrology, Pollution, Remediation*, Kluwer Academic Publishers: Dordrecht.

PART 17

Climate Change

195: Acceleration of the Global Hydrologic Cycle

JOHN ROADS[1], ROBERT OGLESBY[2], FORREST HOFFMAN[3] AND FRANKLIN ROBERTSON[2]

[1]*Scripps Institution of Oceanography, UCSD La Jolla, CA, US*
[2]*NASA/Marshall Space Flight Center, Huntsville, AL, US*
[3]*Oak Ridge National Laboratory, Oak Ridge, TN, US*

The global hydrologic cycle, which can be conceptually described as evaporation of water vapor from the ocean, transport of water vapor by atmospheric winds to land regions, condensation, and precipitation of atmospheric water back to the surface, and subsequent transport, by streams, of this water back to the ocean, may intensify or accelerate in the future, as the planet warms owing to an increasing greenhouse resulting from increasing emissions of anthropogenic gases. As in many previous studies, increased hydrologic fluxes are found in a 100-year integration of the NCAR Coupled Climate System Model when forced with increasing CO_2 levels. That is, as the earth warms, the evaporation increases, which not only increases the precipitation rate, but also the subsequent moisture convergence to land regions. However, if this increase in the cycling rate by various processes is measured with respect to the changing atmospheric and surface water reservoirs, then the atmospheric hydrologic cycle may appear to be actually deaccelerating since the atmospheric reservoir in this model is influenced more by the temperature changes than the actual transformation processes are. By contrast the land surface water reservoirs are less affected and thus there does appear to be an accelerated land surface hydrologic cycle, especially at high latitudes, and a coupled land-atmosphere acceleration, at least over land regions. There are also important regional differences.

INTRODUCTION

From various modeling studies as well as increasing observational evidence, we know that globally rising CO_2 and other trace gas amounts, resulting from globally increasing population and economic growth, are expected to increase the atmospheric "greenhouse" effect and thus cause an increase in the global surface temperature. The atmospheric "greenhouse" effect, which is necessary for sustaining life on the earth, results from these anthropogenic gases (and water vapor) being relatively transparent to solar radiation but relatively opaque to outgoing planetary thermal radiation; in order to maintain a radiative balance between outgoing planetary radiation and incoming radiation, the surface temperature is therefore much higher than the effective outgoing planetary emission temperature. As the anthropogenic greenhouse gases increase, the atmospheric opacity to outgoing radiation increases still further, thus requiring an increased lower atmosphere temperature as part of the planetary radiative equilibrium. Along with this lower atmosphere temperature increase, modeling studies have also suggested that global evaporation and precipitation will increase.

The water vapor capacity of the atmosphere increases exponentially with increasing temperature, according to the well-known Clausius–Clapeyron relationship. Since evaporation depends on the difference of water vapor at the surface from the relatively drier air aloft, then, other things being equal, higher temperatures tend to result in increased water vapor differences between the atmosphere and surface, which then results in increased evaporation. Indeed, it is generally true that global evaporation reaches a maximum over the ocean where the sea-surface temperature is highest. It is also generally true that evaporation is much greater during the summer than the winter and much greater during the daytime than the nighttime.

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

Evaporated water vapor is transported by atmospheric winds to other regions, leading to increased convergence of this water vapor mass flux in various precipitating regions. Since the atmosphere can only hold so much water vapor, depending upon the local temperature and pressure, excess water vapor condenses into cloud particles, which grow into raindrops through nucleation, condensation, coalescence, and finally by accretion into large particles, which fall as precipitation back to the surface. When the precipitation falls over land, it wets the soil (immediately or later depending on whether it first falls as rain or snow) and ultimately this surface water flows into streams, which discharge back to the ocean. In a climatological steady state, there is as much atmospheric water converged over land regions as is discharged back to the ocean.

An evaporation increase in a world with increased greenhouse gases will, of course, not happen everywhere. For example, current snowmelt-dominated regions tend to release water in great abundance as the seasons transition from winter to summer. If snow is no longer present owing to higher wintertime temperatures, or the summertime surface has desiccated, then evaporation (as well as streamflow and perhaps precipitation) may decrease in these regions as the global temperature increases. The surface water availability for a specific region is the complicating factor. Over the subtropical deserts the surface water availability is minimal and little evaporation occurs in these regions despite them being some of the hottest areas of the planet. In fact, just as one of the main heating mechanisms for the atmosphere comes from the latent heat released when vapor changes into cloud water, so too is evaporation one of the principle mechanisms for cooling the surface, through the latent heat needed to change liquid and solid water to its gaseous vapor phase.

Part of the postulated hydrologic cycle change thus depends upon how the local surface water reservoir is changed with increased temperature. In fact, global change modeling studies indicate there will be regions and times of increased drought as well as regions and times of increased precipitation, moisture convergence, and evaporation (see e.g. Han and Roads, 2004). Part of the regional hydroclimatological change is also related to how the regional atmospheric moisture, surface vegetation, surface wind, atmospheric planetary boundary-layer stability, and so on change.

The time distribution of hydrologic processes may also change. For example, increased tropical sea-surface temperatures are usually associated with more intense and deep convection. Actually, there are many places where precipitation is not only greater during the summer, but is more likely to form from convective activity. However, periods of intense convection are usually followed by even longer times of clear weather and quiescent conditions and the overall rate of the entire cycle may actually be lower. Thus besides amount, the rate at which hydrologic processes occur could also change as the global temperature increases.

Understanding and predicting future characteristics about these hydrologic processes and reservoirs are a primary motivation for the launch of earth observing satellites, development of extensive surface hydrometeorological networks, and establishment of large programs like the World Climate Research Programme's Global Water and Energy-cycle Experiment, which are designed to encourage and coordinate the multiple activities needed to observe, simulate, and predict variations in the global hydrologic cycle. Remote sensing from space has been an important reason for expanding areal coverage of conventional point measurements to cover historically data poor regions such as the global ocean and unpopulated areas.

The global hydrologic cycle is certainly not adequately observed (see e.g. Roads *et al.*, 2003). For example, evaporation is really only measured in a few land regions at specialized instrument sites, using, for example, eddy correlation techniques. Moisture convergence can just barely be determined from measurement of wind and humidity and even then there is some question as to how accurate these calculations are from the scattered measurements. Soil moisture is also only measured at a few land sites. Even over the ocean, freshwater amounts are unknown since we do not really have a good analysis of large-scale ocean salinity. Runoff is known at gauged basins but there are many ungauged sites within even these basins and besides, it is not altogether clear what the contribution of anthropogenic diversions and subsurface flows are to the cycle in many heavily managed regions. Precipitation is probably the most well known and observed hydrologic variable but there are also well-known problems involved with under catch, lack of gauge measurements over the ocean, and so on. Improving observations of the hydrologic cycle are the key to developing better models, which will ultimately be used to simulate and predict the global hydrologic cycle. Despite the lack of long-term observations, it may at least be possible to estimate precipitation variations as a guide for evaluating corresponding model simulations.

One of the most comprehensive models used to simulate and predict the global hydrologic cycle is a coupled general circulation model. General circulations models are based upon the fundamental laws governing conservation of energy and mass as approximated by large-scale finite differences or, equivalently, severely truncated spherical harmonics. These large-scale models also include parameterizations of physical processes, which are large-scale approximations to nature's small-scale physical processes. Although the models are still imperfect, they are being continually improved by comparing available simulations to observations. Many of the early atmosphere only

simulations used observed sea-surface temperatures as a driving mechanism and many studies have focused in on understanding the response of the atmosphere to changed SSTs. In that regard, it should be mentioned that the Hadley Centre has now produced monthly SSTs for the period 1900–2000, which may be useful for developing preliminary estimates of "observed" hydrologic variables. Many of the early ocean only simulations used atmospheric fluxes as the driving mechanism and many studies have focused in on understanding the response of the ocean to different fluxes. Coupled atmosphere-ocean studies are now beginning to be used as the major tool to predict future variations. For these coupled experiments, the only external forcing is radiative, which depends upon predictions of what this composition might be like given projected population and economic growth.

The hydrologic cycle from such a coupled simulation forced by projected increases in greenhouse gases is described here. In particular, starting from initial conditions relevant to the present day, a scenario that assumes a linear rate of increase of CO_2 to an almost doubled amount in 99 years (371–710 ppm) was used here. We examined mean characteristics of the hydrologic cycle during this "doubling", including how much water is in the atmosphere and subsurface, how much water is evaporated and precipitated, and how much water is converged into an area and then subsequently runs off to the ocean and how these processed changed under this scenario. In addition, we were interested in the cycling rate at which this occurred on average and how these rates might change in the future. Section PCM describes the model used for these calculations. Section "PCM hydrologic cycle" describes hydrologic characteristics and Section "Cycling Rate" describes the estimated rates at which these processes occur. Conclusions are provided in Section "Summary". Basically, the globally averaged precipitation and evaporation did increase in the model's increased greenhouse world. However, cycling rates, measured by the rate at which atmospheric processes occurred in comparison to the available atmospheric reservoir actually slowed down in the atmosphere, because of the relative increase in the atmospheric water vapor reservoir. Nonetheless, because of the dominance of surface water changes over land regions, the overall cycling rate did accelerate over land regions.

PCM

The coupled model used here – the Parallel Climate Model (PCM) – makes use of the National Center for Atmospheric Research (NCAR) Community Climate Model (CCM3), the NCAR Land Surface Model (LSM; which is a modern model incorporating vegetation effects like wilting and transpiration), the Department of Energy (DOE) Los Alamos National Laboratory Parallel Ocean Program (POP), the Naval Postgraduate School sea-ice model, and a distributed flux coupler. The CCM3 is a spectral dynamics model, which uses a T42 resolution (approximately 2.8°) with 18 hybrid levels in the vertical. The LSM simulates the biogeophysics of prescribed vegetation types and hydraulic and thermal properties of 12 soil types. The POP uses a grid with a displaced North Pole at an average of 2.3° latitude and longitude in the midlatitudes, with increased latitudinal resolution near the equator of approximately 1.2°. The sea-ice component predicts the evolution of ice thickness, ice concentration, velocity, snow thickness, and surface temperature; and it uses an elastic-viscous-plastic (EVP) ice rheology dynamics. The distributed flux coupler, originally developed for the NCAR Climate System Model (CSM) connects the PCM components and facilities exchange of flux and state variables among the component models (Washington *et al.*, 2000).

A series of climate change experiments were made with PCM; basic descriptions are given in Dai *et al.* (2001a; 2001b). Briefly, a set of historical runs was made starting in 1870 and running to the present. The model components were each spun up separately, and then run in fully coupled mode for 40 years without any anthropogenic forcing, before the first model year of analysis (1870). As described in Dai *et al.* (2001a), the model overall performed quite well in simulating the observed climate over the past 130 years, with the major deficiency being an inability to properly simulate the warming that occurred between the 1920s and the 1940s. This is likely due to the absence of changes in external solar forcing in the model simulation. On the other hand, the model captured remarkably well the rapid warming since the 1970s, which is thought to be due to the rise in anthropogenic greenhouse gases, which are included in the base model run.

A series of runs were then made projecting into the future through the end of the twenty-first century using various postulated greenhouse gas scenarios. Of particular interest for this project, an ensemble of 99-year Business-As-Usual (BAU) scenario runs beginning in the year 2000 were run on the IBM RS/6000 SP parallel computers at Oak Ridge National Laboratory (Hoffman *et al.*, 2004). These simulations used projected atmospheric increases in CO_2 and other trace gases (Dai, 2001a) comparable to the mean of all the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emissions Scenarios (SRES) scenarios (Nakicenovic and Swart, 2000). CO_2 levels increased linearly from ~371 ppm in 2000 to 710 ppm in 2100. Owing to various reasons, only one of these PCM runs had a reasonably complete set of saved hydrometeorological fields and this run was analyzed here. Given the long period of the individual runs, additional ensemble members would have been nice but were not critical for this analysis of gross average properties. It should be noted, however, that moisture convergence was not available and

output runoff appeared to be erroneous. These fields were therefore deduced from the diagnostic P-E. This assumption is certainly more dubious for runoff, because surface water variations can be important, at least on seasonal timescales but for annual and longer time series the approximation is at least more reasonable.

Figure 1 shows the average surface temperature and the temperature differences (last 30 years minus the first 30 years) caused by the increase in greenhouse gases in this model. The temperature increase is greatest at high latitudes, due in part to the snowmelt and reduction in surface albedo, which also makes a major contribution. That is, the reduction in albedo causes increased solar radiation to be absorbed by the earth, which must therefore balance this increased incoming radiation with increased outgoing radiation via increased surface temperature. On the average the ocean temperatures do increase although the relative increase in the land temperatures with respect to the ocean temperatures is much greater. Of disquieting interest is the slight decrease in the North Atlantic, which is a region of deep-water formation. Presumably the ocean dynamics is acting to reduce the temperature in this region. However, as shown in Table 1, on the average, the ocean was still relatively warmer than the colder land. It should be noted here that for this paper, the land mean excluded the permanent ice covered Greenland and Antarctica continents. It should be noted however, that the global mean did include these areas. The time series of land, ocean, and global shows the temperature increase is fairly linear with respect to the linearly increasing greenhouse gases.

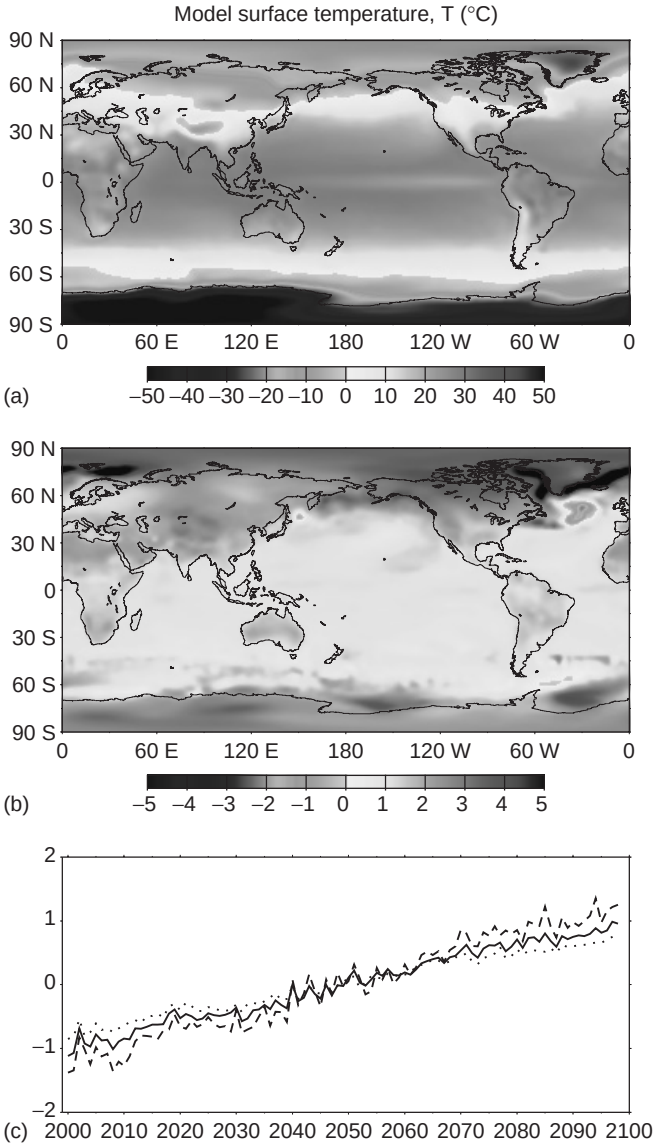


Figure 1 Model surface temperature, T, (C): (a) Mean surface temperature, T, C for the 99-year simulation: (b) difference in the surface temperature, ΔT, C (average of years 69–98 minus average over years 0–29); (c) annual differences from the overall means of the global (solid line), land (dashed line), ocean (dotted line) temperatures for each year of the 99-year simulation. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

PCM HYDROLOGIC CYCLE

Consider the following atmospheric and surface hydrologic cycle mass conservation equations:

$$\frac{\partial Q}{\partial t} = E + MC - P$$
$$\frac{\partial W}{\partial t} = -E - N + P$$
$$\frac{\partial (Q + W)}{\partial t} = MC - N \qquad (1)$$

Q is the total atmospheric water (usually referred to as total precipitable water and includes water vapor as well as cloud and precipitation water), W is the total surface water (over land this is the soil moisture above the saturated zone plus snow water equivalent), E is the evaporation, MC is the atmospheric moisture convergence, N is the surface runoff, and P the precipitation. Over the ocean, the surface water runoff is actually freshwater transport, which is more similar to the atmospheric moisture convergence, since it is dependent upon ocean currents instead of gravitational drainage. It should also be mentioned that the surface water is not well defined over the ocean; instead a more appropriate variable is salinity, which measures how fresh or saline a particular body of water is. It should be noted that the tendency terms are important on diurnal to daily timescales but at longer timescales (interannual to decadal) become relatively small and the magnitude of the reservoirs (Q and W) are determined by complex

Table 1 Global, land (except Antarctica and Greenland), and ocean means and differences (average of years 69–98 minus average over years 0–29) for T, Q, W, P, E, MC, N, P/Q, P/W, E/Q, E/W, MC/Q, N/W, $MC/(Q+W)$, $N/(Q+W)$. NA means this entry is not available

Variable	Years	Land	Ocean	Global
T, C	2000–2029	12.0951	19.4729	14.1900
	2069–2098	13.8914	20.5476	15.5632
	(2069–98)–(2000–29)	1.7964	1.0747	1.3732
	2000–2098	12.9740	20.0002	14.8744
	2000–2098sd	0.7591	0.4545	0.5791
Q, mm	2000–2029	18.4254	27.9137	23.3622
	2069–2098	20.3631	30.4105	25.5517
	(2069–98)–(2000–29)	1.9377	2.4969	2.1896
	2000–2098	19.3878	29.1309	24.4453
	2000–2098sd	0.8094	1.0600	0.9236
P, mm/day	2000–2029	2.3092	3.6740	3.1060
	2069–2098	2.4039	3.7388	3.1809
	(2069–98)–(2000–29)	0.0947	0.0648	0.0749
	2000–2098	2.3564	3.7037	3.1432
	2000–2098sd	0.0469	0.0320	0.0325
E, mm/day	2000–2029	1.7288	4.0230	3.1057
	2069–2098	1.7952	4.1089	3.1806
	(2069–98)–(2000–29)	0.0664	0.0860	0.0749
	2000–2098	1.7621	4.0632	3.1427
	2000–2098sd	0.0295	0.0387	0.0325
N, mm/day	2000–2029	0.5804	NA	NA
	2069–2098	0.6087	NA	NA
	(2069–98)–(2000–29)	0.0283	NA	NA
	2000–2098	0.5943	NA	NA
	2000–2098sd	0.0219	NA	NA
M, mm	2000–2029	185.2886	NA	NA
	2069–2098	185.6194	NA	NA
	(2069–98)–(2000–29)	0.3308	NA	NA
	2000–2098	185.4948	NA	NA
	2000–2098sd	0.6932	NA	NA
S, mm	2000–2029	12.2438	NA	NA
	2069–2098	11.0157	NA	NA
	(2069–98)(2000–29)	−1.2281	NA	NA
	2000–2098	11.6865	NA	NA
	2000–2098sd	0.8208	NA	NA
W, mm	2000–2029	197.5334	NA	NA
	2069–2098	196.6362	NA	NA
	(2069–98)–(2000–29)	−0.8972	NA	NA
	2000–2098	197.1771	NA	NA
	2000–2098sd	1.0535	NA	NA
MC, mm/day	2000–2029	0.5804	−0.3490	0.0000
	2069–2098	0.6087	−0.3701	−0.0001
	(2069–98)–(2000–29)	0.0283	−0.0211	−0.0001
	2000–2098	0.5943	−0.3595	−0.0001
	2000–2098sd	0.0219	0.0117	0.0012
P/Q, day^{-1}	2000–2029	0.1381	0.1433	0.1571
	2069–2098	0.1299	0.1335	0.1470
	(2069–98)–(2000–29)	−0.0082	−0.0098	−0.0101
	2000–2098	0.1341	0.1382	0.1521
	2000–2098sd	0.0038	0.0040	0.0043
P/W, day^{-1}	2000–2029	0.0127	NA	NA
	2069–2098	0.0133	NA	NA
	(2069–98)–(2000–29)	0.0006	NA	NA

continued overleaf

Table 1 (*continued*)

Variable	Years	Land	Ocean	Global
	2000–2098	0.0130	NA	NA
	2000–2098sd	0.0003	NA	NA
E/Q, day^{-1}	2000–2029	0.0957	0.1495	0.1297
	2069–2098	0.0901	0.1403	0.1216
	(2069–98)–(2000–29)	−0.0057	−0.0091	−0.0080
	2000–2098	0.0929	0.1448	0.1257
	2000–2098sd	0.0024	0.0038	0.0034
E/W, day^{-1}	2000–2029	0.0108	NA	NA
	2069–2098	0.0112	NA	NA
	(2069–98)–(2000–29)	0.0004	NA	NA
	2000–2098	0.0110	NA	NA
	2000–2098sd	0.0002	NA	NA
MC/Q, day^{-1}	2000–2029	0.0423	−0.0062	0.0274
	2069–2098	0.0398	−0.0068	0.0253
	(2069–98)–(2000–29)	−0.0026	−0.0006	−0.0021
	2000–2098	0.0411	−0.0066	0.0264
	2000–2098sd	0.0017	0.0006	0.0011
N/W, day^{-1}	2000–2029	0.0048	NA	NA
	2069–2098	0.0051	NA	NA
	(2069–98)–(2000–29)	0.0002	NA	NA
	2000–2098	0.0050	NA	NA
	2000–2098sd	0.0002	NA	NA
$MC/(Q+W)$, day^{-1}	2000–2029	0.0019	NA	NA
	2069–2098	0.0021	NA	NA
	(2069–98)–(2000–29)	0.0001	NA	NA
	2000–2098	0.0020	NA	NA
	2000–2098sd	0.0001	NA	NA
$N/(Q+W)$, day^{-1}	2000–2029	0.0043	NA	NA
	2069–2098	0.0045	NA	NA
	(2069–98)–(2000–29)	0.0002	NA	NA
	2000–2098	0.0044	NA	NA
	2000–2098sd	0.0001	NA	NA

interactions between the other processes. That is not to say, however, that the tendency terms are negligible; seasonal and interannual variations, which measure the seasonal and interannual imbalances in the other processes, can be quite important, especially for surface water. Finally, please note that the last equation is an expression of mass conservation, which shows that atmospheric moisture convergence must equal the surface divergence or runoff (or ocean freshwater divergence) if the freshwater mass is not to increase or decrease locally, at least on the average.

The global distribution of these hydrometeorological processes in the CSM 99 year average is shown in Figure 2, along with the reservoirs. Note the concentration of the atmospheric hydrologic processes near the tropics, where precipitation, evaporation, moisture convergence, and runoff are large. Again, this geographical distribution is a consequence, but only in part, of the nonlinear Clausius–Clapeyron relation. Note that just adjacent to the tropical regions are the subtropical regions, which exhibit some of the largest evaporation regions (ocean) and some of the smallest evaporation regions (land), again despite having some of the highest global surface temperatures. Water availability at the surface as well as in the atmosphere adjacent to the surface, are just as important as temperature changes. For example, the distribution of individual hydrologic processes and reservoirs is certainly different than the temperature distribution. Precipitation is greatest in the intertropical convergence zones, where convergence is greatest, and least in the subtropical regions, where the divergence is greatest. On the average, water is diverged from the ocean to the land. This land convergence of atmospheric moisture is balanced by the surface runoff, which transports the surface water back to the ocean. The surface water is greatest in the regions of permanent ice cover (Greenland and Antarctica).

Although the qualitative distributions of these processes and reservoirs are reasonably realistic (see e.g. Roads *et al.*, 2002) and illustrate the overall complexity in the hydrologic cycle that goes way beyond simple relations to temperature, there are certainly problems with the simulation. For example, note the double intertropical convergence zone, which is a common model defect.

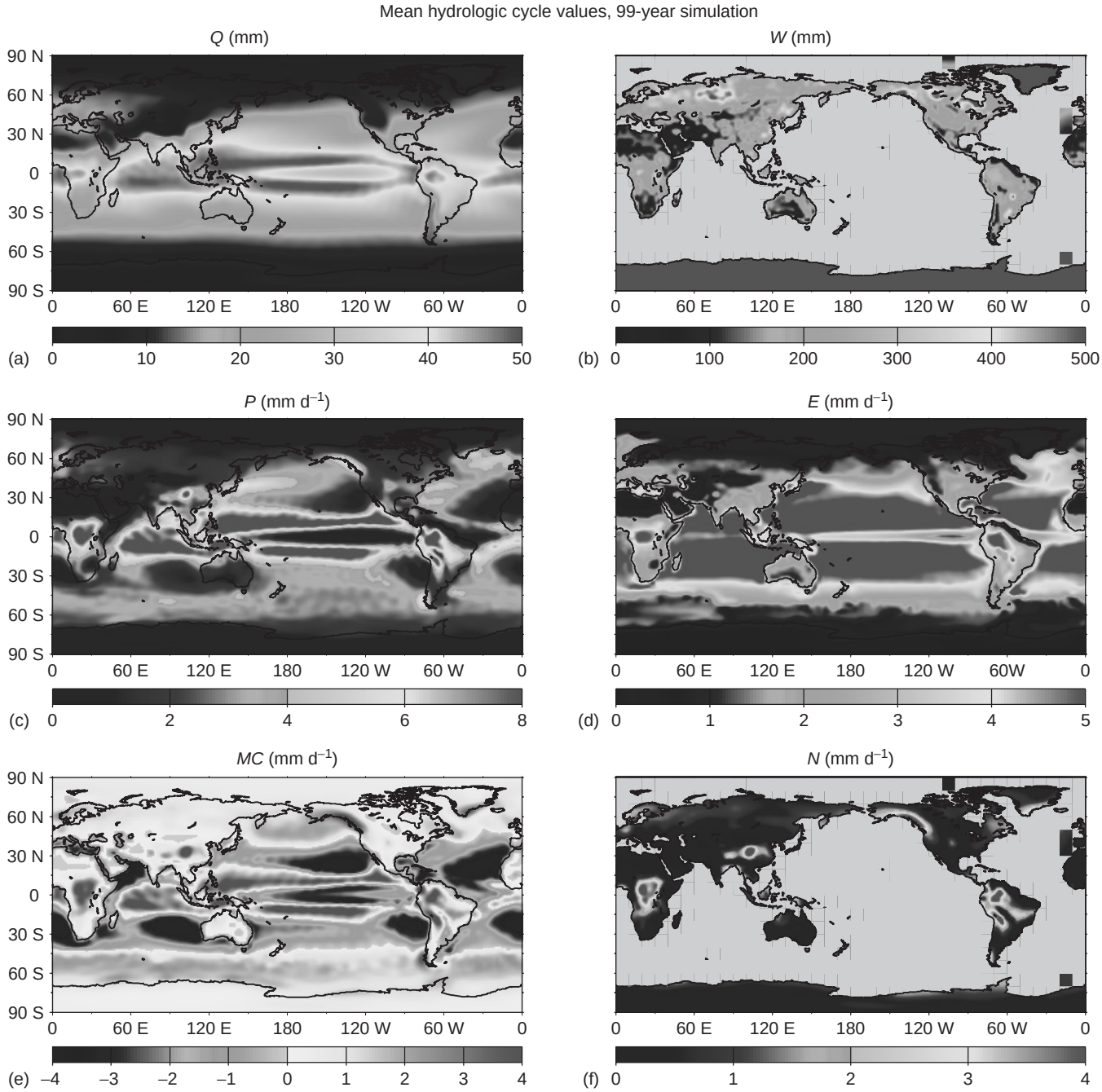


Figure 2 Mean hydrologic cycle values for the 99-year simulation: (a) *Q*, precipitable water, mm; (b) *W*, surface water, mm; (c) *P*, precipitation, mm/day; (d) *E*, evaporation, mm/day; (e) *MC*, moisture convergence, mm/day; (f) *N*, runoff, mm/day. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

There are other subtle defects that reduce our overall confidence in the overall projection, and again, efforts must continue to substantially reduce the uncertainty in the ability of these models to simulate future climate change.

Whether the hydrologic cycle will accelerate in a world with increased greenhouse gases could be defined as a simple difference in these processes. Do evaporation and precipitation increase, and where do these increases occur? How do changes in moisture convergence and runoff contribute to these increases as well as any decreases? How do the atmospheric and surface water reservoirs change? What are the regional differences? Are these regional differences significant? As a measure of the significance of

the changes, the difference of the last 30 year from the first 30-year mean is divided by the estimated standard deviation of 30-year means, which is obtained by finding the standard deviation of annual means and then dividing this value by the square root of 30, which assumes the interannual variations are independent.

As shown in Figure 3, there are certainly many significant differences between these processes at the end of the 99-year run from the beginning of the 99-year run as measured by normalized differences between the last 30 years and the first 30 years. The most significant differences occur in the precipitable water, which increases everywhere, and

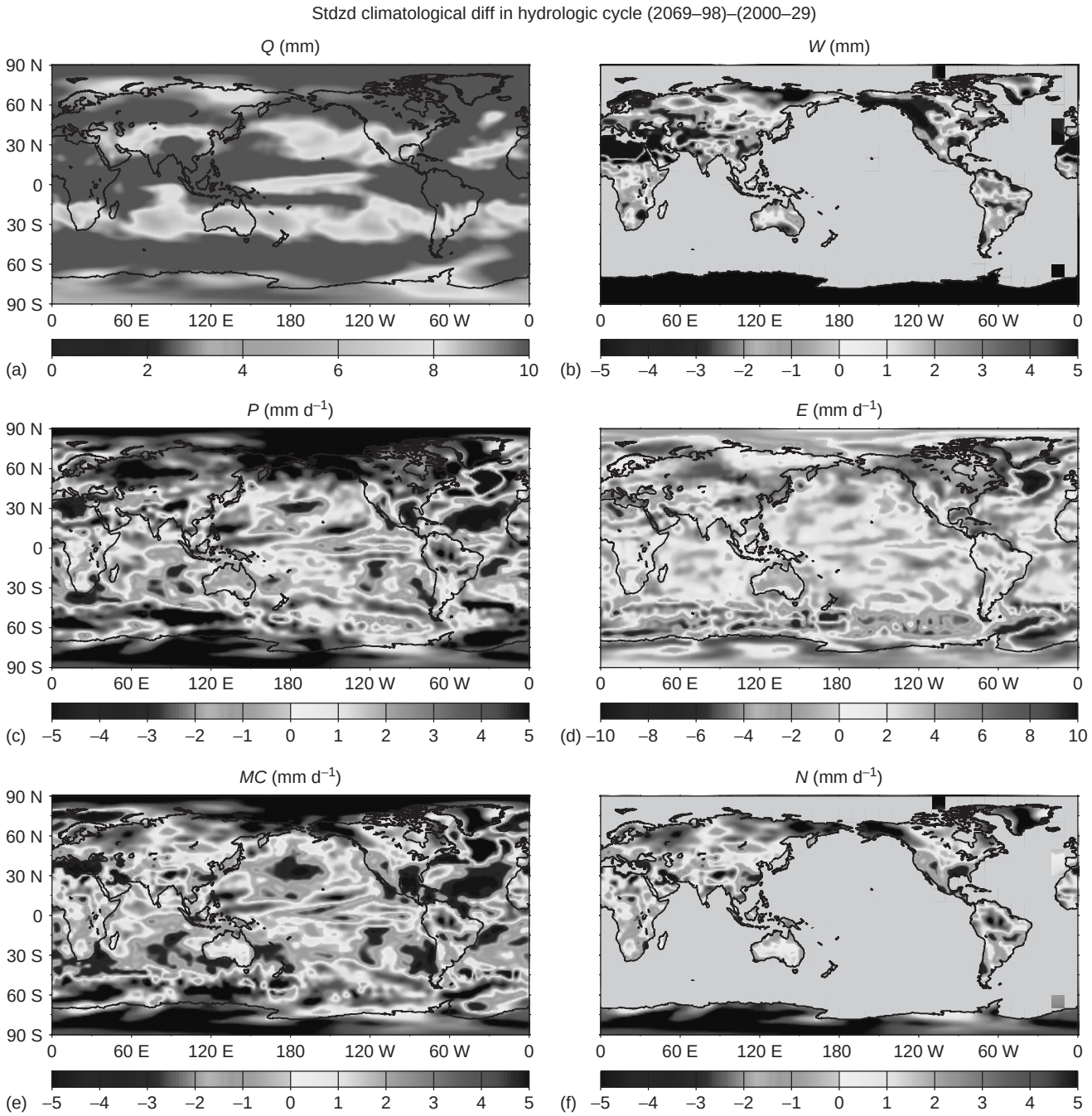


Figure 3 Climatological differences in the hydrologic cycle (average of years 69–98 minus average over years 0–29) divided by estimates of the 30-year mean standard deviation: (a) ΔQ, precipitable water; (b) ΔW, surface water; (c) ΔP, precipitation; (d) ΔE, evaporation; (e) ΔMC, moisture convergence; (f) ΔN, runoff. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

in the evaporation, which increases almost everywhere, except for the North Atlantic and high latitude southern oceans. There are also slight decreases in other regions. Precipitation shows relative increases in high latitudes, excepting the North Atlantic region again, where the evaporation was reduced. The precipitation, however, shows relative decreases in tropical to subtropical latitudes, which is consistent with model simulations covering the twentieth century Kumar *et al.* (2004). Most of the precipitation changes appear to be correlated with the changes in moisture convergence, which on the average show increased convergence in the high latitudes and increased divergence in lower latitudes. Again, there are certainly notable exceptions, including the North Atlantic, where it appears the evaporation changes are more important than the moisture convergence changes. The runoff, which must balance the convergence (and here it does by design), has the same features and perhaps, surprisingly, so does the surface water. Note, for example, the increase in surface water over Antarctica. However, over most tropical and midlatitude land regions the surface water decreases, as was previously shown by Wetherald and Manabe (1999).

Figure 4 shows a time series of the globally averaged, the land averaged, and the ocean averaged differences for these processes. Precipitable water increases linearly by about 2 mm whereas surface water has much stronger variability and overall a slight decrease. Note the gradual increase in global precipitation and evaporation from averages of 3.1 mm day^{-1} to 3.2 mm day^{-1}, which is somewhat smaller than the global value from an ensemble of models forced with +2K and −2K uniform sea-surface temperature changes (Randall *et al.*, 1992). This increase is accompanied by a slight increase in MC over land, along with an increase in runoff, and increased divergence over the ocean. The precipitation variations over land are especially high and have amplified values every 20 years or so, accompanied by shorter period high amplitude variations that may be related to the changes in surface water. These interannual variations deserve further scrutiny, but this is beyond the scope of the present study.

CYCLING RATE

Another measure of the hydrologic cycle acceleration is the cycling rate described previously by Chahine *et al.* (1997; see also Trenberth, 1997, 1998; Roads *et al.*, 1998; Douville *et al.*, 2002; Bosilovich *et al.*, 2004). The cycling rate is a measure of the timescale of the processes, which is obtained here by dividing the first equation by Q, the second by W, and the third by $(Q + W)$:

$$\frac{\partial \ln Q}{\partial t} = \frac{E}{Q} + \frac{MC}{Q} - \frac{P}{Q}$$

$$\frac{\partial \ln W}{\partial t} = -\frac{E}{W} - \frac{N}{W} + \frac{P}{W}$$

$$\frac{\partial \ln(Q + W)}{\partial t} = \frac{MC}{(Q + W)} - \frac{N}{(Q + W)} \qquad (2)$$

Note that the only unit in these equations is an inverse time or rate for the various hydrometeorological processes. These rates are hereafter referred to here as cycling rates (we prefer cycling to the previous recycling designation, which has also been previously used to describe the relative contributions of local evaporation versus transport), since these rates are a measure of how fast the various hydrologic processes occur. So, another measure of whether the hydrologic cycle is accelerating could be related to changes in these cycling rates, which, again, are related to both the change in the individual processes and the changes in the individual reservoirs. Exactly how these rates are calculated may have some influence on their perceived value. For example, instantaneous ratios could be calculated and then averaged, but here only annual means were available and as will be shown, the ratios $[P]/[Q] \approx [P/Q]$. Whether or not significant differences occur when shorter timescales (weather) are considered would be of great interest, but this is also beyond the scope of the present study (and available data).

Figure 5 shows geographical characteristics of the cycling ratio. As shown in Table 1, precipitation and evaporation have overall atmospheric cycling rates (P/Q, E/Q) of. 157 days^{-1} or residence times of 6.6 days, although in snow covered regions, where the precipitable water is small, the residence time can be on the order of 2 days. This is a bit higher than previous estimates (e.g. Roads *et al.*, 1998) in part, because of the excessive precipitation and reduced precipitable water in this model, compared to observations. Evaporation cycling rates are a bit more uniform, although certainly smaller, over subtropical regions, where moisture convergence becomes more important. The cycling rates in the land surface are much smaller, of order 100 days in colder snow covered regions although some regions, such as the tropical land regions, have cycling rates in the surface comparable to atmospheric cycling rates. The overall cycling rates for the total atmospheric and surface mass over the land regions are smaller still, since they are dependent upon atmospheric and surface transport times rather than surface flux and precipitation times.

Figure 6 shows the standardized climatological differences in these cycling rates. For the most part, atmospheric cycling rates calculated from precipitation and evaporation are negative almost everywhere. As shown in Table 1, the average precipitation and evaporation differences are on the order of −0.01 days^{-1}, which is a change in the precipitation cycling times about 6.4 to

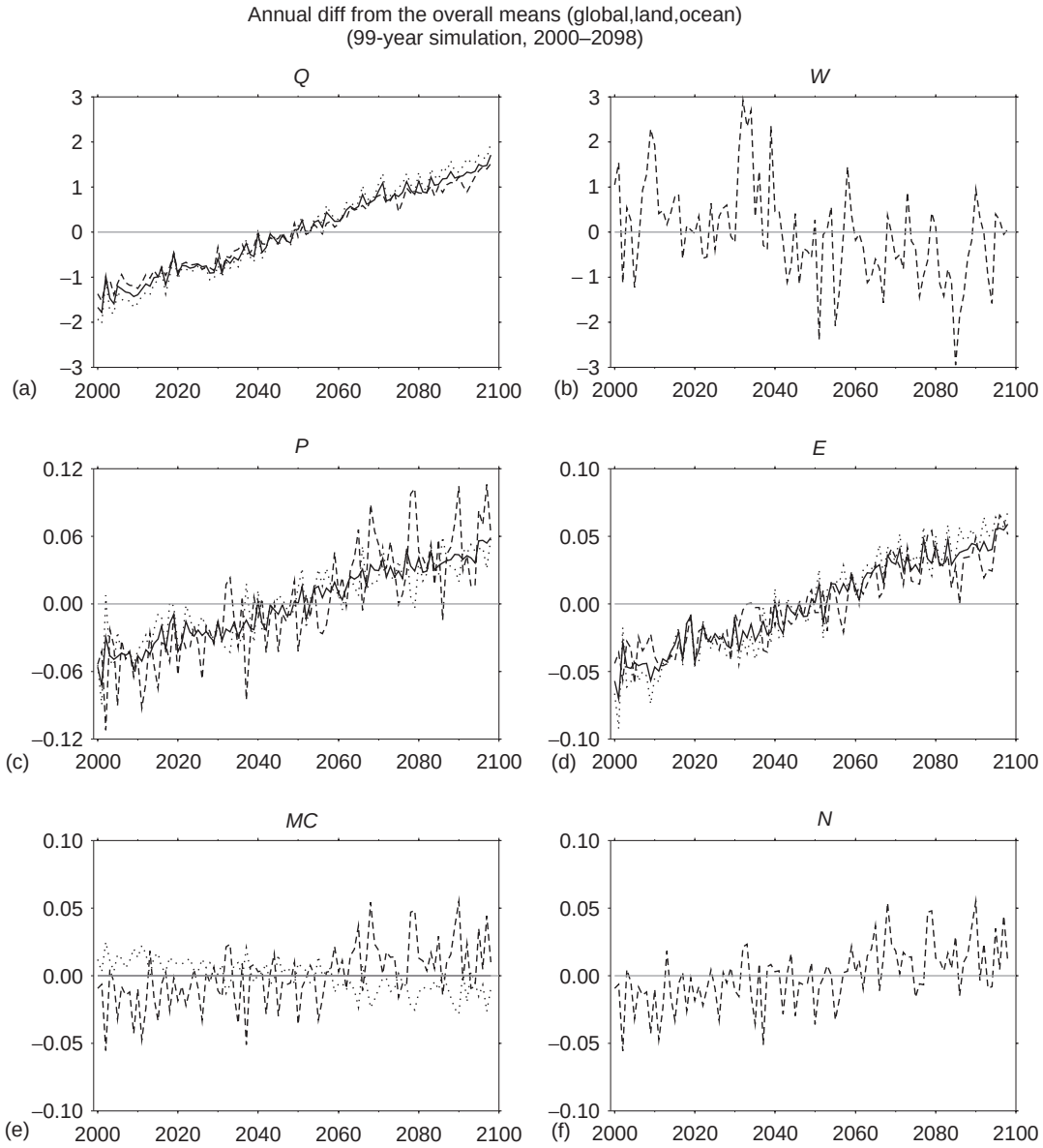


Figure 4 Annual differences from the overall means of the global (solid line), land (dashed line), ocean (dotted line) hydrologic cycle values for each year of the 99-year simulation: (1) ΔQ, precipitable water, mm; (b) ΔW, surface water, mm; (c) ΔP, precipitation, mm/day; (d) ΔE, evaporation, mm/day; (e) ΔMC, moisture convergence, mm/day; (f) ΔN, runoff, mm/day

6.8 days. The cycling rate differences due to moisture convergence are complex, and differences depend not only on changes in moisture convergence but also the precipitable water. Somewhat opposite features occur in the land surface cycling, which as noted previously showed less of an increase in the reservoir (in fact an overall decrease) as the precipitation and evaporation increase. The runoff cycling is related more to the relative increases in runoff in high latitudes and relative decrease in runoff in low latitudes. The overall cycling rate, measured by $MC/(Q + W)$ or $N/(Q + W)$, also shows this high latitude increase and low latitude decrease. A notable exception is the decrease over Antarctica, which as noted previously had a surprisingly substantial increase in the surface water.

Figure 7 shows a time series of the globally averaged cycling rates. Again note that on the average, the atmospheric cycling rate tends to decrease, despite the increasing amount of precipitation, evaporation, while the land cycling rate tends to increase. These changes are related to

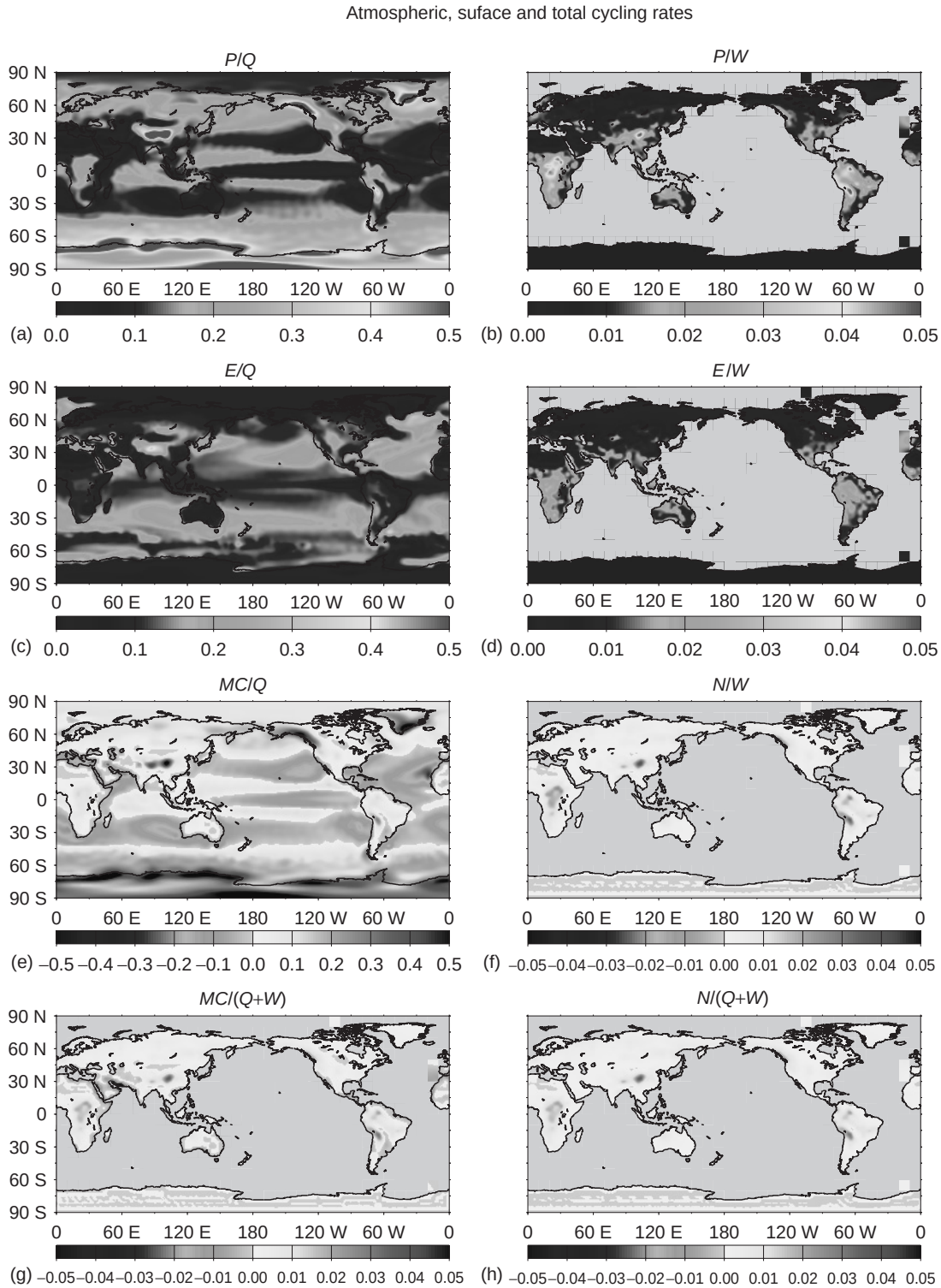


Figure 5 Atmospheric, surface, and total cycling rates, d^{-1}: (a) P/Q; (b) P/W; (c) E/Q; (d) E/W; (e) MC/Q; (f) N/W; (g) $MC/(Q+W)$; (h) $N/(Q+W)$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

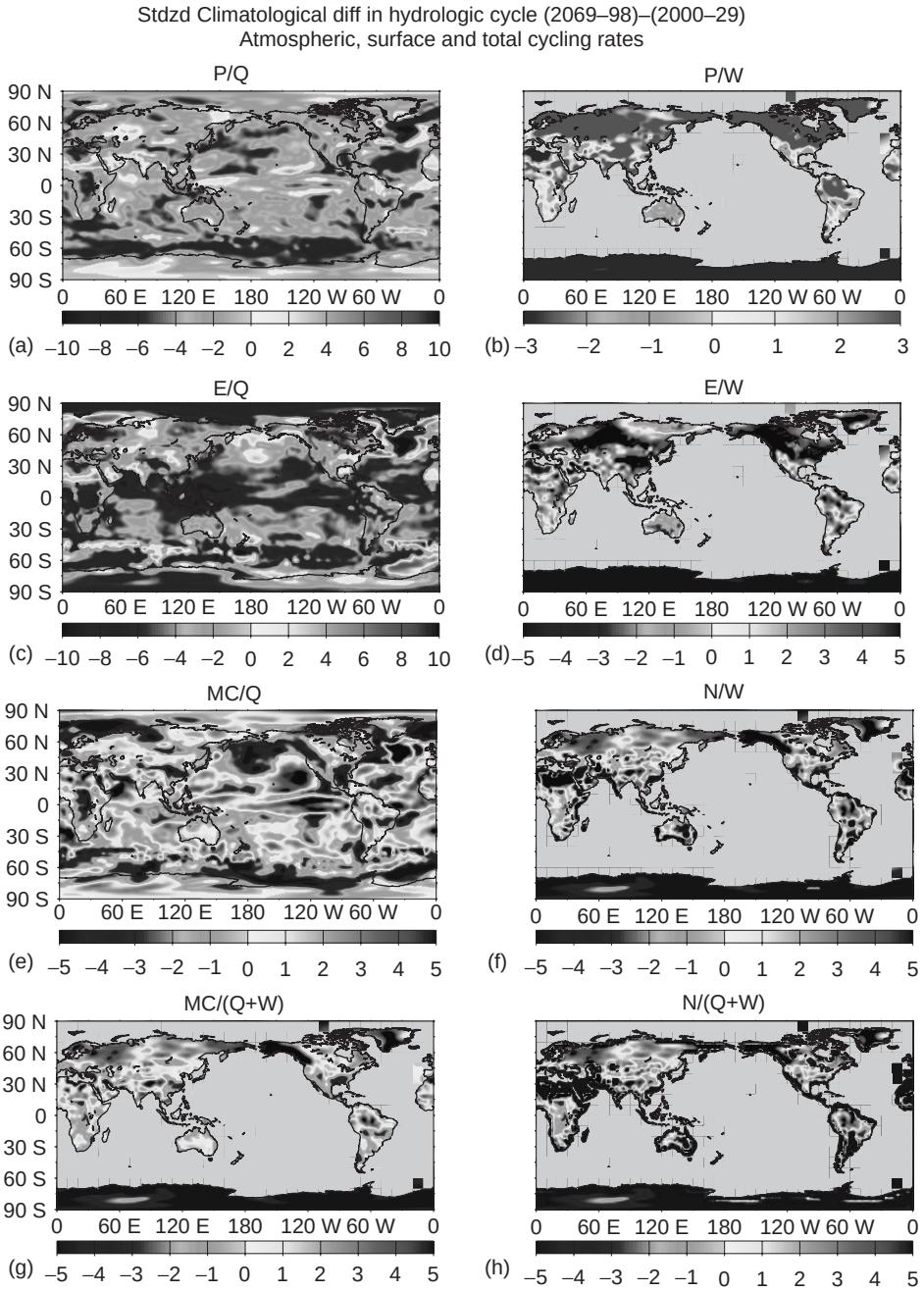


Figure 6 Climatological differences in the hydrologic cycle (average of years 69–98 minus average over years 0–29) atmospheric, surface, and total cycling rates, divided by estimates of the 30-year standard deviation: (a) $\Delta(P/Q)$; (b) $\Delta(P/W)$; (c) $\Delta(E/Q)$; (d) $\Delta(E/W)$; (e) $\Delta(MC/Q)$; (f) $\Delta(N/W)$; (g) $\Delta(MC/(Q+W))$; (h) $\Delta(N/(Q+W))$. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

the changes in the atmospheric reservoir, which increases substantially, and the surface reservoir, which slightly decreases. When the total water mass is considered over land regions, which shows a moderate increase, then the total (atmosphere and land) cycling rate does show a moderate increase or "acceleration", in a world with increased greenhouse gases.

SUMMARY

So, does the hydrologic cycle accelerate? If by acceleration we mean that there should be an increase in a particular process then on the average the evaporation and precipitation increase with increasing temperature and this results in increased transport from ocean to land. However, there are

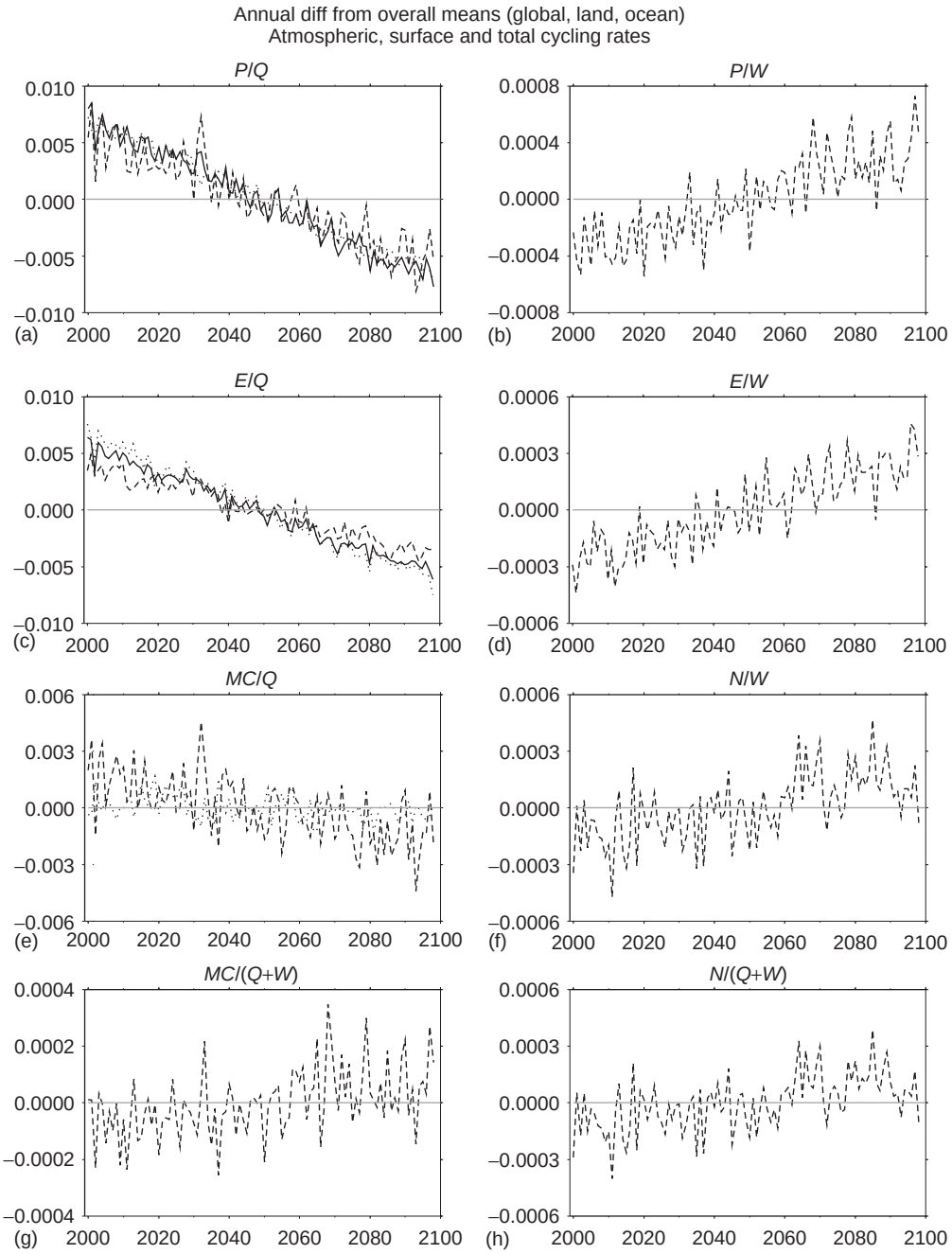


Figure 7 Annual differences from the overall means of the global (solid line), land (dashed line), ocean (dotted line) atmospheric, surface, and total cycling rates, d^{-1}: (a) $\Delta(P/Q)$; (b) $\Delta(P/W)$; (c) $\Delta(E/Q)$; (d) $\Delta(E/W)$; (e) $\Delta(MC/Q)$; (f) $\Delta(N/W)$; (g) $\Delta(MC/(Q+W))$; (h) $\Delta(N/(Q+W))$

certainly many regional differences, which are dependent upon changes in the moisture convergence, land evaporation, surface runoff, and surface water.

If we take into account the relative change in the atmospheric and surface reservoirs then on the average the atmospheric cycling is reduced, owing to the greater water holding capacity of the atmosphere and the surface cycling is enhanced, owing to the small decrease in water holding capacity of the surface water. Again, there are important regional differences. Over high latitude land (except Antarctica), the total cycling rate is increased, owing in part to the more dominant change in the processes relative to the changing surface water, whereas low latitude land does show decreased cycling rates owing more to

decreases in the processes than changes in the atmosphere, surface, or total water reservoirs.

Again, there is also a need to consider averages of instantaneous cycling rates to better understand how robust these measures using annual means are. There are also other measures of hydrologic cycle acceleration that should be considered. For example, what are the changes in the frequency of more extreme weather and short-term climate events? Does precipitation change its distribution from slowly acting processes to one of short duration but more intense processes? Rainfall rates from tropical convection are distinctly different from midlatitude nimbostratus, which has hydrologic implications for soil moisture and runoff. In other words, what are the weather and short-term climate events like in an increased greenhouse world? Understanding whether or not these weather and short-term climate events are reasonable in current large-scale coupled models, however, will require better understanding and improved capability to simulate them. We would also like to note that changes in the hydrologic cycle cannot be readily understood without also understanding associated radiative and sensible heat fluxes, their modes of variability, and their response in increased greenhouse gas climates.

Acknowledgments

This study was supported by NOAA NA17RJ1231 and grants under the NASA Global Water and Energy-Cycle Program, NASA NAG5-11738 and NAG8-1875. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or NASA.

REFERENCES

Bosilovich M., Schubert S. and Walker G. (2004) Global changes of the water cycle intensity. 15th *Symposium on Global Change and Climate Variations*, Seattle, Jan. 2004.

Chahine M.T., Haskins R. and Fetzer E. (1997) Observation of the recycling rate of moisture in the atmosphere: 1988:1994. *Gewex News*, **7**, 1–4.

Dai A., Meehl G.A., Washington W.M., Wigley T.M.L. and Arblaster J.A. (2001b) Ensemble simulation of 21st century climate changes: business as usual vs. CO2 stabilization. *Bulletin of the American Meteorological Society*, **82**, 2377–2388.

Dai A., Wigley T.M.L., Boville B.A., Kiehl J.T. and Buja L.E. (2001a) Climates of the 20th and 21st centuries simulated by the NCAR climate system model. *Journal of Climate*, **14**, 485–519.

Douville H., Chauvin F., Planton S., Royer J.-F., Salas-Melia D. and Tyteca S. (2002) Sensitivity of the hydrological cycle to increasing amounts of greenhouse gasses and aerosols. *Climate Dynamics*, **20**, 45–68.

Han J. and Roads J. (2004) US climate sensitivity simulated with the NCEP regional spectral model. *Climate Change*, **62**, 115–154. doi:10.1023/B:CLIM.0000013675.66917.15.

Hoffman F., Hargrove R., Erickson D. and Oglesby R. (2004) PCM simulations. *Earth Interactions*, (in press).

Kumar A., Yang F., Goddard L. and Schubert S.D. (2004) Differing trends in the tropical surface temperatures and precipitation over land and oceans. *Journal of Climate*, **17**(3), 653–664.

Nakicenovic N. and Swart R. (Eds.) (2000) *IPCC Special Report on Emission Scenarios*, Cambridge University Press: Cambridge, p. 570.

Randall D.A., Cess R.D., Blanchet J.P., Boer G.J., Dazlich D.A., Del Genio A.D., Deque M., Dymnikov V., Galin V., Ghan S.J., Lacis A.A., Le Treut H., Li Z.-X., Liang X.-Z., McAvaney B.J., Meleshko V.P., Mitchell J.F.B., Morcrette J.-J., Potter G.L., Rikus L., Roeckner E., Royer J.F., Schlese U., Sheinin D.A., Slingo J., Sokolov A.P., Taylor K.E., Washington W.M., Wetherlad R.T., Yagai I., and Zhang M.-H. (1992) Intercomparison and interpretation of surface energy fluxes in atmospheric general circulation models. *Journal of the Geophysical Research*, **97**, 3711–3724.

Roads J.O., Chen S.-C., Marshall S. and Oglesby R. (1998) Atmospheric moisture cycling rates. *GEWEX News*, **8**(3), 7–10.

Roads J., Kanamitsu M. and Stewart R. (2002) CSE water and energy budgets in the NCEP-DOE reanalysis II. *Journal of Hydrometeorology*, **3**, 227–248.

Roads J., Lawford R., Bainto E., Berbery E., Chen S., Fekete B., Gallo K., Grundstein A., Higgins W., Kanamitsu M., *et al.* (2003) GCIP water and energy budget synthesis (WEBS). *Journal of the Geophysical Research*, **108**(D18), GCP 4-1 to 4-39, 10.1029/2002JD002583.

Trenberth K. (1997) Atmospheric moisture residence times and Cycling: Implications for how precipitation may change as climate changes. *Gewex News*, **3**(1), 4–6, 16.

Trenberth K.E. (1998) Atmospheric moisture residence times and cycling: Implications for rainfall rates and climate change. *Climatic Change*, **39**, 667–694.

Washington W.M., Weatherly J.W., Meehl G.A., Semtnet A.J. Jr, Bettge T.W., Craig A.P., Strand W.G. Jr, Arblaster J., Wayland V.B., James R. *et al.* (2000) Parallel climate model (PCM) control and transient simulations. *Climate Dynamics*, **16**, 755–774.

Wetherald R.T. and Manabe S. (1999) Detectability of summer dryness caused by greenhouse warming. *Climatic Change*, **43**(3), 495–511.

196: The Role of Water Vapor and Clouds in the Climate System

GRAHAME STEPHENS

Atmospheric Sciences, Colorado State University, Fort Collins, CO, US

The exchanges of water between the different water reservoirs on Earth establishes the so-called hydrological cycle. The influence of this cycle on the energy budget of Earth is central to understanding and predicting climate change. In fact, those processes that characterize the smallest of the water reservoirs - namely the atmospheric branch of the hydrological cycle - play an especially critical role in regulating climate change through feedbacks processes. Two key feedback processes are described. The first deals with the feedback associated with the relation between water vapor, radiation and temperature and the second concerns feedbacks associated with clouds and radiation.

Understanding the climate of Earth and predicting the way it varies in time requires a quantitative understanding of the way water cycles back and forth between its main reservoirs both in the atmosphere and at the Earth's surface. The exchanges of water between these reservoirs establishes the so-called hydrological cycle and it is the influence of this cycle on the energy budget of Earth that is central to understanding and predicting climate change. Those processes that relate to the smallest of the reservoirs of water – namely the atmospheric branch of the hydrological cycle – play an especially critical role in regulating climate change through feedbacks (refer to Figure 1) that occur as a result of the influence of these forms of water on the Earth's energy balance. Changes to the amount and distribution of water vapor and clouds, in particular, such as might occur in response to the radiation imbalances induced by increases in greenhouse gases (referred to as a climate forcing), establish these key feedbacks.

THE PLANET'S ENERGY BALANCE

A global-annual-mean depiction of the Earth's energy balance is presented in Figure 2. The mean radiant energy leaving the planet is balanced by the incoming radiation from the sun ($\sim$342 W m^{-2}). The processes responsible for this balance are the emission of infrared or thermal radiation from the Earth-atmosphere system to space ($\sim$235 W m^{-2}) and the reflection of solar radiation from that atmosphere and surface also back to space ($\sim$107 W m^{-2}). The long-wave radiant energy leaving the planet is referred to as the outgoing long wave radiation (OLR) and is carried by radiation of wavelengths longer than 4 μm. The energy in reflected sunlight lies in those wavelengths shorter than 4 μm and is typically expressed in terms of the ratio between the outgoing solar radiation and the incoming solar radiation referred to as the *planetary albedo* ($= 107/342 = 0.31$).

By summing the energy fluxes that flow to and from the atmosphere in Figure 2, we learn that the atmosphere constantly loses energy by the net exchanges of radiation between the atmosphere and space. According to Figure 2, we estimate this loss to be of approximately 102 W m^{-2}. A more common way of expressing this radiative loss is in terms of the rate of cooling of the atmosphere (e.g. Stephens *et al.*, 1994) and to first order, this loss occurs primarily from the emission of infrared radiation from the atmosphere. This emission in turn is grossly influenced by the absorption and thus emission of infrared radiation (IR) by water vapor and, to a lesser extent, clouds.

The radiant energy loss from the atmosphere is balanced by energy transfer from the surface via convective and turbulent transfer (thermals, 24 W m^{-2}; evapotranspiration, 78 W m^{-2}). The dominant contribution by

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

evapo-transpiration represents the excess latent heat release because of the formation of precipitation in the atmosphere in excess of the energy taken from the surface for evaporation of water at the surface and the energy taken from air to evaporate clouds and precipitation. This net heating of approximately $78\,W\,m^{-2}$ is proportional to the precipitation falling from the atmosphere to the surface as this precipitation, as the global average is in balance with the evaporation from the surface.

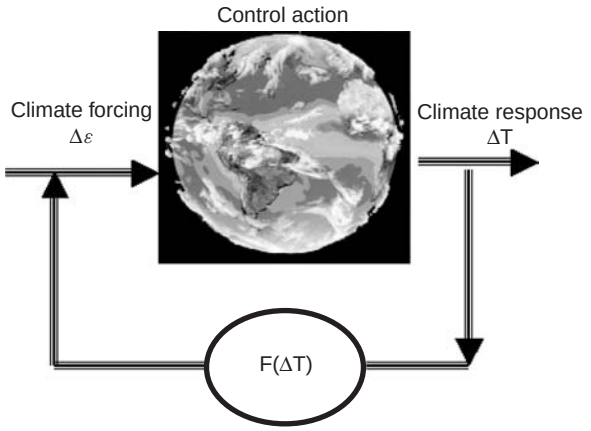


Figure 1 The notion of a climate feedback system. The system is defined by the external forcing $\Delta\varepsilon$ such as associated with the increased buildup of greenhouse gases, the system output or response such as global mean surface temperature change (ΔT), the control action that represents the physical climate system defining the connection between input and output, and processes $F(\Delta T)$ that establish the feedback through the dependence of the process on the response ΔT. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

A simple demonstration of the importance of the link between the radiative processes that govern the energy budget of the planet and the hydrological-related processes that dictate the distribution of water around the planet is provided when considering the approximate balance between the global mean radiation budget of the atmosphere and the latent heating associated with precipitation. The implication of this gross balance is that global-scale changes in the atmospheric cooling, induced for example by the buildup of greenhouse gases, most likely require compensating changes in latent heating and thus fundamental changes in the Earth's hydrological cycle. It is thus appropriate to consider this radiative cooling, controlled by water vapor and clouds, not only as a rudimentary measure of the activity of the Earth's greenhouse effect but also as an indirect measure of the gross activity of the hydrological cycle in heating the atmosphere.

WATER VAPOR, CLOUDS, AND THE PLANET'S ENERGY BALANCE

Water vapor absorbs solar radiation and both absorb and emit infrared radiation. Through absorption of solar radiation, water vapor acts to reduce the planetary albedo and through the absorption and emission of infrared radiation water vapor acts as the principal greenhouse gas, and the infrared emission from the atmosphere (both to the surface and to space) especially in clear sky regions is defined largely by the total water vapor content of the atmosphere (e.g. Chahine, 1992).

Water vapor also represents an essential stage in the hydrological cycle connecting water at the Earth's surface

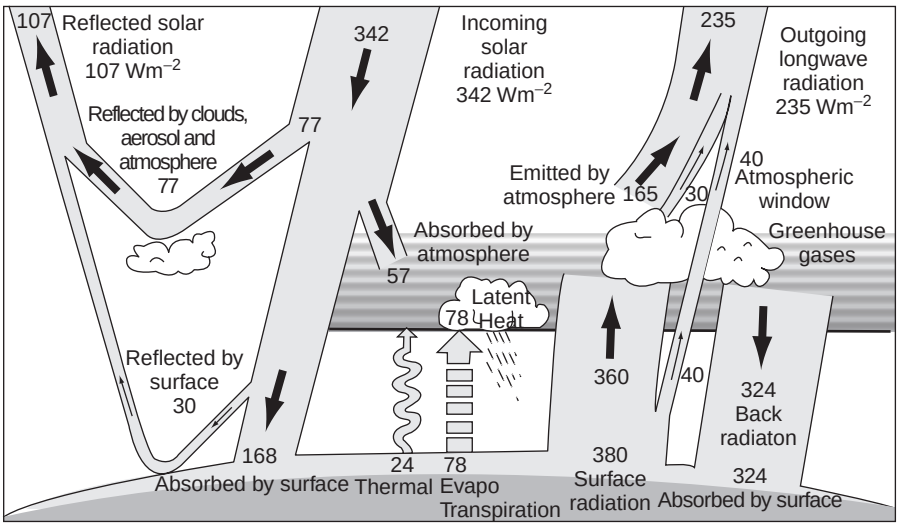


Figure 2 The annual-global mean energy balance of earth. The values given are best estimates of fluxes of energy quoted in $W\,m^{-2}$ (Reproduced from Kiehl and Trenberth, 1997 by permission of American Meteorological Society)

to clouds and the precipitation that returns water back to the surface. As such, the movement of water vapor in the atmosphere by the large-scale windfields of the planet's weather systems grossly influences where clouds form and how much precipitation falls from these systems.

Clouds also affect the planet's energy balance in a manner that is profoundly relevant to climate:

- Clouds reflect more solar radiation to space than the surrounding clear sky thus depleting the amounts of solar energy reaching the surface in these regions. This is sometimes referred to as the *albedo effect*. When viewed from the top of atmosphere (TOA), clouds increase the overall amount of (solar) radiation leaving the planet relative to clear skies.
- Clouds are generally more opaque at infrared wavelengths. As a result, more IR radiation is absorbed in the atmosphere in the presence of clouds than in clear skies. The actual amount of energy emitted to space, however, is decreased (relative to clear skies) because clouds effectively raise the altitude of the level of emission to space to heights characterized by colder temperatures.
- Clouds partition their influence on radiation differently between the atmosphere and surface. Clouds, and high clouds in particular, largely decrease the total atmospheric cooling by increasing the amount of IR energy absorbed from the Earth's surface while producing a compensating cooling at the surface by decreasing the amount of solar radiation that penetrates to the atmosphere.

Analyses of data collected from the Earth-orbiting satellites provide estimates of global distributions of TOA radiative fluxes. Differences between clear and cloudy-sky TOA fluxes (e.g. Harrison *et al.*, 1990) provides a way of estimating the extent that the solar effects of clouds and the long-wave effects of clouds cancel in the sense of the net radiative fluxes formed as the sum of IR and solar fluxes. These satellite observations demonstrate how the short-wave effect generally dominates over the long-wave effect globally. The strongest albedo effect is associated with the persistent maritime stratus off the west coast of continents and storm-track clouds in the summer hemisphere over the mid- and high-latitude Atlantic and Pacific oceans. Although these satellite observations are of immense value in testing global climate models, we are not yet able to determine the separate influences of clouds on the atmosphere and surface without major assumptions.

WATER VAPOR FEEDBACK

In his 1905 correspondence to C.G. Abbott, T.C. Chamberlain notes

> "water vapor, confessedly the greatest thermal absorbent in the atmosphere, is dependent on temperature for its amount, and if another agent, as CO_2, not so dependent, raises the temperature of the surface, it calls into function a certain amount of water vapor which further absorbs heat, raises the temperature and calls forth for more vapor. . .."

Chamberlain's comments provide a clear and concise account of the essential ingredients of the nature of water vapor feedback on the Earth's climate – this feedback requires that the strength of the "thermal absorbent effects" depends in some way on temperature. Although water vapor also absorbs solar radiation, the dominant influence on the energy budget lies in its effects on IR radiation. To first order, this influence is determined by the amount of water vapor integrated through the atmospheric column. Empirical evidence suggests that the mean relative humidity of the atmosphere remains approximately constant and it is through this empirical observation that the water vapor feedback is established (Held and Soden, 2000) since it implies a single, direct relation between column water vapor and temperature.

This constant relative humidity argument is the simple basis of water vapor feedback and more details of the feedback and our current state of its understanding are presented in Held and Soden (2000). According to climate models, this feedback contributes to the major portion of the global warming associated with increasing concentrations of atmospheric CO_2. In reality, however, the water vapor feedback is interfered with by the transport of vapor and by its condensation into clouds – both effects complicate the actual relation between water vapor and temperature. For example, changes to the small amounts of water vapor in the upper atmosphere above 5 km (i.e. in the region of the upper troposphere) produces a disproportionate influence on the radiation emitted to space. Specifically, a 10% change in upper tropospheric absolute humidity has an approximately threefold larger effect on the OLR than does an equivalent percentage change (but significantly larger absolute change) in lower tropospheric water vapor. Unfortunately many factors, over and above temperature, contribute to the observed changes in upper tropospheric water vapor thus greatly complicating our simple ideas of the water vapor feedback. Our ability to understand the specific role of water vapor in these upper layers of the atmosphere continue to be thwarted by our crude ability to observe water vapor at these levels on the global scale.

The effect of water vapor on the formation and evolution of clouds, in turn, adds to the complexity of the water vapor feedback. The actual way clouds interfere with the water vapor feedback process and the specific connection between water vapor, cloudiness, surface temperature, and the greenhouse effect of Earth are still poorly observed and inadequately understood.

CLOUD FEEDBACK

As in Figure 1, feedbacks are generally thought of in terms of a control system and its output (response) given some forcing. Therefore, feedbacks are only meaningful when defined in terms of the output. Although several meaningful measures of output could be considered, it is customary to articulate global climate change and related output in terms of surface temperature. In this context, cloud feedback refers to those cloud processes that in some way depend on temperature and induce a discernible effect on the energy balance that in turn feedback on temperature. Although this simple view of feedback has problems (e.g. Stephens, 2004), it nevertheless serves as a useful framework. We find that generally cloud feedbacks are complex and the processes that establish them are poorly understood. Consequently, cloud feedback now stands as one of the principal obstacles in the way of a meaningful prediction of climate change.

Projecting the salient points noted in Chamberlain's comment onto the topic of cloud feedback exposes a number of important issues that highlight both the essential differences between cloud feedbacks and water vapor feedbacks and offers some explanation for why progress in understanding cloud feedback has been elusive:

1. The "thermal absorbent" character of water is greatly enhanced when water is in its condensed phase. On a molecule-by-molecule basis, water in either the solid or liquid form in the atmosphere absorbs approximately 1000 times more strongly than in the gaseous form. The amount of energy actually absorbed and scattered by clouds also depends on other complicating factors, such as the amounts and vertical organization of clouds and the water contents of clouds and particle sizes, to mention a few.
2. It is reasonably clear what is meant by water vapor in the context of water vapor feedback but the meaning of clouds is much less obvious. As mentioned, a number of cloud properties affect the radiation budget of Earth and all are potentially relevant to the cloud feedback problem. A more extensive list of cloud properties that contribute to the feedback, compared to water vapor, merely reflects the greater complexity of the cloud feedback problem.
3. The dependence of clouds on temperature, and surface temperature specifically, is much less obvious than for water vapor (although complications arise even for water vapor). To first order, this dependence is governed more so by the large-scale motions of the atmosphere and thus on the complex dependence of these motions on temperature. This complexity is arguably the single most important factor to cloud feedback, yet it is generally overlooked or grossly simplified in most cloud feedback studies rendering these inconclusive.
4. Unlike water vapor, clouds also impart an effect on the disposition of solar radiation primarily through their albedo effect. This introduces further complexity to the problem since, when viewed from the TOA, the "thermal absorbent" effects are primarily, although not exactly, offset by the albedo effect in tropical regions. The tendency for long and short-wave processes to approximately balance one another in these regions has confounded our attempts to understand the cloud feedback problem.

These issues and the cloud feedback problem in general can be more succinctly posed in the form of the following questions:

1. Given a fixed distribution of clouds and relevant properties, what is the associated distribution of radiative heating? Our understanding of the radiative transfer processes is sufficiently advanced that the answer to this broad question is straightforward, given the required cloud information. Detailed issues as to how these radiative processes are most efficiently incorporated into large-scale weather and climate models, however, continue to be a focus of research.
2. Given a fixed distribution of heating, what then happens to the distribution of clouds and relevant properties? Our ability to answer this question is far less developed. What is required is a clearer understanding of how the heating and moistening effects of clouds influence the dynamical and thermodynamical properties of the atmosphere, and how these in turn affect the large-scale motions of air and the formation and evolution of clouds. Prediction of the formation and evolution of clouds is poorly handled in models, and our ability to test these predictions with global cloud observations is also limited. A satellite-based program for characterizing global cloudiness that is widely used for this purpose is the International Satellite Cloud Climatology Project (Rossow and Schiffer, 1999).

AEROSOL, CLOUD, AND PRECIPITATION

The atmosphere also contains suspensions of small aerosol particles that not only scatter solar radiation and thus directly affect the radiation balance of the planet (the so-called direct aerosol forcing) but also indirectly influence this energy balance through the effects of aerosol as sources of nucleation of cloud particles. Aerosol particles that nucleate cloud (liquid) droplets or ice crystals are referred to as cloud condensation nuclei (CCN) and ice nuclei (IN) respectively. Not all aerosol particles serve as the nucleus for cloud particle growth, and the ability to activate cloud

particle growth depends on the amount of water vapor in the environment and the chemistry of the nuclei, their size and number concentrations, and other factors.

The concentrations of CCN and IN in air profoundly influence cloud microphysical properties, notably the size and concentrations of cloud particles. Conditions of higher concentrations of aerosol and thus CCN result in clouds composed of smaller but more numerous droplets. The consequences of these changes to cloud microphysics are important to climate for the following reasons:

1. Clouds composed of higher concentrations of smaller droplets reflect more solar radiation than clouds of equivalent water content but composed of a few, larger droplets. This is referred to as the *Twomey effect* (Twomey, 1977). It is observed that mostly shallow clouds, when formed in locally polluted air masses, reflect more radiation than surrounding clouds formed in cleaner air (Figure 3). On the global scale, this effect on the albedo of clouds, although highly uncertain, is thought to be at least as large as the direct effect of aerosol on solar radiation.
2. The Twomey effect described above assumes that the water content does not differ between polluted and unpolluted clouds. This is a dubious assumption since shifting the cloud droplet populations to smaller sizes also inhibits the coalescence rain forming process leading to less precipitation from clouds formed in polluted air masses (Rosenfield, 1999). Thus for the same total mass of water, less precipitation is produced implying a reduction of the precipitation efficiency of clouds and further implying different life-cycle characteristics.

Figure 3 Clouds composed of higher concentrations of small droplets are referred to as *colloidally stable*. For the same total water content, more colloidally stable clouds reflect more solar radiation than do clouds composed of larger droplets. This enhanced reflectivity is observed in this satellite image of clouds off the west coast of California. Shown is an extreme example of the effects of aerosol emitted from the chimney stacks of ships. This higher aerosol air enters the cloud creating local areas of higher CCN, smaller droplets, and higher albedo

The effects of changing aerosol concentrations on clouds causes a chain reaction in a number of processes that affect the microphysics of both clouds and precipitation. The net result of these processes and how they effect different cloud types (such as shallow and deep clouds), and ice (cold) processes versus water (warm) processes are not well understood at this time.

FURTHER READING

Stephens G.L. (1999) Radiative effects of clouds and water vapor. In *Global Energy and Water Cycles*, Browning K.A. and Gurney R.J. (Eds.), Cambridge University Press: pp. 71–90.

Stephens G.L., Slingo A., Webb M.J., Minnett P.J., Daum P.H., Kleiman L., Wittmeyer I. and Randall D.A. (1994) Observations of the Earth's Radiation budget in relation to atmospheric hydrology, Part IV: atmospheric column radiative cooling over the worlds' oceans. *Journal of Geophysical Research*, **99**, 18585–18604.

REFERENCES

Chahine M. (1992) The hydrological cycle and its influence on climate. *Nature*, **359**, 373–378.

Harrison E.F., Minnis P., Barkstrom B.R., Ramanathan V., Cess R.D. and Gibson G.G. (1990) Seasonal variation of cloud radiative forcing derived from the Earth radiation budget experiment. *Journal of Geophysical Research*, **95**, 18687–18703.

Held I. and Soden B.J. (2000) Water vapor feedback and Global warming. *Annual Review of Energy and the Environment*, **25**, 441–475.

Kiehl J.T. and Trenberth K.E. (1997) Earth's annual-global mean energy. *Bulletin of the American Meteorological Society*, **78**, 197.

Rosenfield D. (1999) TRMM observed first direct evidence for smoke from forest fires inhibiting precipitation. *Geophysical Research Letters*, **26**, 3105–3108.

Rossow W.B. and Schiffer R.A. (1999) Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, **80**, 2261–2288.

Stephens G.L. (2005) Cloud feedbacks in the climate system. *Journal of Climate*, **18**, 237–273.

Twomey S. (1977) The influence of pollution on the short-wave albedo of clouds. *Journal of the Atmospheric Sciences*, **34**, 1149–1152.

197: Observed Trends in Hydrologic Cycle Components

HARRY F LINS

United States Geological Survey, Office of Ground Water, Reston, VA, US

Documentation of change in the Earth's climate is accomplished by assessing the rates, magnitude, and distribution of changes in various elements of the climate system, such as the components of the hydrologic cycle. The present section reviews the general character of changes in precipitation, streamflow, and evaporation as determined using systematically collected data through the end of the twentieth century. Precipitation over global land areas increased about 2% during the century, and streamflow also exhibited widespread increases. There was good agreement regionally between the observed precipitation and streamflow increases. The precipitation increases appear to have occurred most commonly in higher intensity categories (>50 mm per day), while the streamflow increases were overwhelmingly observed in the low to moderate range of flows. No systematic increases were observed in peak streamflows. These findings indicate that a general intensification of the hydrologic cycle occurred during the twentieth century, but that this intensification did not result in increased hydrologic extremes.

INTRODUCTION

Climate change is documented by assessing the rates, magnitude, and distribution of changes in climate system components. Although simple in concept, such documentation has proven to be problematic in practice. Existing climate observing systems are capable of only partially answering critical questions associated with climate change, particularly for such hydrologic cycle variables as precipitation, runoff, and evaporation (NRC, 1999). Changes in station location, time of observation, and monitoring equipment, as well as difficulties in operating monitoring stations over multidecadal periods and at many locations worldwide, are all factors that contribute to this problem. Significantly, however, the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, 2001) notes that the "certainty of conclusions that can be drawn about climate from observations depends critically on the availability of accurate, complete and consistent series of observations".

In addition, human alteration of the landscape and riverine systems can significantly affect hydrologic cycle changes, from small watershed to large river basin scales (Shiklomanov and Penkova, 2003). The data needed to account for such confounding effects, such as consumptive water use, reservoir storage, and land use change, are often less well measured or more difficult to obtain than those for precipitation, streamflow, and evaporation. Remotely sensed data are finding increasing use in hydrologic cycle investigations, most particularly in regions of the world where systematic *in situ* data collection networks are sparse or nonexistent. However, such systems are not without their own problems. In many instances, remotely sensed data are less than 20 years in length, limiting their utility for trend assessment, and they generally do not measure quantities that are directly comparable to ground-based observations.

Trend analysis results are also very sensitive to the conditions that existed at the endpoints of the time series. Differences of one or two years in the start or end points in a data time series, even when the time series are otherwise identical, can substantially alter results. Similarly, differences in analytical methods among published studies can compromise the significance of results and complicate summary assessments of change. Douglas *et al.* (2000) noted, for example, that spatial correlation among streamflow observing stations markedly affects the results of

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

trend testing. They found that regional cross correlation of flow records dramatically reduced the effective number of samples available for trend assessment, and that by not taking this influence into account, erroneous conclusions would be drawn with respect to the absence or presence of regional trends.

Despite the many problems associated with the collection and analysis of the hydrologic cycle and related data, studies aimed at identifying trends in hydrologic cycle components at regional to continental scales are increasing. The most extensive work has addressed precipitation trends. Precipitation data have been collected longer and at more sites worldwide than have streamflow and evaporation data. It should be noted, however, that despite its breath of monitoring, measurement problems and bias continue to adversely affect the quality of precipitation data. Even so, the density of land precipitation monitoring networks is sufficient to facilitate global assessments of trend, such as those reported in recent climate change assessments by the IPCC (1996, 2001).

In contrast, investigations of runoff or streamflow trends have been more geographically restricted, with the most comprehensive analyses covering Canada, Europe, the United States, and to a lesser extent, Australia and South America. Comparatively, little streamflow data are available for Africa and Asia and record lengths in these regions tend to be relatively short. Another confounding issue in assessing streamflow trends relates to the influence of human activities on rivers and streams. Flows on most gauged watercourses are modified to some extent by human activities. Humans appropriate more than 50% of accessible renewable water resources globally (Postel *et al.*, 1996). By the late 1980s, there were more than 36 000 large dams worldwide, representing a 700% increase in the standing stock of natural river water (Vörösmarty *et al.*, 1997). In the contiguous United States alone, of the 5.2 million km of rivers, there are only 42 reaches with lengths greater than 200 km that are free flowing (Benke, 1990).

Avoiding potential problems in understanding how the streamflow component of the hydrologic cycle is changing through time is most easily achieved by using streamflow records that reflect natural or near-natural conditions. To this end, some hydrologists and hydrological services have identified climate-sensitive stations within their monitoring networks that reflect unimpaired basin conditions. Unimpaired generally means that there is no overt adjustment of natural streamflow by diversion or augmentation, regulation of the watercourse by a containment structure, or reduction of base flow by groundwater pumping. In practical terms, unimpaired records are generally considered to be those where the degree of human activity in the watershed is small enough so as not to affect significantly the value of monthly mean discharge as computed on the basis of daily mean discharge. The Reference Hydrometric Basin Network (RHBN), assembled by Environment Canada (Harvey *et al.*, 1999), and the Hydro-climatic Data Network (HCDN) of the US Geological Survey (Slack and Landwehr, 1992) are examples of streamgauging networks that have been identified as meeting specific climate sensitivity criteria.

Studies of trends in evaporation, as with streamflow, are limited geographically. Nearly all are based on pan evaporation measurements because pans have been systematically used longer and in more locations than other types of equipment. Historically, pan evaporation measurements have been viewed as an index of potential evaporation or the evaporation that occurs where there is an unlimited supply of water. Recently, however, Golubev *et al.* (2001) developed a method for estimating actual evaporation from the land surface using pan evaporation measurements, a more meaningful quantity when comparing with precipitation and streamflow. Despite this development, it should be noted that pan evaporation estimates are problematic at best. Pan evaporation must be adjusted by seasonally varying coefficients in order to provide estimates of lake evaporation, and these coefficient adjustments are only approximations (Mather, 1974).

A review of the primary recent literature documenting changes in hydrologic cycle components follows. This review does not purport to establish invariant and regionally detailed trends worldwide. Rather, its purpose is to provide a *snapshot* or generalized characterization of changes in precipitation, streamflow, and evaporation as determined using systematically collected data that end around the year 2000. The dynamic nature of hydrologic cycle variables, coupled with the analytical and interpretive limitations associated with monotonic trend tests, constrain the certainty and enduring relevance of specific results.

CHANGES IN PRECIPITATION

The IPCC third assessment (IPCC, 2001) reported that global land precipitation increased about 2% during the twentieth century, but that the increase was not uniform spatially or temporally. This is evident in the seasonal precipitation change maps in Figure 1. In the middle and high latitudes of the Northern Hemisphere, for example, there have been widespread precipitation increases, particularly during the autumn months. However, precipitation decreases have been observed in Europe, Western Russia, and around the Mediterranean. In Canada, precipitation increased by an average of more than 10% during the twentieth century (Mekis and Hogg, 1999), including increases in snowfall (Zhang *et al.*, 2000). Farther south, in the United States, precipitation increased between 5 and 10% over the century. Karl and Knight (1998) reported an increase in the United States of about 10% nationwide,

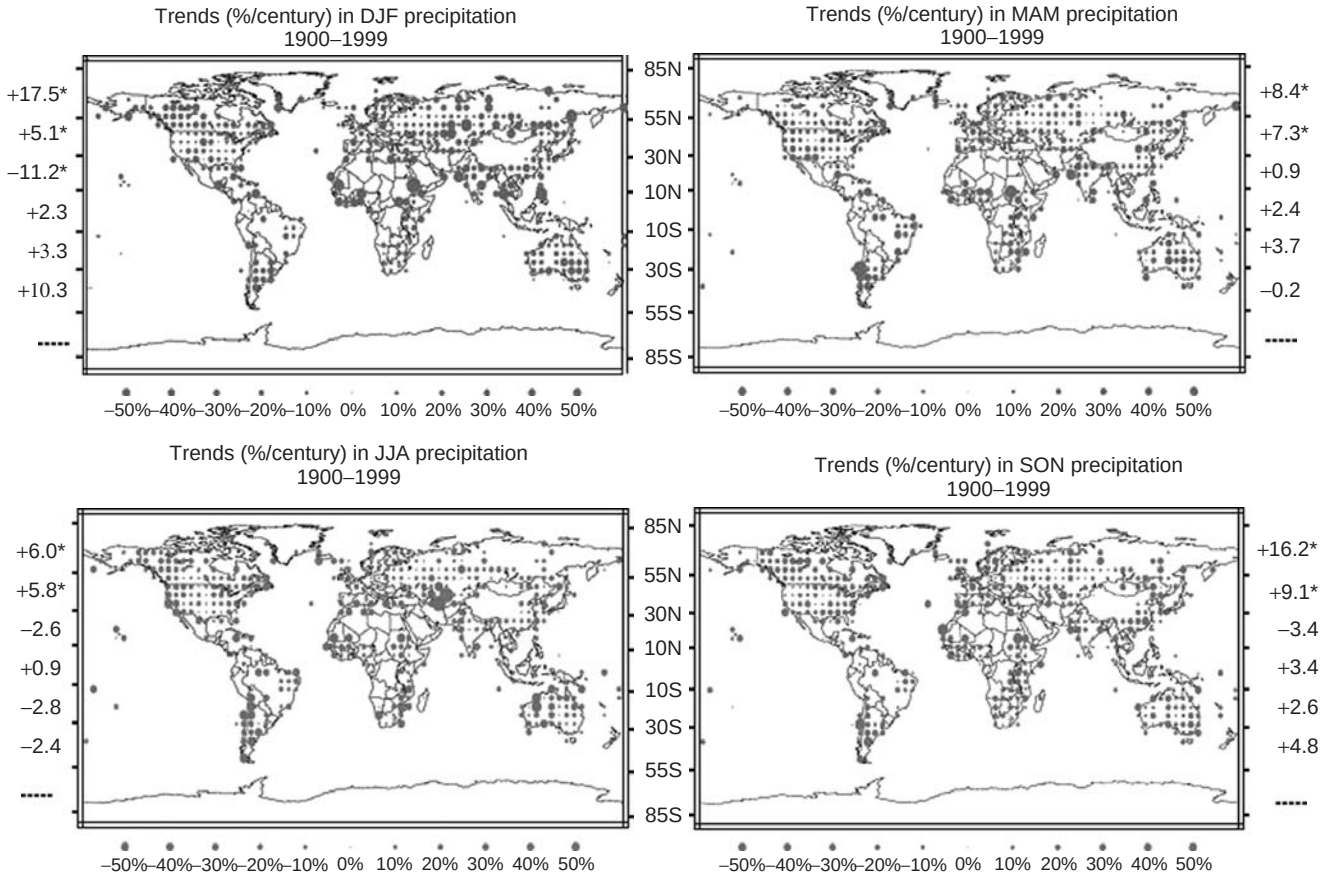


Figure 1 Trends in seasonal precipitation, 1900–1999. Trend magnitude is signified by the size of the circle (*Source:* Intergovernmental Panel on Climate Change (IPCC), 2001). Trend direction is represented by the color of the circle which can be seen in the color version of this image that is available at http://www.mrw.interscience.wiley.com/ehs

and that more than half of this increase was in heavier precipitation categories, that is, the upper 10 percentiles of the precipitation distribution. They also noted significant variability in the pattern of trends regionally and seasonally, with the increases being confined to the spring, summer, and autumn months. Groisman *et al.* (2004) confirmed this finding, adding that the largest trends were observed in the eastern two-thirds of the United States, and primarily in the warm season when intense rainfall events are most frequent. However, Kunkel *et al.* (2003) pointed out that during the late nineteenth century, intense precipitation events were nearly as high as during the late twentieth century. This suggests that the reporting of trends in precipitation since about 1900 may not encompass the full range of natural variability that has characterized the climate system in recent centuries.

During the second half of the twentieth century, annual precipitation decreased slightly in China, although increases were observed over the middle and lower Yangtze River basin (Zhai *et al.*, 1999a,b). Northern Europe and Scandinavia saw increased precipitation, while southern Europe, down to the Mediterranean, experienced a general decrease over these same decades. Coupling these results with those for North America and averaging zonally, annual precipitation increased between 7 and 12% for the zones from 30 °N to 85 °N (IPCC, 2001).

In the Northern Hemisphere subtropical and tropical zones, precipitation tended to decrease during all seasons in the twentieth century. This was particularly true in North Africa and the Asian subcontinent. Kumar *et al.* (1999a,b) found no evidence of a long-term trend in Indian monsoonal rainfall, although they did note significant multidecadal variations.

In the Southern Hemisphere, precipitation trends have also been mixed. Annual rainfall over most of Australia has increased, as have the number of rain days, but precipitation decreases during winter (June, July, and August) have been notable in the eastern and western thirds of the continent. Seasonal differences in trends have also characterized the changes in southern Africa, where warm season precipitation (December–May) has decreased and cool season precipitation (June–November) has increased. Precipitation increases have occurred in most of South America in all seasons with two notable exceptions, Chile and eastern

Brazil, where decreases were dominant during the twentieth century.

It is important to note that most of these reported changes in precipitation are within the measurement error that has been documented for precipitation. UNESCO (1978) and Legates (1987) estimate the bias in precipitation measurements, averaged globally, to be about 11%. Legates (1995) further argued that local increases in air temperature and decreases in wind speed, resulting from climatic variability or from local urbanization, can introduce spurious trends into the precipitation record. This is because the bias in measuring liquid precipitation is lower than for solid precipitation, as it is for lower wind speeds. Moreover, spurious trends can be introduced through monitoring station relocations and instrument changes (Groisman, 1991). It is likely that these factors have contributed to the variable and occasionally contradictory precipitation trend results that have been published.

CHANGES IN STREAMFLOW

Global assessments of trends in streamflow have proven to be more elusive, primarily because network limitations, data access, and data quality issues have hindered attempts at producing a unified global synthesis. However, several studies have provided noncomprehensive, low spatial density looks at worldwide trends in streamflow. Chiew and McMahon (1996) tested for trends in annual streamflow volumes and peak discharges at 142 stations on 6 continents. They found no consistent pattern of widespread trends in either variable, although most of their data ended before 1980. More recently, Kundzewicz *et al.* (2004) and Svensson *et al.* (2004) analyzed trends in annual maximum flow, and in peaks over threshold and annual low flows, respectively. The Kundzewicz *et al.* (2004) results utilized records from 195 stations on 6 continents, and are summarized in Table 1. The overwhelming majority of stations (92%) were located in North America, Europe, and Australia. Among all stations, 70% had no trend in the annual maximum flow, 14% had an increasing trend, and 16% a decreasing trend. Although the actual percentages varied from continent to continent, each exhibited a pattern of trends generally consistent with the aggregate totals. Using a much smaller sample of stations (21), Svensson *et al.* (2004) found a mixed pattern of trends in peaks over threshold, with approximately 30% of the stations exhibiting a trend, and with more downward trends than upward. A very different pattern was apparent in low flows, however (Figure 2).

Table 1 Trends ($p \leq 0.10$) in annual maximum streamflow, by continent, for 195 streamgauging stations worldwide

Region	Number of stations	Number with increasing trend	Number with no trend	Number with decreasing trend
Africa	4	1	1	2
Asia	8	0	5	3
South America	3	0	3	0
North America	70	14	44	12
Australia–Pacific	40	1	34	5
Europe	70	11	50	9
Totals	195	27(14%)	137(70%)	31(16%)

Source: Kundzewicz *et al.*, 2004.

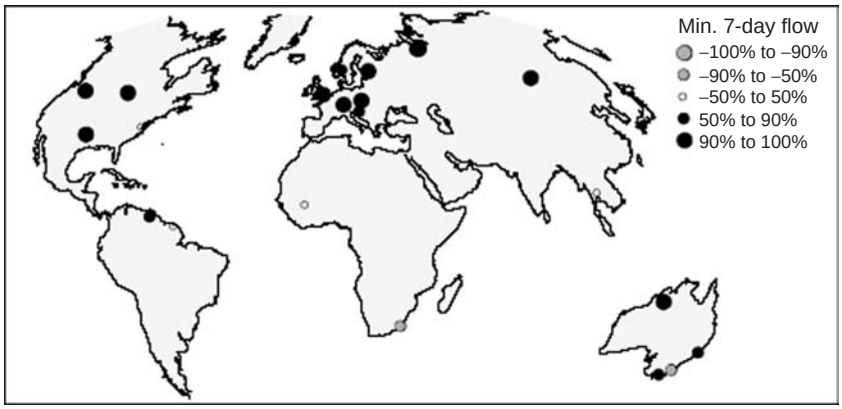


Figure 2 Trends in the annual minimum 7-day mean flow series at 21 stations worldwide. Negative trends are shown as gray circles and positive trends as black circles. The largest circles identify trends significant at $p \leq 0.10$ (*Source:* Svensson *et al.*, 2004)

Approximately 52% of the stations had statistically significant trends in seven-day low flows, and all of these were increasing trends. This implies a reduction in the incidence of hydrologic drought. Importantly, these broad-scale study results are very consistent with those from regionally specific investigations.

The most comprehensive documentation of streamflow change comes from regional studies and, of these, the most complete assessment exists for North America, where several national-scale studies have been published for Canada and the United States. Zhang *et al.* (2001) found that annual mean streamflow generally decreased in Canada between 1947 and 1996, with significant decreases occurring in the southern part of the country, particularly southern British Columbia and Alberta. They also found decreases in monthly mean streamflow for most months of the year except March and April when significant increases were observed. Changes in the frequency distribution of daily streamflow, from low to high, were also evaluated. In southern Canada, as with the annual mean, significant decreases were observed in all percentiles of daily flow. Over northern British Columbia and the Yukon Territory, however, significant increases were identified in the lower flow percentiles. Zhang and his collaborators also noted that the breakup of river ice and the ensuing spring freshet were occurring earlier, especially in British Columbia, and that river freeze-up appeared to be occurring earlier in the autumn in eastern Canada.

In the United States, numerous investigations have found a consistent pattern of streamflow increases across much of the country during the twentieth century. Lettenmaier *et al.* (1994) identified strong increases in monthly mean streamflow in the months from November to April for the years 1948–1988. The largest trend magnitudes were observed in the north-central states. This study also found streamflow decreases in the Pacific Northwest that were consistent with the streamflow decreases in the adjacent provinces of southwestern Canada noted by Zhang *et al.* (2001).

Lins and Michaels (1994) assessed trends in streamflow in the United States by region and by month. The regions were defined by principal components analysis of monthly mean streamflow. A separate analysis was performed for each calendar month. The resulting time series of component scores for each component and month were then tested for trend. Nearly all regions of the country experienced increasing streamflows between 1941 and 1988, but the significant trends were only observed during the autumn to early winter months. No regional trends were observed in the Pacific Northwest and, as Lettenmaier *et al.* found, the north-central region had the strongest trend.

During the 1990s, a series of major flooding events in the United States received significant attention and led to speculation that extreme hydrologic events (including both floods and droughts) were increasing, possibly in response to greenhouse warming. Lins and Slack (1999) evaluated this possibility by testing for trends in the percentiles of annual (on the basis of daily mean) streamflow for periods ranging from 30 to 80 years. They found that trends were most prevalent in the lower half of the frequency distribution, from the annual minimum to median flow, and that the trend was upward at 40–50% of the stations tested. In contrast, they reported a decline in the number of stations having trends in the upper half of the distribution, with the annual maximum flow reporting the least increasing trends (at 10% of the stations). The streamflow increases were observed across much of the United States, but particularly the northeastern quarter of the country, while decreases were detected in most percentiles in the Pacific Northwest and in the Southeast.

Douglas *et al.* (2000) also looked at trends in high and low flows using a regional trend test. They found no evidence for a coherent trend in high flows regionally, but did report upward trends in low flows in much of the northeastern quarter of the United States, supporting the Lins and Slack findings. A subsequent study by Groisman *et al.* (2001), however, appeared to arrive at a different conclusion with respect to trends in high flow. Working with a subset of the same data used in previous streamflow trend analyses, Groisman and his colleagues reported that the largest streamflow increase in the United States occurred in the highest streamflow percentiles, seemingly contradicting all previous work. Upon closer examination, however, what Groisman found was not that there were more stations having trends in the highest streamflow percentiles, or that the magnitude of the trend in percentage terms was greater. Their analysis actually yielded the following result: if one calculated the total volume of water that increased (or decreased) from 1939 to 1999, and then determined what proportion of that increase was produced by increases in the 0–5th percentile bin, 6–10th percentile bin, and so forth up to the 96–100th percentile bin, the bin that contributed most of the increase is the 96–100th percentile bin.

Importantly, this result is not inconsistent with, or contradictory to, the findings of previous studies that found relatively few stations with trends in the annual maximum streamflow. The reason for this is the quantity of water contained in each percentile bin of the annual streamflow distribution. Because streamflow is a lognormally distributed variable, where the annual maximum value is typically two or more orders of magnitude greater than the annual minimum value, the uppermost percentiles correspond to very large quantities of water relative to other parts of the distribution. So, even though the percentage increase in the annual maximum streamflow is relatively small (as noted by Lins and Slack and Douglas *et al.*), the corresponding volume increase can be very large. In other words, a small percentage change in a very large volume of water is a

large value in comparison to a large percentage change in a relatively small volume. Thus, the differences between the Groisman *et al.* findings and those of the others are apparent and interpretive rather than substantive. This situation does, however, underscore the problems that can attend trend assessments on the basis of different analytical approaches.

Finally, McCabe and Wolock (2002) found a significant increase in annual minimum and median daily streamflow around 1970, and a less significant mixed pattern of increases and decreases in annual maximum daily streamflow. These changes were primarily observed in the eastern United States and are consistent with previous studies. Notably, though, McCabe and Wolock observed that the streamflow increases appeared as a step change rather than as a gradual trend, which has important implications (Figure 3). The inference drawn from a gradual trend is that

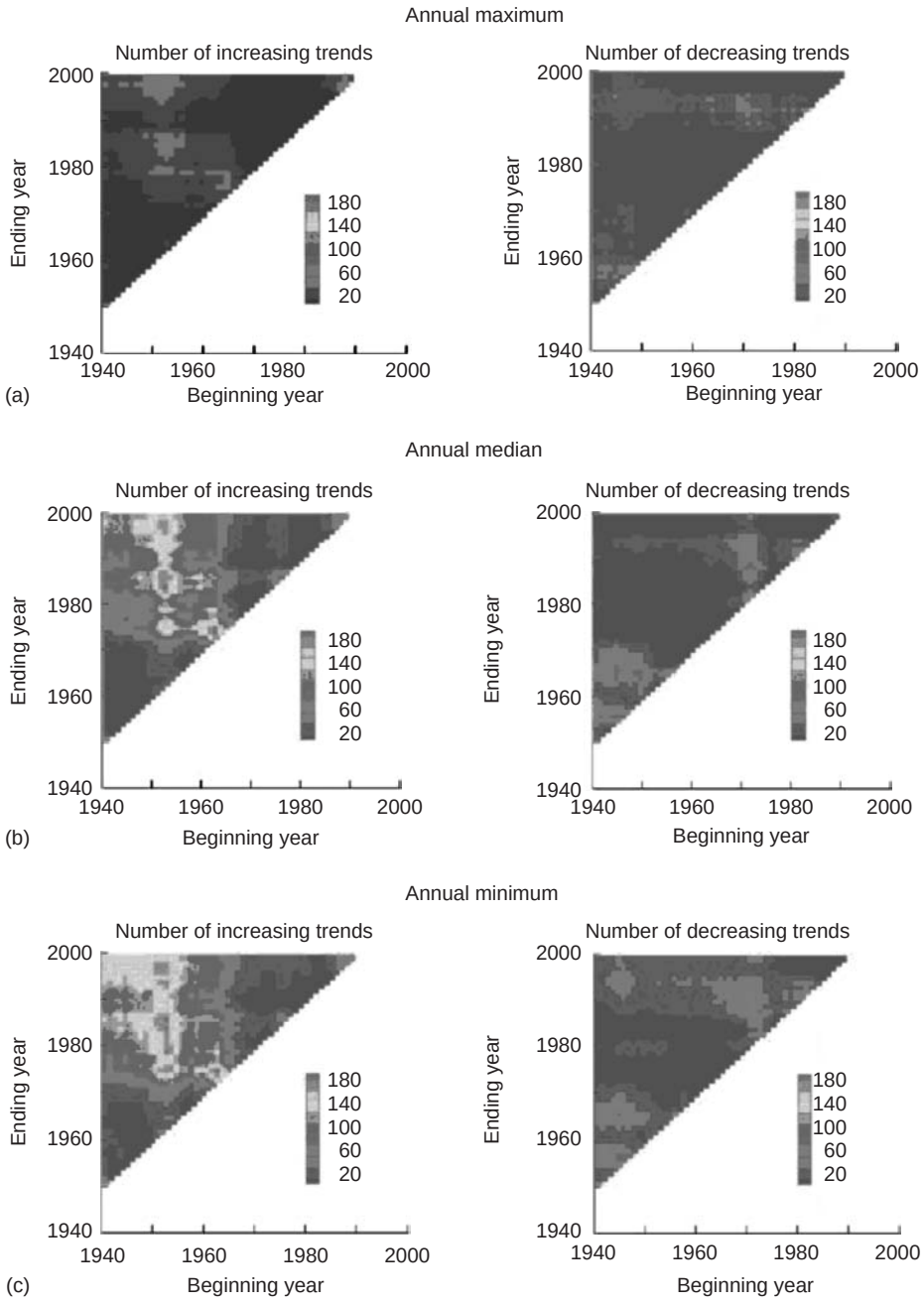


Figure 3 Number of sites with significant ($p \leq 0.05$) increasing and decreasing trends in (a) annual maximum, (b) median, and (c) minimum daily streamflow for various periods at least 10 years in length, and for 400 stations in the United States during 1941–1999 (Reproduced fromMcCabe and Wolock, (2002) by permission of American Geophysical Union (AGU)). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

it is likely to continue into the future, while the implication of a step change is that the climate system has shifted to a new regime that will likely remain stable until a new shift occurs.

Fewer studies of streamflow trends have been published for other continental areas. In South America, discharge data for several major rivers in the southeastern part of the continent covering the period 1901–1995 indicate that streamflow increased after the mid-1960s (Garcia and Vargas, 1998; Genta *et al.*, 1998). This increase was also accompanied by a decrease in the amplitude of the seasonal cycle in most rivers.

In Europe, much of the published work on streamflow trends has focused specifically on flooding. Robson *et al.* (1998) and Robson (2002) evaluated local and national flood series in the United Kingdom, and the effect of climatic variability on trend detection. Their results indicated that more protracted episodes of high flow have occurred during the second half of the twentieth century. However, they found no statistical evidence of a long-term (80–120 years) trend in flooding. An analysis of systematic flood records for winter and summer seasons in central Europe since the middle of the nineteenth century, and longer-term historical records of major floods since the sixteenth century, was performed by Mudelsee *et al.* (2003). They found a decrease in the occurrence of winter floods on the Elbe and Oder rivers over an 80- to 150-year period with no trend in summer flooding. They attribute the winter season decrease in part to a decline in strong freezing events that reduce late winter ice jamming events and consequent higher flood peaks. This study also detected significant long-term changes in flood occurrence between the sixteenth and nineteenth centuries, but concluded that reductions in river length, construction of reservoirs, and deforestation had minor effects on flood frequency. One study, by Kahya and Kalayci (2004), focused on trends in average flow conditions. Using monthly mean streamflow records for 26 basins in Turkey from 1964 to 1994, they found that streamflow had generally decreased in western and southern Turkey, while not changing significantly in eastern Turkey. Finally, Hisdal *et al.* (2001) tested for trends in hydrologic drought using a pan-European data set of more than 600 daily streamflow records encompassing four time periods: 1962–1990, 1962–1995, 1930–1995, and 1911–1995. For most stations tested, no significant changes were detected, and the authors concluded that there was no evidence to indicate that drought conditions, in general, had become more severe or frequent.

CHANGES IN EVAPORATION

Most work on trends in evaporation has been done using pan measurements from the former Soviet Union and the United States. The IPCC Second Assessment Report (1996) described widespread decreases in pan evaporation during the twentieth century using these data. Subsequent to the Second Assessment, several authors noted an inconsistency between the reported decreases in pan evaporation, which was interpreted as a decrease in actual land surface evaporation, and observed increases in both temperature and precipitation in Russia and the United States. Brutsaert and Parlange (1998), for example, noted that the contradiction in the trends of evaporation, temperature, and precipitation was difficult to reconcile in the context of a general intensification of the hydrological cycle over northern extratropical land areas. Lawrimore and Peterson (2000) and Golubev *et al.* (2001) conducted additional studies and arrived at conclusions similar to Brutsaert and Parlange.

Golubev *et al.* (2001) developed a procedure for estimating actual land surface evaporation from pan evaporation measurements using coupled observations of both actual and pan evaporation at multiple field sites in Russia. In applying this procedure to pan data from Russia and the United States, they concluded that the actual evaporation increased over most arid regions of both countries, as well as over humid maritime regions of the eastern United States during the warm season. They also found that the actual evaporation decreased over heavily forested areas of Russia and the northern United States.

SUMMARY AND CONCLUSION

Analyses of trends in the hydrologic cycle indicate that precipitation over global land areas increased about 2% during the twentieth century, and that the streamflow also exhibited widespread increases. There was generally good agreement regionally between the observed trends in precipitation and streamflow. Moreover, some investigations that reported precipitation increases also found that these increases occurred more frequently in higher intensity categories (e.g. >50 mm per day). Notably, in regions where such precipitation increases were observed, there appeared to be increases in low to moderate streamflows. There is no evidence of widespread or systematic increases in peak streamflows, although there is widespread evidence of increases occurring in annual low flows. This pattern of trends indicates that a general intensification of the hydrologic cycle occurred during the twentieth century, but contrary to the hypothesis that such an intensification would result in increased hydrologic extremes (i.e. floods and droughts), it did so in a more benign and beneficial manner.

REFERENCES

Benke A.C. (1990) A perspective on America's vanishing streams. *Journal of the North American Benthological Society*, **9**, 77–88.

Brutsaert W. and Parlange M.B. (1998) Hydrological cycle explains the evaporation paradox. *Nature*, **396**, 30.

Chiew F.H.S. and McMahon T.A. (1996) Trends in historical streamflow records. In *Regional Hydrological Response to Climate Change*, Jones J.A.A., Liu C.M., Woo M.-K. and Kung H.-T. (Eds.), Kluwer Academic Publishers: Amsterdam, pp. 63–68.

Douglas E.M., Vogel R.M. and Kroll C.N. (2000) Trends in floods and low flows in the United States: impact of spatial correlation. *Journal of Hydrology*, **240**, 90–105.

Garcia N.O. and Vargas W.M. (1998) The temporal climatic variability in the 'Rio de la Plata' Basin displayed by the river discharges. *Climatic Change*, **38**, 359–379.

Genta J.L., Perez-Iribarren G. and Mechoso C.R. (1998) A recent increasing trend in the streamflow of rivers in southeastern South America. *Journal of Climate*, **11**, 2858–2862.

Golubev V.S., Lawrimore J.H., Groisman P.Y.a, Speranskaya N.A., Zhuravin S.A., Menne M.J., Peterson T.C. and Malone R.W. (2001) Evaporation changes over the contiguous United States and the Former Soviet Union: a reassessment. *Geophysical Research Letters*, **28**, 2665–2668.

Groisman P.Y. (1991) Unbiased estimates of precipitation change in the Northern Hemisphere extratropics, *Proceedings of the Fifth Conference on Climate Variations, American Meteorological Society*, Denver.

Groisman P.Y., Knight R.W. and Karl T.R. (2001) Heavy precipitation and high streamflow in the contiguous United States: trends in the 20th century. *Bulletin of the American Meteorological Society*, **82**, 219–246.

Groisman P.Y., Knight R.W., Karl T.R., Easterling D.R., Sun B. and Lawrimore J.H. (2004) Contemporary changes of the hydrological cycle over the contiguous United States: trends derived from *in situ* observations. *Journal of Hydrometeorology*, **5**, 64–85.

Harvey K.D., Pilon P.J. and Yuzyk T.R. (1999) *Canada's Reference Hydrometric Basin Network (RHBN): In partnerships in Water Resource Management*, Paper presented at CWRA 51th Annual Conference, Canadian Water Resources Association: Halifax.

Hisdal H., Stahl K., Tallaksen L.M. and Demuth S. (2001) Have streamflow droughts in Europe become more severe or frequent? *International Journal of Climatology*, **21**, 317–333.

IPCC (1996) *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, Houghton J.T., Meira Filho L.G., Callander B.A., Harris N., Kattenberg A. and Maskell K. (Eds.), Cambridge University Press: Cambridge, p. 572.

IPCC (2001) *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Houghton J.T., Ding Y., Griggs D.J., Noguer M., van der Linden P.J., Dai X., Maskell K. and Johnson C.A. (Eds.), Cambridge University Press: Cambridge, p. 881.

Kahya E. and Kalayci S. (2004) Trend analysis of streamflow in Turkey. *Journal of Hydrology*, **289**, 128–144.

Karl T.R. and Knight R.W. (1998) Secular trends of precipitation amount, frequency, and intensity in the United States. *Bulletin of the American Meteorological Society*, **79**, 231–241.

Kumar K.K., Kleeman R., Crane M.A. and Rajagopalan B. (1999a) Epochal changes in Indian monsoon-ENSO precursors. *Geophysical Research Letters*, **26**, 75–78.

Kumar K.K., Rajagopalan B. and Crane M.A. (1999b) On the weakening relationship between the Indian monsoon and ENSO. *Science*, **284**, 2156–2159.

Kundzewicz Z.W., Graczyk D., Maurer T., Przymusińska I., Radziejewski M., Svensson C. and Szwed M. (2004) *Detection of Change in World-Wide Hydrological Time Series of Maximum Annual Flow*, World Climate Applications and Services Programme Report 64, WMO/TD-No. 1239, WMO, p. 35.

Kunkel K.E., Easterling D.R., Redmond K. and Hubbard K. (2003) Temporal variations of extreme precipitation events in the United States: 1895–2000. *Geophysical Research Letters*, **30**, 1900–1903, doi:10.1029/2003GL018052.

Lawrimore J.H. and Peterson T.C. (2000) Pan evaporation trends in dry and humid regions of the United States. *Journal of Hydrometeorology*, **1**, 543–546.

Legates D.R. (1987) A climatology of global precipitation. *Publications in Climatology*, **40**, 84.

Legates D.R. (1995) Precipitation measurement biases and climate change detection, *Proceedings of the Sixth Symposium on Global Change Studies. American Meteorological Society*, Dallas, pp. 168–173.

Lettenmaier D.P., Wood E.F. and Wallis J.R. (1994) Hydro-climatological trends in the continental United States, 1948–1988. *Journal of Climate*, **7**, 586–607.

Lins H.F. and Michaels P.J. (1994) Increasing U.S. streamflow linked to greenhouse forcing. *EOS*, **75**, 281,284–285.

Lins H.F. and Slack J.R. (1999) Streamflow trends in the United States. *Geophysical Research Letters*, **26**, 227–230.

Mather J.R. (1974) *Climatology: Fundamentals and Applications*, McGraw-Hill: New York, p. 412.

McCabe G.J. and Wolock D.M. (2002) A step increase in streamflow in the conterminous United States. *Geophysical Research Letters*, **29**, 2185–2188.

Mekis E. and Hogg W.D. (1999) Rehabilitation and analysis of Canadian daily precipitation time series. *Atmosphere-Ocean*, **37**, 53–85.

Mudelsee M., Borngen M., Tetzlaff G. and Gr–newald U. (2003) No upward trends in the occurrence of extreme floods in central Europe. *Nature*, **425**, 166–168.

NRC (1999) *Adequacy of Climate Observing Systems*, National Academy Press: Washington, p. 51.

Postel S.L., Daily G.C. and Ehrlich P.R. (1996) Human appropriation of renewable fresh water. *Science*, **271**, 785–788.

Robson A.J. (2002) Evidence for trends in UK flooding. *Philosophical Transactions of the Royal Society of London. Series A*, **360**, 1327–1343.

Robson A.J., Jones T.K., Reed D.W. and Bayliss A.C. (1998) A study of national trend and variation in UK floods. *International Journal of Climatology*, **18**, 165–182.

Shiklomanov I.A. and Penkova N.V. (2003) Methods for assessing and forecasting global water use and water availability. In *World Water Resources at the Beginning of the 21st Century*, Shiklomanov I.A. and Rodda J.C. (Eds.), Cambridge University Press: Cambridge, pp. 27–44.

Slack J.R. and Landwehr J.M. (1992) Hydro-Climatic Data Network: a U.S. Geological Survey Streamflow Data Set for the United States for the Study of Climate Variations, 1874–1988, U.S. Geol. Surv. Open-File Rept. 92–129, U.S. Geological Survey, p. 193.

Svensson C., Kundzewicz Z.W. and Maurer T. (2004) *Trends in Flood and Low Flow Hydrological Time Series*, World Climate Applications and Services Programme Report 66, WMO/TD-No. 1241, WMO, p. 26.

UNESCO (1978) *World Water Balance and Water Resources of the Earth*, UNESCO Series Studies and Reports in Hydrology, No. 25, UNESCO, Leningrad, p. 663.

Vörösmarty C.J., Sharma K.P., Fekete B.M., Copeland A.H., Holden J., Marble J. and Lough J.A. (1997) The storage and aging of continental runoff in large reservoir systems of the world. *Ambio*, **26**, 210–219.

Zhai P.M., Ren F.M. and Zhang Q. (1999b) Detection of trends in China's precipitation extremes. *Acta Meteorologica Sinica*, **57**, 208–216.

Zhai P.M., Sun A., Ren F., Liu X., Gao B. and Zhang Q. (1999a) Changes of climate extremes in China. *Climatic Change*, **42**, 203–218.

Zhang X., Harvey K.D., Hogg W.D. and Yuzyk T.R. (2001) Trends in Canadian streamflow. *Water Resources Research*, **37**, 987–998.

Zhang X., Vincent L.A., Hogg W.D. and Niitsoo A. (2000) Temperature and precipitation trends in Canada during the 20th century. *Atmosphere-Ocean*, **38**, 395–429.

198: Role and Importance of Cryospheric Processes in Climate System

JOSEPH R MCCONNELL

Division of Hydrologic Sciences, Desert Research Institute, University and Community College System of Nevada, Reno, NV, US

The cryosphere, which includes seasonal snow, frozen ground, sea ice, glaciers, ice caps, and ice sheets, is one of Earth's five main regimes together with the atmosphere, biosphere, hydrosphere, and pedosphere. It is estimated that the cryosphere contains about 1.8% of all water on Earth but nearly 70% of the freshwater. Interactions between snow and ice cover and the radiation budget make the cryosphere especially important to climate at local to global scales, while strong positive feedback mechanisms mean that the cryosphere is particularly sensitive to climate change. Impurities trapped in glaciers and ice sheets archive detailed, high-time resolution records of the Earth's changing climate and hydrologic cycle over the past million years. These and other records document repeated variations in the size of the cryosphere including very large, long-duration changes such as the glacial and interglacial periods as well as much smaller, though still important changes such as the Medieval Warm Period and Little Ice Age during the last two millennia. The cryosphere is shrinking rapidly in response to the current warming, particularly in the Northern Hemisphere.

INTRODUCTION

The Earth's environment can be divided into five main regimes: atmosphere, biosphere, cryosphere, hydrosphere, and pedosphere. The cryosphere is the frozen portion and it includes snow, frozen ground, sea ice, glaciers, ice caps, and ice sheets. The cryosphere is an integral part of the hydrologic cycle. It is estimated that the cryosphere contains about 1.8% of all water on Earth but nearly 70% of the freshwater. Annual to millennial scale storage of water in glaciers and ice sheets impacts sea level, while seasonal to annual storage of water in mountain snow packs regulates runoff and river flows.

Snow and ice cover at high latitudes dramatically increase albedo there and so enhance poleward temperature gradients. These temperature gradients play an important role in the movement of heat and energy around the globe through atmospheric and oceanic circulation, with net heat transport from the low to high latitudes. Without such transport, the equatorial regions would be 14 °C warmer and the North Pole 25 °C colder (Barry and Chorley, 1992). Because exchanges between the cryosphere and the rest of the hydrologic cycle depend on the change of water to and from the solid phase (ice and snow) and because much of the cryosphere exists near the melting point, the size and behavior of the cryosphere is highly sensitive to climate change. Decreased snow and sea ice cover decrease albedo and so increase radiative heating, resulting in a positive feedback and amplification of global climate change in the polar regions and at higher elevations.

Impurities trapped in glaciers and ice sheets, together with the thickness of snow and ice in annual layers, comprise arguably the most detailed and direct archive of paleoclimatic and paleoenvironmental information available. These archives document changes in the size of the cryosphere over the past million years, including large increases such as the glacial periods during the Pleistocene. The sensitivity of the cryosphere is demonstrated by recent smaller but significant increases in the size of the cryosphere such as the Little Ice Age from A.D. 1400 to 1650 when temperatures were only about 1 °C cooler than it is presently (Graedel and Crutzen, 1995).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

RESERVOIRS IN THE CRYOSPHERE

The cryosphere consists of a number or reservoirs that, in sum, contain about 1.8% of all the water on earth and about 70% of all freshwater resources. The reservoirs include seasonal snow, frozen ground, floating ice such as sea, lake and river ice, glaciers, ice caps, and ice sheets.

In terms of area, seasonal snow cover is the largest component of the cryosphere, covering up to 33% of the Earth's total land surface with about 98% of the total seasonal snow cover located in the Northern Hemisphere. Seasonal snow packs accumulate winter precipitation that is subsequently released through melt, thereby generating surface runoff during the spring and summer when it is needed for agriculture. In many semiarid and arid regions, as much as 90% of the annual surface runoff results from melting of the seasonal snow pack in nearby mountains.

Much of the soil that underlies the high latitude and high elevation regions of the Earth is permanently or seasonally frozen. Although the amount of water stored as ice in seasonally or permanently frozen ground (or permafrost) varies from nearly zero to more than 30%, the total amount of freshwater stored in permafrost is estimated to be 0.3 million km^3 or about 1% of that stored in the Antarctic Ice Sheet (Intergovernmental Panel on Climate Change, 2001). 22.8 million km^2 or nearly 24% of the exposed land of the Northern Hemisphere contains permafrost including over half of Canada and much of the Russian Arctic. Permafrost thickness varies from tens of meters to over a thousand meters and temperatures range from just below freezing to less than $-10\,^{\circ}C$. Permanently frozen ground is usually overlain by an "active layer" that freezes and thaws annually. Freezing and thawing processes impact ecosystem diversity and productivity, and understanding such processes is important for civil engineering and architectural projects in cold regions.

Floating ice includes river, lake, and sea ice. During their respective winter seasons, sea ice covers 14 to 16 million km^2 of the Arctic Ocean and 17 to 20 million km^2 of the Southern Ocean around Antarctica. Conversely, the Arctic minimum in summer sea ice cover is approximately 7 to 9 million km^2, far greater than the 3 to 4 million km^2 minimum found in the Antarctic. Traditional human cultures and biological systems are adapted to the seasonal cycle of sea ice extent. Penguins, polar bears, seals, and other mammals depend entirely on sea ice for habitat.

The large ice sheets in Antarctica and Greenland contain the bulk of the water stored in the cryosphere. With an area of 12.37 million km^2 and mean thickness of 2.5 km, the Antarctic Ice Sheet covers an area half the size of North America and contains 25.71 million km^3 of ice or about 23.3 million km^3 of freshwater. Covering 1.71 km^2, the Greenland ice sheet is about 14% the size of the Antarctic Ice Sheet and has a mean thickness of 1.6 km. The Greenland ice sheet contains some 2.85 million km^3 of ice or about 2.6 million km^3 of freshwater. Glaciers and small ice caps outside the Antarctica have a combined area of about 0.54 million km^2 and a mean thickness of approximately 0.2 km. The total volume of these glaciers and small ice caps is 0.18 million km^3 of ice or 0.16 million km^3 of freshwater (Intergovernmental Panel on Climate Change, 2001).

OBSERVING THE CRYOSPHERE

Seasonal snow packs accumulate winter precipitation that is subsequently released through melt, thereby generating surface runoff during the spring and summer when it is needed for agriculture. In many semiarid and arid regions, as much as 90% of the annual surface runoff results from melting of the seasonal snow pack in nearby mountains. Measurements of the water stored in seasonal snow packs are used to predict spring runoff and also to manage reservoir and water conveyance systems. Systematic measurements of snow depth and snow density on specific dates throughout the snow accumulation season called *snow surveys*, together with empirical relationships between such snow pack measurements and observed stream flows in previous years, have been used for much of the twentieth century to predict spring runoff. More recently, networks of autonomous systems to measure snow depth and snow water equivalence (SWE) have been developed. The SNOWTEL network is operated by the National Resource Conservation Service and provides 15 min measurements of snow depth, SWE, temperature, wind, and soil moisture at more than 750 locations throughout the western United States. SNOTEL data are available in real-time over the World Wide Web (**`http://www.wcc.nrcs.usda.gov/snow/`**). Systematic manual measurements of seasonal snow depth and water content are made routinely at many locations throughout the world (e.g. Central Asia, Andes). Though less extensive than the SNOTEL network in the United States, autonomous snow measurement systems have been deployed in other parts of the world.

While empirically based runoff prediction is often adequate for spring runoff prediction, understanding of the physical processes underlying snow melt, infiltration, runoff generation, and plant physiology in snow-covered regions is required to assess the impact of climate change on hydrologic systems where snowfall occurs. Studies intended to develop such understanding require detailed information on the distribution of snow-covered area (SCA) and SWE. Because of the strong differences in albedo between snow-covered and exposed soils and vegetation, satellite and airborne remote sensing provide an efficient means to measure SCA. Measurements from satellites have been used routinely since the late 1960s to monitor seasonal snow cover. However, while progress has been made in measuring SWE via remote sensing, campaign style measurements

of snow depth and snow density are still required (e.g. NASA's Cold-land Processes Field Experiment).

Because permafrost is defined solely by soil temperature and not by soil moisture content, overlying snow cover, or air temperature, determining the location and extent of permafrost is difficult particularly at regional to hemispheric scales. Systemic monitoring through borehole and probe temperature measurements of the seasonally melting and freezing or active layer that overlies most permafrost is particularly important to assess changes in frozen ground. The Circumpolar Active Layer Monitoring Program (CALM) was established in 1991 under the auspices of the International Permafrost Association to acquire long-term active layer measurements. The network includes more than 80 sites distributed throughout the permafrost regions of the Northern Hemisphere and efforts are under way to establish a similar network of sites for long-term active layer measurements in the Southern Hemisphere (Brown *et al.*, 2000).

Because of the importance of sea ice extent to the shipping industry, surface and airborne sea ice surveys have been conducted routinely for decades. Prior to the satellite era, sea ice extent was derived from observations from ships, aircraft, and land installations. Since the 1970s, satellite measurements such as those from NASA's Nimbus 7 Scanning Multichannel Microwave Radiometer (SMMR), the Defense Meteorological Satellite Program Special Sensor Microwave Imagers (SSMIs), and the National Space Development Agency of Japan's recently launched Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) have been used to monitor sea ice extent. Satellite data from the Southern Ocean show general increases in sea ice extent since the late 1970s.

Measurements from radar and laser altimetry surveys such as those from NASA's recently launched Ice, Cloud and land Elevation Satellite (ICESat) will be used to measure changes in surface elevation. The ICESat mission's primary focus is the cryosphere and it will provide detailed measurements of surface elevation of sea ice. Such data will be used to measure and monitor changes in sea ice thickness. Indeed, ice thickness measurements from submarines, together with surface and other observations, show that sea ice has thinned dramatically in the Arctic over recent decades.

Because glaciers and ice sheets hold nearly 70% of the Earth's freshwater resources and because changes in these cryosphere reservoirs directly impact the sea level, monitoring changes in glaciers and ice sheets is especially important to understanding the cryosphere. Measurements fall into two primary categories: ground-based measurements through snow pits and ice cores, and remotely sensed observations from aircraft and satellites.

Glaciers and ice sheets can be divided into accumulation and ablation regions where surface mass balance is positive and negative respectively, with ice flow within the glacier or ice sheet from the accumulation to the ablation zones. Much like tree rings or lake varves, layers of snow accumulate in the higher elevation regions. Ice cores and snow pits collected in the accumulation zones provide a means to sample these snow layers. Chemical and physical parameter measurements are used to identify annual layers and to document historical changes in atmospheric chemistry (e.g. important greenhouse gas concentrations such as carbon dioxide and methane) and atmospheric fallout from volcanism (e.g. sulfur, lead), industrial pollution (e.g. lead, vanadium, sulfur), biomass burning (e.g. black carbon, ammonia), and dust (e.g. aluminum, calcium, iron) and sea salt (e.g. sodium, magnesium, strontium) transport. Such measurements provide insights not only into natural and perturbed biogeochemical cycles and atmospheric and oceanic circulation but also into historical changes in meteorology (e.g. snow accumulation rate, windiness), climate (e.g. temperature, circulation), and the state of ice sheet or glacier (frequency of summer melt, surface mass balance, ice sheet elevation). Ice cores have been collected from glaciers and ice sheets at least since the 1950s including those collected from lower latitude, high altitude glaciers throughout the world as well as from smaller glaciers and ice caps around the Arctic Basin.

Five ice cores reaching bedrock have been collected in Greenland, including two nearly adjacent cores completed in the early 1990s at the highest point of the ice sheet (72 °N, 38 °W, 3200 m): the US Greenland Ice Sheet Project 2 (GISP2) and the European Greenland Ice Core Program (GRIP) cores. Each over 3000 m in length and analyzed for a broad range of physical and chemical parameters, these cores provided for the first time nearly replicate high-resolution environmental records spanning more than 250 000 years with annual layering identified over the past 110 000 years. In 2003, the European North Greenland Ice Core Project (NGRIP) (75 °N, 42 °W, 2900 m) reached bedrock, providing a 3010-m record spanning ~200 000 years from the northern part of the ice sheet. Hundreds of snow pits and shallow and intermediate (~150 m) cores, including more than 85 cores since 1995 as part of the ongoing US Program for Arctic Regional Climate Assessment (PARCA), have also been collected from most regions of the Greenland ice sheet (Thomas *et al.*, 2001).

A number of cores reaching bedrock have also been collected in Antarctica, including the Siple Dome core (78.5 °S, 106.8 °W, 623 m) in West Antarctica and the Vostock (78.5 °S, 106.8 °E, 3500 m) and the recently completed Dome Concordia (81.7 °S, 148.8 °W, 3200 m) cores in East Antarctica. A coordinated international research program

called the International Trans Antarctic Scientific Expedition (ITASE) is underway. ITASE includes the collection and analysis of a large number of shallow ice cores from all regions of Antarctica spanning the last 200 years. Interpretations of glaciochemical records are complicated by air-to-snow transfer processes and modifications of the chemical or physical record after snow deposition. While far more important for some parameters than others, such transfer processes may modulate the relationship between concentrations in the air and those in the snow and thence preserved in the ice core record. Recent research efforts have been directed toward better understanding and modeling of these transfer processes.

Developments in remote sensing have and will advance our ability to observe the cryosphere in the coming years. For glaciers and ice sheets, aircraft and satellite-mounted laser altimeters such as those aboard NASA's recently launched ICESat will be used to measure changes in surface elevation, with thickening indicating regions of positive mass balance. Similar measurements using radar altimetry will be made from the European Space Agency's CryoSat platform. CryoSat is scheduled for launch in 2004.

Precise gravity measurements from the Gravity Recovery and Climate Experiment (GRACE) satellites will be used to track how water is transported and stored within the Earth's environment, including measuring changes in mass of cryospheric reservoirs such as the ice sheets and permafrost. The GRACE satellites were launched in 2002. GRACE is a joint mission between NASA and German Center for Air and Space Flight.

General background information on the cryosphere, as well as regularly updated satellite and other global cryosphere measurements, are available from the US National Snow and Ice Data Center (**`http://nsidc.org/`**) and the CRYosphere SYStem in Canada (CRYSYS) program (**`http://www.msc.ec.gc.ca/crysys/`**).

GLACIERS AND ICE SHEETS AS PALEOCLIMATE ARCHIVES

Studies of chemical and physical parameters in ice cores form a cornerstone of global climate change research because glaciers and ice sheets contain arguably the most complete archives of paleoclimate information available. Unlike other long-term proxies such as tree ring records and lake sediment and peat bog cores, ice cores collected from glaciers and ice sheets contain direct, very high-resolution records of precipitation and atmospheric chemistry through soluble and insoluble impurities contained within the ice and gases within trapped bubbles. Among other issues, these data are used to study long-term climate variability and recent warming as well as changes in hemispheric-scale atmospheric and oceanic circulation, sea ice extent, anthropogenic pollution, and volcanism (Legrand and Mayewski, 1997; McConnell *et al.*, 2002).

Deep ice core measurements have documented rapid climate changes such as the 7 °C increase in North Atlantic temperature that occurred over a 50-year period at the end of the last glacial period. Measurements of important greenhouse gases like methane, carbon dioxide, and chlorofluorocarbons in bubbles within the snow and ice extend direct atmospheric chemistry measurements from recent decades and so provide historical records of changing atmospheric chemistry and related greenhouse forcing (Raynaud *et al.*, 1993).

Ocean productivity is in a large part modulated by fertilization from atmospheric dust containing iron and other nutrients and is a major component of the global carbon cycle. Recent advances in ice core analytical techniques allow for a sub-annually resolved determination of past atmospheric dust and iron fluxes to the Arctic and Antarctic. Such proxies are important for understanding how both naturally occurring and human-caused changes in dust fluxes in remote regions have altered carbon sequestration in the oceans over recent decades to centuries to millennia. Similar measurements from ice cores may one day provide better understanding of large-scale rapid climate changes such as those that occurred during the last glacial to interglacial transition (National Research Council, 2002).

Annually resolved records of net water accumulation from the accumulation zones of glaciers provide a direct measurement of the past changes in the hydrologic cycle and are important for detecting recent changes in the global hydrologic cycle and for determining the current mass balance of glaciers and ice sheets. Because much of the Earth's water is stored in glaciers and ice sheets, understanding ice sheet mass balance is critical to predicting future sea levels. While it is generally believed that climate warming will accelerate the global hydrologic cycle and result in increased precipitation over many of the colder regions of the Earth, extensive recent measurements of precipitation from ice cores in Greenland do not show any significant increases in recent precipitation. However, determination of significant long-term precipitation trends is difficult because precipitation over much of Greenland is cyclical and highly variable from one location to another.

VULNERABILITY OF CRYOSPHERE TO CLIMATE CHANGE

Interaction between the cryosphere and other components of the hydrologic cycle depends on the change of water to and from the solid phase (ice and snow) and much of the cryosphere exists near the melting point so the size and behavior of the cryosphere is highly sensitive to climate change. There is a strong positive feedback

between the cryosphere and climate through changes to the albedo or surface reflectivity. The albedo of freshly fallen snow, aged snow, and sea ice are approximately 0.9, 0.5, and 0.8 respectively, although sea ice albedo is highly variable since it depends on the presence or absence of snow cover. Conversely, the albedo of bare and vegetated land and open seawater is approximately 0.2, 0.1, and 0.1 respectively. Enhanced snow cover and sea ice increase albedo, decreasing absorption of solar radiation and leading to enhanced cooling. In a positive feedback loop, the enhanced cooling results in reduced melting and so persistence of the snow cover and sea ice. Conversely, lower snow cover and sea ice result in lower albedo and enhanced heating, leading to increased melting and less persistence of snow and sea ice. Impurities, in particular black carbon from combustion, significantly reduce the albedo of snow and ice. Recent evidence suggests that even small increases in black carbon emissions to the atmosphere from human industrial activities deposited with snow may be significantly altering the global climate (Hansen and Nazarenko, 2004).

Seasonal snow packs accumulate winter precipitation that is subsequently released through melt, thereby generating surface runoff during the spring and summer when it is needed for agriculture. In many semiarid and arid regions, as much as 90% of the annual surface runoff results from melting of the seasonal snow pack in nearby mountains. The likely results of climate warming would be increased rainfall in winter and earlier onset of the melt season, both of which result in an earlier release of water from the snow pack reservoir, resulting in higher river flows in winter and early spring before the growing season and lower late spring and summer flows.

Reconstructions using historical surface observations coupled with more recent satellite measurements indicate that the overall snow cover increased over North America from the early 1900s to the mid-1970s, in particular fall snow cover, followed by a pronounced decrease in SCA during the 1980s and 1990s due to more rapid melting in early spring. While the historical data for Eurasia are more sporadic, they indicate little long-term change in fall snow cover, but a rapid reduction in spring snow cover. Note that 98% of the Earth's seasonally SCA is in the Northern Hemisphere. It is likely that the increasing early winter snowfall observed over North America dominated through the mid-1970s, resulting in longer duration snow cover. Since then, however, early springtime warming has dominated (Brown, 2001).

As with other parts of the cryosphere, permafrost is sensitive to the changing climate. Changes in the depth of the active layer and permafrost temperatures are closely related changes in overlying air temperatures, although variations in the depth and persistence of snow cover also play a role since the snow cover insulates the ground from air temperature extremes. Differences in rock and soil type between monitoring sites also modify the responsiveness of the subsurface to changes in air temperature. While consistent, long-term records are relatively few, indications are that active layer depths and borehole temperatures have increased in some parts of the Arctic and at lower latitudes in parallel with increasing air temperatures and decreasing seasonal snow cover, especially in the late 1990s.

By the middle of the twenty-first century, modeling studies indicate that near-surface permafrost area will decline by 12% to 15% and an active layer thickness will increase by 15% to 30% in response to climate warming. Release of methane and other gases by melting permafrost further increases greenhouse gas concentration, resulting in an indirect positive feedback that will accelerate decreases in permanently frozen ground (Intergovernmental Panel on Climate Change, 2001).

By insulating the relatively warm ocean water from the cold polar atmosphere, sea ice regulates the exchange of heat, moisture, and salinity in the polar oceans. The large differences in albedo between ice covered and open ocean lead to a strong positive feedback between warming and decreases in sea ice extent. While seasonal changes in sea ice extent are large in both the Northern and Southern Hemispheres, they are much larger in the Antarctic. Long-term trends in sea ice extent from satellite measurements over the past 25 years show that in most of the Arctic, sea ice season has decreased ($\sim$2.7% per decade) while in much of the Antarctic, sea ice season has increased ($\sim$1.1% per decade) (Cavalieri *et al.*, 1997; Parkinson, 2002). In the latter, trends in sea ice season duration were found to correlate to observed spatial patterns in air temperature trends, with decreases found around the Antarctic Peninsula where strong warming has been observed and increases in sea ice season duration in the Ross Sea region.

Another consequence of decreased sea ice cover, particularly in the Arctic, may be a change in deep water formation. Rejection of brine as sea ice forms has important influences on sea surface salinity, deep water formation rates, and large-scale ocean circulation. Sea ice has also been found to play an important role in the redistribution of freshwater in the North Atlantic (Intergovernmental Panel on Climate Change, 2001).

Sea level and the size of the cryosphere are directly linked because $\sim$1.8% of Earth's water is stored in glaciers and ice sheets. Neglecting thermal expansion of the oceans, isostatic rebound, and other changes that would result from large increases in global temperature but allowing for calculated isostatic rebound and sea water replacing grounded ice, sea level would rise by 7.2 m if the entire Greenland ice sheet melted, 61.1 m if the Antarctic ice sheet melted, and 0.51 m if all glaciers and smaller ice caps melted, respectively (Intergovernmental Panel on Climate Change, 2001).

Also important to global climate is the timing and rate of melt water delivery to the oceans, particularly in the North Atlantic. Numerous studies suggest that melt water production from the Greenland ice sheet and deepwater formation in the North Atlantic played a major role in the well-documented Younger Dryas rapid climate shift to glacial conditions about 10 500 years ago (Intergovernmental Panel on Climate Change, 2001; National Research Council, 2002).

While an understanding of the current mass balance of ice sheets is not complete, models suggest a decrease in the size of the Greenland ice sheet under likely warming scenarios because much of Greenland is relatively close to the melting point, contributing to an increase in sea level (Intergovernmental Panel on Climate Change, 2001). Indeed, repeat altimetry surveys over recent decades suggest that while the upper elevations of the Greenland ice sheet are in overall mass balance (i.e. there is no change in the thickness of the ice sheet in these regions), the margins of the ice sheet are shrinking rapidly (Krabill *et al.*, 2000). Much of this behavior is a result of natural cycles in atmospheric and oceanic circulation that determines the meteorology over Greenland is uncertain and so the current mass balance of the ice sheet is unclear, however. Warmer atmospheric temperatures will likely accelerate the hydrologic cycle and lead to increased precipitation. Because temperatures over most of Antarctica are well below the melting point, it is likely that the Antarctic ice sheet will increase in size under possible global warming scenarios, contributing to a decrease in sea level (Intergovernmental Panel on Climate Change, 2001).

Mountain glaciers are particularly vulnerable to climate change with twentieth century consequences observed throughout the world. As much as 20% of the sea level rise observed during the twentieth century has been attributed to the retreat of mountain glaciers (Intergovernmental Panel on Climate Change, 2001). Increased melting of mountain glaciers has resulted in increased flooding along major river systems such as the Rhine, while melting of glaciers and retreat of permafrost has resulted in increased erosion, debris flows, and avalanches. The complete loss of mountain glaciers, projected to occur as early as the middle of the twenty-first century in some high mountain regions such as the Andes and Himalayas, will have disastrous impacts on subsistence agriculture and hydropower generation that depend on flows from mountain glaciers.

Acknowledgments

Support was provided by SAHRA ("Sustainability of Semi-Arid Hydrology and Riparian Areas"), an NSF Science and Technology Center at the University of Arizona. Additional support came from research grants from the National Science Foundation and the National Aeronautics and Space Administration.

FURTHER READING

Hardy J.P., Albert M.R. and Marsh P. (Eds.) (1999) *Snow Hydrology*, John Wiley & Sons.

Tranter M., Armstrong R., Brun E., Jones G., Sharp M. and Williams M. (Eds.) (1999) *Interactions between the Cryosphere, Climate, and Greenhouse Gases*, IAHS Series of Proceedings and Reports, IAHS.

REFERENCES

Barry R.G. and Chorley R.J. (1992) *Atmosphere, Weather and Climate*, *Sixth Edition*, Routledge: London.

Brown R.D. (2001) Northern hemisphere snow cover variability and change, 1915–97. *Journal of Climate*, **13**, 2339–2355.

Brown J., Hinkel K.M. and Nelson F.E. (2000) The circumpolar active layer monitoring (CALM) program: research designs and initial results. *Polar Geography*, **24**, 165–258.

Cavalieri D.J., Gloersen P., Parkinson C.L., Comiso J.C. and Zwally H.J. (1997) Observed hemispheric asymmetry in global sea ice changes. *Science*, **272**, 1104–1106.

Graedel T.E. and Crutzen P.J. (1995) *Atmosphere, Climate, and Change*, Scientific American Library: New York.

Hansen J. and Nazarenko L. (2004) Soot climate forcing via snow and ice albedos. *Proceedings of the National Academy of Sciences*, **101**, 423–428.

Intergovernmental Panel on Climate Change (2001) *Climate Change 2001: The Scientific Basis*, Cambridge University Press: Cambridge.

Krabill W., Abdalati W., Frederick E., Manizade S., Martin C., Sonntag J., Swift R., Thomas R., Wright W. and Yungel J. (2000) Greenland ice sheet: high-elevation balance and peripheral thinning. *Science*, **289**, 428–430.

Legrand M. and Mayewski P.A. (1997) Glaciochemistry of polar ice cores: a review. *Reviews of Geophysics*, **35**, 219–243.

McConnell J.R., Lamorey G.W. and Hutterli M.A. (2002) A 250-year high-resolution record of Pb flux and crustal enrichment in central Greenland. *Geophysical Research Letters*, **29**, 2130.

National Research Council Committee on Abrupt Climate Change (2002) *Abrupt Climate Change: Inevitable Surprises*, National Academy Press: Washington.

Parkinson C.L. (2002) Trends in the length of the Southern Ocean sea-ice season, 1979–99. *Annals of Glaciology*, **34**, 435–440.

Raynaud D., Jouzel J., Barnola J.M., Chappellaz J., Delmas R.J. and Lorius C. (1993) The ice record of greenhouse gases. *Science*, **259**, 926–934.

Thomas R.H., PARCA Investigators (2001) Program for Arctic Regional Climate Assessment (PARCA): goals, key findings, and future directions. *Journal of Geophysical Research*, **106**, 33691–33705.

199: Role and Importance of Paleohydrology in the Study of Climate Change and Variability

BRENDA EKWURZEL

Global Environment Program, Union of Concerned Scientists, Washington, DC, US

Paleohydrology has advanced our understanding of climate by uncovering the variability of the prehistoric water cycle. Various proxy records provide insight into aspects of the hydrologic cycle that occurred before the instrumental record and allow for paleoclimate reconstructions locally or over large regions. For example, geomorphologic methods can quantify megafloods; coral or sediment core methods create paleodischarge records; rodent middens archive changes in precipitation volume and temperature; closed-basin lake level fluctuations offer quantitative evidence of changes in water balance, and tree-rings provide a high-resolution archive of precipitation or drought. This article reviews representative examples from the field of paleohydrology and demonstrates the role each has played in enhancing our understanding of climate variability.

INTRODUCTION

Time-series measurements span a range of timescales from rapidly repeating satellite measurements to time series that begin with the earliest historical records that continue to the present day. Long historical records, such as the typhoon records in local gazettes kept since 975 A.D. in the Guangdong Province of southern China (Liu *et al.*, 2001; Perkins, 2002b), are rare. The longest surface air temperature record (daily back to 1772 and monthly since 1659) is located in the English Midlands, commonly referred to as the *Central England Temperature Record* (Brabson and Palutikof, 2002). Historical freeze and thaw dates for rivers and lakes exist for the past 150 years; a rare 550-year record persists for Lake Suwa in Japan (Magnuson *et al.*, 2000). Historical records of grape-harvest dates extend back to 1370 in Burgundy France (Chuine *et al.*, 2004). Unfortunately, the time frame of human observation may not be long enough or the records accurate enough for many climate-variability studies. Paleohydrology can be a vital approach for extending our reference time over longer periods where climate cycles, trends, or dramatic shifts may give full expression. Furthermore, paleohydrology can reconstruct extreme hydrologic events that humans have never observed such as the Missoula Flood and the Altay Flood, both with peak discharges of around 20 million $m^3\ s^{-1}$ (O'Connor and Baker, 1992; Baker, *et al.*, 1993). This is roughly equivalent to over 7000 times the average flow rate over the Niagara Falls (i.e. the combined flow of American Falls, Bridal Veil Falls, and Canadian Horseshoe Falls).

Unlike the modern hydrologist who can probe the current hydrologic cycle to record a specific property of interest, the paleohydrologist is often at the mercy of preserved physical, chemical, and biological data that serve as a proxy for hydrologic processes. This article is organized into two broad sections. The first section addresses examples of climate archives that exist in preserved elements of the ancient hydrologic cycle that can be directly sampled today. Reservoirs of old H_2O molecules in the form of glacial ice and old groundwater serve as important archives of climate information. Calcite deposits in caves provide another groundwater archive because they record groundwater geochemistry and surface recharge conditions at the time of formation. The second broad section reviews elements of the hydrologic cycle that are too transient to be preserved as water that can be directly sampled today. Instead, other records serve as proxy evidence for past precipitation, surface water discharge, or the lack of water such as during droughts. This article presents some of the more successful paleohydrology methods that have added new insights to climate

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

variability or that may play a more prominent role in future studies.

LONG RESIDENCE TIME ELEMENTS OF HYDROLOGIC CYCLE: CLIMATE ARCHIVES

Frozen Precipitation As a Climate Archive: Continental Ice Sheets

Precipitation that falls at high altitude or latitude may encounter cold enough temperatures to fall as snow. In regions where a winter's snow survives a summer season, glaciers can form as additional snow accumulates each year and eventually turns into ice. Mountain glaciers and large continental ice sheets archive local precipitation, atmospheric gases, dust, and other constituents through their annual snow accumulation. A proper understanding of the hydrologic conditions that impact the precipitation (e.g. humidity, moisture sources or prevailing storm track patterns, temperature, elevation, global sea level, etc.) is critical for proper paleoclimatic interpretation of the ice-core records. The most famous locations for the longest ice records are in Antarctica and Greenland (Figure 1).

These ice-core records with exquisite resolution have had a profound impact on our interpretation of climate variability during the late Quaternary period. The Greenland Ice core data reach as far back as 150 000 years while the third core drilled at Vostok, Antarctica, extends over the past 420 000 years and the ice core from Dome C Antarctica offers a view of climate variability over the past 740 000 years (Figure 1; Stauffer, 1999; Petit *et al.*, 1999; EPICA, 2004). Extensive research on ice cores has produced significant results such as correlations and lead/lag relationships between the Greenland and Antarctic ice cores (e.g. Blunier *et al.*, 2001). In this article, we focus on the analysis of the Vostok ice-core record as representative of the techniques used and major findings of the ice-core climate archive in general.

The major techniques used to extract climate information from the ice are to measure the stable isotopes of the precipitation, the dissolved constituents in the precipitation, dry deposition on the snow surface (e.g. dust), and to measure the atmospheric trace gases eventually trapped within the ice. The stable isotopes of the water molecule hydrogen (δD (in ‰) $= [(({}^{2}H/{}^{1}H)_{sample}/({}^{2}H/{}^{1}H)_{standard}) - 1] \times 1000$) and oxygen ($\delta^{18}$O (in ‰) $= [(({}^{18}O/{}^{16}O)_{sample}/({}^{18}O/{}^{16}O)_{standard}) - 1] \times 1000$) are measured relative to the Vienna Standard Mean Ocean Water and in general are most depleted in Antarctic precipitation. Details regarding how ice-core researchers reconstruct temperature of the atmosphere from the stable isotopes measured in the ice can be found in Petit *et al.* (1999). The trace gases trapped in air bubbles about 100 m below the snow surface preserve the variability of atmospheric trends (Stauffer, 1999; Petit *et al.*, 1999) which causes an age offset between the ice and the trapped gas in a given layer. This age offset creates some difficulty determining the exact chronology

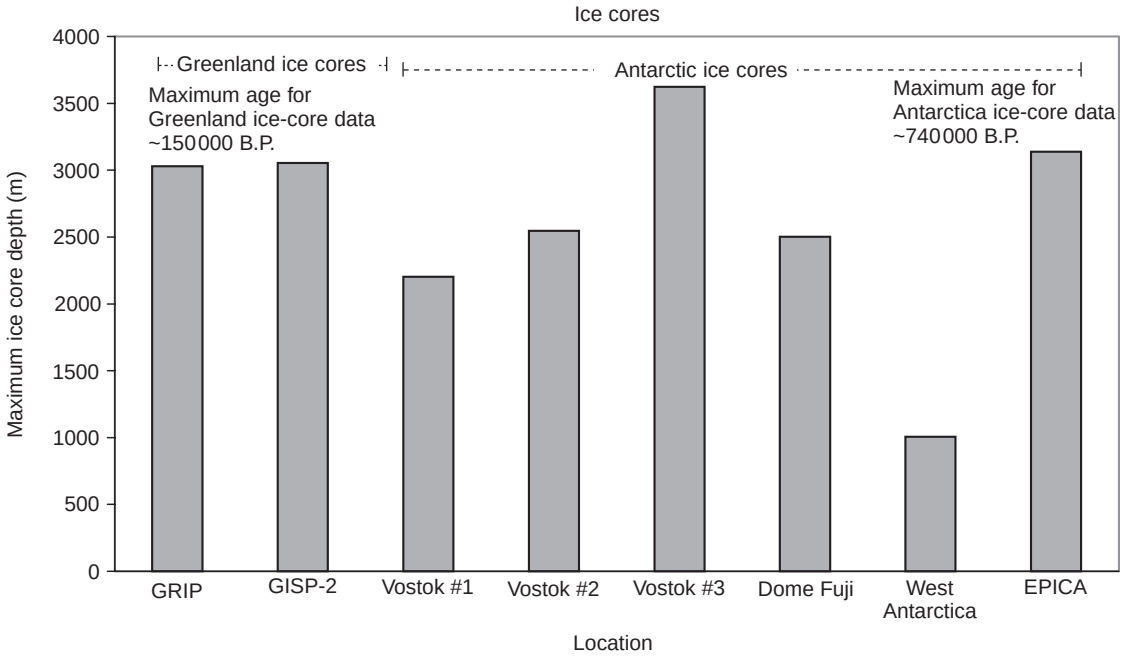


Figure 1 Long ice cores with their associated project names. Their general location (i.e. Greenland or Antarctica), their maximum drilling depth, and regional maximum age limit are depicted. (Adapted from Stauffer (1999), Petit *et al.* (1999) and EPICA (2004)) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

for forcing mechanisms and climate response. However, Petit *et al.* (1999) argue that the new glaciological timescale known as *GT4* for the Vostok core yield age accuracy of ±15 kyr for most of the record and better than ±5 kyr for the last 110 kyr. This age resolution will be accurate enough to see general climate response and most lead-or-lag trends between trace gas measurements and ice measurements.

The ice-core results confirm what has been found in other climate records such as ocean sediment cores and lake sediment cores regarding glacial cycles over the past million years (Figure 2). Ice core glacial-to-interglacial cycles exhibit periodicity similar to Milankovitch orbital forcing periods of 100 000 and 41 000 years (Imbrie *et al.*, 1992; Petit *et al.*, 1999). Dust decreased while temperature, methane, and CO_2 all increased during the glacial-to-interglacial transition (Petit *et al.*, 1999). Methane, in particular, increased significantly during the last half of deglaciation. This sequence, Petit *et al.* (1999) argue, implies that the Milankovitch orbital cycles trigger the climate response which is amplified by the trace gases and then the strong ice-albedo feedback leading to rapid deglaciation.

Groundwater As a Climate Archive

In contrast to the short residence times of water in the atmosphere and in rivers, water that travels through porous media in the subsurface can have much longer residence times. Indeed aquifers such as the Milk River aquifer in Alberta, Canada, and the Great Artesian Basin in Australia are amongst the oldest groundwater (ca. one to four million years old) dated thus far (Phillips *et al.*, 1986; Torgersen *et al.*, 1991; Collon *et al.*, 1998). In the geologic record, periods of erosion erase prior deposition history, leaving only an erosional surface to document the period as compared to the preserved periods of sedimentation. Likewise, confined aquifers can archive periods of active recharge yet leave gaps during periods when recharge does not occur (e.g. during a drought or when covered by ice during a glacial period). Ideal settings for groundwater archives of climate are confined aquifers with a single recharge region where minimal mixing occurs with other formation water during the subsurface transit. Perhaps the most successful groundwater paleohydrology contribution to the evolution of the global climate-variability theory is the mean annual temperature reconstructions of the

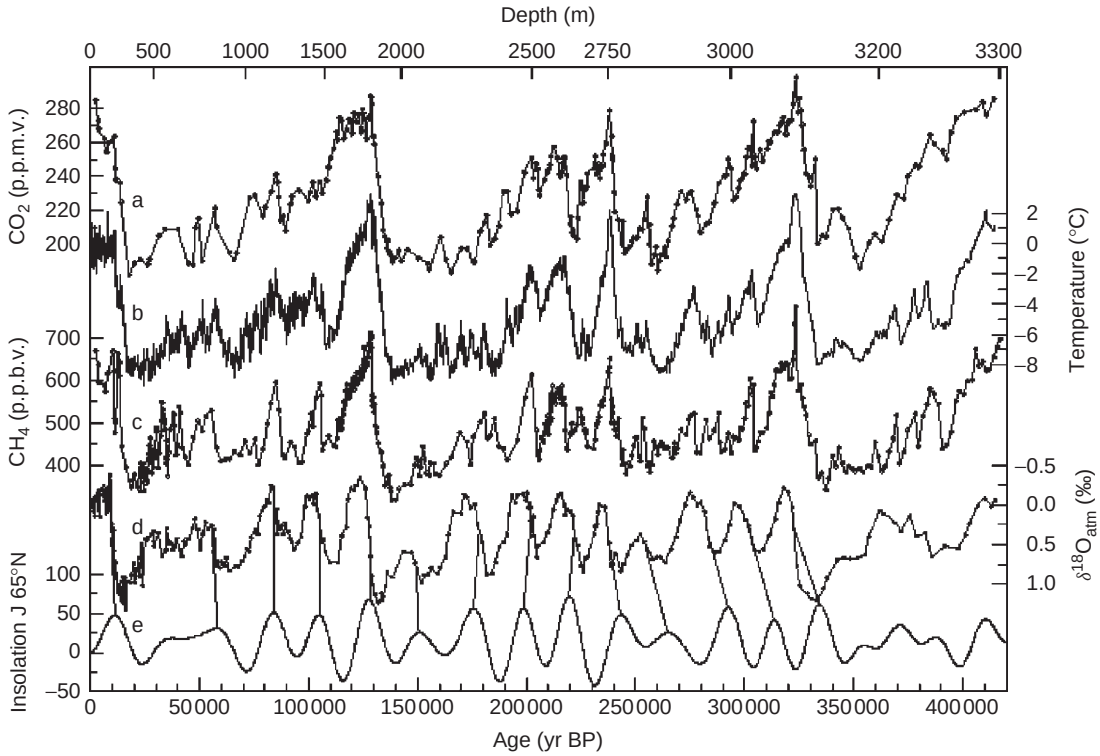


Figure 2 Vostok time series and insolation. Series with respect to time (gt4 timescale for ice on the lower axis, with indications of corresponding depths on the top axis) of (a) CO_2; (b) isotopic temperature of the atmosphere; (c) CH_4; (d) $\delta^{18}O_{atm}$; and (e) mid-June insolation at 65 °N (in Wm^{-2}). The mean resolution of the CO_2 (CH_4) profile is about 1500 (950) years. It goes up to about 6000 years for CO_2 in the fractured zones and in the bottom part of the record, whereas the CH_4 time resolution ranges between a few tens of years to 4500 years. The overall accuracy for CH_4 and CO_2 measurements are ±20 p.p.b.v. and 2–3 p.p.m.v. respectively. $\delta^{18}O_{atm}$ standard is modern air composition. No gravitational correction has been applied. (Reproduced from Petit *et al.* (1999) by permission of Nature)

Last Glacial Maximum (LGM) period on continents. In particular, theories regarding high latitude versus tropical latitude climate response during the LGM came under intense scrutiny and revisions and refinements were made, as we shall see after a brief discussion of the noble gas approach.

Groundwater: Reconstructing Paleotemperature from Dissolved Gases

Noble gases have properties that make them appropriate for archives of the temperature and elevation of the recharge conditions for an aquifer (Stute and Schlosser, 2000). The solubility of noble gases is a direct function of temperature, atmospheric pressure, and salinity of the water. Noble gas solubility increases with decreasing temperature and the sensitivity to this trend is different for each noble gas (i.e. higher mass noble gases are more sensitive.) If the water table is between 5 and 30 m from the surface and the recharge rate is active but not exceeding several hundred mm year^{-1}, the noble gas recharge temperature is about 1 °C different from the ground surface temperature. For further discussion regarding methods to account for other sources of various noble gases, see Stute and Schlosser (2000).

The noble gas hydrologist interested in reconstructing LGM paleotemperatures chooses a confined aquifer in which the recharge zone elevation is well known. Thus, if one knows the confined aquifer recharge elevation, and one assumes the precipitation is not saline, one can unravel the ground temperature at the water table by measuring the noble gas concentrations from wells pulling water up from the appropriate age water for the LGM. The typical approach has been to measure noble gases, stable isotopes, and radiocarbon throughout the confined aquifer (Andrews and Lee, 1979; Heaton *et al.*, 1986; Stute and Deák, 1989; Stute *et al.*, 1992, 1995a, 1995b; Dennis *et al.*, 1997; Ballantine and Hall, 1999; Beyerle *et al.*, 1998; Clark *et al.*, 1998; Stute and Talma, 1998; Edmunds *et al.*, 1998). The radiocarbon data combined with stable carbon isotope measurements and a refined geochemical evolution model is critical for placing the noble gas paleotemperatures into a chronologic reference frame. The water stable isotopes serve as a check on the radiocarbon ages because the glacial ocean $\delta^{18}O$ and $\delta^{2}H$ shifted to more enriched values worldwide (1.3%) as much of the depleted water was locked up in the ice during the LGM (Fairbanks and Matthews, 1978). Note that the water stable-isotope signal may be complicated by other factors such as temperature, storm track patterns, and evaporation, which must be accounted for.

A key distinction between groundwater as a climate archive and annual tree-rings, annual lake varves, ice core, or other high-resolution records is that the climate-variability signal will be smoothed in a confined aquifer archive. The water and dissolved constituents will be dispersed over thousands of years as the water flows through a heterogeneous aquifer (see, for example, Stute and Schlosser, 2000; Figure 6). Groundwater dispersion is the reason researchers usually are only able to distinguish the relative shift in paleotemperature between the LGM and today. Statistically, this is achieved by averaging many measurements for aquifer waters in the most recent few thousand radiocarbon years and compare those with the average of many measurements of water collected from several thousands radiocarbon years around the LGM time period. This is also necessary when one considers the average age error for older radiocarbon dated waters is around 2000 years (Phillips *et al.*, 1989).

Groundwater: LGM Paleotemperature Provokes Further Scrutiny of Tropical SST

One of the central issues that evolved in the paleoclimate debate was whether the tropical sea surface temperatures (SST) were similar to modern values while the polar ocean regions experienced more extreme swings in temperature through time as indicated by foraminiferal tropical sea surface temperature reconstructions. Central to this debate was the difference between tropical SST and the tropical-mountain glacier snowline elevation also known as *equilibrium line altitude.* Geomorphic reconstructions of equilibrium line altitude during the LGM coupled with foraminifera reconstructions of tropical SST implied a LGM lapse rate that would not match known physical principles of the modern lapse rate (Greene *et al.*, 2002). During this debate, several Global Circulation Models had a difficult time holding tropical SST relatively constant while the higher latitude SST cooled during their LGM simulations (Hansen *et al.*, 1984; Manabe and Broccoli, 1985).

As with most paleoclimate discussions, debate evolved as the proxy evidence accumulated. Early paleoclimate work indicated relatively stable tropical SST. After extensive work in the oceans, the Climate: Long-Range Investigation, Mapping, and Prediction (CLIMAP) Project Members (1981) endeavored to reconstruct SST history based on the planktonic foraminifera assemblages (CLIMAP, 1981). CLIMAP predicted little change in tropical SST for the LGM with a global statistical cooling of 1.4 °C and 1.7 °C for February and August respectively (CLIMAP, 1981). Matthews and Poore (1980) argued that tropical SST remained constant throughout the Cenozoic, despite swings in high-latitude SST.

Despite the more sporadic nature of the terrestrial paleoclimate records compared to the synoptic scale oceanographic records, continental reconstructions indicated that the tropics were probably cooler during the LGM. The most notable records were the tropical snowline elevation reconstructions that indicated a significant lowering of the snowline in tropical mountains during the LGM representing a 4 °C to 6 °C cooler temperature (e.g. Broecker and

Denton, 1989). Tropical vegetation records also indicated a greater LGM cooling compared with CLIMAP and the Global Circulation Model research questioned the inconsistency (Rind and Peteet, 1985). This controversy drove researchers to look for other continental proxy records for temperature in the tropics and the search began in earnest for aquifers with archived LGM waters in order to apply the noble gas paleotemperature method.

The paleohydrology work in groundwater archives is still ongoing, but the results thus far confirm that the recharge locations of tropical aquifers record cooler LGM temperatures. Table 1 summarizes the results: (i) tropical aquifers ($<33°$ latitude) were on average $5.8 \pm 1.5\,°C$ cooler during the LGM, and (ii) midlatitude aquifers (33° to 51° latitude) were on average $7.0 \pm 1.8\,°C$ cooler during the LGM (some locations were near the ice sheet). Higher latitudes were usually ice-covered regions during the LGM, which eliminated or restricted recharge to high-latitude aquifers; thus, most of this research has been conducted in tropical to midlatitude aquifers.

The straightforward physics of the noble gas paleotemperature results compelled paleoclimate researchers to reevaluate the CLIMAP results and to search for other proxy records for tropical ocean SST. Guilderson *et al.* (1994) developed an important independent ocean proxy for tropical sea surface temperature based on $\delta^{18}O$ and Sr/Ca data from Barbados corals and found a tropical cooling of around 5 °C at the LGM. More recent faunal reconstructions yield a wider range of tropical SST during the LGM, which reflect the spatial variability of the tropical ocean. The faunal reconstructions suggest as much as 5.5 °C SST depression in the eastern equatorial Pacific and equatorial Atlantic ocean while only a 2 °C cooling was found near Hawaii during LGM (Lee and Slowey, 1999; Mix *et al.*, 1999). Entirely different proxy records for tropical ocean SST yield minimal depression at LGM of about 2 °C based on alkenones (Bard, 1999). In summary, the more recent ocean proxy records for tropical SST suggest a larger cooling (i.e. 2–5 °C) than the original CLIMAP reconstruction (i.e. 1–1.7 °C) of tropical SST during the LGM, based on a wider range of ocean proxy SST interpretations.

Greene *et al.* (2002) worked with a comprehensive set of tropical snowline (i.e. glacier equilibrium line altitude) data to reconstruct the LGM freezing level which was a combination of the lowered sea level, reduction in snowline elevation, and a host of other factors tested with their analysis. Using a single-cell tropical climate model, Greene *et al.* (2002) found that LGM precipitation did not have to be changed significantly but rather that the snowline depression combined with lapse-rate physics implies that the tropical SST had to be around 3 °C cooler during the LGM. This is less than the most extreme ocean proxy SST records, but is consistent with the average of these ocean proxy records and their associated errors, with vegetation or tropical pollen reconstructions for the tropics, with the recharge temperature locations from noble gas groundwater records, and with many Global Circulation Model results (Farrera *et al.*, 1999). The groundwater archive provides a mean annual paleotemperature at a terrestrial elevation that can be directly compared with lapse-rate gradients. The paleohydrologic evidence from the groundwater archives filled an important gap between the sea surface elevation and mountain snowline elevation to confirm that LGM lapse rates adhered to physical principles that create modern lapse rates. Now researchers have a more refined understanding of how precipitation along with temperature influence local equilibrium line altitude (Hostetler and Clark, 1997) while recognizing significant regional variability in tropical ocean SST during the glacial period (Farrera *et al.*, 1999).

Water Supersaturated with Respect to Calcite: Caves As Climate Archives

Settings where water supersaturated with respect to calcite causes precipitation of mineral layers that reflect the water geochemistry create the potential for long paleoclimate records. These records are preserved in calcite veins, stalactites, stalagmites, dripstones, or speleothems (a general term for a mineral deposit formed in a cave by the action of water). Cave-record time spans are on par with continental ice sheets and ocean sediment cores. For example, a core from a calcite vein in a Nevada extensional fault, Devils Hole, has uranium series dates that span from 566 to 60 ka (Ludwig *et al.*, 1992; Winograd *et al.*, 1992). As we shall see in the following examples, the speleothem records found in caves will play a more critical role in climate-variability studies if the paleohydrology of each study location is properly reconstructed. The following paleohydrology aspects must be understood (i) recharge locations; (ii) recharge water geochemistry; (iii) recharge vegetation; (iv) transit time in the groundwater between recharge location and mineral deposition location; (v) groundwater geochemical evolution between recharge and mineral deposition location; and (vi) conditions that impact the isotopic fractionation during the phase change between dissolved phase and solid mineral precipitation (e.g. temperature). The net result is that the ages given by the methods used to reconstruct the dates for mineral precipitation need to be corrected for the transit time between recharge and mineral deposition. Such offsets are critical if we are to understand potential lead-or-lag relationships between individual site responses and global forcing mechanisms.

The Devils Hole calcite-vein data provoked a spirited debate in the paleoclimate community (e.g. Broecker, 1992; Shackleton, 1993; Imbrie *et al.*, 1993; Herbert *et al.*, 2001; Herbert *et al.*, 2002), because if the published chronology was accurate, the $\delta^{18}O$ measured in the calcite mimicked the broad patterns seen in the ice and ocean records

Table 1 Dissolved noble gas reconstructions of the difference between the Last Glacial Maximum (LGM) and Holocene mean annual recharge temperature. Recharge elevations and recharge location are indicated for each study region. Table is arranged in order of increasing latitude

Study location	Holocene noble gas mean temperature	LGM noble gas mean annual temperature (parentheses indicate when dates are not from LGM)	Holocene/LGM noble gas temperature difference (parentheses indicate when dates are not from LGM)	Recharge elevation	Recharge latitude	Recharge longitude	Reference
Maranhão Basin, Brazil	29.6 ± 0.3 °C	24.2 ± 0.3 °C	5.4 ± 0.6 °C	450 ± 50 m	7 °S	41.5 °W	Stute *et al.* (1995a)
Chad Basin, Nigeria	30.5 ± 0.5 °C (present-day groundwater temperature)	22 ± 1.5 °C	8.5 ± 2.0 °C	330 ± 60 m	11 °N	13 °E	Edmunds *et al.* (1998)
Khwad Fan Aquifer, Oman	33.5 ± 1.7 °C	26.6 ± 0.6 °C	6.5 ± 0.6 °C	(Not specified) Entire study region range: 0–2000 m	23.5 °N	58 °E	Weyhenmeyer *et al.* (2000)
Stampriet Aquifer, Namibia	26.5 ± 0.7 °C	21.2 ± 0.7 °C	5.3 ± 0.5 °C	1230 ± 20 m	24 °S	19.5 °E	Stute and Talma (1998)
Carrizo Aquifer, Texas, United States	20.6 ± 0.6 °C	15.8 ± 0.4 °C	5.2 ± 0.7 °C	200 m	29 °N	98.5 °W	Stute *et al.* (1992)
Floridan Aquifer, Georgia, United States	17.7 ± 0.3 °C	13.7 ± 0.3 °C (Ages not well constrained)	4.0 ± 0.6 °C	125 ± 50 m	32.8 ± 0.3 °N	82.7 ± 0.6 °W	Clark *et al.* (1997)
Uitenhage, South Africa	19.5 ± 2 °C	(14 ± 1 °C) (LGM + previous interstadial)	(5.5 ± 2.2 °C)	200 ± 20 m	33.8 °S	25.5 °E	Heaton *et al.* (1986)
San Juan Basin, New Mexico, United States	9.1 ± 0.5 °C	3.6 ± 0.5 °C	5.5 ± 0.7 °C	2000 ± 100 m	36.5 °N	108 °W	Stute *et al.* (1995b)

Dakota Aquifer, Colorado and Kansas, United States	13.5 ± 0.7 °C	(7.0 ± 1 °C) (Not available for LGM but lowest temperature recorded near 12 ± 2 kyr ^{14}C age)	(6 ± 1 °C)	1255 ± 250 m	37.9 ± 0.6 °N	102.2 ± 0.8 °W	Clark *et al.* (1998)
Aquia Aquifer, Maryland, United States	13.7 ± 0.3 °C	4.7 ± 0.5 °C	9.0 ± 0.6 °C	40 ± 2 m	38.9 ± 0.1 °N	76.7 ± 0.2 °W	Aeschbach–Hertig *et al.* (2002)
Glatt Valley Aquifer, Switzerland	7.8 ± 0.6 °C	(Diminished or no recharge at LGM due to ice cover. Nearest well in age (25–30 kyr) is 2.4 ± 1.4 °C)	(At least 5 °C)	540 m	~47.2 ± 0.2 °N (Not specified)	~8.7 ± 0.5 °E (Not specified)	Beyerle *et al.* (1998)
Great Hungarian Plain, Hungary	10.6 ± 0.7 °C	1.9 ± 0.5 °C	8.6 ± 0.9 °C	150 ± 5 m	47.8 °W	21 °E	Stute and Deák (1989)
Bavaria Germany (several different aquifers)	8.1 ± 0.5 °C	(3.0 ± 1.1 °C) (LGM + previous interstadial)	(5.1 ± 1.2 °C)	Not applicable (several different aquifers)	49.5 ± 1.5 °N (several different aquifers)	9 ± 1.2 °E (several different aquifers)	Rudolph *et al.* (1984)
Chalk Aquifer, United Kingdom	9.5 °C	(Diminished or no recharge at LGM due to ice cover. Nearest well in age (>20 kyr) is 5.0 °C)	(At least 4.5 °C)	112.5 ± 12.5 m	51.3 ± 0.1 °N	1.4 ± 0.2 °E	Elliot *et al.* (1999)
Oligocene Aquifer, Mazovian Basin, Poland	5.9 ± 0.7 °C	0.7 ± 0.4 °C (Ages not well constrained)	5.0 ± 0.7 °C	75 ± 25 m	51.4 ± 0.2 °N	22 °E	Zuber *et al.* (2000)
Bunter Sandstone Aquifer, United Kingdom	9.6 ± 1.4 °C	(Diminished or no recharge at LGM due to permafrost cover. Nearest well in age (22 kyr) is 3.4 ± 1.0 °C)	(At least 6.2 ± 1.7 °C)	(Not specified)	53.4 °N	0.5 °E	Andrews and Lee (1979)

with a subtle twist. Winograd *et al.* (1992) pointed out the subtle difference between the records, remarking that the Devils Hole $\delta^{18}O$ curve suggests variability in glacial period duration. At Devils Hole, the glacial-interglacial periods became aperiodic and increased in length until the present (i.e. 80 000 and 130 000 years long) while interglacial periods lasted on the order of 20 000 years (Winograd *et al.*, 1992; Winograd, 2002). The aperiodic nature of the record along with a prominent deglaciation and glacial cycle occurring near 450 to 350 ka plus the difference in timing between this record and the Marine Isotope Stage 6 created much controversy and debate.

Reducing the uncertainties with the Devil's Hole chronology may lie in increasing our understanding of the hydrologic conditions of the past. Potential recharge locations for Devils Hole include several nearby ranges that have different flow path distances (80 to 160 km) from the calcite-core location (Winograd *et al.*, 1992). The interpretation of the groundwater geochemistry profoundly impacts the age assignment. Assumptions regarding paleorecharge conditions for temperature, vegetation, partial pressure of soil CO_2 and pH combined with measurements for the stable carbon isotope and radiocarbon evolution in the groundwater are required for accurate radiocarbon residence time calculations. Winograd *et al.* (1992) report uncertainty in radiocarbon residence time estimates for Devils Hole that range between 10 000 and 30 000 years. Winograd *et al.* (1992); Winograd (2002) point out that the U-series dating suggests the Devils Hole deglaciation cycles, as indicated by the calcite $\delta^{18}O$ curve, precede deglaciation dates for both the Vostok ice core and the orbitally tuned global ice volume record (known as SPECMAP) that is based on ocean cores. Winograd *et al.* (1992) emphasize that adding on the groundwater residence time will only increase the lead time for the onset of deglaciation at Devils Hole.

A further intriguing aspect of the Devils Hole record is that the concurrent calcite $\delta^{13}C$ measurements match the broad pattern of the $\delta^{18}O$ curve variations with the important observation that all the changes in $\delta^{13}C$ lead the $\delta^{18}O$ shifts. Coplen *et al.* (1994) were unsure whether the $\delta^{13}C$ variability reflected isotopic shifts in global atmospheric CO_2 or vegetation changes in Nevada at the recharge locations for Devils Hole. Refinement of global and regional lead/lag relationships is the focus of climate-variability research. Recent high-resolution ice-core data depict millennial-scale warming in Antarctica that precede by 1500 to 3000 years warming at Greenland over the past 90 ka B.P. (Blunier and Brook, 2001). The Devils Hole $\delta^{13}C$ variability may reflect either atmospheric CO_2 or vegetation shifts, but both would have important implications for climate variability.

Herbert *et al.* (2001) posit that the Devils Hole record represents a regional response and does not contradict the Milankovitch theory for the ice ages. They cite their own research into sea surface temperature reconstructions for the California Current, which warmed 10 to 15 ka in advance of deglaciation. Herbert *et al.* (2001) show no such sea surface temperature warming south of the modern California Current location. Their interpretation is that the Devils Hole and California Current region have responded to reorganized wind systems over the North Pacific when the continental ice sheets were large and represent a regional response (Figure 3). Imbrie *et al.* (1993) found through spectral analysis that the Devils Hole chronology has peaks in the spectrum that occur at 91 kyr and 42 kyr, with adjacent peaks at 27 kyr and 23 kyr. They suggest that coherent concentrations of variance at or near orbital periods support orbital influence.

The lead-and-lag relationships are critical for climate research, which makes the Devils Hole chronology very important. Refinement of chronology may be critical for placing the seminal record of Devils Hole into a regional and global context. Certainly, changes in recharge conditions in a semiarid setting would have a profound impact on groundwater travel times. Furthermore, the two nearby mountain ranges have varying flow path distances to the discharge location and the relative mixture of these waters could impact the Devils Hole chronology. The major filter between the climate signal at the surface that paleoclimatologists seek and the calcite-vein record is a complex groundwater regime that beckons for improved understanding. Refinement of the paleohydrology for the Devils Hole story is a rich area for future research that may be important for unraveling the regional, global, as well as orbital teleconnections of the land–ocean–atmosphere system.

To eliminate dating uncertainties, such as those associated with the 80 to 160 km distance between recharge locations and the Devils Hole calcite vein (Winograd *et al.*, 1992), the ideal study location may be one with the shortest possible residence time (e.g. less than a year) between recharge and mineral precipitation. Some locations where speleothems that have been studied have a short residence time between recharge and speleothem formation. One recent study fills the gap between where the Devils Hole record terminates and modern time. The Devils Hole calcite-vein surface has a U-series date of about 60 000 years before present (Ludwig *et al.*, 1992). Speleothem deposits in the Cave of Bells in southern Arizona yielded U-series dates extending from 9000 to 55 000 years before present (Wagner *et al.*, 2002). Common to many speleothem studies, the $\delta^{13}C$ record may be difficult to interpret due to lack of constraint on the degassing of carbon dioxide when the water enters the cave atmosphere and possible disequilibrium with soil CO_2 in the recharge zone (Baker *et al.*, 1997). Fortunately, the stable isotopes of hydrogen and oxygen in water often exhibit more conservative behavior in speleothems. The Cave of

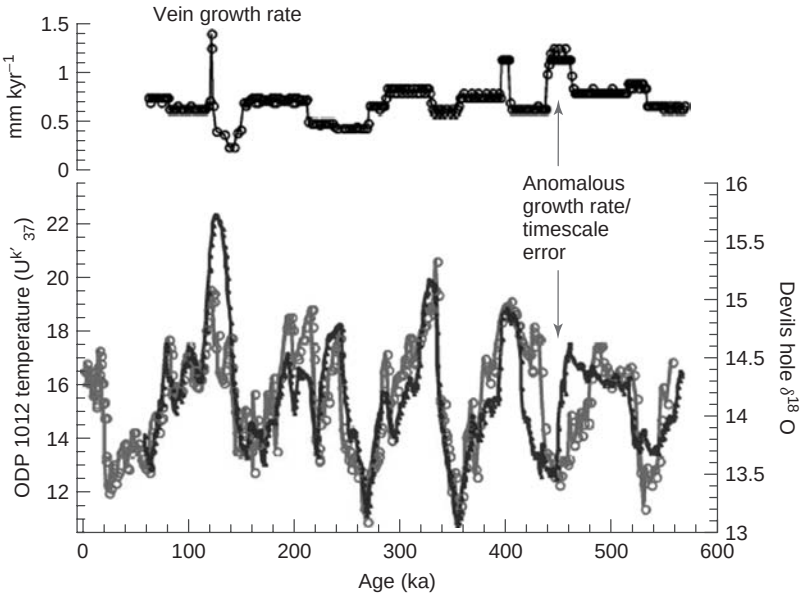


Figure 3 Comparison of a representative alkenone record (ODP 1012) with the Devils Hole δ^{18}O record, on independent timescales. The alkenone record represents sea surface temperature along the California Margin. The good match, in particular the rise in temperature and δ^{18}O of both records before deglaciation, suggests that regional, not global, temperature changes dominate the isotopic composition of the terrestrial record. The growth rate of the Devils Hole vain inferred from the Devils Hole timescale contains an anomalous interval beginning at 438 ka, suggestive of a dating anomaly. (Reproduced from Herbert *et al.* (2001) with permission of American Association for the Advancement of Science) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Bells speleothems δ^{18}O record reveal rapid changes during the last deglaciation that appear synchronous, within dating error, with the deglacial sequence of the GISP-2 δ^{18}O Greenland ice-core record (Wagner *et al.*, 2002). The rapid transit between recharge and speleothem deposition at Cave of Bells increases our confidence that the δ^{18}O record more directly reflects δ^{18}O of recharge water conditions at the surface (Wagner *et al.*, 2002). Speleothem research promises to play an important role in climate-variability studies in the future, particularly if the paleohydrology for each speleothem setting is well understood.

TRANSIENT RESIDENCE TIME ELEMENTS OF HYDROLOGIC CYCLE: PROXY RECORDS FOR PALEOHYDROLOGY

Precipitation History

Rodent Middens: Proxy for Precipitation

Locations where climate is extreme can serve as ideal places for archives because climate shifts create a dramatic and unmistakable local response. Semiarid and arid regions often have sharp climate gradients where shifts toward more pluvial (i.e. wetter) periods result in the sudden appearance of vegetation or a shift in vegetation species. Rodent middens have become an important source of information in arid and semiarid regions where lakes, glaciers, and other rich paleoclimate archives are often not available. Fossil rodent middens, as defined by Latorre *et al.* (2002), are "amalgamations of rodent feces and plant, insect, and vertebrate remains encased in hardened urine (amberat) and commonly preserved in rock shelters, caves, and crevices". The plant record preserved in a midden reflects the vegetation within about 100 meters of the midden site, the typical foraging range for most rodent species (Betancourt *et al.*, 2000). Packrat middens in the southwest United States, stick-nest rat middens in Australia, hyrax and dassie rat middens in Africa, and a variety of rodents (genera *Abrocoma, Lagidium, Octodontomys*, and *Phyllotis*) in South America have proven to be valuable archives for paleoclimate research (Pearson and Betancourt, 2002).

A successful reconstruction of the history of precipitation requires a number of steps. After sorting through plant remains in middens, the vegetation zones for the region must first be carefully mapped out and understood in terms of temperature, precipitation volume, elevation, and latitude or longitude factors. For example, Figure 4 depicts the major vegetation zones used to study rodent middens from Calama-Salar de Atacama basins in northern Chile. Second, specimens from the largest size middens are identified to the highest taxonomic resolution possible and a portion of the midden is extracted for pollen identification (Lyford *et al.*, 2002). Third, fecal pellets or plant remains from middens are dated by conventional or Accelerator

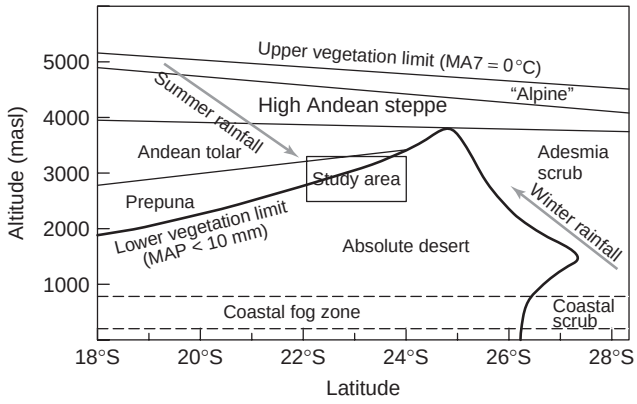


Figure 4 Example of a diagram used for the major vegetation zones found in northern Chile for a rodent midden study. (Reproduced from Latorre *et al.* (2002) by permission of Geological Society of America)

Mass Spectrometry (AMS) radiocarbon dating methods (Betancourt *et al.*, 2000; Lyford, *et al.*, 2002; Latorre, *et al.*, 2002). The final step is to correlate modern vegetation-assemblage patterns that reflect precipitation, temperature, and elevation aspects with midden vegetation changes. At a given midden location, presumably over the radiocarbon time period, the major orography, latitude, and longitude of the site have remained relatively constant. Thus, any changes in vegetation detected in the midden represent local shifts in precipitation and temperature.

The limitations of this method are that the middens are hard to find leaving large gaps in the stratigraphic record, and that pollen from middens often suffers from poor spatial and taxonomic resolution (Pearson and Betancourt, 2002). The strengths of the method are that middens are available in semiarid regions, the macrofossils have high spatial resolution and excellent taxonomic resolution while the available pollen provides quantitative vegetation reconstruction (Pearson and Betancourt, 2002).

Rodent Middens: Evidence for Climate Variability and Regional Teleconnections

The above methods have been successfully applied to reconstruct precipitation history for many arid regions of the world. Presented below is a case study in Chile that illustrates how the paleohydrologic evidence refined prevailing theories of the major climate-forcing processes for the region. The midden record in the Atacama Desert of northern Chile spans over the past 40 000 years, but is strongest with regard to stratigraphic continuity and reliable radiocarbon dates over the past 22 000 years. The present extreme climate conditions make the Atacama Desert part of the distinct group of driest regions on earth. The South Pacific Anticyclone restricts Pacific moisture transport, while the high elevation of the Andes blocks most of the moisture from the Amazon Basin to help create the extreme desert (Placzek *et al.*, 2001). Heating in summer of the warm-cored, closed, anticyclone develops the Bolivian High over the Altiplano which intensifies the trade winds and creates the South American Summer Monsoon precipitation (Zhou and Lau, 1998; Placzek *et al.*, 2001).

The Atacama paleohydrology record differs from the central Andes paleoclimate reconstructions regarding the timing and maximum wetness and aridity since the LGM (Betancourt *et al.*, 2000). The rodent midden record for the Atacama indicates that between 16.2 and 10.5 kyr B.P. precipitation was 70 to 100 mm per year at 2600 m elevation and 100 to 150 mm per year at 3000 to 3200 m elevation (Betancourt *et al.*, 2000; Latorre *et al.*, 2002). These elevations today experience mean annual precipitation of 20 mm per year at the lower elevation and 50 to 70 mm per year at the higher elevation. Another advantage of rodent midden vegetation records is that the seasonality of precipitation can be inferred which is often not possible with lake level, or paleogroundwater reconstructions. This is possible because a plant classification system exists (i.e. C3, C4, or Crassulacean acid metabolism (CAM)) that is based on photosynthetic pathway processes that result in different carbon isotope signatures. The dominance of summer-flowering herbs, C4 plants, and southward displacement of northern species in the rodent midden record during the wettest period over the past 22 000 years (16.2–10.5 kyr B.P.) is strong evidence for increased summer precipitation events (Betancourt *et al.*, 2000; Latorre *et al.*, 2002).

Prior to the discovery of the rodent midden record, researchers linked fluctuations in the South American Summer Monsoon precipitation to fluctuations in summer insolation in the Southern Hemisphere (Rech *et al.*, 2002). Yet, the period for wettest conditions at the Atacama Desert coincides with a minimum in austral-summer insolation (Figure 5). Local records around Atacama all indicate a pluvial period between 10.2 and 9 kyr B.P. (Betancourt *et al.*, 2000; Latorre, *et al.*, 2002; Rech *et al.*, 2002; Rech *et al.*, 2003). This has led researchers to look for larger teleconnections to explain the Atacama pluvial period. During El Niño events, westerly wind anomalies inhibit moisture advection from the east to the western part of the Bolivian Altiplano and central Atacama. Whereas La Niña conditions favor strong easterly winds that create convection and precipitation in the region. Combined paleohydrologic evidence led Betancourt *et al.* (2000) to conclude that strengthened Pacific trade winds (easterlies) during persistent La Niña conditions of the late glacial-early Holocene period created the pluvial period of the Atacama region.

The rodent midden record has contributed significant paleoprecipitation evidence regarding temporal resolution, a quantitative precipitation volume reconstruction, and links to the dominant season for increased precipitation (summer). The debate over why sustained millennial pluvial

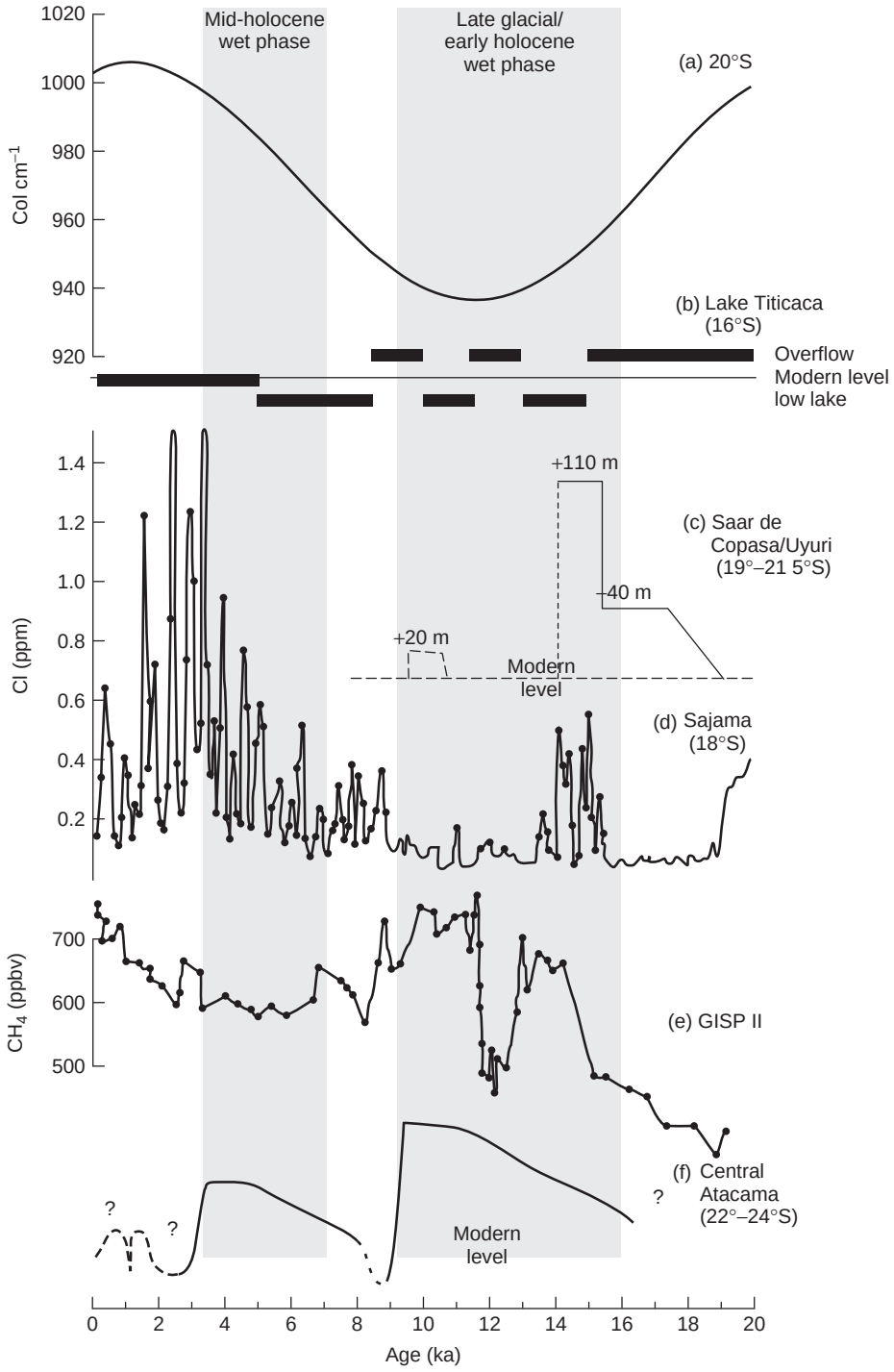


Figure 5 South American summer insolation at 20 °S and paleoclimatic records from the central Andes (a–d) and the global methane record (e) and wetland records for groundwater discharge reconstructions (f) figures. (Reproduced from Rech *et al.* (2002) by permission of Geological Society of America)

periods occurred in the Atacama region will continue, but certainly this evidence points to global teleconnections as potential mechanisms such as the Walker circulation combined with orbital scale forcing such as insolation anomalies.

Tree-rings: High-resolution Proxy Record for Precipitation

Tree-rings perhaps come closest as the ideal tool for terrestrial paleoclimate research. They satisfy many of the characteristics needed: high temporal resolution (annual and

even seasonal in some cases), wide spatial distribution, and clear climate-forcing parameters. Tree-ring width strongly correlates with precipitation and temperature trends. In order to create precipitation reconstructions, cores taken from several trees at each site are detrended to remove tree aging effects and the detrended core widths are averaged to reduce nonclimatic noise specific to individuals (Meko *et al.*, 1993). To further illustrate the tree-ring method, several studies in western North America serve as representative examples.

To establish the validity of tree-ring precipitation reconstruction, Dettinger *et al.* (1998) compared the tree-ring record with the gridded monthly precipitation anomalies from the instrument record (1880–1965) for the western United States region. Zonally averaged empirical orthogonal functions of the instrumental precipitation data reveal an interannual (3–7 years) and decadal "north–south seesaw of precipitation pivoting near 40 °N" that are strongly linked with sea level pressure variations (Cayan *et al.*, 1998; Dettinger *et al.*, 1998). The tree-rings matched the same pattern: (i) Southwestern US precipitation increased during El Niño events while Pacific Northwest precipitation increased during La Niña events (Dettinger *et al.*, 1998). Similar patterns persist over the entire 260-year tree-ring chronology. In contrast to the instrument record, the tree-ring chronology is long enough to also suggest a very low frequency variation with greater than 30-year periods (Dettinger *et al.*, 1998). As more tree-ring records are produced around the world, perhaps a new understanding of climate teleconnections over the very low frequency time periods will emerge.

Surface Water Discharge Reconstructions

Paleodischarge: Sediment and Coral Records of Paleodischarge

Sediment records are strongest for recording flood events and weakest at preserving drought occurrences. It is difficult to generate a quantitative discharge volume from sediment records. However, flood frequency is available through various geomorphic and geochemical dating techniques and provides another important method for climate-variability research. A few representative examples follow from three depositional environments: floodplain, estuary, and ocean.

Floodplain sediment cores combined with ^{210}Pb dating can provide annual resolution of large discharge events that deposited sediments in floodplains over the last century. Aalto *et al.* (2003) employed X-ray analysis and clay-normalized ^{210}Pb activity of floodplain sediment cores to correlate 20–80 cm thick packages of low-energy silt-rich sediment packages. The cores were collected from the Beni and Mamore river basins in the Andean–Amazonian foreland that comprises 720 000 km^2 of the Amazon Basin. The authors interpreted the sediment record as a series of crevasse splays formed during levee failure. Aside from a minimum flood threshold of 6000 m^3 s^{-1}, the method does not yield further refinement of flood discharge volumes. Instead, the power of the annual-resolution ^{210}Pb dating yielded a tight correlation for the timing of crevasse-splay occurrences with the cold-phase El Niño/Southern Oscillation (ENSO) or La Niña events. Aalto *et al.* (2003) found that rapidly rising floods were the preconditions for forming the crevasse-splay deposits and found no flooding after 1974 signaling a shift in response to ENSO forcing over the latter decades of the last century. Similar paleohydrologic records of discharge will be important for gathering many different watershed responses to climate-forcing mechanisms such as ENSO. Theoretically, the future may yield classification of various watershed characteristics that respond in predictable ways to variable climate-forcing processes such as ENSO.

Stratigraphic and radiocarbon dating techniques were used to form the chronostratigraphy of marsh sediments in the San Francisco Bay Estuary study by Goman and Wells (2000). The Sacramento and San Joaquin river drainages comprise over 40% of California's surface area and together provide 90% of the freshwater flow into the San Francisco Bay. Goman and Wells (2000) chose core locations to be at the confluence of the two rivers (Browns Island) and the modern mixing zone (Peyton Hill marsh) between freshwater and saline ocean water. Through a combination of techniques to characterize each depositional facies (i.e. grain-size analysis, organic content, iron concentration, vegetation identification of macrofossils, and Accelerator Mass Spectrometry (AMS) radiocarbon dates on seeds) the authors were able to identify periods of sedimentation from freshwater floods that outpaced sea-level rise versus other periods of more saline conditions with decreased freshwater outflow from the rivers. The vegetation analysis allowed for salinity reconstruction of the estuary. Goman and Wells (2000) identified a 2000-year period beginning in 3800 calendar year B.P. where the Sacramento and San Joaquin rivers had higher flows than modern discharge. These dates correlate with other studies that record evidence of increased precipitation as recorded by lake levels (e.g. Mono Lake and a shallow lake that existed in the Mojave Desert). The two largest flood events, preserved in the sediment, occurred at 3600 and 530 calendar year B.P. In addition to extreme floods, the record preserved the sustained 2000-year period of increased river discharge. As a climate-forcing mechanism, Goman and Wells (2000) proposed deep penetration of cyclonic winter storms further south than today due to a significant low pressure system formed along the west coast from a weakened subtropical high in the eastern North Pacific. The longer records provided by paleohydrologic research enable researchers to evaluate different climate

patterns that have not been recorded by the relatively short-term instrumental record.

Researchers have measured the Ba/Ca in Australia's Great Barrier Reef corals to reconstruct sediment discharge from the Burdekin River (McCulloch *et al.*, 2003). Barium sorbed to river suspended particles will desorb once the freshwater encounters the higher salinity ocean water. McCulloch *et al.* (2003) exploited the fact that corals incorporate the local seawater Ba/Ca ratio into their skeletons to track the times in the past when the terrestrial river source of barium swamped the typical seawater Ba/Ca ratio. The researchers analyzed a core from a 250-year-old coral to create a proxy record for local flood discharge events. They checked the proxy method and found that the spikes in coral Ba/Ca record correlated with historical Burdekin river discharge recorded since 1921. The longer coral record suggests that natural climate variability was overwhelmed by the impact of European settlement since 1862. Changes in grazing and agricultural land-use practices created an increase in coral Ba/Ca ratios by 5–10 orders of magnitude over pre-settlement ratios (McCulloch *et al.*, 2003). The exciting prospect of this novel technique is that the widespread distribution of corals in tropical latitudes may permit the extensive reconstruction of high-resolution paleodischarge records. These reconstructions may prove critical for future climate-variability studies.

Extreme Floods: Paleohydrology Reconstruction of the Earth's Largest Floods

At various points in the Earth's history, ice sheets captured and retained significant amounts of freshwater for many thousands of years. The growth, prolonged existence, and decay of these ice sheets created significant feedbacks into the global climate system. Ice sheets increase the global albedo thereby altering the world's energy balance, and ice sheets over a kilometer thick can alter major atmospheric circulation (e.g. Herbert *et al.*, 2001). Ice sheets also tend to form dams that capture meltwater into glacial lakes, some of which have been enormous. Breaches of dams retaining these immense glacial lakes created megafloods that delivered freshwater to sensitive ocean basins which decreased deep-water formation and ultimately impacted global climate (Leverington *et al.*, 2000; Fisher *et al.*, 2002; Perkins, 2002a; Fisher, 2003).

The paleohydrologist who studies extreme floods confronts a common problem in geology: are modern processes the key to understanding the past or are some events so extreme that notions of uniformitarianism do not apply (Baker, 2002a)? Most of the largest floods on earth have occurred when dams retaining major bodies of water have been breached, as has been the case with natural dams storing large volumes of meltwater from ice-covered regions.

Megafloods have altered the landscape through major erosion and have left depositional features for the geomorphologist to decipher. A combination of geomorphology (e.g. scour marks, streamlined depositional features, boulders, giant current dunes, washload, slope-area features) and principles of fluid hydraulics (e.g. Chezy and Manning equations and Froude number) provide tools for paleohydraulic peak flow reconstructions (Baker, 2002b). Large boulders transported and ultimately deposited with associated downstream scour marks help quantify flow velocity and bed shear stress values (Baker and Ritter, 1975). Equally important are the relationships between the range of sediment sizes carried as washload, suspended load, and bedload for reconstructing sustained bed shear stresses. Geomorphic evidence for washload sizes can be found in deposits located high above the paleochannel floor (O'Connor, 1993). Paleoflood water-surface levels for peak discharge have been determined in settings where the geometry of the valley and scoured divide crossings were mapped in detail for paleohydraulic calculations (Baker, *et al.*, 1993). The reconstructed hydraulic properties of the two largest floods on Earth are breathtaking. The Pleistocene Missoula floods of the Channeled Scabland region of Washington State and the Altay Mountain floods of Siberia had peak flow velocities of around $25\,m\,s^{-1}$, peak discharges of around $2 \times 10^7\ m^3\ s^{-1}$, maximum bed shear stresses of 10 000 to 40 000 $N\,m^{-2}$, and peak stream power per unit area values of 2×10^5 to $1 \times 10^6\ W\,m^{-2}$ respectively (Baker, 2002b). When one considers that the Sverdrup unit used by oceanographers for ocean currents flow volume ($1\,Sv = 10^6\ m^3\ s^{-1}$) is an order of magnitude smaller, one could imagine that flood flow volumes on the order of $10^7\ m^3\ s^{-1}$ could impact global climate.

Extreme Floods: Paleohydrologic Links to Younger Dryas Cold Period

In contrast to the more subtle, yet significant climate impact of water as it continuously travels through the hydrologic cycle, the extreme floods around 13 000 and 11 350 years ago perhaps represent the most direct and spectacular links between hydrology and global climate. Paleoclimatologists have gathered strong evidence for glacial and interglacial cycles, improved dating for these transitions, and established links to Milankovitch solar insolation forcing along with a host of other indicators for the synchronous, filtered, or lagged responses in ice, ocean, land, flora, and fauna. The most intensely studied glaciation-and-deglaciation cycle is the last one, since so many of the paleo-indicators have not yet been erased from the record. Even if the first-order forcing is understood for major climate trends such as glacial cycles, there are features recorded in the paleoclimate record that reveal specific paleotectonic, paleolatitude, paleohydrologic, and other conditions that make each glacial cycle unique. For example, a prominent cold

period around 13 000 to 11 000 years ago stood out in contrast to the overall interglacial climate rebound from the LGM around 18 000 years ago. This period is known as the *Younger Dryas*.

Mounting paleoclimate evidence has emerged for the Younger Dryas. Where sediment cores show dramatic increases in ice rafted debris indicative of large "iceberg armadas" that traversed the North Atlantic, both ocean surface salinity and SST in the region decreased (Broecker *et al.*, 1988, 1990). North Atlantic Deep Water formation significantly decreased, effectively diminishing the northeastward movement of warm, tropical ocean waters that normally kept Europe relatively warm (Broecker and Denton, 1989). Dramatic shifts in lake sediment core pollen records indicated a systematic change toward colder vegetation species and surges in mountain glaciers.

Ironically, it was probably the release of water that had been stored in the frozen state for so many years over North America that created the cold excursion during the earth's overall warming trend. Recent paleohydrology evidence, summarized in Table 2, suggests the sequence of hydrologic processes that eventually led to the Younger Dryas cold period. On the North American continent during the glacial maximum 20 000 to 18 000 years ago, the Laurentide Ice Sheet extended over most of Canada east of the Rockies and southward into present-day Iowa and Illinois, and was thickest (~5 km) over present Hudson Bay (Leverington *et al.*, 2000; Fisher *et al.*, 2002; Perkins, 2002a; Fisher, 2003). Figure 6 depicts the Laurentide Ice sheet and glacial lake Agassiz with arrows delineating the three major drainage outlets used at different times as the ice sheet

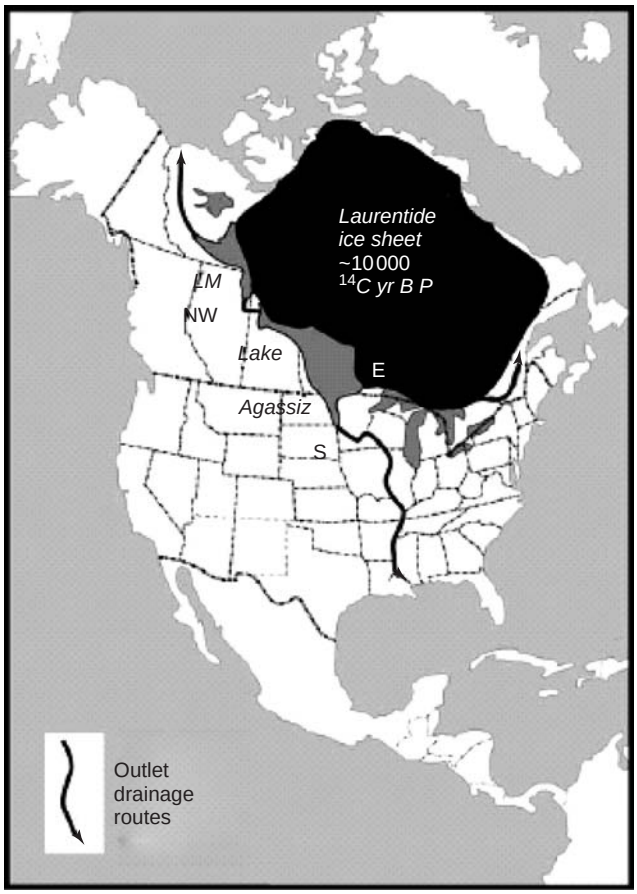


Figure 6 Location of glacial Lake Agassiz and its three main outlets. Abbreviations: S, south; E, east; NW, northwest; LM, glacial Lake McConnell. (Reproduced from Fisher (2003) by permission of Elsevier)

Table 2 Example of a direct link between hydrology and global climate change. Events surrounding the Younger Dryas and Preboreal Oscillation cold periods (after Leverington *et al.*, 2000; Fisher *et al.*, 2002; Perkins, 2002a; Fisher, 2003)

Time period (years B.P.)	Paleohydrologic conditions
20 000–18 000	Laurentide Ice Sheet at maximum extent
Prior to 13 000	Laurentide Ice Sheet retreats past the divide that separates Mississippi watershed from northward drainage. Meltwater pools against ice sheet to form glacial Lake Agassiz (~134 000 km^2). Overflow spilled down the Mississippi River drainage valley.
13 000	New spillway to the east was breached creating a sudden Lake Agassiz level drop (~100 m) and 9500 km^3 of freshwater flooded through the Great Lakes and exited the St. Lawrence River into the North Atlantic. Over the next several decades, freshwater discharge remained twice as large as the discharge prior to the breach down the St. Lawrence River.
13 000–11 400	Younger Dryas 1600 year cold period.
11 400–11 350	Laurentide Ice Sheet grew, extended southward and blocked St. Lawrence drainage to the North Atlantic. Meltwater again flowed down the Mississippi River valley.
	North Atlantic Deep Water formation increased and a return to warmer conditions prevailed.
11 350–11 100	Preboreal Oscillation occurred due to a 50-m drop in Lake Agassiz level over several years. Water drained out the Mackenzie River valley into the Arctic Ocean. This freshwater froze into sea-ice and circulated out through Fram Strait and again into the North Atlantic Ocean where it melted and shut down deep water formation.
8500–8100	Laurentide Ice Sheet collapsed over Hudson Bay and released 163 000 km^3 of water through Hudson Bay and into the North Atlantic, raising global sea level by 0.5 m. The next 400 years was a cold period (Figure 7).

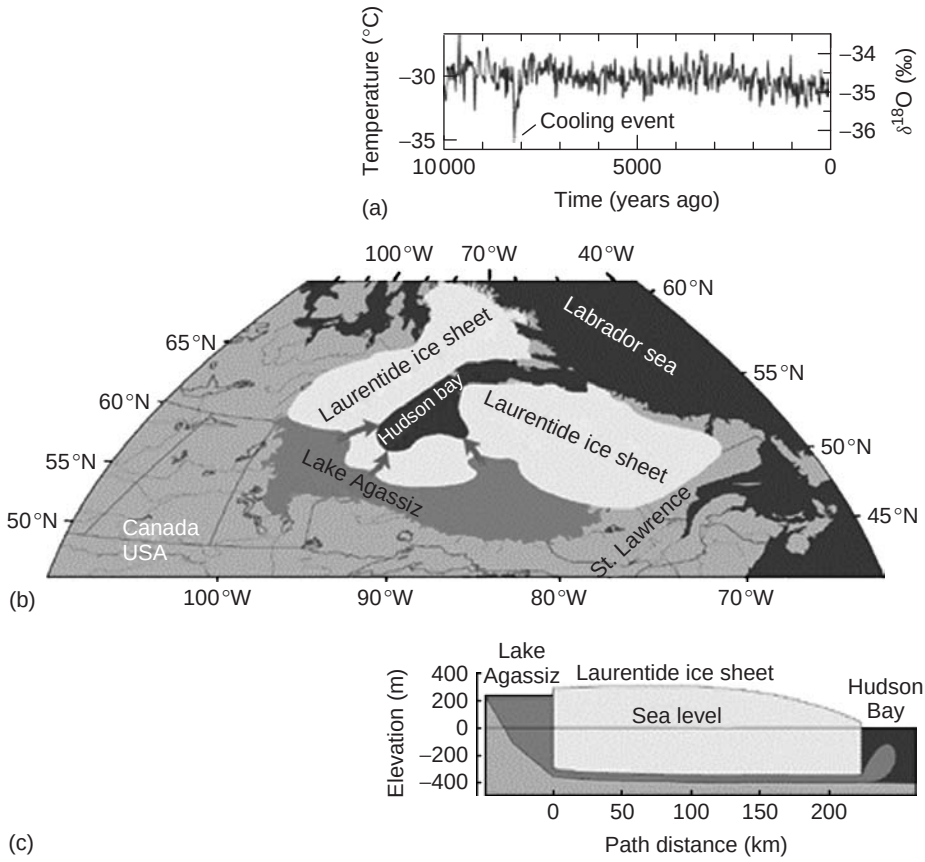


Figure 7 The climate, the lake, and the dam (a) The climate record from ice cores in central Greenland reveals an abrupt cooling event about 8200 years ago. Data from Dansgaard *et al.* (1993) (b) Lake Agassiz formed at the southern margin of the disintegrating Laurentide Ice Sheet and released its stored water to Hudson Bay. Possible flood routes are indicated by arrows. Many more routes are possible (c) At the time of the flood, the ice dam was probably several hundred kilometers wide. Subglacial drainage from the lake to Hudson Bay would have started when the pressure of lake water approached that for flotation of the dam. (Reproduced from Clarke *et al.* (2003) by permission of American Association for the Advancement of Science) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

covered or retreated from major continental drainage divides.

Over the last decade, researchers have debated whether a large volume of freshwater released into sensitive regions could impact the world's climate enough to reverse a global warming trend for over a millennium. North Atlantic Deep Water, formed in the subpolar seas of the North Atlantic (Greenland Sea, Iceland Sea, Norwegian Sea, and Labrador Sea), drives the global circulation system that distributes heat around the world's oceans. Deep water is created in the North Atlantic rather than in the similarly cold latitudes of the North Pacific because of the higher salinity of the North Atlantic (~35 psu vs. ~32 psu). North Atlantic Deep Water forms as saline, cold surface water cools, which increases the density and causes it to sink (Broecker *et al.*, 1988, 1989, 1990). A sufficiently large influx of cold, fresh, and more buoyant water to this region would decrease the surface density, thereby capping the North Atlantic and preventing deep-water formation during the winter months. These deep-water formation sites were directly affected as Lake Agassiz emptied into the St. Lawrence River valley and into the Mackenzie River to exit the Arctic Ocean through Fram Strait. When Lake Agassiz drained instead into the Mississippi River to the Gulf of Mexico, there would likely have been less impact on global climate. The recent paleohydrology work on Lake Agassiz provided convincing evidence for the timing and volume of flood events and provided concurrent terrestrial evidence for triggering events recorded in North Atlantic Ocean sediment cores during the Younger Dryas.

Droughts

The extreme events of hydrology, floods and droughts, both can wreak havoc on flora, fauna, and human interests such as water resources, agriculture, livestock, or property. A

better understanding of the climate-forcing parameters and regional responses may improve forecasting, and hopefully lead to improved seasonal preparation and damage mitigation from such extreme events. Drought cycles that extend further back in time than the instrumental records have the potential to confirm patterns that have emerged in the instrument record or uncover some extreme forcing conditions not yet experienced. As stated by Cook *et al.* (1999) "There are often too few realizations of proposed forcing mechanisms of drought in the short instrumental records to test any of them in a statistically rigorous way." Thus far, we have discussed the many proxy records for the presence of water. Drought, or the lack of water, changes the available climate proxy records. The fluctuations of closed-basin lakes are a sensitive and long-term indicator of drought, and vegetation response has also proven to be a successful proxy record of a lack of water. Ultimately, tree-ring records provide not only a sensitive indicator of drought but also provide a high annual resolution record over a millennial timescale.

Droughts: Closed-basin Lake Levels

Reconstruction of the fluctuations of closed-basin lakes is a sensitive indicator of drought. Closed-basin lake levels provide quantitative evidence for the changes in the entire basin water balance, which is otherwise difficult to obtain. All the various methods have several common elements: identify the paleoshoreline elevation, date the prehistoric lake level, and check for the possible tectonic shifts that may have influenced the shoreline elevation.

Most hydrographically closed lakes studies measured the multicentury to millennial-scale lake level changes (Komatsu *et al.*, 2001; Licciardi, 2001; Oviatt, 1997; Ridge and Larsen, 1990). The more recent lake level reconstruction of the past 3800 years at Mono Lake, located east of the Sierra Nevada Mountains in east-central California, illustrates this method (Stine, 1990 and Stine, 1994). Several rivers flow into Mono Lake and drop their sediment load to form deltas. The transgressive and regressive sequences in the deltaic sediments preserve the former fluctuations of the lake level. A progressive upward fining sediment sequence reflects the transition from a high-energy stream to the silts of deep lake water. As the lake recedes, in contrast, the same location would have an upwardly coarsening sediment sequence. Past Mono Lake high stands are preserved as berms, bars, clifflines, or pronounced breaks in vegetation. The engineer's level and rod was used to survey the prior lake levels.

Chronology at Mono Lake was provided by volcanic ash deposits, charcoal, or other organic debris that can be radiocarbon dated, and relict tree stumps. If the death of the tree stump is associated with shoreline transgression, then the radiocarbon age of the outer growth-ring provides the age of transgression. The annual tree-rings give the number of years the lake level was below the basal root elevation of the relict stump. Two severe and long-term drought periods were identified this way for Mono Lake; an over 200 year drought that ended around 1112 A.D. and a 140 year drought that ended around 1350 A.D. Stine (1994) linked these extreme droughts to a reorientation of midlatitude storm tracks that may have been associated with the warmer temperatures of the medieval period. Such findings could have profound implications for the future of California water resources if further warming occurs over the next century.

Droughts: Tree-ring Reconstructions

Lack of water during the growing season will restrict the width of a tree-ring for that year. Regional drought reconstructions rely on statistical techniques to find patterns of drought in a region. The measured ring widths of cores are handled the same way as described above for the paleoprecipitation research (Meko *et al.*, 1993). Similar methods have led to successful regional drought chronologies in the northeast United States (Cook and Jacoby, 1977), southeast United States (Cook *et al.*, 1988; Stahle *et al.*, 1985, 1988), western United States (Gray *et al.*, 2003; Graumlich, 1993; Stahle and Cleaveland, 1988; Stockton and Meko, 1975) and other parts of the world.

Recent efforts to create continental-scale drought records by combining the tree-ring chronologies to produce statistical data that correlate favorably with the Palmer Drought Severity Index (PDSI) based on the instrument record back to 1895 (Meko *et al.*, 1993; Cook *et al.*, 1996, 1999). These methods give confidence for reconstructions beyond the instrumental record. The primary elements of continental-scale drought reconstruction are as follows and are based upon the United States continental-scale reconstructions of Meko *et al.* (1993) and Cook *et al.* (1996, 1999). First, both the PDSI data and tree-ring data are converted to the same coarse grid scale for a more direct comparison of the derived results. For example, in the Cook *et al.* (1999) case, this resulted in a reduced PDSI network to 15% of the original size while the same grid for the tree-ring data had some data gaps or spatial holes in the Great Plain States due to lack of local tree-ring sites. Second, despite the known northward shifts in the timing of the growing season which can be directly compared on a regional scale between monthly PDSI and tree-ring grid data, they used the summer PDSI as the most comprehensive drought index in order to compare with the tree-ring grid across the entire continental scale. Third, the investigators employed various regression and statistical techniques to compare the tree-ring data for the same time period as the instrument-based PDSI results. Finally, the tree-ring grid data were put through various rotated principal component analysis techniques to search for factors that may be linked to patterns for drought variability. The extremely high correlation

between these two drought indicators for the period of historical record strongly supported the validity of the tree-ring chronologies further back in time. Perhaps more than any other paleohydrology method, the high-resolution tree-ring record with its extensive spatial and temporal distribution of measurements yields continental-scale reconstructions that are useful to global climate modelers.

Droughts: Climate Significance of Frequency and Spatial Patterns

Severe multiyear droughts have occurred in the United States about twice per century over the past 400 years and the causes have remained elusive (Woodhouse and Overpeck, 1998). The major findings of the United States continental-scale tree-ring drought reconstructions are (i) the 1930s drought (predominantly the northern Great Plains) was the most severe since 1700; (ii) the next most severe droughts occurred around 1820, 1860, 1950 (predominantly the southwest), and 1977; and (iii) the wettest periods clustered around 1820–1840 and 1900–1920 periods (Cook *et al.*, 1999; Gray *et al.*, 2003). The success of the tree-ring paleohydrology reconstruction of PDSI has led to a new analysis on the causes of persistent drought.

Earlier research suggested a strong link between La Niña conditions and drought in the southwest United States, but not consistently for each drought period (Cole and Cook, 1998). Hence the search for concurrent extratropical forcing conditions or lag responses continued. Cole *et al.* (2002) compared the updated record of ENSO, the tree-ring PDSI reconstruction, and the extended Pacific Decadal Oscillation (PDO) and found clear links between the oscillation phases and persistent drought conditions in the United States. The PDO is an index based on trends in SST in the North Pacific. The paleoclimate reconstructions showed that long-period La Niña events existed in earlier centuries in contrast to the typical short-period La Niña events prevalent in the twentieth-century record. Also, Cole *et al.* (2002) found a spatial shift in the drought centers from the Arizona/California region for the earlier periods to Texas/New Mexico region in the twentieth-century (Figure 8). La Niña conditions when the PDO remains in a negative state correlate with a persistent multiyear drought (Cole *et al.*, 2002). Such paleoclimate revelations may help climate forecasters to predict when economically devastating multiyear drought conditions are a high probability by tracking current ENSO and PDO conditions.

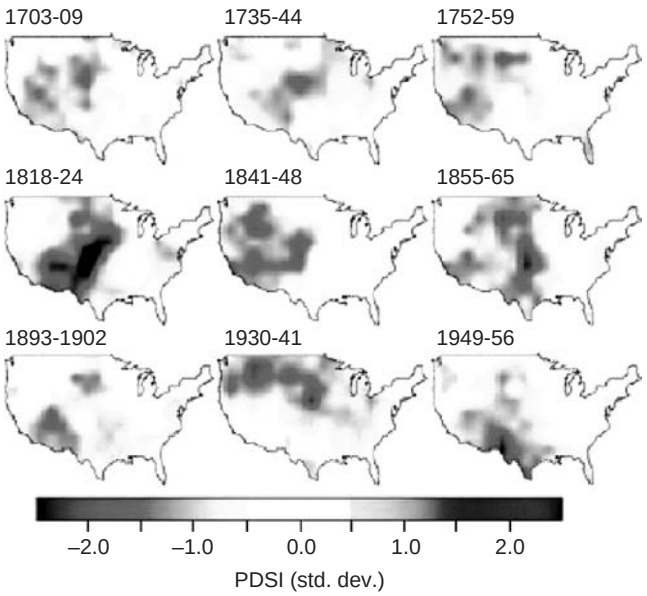


Figure 8 Mapped patterns of reconstructed Palmer Drought Severity Index (PDSI) for the multiyear droughts between 1700 and 1978. (Reproduced from Cole *et al.* (2002) by permission of American Geophysical Union) A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

CONCLUSIONS

Water plays a critical role in climate, and paleohydrology has helped elucidate climate variability prior to human recording of climate data. Over the past several decades, paleohydrology has made important contributions to climate knowledge such as refining the tropical terrestrial temperature shift between today and the LGM that helped resolve some of the conflicting tropical ocean proxy records. Often it is the extremes of the hydrological cycle, such as floods or drought that capture the public's attention when loss of crops, livestock, property or life occurs. Understanding the conditions in the past that created multidecadal megadroughts or extreme flooding may increase our chances for earlier detection and possible mitigation if similar ocean–atmosphere forcing conditions arise in the future. We expect that as more proxy records for the hydrologic cycle are discovered, paleohydrology will continue to advance our understanding of climate variability.

Acknowledgment

I wish to thank Soroosh Sorooshian who started me on this path of paleohydrology review. This chapter benefited from discussions with V. R. Baker, J. E. Cole, and J. D. M. Wagner. Review comments by F. Phillips and V. A. Childers significantly improved the manuscript.

REFERENCES

Aalto R., Maurice-Bourgoin L., Dunne T., Montgomery D.R., Nittrouer C.A. and Guyot J-L. (2003) Episodic sediment

accumulation on Amazonian flood plains influenced by El Niño/Southern Oscillation. *Nature*, **425**, 493–497.

Aeschbach-Hertig W., Stute M., Clark J.F., Reuter R.F. and Schlosser P. (2002) A paleotemperature record derived from dissolved noble gases in groundwater of the Aquia Aquifer (Maryland, USA). *Geochimica et Cosmochimica Acta*, **66**, 797–817.

Andrews J.N. and Lee D.J. (1979) Inert gases in groundwater from the Bunter Sandstone of England as indicators of age and palaeoclimatic trends. *Journal of Hydrology*, **41**, 233–252.

Baker V.R. (2002a) The Study of Superfloods. *Science*, **295**, 2379–2380.

Baker V.R. (2002b) High-energy megafloods: planetary settings and sedimentary dynamics. *Special Publications International Association of Sedimentology*, **32**, 3–15.

Baker V.R., Benito G. and Rudoy A.N. (1993) Paleohydrology of late Pleistocene superflooding, Altay Mountains, Siberia. *Science*, **259**, 348–350.

Baker A., Ito E., Smart P.L. and McEwan R.F. (1997) Elevated and variable values of ^{13}C in speleothems in a British cave system. *Chemical Geology*, **136**, 263–270.

Baker V.R. and Ritter D.F. (1975) Competence of rivers to transport coarse bedload. *Geological Society of America Bulletin*, **86**, 975–978.

Ballantine C.J. and Hall C.M. (1999) Determining palaeotemperature and other variables using noble gas concentrations in water. *Geochimica et Cosmochimica Acta*, **63**, 2315–2336.

Bard E. (1999) Ice age temperatures and geochemistry. *Science*, **284**, 1133–1134.

Betancourt J.L., Latorre C., Rech J.A., Quade J. and Rylander K.A. (2000) A 22 000-year record of monsoonal precipitation from northern Chile's Atacama Desert. *Science*, **289**, 1542–1546.

Beyerle U., Purtschert R., Aeschbach-Hertig W., Imboden D.M., Loosli H.H., Wieler R. and Kipfer R. (1998) Climate and groundwater recharge during the last glaciation in an ice-covered region. *Science*, **282**, 731–734.

Blunier T. and Brook E.J. (2001) Timing of millennial-scale climate change in Antarctica and Greenland during the last glacial period. *Science*, **291**, 109–112.

Brabson B.B. and Palutikof J.P. (2002) The evolution of extreme temperatures in the Central England temperature record. *Geophysical Research Letters*, **29**, 2163, doi:10.1029/2002GL015964.

Broecker W.S. (1992) Upset for Milankovitch theory. *Nature*, **359**, 779–780.

Broecker W.S., Andree M., Wolfli W., Oeschger H., Bonani G., Kennett J. and Peteet D. (1988) The chronology of the last deglaciation: implications to the cause of the younger Dryas event. *Paleoceanography*, **3**, 1–19.

Broecker W.S., Bond G., Klas M., Bonani G. and Wolfli W. (1990) A salt oscillator in the glacial Atlantic? I. The concept. *Paleoceanography*, **5**, 469–477.

Broecker W.S. and Denton G.H. (1989) The role of ocean-atmosphere reorganizations in glacial cycles. *Geochimica et Cosmochimica Acta*, **53**, 2465–2501.

Broecker W.S., Kennett J., Flower B., Teller J.T., Trumbore S., Bonani G. and Wolfli W. (1989) Routing of meltwater from the laurentide ice sheet during the younger Dryas cold episode. *Nature*, **341**, 318–321.

Cayan D.R., Dettinger M.D., Diaz H.F., Graham N.E. (1998) Decadal variability of precipitation over western North America. *Journal of Climate*, **11**, 3148–3166.

Chuine I., Yiou P., Viovy N., Seguin B., Daux V. and Ladurie E.L.-R. (2004) Historical phenology: grape ripening as a past climate indicator. *Nature*, **432**, 289–290.

Clark J.F., Davisson M.L., Hudson G.B. and Macfarlane P.A. (1998) Noble gases, stable isotopes, and radiocarbon as tracers of flow in the Dakota Aquifer, Colorado and Kansas. *Journal of Hydrology*, **211**, 151–167.

Clark J.F., Stute M., Schlosser P., Drenkard S. and Bonani G. (1997) A tracer study of the Floridan Aquifer in southeastern Georgia: implications for groundwater flow and paleoclimate. *Water Resources Research*, **33**, 281–289.

Clarke G., Leverington D., Teller J. and Dyke A. (2003) Superlakes, megafloods, and abrupt climate change. *Science*, **301**, 922–923.

Climate: Long-Range Investigation, Mapping, and Prediction (CLIMAP) Project Members (1981) Seasonal reconstruction of the Earth's surface at the last Glacial maximum. *Geological Society of America Map Chart Series*, **36**, 1–18.

Cole J.E. and Cook E.R. (1998) The changing relationship between ENSO variability and moisture balance in the continental United States. *Geophysical Research Letters*, **25**, 4529–4532.

Cole J.E., Overpeck J.T. and Cook E.R. (2002) Multiyear La Niña events and persistent drought in the contiguous United States. *Geophysical Research Letters*, **29**, 1647, 10.1029/2001GL013561.

Collon P., Kutschera W., Davids B., Fauerbach M., Harkewics R., Morrissey D., Sherrill B., Steiner M., Pardo R., Paul M., *et al.* (1998) First attempt at dating groundwater from the Great Artesian basin of Australia with ^{81}Kr using AMS, presented at ICOG-9, *Ninth International Conference on Geochronology, Cosmochronology and Isotope Geology*, Beijing, 20–26 August 1998.

Cook E.R. and Jacoby G.C. Jr (1977) Tree-ring-drought relationships in the Hudson Valley, New York. *Science*, **198**, 399–401.

Cook E.R., Kablack M.A. and Jacoby G.C. (1988) The 1986 drought in the southeastern United States: How rare an event was it? *Journal of Geophysical Research*, **93**, 14,257–14,260.

Cook E.R., Meko D.M., Stahle D.W. and Cleaveland M.K. (1996) Tree-ring reconstructions of past drought across the coterminous United States: tests of a regression method and calibration/verification results. In *Tree Rings, Environment, and Humanity*, Dean J.S., Meko D.M. and Swetnam T.W. (Eds.), Radiocarbon: pp. 155–169.

Cook E.R., Meko D.M., Stahle D.W. and Cleaveland M.K. (1999) Drought reconstructions for the continental United States. *Journal of Climate*, **12**, 1145–1162.

Coplen T.B., Winograd I.J., Landwehr J.M. and Riggs A.C. (1994) 500,000-year stable carbon isotopic record from Devils Hole, Nevada. *Science*, **263**, 361–365.

Dansgaard W., Johnsen S.J., Clausen H.B., Dahl-Jensen D., Gundestrup N.S., Hammer C.U., Hvidberg C.S., Steffensen J.P., Sveinbjörnsdottir A.E., Jouzel J., *et al.* (1993) Evidence

for general instability of past climate from a 250-kyr ice-core record. *Nature*, **364**, 218–220.

Dennis F., Andrews J.N., Parker A., Poole J. and Wolf M. (1997) Isotopic and noble gas study of chalk groundwater in the London Basin, England. *Applied Geochemistry*, **12**, 763–773.

Dettinger M.D., Cayan D.R., Diaz H.F. and Meko D.M. (1998) North-south precipitation patterns in western North America on interannual-to-decadal timescales. *Journal of Climate*, **11**, 3095–3111.

Edmunds W.M., Fellman E., Baba Goni I., McNeill G.W. and Harkness D.D. (1998) Groundwater, palaeoclimate and palaeorecharge in the southwest Chad Basin, Borno State, Nigeria. *Isotope Techniques in Studying Past and Current Environmental Changes in the Hydrosphere and the Atmosphere*, IAEA: Vienna, pp. 693–707.

Elliot T., Andrews J.N. and Edmunds W.M. (1999) Hydrochemical trends, palaeorecharge and groundwater ages in the fissured Chalk aquifer of the London and Berkshire Basins, UK. *Applied Geochemistry*, **14**, 333–363.

EPICA community members (2004) Eight glacial cycles from an Antarctic ice core. *Nature*, **429**, 623–628.

Fairbanks R.G. and Matthews R.K. (1978) The marine oxygen isotope record in Pleistocene coral, Barbados, West Indies. *Quaternary Research*, **10**, 181–196.

Farrera I., Harrison S.P., Prentice I.C., Ramstein G., Guiot J., Bartlein P.J., Bonnefille R., Bush M., Cramer W., von Grafenstein U., *et al.* (1999) Tropical climates at the last Glacial Maximum: a new synthesis of terrestrial palaeoclimate data. 1. Vegetation, lake-levels and geochemistry. *Climate Dynamics*, **15**, 823–856.

Fisher T.G. (2003) Chronology of glacial Lake Agassiz meltwater routed to the Gulf of Mexico. *Quaternary Research*, **59**, 271–276.

Fisher T.G., Smith D.G. and Andrews J.T. (2002) Preboreal oscillation caused by a glacial Lake Agassiz flood. *Quaternary Science Reviews*, **21**, 873–878.

Goman M. and Wells L. (2000) Trends in river flow affecting the northeastern reach of the San Francisco Bay Estuary over the past 7000 years. *Quaternary Research*, **54**, 206–217. doi:10.1006/qres.2000.2165.

Graumlich L.J. (1993) A 1000-year record of temperature and precipitation in the Sierra Nevada. *Quaternary Research*, **39**, 249–255.

Gray S.T., Betancourt J.L., Fastie C.L. and Jackson S.T. (2003) Patterns and sources of multidecadal oscillations in drought-sensitive tree-ring records from the central and southern Rocky Mountains. *Geophysical Research Letters*, **30**, 1316, doi:10.1029/2002GL016154.

Greene A.M., Seager R. and Broecker W.S. (2002) Tropical snowline depression at the last glacial maximum: comparison with proxy records using a single-cell tropical climate model. *Journal of Geophysical Research*, **107**, ACL41–ACL418, (DOI 10.1029/2001JD000670).

Guilderson T.P., Fairbanks R.G. and Rubenstone J.L. (1994) Tropical temperature variations since 20,000 years ago: modulating interhemispheric climate change. *Science*, **263**, 663–665.

Hansen J., Lacis A., Rind D., Russell G., Stone P., Fung I., Ruedy R. and Lerner J. (1984) Climate sensitivity: analysis of feedback mechanisms. In *Climate Processes and Climate Sensitivity*, Hansen J.E. and Takahashi T. (Eds.), Geophysical Monograph 29, American Geophysical Union: Washington, pp. 130–163.

Heaton T.H.E., Talma A.S. and Vogel J.C. (1986) Dissolved gas palaeotemperatures and ^{18}O variations derived from groundwater near Uitenhage, South Africa. *Quaternary Research*, **25**, 79–88.

Herbert T.D., Schuffert J.D., Andreasen D. and Heusser L. (2002) Response: the California current, Devils Hole, and Pleistocene climate. *Science*, **296**, 7.

Herbert T.D., Schuffert J.D., Andreasen D., Heusser L., Lyle M., Mix A., Ravelo A.C., Stott L.D. and Herguera J.C. (2001) Collapse of the California Current during glacial maxima linked to climate change on land. *Science*, **293**, 71–76.

Hostetler S.W. and Clark P.U. (1997) Climatic controls of western U.S. glaciers at the last glacial maximum. *Quaternary Science Review*, **16**, 505–511.

Imbrie J., Boyle E.A., Clemens S.C., Duffy A., Howard W.R., Kukla G., Kutzbach J., Martinson D.G., McIntyre A., Mix A.C., *et al.* (1992) On the structure and origin of major glaciation cycles 1. Linear responses to Milankovitch forcing. *Paleoceanography*, **7**, 701–738.

Imbrie J., Mix A.C., Martinson D.G. (1993) Milankovitch theory viewed from Devils Hole. *Nature*, **363**, 531–533.

Komatsu G., Brantingham P.J., Olsen J.W. and Baker V.R. (2001) Paleoshoreline geomorphology of Böön Tsagaan Nuur, Tsagaan Nuur and Orog Nuur: the Valley of Lakes, Mongolia. *Geomorphology*, **39**, 83–98.

Latorre C., Betancourt J.L., Rylander K.A. and Quade J. (2002) Vegetation invasions into absolute desert: a 45 000 yr rodent midden record from the Calama-Salar de Atacama basins, northern Chile (lat 22°-24 °S). *GSA Bulletin*, **114**, 349–366.

Lee K.E. and Slowey N.C. (1999) Cool surface waters of the subtropical North Pacific Ocean during the last glacial. *Nature*, **397**, 512–514.

Leverington D.W., Mann J.D. and Teller J.T. (2000) Change in the bathymetry and volume of glacial Lake Agassiz between 11 000 and 9300^{14}C yr B.P. *Quaternary Research*, **54**, 174–181.

Licciardi J.M. (2001) Chronology of latest Pleistocene lake-level fluctuations in the pluvial Lake Chewaucan basin, Oregon, USA. *Journal of Quaternary Science*, **16**, 545–553.

Liu K.-B., Shen C. and Louie K.-S. (2001) A 1,000-year history of typhoon landfalls in Guangdong, southern China, reconstructed from Chinese historical documentary records. *Annals of the Association of American Geographers*, **91**, 453–464.

Ludwig K.R., Simmons K.R., Szabo B.J., Winograd I.J., Landwehr J.M., Riggs A.C. and Hoffman R.J. (1992) Mass-spectrometric ^{230}Th-^{234}U-^{238}U dating of the Devils Hole calcite vein. *Science*, **258**, 284–287.

Lyford M.E., Betancourt J.L. and Jackson S.T. (2002) Holocene vegetation and climate history of the northern Bighorn Basin, southern Montana. *Quaternary Research*, **58**, 171–181.

Magnuson J.J., Robertson D.M., Benson B.J., Wynne R.H., Livingstone D.M., Arai T., Assel R.A., Barry R.G., Card V., Kuusisto E., Granin N.G., Prowse T.D., Stewart K.M., Vuglinski V.S. (2000) Historical trends in lake and river ice cover in the northern hemisphere. *Science*, **289**, 1743–1746.

Manabe S. and Broccoli A.J. (1985) A comparison of climate model sensitivity with data from the last glacial maximum. *Journal of Atmospheric Science*, **42**, 2643–2651.

Matthews R.K. and Poore R.Z. (1980) Tertiary $\delta^{18}O$ record and glacio-eustatic sea-level fluctuations. *Geology*, **8**, 501–504.

McCulloch M., Fallon S., Wyndham T., Hendy E., Lough J. and Barnes D. (2003) Coral record of increased sediment flux to the inner Great Barrier Reef since European Settlement. *Nature*, **421**, 727–730.

Meko D., Cook E.R., Stahle D.W., Stockton C.W. and Hughes M.K. (1993) Spatial patterns of tree-growth anomalies in the United States and Southeastern Canada. *Journal of Climate*, **6**, 1773–1786.

Mix A.C., Morey A.E. and Pisias N.G. (1999) Foraminiferal faunal estimates of paleotemperature: Circumventing the no-analog problem yields cool ice age tropics. *Paleoceanography*, **14**, 350–359.

O'Connor J.E. (1993) Hydrology, hydraulics and sediment transport of Pleistocene Lake Bonneville flooding on the Sanke River, Idaho. *Geological Society of America Special Paper*, **274**, 1–83.

O'Connor J.E. and Baker V.R. (1992) Magnitudes and implications of peak discharges from glacial Lake Missoula. *Geological Society of America Bulletin*, **104**, 267–279.

Oviatt C.G. (1997) Lake Bonneville fluctuations and global climate change. *Geology*, **25**, 155–158.

Pearson S. and Betancourt J.L. (2002) Understanding arid environments using fossil rodent middens. *Journal of Arid Environments*, **50**, 499–511.

Perkins S. (2002a) Once upon a lake: the life, times, and demise of the world's largest lake. *Science News*, **162**, 283–284.

Perkins S. (2002b) Climate: Chinese records show typhoon cycles. *Science News*, **162**, 174.

Petit J.R., Jouzel J., Raynaud D., Barkov N.I., Barnola J.-M., Basile I., Bender M., Chappellaz J., Davis M., Delaygue G., *et al.* (1999) Climate and atmospheric history of the past 420 000 years from the Vostok ice core, Antarctica. *Nature*, **399**, 429–436.

Phillips F.M., Bentley H.W., Davis S.N., Elmore D. and Swannick G.B. (1986) Chlorine-36 dating of very old ground water II: Milk River aquifer, Alberta. *Water Resources Research*, **22**, 2003–2016.

Phillips F.M., Tansey M.K. and Peeters L.A. (1989) An isotopic investigation of groundwater in the central San Juan Basin, New Mexico: carbon 14 dating as a basis for numerical flow modeling. *Water Resources Research*, **25**, 2259–2273.

Placzek C., Quade J. and Betancourt J.L. (2001) Holocene lake-level fluctuations of Lake Aricota, southern Peru. *Quaternary Research*, **56**, 181–190.

Rech J.A., Pigati J.S., Quade J. and Betancourt J.L. (2003) Re-evaluation of mid-Holocene deposits at Quebrada Puripica, northern Chile. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **194**, 207–222.

Rech J.A., Quade J. and Betancourt J.L. (2002) Late Quaternary paleohydrology of the central Atacama Desert (lat 22°-24 °S), Chile. *GSA Bulletin*, **114**, 334–348.

Ridge J.C. and Larsen F.D. (1990) Re-evaluation of Antevs' New England varve chronology and new radiocarbon dates of sediments from glacial Lake Hitchcock. *Geological Society of America Bulletin*, **102**, 889–899.

Rind D. and Peteet D. (1985) Terrestrial conditions at the last glacial maximum and CLIMAP sea surface temperature estimates: are they consistent? *Quaternary Research*, **24**, 1–22.

Rudolph J., Rath H.K. and Sonntag C. (1984) Noble gases and stable isotopes in ^{14}C-dated paleowaters from central Europe and the Sahara. *Isotope Hydrology*, IAEA-SM-270, IAEA: Vienna, pp. 467–477.

Shackleton N.J. (1993) Last interglacial in Devils Hole. *Nature*, **362**, 596.

Stahle D.W. and Cleaveland M.K. (1988) Texas drought history reconstructed and analyzed from 1698–1980. *Journal of Climate*, **1**, 59–74.

Stahle D.W., Cleaveland M.K. and Hehr J.G. (1985) A 450-year drought reconstruction for Arkansas, United States. *Nature*, **316**, 530–532.

Stahle D.W., Cleaveland M.K. and Hehr J.G. (1988) North Carolina climate changes reconstructed from tree rings: A.D. 372–1985. *Science*, **240**, 1517–1519.

Stauffer B. (1999) Cornucopia of ice core results. *Nature*, **399**, 412–413.

Stine S. (1990) Late Holocene fluctuations of Mono Lake, eastern California. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **78**, 333–381.

Stine S. (1994) Extreme and persistent drought in California and Patagonia during medieval time. *Nature*, **369**, 546–549.

Stockton C.W. and Meko D.M. (1975) A long-term history of drought occurrence in western United States inferred from tree rings. *Weatherwise*, **28**, 244–249.

Stute M., Clark J.F., Schlosser P., Broecker W.S. and Bonani G. (1995b) A high altitude continental palaeotemperature record derived from noble gases dissolved in groundwater from the San Juan Basin, New Mexico. *Quaternary Research*, **43**, 209–220.

Stute M. and Deák J. (1989) Environmental isotope study (^{14}C, ^{13}C, ^{18}O, D, noble gases) on deep groundwater circulation systems in Hungary with reference to palaeoclimate. *Radiocarbon*, **31**, 902–918.

Stute M., Forster M., Frischkorn H., Serejo A., Clark J.F., Schlosser P., Broecker W.S. and Bonani G. (1995a) Cooling of tropical Brazil (5 °C) during the last glacial maximum. *Science*, **269**, 379–383.

Stute M. and Schlosser P. (2000) Atmospheric noble gases. In *Environmental Tracers in Subsurface Hydrology*, Cook P. and Herczeg A.L. (Eds.), Kluwer Academic Publishers: Boston, pp. 349–377.

Stute M., Schlosser P., Clark J.F. and Broecker W.S. (1992) Palaeotemperatures in the southwestern United States derived from noble gas measurements in groundwater. *Science*, **256**, 1000–1003.

Stute M. and Talma S. (1998) Glacial temperatures and moisture transport regimes reconstructed from noble gases and O-18, Stampriet aquifer, Namibia. *Isotope Techniques in Studying Past and Current Environmental Changes in the Hydrosphere and the Atmosphere*, IAEA: Vienna, pp. 307–318.

Torgersen T., Habermehl M.A., Phillips F.M., Elmore D., Kubik P., Jones B.G., Hemmick T. and Gove H.E. (1991) Chlorine-36 dating of very old groundwater: III. Further

studies in the Great Artesian Basin, Australia. *Water Resources Research*, **29**, 1875–1877.

Wagner J.D., Cole J.E., Beck J.W., Patchett P.J. and Peachey W.D. (2002) Record of abrupt deglaciation in the arid southwest United States from speleothem deposits. *EOS Transactions American Geophysical Union*, **83**, Fall meeting supplement, Abstract PP62A-0335.

Weyhenmeyer C.E., Burns S.J., Waber H.N., Aeschbach-Hertig W., Kipfer R., Loosli H.H. and Matter A. (2000) Cool glacial temperatures and changes in moisture sources recorded in Oman groundwaters. *Science*, **287**, 842–845.

Winograd I.J., Coplen T.B., Landwehr J.M., Riggs A.C., Ludwig K.R., Szabo B.J., Kolesar P.T. and Revesz K.M. (1992) Continuous 500 000-year climate record from vein calcite in Devils Hole, Nevada. *Science*, **258**, 255–260.

Winograd I.J. (2002) The California Current, Devils Hole, and Pleistocene climate. *Science*, **296**, 7.

Woodhouse C.A. and Overpeck J.T. (1998) 2000 years of drought variability in the central United States. *Bulletin of the American Meteorological Society*, **79**, 2693–2714.

Zhou J. and Lau K.M. (1998) Does a monsoon climate exist over South America? *Journal of Climate*, **11**, 1020–1040.

Zuber A., Weise S.M., Osenbrück K., Pajnowska H. and Grabczak J. (2000) Age and recharge pattern of water in the Oligocene of the Mazovian basin (Poland) as indicated by environmental tracers. *Journal of Hydrology*, **233**, 174–188.

200: Changes in Regional Hydroclimatology and Water Resources on Seasonal to Interannual and Decade-to-Century Timescales

DAVID MEKO

Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ, US

Precipitation, runoff, snowpack, and other elements of the hydrologic cycle respond to climate change on many different timescales. Responses are typically regional rather than local and are linked to large-scale features of the general circulation of the atmosphere. Changes in regional hydroclimatology are important to water resources in most parts of the world, but especially in arid and semiarid regions that are often plagued by chronic shortfalls of water. Instrumental records of precipitation and streamflow indicate that changes on decadal-and-longer timescales have been large enough to be of practical importance, and paleoclimatic records suggest that even larger changes have occurred over the past few centuries. Tree-ring records are perhaps the most useful of the paleoclimatic resources for extracting hydroclimatic estimates with annual resolution. Paleolimnological data has also figured prominently in pointing out possible large changes in regional hydroclimatology prior to the start of the tree-ring record. In North America, multiple types of proxy records suggest low-frequency variance in runoff may have been amplified about 500 years ago, with associated multidecadal periods of drought and wetness.

INTRODUCTION

Change is inherent to the hydroclimatology of a water basin or region. From the standpoint of water resources, the record of past changes in regional hydroclimatology is important in providing information on the natural range of variability of water supply. This information can be useful in probabilistic estimates of likely future stresses on water resources, as well as in understanding the dynamics of climate change and interactions of climate variation and hydrology. Changes in regional hydroclimatology can be tracked with instrumental data back to the late 1800s in some places. Paleoclimatic data can extend this record to several more centuries with annual resolution and thousands of years with reduced resolution and spatial coverage.

Changes in regional hydroclimatology are ultimately of interest because of their potential impact on man and his environment. Paleoclimatic studies point to the decline of entire civilizations at least partly in response to droughts of prehistoric times. In recent decades, large fluctuations in precipitation and river runoff have been associated with stresses on water resources systems. As the gap between water demand and available supply narrows in many parts of the world, it is becoming increasingly important to understand the full range of regional hydroclimatic variability and to identify its causes.

Variables

Hydroclimatic change can be quite inclusive in encompassing change in virtually any element of the hydrologic cycle (*see* **Chapter 2, The Hydrologic Cycles and Global Circulation, Volume 1**). Examples are precipitation, temperature, runoff, evapotranspiration, and recharge. A change might affect amounts (e.g. total annual runoff), as well as rates of processes (e.g. intensity of precipitation). Frequencies and return intervals of events, such as floods and droughts are also of interest.

Definitions of Change

Change in the most general sense is any form of deviation from constancy. Year-to-year variations in runoff, for

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

example, would be classified as "change" by this definition regardless of the size of the variations. In contrast are the more stringent statistical definitions of change used in climatology. For example, climate change has been defined as "a statistically significant variation in either the mean state of the climate or in its variability, persisting for an extended period (typically decades or longer)" (IPCC, 2001). By the strict definition, a variation in runoff would need to be evaluated in the context of the statistical distribution properties of runoff before being classified as a change. But a change may well be practically significant as far as its effect on the water resources of the region without being statistically significant (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1** and **Chapter 7, Methods of Analyzing Variability, Volume 1**).

Timescales

The timescale of a hydroclimatic change refers to the interval of time over which the change occurs. Changes on interannual to century timescales are most relevant to water resources planning and operations. It is well known, however, that the Earth's climate has also varied significantly on timescales of thousands to millions of years (*see* **Chapter 34, Climate Change – Past, Present and Future, Volume 1**). The most recent past is generally considered most relevant to water resources planning. Certainly the period since the end of the last ice age, some 10 000 to 15 000 years ago, is more relevant than earlier epochs, when boundary conditions of regional hydroclimatology in many places were much different than today. The highest-quality paleoclimatic data are restricted to about the last 1000 to 2000 years, and spatial coverage and time resolution increase greatly over even that short interval.

In general, the time series for studying hydroclimatic change must be much longer than the timescale of interest. It follows that many years of data are needed to study interannual changes, many decades of data to study decadal changes, and so forth. Regionally extensive instrumental records of precipitation, streamflow, and other variables are generally on the order of a few decades to a century in length, and so are unsuitable for addressing the longer timescales of hydroclimatic change. Tree-ring records can extend the record of moisture variations back to a few centuries, and other paleoclimatic data (e.g. trace elements and stable isotopes in dated lake sediments) can extend the record to a few millennia in scattered locations (*see* **Chapter 199, Role and Importance of Paleohydrology in the Study of Climate Change and Variability, Volume 5**).

Spatial Scale

The spatial scale of changes in regional hydroclimatology is governed by features of the general circulation of the atmosphere. Because these features are dynamically interrelated over great distances, hydroclimatic changes tend to be spatially coherent over hundreds to thousands of kilometers, and changes in widely separate areas tend to be statistically correlated. Anomalous positioning of ridges and troughs in the atmospheric circulation is generally associated with widespread drought or wetness. For example, an amplified upper level ridge over the west coast of the United States in January 2003 was associated with exceptionally dry conditions across the southern states and a contrast of warmth in the west with cool conditions in the east (Figure 1).

The overriding importance of atmospheric circulation patterns to spatial features of runoff variation must be considered in any attempt to define hydroclimatic "regions". Boundaries of regions of similar hydroclimate generally do not correspond to watershed boundaries (Lins, 1997). There may be an overlap, however, when watershed divides also happen to be major impediments to atmospheric flow or to affect the anchoring of troughs and ridges in the general circulation of the atmosphere.

Spatial scale of hydroclimatic changes is important to water resources (*see* **Chapter 3, Hydrologic Concepts of Variability and Scale, Volume 1**). For large basins, size can act as a buffer against the effects of drought, as a drought in one part of the basin might coincide with a wet period in another part. Paleoclimatic data emphasize the importance of sample period in the evaluation of spatial scales of hydroclimatic change. For example, tree rings identify a North American "megadrought" in the 1500s to much greater extent than any drought in the period covered by precipitation and streamflow records (Stahle *et al.*, 2000).

DATA

Instrumental

Precipitation and Temperature

The most regionally extensive direct evidence for hydroclimatic change comes from continuous records of precipitation and temperature. These two variables are particularly useful because of the existence of long instrumental records. Precipitation is more directly relevant to hydroclimatology, but temperature records are also useful as temperature is correlated with potential evapotranspiration and plays a role in processes that affect the timing of runoff (*see* **Chapter 41, Evaporation Modeling: Potential, Volume 1** and **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**).

Instruments that could reliably measure precipitation and temperature were developed by the late seventeenth and the early eighteenth centuries (*see* **Chapter 35, Rainfall Measurement: Gauges, Volume 1**). But few continuous instrumental records exist from the late seventeenth and eighteenth centuries, and in many regions

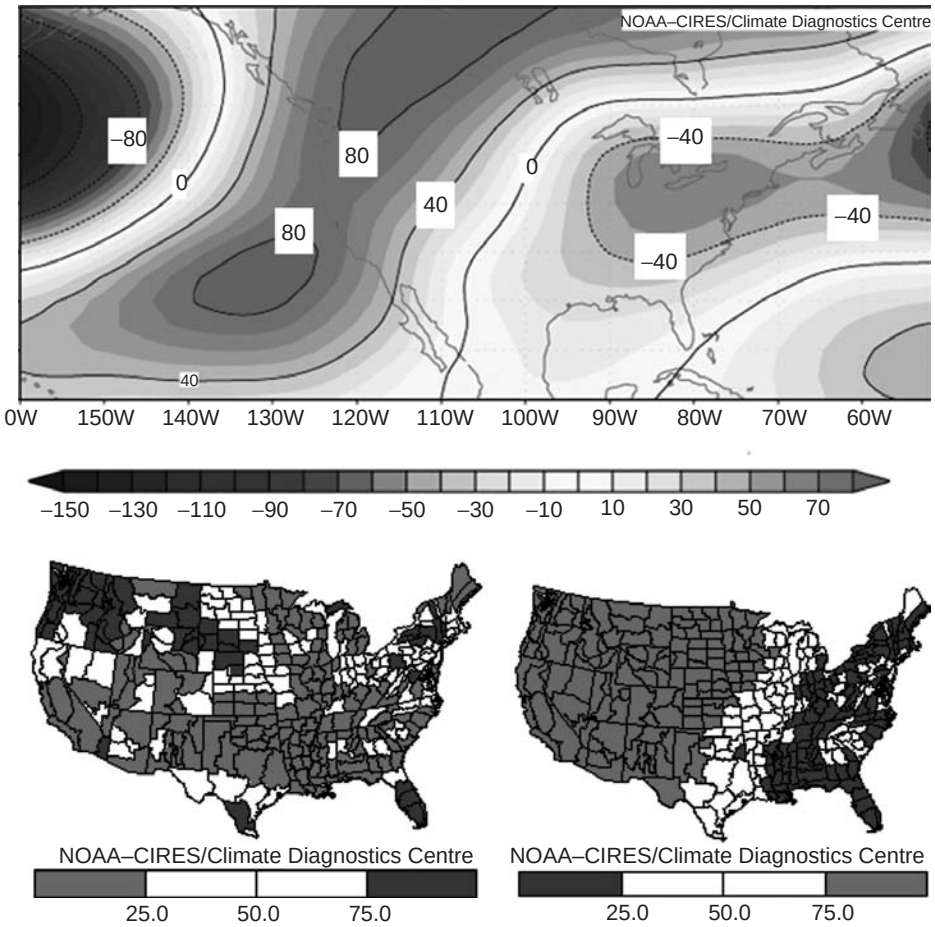


Figure 1 Anomalous climatic patterns in January 2003, an extreme drought month in the southwestern United States. (a): 500 mb geopotential height anomaly (m); (b) left: precipitation percentile by climate division, bottom right: temperature percentile by climatic division. Percentiles computed on the period 1895 to 1999. Images provided by the NOAA-CIRES Climate Diagnostics Center, Boulder Colorado from their web site at **`http://www.cdc.noaa.gov/`**. Data from Climate Division Dataset (NCDC, 1994). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

records exist for only part of the twentieth century. Multiple stations are needed for regional assessment, especially in mountainous regions and other places where climatic gradients are sharp. Various networks of station data have been developed for use in climatology (see Section "Sources of data"). One example is the Global Historical Climatology Network (GHCN), with more than 6000 long-term monthly precipitation and temperature stations specifically screened and processed for the monitoring and detection of climatic change (Vose *et al.*, 1992). The earliest series in the GHCN starts in 1697, but coverage does not become sufficiently dense for regional hydroclimatic studies until the late 1800s or early 1900s. To facilitate analyses at regional and larger scales, quality-controlled gridded precipitation networks have also been developed.

In addition to precipitation and temperature, hydroclimatic studies have made wide use of drought indices. A drought index is a mathematical index that distills the information from more basic measurements (e.g. precipitation and temperature) to summarize the severity of drought. Two commonly used drought indices are the standardized precipitation index, or SPI that is based entirely on precipitation (Guttman, 1999), and the Palmer Drought Severity Index, or PDSI that combines the effects of precipitation and temperature through a simple water-balance accounting model (Palmer, 1965).

Streamflow

Streamflow records are measurements of the discharge, or flow, of rivers and streams and are usually derived from a stage-discharge curve that relates the water level in the river to the volume rate of flow past the gage. Streamflow records worldwide are considerably less numerous than precipitation records, but have the advantage for regional hydroclimatology of integrating moisture conditions over space on

a watershed-scale, and of directly providing an approximation to the residual of precipitation and evapotranspiration. Gaged records of stream discharge are sometimes flawed as indicators of hydroclimatic change because of various factors, including watershed changes, irrigation losses, instability in the stage–discharge relationship, and changes in evaporative losses and timing of flows due to reservoir storage (*see* **Chapter 191, Environmental Flows: Managing Hydrological Environments, Volume 5**). A network of minimally distorted streamflow records with coverage to the late 1800s has been assembled for hydroclimatic studies in the United States (Slack *et al.*, 1988).

Where high-quality stream gage data is unavailable, an alternative is gaged streamflow that has been adjusted for removal of known effects of man by adding back in recorded or estimated diversions of water, and so on. These so-called "natural-flow", "unimpaired-flow", or "virgin-flow" records are sometimes available from water management agencies and can be quite useful in hydroclimatic analysis (e.g. Hidalgo and Dracup, 2003).

Lake Levels

The water level of lakes has long been recognized to respond to changes in climate. Factors affecting the sensitivity of lake-level to climatic fluctuations include the size of the drainage area, the presence or absence of an outlet, the water-balance type, the basin relief, and the extent of human disturbance. The most sensitive lakes for recording hydroclimatic changes are generally those lacking an outlet. Many instrumental records of lake-level change extend back to the nineteenth or early twentieth centuries. For example, the record for the Dead Sea starts in 1800; the Caspian Sea in 1851; and the Great Salt Lake, US, in 1880. Globally, many closed lakes relatively free from human disturbance show broadly parallel patterns of behavior, with high levels in the latter part of the nineteenth century, low levels in the 1920s to 1940s, followed by higher levels to at least the early 1980s (Street-Perrott and Harrison, 1985).

Paleoclimatic Data

Tree Rings

Indices of the ring widths of moisture-sensitive trees are the most widely used type of paleoclimatic data for studying changes in hydroclimatology on interannual-to-decadal timescales. The period of analysis from living trees is limited only by the maximum tree age that typically exceeds 300 years for useful species. The period can be extended much further – to thousands of years – using deadwood material from snags, buried logs, and other sources. Tree rings can be dated exactly to the year, enabling an accurate resolution of interannual hydroclimatic variations. The strength of hydrologic signal depends on the sensitivity of growth of the particular tree species or specimen to the parameter of interest. Tree rings generally provide opportunistic sampling of important runoff-producing areas in semiarid regions because of the tendency for undisturbed forests to be in the mountains. Moreover, the wide distribution of suitable tree species over the temperate latitudes facilitates regional analysis.

The biological basis of the relationship between tree rings and climate is covered thoroughly by Fritts (1976); aspects relevant to dendrohydrology are covered by Meko *et al.* (1995). Dendrohydrologic reconstructions are statistical estimates of hydrologic variables from networks of tree-ring chronologies. The most commonly reconstructed variables are precipitation, streamflow, and lake levels. The development of statistical reconstruction methods has been an active area of research in dendrohydrology. Multivariate methods, including principal components analysis and canonical correlation analysis were first applied in the reconstruction of lake levels (Stockton and Fritts, 1973) and soon after to streamflow records (Stockton and Jacoby, 1976). The statistical methodology of dendrohydrologic reconstruction as of the early 1990s has been thoroughly reviewed by Loaiciga *et al.* (1993). Some recent extensions are neural networks as an alternative reconstruction model (Woodhouse, 1999), cross validation in model selection (Hidalgo *et al.*, 2000), and "noise-added" reconstructions to facilitate probabilistic interpretation of reconstructed flow series (Meko *et al.*, 2001).

Several statistical issues remain key concerns in the evaluation of tree-ring reconstructions of hydrologic or climatic variables. Measured widths of tree rings generally have a size-related or age-related trend that must be statistically removed, and the approach to removal of this growth trend determines the lowest frequency of climatic signal that can be identified. A tree-ring chronology is an average over many trees at a site, and the sample-depth, or number of trees, generally decreases back in time in accordance with the age distribution of available trees. This sample-size heterogeneity necessitates safeguards against loss of signal and possible inflation of variance due to the changing sample size. The problem becomes more complicated when chronologies are formed by merging living-tree samples from sites with known hydrologic settings with archaeological samples whose original setting is unknown. Statistical issues in application of tree-ring data in hydrologic reconstruction are discussed further by Cook and Kairiukstis (1990).

Although the tree-ring variable of choice for most dendrohydrologic reconstructions has been the annual index of precisely dated, measured ring widths, other tree-ring indicators have occasionally been used. Stable isotopes of hydrogen and carbon in dated tree rings, for example, have shown some promise in hydroclimatology (Leavitt and Wright, 2002). In a radically different type of application, tree-ring counts alone, without precise dating and width

measurement, can yield useful hydrologic information. For example, ring counts of radiocarbon-dated submerged stumps in Mono Lake, California, delineated two epic droughts – periods in which trees were able to take root and grow for more than a century in what is now a submerged lake (Stine, 1994).

Lake Cores

Just as the measured stages of closed basin lakes are a record of modern-day changes in aridity, the past stages of such lakes inferred from cores of lake sediments are a record of long-term changes in aridity (*see* **Chapter 110, Paleolimnology and Paleohydrology, Volume 3**). The climatic record obtainable from lake cores is potentially much longer than that provided by tree rings, and can cover the entire time period the lake existed within a closed basin. Many lake-core records extend well before the start of the Holocene, although dating can be problematic beyond about 20 000 to 30 000 years (Street-Perrott and Harrison, 1985). Biological and chemical properties imprinted by animals and plants that lived in the lakewater are one source of information in lake cores. Ostracodes are very small animals, rarely large enough to be seen by the naked eye, that have been present on Earth since the Ordovician Period. They occur today in great numbers in lakes, rivers, and seas, and their shells are preserved in sediments. Diatoms are unicellular algae whose cell walls are made of silica, and whose fossils are often well preserved in lake and marine sediments. Particular species of ostracodes and diatoms have restricted ranges tolerance to salinity and water depth. Changes in species composition can be interpreted as a signal for changes in salinity and water depth that in turn are related to changes in lake volume and aridity. The Mg/Ca ratio in ostracode shells has been found to be an especially robust indicator for changes in lake salinity (Yu *et al.*, 2002).

Interpretation of the Mg/Ca ratio and other measured quantities from the core as time series requires dating of the cores, and there is often considerable uncertainty in dates. Radiocarbon dating and magnetostratigrapy have been used with considerable success in lake-level studies (Street-Perrott and Harrison, 1985). The resolution of the hydroclimatic signal is limited by the sedimentation rate and the accuracy of the dating technique. Even for lakes with high accumulation rates, however, dating by ^{14}C is accurate to only about ± 50 years.

If sediments are laminated (varved), much higher resolution is possible. If varves are unequivocally annual, dating is reduced to counting the annual layers, and the potential exists for millennia-long records of hydroclimatology on interannual to century timescales (Overpeck, 1996). Varved lake sediments are the exception rather than the rule in semiarid environments, however, so that the potential for regional information in some of the more critical water-limited environments is limited.

Other

Other paleoclimatic indicators have met with some success in inferring past hydroclimatic conditions (*see* **Chapter 199, Role and Importance of Paleohydrology in the Study of Climate Change and Variability, Volume 5**). Past lake levels have also been inferred from geomorphic interpretation on the basis of surveying and description of landforms along lake shores (Street-Perrott and Harrison, 1985). Flora composition inferred from pollen and plant parts in radiocarbon-dated packrat middens has been used to infer wet or dry conditions relative to the present (Betancourt *et al.*, 1990). Periods of reactivation of sand dunes – and drought – can be identified by ^{14}C dating of organic matter or bone fragments from buried soils, and accuracy to perhaps ± 30 years is now possible with optically stimulated luminescence and single aliquot regeneration (OSL-SAR) dating of quartz grains (Forman and Pierson, 2003; Wolfe *et al.*, 2001). Geomorphic interpretation of backwater deposits and other features have been used for paleoflood estimation. Coring studies for hydroclimatic information have not been restricted to lake sediments. Relative abundance of different taxa of fossilized testate amoeba in cores of peat bogs, for example, has been used to infer changes in water-table depth near Lake Michigan (Booth and Jackson, 2003). It should be noted that while cores provide a continuous history of hydrologic variation, other methods often provide just a snapshot of conditions at a particular time.

SOURCES OF DATA

Instrumental and paleoclimatic data for studying hydroclimatic change are readily available online. Precipitation data, including the Global Historical Climatology Network (GHCN) baseline global data set for monitoring climate change, are available from National Oceanic and Atmospheric Administration (NOAA's) National Climatic Data Center (URL: **`http://www.ncdc.noaa.gov/oa/ncdc.html`**).

Online mapping of climatic parameters, including precipitation, temperature, and geopotential height fields, is possible through the NOAA-CIRES Climate Diagnostics Center (URL: **`http://www.cdc.noaa.gov/`**). Indices of sea surface temperature (SST) and atmospheric circulation can be obtained from the National Center for Atmospheric Research (URL: **`http://www.cgd.ucar.edu/cas/catalog/climind/`**) and the University of Washington's Joint Institute for the Study the Atmosphere and Ocean (JISAO) (URL: **`http://tao.atmos.washington.edu/data_sets/`**). The JISAO site is also an excellent source of gridded precipitation datasets. Gaged streamflow time series for the United States can be found at a web site of the US Geological Survey (USGS) (URL: **`http://water.usgs.gov/data.html`**). Included on the USGS site is a hyperlink to the Hydro-climatic Data Network (HCDN) of Slack

et al. (1988). The HCD consists of streamflow records unaffected by artificial diversions, storage, or other works of man; the records span the period 1874 to 1988 and the dataset includes some 1659 sites throughout the United States and its territories. No single online source exists for the so-called "natural-flow" streamflow records, as these are usually created and managed by various agencies. For example, natural-flow data for the Sacramento and San Joaquin River systems, California, can be obtained from the California Department of Water Resources (URL: `http://cdec.water.ca.gov/cgi-progs/iodir/WSIHIST`).

A clearinghouse for paleoclimatic data and reconstructions is NOAA's World Data Center for Paleoclimalogy (URL: `http://www.ngdc.noaa.gov/paleo/`). The "Paleoclimatic Data" link from that site leads to a vast assortment of data sets relevant to hydroclimatology. Among these are tree-ring reconstructions of streamflow and precipitation, lake-level records, and a unique dataset of gridded drought index reconstructed from tree rings for North America (Cook *et al.*, 1999).

PRECIPITATION

Studies too numerous to list here have addressed regional changes in precipitation around the world. Significant correlation is often found between time series of regional precipitation and various indices of atmospheric circulation and SST. Gridded precipitation for western North America indicates that, regionally, a significant part (20–50%) of the variance of annual precipitation is at decadal timescales, and that spatial modes of variability are on the order of 1000 km in size (Cayan *et al.*, 1998). Wet and dry decades are associated with changes in the frequency of atmospheric circulation modes, such as the Pacific-North-America (PNA) pattern, and some low-frequency components appear to be linked to global-scale climate processes. Significant correlations have been reported between regional precipitation in the United States and the El Nino/Southern Oscillation (ENSO) phenomenon, which reflects SST and surface pressure differences across the equatorial Pacific on timescales of a few years (3–7 years). Although significant correlations can be found between regional precipitation and individual circulation or SST indices, stronger relationships emerge with combinations of indices. Apparently, the ENSO signal in precipitation in North America is complicated by the influence from SST anomalies in the North Pacific – as reflected in the Pacific Decadal Oscillation (PDO), which acts on multidecadal timescales (McCabe and Dettinger, 2000). It is not clear that ENSO and the PDO represent independent modes of SST behavior, and indeed some evidence suggests that the PDO is a low-frequency manifestation of ENSO (e.g. Newman *et al.*, 2003). Influence on North America precipitation may also be exerted from North Atlantic SST through the Atlantic Multidecadal Oscillation (AMO), on timescales of 65 to 80 years (Enfield *et al.*, 2001). In Europe and Africa, a north–south atmospheric pressure oscillation known as the *North Atlantic Oscillation* (NAO) has been linked with multidecadal variations in precipitation.

Tree rings have been used for quantitative reconstruction of regional precipitation in the United States since the mid-twentieth century. Early tree-ring studies focused on the semiarid western United States, where most of the developed tree-ring chronologies were located. Recent decades have seen the expansion of regional tree-ring studies to other parts of the globe where water is in short supply, such as Jordan and Turkey. A fairly complete inventory of recent regional precipitation reconstructions, with data and references, is available at the NOAA World Data Center for Paleoclimatology (see Section "Sources of Data"). Earlier reconstructions are described in Stockton *et al.* (1985) and other sources (see Section "Further Reading"). A recent example serves to illustrate typical methodology and interpretation.

Example from Tree Rings: Southern Manitoba, Canada

Hydroclimatic changes in southern Manitoba, Canada, since 1409 A.D. have been reconstructed from the widths of tree rings by St. George and Nielsen (2002). The region for this study is a 100-km long corridor along the Red River (Figure 2). Tree-ring samples were taken from a variety of sources: living trees (*Quercus macrocarpa*), timbers from historical buildings and Euro-Canadian archaeological sites, and subfossil logs from alluvial sections along river. The assorted tree-ring material gave a time coverage to 1720 A.D. from living trees and to 1286 A.D. from subfossil samples.

Data Treatment and Reconstruction Modeling

To preserve as much low-frequency climate information as possible, an age-related trend was removed using regional curve standardization (RCS) in converting ring width to a dimensionless index of the growth of each tree (Briffa *et al.*, 1992). Indices for different samples at a site were averaged together to form a single tree-ring chronology, and sample replication was judged adequate to allow the chronology to be reliably used for hydroclimatic reconstruction back to 1409 a.d.

Exploratory correlation analysis with monthly precipitation series identified the 12-month sum of precipitation ending with July of the growth year as a reasonable hydrologic variable for reconstruction. A regression model ($R^2 = 0.43$) relating the annual precipitation to the tree-ring chronology was calibrated over the period 1896 to 1996 and validated using cross validation (Michaelsen, 1987).

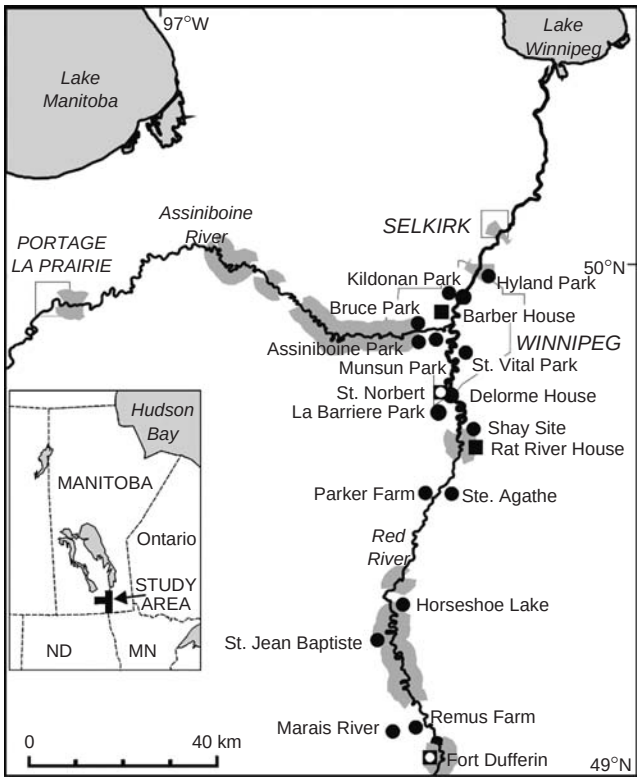


Figure 2 The Assiniboine and Red River in southern Manitoba and the tree-ring sampling network for reconstruction of annual precipitation at Winnipeg, Canada. Tree-ring samples from living-tree sites (circles), historical buildings and archaeological sites (squares), and subfossil logs from alluvium along rivers (shading) (Reprinted from St George & Nielsen 2002. ©2002, with permission from Elsevier)

Hydroclimatic Changes

The reconstructed annual precipitation, 1409 to 1998 A.D., plotted in Figure 3, suggests that hydroclimatic conditions after the establishment of the first Euro-Canadian settlement in Manitoba (1811 A.D.) are not representative of the conditions since 1409 A.D. Precipitation anomalies are more persistent and precipitation is more variable in the first half of the reconstruction than the second half. Moreover, the most extreme precipitation values fall within the first half of the reconstruction, and all years with reconstructed precipitation greater than two standard deviations above the instrumental mean occur before 1800 A.D.

The extended wet period near 1600 A.D. is synchronous with wet epochs elsewhere in the northern Plains (North Dakota, Minnesota) as inferred from limnological studies, and suggests that hydroclimatic shifts in this region can last for several decades and cover several thousand square kilometers. Interestingly, the early part of the wet period overlaps with a well-documented megadrought that apparently affected watersheds across much of northern Mexico and the southwestern United States (Stahle *et al.*, 2000). The results imply that climatic case studies in regional drought and flood planning based exclusively on experience during the twentieth century may dramatically underestimate true worst-case scenarios for southern Manitoba.

RUNOFF

Like precipitation, runoff has been shown to have a large spatial structure and to be significantly correlated with indices of atmospheric circulation and SST. Strong linkages between regional streamflow variations and indices of atmospheric circulation and SST have been demonstrated on global, regional, and basin scales (e.g. Dettinger and Diaz, 2000; Jury, 2003; Eltahir, 1996; Hidalgo and Dracup, 2003).

Much research has addressed circulation and SST linkages to runoff variation in western North America. The ENSO phenomenon is associated with opposite-sign streamflow anomalies in the southwestern and northwestern United States (Cayan and Peterson, 1989), and streamflow in widely separated parts of the United States is significantly related to the extreme modes of southern Oscillation variability (Dracup and Kahya, 1994).

The large spatial scale of runoff variations can be illustrated with natural-flow records for the Colorado River and Sacramento River, two rivers of vital importance to water resources in western United States (Figures 4 and 5). Despite being separated by some 1000 km, these basins are simultaneously in drought more often than expected by chance alone. Perhaps the best example of joint drought is the water-year 1977, when flow reduced to less than 34% of normal on both rivers. Regionally, low streamflow events such as these are typically associated with departures from normal atmospheric circulation over wide parts of the Northern Hemisphere. The winters of 1976 and 1977 were so unusual as to be the focus of a case study of anomalous atmospheric circulation (Namias, 1978). The winter was characterized by amplification of the normal seasonal 500-mb ridge over the western United States, with teleconnections to anomalously strong troughs over the Aleutians and the eastern United States.

Tree rings are frequently used to extend runoff records before the start of the instrumental period. The usual approach is to calibrate tree records with either a gaged-flow or natural-flow record, and substitute earlier tree-ring data into the equation to generate a long-term reconstruction. Western North America has been the geographic focus of many such reconstructions. Findings on contrasts of twentieth century and earlier hydroclimatology from seven of these reconstructions are summarized in Table 1. Note that basins studied are highly variable in size, with mean annual flows varying by a factor of almost 500. A common theme is the relative wetness of the twentieth century, especially the early decades. Identification of a common period of "worst" drought in these records is somewhat problematical because of the uneven time coverage.

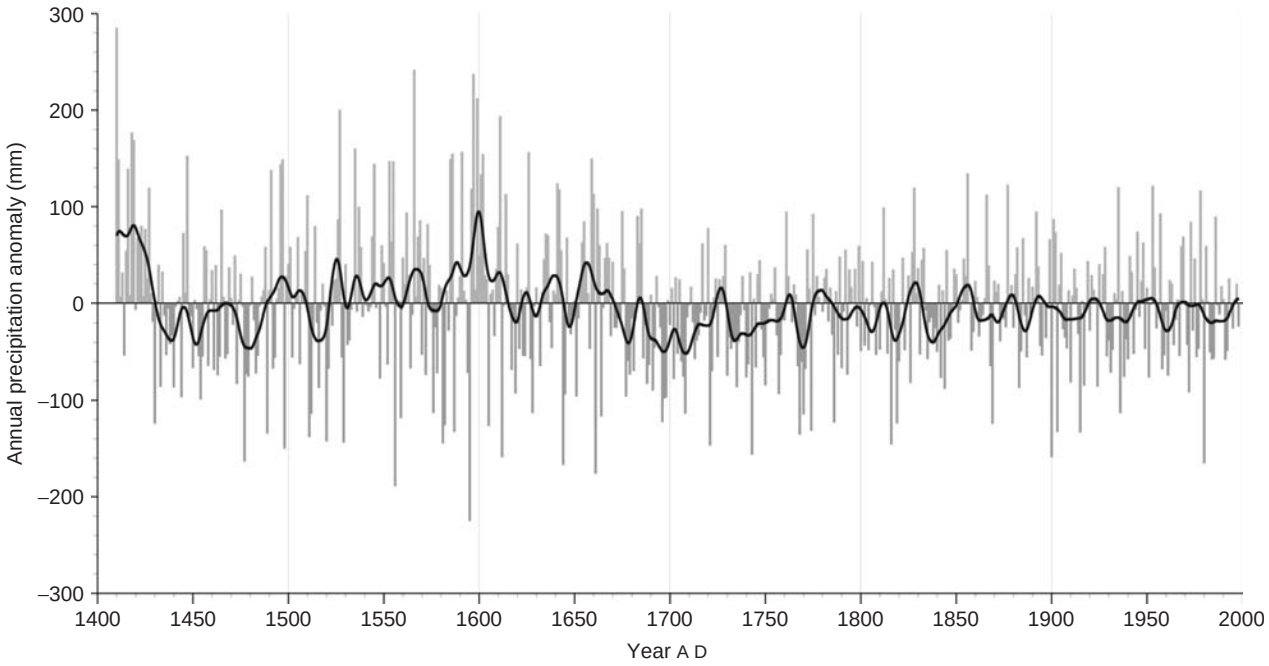


Figure 3 Tree-ring reconstruction of annual (August–July) precipitation, 1409 to 1998 A.D., at Winnipeg, Canada. Precipitation as departures from 1961 to 1990 mean. Smooth curve is 15-year running mean (Reprinted from St George & Nielsen 2002. ©2002, with permission from Elsevier)

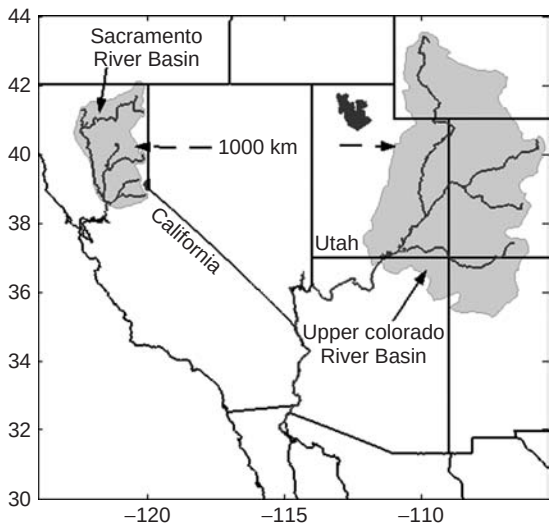


Figure 4 Drainage basins of Upper Colorado and Sacramento Rivers. A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

Example from Tree Rings: Colorado River Basin, US

Tree-ring records have been especially useful for pointing out important components of long-term variability of runoff in the Upper Colorado River basin. The annual flow for the Colorado River at Lees Ferry, Arizona, has been reconstructed back by Stockton and Jacoby (1976) from a network of some 17 tree-ring sites strategically located relative to major runoff-producing areas. The reconstruction, generated with a regression model ($R^2 = 0.87$) that used principal components of tree-ring chronologies as predictors, covers the years 1520 to 1961 and shows great variability of runoff on decadal timescales (Figure 6). The 20-year running mean reaches an all-time high in the early 1900s, which happens to coincide with the period used as a baseline to allocate the waters of the river. The smoothed series illustrates the importance of considering the long-term record in assessing range of severity of extended droughts: flows in the late 1500s average just 80% of the long-term mean reconstructed flows. The tree-ring record also suggests a bias in the observed long-term mean flow, with the mean from the tree-ring record 1520 to 1961 just 90% of the observed mean for 1906 to 2001.

It is important to keep in mind that tree-ring reconstructions are estimates rather than measurements of streamflow, and that the accuracy of the estimates depends on many factors, including the basic quality of the tree-ring data, the method of processing that data, and the statistical reconstruction model. An alternative reconstruction for the Colorado River at Lees Ferry, Arizona, using a slightly different method of data reduction than Stockton and Jacoby (1976) shows even more severe drought conditions in the late 1500s than indicated in Figure 6 (Hidalgo *et al.*, 2000). An assessment of tree-ring records by Stahle *et al.* (2000) suggests that the low flows of the 1500s on the

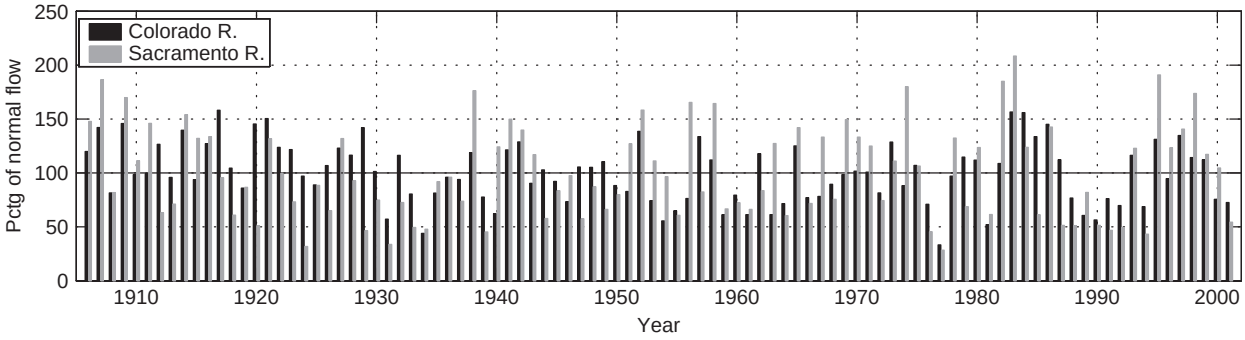


Figure 5 Annual (water-year) totals of natural flow of the Sacramento River, California, and the Colorado River at Lees Ferry, Arizona. Means are for period 1906 to 2001. Colorado River data from US Bureau of Reclamation; Sacramento River data from California Department of Water Resources (see Data Sources)

Table 1 Hydroclimatic changes inferred from tree-ring reconstructions of streamflow in western North America

River /reference[a]	Mean flow[b] ($m^3\,sec^{-1}$)	Reconstruction Period[c]	Hydroclimatic change[d]
Colorado R, AZ	590.22	1520 to 1961	Longest period of high-flow years 1907 to 1930
Stockton and Jacoby (1976)			Drought in late 1500s longer and more severe than any in the twentieth century
Salt R, AZ	24.50	1580 to 1979	Frequency of high-flow relatively high in the twentieth century
Smith and Stockton (1981)			Most severe sustained periods of low flow before the twentieth century
Gila R, AZ	13.08	1663 to 1985	Clustering of high-flow years in the early twentieth century
Meko and Graybill (1995)			Drought in 1950s most severe on record
Middle Boulder Ck, CO	1.55	1703 to 1987	Low flows less persistent in the twentieth century than in the nineteenth century
Woodhouse (2001)			Multiyear drought in the 1840s more severe than any in the twentieth century
Sacramento R, CA	707.49	869 to 1977	Lowest average (10–50 yr) means near 1300 A.D.
Meko *et al.* (2001)			Extreme single-year low flows in late 1500s
Yellowstone R, WY	85.58	1706 to 1977	Relatively high mean flows in the twentieth century (except 1930s)
Graumlich *et al.* (2003)			Drought of 1930s more severe than any reconstructed drought
South Saskatchewan R, Canada;	302.81	1470 to 1992	Relatively high mean flows in the twentieth century
Case and MacDonald (2003)			Droughts more severe in the 1840s than in the twentieth century

[a]river for which flow reconstructed, followed by reference to study.
[b]mean daily observed flow for period used to calibrate reconstruction (all greater than 50 years).
[c]time period of reconstruction (A.D.).
[d]selected outstanding findings, with emphasis on comparison with the twentieth century.

Colorado River are just one component of a much larger-scale phenomenon.

Example from Tree Rings: White River Basin, Arkansas, Southeastern United States

In this second example, summer (JJA) runoff of the White River, Arkansas, was reconstructed from tree rings of bald cypress (*Taxodium distichum*) by Cleaveland (2000). In contrast to the Colorado River basin, the White River basin is less arid, has less extreme climate gradients, and a seasonal flow regime not dominated by mountain snowmelt. The White River, a major tributary of the Mississippi River, drains some 66 000 km^2 in Arkansas and Missouri (Figure 7).

Bald cypress is an extremely long-lived species, and the three tree-ring chronologies for the study extend to earlier than 1200 A.D. The gaged-flow record for the White River at Clarendon begins in 1900, and has a cool-season dominant regime with lowest monthly mean flows

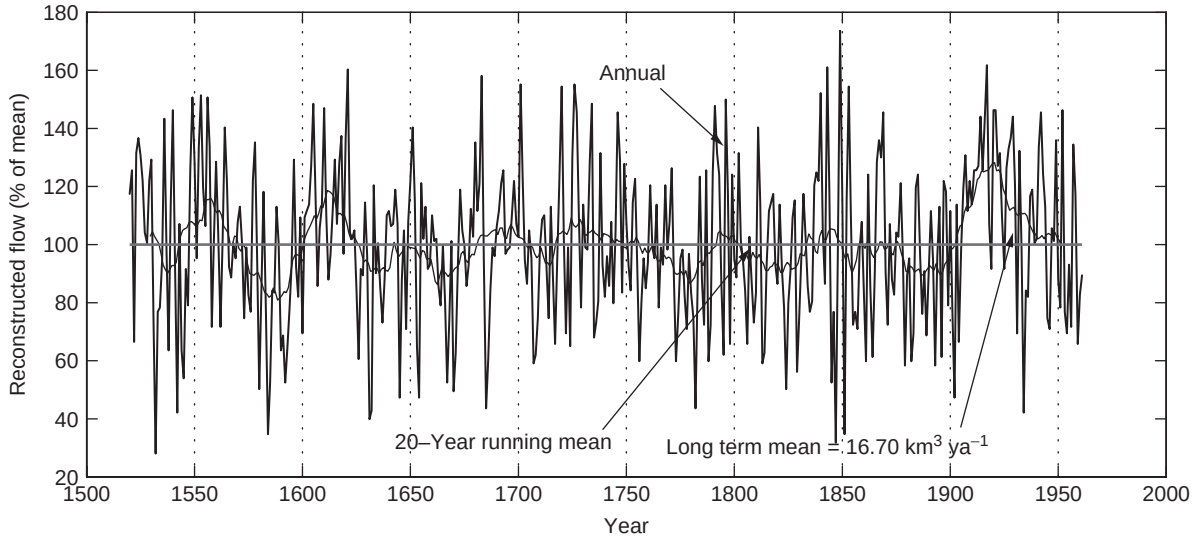


Figure 6 Tree-ring extension of annual flow of Colorado River at Lees Ferry (data from Stockton and Jacoby, 1976). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

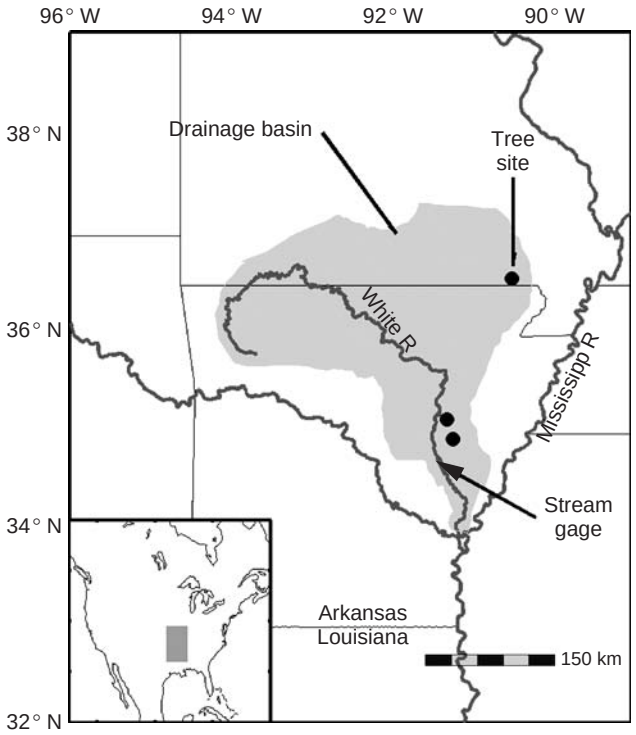


Figure 7 Map showing location of drainage basin, stream gage, and tree-ring sites for reconstruction of White River, Arkansas, summer flow (Reproduced from Cleaveland (2000) by permission of Hodder Arnold). A color version of this image is available at http://www.mrw.interscience.wiley.com/ehs

in the summer months. Bald cypress from this region are responsive mainly to growing-season rainfall, however, and it is this summer component of flow that was reconstructed. The reconstruction was generated with a linear regression model ($R^2 = 0.67$) calibrated during 1931 to 1985 and validated during 1900 to 1930.

Hydroclimatic Changes

The reconstruction of summer (JJA) mean daily flow, 1023 to 1985 A.D., contains extended dry periods and wet periods that exceed anything in the modern record (Figure 8). One focus of the interpretation was on the tendency for persistence of flows above and below thresholds of wetness and drought defined by percentiles of the long-term reconstruction. A difference in persistence regimes was identified: low flows were persistent from year-to-year and high flows were random. An example of the persistence of low flows is the sequence of 10 consecutive years below the median from 1449 to 1458 A.D. In contrast, the modern gaged record contains no examples of such recurrently low summer flows.

Mean and variance are basic properties tied to the concept of stationarity of hydrologic time series (Salas *et al.*, 1980). A *t*-test of the reconstructed flows plotted in Figure 8 revealed no significant changes in mean from one century to the next. A nonparametric test (Conover, 1980) did, however, indicate significant changes in variance. The twentieth century had anomalously high variance, with a significant increase from the nineteenth century.

In discussing the differences between flow characteristics in the twentieth century and earlier centuries, Cleaveland (2000) leaves open the possibility that human modification of the watershed may have had some impact on the results of the study. For example, the loss of wetland forests would tend to reduce infiltration and speed up

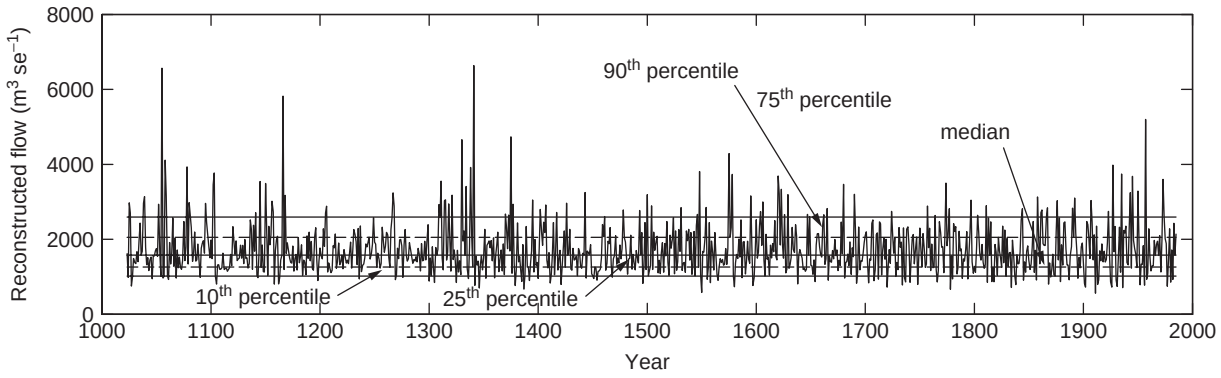


Figure 8 Total summer (JJA) mean daily streamflow of White River at Clarendon, Arkansas, reconstructed from tree rings (Reproduced from Cleaveland (2000) by permission of Hodder Arnold)

overland flow of runoff, possibly making low flows more intense by reducing base flow (*see* **Chapter 111, Rainfall Excess Overland Flow, Volume 3**; **Chapter 187, Land Use Impacts on Water Resources – Science, Social and Political Factors, Volume 5**).

ARIDITY

Annual resolved, calibrated, and validated reconstructions, as generated from tree rings, are useful for evaluating long-term statistical properties of hydrologic variables, but more qualitative information on relative wetness and dryness can also be useful for placing the instrumental record in context. Aridity information derived from lake cores is particularly relevant in assessing multidecadal change (*see* **Chapter 110, Paleolimnology and Paleohydrology, Volume 3**). Lake-core data is also somewhat complementary to tree rings in that, longer-term and possibly lower-frequency hydrologic information can be extracted from the lake cores. Lake-core time series of various quantities, some related to aridity, are available from the World Data Center for Paleoclimatology (see Section "Sources of data").

Example from Paleolimnology: Northern Great Plains, US

Multi-proxy paleolimnological data were used to infer long-term changes in aridity in the northern Great Plains, US (Yu *et al.*, 2002). The paleoclimatic samples were salinity/drought proxies in sediment cores from four lakes in North Dakota and Minnesota (Figure 9). The chronology for the core was established by Accelerator Mass Spectrometer (AMS) ^{14}C dating at a series of depths in the core, and then interpolating between those depths with a straight line fit to age against depth.

Data Treatment

The core from Rice Lake (Figure 9) was the central focus of the study. The top 5.5 m of that core, covering the last 2100 years, was sampled at decadal time intervals for ostracodes. Carbonate shell fragments were subjected to isotope analysis for the computation of $^{18}O/^{16}O$ and $^{13}C/^{12}C$ ratios, and the acid residual from the isotope analysis was analyzed for various trace elements, including Mg and Ca. The hydroclimatic interpretation was centered on time variations in the ostracode-shell Mg/Ca ratio, which indirectly is a proxy for aridity (see Section "Data").

Smoothed series of the Mg/Ca ratio from the four lakes were averaged together after a resampling of the original

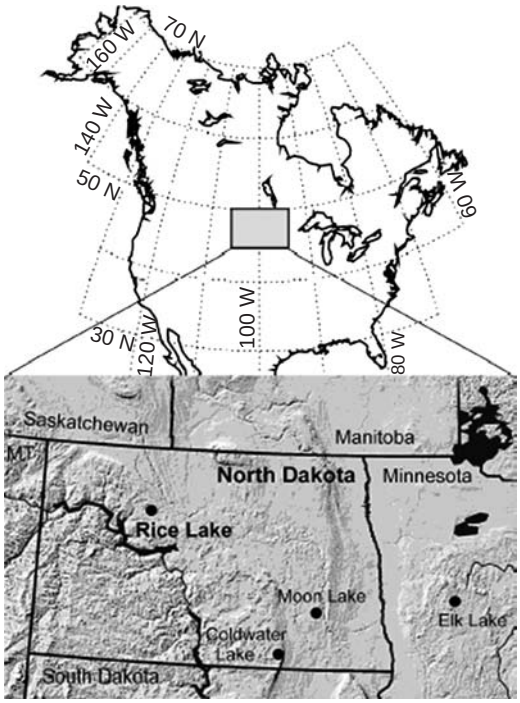


Figure 9 Lakes yielding a 2100-year proxy record of wet and dry conditions in the northern Great Plains from sediment cores (Reproduced from Yu *et al.*, (2002) by permission of Hodder Arnold)

time series at even 10-year intervals to arrive at the regional aridity index. The resolution in the resulting time series is too crude to track interannual fluctuations in aridity, but is suitable for identifying multidecadal fluctuations. The inferred drought signal is qualitative rather than quantitative, as the Mg/Ca ratio is not calibrated statistically with any particular instrumental time series of aridity.

Hydroclimatic Changes

The composite time series of lake aridity index has several upward excursions interpreted as hydrologic droughts (Figure 10). Key features are dry periods lasting some 400 years each, separated much shorter wet intervals. The earliest dry period, the "medieval climate anomaly" (MCA) has two aridity peaks that happen to correspond to multidecadal droughts in the southern Sierra Nevada of California (Stine, 1994, 1998). Another notable feature is the peak in dryness about 300 years ago, in a period that is generally considered the Little Ice Age (LIA). The results imply that much of the LIA was relatively arid in the northern Great Plains and that rhythms of aridity may have occurred in this region on a longer timescale than typically amenable to study with tree rings.

SNOWPACK

Snow accumulation is an interesting climatic index as it responds to both precipitation and temperature variations (*see* **Chapter 159, Snow Cover, Volume 4**; **Chapter 160, Energy Balance and Thermophysical Processes in Snowpacks, Volume 4**). Regional hydroclimatic variability as reflected by changes in snow accumulation in the mountainous western United States has been studied by Cayan (1996), who used the snow water equivalent (SWE) on April 1 as a summary variable. Snow course data extending back to the 1920s indicated considerable year-to-year variability in SWE, with some years having anomalously high or low accumulation across multiple basins. A notable shift in the time series occurred in 1977, which coincides with a widely reported regime change in the climate of the North Pacific sector (Ebbesmeyer *et al.*, 1991). One characteristic of the shift is decreased snow accumulation in the northern watersheds of the western United States and increased accumulation in the southern watersheds. Cayan (1996) links the observed snowpack changes to stronger North Pacific lows in the winters and increased occurrences of the warm phase of ENSO.

Another aspect of snowpack important to regional water resources is the timing of snowmelt (*see* **Chapter 114, Snowmelt Runoff Generation, Volume 3**). Changes in timing can affect the operation of reservoirs systems, which frequently must respond to competing interests of water storage and flood control. An interesting study combining phenological and hydrologic data suggests that a change toward earlier snowmelt has occurred in the western United States since the late 1970s, and that the change resulted from a 1–3 °C increase in spring temperatures over western North America (Cayan *et al.*, 2001).

Snowpack data are being increasingly used in dendrohydrology as longer, quality-controlled SWE datasets are becoming available. Snowpack for the Gunnison Basin in western Colorado has been reconstructed from tree rings back to 1569 A.D. (Woodhouse, 2003). years of persistent low April-1 SWE are not distributed evenly through the record, and the twentieth century is notable for reduced frequency of persistently low SWE events compared with the long-term record.

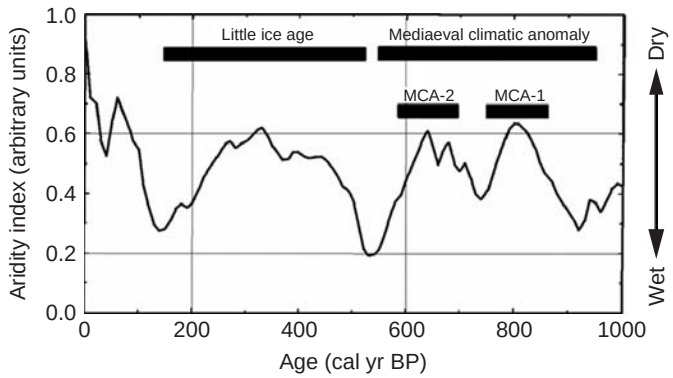


Figure 10 Aridity index for the Northern Great Plains of North America from salinity/drought profiles Rice, Moon, Coldwater, and Elk Lakes. The index shows a double-peak pattern during the Mediaeval Climate Anomaly (MCA; Stine, 1998). Note that time increases toward left on the time axis, with years before present (BP) representing years before 1950 A.D. (Reproduced from Yu *et al.*, (2002) by permission of Hodder Arnold)

FLOOD FREQUENCY

Changes in the amount and timing of runoff also affect the frequency of floods (*see* **Chapter 125, Rainfall-runoff Modeling for Flood Frequency Estimation, Volume 3**). For many years the subject of climatic change was not considered relevant to flood-frequency analysis, but that viewpoint has changed. Subtle shifts in the locations of the low and high pressure anomalies in the North Pacific have been shown to affect the occurrence of large winter floods in the southwestern United States (Ely *et al.*, 1994). The probability distribution of partial duration flood series, moreover, can differ from one climatic regime to another (Hirschboeck, 1987). Paleoclimatology has been useful in extending information on flood history beyond the instrumental record (e.g. Tardif and Bergeron, 1997; Enzel *et al.*, 1993). Other paleoclimatic flood evidence, as well as the general topic of climate variability and frequency of floods at decadal to millennial timescales has been reviewed Redmond *et al.* (2002).

CAUSES

The root causes of changes in regional hydroclimatology on various timescales are still largely unknown. The search for "causes" proceeds to a deeper level than mere associations (e.g. correlations between precipitation and SST indices), and often centers on possible "forcing factors" of change. One view is that no forcing factors need be invoked, and that variations on interannual-to-century timescales naturally evolve from the inherent random variability of the atmosphere or from inherent variability in the ocean dynamical system. Hypotheses for outside forcing include solar variability and variability in volcanic aerosol loading of the stratosphere. Paleoclimatic data from North America hint at a solar-variability link on large spatial scales (Cook *et al.*, 1997; Yu *et al.*, 2002), although statistical evidence is weak and demonstrated mechanism is lacking. It is of course possible that some combination of the above factors (and possibly others) has influenced the hydroclimatic variations evident in the instrumental and paleoclimatic record (Goodess *et al.*, 1992).

The possible anthropogenic influence on recent and future hydroclimatic changes is a topic of much recent research. Coincident with the increase in atmospheric greenhouse gases and aerosol loading of the troposphere since the preindustrial period are a number of hydroclimatic changes. With varying degrees of certainty, it can be said that changes in the twentieth century include (i) an increase (5–10%) in continental precipitation for much of the Northern Hemisphere, (ii) an increase in heavy precipitation events at high latitudes, and (iii) increased summer drying and incidence of drought in some areas (IPCC, 2001) (*see* **Chapter 195, Acceleration of the Global Hydrologic Cycle, Volume 5** and **Chapter 33, Human Impacts on Weather and Climate, Volume 1**).

Acknowledgments

Research reported here was partly supported by grant ATM-0080834 from the National Science Foundation.

FURTHER READING

Beniston M. and Innes J.L. (Eds.) (1998) *The Impacts of Climate Variability on Forests*, Springer: New York.

Bradley R.S. and Jones P.D. (Eds.) (1992) *Climate Since A.D. 1500*, Routledge: London.

Cayan D.R., Redmond K.T. and Riddle L.G. (1999) ENSO and hydrologic extremes in the western United States. *Journal of Climate*, **12**, 2881–2893.

Cook E.R., Briffa K.R., Meko D.M., Graybill D.A. and Funkhouser G. (1995) The "segment length curse" in long tree-ring chronology development for paleoclimatic studies. *The Holocene*, **5**(2), 229–237.

Dettinger M.D., Cayan D.R., Diaz H.F. and Meko D.M. (1998) North-south precipitation patterns in western North America on interannual-to-decadal timescales. *Journal of Climate*, **11**, 3095–3111.

Diaz H.F. (Ed.) (2003) *Climate Variability and Change in High Elevation Regions: Past, Present and Future*, Kluwer Academic Publishers: Boston.

Diaz H.F. and Markgraf V. (Eds.) (1992) *El Nino, Historical and Paleoclimatic Aspects of the Southern Oscillation*, Cambridge University Press: Cambridge.

Diaz H.F. and Markgraf V. (Eds.) (1999) *El Nino and the Southern Oscillation: Multiscale Variability and Societal Impacts*, Cambridge University Press: Cambridge.

Diaz H.F. and Morehouse B. (Eds.) (2003) *Climate and Water: Transboundary Challenges in the Americas*, Kluwer Academic Publishers: Dordrecht.

Ely L.L. (1997) Response of extreme floods in the southwestern United States to climate variations in the late Holocene. *Geomorphology*, **19**, 175–201.

McCabe G.J. and Dettinger M.D. (2002) Primary modes and predictability of year-to-year snowpack variations in the Western United States from teleconnections with Pacific Ocean climate. *Journal of Hydrometeorology*, **3**(1), 13–25.

Meko D.M., Hughes M.K. and Stockton C.W. (1991) Climate change and climate variability: the paleo record. *Managing Water Resources in the West under Conditions of Climate Uncertainty*, National Academy Press, pp. 71–100.

Meko D.M. and Stockton C.W. (1984) Secular variations in streamflow in the western United States. *Journal of Climate and Applied Meteorology*, **23**(6), 889–897.

Osborn T.J., Briffa K.R. and Jones P.D. (1997) Adjusting variance for sample-size in tree-ring chronologies and other regional mean timeseries. *Dendrochronologia*, **15**, 89–99.

Overpeck J. (1995) *Paleoclimatology and Climate System Dynamics*, Reviews of Geophysics, Supplement: U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, pp. 863–871.

Wigley T.M.L., Briffa K.R. and Jones P.D. (1984) On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. *Journal of Climate and Applied Meteorology*, **23**, 201–213.

REFERENCES

Betancourt J.L., Van Devender T.R. and Martin P.S. (Eds.) (1990) *Packrat Middens: the Last 40 000 years of Biotic change*, The University of Arizona Press: Tucson.

Booth R.K. and Jackson S.T. (2003) A high-resolution record of late-Holocene moisture variability from a Michigan raised bog, USA. *The Holocene*, **13**(6), 863–876.

Briffa K.R., Jones P.D. and Schwwingruber F.H. (1992) Tree-ring density reconstructions of summer temperature patterns across western North America since 1600. *Journal of Climate*, **5**, 735–754.

Case R.A. and MacDonald G.M. (2003) Tree-ring reconstructions of streamflow for three Canadian prairie rivers. *Journal of the American Water Resources Association*, **39**, 703–716.

Cayan D.R. (1996) Interannual climate variability and snowpack in the western United States. *Journal of Climate*, **9**, 928–948.

Cayan D.R., Dettinger M.D., Diaz H.F. and Graham N.E. (1998) Decadal variability of precipitation over western North America. *Journal of Climate*, **11**, 3148–3165.

Cayan D.R., Kammerdiener S.A., Dettinger M.D., Caprio J.M. and Peterson D.H. (2001) Changes in the onset of spring in the western United States. *Bulletin of the American Meteorological Society*, **82**(3), 399–415.

Cayan D.R. and Peterson D.H. (1989) *The Influence of North Pacific Atmospheric Circulation on Streamflow in the West*, Geophysical monograph 55, American Geophysical Union, pp. 375–397.

Cleaveland M.K. (2000) A 963-year reconstruction of summer (JJA) streamflow in the White River, Arkansas, USA, from tree rings. *The Holocene*, **10**(1), 33–41.

Conover W. (1980) *Practical Nonparametric Statistics*, *2nd Edition*, John Wiley & Sons: New York.

Cook E.R. and Kairiukstis L.A. (Eds.) (1990) *Methods of Dendrochronology: Applications in the Environmental Sciences*, Kluwer Academic Publishers: Boston.

Cook E.R., Meko D.M., Stahle D.W. and Cleaveland M.K. (1999) Drought reconstructions for the continental United States. *Journal of Climate*, **12**, 1145–1162.

Cook E.R., Meko D.M. and Stockton C.W. (1997) A new assessment of possible solar and lunar forcing of the bi-decadal drought rhythm in the western United States. *Journal of Climate*, **10**, 1343–1356.

Dettinger M.D. and Diaz H.F. (2000) Global characteristics of stream flow seasonality and variability. *Journal of Hydrometeorology*, **1**, 289–310.

Dracup J.A. and Kahya E. (1994) The relationships between U.S. streamflow and La Nina events. *Water Resources Research* **30**(7), 2133–2141.

Ebbesmeyer C.C., Cayan D.R., McLain D.R., Nichols F.H., Peterson D.H. and Redmond K.T. (1991) 1976 step in the Pacific climate: forty environmental changes between 1968–1975 and 1974–1984. *Proceedings of the Seventh Annual Pacific Climate Workshop*, California Department of Water Resources, Interagency Ecological Studies Program: Report 26, pp. 115–126.

Eltahir E.A.B. (1996) El Nino and the natural variability in the flow of the Nile river. *Water Resources Research*, **32**, 131–137.

Ely L.L., Yehouda E. and Cayan D.R. (1994) Anomalous North Pacific atmospheric circulation and large winter floods in the southwestern United States. *Journal of Climate*, **7**(6), 977–987.

Enfield D.B., Mestas-Nunez A.M. and Trimble P.J. (2001) The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental U.S. *Geophysical Research Letters*, **28**(10), 2077–2080.

Enzel Y., Ely L.L., House P.K., Webb R.H. and Baker V.R. (1993) Paleoflood evidence for upper bound to flood magnitudes in the Colorado River Basin. *Water Resources Research*, **29**, 2287–2297.

Forman S.L. and Pierson J. (2003) Formation of linear and parabolic dunes on the eastern Snake River Plain, Idaho, in the nineteenth century. *Geomorphology* **56**, 189–200.

Fritts H.C. (1976) *Tree Rings and Climate*, Academic Press: London.

Goodess C.M., Palutikof J.P. and Davies T.D. (1992) *The Nature and Causes of Climate Change*, Belhaven Press: London.

Graumlich L.J., Pisaric M.F.J., Waggoner L.A., Littell J.S. and King J.C. (2003) Upper Yellowstone river flow and teleconnections with Pacific basin climate variability during the past three centuries. *Climatic Change*, **59**, 245–262.

Guttman N.B. (1999) Accepting the standardized precipitation index: a calculation algorithm. *Journal of the American Water Resources Association*, **35**(2), 311–322.

Hidalgo H.G. and Dracup J.A. (2003) ENSO and PDO effects on hydroclimatic variations of the upper Colorado river basin. *Journal of Hydrometeorology*, **4**, 5–23.

Hidalgo H.G., Piechota T.C. and Dracup J.A. (2000) Alternative principal components regression procedures for dendrohydrologic reconstructions. *Water Resources Research* **36**(11), 3241–3249.

Hirschboeck K.K. (1987) Hydroclimatologically-defined mixed distributions in partial duration flood series. In *Hydrologic Frequency Modeling*, Sing V.P. (Ed.), D. Reidel Publishing Company, pp. 199–212.

IPCC (2001) *Climate Change 2001: The Scientific Basis*, Houghton J.T. (Ed.), Cambridge University Press: Cambridge.

Jury M.R. (2003) The coherent structure of African river flows: composite climate structure and the Atlantic circulation. *Water SA*, **29**(1), 1–10.

Leavitt S.W. and Wright W.E. (2002) Spatial expression of ENSO, drought, and summer monsoon in seasonal δ^{13}C of ponderosa pine tree rings in southern Arizona and New Mexico. *Journal of Geophysical Research*, **107**(D18), 1–10.

Lins H.F. (1997) Regional streamflow regimes and hydroclimatology of the United States. *Water Resources Research*, **33**(7), 1655–1667.

Loaiciga H.A., Haston L. and Michaelsen J. (1993) Dendrohydrology and long-term hydrologic phenomena. *Reviews of Geophysics*, **31**, 151–171.

McCabe G.J. and Dettinger M.D. (2000) Decadal variations in the strength of ENSO teleconnections with precipitation in the

western United States. *International Journal of Climatology*, **19**, 1399–1410.

Meko D.M. and Graybill D.A. (1995) Tree-ring reconstruction of upper Gila river discharge. *Water Resources Bulletin*, **31**(4), 605–616.

Meko D.M., Stockton C.W. and Boggess W.R. (1995) The tree-ring record of severe sustained drought. *Water Resources Bulletin*, **31**(5), 789–801.

Meko D.M., Therrell M.D., Baisan C.H. and Hughes M.K. (2001) Sacramento river flow reconstructed to A.D. 869 from tree rings. *Journal of the American Water Resources Association*, **37**(4), 1029–1040.

Michaelsen J. (1987) Cross-validation in statistical climate forecast models. *Journal of Climate and Applied Meteorology*, **26**, 1589–1600.

Namias J. (1978) Multiple causes of the North American abnormal winter, 1976–77. *Monthly Weather Review*, **106**(3), 279–295.

NCDC (1994) *Time Bias Corrected Divisional Temperature-Precipitation-Drought Index; Documentation for Dataset TD-9640*, Available from DBMB, NCDC, NOAA: Asheville.

Newman M., Combo G.P. and Alexander M.A. (2003) ENSO-forced variability of the Pacific decadal oscillation. *Journal of Climate*, **16**(23), 3853–3857.

Overpeck J.T. (1996) Varved sediment records of recent seasonal to millennial-scale environmental variability. In *Climatic Variations and Forcing Mechanisms of the Last 2000 years*, Jones P.D. and Bradley R.S. (Eds.), Springer-Verlag: Berlin, pp. 479–498.

Palmer W.C. (1965) *Meteorological Drought*, United States Department of Commerce, Weather Bureau, Research Paper No. 45.

Redmond K.T., Enzel Y., Kyle H.P. and Biondi F. (2002) Climate variability and flood frequency at decadal to millennial time scales. *Ancient Floods, Modern Hazards: Principles and Applications of Paleoflood Hydrology*, Water Science and Application, No. 5, American Geophysical Union: pp. 21–45.

Salas J.D., Delleur J.W., Yevjevich V.M. and Lane W.L. (1980) *Applied Modeling of Hydrologic Time Series*, Water Resources Publications: Littleton.

Slack J.R., Lumb A.M. and Landwehr J.M. (1988) *Hydro-Climatic Data Network (HCDN): Streamflow Data Set, 1874–1988*, USGS Water-Resources Investigations Report 93–4076, U.S. Government Printing Office.

Smith L.P. and Stockton C.W. (1981) Reconstructed streamflow for the Salt and Verde rivers from tree-ring data. *Water Resources Bulletin*, **17**(6), 939–947.

St. George S. and Nielsen E. (2002) Hydroclimatic change in southern Manitoba since A.D. 1409 inferred from tree rings. *Quaternary Research*, **58**(2), 103–111.

Stahle D.W., Cook E.R., Cleaveland M.K., Therrell M.D., Meko D.M., Grissino-Mayer H.D., Watson E. and Luckman B.H. (2000) Tree-ring data document 16th century megadrought over North America. *EOS Transactions*, **81**(12), 121–125.

Stine S. (1994) Extreme and persistent drought in California and Patagonia during mediaeval time. *Nature*, **369**, 546–549.

Stine S. (1998) Mediaeval climate anomaly in the Americas. In *Water, Environment and Society in Times of Climatic Change*, Issar A.S. and Brown N. (Eds.), Kluwer, pp. 43–67.

Stockton C.W., Boggess W.R. and Meko D.M. (1985) Climate and tree rings. In *Paleoclimate Analysis and Modeling*, Hecht A.D. (Ed.), John Wiley & Sons, pp. 71–150.

Stockton C.W. and Fritts H.C. (1973) Long-term reconstruction of water level changes for Lake Athabasca by analysis of tree rings. *Water Resources Bulletin*, **9**(5), 1006–1027.

Stockton C.W. and Jacoby G.C. (1976) *Long-Term Surface-Water Supply and Streamflow Trends in the Upper Colorado River Basin*, Lake Powell Research Project Bulletin No. 18, National Science Foundation.

Street-Perrott F.A. and Harrison S.P. (1985) Lake levels and climate reconstruction. In *Paleoclimate Analysis and Modeling*, Chap. 7, Hecht A.D. (Ed.), John Wiley & Sons, pp. 291–340.

Tardif J. and Bergeron Y. (1997) Ice-flood history reconstructed wit tree-rings from the southern boreal forest limit, western Quebec. *The Holocene*, **7**, 291–300.

Vose R.S., Schmoyer R.L., Steurer P.M., Peterson T.C., Heim R., Karl T.R. and Eischeid J. (1992) *The Global Historical Climatology Network: Long-Term Monthly Temperature, Precipitation, Sea Level Pressure, and Station Pressure Data*, ORNL/CDIAC-53, NDP-041, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory: Oak Ridge.

Wolfe S.A., Huntley P.P., Ollerhead D.J., Sauchyn D.J. and MacDonald G.M. (2001) Late 18th century drought-induced sand dune activity, Great Sand Hills, Saskatchewan. *Canadian Journal of Earth Sciences*, **38**, 105–117.

Woodhouse C.A. (1999) Artificial neural networks and dendroclimatic reconstructions: an example from the Front Range, Colorado, USA. *The Holocene*, **9**(5), 521–529.

Woodhouse C.A. (2001) A tree-ring reconstruction of streamflow for the Colorado Front Range. *Journal of the American Water Resources Association*, **37**(3), 561–569.

Woodhouse C.A. (2003) A 431-year reconstruction of western Colorado snowpack from tree rings. *Journal of Climate*, **16**(10), 1551–1561.

Yu Z., Ito E., Engstrom D.R. and Fritz S.C. (2002) A 2100-year trace-element and stable-isotope record at decadal resolution from Rice lake in the northern Great Plains, USA. *The Holocene*, **12**(5), 605–617.

201: Land-Atmosphere Models for Water and Energy Cycle Studies

BART NIJSSEN[1] AND LUIS A BASTIDAS[2]

[1]Departments of Hydrology and Water Resources/Civil Engineering and Engineering Mechanics, University of Arizona, Tucson, AZ, US
[2]Department of Civil and Environmental Engineering/Utah Water Research Laboratory, Utah State University, Logan, UT, US

Land surface models simulate the temporal evolution of the energy and water balance at and near the land surface. These models can be used coupled to an atmospheric model, in which case feedbacks between the land surface and atmosphere are explicitly represented, or they can be used in a stand-alone, uncoupled mode, in which case observed or simulated atmospheric conditions act as time-varying boundary conditions. This article reviews the basic layout of land surface models, their history, and their uses, specifically in uncoupled simulations. In this context, the challenges of model performance evaluation and the role of model intercomparison projects are discussed. The article ends with a look toward the future and the next series of developments that land surface models are likely to undergo.

INTRODUCTION

About 70% of the Earth's surface is covered by oceans, seas, lakes, and other large bodies of open water. As such, much of the interaction between the Earth's surface and the atmosphere occurs as an exchange of energy, mass, and momentum across a liquid/gas interface. Immediately above this interface, the air is saturated with moisture because the supply of water is essentially unlimited and moisture can be freely exchanged. As a result, the sea surface temperature (SST) defines both the thermal energy state of the sea surface and the amount of moisture in the air immediately above it, two important lower boundary conditions for atmospheric circulation models (Betts *et al.*, 1996).

Over land, the situation is more complicated because the supply of moisture is neither unlimited nor can moisture be exchanged freely. The amount of water available for exchange with the atmosphere is a function of the amount of water stored in the soil, which varies with time, as well as the type of soil and its vegetative cover. Vegetation actively regulates the exchange of moisture with the atmosphere through stomatal control. Stomata are small openings on leaf surfaces through which a plant exchanges carbon dioxide, oxygen, and water with the atmosphere. Stomatal control allows a plant to limit the amount of transpiration at times when moisture is in limited supply. As a result, the air above the land–atmosphere interface is often unsaturated, and knowledge of the land surface temperature alone is generally insufficient to determine the moisture state of the atmosphere close to the land surface.

Because of the large specific heat of water, active mixing, and the sheer amount of water stored in the world's oceans, the ocean surface provides a much larger amount of heat storage than does the land surface. In addition, ocean currents can transport large amounts of heat over long distances, while horizontal heat transport through the land surface is negligible. Instead, the atmosphere provides a means of heat transport between different locations at the land surface.

Because of its smaller role in heat storage and transport, and because of its smaller spatial extent, the land surface is arguably of lesser importance as a lower boundary condition to the atmosphere than the ocean surface. However, for many important exchange processes, the land surface exhibits greater spatial and temporal variability in controlling characteristics than the ocean. For example,

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

soil hydraulic properties and water-holding capacities vary greatly over short distances, as does vegetation. Vegetation typically exhibits a seasonal cycle of growth and decline, and surface albedo varies as a function of time and place, with direct effects on the amount of energy reflected and absorbed at the land surface. In addition, even though in many cases the effects of the land surface on the atmosphere are mostly local, the land boundary condition is of special interest because this is where people live.

Although the exchange mechanisms of mass, energy, and momentum between the land surface and the atmosphere are reasonably well understood qualitatively, it is precisely this variability of land surface properties and states in both space and time that complicates the development of a quantitative description of land–atmosphere exchange processes. A quantitative description of the mass, energy, and momentum exchanges, as well as the spatial distribution and temporal evolution of mass and energy storage at the land surface is the domain of land surface models. These models serve both as a representation of the lower boundary in atmospheric models and as stand-alone models to study and predict land surface fluxes and states.

The land surface continually exchanges energy with the overlying atmosphere. Incident shortwave radiation, either in the form of diffuse or direct beam radiation, originates from the sun, and is partly absorbed by the land surface. Direct beam radiation is shortwave radiation that is not scattered by the atmosphere and comes from the direction of the solar disk, while diffuse radiation is scattered by the atmosphere and emanates from all directions of the sky. Globally, only 45% of the shortwave or solar radiation at the top of the atmosphere is absorbed by the land surface. 25% is reflected by the atmosphere, 35% is absorbed by the atmosphere, and 5% is reflected by the Earth's surface (Trenbert, 1992). Note that these figures reflect global average values, including both land and ocean surfaces. Absorbed shortwave radiation (R_s) at the surface is given by

$$R_s = (1 - \alpha) R_s^{\downarrow} \tag{1}$$

where α is the shortwave albedo, and $R_s^{\downarrow}$ is the incoming shortwave radiation at the land surface. Absorption of solar energy by the atmosphere and the land surface leads to an increase in heat storage and hence temperature. Any body emits radiation, so both the land surface and the atmosphere reradiate some of the energy absorbed from the sun. The land surface and atmosphere emit radiation at a much longer wavelength than the sun, because of their lower temperatures. This emitted radiation is referred to as long-wave radiation. Part of the energy emitted by the atmosphere is intercepted by the land surface and consequently the radiative energy incident on the land surface is the sum of the incident shortwave and long-wave radiation. The land surface itself emits long-wave radiation at a rate of

$$R_L^{\uparrow} = \varepsilon \sigma {T_s}^4 \tag{2}$$

where $R_L^{\uparrow}$ is the emitted long-wave radiation, ε the emissivity of the land surface, σ the Stefan-Boltzman constant, and T_s the land surface temperature.

The land surface and overlying atmosphere also exchange energy by means of turbulent transport, in which eddies of varying sizes transport heat, moisture, and momentum. The energy exchanged through the transport of warm air parcels is termed *sensible heat* (H), and is a function of the temperature gradient near the land surface

$$H = -\rho c_p K_h \frac{\partial T}{\partial z} \tag{3a}$$

where ρ is the air density, c_p is the specific heat of air at constant pressure, $\partial T/\partial z$ is the temperature gradient near the surface, and K_h is a turbulent exchange coefficient or eddy diffusivity. Equation (3a) is often written in the integral form

$$H = -\rho c_p \frac{(T_a - T_s)}{r_{ah}} \tag{3b}$$

where T_a is the air temperature, and r_{ah} is an aerodynamic resistance to sensible heat exchange between the levels where T_a and T_s are measured or calculated.

Energy exchanged through turbulent transport of water vapor is termed *latent heat exchange* (LE). The phase change that occurs when water is evaporated from open water surfaces, transpired by vegetation or sublimated from snow and ice, requires a large amount of energy. This energy is released when the water vapor condenses, which can happen large distances away from the original place of evaporation. LE between the land surface and the atmosphere can be described as

$$LE = -\lambda \rho K_v \frac{\partial q}{\partial z} = -\frac{\rho c_p}{\gamma} K_v \frac{\partial e}{\partial z} \tag{4a}$$

where λ is the latent heat of vaporization, K_v is the eddy diffusivity for vapor exchange, q is specific humidity, e is vapor pressure, and γ is the psychrometric constant ($c_p p/0.622\lambda$ with p the atmospheric pressure). Similar to equation (3b), this equation can also be written in the integral form as

$$LE = -\frac{\rho c_p}{\gamma} \frac{(e_a - e_*)}{r_v} \tag{4b}$$

where e_a is the vapor pressure of the air, e_* is the vapor pressure near the surface, and r_v is an effective resistance to vapor exchange between the levels where e_a and e_* are measured or calculated.

The land surface not only exchanges energy with the overlying atmosphere, but also exchanges heat with deeper soil layers, largely through conduction. This heat transport (G) is generally described by the heat flux equation

$$G = -k_s \frac{\partial T}{\partial z} \tag{5}$$

where k_s is the soil thermal conductivity and T is the temperature in the soil.

Because energy is conserved, any imbalance between incoming and outgoing fluxes of energy will result in a change of the energy stored at the ground surface, largely expressed as a change in land surface temperature. Thus, the change in energy stored at the land surface (S_E) is given by

$$\frac{dS_E}{dt} = R_n - H - LE - G \tag{6}$$

where R_n is the net radiation given by

$$R_n = (1 - \alpha)R_s^{\downarrow} + R_L^{\downarrow} - R_L^{\uparrow} \tag{7}$$

where $R_L^{\downarrow}$ is the long-wave radiation flux incident on the land surface. For surfaces with small thermal heat capacities, the amount of energy storage is small and the net radiation is balanced by the sensible, latent and ground heat fluxes.

Conservation of mass can be used to develop a mass balance equation for a segment of the land surface similar to equation (6). In hydrology, the mass conservation of water is of primary interest. In this case, the change in the amount of water stored in a section of the land surface or control volume equals the net flow of water into and out of this segment. Storage changes consist of changes in water stored in the soil, on the surface and in the vegetation canopy if one is present. Surface water storage can consist of liquid water in lakes, ponds, and depressions, or of water stored in solid form as snow or ice. The main moisture flux into the control volume is precipitation, while evapotranspiration, surface runoff, and percolation to deeper soil layers are the main mass fluxes out of the control volume. In its simplest form, the mass conservation equation for water or water balance equation takes the form

$$\frac{dS_W}{dt} = P - ET - Q_s - Q_g \tag{8}$$

where S_W is the amount of moisture stored at and near the surface per unit of land surface, P is the precipitation rate, ET is the evapotranspiration rate, Q_s is the surface runoff, and Q_g the percolation of water to deeper soil layers. The evapotranspiration term links the water balance and energy equation, because the latent heat-flux term in equation (6) is the energy equivalent of the evapotranspiration term in equation (8). Consequently, the energy and water balance at the land surface are tightly linked. The conservation equations (6) and (8) form the basis for all land surface models.

Land surface models typically divide the land surface in a large number of model elements, for each of which the energy and moisture balance terms are calculated. Because land surface schemes were initially developed to simulate feedbacks between the land surface and the atmosphere, they largely focus on the vertical exchange of energy and moisture and tend to neglect horizontal exchanges between the model cells. Some schemes include routing models that will transport the runoff generated at the level of individual grid cells through an idealized channel network to simulate river discharge at a number of locations in the basin.

EVOLUTION OF LAND SURFACE MODELING

Richardson (1922), in his book *Weather prediction by numerical process*, already pointed out that "the atmosphere and the upper layers of the soil or sea form together a united system". He identified three different varieties of earth surface that should be represented, namely, sea, bare soil, and soil covered by vegetation. Richardson suggested that the combination of these surface types in a single model cell (200×200 km in his case) could be represented by area-weighted, average constants, with "due regard being paid to the season and the customary times of ploughing, harrowing, sowing, and the like". Heat and water transfer in the soil could be represented by finite difference approximations of the differential equations governing their transport. The soil layers were discretized so that the layers were progressively thicker with increasing depth to allow for a detailed treatment of the steep gradients near the surface (Richardson suggested layer boundaries at depths of 0, 0.0192, 0.0639, 0.191, 0.536, and 1.475 m below the surface). Richardson based much of the empirical part of his model on observations in the British Isles. These observations indicated that during the wintertime, which tends to be wet and damp in northwestern Europe, evaporation from bare soil was similar to that from an open water surface. During the summer, soil evaporation was to be modeled explicitly as a function of moisture and vapor transport in the unsaturated soil. The equation that Richardson suggested is the same as the Richards equation in soil water transport, except for the additional vapor flux term (note that Richards published his equation for moisture transport in unsaturated media almost a decade later (Richards, 1931)). Evaporation from vegetated surfaces was represented differently. Richardson recognized that "a portion of rain, and the greater part of dew, is caught on foliage and evaporated without ever reaching the soil". Citing work by Brown and Escombe (1900) and Brown and Wilson (1905), he suggested that transpiration through leaf surfaces might be represented analogously to

electric conduction, in that "the rate of transpiration may be said to be inversely as the 'resistance' of the stomata" (Richardson, 1922). Transpiration is then a function of this resistance and the vapor gradient between the leaf cavities and the surrounding air. Scaling from a single leaf to the leaf canopy is done through an increase in the conductance. Transpiration diminishes when soil moisture falls below a certain level.

The treatise by Richardson (1922) is remarkable given that he outlined a blueprint for a numerical weather prediction (NWP) model, including a land surface model, before the advent of modern computers. Richardson discusses an example on "computing forms" for a single 6-h time step, in which "all computations were worked twice and compared and corrected" and "multiplications were mostly worked using a 25 cm slide rule". However, given these computational resources, near real-time simulation of weather and land surface conditions did not make further significant advances until the development of modern computers.

Pitman (2003) has given an extensive review of the development of modern land surface schemes designed for climate models. Following Sellers *et al.* (1997a), he considers three generations of land surface models in climate models. The first-generation model is typified by the "Manabe bucket model" (Manabe, 1969), and is much simpler than the land surface scheme envisioned by Richardson in 1922. This model ignored the heat flux into the soil, assumed a globally constant soil depth and water-holding capacity, and evaporation was limited when the soil moisture fell below a certain threshold. Vegetation, and its role in controlling evaporation from the land surface, was not represented explicitly. Runoff was generated when the moisture storage capacity was exceeded.

Second-generation models, starting with the work by Deardorff (1978), are characterized by an explicit treatment of vegetation and a more realistic representation of evaporation, including evaporation from interception storage and a separation of soil evaporation and transpiration. The soil is typically divided into a number of layers, and vertical moisture transport between the layers is modeled on the basis of a simplified version of Richards equation. These models also allow for a dynamic representation of the ground heat flux, initially based on the force-restore method (Bhumralkar, 1975; Deardorff, 1978), and later on finite difference approximations of the heat-flux equation. As such, they are much closer to the land surface schemes envisioned by Richardson (1922), than the first-generation models. Runoff is typically produced through a fast response or event response mechanism that simulates surface runoff, and a slow runoff mechanism that simulates subsurface runoff. When these land surface models are employed in climate models, all runoff is directly moved to the ocean, unless an explicit routing scheme is employed.

The production of runoff and the partitioning between fast and slow response mechanisms varies widely between the existing land surface schemes as is evident from model intercomparison projects such as the Project for the Intercomparison of Land Surface Parameterization Schemes (PILPS) (e.g. Lohmann *et al.*, 1998; Bowling *et al.*, 2003; Nijssen *et al.*, 2003). As a result of the strong interaction between the energy and moisture balance, these differences in runoff production and the resulting differences in moisture storage, have a direct impact on the partitioning of the available energy in latent and sensible heat (Pitman, 2003).

Although the second-generation land surface schemes include much more process understanding than the first-generation schemes, Pitman (2003) raises the point that in the case of coupled simulations "it is an open question as to what extent changing a first-generation scheme to a second-generation scheme has really improved the simulation of climate." However, he concludes that there is a definite majority view in the land surface modeling community that second-generation models perform better than first-generation models. This is supported by the observation that in model intercomparison studies they tend to perform better, as a group, than first-generation models.

Development of third-generation models started in the late 1980s and is currently (2005) an active area of model development and research. These models are characterized by a much more complete representation of plant physiological processes, in particular, the representation of plant photosynthesis and canopy conductance. Consequently, third-generation models incorporate a representation of the carbon cycle, and its interaction with the water and energy cycle (Pitman, 2003). Although these third-generation land surface models can respond to changes in atmospheric conditions by growing bigger trees or larger leaves, they typically do not allow for a succession of vegetation types in response to changes in atmospheric conditions. Dynamic vegetation models, which allow such a succession of vegetation types, typically operate at longer timescales than the land surface schemes, but a combination of the two modeling approaches will be important to simulate long-term climate dynamics and their interaction with land surface processes.

Land surface models used in uncoupled studies, that is, without being interactively coupled to an atmospheric model, are essentially the same as those in climate and NWP models. The main differences are that they are typically operated at higher spatial resolutions and that water is often routed through some idealized channel network to simulate river discharge at selected locations within a basin. Because of the large spatial extent of the individual grid elements in most land surface applications (spatial resolutions are typically 1/8° or larger), horizontal exchanges of moisture between grid elements over the surface or through the subsurface have generally been neglected. However, as the

size of model elements decreases as a result of increased computing resources and increased data availability from remote sensing, the assumption that land surface model elements do not interact becomes more tenuous. As a result, new approaches are being developed to allow for horizontal exchange between model elements and to better represent saturated groundwater processes.

ROLE OF LAND–ATMOSPHERE MODEL EXPERIMENTS IN REDUCING ERRORS AND UNCERTAINTIES

Land surface model validation and intercomparison projects such as PILPS (Henderson-Sellers *et al.*, 1993, 1995) and the Global Soil Wetness Project (GSWP) (Dirmeyer *et al.*, 1999) helped to assess the previous and current generations of land surface models. PILPS has made major contributions to their evaluation at the local scale by using observed atmospheric boundary conditions or forcings and high quality measured fluxes for validation of the simulated processes. In GSWP, on the other hand, the land surface models were used in a mode much closer to their application in global climate models (GCMs), that is, the models were implemented on a global grid with a resolution coarser than the typical spatial scale of surface processes. Both of these projects have proven very useful in identifying model strengths and weaknesses and need to be continued with the next generation of land surface models for the validation of added processes. However, they also need to be supplemented by new experiments.

Results of field experiments that measure surface fluxes and variables as local point values or on a regional scale are regularly used to test surface parameterization schemes. For example, ARME data over the Amazon basin (Shuttleworth *et al.*, 1984) have been used in validating SiB (Sellers *et al.*, 1989), ISBA (Noilhan *et al.*, 1993), and the ECMWF surface model (Viterbo and Beljaars, 1995). FIFE data over the Konza Prairie in Kansas, United States (Sellers *et al.*, 1988, 1992), and Cabauw data in the Netherlands (Beljaars and Bosveld, 1997) have been used to validate the ECMWF model (Betts *et al.*, 1993; Beljaars and Viterbo, 1994) and, in the case of FIFE, to validate ISBA. HAPEX/MOBILHY data in southern France (André *et al.*, 1986) have been used to test the ISBA model and the ECMWF model. SEBEX data, over the Sahelian region, have been used to validate the ECMWF model (Beljaars and Viterbo, 1998). Data from several sites in Illinois have been used to improve the parameterization of the NCEP Noah model (Mitchell *et al.*, 2000). At most of these field sites, snow is not a major component of the hydrological cycle. Field campaigns in the Canadian boreal forest (BOREAS, Sellers *et al.*, 1997b) and the Scandinavian forest (NOPEX, Halldin *et al.*, 1998) and more currently the Cold Land Processes Experiment (CLPX) in northern Colorado are filling that gap.

As applications of land surface models change from the plot scale to the regional scale and from an uncoupled to a coupled mode, new experiments must be designed, which will specifically address the representation of subgrid-scale variability of the surface and the feedbacks between surface processes and the atmosphere. The difficulties encountered in attempting to compare simulations and observations for these different applications can be illustrated, for instance, for the case of soil moisture. In some of the PILPS experiments, local soil moisture measurements could be compared with simulated soil moisture values on the basis of model simulations that were forced with locally measured meteorological variables (e.g. Shao and Henderson-Sellers, 1996). However, when each model grid cell represents a large area rather than a small plot, it is not obvious how model simulated soil moisture is related to soil moisture measured at a specific location. This is further complicated because soil moisture, though it goes by the seemingly unambiguous name, has a necessarily model-specific meaning (see the Section on "Model performance evaluation"). In GSWP-1, only the soil moisture anomalies at a spatial resolution of $1^\circ \times 1^\circ$ could be evaluated, because the models used different water-holding capacities, which were all different from the one that was deduced from observations (Entin *et al.*, 1999). When a land surface model is coupled to an atmospheric column model, the relationship between soil moisture and the atmospheric conditions can again be explored at the local scale. This setup was used by Douville *et al.* (2000) to determine how the assimilation of near-surface temperature and humidity can improve the simulated soil moisture. Finally, with coordinated GCM sensitivity experiments the impact of the different approaches to soil moisture modeling on climate change could be evaluated in a coupled environment (Crossley *et al.*, 2000).

Unfortunately, we are not presently able to say how the results obtained in PILPS can be related to the uncertainty in soil moisture changes found for a climate with increased greenhouse gas concentrations. Within this context, the Global Energy and Water Cycle Experiment (GEWEX) Global Land Atmosphere System Studies (GLASS) working group intends to foster an evaluation of the next generation of land surface models and to coordinate their evaluation in different applications. GLASS is also serving as an interface between the land surface community and other GEWEX projects. The structure of GLASS highlights the spatial scales at which the schemes are applied and the degree of interaction allowed with the atmosphere. GLASS also supports ALMA (Assistance for Land Surface Modelling Activities), which provides an infrastructure and technical support for these model intercomparison projects. To facilitate the exchange of forcing data for land surface schemes and the results produced by these schemes, ALMA has established data standards. The aim is to have a data exchange format that is stable but still general and

flexible enough to evolve with the needs of land surface schemes. This should ensure that ALMA-compatible data exchange procedures only need to be implemented once in a land surface model and that future model intercomparisons of land surface schemes can be run more efficiently. PILPS 2e (Bowling *et al.*, 2003), PILPS C1, and the current (2005) PILPS San Pedro Experiments have made use of those standards.

Another current experiment run under GLASS, the Global Land-Atmosphere Coupling Experiment (GLACE) is a multimodel intercomparison experiment that focuses on the ability of the land surface state to affect rainfall generation and other atmospheric processes. The experiment aims to quantify the strength of land–atmosphere coupling in the different global atmospheric models used for weather and climate studies. The hope is that the development of a "table" of coupling strengths would aid in the interpretation of the many land–atmosphere interaction studies now appearing in the literature.

APPLICATIONS OF UNCOUPLED LAND SURFACE MODELS

A problem in coupled land surface/atmosphere simulations is that errors in atmospheric conditions will cause errors in the energy and moisture states of the land surface. Similarly, errors in land surface conditions will cause errors in the energy and moisture states of the atmosphere. Consequently, the land surface and atmospheric states in a coupled model system will typically drift away from observed conditions and develop a climatology or long-term average state, significantly different from the observed. This is particularly problematic for real-time weather and climate prediction models, because the future state of the atmosphere is at least in part dependent on the current state of the atmosphere, ocean, and land surface. Hence, errors in the characterization of the current state diminish the reliability and accuracy of model forecasts.

For this reason, observed conditions are often used as boundary or initial conditions. For example, observed sea surface temperatures, which can be derived from satellite observations, are often used as boundary conditions in coupled land surface/atmosphere models that do not include an explicit ocean model. NWP is an initial value problem, in which the model forecasts a future state on the basis of the initial state (Kalnay, 2003). Because of the relative scarcity of atmospheric observations, particularly of the conditions in the upper levels of the atmosphere, a complete specification of the initial state is not possible based on observations alone. Hence, information, for example, based on climatology or a previous model state, is often used as an additional source of information. In NWP, the process that produces the initial state is referred to as analysis. Data assimilation is an analysis technique that integrates real-world observations and model states, to determine the best possible representation of the state of the system. For a discussion of data assimilation methods in the atmospheric sciences, see for example Daley (1991) or Kalnay (2003).

In a global context, the state of the land surface is less important than the state of the oceans for forecasting future states of the atmosphere. However, a number of studies have demonstrated the role of the current state of the land surface in determining future conditions (e.g. Garratt, 1993; Koster *et al.*, 1995, 2003; Brubaker and Entekhabi, 1996). Consequently, not only atmospheric and oceanic observations should be assimilated into NWP and earth systems models, but land surface states should be assimilated as well. Unfortunately, direct observations of land surface states, such as soil moisture, soil temperature, and snow water equivalent are made only in a few locations. Even then they are rarely available at the same scale as the computational elements in NWP and earth systems models. Hence, an alternative approach has been developed in which land surface models, forced with observed atmospheric conditions, are used to produce continuous fields, continuous in both space and time, of land surface states. These land surface states in turn can then be assimilated into NWP models. A prominent example of this methodology is the Land Data Assimilation System (LDAS) started in 1999, which had as its original goal the development of real-time, hourly, distributed, uncoupled land surface simulations for the United States domain at 0.125° resolution (Mitchell *et al.*, 1999). This has led to the subsequent development of a Global LDAS or GLDAS (Rodell *et al.*, 2004). As part of LDAS, a number of physically based land surface models are run on the basis of model-independent, observation-based atmospheric forcings. Mitchell *et al.* (1999) describe the goals of the LDAS project as "(i) improve land surface model physics by sharing algorithms, methods, and insights against a backdrop of joint intercomparison and validation, (ii) identify causes of and reduce extent of the spread in surface water fluxes and surface water storage typically seen in land surface model intercomparisons, (iii) reduce the uncertainty in land surface water budget estimates, (iv) utilize several new real-time GCIP-sponsored forcing and validation products, (v) compare uncoupled LDAS with traditional coupled 4DDA, (vi) support water resource application (water supply and agriculture), and (vii) provide land surface initial conditions (e.g. soil moisture and snowpack) for both (a) predictability studies of the role of sea versus land lower boundary conditions on seasonal forecasts and (b) real-time weather and climate model predictions on several timescales (days, weeks, seasons)".

The geographical extent of the original LDAS project covered a large part of North America, with the main focus on the contiguous United States. To provide global fields of land surface states, a GLDAS was developed

jointly by the National Aeronautics and Space Administration (NASA), Goddard Space Flight Center (GSFC), and the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP) (Rodell *et al.*, 2004). This latter system can be operated on variable spatial resolutions. Koster *et al.* (2004) have demonstrated the potential benefit of realistic initialization of land surface states on subseasonal forecasts in coupled simulations.

As a further step in engineering computational environments that can take advantage of multiprocessor computing clusters and large amounts of real-time or near real-time atmospheric and land surface inputs, NASA has developed a land information system (LIS), in which multiple land surface models are embedded (Tian *et al.*, 2004; Kumar *et al.*, 2004). This software environment allows simulations to be run at variable spatial and temporal resolutions and, at least to some extent, automates the necessary interpolation and averaging of model inputs.

While the above-mentioned data assimilation systems simulate most land surface moisture and energy states, the Snow Data Assimilation System (SNODAS) developed by the NOAA National Weather Service (NWS) National Operational Hydrologic Remote Sensing Center (NOHRSC), was specifically developed to "to provide a physically consistent framework for integrating the wide variety of snow data that is available at various times" (Carroll *et al.*, 2001). This system is driven with near real-time surface weather observations and the output from NWP models, while observations of snow water equivalent, snow depth, and snow areal extent are assimilated to update the model states. SNODAS products are made available on an operational basis through the National Snow and Ice Data Center (NSIDC) since October 2003 (Barrett, 2003).

Uncoupled land surface simulations can also provide a valuable tool in the study of predictability within the earth system at seasonal and interannual scales. Predictability can be defined as the extent to which the future state or flux of a system can be predicted on the basis of its current state. Not all land surface or atmospheric states and fluxes demonstrate the same level of predictability. Even for a given land surface state or flux, the level of predictability may vary as a function of place and time. For example, Maurer *et al.* (2003) used a 50-year uncoupled simulation with the Variable Infiltration Capacity (VIC) land surface model (Liang *et al.*, 1994; Cherkauer *et al.*, 2003) to study the predictability of runoff in the Mississippi Basin in the United States. The 50-year simulation was forced with a station-based forcing data set at a spatial resolution of 0.125° and a temporal resolution of 3 h (Maurer *et al.*, 2002). Using the Southern Oscillation Index (SOI) and Atlantic Oscillation (AO) as surrogates for climate predictability, they found that, in the Mississippi Basin, soil moisture is the dominant source of runoff predictability at lead 0 in all seasons. Lead 0 here refers to the season at the start of which the soil moisture is known and represents an average lead time of 1.5 months (note that the analysis in Maurer *et al.* (2003) was done at a seasonal timescale). The only exception to this was the summer season (June–August) in the western, more mountainous, part of the basin. During this season and in this region, snow water equivalent formed the dominant source of predictability. For lead times of one to three seasons, the climate indicators provided a small, but significant predictability for winter (December–February) runoff, especially in the eastern part of the basin.

Uncoupled land surface simulations have also been used in scenario evaluation studies. For example, Matheussen *et al.* (2000) used the VIC model to simulate the effects of land cover change on streamflow in the interior Columbia River Basin (United States and Canada). They compared model results on the basis of two different vegetation scenarios to determine vegetation-related changes in runoff, evaporation, and snow accumulation. The vegetation scenarios consisted of historical vegetation conditions, estimated for 1900 prior to widespread European settlement of the basin, and current vegetation conditions, estimated for 1990. An unanswered question at this point is how well the model-predicted sensitivities to changes in land cover mirror the observed changes.

Nijssen *et al.* (2001) used the same land surface model to evaluate the effects of climate change on nine, continental river basins. Changes in climate were based on the differences between control runs and climate change runs from a number of GCMs. In addition, they studied the effect of a fixed change in temperature (+2 °C) or precipitation (+10%) in any given season on the runoff and evaporation in all seasons. Although it could be argued that the effects of climate change on hydrology would be better represented in coupled land surface/atmosphere models, the output from these coupled models generally shows significant biases in the simulation of both temperature and precipitation under current climate conditions. These biases are often so large that direct application of the modeled meteorology in a land surface model is not meaningful (e.g. Doherty and Mearns, 1999). Consequently, uncoupled simulations allow for the application of a bias correction to the GCM output prior to the simulation of the land surface hydrology. Furthermore, the spatial resolution of most GCMs, particularly when operated in climate mode, is significantly coarser than the resolution supported by most uncoupled land surface models. GCM output is therefore "downscaled" to the resolution of the land surface model. In Nijssen *et al.* (2001), GCM model output was bias corrected by calculating the differences in temperature and precipitation between a control run and the climate change run, and imposing these differences on a historic, station-based forcing data set. Wood *et al.* (2002) discusses a probability mapping method for bias correcting GCM output. The

probability mapping method maps the empirical cumulative density function of a GCM predicted variable to the empirical cumulative density function of historic observations of this same variable. Although this method does not accommodate changes in the shape of the empirical density function with time (everything is mapped to the historic cumulative density function), it provides a more flexible method for bias correcting than a simple superposition of differences. Wood *et al.* (2004) evaluated the hydrological implications of six different techniques for downscaling climate model output.

MODEL PERFORMANCE EVALUATION

Advances in our understanding of the processes that control biosphere–atmosphere interactions are spurring the development of increasingly sophisticated land surface models that aim to represent the transfer of mass, energy, and momentum between a vegetated surface and the atmosphere (Dickinson *et al.*, 1986; Sellers *et al.*, 1986). The use of these land surface models is being extended beyond their original implementation in general circulation models to diverse applications in hydrology, biogeochemistry, and ecology. Model improvements include better representations of the vegetation physiology and attempts to represent surface heterogeneity at the GCM subgrid scale. However, model predictions are biased due to simplified model physics, errors in input forcing, limited knowledge of parameter values, and inadequate treatment of subgrid-scale spatial variability. Consequently, even state-of-the-art land surface model predictions are erroneous in space and time. When coupled with GCMs to predict future climate scenarios, these biases may lead to systematic or long-term errors and may render unrealistic or unreliable forecasts of land surface hydrology.

Land surface models are largely defined by: (i) parameters, which are generally time-invariant descriptions of surface characteristics (such as percent-vegetated cover, soil porosity, surface roughness, etc.); and (ii) states, which are storages of water and energy that are propagated in time by the model physics. Although some model parameters may be measured or estimated directly (albedo, percent-vegetated cover, leaf area index, etc.) at the patch or grid scale, other parameters (soil hydraulic conductivity, stomatal resistance, aerodynamic resistance, etc.) are not easily measured (even at the patch scale) and are difficult to interpolate to large GCM size scale. The majority of land surface schemes currently use look-up tables to define large-scale parameter values associated with specific soil-vegetation regimes. For these parameters, a reasonable range estimated from published values may be the best estimation of a grid average value.

Originally, model intercomparison studies assumed that parameters with the same physical interpretation should have the same value in all land surface models. However, land surface models are used on a GCM grid scale and, hence, effective model parameter values are required to model grid-scale average processes (e.g. Bastidas *et al.*, 1999; Beven, 1995; Brewer and Wheatcraft, 1994; Gupta *et al.*, 1999; Sorooshian *et al.*, 1999). These effective values are, by their very nature, dependent on the particular parameterization that varies from model to model. Consequently, even though the parameters might share the same name and conceptual representation among different land surface models, their model-specific values vary.

The increase in complexity of process representations has resulted in large numbers of model parameters. The actual number depends on the number of soil and canopy layers and the complexity of the associated process descriptions. Most of these parameters vary in space, and some also vary in time. For example, soil hydraulic properties must be specified to establish the relationship between hydraulic conductivity, soil metric potential, and soil water content. The values of these soil properties significantly affect model behavior, implying that land surface models need to be calibrated or regularly corrected, particularly if the model output has significant sensitivity to variations in the parameter. In that case, an adjustment or calibration of the initial parameter value may be needed to better match observations.

In a discussion of environmental modeling, Beck (2002a) points out that the contention once was that any model requiring calibration against past observations was thereby rendered inferior, and certainly incapable of extrapolation to conditions not previously observed. In contrast, it was assumed that a model based *truly* on the Laws of Physics would not suffer from such difficulties, because these models would employ physical constants whose values are known *a priori* and which are universally applicable. Ideally, these models would contain no parameters without immediate physical meaning. All their parameters could be determined through independent measurements in the field; none would require calibration. Beck (2002a) argues that such a happy state of affairs is not attainable in practice. To believe otherwise has been an illusion, albeit an extraordinarily useful one for the development of models.

In addition, even the most advanced data acquisition platforms in operation in 2005 do not adequately capture the spatial heterogeneity of geophysical systems and biological states of the environment (Beck, 2002b). As a result, land surface models contain a profusion of unobserved state variables. Reconciling model with observed behavior to improve understanding is quintessentially an issue of demonstrating, beyond reasonable doubt, that a match of the two approximations of the truth has not been achieved at the expense of imposing absurd values on the model parameters.

Model performance is typically evaluated by calculating the error of the simulation, which is usually defined as the difference between the observed and the computed series. This difference has two sources: observation errors and model errors. The model error is the combined result of the model structural error and the error due to improper identification of the parameter set:

$$\begin{matrix}\text{Total}\\ \text{error}\end{matrix} = \begin{matrix}\text{Data}\\ \text{error}\end{matrix} + \begin{matrix}\text{Model}\\ \text{structural}\\ \text{error}\end{matrix} + \begin{matrix}\text{Parameter}\\ \text{specification}\\ \text{error}\end{matrix} \quad (9)$$

The identification and elimination of the latter error is one of the goals of the parameter estimation procedures. The reduction of the model structural error (the total elimination of this error would not be feasible) is a task for the developers of the specific models. Some of the problems with the model structure are usually identified if a proper calibration procedure is used. The data error is related to the way the information is collected, the instrumentation used, and its precision. In general, the identification of these separate errors is a difficult task. One approach to estimate the data error is the use of artificial neural networks (ANN) because they do not have a dependence on model structure (Bastidas *et al.*, 2003).

By using parameter estimation procedures, the overall error in the predicted fluxes can be reduced by as much as 20 to 40%, on average, for the heat fluxes at different locations throughout the world (e.g. Bastidas *et al.*, 1999, 2001, 2003; Demarty *et al.*, 2004; Demarty *et al.*, 2005; Gupta *et al.*, 1999; Hogue *et al.*, 2005; Leplastrier *et al.*, 2002; Rosolem *et al.*, 2005; Sen *et al.*, 2001; Yang *et al.*, 2005; Xia *et al.*, 2002). This finding is of importance for the comparison of different models. A fair comparison is possible only if the best solutions from different models are compared. Only the elimination of the improper parameter identification-induced error allows for the comparison of the model structural error.

As discussed, observations can be used to constrain the models, that is, to bound the parameter values so that the model outputs are consistent with the field observations. To attain this consistency, different parameter sets should be obtained for different environmental conditions and environments. Remote sensing has the potential to provide information about the space-time variations on the land surface processes. This is of particular relevance because this information can be used to parameterize land surface models and to derive estimates of the latent heat flux (see e.g. Bastiaanssen *et al.*, 1994; Franks and Beven, 1999; Kustas and Humes, 1996; Lakshmi, 2000; Pelgrum and Bastiaanssen, 1996; Wood and Lakshmi, 1993).

Liu *et al.* (2003) and Liu *et al.* (2004) have shown that proper parameter estimation is of paramount importance when a land surface model is coupled to an atmospheric model. In that case, parameter sensitivities and any deficiency associated with the parameter values are amplified through the coupling. Using the NCAR-LSM coupled to a single column atmospheric model, they found that a small, intentional, but incorrect parameterization for canopy evaporation produced wild oscillations in model output when the canopy parameters were changed beyond a very limited range (Liu *et al.*, 2003). The original parameterization was introduced to speed up the convergence of the solutions in an off-line mode, where these oscillations did not appear. Using the same model and a multiple objective sensitivity analysis procedure (several fluxes analyzed simultaneously), they established that parameter sensitivity changed significantly between coupled and off-line modes (Liu *et al.*, 2004). However, coupled modeling also imposes strong land surface forcing biases predicted by the atmospheric models on the land surface model. These biases in precipitation and radiation can overwhelm the behavior of land surface model physics (Dirmeyer, 2001).

INTO THE FUTURE

Workshops organized during the past few years by GEWEX, the European Centre for Medium-Range Weather Forecasts (ECMWF), the International Geosphere-Biosphere Programme (IGBP), and others have pointed out a number of directions in which innovation in the next generation of land surface models will be necessary. A number of these have also been identified by Pitman (2003) in his review of land surface models. Specifically, innovations in land surface models must focus on the following aspects:

1. representation of the full carbon cycle;
2. use of the catchment as the basic unit of land surface hydrological processes rather than an arbitrary grid cell;
3. changes in root distribution as a result of changes in CO_2 concentrations;
4. representation of heterogeneity and subgrid-scale processes, in other words, larger importance of the horizontal complexity of the surface;
5. increased recognition of the significance of chemistry in land surface processes and consequently its representation in land surface models;
6. use of new remotely sensed variables for data assimilation;
7. improved representations of feedbacks between surface and atmosphere.

To these we would like to add the following:

8. model structures to represent spatial variability at a scale appropriate for the dominant processes and required output – a multiple resolution framework;

9. new methods for a more realistic representation of spatial variability; either deterministic or statistical; either explicit or in the form of subgrid representations
10. improved methods for comparing observations and predictions of spatial response.

At the time of this writing (2005), the land surface modeling community has successfully met the challenges outlined by Richardson in 1922. However, as listed above, many challenges and research questions remain, particularly in the development of more robust parameterizations of surface processes, in the assessment of model sensitivities, and in the development of a better understanding of the relationship between uncoupled and coupled model simulations. These questions are not merely academic. Land surface models have come to play an important role in our attempts to evaluate and predict changes in a rapidly changing world. Better land surface models, both for coupled and uncoupled simulations can play their part in anticipating change and in making us a better steward of our planet.

REFERENCES

André J.C., Goutorbe J.P. and Perrier A. (1986) HAPEX-MOBILHY: A hydrologic atmospheric experiment for the study of water budget and evaporation flux at the climatic scale. *Bulletin of the American Meteorological Society*, **67**, 138–144.

Barrett A.P. (2003) *National Operational Hydrologic Remote Sensing Center SNOw Data Assimilation System (SNODAS) Products at NSIDC*, NSIDC Special Report, 11, National Snow and Ice Data Center, Boulder, Digital Media.

Bastiaanssen W.G.M., Hoekman D.H. and Roebling R.A. (1994) A methodology for the assessment of surface resistance and soil water storage variability at mesoscale based on remote sensing measurements, a case study with the HAPEX-EFEDA data. *IAHS Special Publications*, **2**, 66.

Bastidas L.A., Gupta H.V. and Sorooshian S. (2001) Bounding parameters of land surface schemes with observational data. In *Observations and Modeling of the Land Surface Hydrological Processes*, Lakshmi V., Albertson J. and Schaake J. (Eds.), Water Science and Application Volume 3, American Geophysical Union: pp. 65–76.

Bastidas L.A., Gupta H.V. and Sorooshian S. (2003) Parameter, Structure and Performance Evaluation for Land Surface Models. In *Advances in the Calibration of Watershed Models*, Duan Q., Gupta H.V., Sorooshian S., Rousseau A. and Turcotte R. (Eds.), AGU: pp. 229–238.

Bastidas L.A., Gupta H.V., Sorooshian S., Shuttleworth W.J. and Yang Z.-L. (1999) Sensitivity Analysis of a Land Surface Scheme using Multi-criteria Methods. *Journal of Geophysical Research*, **104**, 19481–19490.

Beck M.B. (2002a) Introduction. In *Environmental Foresight and Models: A Manifesto*, Beck M.B. (Ed.), Elsevier: pp. 3–9.

Beck M.B. (2002b) We have a problem. In *Environmental Foresight and Models: A Manifesto*, Beck M.B. (Ed.), Elsevier: pp. 11–33.

Beljaars A.C.M. and Bosveld F.C. (1997) Cabauw data for the validation of land surface parameterization schemes. *Journal of Climate*, **10**, 1172–1193.

Beljaars A.C.M. and Viterbo P. (1994) The sensitivity of winter evaporation to the formulation of aerodynamic resistance in the ECMWF model. *Boundary Layer Meteorology*, **71**, 135–149.

Beljaars A.C.M. and Viterbo P. (1998) Soil moisture-precipitation interaction: experience with two land surface schemes in the ECMWF model. In *Global Water and Energy Cycles*, Browning K.A. and Gurney R.J. (Eds.), Cambridge University Press, Cambridge, United Kingdom, pp. 223–235.

Betts A.K., Ball J.H. and Beljaars A.C.M. (1993) Comparison between the land surface response of the ECMWF model and the FIFE-1987 data. *Quarterly Journal of the Royal Meteorological Society*, **119**, 975–1001.

Betts A.K., Ball J.H., Beljaars A.C.M., Miller M.J. and Viterbo P.A. (1996) The land surface-atmosphere interaction: A review based on observational and global modeling perspectives. *Journal of Geophysical Research*, **101**, 7209–7225.

Beven K.J. (1995) Linking parameters across scales: subgrid parameterizations and scale dependent hydrological models. *Hydrological Processes*, **9**, 507–525.

Bhumralkar C.M. (1975) Numerical experiments on the computation of ground surface temperature in an atmospheric circulation model. *Journal of Applied Meteorology*, **14**, 1246–1258.

Bowling L.C., Lettenmaier D.P., Nijssen B., Graham L.P., Clark D.B., El Maayar M., Essery R., Goers S., Gusev Y.M., Habets F., *et al.* (2003) Simulation of high-latitude hydrological processes in the Torne–Kalix basin: PILPS Phase 2(e): 1. Experiment description and summary intercomparisons. *Global and Planetary Change*, **38**, 1–30.

Brewer K.E. and Wheatcraft S.W. (1994) Including multi-scale information in the characterization of hydraulic conductivity distributions. In *Wavelets in Geophysics*, Foufoula-Georgiou E. and Kumar P. (Eds.), Academic Press: San Diego, pp. 213–248.

Brown H.T. and Escombe F. (1900) Static diffusion of gases and liquids in relation to the assimilation of carbon and translocation in plants. *Philosophical Transactions of the Royal Society London, Series B*, **193**, 223–291.

Brown H.T. and Wilson W.E. (1905) On the thermal emmissivity of a green leaf in still and moving air. *Proceedings of the Royal Society of London, Series B Containing Papers of a Biological Character*, **76**, 122–137.

Brubaker K.L. and Entekhabi D. (1996) Analysis of feedback mechanisms in land-atmosphere interaction. *Water Resources Research*, **32**, 1343–1358.

Carroll T., Cline D., Fall G., Nilsson A., Li L. and Rost A. (2001) NOHRSC Operations and the simulation of snow cover properties for the coterminous U.S. *Proceedings of the 69th Annual Meeting of the Western Snow Conference*, Sun Valley, pp. 1–14.

Cherkauer K.A., Bowling L.C. and Lettenmaier D.P. (2003) Variable Infiltration Capacity (VIC) cold land process model updates. *Global and Planetary Change*, **38**, 151–159.

Crossley J., Polcher J., Cox P., Gedney N. and Planton S. (2000) Uncertainties linked to land-surface processes in climate change simulations. *Climate Dynamics*, **16**, 949–961.

Daley R. (1991) *Atmospheric Data Analysis*, Cambridge University Press: Cambridge, p. 457.

Deardorff J.W. (1978) Efficient prediction of ground surface temperature and moisture, with inclusion of a layer of vegetation. *Journal of Geophysical Research*, **83**, 1889–1903.

Demarty J., Ottlé C., Braud I., Frangi J.P., Bastidas L.A. and Gupta H.V. (2004) Using a multi-objective sensitivity analysis to calibrate the SiSPAT-RS Model. *Journal of Hydrology*, **287**, 214–236.

Demarty J., Ottlé C., Braud I., Olioso A., Frangi J.P., Gupta H.V. and Bastidas L.A. (2005) Constraining a physically based SVAT model with surface water content and thermal infrared brightness temperature measurements using a multiobjective approach. *Water Resources Research*, **V41**(1), W01011, doi:10.1029/2004WRR003695.

Dickinson R.E., Henderson-Sellers A., Kennedy P.J. and Wilson M.F. (1986) *Biosphere Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model*, NCAR Technical Note, NCAR, TN275 + STR, p. 69.

Dirmeyer P.A. (2001) Climate drift in a coupled land-atmosphere model. *Journal of Hydrometeorology*, **2**, 89–100.

Dirmeyer P.A., Dolman A.J. and Sato N. (1999) The pilot phase of the global soil wetness project. *Bulletin of the American Meteorological Society*, **80**, 851–878.

Doherty R. and Mearns L. (1999) *A Comparison of Simulations of Current Climate from Two Coupled Atmosphere-ocean Global Climate Models Against Observations and Evaluation of Their Future Climates*, Report to the National Institute for Global Environmental Change, Environmental and Societal Impacts Group/National Centers for Atmospheric Research, Boulder, Colorado.

Douville H., Viterbo P., Mahfouf J.-F. and Beljaars A.C.M. (2000) Evaluation of the optimum interpolation and nudging techniques for soil moisture analysis using FIFE data. *Monthly Weather Review*, **128**, 1733–1756.

Entin J.K., Robock A., Vinnikov K.Y., Zabelin V., Liu S., Namkhai A. and Adyasuren T. (1999) Evaluation of soil wetness project soil moisture simulations. *Journal of the Meteorological Society of Japan*, **77**, 183–198.

Franks S.W. and Beven K.J. (1999) Conditioning a multiple-patch SVAT model using uncertain time-space estimates of latent heat fluxes as inferred from remotely sensed data. *Water Resources Research*, **35**, 2751–2761.

Garratt J.R. (1993) Sensitivity of climate simulations to land-surface and atmospheric boundary layer treatments – a review. *Journal of Climate*, **6**, 419–448.

Gupta H.V., Bastidas L.A., Sorooshian S., Shuttleworth W.J. and Yang Z.-L. (1999) Parameter estimation of a land surface scheme using multi-criteria methods. *Journal of Geophysical Research*, **104**, 19491–19504.

Halldin S., Gottschalk L., van de Griend A.A., Gryning S.E., Heikinheimo M., Hogstrom U., Jochum A. and Lundin L.C. (1998) NOPEX – a northern hemisphere climate processes land surface experiment. *Journal of Hydrology*, **212/213**, 172–187.

Henderson-Sellers A., Pitman A., Love P., Irranejad P. and Chen T. (1995) The project for intercomparison of land-surface parameterization schemes (PILPS): Phases 2 and 3. *Bulletin of the American Meteorological Society*, **94**, 489–503.

Henderson-Sellers A., Yang Z.-L. and Dickinson R.E. (1993) The project for intercomparison of land-surface parameterization schemes. *Bulletin of the American Meteorological Society*, **74**, 1335–1349.

Hogue T.S., Bastidas L.A., Gupta H.V., Sorooshian S. and Mitchell K. (2005) Evaluation and transferability of the Noah land land surface model in semiarid environments. *Journal of Hydrometeorology*, **6**, 68–84.

Kalnay E. (2003) *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press: Cambridge.

Koster R.D. and Suarez M.J. (1995) Relative contributions of land and ocean processes to precipitation variability. *Journal of Geophysical Research*, **100**, 13775–13790.

Koster R.D., Suarez M.J., Higgins R.W. and van den Dool H.M. (2003) Observational evidence that soil moisture variations affect precipitation. *Geophysical Research Letters*, **30**, 45/1–45/4.

Koster R.D., Suarez M.J., Liu P., Jambor U., Berg A., Kistler M., Reichle R., Rodell M. and Famiglietti J.S. (2004) Realistic initialization of land surface states: Impacts on subseasonal forecast skill. *Journal of Hydrometeorology*, **5**(6), 1049–1063.

Kumar S.V., Peters-Lidard C.D., Tian Y., Houser P.R., Geiger J., Olden S., Lighty L., Eastman J.L., Doty B., Dirmeyer P., *et al.* (2004) Land Information System – An interoperable framework for high resolution land surface modeling. *Environmental Modelling and Software*, submitted.

Kustas W.P. and Humes K.S. (1996) Sensible heat flux from remotely sensed data at different resolutions. In *Scaling up in Hydrology Using Remote Sensing*, Stewart J.B., *et al.* (Eds.), John Wiley and Sons: New York, p. 255.

Lakshmi V. (2000) A simple surface temperature assimilation scheme for use in land surface models. *Water Resources Research*, **36**, 3687–3700.

Leplastrier M., Pitman A.J., Gupta H. and Xia Y. (2002) Exploring the relationship between complexity and performance in a land surface model using the multi-criteria method. *Journal of Geophysical Research*, **107**, 4443, doi: 10.1029/2001JD000931.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, **99**, 14415–14428.

Liu Y., Bastidas L.A., Gupta H.V. and Sorooshian S. (2003) Impacts of a parameterization deficiency on off-line and coupled land surface model simulations. *Journal of Hydrometeorology*, **4**, 901–914.

Liu Y., Gupta H.V., Sorooshian S., Bastidas L.A. and Shuttleworth W.J. (2004) Constraining land surface and atmospheric parameters of a locally coupled model using observational data. *Journal of Geophysical Research*, **109**, D21101, doi:10.1029/2004JD004730.

Lohmann D., Lettenmaier D.P., Liang X., Wood E.F., Boone A.A., Chang S., Chen F., Dai Y., Desborough C., Dickinson R.E., *et al.* (1998) The project for intercomparison of land-surface parameterization schemes (PILPS) phase 2(c) Red-Arkansas river basin experiment: 3. Spatial and temporal analysis of water fluxes. *Global and Planetary Change*, **20**, 161–179.

Manabe S. (1969) Climate and the ocean circulation: 1. The atmospheric circulation and the hydrology of the Earth's surface. *Monthly Weather Review*, **97**, 739–805.

Matheussen B., Kirschbaum R.L., Goodman I.A., O'Donnell G.M. and Lettenmaier D.P. (2000) Effects of land-cover change on streamflow in the interior Columbia river basin (USA and Canada). *Hydrological Processes*, **14**, 867–885.

Maurer E.P. and Lettenmaier D.P. (2003) Predictability of seasonal runoff in the Mississippi river basin. *Journal of Geophysical Research*, **108**(D16), 8607, doi:10.1029/2002JD002555.

Maurer E.P., Wood A.W., Adam J.C., Lettenmaier D.P. and Nijssen B. (2002) A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States. *Journal of Climate*, **15**, 3237–3251.

Mitchell K.E., Houser P.R., Wood E.F., Schaake J.C., Tarpley J.D., Lettenmaier D.P., Higgins R.W., Marshall C., Lohmann D., Ek M., *et al.* (1999) GCIP Land Data Assimilation System (LDAS) project now underway. *GEWEX News*, **9**, 3–6.

Mitchell K., Lin Y., Rogers E., Marshall C., Ek M., Lohmann D., Schaake J., Tarpley D., Grunmann P., Maninkin G., Duan Q. and Koren V. (2000) Recent GCIP-sponsored advancements in coupled land-surface modeling and data assimilation in the NCEP ETA mesoscale model, *AMS 15th Conference on Hydrology, 9–14 January*, Long Beach, pp 180–183.

Nijssen B., Bowling L.C., Lettenmaier D.P., Clark D.B., El Maayar M., Essery R., Goers S., Gusev Y.M., Habets F., van den Hurk B., *et al.* (2003) Simulation of high-latitude hydrological processes in the Torne–Kalix basin: PILPS phase 2(e): 2. Comparison with observations. *Global and Planetary Change*, **38**, 31–53.

Nijssen B., O'Donnell G.M., Hamlet A.F. and Lettenmaier D.P. (2001) Hydrologic sensitivity of global rivers to climate change. *Climatic Change*, **50**, 143–175.

Noilhan J., Mahfouf J., Manzi A. and Planton S. (1993) Validation of land-surface parameterizations: developments and experience at the French weather service. *Proceedings Seminar ECMWF, 7–11 September 1992*, ECMWF: Reading, pp. 125–158.

Pelgrum H. and Bastiaanssen W.G.M. (1996) An intercomparison of techniques to determine the area-averaged latent heat flux from individual in situ observations: A remote sensing approach using the European Field Experiment in a desertification-threatened area data. *Water Resources Research*, **32**, 2775–2786.

Pitman A.J. (2003) The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology*, **23**, 479–510, doi: 10.1002/joc.893.

Richards A.L. (1931) Capillary conduction of liquids through porous media. *Physics*, **1**, 316–333.

Richardson L.F. (1922) *Weather prediction by numerical process*, Cambridge University Press: Cambridge, p. 236.

Rodell M., Houser P.R., Jambor U., Gottschalck J., Mitchell K., Meng C.-J., Arsenault K., Cosgrove B., Radakovich J., Bosilovich M., *et al.* (2004) The global land data assimilation system. *Bulletin of the American Meteorological Society*, **85**, 381–394.

Rosolem R., Bastidas L.A., Shuttleworth W.J. and Goncalvez L.G. (2005) *Evaluation of the Effect of Selective Logging on the Energy-water and Carbon Exchange Processes and Parameterizations for Tropical Forests in the SiB2 Model*, IUGG Meeting, Foz de Iguacu.

Sellers P.J., Dickinson R.E., Randall D.A., Betts A.K., Hall F.G., Berry J.A., Collatz G.J., Denning A.S., Mooney H.A., Nobre C.A., *et al.* (1997a) Modelling the exchanges of energy, water and carbon between continents and the atmosphere. *Science*, **275**, 502–509.

Sellers P.J., Hall F.G., Asrar G., Strebel D.E. and Murphy R.E. (1988) The First ISLSCP Field Experiment (FIFE). *Bulletin of the American Meteorological Society*, **69**, 22–27.

Sellers P.J., Hall F.G., Asrar F., Strebel D.E. and Murphy R.E. (1992) An overview of the first international satellite land surface climatology project (ISLSCP) field experiment (FIFE). *Journal of Geophysical Research*, **97**, 18345–18371.

Sellers P.J., Hall F.G., Kelly B., Baldocchi D., Black A., Berry J., Ryan M., Ranson K.J., Crill P.M. and Lettenmaier D.P. *et al.* (1997b) BOREAS in 1997: experiment overview, scientific results, and future directions. *Journal Geophysical Research, Atmospheres*, **102**, 28731–28769.

Sellers P.J., Mintz Y., Sud Y.C. and Dalcher A. (1986) A Simple Biosphere Model (SIB) for use within general circulation models. *Journal of the Atmospheric Sciences*, **43**, 505–531.

Sellers P.J., Shuttleworth W.J., Dorman J.L., Dalcher A. and Roberts J.M. (1989) Calibrating the simple biosphere model for Amazonian tropical forest using field and remote sensing data. Part I: Average calibration with field data. *Journal of Applied Meteorology*, **28**, 727–759.

Sen O.L., Bastidas L.A., Shuttleworth W.J., Yang Z.-L., Gupta H.V. and Sorooshian S. (2001) Impact of field calibrated vegetation parameters on GCM climate simulations. *Quarterly Journal of the Royal Meteorological Society, Part B*, **127**, 1199–1224.

Shao Y. and Henderson-Sellers A. (1996) Modelling soil moisture: a project for intercomparison of land surface parameterization schemes phase 2(b). *Journal of Geophysical Research*, **101**, 7227–7250.

Shuttleworth W.J., Gash J.H.C., Lloyd C.R., Moore C.J., Roberts J., *et al.* (1984) Eddy correlation measurements in Amazon forest. *Quarterly Journal of the Royal Meteorological Society*, **110**, 1143–1162.

Sorooshian S., Bastidas L.A. and Gupta H.V. (1999) Application of multi-objective optimization algorithms for hydrologic model identification and parameterization, 2nd *International Conference on Multiple Objective Decision Support Systems for Land, Water, and Environmental Management, August 1–5 1999*, Brisbane.

Tian Y., Peters-Lidard C.D., Kumar S., Geiger J., Houser P.R., Eastman J.L., Dirmeyer P., Doty B. and Adams J. (2004) High Performance Land Surface Modeling with a Beowulf Cluster. *Computing in Science and Engineering*, submitted.

Trenbert K.E. (1992) *Climate system modeling*, Cambridge University Press: Cambridge, p. 788.

Viterbo P. and Beljaars A.C.M. (1995) An improved land surface parameterization scheme in the ECMWF model and its validation. *Journal of Climate*, **8**, 2716–2748.

Wood E.F. and Lakshmi V. (1993) Scaling water and energy fluxes in climate systems: three land-atmosphere modeling experiments. *Journal of Climate*, **6**, 839–857.

Wood A.W., Leung L.R., Sridhar V. and Lettenmaier D.P. (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216.

Wood A.W., Maurer E.P., Kumar A. and Lettenmaier D.P. (2002) Long range experimental hydrologic forecasting for the Eastern U.S.. *Journal of Geophysical Research*, **10**, 4429, doi:10.1029/2001JD000659.

Xia Y., Pitman A.J., Gupta H.V., Leplastrier M., Henderson-Sellers A. and Bastidas L.A. (2002) Calibrating a land surface model of varying complexity using multi-criteria methods and the Cabauw data set. *Journal of Hydrometeorology*, **3**, 181–194.

Yang K., Koike T., Ye B. and Bastidas L.A. (2005) Inverse analysis of the role of soil vertical heterogeneity in controlling surface soil state and energy partition. *Journal of Geophysical Research*, **110**(D8), D08101, doi:10.1029/2004JD005227.

202: Use of Climate Information in Water Resources Management

HOLLY C HARTMANN

Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, US

Failures in water resources management, evolving societal objectives, and advances in the scientific understanding of hydroclimatic systems have led to reassessment of techniques traditionally used in support of water resources management policies and practices. Climate information, encompassing a variety of timescales, is clearly useful for water resources management, yet remains underutilized. This article examines historical climate information provided by instrumental records and paleoclimatological indicators, and prognostications of future climate conditions provided by climate forecasts and assessments of climate change, and issues related to their use in water resources management. Two issues are of central importance: nonstationarity and uncertainty. The water resources community must accept that in many cases one of the fundamental assumptions underlying classical hydrology is simply incorrect. However, the water resources management community need not assess climate information and its use in isolation. Networks of researchers, including hydroclimatic and social scientists, are actively seeking more integrated involvement with the water resources management community, are adapting their research agenda and activities to address the needs of decision makers and stakeholders, and are committed to developing practical products and decision-support tools.

INTRODUCTION

Water resource managers have long been incorporating information about climate variability in their decisions. The tremendous, regionally ubiquitous investments in infrastructure to reduce flooding (e.g. levees, reservoirs) or assure reliable water supplies (e.g. reservoirs, groundwater development, irrigation systems, water allocation, and transfer agreements) reflect societal goals to mitigate the impacts of climate variability at multiple time and space scales. Myriad engineering and hydrology manuals (e.g. Linsley *et al.*, 1975; Water Resources Council, 1981; Maidment, 1993) are replete with techniques for estimating expected climate and hydrologic conditions (jointly referred to as hydroclimatic conditions) and their variability.

However, accumulated experience, evolving societal objectives, and advances in the scientific understanding of hydroclimatic systems have complicated the use of traditional techniques in support of water resources management. Many communities have faced multiple events earlier thought to have low probabilities of occurrence (e.g. National Research Council NRC, 1995). Extreme conditions (e.g. floods, droughts) are by definition rare events and their likelihood of recurrence is difficult to judge with data collected over only the few decades often available (Stedinger *et al.*, 1993). Also, as instrumental records grow in length, long-term shifts in streamflows can be observed (Lettenmaier *et al.*, 1994; Lins and Slack, 1999; Douglas *et al.*, 2000) leading to questions about the relative impacts of shifts in climate conditions, land use, and river hydraulics.

Further, the growing financial, political, social, and environmental costs of infrastructure options have shifted the focus of large water management institutions to optimizing operations of existing projects (Bureau of Reclamation, 1992; Beard, 1993; Stakhiv, 2003). In many cases, this includes improving project returns (both economic and noneconomic) outside the range of conditions considered in original procedures (e.g. optimizing returns under average conditions in addition to reducing damages during extreme events). Further, increasingly diverse and often conflicting demands prompt searches for additional potential trade-offs

Encyclopedia of Hydrological Sciences. Edited by M G Anderson.

among interests by considering a broader range of hydroclimatic conditions (e.g. Congressional Budget Office, 1997; Pulwarty and Melis, 2000).

Finally, the evolving understanding of earth dynamics has changed perspectives about potential climate variability. Extremely long time series of paleoclimatological indicators (e.g. *see* **Chapter 199, Role and Importance of Paleohydrology in the Study of Climate Change and Variability, Volume 5**) have made clear that climate and water supplies in many regions are more variable than indicated by instrumental records alone, with periods of extreme drought or wetness lasting from several years to several decades, albeit often interrupted by more typical conditions. Also, climate is now recognized as a chaotic process shifting among distinct regimes with statistically significant differences in average conditions and variability (Hansen *et al.*, 1997). Further, myriad studies related to global warming are becoming more confident in their conclusions that the future portends statistically significant changes in hydroclimatic averages and variability.

Until the last decade, climate was viewed largely as a collection of random processes. This paradigm informs much of the water resource management practices developed over the past 50 years, which persist today. However, climate variability and long-term change are now considered by the scientific community to result from phenomena that can be understood at a fundamental level and that, in some cases, can be predicted with long lead-times. This revolutionary paradigm shift calls into question water management practices developed under assumptions that random chance and time-invariant probability distributions are key drivers of variability.

The objective of this article is to examine several categories of climate information and issues related to their use in water resources management. Because water resources management takes place in a complex milieu in which climate is but one, and perhaps not the most important factor, this article also examines recent efforts to more effectively integrate climate research and water resources management. Finally, this article offers recommendations for the water resources management community in their attempts to improve their use of advanced climate information.

CONSIDERING CLIMATE

Traditionally, instrumental observations of temperature, precipitation, and streamflow, among other variables, measured over years to decades comprised the bulk of climate information used in support of water resource management decisions. Climate information has evolved to encompass a broad range of timescales, as illustrated in Figure 1, each having its own conceptual and analytical approaches. Historical climate and paleoclimatological information rely on direct observations or proxies to describe how conditions have varied in the past. Climate forecasts and climate-change information address a future, impossible to know with certainty before it occurs.

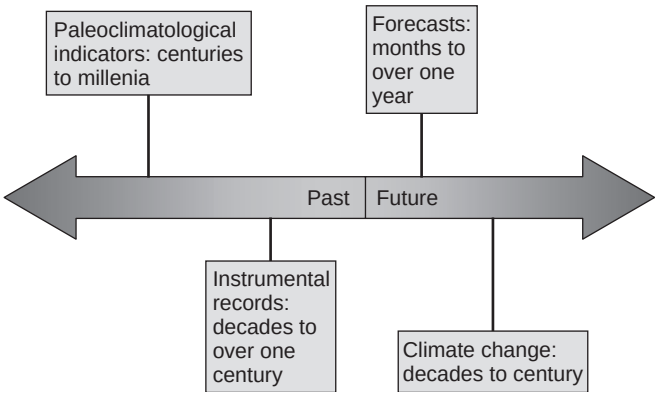


Figure 1 Climate information encompasses paleoclimatological indicators, instrumental records, climate forecasts, and climate change scenarios

Historical Climate Variability

Instrumental Records

The early design and operation of water management systems were based largely on trial and error, for example, repeated reconstructions of levees along the Illinois River (Thompson, 2002) and rebuilding of shipping harbors (Shallot, 2003). The recurring need to "fix on failure" and the increasing complexity of water management systems led to the installation of instruments for gauging streamflows and the development of analytical techniques for evaluating water supply reliability and hydrologic risk. Lettenmaier (2003) tracks the concurrent evolution of methods to analyze historical time-series and approaches for water resource management from the perspective of management capacity to incorporate climate variability. Standard practice has been to embed climate variability through analytical techniques that use observed hydrologic time series such as variants of mass curve analysis to size reservoir systems and frequency analysis to assess hydrologic risk (Hirsch *et al.*, 1993; Salas, 1993; Stedinger *et al.*, 1993). Direct use of climate information includes hydrologic and hydraulic routing of "design storms" of various magnitudes and likelihoods and the use of rainfall-runoff models in which meteorological variables drive watershed response (Lettenmaier and Wood, 1993; Urbanas and Roesner, 1993).

Notwithstanding past successes, there exist serious challenges in the use of instrumental records in water resources management. First, instrumental records have often been too short to adequately express climate variability and resulting impacts. A classic example concerns the allocation of Colorado River flows in 1922 (Rhodes *et al.*, 1984;

Hobbs, 2004). Understandably, few instrumental observations were available to reliably estimate water supply availability; climate variability was accommodated by the use of a 10-year average to monitor compliance with flow obligations. Observations over the ensuing decades have revealed flows prior to 1922 to be unusually high with unusually low variability, leading to overestimation of Colorado River flows and their reliability, over allocation of available water supplies, and costly interstate conflict. Another clear example concerns the American River in California, where 10 of the largest 13 floods since 1905 have occurred only since 1950 (NRC, 1995).

Second, as instrumental records have grown longer, they increasingly belie one of the fundamental assumptions behind most extant water resources management – stationarity. Stationary time series have time-invariant statistical characteristics (e.g. mean, variance), meaning that different parts of the historical record can be considered equally likely. Further, within the limits posed by sampling, statistics computed from stationary time series can be used to define a probability distribution that will also then faithfully represent expectations for the future (Salas, 1993). Yet, long climate and hydrology time series often show trends (e.g. Baldwin and Lall, 1999; Olsen *et al.*, 1999) or persistent regimes, that is, periods characterized by distinctly different statistics (e.g. Angel and Huff, 1995; Quinn, 1981, 2002) with consequences for the estimation of hydrologic risk (Olsen *et al.*, 1998). Observed regimes and trends can have multiple causes, including climatic changes, watershed and river transformations, and management impacts (e.g. irrigation return flows, trans-basin water diversions).

Finally, instrumental records occur at timescales directly experienced by individuals and embodied in institutional memory. Experience may provide cognitive advantages for perceiving the connections between climate variability, hydrologic impacts, and water resources management. However, the risk in water management is that direct experience associated with specific events (i.e. the conditions and success or failures of actions taken during those periods) can "anchor" perceptions during subsequent events, precluding full appreciation of variability of climate patterns, effects on hydrology, and implications for water management. For example, the extremely high flows leading to the placement of plywood extensions on the massive Glen Canyon Dam and fears for the structure's integrity (Rhodes *et al.*, 1984) are firmly fixed in the memory of hydrologists and water managers familiar with the episode and affect both water supply forecasts and management decisions (Pagano *et al.*, 2004).

Multiple techniques exist to more effectively exploit instrumental records and to more accurately represent potential hydroclimatic variability and uncertainty. Walker and Douglas (2003) review statistical techniques for detecting and accommodating different types of nonstationarity (e.g. estimating nonstationary hydrologic risk) in water management applications. For example, for streamflows showing interannual persistence, computing hydrologic risk using concepts of "time to first occurrence" rather than "time between occurrences" can reduce the risk of failure over a system's design life (Douglas *et al.*, 2002). Stochastic hydrology techniques use various forms of autoregressive models to generate multiple synthetic streamflow time series with statistical characteristics matching available observations. Early stochastic hydrology work challenged the assumption of stationarity (Mandelbrot and Wallis, 1969; Klemes, 1974; Potter, 1976), but ultimately had little impact on water management practices (Lettenmaier, 2003). Exceptions can be found, however. In estimating the risk of low flows for the Sacramento River Basin in California, the Bureau of Reclamation (Frevert *et al.*, 1989) generated twenty 1000-year streamflow time series matching selected statistics of observed flows (adjusted to compensate for water management impacts on natural flows); the nonexceedance probabilities of low flows were computed by counting the occurrences of low flows within 1- through 10-year intervals for all twenty 1000-year sequences. The U.S. Army Corps of Engineers (1992) uses a similar approach to estimate flood magnitudes with return periods exceeding 1000 years, using Monte Carlo sampling from within the 95% confidence limits of a Log Pearson III distribution developed by synthesizing multiple streamflow time series. Several practical statistical approaches were demonstrated using the long instrumental record for Lake Erie water level records (Herche and Hartmann, 1992; Kite, 1992; Kubik, 1992; Potter, 1992; Privalsky, 1992).

Conceptual rainfall-runoff models offer some advantages over statistical techniques in using instrumental climate records; these models represent, with varying levels of complexity, the transformation of rainfall and other meteorological forcing variables (e.g. air temperature, humidities) to watershed runoff and streamflow, including accounting for hydrologic storage conditions (e.g. snowpack water storage, soil moisture, groundwater storage). Recent models use spatially distributed representations of local topography, soil and land-cover characteristics, vegetation dynamics, and the surface energy balance (Wigmosta *et al.*, 1994; Ivanov *et al.*, 2004). These models can be used to assess the impacts and implications of various climate scenarios, by using historic meteorological time series as input, generating hydrologic time series, and then using those hydrologic scenarios as input to hydraulic routing and water management models. This approach enables consideration of current landscape and river channel conditions, which may be quite different from that embodied in early instrumental records, and which can dramatically alter a watershed's hydrologic behavior (Vorosmarty *et al.*, 2004). The use of multiple input time series or system parameterizations enables a probabilistic assessment of an ensemble

of scenarios. For example, Carpenter and Georgakakos (2001) and Yao and Georgakakos (2001) linked an ensemble streamflow simulation model and representations of the hydraulic system to assess the capabilities of extant and alternative reservoir operation schemes for the Folsom River, California.

Paleoclimatological Indicators

As described by Ekwurzel (**Chapter 199, Role and Importance of Paleohydrology in the Study of Climate Change and Variability, Volume 5**), many paleoclimatological techniques have shed light on climate conditions prior to instrumental observations. Glacier ice, groundwater chemistry, cave and coral formations, lakebed sediments, and tree-rings, among others, have each been used to infer climate conditions occurring hundreds and even thousands of years ago. Paleoclimatological information can put modern extreme climate events in a broader perspective; studies of paleoclimatological indicators have identified flooding events far exceeding any measured in modern times and the recurrence of severe droughts over large regions and lasting many years. Paleoclimatological studies have also helped to establish teleconnections between climatic conditions across oceans and continents and have led to the identification of oceanic and atmospheric patterns favoring persistent drought.

There are several overarching messages for water management professionals provided by paleoclimatological information. First, past climate conditions in a region may have been more extreme and persistent than indicated by historical records, suggesting that similarly severe conditions can recur; analyses based only on instrumental records underestimate hydrologic risk. For example, in reconstructions of US Southwest cool-season precipitation (the region's strongest determinant of annual streamflows), Ni *et al.* (2002) found five 5-year periods drier than the severe drought of 1999–2003 in northern Arizona, when precipitation was only 69% of the average over the years 1000–1988. The most extreme five-year period (1666–1670) experienced only 47% of the millennial average. Using slightly different methods, Salzer and Kipfmueller (in press) found the late 1500s to be the most severe drought period in that same region, on timescales from single-year drought (1584) to 10-year drought (1583–1592). However, the sixteenth-century droughts appear mild compared to multiple persistent droughts reflected in tree-ring records exceeding 2000 years (Grissino-Mayer, 1996).

Second, in some cases, the same climate processes and forcing conditions leading to paleoclimatological extremes continue to affect climate variability today, reinforcing prospects that similar extremes may recur. As presented in more detail by Ekwurzel (*see* **Chapter 199, Role and Importance of Paleohydrology in the Study of Climate Change and Variability, Volume 5**), links have been established between the El Nino/Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO) and tree-ring reconstructions of the Palmer Drought Severity Index (PDSI), among others (Cole *et al.*, 2002; Gray *et al.*, 2003), although the characteristics of some patterns differ (e.g. longer La Nina events in earlier centuries compared to those of the twentieth century) and issues remain regarding independent evaluation of reconstructed variables (i.e. tropical sea surface temperatures, continental precipitation and temperature).

Finally, climate nonstationarity is the rule, not the exception. The normal behavior of regional climate consists of transitioning among persistent regimes having significantly different statistical characteristics. Further, these regimes can differ in several ways, including their central tendency, variance, persistence, and intra-annual seasonality. There is no "normal" climate to which any deviations must return.

To date, the use of paleoclimatological information in water management remains more prospect than practice. While paleoclimatological indicators provide a longer perspective about past climate regimes and extremes, they have less spatiotemporal specificity, less precision, and greater estimation uncertainty than the direct observations comprising the instrumental hydroclimatic record. Additionally, from a cognitive perspective, paleoclimatological information is far removed from personal experience or institutional memory, posing difficulties for individuals to comprehend within their existing paradigms. For example, in a workshop with consulting engineers, a 300-year (1700–1999) paleoclimatologic reconstruction of the PDSI was used to explore how concepts of climate nonstationarity were or could be incorporated into civil engineering design (Hartmann *et al.*, in review). The entire group was dismissive of the relevance of paleoclimatological data for water management, due to their confidence in the stationarity assumption and, in a seeming contradiction, skepticism about the similarity between past and present hydroclimatic teleconnections. Further, water managers are typically focused on specific planning horizons (e.g. a 20-year infrastructure development plan). Even when interested in paleoclimatological information, they have difficulty linking general long-term (e.g. over 1000 years) risks with specific risks defined within their planning horizon. For example, should they use statistics derived from (i) the entire paleoclimatological record to reflect the maximal potential variability; (ii) the most recent decades to reflect that, with evidence of recent strong trends, the near term may be a better predictor than the distant past; or (iii) specific periods from the distant past to reflect that secular patterns may be more likely to dominate upcoming decades?

However, in the US West, as the region's severe drought (beginning in the mid- to late-1990s) has evolved, so has the water resources sector's interest in considering paleoclimatological information. Individuals and organizations

are actively exploring the relevance of paleoclimatological information for water resources issues, including requesting paleoclimatological information and background briefings from scientists (G. Garfin, University of Arizona, personal communication, 2004). Whether the information and its implications for water resources management will be integrated into the sector's cognitive framework (e.g. acknowledging that design of long-lived infrastructure or policies must accommodate nonstationarity) remains to be seen. Fully integrated use of paleoclimatological information will require the development of prospective scenarios, sensitivity studies of infrastructure design and operating plans (e.g. regulation rules), and water resources management policies and practices that accommodate the increased variability posited by the paleoclimatological record. These efforts require scientific research support through development of paleoclimatological time series at higher spatial and temporal resolution (e.g. Ni *et al.*, 2002), establishing that the climatic patterns and teleconnections leading to specific paleoclimatological conditions are still operative, development of methods to identify the time periods and statistical characteristics of paleoclimatological information that have relevance for specific planning horizons, and further identification of the implications for water supplies of long-term climate variability and nonstationarity reflected in paleoclimatological information.

Climate Forecasts

Virtually all water management decisions require some sort of climate forecast. Even the consideration of historical information to assess hydrologic risk or estimate typical conditions represents an expectation about possibilities for future conditions. Seasonal climate forecasts have long promised the potential to improve water management responses to the vagaries of climate variability (Changnon and Vonnhame, 1986), but have often played only a marginal role in real-world decision making (Changnon, 1990; Sonka *et al.*, 1992; Pulwarty and Redmond, 1997; Callahan *et al.*, 1999; Pulwarty and Melis, 2000; Pagano *et al.*, 2001, 2002; Hartmann *et al.*, in review). However, recent advances in modeling and predictive capabilities continue to motivate the use of climate forecasts. In the United States, the Pacific Northwest, California, and the Southwest are strong candidates for the use of long-lead forecasts because ENSO and PDO signals are particularly strong in these regions and each region's water supplies are closely tied to accumulation of winter snowfall, amplifying the impacts of climatic variability (Redmond and Koch, 1991). Climate forecasts in these regions show substantial skill, even at lead times approaching 8 months for the winter season, especially for temperature (Hartmann *et al.*, 2002a).

A variety of seasonal climate forecast products are available (Hartmann *et al.*, 2002b); they are typically generated by combining results, often subjectively but sometimes objectively, from a variety of techniques (including statistical and conceptual modeling) and sometimes making adjustments based on current observations, recent research, and expert judgment (Goddard *et al.*, 2001; Hartmann *et al.*, 2002b). The official US products are issued by the National Weather Service (NWS) Climate Prediction Center (CPC). Climate forecasts cover varying periods and are issued with different lead times. For example, each month, the CPC issues a suite of forecasts, including one 1-month outlook and a series of 13 3-month outlooks each offset by 1 month. Climate forecasts are typically issued in probabilistic terms and in reference to the observed frequency distribution of a specific historical period. For example, the CPC seasonal climate forecast maps show the likelihood of occurrence for average air temperature or total precipitation over the period to fall within tercile categories defined by the upper, middle, or lower third of conditions reflected in the historical record from 1971 to 2000. An alternative product expresses the climate outlooks using probability of exceedance curves, using different rules shifting probabilities among the tercile categories (Barnston *et al.*, 2000). The probability forecasts require additional information to place their information in context. For example, the CPC provides maps of temperature and precipitation probabilities, maps and tables of historical probability class limits, a text discussion of the rationale behind the forecast, and skill maps for some of the analytical techniques used in producing the forecast. Finally, it is important to note that seasonal climate forecasts should be considered "forecasts of opportunity", that is, possible only when specific climate patterns (e.g. ENSO) exist. Often, no techniques have shown skill for some regions, seasons, and lead times, so forecasters decline to issue an actual prediction, substituting a statement intended to convey complete uncertainty, although using confusing terminology (e.g. "climatology", "equal probability") (Hartmann *et al.*, 2002a).

Water management processes (e.g. reservoir regulation rules) typically require estimates of hydrologic conditions (e.g. monthly or seasonal streamflow volumes), rather than outlooks of seasonal precipitation or temperature. Researchers and operational hydrologic forecasters have long sought to incorporate climate forecasts into water supply forecasts (Pagano and Garen, 2003). The Columbia Basin Interagency Committee (CBIAC) initially considered the use of climate outlooks shortly after the NWS began issuing 30-day climate outlooks in 1953 (CBIAC, 1955), but concluded at that time and in later assessments (CBIAC, 1964) that forecast skill was too low for practical use, the risks of very poor forecasts were too great compared to the use of climatological average conditions for streamflow forecasting, and the uncertainties introduced by downscaling regional climate outlooks to the scale of small basins were too great. Instead, hydrologic forecasts have long been based on regression approaches (Hartmann *et al.*, 2002b)

because the relationship between snowpack and subsequent streamflow is so strong that hydrologists have often considered it to be deterministic (Pagano and Garen, 2003); in the infrequent cases where climate information has been integrated into hydrologic forecasts, climate indices (e.g. the intensity of the Southern Oscillation index [SOI]) have been used as regression predictor variables (Hartmann *et al.*, 2002b; Pagano and Garen, 2003).

Further, because the relationships between climate variables (temperatures and precipitation) and streamflow show greater variance, they are more appropriately considered in probabilistic terms. Hydrologic forecasters have been reluctant to issue probabilistic forecasts of water supplies and streamflows, considering them to be too vague to be useful, especially where reservoir operating rules require single streamflow values. However, that trend is changing, albeit slowly (Hartmann *et al.*, 2003); the NWS is committed to providing hydrologic forecasts in probabilistic terms (NWS, 1999), because they enable quantitative estimation of inevitable uncertainties associated with inherently chaotic weather and climate systems. From a decision-making perspective, probabilistic forecasts are more informative because they explicitly communicate uncertainty and are more useful because they can be directly incorporated into risk-based calculations (e.g. expected consequences).

Increasingly, hydrologic forecasters have turned to conceptual rainfall-runoff models to create hydrologic forecast ensembles using several meteorological sequences to produce several hydrologic predictions, keeping constant the hydrologic model's initial conditions (e.g. snowpack or soil moisture). Termed *ensemble streamflow prediction* (ESP), early implementations (Day, 1985) considered each ensemble member (i.e. trace) independent and equally likely; however, persistent climate patterns leading to nonstationarity mean that some meteorological sequences are more or less likely during certain periods. Hamlet and Lettenmaier (1999a) implemented an ESP system for the US Pacific Northwest using conditional sampling, reflecting the conditional hydrologic response of the Columbia River Basin to the ENSO state, which was further conditioned by epochal PDO forecasts. They used historic meteorological data, gridded at a one-degree spatial resolution, to drive the spatially distributed macroscale hydrology model (Liang *et al.*, 1994; Cherkauer and Lettenmaier, 2002) over the entire Columbia River Basin. Ensemble traces, consisting of precipitation and air temperature time series organized by water years (October–September), were associated with one of six predefined PDO/ENSO categories in a 50-year historical record. Given an anticipated PDO/ENSO climate signal for the coming water year, resampled meteorological traces were then used as "forecasts" to drive the VIC model on the basis of the initial soil and snow conditions as of the forecast date. Use of PDO/ENSO categories obviates the need for predicting ENSO intensity with great accuracy, an element of ENSO forecasts currently problematic at long-lead times.

Alternatively, meteorological sequences can be selected on the basis of consistency with weather and climate forecasts. Several researchers (Perica, 1998; Perica *et al.*, 1999; Smith *et al.*, 1992; Croley and Lee, 1993; Croley, 1996, 1997) have developed procedures for efficiently considering multiple meteorological forecasts by restructuring the set of possible future scenarios. Croley (1997) uses all possible meteorological sequences as hydrologic model input, then biases the resulting hydrologic ensemble through differential weighting of the traces to match the meteorological forecast probabilities. Croley (2000) provides software for combining meteorological forecasts that specify multiple event probabilities (e.g. with variable probabilities for several categories as in the NWS seasonal climate outlooks) and most-probable event outlooks (e.g. seasonal climate outlooks issued by the Canadian Meteorological Center). The South Florida Water Management District uses Croley's approach for weighting ESP traces, according to climate outlook probabilities, in their water supply forecasts (Cadavid *et al.*, 1999). Other methodological enhancements have been developed and implementation issues remain (Hartmann *et al.*, 2003), but ESP hydrologic forecasts provide information not available from regression approaches, even at their earliest lead times (Franz *et al.*, 2003).

Incorporation of probabilistic climate outlooks into regression-based water supply outlooks (e.g. Modini, 2000) have proved more operationally demanding yet no more accurate than simply using climate indices as predictor variables within regression equations (Pagano and Garen, 2003). However, the poor performance of statistical streamflow forecasts in unusually warm years (e.g. water year 2004 in the US West) highlights one of the critical advantages of forecasting systems based on conceptual rainfall-runoff models: their robustness to changing climate regimes, for example, the increasingly warm winters in the US West. This situation also highlights that even operational practices can be affected by nonstationarity, as forecast system performance degrades under climatic transitions.

Realistically, climate outlooks and advancements are typically initially incorporated into water management decisions informally, using subjective, *ad hoc* procedures on the initiative of individual water managers (Changnon, 2000; Rayner *et al.*, 2001; Pagano *et al.*, 2002). While improvised, those decisions are not necessarily insignificant. For example, the Salt River Project, among the largest water management agencies in the Colorado River Basin and primary supplier to the Phoenix metropolitan area, decided in August 1997 to substitute groundwater withdrawals with reservoir releases, expecting increased surface runoff during a wet winter related to El Nino. With that decision, they

risked losses exceeding $4 million in an attempt to realize benefits of $1 million (Pagano *et al.*, 2002). However, because these informal processes are based in part on confidence in the predictions, overconfidence in forecasts can be even more problematic than lack of confidence, as a single incorrect forecast that provokes costly shifts in operations can devastate user confidence in subsequent forecasts (e.g. Glantz, 1982). Single forecasts, even if not used, can reinforce misperceptions of generally poor climate forecast capabilities; in the winter of 2000/01, climate outlooks for the Pacific Northwest indicated a low chance for dry conditions, yet the region experienced the driest year on record; the water management community felt fortunate that they had not considered the climate forecasts (Pagano and Garen, 2003).

Ongoing relationships among researchers, operational forecasters, and decision makers are essential to ensure appropriate interpretation and application of evolving forecast products. Considering the diversity of water management circumstances, the continually evolving nature of predictive capabilities, the variety of forecast performance requirements and criteria, and workloads facing the research and operational forecasting communities, personalized support for individual water managers on a broad scale is impractical. However, dynamic decision-support tools implemented using Internet technology and accessing diverse data sources enable individual decision makers to perform their own data exploration and analyses, customized to reflect their specific situations and concerns with sufficient expert guidance to transform data and information into practical knowledge (e.g. **`http://fet.hwr.arizona.edu/ForecastEvaluationTool/`**). Experience demonstrates that a decision-support tool focused on forecast evaluation can help decision makers in their strategic thinking about essential forecast attributes; requisite performance thresholds; relationships among forecast quality, utility, and value; and the potential utility and value of forecast improvements (Hartmann *et al.*, 2004).

Alternatively, end-to-end decision-support tools that embody unique resource management circumstances enable formal, and more objective, linkages between climatological, hydrologic, and institutional processes. Typically, these end-to-end tools are developed for organizations making decisions with high impact (e.g. state or national agencies) or high economic value (e.g. hydropower production), and which possess the technical and managerial abilities to efficiently exploit research advances. A rapidly growing literature (Scott *et al.*, 2000; Hamlet and Lettenmaier, 2000; Hamlet *et al.*, 2002; Carpenter and Georgakakos, 2001; Yao and Georgakakos, 2001; Georgakakos, 2002a; Clark *et al.*, 2002) describes end-to-end studies that evaluate the implications of using hydroclimatic forecasts in the operation of large river systems. Used in design studies (Lee, 1999; Davis and Pangburn, 1999), they offer a way forward for decision makers reluctant to shift from statistical analysis of historical data based only on theoretical advantages.

Climate Change

Prospects for climate change due to global warming have moved from the realm of speculation to general acceptance (Intergovernmental Panel on Climate Change (IPCC), 1990; IPCC, 1995a, IPCC, 2001a). The potential impacts on water resources and their implications for management have been central topics of concern in climate-change assessments (e.g. Environmental Protection Agency, 1989; IPCC, 1995b, 2001b; National Assessment Synthesis Team, 2000; Gleick and Adams, 2000; Barnett *et al.*, 2004). The near-universal analytical approach has been one of sensitivity analysis (Lettenmaier, 2003):

1. downscaling outputs from a dynamic general circulation model of global land–atmosphere–ocean system to generate regional- or local-scale meteorological time series over many decades,
2. using the meteorological time series as input to rainfall-runoff models to generate hydrologic time series,
3. using the hydrologic scenarios as input to water management models, and
4. assessing differences among baseline and change scenarios using a variety of metrics.

Estimates of prospective impacts of climate change on precipitation have been mixed, leading in many cases to increasing uncertainty about the reliability of future water supplies. However, where snow provides a large fraction of annual water supplies, prospective temperature increases dominate hydrologic impacts, leading to stresses on water resources and increased hydrologic risk. Higher temperatures effectively shift the timing of the release of water stored in the snowpack "reservoir" to earlier in the year, reducing supplies in summer when demands are the greatest, while also increasing the risk of floods due to rain-on-snow events. Early studies on the Great Lakes system (Croley, 1990; Hartmann, 1990) demonstrated additional impacts of higher temperatures – greater lake heat storage leading to loss of ice cover, increased winter lake evaporation, lower lake levels, and potential failure to meet Lake Ontario regulation objectives under extant operating rules – and have been confirmed by more recent analyses (Lee *et al.*, 1994; Lee *et al.*, 1997; Sousounis *et al.*, 2000; Lofgren *et al.*, 2002). For other regions, extant water management systems for large river basins appear to be effective for all but the most severe scenarios (Hamlet and Lettenmaier, 1999b; Lettenmaier *et al.*, 1999), although in many regions, extensive detailed studies of the ability of existing reservoir systems and operational regulations to meet water management goals under changed climates are fairly recent

(e.g. Saunders and Lewis, 2003; Christensen *et al.*, 2004; Payne *et al.*, 2004; VanRheenen *et al.*, 2004).

Cognitively, climate-change information is difficult to integrate into water resources management. First, within the water resources engineering community, the stationarity assumption is a fundamental element of professional training; current hydrologic analysis techniques used in practice are seen as generally sufficient (e.g. Matalas, 1997; Lins and Stakhiv, 1998), especially in the context of slow policy and institutional evolution (Stakhiv, 2003). Second, the century timescales of climate change exceed typical planning and infrastructure design horizons and are remote from human experience. However, in the face of circumstances nearing or exceeding the effectiveness of existing management paradigms, individuals can become more cognizant of the need to consider climate change. In the US Southwest, over 1999–2004, Lake Powell levels declined faster than previously considered scenarios of extreme sustained drought (e.g. Harding *et al.*, 1995; Tarboton, 1995), from full to only 38% capacity in November 2004 (Bureau of Reclamation, 2004); the spatial extent, cascading ecological effects, and the rate of changes in the forested landscape on the Colorado Plateau over that same period have exceeded all prior expectations (7th Biennial Meeting on Integrating Science and Management on the Colorado Plateau, US Geological Survey, 4–7 November, 2003, Flagstaff, Arizona). Resource managers, policymakers, and the general public are now actively seeking scientific guidance in exploring how management practices can be more responsive to the uncertainties associated with a changing climate. Finally, even individuals trying to stay up-to-date can face confusion in conceptually melding the burgeoning climate change impacts literature. Assessments are often repeated as general circulation and hydrologic model formulations advance, or as new models become available throughout the research community. Further, assessments can employ a variety of techniques for downscaling; transposition techniques (e.g. Croley *et al.*, 1998) are more intuitive than the often mathematically complex statistical and dynamical downscaling techniques (e.g. Clark *et al.*, 1999; Westrick and Mass, 2001; Wood *et al.*, 2002; Benestad, 2004). The multiplicity of scenarios and vague attribution of their prospects for occurrence, which depend so strongly on feedbacks among social, economic, political, technological, and physical processes, further complicate conceptual integration of climate-change impacts assessment results in a practical water management context.

There is a rapidly growing literature on the consideration of climate change in water resources management (Frederick *et al.*, 1997; Gamble *et al.*, 2003; Lettenmaier, 2003; Loomis *et al.*, 2003; Snover *et al.*, 2003; Stakhiv, 2003; Ward *et al.*, 2003). Some (Matalas, 1997) contend that existing approaches are sufficient for water resource management planning and risk assessment because they contain safety factors. However, an inescapable message for the water resource management community is the inappropriateness of the stationarity assumption in the face of climate change. While precipitation changes may remain too uncertain for consideration in the near term, temperature increases are more certain and can have strong hydrologic consequences.

CONSIDERING WATER MANAGEMENT

Governments have made large investments to improve climate information and understanding over the past decades through satellites, *in situ* measuring networks, supercomputers, and research programs. However, there has been broad disappointment in the extent to which improvements in hydroclimatic science from large-scale research programs have affected resource management practices in general (Pielke, 1995, 2001; NRC, 1998a, 1999a), and water resource management in particular (NRC, 1998b, 1999b,c). Ensuring that improved data products, conceptual models, and predictions (forecasts and scenarios) are useful to the water resources management community has explicitly been identified as an important objective by several national and international programs (Endreny *et al.*, 2003 and also **Chapter 203, A Guide to International Hydrologic Science Programs, Volume 5**).

The water resources management milieu is complex and diverse, and climate influences are only one factor among the many affecting water management policies and practices. Many reasons exist for the slow adoption of advanced scientific information in water management including the lack of familiarity with available information, disconnects between the specific information available (e.g. variables, spatiotemporal timescales) and those relevant to decision makers, skepticism about the quality and applicability of information, and institutional impediments (Changnon, 1990; Pulwarty and Redmond, 1997; Pagano *et al.*, 2001, 2002; Jacobs, 2002; Jacobs and Pulwarty, 2003). Repeatedly, reviews of hydroclimatic science and water resources management (NRC, 1999a,b,c, 2001, 2004) recognize that more effective application of evolving hydroclimatic information requires coordinated efforts among the research, operational product generation, and water management communities.

Several ongoing efforts are leading the way forward to establish more effective ways of incorporating climate information into water resources management (Pulwarty, 2002; Office of Global Programs, 2004). While diverse in their details, all link natural variability, analytical and predictive technologies, and water management decisions within an end-to-end context extending from data through large-scale analyses and predictions (forecasts or scenarios), prediction evaluation, impacts assessment, applications, and evaluations of applications (e.g. Young, 1995; Miles *et al.*,

2000). For example, Georgakakos *et al.* have focused on linking climate forecasts with the management of complex river and reservoir systems (Georgakakos, 2002b). Key aspects of their approach include: (i) identifying the response of the hydrologic systems to climatic variability (Georgakakos *et al.*, 1998; Yao and Georgakakos, 2001); (ii) integrating meteorological forecasts into hydrologic forecasts (Carpenter and Georgakakos, 2001; Georgakakos and Krzysztofowicz, 2001); (iii) working with water managers to determine preferences when competing demands require trade-offs (Yao and Georgakakos, 2001); (iv) quantifying the impacts of alternative operational scenarios (Yao and Georgakakos, 1993); and (v) transferring end-to-end decision-support tools to water management entities (Georgakakos, 2002b).

However, water management situations often involve many different and increasingly competitive interests (e.g. energy production, reservoir recreation, downstream ecology). In these circumstances, effective integration of climate information into actual management practices and policies requires understanding what various interest groups value and how they interact and frame their problems (Pulwarty and Melis, 2000). For example, reliability of water supplies, as a key value in the US West, fosters decisions based more on "fear of failure" rather than "prospects for success" (Pagano *et al.*, 2001, 2002; Pagano and Garen, 2003). In contexts with large numbers of actors facing contentious issues, the barriers to the use of climate information can be formidable (Pulwarty and Redmond, 1997). However, the "use" of climate information, when broadly defined (Ray, 2004), may be more extensive than first perceived. Ray's typology of use of climate information includes:

1. *Consultation*. If individuals aren't receiving and looking at information, they lack the means for even this most basic type of use. If conditions aren't problematic, this use may be sufficient for most of the water management community.
2. *Consideration*. Information is confirmed or rejected by comparison with prior knowledge and experience-based personal judgment and mental models. Consideration requires prior consultation.
3. *Incorporation*. Information is generally required in specific forms that match decision processes, including the required variable (e.g. seasonal total precipitation, number of days between rain events), availability at the proper time within a decision calendar, appropriate spatial and temporal scale, and sufficient confidence in the information, among others. Incorporation requires prior consultation and consideration.
4. *Dialog about risks with stakeholders*. Risk dialog requires consultation and consideration, but can occur and be important even without explicit incorporation of climate information (e.g. because mitigation options are not available). For example, even if reservoir regulation rules cannot be used to mitigate anticipated drought impacts on water supply availability, it is useful for junior water rights holders to understand the risks they face in experiencing shortages of irrigation water.

Several projects, in the US Pacific Northwest, Southwest, and Great Lakes regions, embrace the complexities involved in integrating hydroclimate research and information in water resources management (Gamble *et al.*, 2003). The Great Lakes region has a long history of sustained relationships among hydroclimate researchers, operational water management entities, and stakeholders affected by management decisions. Efforts in the Pacific Northwest and Southwest are more recent and explicitly designed to be ongoing, dynamic processes; they are particularly useful for assessing the effectiveness of specific strategies for working with the water management community as new perceptions and practices emerge with regard to responding to climate variability and, increasingly, climate change. All three projects have developed an extensive literature (see Gamble *et al.*, 2003; Bales *et al.*, 2004; Lemos and Morehouse, 2005; Hartmann *et al.*, in review) detailing their methods for mutual knowledge development among researchers, operational information providers, decision makers, and stakeholders, which can provide guidance for others seeking to advance the use of climate information in water resource management in other contexts.

Key processes that facilitate the use of climate information include the use of highly structured workshops and collaborative prediction efforts (e.g. Garfin and Morehouse, 2001) to develop consensus forecasts or scenarios, establishment of permanent collaborative groups that meet repeatedly or network over extended periods (e.g. annual briefing workshops at the beginning of each water year; Gamble *et al.*, 2003), and exploiting environmental events that trigger interest in climate information and its application. Effective matching of water resource management professionals with the specific issue is critical as well. For example, innovative mid-level managers acting on their own initiative have proven effective at incorporating experimental climate forecasts in water resources operations and planning, and they have unique insights valuable for evaluating the appropriateness and effectiveness of particular hydroclimatic products (e.g. Changnon, 2000; Rayner *et al.*, 2001; Pagano *et al.*, 2002; Gamble *et al.*, 2003; Hartmann *et al.*, in review). However, experience in the Pacific Northwest demonstrated that the risk-averse tendencies of mid-level water managers makes them ineffective vectors for integrating climate-change information into water management policies and practices; rather, high-level decision makers have more capacity to use climate-change information (Climate Impacts Group, 2001; Gamble *et al.*, 2003). These projects and other assessments (e.g. Jacobs, 2002;

Wilson, 2002; Jacobs and Pulwarty, 2003) also confirm the importance of systematic approaches to knowledge development and better understanding of institutions from a dynamic perspective.

CONCLUSIONS AND RECOMMENDATIONS

Climate information, encompassing a variety of timescales, is clearly useful for water resource management, yet remains underutilized. Advances in climatological information and understanding are uneven; some regions have much information available, while others do not. It is incumbent on members of the water resource management community to know the state of climatic understanding for their region. For example, for their region, is there evidence of climate or hydrologic regimes or trends? Is there paleoclimatological evidence of extended drought? Are good climate forecasts available, and for what seasons and lead times? What are the anticipated climate and hydrologic impacts of global warming?

However, the water resource management community need not assess climate information and its use in isolation. Networks of researchers including hydroclimatic and social scientists are actively seeking more integrated involvement with the water resource management community, are adapting their research agenda and activities to address the needs of decision makers and stakeholders, and are committed to developing practical products and decision-support tools. Individuals in the water resource community should seek out these research groups to define the appropriate end-to-end chain connecting climate variability to hydrologic impacts and water management implications; the status of current climate information products vis-à-vis their practical use in water management practice; loci within various water management decision processes where climate uncertainty and information has potential to change outcomes; indicators to be used as triggers for management action (e.g. in drought plans); and options for increasing the flexibility of policies, planning processes, and operational practices. Members of a region's water resource management community, from high-level policy makers to operational field personnel should participate in two types of permanent collaborative groups: (i) those that provide ongoing dynamic integrated climate assessments; and (ii) those that focus on innovative water management practices, planning, and policies. However, because traditional funding mechanisms for hydroclimatic research limit technology transfer activities (Georgakakos, 2002b), the water resource management community must be willing to participate in cooperative agreements, provide incentives for conducting decision context-specific research, and commit to systematic institutional knowledge development (e.g. personnel training).

Although many issues are involved in better utilizing climate information in water resources management, two issues are of central importance: nonstationarity and uncertainty. The water resources community must accept that, in many cases, one of the fundamental assumptions underlying classical hydrology is simply incorrect. Water resource management practices and policies should begin with a presumption of nonstationarity, and justify their use of stationarity concepts in hydrologic analyses with strong evidence. Some (e.g. Matalas, 1997) contend that the stationarity assumption should not be completely disregarded, but it is clear that regimes and trends in hydroclimatic conditions are the rule, not the exception. Where applicable, the water resources management community must also recognize the important hydrologic implications of warmer temperatures both at the scale of extended climate forecasts and climate change.

The message that the water resources management community must better appreciate and accommodate uncertainty is not new (Matalas and Fiering, 1977). The computation of "best estimates" using instrumental records is clearly insufficient. Also, deterministic streamflow synthesis, simulation, and forecasting must be replaced with approaches that explicitly express uncertainty and risk, whether in terms of probabilities or using scenarios. Both scenario-based approaches and stochastic hydrology techniques, especially those that weigh more recent observations or observations from similar regimes more heavily, should be considered. While standard practices for scenario-based approaches have not yet emerged, practices to avoid are clear. For example, in using instrumental or paleoclimatological analog scenarios, assessments should search for "rhymes" rather than repeats of past situations. Further, assessments should not become fixed on the details of specific scenarios and their consequences, but focus on implications that are consistent (e.g. warming) versus implications that are more uncertain (e.g. precipitation). Finally, the water resource management community needs to better communicate uncertainty and risk generally, and in specific decision contexts, with stakeholders and the public.

Increasing the utility of hydroclimatic information is ultimately an interactive, iterative process involving researchers, forecasters, and the water resources management community over extended periods and multiple hydroclimatic events. Focusing on incremental improvements provides opportunities for sustained interactions as new analytical techniques, assessment processes, information products, and decision-support tools become available. Progress in hydroclimatic research and water resource management is neither easy nor assured. However, working together offers opportunities for progress unlikely to be realized independently.

Acknowledgments

The author gratefully acknowledges the support provided by grant NA16GP1577 from the National Oceanic and Atmospheric Administration (NOAA) Office of Global Programs and NSF-STC Grant EAR-9876800 for SAHRA, the Science and Technology Center for Sustainability of semi-Arid Hydrology and Riparian Areas.

REFERENCES

Angel J.R. and Huff F.A. (1995) Seasonal distribution of heavy rainfall events in the Midwest. *Journal of Water Resources Planning and Management*, **121**, 110–115.

Baldwin C.K. and Lall U. (1999) Seasonality of streamflow. *Water Resources Research*, **35**, 1143–1154.

Bales R.C., Liverman D.M. and Morehouse B.J. (2004) Integrated assessment as a step toward reducing climate vulnerability. *Bulletin of the American Meteorological Society*, **85**, 1727–1734.

Barnett T., Malone R., Pennell W., Astammer D., Demter B. and Washington W. (2004) The effects of climate change on water resources in the West: introduction and overview. *Climatic Change*, **62**, 1–11.

Barnston A.G., He Y. and Unger D.A. (2000) A forecast product that maximizes utility for state-of-the-art seasonal climate prediction. *Bulletin of the American Meteorological Society*, **81**, 1271–1279.

Beard D. (1993) *Blueprint for Reform: The Commissioner's Plan for Reinventing Reclamation*, Bureau of Reclamation: Washington.

Benestad R.E. (2004) Empirical-statistical downscaling in climate modeling. *EOS, Transactions, American Geophysical Union*, **85**, 417–422.

Bureau of Reclamation (1992) *A Long Term Framework for Water Resource Management, Development, and Protection*, U.S. Department of Interior: Washington.

Bureau of Reclamation (2004) *Drought Conditions in the West Continue*, U.S. Department of Interior: Washington. URL: `http: //www.usbr.gov/uc/feature/drought.html`

Cadavid L.G., VanZee R., White C., Trimble P. and Obeysekera J.T.B. (1999) Operational planning in south Florida using climate forecast. In *Proceedings of the 19th Annual American Geophysical Union Hydrology Days*, H.J., Morel-Seytoux (Ed.), Colorado State University: Ft. Collins.

Callahan B., Miles E. and Fulharty D. (1999) Policy implications of climate forecasting for water resources management in the Pacific Northwest. *Policy Sciences*, **32**, 269–293.

Carpenter T.M. and Georgakakos K.P. (2001) Assessment of Folsom Lake response to historical and potential future climate scenarios, I, forecasting. *Journal of Hydrology*, **249**, 148–175.

CBIAC (1955) *Use of 30-day Weather Outlook in Forecasting Runoff in the Columbia River Basin*, Columbia Basin Interagency Committee: Portland.

CBIAC (1964) *Derivation of Procedures for Forecasting Inflow to Hungry Horse Reservoir, Montana*, Columbia Basin Interagency Committee: Portland.

Changnon S.A. (1990) The dilemma of climatic and hydrologic forecasting for the Great Lakes. In *Proceedings of The Great Lakes Water Level Forecast and Statistics Symposium*, Hartmann H.C. and Donahue M.J. (Eds.), Great Lakes Commission: Ann Arbor, pp. 13–25.

Changnon S.A. (Ed.) (2000) *El Nino, 1997–1998: The Climate Event of the Century*, Oxford University Press: Oxford.

Changnon S.A. and Vonnhame D.R. (1986) Use of climate predictions to decide a water management problem. *Water Resources Bulletin*, **22**, 649–652.

Cherkauer K.A. and Lettenmaier D.P. (2002) Hydrologic effects of frozen soils in the upper Mississippi River Basin. *Journal of Geophysical Research, Atmospheres*, **104**, 19599–19610.

Christensen N.S., Wood A.W., Voisin N., Lettenmaier D.P. and Palmer R.N. (2004) Effects of climate change on the hydrology and water resources of the Colorado River Basin. *Climatic Change*, **62**, 337–363.

Clark M.P., Hay L.E., McCabe G.J., Leavesley G.H., Sereze M.C. and Wilby R.L. (2002) The use of weather and climate information in forecasting water supply in the western United States. *Water and Climate in the Western United States*, University Press of Colorado: Boulder.

Clark M.P., Hay L.E., McCabe G.J., Leavesley G.H. and Wilby R.L. (1999) Towards the use of atmospheric forecasts in hydrologic models, I, Forecast drift and scale dependencies. *EOS Transactions AGU*, **80**, Fall Meeting Supplement, Abstract H32G-10, F406–F407.

Climate Impacts Group (2001) *Climate and Water Policy Workshop Executive Summary*, Joint Institute for the Study of the Atmosphere and Oceans, University of Washington: Seattle.

Cole J.E., Overpeck J.T. and Cook E.R. (2002) Multiyear La Nina events and persistent drought in the contiguous United States. *Geophysical Research Letters*, **29**, 1647.

Congressional Budget Office (1997) *Water Use Conflicts in the West: Implications of Reforming the Bureau of Reclamation's Water Supply Policies*, Congressional Budget Office: Washington.

Croley T.E. (1990) Laurentian Great Lakes double-CO_2 climate change hydrological impacts. *Climatic Change*, **17**, 27–48.

Croley T.E. II (1996) Using NOAA's new climate outlooks in operational hydrology. *Journal of Hydrologic Engineering*, **1**, 93–102.

Croley T.E. II (1997) Mixing probabilistic meteorology outlooks in operational hydrology. *Journal of Hydrologic Engineering*, **2**, 161–168.

Croley T.E. II (2000) *Using Meteorology Probability Forecasts in Operational Hydrology*, American Society of Civil Engineers Press: Reston.

Croley T.E. II. and Lee D.H. (1993) Evaluation of Great Lakes net basin supply forecasts. *Water Resources Research*, **29**, 267–282.

Croley T., Quinn F., Kunkel K. and Changnon S. (1998) Great Lakes hydrology under a transposed climate. *Climatic Change*, **38**, 405–433.

Davis R.E. and Pangburn T. (1999) Development of new snow products for operational water control and management in the Kings River Basin, California. *EOS Transactions AGU*, **81**, Spring Meeting Supplement, Abstract H22D-07, S110.

Day G.N. (1985) Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management*, **111**, 157–170.

Douglas E.M., Vogel R.M. and Kroll C.N. (2000) Trends in flood and low flows across the U.S. *Journal of Hydrology*, **240**, 90–105.

Douglas E.M., Vogel R.M. and Kroll C.N. (2002) Impact of streamflow persistence on hydrologic design. *Journal of Hydrologic Engineering*, **7**, 220–227.

Endreny T., Felzer B., Shuttleworth J.W. and Bonell M. (2003) Policy to coordinate watershed hydrological, social, and ecological needs: the HELP Initiative. In *Water: Science, Policy, and Management*, Lawford R., Fort D., Hartmann H. and Eden S. (Eds.), American Geophysical Union: Washington, pp. 395–411.

Environmental Protection Agency (1989) *The Potential Effects of Global Climate Change on the United States. Report to Congress*, Smith J.B. and Tirpak D. (Eds), EPA Office of Policy, Planning and Evaluation: Washington.

Franz K., Hartmann H.C., Sorooshian S. and Bales R. (2003) An evaluation of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River Basin. *Journal of Hydrometeorology*, **4**, 1105–1118.

Frederick K., Major D. and Stakhiv E. (Eds.) (1997) *Climate Change and Water Resources Planning Criteria*, Kluwer Academic Publishers: Dordrecht.

Frevert D.K., Cowan M.S. and Lane W.L. (1989) Use of stochastic hydrology in reservoir operation. *Journal of Irrigation and Drainage Engineering*, **115**, 334–343.

Gamble J.L., Furlow J., Snover A.K., Hamlet A.F., Morehouse B.J., Hartmann H. and Pagano T. (2003) Assessing the impact of climate variability and change on regional water resources: the implications for stakeholders. In *Water: Science, Policy, and Management*, Lawford R., Fort D., Hartmann H.,and Eden S. (Eds.), American Geophysical Union: Washington, pp. 341–368.

Garfin G. and Morehouse B. (Eds.) (2001) *Proceedings, 2001 Fire and Climate Workshops*, Institute for the Study of Planet Earth, University of Arizona: Tucson.

Georgakakos K.P. (2002a) Climate forecasts and water resources management: a fertile field for hydroinformatics, *Proceedings of the Fifth International Conference on Hydroinformatics, Volume Two: Software Tools and Management Systems*, Cardiff, pp. 797–804.

Georgakakos K.P. (2002b) US corporate technology transfer in hydrometeorology. *Journal of Hydroinformatics*, **4**, 3–13.

Georgakakos K.P. and Krzysztofowicz R. (Eds.) (2001) Special issue on probabilistic and ensemble forecasting. *Journal of Hydrology* **249**, 1–196.

Georgakakos A.P., Yao H., Mullusky M.G. and Georgakakos K.P. (1998) Impacts of climate variability on the operational forecast and management of the Des Moines River basin. *Water Resources Research*, **34**, 799–821.

Glantz M.H. (1982) Consequences and responsibilities in drought forecasting: the case of Yakima, 1977. *Water Resources Research*, **18**, 3–13.

Gleick P.H. and Adams D.B. (2000) *Water: The Potential Consequences of Climate Variability and Change for Water Resources of the United States*, Pacific Institute: Oakland.

Goddard L., Mason S.J., Zebiak S.E., Ropelewski C.F., Basher R. and Cane M.A. (2001) Current approached to seasonal-to-interannual climate predictions. *International Journal of Climatology*, **21**, 1111–1152.

Gray S.T., Betancourt J.L., Fastie C.L. and Jackson S.T. (2003) Patterns and sources of multidecadal oscillations in drought-sensitive tree-ring records from the central and southern Rocky Mountains. *Geophysical Research Letters*, **30**, 1316.

Grissino-Mayer H. (1996) A 2129-year reconstruction of precipitation for northwestern New Mexico, U.S.A. In *Tree Rings, Environment and Humanity*, Dean J.S., Meko D.M. and Swetnam T.W. (Eds.), Radiocarbon: Tucson, pp. 191–204.

Hamlet A.F., Huppert D. and Lettenmaier D.P. (2002) Economic value of long-lead streamflow forecasts for Columbia River hydropower. *Journal of Water Resources Planning and Management*, **128**, 91–101.

Hamlet A.F. and Lettenmaier D.P. (1999a) Columbia River streamflow forecasting based on ENSO and PDO climate signals. *Journal of Water Resources Planning and Management*, **125**, 333–341.

Hamlet A.F. and Lettenmaier D.P. (1999b) Effects of climate change on hydrology and water resources of the Columbia River Basin. *Journal of the American Water Resources Association*, **35**, 1597–1623.

Hamlet A.F. and Lettenmaier D.P. (2000) Long range forecasting and its use for water management in the Pacific Northwest region of North America. *Journal of Hydroinformatics*, **2**, 163–182.

Hansen J., Sato M., Ruedy R., Lacis A., Asamoah K., Beckford K., Borenstein S., Brown E., Cairns B., Carlson B., *et al.* (1997) Forcings and chaos in interannual to decadal climate change. *Journal of Geophysical Research*, **102**, 25679–25720.

Harding B.J., Sangoyomi T.B. and Payton E.A. (1995) Impacts of severe sustained drought on Colorado River water resources. *Water Resources Bulletin*, **31**, 815–824.

Hartmann H.C. (1990) Impacts on Laurentian Great Lakes levels. *Climatic Change*, **17**, 49–68.

Hartmann H.C., Bales R. and Sorooshian S. (2002b) Weather, climate, and hydrologic forecasting for the U.S. Southwest: a survey. *Climate Research*, **21**, 239–258.

Hartmann H.C., Bradley A. and Hamlet A. (2003) Advanced hydrologic predictions for improving water management. In *Water: Science, Policy, and Management*, Lawford R., Fort D., Hartmann H. and Eden S. (Eds.), American Geophysical Union: Washington, pp. 285–307.

Hartmann H.C., Garfin G.M., Morehouse B., Sorooshian S., Vásquez-León M. and Bales R. (in review) Forecast assessment: a key element in stakeholder-driven integrated climate assessments. *Global Environmental Change*.

Hartmann H.C., Imam B., Lay E., Lamb D. and Sorooshian S. (2004) A tool for improving natural resource management under climate uncertainty. *Proceedings, 29th Climate Diagnostics Workshop*, National Weather Service: Silver Spring.

Hartmann H.C., Pagano T.C., Bales R. and Sorooshian S. (2002a) Confidence builders: evaluating seasonal climate forecasts from user perspectives. *Bulletin of the American Meteorological Society*, **83**, 683–698.

Herche L.R. and Hartmann H.C. (1992) Estimation of Great Lakes water level statistics: conditioning via "the bootstrap". *Journal of Great Lakes Research*, **18**, 218–228.

Hirsch R.M., Helsel D.R., Cohm T.A. and Gilroy E.J. (1993) Statistical treatment of hydrologic data. In *Handbook of Hydrology*, Chap. 17, Maidment D.R. (Ed.), McGraw-Hill.

Hobbs G. (2004) The role of climate in shaping western water institutions. *Denver University Water Law Review*, **7**, 101–145.

IPCC (1990) *Scientific Assessment of Climate Change: Report of Working Group I to the First Assessment Report of the IPCC*, Cambridge University Press: Cambridge.

IPCC (1995a) *Climate Change 1995: IPCC Second Assessment*, Cambridge University Press: Cambridge.

IPCC (1995b) *Impacts, Adaptations and Mitigations: Contributions of Working Group II to the Second Assessment Report of the IPCC*, Cambridge University Press: Cambridge.

IPCC (2001a) *Climate Change 2001: Synthesis Report. Third Assessment Report of the IPCC*, Cambridge University Press: Cambridge.

IPCC (2001b) *Impacts, Adaptations, and Vulnerability: Contribution of Working Group II to the Third Assessment Report of the IPCC*, Cambridge University Press: Cambridge.

Ivanov V.Y., Vivoni E.R., Bras R.L. and Entekhabi D. (2004) Preserving high-resolution surface and rainfall data in operational-scale basin hydrology: a fully-distributed physically-based approach. *Journal of Hydrology*, **298**, 80–111.

Jacobs K. (2002) *Connecting Science, Policy, and Decision-Making: A Handbook for Researchers and Science Agencies*, Office of Global Programs, National Oceanic and Atmospheric Administration: Silver Spring.

Jacobs K. and Pulwarty R. (2003) Water resource management: science, planning and decision-making. In *Water: Science, Policy, and Management*, Lawford R., Fort D., Hartmann H. and Eden S. (Eds.), American Geophysical Union: Washington, pp. 177–204.

Kite G.W. (1992) Spectral analysis of selected Lake Erie levels. *Journal of Great Lakes Research*, **18**, 207–217.

Klemes V. (1974) The Hurst phenomenon: a puzzle? *Water Resources Research*, **10**, 675–687.

Kubik H.E. (1992) Annual extreme lake level elevations by total probability theorem. *Journal of Great Lakes Research*, **18**, 202–206.

Lee D.H. (1999) Institutional and technical barriers to implementing risk-based water resources management: a case study. *Journal of Water Resources Planning and Management*, **125**, 186–193.

Lee D.H., Croley T.E. II and Quinn F.H. (1997) Lake Ontario regulation under transposed climates. *Journal of the American Water Resources Association*, **33**, 55–69.

Lee D.H., Quinn F.H., Sparks D. and Rassam J.C. (1994) Modification of Great Lakes regulation plans for simulation of maximum Lake Ontario outflows. *Journal of Great Lakes Research*, **20**, 569–582.

Lemos M.C. and Morehouse B.J. (2005) The co-production of science and policy in integrated climate assessments. *Global Environmental Change*, **15**, 57–68.

Lettenmaier D.P. (2003) The role of climate in water resources planning and management. In *Water: Science, Policy, and Management*, Lawford R., Fort D., Hartmann H. and Eden S. (Eds.), American Geophysical Union: Washington, pp. 247–266.

Lettenmaier D.P. and Wood E.F. (1993) Hydrologic forecasting. In *Handbook of Hydrology*, Chap. 26, Maidment D.R. (Ed.), McGraw-Hill.

Lettenmaier D., Wood A., Palmer R., Wood E. and Stakhiv E. (1999) Water resources implications of global warming: a U.S. regional perspective. *Climatic Change*, **43**, 537–579.

Lettenmaier D.P., Wood E.F. and Wallis J.R. (1994) Hydro-climatological trends in the continental United States, 1948-88. *Journal of Climate*, **7**, 586–607.

Liang X., Lettenmaier D.P., Wood E.F. and Burges S.J. (1994) A simple hydrologically based model of land surface water and energy fluxes for GSMs. *Journal of Geophyscial Research, Atmospheres*, **99**, 14415–14428.

Lins H.F. and Slack J.R. (1999) Streamflow trends in the United States. *Geophysical Research Letters*, **26**, 227–230.

Lins H.F. and Stakhiv E.Z. (1998) Managing the nation's water in a changing climate. *Journal of the American Water Resources Association*, **34**, 1255–1264.

Linsley R.K., Kohler M.A. and Paulhus J.L.H. (1975) *Hydrology for Engineers*, *Second Edition*, McGraw Hill Book Company: New York.

Lofgren B.M., Quinn F.H., Clites A.H., Assel R.A., Eberhardt A.J. and Luukkonen C.L. (2002) Evaluation of potential impacts on Great Lakes water Resources based on climate scenarios of two GCMs. *Journal of Great Lakes Research*, **28**, 537–554.

Loomis J., Koteen J. and Hurd B. (2003) Economic and institutional strategies for adapting to water-resource effects of climate change. In *Water and Climate in the Western United States*, Lewis W. (Ed.), University Press of Colorado: Boulder, pp. 235–249.

Maidment D.R. (Ed.) (1993) *Handbook of Hydrology*, McGraw-Hill: New York.

Mandelbrot B.B. and Wallis J.R. (1969) Some long-run properties of geophysical records. *Water Resources Research*, **5**, 321–340.

Matalas N.C. (1997) Stochastic hydrology in the context of climate change. *Climatic Change*, **37**, 89–101.

Matalas N.C. and Fiering M.B. (1977) Water resources system planning.In *Climate, Climatic Change, and Water Supply*, National Academy Press: Washington, pp. 99–110.

Miles E.L., Snover A.K., Hamlet A.F., Callahan B. and Fluharty D. (2000) Pacific northwest regional assessment: the impacts of climate variability and change on the water resources of the Columbia river basin. *Journal of the American Water Resources Association*, **36**, 399–420.

Modini G.C. (2000) Long-lead precipitation outlook augmentation of multi-variate linear regression streamflow forecasts, *Proceedings of the 68th Annual Western Snow Conference*, Port Angeles, pp. 57–68.

National Assessment Synthesis Team (2000) *Climate Change Impacts on the United States: The Potential Consequences of Climate Variability and Change*, U.S. Global Change Research Program: Washington.

Ni F., Cavazos T., Hughes M.K., Comrie A.C. and Funkhouser G. (2002) Cool-season precipitation in the Southwestern USA since AD 1000: comparison of linear and nonlinear techniques for reconstruction. *International Journal of Climatology*, **22**, 1645–1166.

NRC (1995) *Flood Risk Management and the American River Basin: An Evaluation*, National Academy Press: Washington.

NRC (1998a) *GCIP: A Review of Progress and Opportunities*, National Academy Press: Washington.

NRC (1998b) *Hydrologic Sciences: Taking Stock and Looking Ahead*, National Academy Press: Washington.

NRC (1999a) *Making Climate Forecasts Matter*, National Academy Press: Washington.

NRC (1999b) *A Vision for the National Weather Service: Road Map for the Future*, National Academy Press: Washington.

NRC (1999c) *Hydrologic Science Priorities for the U.S. Global Change Research Program: An Initial Assessment*, National Academy Press: Washington.

NRC (2001) *Envisioning the Agenda for Water Resources Research in the Twenty-first Century*, National Academy Press: Washington.

NRC (2004) *Confronting the Nation's Water Problems: The Role of Research*, National Academy Press: Washington.

NWS (1999) *Strategic Plan for Weather, Water, and Climate Services: 2000-2005*, Department of Commerce: Washington.

Office of Global Programs (2004) *Regional Integrated Sciences and Assessments*, National Oceanic and Atmospheric Administration. **http://www.risa.ogp.noaa.gov**, 17 March 2004.

Olsen J.R., Lambert J.H. and Haimes Y.Y. (1998) Risk of extreme events under nonstationary conditions. *Risk Analysis*, **18**, 497–510.

Olsen J.R., Stedinger J.R., Matalas N.C. and Stakhiv E.Z. (1999) Climate variability and flood frequency estimation for the upper Mississippi and lower Missouri rivers. *Journal of the American Water Resources Association*, **35**, 1509–1523.

Pagano T.C. and Garen D.C. (2003) Use of climate information in official western U.S. water supply forecasts. In *Proceedings, World Water and Environmental Resources Congress*, Bizier P. and DeBarry P. (Eds.), American Society of Civil Engineers: Reston, CD-ROM.

Pagano T.C., Garen D. and Sorooshian S. (2004) Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002. *Journal of Hydrometeorology*, **5**, 896–909.

Pagano T.C., Hartmann H.C. and Sorooshian S. (2001) Using climate forecasts for water management: Arizona and the 1997-98 El Nino. *Journal of the American Water Resources Association*, **37**, 1139–1153.

Pagano T.C., Hartmann H.C. and Sorooshian S. (2002) Use of climate forecasts for water management in Arizona: a case study of the 1997-98 El Niño. *Climate Research*, **21**, 59–269.

Payne J.T., Wood A.W., Hamlet A.F., Palmer R.N. and Lettenmaier D.P. (2004) Mitigating the effects of climate change on the water resources of the Columbia River Basin. *Climatic Change*, **62**, 233–256.

Perica S. (1998) Integration of meteorological forecasts/climate outlooks into an ensemble streamflow prediction system. In *Preprints, 14th Conference on Probability and Statistics in the Atmospheric Sciences*, American Meteorological Society: Boston, pp. 130–133.

Perica S., Schaake J. and Seo D.J. (1999) National weather service river forecast system (NWSRFS) operational procedures for using short and long range precipitation forecasts as input to ensemble streamflow prediction (ESP). In *Preprints, 14th Conference on Hydrology*, American Meteorological Society: Boston.

Pielke R.A. Jr (1995) Usable information for policy: an appraisal of the U.S. global change research program. *Policy Sciences*, **38**, 39–77.

Pielke R.A. Jr (2001) *The Development of the U.S. Global Change Research Program: 1987 to 1994*, Policy Case Study, National Center for Atmospheric Research: Boulder.

Potter K.W. (1976) Evidence of nonstationarity as a physical explanation of the Hurst phenomenon. *Water Resources Research*, **12**, 1047–1052.

Potter K.W. (1992) Estimating the probability distribution of annual maximum levels on the Great Lakes. *Journal of Great Lakes Research*, **18**, 229–235.

Privalsky V. (1992) Statistical analysis and predictability of Lake Erie water level variations. *Journal of Great Lakes Research*, **18**, 236–243.

Pulwarty R.S. (2002) *Regional Integrated Sciences and Assessment Program*, Office of Global Programs, National Oceanic and Atmospheric Administration: Silver Spring.

Pulwarty R. and Melis T. (2000) Climate extremes and adaptive management on the Colorado River. *Journal of Environmental Management*, **63**, 307–324.

Pulwarty R.S. and Redmond K.T. (1997) Climate and salmon restoration in the Columbia River basin: the role and usability of seasonal forecasts. *Bulletin of the American Meteorological Society*, **78**, 381–397.

Quinn F.H. (1981) Secular changes in annual and seasonal Great Lakes precipitation, 1854–1979, and their implications for Great Lakes water resources studies. *Water Resources Research*, **17**, 1619–1624.

Quinn F.H. (2002) Secular changes in Great Lakes water level changes. *Journal of Great Lakes Research*, **28**, 451–465.

Ray A.J. (2004) *Linking Climate to Multi-purpose Reservoir Management: Adaptive Capacity and Needs for Climate Information in the Gunnison Basin, Colorado*, Ph.D. Dissertation, Geography Department, University of Colorado, Boulder.

Rayner S., Lach D., Ingram H. and Houck M. (2001) *Why Water Resource Managers Don't Use Climate Forecasts*, International Research Institute on Climate Prediction: Palisades.

Redmond K.T. and Koch R.W. (1991) Surface climate and streamflow variability in the western United States and their relationship to large-scale circulation indices. *Water Resources Research*, **27**, 2381–2399.

Rhodes S., Ely D. and Dracup J. (1984) Climate and the Colorado River: limits to management. *Bulletin of the American Meteorological Association*, **65**, 682–691.

Salas J.D. (1993) Analysis and modeling of hydrologic time series. In *Handbook of Hydrology*,Chap. 19, Maidment D.R. (Ed.), McGraw-Hill: New York.

Salzer M.W. and Kipfmueller K.F. (in press) Reconstructed temperature and precipitation on a millennial timescale from tree-rings in the Southern Colorado Plateau. *Climatic Change*, in press.

Saunders J.F. III and Lewis W.M. Jr (2003) Implications of climatic variability for regulatory low flows in the South Platte River Basin, Colorado. *Journal of the American Water Resources Association*, **39**, 33–45.

Scott M.J., Vail L.W., Jaksch J.A. and Anderson K.K. (2000) Considerations for management of irrigation water with climate variability. *EOS Transactions AGU*, **81**, Spring Meeting Supplement, Abstract H32B-06, S205.

Shallot T. (2003) Success through failure: Army science in harbor construction - 1820 to 1860. *Water Resources Impact*, **5**, 5–8.

Smith J.A., Day G.N. and Kane M.D. (1992) Nonparametric framework for long-range streamflow forecasting. *Water Resources Bulletin*, **118**, 82–92.

Snover A.K., Hamlet A.F. and Lettenmaier D.P. (2003) Climate change scenarios for water planning studies: pilot applications in the Pacific Northwest. *Bulletin of the American Meteorological Society*, **84**, 1513–1518.

Sonka S.T., Changnon S.A. and Hofing S.L. (1992) How agribusiness uses climate predictions- implications for climate research and provision of predictions. *Bulletin of the American Meteorological Society*, **73**, 1999–2008.

Sousounis P., Albercook G., Allen D., Andresen J., Brooks A., Brown D., Cheng H.H., Davis M., Lehman J., Lindeberg J., *et al.* (2000) *Preparing for a Changing Climate: The Potential Consequences of Climate Variability and Change for the Great Lakes*, U.S. Global Change Research Program: Washington.

Stakhiv E. (2003) What can water managers do about climate variability and change? In *Water and Climate in the Western United States*, Lewis W. (Ed.), University Press of Colorado: Boulder, pp. 131–142.

Stedinger J.R., Vogel R.M. and Foufoula-Georgiou F. (1993) The frequency analysis of extreme events. In *Handbook of Hydrology*, Chap. 18, Maidment D.R. (Ed.), McGraw-Hill: New York.

Tarboton D.G. (1995) Hydrologic scenarios for severe sustained drought in the Southwestern United States. *Water Resources Bulletin*, **35**, 803–813.

Thompson J. (2002) *Wetlands Drainage, River Modification, and Sectoral Conflict in the Lower Illinois Valley, 1890-1930*, Southern Illinois University Press: Carbondale.

Urbanas B.R. and Roesner L.A. (1993) Hydrologic design for urban drainage and flood control. In *Handbook of Hydrology*, Chap. 28, Maidment D.R. (Ed.), McGraw-Hill: New York.

U.S. Army Corps of Engineers (1992) *Guidelines for Risk and Uncertainty Analysis in Water Resources Planning, Volumes I and II*, IWR Report 92-R-1, 92-R-2, Institute for Water Resources, Fort Belvoir.

VanRheenen N., Wood A.W., Palmer R.N. and Lettenmaier D.P. (2004) Potential implications of PCM climate change scenarios for Sacramento-San Joaquin River basin hydrology and water resources. *Climatic Change*, **62**, 257–281.

Vorosmarty C., Lettenmaier D., Leveque C., Meybeck M., Pahl-Wostl C., Alcamo J., Cosgrove W., Grassl H., Hoff H., Kabat P., *et al.* (2004) Humans transforming the global water system. *EOS, Transactions, American Geophysical Union*, **85**, 509–514.

Walker F.R. and Douglas E.M. (2003) Identifying hydrologic variability and change for strategic water system planning and design. In *Water: Science, Policy, and Management*, Lawford R., Fort D., Hartmann H. and Eden S. (Eds.), American Geophysical Union: Washington, pp. 267–284.

Ward R.C., Pielke R. Sr and Salas J. (Eds.) (2003) Special issue: is global climate change research relevant to day-to-day water resources managers? *Water Resources Update*, **124**, 1.

Water Resources Council (1981) *Guidelines for Determining Flood Risk Frequency, Revised Bulletin 17B*. U.S. Water Resources Council: Washington.

Westrick K.J. and Mass C.F. (2001) An evaluation of a high resolution hydrometeorological modeling system for prediction of a cool-season flood event in a coastal mountainous watershed. *Journal of Hydrometeorology*, **2**, 161–180.

Wigmosta M.S., Vail L.W. and Lettenmaier D.P. (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, **30**, 1665–1669.

Wilson J. (2002) Scientific uncertainty, complex systems, and the design of common-pool institutions. *The Drama of the Commons*, National Academy Press: Washington, pp. 327–360.

Wood E., Maurer E.P., Kumar A. and Lettenmaier D.P. (2002) Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research, Atmospheres*, **107**, 4423–4429.

Yao H. and Georgakakos A. (1993) New control concepts for uncertain water resources systems: 2, reservoir management. *Water Resources Research*, **29**, 1517–1526.

Yao H. and Georgakakos A. (2001) Assessment of Folsom Lake response to historical and potential future climate scenarios, II, reservoir management. *Journal of Hydrology*, **249**, 176–196.

Young R.A. (Ed.) (1995) Special issue: managing the Colorado River in a severe sustained drought. *Water Resources Bulletin*, **35**, 779–944.

203: A Guide to International Hydrologic Science Programs

RG LAWFORD[1] AND S EDEN[2]

[1] *International GEWEX Project Office, Silver Spring, MD, US*
[2] *US Climate Change Science Program Office, Washington, DC, US*

Building on the successes of a few mature organizations dealing with the hydrological sciences and the International Hydrologic Decade of the 1960s, the number of water research and applications programs has grown rapidly in the past two decades. This growth can be attributed to new global research priorities (e.g. climate change), new technologies (e.g. satellites and high speed computers), new environmental awareness (e.g. globalization and its environmental impacts), and a growing awareness of the limits to available safe water in many parts of the world. Although some progress was made towards better structuring of hydrologic research activities during preparations for the International Water Decade (1981–1990), the relative roles and relationships among the plethora of groups and projects can still be confusing unless the range of activities and responsibilities is viewed in a broader context. This overview provides some context for these programs.

In general, international hydrologic research activities are undertaken within a programmatic framework developed by a community of scientists and coordinated by international scientific organizations and government agencies. Because international frameworks and programs do not fund research directly, they must rely on nations for funding support and experts. Frequently, national and international programs are implemented in an iterative and symbiotic way whereby national programs provide a core set of projects, while international programs provide the intellectual frameworks and organizational structures. Effective implementation of international hydrologic science programs in this context relies on mutual trust and respect, and on the clear articulation and pursuit of a shared vision.

This article provides an overview of the international hydrological science programs and their priorities, and identifies related global hydrometeorological and applications programs that either directly or indirectly provide scientific support for water management. It also describes mechanisms that facilitate the application of hydrologic science to regional and national water resource issues and those that inform the public and decision makers of the results from these international science programs.

INTRODUCTION

International hydrological science has developed rapidly in the past two decades. This development has taken a unique path because of the nature of hydrologic issues. Until recently, many hydrologic problems focused on individual basins or catchments; consequently, the requirement for international collaboration was limited. Institutions developed independently such that authorities and responsibilities for water resources varied from nation to nation. In some countries the control of water was highly centralized, with monitoring and forecast services maintained at the national level; in other countries, responsibilities were delegated to regional or county levels or were never fully developed. These differences affect data collection and archival issues, making it more difficult to assemble consistent data sets. Data problems are compounded by policies that limit access in the name of national security or revenue generation. Collaboration on transboundary waters was largely limited to bilateral structures, such as the International Joint Commission (United States–Canada).

Encyclopedia of Hydrological Sciences. Edited by M G Anderson, 2005.

On the science side, hydrology has always been highly multidisciplinary. It has drawn on diverse groups such as civil engineering, which emphasizes models and application of hydrologic information to infrastructure design. It has also entrained geographical sciences, with their spatial and inventory perspectives. New complex system challenges are widening the range of disciplinary perspectives with useful contributions to hydrological science. International organizations, programs, and initiatives have received more attention and have accelerated development and more effective coordination to overcome this fragmentation as a result of two special programs; the International Hydrological Decade (IHD from 1965–1974) and the International Water Decade (1981–1990).

Internationally, hydrologic research is undertaken frequently within a programmatic and scientific framework developed by a community of scientists and coordinated by an international science organization. These organizations can be academic, governmental, or private. They play a major role in international science because they bring together experts from a broad range of specialities to address a full spectrum of science questions and provide a forum for science discussions and program formulation.

These international frameworks and programs do not fund research directly and must rely on individual nations to fund the activities and experts required to make the program a success. International programs continue to have an important role in hydrology where, historically, research has focused on basin scale phenomena. They give researchers access to a much broader community of experts, allow for a coordinated and comprehensive approach to problems so nations can use their funds more effectively, and advance and improve the science by making national contributions compatible with increasing implicit global standards.

Some international hydrology research programs progress by coordinating, influencing, and guiding national research activities to achieve the goals set out in their internationally approved science plans. This approach has been successful where strong leadership is present. Frequently, scientific leadership is provided through both a steering committee composed of active and knowledgeable scientists and a project office or secretariat that tracks progress and ensures that participants are contributing to the needs of the program in a timely and effective manner.

Some science efforts at the international level involve large projects that entrain a number of scientists from different regions to work on common questions. Because of their size and the opportunities they provide for international collaboration, these projects often attract some of the best and most visionary researchers. International project leaders and frequently facilitators with an ability to draw together international scientists to develop science plans, coordinate implementation, and interpret research results in a broad context.

The coordination of science within an international framework can be challenging because it relies on trust and the development and pursuit of a shared vision. Even though scientific approaches are apolitical, the issues addressed can be politically sensitive (e.g. the construction of new dams), so that trust is essential at all levels. A challenge for some programs arises from a reluctance of suitable scientific experts to work in international coordination roles for this type of research.

In hydrologic sciences, research is also coordinated through United Nations (UN) Water, and its special agencies, such as the World Meteorological Organization (WMO) and the United Nations Education, Scientific and Cultural Organization (UNESCO). Generally, the largest contributions for hydrological and related climate programs come from the United States, Europe, and Japan. There appears to be a second tier of funders, including Brazil, China, Canada, and Australia with other countries contributing as they can. All the three major funders maintain strong space programs and consider hydrological and water cycle issues as part of their observations and research related to the Earth's system. Another coordination mechanism is the International Council for Science (ICSU) and its International Association for Hydrological Sciences (IAHS), which works directly with the scientists.

This article provides an overview of the international hydrological science programs and briefly identifies related hydrological and hydrometeorological application programs that directly or indirectly provide scientific support for water management. This discussion is limited to programs that are global or at least multinational in scope. It also describes international mechanisms that facilitate the application of hydrologic science to regional and national water resource issues.

Before the individual programs and initiatives are discussed, the primary broad research frameworks and coordinating organizations will be described (Sections "Factors and principles governing the development of international hydrologic science programs" and "Umbrella programs for hydrologic research"). Program descriptions are divided into three sections: the Section on "Programs focused on observations and data management", the Section "Programs focused on process and modeling studies and basic science", and the Section "Knowledge transfer: applications, assessments, applied research, and outreach" discusses assessments, applications, and outreach. Most programs and initiatives have activities in two or more of these categories, although one is usually a priority. However, for this article, they are classified according to their primary purpose. The following review draws heavily on Lawford *et al.* (2003) and web-based sources (the web sites are listed in Appendix A).

FACTORS AND PRINCIPLES GOVERNING THE DEVELOPMENT OF INTERNATIONAL HYDROLOGIC SCIENCE PROGRAMS

In the past the international hydrologic sciences may have developed more slowly than meteorological and oceanographic sciences because water resource problems are primarily experienced at local, regional, basin, and/or national scales and, with the exception of transboundary problems, are addressed by national or state governments. Issues such as transboundary river flows are often viewed as sensitive and may involve sociopolitical and legal disputes. Even though science itself strives to be apolitical and objective, the funding and requirements for science do influence the directions in which science develops. Although international coordination of hydrological research has taken place since at least 1922, the demands and opportunities for global-scale hydrologic science have appeared only in the last few decades. IHD was a landmark event for international water science, because a network of research basins was established around the world. Since then hydrologic systems have come to be viewed as a global concern, rather than just a mosaic of disconnected local and regional problems. Nations without sufficient water resources are becoming aware that they need to be open to new strategies for overcoming shortages in supply. In addition, the advent of advanced computing systems has accelerated the development of data assimilation and prediction capabilities. Scientific research and technological developments can provide the scientific basis and infrastructure needed to advance international cooperation in hydrology. Factors that contribute to this change in perspective are discussed below.

The View from Space

Since their advent in the late 1960s, meteorological and other Earth satellites have gradually been changing our view of the hydrological cycle. The ability to look at the Earth from space has given us a more holistic appreciation of the Earth's global water cycle. Hydrologic sciences are an essential element of Earth science as it expands its boundaries to gain an integrated understanding of the functioning of the Earth system.

Global Environmental Change

A major impetus for international hydrological science programs in recent years has been the concern about the impacts of global change associated with global climate change and with growing demands for water and other resources arising from increasing populations and economic development. Large, shared resources, such as the Arctic Sea, are being polluted by multiple countries, and the anthropogenic alteration of such a global-scale ecosystem is likely to result in correspondingly global-scale consequences (IPCC, 2001). Arguably, the major issue pushing hydrologists to develop and provide a global perspective is the issue of global environmental change, of which climate change is a significant aspect. It is clear that the uncertainties arising from issues of climate variability and change cannot be understood without a complete understanding of the global water cycle. Furthermore, a broader base for hydrologic research is needed to understand the scientific basis for strategies to meet the goals of Rio 21 and the UN Millennium Development Goals (MDGs). Indeed, the globalization of hydrologic research is coming about partly because of the need to encourage more nations to share hydrologic data to improve models for global analysis and assessment and partly because many contiguous countries are facing similar environmental challenges as they share international river basins.

Sustainable Development

Through its Commission on Sustainable Development (CSD), the UN provides a major policy framework for addressing international water issues. This approach supports the principles embodied in the UN charter, namely, the promotion of peace and equality among the people of the world and the protection of societies from threats that may lead to widespread loss. The adoption of the principle of sustainable development as outlined in the "The Brundtland Report" (World Commission on Environment and Development, 1987) is widely regarded as a feasible path to achieve long-term global stability. In order to clarify its approach to water and other renewable resources, the UN has structured many of its new initiatives around the theme of sustainable development. Furthermore, the linked concerns of global environmental change and meeting the growing need for underserved communities for access to clean water are motivating applied research in support of sustainable resource management, such as "Integrated Water Resource Management (IWRM)".

International and Intersectoral Dialogue

The emergence of organizations that promote and facilitate international and intersectoral dialogue has stimulated the communication needed to advance hydrologic science and its application to real problems. The World Water Council (WWC) and Global Water Partnership (GWP) are examples of organizations attempting to draw groups together to identify issues that transcend national boundaries and to develop policy options. These organizations use cross-sectoral dialogue as well as dialogues within international fora (such as the World Water Forum) to address common water issues. Some of these issues are

short-term, such as the need for flood warning systems and data inputs to aid tactical water management decisions. Other issues require long-term approaches because they are linked with evolving factors such as depletion of groundwater reserves, increasing demands for surface water, and declining water quality. Dire, long-term consequences can result from the cumulative effects of poor short-term water management decisions, especially when these are superimposed on land use and socio-economic choices that increase vulnerability to risks like floods or water-mediated diseases, and on the uncertain, but potentially important effects of climate change. Hydrologic sciences have made substantive progress in areas where international collaboration has been strong. National and international assessments have relied on dialogue to bring together information that traditionally has been fragmented and available only locally. The hydrologic components of global change research in particular have benefited from international collaboration (e.g. Dialogue on Water and Climate (DWC) and Intergovernmental Panel on Climate Change (IPCC)).

Capacity Building

Capacity building in less developed countries involves the development of the capability to develop and operate advanced observations and prediction systems, independent data providers and programs of water and environmental management, or to improve their national services. Capacity building also includes education and research capabilities. Capacity building goals frequently include economic and related technological gaps between developing and developed countries. The contributions of science are of greatest value when the research provides techniques and information to deal with these obstacles. Increasingly, international science programs that have an element devoted to capacity building and technology transfer, in addition to improving scientific understanding, are the most effective in achieving their goals.

The following sections provide information on coordinating bodies, the umbrella hydrologic science programs, and the projects and activities that take place under each organization.

UMBRELLA PROGRAMS FOR HYDROLOGIC RESEARCH

The UN provides an institutional structure for initiation, coordination, and application of international science efforts. UN Water provides the institutional structure for hydrological research and related activities. Other international bodies, principally ICSU and its affiliated programs, provide additional nonpolitical, nongovernmental scientific perspectives. They identify critical scientific issues and develop frameworks for addressing them through coordinated programs. Additionally, a number of nongovernmental partnerships are using hydrologic sciences in their deliberations.

International Scientific Associations

ICSU, through its several member unions provides an umbrella for international scientific activities in numerous fields. A nongovernmental organization founded in 1931, ICSU has been bringing scientists from a wide spectrum of natural and human sciences together to address major international and interdisciplinary issues including hydrology and water since its founding. ICSU mobilizes the expertise and knowledge of the international scientific community, stimulates constructive debate, and acts as an authoritative independent voice for international science and scientists.

IAHS was founded in 1922 as a branch of the International Union of Geodesy and Geophysics (IUGG) to promote research in the hydrologic sciences. IAHS became part of ICSU when the organization was formed. Through its eight committees and working groups, IAHS provides hydrologic scientists with a forum for planning and reviewing scientific projects. The committees, many of which are discussed in more detail later in this section, include the International Commission on Surface Water (ICSW), the International Commission on Groundwater (ICGW), the International Commission on Continental Erosion (ICCE), the International Commission on Snow and Ice (ICSI), the International Commission on Water Quality (ICWQ), the International Commission on Water Resources Systems (ICWRS), the International Commission on Remote Sensing (ICRS), the International Commission on the Coupled Land–Atmosphere System (ICCLAS), and the International Commission on Tracers (ICT). In addition, IAHS frequently relies on Working Groups, such as the Hydrology 2020 Working Group, the IAHS/WMO Working Group for Global Energy and Water Cycle Experiment (GEWEX), and the Prediction in Ungauged Basins (PUB) Working Group to address emerging issues in hydrology. IAHS promotes scientific exchange by organizing major scientific conferences, distributing a newsletter, and publishing a series of "red books" of conference and symposium proceedings.

Under the ICSU umbrella, several programs have been initiated with a focus on understanding the changes that are occurring to the Earth System and the implications of these changes for global sustainability. These include the International Geosphere Biosphere Programme (IGBP), the International Human Dimensions Programme on Global Environmental Change (IHDP), and DIVERSITAS, a scientific program on global biodiversity. ICSU is also a sponsor, along with UNESCO and WMO, of the World Climate Research Programme (WCRP). In 2001, these global environmental programs formed a coordinated superstructure called the Earth System Sciences Partnership (ESSP), a

partnership that provides a mutually agreed mechanism for these programs to work together towards shared science objectives in areas of carbon, water, food security, and health. Hydrologic sciences are fundamental components of the ESSP programs.

The UN System

The UN system provides a framework for identifying issues that require information from scientific inquiry, establishing international goals, and initiating and coordinating implementation activities. As indicated below, many individual UN agencies are responsible for programs and projects that relate to water. Cooperation within the UN system on freshwater issues is coordinated through UN Water. Generally, UN agencies work with the appropriate national government agencies that play a critical role in determining how science can be applied to specific issues.

WMO was established in 1951 as a specialized agency of the UN. It facilitates international cooperation in planning and establishing networks for making meteorological, hydrological, and other observations and in promoting the rapid exchange of meteorological information, the standardization of meteorological observations, and the uniform publication of observations and statistics. It also furthers the application of meteorology to aviation, shipping, water problems, agriculture, and other human activities; promotes operational hydrology; and encourages research and training in meteorology. The main WMO program dealing with hydrology is the Hydrology and Water Resources Programme (HWRP) and its supervisory body is the Commission on Hydrology (CHy).

The International Hydrological Programme (IHP) of UNESCO addresses a wide range of priority water resources from physical, ecological, and socioeconomic perspectives. Its research is directed to improve the scientific and technological basis for the management of water resources, including the protection of the environment. IHP also maintains a growing number of centers of excellence, regional research and training centers, and directed research projects. Some of these are described more fully in later sections.

The Food and Agriculture Organization (FAO) of the UN was founded in 1945 with a mandate to raise levels of nutrition and standards of living, to improve agricultural productivity, and to better the condition of rural populations. In the realm of hydrologic science, the FAO cosponsors several observational, data management and research programs with other UN organizations, ICSU programs, and nongovernmental organizations.

The United Nations Environmental Programme (UNEP), which was established in 1972, relies on hydrologic and other environmental sciences to address problems of the human environment. UNEP supports the Global International Waters Assessment (GIWA), the Global Environmental Monitoring System (GEMS/Water) Freshwater Quality System, the International Environmental Technology Center (IETC), and the World Conservation Modeling Center (WCMC).

Joint and Multisponsor Programs

Since its inception in 1979, the World Climate Programme (WCP) has been cosponsored by WMO, UNEP, UNESCO Intergovernmental Oceanographic Commission (IOC), and ICSU. The WCP includes four components: the World Climate Data and Monitoring Programme, World Climate Impact Assessment and Response Strategies Programme, World Climate Applications and Services Programme, and WCRP.

WCRP was established in 1980 to develop the fundamental scientific understanding of the physical climate system and climate processes (including hydrologic processes) needed to determine the extent to which climate can be predicted and the extent of humans' influence on climate (WCRP, 1984). One of WCRP's four major projects, GEWEX, focuses on closing water and energy budgets on global and regional scales and on enhancing the value of climate information for water resource management. WCRP research also provides the scientific foundation for meeting the research challenges posed in international policy frameworks such as the United Nations Framework Convention on Climate Change (UNFCCC) and Agenda 21.

Other components of the UN system that support water science and related activities include: United Nations Development Programme (UNDP), United Nations International Children's Emergency Fund (UNICEF), United Nations University (UNU), World Health Organization (WHO), International Atomic Energy Agency (IAEA), and the World Bank. UN Regional commissions with responsibilities in water include: the Economic Commission for Europe (ECE), the Economic Commission for Africa (ECA), the Economic and Social Commission for West Asia (ESCWA), the Economic Commission for Asia and the Pacific (ESCAP), and the Economic Commission for Latin America (ECLAC). Specific issue-focused conventions with administrative bodies established by the UN include the International Strategy for Disaster Reduction (ISDR), United Nations Framework Convention on Climate Change (UNFCCC), Convention on Biological Diversity (CBD), and United Nations Convention to Combat Desertification (UNCCD). The water-related activities of many of these are described in the following sections.

Partnerships for Research

Increasingly, nongovernmental organizations are partnering with governments and intergovernmental organizations that develop or apply the results of hydrological research in

order to achieve important goals in the area of water. A very brief introduction to these partnerships appears below.

The World Bank, UNDP, and the Swedish International Development Agency (SIDA) created GWP in 1996. It provides a policy framework for coordinating financial, technical, policy, and human resources to address critical issues related to sustainable water management and it has attracted many partners globally.

The WWC was established in 1996 as a nonprofit, nongovernmental umbrella organization dedicated to strengthening international and national efforts to develop methods for improved management of the world's water resources. It organizes a triennial conference known as the *World Water Forum* (WWF) and an associated Ministerial conference to bring attention to water issues.

The DWC was initiated in the Netherlands in 2001 to improve the capacity in water resources management to cope with the impacts of increasing variability of the world's climate. It established a global platform through which policy makers and water resources managers have better access to, and make better use of, information generated by climatologists and meteorologists. In 2003, it reestablished itself as the Cooperative Programme on Water and Climate (CPWC).

Regional intergovernmental organizations facilitate international hydrologic research at the regional level. The ESSP's Asia–Pacific Network for Global Change Research (APN) is an intergovernmental network for the promotion of global change research and links between science and policy making in the Asia–Pacific Region. The Inter-American Institute for Global Change Research (IAI) is an intergovernmental organization supported by 19 countries in the Americas that carries out work related to water resources and environmental issues.

PROGRAMS FOCUSED ON OBSERVATIONS AND DATA MANAGEMENT

Historically, the understanding of physical water cycle processes has been derived from observations at the watershed or catchment scale. As previously discussed, with the advent of satellites and their ability to observe large-scale hydrologic phenomena, the water cycle is increasingly being viewed as a global system. Furthermore, water cycle processes at all scales are fundamental to the climate system; consequently, the hydrologic sciences are an important aspect of climate research. Recognition of this fact has led to new requirements for, and approaches to, global observations. In addition, global and regional hydrological models are needed to support climate studies, and development of these models requires specialized data sets. Most national data programs are maintained for reasons other than global-scale sustainability or environmental understanding. In particular, they usually are designed for national purposes that include planning, national prediction services, developing local design standards, and monitoring the implementation of policies and standards. Yet, transboundary policies and management rely on continued access to multinational data and products to provide baseline information on water supply and the factors that may influence future water availability.

In order to address data issues, many international organizations with programs focused on water-related issues must include observations, data collection, and data management among their functions. For example, UN agencies use data extensively in their research programs to contribute directly to research for developing scientific understanding of the physical systems (e.g. WCRP), while others oversee the implementation and maintenance of observing systems and promote the sharing of data and technology transfer on a global basis (e.g. the Global Observing System (GOS) components: see Section "Monitoring and observational programs").

Experiments and Field Campaigns

Some international programs develop global hydrometeorological data sets for use in water management, precipitation forecasting, and climate monitoring. The potential of this approach has increased significantly with the availability of global data sets from polar-orbiting satellites.

In 1990, the WCRP launched GEWEX with the goal of determining energy and water budgets for the globe (WCRP, 1990). Individually existing projects and newly initiated activities were drawn together to ensure that the most appropriate satellite data were used to produce global water vapor, cloud, and precipitation fields. In order to relate these data sets to regional water problems and to obtain regional data sets for evaluating the satellite products, Continental Scale Experiments (CSEs) were established in a number of large basins around the world. Over the past decade, GEWEX scientists have acquired research data sets to examine a range of land, surface hydrology, and atmospheric processes. The GEWEX CSEs include the GEWEX Americas Prediction Project (GAPP) in the United States, the Baltic Sea Experiment (BALTEX), the GEWEX Asian Monsoon Experiment (GAME) in East Asia, the Large-scale Biosphere Atmosphere Experiment in Amazonia (LBA) primarily in Brazil, the Mackenzie GEWEX Study (MAGS) in Canada, the Murray Darling Basin (MDB) in Australia, the La Plata Basin (LPB) in South America and, most recently, the African Monsoon Multidisciplinary Analysis (AMMA) in west Africa. Data from these regional studies contribute to algorithm development and testing, and provide a baseline for evaluating the reliability of global products. Research in these programs contribute to land surface and hydrological process understanding and to the GEWEX Water Resources Application Project (WRAP).

An important and relatively new component of the WCRP is the Coordinated Enhanced Observing Period (CEOP) Phase I. This is a cooperative effort carried out to compile global-scale continental data sets for the 2001–2004 time period using the new generation of satellites as well as operational satellite data products, reference site data, and model derived data. The data sets, collected in different CSE areas representing different hydrologic regimes and supplemented with model output and satellite data, enable modelers to test the transferability of parameterizations to other, unsampled but hydrologically similar locations. The goal is to improve the prediction of climate variability and its effects on water resources in various regions of the world. Details on CEOP are provided in Stewart *et al.* (2001) and Koike (2004).

Monitoring and Observational Programs

Observational programs continue to play a central role in the hydrological sciences. Data is needed for the analysis, development, and evaluation of models required to address water issues related to consumption, sanitation, irrigation, hydropower production, water transport, floods, and droughts.

Observational networks are maintained by national governments to support objectives that may be national in scope, such as maintaining national forecast services and monitoring the state of national resources. Data from these national networks are the foundation of many global observing programs. For example, the WMO supports the collection and analysis of data from national networks of hydrological and meteorological stations and promotes data sharing. Historically, the ability to measure streamflow, surface water storage, and precipitation has been the basis for the development of engineering criteria for dams, reservoirs, and other water works. In addition, the ability to analyze water samples for chemical and biological constituents has enabled environmental regulation programs.

The planning of networks for different purposes has taken place relatively independently, with many meteorological measurements taken at airports and streamflow measurements taken at accessible areas along riverbanks. During the past decade, a great deal of effort has been directed toward more integrated network planning. The following review of organizations addressing measurement needs discusses those focusing on *in situ* measurement first, followed by activities that are bringing both *in situ* and space-based measurements into a common framework.

In Situ Measurements

The GOS programs, which include the Global Climate Observing System (GCOS), Global Terrestrial Observing System (GTOS), and Global Ocean Observing System (GOOS), were established in the early 1990s. They address the data needs of resource managers, scientists, and policy makers to monitor the Earth system and to provide a capability to monitor the Earth's climate. The purpose of GCOS is to ensure that the comprehensive, long-term observations and information (atmospheric, oceanic, terrestrial, and cryospheric) needed to improve our capability to detect, predict, and assess climate variability and change are obtained and made available to all potential users. GCOS facilitates the coordination and collaboration of national and international observational programs. Through a strategy of observations, modeling, and analysis of terrestrial ecosystems, GTOS focuses on the very diverse surface observational systems needed for land areas. It was established to provide necessary data on surface freshwater hydrology (including the cryosphere and nonconfined subsurface hydrology) and other land processes to detect, quantify, and give early warning of changes in terrestrial ecosystems, with the ultimate goal of supporting sustainable development and improving human welfare.

GCOS, GTOS, and Global Terrestrial Network for Hydrology (GTN-H) receive advice on data needs from a joint scientific advisory panel known as the *Terrestrial Observation Panel for Climate* (TOPC). On the basis of a meeting of experts (reported by Cihlar *et al.*, 2000), a global hydrological observation network for climate was established to share hydrologic information that would meet policy and science requirements in the area of climate research. This network, known as the *Global Terrestrial Network for Hydrology* (GTN-H), has now been adopted as an active international project under the sponsorship of GTOS, GCOS, and WMO.

Many real-time hydrometeorological data are acquired through the WMO's World Weather Watch (WWW) Programme. This global monitoring and assessment program maintains a meteorological database and distributes data and products including basic meteorological, hydrologic, oceanographic, and other environmental data and weather analysis, warnings, and forecasts.

The World Hydrological Cycle Observing System (WHYCOS) is a developing program based, in concept, on a global network of reference stations that transmit hydrological and meteorological data in near-real time, via satellites, to national hydrologic and regional centers and regional forecast centers. These data enable the provision of constantly updated national and regionally distributed databases of consistently high quality. WHYCOS supplements existing hydrological observing networks and supports the establishment and enhancement of information systems that can supply reliable water-related data to resource planners, decision makers, scientists, and the general public. It produces consistent regional data sets for use in preparing products for water resources assessment and management. WHYCOS is developing regional components known as *regional HYCOSs* to meet the regional priorities expressed by the

participating countries. In its initial phase WHYCOS has focused on establishing components in international river basins, in the catchment areas of enclosed seas, and in regions of Africa that are poorly served with hydrological information.

The Global Network of Isotopes in Precipitation (GNIP) was initiated in 1958 by the IAEA in collaboration with WMO. The objective of the program is to collect systematic data on isotope content of precipitation on a global scale. Regular intercomparisons among the national laboratories using standards developed by the United States Geological Survey (USGS) and statistical evaluations of the data are organized by the IAEA. The IAEA also sponsors the monitoring and analysis of isotopes in surface and groundwater, thereby enabling climatological (interpretation of paleorecords), atmospheric (validation of global circulation models), and hydrological (large regional and global-scale water balances) studies. An IAEA/UNESCO Joint International Isotopes in Hydrology Program (JIIHP) was initiated in 2003 to improve implementation and coordination of the isotope hydrology programs of both agencies. Integrated isotope databases allow for the gathering, storage, and dissemination of isotope, chemical, hydrogeological, and geographical data of water studies around the world.

A project within the International Shared Aquifer Resources Management (ISARM) program, jointly sponsored by the International Association of Hydrogeologists (IAH), UNESCO, FAO, and others, is developing pilot groundwater monitoring networks for selected aquifers that extend across country borders. UNESCO IHP initiated an inventory of groundwater contamination through its National Committees and Regional Offices of Science and Technology, and it has been exploring the need for a consistent groundwater monitoring program.

Satellite Capabilities

While slow to penetrate the field of hydrology, satellites are now making major contributions to the field. This is particularly true for the last decade, when the number of Earth-observing satellites increased markedly and hydrological research began to make more use of satellite products. For example, hydrology has made extensive use of precipitation data from satellites, and water management recognizes the great potential of the development of high-resolution precipitation products. In addition, development of high-resolution global land surface data sets has provided an incentive to hydrological modelers to develop distributed hydrological models that can take full advantage of these products. Of course, calibrating these new models is one of the major challenges facing hydrology today.

Currently, space agencies have developed a range of sensing technologies that provide new ways of looking at the Earth and new opportunities for deriving hydrological variables. For example, the planned suite of research satellites includes gravimetric measurements that can provide insight on groundwater and surface water levels, microwave sensors that can provide information on soil moisture, and radars, synthetic aperture radars, and surface lidars that can provide high-resolution information on topography and vegetation cover. Space agencies that are taking the lead in dealing with these satellite issues include the US National Aeronautics and Space Administration (NASA), European Space Agency (ESA), Japan Aerospace Exploration Agency (JAXA), and the Chinese Space Agency (CSA). NASA provides funding for a national research program to exploit the development of these satellites and for the assessment of uses of the data they provide. While other space agencies do not fund hydrological research, they all dialogue with the research community and support research by making data available.

The Committee on Earth Observation Satellites (CEOS) is an international organization charged with coordinating international civil Earth-observing missions and the interaction of these programs with users of satellite data worldwide. CEOS ensures that critical scientific questions relating to Earth observation and global change are covered, and satellite missions do not unnecessarily overlap each other.

Integrated Satellite/In Situ Measurements Activities

The Integrated Global Observing Strategy – Partnership (IGOS-P) is a partnership among international organizations, space agencies, the scientific community, and international research programs. IGOS-P was initiated in 1998 to plan and coordinate observations of the planetary environment, including *in situ* and space-based measurements. The IGOS Partners have adopted a thematic approach with joint planning activities to address particular categories, crosscutting themes, or domains of observations, such as oceans, disaster management, or carbon storage and cycling. In 2003, the IGOS Partners adopted an Integrated Global Water Cycle Observations (IGWCO) theme that provides a framework for water-related observations. Initially, coordination activities are being directed at three priorities: precipitation, surface hydrology, and water resource applications such as irrigation. The theme also provides an umbrella for CEOP and its links with the space agencies. The CEOS Strategic Implementation Team leads the development of the space component of IGOS-P while the GOS programs and their sponsors are leading the development of an *in situ* component.

In 2003, at the first Earth Observing Summit, nations and international environmental programs launched the development of a new observing system paradigm known as the *Global Earth Observing System of Systems* (GEOSS). At the time of this writing, more than 60 nations have joined this effort along with a number of international programs and

the European Union. Participating nations have affirmed the need for timely, high-quality, long-term, global information as a basis for sound decision making, and have committed themselves to develop a comprehensive, coordinated, and sustained Earth observation system. The system builds on the work of IGOS-P and other international observational programs.

Data Products, Distribution, and Storage

An important function of international programs is the production of consistent data sets that can be used with confidence in all parts of the world. For example, GEWEX, through its Global Radiation Panel (GRP), produces a range of global water cycle data sets from satellite observations. One product that is used extensively is the Global Precipitation Climatology Project (GPCP) series of monthly precipitation maps. These data sets are being used to validate general circulation and climate models, to study the global hydrological cycle, and to diagnose the variability of the global climate system. These studies have been aided by extensive efforts to calibrate data sets by eliminating the effects of factors, such as orbital drift, that change the intensity of the signal arriving at the surface.

A number of data and information centers have been established to archive global data products and make them available to all nations. In the past, international programs have encouraged the adoption of international standards for measurement and data processing methodologies, data archival and exchange, and metadata documentation. Consequently, many nations support centers for collecting, standardizing, and distributing global environmental data. While distributed archives and networks accessed through a web-based portal are becoming more common, centralized data centers such as the Global Runoff Data Center (GRDC) continue to play an important role.

Data Centers

The GRDC was formed under the auspices of WCRP and WMO in 1988. GRDC focuses on the acquisition and dissemination of river discharge data on a global scale and the provision of data products and specialized services for the research community, water managers, and water-related programs of the specialized agencies of the UN.

The Global Precipitation Climatology Center (GPCC) is the central *in situ* data component of the GPCP and fulfills other data functions in support of WCRP and WMO. The infrastructure for implementing international data management activities often emerges from activities supported by national governments. The GPCC is an example of this strategy. Specific functions of the GPCC are the collection of rain-gauge-measured precipitation data worldwide, quality control of these data, and calculation of areal mean totals based on the conventionally measured data over land.

For the jointly sponsored GOS programs, a Global Observing Systems Information Center (GOSIC) was established at the College of Marine Studies of the University of Delaware, US, to provide a single entry point for users seeking data and information produced by the global observing systems GCOS, GOOS, and GTOS.

Data Programs

Within the UN system, the HWRP of WMO provides for the collection and analysis of hydrological data as a basis for assessing and managing freshwater resources. The activities under the HWRP concentrate on the measurement of basic hydrological elements from networks of hydrological and meteorological stations; the collection, processing, storage, retrieval, and publication of hydrological data, including data on the quantity and quality of both surface water and groundwater; the provision of such data and related information for use in planning and operating water resources projects; and the installation and operation of hydrological forecasting systems. The program also promotes increased collaboration between national hydrologic services and national meteorological services, particularly in the provision of timely and accurate hydrological forecasts.

The GEMS/Water Programme, initiated by UNEP in 1977, provides authoritative information on the state and trends of inland water quality over the globe. These data are required to support global environmental assessments and decision-making processes. More than 100 countries participate in GEMS/Water. Data records extend from 1977 to the present. Since 1978, GEMS/Water has been hosted at Environment Canada's National Water Research Institute.

Socioeconomic data and data on water use are needed to support water policy decisions. Examples of products that are available include FAO's Cropwat and Climwat programs (described in Section "Water and Food") and WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP), which maintains a database containing water supply and sanitation coverage estimates, along with all the household survey data from which these estimates were derived.

PROGRAMS FOCUSED ON PROCESS AND MODELING STUDIES AND BASIC SCIENCE

The concept of the global water cycle links the physical components of the hydrological cycle, including surface water, groundwater, and atmospheric variables. This concept, which served as the basis of a comprehensive review by Hornberger *et al.* (2001), also has been used as a basis for defining the requirements for water cycle variables (IGOS, 2004). The following section reviews hydrologic research programs that address this cycle as well as the needs of the water management community.

Issues Involved in Hydrological Research

Surface Water

The ICSW of IAHS promotes research in surface water hydrology and its interaction with other aspects of the hydrological cycle. Priority issues include flood and drought prediction, mitigation, and forecasting, but increasing priority is given to interdisciplinary research, including in-stream ecology, wetland ecology, hydrology and health, and knowledge building to reduce international conflict in water.

Initiated in 2003, the Working Group on PUB is a new IAHS program focused on improving the accuracy of predictions in ungauged river basins. PUB is expected to provide an indication of the value of existing and planned measurement networks and new satellite missions, as well as to advance understanding of how river basins function in various parts of the world. PUB functions as a loosely coordinated global network of groups of experts associated with particular regions, sectors, and disciplines.

GAPP supports the Hydrological Ensemble Prediction Experiment (HEPEX) and the Model Parameter Estimation Experiment (MOPEX), which investigates techniques for the *a priori* estimation of the parameters used in hydrologic and coupled models. MOPEX is developing a comprehensive database that contains many years of historical hydrometeorological time series data for select basins and hydrologic parameters for basins worldwide.

Groundwater

The IAHS ICGW focuses on groundwater hydrology, including the scientific basis for groundwater resource assessment and groundwater management, and on new technologies and methodologies that are useful in groundwater quality studies. ICGW working groups deal with groundwater quality, environmental risk assessment, numerical modeling, innovative field and laboratory measurement techniques, advanced field data interpretation methods, characterization of heterogeneous media, and management of the uncertainty of groundwater resources.

Since 1978, the International Ground Water Modeling Center (IGWMC) has advised on groundwater modeling problems and distributed groundwater modeling software, undertaken practical research on applied areas of groundwater hydrology and modeling, and provided technical assistance on problems related to groundwater modeling. It also supports and advances the appropriate use of quality-assured models and groundwater resource protection and management. In addition, the International Groundwater Resource Assessment Center (IGRAC), a UNESCO center of excellence, conducts research and provides expert advice on groundwater issues.

Water Quality

The ICWQ of IAHS deals with issues related to the water quality of hydrological systems, including assessment and management. The commission currently has two working groups: Nutrient Cycling and Management and Urban Water Quality. The commission works with UNESCO IHP on issues of water supply, water quality and human health, integrated assessment of water resources, changes in the water quality of receiving waters as a consequence of land-based activities, scaling issues in water quality, and development and protection of vulnerable environments.

Climate

Because of the continental to global nature of atmospheric processes, hydrology must be able to address the climate-hydrology interface at all scales up to the global level. Conversely, because of the profound implications of climate variability and change for water resources, some water programs have adopted climate issues at the local scale as a priority theme. International research on the climate–hydrology interface generally takes the form of multinational hydrologic research programs that interface with large-scale climate modeling projects.

GEWEX deals specifically with the role of the water cycle in climate. Its main thrust is the production and analysis of global and regional data sets to support climate model development and climate process research, as well as the analysis and modeling of the global water cycle. GEWEX contributes to prediction capabilities on the monthly to interannual time scales. One central research question for GEWEX revolves around the hypothesis that the global water cycle is accelerating and humans play a significant role in that process.

CLIVAR, which was launched in 1995, strives to describe, analyze, model, and predict climate variability at time scales from a season to a century. As part of WCRP, CLIVAR provides insights into the working of the climate system with a primary focus on the atmosphere and the ocean and their interactions. It also focuses on the development of prediction techniques that will allow prediction on seasonal to century time scales. CLIVAR also complements other parts of WCRP such as GEWEX, through its monsoon system studies, and IGBP, through its Earth system modeling.

Past Global Changes (PAGES) provides a unified framework for paleo-environmental research on past climates and extreme events such as floods and droughts by facilitating international cooperation. This coordination embraces paleo aspects of the interactive hydrologic, atmospheric, oceanic, chemical, and biological processes that regulate the Earth system, concentrating on those aspects that relate to potential future changes of importance to humans.

Global Continental PalaeoHydrology (GLOCOPH) is a commission of the International Union for Quaternary Research (INQUA) established to study global paleohydrology during the last 20 000 years. GLOCOPH studies characteristics of past hydrological cycles and their role

in past environmental changes useful for validating the response of global or regional climate models in different environmental settings and prediction of future environmental change.

Land Surface Processes

In 2002, IGBP adopted a new program structure in which terrestrial water issues are addressed by projects in the land, land-atmosphere, and land-ocean themes. Some of this work builds on the success of the Biological Aspects of the Hydrologic Cycle (BAHC) program, a major research element of the IGBP during the period 1988 to 2002. BAHC studies provided a basis for understanding the effects of land use changes on runoff and for assessing the consequences of climate change for water resources. The integrated land-atmosphere project, planned for launch in 2004, supersedes the existing land-based project, Global Change Terrestrial Ecosystems (GCTE). GCTE studied the effects of changes in climate, atmospheric composition, and land use on the structure and functioning of terrestrial ecosystems, and analyzed how these effects lead to feedbacks to the atmosphere and the physical climate system.

Under IGBP's new structure, terrestrial studies will be undertaken in the Integrated Land Ecosystem-Atmosphere Process Study (iLEAPS) as well as in the ongoing Land Use and Land Cover Change (LUCC) and Land Ocean Interactions in the Coastal Zone (LOICZ) projects. iLEAPS will provide an understanding of how interacting physical, chemical, and biological processes transport and transform energy and matter through the land-atmosphere interface. iLEAPS will focus on the processes of land-atmosphere exchange of energy and matter, emphasizing feedbacks and interactions between these two components of the Earth system. LUCC research addresses the problem of land use and land cover dynamics through comparative case study analysis, empirical observations and diagnostic models, and integrated regional and global modeling. Since 1993, LOICZ research has focused on areas where land, ocean, and atmosphere interact. While scientific research remains central to LOICZ, during the next phase there is an increased focus on integrated analysis and synthesis, including the human dimensions.

To obtain a better understanding of the global hydrologic cycle, comprehensive watershed, river basin, and continental-scale process studies are needed. If properly coordinated and planned, process studies can potentially provide observational data in a standardized (i.e. transferable) format and can also improve the understanding of various aspects of the hydrologic cycle. Examples of successful process studies include: the First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE) in the United States and the Hydrology-Atmosphere Pilot Experiment in the Sahel (HAPEX-SAHEL). These projects, carried out as part of the GEWEX ISLSCP experiment have provided a scientific base, professional relationships, and infrastructure for advancing international cooperation in surface hydrology.

ICCLAS (International Commission on Coupled Land-Atmosphere System) promotes the evaluation and usage of outputs generated by global/regional coupled land-atmosphere models for water resources applications, including the management of floods and droughts. Specific topics of study include: the hydrological impacts of future climate, regional hydrology of climate-sensitive regions, especially semiarid and arid areas; historic hydroclimatological variability of relevant regions, catchment and regional scales; hydroclimatically relevant decision-making and risk assessment; and coupled hydrometeorological modeling.

Snow/Ice

The Climate and Cryosphere (CliC) project, initiated by WCRP in March 2000, seeks to assess and quantify the impacts of climatic variability and change on the cryosphere and their feedbacks to the climate system, and to determine the stability of the global cryosphere. Hydrologic expertise is required to address many of its principal scientific questions, such as glacier contributions to global sea level change and the energy and water cycle in regions with land ice, snow cover, and frozen ground. CliC builds on the success of Arctic Climate System Study (ACSYS), which had a focus on the hydrometeorological processes controlling runoff into the Arctic Ocean.

The IAHS ICSI is concerned with the study of snow and ice in all its forms. ICSI has four general priorities: River, Lake and Sea Ice; Seasonal Snow Cover and Avalanches; Glaciers and Ice Sheets; and Ice as a Material (including atmospheric ice, permafrost, and extraterrestrial ice). ICSI working groups also oversee the Snow Models Intercomparison Project (SNOWMIP) that compares the structure and functioning of existing snow models; Snow and Climate Working Group (SCWG) that examines the relationships between snow, ice, atmosphere, and climate on a global scale; River and Lake Ice Working Group (RLIWG) that identifies and quantifies physical and chemical processes in ice-covered freshwater basins, and relates these to the ecology of the systems themselves (e.g. life cycles of fish); and Snow-Vegetation Interactions Working Group (SVIWG) that defines the nature of snow–vegetation interactions and feedback relationships. ICSI also is responsible for UNEP's World Glacier Monitoring Service (WGMS) and has several working groups addressing regional glaciers (Andes and Himalayas) and mass balance techniques for developing new methods for monitoring glaciers.

Extreme Events

The WMO's World Weather Research Programme (WWRP), which seeks to improve weather forecasts on 1- to 14-day time scales, launched The Observing System

Research and Predictability Experiment (THORPEX) to develop a system for the prediction of high impact weather such as hurricanes, severe thunderstorms and associated flash floods, dust storms, and tornados. THORPEX will develop assimilation systems and carry out observing system tests to verify and evaluate the value of experimental and operational remote sensing and *in situ* observing systems in improving forecasts.

Methodologies and Techniques:

IAHS Commissions that focus on methodology cut across the other IAHS Commissions. The ICRS focuses on the use of remote sensing, data transmission, and Geographic Information Systems (GIS) in applications related to hydrological sciences. The Commission currently has two sections, Remote Sensing and GIS. ICT, established in 1991, promotes the development and improvement of artificial and natural tracer methods, and the dissemination of tracer methods to the hydrological community.

Experiments

An important approach to international science involves the launch of focused projects that allow researchers to work together. (Data collection and management aspects of these experiments are described in Section "Experiments and Field Campaigns".) The GEWEX program undertook an internationally coordinated effort to establish a number of focused projects including CSEs. The GEWEX Hydrometeorology Panel (GHP) and its CSEs develop analyses of regional water and energy budgets, and seek to improve the prediction of hydrometeorological phenomena. Over the past decade, scientists involved with the GEWEX program have contributed to intensive studies of specific hydroclimatological regimes (Lawford *et al.*, 2004). These continental-scale river basin studies will provide improved observations and coupled land-atmosphere models. In general, the scientific developments from this research find their way into operational services through the national agencies responsible for flood prediction, weather forecasting, and water resource management.

The ESSP, formed in 2001, facilitates collaboration among the four ICSU global environmental change research programs (DIVERSITAS, IGBP, IHDP, and WCRP) for the integrated study of the Earth system, the changes that are occurring in the system, and the implications of these changes for global sustainability. The programs are planning collaborative activities on Earth system analysis and modeling, water (see Section "Hydrologic Process Studies"), carbon, food security, and health. The water study, known as the Global Water System Project (GWSP), will address how changes in the Global Water System (GWS) (e.g. the system of dams and reservoirs) have influenced patterns of runoff and contributed to possible long-term changes in climate. In addition, they sponsor START (Global Change System for Analysis, Research and Training), a program that carries out capacity building, regional research, and networking activities in Africa, Asia, and the Pacific. START currently is promoting and facilitating the development of a small set of integrated regional studies including LBA, AMMA, and the Monsoon Asia Integrated Regional Study (MAIRS).

Hydrological Contributions to Earth System Modeling

Precipitation and the hydrologic cycle components are major drivers in Earth system modeling. A major international Earth system modeling effort to which hydrologic science makes key contributions is the Global Analysis, Integration and Modelling (GAIM). GAIM's goal is to improve the understanding of how the Earth functions as a system through the use of both data and models. Because of its complexities, Earth system modeling presents a challenge to GAIM and other IGBP and WCRP projects. GAIM has traditionally focused its efforts on global biogeochemical cycles and their links to the hydrologic cycle and to the physical climate system, for natural systems as well as in the context of anthropogenic influences.

WCRP is developing the Coordinated Observing and Prediction of the Earth System (COPES) strategy with the goal of facilitating the prediction of climate and Earth System variability and change for use in an increasing range of practical applications of direct benefit to society. To achieve this goal, COPES will determine what aspects of the climate/Earth system are predictable and those that are not, at weekly, seasonal, interannual, and decadal through to century time scales. In these efforts, COPES will utilize improving observing systems, data assimilation techniques, and models of the climate/Earth system.

Major national efforts to develop Earth system models are under way. For example, in the United States the Earth System Modeling Framework (ESMF) group is developing modeling capabilities to address climate, numerical weather prediction, data assimilation, and other Earth science applications through a multiagency collaboration that includes many of the major national climate, weather, and data assimilation efforts. An ESMF collaborating partner is the Programme for Integrated Earth System Modeling (PRISM), funded by the European Union, which draws together modeling efforts in individual European countries. In Japan, the Frontier Research System for Global Change is carrying out a project with similar aims.

Hydrologic Process Studies

The scope of UNESCO's IHP is broadly defined and covers many categories of water science research and related activities. Its two primary study programs, Flow Regimes from

International Experimental and Network Data (FRIEND) and Hydrology for the Environment, Life and Policy (HELP) rely heavily on hydrological process studies. The purpose of IHP is to improve the scientific and technological basis for the development of methods and expertise for rational management of water resources including environmental protection, and to integrate developing countries into worldwide ventures of research and training. IHP was put in place at the end of the IHD. It has been developed as a series of multiyear programs with IHP-V ending in 2001 and IHP-VI commencing in 2002.

The FRIEND program was initiated in 1985 in Europe. By 2004, FRIEND had grown to involve research institutes, universities, and operational agencies from more than 90 countries. FRIEND is an international research program whose goal is to develop an understanding of the hydrologic variability among regions of the world. Its research program is a collection of international studies in regional hydrology aimed at improving knowledge of hydrological processes and flow regimes through the exchange of data, knowledge, and techniques. This is accomplished in a cooperative manner through mutual exchange of data, techniques, and knowledge among FRIEND scientific partners. FRIEND is active on all continents except North America, where its activities are only now being discussed.

In 2001, ESSP started planning the GWSP, which was formally launched in 2004. GWSP promotes research to improve the knowledge of the interactions of the GWS with the environment over a range of time and space scales. In this context, the GWS is defined as the totality of natural processes, biogeochemical and ecological interactions, human engineering and policy interventions, and domestic and industrial use, and recycling. GWS is seen as both a driver of environmental change and a recipient of the impacts of different external changes.

KNOWLEDGE TRANSFER: APPLICATIONS, ASSESSMENTS, APPLIED RESEARCH, AND OUTREACH

Research is one component of the complex, dynamic environment in which global water-related issues are addressed. International, national, and regional hydrologic research programs are turned to societal benefit through development, education, and institutional change. Effective knowledge transfer underpins these essential activities. Informing policy makers is a critical role for scientific programs. At the policy level, exercises such as the World Water Forum are important to ensure that substantive and reliable hydrological information is brought forward to help in setting the broader sustainable development agenda. Furthermore, getting the best available hydrological information to the right people at all levels of decision-making can have a significant impact on quality of life in the present and water sustainability in the future.

Applications

Scientists are encouraged to develop pathways whereby their research results can be used in more effective decision-making. UNESCO's HELP focuses on in-depth applications in individual basins by characterizing both environmental and anthropogenic/policy concerns and relating current scientific understanding to these needs. HELP, a joint UNESCO/WMO initiative, is led by UNESCO IHP. The HELP network of basins is expanding from a 25-basin pilot phase launched in 2000 to a much larger number of basins in its full implementation in 2004. From the technical perspective, the broad objectives of HELP are to strengthen field-oriented, experimental hydrology using drainage basins (on scales up to 10^4 to 10^6 km^2) in order to address physical (hydrological, climatological, ecological) and nonphysical (technical, sociological, economics, administrative, law) issues.

New techniques developed through research are transferred to service for society through a process referred to as *technology transfer*. The transfer of new technologies to operational services is an essential process, but is one that requires a great deal of planning and nurturing. Case studies that demonstrate the importance of information sharing and transfer are under way through the assessment activities of World Water Assessment Project (WWAP) and collaborative research activities such as GEWEX. Through the efforts of the WMO's HWRP and CHy, national hydrometeorological operational agencies are kept informed about new technologies. WMO programs that facilitate the transfer of models, technologies, and data among nations also make an important contribution to these efforts.

Decision makers frequently may be reluctant to apply the latest research in their decision processes. For example, water resource managers must consider legal, institutional, and cultural factors as well as the timing and form of information, equipment and system constraints, and the risks and liabilities of deviating from an established practice. In order to design useful applications, studies are needed on the use of information in management and policy decisions related to water. IHDP, which was launched in 1990, promotes study of the institutional and human factors that affect society's response to global change. Within the domain of water science, IHDP considers issues such as the role of governance in water management and the implications of private versus public ownership of water. Studies of these issues are important for understanding the context for hydrological research applications.

Frequently, programs that apply the results of water research focus on the needs of specific sectors or address a specific subset of water-related issues. The following

subsections discuss projects that focus on critical information needs across a range of sectors and issues.

Water and Development

National and regional economies often depend on balancing the supply and demand for water. In many regions, the decreasing amounts of uncommitted water constrain economic development. (See e.g. Boehmer *et al.*, 2001). Nations and regions constrained in this way represent challenges to global development, stability, and sustainability; therefore, issues of water for development are global concerns that require support from science.

UNDP operates in 166 countries, where it helps to strengthen national capacities to alleviate poverty. Resolving water issues is central to poverty alleviation and human and ecosystem health. Key concerns include combating desertification and helping vulnerable populations adapt to climate change. UNDP advances its programmatic goals in these areas by applying hydrologic science in collaboration with other international bodies.

Water and Health

Clean water is critical for life and health, and the supply of safe water for human consumption is a primary social concern. By the current estimate, water-related illness accounts for more than 2.2 million deaths per year worldwide (UNESCO-WWAP, 2003). The provision of clean water for drinking and sanitation services is a central theme in the MDGs.

The Water, Sanitation and Health Program (WSHP) of WHO addresses aspects of water, sanitation, and hygiene where the health burden is high, knowledge levels are low, and where interventions could make a major difference. Applied research in hydrologic sciences addresses issues related to drinking water quality, bathing water, water supply and sanitation monitoring, sanitation and hygiene development, water-associated pathogens, wastewater use, health in water resources development, and emerging issues in water and infectious diseases.

Water and Food

Water management is a critical element of food production, especially when agricultural irrigation is used to supplement inadequate rainfall. The Water Resources, Development and Management Service (AGLW) of the FAO addresses the sustainable use and conservation of water in agriculture. It assesses water resources, monitors agricultural use, assists in water policy formation, and promotes irrigated agriculture and efficient water use through management innovations, modernization, and institutional reforms. The FAO relies on research products and observations for its advisory services. It maintains a global information system to support irrigation management through Cropwat, a decision-support system that estimates water balance components such as evapotranspiration, crop water requirements, and crop irrigation requirements. Cropwat is used in combination with FAO's global climatic database to optimize irrigation scheduling and water allocation for various crops worldwide.

The Consultative Group on International Agricultural Research (CGIAR), created in 1971, supports a system of 16 Future Harvest Centers that work in more than 100 countries to use hydrologic and other sciences to help reduce hunger and poverty, improve human nutrition and health, and protect the environment. Through its Challenge Program on Water and Food, CGIAR sponsors research on knowledge and methods for growing more food with less water. This research has a geographical focus provided by Benchmark Basins.

The International Water Management Institute (IWMI), a Future Harvest Center, focuses on the sustainable use of water and land resources in agriculture and on the water needs of developing countries. IWMI also identifies priority water management and food security issues and develops, tests, and promotes management practices and tools that can be used to more effectively manage water and land resources and address water scarcity issues.

ESSP is addressing the link between water and food through the Global Environmental Change and Food Systems (GECAFS) project launched in 2004. Research results and databases resulting from this project will contribute to improving food provision and minimizing environmental impacts of food production, distribution, and waste.

Water, Weather, and Climate

For over 65 years, the WMO and its predecessor, the International Meteorological Organization, have supported national hydrological services, river basin authorities, and institutions responsible for water management in a wide range of activities through HWRP. HWRP promotes activities in operational hydrology and furthers close cooperation between meteorological and hydrological services. HWRP also promotes improvements in the capabilities of developing countries to independently assess their water resources on a continuous basis, to respond to threats of floods and droughts, and to meet the requirements for water for various uses. In addition, HWRP facilitates the sharing of new models, technologies, and data among national operational hydrometeorological agencies and services.

The WCP (described in Section "Joint and Multisponsor Programs") has undertaken a project that applies advances in data, climate knowledge, and prediction capabilities to limit the negative impacts of climate variability and improve planning. This project, the Climate Information and Prediction Services (CLIPS), includes an expert team on climate and hydrology that monitors scientific developments and provides guidance on the use of climate predictions in water resources and flood management systems. Its main activities are training, demonstration/pilot

projects, liaison with research programs, and networking. WCP-Water also undertakes projects that utilize results from hydrological research.

Water and Environment

Water is a central determinant of ecosystem health and a major medium of transport for pollution. International programs in the hydrologic sciences that address environmental issues have multiplied in the past two decades. Since its inception in ICSU in approximately 1994, the SCOWAR (Scientific Committee on Water Resources) provided a scientific focus for addressing water issues. In 1996–1997, the committee narrowed its focus to the environmental aspects of water management. SCOWAR identified knowledge gaps and used existing databases to make ecological predictions possible. GWSP (described in Section "Hydrologic Process Studies") replaced SCOWAR in 2003.

DIVERSITAS, an international global environmental change research program under ICSU, promotes integrative biodiversity science by linking biological, hydrological, ecological, and social disciplines in order to produce an understanding of biodiversity loss, and to draw out the implications for conservation and sustainable biodiversity policies.

The UNEP Freshwater Unit supports the goals of environmentally sustainable development and enhanced environmental quality by promoting the integrated management and use of freshwater resources. Its activities include the development of curriculum and training material and the organization of international and regionally focused workshops. It uses hydrologic information and science in its environmental inventory, analysis, diagnosis, and action planning activities, and as a basis for promoting appropriate technologies.

Since its inception in 1991, the Global Environment Facility (GEF), administered by the World Bank in partnership with UNDP and UNEP, has provided grants to developing countries for water projects and activities, including transboundary water projects (e.g. groundwater, land degradation, and persistent organic pollutants) that protect the global environment and promote sustainable livelihoods. Many of these projects have had a substantial hydrological research component.

Assessments

Assessments provide a baseline understanding of the current status of a resource from which the future is projected. They can be used to provide insight on the nature and role of change. Water research is needed to provide accurate, comprehensive, and reliable assessments. The World Water Assessment Project (WWAP) undertakes assessments of the state of the world's freshwater resources and ecosystems; identifies critical issues and problems; develops indicators and measures progress towards achieving sustainable use of water resources; helps countries develop their own assessment capacity; documents lessons learned; and publishes a World Water Development Report (WWDR) at regular intervals. In support of these activities, the WWAP develops data sets and supporting information technologies for data interpretation, comparative trend analyses, and modeling. Although the program primarily focuses on terrestrial freshwater, it also addresses problems in marine near-shore environments and coastal zone regions where waters receive pollution and sediments from land-based sources or are under threat from the flooding of rivers and potential sea level rise. WWAP is the flagship program of UN Water, which coordinates UN initiatives involving freshwater.

IPCC carries out comprehensive climate assessments that cover the water-climate links. Established by WMO and the UNEP in 1988, the IPCC assesses the scientific understanding of human-induced climate change, its potential impacts, and options for adaptation and mitigation. IPCC assessments are based mainly on published scientific/technical literature. Many of the studies of the effects of climate on water resources, such as the US National Assessment of the potential consequences of climate variability and change on water (Gleick and Adams, 2000) have drawn heavily from the IPCC assessments.

Other UNEP-sponsored assessment programs include the Global Resource Information Database (GRID) and GIWA. GRID was established in 1985 as part of GEMS (described in Section "Data Programs"). It is a network of centers that collect and disseminate information on the state of natural resources worldwide through the open exchange of global and regional environmental geo-referenced data sets. Twelve cooperating GRID centers contribute data to this database system. GIWA was initiated in 1999 as a four-year project to produce a comprehensive and integrated global assessment of international waters. It has focused on the ecological status of 66 water areas in the world and the causes of their environmental problems.

IGRAC is part of a family of global information centers and GOSs for improving the understanding and management of freshwater resources. IGRAC, which was launched in 1999 by UNESCO, WMO, and the Netherlands, is also one of the UNESCO global centers of excellence. It is developing an interactive Global Groundwater Information System and groundwater assessment guidelines and protocols.

Nongovernmental institutes also undertake assessment activities with global scope that rely on water research. The World Resources Report (WRR), a biennial report series started in 1986, produced by the World Resource Institute and its partners, UNDP, UNEP, and the World Bank, assesses global environmental trends including, agriculture and food, coastal, marine and inland waters, freshwater, and biodiversity. Other reviews of the state of global and

regional water resources include: the biennial series, *The World's Water,* produced by the Pacific Institute (Gleick *et al.*, 2002), and the Worldwatch Institute's "Pilot Analysis of Global Freshwater Ecosystems" (Revenga *et al.*, 2000).

Applied Research: Regional and National Research Institutes and Centers

Applied research makes the critical link between science and benefits to society, but the application process is not always straightforward. The distinction between basic and applied research frequently allows for large areas of overlap. The results of basic research can find immediate operational uses and the application of studies carried out at centers for applied research sometimes does not occur. This overlap is evident in the missions of many national and regional research centers and programs.

UNESCO has established regional centers for research and training that provide facilities for capacity building, training, and research focused on regional needs. The following is a list of existing centers that focus on the area of water science:

- International Research and Training Center on Erosion and Sedimentation (IRTCES), Beijing, China;
- Regional Humid Tropics Hydrology and Water Resources Center for South-East Asia and the Pacific (WRCSEAP), Kuala Lumpur, Malaysia;
- Regional Water Centre for the Humid Tropics of Latin America and the Caribbean (CATHALAC), Panama City, Panama;
- Regional Center on Urban Water Management (RCUWM), Teheran, Islamic Republic of Iran; and
- Regional Center for Training and Water Studies of Arid and Semi-arid Zones (RCTWS), Egypt.

Several new centers have been proposed and are in various stages of development. These include a Regional Center for the Management of Shared Groundwater Resources, Utrecht, The Netherlands (under the auspices of UNESCO and WMO); a Regional Center on Urban Water Management, Santafé de Bogotá; a Regional enter for Ecohydrology, Warsaw, Poland; and a Regional Center on Drought in Africa, South Africa.

Many of the research and information programs that address the problems and opportunities outlined in the sections above on "Water and Development", "Water and Health", "Water and Food", "Water, Weather, and Climate", and "Water and Environment" are carried out by agencies through centers supported by national governments. The following partial list identifies some of these principal centers.

Hydrologic research at the Centre for Ecology and Hydrology (CEH) builds on the former Institute of Hydrology (IH), which led hydrologic research in the United Kingdom since its inception in 1968. Research undertaken at CEH extends from water and energy fluxes to processes such as evaporation, interception, and infiltration, to mathematical modeling of the hydrological cycle and chemical processes above and below ground.

Established in 1974, the National Institute for Environmental Studies (NIES) in Japan conducts basic research on environmental sciences, including water sciences, from the perspective of various disciplines, such as physics, chemistry, biology, health sciences, engineering, and economics. It supports the Center for Global Environmental Research (GER), which provides research support facilities and monitoring data for integrated environmental research.

The Stockholm Environment Institute (SEI) has conducted water research since 1989. It developed the Water Evaluation and Assessment Planning Tool (WEAP) and has used it to develop scenarios that have advanced sustainable water development.

Commonwealth Scientific and Industrial Research Organization (Land and Water) (CSIRO) in Australia works with its partners to find scientific solutions to water problems through multidisciplinary research, applying knowledge and research results from hydrology, hydrogeology, soil science, ecology, atmospheric science, remote sensing, and the social sciences.

The Global Hydrology and Climate Center (GHCC) brings together the NASA Marshall Space Flight Center (MSFC), the State of Alabama's Space Science and Technology Alliance (SSTA), and the Universities Space Research Association (USRA) to undertake hydrologic research that contributes to understanding and predicting the Earth's global water cycle, its connections to climate variability and weather, and its interactions with society. The National Oceanic and Atmospheric Administration (NOAA) and NASA also sponsor individual university centers for studies of the application of satellite and climate data to water resources management.

The German Federal Ministry of Education and Research (BMBF) sponsors the Global Change in the Hydrologic Cycle (GLOWA) Program to develop simulation-tools and instruments as a basis for sustainable water management at the regional level (river basins of approx. 100 000 km^2). Science themes include natural variability of precipitation, variations caused by human activities, and their effect on the hydrological cycle; interactions between the hydrological cycle, the biosphere, and land use; and water availability and conflicting water uses. In addition to research, GLOWA projects train scientists in developing countries and develop mechanisms for knowledge transfer.

The National Water Research Institute of Environment Canada, which has branches in Burlington, Ontario, and Saskatoon, Saskatchewan, undertakes a wide range of water research including hydrology, hydrogeology, aquatic

ecology, and limnology. It specializes in hydrologic processes in the Great Lakes and the Arctic.

Outreach

As used here, the term outreach encompasses a range of activities from the conferences and newsletters of professional societies to fostering technical competence in developing countries by international organizations. It includes efforts to raise awareness in the general public and provide access to repositories of information, specialized training workshops, grassroots institution building, ministerial briefings, forums for discussion of controversial issues, and similar activities. Critical elements in outreach include information exchange, education/training, and capacity building.

Public outreach regarding the value of hydrologic research and the uses of water science is important for international water programs because they increase awareness of the benefits and results and promote use of tools developed or improved as a result of water research and build a constituency for research. Outreach activities in education prepare the human resources needed to support continued progress in water science. Along with technical training and institutional capacity building, they also increase the abilities of nations to take advantage of the benefits of research and provide more equal access to those benefits. Outreach within and among water scientists and other water professionals facilitates the exchange of information vital to scientific and technical progress.

Outreach is a component of many organizations and programs, but often it is most effective when carried out by groups that give priority to this activity. The following sections provide a sample of the range of programs, organizations, and networks committed to outreach activities.

Education, Training, and Capacity Building

Two academic institutions within the UN system specifically address education, training, and capacity building in the hydrologic sciences. UNESCO and the Institute for Infrastructural, Hydraulic and Environmental Engineering (IHE) Delft have recently established an Institute for Water Education (IWE) in the Netherlands to make the knowledge of water management more accessible worldwide. IHE has offered training to water professionals from developing countries since 1957 and has expanded its curriculum to encompass the multidisciplinary needs of integrated water resources management. The United Nations University-International Network on Water, Environment and Health (UNU-INWEH) was formed in 1996 to strengthen water management capacity through training, particularly of experts from developing countries, and to provide on-the-ground support for projects.

START (see also Section "Experiments") contributes to international hydrologic sciences by building regional research capacity and providing training for regional scientists. It promotes research in the regional origins and impacts of global environmental changes including changes in the water cycle. START is a system of interconnected regional research networks (RRNs), regional research centers (RRCs), and regional research sites (RRSs) being developed by the ESSP. Each of the RRCs serves as the information center for the RRN and provides for the synthesis of results from various disciplines concerned with global change phenomena. The RRSs are institutes in the region with specialized expertise that allows them to carry out research on core projects.

There are numerous international education, training, and capacity building activities associated with universities. For example, the Water, Engineering and Development Centre (WEDC), founded in 1971 at Loughborough University in the United Kingdom, offers education, training, research, and consultancy related to the planning, provision, and management of infrastructure for development in low- and middle-income countries. Some university-sponsored outreach programs focus on young students, such as the Global Rivers Environmental Education Network (GREEN), initiated in 1984 at the University of Michigan, which provides opportunities for youth to learn more about water quality at the local watershed scale and to contribute to local water management. Other activities have broader aims; the Universities Council on Water Resources (UCOWR) involves universities in the United States and throughout the world in education, research, and public service activities relating to water resources, including information support for policy development.

Information Exchange

Scientific and professional organizations, through regular meetings, newsletters, journals, and other communication mechanisms, traditionally have served as conduits for information exchange. At the same time, symposia, policy briefings, newsletters, and brochures, produced by organizations that operate at the science-policy interface, have provided the public with scientific information. With the spread of the Internet throughout the world, electronic networks are emerging as the major mechanism for information sharing. Web pages, electronic clearinghouses, and "bulletin boards" have joined traditional professional associations, advisory committees, workshops, and conferences in the information toolbox. In addition, programs aimed specifically at developing dialogues among scientists, policy makers, and stakeholders, such as the DWC have emerged, which take advantage of all these tools. The following list, while incomplete, does provide an overview of many organizations that facilitate the exchange of hydrologic and related information, often within an applications context.

Professional Associations

The IAHS is a professional association of scientists from the broad range of hydrologic sciences (see Section "International scientific associations" for details).

The International Water Association (IWA) is a global network of water professionals, spanning the continuum between research and practice, and covering all facets of the water cycle. Formerly the International Water Quality Association, the IWA focuses on the science and management of drinking water, wastewater, stormwater, and the conservation of water resources.

Since its formation in 1972, the International Water Resources Association (IWRA) has worked to improve water management worldwide through dialogue, education, and research. It actively promotes the sustainable management of water resources by improving the understanding of the physical, biological, chemical, institutional, and socioeconomic aspects of water.

The International Association for Environmental Hydrology (IAEH) is a worldwide association of environmental hydrologists that enables environmental hydrologists and engineers, especially in developing countries, to share technical information and exchange ideas related to hydrology and the environment.

IAH is a scientific and educational organization, founded in 1956, whose aims are to promote research on the understanding and management of groundwater.

Regional and Subspecialty Organizations and Associations

The International Network of Basin Organizations (INBO) facilitates exchanges of experiences and expertise among organizations interested in river basin management. It promotes sound water management, facilitates the implementation of basin management tools, and promotes information and training programs for people involved in water management.

The International Association of Hydraulic Engineering and Research (IAHR) promotes, advances and exchanges of knowledge of engineering techniques for water resources, river and coastal hydraulics, risk analysis, energy, environment, disaster prevention, and industrial processes. IAHR also operates the European Engineering Graduate School in Stuttgart, Germany.

The International Association for Sediment Water Science (IASWS) promotes, encourages, and recognizes excellence in scientific research related to sediments and their interactions with water and biota in fluvial, lacustrine, and marine systems. It also fosters collaborative research and dialogue among Earth scientists, biologists, chemists, and environmental engineers with an interest in sediment–water interactions.

The International Coordinating Committee on Reservoir Sedimentation (ICCORES), founded in 1992, brings together experts from water-related international associations to promote the combined action of various disciplines and experiences gained in different countries for addressing the problems of reservoir sedimentation.

The International Commission on Large Dams (ICOLD) provides a forum for the exchange of knowledge and experience in dam engineering. Founded in 1928, it supports the safe, efficient, economic, and environmentally friendly construction of dams.

The International Commission on Irrigation and Drainage (ICID), established in 1950, is dedicated to enhancing the worldwide supply of food and fiber by improving water and land management, especially irrigation, drainage, and flood management techniques.

The International Association of Theoretical and Applied Limnology (Societas Internationalis Limnologiae, SIL), which was founded in 1922, promotes and communicates new and emerging knowledge among limnologists to advance the understanding of inland aquatic ecosystems and their management.

Since its founding in 1999, the International Water History Association (IWHA) has established a network of experts to explore the complex processes shaping water resource use.

The Small Island Developing States (SIDS) Network (SIDSnet) connects 43 states in the Pacific, Caribbean, Atlantic, Indian Ocean, Mediterranean, and African island nations. It encourages the use of information and communication technologies to enable SIDS to attain sustainable development objectives. The project was launched in 1998 through UNDP's Sustainable Development Networking Programme (SDNP) and the Alliance of Small Island States (AOSIS).

The Inter-Islamic Network on Water Resources Development and Management (INWRDAM) is an intergovernmental, autonomous organization operating under the umbrella of the Standing Committee on Scientific and Technological Cooperation (COMSTECH) of the Organization of the Islamic Conference (OIC). Established in 1987, INWRDAM is hosted in Amman by the government of Jordan.

MEWIN (Middle East Water Information Network), founded in 1994, improves access to information and data on Middle East water resources and issues, and encourages cooperative planning, use, and management of the resource.

Public Information and Advocacy

The Pacific Institute for Studies in Development, Environment and Security has a Water and Sustainability Program that integrates several major projects to address freshwater problems across applications from agricultural production to ecosystem health. It publishes *The World's Water* (see Section "Assessments") and maintains an open-access database on which this biannual assessment is based.

Since its inception in 1974, the Worldwatch Institute has undertaken interdisciplinary research with a global focus on water and interactions among key environmental, social, and economic trends.

The World Wildlife Fund (WWR) – Living Waters Program provides background information on freshwater and WWF's objectives to safeguard freshwater systems.

The International Rivers Network (IRN) was initiated in 1985 by environmental activists to halt and reverse the degradation of river systems, promote sustainable alternatives to damming and channeling rivers, and increase awareness of the value of rivers and their importance for society.

LakeNet is a global network of individuals and organizations in more than 90 countries working for the conservation and sustainable management of lakes. The LakeNet Secretariat is a US-based nonprofit organization.

Electronic Networks and Web Portals

The WaterWeb Consortium is dedicated to improving communications and information access to help address the world's water problems. It maintains a listing of web links containing web-based data inventories and local network portals.

The UNESCO Water Portal provides links to the current UNESCO and UNESCO-led programs on freshwater and serves as an interactive entry point for websites of water-related organizations, including water links, water events, learning modules, and other on-line resources.

The Global Applied Research Network (GARNET) facilitates the sharing of applied research information between researchers working in the water and sanitation sector.

The International Waters Learning Exchange and Resource Network (IW: LEARN) mission is to build an Internet-based "global knowledge community" to protect, restore, and sustain the world's aquifers, great lakes and river basins, coastal zones, seas, and oceans, and emphasizes building capacity among transboundary water resource projects.

SUMMARY

As outlined in the preceding pages, international hydrological sciences have followed a unique development path. Opportunities for international coordination have expanded greatly in the past two decades as world attention has focused on issues of environment, climate, and freshwater sustainability. New capabilities of satellite systems and the gains that can be realized from better integration of observation systems are driving the development of coordination mechanisms and integrating structures. This section considers some factors that will continue to affect the development of international hydrological sciences in the coming decades and considers emerging structures that may strengthen the coordination needed to realize its promise.

Challenges in the Coordination of International Science

There is a clear requirement for continued growth of international hydrologic sciences. Currently, several stable and highly respected organizations, such as the IAHS, bring together hydrologists from many different specialties and provide a forum for them to discuss the advances in their science. In addition, there are a number of organizations that focus on water policy and the use of hydrologic science to better understand specific issues. The emphasis of the UN on issues related to climate change and sustainable development has provided new opportunities for hydrology that supports large-scale understanding of the climate and the Earth system, or that supports the assessment, monitoring, and forecasting of water availability. In addition, the interest of the space sector in increasing the use of satellite data for water management is leading to new opportunities for hydrologic research. The broad interest of international programs in some branches of hydrologic science has supported growth in those sectors. However, as the competition for funds increases, it is unclear whether the past rate of growth can continue.

Hydrologic sciences have always been highly multidisciplinary. Some efforts have been made to define the boundaries of hydrologic science (e.g. NAS, 1991), but these efforts have been only partially successful. The range of disciplines required to work in an integrated way on water issues is growing, and the hydrologic sciences are drawing on a wider range of scientific groups to address complex problems. In addition, the hydrological sciences are commonly applied to decision-making in nonscientific arenas, where understanding of the decision context can be as influential as scientific understanding.

In addition, the hydrologic sciences must deal with scale issues associated with linking phenomena from the watershed (catchment) to global scales and from seconds to centuries. Global-scale hydrology and local hydrology require different approaches; for example, at the watershed scale, heterogeneities persist in ways not seen in oceanography or meteorology. When dealing with large-scale problems, it is necessary to summarize these heterogeneities by relatively simple parameterizations. Beyond the scientific challenge this implies, it introduces tensions within the hydrologic sciences because some hydrologists can feel uncomfortable that the parameterizations are not faithful to the underlying physics. On the other hand, parameterizations that can be used in coupled land-atmosphere models are necessary to ensure that hydrological processes are properly accounted for in climate models.

Other challenges that are pronounced in the hydrologic sciences include the essential reliance on *in situ* data collection and the issues of network maintenance, data standards, and access of such reliance entails, including the exchange of sensitive water-related data across national

borders. All these difficulties are at the root of many of the challenges faced by the international hydrological science community. Among the most pressing challenges are:

1. *Funding and prioritization:* There is a pressing need to convert competition to synergy and efficiency. Since World War II, the United States and other developed countries have tended to be strong players in international programs in which there is a major atmospheric component. However, the lack of US involvement in some mainstream programs for several decades, such as UNESCO, has led to a reduction of US international contributions in these program areas. In spite of this reduction, the United States has some excellent applications programs at the regional and national levels (e.g. NOAA's Regional Integrated Science and Assessment projects). The satellite agencies of the world, especially NASA, have been in the forefront of encouraging a continental and global-scale perspective of surface water distribution. The challenge is to forge a unified scientific basis for the study of global hydrological systems.
2. *Networks and access to data:* The need for a framework for water cycle science and observations is becoming urgent because of the erosion of national hydrometric observational networks and gaps in the data sets held by global data centers (e.g. GRDC). In many countries, routine hydrological observations are made that could be used for freshwater assessments, monitoring, and climate research. However, an international hydrological network that operates on a standardized set of agreed-upon procedures for data collection, dissemination, analysis, and use does not exist. Many countries have very sparse hydrometeorological measurements. Furthermore, some countries will not release their data, because of national data policy restrictions. The United States has been a leader and role model in making its hydrologic information available to all users (see e.g. **`http://www.water.usgs`**) at a time when even some developed countries (e.g. Canada and Belgium) have removed their hydrologic information from public domain. An overall observational strategy for hydrometric measurements over land must be developed and nations should be encouraged to make commitments to support it. The free and open access to hydrometric data should be adopted as a basic principle for this network, as required by most interpretations of WMO resolution (CgXIII Resolution #25 adopted in 1999). The lack of free and open data exchange is a significant barrier to the international science community in many areas, including climate modeling (see, for example, the discussion by the Ad Hoc Group on Global Water Data Sets of the International Association of Hydrological Sciences, 2001). Remote sensing methods, primarily employing satellites, are increasing the instantaneous availability of information globally. Remote sensing is sometimes viewed as the solution to these problems; however, *in situ* measurements also are critical for the development and validation of estimates derived from remote sensing data and the interpretation of their implications.
3. *Developing a hydrologic perspective:* The growth of a recognized voice for international hydrology has not been as rapid as it has been in the fields of meteorology and ocean science. While the UN has one agency for meteorology, at least 13 agencies have some responsibility for water. On the services side, the national hydrologic services have not had the same requirement for data sharing and joint approaches to research as the meteorological services have because predictions depended only on information from local and upstream watersheds rather than the entire globe. When basins extend across national boundaries, the sharing of data between the upstream and the downstream countries is limited in some cases. New prediction capabilities, advances in transferability of models, and new integrated approaches to water management are slowly changing this attitude.
4. *Strategies for integration:* Recently, Europe and other countries have adopted policies that will lead to a stronger emphasis on integrated river basin management and a need to bring together all aspects of water issues with land management and socioeconomic considerations. For example, water use is highly dependent on water quality, which in turn is closely tied to land use and land use changes. To be effective, future water policy will need to be integrated with land use policies. Integration will require surface water quantity, quality, and groundwater issues to be considered and resolved in a more comprehensive knowledge framework. Research and data collection activities will need to become more multidisciplinary and to expand their perspectives to provide an adequate base for this integrated approach, which in turn will lead to further demands for integration.
5. *Efficient support for water managers:* Increases in the application of automation, satellite data, assimilation and prediction techniques, and the Internet are leading to higher expectations for efficiency in managing water systems. Better integration of scientific understanding into decision processes is needed. Choices will be influenced by the values held by the national and regional governments overseeing the decision process. The assessment of the options requires inputs from the water science community and guidance on criteria for determining the advantage of one strategy over another.
6. *Bridging language gaps:* Hydrology is essentially an interdisciplinary science. It is becoming even more

complex as it is pressed to address high priority science questions in a global context. The languages of the individual disciplines concerned with the water cycle can facilitate communication among its specialists, but impede communication with others. As the needs force cooperation among an expanding range of disciplines, the broad, social challenge of communicating across disciplines grows. Some programs, such as the GWSP, are addressing this challenge explicitly through the development of a lexicon. Other strategies involve projects that demand direct interactions and the development of long-term relationships among scientists of different disciplines.

Long-term relationships also appear to be essential for science to have an impact on decision-making. If non-scientists and scientists speak different languages, as C.P. Snow suggested, the hydrologic sciences must develop more effective mechanisms for communicating with research users from other backgrounds and for applying scientific results to the solution of practical problems.

7. *Emerging structures:* In the area of environmental sciences, the need for greater integration has been realized and led to the formation of the ESSP in 2001. This partnership brings together expertise from climate, hydrology, biology, aquatics, and human dimensions (DIVERSITAS, IGBP, IHDP, and WCRP) to address priorities for humanity. While ESSP studies will involve macroscale hydrology, there are some aspects of hydrology and water science at the watershed scale that may find only limited involvement in such efforts. ESSP studies also will increase the demand for socioeconomic data in support of hydrological research. This raises the question, "Is there a need to develop a new interdisciplinary framework that will allow hydrologists who work at the conventional watershed scales to become more integrated into the global perspective on water problems, or does the current suite of associations, committees, organizations, and programs outlined in this article provide adequate coordination?" If a new framework is needed, what form should it take?

To deal with the growing realization that water will be one of the critical environmental issues of the twenty-first century, many programs have emerged that require hydrological expertise. Many of these programs are outside the traditional governmental framework for water organizations and involve partnerships with the private sector. It is not clear whether these partnerships will lead to different perspectives on water (e.g. changes in the view that water is a public good and each human should have access to a certain quantity of safe water as a basic human right). However, it is likely that these partnerships will also extend into the research area and influence future research priorities. The hydrologic research community is not entirely prepared for this possibility.

Appendix A ACRONYMS AND WEBSITES

ACSYS – Arctic Climate System Study

AGLW – Water Resources, Development and Management Service (FAO)
(**http://www.fao.org/ag/agl/aglw/**)

AMMA – African Monsoon Multidisciplinary Analysis
(**http://www.ofps.ucar.edu/amma/amma_summary.htm**)

AOSIS – Alliance of Small Island States
(**http://www.sidsnet.org/aosis/**)

APN – Asian Pacific Network for Global Change Research
(**http://www.apn.gr.jp/indexe.html**)

BAHC – Biological Aspects of the Hydrological Cycle
(**http://www.pik-potsdam.de/~bahc/**)

BALTEX – Baltic Sea Experiment
(**http://w3.gkss.de/baltex/**)

BMBF – German Federal Ministry of Education and Research

CATHALAC – Water Center for the Humid Tropics of Latin America and the Caribbean
(**http://www.cathalac.org/**)

CBD – Convention on Biological Diversity

CEH – Center for Ecology and Hydrology
(**http://www.nwl.ac.uk/ih/**)

CEOP – Coordinated Enhanced Observing Period
(**http://www.ceop.net** and **http://www.usask.ca/geography/MAGS/GHP/ceop_issues.html**)

CEOS – Committee for Earth Observation Satellites
(**http://www.ceos.org/**)

CGIAR – Consultative Group on International Agricultural Research
(**http://www.cgiar.org/** and **http://www.waterforfood.org/**).

CHy – Commission on Hydrology

CLIVAR – Climate Variability and Predictability
(**http://www.clivar.org/**)

CliC – Climate and Cryosphere Project
(**http://clic.npolar.no/**)

CLIPS – Climate Information and Predictions Services
(**http://www.wmo.ch/web/wcp/clips2001/html/**)

COMSTECH – Standing Committee on Scientific and Technological Cooperation
(**http://www.comstech.org.pk/**)

COPES – Climate Observing and Prediction in the Earth System

CPWC – Cooperative Programme on Water and Climate
(**http://www.waterandclimate.org/home.asp**)

CSA – Chinese Space Agency

CSD – Commission on Sustainable Development
CSE – Continental Scale Experiment (**http://www.gewex.com/cseslocation.html**).
CSIRO – Commonwealth Scientific and Industrial Research Organization (Land and Water: **http://www.clw.csiro.au/index_alt.html**)
DIVERSITAS (See: **http://www.diversitas-international.org/**)
DWC – Dialogue on Water and Climate (**http://www.waterandclimate.org/home.asp**)
ECA – Economic Commission for Africa
ECE – Economic Commission for Europe
ECLAC – Economic Commission for Latin America
ESA – European Space Agency
ESCAP – Economic Commission for Asia and the Pacific
ESCWA – Economic and Social Commission for West Asia
ESMF – Earth System Modeling Framework
ESSP – Earth System Sciences Partnership (**http://www.ess-p.org/**)
FAO – Food and Agriculture Organization (**http://www.fao.org/**)
FIFE – First ISLSCP Field Experiment (**http://www.esm.versar.com/fife/fifehome.htm**)
FRIEND – Flow Regimes from International Experimental and Network Design (**http://www.nwl.ac.uk/ih/www/research/bfriend.html**)
GAIM – Global Analysis, Integration and Modelling (**http://gaim.unh.edu/**)
GAME – GEWEX Asian Monsoon Experiment (**http://www.hyarc.nagoya-u.ac.jp/game/**)
GAPP – GEWEX Americas Prediction Project (**http://www.ogp.noaa.gov/mpe/gapp/**)
GARNET – Global Applied Research Network (**http://www.lboro.ac.uk/departments/cv/wedc/garnet/grntover.html**)
GCOS – Global Climate Observing System (**http:xs//www.wmo.ch/web/gcos/gcoshome.html**)
GCTE – Global Change Terrestrial Ecosystems (**http://www.gcte.org/**)
GECAFS – Global Environmental Change and Food Systems (**http://www.gecafs.org/**)
GEF – Global Environmental Facility (**http://www.gefweb.org/**)
GEMS – Global Environmental Monitoring System (**http://www.gemswater.org/**)
GEOSS – Global Earth Observing System of Systems (**http://earthobservations.org/default.asp**)
GER – Global Environmental Research
GEWEX – Global Energy and Water Cycle Experiment (**http://www.gewex.org/**)
GHCC – Global Hydrology and Climate Center (**http://www.ghcc.msfc.nasa.gov/**)
GHP – GEWEX Hydrometeorology Panel (**http://ecpc.ucsd.edu/projects/ghp/ghp.html**)
GIS – Geographic Information System
GIWA – Global International Waters Assessment (**http://www.giwa.net/**)
GLOCOPH – Global Continental PalaeoHydrology (**http://www.geodata.soton.ac.uk/glocoph/glocoph.html**)
GLOWA-Global Change in the Hydrological Cycle (**http://www.glowa.org/**)
GMS-Geostationary Meteorological Satellites
GNIP – Global Network of Isotopes in Precipitation (**http://isohis.iaea.org/**)
GOOS – Global Ocean Observing System (**http://ioc.unesco.org/goos/**)
GOS – Global Observing System (**http://www.gos.udel.edu/aboutgosic.htm**)
GOSIC – Global Observing Systems Information Center
GPCC – Global Precipitation Climatology Center (**http://www.dwd.de/en/FundE/Klima/KLIS/int/GPCC/GPCC.htm**)
GPCP – Global Precipitation Climatology Project (**http://cics.umd.edu/∼ yin/GPCP/**)
GRDC – Global Runoff Data Center (**http://www.bafg.de/grdc.htm**)
GREEN – Global Rivers Environmental Education Network (**http://www.green.org/**)
GRID – Global Resource Information Database (**http://www.unep.org/newdraft/unep/eia/ein/grid/home.htm**)
GRP – GEWEX Radiation Panel (**http://grp.giss.nasa.gov/**)
GTN-H – Global Terrestrial Network for Hydrology (**http://www.fao.org/gtos/gt-netHYD.html**)
GTOS – Global Terrestrial Observing System (**http://www.fao.org/gtos/**)
GWP – Global Water Partnership (**http://www.gwpforum.org/servlet/PSP**)
GWS – Global Water System
GWSP – Global Water System Project (**http://www.gwsp.org/**)
HAPEX – SAHEL - Hydrology-Atmosphere Pilot Experiment in the Sahel
HELP – Hydrology for the Environment, Life and Policy (**http://www.unesco.org/water/ihp/help/**)
HEPEX-Hydrological Ensemble Prediction Experiment
HOPC – Hydrology Observation Panel for Climate
HWRP – Hydrology and Water Resource Programme (**http://www.wmo.ch/web/homs/hwrpframes.html**)
IAEA – International Atomic Energy Agency (**http://www.iaea.org/**)

IAEH – International Association for Environmental Hydrology
(**http://www.hydroweb.com/**)
IAH – International Association of Hydrogeologists
(**http://www.iah.org/**)
IAHR – International Association of Hydraulic Engineering and Research
(**http://www.iahr.net/site/index.html**)
IAHS – International Association of Hydrological Sciences
(**http://www.cig.ensmp.fr/~iahs/**)
IAI – Inter-American Institute for Global Change Research
(**http://www.iai.int/**)
IASWS – International Association for Sediment Water Science
ICCE – International Commission on Continental Erosion
ICCLAS – International Commission on Coupled Land-Atmosphere System
ICCORES – International Coordinating Committee on Reservoir Sedimentation
ICGW – International Commission on Groundwater
(**http://www.iasws.com/**)
ICID – International Commission on Irrigation and Drainage
(**http://www.icid.org/index_e.html**)
ICOLD – International Commission on Large Dams
(**http://www.icold-cigb.org/**)
ICRS – International Commission on Remote Sensing
(**http://hydrolab.arsusda.gov/~jritchie/**)
ICSI – International Commission on Snow and Ice
(**http://www.glaciology.su.se/ICSI/**)
ICSU – International Council for Science
(**http://www.icsu.org/index.php**)
ICSW – International Commission on Surface Water
ICT – International Commission on Tracers
ICWQ – International Commission on Water Quality
ICWRS – International Commission on Water Resources Systems
IETC – International Environmental Technology Center
(**http://www.unep.or.jp/**)
IGRAC – International Groundwater Resources Assessment Center
(**http://igrac.nitg.tno.nl/homepage.html**)
IGBP – International Geosphere Biosphere Programme
(**http://www.igbp.kva.se/cgi-bin/php/frameset.php**)
IGOS-P – Integrated Global Observing Strategy – Partnerships
(**http://ioc.unesco.org/igospartners**)
IGRAC – International Groundwater Resources Assessment Centre
(**http://igrac.nitg.tno.nl/homepage.html**)
IGWMC – International Groundwater Modeling Center
(**http://www.mines.edu/igwmc/**)
IGWCO – Integrated Global Water Cycle Observing theme (IGOS)
(**http://ioc.unesco.org/igospartners/Water.htm**)
IH – Institute of Hydrology
IHD – International Hydrological Decade
IHDP – International Human Dimensions Programme on Global Environmental Change
(**http://www.ihdp.uni-bonn.de/**)
IHE – Institute for Infrastructural, Hydraulic and Environmental Engineering
IHP – International Hydrological Programme
(**http://www.unesco.org/water/ihp/index.shtml**)
iLEAPS – Integrated Land Ecosystem-Atmosphere Process Study
(**http://www.atm.helsinki.fi/ILEAPS/**)
INBO – International Network of Basin Organizations
(**http://www.riob.org/**)
INQUA – International Union for Quaternary Research
INWEH – International Network on Water, Environment and Health
(**http://www.inweh.unu.edu/inweh/**)
INWRDAM – Inter-Islamic Network on Water Resources Development and Management
(**http://amon.nic.gov.jo/inwrdam/**)
IOC – Intergovernmental Oceanographic Commission
(**http://oceanportal.org/**)
IPCC – Intergovernmental Panel on Climate Change
(**http://www.ipcc.ch/**)
IRTCES – International Research and Training Center on Erosion and Sedimentation
ISARM – International Shared Aquifer Resources Management
ISDR – International Strategy for Disaster Reduction
ISLSCP – International Satellite Land Surface Climatology Project
(**http://www.gewex.org/islscp.html**)
ISSC-International Social Science Council
(**http://www.unesco.org/ngo/issc/**)
IRN – International Rivers Network
(**http://www.irn.org/index.asp?id=/basics/about.html**)
IUGG – International Union of Geodesy and Geophysics
(**http://www.iugg.org/**)
IUCN-The World Conservation Union
(**http://www.iucn.org/**)
IWA – International Water Association
(**http://www.iwahq.org.uk/template.cfm?name=home**)
IWE – Institute for Water Education
(**http://www.ihe.nl/vmp/contentsHomePage.html**)
IWHA – International Water History Association
(**http://www.iwha.net/**)

IW – LEARN – International Waters Learning Exchange and Resource Network (**http://www.iwlearn.org/**)

IWMI – International Water Management Institute (**http://www.iwmi.cgiar.org/**)

IWRA – International Water Resources Association (**http://www.iwra.siu.edu/about/index.html**)

IWRM – Integrated Water Resource Management

IPO – International Project Office

JAXA – Japan Aerospace Exploration Agency

JIIHP – Joint International Isotopes in Hydrology Program (**http://www.unesco.org/water/ihp/isotopes.shtml**)

JMP – Joint Monitoring Programme (for Water Supply and Sanitation) (**http://www.wssinfo.org/en/welcome.html**)

LBA – Large-scale Biosphere Atmosphere Experiment in Amazonia

LOICZ – Land Ocean Interactions in the Coastal Zone (**http://wwwold.nioz.nl/loicz/**)

LPB – La Plata Basin Project

LUCC – Land Use and Land Cover Change

MAGS – Mackenzie GEWEX Study (**http://www.usask.ca/geography/MAGS/**)

MAIRS – Monsoon Asia Integrated Regional Study

MDB – Murray Darling Basin

MDG – Millennium Development Goal

MEWIN – Middle East Water Information Network (**http://water1.geol.upenn.edu/**)

MOPEX – Model Parameter Estimation Experiment (**http://www.nws.noaa.gov/oh/mopex/**)

MSFC – Marshall Space Flight Center (**http://www.msfc.nasa.gov/**)

NAS – National Academy of Science

NASA – National Aeronautics and Space Administration (**http://www.nasa.gov/home/**)

NIES – National Institute for Environmental Studies (**http://www.nies.go.jp/index-j.html**)

NOAA – National Oceanic and Atmospheric Administration (**http://www.noaa.gov/**)

OIC – Organization of the Islamic Conference

PAGES – Past Global Changes (**http://www.pages.unibe.ch/about/about.html**)

PRISM – Programme for Integrated Earth System Modeling

PUB – Prediction in Ungauged Basins (**http://cee.uiuc.edu/research/pub/default.asp.**)

RLIWG – River and Lake Ice Working Group

RRC – Regional Research Center

RRN – Regional Research Network

RRS – Regional Research Site

SCOWAR – Scientific Committee on Water Resources

SCWG – Snow and Climate Working Group

SDNP – Sustainable Development Networking Programme

SEI – Stockholm Environment Institute (**http://www.sei.se/water/overview.html**)

SIDA – Swedish International Development Agency (**http://www.sida.se/Sida/jsp/polopoly.jsp?d=107**)

SIDS – Small Island Developing States (**http://www.sidsnet.org/index.html**)

SIL – Societas Internationalis Limnologiae

SNOWMIP – Snow Models Inter-comparison Project

SSTA – Space Science and Technology Alliance

START – Global Change System for Analysis, Research and Training (**http://www.start.org/**)

SVIWG – Snow–Vegetation Interactions Working Group

THORPEX – The Observing System Research and Predictability Experiment (**http://www.mmm.ucar.edu/uswrp/programs/thorpex.html**)

TOPC – Terrestrial Observation Panel for Climate (**http://www.fao.org/gtos/TOPC.html**)

UCOWR – University Council on Water Resources (**http://www.uwin.siu.edu/ucowr/index.html**)

UN – United Nations (**http://www.un.org**)

UNCCD – United Nations Convention to Combat Desertification

UNDP – United Nations Development Programme (**http://www.undp.org**)

UNEP – United Nations Environmental Programme (freshwater programs: **http://www.unep.org/unep/program/natres/water/fwu/home.htm**)

UNESCO – United Nations Education, Scientific and Cultural Organization (**http://www.unesco.org**)

UNFCCC – United Nations Framework Convention on Climate Change (**http://unfccc.int/**)

UNICEF-United Nations International Children's Emergency Fund (**http://www.unicef.org/**)

UNU – United Nations University

USEPA – United States Environmental Protection Agency (**http://www.epa.gov/**)

USGS – United States Geological Survey (**http://www.usgs.gov/**)

USRA – Universities Space Research Association (**http://www.usra.edu/**)

WCMC – World Conservation Modeling Center (**http://www.unep-wcmc.org/**)
WCP – World Climate Programme (**http://www.wmo.ch/web/wcp/wcp-home.html**).
WCRP – World Climate Research Programme (**http://www.wmo.ch/web/wcrp/wcrp-home.html**)
WEAP – Water Evaluation and Assessment Planning (**http://www.weap21.org/**)
WEDC – Water Engineering and Development Center (**http://wedc.lboro.ac.uk/index.htm**)
WGCM – Working Group on Coupled Modelling (**http://www.wmo.ch/web/wcrp/wgcm.htm**)
WGMS – World Glacier Monitoring Service (**http://www.geo.unizh.ch/wgms/**)
WHO – World Health Organization (Water, Sanitation and Health: **http://www.who.int/water_sanitation_health/en/**)
WHYCOS – World Hydrological Cycle Observing System (**http://www.wmo.ch/web/homs/projects/whycos.html**)
WMO – World Meteorological Organization (**http://www.wmo.ch/indexflash.html**)
WRAP – Water Resources Application Project
WRCSEAP – Regional Humid Tropics Hydrology and Water Resources Center for South-East Asia and the Pacific
WRR – World Resources Report
WSHP – Water, Sanitation and Health Program
Worldwatch – Worldwatch Institute (**http://www.worldwatch.org/**)
WWAP – World Water Assessment Project (**http://www.unesco.org/water/wwap/**)
WWC – World Water Council (**http://www.worldwatercouncil.org/**)
WWDR – World Water Development Report (**http://www.unesco.org/water/wwap/wwdr/index.shtml**)
WWF – World Water Forum
WWR – World Wildlife Fund (Freshwater Initiative: **http://www.panda.org/about_wwf/what_we_do/freshwater/index.cfm**)
WWRP – World Weather Research Programme (**http://box.mmm.ucar.edu/uswrp/wwrp/wwrp.html**)
WWW – World Weather Watch (**http://www.wmo.ch/web/www/www.html**))

FURTHER READING

Global Water Partnership (2000) *Making Every Drop Count, Sustainable Development International: Strategies and Technologies for Agenda 21 Implementation*, ICG Publishing.

Van den Heuvel M. and Willemse E. (2001) Achieving water security: making water everybody's business. *Sustainable Development International: Strategies and Technologies for Agenda 21 Implementation*.

REFERENCES

Ad Hoc Group on Global Water Data Sets of the International Association of Hydrological Sciences (Vorosmarty C., Askew A., Grabs W., Barry R.G., Birkett C., Doll P., Goodison B., Hall A., Jenne R., Kitaev L., *et al.*) (2001) Global water data: a newly endangered species. *EOS Transactions of the American Geophysical Union*, **82**(5), 54–58.

Boehmer K., Memon A. and Mitchell B. (2001) Towards sustainable water management in southeast Asia: experiences from Indonesia and Malaysia. *Water Resources Journal of the UN Economic and Social Commission for Asia and the Pacific*, 1–30.

Cihlar J., Grabs W. and Landwehr J. (2000) *Establishment of a Global Hydrological Observation Network for Climate: Report of the GCOS/GTOS/HWRP Expert Meeting*, GCOS Report 63, GTOS Report 26, WMO/TD-No.1047, WMO, Geneva, p. 93, Internet availability at **http://www.wmo.ch/web/homs/geisenheim.pdf** and **http://www.fao.org/GTOS/gtospub26.htm**.

Gleick P.H. and Adams B. (2000) *Water: the Potential Consequences of Climate Variability and Change for the Water Resources of the United States*, USGCRP, p. 151.

Gleick P.H., Burns W.C.G., Chalecki E.L., Cohen M., Cushing K.K., Mann A.S., Reyes R., Wolff G.H. and Wong A.K. (2002) *The World's Water 2002–2003*, Biennial Report on Freshwater Resources, Island Press.

Hornberger G.M., Aber J.D., Bahr J., Bales R.C., Bevan K., Foufoula-Georgiou E., Katulo G., Kinter J.L. III, Koster R.D. and Lettenmaier D.P. and (2001) *A Plan for a New Science Initiative on the Global Water Cycle*, US Water Cycle Study Group of the US Global Change Research Program (USGCRP), USGCRP Office: Washington.

Integrated Global Observing Strategy (IGOS) Partnership (2004) *A Global Water Cycle Theme for the IGOS Partnership*, European Space Agency, p. 100.

Intergovernmental Panel on Climate Change (2001) *Climate Change 2001, The Scientific Basis*, Cambridge University Press, p. 443.

Koike T. (2004) *The Coordinated Enhanced Observing Period – An Initial Step for Integrated Global Water Cycle Observation*, WMO Bulletin 53, No. 2, WMO: Geneva.

Lawford R.G., Landwehr J.M., Sorooshian S. and Whitaker M.P.L. (2003) International hydrological science programs and global water issues. Chapter 11. In *Water: Science, Policy and Management*, Lawford R., Fort D., Hartmann H. and Eden S. (Eds.), AGU: Washington, pp. 223–246.

Lawford R.G., Stewart R., Roads J., Isemer H.-J., Manton M., Marengo J., Yasunari T., Benedict S., Koike T. and Williams S. (2004) Advancing global and continental-scale hydrometeorology: contributions of the GEWEX hydrometeorology panel. *Bulletin of the American Meteorological Society*, **85**(12), 1917–1930.

National Academy of Sciences (1991) *Opportunities in the Hydrologic Sciences*, NAS Press: Washington, p. 348.

Revenga C., Brunner J., Henninger N., Kassem K. and Payne R. (2000) *Pilot Analysis of Global Freshwater Ecosystems (PAGE)*, World Resources Institute: Washington.

Stewart R., Leese J. and Koike T. (2001) *CEOP Science Plan and Overall Strategy*, **http://www.usask.ca/geography/MAGS/GHP/ceop_issues.html**.

UNESCO-WWAP (2003) *Water for People, Water for Life*, The United Nations World Water Development Report, UNESCO and Berghahn Books: Barcelona, p. 576.

World Climate Research Programme (1984) *Scientific Plan for the World Climate Research Programme*, WCRP Publication No. 2, WMO/TD-No. 6, World Meteorological Organization: Geneva.

World Climate Research Programme (1990) *Scientific Plan for the Global Energy and Water Cycle Experiment*, WCRP Publication No. 40, WMO/TD-No. 376, World Meteorological Organization: Geneva.

World Commission on Environment and Development (1987) *Our Common Future*, Oxford University Press: Oxford and New York.

Index